Homework 2

**#Set up your workspace:**

precip <- read.csv(file="C:/Users/Emily/Documents/enield/homework2/data/precip.csv")

precipEN <- read.csv(file="C:/Users/Emily/Documents/enield/homework2/data/precipEN.csv")

rwi <- read.csv(file="C:/Users/Emily/Documents/enield/homework2/data/rwi.csv")

rvalues <- read.csv(file="C:/Users/Emily/Documents/enield/homework2/data/rvalues.csv")

library(ggplot2)

library(dplyr)

library(tidyr)

library(magrittr)

temp.data <- gather(data=precip, key=month, value=precip, -year)

head(temp.data)

tidy.precip <- arrange(temp.data, year, month)

head(tidy.precip)

temp.data <- mutate(tidy.precip, date.str = paste(year, month, '01'))

head(temp.data)

temp.data <- mutate(temp.data, date=as.Date(date.str, '%Y %B %d'))

head(temp.data)

**#Plot of the monthly average rainfall:**

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='Precipitation (in)')

**#Group two spreadsheets:**

rwi.precip <- left\_join(precip, rwi, by='year')

cor(rwi.precip$rwi, rwi.precip$Jan)

cor(rwi.precip$rwi, rwi.precip$Jan, use='complete.obs')

cor(rwi.precip$rwi, rwi.precip$Jul, use='complete.obs')

cor(rwi.precip$rwi, lag(rwi.precip$Dec), use = 'complete.obs')

rwi.precip %$% cor(rwi, lag(Dec), use='complete.obs')

rwi.precip <- left\_join(rwi, tidy.precip, by='year') %>% filter(year>=1895)

rwi.precip %>% filter(month=='Jan') %$% cor(rwi, precip, use='complete.obs')

rwi.precip %>% filter(month=='Dec') %$% cor(rwi, lag(precip),use='complete.obs')

rwi.precip %>% group\_by(month) %>% summarise(r=cor(rwi, precip))

**#Scatterplot of Jan. Precip vs Tree Ring Width**

ggplot(data=precipEN) + geom\_point(aes(x=Jan, y=rwi)) + ylab('Tree Ring Width ')+xlab('Precip in Jan. (inches)')

**#Scatterplot of August. Precip vs Tree Ring Width**

ggplot(data=precipEN) + geom\_point(aes(x=Aug, y=rwi)) + ylab('Tree Ring Width ')+xlab('Precip in Aug. (inches)')

**#Lagged Correlation:**

rwi.precip %>% filter(month=='Jun') %$% cor(rwi, lag(precip), use='complete.obs')

**#Plot the two r values per month**

ggplot(data=rvalues) +

geom\_point(aes(x=Month, y=r, color='r'), size=3) +

geom\_point(aes(x=Month, y=rlagged, color='rlagged'), size=3) +

ylab('R Value') + xlim(0, 12) + ylim(0, 0.45) + scale\_x\_continuous(breaks=0:12)

**Part 1 Questions**

Precipitation is highest in the summer months (July, August) and begins to decline in the fall months. It is relatively constant in the winter months (Jan-Mar) at approximately 0.8 inches and it is at its lowest in the spring months (Apr-Jun). This pattern is shown in the figure below.

![Rainfall Plot](images/Rainfall.png)

The current December precipitation values has a low correlation with tree ring width of that year because the growing season of the tree is over. So this extra water will not influence the growth of the tree for that current year. The previous December precipitation values have a stronger correlation to the current tree year growth because this water (or snow) is stored right before the growing season begins. One can look at the graph and see where the growing season of the tree begins and what month’s precipitation affects it the most. The months that have the R-value greater than the R-lagged value are part of the growing season (Jan to June of the current year). The months that have the R-lagged value greater than the R-value influence the growing season by accumulating water (July to December of the previous year).

![R Plot](images/Rplot.png)

This relationship can be seen in the scatter plots of the current January vs tree ring width and of the current December vs the tree ring width. The scatter plot of the current January shows a greater correlation than that of the current August. This argues that the current January is part of the tree’s growing season but not the current August.

![Jan](images/Jan.png)

![Aug](images/Aug.png)

In summary the tree grows in the summer (June to August) and uses the water that it stores from July of the previous year to June of the current year. The more water that falls in these months the greater the tree can grow. This would make it appropriate to use a linear correlation to describe the relationship between tree ring growth and precipitation. The correlation won’t be strong because there are other factors that influence tree growth.

**Part 2 Questions**

1. The following statement is true: A positive correlation between tree ring width and January precipitation might happen if wetter (dryer)-than-average Januarys tend to have bigger (smaller)-than-average tree ring width.

A positive correlation means that as one variable increases (or decreases) the other variable also increases (or decreases). In this statement as the rainfall increased, the tree ring width also increased. Therefore they are positively correlated.

1. In general as the age of the second car increases the price decreases. This describes a negative correlation between the two. The price decreases as a car gets older because it has more kilometers on it and more parts fall apart due to wear and tear making its monetary value decrease.

As antique cars age they increase in value. So there is a positive correlation between the two. Antique cars represent a small proportion of the overall car market so if they were included with the rest of the aging cars than the correlation between price and age will still be negative but the magnitude of the r value would decrease.

1. False. The magnitude of the correlation coefficient tells you the overall strength of your relationship. R=0.9 does not translate to 90% of the points being highly correlated. This is a measurement of the overall spread of the data. It takes into account all of the points so if 90% of the data lines up almost perfectly but 10% is highly scattered this will not produce an R-value of 0.9 as there is a large spread in the overall data set.
2. The correlation of 0.4 between the average household income and rates of teenage pregnancy is higher than the actual relationship between the individual household income and teen pregnancy rate. This is due to the Ecological Correlation which states that the aggregation of data changes the correlation of the data.
   1. Grade point average in freshman year and sophomore year in college. R=0.6. Between freshman and sophomore year the studying habits of most students will not change significantly. A person with a GPA of 3.2 will most likely maintain that GPA in the next year.
   2. Grade point average in freshman year and senior year in college. R=0.3. Between freshman and senior year of college the studying habits of an individual will most likely improve so their overall GPA will be higher than their freshman score. But other students might decrease in their overall GPA as their classes become more difficult. This justifies R=0.3
   3. Length and weight of two-by-four boards of lumber. R=0.95. The volume of a 2x4 and its weight is strongly correlated. As the length increase there is more wood therefore it should be heavier. There might be small variations because of the width.
3. Studies find a negative correlation between hours spent watching television and scores on reading tests. Does watching television make people less able to read? Discuss briefly.

A correlation between watching TV and reading test scores does not mean that watching TV causes people to perform badly on reading test it just means that there is an association between the two. In general the more TV that a person watches the poorer they perform on reading tests. This could be because a person spends more of their free time watching TV than reading which makes them perform worse. In order to get a complete sense of the picture more variables need to be taken into consideration to see if there could be other causes. This can include the hours spent reading vs the hours watching TV.