Analysis & Insights

Submitted by:

Neha A Varshney

(Varshney.n@husky.neu.edu)

Purpose

- This Presentation is to provide the brief overview of my analysis, answering the given questions
- I understand that, mostly the business communicates outlines via the presentations.
- The detailed answers with code snippets are attached in the jupyter notebook

Problem Definition

- The city of Seattle only receives 911 calls for four reasons
- a hot latte spills all over your lap
- Beavers attack i.e unsuspecting passersby's (watch out for those beavers!)
- Seal attacks (can't be too careful)
- Marshawn Lynch sightings (people get very excited and choose to call 911 for some reason).

Objective

- The objective is to perform analysis on this data set and extract insights answering the given questions.
- The technical language used: Python 3.7
- Libraries used : pandas, numpy, pylab, scikit-learn, matplotlib, seaborn, plotly
- Platform: Jupyter Notebook

- **QUESTION 1:**

1.A: What is the most common reason for calling 911?

1.B: Display these results graphically

Answer 1.A :

- The reason for which maximum number of calls are made, will be the most common reason to call 911.
- Here, on performing value count on the type feature, the maximum count i.e 508 is for ***the Beaver Accident***.

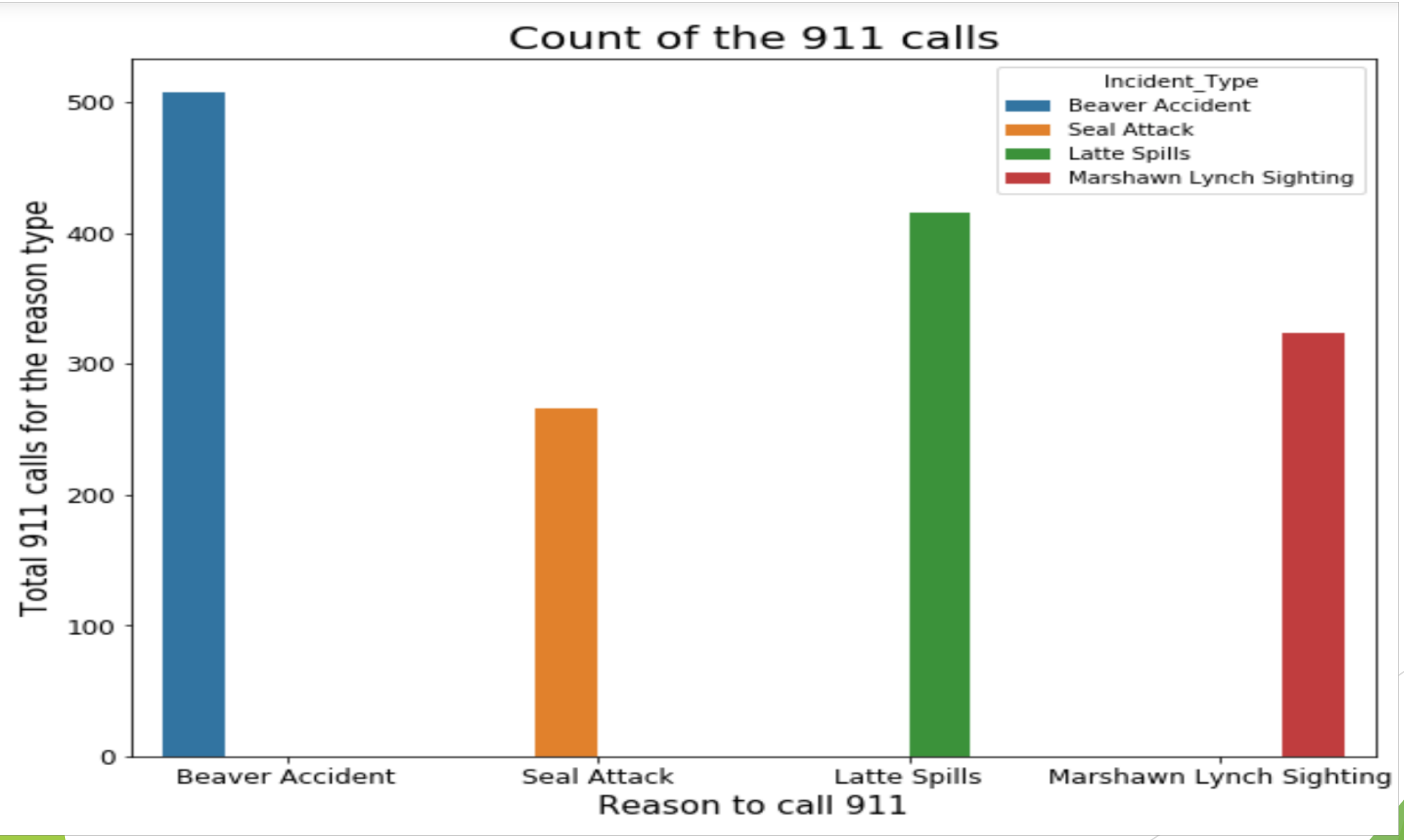
```
Printing the count of reasons, i.e total reports under each reason


Beaver Accident              508
Latte Spills                 416
Marshawn Lynch Sighting      324
Seal Attack                  266
Name: Incident_Type, dtype: int64


Normalizing the value count as rate, to get the clear and relative understanding


Beaver Accident              0.335535
Latte Spills                 0.274769
Marshawn Lynch Sighting      0.214003
Seal Attack                  0.175694
Name: Incident_Type, dtype: float64
```

Answer 1.B:


Count of the 911 calls
Incident_Type
Beaver Accident
Seal Attack
Latte Spills
Marshawn Lynch Sighting
Total 911 calls for the reason type
0
100
200
300
400
500
Beaver Accident
Seal Attack
Latte Spills
Marshawn Lynch Sighting
Reason to call 911

- Question 2

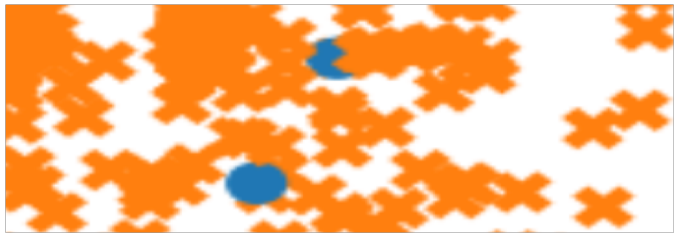
2.A: Please create a graph of the 911 calls using the 'Latitude' and 'Longitude' (graph type is up to you) (differentiate call type using colors)
2.B: Are there any data points that look mislabeled?

Answer 2.A:

- Created the scatter plot with x axis as Incident_Latitude and the y axis as Incident_Longitude
- Scatter plots show how much one variable is affected by another i.e here, the relation of reason of the calls to latitiude and longitude of the incident.
- Incident_Type i.e reason type as the hue value
- Please refer to figure titled: « *911 Calls (Latitude & Longitude) differentiated by call type*
- This provides the insight that the 911 call types are segmented across the location(latitude and longitude)

Answer 2.B:

- In the scatter plot for answer 2.A, there are few data points present in the segment (cluster) of a different color (for example)



- Here, blue data points are potentially mislabeled.
- As the color of the cluster represent the call type, therefore Yes, there are potentially mislabeled data points
- *If there is more information about the data, I can do further interpretations regarding these data points*
- Please refer to plot with title: *911 Calls (Latitude & Longitude) differentiated by call type*

Question 3

3.A: If we were to use only 'Latitude' and 'Longitude', could we make an intelligent decision as to why a resident dialed 911? (In other words, if we take off the labels - can we still determine which category a 911 call would most likely fall into?) Please describe this algorithm and your reason for choosing it.
3.B: Does the algorithm chosen utilize Euclidean distance? Should we be concerned that 'Latitude' and 'Longitude' are not necessarily Euclidean?
3.C: Please display the results of your algorithm, along with the associated code
3.D: Please display the number of correct categorizations
3.E: What insight can we extract from this analysis?

Answer 3.A:

- The segmentation of call type as seen in the question 2.A, clearly indicates that latitude and longitude can be used to to decide on the reason of the call.
- For example, I am a 911 responder, if I know that calls from location A are due to beaver accident.
- Now, I got a new call which is from the next building of location A. It will be very probable that this call is for beaver accident too.
- Also, in the analysis it is seen that:
- In case when, multiple calls are made from one location, they have reported the same reason i.e **location as a proxy of attack(reason)**
- Therefore, if we take off the labels - can we still determine which category a 911 call would most likely fall into.
- **Algorithm : *K-Means Clustering***
- **Reason being, The Variance(spread of information) of a reason of the call is mostly retained in a cluster, i.e in a segment**

Answer 3.B:

- k-means generates clusters based on the Euclidean distance between points— meaning the straight-line distance between two pins in the map.

- But as we know, the Earth isn't flat so this approximation will affect the clusters being generated

- Instead, we should be using Geographical (spatial) distance i.e the distance measured along the surface of the earth i.e calculating lengths of the shortest curve between two points along the surface of the Earth.

- Hierarchical clustering, PAM, CLARA, and DBSCAN are the popular examples of using spatial distances.

Answer 3.(C,D)

- The associated code and visualizations are described in the notebook
- The algorithm uses the k means method of the scikit-learn library
- Here, the objective is to categorize calls on the basis of latitude and latitude and as there are 4 different type of calls.
- Therefore, number of clusters are taken to be 4
- Otherwise, Elbow plot helps to discover the cluster capturing the maximum variance.

Answer 3.E , elaboration on the next slide

- The Statistical Analysis of the K-Means shows that:

```
Beaver Accident
0    499
1      6
3      2
2      1
Name: Cluster, dtype: int64
Latte Spills
1    416
Name: Cluster, dtype: int64
Marshawn Lynch Sighting
2    258
1     47
3     19
Name: Cluster, dtype: int64
Seal Attack
3    240
2     18
1      5
0      3
Name: Cluster, dtype: int64
```

Wrongly classified values i.e categorised in a different cluster

Answer 3.E

- According to the stats, the potential cluster for:
- Beaver Accident is 0, Latte Spills is 1, Marshawn Lynch Sighting is 2, Seal Attack is 3
- 15% of Marshawn Lynch Sighting are categorized as Latte Spills
- This is also supported by the, scatter plot in 2.A, there is a overlap between red and green i.e Latte Spills and Marshawn Lynch Sighting

- Latte Spills is the cluster with least spread out distance
- To infer with the problem statement, in come cases people called out to 911 due to latte spills but maybe did not close the reason for some reason. Or this maybe a error in dataset. Cross Checking with business person is the best idea here.
- Seal Attacks are also some times reported as the Marshawn Lynch Sighting

- Getting more data and analyzing, I will be able to provide more probable insights on this.

- *THANK YOU!*