SDS Project - MetroBike

We chose the Austin City MetroBike data set because we were curious to discover the different ways the MetroBikes have been used in Austin. We see many people using the CapMetro bike and scooter systems around the city and campus because it is a fairly accessible and relatively inexpensive mode of transportation.

According to Movability (https://austin.bcycle.com/blog/2021/01/26/metrobike-the-bike-share-you-love-with-a-new-name-and-new-features!), the city of Austin partnered up with CapMetro to transform B-Cycle, a bikeshare program, into MetroBike in 2021. MetroBike aims to strengthen the city's transportation methods and work toward objectives laid out by the Austin Strategic Mobility Plan by increasing accessibility to bikes. The program offers different payment plans, price ranges, and bike types depending on the type of experience travelers want. Analyzing this database will allow us to better understand user preferences on the MetroBike program, such as peak usage hours or how long a trip will take, which can allow us to better allocate our transportation resources and understand the dynamics of city public transportation to optimize individual needs as well.

Each unique row of our data set represents one trip that was taken on one bike. The variables that we chose for this project were:

1. ***Trip Id*** (Each MetroBike trip has an individual identifier number),
2. ***Bike Id*** (Each bike's id number),
3. ***Trip Duration*** (How long each bike trip is in minutes),
4. ***Day*** (The day of the week a trip occurred on - either a weekday or a weekend),
5. ***Checkout Time*** (The hours, minutes, and seconds that a bike was checked out for a trip),
6. ***Bike Type*** (The type of bike used - either classic or electric),
7. ***Season*** (The date used split into Fall/Winter/Spring/Summer), and
8. ***Year*** (The year the bike trip occurred).

Research Questions:

1. **Is there a relationship between checkout time and the season in which the trip takes place?** We believe that there will be a relationship - the colder the season (so fall and especially winter), the earlier the checkout time will be, as people will be less likely to go out and explore as the days get colder and darker sooner.
2. **How does Year affect the Trip Duration.** We expect trip duration to trend higher in 2019 and 2020 than other years in the dataset because of the Covid-19 lockdown. Our rationale is that the pandemic gave people more time and therefore encouraged them to do things outdoors and/or socially distanced; bike riding fit both these criteria.
3. **Is there a relationship between Bike Type and type of day?** We predict that there is a relationship between the type of bike used and the type of day the trip occurred on. Our expectation is that electric bike use might be higher on weekdays since people often have someplace to be on weekdays (ie. work or school), and electric bikes can reach higher speeds with less effort.

---

2. Methods

Uploading the packages and data we need

```
#Download the data and store it
BikeData <- read_csv("Austin_MetroBike_Trips_20240228.csv")
dim(BikeData)
```

```
## [1] 35320    14
```

We start with 35320 rows and 14 columns (variables).

Creating our new variables and cleaning the data set:

```
#Add Seasons
BikeData <- BikeData |> mutate(Season = case_when(
  Month %in% c(12,1,2) ~ "Winter",
  Month %in% c(3,4,5) ~ "Spring",
  Month %in% c(6,7,8) ~ "Summer",
  Month %in% c(9,10,11) ~ "Fall"
)) |>
#Edit the time to take away the colons
  mutate(Time = BikeData$`Checkout Time` |>
  str_remove(":") |>
  str_remove(":")) |>
# Convert Checkout.Date values to days of the week
  mutate(Weekday = ifelse(weekdays(as.Date(`Checkout Date`)) %in% c("Saturday", "Sunday"), "Weekend", "Weekday"))

#Change the var type of time to numerical data}
BikeData$Time <- as.numeric(BikeData$Time)
BikeData$Time <- BikeData$Time %/% 10000 + ((BikeData$Time%%10000)%/%100)/60

#Since we won't be working with every column in this data set, we can create a separate data set to manipulate wi
th only the variables we want to explore.
MetroBike <- BikeData |> select(
  TripId = `Trip ID`, BikeId = `Bicycle ID`,
  Duration = `Trip Duration Minutes`,
  DayType = Weekday, Time,
  BikeType = `Bike Type`,
  Season, Year) |>
  mutate(BikeType = recode(BikeType, "Classic" = "classic", "Electric" = "electric"))

summary(MetroBike)
```

```
##      TripId            BikeId             Duration          DayType
##  Min.   :20877893   Length:35320       Min.   :    2.00   Length:35320
##  1st Qu.:22343674   Class :character   1st Qu.:   16.00   Class :character
##  Median :23854387   Mode  :character   Median :   30.00   Mode  :character
##  Mean   :24301953                      Mean   :   64.67
##  3rd Qu.:25580551                      3rd Qu.:   53.00
##  Max.   :32189049                      Max.   :22993.00
##       Time         BikeType            Season               Year
##  Min.   : 0.00   Length:35320       Length:35320       Min.   :2019
##  1st Qu.:12.82   Class :character   Class :character   1st Qu.:2020
##  Median :15.57   Mode  :character   Mode  :character   Median :2021
##  Mean   :15.35                                         Mean   :2021
##  3rd Qu.:18.32                                         3rd Qu.:2021
##  Max.   :23.97                                         Max.   :2023
```

From the summary, we can see that there aren't any NA values. However, the maximum value for the Trip Duration Minutes is very high compared to the 3rd quartile value, suggesting that there will be outliers and high values we have to take into account when we analyze the data.

```
#Check our final data set
dim(MetroBike)
```

```
## [1] 35320     8
```

```
MetroBike |> mutate_all(is.na) |> summarize_all(sum)
```

```
## # A tibble: 1 × 8
##   TripId BikeId Duration DayType  Time BikeType Season  Year
##    <int>  <int>    <int>   <int> <int>    <int>  <int> <int>
## 1      0      0        0       0     0        0      0     0
```

After we successfully added and selected the variables we need, we now have 8 columns with 35320 rows. Also, as we can see, there are 0 NAs in our data set. Now, our data is tidy, as each data point has its own cell, and every observation has its own row.

3. Results

Univariate Distributions - Amanda

```
#Univariate Distribution 1: Checkout Time
MetroBike |> ggplot() +
  geom_histogram(aes(x = Time), color = "black", fill = "lavender", binwidth = 1, center = 0.5) +
  scale_x_continuous(limits = c(0, 24), breaks = seq(0, 24, 1)) +
  labs(
    title = "Distribution of Checkout Time",
    x = "Checkout Time (Hours after Midnight)",
    y = "Count",
    caption = "Figure 1"
  ) +
  guides(fill="none")
```

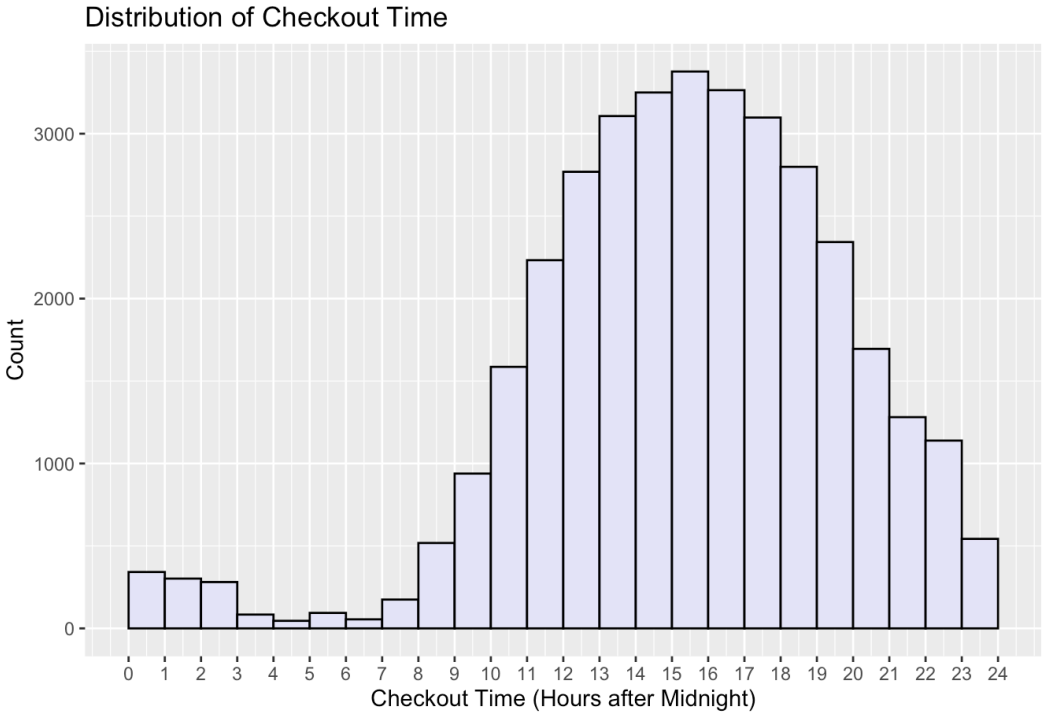


Figure 1

```
MetroBike |> summarize(median = median(Time), IQR = IQR(Time))
```

```
## # A tibble: 1 × 2
##   median   IQR
##    <dbl> <dbl>
## 1   15.6   5.5
```

From the graph, we can see how the checkout times have a fairly bell shaped distribution. This shape and distribution makes sense, as the early hours of 2-6 am would have the least amount of people taking rides, as they are usually sleeping at that time. Additionally, many more people will need a way for transportation in the middle of the day. From our summary statistics, we can see that the median time for people to check out a bike would be at about 3pm, which we can reasonably assume to be when people start coming out of school or jobs, and need to go somewhere. The IQR of the checkout time is a range of about about 5.5 hours.

```
#Univariate Distribution 2: Season
MetroBike |>
  ggplot() +
  geom_bar(aes(x=Season, fill = Season)) +
  labs(
    title = "Distrbution of Trips by Season",
    x = "Season",
    y = "Count",
    caption = "Figure 2"
  ) +
  guides(fill="none") +
  scale_x_discrete(limits = c("Spring","Summer","Fall", "Winter"))
```

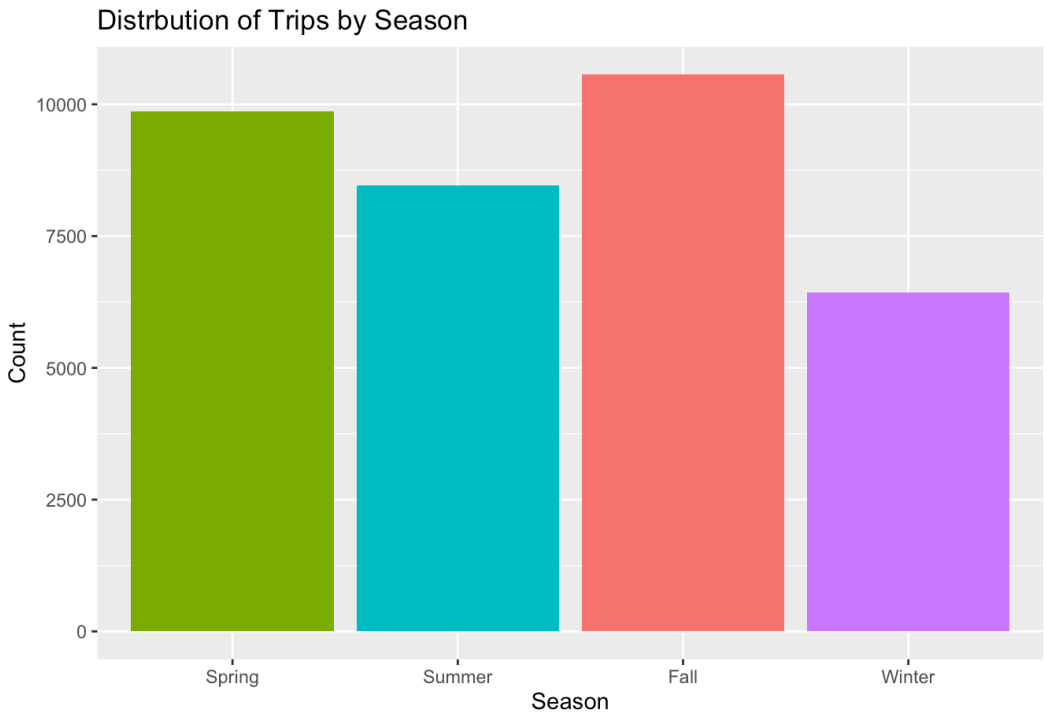


Figure 2

```
MetroBike |> group_by(Season) |> count()
```

```
## # A tibble: 4 × 2
## # Groups:   Season [4]
##   Season     n
##   <chr>  <int>
## 1 Fall   10565
## 2 Spring  9869
## 3 Summer  8461
## 4 Winter  6425
```

From the data, we can see that fall is the most popular season for tourists to come use the Metrobike - there was a total of *10565* Single Trips bought - while winter was the least popular season. There were 9869 bought in spring, 8461 bought in summer, and 6425 bought in winter.

Bivariate Distribution - Amanda

Research Question: Is there a relationship between checkout time and the season in which the trip takes place?

```
#Show the relationship between Season and the time bikes were checked out.
MetroBike |>
  ggplot() +
  geom_histogram(aes(x = Time, fill = Season),  color = "black", binwidth = 1, center = 0.5) +
  facet_wrap(~Season) +
  scale_x_continuous(limits = c(0, 24), breaks = seq(0, 24,2)) +
  guides(fill = "none") +
  labs(
    title = "Distribution of Checkout Time across Seasons",
    x = "Time (Hours after Midnight)",
    y = "Count",
    caption = "Figure 3"
  )
```

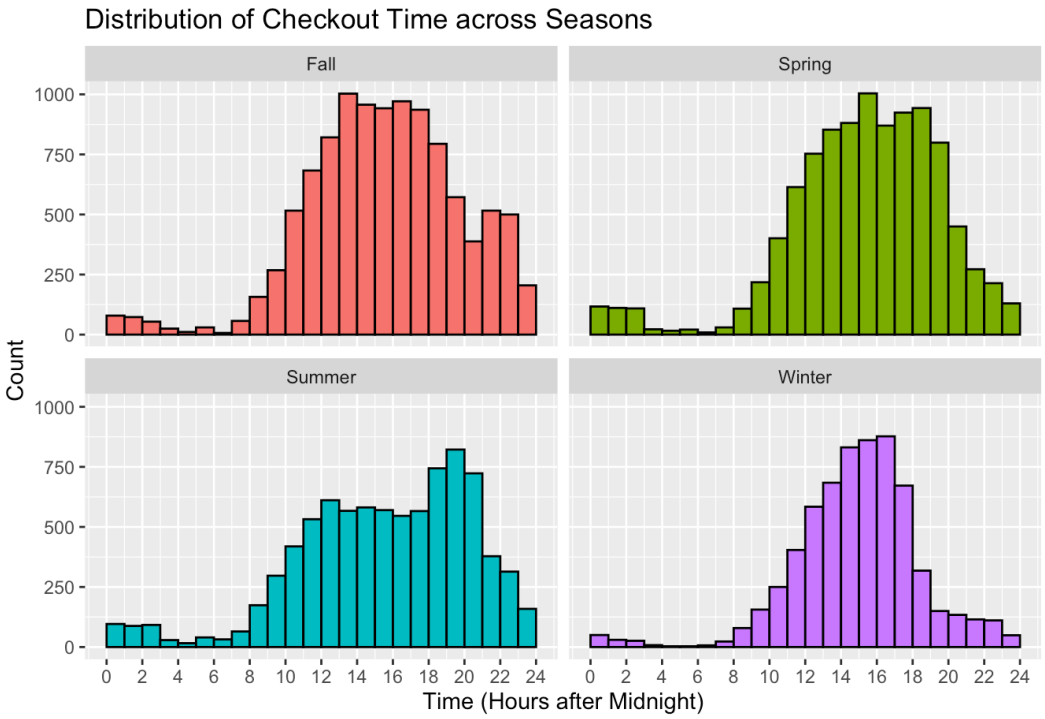


Figure 3

```
MetroBike |> group_by(Season) |> summarize(median = median(Time), IQR = IQR(Time), mean = mean(Time))
```

```
## # A tibble: 4 × 4
##   Season median   IQR  mean
##   <chr>   <dbl> <dbl> <dbl>
## 1 Fall     15.6  5.53  15.5
## 2 Spring   15.7  5.45  15.3
## 3 Summer   16.0  6.95  15.5
## 4 Winter   15.1  3.95  14.9
```

From the visualization, we can see how during the winter months, there is a steep decline later during the day after around 6pm, as it would be too cold. In summer, there was an increase in the number of users at around 8-10pm, which is when the temperature is cooler. In spring, the use peaked at around 5 to 6pm where more people used a MetroBike, as the weather is generally nicer then. The graph clearly shows the various distributions of checkout time according to the seasons, and the summary statistics proves how the median checkout time is earliest during winter (3pm), while fall and spring have similar median checkout times (3:36pm and 3:40 pm respectively), and summer has the latest checkout time at around 4 pm. Summer also has the largest range of checkout times, while winter has the smallest range.

```
#Add Year into the visualization of season and checkout time
MetroBike |>
  ggplot() +
  geom_boxplot(aes(y = Time, x = Season, fill = Season)) +
  facet_wrap(~Year) +
  guides(fill = "none") +
  labs(
    title = "Distribution of Checkout Time across Year",
    x = "Season",
    y = "Time (hhmmss)",
    caption = "Figure 4"
  ) +
scale_x_discrete(limits = c("Spring","Summer","Fall", "Winter")) +
  scale_y_continuous(limits = c(0, 24), breaks = seq(0, 24, 4))
```

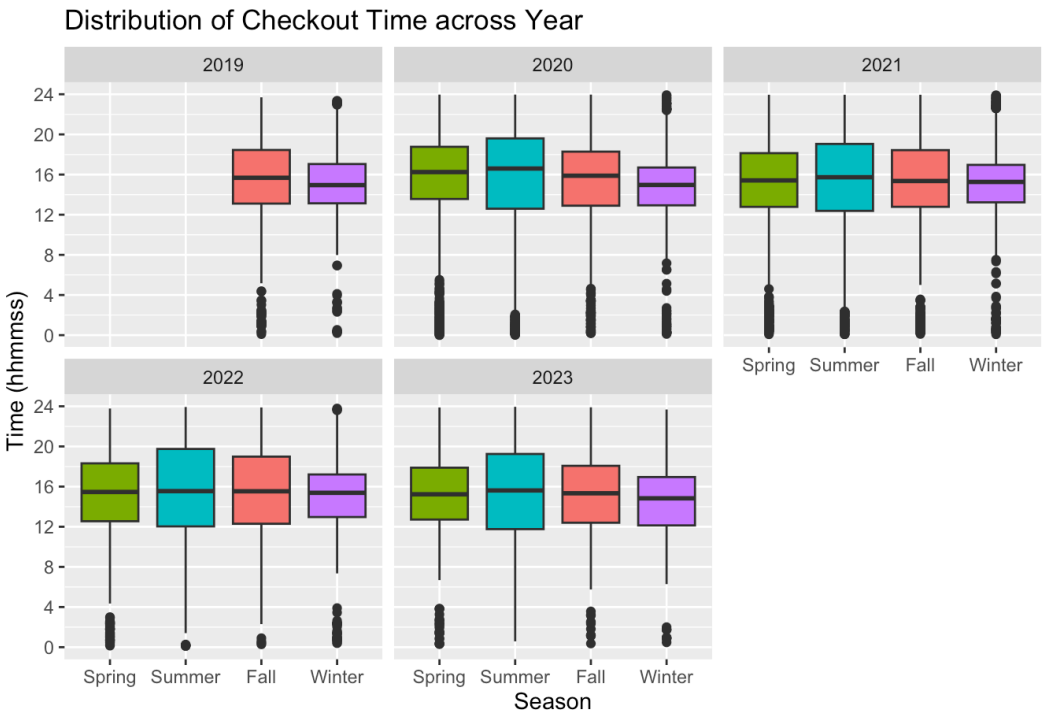


Figure 4

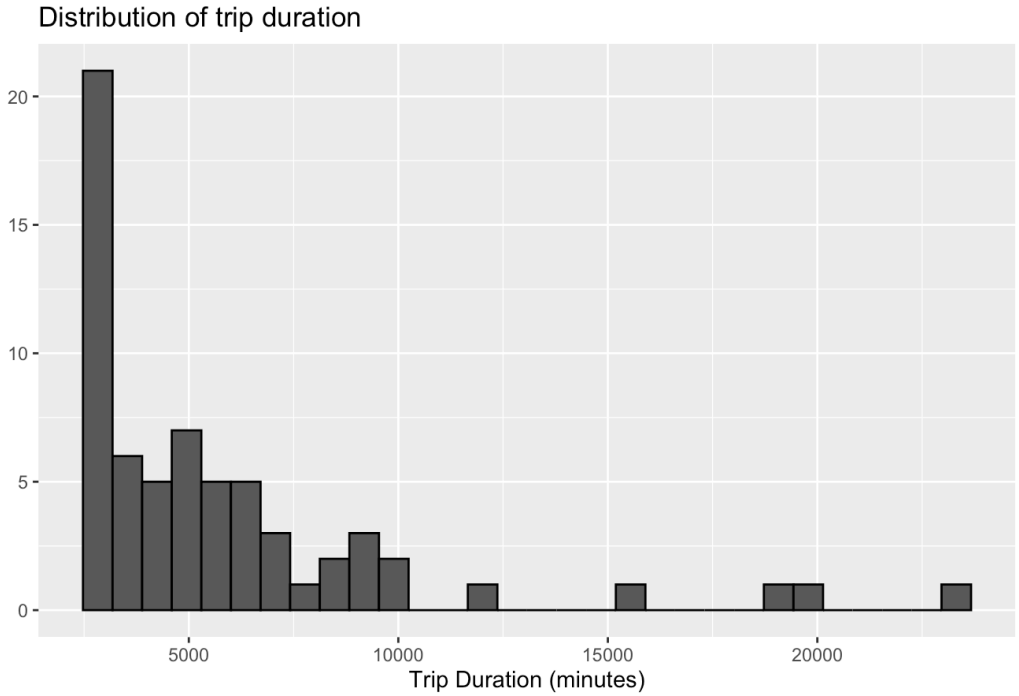
```
MetroBike |> group_by(Year, Season) |> summarize(median = median(Time), IQR = IQR(Time))
```

```
## # A tibble: 18 × 4
## # Groups:   Year [5]
##     Year Season median   IQR
##    <dbl> <chr>   <dbl> <dbl>
##  1  2019 Fall     15.7  5.34
##  2  2019 Winter   15.0  3.91
##  3  2020 Fall     15.9  5.38
##  4  2020 Spring   16.2  5.2
##  5  2020 Summer   16.6  7
##  6  2020 Winter   15.0  3.77
##  7  2021 Fall     15.4  5.65
##  8  2021 Spring   15.4  5.35
##  9  2021 Summer   15.7  6.67
## 10  2021 Winter   15.3  3.73
## 11  2022 Fall     15.5  6.68
## 12  2022 Spring   15.5  5.77
## 13  2022 Summer   15.6  7.72
## 14  2022 Winter   15.4  4.24
## 15  2023 Fall     15.3  5.67
## 16  2023 Spring   15.2  5.17
## 17  2023 Summer   15.6  7.49
## 18  2023 Winter   14.8  4.82
```

This graph was organized in chronological order, starting from fall of 2019 until the winter of 2023. Overall, although we can see the different distributions across seasons within each year, across the years the distributions do not vary considerably. The median and range of each season does not change much from 2019 to 2023, and the shape stays fairly consistent as well.

Univariate Distributions - Ava

```
# Univariate 1: trip duration (<2500, excludes 65 observations)
MetroBike |>
  filter(`Duration` > 2500) |>
  ggplot() +
  geom_histogram(aes(x = `Duration`), color = "black") +
  labs(
    x = "Trip Duration (minutes)",
    y = " ",
    title = "Distribution of trip duration",
    caption = "Note: 65 observations not included in order to fit the distribution.
    Figure 5"
  )
```



Note: 65 observations not included in order to fit the distribution.
Figure 5

```
# summary statistics
MetroBike |>
  filter('Duration' > 2500) |>
  # the data is clearly skewed, so using median and IQR
  summarize(median = median(Duration), IQR = IQR(Duration))
```

```
## # A tibble: 1 × 2
##   median   IQR
##    <dbl> <dbl>
## 1     30    37
```

This graph tells us that a large proportion of trips had a duration under 5000 minutes. This histogram is highly skewed right. The range of Trip Durations is too wide to be seen on a graph properly; here are the distributions for trip duration times below 2500 minutes. This number was chosen because there were very few outliers after 2500 minutes.

```
# Univariate 1: trip duration (natural log, includes all observations)
MetroBike |>
  ggplot() +
  geom_histogram(aes(x = `Duration`), color = "black") +
  labs(
    x = "Trip Duration (minutes)",
    y = " ",
    title = "Distribution of trip duration (natural log fitted onto x axis)",
    caption = "Figure 6"
  )   +
  scale_x_continuous(trans = "log10")
```

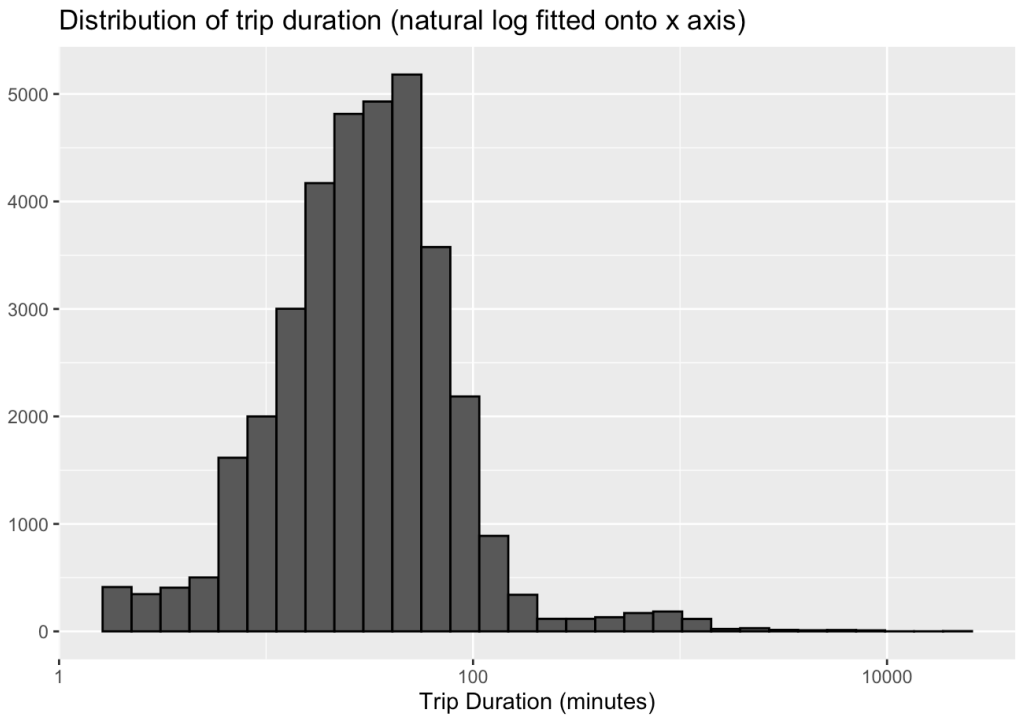


Figure 6

```
# summary statistics
MetroBike |>
  filter('Duration' > 2500) |>
  # the data is rather normal, so using mean and SD
  summarize(mean = mean(Duration), 'standard deviation' = sd(Duration))
```

```
## # A tibble: 1 × 2
##    mean `standard deviation`
##   <dbl>                <dbl>
## 1  64.7                 328.
```

This histogram is much closer to a normal distribution, though it is still clearly skewed right. Because we had to cut the duration short, the first graph is not representative of our data. This is trip duration fitted on the log scale in order to see all of the values graphed.

```
# Univariate 2: year
MetroBike |>
  ggplot() +
  geom_bar(aes(x = Year)) +
  labs(
    x = "Year",
    y = "Number of Rides",
    title = "Number of recorded bike rides per year",
    caption = "Figure 7"
  )
```

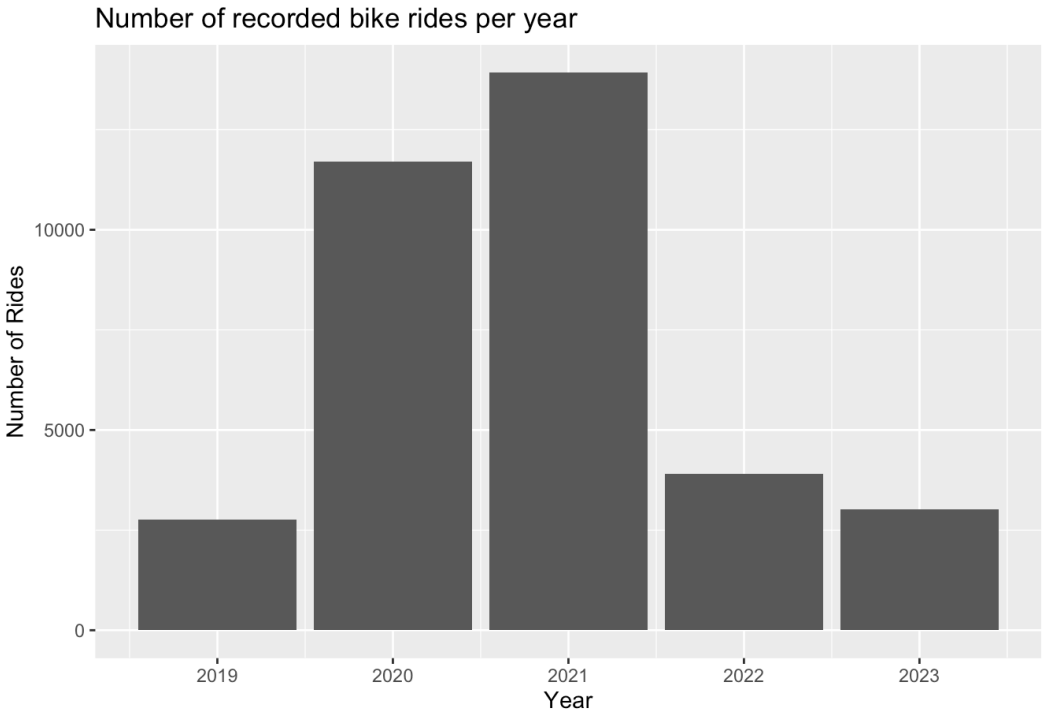


Figure 7

```
# summary statistics: frequency by year
MetroBike |>
  group_by(Year) |>
  # count and proportion
  summarize(count = n(),
            proportion = n() / nrow(MetroBike))
```

```
## # A tibble: 5 × 3
##    Year count proportion
##   <dbl> <int>      <dbl>
## 1  2019  2769     0.0784
## 2  2020 11704     0.331
## 3  2021 13924     0.394
## 4  2022  3911     0.111
## 5  2023  3012     0.0853
```

There were significantly more bike rides in 2020 and 2021 than in the years 2019, 2022, and 2023. 2021 saw the most bike rides.

Bivariate Distribution - Ava

```
# Bivariate 1: year vs. trip duration
MetroBike |>
  ggplot() +
  geom_point(aes(x = Year, y = `Duration`)) +
  labs(
    x = "Year",
    y = "Trip Duration (minutes)",
    title = "Trip durations of rides (minutes) by year",
    caption = "Figure 8"
  )
```

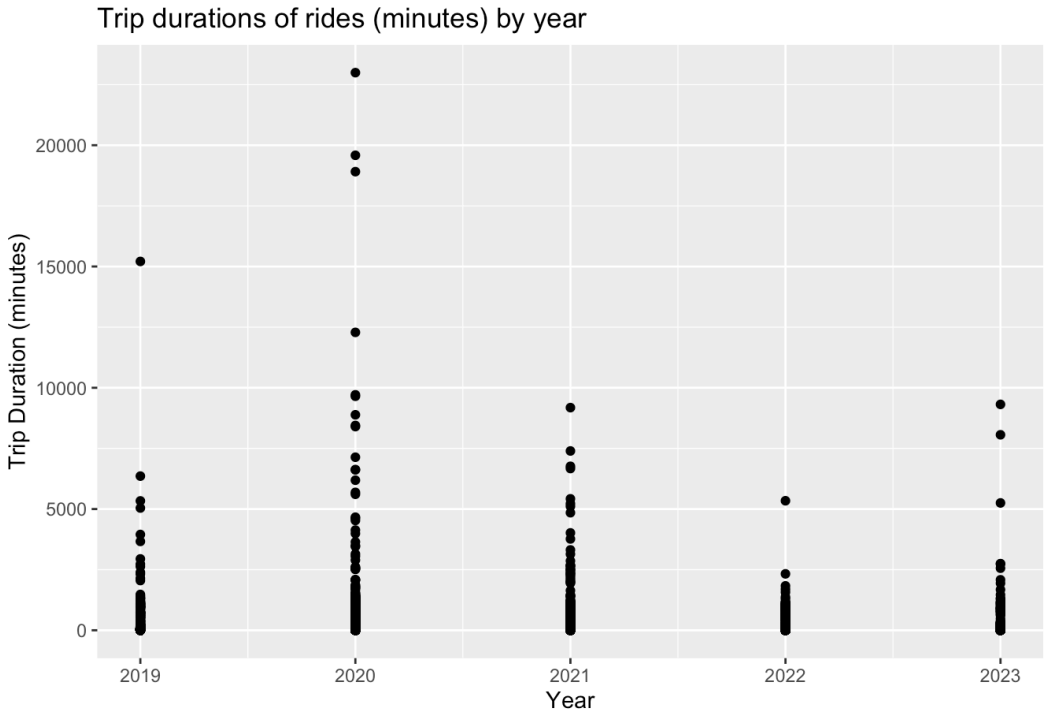


Figure 8

```
# summary stats
MetroBike |>
  group_by(Year) |>
  # most of the distributions are skewed, so using median/IQR
  summarize(median = median(Duration),
          IQR = IQR(Duration))
```

```
## # A tibble: 5 × 3
##    Year median   IQR
##   <dbl>  <dbl> <dbl>
## 1  2019     26    31
## 2  2020     34    38
## 3  2021     30    37
## 4  2022     27    34
## 5  2023     27    36
```

From this graph, we notice that the majority of bike trips trended higher or had a larger range in 2019 and 2020.

```
# time duration vs. checkout time by year
MetroBike |>
  ggplot() +
  geom_point(aes(x = `Time`, y = `Duration`), alpha = 0.2) +
  scale_y_continuous(limits = c(0, 2000), breaks = seq(0, 2000, 250)) +
  labs(
    x = "Checkout Time",
    y = "Trip Duration (minutes)",
    title = "Trip Duration vs. Checkout Time, by Year",
    caption = "Figure 9"
  ) +
  facet_wrap(~Year)
```

```
## Warning: Removed 82 rows containing missing values (`geom_point()`).
```

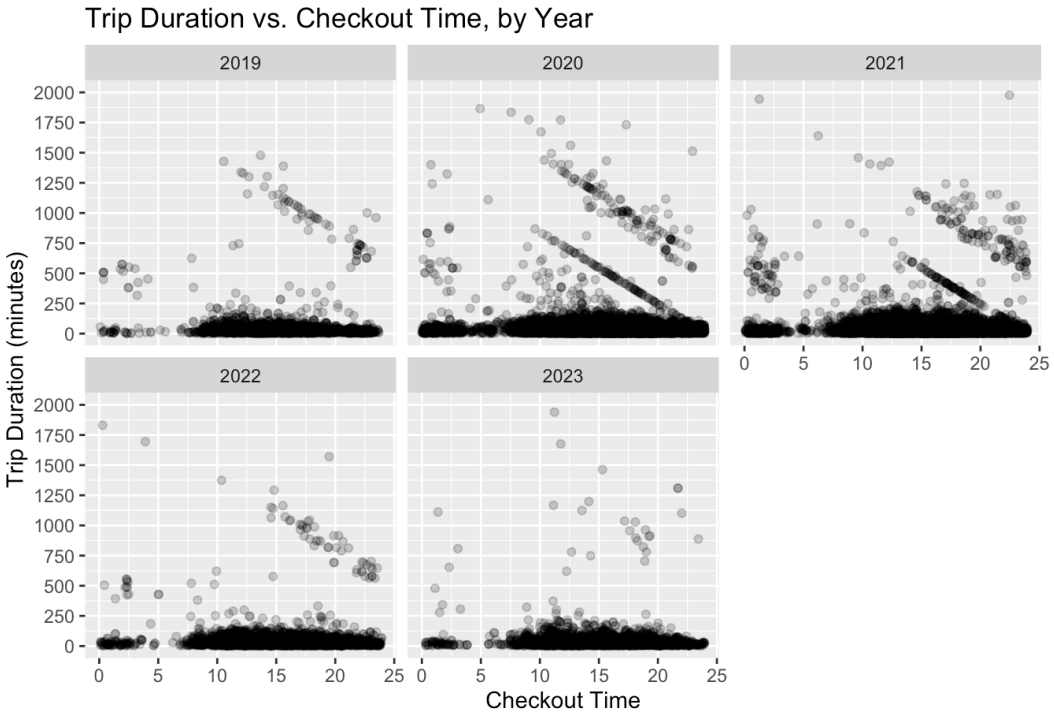


Figure 9

```
# summary stats
MetroBike |>
  group_by(Year) |>
  summarize('Mean Checkout Time' = mean(Time),
          'SD Checkout Time' = sd(Time),
          'Mean Trip Duration' = mean(Duration),
          'SD Trip Duration' = sd(Duration))
```

```
## # A tibble: 5 × 5
##    Year `Mean Checkout Time` `SD Checkout Time` `Mean Trip Duration`
##   <dbl>                <dbl>              <dbl>                <dbl>
## 1  2019                 15.6               4.07                 77.5
## 2  2020                 15.4               4.40                 75.3
## 3  2021                 15.3               4.33                 58.3
## 4  2022                 15.3               4.22                 52.2
## 5  2023                 15.1               4.10                 57.5
## # ℹ 1 more variable: `SD Trip Duration` <dbl>
```

Since the graph before this one doesn't tell us much, here is Trip Duration and Checkout Time plotted against each other and facet-wrapped by Year. This graph highlights some clusters of duration/time values. In 2020 and 2021, there are some "clumps" of data where there is a strong, positive correlation between Checkout Time and Trip Duration. Note: R omitted 82 values because they fell outside of our defined range for the x-axis; this was done to fit the data on the graph.

```
# bivariate: distribution of trip duration, facet wrapped by year and on log scale
MetroBike |>
  ggplot() +
  geom_histogram(aes(`Duration`), color = "black") +
  facet_wrap(~Year) +
  scale_x_continuous(trans = "log10") +
  labs(
    title = "Trip Duration on the Log Scale by Year",
    x = "Log of Trip Duration",
    y = "Count",
    caption = "Figure 10"
  )
```

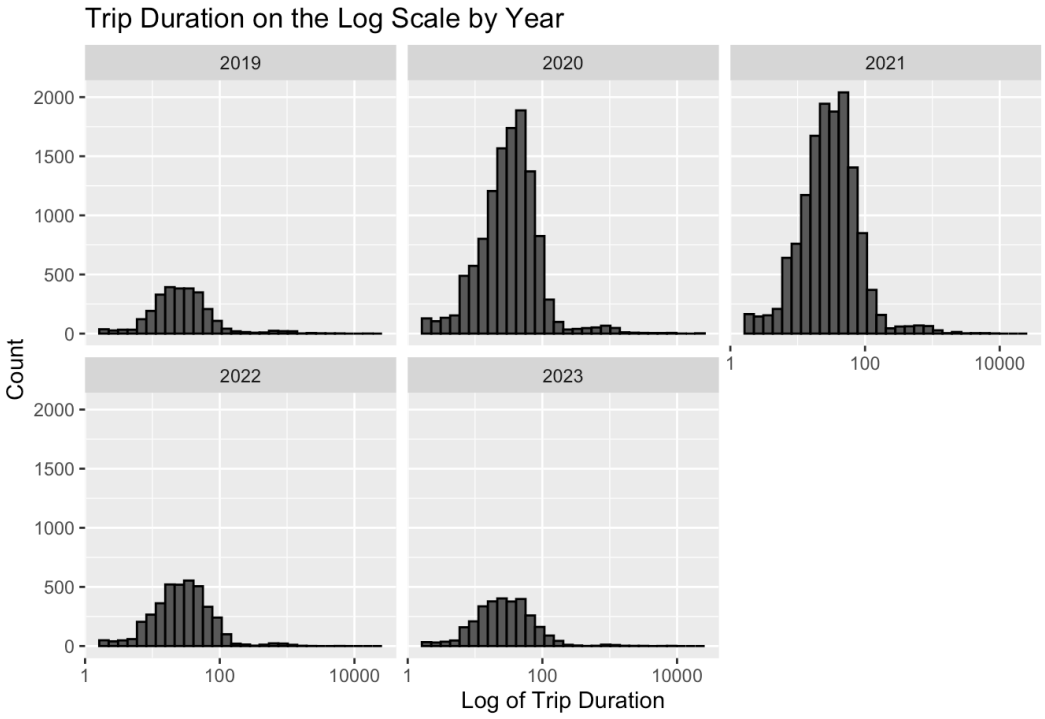


Figure 10

These graphs show us that the distribution of trip duration was most normal in 2019, 2021, and 2023. Out of these, 2019 has the least right skew, as it does not contain some extremely high values.

Univariate Distribution - Zoe

Investigation of the `BikeType` variable: Univariate Visualization

```
MetroBike|>
  # Make a bar graph of the different bike types
  ggplot()+
  geom_bar(aes(x=BikeType), fill = "coral")+
  # Add title and labels
  labs(title="Counts of Different Bike Types Used",
       x="Bike Type",
       y= "Number of Trips",
       caption = "Figure 11")
```

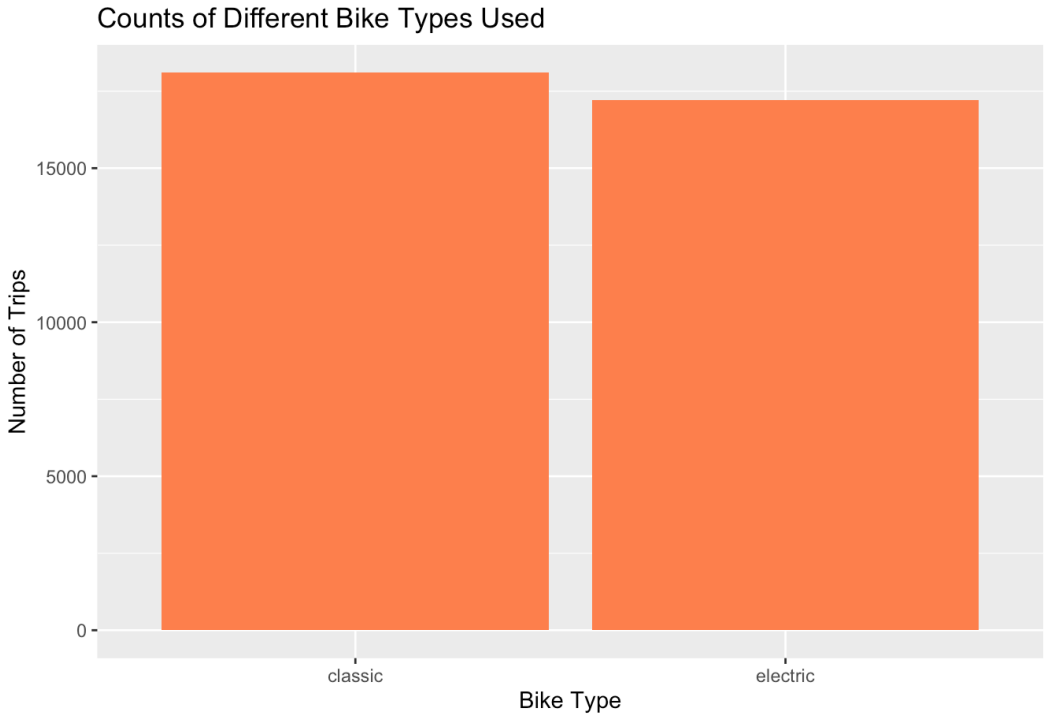


Figure 11

```
# Summary statistics
MetroBike |>
   # Split the data in groups
  group_by(BikeType) |>
  # Summarize per bike type
  summarize(count = n(),
            proportion = n() / nrow(MetroBike))
```

```
## # A tibble: 2 × 3
##   BikeType count proportion
##   <chr>    <int>      <dbl>
## 1 classic  18102      0.513
## 2 electric 17218      0.487
```

There is a relatively equal distribution of trips using classic and electric bikes, with a slightly greater amount of trips with classic bikes present. 18,102 trips, or 51.3% of the "Single Trip" journeys, used classic bikes. 17,218 trips, or 48.7% of the "Single Trip" journeys, used electric bikes.

Investigation of the Day Type variable: Univariate Visualization

```
MetroBike|>
  # Make a bar graph of the different day types
  ggplot()+
  geom_bar(aes(x=DayType), fill = "lightgreen")+
  # Add title and labels
  labs(title="Counts of Different Day Types",
       x="Day Type",
       y= "Number of Trips",
       caption = "Figure 12")
```

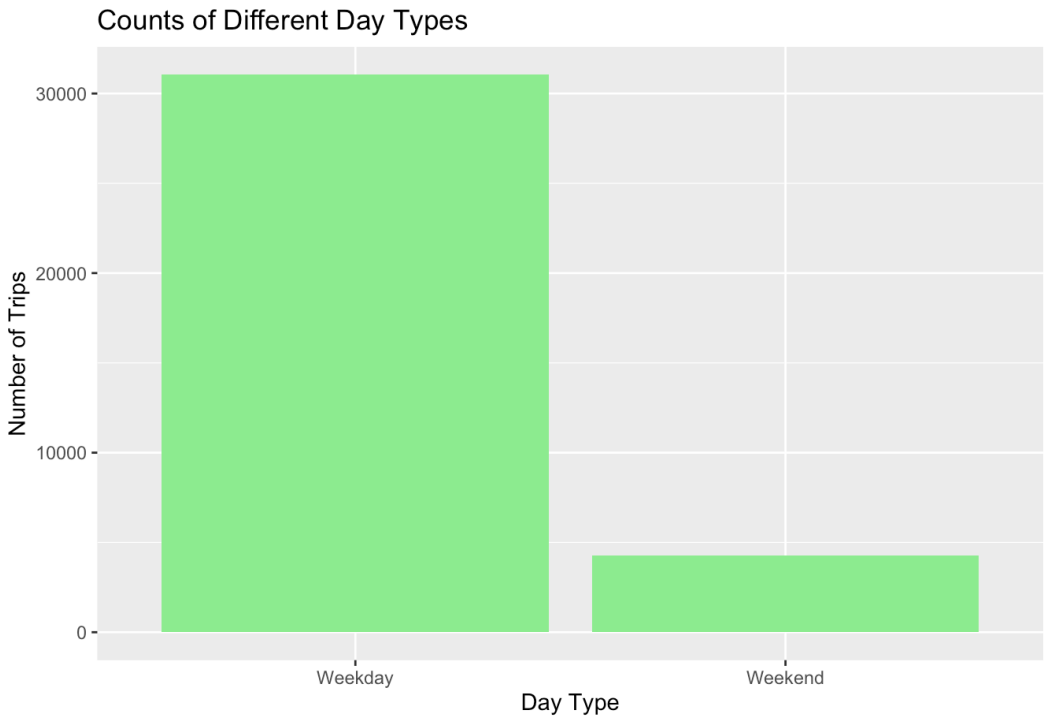


Figure 12

```
# Summary statistics
MetroBike |>
   # Split the data in groups
  group_by(DayType) |>
  # Summarize per bike type
  summarize(count = n(),
            proportion = n() / nrow(MetroBike))
```

```
## # A tibble: 2 × 3
##   DayType count proportion
##   <chr>   <int>      <dbl>
## 1 Weekday 31052      0.879
## 2 Weekend  4268      0.121
```

The majority of the trips in the dataset occurred on a weekday. 31,052 trips, or 87.9% of the trips, happened on a weekday. 4,268 trips, or 12.1% of the trips, took place on a weekend. This makes sense since we defined 5 days of the week to be weekdays, so we'd expect a considerably larger sample of weekday observations. The smaller number of weekend observations should be kept in consideration during further analysis.

Bivariate Distribution - Zoe

Investigation of the Day Type variable and the Bike Type variable: Multivariate Visualization #1

```
MetroBike|>
  mutate(BikeType=recode(BikeType, "classic"="Classic","electric"="Electric"))|>
  # Make a segmented bar graph using the two variables
ggplot() +
  # Analyze the proportion of bike types per day type
  geom_bar(aes(x = DayType, fill = BikeType), position = "fill") +
  # Add main title, axis labels, and legend title
labs(title="Proportion of Bike Types for Each Day Type",
     x= "Day Type",
     y = "Proportion of Bike Types",
     fill = 'Bike Type',
     caption = "Figure 13")
```

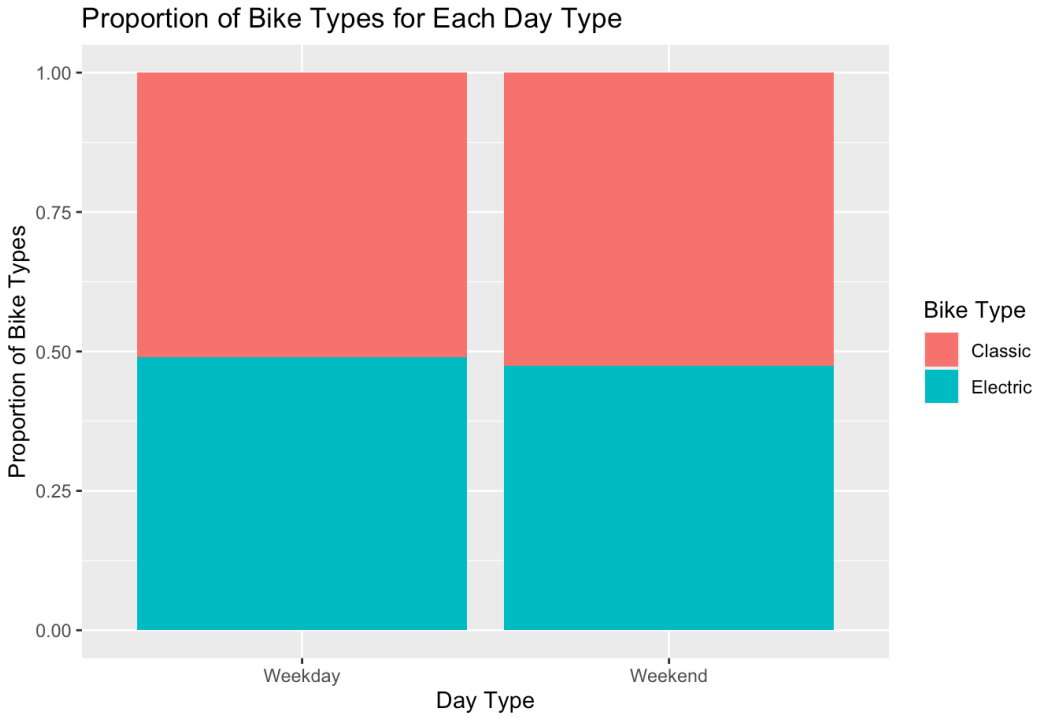


Figure 13

```
# Find frequencies
table(MetroBike$BikeType, MetroBike$DayType)
```

```
##
##            Weekday Weekend
##   classic    15859    2243
##   electric   15193    2025
```

```
# Find proportions (from frequency table)
prop.table(table(MetroBike$BikeType, MetroBike$DayType), 2)
```

```
##
##              Weekday   Weekend
##   classic  0.5107239 0.5255389
##   electric 0.4892761 0.4744611
```

The proportion of bike types across the day types is highly similar, with a slightly greater presence of classic bikes in both day types. Out of all the "Single Trip" journeys that occurred on the weekday, 51.1% of the trips, or 15,859 trips, used a classic bike, and 48.9% of the trips, or 15,193 trips, used an electric bike. Out of all the "Single Trip" journeys that occurred on the weekend, 52.6% of the trips, or 2,243 trips, used a classic bike, and 47.4% of the trips, or 2,025 trips, used an electric bike. The fairly equivalent set of proportions across day types suggest that there is no relationship between the two variables.

Investigation of the Day Type variable and the Bike Type variable: Multivariate Visualization #2

```
MetroBike |>
# Create a ggplot
ggplot() +
  # Create histograms and define mapping aesthetics
  geom_histogram(aes(x = log(Duration)),
  # Color bins green and outline in black
  fill="lightgreen", color="black",
  # Set binwidth and center
  binwidth=0.3)+
  # Adjust the tick marks of the x-axis
  scale_x_continuous(breaks = seq(0,10.5,1))+
  # Facet per category and display in two columns
  facet_wrap(~DayType + BikeType, ncol = 2)+
  # Add title and labels
  labs(title="Log of Trip Duration Grouped by Day Type and Bike Type",
       x="Log of Trip Duration",
       y="Number of Trips",
       caption = "Figure 14")
```

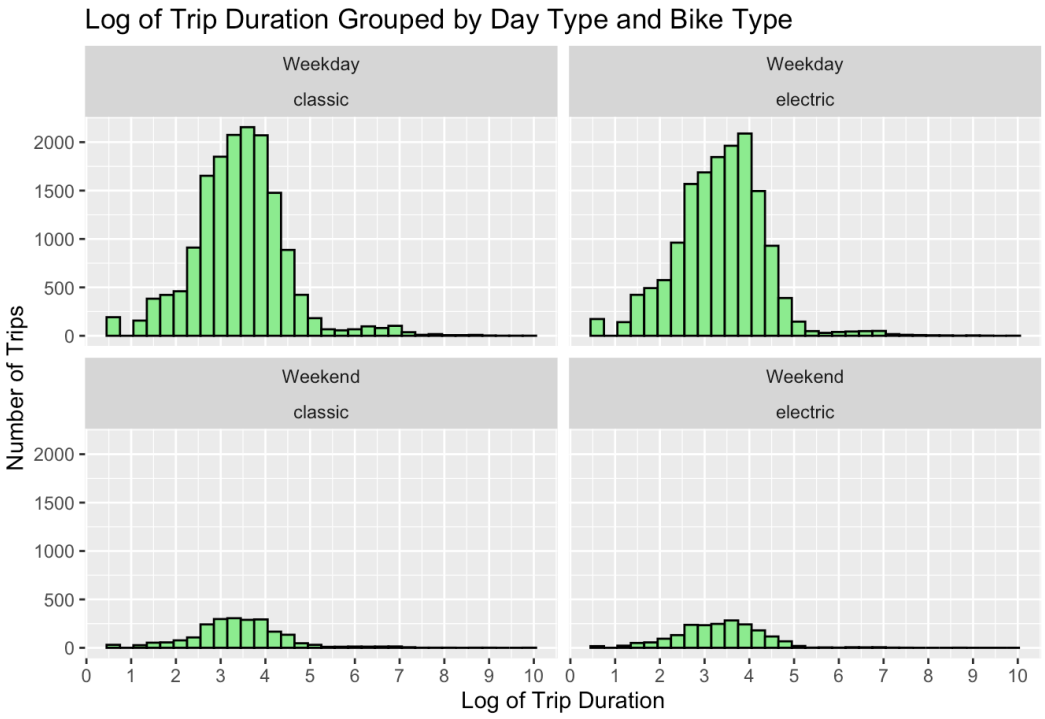


Figure 14

```
MetroBike |>
  # Split the data in groups
  group_by(BikeType, DayType) |>
  # Summarize median and IQR per group
  summarize(median = median(log(Duration)), IQR = IQR(log(Duration)))
```

```
## # A tibble: 4 × 4
## # Groups:   BikeType [2]
##   BikeType DayType median   IQR
##   <chr>    <chr>    <dbl> <dbl>
## 1 classic  Weekday   3.43  1.16
## 2 classic  Weekend   3.37  1.12
## 3 electric Weekday   3.40  1.20
## 4 electric Weekend   3.37  1.22
```

Across all day type and bike type combinations, the distribution of the log of trip duration seems to be highly similar. All distributions are skewed to the right, with a larger concentration of trips on the lower end of the log of trip duration values (0 to 5) than the higher end. The median log of trip duration for the trips using classic bikes on weekdays was 3.43. The corresponding IQR was 1.16. The median log of trip duration for the trips using classic bikes on weekends was 3.37. The corresponding IQR was 1.11. The median log of trip duration for the trips using electric bikes on weekdays was 3.40. The corresponding IQR was 1.20. The median log of trip duration for the trips

using electric bikes on weekends was 3.37. The corresponding IQR was 1.22. It is most common for the log of trip duration value to be 3 to 4 across all four distributions. However, it is also quite common to have a log of trip duration value from 2.5 to 3 for "Single Trip" excursions that occurred during the weekend. Regardless of the bike type or day type, it is less common to have a log of trip duration value greater than 5. The general similarity between the four distributions suggest there is no relationship between the bike type and day type with the log-transformed trip duration.

---

4. Discussion

1. **Takeaways:** One main takeaway from our analysis is that there is a relationship between the bike checkout time and the season in which a trip took place. Checkout times during winter occurred less often in the evening than compared to other seasons. There is also a relationship between the year in which a trip occurred and its duration, with trips in 2020 having the longest trip duration. We concluded that there does not seem to be a relationship between the type of bike used and the type of day in which a trip takes place. Since the dataset only consists of information connected to "Single Trip" passes, the takeaways are mainly applicable to "Single Trip" journeys and are not entirely representative of all MetroBike trips.
    1. **Is there a relationship between checkout time and the season in which the trips take place? (Amanda)** There does seem to be a relationship between the season and checkout time. From graph 3, we can see how winter had the earliest checkout times, experiencing a steep decline in the number of checkouts after around 5pm, while the other (warmer) months had a much more smooth distribution of checkout times, as well as having more people checking out at any given time. According to the summary statistics for graph 3, summer has the largest spread in time, while winter has the smallest, and the spring and fall seasons have fairly similar distributions. These observations make sense with what we predicted, as we thought there would be less trips taken out in the cold during winter, and during summer people have more free time to travel at any time during the day. One surprising detail I found was that the checkout time data did not fluctuate significantly across years - although the number of trips and the trip duration did change significantly between covid and non-covid years, the actual time people decided to start their journeys (as opposed to how long/how many people) was pretty stable.
    2. **How does Year affect the Trip Duration? (Ava)** Year seems to affect Trip Duration in that there are longer trips in 2020. This is most apparent in Figure 8, where the range for trip duration is clearly the largest by far for the year 2020. Additionally, there were significantly more rides in 2020 and 2021. These findings were aligned with my predictions about bike usage going up during the COVID-19 lockdown since socially distanced and/or outdoor activities shot up in popularity during this time. This conclusion is further proven by Figure 10, which shows that 2020 and 2021's distributions of duration had significantly higher ranges than 2019, 2022, and 2023. Thus, we can say there is a relationship between Year and Trip Duration, though it isn't necessarily numeric. Trip Duration peaks the closer the Year value is to 2020 and 2021 (thinking of Year as a numeric variable); or, the years 2020 and 2021 had longer trip duration overall (thinking of Year as a categorical variable).
    3. **Is there a relationship between Bike Type and Day Type? (Zoe)** There does not seem to be a relationship between the type of bike used during a trip and the type of day on which the trip occurred. Looking at Figure 13, it seems that the proportion of bike types are fairly equal across both day types. Out of all the "Single Trip" journeys that occurred on the weekday, 51.1% of the trips, or 15,859 trips, used a classic bike, and 48.9% of the trips, or 15,193 trips, used an electric bike. Similarly, 52.6% of the weekend "Single Trip" trips, or 2,243 trips, used a classic bike, and 47.4% of the trips, or 2,025 trips, used an electric bike. The closeness in bike type proportions across day types suggest that there is not a relationship between the type of day a trip took place and the type of bike used. This conclusion contradicts our expectation that there would be a large amount of trips conducted with electric bikes on the weekdays. Across both the weekend and weekday trips, there was a slightly higher use of classic bikes. We found this surprising because we expected people to gravitate towards newer technology, like the electric bike. However, it is possible that some people would rather stick with what they're familiar with, leading to that slight majority of classic bikes. Figure 14 further explored the bike type and day type variables to investigate their relationship with the log of trip duration. The similar shape and trends of all four distributions suggested that there was no relationship between the variables. Lastly, it should be noted that the sample size of weekend trips was much less than the sample size of weekday trips, so the applicability of our findings is somewhat limited.
2. One ethical concern in our findings is that they were derived from publicly sourced data. Though this is not inherently a concern, the people who "created" these data by riding MetroBikes may not know their data is being collected and published for free use electronically. Despite there being no identifying data connected to the MetroBike observations, nowhere does it say that riders are told their data is being collected and made easily accessible by anyone—e.g., three random college students. However, our results can positively affect the community in that they reflect the Austin community's consistent use—and thus demonstrated need—for the MetroBike system. Despite the decline of its use after 2021, the data shows a regular use of MetroBikes in Austin, even post-lockdown.

---

5. Reflection

1. The most challenging aspect of the project was cleaning the dataset so that we could work with the data we needed. Although we were lucky that we did not have to deal with NA values, we did have to select the variables we needed and transform some of the variables into a form that we could analyze. Additionally, we had to use the data given to us and manipulate them into fitting the guidelines for our

project. For example, we had to brainstorm ways to make a date fit a maximum-4-categories categorical variable. At first, we wanted to change the date to days of the week, but there were more than 4 days in the week (7), so we had to then change the variable again to just weekday or weekend, which we then had to discuss because of the unequal distribution of days within the two categories. Another obstacle we encountered while cleaning was the question of how to show the checkout time, which was originally given as hh:mm:ss. At first we simply took away the colons and stored it as a numerical variable in the form hhmmss, but then we had to perform the mathematical operators to transform the checkout time in terms of number of hours after midnight, taking into account the minutes as well.

2. This process has taught us valuable lessons about tidying, cleaning, and understanding the nuance of large data sets. When transforming variables such as `Date` into seasons, and day type (weekend/weekday), we had to apply data manipulation principles to the limitations of how we record time—what seasons are when; or how a week is divided into weekday/weekend. Our process when manipulating `Checkout Time` was similar in that we had to change the data we were given into something a ggplot function could process as linear. The data manipulation process for just about every variable we used taught us how to troubleshoot and adjust given data to answer questions with nuance beyond the data at face value.

3. First and foremost, thank you to Professor Guyot for providing us the framework and toolkit to independently conduct analyses in R. Thank you to the City of Austin for this MetroBike data, as well as Ciara, Vaishnavi, and Dustyn for their clarifying notes on codes. Additionally, as a team, we worked together on all of the questions except the ones relating to our individual research question - for example, we worked on the introduction, methods, discussion, and reflection sections together as a group, but we worked on our question for our graphs and visualizations in the results section, which are clearly labeled by our names.

4. Citations:

    1. “The Bike Share You Love, with a New Name and New Features!” Metrobike, austin.bcycle.com/blog/2021/01/26/metrobike-the-bike-share-you-love-with-a-new-name-and-new-features! Accessed 28 Mar. 2024.

    2. City of Austin, Texas - data.austintexas.gov. “Austin MetroBike Trips: Open Data: City of Austin Texas.” Data.AustinTexas.Gov - The Official City of Austin Open Data Portal, 12 Feb. 2024, data.austintexas.gov/Transportation-and-Mobility/Austin-MetroBike-Trips/tyfh-5r8s/about_data.

    3. GfG. “Convert Date to Day of Week in R.” GeeksforGeeks, GeeksforGeeks, 23 May 2021, www.geeksforgeeks.org/convert-date-to-day-of-week-in-r/.

    4. “Remove Matched Patterns - Str_remove.” - Str_remove • Stringr, stringr.tidyverse.org/reference/str_remove.html. Accessed 28 Mar. 2024.

    5. robk@statmethods.net, Robert Kabacoff -. “Operators.” Quick-R: Operators, www.statmethods.net/management/operators.html#:~:text=How%20can%20I%20perform%20integer,yields%20a%20quotient%20of%2(Accessed 28 Mar. 2024.