Data cleansing and preprocessing

Name of the Student

Name of the University

Authors note

# Part A

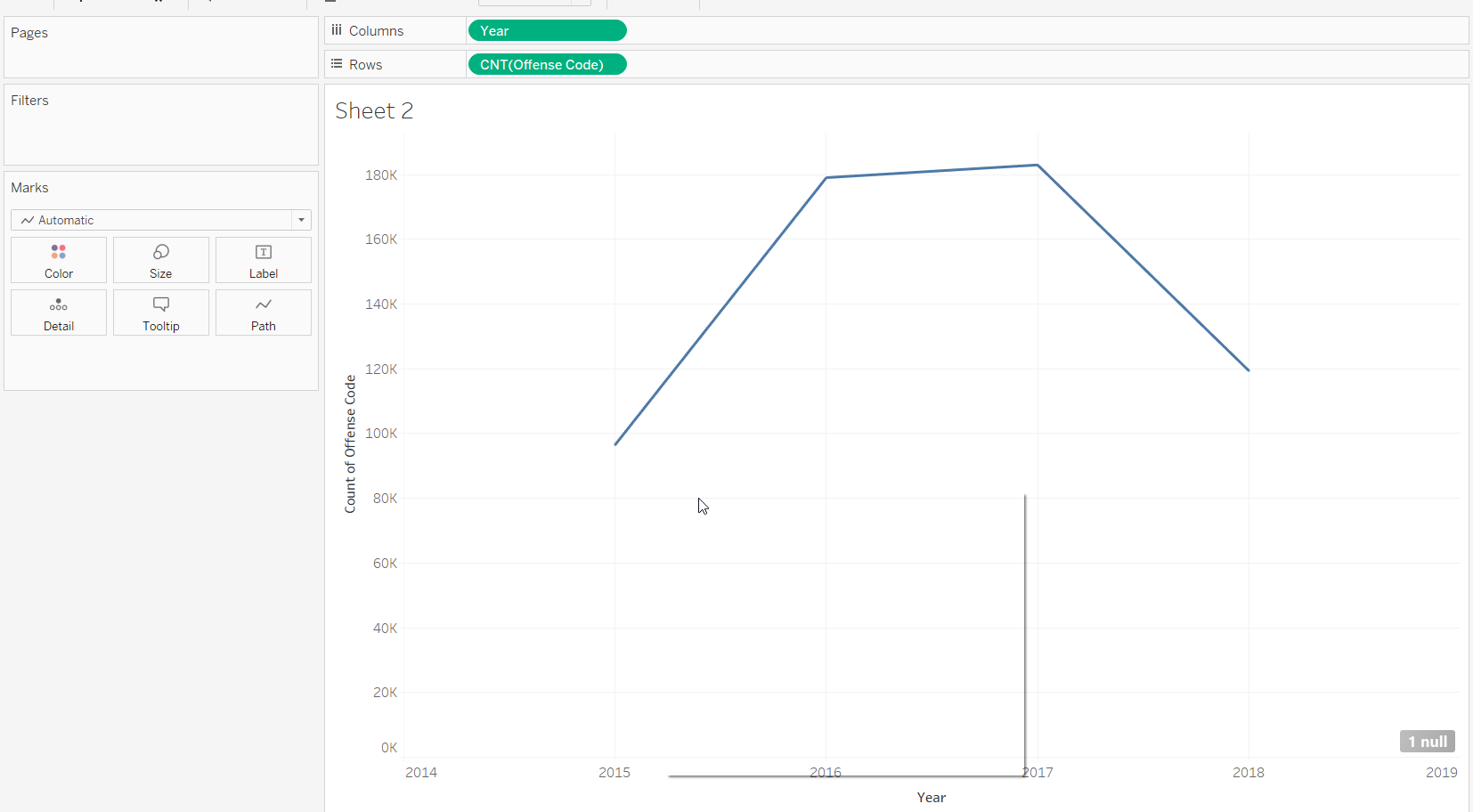
In the field of data analysis, the data cleansing and data preprocessing are considered as one of the crucial stages of the extracting data insights from the collected data. The collected data from the internal or external sources can be of structured, unstructured or semi structured in nature that will be hard to ingest in the tools to get the required patterns or insights. While inspecting the collected data it is important to ensure that data types of the different attributes in the dataset are appropriate and if required changed to the appropriate one.

Such as names and crime descriptions in the data files should be of the form strings, recorded date and year should be of date time type in order make the database engine or other data analysis tools. If required new calculated columns can be added to the dataset using the different tools available in the tableau or other data analysis tools along with the transformation of the dataset. For different attributes the null or unacceptable values in the columns can lead to inconsistent results from the used data set in the analysis. through the measure option for the different attributes, it is possible to use the aggregate (SUM, average) or distinct values to generate plot from the data using tableau tool. Tableau also provides the filter option in order to eliminate the null or NaN values as well as concentrating on the specific points for the exploration.

# **Part** B

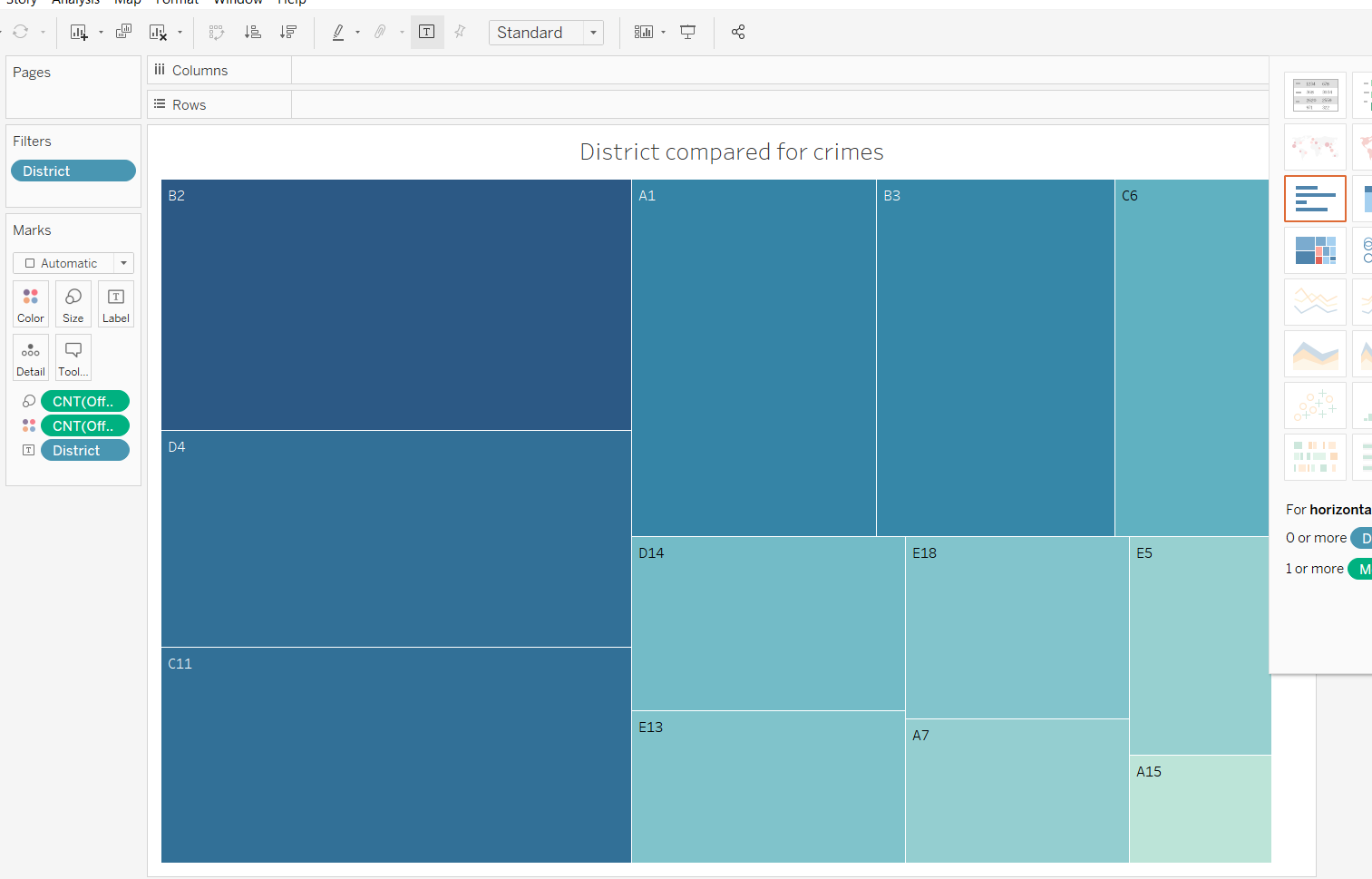
At first the null values are removed or deselected as the part of the data cleansing as well as the type of the values for different variables are checked. Finally both the data files are merged depending on one column which lead to generating the integrated plots from both the database.

In the following line plot, the trend of crimes over the years is investigated.



From the above plot it is clear that maximum number of crimes are recorded in the year of 2017. Lowest number of crimes are recorded in the years of 2015.

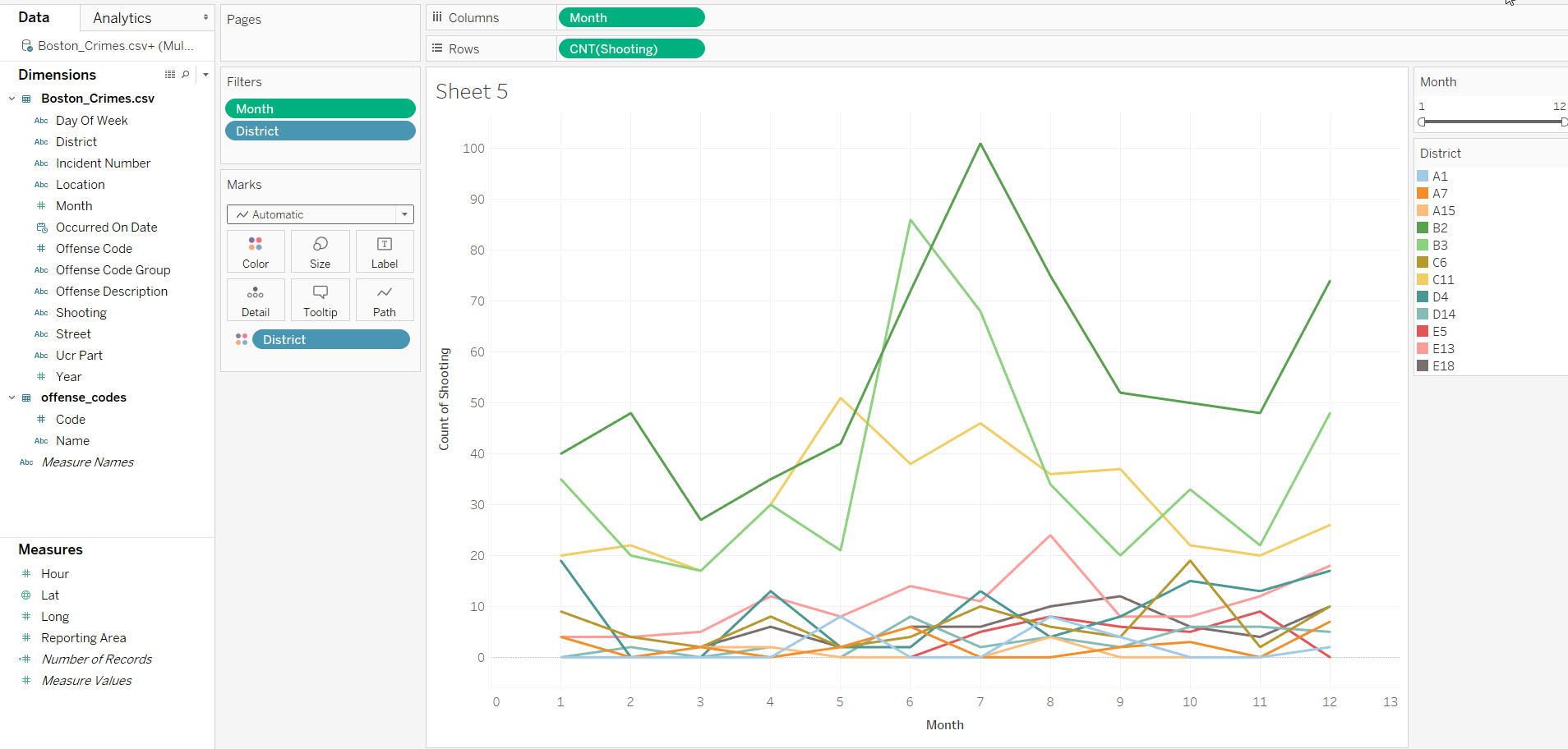
Following is another plot that depicts the areas or districts that have higher number of crimes in their locality.



From the above heat map plot, it can be stated that maximum number of crimes are reported for B2 and lowest number of crimes are recorded for A15. D4, C11 are the other important areas where the police department needs to concentrate in order to maintain better law and order inside Boston.

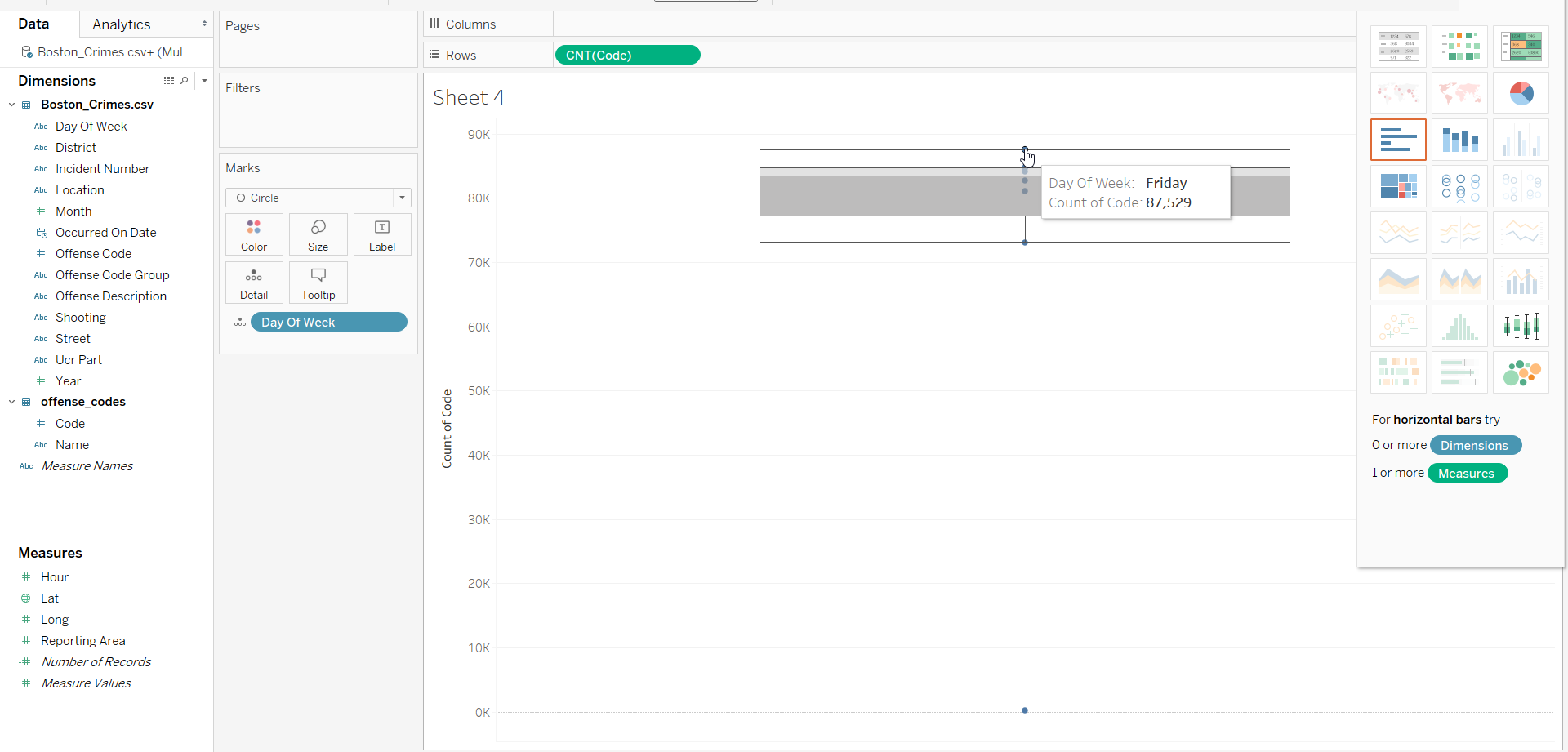
# Part C

In this section the count of the shooting incidents are compared for different districts and different months.



Here, it can be seen that that maximum number of shooting incidents are recorded from the 5th to 8th months in the districts B2, B3, C11.

In the next plot, box plots are utilized in order to find out the day with most recorded crimes in Boston.



Here, in the above box plot, on Friday most number of crimes are recorded.

# Bibliography

Datig, I. and Whiting, P., 2018. Telling your library story: tableau public for data visualization. *Library Hi Tech News*.

Ko, I. and Chang, H., 2017. Interactive visualization of healthcare data using tableau. *Healthcare informatics research*, *23*(4), pp.349-354.

O'Brien, A. and Stone, D., 2020. Yes, You Can Import, Analyze, and Create Dashboards and Storyboards in Tableau!. *Journal of Emerging Technologies in Accounting*, pp.0000-0000.

Town, P. and Thabtah, F., 2019. Data analytics tools: a user perspective. *Journal of Information & Knowledge Management*, *18*(01), p.1950002.