AUSTIN

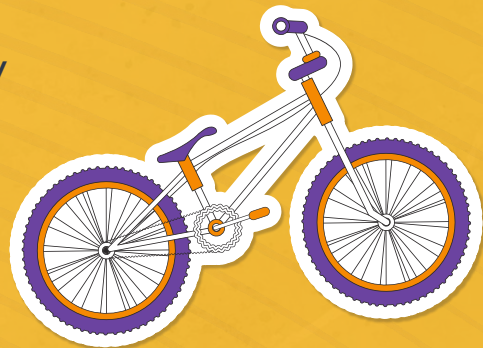# INVESTIGATION OF
Pay As You Ride
METROBIKE TRIPS

Ava Lim, Amanda Zhang, and Zoe Ly

BIKE
BIKE

RIDE
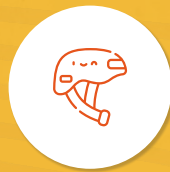RIDE
RIDE

Our Motivations

Large presence in Austin

- Fairly accessible and inexpensive mode of transportation

- Seen all over the city and UT campus

Explore how the bikes are used

- Interested in understanding user preferences

- Want to identify trends that allow for better allocation of resources or optimize user needs

BIKE
BIKE

RIDE
RIDE

Cleaning and Formatting our Data

Variables

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

Formatting

- Season
- Checkout Time
- Day

Cleaning

- Filtered for Single Trip Data (Pay as you Ride Single Trip)

- No NA values!

- Original Dataset had 14 columns, we selected 8 of them. 35320 Observations Total

```r
#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)
```
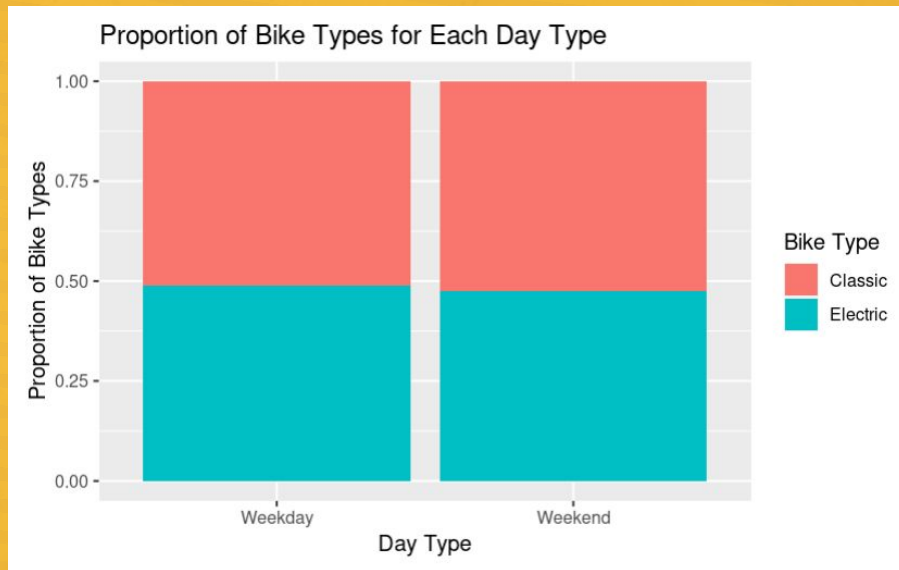
Is there a relationship between the type of bike used and the type of day in which the trip occurred?

Proportion of Bike Types for Each Day Type

Frequencies of Bike Types Per Day Type

	Weekday	Weekend
classic	15859	2243
electric	15193	2025

Proportions of Bike Types Per Day Type

	Weekday	Weekend
classic	0.5107239	0.5255389
electric	0.4892761	0.4744611

There seems to be no relationship between the type of bike used and the type of day on which the trip occurred.
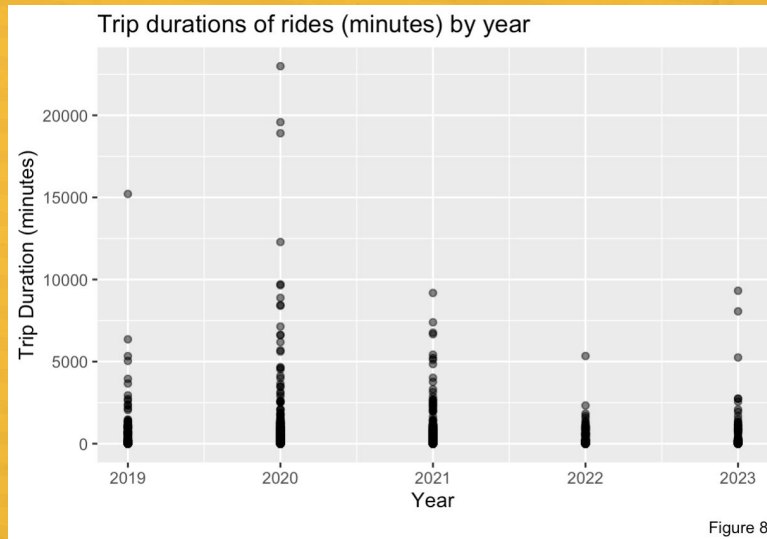
Is there a relationship between the year a trip occurred and its length?

There seems to be a relationship between Year and Trip Duration, as the spread of Trip Duration is noticeably larger in the years 2020 and 2021.

Figure 8

Figure 11
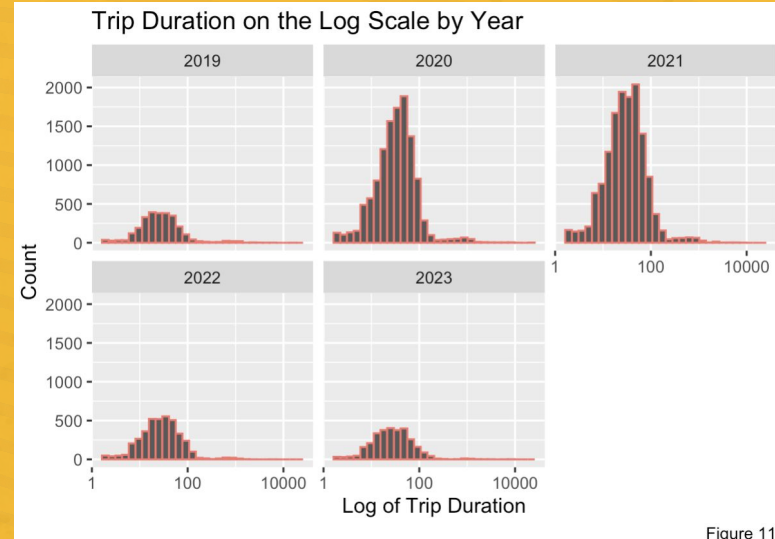
Note that this relationship is not linear, and that with 5 discrete years, Year could be considered as either a numeric or categorical variable.

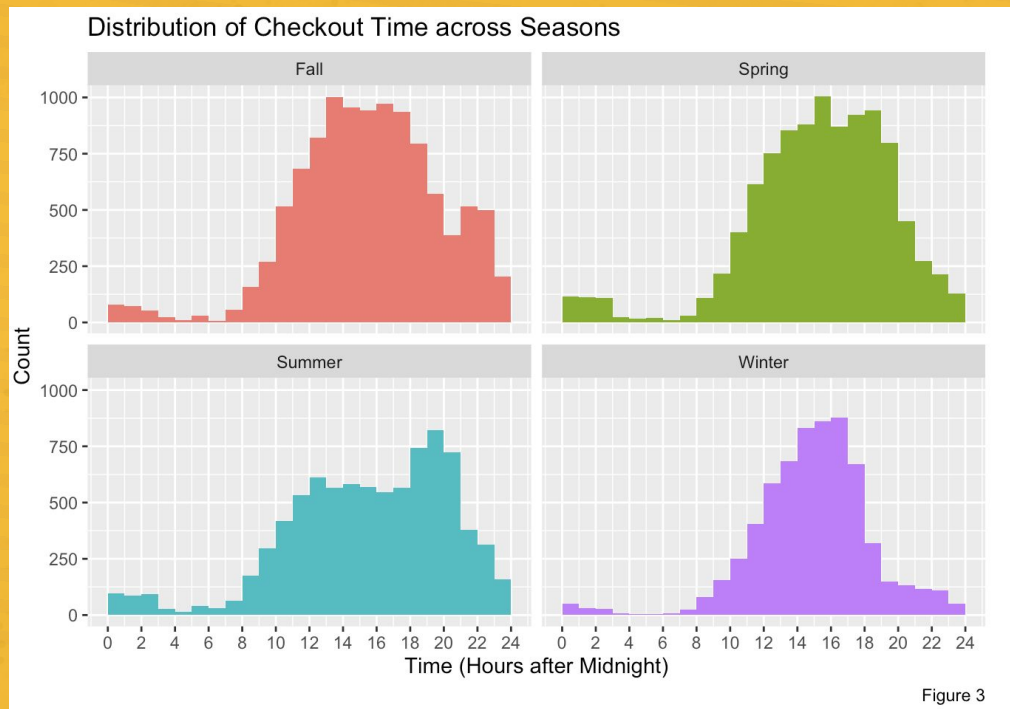# Is there a relationship between the Season and Checkout Time of every trip?

Figure 2
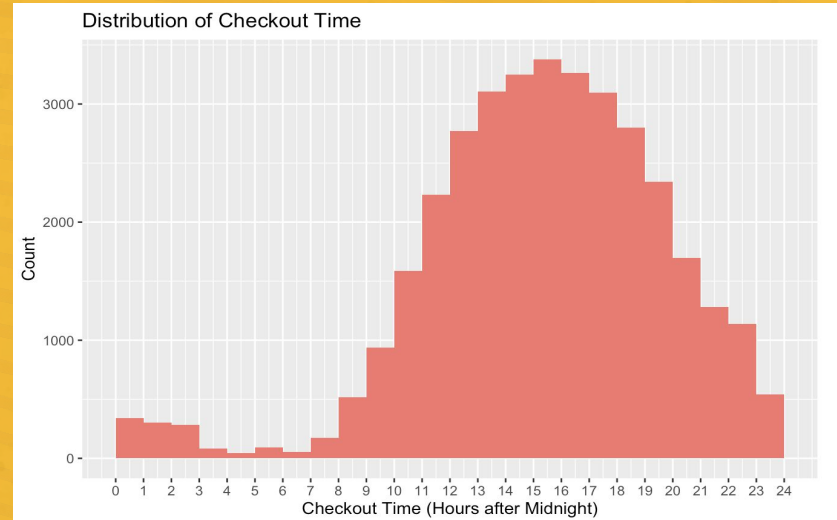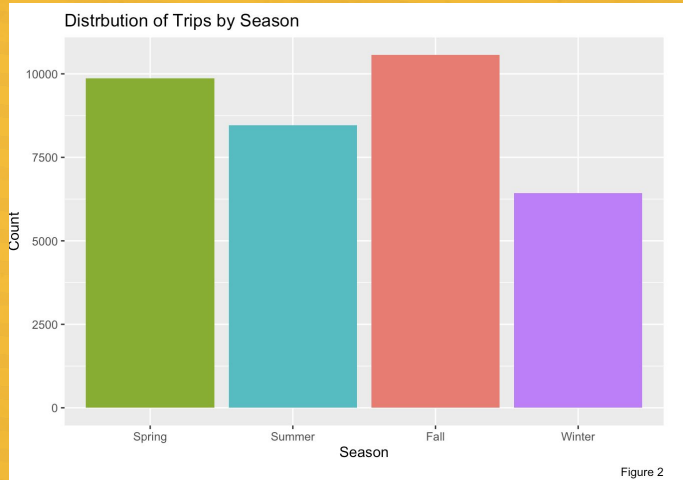
These distributions of Season and Checkout Time follow a pretty expected distribution that can be explained logically.
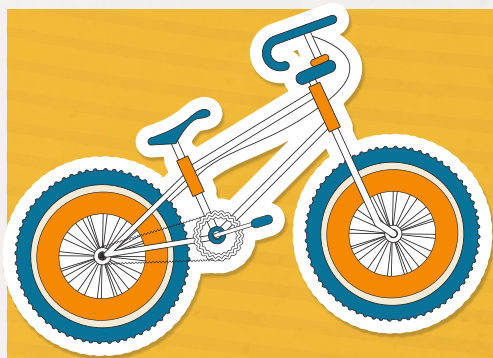
RIDE
RIDE
RIDE

(cont.) Some relationship between Season and Checkout Time

Distribution of Checkout Time across Seasons

Figure 3

Although the distribution across seasons follows mainly the same shape or distribution, we can see how each season has its own details and unique points, especially winter.

Conclusion!

Whether a trip **occurs on a weekend or weekday** most likely does not influence the **proportion of classic or electric bikes** used in a day, though we can conclude that **checkout times** tend to be earlier in the day in **Winter** months, and that the **year** during which a ride took place can affect its **duration** based on this dataset.