การพัฒนาระบบตรวจสอบคุณภาพข้อมูล

ศักดา เลิศพิพัฒน์วาณิชย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)
คณะสถิติประยุกต์
สถาบันบัณฑิตพัฒนบริหารศาสตร์
2564

การพัฒนาระบบตรวจสอบคุณภาพข้อมูล

ศักดา เลิศพิพัฒน์วาณิชย์

คณะสถิติประยุกต์

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.ปรีชา วิจิตรธรรมรส)

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาแล้วเห็นสมควรอนุมัติให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร.เดือนเพ็ญ ธีรวรรณวิวัฒน์)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ปรีชา วิจิตรธรรมรส)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.อานนท์ ศักดิ์วรวิชญ์)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.รมิดา ศรีเหรา)

..... คณบดี

(ผู้ช่วยศาสตราจารย์ ดร.ปราโมทย์ ลือนาม)

_____/_____/_____

บทคัดย่อ

ชื่อวิทยานิพนธ์	การพัฒนาระบบตรวจสอบคุณภาพข้อมูล
ชื่อผู้เขียน	ศักดา เลิศพิพัฒน์วาณิชย์
ชื่อปริญญา	วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)
ปีการศึกษา	2564

ตั้งแต่แต่ดีตจนถึงปัจจุบัน นักวิจัยได้มีการศึกษาหาแนวทางในการแก้ปัญหาคุณภาพของข้อมูลที่ไม่มีคุณภาพมาอย่างต่อเนื่องแต่ก็ยังไม่สามารถจัดการปัญหาดังกล่าวได้ เนื่องจากความก้าวหน้าด้านเทคโนโลยีรสารสนเทศอย่างก้าวกระโดด ทำให้เกิดการเปลี่ยนถ่ายยุคของข้อมูลแบบเดิมไปสู่ยุคของข้อมูลขนาดใหญ่ ที่ข้อมูลมาจากหลากหลายแหล่ง หลากหลายประเภท และมีจำนวนมหาศาล การเปลี่ยนแปลงนี้ทำให้การบริหารคุณภาพของข้อมูล ทำได้ยาก และซับซ้อนมากยิ่งขึ้น ทำให้ต้องใช้แรงงานและเวลาจำนวนมาก อีกทั้งยังต้องอาศัยความรู้ความเข้าใจในการตรวจสอบคุณภาพของข้อมูล ส่งผลให้การตรวจสอบคุณภาพข้อมูลมักจะถูกละเลย จนก่อให้เกิดผลกระทบเชิงลบต่าง ๆ ตามมาภายหลัง

การทดแทนแรงงานในการตรวจสอบคุณภาพข้อมูลโดยการใช้ระบบคอมพิวเตอร์เป็นอีกทางเลือกหนึ่งที่จะช่วยให้ลดการใช้แรงงานและความซับซ้อนในการตรวจสอบคุณภาพข้อมูล งานวิจัยนี้จึงมีแนวคิดในการออกแบบและพัฒนาระบบตรวจสอบคุณภาพข้อมูล โดยพัฒนาโดยใช้ภาษา Python ซึ่งเป็นภาษาที่ใช้กันอย่างแพร่หลายในแวดวงวิศวกรรมในปัจจุบัน โดยผลสำรวจความพึงพอใจในการใช้งานระบบตรวจสอบคุณภาพข้อมูลแสดงให้เห็นว่าผู้ใช้งานมีความพึงพอใจ เนื่องด้วยสามารถช่วยลดเวลาในการทำงานในการตรวจสอบคุณภาพข้อมูลได้เป็นอย่างดี อีกทั้งยังช่วยเพิ่มคุณภาพของงานได้ดียิ่งขึ้นอีกด้วย

ABSTRACT

Title of Thesis	The development of data quality audit system
Author	Sakda Loetpipatwanich
Degree	Master of Science (Applied Statistics)
Year	2021

There is much research in the last decade, but researchers have continued to study solutions to the problem of poor data quality but have not been able to deal with it. Due to the leap in information technology advancement causing the transition of the traditional data age to the age of Big Data where information comes from a variety of sources, various types, and there are enormous numbers. This change made data quality management more difficult and complex which cause a lot of labor and time. It also requires knowledge and understanding in examining the quality of information. As a result, data quality checks are often ignored until causing various negative effects afterward.

The replacement of labor in data quality auditing by using a computerized system is another option to reduce the labor and complexity of performing data quality checks. Therefore, this research has the concept of designing and developing an information quality auditing system. It was developed using Python, the most widely used programming language in engineering today. The results of the satisfaction survey on the system showed that the users were satisfied because the

system can greatly reduce the time spent on data quality checks and also improves the quality of work even further

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้เนื่องจากได้รับความกรุณาอย่างสูงจาก ผศ. ดร.ปรีชา วิจิตรธรรมรส อาจารย์ที่ปรึกษางานวิจัย ที่กรุณาให้คำแนะนำปรึกษา ตลอดจนปรับปรุงแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่อย่างดี อีกทั้งยังให้ข้อคิดที่เป็นประโยชน์ตลอดมา จนวิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ ผู้เขียนจึงใคร่ขอขอบพระคุณอาจารย์เป็นอย่างสูงด้วยความเคารพยิ่ง

ขอขอบพระคุณ รศ. ดร.เดือนเพ็ญ ธีรวรรณวิวัฒน์ และ ผศ. ดร.รมิดา ศรีเหรา ที่สละเวลาเป็นกรรมการสอบวิทยานิพนธ์ ช่วยให้คำแนะนำและพิจารณาตรวจสอบให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์ยิ่งขึ้น ขอขอบพระคุณ ผศ. ดร.อานนท์ ศักดิ์วรวิชญ์ ที่สละเวลาเป็นกรรมการสอบวิทยานิพนธ์และให้ข้อคิดเห็นในการตรวจสอบคุณภาพข้อมูล ขอขอบพระคุณคณาจารย์ทุกท่านที่ได้ถ่ายทอดวิชาความรู้อันเป็นประโยชน์ยิ่งให้แก่ผู้เขียน ขอบคุณเจ้าหน้าที่คณะสถิติประยุกต์ทุกท่านที่ให้ความช่วยเหลือในการติดต่อประสานงานในเรื่องต่างๆ

สุดท้ายนี้ ผู้วิจัยขอขอบคุณครอบครัวที่เป็นกำลังใจให้เสมอ คือน้องพัชรผู้เป็นภรรยาที่ให้ความห่วงใย ช่วยเป็นกำลังใจ ส่งเสริม สนับสนุนทุกด้าน และเป็นแรงบัลดาลใจอันสำคัญตลอดมา จนประสบผลตามที่ตั้งใจ ทำให้การศึกษาครั้งนี้สำเร็จลุล่วงได้ตามที่ตั้งใจทุกประการ

ศักดา เลิศพิพัฒน์วาณิชย์

ธันวาคม 2564

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ปัญหาคุณภาพข้อมูลที่ไม่ดี เป็นปัญหาที่พบมาตั้งแต่ยุคแรกของการใช้ข้อมูลในการประกอบการตัดสินใจ ยิ่งในยุคปัจจุบันข้อมูลยิ่งมีความสำคัญมากยิ่งขึ้น ทำให้การปรับปรุงคุณภาพข้อมูลในองค์กรถือเป็นประเด็นหลักที่องค์กรต่าง ๆ ให้ความสนใจ เนื่องการพบว่าข้อมูลที่มีคุณภาพนั้น มีอิทธิพลที่สำคัญต่อประสิทธิภาพขององค์กร (Gorla, Somers, & Wong, 2010) การตัดสินใจที่มีคุณภาพย่อมเกิดจากแหล่งข้อมูลที่มีคุณภาพ (Janssen, van der Voort, & Wahyudi, 2017) นอกจากนี้คุณภาพข้อมูลที่ไม่ดียังส่งผลถึงระบบปัญญาประดิษฐ์ที่เริ่มมีการนำมาใช้อย่างแพร่หลายในปัจจุบัน โดยงานวิจัยพบว่าปัญหาคุณภาพข้อมูลก่อให้เกิดข้อจำกัดในเรื่องประสิทธิภาพการทำงานของปัญญาประดิษฐ์ ซึ่งก่อให้เกิดผลกระทบร้ายแรงได้ (Slota et al., 2020) อีกทึ่งยังพบว่าในกรณีของ Amazon Mechanical Turk ที่เป็นแพลตฟอร์มสำหรับทำ Crowdsourcing ซึ่งเป็นแหล่งข้อมูลที่ใช้ในการศึกษาวิจัยด้านจิตวิทยา ได้ตรวจพบว่าข้อมูลที่เข้าสู่ระบบมีคุณภาพที่แย่ลงเรื่อย ๆ ส่งผลต่องานวิจัยด้านจิตวิทยา เนื่องจากนักวิจัยเชื่อว่าข้อมูลที่มาจาก Amazon Mechanical Turk เป็นข้อมูลที่มีคุณภาพ จึงละเลยการตรวจสอบคุณภาพของข้อมูลก่อนนำไปใช้ในงานวิจัย ทำให้งานวิจัยที่ได้นั้นผิดเพี้ยนไปจากความเป็นจริง (Chmielewski & Kucker, 2020) นอกจากนี้ยังมีการศึกษาวิจัยผลกระทบจากคุณภาพข้อมูลที่ไม่ดี โดยพบว่าโครงการบูรณาการข้อมูล ร้อยละ 88 ล้มเหลว หรืองบประมาณบานปลายอย่างมีนัยสำคัญ ในขณะที่องค์กรร้อยละ 33 มีการยกเลิก หรือเลื่อนการติดตั้งระบบสารสนเทศ เพราะปัญหาคุณภาพข้อมูล นอกจากนี้ยังมีการประเมินการสูญเสียค่าใช้จ่ายประมาณ 1,800 ล้านบาทต่อปี ในสหรัฐอเมริกาในการใช้ข้อมูลกำหนดกลุ่มเป้าหมายที่ผิดพลาด นอกจากนี้ปัญหาด้านคุณภาพข้อมูลเป็นปัญหาหลักของการล้มเหลวในระบบบริหารลูกค้าสัมพันธ์ ขณะที่บริษัทน้อยกว่าร้อยละ 50 ที่มั่นใจว่าข้อมูลของต้นเองมีคุณภาพ และมีบริษัทเพียงร้อยละ 15 ที่มีความมั่นใจในคุณภาพของข้อมูลที่ได้จากภายนอก นอกจากนี้ยังพบว่าองค์กรทั่วไปมักประเมินคุณภาพของข้อมูลดีเกินไป แต่ประเมินค่าใช้จ่ายจากความผิดพลาดน้อยเกินไป (Haug, Zachariassen, & Liempd, 2011)

จากปัญหาคุณภาพข้อมูลดังที่กล่าวมาข้างต้น นักวิจัยได้มีการศึกษาหาแนวทางในการแก้ปัญหาอย่างต่อเนื่อง โดยในงานวิจัยสามารถพบตัวชี้วัดคุณภาพข้อมูลได้มากถึง 48 ตัวชี้วัด กระจายในหลากหลายมิติ (Sebastian-Coleman, 2013) อันเป็นผลจากความก้าวหน้าด้านเทคโนโลยีสารสนเทศอย่างก้าวกระโดด ที่ข้อมูลมาจากหลากหลายแหล่ง หลากหลายประเภท และมีจำนวนมหาศาล ทำให้การบริหารคุณภาพของข้อมูลซับซ้อนมากยิ่งขึ้น ทำให้ต้องใช้แรงงานและเวลาจำนวนมาก อีกทั้งยังต้องอาศัยความรู้ความเข้าใจในการตรวจสอบคุณภาพของข้อมูลในมิติต่างๆ อีกทั้งระบบตรวจสอบคุณภาพข้อมูลมักพบในระบบฐานข้อมูลเชิงการค้าที่มีราคาสูง ทำให้นักวิจัยหรือผู้ใช้งานทั่วไปเข้าถึงระบบดังกล่าวได้ยาก ส่งผลให้การตรวจสอบคุณภาพข้อมูลมักจะถูกละเลย จนก่อให้เกิดผลกระทบเชิงลบตามมาภายหลัง

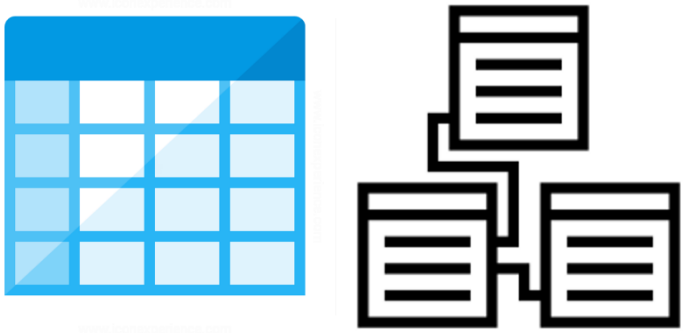
จากที่กล่าวมาข้างต้น งานวิจัยนี้จึงมีแนวคิดในการคัดเลือกตัวชี้วัดคุณภาพข้อมูลพื้นฐานที่ทำได้ง่ายในเชิงปฏิบัติ และใช้ได้กับทุกชุดข้อมูล แล้วจึงนำมาออกแบบและพัฒนาระบบตรวจสอบคุณภาพข้อมูล โดยพัฒนาโดยใช้ภาษา Python ซึ่งเป็นภาษาที่ใช้กันอย่างแพร่หลายในแวดวงวิศวกรรมในปัจจุบัน เนื่องจากเป็นภาษาที่อ่านเข้าใจได้ง่าย ขนาดเล็ก และมีส่วนเสริมต่าง ๆ ให้เลือกใช้งานจำนวนมาก (Srinath, 2017) ผู้วิจัยมีความคาดหวังที่จะสร้างทางเลือกในการตรวจสอบคุณภาพข้อมูลให้แก่ผู้ทำงานด้านข้อมูล เพื่อพัฒนาคุณภาพของสารสนเทศที่เกิดจากโครงการด้านข้อมูลต่าง ๆ

1.2 วัตถุประสงค์ของการศึกษา

1. เพื่อศึกษาและคัดเลือกตัวชี้วัดคุณภาพของชุดข้อมูลอย่างง่าย
2. เพื่อออกแบบและพัฒนาระบบตรวจสอบคุณภาพชุดข้อมูล
3. เพื่อทดสอบประสิทธิภาพระบบตรวจสอบคุณภาพชุดข้อมูล

1.3 ขอบเขตการศึกษา

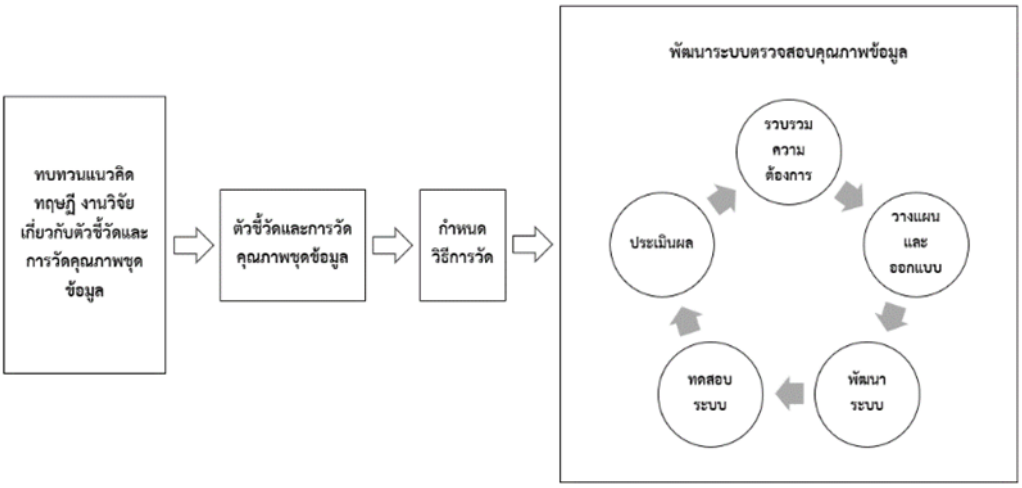
ในการศึกษานี้จะทำการศึกษาวิธีการวัดคุณภาพชุดข้อมูลในมิติต่าง ๆ เพื่อให้ได้วิธีการวัดที่สามารถนำมาใช้ในการตรวจสอบคุณภาพของชุดข้อมูลที่ใช้ได้ในเชิงปฏิบัติ โดยชุดข้อมูลที่ศึกษานี้จะใช้ชุดข้อมูลเชิงเดี่ยวหรือชุดข้อมูล 1 ตาราง ไม่รวมถึงชุดข้อมูลเชิงสัมพันธ์ โดยระบบที่พัฒนาในเบื้องต้นจะรองรับระบบปฏิบัติการ MacOS และ Linux ซึ่งผลลัพธ์สามารถแสดงผลได้ทั้งในรูปแบบรายงาน และรูปแบบที่คอมพิวเตอร์สามารถเข้าใจได้เพื่อนำไปใช้ในกระบวนการต่าง ๆ ในด้านวิศวกรรมข้อมูลต่อไป



ภาพที่ 1.1 ชุดข้อมูลเชิงเดี่ยว(ซ้าย) ชุดข้อมูลเชิงสัมพันธ์(ขวา)

1.4 กรอบแนวคิดของการศึกษา

ในการดำเนินการวิจัย ผู้วิจัยมุ่งเน้นศึกษาตัวชี้วัดคุณภาพข้อมูลของชุดข้อมูลเชิงเดี่ยว โดยหลังจากได้ตัวชี้วัดแล้วจึงนำไปพัฒนาเป็นระบบตรวจสอบคุณภาพข้อมูล ซึ่งมีรายละเอียด ดังภาพที่ 1.2



ภาพที่ 1.2 กรอบแนวคิดของการศึกษา

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับในการศึกษาครั้งนี้ คือ

1. มีตัวชี้วัด และวิธีการวัดคุณภาพชุดข้อมูลเชิงเดี่ยวในเชิงปฏิบัติ

2. เป็นแนวทางในการพัฒนาระบบตรวจสอบคุณภาพข้อมูล

3. เพิ่มช่องทางให้แก่ผู้ที่ต้องทำงานด้านข้อมูลเข้าถึงระบบในการบริหารจัดการคุณภาพข้อมูลได้ง่ายขึ้น และสามารถปรับแต่งให้เหมาะสมกับความต้องการของผู้ใช้งานได้

บทที่ 2

งานวิจัยที่เกี่ยวข้อง

ในบทนี้ผู้วิจัยได้นำเสนอความคิดที่ได้จากการศึกษา เอกสาร วารสาร บทความทางวิชาการ เพื่อใช้เป็นกรอบแนวคิดในการศึกษาวิจัย โดยมุ่งเน้นไปที่แนวคิดเกี่ยวกับการวัดคุณภาพข้อมูล

2.1 ทฤษฎีที่เกี่ยวข้อง

Wang และ Strong (Wang & Strong, 1996) ได้พัฒนากรอบการทำงาน และเรียบเรียงมิติคุณภาพข้อมูลที่สำคัญ โดยเป็นการรวบรวมข้อมูลจากผู้ใช้งานข้อมูล แทนที่จะเป็นข้อมูลในเชิงทฤษฎี โดยในกรอบการทำงานนี้ ได้จัดมิติคุณภาพข้อมูลไว้ใน 4 ด้าน ได้แก่ ด้านของคุณภาพโดยเนื้อแท้ของข้อมูล (Instinct Data Quality) ด้านของคุณค่าของตัวข้อมูลเอง (Contextual Data Quality) ด้านของการเป็นตัวแทนของข้อมูล (Representational Data Quality) และด้านของการเข้าถึงได้ของข้อมูล (Accessibility Data Quality) โดยมีรายเอียดดังนี้

1) ด้านของคุณภาพโดยเนื้อแท้ของข้อมูล (Instinct Data Quality) ประกอบด้วยมิติคุณภาพข้อมูล ดังนี้

1.1) มิติด้านความแม่นยำ

1.2) มิติด้านความตรงประเด็น

1.3) มิติด้านความน่าเชื่อถือ

1.4) มิติด้านชื่อเสียง

2) ด้านของคุณค่าของตัวข้อมูลเอง (Contextual Data Quality) ประกอบด้วยมิติคุณภาพข้อมูล ดังนี้

2.1) มิติด้านการสร้างมูลค่าเพิ่ม

2.2) มิติด้านความสัมพันธ์กัน

2.3) มิติด้านความทันเวลา

2.4) มิติด้านความสมบูรณ์

2.5) มิติด้านความเหมาะสมของจำนวนข้อมูล

3) ด้านของการเป็นตัวแทนของข้อมูล (Representational Data Quality) ประกอบด้วยมิติคุณภาพข้อมูล ดังนี้

3.1) มิติด้านสามารถในการตีความ

3.2) มิติด้านความง่ายต่อความเข้าใจ

3.3) มิติด้านการเป็นตัวแทนที่มีความคงเส้นคงวา

3.4) มิติด้านการเป็นตัวแทนที่มีความกระชับ

4) ด้านของการเข้าถึงได้ของข้อมูล (Accessibility Data Quality) ประกอบด้วยมิติคุณภาพข้อมูล ดังนี้

4.1) มิติด้านการเข้าถึงได้ของข้อมูล

4.2) มิติด้านความปลอดภัยในการเข้าถึง

Redman(1997) ได้กล่าวถึงมิติที่เกี่ยวข้องกับคุณภาพของข้อมูลดังนี้

1) ด้านตัวแบบข้อมูล

1.1) มิติด้านเนื้อหา

1.2) มิติด้านความเกี่ยวข้องกับข้อมูล

1.3) มิติด้านความสามารถในการได้คุณค่า

1.4) มิติด้านความชัดเจนของความหมาย

2) ด้านระดับของรายละเอียด

2.1) มิติด้านความเป็นหน่วยย่อย

2.2) มิติด้านความแม่นยำของขอบเขตคุณสมบัติ

3) ด้านองค์ประกอบ

3.1) มิติด้านความเป็นธรรมชาติ คุณลักษณะของข้อมูลควรมีความง่าย ซึ่งสอดคล้องกับความเป็นจริง และแต่ละคุณสมบัติแต่ละตัวควรสะท้อนความเป็นจริงอย่างใดอย่างหนึ่ง

3.2) มิติด้านความเป็นลักษณะเฉพาะ แต่ละคุณสมบัติควรจะสามารถแยกแยะ และแตกต่างจากคุณสมบัติตัวอื่น

3.3) มิติด้านเป็นเนื้อเดียวกันของข้อมูล

3.4) มิติด้านความจำเป็นขั้นต่ำซ้ำซ้อน

4) ด้านความคงเส้นคงวา

4.1) มิติด้านความคงเส้นคงวาเชิงระบบ ขององค์ประกอบต้นแบบ

4.2) มิติด้านความคงเส้นคงวาเชิงโครงสร้างของคุณลักษณะระหว่างประเภทเอกลักษณ์

5) ด้านปฏิกิริยาต่อการเปลี่ยนแปลง

5.1) มิติด้านความเข็มแข็ง

5.2) มิติด้านความหยืดหยุ่น

6) ด้านคุณค่าของข้อมูล

6.1) มิติด้านความแม่นยำ

6.2) มิติด้านความสมบูรณ์

6.3) มิติด้านความเป็นปัจจุบัน

6.4) มิติด้านความแน่นอน

7) ด้านการเป็นตัวแทนของข้อมูล

7.1) มิติด้านความเหมาะสม

7.2) มิติด้านความสามารถในการตีความความสามารถในการปรับเปลี่ยน

7.3) มิติด้านความแม่นยำของรูปแบบ

7.4) มิติด้านความหยืดหยุ่นของรูปแบบ

7.5) มิติด้านความสามารถที่เป็นตัวแทนของค่าว่าง

7.6) มิติด้านประสิทธิภาพการใช้พื้นที่เก็บข้อมูล

7.7) มิติด้านตัวอย่างทางกายภาพ สอดคล้องกับรูปแบบที่ข้อมูลเป็น

Batini และคณะ (2009) ได้ศึกษาทบทวนวรรณกรรมที่เกี่ยวข้องกับการประเมินและปรับปรุงคุณภาพข้อมูลรวมถึงงานของ Wang (1996) และ Redman (1997) โดยได้บรรยายสรุปถึงมิติคุณภาพ ตัวชี้วัด และการวัด โดยมีรายละเอียด ดังนี้

1) มิติด้านความแม่นยำ (Accuracy)

ตัวชี้วัด 1: ความแม่นยำของโครงสร้าง = $\frac{\text{จำนวนข้อมูลที่ถูกต้อง}}{\text{จำนวนข้อมูลทั้งหมด}}$

ตัวชี้วัด 2: จำนวนข้อมูลที่ถูกต้องที่ถูกส่งมอบ

2) มิติด้านความสมบูรณ์ (Completeness)

ตัวชี้วัด 1: ความสมบูรณ์ของข้อมูล = $\frac{\text{จำนวนข้อมูลที่ไม่เป็นค่าว่าง}}{\text{จำนวนข้อมูลทั้งหมด}}$

ตัวชี้วัด 2: ความสมบูรณ์ของข้อมูล = $\frac{\text{จำนวนข้อมูลที่ส่งมอบ}}{\text{จำนวนข้อมูลที่คาดหวัง}}$

ตัวชี้วัด 3: สำรวจผู้ใช้งาน – แบบสอบถาม

3) มิติด้านความคงเส้นคงวา (Consistency)

ตัวชี้วัด 1: ความคงเส้นคงวา = $\frac{\text{จำนวนข้อมูลที่คงเส้นคงวา}}{\text{จำนวนข้อมูลทั้งหมด}}$

ตัวชี้วัด 2: ความคงเส้นคงวา =จำนวนการเข้ารหัสที่แตกต่างกัน

4) มิติด้านความทันเวลา (Timeliness)

ตัวชี้วัด 1: สัดส่วนการประมวลผลข้อมูลได้ภายในเวลาที่กำหนดไว้

ตัวชี้วัด 2: สำรวจผู้ใช้งาน - แบบสอบถาม

5) มิติด้านความเป็นปัจจุบัน (Currency)

ตัวชี้วัด 1: วันที่ปรับปรุงข้อมูลล่าสุด

ตัวชี้วัด 2: ความเป็นปัจจุบัน = เวลาที่ร้องขอ - เวลาปรับปรุง

ตัวชี้วัด 3: ความเป็นปัจจุบัน = อายุ + (เวลาส่งมอบ - เวลานำเข้า)

ตัวชี้วัด 4: สำรวจผู้ใช้งาน - แบบสอบถาม

6) มิติด้านการเปลี่ยนแปลง (Volatility)

ตัวชี้วัด 1: ความยาวของช่วงเวลาที่ข้อมูลไม่มีการเปลี่ยนแปลง

7) มิติด้านความมีลักษณะเฉพาะ (Uniqueness)

ตัวชี้วัด 1: จำนวนข้อมูลซ้ำซ้อน

8) มิติด้านความเหมาะสมของจำนวนข้อมูล (Appropriate amount of data)

ตัวชี้วัด 1: จำนวนที่เหมาะสมของข้อมูล = $Min\left(\frac{\text{จำนวนของข้อมูลที่ได้รับ}}{\text{จำนวนของข้อมูลที่ต้องการ}}\right)$

ตัวชี้วัด 2: สำรวจผู้ใช้งาน - แบบสอบถาม

9) มิติด้านการเข้าถึง (Accessibility)

ตัวชี้วัด 1: การเข้าถึง = $MAX\left(0; 1 - \left(\frac{\text{เวลาส่งมอบ} - \text{เวลาร้องขอ}}{\text{กำหนดเวลา} - \text{เวลาร้องขอ}}\right)\right.$

ตัวชี้วัด 2: การเข้าถึง = $\frac{\text{จำนวนการเชื่อมโยงที่เสียหาย}}{\text{จำนวนการติดตั้งที่เสียหาย}}$

ตัวชี้วัด 3: สำรวจผู้ใช้งาน – แบบสอบถาม

10) มิติด้านความน่าเชื่อถือ (Credibility)

ตัวชี้วัด 1: จำนวนของข้อมูลที่เป็นค่ามาตรฐาน

ตัวชี้วัด 2: สำรวจผู้ใช้งาน - แบบสอบถาม

11) มิติด้านการตีความ (Interpretability)

ตัวชี้วัด 1: จำนวนข้อมูลที่สามาถตีความได้ หรือ เอกสารอธิบาย

ตัวชี้วัด 2: สำรวจผู้ใช้งาน - แบบสอบถาม

12) มิติด้านประโยชน์ (Usability)

ตัวชี้วัด 1: สำรวจผู้ใช้งาน - แบบสอบถาม

13) มิติด้านการได้มาของข้อมูล (Derivation Integrity)

ตัวชี้วัด 1: เปอร์เซ็นต์ของการคำนวณที่ถูกต้องของข้อมูลที่ได้รับตามสูตรที่ได้มาหรือคำจำกัดความการคำนวณ

14) มิติด้านความกระชับ (Conciseness)

ตัวชี้วัด 1: ความลึกของเพจ

ตัวชี้วัด 2: สำรวจผู้ใช้งาน - แบบสอบถาม

15) มิติด้านการบำรุงรักษา (Maintainability)

ตัวชี้วัด 1: จำนวนเพจที่ไม่มีข้อมูลพื้นฐาน

16) มิติด้านการใช้งาน (Applicability)

ตัวชี้วัด 1: จำนวนเพจที่ใช้งานไม่ได้

17) มิติด้านความสะดวก (Convenience)

ตัวชี้วัด 1: เส้นทางการนำทางที่ใช้งานยาก = $\frac{\text{เส้นทางการนำทางที่สูญหาย}}{\text{การนำทางที่ถูกขัดจังหวะ}}$

18) มิติด้านความเร็ว (Speed)

ตัวชี้วัด 1: ความเร็วในการตอบสมองของเครื่องแม่ข่าย และเครือข่าย

19) มิติด้านความครอบคลุม (Comprehensiveness)

ตัวชี้วัด 1: สำรวจผู้ใช้งาน - แบบสอบถาม

20) มิติด้านความชัดเจน (Clarity)

ตัวชี้วัด 1: สำรวจผู้ใช้งาน - แบบสอบถาม

21) มิติด้านสามารถตรวจสอบได้ (Traceability)

ตัวชี้วัด 1: จำนวนเพจที่ไม่มีผู้แต่ง หรือไม่มีแหล่งที่มี

22) มิติด้านความปลอดภัย (Security)

ตัวชี้วัด 1: จำนวนที่เข้าสู่ระบบแบบไม่ปลอดภัย

ตัวชี้วัด 2: สำรวจผู้ใช้งาน - แบบสอบถาม

23) มิติด้านความถูกต้อง (Correctness)

ตัวชี้วัด 1: สำรวจผู้ใช้งาน - แบบสอบถาม

24) มิติด้านจุดมุ่งหมาย (Objectivity)

ตัวชี้วัด 1: สำรวจผู้ใช้งาน - แบบสอบถาม

25) มิติด้านความสัมพันธ์ (Relevancy)

ตัวชี้วัด 1: สำรวจผู้ใช้งาน - แบบสอบถาม

26) มิติด้านชื่อเสียง (Reputation)

ตัวชี้วัด 1: สำรวจผู้ใช้งาน - แบบสอบถาม

27) มิติด้านความยากง่ายในการปฏิบัติการ (Ease of operation)

ตัวชี้วัด 1: สำรวจผู้ใช้งาน - แบบสอบถาม

28) มิติด้านการมีปฏิสัมพันธ์ (Interactivity)

ตัวชี้วัด 1: จำนวนแบบฟอร์ม – จำนวนเพจเฉพาะบุคคล

Sebastian (2013) ได้กล่าวถึงรายละเอียดของตัวชี้วัดในแต่ละมิติที่สำคัญ โดยมีรายละเอียดตามตารางที่ 2.1

ตารางที่ 2.1 รายละเอียดมิติคุณภาพข้อมูลของ Sebastian

หมายเลข	**มิติคุณภาพ**	**ประเภทการวัด**	**คำอธิบายการวัด**	**วัตถุประสงค์ของการวัด**	**ประเภทของการประเมิน**
1	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-มีคำอธิบายและข้อมูลอ้างอิงเพียงพอ	ประเมินคำอธิบายและข้อมูลอ้างอิงว่ามีเพียงพอหรือไม่	ภาพรวมของเนื้อหาภายในฐานข้อมูล	ประเมินครั้งแรก
2	ความคงเส้นคงวา	ความคงเส้นคงวาของการจัดรูปแบบภายในเขตข้อมูลข้อมูล	ประเมินคุณสมบัติของแต่ละเขตข้อมูลและความคงเส้นคงวาของการจัดรูปแบบภายในเขตข้อมูลข้อมูล	แบบจำลองข้อมูล	ประเมินครั้งแรก
3	ความถูกต้อง/ความคงเส้นคงวา	ความคงเส้นคงวาของการจัดรูปแบบระหว่างตารางข้อมูล	ประเมินคุณสมบัติของแต่ละเขตข้อมูลและความคงเส้นคงวาของการจัดรูปแบบภายในเขตข้อมูลข้อมูลระหว่างตารางข้อมูลที่เป็นประเภทเดียวกัน	แบบจำลองข้อมูล	ประเมินครั้งแรก
4	ความคงเส้นคงวา	ความคงเส้นคงวาในการใช้ค่าตั้งต้นสำหรับเขตข้อมูล	ประเมินคุณสมบัติของแต่ละเขตข้อมูลและค่าตั้งต้นที่ถูกกำหนดให้ในแต่ละเขตข้อมูล	แบบจำลองข้อมูล	ประเมินครั้งแรก

หมายเลข	มิติคุณภาพ	ประเภทการวัด	คำอธิบายการวัด	วัตถุประสงค์ของการวัด	ประเภทของการประเมิน
5	ความถูกต้อง/ความคงเส้นคงวา	ความคงเส้นคงวของค่าตั้งต้นระหว่างข้อมูลตาราง	ประเมินคุณสมบัติของแต่ละเขตข้อมูลและค่าตั้งต้นที่ถูกกำหนดให้ในแต่ละเขตข้อมูลระหว่างตารางข้อมูลที่เป็นเขตข้อมูลประเภทเดียวกัน	แบบจำลองข้อมูล	ประเมินครั้งแรก
6	ความทันเวลา	ข้อมูลถูกประมวลผลและส่งมอบตรงตามเวลาที่กำหนด	เปรียบเทียบจำนวนข้อมูลที่ส่งจริงกับจำนวนข้อมูลที่ต้องส่งตามเวลาที่กำหนด	กระบวนการ/ตรงตามเวลาที่กำหนด	ประเมินครั้งแรก
7	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-ความพร้อมสำหรับประมวลผล	สำหรับไฟล์, ยืนยันว่าเขตข้อมูลทั้งหมดพร้อมสำหรับประมวลผล (ถ้าเป็นไปได้ให้ตรวจสอบรอบของข้อมูลด้วย)	การรับข้อมูล	กระบวนการควบคุม
8	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-นับจำนวนแถวข้อมูล	สำหรับไฟล์, เปรียบเทียบจำนวนแถวข้อมูลที่นับได้กับจำนวนแถวข้อมูลในเอกสารควบคุม	การรับข้อมูล	กระบวนการควบคุม
9	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-สรุปจำนวนเขตข้อมูลทั้งหมด	สำหรับไฟล์, เปรียบเทียบจำนวนเขตข้อมูลที่นับได้กับจำนวนเขตข้อมูลใน	การรับข้อมูล	กระบวนการควบคุม

หมายเลข	มิติคุณภาพ	ประเภทการวัด	คำอธิบายการวัด	วัตถุประสงค์ของการวัด	ประเภทของการประเมิน
			เอกสารควบคุม		
10	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-เปรียบเทียบขนาดข้อมูลในอดีต	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบขนาดข้อมูลที่ได้รับกับขนาดของข้อมูลที่เคยได้รับว่าใกล้เคียงกันหรือไม่	เงื่อนไขการรับข้อมูล	การวัดภายในกระบวนการ
11	ความสมบูรณ์	ความสมบูรณ์ของแถวข้อมูล-ความยาว	จำนวนความยาวของรายการตรงตามที่กำหนดไว้	เงื่อนไขการรับข้อมูล	การวัดภายในกระบวนการ
12	ความสมบูรณ์	ความสมบูรณ์ของเขตข้อมูล-ไม่มีฟิลที่สามารถเป็นค่าว่างได้	ไม่มีเขตข้อมูลที่ยอมให้เป็นค่าว่างได้	เงื่อนไขการรับข้อมูล	กระบวนการควบคุม
13	ความถูกต้อง/ความคงเส้นคงวา	ความถูกต้องของชุดข้อมูล-ไม่มีข้อมูลซ้ำซ้อน	ข้อมูลสามารถถูกแยกแยะได้ และทำการลบข้อมูลที่ซ้ำซ้อนออก	เงื่อนไขการรับข้อมูล	กระบวนการควบคุม
14	ความถูกต้อง/ความคงเส้นคงวา	ความถูกต้องของชุดข้อมูล-ตรวจสอบความสมเหตุสมผลของแถวข้อมูลที่ซ้ำซ้อน	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบสัดส่วนของจำนวนแถวข้อมูลที่ซ้ำซ้อน กับจำนวนรายรายบันทึกทั้งหมดในชุดข้อมูลก่อนหน้า	เงื่อนไขการรับข้อมูล	กระบวนการควบคุม
15	ความสมบูรณ์	ความสมบูรณ์	ตรวจสอบความสม	เงื่อนไขการ	การวัด

หมายเลข	มิติคุณภาพ	ประเภทการวัด	คำอธิบายการวัด	วัตถุประสงค์ของการวัด	ประเภทของการประเมิน
		ของเนื้อหาเขตข้อมูล-ค่าตั้งต้นจากแหล่งข้อมูล	เกตุสมผล, เปรียบเทียบจำนวนและสัดส่วนของค่าตั้งต้นแถวข้อมูล สำหรับเขตข้อมูลที่สำคัญให้กำหนดเกณฑ์หรือจำนวนในอดีตและสัดส่วน	รับข้อมูล	ภายในกระบวนการ
16	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูลบนพื้นฐานของเงื่อนไขที่กำหนด	ตรวจสอบให้แน่ใจว่าค่าต่ำสุด- สูงสุดของค่าเขตข้อมูลวันที่อยู่ในขอบเขตที่กำหนดสำหรับการนำเข้าข้อมูล	เงื่อนไขการรับข้อมูล	กระบวนการควบคุม
17	ความสมบูรณ์	ความสมเหตุสมผลของชุดข้อมูลบนพื้นฐานของเงื่อนไขที่กำหนด	ตรวจสอบให้แน่ใจว่าค่าต่ำสุด - สูงสุดของค่าวันที่ในเขตข้อมูลค่าวันที่ที่สำคัญสอดคล้องกับเกณฑ์ที่สมเหตุสมผล	เงื่อนไขการรับข้อมูล	การวัดภายในกระบวนการ
18	ความสมบูรณ์	ความสมบูรณ์ของฟิลข้อมูล-การได้รับข้อมูลที่เขตข้อมูลที่สำคัญสูญหาย	ตรวจสอบรายการข้อมูลของเขตข้อมูลที่สำคัญก่อนการประมวลผลรายการข้อมูล	เงื่อนไขการรับข้อมูล	กระบวนการควบคุม
19	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-แถวข้อมูล	นับจำนวนแถวข้อมูลคงเหลือหลังผ่านการประมวลผล รวมทั้ง	การประมวลผลข้อมูล	กระบวนการควบคุม

หมายเลข	มิติคุณภาพ	ประเภทการวัด	คำอธิบายการวัด	วัตถุประสงค์ของการวัด	ประเภทของการประเมิน
		คงเหลือหลังผ่านการประมวลผล	แถวข้อมูลที่ซ้ำซ้อน		
20	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-เหตุผลของการตัดแถวข้อมูลออก	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบจำนวนและสัดส่วนสวนของแถวข้อมูลที่ถูกตัดออกด้วยเหตุจำเพาะกับเกณฑ์ที่กำหนด หรือสัดส่วนที่เกิดขึ้นในอดีต	การประมวลผลข้อมูล	การวัดภายในกระบวนการ
21	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูลหลังผ่านการประมวลผล-สัดส่วนของข้อมูลเข้า-ข้อมูลออก	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบสัดส่วนของข้อมูลเข้า-ข้อมูลออก จากกระบวนการในอดีต	การประมวลผลข้อมูล	การวัดภายในกระบวนการ
22	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูลหลังผ่านการประมวลผล-สัดส่วนของข้อมูลเข้า-จำนวนเขตข้อมูล	จำนวนเขตข้อมูลคงเหลือหลังผ่านการประมวลผล สำหรับสถานการณ์ที่สมดุล	การประมวลผลข้อมูล	กระบวนการควบคุม
23	ความสมบูรณ์	ความสมบูรณ์ของเนื้อหาเขตข้อมูล-สัดส่วน	จำนวนเขตข้อมูลตรวจสอบความสมเหตุสมผล,	เนื้อหา/จำนวนคุณลักษณะ	การวัดภายในกระบวนกา

หมายเลข	มิติคุณภาพ	ประเภทการวัด	คำอธิบายการวัด	วัตถุประสงค์ของการวัด	ประเภทของการประเมิน
		ผลรวมของจำนวนเขตข้อมูล	เปรียบเทียบสัดส่วนของผลรวมจำนวนเขตข้อมูลขาเข้าปละขาออกจากกระบวนการจากการประมวลผลชุดข้อมูลในอดีตสำหรับสถานการณ์ที่ไม่สมดุล		ร
24	ความสมบูรณ์	ความสมบูรณ์ของเนื้อหาเขตข้อมูล-ค่าตั้งต้นจากแหล่งที่มา (ประเภทย่อยของข้อ 33 โปรไฟล์หลายเขตข้อมูล)	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบจำนวนและสัดส่วนของแถวข้อมูลตั้งต้น สำหรับเขตข้อมูลที่ได้รับ กับเกณฑ์ที่กำหนด หรือจำนวนหรือสัดส่วนในอดีต	การประมวลผลข้อมูล	การวัดภายในกระบวนการ
25	ความสมบูรณ์	ระยะเวลาในการประมวลผลข้อมูล	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบเวลาที่ใช้ในการประมวลผล กับเวลาที่ใช้ในการประมวลผลในอดีต หรือตามระยะเวลาที่	การประมวลผลข้อมูล	การวัดภายในกระบวนการ

หมายเลข	มิติคุณภาพ	ประเภทการวัด	คำอธิบายการวัด	วัตถุประสงค์ของการวัด	ประเภทของการประเมิน
			กำหนดไว้		
26	ความทันเวลา	ความพร้อมของข้อมูลเมื่อต้องการประมวลผล	เปรียบเทียบจำนวนข้อมูลที่มีอยู่ ณ เวลาที่ต้องประมวลผล กับจำนวนข้อมูลที่ควรจะเป็น	กระบวนการ/ตรงตามเวลาที่กำหนด	การวัดภายในกระบวนการ
27	ความถูกต้อง	ตรวจสอบความถูกต้อง, เขตข้อมูลเดี่ยว, รายละเอียดผลลัพท์	เปรียบเทียบค่าบนข้อมูลที่เข้ามาว่าถูกต้องตามขอบเขตที่กำหนดหรือไม่	เนื้อหา/จำนวนแถวข้อมูล	การวัดภายในกระบวนการ
28	ความถูกต้อง	ตรวจสอบความถูกต้อง, สะสม	รวบรวมผลลัพท์จากการตรวจสอบความถูกต้องทั้งหมด เปรียบเทียบกับผลลัพธ์ในอดีต	เนื้อหาสรุป	การวัดภายในกระบวนการ
29	ความถูกต้อง/ความคงเส้นคงวา	ตรวจสอบความถูกต้อง, หลายเขตข้อมูลภายในตารางข้อมูล, รายละเอียดผลลัพธ์	เปรียบเทียบข้อมูลกับเขตข้อมูลที่มีความสัมพันธ์กันในตารางข้อมูลเดียวกัน	เนื้อหา/จำนวนแถวข้อมูล	การวัดภายในกระบวนการ
30	ความคงเส้นคงวา	ความคงเส้นคงวาของโปรไฟล์เขตข้อมูล	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบการกระจายของค่าในแถวข้อมูล กับข้อมูลใน	เนื้อหา/จำนวนแถวข้อมูล	การวัดภายในกระบวนการ

หมายเลข	มิติคุณภาพ	ประเภทการวัด	คำอธิบายการวัด	วัตถุประสงค์ของการวัด	ประเภทของการประเมิน
			อดีต		
31	ความคงเส้นคงวา	ความคงเส้นคงวาของเนื้อหาในชุดข้อมูล	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบจำนวนค่าที่แตกต่างกันของคุณลักษณะที่ปรากฏในชุดข้อมูล กับข้อมูลในอดีต	สรุปเนื้อหา	การวัดภายในกระบวนการ
32	ความคงเส้นคงวา	ความคงเส้นคงวาของเนื้อหาในชุดข้อมูล, สัดส่วนของจำนวนค่าที่แตกต่างกันของสองคุณลักษณะที่ปรากฏ	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบสัดส่วนระหว่างค่าที่แตกต่างกันของเขตข้อมูลที่สำคัญ กับสัดส่วนในอดีต	สรุปเนื้อหา	การวัดภายในกระบวนการ
33	ความคงเส้นคงวา	ความคงส่วนคงวาของโปรไฟล์แบบหลายเขตข้อมูล	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบค่าการกระจายของข้อมูลของแต่ละเขตข้อมูล	เนื้อหา/จำนวนแถวข้อมูล	การวัดภายในกระบวนการ
34	ความคงเส้นคงวา	ความคงเส้นคงวาของลำดับเหตุการณของกฎเกณฑ์ทางธุรกิจ	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบค่าวันที่กับลำดับเหตุกาณ์ของกฎเกณฑ์ทางธุรกิจ	เนื้อหา/วันที่ของเนื้อ	การวัดภายในกระบวนการ
35	ความคงเส้นคงวา	ความคงเส้นคงวาของเวลาที่ใช้	ตรวจสอบความสมเหตุสมผล,	เนื้อหา/วันที่ของเนื้อ	การวัดภายใน

หมายเลข	มิติคุณภาพ	ประเภทการวัด	คำอธิบายการวัด	วัตถุประสงค์ของการวัด	ประเภทของการประเมิน
			เปรียบเทียบความคงเส้นคงวาของเวลาที่ใช้ในกระบวนการ กับเวลาที่ใช้ในอดีต		กระบวนการ
36	ความคงเส้นคงวา	ความคงเส้นคงวาของผลการคำนวณภายในเขตข้อมูลกับเขตข้อมูลที่ใกล้เคียงกัน	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบผลการคำนวณภายในเขตข้อมูล สัดส่วนของผลรวม และ ค่าเฉลี่ยของผลรวม เทียบกับค่าที่คำนวณได้ในอดีต	เนื้อหา/จำนวนคุณลักษณะ	การวัดภายในกระบวนการ
37	ความคงเส้นคงวา	ความคงเส้นคงวาของจำนวนแถวข้อมูลตามช่วงเวลาที่กำหนด	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบจำนวนแถวข้อมูลตามช่วงเวลาที่กำหนด เช่น รายเดือน รายตีมาส รายปี เทียบกับจำนวนในอดีต	เนื้อหา/การรวมหน่วยเวลา	การวัดอย่างต่อเนื่อง
38	ความคงเส้นคงวา	ความคงเส้นคงวาของจำนวนเขตข้อมูลตามช่วงเวลาที่กำหนด	ตรวจสอบความสมเหตุสมผล, เปรียบเทียบจำนวนเขตข้อมูลตามช่วงเวลาที่กำหนด เช่น รายเดือน รายตีมาส รายปี เทียบกับจำนวนในอดีต	เนื้อหา/การรวมหน่วยเวลา	การวัดอย่างต่อเนื่อง

หมายเลข	มิติคุณภาพ	ประเภทการวัด	คำอธิบายการวัด	วัตถุประสงค์ของการวัด	ประเภทของการประเมิน
39	ความถูกต้อง/ความสมบูรณ์	ความสมบูรณ์ในการอ้างอิงความสัมพันธ์	ยืนยันความสมบูรณ์ระหว่างตารางข้อมูลอ้างอิงกับตารางข้อมูลที่เรียกใช้ เพื่อระบุแถวข้อมูลที่หาแหล่งอ้างอิงไม่ได้	เนื้อหาระหว่างตารางข้อมูล	การวัดอย่างต่อเนื่อง
40	ความถูกต้อง/ความสมบูรณ์	ความสมบูรณ์ในการอ้างอิงความสัมพันธ์	ยืนยันความสมบูรณ์ระหว่างตารางข้อมูลที่เรียกใช้ กับตารางข้อมูลอ้างอิง เพื่อระบุรายการอ้างอิงที่ไม่พบในตารางข้อมูลที่เรียกใช้	เนื้อหาระหว่างตารางข้อมูล	การวัดอย่างต่อเนื่อง
41	ความถูกต้อง/ความถูกต้อง	ตรวจสอบความถูกต้อง, ระหว่างตารางข้อมูล, รายละเอียดของผลลัพธ์	เปรียบเทียบค่าที่จับคู่กันได้ หรือความสัมพันธ์เชิงธุรกิจระหว่างตารางข้อมูล เพื่อให้แน่ใจว่าข้อมูลได้รับการดูแลอย่างเหมาะสม	เนื้อหาระหว่างตารางข้อมูล	การวัดอย่างต่อเนื่อง
42	ความถูกต้อง/ความคงเส้นคงวา	ความคงเส้นคงวาของโปรไฟลด์แบบหลายเขตข้อมูลระหว่างตารางข้อมูล	ตรวจสอบความสมเหตุสมผลระหว่างตารางข้อมูล, เปรียบเทียบค่าการกระจายของแถวข้อมูล หรือ ตารางที่สัมพันธ์กัน กับสัดส่วนในอดีต เพื่อทดสอบ	เนื้อหาระหว่างตารางข้อมูล	การวัดอย่างต่อเนื่อง

หมายเลข	มิติคุณภาพ	ประเภทการวัด	คำอธิบายการวัด	วัตถุประสงค์ของการวัด	ประเภทของการประเมิน
			ตามกฎเกณฑ์เชิงธุรกิจ		
43	ความถูกต้อง/ความคงเส้นคงวา	ความคงเส้นคงวาของลำดับเหตุการณ์ กับกฎเกณฑ์เชิงธุรกิจ	ตรวจสอบความสมเหตุสมผลระหว่างตารางข้อมูล, เปรียบเทียบค่าของข้อมูลตามช่วงเวลาที่กำหนดในกฎเกณฑ์เชิงธุรกิจ	เนื้อหา/ลำดับเหตุการณ์/ระหว่างตารางข้อมูล	การวัดอย่างต่อเนื่อง
44	ความถูกต้อง/ความคงเส้นคงวา	ความคงเส้นคงวาระของผลการคำนวนภายในเขตข้อมูลระหว่างตาราง	ตรวจสอบความสมเหตุสมผลระหว่างตารางข้อมูล, เปรียบเทียบผลการคำนวนภายในเขตข้อมูลกับเขตข้อมูลในตารางอื่นที่มีความสัมพันธ์กัน	เนื้อหาระหว่างตารางข้อมูล/จำนวนคุณลักษณะ	การวัดอย่างต่อเนื่อง
45	ความถูกต้อง/ความคงเส้นคงวา	ความคงเส้นคงวาระของจำนวนเขตข้อมูลตามช่วงเวลาที่กำหนด	ตรวจสอบความสมเหตุสมผลระหว่างตารางข้อมูล, เปรียบเทียบจำนวนเขตข้อมูล สัดส่วนรวมทั้งหมดตามช่วงเวลาที่กำหนด เช่น รายเดือน รายไตรมาส รายปี กับตารางข้อมูลที่เกี่ยวข้อง	เนื้อหา/ระหว่างตารางข้อมูล/การรวมของหน่วยเวลา	การวัดอย่างต่อเนื่อง
46	ความคงเส้นคง	เปรียบเทียบ	เปรียบเทียบผลลัพธ์	ภาพรวม	การวัด

หมายเลข	มิติคุณภาพ	ประเภทการวัด	คำอธิบายการวัด	วัตถุประสงค์ของการวัด	ประเภทของการประเมิน
	วา	ความคงเส้นคงวากับตัวเทียบภายนอก	การวัดคุณภาพข้อมูลกับตัวเทียบภายนอก เช่น อุตสาหกรรมภายนอก หรือการวัดระดับชาติที่ข้อมูลใกล้เคียงกัน	เนื้อหาของฐานข้อมูล	อย่างต่อเนื่อง
47	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-เพียงพอสำหรับวัตถุประสงค์ที่กำหนดไว้ในภาพรวม	เปรียบเทียบเนื้อหาของฐานข้อมูลในภาพรวม กับความต้องการของผู้ใช้งานข้อมูล	ภาพรวมเนื้อหาของฐานข้อมูล	การวัดอย่างต่อเนื่อง
48	ความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-เพียงพอสำหรับการวัดและควบคุมในภาพรวม	ประเมินประสิทธิภาพในการวัดและควบคุม	ภาพรวมเนื้อหาของฐานข้อมูล	การประเมินอย่างต่อเนื่อง

Earley และคณะ (2017) ศึกษารวบรวมมิติคุณภาพข้อมูล โดยได้กำหนดมิติคุณภาพเป็น 2 ส่วน ได้แก่ มิติหลัก และมิติอื่น ๆ โดยมีรายละเอียด ดังนี้

1) มิติหลัก

1.1) ความสมบูรณ์ สัดส่วนของข้อมูลที่เก็บเทียบกับส่วนที่ควรจะเป็น

ตัวชี้วัด : ความสมบูรณ์ของข้อมูล = $\frac{\text{จำนวนข้อมูลที่ส่งมอบ}}{\text{จำนวนข้อมูลที่คาดหวัง}}$

1.2) ความเป็นเอกลักษณ์ ไม่มีข้อมูลที่ถูกบันทึกเหมือนกันซ้ำซ้อน

ตัวชี้วัด : จำนวนข้อมูลซ้ำซ้อน

1.3) ทันเวลา ระดับการแสดงข้อมูลที่แสดงความเป็นจริงในช่วงเวลาที่ต้องกา

ตัวชี้วัด : ความทันเวลา = $\frac{\text{จำนวนข้อมูลที่มาทันเวลา}}{\text{จำนวนข้อมูลทั้งหมด}}$

1.4) ความถูกต้อง ข้อมูลถูกต้องตามที่ถูกนิยามไว้ (รูปแบบ โครงสร้าง ขอบเขต)

ตัวชี้วัด : ความถูกต้อง = $\frac{\text{จำนวนข้อมูลที่ถูกต้องตามเกณฑ์ที่กำหนด}}{\text{จำนวนข้อมูลทั้งหมด}}$

1.5) ความแม่นยำ ระดับความถูกต้องของข้อมูลในโลกความเป็นจริง หรือกิจกรรมที่ถูกอธิบาย

1.6) ความแน่นอน การไม่ปรากฏถึงความแตกต่าง เมื่อเปรียบเทียบกับสิ่งที่อธิบายสิ่งเดียวกัน

2) มิติอื่น ๆ

2.1) ความสามารถในการใช้งาน ข้อมูลสามารถเข้าใจได้ ง่าย ตรงไปตรงมา เข้าถึงได้ บำรุงรักษาได้ มีระดับความเที่ยงตรงที่เหมาะสม

2.2) ปัญหาเรื่องเวลา

2.3) ความหยืดหยุ่น สามารถแยกแยะ หรือจัดกลุ่มได้ ง่ายต่อการจัดการ

2.4) ความลับ ข้อมูลได้รับการควบคุม ป้องกัน มีกระบวนการด้านความปลอดภัย สามารถยืนยันความถูกต้องได้

2.5) คุณค่า ข้อมูลมีประโยชน์

จากทฤษฎีเบื้องต้นผู้วิจัยได้สรุปแนวคิดดังกล่าว โดยสามารถคัดเลือกตัวชี้วัดที่สามารถคำนวณได้จากตัวชุดข้อมูลเอง โดยไม่รวมตัวชี้วัดที่คำนวณได้จากการตอบแบบสอบถาม ได้ 4 มิติ รวม 7 ตัวชี้วัด

2.2 มิติด้านความสมบูรณ์

Wang และ Strong (Wang & Strong, 1996) ได้ให้ตัวอย่างลักษณะข้อมูลในมิติด้านความสมบูรณ์ เช่น ระบบบันทึกข้อมูลต้องมีการบันทึกข้อมูลได้อย่างครบถ้วน ในกรณีที่ระบบบันทึกข้อมูลมีปัญหา ต้องสามารถฟื้นฟูระบบและส่งข้อมูลที่ขาดหายไปได้ หรืออีกกรณีหนึ่ง คือ ข้อมูลครบตามการออกแบบ มีการเก็บข้อมูลที่ครบถ้วนตามวัตถุประสงค์ เช่น ข้อมูลผู้ป่วย ต้องมีการเก็บ ชื่อ อายุ ที่อยู่ของผู้ป่วย ถ้าไม่ได้ออกแบบให้เก็บข้อมูลเหล่านี้ ถือว่าข้อมูลไม่สมบูรณ์จากออกแบบ ในขณะที่งานวิจัยอื่นๆ ก็ให้คำอธิบายใกล้เคียงกันคือข้อมูลต้องไม่เป็นค่าว่าง (Batini, Cappiello, Francalanci, & Maurino, 2009) หรือ ไม่มีข้อมูลที่ยอมให้เป็นค่าว่างได้ (Sebastian-Coleman, 2013) โดยจากงานวิจัยดังกล่าวสามารถกำหนดออกมาเป็นตัวชี้วัดและวิธีการคำนวณได้ดังนี้

1) ตัวชี้วัด – สัดส่วนความครบถ้วนของข้อมูลทั้งหมด

สัดส่วนความครบถ้วนของข้อมูลทั้งหมด $= 1 - \left(\frac{\text{จำนวนข้อมูลสูญหายทั้งหมด}}{\text{จำนวนข้อมูลทั้งหมด}}\right)$

Transaction unique id	Price	Date_of_Transfer	Property Type	Old/New	Duration	Town/City
{178CAFC0-F870-4777-8503-F304B47FD80D}	35000	1995-06-22 00:00		N		NORTHAMPTON
{66965E71-414E-4DA5-98A9-F304B75E12D0}		1995-03-30 00:00				
{EDC2913A-433E-406E-A30D-F304D0DD9C3F}		1995-02-28 00:00				
{681FDDBD-5816-47BE-BA8F-F304E02B8F26}		1995-08-16 00:00				BLACKPOOL
{ABDA7B7B-0820-41B6-AF7F-F694AF3BCD5C}		1995-01-11 00:00		N		LYME REGIS
{A30F3313-D08F-4F00-BCCB-F694DA91D30A}	30000	1995-11-30 00:00		N		LEIGH
{21A78385-2D67-4BD3-8A30-F694E520F433}	248000			N		LONDON

ภาพที่ 2.1 ลักษณะข้อมูลสูญหาย โดยไม่พิจารณาสดมภ์ข้อมูลที่สูญหายทั้งสดมภ์

2) ตัวชี้วัด – สัดส่วนสดมภ์ที่มีข้อมูลครบถ้วน

สัดส่วนสดมภ์ที่มีข้อมูลครบถ้วน $= \left(\frac{\text{จำนวนสดมภ์ที่มีข้อมูลครบถ้วน}}{\text{จำนวนสดมภ์ทั้งหมด}}\right)$

Transaction_unique_id	Price	Date_of_Transfer	Property Type	Old/New	Duration	Town/City
{178CAFC0-F870-4777-8503-F304B47FD80D}	35000	1995-06-22 00:00	T	N	F	NORTHAMPTON
{66965E71-414E-4DA5-98A9-F304B75E12D0}		1995-03-30 00:00	S	N	F	
{EDC2913A-433E-406E-A30D-F304D0DD9C3F}	59000	1995-02-28 00:00		N	F	MANCHESTER
{681FDDBD-5816-47BE-BA8F-F304E02B8F26}	25000	1995-08-16 00:00	F	N	L	BLACKPOOL
{ABDA7B7B-0820-41B6-AF7F-F694AF3BCD5C}	63000	1995-01-11 00:00	S	O		
{A30F3313-D08F-4F00-BCCB-F694DA91D30A}	30000		S	N	L	LEIGH
{21A78385-2D67-4BD3-8A30-F694E520F433}	248000	1995-05-24 00:00	F	O	L	LONDON

ภาพที่ 2.2 ลักษณะข้อมูลสูญหาย

3) ตัวชี้วัด – สัดส่วนสดมภ์ที่มีข้อมูล

สัดส่วนสดมภ์ที่มีข้อมูล = 1 - $\left(\frac{\text{จำนวนสดมภ์ที่มีข้อมูลสูญหายทั้งหมด}}{\text{จำนวนสดมภ์ทั้งหมด}}\right)$

Transaction_unique_id	Price	Date_of_Transfer	Property Type	Old/New	Duration	Town/City
{178CAFC0-F870-4777-8503-F304B47FD80D}	35000	1995-06-22 00:00			F	
{66965E71-414E-4DA5-98A9-F304B75E12D0}	90000	1995-03-30 00:00			F	
{EDC2913A-433E-406E-A30D-F304D0DD9C3F}	59000	1995-02-28 00:00	S		F	
{681FDDBD-5816-47BE-BA8F-F304E02B8F26}	25000		F		L	
{ABDA7B7B-0820-41B6-AF7F-F694AF3BCD5C}	63000		S		F	
{A30F3313-D08F-4F00-BCCB-F694DA91D30A}	30000	1995-11-30 00:00	S		L	
{21A78385-2D67-4BD3-8A30-F694E520F433}	248000	1995-05-24 00:00	F		L	

ภาพที่ 2.3 ลักษณะสดมภ์ข้อมูลที่มีข้อมูลสูญหายทั้งสดมภ์

4) ตัวชี้วัด – สัดส่วนแถวข้อมูลที่มีข้อมูลครบถ้วนทุดสดมภ์

สัดส่วนแถวข้อมูลที่มีข้อมูลครบถ้วนทุดสดมภ์= ((จำนวนแถวข้อมูลที่มีข้อมูลครบถ้วน)/ จำนวนแถวข้อมูลทั้งหมด)

Transaction_unique_id	Price	Date_of_Transfer	Property Type	Old/New	Duration	Town/City
{178CAFC0-F870-4777-8503-F304B47FD80D}	35000	1995-06-22 00:00	T		F	NORTHAMPTON
{66965E71-414E-4DA5-98A9-F304B75E12D0}	90000	1995-03-30 00:00	S	N	F	LIVERPOOL
{EDC2913A-433E-406E-A30D-F304D0DD9C3F}		1995-02-28 00:00	S		F	
{681FDDBD-5816-47BE-BA8F-F304E02B8F26}	25000	1995-08-16 00:00	F	N	L	BLACKPOOL
{ABDA7B7B-0820-41B6-AF7F-F694AF3BCD5C}	63000	1995-01-11 00:00	S		F	LYME REGIS
{A30F3313-D08F-4F00-BCCB-F694DA91D30A}	30000	1995-11-30 00:00	S		L	LEIGH
{21A78385-2D67-4BD3-8A30-F694E520F433}	248000	1995-05-24 00:00	F	O	L	LONDON

ภาพที่ 2.4 ลักษณะแถวข้อมูลที่สมบูรณ์

2.3 มิติด้านความคงเส้นคงวา

Wang และ Strong (Wang & Strong, 1996) ได้ให้ตัวอย่างลักษณะข้อมูลในมิติด้านความคงเส้นคงวา เช่น มีการแสดงผลในทศนิยมสองตำแหน่ง ก็ต้องมีการแสดงในรูปแบบนี้ตลอด โดย Sebastian-Coleman (2013) ได้ให้ตัวอย่างลักษณะข้อมูลในมิตินี้ว่า จะต้องเป็นข้อมูลที่มีรูปแบบภายในสดมภ์ข้อมูลนั้นๆ เหมือนกันทั้งสดมภ์จึงจะเรียกได้ว่าสดมภ์นั้นๆ มีความคงเส้นคงวา โดยจากงานวิจัยดังกล่าวสามารถกำหนดออกมาเป็นตัวชี้วัดและวิธีการคำนวณได้ดังนี้

1) ตัวชี้วัด – สัดส่วนจำนวนสดมภ์ข้อมูลที่มีรูปแบบข้อมูลภายในเหมือนกัน

สัดส่วนจำนวนสดมภ์ข้อมูลที่มีรูปแบบข้อมูลภายในเหมือนกันทั้งหมด

= ((จำนวนสดมภ์ข้อมูลที่มีรูปแบบข้อมูลภายในเหมือนกันทั้งหมด)/จำนวนสดมภ์ข้อมูลทั้งหมด

Transaction unique id	Price	Date_of_Transfer	Property Type	Old/New	Duration	Town/City
{178CAFC0-F870-4777-8503-F304B47FD80D}	35000	1995-06-22 00:00	T	N	F	NORTHAMPTON
{66965E71-414E-4DA5-98A9-F304B75E12D0}	90000	1995-03-30 00:00	S	N	F	LIVERPOOL
{EDC2913A-433E-406E-A30D-F304D0DD9C3F}	59000	1995-02-28 00:00	S	N	F	MANCHESTER
{681FDDBD-5816-47BE-BA8F-F304E02B8F26}	25000	1995-08-16 00:00	F	N	L	BLACKPOOL
{ABDA7B7B-0820-41B6-AF7F-F694AF3BCD5C}	63000	1995-01-11 00:00	S	N	F	LYME REGIS
{A30F3313-D08F-4F00-BCCB-F694DA91D30A}	30000	1995-11-30 00:00	S	N	L	LEIGH
Data Pattern	XXXXX	XXXX-XX-XX XX:XX	X	X	X	XXXXX

ภาพที่ 2.5 ลักษณะรูปแบบข้อมูลในแต่ละสดมภ์

2.4 มิติด้านความแม่นยำ

Wang และ Strong (Wang & Strong, 1996) ได้ให้ตัวอย่างลักษณะข้อมูลในมิติด้านความแม่นยำไว้ว่าในกรณีที่ชุดข้อมูลเดียวกัน มีหลายแหล่งข้อมูล ทุกแหล่งข้อมูลต้องแสดงข้อมูลที่ตรงกัน ถ้าแสดงข้อมูลไม่ตรงกันแสดงว่าข้อมูลไม่มีความแม่นยำ นอกจากนี้ Batini et al. (2009) ยังอธิบายไว้ว่ายังหมายถึงจำนวนข้อมูลที่ถูกต้องที่ถูกส่งมอบ จึงตีความได้ว่าการส่งข้อมูลที่มีจำนวนถูกต้องนั้นมีได้สองลักษณะคือการส่งข้อมูลไม่ครบจำนวน และการส่งข้อมูลที่เกินจำนวน โดยที่นี้จะหมายถึงการส่งข้อมูลที่ซ้ำซ้อนกัน โดยจากงานวิจัยดังกล่าวสามารถกำหนดออกมาเป็นตัวชี้วัดและวิธีการคำนวณได้ดังนี้

1) ตัวชี้วัด – สัดส่วนจำนวนแถวข้อมูลที่ไม่มีข้อมูลซ้ำซ้อน

สัดส่วนจำนวนแถวข้อมูลที่ไม่มีข้อมูลซ้ำซ้อน

= 1 - (จำนวนแถวข้อมูลที่มีค่าเหมือนกันทุกประการ/จำนวนแถวข้อมูลทั้งหมด)

Transaction_unique_id	Price	Date_of_Transfer	Property Type	Old/New	Duration	Town/City
{178CAFC0-F870-4777-8503-F304B47FD80D}	35000	1995-06-22 00:00	T		F	NORTHAMPTON
{178CAFC0-F870-4777-8503-F304B47FD80D}	35000	1995-06-22 00:00	T		F	NORTHAMPTON
{EDC2913A-433E-406E-A30D-F304D0DD9C3F}	59000	1995-02-28 00:00	S		F	MANCHESTER
{681FDDBD-5816-47BE-BA8F-F304E02B8F26}	25000	1995-08-16 00:00	F		L	BLACKPOOL
{ABDA7B7B-0820-41B6-AF7F-F694AF3BCD5C}	63000	1995-01-11 00:00	S		F	LYME REGIS
{A30F3313-D08F-4F00-BCCB-F694DA91D30A}	30000	1995-11-30 00:00	S		L	LEIGH
{21A78385-2D67-4BD3-8A30-F694E520F433}	248000	1995-05-24 00:00	F		L	LONDON

ภาพที่ 2.6 ลักษณะแถวข้อมูลที่ซ้ำกัน

2.5 มิติด้านความถูกต้อง

Earley (2017) ได้ให้คำอธิบายข้อมูลในมิตินี้ว่าข้อมูลต้องมีความถูกต้องตามที่นิยามไว้ โดยต้องมีการกำหนดกฎเกณฑ์ที่แน่นอน ซึ่งสอดคล้องกันกับงานวิจัยของ Batini et al. (2009) โดยจากงานวิจัยดังกล่าวสามารถกำหนดออกมาเป็นตัวชี้วัดและวิธีการคำนวณได้ดังนี้

1) ตัวชี้วัด – สัดส่วนจำนวนข้อมูลที่มีรูปแบบข้อมูลถูกต้องตามที่กำหนด

สัดส่วนจำนวนข้อมูลที่มีรูปแบบข้อมูลถูกต้องตามที่กำหนด

$$= \left(\frac{\text{จำนวนข้อมูลที่มีรูปแบบข้อมูลถูกต้องตามที่กำหนด}}{\text{จำนวนแถวข้อมูลทั้งหมด}} \right)$$

Transaction unique id	Price	Date_of_Transfer	Property Type	Old/New	Duration	Town/City
{178CAFC0-F870-4777-8503-F304B47FD80D}	35000	1995-06-22 00:00	T	N	F	NORTHAMPTON
{66965E71-414E-4DA5-98A9-F304B75E12D0}	90000	1995-03-30 00:00	S	N	F	LIVERPOOL
{EDC2913A-433E-406E-A30D-F304D0DD9C3F}	59000	1995-02-28 00:00	S	N	F	MANCHESTER
{681FDDBD-5816-47BE-BA8F-F304E02B8F26}	25000	1995-08-16 00:00	F	N	L	BLACKPOOL
{ABDA7B7B-0820-41B6-AF7F-F694AF3BCD5C}	63000	1995-01-11 00:00	S	N	F	LYME REGIS
{A30F3313-D08F-4F00-BCCB-F694DA91D30A}	30000	1995-11-30 00:00	S	N	L	LEIGH
Data Consumer Defined Rules	0 - 100000	^([0-9])*	(T|S|F)	(O|N)	(F|L)	^([A-Z])*

ภาพที่ 2.7 ลักษณะรูปแบบข้อมูลที่กำหนด

ตารางที่ 2.2 ตารางสรุปมิติ ตัวชี้วัด และวิธีการวัด B (Batini), S (Sebastian) , E (Earley)

มิติ	ตัวชี้วัด	วิธีการวัด	B	S	E	ตัวชี้วัดที่ใช้ในการวิจัยนี้
มิติด้านความแม่นยำ	จำนวนข้อมูลที่ถูกต้องที่ถูกส่งมอบ	คำนวณโดยตรงจากชุดข้อมูล 1 ตาราง	X			X
มิติด้านความสมบูรณ์	ความสมบูรณ์ของข้อมูล (ข้อมูลที่ไม่เป็นค่าว่าง)	คำนวณโดยตรงจากชุดข้อมูล 1 ตาราง	X	X	X	X
มิติด้านความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-มีคำอธิบายและข้อมูลอ้างอิงเพียงพอ	แบบสำรวจ		X		
มิติด้านความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-ความพร้อมสำหรับประมวลผล	แบบสำรวจ		X		
มิติด้านความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-สรุปจำนวนเขตข้อมูลทั้งหมด (สัดส่วนสดมภ์ที่มีข้อมูล)	คำนวณโดยตรงจากชุดข้อมูล 1 ตาราง		X		X
มิติด้านความสมบูรณ์	สัดส่วนแถวข้อมูลที่มีข้อมูลครบถ้วนทุดสดมภ์	คำนวณโดยตรงจากชุดข้อมูล 1 ตาราง				X
มิติด้านความสมบูรณ์	สัดส่วนสดมภ์ที่มีข้อมูลครบถ้วน	คำนวณโดยตรงจากชุดข้อมูล 1 ตาราง				X
มิติด้านความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-เปรียบเทียบขนาดข้อมูลในอดีต	คำนวณจากชุดข้อมูลหลายตารางที่สัมพันธ์กัน		X		
มิติด้านความสมบูรณ์	ความสมบูรณ์ของเนื้อหาเขตข้อมูล-ค่าตั้งต้นจากแหล่งข้อมูล	คำนวณจากชุดข้อมูลหลายตารางที่สัมพันธ์กัน		X		

มิติ	ตัวชี้วัด	วิธีการวัด	B	S	E	ตัวชี้วัดที่ใช้ในการวิจัยนี้
มิติด้านความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูลบนพื้นฐานของเงื่อนไขที่กำหนด	แบบสำรวจ		X		
มิติด้านความสมบูรณ์	ความสมบูรณ์ของฟิลข้อมูล-การได้รับข้อมูลที่เขตข้อมูลที่สำคัญสูญหาย	แบบสำรวจ		X		
มิติด้านความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-แถวข้อมูลคงเหลือหลังผ่านการประมวลผล	คำนวณจากชุดข้อมูลหลายตารางที่สัมพันธ์กัน		X		
มิติด้านความสมบูรณ์	ความสมบูรณ์ของเนื้อหาเขตข้อมูล-สัดส่วนผลรวมของจำนวนเขตข้อมูล	แบบสำรวจ		X		
มิติด้านความสมบูรณ์	ความสมบูรณ์ในการอ้างอิงความสัมพันธ์	คำนวณจากชุดข้อมูลหลายตารางที่สัมพันธ์กัน		X		
มิติด้านความสมบูรณ์	ความสมบูรณ์ของชุดข้อมูล-เพียงพอสำหรับวัตถุประสงค์ที่กำหนดไว้	แบบสำรวจ		X		
มิติด้านความคงเส้นคงวา	ความคงเส้นคงวา (จำนวนข้อมูลที่มีรูปแบบเหมือนกัน)	คำนวณโดยตรงจากชุดข้อมูล 1 ตาราง	X	X	X	X
มิติด้านความคงเส้นคงวา	ความคงเส้นคงวา (จำนวนข้อมูลที่มีรูปแบบต่างกัน)	คำนวณโดยตรงจากชุดข้อมูล 1 ตาราง	X			
มิติด้านความคงเส้นคงวา	ความคงเส้นคงวาของการจัดรูปแบบระหว่าง	คำนวณจากชุดข้อมูลหลายตาราง		X		

มิติ	ตัวชี้วัด	วิธีการวัด	B	S	E	ตัวชี้วัดที่ใช้ในการวิจัยนี้
	ตารางข้อมูล	ที่สัมพันธ์กัน				
มิติด้านความคงเส้นคงวา	ความคงเส้นคงวาในการใช้ค่าตั้งต้นสำหรับเขตข้อมูล	คำนวณจากชุดข้อมูลหลายตารางที่สัมพันธ์กัน		X		
มิติด้านความคงเส้นคงวา	ความคงเส้นคงวาของเวลาที่ใช้	แบบสำรวจ				
มิติด้านความคงเส้นคงวา	ความคงเส้นคงวาของผลการคำนวณภายในเขตข้อมูลกับเขตข้อมูลที่ใกล้เคียงกัน	คำนวณจากชุดข้อมูลหลายตารางที่สัมพันธ์กัน		X		
มิติด้านความคงเส้นคงวา	ความคงเส้นคงวาของจำนวนแถวข้อมูลตามช่วงเวลาที่กำหนด	คำนวณจากชุดข้อมูลหลายตารางที่สัมพันธ์กัน		X		
มิติด้านความทันเวลา	สัดส่วนการประมวลผลข้อมูลได้ภายในเวลาที่กำหนด	แบบสำรวจ	X	X		
มิติด้านความทันเวลา	ความพร้อมของข้อมูลเมื่อต้องการประมวลผล	แบบสำรวจ		X		
มิติด้านความเป็นปัจจุบัน	วันที่ปรับปรุงข้อมูลล่าสุด	คำนวณจากชุดข้อมูลหลายตารางที่สัมพันธ์กัน	X		X	
มิติด้านการเปลี่ยนแปลง	ความยาวของช่วงเวลาที่ข้อมูลไม่มีการเปลี่ยนแปลง	คำนวณจากชุดข้อมูลหลายตารางที่สัมพันธ์กัน	X			
มิติด้านความเหมาะสมของจำนวนข้อมูล	จำนวนที่เหมาะสมของข้อมูล	แบบสำรวจ	X			
มิติด้านการเข้าถึง	การเข้าถึงข้อมูล	แบบสำรวจ	X			

มิติ	ตัวชี้วัด	วิธีการวัด	B	S	E	ตัวชี้วัดที่ใช้ในการวิจัยนี้
มิติด้านความน่าเชื่อถือ	จำนวนของข้อมูลที่เป็นไปตามค่ามาตรฐาน	แบบสำรวจ	X			
มิติด้านการตีความ	จำนวนข้อมูลที่สามาถตีความได้ หรือ เอกสารอธิบาย	แบบสำรวจ	X			
มิติด้านการได้มาของข้อมูล	เปอร์เซ็นต์ของการคำนวณที่ถูกต้องของข้อมูลที่ได้รับตามสูตรที่ได้มาหรือคำจำกัดความการคำนวณ	แบบสำรวจ	X			
มิติด้านความเร็ว	ความเร็วในการตอบสมองของเครื่องแม่ข่ายและเครือข่าย	แบบสำรวจ	X			
มิติด้านความถูกต้อง	ความถูกต้อง ข้อมูลถูกต้องตามที่ถูกนิยามไว้	คำนวณโดยตรงจากชุดข้อมูล 1 ตาราง		X	X	X

บทที่ 3

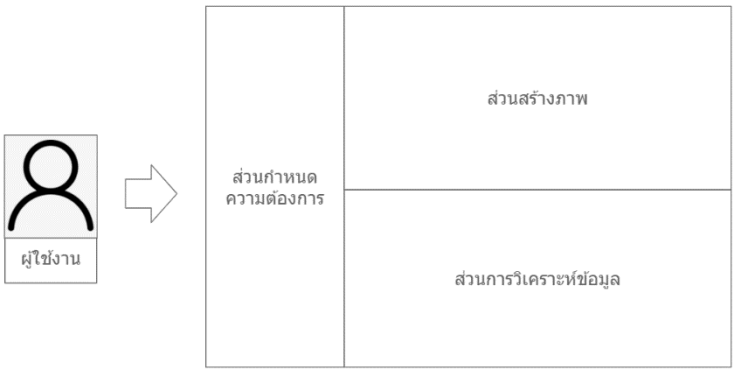
วิธีการดำเนินงาน

การพัฒนาระบบตรวจสอบคุณภาพข้อมูลนี้ ได้นำตัวชี้วัดคุณภาพข้อมูลที่ได้ศึกษาไว้แล้วในบทที่ 2 ผนวกกับตัวชี้วัดเพิ่มเติมที่คาดว่าจะมีประโยชน์ต่อผู้ใช้งาน โดยในบทนี้จะเสนอรายละเอียดดังกล่าว ดังนี้

3.1 การออกแบบและพัฒนาระบบ

3.1.1 โครงสร้างการทำงานของระบบ

ในส่วนนี้ผู้วิจัยจะกล่าวถึงรายละเอียดในการออกแบบระบบ ซึ่งสามารถแบ่งได้ ดังนี้



ภาพที่ 3.1 โครงสร้างการทำงานของระบบ

ในการออกแบบทางสถาปัตยกรรมของระบบ นั้น ได้มีการแบ่งส่วนการทำงานออกเป็น 3 ส่วน ได้แก่ ส่วนกำหนดความต้องการ ส่วนการวิเคราะห์ข้อมูล และส่วนสร้างภาพข้อมูล

1) ส่วนกำหนดความต้องการ

ในส่วนกำหนดความต้องการนี้ เป็นส่วนที่ผู้ใช้งานสามารถกำหนดค่าต่าง ๆ ได้ว่าต้องการทำงานในส่วนใด โดยสามารถแบ่งได้ เป็น 4 ส่วน ได้แก่ ส่วนตั้งค่าต่าง ๆ สำหรับการสร้างโปรไฟล์ข้อมูล อาทิเช่น ตำแหน่งที่วางไฟล์ข้อมูล ชื่อโปรไฟล์ เป็นต้น ส่วนต่อมาคือการตั้งค่าสำหรับตรวจสอบข้อมูล ในส่วนนี้ผู้ใช้สามารถกำหนดได้ว่าลักษณะข้อมูลที่ถูกต้องนั้นมีลักษณะอย่างไง อาทิเช่น ข้อมูลต้องอยู่ไหนพิสัยเท่าไหร่ รูปแบบแบบใด เป็นค่าสูญหายได้หรือไม่ เป็นต้น ส่วนสุดท้ายจะ

เป็นการตั้งค่าผลลัพธ์ที่ต้องการว่าต้องการในรูปแบบใด ระหว่างรายงานที่มีการสร้างภาพข้อมูลสำหรับ กรณีที่ผู้ใช้ต้องการนำไปทบทวนด้วยตัวเอง หรือผลลัพธ์ในรูปของข้อมูลโครงสร้างแบบ JSON ที่ใช้สำหรับนำไปใช้งานต่อในการพัฒนาระบบอัตโนมัติ

2) ส่วนการวิเคราะห์ข้อมูล

ในส่วนของการวิเคราะห์ข้อมูล สามารถแบ่งได้เป็น 2 ส่วน คือ การวิเคราะห์ข้อมูลเพื่อสร้างเป็นโปรไฟล์ และการวิเคราะห์ข้อมูลเพื่อตรวจสอบคุณภาพข้อมูลที่ได้มีการกำหนดไว้ ในส่วนของการวิเคราะห์ข้อมูลเพื่อสร้างเป็นโปรไฟล์นั้น จะประกอบไปด้วย การวิเคราะชุดข้อมูลในภาพรวม ซึ่งผลลัพธ์ที่ได้จะแสดงข้อมูล อาทิเช่น จำนวนแถวข้อมูล จำนวนสดมภ์ข้อมูล จำนวนข้อมูลซ้ำซ้อน และจำนวนข้อมูลสูญหาย เป็นต้น นอกจากนี้ยังมีการวิเคราะห์ลงไปในระดับรายสดมภ์ ซึ่งจะแสดงผลลัพธ์ อาทิเช่น ประเภทข้อมูล จำนวนข้อมูลจำเพาะ จำนวนข้อมูลสูญหาย จำนวนรูปแบบของข้อมูล จำนวนข้อมูลที่มีค่าเหมือนกัน เป็นต้น สำหรับการวิเคราะห์เพื่อตรวจสอบคุณภาพข้อมูลนั้น จะเป็นการวิเคราะห์ตามที่ผู้ใช้งานได้กำหนดไว้ในตอนต้น โดยผลลัพธ์ที่ได้ จะแสดงข้อมูล อาทิเช่น จำนวนรายการที่ผู้ใช้กำหนดไว้ จำนวนข้อมูลที่ผ่านเงื่อนไขที่กำหนด เป็นต้น

3) ส่วนการสร้างภาพข้อมูล

ในส่วนการสร้างภาพข้อมูลนี้ จะเป็นการนำข้อมูลจากส่วนการวิเคราะห์ข้อมูล มาสร้างภาพข้อมูล โดยมีการจัดรูปแบบแบบรายงาน โดยรายงานนี้ถูกสร้างในรูปแบบของเอกสาร HTML(Hypertext Markup Language๗ ผู้ใช้งานสามารถเปิดเอกสารนี้ได้โดยใช้อินเตอร์เน็ตบราวเซอร์ทั่วไปได้ รวมถึงสามารถเปลี่ยนรูปแบบรายงานเป็นเอกสาร PDF(Portable Document Format) ไฟล์ได้ตามต้องการ

3.1.2 การประมวลผลข้อมูล

ในส่วนของการประมวลผลข้อมูล ผู้วิจัยได้ใช้วิธีการประมวล 2 แบบ โดยแบ่งได้ ดังนี้

1) การประมวลผลเพื่อคำนวณคุณภาพข้อมูลในมิติต่าง ๆ

ในการประมวลผลในลักษณะนี้ ผู้วิจัยใช้วิธีการคำนวณโดยการกำหนดสมการลงบนรหัสโปรแกรมโดยตรง

2) การประมวลผลเพื่อคำนวณค่าทางสถิติ

ในการประมวลผลในลักษณะนี้ ผู้วิจัยใช้การคำนวณจากฟังชั่นสำเร็จรูปของไลบรารี่ที่มีชื่อว่า Pandas ซึ่งใช้งานกันทั่วไปในงานด้านวิเคราะห์ข้อมูล ค่าทางสถิติที่ใช้วิธีนี้ในการคำนวณ เช่น ค่าเฉลี่ย ต่าง ๆ ค่าเบี่ยงเบนมาตรฐาน และค่าความแปรปวนเป็นต้น

3.1.3 การแสดงผลตัวชี้วัดที่ได้จากการศึกษา

1) มิติด้านความสมบูรณ์

1.1) สัดส่วนความครบถ้วนของข้อมูลทั้งหมด

1.2) สัดส่วนสดมภ์ที่มีข้อมูลครบถ้วน / จำนวนสดมภ์ที่เป็นค่าว่างทั้งหมด

1.3) สัดส่วนสดมภ์ที่มีข้อมูล

1.4) สัดส่วนแถวข้อมูลที่มีข้อมูลครบถ้วนทุดสดมภ์

2) มิติด้านควาคงเส้นคงวา

2.1) สัดส่วนจำนวนสดมภ์ข้อมูลที่มีรูปแบบข้อมูลภายในเหมือนกันทั้งหมด

3) มิติด้านความแม่นยำ

3.1) สัดส่วนจำนวนแถวข้อมูลที่ไม่มีข้อมูลซ้ำซ้อน

4) มิติด้านความถูกต้อง

4.1) สัดส่วนจำนวนข้อมูลที่มีรูปแบบข้อมูลถูกต้องตามที่กำหนด

3.1.4 การแสดงผลค่าทางสถิติ

1) จำนวนแถวข้อมูลทั้งหมด

2) จำนวนแถวข้อมูลทั้งหมดหลังนำข้อมูลซ้ำซ้อนออก

3) จำนวนสดมภ์ทั้งหมด

4) สัดส่วนความเป็นเอกลักษณ์ของข้อมูลแต่ละสดมภ์

6) จำนวนความยาวต่ำสุด – สูงสุดของข้อมูลแต่ละสดมภ์

7) ข้อมูลที่มีความถี่มากที่สุดของแต่ละสดมภ์

8) รูปแบบข้อมูลที่มีความถี่มากที่สุดของแต่ละสดมภ์

11) จำนวนข้อมูลที่มีความแตกต่างกัน กรณีเป็นข้อมูลเชิงตัวเลข

12) ค่าต่ำสุด - สูงสุด

13) ค่ามัชฌิมเลขคณิต ค่ามัธยฐาน และค่าฐานนิยม

14) ค่าความแปรปวน

15) ค่าเบี่ยงเบนมาตรฐาน

16) ควอไทล์ที่ 1 อินเตอร์ควอไทล์ ควอไทล์ที่ 3

17) ค่าขั้นต่ำ - ค่าขั้นสูง

18) ค่านอกเกณฑ์ด้านลบ - ด้านบวก

19) จำนวนข้อมูลที่ได้ทำการตรวจสอบ

20) จำนวนที่ผ่านการตรวจสอบ

21) สัดส่วนขข้อมูลที่ผ่านการตรวจสอบ

3.1.5 การแสดงผลอื่น ๆ

1) วันที่วิเคราะห์ข้อมูล

2) รุ่นของระบบ

3) รหัสรายงาน

4) ชื่อไฟล์ข้อมูลที่ทำการวิเคราะห์

5) ชื่อสดมภ์ที่เป็นคีย์ต้น

9) ภาพแสดงฮิสโตแกรม ของแต่ละสดมภ์ กรณีเป็นข้อมูลเชิงตัวเลข

10) ภาพแสดงบ็อกพล็อต ของแต่ละสดมภ์ กรณีเป็นข้อมูลเชิงตัวเลข

3.1.6 การแสดงผลในรูปแบบรายงาน

ในส่วนนี้ผู้วิจัยได้ใช้ผลลัพธ์ที่เกิดขึ้นจากระบบมาแสดงผลในรูปแบบของรายงาน โดยประกอบ 2 ส่วน ได้แก่

1) การสร้างภาพข้อมูล

ในส่วนนี้ผู้วิจัยได้ออกแบบให้การสร้างภาพข้อมูล นำข้อมูลรูปแบบ JSON(JavaScript Object Notation) นำมาสร้างภาพผ่าน Python Package ที่ชื่อว่า Matplotlib โดยภาพข้อมูลที่สร้างด้วยวิธีการนี้ประกอบไปด้วย ภาพฮีสโตแกรม ภาพบ๊อกพล็อต ภาพกราฟแท่ง และภาพสัดส่วนวงกลม

2) การแสดงผลหน้ารายงาน

ในส่วนนี้ผู้วิจัยได้ออกแบบให้การสร้างหน้ารายงาน ด้วยการสร้างเอกสารHTML ผ่านภาษา Python เพื่อให้สามารถเปิดได้ผ่าน web browser และง่ายต่อการบันทึกเป็นเอกสาร PDF

3.2 การทดสอบการใช้งานระบบ

ในงานวิจัยนี้ได้นำระบบที่พัฒนาขึ้นมาประยุกต์ใช้กับการตรวจสอบคุณภาพข้อมูล 2 กรณี ได้แก่

1) การวิเคราะห์เพื่อสำรวจข้อมูลเบื้องต้น ในกรณีนี้จะเป็นการนำชุดข้อมูลมาวิเคราะห์ว่าคุณภาพข้อมูลเบื้องต้นเป็นอย่างไร ก่อนนำไปทำการวิเคราะห์อื่น ๆ ต่อไป

2) การวิเคราะห์เพื่อตรวจสอบคุณภาพข้อมูลแบบกำหนดเงื่อนไข โดยนำผลวิเคราะห์จากการสำรวจข้อมูลเบื้องต้นมาเป็นเกณฑ์กำหนดเงื่อนไขการตรวจสอบต่าง ๆ

โดยทั้ง 2 กรณีนี้จะใช้ชุดข้อมูลเดียวกันโดยเป็นข้อมูลซื้อขายของออนไลน์ในช่วงเวลา เดือนธันวาคม ปี 2010 ถึงเดือนธันวาคม ปี 2011 ของร้านค้าในสหราชอาณาจักร ที่เผยแพร่โดย

มหาวิทยาลัยลอนดอนเซ้าท์แบงค์ (London South Bank University) ในเว็บไซต์ www.kaggle.com โดยชุดข้อมูลมีรายละเอียดเบื้องต้นดังนี้

1) ข้อมูลประกอบด้วยสดมภ์ 8 สดมภ์ ได้แก่ InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID และ Country

2) ข้อมูลมีจำนวน 541,909 รายการ

การวิเคราะห์เพื่อสำรวจคุณภาพข้อมูลเบื้องต้น มีขั้นตอนดังนี้

1) ให้ระบบทำการตรวจสอบคุณภาพข้อมูลโดยยังไม่มีการกำหนดเงื่อนไข

2) นำผลลัพธ์ที่แสดงในรูปแบบรายงานมาวิเคราะห์คุณภาพข้อมูลจากโปรไฟล์ข้อมูลที่ได้

การวิเคราะห์เพื่อตรวจสอบคุณภาพข้อมูลแบบกำหนดเงื่อนไข มีขั้นตอนดังนี้

1) นำผลการวเคราะห์ในข้อ 3.2.1 มากำหนดเงื่อนไขการตรวจสอบคุณภาพข้อมูลในส่วนที่ 2

2) ให้ระบบตรวจสอบคุณภาพข้อมูลตามเงื่อนไข

3) วิเคราะห์ผลการตรวจสอบคุณภาพข้อมูล

บทที่ 4
รายงานการทดสอบ

4.1 รายงานผลการทดสอบระบบ

การวิเคราะห์เพื่อสำรวจข้อมูลเบื้องต้น ได้ผลการทดสอบดังนี้

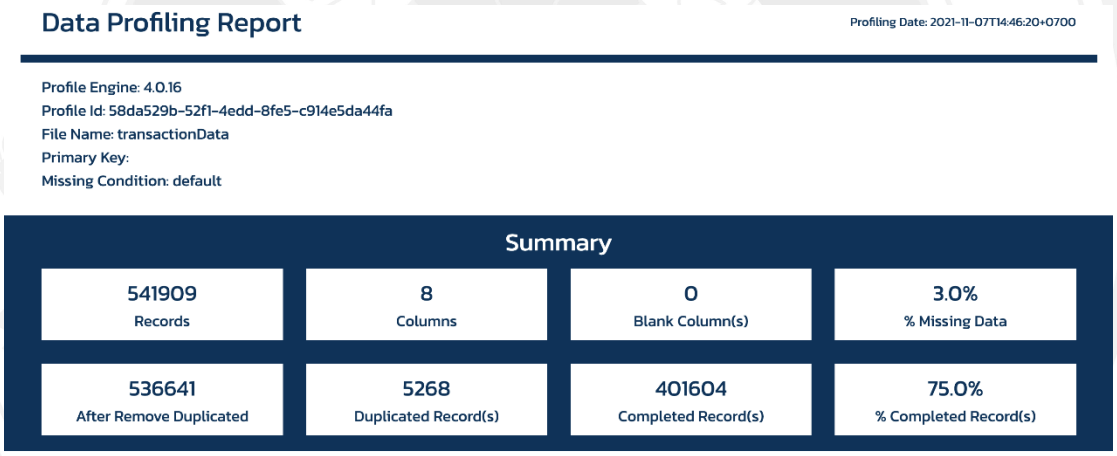
1) ขนาดข้อมูลประมาณ 45 MB

2) ข้อมูลมีจำนวน 8 สดมภ์ จำนวนรายการ 541,909 รายการ

3) ระบบเวลาที่ใช้ 10:50 นาที ในการสร้างโปรไฟล์ข้อมูล

โดยรายละเอียดโปรไฟล์ข้อมูลมีรายละเอียดดังนี้

1) รายละเอียดในภาพรวมของข้อมูล

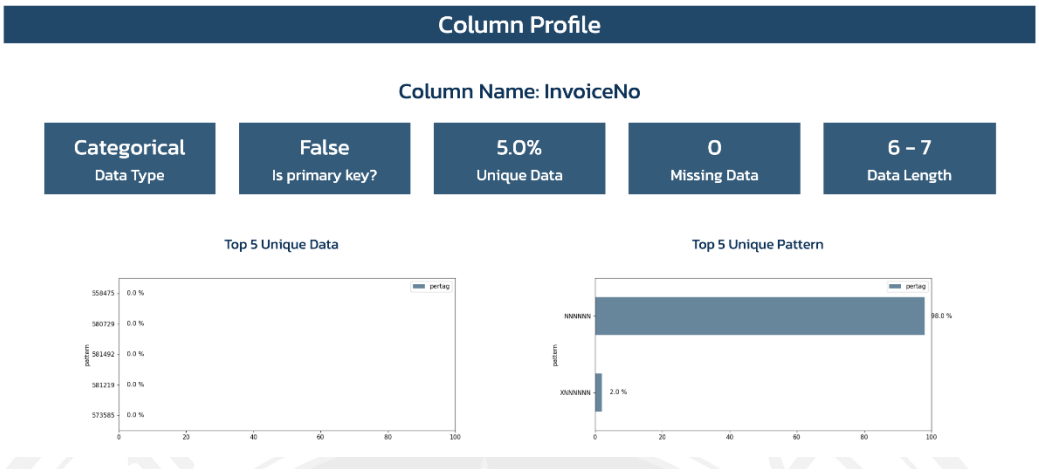


ภาพที่ 4.1 รายละเอียดภาพรวมของข้อมูล

จากการตรวจสอบภาพรวมของข้อมูลของระบบ พบว่าชุดข้อมูลมีข้อมูลสูญหายจำนวน 3% และมีข้อมูลซ้ำซ้อนจำนวน 5,268 รายการ และมีจำนวนรายการที่มีข้อมูลครบถ้วนทุกสดมภ์จำนวน 401,604 รายการ หรือคิดเป็น 75% ของจำนวนรายการทั้งหมด

2) รายละเอียดของแต่ละสดมภ์

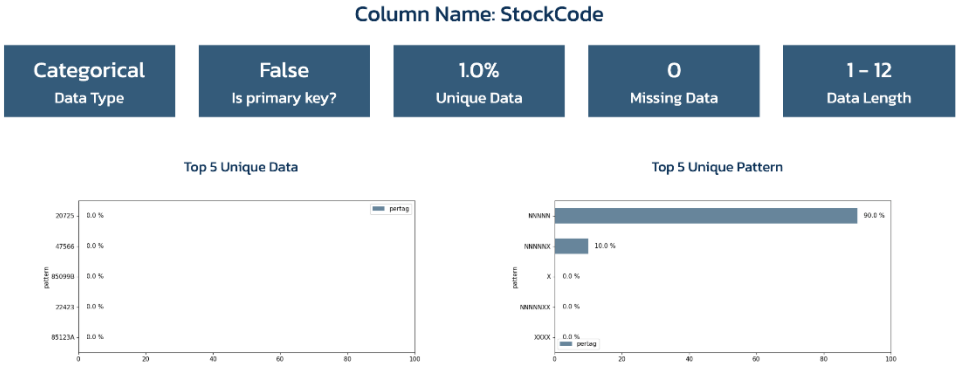
2.1) สดมภ์ – InvoiceNo



ภาพที่ 4.2 ผลลัพธ์การตรวจสอบของสดมภ์ InvoiceNo

จากการตรวจสอบของระบบพบว่า สดมภ์ InvoiceNo หรือ เลขที่ใบเสร็จ เป็นสดมภ์ประเภทหมวดหมู่ โดยมีค่าความเฉพาะของข้อมูลอยู่ที่ 5% หมายความว่าข้อมูลในสดมภ์นี้สามารถที่จะมีค่าซ้ำกันได้ กล่าวคือในแต่ละ InvoiceNo สามารถมีได้หลายรายการ นอกจากนี้ยังพบว่าไม่มีข้อมูลสูญหาย และความยาวของข้อมูลอยู่ที่ 6 ตัวอักษร จำนวน 98% และจำนวน 7 ตัวอักษร จำนวน 2% ซึ่งถ้าพบข้อมูลในลักษณะจำเป็นต้องมีการตรวจสอบรูปแบบของข้อมูลว่า 2% เป็นรูปแบบข้อมูลที่ถูกต้องหรือไม่

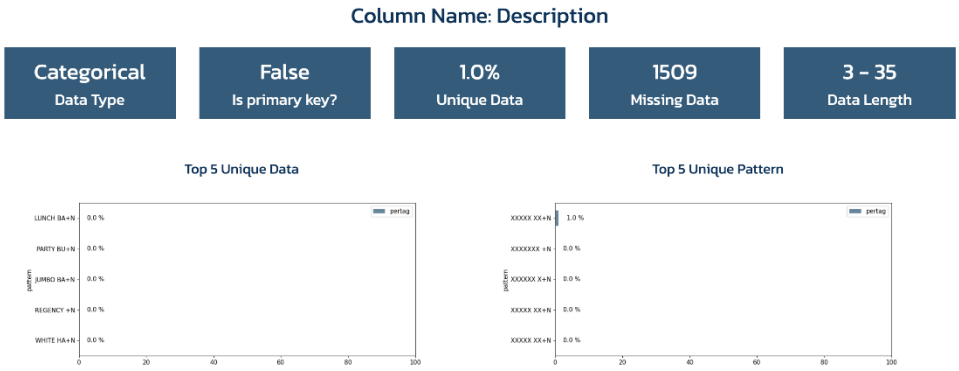
2.2) สดมภ์ – StockCode



ภาพที่ 4.3 ผลลัพธ์การตรวจสอบของสดมภ์ StockCode

จากการตรวจสอบของระบบพบว่า สดมภ์ StockCode หรือ รหัสสินค้า เป็นสดมภ์ประเภทหมวดหมู่ โดยมีค่าความเฉพาะของข้อมูลอยู่ที่ 1% หมายความว่าข้อมูลในสดมภ์นี้สามารถที่จะมีค่าซ้ำกันได้ กล่าวคือในแต่ละ Stockcode สามารถมีได้หลายรายการ นอกจากนี้ยังพบว่าไม่มีข้อมูลสูญหาย และความยาวของข้อมูลอยู่ที่ 5 ตัวอักษร จำนวนประมาณ 90% จำนวน 6 ตัวอักษร จำนวนประมาณ 10% และยังพบจำนวนอักษร 1, 7 และ 4 ตัวอักษร จำนวนเล็กน้อย ซึ่งถ้าพบข้อมูลในลักษณะจำเป็นต้องมีการตรวจสอบรูปแบบของข้อมูลว่าเป็นรูปแบบข้อมูลที่ถูกต้องหรือไม่

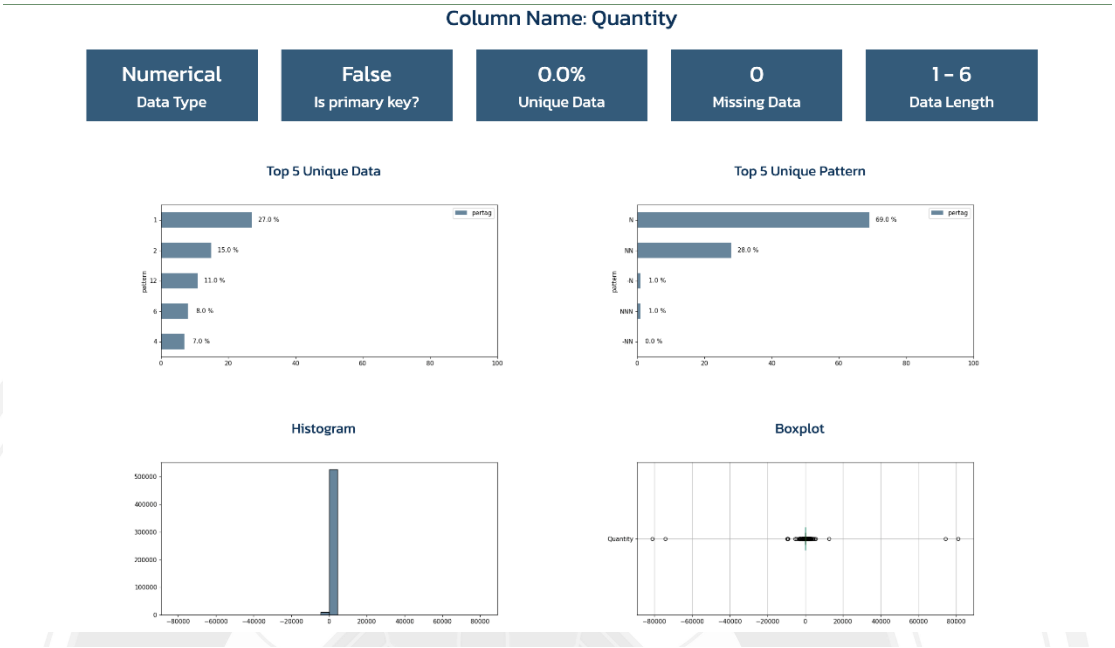
2.3) สดมภ์ – Description



ภาพที่ 4.4 ผลลัพธ์การตรวจสอบของสดมภ์ Description

จากการตรวจสอบของระบบพบว่า สดมภ์ Description หรือ คำอธิบาย เป็นสดมภ์ประเภทหมวดหมู่ โดยมีค่าความเฉพาะของข้อมูลอยู่ที่ 1% หมายความว่าข้อมูลในสดมภ์นี้สามารถที่จะมีค่าซ้ำกันได้ กล่าวคือในแต่ละ Description สามารถมีได้หลายรายการ นอกจากนี้ยังพบว่ามีข้อมูลสูญหายจำนวน 1,509 รายการ ซึ่งต้องมีการตรวจสอบข้อมูลต่อไปว่าเป็นเพราะเหตุใด ในขณะที่ความยาวหรือรูปแบบของข้อมูลในสดมภ์นี้ไม่ถูกนำมาพิจารณา เนื่องจากเป็นสดมภ์ที่ใช้ในการอธิบายสดมภ์ StockCode

2.4) สดมภ์ – Quantity



ภาพที่ 4.5 ผลลัพธ์การตรวจสอบของสดมภ์ Quantity

จากการตรวจสอบของระบบพบว่า สดมภ์ Quantity หรือ จำนวนสินค้าที่ซื้อ เป็นสดมภ์ประเภทตัวเลข โดยมีค่าความเฉพาะของข้อมูลอยู่ที่ 0% หมายความว่าข้อมูลในสดมภ์นี้สามารถที่จะมีค่าซ้ำกันได้ กล่าวคือในแต่ละรายการสามารถมี จำนวนสินค้าที่ซื้อ ซ้ำกันได้ นอกจากนี้ยังพบว่าไม่มีข้อมูลสูญหาย และความยาวของข้อมูลอยู่ที่ 1 ถึง 6 หลัก โดยมีจำนวน 1 หลัก 69% หมายความว่าลูกค้าจะซื้อสินที่จำนวน 1 ถึง 9 ชิ้น เป็นจำนวนถึง 69% ในขณะที่ซื้อสินค้าจำนวน 10 ถึง 99 ชิ้น จำนวน 28% นอกจากนี้ยังพบการซื้อสินค้ามากกว่า 100 ชิ้นจำนวนเล็กน้อย และพบจำนวนสินค้าที่ซื้อเป็นค่าติดลบ(-NN) ซึ่งต้องมีการตรวจสอบข้อมูลต่อไปว่าเป็นข้อมูลที่ถูกต้องหรือไม่

Metrics

Metrics	Value
distinct value	722
min	-80995.0
max	80995.0
mean	9.62
mode	1.0
median	3.0
Variance	48018.03
Standard deviation	219.13

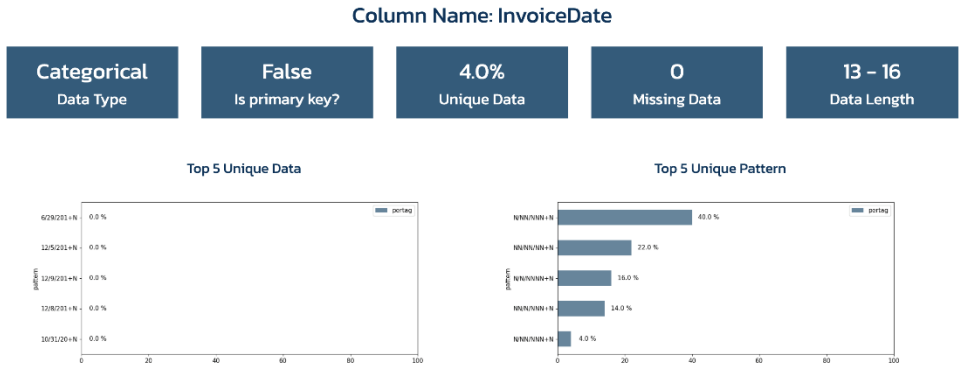
Metrics

Metrics	Value
median	3.0
first quartile	1.0
third_quartile	10.0
iqr	9.0
minimum	-3.5
maximum	14.5
upper outliner datapoint	69340
lower outliner datapoint	4413

ภาพที่ 4.6 ผลลัพธ์การตรวจสอบของสดมภ์ Quantity

นอกจากนี้เนื่องจากเป็นข้อมูลสดมภ์ประเภทตัวเลข ระบบจะทำการคำนวณค่าทางสถิติมาให้เพื่อพิจารณา โดยพบว่าค่าต่ำสุดอยู่ที่ -80,995 ค่าสูงสุดอยู่ที่ 80,995 ซึ่งเป็นค่าที่ไม่ปกติ โดยเมื่อพิจารณารวมกับค่าทางสถิติอื่นๆ เช่น ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน และค่าควอไทล์ที่ 3 จะเห็นชัดแจนว่าเป็นค่าที่ผิดปกติ ต้องมีการนำออกจากการวิเคราะห์ในลำดับต่อไป

2.5) สดมภ์ – InvoiceDate

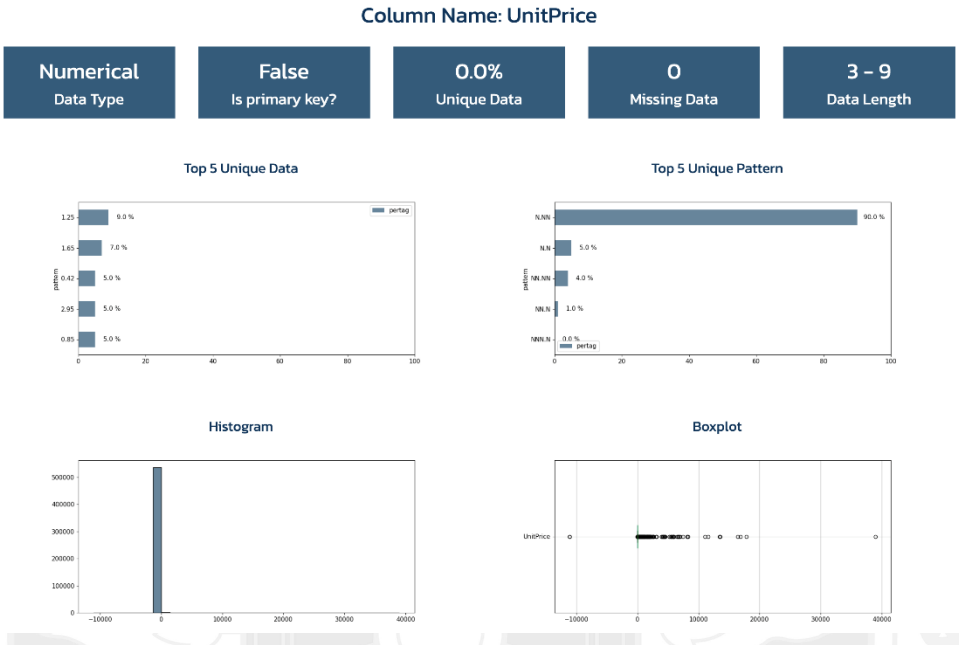


ภาพที่ 4.7 ผลลัพธ์การตรวจสอบของสดมภ์ InvoinceDate

จากการตรวจสอบของระบบพบว่า สดมภ์ InvoiceDate หรือ วันที่ใบเสร็จ เป็นสดมภ์ประเภทหมวดหมู่ โดยมีค่าความเฉพาะของข้อมูลอยู่ที่ 4% หมายความว่าข้อมูลในสดมภ์นี้สามารถที่จะมีค่าซ้ำกันได้ กล่าวคือในแต่ละวันสามารถมีวันที่ ๆ ของใบเสร็จเดียวกันได้หลายรายการ นอกจากนี้ยังพบว่าไม่มีข้อมูลสูญหายจำนวน 1,509 รายการ ในขณะที่ความยาวหรือรูปแบบของข้อมูลในสดมภ์นี้อยู่ที่ 13 ถึง 16 ตัวอักษร โดยข้อมูลในสดมภ์นี้เป็นรูปแบบข้อมูลวันและเวลา ซึ่งควรจะมีข้อมูลรูปแบบเดียวกันทั้งหมด โดยเบื้องต้นพบว่าสาเหตุที่รูปแบบแตกต่างกันเป็นเพราะข้อมูลวันและเวลาในชุดข้อมูลนี้มีการตัด 0 ออกจากข้อมูลหลักหน่วย เช่น วันที่ 8 เดือน 8 ปี 2011 ในข้อมูล

ชุดนี้จะเก็บข้อมูลเป็น 8/8/2011 แทนที่จะเป็น 08/08/2011 ตามรูปแบบมาตรฐาน ดังนั้นจึงแนะนำให้แปลงเป็นรูปแบบมาตรฐานก่อนการตรวจสอบข้อมูล เพื่อแยกข้อมูลที่ผิดรูปแบบได้ชัดเจนมากขึ้น

2.6) สดมภ์ – UnitPrice



ภาพที่ 4.8 ผลลัพธ์การตรวจสอบของสดมภ์ UnitPrice

จากการตรวจสอบของระบบพบว่า สดมภ์ UnitPrice หรือ ราคาต่อหน่วย เป็นสดมภ์ประเภทตัวเลข โดยมีค่าความเฉพาะของข้อมูลอยู่ที่ 0% หมายความว่าข้อมูลในสดมภ์นี้สามารถที่จะมีค่าซ้ำกันได้ กล่าวคือในแต่ละรายการสามารถมี ราคาต่อหน่วย ซ้ำกันได้ นอกจากนี้ยังพบว่าไม่มีข้อมูลสูญหาย และความยาวของข้อมูลอยู่ที่ 3 ถึง 9 หลัก โดยมีจำนวน 1 หลัก 95% หมายความว่าราคาสินค้าที่ถูกซื้อจะอยู่ที่ราคา 1 ถึง 9 เป็นจำนวนถึง 95% ในขณะที่ราคาสินค้า 10 ถึง 99 จำนวน 5% และพบการซื้อสินค้าที่มีราคามากกว่า 100 จำนวนเล็กน้อย

Metrics

Metrics	Value
distinct value	1630
min	-11062.06
max	38970.0
mean	4.63
mode	1.25
median	2.08
Variance	9454.28
Standard deviation	97.23

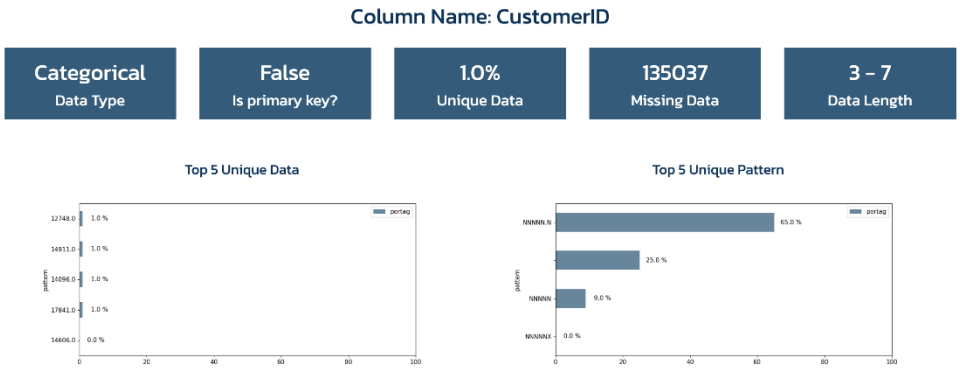
Metrics

Metrics	Value
median	2.08
first quartile	1.25
third_quartile	4.13
iqr	2.88
minimum	-0.19
maximum	5.57
upper outliner datapoint	85419
lower outliner datapoint	2

ภาพที่ 4.9 ผลลัพธ์การตรวจสอบของสดมภ์ UnitPrice

นอกจากนี้เนื่องจากเป็นข้อมูลสดมภ์ประเภทตัวเลข ระบบจะทำการคำนวณค่าทางสถิติมาให้เพื่อพิจารณา โดยพบว่าค่าต่ำสุดอยู่ที่ -11,062.06 ซึ่งเป็นค่าที่ไม่ปกติ โดยเมื่อพิจารณารวมกับค่าทางสถิติอื่นๆ เช่น ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน และค่าควอไทล์ที่ 3 จะเห็นชัดแจนว่าเป็นค่าที่ผิดปกติ ต้องมีการนำออกจากการวิเคราะห์ในลำดับต่อไป

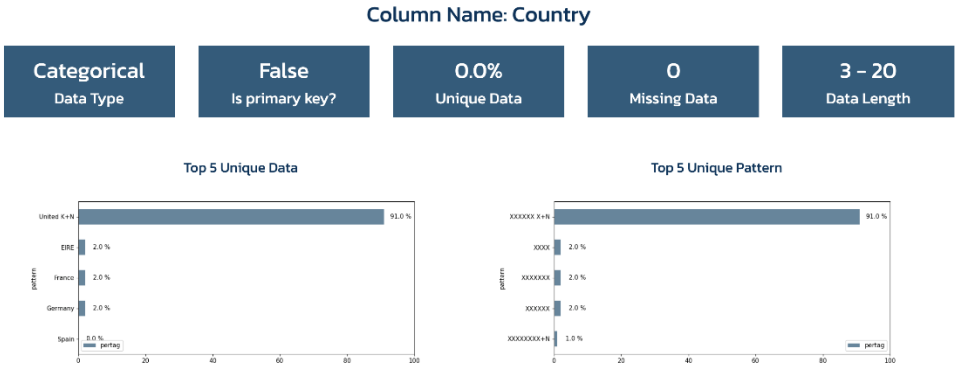
2.7) สดมภ์ – CustomerID



ภาพที่ 4.10 ผลลัพธ์การตรวจสอบของสดมภ์ CustomerID

จากการตรวจสอบของระบบพบว่า สดมภ์ CustomerID หรือ รหัสลูกค้า เป็นสดมภ์ประเภทหมวดหมู่ โดยมีค่าความเฉพาะของข้อมูลอยู่ที่ 1% หมายความว่าข้อมูลในสดมภ์นี้สามารถที่จะมีค่าซ้ำกันได้ กล่าวคือในแต่ละรหัสลูกค้าสามารถมีได้หลายรายการ นอกจากนี้ยังพบว่ามีข้อมูลสูญหายจำนวน 13,5037 รายการ หรือประมาณ 25% ซึ่งอาจเกิดจากการซื้อสินค้าจากลูกค้าที่ไม่ใช่สมาชิก และความยาวของข้อมูลอยู่ที่ 3 ถึง 7 ตัวอักษร

2.8) สดมภ์ – Country



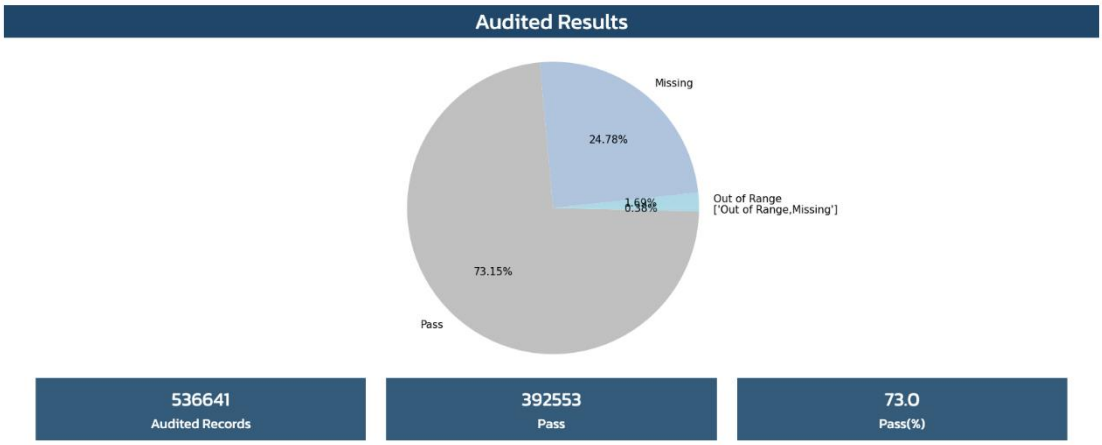
ภาพที่ 4.11 ผลลัพธ์การตรวจสอบของสดมภ์ Country

จากการตรวจสอบของระบบพบว่า สดมภ์ Country หรือ ประเทศ เป็นสดมภ์ประเภทหมวดหมู่ โดยมีค่าความเฉพาะของข้อมูลอยู่ที่ 0% หมายความว่าข้อมูลในสดมภ์นี้สามารถที่จะมีค่าซ้ำกันได้ กล่าวคือในแต่ละประเทศลูกค้าสามารถมีได้หลายรายการ นอกจากนี้ยังพบว่าไม่มีข้อมูลสูญหาย และมีความยาว 3 ถึง 20 ตัวอักษร โดยสดมภ์นี้เป็นลักษณะชื่อของประเทศจึงไม่นำความยาวและรูปแบบของตัวอักษรมาพิจารณา

การวิเคราะห์ผลการตรวจสอบคุณภาพข้อมูลตามเงื่อนไขที่กำหนด

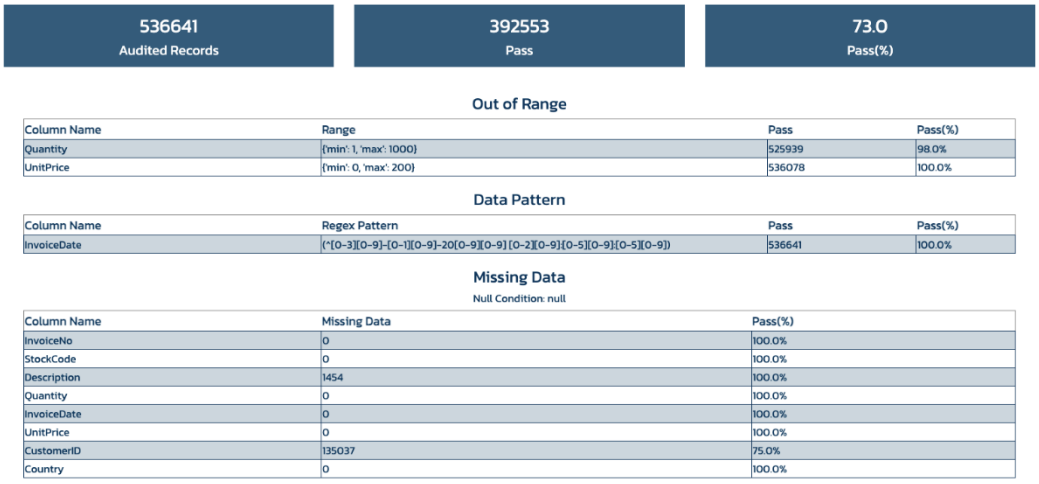
จากผลการวิเคราะห์คุณภาพข้อมูลเบื้องต้น สามารถกำหนดเงื่อนไขการตรวจสอบคุณภาพข้อมูลได้ดังนี้

1) ต้องไม่มีข้อมูลซ้ำซ้อน
2) ต้องไม่มีข้อมูลสูญหาย
3) สดมภ์ InvoiceDate ต้องมีรูปแบบ YYYY-MM-DD HH:MM:SS
4) สดมภ์ Quantity ต้องมีค่าต่ำสุดอยู่ที่ 1 และมากสุดอยู่ที่ 1,000
5) สดมภ์ UnitPrice ต้องมีค่าต่ำสุดอยู่ที่ 0 และมากสุดอยู่ที่ 200



ภาพที่ 4.13 ผลลัพธ์จากการตรวจสอบตามเงื่อนไขในรูปแบบกราฟวงกลม

ผลลัพธ์ในภาพรวมแสดงให้เห็นว่า มีจำนวนข้อมูลที่ถูกตรวจสอบจำนวน 536,641 รายการ ซึ่งเป็นจำนวนที่ตัดข้อมูลที่ซ้ำซ้อนออกแล้ว หลังจากนั้นจึงนำมาตรวจสอบตามเงื่อนไขที่กำหนด โดยพบว่ามีจำนวนที่ผ่านเงื่อนไขทั้งหมดจำนวน 392,553 รายการ หรือคิดเป็น 73.15% โดยรายการที่ตัดการจะเป็นรายการที่มีข้อมสูญหาย 24.78% ข้อมูลที่อยู่นอกพิสัย 1.69% และข้อมูลรายการที่มีทั้งข้อมูลสูญหายและนอกพิสัยด้วยอยู่ 0.38%



536641
Audited Records

392553
Pass

73.0
Pass(%)

Out of Range

Column Name	Range	Pass	Pass(%)
Quantity	{'min': 1, 'max': 1000}	525939	98.0%
UnitPrice	{'min': 0, 'max': 200}	536078	100.0%

Data Pattern

Column Name	Regex Pattern	Pass	Pass(%)
InvoiceDate	(^[0-3][0-9]-[0-1][0-9]-20[0-9][0-9] [0-2][0-9]:[0-5][0-9]:[0-5][0-9])	536641	100.0%

Missing Data

Null Condition: null

Column Name	Missing Data	Pass(%)
InvoiceNo	0	100.0%
StockCode	0	100.0%
Description	1454	100.0%
Quantity	0	100.0%
InvoiceDate	0	100.0%
UnitPrice	0	100.0%
CustomerID	135037	75.0%
Country	0	100.0%

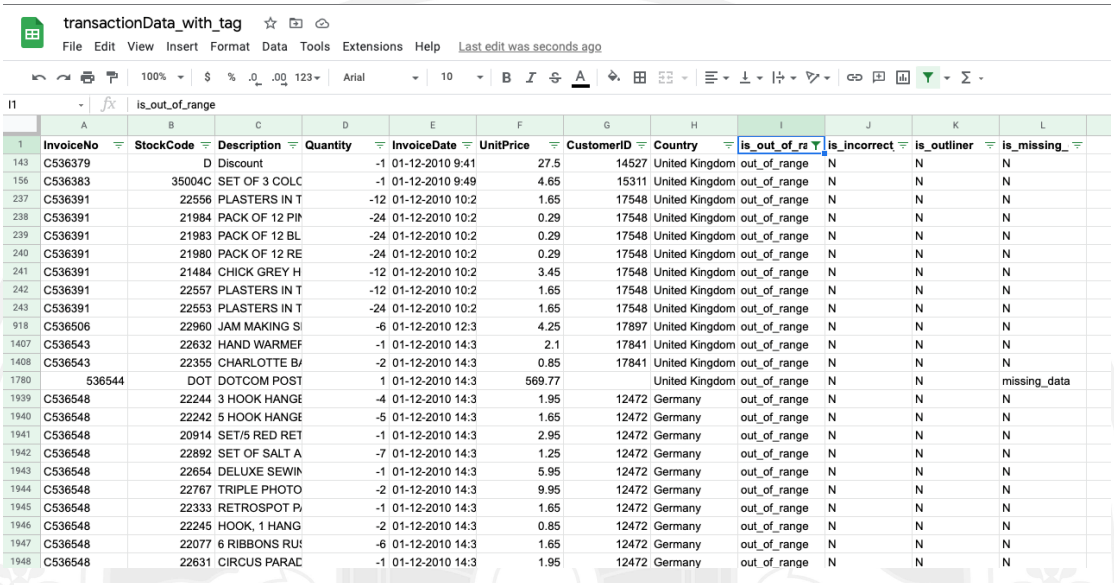
ภาพที่ 4.12 ผลลัพธ์จากการตรวจสอบตามเงื่อนไข

ผลลัพธ์การตรวจสอบคุณภาพข้อมูลแยกตามเงื่อนไขที่กำหนด

1) การตรวจสอบพิสัยของข้อมูลในสดมภ์ Quantity พบว่ามีข้อมูลผ่านตามเงื่อนไขจำนวน 98% ในขณะที่ข้อมูลในสดมภ์ผ่านตามเงื่อนไข 100%

2) การตรวจสอบรูปแบบข้อมูลในสดมภ์ InvoiceDate พบว่ามีข้อมูลผ่านตามเงื่อนไข 100%

3) การตรวจสอบข้อมูลสูญหายในแต่ละสดมภ์ พบว่ามี 2 สดมภ์ที่มีข้อมูลสูญหาย ได้แก่ สดมภ์ CustomerID โดยมีจำนวนข้อมูลสูญหาย 135,037 หรือคิดเป็น 25% จากข้อมูลทั้งหมด และสดมภ์ Description โดยมีจำนวนข้อมูลสูญหาย 1,454 หรือคิดเป็น 0.002% จากข้อมูลทั้งหมด



transactionData_with_tag

File Edit View Insert Format Data Tools Extensions Help Last edit was seconds ago

I1 is_out_of_range

	A	B	C	D	E	F	G	H	I	J	K	L
1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	is_out_of_ra	is_incorrect	is_outliner	is_missing_
143	C536379	D	Discount	-1	01-12-2010 9:41	27.5	14527	United Kingdom	out_of_range	N	N	N
156	C536383	35004C	SET OF 3 COLC	-1	01-12-2010 9:49	4.65	15311	United Kingdom	out_of_range	N	N	N
237	C536391	22556	PLASTERS IN T	-12	01-12-2010 10:2	1.65	17548	United Kingdom	out_of_range	N	N	N
238	C536391	21984	PACK OF 12 PIN	-24	01-12-2010 10:2	0.29	17548	United Kingdom	out_of_range	N	N	N
239	C536391	21983	PACK OF 12 BL	-24	01-12-2010 10:2	0.29	17548	United Kingdom	out_of_range	N	N	N
240	C536391	21980	PACK OF 12 RE	-24	01-12-2010 10:2	0.29	17548	United Kingdom	out_of_range	N	N	N
241	C536391	21484	CHICK GREY H	-12	01-12-2010 10:2	3.45	17548	United Kingdom	out_of_range	N	N	N
242	C536391	22557	PLASTERS IN T	-12	01-12-2010 10:2	1.65	17548	United Kingdom	out_of_range	N	N	N
243	C536391	22553	PLASTERS IN T	-24	01-12-2010 10:2	1.65	17548	United Kingdom	out_of_range	N	N	N
918	C536506	22960	JAM MAKING SI	-6	01-12-2010 12:3	4.25	17897	United Kingdom	out_of_range	N	N	N
1407	C536543	22632	HAND WARMEF	-1	01-12-2010 14:3	2.1	17841	United Kingdom	out_of_range	N	N	N
1408	C536543	22355	CHARLOTTE BA	-2	01-12-2010 14:3	0.85	17841	United Kingdom	out_of_range	N	N	N
1780	536544	DOT	DOTCOM POST	1	01-12-2010 14:3	569.77		United Kingdom	out_of_range	N	N	missing_data
1939	C536548	22244	3 HOOK HANGE	-4	01-12-2010 14:3	1.95	12472	Germany	out_of_range	N	N	N
1940	C536548	22242	5 HOOK HANGE	-5	01-12-2010 14:3	1.65	12472	Germany	out_of_range	N	N	N
1941	C536548	20914	SET/5 RED RET	-1	01-12-2010 14:3	2.95	12472	Germany	out_of_range	N	N	N
1942	C536548	22892	SET OF SALT A	-7	01-12-2010 14:3	1.25	12472	Germany	out_of_range	N	N	N
1943	C536548	22654	DELUXE SEWIN	-1	01-12-2010 14:3	5.95	12472	Germany	out_of_range	N	N	N
1944	C536548	22767	TRIPLE PHOTO	-2	01-12-2010 14:3	9.95	12472	Germany	out_of_range	N	N	N
1945	C536548	22333	RETROSPOT PA	-1	01-12-2010 14:3	1.65	12472	Germany	out_of_range	N	N	N
1946	C536548	22245	HOOK, 1 HANG	-2	01-12-2010 14:3	0.85	12472	Germany	out_of_range	N	N	N
1947	C536548	22077	6 RIBBONS RUS	-6	01-12-2010 14:3	1.65	12472	Germany	out_of_range	N	N	N
1948	C536548	22631	CIRCUS PARAD	-1	01-12-2010 14:3	1.95	12472	Germany	out_of_range	N	N	N

ภาพที่ 4.14 ผลลัพธ์ในรูปแบบไฟล์ตาราง

นอกจากนี้ระบบยังให้ผลลัพธ์ในรูปแบบไฟล์ตารางที่ผู้ใช้งานสามารถนำไปกรองข้อมูลเพิ่มเติมตามเงื่อนไขที่กำหนดได้โดยระบบจะทำการเพิ่มสดมภ์มาให้จำนวน 4 สดมภ์ ได้แก่ is_out_of_range, is_incorrect_pattern, is_outliner, is_missing_data โดยแต่ละสดมภ์จะเป็นตัวบอกว่ารายการนั้น ๆ ผ่านหรือไม่ผ่านเงื่อนไขใด นอกจากนี้ยังสามมารถนำผลลัพธ์ในข้อ 4.1.1 และ 4.1.2 สามารถนำมาสรุปอยู่ในมิติของคุณภาพข้อมูลได้ดังนี้

1) มิติด้านความสมบูรณ์

1.1) สัดส่วนความครบถ้วนของข้อมูลทั้งหมด

จำนวนข้อมูลทั้งหมด = 536,641 x 8 = 4,293,128

จำนวนข้อมูลสูญหายทั้งหมด = 1,454 + 135,037 = 136,491

สัดส่วนความครบถ้วนของข้อมูลทั้งหมด = 1 - (136,491/4,293,128) = 0.97

1.2) สัดส่วนสดมภ์ที่มีข้อมูล

จำนวนสดมภ์ทั้งหมด = 8

จำนวนสดมภ์ที่มีข้อมูล = 8

สัดส่วนสดมภ์ที่มีข้อมูล = 8/8 =1.0

1.3) สัดส่วนสดมภ์ที่มีข้อมูลครบถ้วน

จำนวนสดมภ์ทั้งหมด = 8

จำนวนสดมภ์ที่มีข้อมูลครบถ้วน = 6

สัดส่วนสดมภ์ที่มีข้อมูลครบถ้วน = 6/8

= 0.75

1.4) สัดส่วนแถวข้อมูลที่มีข้อมูลครบถ้วนทุกสดมภ์

จำนวนแถวข้อมูลทั้งหมด = 536,641

จำนวนแถวข้อมูลที่มีข้อมูลครบถ้วน = 401,604

สัดส่วนแถวข้อมูลที่มีข้อมูลครบถ้วนทุกสดมภ์ = 401,604/536,641 = 0.75

2) มิติด้านความคงเส้นคงวา

2.1) สัดส่วนจำนวนสดมภ์ข้อมูลที่มีรูปแบบข้อมูลภายในเหมือนกันทั้งหมด

จำนวนแถวข้อมูลทั้งหมด = 536,641

จำนวนแถวข้อมูลที่มีรูปแบบถูกต้อง = 536,641

สัดส่วนจำนวนสดมภ์ข้อมูลที่มีรูปแบบข้อมูลภายในเหมือนกันทั้งหมด

= 536,641/536,641= 1.0

3) มิติด้านความแม่นยำ

3.1) สัดส่วนจำนวนแถวข้อมูลที่ไม่มีข้อมูลซ้ำซ้อน

จำนวนแถวข้อมูลทั้งหมด = 536,641

จำนวนแถวข้อมูลซ้ำซ้อน = 5,268

สัดส่วนจำนวนแถวข้อมูลที่ไม่มีข้อมูลซ้ำซ้อน = 1 – (5,268/536,641)

= 0.99

4) มิติด้านความถูกต้อง

4.1) สัดส่วนจำนวนข้อมูลที่มีรูปแบบข้อมูลถูกต้องตามที่กำหนด

จำนวนสดมภ์ที่ทำการตรวจสอบ = 3

จำนวนข้อมูลทั้งหมดที่ทำการตรวจสอบ = 536,641 x 3 = 1,609,923

จำนวนข้อมูลที่ผ่านการทดสอบ = 525,939 + 536,078 + 536,641

= 1,598,658

สัดส่วนจำนวนข้อมูลที่มีรูปแบบข้อมูลถูกต้องตามที่กำหนด

= 1,598,658/1,609,923

= 0.99

บทที่ 5

สรุปผล อภิปรายผล และข้อเสนอแนะ

5.1 สรุปผล

จากผลการศึกษาและพัฒนานี้ พบว่าสามารถคัดเลือกตัวชี้วัดที่สามารถวัดคุณภาพของชุดข้อมูลในเชิงปฏิบัติได้ จำนวน 7 ตัวชี้วัดพร้อมวิธีการคำนวณ ได้แก่ จำนวนข้อมูลสูญหายแต่ละสดมภ์ สัดส่วนของข้อมูลสูญหาย จำนวนสดมภ์ข้อมูลที่มีข้อมูลสูญหายโดยสมบูรณ์ สัดส่วนจำนวนแถวข้อมูลที่สมบูรณ์ สัดส่วนจำนวนรูปแบบข้อมูล จำนวนแถวข้อมูลที่ซ้ำกันโดยสมบูรณ์ และสัดส่วนรูปแบบข้อมูลที่ถูกต้อง โดยผู้วิจัยได้นำตัวชี้วัดเหล่านี้มาพัฒนาเป็นระบบตรวจสอบคุณภาพข้อมูลโดยใช้ภาษา Python ในการพัฒนา ซึ่งได้ออกแบบระบบออกเป็น 3 ส่วน ได้แก่ ส่วนกำหนดความต้องการ ส่วนวิเคราะห์ข้อมูล และส่วนสร้างภาพข้อมูล โดยการแยกส่วนนี้เพื่อให้ง่ายต่อการแก้ไข ปรับปรุงระบบในอนาคต ซึ่งภายหลังเมื่อได้พัฒนาและทดสอบระบบเสร็จสินแล้ว พบว่าระบบสามารถทำงานได้ โดยสามารถลดเวลาในการตรวจสอบข้อมูลและผลลัพธ์ทที่ได้ยังช่วยในการวิเคราะห์แยกแยะคุณภาพของข้อมูลได้เป็นอย่างดี

5.2 อภิปรายผล

จากผลสรุปที่ได้จากการศึกษาและพัฒนานี้ จะเห็นว่าการพัฒนาระบบตรวจสอบคุณภาพข้อมูลโดยใช้เทคโนโลยีด้านคอมพิวเตอร์เข้ามาช่วยในการทำงาน ทำให้สามารถลดเวลาในการทำงานและเพิ่มคุณภาพในการทำงานได้เป็นอย่างดี โดยเฉพาะอย่างยิ่งในกรณีที่มีข้อมูลจำนวนมาก จะช่วยให้ลดเวลาในการทำงานได้อย่างชัดเจน แต่ทั้งนี้ยังต้องอาศัยแรงงานบางส่วนในการกำหนดรายละเอียดการตรวจสอบคุณภาพข้อมูลที่อาจจะมีเพิ่มเติมในแต่ละชุดข้อมูล โดยอาจจะมีการกำหนดเกณฑ์การวัดที่เป็นมาตราฐานไว้ส่วนหนึ่งที่สามารถใช้กับชุดข้อมูลโดยทั่วไปได้ กับอีกส่วนหนึ่งที่เฉพาะสำหรับชุดข้อมูลแต่หละประเภท ในด้านประสิทธิภาพในการประมวลผลนั้น ความเร็ว และขนาดของข้อมูลที่รองรับได้นั้น จะอยู่ที่ประสิทธิภาพของเครื่องคอมพิวเตอร์ที่ทำการติดตั้งระบบ โดยพบว่าในคอมพิวเตอร์ส่วนบุคคลสามารถใช้ประมวลผลข้อมูลที่ไม่มีการบีบอัดขนาด 500MB ได้โดยไม่เกิดข้อผิดพลาดแต่ประการใด ซึ่งถือว่าเป็นขนาดข้อมูลที่เพียงพอต่อการใช้งานทั่วไป เนื่องจาก

โดยทั่วไปแล้วข้อมูลที่ขนาดใหญ่จริง ๆ จะเป็นข้อมูลประเภทการทำรายการ (Transaction Data) หรือข้อมูลประเภทบันทึกของระบบ (Logging/Sensor Data) ที่อาจจะขนาด 1 GB ถึง หลาย ๆ TB ซึ่งเป็นส่วนน้อยในระบบฐานข้อมูล ซึ่งในกรณีนี้อาจจะต้องใช้เทคโนโลยีการประมวลผลข้อมูลขนาดใหญ่เข้ามาช่วย หรืออาจจะนำข้อมูลขนาดใหญ่เหล่านี้มาแบ่งให้เล็กลง เช่น รายเดือน หรือรายปี ก็จะยังสามารถใช้ระบบที่พัฒนานี้ได้เช่นกัน

5.3 ข้อเสนอแนะ

เนื่องจากการพัฒนาระบบตรวจสอบข้อมูลนี้ ถูกพัฒนาให้รองรับระบบปฏิบัติการ Linux และ MacOS เป็นหลัก เนื่องจากระบบดังกล่างมักถูกใช้ในการทำงานของระบบวิศกรรมข้อมูลแบบอัติโนมัติ จึงทำให้ผู้ใช้งานบางส่วนไม่สามารถใช้งานได้บนระบบปฏิบัติการ Microsoft Windows ผู้วิจัยและพัฒนาจึงเห็นว่าในกรณีนี้ อาจจะมีการพัฒนาระบบในรูปแบบของเว็บแอพพลิเคชั่นอีกช่องทางหนึ่ง เพื่อให้ง่ายต่อการใช้งานแก่ผู้ใช้งานทั่วไป นอกจากนี้ในการเพิ่มประสิทธิภาพในการทำงานของระบบให้รองรับการประมวลผลข้อมูลขนาดใหญ่ หรือรองรับการทำงานโดยเชื่อมต่อกับระบบฐานข้อมูลโดยตรงก็สามารถทำได้ด้วยเช่นเดียวกัน โดยทำการปรับแก้ไขรหัสโปรแกรมของระบบในส่วนของการประมวลผลข้อมูลจาก Pandas เป็น Database Engine ก็ได้เช่นเดียวกัน ทำให้การเชื่อมต่อระบบตรวจสอบคุณภาพข้อมูลกับระบบฐานข้อมูลมีประสิทธิภาพสูงสุด นอกจากนี้ในยังสามารถเพิ่มเติมเทคนิคการวิเคราะห์ต่าง ๆ ให้แก่ระบบเพื่อเพิ่มศักยภาพของระบบ อาทิเช่น การวิเคราะห์ข้อมูลที่มีปริมาณร้อย (Sparseness) หรือ การวิเคราะห์ความคล้ายกันของข้อมูล (Similarity) ที่ทำการวิเคราะห์ เป็นต้น

บรรณานุกรม

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ***ACM Comput. Surv.***, *41*(3), Article 16. doi:10.1145/1541880.1541883

Chmielewski, M., & Kucker, S. C. (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. ***Social Psychological and Personality Science***, *11*(4), 464-473. doi:10.1177/1948550619875149

Earley, S. H. D. D. M. A. (2017). ***DAMA-DMBOK : data management body of knowledge***.

Gorla, N., Somers, T. M., & Wong, B. (2010). Organizational impact of system quality, information quality, and service quality. ***The Journal of Strategic Information Systems***, *19*(3), 207-228.

Haug, A., Zachariassen, F., & Liempd, D. (2011). The costs of poor data quality. ***Journal of Industrial Engineering and Management***, *4*. doi:10.3926/jiem..v4n2.p168-193

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. ***Journal of Business Research***, *70*, 338-345.

Sebastian-Coleman, L. (2013). ***Measuring Data Quality for Ongoing Improvement***.

Slota, S. C., Fleischmann, K. R., Greenberg, S., Verma, N., Cummings, B., Li, L., & Shenefiel, C. (2020). ***Good systems, bad data?: Interpretations of AI hype and failures***. Proceedings of the Association for Information Science and Technology, 57(1), e275.

Srinath, K. R. (2017). *Python* ***The Fastest Growing Programming Language***.

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. ***Journal of Management Information Systems***, *12*(4), 5-33. doi:10.1080/07421222.1996.11518099

ภาคผนวก

ภาคผนวก ก

ชุดคำสั่งสำหรับตรวจสอบข้อมูลเบื้องต้น

```
import sakdas as sd
import pandas as pd

df = pd.read_csv('/Users/sakdaloetpipatwanich/Documents/sakdas_prod/data.csv')

get_profile =
sd.load(df,'transactionData','/Users/sakdaloetpipatwanich/Documents/sakdas_prod')
```

ภาคผนวก ข

ชุดคำสั่งสำหรับตรวจสอบข้อมูลแบบกำหนดเงื่อนไข

```
import sakdas as sd
import pandas as pd

df =
pd.read_csv('/Users/sakdaloetpipatwanich/Documents/sakdas_prod/data_formated_f
ull.csv')


auditing_config = {'audit':{
    'audit_missing_value': True,
    'audit_data_pattern':[
                {'column_name':'InvoiceDate', 'regex_pattern': '(^[0-3][0-9]-[0-1][0-9]-20[0-
9][0-9] [0-2][0-9]:[0-5][0-9]:[0-5][0-9])'}
            ],
    'audit_outlier': False,
    'audit_data_range' : [{'column_name':'Quantity', 'min': 1, 'max': 1000}
                        ,{'column_name':'UnitPrice', 'min': 0, 'max': 200}]
    }
}

auditDataset =
sd.load(df,'transactionData','/Users/sakdaloetpipatwanich/Documents/sakdas_prod',
auditing_config = auditing_config)
```

ภาคผนวก ค

ตัวอย่างผลลัพธ์จากระบบในรูปแบบ JSON

```
{
    "profile_engine": "4.0.16",
    "profile_id": "1625a9d5-bade-4dc1-bc2e-7c2141783712",
    "data_name": "transactionData",
    "profiling_datetime": "2021-11-22T17:05:32+0700",
    "number_of_record": 541909,
    "number_of_record_after_dedup": 536641,
    "number_of_dupRecord": 5268,
    "column_list": [
        {
            "columnName": "InvoiceNo",
            "dataType": "Categorical"
        },
        {
            "columnName": "StockCode",
            "dataType": "Categorical"
        },
        {
            "columnName": "Description",
            "dataType": "Categorical"
        },
        {
            "columnName": "Quantity",
            "dataType": "Numerical"
        },
        {
            "columnName": "InvoiceDate",
            "dataType": "Categorical"
```

```
    },
    {
        "columnName": "UnitPrice",
        "dataType": "Numerical"
    },
    {
        "columnName": "CustomerID",
        "dataType": "Categorical"
    },
    {
        "columnName": "Country",
        "dataType": "Categorical"
    }
],
"number_of_primary_key": 0,
"primary_key_column": "",
"number_of_column": 8,
"number_of_complete_record": 401604,
"ratio_of_complete_record": 0.75,
"ratio_of_complete_column": 0.75,
"number_of_blank_column": 0,
"number_of_complete_column": 6,
"number_of_missing_data": 136491,
"ratio_of_missing_data": 0.03,
"column_profile": {
    "InvoiceNo": {
        "dataType": "Categorical",
        "is_primary_key": false,
        "distinct_value": 25900,
        "ratio_distinct_value": 0.05,
        "missing_value": 0,
```

"ratio_missing_value": 0.0,
"data_lenght_min": 6,
"data_lenght_max": 7,
"data_lenght_mean": 6.02,
"data_lenght_mode": 6,
"data_lenght_median": 6.0,
"data_lenght_std": 0.13,
"data_lenght_var": 0.02,
"top_5_data_value": [
{
"data_value": "573585",
"count_data_value": 1114,
"value_ratio": 0.0
},
{
"data_value": "581219",
"count_data_value": 749,
"value_ratio": 0.0
},
{
"data_value": "581492",
"count_data_value": 731,
"value_ratio": 0.0
},
{
"data_value": "580729",
"count_data_value": 721,
"value_ratio": 0.0
},
{
"data_value": "558475",

```
            "count_data_value": 705,
            "value_ratio": 0.0
        }
    ],
    "top_5_data_pattern": [
        {
            "pattern": "NNNNNN",
            "ratio": 0.98
        },
        {
            "pattern": "XNNNNNN",
            "ratio": 0.02
        }
    ],
"StockCode": {
    "dataType": "Categorical",
    "is_primary_key": false,
    "distinct_value": 4070,
    "ratio_distinct_value": 0.01,
    "missing_value": 0,
    "ratio_missing_value": 0.0,
    "data_lenght_min": 1,
    "data_lenght_max": 12,
    "data_lenght_mean": 5.09,
    "data_lenght_mode": 5,
    "data_lenght_median": 5.0,
    "data_lenght_std": 0.36,
    "data_lenght_var": 0.13,
    "top_5_data_value": [
        {
            "data_value": "85123A",
```

```
            "count_data_value": 2301,
            "value_ratio": 0.0
        },
        {
            "data_value": "22423",
            "count_data_value": 2192,
            "value_ratio": 0.0
        },
        {
            "data_value": "85099B",
            "count_data_value": 2156,
            "value_ratio": 0.0
        },
        {
            "data_value": "47566",
            "count_data_value": 1720,
            "value_ratio": 0.0
        },
        {
            "data_value": "20725",
            "count_data_value": 1626,
            "value_ratio": 0.0
        }
    ],
    "top_5_data_pattern": [
        {
            "pattern": "NNNNN",
            "ratio": 0.9
        },
        {
            "pattern": "NNNNNX",
```

```
            "ratio": 0.1
        },
        {
            "pattern": "XXXX",
            "ratio": 0.0
        },
        {
            "pattern": "NNNNNXX",
            "ratio": 0.0
        },
        {
            "pattern": "X",
            "ratio": 0.0
        }
    ],
"Description": {
    "dataType": "Categorical",
    "is_primary_key": false,
    "distinct_value": 4211,
    "ratio_distinct_value": 0.01,
    "missing_value": 1454,
    "ratio_missing_value": 0.0,
    "data_lenght_min": 1,
    "data_lenght_max": 35,
    "data_lenght_mean": 26.36,
    "data_lenght_mode": 30,
    "data_lenght_median": 27.0,
    "data_lenght_std": 5.58,
    "data_lenght_var": 31.13,
    "top_5_data_value": [
        {
```

```
            "data_value": "WHITE HANGING HEART T-LIGHT HOLDER",
            "count_data_value": 2357,
            "value_ratio": 0.0
        },
        {
            "data_value": "REGENCY CAKESTAND 3 TIER",
            "count_data_value": 2189,
            "value_ratio": 0.0
        },
        {
            "data_value": "JUMBO BAG RED RETROSPOT",
            "count_data_value": 2156,
            "value_ratio": 0.0
        },
        {
            "data_value": "PARTY BUNTING",
            "count_data_value": 1720,
            "value_ratio": 0.0
        },
        {
            "data_value": "LUNCH BAG RED RETROSPOT",
            "count_data_value": 1625,
            "value_ratio": 0.0
        }
    ],
    "top_5_data_pattern": [
        {
            "pattern": "XXXXX XXX XXXXXXXX XXXXXX",
            "ratio": 0.01
        },
        {
```

```
            "pattern": "XXXXX XXX XXX XXXXXXXXX",
            "ratio": 0.01
        },
        {
            "pattern": "XXXXX XXXXXXX XXXXX X-XXXXX XXXXXX",
            "ratio": 0.0
        },
        {
            "pattern": "XXXXX XXXXXX XXXXX XXXX XXXXX",
            "ratio": 0.0
        },
        {
            "pattern": "XXXXXXX XXXXX XXXXXXXXX,XXXX",
            "ratio": 0.0
        }
    ],
"Quantity": {
    "dataType": "Numerical",
    "is_primary_key": false,
    "distinct_value": 722,
    "ratio_distinct_value": 0.0,
    "missing_value": 0,
    "ratio_missing_value": 0.0,
    "data_lenght_min": 1,
    "data_lenght_max": 6,
    "data_lenght_mean": 1.33,
    "data_lenght_mode": 1,
    "data_lenght_median": 1.0,
    "data_lenght_std": 0.51,
    "data_lenght_var": 0.26,
    "min_value": -80995.0,
```

"max_value": 80995.0,
"mean": 9.62,
"median": 3.0,
"mode": 1.0,
"variance": 48018.03,
"std": 219.13,
"first_quartile": 1.0,
"third_quartile": 10.0,
"iqr": 9.0,
"minimum": -3.5,
"maximum": 14.5,
"negative_outliner_datapoint": 69340,
"positive_outliner_datapoint": 4413,
"top_5_data_value": [
{
"data_value": 1,
"count_data_value": 144495,
"value_ratio": 0.27
},
{
"data_value": 2,
"count_data_value": 81245,
"value_ratio": 0.15
},
{
"data_value": 12,
"count_data_value": 60858,
"value_ratio": 0.11
},
{
"data_value": 6,

```
            "count_data_value": 40656,
            "value_ratio": 0.08
        },
        {
            "data_value": 4,
            "count_data_value": 38393,
            "value_ratio": 0.07
        }
    ],
    "top_5_data_pattern": [
        {
            "pattern": "N",
            "ratio": 0.69
        },
        {
            "pattern": "NN",
            "ratio": 0.28
        },
        {
            "pattern": "NNN",
            "ratio": 0.01
        },
        {
            "pattern": "-N",
            "ratio": 0.01
        },
        {
            "pattern": "-NN",
            "ratio": 0.0
        }
    ],
```

```
"InvoiceDate": {
    "dataType": "Categorical",
    "is_primary_key": false,
    "distinct_value": 23260,
    "ratio_distinct_value": 0.04,
    "missing_value": 0,
    "ratio_missing_value": 0.0,
    "data_lenght_min": 19,
    "data_lenght_max": 19,
    "data_lenght_mean": 19.0,
    "data_lenght_mode": 19,
    "data_lenght_median": 19.0,
    "data_lenght_std": 0.0,
    "data_lenght_var": 0.0,
    "top_5_data_value": [
        {
            "data_value": "31-10-2011 14:41:00",
            "count_data_value": 1114,
            "value_ratio": 0.0
        },
        {
            "data_value": "08-12-2011 09:28:00",
            "count_data_value": 749,
            "value_ratio": 0.0
        },
        {
            "data_value": "09-12-2011 10:03:00",
            "count_data_value": 731,
            "value_ratio": 0.0
        },
        {
```

```
            "data_value": "05-12-2011 17:24:00",
            "count_data_value": 721,
            "value_ratio": 0.0
        },
        {
            "data_value": "29-06-2011 15:58:00",
            "count_data_value": 705,
            "value_ratio": 0.0
        }
    ],
    "top_5_data_pattern": [
        {
            "pattern": "NN-NN-NNNN NN:NN:NN",
            "ratio": 1.0
        }
    ],
"UnitPrice": {
    "dataType": "Numerical",
    "is_primary_key": false,
    "distinct_value": 1630,
    "ratio_distinct_value": 0.0,
    "missing_value": 0,
    "ratio_missing_value": 0.0,
    "data_lenght_min": 3,
    "data_lenght_max": 9,
    "data_lenght_mean": 3.99,
    "data_lenght_mode": 4,
    "data_lenght_median": 4.0,
    "data_lenght_std": 0.31,
    "data_lenght_var": 0.1,
    "min_value": -11062.06,
```

"max_value": 38970.0,
"mean": 4.63,
"median": 2.08,
"mode": 1.25,
"variance": 9454.28,
"std": 97.23,
"first_quartile": 1.25,
"third_quartile": 4.13,
"iqr": 2.88,
"minimum": -0.19,
"maximum": 5.57,
"negative_outliner_datapoint": 85419,
"positive_outliner_datapoint": 2,
"top_5_data_value": [
{
"data_value": 1.25,
"count_data_value": 49750,
"value_ratio": 0.09
},
{
"data_value": 1.65,
"count_data_value": 37627,
"value_ratio": 0.07
},
{
"data_value": 0.85,
"count_data_value": 28182,
"value_ratio": 0.05
},
{
"data_value": 2.95,

```
        "count_data_value": 27350,
        "value_ratio": 0.05
    },
    {
        "data_value": 0.42,
        "count_data_value": 24277,
        "value_ratio": 0.05
    }
],
"top_5_data_pattern": [
    {
        "pattern": "N.NN",
        "ratio": 0.9
    },
    {
        "pattern": "N.N",
        "ratio": 0.05
    },
    {
        "pattern": "NN.NN",
        "ratio": 0.04
    },
    {
        "pattern": "NN.N",
        "ratio": 0.01
    },
    {
        "pattern": "NNN.N",
        "ratio": 0.0
    }
],
```

```
"CustomerID": {
    "dataType": "Categorical",
    "is_primary_key": false,
    "distinct_value": 4372,
    "ratio_distinct_value": 0.01,
    "missing_value": 135037,
    "ratio_missing_value": 0.25,
    "data_lenght_min": 3,
    "data_lenght_max": 7,
    "data_lenght_mean": 5.99,
    "data_lenght_mode": 7,
    "data_lenght_median": 7.0,
    "data_lenght_std": 1.74,
    "data_lenght_var": 3.01,
    "top_5_data_value": [
        {
            "data_value": 17841.0,
            "count_data_value": 7812,
            "value_ratio": 0.01
        },
        {
            "data_value": 14911.0,
            "count_data_value": 5898,
            "value_ratio": 0.01
        },
        {
            "data_value": 14096.0,
            "count_data_value": 5128,
            "value_ratio": 0.01
        },
        {
```

```
            "data_value": 12748.0,
            "count_data_value": 4459,
            "value_ratio": 0.01
        },
        {
            "data_value": 14606.0,
            "count_data_value": 2759,
            "value_ratio": 0.01
        }
    ],
    "top_5_data_pattern": [
        {
            "pattern": "NNNNN.N",
            "ratio": 0.75
        },
        {
            "pattern": "",
            "ratio": 0.25
        }
    ],
"Country": {
    "dataType": "Categorical",
    "is_primary_key": false,
    "distinct_value": 38,
    "ratio_distinct_value": 0.0,
    "missing_value": 0,
    "ratio_missing_value": 0.0,
    "data_lenght_min": 3,
    "data_lenght_max": 20,
    "data_lenght_mean": 13.37,
    "data_lenght_mode": 14,
```

```
"data_lenght_median": 14.0,
"data_lenght_std": 2.16,
"data_lenght_var": 4.66,
"top_5_data_value": [
    {
        "data_value": "United Kingdom",
        "count_data_value": 490300,
        "value_ratio": 0.91
    },
    {
        "data_value": "Germany",
        "count_data_value": 9480,
        "value_ratio": 0.02
    },
    {
        "data_value": "France",
        "count_data_value": 8541,
        "value_ratio": 0.02
    },
    {
        "data_value": "EIRE",
        "count_data_value": 8184,
        "value_ratio": 0.02
    },
    {
        "data_value": "Spain",
        "count_data_value": 2528,
        "value_ratio": 0.0
    }
],
"top_5_data_pattern": [
```

```
            {
                "pattern": "XXXXXX XXXXXXX",
                "ratio": 0.91
            },
            {
                "pattern": "XXXXXX",
                "ratio": 0.02
            },
            {
                "pattern": "XXXXXXX",
                "ratio": 0.02
            },
            {
                "pattern": "XXXX",
                "ratio": 0.02
            },
            {
                "pattern": "XXXXXXXXXXX",
                "ratio": 0.01
            }
        ]
    }
  }
}
```

ประวัติผู้เขียน

ชื่อ-นามสกุล ศักดา เลิศพิพัฒน์วาณิชย์

ประวัติการศึกษา วิทยาศาสตร์บัณฑิต
มหาวิทยาลัยรามคำแหง
ปีที่สำเร็จการศึกษา พ.ศ. 2553

ประสบการณ์การทำงาน พ.ศ. 2562 - ปัจจุบัน
ผู้เชี่ยวชาญแพลตฟอร์มข้อมูล
บริษัท แอสเซนด์ กรุ๊ป จำกัด
กรุงเทพมหานคร 101400

พ.ศ. 2560-2562
วิศกรข้อมูล, ผู้นำทีมวิศกรข้อมูล
บริษัท เซอร์ทิส จำกัด
กรุงเทพมหานคร 101110

พ.ศ. 2558-2560
วิศวกรเทคโนโลยีสารสนเทศ 2
สำนักงานพัฒนารัฐบาลดิจิทัล (องค์การมหาชน)
กรุงเทพมหานคร 10400

พ.ศ. 2557-2558
นักวิเคราะห์ระบบ
บริษัท โปรซอฟท์ คอมเทค จำกัด
กรุงเทพมหานคร 10240