DSO545-Homework5

Xu Zhang

April 14, 2018

Case1

(1)Use rvest to scrape the table of the 100 most viewed YouTube videos from the following Wikipedia page (link: https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos).

```
page="https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos"

youtube = page %>%
  read_html()%>%
  html_nodes(xpath = '//*[@id="mw-content-text"]/div/table[1]') %>%
  html_table(fill = TRUE)

youtube = youtube[[1]]
youtube_new=youtube[1:100,1:6]
head(youtube_new)
```

```
##   Rank                                      Video name
## 1   1.                                  "Despacito"[9]
## 2   2.                               "See You Again"[14]
## 3   3.                                "Shape of You"[21]
## 4   4.                               "Gangnam Style"[22]
## 5   5.                                 "Uptown Funk"[27]
## 6   6. "Masha and the Bear: Recipe for Disaster"[28]
##                     Uploader / artist Views (billions)       Upload date
## 1  Luis Fonsi featuring Daddy Yankee             5.04  January 12, 2017
## 2 Wiz Khalifa featuring Charlie Puth             3.51     April 6, 2015
## 3                          Ed Sheeran             3.44  January 30, 2017
## 4                                 Psy             3.14     July 15, 2012
## 5   Mark Ronson featuring Bruno Mars             3.01 November 19, 2014
## 6                          Get Movies             2.98  January 31, 2012
##   Notes
## 1   [B]
## 2   [C]
## 3   [D]
## 4   [E]
## 5   [F]
## 6   [G]
```

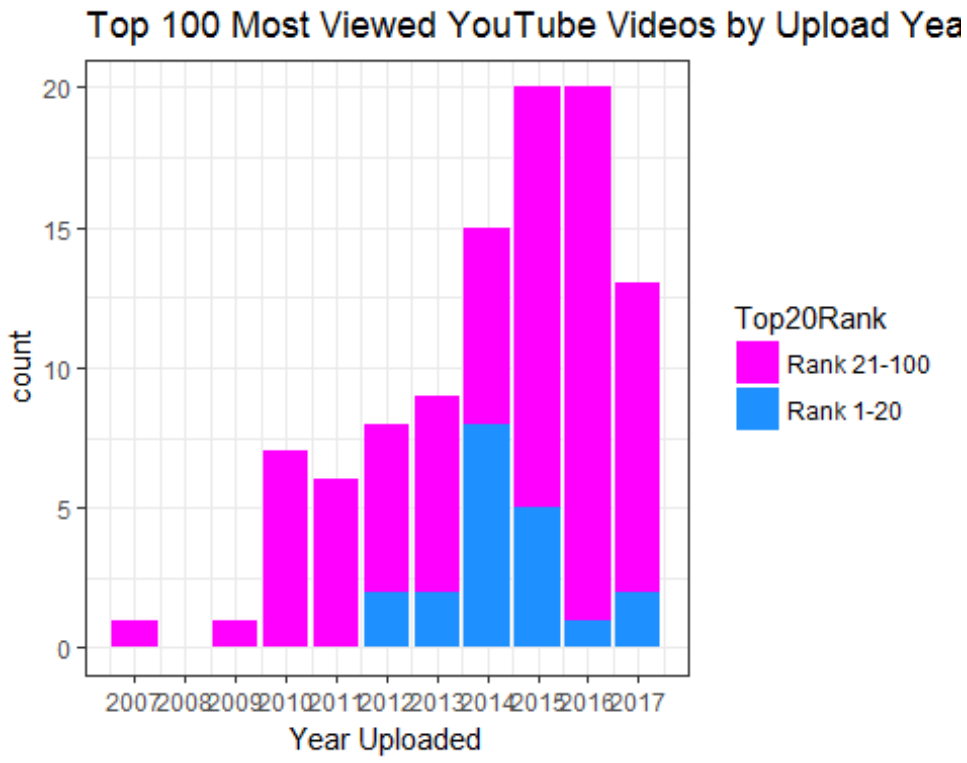
(2) How many of the videos in the top 100 include Justin Bieber as the uploader / artist?

```
nrow(youtube_new %>%
       filter(`Uploader / artist`=="Justin Bieber"))
```

```
## [1] 3
```

(3) Create an EXACT copy of the following graph. Use colors magenta and dodgerblue, and for the theme use theme_bw().

```
youtube_new$Rank=as.numeric(youtube_new$Rank)
youtube_new$`Upload date`=mdy(youtube_new$`Upload date`)

youtube_new2=youtube_new %>%
  mutate(Year=year(`Upload date`),
         Top20Rank=ifelse(Rank<=20,"Rank 1-20","Rank 21-100")) %>%
  group_by(Year,Top20Rank) %>%
  summarise(count=n())

lev=levels(factor(youtube_new2$Top20Rank))
lev=lev[c(2,1)]
youtube_new2$Top20Rank=factor(youtube_new2$Top20Rank,levels = lev)

ggplot(youtube_new2,aes(x=Year,y=count,fill=Top20Rank))+
 geom_col(position = "stack")+
 scale_fill_manual(values = c("magenta","dodgerblue"))+
 scale_x_continuous(breaks = seq(2007,2017,1))+
 theme_bw()+
 labs(title="Top 100 Most Viewed YouTube Videos by Upload Year",
       x="Year Uploaded")
```



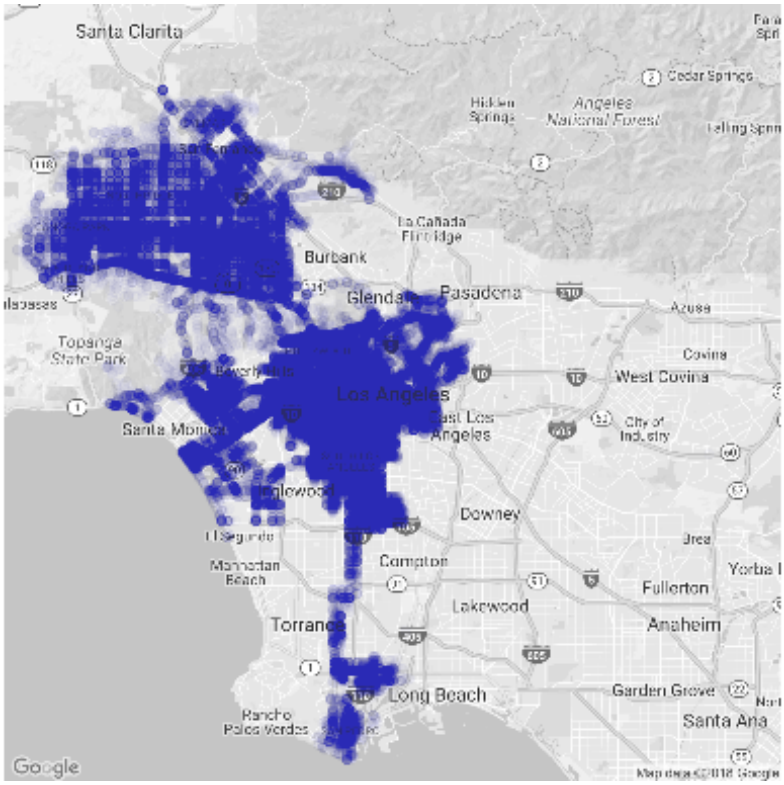
Case2

1. Setup a black and white map for LA. Save it to a variable called LosAngeles. Set yourmaptype to roadmap and the zooming parameter to 10.

```
LosAngeles=qmap(location = "Los Angeles",zoom = 10,maptype = "roadmap",color
= "bw")

## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=Los+An
geles&zoom=10&size=640x640&scale=2&maptype=roadmap&language=en-EN&sensor=fals
e

## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?ad
dress=Los%20Angeles&sensor=false

## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```

2. Reproduce the following map to show where the traffic collisions in LA were from 2008 till 2013? Choose an appropriate alpha (0.01), use color dark blue. Do you drive in LA? Describe what you see from the map you produced.

```
LosAngeles+
  geom_point(data = collision,
             aes(x=LON,y=LAT),
             color="dark blue",
             alpha=0.01)

## Warning: Removed 3454 rows containing missing values (geom_point).
```

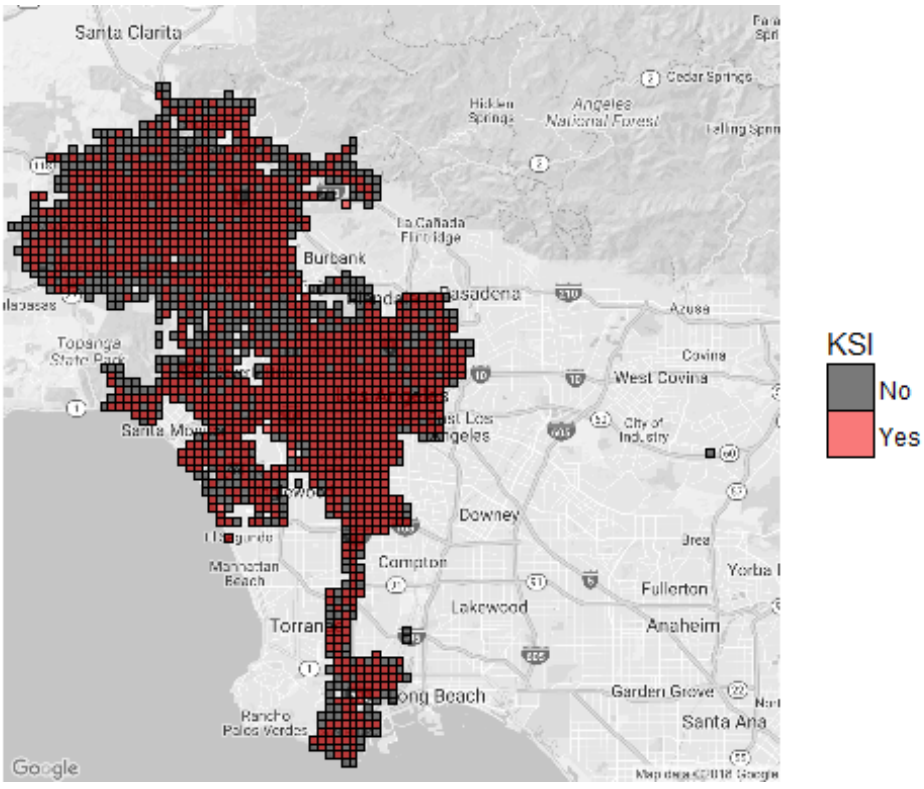


From the map I found collisions were occured around the downtown and north-west Los Angeles. Also the freeway to Long Beach is accident-prone area.

3. In traffic collision analysis, create a metric KSI (killed and severely injured) to identify the severeness of an accident. Create a new column KSI, put in Yes for those accidents with HighestDegreeofInjury of Fatal or Severe Injury. Mark No for other cases. Recreate the following map. (Use size = 0.01, bins=100, alpha=0.5, red color for Yes, and black color for No)

```
collision2=collision %>%
  mutate(KSI=ifelse(HighestDegreeofInjury %in% c("Fatal","Severe Injury"),"Yes","No"))

LosAngeles+
  stat_bin2d(data = collision2,
             aes(x=LON,y=LAT,fill=KSI),
             bins = 100,alpha=0.5,size=0.01,col="black")+
  scale_fill_manual(values = c("black","red"))
```

```
## Warning: Removed 3454 rows containing non-finite values (stat_bin2d).
```



4. Now we focus on central Los Angeles areas and collisions with “KSI=YES” cases only. Reproduce a copy of the following graphs. (Use bins=5, and zooming parameter =12)

```
LosAngeles2=qmap(location = "Los Angeles",zoom = 12,maptype = "roadmap",color = "bw")
```

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=Los+An
geles&zoom=12&size=640x640&scale=2&maptype=roadmap&language=en-EN&sensor=fals
e

## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?ad
dress=Los%20Angeles&sensor=false

## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```

```
collision3=collision2 %>%
  filter(KSI=="Yes")

lev2=levels(factor(collision3$CollisionDayofWeek))
lev2=lev2[c(2,6,7,5,1,3,4)]
collision3$CollisionDayofWeek=factor(collision3$CollisionDayofWeek,levels = l
ev2)

LosAngeles2+
  stat_density2d(data = collision3,
                 aes(x=LON,y=LAT,
                     fill=..level..,
                     alpha=..level..),
                 bins=5,
                 geom = "polygon")+
  scale_fill_gradient(low= "black", high = "red",guide = FALSE)+
  facet_wrap(~CollisionDayofWeek,nrow = 3)+
  ggtitle("KSI Collisions by Day of the Week")+
  theme(legend.position = 'none')
```

```
## Warning: Removed 3294 rows containing non-finite values (stat_density2d).
```

KSI Collisions by Day of the Week