# INFX 573: Problem Set 3 - Data Analysis

*Rajendran Seetharaman*

*Due: Thursday, October 26, 2017*

**Collaborators:**

**Instructions:**

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset3.Rmd` file from Canvas. Open `problemset3.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset3.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the R Markdown file to `YourLastName_YourFirstName_ps3.Rmd`, knit a PDF and submit the PDF file on Canvas.

**Setup:**

In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
```

**Problem 1: Flight Delays**

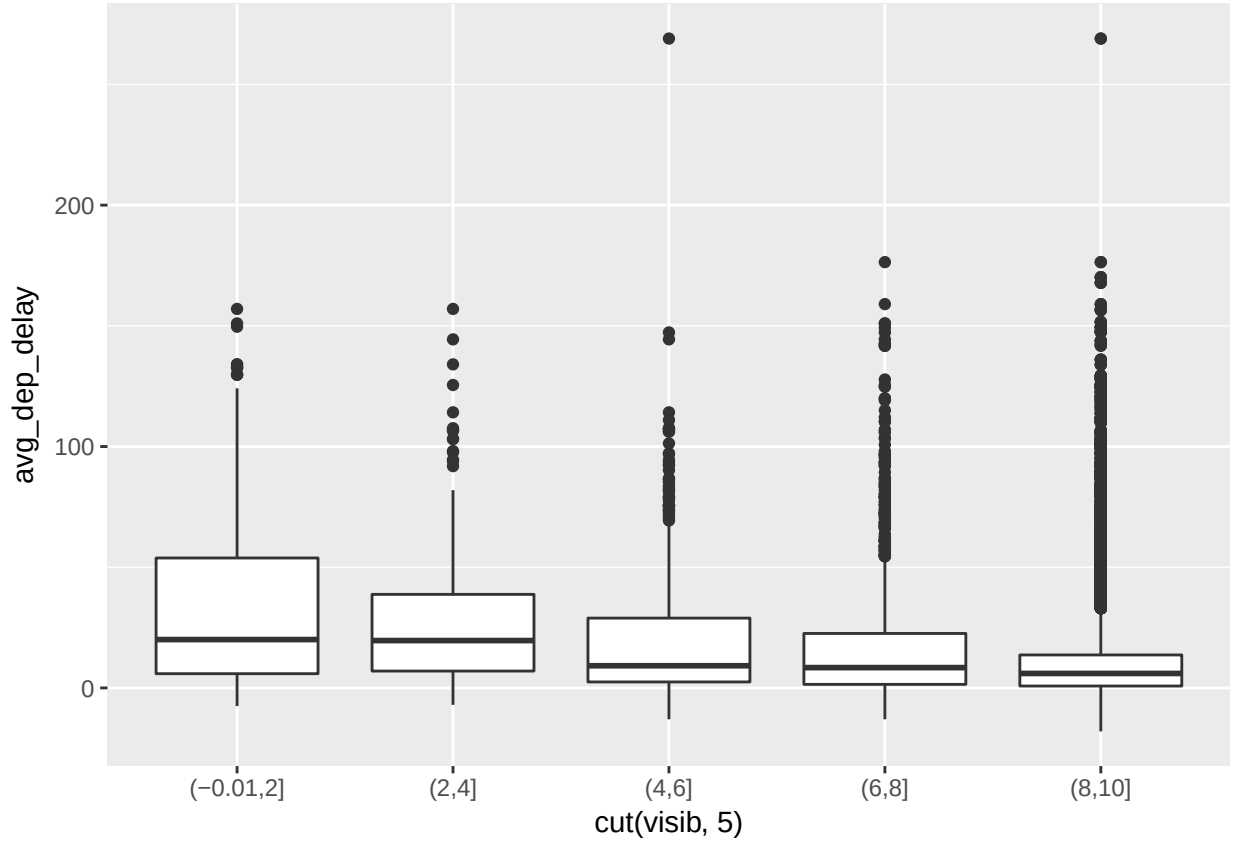
Flight delays are often linked to weather conditions. How does weather impact flights from NYC? Utilize both the `flights` and `weather` datasets from the `nycflights13` package to explore this question. Include at least two visualizations to aid in communicating what you find.

Ans. I first tried to analyze the relationship that the visibility level at a particular time of the day has with the arrival and departure delays of the particular flight. My observation from the data is that as the visibility level increases, the mean arrival and departure delays go down. The median arrival/departure delays as well as the interquartile range for low visibility flights is higher than higher visibility flights. As the visibility increases, both tend to go down. This might be because pilots can navigate better in good visibility conditions; reducing flight times and hence delays.

I then tried to analyse the relationship between humidity at a particular time and the mean arrival and departure delays at that time. I observed that as the humidity increases, the median of the avg arrival/departure delays tends to slightly increase after the humidity exceeds 60% approximately. When the humidity levels exceed 60%, it is an indication that it might be raining. The variability (IQR) in the arrival/departure delays also increases significantly as the humidity exceeds 60%. This might be because rainy weather conditions increase the probability of flight delays.

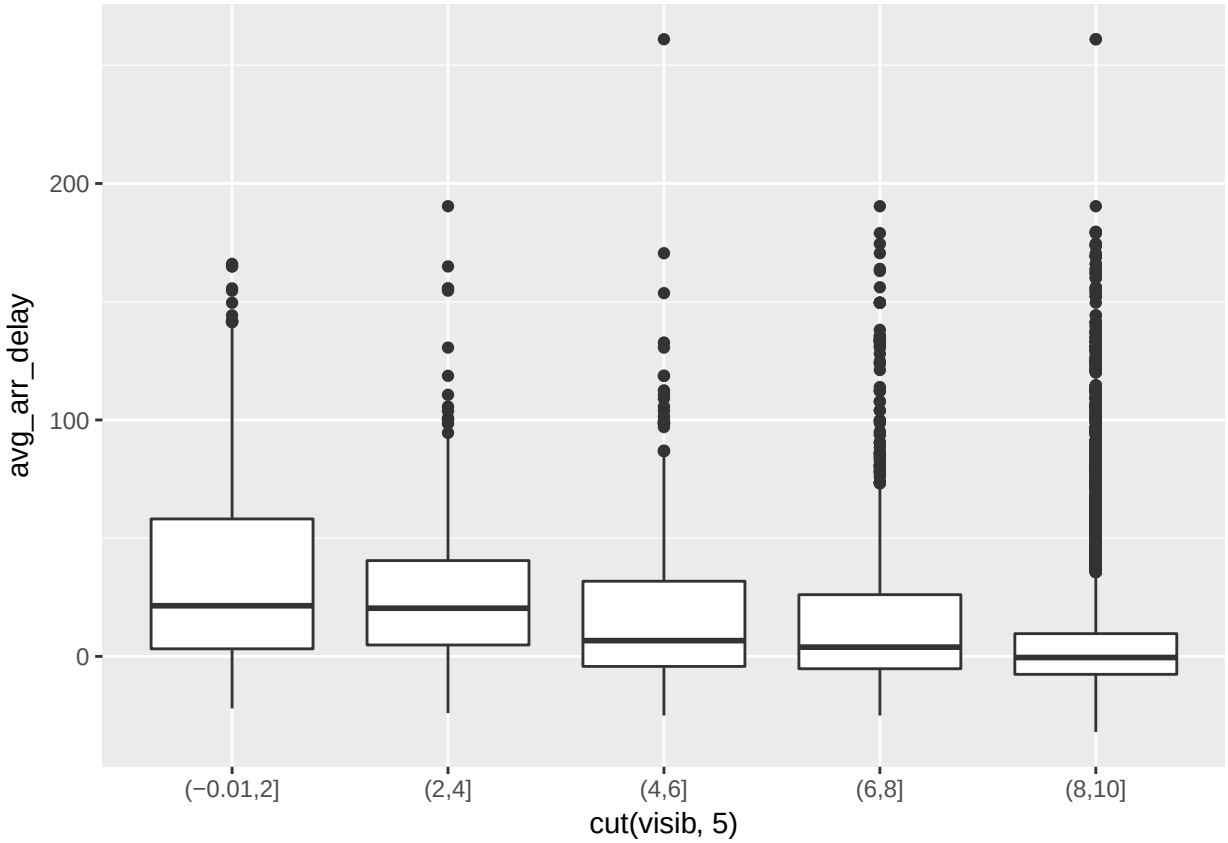
```
library(nycflights13)
# load flight and weather data
flights <- nycflights13::flights
weather <- nycflights13::weather

# group flights by the flight time & day and compute mean arrival & departure delays
flights_delay_summary <- flights %>% group_by(time_hour) %>%
  summarize(avg_dep_delay=mean(dep_delay,na.rm=TRUE),
  avg_arr_delay=mean(arr_delay,na.rm = TRUE)) %>%
  select(time_hour,avg_dep_delay,avg_arr_delay)

# join the newly computed flight dataset and weather dataset by the timestamp
weather_flights <- inner_join(flights_delay_summary,
  weather,by="time_hour")

#boxplots of mean arrival and departure delays for each grouping of visibility levels
ggplot(weather_flights,aes(x=cut(visib,5),y=avg_dep_delay))+geom_boxplot()
```

```
## Warning: Removed 39 rows containing non-finite values (stat_boxplot).
```



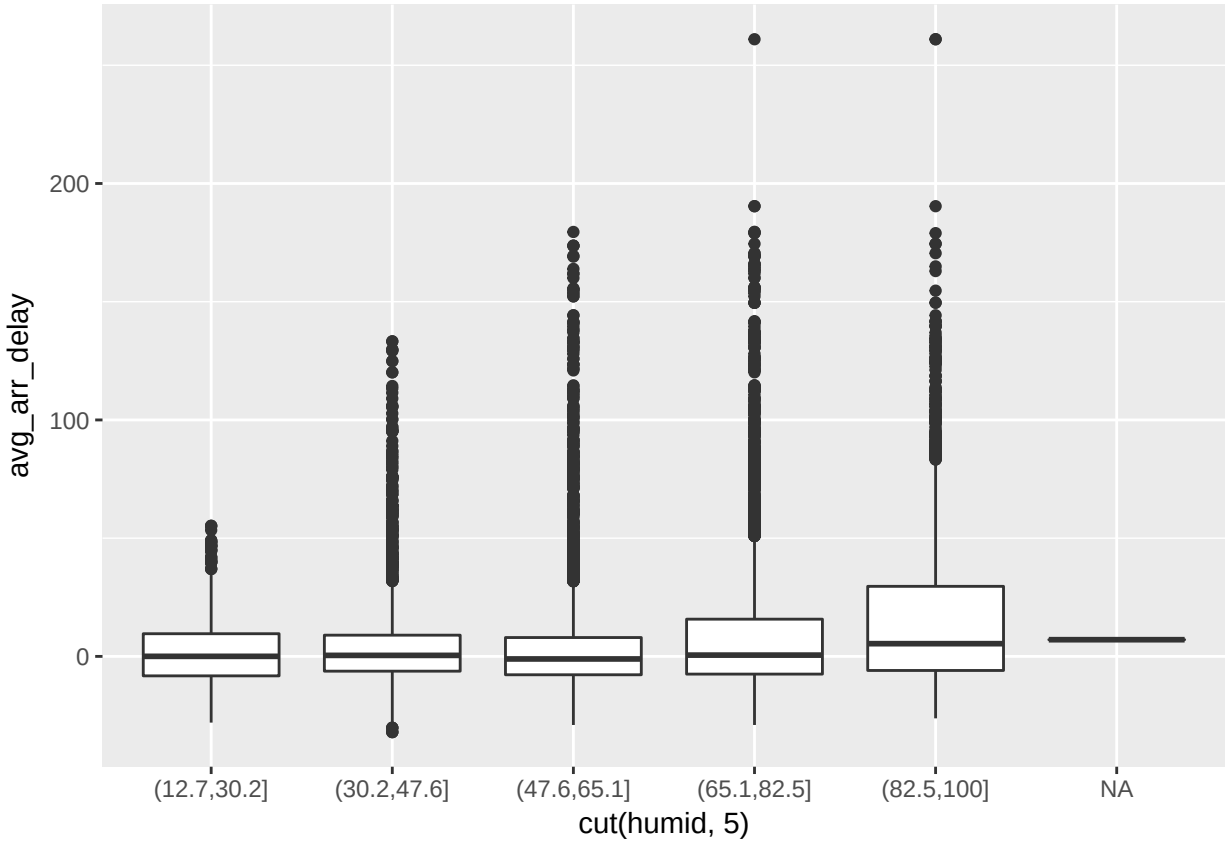

```
ggplot(weather_flights,aes(x=cut(visib,5),y=avg_arr_delay))+geom_boxplot()
```

```
## Warning: Removed 42 rows containing non-finite values (stat_boxplot).
```



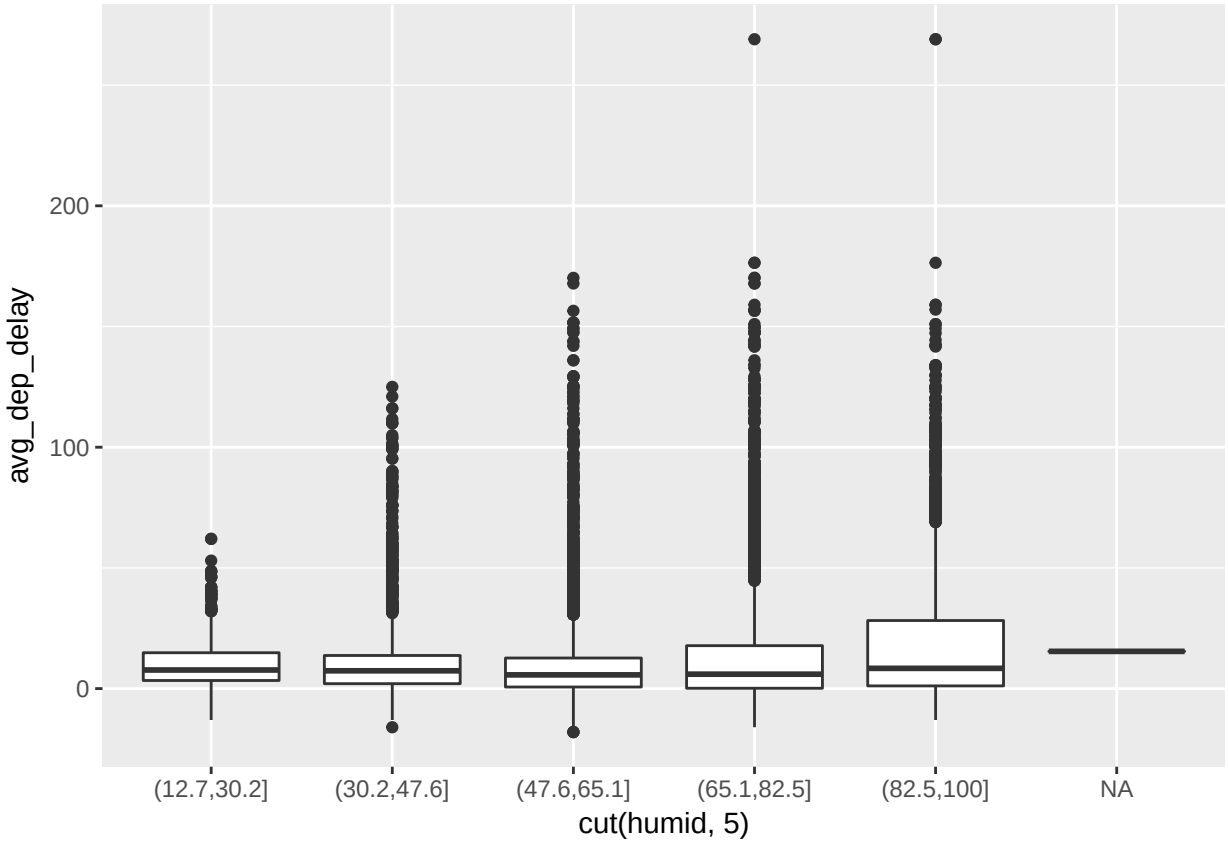

```
#boxplots of mean arrival and departure delays for each grouping of humidity levels
ggplot(weather_flights,aes(x=cut(humid,5),y=avg_arr_delay))+geom_boxplot()
```

```
## Warning: Removed 42 rows containing non-finite values (stat_boxplot).
```




```
ggplot(weather_flights,aes(x=cut(humid,5),y=avg_dep_delay))+geom_boxplot()
```

```
## Warning: Removed 39 rows containing non-finite values (stat_boxplot).
```




**Problem 2: 50 States in the USA**

In this problem we will use the `state` dataset, available as part of the R statistical computing platforms. This data is related to the 50 states of the United States of America. 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 a data analysis.**

Ans. The state data set contains information pertaining to the 50 states of the US. It contains facts and figures like average income, population, illiteracy rates etc.

The dataset contains the following variables-

abb- A 2 letter abbreviation for each state area- Numeric area of state in square miles latitude- Latitude of geographic center of state longitude- Longitude of geographic center of state division- Factor which gives state divisions. It has values corresponding to- (New England, Middle Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain, and Pacific) name- Name of the state region- Factor which gives the region of the state (Northeast, South, North Central, West) Population- Population estimate of state as of July 1st, 1975 Income- Per capita income (1974) Illiteracy - Percent of population which is illiterate (1970) Life.Exp - Life expectency in years between 1969 and 1971 Murder- Murder and non negligient manslaughter per 100000 people in 1976 HS.Grad - Percent high school graduates in 1970 Frost- mean number of days with minimum temperature below freezing (1931-1960) in capital or large city

Source- https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/state.html

```r
#merging data from all the state data sources
df <- state.x77
df1 <- data.frame(df)
df1 <- df1 %>% mutate(name=row.names(df1),abb=state.abb,
  latitude=state.center$x,longitude=state.center$y,
  division=state.division,region=state.region)
```

**(b) Suppose you want to explore the relationship between a state's `HS Grad` rate and other characteristics of the state, for example income, illiteracy rate, and more. Begin by examine the bivariate relationships present in the data. What does your analysis suggest might be important varibles to consider in building a model to explain variation in highschool graduation rates?**

Ans. I first tried to examine the relationship between HS Grad and Illiteracy. The relationship that I observed was that the variables have an inverse relationship with each other. For the states which have high HS Grad rates, the illiteracy rates are low which is as expected. This is supported by the scatterplot as well as the high negative pearson correlation coefficient of -0.65.

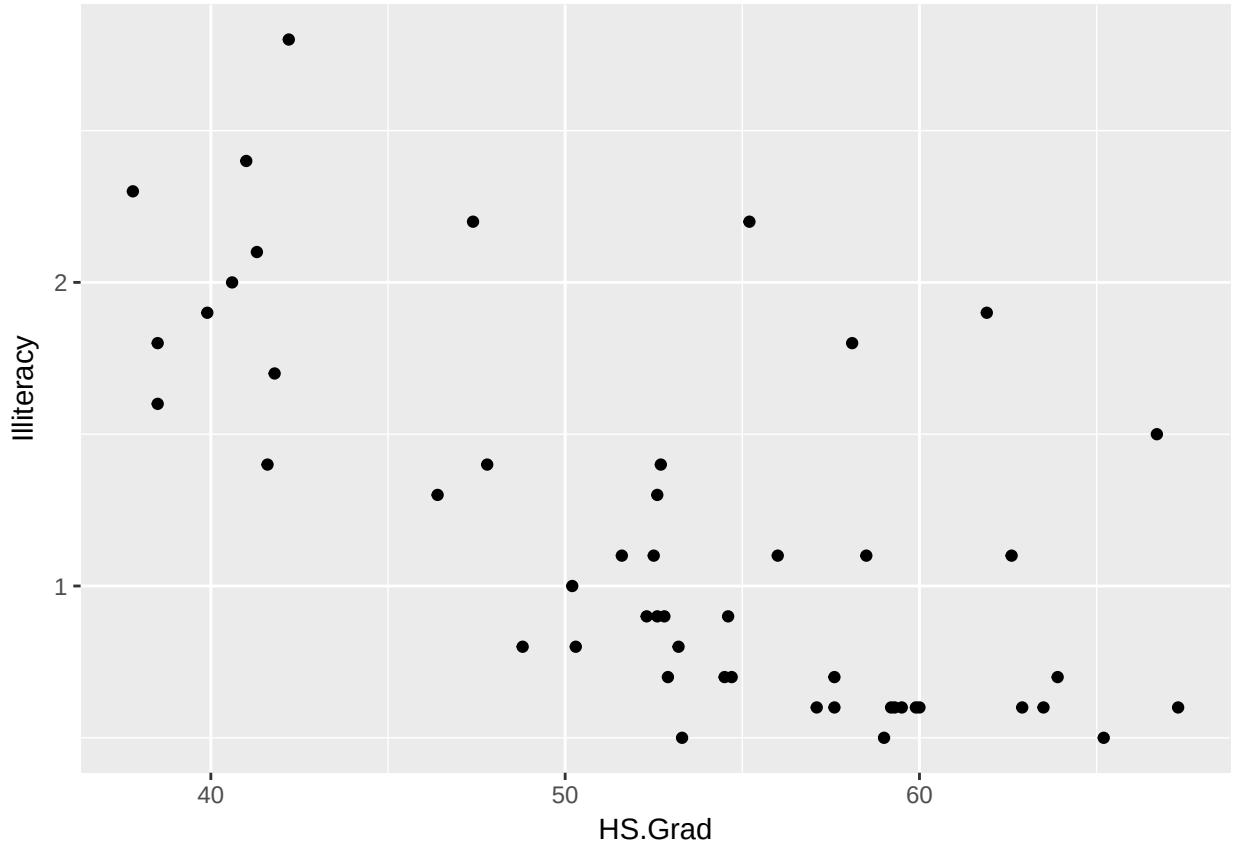
I also tried to examine the relationship between the HS Grad rates and Income variables. The data shows that these variables are correlated and have a positive relationship. This in in line with the general intution that educated individuals have better career prospects and hence earn better. The pearson correlation coefficient of 0.61 and the scatterplot support this hypothesis.

When I tried to examine the relationship between population of a state and HS Grad rates, I was unable to ascertain a strong relationship betweent the variables. The r value was also -0.09 which indicates a very low or no correlation between the variables.

Finally, I tried to examine the relationship between the region and divisions that a state falls in. I observed that west on an average has the highest HS grad rates and th south has the lowest. The west is 18 percentage points greater than that south. Also, within each region, there is a fluctuation in Avg. grad rates within each division.

To conclude, the mportant variables to consider while building a model would be income, illiteracy, region, and division.

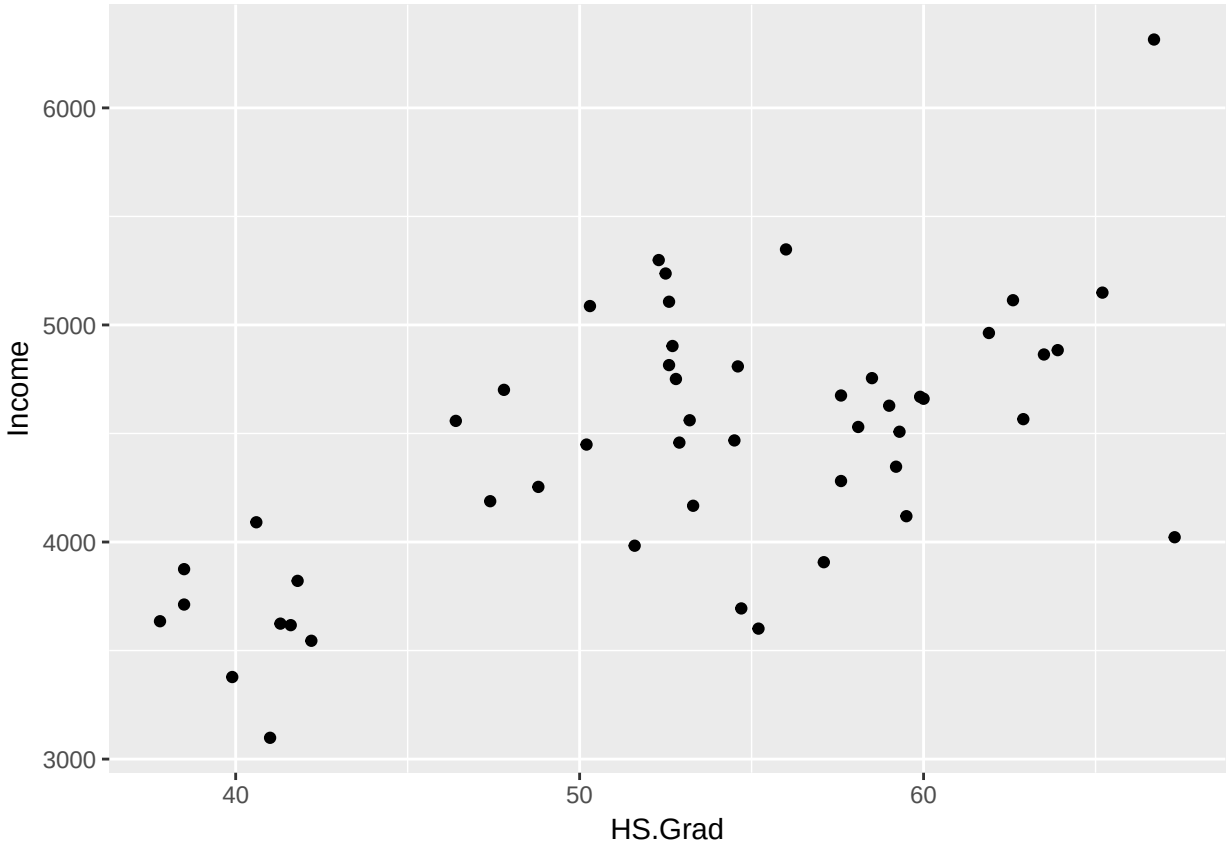
```r
# Analysing relationship between HS Grad and illiteracy
ggplot(df1,aes(x=HS.Grad,y=Illiteracy))+geom_point()
```




```
cor.test(x=as.array(df1$HS.Grad),y=as.array(df1$Illiteracy),
  method="pearson",use="complete.obs")
```

```
##
##  Pearson's product-moment correlation
##
## data:  as.array(df1$HS.Grad) and as.array(df1$Illiteracy)
## t = -6.0408, df = 48, p-value = 2.172e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7908657 -0.4636561
## sample estimates:
##       cor
## -0.6571886
```

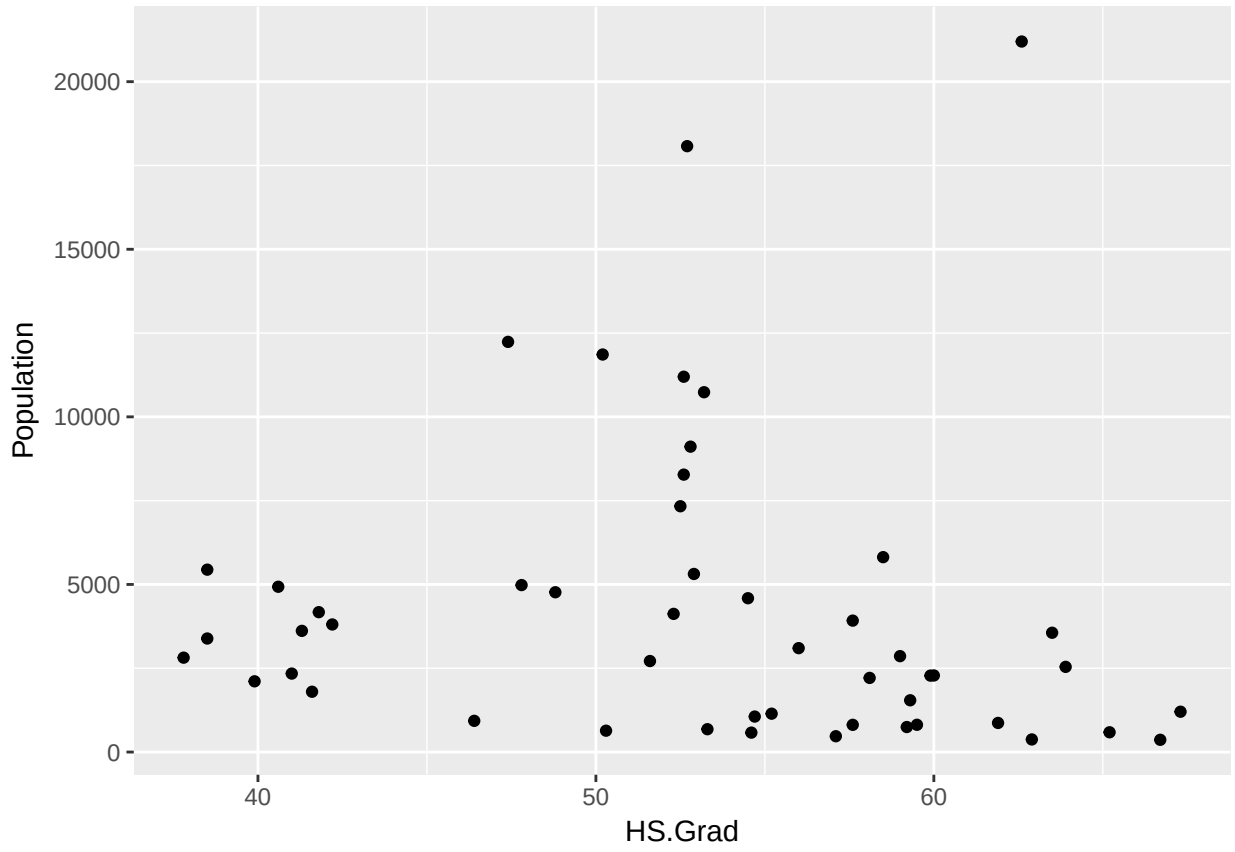
```
# Analysing relationship between HS Grad and income
ggplot(df1,aes(x=HS.Grad,y=Income))+geom_point()
```




```
cor.test(x=as.array(df1$HS.Grad),y=as.array(df1$Income),
  method="pearson",use="complete.obs")
```

```
##
##  Pearson's product-moment correlation
##
## data:  as.array(df1$HS.Grad) and as.array(df1$Income)
## t = 5.4738, df = 48, p-value = 1.579e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4128194 0.7660866
## sample estimates:
##       cor
## 0.6199323
```

```
# Analysing relationship between HS Grad and population
ggplot(df1,aes(x=HS.Grad,y=Population))+geom_point()
```




```
cor.test(x=as.array(df1$HS.Grad),y=as.array(df1$Population),
  method="pearson",use="complete.obs")
```

```
##
##  Pearson's product-moment correlation
##
## data:  as.array(df1$HS.Grad) and as.array(df1$Population)
## t = -0.68569, df = 48, p-value = 0.4962
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3667823  0.1849277
## sample estimates:
##         cor
## -0.09848975
```

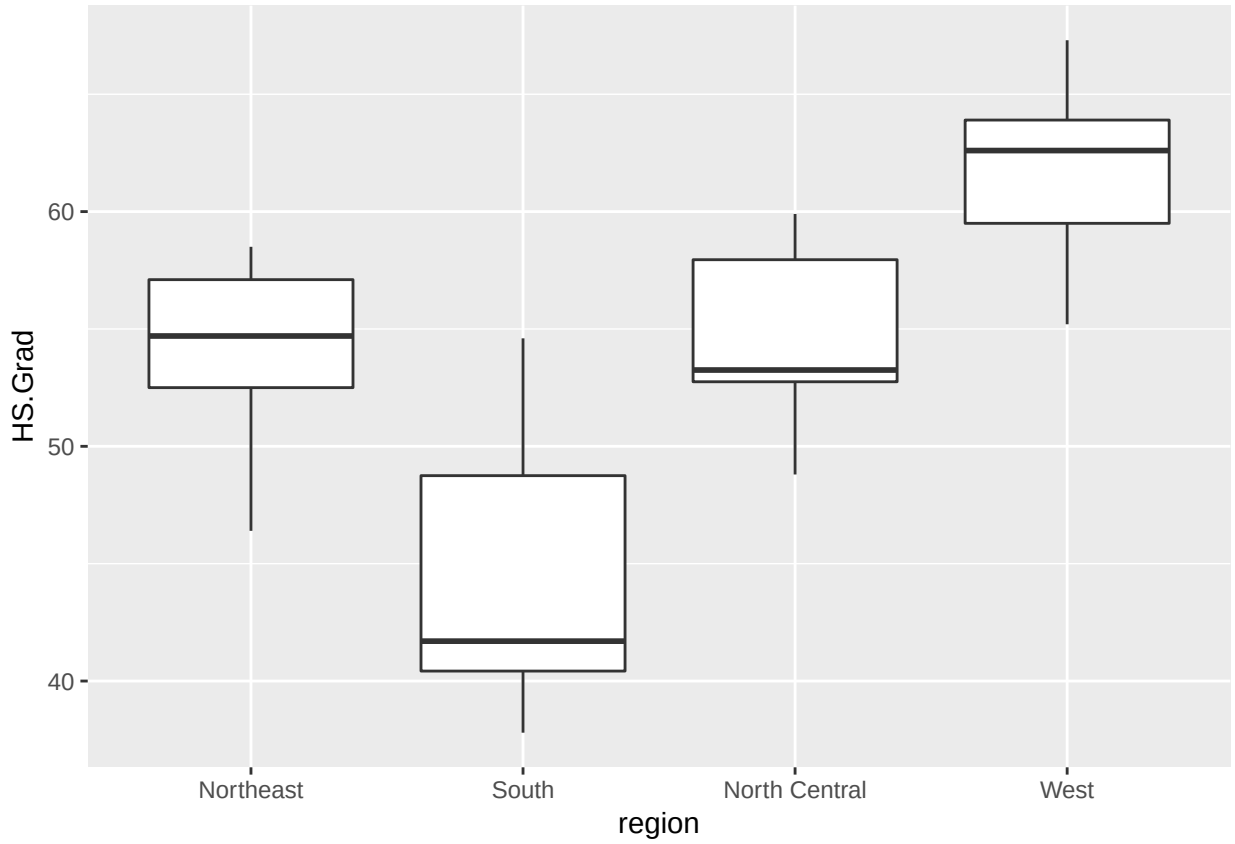
```
# Analysing relationship between HS Grad and region, division
df1 %>% group_by(region) %>%
  summarise(avg_HS_Grad=mean(HS.Grad,na.rm=TRUE))
```

```
## # A tibble: 4 x 2
##          region avg_HS_Grad
##          <fctr>       <dbl>
## 1     Northeast    53.96667
## 2         South    44.34375
## 3 North Central    54.51667
## 4          West    62.00000
```

```
df1 %>% group_by(region,division) %>%
  summarise(avg_HS_Grad=mean(HS.Grad,na.rm=TRUE))
```

```
## # A tibble: 9 x 3
## # Groups:   region [?]
##          region           division avg_HS_Grad
##          <fctr>             <fctr>       <dbl>
## 1     Northeast        New England    55.05000
## 2     Northeast    Middle Atlantic    51.80000
## 3         South     South Atlantic    45.72500
## 4         South East South Central    40.65000
## 5         South West South Central    45.27500
## 6 North Central East North Central    53.20000
## 7 North Central West North Central    55.45714
## 8          West           Mountain    61.41250
## 9          West            Pacific    62.94000
```

```
ggplot(df1,aes(y=HS.Grad,x=region))+geom_boxplot()
```



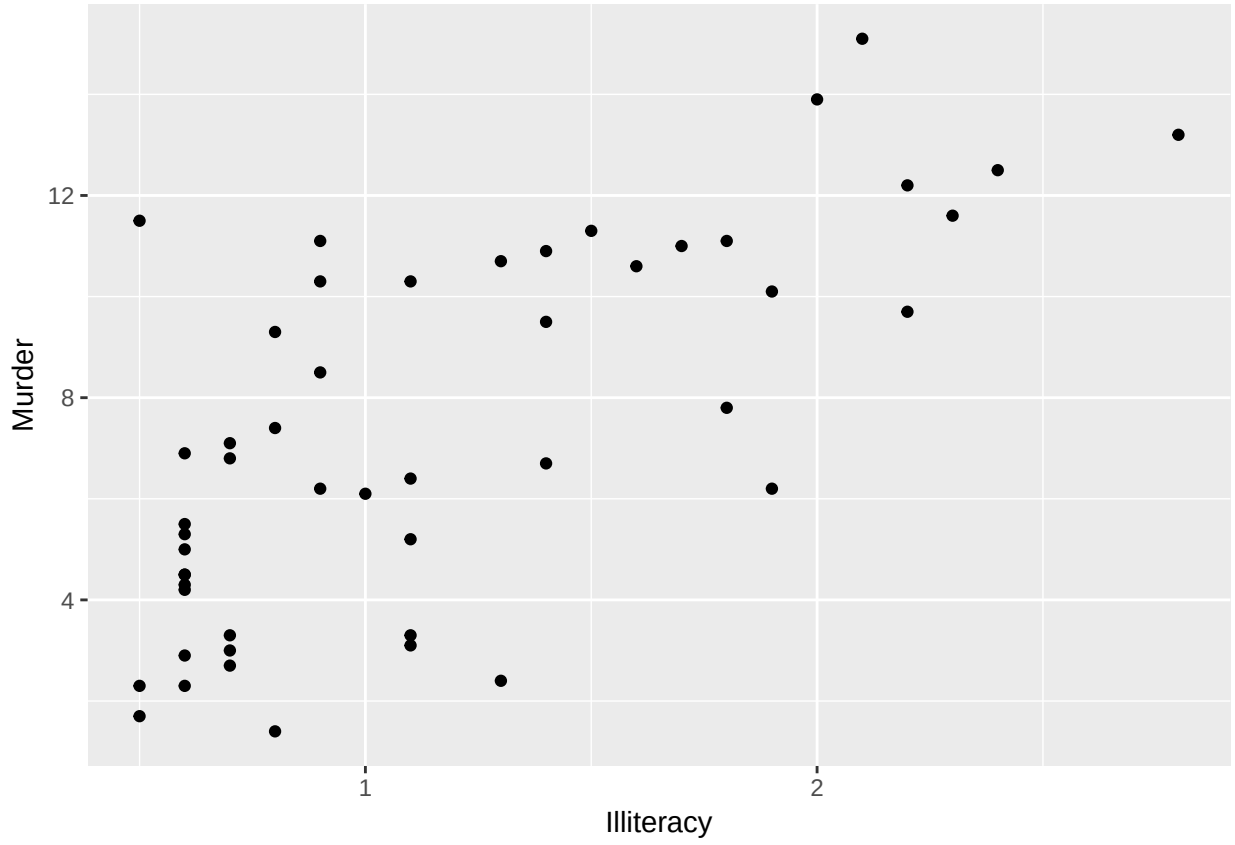

**(c) Develop a new research question of your own that you can address using the `state` dataset. Clearly state the question you are going to address. Provide at least one visualization to support your exploration of this question. Discuss what you find.**

Ans. Is there a relationship between illiteracy and murder rates in a state? Do states with high illiteracy proportions have high murder rates.

The data shows that these variables are correlated and have a positive relationship. States having high

illiteracy proportions do show indications of high murder rates. This indicates a correlation but not necessarily a causation. The pearson correlation coefficient of 0.70 and the scatterplot support this hypothesis.

```
#analysing relationship between illiteracy proportions and murder rates

ggplot(df1,aes(x=Illiteracy,y=Murder))+geom_point()
```




```
cor.test(x=as.array(df1$Illiteracy),y=as.array(df1$Murder),
  method="pearson",use="complete.obs")
```

```
##
##  Pearson's product-moment correlation
##
## data:  as.array(df1$Illiteracy) and as.array(df1$Murder)
## t = 6.8479, df = 48, p-value = 1.258e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5279280 0.8207295
## sample estimates:
##       cor
## 0.7029752
```

### Problem 3: Income and Education

The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.

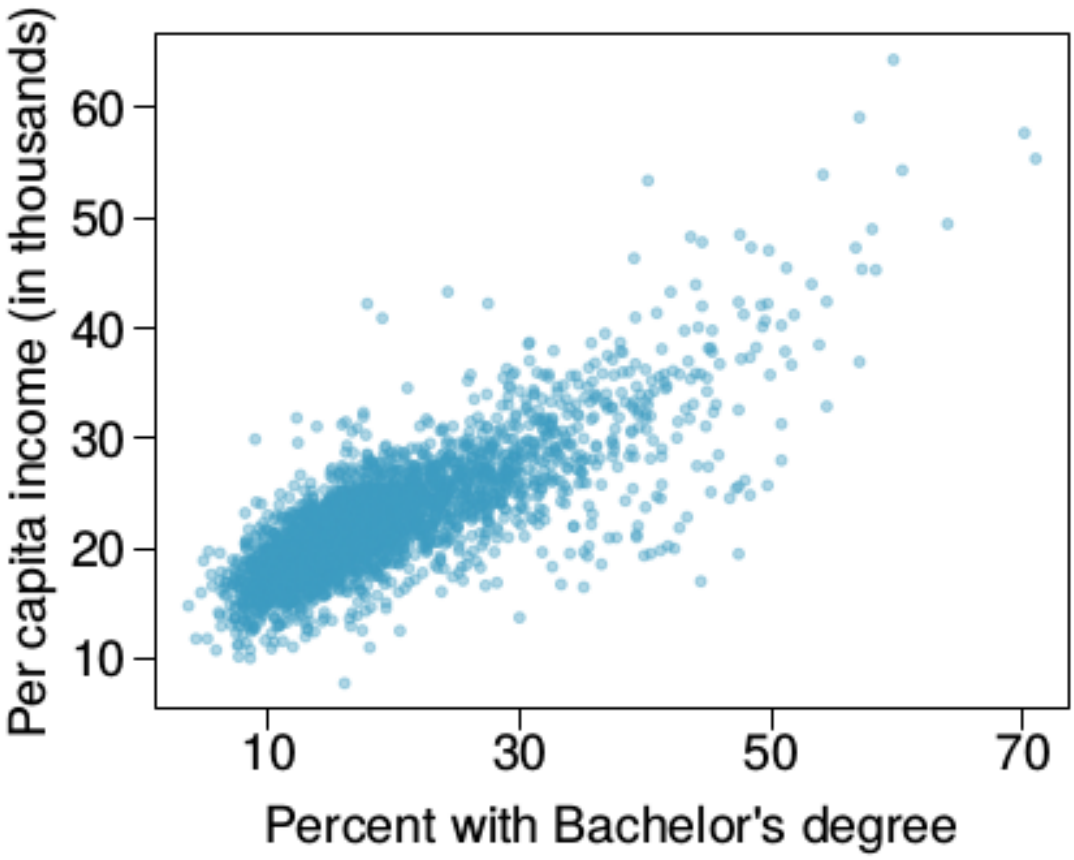



Figure 1: Income and Education in the US.

**(a) What are the explanatory and response variables?**

Ans. The explanatory variable here is the Percent of people with a Bachelors degree. The response variable here is the Per capita Income (In Thousands).

**(b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.**

Ans. The variables seem to have a positive correlation with each other and the relationship seems to look like a positive linear relationship. There are some unusual observations. There are observations in this data which have a low per capita income despite having a large percent of people with a Bachelors degree. There are also data points which have a high per capita income with a fewer percent of people with a bachelors degrees.

**(c) Can we conclude that having a bachelors degree increases ones income? Why or why not?**

Ans. We cannot conclude that having a bachelors degree most definately increases ones income. There might also be several other variables/factors other than having a bachelors degree which might increase ones income, for example- a persons intelligence, business acumen, communication skills, hard work, gambling skills etc. It could also be that a third variable like a persons intelligence helps increase one's income and leads him/her to also attain a bachelors degree.