Statistical Analysis Plan: King County Road Accident Analysis

I. Background and Rationale of Study

This statistical analysis plan will analyze factors related to the severity and distance of the road affected by accidents in King County, Washington. We will employ the *US Accidents (2016 - 2021)* data set, and explore variables such as rainfall, windspeed, visibility, time of day, and locations of interest to determine their effect on road accidents.

II. Objectives and Hypotheses

With this project, 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, windspeed, and visibility on the overall severity of the 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?

III. Study Overview

Our analysis is based on the *US Accidents (2016 - 2021)* data set, originally published to Kaggle by Sobhan Moosavi, a scientist at Lyft. This data set covers 49 states of the United States, and spans February 2016 to December 2021. The data has been aggregated from multiple data providers and APIs that provide streaming traffic event data, and are sourced from organizations such as the US and state departments of transportation, law enforcement, traffic cameras, and traffic sensors.

For the purposes of our analysis, we will focus on King County, Washington. This subset of data contains 15,903 rows and spans June 2016 to December 2021. See Table 1 in the appendix for a list of all variables and definitions.

IV. Statistical Analysis

In this section, we will explain each of the aforementioned questions in greater detail.

Question 1: What is the relationship between the distance of the road affected by the accident and severity of the accident?

In this part of our analysis, we will analyze the relationship between the *Distance(mi)* variable and the *Severity* variable. *Distance(mi)* is the distance of the road affected in miles by the accident and *Severity* is a number between 1 and 4 where 1 indicates the least impact on traffic and 4 indicates a significant impact on traffic.

The core purpose of this analysis is to help Lyft modify their pricing algorithm accordingly by identifying areas where the severity of the accident is higher if its statistically related to distance of the road affected. Our hypothesis is as follows:

- H_0: There is no difference between the average distances affected by the severity of the accident
- H_1: There is a difference between the average distances affected by the severity of the accident

Initially, we will divide our dataset in two groups: severe and not severe accidents, where severe accidents are defined as having a severity value greater than 2, and not severe accidents have a value less than or equal to 2. Once this is done, we will analyze the variance of the distance column in our groups. If the variance is the same, we will apply the two sample t-test with equal variance. If the variance is unequal, we will employ the two sample t-test with unequal variance. We will be using a significance level of 0.05 and try to control for power through manipulating the sample size.

Finally, for cross validation purposes we will also perform an ANOVA test. In this part, we will divide the data in four groups based on the value of severity. This test will help verify our results from our initial two-sample t-test.

Question 2: What is the effect of rainfall, windspeed, and visibility on the overall severity of the accident?

For this question, we will be focusing on the following variables: *Wind_Speed(mph)*, *Precipitation(in)*, and *Severity*.

Prior to beginning our analysis, we examined the variables of interest and checked for any missing values. We observed some missing values in wind speed and precipitation, but chose to remove these values after looking at the overall number of rows in the dataset.

In order to visualize the relationship between wind speed, precipitation, and the severity of the accident, we will create a violin plot. To observe the actual values, we will also compute the mean, standard deviation, and variance across different severity levels to see if there is a significant difference.

In this case, we are working with one categorical variable (severity) and other numerical variables (precipitation and wind speed). We will use the ANOVA test to check for differences among the means of the different severity levels by examining the amount of variation within each sample, relative to the amount of variation between the samples. In order to use the ANOVA test, we will validate the following assumptions:

- The responses for each factor level have a normal population distribution
- These distributions have the same variance
- The data are independent

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?

The purpose of this analysis is to determine when more severe accidents occur; specifically comparing midnight hours and evening transit hours. To evaluate this, we defined two categories: midnight hours, and evening transit hours. Midnight hours are defined as any time between 11 pm and 1 am, and evening transit hours are defined as any time between 5 pm and 7 pm. The hours were derived from the table field *Start_Time* and *End_Time* of the accident.

We created two corresponding categorical variables: *Midnight* (midnight hours) and *Office* (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:

- H_0: There is no difference between the average severity between midnight accidents and evening transit hour accidents
- H_1: There is a difference between the average severity between midnight accidents and evening transit hour accidents

We will thoroughly validate each of the assumptions below to ensure we can successfully employ the one-sample t-test:

1. The data is collected from a representative, randomly selected portion of the total population.
2. Data should follow a continuous or discrete scale of measurement.
3. Means should follow the normal distribution, as well as the population.
4. Independence of the observations. Each subject should belong to only one group.

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.

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

For this question, we are interested in understanding accident hotspot locations within King County. We will be working with the following 12 variables pertaining to locations of interest: *Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Turning_Loop*.

We will use the independent samples t-test. It checks whether the unknown population means of a given pair of groups are equal. It allows one to test the null hypothesis that the means of two groups are equal. In this case, the two groups are events where accidents occur.

We will validate each of the following assumptions, required for the independent samples t-test:

1. Independence: While dependent samples are paired measurements for one set of items, independent samples are measurements made on two different sets of items
    a. If the values in one sample affect the values in the other sample, then the samples are dependent
    b. If the values in one sample reveal no information about those of the other sample, then the samples are independent
2. Normality: The sample mean is normally distributed for large sample sizes, regardless of the distribution from which we sample.
    a. The mean and standard deviation (both finite) of the sampling distribution of the sample mean (Where *X bar* represents the sampling distribution of the sample mean of size n each, and *mu X bar* and *sigma X bar* are the mean and standard deviation of the population respectively). The distribution of the sample tends towards the normal distribution as the sample size increases.

3. Homogeneity: An *F-test* is used to test whether two population variances are equal. The null and alternative hypotheses for the test are:
    a. H_0: σ12 = σ22 (the population variances are equal)
    b. H_1: σ12 ≠ σ22 (the population variances are not equal)

When running the test, we will employ the independent samples t-test, and evaluate the following hypotheses:

- H_0: $\mu_1 = \mu_2$ (population mean of dataset 1 is equal to dataset 2)
- H_1: $\mu_1 \neq \mu_2$ (population mean of dataset 1 is different from dataset 2)

If the p-value associated with the t-test statistic is lesser than the set significance level (we have set alpha = 0.05), we reject H_0 which suggests that the mean severity of accidents of the two groups is equal.

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

To answer this question, we will create three linear regression models (full model, reduced model, further reduced model) to evaluate which parameters are important in predicting the length of the road extent affected by the accident. The variable *Distance (mi)* will be the response variable, as it measures the length of the road affected by the accident.

Model 1 is the full model and contains all predictors of interest:

log(Distance(mi)) ~ Temperature + Humidity + Pressure + Visibility + Wind_Speed + Amenity + Bump + Crossing + Give_Way + Junction + No_Exit + Railway + Roundabout + Station + Stop + Traffic_Calming + Traffic_Signal + Turning_Loop + Sunrise_Sunset + Hours

Model 2 is a reduced model and contains a subset of predictors from Model 1 that were found to be significant:

log(Distance(mi)) ~ Temperature + Humidity + Pressure + Visibility + Wind_Speed + Amenity Crossing + Give_Way + Junction + No_Exit + Railway + Roundabout + Station + Stop + Traffic_Signal + Sunrise_Sunset + Hours

Model 3 is further reduced from Model 2 and does not contain any road signs or structures:

log(Distance(mi)) ~ Temperature + Humidity + Pressure + Visibility + Wind_Speed + Sunrise_Sunset + Hours

For each model, we will:

- Validate the following assumptions: linearity, independence, normality, constant variance
    - Transform the dependent variable (distance) to natural log(distance) in order to ensure these assumptions are met
- Calculate the estimated coefficients for each predictor of interest
- Implement a t-test and observe p-values (using a significance level of 0.05) to determine which predictor variables are significant and which predictor variables are not significant

We will compare all three models using a multiple F-test and a global F-test to determine the most appropriate model.

V. Appendix

Table 1: Description of all Variables in Dataset

#	Attribute	Description	Nullable
1	ID	This is a unique identifier of the accident record.	No
2	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).	No
3	Start_Time	Shows start time of the accident in local time zone.	No
4	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	No
5	Start_Lat	Shows latitude in GPS coordinate of the start point.	No
6	Start_Lng	Shows longitude in GPS coordinate of the start point.	No
7	End_Lat	Shows latitude in GPS coordinate of the end point.	Yes
8	End_Lng	Shows longitude in GPS coordinate of the end point.	Yes
9	Distance(mi)	The length of the road extent affected by the accident.	No
10	Description	Shows natural language description of the accident.	No
11	Number	Shows the street number in address field.	Yes

12	Street	Shows the street name in address field.	Yes
13	Side	Shows the relative side of the street (Right/Left) in address field.	Yes
14	City	Shows the city in address field.	Yes
15	County	Shows the county in address field.	Yes
16	State	Shows the state in address field.	Yes
17	Zipcode	Shows the zipcode in address field.	Yes
18	Country	Shows the country in address field.	Yes
19	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	Yes
20	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	Yes
21	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	Yes
22	Temperature(F)	Shows the temperature (in Fahrenheit).	Yes
23	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	Yes
24	Humidity(%)	Shows the humidity (in percentage).	Yes
25	Pressure(in)	Shows the air pressure (in inches).	Yes
26	Visibility(mi)	Shows visibility (in miles).	Yes
27	Wind_Direction	Shows wind direction.	Yes
28	Wind_Speed(mph)	Shows wind speed (in miles per hour).	Yes

29	Precipitation(in)	Shows precipitation amount in inches, if there is any.	Yes
30	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Yes
31	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	No
32	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	No
33	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	No
34	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	No
35	Junction	A POI annotation which indicates presence of junction in a nearby location.	No
36	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	No
37	Railway	A POI annotation which indicates presence of railway in a nearby location.	No
38	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	No
39	Station	A POI annotation which indicates presence of station in a nearby location.	No
40	Stop	A POI annotation which indicates presence of stop in a nearby location.	No
41	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	No
42	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.	No

43	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	No
44	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	Yes
45	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	Yes
46	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	Yes
47	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	Yes