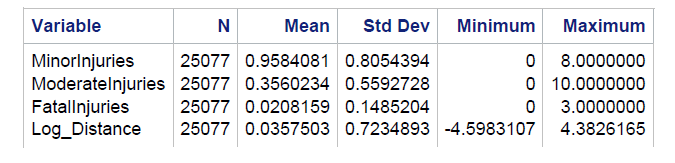
**Multivariate Data Analysis of Traffic Crash Characteristics and Injury Severity**

**1. Introduction**

Traffic crashes continue to be a leading public safety concern globally, resulting in injuries, fatalities, and property damage. This report investigates the relationships between key crash characteristics and the severity of injuries sustained in traffic crashes. The project specifically aims to evaluate the propositions that traffic crashes involving speeding and hit-and-run incidents result in more severe or fatal injuries, and that city property damage correlates with increased injury severity. The analysis was carried out using a multivariate statistical approach, applying descriptive statistics, factor analysis, multiple regression, and cluster analysis to identify patterns and relationships in the crash data.

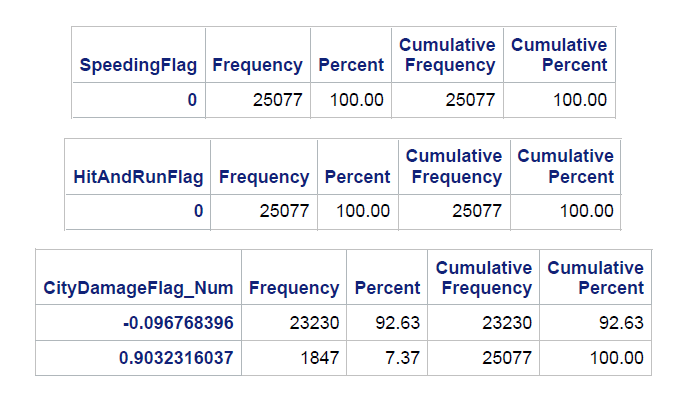
The dataset analyzed was sourced from the San Jose traffic crash records, comprising 25,077 observations and multiple variables capturing injury outcomes, crash factors, and geographic indicators. The dependent variables include Minor Injuries, Moderate Injuries, and Fatal Injuries, while the independent variables consist of SpeedingFlag, HitAndRunFlag, CityDamageFlag, Log-Transformed Distance, and Intersection Number. By employing a systematic approach to data exploration and modeling, this report uncovers insights into injury severity patterns and identifies opportunities for safety improvements.

**2. Descriptive Statistics**

****

Descriptive statistics serve as the foundation for understanding the distribution and behavior of the key variables in the dataset. The analysis begins with summarizing the dependent variables: Minor Injuries, Moderate Injuries, and Fatal Injuries. On average, minor injuries were most prevalent in crashes, with a mean value of 0.9584 and a standard deviation of 0.8054. Moderate injuries were less frequent, with a mean of 0.3560 and a standard deviation of 0.5593, while fatal injuries had the lowest occurrence, with a mean of 0.0208 and a standard deviation of 0.1485.

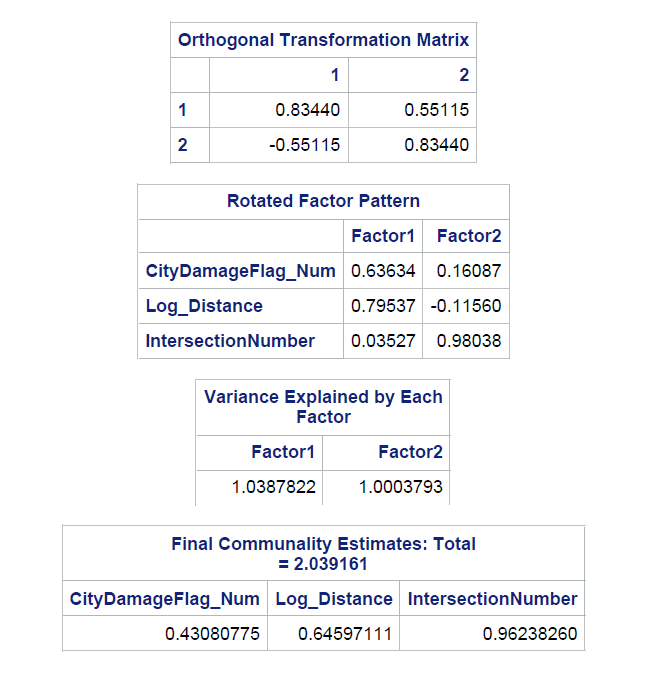
The Log-Transformed Distance variable, which was derived from the original crash distance, had a mean value of 0.0358 and a standard deviation of 0.7235. The use of a log transformation improved the interpretability of this metric variable by addressing skewness.



A frequency analysis was performed for the binary variables, including SpeedingFlag, HitAndRunFlag, and CityDamageFlag. Interestingly, both the SpeedingFlag and HitAndRunFlag variables contained only zero values, indicating that data for speeding and hit-and-run incidents were missing. However, the CityDamageFlag variable revealed more significant results: approximately 92.63% of crashes had no city property damage, value -0.0967, while 7.37% involved city property damage, value 0.9032. While city property damage represents a small proportion of crashes, its impact on injury severity becomes evident in later sections.

**3. Factor Analysis**

To identify latent factors that may influence injury severity, an Exploratory Factor Analysis (EFA) was conducted using the Principal Components method with Varimax rotation. Factor analysis allows for reducing the dimensionality of the data and understanding the underlying structure of the observed variables. The analysis retained two factors, which explained a significant portion of the variance in the dataset.



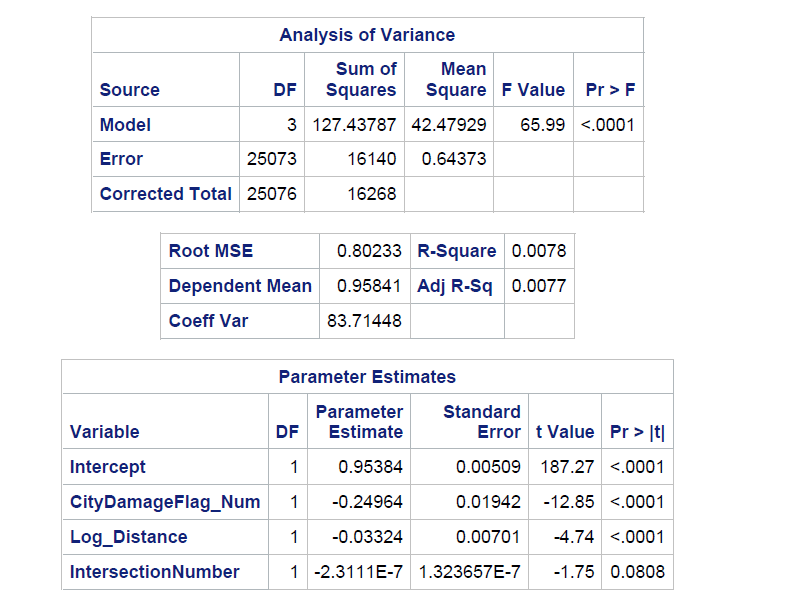
The first factor, which we label as the Distance Factor, showed strong loadings for the Log\_Distance variable, with a loading value of 0.7954. This suggests that the geographical distance of the crash is a key determinant in injury outcomes. The second factor, termed the Intersection Factor, demonstrated high loadings for IntersectionNumber, with a loading value of 0.9804. This indicates that crashes occurring near intersections are associated with unique patterns of injury severity.

Interestingly, the CityDamageFlag\_Num variable loaded onto both factors, indicating its relevance across different dimensions of crash conditions. The total variance explained by the two factors was 2.039, with Factor 1 accounting for slightly more variance than Factor 2. These findings provide a basis for understanding how crash location and city property damage contribute to injury outcomes.

**4. Multiple Regression Analysis**

To evaluate the relationships between crash characteristics and injury severity, multiple regression analysis was performed for each dependent variable—Minor Injuries, Moderate Injuries, and Fatal Injuries. The independent variables included CityDamageFlag\_Num, Log-Transformed Distance, and IntersectionNumber. Below are the results and interpretations for each regression model.

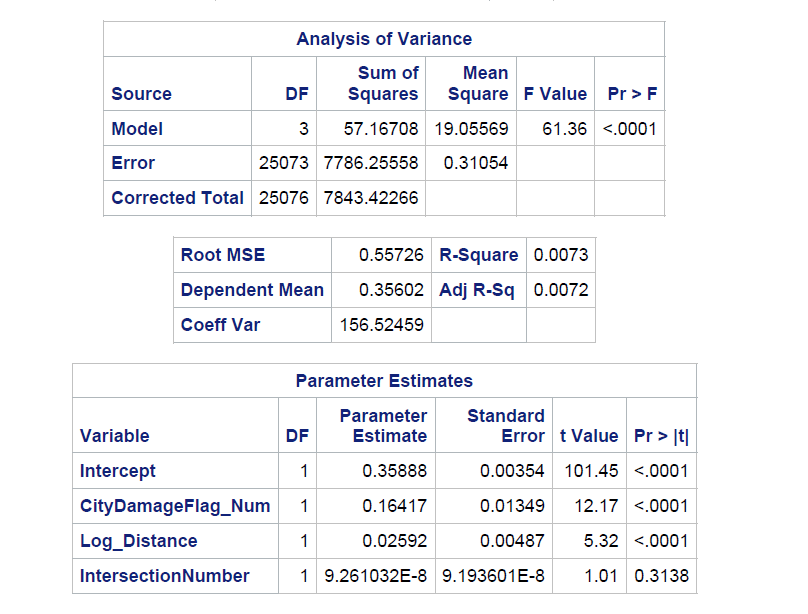
**4.1 Minor Injuries Regression**

****

The CityDamageFlag\_Num variable had a coefficient of -0.2496, p < 0.0001, indicating a negative relationship between city property damage and minor injuries. This suggests that crashes involving city property damage are less likely to result in minor injuries, potentially because these crashes are more severe.

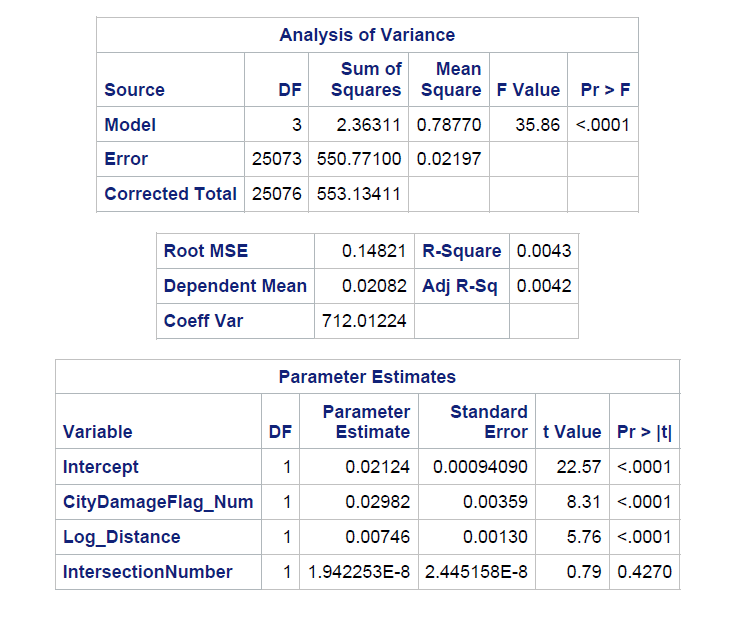
The Log\_Distance variable had a small but statistically significant negative coefficient of -0.0332, p < 0.0001, implying that crashes occurring farther from a reference point are slightly less likely to result in minor injuries. IntersectionNumber was not found to be a significant predictor in this model.

**4.2 Moderate Injuries Regression**

In this model, CityDamageFlag\_Num emerged as a significant positive predictor, with a coefficient of 0.1642, p < 0.0001. This indicates that crashes involving city property damage are associated with an increased likelihood of moderate injuries.

The Log\_Distance variable also exhibited a positive relationship, with a coefficient of 0.0259, p < 0.0001. This suggests that crashes occurring farther from the reference point tend to have a slightly higher likelihood of moderate injuries. As with the minor injuries model, IntersectionNumber was not statistically significant.

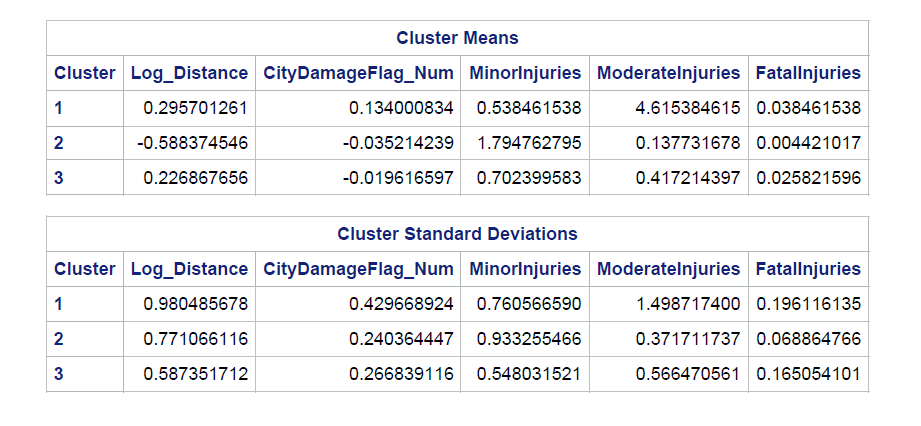
**4.3 Fatal Injuries Regression**

The CityDamageFlag\_Num variable was a significant positive predictor, with a coefficient of 0.0298, p < 0.0001. While the effect size is small, this result indicates that crashes involving city property damage slightly increase the likelihood of fatal injuries.

The Log\_Distance variable also showed a statistically significant positive relationship, with a coefficient of 0.0075, p < 0.0001. This suggests that geographical distance plays a minor role in predicting fatal injuries. Similar to the previous models, IntersectionNumber was not significant.

**5. Cluster Analysis**

Cluster analysis was conducted to identify groups of crashes with similar characteristics. Using the FASTCLUS procedure, three clusters were identified based on the variables Log\_Distance, CityDamageFlag\_Num, Minor Injuries, Moderate Injuries, and Fatal Injuries.



**Cluster 1: High Moderate Injuries**

Cluster 1 is characterized by crashes with a high number of moderate injuries. The average Log\_Distance for this cluster is 0.29, while CityDamageFlag\_Num is positive at 0.13. The mean number of moderate injuries in this cluster is 4.62, indicating that this group represents crashes of moderate severity.

**Cluster 2: Higher Minor Injuries**

Cluster 2 includes crashes with a higher prevalence of minor injuries. The average Log\_Distance is negative at -0.59, and CityDamageFlag\_Num is slightly negative at -0.03. The mean number of minor injuries in this cluster is 1.79, making it distinct from the other clusters.

**Cluster 3: Balanced Injuries**

Cluster 3 represents crashes with balanced numbers of minor and moderate injuries. The average Log\_Distance is 0.23, and CityDamageFlag\_Num is close to zero at -0.02. This group has an average of 0.42 moderate injuries and 0.96 minor injuries.

**6. Findings and Conclusions**

The findings from this analysis provide valuable insights into the relationships between crash characteristics and injury severity. City property damage emerged as a key predictor, positively correlating with moderate and fatal injuries while negatively associating with minor injuries. This indicates that crashes involving city property damage are generally more severe. Geographic factors, as captured by Log\_Distance, also play a role, albeit with smaller effect sizes.

**7. Recommendations**

1. **Target High-Severity Crash Locations**: Focus on intersections and areas with higher Log\_Distance values where moderate and fatal injuries are prevalent.
2. **Implement Safety Measures for City Property Damage**: Given its correlation with severe injuries, enforcing safety measures in urban areas can help reduce crash severity.
3. **Cluster-Specific Interventions**: Tailor safety initiatives to address the needs of specific clusters, such as high-moderate injury crashes.

**8. Conclusion**

This report highlights the critical factors influencing injury severity in traffic crashes. The findings provide actionable insights that can help improve traffic safety measures and inform data-driven decision-making for urban planners and policymakers. Addressing data gaps and implementing targeted interventions will be crucial in reducing traffic-related injuries and fatalities.