Exploration

Lyft Data Exploration

Here we explore the data I've collected while driving for Lyft. To record this data, I have a Google Spreadsheet which I keep open in the background of my phone while I'm out driving. I record the start and end time of each ride, in addition to my odometer reading and gas usage (expressed in decimal gallons). It took a dozen rides or so to calibrate my process, but now I have a flow worked out so I only record the necessary information while I'm stopped.

Load Libraries

```
library(colorspace)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0

## -- Conflicts ---------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(boot)
```

Load Data and Format

I also have kept track of all the gasoline I've purchased (kept in a separate Google Spreadsheet). Here is the weighted (by gallons purchased) average price of gas.

```
gasPrice = 3.5006 # $ / gal
```

Next, we load the data. The difference between `data_all` and `data_drv` is that `data_all` includes *everything*, including all driving that takes place when I don't have any passengers (e.g. when I'm searching for a ride). For the most part, `data_drv` is of most importance.

```
data_all = read_csv("CleanLyftData_All.csv", col_types = cols())

data_all$DOW = factor(data_all$DOW, levels = seq(1,7), labels = c("Mon", "Tue", "Wed", "Thur", "Fri", "S
data_all$Month = factor(data_all$Month, levels = seq(1,12), labels = seq(1,12))
data_all$StartLocation = factor(data_all$StartLocation)
```

```r
data_all$Period = factor(data_all$Period)
data_all$Movement = factor(data_all$Movement)
data_all$Origin = factor(data_all$Origin)
data_all$Goal = factor(data_all$Goal)
bins = seq(6, 22, 2)
data_all$StartTimeBin = cut(data_all$StartTime, breaks = c(bins, 24), labels = bins)

data_drv = read_csv("CleanLyftData_Drives.csv", col_types = cols())

data_drv$DOW = factor(data_drv$DOW, levels = seq(1,7), labels = c("Mon", "Tue", "Wed", "Thur", "Fri", "S
data_drv$Month = factor(data_drv$Month, levels = seq(1,12), labels = seq(1,12))
data_drv$StartLocation = factor(data_drv$StartLocation)
data_drv$Period = factor(data_drv$Period)
data_drv$Movement = factor(data_drv$Movement)
data_drv$Origin = factor(data_drv$Origin)
data_drv$Goal = factor(data_drv$Goal)
data_drv$StartTimeBin = cut(data_drv$StartTime, breaks = c(bins, 24), labels = bins)


paste(nrow(data_drv), "drives")
```

```
## [1] "199 drives"
```

```r
data_all %>%
  filter(Session > max(data_drv$Session) - 3) %>%
  group_by(Session) %>%
  summarise(Wage = paste("$", round(sum(Earnings + Tips, na.rm = T) * (60 / sum(Duration)), 2), sep="")
            AdjWage = paste("$", round(sum(Earnings + Tips - GasUsage * gasPrice, na.rm = T) * (60 / su
            Revenue = sum(Earnings + Tips, na.rm = T),
            GasCost = round(sum(GasUsage) * gasPrice, 2)) %>%
  as.matrix()
```

```
##      Session Wage     AdjWage  Revenue GasCost
## [1,] "41"    "$18.31" "$16.06" "50.07" "9.31"
## [2,] "42"    "$17.46" "$14.79" "30.96" "7.21"
## [3,] "43"    "$26.15" "$22.88" "54.65" "8.61"
```

Totals

Let's view a handful of interesting summary statistics.

```r
c("Total Distance" = sum(data_drv$AdjDistance, na.rm = T),
  "Total Hours" = sum(data_drv$AdjDuration, na.rm = T) / 60,
  "Drives per Session" = nrow(data_drv) / max(data_drv$Session))
```

```
##     Total Distance        Total Hours Drives per Session
##        2138.650000          89.436833           4.627907
```

```r
c("Total Days" = as.integer(max(data_drv$Date) - min(data_drv$Date)),
  "Total Passengers" = sum(data_drv$Passengers),
  "Mean Hours per Week" = round((sum(data_drv$AdjDuration, na.rm = T) / 60) / (as.integer(max(data_drv$
  "Mean Drives per Week" = round(nrow(data_drv) / (as.integer(max(data_drv$Date) - min(data_drv$Date))
```

```
##          Total Days     Total Passengers  Mean Hours per Week
##              104.00               322.00                 6.02
```

```
## Mean Drives per Week
##                13.39
```

```r
c("Mean Revenue per Drive" = round(sum(data_drv$Earnings + data_drv$Tips) / nrow(data_drv), 2),
  "Mean Gas Expenditure per Drive" = round(sum(data_drv$GasUsage * gasPrice) / nrow(data_drv), 2),
  "Mean Gas Expenditure per Session" = round(sum(data_drv$GasUsage * gasPrice) / max(data_drv$Session),
```

```
##           Mean Revenue per Drive   Mean Gas Expenditure per Drive
##                             9.63                             1.22
## Mean Gas Expenditure per Session
##                             5.65
```

What's the median tip amount among those who tip?

```r
quantile(data_drv[data_drv$Tips > 0, ]$Tips, 0.5)
```

```
##   50%
## 2.965
```

Preview Data

```r
tail(data_drv)
```

```
## # A tibble: 6 x 36
##   Session Date                Period Movement Distance Duration Passengers
##     <dbl> <dttm>              <fct>  <fct>       <dbl>    <dbl>      <dbl>
## 1      42 2019-09-28 00:00:00 Drive  Drive        3.71    10.5           1
## 2      42 2019-09-28 00:00:00 Drive  Drive        1.07     6.33          4
## 3      43 2019-10-01 00:00:00 Drive  Drive        3.52    15.8           1
## 4      43 2019-10-01 00:00:00 Drive  Drive       13.5     23.3           2
## 5      43 2019-10-01 00:00:00 Drive  Drive        4.29    13.2           1
## 6      43 2019-10-01 00:00:00 Drive  Drive       23.6     34.1           1
## # ... with 29 more variables: Earnings <dbl>, Tips <dbl>, Shared <lgl>,
## #   TrueShared <dbl>, Conversation <lgl>, Origin <fct>, Goal <fct>,
## #   RatingConversation <dbl>, RatingRoute <dbl>,
## #   RatingComfortability <dbl>, DOW <fct>, Month <fct>, StartTime <dbl>,
## #   PickupTime <dbl>, EndTime <dbl>, TimeLabel <chr>, StartLocation <fct>,
## #   StartGas <dbl>, EndGas <dbl>, GasUsage <dbl>, Wage <dbl>,
## #   RatingSum <dbl>, RatingMean <dbl>, AdjDuration <dbl>,
## #   AdjDistance <dbl>, AdjGas <dbl>, Position <chr>, AdjWage <dbl>,
## #   StartTimeBin <fct>
```

Visualizations

Just a short cut for style.

```r
theme = theme_minimal()
```

Wage Distribution

```r
maint = (600 + 30 + 100 + 300) / 5000 # depreciation + oil + service + parts per the next 5,000 miles
```

Next we'll create a few measures of earnings, becoming more specific in terms of what's included as the list goes down. The last measure is likely the most realistic of what I earn. 95% boostrapped Confident Intervals are listed as well.

```
b = boot((data_drv$Earnings + data_drv$Tips) * (60/data_drv$Duration), function (v, ix) mean(v[ix]), R =
ci = boot.ci(b, type = "perc")$perc[1, ]
sprintf("$%0.2f (95%% CI: $%0.2f - $%0.2f)", b$t0, ci[4], ci[5])
```

```
## [1] "$36.28 (95% CI: $34.58 - $38.08)"
```

```
b = boot(data_drv$Earnings * (60/data_drv$Duration), function (v, ix) mean(v[ix]), R = 5000)
ci = boot.ci(b, type = "perc")$perc[1, ]
sprintf("$%0.2f (95%% CI: $%0.2f - $%0.2f)", b$t0, ci[4], ci[5])
```

```
## [1] "$32.30 (95% CI: $30.84 - $33.90)"
```

```
b = boot((data_drv$Earnings + data_drv$Tips) * (60/data_drv$AdjDuration), function (v, ix) mean(v[ix]),
ci = boot.ci(b, type = "perc")$perc[1, ]
sprintf("$%0.2f (95%% CI: $%0.2f - $%0.2f)", b$t0, ci[4], ci[5])
```

```
## [1] "$23.13 (95% CI: $21.75 - $24.55)"
```

```
b = boot((data_drv$Earnings + data_drv$Tips - data_drv$AdjGas * gasPrice) * (60/data_drv$AdjDuration),
ci = boot.ci(b, type = "perc")$perc[1, ]
sprintf("$%0.2f (95%% CI: $%0.2f - $%0.2f)", b$t0, ci[4], ci[5])
```

```
## [1] "$19.16 (95% CI: $17.85 - $20.53)"
```

```
b = boot((data_drv$Earnings + data_drv$Tips - data_drv$AdjGas * gasPrice - maint * data_drv$AdjDistance
ci = boot.ci(b, type = "perc")$perc[1, ]
sprintf("$%0.2f (95%% CI: $%0.2f - $%0.2f)", b$t0, ci[4], ci[5])
```

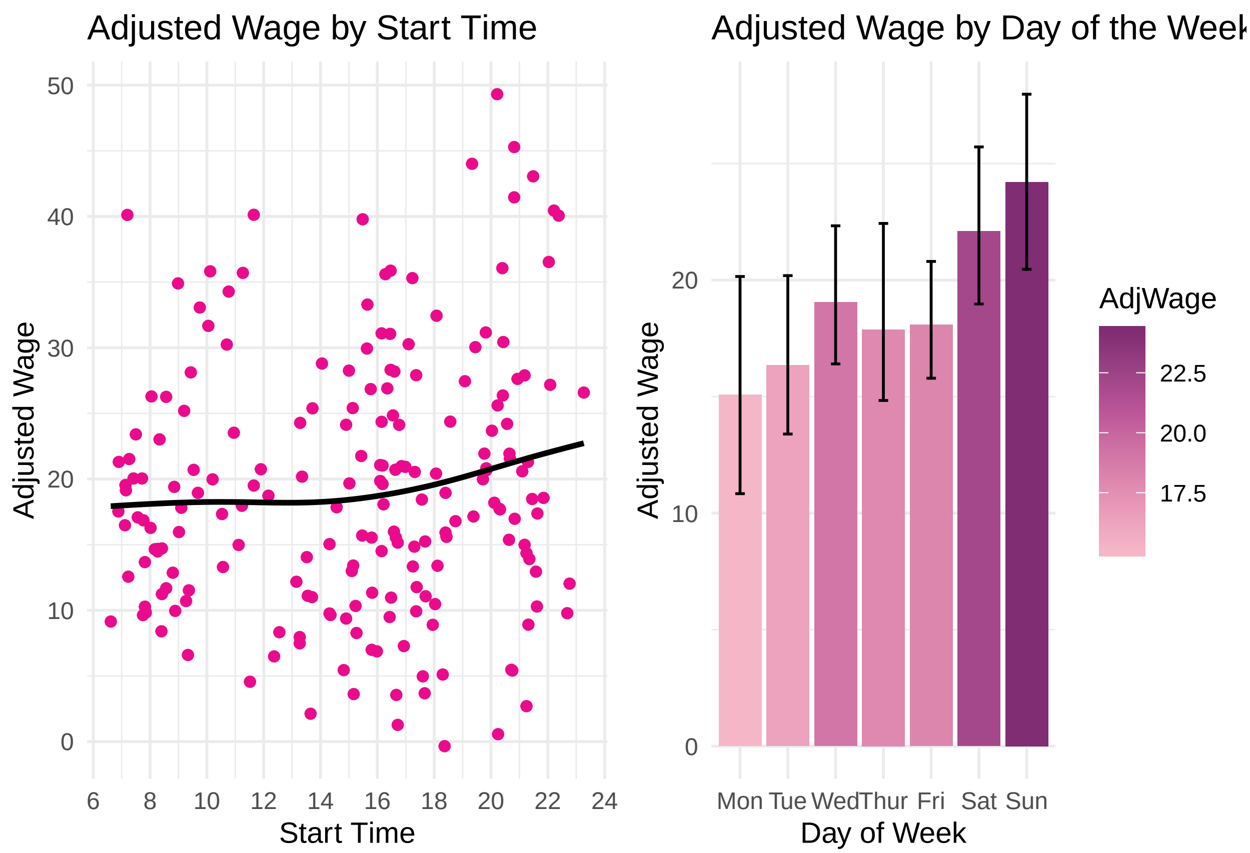
```
## [1] "$14.72 (95% CI: $13.40 - $16.00)"
```

Driving Days and Times

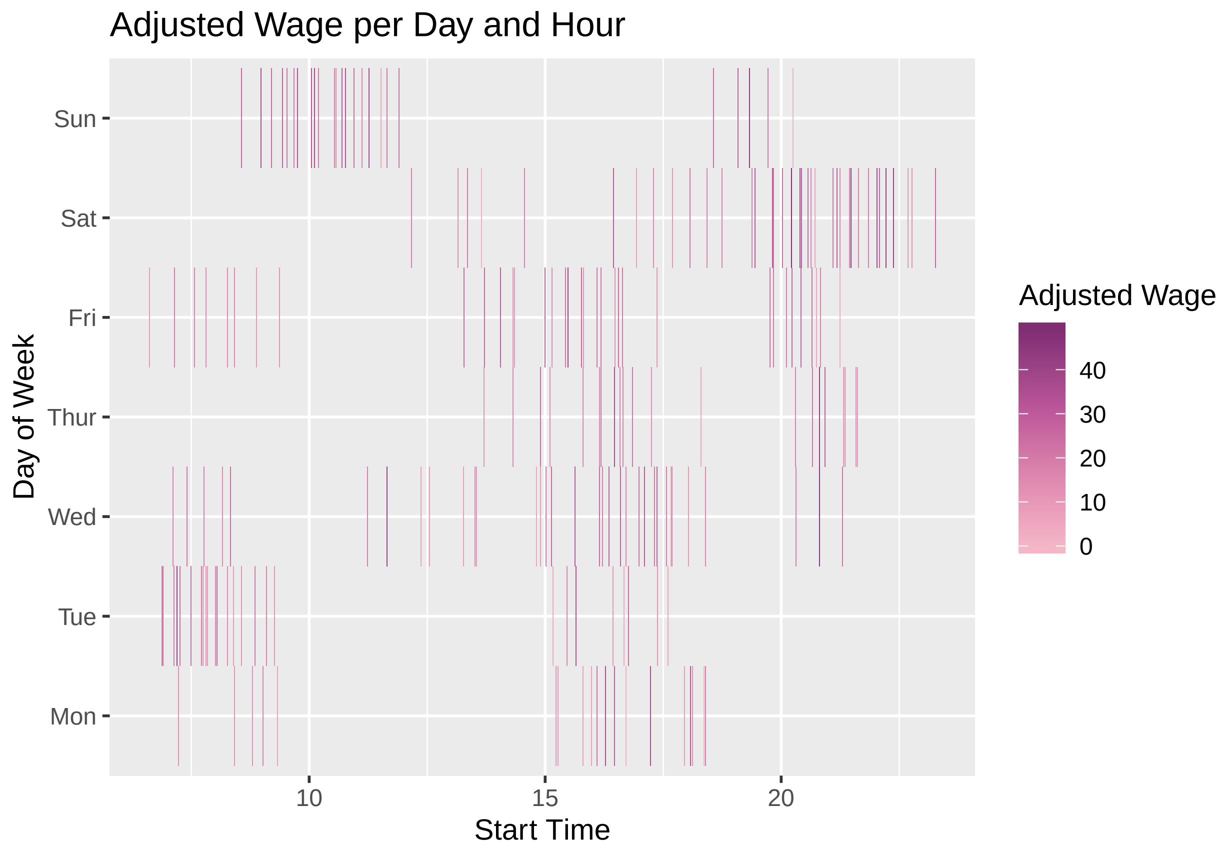
How do earnings correlate with time and day of the ride?

```
p1 = ggplot(data_drv, aes(StartTime, AdjWage)) + theme +
  geom_point(color = "#EA0B8C") +
  geom_smooth(method = mgcv::gam, formula = y ~ s(x, bs = "gp", k = 10), se = F, color = "black") +
  labs(title = "Adjusted Wage by Start Time", x = "Start Time", y = "Adjusted Wage") +
  scale_x_continuous(breaks = seq(4, 24, 2))

res = data.frame()

for (dow in levels(data_drv$DOW)) {

  f = function(data, indices) {
    return(mean(data[indices]))
  }

  d = data_drv[data_drv$DOW == dow, ]
  b = boot(d$AdjWage, statistic = f, R = 1000)

  r = boot.ci(b, type = "bca")$bca
```

```r
  res[dow, c("lb", "AdjWage", "ub")] = list(r[1, 4], b$t0, r[1, 5])
}

p2 = rownames_to_column(res, "DOW") %>%
  ggplot() + theme +
  geom_col(aes(DOW, AdjWage, fill = AdjWage)) +
  geom_errorbar(aes(DOW, ymin = lb, ymax = ub, width=0.2)) +
  labs(title = "Adjusted Wage by Day of the Week", y = "Adjusted Wage", x = "Day of Week") +
  scale_x_discrete(limits = levels(data_drv$DOW)) +
  scale_fill_continuous_sequential(palette = "Magenta", begin = 0.1, end = 0.9, rev = T)

p = grid.arrange(p1, p2, ncol=2)
```




```r
ggsave("StartTime_DOW.png", p, width = 10, height = 6)

ggplot(data_drv) +
  geom_tile(aes(x = StartTime, y = DOW, fill = AdjWage)) +
  labs(x = "Start Time", y = "Day of Week", fill = "Adjusted Wage", title = "Adjusted Wage per Day and I
  scale_fill_continuous_sequential(palette = "Magenta", begin = 0.1, end = 0.9, rev = T)
```



Effect of day on wage?

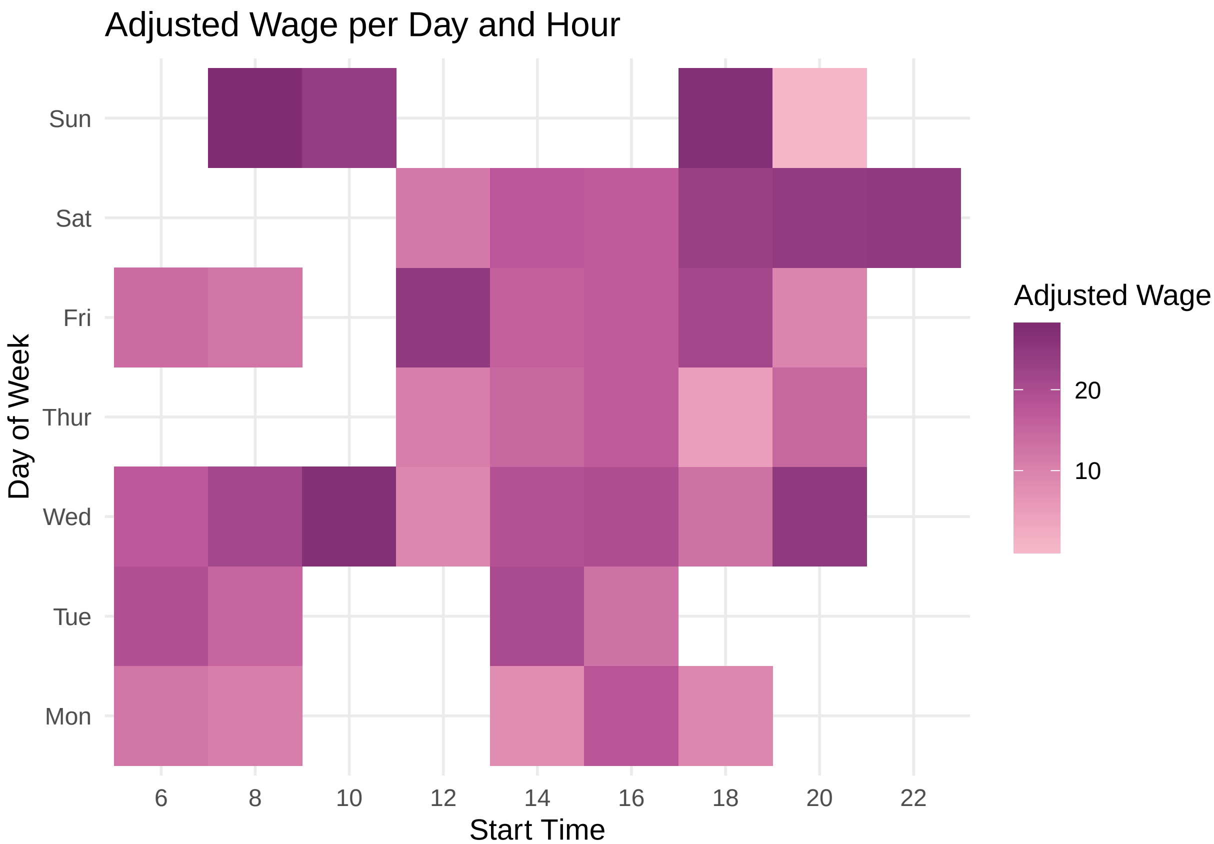
```
summary(aov(AdjWage ~ DOW, data = data_drv))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## DOW           6   1558  259.69    2.85 0.0111 *
## Residuals   192  17497   91.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Average Wage given Days and Times

Look at the big picture of day and time on wage.

```
p = data_drv %>%
  group_by(StartTimeBin, DOW) %>%
  summarise(m = sum(Earnings + Tips - gasPrice * AdjGas) * (60 / sum(AdjDuration))) %>%
  ggplot() + theme +
  geom_tile(aes(x = StartTimeBin, y = DOW, fill = m)) +
  labs(x = "Start Time", y = "Day of Week", fill = "Adjusted Wage", title = "Adjusted Wage per Day and I
  scale_fill_continuous_sequential(palette = "Magenta", begin = 0.1, end = 0.9, rev = T)

p
```




```r
ggsave("WageDayandTime.png", p, width = 6, height = 4)
```

Duration on Wage

Does Lyft pay what it claims to? Do longer rides pay more or less than shorter rides?

```r
summary(MASS::rlm(Earnings ~ Distance + Duration, data = d))
```

```
##
## Call: rlm(formula = Earnings ~ Distance + Duration, data = d)
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0082082 -0.0037484  0.0008961  0.0037981  4.9969493
##
## Coefficients:
##             Value     Std. Error t value
## (Intercept)    0.0062    0.0038      1.6398
## Distance       0.6527    0.0003   1878.0480
## Duration       0.2246    0.0004    612.9634
##
## Residual standard error: 0.005603 on 22 degrees of freedom
```

```r
summary(lm(Earnings ~ Distance + Duration, data = d))
```

```
##
## Call:
## lm(formula = Earnings ~ Distance + Duration, data = d)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4066 -0.3514 -0.2658 -0.1477  4.6547
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.52950    0.71770   0.738  0.46845
## Distance     0.64045    0.06649   9.632 2.38e-09 ***
## Duration     0.21493    0.07008   3.067  0.00564 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.052 on 22 degrees of freedom
## Multiple R-squared:  0.9793, Adjusted R-squared:  0.9774
## F-statistic: 520.3 on 2 and 22 DF,  p-value: < 2.2e-16
```

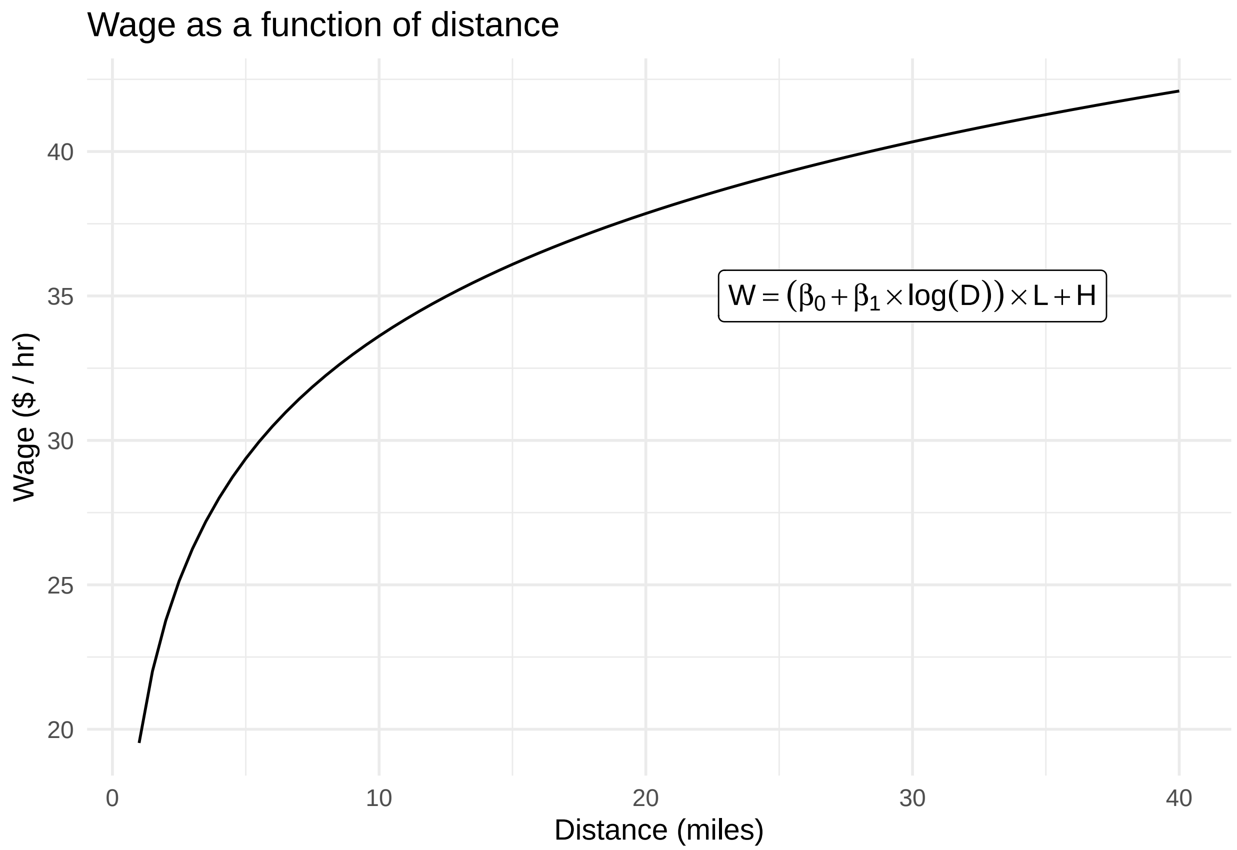
```r
miles = 0.6525
hours = 0.225 * 60
```

```r
data_drv$Speed = data_drv$Distance / (data_drv$Duration/60)
```

```r
summary(lm(Speed ~ log(Distance), data = data_drv))
```

```
##
## Call:
## lm(formula = Speed ~ log(Distance), data = data_drv)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2228  -4.7416  -0.4905   4.5013  20.1800
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.2341     0.9399   9.825   <2e-16 ***
## log(Distance)   9.3768     0.5176  18.117   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.016 on 197 degrees of freedom
## Multiple R-squared:  0.6249, Adjusted R-squared:  0.623
## F-statistic: 328.2 on 1 and 197 DF,  p-value: < 2.2e-16
```

```r
D = seq(1, 40, 0.5)

W = (9.234 + 9.377 * log(D)) * 0.6525 + 13.5

p = ggplot() + theme +
  geom_line(aes(D, W)) +
  geom_label(aes(x = 30, y = 35, label = "W == (beta[0] + beta[1] %*% log(D)) %*% L + H"), parse = T) +
  labs(x = "Distance (miles)", y = "Wage ($ / hr)", title = "Wage as a function of distance")

p
```




```r
ggsave("WageDistance.png", p, width = 5, height = 3)
```

Time Labels

How are my artificial time labels correlated with wage? For the rule set, see the `CleanData.R`.

```r
table(data_drv$TimeLabel)
```

```
##
## AfternoonCommute   EveningCommute   MorningCommute        Nightlife
##                6               20               26               25
##            Other          Tourism
##              116                6
```

Do nightlife rides pay more than non-nightlife rides, as some online articles claim?

```r
t.test(data_drv[data_drv$TimeLabel == "Nightlife", ]$AdjWage, data_drv[data_drv$TimeLabel != "Nightlife
```

```
##
##  Welch Two Sample t-test
##
## data:  data_drv[data_drv$TimeLabel == "Nightlife", ]$AdjWage and data_drv[data_drv$TimeLabel != "Nig
## t = 2.3751, df = 28.553, p-value = 0.02448
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.8025558 10.8019383
## sample estimates:
## mean of x mean of y
```

```
## 24.33217 18.52992
```

Are there differences between commutes?

```r
d = data_drv[data_drv$TimeLabel %in% c("MorningCommute", "AfternoonCommute", "EveningCommute"), ]
anova(lm(AdjWage ~ TimeLabel, data = d))
```

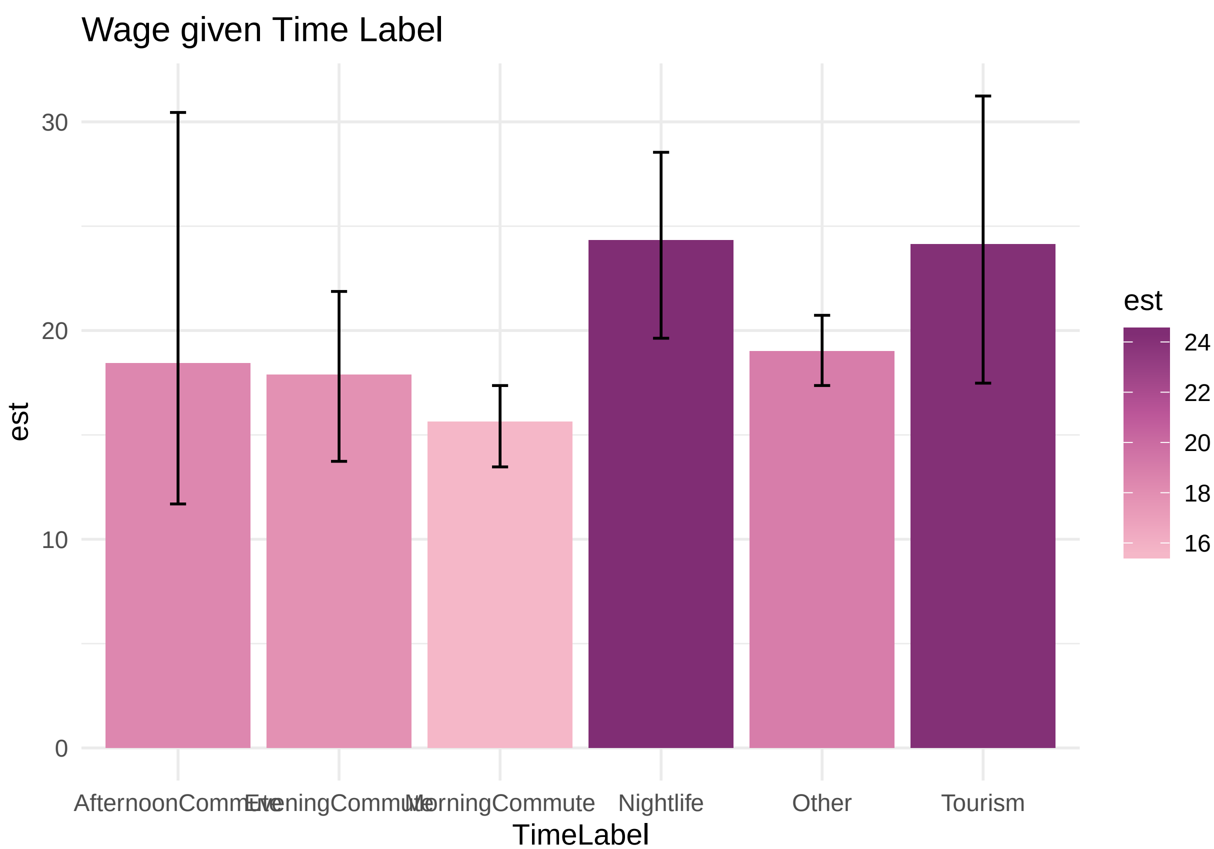
```
## Analysis of Variance Table
##
## Response: AdjWage
##           Df  Sum Sq Mean Sq F value Pr(>F)
## TimeLabel  2   75.47  37.735  0.5916 0.5573
## Residuals 49 3125.24  63.780
```

```r
TukeyHSD(aov(lm(AdjWage ~ TimeLabel, data = d)))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(AdjWage ~ TimeLabel, data = d))
##
## $TimeLabel
##                                       diff        lwr      upr     p adj
## EveningCommute-AfternoonCommute -0.5523929  -9.537072 8.432286 0.9879035
## MorningCommute-AfternoonCommute -2.8117570 -11.553926 5.930412 0.7185731
## MorningCommute-EveningCommute   -2.2593641  -8.000314 3.481586 0.6108740
```

View all the labels are their wages.

```r
res = data.frame()

for (tl in unique(data_drv$TimeLabel)) {

  f = function(data, indices) {
    return(mean(data[indices]))
  }

  b = boot(data_drv[data_drv$TimeLabel == tl, ]$AdjWage, statistic = f, R = 1000)

  r = boot.ci(b, type = "bca")$bca

  res[tl, c("lb", "est", "ub")] = list(r[1, 4], b$t0, r[1, 5])
}

rownames_to_column(res, "TimeLabel") %>%
  ggplot() + theme +
  geom_bar(aes(TimeLabel, est, fill = est), stat="identity") +
  geom_errorbar(aes(TimeLabel, ymin = lb, ymax = ub, width=0.1)) +
  labs(title = "Wage given Time Label") +
  scale_fill_continuous_sequential(palette = "Magenta", begin = 0.1, end = 0.9, rev = T)
```



Drive Position

Do the first and last drives of a session differ in wage from the rides that come in-between?

```
res = data.frame()

for (pos in unique(data_drv$Position)) {

  f = function(data, indices) {
    return(mean(data[indices]))
  }

  b = boot(data_drv[data_drv$Position == pos, ]$AdjWage, statistic = f, R = 1000)

  r = boot.ci(b, type = "bca")$bca

  res[pos, c("lb", "AdjWage", "ub")] = list(r[1, 4], b$t0, r[1, 5])
}

p = rownames_to_column(res, "Position") %>%
  ggplot() + theme +
  geom_col(aes(Position, AdjWage, fill = AdjWage)) +
  geom_errorbar(aes(Position, ymin = lb, ymax = ub, width=0.1)) +
  labs(title = "Adjusted Wage given Drive Position", y = "Adjusted Wage") +
  scale_x_discrete(limits = c("First", "Middle", "Last")) +
  scale_fill_continuous_sequential(palette = "Magenta", begin = 0.1, end = 0.9, rev = T)
```

```
p
```


```
ggsave("WagePosition.png", p, width = 6, height = 4)
```

It appears as though the last ride is likely to earn less than the other two types of rides.

```
summary(aov(AdjWage ~ Position, data = data_drv))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Position      2   3388  1694.2    21.2 4.64e-09 ***
## Residuals   196  15667    79.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

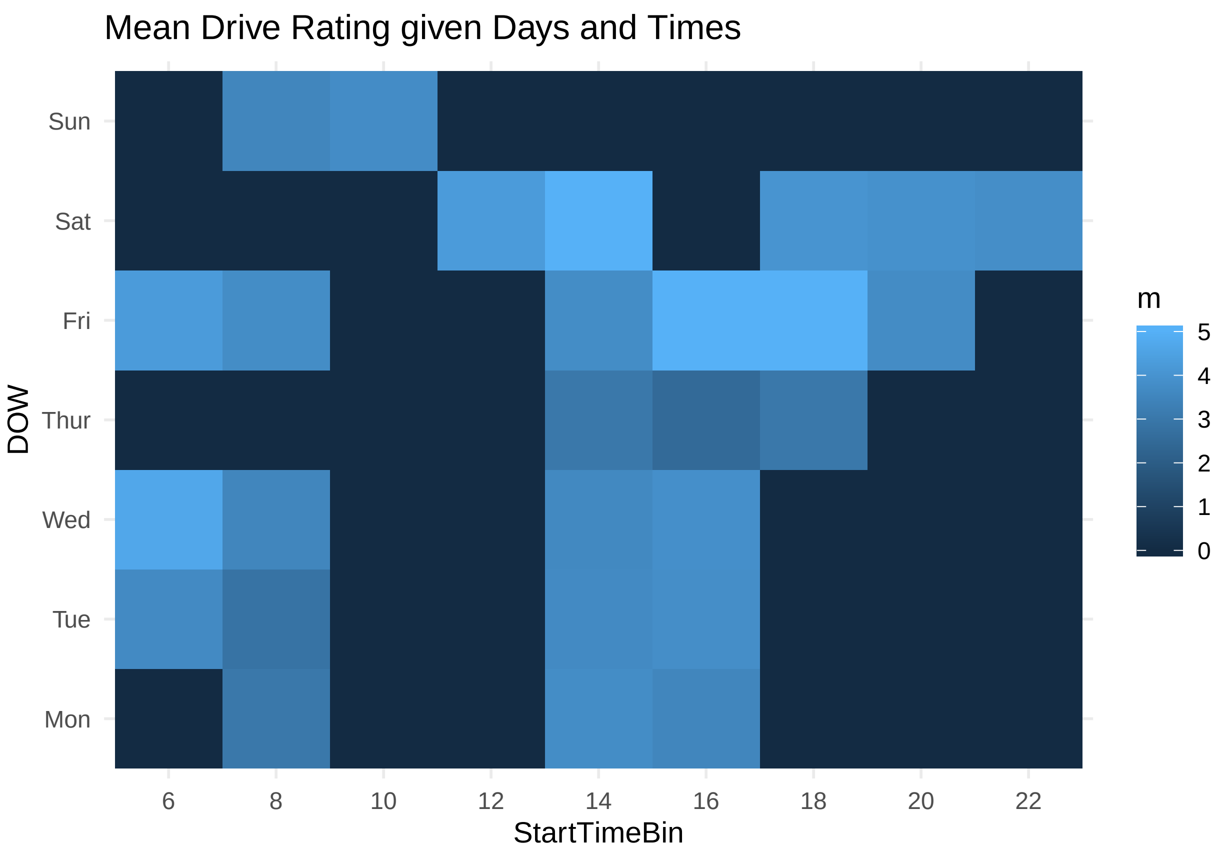
```
TukeyHSD(aov(AdjWage ~ Position, data = data_drv))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = AdjWage ~ Position, data = data_drv)
##
## $Position
##                   diff       lwr        upr     p adj
## Last-First   -4.104247 -8.657857  0.4493634 0.0867201
## Middle-First  5.819957  2.036716  9.6031971 0.0010383
## Middle-Last   9.924204  6.140963 13.7074441 0.0000000
```

Drive Ratings

Only recently have I started to log satisfaction of the drive. More on this to come.

```
data_drv %>%
  group_by(StartTimeBin, DOW) %>%
  summarise(m = mean(RatingConversation, na.rm = T)) %>%
  complete(DOW, fill = list(m = 0)) %>%
  ggplot() + theme +
  geom_tile(aes(x = StartTimeBin, y = DOW, fill = m)) +
  labs(title = "Mean Drive Rating given Days and Times")
```



Start Location

What are the top 3 starting locations? How many starting locations have only been visited once? How many total are there?

```
SL.counts = data_drv %>%
  group_by(StartLocation) %>%
  summarise(n = n()) %>%
  arrange(desc(n))

head(SL.counts, 3)
```

```
## # A tibble: 3 x 2
##   StartLocation     n
##   <fct>         <int>
```

```
## 1 UTC                24
## 2 East Village       12
## 3 Pacific Beach      10
```

```
sum(SL.counts$n == 1)
```

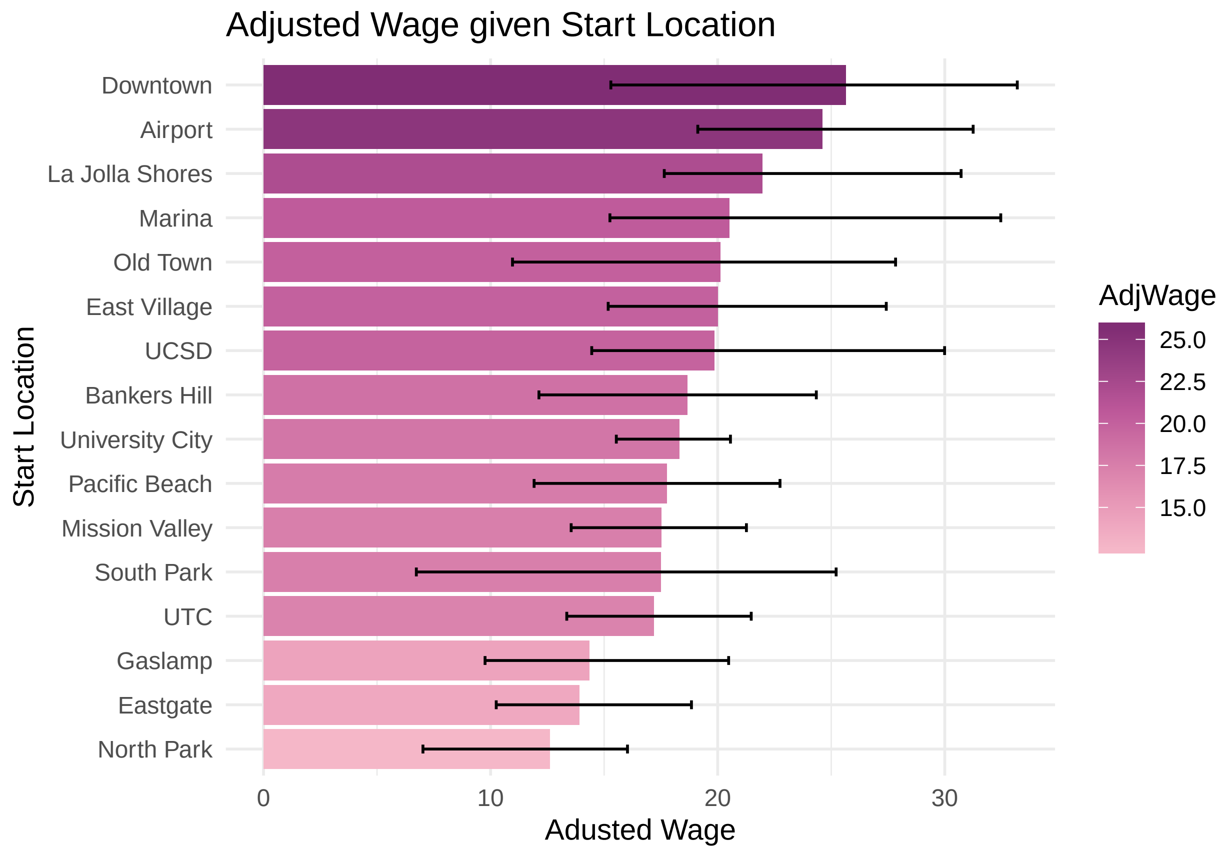
```
## [1] 19
```

```
nrow(SL.counts)
```

```
## [1] 51
```

How does where the ride is started correlate with how much I'll earn?

```
res = data.frame()

for (loc in unique(data_drv$StartLocation)) {

  f = function(data, indices) {
    return(mean(data[indices]))
  }

  d = data_drv[data_drv$StartLocation == loc, ]$AdjWage

  if (length(d) >= 5) {
    b = boot(data_drv[data_drv$StartLocation == loc, ]$AdjWage, statistic = f, R = 1000)

    r = boot.ci(b, type = "bca")$bca

    res[loc, c("lb", "AdjWage", "ub")] = list(r[1, 4], b$t0, r[1, 5])
  }
}

d = rownames_to_column(res, "StartLocation")

p = d %>%
  ggplot() + theme +
  geom_col(aes(StartLocation, AdjWage, fill = AdjWage)) +
  geom_errorbar(aes(StartLocation, ymin = lb, ymax = ub, width=0.2)) +
  scale_x_discrete(limits = d[order(d$AdjWage), "StartLocation"]) +
  scale_fill_continuous_sequential(palette = "Magenta", begin = 0.1, end = 0.9, rev = T) +
  coord_flip() +
  labs(title = "Adjusted Wage given Start Location", x = "Start Location", y = "Adusted Wage")

p
```




```r
ggsave("StartLocationWage.png", p, width = 10, height = 6)
```

Wait Times

First, filter a new dataset to only wait times between rides.

```r
wd = data_all %>% filter(Period == "Search" & Movement != "Home")
```

```r
library(mgcv)
```

```
## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.8-29. For overview type 'help("mgcv-package")'.
```

Day of the week does not seem to be correlated with whether I'll have to wait between rides.

```r
summary(aov(glm(I(Duration == 0) ~ DOW, data = wd, family = "binomial")))
```

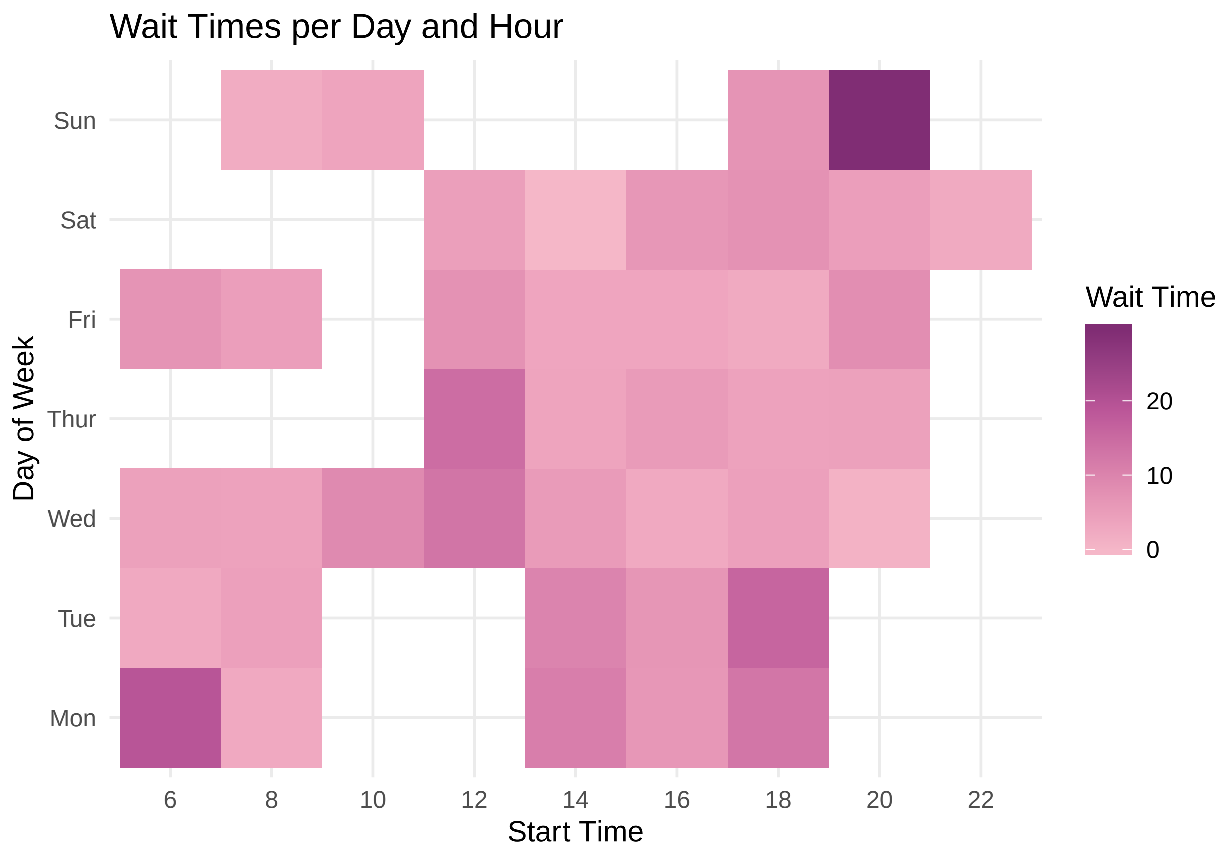
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## DOW           6   1.17   0.195   1.204  0.305
## Residuals   219  35.47   0.162
```

Day of the week does not seem to be correlated with how much I'll have to wait between rides.

```r
summary(aov(glm(Duration ~ DOW, data = wd[wd$Duration > 0, ], family = Gamma(link = "log"))))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## DOW           6    587   97.79   1.719  0.119
## Residuals   173   9844   56.90
```

```r
p = data_all %>%
  filter(Period == "Search" & Movement != "Home") %>%
  group_by(StartTimeBin, DOW) %>%
  summarise(m = mean(Duration)) %>%
  ggplot() + theme +
  geom_tile(aes(x = StartTimeBin, y = DOW, fill = m)) +
  labs(x = "Start Time", y = "Day of Week", fill = "Wait Time", title = "Wait Times per Day and Hour") +
  scale_fill_continuous_sequential(palette = "Magenta", begin = 0.1, end = 0.9, rev = T)

p
```

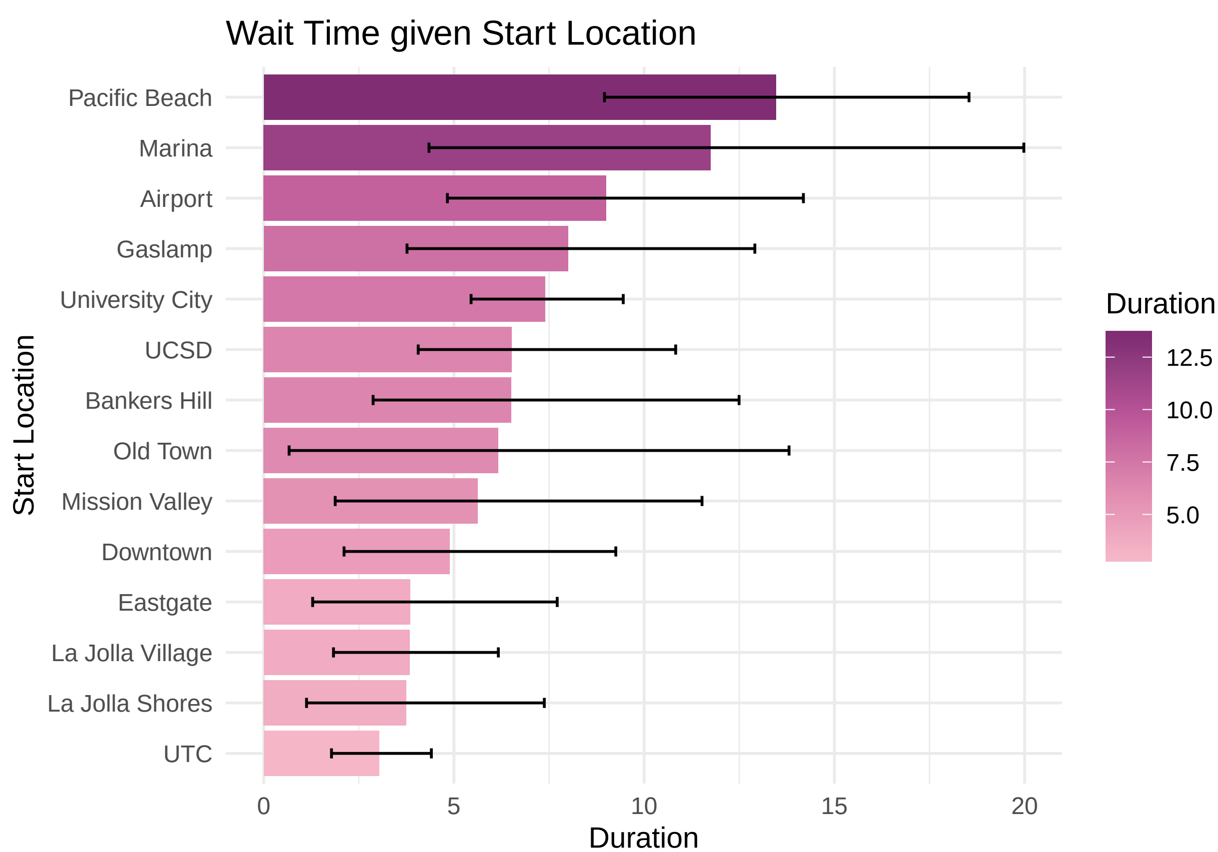



```r
ggsave("WaitDayandTime.png", p, width = 6, height = 4)
```

Wait Times for Location

```r
res = data.frame()

mask = data_all$Period != "Drive"

for (loc in unique(data_all[mask, ]$StartLocation)) {

  f = function(data, indices) {
```

```
    return(mean(data[indices]))
  }

  d = data_all[mask & data_all$StartLocation == loc, ]$Duration

  if (length(d) > 5) {
    b = boot(data_all[mask & data_all$StartLocation == loc, ]$Duration, statistic = f, R = 1000)

    r = boot.ci(b, type = "bca")$bca

    res[loc, c("lb", "Duration", "ub")] = list(r[1, 4], b$t0, r[1, 5])
  }
}

d = rownames_to_column(res, "StartLocation")

p = d %>%
  ggplot() + theme +
  geom_col(aes(StartLocation, Duration, fill = Duration)) +
  geom_errorbar(aes(StartLocation, ymin = lb, ymax = ub, width=0.2)) +
  scale_x_discrete(limits = d[order(d$Duration), "StartLocation"]) +
  scale_fill_continuous_sequential(palette = "Magenta", begin = 0.1, end = 0.9, rev = T) +
  coord_flip() +
  labs(title = "Wait Time given Start Location", x = "Start Location", y = "Duration")

p
```



Determinants

Does having conversation with the passenger make a difference on how much they tip? What if it's a shared ride? What about their interaction?

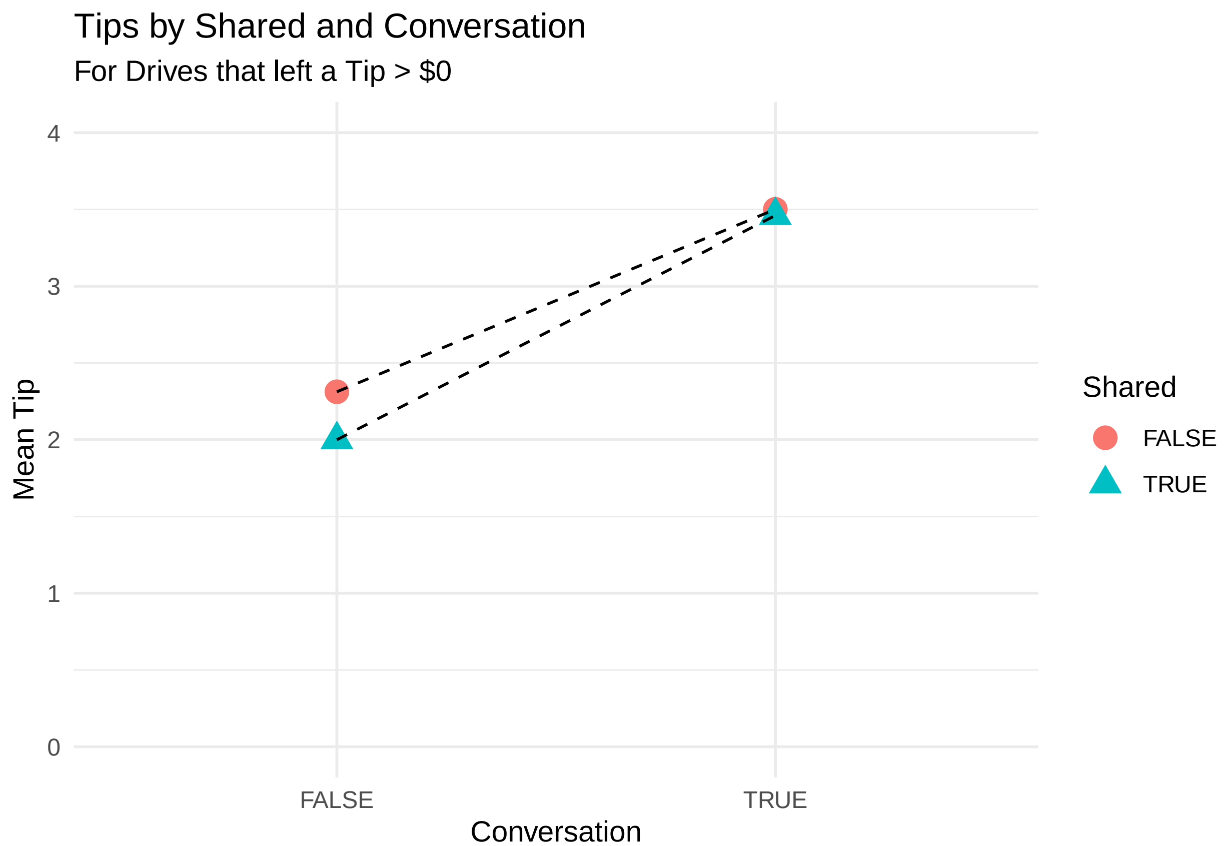
```
mean(data_drv$Tips > 0)
```

```
## [1] 0.3115578
```

```
mean(data_drv[data_drv$Tips > 0, ]$Tips)
```

```
## [1] 3.196935
```

```
p = data_drv %>%
  filter(Tips > 0) %>%
  group_by(Conversation, Shared) %>%
  summarise(avg = mean(Tips)) %>%
  ggplot() + theme +
  geom_point(aes(Conversation, avg, color = Shared, shape = Shared), size=4) +
  geom_line(aes(Conversation, avg, group = Shared), lty=2) +
  lims(y = c(0, 4)) +
  labs(x = "Conversation", y = "Mean Tip", title = "Tips by Shared and Conversation", subtitle = "For D

p
```




```
ggsave("TipsInteraction.png", p, width = 6, height = 5)
```

```
tip.model = glm(Tips ~ Distance + Duration + Passengers + Goal + Origin, data = data_drv %>% filter(Tip

tip.model.c = glm(Tips ~ Conversation + Distance + Duration + Passengers + Goal + Origin, data = data_d
```

```r
a = anova(tip.model, tip.model.c, "Chisq")
a
```

```
## Analysis of Deviance Table
##
## Model 1: Tips ~ Distance + Duration + Passengers + Goal + Origin
## Model 2: Tips ~ Conversation + Distance + Duration + Passengers + Goal +
##     Origin
##   Resid. Df Resid. Dev Df Deviance
## 1        45     5.3919
## 2        44     4.5237  1  0.86818
```

```r
pchisq(a$Deviance[2], a$Df[2], lower.tail = F)
```

```
## [1] 0.3514597
```

```r
tip.model = glm(Tips ~ Distance + Duration + Passengers + Goal + Origin, data = data_drv %>% filter(Tips

tip.model.c = glm(Tips ~ Shared + Distance + Duration + Passengers + Goal + Origin, data = data_drv %>%
a = anova(tip.model, tip.model.c, "Chisq")
a
```

```
## Analysis of Deviance Table
##
## Model 1: Tips ~ Distance + Duration + Passengers + Goal + Origin
## Model 2: Tips ~ Shared + Distance + Duration + Passengers + Goal + Origin
##   Resid. Df Resid. Dev Df  Deviance
## 1        45     5.3919
## 2        44     5.3838  1 0.0080384
```

```r
pchisq(a$Deviance[2], a$Df[2], lower.tail = F)
```

```
## [1] 0.9285598
```

```r
tip.model = glm(Tips ~ Conversation + Shared + Distance + Duration + Passengers + Goal + Origin, data =

tip.model.c = glm(Tips ~ Conversation + Shared + Conversation:Shared + Distance + Duration + Passengers
a = anova(tip.model, tip.model.c, "Chisq")
a
```

```
## Analysis of Deviance Table
##
## Model 1: Tips ~ Conversation + Shared + Distance + Duration + Passengers +
##     Goal + Origin
## Model 2: Tips ~ Conversation + Shared + Conversation:Shared + Distance +
##     Duration + Passengers + Goal + Origin
##   Resid. Df Resid. Dev Df  Deviance
## 1        43     4.5186
## 2        42     4.5153  1 0.0033261
```

```r
pchisq(a$Deviance[2], a$Df[2], lower.tail = F)
```

```
## [1] 0.9540096
```

```r
tip.model = glm(I(Tips > 0) ~ Distance + Duration + Passengers + Goal + Origin, data = data_drv, family

tip.model.c = glm(I(Tips > 0) ~ Conversation + Distance + Duration + Passengers + Goal + Origin, data =
coef(tip.model.c)["ConversationTRUE"]
```

```
## ConversationTRUE
##         1.582665
```

```r
a = anova(tip.model, tip.model.c, "Chisq")
a
```

```
## Analysis of Deviance Table
##
## Model 1: I(Tips > 0) ~ Distance + Duration + Passengers + Goal + Origin
## Model 2: I(Tips > 0) ~ Conversation + Distance + Duration + Passengers +
##     Goal + Origin
##   Resid. Df Resid. Dev Df Deviance
## 1       181     215.62
## 2       180     197.01  1    18.61
```

```r
pchisq(a$Deviance[2], a$Df[2], lower.tail = F)
```

```
## [1] 1.603434e-05
```

```r
tip.model = glm(I(Tips > 0) ~ Distance + Duration + Passengers + Goal + Origin, data = data_drv, family

tip.model.c = glm(I(Tips > 0) ~ Shared + Distance + Duration + Passengers + Goal + Origin, data = data_
coef(tip.model.c)["SharedTRUE"]
```

```
## SharedTRUE
##  -1.002508
```

```r
a = anova(tip.model, tip.model.c, "Chisq")
a
```

```
## Analysis of Deviance Table
##
## Model 1: I(Tips > 0) ~ Distance + Duration + Passengers + Goal + Origin
## Model 2: I(Tips > 0) ~ Shared + Distance + Duration + Passengers + Goal +
##     Origin
##   Resid. Df Resid. Dev Df Deviance
## 1       181     215.62
## 2       180     209.45  1    6.166
```

```r
pchisq(a$Deviance[2], a$Df[2], lower.tail = F)
```

```
## [1] 0.01302274
```

```r
tip.model = glm(I(Tips > 0) ~ Conversation + Shared + Distance + Duration + Passengers + Goal + Origin,

tip.model.c = glm(I(Tips > 0) ~ Conversation + Shared + Conversation:Shared + Distance + Duration + Pas
a = anova(tip.model, tip.model.c, "Chisq")
a
```

```
## Analysis of Deviance Table
##
## Model 1: I(Tips > 0) ~ Conversation + Shared + Distance + Duration + Passengers +
##     Goal + Origin
## Model 2: I(Tips > 0) ~ Conversation + Shared + Conversation:Shared + Distance +
##     Duration + Passengers + Goal + Origin
##   Resid. Df Resid. Dev Df Deviance
## 1       179     188.89
## 2       178     188.87  1 0.016166
```

```r
pchisq(a$Deviance[2], a$Df[2], lower.tail = F)
```

```
## [1] 0.898824
```