Programming Exercise 1: Tableau Public

Parikh Malav Arunkumar

30708974

Lab 04 Monday 8 AM

Fahimeh Sadat Saleh and Pratik Bhumkar

# **Data Wrangling**

The raw data provided contains a date\_time column and contains year, month, date and time so there is no need to keep the date\_time column and hence it has been removed for making the data concise. Similarly, there are separate columns for latitude and longitude alongside a column named location which gives a combined latitude and longitude so there is no need to keep that location column and hence has been removed. Raw data contains a column named sensor\_name which is a short reference to sensor\_description but is of no relevance and hence has been removed. The blank values in raw data have been replaced with ‘-’ to notify that there is no data for that cell. The column names are not in the same format and are inconsistent as one file shows sensor names as sensor\_description while second file shows it as sensor\_name which has been changed to sensor\_description in both files. Similarly, sensor\_id and sensor\_ID are inconsistent and have been changed to sensor\_id in both files.

The data is loaded in Tableau and a join has been applied between Pedestrian counting system sensor location and Pedestrian counting system 2019. A left join is used here because we want to consider the hourly counts of each sensor in 2019 even if there are no counts. The image below shows the data in its current state.

A screenshot of a computer

Description automatically generated

A screenshot of a social media post

Description automatically generated

# **Data Exploration**

While exploring the data in Tableau some errors were found. The first error being the naming conflict for three sensor ids namely 46, 54 and 60. As shown in the images below, the names alongside these 3 sensor ids are specified in two different ways.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

This data inconsistency is cleaned in Excel by applying filter to the columns and replacing the conflicting names with the ones provided in the sensor locations file and the cleaned entries are shown below.

A screenshot of a computer

Description automatically generated

Another error is with sensor id 47 where the sensor description is duplicated and written twice as shown in image below.

A screenshot of a computer

Description automatically generated

This error is removed in Excel and has been given the name correctly as shown in the image below.

A screenshot of a computer

Description automatically generated

The sensor id 14 shows an installation date of 25/09/2019, but there are records of hourly counts at sensor id 14 before 25/09/2019. There is no note added of the sensor being upgraded or a new one is being installed instead of the old one. And, the sensor id 25 shows that the sensor at id 14 is relocated to sensor id 25 so there should be a note mentioning the information about the sensor being upgraded or a new one being installed. Because if it is not the case than sensor id 14 should have a status of inactive and subsequently should not have any data after the date October 2 to justify that the senor id 14 is removed. But, the status of sensor id 14 is active and the file contains data after October 2 which suggests that the sensor id 14 is either replaced with a new one or upgraded on September 25. So, there should be a note added against sensor Id 14 suggesting it is upgraded to make the data consistent and meaningful.

A screenshot of a computer

Description automatically generated

Thus, a new note of the sensor being upgraded on 25/09/2019 has been posted on the notes section of sensor id 14.

# **Answering Questions**

1. *How does pedestrian traffic volume vary in different parts of Melbourne?*

The visualization provided below will help in answering the above question. The below shown visualization is a map which has been plotted using the longitude and longitude provided and the various dots show the locations of all the sensors irrespective of its current status. The size of the dots varies based on the number of hourly counts starting from 331,209 to 13,934,140 divided amongst 4 different sizes. The groups are divided into 4 different types of clusters based on the hourly counts where dark blue cluster are the not clustered once having zero hourly as those sensors are inactive or have been removed. The visualization also shows names of the sensors along with the average hourly counts. The visualization clearly shows that the pedestrian traffic is highest in a certain region which is Melbourne Central Business District (CBD). As the region is the heart of Melbourne, the pedestrian traffic is the highest in and around CBD with multiple offices being located there alongside many restaurants, multiplexes which garners a lot of public attraction. The regions away from the CBD have substantially lesser number of pedestrian volumes every hour. It can be suggested from the visualization that the yellow spots are the places that have the highest number of pedestrians and the blue spots are the places with least pedestrians across Melbourne.

A map of a computer

Description automatically generated

1. *How does the day of the week and time affect overall pedestrian traffic volume?*

The visualization provided below will help answering the above question. The below shown visualization shows an area chart distributed across time on x-axis shown in 24-hour format and y-axis showing the total hourly counts. The different colors represent different days of the week as shown in the legend. The visualization clearly shows that every single day be it a weekday or weekend, the pedestrian count across Melbourne is at peak at 5 PM. It is the peak time of people going home from work or going out to enjoy resulting in the highest hourly counts across any day of the week. While the time period from midnight 12 to wee hours the number of pedestrians is negligible as compared to any other time in a day. It can be seen in the chart that the distribution of pedestrians across any day any time is similar except at nighttime. The pedestrian traffic at weekends especially Friday and Saturday are more after 5 PM up to midnight 12. From the visualization, it can be suggested that Friday is the busiest day across the week and 5 PM is the busiest hour on every day across the week. While Sunday has the least spread across time suggesting Sunday is the least busiest day and morning 4 AM is the least busiest hour on every day across the week.

A screenshot of a map

Description automatically generated