BIOST 557 A Winter 2022: Applied Statistics & Experimental Design

US Road Accident Analysis

Overview

We will be statistically analysing the US road accidents dataset which records the road accidents across the US and includes various factors describing and accompanying the accident

This analysis can be used to make policy recommendations / business decisions / insights:

1. Public work orgs can improve road conditions and manage traffic rules accordingly
2. Companies like Uber, Lyft can re-route cabs and dynamically update pricing based on changing traffic patterns
3. Safety recommendations:
    a. Car drivers are provided with better live updates, and provide insight about locations where they can be more vigilant (accident hotspots)
    b. Help pedestrians feel safer while crossing roads / walking in certain regions

Question 1: Does the distance of the road affected by traffic depend on the severity of the accident?

Approach:

Hypothesis testing where,

H0: There is no difference between the means of the - distance of the road affected by the severity of the accident

H1: There is a difference between the means of the - distance of the road affected by the severity of the accident

Testing strategy - Perform two sample t-test where we divide the severity of accident into 2 categories of severe and not severe accidents, and check for statistical significance between between the groups. Results would indicate whether the distance of the road affected is dependent on severity of the accident

Question 2: What is the effect of rainfall and wind speed on the severity of accident?

Approach:

Perform Exploratory Data Analysis (EDA) to evaluate the relations of these natural factors on the severity of accident

Current challenge:

The dataset is huge and we’re figuring out an efficient way to parse it

(maybe using Spark)

Question 3: Comparing peak hours in a day during which accidents occur - office leaving hours vs. midnight

Approach:

A/B testing to compare accidents during office leaving hours and sunset.

V1: Road accidents are higher during midnight as compared to office leaving hours.

V2: Road accidents are higher during office leaving hours compared to midnight.

Question 4: What are some of the most accident hotspot locations?

Approach:

Using the one-sided ANOVA test, we want to see if there is a significant difference in the most popular accident hotspot locations across cities

For instance, if the mean of the column for the accident location is a stop sign for Seattle, we want to check if that is the same for say, Redmond

To reiterate, the categorical variables are popular cities in King County

Since we have granular data, we can compare means across popular cities like Seattle, Redmond, Bellevue and Kirkland

Question 5: Multiple linear regression for all variables

Approach:

1) Validate assumptions of multiple linear regression
- Multicollinearity
- Constant variance
- No autocorrelation
- Normality of errors
- Exogeneity

* Feature selection to identify severity of accidents near road signs.

* F test