Hubway Visualization Challenge

Nam Pham

December 18, 2016

```
library(knitr)
library(dplyr)
library(readr)
library(ggplot2)
library(maptools)
#library(RNeo4j)
library(stringr)
library(ggmap)
library(kableExtra)
```

```
csv.folder = "C:/Users/namhpham/Documents/Personal files/R workspace"
file.stations = paste(csv.folder, "hubway_stations.csv", sep="/")
file.trips    = paste(csv.folder, "hubway_trips.csv"   , sep="/")
df.stations = read_csv(file=file.stations)
df.trips    = read_csv(file=file.trips)
```

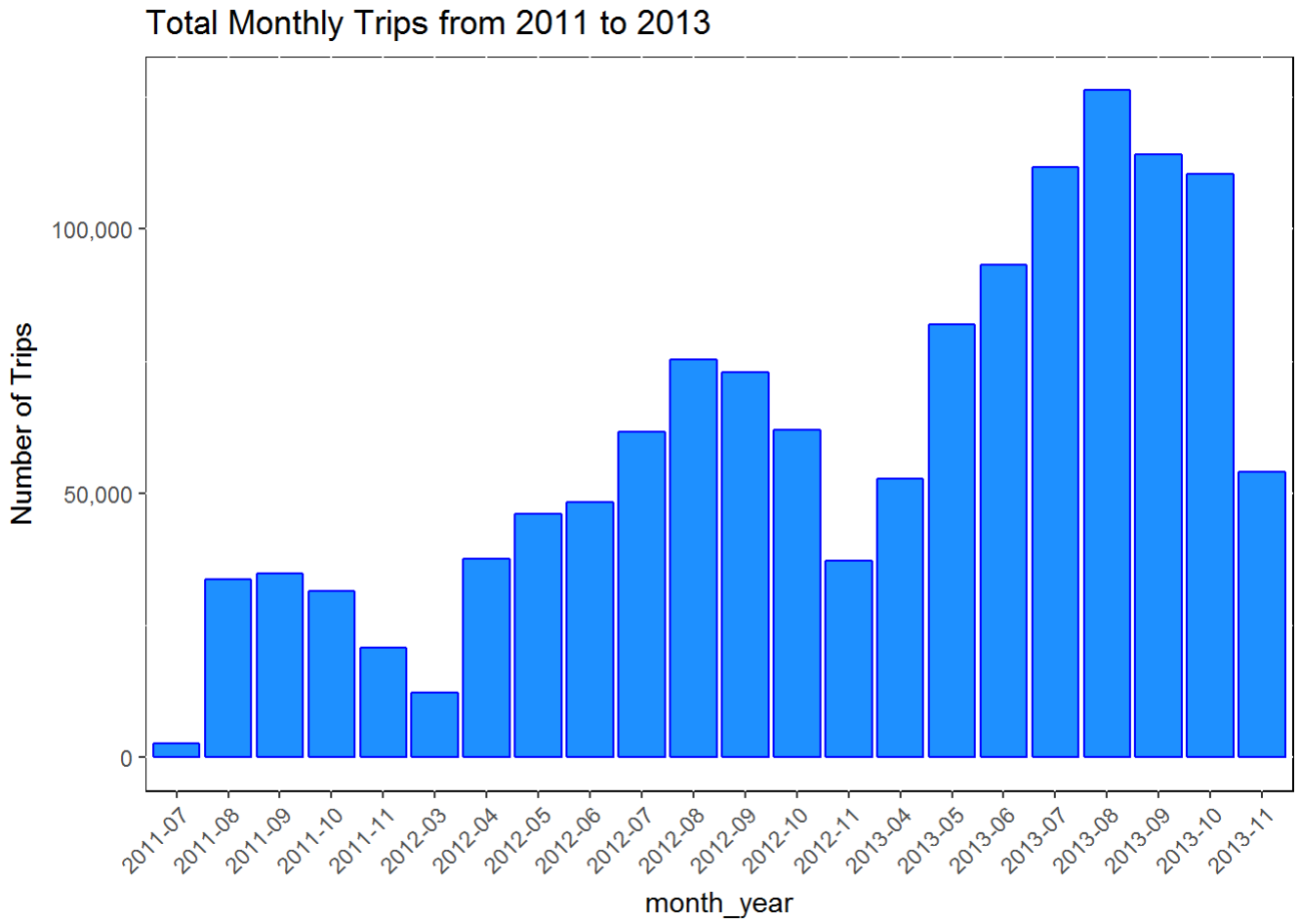
```
df.trips %>%
  filter(!is.na(strt_statn),
         !is.na( end_statn),
         duration > 300) %>%           #include trip > 1 minute
  mutate(age=2013-birth_date) %>%     #add age variable at the time of dataset
  mutate(raw.date=as.Date(gsub( " .*$", "", start_date), "%m/%d/%Y")) %>%
  #extract the date before the white space and convert to Date type
  mutate(duration=duration/60)%>% #convert duration to minutes
  mutate(zip_code=str_pad((gsub("'", '',zip_code)), 5, pad="0")) %>%
  #remove ' in the beginning and add a leading 0 if zipcode has less than 4 letter
  select(seq_id, hubway_id, strt_statn, end_statn,
         start_date, end_date, subsc_type,
         duration, zip_code, birth_date, gender, age, raw.date, bike_nr)%>%
         {.}->df.trips
```

```
df.trips%>%
  group_by(end_statn)%>%
  summarize(incoming_trip=length(end_statn))-> count_trips
  #count number of times that end_station appears, grouped by itself

df.stations%>%
  mutate(incoming_trip=count_trips$incoming_trip)%>%
  #add new column of incoming trips to existing station, because station_id in both tables are a
rranged from 3->145, making it possible to merge
  {.}->df.stations.graph
```

Greater Boston area is adopting Hubway biking system. Since 2011, number of trips has increased over time, with heavy concentration in summer months

```
df.trips %>%
  mutate(month_year=format(raw.date,"%Y-%m")) %>%
  ggplot(aes(x=month_year,y=(unique(seq_id))))+
  #scale_x_date(format="%Y-%m")+
  stat_summary(fun.y=length,geom="bar",fill="dodgerblue1",color="blue")  +
  #count the length of unique trip id, which is total trips
  theme(axis.text.x=element_text(angle=45,hjust=1))+
  #rotate the x-axis by 90 degree for easier viewing
  ggtitle("Total Monthly Trips from 2011 to 2013") +
  scale_y_continuous(labels = scales::comma) +
  theme(panel.background = element_rect(fill='white', colour='black'))+
  ylab("Number of Trips")
```



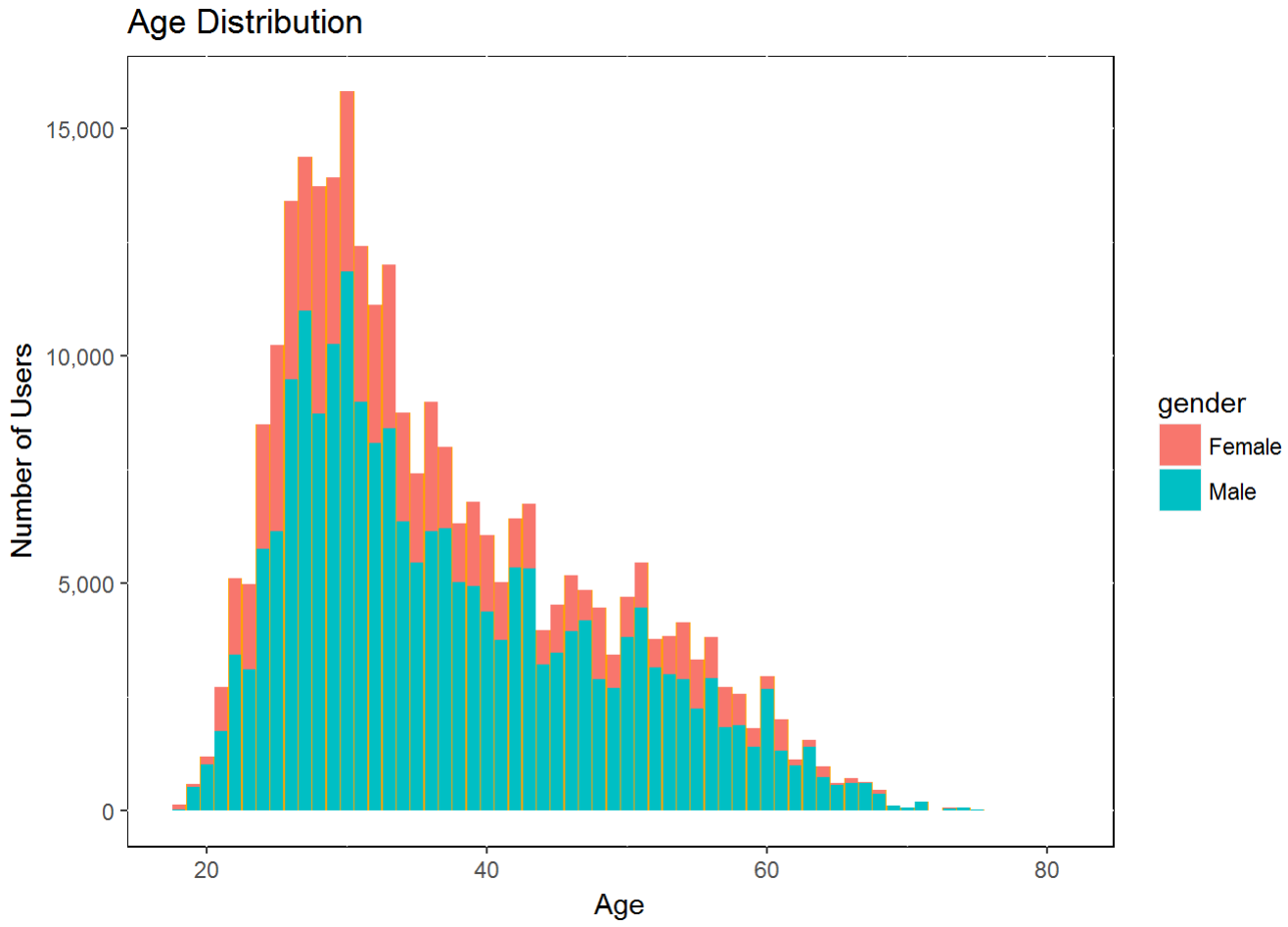
Hubway usage time is greater on weekend, while weekday trips are shorter

Days_of_Week	Average_Minutes	Average Trips
Saturday	28.78	1563

Days_of_Week	Average_Minutes	Average Trips
Sunday	28.50	1414
Friday	21.84	1569
Monday	21.65	1546
Thursday	18.52	1576
Wednesday	18.37	1597
Tuesday	18.20	1540

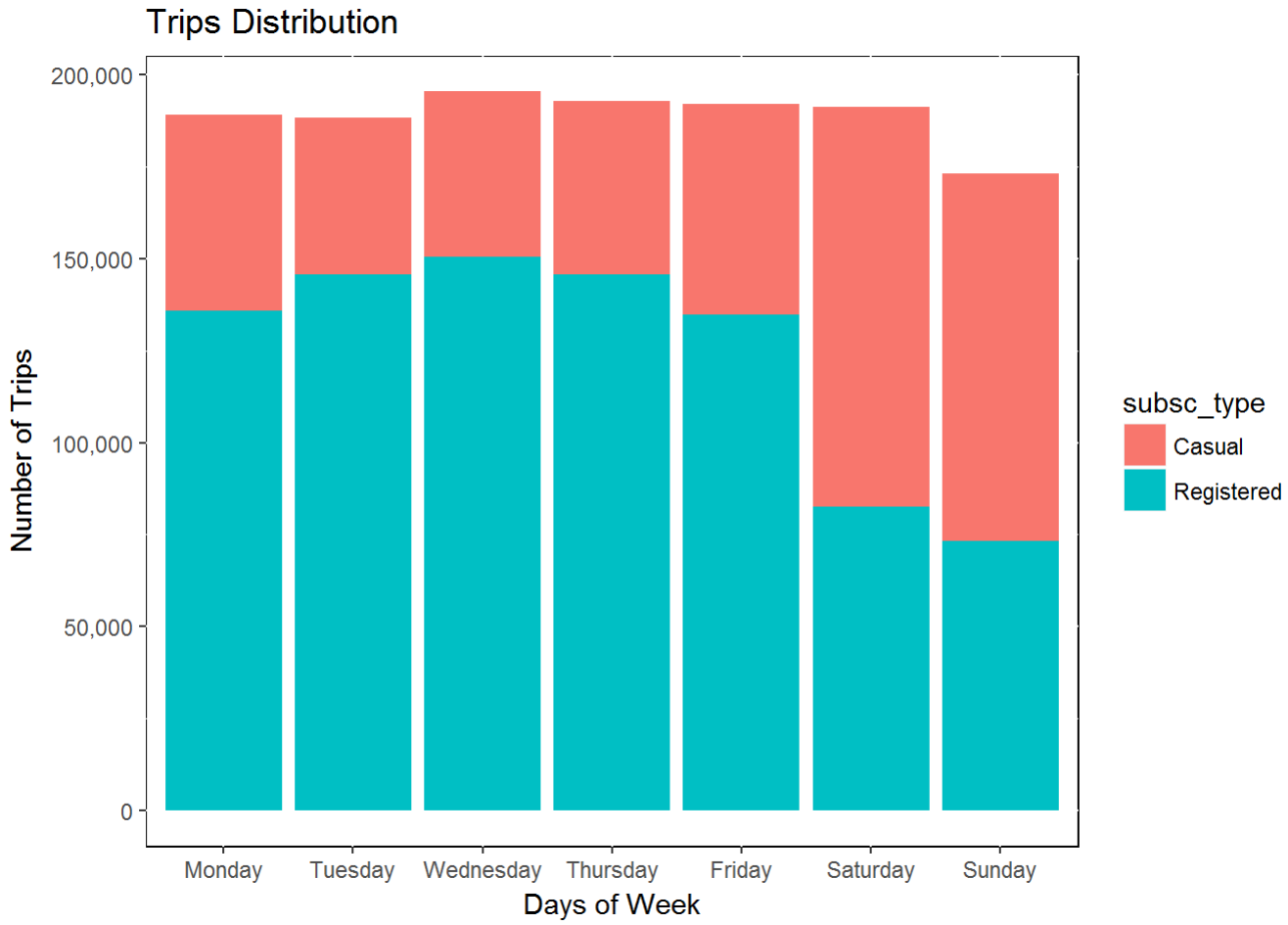
Typical users of Hubway biking system are males in their 30s

```
df.trips %>%
  filter(!is.na(gender),!is.na(age))%>%
  ggplot(aes(age)) +
  geom_histogram(binwidth=1, fill="orange")+
  #geom_density(col="blue")+
  ylab("Number of Users")+
  xlab("Age")+
  geom_bar(aes(fill=gender)) +
  theme(panel.background = element_rect(fill='white', colour='black'))+
  scale_y_continuous(labels = scales::comma) +
  ggtitle("Age Distribution")
```



Most weekday bikers are registered users, who presumably use Hubway frequently to commute to work. Weekend trips are more utilized by one-time, casual users

```
df.trips %>%
  #mutate(duration_type=ifelse(df.trips$duration >=30, "Long Trip", "< 30 min"))%>%
  mutate(Weekday=weekdays(raw.date)) %>% #extract weekday from raw date
  #if duration is longer than 20 minutes, then "Long trip", otherwise it is short
  ggplot  (aes(x   =Weekday))+
  scale_x_discrete(limits=c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunda
y"))+
  geom_bar(aes(fill=subsc_type)) +
  ylab("Number of Trips")+
  scale_y_continuous(labels = scales::comma) +
  xlab("Days of Week") +
  theme(panel.background = element_rect(fill='white', colour='black'))+
  ggtitle("Trips Distribution")
```



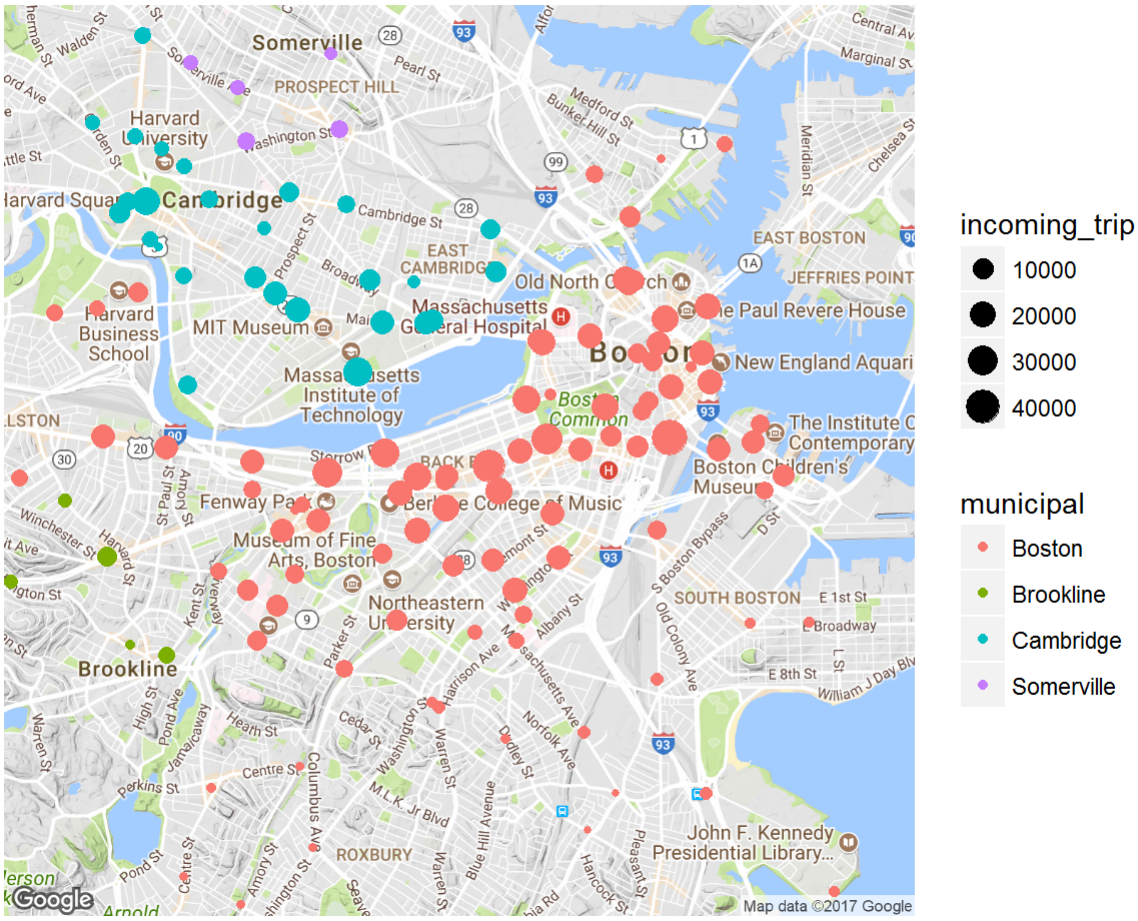
Boston and Cambridge are the hotspots for Hubway. These two cities account for the highest incoming trips, from both out-of-city and within-city

```
qmap(location = "Back Bay, MA", zoom = 13, maptype = "terrain") + geom_point(data = df.stations.
graph,
    aes(x = lng, y = lat, col = municipal, size = incoming_trip)) + ggtitle("Top Destinations fo
r Hubway users")
```

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=Back+Bay,+MA&zoom=13&size
=640x640&scale=2&maptype=terrain&language=en-EN&sensor=false
```

```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=Back%20Bay,%2
0MA&sensor=false
```

Top Destinations for Hubway users



Biking is especially popular along Commonwealth Ave, where protected bike lanes are installed

Improving the System

- Add more bikes to Somerville and Brookline areas
- Install more bike racks in Financial District/South Station (Boston) and Harvard/MIT (Cambridge), the stations with top incoming traffics
- Expand the Hubway system along bike trails and tourist attractions, such as Northern Strand Community Trail (Bike to the Sea Trail)