**Task 2**

For this task, Python is used with library Pandas.

The label “Flags” and “Flag Text” are obtained in Task 1, so they are used directly in this task.

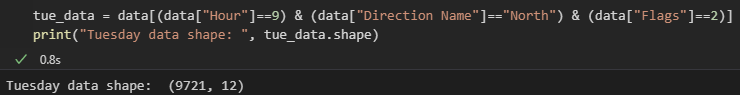
First step is read the file and convert the label “Date” to type *datetime*, which makes further operation convenient.



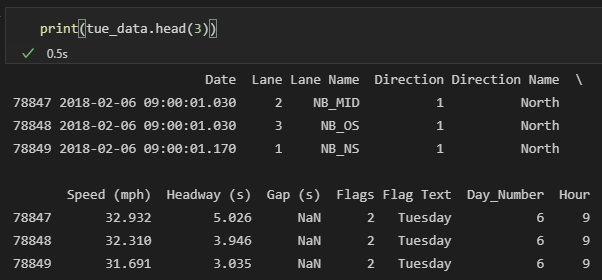
Second, create a new label called “Day\_Number” to store the day corresponding to the label “Date”. For example, “Date: 2018-02-02 00:00:03.050000” is “Day\_Number: 2”. And also create a new label called “Hour” to store the specific hour corresponding to the label “Date”. For example, “Date: 2018-02-02 00:00:03.050000” is “Hour: 0”.



Third, filter the data that satisfied the requirements of Tuesday between 9:00 and 10:00, and in North lanes. It has 9721 rows and 12 columns.



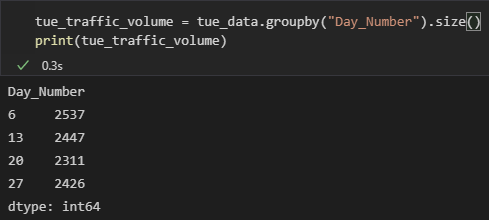
Now the first three rows of filtered data:



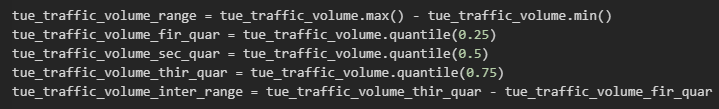
Now there are two circumstances, first is dividing the data depend on label “Day\_Number” only, this gives four groups of data. Second is dividing the data depend on “Day\_Number” and “Lane Name”, this gives 12 groups of data.

*Four groups:*

Use groupby() function and size() function to get the traffic volume of each Tuesday.



Finally, use functions max()-min() to get range, quantile(0.25) to get first quartile, quantile(0.5) to get second quartile, quantile(0.75) to get third quantile and quantile(0.75)-quantile(0.25) to get interquartile range.

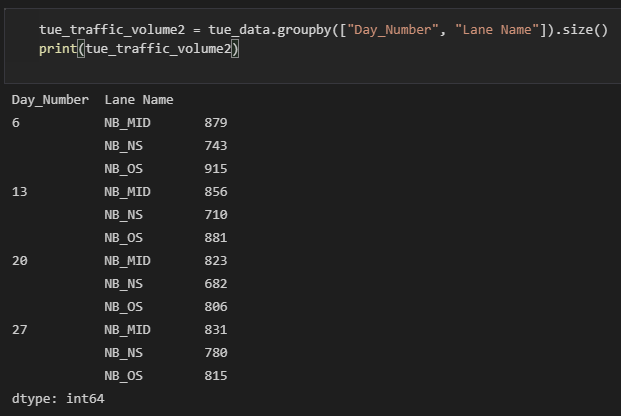


The result of 4 groups:

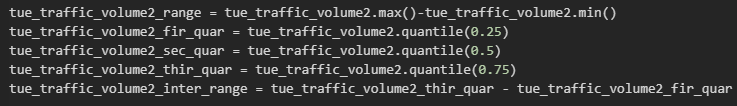
|  |  |
| --- | --- |
| Range | 226 |
| 1st Quartile | 2397.25 |
| 2nd Quartile | 2436.5 |
| 3rd Quartile | 2469.5 |
| Interquartile Range | 72.25 |

*Twelve Groups*

Use groupby() function and size() function to get the traffic volume of each lane in each Tuesday.



Use same steps as 4 groups to get range, first quartile, second quartile, third quartile and interquartile range.



The result of 12 groups:

|  |  |
| --- | --- |
| Range | 233 |
| 1st Quartile | 770.75 |
| 2nd Quartile | 819 |
| 3rd Quartile | 861.75 |
| Interquartile Range | 91 |

**Interpretation:**

Range and interquartile distance both describe the dispersion of data. However, range is easily affected by outliers. For interquartile distance, outliers are excluded, but only partial data is used. The analysis is performed by combining these values and described in 4 groups and 12 groups respectively.

*Four Groups*

Both of the values of range and interquartile range are not too large, which means that the traffic volume in the north direction every Tuesday between 9am to 10am of the month does not vary much. It can be inferred that if there are no special circumstances (such as festivals or emergencies), the traffic volume on every Tuesday between 9am to 10am in the following months should be around the average value (). In addition, there may not many outliers on this specific traffic volume of north direction.

*Twelve Groups*

Both of the values of range and interquartile range are not too large, which means that the traffic volume in the north direction of **each lane** every Tuesday between 9am to 10am of the month does not vary much. It can be inferred that if there are no special circumstances (such as festivals or emergencies), the traffic volume of **each lane** on every Tuesday between 9am to 10am in the following months should be around the average value (). In addition, there may not many outliers on this specific traffic volume of north direction.

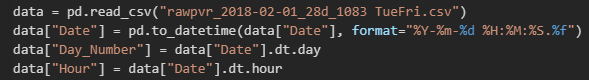
**Task 3**

For this task, Python is used with libraries Pandas and Matplotlib.

Tuesday is chose to visualize the average traffic volume for each hour.

The label “Flags” and “Flag Text” are obtained in Task 1, so they are used directly in this task.

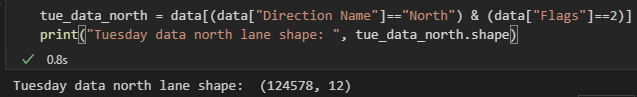
The first and second steps are the same as task 2 first two steps. Following is the codes:



From third step, there are two circumstances. First is north direction, and second is south direction.

*North Direction*

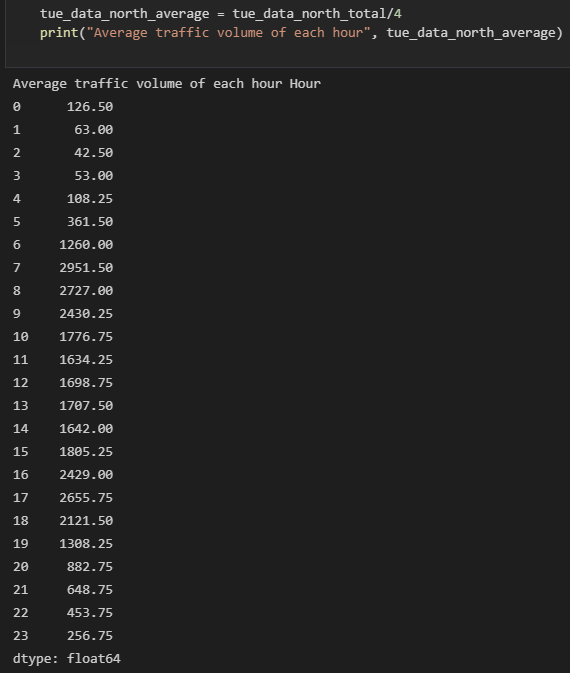
Third, filter the data that satisfied the requirements of Tuesday and in North lanes. It has 124578 rows and 12 columns.



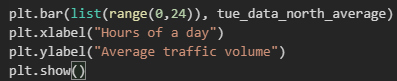
Use groupby() function and size() function to get the traffic volume of each hour on Tuesday.



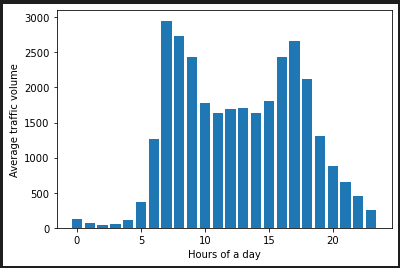
The values currently obtained are the total traffic volume of each hour, the average value should be divided by 4.



Then visualize the values as bar plot

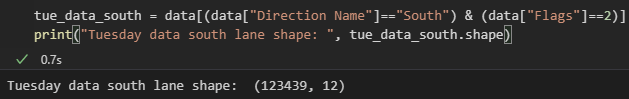


The final result of North Lanes average traffic volume per hour:



*South Direction*

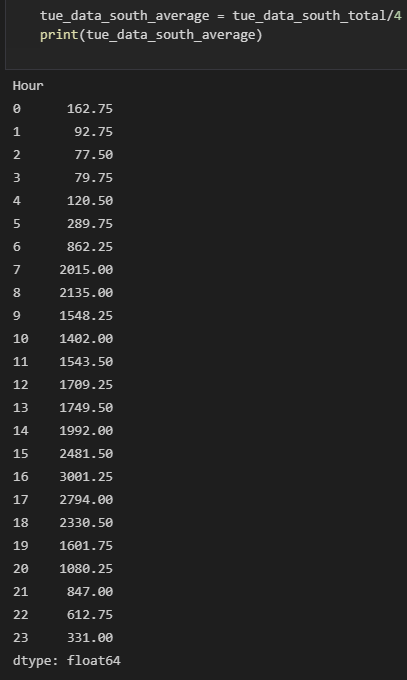
Third, filter the data that satisfied the requirements of Tuesday and in South lanes. It has 123439 rows and 12 columns.



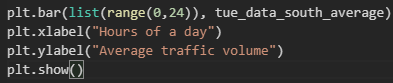
Use groupby() function and size() function to get the traffic volume of each hour on Tuesday.



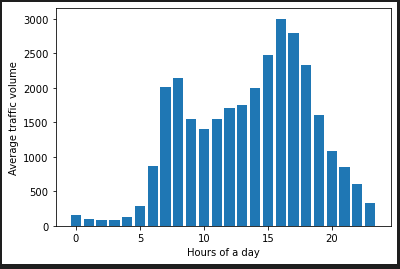
The values currently obtained are the total traffic volume of each hour, the average value should be divided by 4.



Then visualize the values as bar plot



The final result of South Lanes average traffic volume per hour:



**Interpretation:**

As can be seen from the two bar plot above, for Tuesday, the general trend of north direction and south direction traffic volume is the same. There is a low peak time at 0 to 4 o'clock and a peak time at 7 to 9 o'clock, and 15 to 18 o'clock. For the north direction, the lowest peak time occurs at 2:00 with a value of 42.5 and the highest peak time occurs at 7:00 with a value of 2951.5. For the south direction, the lowest peak time also occurs at 2:00 with a value of 77.50 and the highest peak time occurs at 16:00 with a value of 3001.25.

It can be inferred that more people go to work in the north direction, and similarly more people go home after work in the south direction. And most people go to work at 7:00 and 8:00, and leave work at 16:00 to 17:00.

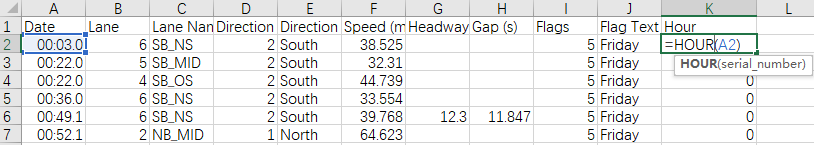
**Task 4**

For this task, excel is used.

Tuesday is chose to visualize the average traffic volume for each hour.

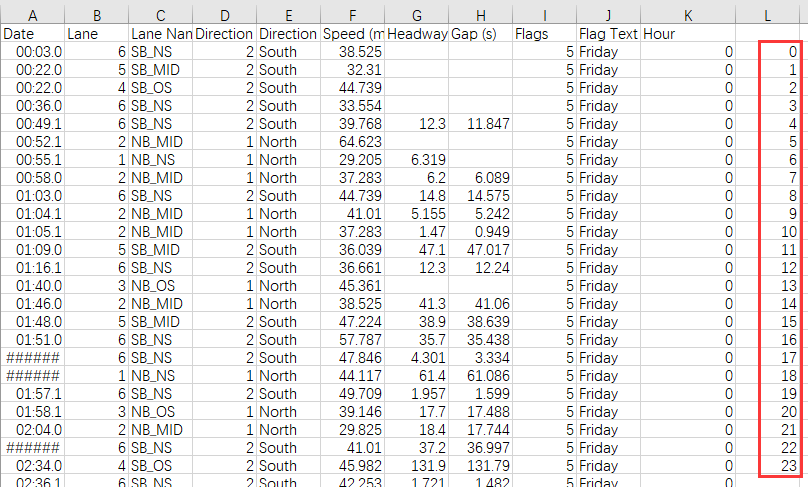
The label “Flags” and “Flag Text” are obtained in Task 1, so they are used directly in this task.

First create a new label called “Hour”, this column stores the specific hour corresponding to label “Date”. This can be done by using HOUR() function



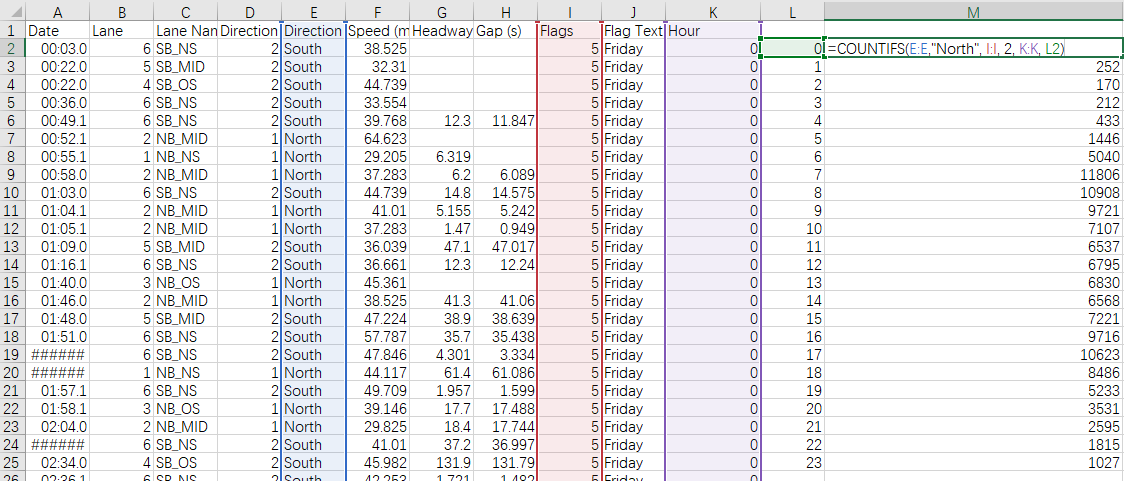
Second create a column with 24 rows, with each cell filled with a single number from 0 to 23. This can be done easily by enter 0 and 1 in first two rows and then apply to 23th row.

From third step, there are two circumstances. First is north direction, and second is south direction.

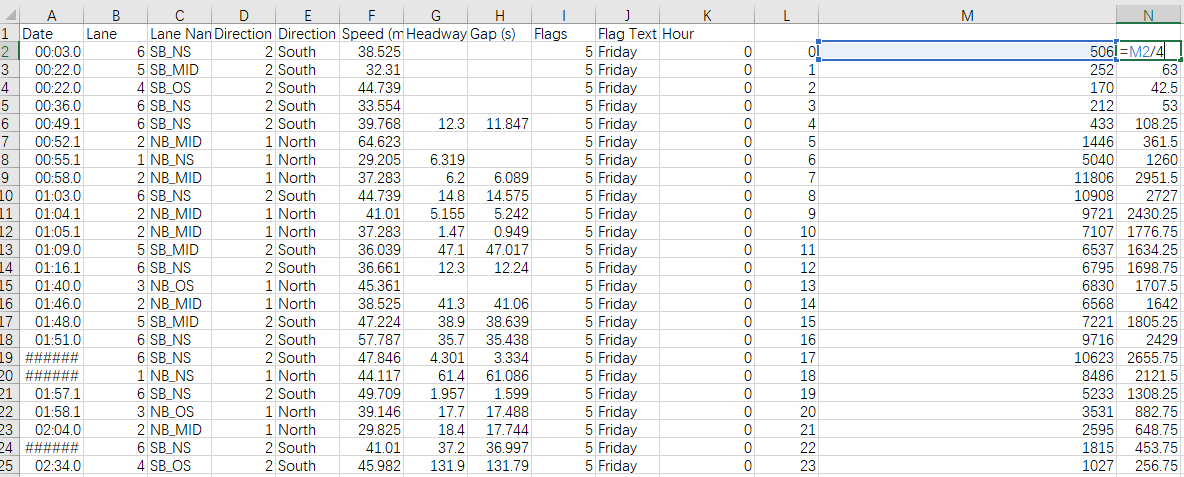


*North Direction*

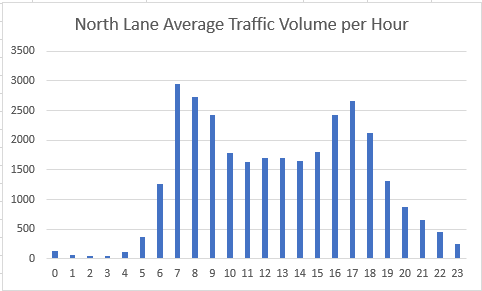
Third use COUNTIFS() function to count the data that satisfied the requirements of north lanes, Tuesday and each hour (from 0 to 23) of a day.



Now the current data from M2 to M25 is the total traffic volume of each hour every Tuesday. To obtain the average traffic volume, these data need to be divided by 4.

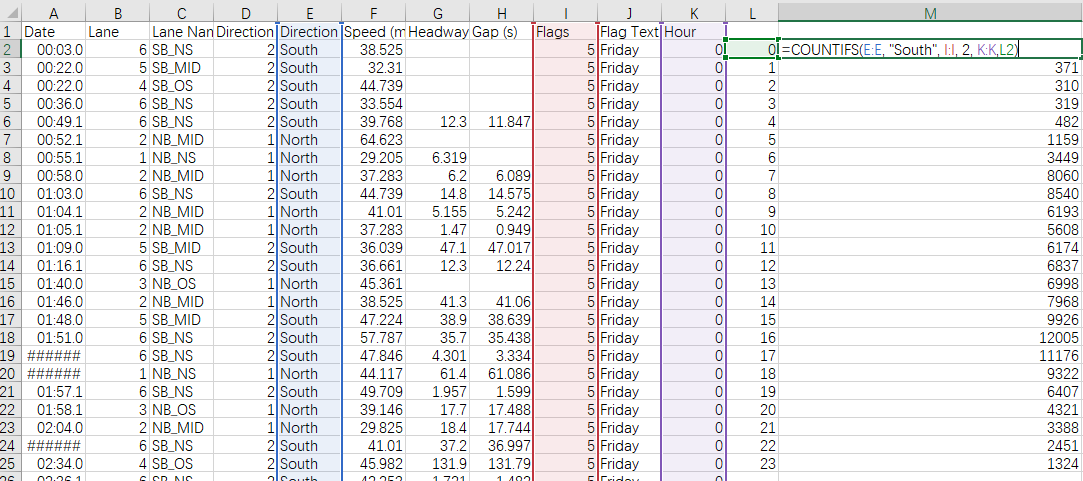


Finally, select L column and N column, go to Insert -> Insert Column or Bar Chart -> More Column Charts… -> Clustered Column then select the second chart to obtain the bar plot. The bar plot is:

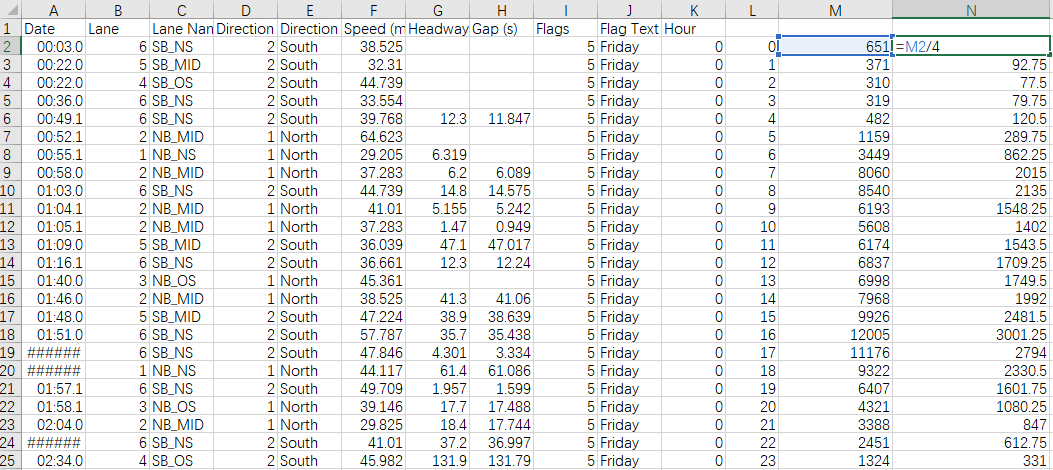


*South Direction*

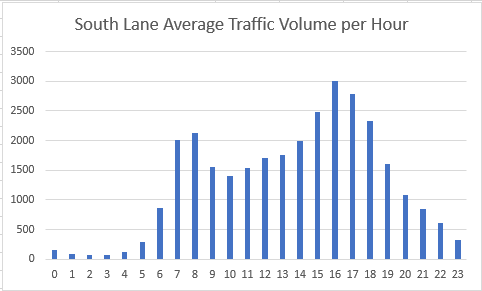
Third use COUNTIFS() function to count the data that satisfied the requirements of south lanes, Tuesday and each hour (from 0 to 23) of a day.



Now the current data from M2 to M25 is the total traffic volume of each hour every Tuesday. To obtain the average traffic volume, these data need to be divided by 4.



Finally, select L column and N column, go to Insert -> Insert Column or Bar Chart -> More Column Charts… -> Clustered Column then select the second chart to obtain the bar plot. The bar plot is:



**Python and Excel Comparison**

**Similarity**

They can all be used for further processing, analysis and visualization of the data. Both have functions already defined and can be used directly. Both can operate on csv files.

**Python**

Advantage:

1. In addition to some self-contained functions, it is easy to download third-party libraries to make data processing simpler and more diverse (suck as machine learning).

2. More efficient when handling large amounts of data.

3. By writing the code once, repetitive operations can be performed.

4. The code is readable and can be saved, so it can be disseminated and improved.

5. It can be used in multiple platforms.

6. Can read and process many types of files

Disadvantage

1. Relatively difficult to learn.

2. May face various bugs.

**Excel**

Advantage

1. It has a graphical interface. Process data with a mouse click without writing a lot of code. More intuitive operation.

2. Easy to learn, excel can be used by short learning time.

3. High penetration rate, most computers have Excel.

Disadvantage

1. It gets laggy when processing large amounts of data

2. When performing the same operation on data, the operation needs to be repeated.

3. The process is not recorded after processing.