FIT5147 Data Exploration and visualisation
Programming Exercise 1- Tableau

Student Name: Arushi Tejpal
Student ID: 28130006
Tutorial time: Friday 4pm
Tutor Name: Farah Tasnuba Kabir, Jie (Lewis) Liu

DATA WRANGLING

In this assignment I will analyse pedestrian traffic in the City of Melbourne and answer if different locations, day and time will affect the pedestrian traffic in Melbourne.
The initial step to investigate this data first clean the data.
Data wrangling is the method of gathering, cleaning, selecting and transforming data to help analyse and form conclusions from the data.

1.1 Data Wrangling Steps:

Step 1:
In the file *" Pedestrian_counting_system_sensor_location.csv"* I updated the location for Sensor_id 16 to the sensor_id 53's location as direction_1: WEST and direction_2: as EAST.

Step 2:
I then deleted sensor_id: 38 and sensor_id 32 as those devises had been removed in 2017.

Step 3:
Sensor_id : 13 has null categorical values in direction 1 and direction 2 therefore this row needs to also be deleted.

Step 4:
Sensors that were installed after 2019 needed to also be deleted as we were using data from 2019. Sensor id: 66, 65 and 63 were removed from the file as they were installed in year 2020.

Step 5:
The Naming:

In the file "pedestrian_counting_system_2019.csv" file, the columns in sensor_name is the same as sensor_description in the *"Pedestrian_counting_system_sensor_location.csv"* file therefore you are able to join the sensor description and sensor name;
However, I have found some errors as sensor_id: 59 was named building 80 "Building 80 RMIT" in the *Pedestrian_counting_system_sensor_location.csv* file
but was named "Swanston St- RMIT Building 80" in the "pedestrian_counting_system_2019.csv" file, many of these errors were found.
Therefore, forming a join with the names attribute would allow us to lose our data as they that did not match.

Step 7:
The time was showing '0' for 12pm as the time on the file was not AM and PM. I was able to convert the time from excel. Seen in figure 1.1

Time

C	D	E	F	G	H	I
Month	Mdate	Day	Time	Sensor_ID	Sensor_Nam	Hourly_Counts
November	1	Friday	5:00:00 PM	34	Flinders St-S	300
November	1	Friday	5:00:00 PM	39	Alfred Place	604
November	1	Friday	5:00:00 PM	37	Lygon St (Ea	216
November	1	Friday	5:00:00 PM	40	Lonsdale St-	627
November	1	Friday	5:00:00 PM	36	Queen St (W	774
November	1	Friday	5:00:00 PM	29	St Kilda Rd-A	644
November	1	Friday	5:00:00 PM	42	Grattan St-S	453
November	1	Friday	5:00:00 PM	43	Monash Rd-S	387
November	1	Friday	5:00:00 PM	44	Tin Alley-Sw	27
November	1	Friday	5:00:00 PM	35	Southbank	2691
November	1	Friday	5:00:00 PM	45	Little Collins	2173
November	1	Friday	5:00:00 PM	46	Pelham St (S	203
November	1	Friday	5:00:00 PM	47	Melbourne C	2354
November	1	Friday	5:00:00 PM	48	QVM-Queen	358
November	1	Friday	5:00:00 PM	49	QVM-Therry	161
November	1	Friday	5:00:00 PM	50	Faraday St-L	502
November	1	Friday	5:00:00 PM	51	QVM-Frankli	159
November	1	Friday	5:00:00 PM	52	Elizabeth St-	914
November	1	Friday	5:00:00 PM	54	Lincoln-Swar	276
November	1	Friday	5:00:00 PM	55	Elizabeth St-	2070
November	1	Friday	5:00:00 PM	56	Lonsdale St -	789
November	1	Friday	5:00:00 PM	57	Bourke St Br	2122
November	1	Friday	5:00:00 PM	58	Bourke St - S	2528
November	1	Friday	5:00:00 PM	59	Swanston St	485
November	1	Friday	5:00:00 PM	61	Swanston St	1240
November	1	Friday	5:00:00 PM	62	La Trobe St (	229
November	1	Friday	6:00:00 PM	4	Town Hall (W	2950

Figure 1.1 Changed "Time" column to AM/PM time.

Step 6:

Finally we have to link both data sets by inner joining sensor id from both sets as seen in figure 1.2.

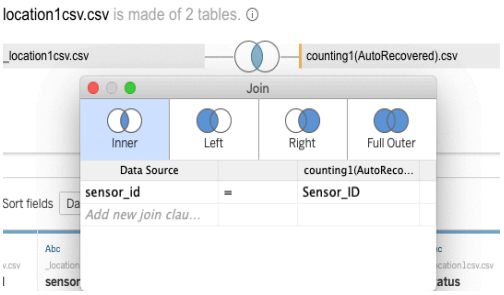

Fig.1.2 inner join with sensor id on Tableau.

Naming error:

If inner joined occurred with the sensor name and sensor description without name changes then no outlier would be found in the box plot but if joined occur with sensor_id then outlier found would be sensor id 22, seen in figure 1.3.

This shows that some of the data is lost with joining the names as they are both different in the different files.

Furthermore, it is found that there is a count of 58 values for inner join with sensor id and a count of 53 values with inner join with sensor_name. which further proves that

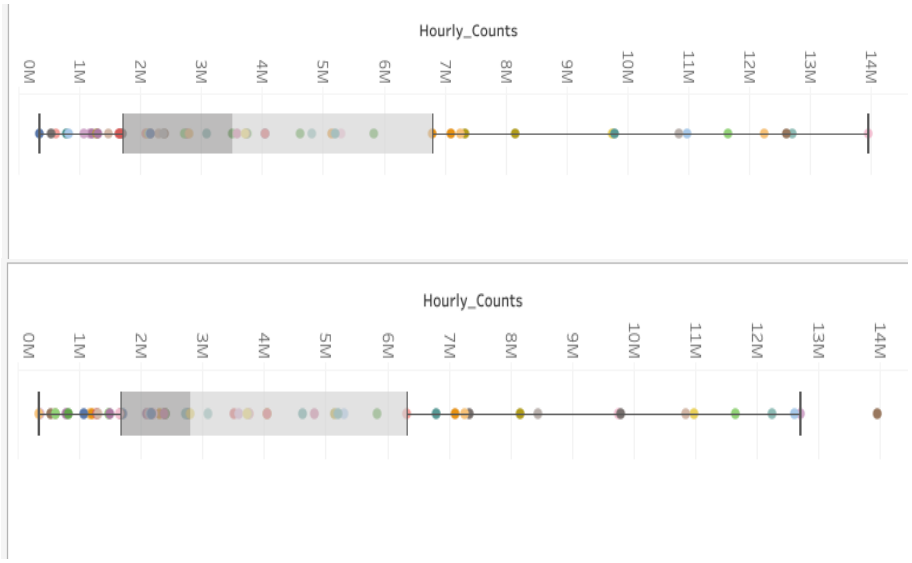

Fig.1.3. Right side: inner join with sensor_name box plot. Left side inner join with sensor id box plot.

DATA EXPLORATION

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

From figure 2.1 you are able to see the map of Melbourne city and a large cluster of circles among the centre of the city, mainly Elizabeth St. and Swanston St.

You are also able to see a clusters around train stations.

The bigger the circle the more hourly counts which shows there's more traffic. As you move out from the centre of the city the cluster becomes smaller which concludes that there is less pedestrian traffic on the outer parts of the city and more traffic volume in the centre of the city.

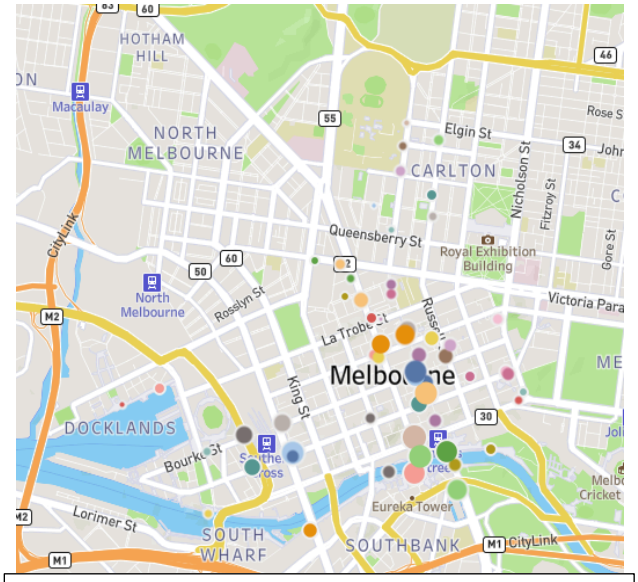

Fig.2.1. Map of City of Melbourne. Street view. The size of circles represent the hourly count and the colour represents the different sensor_id's.

From figure 2.2 you are able to see a bar graph with all sensor id's and names

The maximum hourly count is sensor id: Flinders St. – Elizabeth St. with hourly count 13,934,140.
The minimum hourly count is sensor_id 62 Latrobe St. (North) with an hourly count of 331,209. The average hourly count is 4,593,304.74

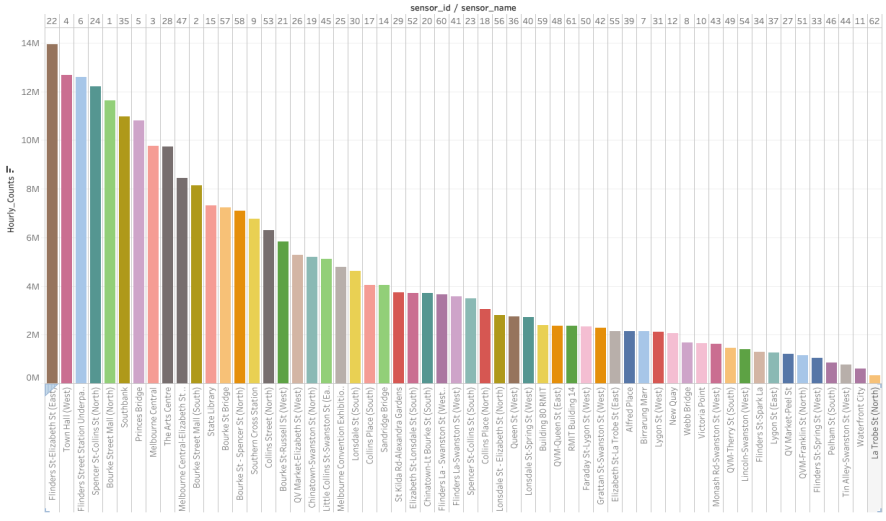

Fig.2.1 Bar chart of all sensors in Melbourne City vs Hourly count, descending order.

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

In figure 3.1 it shows the maximum day is on Friday with 2211690 hourly count with the minimum day on Sunday with 40769 hourly count. The average hourly count is 656186.39. Which helps conclude that the busiest day in 2019 was Friday and the least busiest day is Sunday with the least traffic volume.

The Box plot graph shows the outliers found on Saturday which Is found to be sensor id 22 which is Flinders St. and sensor id 4 which is Town Hall.

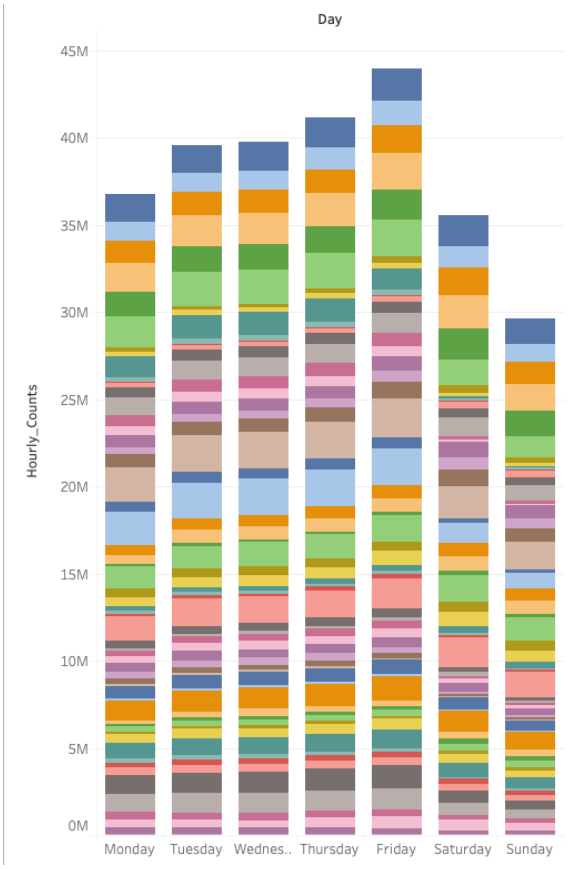

Fig.3.1 Bar chart of hourly count in each day for all sensor id.

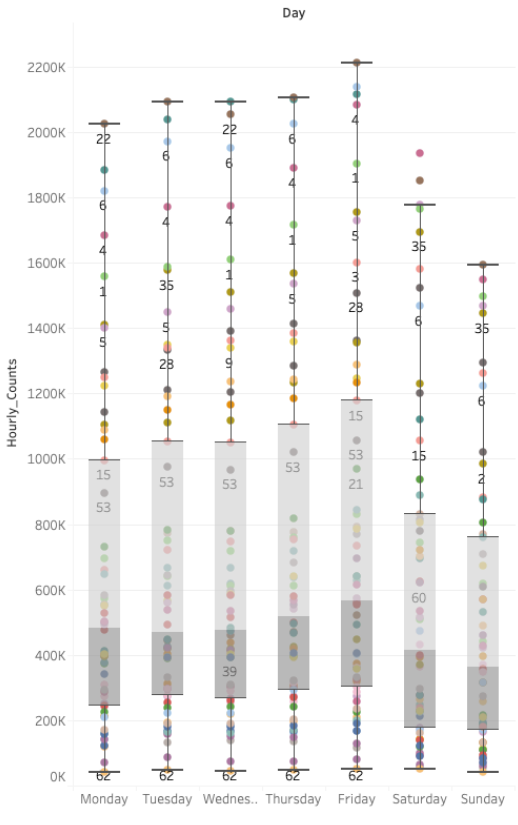

Fig.3.2 Box plot of hourly count in each day for sensor id.

The Bar graph in Figure 3.3 shows the most pedestrian traffic to be found at around 5PM and the least busiest at 4AM.
From 4am to 6am it is quiet low then at 7am the traffic has a drastic increase. 7AM 11AM and 8AM are equally as busy.

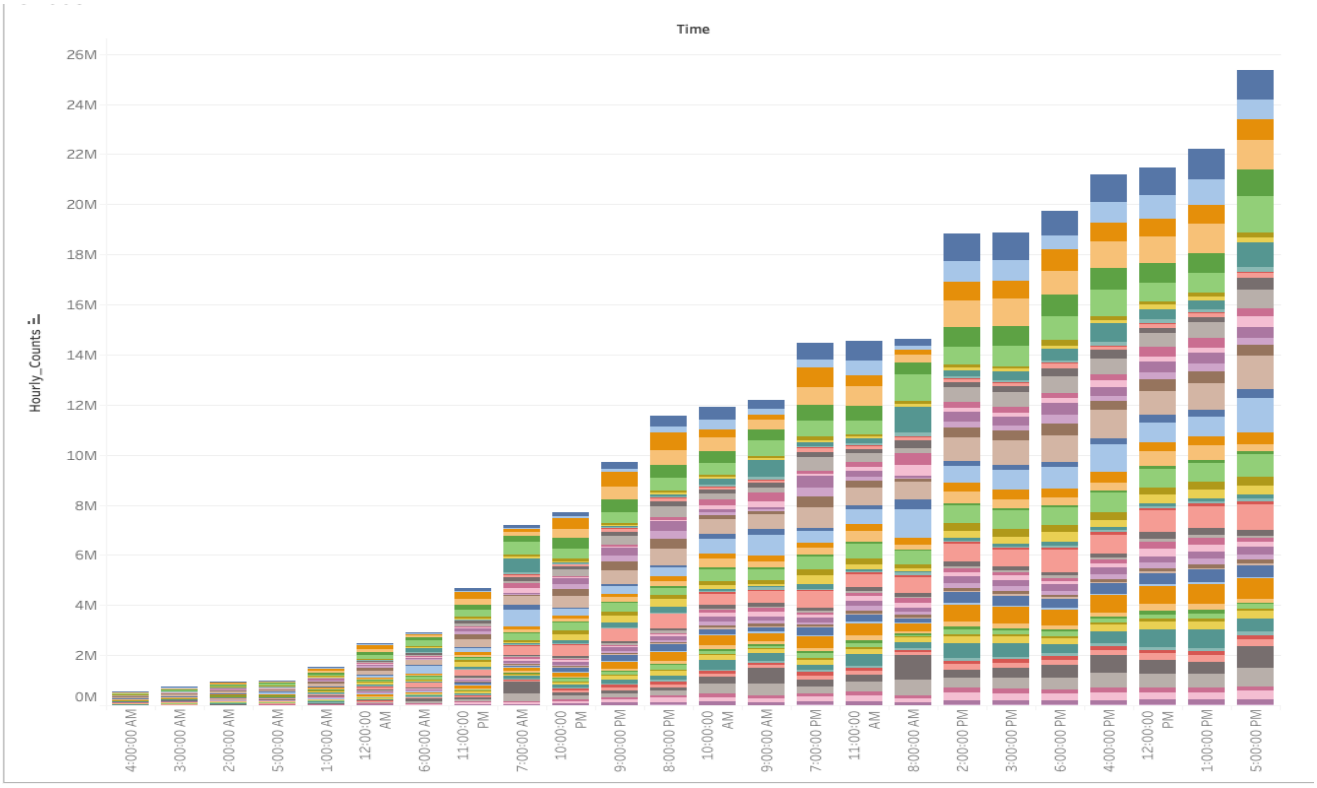

Fig.3.3. Bar graph of hourly counts with tome of each hour AM/PM. Ascending order.