ANLY501 Project Assignment 2

Jun Wang
Junke Wang
Yanchen Wang

Section1: Basic Statistical Analysis and Data Cleaning Insight

Further Data Cleaning:

- **Car Accident Data:**

For the car crash data set, in project assignment 1, we set the attribute "SPEEDING_INVOLVED" as a categorical (binary) attribute with “0” or “1”, where “0” mean speeding was not a factor and “1” means it was a factor of car accidents.

In this project assignment, we further clean this data by first checking the missing and incorrect value, then checking for the data outliers, and removes all the noises to form a cleaned data. First of all, there is no issue of missing data except for attribute “ADDRESS”. However, since we are not able to fill in the missing address, so the we decided to leave the attribute “ADDRESS” as it is. Note that the attribute “ADDRESS” is not significant in our data, since we also have attributes “LONGITUDE” and “LATITUDE” to identified the location of one specific car accident.

After checking for missing values, we decided to modify some attributes in car accident data before conduct outlier test. The three unknown attributes “UNKNOWNINJURIES_BICYCLIST”, “UNKNOWNINJURIES_DRIVER”, and “UNKNOWNINJURIES_PEDESTRIAN” are unrelated to our study, therefore, we dropped them from the original dataset. In the original dataset, it separates the number of injury into 9 attributes with different level of injury-severity, as well as different types of people. Therefore, we decided to merge all 9 attributes into one attribute named “Total_injuries” which is the sum of attributes “MAJORINJURIES_BICYCLIST”, “MAJORINJURIES_DRIVER”, “MAJORINJURIES_PEDESTRIAN”, “MINORINJURIES_BICYCLIST”, “MINORINJURIES_DRIVER”, “MINORINJURIES_PEDESTRIAN”, “FATAL_BICYCLIST”, “FATAL_DRIVER”, and “FATAL_PEDESTRIAN”. In addition to total number of injury, we added another merged attribute called “Total_involved” which is the sum of attributes “TOTAL_BICYCLES”, “TOTAL_PEDESTRIANS”, and “TOTAL_VEHICLES”. The “Total_involved” is the total number of car, bicycles, vehicles, and pedestrians involved in one specific accidents.

Then with the modified car accident data, we decided to use the IQR method to check for the outliers. Since most of our numerical variables are less than 10 of which has a significant portion of values equal to 0. So, if we use the original IQR method which has IQR equals the difference between the 3^{rd} and 1^{st} quantiles, the value for IQR are all 0 that would not allow us to

check the outliers. So we decided to modify the IQR by setting it equals the difference between 95 and 5 percentiles. The lower bound equals 1.5 modified IQR lower than the 5 percentile, and the upper bound equals 1.5 modified IQR larger than the 95 percentile. We identified the outliers as numbers outside of the boundary, and remove the entire record from the original data set to the outlier data set. We decided to remove all the outliers since all of them are extremely rare events which are considered as unrelated records to the weather condition. For instance, one of the accident has a number of pedestrian as 11, which is very unlikely occurred. We applied our modified IQR method to two attributes we merged above, "Total_injuries" and "Total_involved", which is the representation of other 13 attributes (when we compute the modified IQR values for those attributes, most of them are still 0, therefore, we decide to use the merged attribute to test for outliers).

- **Weather Condition Data:**

For the weather condition data, in project assignment 1, we already handled with incorrect and missing value for attribute "SNOW" and "PRCP". For the attribute "SNOW", we first set it to 0 if the minimum temperature was above 32 degrees (above freezing temperature) and then for those entries still missing value, we used mean to replace the space. For the attributes "PRCP", we replaced missing values and noise values by using the mean of other two stations for that day.

Similar to car accident data, in this project assignment, we further clean weather condition data by first checking the missing and incorrect value for the left over attributes "TMAX" and "TMIN", then checking outliers for all 4 attributes, and removes all the noises to obtain a cleaned data. We used the same strategy as last project assignment to fill in the missing values of "TMAX" and "TMIN". We replaced missing values by using the mean of other two stations for that day. Then, we check for the incorrect values. For each attribute, we defined that one value is incorrect if the largest difference among three stations is greater than 10, and we replaced the incorrect value by the mean of the other two station as well. After filling all the missing and replacing all the incorrect values, there are still 12 missing value due to the fact that on those day, two of the station does not have data, so we were not able to take the average. So we searched the history weather condition for those still missing values and filled in manually.

We believed there is no outlier in the weather condition data. After sorting all the attribute, we have the following results:

	PRCP (inch)	SNOW (inch)	TMAX(Fahrenheit)	TMIN(Fahrenheit)
Minimum Value	0	0	16	2
Maximum Value	3.37	14	100	81

Since all the minimum and maximum values for weather attributes are reasonable, we considered that there is no outlier need to be removed in our data set. (Note that on 1/23/2016, there was a blizzard and the maximum value of snowfall is occurred on that day.)

Basic Statistical Computation and Explanation:

After the further cleaning phase, we computed the mean, median, and standard deviation for all the numerical attributes for both car accident data and weather condition data (Note that although attributes "LONGITUDE" and "LATITUDE" in car accident data are numerical, it is meaningless to compute the mean, median, or standard deviation). The following are the results for statistical computation:

Statistic Summary for Attributes of Car Accident Data

Attribute Name	Mean	Median	Standard Deviation
BICYCLISTSIMPAIRED	7.93551824602e-05	0.0	0.00890779912297
DRIVERSIMPAIRED	0.0102028091735	0.0	0.100492347263
FATAL_BICYCLIST	9.06916370974e-05	0.0	0.0095227838432
FATAL_DRIVER	0.000872907007063	0.0	0.0299135078133
FATAL_PEDESTRIAN	0.000374103003027	0.0	0.0193381242619
MAJORINJURIES_BICYCLIST	0.000691523732868	0.0	0.0262877448214
MAJORINJURIES_DRIVER	0.099545408169	0.0	0.340294956154
MAJORINJURIES_PEDESTRIAN	0.0024713471109	0.0	0.0634796841067
MINORINJURIES_BICYCLIST	0.00604233032162	0.0	0.0780801652342
MINORINJURIES_DRIVER	0.156227681355	0.0	0.406943764852
MINORINJURIES_PEDESTRIAN	0.0182630284205	0.0	0.181553377997
PEDESTRIANSIMPAIRED	0.000748206006054	0.0	0.0346569528279
TOTAL_BICYCLES	0.0103955289023	0.0	0.10363841019
TOTAL_GOVERNMENT	0.117910464681	0.0	0.340456846559
TOTAL_PEDESTRIANS	0.0175941775969	0.0	0.139990211292
TOTAL_TAXIS	0.0927775447507	0.0	0.305832933921
TOTAL_VEHICLES	1.97655621181	2.0	0.528876375286
Total_injuries	0.284579020757	0.0	0.544233371982
Total_involved	2.00454591831	2.0	0.510305503602

Statistic Summary for Attributes of Weather Condition Data

Attribute Name	Mean	Median	Standard Deviation
PRCP	0.12322122571	0.0	0.318740130212
SNOW	0.0626557050324	0.0	0.681521792785
TMAX	67.6386397608	70.0	18.4635887589
TMIN	48.2669407075	49.0	17.57507367

One significant is that for the car accident data, most of the attributes, by their nature, have the median of 0 and mean and standard deviation close to 0. As mentioned previously, most of the inputs for each attributes in car accident data are less then 10, and we have over 90,000 accident records, among which most of the records are 0. For instance, not every accident would have bicyclist or pedestrian involved, so many of the records are 0. When we computing the mean, we need to use the number of bicyclist or pedestrian over the total of over 90,000 accidents, which will obviously give a closed-zero value. Similar arguments for the rest of the attributes in car accidents data. In addition, the median of "TOTAL_VEHICLES" and "Total_involved" are 2, which is reasonable, since for every accident there usually have 2 parties i.e. 2 vehicles involve. Even though there were records that have more than 2 vehicles, with a relatively large denominator (in this case, more than 90,000), the mean should still around 2.

For weather condition data, the median for rainfall and snowfall were also be 0, and the mean and standard deviation are around 0. Since during the four-year time period (2013-2016), the rainy and snowing days are relatively small compare to the total number of days. Then, by the same argument used above for car accident data, it is reasonable that the statistical numbers are around 0. Other than that, the mean and median for maximum and minimum temperatures are

close to each other, where the median is a little greater than mean. This shows that the distributions of maximum and minimum temperatures are both slightly skewed to the left. (This is also shown in the next section by plotting the histograms for "TMAX" and "TMIN")

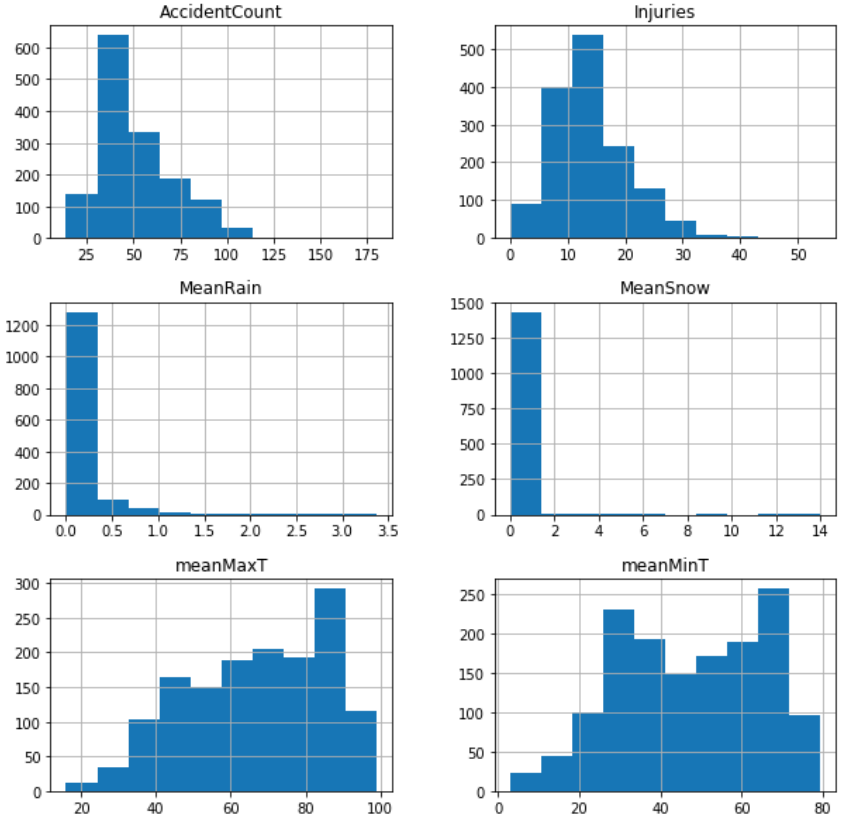
Data Binning:

We used the equal-depth binning method to bin the attributes "TMAX" and "TMIN". By setting the 33 and 66 percentiles, we binned the temperature attributes into three bins with equal numbers of records. Since there are four seasons during the year, but spring and fall are relatively shorter than summer or winter. In addition, temperatures during spring and fall are similar and close to each other. Therefore, using the equal-depth method, we binned the record into summer, spring plus fall, and winter.

Section2: Histograms and Correlations

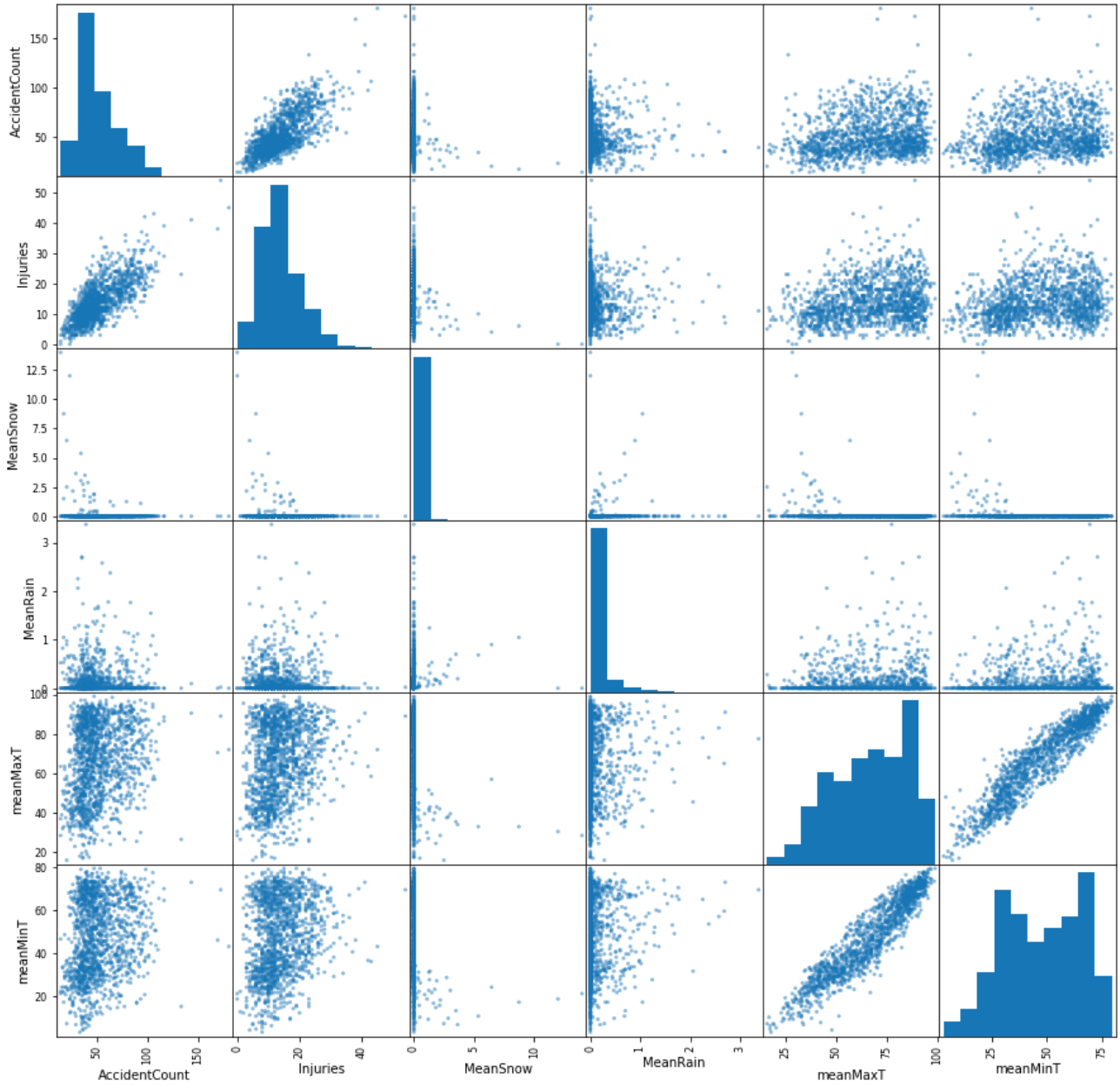
Before plotting the histograms and scatterplot, we decided to merge our two datasets together in order to obtain a better comparison across the datasets. For the weather condition data, we combine the three stations information into one average value for rainfall, snowfall, maximum temperature, and minimum temperature. For the car accidents data, we count the number of total accidents and total number of injuries per day. Finally, we merge the two datasets together with each record represent one specific day. So during each day of 2013-2016, we will have attributes "AccidentCount", "Injuries", "MeanRain", "MeanSnow", "meanMaxT", "meanMinT".

Histograms:



We plotted the histograms for all the attributes in the merged data. From the histograms above, a large number of accident count per day is around 30-40, and a large number of injuries per day is around 10-15. We also can see that the distributions of total number of accident and total number of injuries are similar (both skewed to the right). This is because when the number of accidents increase, it is most likely that the number of injuries would increase as well, since they are positively correlation, and this will further show by the scatterplots. For the rainfall and snowfall histograms, we can see that most of the days has no rain or snow. This indicate that during the four years, the number of sunny day is significantly higher than the rainy and snowing day. For maximum and minimum average temperatures, the distributions are both slightly skewed to the left, because they have a strong correlation by nature.

Scatterplots:



The most obvious correlation shown in the above scatterplot matrix is between the maximum and minimum average temperatures. As mentioned above in the histogram section, these two attributes by nature has a strong positive correlation. In addition, with the similar explanation above, there is a significant correlation between total number of car accident and total number of injuries. Then, the rest of the scatterplots do not show a clear pattern or a specific relationship among the attributes. For instance, we were expecting a positive correlation between rainfall/snowfall and number of accidents, a strong correlation between temperature and number of accidents, as well as a significant correlation between rainfall/snowfall and number of injuries. However, none of the above expectation was shown by the scatterplot matrix. Therefore, we would conduct further hypothesis tests regarding to those arguments.

Section3: Cluster Analysis

Overall Result for Cluster Analysis:

In the car accident and weather data set, we found out the silhouette coefficient is highest when number of clusters is 2. This number of cluster makes sense because in the car accident data set, number of injuries is mostly zero and number of parties involved is mostly two (two-car accident). So that we could cluster them into two levels, low and high. In the weather dataset, maximum temperature and minimum temperature also make sense to cluster them into two clusters i.e. one captures winter temperature and one captures summer temperature. For the rainfall, most of the days didn't rain and it makes sense to put them into low and high two clusters.

- **Car Accident Data Cluster Analysis:**

In the car accident data set, we used the attributes "Total_injuries", "Total_involved", and "SPEEDING_INVOLVED". As mentioned in section1 above, the "Total_injuries" is sum of major, minor, fatalities of pedestrians, cyclists and drivers. "Total_involved" is sum of total pedestrians, total bicycles and total vehicles. We use those two variables because most of instances in those sub attributes are zero or one i.e. very small value that it's very hard to cluster an attributes with most values = 0 and very few of them = 1 or larger. In this case, we aggregate similar attributes into one attribute and do cluster analysis on the aggregated attributes.

First, by conducting three cluster analyses on car accident data, the ***centroids*** of the clusters in the original data are as follow:

- The centroids by using K-means:
  [0.484192037470726, 2.0995316159250588, 1.0]
  [0.2806376657456967, 2.0026704276152274, 0.0]
- The centroids by using Ward:
  [0.2927028128821851, 1.9990827558092132, 0.0]
  [0.46808510638297873, 2.0319148936170213, 1.0]
- The centroids by using DBSCAN (number of clusters in DBSCAN is 3):
  [6.0, 4.0, 0.0]
  [0.29011841567986935, 1.9972437729685586, 0.0]
  [0.4187192118226601, 2.2167487684729066, 1.0]

(Note that DBSCAN only has eps and the centroid [6.0, 4.0, 0.0] might be an outlier and we could see it in the plot later in this section.) For each centroid, the first column is number of injuries, second column is number of parties involved, and the third column is speeding

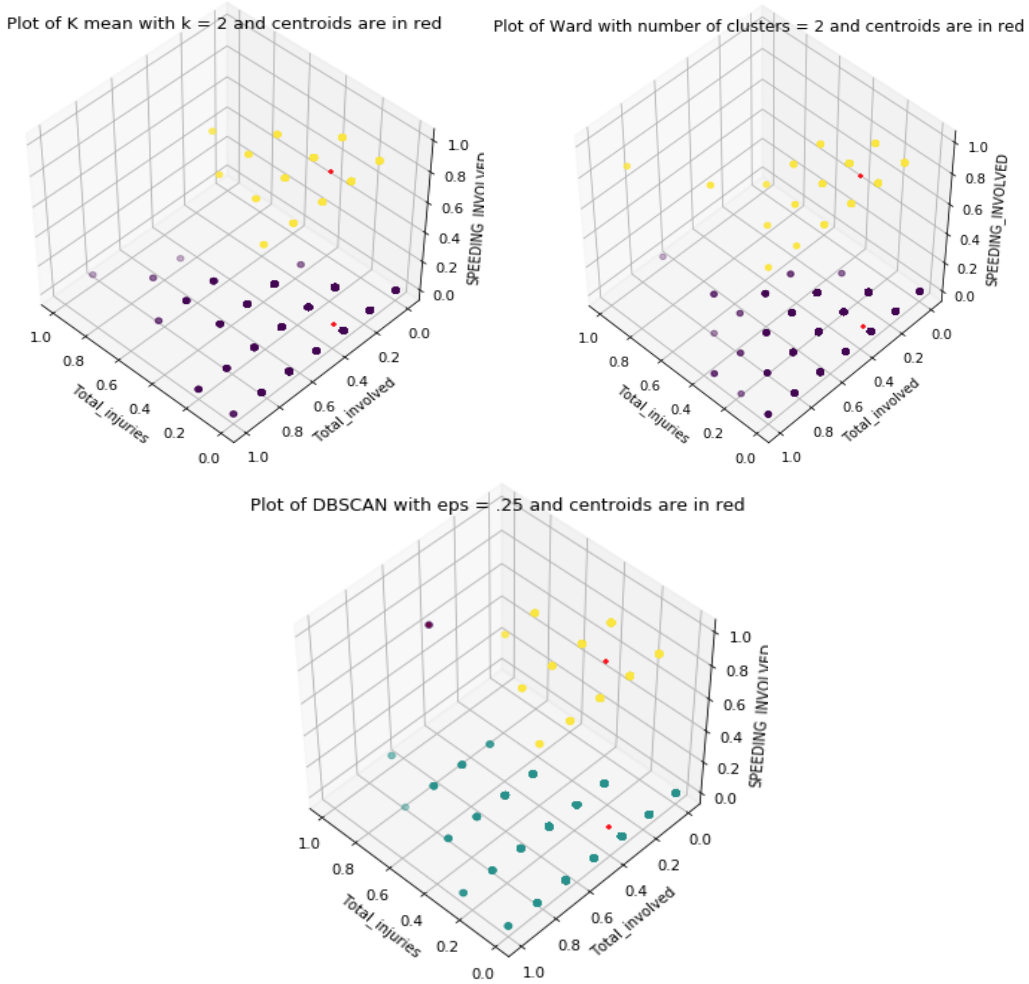
involved, which is a binary variable. In addition, for each cluster, the third column is 0 and 1. Here, 0 means speeding was not a factor of the accident and 1 means speeding was a factor of the accident. From the cluster, we can see that accidents with speeding involved had more number of parties involved and number of injuries on average. Overall, with different three cluster analyses, the centroids are about the same in each cluster.

Second, we used the ***Silhouette coefficient*** to evaluate the quality of our clusters. The list of results for the Silhouette coefficient for each case of three different cluster method are as follows:

- For K-means with k= 2, the Silhouette coefficient is 0.82874215613
- For K-means with k= 2, the Silhouette coefficient is 0.772950271498
- For Ward with k = 2, the Silhouette coefficient is 0.82497653045
- For Ward with k = 3, the Silhouette coefficient is 0.780263673914
- For DBSCAN with eps = 0.2, the Silhouette coefficient is 0.802258860098
- For DBSCAN with eps = 0.25, the Silhouette coefficient is 0.860318858337
- For DBSCAN with eps = 0.3, the Silhouette coefficient is 0.854888059573
- For DBSCAN with eps = 0.35, the Silhouette coefficient is 0.852529211487

From the above results, we can see that the Silhouette coefficient is highest for k = 2 in K-means and Ward and for eps = .25 in DBSCAN.

Finally, we plotted each cluster and get the ***plots*** as follows:



From the above plots, we can see that points are discreetly distributed because there were only a few unique values for "Total_injuries" and "Total_invloved". From the centroids, we can see the centroid when Speeding = 1 has bigger "Total_injuries" and bigger "Total_involved" than Speeding = 0 and this reflects the result from the clustering part that speeding is causes more injuries and more number of parties involved on average. In both k mean and ward, we can see the clusters and centroids are about the same in the two plots. In DBSCAN, we can see there is one point at upper left corner and that point is the only point in that cluster so that it could be the outlier. Other than the outlier in DBSCAN, plots of k mean, ward and DBSCAN are very similar with each other.

- **Weather Condition Data Cluster Analysis:**

In weather data set, we used the attribute 'TMAX', 'TMIN' and 'PRCP'. We didn't use the attribute 'SNOW' in cluster analysis because majority of days didn't snow at all so that the values in that attribute mostly are zero. When conducting cluster analysis, we merged data by date by taking mean of records from three stations on that day.

After merging the data, we applying the same steps from car accident data to the weather condition data. First, we conducted three cluster analyses on the data, the ***centroids*** of the clusters in the original data are as follow:

- The centroids by using K-means are:
  [0.14284911370514472, 81.76977950713359, 61.96952010376135]
  [0.09734057971014495, 50.20434782608696, 31.441304347826087]
- The centroids by using Ward are:
  [0.08262419871794878, 53.79086538461539, 34.58052884615385]
  [0.1725887652358239, 84.15182829888712, 64.70906200317965]
- The centroid of DBSCAN (Number of clusters in DBSCAN is 2):
  [2.716666666666667, 61.5, 50.5]
  [0.11779872058487537, 66.86943111720356, 47.54763536668951]

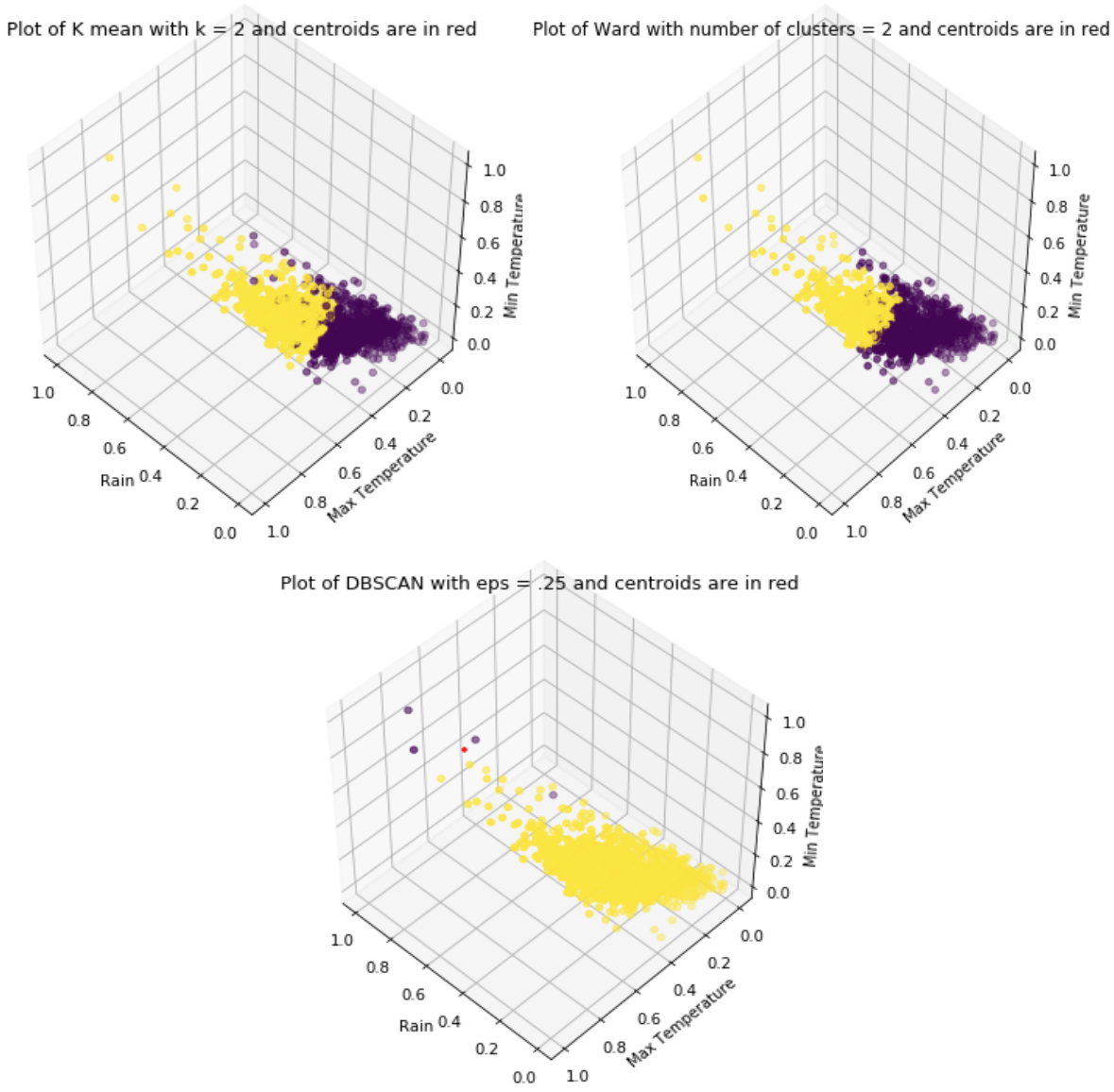
For each centroid, the first column is rainfall, second column is maximum temperature and the third column is minimum temperature. From the maximum and minimum temperature, we can see the difference between two clusters is pretty big and we can see that one is for summer temperature and the other one is for winter temperature. We can also see that there was more raining in summer than winter, which reflects the truth. In K-means and Ward, the centroids are about the same. In DBSCAN, the first column in the first centroid (rainfall amount) is significantly different from the centroids in k mean and ward. This centroid might be an outlier and we should be able to see it clearer in the plot later

Second, we used the ***Silhouette coefficient*** to evaluate the quality of our clusters. The list of results for the Silhouette coefficient for each case of three different cluster method are as follows:

- For K-means with k= 2, the Silhouette coefficient 0.562689103907
- For K-means with k= 3, the Silhouette coefficient 0.481463067128
- For Ward with k = 2, the Silhouette coefficient is 0.533078795742
- For Ward with k = 3, the Silhouette coefficient is 0.466275647485
- For DBSCAN with eps = 0.15, the Silhouette coefficient is 0.5103764991
- For DBSCAN with eps = 0.2, the Silhouette coefficient is 0.548403217818
- For DBSCAN with eps = 0.25, the Silhouette coefficient is 0.549280024936
- When eps = 0.3 Number of cluster is 1 and score is not available

We can see that the silhouette coefficient is highest for k = 2 in k mean and ward and for eps = .25 in DBSCAN. When eps = .3 in DBSCAN, number of cluster becomes 1.

Finally, we plotted each cluster and get the ***plots*** as follows:



From the above plots, we can see that the centroids in red are covered by the points in the sample. Unlike the plots in car accident dataset, points here are more continuously distributed because there are a lot more unique values in temperatures than the values in car accident data. In K-means and Ward, the clusters are about the same that two clusters represent low and high temperature and in the cluster with high temperature, we can see that rain amount is bigger, which represents that summer has more rainfall. In DBSCAN, we can see majority of points are in one cluster and the other cluster only contains a couple points with very large amount of rainfall.

Section4: Association Rules / Frequent Itemset Mining Analysis

In this section, we used the reference from *https://pypi.python.org/pypi/apyori/1.1.1*. we merged the car accident data and weather data together. So the merged data has the number of accidents on a specific day, temperature, perception and snowing. In order to run the association rule, we binned perception and snowing together to create a new attribute called "niceWeather". This attribute is 1 if either rainfall or snowfall is greater than 0.1 inches and 0 otherwise. In the previous part, one of our guesses is that there might be some impact of raining or snowing on number of accidents for that day. We also binned number of accident into two bins using equi-depth binning strategy.

As a result, we obtained that, with the equi-depth bins, about a quarter of days doesn't have raining or snowing. First of all, for the support level, {niceWeather0.0, accidentBin0.0} has a support level of 0.378 and {niceWeather0.0} and {accidentBin0.0} have support level 0.7526 and 0.5086. If those two are independent then {niceWeather0.0, accidentBin0.0} should have support level = 0.7526*0.5086 = 0.3828 ≠ 0.378 so that those two attributes are not independent. Secondly, as for the confidence level, {niceWeather0.0} → {accidentBin0.0} has confidence level of 0.5023 and {niceWeather1.0} → {accidentBin0.0} has confidence level of 0.5278 that probability of accidentBin=0.0 given weather was nice is 0.5023 and probability of accidentBin=0.0 given weather was bad is 0.5278. This gives us a sense that the number of accidents is smaller on a rain or snow day. On the other hand, {niceWeather0.0} → {accidentBin1.0} has confidence level of 0.4977 and {niceWeather1.0} → {accidentBin1.0} has confidence level of 0.4722 that probability of accidentBin=1.0 given weather was nice is 0.4977 and probability of accidentBin=0.0 given weather was bad is 0.4722. This also gives us a sense that the number of accidents is smaller on a rain or snow day.

Follows are the results from Apriori algorithm with different support levels:

Support level = .05

support_set	support_level
accidentBin0.0	0.508591065
accidentBin1.0	0.491408935
niceWeather0.0	0.75257732
niceWeather1.0	0.24742268
accidentBin0.0niceWeather0.0	0.378006873
niceWeather1.0accidentBin0.0	0.130584192
accidentBin1.0niceWeather0.0	0.374570447
niceWeather1.0accidentBin1.0	0.116838488

From	To	conf_level
	accidentBin0.0	0.508591065
	accidentBin1.0	0.491408935
	niceWeather0.0	0.75257732
	niceWeather1.0	0.24742268
accidentBin0.0	niceWeather0.0	0.743243243
niceWeather0.0	accidentBin0.0	0.502283105
accidentBin0.0	niceWeather1.0	0.256756757
niceWeather1.0	accidentBin0.0	0.527777778
accidentBin1.0	niceWeather0.0	0.762237762
niceWeather0.0	accidentBin1.0	0.497716895
accidentBin1.0	niceWeather1.0	0.237762238
niceWeather1.0	accidentBin1.0	0.472222222

Support level = .1

support_set	support_level
accidentBin0.0	0.508591065
accidentBin1.0	0.491408935
niceWeather0.0	0.75257732
niceWeather1.0	0.24742268
accidentBin0.0niceWeather0.0	0.378006873
niceWeather1.0accidentBin0.0	0.130584192
accidentBin1.0niceWeather0.0	0.374570447
niceWeather1.0accidentBin1.0	0.116838488

From	To	conf_level
	accidentBin0.0	0.508591065
	accidentBin1.0	0.491408935
	niceWeather0.0	0.75257732
	niceWeather1.0	0.24742268
accidentBin0.0	niceWeather0.0	0.743243243
niceWeather0.0	accidentBin0.0	0.502283105
accidentBin0.0	niceWeather1.0	0.256756757
niceWeather1.0	accidentBin0.0	0.527777778
accidentBin1.0	niceWeather0.0	0.762237762
niceWeather0.0	accidentBin1.0	0.497716895
accidentBin1.0	niceWeather1.0	0.237762238
niceWeather1.0	accidentBin1.0	0.472222222

Support level = .25

support_set	support_level
accidentBin0.0	0.508591065
accidentBin1.0	0.491408935
niceWeather0.0	0.75257732
accidentBin0.0niceWeather0.0	0.378006873
accidentBin1.0niceWeather0.0	0.374570447

From	To	conf_level
	accidentBin0.0	0.508591065
	accidentBin1.0	0.491408935
	niceWeather0.0	0.75257732
accidentBin0.0	niceWeather0.0	0.743243243
niceWeather0.0	accidentBin0.0	0.502283105
accidentBin1.0	niceWeather0.0	0.762237762
niceWeather0.0	accidentBin1.0	0.497716895

Section5: Hypothesis Testing

- **Overall Information:**

For this section, we come up with 4 hypotheses in total and apply 7 methods to them. Each hypothesis is tested by one or more method. In the following paragraphs, we will state our hypothesis, explain the applied method or methods, the reasons we chose one specific method, as well as the results of each test.

- **Preparation:**

Just like previous sections, we merged our two datasets and adjust the data in order to improve the performance of the later tests. We merged two cleaned datasets based on the specific date. That is, the very first attribute column is the data for year 2013 to 2016, and each row contains data of rainfall, snow, maximum temperature, minimum temperature, number of accidents, total number of injuries on that day. By the nature of our arrangement, the merged data has 1461 rows (in a length of 4 years). In addition, we also create a new attribute named 'injPERacci' which represents the number of injuries per accident and manually find the outliers in the merged data by sorting the attributes, and removed all the noises. Finally, we created bins for different attributes in order to better conduct each hypothesis test. "maxTbin" and "minTbin" are the bins for maximum and minimum temperature using equi-depth bin into three bins as low, medium and high and mark them by 1,2,3. "rainBin" and "snowBin" are bins for rainfall and snowfall for that day. "rainBin" is marked as 1 if the amount of rainfall on that day was greater than 0.1 inches and "snowBin" is marked as 1 if the amount of snowfall was greater than 0.1 inches. Next, "niceWeather" aggregates "rainBin" and "snowBin". This bin is 1 if any one of "rainBin" or "snowBin" is 1. So that this represent if that day was clear. "injuryBin", "rateBin" and "accidentBin" are bins for "totalInjury", "injPERacci" and "AccidentCount" by using equi-depth binning strategy into two bins and mark them as 0 meaning low and 1 meaning high.

- ***Hypothesis 1:***

There is a significant impact of weather condition on number of accidents.

Method: T Test

Reason:

For this hypothesis, we want to compare if there is a significant difference between the mean of accidents on a nice day and the mean of accidents on a rainy or snowy day. Since the variances of two populations are unknown, we are using Student's t test with assumption of equal variance between two samples. In particular, we want to test if the means of two populations, car accidents in a nice day and car accidents in a rainy or snowy day, are equal.

Procedure:

In this test, H_0 and H_a are:

H_0: The means of numbers of accidents are the same on both nice and bad weather days.

Ha: The means of numbers of accidents are not the same on nice and bad weather days.

Two columns from the merged and binned data are used in this test: 'niceWeather' and 'AccidentCount'. The 'nice weather' indicates that there is no rain or snow on a specific date. Number of Accident is count by number of rows in the car accident record dataset with a specific date. Then t test is applied to two groups of accident counts.

Result:

The result of the test is evaluated by p value. The turnout p value is 0.007428, which is small enough to reject H_0. Therefore, it can be concluded that there is a significant impact of weather condition on total number of accidents during a specific day. We also take a look at the means for both groups. The numbers of accidents on nice weather days has a mean of 52.25 and the numbers of accidents on bad weather days has a mean of 48.93. As a result, nice weather increases the number of car accidents per day.

- ***Hypothesis 2:***

There is a linear relationship between maximum temperature and number of car accidents.

Method: Linear Regression

Reason:

Since we believe that as temperature goes higher, people are more likely to drive to go outside, we expect to see a positive linear relationship between maximum temperature and number of accidents. Moreover, maximum temperature and number of accidents are continuous variables that they can be used in a linear regression model.

Procedure:

In this test, the independent variable(x) the mean maximum temperature among three stations for each day and dependent variable(y) is the number of accident each day, and therefore, the attributes used are 'meanMaxT' and 'AccidentCount'. The 'meanMaxT' attribute is calculated by getting the mean of maximum temperature in three weather stations with different locations. After that, x and y are fitted into a linear regression and the resultant slope, R^2 and p value are printed.

Result:

Slope = 0.224, p value = 3.19e-15, $R^2 = 0.0412$. $R^2 = 0.0412$ means that only 4.12% variation of maximum temperature can be explained by this linear model. This is too low to prove that there is a linear relationship between maximum temperature and number of accidents.

- ***Hypothesis 3:***

Among maximum temperature, minimum temperature, rainfall, snow and accident count, which combination of these factors affect number of injuries per accident the most?

Method: SVM and KNN (Lazy Learner Method)

Reason:

In predictive models, the goal is on model accuracy. We want to find a combination that has maximum accuracy in both training and test set. In this part, we want to find out factors affecting number of injuries per accident i.e. severity of accidents. There are many factors and the question is in high dimension. Therefore, classifiers are used to test the results instead of the hypothesis tests since they work better with high dimension data. Also, all the factors are numeric and continuous, and among the five classifiers required only SVM and KNN can deal with not binned data. As a result, SVM and KNN are used to get a test result.

Procedure:

For both SVM and KNN classification, we select only the factors and target variable and separate both into train and validate sets. The factors data is normalized after selection because they are not binned. The ratio I chose to separate them is 0.2: 0.8 for validate and train sets. After that, cross validation is done and accuracy is calculated. The final print outs are accuracy score, confusion matrix and classification report. Moreover, ROC curves are drawn. Note that this

procedure uses part of professor's code from Week 6 in class exercise. Finally, different attributes are tested in the above procedure and get accuracy score for each combination.

Result:

In these classifications, accuracy scores evaluate the test results of different combination. I tested almost all combinations and **some** of the results are as following:

Classify injury per accident by maximum temperature, rainfall and snow:
h3Test(myData,['meanMaxT','MeanRain','MeanSnow'], ['rateBin'])

```
SVM: 0.474226 (0.020554)
SVM Accuracy Score
0.498281786942
Confusion Matrix
[[  0 146]
 [  0 145]]
             precision    recall  f1-score   support

       -1.0       0.00      0.00      0.00       146
        1.0       0.50      1.00      0.67       145

avg / total       0.25      0.50      0.33       291
```

```
KNN: 0.515495 (0.035598)
KNN Accuracy Score
0.487972508591
Confusion Matrix
[[ 32 114]
 [ 35 110]]
             precision    recall  f1-score   support

       -1.0       0.48      0.22      0.30       146
        1.0       0.49      0.76      0.60       145

avg / total       0.48      0.49      0.45       291
```

Classify injury per accident by minimum temperature, rainfall and snow:
h3Test(myData,['meanMinT','MeanRain','MeanSnow'], ['rateBin'])

```
SVM: 0.474226 (0.020554)
SVM Accuracy Score
0.498281786942
Confusion Matrix
[[  0 146]
 [  0 145]]
             precision    recall  f1-score   support

       -1.0       0.00      0.00      0.00       146
        1.0       0.50      1.00      0.67       145

avg / total       0.25      0.50      0.33       291
```

```
KNN: 0.505172 (0.038535)
KNN Accuracy Score
0.501718213058
Confusion Matrix
[[ 38 108]
 [ 37 108]]
             precision    recall  f1-score   support

       -1.0       0.51      0.26      0.34       146
        1.0       0.50      0.74      0.60       145

avg / total       0.50      0.50      0.47       291
```

Classify injury per accident by rainfall, snow and accident count:
h3Test(myData,['MeanRain','MeanSnow','AccidentCount'], ['rateBin'])

```
SVM: 0.474226 (0.020554)
SVM Accuracy Score
0.498281786942
Confusion Matrix
[[  0 146]
 [  0 145]]
             precision    recall  f1-score   support

       -1.0       0.00      0.00      0.00       146
        1.0       0.50      1.00      0.67       145

avg / total       0.25      0.50      0.33       291
```

```
KNN: 0.505136 (0.015476)
KNN Accuracy Score
0.518900343643
Confusion Matrix
[[ 41 105]
 [ 35 110]]
             precision    recall  f1-score   support

       -1.0       0.54      0.28      0.37       146
        1.0       0.51      0.76      0.61       145

avg / total       0.53      0.52      0.49       291
```

Classify injury per accident by rainfall and snow:
h3Test(myData,['MeanRain','MeanSnow'], ['rateBin'])

```
SVM: 0.502557 (0.032526)
SVM Accuracy Score
0.494845360825
Confusion Matrix
[[144   2]
 [145   0]]
             precision    recall  f1-score   support

       -1.0       0.50      0.99      0.66       146
        1.0       0.00      0.00      0.00       145

avg / total       0.25      0.49      0.33       291
```

```
KNN: 0.506034 (0.033844)
KNN Accuracy Score
0.508591065292
Confusion Matrix
[[61 85]
 [58 87]]
             precision    recall  f1-score   support

       -1.0       0.51      0.42      0.46       146
        1.0       0.51      0.60      0.55       145

avg / total       0.51      0.51      0.50       291
```

Classify injury per accident by rainfall, snow, maximum temperature and accident count:
h3Test(myData,['MeanRain','MeanSnow', 'meanMaxT','AccidentCount'], ['rateBin'])

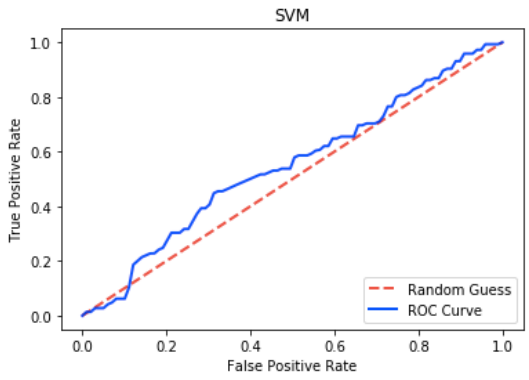
```
SVM: 0.585868 (0.040101)
SVM Accuracy Score
0.567010309278
Confusion Matrix
[[99 47]
 [79 66]]
             precision    recall  f1-score   support

       -1.0       0.56      0.68      0.61       146
        1.0       0.58      0.46      0.51       145

avg / total       0.57      0.57      0.56       291
```

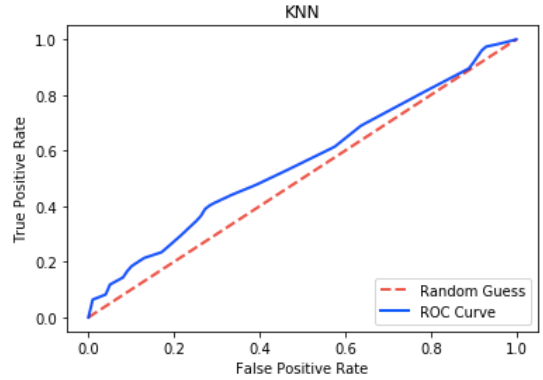
```
KNN: 0.570469 (0.043239)
KNN Accuracy Score
0.553264604811
Confusion Matrix
[[97 49]
 [81 64]]
             precision    recall  f1-score   support

       -1.0       0.54      0.66      0.60       146
        1.0       0.57      0.44      0.50       145

avg / total       0.56      0.55      0.55       291
```

However, the combination with **the highest accuracy score** is:
Classify injury per accident by rainfall, snow and accident count:
h3Test(myData,['MeanRain','meanMaxT','AccidentCount'], ['rateBin'])

```
SVM: 0.580703 (0.041837)
SVM Accuracy Score
0.567010309278
Confusion Matrix
[[99 47]
 [79 66]]
             precision    recall  f1-score   support

       -1.0       0.56      0.68      0.61       146
        1.0       0.58      0.46      0.51       145

avg / total       0.57      0.57      0.56       291
```

```
KNN: 0.575626 (0.042908)
KNN Accuracy Score
0.553264604811
Confusion Matrix
[[97 49]
 [81 64]]
             precision    recall  f1-score   support

       -1.0       0.54      0.66      0.60       146
        1.0       0.57      0.44      0.50       145

avg / total       0.56      0.55      0.55       291
```





From the last two tests, it is obvious that removing the attribute ‘MeanSnow’ does not change the accuracy much, therefore the last combination reflects information the most. All the examples of less accurate case shown above is similar to the best combination but with one attribute

difference, therefore it can show that the best combination do have the necessary amount and choice of attribute. From the ROC curves for SVM and KNN, we can see that they have the same overall shape, which means that they have similar accuracy and false positive rate. This can also be seen from the accuracy scores and the confusion matrices directly. The difference in accuracy scores is around 0.005 and the confusion matrices look almost the same. All the above information shows that SVM and KNN have about the same performance in this test. Also, it can be concluded that the combination of rainfall, maximum temperature and accident count has the most impact on number of injuries per accident.

- ***Hypothesis 4:***

Among high/median/low maximum temperature, high/median/low minimum temperature, high/low rainfall, high/low snow and nice/bad weather, which combination of these factors affect number of injuries the most?

Method: Decision Tree, Random Forest, Naïve Bayes

Reason:

This question seems similar to the question in Hypothesis 3, but there is a difference that all the factors in this question are binned. The three methods I chose to test this hypothesis are all work well with high dimension data and with binned data and target.

Procedure:

The procedure of this test is almost the same as hypothesis 3. The difference is that since all the factors are binned, the normalization process does not exist. The other part of the code and printouts are the same as hypothesis 3.

Result:

In these classifications, accuracy scores evaluate the test results of different combination. I tested almost all combinations and **some** of the results are as following:

Classify number of injuries by minimum temperature bin, maximum temperature bin and nice weather bin:
h4Test(myData,['minTbin','maxTbin','niceWeather'], ['injuryBin'])

```
Decision Tree: 0.595299 (0.054113)
Decision Tree Accuracy Score
0.54295532646
Confusion Matrix
[[82 53]
 [80 76]]
             precision    recall  f1-score   support

       -1.0       0.51      0.61      0.55       135
        1.0       0.59      0.49      0.53       156

avg / total       0.55      0.54      0.54       291
```

```
Random Forest: 0.590156 (0.047555)
Random Forest Accuracy Score
0.570446735395
Confusion Matrix
[[81 54]
 [71 85]]
             precision    recall  f1-score   support

       -1.0       0.53      0.60      0.56       135
        1.0       0.61      0.54      0.58       156

avg / total       0.58      0.57      0.57       291
```

```
Naive Bayes: 0.589287 (0.032575)
Naive Bayes Accuracy Score
0.591065292096
Confusion Matrix
[[87 48]
 [71 85]]
             precision    recall  f1-score   support

       -1.0       0.55      0.64      0.59       135
        1.0       0.64      0.54      0.59       156

avg / total       0.60      0.59      0.59       291
```

Classify number of injuries by maximum temperature bin and snow bin:
h4Test(myData,['maxTbin','snowBin'], ['injuryBin'])

```
Decision Tree: 0.581594 (0.038970)
Decision Tree Accuracy Score
0.59793814433
Confusion Matrix
[[103  32]
 [ 85  71]]
             precision    recall  f1-score   support

       -1.0       0.55      0.76      0.64       135
        1.0       0.69      0.46      0.55       156

avg / total       0.62      0.60      0.59       291
```

```
Random Forest: 0.581594 (0.038970)
Random Forest Accuracy Score
0.563573883162
Confusion Matrix
[[ 54  81]
 [ 46 110]]
             precision    recall  f1-score   support

       -1.0       0.54      0.40      0.46       135
        1.0       0.58      0.71      0.63       156

avg / total       0.56      0.56      0.55       291
```

```
Naive Bayes: 0.592706 (0.040881)
Naive Bayes Accuracy Score
0.59793814433
Confusion Matrix
[[103  32]
 [ 85  71]]
             precision    recall  f1-score   support

       -1.0       0.55      0.76      0.64       135
        1.0       0.69      0.46      0.55       156

avg / total       0.62      0.60      0.59       291
```

Classify number of injuries by rain bin and snow bin:
h4Test(myData,['rainBin','snowBin'], ['injuryBin'])

```
Decision Tree: 0.551540 (0.041687)
Decision Tree Accuracy Score
0.536082474227
Confusion Matrix
[[  0 135]
 [  0 156]]
             precision    recall  f1-score   support

       -1.0       0.00      0.00      0.00       135
        1.0       0.54      1.00      0.70       156

avg / total       0.29      0.54      0.37       291
```

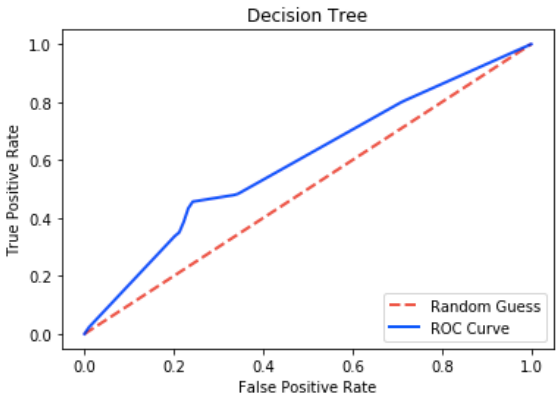
```
Random Forest: 0.551540 (0.041687)
Random Forest Accuracy Score
0.536082474227
Confusion Matrix
[[  0 135]
 [  0 156]]
             precision    recall  f1-score   support

       -1.0       0.00      0.00      0.00       135
        1.0       0.54      1.00      0.70       156

avg / total       0.29      0.54      0.37       291
```

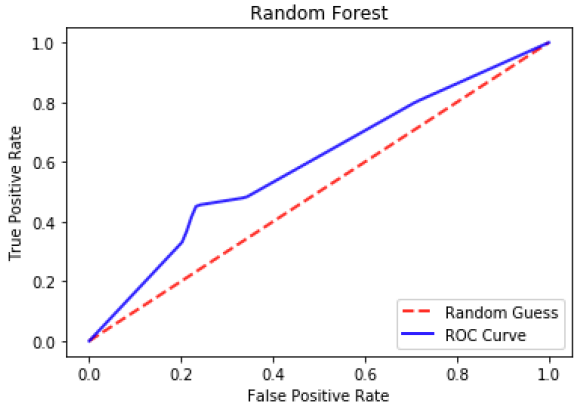
```
Naive Bayes: 0.457884 (0.032962)
Naive Bayes Accuracy Score
0.46735395189
Confusion Matrix
[[134   1]
 [154   2]]
             precision    recall  f1-score   support

       -1.0       0.47      0.99      0.63       135
        1.0       0.67      0.01      0.03       156

avg / total       0.57      0.47      0.31       291
```

However, the combination with **the highest accuracy score** is:
Classify number of injuries by minimum temperature bin, rain bin and snow bin:
h4Test(myData,['maxTbin','rainBin','snowBin'], ['injuryBin'])

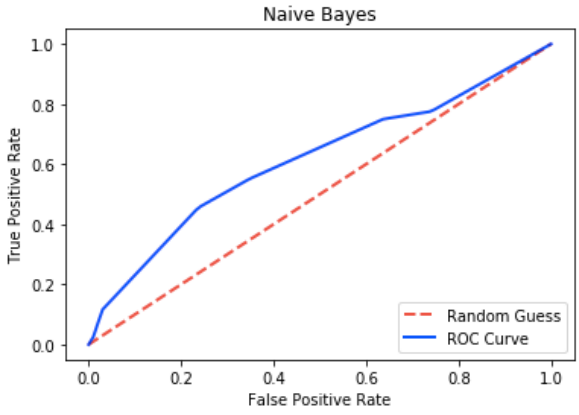
```
Decision Tree: 0.602210 (0.035512)
Decision Tree Accuracy Score
0.563573883162
Confusion Matrix
[[89 46]
 [81 75]]
             precision    recall  f1-score   support

       -1.0       0.52      0.66      0.58       135
        1.0       0.62      0.48      0.54       156

avg / total       0.58      0.56      0.56       291
```

```
Random Forest: 0.598762 (0.038452)
Random Forest Accuracy Score
0.563573883162
Confusion Matrix
[[89 46]
 [81 75]]
             precision    recall  f1-score   support

       -1.0       0.52      0.66      0.58       135
        1.0       0.62      0.48      0.54       156

avg / total       0.58      0.56      0.56       291
```






```
Naive Bayes: 0.587548 (0.040161)
Naive Bayes Accuracy Score
0.59793814433
Confusion Matrix
[[103  32]
 [ 85  71]]
             precision    recall  f1-score   support

       -1.0       0.55      0.76      0.64       135
        1.0       0.69      0.46      0.55       156

avg / total       0.62      0.60      0.59       291
```



All the examples of less accurate case shown above is similar to the best combination but with one attribute difference, therefore it can show that the best combination do have the necessary amount and choice of attribute. The differences of accuracy are not large from each other. From the ROC curves for Decision Tree, Random Forest and Naïve Bayes, we can see that Decision Tree and Random Forest have the same overall results, which makes sense because Random Forest is basically a group of bagged trees. This can also be seen from the accuracy

scores and the confusion matrices directly that two methods turn out exactly the same accuracy scores and confusion matrices. All the above information shows that Decision Tree and Random Forest have the same performance in this test. The Naïve Bayes method, however, shows a curvier ROC curve. Comparing the ROC curve between Decision Tree and the Naïve Bayes, it can be suggested that Naïve Bayes has a better performance in this test, which is also shown in the accuracy score. Also, it can be concluded that the combination of maximum temperature bin, rain bin and snow bin has the most impact on number of injuries per day.

Section6: Writeup

For this project, we started with the interests of studying the effect of weather conditions on car crashes during several years' period over different regions of the United States. Then, because of the limitation of data collection and time constrain, we finalized our study as investigating the relationship between the weather conditions and the number of motor vehicle crashes, injuries, and fatalities that occurred in Washington, D.C. between year 2013 and 2016. In our study, we used and analyzed the impact of weather conditions on the daily number of motor vehicle crashes, injuries, and fatalities. In addition, instead of studying the entire country, we plan to focus on Washington D.C. as a representation of the urban area, since we believe that there is a significant difference in results between urban area and rural area. We believe that it is meaningful to study this relationship because the explicit model we conduct would allow relative people to predict or forecast the number of car accidents based on the specific weather condition. For instance, if we were able to predict the specific increase of accidence number that would occur when the amount of rainfall in the future is predicted, then relevant people would be able to be prepared ahead. As a results, people such as police officers and emergency medical service personals would be better arranged accordingly to different daily weather condition.

During the phase of data collection, we obtained both of our data set from the official government websites, using *http://opendata.dc.gov/datasets/crashes-in-dc/data* for our daily car accidents data and using *https://www.ncdc.noaa.gov* for our daily weather condition data. We believe that the official government websites would provide relatively reliable and complete data set than other available places. These relatively reliable and complete data sets would allow us to conduct future data cleaning and analysis. Moreover, since the API for the data pages are available to the public, so the data is accessible compare to other sources. Overall, the car accident data we collected contains 28 attributes and weather condition data contains 4 attributes (one of the attributes is called "data type" that contains 15 types of data which are untimely serve as 4 different attributes after we rearrange and clean the data).

In general, for the original car accidents data set, it contains the attributes of the date of occurrence, the location of the accident, number of car involved in the accident, number of injury, and number of death. On the other hand, for the weather conditions data set, the variables contain date, location of the weather stations, minimum temperature, maximum temperature, precipitation, and snowfall amount. All the variables are useful since they would help us to have a better understanding of the relationship we want to investigate. First of all, the date and location variables allow us to match the weather condition with the car accidents. Then, the rest of the variables for car accidents data (number of car, number of injury, and number of death) could let us define or verified the scale of the accident, whether it is a major accident or a minor accident. Similarly, the rest of the variables for weather condition data (max/min temperature, precipitation, and snowfall) would allow us to interpret and estimate the specific weather

condition during that date. Based on the attributes of our two datasets, we first listed several possible hypotheses for our study:

- The weather conditions have great impact on the number of car crashes, injuries, and fatalities.
- Weather conditions and number of car crashes, injuries, and fatalities have a positive correlation.
- Higher temperature has a negative influence to drivers' condition.
- Extreme temperature has a positive correlation to number of accidents.
- When the rainfall amount increase, the number of car crashes, injuries, and fatalities also will increase.
- The rainfall has a more negative influence to drivers' condition than the snowfall.

After came up with the above hypotheses, before we do further analysis and conduct hypotheses text, we first need to clean our datasets. We decided to clean our data separately and merged both data for future analysis and testing after detect and fill in the missing values, detect and replace the incorrect values, and find out and remove the outliers. When checking the missing values, we simply just found out the empty entries for each attribute. When checking for the correctness for each attribute in the dataset, we manually chose the reasonable boundary, found out the entries that fall out of the limits, and defined it as an incorrect value. When finding the outliers, we used our modified IQR method, which replaced the IQR values by setting it equals the difference between 95 and 5 percentiles. The lower bound equals 1.5 modified IQR lower than the 5 percentile, and the upper bound equals 1.5 modified IQR larger than the 95 percentile. When fixing the data, we generally use the mean for that attribute to fill in the missing values and replace the incorrect entries, and for the outliers, we decided to remove all of them.

After finished all the data cleaning steps, we computed the basic statistical variables such as mean, median, and standard deviation for all the numerical attributes for both cleaned car accident data and cleaned weather condition data, in order to have an overall sense of the distribution for each attribute in the datasets. One significant result is that for the car accident data, most of the attributes, by their nature, have the median of 0 and mean and standard deviation close to 0 because most of the inputs for each attributes in car accident data are less then 10, and we have over 90,000 accident records, among which most of the records are 0.

In addition, we conducted clustering on our dataset separately as well. In the car accident and weather data set, we found out the silhouette coefficient is highest when number of clusters is 2. After conducting cluster analysis, in car accident data set, we found out that speeding would cause more injuries and more parties involved in an accident. In the weather data set, the clusters were done in maximum temperature, minimum temperature and rainfall amount on a daily basis. We found out there are two distinct two clusters, one is for summer temperature and the other one is for winter temperature and there was more raining in summer than winter.

Then, in order to have a better comparison between the car accident data and weather condition data, we decided to merged our two cleaned datasets into one big data which is based on the specific date. That is, the very first attribute column is the data for year 2013 to 2016, and each row contains data of rainfall, snow, maximum temperature, minimum temperature, number of accidents, total number of injuries on that day. By the nature of our arrangement, the merged data has 1461 rows (in a length of 4 years). Next, we plotted the histograms and scatterplots matrix for every attributes in the new merged dataset. From the histograms and scatterplots, we can see that the distributions of total number of accident and total number of injuries are similar

and both skewed to the right and as the number of accidents increase, the number of injuries would increase as well, which indicates the positively correlation between them. Moreover, we can see that most of the days has no rain or snow. This indicate that during the four years, the number of sunny day is significantly higher than the rainy and snowing day. Finally, the most obvious correlation is between the maximum and minimum average temperatures, and both have the distributions are both slightly skewed to the left. Also note that, most of the scatterplots do not show a clear pattern or a specific relationship among the attributes.

The last analysis before conducting hypothesis test is the association rule. From the supporting set, we concluded that weather and number of car accidents are not dependent. In confidence set, probability of accidentBin=0.0 given weather was nice is 0.5023 and probability of accidentBin=0.0 given weather was bad is 0.5278. probability of accidentBin=1.0 given weather was nice is 0.4977 and probability of accidentBin=0.0 given weather was bad is 0.4722. Those results give us a sense that the number of accidents is smaller on a rain or snow day which is consistence with the result of our first hypothesis test in the later section.

After all the previous analysis (basic statistical computations, histograms, scatterplots, clustering, association rule), we improved and updated our list of hypotheses as follows:

- There is a significant impact of weather condition on number of accidents.
- There is a linear relationship between maximum temperature and number of car accidents.
- Among maximum temperature, minimum temperature, rainfall, snow and accident count, which combination of these factors affect number of injuries per accident the most?
- Among high/median/low maximum temperature, high/median/low minimum temperature, high/low rainfall, high/low snow and nice/bad weather, which combination of these factors affect number of injuries the most?

For the first hypothesis, we wanted to investigate if the mean of two populations, car accidents in a nice day and car accidents in a rainy or snowy day, are equal. After conducting Student's t test, the p value of the test was 0.007428 which is so low that we reject H_0 that the two means are not equal. The numbers of accidents on nice weather days has a mean of 52.25 and the numbers of accidents on bad weather days has a mean of 48.93. We concluded that there is a statistical significant of the means between two groups that numbers of accidents on nice weather days is significantly higher than numbers of accidents on bad weather days. This result is pretty surprising because our initial guess is that raining and snowing would cause more road hazards that would cause more accidents. The following list contains some possible explanations of our result:

- Our data is car accidents only in District of Columbia where majority roads are streets with low speed limit. So that our data doesn't truly reflect car accidents on highways.
- During raining or snowing, drivers would usually pay extra attention on road conditions and traffic around them.
- On a rainy or snowy day, if it is during weekend, people would prefer to stay at home so that there would be less cars on road so that number of accidents would be smaller because less cars on road.

For the second hypothesis, we intend to find out the relationship between temperature and number of accidents. We expected to see a positive linear relationship between maximum temperature and number of accidents because we believe that as temperature goes higher, people are more likely to drive and go outside. Therefore, in this part, we tried to fit a linear regression model between maximum temperature and number of accidents. After fitting the regression line,

we obtained R^2 as 0.0412 which indicated that only 4.12% variation of maximum temperature can be explained by the linear model. Hence, we were not able to find a linear relationship between maximum temperature and number of car accidents.

We then conducted predictive analysis to test our third and fourth hypothesis. In predictive models, the goal is to compare the model accuracy and used the model with the highest accuracy level. We were trying to find a combination that has maximum accuracy in both training and test set. In our third hypothesis, we wanted to find out what factors affect number of injuries per accident the most. We used SVM (Support Vector Machine) and KNN (k-nearest neighbors) and after trying all possible combinations, we found out that rainfall amount, maximum temperature and number of accidents on that day are the factors affecting injuries per accident the most. Moreover, in the fourth hypothesis, we tried to evaluate what influence number of injuries the most. We used Decision Tree, Random Forest and Naïve Bayes to run the analysis and find out that maximum temperature, rain and snow influence number of injuries the most. In both hypotheses, the accuracies are about 0.6, which is a little bit low for predictive analysis and we found out that there almost every attribute affects number of injuries and injuries per accident. Weather is only a small factor of causing accident and our data only contains weather and accidents that we couldn't further evaluate other possible reasons of car accidents.