Analysis of Bikeshare Data in 2018: User Type, Trip Duration, and Usage Patterns

Muhammad Enrizky Brillian_1009713712

2023-04-04

1. Merging Datasets from multiple CSVs

We merge multiple CSV files from each quarter in 2018 into a compound CSV file that consists of the bikeshare data for the whole year in 2018.

2. Description of the Dataset

```
## [1] "trip_id"               "trip_duration_seconds" "from_station_id"
## [4] "trip_start_time"       "from_station_name"     "trip_stop_time"
## [7] "to_station_id"         "to_station_name"       "user_type"
```

1. Trip ID: A unique identifier for each trip
2. Trip Duration (Seconds): The length of the trip in seconds
3. From Station ID: The unique identifier for the starting station of the trip
4. Trip Start Time: The date and time the trip began
5. From Station Name: The name of the starting station for the trip
6. Trip Stop Time: The date and time the trip ended
7. To Station ID: The unique identifier for the ending station of the trip
8. Bike ID: The name of the ending station for the trip
9. User Type: A categorical variable indicating whether the user is a "Annual Member" (annual pass holder of the bikeshare program) or a "Casual Member"(24 or 72 hour pass holders of the bikeshare program)

3. The Background of the Data

The bikeshare data for 2018 was collected by the City of Toronto, specifically by the Transportation Services division. The data was collected in the context of a bikeshare program called Bike Share Toronto, which is a public bicycle sharing system in Toronto, Canada. Bike Share Toronto provides access to bicycles at various stations throughout the city, allowing users to rent a bike for short trips and return it to any station within the system. The data collected includes information about bike trips, such as trip duration, start and end station, and user type, and other variables.This data is publicly available through the City of Toronto's Open Data portal for research and analysis purposes.

The dataset can be used for analysis and insights into the usage patterns and trends of the bikeshare program in Toronto during the year 2018. It is typically used by researchers, policymakers, and urban planners to analyze bike usage patterns, evaluate the effectiveness of the bikeshare program, and inform transportation planning and policy decisions. Overall, the bikeshare data for 2018 is a valuable resource for investigating bike usage patterns, user behavior, and system performance in Toronto, and can contribute to evidence-based decision making in transportation planning and policy.

4. What is the overall research question?

As researchers, our overall research question is to investigate the usage patterns and trends of bikeshare data in Toronto for the year 2018. We aim to analyze various variables, such as trip duration, trip start time, user type, and station information, to gain insights into how the bikeshare system was utilized by different user groups, and to identify any patterns or trends that may emerge from the data. Specifically, our research question is:

- What are the summary statistics for trip duration by user type, and how do they compare?
- How does the total number of trips taken by each user type vary by month?
- What are the top 10 most popular starting stations for bike rides in the dataset?
- How many trips lasted 30 minutes or more, and how many lasted 45 minutes or more, for each user type?
- What is the trend of total trips by hour of day, and how does it relate to peak usage times?
- How does monthly trip volume vary by user type, and what is the overall trend over the year?

5. Tables

5.1 Trip Duration (in Seconds) Summary By User Type

Table 1: Summary of Trip Duration (in seconds) by User Type

user_type	Number_Trips	Average_duration	Median_duration	Min_duration	Max_duration
Annual Member	1572980	725.0167	600	60	55077
Casual Member	349975	2032.4958	1220	60	54971

From the table, we can see that the "Annual Member" user type has a significantly larger number of trips (1,572,980) than the "Casual Member" user type (349,975). However, the "Casual Member" user type has a higher average trip duration (2,032.4958 seconds) compared to the "Annual Member" user type (725.0167 seconds). The median duration for both user types is lower than the average, which indicates that there are some trips with much higher durations, skewing the average upwards.

5.2 Table of Total Trips by Month for each user type

Table 2: Total Trips by Month for each user type

Month	Annual Member	Casual Member
Jan	42469	1390
Feb	47276	2455
Mar	78564	6405
Apr	82194	12589
May	160989	51761
Jun	186463	64374
Jul	215835	70481
Aug	209770	71449
Sep	207789	47212
Oct	160326	15553
Nov	100675	3612
Dec	80630	2694

This table can provide insights into how usage of the bike share service varies by month and by user type. For example, it appears that usage by both annual and casual members increases during the warmer months, as the number of trips taken by both groups is much higher in the summer months (June through August) compared to the winter months (December through February). It also appears that annual members take many more trips overall than casual members.

For Annual Members, the month with the highest number of trips is July, with 215,835 trips. The month with the lowest number of trips is December, with 80,630 trips. For Casual Members, the month with the highest number of trips is May, with 51,761 trips. The month with the lowest number of trips is January, with 1,390 trips.

5.3 The Top 10 Most Popular Starting Stations

Table 3: Top 10 Most Popular Starting stations

Starting Station	num_trips
York St / Queens Quay W	24017
Bay St / Queens Quay W (Ferry Terminal)	22743
Union Station	19869
Bay St / Wellesley St W	19184
Sherbourne St / Wellesley St E	19131
Front St W / Blue Jays Way	17282
Princess St / Adelaide St E	17089
Dundas St W / Yonge St	17054
Bay St / College St (East Side)	16965
Bathurst St/Queens Quay(Billy Bishop Airport)	16794

The table is sorted in descending order based on the number of trips, so the station with the highest number of trips is listed first. According to the table, the most popular starting station is "York St / Queens Quay W" with 24,017 trips, followed by "Bay St / Queens Quay W (Ferry Terminal)" with 22,743 trips, and "Union Station" with 19,869 trips.

This table can be useful for bike-sharing companies or city officials to identify the most popular locations for bike rides and allocate their resources accordingly, such as providing more bikes or improving bike infrastructure in those areas.

5.4 Trip duration More than 30 minutes and More than 45 minutes

Table 4: Trip Duration more than 30 minutes and 45 minutes.

user_type	>= 30 mins	>= 45 mins
Annual Member	32647	10364
Casual Member	87792	51791

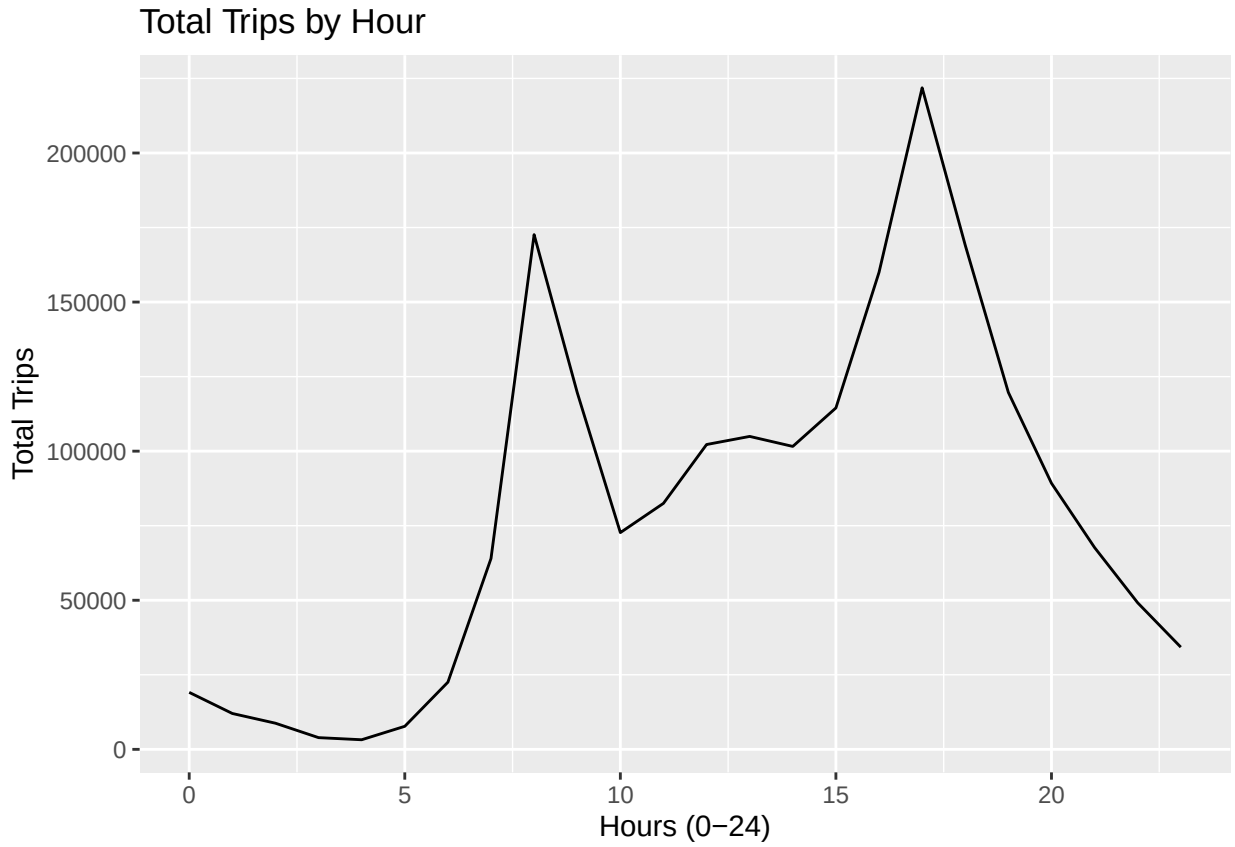
The table "Trip Duration more than 30 minutes and 45 minutes" shows the number of trips taken by each user type (Annual Member and Casual Member) that lasted 30 minutes or more and 45 minutes or more, respectively.

For Annual Members, there were 32,647 trips that lasted 30 minutes or more, and 10,364 trips that lasted 45 minutes or more. For Casual Members, there were 87,792 trips that lasted 30 minutes or more, and 51,791 trips that lasted 45 minutes or more.

This information could be useful for analyzing the usage patterns of different user types and identifying potential areas for improvement in the bike sharing system. For example, if a significant number of casual users are taking longer trips, the bike sharing company could consider offering different pricing plans or incentives to encourage shorter trips and more frequent use. Additionally, the data could help inform decisions about bike station placement and bike fleet size based on usage patterns.

6. Graphs

6.1 Line Chart of Total Trips by Hour

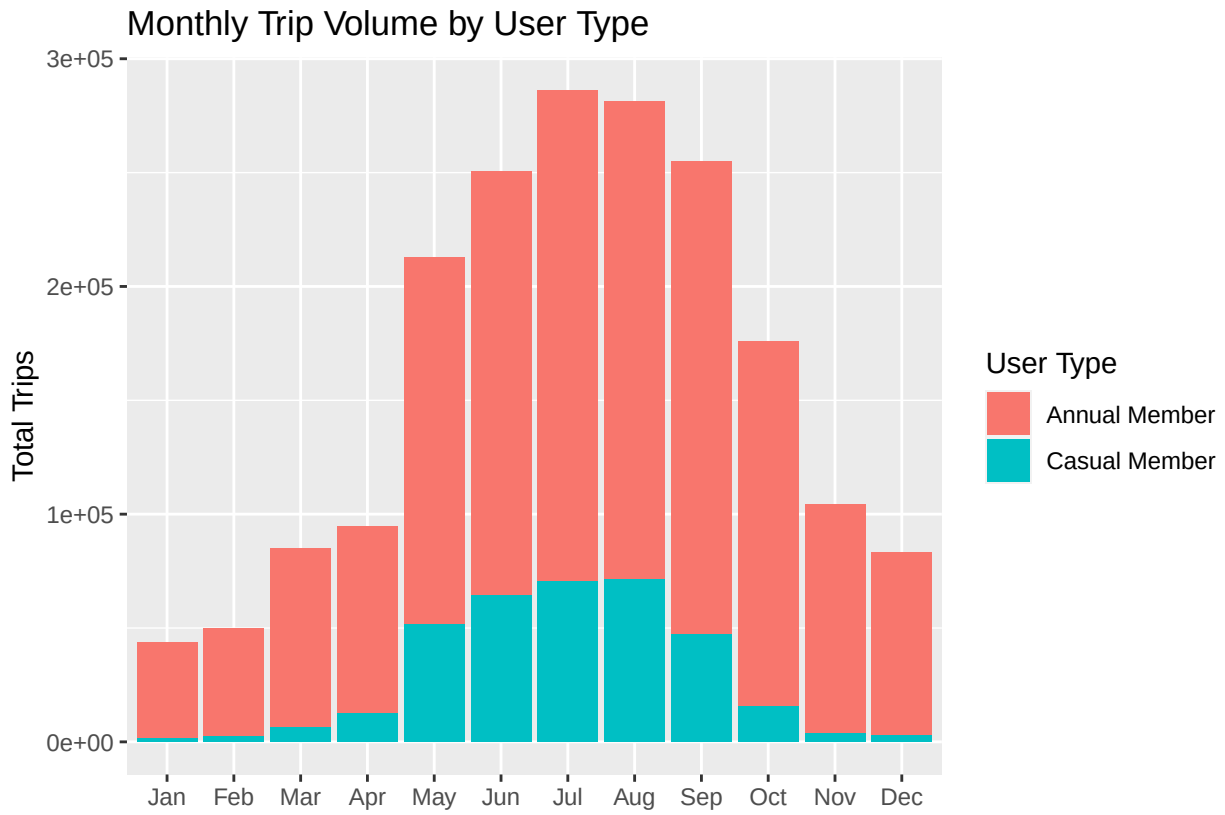


The Line Chart of Total Trips by Hour displays the total number of trips taken by hour of day for the bikeshare data in 2018. The x-axis represents the hour of the day in a 24-hour format, while the y-axis represents the total number of trips taken during that hour.

The line shows the trend of the total number of trips throughout the day. It appears that the number of trips starts to increase from around 5 AM, peaks around 8 AM, drops gradually until noon, picks up again in the afternoon until around 5 PM, and then starts to decrease steadily until midnight.

This pattern indicates that the bikeshare service is mostly used for commuting to work or school during the weekdays, with higher demand during rush hours in the morning and evening. The chart could be useful to help the bikeshare company plan for bike availability and schedule maintenance based on peak usage times.

6.2 Stacked bar plot of trip volume by user type and month



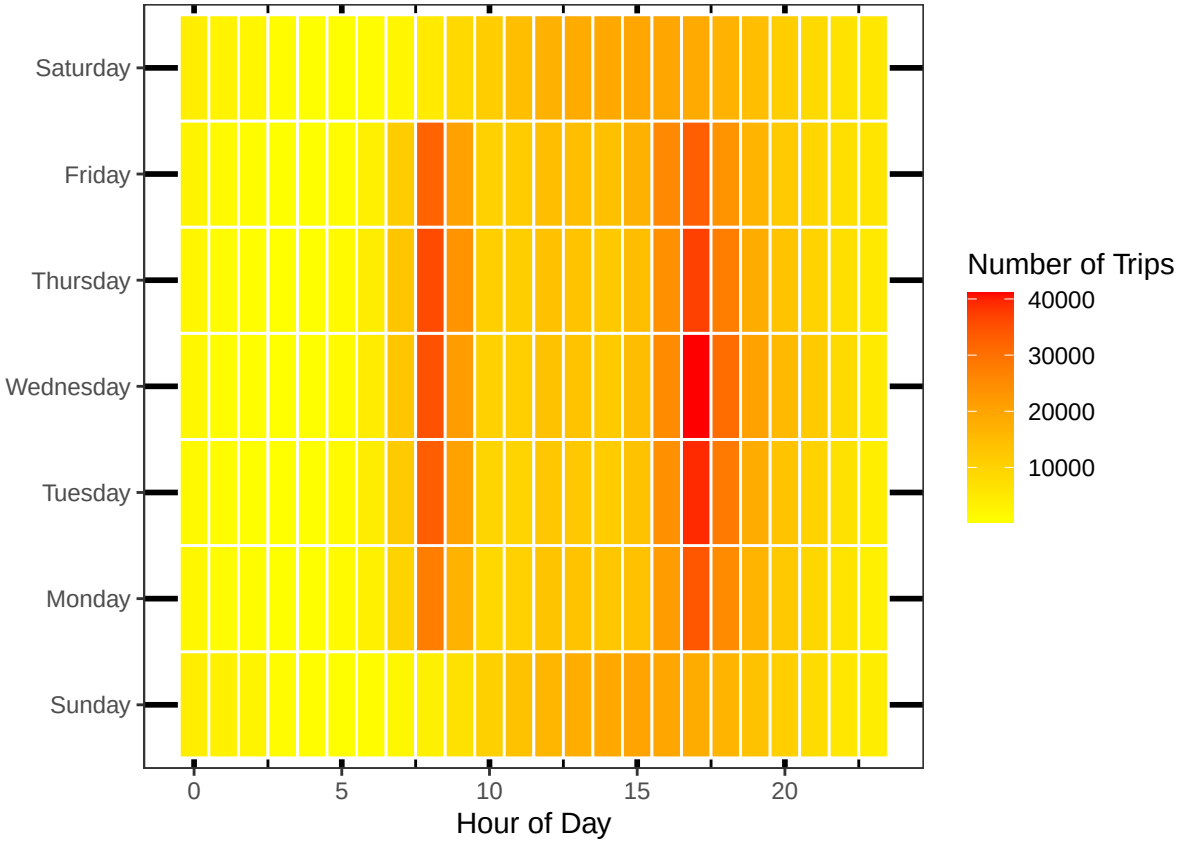
The stacked bar plot shows the trend of monthly trip volume by user type for the Bikeshare data in 2018. The X-axis represents the months, and the Y-axis represents the total number of trips. The bars are stacked based on the user type, with casual users in red and annual members in blue.

From this plot, we can see that the number of trips taken by Annual Members is consistently higher than the number of trips taken by Casual Members throughout the year. We can also observe that the number of trips taken by both user types peaks in the summer months (June to August) and declines during the winter months (December to February). This suggests that the bikeshare program is more popular during the warmer months of the year.

We can also see that the number of trips generally increased from January to July, with the highest number of trips in July, followed by a gradual decrease in the number of trips from August to December. The plot also shows a dip in the number of trips in November and December, which may be due to seasonal factors such as colder weather.

Overall, this plot provides an overview of the distribution of trips by user type over the months, which can help stakeholders make informed decisions about their marketing strategies and resource allocation.

6.3 Heatmap of trips by day of the week



This heatmap represents the trend of the number of bike trips taken on each day of the week and at each hour of the day. The x-axis of the heatmap represents the hour of the day, while the y-axis represents the day of the week. The cells of the heatmap are colored according to the number of bike trips that occurred at each hour of the day on each day of the week. The color scale ranges from yellow (indicating low trip volumes) to red (indicating high trip volumes).

From the Heatmap, we can observe that the bike usage is highest during the weekdays, particularly on Tuesday, Wednesday, and Thursday. During these weekdays, the highest usage hours are the morning commute hours from 7-9 am and evening hours from 4-6 pm. On weekends, the usage of bikes is relatively lower than weekdays, and there is a slightly different pattern in bike usage. Bike usage is highest during the afternoon hours, particularly from 12 pm to 4 pm on both Saturday and Sunday.

Overall, this heatmap provides an insight into the bike usage pattern during different times of the week, which can help the bikeshare company to plan its resources and services accordingly.

7. Hypothesis Testing

Trip duration Between Trips taken on weekdays versus weekends.

We would like to determine if there is a significant difference in trip duration between trips taken on weekdays versus weekends.

Hypothesis:

- H_0: **The average trip duration on weekdays (μ_d) is equal to the average trip duration on weekend** (μ_e), i.e., $\mu_d = \mu_e$

- H_1: **The average trip duration on weekdays (μ_d) is not equal to the average trip duration on weekend** (μ_e), i.e., $\mu_d \neq \mu_e$

To test this hypothesis, we can use a two-sample t-test to compare the mean trip duration of weekday trips with the mean trip duration of weekend trips. We can split the data into two groups: weekday trips (Monday to Friday) and weekend trips (Saturday and Sunday).

We can then calculate the p-value associated with the t-statistic, which is the probability of observing a t-value as extreme or more extreme than the one calculated, assuming the null hypothesis is true. If the p-value is less than our significance level (typically 0.05), we can reject the null hypothesis and conclude that there is a significant difference in trip duration between weekday and weekend trips.

We can also calculate a confidence interval for the difference between the means to estimate the range of values in which the true difference between the means is likely to fall.

Overall, this hypothesis testing approach allows us to determine whether there is a statistically significant difference in trip duration between trips taken on weekdays versus weekends, and provides us with an estimate of the magnitude of this difference.

```
##
##  Welch Two Sample t-test
##
## data:  weekdays$trip_duration_seconds and weekends$trip_duration_seconds
## t = -112.86, df = 579131, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -383.8696 -370.7647
## sample estimates:
## mean of x mean of y
##  873.7795 1251.0967
```

The Welch Two Sample t-test shows a significant difference between the mean trip durations of weekdays and weekends. The p-value (2.2e-16) is less than the significance level of 0.05, indicating strong evidence against the null hypothesis of no difference. The confidence interval (-383.8696, -370.7647) also does not include 0, further supporting the conclusion that there is a significant difference between the two groups. Therefore, we can conclude that there is a significant difference in trip duration between trips taken on weekdays versus weekends. Specifically, the mean trip duration on weekends (1251.0967 seconds) is longer than that on weekdays (873.7795 seconds).

8. Bootstrapping

Estimating the average trip duration for casual members

We could use bootstrapping to estimate the sampling distribution of the mean and compute a confidence interval.

To perform bootstrapping, we can take repeated samples of the trip durations of casual members from the dataset with replacement, each time calculating the mean trip duration. We can then calculate the average of these means data to estimate the average trip duration for casual members. Finally, we also estimate confidence intervals from those means data.

```
## Mean trip duration for casual members: 2032.496
```

```
## 95% confidence interval: (  2022.838 - 2042.171  )
```

Based on the result, the mean trip duration for casual members is 2032.496 seconds. The 95% confidence interval is (2022.838 - 2042.171) seconds. This means that we are 95% confident that the true mean trip duration for casual members falls within this range.

9. Non-Linear Regression Analysis

Analyzing the relationship between the number of trips and hour of the day.

In this case, we would treat the hour of the day as a continuous variable and fit a non-linear regression model with the number of trips as the dependent variable and the hour of the day as the independent variable. The resulting model would give us information on the direction and strength of the relationship between the hour of the day and the number of trips.

```
##
## Call:
## lm(formula = num_trips ~ trip_start_hour + I(trip_start_hour^2) +
##     I(trip_start_hour^3), data = trips_by_hour)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38269 -21133  -9337   9561 103476
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          11345.41   25688.88   0.442  0.66348
## trip_start_hour      -5051.71    9882.68  -0.511  0.61483
## I(trip_start_hour^2)  2225.45    1011.04   2.201  0.03964 *
## I(trip_start_hour^3)   -86.35      28.87  -2.991  0.00721 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36500 on 20 degrees of freedom
## Multiple R-squared:  0.6909, Adjusted R-squared:  0.6445
## F-statistic:  14.9 on 3 and 20 DF,  p-value: 2.501e-05
```

Explanation

The non-linear regression model suggests that there is a significant relationship between the number of bike trips and the hour of the day, and that this relationship is best explained by a third-order polynomial function. **The coefficient of determination (R-squared) of 0.69 indicates that the model explains 69% of the variation in the data.** The p-values for the coefficients of the quadratic and cubic terms are both significant, suggesting that the non-linear terms are needed to better explain the relationship between the number of bike trips and the hour of the day. However, the p-value for the coefficient of the linear term is not significant, suggesting that there may not be a significant linear relationship between the number of bike trips and the hour of the day.

Intrepreting Regression Parameter

The regression parameters in this case refer to the coefficients of the independent variables in the multiple regression equation.

For this non-linear regression model, we have the following coefficients:

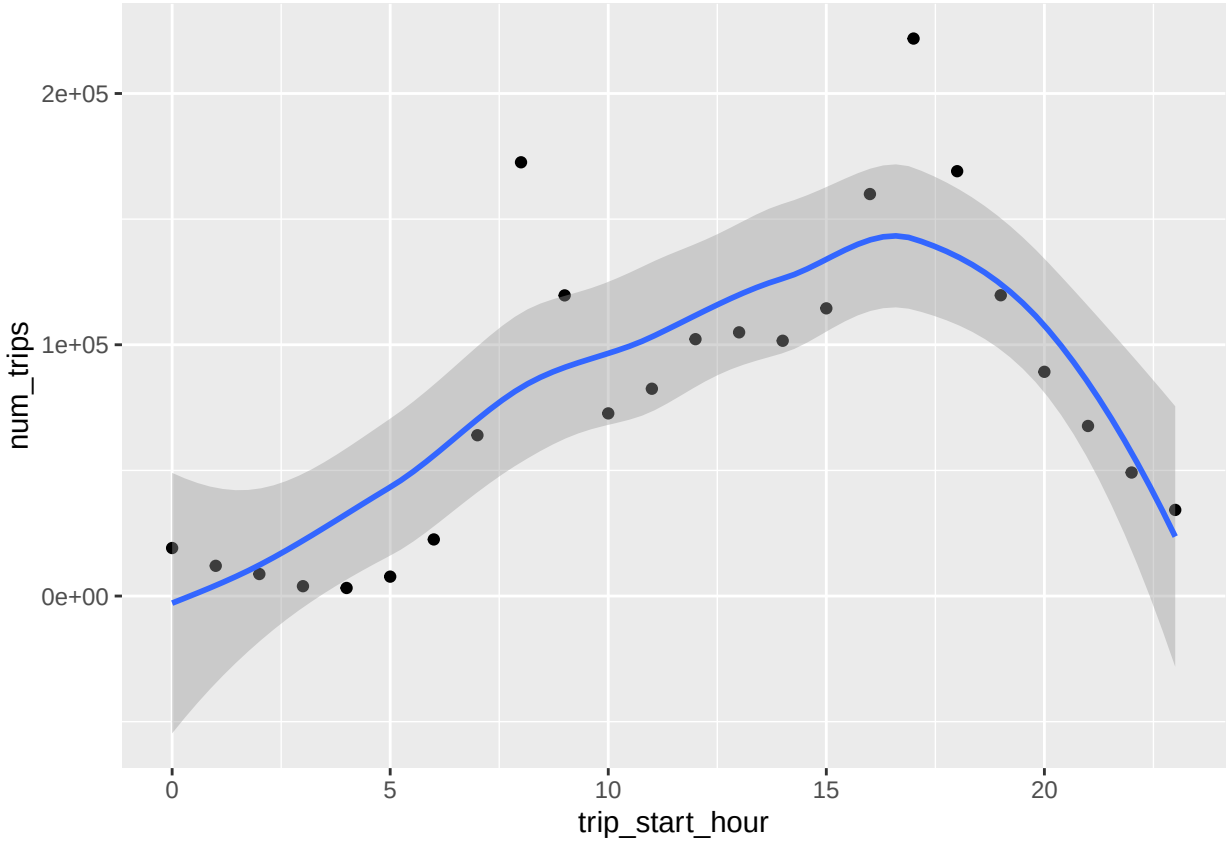
- The intercept coefficient represents the value of the dependent variable (num_trips) when all independent variables are equal to zero. In this case, the intercept coefficient is 11345.41, which means that when the trip_start_hour, trip_start_hour squared, and trip_start_hour cubed are all zero, the expected value of num_trips is 11345.41.
- The coefficient of trip_start_hour represents the linear effect of the start hour on the number of trips. In this case, the coefficient is negative (-5051.71), which suggests that as the start hour increases by one unit, the expected value of num_trips decreases by 5051.71.

- The coefficient of I(trip_start_hourˆ2) represents the quadratic effect of start hour on the number of trips. In this case, the coefficient is positive (2225.45), which suggests that the relationship between start hour and num_trips is curvilinear, with a maximum value of num_trips occurring at a specific start hour.
- The coefficient of I(trip_start_hourˆ3) represents the cubic effect of start hour on the number of trips. In this case, the coefficient is negative (-86.35), which suggests that the curvilinear relationship between start hour and num_trips is concave down, with the rate of decrease in num_trips slowing down as start hour increases.

Overall, this non-linear regression suggests that the relationship between start hour and the number of trips is not simply linear, but rather a curvilinear relationship with a peak at a specific start hour. Additionally, the cubic term indicates that this curvilinear relationship is concave down, meaning that the rate of decrease in the number of trips as start hour increases slows down as start hour increases.

Overall, this suggests that there is a non-linear relationship between the hour of the day and the number of trips taken, and this relationship is well captured by the polynomial regression model.

Plotting the Result



10. Cross Validation

Analyzing between Trip Duration (in hour) and hour of the day

In this case, we can use cross-validation to analyze the relationship between trip duration and hour of the day in the bikeshare data.

The idea behind this approach is to split the dataset into two parts, with one part used to train the model and the other part used to validate it. We can then measure the accuracy of the model on the validation set and adjust the model as needed to improve its performance.

To use cross-validation to analyze the relationship between trip duration (in hour) and hour of the day, we would start by splitting the dataset into a training set and a validation set. We could then use **Non-linear regression** to build a model that predicts trip duration based on the hour of the day, using the training set. Once we have built the model, we would use the validation set to evaluate its performance.

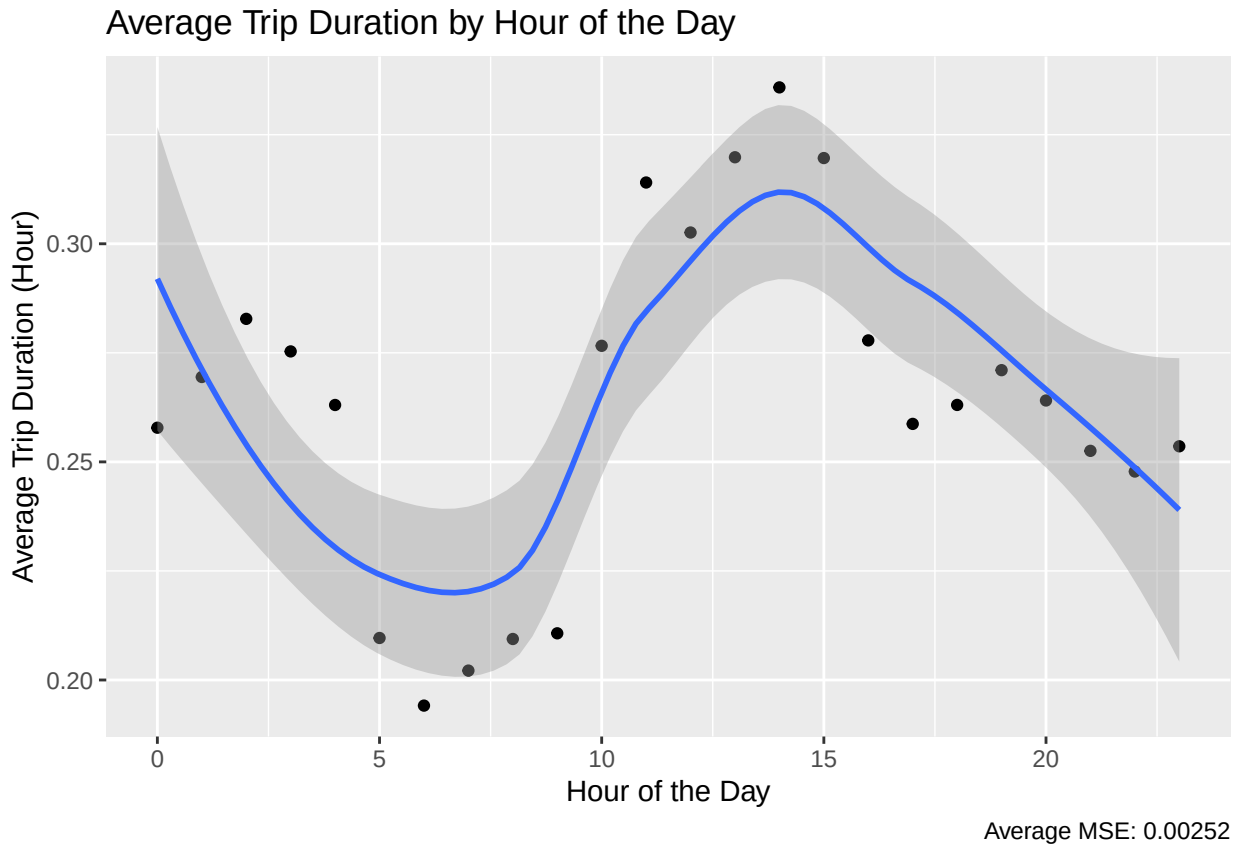
Our approach is k-fold cross-validation, where the dataset is divided into k equally-sized subsets. We then train the model on k-1 of the subsets and use the remaining subset for validation. We repeat this process k times, so that each subset is used for validation once.

```
## The Average MSE is: 0.002516312
```

Conclusion

The average MSE of 0.002516312 is quite small, which suggests that the non-linear regression model is fitting the data quite well. This means that there is likely a relationship between the hour of the day and the average trip duration (in hour), and that this relationship can be captured by the non-linear regression model.

Plotting the Result



11. Summary of Research

Based on the information presented in the report, the following are the key findings:

- The bikeshare service is mostly used by annual members, who take significantly more trips than casual members.
- The average trip duration for casual members is higher than that of annual members, but both groups have a similar median duration.
- Usage of the bikeshare service increases during the warmer months (June through August) compared to the winter months (December through February).
- The most popular starting station is "York St / Queens Quay W" with 24,017 trips.
- Annual members mostly use the bikeshare service for commuting to work or school, with higher demand during rush hours in the morning and evening.
- Casual members tend to take longer trips compared to annual members, with a significant number of trips lasting 30 minutes or more.
- The bikeshare company could consider offering different pricing plans or incentives to encourage shorter trips and more frequent use by casual members.

We then analysed trip duration when accounting for user type and this reveals that casual members travel nearly double the amount of time compared to annual members. Coming in at nearly 34 minutes on average compared to an annual member's 12 minutes. We thought that this was rather odd but upon investigation of the payment options given by bike share toronto we found out that a casual member could purchase a day pass for 90 minute rides compared to an annual member who could only purchase passes for up to 30 minutes and 45 minutes. We also revealed that annual members took nearly 4.5 times more trips compared to casual members, coming in at 1572980 trips versus 349975 trips.

Overall, the report provides insights into the usage patterns and behavior of users of the bikeshare service, which could help inform decisions about bike station placement, bike fleet size, pricing plans, and incentives to encourage more usage.

12. Appendix

```
# 1. Merging Datasets from multiple CSVs
# loading libraries
library(tidyverse)
library(dplyr)
library(lubridate)
library(knitr)
#Setting the working Directory
setwd("~/STAA57_Project")
# Create empty data frame
bikeshare_data <- data.frame()
# Loop through all CSV files in folder
for (file in list.files(pattern=".csv")) {
  temp_data <- read.csv(file, header=TRUE)  # Read in CSV file
  bikeshare_data <- rbind(bikeshare_data, temp_data)  # Bind data to the df
  rm(temp_data)  # Remove original data to free up memory}

# 2. Description of the Dataset
names(bikeshare_data)

# 3. The  Background of the Data: No Code in this Part

# 4. Overall Research Question: No Code in this Part

# 5 Tables
## 5.1 Trip Duration (in Seconds) Summary By User Type
```

```r
# Create summary statistics of trip duration by user type
trip_duration_summary <- bikeshare_data %>% group_by(user_type) %>%
  summarise(
    Number_Trips = n(), Average_duration = mean(trip_duration_seconds),
    Median_duration = median(trip_duration_seconds),
    Min_duration = min(trip_duration_seconds),
    Max_duration = max(trip_duration_seconds))

# Display the summary table
kable(trip_duration_summary,
  caption = "Summary of Trip Duration (in seconds) by User Type",
  align=c("l","c","c","c","c"))

## 5.2 Table of Total Trips by Month for each user type

# Convert trip_start_time to a datetime object
bikeshare_data$trip_start_time <- as.POSIXct(bikeshare_data$trip_start_time,
                                             format = "%m/%d/%Y %H:%M")

# Extract month from trip_start_time
bikeshare_data$trip_month <- month(bikeshare_data$trip_start_time, label = T)

# Create a table of trip count by month
trip_count_by_month <- bikeshare_data %>% drop_na() %>%
  count("Month"=trip_month, user_type) %>%
  pivot_wider(names_from = user_type, values_from = n)

# Print the table
kable(caption = "Total Trips by Month for each user type",trip_count_by_month)

## 5.3 The Top 10 Most Popular Startinh Stations

# Group the data by starting station and count the number of trips
starting_stations <- bikeshare_data %>%
  group_by("Starting Station"=from_station_name) %>%
  summarize(num_trips = n()) %>% arrange(desc(num_trips))

# Select the top 10 most popular starting stations
top_starting_stations <- head(starting_stations, 10)

# Create the table
kable(top_starting_stations, caption= "Top 10 Most Popular Starting stations")

## 5.4 Trip duration More than 30 minutes and More than 45 minutes

bikeshare_data %>%
  filter(trip_duration_seconds > 1800) %>% group_by(user_type) %>%
  summarise(">= 30 mins" = n()) %>% full_join(bikeshare_data %>%
      filter(trip_duration_seconds > 2700) %>% group_by(user_type) %>%
        summarise(">= 45 mins" = n()), by = "user_type") %>%
  kable(caption = "Trip Duration more than 30 minutes and 45 minutes.")

## 6.1 Line Chart of Total Trips by Hour
```

```r
# Extract day and hour from trip_start_time
bikeshare_data <- bikeshare_data %>%
  mutate(trip_start_hour = hour(trip_start_time),
         trip_day = wday(trip_start_time, label = T, abbr = F))

# summarize data by hour of day
hourly_trips <- bikeshare_data %>%
  group_by(trip_start_hour) %>%
  summarize(total_trips = n())

# create line chart of total trips by hour
ggplot(hourly_trips, aes(x = trip_start_hour, y = total_trips)) +
  geom_line() + ggtitle("Total Trips by Hour") +
  xlab("Hours (0-24)") + ylab("Total Trips")

## 6.2 Stacked bar plot of trip volume by user type and month

# group by month and user type and calculate the total trips
monthly_user_trips <- bikeshare_data %>% group_by(trip_month, user_type) %>%
  summarize(total_trips = n())

# create the stacked bar plot
ggplot(monthly_user_trips, aes(x=trip_month, y=total_trips, fill = user_type))
+ geom_bar(stat = "identity") +
  labs(title = "Monthly Trip Volume by User Type",
       x = "", y = "Total Trips", fill = "User Type")

## 6.3 Heatmap of trips by day of the week

# Aggregate the data by day of the week and hour
heatmap_data <- bikeshare_data %>% group_by(trip_start_hour, trip_day) %>%
  summarize(trip_count = n())

# Create the heatmap
ggplot(heatmap_data, aes(x=trip_start_hour, y=trip_day, fill = trip_count)) +
  geom_tile(color = "white", size = 0.5) +
  scale_fill_gradient(low = "yellow", high = "red") +
  labs(x = "Hour of Day", y = "", fill = "Number of Trips") + theme_bw() +
  theme(axis.text.x = element_text(angle = 0, hjust = 0.5),
        panel.grid.major = element_line(color = "black", size = 1),
        panel.grid.minor = element_line(color = "black", size = 0.5))

# 7. Hypothesis Testing
## Trip duration Between Trips taken on weekdays versus weekends.

# Subset data to separate trips taken on weekdays and weekends
weekdays <- subset(bikeshare_data, !trip_day %in% c("Saturday", "Sunday"))
weekends <- subset(bikeshare_data, trip_day %in% c("Saturday", "Sunday"))
# Conduct two-sample t-test
t.test(x=weekdays$trip_duration_seconds, y=weekends$trip_duration_seconds,
       var.equal = FALSE)

# 8. Bootstrapping
```

```r
## Estimating the average trip duration for casual members

# Define a function to calculate the average trip duration
mean_duration <- function(data) {mean(data$trip_duration_seconds)}

# Subset data for casual members
casual_data <- bikeshare_data %>% filter(user_type == "Casual Member")

# Use bootstrapping to estimate the mean trip duration
set.seed(123)
n_boot <- 100
bootstrap_means <- replicate(n_boot, casual_data %>%
                             sample_n(nrow(casual_data), replace=TRUE) %>%
                             mean_duration())

# Calculate the 95% confidence interval
ci <- quantile(bootstrap_means, c(0.025, 0.975))

# Print the results
cat("Mean trip duration for casual members:", mean(casual_data$trip_duration_seconds), "\n")
cat("95% confidence interval:","( ", ci[1], "-", ci[2]," )", "\n")

# 9. Non-Linear Regression Analysis
## Analyzing the relationship between the number of trips and hour of the day.

# Group the data by hour and count the number of trips for each hour
trips_by_hour <- bikeshare_data %>% group_by(trip_start_hour) %>%
  summarize(num_trips = n())

# Use Poly regression to analyze between the total trips and hour of the day
model <- lm(num_trips ~ trip_start_hour + I(trip_start_hour^2)
            + I(trip_start_hour^3), data = trips_by_hour)

# Print the summary of the model
summary(model)

# Print the summary of the model
summary(model)
# Plotting The Result
ggplot(trips_by_hour, aes(x=trip_start_hour, y=num_trips)) +
  geom_point() + stat_smooth()

# 10. Cross Validation
## Analyzing between Trip Duration (in hour) and hour of the day
# Calculate average trip duration (in hour) by hour of the day
avg_duration <- bikeshare_data %>% group_by(trip_start_hour) %>%
  summarize(avg_trip_duration = mean(trip_duration_seconds/3600))

# Define the model formula
formula <- avg_trip_duration ~ poly(trip_start_hour,degree = 3, raw = T)

# Define the number of folds for cross-validation
k <- 10
```

```r
# Initialize a vector to store the mean squared errors for each fold
cv_mse <- rep(0, k)

# Define the indices for each fold
set.seed(123)
folds <- cut(seq(1, nrow(avg_duration)), breaks = k, labels = FALSE)

# Perform k-fold cross-validation
for (i in 1:k) {
  # Split the data into training and test sets for this fold
  test_indices <- which(folds == i, arr.ind = TRUE)
  test_set <- avg_duration[test_indices, ]
  train_set <- avg_duration[-test_indices, ]

  # Fit the model to the training set
  model <- lm(formula, data = train_set)

  # Use the model to predict the test set
  predictions <- predict(model, newdata = test_set)

  # Calculate the mean squared error for this fold
  mse <- mean((predictions - test_set$avg_trip_duration)^2)

  # Store the mean squared error in the cv_mse vector
  cv_mse[i] <- mse}

# Calculate the average mean squared error over all folds
avg_mse <- mean(cv_mse)

# Print the average mean squared error over all folds
cat("The Average MSE is:",avg_mse,"\n")
```