The Data

The data set selected for this project was the 311 data, which was filtered to only include complaints relating to snowfall from 2010 – 2018. This dataset was chosen due to the author's familiarity with the data after utilizing it on a previous project. This data primarily tracked complaints made regarding sidewalks and uncleared roads. This feature was further broken down by Borough and, more specifically, community boards. The street and cross streets of the complaint were also included in the complaint. The complaint was either received by phone or online, and this was also recorded in the data. Other variables initially selected for data analysis were; descriptor, action taken, and days needed. Initially, the author believed that the days needed could be predicted based on the other variables. However, after examining the data available, it was assumed that action taken was a more suitable variable to predict and build a model based on the complaint received.

Data Cleaning

After examining the variables initially present in the dataset, the columns Agency Name, Closed Date, Created Date, Cross Street 1, Cross Street 2, Location Type, Resolution, Action Updated Date, Resolution Description, Status, Street Name, and Unique Key were removed. Once the data was loaded into python and the correlation matrix was created, the author observed that Borough and community board were highly correlated; however, the author believed they were both required to determine if an action would be taken. It was also determined that action taken would be converted from the sanitation report code to the values: action taken, and action not taken. The original data file contained approximately 80,000 complaints. After removing any row with missing data from the file, the number of records in the CSV file was reduced to approximately 54,000; this was verified by using pandas' dropna function. The final step taken to improve the quality of the data was the removal of any record with a response time greater than 17 days since, after such a lengthy period of time, even if there were a response, it would result in no action taken. Cross Street 2 was removed due to the perfect correlation with Cross Street 1; Cross Street 1 was also removed. The Street Name was removed due to its lack of importance to the action taken. Keeping the street name would also require further data processing and transformation with minimal gained benefit. Research done while analyzing the data indicated that statistically, the Borough is less critical than the Community Board the complained originated from as it

relates to the time taken to respond to most complaints affecting the action taken. To create the model to predict whether an action was (not) taken, six columns were used: Borough, Days Needed to Resolve Complaint, Community Board, Descriptor, Action Taken, Open Data Channel Type.

Data Transformation & Validation

The six columns in the data file initially were an object data type except for 'Days Needed,' which was an int64 data type. It was determined that a categorical model would be built, requiring data transformation of the data present in the panda data frame to the category data type. This transformation allowed the data to be used in categorical data models. The transformation of the data went smoothly due to the similarity in data types for 5 of the six columns used. A correlation matrix was then created to determine if any of the variables had a significant correlation. The descriptor and action taken had a correlation value of approximately 0.6, and the descriptor and days taken had a correlation value of 0.4. A level of correlation was expected between these variables since the data shows that specific complaints disproportionally result in action or no action. This is especially true when placed in the context of a particular Borough Community Board. The correlation between the descriptor and the days needed was also expected. Specific snow complaints which are represented by a specific descriptor are considered a high priority. At the same time, others are regarded as a low priority or require a high priority complaint to be addressed. Once the high priority complaints are addressed, more areas become accessible to assess the action necessary for lower priority complaints. There is also a chance that the responsible party may address low-level offenses, or a change in temperature would result in the complaint being moot before it is investigated. The final task for the data was removing the variable the model was predicting from the original data. During the supervised and unsupervised model, it was observed that particular models required this data to be stored in different data types.

Data Visualization & Analysis

Several visualizations were created during assignment 1 for data exploration. A figure was created plotting 'Days Needed to Resolve a Complaint' vs. 'The Number of Complaints Resolved.' This figure showed that the relationship between these two variables resulted in a chart resembling an exponential decay curve. Further examination shows that approximately 80% of all complaints are resolved within seven days of receiving the

complaint, with 40% of complaints being resolved within two days of receiving the complaint. One of the most surprising results from the visualizations created to compare the variables was the realization that Staten Island only has four community boards a disproportionately lower number than other boroughs in New York City. It was also observed that sidewalk complaints accounted for approximately 68% of the total complaints, street cleaning complaints accounted for 30% of all complaints, and the remaining two categories accounting for approximately 2% of all complaints. Given the type of complaint, the number of days to resolve a complaint, and the community board, the author of this paper assumes that based on clustering, it would be easier to predict whether an action will be taken or not taken when the data from these categories are collectively considered. The final observation of the grouping of complaints by community board showed the number of total complaints received appears to be evenly distributed among the community boards. This observation has led the author to conclude that while responses may vary across the city, snow events tend to affect everyone similarly living in New York City, and these people tend to have the same expectations about their neighbors and city's response to cleaning up after a snowstorm.

Supervised Learning

The two algorithms chosen for supervised learning were the k nearest Neighbor and the decision tree algorithm. The k nearest Neighbor is an algorithm that can be used to solve both classification and regression problems. It is built on the principle that things close to each other will have similar characteristics or features. The similarity of these variables is based on the mathematical principle about the distance between points, with the algorithms default being the straight-line distance between two points. Utilizing the training data with the default parameters, the model had an accuracy score of 0.78. The training data set was then split to determine the cross-validation value. The training set was then used to determine the optimal number of nearest neighbors for this model. The result of this analysis was nine, and the model generated after including this parameter resulted in an increased accuracy score of 0.83. It was also observed that the training data accuracy score tended to be slightly higher than the test data score. In utilizing both the default and adjusted nearest neighbor algorithm, the following results were achieved or observed:

Changing the power to 1 had a negligible impact on the model's accuracy score in comparison to the default Nearest Neighbor.

- Graphing the training accuracy vs. testing accuracy for varying nearest neighbor values shows that the training model always slightly outperforms the testing model. The graph also shows that there is very little gained as it relates to improving the accuracy of the model by adjusting the nearest neighbor value.
- The precision score of 0.83 indicates that the model predicts 83 out of every 100 values.
- The recall score of 0.89 indicates that if the model picked 100 values, it would be correct 89 times.
- The f-score taking both precisions and recall into account estimates the model's precision as 0.85, meaning that the model predicts 85 of every 100 values correctly when both recall and precision are considered in the calculation of the value.

Decision trees similar to k nearest Neighbor is an algorithm used to solve categorical and regression problems. This is primarily achieved by splitting the data in several yes, no, or if then else decisions. The depth of the three depends on the user preference and available choices based on the number of options in the data. The results of this algorithm can be visually thought of as a family tree or a workflow diagram. The decision tree is also not randomly generated but based on low entropy, the result that returns the most significant gain in information based on what is already known. Utilizing the training data to determine the optimal depth of the tree showed that when the depth was set to 1, the model had an accuracy score of 0.81. However, when the optimal max depth value of 8 was utilized, the model's accuracy was 0.83; this was the same level of accuracy achieved with the optimized nearest neighbor algorithm. Cross-validation scores were also calculated to verify the result. It was also observed that the training data accuracy score tended to be slightly higher than the test data score, except for the optimal depth where they overlapped. In utilizing both the default and adjusted decision tree algorithm, the following results were achieved or observed:

- Adjusting the splitter from the default setting to random has a negligible impact on the model's accuracy score in comparison to the default decision tree.
- Graphing the training accuracy vs. testing accuracy for varying maximum depth levels resulted in 8 being identified as the depth which the training accuracy and testing accuracy intersect. The graph also shows a maximum depth increases the training model accuracy by a single percentage point while creating a divide between

the testing and training accuracy. The difference in accuracy graphically appeared only to be 1-2 percent.

- The precision score of 0.83 indicates that the model predicts 83 out of every 100 values.
- The recall score of 0.91 indicates that if the model picked 100 values, it would be correct 91 times.
- The f-score taking both precision and recall into account estimates the model's precision as 0.87, meaning that the model predicts 85 of every 100 values correctly when both recall and precision are considered in the calculation of the value. • The Decision Tree, when optimized by the user appears to outperform the Nearest Neighbor based on the precision, recall, and f-score.

Unsupervised Learning

In the final project, PCA for feature selection was utilized to determine the variables required to retain 95% of the variance. It was determined utilizing PCA for feature selection that all the variables were needed to account for 95% of the variance. The data from PCA for feature selection was not used with the two supervised learning algorithms to determine if it would have resulted in better results since none of the features were identical.

The DBSCAN is an unsupervised machine learning algorithm which groups data points together based on the distance separating the points. The DBSCAN with PCA vs. without PCA showed that with PCA, the data was divided into 44 clusters, while without PCA, 1123 clusters were required. It was also notable that the non-PCA implementation had 6008 noise points, while only 370 were present in the PCA version. The k-mean elbow method is a technique used in unsupervised learning to determine the optimal number of clusters or k value. The k-mean implementation showed little difference between the PCA and non-PCA implementation; the PCA had a k elbow value of 3 while the non-PCA had a value of 4. Agglomerate/Hierarchical algorithm is one of the more successful unsupervised clustering algorithms. It works by starting with individual data points, then pairing these data points, which are then paired by groups, until all the data is in one large group. Utilizing the Agglomerate/Hierarchical algorithm with and without PCA appeared to return inconclusive or results that were beyond the investigator's ability to interpret.

Lessons Learned & Conclusions

After conducting data cleaning, unsupervised, and supervised machine learning, I have drawn the following conclusions:

- Machine learning is not a substitute for developing a well thought out theory.
- Good data selection practices remain essential, and you should be able to justify in 1-2 sentences any feature selected for your model. You should also be able to explain the exclusion of any element and, if possible, have a data feature in mind that could improve your model if available.
- Selecting an algorithm to build your model and the optimization process should be taken seriously; this can affect the quality of your results in the testing phase.
- The decision to choose unsupervised vs. supervised learning should be based on the availability of data and your understanding or lack of knowledge about the associative properties between the features.

In conclusion, if I had to redo this project the change, I would make predicting the number of days to respond to a snow complaint.