# Problem Set 11

1.

(a)

```
# (a)
```

```{r}
library(ggplot2)
# Read the data
unusual <- matrix(scan("https://mtrosset.pages.iu.edu/StatInfeR/Data/unusual.dat"), ncol = 2, byrow = TRUE)
```

```{r}
unusual.df <- data.frame(unusual)
ggplot(unusual.df, aes(sample = X1)) + stat_qq() + ggtitle("normal probability plot of the x values")
```

From what it looks like the data doesn't have a extreme skeweness or outliers that would contradict approximate normality. So we can say that the x values are approximately normal.

(b)

```
# (b)
```

```{r}
ggplot(unusual.df, aes(sample = X2)) + stat_qq() + ggtitle("normal probability plot of the y values")
```

The line is approximately close to a straight line without any extreme skeweness or outliers. So there is a high probability that it was drawn from a normal distribution. We can that the y values are approximately normal.

(c)

```
# (c)

There is another way to check for normality in bivariate data which is by plotting a scatter diagram of the two columns and see if there is an ellipse like relationship that can be superimposed on to it
```

```{r}
source("https://mtrosset.pages.iu.edu/StatInfeR/binorm.R")
binorm.scatter(cbind(unusual.df$X1, unusual.df$X2))
```

One thing we notice is that there is no obvious relationship between the two variables but that doesn't mean that the two variables are not normal. We are interested to know if the values appear to have been drawn from a bivariate normal distribution and from what we can see, it doesn't look like the data was drawn from a bivariate normal because the ellipse does a horrible job of describing the shape of the data as there is a line of data going from top left to bottom right which is being left out. For the data to be considered a bivariate distribution, the ellipse should describe the data as close as possible.

```{r}
cor(unusual.df$X1, unusual.df$X2)
```

```
[1] 0.3244167
```

2.

If we are looking at how the performance decreases following the praises in terms of regression, then it's probably because the performance regresses back to the mean. The first performance was very likely not due to praise and possibly due to a high level of focus and balance in the moment. However, after the praise the focus may have been disturbed or the individual couldn't keep up the same level of high attention for a prolonged period of time or it could be due to any other factor as well. Thus, after some good performance, there is a natural tendency to regress back to the average again.

3.

(a) I think this suggestion is not accurate because the correlation is present but weak at 0.5. So the scores don't exactly carry over the same. Additionally, every student performs differently on every test. There could be other factors that can influence Jill's score to be lower also. Moreover, the average score for test 2 is 64 points with a standard deviation of 12 points. Because Jill scored 80 on test 1 and is closer to the average, it would make sense to assign Jill a closer value to the average in test 2. So I would advise the professor to assign 67 points to Jill in test 2.

(b) Again, the correlation is positive but is not large enough to say that the scores will always be 1 standard deviation above the mean for Jack. Additionally, the mean being 75 points with a standard deviation of 10, it's more likely that Jack's score 75 instead of the +1 SD value 85. Moreover, the unpredictable or hidden factors that can influence their score is not known, so if we should take those factors also into account, then I would advise the professor to assign a score of 80 to Jack on test 1.

4.

(a) So on average every time a team seems to be winning 81 games in a season with plus or minus 12 games. This also means rather than scoring 98 wins the next season, their number of wins will be closer to the average which is 81. Even if we were to base our decision solely on the fact that the top team won 98 games last season, then we have to see if every situation in the next season plays out the same way as last season. But that's usually not the case because the other teams would have up-ed their game and adapted to Los Angeles Angels' strategies. If we were to quantify this relationship between seasons using statistical methods, we find that there is only a 0.54 strength in proportionality between the two. This means that there is some connection for the next season based on the previous season but it does not translate to a perfect carryover of wins from one season to the next. There's still a lot of variability and unpredictability in each team's performance.

(b)

```
# (b)

There is a regression formula that we can make use of to make predictions using only the given data
```{r}
mean_wins <- 81
correlation <- 0.54
actual_wins_2014 <- 98

predicted_wins_2015 <- mean_wins + (correlation * (actual_wins_2014 - mean_wins))
predicted_wins_2015
```
```

```
[1] 90.18
```

(c) The regression predictions that you see here are only based on the pattern in each team's performance in the previous seasons. There are always factors that are unpredictable and so there is a high chance for an outlier to appear and win 96 games. They are most likely to win 91 games but we are not rejecting the fact that an unforeseen circumstance or a hidden factor can change this estimate. What the regression does is take the estimate closer to the average for the next season but we should not reject possible outliers. Perhaps having more data will help the model make a better estimate but with the limited data at hand, this estimation isn't misplaced.

5.

```
```{r}
adults <- read.table("adults.txt", sep = " ", header = TRUE)
adults
```
```

(a)

```
```{r}
library(ggplot2)

ggplot(adults, aes(x = Height., y = Weight)) + geom_point() + geom_smooth(method = "lm", se = FALSE) + labs(x = "Height", y
= "Weight") + ggtitle("Scatterplot of Height vs Weight with Regression Line")
```

(b)

```
```{r}
ggplot(adults, aes(x = Weight, y = Height.)) + geom_point() + geom_smooth(method = "lm", se = FALSE) + labs(x = "Height", y
= "Weight") + ggtitle("Scatterplot of Height vs Weight with Regression Line")

```
```

(c)
```

```r
lm_model <- lm(Weight ~ Height., data = adults)
height_predict <- 180
new_data <- data.frame(Height. = height_predict)
w <- predict(lm_model, newdata = new_data)
w
```

```
       1
92.58016
```

(d)

```r
lm_model2 <- lm(Height. ~ Weight, data = adults)
new_data2 <- data.frame(Weight = w)
predict(lm_model2, newdata = new_data2)
```

```
       1
171.1131
```

(e) The answer to (d) is not 180 even though 92.5 was predicted for a person with a height of 180 cm because we are fitting two different models one is with y = height and the other is y = weight. If we look at the model's fit in the graphs, we see that for the height 180, the weight is around 92 and in the other model for the weight 92.5, height is around 171. That is why in our second model we see that the weight 92.5 instead of getting 180, we get 171. If we want the model to be as close as possible to 180, then we will need more quality to feed the model.