Data Wrangling and Visualization

Divya Krishnan

Monday, November 2, 2015

Data Wrangling and Visualization

```
# Stardard libraries
library(tidyr)
library(dplyr)
library(ggplot2)
```

Exploring the NYC Flights Data In this problem set we will continue to use the data on all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013. You can find this data as part of the `nycflights13` R package. Data includes not only information about flights, but also data about planes, airports, weather, and airlines. Load the data and use it to answer the following questions.

(a) Flights are often delayed. Perform an exploratory data analysis to address each of the following questions:

- What was the worst day to fly out of NYC in 2013 if you dislike delayed flights?
  31st January was the worst day to fly out of NYC, if you dislike delayed flights.

```
library(nycflights13)
flights<-nycflights13::flights
# Computing total departure delay for each day
flights %>%
  group_by(year,month,day) %>%
    summarise(totdelay=sum(dep_delay,na.rm=TRUE)) %>%
      arrange(desc(totdelay))
```

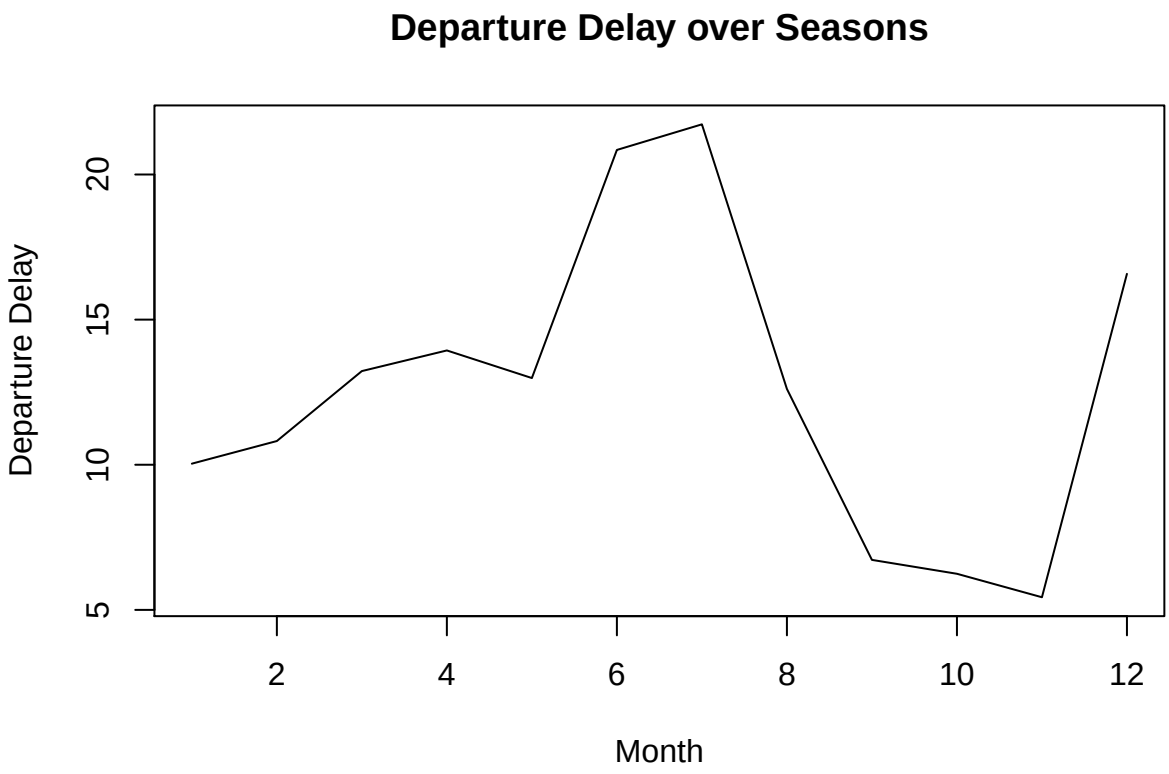
```
## Source: local data frame [365 x 4]
## Groups: year, month [12]
##
##     year month   day totdelay
##    (int) (int) (int)    (dbl)
## 1   2013     1    31    24159
## 2   2013     1    30    22956
## 3   2013     1    16    21044
## 4   2013     1    25    19424
## 5   2013     1    24    17733
## 6   2013     1    13    16137
## 7   2013     1    28    13004
## 8   2013     1     2    12958
## 9   2013     1    22    11062
## 10  2013     1     3     9933
## ..   ...   ...   ...      ...
```

```r
# Computing mean departure delay for each day
flights %>%
  group_by(year,month,day) %>%
    summarise(meandelay=mean(dep_delay,na.rm=TRUE)) %>%
      arrange(desc(meandelay))
```

```
## Source: local data frame [365 x 4]
## Groups: year, month [12]
##
##     year month   day meandelay
##    (int) (int) (int)     (dbl)
## 1   2013     1    31  28.65836
## 2   2013     1    30  28.62344
## 3   2013     1    16  24.61287
## 4   2013     1    25  21.89853
## 5   2013     1    13  19.87315
## 6   2013     1    24  19.46542
## 7   2013     1    28  15.13853
## 8   2013     1     2  13.85882
## 9   2013     1    22  12.49944
## 10  2013     1     1  11.54893
## ..   ...   ...   ...       ...
```

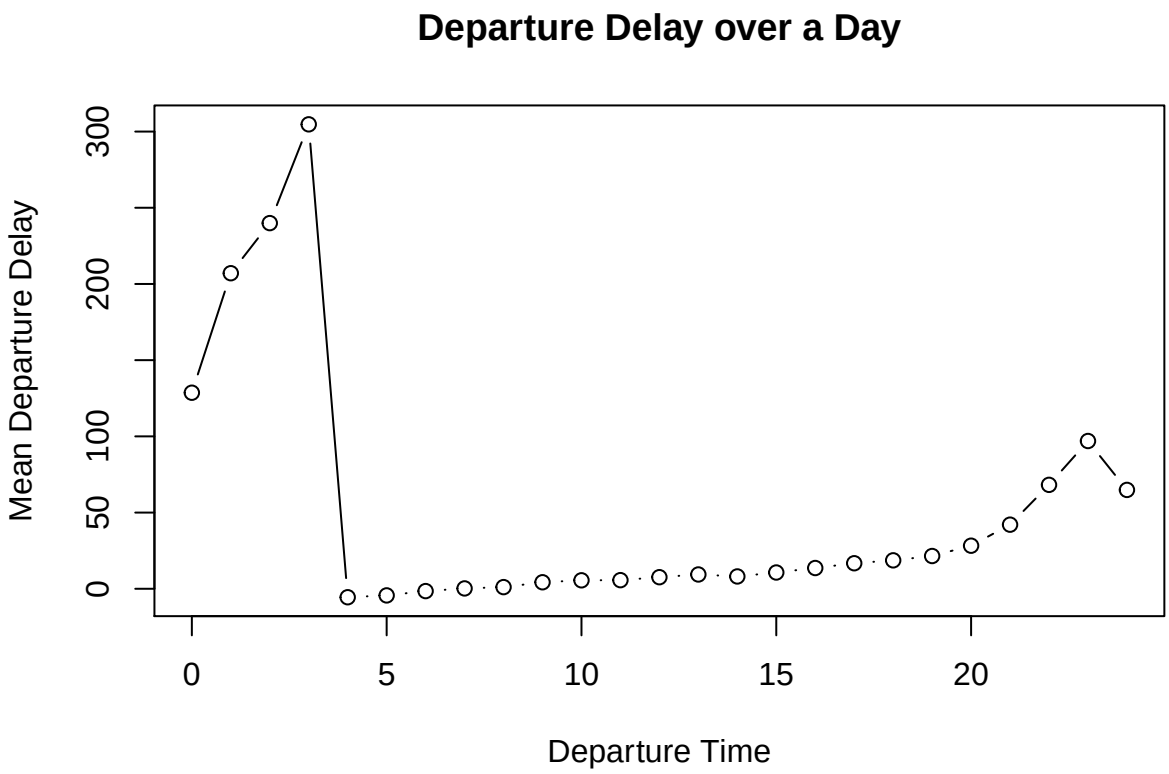
- Are there any seasonal patterns in departure delays for flights from NYC?
  It is observed that there is highest mean departure delay in the summer months of June,July and August and lowest in fall months of Sep,Oct and Nov. There is a slight increase in mean departure delay in the months of spring, namely Mar,Apr and May. The mean departure delay is very high in Dec (holiday month) but not so much in the other winter months of Jan and Feb.

```r
# Computing mean delay according to month
seasonalDelay<-flights %>%
      group_by(month) %>%
        summarise(seasonalDepDelay=mean(dep_delay,na.rm=TRUE))
par(mfrow=c(1,1))
# Plotting mean delay across seasons
plot(seasonalDelay$seasonalDepDelay ~ seasonalDelay$month,type="l",
     xlab="Month",ylab="Departure Delay",main="Departure Delay over Seasons")
```



- On average, how do departure delays vary over the course of a day?
  The mean departure delay over the course of a day peaks between 3-4 am and then drops drastically between 4-5 am and then has a slow and steady growth till 10 pm.

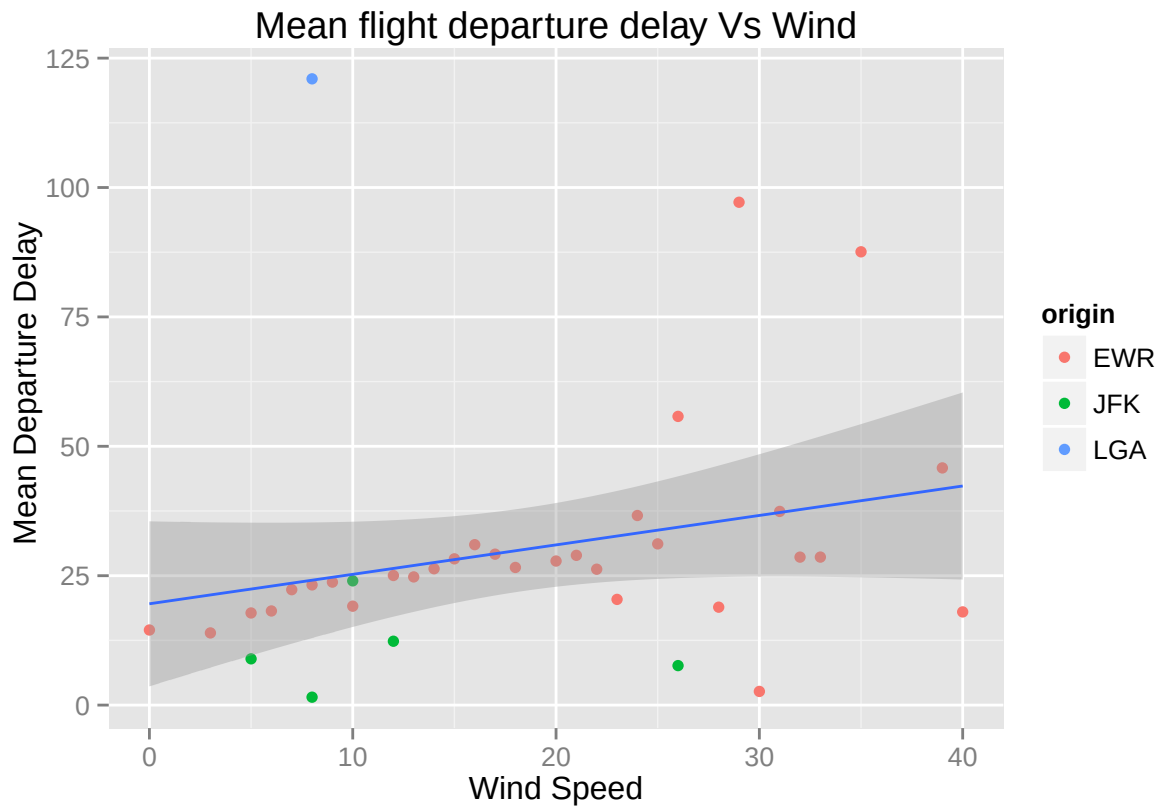
```
# Computing mean departure delay over a day (measured by hour)
acrossADay<-flights %>%
  group_by(hour) %>%
    summarise(delayDepTime=mean(dep_delay,na.rm=TRUE))
# Plotting mean departure delay Vs Hour
plot(acrossADay$delayDepTime ~ acrossADay$hour,type="b",
     xlab="Departure Time",ylab="Mean Departure Delay",
     main="Departure Delay over a Day")
```



(b) Flight delays are often linked to weather conditions. How does weather impact flights from NYC? Utilize both the `flights` and `weather` datasets to explore this question. Include at least one visualization to aid in communicating what you find. The data and the visualization show that as the wind speed increases, the mean departure delay for flights increases. There is a good linear relationship between Wind Speed and Mean departure delay. The mean departure delay for flights originating from EWR airport definitely have a good linear relationship except for some outliers. But the same can't be said about flights originating from JFK. There is only one flight originating from LGA, hence can't comment much about that airport as well.

```
# Importing weather data
weather<-nycflights13::weather
# Departure delay for each hour
depDelayHour<-flights %>%
  group_by(origin,year,month,day,hour) %>%
    summarise(meanDelay=mean(dep_delay,na.rm=TRUE))
# Joing weather and Departure delay for each hour
depDelayWeather<-weather %>%
  inner_join(depDelayHour,c("origin","year","month","day","hour"))
# Grouping by wind speed (rounded) and origin
windVsDelay<-depDelayWeather %>%
  group_by(windSpeed=round(wind_speed,0),origin) %>%
    summarise(windDelay=mean(meanDelay,na.rm=TRUE))
# Removing the Wind Speed=1048, which seems to be an outlier
windVsDelay<-subset(windVsDelay,windVsDelay$windSpeed<1000)
```

```
# Plotting Mean flight departure delay across wind speed
g<-ggplot(windVsDelay, aes(x=windSpeed, y=windDelay))
g<-g + geom_point(aes(color=origin),size=2)
g<-g + geom_smooth(method="lm")
g <- g + labs(title="Mean flight departure delay Vs Wind",
              x="Wind Speed", y=" Mean Departure Delay")
g
```



(c) Flight performance may also be impacted by the aircraft used. Do aircrafts with certain characteristics (e.g. manufacturer) demonstrate better performance? Utilize both the `flights` and `planes` datasets to explore this question. Include at least one visualization to aid in communicating what you find. The visualizations compared Mean departure and arrival delay with the year of manufacture of the plane. There is very weak or hardly any correlation between year of manufacture and mean departure and arrival delay. Also the 'Fixed wing multi engine' seems to be the most common type of plane and one that is mostly manufactured in the recent times. It can be said that the year of the manufacture hardly has any effect on the performance of the plane.

```
# Importing planes dataset
planes<-nycflights13::planes
# Computing mean delays,air time and distance
depDelayTailnum<-flights %>%
  group_by(tailnum) %>%
    summarise(meanDepDelay=mean(dep_delay,na.rm=TRUE),
              meanArrDelay=mean(arr_delay,na.rm=TRUE),
```

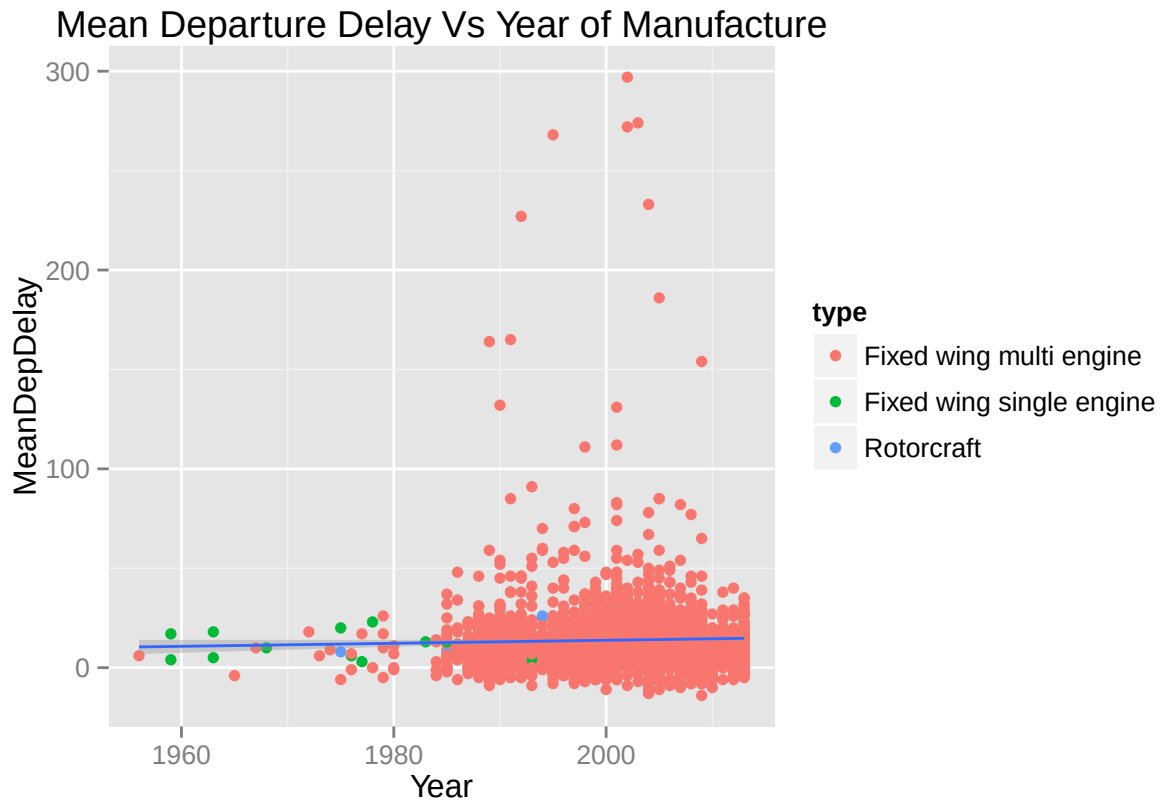
```
                  meanAirTime=mean(air_time,na.rm=TRUE),
                    meanDistance=mean(distance,na.rm=TRUE))
# Joining planes and summarised data set of flights
depDelayPlanes<-depDelayTailnum %>%
                inner_join(planes,by="tailnum")

# Removing rows with missing year value
depDelayPlanes.mod<-depDelayPlanes[complete.cases(depDelayPlanes[,6]),]

# Plotting Year of Manufacture and Mean departure delay
g<-ggplot(depDelayPlanes.mod, aes(x=year, y=round(meanDepDelay,0)))
g<-g + geom_point(aes(color=type),size=2)
#g<-g + scale_color_discrete(name="type")
g<-g + geom_smooth(method="lm")
g <- g + labs(title="Mean Departure Delay Vs Year of Manufacture",
              x="Year", y="MeanDepDelay")
g
```

```
## Warning: Removed 6 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```

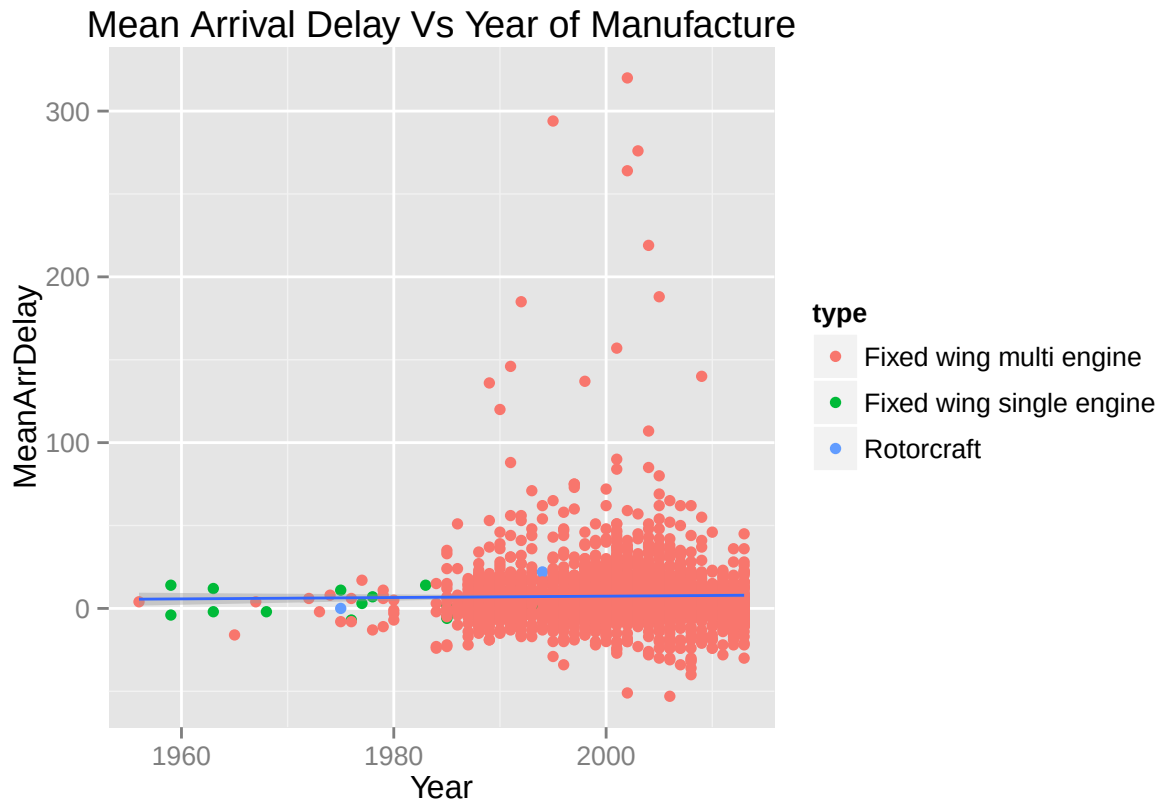



```
# Plotting Year of Manufacture and Mean arrival delay
g<-ggplot(depDelayPlanes.mod, aes(x=year, y=round(meanArrDelay,0)))
```

```
g<-g + geom_point(aes(color=type),size=2)
#g<-g + scale_color_discrete(name="type")
g<-g + geom_smooth(method="lm")
g <- g + labs(title="Mean Arrival Delay Vs Year of Manufacture",
              x="Year", y="MeanArrDelay")
g
```

```
## Warning: Removed 6 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



Data Wrangling and Visualization In this problem we will use the `state` dataset, available as part of the R statistical computing platforms. Load the data and use it to answer the following questions.

(a) Describe the data and each variable it contains. Tidy the data, preparing it for an exploratory data analysis. The "state" data set has the following variables - state.abb: abbreviations for the state names. state.area: state areas (in square miles). state.center: approximate geographic center of each state in negative longitude and latitude. Alaska and Hawaii are placed just off the West Coast. state.division: state divisions (New England, Middle Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain, and Pacific). state.name: full state names. state.region: region (Northeast, South, North Central, West) that each state belongs to. state.x77: Matrix giving the following statistics in the respective columns. Population: population estimate as of July 1, 1975 Income: per capita income (1974) Illiteracy: illiteracy (1970, percent of population)

Life Exp: life expectancy in years (1969–71) Murder: murder and non-negligent manslaughter rate per 100,000 population (1976) HS Grad: percent high-school graduates (1970) Frost: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city Area: land area in square miles[1]

-Reference [1] State {datasets}. (n.d.). Retrieved November 2, 2015, from http://www.inside-r.org/r-doc/datasets/state

```
# Getting state dataset
data(state)
# Creating a data frame of the various variables
  # inside state data set except the matrix variable state.x77
state<-data.frame(abb=state.abb,area=state.area,
                  longitude=state.center$x,latitude=state.center$y,
                  division=state.division,name=state.name,
                  region=state.region,stringsAsFactors=FALSE)
# Creating a data frame for state.x77 variable
stateStats<-as.data.frame(state.x77)

# Tidying state dataset
state<-tbl_df(state)
state.tidy<-state %>%
              gather(locationType,locationValue,-c(abb,area,division,name,region))
# Recasting division, region and name to character
state.tidy$division<-as.character(state.tidy$division)
state.tidy$region<-as.character(state.tidy$region)
state.tidy$locationType<-as.character(state.tidy$locationType)

# Rearranging columns
state.tidy<-subset(state.tidy,
                   select=c(name,abb,area,division,
                            region,locationType,locationValue))

# Tidying state.x77 dataset

# Getting state names from the row names
stateStats$name<-rownames(state.x77)
# Setting row names to integer values
rownames(stateStats)<-c(1:nrow(stateStats))
# Gathering columns except name into rows
stateStats.tidy<-tbl_df(stateStats %>%
  gather(statType,value,-name))

# Joining the 2 data frames of state
stateMain<-state.tidy %>%
            inner_join(stateStats.tidy,by="name")

# Changing the Area value in statType to 'Land Area'
stateMain$statType<-as.character(stateMain$statType)
stateMain$statType<-sub("Area","Land Area",stateMain$statType)

# Including area column as 'Total area' category in the statType column

# Spreading out the statType column
stateMain.tidy<-stateMain %>%
```

```
  spread(statType,value)
# Renaming area column as 'Total Area'
colnames(stateMain.tidy)[3]<-"Total Area"

stateMain.tidy<-stateMain.tidy %>%
  gather(statType,value,-c(name,abb,division,region,locationType,locationValue))

# Tidied final state datatset
stateMain.tidy
```

```
## Source: local data frame [900 x 8]
##
##          name   abb           division region locationType locationValue
##         (chr) (chr)              (chr)  (chr)        (chr)         (dbl)
## 1     Alabama    AL East South Central  South     latitude       32.5901
## 2     Alabama    AL East South Central  South    longitude      -86.7509
## 3      Alaska    AK            Pacific   West     latitude       49.2500
## 4      Alaska    AK            Pacific   West    longitude     -127.2500
## 5     Arizona    AZ           Mountain   West     latitude       34.2192
## 6     Arizona    AZ           Mountain   West    longitude     -111.6250
## 7    Arkansas    AR West South Central  South     latitude       34.7336
## 8    Arkansas    AR West South Central  South    longitude      -92.2992
## 9  California    CA            Pacific   West     latitude       36.5341
## 10 California    CA            Pacific   West    longitude     -119.7730
## ..         ...   ...               ...    ...          ...           ...
## Variables not shown: statType (fctr), value (dbl)
```

(b) Develop one research question of your own that you can address using the state dataset. Clearly state the question you are going to address. Provide at least two visualizations to support your exploration of this question. Discuss what you find.

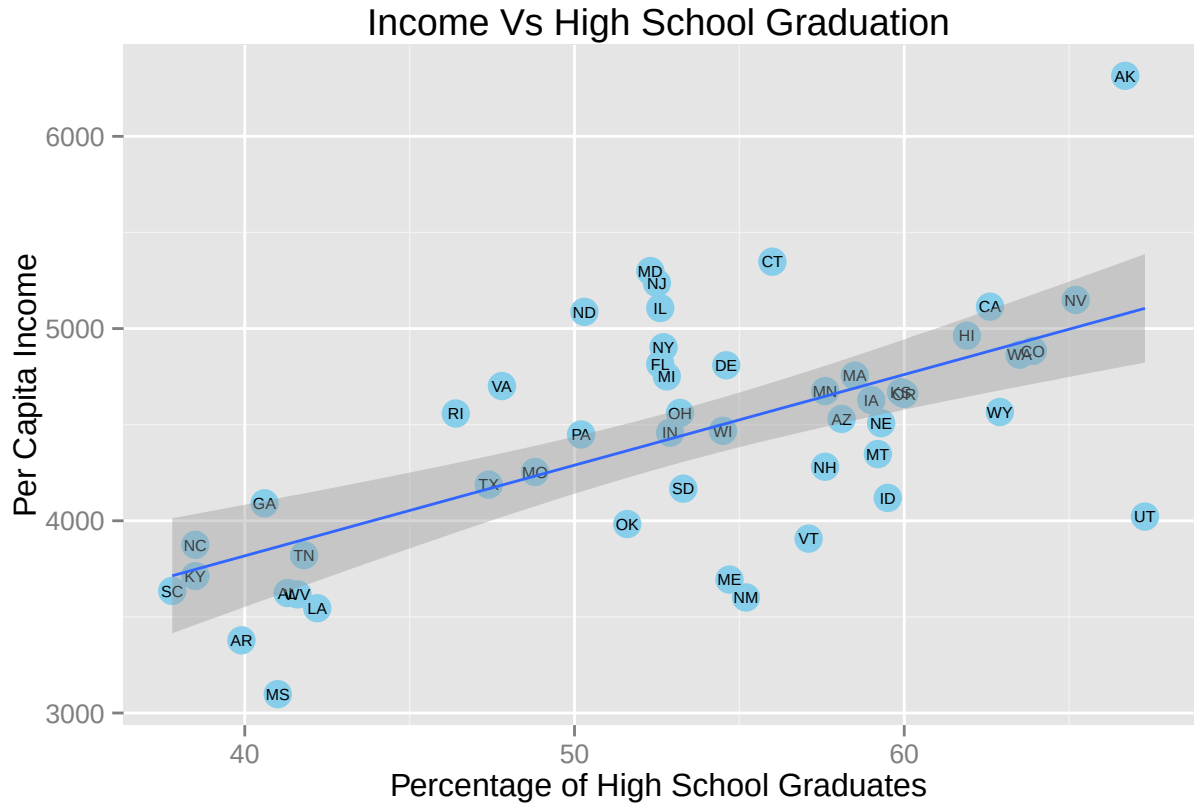
Research Question: Does the affect of education level on per capita income vary according to the region of the state? Education level can be measured using the Percentage of High School Graduates in the state. The per capita income in a state is present in the state dataset.

The visualization of 'Income Vs High School Graduation(Across Regions)' depicts that as the percentage of high school graduates increases, the per capita income of the state also increases. There seems to be a strong positive linear relationship between 'Percentage of High School Graduates' and 'Per Capita Income'. It can also be observed that the West region has majority of the states having higher percentage of high school graduates and higher per capita income whereas the South region has majority of the states having lower percentage of high school graduates and lower per cpaita income. The North Central and Northeast regions have medium percentage of high school graduates and medium per capita income. The visualizations of 'Income Vs High School Graduation(Across Regions Individually)' provides a closer examination of the regions individually. The visualization results of south and west region concur with the previous visualization conclusions. But it can be noticed that Northeast region has a higher variance than the North Central region. Also, there is hardly any linear relationship between percentage high school graduates and income level in North Central region and there is a very small negative linear relationship between percentage high school graduates and income level in Northeast region.

```
# Spreading the state dataset
stateData<-stateMain.tidy %>%
  spread(statType,value)
# Spreadin the location rows into columns
```

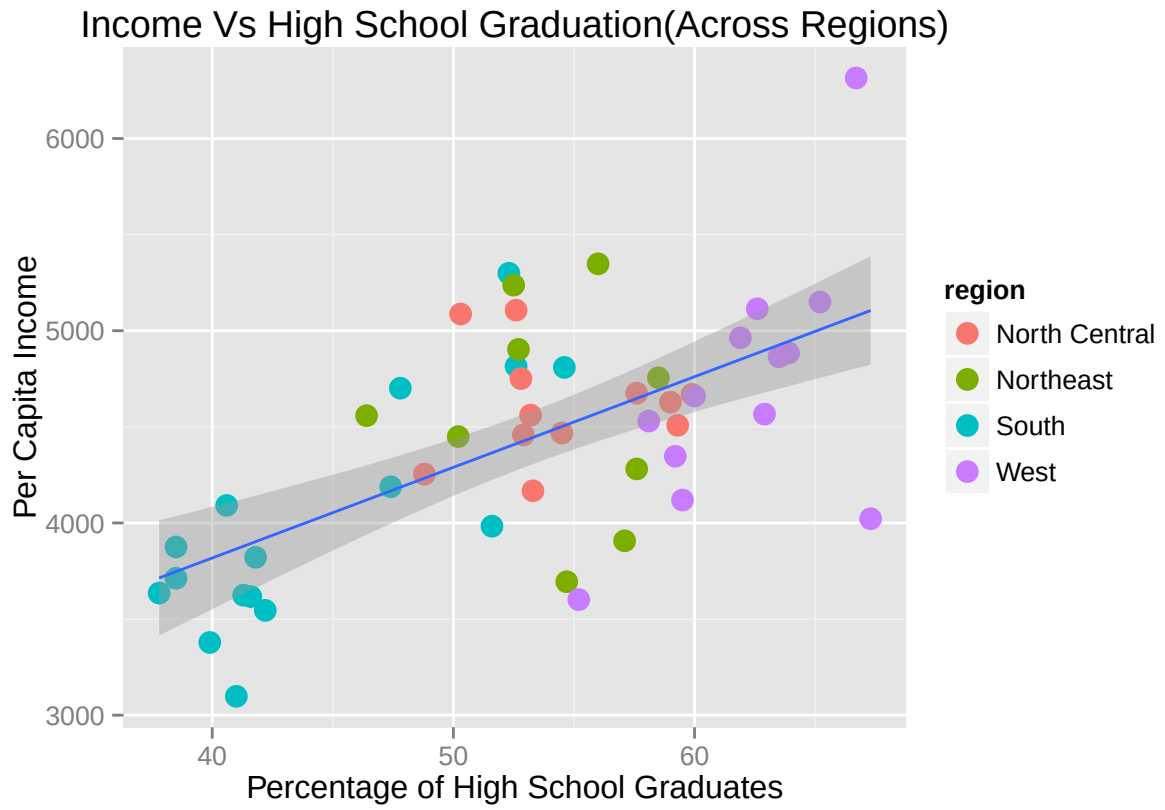
```
stateData<-stateData %>%
            spread(locationType,locationValue)
# Renaming the colnames according to convention
colnames(stateData)<-c("name","abb","division","region",
                       "totalArea","frost","hsGrad","illiteracy",
                       "income","landArea","lifeExpectancy","murder",
                       "population","latitude","longitude")
# Creating dataset about education,income and regions
st<-stateData %>%
  select(abb,region,hsGrad,income)

# Plot of Income Vs High School Graduation for each state
g<-ggplot(st,aes(x=hsGrad,y=income))
g<-g + geom_point(color="skyblue",size=5)
g<-g + geom_text(aes(label=abb), size=2)
g<-g + stat_smooth(method="lm")
g<-g + labs(title="Income Vs High School Graduation",
            x="Percentage of High School Graduates", y="Per Capita Income")
g
```

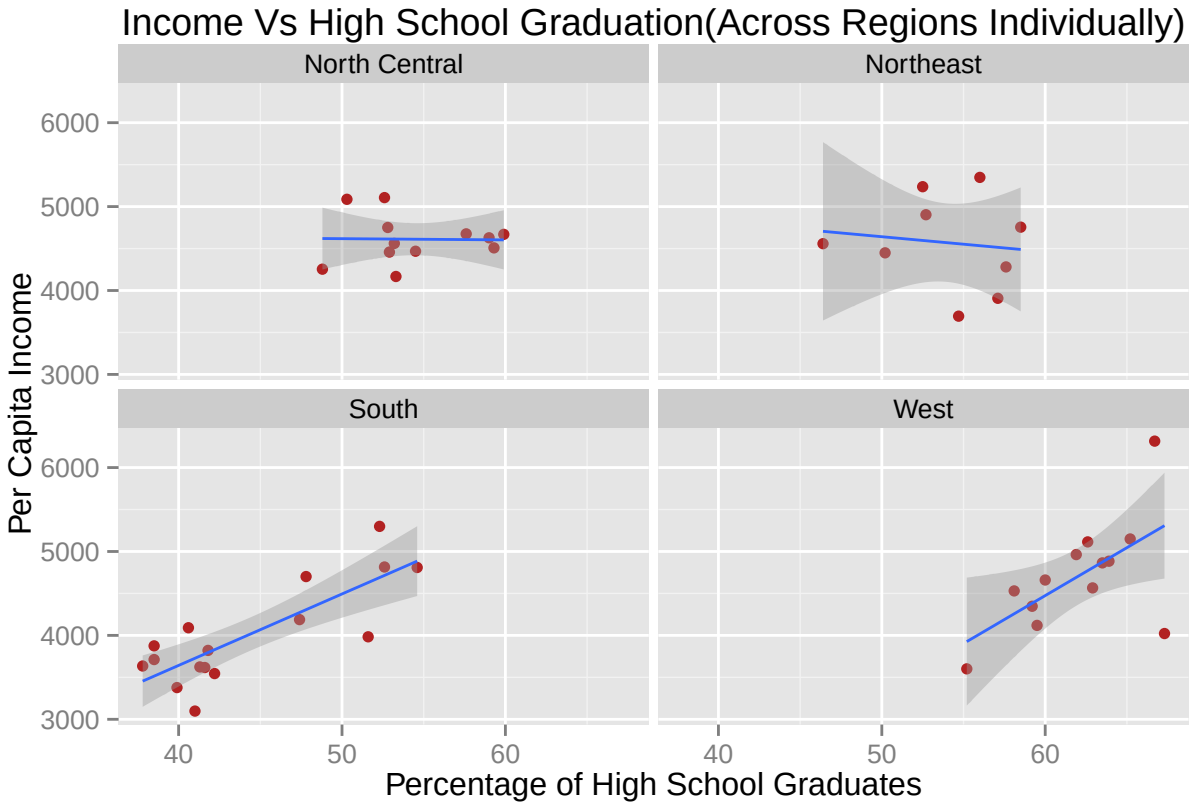



```
# Plot of Income Vs High School Graduation(Across Regions)
g<-ggplot(st,aes(x=hsGrad,y=income))
g<-g + geom_point(aes(color=region),size=4)
g<-g + stat_smooth(method="lm")
g<-g + labs(title="Income Vs High School Graduation(Across Regions)",
```

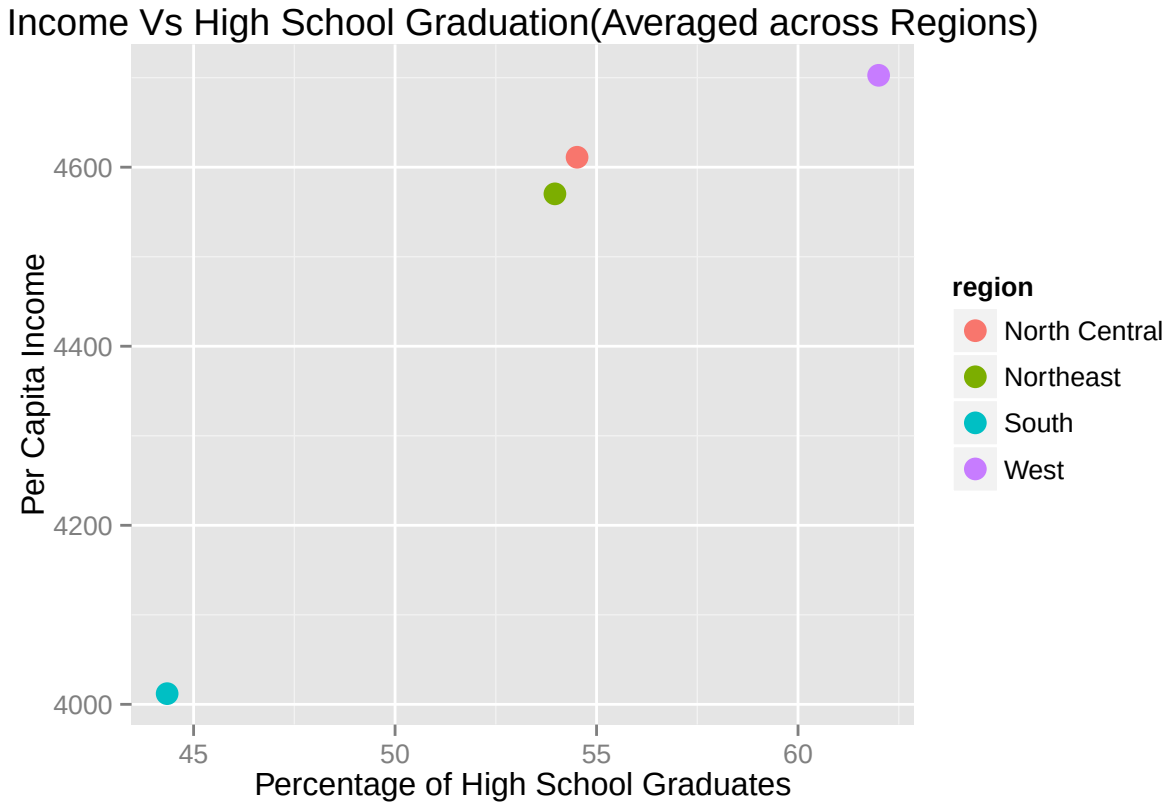
```
            x="Percentage of High School Graduates", y="Per Capita Income")
g
```




```
# Plot of Income Vs High School Graduation(Across Regions Individually)
g<-ggplot(st, aes(hsGrad,income))
g<-g + geom_point(color="firebrick")
g<-g + geom_smooth(method="lm")
g<-g + facet_wrap(~region, ncol=2)
g<-g + labs(title="Income Vs High School Graduation(Across Regions Individually)",
            x="Percentage of High School Graduates", y="Per Capita Income")
g
```




```
# Computing average graduation and income
stRegion<-st %>%
            group_by(region) %>%
              summarise(meanHsGrad=mean(hsGrad,na.rm=TRUE),meanIncome=mean(income,na.rm=TRUE))
# Plotting Mean Income Vs Mean High School Graduation(Across Regions)
g<-ggplot(stRegion,aes(x=meanHsGrad,y=meanIncome))
g<-g + geom_point(aes(color=region),size=4)
g<-g + labs(title="Income Vs High School Graduation(Averaged across Regions)",
            x="Percentage of High School Graduates", y="Per Capita Income")
g
```


Income Vs High School Graduation(Averaged across Regions)
Per Capita Income
4600
4400
4200
4000
45
50
55
60
Percentage of High School Graduates
region
North Central
Northeast
South
West