Course Final Project

Chris Hilger

10-19-17

Abstract (typically 250 words or less)

In a few short years, ridesharing companies like Uber have revolutionized the way people travel in cities. However, I was curious to know if there were any trends how individuals use Uber. This study investigated the frequency of Uber pickups by time of day, day of week, and lat/long location in the city of New York during April – September 2015. In all, the dataset contains over 4 million Uber pickups and was obtained through a Kaggle dataset acquired via a Freedom of Information Act request.

Data exploration techniques were used and then heatmaps of pickup locations New York City on Weekday Mornings (Monday, Tuesday, Wednesday, 6-8 AM) vs Party Evenings (Thursday, Friday, Saturday, 9-11 PM) were created.

I was able to find which day and time are most popular, and which locations in New York City are the most popular pickup points on Weekday Mornings vs Party Evenings.

From these, we know have a better understanding of how customers use Uber to travel around New York City.

Introduction (1-2 paragraphs)

This data set contains over 4 million rides in New York City during April – September 2015. With so many rides, it may be possible to draw insights regarding how the traffic flows throughout the city at specific points in time. From knowing this, one could better predict traffic in the city, and where to instruct Uber drivers to travel to in order to increase efficiency of the Uber network.

Since the dataset was posted on Kaggle, a lot of work has been done to investigate trends in the data. These include simple histograms of ride frequency by time of day and day of the week, and also visualizations to answer more complex questions such as if Uber is taking away rides from Yellow Cab Taxis <https://fivethirtyeight.com/features/uber-is-taking-millions-of-manhattan-rides-away-from-taxis/>

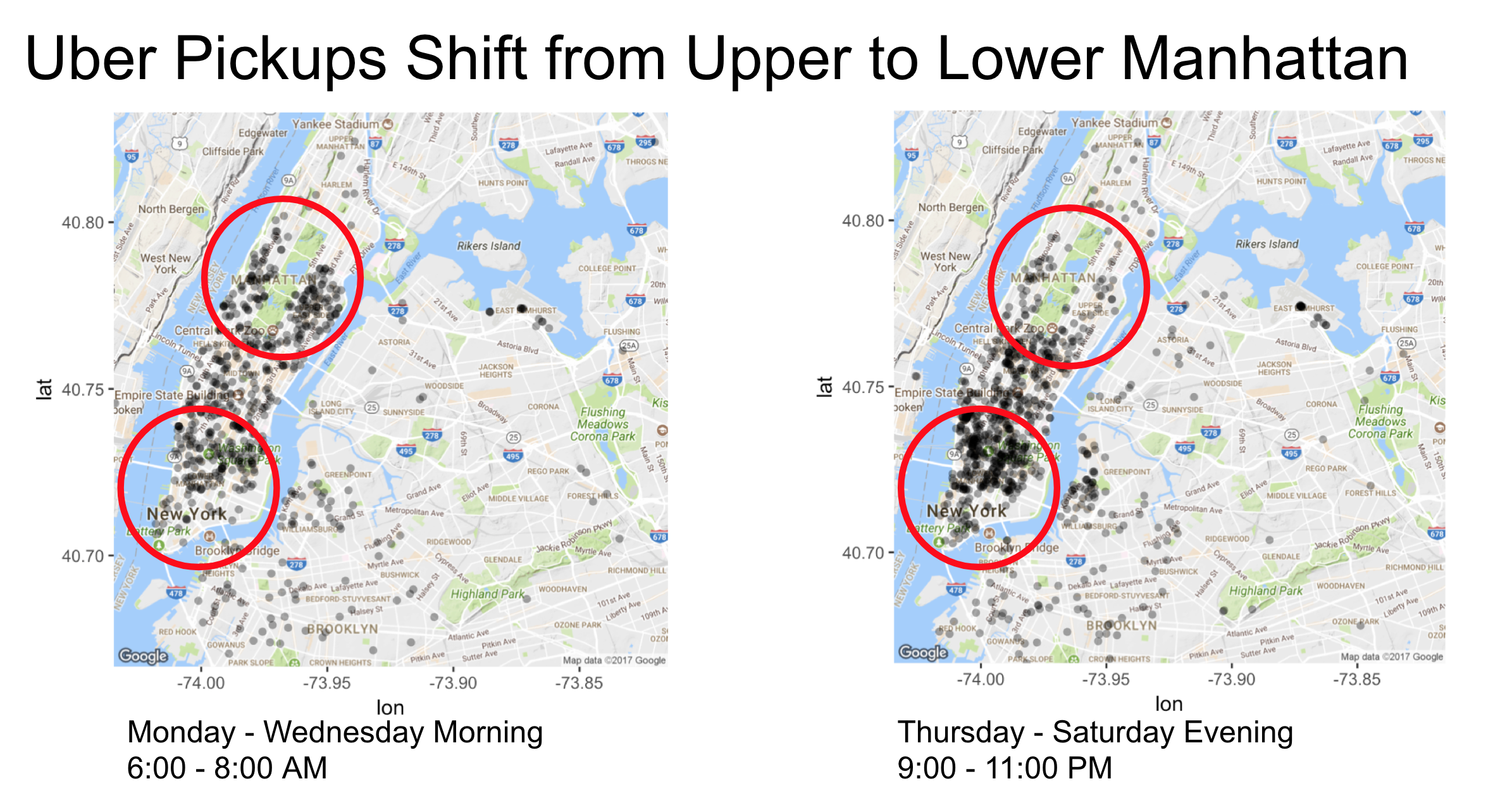
However, none of the existing analysis asked the simple question about where traffic flows in the city at one part of day to the next. The goal of this project was to see how traffic flows between the period of Weekday Mornings and Party Evenings.

Methods (2-4 paragraphs with code interspersed)

The source of the data was a Kaggle dataset obtained via a FOIL request <https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city> The date of collection for the data was April – September 2015. The data were automatically collected via the Uber app. A FOIL request then forced Uber to share this data with the public. The total sample size included 4 million uber rides, offering pickup date, time and lat/long of pickup. The main output is a general visualization of how pickup locations change over time, laid on top of a map. Some covariates might have been major events in the city, weather, and traffic abnormalities (accidents). Luckily this was a very clean data set. I did not have missing values, however, I did have to spend a considerable amount of time parsing the data/time data into a useable format. No hard statistical methods were used for this project, since I was creating a general visualization instead of a hard statistical output.

Results

The main output of interest was to see how pickup locations moved in the city based on the time of day and day of week. Specifically, I investigated Monday - Wednesday Mornings, 6:00 - 8:00 AM and Thursday – Saturday Evenings, 9:00 – 11:00 PM. As you can see in the visuals below, the morning pickups at the beginning of the week are concentrated much more in the Upper East Side and Upper West Side, while the pickups in the weekdays toward the end of the week in the evening are located more in lower Manhattan.



The code to generate these plots is below:

partyevenings <- sampledata %>%

filter(dayofweek %in% c("Thurs", "Fri", "Sat") & hour %in% c(21, 22, 23))

mapPoints <- ggmap(map) +

geom\_point(data = partyevenings, size = 2 , stroke = 0, shape = 16, alpha = .4, aes(x = Lon, y = Lat), show.legend = FALSE)

mapPoints

workmornings <- sampledata %>%

filter(dayofweek %in% c("Mon", "Tues", "Wed") & hour %in% c(6, 7, 8))

#map <- get\_map(location = 'New York City', zoom = 12)

mapPoints <- ggmap(map) +

geom\_point(data = workmornings, size = 2 , stroke = 0, shape = 16, alpha = .4, aes(x = Lon, y = Lat), show.legend = FALSE)

mapPoints

Discussion (1-2 paragraphs)

From this visual, it is clear that yes, pickup locations do vary depending of the time of day and day of the week. Specifically, more pickups are made on in the Upper East Side and Upper West side during the mornings early in the week. Then, the pickups shift to lower Manhattan for evenings later in the week and on the weekend. This is important because it shows that if you are a uber driver, you should plan to make more pickups in the Upper East and West side in the mornings early in the week, and then more pickups in lower Manhattan if you are working later in the week in the evenings.

In the future, I could include weather data to see weather has a major impact of number of pickups at a given time or location of specific pickups. For example, when it is raining, there might be a spike in pickups, particularly in locations around the city that are not near subway stations. The idea being that if it starts to rain, and there isn’t a subway stop around, people will request an Uber to stay dry.

References

Kaggle data set: https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city

Project Code

```{r}

#load in all these libraries

library(ggplot2)

library(gdata)

library(dplyr)

library(lubridate)

library(scales)

library(tidyr)

library(DT)

library(ggthemes)

library(ggmap)

```

---

title: "Final Project"

output: html\_document

---

```{r}

library(gdata)

#read in each csv file, and then bind it into a massive data frame

mydata1 <- read.csv("uber-raw-data-apr14.csv")

mydata2 <- read.csv("uber-raw-data-aug14.csv")

mydata3 <- read.csv("uber-raw-data-jul14.csv")

mydata4 <- read.csv("uber-raw-data-jun14.csv")

mydata5 <- read.csv("uber-raw-data-may14.csv")

mydata6 <- read.csv("uber-raw-data-sep14.csv")

alldata <- rbind(mydata1, mydata2, mydata3, mydata4, mydata5, mydata6)

```

```{r}

#properly parse out the date and time, and save each relevent piece of the data to a new column

alldata$Date.Time <- as.POSIXct(alldata$Date.Time, format = "%m/%d/%Y %H:%M:%S")

alldata$Time <- format(as.POSIXct(alldata$Date.Time, format = "%m/%d/%Y %H:%M:%S"), format="%H:%M:%S")

alldata$Date.Time <- ymd\_hms(alldata$Date.Time)

alldata$day <- factor(day(alldata$Date.Time))

alldata$month <- factor(month(alldata$Date.Time, label = TRUE))

alldata$year <- factor(year(alldata$Date.Time))

alldata$dayofweek <- factor(wday(alldata$Date.Time, label = TRUE))

alldata$hour <- factor(hour(hms(alldata$Time)))

alldata$minute <- factor(minute(hms(alldata$Time)))

alldata$second <- factor(second(hms(alldata$Time)))

```

```{r}

#plot of all overall number of rides for the months we have data for

ggplot(data = alldata) +

geom\_bar(mapping = aes(x=month)) +

ggtitle("Trips by Month")

```

```{r}

#plot of number of trips by hour

ggplot(data = alldata) +

geom\_bar(mapping = aes(x=hour)) +

ggtitle("Trips by hour")

```

```{r}

#plot of number of trips by the day of the week

ggplot(data = alldata) +

geom\_bar(mapping = aes(x=dayofweek)) +

ggtitle("Trips by day of week")

```

```{r}

#combined visual looking at day of the week and time of day. This reveals a bimodal distribution for some days

ggplot(data = alldata) +

geom\_bar(mapping = aes(x=hour)) +

ggtitle("Trips by day of week by hour") +

facet\_wrap(~dayofweek, ncol = 2) +

scale\_x\_discrete(breaks=seq(0,24,2))

```

```{r}

#add the count of entries for each month to the data frame

alldata$weights <- ""

alldata$weights[alldata$month=="Apr"] <- nrow(mydata1)

alldata$weights[alldata$month=="Aug"] <- nrow(mydata2)

alldata$weights[alldata$month=="Jul"] <- nrow(mydata3)

alldata$weights[alldata$month=="Jun"] <- nrow(mydata4)

alldata$weights[alldata$month=="May"] <- nrow(mydata5)

alldata$weights[alldata$month=="Sep"] <- nrow(mydata6)

```

```{r}

#sample a portion of the data in proportion to the count of pickups for that month

set.seed(300)

sampledata <- alldata %>% sample\_n(size = 10000, weight = as.numeric(weights))

```

```{r}

#draw the map with a sample of the pickup points

library(ggmap)

bbox <- make\_bbox(c(-74.05,-73.8), c(40.68, 40.82), f = 0.01)

map <- get\_map(bbox)

mapPoints <- ggmap(map) +

geom\_point(data = sampledata, size = 1, stroke = 0, shape = 16, alpha = .3, aes(x = Lon, y = Lat), show.legend = FALSE)

mapPoints

```

```{r}

#Map for party evenings, Thurs-Saturday nights

partyevenings <- sampledata %>%

filter(dayofweek %in% c("Thurs", "Fri", "Sat") & hour %in% c(21, 22, 23))

mapPoints <- ggmap(map) +

geom\_point(data = partyevenings, size = 2 , stroke = 0, shape = 16, alpha = .4, aes(x = Lon, y = Lat), show.legend = FALSE)

mapPoints

```

```{r}

#Draw the map for workday mornings, Mon-Wed 6-8 AM

workmornings <- sampledata %>%

filter(dayofweek %in% c("Mon", "Tues", "Wed") & hour %in% c(6, 7, 8))

mapPoints <- ggmap(map) +

geom\_point(data = workmornings, size = 2 , stroke = 0, shape = 16, alpha = .4, aes(x = Lon, y = Lat), show.legend = FALSE)

mapPoints

```