Data Exploration and Visualization
Programming Exercise - 1

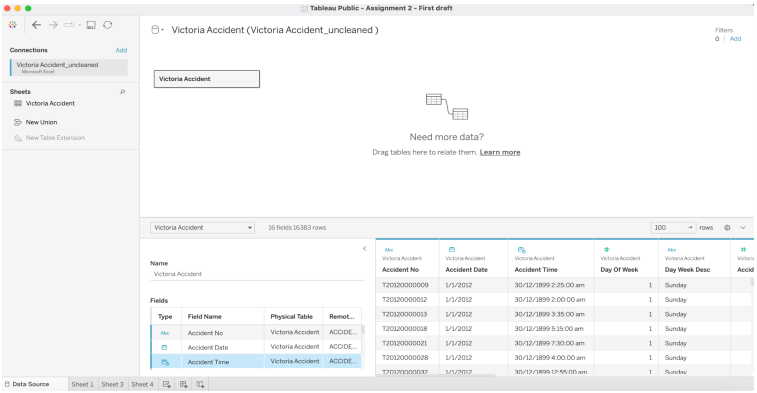
Name: Subbulakshmi Natarajan
Student ID: 34069178

1. Data checking and cleaning (i.e., Steps 1 to 3)

- **Image of loading the dataset**

I changed my CSV file format to Excel format since I got an error when I was uploading the CSV file through text file format. We can see that the accident time data type is changed to date and time it can be said to be one of eros since we have a separate column as accident date which is there in the dataset.

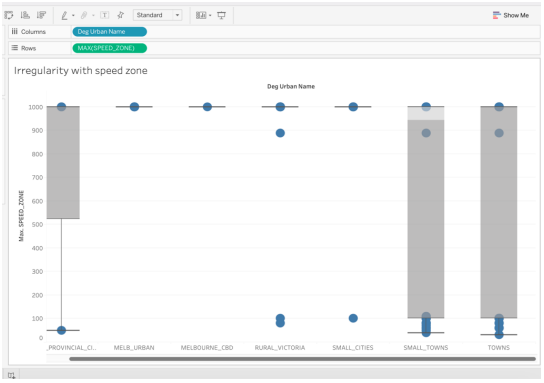
This is image of what it looks like when I load it into tableau is below:



- **Irregularities and errors**

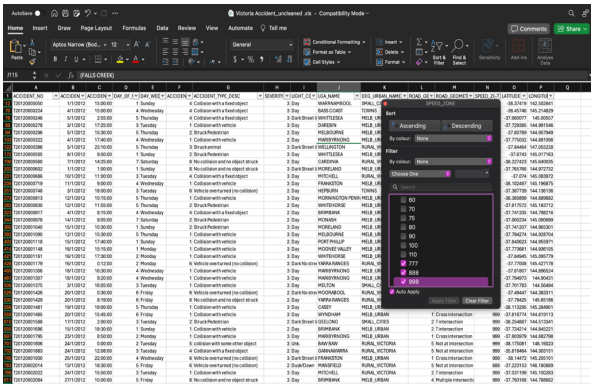
1) Speed Zone (Not on scale limit 0-110)

Using Tableau, I found the irregularities or errors through a box plot for the speed zone to depict the maximum speed that each city has taken. It can clearly be seen that the speed limit has crossed the scale like 888,999 and 777 which is not included in the speed scale. In the diagram, we can see that the speed zone I have taken is the maximum speed zone according to the urban name we can see that there are some data points which say beyond the speed scale as I have highlighted them in the picture. Below is the diagram for the following:

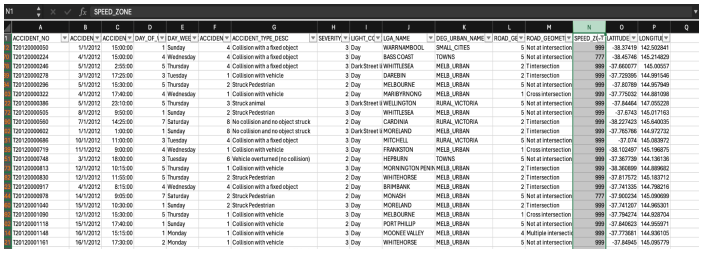


To correct this error, I have used Excel for clearing this unwanted data and made it null values now after the data is cleared here is how the figure looks:

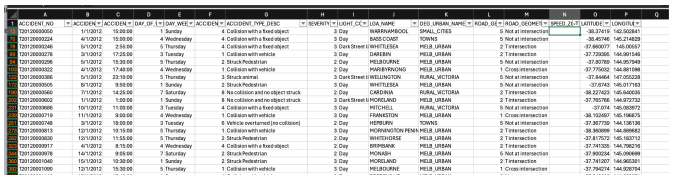
Step 1: I filtered the irrelevant data that is 777, 888 and 999.



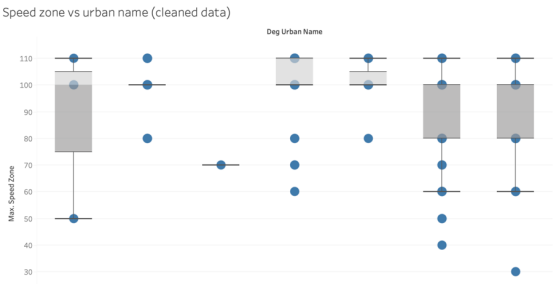
Step 2: Filtered the 777,888,999



Step 3: Removed the irrelevant data and make as null values.

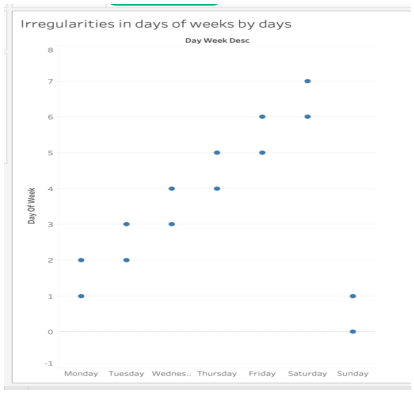


After the data is cleaned the figure below in Tableau:



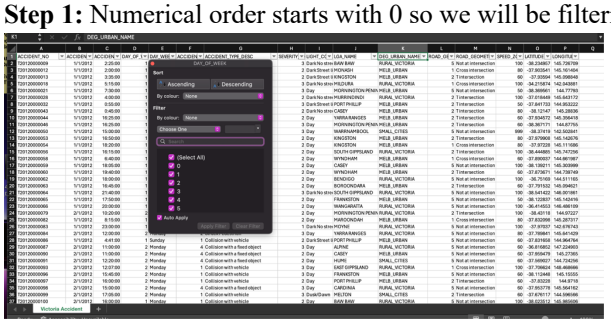
2) Days of week numerically the data is wrongly numbered according to the numerical order.

In the tableau, I checked the day of the week with the day of the week desc that is Sunday, Monday etc. When I noticed that it says 0 to 7. According to me, the week should start from 1- Sunday till 7 - Saturday we have a total of 7 days in a week. Below is the graph which shows irregularity in days of weeks by days which clearly shows that Sunday and Monday have week 1 so we can technically say it is wrong which is showing in the graph below:

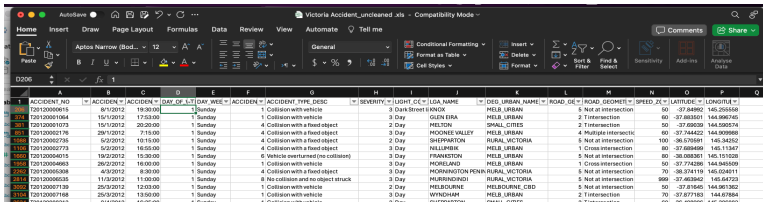


To clean this data, we will be using Excel to adjust the number of days in the column of day of the week.

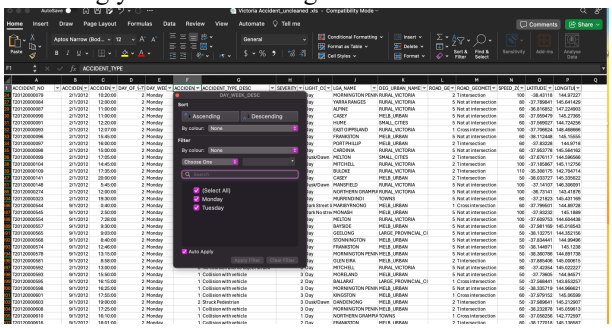
Step 1: Numerical order starts with 0 so we will be filtering with each day of the week and removing the 0.



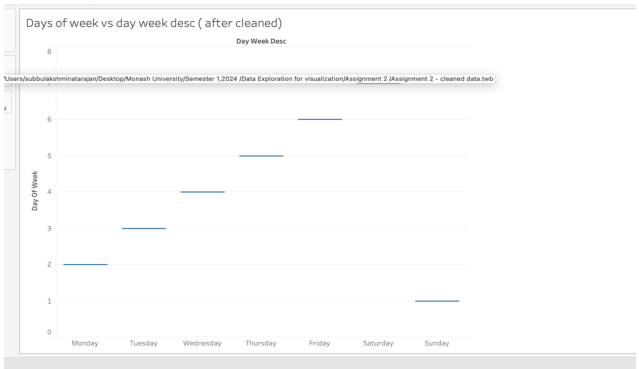
Step 2: Replace 0 to 1 for Sunday to make it a numerical order.



Step 3: You can see that from Monday onwards there is Tuesday which is in numerical order, so I adjusted accordingly for the following week so that it doesn't have irregularities.

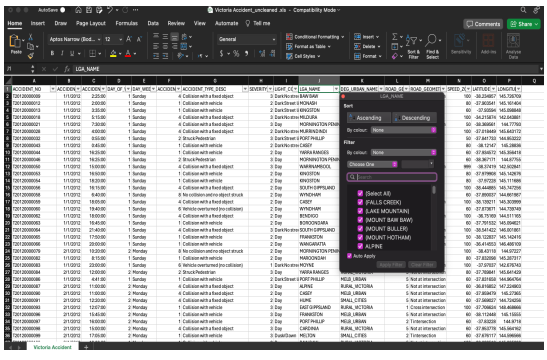


Data Cleaned after this graph which shows clearly:

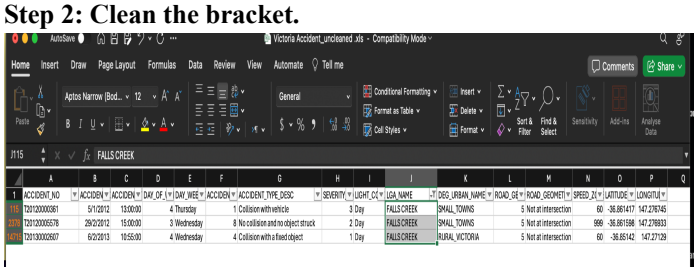


3) Removing brackets in the urban name

Step 1: Remove the bracket for Fall Creek, Lake Mountain, Mount Baw Baw, Mount Buller, and Mount Hotham.



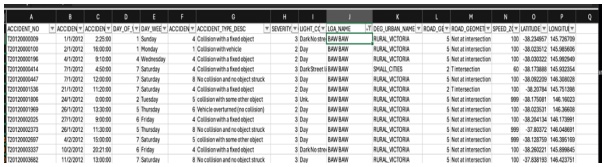
Step 2: Clean the bracket.



3) Spelling error for Baw Baw in the column LGA_NAME

To clean I used Excel for it.

Step 1: You can see Baw Baw filtered in the dataset.



Step 2: Now change Mount Baw Baw in the dataset.

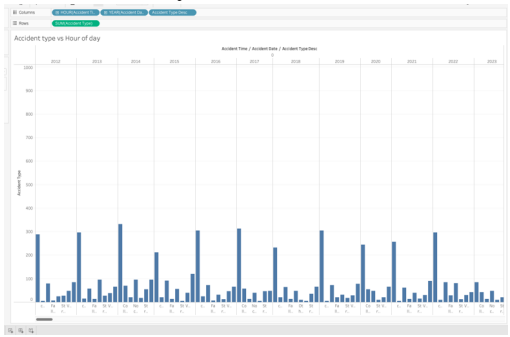
CIDENT_NO	ACCIDEN	ACCIDEN	DAY_OF_	DAY_WEE	ACCIDEN	ACCIDENT_TYPE_DESC	SEVERITY	LIGHT_CC	LGA_NAME	DEG_URBAN_NAME
20120000009	1/1/2012	2:25:00	1	Sunday	4	Collision with a fixed object	3	Dark No stre	MOUNT BAW BAW	RURAL_VICTORIA
20120000100	2/1/2012	16:00:00	1	Monday	1	Collision with vehicle	2	Day	MOUNT BAW BAW	RURAL_VICTORIA
20120000196	4/1/2012	9:10:00	4	Wednesday	4	Collision with a fixed object	2	Day	MOUNT BAW BAW	RURAL_VICTORIA
20120000414	7/1/2012	4:50:00	7	Saturday	4	Collision with a fixed object	3	Dark Street li	MOUNT BAW BAW	SMALL_CITIES
20120000447	7/1/2012	12:00:00	7	Saturday	8	No collision and no object struck	3	Day	MOUNT BAW BAW	RURAL_VICTORIA
20120001536	21/1/2012	11:20:00	7	Saturday	4	Collision with a fixed object	2	Day	MOUNT BAW BAW	RURAL_VICTORIA
20120001806	24/1/2012	0:00:00	2	Tuesday	5	collision with some other object	3	Unk.	MOUNT BAW BAW	RURAL_VICTORIA

From this screenshot, we can see that we have removed all the errors from the column from LGA_name to have clean data.

2. Data exploration (i.e., Step 4)

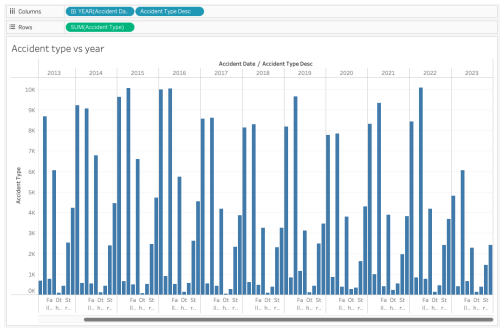
Q1.

- **Hour of day**



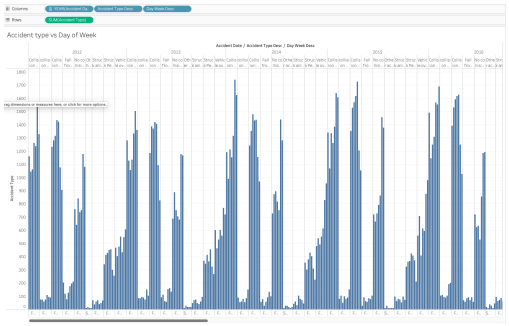
Each year the accident type for each year fluctuated according to accident type in each year. It has been said that the year said to me the highest according to other years and accident types. When it comes to accident type it can be said that vehicles have been recorded the highest. But this data clearly says that we cannot see each hour how many accident types have happened in each year that has been recorded.

- **Year**



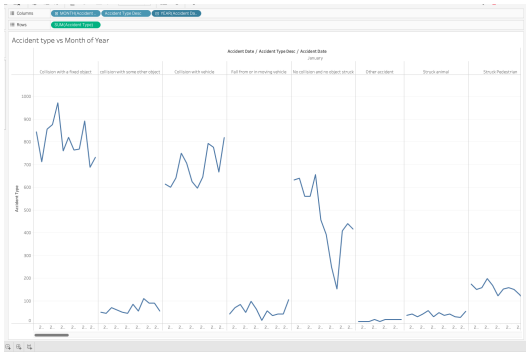
Each year the accident type for each year fluctuated according to accident type in each year. It has been said that the year said to me the highest according to other years and accident types. When it comes to accident type it can be said that vehicles have been recorded the highest.

- **Day of week**



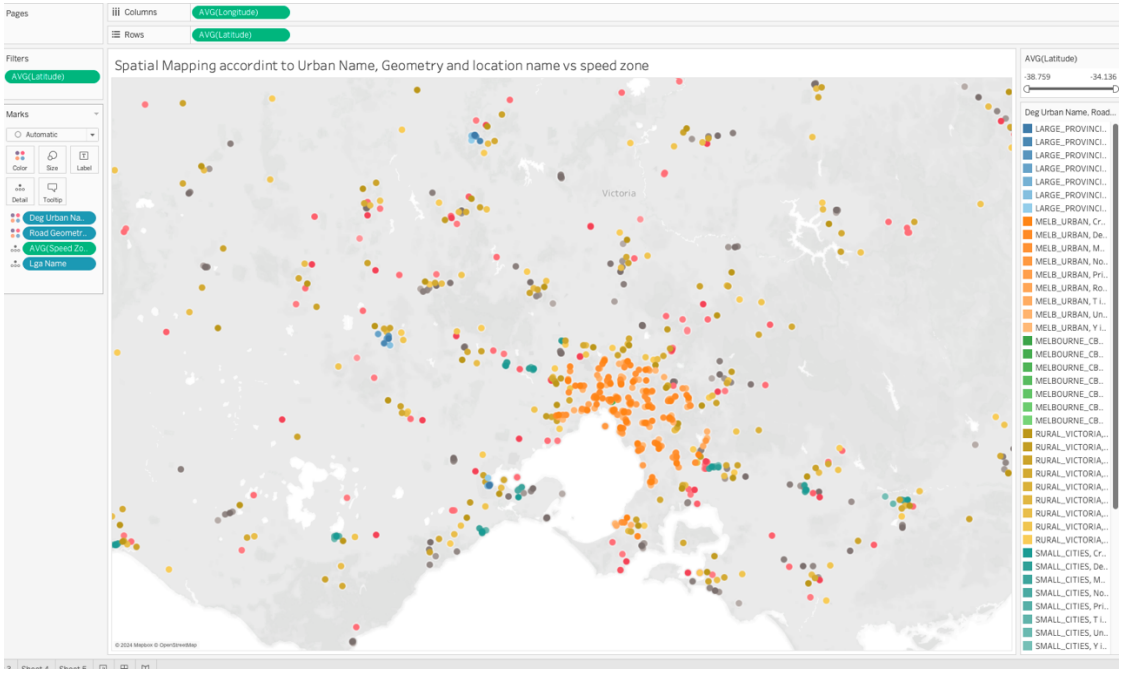
Each week in the years it says the same as above that year 2022 is the same.

- **Month of Year**



Each month of year by accident type it can be said by the scatter plot we can fluctuation of the graph due to increase and decrease of the accident.

Q2.



From the spatial graph, we can clearly identify the speed zone in each region and cities that can be seen clearly in this graph.

Challenge:

We cannot identify due to the data quality of the dataset we cannot specify the road support for this map.