Rental Bicycle Analysis to Design a Marketing Strategy

Furqan Ali Khan, Uday Bhaskar Gonnabathula, Rajkumar Gara,
Madhurima Madiraju, Vinay Danepalli

Department of Computer Science, Purdue University Fort Wayne

CS 59000 Data Analytics in Business Using R

Dr. Adolfo Coronado

December 11, 2022

Abstract

In this case study, we will be analysing a public dataset for a Bicycle rental company provided by Kaggle. We will be using R programming language for this analysis because of its easy statistical analysis tools and data visualizations.

We have analysed and visualised ride patterns in 12 months to gain insights on possible marketing strategies for converting casual riders into members.

We will determine how the "casual" and annual customers differ in order to develop a targeted and effective marketing message that would encourage the "casual" customers to switch to an annual membership.

The dataset comprises more than 5.7 million observations spread over the financial year (FY) 2022. We will be analysing the complete data to fetch insights about how the members and casual users use the services and how it varied over the months and different seasons. These insights help us to predict the future growth of company and how the revenue of the company would be if most of the casual users are converted into members. We will be comparing the ridership from the previous year FY 2021 and how the company has done after covid in the FY 2022. Using this data, we will be predicting the growth and future of the company. How changes in different parameters effect the growth of this company.

TABLE OF CONTENTS

Introduction

Cyclistic introduced a popular bike-share program in 2016. The initiative has expanded since then to include a fleet of 5,824 bicycles that are geotracked and locked into a system of 692 stations throughout Chicago. The bikes may be released from one station and brought back to any other station in the network.

Up to this point, Cyclistic's marketing approach focused on raising public awareness and appealing to a wide range of consumer groups. The price plans' flexibility, which included single-ride passes, full-day passes, and annual memberships, was one strategy that assisted in making these things possible. Casual riders are those who buy one-ride or all-day passes from the company. Cyclistic members and subscribers are customers who purchase annual memberships.

According to the director of marketing (stakeholder), the company's future prosperity rests on increasing the number.

Using the data we can create a marketing plan based on these findings to turn casual riders into annual members and look toward other parameters to improve business and revenue.

https://www.kaggle.com/datasets/gunnarn/chicago-bicycle-rent-usage

The parameters include bike types, bike pickup location, drop-off location, start and end time, the duration of the rental, and the latitude and longitude of the stations.

- ride_id - This is a unique ID that is associated with each bike. We are not using this as part of the data analysis since we do not have the database that associates bikes with their specific ID.

- Rideable_type - There are different types of bikes, like docked bikes, electric bikes, classic bikes, etc

- Started_at - This is a string that gives the starting date and time of the ride. From this, we can also extract ride length, day, month, etc

- Ended_at - This is a string that gives the ending date and time of the ride.
- Start_station_name - This gives the name of the street where the ride has been started.
- Start_station_id - This is a unique ID that’s associated with the location.
- End_station_name - This gives the name of the street where the ride has ended.
- End_station_id - This is a unique ID that’s associated with the location.
- Start_lat - Gives the latitude of the starting location
- Start_lng - Gives the longitude of the starting location
- End_lat - Gives the latitude of the starting location
- End_lng - Gives the longitude of the starting location

- Member_casual - The consumers of the bike are divided into two types - casual and member. Members have annual subscriptions. Casual riders are those who rent the bike on an hourly basis whenever they need it.

Research Topics

1. Which locations have the highest number of users? For example, bicycle usage in the downtown area vs suburbs.
We are using a descriptive model to analyse this research question. By definition, the descriptive model analyses the previous data to see the upward or downward trend in the data. We use illustrations to depict the highest number of users in various areas like downtown and suburbs. This data is helpful to predict areas with heavy demand.

We plan to use clustering like k means clustering to plot the location data. Then using the mode, we will find the location with the highest sales.

2. How do Location and season affect the rental numbers? People from cold regions are less likely to rent a bicycle in winter.
The number of rentals based on the location and season can be analysed using a linear regression model. Input to the model is the location and seasonal data. The model randomly distributes the input data into training and testing data. We can use training data for training the model as well as cross-validating the model with the k-fold cross-validation method. The testing data can be used to evaluate the accuracy or metrics of the model. We can use this model to predict the number of rentals in various locations under different seasons. This data is helpful to predict the demand for bicycles in all areas.

We want to use a regression model here. We will take the sales from different months, check the weather's effect on them, and predict the sales for next season. This should help with the pricing strategy.

3. A number of people rent over the weekend vs weekdays, as well as during the holiday season.
We are using a regression model to analyse the number of people renting over weekends and weekdays. On weekdays, people must be using to commute so we are running a regression.

4. A number of people rent over the weekend vs weekdays, as well as during the holiday season.
We are using a regression model to analyse the number of people renting over weekends and weekdays. On weekdays, people must be using to commute so we are running a regression.

Data Tidying

It's typically said that cleaning and preparing the data takes up 80% of data analysis time. And it's not just the first step; throughout the course of the research, it must be repeated numerous times when new issues are identified or fresh data is gathered. In order to understand the issue, this study concentrates on a minor but significant part of data cleaning that I refer to as data tidying: organizing datasets to make analysis easier.

A systematic approach to arrange data values inside a dataset is provided by the tidy data principles. Because you don't have to start from scratch each time and invent the wheel, a standard makes initial data cleansing simpler. The tidy data standard was created to make it easier to explore and analyse the data at first, as well as to make it simpler to create data analysis tools that function well together. The translation of modern tools is frequent. The output from one tool must be time-consumingly munged before being entered into another. Together, tidy datasets and tidy tools facilitate data analysis, allowing you to concentrate on the intriguing domain problem rather than the uninteresting logistics of the data.

While all tidy datasets are similar, each messy dataset is messy in its own unique way. A defined method of connecting a dataset's physical form, or its structure, with its semantics is provided by tidy datasets (its meaning). In this part, I'll introduce some common terminology for characterizing a dataset's structure and semantics before defining tidy data using those terms.

Prepare/Process Data

To evaluate and spot trends, Cyclist's historical trip data (from April 2021 to March 2022) was provided (internal/first-party data provided as .csv files).

Each dataset is copied and the original files are stored in a separate directory in case the originals need to be referenced.

We have merged monthly files into a single .csv file and we have summarized the data to observe the number of rows and observation names.

De-identified User IDs, user types, bike types, and start and end information are all included in this data (times, positions, station names & IDs).

We have removed the columns with unnecessary variables like ride id, starting id, ending id and have omitted rows with NA data (No data) and have added required variables like ride length and season.

We checked for the unique values in all the columns and omitted any extra values not required for analysis. For ex: On checking unique values on start station name and end station name we found out that there was a “ ”(not named) station in end stations names, we decided to drop the blank value.

We have plotted different graphs to observe the variations in the number of ridings during weekends, weekdays and during different seasons.

We have chosen suitable models to predict the future trend.

Data transformation

After cleaning the data, we still required a lot of values to proceed with the analysis and modelling. On exploring the dataset well, we were able to derive different columns using the available data. We came across different functions such as difftime, distGeo, time2season to be able to get the desired values required for modelling.

Added ride length column using difftime function using time stamp available in data.
Provided started at and ended at timestamp to the difftime function to calculate ride length.
bicycle_data$ride_length <- difftime(bicycle_data$ended_at, bicycle_data$started_at)

Provided the started at time to POSIX function to change the format and used as.Date function to get the date out of timestamp.
date <- as.POSIXct(bicycle_data$started_at, format = "%Y-%m-%d %H:%M:%S")
datemonth <- as.Date(date, format="%Y-%m-%d")

Provided the start_lat, start_lang, end_lat,end_lng values as two separate matrix and used distGeo function which provides us with distance travelled in miles between the given coordinates.
bicycle_data$distance<-distGeo(matrix(c(bicycle_data$start_lng, bicycle_data$start_lat), ncol=2), matrix (c(bicycle_data$end_lng, bicycle_data$end_lat), ncol=2))

Added a column to check whether the day is weekend or weekday using date. We have used an if else statement to classify the date.
bicycle_data$week <- ifelse(weekdays(date) %in% c("Saturday", "Sunday"), "weekend", "weekday")

Using time2season(date) created a season column. Passed the date value derived by converting into POSIX format and extracting just the date value in ***"%Y-%m-%d"*** format.
bicycle_data$season <- time2season(datemonth, out.fmt = "seasons", type="default")

Found the count of ridership at each station to predict the future ridership count.

New Dataset

• Rideable_type • Started_at • Ended_at • Start_station_name • End_station_name • Start_lat • Start_lng	• End_lat • End_lng • Member_casual • Ride length • Distance • Season • Month • Day_of_week

Data Visualization

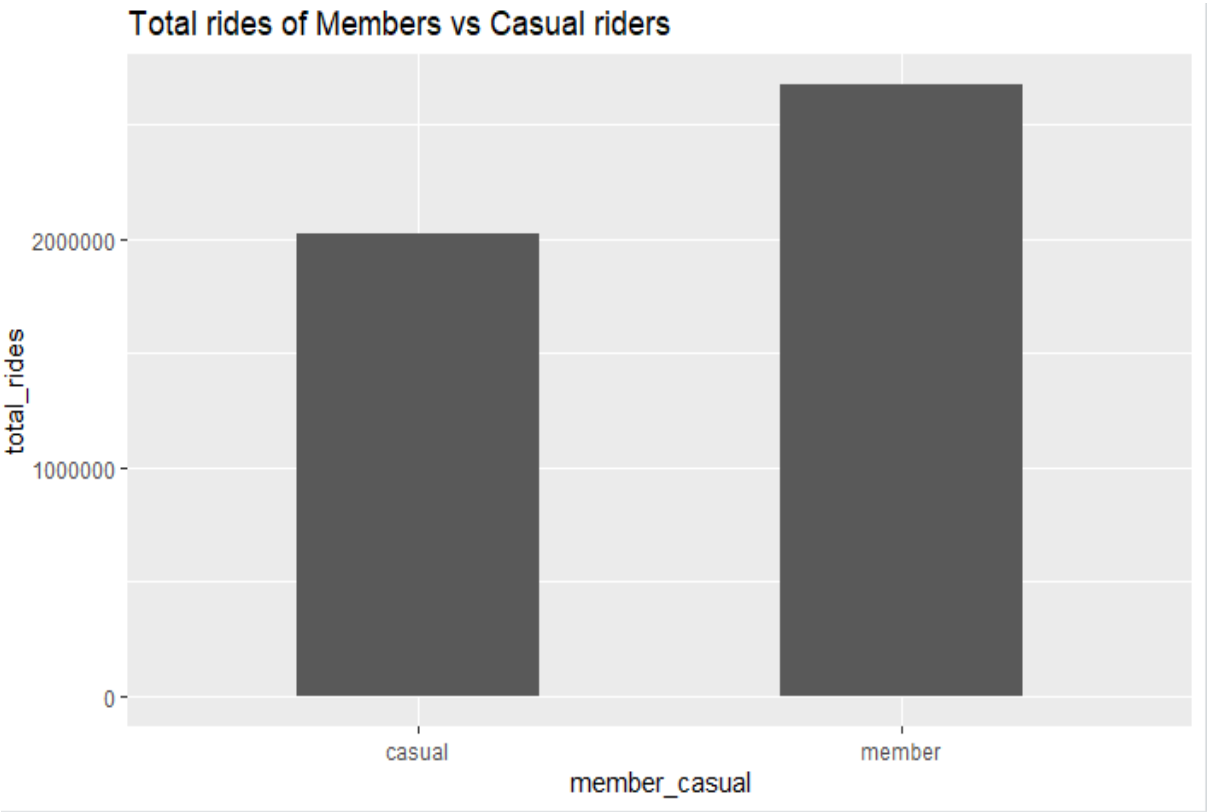
The graphic depiction of data and information is known as data visualization. Data visualization tools offer an easy approach to observe and comprehend trends, outliers, and patterns in data by employing visual elements like charts, graphs, and maps. It also gives staff members or business owners a great way to deliver facts to non-technical audiences without confusing them.

To analyse vast volumes of data and make data-driven decisions in the world of big data, tools and technology for data visualization are crucial.

1. Total Rides of Members vs Casual Riders

Considering the FY 2020-21, we had the data of 5.7M rides. On data tidying we were able to reduce the data to 4.7M by removing the NA values and station with no names.

There are two types of riders: Members and Casual Users.
Among the total rides 4,701,587, there are 2,026,997 rides taken casual users and 2,674,590 members.

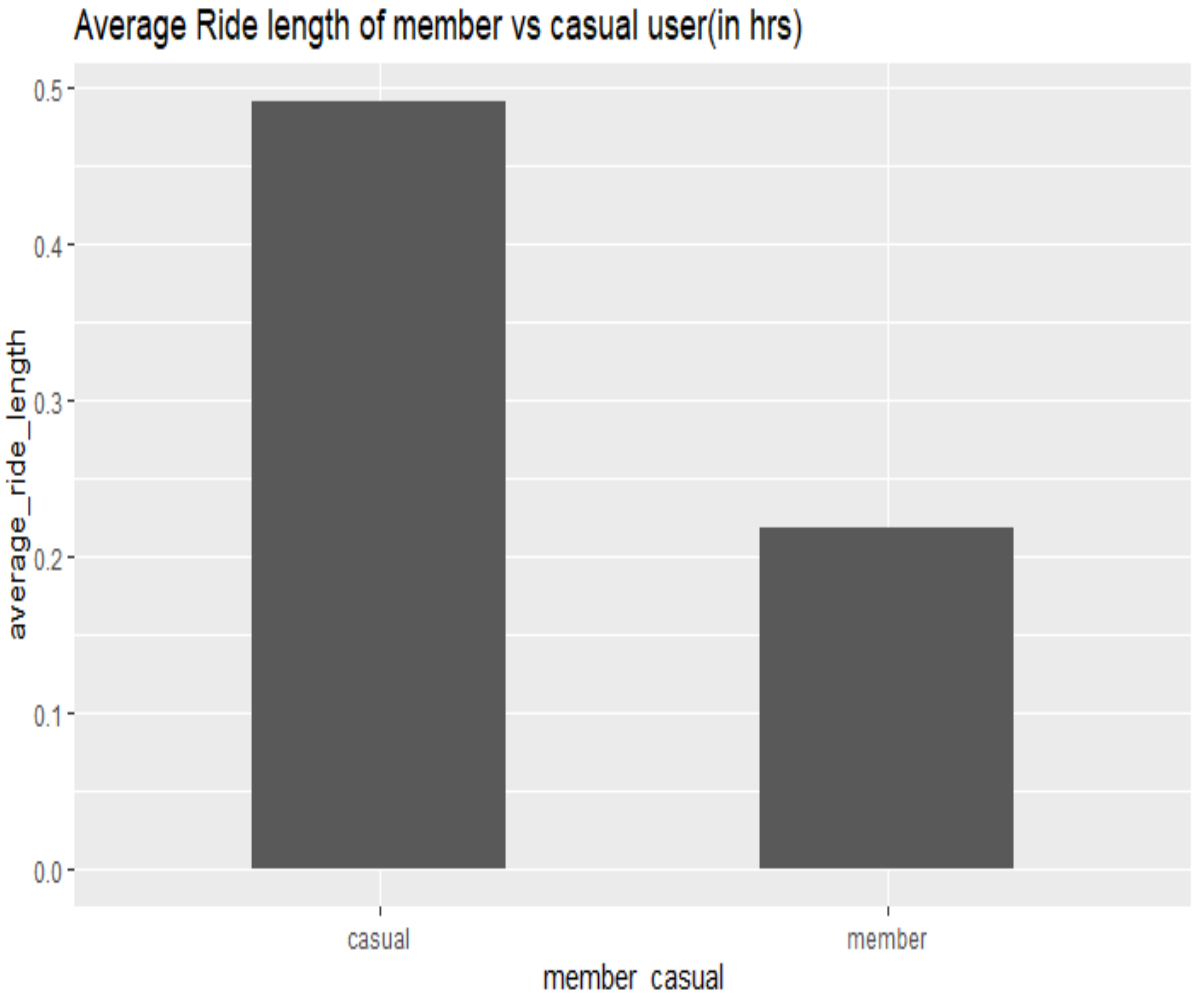


2. Average Ride Length of Member vs Casual user

From the data we derived a new column ride length using the difftime function we gave us difference in time stamps between the start time and end time of the ride.
The average ride time, median, max and min ride time was computed for the complete dataset.

On comparing average ride length by grouping member_casual column, we see that even though the members using the bikes is higher still the average ride length of casual users is more than double.
Through this data we can interpret that casual riders use bikes for longer durations which would mean that bikes are bring used for exploring the city.

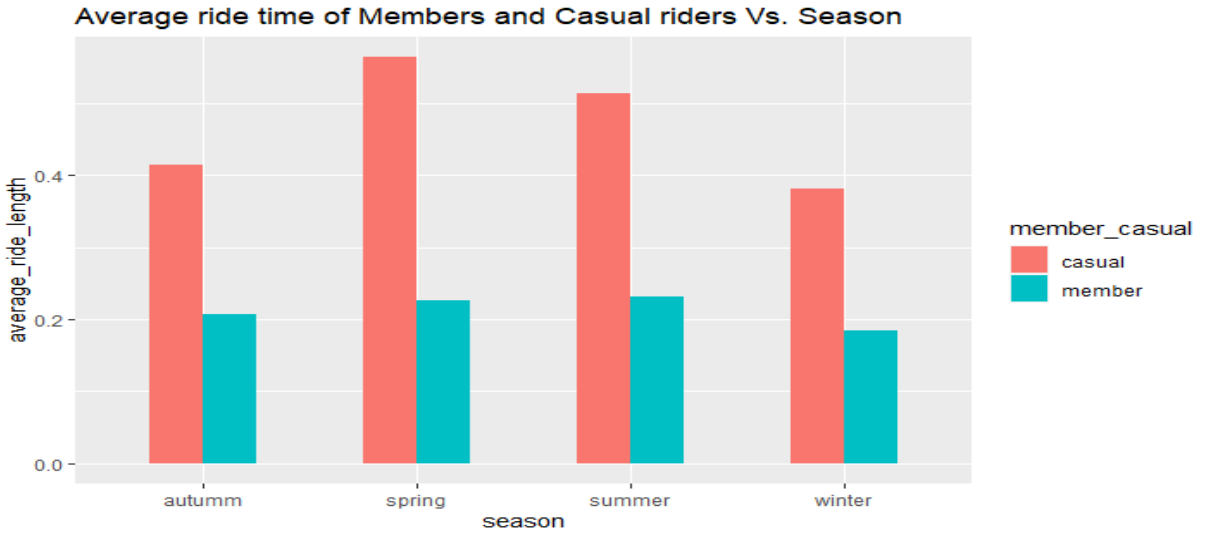
Members usage duration is much lower than casual users. This trend shows that members utilize the bikes for commuting between their homes to office/university. Casual users form a larger chunk of the company's revenue.



3. Total Rides of Members and Casual Riders Vs Season

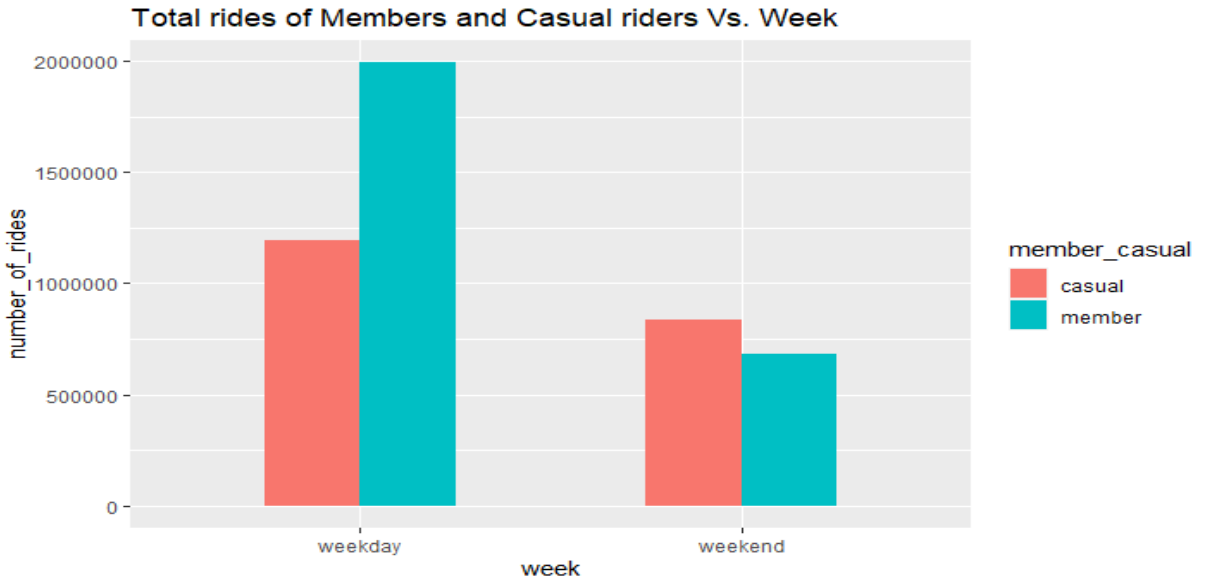
On grouping the data by seasons, we can visualise the seasonal trend of ridership among the members and casual users. It is a clear win that the duration of usage is highest by casual users. If we look at the seasonal usage, we see that the ridership during winters among members is 3 times than casual users whereas in summers casual users have an upper hand compared to members. The overall result shows that other than summers, during other seasons casual users tend to use the services lesser.

Converting more casual users to Members will help increase the revenue during winters.



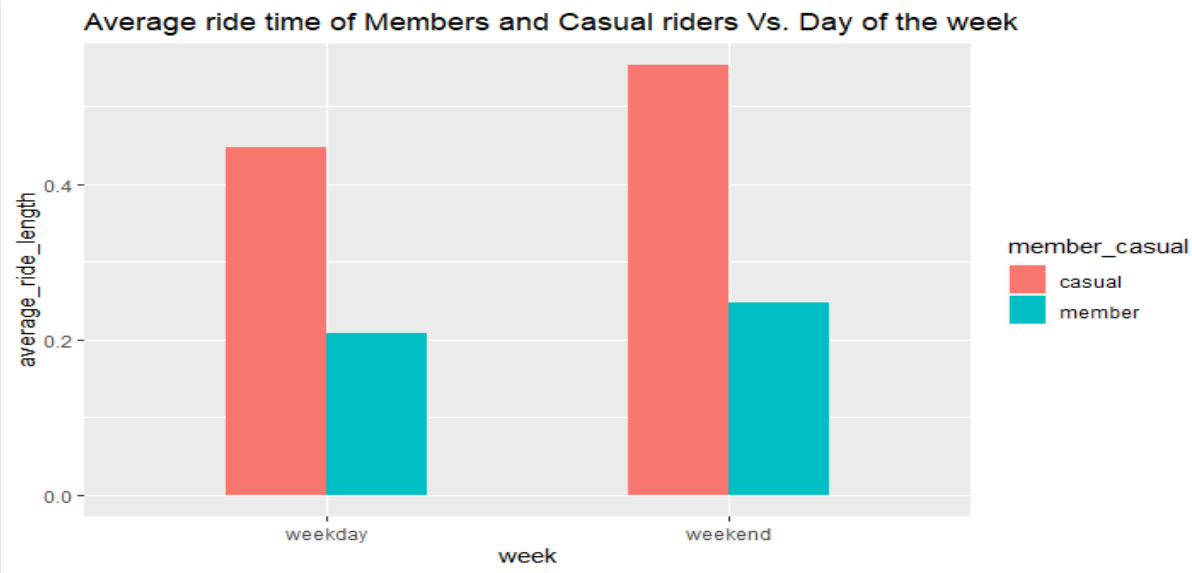
4. Total Rides of Members and Casual Riders Vs Day of Week

The total number of rides taken by casual users is higher than that of members during the weekends.



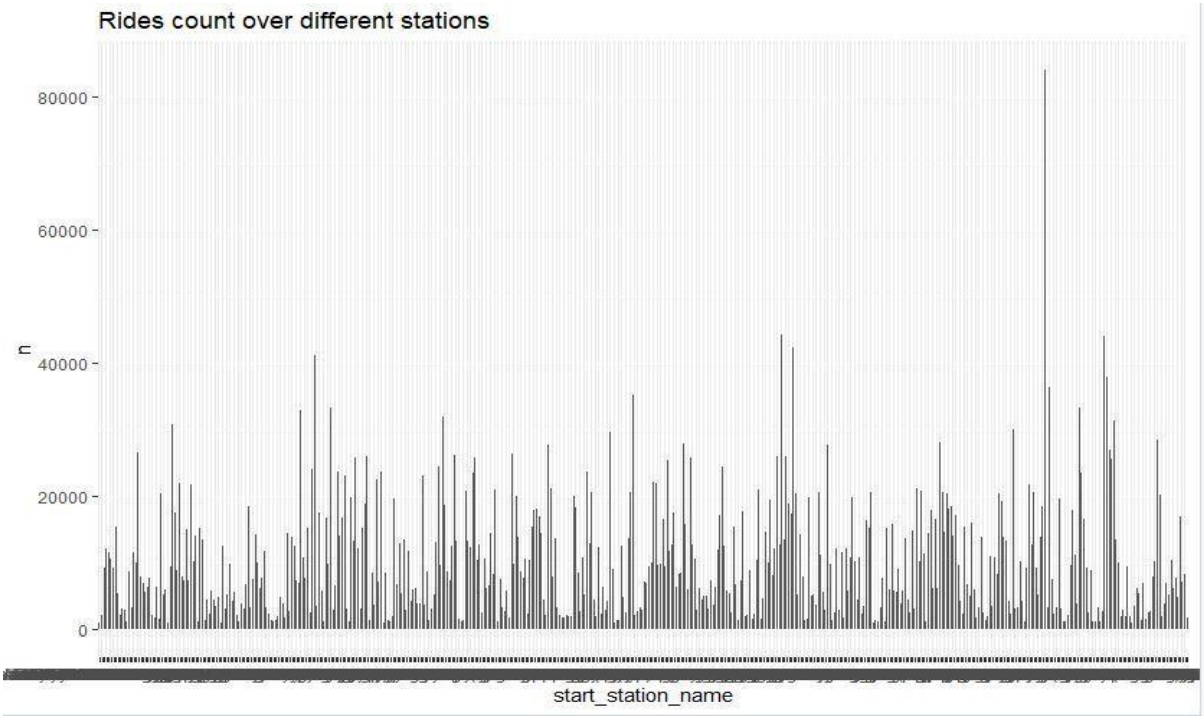
5. Average Ride Time of Members and Casual Riders Vs Day of the week

Average ride time has always been higher among the casual users irrespective whether it is a weekday or weekend.
The number of rides over the weekday's averages about 650,000*5 rides for weekdays and average of 750,000*2 rides.



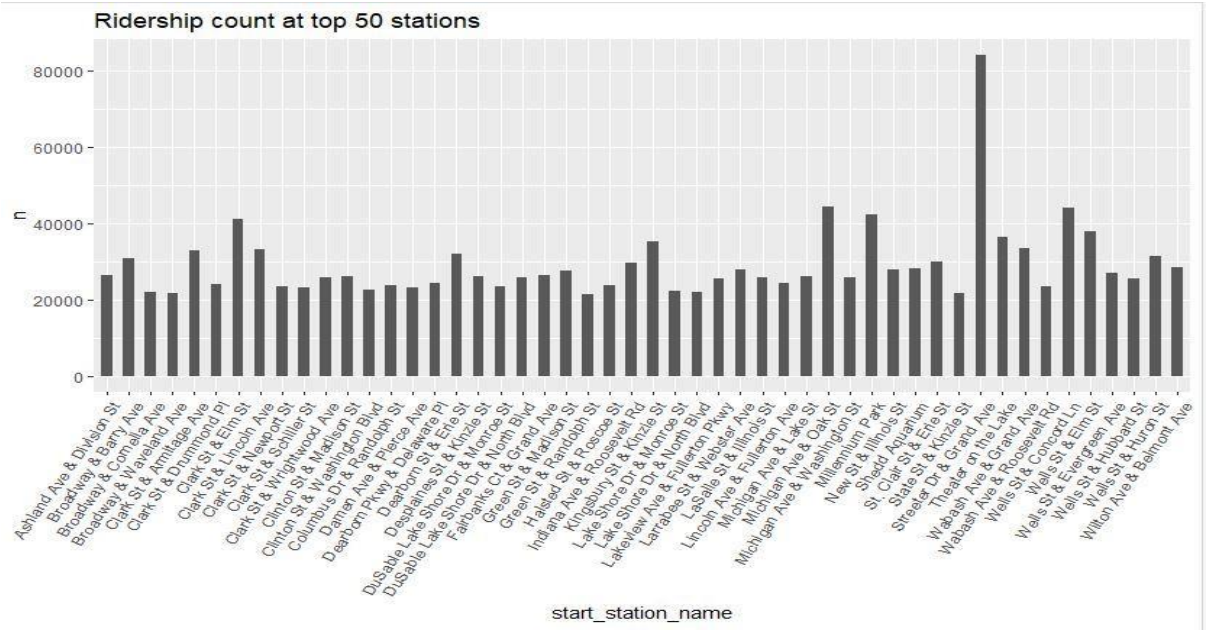
6. Count of Riders at Different Stations

We analysed the number of rides being taken at different stations. There are more than 80,000 rides at Grand Avenue station in Chicago compared to average number of rides among top 50 locations which is about 20,000 rides at each.



7. Count of Riders at Top 50 Stations

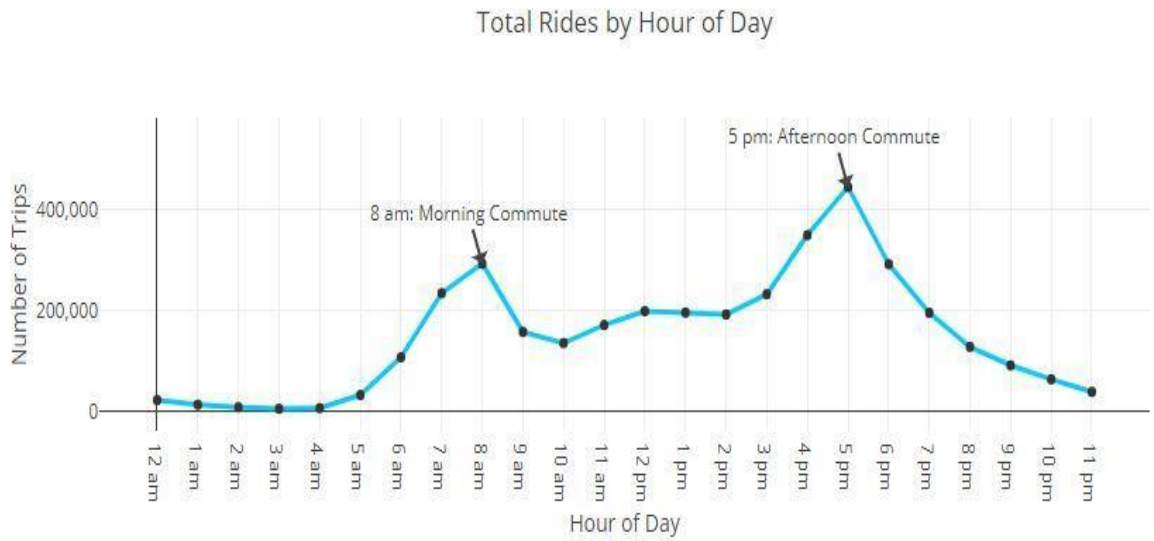
We analysed the number of rides being taken at different stations. There are more than 80,000 rides at Grand Avenue station in Chicago compared to average number of rides among top 50 locations which is about 20,000 rides at each.



8. Total Rides by the Hour of Day

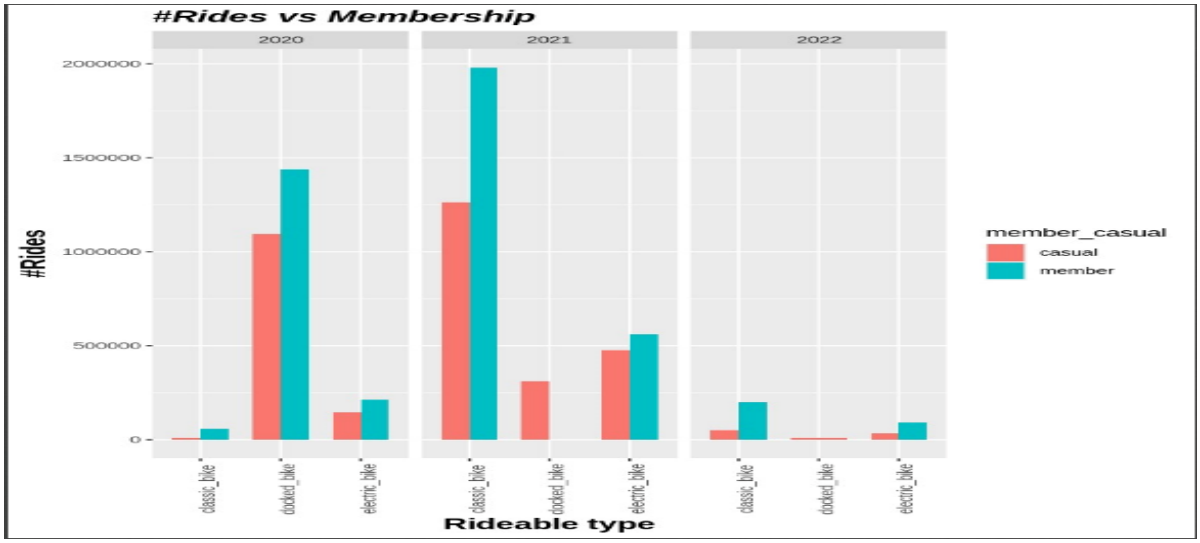
We see that the ridership count peaks during 8AM in the mornings and 5PM in the evening. This trend shows that most of the rides are taken by Professionals/Students who use these bikes to commute.

The stats show that the number of rides taken between 9PM to 6AM have been very low.



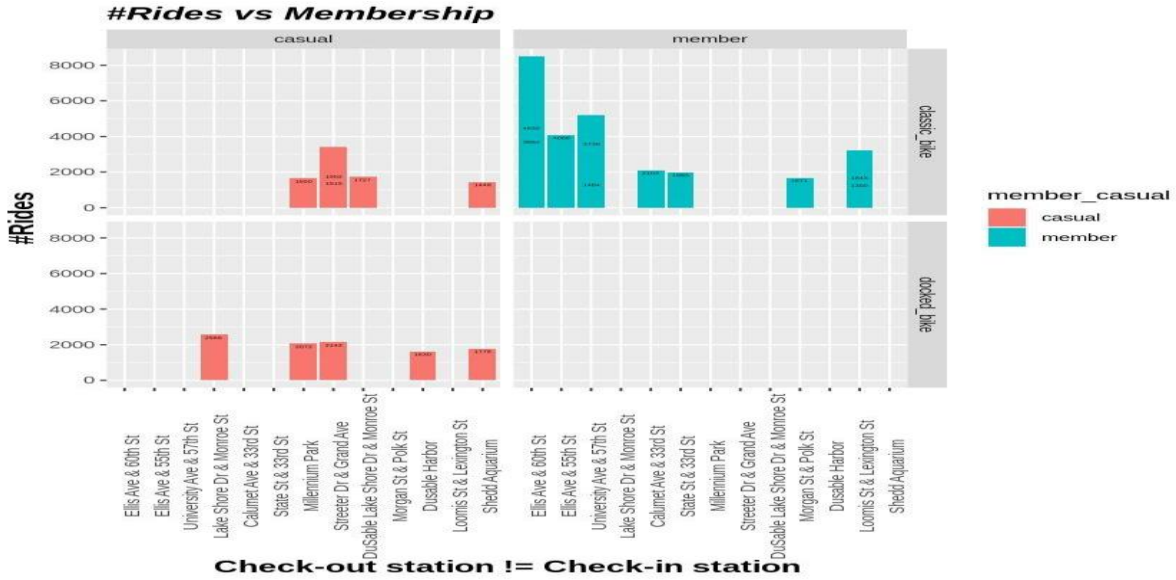
9. Total Rides vs Membership vs Rideable type

In this Visualisation, we see the utilization of different bikes based on member category over the period of April 2020 to July 2022. We see that docked bikes have a lot of riders during FY2020-21 but on the addition of lot of classic bikes by the company, more riders got accustomed to use classic bikes over docked bikes.



10. Total Rides vs Membership vs Rideable type

Here we compare the number of people whose start station is not same as the end station. We see the utilization of docked bike vs classic bike based on member category.



Data Modelling

1. To predict which type of bike will be used at a particular station

Multinomial logistic regression is a classification technique for statistical problems with more than two possible discrete outcomes, which generalizes logistic regression to multiclass problems. Accordingly, it is a model that is used to predict the probabilities of the various outcomes of a categorically distributed dependent variable given a set of independent variables.

Advantages:

– Helps to understand the relationships among the variables present in the dataset.

– Simultaneous Models result in smaller standard errors for the parameter estimates than when fitting the logistic regression models separately.

– The choice of reference class has no effect on the parameter estimates for other categories.

Using Multinomial Logistic regression, we predict which type of bike will be used at a particular station. Station name was given as one of the inputs to predict the rideable type.

This model predicted the ride type based on the station with an accuracy of 72.8%.

X <- station_name

Y <- rideable_type

```
> # Calculating accuracy - sum of diagonal elements divided by total obs
> round((sum(diag(tab))/sum(tab))*100,2)
[1] 72.8
> tab
```

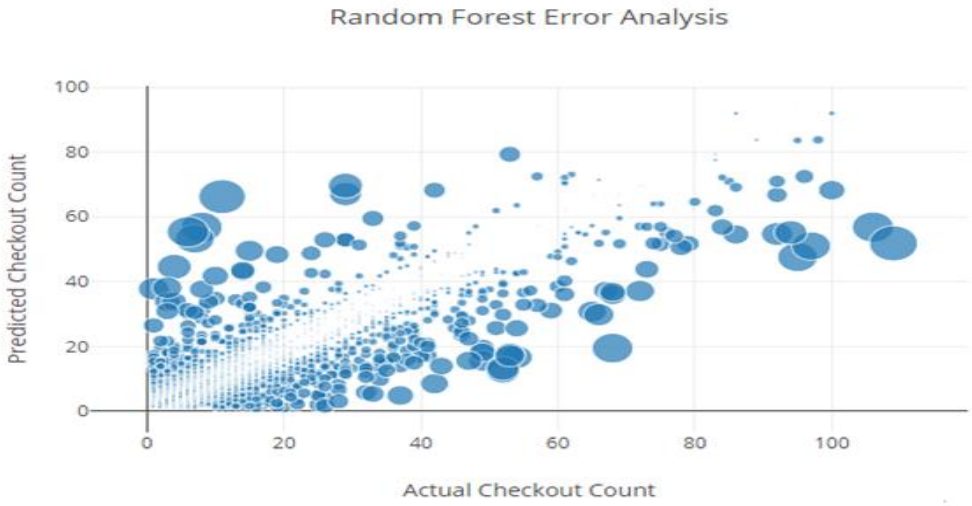
	classic_bike	docked_bike	electric_bike
classic_bike	2113	0	17
docked_bike	302	1	5
electric_bike	491	1	70

2. To predict the checkout count based on season, weekdays, rideable type, member

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. The steps for the Random Forest algorithm are as follows:

- Choose k data points at random from the practice set.
- Create a decision tree with these k data points in it.
- Repeat steps 1 and 2 for the number N of trees you want to construct.

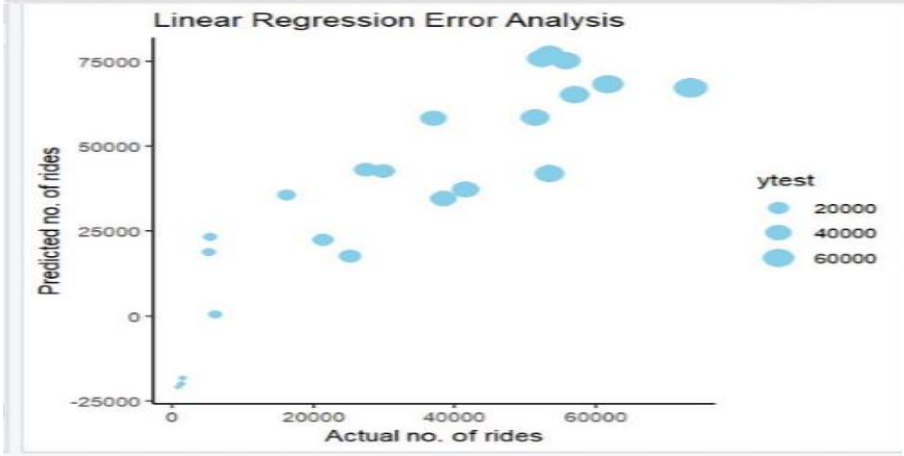
Using Random Forest Regression, we predicted the checkout count based on all other factors. We got an accuracy of 72.04%. Which explains that the model predicted the checkout count correctly 72.04% of the times.




```
> Mean_Squared_Error
[1] 39.35216
> r2
[1] 0.7204987
> Mean_Absolute_Error
[1] 3.732972
> Median_Absolute_Error
[1] 2.076043
```

We tried **Linear regression** for the same research question.

```
> # Calculating accuracy - sum of diagonal elements divided by total obs
> round((sum(diag(tab))/sum(tab))*100,2)
[1] 13.64
```

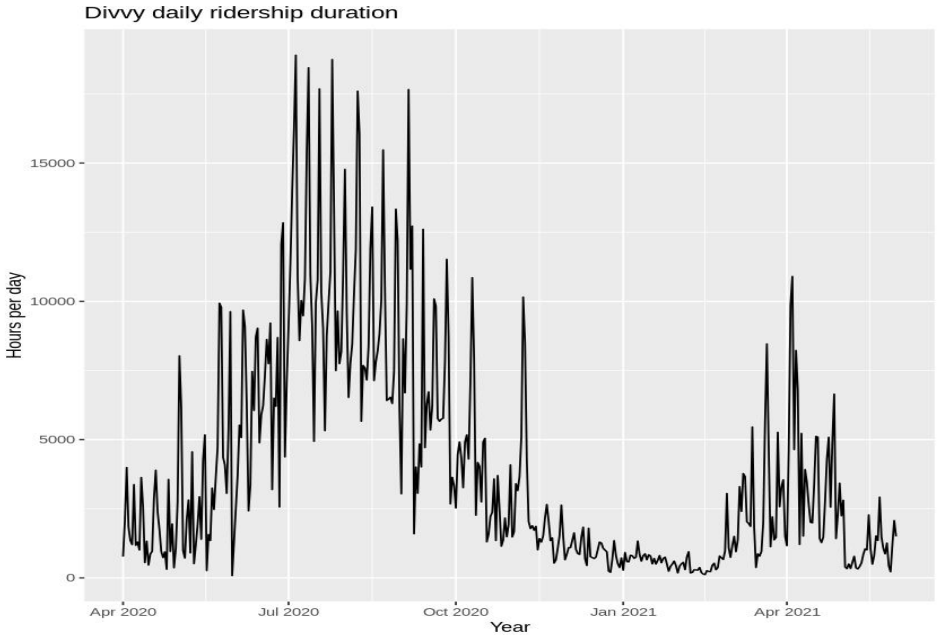


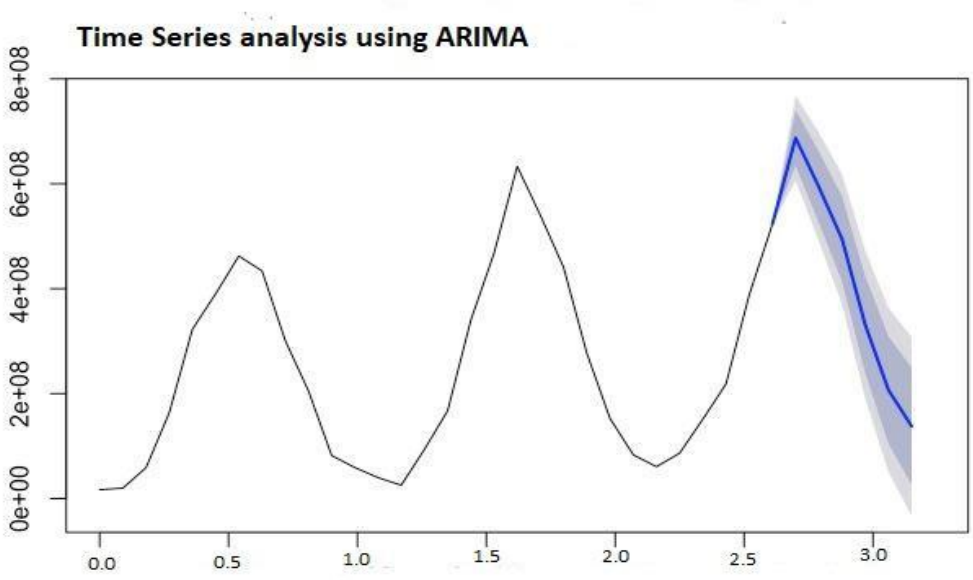
After comparing Linear Regression and Random forest's R square vale, we found that Random Forest has much better performance for this prediction.

3. To understand the effect of weather on the ride rental

ARIMA is a method for forecasting or predicting future outcomes based on a historical time series. It is based on the statistical concept of serial correlation, where past data points influence future data points.

ARIMA models can take into account trends, cycles, seasonality, and other non-static types of data when making forecasts. ARIMA forecasting is achieved by plugging in time series data for the variable of interest. Statistical software will identify the appropriate number of lags or amount of differencing to be applied to the data and check for stationarity. It will then output the results, which are often interpreted similarly to that of a multiple linear regression model.





Dashboard

We have created an interactive dashboard using the Shiny package available in R to display the visualisations we have.

Conclusion

- Members usage duration is much lower than casual users. Casual users form a larger chunk of the company's revenue.
- During the weekends casual riders use the bike sharing service the most.
- The casual riders have a higher ride length when compared to members which shows that members are usually using the services to commute between particular locations every day.
- We observed that some locations have less than 1000 rides throughout the year so those stations could be closed.
- The Grand Ave station has the highest foot fall and there are about 360 stations with more than average number of users.
- The other stations where the footfall is less than 5000 rides per year, we can add attractive discount for people travelling from less ridership locations to higher ridership locations so that there is availability of bikes at those stations.
- During the office hours the ridership is very high, instead of having bikes in less used location we can move them to high ridership locations.
- Based on the Random Forest Model, we can predict the number of checkouts depending on the weather.
- Based on the Multinomial Logistic regression model, we can predict which bike is being used in which station.
- Based on Time series Analysis we realized that the rentals during winter drop, we need to think of membership plan for the summer months so that we could get in more members.