by CommunistSquared, Wikimedia Commons

Analysis of SDOT Accidents in Cold Conditions (Applied Data Science Capstone by IBM/Coursera)

Introduction

In this project, we will be digging deeper into the accident data provided to us from the Seattle Department of Transportation (SDOT). The question we aim to answer is this: ***What areas in Seattle are most prone to accidents when affected by snow and ice?***

We are confident that our findings would be of good use to the SDOT and its associated government agencies (including, but not limited to the Seattle Police Department). For one thing, it would result in more efficient allocation of their resources to prevent injuries and damage to property caused by accidents in these areas.

As an effect, the general populace of the city of Seattle would be safer, knowing those who watch over them have more information at their disposal.

My Jupyter Notebook is found here.

Data

You can find the Dataset by Clicking here. You can also find the Metadata by Clicking here

The data we have on hand is quite extensive, providing us with relevant attributes such as the location and time of the accident, as well as the consequent weather and lighting conditions when the event occurred. There are also details indicating how many pedestrians were involved, or if one or more cars that were affected were parked or not.

Breakdown

For this particular case, the Dependent Variable to predict will be the severity of the accident, or **SEVERITYCODE**. Note that the Severity Codes in the provided CSV file are only limited to two values: 1 and 2. According to the supplementary Metadata file, **SEVERITYCODE 1** corresponds to **Property Damage** and **2** corresponds to **Injury**. Thus, the model we will create will be geared towards ***predicting the likelihood of a potential accident to lead to Property Damage or Injury.***

NOTE: There were other Severity Codes disclosed on the Metadata: 0 (Unknown), 2b (Serious Injury), and 3 (Fatality). We believe these were omitted because the dataset already had thousands of rows to work on without these Severity Codes. 0 (Unknown) was most likely omitted because of its potential to confuse the model. Finally, we believe 2 (Injury) would be of utmost importance because we don't want to wait for serious injuries or fatalities to happen before we work on data. In other words if we could prevent injury we could definitely prevent worse things from happening.

We have chosen to work with the following Attributes:

- **X, Y** - Coordinates of the general location of the collision.
- **JUNCTIONTYPE** - Category of junction at which collision took place
- **WEATHER** - A description of the weather conditions during the time of the collision.
- **ROADCOND** - The condition of the road during the collision.
- **LIGHTCOND** - The light conditions during the collision.
- **SPEEDING** - Whether or not speeding was a factor in the collision.

These descriptions were gathered from the supplementary Metadata file.

These are the columns we dropped, due to their lesser relevance, string properties, or because they are details of the accident after the fact: **LOCATION, PERSONCOUNT, VEHCOUNT, INCDATE, INCDTTM, OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, ADDRTYPE, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, Duplicate SEVERITYCODE, COLLISIONTYPE, PEDCOUNT, PEDCYLCOUNT, SDOT_COLCODE, SDOT_COLDESC, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR, SEVERITYDESC**

Data Selection/Cleaning

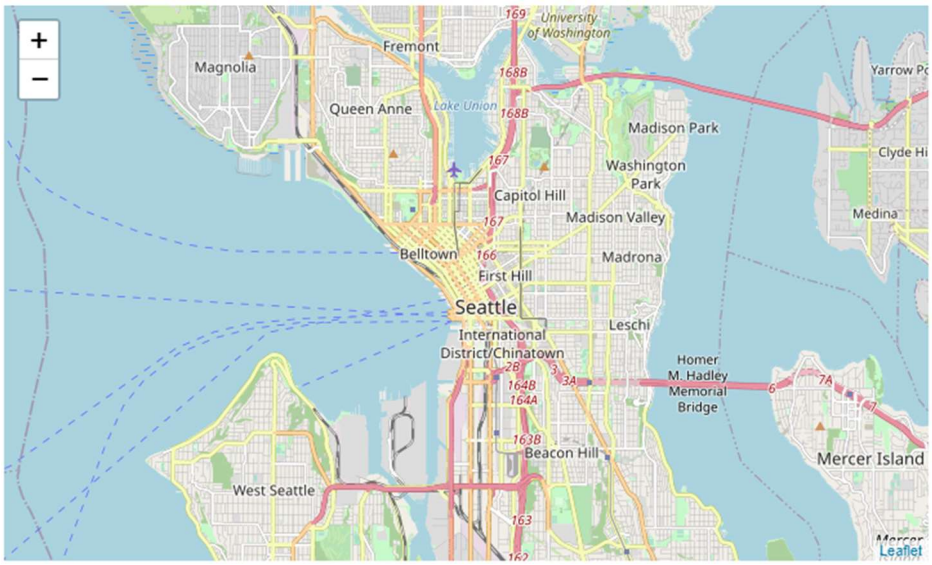
After loading the CSV we loaded the precise columns that we need, as mentioned earlier. These will be in a dedicated dataframe for us to work with independent of the raw data straight from the CSV file. In the dedicated datafame df, we converted NaN values in Column SPEEDING to 'N'. After that, we dropped all other Rows that had the values NaN, 'Unknown' and 'Other'. From there we drew out all rows with ROADCOND of 'Snow/Slush' and 'Ice'. This takes us from a dataset of more than 100000 rows to a little less than 1900 rows.

Methodology / Results

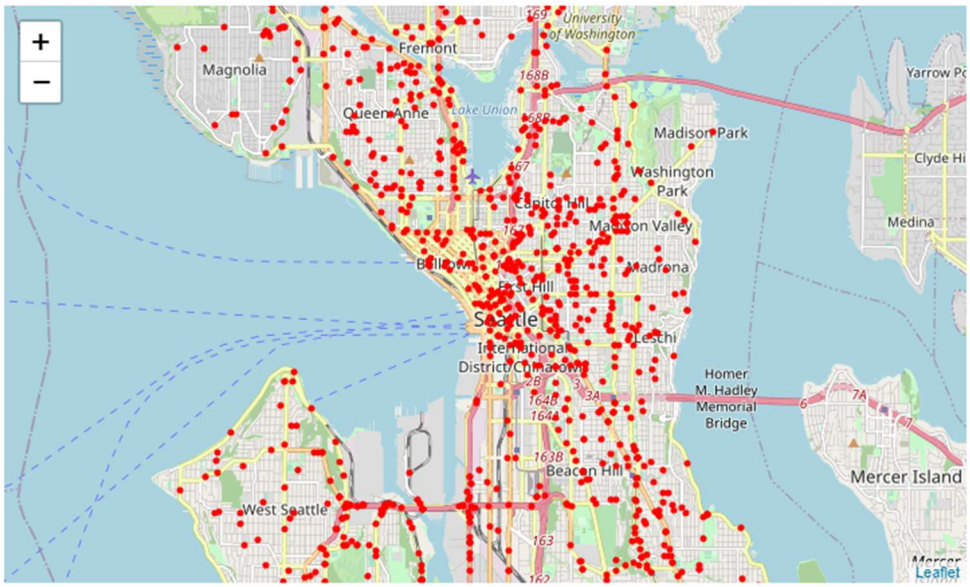
Visualization

We brought the numbers down that low to save on resources while still working on a relevant issue. Part of our analysis involved us using Folium. Using this tool we are able to see where vehicular accidents happen in Seattle, particularly those involving snow and ice.

After loading Folium, we proceeded with loading a generic map of Seattle.

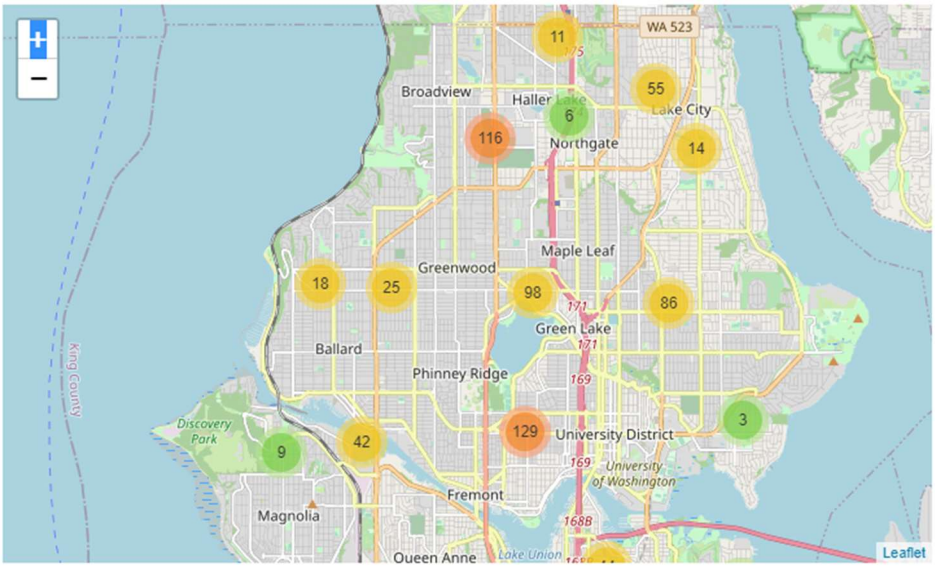


We went ahead and plotted each accident as detailed in df_snow. Imagine if we forced resources to plot all 150000 accidents!

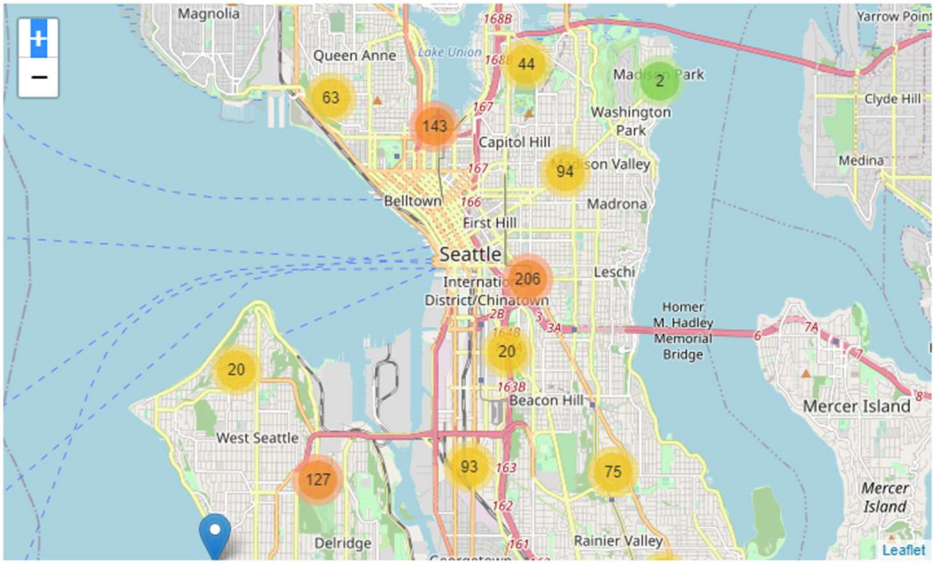


With this we could have gone through each point and make conclusions from here, but we opted to use Folium to group points for us.

From a zoom level of 12, we can already determine areas that stick out more. We have decided to consider the Orange spots.



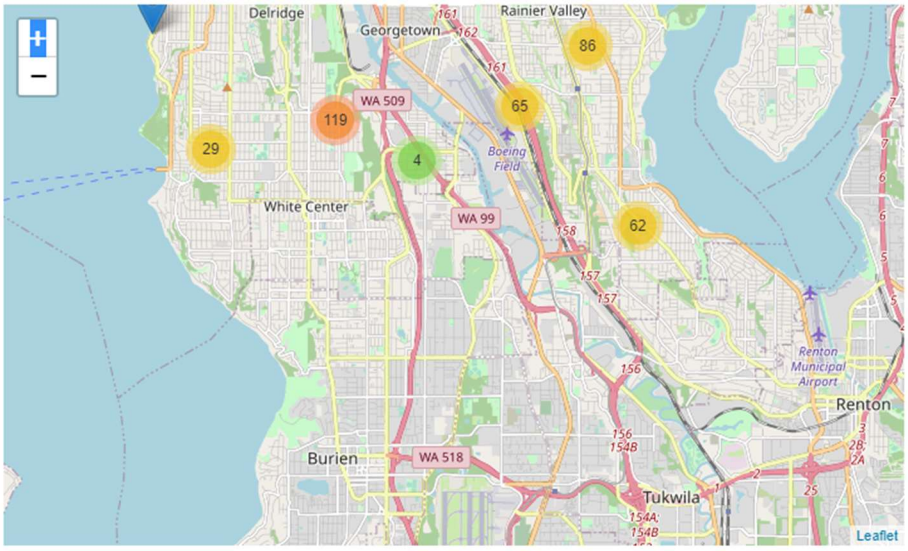
1. 116 recorded accidents in the area between Broadview & Haller Lake.
2. 129 recorded accidents in the area surrounded by Phinney Ridge, University District, and Fremont.



3. 143 recorded accidents at the south end of Lake Union, close to Capitol Hill and north of Belltown.

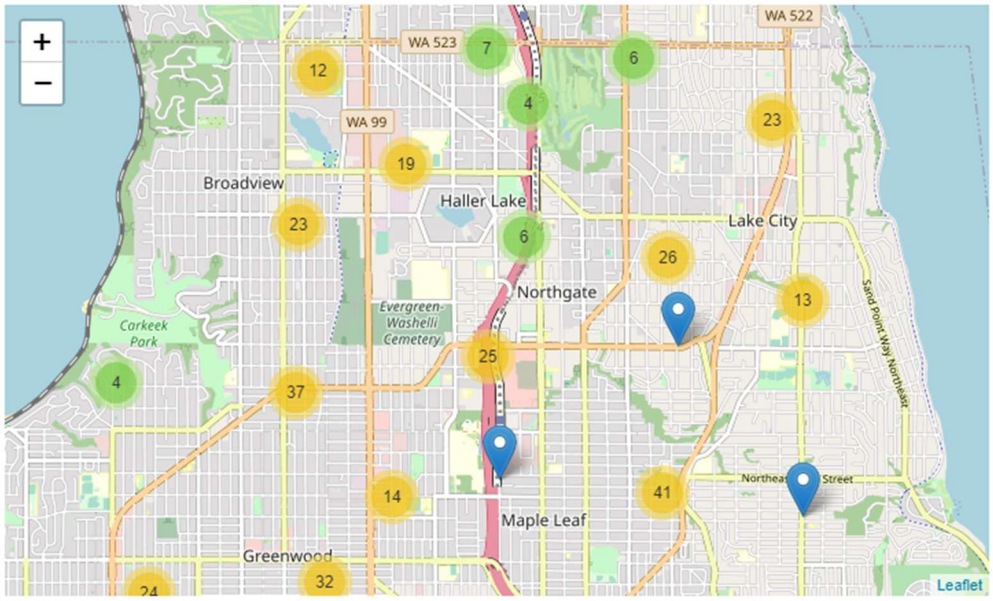
4. 206 recorded accidents in the area just north of the intersection of I5 & I90, close to the International District/Chinatown.

5. 127 recorded accidents in the area between West Seattle and Delridge.

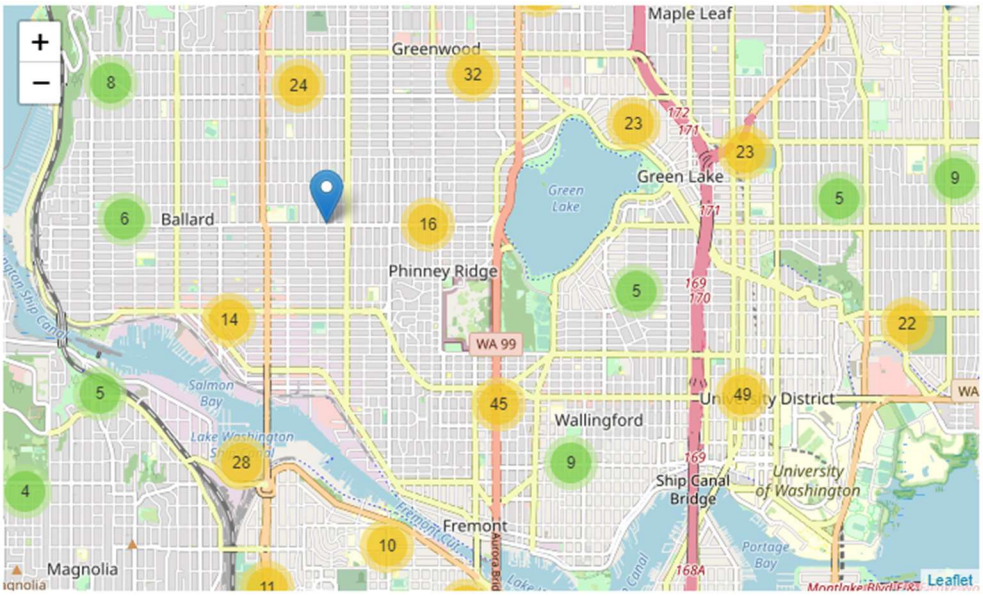


6. 119 recorded accidents in the area north of White Center, to the east of WA 509.

After increasing the map's zoom level to 13, we took a look at the spots with 30 or more recorded accidents, so as to give more precise areas of interest.

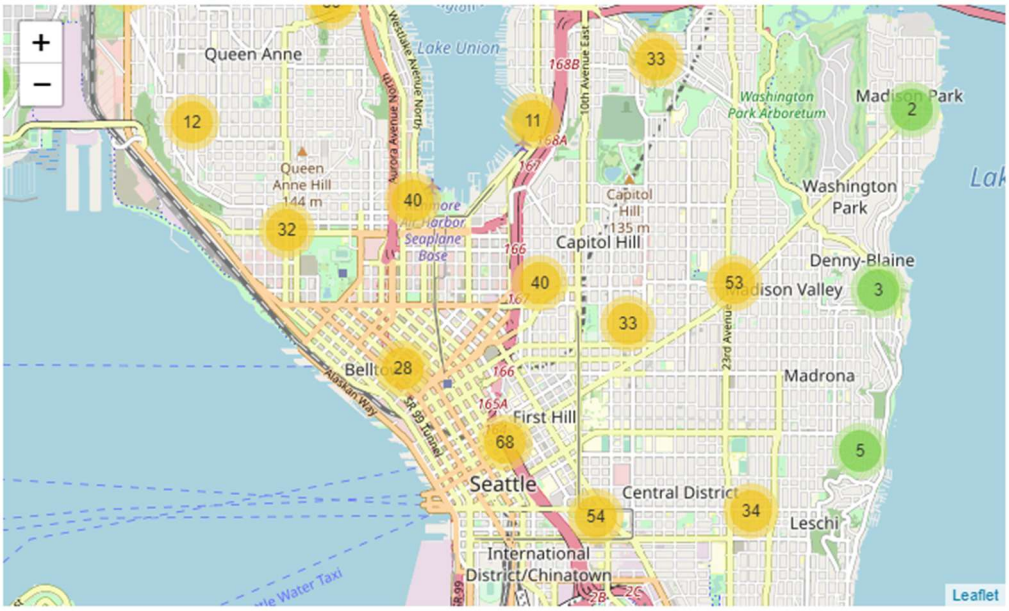


1. 37 recorded accidents close to the junction of N 105th and Greenwood Avenue N.
2. 41 recorded accidents in the area where Lake City Way NE and Ravenna Avenue N merge.
3. 32 recorded accidents close to the junction of N 85th and Aurora Avenue.



4. 45 recorded accidents near the intersection of Aurora Avenue N, N 46th, and Green Lake Way N.

5. 49 recorded accidents in the University District area, close to the junction of 15th Avenue NE and N 50th, and also close to the junction of Roosevelt Way NE and N 45th.



6. 35 recorded accidents in the north Queen Anne area, close to the Lake Washington Ship Canal and Aurora Bridge.

7. 33 recorded accidents in the area between Boyer Avenue E and 10th Avenue E.

8. 40 recorded accidents in the area between Aurora Avenue N and Westlake Avenue N, close to Kenmore Air Harbor Seaplane Base.

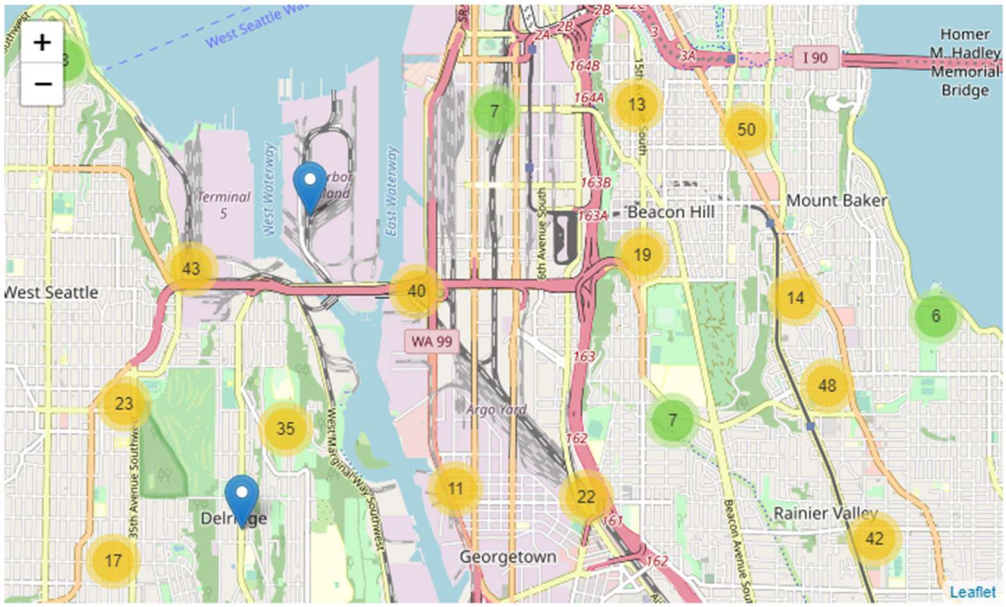
9. 40 recorded accidents in the southern Capitol Hill area, between I 5 Express and East Olive Way.

10. 33 recorded accidents close to East Madison Street and East Pine Street.

11. 68 recorded accidents in the intersection of the I 5 Express, and James Street.

12. 54 recorded accidents close to where Boren Avenue and South Jackson Street connect, close to Bailey Gatzert Elementary School.

13. 34 recorded accidents between 23rd Avenue and Martin Luther King Jr Way, close to East Yesler Way.



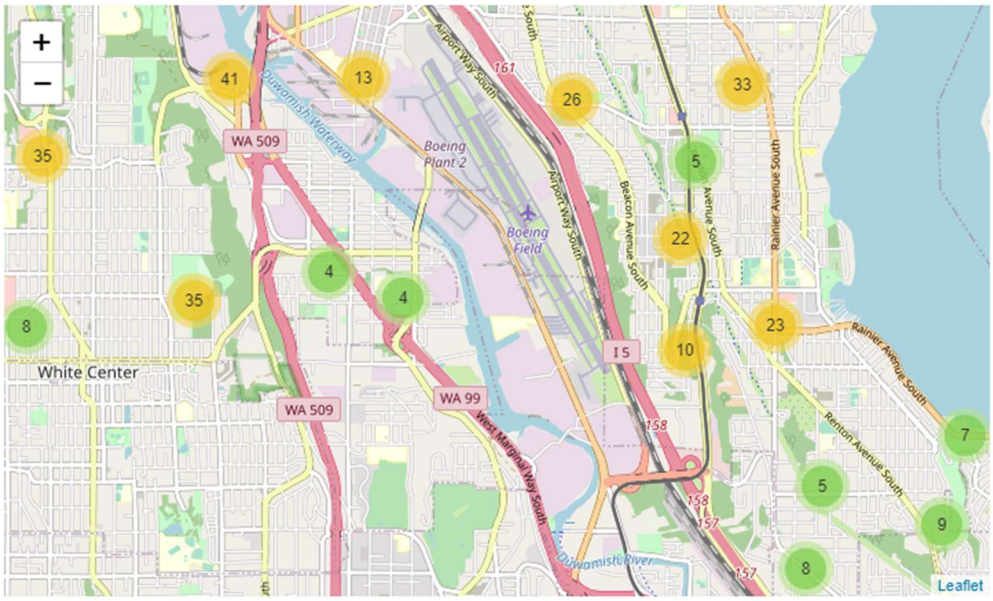
14. 50 recorded accidents south of Judkins Park, north of South College Street, along Rainer Avenue South.

15. 43 recorded accidents to the southeast of Terminal 5, close to the intersection of the West Seattle Bridge, Harbor Avenue Southwest, and Fauntleroy Way Southwest.

16. 40 recorded accidents close to the intersection of the Alaskan Freeway and the West Seattle Bridge.

17. 35 recorded accidents along Delridge Way Southwest, in the area between the West Seattle Golf Course and Puget Park.

18. 42 recorded accidents close to the intersection of Martin Luther King Jr Way and South Orcas Street.



19. 41 recorded accidents close to the Duwamish Waterway, Duwamish Trail, and the 1st Avenue South Bridge.

20. 35 recorded accidents near the intersection of Southwest Holden Street and Delridge Way Southwest.

21. 35 recorded accidents south of Westcrest Park, close to Olson Place Southwest.

22. 33 recorded accidents in the east area of Othello, close to Rainier Avenue South.

Additional Analysis

As a supplement to the above analysis we looked at some of the other columns on their own for more details on where SDOT can work on to avoid more accidents in case of cold weather.

```
df_snow['JUNCTIONTYPE'].value_counts()
```

```
Mid-Block (not related to intersection)              1198
At Intersection (intersection related)                340
Mid-Block (but intersection related)                  244
Driveway Junction                                      55
At Intersection (but not related to intersection)      25
Ramp Junction                                           3
```

```
df_snow['WEATHER'].value_counts()
```

```
Clear                       697
Snowing                     636
Overcast                    340
Raining                     103
Sleet/Hail/Freezing Rain     43
Fog/Smog/Smoke               35
Blowing Sand/Dirt             8
Severe Crosswind              3
Name: WEATHER, dtype: int64
```

```
df_snow['LIGHTCOND'].value_counts()
```

```
Dark - Street Lights On     850
Daylight                    846
Dawn                         81
Dusk                         35
Dark - No Street Lights      31
Dark - Street Lights Off     22
Name: LIGHTCOND, dtype: int64
```

This tells us that there can be as many snow/ice related accidents in daylight as there are at night (regardless of whether street lights are on/present or not). There can also be as many of these accidents in clear weather as there would be in snowing weather.

However, accidents concerning Mid-Blocks that are not related to intersection stand out.

Modelling: K-Nearest Neighbors

We have also prepared a KNN model to predict the potential Severity of an accident given the Location, Junction Type, Weather/Road/Light Conditions, as well as if the vehicle in question was speeding.

Through testing we have determined the best K value to be 12 neighbors. This gives us an accuracy score of 0.7989276139410187.

Model Preparation

Before setting up a model we've had to convert the dataframe into an array of workable float and integer values. Consequently we assigned all these independent variables to X, and the dependent variable SEVERITYCODE into y.

Train/Test Split

True to the methods we have been taught throughout these courses, we have established a Train/Test Split for the model to 'play' with. We used iteration (as was used in previous courses) in determining the best value for K; but, instead of iterating up until 10 neighbors we thought to push the loop to 100 to see if there was anything we were missing.

Again, you are welcome to click here to access the notebook where I have the actual model up and running.

Discussion / Conclusion

Using the methods and approaches which have been taught to us over the past months, We have been able to convert raw data into actual information that could prove useful, and I hope, vital to our authorities in the Seattle Department of Transportation.

- There are **6 main areas** that could be considered first to bring down the number of cold weather vehicular accidents. Drilling down further, we can point out **22 more specific areas** for consideration, with specific streets, junctions, and intersections.

- There are no specific weather or light conditions that stand out, but we can say that **a majority of vehicular accidents in cold weather involve Mid-blocks**, which are not related to any intersection. Analysts can find this particular junction type in the areas specified and work on preventing accidents from happening in those areas.

- Using the **K-Nearest Neighbor** model, we can input accident data (Coordinates, Light/Weather/Road Conditions) and determine the severity of the incident.

There is much more to be extracted from the dataset! We could certainly use the entire dataset of 150000 records to determine where most accidents happen in Seattle, regardless of road condition.

However, I would recommend extracting specific information and making analysis from there, and then working up to answering more relevant questions with more precise answers, from more data.

On a personal note I could have used more time in coming up with more information and more interesting details. This report has opened my eyes to what a simple jab at data analysis, visualization with Machine Learning and Python can add value to people and entire communities.

Thank you for your time. Onward and upward!