Date: August 5, 2023

Team: Meral Basit, Chad Jasso, Quyen Ha

Topic: Bike Safety by the Numbers: Bicyclist Fatality Analysis

I. INTRODUCTION

A. Data Source

Since 1975, the National Highway Traffic Safety Administration ("NHTSA") annually publishes the Fatality Analysis Reporting System ("FARS"). FARS is a census of fatal crashes with a set of files documenting all qualifying fatalities that occurred in the U.S.

To qualify as a FARS case, the crash must involve a vehicle or pedestrian traveling on a traffic way open to the public and must have resulted in the death of the victim within 30 days of the crash.[1] Along with various datasets, FARS also provides a comprehensive user's manual, with variable and code definitions.

B. Primary and Secondary Data Description

Our primary data is *pbtype*—this data contains information about crashes between motor vehicles and pedestrians, people on personal conveyances (e.g., a truck driver using their truck for personal purposes when they are not on-duty)[2], and bicyclists. There is one record for each pedestrian, bicyclist, or person on a personal conveyance.

Our secondary data is *accident*—this data contains information about crash characteristics and environmental conditions at the time of the crash. There is one record per crash.

We assume that police filed reports for a representative majority of crashes involving bicyclists within the United States. Our team is aware of the social factors that may cause people to not report a crash to the police, and for a more exhaustive analysis, we would want to corroborate the findings with other data sources.

[1] Fatality Analysis Reporting System Analytical User's Manual, 1975-2021. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813417.

[2] https://www.fmcsa.dot.gov/regulations/hours-service/personal-conveyance ("Personal conveyance is the movement of a commercial motor vehicle (CMV) for personal use while off-duty. A driver may record time operating a CMV for personal conveyance as off-duty only when the driver is relieved from work and all responsibility for performing work by the motor carrier.").

II. DATA BUILD

A. Primary Data Build

Initially, we considered the period of analysis to be from 2014 to 2021, as 2014 is the first year in which *pbtype* is available, and 2021 is the latest available year of data. However, our initial data exploration revealed that the 2014 data lacked variables that are (1) included in all subsequent years of data and (2) relevant to our analysis. Therefore, we dropped the 2014 data.

Further, we found that there was a large-scale adoption of new coding practices from 2016 onward. Therefore, we ultimately dropped the 2015 data and decided to modify the relevant period to be 2016 to 2021.

Since our analysis only focuses on fatalities associated with bicyclists, we removed all observations where the variable PBPTYPENAME is not “Bicyclist.” PBTYPENAME describes the role of the person involved in the crash but may not denote the actual deceased victim. In our report, we assumed that PBTYPENAME denotes the deceased individual. In addition, we also removed variables that only pertain to pedestrians and people on personal conveyances.[3]

At the end, we had 27 remaining variables and 5,469 observations representing individuals involved in bicyclist fatalities from 2016 to 2021.

B. Secondary Data Build

We chose the same time period for *accident* as we did for *pbtype*, as *accident* is supplementing *pbtype*. As the FARS data manual instructed, we used ST_CASE to merge pbtype with *accident* for each year.[4]

III. DEMOGRAPHIC OF INDIVIDUALS INVOLVED IN BICYCLIST FATALITIES

The demographics we decided to explore were age and sex. Our hypothesis was that there would be an increase in the number of middle-aged individuals who were involved in fatal crashes, which we defined as between ages 40-60, for both male and female bicyclists.

In terms of data quality, 1.4% of entries were marked with one of two codes for “unknown” ages.[5] Additionally, less than one percent of entries were marked with one of three categories denoting an “unknown” sex.[6] After further exploration, we decided to exclude entries with unknown sex and/or age from this analysis. There is a risk that the incidents associated with

[3] *See* code included in https://github.com/UC-Berkeley-I-School/Project2_Basit_Jasso_Ha for our data cleaning process.

[4] The merge is a many-to-one merge—*pbtype* is unique by each person involved in each crash (i.e., by PER_NO, ST_CASE, and YEAR), whereas *accident* is unique by crash and year (i.e., by ST_CASE and YEAR).

[5] Unknown Age Codes: 998 (“Not Reported”; 1.1% of entries), 999 (“Unknown” from 2016-2017, “Reported as Unknown” from 2018-Later; 0.29% of entries).

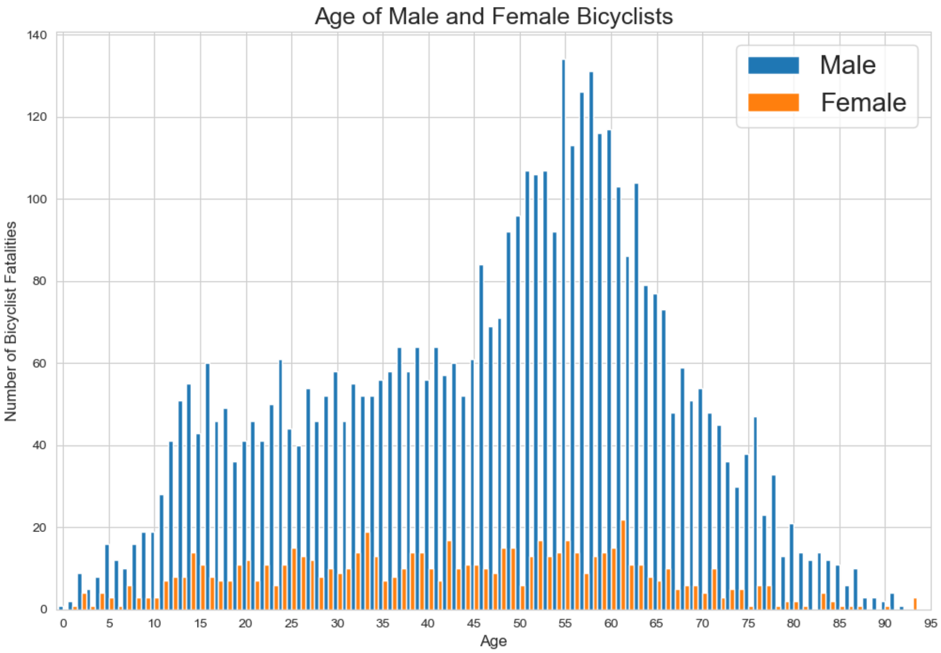
[6] “Unknown” categories that were not included in analysis: Not Reported (0.55% of entries), Reported as Unknown (0.13% of entries), Unknown (0.09% of entries).

unknown sexes and ages are more catastrophic (possibly resulting in a bicyclist whose demographics are hard to discern), which our team is aware of, and will investigate further at a later time.

As **Figure 1** demonstrates, our null hypothesis was proven false. We did see an increase in the number of male cyclists involved in accidents, roughly from ages 45-65. However, the number of female cyclists involved in accidents remained relatively stable through ages 25-60.

In general, a far greater number of male cyclists were involved in accidents with motor vehicles than female cyclists (4635 males : 738 females). This disparity could be a result of a different base number of male and female bicyclists, or from a difference in rates of police documentation between the sexes. We recommend examining accident data from a different source, as well as data on male and female bicyclists rates, to further explore these findings.

Figure 1
Number of Bicyclist Fatalities by Sex and Age



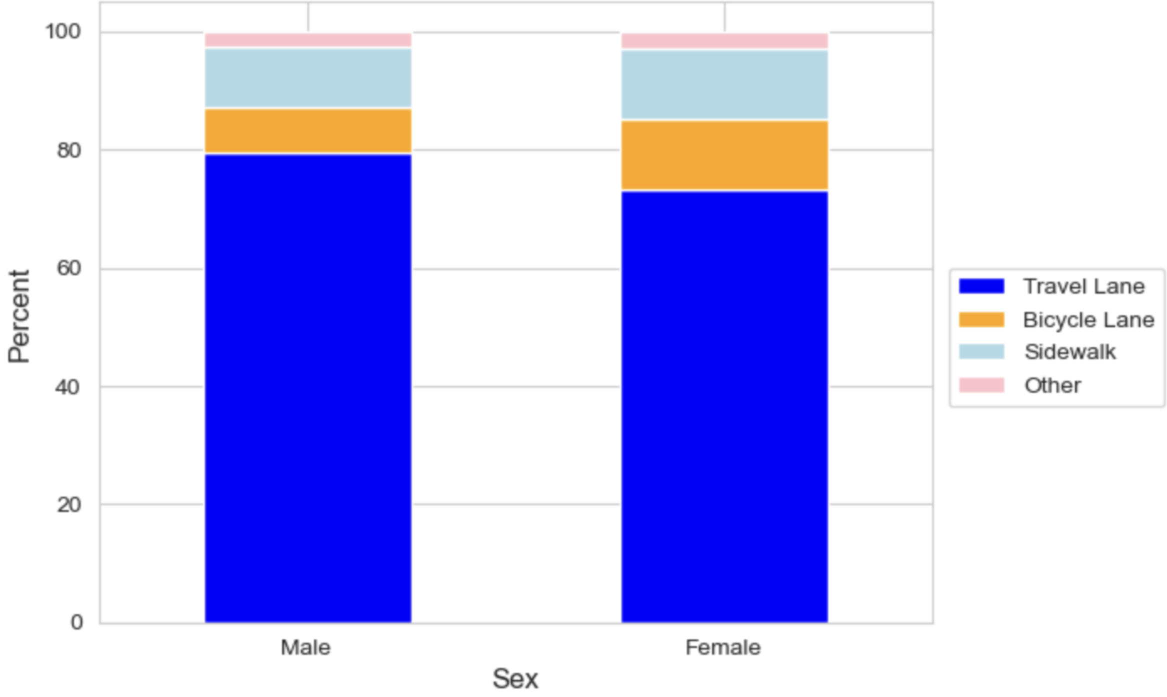
Differences in distribution within each categorical variable between males and females were also explored.[7] For example, 80% of females were involved in a crash where a crosswalk was not noted, compared to 82% of males. Our null hypothesis was that distributions within each variable would differ by no more than 5% between males and females, for all variables.

The distribution within most variables varied minorly between the sexes, with an average difference in distribution of 0.46%. However, there was a substantially larger difference in the distribution of bike positions (variable BIKEPOS) between the sexes, as shown in **Figure 2**.

[7] Columns 6 through 27 were explored in this analysis, as columns 1-5 do not contain data pertinent to crash characteristics.

Upon first contact with a motor vehicle, 6.5% more males were in a travel lane than females. In contrast, 4.55% more females were in a bike lane than males. The difference of 6.5% between males and females within the variable BIKEPOS proved our null hypothesis false. This large difference in bike position upon impact between males and females indicates a need for further exploration into this topic.

Figure 2
Bike Position Distribution



IV. ENVIRONMENTAL FACTORS OF BICYCLIST FATALITIES

In this section, we considered the following environmental factors: light (i.e., the level of light that existed at the time of the crash) and weather (i.e., the prevailing atmospheric conditions that existed at the time of the crash) conditions, both of which are available in the supplemental *accident* data.[8]

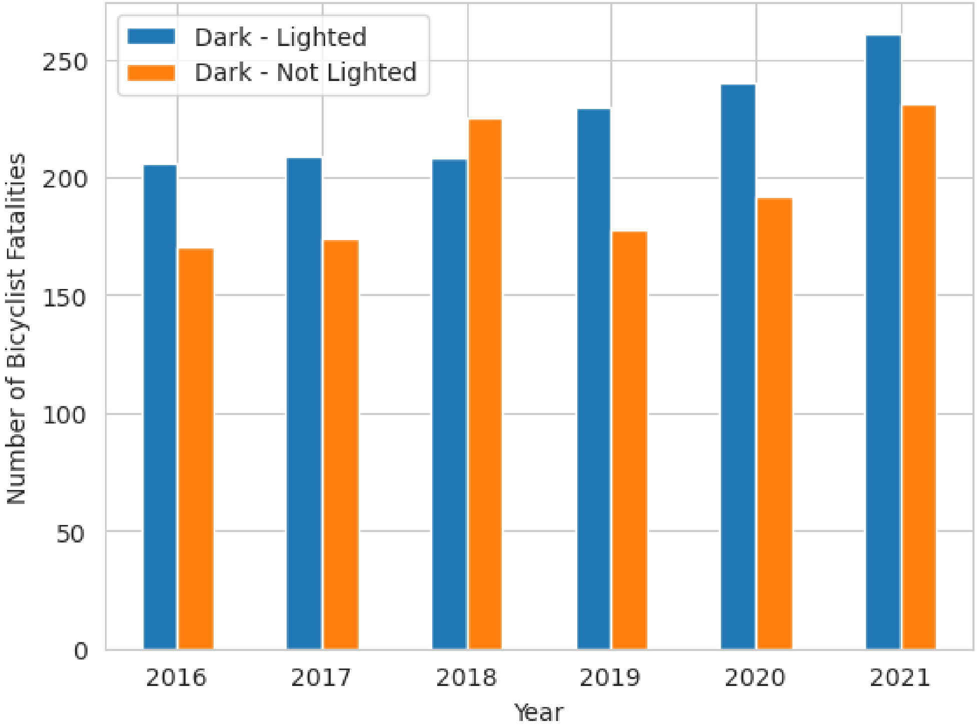
In particular, our first null hypothesis is that at night, there would be more bicyclist fatalities in unlighted streets compared to lighted ones.[9] However, as shown in **Figure 3**, the number of annual bicyclist fatalities is typically higher for lighted streets (except in 2018) compared to unlighted streets.

However, this analysis has many limitations, and we believe that expanding to other data sources is necessary to reach a more definitive conclusion. For example, **Figure 3** may be skewed if

[8] The variable for light condition is LGT_CONDNAME. The variable for weather conditions is WEATHERNAME.
[9] We assume lighted streets at night are associated with "Dark - Lighted", and unlighted streets at night are associated with "Dark - Not Lighted" for variable LGT_CONDNAME.

locations with more bicyclists (such as Portland or San Francisco)[10] also have more lighted streets than unlighted ones. To be more comprehensive, we would need to obtain data pertaining to the prevalence of lighted versus unlighted streets for each location (e.g., a city, county, or census area) and normalize the FARS crash data with this supplemental data.

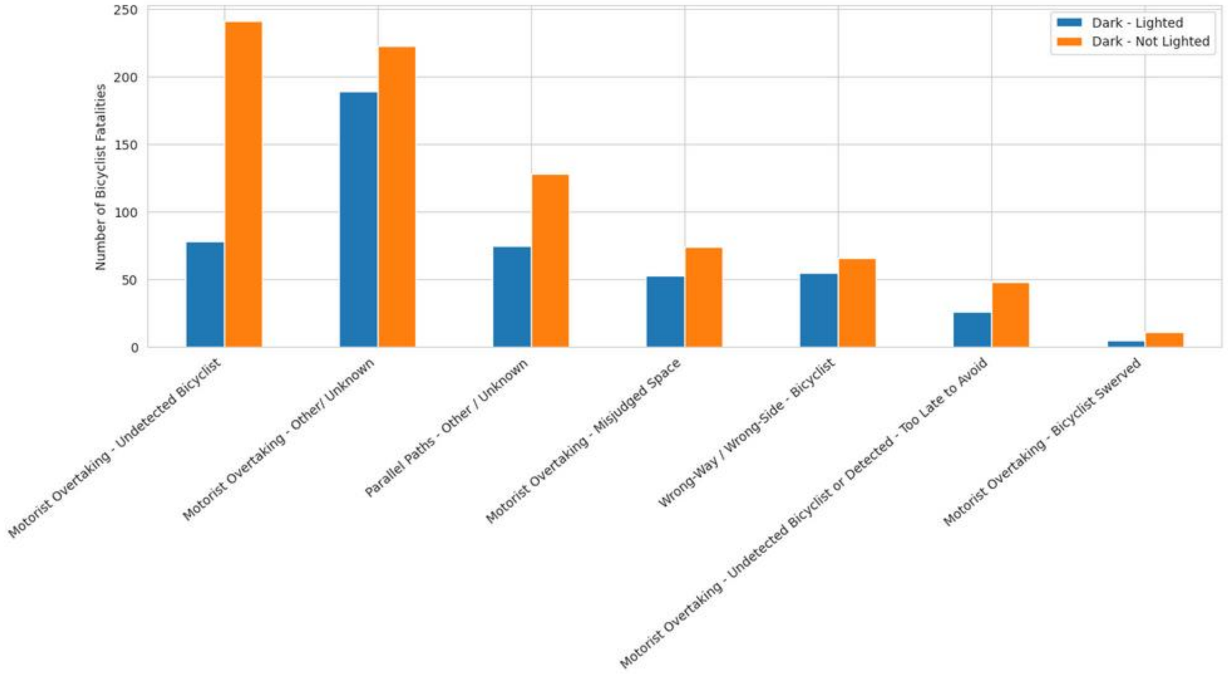
Figure 3
Number of Bicyclist Fatalities by Lighting Group



We further investigated crash types in relation to lighting groups. We focused on crash types in which the number of bicyclist fatalities associated with unlighted streets are *higher* than lighted streets.[11] As **Figure 4** demonstrates, *all* crash type categories in which a motorist was overtaking a bicyclist have higher bicyclist fatalities in unlighted streets. This finding suggests that the lack of visibility in unlighted streets at night may positively correlate with fatal motorist overtakings of bicyclists.

[10] https://anytimeestimate.com/research/most-bike-friendly-cities-us-2022/.
[11] We exclude crash types for which the difference between the number of bicyclist fatalities associated with unlighted streets and lighted streets is lower than five.

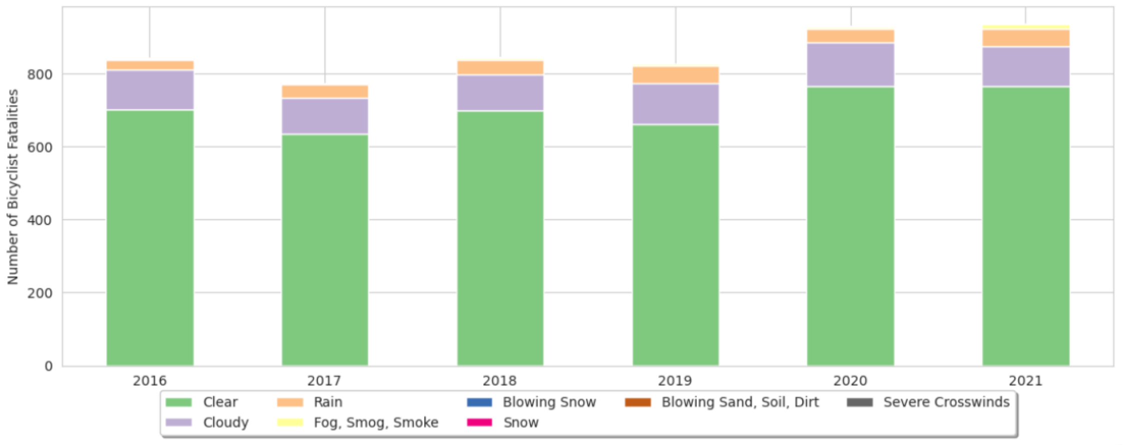
Figure 4
Number of Bicyclist Fatalities by Lighting Group and Crash Type
Limited to where Fatalities on Unlighted Streets are Higher than Lighted Streets



We also hypothesized that most bicyclist fatalities occurred during adverse weather conditions, which include the following conditions coded in the FARS data: Blowing Sand, Soil, Dirt; Blowing Snow; Fog, Smog, Smoke; Rain, Severe Crosswinds, and Snow. **Figure 5** shows that our hypothesis is false, as most bicyclist fatalities occurred during the Clear weather condition. The second most common weather condition is Cloudy.[12] To expand this analysis, we propose getting more granular weather data (e.g., specific levels of rain, snow, etc.) for a given time and location and using such data to investigate whether certain weather-location combinations are more likely to lead to bicyclist fatalities.

[12] In this analysis, we excluded unknown weather conditions (i.e., WEATHERNAME does not equal the following values: Not Reported, Other, Reported as Unknown, Unknown), which accounts for ~5.6% of all observations.

Figure 5
Number of Annual Bicyclist Fatalities by Weather Condition



V. CHARACTERISTICS OF BICYCLIST FATALITIES

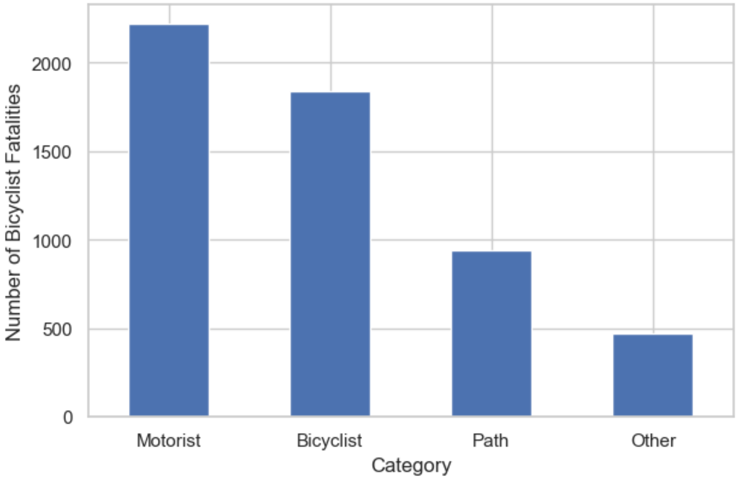
In this section, we explore potential correlations among the factors contributing to bicycle accidents. Variables available to us included the crash type, crash location, position of the bicyclist, direction of travel, and crash group for bicyclists involved in the incident. With these variables, we aimed to explore whether crash type, location, bicyclist's position, initial direction of travel, and crash group significantly contribute to bicycle accidents, and what the relationships between these factors are. Our initial hypothesis was that motorists were the leading cause of bicyclist fatalities and that those fatalities were most likely to occur during left turns across lanes.

At the outset of our analysis, we recognized that the crash group variable, although valuable in its specificity, required some refinement to facilitate effective quantification. Therefore, we conducted a data cleaning process to combine certain variables. Initially comprising 78 distinct crash group categories, we streamlined them into four broader groups (1) Incidents caused by the motorist; (2) Incidents caused by the cyclist; (3) Incidents caused by path-related factors; and (4) “Other” category incorporating all remaining cases.:[13]

In-line with our initial hypothesis, **Figure 6** reveals that most bicyclist fatalities are attributable to crashes caused by motorists.

[13] *See* code included in https://github.com/UC-Berkeley-I-School/Project2_Basit_Jasso_Ha for how we streamlined the crash group categorization.

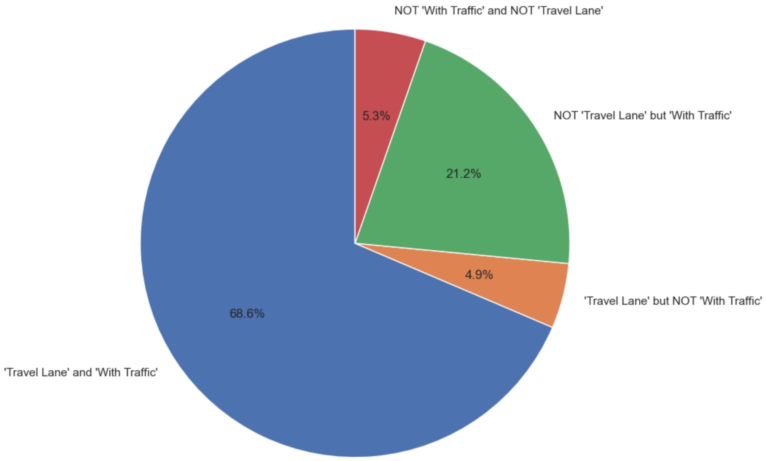
Figure 6
Number of Bicyclist Fatalities by Crash Group Category



However, among the motorist incidents, no specific crash group determinant was evident, with the largest sub-category remaining undefined (“Other/Unknown”).

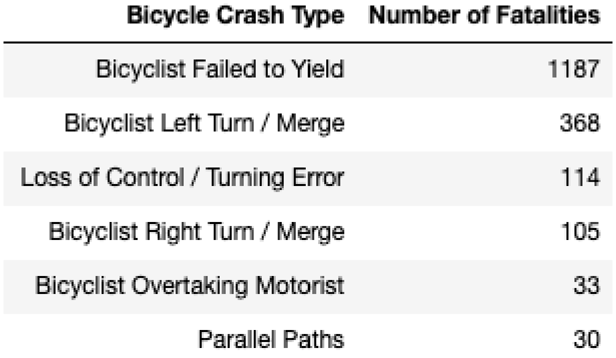
Consistent with our expectations, as **Figure 7** demonstrates, most of these incidents took place in travel lanes rather than bike lanes and involved bicyclists traveling in the same direction as traffic, which aligns with the common scenario for bicyclists in the United States.

Figure 7
Percentage of Bicyclist Fatalists Caused by Motorists by Direction of Traffic



While most bicyclist fatalities were caused by motorists, we also investigated bicyclist-caused fatalities. Focusing on bicyclist-caused fatalities allows us to identify specific factors and behaviors that may contribute to these incidents and propose targeted measures to address them effectively in future analyses. As **Figure 8** illustrates, amongst bicyclist-caused fatalities, the highest frequency of occurrence was associated with bicyclists failing to yield.

Figure 8
Number of Bicyclist Fatalities by Bicycle Crash Group



Bicycle Crash Type	Number of Fatalities
Bicyclist Failed to Yield	1187
Bicyclist Left Turn / Merge	368
Loss of Control / Turning Error	114
Bicyclist Right Turn / Merge	105
Bicyclist Overtaking Motorist	33
Parallel Paths	30

VI. CONCLUSIONS

Given the limited scope of this report, we are not making any specific recommendations to improve bike safety based on any analysis. Instead, our team recommends that we expand the scope of this report and specifically dedicate more resources into understanding the following topics:

- Higher proportion of middle-aged males in crashes.
- Discrepancy in bike position upon impact between males and females.
- The lack of visibility in unlighted streets at night impacts fatal motorist overtakings of bicyclists.
- The primary contributors to bicyclist fatalities were drivers operating motor vehicles.
- In situations where the bicyclist was found responsible, the predominant factor leading to fatalities was their failure to yield.

In conclusion, this research project highlights the essential areas for future analyses and underscores the need to broaden the scope of our inquiry into bicyclist safety. While specific recommendations are withheld due to the report's limited focus, it is evident that a deeper exploration of topics such as the involvement of middle-aged males in crashes, sex-related differences in bike impact positions, the correlation between unlighted streets and fatal motorist overtakings, and the pivotal role of drivers in bicyclist fatalities is imperative. Additionally, understanding the factors contributing to fatalities where bicyclists are found responsible, particularly their failure to yield, warrants further attention. By dedicating resources to these critical aspects, we can pave the way for targeted interventions and strategies that improve bicyclist safety.