Hao-FinalExam

```
library(dplyr)
library(ggplot2)
setwd("~/Google Drive/CUNY/git/DATA606/FinalExam")
```

Part1

a. Describe the two distributions (2 pts)

Distribution A has a mean of 5.05 and standard deviation of 3.22 as given and is skewed to the right with a minimum value of zero. It may be a log normal distribution.

Distribution B has a mean of 5.04 and standard deviation of 0.58 as given and is closer to a normal distribution.

b. Explain why the means of these two distributions are similar but the standard deviations are not (2 pts)

Distribution A is comprised of the observed variable; whereas Distribution B is comprised only of the *mean* of 500 samples of size 30. Since Distribution B is based on the mean of Distribution A, it is not surprising that the mean of Distribution B is close to the mean of Distribution A. However, the standard deviation of Distribution B is only measuring the variation of the estimated means of Distribution A. In fact, the standard deviation of Distribution B is the standard error of Distribution A, which is the standard deviation of Distribution A divided by the square root of the sample size, as shown in the calculation below:

```
3.22 / sqrt(30)
```

```
## [1] 0.5878889
```

c. What is the statistical principal that describes this phenomenon (2 pts)?

The central limit theorem, which states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, if the sample size is large enough (n > 30 when the underlying distribution is roughly normally distributed or greater if not).

Part2

```
options(digits=2)
data1 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68))
data2 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74))
data3 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73))
data4 <- data.frame(x=c(8,8,8,8,8,8,8,19,8,8,8),
                    y=c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.5,5.56,7.91,6.89))
```

For each column, calculate (to two decimal places):

a. The mean (for x and y separately; 1 pt)

```
df = data.frame(rbind(cbind(set = 'data1', data1),
                      cbind(set = 'data2', data2),
                      cbind(set = 'data3', data3),
                           cbind(set = 'data4', data4)))
df_mean = df %>% group_by(set) %>%
  summarise(mean_x = mean(x), mean_y = mean(y))
print.data.frame(df_mean, digits = 2)
```

```
##     set mean_x mean_y
## 1 data1      9    7.5
## 2 data2      9    7.5
## 3 data3      9    7.5
## 4 data4      9    7.5
```

b. The median (for x and y separately; 1 pt)

```
df_median = df %>% group_by(set) %>%
  summarise(median_x = median(x), median_y = median(y))
print.data.frame(df_median, digits = 2)
```

```
##     set median_x median_y
## 1 data1        9      7.6
## 2 data2        9      8.1
## 3 data3        9      7.1
## 4 data4        8      7.0
```

c. The standard deviation (for x and y separately; 1 pt)

```
df_sd = df %>% group_by(set) %>%
  summarise(sd_x = sd(x), sd_y = sd(y))
print.data.frame(df_sd, digits = 2)
```

```
##     set sd_x sd_y
## 1 data1  3.3    2
## 2 data2  3.3    2
## 3 data3  3.3    2
## 4 data4  3.3    2
```

For each x and y pair, calculate (also to two decimal places; 1 pt):

d. The correlation (1 pt)

```
df_cor = df %>% group_by(set) %>%
  summarise(cor = cor(x, y))
print.data.frame(df_cor, digits = 2)
```

```
##     set  cor
## 1 data1 0.82
## 2 data2 0.82
## 3 data3 0.82
## 4 data4 0.82
```

e. Linear regression equation (2 pts)

The regression equation for each data set is: yˆ = intercept + slope * x

```
df_lm = df %>% group_by(set) %>%
  do(model = lm(y ~ x, data = .)) %>%
  mutate(intercept = summary(model)$coeff[1],
         slope = summary(model)$coeff[2]) %>%
  select(-model)
print.data.frame(df_lm, digits = 2)
```

```
##     set intercept slope
## 1 data1         3   0.5
## 2 data2         3   0.5
## 3 data3         3   0.5
## 4 data4         3   0.5
```

f. R-Squared (2 pts)

```
df_rsq = df %>% group_by(set) %>%
  do(model = lm(y ~ x, data = .)) %>%
  mutate(rsq = summary(model)$r.squared) %>%
  select(-model)
print.data.frame(df_rsq, digits = 2)
```

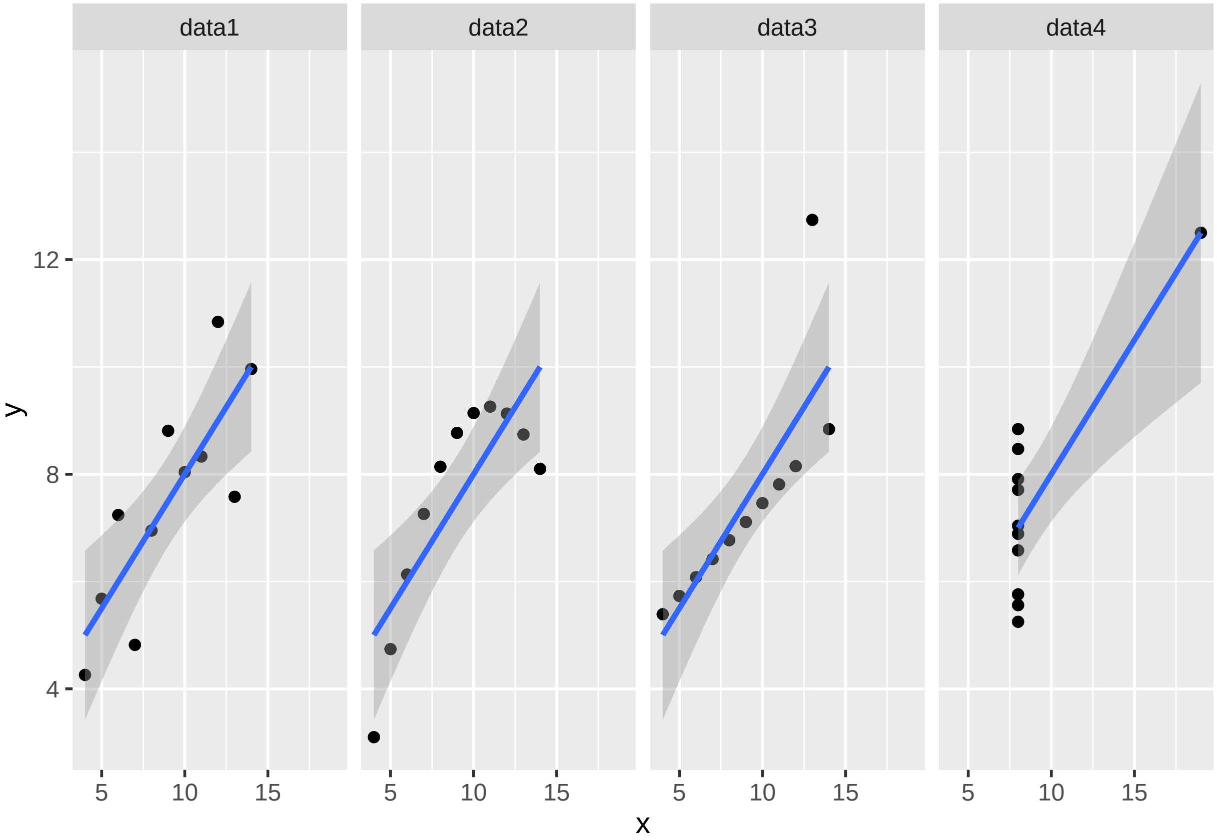
```
##     set  rsq
## 1 data1 0.67
## 2 data2 0.67
## 3 data3 0.67
## 4 data4 0.67
```

For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts)

Based on the charts below:

- Data1: Yes, the data points are relatively linearly arranged
- Data2: No, the data clearly show a non-linear relationship
- Data3: Maybe, while the data exhibit a strong linear relationship, there is one outlier with severe leverage that significantly steepens the regression line. The analyst would need to determine if the outlier is legitimate or not. If not and it can be safely removed, then linear regression is appropriate. If the outlier is legitimate, then whether linear regression is appropriate may depend on the specific question being answered
- Data4: No, the data cannot be described by a function and thus linear regression could not be applied

```
df %>% ggplot(aes(x = x, y = y)) +
  geom_point() +
  facet_grid(. ~ set) +
  stat_smooth(method = 'lm')
```



Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create. (2 pts)

As made clear in this exam, sets of data can have similar or even identical summary statistics, e.g. mean, median, sd, cor, and even produce the same linear regression models, but can be drastically different from one another. However, the simple plots above clearly illustrate the dramatically different nature of the four data sets in question.

In addition, visualization is very helpful for data exploration and help identify patterns (e.g. seasonality), outliers (e.g. faulty data, exceptional data), and other important information that might be be captured in summary statistics or even while looking at the raw tabular data.

Furthermore, visualization is also very helpful for model diagnostics, e.g. to ensure that the assumptions and criteria of a model are being met. For example, with regression, an analyst would evaluate 1) qq-plots to check if errors are normally distributed, 2) plots of errors vs. observations to ensure errors are randomly distributed and no identifiable structure/pattern remains in the errors, 3) plots of errors vs. order of observation to ensure issues like heteroskedasticity are not present.