Term Project Proposal: US Accidents

จัดทำโดยกลุ่ม ‘ต่อไป’

63070501017 นายณธรรศ สุวรรณพงษ์

63070501068 นายสราวุฒิ นุชนารถ

63070501069 นายสัณหณัฐ พรมจรรย์

63070501077 นายสุทธิพงษ์ ปัญญาดี

โครงงานชิ้นนี้เป็นส่วนหนึ่งของวิชา Data Models

ภาคเรียนที่ 2 ปีการศึกษา 2564

คณะวิศวกรรมศาสตร์ สาขาวิศวกรรมคอมพิวเตอร์

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

Abstract

- ใช้วิเคราะห์ข้อมูลเกี่ยวกับข้อมูลอุบัติเหตุทางรถยนต์ทั่วประเทศ ซึ่งครอบคลุม 49 รัฐของสหรัฐอเมริกา ปี 2016 - 2021 ใช้ API หลายตัวที่ให้ข้อมูลเหตุการณ์การจราจร โดยจะนำข้อมูลมาใช้ วิเคราะห์ข้อมูลของการเกิดอุบัติเหตุด้วยปัจจัยต่างๆ ด้วยข้อมูลสภาพแวดล้อม ความรุนแรงของอุบัติเหตุ ตำแหน่งละติจูด ลองจิจูดที่เกิดอุบัติเหตุ ช่วงถนนที่ได้รับผลกระทบ ช่วงเวลาที่เกิดขึ้นและสิ้นสุด เกิดที่เมืองไหน รัฐไหน อุณหภูมิ สภาพการมองเห็น ความชื้น ความดัน ความแรงลม ทิศทางลม สภาพอากาศ สถานที่ที่เกิดเหตุมีทางแยกหรือตรอกซอกซอยไหม บริเวณนั้นมีสนามบินอยู่ รางรถไฟ สถานีไหม มีถนนทางตัน ป้ายสัญญาณไฟจราจร ป้ายให้หยุด บริเวณดังกล่าวหรือไม่ มาทำการ Visualization จากนั้นวิเคราะห์ข้อมูลหาส่วนที่เป็นปัจจัยการเกิดอุบัติเหตุต่างๆใน USA พร้อมทั้งหาเหตุผลประกอบ จะทำให้เราทราบว่าปัจจัยต่างๆ สภาพสิ่งแวดล้อม สภาพอากาศ ในช่วงเวลาไหนที่ทำให้เกิดอุบัติเหตุ และช่วยในการตัดสินใจลงทุน ทั้งด้านสินค้า ประกัน หรือการแก้ไขสภาพแวดล้อมที่เกิดขึ้น ทำให้ลดความเสี่ยงที่จะเกิดอุบัติเหตุจากปัจจัยต่างๆได้

Introduction

- อัตราการเกิดอุบัติเหตุบนถนนยังมีอยู่มากในปัจจุบัน จึงเป้นอีกเรื่องหนึ่งที่น่าสนใจ กลุ่มของเราจึงต้องการที่จะทราบว่า ในช่วงเวลาไหนของวัน ถนนแบบไหน ความร้ายแรงของอุบัติเหตุ มาวิเคราะห์แนวโน้มของการเกิดอุบัติเหตุ เพื่อหาแนวทาง หรือวิธีการรับมือที่เกิดขึ้น ลดจำนวนอุบัติเหตุที่จะเกิดขึ้น หรือกระทั่งลดระดับความรุนแรง กลุ่มของเราจึงได้ทำการหยิบยก Dataset ที่ได้บันทึกรายละเอียดการเกิดอุบัติเหตุทั้งข้อมูล Quantitative ประกอบด้วย เวลาที่เกิดอุบัติเหตุ ช่วงเวลาที่เกิดอุบัติเหตุ ความยาวเส้นกั้นถนน อุณหภูมิ ความเร็วลม ความชื้น ความดันอากาศ ระยะการมองเห็น และข้อมูล Categorical ประกอบด้วย ระดับความรุนแรง ชื่อถนน เมือง ขนาดถนน ทิศทางลม ซึ่งกลุ่มของเรามองว่ามีรายละเอียดที่เพียงพอจะมาเป็นข้อมูลที่จะนำมา Visualization เพื่อดูการกระจายตัว วิเคราะห์ผล และสร้างเป็นโมเดลขึ้นมา เพื่อตอบโจทย์ปัญหาของเรา ช่วยในการตัดสินใจ และจัดการเหตุการณ์ในแต่ละช่วงเวลา ลดความเสี่ยงที่จะเกิดอุบัติเหตุ และลดระดับความรายแรง

Data Source Description

- โดยข้อมูลที่เรานำมาทำการวิเคราะห์เพื่อแก้ไขปัญหานี้ก็คือ ข้อมูลอุบัติเหตุในสหรัฐอเมริกา โดยข้อมูลอุบัติเหตุถูกรวบรวมตั้งแต่เดือนกุมภาพันธ์ 2559 ถึงธันวาคม 2564 ใน Dataset ของที่เรานำมาศึกษา มีทั้งข้อมูลในส่วนของปัจจัยการเกิดอุบัติเหตุต่างๆใน USA ซึ่งใน Data Set มีข้อมูลดังนี้ ข้อมูล Quantitative ประกอบด้วย เวลาที่เกิดอุบัติเหตุ ช่วงเวลาที่เกิดอุบัติเหตุ ความยาวขอบเขตถนน ละติจูด ลองจิจูด อุณหภูมิ ลมหนาว ความเร็วลม ความชื้น ความดันอากาศ ปริมาณน้ำฝน ระยะการมองเห็น ทัศนวิสัย และข้อมูล Categorical ประกอบด้วย ไอดีระบุ ระดับความรุนแรง คำอธิบาย สภาวะอากาศ เลขที่ถนน เลนถนน ชื่อถนน เมือง เขต สถานะ ประเทศ รหัสไปรษณีย์ ทิศทางลม รางรถไฟ สถานี สนามบิน การชน ทางแยก ทางข้าม ป้ายให้ทาง ทางตัน วงเวียน วงเลี้ยว ป้ายจราจร ไฟจราจร สิ่งอำนวยความสะดวก พระอาทิตย์ขึ้น-ตก ลูกระนาด

โดยข้อมูลData อ้างอิงนำมาจาก

- https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

Problem Statement

- เพื่อนำไปใช้ในการศึกษาการลงทุน หรือการก่อสร้างสำหรับผู้ที่ริเริ่มมีความสนใจเกี่ยวกับความปลอดภัยในการขับรถยนต์ของ USA โดย Dataset ที่เรานำมาศึกษาจะบ่งบอกถึงข้อมูลต่างๆที่มีปัจจัยเกี่ยวกับการเกิดอุบัติเหตุ สถานที่ เวลา สภาพอากาศ ซึ่งสามารถนำไปเป็นความรู้เพื่อประกอบการตัดสินใจในการแก้ไขจัดการพื้นที่สัญจร และสามารถบ่งบอกปัญหาสภาพแวดล้อม เราจึงจำเป็นต้องนำข้อมูลเหล่านี้มาวิเคราะห์ข้อมูลการเกิดอุบัติเหตุต่างๆ เพื่อนำมาใช้ในการวางแผนธุรกิจได้

Proposed Analytic Technique

- นำข้อมูลใน Dataset ที่เรานำมาศึกษามาทำการ Visualization ว่ามีการกระจายตัวของข้อมูล และ ข้อมูลมีความสัมพันธ์กันไหม โดยการทำ Visualization เริ่มจากนำข้อมูลที่นำมาศึกษามาจัดระเบียบข้อมูลและนำมาพล็อตกราฟเพื่อดูองค์ประกอบโดยรวมของข้อมูลใน Dataset จากนั้นเราจะนำข้อมูลที่ได้มาหาแนวโน้มของการเกิดอุบัติเหตุ เพื่อหาแนวทาง หรือวิธีการรับมือที่เกิดขึ้น ลดจำนวนอุบัติเหตุที่จะเกิดขึ้น หรือกระทั่งลดระดับความรุนแรง โดยจะมีเทคนิคดังนี้

1.) การจัดการข้อมูล

ใน R-studio มี library การจัดการข้อมูลคือ dplyr มาจัดระเบียบข้อมูลให้สามารถใช้งาน และตรวจสอบได้ง่ายขึ้น

- mutate เพื่อจัดข้อมูลให้เป็นช่วงตามสิ่งที่เราต้องการ
- Select การเลือกใช้ข้อมูลที่จำเป็นเพื่อหได้ข้อมูลที่เราต้องการมาใช้งาน
- Group by รวมข้อมูลเข้าด้วยกันเพื่อนำมาใช้คำนวณหาค่า count sum และอื่นๆ
- Arrange ใช้เพื่อให้เราแสดงข้อมูลที่มีการจัดเรียงตามที่เราต้องการ
- mutate if และ summary เพื่อตรวจสอบภาพรวมของข้อมูล จำนวน ค่าต่ำสุด ค่าสูงสุด และค่ากลาง
- summarize การสรุปข้อมูล
- Unite และ Separate การแยก และ รวมข้อมูล เข้าด้วยกัน
- Gather และ Spread เพื่อแยกแถวย่อย หรือ รวมแถวย่อยนั้นๆ
- Filter เพื่อหาข้อมูลในช่วงที่ต้องการ

2.) การพล็อตกราฟ ggplot2

ใน r studio จะมี library การพล็อตกราฟหลายอย่างด้วยคำสั่งเบื้องต้นคือ ggplot ต่อด้วยฟังชันที่ต้องการพล็อต ในที่นี้เราคาดว่าจะใช้อยู่ 2 ส่วนหลักๆคือ geom_bar เพื่อดูข้อมูลในส่วนต่างๆ และ geom_histrogram เพื่อดูการกระจายตัวของข้อมูล และสามารถหาความสัมพันธ์ของข้อมูลด้วย datarealationship วัดด้วยค่า chi-squared

3.) probability distribution

เป็นการนำข้อมูลที่มีอยู่จนถึงปัจจุบันมาคำนวณความน่าจะเป็นของอนาคตเพื่อเป็นตัวช่วยการรองรับการตัดสินใจได้โดยใช้ PDF CDF quartile มาเป็นเครื่องมือช่วย