**DATA MANAGEMENT PROJECT REPORT**

(Project Semester: August-December 2019)



***DIVVY BIKESHARE 2018 ANALYSIS***

Submitted by

Sourin Ghosh

11907705

Program and Section: B.Tech(CSE), KM004

Course Code: INT217

Under the Guidance of

**Sandeep Kaur**

**Discipline of CSE/IT**

**Lovely School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that SOURIN GHOSH bearing Registration no. 11907705 has completed INT217 project titled, **“BIKE SHARE 2018 ANALYSIS”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

**Sandeep Kaur**

**School of Computer Science & Engineering**

**Lovely Professional University**

**Phagwara, Punjab.**

Date: 15 December 2021

**DECLARATION**

I, Sourin Ghosh, student of B.Tech CSE under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 15 December 2021

Registration No.: 11907705 Sourin Ghosh

**ACKNOWLEDGEMENT**

I would like to express my special thanks of gratitude to my teacher Mrs. Sandeep Kaur who gave me the golden opportunity to do this wonderful project of analysis of the data of a bikeshare in Chicago and Ney York namely “DIVVY BIKESHARE 2018” which also helped me in doing a lot of research and I came to know about so many new things. I am thankful to them.Secondly, I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

Furthermore, I would also like to acknowledge with much appreciation the crucial role of the Tableau Builder software and MS Excel who gave the tools to complete the task of building a Dashboard. A special thanks goes to my class and YouTube for providing the necessary information about the working of software. Last but not least, many thanks go to the Kaggle website for providing the raw data for the project.

**Table of Content**

1. Introduction

2. Scope of the Analysis

3. Source of dataset

4. ETL process

5. Analysis on dataset (for each analysis)

1. Introduction
2. General Description
3. Specific Requirements, functions and formulas
4. Analysis results
5. Visualization

6. References

7. Bibliography

**INTRODUCTION**

**Data Analysis** is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

This dataset contains two worksheets one having bikeshare data of Chicago city and other having the bikeshare data of New York city

.

BIKE SHARE 2018 ANALYSIS contains the following data fields: -

* Trip Duration – Total duration of bike being used in seconds
* Start Time – Exact date and time when a user unlocked the bike for rent
* Stop Time – Exact date and time when user leaves the bike after use
* Start Hour – Exact hour when user unlocked the bike for rent
* Weeknum – Weekday represented in number format
* Weekday – The weekday on which bike is taken
* Start Station ID – Station ID of station from where user started the trip
* Start Station Name – Name of station from where user started the trip
* End Station ID - Station ID of station where user ended the trip
* End Station Name - Name of station where user ended the trip
* Bike ID – Unique ID of a particular bike
* User Type – Type of user whether he/she is subscriber or customer
* Gender – Gender of user renting the bike for a trip

**SCOPE OF ANALYSIS**

The bike share company wants to see and analyze the sales trend and work upon the lagging segments and outperforming employees accordingly. The Analytics team also wants to create analyze the database in depth to help the company grow exponentially. The Analytics team wishes to answer the following objectives for both Chicago and New York: -

1. Total duration of bike usage in New York and Chicago separately.

2. Gender distribution

3. Type of user using the bike

4. Top 10 busy stations in New York and Chicago separately

5. Average trip durations according to weekdays

6. Total trips according to hours

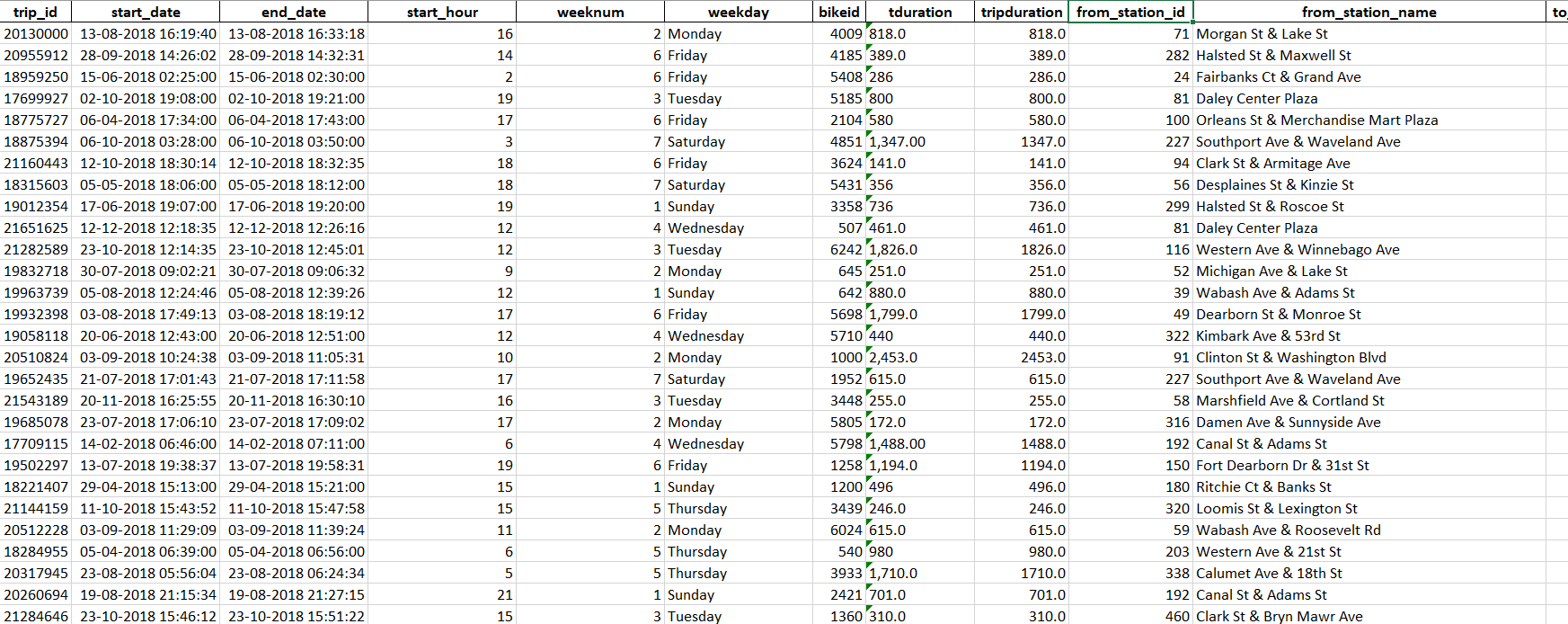
Aim of this project is to answer the above objectives in the form of visualization by creating a dashboard to convey the answers effectively and efficiently.

**ETL PROCESS**

In computing, extract, transform, load (ETL) is a process in database usage to prepare data for analysis, especially in data warehousing. Data extraction involves extracting data from homogeneous or heterogeneous sources, while data transformation processes data by transforming them into a proper storage format/structure for the purposes of querying and analysis; finally, data loading describes the insertion of data into the final target database such as an operational data store, a data mart, or a data warehouse. A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions.

Precisely, ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. ETL stands for Extract, Transform and Load.

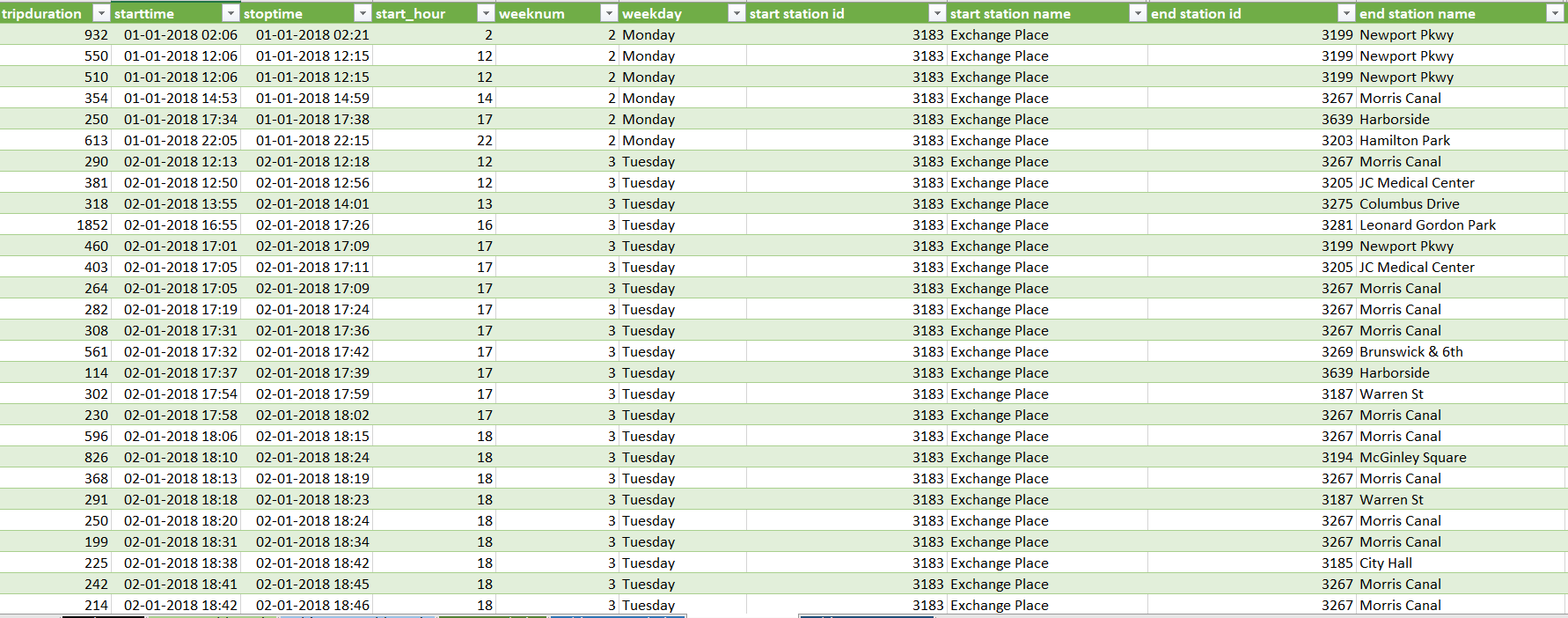
Before ETL, the dataset looked like this. **This data is taken from Kaggle.**



Through the process of ETL, we are going to clean the dataset and bring all the entities to their proper data format.

**Step 1: Removing the blank cells from the dataset.**

For this, select the whole dataset. Go to Find and Select in the Home tab of excel. Select Go to Special from the drop-down menu and then tick the blank option. All the blank cells will be selected. Then go to Delete option in the home tab again and select Delete Rows from the drop-down menu. This will remove any rows with blank cells.



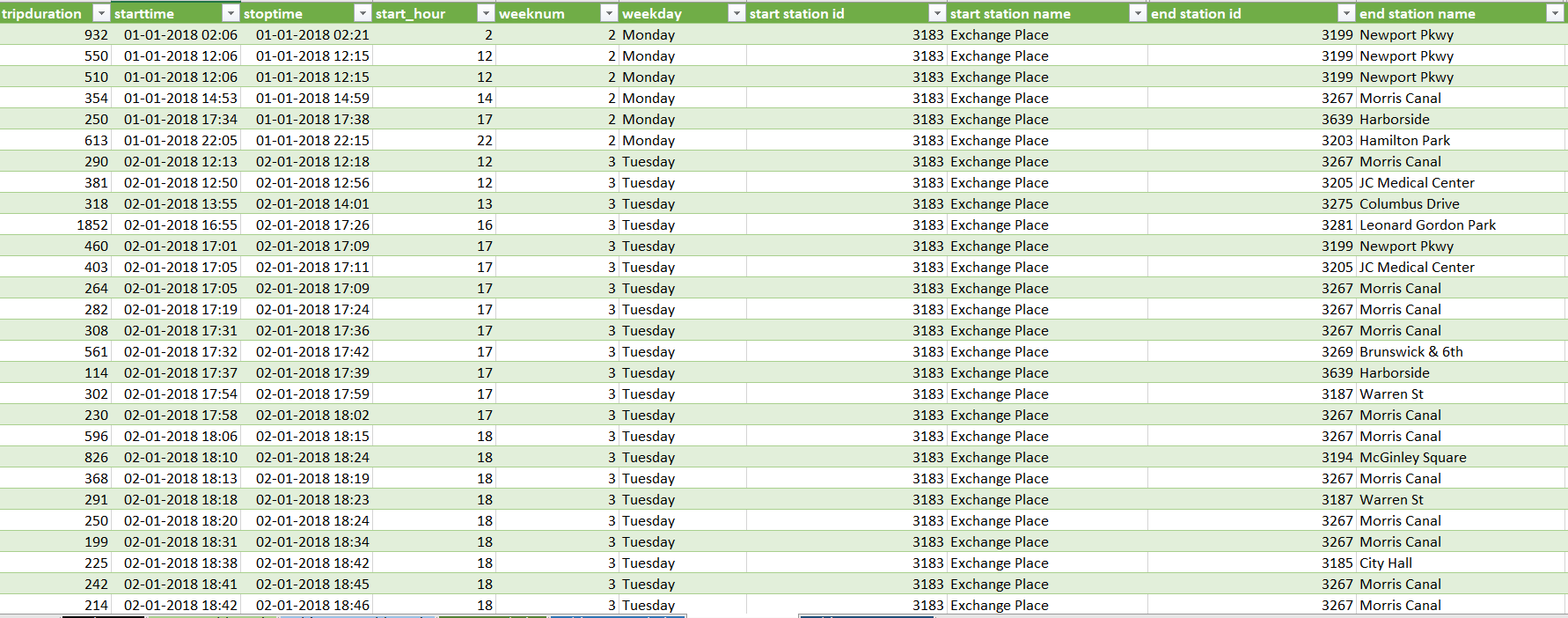
**Step 2: Removing columns which are not properly defined or not crucial to our analysis.**

For this we will columns which are redundant like the column with just the index numbers. For this we will select that particular column and then go to delete option in the home tag and then select Delete Columns from the drop-down menu.



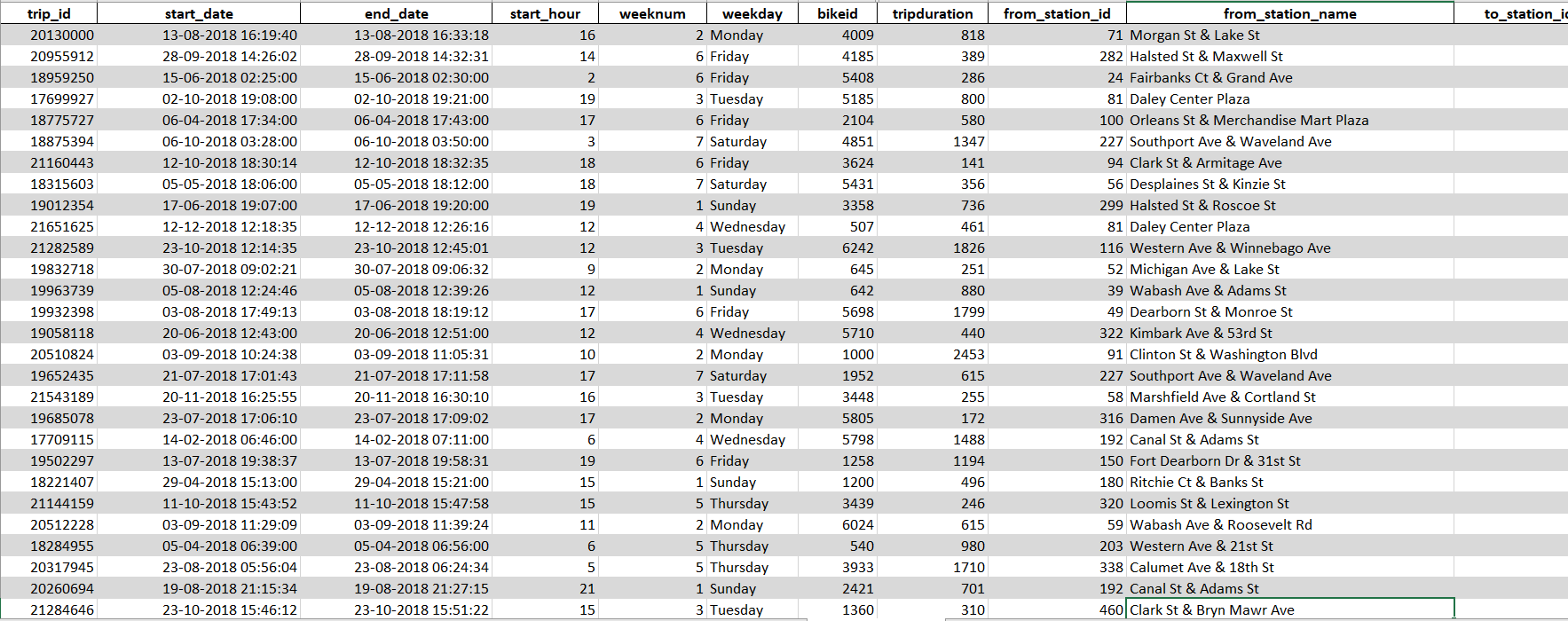
**Step 3: Giving proper and appropriate column names.**

The dataset does not have proper columns so our next step would be to giver proper column names to the columns wherever required.



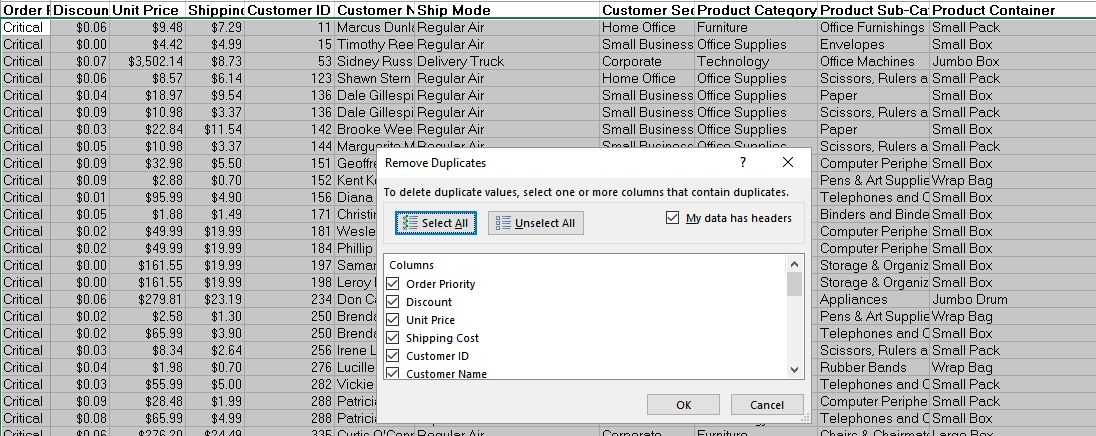
**Step 4: Improvising Proper Data Formatting**

Without proper Data Formatting, proper analysis will not take place. So, we will bring down certain columns to their proper format. For example, the dates should be in the date format and price and sales should be in currency format for better results.



**Step 5: Removing Duplicate Values**

It might be possible that our data may be containing duplicate values which may hinder in precise analysis. So, our last task in ETL will be removing duplicate values and making our data perfect for analysis.



**ANALYSIS OF DATASET**

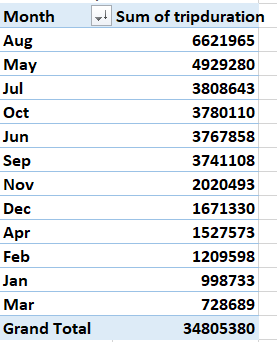
1. **Total duration of bike usage in New York and Chicago separately:**

**Description:**

In this we calculate total duration of trip every month in seconds like how many seconds of trips occurred in a month. For this we sum up the total duration of bike used in a trip in a specific month. From this we can get to know in which month people are going on trips more

**Results:**

**For NEY YORK:**



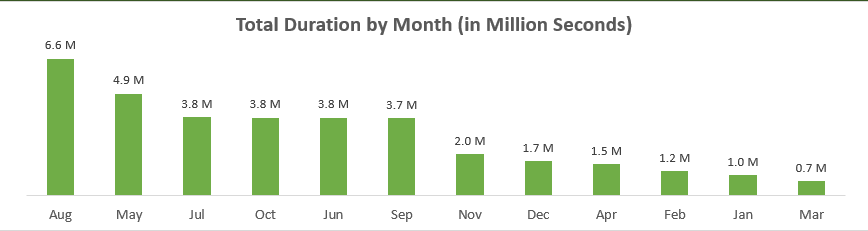
**For CHICAGO:**



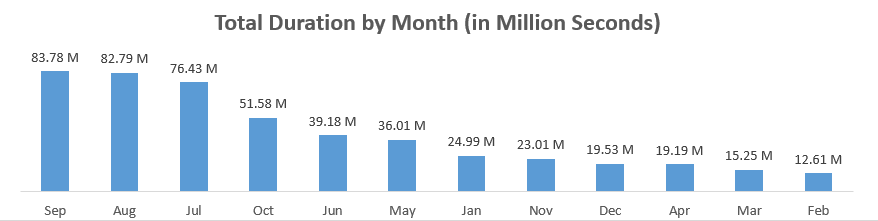
**Visualization:**

The results are then visualized in the form of a bar graph for both Ney York and Chicago

**For NEY YORK:**



**For CHICAGO:**



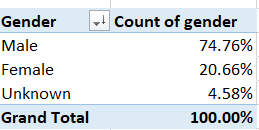
1. **Gender distribution:**

**Description:**

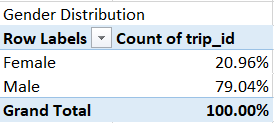
It is used for calculating the current trend of the number of trips taken by male gender and number of trips taken by female gender basically gender distribution

**Results:**

**For NEY YORK:**



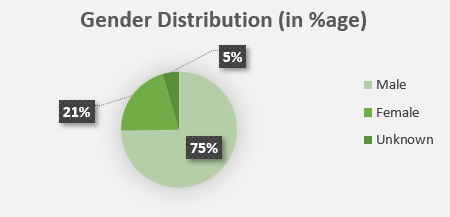
**For CHICAGO:**



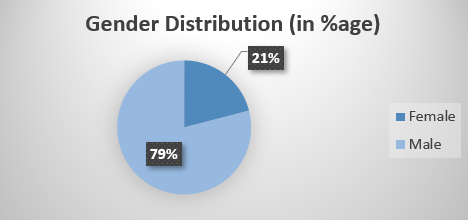
**Visualization:**

We will use pie chart to visualize the distribution.

**For NEY YORK:**



**For CHICAGO:**



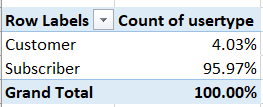
1. **Type of user using the bike**

**Description:**

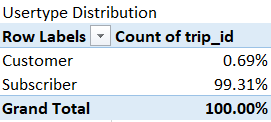
Describes the total percentage of subscriber using the bike and percentage of normal customer taking a ride based on this company determines amount of people taking subscriptions.

**Results:**

**For NEY YORK:**



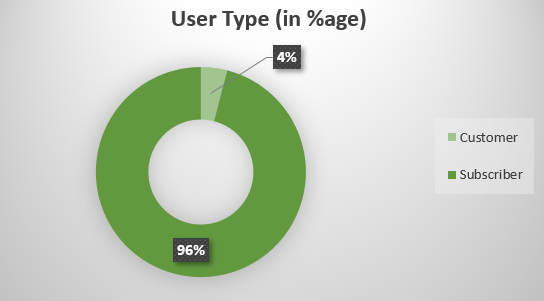
**For CHICAGO:**



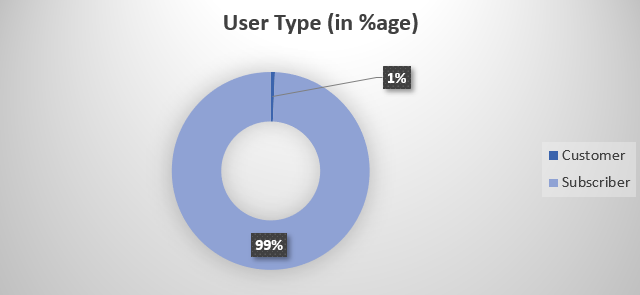
**Visualization:**

The results are visualized with the help of pie chart

**For NEY YORK:**



**For CHICAGO:**



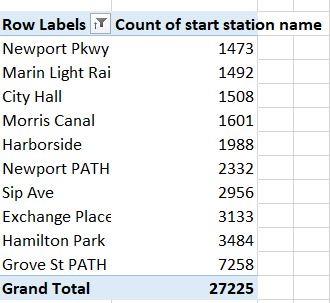
1. **Top 10 busy stations in New York and Chicago separately**

**Description:**

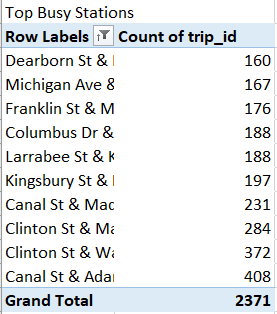
Determines the top 10 busy stations based on most number of trips taken from a particular station

**Results:**

**For NEY YORK:**



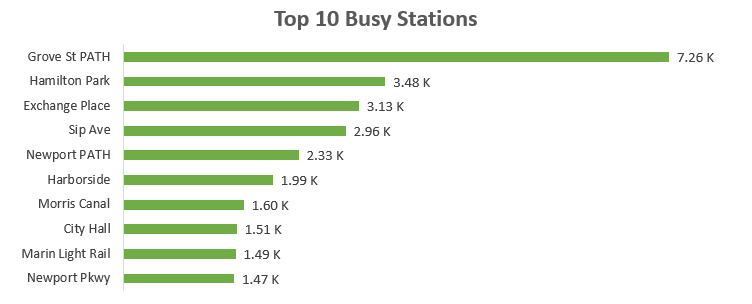
**For CHICAGO:**



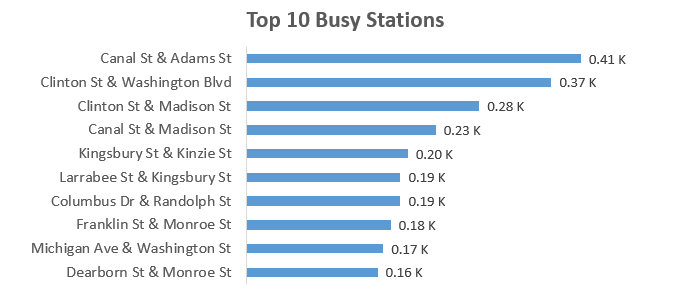
**Visualization:**

We visualize the above results with the help a line graph

**For NEY YORK:**



**For CHICAGO:**



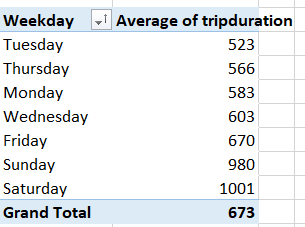
1. **Average trip durations according to weekdays**

**Description:**

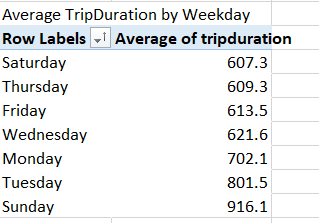
It finds average trips taken on weekdays for each weekday it gives the data of how many trips have been taken in average on a particular weekday

**Results:**

**For NEY YORK:**



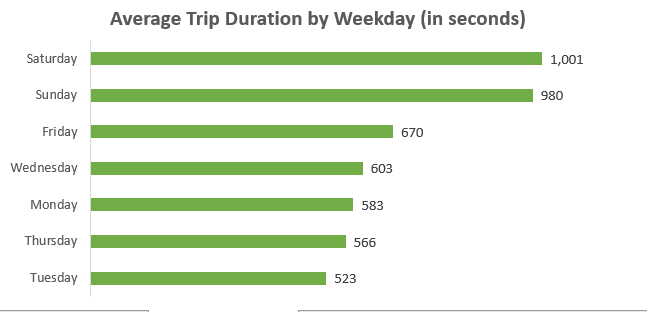
**For CHICAGO:**



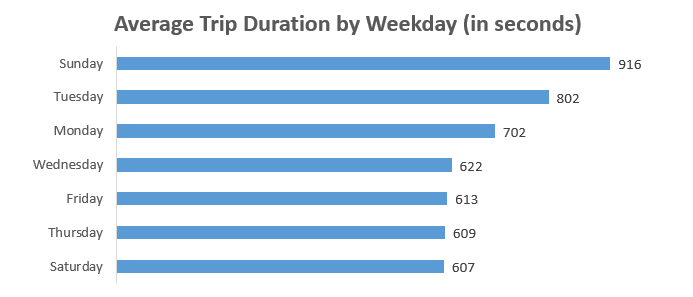
**Visualization:**

The results are visualized in the form of stacked bar graph.

**For NEY YORK:**



**For CHICAGO:**



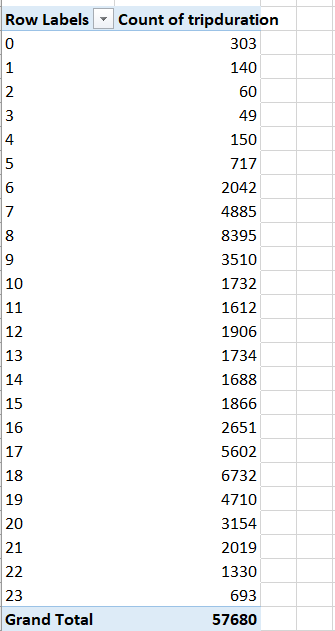
1. **Total trips according to hours**

**Description:**

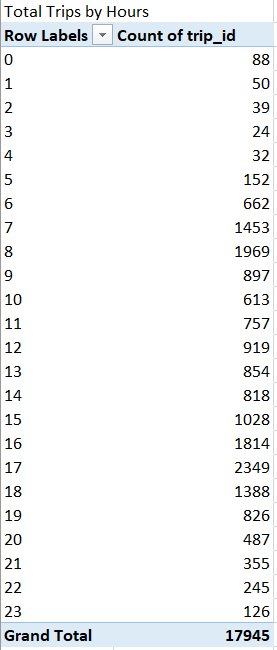
In this we analyze for a particular hour like for one hour or two hour how many trips have been done by users basically analyzing total number of trips according to hours

**Results:**

**For NEY YORK:**

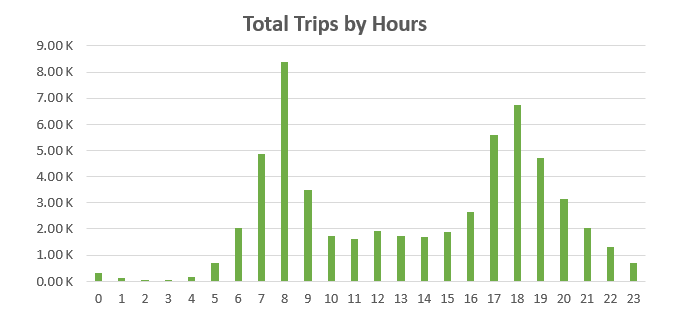


**For CHICAGO:**

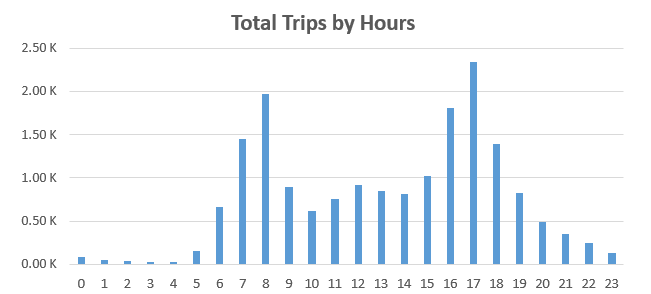


**Visualization:**

**For NEY YORK:**

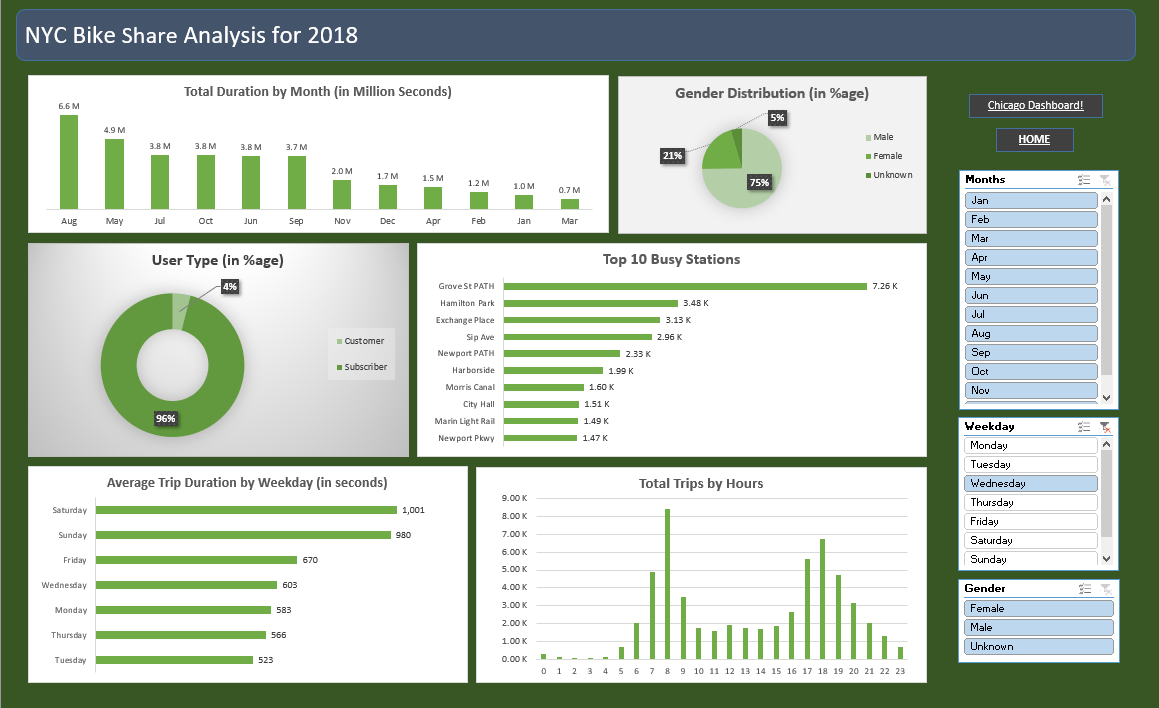


**For CHICAGO:**

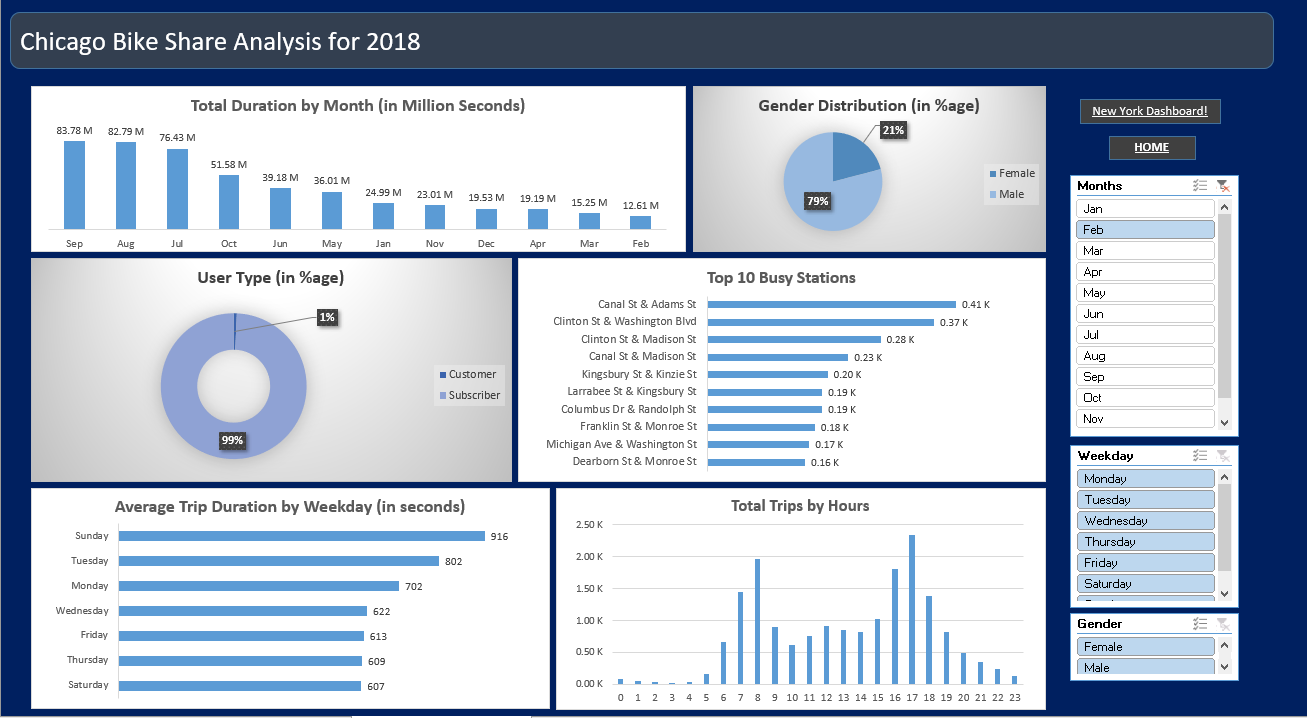


**Final Dashboard**

**For NEY YORK**



**For CHICAGO:**



**REFERENCES AND BIBLIOGRAPHY**

* The first reference would be Kaggle.com for providing me with the dataset.
* 2nd reference would be the myanimelist.com for giving me the idea to work on.
* 3rd reference would be tutorials point.com for clarifying my doubts and queries in the topic.
* Tableau.com also helped me in guiding me through the use of tableau in cleaning of data.
* Lastly the practical’s provided by the course teacher were the key player in making of this project.