บทที่ 1

บทนำ

# 1.1 ความเป็นมาและความสำคัญของปัญหา

ในยุคปัจจุบันที่โลกและธุรกิจกำลังถูกขับเคลื่อนด้วยข้อมูลจำนวนมหาศาลจนทำให้คำว่า “Big Data” เป็นที่รู้จักมากขึ้น หลายคนอาจเคยได้ยินคำว่า “Big Data” และแปลความหมายตรงตัวว่าคือข้อมูลที่มีขนาดใหญ่เพียงเท่านั้น ไม่ได้สนใจว่าจะนำข้อมูลเหล่านั้นไปใช้ประโยชน์ได้อย่างไร และวิธีการจัดการข้อมูลจำนวนมากนี้ทำอย่างไร ทำให้ข้อมูลที่มีไร้ซึ่งประโยชน์ และกลายเป็นขยะข้อมูลสูญเปล่า จึงทำให้มีนักพัฒนาต่าง ๆ ได้พัฒนาภาษาคอมพิวเตอร์ขึ้นมา เพื่อนำมาประมวลผลและจัดการกับสิ่งเรียกว่า “Big Data” ภาษา R เป็นหนึ่งในภาษาที่ได้รับความนิยมในการนำมาจัดการกับข้อมูลขนาดใหญ่ เนื่องจากเป็นภาษาที่เข้าใจง่าย มีความสามารถในการวิเคราะห์ Data Frame ซึ่งเป็นหัวใจของการทำงาน Data Analysis การใช้งานเพื่อวิเคราะห์ Data Sciences อีกทั้งยังรองรับระบบปฏิบัติการหลายระบบ ปฏิบัติการได้ทั้ง Windows, Mac, OS หรือ Linux และเหตุผลสำคัญที่สุดคือสามารถดาวน์โหลด R และ RStudio มาใช้งานได้ฟรีเนื่องจากเป็น Open Source

เนื่องจากภาษา R มีความสามารถในการวิเคราะห์ข้อมูล แต่ยังมีข้อจำกัดหลายประการ ไม่ว่าจะเป็นการเชื่อมต่อกับฐานข้อมูลหรือการทำ Visualization หากต้องการเพิ่มประสิทธิภาพในการวิ  
เคราะห์ข้อมูล จำเป็นจะต้องใช้แพ็กเกจอื่นร่วมด้วยซึ่ง Dplyr เป็นแพ็กเกจหนึ่งที่สามารถติดตั้งและใช้งานได้ง่าย มีแหล่งข้อมูลจำนวนมากให้ศึกษา และสามารถเชื่อมต่อกับฐานข้อมูลได้สะดวก ด้วยเหตุผลดังกล่าวนี้ จึงทำให้คณะผู้จัดทำเลือกแพ็กเกจนี้มาใช้ในการวิเคราะห์ข้อมูลเป็นหลัก และทำ  
การศึกษาแพ็กเกจที่สามารถทำงานร่วมกับ Dplyr เพื่อนำไปใช้ในการวิเคราะห์ข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพ โดยเรียกใช้ความสามารถของแพ็กเกจ Dplyr และแพ็กเกจอื่น ๆ ร่วมกัน และเมื่อเปรียบเทียบกับ R ที่ไม่มีการเรียกใช้แพ็กเกจแล้ว สามารถวัดประสิทธิภาพการทำงานของแพ็กเกจได้ โดยทางคณะผู้จัดทำได้เลือกข้อมูลเกี่ยวกับการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 หรือที่เรียกว่า “COVID-19” มีข้อมูลที่เกี่ยวข้องกับไวรัสชนิดนี้เป็นจำนวนมาก ไม่ว่าจะเป็นจำนวนผู้ป่วยที่เพิ่มขึ้น จำนวนผู้เสียชีวิต จำนวนผู้ที่รักษาหาย เป็นต้น นับได้ว่าเป็นข้อมูลขนาดใหญ่และมีแนวโน้มที่จะเปลี่ยนแปลงตลอดเวลาในอนาคต

ทางผู้จัดทำจึงเล็งเห็นว่า หากสามารถปรับปรุงแพ็กเกจโดยดึงประสิทธิภาพออกมาใช้งานได้อย่างครบถ้วน จะส่งผลให้นำไปใช้งานง่ายยิ่งขึ้น ลดขั้นตอนการเรียกใช้งานจากหลาย ๆ แพ็กเกจ ซึ่งจะเป็นประโยชน์แก่ผู้ที่ต้องการใช้งานภาษา R เพื่อวิเคราะห์ข้อมูลที่มีความหลากหลายมากขึ้น และสามารถแสดงข้อมูลต่าง ๆ ให้กับคนทั่วไปสามารถเข้าถึงข้อมูลจำนวนมากนี้ ออกมาในรูปแบบ

ที่ทุกคนสามารถเข้าใจได้ง่าย ซึ่งจะเป็นประโยชน์อย่างยิ่งต่อการทำกิจกรรมต่าง ๆ ในการดำเนินชีวิตประจำวัน เช่น การลงทุนทำธุรกิจ การวางแผนชีวิตรับมือกับปัญหาที่จะเกิดขึ้น เป็นต้น

# 1.2 วัตถุประสงค์

1. เพื่อศึกษาวิธีการวิเคราะห์ข้อมูลโดยการใช้ภาษา R
2. เพื่อศึกษาความสามารถในการทำงาน ความสามารถในการรองรับขนาดของข้อมูล วิธีการจัดการข้อมูล และวิธีการใช้คำสั่งของแพ็กเกจ Dplyr
3. เพื่อทดลองวิเคราะห์ข้อมูลขนาดใหญ่ โดยใช้ภาษา R และใช้แพ็กเกจที่เลือกคือ Dplyr เพื่อทดสอบการทำงานในเชิงลึกต่อไปซึ่งข้อมูลที่ใช้ในการทดสอบคือข้อมูลเกี่ยวกับโรคติดเชื้อไวรัสโคโรนา 2019
4. เพื่อทดลองและเปรียบเทียบประสิทธิภาพการวิเคราะห์และแสดงผลข้อมูลระหว่างฟังก์ชันการวิเคราะห์ข้อมูลขนาดใหญ่ที่พัฒนาด้วยภาษา R แบบไม่ใช้แพ็กเกจ Dplyr และฟังก์ชันการวิเคราะห์ข้อมูลขนาดใหญ่ที่พัฒนาด้วยแพ็กเกจ Dplyr
5. เพื่อศึกษาแนวทางการนำข้อมูลจากการวิเคราะห์มาทำการแสดงเป็น Visualization ในรูป

แบบต่าง ๆ ตามความเหมาะสมของข้อมูล

# 1.3 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้รับความรู้ความเข้าใจเกี่ยวกับการใช้ภาษา R การใช้แพ็กเกจ Dplyr และการใช้แพ็กเกจอื่น ๆ ที่ช่วยเสริมประสิทธิภาพการทำงานของแพ็กเกจ Dplyr
2. สามารถนำข้อมูลขนาดใหญ่มาวิเคราะห์และแสดงผลในรูปแบบที่เข้าถึงได้ง่าย
3. ได้นำความรู้จากการเรียนภายในห้องเรียนมาประยุกต์ใช้งานเพื่อวิเคราะห์ข้อมูลจริง
4. สามารถเปรียบเทียบประสิทธิการทำงานของแพ็กเกจโดยการทำ Profiling ได้เพื่อเป็น

แนวทางในการเลือกแพ็กเกจอื่น ๆ มาใช้งาน

# **1.4** ข้อจำกัดของโครงงาน

1. แพ็กเกจไม่รองรับการวิเคราะห์ข้อมูลแบบเรียลไทม์ ข้อมูลที่มีการเปลี่ยนแปลงตลอดเวลา
2. ไม่สนับสนุนการเชื่อมต่อกับ Hadoop การประมวลผลใน Cluster
3. เนื่องจากใช้หลายแพ็กเกจร่วมกัน จึงทำให้ต้องมีการศึกษาเพิ่มเติมเกี่ยวกับแนวทางการ

ประยุกต์ใช้ร่วมกันของแพ็กเกจต่าง ๆ

# 1.5 ขอบเขตของโครงงาน

1. สามารถแสดงให้เห็นถึงความแตกต่างระหว่างการเรียกใช้แพ็กเกจ Dplyr และการใช้ภาษา R ในการวิเคราะห์ข้อมูลโดยตรง
2. สามารถนำแพ็กเกจ Dplyr ที่เลือกมาใช้ในการวิเคราะห์ข้อมูลร่วมกับแพ็กเกจอื่น เพื่อแสดงประสิทธิภาพในการทำงานร่วมกันออกมา
3. สามารถแสดงข้อมูลจากการวิเคราะห์ออกมาในรูปแบบที่สามารถเข้าใจได้ง่าย

# **1.6** แผนการดำเนินงานของวิชาโครงงาน

โดยมีระยะเวลาดำเนินงานทั้งหมด 2 เทอม

ตาราง 1.1 แผนการดำเนินงานระยะเวลา 2 เทอม

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **แผนการดำเนินงาน** | **ส.ค.** | **ก.ย.** | **ต.ค.** | **พ.ย.** | **ม.ค.** | **ก.พ.** | **มี.ค.** |
| **1. ค้นหาข้อมูลที่ต้องการนำมาใช้งาน** |  |  |  |  |  |  |  |
| 1.1 หาข้อมูลที่น่าสนใจ |  |  |  |  |  |  |  |
| 1.2 หาแหล่งที่มาของข้อมูล |  |  |  |  |  |  |  |
| **2. ศึกษาเครื่องมือที่ต้องการใช้งาน** |  |  |  |  |  |  |  |
| 2.1 ทดลองใช้งานแพ็กเกจ  Dplyr |  |  |  |  |  |  |  |
| 2.2 ทดลองใช้งาน Profiling  ใน RStudio |  |  |  |  |  |  |  |
| 2.3 ศึกษาการทำVisualization  และเครื่องมือที่ใช้ในการทำ |  |  |  |  |  |  |  |
| 2.4 ศึกษาทฤษฎีที่เกี่ยวข้อง |  |  |  |  |  |  |  |
| **3.ศึกษาและทำความเข้าใจข้อมูลที่ต้องการนำมาใช้งาน** |  |  |  |  |  |  |  |
| 3.1 เลือกข้อมูลที่ต้องการนำ  มาใช้งาน |  |  |  |  |  |  |  |
| 3.2 กำหนดช่วงเวลาของข้อ  มูลที่ต้องการนำมาใช้งาน |  |  |  |  |  |  |  |

ตาราง 1.1 แผนการดำเนินงานระยะเวลา 2 เทอม (ต่อ)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **แผนการดำเนินงาน** | **ส.ค.** | **ก.ย.** | **ต.ค.** | **พ.ย.** | **ม.ค.** | **ก.พ.** | **มี.ค.** |
| 3.3 จัดเตรียมข้อมูลให้พร้อมใช้งาน |  |  |  |  |  |  |  |
| 3.4 ออกแบบแบบจำลอง  แสดงความสัมพันธ์ของข้อมูลอธิบายโครงสร้างของฐานข้อมูล |  |  |  |  |  |  |  |
| 3.5 แสดงความสัมพันธ์ของ  ข้อมูล |  |  |  |  |  |  |  |
| **4. นำข้อมูลที่ได้จัดเก็บลงใน Local**  **Server** |  |  |  |  |  |  |  |
| 4.1 เลือก Local server ที่จะ  นำมาใช้งาน |  |  |  |  |  |  |  |
| 4.2 นำข้อมูลมาจัดเก็บใน  Local server |  |  |  |  |  |  |  |
| **5. สรุปผลจากการศึกษาข้อมูลที่**  **ต้องการใช้งานและหาวิธีแก้ไขปัญหา** |  |  |  |  |  |  |  |
| 5.1 แลกเปลี่ยนความรู้ที่ได้  จากการทดลอง |  |  |  |  |  |  |  |
| 5.2 นำข้อมูลใน Local server มาใช้งาน |  |  |  |  |  |  |  |
| 5.3 เลือกแพ็กเกจเสริมที่  ต้องการนำมาใช้งานร่วมกับแพ็กเกจหลัก |  |  |  |  |  |  |  |
| **6. พัฒนากลุ่มแพ็กเกจให้สามารถใช้งานได้หลากหลายยิ่งขึ้น** |  |  |  |  |  |  |  |
| **7. นำแพ็กเกจที่เลือกมาทดลองใช้งานกับข้อมูลขนาดเล็ก** |  |  |  |  |  |  |  |

ตาราง 1.1 แผนการดำเนินงานระยะเวลา 2 เทอม (ต่อ)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **แผนการดำเนินงาน** | **ส.ค.** | **ก.ย.** | **ต.ค.** | **พ.ย.** | **ม.ค.** | **ก.พ.** | **มี.ค.** |
| 7.1 ชุดข้อมูลผู้ติดเชื้อไวรัสโคโรนา 2019 |  |  |  |  |  |  |  |
| 7.2 ชุดข้อมูล GDP |  |  |  |  |  |  |  |
| 7.3 ชุดข้อมูลอุณหภูมิ |  |  |  |  |  |  |  |
| 7.4 ชุดข้อมูลอันดับสาธารณสุข |  |  |  |  |  |  |  |
| 7.5 นำผลการวิเคราะห์ข้อมูล  ขนาดเล็กมาใช้ในการทำ Visualization |  |  |  |  |  |  |  |
| **8. เปรียบเทียบประสิทธิภาพของแพ็กเกจ** |  |  |  |  |  |  |  |
| 8.1 เปรียบเทียบประสิทธิภาพ  ความแม่นยำ (ใช้การคำนวณทางคณิตศาสตร์) |  |  |  |  |  |  |  |
| 8.2 เปรียบเทียบประสิทธิภาพ  การทำงาน (ใช้ประกอบการทำ Profiling) |  |  |  |  |  |  |  |
| **9. สรุปผลจากการเปรียบเทียบ**  **ประสิทธิภาพแพ็กเกจ** |  |  |  |  |  |  |  |
| **10. สรุปผลจากการวิเคราะห์ข้อมูล**  **ขนาดเล็ก** |  |  |  |  |  |  |  |
| **11. สรุปปัญหาและข้อเสนอแนะ**  **เพื่อนำไปใช้พัฒนาต่อ** |  |  |  |  |  |  |  |
| **12. นำแพ็กเกจที่เลือกมาทดลองใช้งานกับข้อมูลขนาดใหญ่** |  |  |  |  |  |  |  |
| **13. สรุปผลจากการวิเคราะห์ข้อมูล**  **ขนาดใหญ่** |  |  |  |  |  |  |  |
| **14. สรุปปัญหาและข้อเสนอแนะ**  **เพื่อนำไปใช้ในการพัฒนาต่อไป** |  |  |  |  |  |  |  |

ตาราง 1.1 แผนการดำเนินงานระยะเวลา 2 เทอม (ต่อ)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **แผนการดำเนินงาน** | **ส.ค.** | **ก.ย.** | **ต.ค.** | **พ.ย.** | **ม.ค.** | **ก.พ.** | **มี.ค.** |
| **15. จัดทำรูปเล่มรายงาน** |  |  |  |  |  |  |  |
| 15.1 บทที่ 1 |  |  |  |  |  |  |  |
| 15.2 บทที่ 2 |  |  |  |  |  |  |  |
| 15.3 บทที่ 3 |  |  |  |  |  |  |  |
| 15.4 บทที่ 4 |  |  |  |  |  |  |  |
| 15.5 บทที่ 5 |  |  |  |  |  |  |  |