BIOST 557 A Winter 2023: Applied Statistics & Experimental Design

King County Road Accident Analysis

Aditi Kharkwal, Akshit Miglani, Ameya Bhamare, Danish Nadeem, Tejal Kolte

Project Motivation

We will analyze factors (such as rainfall, time of day, and locations of interest) related to the severity and distance of the road affected by accidents in King County, Washington

This analysis can be used for policy recommendations, business decisions, insights:

- Public work organizations can improve road conditions and manage traffic rules accordingly
- Companies like Uber and Lyft can reroute cabs and dynamically update pricing based on changing traffic patterns
- Safety recommendations
    - Drivers are provided with better live updates and insight about accident hotspots
    - Help pedestrians feel safer when crossing roads and walking in certain areas

Data Overview

- *US Accidents (2016 - 2021)* data set (published to Kaggle by a scientist at Lyft)
  - Covers 49 states
  - Spans February 2016 - December 2021
  - Aggregated from multiple data providers and APIs
    - US and State Department of Transportation
    - Law Enforcement
    - Traffic Cameras, Traffic Sensors

- We focused on King County, Washington
  - Data for 15,903 unique accidents

Objectives and Hypotheses

We aim to gain a better understanding of factors relating to road accidents and investigate how different variables can impact the severity and distance of the road impacted by an accident. Specifically, we will explore the following questions:

1. What is the relationship between the distance of the road affected by the accident and severity of the accident?
2. What is the effect of rainfall and windspeed on the overall severity of an accident?
3. Is there a relationship between the severity of an accident and the time of day during which it occurs? Is there a significant difference between the severity of accidents that occur around midnight and those that occur around evening transit hours?
4. Do locations of interest, such as stop signs, roundabouts, bumps, and crossings have a significant impact on the severity of accidents?
5. What parameters are the most important in predicting the length of the road extent affected by the accident?

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

Variables:

- Distance(mi): The distance of the road affected in miles by the accident
- Severity: Number between 1 and 4 where 1 indicates the least impact on traffic (in terms of time) and 4 indicates a significant impact on traffic

Significance Level = 0.05

Null Hypothesis: There is no difference between the average distances of the road affected by the severity of the accident

Alternate Hypothesis: There is a difference between the average distances of the road affected by the severity of the accident

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

Approach:

- We divided our dataset into two parts based on severity. One group had severity values 1 and 2 and the other had 3 and 4
- After this we conducted two sample t-test with unequal variance. **Group 1 (not severe)** had variance 1.1 and **Group 2 (severe)** had variance 1.3

Results and Conclusion:

- T-statistic value = -5.297 & P-value = 1.226e-07
- Since p-value is less than 0.05 we **reject the null hypothesis**
- For purpose of validation we also conducted one way ANOVA test and got the following results test-statistic = 62.825 and P-value = 2.252e-40
- Since P-value is less than 0.05 we **reject the null hypothesis**

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

Approach: Perform Exploratory Data Analysis (EDA) and ANOVA test to evaluate the relations of these natural factors on the severity of accidents

Variables:

- Wind_Speed(mph): Shows wind speed (in miles per hour)
- Precipitation(in): Shows precipitation amount in inches, if there is any
- Severity: Number between 1 and 4 where 1 indicates the least impact on traffic (in terms of time) and 4 indicates a significant impact on traffic

Null Hypothesis: There is no difference between the means of the precipitation/wind speed affected by the severity of the accident

Alternate Hypothesis: There is a difference between the means of the precipitation/wind speed affected by the severity of the accident (Significance Level = 0.05)

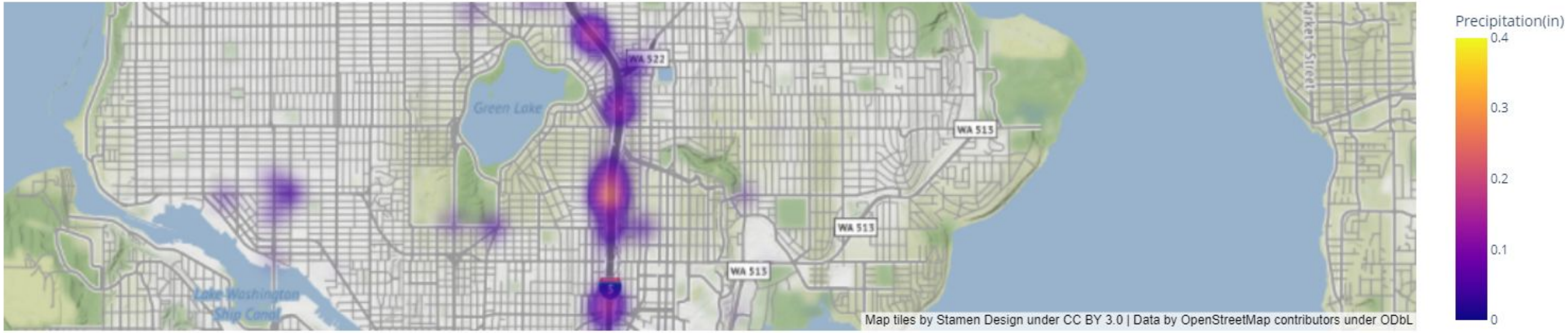
Test performed: One way ANOVA test

Result: The p-value corresponding to the F-value of 7.8 at a significance level of 0.05 is less than 0.05.

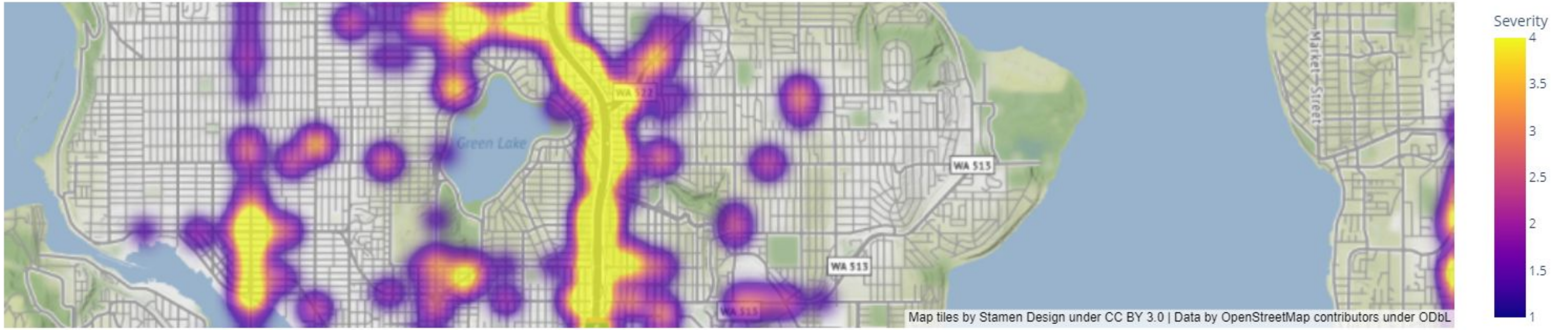
Conclusion: We found statistically significant evidence to reject the null hypothesis

EDA

There appear to be more accidents in locations where there is more precipitation (first map) (there seems to be a correlation from the outlook)


Green Lake
WA 522
WA 513
WA 513
WA 513
Lake Washington Ship Canal
Market Street
Map tiles by Stamen Design under CC BY 3.0 | Data by OpenStreetMap contributors under ODbL
Precipitation(in)
0.4
0.3
0.2
0.1
0

Precipitation Heat Map around Green Lake


Green Lake
WA 522
WA 513
WA 513
WA 513
Market Street
Map tiles by Stamen Design under CC BY 3.0 | Data by OpenStreetMap contributors under ODbL
Severity
4
3.5
3
2.5
2
1.5
1

Severity Heat Map

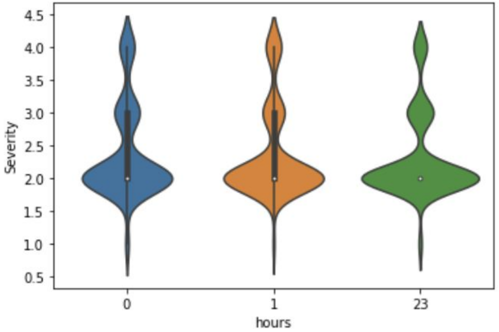
Question 3: Is there a relationship between the severity of an accident and the time of day during which it occurs? Is there a significant difference between the severity of accidents that occur around midnight and those that occur around evening transit hours?

Main Idea: The purpose of this analysis is to determine when more severe accidents occur; specifically comparing midnight hours and evening transit hours. We will perform a one sample t-test to compare the mean of one group against the specified mean generated from a population. Our hypothesis is as follows:

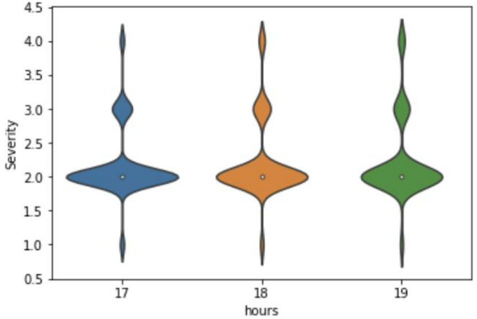
Null Hypothesis: There is no difference between the average severity between midnight accidents and evening transit hour accidents

Alternate Hypothesis: There is a difference between the average severity between midnight accidents and evening transit hour accidents

When evaluating the results of the one-sample t-test, we will compare the p-value of the sample under the null hypothesis. Considering the 5% level of significance, if the p-value would be less than 0.05, we would reject the null hypothesis.



Distribution of severity of accidents during midnight hours (11pm - 1am)



Distribution of severity of accidents during office transit hours (5pm - 7pm)

Question 3: Is there a relationship between the severity of an accident and the time of day during which it occurs? Is there a significant difference between the severity of accidents that occur around midnight and those that occur around evening transit hours?

Approach: Created 2 categorical variables *Office_transit* and *Midnight* which had accident severity records from accidents occurred between 5pm-7pm and 11pm-1am respectively. With an array of severity of accidents from both the categorical variables we performed a one sample t-test to compare the mean of one group against the specified mean generated from a population.

Result and conclusion:

- T-statistic value = 6.422 and P-value = 1.507e-10
- We reject the null hypothesis since p-value is significantly less than 0.05 which is the significance level we have assumed for our testing.
- Interpretation: There is a statistically significant difference in the average of severity of accidents which occur at midnight vs accidents which occur during office transit hours.
- The Data visualization was not enough to conclude or find out the difference between the severity of accidents between the 2 groups however t test was a suitable test to arrive at this conclusion.

Question 4: Do locations of interest, such as stop signs, roundabouts, bumps, and crossings have a significant impact on the severity of accidents?

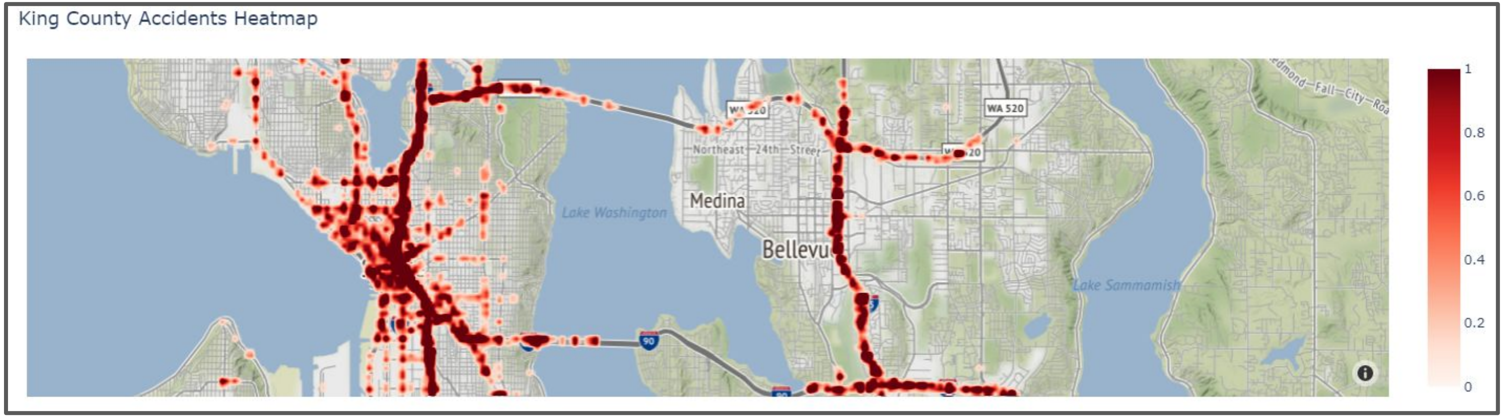
Null hypothesis: Mean severity of accidents at locations of interest (stop signs, roundabouts, bumps, etc) is equal to the mean severity of accidents at other locations.

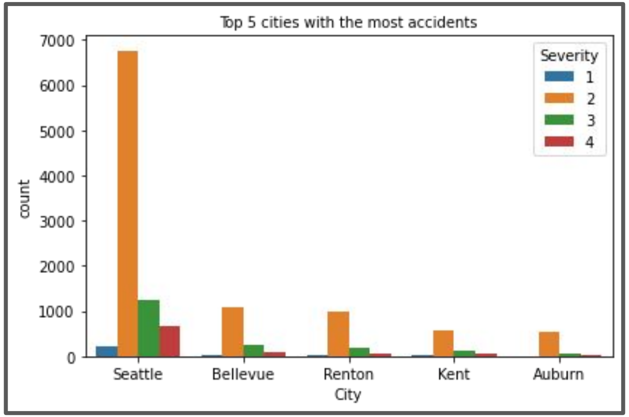
Test performed: Independent sample unequal variances t-test.

Result: The p-value corresponding to the t-statistic of **-7.96** at a significance level of **0.05** is **1.74e-15**.

Conclusion: We found statistically significant evidence to reject the null hypothesis. This suggests the alternative hypothesis that the mean severity of accidents between the two groups are unequal.

EDA: Accident Hotspot Locations





Question 5: What parameters are the most important in predicting the length of the road extent affected by the accident?

Approach: Created 3 linear regression models

- ***Full Model (Model 1):*** **All parameters of interest**

  E[ln(Distance(mi))] ~ Severity + Side + Temperature + Wind_Chill + Humidity + Pressure + Visibility + Wind_Speed + Amenity + Bump + Crossing + Give_Way + Junction + No_Exit + Railway + Roundabout + Station + Stop + Traffic_Calming + Traffic_Signal + Turning_Loop + Civil_Twilight + Nautical_Twilight + Astronomical_Twilight + Sunrise_Sunset + Weekday + Hours

- ***Reduced Model (Model 2):*** **Parameters we hypothesized to be the most important**

  E[ln(Distance(mi))] ~ Severity + Side + Temperature + Wind_Chill + Humidity + Pressure + Amenity + Crossing + No_Exit + Roundabout + Station + Stop + Traffic_Signal + Astronomical_Twilight + Weekday + Hours

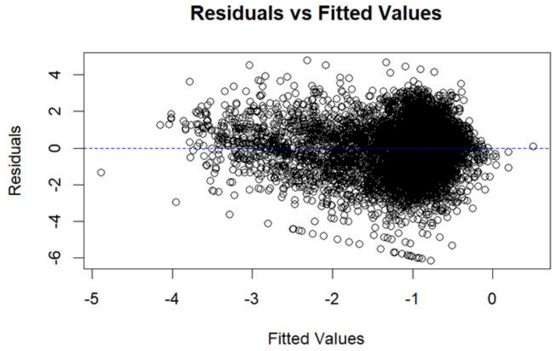
- ***Further Reduced Model (Model 3):*** **No road signs or structures**

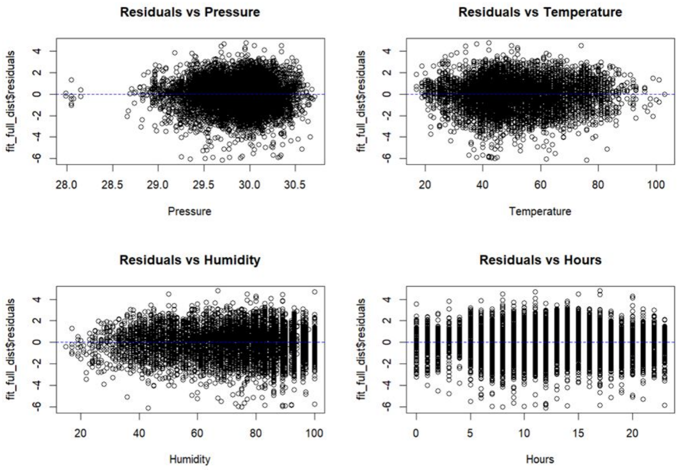
  E[ln(Distance(mi))] ~ Severity + Temperature + Wind_Chill + Humidity + Pressure + Astronomical_Twilight + Weekday + Hours

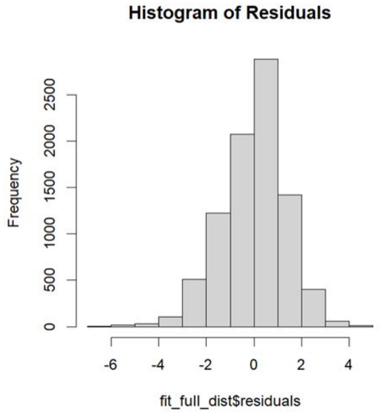
Question 5: What parameters are the most important in predicting the length of the road extent affected by the accident?

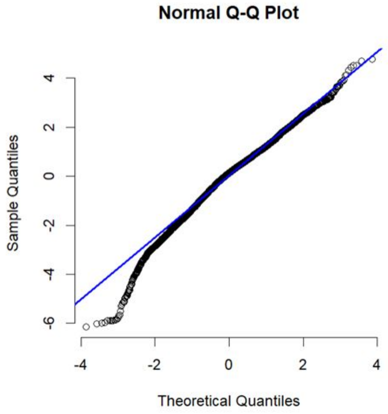
Approach: Checked assumptions for inference

- Independence - *met*
- Linearity - *approximately met*
- Constant Variance - *not met*
  - Used Robust Standard Errors
- Normality - *approximately met*









Question 5: What parameters are the most important in predicting the length of the road extent affected by the accident?

Approach: For each model, we:

- Calculated the estimated coefficient for each predictor
- Implemented a t-test using robust standard errors and observed p-values (using a significance level of 0.05) to determine which predictor variables are significant

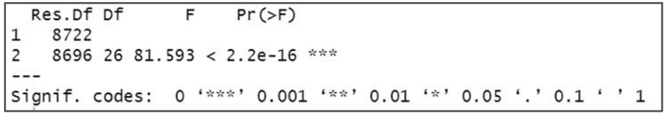
To assess and compare these models, we:

- Used a Global F-test on Model 1 to test the null hypothesis that the coefficients on all predictors of interest equal 0
- Used a multiple F-test to compare Model 1 and Model 2 and test the null hypothesis that the additional predictors (Visibility, Wind_Speed, Bump, Give_Way, Junction, Railway, Traffic_Calming, Turning_Loop, Civil_Twilight, Nautical_Twilight, Sunrise_Sunset) equal 0
- Used a multiple F-test to compare Model 2 and Model 3 and and test the null hypothesis that the additional predictors (Amenity, Crossing, No_Exit, Roundabout, Station, Stop, Traffic_Signal) equal 0
- Computed adjusted R-squared values for each model
- Used adjusted R-squared values and results of F-tests to determine most appropriate model

Question 5: What parameters are the most important in predicting the length of the road extent affected by the accident?

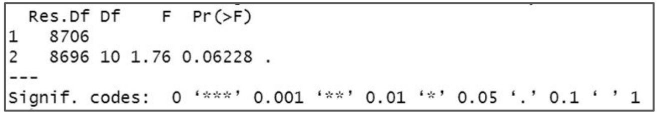
Results

- **Global F-test:** reject the null hypothesis. Thus, we have evidence to conclude that that the coefficients on all predictors of interest do not equal 0.



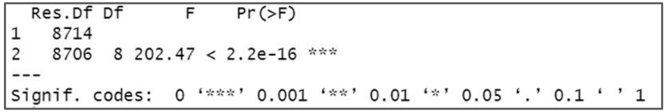
```
  Res.Df Df      F    Pr(>F)    
1   8722                        
2   8696 26 81.593 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Multiple F-test: Comparing Model 1 and Model 2:** fail to reject the null hypothesis. Thus, we have evidence to conclude that the coefficients on the following predictors equal 0: Visibility, Wind_Speed, Bump, Give_Way, Junction, Railway, Traffic_Calming, Turning_Loop, Civil_Twilight, Nautical_Twilight, Sunrise_Sunset



```
  Res.Df Df    F  Pr(>F)  
1   8706                  
2   8696 10 1.76 0.06228 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Multiple F-test: Comparing Model 2 and Model 3:** reject the null hypothesis. Thus, we have evidence to conclude that the coefficients on the following predictors do not all equal 0: Amenity, Crossing, No_Exit, Roundabout, Station, Stop, Traffic_Signal



```
  Res.Df Df      F    Pr(>F)    
1   8714                        
2   8706  8 202.47 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Adjusted R-Squared**

Model	Adjusted R-Squared
Model 1	0.1863
Model 2	0.1856
Model 3	0.03513

Question 5: What parameters are the most important in predicting the length of the road extent affected by the accident?

Conclusion: Reduced Model (Model 2) is the most appropriate model

- Most important parameters: Severity, Side, Temperature, Wind_Chill, Humidity, Pressure, Amenity, Crossing, No_Exit, Roundabout, Station, Stop, Traffic_Signal, Astronomical_Twilight, Weekday, Hours

Estimated Model:

E[ln(Distance(mi))] = -0.086 x Severity + 0.614 x Side + -0.042 x Temperature + 0.034 x Wind_Chill + -0.005 x Humidity + -0.532 x Pressure + -0.634 x Amenity + -0.675 x Crossing + -0.859 x No_Exit + 1.014 x Roundabout + -0.332 x Station + -0.504 x Stop + -0.518 x Traffic_Signal + 0.075 x Astronomical_Twilight + 0.059 x Weekday + 0.021 x Hours

Questions?