# DSC 530, Data Exploration and Analysis Final Project Summary

## Zicheng (Chance) Xu

## Date: 8/5/2024

The hypothesis question explored in analysis for the final project is: "Is it more likely to have severe car accidents during the night than during the day?" This hypothesis investigates whether the time of day, based on the astronomical twilight, has a significant impact on the severity of car accidents, as I was contemplating factors such as visibility when it turns dark will reduce the reaction time of the driver that could result in more severe car accidents than during the day. The total number of observations is 186,760 car accidents that happened from Feb 2016 to Mar 2023 in Miami, FL.

In the analysis, I used multiple steps, starting off with histograms where I found that most of the accidents were graded at level 2 severity out of 4 levels, where 1 represents the least severity impact to the road and 4 represents the most. In addition, there were more than 25,000 records having distance impact of 0 mile, meaning no effect to the road at all. I found 1% of the dataset where the temperature recorded was below 0 Fahrenheit. Given unknown cause of the outliers and the small percentage, I removed them from the dataset. In addition, there were about 0.31% of the dataset that had recorded humidity less than 30% which I removed them because Miami, FL is located at the humid subtropical climate zone with annual average humidity of around 60% and should be skewed to the right according to internet resources that reported extreme humidity of over 90%, compared to the driest state of Nevada with annual average humidity of 38.3%. According to the EDA analysis, none of the five numerical variables that are helpful for the hypothesis testing is normally distributed, but the mean of both Distance(mi) and Severity are greater during the night than during the day. PMF and CDF both indicate night car accidents tend to have larger impact. According to the analytical distribution, the time between car accidents in minutes is close to exponential distribution with large lambda value. Scatter plots as well as correlation indicate wind speed and temperature as well as precipitation and humidity have weak positive correlation. Hypothesis testing indicates it is indeed more likely to have more severe car accidents during the night than during the day, as the p value is much less than 0.05 for both Chi-squared and one-sided mean diff testing.

I feel the time series analysis is missed during the analysis as the data was collected over time and time series analysis could help me identify some patterns or trends if there were any. Also, other potentially influential factors, such as weather conditions, road type, and driver behavior, were not included. These variables could provide additional insights into the factors contributing to severe car accidents.

Incorporating additional variables like weather conditions, road types, and traffic density could have enhanced the analysis. These variables might interact with the time of day and accident severity, providing a more comprehensive understanding of the causes of severe car accidents.

One assumption made was that the relationship between the time of day and accident severity is direct. However, the relationship could be more complex, potentially involving third factors and confounding variables that were not considered.

Handling missing data and identifying appropriate methods for dealing with outliers were challenging. Ensuring the accuracy and reliability of the statistical tests conducted was another critical aspect. The chi squared test helped me reject the null hypothesis about the distribution of the severity between the day and night data, however I had to do one sided mean diff test to confirm it’s more likely to have more severe car accidents during the night than during the day.