INFO 2950 Final Report

October 3, 2020

1 Pedicyclist Crashes

Isabel Frank- **irf72**
Lavona Harper - **leh224**
May 17th, 2020

The transportation industry has evolved due to the increase of technology and innovation, so there are many options for transportation to get around town. Cycling remains a popular choice of transportation as growing urban cities begin to crowd that is appealing commuting and exercise. Since cyclist are still on the streets, it is important that we are continuing to find ways to reduce the risk that people run when they hop on their bikes to go to work every morning or take their families for a Sunday ride. To reduce the risk that cyclists face, we must investigate the common factors that contribute to increasing pedicyclist fatalities. Some factors that increase fatalities include: increasing bike sales, season, time of day, day of the week, and population size. This report analyzes and discusses the different factors that could contribute to pedicyclist crashes. Our hypothesis determines if there is a correlation between the the common factors listed above and cycling fatalities. *Pedicyclist Crashes* applies data science methods to clean,combine, and analyze several datasets to determine the culprits of cycling fatalities.

1.1 1. INTRODUCTION

It's a sunny morning in California and our family is on our way to the beach. The freeway is packed with cars and the streets are crowded. It seems that everyone had planned to go to the beach today. It is normal to see cyclists roaming around in beach cities, so one would think that drivers in the area would be more cautious. At least that's what we were convinced of until we witnessed an unfortunate accident. In the right lane, there was cyclist waiting at the stop light with the rest of the cars. When the light turned green, the cyclist sped up, but not fast enough. The car behind him gently bumped his tire, but the force of the man's truck sent the cyclist flying, far from his bike. All we saw was the agony in his face as his body hit the ground. Our car drove past him and that was the last we saw of this innocent man. We could only think, *Is this man going to be ok?* One accident could lead to a man's death, simply because people are not taking precautions as drivers and cyclists.

1.1.1 1.1 Purpose

The purpose of this detailed report is to identify commonalities that may be contributing to cycling fatalities. Analyzing these commonalities will help public safety organizations understand necessary changes to increase the safety of the cycling community and also educating cyclist about factors that may be increasing their risk of injury or death. Technology has provided the world with

resources that can help predict traffic flow and weather. The reality is that there are some factors that people have no control over until they are in a given situation. For example, it's January and the weather app predicts it's going to be sunny all day. Someone who wouldn't normally take their bike to work decides to ride that day. It isn't until 6pm that the clouds begin to settle in and rain starts sprinkling out of the sky. Now this person is stuck biking home with increased risk of injury. This scenario shows that there is ultimately no way to predict an accident, but it is important that every cyclist knows that they can reduce their risk with awareness of trends and precautions.

1.1.2 1.2 Key Objectives

The key objectives of this analysis is to look at cycling fatalities to determine: 1. Were there any significant differences regarding cycling between different calendar years? 2. Age and gender have an influence on cycling fatalities? If so, which age group and gender? Why is that? 3. A specific time of day or season has an effect on cycling fatalities?

1.1.3 1.3 Significance

The significance of this report is to provide the frequent cyclist with information regarding cycling fatalities. This information is meant to make cyclist more aware of probable risks and prepare for the worst case scenario. It is important that cyclists and public safety organizations are aware of these risks, so they can take necessary precautions to prevent accidents.

1.2 2. DATA DESCRIPTION

Several datasets were collected and examined from credible sources to determine which factors may contribute to cycling fatalities. The following sections will provide more details on each of the datasets and its usage in our analysis.

1.2.1 2.1 National Highway Traffic and Safety Administration (NHTSA)

The National Highway Traffic Safety Administration (NHTSA) is associated with and funded by the US Department of Transportation and provides datasets on vehicle accidents across America. The following source is an official 2016 report from the NHTSA. Many of the tables have observed rows that identify age, state, city, vehicle type and alcohol content and measure attributes like fatalities and population to those variables. Together all these values can help determine links to higher fatality rates. Some of the information displayed for population and ages comes from the US Census and while they do their best to provide accurate information, it is hard to control the accuracy of the information that people provide, therefore it can have a slight affect on the results. The people who fill out the Census report know that the information they provide will be used for necessary population counts, but other than that it can be assumed that people may not be actively aware that information like the time of day cyclist are out is being used to analyze fatalies and increase road safety. However, if there wasn't data like this being collected, then the NHTSA wouldn't know to put a new streetlight in an intersection that is causing more fatalities. The purpose of this data is to enforce standards with state and local governments to ensure safety and take action to reduce deaths and injuries due to road and vehicle related accidents. From this source, we extracted several tables to use in our analysis.

The first dataset was created to display the total fatalities and compare it to the total number and percentage of pedicyclist fatalities between 2007 and 2016. The original dataset contains 9 years and 4 columns. This data set was used as is and only a few adjustments (e.g. dropping commas

and percent signs) were made to the dataset to avoid running into issues when creating graphs. The table columns extracted include the following:
- **Year** = year of crash - **Total Fatalities** = number of total traffic deaths that year - **Pedicyclist Fatalities** = number of pedicyclist deaths - **Percentage of Total Fatalities** = proportion of cycling fatalities to total fatalities in a given year

The second dataset was created to merge two pieces of information displayed by the NHTSA to analyze season and day of the week to time of day. This dataset displays the percentage of pedicyclist deaths in relation to season and day of the week based on time of day. The original dataset included land use, pedicyclist location, and light conditions, but that information was not used in our analysis. The merged dataset contains 7 time frames and 7 columns. To create this dataset, we removed the total column and convert the percentages to decimals. The dataset was then transposed and altered to mimic a more structured and formal data table. Once this was done to both the seasons and days of the week tables, the two tables were merged to create the table needed for our analysis. The table columns extracted include the following:
- **Time Frames** = time cyclist crashes occured (expressed with military time) - **Jan-Feb, Dec (Winter)** = % of crashes in winter - **Mar- May (Spring)** = % of crashes in spring - **Jun-Aug (Summer)** = % of crashes in summer - **Sep- Nov (Fall)** = % of crashes in fall - **Weekday** = % of crashes Monday- Friday - **Weekend** = % of crashes Saturday and Sunday - **Total** = % of all crashes in a given timeframe

The third dataset contains the demographic information of cyclists that were killed. The original dataset contained 20 age groups and 10 columns. To clean this dataset, 3 rows were removed that included grouped and redundant information that was not necessary to include in our analysis. This table is necessary and will be one of the primary datasets analyzed to answer a key objective.The table columns extracted include the following: - **Age (Years)** = age Range of the pedicyclist fatality - **Male Killed** = number of males killed for a given age group - **Male Population (thousands)** = number of males in a given age group - **Male Fatality Rate** = number of male pedicyclists killed per million residents of a given age group - **Female Killed** = number of male pedicyclists killed - **Female Population (thousands)** = number of females in a given age group - **Female Fatality Rate** = number of female pedicyclists killed per million residents of a given age group - **Total Killed Rate** = number of pedicyclists killed in a given age group - **Total Fatality Rate** = number of pedicyclists per 1 million residents in a given age group

The fourth dataset contains a list of cities in the United States with populations greater than or equal to 500,000, and their fatality rates. The original dataset had percentages in the Percentage of Total Traffic Fatalities and commas in Resident Population. To clean this dataset, commas and % symbols were removed in order to make these values into floats so that mathematics could be used. This table will help tell whether or not there is a correlation between Resident Popualtion and Percentage of Total Trafic Fatalities. - **Resident Population** = the resident population in that given city - **Fatality Rate per 1 Million Population Pedalcyclist**= the number of pedalcyclist killed per million people in a given city

1.2.2 2.2 National Bicycle Dealer Association (NBDA)

The National Bicycle Dealers Association collects information from many bike reailers to configure annual bike sales. They are a non-profit that is funded by bike retailers through a partner relationship. The NBDA provides research and education to specialty bike retailers to help boost their sales and reach maximum profit that year by analyzing trends in data. There are three tables in the

data that are used to reach conclusions about bike sales. The first table shows bike sale per million of bikes of all wheel sizes vs bikes with a particular wheel size (columns) for all years between 1981 and 2015(rows). The second tables the different types of bikes (rows) and its percentage of sales in a given year between 2005 and 2012 (columns). The third table shows the number of specialty bike stores (columns) between 2000 and 2015 (rows). The data in these tables are useful in some way to bike sales as it gives insight to the retailer on how to market to their customers to increase profits and can help predict sales for the next year given the previous circumstances. While this source provides analysis of the given tables, the raw data will be used as predictors in this project and the preprocessed data will not be be used. The people that purchased bikes were not aware that their data was being recorded for such large scale purposes. From this source, we extracted the table that had recorded bike sales per year from 1981-2015.

The dataset was created to help bicycle retailers be aware of trends in bike sales to help them boost sales from year to year. However, this dataset counts sales that were made in bikeshops and does not take into account that people may resell their bike or purchase from websites such as Craigslist. If there is a particular year that bike sales increased, then there is a possibilty that the increased number of cyclist may have an effect on cycling fatalities. The table columns extracted include the following: - **Year** = year being observed - **Bicycles Sold (Millions) 20" wheels** = numbers of bikes sold with a particular wheel type - **Bicycles Sold (Millions), all wheel sizes** = number of all bikes sold

This data set was merged with annual bike fatalies because it is also indexed with respect to years.

1.2.3 2.3 Wikipedia - Cycling at the Summer Olympics

This Wikipedia page displays the cycling medal counts by country for each Olypmic Games. We will be using specific counts for medals and the only way these numbers would be different than reported is if there was a mistake in reporting. People may be inspired to start cycling more if they feel inspired by people winning in the Olympic games (e.g. if a Boston Native won a medal in an Olympic event, will it inspire Bostonians to bike more?). There might be issues quantifying success at the Olympic games. A question to ask is if this is a good metric that would be representive of motivation in professional athletes, or would the Tour de France be more telling? In addition, the events hosted are slightly different each year therefore some numbers may be more significant than others. From this source, we extract medal counts by country for each Olympic year.

This dataset was created as a historic reference to keep track of the number of medals each country has earned in the Olympics for cyclying from 1896 to 2016. Since we only require information from the US, we cleaned this dataset by extracting the column with medals earned from only the US. In addition, to make the years into rows to be consistent with other data sets, this dataset was transposed. The data came from the Union Cycliste Internationale (UCI), a French organization that provides information about cycling and cycling events. The table columns extracted include the following:
- **Year** = olympic year - **Medals Earned** = count of medals earned by the US in cycling

This data set was merged with annual bike sales and fatalities to combine all of the information with respect to time in one dataframe.

1.3 3. DATA ANAYLYSIS AND SIGNIFICANCE

From the questions mentioned in 1.2, we will attempt to answer them by comparing different pieces of data to gauge significance.

1.3.1 3.1 Age and Gender

To visualize the relationship with pedicyclist fatality rates with respect to age and gender, a bar graph is used to describe the relationship. In order to formalize whether or not there is a relationship, a t-test is used to test significance.

Code An assumption for a t-test is that variances of both groups are equal. From using numpy.var() for both groups, one can see that the variances are not equal, thus, Welch's t-test was used in the stats.ttest_ind() function

Visual

Hypothesis Testing With the Welch's t-test, we will test whether or not the mean number of deaths per females is significantly different than the average number of deaths for males.

$$H_0 : \mu_{\text{death age groups, Female}} = \mu_{\text{death age groups, Male}}$$

$$H_A : \mu_{\text{death per age groups, Female}} \neq \mu_{\text{death per age groups, Male}}$$

The calculated test statistic was 6.55, and the p-value was 4.86e-05. Thus, at $\alpha = 0.05$, we reject the null hypothesis. Because the null hypothesis is rejected, one can say that the mean number of deaths per age group for males and females is significantly different.

Overall Significance From both the graph and the hypothesis test, we can conclude that there is in fact quite a difference between the rate at which males are being killed in cycling accidents versus the rate at which females are being killed. In the event of a cycling accident, the driver is likely not going to be more likely to hit a male simply because they are a male, as many drivers don't pay particular attention to the gender of a cyclist on the road. However, there is a signifcant difference, and this could be associatiated with the behavior of male cyclists compared to female cyclists. Do female cyclists do a better job with rules of the road (e.g. staying near the shoulder and signaling)? Do male cyclists tend to have more careless behavior such as electing to not wear a helmet, which the lack of protection will in turn increase the chances of dying in the accident?

1.3.2 3.2 Bike Sales and Olympic Medals Earned

In order for someone to invest in purchasing a bicycle, there is likely to be a reason for this motivation. From data from 1992 to 2016, we collected the amount of bikes sold (in millions) from shops in the United States as well as the number of medals the United States earned in cycling during the Olympic Years. To determine if there were more bikes sold in years a t-test was used to test difference in means of annual sales in summer Olympic years versus non-summer Olympic Years.

Coding Mean and variances were calculated to determine if there was significantly different variances, and there was not.

Visual

Hypothesis Testing With a t-test, we will test whether or not the mean number of bike sales in Olympic years is different from mean number of bike sales in non-Olympic years.

$$H_0 : \mu_{\text{Annual Number of Bikes Sold | Olympic year}} = \mu_{\text{Annual Number of Bikes SOld | Not Olympic Year}}$$
$$H_0 : \mu_{\text{Annual Number of Bikes Sold | Olympic year}} \neq \mu_{\text{Annual Number of Bikes SOld | Not Olympic Year}}$$

From calculating by hand, the test statistic is 0.6997, and the p-value is 0.497, so we fail to reject the null hypothesis at $\alpha = 0.05$. Thus, we can conclude that there is no significant difference in the number of bikes sold whether or not it is an Olympic year.

Overall Significance Some may get inspired from watching other people in their nation succeed at the global level. For some, seeing an Olympian from their hometown may inspire them to start cycling as well. While the number of sales was not significantly different, it is worth noting in 1992, 2000, 2008, and 2012 were all Olympic years and the bike sales were at a local maximum. So, in an Olympic year, more people might buy a bike than the previous year, but the excitement wears off the following year. While on average there are about 600,000 more bikes sold in an Olympic year, it is not significantly different.

1.3.3 3.3 Weekend and Weekday Cycling

Cycling happens at all all hours of the day, and each day of the week. Some people are consistant communter bikers to work, and others ride recreationally outside of rush hours. On one hand, an argument could be made that there are more bikers and families riding on the weekends as an activity. On the other hand, there may be fewer commuters, but drivers are more impatient during rush hour which could lead them to making mistakes that could kill a cyclist. To compare if there is a significant difference, between the two groups, a t-test will be used.

Coding Mean and variances were calculated, and variances were not equal. So, a Welch's t-test will be used.

Visual

Hypothesis Testing With Welch's test, we will test whether or not there is a difference in percentage of fatalities that happen during the week versus over the weekend.

$$H_0 : \mu_{\text{Weekday Deaths Proportion - Weekend Deaths}} = 0$$
$$H_A : \mu_{\text{Weekday Deaths Proportion- Weekend Deaths}} \neq 0$$

At $\alpha = 0.05$, we fail to reject the null hypothesis because the test statistic of 0.12 and the p-value is 0.91. Thus, we can conclude that there are about equal number of deaths during the week and deaths over the weekend among pedalcyclists.

Overall Significance The week might be full with predominantly experienced, commuter cyclists who deal with heavier traffic, whereas the weekend may have many families who are not regular riders and less attuned to road etiquette. Because the rates are about equal, these reasons for cycling fatalies may be equally legitamate.

1.3.4 3.4 Pedicyclist Fatalities in Different Times of Day

Times of day can make road conditions in the same area quite different. While visibility may be better during broad daylight, there is more traffic, so more people are on the road who could potentially hit a cyclist. However, in the dark, there may be fewer people on the road who can hit a cyclist, but this is at the consequence of being less visible. With daylight and traffic being the two main conditions for the road at a given hour, we would like to see if there is at any point during the day where there are significantly more deaths. To do this comparison, an ANOVA will be used.

Coding In the original dataset, time frames were grouped into eight 3-hour increments, but for this anaylsis time frames were grouped into four 6-hour groups

In military time, these groups are 0:00-5:59, 6:00-11:59, 12:00- 17:59, 18:00-23:59. In other words, early hours before dawn and rush hour, rushour through midday, midday through the end of rush hour, dark and late hours after rush hour.

When calculating variances, they are not all equal, so they an assumption of the ANOVA test is violated. The tests can still be run, but a violated assumption may make a finding invalid.

Visual

Hypothesis Testing We will use an ANOVA to see if all the percentages are different from one another in each of our new time frame groups.

$$H_0 : \text{All rates of pedacyling accidents are equal in each of the time frames.}$$

$$H_A : \text{At least on of the pedacycling accident rates is not equal to all of the rest.}$$

The test statistic is 6.27, and the p-value = 0.54. Thus, we fail to reject the null hypothesis, and we cannot say that there is a rate that is significantly different from all the rest. However, because an assumption was violated, our conclusions may not be valid.

Overall Significance While we cannot reject at the hard cutoff of $\alpha = 0.05$, it is worth noting that it is not quite a uniform distribution. From the graph above, one can see that between 18:00-20:59 there is a spike of deaths in the winter. This might be due to daylight ending sooner in the winter. People's daily routines in terms of time are not changing, so for somebody who rides after work regularly might not be stopped due to less daylight for the sake of keeping routine. However, visiibility is worse, and that can cause more danger.

1.3.5 3.5 City Population and Pedicyclist Fatalities

In urban settings, bike accidents are more likely to occur because there is inherently more traffic, and generally less space for cyclists. So, doing a test of slope regression test may be worthwhile to see if there a relation exists.

Coding Resident Populations were converted to counts in millions, and the test of the regression slope was only done for populations greater than 2 million. This was done to move past the sporadic behavior in the lower populations and to where the rates move in a more regular trend.

Visual

Hypothesis Test A test of regression slope will be used to determine if there is a relationship between population and fatality rates among pedacyclists.

$$H_0 : \beta_0 = 0, \text{There is no relation between city population and pedicyclist fatalities}$$

$$H_A : \beta_0 \neq 0, \text{There exists a relation between city population and pedicyclist fatalities}$$

After running stats.linregress(), we find that the slope=-0.09, r-squared = -0.18, and the p-value = 0.82. At $\alpha = 0.05$, we fail to reject the null hypothesis. Thus, we can conclude that there is no relationship between the resident population and bike fatalities.

Overall Significance Even though many accidents happen in urban settings, there is not a positive nor significant correlation between population and fatality rates. This could be due to the independence of each of the data points- different cities. The low r-squared value implies that much of the varaition in pedicycling fatalites is due to factors other than population. Each city was built independently, so infrastructure and urban planning could be quite different. A large city could have a low fatality rate because the city could have been built in such a way that was friendly to bikers. Even though a larger city might have stronger risks (e.g. heavy traffic), they may have more funding to build better infrastructure that makes for a more friendly area to cyclists.

1.3.6 3.6 Pedicyclist Fatalities and Bikes Sales

If people more people are buying bikes sales, does that mean that total fatalities will increase because more people are on bikes? A linear regression model was used to test this relationship in the years 2007-2016.

Coding The table that combines all the data that have the commonality of year, and because this data came from different sources, some cells are more populated than others. Thus, only 2007-2016 had information for both bike sales and Total Fatalities. So, a new dataframe was created to support this analysis.

Visual

Hypothesis Test A test of regression slope will be used to determine if there is a relationship between total cycling fatalities and number of bikes sold in a year.

$$H_0 : \beta_0 = 0, \text{There is no relation between pedicyclist fatalities and bike sales}$$

$$H_A : \beta_0 \neq 0, \text{There exists a relation between pedicyclist fatalities and bike sales}$$

After running stats.linregress(), we find that the slope= 482.04, the r-squared is 0.26, and the p-value= 0.49. At $\alpha = 0.05$, we fail to reject the null hypothesis. Thus, we can conclude that there is no relationship between pedicyclist fatalities and bike sales in a given year.

Overall Significance From looking at the graph, one can see that there is very little relationship or similar behavior in bike sales and total fatalities through these 9 years. Even though there was a local maximum in bike sales for many Olympic years, there were not peaks in pedicyclists fatalaities in those years. Due to a low r-squared value, one cannot conclude that much of the bike fatalities are due to bike sales. In context, this could mean that the spikes in bike sales might not be from people new to cycling joining the market. It could be that already committed riders felt inspired,

and decided to purchase a new bike. These seasoned riders are likely to be quite familiar with safety precautions. So, the variation in total fatalities is due to many factors other than bike sales.

1.4 4. Linear Regression Model

From the significance tests and visuals shown above, it can be seen that age, gender, and time of day have an impact on bike fatalities. Given the constraints of the data and the lack of crossover, a single model cannot be made combining all of the predictors. However, there are three models that can be created that show these relationships.

Model Based on Age and Gender Fatality Rates

$$\hat{\text{Total Fatality Rate}} = 0.023 \text{ Age (Years)} + 0.46 \text{ Male Fatality Rate} + 0.73 \text{ Female Fatality Rate} - 0.06$$

$$R^2 = 0.99$$

With high correlation, this model shows that the total fatality rate is heavily explained by the age and the rates of males and females. From the coefficients in the model, one can say that for every 1 fatality increase in a million residents, this will make the total fatility rate rise by 0.46 and 0.73 for males and females respectively. For every increase in overall age group, the total increase in fatality rate would rise by 0.023.

Model Based on Male and Female Gender Populations

$$\hat{\text{Total Number of Fatalities}} = -4.43 \text{ Age (Years)} - 0.02 \text{ Male Population (thousands)} + 0.03 \text{ Female Population (th}$$

$$R^2 = 0.39$$

As shown before with a hypothesis test, the average fatality rate of males is significantly higher than that of females. With an increase of 1000 male residents, according to this model, will actually decrease the total number of fatalities, holding all else constant. In addition, with all else held constant, the increase in female population will cause in ancrease in total fatalities as seen with the positive coefficient. This is counter-intuitive considering that males were significantly more at risk for cycling accidents. The R^2 value in this model is 0.39, which does not imply a strong correlation. This means that 61% of the total variance in Total Number of Fatalities is explained by factors other than age and gender. As different as the fatality rates may be, this model shows that even with a significant increase in male population, that the total number of fatalities would not increase, holding all else constant.

Model Based on Time Frames

$$\text{Propotional difference in bike fatalities} = \frac{-4.25(0:00\text{-}2:59) - 7.25\ (3:00\text{-}5:59) + 0.75\ (6:00\text{-}8:59) - 1.25\ (9:00\text{-}11:59) - 1.25\ (12:00\text{-}14:59) + 1.75\ (15:00\text{-}17:59) + 9.75\ (18:00\text{-}20:59) + 1.75\ (21:00\text{-}23:59) + 12.25}{-4.25(0:00\text{-}2:59) - 7.25\ (3:00\text{-}5:59) + 0.75\ (6:00\text{-}8:59) - 1.25\ (9:00\text{-}11:59) - 1.25\ (12:00\text{-}14:59) + 1.75\ (15:00\text{-}17:59) + 9.75\ (18:00\text{-}20:59) + 1.75\ (21:00\text{-}23:59) + 12.25}$$

This model is meant to be used to assess difference in risk when comparing two timeframes. There are 2 inputs meant to be used for this model– Time Frame A and Time Frame B. Time Frame A will use the numerator and plug in a 1 in the numerator for that particular time frame. Similarly, Time Frame B is used in the denominator; plug in a 1 for that particular time frame. This will return a proportion that says that A is some amount times worse than B in terms of fatality rates. So, if A > B, then A is a more risky time frame, and vice versa. This model could be useful to assess comparative risk and make a more informed decision of what hour to bike in.

2 5. Conclusion

2.1 5.1 Summary of Findings

From collecting data, cleaning, and analyzing, we were able to gain insight into our objective and hypothesis. We had initially predicted that factors such as increase in bike sales, a summer Olympic year, time of day and week, population size, as well as age and gender demographics all have positive correlation with pedicyclist fatality rates. However, after visualizing data and running hypothesis tests, not all of these correlations were true.

Were there any significant differences regarding cycling between different calendar years? Some years were Olympic years, while others were not, and this caused local maxima on most of these years for annual bike sales. However, on average there was not a significant difference. In addition, bike sales and fatalities did not have a strong correlation. This may seem counter intuitive because with more bikes purchased, this might mean that there is probably more ridership on the roads, and this means that there are more chances for an accident to occur. From the data we collected, there was more we can see that there was fluctation in bike sales and fatalities, but there were no strong correlations that were unveiled when making signifcance tests.

Do age and gender have an influence on cylcling fatalities? If so, which age group and gender? Why is that? From significance testing, we concluded that males had significantly higher fatality rates. In addition, from the visuals, the fatality rates are significantly worse for men between the ages of 50-64. A reasonable explanation for this could be due either significantly more ridership among males or a significant difference in behavior for male cyclists. If females were cycling less, than this would inherently yield fewer fatalities per million residents. Some differences in behavior for male cyclists could be that they take fewer precautions while biking or electing to not wear a helmet, and this would lead to a higher likelihood of death during an accident.

Was there a specific time of day or a season that had an effect on cycling fatalities? No hypothesis test concluded that there were statistically significant differences in fatalites when comparing weekend and weekdays, time of day, or seasons within those time frames. The weekend and weekday comparison showed that there was no significant difference in accidents. While the ANOVA just barely failed to reject the null hypothesis test that all the time of days had equal fatlity, it is worthy to note that the time frame fataliteis were not uniform. The 18:00-20:59 timeframe had more deaths. In addition, the winter season did not always have the highest number of deaths, and in that time frame, it was much higher than the other seasons. This could potentially be associated with daylight hours getting shortened, but people are maintaining their daily routines. If people had had the routine of biking for work recreationally, they may have the desire to maintain it, even if it is getting darker at that hour. This darkness reduces visibility which can lead to an accident.

2.2 5.2 Future Analysis

To further analyze this dataset, more comparisons with different predictors could be done to potentially discover predictors that are relevant to total fatality rates in cycling. Some of these relationships could be analysing what types of vehicles are responsible for causing accidents, or which cities have a higher fatality rate than one would expect for a city of their size. This analysis would unveil which particular cities are above average and below average for fatalities among pedalcyclists. To analyze these associations further, it would be necessary to use data for be more specific. While this summary data was able to provide some level of insight, using data that is more specific (e.g. a list of all accidents with all of the information regarding these factors) would be able to perform analyses that are more specific. In addition, with a dataset that is more granular, it would be more possible to make a single model with all predictors rather than having seperate models dealing with different aspects of cycling.

2.3 5.3 Acknowledgements

Thank you to all the TA's who helped answer all of our questions on Campuswire and were available during office hours. A special thank you to Professor Mimno for sacrificing his dignity for our memory. We had some issues using github, so we had to improvise to share our progress and materials. The final of our source code/data cleaning has been uploaded and can be accessed below:

https://github.coecis.cornell.edu/leh224/INFO-2950-Final-Project.git

Thank you!