Capstone Project 1: Data Wrangling

1. Data Collection and Importing

The dataset was loaded as a *.csv* file. Since the data was originally collected using an API, the column headers were specified correctly with no additional parsing being required. Once the data was loaded into a dataframe, the size was verified as ***212,760 rows*** and ***40 columns***.

2. Data Cleaning

a. **Selecting initial variables**: Due to a large number of initial columns, the dataset needed to be filtered. Univariate analysis was conducted in order to determine these variables. More details regarding the analysis can be found in this jupyter notebook. Here is a list of such variables and the corresponding reasons to remove them:

   i. **Too many unique categories:** Variables such as *OBJECTID*, *INCKEY*, *COLDETKEY*, and *REPORTNO* contained unique identification numbers for each incident, essentially representing categorical data. Due to 212,760 unrelated categories being present, these features would not provide any useful information. The variables for lane segment in which the collision occurred (*LANESEGKEY*) and specific crosswalk (*CROSSWALKKEY*) contained 2084 and 2284 unique categories respectively which would be difficult to analyze. Also, both variables contained a '0' category which contained over 200,000 of the instances making them undesirable for predicting different categories of severity.

   ii. **Too many missing values:** This is the largest category of variables that were excluded were easily identified using the *.info()* built-in Python function which displays a list of the number of 'non-null' values for each column . The variable *INTKEY* stored the collision intersection but contained only 68,578 entries. Similarly *EXCEPTRSNCODE* and *EXCEPTRSNDESC* both contained less than 100,000 values. Other variables such as whether the driver was attentive (*INATTENTIONIND*) and whether the driver was speeding (*SPEEDING*) were meaningful features but contained only 29,116 and 9,492 values respectively. *PEDROWNOTGRNT* showed whether or not the pedestrian right of way was granted but contained only 4983 values, while *SDOTCOLNUM* which contained specific numbers assigned to each collision by SDOT (Seattle Department of Transportation) contained just 127,205 entries.

   iii. **Redundant variables:** The *LOCATION* variable contained specific addresses but latitude and longitude information provided more accurate location data. *INCDATE* contained date information but *INCDTTM* contained both date and time so the former was excluded. *SDOT_COLCODE* contained collision codes representing types of collisions assigned by SDOT. *SDOT_COLDESC* contained the same information but had text descriptions making it more useful. Similarly,

ST_COLDESC was chosen over *ST_COLCODE*. Note that 'ST' stands for state as opposed to 'SDOT'. For the same reason, *SEVERITYCODE* was excluded while *SEVERITYDESC* was retained as the target variable for our prediction.

Note that this was the initial feature selection exercise to remove variables that provided little information. More variables could be excluded following further analysis.

b. **Dropping variables and renaming columns**: The remaining variables were renamed to more readable names. For example, *INCDTTM* was renamed to 'Incident Data and Time' while *VEHCOUNT* was renamed to 'Number of Vehicles Involved'.

c. **Handling missing values**: 11 columns contained null values which were visualized as a matrix using Python's *msno* library. Variables such as weather and road condition had missing values for the exact same rows while latitude and longitude had a similar connection. Other variables such as junction type had a unique set of rows containing null values. Below, different methods for handling missing values are described.
    i. **Dropping rows for variables containing few missing values:** A total of 7,366 entries were missing for both latitude and longitude. Since this was small compared to the initial 212,760 instances, simply dropping these rows would have little impact on the dataset. For address type, there were fewer missing values but it turned out that those instances overlapped with the same missing rows as the coordinates. Hence it was safe to remove these rows as well.
    ii. **Replacing missing entries with 'unknown' category:** For weather, light conditions, road conditions, DUI, junction type and collision type, the missing entries were converted to an 'unknown' category since that category already existed for these columns. However, this caused the 'unknown' categories to now have a significant number of instances. In some cases, they contained greater than 30,000 instances which amounts to more than 15% of the instances for a variable. In order to handle these 'unknown' values, more detailed analysis as well as multivariate analysis needs to be done with respect to the severity variable. This will be covered in the exploratory data analysis section.

d. Dealing with outliers: The numerical variables were analyzed to determine if there were any obvious outliers as a result of mistakes made during data collection. However, they were no such variables. Any extreme values looked to be naturally present and could be useful for further analysis.

3. Timeseries Variables Extraction

The date and time variable was split into five different variables - Year, Month, Date, Hour, Minute - for time series analysis.

The dataset following the data wrangling process contained ***205,394 rows and 27 columns***.