Devoir1

2024-09-28

Librairies and date importation

```
mydata <- read.csv("C:/Users/Admin/OneDrive/Desktop/Automne 2024/Statistical modeling/GitPractice/Practi
#head(mydata)
```

Cleaning and data preparation

Cleaning function

```
preproc_dat <- function(mydata){
  dat <- mydata

  #Changing variables (mem,wday,rushhours) to factors, creating var to code for weekend and rushmod (mo
  dat <- mutate(dat, wend = as.factor(as.integer(dat$wday %in% c("Saturday","Sunday"))))  #creates vari
  dat <- mutate(dat, rushmod = rushhour)
  dat$rushmod[dat$wend==1] <- 3
  dat <- mutate(dat,across(c(mem,wday,rushhour,rushmod), as.factor))

  return(dat)
}
```

```
dat  <- preproc_dat(mydata)
```

Data preparation and exploration

New variables create the variables:
2 new variables for day and time, rain(rain vs. no rain), trip duration(trip_short = trip $<$ 30min, trip_med = trip $>$ 30min && trip $<$ 60min), month from day

```
# 2 new variables for day and time
# Convert 'dep' column to date-time format
dat <- mutate(dat, dep = ymd_hms(dep))

# Create new columns for 'day' and 'time'
dat <- mutate(dat, day = as.Date(dep), time = format(dep, "%H:%M:%S"))

# Create a "rain vs no rain" variable
dat <- dat %>%
  mutate(rain = ifelse(prec > 0, "rain", "no rain"))

# Categorize trip duration into three categories: short, medium, and long
dat <- dat %>%
  mutate(trip_category = case_when(
    dur < 30 * 60 ~ "trip_short",       # Trip duration less than 30 minutes
    dur >= 30 * 60 & dur < 60 * 60 ~ "trip_med",   # Trip duration between 30 and 60 minutes
```

```
    dur >= 60 * 60 ~ "trip_long"        # Trip duration 60 minutes or more
  ))


# Extract the month from the "day" column
dat <- dat %>%
  mutate(month = month(day))  # Extract month and label with abbreviated month names

#factors
dat$mem <- as.factor(dat$mem)
dat$trip_category <- as.factor(dat$trip_category)
dat$rain <- as.factor(dat$rain)


# View the first few rows of the dataset to confirm the new variables
head(dat)
```

```
##                   dep  dur mem   wday temp prec rushhour wend rushmod
## 1 2021-05-28 07:56:26 2573   0 Friday  7.8    0        1    0       1
## 2 2021-05-28 07:04:16  602   0 Friday  7.8    0        1    0       1
## 3 2021-05-21 07:50:03 1131   0 Friday 25.3    0        1    0       1
## 4 2021-05-28 08:36:47  782   0 Friday  7.8    0        1    0       1
## 5 2021-05-07 08:50:12  690   0 Friday  9.7    0        1    0       1
## 6 2021-05-07 11:03:24  969   0 Friday  9.7    0        3    0       3
##          day     time    rain trip_category month
## 1 2021-05-28 07:56:26 no rain      trip_med     5
## 2 2021-05-28 07:04:16 no rain    trip_short     5
## 3 2021-05-21 07:50:03 no rain    trip_short     5
## 4 2021-05-28 08:36:47 no rain    trip_short     5
## 5 2021-05-07 08:50:12 no rain    trip_short     5
## 6 2021-05-07 11:03:24 no rain    trip_short     5
```

```
dat$trip_category <- as.factor(dat$trip_category)
levels(dat$trip_category)
```

```
## [1] "trip_long"  "trip_med"   "trip_short"
```

```
# Check the distribution of the rushmod variable
table(dat$rushmod, dat$wend)
```

Check the distribution of the rushmod variable

```
##
##       0   1
##   1 300   0
##   2 300   0
##   3 300 360
```

####Exploration

```
summary(dat)
```

```
##       dep                              dur         mem            wday
##  Min.   :2021-05-01 08:19:59.00   Min.   :  67.0   0:630   Friday   :180
##  1st Qu.:2021-06-15 11:30:01.75   1st Qu.: 448.0   1:630   Monday   :180
##  Median :2021-08-01 00:35:39.50   Median : 733.0           Saturday :180
```

```
##  Mean   :2021-08-01 03:44:54.00   Mean   : 953.1          Sunday   :180
##  3rd Qu.:2021-09-14 18:06:04.75   3rd Qu.:1188.2          Thursday :180
##  Max.   :2021-10-31 23:01:07.00   Max.   :6807.0          Tuesday  :180
##                                                           Wednesday:180
##       temp            prec        rushhour wend    rushmod      day
##  Min.   : 4.90   Min.   : 0.000   1:420    0:900   1:300   Min.   :2021-05-01
##  1st Qu.:15.60   1st Qu.: 0.000   2:420    1:360   2:300   1st Qu.:2021-06-15
##  Median :18.90   Median : 0.000   3:420            3:660   Median :2021-07-31
##  Mean   :18.47   Mean   : 1.486                            Mean   :2021-07-31
##  3rd Qu.:22.00   3rd Qu.: 0.300                            3rd Qu.:2021-09-14
##  Max.   :28.20   Max.   :37.000                            Max.   :2021-10-31
##
##      time                rain        trip_category      month
##  Length:1260        no rain:820   trip_long :  16   Min.   : 5.0
##  Class :character   rain   :440   trip_med  : 123   1st Qu.: 6.0
##  Mode  :character                 trip_short:1121   Median : 7.5
##                                                     Mean   : 7.5
##                                                     3rd Qu.: 9.0
##                                                     Max.   :10.0
##
```

We have as much as member than non-member riders

```
table(dat$mem)
```

```
##
##   0   1
## 630 630
```

```
table(dat$wday)
```

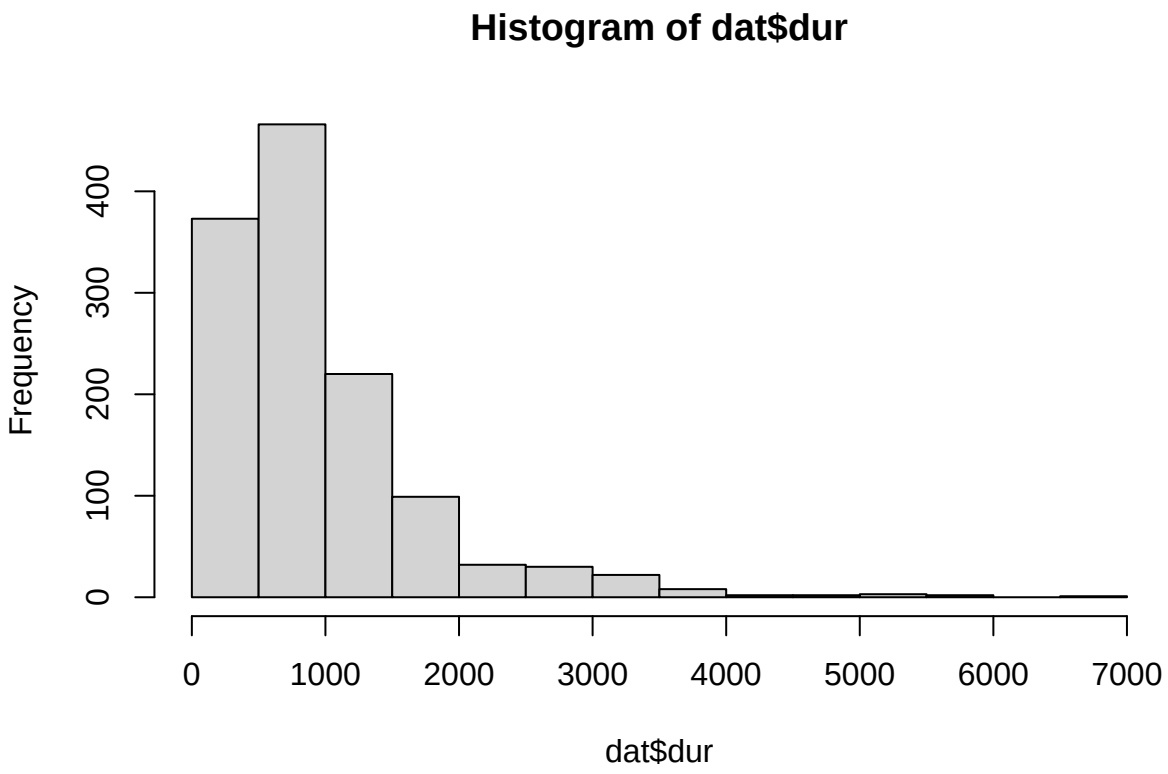
```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##       180       180       180       180       180       180       180
```

Graphs

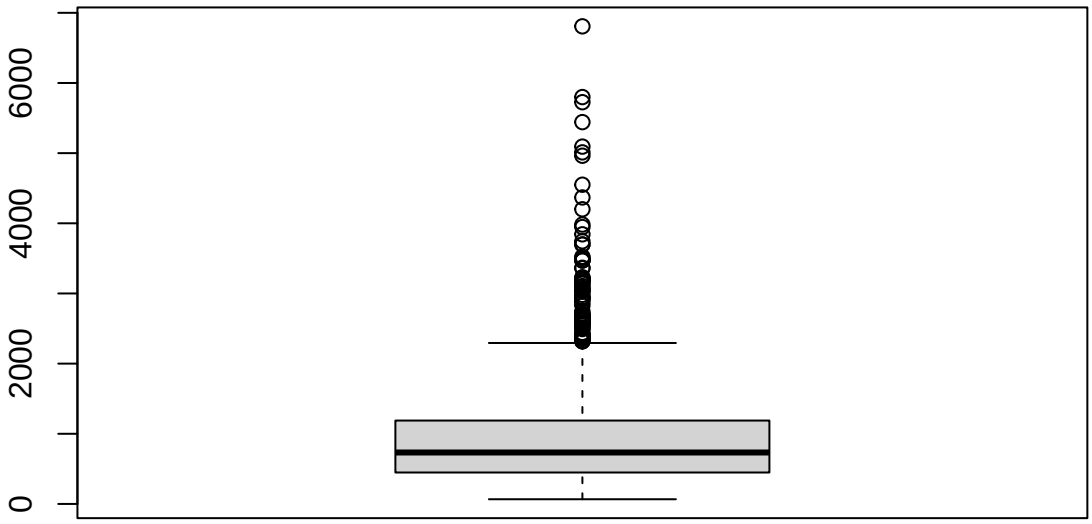
```
#Checking if any columsn have NA, nope! What a clean dataset.
colSums(is.na(dat))
```

```
##           dep           dur           mem          wday          temp
##             0             0             0             0             0
##          prec      rushhour          wend       rushmod           day
##             0             0             0             0             0
##          time          rain trip_category         month
##             0             0             0             0
```

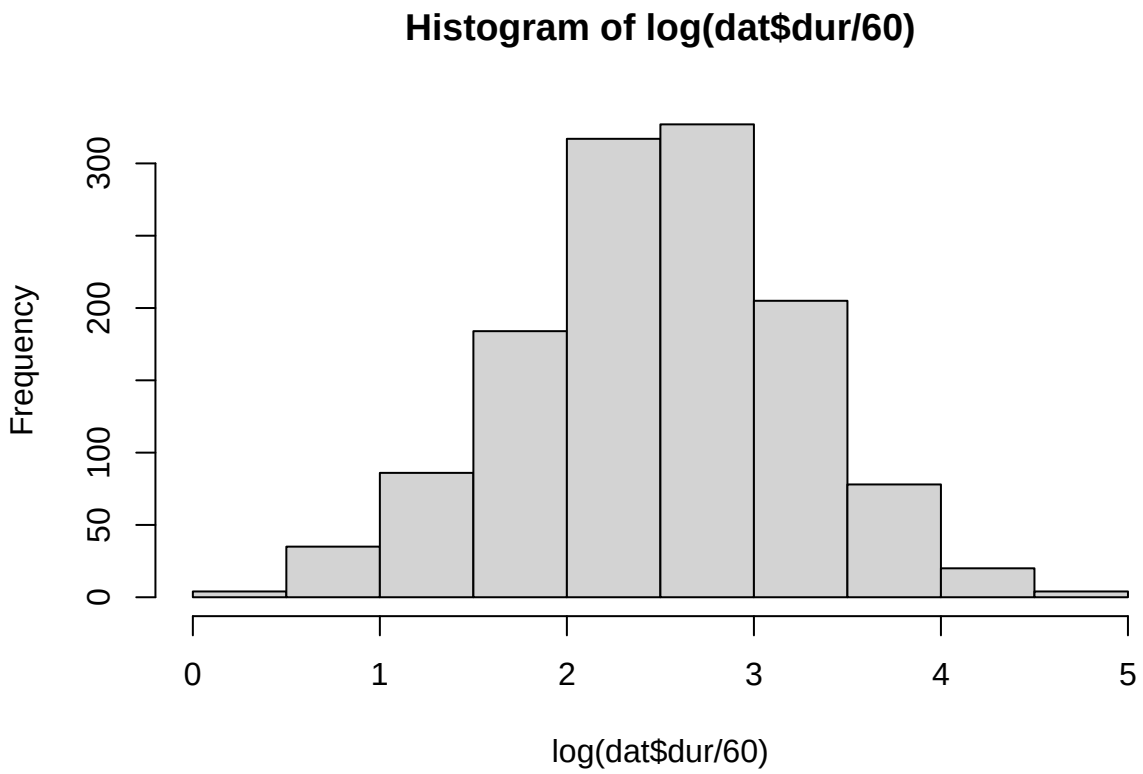
```
#Checking for data distribution
hist(dat$dur)
```



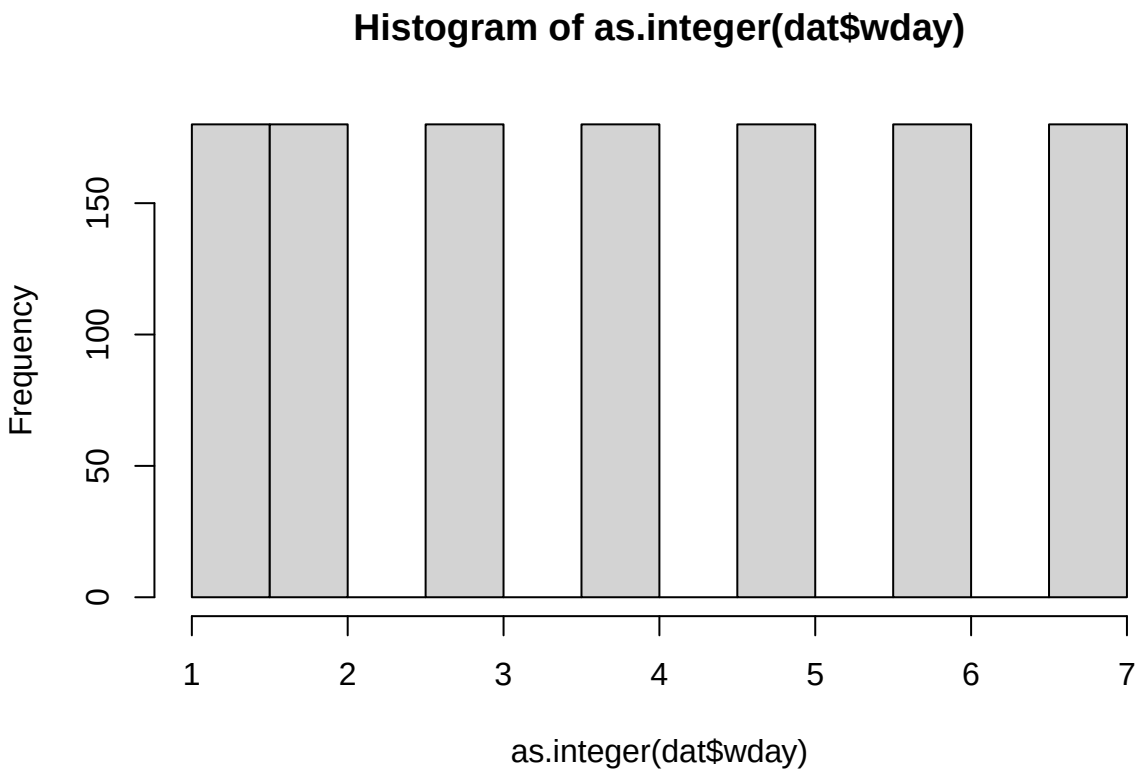

```
boxplot(dat$dur)
```



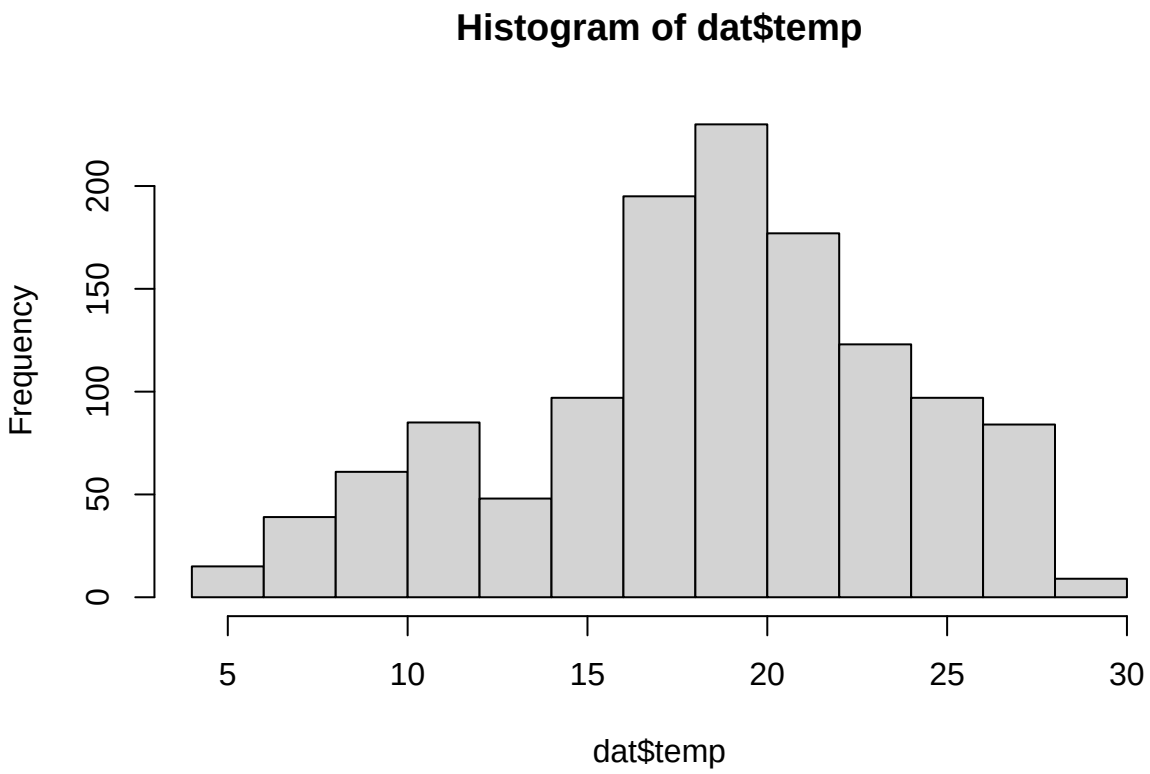

```
hist(log(dat$dur/60))          #seems like this normalizes the data nicely
```



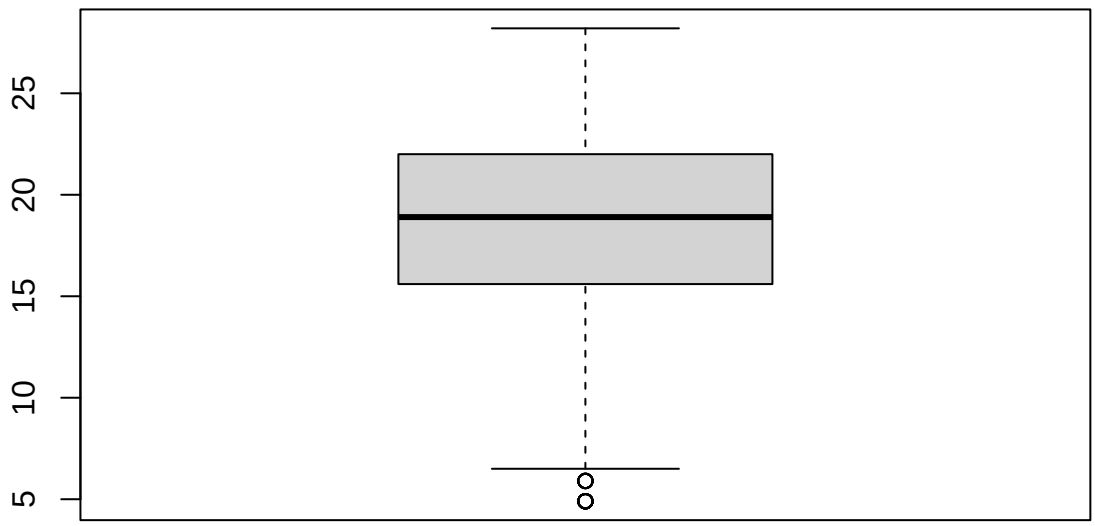

```r
hist(as.integer(dat$wday))    #all days appear equally represented
```



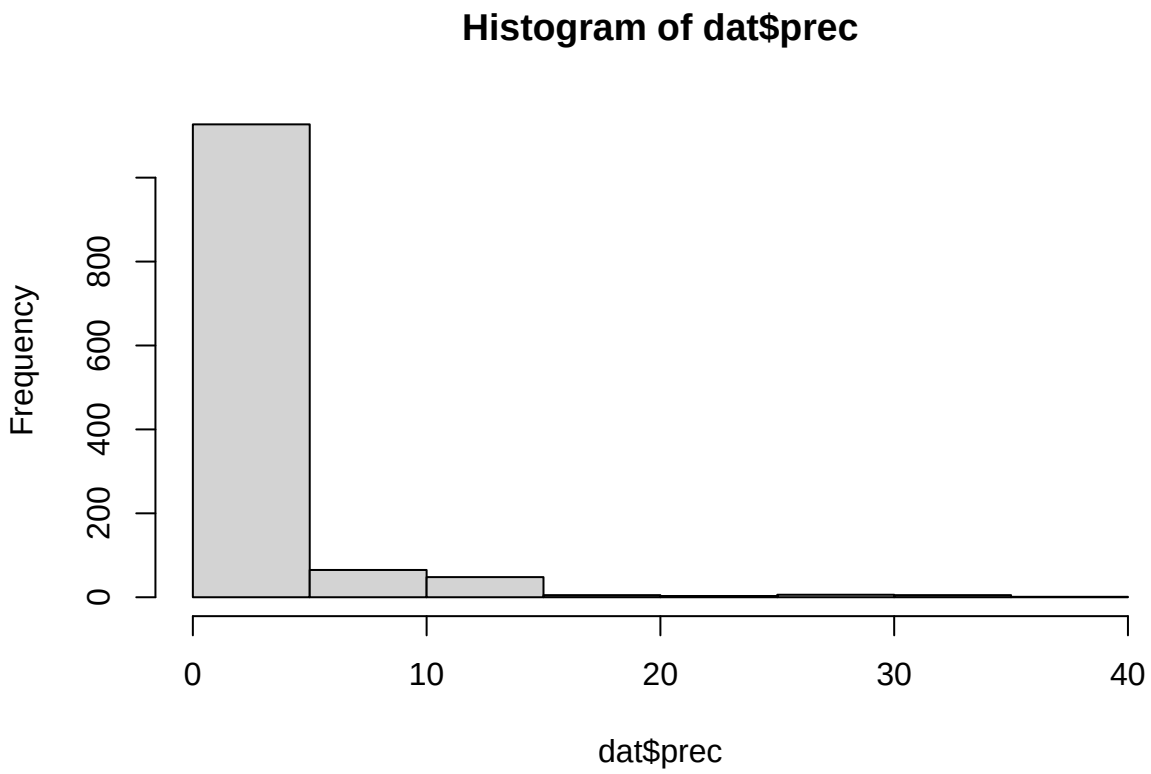

```r
hist(dat$temp)                    #temperature appears almost normally distributed
```




```
boxplot(dat$temp)
```




```r
hist(dat$prec)                #Rain appears highly non normal as well as having multiple 0 entries, sug
```




```
100*sum(dat$prec == 0)/length(dat$prec) #proportion or days without rain.
```

```
## [1] 65.07937
```

Data analysis

Statistic of trip duration for mem=1 and mem=0 We need to calculate the average trip duration for both members and non-members. Then, we'll adjust for weekend vs. non-weekend trips to see if the patterns hold

```
## # A tibble: 2 x 9
##   mem   count avg_duration median_duration sd_duration min_duration max_duration
##   <fct> <int>        <dbl>           <dbl>       <dbl>        <int>        <int>
## 1 0       630        1109.            828.        884.          114         6807
## 2 1       630         797.            639         635.           67         5442
## # i 2 more variables: Q1_duration <dbl>, Q3_duration <dbl>
```

Statistic on trip duration when we adjust for weekend vs. non-weekend usage (wend= 0 or 1)

```
## `summarise()` has grouped output by 'mem'. You can override using the `.groups`
## argument.
```

```
## [1] "Average trip duration for members vs non-members:"
```

```
## # A tibble: 2 x 2
##   mem   avg_duration
##   <fct>        <dbl>
## 1 0            1109.
```

```
## 2 1             797.
```

```
## [1] "Average trip duration for members vs non-members, separated by weekend/non-weekend:"
```

```
## # A tibble: 4 x 3
## # Groups:   mem [2]
##   mem   wend  avg_duration
##   <fct> <fct>        <dbl>
## 1 0     0            1091.
## 2 0     1            1154.
## 3 1     0             787.
## 4 1     1             823.
```

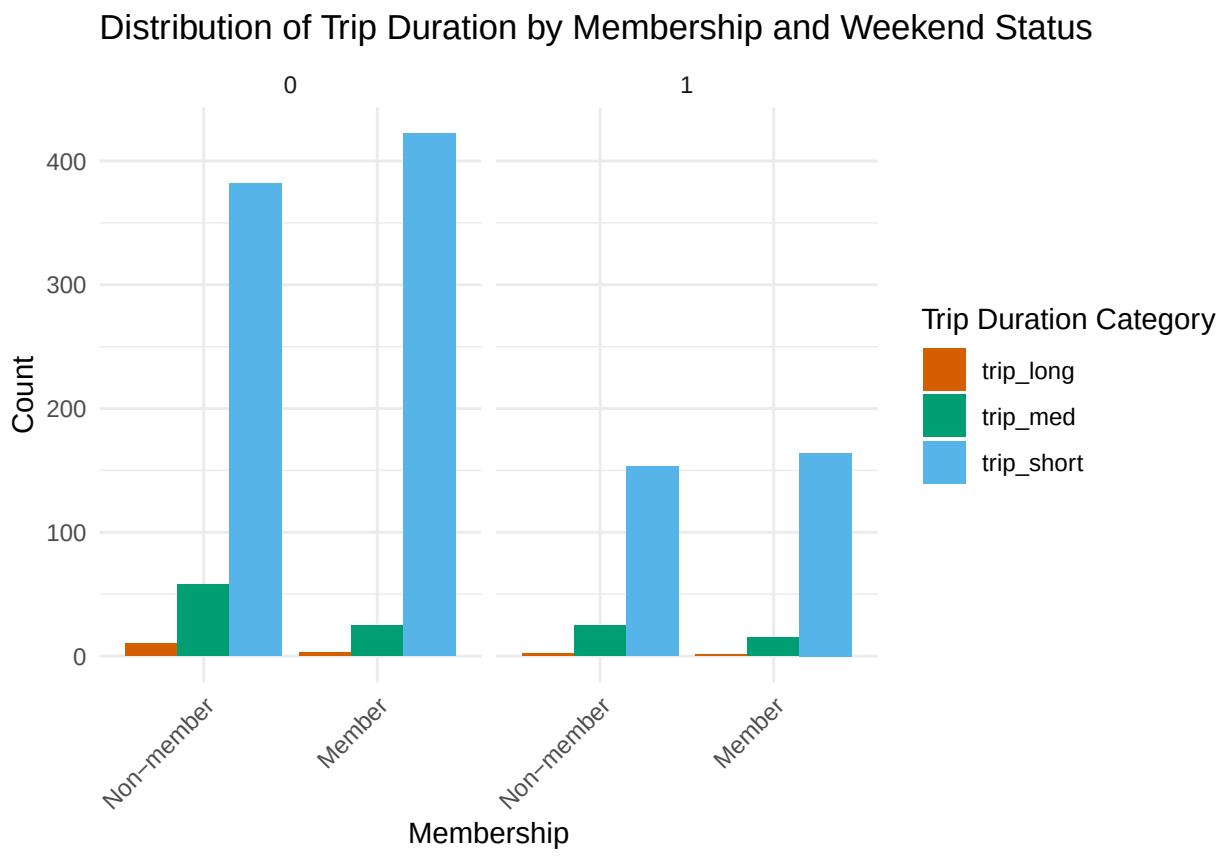
Average trip duration for non members on weekend (wend=1) is higher than non members on weekdays (1153.80 vs 1090.89) . The same remarks goes for members, they tend to bike more on weekends. We also notice that non members bike more than members. Non members also bike for a longer time than members. Even after checking for weekends and weekdays. We can do a way ANOVA to assess the mean difference of duration for members vs non-members accounting for wend.

Distribution of Duration with other variable

```
library(ggplot2)
```

```
## Warning: le package 'ggplot2' a été compilé avec la version R 4.3.3
```

```
# library(tidyr)
ggplot(dat, aes(x = factor(mem, labels = c("Non-member", "Member")), fill = trip_category)) +
  geom_bar(position = "dodge") +
  facet_wrap(~ wend) +  # Facet by weekend status (wend)
  labs(title = "Distribution of Trip Duration by Membership and Weekend Status",
       x = "Membership", y = "Count", fill = "Trip Duration Category") +
  scale_fill_manual(values = c("trip_short" = "#56B4E9", "trip_med" = "#009E73", "trip_long" = "#D55E00
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Two-Sample t-test (Comparing Two Groups: Members vs. Non-members)

To assess if BIXI members have shorter trips than non-members

```
# Perform a two-sample t-test

t_test_result <- t.test(dur ~ mem, data = dat)

# Print the result of the t-test
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  dur by mem
## t = 7.1851, df = 1142.1, p-value = 1.21e-12
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  226.4783 396.6328
## sample estimates:
## mean in group 0 mean in group 1
##       1108.8683        797.3127
```

Based on the very small p-value, we reject the null hypothesis and conclude that non-members take significantly longer trips on average compared to members. The difference in mean trip durations is statistically significant, with non-members taking trips that are 3.78 to 6.61 minutes longer on average than members. ##### when include wend

```r
# Model 1: Trip duration ~ membership (mem)
model1 <- lm(dur ~ mem, data = dat)

# Model 2: Trip duration ~ membership (mem) + weekend (wend) + interaction
model2 <- lm(dur ~ mem * wend, data = dat)

# Compare the two models using ANOVA
anova_comparison <- anova(model1, model2)

# Print the ANOVA comparison result
print(anova_comparison)
```

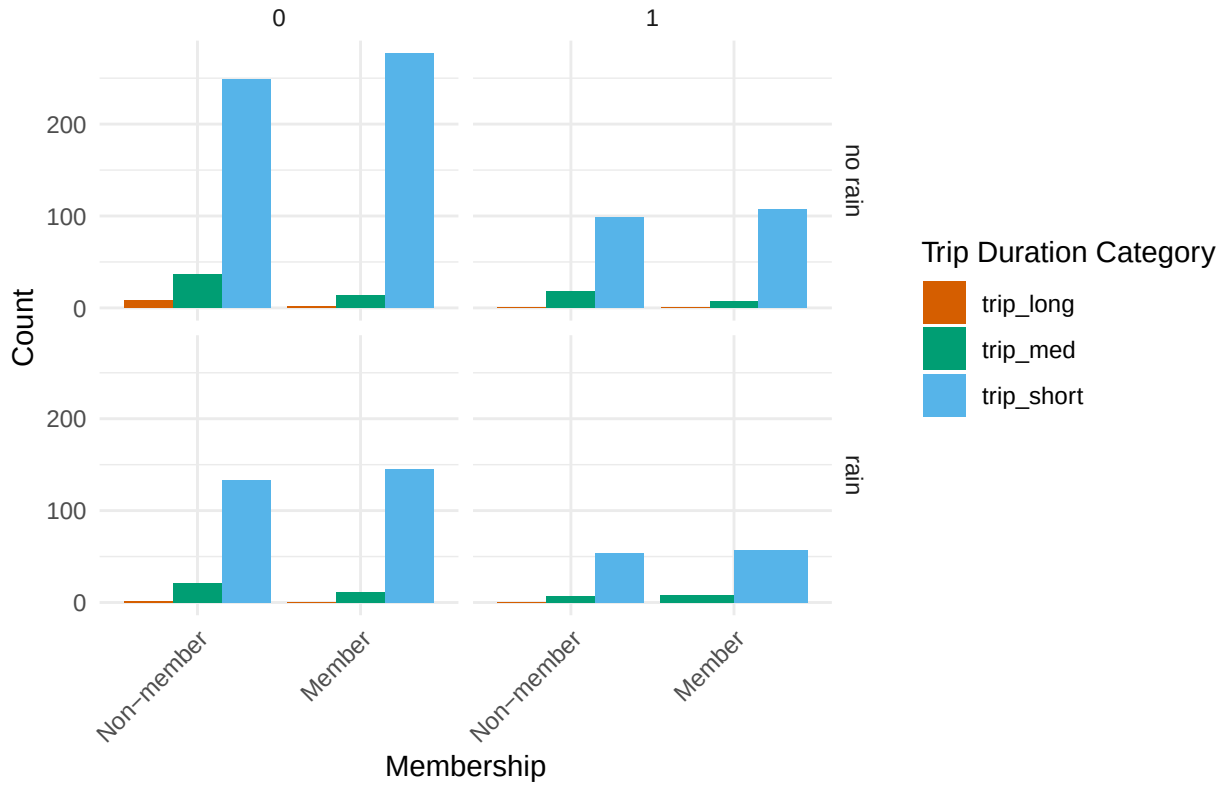
```
## Analysis of Variance Table
##
## Model 1: dur ~ mem
## Model 2: dur ~ mem * wend
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1   1258 745076851
## 2   1256 744404629  2    672222 0.5671 0.5673
```

Non-significant p-value (0.5673): The addition of the weekend status (wend) and the interaction between membership and weekend status (mem * wend) does not significantly improve the model's ability to explain trip duration. This means that, based on this analysis, weekend status (wend) does not have a significant effect on trip duration, nor does it significantly interact with membership (mem) to affect trip duration.

With weather (rain or no rain)

the distribution of trip duration for members vs. non-members, broken down by whether the trip happened on a weekend or weekday and whether it was raining.

```r
ggplot(dat, aes(x = factor(mem, labels = c("Non-member", "Member")), fill = trip_category)) +
  geom_bar(position = "dodge") +
  facet_grid(rain ~ wend) +  # Switch the facets (rain on rows, wend on columns)
  labs(title = "Distribution of Trip Duration by Membership, Rain, and Weekend Status",
       x = "Membership", y = "Count", fill = "Trip Duration Category") +
  scale_fill_manual(values = c("trip_short" = "#56B4E9", "trip_med" = "#009E73", "trip_long" = "#D55E00
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Model 3: Trip duration ~ membership (mem) + weekend (wend) + rain + interactions
model3 <- lm(dur ~ mem * wend * rain, data = dat)

# Perform the ANOVA to compare models
anova_comparison_rain <- anova(model2, model3)

# Print the ANOVA comparison result
print(anova_comparison_rain)
```

```
## Analysis of Variance Table
##
## Model 1: dur ~ mem * wend
## Model 2: dur ~ mem * wend * rain
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1   1256 744404629
## 2   1252 743050991  4   1353639 0.5702 0.6843
```

Non-significant p-value (0.6843): This means that adding rain and its interactions with membership and weekend status does not significantly improve the model's fit. Therefore, rain does not have a statistically significant impact on trip duration in this dataset. Based on this analysis, rain does not seem to be an important factor in explaining trip duration, and you can stick with the simpler Model 2 (which includes only membership and weekend status).

With rushmod

The linear model will show if there are significant differences in trip durations between:

`AM rush hour (rushmod = 1).`

PM rush hour (rushmod = 2).
Non-rush hour (rushmod = 3).

```
# Filter the data to only include weekday trips (wend == 0)
weekday_data <- dat %>% filter(wend == 0)
# Convert rushmod to a factor for modeling
weekday_data$rushmod <- as.factor(weekday_data$rushmod)
# Fit a linear model to compare trip durations based on rush hour categories
rush_hour_model <- lm(dur ~ rushmod, data = weekday_data)

# Print the summary of the model
summary(rush_hour_model)
```

```
##
## Call:
## lm(formula = dur ~ rushmod, data = weekday_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -920.2 -479.8 -206.4  217.5 4775.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   790.53      44.77  17.657  < 2e-16 ***
## rushmod2      212.71      63.32   3.360 0.000814 ***
## rushmod3      232.72      63.32   3.676 0.000251 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 775.5 on 897 degrees of freedom
## Multiple R-squared:  0.01817,    Adjusted R-squared:  0.01598
## F-statistic: 8.299 on 2 and 897 DF,  p-value: 0.0002685
```

The p-values for both rushmod2 (PM rush hour) and rushmod3 (non-rush hour) are highly significant ($p = 0.000814$ and $p = 0.000251$, respectively). This suggests that both PM rush hour and non-rush hour trip durations are significantly longer than AM rush hour trip durations.

```
# Perform Tukey HSD post-hoc test for pairwise comparisons
rushmod_comparison <- TukeyHSD(aov(rush_hour_model))

# Print the pairwise comparison results
print(rushmod_comparison)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = rush_hour_model)
##
## $rushmod
##          diff        lwr      upr     p adj
## 2-1 212.71333   64.07235 361.3543 0.0023439
## 3-1 232.72000   84.07902 381.3610 0.0007347
## 3-2  20.00667 -128.63431 168.6476 0.9464549
```

PM rush vs. AM rush (2-1):

```
Difference: 212.71 seconds longer in PM rush hour compared to AM rush hour.
```

```
p-value: 0.0023 (significant), meaning the trip durations during PM rush hour are significantly longer
```

Non-rush vs. AM rush (3-1):

```
Difference: 232.72 seconds longer in non-rush hour compared to AM rush hour.
p-value: 0.0007 (significant), meaning the trip durations during non-rush hours are significantly longe
```

Non-rush vs. PM rush (3-2):

```
Difference: 20.01 seconds longer in non-rush hour compared to PM rush hour.
p-value: 0.9465 (not significant), meaning there is no significant difference between PM rush hour and
```

Conclusion:

```
AM rush hour trips are significantly shorter than both PM rush hour and non-rush hour trips.
There is no significant difference between PM rush hour and non-rush hour trip durations.
```