IMT 573: Problem Set 3 - Data Analysis

Naga Soundari Balamurugan

Due: Tuesday, October 23, 2018 11:59AM

Collaborators: Jayashree Raman

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. If you are using more than just a standard function that you found from another source, please credit the source in the comments. For example:
4. Collaboration on problem sets is acceptable, and even encouraged, but students must turn in an individual write-up in their own words and their 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` or `Knit Word`, rename the R Markdown file to `YourLastName_YourFirstName_ps3.Rmd`, knit a PDF or DOC and submit both the PDF/DOC and the Rmd file on Canvas.

Setup:

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

```
# Load standard libraries
library(tidyverse)
library(nycflights13)
library(jsonlite)
library(kableExtra)
```

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.

Importing Data:

```
#Loading flight data
#List all the functions in the nycflights13 package
ls("package:nycflights13")
```

```
## [1] "airlines" "airports" "flights"  "planes"   "weather"
```

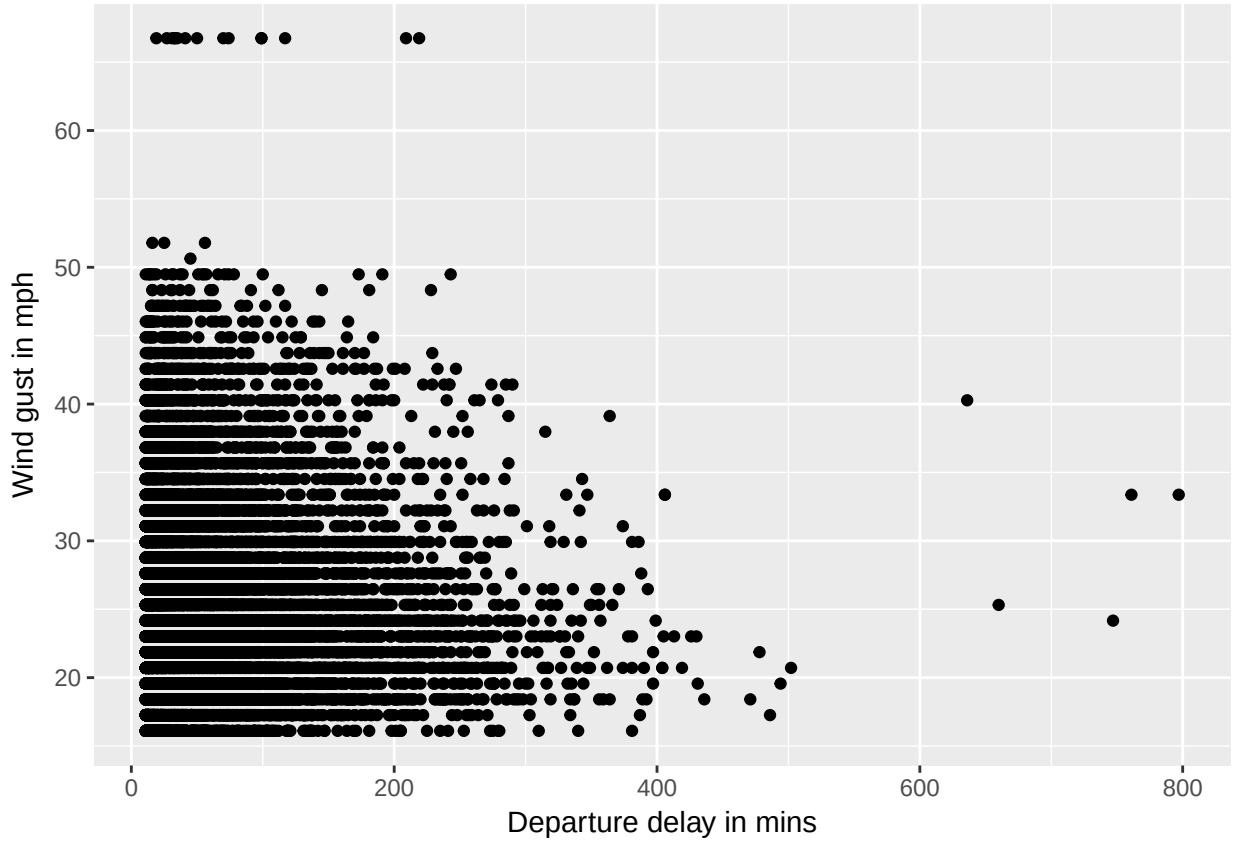
```
#Read in flights and Weather dataset
flightsData <- nycflights13::flights
weatherData <- nycflights13::weather

#Step 1: Merge flights data with weather data
completeData <- merge(flightsData, weatherData, by = c("origin", "time_hour", "year",
                                                       "month", "day", "hour"))

#Step 2: Filter the rows that has departure delays
dataWithDelay <- completeData %>% na.omit() %>% filter(dep_delay > 10)
```

Step 1: Both the datasets are imported and merged to form a single dataframe by the columns origin and time_hour.

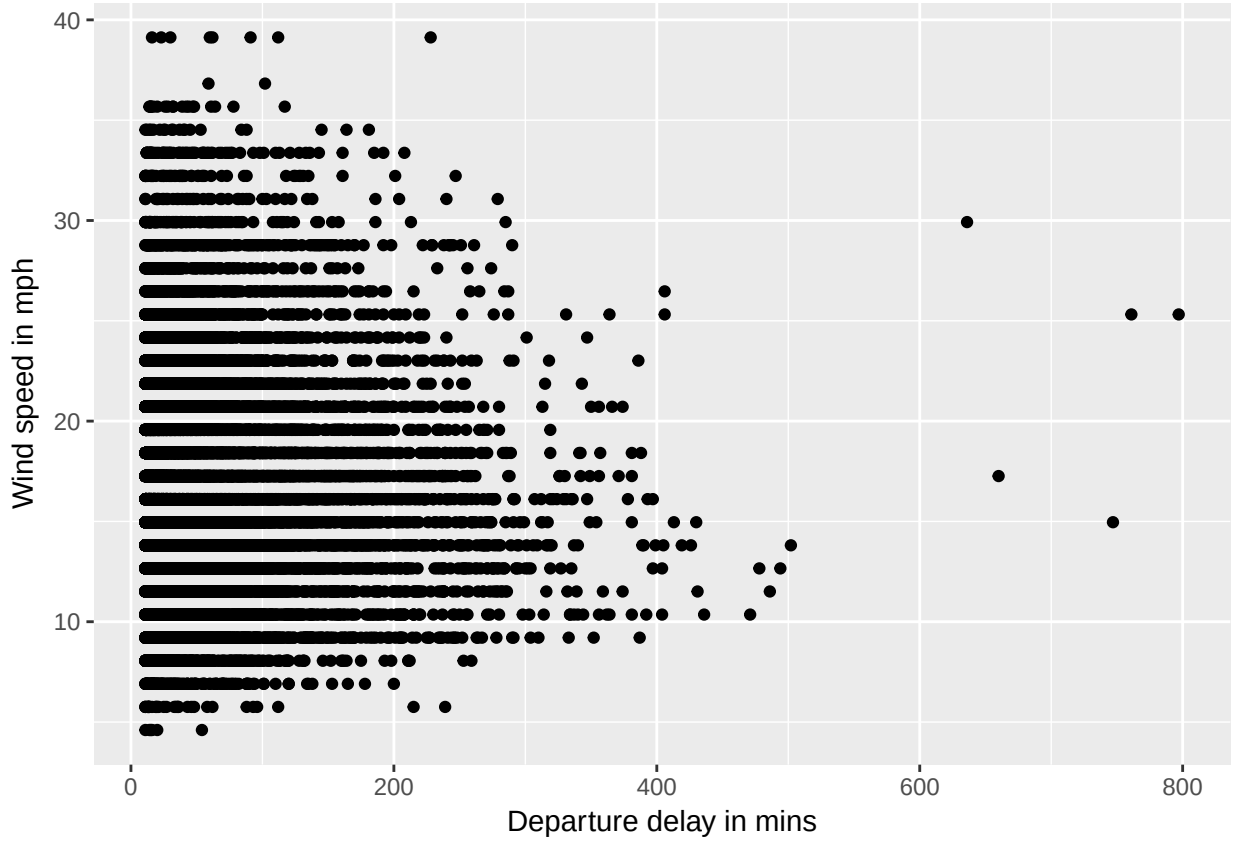
Step 2: The dataset is filtered on basis of dep_delay that are greater than 10 minutes. I opted to do the this filtration as we have the weather details for the NewYork airports and delays lesser than 10 minutes are not that significant. Also I have removed the rows that has NA as it does not provide any significance.

I would like to focus on the wind_gust and wind_speed as those would affect the takeoff in majority of cases.

```
#Delay vs Wind Gust
delayWithGust <- ggplot(data = dataWithDelay, aes(x = dep_delay, y = wind_gust)) +
  geom_point(stat = "identity") + xlab("Departure delay in mins") +
  ylab("Wind gust in mph")
delayWithGust
```

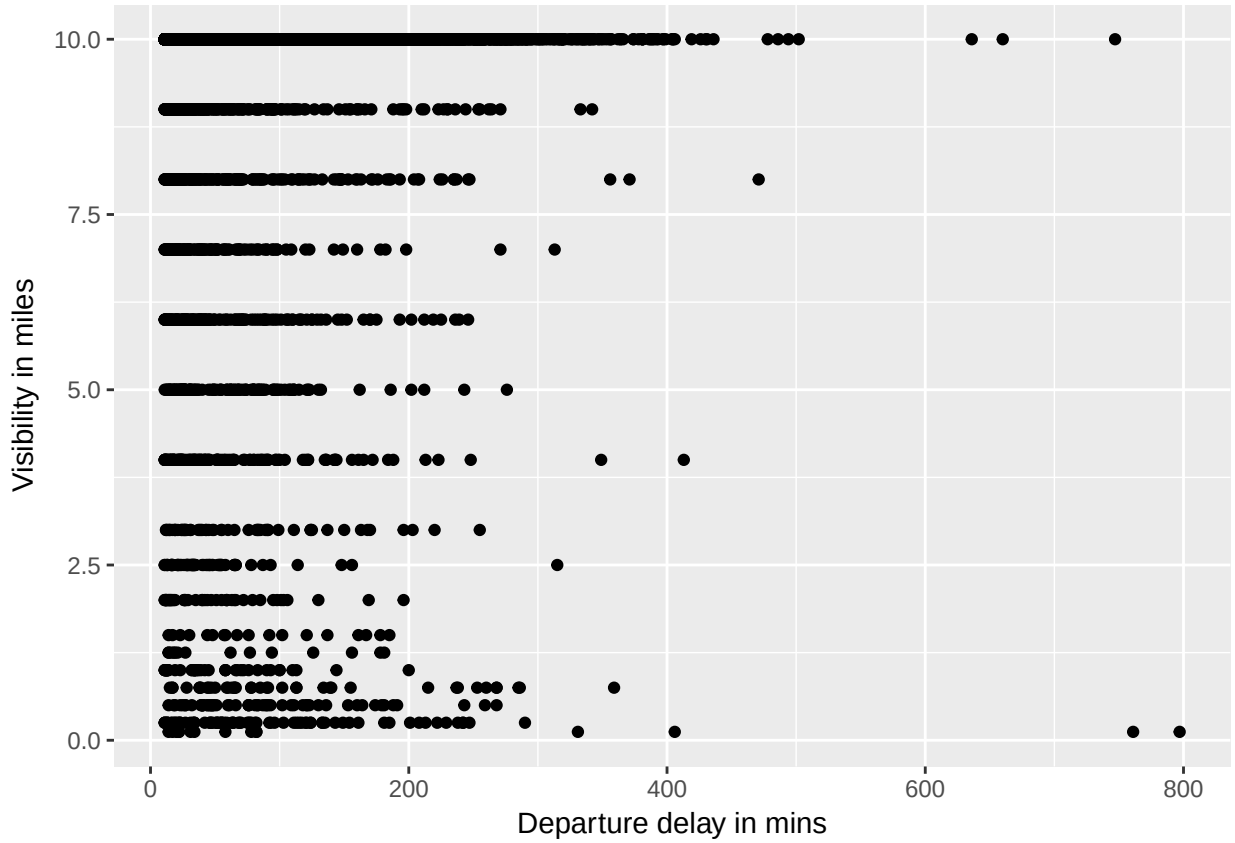


```
delayWithSpeed <- ggplot(data = dataWithDelay, aes(x = dep_delay, y = wind_speed)) +
  geom_point(stat = "identity") + xlab("Departure delay in mins") +
  ylab("Wind speed in mph")
delayWithSpeed
```

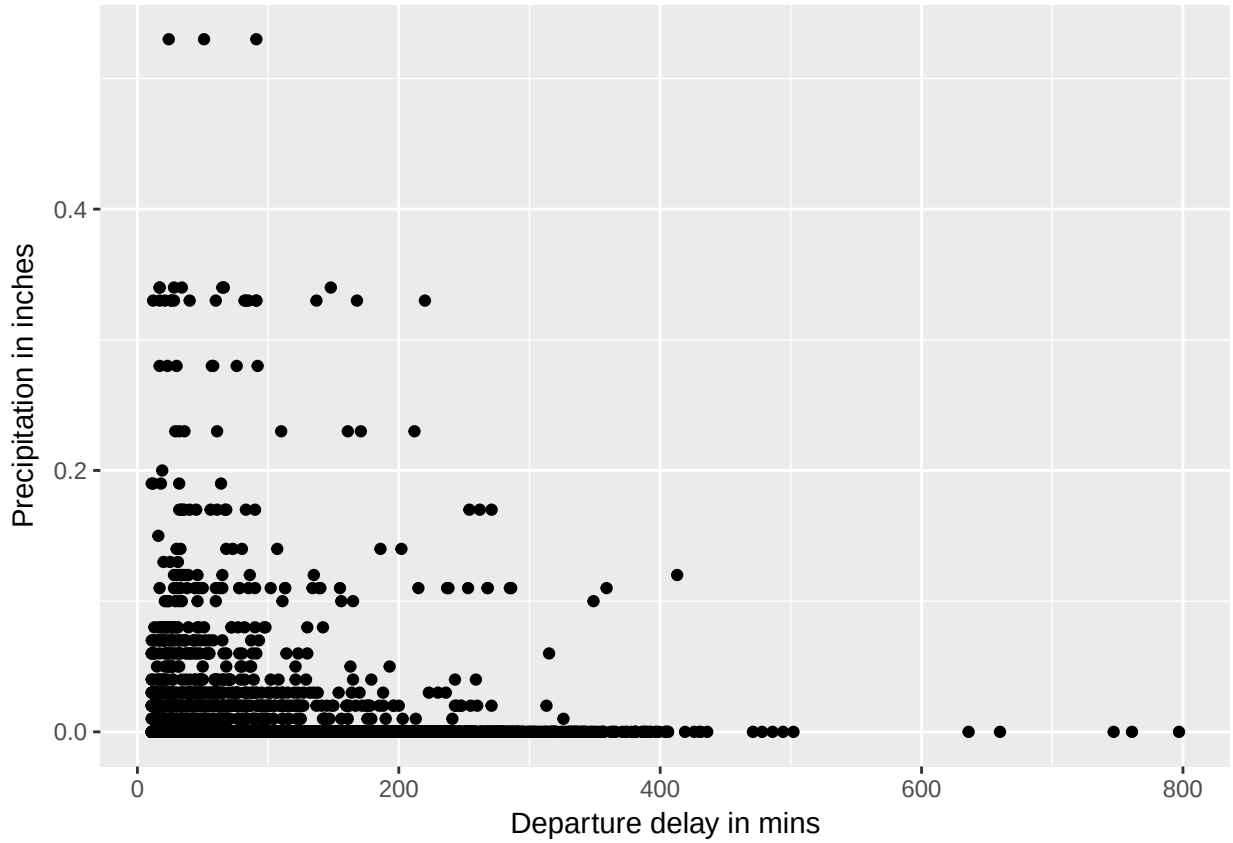


From the above plots, we see that there is no good correlation between departure delay and wind speed/gust. The data points are random. So lets explore for few other factors like visibility, precipitation, pressure etc.,

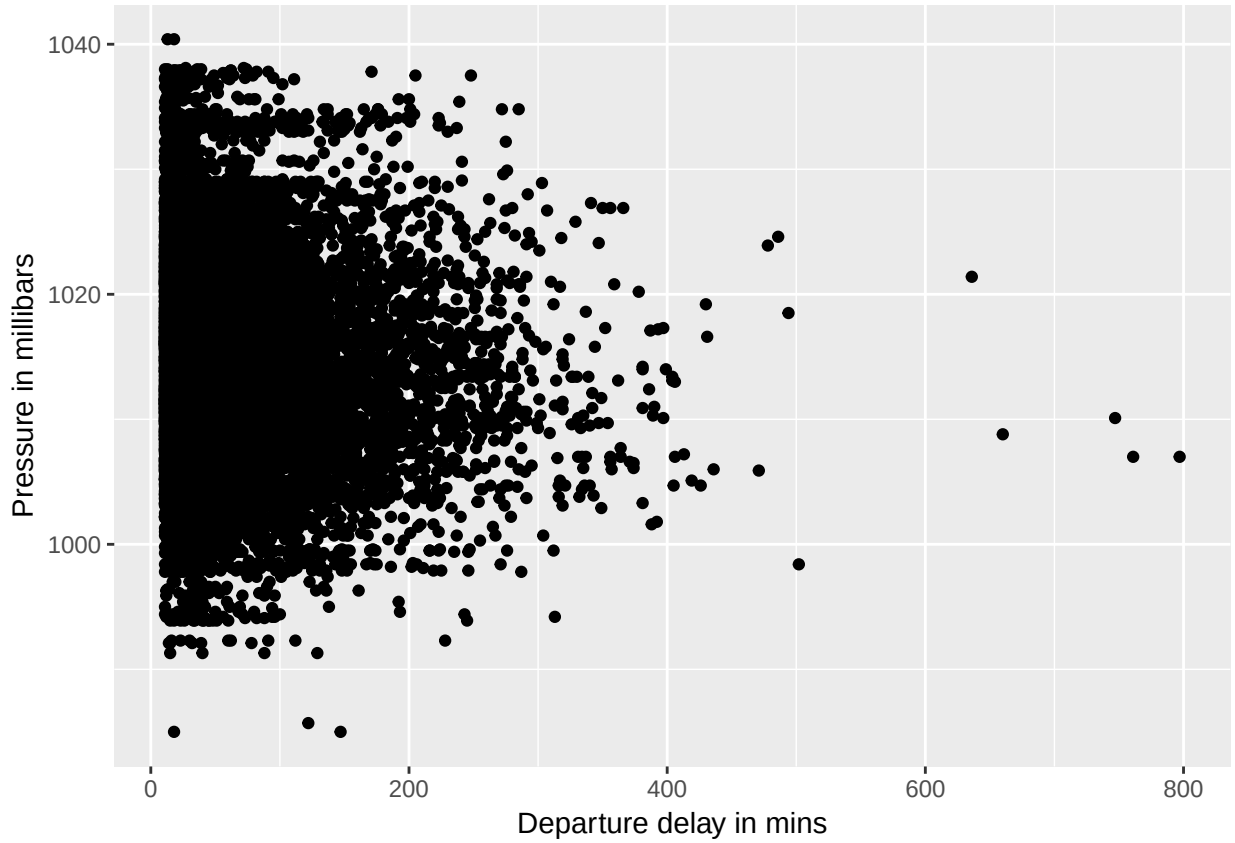
```
delayWithVisib <- ggplot(data = dataWithDelay, aes(x = dep_delay, y = visib)) +
  geom_point(stat = "identity") + ylab("Visibility in miles") +
  xlab("Departure delay in mins")
delayWithVisib
```



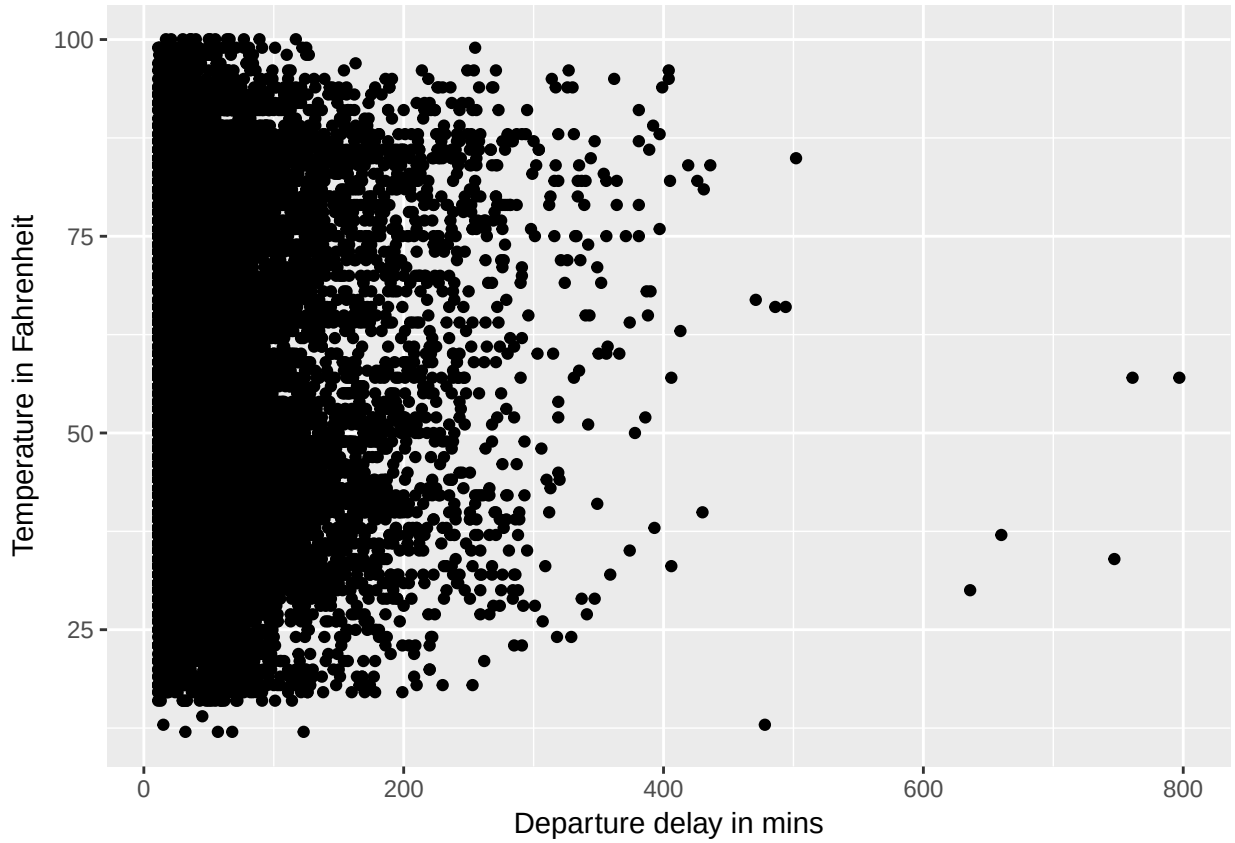

```
delayWithPrecip <- ggplot(data = dataWithDelay, aes(x = dep_delay, y = precip)) +
  geom_point(stat = "identity") + ylab("Precipitation in inches") +
  xlab("Departure delay in mins")
delayWithPrecip
```



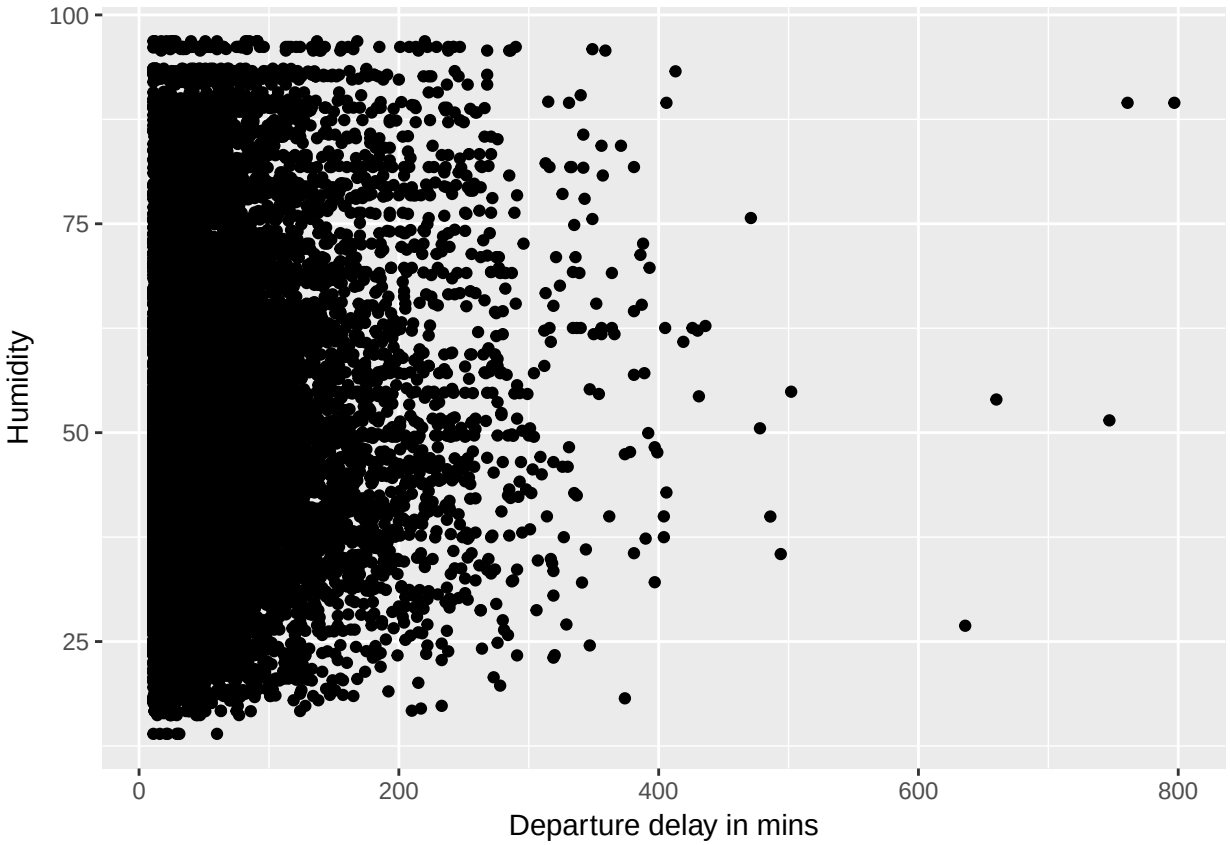

```
delayWithPressure <- ggplot(data = dataWithDelay, aes(x = dep_delay, y = pressure)) +
  geom_point(stat = "identity") + ylab("Pressure in millibars") +
  xlab("Departure delay in mins")
delayWithPressure
```




```
delayWithTemp <- ggplot(data = dataWithDelay, aes(x = dep_delay, y = temp)) +
  geom_point(stat = "identity") + ylab("Temperature in Fahrenheit") +
  xlab("Departure delay in mins")
delayWithTemp
```




```
delayWithHumid <- ggplot(data = dataWithDelay, aes(x = dep_delay, y = humid)) +
  geom_point(stat = "identity") + ylab("Humidity") + xlab("Departure delay in mins")
delayWithHumid
```



Among the above plots, only the precipitation has a good correlation with the departure delay. Though the departure delay in minutes increases as the precipitation decreases which is contradictory. Thus this needs to be digged deep to find the reason behind it.

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.

```
#Load the state dataset
stateData <- state.x77

no_of_rows <- nrow(stateData)
no_of_cols <- ncol(stateData)

#Change it as data frame
stateDataDF <- as.data.frame(stateData)

#Assign the row names to a new column names City
stateDataDF$City <- rownames(stateDataDF)

#Rearrange the columns
```

```
stateDataDF <- stateDataDF[c("City", "Population", "Income", "Illiteracy", "Life Exp", "Murder",
                             "HS Grad", "Frost", "Area")]
```

The dataset has 8 columns with 50 rows which denotes 50 states of United states. This dataset describes various factors of the states like population, average income, illiteracy rate etc., Tidy the data by converting it as a data frame and then creating a new column named 'City'. The rownames that has the list of the states is assigned to this column. The following are the column details.

- 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

As all the columns are of type integer, lets leave it that way.

(b) Suppose you want to explore the relationship between a state's `Murder` rate and other characteristics of the state, for example population, illiteracy rate, and more. Begin by examining the bivariate relationships present in the data. What does your analysis suggest might be important variables to consider in building a model to explain variation in murder rates?

```
#run correlation test
correlation <- cor(stateData)
kable(correlation, "latex") %>% kable_styling(bootstrap_options = c("striped", "hover"))
```

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Population	1.0000000	0.2082276	0.1076224	-0.0680520	0.3436428	-0.0984897	-0.3321525	0.0225438
Income	0.2082276	1.0000000	-0.4370752	0.3402553	-0.2300776	0.6199323	0.2262822	0.3633154
Illiteracy	0.1076224	-0.4370752	1.0000000	-0.5884779	0.7029752	-0.6571886	-0.6719470	0.0772611
Life Exp	-0.0680520	0.3402553	-0.5884779	1.0000000	-0.7808458	0.5822162	0.2620680	-0.1073319
Murder	0.3436428	-0.2300776	0.7029752	-0.7808458	1.0000000	-0.4879710	-0.5388834	0.2283902
HS Grad	-0.0984897	0.6199323	-0.6571886	0.5822162	-0.4879710	1.0000000	0.3667797	0.3335419
Frost	-0.3321525	0.2262822	-0.6719470	0.2620680	-0.5388834	0.3667797	1.0000000	0.0592291
Area	0.0225438	0.3633154	0.0772611	-0.1073319	0.2283902	0.3335419	0.0592291	1.0000000

From the above correlation table, we see that the murder is highly correlated with **Illiteracy and Life Expectancy**. It is positively correlated with Illiteracy(0.702) and negatively correlated with Life Expectancy(-0.78).

(c) Choose one variable and fit a simple linear regression model, $Y = \beta_1 X + \beta_0$, using the `lm()` function in R. Describe your results.

```
#Build Linear model
lm_Murder_Illiteracy <- lm(Murder ~ Illiteracy, data = stateDataDF)
summary(lm_Murder_Illiteracy)
```

```
##
## Call:
## lm(formula = Murder ~ Illiteracy, data = stateDataDF)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5315 -2.0602 -0.2503  1.6916  6.9745
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3968     0.8184   2.928   0.0052 **
## Illiteracy    4.2575     0.6217   6.848 1.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.653 on 48 degrees of freedom
## Multiple R-squared:  0.4942, Adjusted R-squared:  0.4836
## F-statistic: 46.89 on 1 and 48 DF,  p-value: 1.258e-08
```

```
lm_Murder_LifeExp <- lm(Murder ~ `Life Exp`, data = stateDataDF)
summary(lm_Murder_LifeExp)
```

```
##
## Call:
## lm(formula = Murder ~ `Life Exp`, data = stateDataDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7272 -1.6733 -0.1734  1.4909  4.8680
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  159.576     17.579   9.078 5.45e-12 ***
## `Life Exp`    -2.147      0.248  -8.660 2.26e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.33 on 48 degrees of freedom
## Multiple R-squared:  0.6097, Adjusted R-squared:  0.6016
## F-statistic: 74.99 on 1 and 48 DF,  p-value: 2.26e-11
```

Both the above model shows a statitically significant relationship between the variables as the p-values are less than 0.05 and has a high R-squared values. But the relationship between murder rate and life expectancy is more obvious and one depends on another. Thus it is not interesting enough to explore. Because in general, Murder rate and lefe expectancy are more like confounding variables. Thus Illiteracy rate would be an important variable to consider.

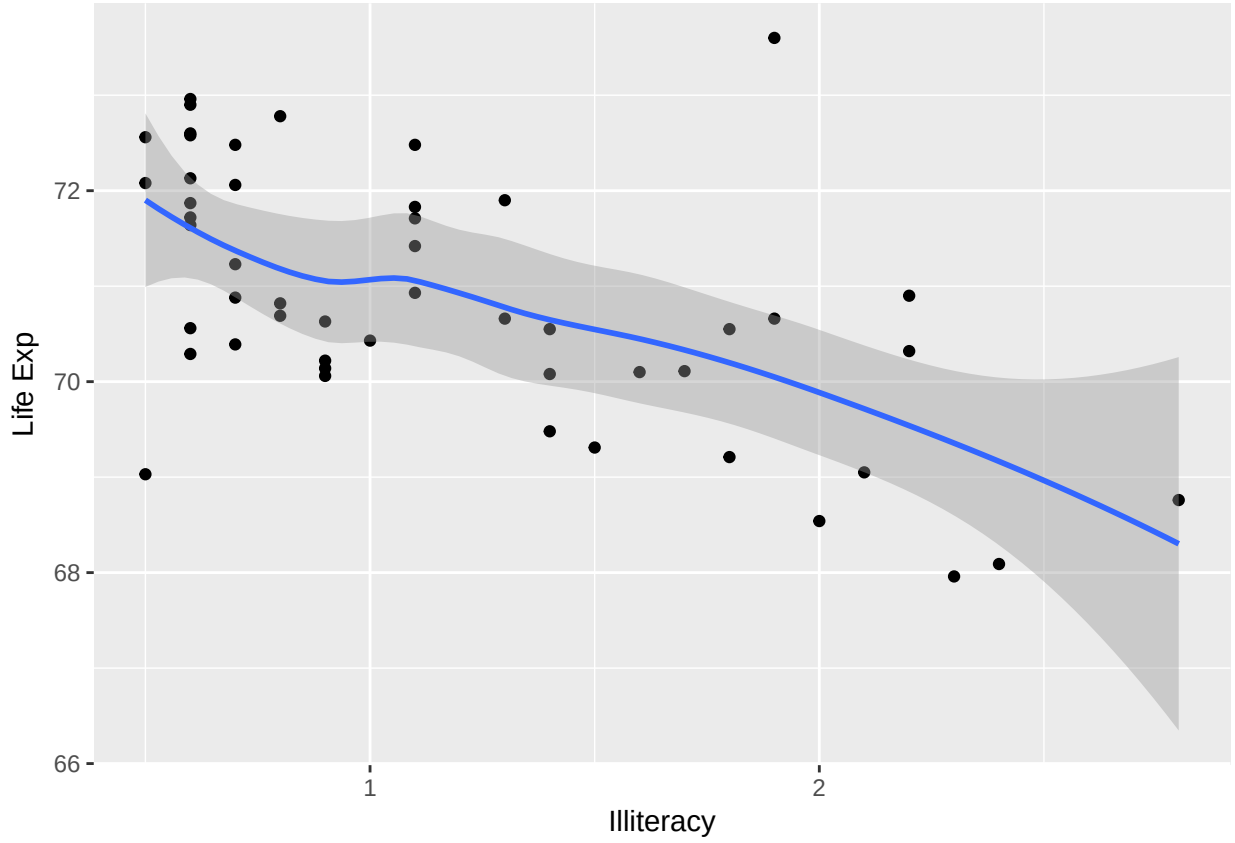
(d) Develop a new research question of your own that you can address using the `state` dataset. Clearly state the question you are going to address. Provide at least one visualizations to support your exploration of this question. Discuss what you find.

I would like to explore the relationship between illitercay rate and life expectancy. *Does the illiteracy rate affect the life expectancy?* People who are illiterate might have low income which could lessen their life expectancy. Lets explore it in the below graph.

```
#Illiteracy vs life expectancy
plotIlliteracy <- ggplot(data = stateDataDF, aes(x = Illiteracy, y = `Life Exp` )) +
  geom_point(stat = "identity") + geom_smooth()
```

```
plotIlliteracy
```

```
## `geom_smooth()` using method = 'loess'
```




```
#Linear model for illiteracy rate and life expectancy
lm_illit_lifeExp <- lm(`Life Exp` ~ Illiteracy, data = stateDataDF)
summary(lm_illit_lifeExp)
```

```
##
## Call:
## lm(formula = `Life Exp` ~ Illiteracy, data = stateDataDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7169 -0.8063 -0.0349  0.7674  3.6675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  72.3949     0.3383 213.973  < 2e-16 ***
## Illiteracy   -1.2960     0.2570  -5.043 6.97e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.097 on 48 degrees of freedom
## Multiple R-squared:  0.3463, Adjusted R-squared:  0.3327
## F-statistic: 25.43 on 1 and 48 DF,  p-value: 6.969e-06
```

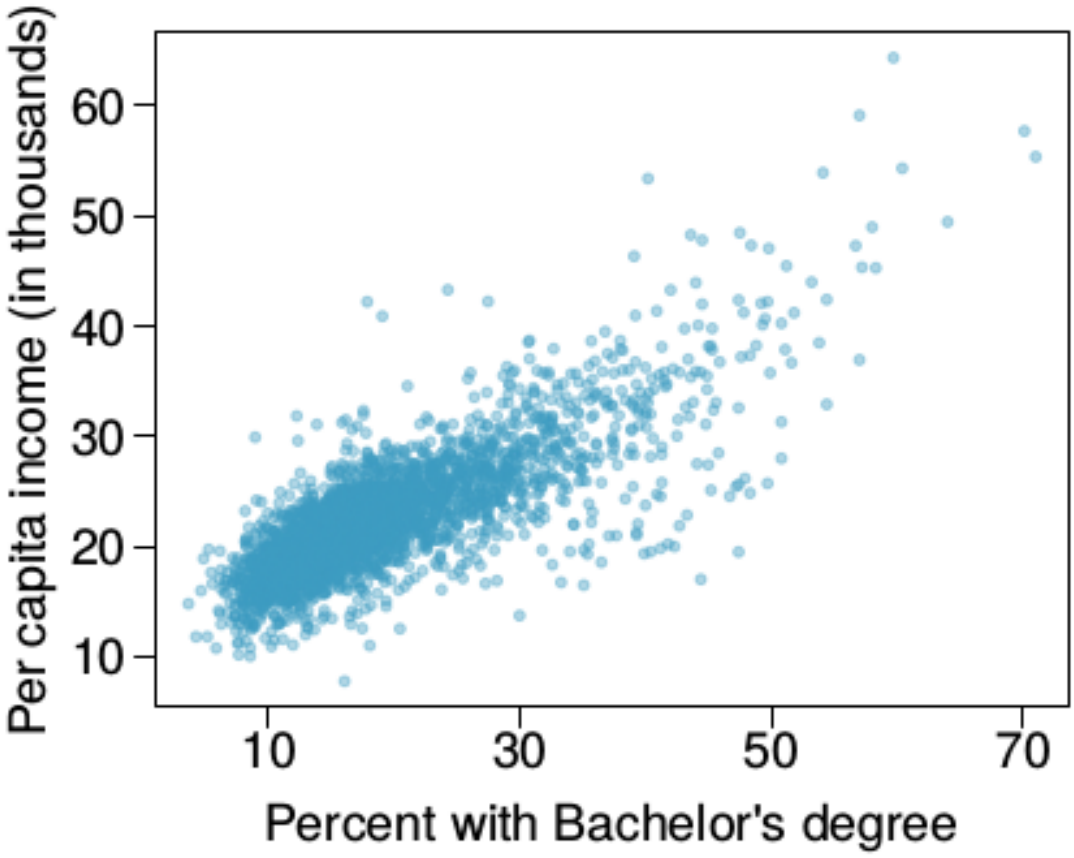


Figure 1: Per Capita Income and Education

From the above graph, we can see that the life expectancy decreases as the illiteracy rate increases. Also the linear model reproves this relationship with a significant p-value(6.969e-06) and coefficient value(-1.2960). This relationship might be cause of people with no education might land in manwork related job which could reduce their life expectancy.

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 bachelors degree in 3,143 counties in the US in 2010.

(a) What are the explanatory and response variables?

The percent with bachelor's degree(x axis) is the explanatory variable and Per capita income(y axis) is the response variable. This is because the per capita income depends on the percent with bachelor's degree. The reverse case might be true, but as the bachelor's degree is in the x-axis, that should be the explanatory variable.

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

Both the variables are positively correlated. The per capita income increases with the increase in percent with bachelor's degree. Most of the data points fall in the 10 - 30 percent bachelor's degree bucket. In the range 30 - 50 percent with bachelor's degree not all data points have a higher per capita income. There are many data points in this range that has low per capita income(<30K). Only few data points show a higher per capita income. Also we could see very few data points in the range 50 to 70 percent with bachelor's degree.

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

By seeing roughly we can say that the income increases if one has a bachelors degree. But as we take a detailed look into plot, we could come to the conclusion that *having a bachelors degree does not increase ones income*. This is because of various reasons like, * Majority of data points in the lower per capita income range. Since there is no even spread of data points, we couldnt decide on the correlation. * As mentioned in (b), in the range of 30% - 50% with bachelor's degree, most of the data points has the same per capita income as 10-30 percent with bachelor's degree range. Thus there is no significant increase in per capita income as of the increase in bachelor's degree percentage. * In addition, we have only fewer data points in the 50-70% range using which significant correlations cannot be found.