Bicycles Accident Analysis

by Michael Rusin

102 people

... are killed in a car accident
per day!*

* https://en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year

Introduction (business problem)

This report will try to analyze the best location and time for cycling activities. Specifically, this project will target stakeholders interested in **cycling activities** such as individual cyclists, cycling communities, and companies/sponsors/event organizers of cycling activities.

During a pandemic, many people choose cycling as an alternative to sports. Cycling is considered the safest way to exercise because of minimal contact with other people. There are many accidents involving cyclists. Cyclists also need protection and a sense of security while on the road. Cycling is not only interpreted as transportation activity, but also sports and recreational activities. When an accident occurs, car drivers are still protected by car frames and car safety technology in comparison. So, the chances of surviving or being injured are still relatively low compared to cyclists. Cyclists are only protected by wearing helmets on their heads. When an accident occurs, their bodies, feet, and hands have the potential to be injured.

This project will assist the Seattle Department of Transportation (SDOT) to provide different traffic signs in accident-prone areas for cyclists. This project will also help the cyclist community[*1] like **Cascade Bicycle Club, COGS (Cyclists Of Greater Seattle), Brake the Cycle, etc.** to find out the right track and time to hold a cycling event.

Many events are held by many cyclist communities, like Cascade Bicycle Club, for example. This club hosts **several major riding events**[*2] every year including Chilly Hilly, Seattle Bike-n-Brews, Ride for Major Taylor, Flying Wheels Summer Century, Woodinville Wine Ride, Seattle Night Ride, the Red-Bell 100, Seattle to Portland (STP), Ride from Seattle to Vancouver and Party (RSVP), Ride Around Washington (RAW), High Pass Challenge (HPC), and Kitsap Color Classic (KCC).

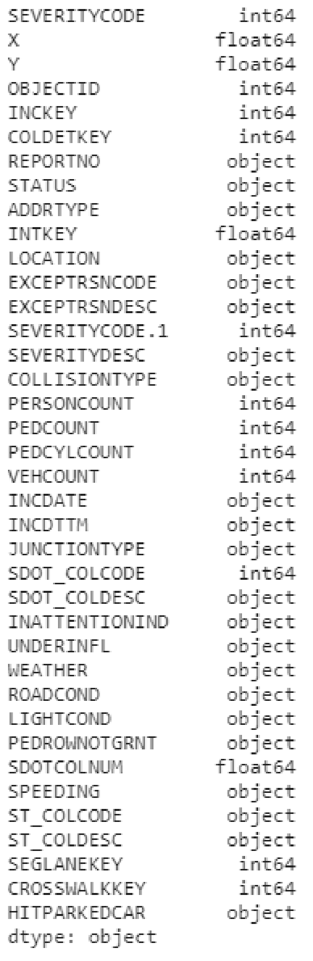
This project can help **companies, sponsors, and event organizers** to create **safe cycling events** for all participants.

[*1] http://wabikes.org/growing-bicycling/resources/bike-clubs-in-washington-state/

[*2] https://en.wikipedia.org/wiki/Cascade_Bicycle_Club#Major_events

Data understanding

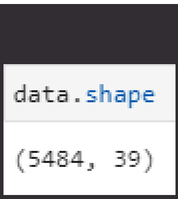
The original dataset can be found here:
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv



```
SEVERITYCODE          int64
X                   float64
Y                   float64
OBJECTID              int64
INCKEY                int64
COLDETKEY             int64
REPORTNO             object
STATUS               object
ADDRTYPE             object
INTKEY              float64
LOCATION             object
EXCEPTRSNCODE        object
EXCEPTRSNDESC        object
SEVERITYCODE.1        int64
SEVERITYDESC         object
COLLISIONTYPE        object
PERSONCOUNT           int64
PEDCOUNT              int64
PEDCYLCOUNT           int64
VEHCOUNT              int64
INCDATE              object
INCDTTM              object
JUNCTIONTYPE         object
SDOT_COLCODE          int64
SDOT_COLDESC         object
INATTENTIONIND       object
UNDERINFL            object
WEATHER              object
ROADCOND             object
LIGHTCOND            object
PEDROWNOTGRNT        object
SDOTCOLNUM          float64
SPEEDING             object
ST_COLCODE           object
ST_COLDESC           object
SEGLANEKEY            int64
CROSSWALKKEY          int64
HITPARKEDCAR         object
dtype: object
```

Based on definition of our problem, features or columns that will influence our analysis are:
1. **Location**: Latitude (X), Longitude (Y), Address Type (ADDRTYPE)
2. **Severity**: A code that corresponds to the severity of the collision (SEVERITYCODE), a detailed description of the severity of the collision (SEVERITYDESC)
3. **Person Count**: Total number of people involved (PERSONCOUNT), number of bicycles involved in the collision (PEDCYLCOUNT)
4. **Date**: The date and time of the incident (INCDTTM)
5. **Condition**: Description of the weather conditions (WEATHER), condition of the road (ROADCOND), light conditions during the collision (LIGHTCOND)

Conditional Selection — Only show data that bicycles involved in the collision
In the explanation of the problem above, we will help solve the problem for cyclists. We limit data on car accidents involving cyclists. So the PEDCYLCOUNT column must be greater than zero.



```
data.shape

(5484, 39)
```

DataFrame Shape
The dataset used is (5484 rows, 38 columns). Not all columns will be used, will be selected according to the data description above.

DataFrame Data Type and Missing Values
In the description below, we will know the data types and missing values and their presentations.

Methodology

1. Features Selection
Not all features are used for analysis in this project. Thus, only some data is displayed and analyzed.

2. Handling Missing Values
Missing values will interfere with the prediction and analysis results. So, we need to handle the missing values by deleting them or filling them in. If there are not too many missing values, we can choose the option to delete them.

3. Handling Duplicates Values
Duplicate values will also interfere with the analysis and prediction results. First, we need to detect the number of duplicate values in the dataset. Next, these duplicate values need to be removed to make the dataset cleaner.

4. Convert 'INCDTTM' Column to Datetime Type
'ICDDTM' Column needs to be changed in the DateTime type. Because by converting it to a DateTime type, we can extract hour, day, month, and year data. These data can help us to analyze data more deeply.

5. Exploratory Data Analysis (EDA)
After cleaning the data, we can run the exploratory data analysis. The analysis framework follows the problem we have defined, namely finding the best time and location for cycling activities.
First, the data will be explored and analyzed based on data related to time, such as the hour, day, month, year, and weather. The data is visualized to get an overview of the best time to hold a cycling event. Second, looking for an overview of the conditions for the best place to hold a cycling event. The data visualized include light conditions, road conditions, and address types.

6. Model Building
The machine learning model used in this project is logistic regression. Why use logistic regression? First, the data is binary. Second, we need probabilistic results to find out the time and place conditions that are most likely to cause injury collision. Before model building, the data will be encoded using the one-hot encoder and split into training and testing data.