# **DSC530- T301 Data Exploration and Analysis**

# **Term Project**

# **Lenin Kamma**

* Statistical/Hypothetical Question
  + Do most of the fatal accidents happen on Monday?

Accidents happen on different days of the week. But there is a perception that most of the fatal accidents occur on Monday. I want to examine this claim based on the observations.

* Outcome of your EDA

I have taken the dataset from Kaggle.com to perform the analysis. 5 different variables have been taken and null values have been removed from the final dataset. Some of the categorical variables have been transformed and assigned a numerical value. There is no outlier in the day\_of\_week values. Most of the accidents have been observed as slight. Also, I observed that a very few accidents happened when the speed limit was 20 miles. In this project I am focusing on fatalities on Monday. So these outliers have no impact on the outcome. I created a new dataframe with day\_of\_week and total fatal accidents and used this to draw scatter plots and complete the data exploration.

**Observations**

* The Fatalities are found to be high on Saturday
* Fatalities are lowest on Wednesday
* The Pearson correlation coefficient between the day of the week and fatalities is 0.357, which indicates not so high correlation
* The covariance is found to be 7.28333333e+01, which is positive suggesting that both variables change in the same direction with each other
* The chi-statistic is 5.8974 and degrees of freedom is 12. The p-value is 0.92

The high p-value indicates, we expect to see the observed total deviation, or more about 92% of the time.

* Built a multiple regression model to find the effect of Road\_type and Speed\_limt on the Accident Severity. The result of the analysis are

Intercept: 2.887265567557953

Coefficients: [0.02644138 -0.00240911]

* The negative coefficient value (-0.002) indicates that as the speed increases the accident severity value decrease, which means there are more fatal(value 1) accidents at higher speed limit
* The positive coefficient value (0.026) indicates that as the road\_typ value increases, the accident severity value increases. Surprisingly, as road condition changes from dry to flood, the severity of accidents reduced
* What do you feel was missed during the analysis?

I think the analysis missed other factors in the dataset such as speed of the vehicle to find the true correlation with fatality

* Were there any variables you felt could have helped in the analysis?

Although I took day\_ of\_ week for analysis, the time of accident could have helped complete the analysis more accurately

* Were there any assumptions made you felt were incorrect?

None

* What challenges did you face, what did you not fully understand?

I had to transform some of the categorical variables and remove some variables which I found add no value to the analysis. I had to create new data frame to find the total fatality count by day of the week.

* Submit a link to your repository to the assignment link during the final week of class.

Submitted