

**INTRODUCTION TO DATA MANAGEMENT PROJECT REPORT**

(Project Semester August-November 2020)

***BIKE SHARE ANALYSIS (USAGE OF EXCEL)***

Submitted by

Naidu Nagendra Babu

Registration No: 11805086 Programme:B.Tech

Roll no:B44

Section: KM076 Course Code: INT217

Under the Guidance of

**MS VASUDHA**

**Discipline of CSE/IT**

**Lovely School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

**CONTENTS OF THE REPORT**

* Cover page
* Certificate
* Declaration
* Acknowledgement
* Table of Content
  1. Introduction
  2. Scope of the Analysis
  3. Existing System
  4. Source of dataset
  5. ETL process
  6. Analysis on dataset (for each analysis)
     1. Introduction
     2. General Description
     3. Specific Requirements, functions and formulas
     4. Analysis results
     5. Visualization
  7. Future scope
  8. References
  9. Bibliography

**CERTIFICATE**

This is to certify that NAIDU NAGENDRA BABU bearing Registration no. 11805086 has completed INT 217 project titled, **“BIKE SHARE ANALYSIS CHICAGO (Usage of EXCEL)”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

**Signature and Name of the Supervisor Designation of the Supervisor**

**School of Computer Science and Engineering**

Lovely Professional University Phagwara, Punjab.

**DECLARATION**

I, NAIDU NAGENDRA BABU student of B.Tech 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: Signature:

Registration No. 11805086 Name of the student: Naidu Nagendra Babu

Acknowledgement

*I am grateful to my instructor MS.VASUDHA.. She has truly been a great source of inspiration, constructive criticism, insight and input. The knowledge she has imparted and the patience they have displayed was vital in completing this study. I am thankful for the immense encouragement and morale support that she graciously provided during this study.*

*I owe much of my academic and personal success to my parents, who, by example, provided me with the motivation and courage to pursue this course in the field of my interest. Special thanks to all my friends, near and far, for their love and support that made me finish my Project and its report successfully.*

Naidu Nagendra babu

**Table of Contents**

**Introduction**

**Scope Of Analysis**

**Objective**

**Existing System(Drawbacks and Alternatives)**

**Sources of Data Set**

**ETL Process**

What is ETL?

Why do we need ETL?

ETL process in Data Warehouses Modern ETL Process

**Analysis on Dataset**

Introduction

Concepts of Excel used in Analysing the Data. Specific requirements ,functions and formulas Analysis of Results

Visualization

**Future Scope**

**References**

**Bibliography**

# INTRODUCTION ON BIKE SHARE ANALYSIS:

Over the past decade, bicycle-sharing systems have been growing in number and popularity in cities across the world. Bicycle-sharing systems allow users to rent bicycles for short trips, typically 30 minutes or less. With the latest technologies, it is easy for a user of the system to access a dock within the system to unlock or return bicycles. These technologies also provide a wealth of data that can be used to explore how these bike-sharing systems are used.

In this project, we will perform an exploratory analysis on data provided by Motivate, a bike- share system provider for many major cities in the United States. We will compare the system usage between three large cities: New York City, Chicago, and Washington, DC. We will also see if there are any differences within each system for those users that are registered, regular users and those users that are short-term, casual users.

## WHAT IS MEANT BY BIKE SHARE ANALYSIS:



A **bicycle-sharing system**, **public bicycle system**, or **bike-share scheme**, is a service in which bicycles are made available for shared use to individuals on a short term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" and return it at another dock belonging to the same system. Docks are special bike racks that lock the bike, and only release it by computer control. The user enters payment information, and the computer unlocks a bike. The user returns the bike by placing it in the dock, which locks it in place. Other systems are dockless.. For many system.Smartphone mapping apps Show nearby available bikes and open docks.

## HISTORY OF BIKE SHARING:

The first bike sharing projects were initiated by local community organisations, or as charitable projects intended for the disadvantaged, or to promote bicycles as a non-polluting form of transport, or they were business enterprises to rent out bicycles.

The first documented bike-share project began in Europe in 1965, the group painted fifty bicycles white and placed them unlocked for everyone to use freely.

Ernest Callenbach's novel *Ecotopia*(1975) illustrated the idea. In the utopian novel of a society that does not use fossil fuels, Callenbach describes a bicycle sharing system which is available to inhabitants and is an integrated part of the public transportation system.

In 1995 a system of 300 bicycles using coins to unlock the bicycles in the style of shopping carts was introduced in Copenhagen . It was initiated by Morten Sadolin and Ole Wessung. The idea was developed by both Copenhageners after they were victims of bicycle theft one night in 1989.

Bike share technology has evolved over the course of decades. Of the world's 15 biggest public bike share programs, 13 are in China. In 2012, the biggest are in Wuhan and Hangzhou, with around 90,000 and 60,000 bikes respectively.

# SCOPE OF ANALYSIS:

In order to cover small distances in quick time and for the benefit of the health purposes it functions effectively.Not only do bike-share systems intend to reduce traffic congestion, they also aim to reduce air pollution through decreased automobile usage. The study on D.C.'s Capital Bikeshare estimated that the reduction in traffic congestion would be equivalent to roughly $1.28 million in annual benefits, accrued through the reduction in congestion- induced CO2 emissions. A separate study of transportation in Australia estimated that 1.5 kilograms of CO2 equivalent emissions are avoided by an urban resident who travels 5 kilometerr by cycling rather than by car during rush hour periods.



Beside the people who ride without a helmet and injury their head, bicycle-sharing systems has a positive health effects to people overall.Cycling is a good way for exercise and stress relief. It can increase recreation and improve sociability of a city, which make people live more happy and relax. The report from Centers for Disease Control and Prevention (CDC) point out that cycling also help preventing disease like obesity, heart disease (can reduce up to 82%) and diabetes (can reduce up to 58%). Therefore, bicycle-sharing systems has a positive effect on metal and physical health, which attract more people to use.

**Objectives of this project:**

* Most Common Hours
* Most Popular Stations
* Top and Avg High Duration
* User and Gender Analysis
* Common Start Week Day

**Some more objectives of this project:**

* To good hand on excel.
* To learn the ETL process
* How to link one sheet to another
* How to use pivot table
* Learn to make dashboard.
* To make different type of graphs
* To learn how to fetch data from other source to excel.
* How to protect Workbook

# 3.EXISTING SYSTEM:

**LIMITATIONS OF EXISTING SYSTEM:**

Despite their theoretical and observed benefits, bike-share programs have come under attack as their presence has grown throughout the world. Much of this criticism has focused on the use of public funding - concerned critics posit that the use of tax dollars for bike-share programs should instead be diverted towards building or maintaining roads and other services that more residents use on a daily basis. However, this argument relies on a faulty assumption that taxpayer money is a significant source of bike-share funding. An analysis by People for Bikes, an organization that advocates for new and safe bike infrastructure, found that public investment in Salt Lake City’s Greenbike and Denver's [B-Cycle](https://en.wikipedia.org/wiki/B-cycle) programs was significantly less than traditional public transit (e.g. bus or rail) in those same cities, on a per-trip basis. Both Greenbike and B-Cycle's publicly funded subsidies amount to 10 percent or less of the total cost of one trips. In contrast, Salt Lake City's bus and rail system (UTA) relies on 80 percent public funding for a single trip. Other critics claim that bike-share programs fail to reach more low-income communities. Some efforts have attempted to address this issue, such as New York City's Citi Bike's discounted membership program, which is aimed at increasing ridership among low- income residents. However, around 80 percent of study respondents reported that they had no knowledge of the program's discount.

*UCED CAR PARKING:*

Bike-share programs, especially the earlier services that required docking areas along urban streets, may encroach upon the space available for on-street car parking. Reduced car parking is therefore a negative externality, which is off-set by six to eight bikes fitting into one car park. As bike-share companies have transitioned into dockless programs, this effect may have been reduced.

**4** .**SOURCES OF DATASET:**

I have took the data from the Kaggle website. here I could get the much relevant data where I could perform the operations on the data very well. It also provides various projects on Artificial Intelligence,Machine Learning,Mobile development ,Web development,Android development.The data set that I took clearly explains about all the necessary things that I want to perform the operations.This website is one of the most popular ,where one can learn,perform and their strengths.

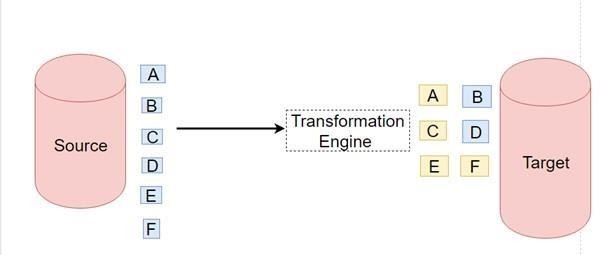
Link: https://www.kaggle.com/deepak525/us-bike-share-analysis

# 5.EXTRACTION TRANFORM AND LOAD:

#### ETL is an abbreviation of Extract, Transform, and Load. In this process, an ETL tool extracts the data from different RDBMS source systems then transforms the data like applying calculations, concatenations, etc. and then load the data into the Data Warehouse system.

It’s tempting to think a creating a Data Warehouse is simply extracting data from multiple sources and loading into database of a Data Warehouse. This is far from the truth and requires a complex ETL process. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.

In order to maintain its value as a tool for decision makers, Data Warehouse system needs to change with business changes. ETL is a recurring activity (daily, weekly, monthly) of a Data Warehouse system and needs to be agile, automated, and well documented.



**Why do we need ETL?**

There are many reasons for adopting ETL in the organization:

It helps companies to analyze their business data for taking critical business decisions.

Transactional databases cannot answer complex business questions that can be answered by ETL.

A Data warehouse provides a common data repository.

ETL provides a method of moving the data from various sources into a data warehouse.

As data sources change, the Data Warehouse will automatically update.

Well-designed and documented ETL system is almost essential to the success of a Data Warehouse project.

Allow verification of data transformation, aggregation and calculations rules.

ETL process allows sample data comparison between the source and the target system.

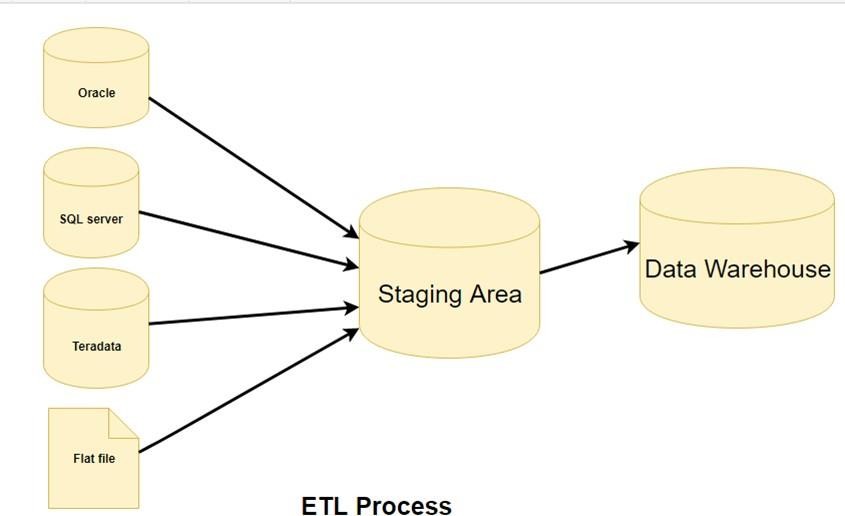
ETL process can perform complex transformations and requires the extra area to store the data.

ETL helps to Migrate data into a Data Warehouse. Convert to the various formats and types to adhere to one consistent system.

ETL is a predefined process for accessing and manipulating source data into the target database.

ETL offers deep historical context for the business.

It helps to improve productivity because it codifies and reuses without a need for technical skills.



**ETL Process in Data Warehouses**

#### ETL is a 3-step process: Step – 1: Extraction

Step – 2: Transformation

Step – 3: Loading

**Extraction:**

In this step, data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Data Warehouse

database, rollback will be a challenge. Staging area gives an opportunity to validate extracted data before it moves into the Data Warehouse.

Data Warehouse needs to integrate systems that have different DBMS, Hardware, Operating Systems and Communication Protocols. Sources could include legacy applications like Mainframes, customized applications, Point of contact devices like ATM, Call switches, text files, spreadsheets, ERP, data from vendors, partners amongst others

Hence one needs a logical data map before data is extracted and loaded physically. This data map describes the relationship between sources and target data.

***Three Data Extraction methods***

#### Full Extraction

* 1. Partial Extraction – without update notification.
  2. Partial Extraction – with update notification

Irrespective of the method used, extraction should not affect performance and response

time of the source systems. These source systems are live production databases. Any slow down or locking could effect company’s bottom line.

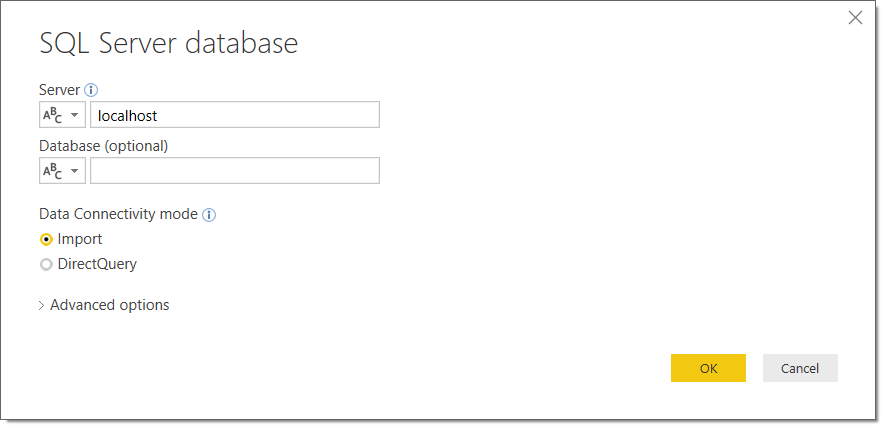
**Some validations are done during Extraction**

* Reconcile records with the source data
* Make sure that no spam/unwanted data loaded
* Data type check
* Remove all types of duplicate/fragmented data
* Check whether all the keys are in place or not

**Extraction of My data from the server:**

The data for the sales analysis I have extracted the data from the sql server database of edx site belongs to the united states of America.

To extract the data from the sql server you need to go to data in excel and the click on get data and then choose get data from the data base and then press from sql server



In the server you need to enter: msedxeus.database.windows.net In the data base enter : DAT206x01

Then choose multiple option there will be sales data table , manufacturer data table , Locations data table, products data table.

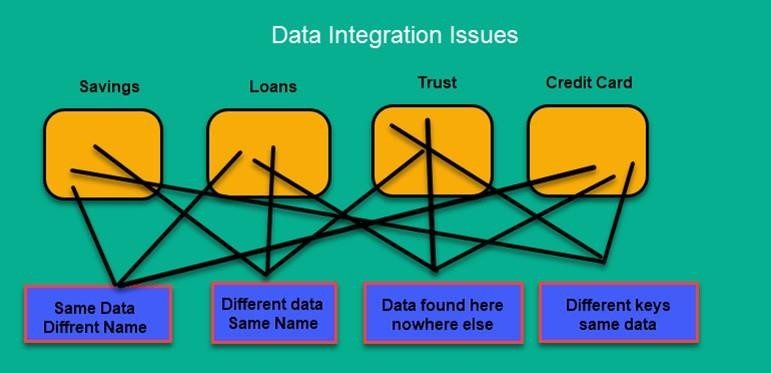
The next step is to transform the data.

**Transform :**

Data extracted from source server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data such that insightful BI reports can be generated.

In this step, you apply a set of functions on extracted data. Data that does not require any transformation is called as **Direct move** or **Pass through data.**

In transformation step, you can perform customized operations on data. For instance, if the user wants sum-of-sales revenue which is not in the database. Or if the first name and the last name in a table is in different columns. It is possible to concatenate them before loading



**Validations are done during this stage**

Filtering – Select only certain columns to load

Using rules and lookup tables for Data standardization Character set conversion and encoding handling

Conversion of units of measurements like Date Time Conversion, Currency conversions, Numerical Conversions, etc.

Data threshold validation check. For example, age cannot be more than two digits. Data flow validation from the staging area to the intermediate tables.

Required fields should not be left blank.

Cleaning ( for example, mapping NULL to 0 or Gender Male to “M” and Female to “F” etc.)

Split a column into multiples and merging multiple columns into a single column. Transposing rows and columns

Use lookups to merge data

Using any complex data validation (e.g., if the first two columns in a row are empty then it automatically reject the row from processing)

**Loading :**

Loading data into the target data warehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period. Hence, load process should be optimized for performance.

In case of load failure, recover mechanisms should be configured to restart from the point of failure without data integrity loss. Data Warehouse admins need to monitor, resume, cancel loads as per prevailing server performance.

**Types of Loading**

**Initial Load** – Populating all the Data Warehouse tables

**Incremental Load** – Applying ongoing changes as when needed periodically.

#### **Full Refresh** – Erasing the contents of one or more tables and reloading with fresh data.

**Load Verification**

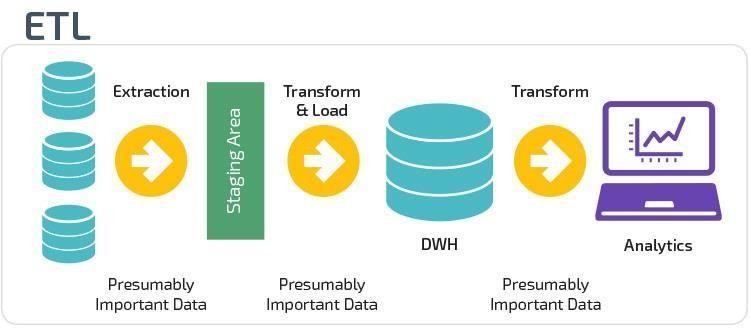
Ensure that the key field data is neither missing nor null.

#### 1.Test modelling views based on the target tables.

#### 2.Check that combined values and calculated measures.

#### 3.Data checks in dimension table as well as history table.

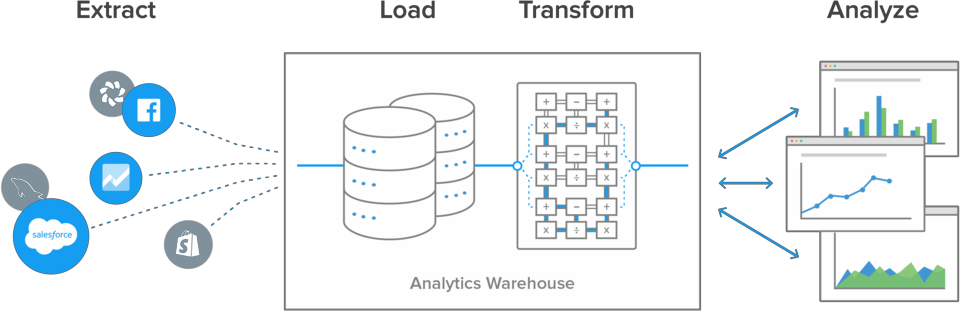
4.Check the BI reports on the loaded fact and dimension table.



**Modern ETL Process**

Modern technology has changed most organizations approach to ETL, for several reasons. The biggest is the advent of powerful analytics warehouse like Amazon Redshift and Google BigQuery. These newer cloud-based analytics databases have the horsepower to perform transformations in place rather than requiring a special staging area.

Also, data today is frequently analyzed in raw form rather than from preloaded OLAP summaries. This has led to the development of lightweight, flexible, and transparent ETL systems with processes that look something like this:



The biggest advantage to this setup is the transformations and data modelling happen in the analytics database. This gives the BI team, data scientist, and analysts greater control over how they work with it, in a common language they all understand.

**ETL PROCESS ON MYDATA:**

The data is extracted from the Kaggle.com The format of the data available on the website was csv format to perform analysis. The data were from 2016 Jan to 2016 dec. After extraction of data the cleaning process is performed all null values were changed into standardized value. The identifiers were made according to the need and representation. The attribute having date as columns were changed accordingly as they were not proper and jumbled . I have also added some of the column such as day, start of month in order to evaluate the data easily.It helps in understanding the data quite easily. Finally atlast I have converted the csv file excel format.

# 6.ANALYSIS ON DATASET:

Analysis is done on five objectives naming analysis on individual most start month, analysis of total sum of travel time and average of travel time, analysis of total number of user types

Whether they are customers or subscribers, analysis on the most common start station where the user picks up the bike and where the user drops the bike, analysis of most common start and end station,analysis on the most common start hour and the total trips done.

**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.

The process of data analysis:

* Data requirements
* Data collection
* Data processing
* Data cleaning
* Modeling and algorithms

**Specific Requirements, functions and formulas:**

We use different formulas for data cleaning, data processing and modelling and algorithms.

Main function of excel that I have used are

* Power pivot
* Pivot table
* Conditional formatting
* Create graph using power pivot
* Create relation between different table
* Function analyser (Slicer)
* Different types of graph
* Different functionality of graph

**Data visualization:**

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data in the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Common general types of data visualization:

* Charts
* Tables
* Graphs
* Maps
* Dashboard

### INTRODUCTION:

The first most analysis that I have done is the most common start hour of the whole year in CHICAGO .We may start travel at any point of the time.So I generally took the most common point of hour where the most people do start their journey. I did along with the trip id or bike id of the person who had started the journey from the particular station.

### SPECIFIC REQUIREMENTS,FUNCTIONS AND FORMULAS:

The analysis requires Microsoft excel 2016 or above, requires pivot table and different type of graphs. The function and formula of sum and product is used for calculation of percentage. It requires curiosity and lots of interest in analysis

### ANALYSIS:

### 

**6.1.4:VISUALIZATION:**

**INTRODUCTION:**

The first most analysis that I have done is the most common start station of the whole year in CHICAGO .We may start travel at any station of the time.So I generally took the most common point of stationwhere the most people do start their journey. I did along with the trip id or bike id of the person who had started the journey from the particular station.

### SPECIFIC REQUIREMENTS,FUNCTIONS AND FORMULAS:

The analysis requires Microsoft excel 2016 or above, requires pivot table and different type of graphs. The function and formula of sum and product is used for calculation of percentage. It requires curiosity and lots of interest in analysis.

### ANALYSIS:

### 

**6.2.4.VISUALIZATION :**

### 

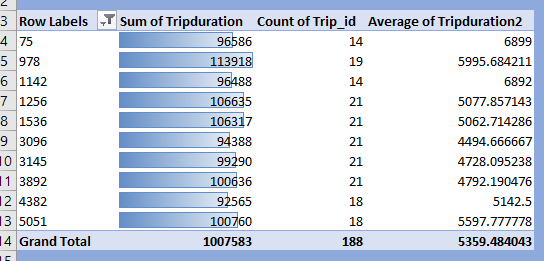
**INTRODUCTION:**

The first most analysis that I have done is the sum and average of all the trpsof the whole year in CHICAGO .We may start travel at any point of the time and at any time.So I generally took start time and stop time where the most people do start their journey. I did along with the trip id or bike id of the person who had started the journey from the particular station.

### SPECIFIC REQUIREMENTS,FUNCTIONS AND FORMULAS:

The analysis requires Microsoft excel 2016 or above, requires pivot table and different type of graphs. The function and formula of sum and product is used for calculation of percentage. It requires curiosity and lots of interest in analysis.

**ANALYSIS:**

****

**6.3.4.VISUALIZATION:**

**INTRODUCTION:**

The first most analysis that I have done is user type,their basic info about gender ,whether they are customers or subcribers of the whole year in CHICAGO . I did along with the trip id or bike id of the person who had started the journey from the particular station.

### SPECIFIC REQUIREMENTS,FUNCTIONS AND FORMULAS:

The analysis requires Microsoft excel 2016 or above, requires pivot table and different typeof graphs. The function and formula of sum and product is used for calculation of percentage. It requires curiosity and lots of interest in analysis.

### ANALYSIS:

### 

### VISUALIZATION:

* + 1. **INTRODUCTION:**

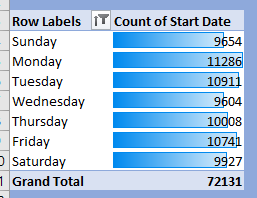
The first most analysis that I have done is the most common start month and day name of the

whole year in CHICAGO .We may start travel in any month of the year and on any day.So I generally took the most common months and days where the most people do start their journey. I did along with the trip id or bike id of the person who had started the journey from the particular station.

**SPECIFIC REQUIREMENTS,FUNCTIONS AND FORMULAS:**

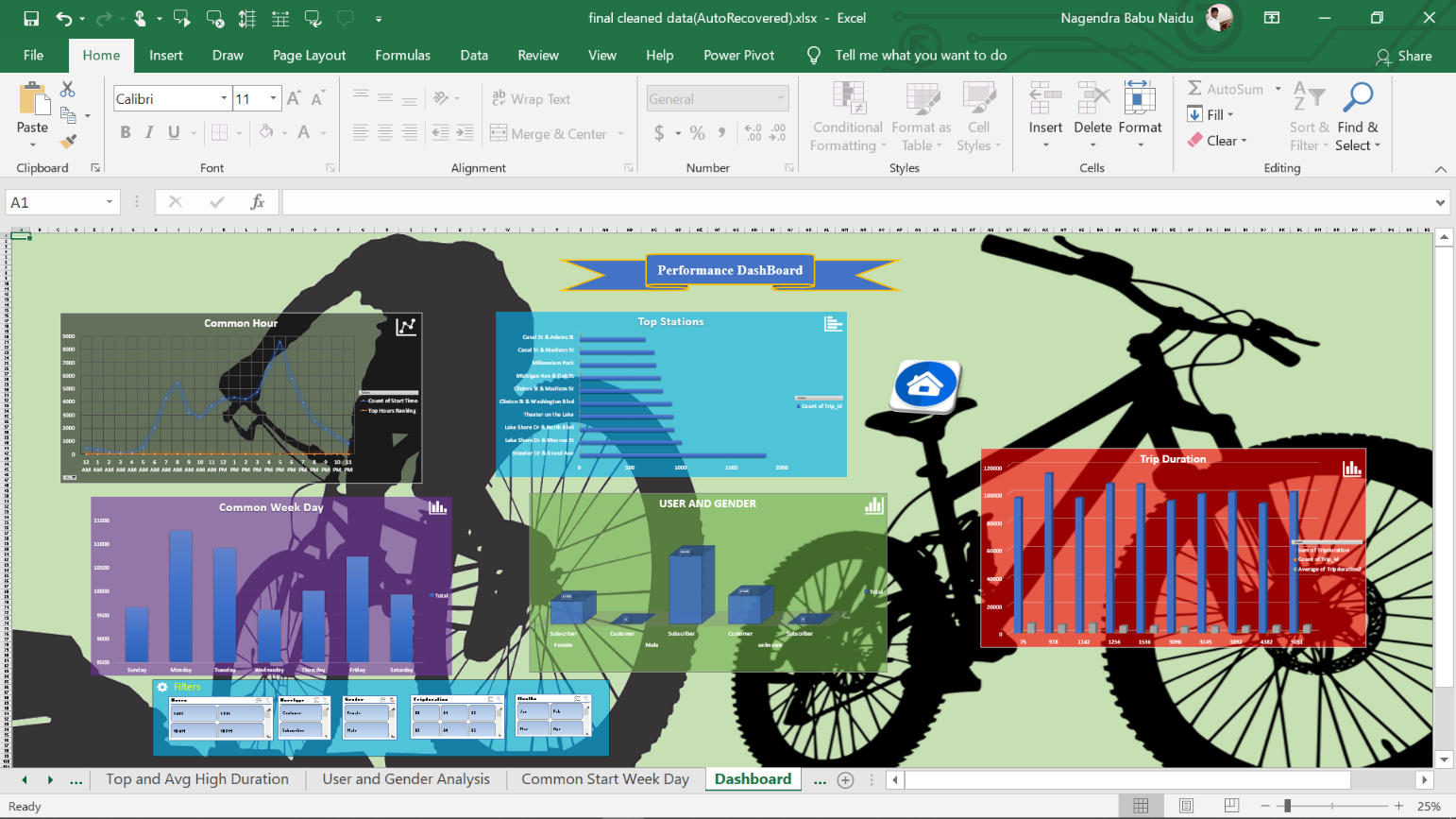
The analysis requires Microsoft excel 2016 or above, requires pivot table and different type of graphs. The function and formula of sum and product is used for calculation of percentage. It requires curiosity and lots of interest in analysis.

**ANALYSIS:**



**VISUALIZATION:**

**DashBoard:**

****

**DrawBacks:**

* No Dynamic Update of details

**FUTURE SCOPE:**

Today, U.S. bike share members have to receive specialized key cards to unlock bikes within their city’s system. A New York Citi Bike member can’t use his or her membership while visiting, say, Boston or Chicago.

In the future, bike share could operate seamlessly with a city’s transit system cards and also work across different cities, according to Jay Walder, CEO of the nation’s largest bike share operator, which [changed its name](http://www.motivateco.com/news/2015/01/14/alta-bicycle-share-becomes-motivate) to [Motivate](http://www.motivateco.com/) (formerly called Alta Bicycle Share) yesterday.

* FLICKR USER [OMAR RAWLINGS](https://www.flickr.com/photos/oc_rawlings/9125396985)

“It’s great that bike share started as a sort of independent thing. It got going quickly that way, the growth has been phenomenal. But it’s maturing and evolving,” says Walder, the former chairman of New York City’s MTA. “We want to see how it actually integrates into other forms of transportation.”

Portland-based Alta Bicycle Share started out with 400 bicycles in Washington, D.C., four years ago and now operates a total of 16,000 bicycles [in 10 metro areas](http://www.motivateco.com/locations) in the U.S., Canada, and Australia, reflecting the huge growth of the bike share model globally. But the company struggled to manage New York City’s Citi Bike, which became the nation’s largest bike share when it launched in 2013. [A recent report](https://comptroller.nyc.gov/newsroom/comptroller-stringer-audit-finds-poor-maintenance-shoddy-oversight-of-the-citi-bike-program/)cited technology glitches and maintenance failures, and others questioned whether the business model, which did not rely on any city funds, worked at its existing membership rate.

In October, new investors, including Equinox and the real estate firm Related Companies, acquired the company, moved its headquarters to Brooklyn, and brought on Walder (whose first hire was a vice president for technology, which the company somehow didn’t have before). In New York, they also announced Citi Bike would expand from 6,000 to 12,000 bikes by 2017–and raised membership fees.

Beyond fixing bicycles and docking stations and tweaking the basic user experience, Walder says he’s now focused on the bigger picture, especially improving connectivity for bike share users by giving them more real-time information and a better ability to report problems and contribute to the improving the system. “It’s about … leveraging a lot of the goodwill that exists with bike share and the way that people really want to help,” he says.

The company also is thinking about how bike share can become part of larger regional transit systems, integrating with existing transit system smart cards and eventually migrating everything to smartphones, says Walder. (Even New York City, where Metro Cards are extremely old-school compared to most cities’ chip-based cards, is [looking at moving](https://www.fastcompany.com/3024544/fast-feed/so-long-metrocard-nyc-subway-cards-will-soon-see-their-last-swipes) from swipes to smartphones by 2019).

**8.References**

1. Youtube.com
2. Google.com
3. Wikipedia.com

# 9.BIBLIOGRAPHY:

1. [http://www.data.gov.in](http://www.data.gov.in.l/)
2. <https://products.office.com/en-in/excel>
3. <https://kaggle.com/datasets>
4. <https://www.excel-easy.com/>
5. [https://www.guru99.com/excel-tutorials](https://www.guru99.com/excel-tutorialsl)
6. <https://www.tutorialspoint.com/excel/>
7. <https://digital.com/blog/excel-tutorials/>
8. <https://www.udacity.com/excel>
9. <https://www.techonthenet.com/excel/index.php>
10. <https://www.javatpoint.com/excel-tutorial>