|  |  |
| --- | --- |
| **Date:** | July 23, 2023 |
| Re: | Research Proposal for Project 2 |
| Team: | Meral Bassit, Chad Jasso, Quyen Ha |
| Github: | https://github.com/UC-Berkeley-I-School/Project2\_Basit\_Jasso\_Ha |

# Primary and supplement Dataset

## Primary and Supplemental Data Description

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]](#footnote-1)

We will use a FARS dataset named “pbtype” as the primary dataset, which includes FARS data related to pedestrians, bicyclists, and people on personal conveyances. We will further use the annual resident population by state from the Federal Reserve Bank of St. Louis (“FRED”) as the supplemental dataset to normalize state population in the primary pbtype data.[[2]](#footnote-2)

## Initial Primary Data Visualization

**Figure 1: Number of People Involved in Fatal Crashes by Year[[3]](#footnote-3)**

A graph of a number of years

Description automatically generated

**Figure 2: Number of People Involved in Fatal Crashes by State**

A graph with blue lines

Description automatically generated

## Primary Data’s Variables to Explore

The following columns will allow us to obtain meaningful observations regarding the characteristics of the cyclists who are involved in accidents:

1. PBAGE/PBAGENAME: represent the age of the person involved in the incident.
2. PBSEX/PBSEXNAME: represent the gender code of the person involved in the incident.

The following columns will allow us to better understand the environment where the accidents occurred:

1. STATE/STATENAME: represent the state where the incident occurred.
2. PBCWALK/PBCWALKNAME: represent whether a marked crosswalk was present at the incident.
3. PBSWALK/PBSWALKNAME: represent whether a sidewalk was present at the incident.
4. PBSZONE/PBSZONENAME: represent whether the incident occurred in a school zone.

The following columns will allow us to gain insights into the attributes of the bicycle accidents:

1. BIKECTYPE/BIKECTYPENAME: represent the crash type.
2. BIKELOC/BIKELOCNAME: represent the crash location.
3. BIKEPOS/BIKEPOSNAME: represent the bicyclist’s position of the bicyclist.
4. BIKEDIR/BIKEDIRNAME: represent the bicyclist’s initial direction of.
5. BIKECGP/BIKECGPNAME: represent general groupings of BIKECTYPE/BIKECTYPENAME.

# Suggested Report Outline

1. Introduction & Hypothesis
   1. Hypothesis: there is a commonality for fatal crashes involving bicyclists.
2. Data Description & Build
   1. Description of the primary and supplementary data.
   2. Description of data cleaning, build, and assumptions.
3. Analysis
   1. Characteristics of bicyclists/environment/crashes.
4. Conclusion
   1. Conclusion about hypothesis, and identify areas for improvements.

1. Fatality Analysis Reporting System Analytical User’s Manual, 1975-2021. <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813417>. [↑](#footnote-ref-1)
2. *See e.g.,* [https://fred.stlouisfed.org/release/tables?rid=118&eid=259194&od=2016-01-01#](https://fred.stlouisfed.org/release/tables?rid=118&eid=259194&od=2016-01-01). [↑](#footnote-ref-2)
3. In this section, people include pedestrian, bicyclists, and people on personal conveyances who were involved in a FARS qualifying crash. [↑](#footnote-ref-3)