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(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.1.0       v purrr   0.2.5
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       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 speparate Google Spreadsheet). Here is the weighted (by gallons purchased) average price of gas.

```
gasPrice = 3.433 # $ / 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)
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)

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

data_drv = data_drv[data_drv$Cancel == F, ]

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)

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

```
## [1] "120 drives"
```

```
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="")
            Revenue = sum(Earnings + Tips, na.rm = T),
            GasCost = round(sum(GasUsage) * gasPrice, 2)) %>%
  as.matrix()
```

```
##      Session Wage     Revenue GasCost
## [1,] "25"    "$18.38" "43.52" "7.28"
## [2,] "26"    "$16.73" "33.54" "6.59"
## [3,] "27"    "$24.74" "47.51" "7.59"
```

As one may see, I've given about 120 rides in just under 30 driving sessions (a session is the period in which I have Lyft turned on and am accepting rides).

Preview Data

```
tail(data_drv)
```

```
## # A tibble: 6 x 36
##   Session Date                Period Movement Distance Duration Passengers
##     <dbl> <dttm>              <fct>  <fct>       <dbl>    <dbl>      <dbl>
## 1      27 2019-08-24 00:00:00 Drive  Drive        4.22     5.9           1
## 2      27 2019-08-24 00:00:00 Drive  Drive        9.99    15.2           1
## 3      27 2019-08-24 00:00:00 Drive  Drive        1.76     5.73          2
## 4      27 2019-08-24 00:00:00 Drive  Drive        1.06     9.47          2
## 5      27 2019-08-24 00:00:00 Drive  Drive        1.39     8.3           4
## 6      27 2019-08-24 00:00:00 Drive  Drive        7.06    13.8           3
## # ... with 29 more variables: Earnings <dbl>, Tips <dbl>, Cancel <lgl>,
## #   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>
```

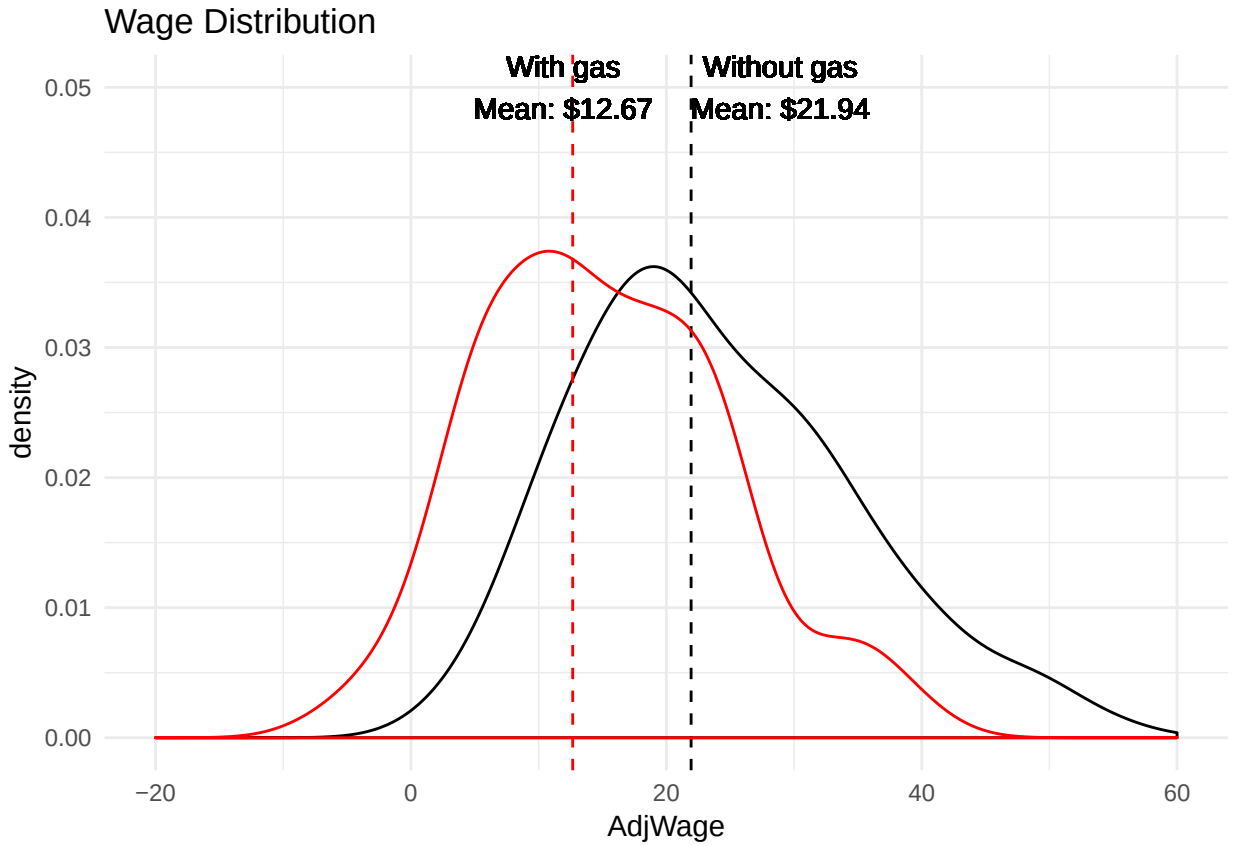
Visualizations

Just a short cut for style.

```
theme = theme_minimal()
```

Wage Distribution

```
maint = (700 + 30 + 100 + 300) / 5000 # depreciation + oil + service + parts perthe next 5,000 miles

avg = sum(data_drv$Earnings + data_drv$Tips) * (60/sum(data_drv$AdjDuration))
maintInc = (data_drv$Earnings + data_drv$Tips - gasPrice * data_drv$AdjGas - maint * data_drv$AdjDistan
avgMaintInc = sum(data_drv$Earnings + data_drv$Tips - gasPrice * data_drv$AdjGas - maint * data_drv$Adj

ggplot(data_drv) + theme +
  geom_density(aes(AdjWage)) +
  geom_vline(xintercept = avg, lty=2) +
  geom_text(aes(avg, y = 0.05, label = paste("Without gas\nMean: $", round(avg, 2), sep="")), nudge_x =
  geom_density(aes(maintInc), col="red") +
  geom_vline(xintercept = avgMaintInc, col="red", lty=2) +
  geom_text(aes(avg, y = 0.05, label = paste("With gas\nMean: $", round(avgMaintInc, 2), sep="")), nudg
  lims(x=c(-20,60)) +
  labs(title = "Wage Distribution")
```




```
b = boot((data_drv$Earnings + data_drv$Tips - gasPrice * data_drv$AdjGas - maint * data_drv$AdjDistance)
b$t0
```

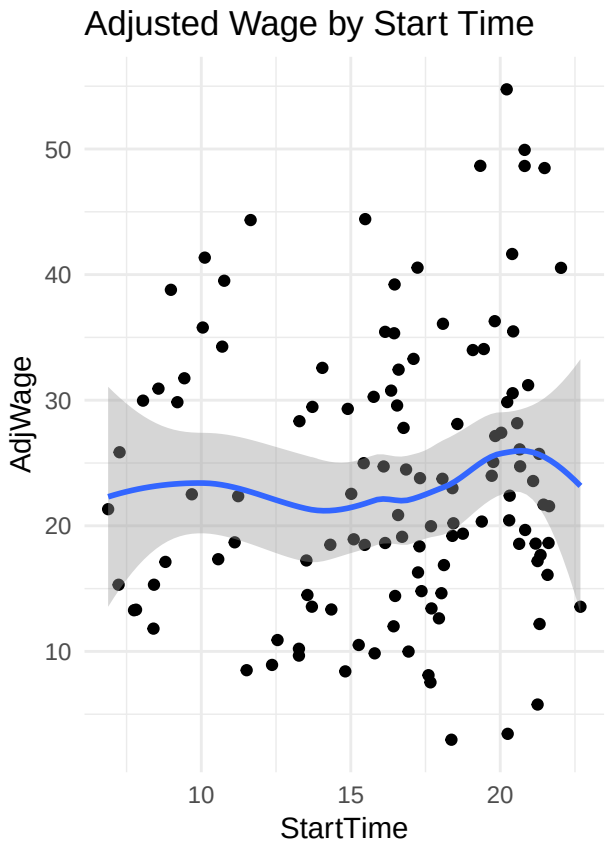
```
## [1] 15.05581
```

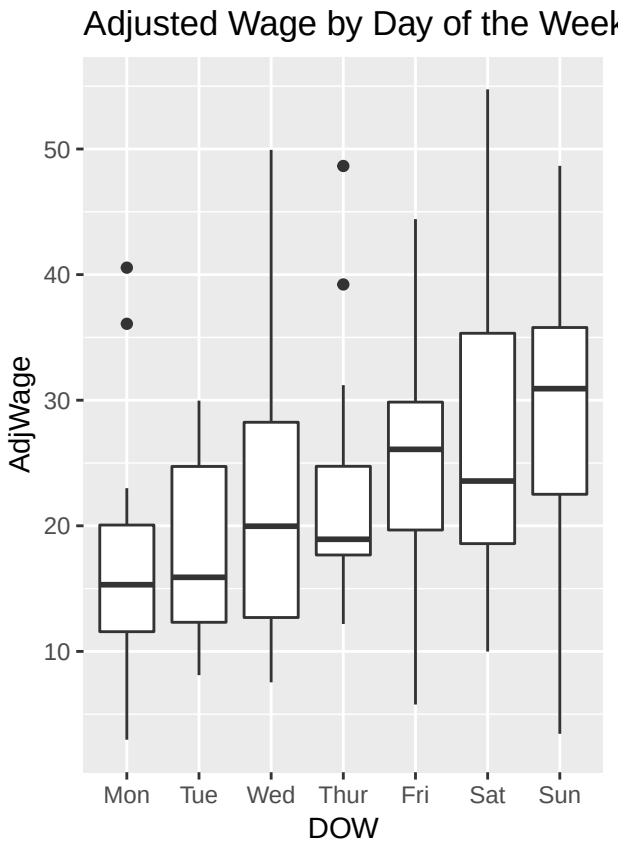
```
boot.ci(b, type = "bca")$bca[1, c(4, 5)]
```

```
##
## 13.34051 16.83380
```

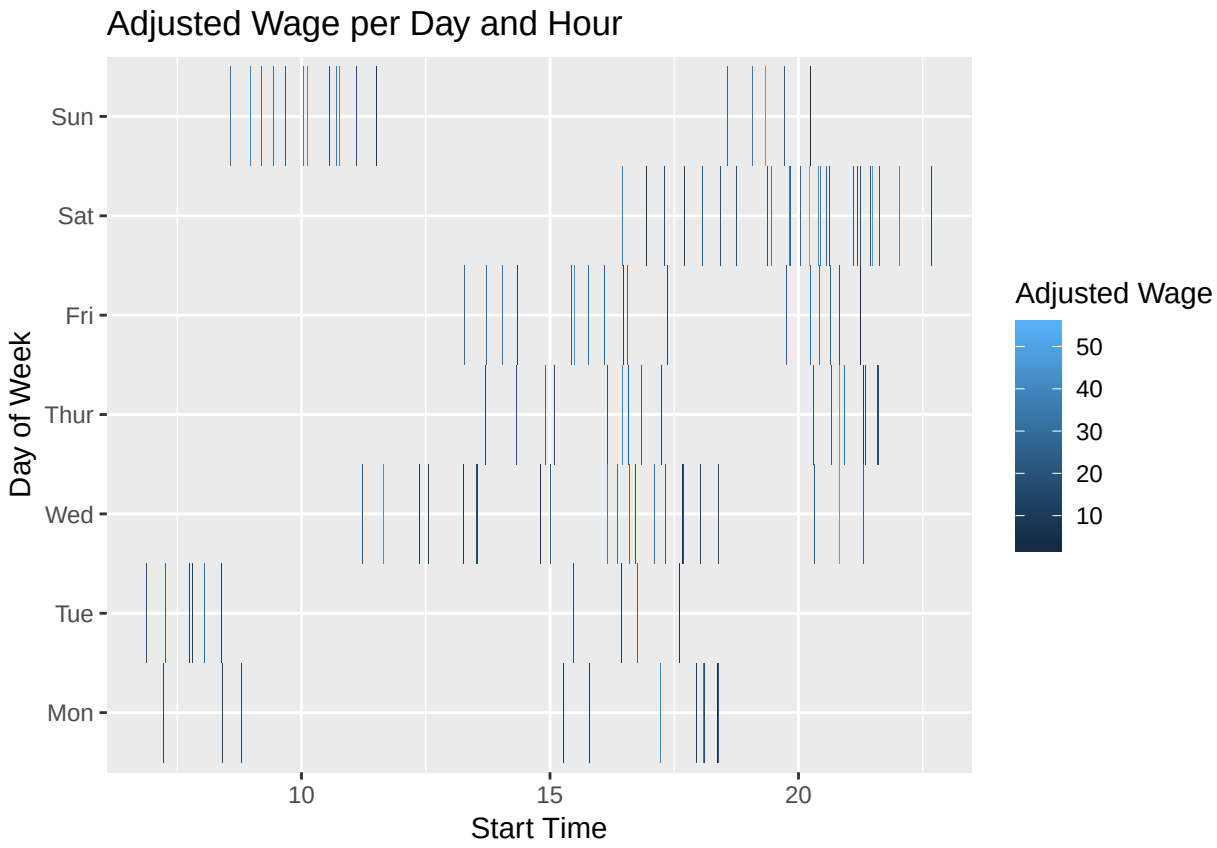
Driving Days and Times

```
p1 = ggplot(data_drv, aes(StartTime, AdjWage)) + theme +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(title = "Adjusted Wage by Start Time")

p2 = ggplot(data_drv) +
  geom_boxplot(aes(DOW, AdjWage)) +
  labs(title = "Adjusted Wage by Day of the Week")

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



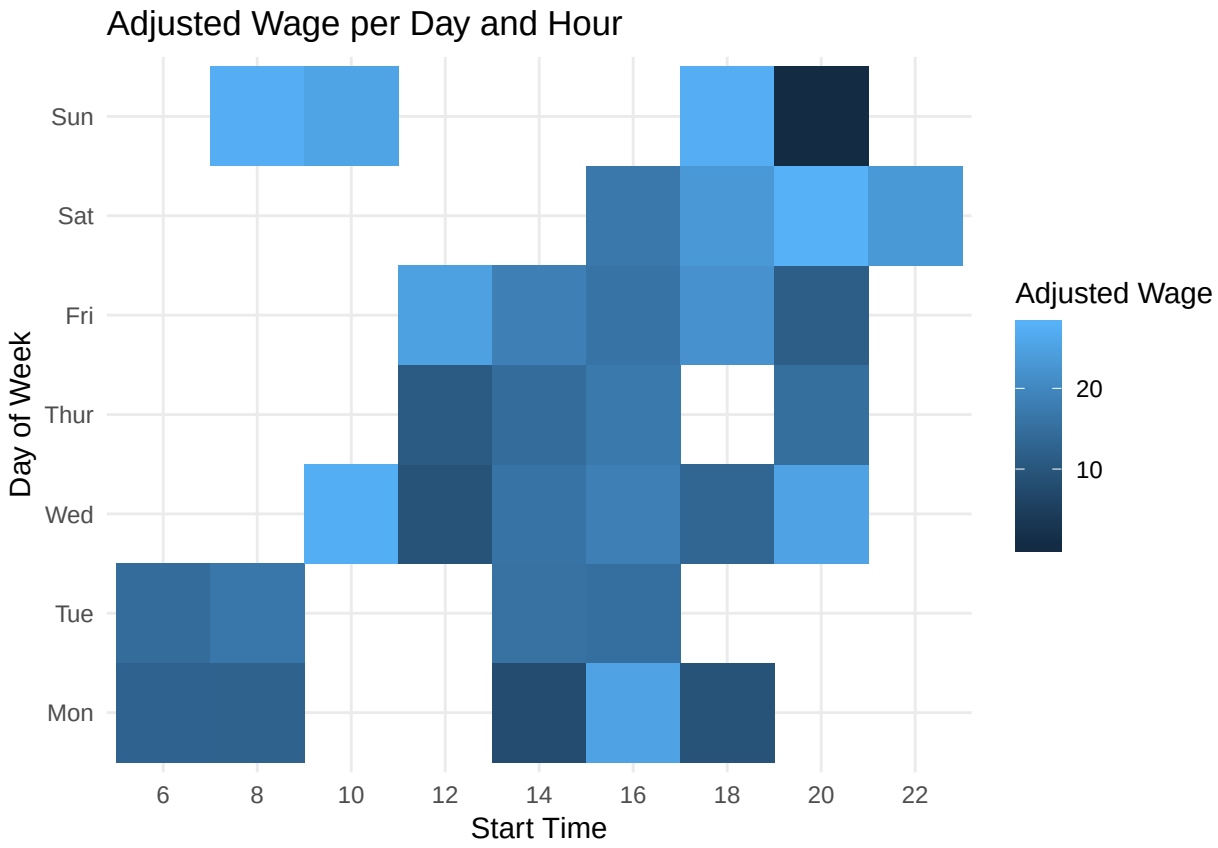



```
ggplot(data_drv) +
  geom_tile(aes(x = StartTime, y = DOW, fill = AdjWage)) +
  labs(x = "Start Time", y = "Day of Week", fill = "Adjusted Wage", title = "Adjusted Wage per Day and
```



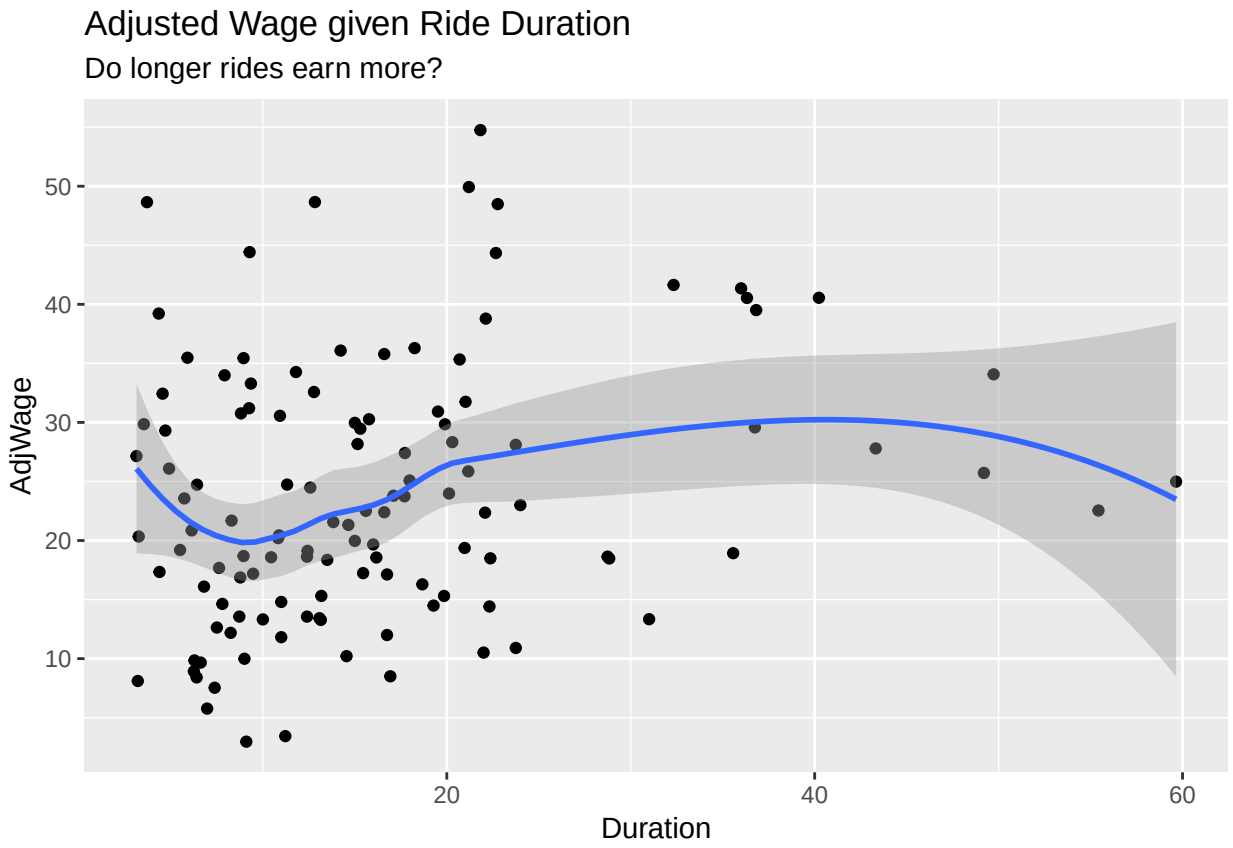
Average Wage given Days and Times

```
bins = seq(6, 22, 2)

data_drv$StartTimeBin = cut(data_drv$StartTime, breaks = c(bins, 24), labels = bins)

data_drv %>%
  group_by(StartTimeBin, DOW) %>%
  summarise(m = sum(Earnings + Tips - gasPrice * AdjGas) * (60 / sum(AdjDuration)), n = n()) %>%
  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
```



Duration on Wage

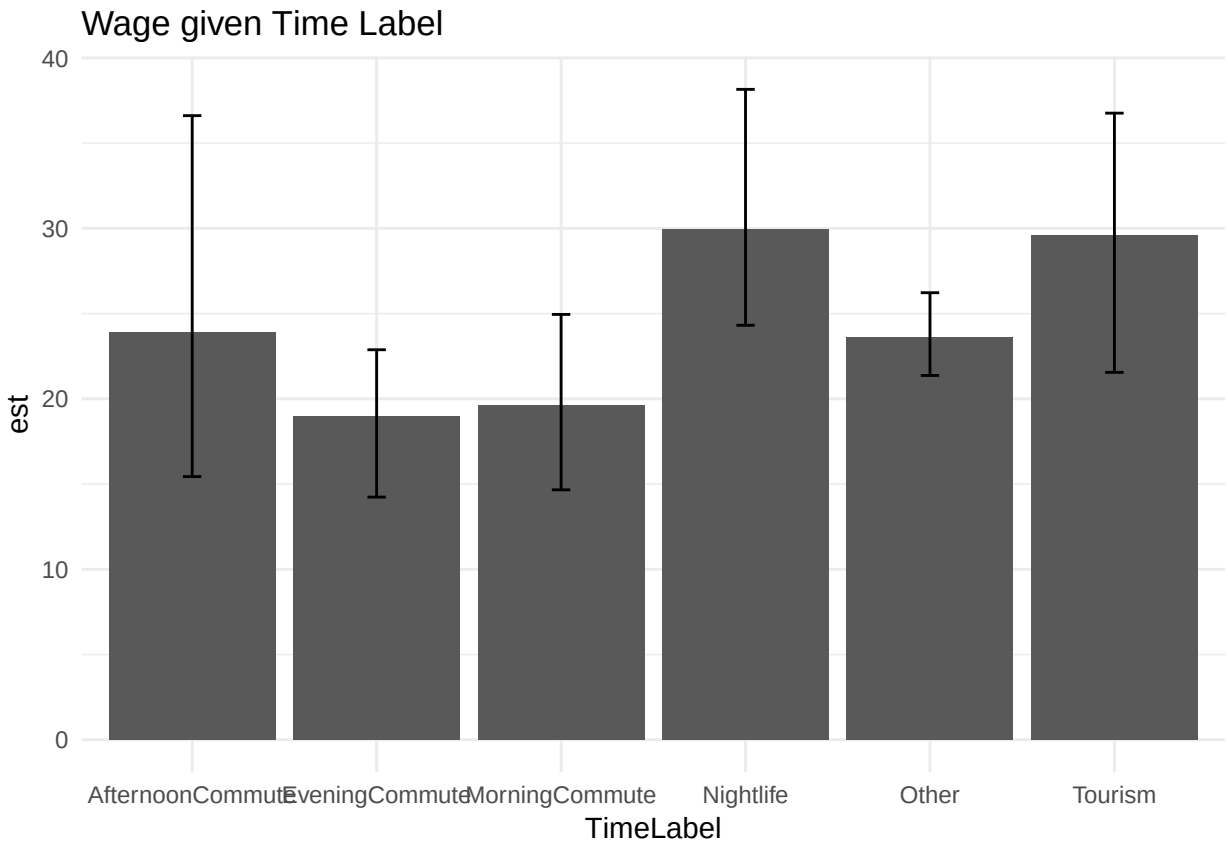
```
ggplot(data_drv, aes(Duration, AdjWage)) +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(title = "Adjusted Wage given Ride Duration", subtitle = "Do longer rides earn more?")
```



Time Labels

This is a WIP.

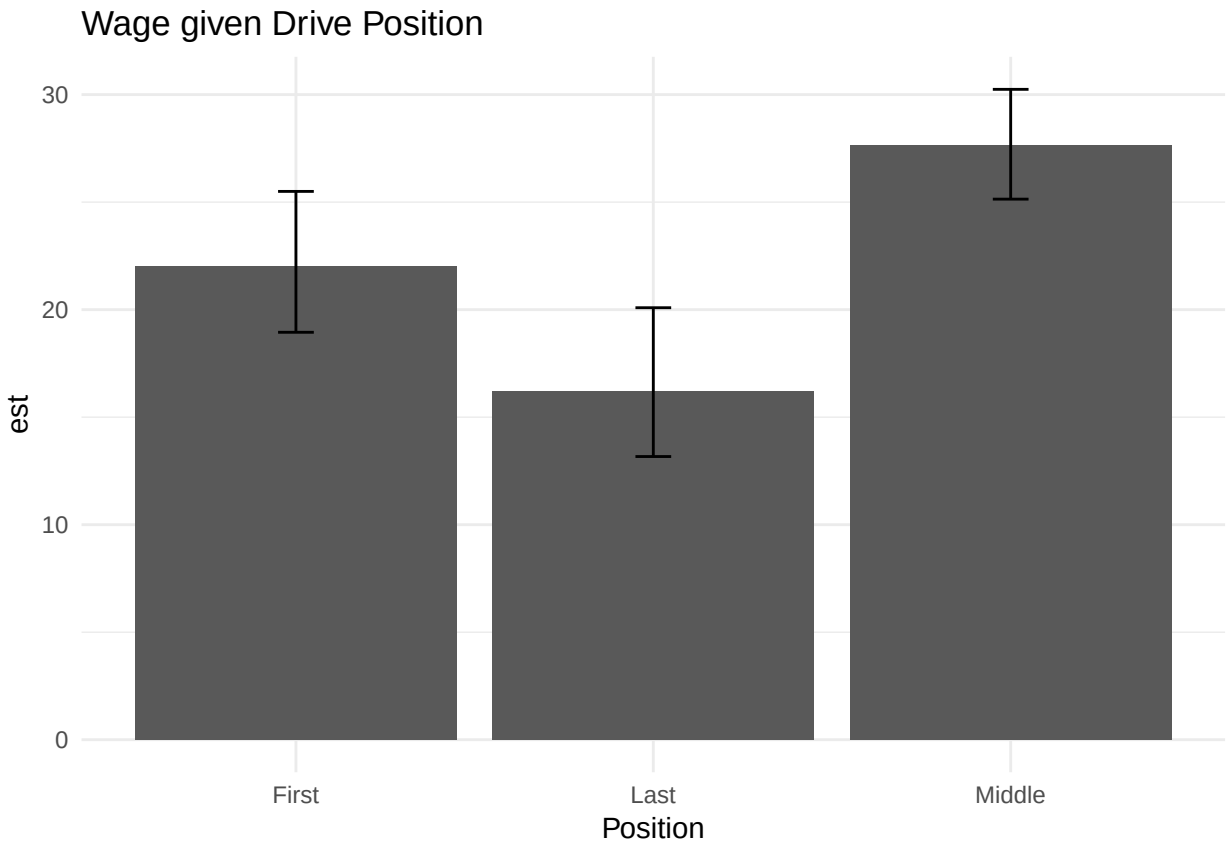
```
res = data.frame()

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), stat="identity") +
  geom_errorbar(aes(TimeLabel, ymin = lb, ymax = ub, width=0.1)) +
  labs(title = "Wage given Time Label")
```



Drive Position

Do the first and last drives of a session differ 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", "est", "ub")] = list(r[1, 4], b$t0, r[1, 5])
}

rownames_to_column(res, "Position") %>%
  ggplot() + theme +
  geom_bar(aes(Position, est), stat="identity") +
  geom_errorbar(aes(Position, ymin = lb, ymax = ub, width=0.1)) +
  labs(title = "Wage given Drive Position")
```

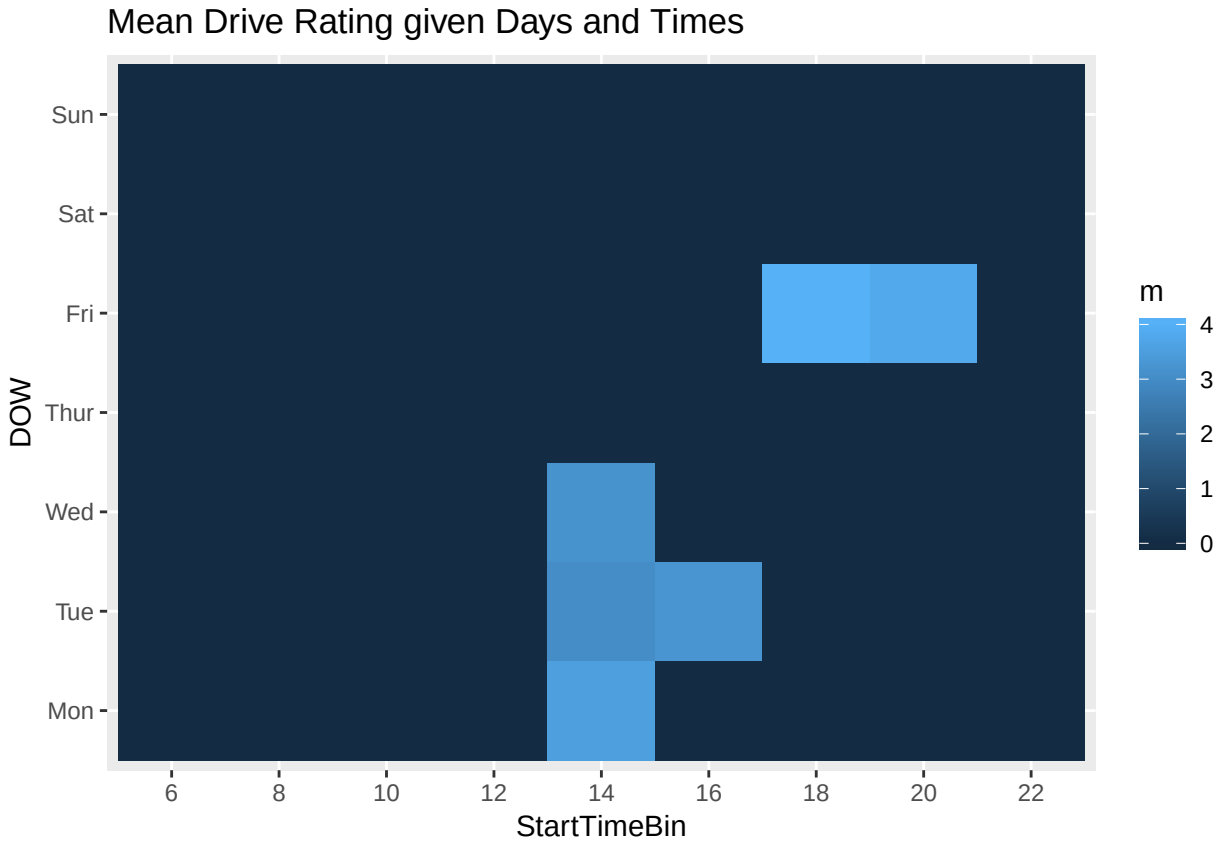


It appears as though the last ride is likely to earn less than the other two types of rides.

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(RatingMean)) %>%
  complete(DOW, fill = list(m = 0)) %>%
  ggplot() +
  geom_tile(aes(x = StartTimeBin, y = DOW, fill = m)) +
  labs(title = "Mean Drive Rating given Days and Times")
```



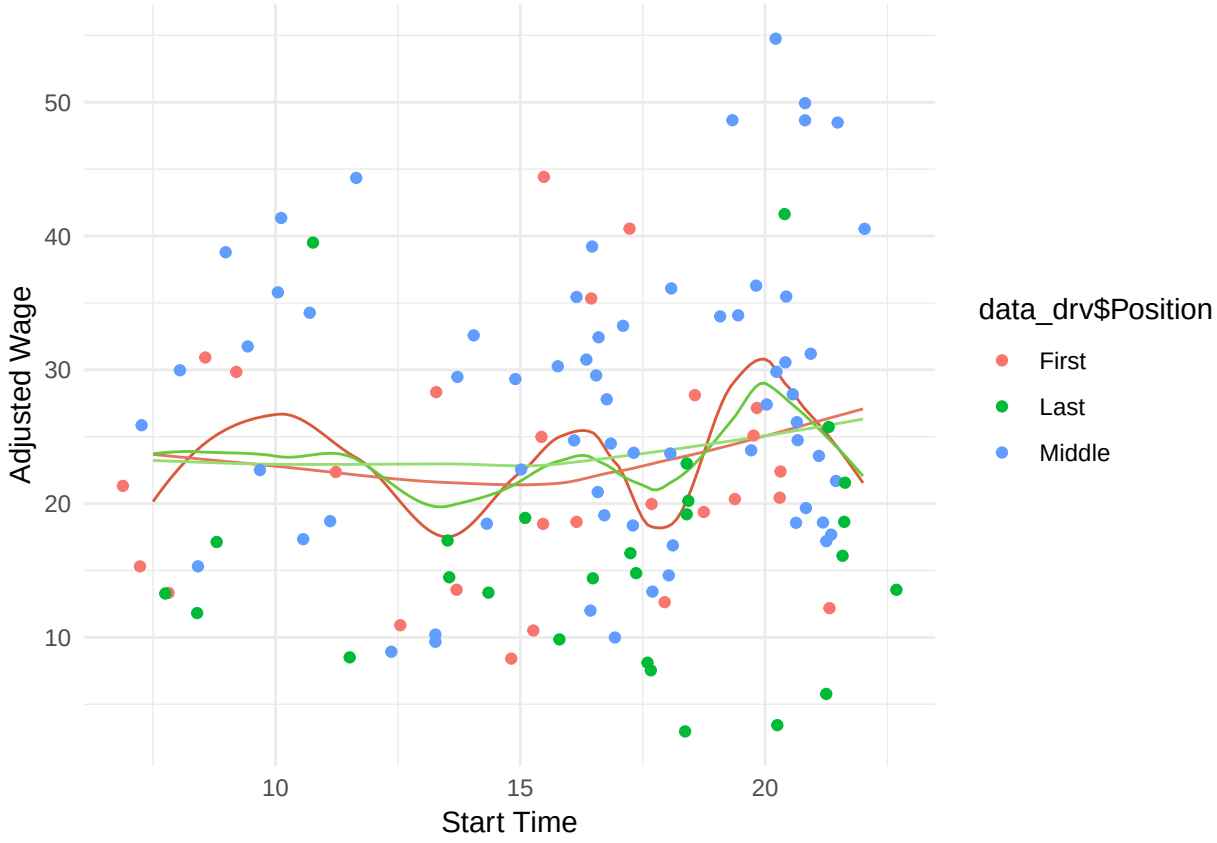
Local Regression

Local and global regression patterns given the hour of the day on wage.

```
m1 = loess(AdjWage ~ StartTime,
           span = 0.4,
           degree = 2,
           data = data_drv,
           family = "gaussian")

m2 = loess(AdjWage ~ StartTime,
           span = 1.0,
           degree = 2,
           data = data_drv,
           family = "gaussian")

m3 = loess(AdjWage ~ StartTime,
           span = 0.4,
           degree = 1,
           data = data_drv,
           family = "gaussian")

m4 = loess(AdjWage ~ StartTime,
           span = 1.0,
           degree = 1,
           data = data_drv,
```

```
          family = "gaussian")
```

```
x = seq(7.5, 22, 0.1)
d = data.frame(x = rep(x, 4),
               y = c(predict(m1, x), predict(m2, x), predict(m3, x), predict(m4, x)),
               k = rep(c(1, 2, 3, 4), each = length(x)))

ggplot() + theme +
  geom_line(aes(d$x, d$y, group = d$k), color = rep(c("#db593b", "#e3765d", "#69c93c", "#95de73"), each
  geom_point(aes(data_drv$StartTime, data_drv$AdjWage, color=data_drv$Position)) +
  labs(x = "Start Time", y = "Adjusted Wage")
```

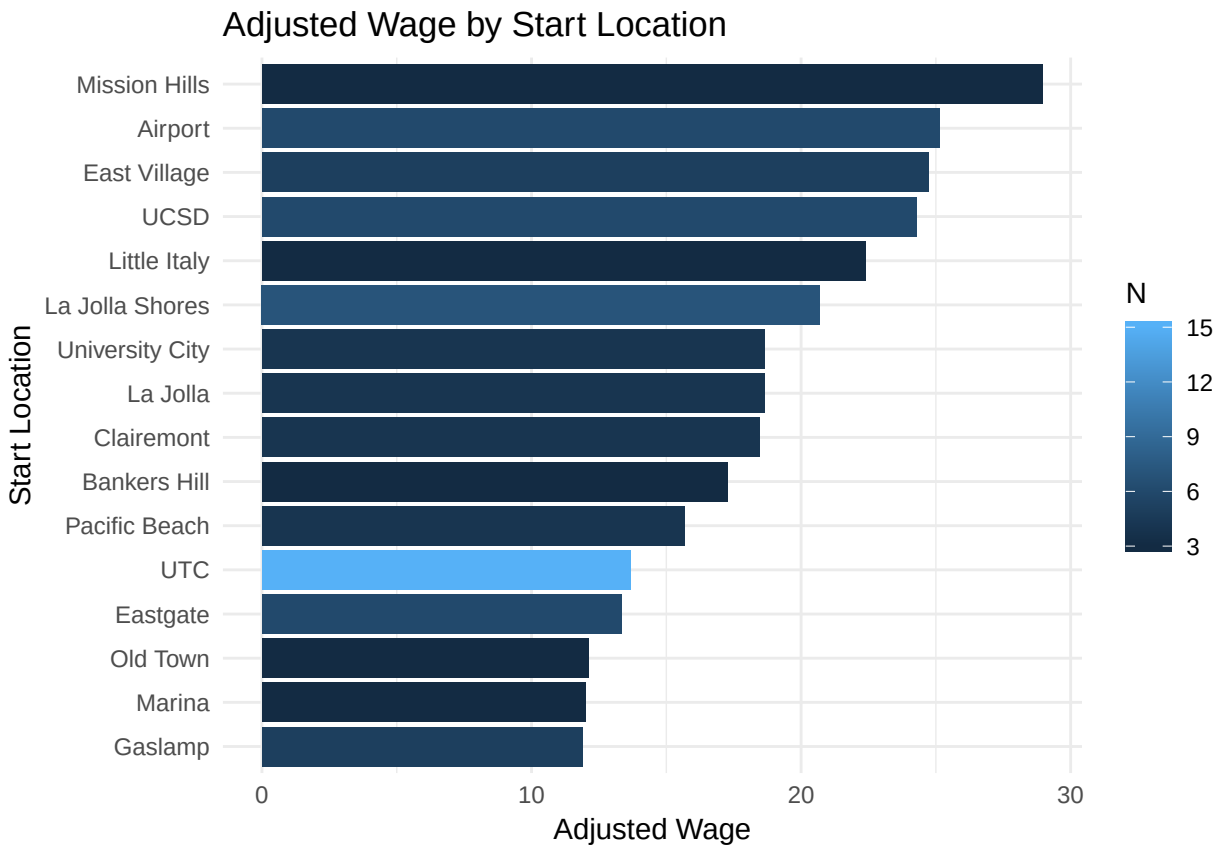


Start Location

How does where the ride is started affect how much I'll earn? Bootstrapped error bars to come.

```
sl_wage = data_drv %>%
  group_by(StartLocation) %>%
  summarise(AdjWage = round(sum(Earnings + Tips - gasPrice * AdjGas) * (60 / sum(AdjDuration)), 2),
            SD = sd((Earnings + Tips - gasPrice * AdjGas) * (60 / AdjDuration)),
            AvgDur = mean(AdjDuration),
            N = n()) %>%
  filter(N > 2)

sl_wage %>%
  arrange(AdjWage) %>%
```

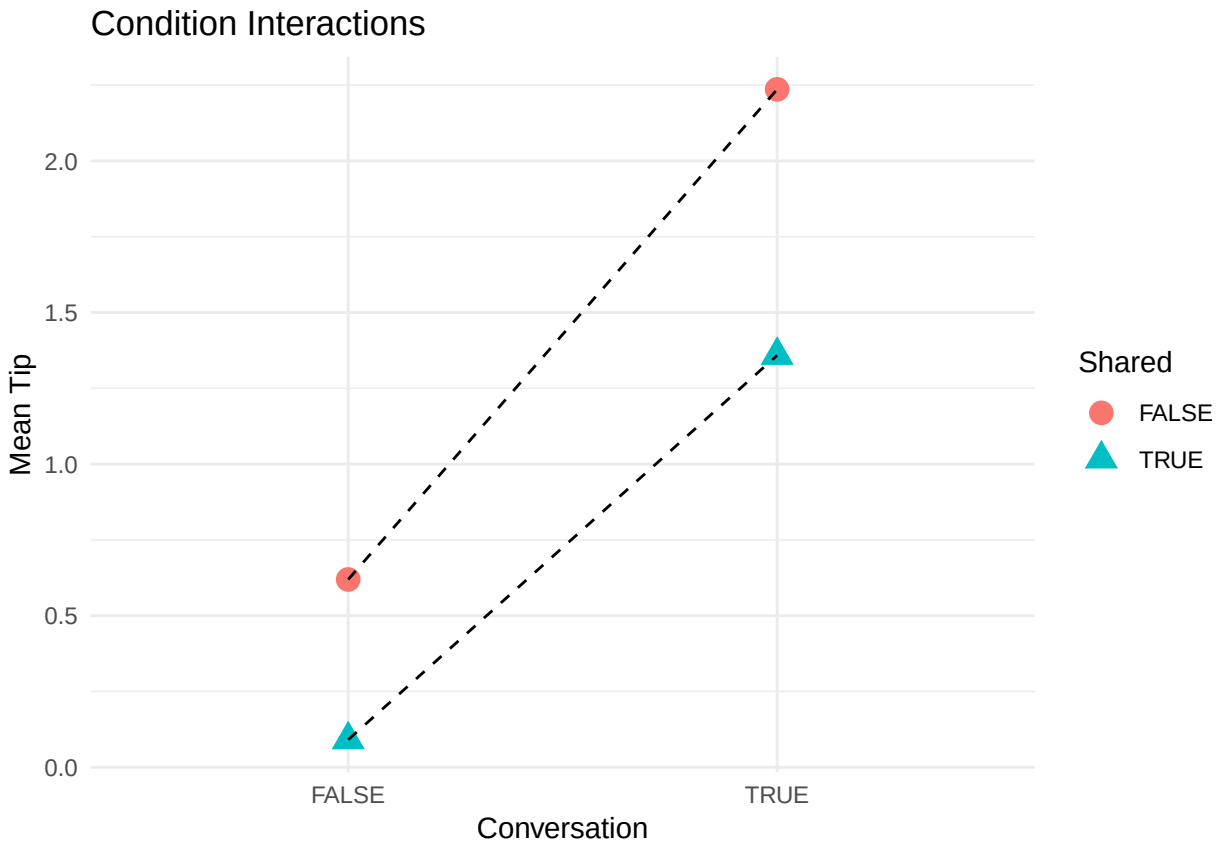
```
  mutate(StartLocation = factor(StartLocation, StartLocation)) %>%
  ggplot() + theme +
  geom_bar(aes(StartLocation, AdjWage, fill = N), stat="identity") +
  # geom_errorbar(aes(StartLocation, ymin = AdjWage - SD / sqrt(N), ymax = AdjWage + SD / sqrt(N), widt
  labs(x = "Start Location", y = "Adjusted Wage", title = "Adjusted Wage by Start Location") +
  coord_flip()
```



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?

```
data_drv %>%
  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) +
  labs(x = "Conversation", y = "Mean Tip", title = "Condition Interactions")
```



Totals

```
sum(data_drv$AdjDistance, na.rm = T)
```

```
## [1] 1320.66
```

```
sum(data_drv$AdjDuration, na.rm = T) / 60
```

```
## [1] 54.912
```

About 1,300 miles driven for Lyft, and 55 hours worth of my time.