# lab2_homework

### *Pranab*

### *February 16, 2015*

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

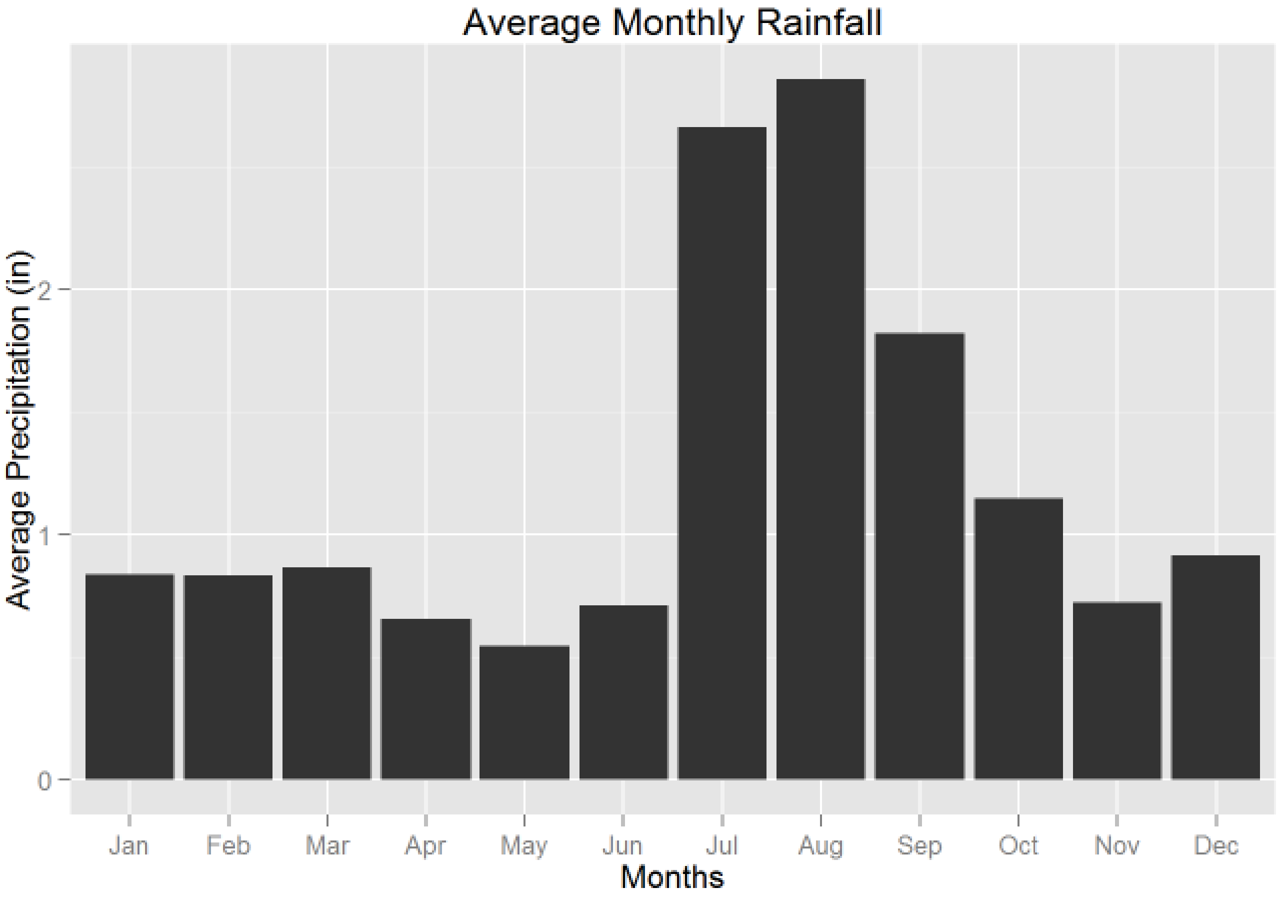
```
precip <- read.csv('precip.csv')
rwi <- read.csv ('rwi.csv')
```

# Part 1: Analysis of El Malpais Data:

1. Monthly Rainfall Pattern:

The monthly rain fall pattern for El Malpais may be observed through the average monthly rainfall bar diagram given below. Maximum rain fall generally occcurs around July, August and September. August, on an avaregae, receives the maximum rainfall at around 2.8 inches while July and September receives around 2.6 and 1.8 inches respectively. The rainfall minima occurs during the month of May, receiving around 0.5 inches on avarage.

```
temp.data <- gather(data=precip, key=month, value=precip, -year)
tidy.precip <- arrange(temp.data, year, month)
temp.data <- group_by(tidy.precip, month)
temp.data <- summarize(temp.data, precip=mean(precip, na.rm=TRUE))
ggplot(temp.data, aes(month, precip)) + geom_bar(stat='identity') + labs(y="Average Precipitation (in)", x="Months", title="Average Monthly Rainfall")
```

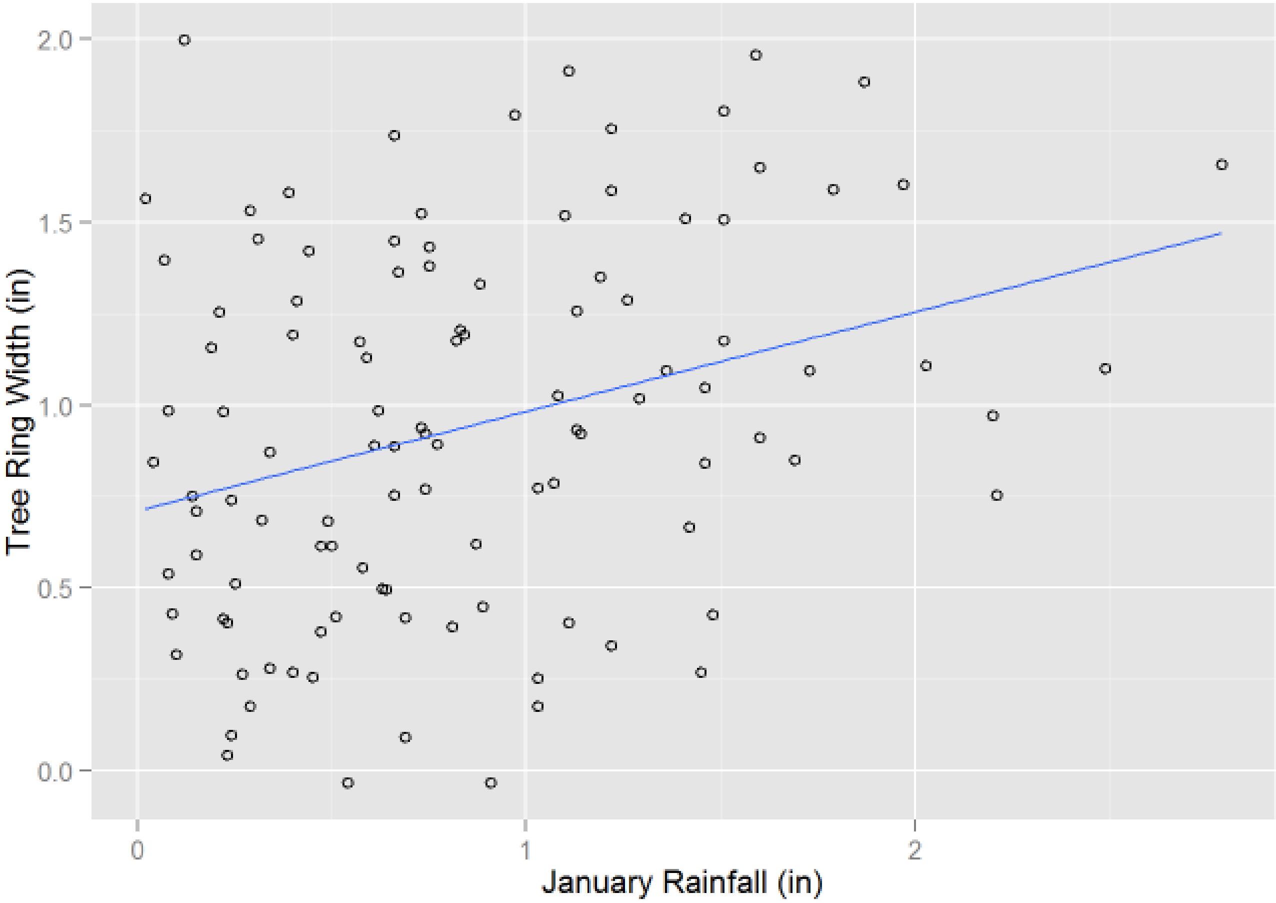


2. Correlation Betweeb Tree Ring Width and Rainfall:

```
rwi.precip <- left_join(precip, rwi, by='year')
ggplot(data=rwi.precip, aes(x=Jan, y=rwi)) + geom_point(shape=1)+labs(y="Tree Ring Width (in)", x="January Rainfall (in)")+ geom_smooth(method=lm, se=FALSE)
```

```
## Warning: Removed 10 rows containing missing values (stat_smooth).
```

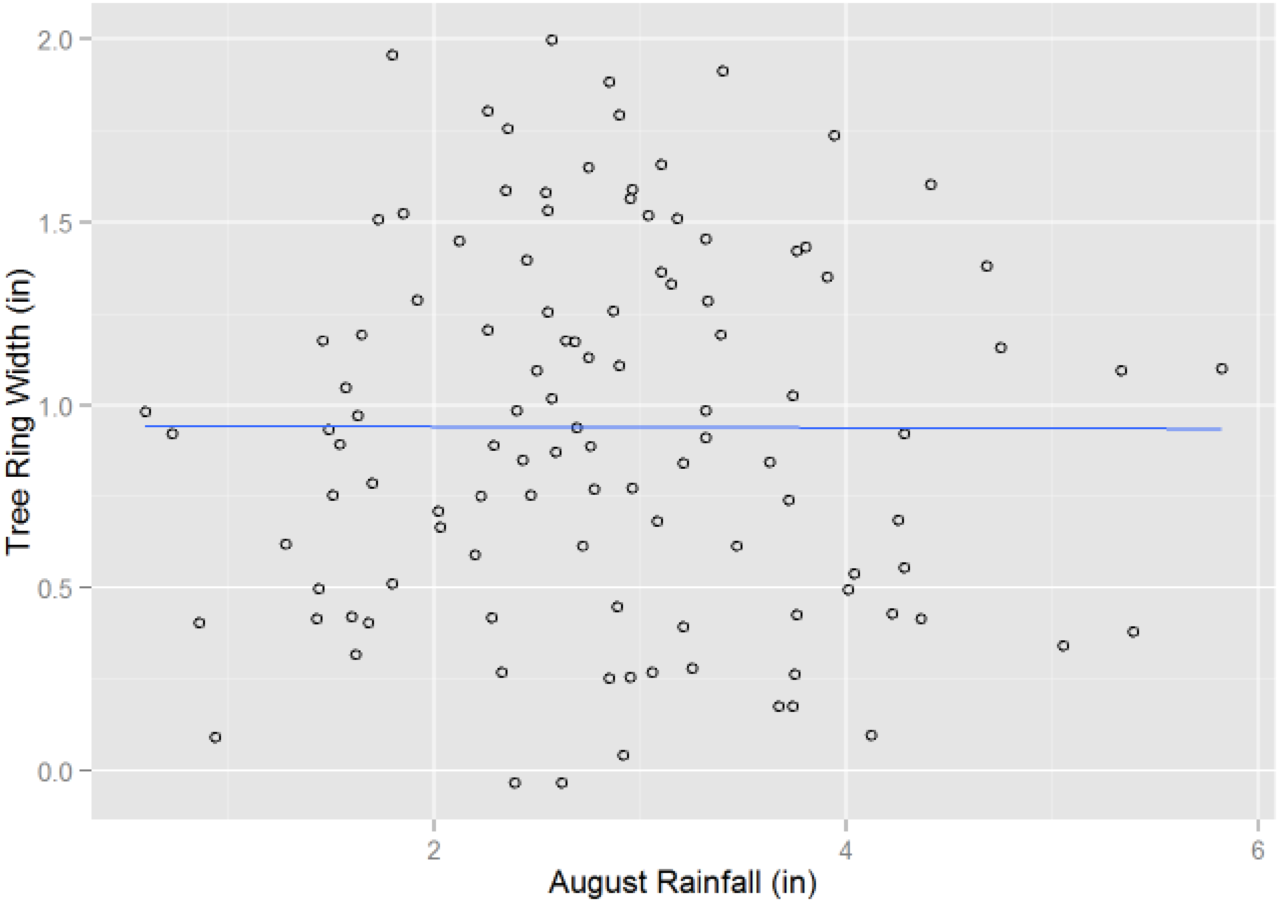
```
## Warning: Removed 10 rows containing missing values (geom_point).
```



```
rwi.precip <- left_join(precip, rwi, by='year')
ggplot(data=rwi.precip, aes(x=Aug, y=rwi)) + geom_point(shape=1)+labs(y="Tree Ring Width (in)", x="A
ugust Rainfall (in)") + geom_smooth(method=lm, se=FALSE)
```

```
## Warning: Removed 10 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 10 rows containing missing values (geom_point).
```



From the two scatterplots depicting correlation between observed tree ring widths and corresponding January and August rainfalls, it may be obserded that while the correlation between rainfall and tree ring width has somewhat moderate relation during January but it has almost no relationship (infact, a little negative) during August. This is highlighted by the slope of the regression lines drawn in these figures. Thus, from this two observations, it may be said that the tree ring growth is more correlated with rainfall around the month of January than Around August.

3. How appropriate is Linear Correlation as a descriptor of correlation between tree ring width and monthly rainfall:

Among other climatic factors, tree ring growth is affetced by temperature and rainfall. SOme of the other factors that affect a tree ring width are tree age and previous year’s growth. If a tree has seen vigorous growth in the preceeding year, then it may have somewhat less growth in the current year. Thus, the ring width is a function of all these different functions. Hence, if we try to express the correlation between the ring growth and one of these factors (rainfall in thi scase), it may not bring out the essence of this complex relationship.

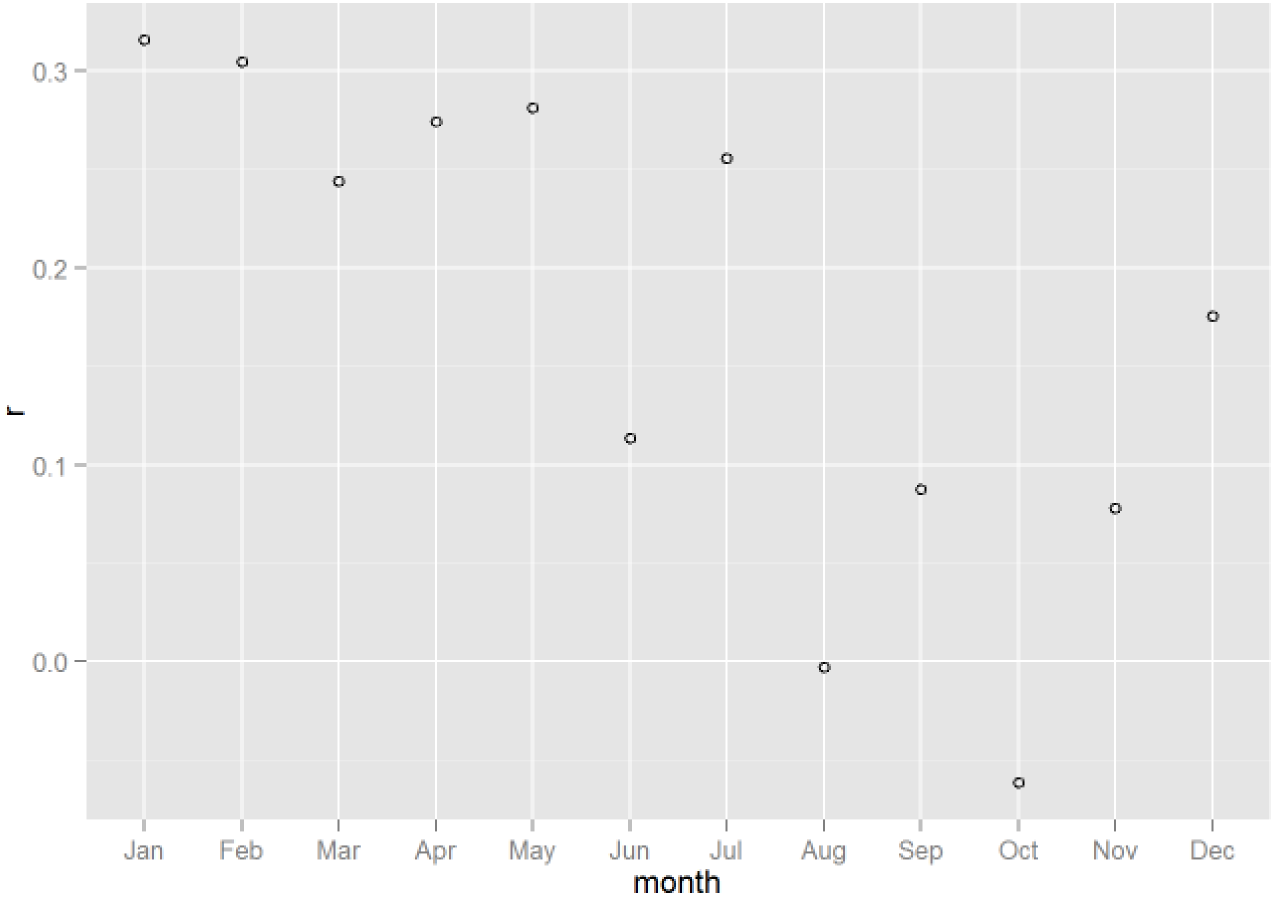
(Reference: http://www.climatedata.info/Proxy/Proxy/treerings_introduction.html (http://www.climatedata.info/Proxy/Proxy/treerings_introduction.html))

4. Correlation between rain fall each month and tree ring growth for present and previous year:

```
temp.data <- gather(data=precip, key=month, value=precip, -year)
tidy.precip <- arrange(temp.data, year, month)
rwi.precip <- left_join(rwi, tidy.precip, by='year') %>% filter(year>=1895)
cor_curr <- rwi.precip %>% group_by(month) %>% summarise(r=cor(rwi, precip))
cor_lag <- rwi.precip %>% group_by(month) %>% summarise(r=cor(rwi, lag(precip), use='complete.obs'))
```

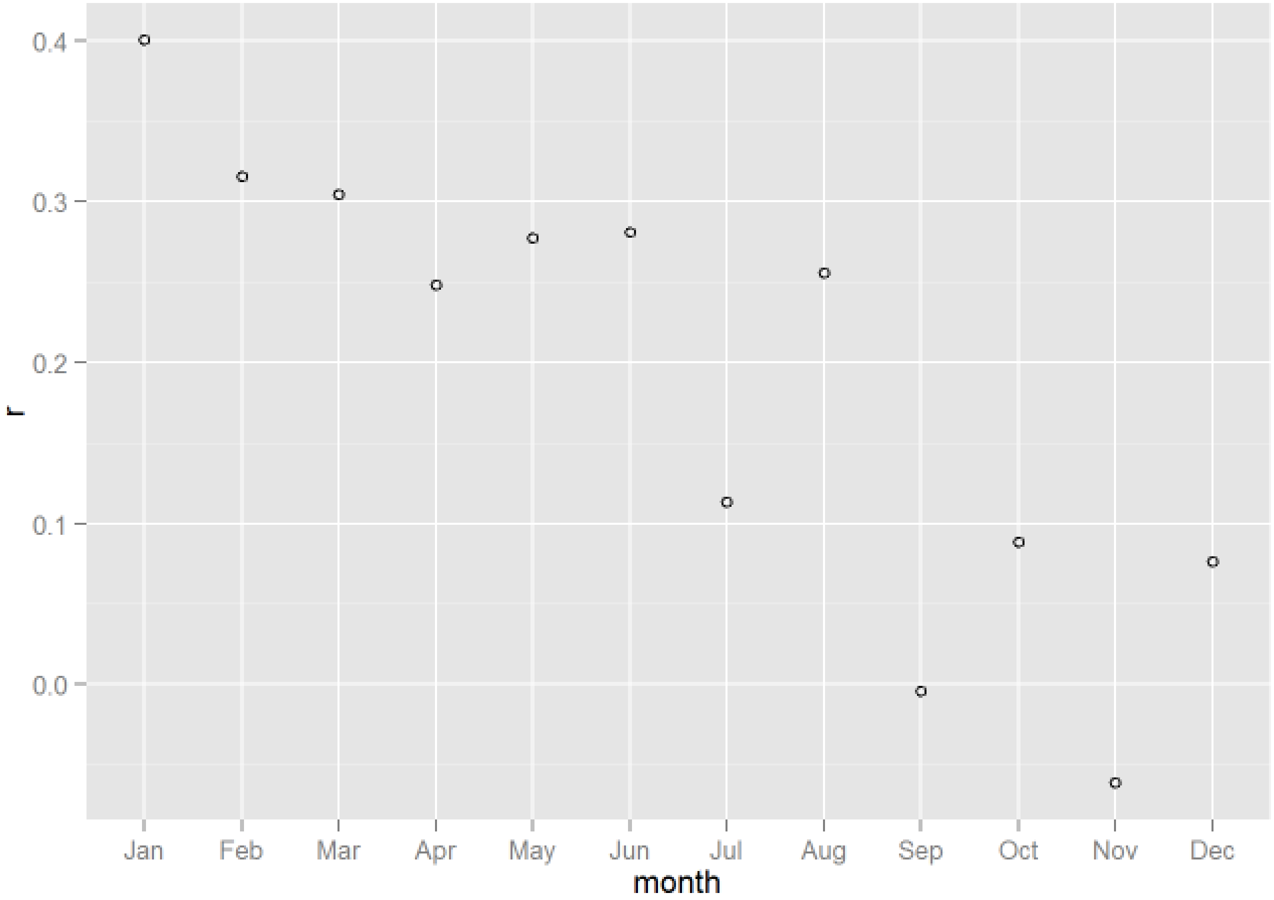
The plot between current month and ring width is:

```
ggplot(cor_curr, aes(x=month, y=r)) + geom_point(shape=1)
```



The plot between lagged time and ring width is:

```
ggplot(cor_lag, aes(x=month, y=r)) + geom_point(shape=1)
```



From the two scatter plots, it may be said that the correlation is more closely related between lagged month and tree ring growth. Thus this period should be chosen for plotting the relationship between precipitation and tree ring growth more realistically. The ring growth in present year is a function of the precipitation received by the tree during the previous year.Thus, from this point of view as well, the lagged time period is more appropriate for exploring the correlation between preciptation and tree ring growth.

# Part 2: Getting the interpretation of Correlation correct:

1. The second statement is true in th econtext of th efirst question. Let us see why the first statement is False before we explain why th esecond one is Correct.

If according to the first statement, even if January is above the average in receiving rainfall, it does not guarantee there will be a high corresponding ring growth. This fact does not correspond with the ring growth. Thus, this fact does not imply a correlation and the statement is False.

According to second statement, if wetter(dryer) than average Januarys have bigger(smaller) ring growths, then the increasing (decreasing) rainfall correspnds with high (low) ring growth which will implya positive correlation. Which makes this statement True.

2. The price of second hand cars goes down as they get old. Thus, an increase in Car’s price causes a decrese in it’s price. Thus, in this case the correlation will be Negative.

While in case of an Antique item, the price goes up with age of th eitem. So in case of an antique car, the age and price are Positively correlated. Thus, if antiques are included in sample, then the previous negative correlation will become less negative. But given that the antiques are rare compared to general used cars, the correlation will still be negative.

3. The correlation coeficient explains the proportion of variance in Y variable that may be accounted for by knowing

the value of X or the vice-versa. Thus, a correlation coefficient of 0.9 means that 90% of the variance in either variable may be accounted for by knowing the other one. Thus, this statement is False.

4. The economic study mentioned in the question finds a correlation of 0.4 between household income and teenage pregnancy rate.

As per various studies conducted by researchers, a strong correlation between low economic status and teenage pregnancy has been found. As per such a study published in Population Reference Beureau, I=teenage birth rates in Mississippi was 61 per 1000 women n the 15-19 age group while the same for New Hampshire was 18 in year 2005 (http://www.prb.org/Publications/Articles/2012/us-teen-birthrate-income.aspx (http://www.prb.org/Publications/Articles/2012/us-teen-birthrate-income.aspx)).

This correlation value takes into account all the states irrespective of their income conditions. Hence it cannot account for the regional variation that prevails in the sample.

(http://economix.blogs.nytimes.com/2012/04/03/income-inequality-and-teenage-pregnancy/ (http://economix.blogs.nytimes.com/2012/04/03/income-inequality-and-teenage-pregnancy/))

5. Possible correlation coefficient between GPA in Freshman and Sophomore year may have some weak positive correlation as generally brighter students are expected to keep up their good grades through the next year. Thus, the correlation coefficient could be 0.6.

The time gap between the Freshmen and Senior year in college is 4 years and many things that decide a student's grade may change. Thus, the correlation may be a weak positive relation of 0.3.

For lumber boards, the weight depends on volume of timber, which in turn, is a function of length. Thus, the correlation coefficient in this case could be 0.95.

6. The fact that hours spent on watching television and scores on reading tests are negatively correlated may not imply a causal relationship. Scores on reading tests depend on practice. Several other factors such as education level, guidance also matter in obtaining a good score. Thus, to assume that watching television lowers one's capability to read will not be based on facts. This has also been proven by scientific studies.

(http://ecademy.agnesscott.edu/~mzavodny/documents/EER_tv_001.pdf (http://ecademy.agnesscott.edu/~mzavodny/documents/EER_tv_001.pdf))