Exploratory Data Analysis - Chicago Taxi Trips

Contents

A in-depth exploratory data analysis for the Chicago taxi dataset. In order to efficiently perform exploratory data analysis, we used a subset containing 1% of the data for each of the provided training datasets so that our subset's yearly proportions were the same as those of the full data's. Hourly, daily, monthly and yearly trends were compared over repeated subsets to demonstrate that the trends in each subset were representative of the full data's trends. It should be noted that the EDA plots and summary statistics examined in this section of the report are all run on the same 1% subset.

Data

```
library(data.table)
library(dplyr)
library(magrittr)
library(MASS)
library(ggplot2)
library(gridExtra)
library(kableExtra)
```

```
taxi_2013 <- fread('subset_2013.csv')
taxi_2014 <- fread('subset_2014.csv')
taxi_2015 <- fread('subset_2015.csv')
taxi_2016 <- fread('subset_2016.csv')
taxi_2017 <- fread('subset_2017.csv')

taxi_2013[, 12 := NULL]
names(taxi_2014) <- names(taxi_2013)
names(taxi_2015) <- names(taxi_2013)
names(taxi_2016) <- names(taxi_2013)
names(taxi_2017) <- names(taxi_2013)

taxi_df <- data.table(rbindlist(list(taxi_2013, taxi_2014, taxi_2015, taxi_2016)),
                      Year = rep(c(2013, 2014, 2015, 2016), times = c(nrow(taxi_2013), nrow(taxi_2014),

taxi_df <- na.omit(taxi_df)
taxi_2017 <- na.omit(taxi_2017)
```

Note that we have an extra column in the taxi_2013 data so we removed it.

Previewing the data

Trip ID

Taxi ID

Trip Start Timestamp

Trip End Timestamp

Trip Seconds

Trip Miles

Pickup Census Tract

Dropoff Census Tract

Pickup Community Area

Pickup O'Hare Community Area

Dropoff Community Area

Fare

Tips

Tolls

Extras

Trip Total

Payment Type

Company

Pickup Centroid Latitude

Pickup Centroid Longitude

Pickup Centroid Location

Dropoff Centroid Latitude

Dropoff Centroid Longitude

Dropoff Centroid Location

Year

8c98866acf97e2d43363c1e62d408dfb63368a8b

6b50c5bb5761c3cb4853a8ca215dee42e63e922c7ebb0ce74cd52a2b3076c37ed38a45d19077a78a0bd19fe39c5e2efce4ed1ad13af2f2b67

03/15/2013 5:15:00 PM

03/15/2013 5:15:00 PM

660

0.0

17031281900

17031081500

28

0

8

$7.45

$5.00

$0.00

$1.00

$13.45

Credit Card

Taxi Affiliation Services

41.87926

-87.64265

POINT (-87.642648998 41.8792550844)

41.89251

-87.62621

POINT (-87.6262149064 41.8925077809)

2013

8c98ba8af921ff20f80b4a64ea5d4b130f5360c0

ed7090d32800eec667900e884a935a298e55efaaaea9d66f2753858efee1178aeb3a874b413c4db7adf5e5c748efbe41682d4625acd9d540c

02/08/2013 7:45:00 PM

02/08/2013 7:45:00 PM

360

1.5

17031081800

17031839100

8

0

32

$6.65

$2.00

$0.00

$0.00

$8.65

Credit Card

41.89322

-87.63784

POINT (-87.6378442095 41.8932163595)

41.88099

-87.63275

POINT (-87.6327464887 41.8809944707)

2013

8c98f469ae3f77d7ab46e41e3ae0491aba8d7ebc

bf426bcd9e203d196e477b945f87932c33c7f177d884ab76aad30e26723e7b5e1f62423c71acb8b299958af0de1c7fa195a45d71f2e67ec8df

05/04/2013 8:45:00 PM

05/04/2013 8:45:00 PM

480

0.0

17031831000

17031062700

22

0

6

$7.25

$0.00

$0.00

$1.00

$8.25

Cash

Taxi Affiliation Services

41.91601

-87.67510

POINT (-87.6750951155 41.9160052737)

41.93609

-87.66611

POINT (-87.6661106945 41.9360865352)

2013

8c9a454a2e2fe9f04471f266c71d4a19ec413f29

c0250f358cae01c5319aeb7b39827e53f9a2259eb32e4c2fba048fd9e7d0a69ebc41c807016fb9f2d815b8a248163904bdbad4ddbe6196d70

09/11/2013 12:45:00 PM

09/11/2013 1:00:00 PM

420

0.0

17031320100

17031081800

32

0

8

$6.25

$2.00

$0.00

$0.00

$8.25

Credit Card

Blue Ribbon Taxi Association Inc.

41.88499

-87.62099

POINT (-87.6209929134 41.8849871918)

41.89322

-87.63784

POINT (-87.6378442095 41.8932163595)

2013

8c9a71715c08f9be6b70f6ccdbdb55e8b64c725d

8b07f9156e568a37d362463c84dbd1118b4eeb753bae502fded3dd7ba0040f5476ebbabb0428fc752b6b1e3de90c50ae9f33c396493a873a

12/14/2013 9:00:00 PM

12/14/2013 9:30:00 PM

1620

0.0

17031241400

17031070400

24

0

7

$16.25

$3.45

$0.00

$1.00

$20.70

Credit Card

Choice Taxi Association

41.90603

-87.67531

POINT (-87.6753116216 41.906025969)

41.92897

-87.65616

POINT (-87.6561568309 41.9289672664)

2013

8c9b5278159c4c8a3fba3292b4bb326fc927fed1

d5ca4708ec7536df92968ff54683f27934c4c48beb482f831315a48b6e116dac3b944c0765bbe76b073f6cd5696a70e5b500d200db836c698

02/01/2013 10:45:00 AM

02/01/2013 11:00:00 AM

300

0.8

17031081403

17031081201

8

0

8

$5.25

$1.00

$0.00

$0.00

$6.25

Credit Card

Dispatch Taxi Affiliation

41.89092

-87.61887

POINT (-87.6188683546 41.8909220259)

41.89916

-87.62621

POINT (-87.6262105324 41.8991556134)

2013

Analysis of predictors over time

Before we begin let's create some datetime variables for our time.

```
date_time_2013 <- strptime(taxi_2013$`Trip Start Timestamp`, format = '%m/%d/%Y %I:%M:%S %p')
date_time_2014 <- strptime(taxi_2014$`Trip Start Timestamp`, format = '%m/%d/%Y %I:%M:%S %p')
date_time_2015 <- strptime(taxi_2015$`Trip Start Timestamp`, format = '%m/%d/%Y %I:%M:%S %p')
date_time_2016 <- strptime(taxi_2016$`Trip Start Timestamp`, format = '%m/%d/%Y %I:%M:%S %p')
date_time <- strptime(taxi_df$`Trip Start Timestamp`, format = '%m/%d/%Y %I:%M:%S %p')
date_time2 <- strptime(taxi_df$`Trip End Timestamp`, format = '%m/%d/%Y %I:%M:%S %p')

# Add month column
taxi_df[, Month := as.numeric(strftime(date_time, '%m'))]
```

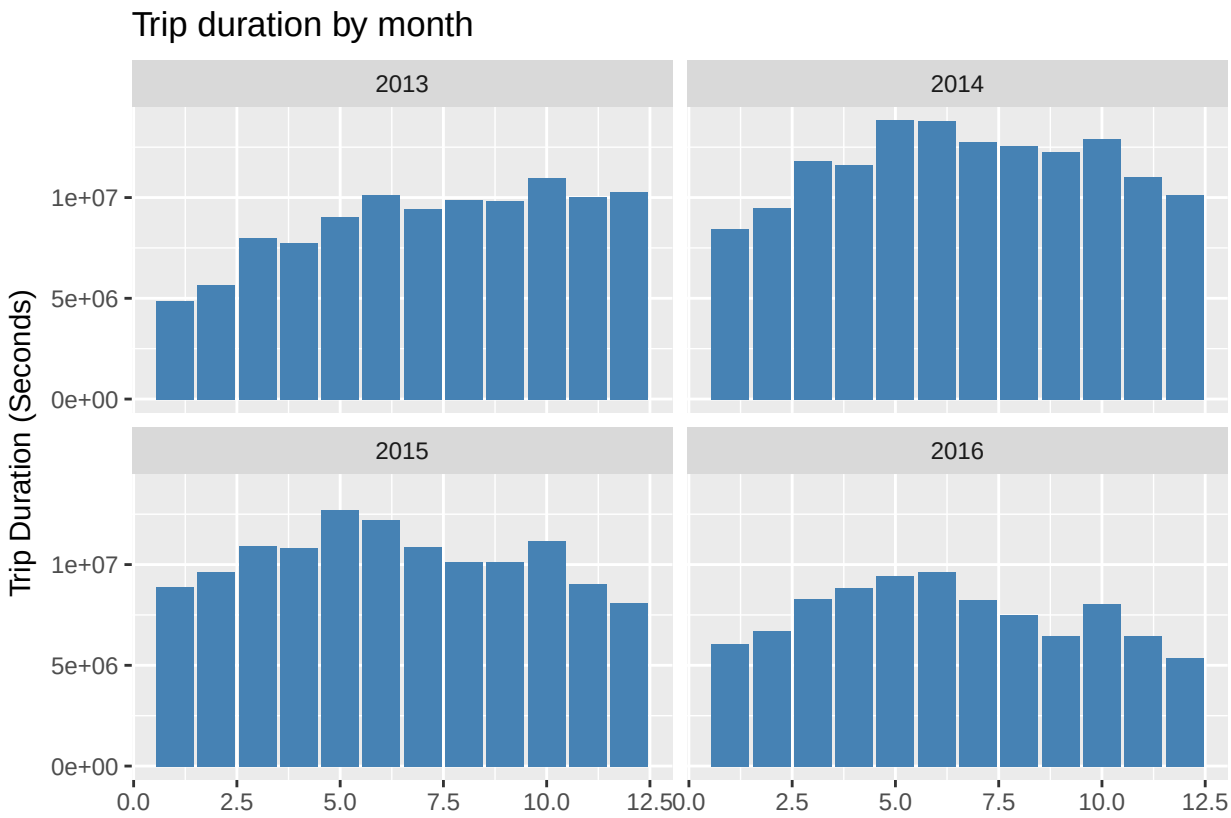
By Month

Let's begin by doing a breakdown by the month of year.

Trip Duration

The first variable we want to examine is trip duration and how it varies from month to month.

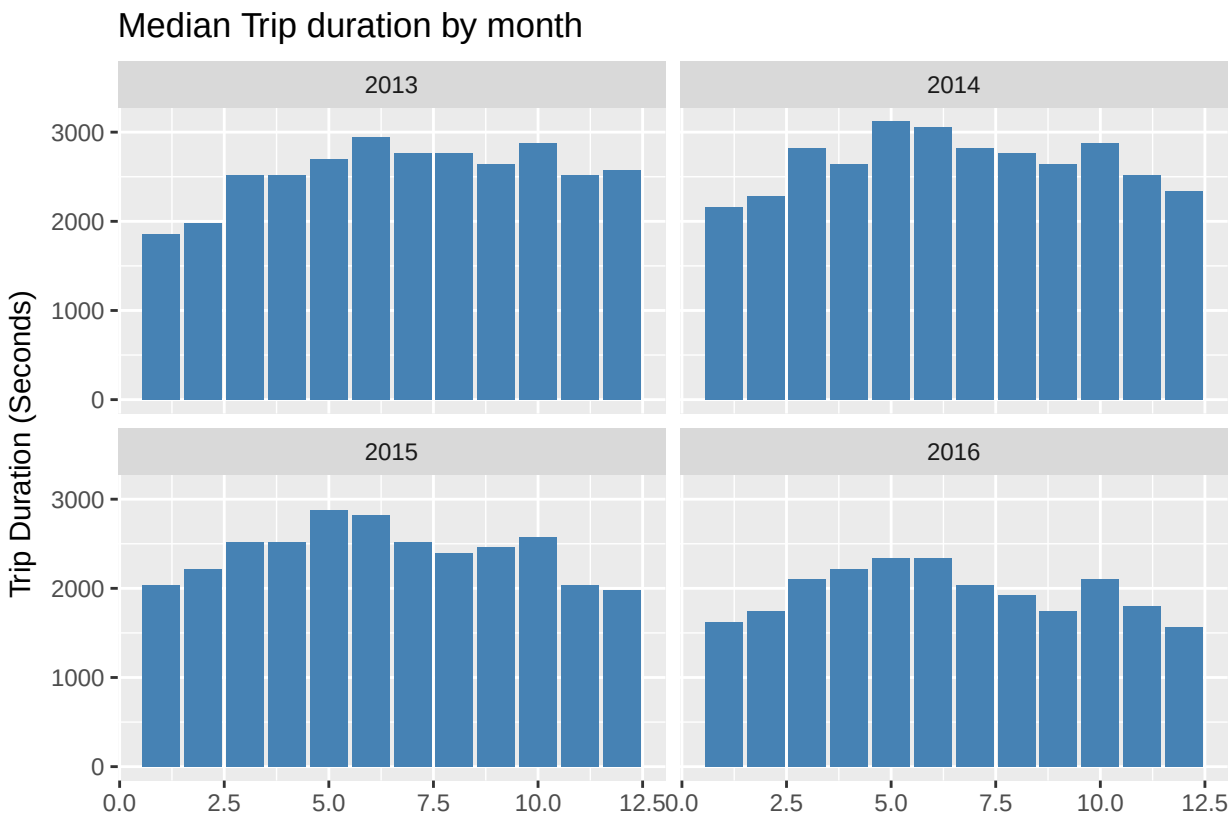
```
g1 <- taxi_df[, .(TripDuration = `Trip Seconds`, Month), by = Year] %>%
  ggplot(aes(Month, TripDuration)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = '', y = 'Trip Duration (Seconds)', title = 'Trip duration by month')
g1
```



The first observation we can make is that the duration of trips increased through 2014 and then started decreasing. The general shape of the histograms have remained the same with May, June, and October

having the longest trips and January having the shortest trips. This is total trip duration so let's see if that is due to a high trip count or if the general taxi ride is longer those months. We will make the same graph but this time with median taxi trip duration.

```
g2 <- taxi_df[, .(TotalTrip = sum(`Trip Seconds`)), by = list(`Taxi ID`, Month, Year)][, .(Median = medi
  ggplot(aes(Month, Median)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = '', y = 'Trip Duration (Seconds)', title = 'Median Trip duration by month')
g2
```

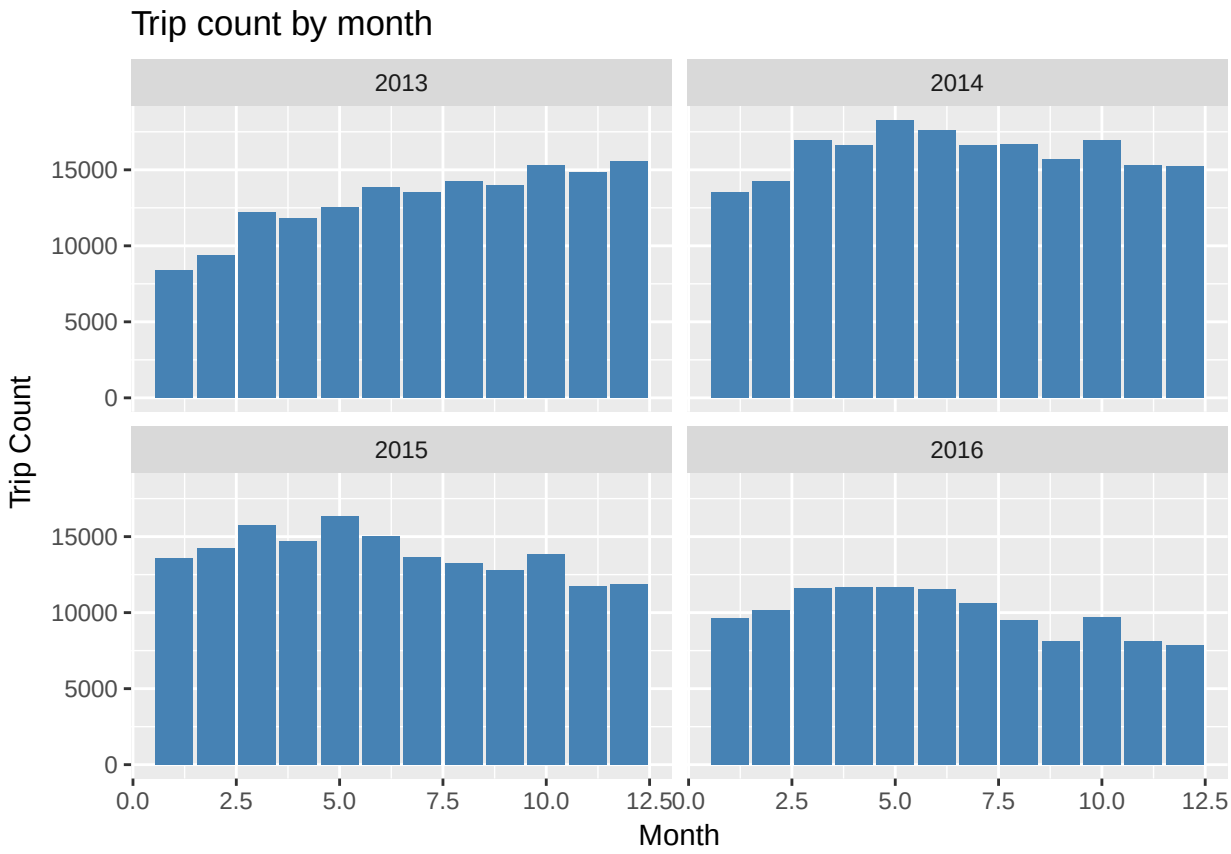


Looking at the median trip durations, we still see that the peaks are generally in May, June, and October and that the low points are in January. The difference is not as big though.

Trip Count

Now let's investigate the number of trips that have been taken each month.

```
g3 <- taxi_df[, .(Month), by = Year][, .(TripCount = .N), by = c('Year', 'Month')] %>%
  ggplot(aes(Month, TripCount)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = 'Month', y = 'Trip Count', title = 'Trip count by month')
g3
```

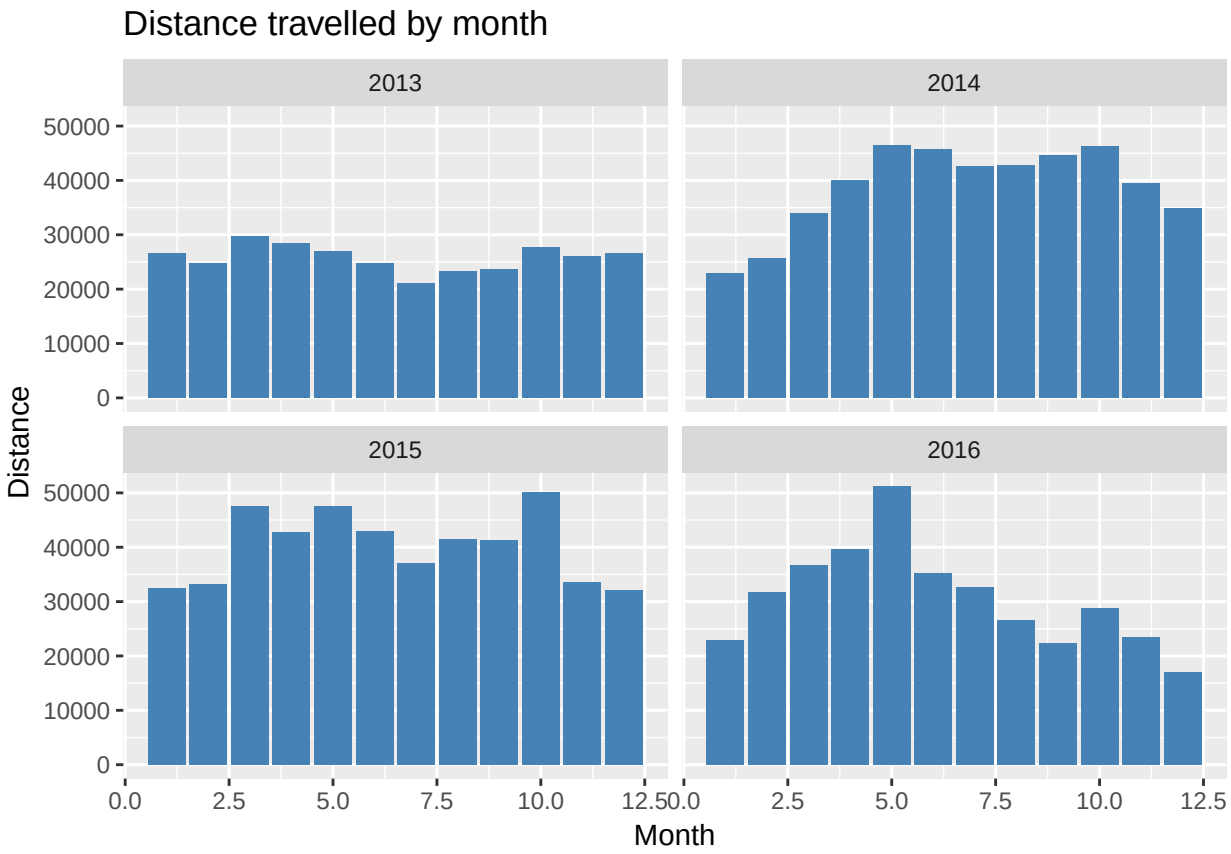


We can see that the trip count follows the same trend as trip duration, where 2014 is the higheest and procees to decrease every year after. We can still find that May, June, and October are the peak months and January is the low.

Trip Distance

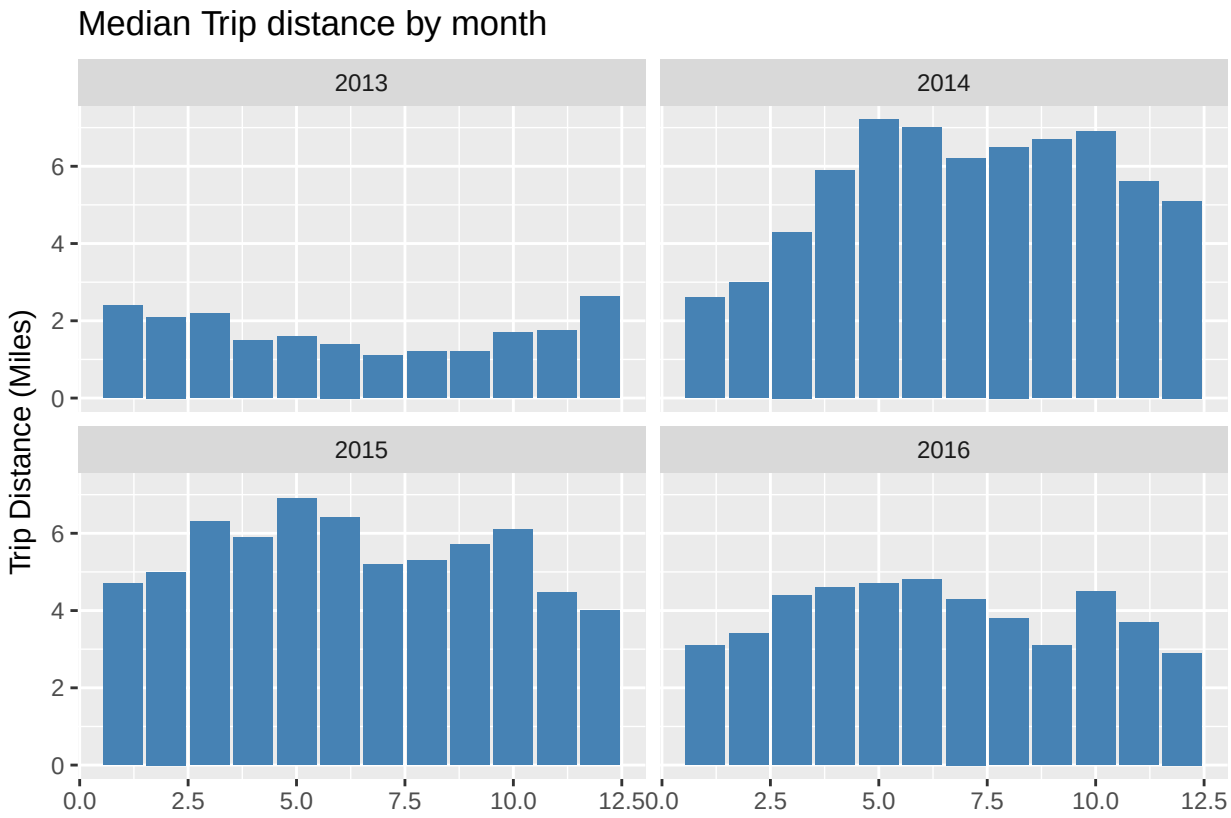
The last variable we want to look into is the trip distance for each month.

```
g4 <- taxi_df[, .(TripDistance = `Trip Miles`, Month), by = Year] %>%
  ggplot(aes(Month, TripDistance)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = 'Month', y = 'Distance', title = 'Distance travelled by month')
g4
```



The total trip distances increased from 2013-2014, but then decreases every year after. January appears to still be the low month, however, May, June, and October don't appear to be distinguishable peaks anymore. Now let's take a look at the median trip distance.

```
g5 <- taxi_df[, .(TotalTrip = sum(`Trip Miles`)), by = list(`Taxi ID`, Month, Year)][, .(Median = media
  ggplot(aes(Month, Median)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = '', y = 'Trip Distance (Miles)', title = 'Median Trip distance by month')
g5
```



The median trip distance shows a large spike from 2013-2014 and then a large decrease from 2015-2016.

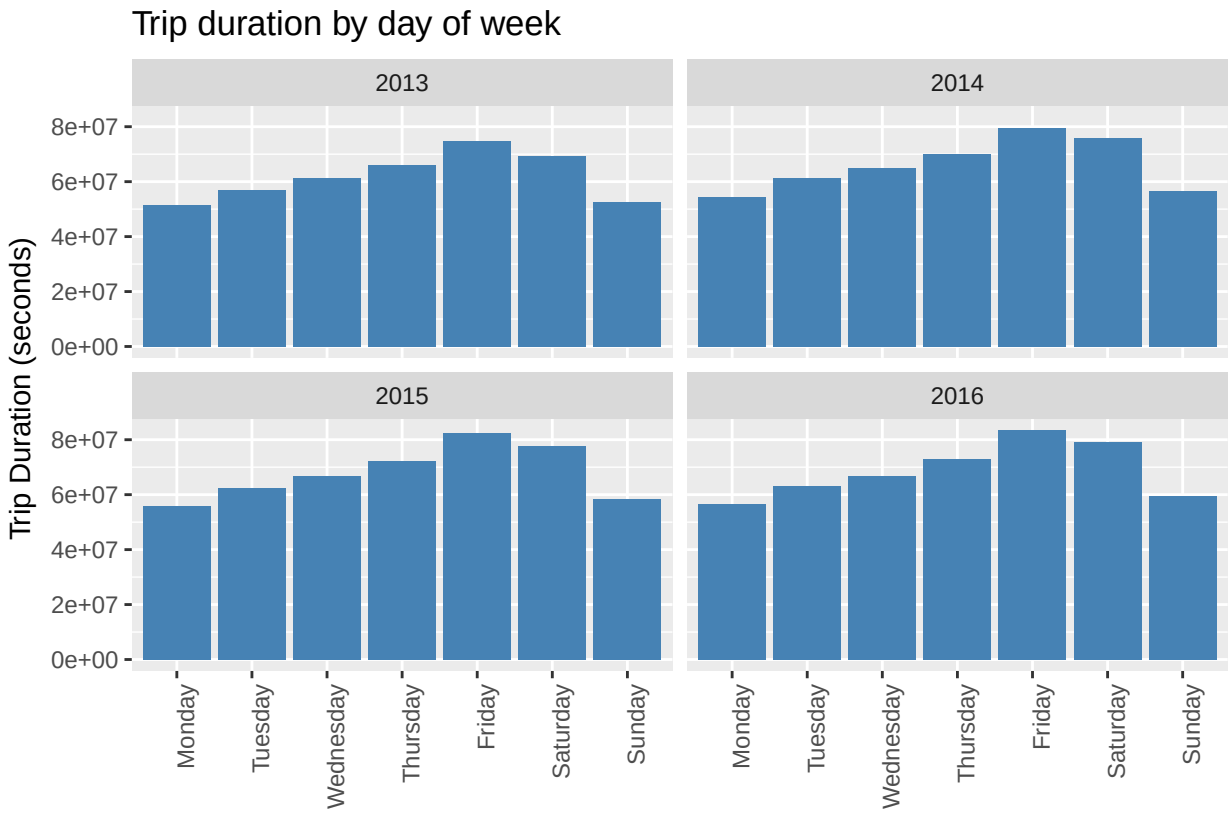
By Day of the week

We now do a closer inspection on the same variables, but now we look at the trends by the day of the week.

Trip Duration

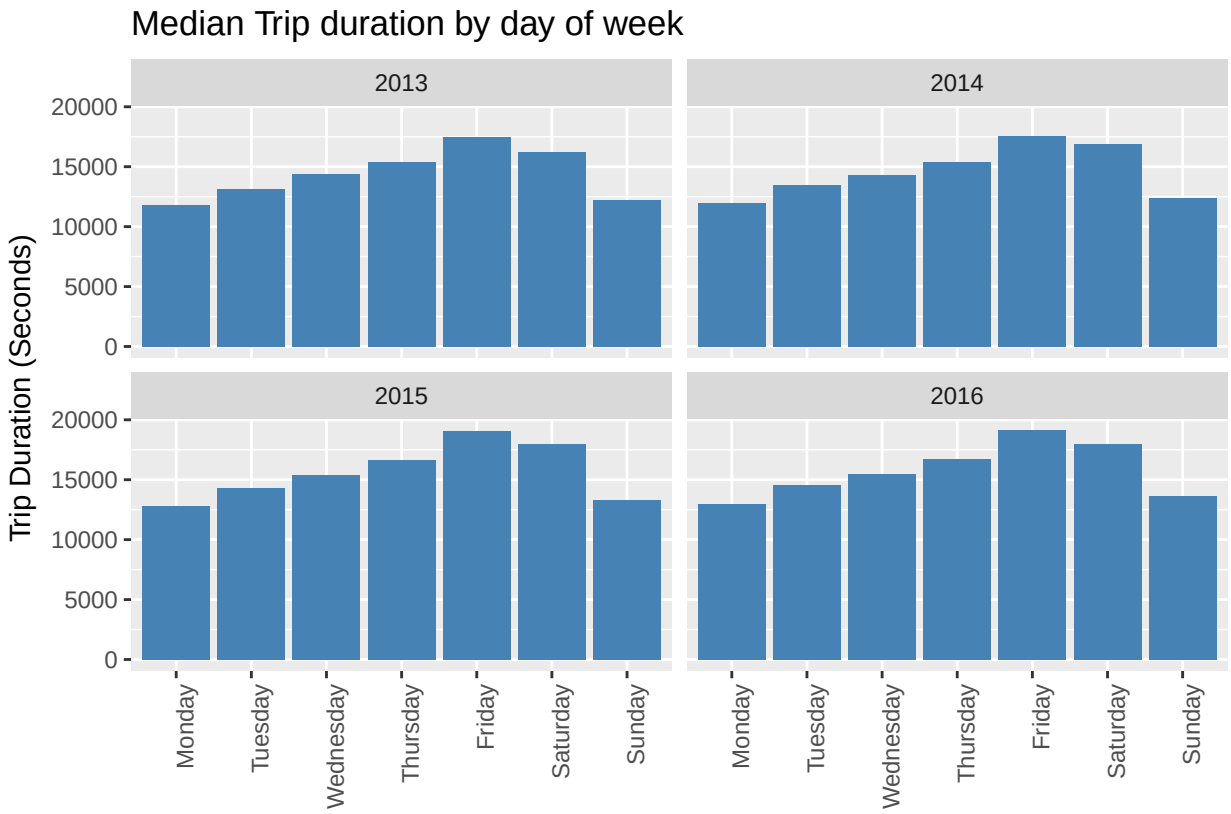
We will start with trip duration again.

```
g6 <- taxi_df[, .(TripDuration = `Trip Seconds`,
                  Day = factor(weekdays(as.Date(date_time)), levels = c('Monday', 'Tuesday', 'Wednesda
  ggplot(aes(Day, TripDuration)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = '', y = 'Trip Duration (seconds)', title = 'Trip duration by day of week') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
g6
```



We see that trips duration is the most on Fridays and Saturdays for all years. Like our previous analysis with trip duration by month, we want to also look at the median trip duration to assess individual rides.

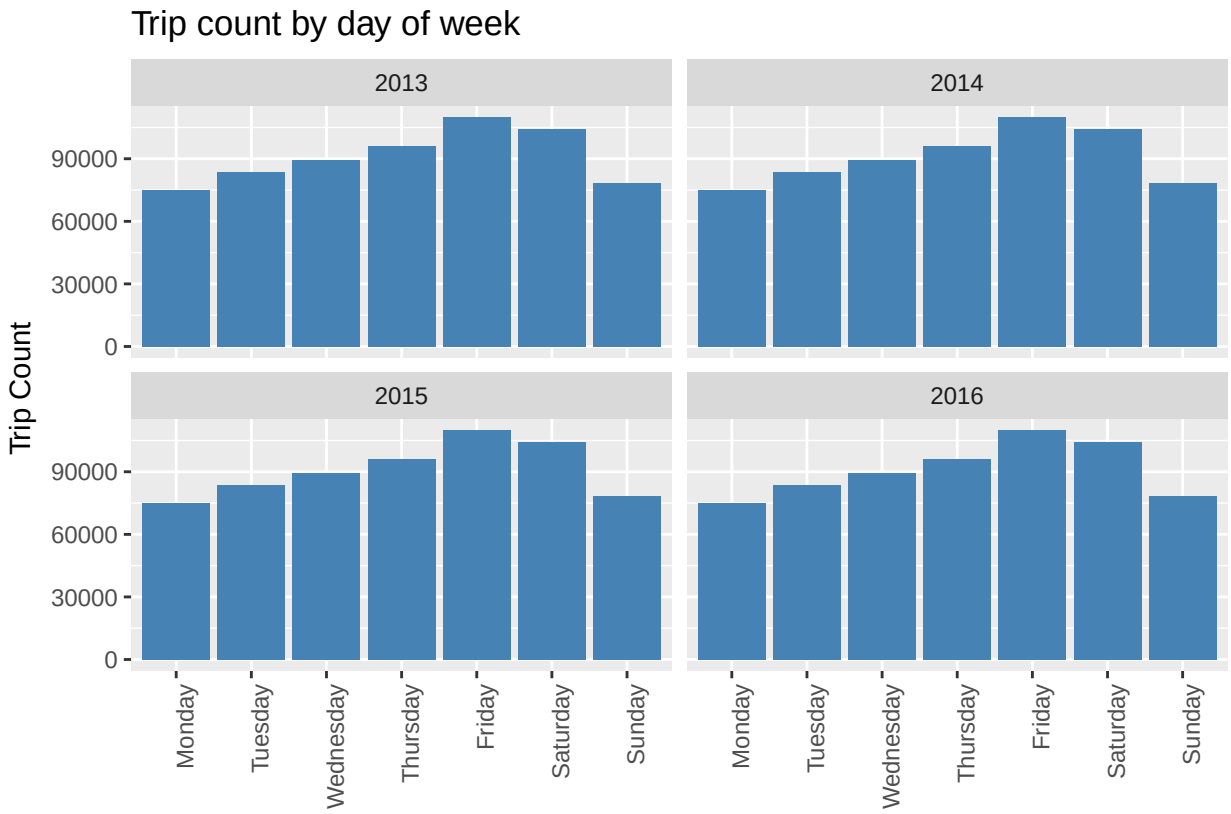
```
g7 <- taxi_df[, .(TaxiID = `Taxi ID`,
                  TripDuration = `Trip Seconds`,
                  Day = factor(weekdays(as.Date(date_time)), levels = c('Monday', 'Tuesday', 'Wednesday
  ggplot(aes(Day, Median)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = '', y = 'Trip Duration (Seconds)', title = 'Median Trip duration by day of week ') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
g7
```



The median trip duration seems to follow the same trend as total trip duration, having the high points on Fridays and Saturdays, so we can expect individual taxi trip durations to behave in this pattern. It makes sense that Friday and Saturday see the longest trip durations as that is when people are more likely to go out and grab drinks or travel to the to and from the airport.

Trip Count

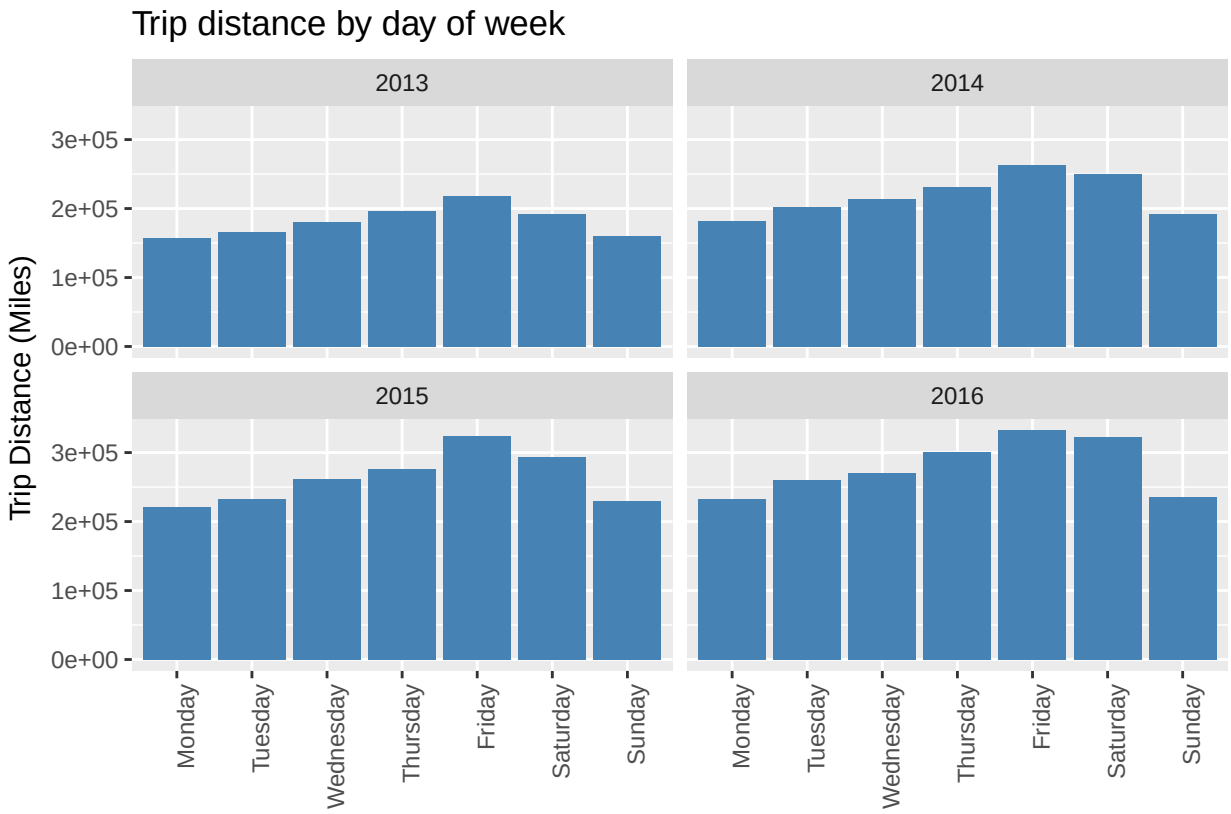
```
g8 <- taxi_df[, .(Day = factor(weekdays(as.Date(date_time)), levels = c('Monday', 'Tuesday', 'Wednesday
  ggplot(aes(Day, TripCount)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = '', y = 'Trip Count', title = 'Trip count by day of week') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
g8
```



As expected, the trip count follows the same trend as the total trip duration and also the median trip duration with the peaks being on Friday and Saturdays.

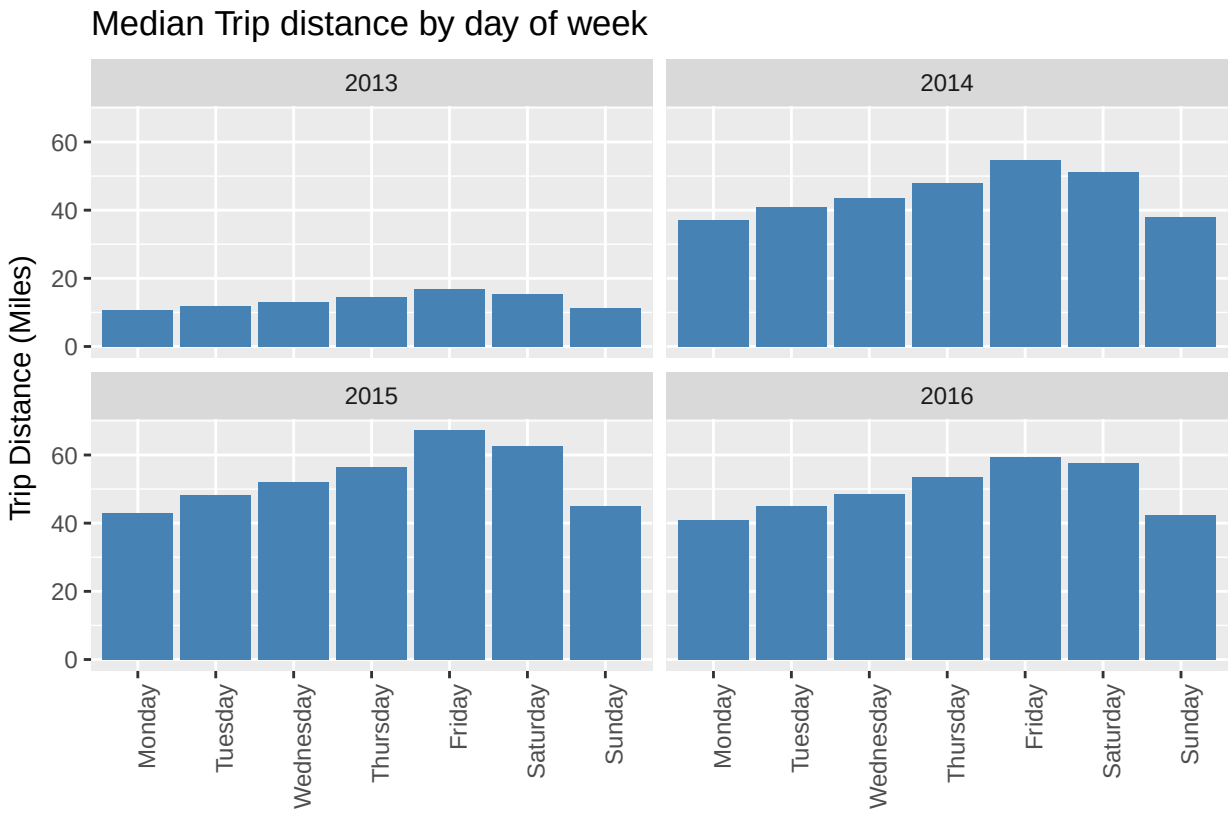
Trip Distance

```
g9 <- taxi_df[, .(TripDistance = `Trip Miles`,
                  Day = factor(weekdays(as.Date(date_time)), levels = c('Monday', 'Tuesday', 'Wednesda
  ggplot(aes(Day, TripDistance)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = '', y = 'Trip Distance (Miles)', title = 'Trip distance by day of week') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
g9
```



The trip distance peaks on Fridays and Saturdays but one interesting thing is the trip distance has seemed to increase over the years while the trip count stayed the same across all years. Let's see if the median trip distance increases also.

```
g10 <- taxi_df[, .(TaxiID = `Taxi ID`,
                   TripDistance = `Trip Miles`,
                   Day = factor(weekdays(as.Date(date_time)), levels = c('Monday', 'Tuesday', 'Wednesday
  ggplot(aes(Day, Median)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = '', y = 'Trip Distance (Miles)', title = 'Median Trip distance by day of week') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
g10
```



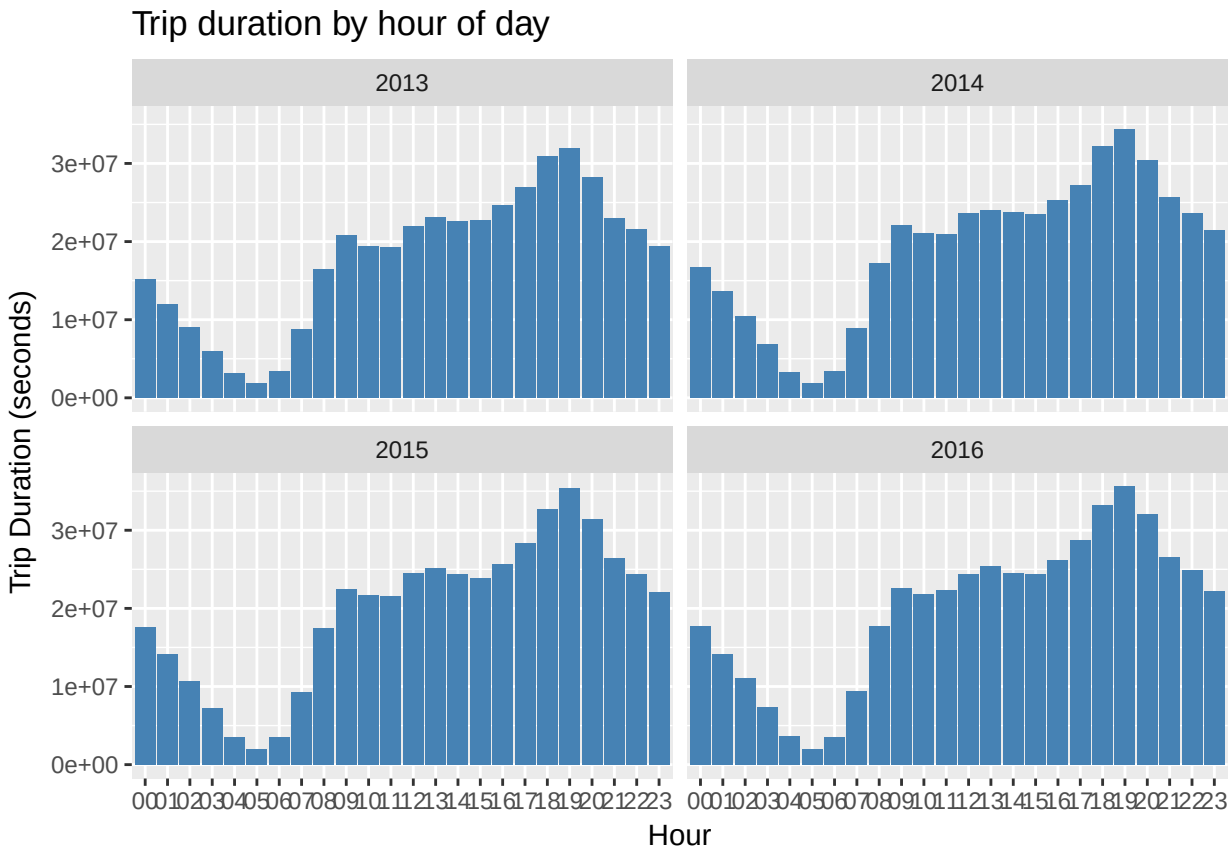
Not much appears to change in this plot except that the 2013 median trip distance is very low.

By Hour of day

Lastly we want to break it down even further and analyze trends by the hour of any day.

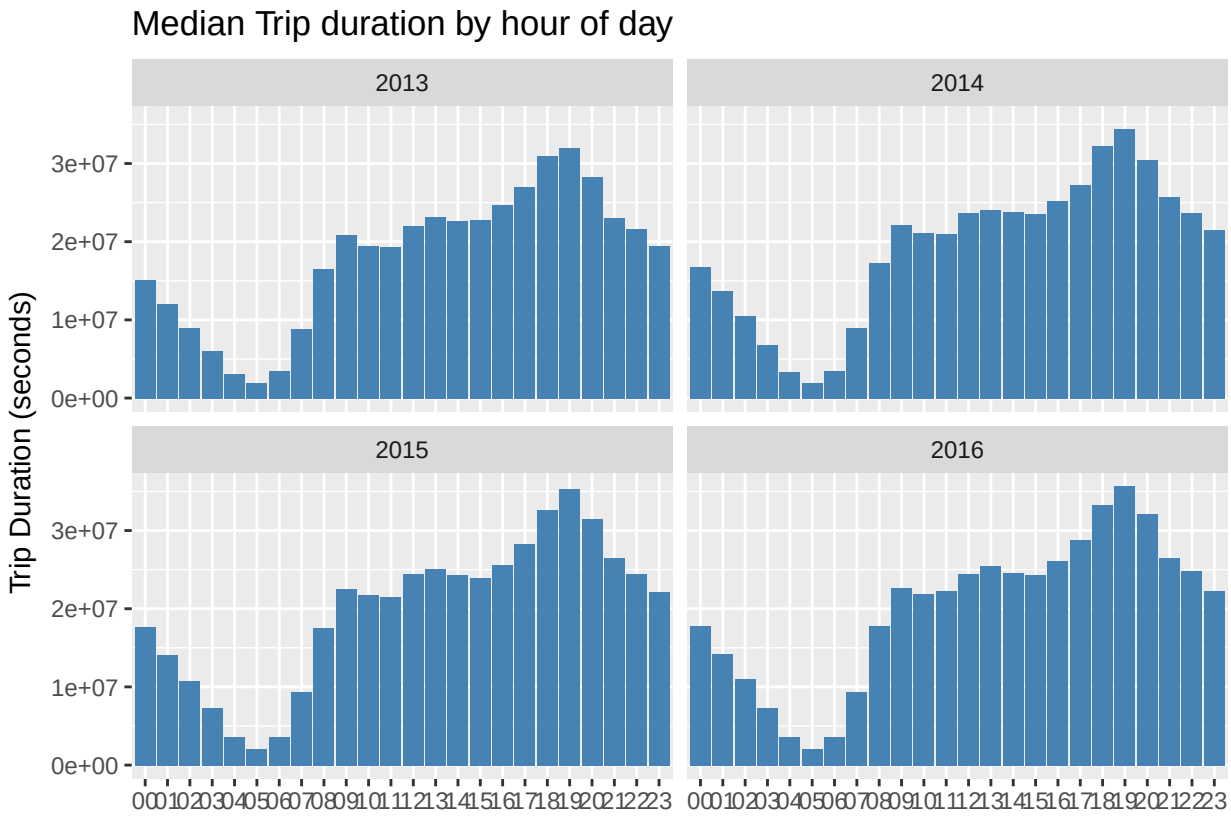
Trip Duration

```
g11 <- taxi_df[, .(TripDuration = `Trip Seconds`, Hour = strftime(date_time, '%H')), by = Year] %>%
  ggplot(aes(Hour, TripDuration)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = 'Hour', y = 'Trip Duration (seconds)', title = 'Trip duration by hour of day')
g11
```



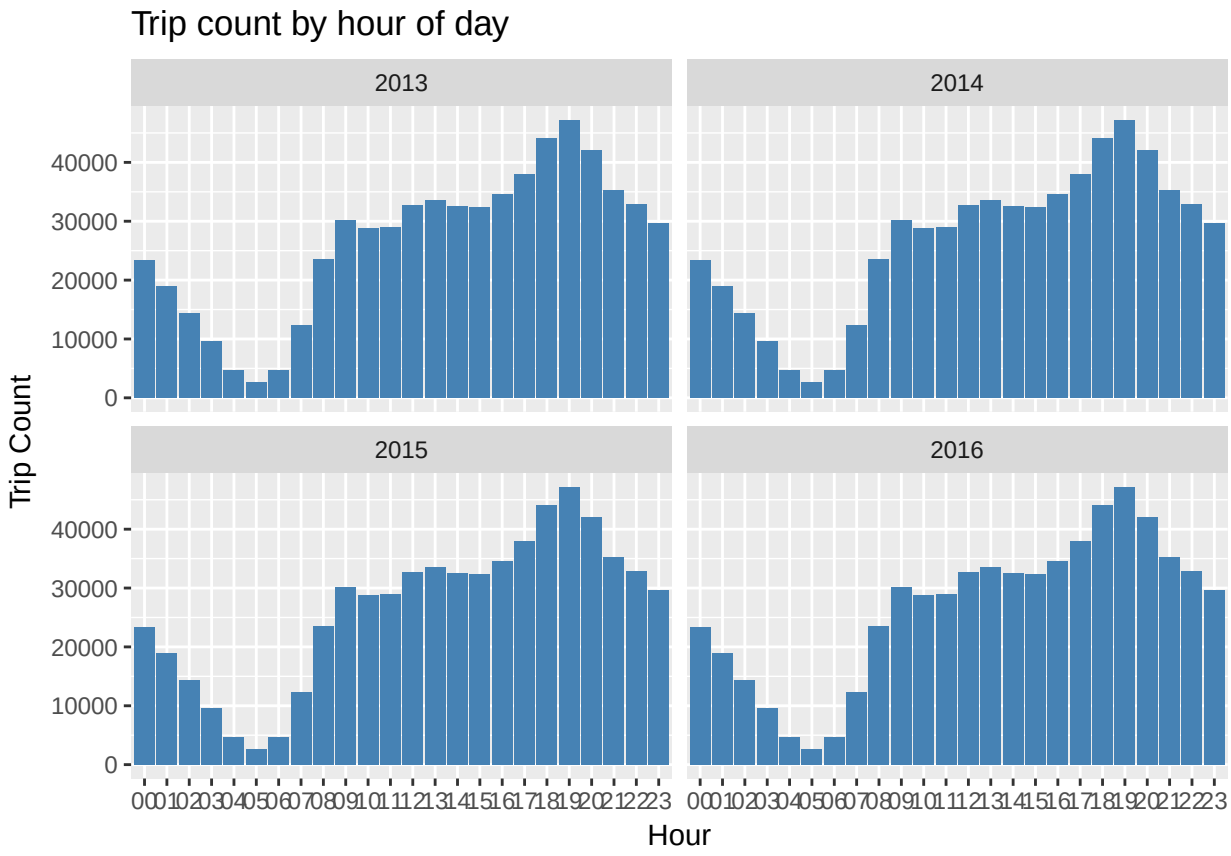
6pm and 7pm appears to be the peaks for trip duration for all years and there appears to be only a slight increase in trip duration after 2013. Let's take a look at the median trip duration.

```
g12 <- taxi_df[, .(TripDuration = `Trip Seconds`,
                   Hour = strftime(date_time, '%H')), by = Year] %>%
  ggplot(aes(Hour, TripDuration)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = '', y = 'Trip Duration (seconds)', title = 'Median Trip duration by hour of day')
g12
```



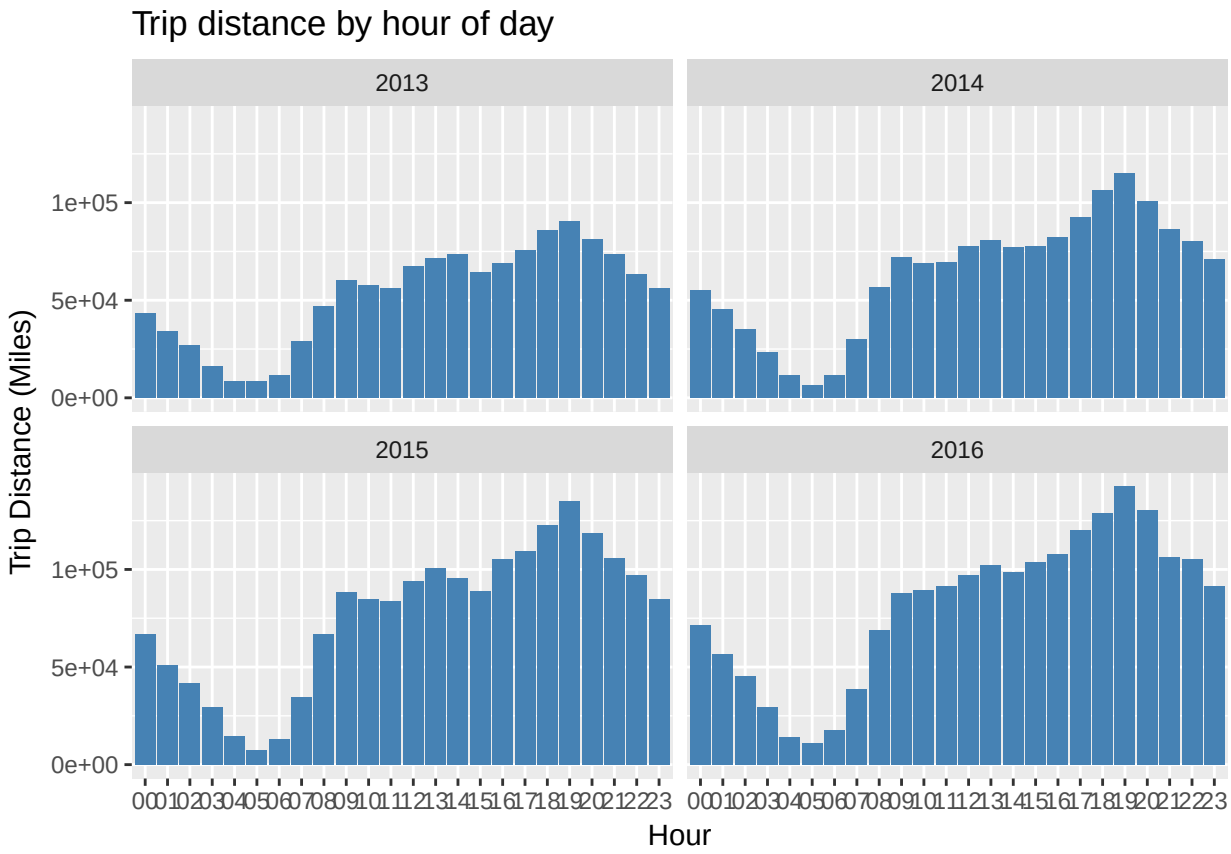
Trip Count

```
g13 <- taxi_df[, .(Hour = strftime(date_time, '%H')), by = Year][, .(TripCount = .N), by = c('Year', 'H
  ggplot(aes(Hour, TripCount)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = 'Hour', y = 'Trip Count', title = 'Trip count by hour of day')
g13
```



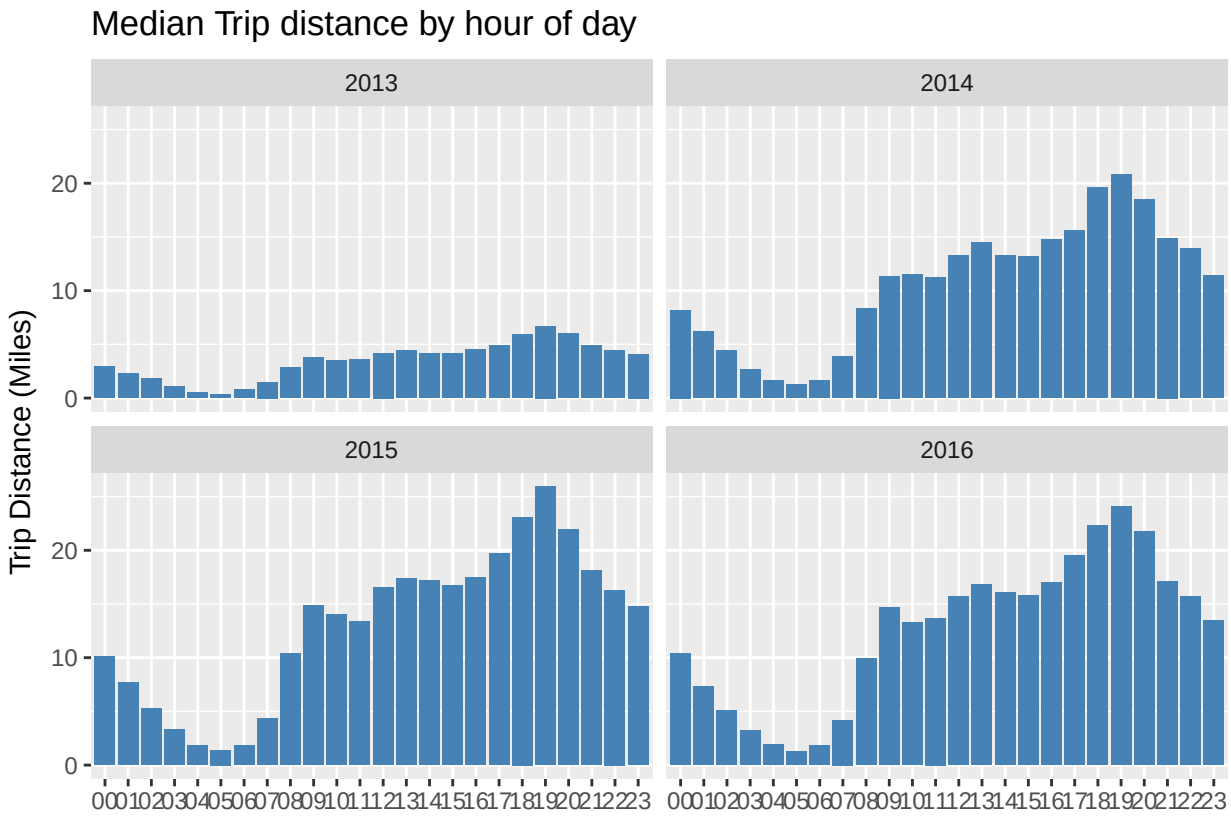
Trip Distance

```
g14 <- taxi_df[, .(TripDistance = `Trip Miles`, Hour = strftime(date_time, '%H')), by = Year] %>%
  ggplot(aes(Hour, TripDistance)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = 'Hour', y = 'Trip Distance (Miles)', title = 'Trip distance by hour of day')
g14
```



And the median trip distance.

```
g15 <- taxi_df[, .(TaxiID = `Taxi ID`,
                   TripDistance = `Trip Miles`,
                   Hour = strftime(date_time, '%H')), by = Year][, .(TotalTrip = sum(TripDistance)), by =
  ggplot(aes(Hour, Median)) +
  geom_bar(stat = 'identity', fill = 'steelblue') +
  facet_wrap(~ as.factor(Year)) +
  labs(x = '', y = 'Trip Distance (Miles)', title = 'Median Trip distance by hour of day')
g15
```


Median Trip distance by hour of day
2013
2014
2015
2016
Trip Distance (Miles)
0
10
20
00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23