DAV5300 Fall 2024 - Final Exam

Name: Tushar Ahuja

Part I

Please put the answers for Part I next to the question number:

1. B
2. A
3. D
4. B
5. B
6. E
7. D
8. E
9. B
10. C

Part II

a. Describe the two distributions (2 pts).

Answers:-

- **Graph A (Observations):**

1. Right-skewed distribution
2. Spans from approximately 0 to 20
3. Peak around 5
4. Shows the raw data distribution with larger spread

- **Graph B (Sampling Distribution):**

1. Approximately normal/bell-shaped distribution
2. Much narrower range (approximately 3.5 to 6.5)
3. More symmetric than Graph A
4. Represents distribution of sample means

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

Answers:-

- The means are similar because the sampling distribution (Graph B) represents the means of samples from the original distribution (Graph A), and by the unbiased sampling property, the mean of sample means equals the population mean.
- The standard deviations are different because Graph B shows the sampling distribution of means, which has a smaller standard deviation than the original distribution by a factor of $1/\sqrt{n}$ (where n=30). This is why Graph B appears more compressed horizontally than Graph A.

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

Answers:-

- The Central Limit Theorem (CLT) is the fundamental statistical principle that explains the phenomenon shown in these graphs. Here's a comprehensive explanation of these two graphs:

Core Principle: The Central Limit Theorem states that regardless of the shape of the original population distribution, the sampling distribution of the mean will be approximately normally distributed for sufficiently large sample sizes (typically n 30). This explains why Graph B shows a normal distribution despite Graph A being right-skewed.

Mathematical Properties: The Central Limit Theorem specifies exact relationships between the population and sampling distribution parameters:

- The mean of the sampling distribution ($\bar{x}$) equals the population mean ( )
- The standard error ($\bar{x}$) equals the population standard deviation divided by $\sqrt{n}$: $\bar{x} = /\sqrt{n}$
- In this case, with n=30, the standard deviation of Graph B (0.58) is approximately equal to the standard deviation of Graph A (3.22) divided by $\sqrt{30}$.

Practical Implications and Applications:

The Central Limit Theroem (CLT) has crucial implications for statistical inference:

- It allows us to use normal probability calculations even when the underlying population is not normal
- It provides the theoretical foundation for constructing confidence intervals and conducting hypothesis tests about population means
- It explains why the sampling distribution (Graph B) becomes more "well-behaved" and predictable than the original population distribution (Graph A)
- This theorem is what enables us to make reliable statistical inferences about population parameters from sample statistics

This fundamental principle underlies most parametric statistical procedures and explains both the shape transformation we observe (from skewed to normal) and the reduction in variability shown in the graphs.

Part III

Consider the three datasets, each with two columns (x and y), provided below.

```
## $x_mean
## [1] 54.26
##
## $y_mean
## [1] 47.83
##
## $x_median
## [1] 53.33
##
## $y_median
## [1] 46.03
##
## $x_sd
## [1] 16.77
##
## $y_sd
## [1] 26.94
##
## $correlation
## [1] -0.06447
##
## $slope
##       x
## -0.1036
##
## $intercept
## (Intercept)
##       53.45
##
## $rsquared
## [1] 0.004157
```

```
## $x_mean
## [1] 54.27
##
## $y_mean
## [1] 47.84
##
## $x_median
## [1] 53.14
##
## $y_median
## [1] 46.4
##
## $x_sd
## [1] 16.77
##
## $y_sd
## [1] 26.94
```

```
##
## $correlation
## [1] -0.06898
##
## $slope
##       x
## -0.1108
##
## $intercept
## (Intercept)
##       53.85
##
## $rsquared
## [1] 0.004758
```

```
## $x_mean
## [1] 54.27
##
## $y_mean
## [1] 47.83
##
## $x_median
## [1] 53.34
##
## $y_median
## [1] 47.54
##
## $x_sd
## [1] 16.77
##
## $y_sd
## [1] 26.94
##
## $correlation
## [1] -0.06413
##
## $slope
##      x
## -0.103
##
## $intercept
## (Intercept)
##       53.43
##
## $rsquared
## [1] 0.004112
```

Fill in the cells of the table below (to three decimal places).

	Data 1		Data 2		Data 3	
	x	y	x	y	x	y
Mean	54.260	47.830	54.270	47.840	54.270	47.830
Median	53.330	46.030	53.140	46.400	53.340	47.540
SD	16.770	26.940	16.770	26.940	16.770	26.940
r	-0.064		-0.068		-0.064	
Intercept	53.450		53.850		53.430	
Slope	-0.103		-0.110		-0.103	
R-Squared	0.004		0.005		0.004	

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!

Data set 1 Yes or No

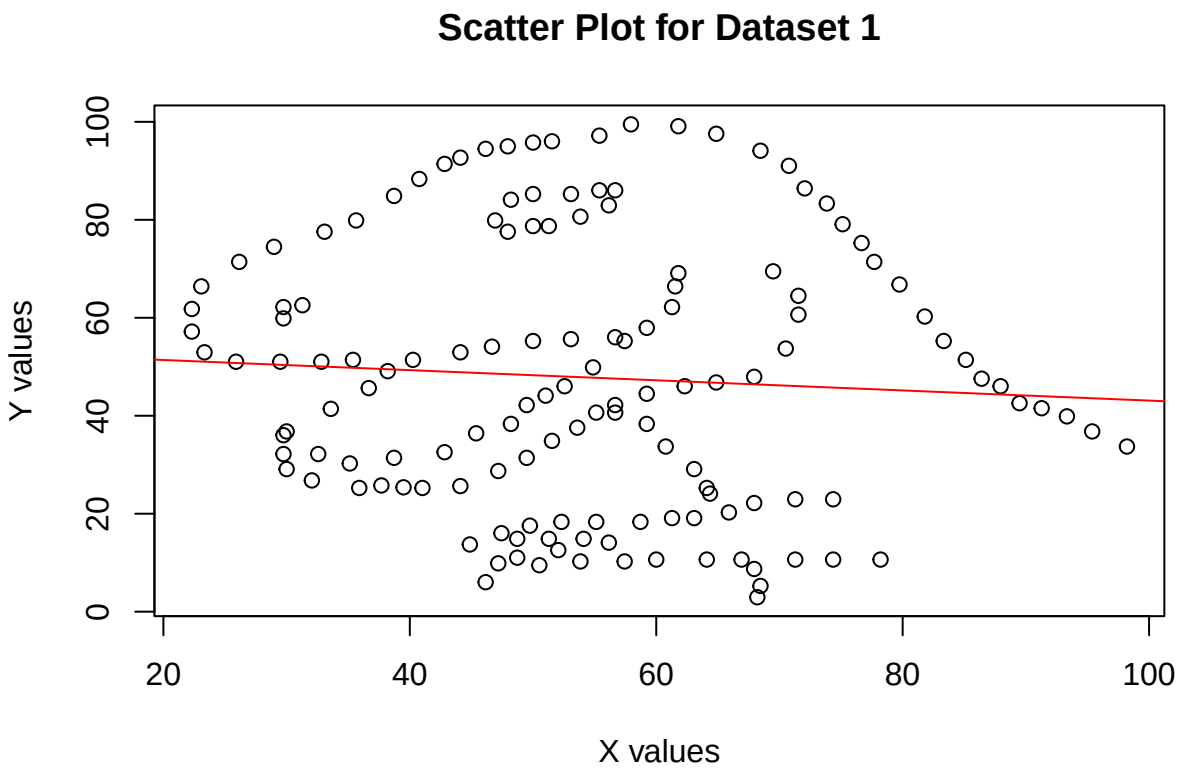
- No

Why?

1. The correlation between the variables x and y in Dataset 1 is -0.064, which is very close to 0, indicating a very weak or no linear relationship between the two variables. A strong linear relationship is needed for a meaningful linear regression model.
2. Additionally, the R-squared value of 0.004157 is extremely low, suggesting that the model only explains 0.4157% of the variance in the data, which is not sufficient for reliable predictions.

Therefore, given these metrics linear regression might not provide a good fit.

```
# Plot to visualize the weak relationship
plot(data1$x, data1$y, main="Scatter Plot for Dataset 1", xlab="X values", ylab="Y values")
abline(lm(data1$y ~ data1$x), col="red") # Adding regression line
```



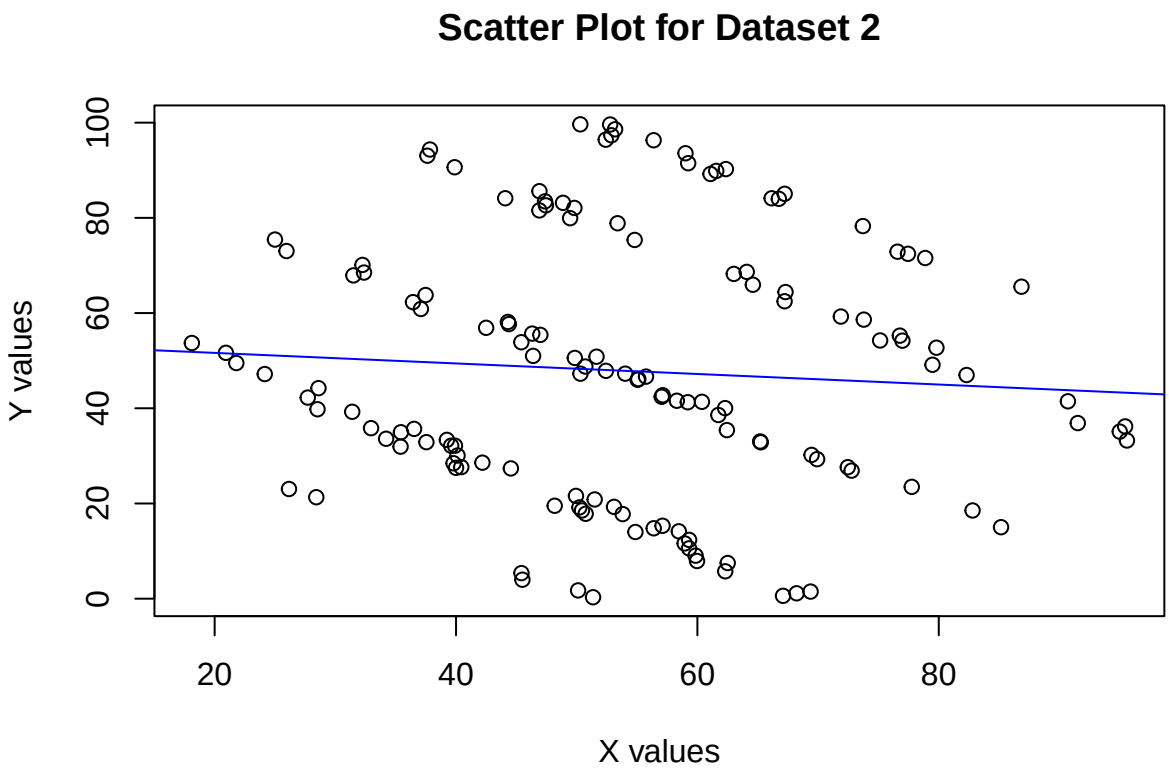
Data set 2 Yes or No

- No

Why?

1. The correlation between x and y in Dataset 2 is -0.068, which is also very weak. A correlation close to zero indicates no meaningful linear relationship.
2. The R-squared value of 0.004758 is similarly very low, meaning the model does not explain much of the variance in the data. Therefore, given these statistics the linear regression model is unlikely to provide valuable insights.

```
# Plot to visualize the weak relationship
plot(data2$x, data2$y, main="Scatter Plot for Dataset 2", xlab="X values", ylab="Y values")
abline(lm(data2$y ~ data2$x), col="blue") # Adding regression line
```



Data set 3 Yes or No

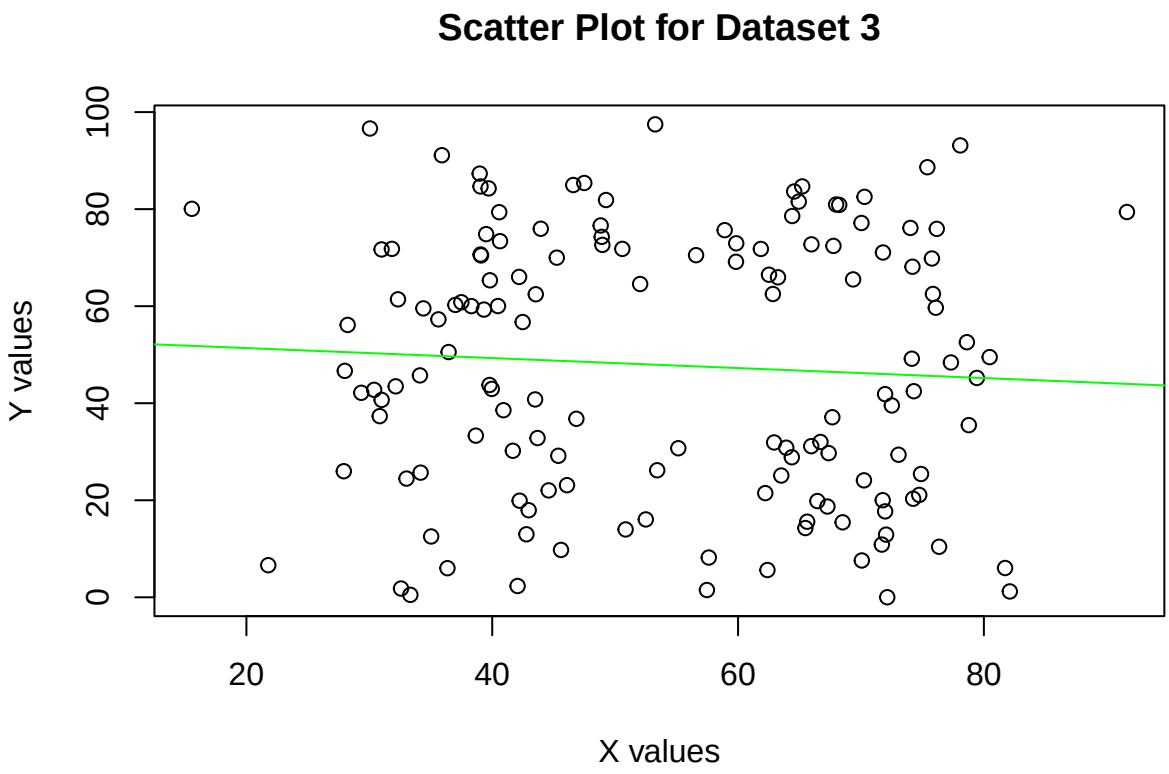
- No

Why?

1. The correlation for Dataset 3 is -0.064, which again indicates a very weak or no linear relationship between the two variables.
2. The R-squared value of 0.004112 is very low, indicating that the regression model doesn't capture much of the variation in y.

Similar to the previous two datasets, these results show that the linear regression model would not be appropriate for making predictions.

```
# Plot to visualize the weak relationship
plot(data3$x, data3$y, main="Scatter Plot for Dataset 3", xlab="X values", ylab="Y values")
abline(lm(data3$y ~ data3$x), col="green") # Adding regression line
```



In all three datasets, the correlation values are very weak (close to 0), and the R-squared values are extremely low, suggesting that a linear regression model is not appropriate for any of these datasets. The linear regression model would not be useful for prediction or understanding the relationship between x and y.

Why it is important to include appropriate visualizations when analyzing data? Be sure to ground your reasoning in the context of the analyses completed above. Include any visualization(s) you create.

Answer:-

- Visualizations are a crucial component of data analysis for several reasons, especially in the context of the analyses we completed above for the three datasets. Here's why they are important:

1. Understanding Relationships Between Variables

Visualizations, such as scatter plots, help us easily identify relationships between variables. In the context of the analyses, by plotting x against y for each dataset, we can visually observe how strongly or weakly the variables are related. Even though the correlation values were very weak (close to 0) in the datasets, visualizing the data can confirm this. For example:

Dataset 1: The scatter plot reveals that the data points are widely spread, and no clear trend (linear or otherwise) can be observed. This aligns with the low correlation of -0.064 and the low R-squared value, suggesting no meaningful linear relationship.

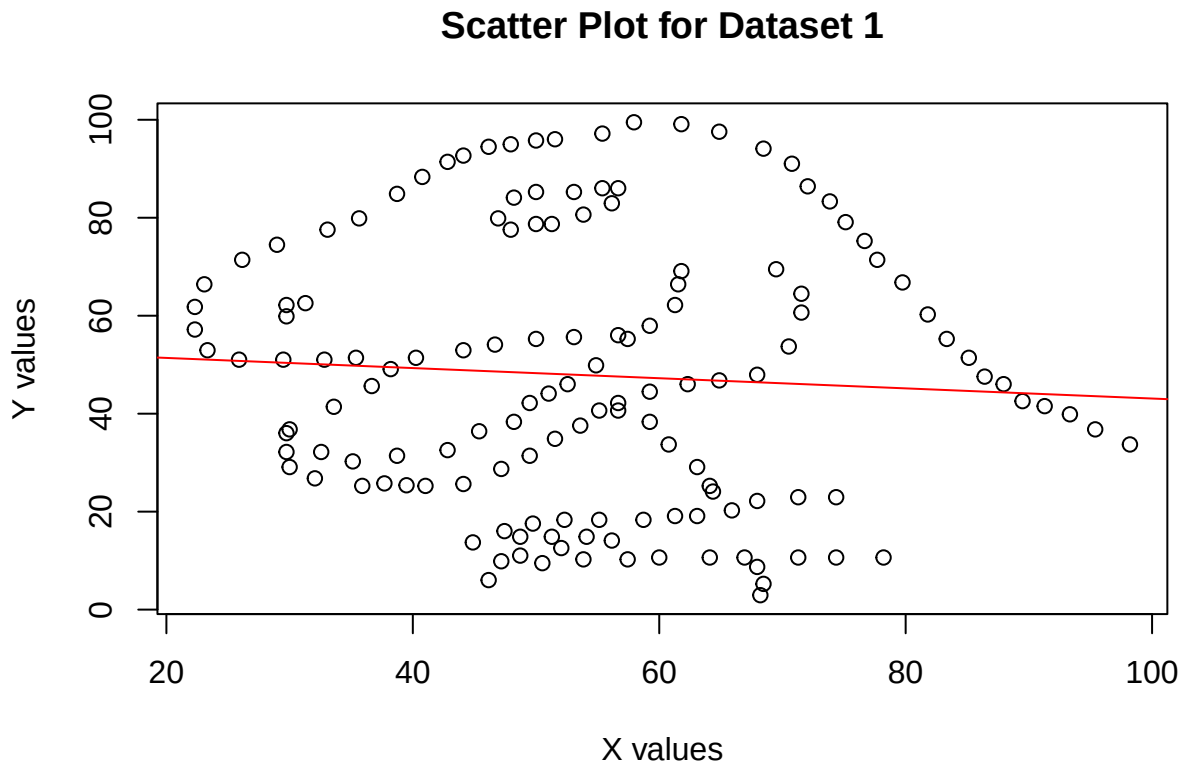
Dataset 2: Similarly, the scatter plot shows random scatter of points with no obvious trend, which further confirms that a linear regression model would not be effective.

Dataset 3: The weak relationship is visually evident, helping us reinforce that linear regression is not appropriate.

Without such plots, it would be harder to grasp the nature of the relationship between the variables and whether applying linear regression is sensible.

Example for Dataset 1

```
# Scatter Plot for Dataset 1
plot(data1$x, data1$y, main="Scatter Plot for Dataset 1", xlab="X values", ylab="Y values")
abline(lm(data1$y ~ data1$x), col="red") # Adding regression line
```



2. Highlighting Outliers or Patterns

Visualizations can help detect outliers, trends, or non-linear patterns that might not be obvious from summary statistics alone. In our case, even though the correlation is low, the scatter plots could reveal clusters or patterns in the data that suggest further exploration. This could indicate that another type of model, such as a polynomial regression or decision tree, might be more appropriate.

3. Interpreting Model Fit

The regression line added to the scatter plot provides a visual representation of how well the linear model fits the data. In the case of all three datasets, we can see that the regression lines (in red, blue, and green) are not fitting the data well, which visually reinforces the conclusion that the linear regression model does not adequately describe the relationship between x and y.

Even though the regression lines appear, the visual mismatch between the data points and the line (in all datasets) confirms the poor fit indicated by the very low R-squared values.

4. Facilitating Communication of Results

Visualizations make it easier to communicate your findings to both technical and non-technical stakeholders. Rather than relying solely on statistical terms (like correlation and R-squared), visualizations provide intuitive insights. For example, a non-expert can quickly see from a scatter plot whether a linear trend is visible, even without understanding the exact correlation value.

5. Supporting Further Exploration

Visualizing data can reveal areas that might need further analysis. If the visual plots showed any patterns (e.g., clusters, non-linearity), we could dive deeper into the analysis and explore alternative models. In this case, visualizing the weak correlations and lack of clear trends can guide us away from using linear regression and towards other statistical methods or transformations.

Summary Visualizations are critical because they help us:

- Understand relationships between variables, confirming statistical results like correlation.
- Detect outliers or patterns that may suggest the need for other models.
- Interpret model fit and whether a regression model is appropriate.
- Communicate findings effectively to stakeholders.
- Support further analysis by guiding our focus on potential areas of interest.

In this analysis, scatter plots were particularly useful for demonstrating the weak relationships between x and y, which reinforced the conclusion that linear regression would not be suitable for any of the datasets.