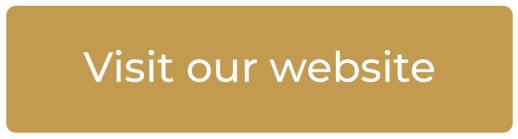


**TASK**

**Exploratory Data Analysis on the \*\*\*\*\*\*\*\*\*\*\*\* Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Summary of the data set

The data set at hand consists of 3 columns namely small\_area , police\_precinct, weight, and a handful of records. It is about the weight of people in small areas.

**DATA CLEANING**

# SUMMARY OF THE METHODS AND VISUALIZATIONS DONE DURING DATA CLEANING

-Renamed the ‘Small Location’ and printed out the description of the data set and it displayed the changes a success.

-Printed out the summary of static of the data set and it displayed the count, mean, std, min, 25%, 50%, 75% and max of the ‘Small Location’ and the weight as listed on the static table.

-Calculated the skewness and kurtosis of the ‘Small Location’ feature and discovered that the skewness is 0,783857 and the Kurtosis is -1,163613.

-Replaced all the categorical variables with actual values which is the ‘Police Precinct’.

-Got the dummies for the ‘Small Location’ in the data set.

MISSING DATA

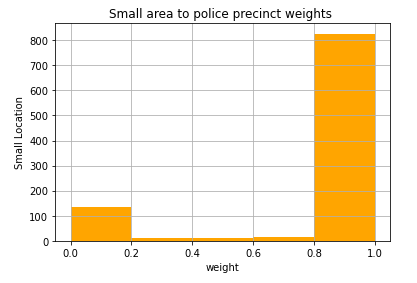
# ANY MISSING DATA? HOW DID YOU HANDLE IT

When I was in a process of cleaning my data set I used the built-in function to find the missing data of the data set that we are working on , and then after that called it to display ‘True’ if there is missing data and ‘False’ if there is no missing data. The results output that there was no ‘True’, meaning that the data set doesn’t consist of any missing data. I then used the isnull() built-in function and the sum() built in function which will collaborate and output the total number of data missing in the data set, just to be sure that there is no missing data for sure.

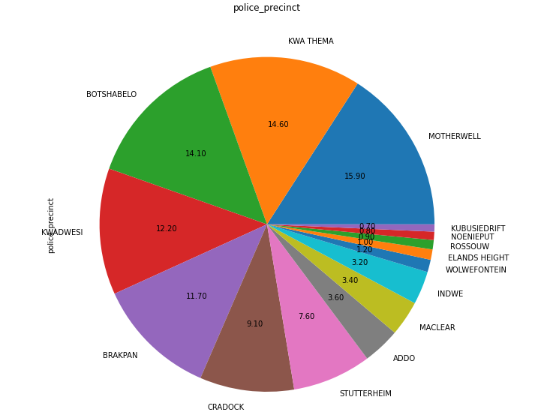
DATA STORIES AND VISUALIZATIONS

# THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALIZATIONS OF THE DATA  
First: Plotted a bar graph to visualize the outcome of the weight in small areas of the graph and founded that the most frequent weight is between 8.0 to 10.

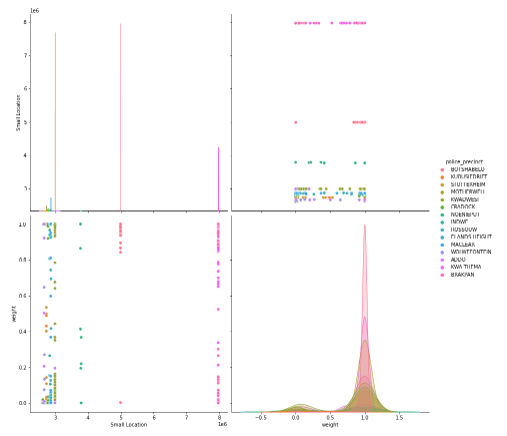
Telling us that most police precinct weigh between 8.0 to 10.



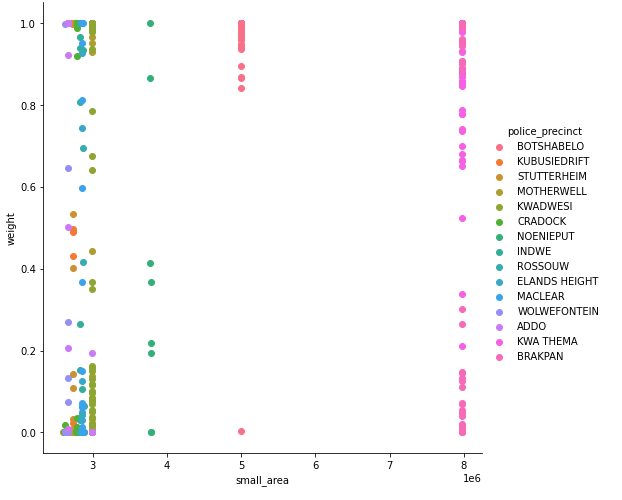
Second: Plotted a Pie chart to view which areas are involved more and came to conclusion of that a small area called ‘Motherland’ is more frequent in this survey and most of the police precinct weight is from the area ‘Motherland’ therefor this area has the highest impact on the results.



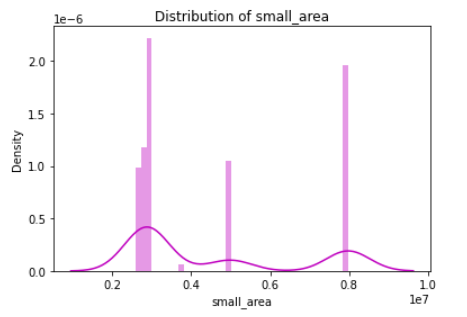
Third: Plotted the pair-plot to display the data set and the relationships at action.



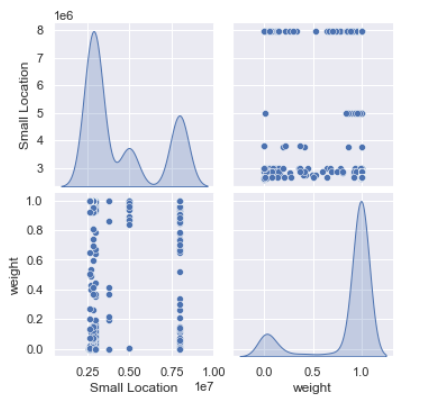
Fourth: Plotted the Face-grid for the data, coloured it and added a legend to it so it could display the relationship that is taking place in the data set.



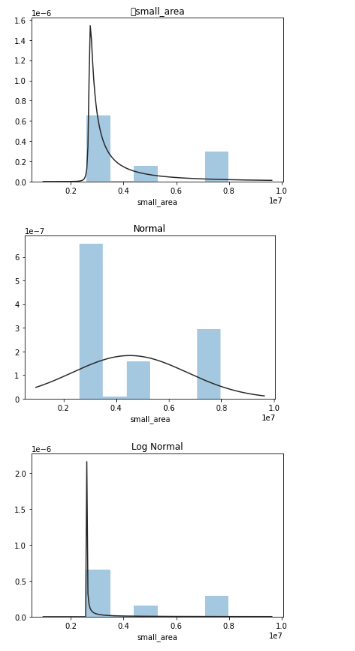
Sixth: Plotted the dist-plot to display the number of small\_area observation for a categorical variable using bars,It displayed that the small areas between 0,2 and 0,4 had the highest density.



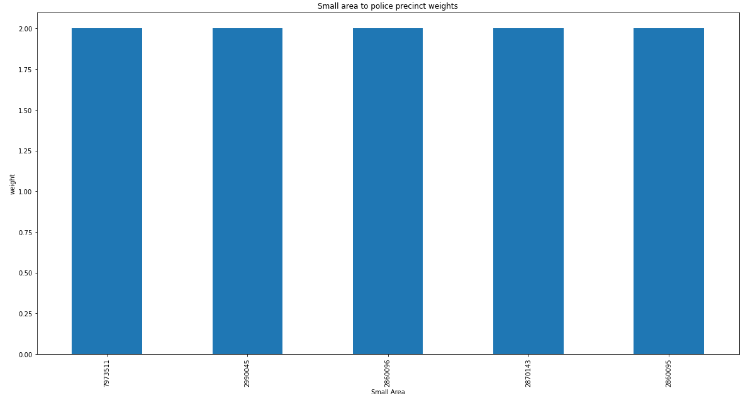
Seventh: Plotted a density for the data set. On the first plot it shows that when analyzing the small location the highest density is 8 and the lowest density is 5. and there is a scatter plot that demonstrates the same thing but in a different plot. When visualizing the weight it shows that the highest density is 1 and the lowest is between 0.1 and 0.2,



Eight: Plotted 3 dist-plot graphs to see the behaviour of the small area column in different graphs namely the Normal and the Log Normal graph.



Nine: Plotted a bar graph that displays the Distribution of weight in small areas. It shows that the average weight for each area reaches 2.0.



# ENSURE THIS DOCUMENT IS NEAT AND CAN BE ADDED IN YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY : Nombuso Dlamini**

