Jack Urso

CSC 364

Professor St. Clair

April 13th, 2021

Final Project (Part 2)

- Questions That Were Answered:

1.) What are the similarities between the weather and violent crime in Chicago?

- Why?
  - I was interested in knowing if there were any patterns between the weather and violent crime in Chicago. Plus, you are always hearing of more violent crimes on warmer days of the year, and I wanted to see if this was true. And if so, I wanted to look at the other attributes that may have contributed to violent crime as well.

2.) When looking at the time and weather, what is the probability of a homicide occurring?

- Why?
  - I wanted to see what attributes contribute to a higher probability of a homicide occurring. For example, if the day falls on a holiday, is there a higher chance of someone being murdered? Or what months correspond to a higher probability of homicide? Also, the algorithm used for answering this question could easily be used by the police to prioritize policing in certain situations.

- Summary of Original Dataset:

- Chicago Crime 2014 and 2015 (2 Datasets)
  - Contains all the cases that happened in that year. For each case ID, it contains the attributes pertain to that case, such as the case number, block, data/time, crime description, and location (latitude and longitude).
  - Link: https://data.world/publicsafety/chicago-crime

- Weather Data 2012-2017 (4 Datasets)
  - The weather data was split into 4 different datasets, including temperature, humidity, wind speed, and weather description.
  - Each dataset is an hourly weather report between the years of 2012-2017. For each hour of the day, it contains a reading of the measurement for each city in the United Sates, such as New York, Los Angeles, and Chicago obviously.
  - Link: https://www.kaggle.com/selfishgene/historical-hourly-weather-data?select=temperature.csv

- Holidays 2014-2015:
    - This was not an official dataset, but I still thought I should include it. It contains a list of holidays the happened between the years of 2014-2015 with the date.
    - Link: https://www.calendarpedia.com/holidays/federal-holidays-2015.html

➢ Preprocessing:

- Combining All the Original Datasets:
    - For all the original datasets, I had to combine into 1 time series dataset. For each date of the years between 2014-2015, I included the following attributes: day of year (1-365), day of week (1-7), average violent crime time (military time), holiday (0 or 1), temperature (Kevin), humidity (%), wind speed (mph), weather description, # of homicides, # of assaults, # of batteries, # of robberies, # of kidnappings, # of rapes, and total # of violent crimes.
    - This took a lot of time, roughly around 20 hours! I used many techniques for preprocessing these the datasets on excel, such as look ups, indexing, filtering, conversion, conditional statements, splitting/merging, and many excel functions.
    - For the weather data, I got the daily high for all the datasets besides the weather description dataset. For the weather description, I took the most common weather description for each day. And, with the crime dataset, I filter out all the violent crimes for each of the years. From the violent crimes, I found the number of specific types of violent crimes and the number of total violent crimes. And, for all the holidays, I put them in manually one at a time.
    - From this preprocessed dataset, I split up into 2 other preprocessed datasets for each of the algorithms.

- K-Means Dataset (Final Preprocessed Dataset):
    - Because K-Means requires continuous values, I had to convert all the categorical attributes into continuous values. Since most of my attributes were already continuous, it was mostly copy and pasting. The only attribute that I had to convert to continuous was the weather description. So, for the weather description, I generalized the weather into five different attributes, including fog, snow, rain, sunny, and cloudy. Thus, I basically converted the weather description into a binary value assigned to one of attributes. Lastly, I omitted any time series attributes and the holiday attribute because I was just looking at weather and crimes.
    - Besides performing preprocessing in excel, I also perform some of the preprocessing in code for the algorithm. Because I did not want the algorithm favoring any specific attributes, I normalized my data by putting it on a scale between 0 and 1. Also, my algorithm was a little weird with the attribute names, so I also changed them within the algorithm.

- Bayesian Probability Dataset (Final Preprocessed Dataset)
    - Although you can perform Bayesian probability with continuous values, I did it with categorical values instead. I wanted the dataset to look like the example dataset we went over in class. So, I converted all the continuous values into categorical values. The time attributes where easily converted to categorical attribute by conditional statements. But, for the weather attributes, I had to sort all of them in order and find ranges that I could classify as low, medium, and high. Lastly, for the target class, there were only two values, false or true. If a

homicide occurred, then it would be true. Since I was mostly looking at the time and weather attributes, I did not have to include the violent crime attributes except for the homicide one.

- I did not perform any preprocessing in the algorithm itself. However, I only include the values and not the attribute names.

➢ Data Mining:

- K-Means:

  - Why K-Means:

    + K-Means is good at finding groupings and similarities within a dataset, so it would be a great algorithm to implement to see how the weather correlates to violent crime. Since most of my attributes are continuous, K-Means would require less preprocessing and the algorithm is a lot easier to implement compared to others.

  - Implementation:

    + I used the 'Jupyter Notebook (Anaconda 3)' to implement K-Means with. It allowed me to view my output as I was coding and work multiple sections of code at a time. In addition, the environment was able to easily output graphs and tables without any problems.

    + Referenced Code (2 Links):

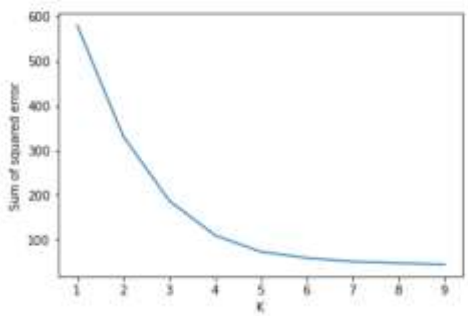
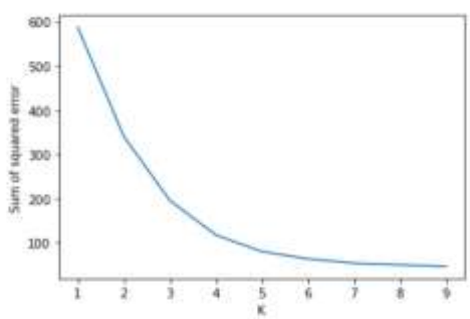
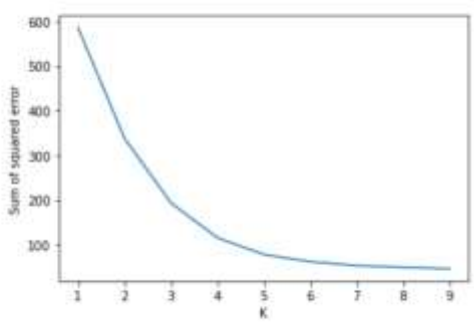
    https://www.kaggle.com/funxexcel/p1-sklearn-k-means-example
    https://github.com/codebasics/py/blob/master/ML/13_kmeans/13_kmeans_tutorial.ipynb

  - Steps:

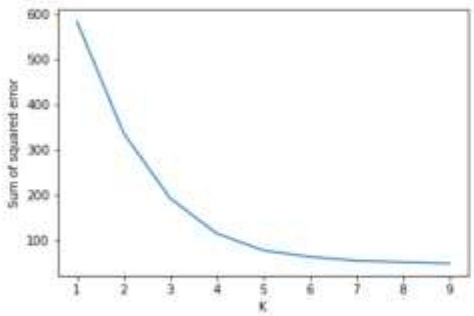
    1.) Read in the dataset.
    2.) Preprocessing the dataset:
        - Changing attribute names.
        - Normalizing the dataset using MinMaxScaler() [0-1 Scale]
    3.) Run K-Means Algorithm 7 times (each subset):
        - Use the "elbow method" to find k value.
        - Then,
        a.) With the given 'k' value, randomly choosing an initial centroid for each cluster.
        b.) Assign each observation to its nearest centroid.
        c.) Update the centroids as being the center in their respective observation.
        d.) We repeat steps B-C over and over until there is no further change in the clusters.
    4.) Print all the centroids.
        - Where I performed the cluster analysis.

- Results:

  + Here are all the elbow graphs and the clusters that were associated with the different violent crime attributes.

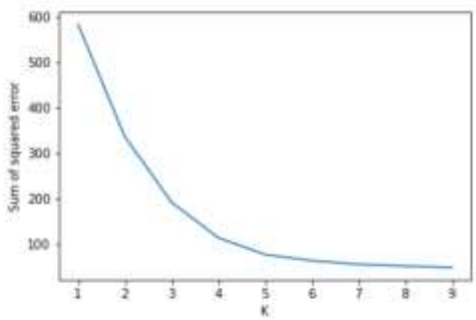
Elbow Graphs:	Clusters (Centroid Values)
Number of Homicides:  	Cluster 1: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Homicides 0.65113, 0.93982, 0.34576, -0.0, -0.0, 1.0, 0.0, 0.0, 0.11947, Cluster 2: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Homicides 0.58725, 0.90289, 0.28463, -0.0, 0.0, 0.0, 1.0, -0.0, 0.12948, Cluster 3: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Homicides 0.6043, 0.88054, 0.33779, -0.0, 0.0, 0.0, 0.0, 1.0, 0.11675, Cluster 4: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Homicides 0.63765, 0.95553, 0.32773, 0.72941, 0.27059, 0.0, 0.0, 0.0, 0.14118,
Number of Assaults:  	Cluster 1: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Assaults 0.6043, 0.88054, 0.33779, -0.0, 0.0, 0.0, 0.0, 1.0, 0.40738, Cluster 2: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Assaults 0.58725, 0.90289, 0.28463, -0.0, 0.0, 0.0, 1.0, -0.0, 0.425, Cluster 3: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Assaults 0.65113, 0.93982, 0.34576, -0.0, -0.0, 1.0, 0.0, 0.0, 0.44922, Cluster 4: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Assaults 0.63765, 0.95553, 0.32773, 0.72941, 0.27059, 0.0, 0.0, 0.0, 0.39589,
Number of Batteries:  	Cluster 1: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Battery 0.63765, 0.95553, 0.32773, 0.72941, 0.27059, 0.0, 0.0, 0.0, 0.38078, Cluster 2: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Battery 0.58725, 0.90289, 0.28463, -0.0, 0.0, 0.0, 1.0, -0.0, 0.40712, Cluster 3: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Battery 0.65113, 0.93982, 0.34576, -0.0, -0.0, 1.0, 0.0, 0.0, 0.4018, Cluster 4: ---------- Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Battery 0.6043, 0.88054, 0.33779, -0.0, 0.0, 0.0, 0.0, 1.0, 0.38248,

Number of Robberies:



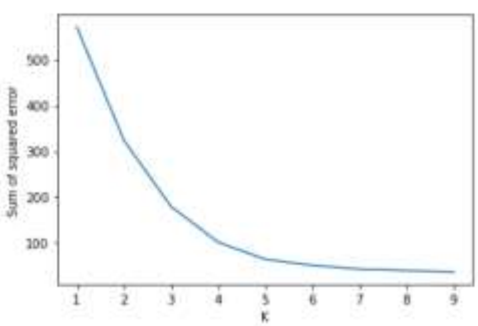

```
Cluster 1:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Robbery
0.63765, 0.95553, 0.32773, 0.72941, 0.27059, 0.0, 0.0, 0.0, 0.40071,
Cluster 2:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Robbery
0.58725, 0.90289, 0.28463, -0.0, 0.0, 0.0, 1.0, -0.0, 0.38894,
Cluster 3:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Robbery
0.6043, 0.88054, 0.33779, -0.0, 0.0, 0.0, 0.0, 1.0, 0.3933,
Cluster 4:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Robbery
0.65113, 0.93982, 0.34576, -0.0, -0.0, 1.0, 0.0, 0.0, 0.39558,
```

Number of kidnappings:



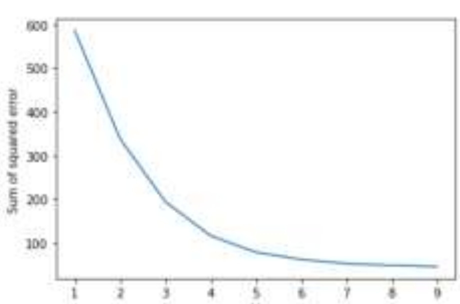

```
Cluster 1:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Kidnapping
0.58725, 0.90289, 0.28463, -0.0, 0.0, 0.0, 1.0, -0.0, 0.11246,
Cluster 2:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Kidnapping
0.6043, 0.88054, 0.33779, -0.0, 0.0, 0.0, 0.0, 1.0, 0.12315,
Cluster 3:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Kidnapping
0.65113, 0.93982, 0.34576, -0.0, -0.0, 1.0, 0.0, 0.0, 0.09204,
Cluster 4:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Kidnapping
0.63765, 0.95553, 0.32773, 0.72941, 0.27059, 0.0, 0.0, 0.0, 0.12,
```

Number of Rapes:




```
Cluster 1:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Rapes
0.6043, 0.88054, 0.33779, -0.0, 0.0, 0.0, 0.0, 1.0, 0.07512,
Cluster 2:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Rapes
0.58725, 0.90289, 0.28463, -0.0, 0.0, 0.0, 1.0, -0.0, 0.08718,
Cluster 3:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Rapes
0.65113, 0.93982, 0.34576, -0.0, -0.0, 1.0, 0.0, 0.0, 0.07603,
Cluster 4:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, Rapes
0.63765, 0.95553, 0.32773, 0.72941, 0.27059, 0.0, 0.0, 0.0, 0.0877,
```

Number of Violent Crimes:




```
Cluster 1:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, VCrimes
0.59373, 0.9489, 0.35242, -0.0, 0.16912, 0.83088, 0.0, 0.0, 0.46537,
Cluster 2:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, VCrimes
0.58725, 0.90289, 0.28463, -0.0, 0.0, 0.0, 1.0, -0.0, 0.49273,
Cluster 3:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, VCrimes
0.6043, 0.88054, 0.33779, -0.0, 0.0, 0.0, 0.0, 1.0, 0.46802,
Cluster 4:
----------
Temp, Humidity, Wind, Fog, Snow, Rain, Sunny, Cloudy, VCrimes
0.75855, 0.94145, 0.30645, 1.0, -0.0, 0.0, 0.0, 0.0, 0.5205,
```

- Analysis from the Results:

  + Firstly, there tends to be a correlation between temperature and violent crime. If the temperature is higher, then there tends to be more violent crime in general. This is observable when you are looking at the number of assaults.

  + With the humidity, higher measurements have a slight correlation with increase violent crime. When looking at robberies, there is a noticeable relation between the attributes. Since one of the depends to humidity is temperature, we could be looking at a relationship between the temperature and violent crime.

  + When looking at the wind-speed, I do not see any relationship between the measurement and violent crimes. But, with kidnapping, it looks like a higher wind-speed leads to a lower number of kidnappings.

  + With the weather attributes (Fog, …, Cloudy) values, they tend to be all over the place and I was not able to get any correlation with the violent crime. This is probably due to having to generalizing the weather description.

- Bayesian Probability:

  - Why Bayesian Probability:

    + Bayesian probability is the best algorithm for finding the probabilities within a dataset. Therefore, it would be perfect for my data and finding the probability of a homicide occurring given certain attribute values. In addition, it is also simple to implement and easy to understand compared to using a neural network.

  - Implementation:

    + For the Bayesian Probability algorithm, I did not have to use any special environment or software to run my code. Simply, I just used the regular PyCharm IDE to run my algorithm on and it worked fine.

    + Referenced Code (1 Link):

    https://gist.github.com/TejeshJadhav/aac97170a0b4f2bc30276880c36b5ea6

- Steps:
  1.) Performs Bayesian Probability given the attribute values from the user (Option 1):
      - Asks the user for the time and weather attributes. And, whether to print the values from the algorithm.
      - Call the Naive Bayes function with the given above values.
      a.) Reads the dataset.
      b.) Finds the necessary values to perform naive bayes, such as the prior probability, likelihood probability, and the normalizing constant.
      c.) Lastly, the function returns an array with, such that [P(Homicide = false | X), P(Homicide = true | X)].

  2.) Or, performs Bayesian Probability on different combinations of attribute values from a specific attribute value (Option 2):
      - Reads the dataset.
      - Populates every possible value with every column.
      - Populates all the values in each column.
      - Finds the average Bayesian Probability of every random combination of attribute values for a certain number of iterations.
      * This code is exclusively for perform data analysis on the Bayesian Probability results.

- Results:

  + Here is an example of the first part of the algorithm. By me showing you an example, it will give you a better idea on how the algorithm works. And it will give you a better idea on how the results were gathered in the second part of the algorithm.

Output:

Enter Month, Day, Time, Holiday, Temp, Humidity, Wind, Weather to determine Whether a homicide occurs:
jan,sun,morning,FALSE,cool,med,med,mist

Prior Probability, P(A):
P(Homicide = true) = 0.6835616438356165
P(Homicide = false) = 0.31643835616438354

Likelihood Probabilities, P(C|A):
P(Month = jan | Homicide = true) = 0.06813627254509018
P(Month = jan | Homicide = false) = 0.12121212121212122
P(Day = sun | Homicide = true) = 0.1623246492985972
P(Day = sun | Homicide = false) = 0.09956709956709957
P(Time = morning | Homicide = true) = 0.08817635270541083
P(Time = morning | Homicide = false) = 0.08658008658008658
P(Holiday = FALSE | Homicide = true) = 0.9719438877755511
P(Holiday = FALSE | Homicide = false) = 0.9696969696969697
P(Temp = cool | Homicide = true) = 0.14228456913827656
P(Temp = cool | Homicide = false) = 0.3203463203463203
P(Humidity = med | Homicide = true) = 0.24448897795591182
P(Humidity = med | Homicide = false) = 0.19913419913419914
P(Wind = med | Homicide = true) = 0.7294589178356713
P(Wind = med | Homicide = false) = 0.7229437229437229
P(Weather = mist | Homicide = true) = 0.0781563126252505
P(Weather = mist | Homicide = false) = 0.06060606060606061

Normalizing Constant, P(C):
P(C) = 2.487230482709409e-06

Probabilities, P(C|A)*P(A):
P(X | Homicide = true)*P(Homicide = true) = 1.2850376187962756e-06
P(X | Homicide = false)*P(Homicide = false) = 8.961716895232569e-07

Posterior Model, P(A|C) = (P(C|A)*P(A))/P(C):
P(Homicide = true | X) = 0.5166540164771737
P(Homicide = false | X) = 0.3603090649432023

+ Now, I am going to show you my results from the second part of the algorithm. It gives you better visualization on how probabilities are related to the attribute values. I ran 500 iterations on each attribute value, and it had an execution time of 3 minutes and 45 seconds. If I would have checked all the different combinations (not randomly), it would have taken 23 minutes and 30 seconds.

<u>Month</u>:	<u>Day</u>:	<u>Time</u>:	<u>Holiday</u>:
When Month = jan: T: 0.5432020808557333 F: 0.45660858649797165 When Month = feb: T: 0.4849664964450578 F: 0.509450919354264 When Month = march: T: 0.5506981985398132 F: 0.4386282218635036 When Month = april: T: 0.634051287330271 F: 0.36689932045955465 When Month = may: T: 0.7298903497934579 F: 0.26892591433347157 When Month = june: T: 0.8225144803535551 F: 0.18154634159178612 When Month = july: T: 0.7654609673753276 F: 0.23748582050391429 When Month = aug: T: 0.8262894206979691 F: 0.1749535192929943 When Month = sept: T: 0.8156671842778753 F: 0.18343113073198963 When Month = oct: T: 0.6955040030902706 F: 0.30610551072813263 When Month = nov: T: 0.6534224449384758 F: 0.34807702659856243 When Month = dec: T: 0.6665606085110268 F: 0.3333715669013893	When Day = mon: T: 0.6549355289923862 F: 0.3545427108592495 When Day = tues: T: 0.6215100651543048 F: 0.37318992801461665 When Day = wed: T: 0.6221577929007917 F: 0.3821646837716259 When Day = thur: T: 0.6438576326152295 F: 0.35586938896447595 When Day = fri: T: 0.7042407647113689 F: 0.29184002381751856 When Day = sat: T: 0.7784239856440556 F: 0.2264752959554485 When Day = sun: T: 0.7718886832202554 F: 0.22485894458695097 When Day = sun: T: 0.7717452086064622 F: 0.22859716004110214	When Time = morning: T: 0.6843766285238713 F: 0.3158950462835364 When Time = evening: T: 0.6919243096402201 F: 0.30571892920587185	When Holiday = true: T: 0.6729719274262933 F: 0.3251383605775533 When Holiday = false: T: 0.6986546753301699 F: 0.30181105109466416

Temperature:	Humidity:	Wind Speed:	Weather Description:
When Temp = cool: T: 0.4889805029761455 F: 0.5203620605958384 When Temp = med: T: 0.7165654090652018 F: 0.28912957116951965 When Temp = hot: T: 0.8055577116192295 F: 0.19495178931444915	When Humidity = low: T: 0.663742643395671 F: 0.33537479259863257 When Humidity = med: T: 0.7145021464996203 F: 0.2824326388714702 When Humidity = high: T: 0.7052616334232921 F: 0.2967060537398226	When Wind = weak: T: 0.7014827115515998 F: 0.30080632652963124 When Wind = med: T: 0.6983034314428368 F: 0.30622893282087665 When Wind = strong: T: 0.6888006035972928 F: 0.30760240384345927	When Weather = broken clouds: T: 0.6223079970301021 F: 0.3832417918031694 When Weather = few clouds: T: 0.6883259281601826 F: 0.30557100500574685 When Weather = fog: T: 0.7462043572525269 F: 0.25093065850704716 When Weather = haze: T: 1.0248844428100314 F: 0.0 When Weather = heavy intensity rain: T: 0.8390463641658715 F: 0.15473058980473653 When Weather = light rain: T: 0.6418803152239801 F: 0.36228522577075667 When Weather = light snow: T: 0.5822104750788625 F: 0.4196347425773645 When Weather = mist: T: 0.7310984081788544 F: 0.2660163515086845 When Weather = moderate rain: T: 0.8277102860803767 F: 0.17213990743629026 When Weather = overcast clouds: T: 0.6316057491675583 F: 0.3696516346120241 When Weather = scattered clouds: T: 0.7384361412783472 F: 0.2658653365572203 When Weather = sky is clear: T: 0.7010688304007228 F: 0.30090107487258655 When Weather = snow: T: 0.5498224501317128 F: 0.44544894369856286

- Analysis of Results:

  + When looking at the months, there is a noticeable difference between the probabilities. The probability of a homicide occurs goes up during the warmer months and goes down with colder months. The month of June has the highest probability while the month of February has the lowest probability.

  + Like the months, there is an observable difference between the probabilities. During the weekend, there is a higher probability of a homicide occurring. Saturday has the highest probability while the Tuesday has a lowest probability.

  + There is not a noticeable difference between the probabilities between a homicide occurring during the morning and evening. However, from the results, it looks like the evening has a slight increase on the probability.

+ Surprisingly, there is a slight decrease in probability of a homicide occurring on a holiday compared to a non-holiday. I thought there was going to be a major difference between the probabilities.

+ Like the K-Means clustering results, there is an observable difference between temperature and whether a homicide occurs. The higher probability corresponds to a hotter temperature while the lower probability corresponds to a cooler temperature.

+ Again, like the K-Means clustering results, there is not a major difference between humidity probabilities and the wind speed probabilities. There are only slight differences between the attribute values.

+ Shockingly, there are many differences between the attribute values and their probabilities. First, since there were not many haze values in the dataset, you should ignore this attribute value. Strangely, there is a higher probability of a homicide occurring when it is raining compared to when the sky is clear; in fact, moderate rain and heavy intensity rain has the highest probabilities. In addition, it looks like snow and light snow have the lowest probabilities of a homicide occurring.

- Conclusion:

  - Ultimately, I think I got some pretty good results from both algorithms. However, I felt like I got better results from the Bayesian Probability compared to K-Means. This is probably because Bayesian Probability was a little bit more intuitive and easier to implement. Also, I thought the results from Bayesian Probability are a lot easier to interpret compared the results from K-Means. Also, Bayesian Probability help me understand the K-Means results a little bit better.
  - Overall, I am completely satisfied with my work, including the preprocessing, algorithm implantation, and results. If I were grading myself on a scale of 10, I would give myself a 10 on preprocessing, 10 on algorithm implementations, and an 8.5 on results. In the end, I believe the I was able to answer both questions successfully.

- Final Preprocessed Datasets and Code:

  - Gmail Account Contains Everything:
    Username: csc364finalproject@gmail.com
    Password: [redacted]mers
  - Since I wrote most of the code myself, I suggest you look at the project folders located in the google drive. From the links I mentioned above, I only used a small amount of code from the those and most of the code was written by me. For the K-Means algorithm, I ran it with ‘Jupyter Notebook (Anaconda 3)’. For the Bayesian Probability algorithm, it was just implemented with a simple Python file. So, you can use anything you want for that.
  - If you have any trouble, just email me at jrurso@noctrl.edu.