**NYC Motor Vehicle Accident Data Analysis**

ISYS 4193 Final Project Report

Professor EmmaLe Davis

December 7th, 2023

The following report covers insights we discovered related to NYC motor vehicle accident data. Using a dataset that has over 480,000 accidents reported from 2019-2022, our findings highlight key issues to be brought to the attention of the NYPD and NYCDOT. We aim to illustrate these issues via data analysis techniques learned this semester, specifically SAS EG and Tableau.

**Tham Cao**, Coordinator & Data Analyst. Supply Chain & Information Systems. thamcao@uark.edu.

**Georgia Smith**, Data Analyst. Accounting & Information Systems. gms007@uark.edu.

**Abigail Butler**, Data Analyst. Marketing & Information Systems. afbutler@uark.edu.

**Amit Patil**, Computer Scientist & Data Analyst. Data Science. ampatil@uark.edu.

Table of Contents

[**Executive Summary** 3](#_Toc153457829)

[**Problem and Motivation** 3](#_Toc153457830)

[**Approach** 3](#_Toc153457831)

[**Dataset** 4](#_Toc153457832)

[**Tools and Analytics** 4](#_Toc153457833)

[**Results and Recommendations** 5](#_Toc153457834)

[**Contributions and Uniqueness** 6](#_Toc153457835)

[**Appendix** 8](#_Toc153457836)

[**References** 19](#_Toc153457837)

## **Executive Summary**

For the past two months, our team has worked diligently on analyzing a large dataset comprised of approximately 487,000 different motor vehicle accidents in New York City from 2019-2022. This data includes the type of vehicles involved in the accident, the time and date the accident occurred, the location of the accident, the number of injuries/casualties, and types of people involved. We then used our statistical analysis knowledge to draw conclusions about relationships between the variables to present to the NYPD and NYCDOT. We came to multiple conclusions, including that accidents in the summer are deadlier/cause more injuries, deaths spike around 3-5 AM, motorcyclist injuries were most common around 9-10 PM, total injuries and deaths significantly increased during the COVID-19 pandemic, and Staten Island had the highest average injuries and deaths per accident.

We would like to note that this data does not include accidents that did not result in injuries or casualties, as well as accidents that resulted in less than $1,000 in damages (Police Department (NYPD), 2023). Notably, we also had to clean the data to get rid of errors, which reduced the number of accidents from over 2 million to just over 480,000.

## **Problem and Motivation**

New York City is the densest and most populous city in the United States. The megalopolis has a wide array of alternative transportation methods, such as the subway system, buses, and cycling, but even with notoriously high levels of traffic congestion, many individuals drive. Given the number of individuals that live in, visit, and commute to the city, safety concerns must be one of the city’s highest priorities. As such, our team wanted to explore the number and frequency of motor vehicle collisions throughout the city, as well as the effects on the types of people involved. Our hope is that our insights can shed light on the severity of motor vehicle accidents in New York City and help to implement solutions to decrease the overall number of injuries and casualties the city experiences.

## **Approach**

Our approach was to first explore the data set, then identify problem areas. The dataset was quite large and unorganized, so our computer scientist Amit cleaned the data to help reduce the number of statistical errors we were encountering. This action took the file from over two million rows down to a little over 480,000 rows, which made analysis much more manageable.

Each of us had a variety of relationships we wanted to explore. Tham focused on leading contributing factors, Amit focused on accidents by season and borough, Georgia focused on accidents by month and the COVID-19 pandemic, and Abigail focused on the times at which accidents happened. By delegating tasks, we were each able to create Tableau storyboards to present in-class for constructive criticism. From there, we began to brainstorm business questions and subsequent analysis types for our data oversight and insight outline.

About only half of our original business questions were feasible for analysis. We had to consider that we did not have reliable data on population density based on boroughs, crime rates, and other factors that had we included, would have pushed back our project timeline. To stay in scope, we focused on the data strictly within the file and delved further into their relationships. We also noticed that the columns for vehicle type and contributing factors contained many typos and overly specific categories, and it would’ve been far too time consuming to clean and organize these columns. We decided not to run hypothesis tests or regression analysis on these columns.

## **Dataset**

The dataset we used was “Motor\_Vehicle\_Collisions\_-\_Crashes\_20231029.csv” (Police Department (NYPD), 2023). This is a structured dataset that contains approximately 2.05 million rows, of which each row is an accident where at least $1,000 worth of damages or an injury or death occurred. The dataset also contains 29 columns, which contained information on when and where an accident occurred, how many people were injured/killed, what types of people were injured/killed, causes for the accident, and what types of vehicles were involved. While cleaning the data, Amit added the following columns: “YEAR”, “MONTH”, “SEASON”, and “COVID STATUS”. See [Figure 1](#Figure_1) for the full data dictionary.

## **Tools and Analytics**

For data wrangling, Amit used Python through an IPYNB notebook, with heavy use of the pandas package. We first wanted to look at how many crashes occurred in each year, but before we could do that, we had to cast the “CRASH DATE” column to the datetime64[ns] type and create a new column for year. Through counting the number of crashes in each year, we could determine that our dataset had very few values for years between 2012-2018 and 2023 compared to other years, so we decided to filter these years out.

We then decided to examine the null values. By counting the nulls in each column, we noticed that location, contributing factor, and vehicle type columns had a high number of nulls. We also noticed that there was 1 row where “NUMBER OF PERSONS INJURED” had a null value. We inspected this row and we saw that 1 cyclist was injured, so we changed the null to a 1. We also noticed that “LOCATION” had values of “(0.0, 0.0)”, and “LATITUDE” and “LONGITUDE” had values of 0. Although these are legitimate coordinates, they were located far away from New York City, so we assumed that these were supposed to represent unknown locations and replaced them with nulls. Finally, we noticed that the contributing factor columns had values of “Unspecified”, which were replaced with nulls.

Columns for month and season were also created, to make hypothesis testing on the time of year significantly easier. A column for COVID status was also added. The national emergency for the spread of COVID was declared on March 3rd, 2020, so accidents that occurred before this date were considered “Before COVID”. Since our data ends before the national emergency was lifted on May 11th, 2023, all other dates are considered “During COVID”.

Once our data was sufficiently cleaned, we decided to use Tableau to simultaneously create visualizations and conduct exploratory data analysis. All other analyses were done using either ANOVA tests or t-tests through SAS EG, with an alpha set at 0.05. We attempted to create regression lines to predict mean injuries/deaths per accident, but since the adjusted R^2 was less than 0.02 for both regression lines, we decided not to include these in our results.

## **Results and Recommendations**

It is important to note that although we decided not to run any hypothesis testing for contributing factors, we knew it was still a very important element of the dataset and so included visuals within our Tableau visualization to find the leading cause of accidents. This turned out to be distracted driving, which also resulted in the most injuries; however, unsafe speedy change resulted in the most deaths ([Figure 2](#Figure_2)). We also didn’t run any hypothesis tests on vehicle types due to incohesive formatting, spelling errors, etc.

We started by looking at whether the season the accident occurred in has an impact. Our ANOVA test for mean injuries per accident by season resulted in a p-value of <0.0001, which means we rejected the null hypothesis that all seasons have equal mean injuries per accident ([Figure 3](#Figure_3)). Running Tukey’s HSD test showed that every season had a significantly different mean injuries per accident ([Figure 4](#Figure_4)). Our Tableau visualization shows that summer has the highest average and winter has the lowest ([Figure 5](#Figure_5)). Our ANOVA test for mean deaths per accident by season resulted in a p-value of 0.0237, which means we rejected the null hypothesis that all seasons have equal mean deaths per accident ([Figure 6](#Figure_6)). Unlike injuries, Tukey’s HSD test shows that the only two seasons that had a significant difference in rates were summer and spring ([Figure 7](#Figure_7)). Our Tableau visualization shows that summer has the highest average and spring has the lowest ([Figure 8](#Figure_8)).

We also investigated whether the COVID-19 pandemic had an effect on the average number of injuries as well as deaths. We ran two t-tests using COVID status as the independent variable for both tests, and number of persons injured for the dependent variable of the first test ([Figure 9](#Figure_9)) and the number of persons killed as the dependent variable for the second test ([Figure 10](#Figure_10)). The output for both tests showed a Folded-F p-value of <.0001, which indicated that the variables had unequal variances. Looking at Satterthwaite, given the unequal variances, the p-value is also <.0001 for both tests. As such, we rejected the null hypothesis that before-COVID and during-COVID years have equal means. To our surprise, and in support of the alternative hypothesis, the Tableau visualization showed an increase in both the average number of injuries and total number of injuries during the pandemic ([Figure 11](#Figure_11)). The average number of deaths and total number of deaths also increased during the pandemic ([Figure 12](#Figure_12)).

We then looked at whether the borough has an impact. Our ANOVA test for mean injuries per accident by borough resulted in a p-value of <0.0001, which means we rejected the null hypothesis that every borough has equal mean injuries per accident ([Figure 13](#Bookmark1)). Tukey’s HSD test shows that every borough had a significantly different mean injuries per accident ([Figure 14](#Figure_14)). Our Tableau visualization shows that Staten Island has the highest average value and Manhattan has the lowest ([Figure 15](#Figure_15) & [Figure 16](#Bookmark4)). Our ANOVA test for mean deaths per accident by borough resulted in a p-value of 0.2973, which means we fail to reject the null hypothesis that every borough has an equal mean deaths per accident ([Figure 17](#Bookmark5)).

To examine if time period has an impact, we first had to create a variable for time period. We did this in SAS EG by defining “Morning” as 6:00 AM – 11:59 AM, “Afternoon” as 12:00 PM – 4:59 PM, “Evening” as 5:00 PM – 7:59 PM, and “Night” as all other time values. Running an ANOVA test for mean injuries per accident resulted in a p-value of <0.0001, which meant we rejected the null hypothesis that every time period had an equal mean injuries per accident ([Figure 18](#Bookmark6)). Tukey’s HSD test showed that every time period has a significantly different mean injuries per accident ([Figure 19](#Bookmark7)). Our Tableau visualization showed that the greatest number of people injured, on average, was at “Night”, occurring specifically at 9 PM ([Figure 20](#Figure_20)). Our ANOVA test for mean deaths per accident resulted in a p-value of <0.0001, which meant we rejected the null hypothesis that every time period has an equal mean deaths per accident ([Figure 21](#Figure_21)). Tukey’s HSD test shows that night is significantly different from all other time periods, and evening and afternoon are significantly different from each other ([Figure 22](#Figure_22)).

Further research should focus on these variables as well as any possible relationships, as we looked at them independently for the most part. Doing so will help paint a more complete picture and see if there is any underlying correlation between independent variables. A specific aspect that could be looked into is whether internal factors within the dataset also contributed to the causes of accidents (e.g., if distracted driving was caused due to morning glare from sun). Our team used the limited information we had, but if we were able to cross check external variables such as the number of pedestrians, bicycle lanes, density, etc., then we might have been able to come up with more conclusive recommendations.

## **Contributions and Uniqueness**

We thoroughly analyzed this dataset to provide NYPD and NYCDOT with insights into what kinds of accidents happen, when they happen, and why they happen, among other things. Our insights and solutions can also be used to find ways to prevent accidents from occurring and help reduce injuries and fatalities.

We were surprised to find that Manhattan had surprisingly low numbers of injuries and deaths per accident. Manhattan is the urban core of New York City, and it contains many of New York City’s notable landmarks. A possible reason for the reduced severity of collisions in this borough is its walkability (Damashek, 2022). Walkable environments have been shown to reduce the number of fatalities, especially for pedestrians (Walker, 2018). They typically go hand in hand with lower speeds, which has a dramatic effect on making accidents less severe. One study found that “a 10% reduction in the average speed resulted in 19% fewer injury crashes, 27% fewer severe crashes, and 34% fewer fatal crashes” (National Association of City Transportation Officials, 2020). Our team recommends that NYCDOT should reduce speed limits and improve walkability in the other boroughs.

Our team also found that vehicular accidents and injuries were overwhelmingly caused by distracted driving. This is very surprising due to strict laws regarding distracted driving that the state of New York has implemented. For example, VTL 1225 (d) “bars the use of portable electronic devices while driving”, which was enacted in response to cell phone use becoming the most common form of distracted driving (Block O’Toole & Murphy, n.d.). There are already fees and point reductions in place for these primary offenses, but harsher punishments should be set for distracted driving.

We also recommend that the NYPD increase police patrols, as this has been shown to decrease fatalities (Farmer, Cline & Campbell, PLLC., 2019). Due to limited funds, we suggest that these increased patrols be concentrated at certain times and boroughs. For example, since a spike in deaths occur from 3 am to 5 am, patrols should be increased at these times. Targeting more patrols in high injury boroughs like Staten Island and Brooklyn could also reduce the average number of injuries and deaths there. This measure aims to address specific risk factors, creating a tangible impact on accident rates in identified hotspots.

Finally, since we found that motorcyclists were the most injured group, we believe that an awareness campaign would create a significant reduction in injuries. For a pilot program, NYCDOT should place targeted “share the road” signs and driver safety advertisements in boroughs with high accident numbers to test if they truly help decrease overall accidents and injury rates. This pilot program will provide valuable insights into the potential success of such initiatives.

However, it is very important to note that there are discrepancies and nuances due to our removal of nulls within the dataset (a relatively large chunk of data) and because of the nature of the dataset which is only police-submitted accident reports. The dataset does not include instances where the police was not involved, accidents that didn’t cause more than $1000 in damages, or accidents that did not result in an injury or death. In short, this means the dataset does not contain every car collision that occurred in New York City during the years we looked at. This fact can mean that our findings may not be truly reflective or 100% representative of the entirety of New York City vehicular accidents. However, the findings and suggestions provided can still be use as a starting point on initiatives and greater efforts to reduce motor vehicular accidents, deaths, and injuries in New York City.

## **Appendix**

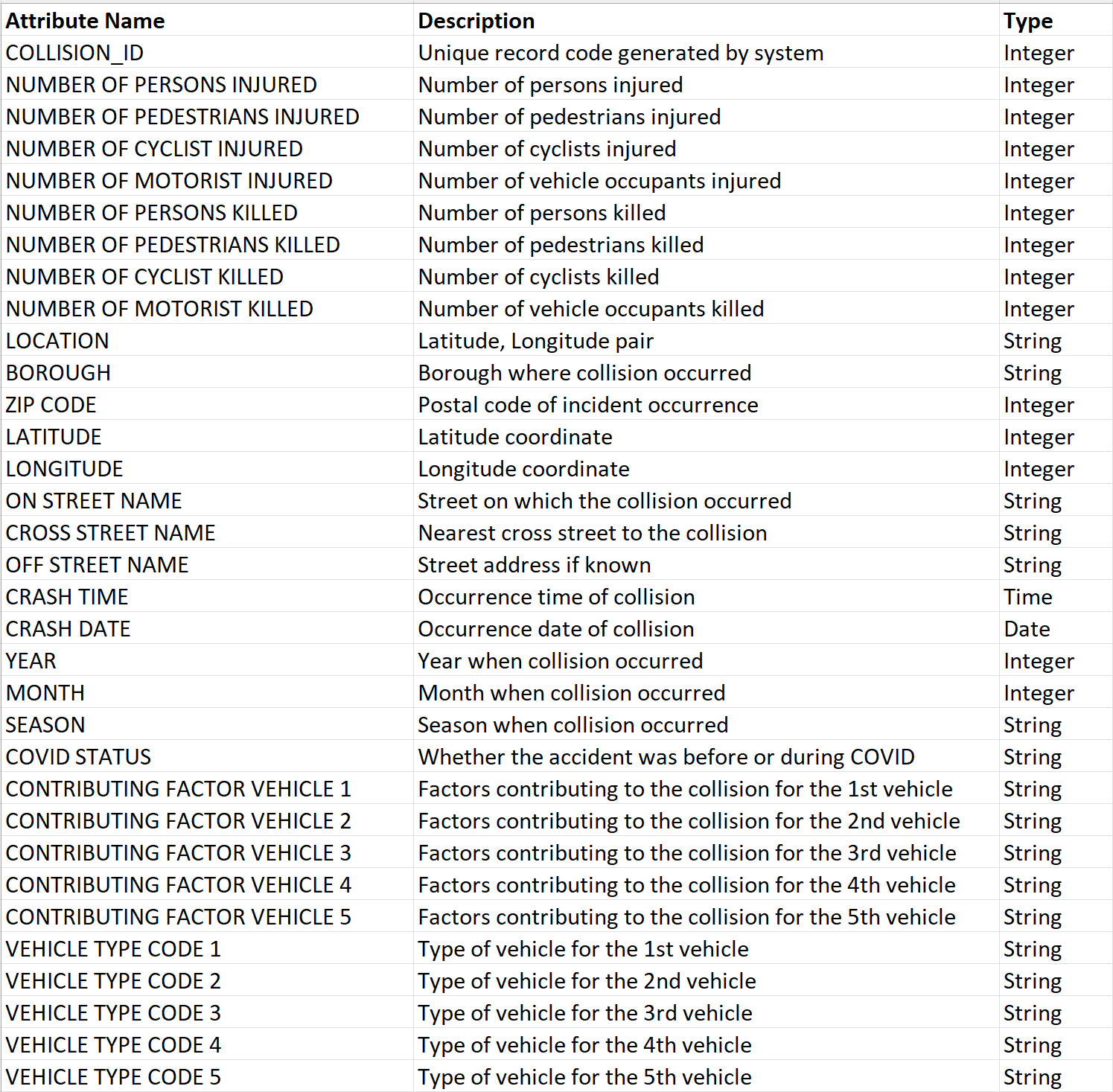


Figure 1: Data Dictionary

A graph with numbers and text

Description automatically generated

Figure 2: Tableau of total accidents by contributing factor

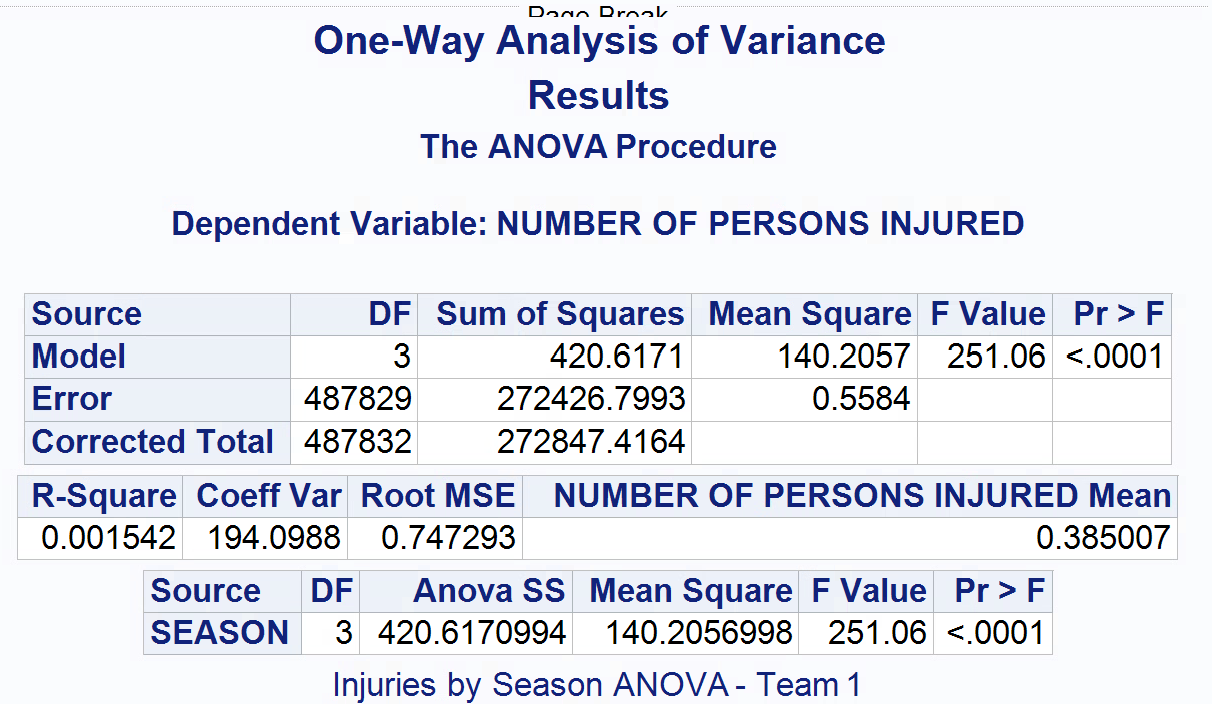


Figure 3: ANOVA test for mean injuries per accident by season



Figure 4: Tukey’s HSD test for mean injuries per accident by season

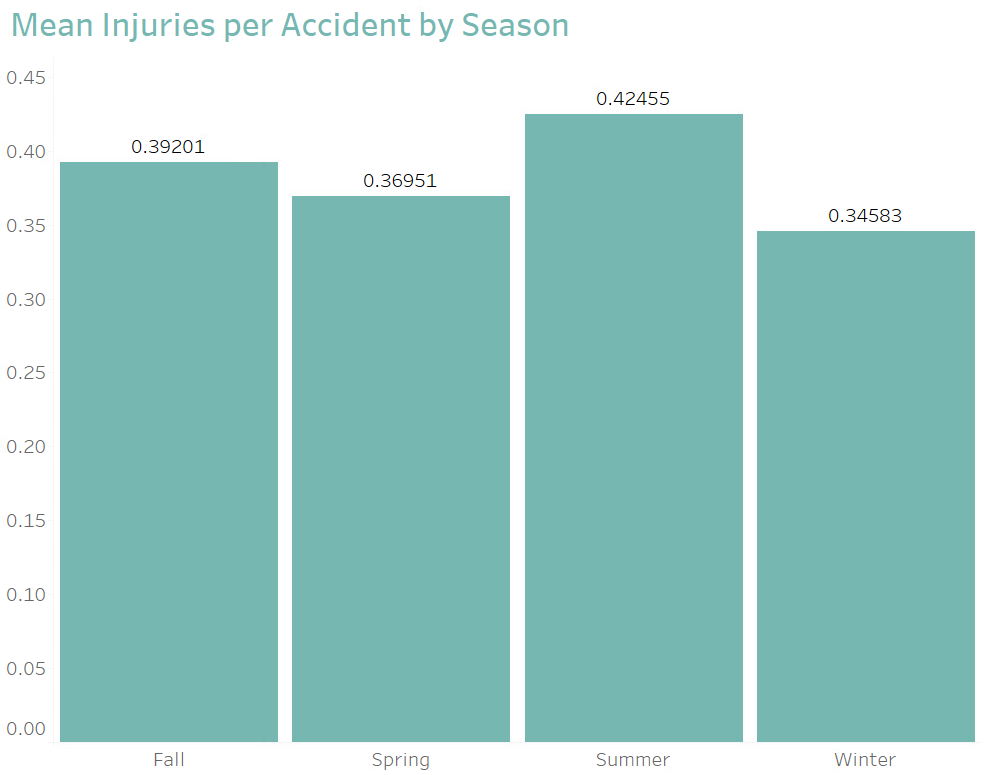


Figure 5: Tableau visualization for mean injuries per accident by season

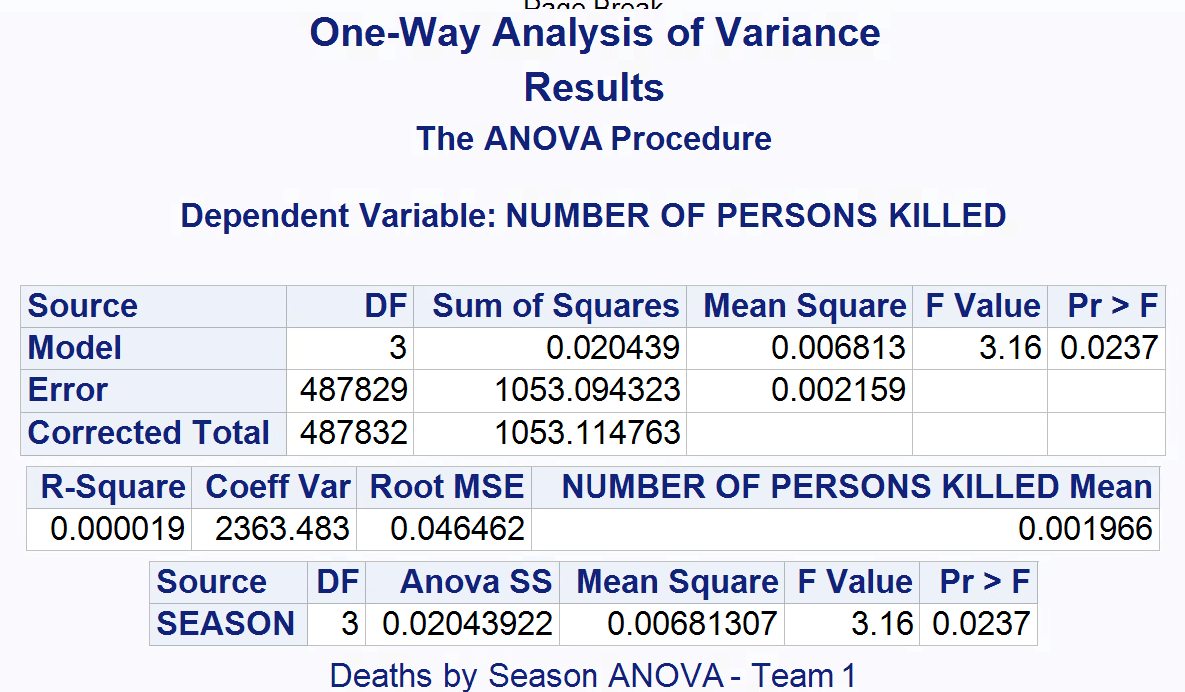


Figure 6: ANOVA test for mean deaths per accident by season



Figure 7: Tukey’s HSD test for mean deaths per accident by season

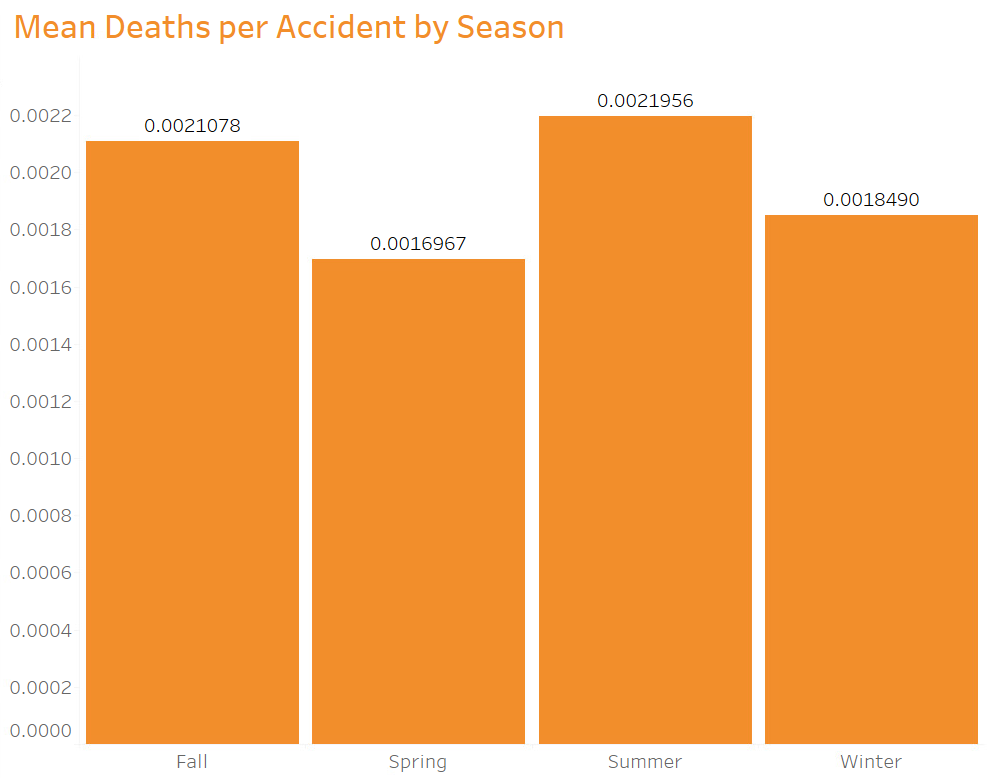


Figure 8: Tableau visualization for mean deaths per accident by season

A screenshot of a test

Description automatically generated

Figure 9: t-test for mean injuries per accident using COVID status

A screenshot of a test

Description automatically generated

Figure 10: t-test for mean deaths per accident using COVID status

A graph with numbers and a bar

Description automatically generated

Figure 11: Tableau visualization for mean injuries per accident by COVID status, with a data callout for total number of injuries by COVID status

A graph with numbers and a bar

Description automatically generated

Figure 12: Tableau visualization for mean deaths per accident by COVID status, with a data callout for total number of deaths by COVID status

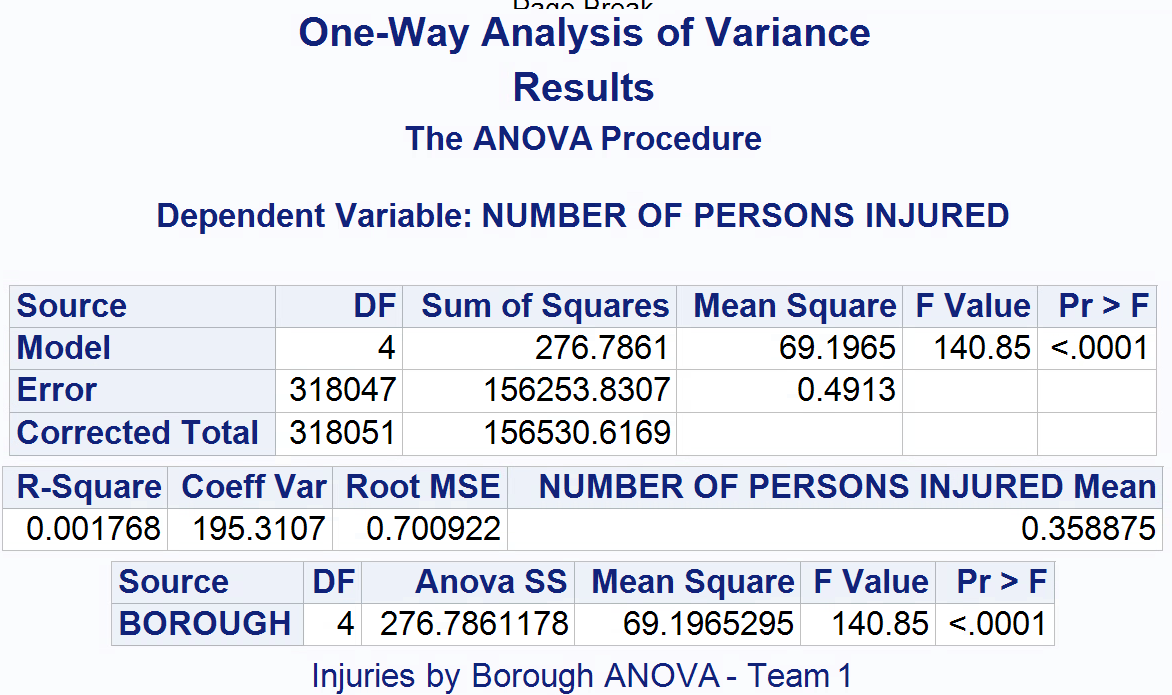


Figure 13: ANOVA test for mean injuries per accident by borough

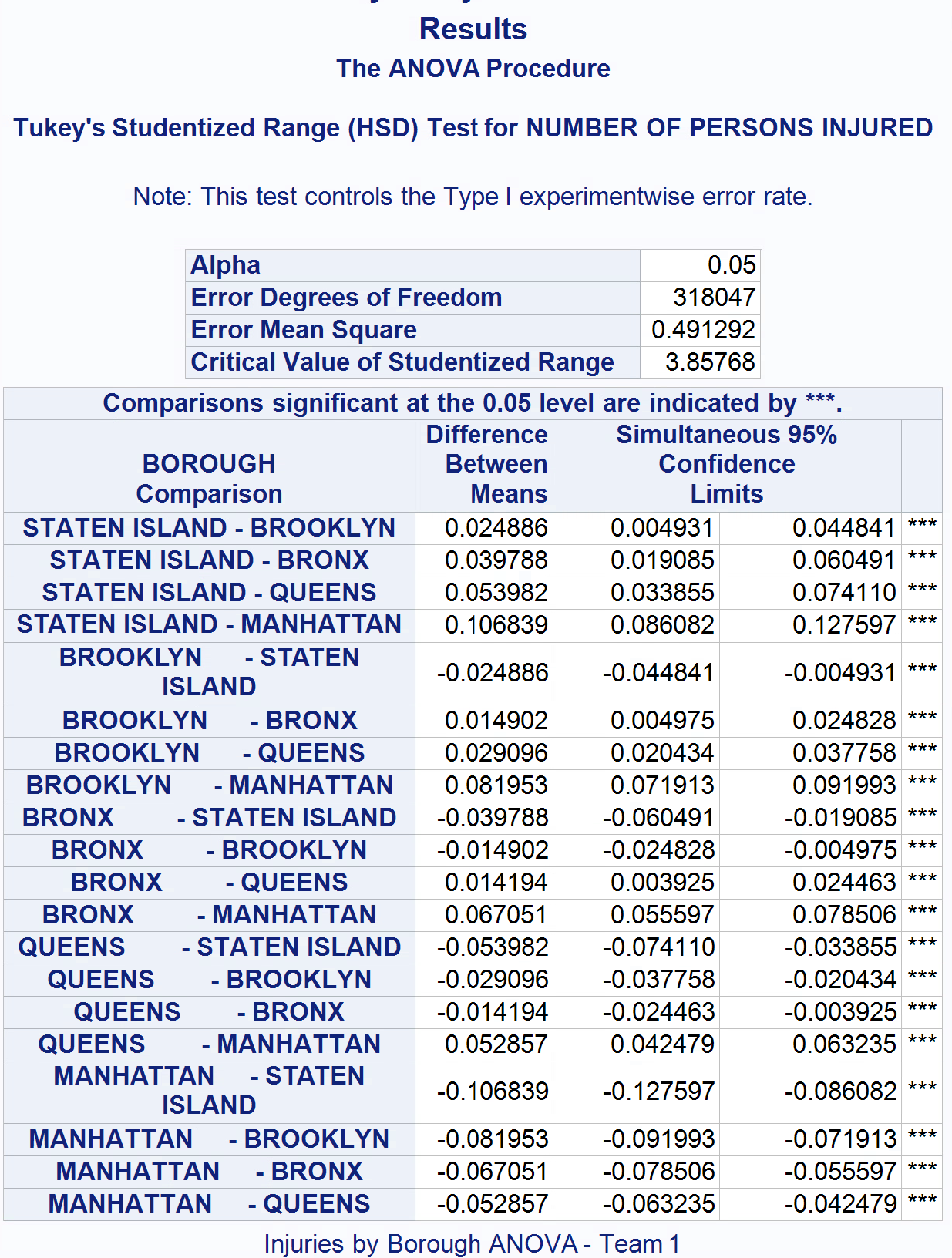


Figure 14: Tukey’s HSD test for mean injuries per accident by borough

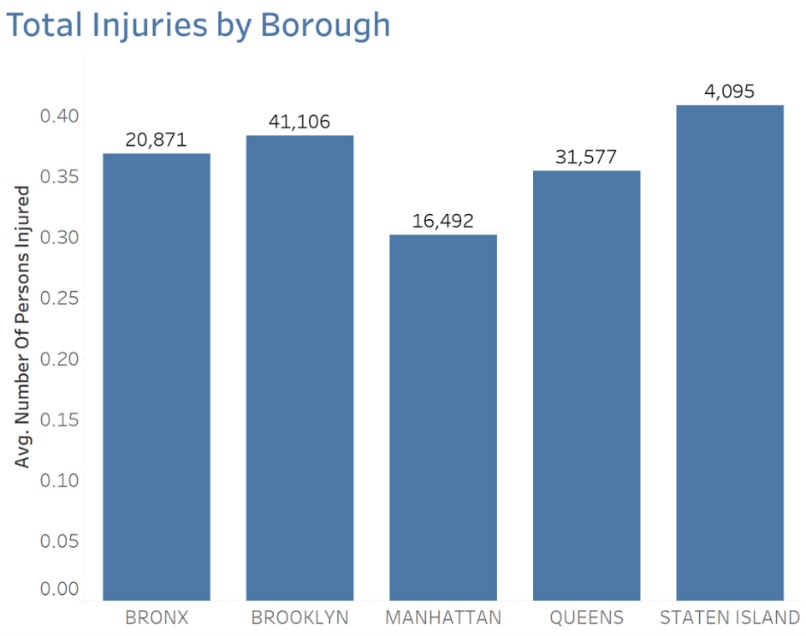


Figure 15: Tableau visualization for mean injuries per accident by borough

A map of different states

Description automatically generated

Figure 16: Manually created visualization of total accidents by borough

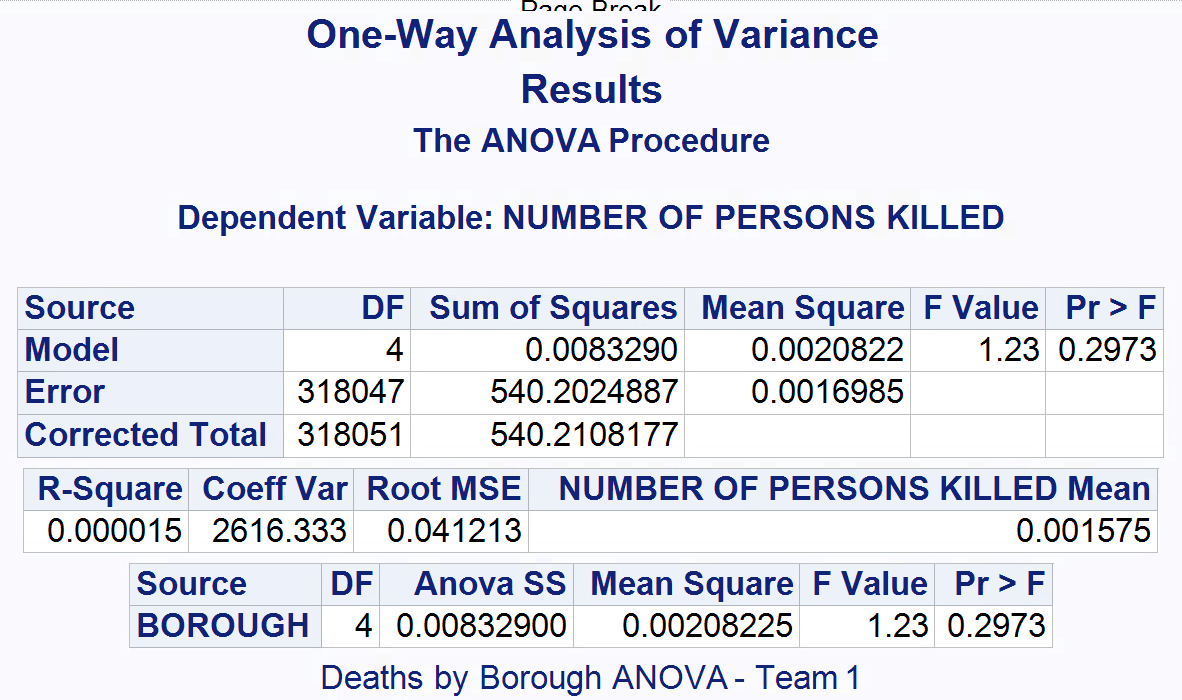


Figure 17: ANOVA test for mean deaths per accident by borough

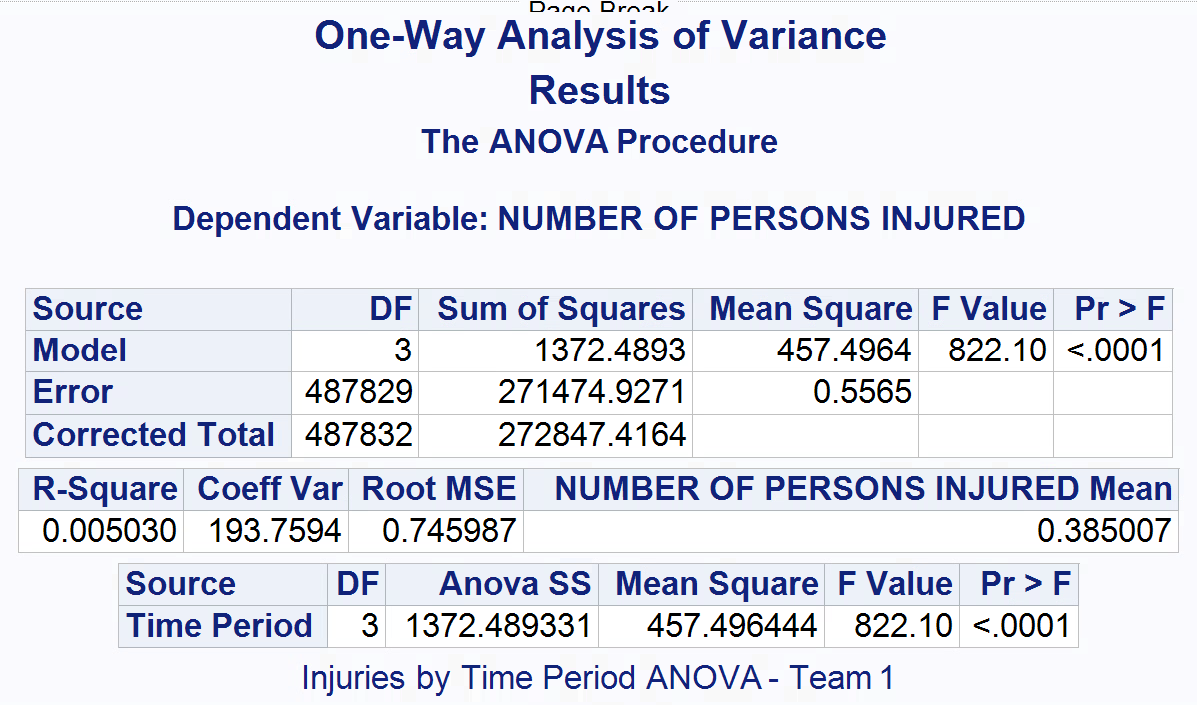


Figure 18: ANOVA test for mean injuries by Time Period ANOVA

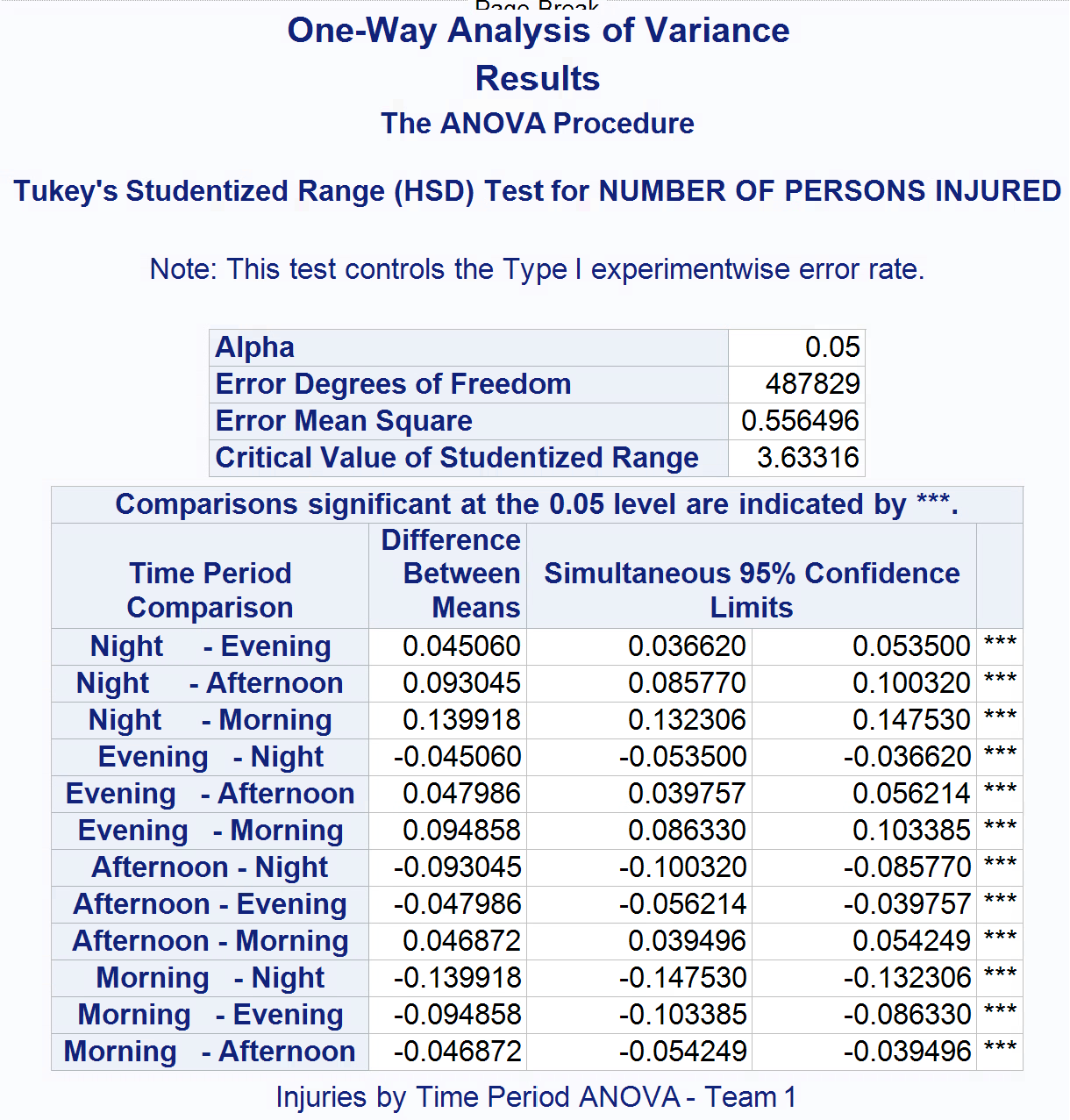


Figure 19: Tukey’s HSD test for mean injuries per accident by time period

A graph with a line

Description automatically generated

Figure 20: Tableau visual of injuries by the hour

A screenshot of a data

Description automatically generated

Figure 21: ANOVA test for mean deaths per accident by time period

A screenshot of a test results

Description automatically generated

Figure 22: Tukey’s HSD test for mean deaths per accident by time period

## **References**

Damashek, J. (2022, April 7). Most dangerous intersections in Manhattan for car accidents. Hecht, Kleeger &amp; Damashek, P.C. https://lawyer1.com/blog/most-dangerous-intersections-in-manhattan-for-car-accidents/

*Distracted driving: a serious problem in New York*. (n.d.). Block O’Toole & Murphy. https://www.blockotoole.com/articles/distracted-driving-a-serious-problem-in-new-york/#:~:text=Generally%20speaking%2C%20VTL%201225(c,other%20applications%20on%20mobile%20devices.

Farmer, Cline & Campbell, PLLC. (2019, March 19). *Greater police presence leads to reduction in traffic fatalities: Farmer, Cline & Campbell, PLLC: Charleston, West Virginia*. Farmer, Cline & Campbell, PLLC. https://www.farmerclinecampbell.com/articles/greater-police-presence-leads-to-reduction-in-traffic-fatalities/

Police Department (NYPD). (2023). *Motor vehicle collisions - crashes | NYC Open Data*

[Dataset]. NYC OpenData. https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

*Speed kills*. National Association of City Transportation Officials. (2020, August 28). https://nacto.org/publication/city-limits/the-need/speed-kills/

Walker, A. (2018, January 31). How walk-first cities are saving lives. Curbed.

https://archive.curbed.com/2018/1/31/16952856/walking-traffic-deaths-pedestrian-streets