Assignment-based Subjective Questions

1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Here are the insights:

- Weather Situation: Demand decreases significantly in bad weather
- Season: Spring season is negatively correlated with demand
- Season: Demand increases in winter season and clear weather
- Weekday: Demand is low on tuesdays

2: Why is it important to use drop_first=True during dummy variable creation?
By dropping one dummy variable column, we reduce multicollinearity as it can be derived from other columns.

3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Temperature is highest correlated with target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Linearity: Linearity was validated by looking at the scatter plot of feature and target variable
Homoscedasticity: It was verified by plotting residual terms against one of feature variables in a scatter plot. It showed no visible patterns and data was uniformly scattered around y = 0 line
Error terms normally distributed: This was verified by plotting histogram of error terms
Multicollinearity: This was verified by plotting correlation heatmap of feature variables

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

1. Atemp: Feeling temperature. Positively correlated. Coefficient: 0.46
2. Year: Demand increasing year on year. Coef: .24
3. Bad Weather: Negatively correlated. Termed as snow in feature variables. coeff: -.21

General Subjective Questions

1: Explain the linear regression algorithm in detail.
Linear Regression is a supervised learning method which computes the linear relationship between dependent variable and independent variable (one or multiple) by training on a precomputed dataset.

Following are the steps to perform linear regression.

- **Data Preparation**: Prepare the data which involves data cleaning, conversion of categorical columns into numerical ones etc

2: Explain the Anscombe's quartet in detail.
"Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed."
-Wikipedia
Anscombe's quartet comprises 4 datasets which contain identical summary statistics but look very different when plotted on a graph.

Below are 4 data sets

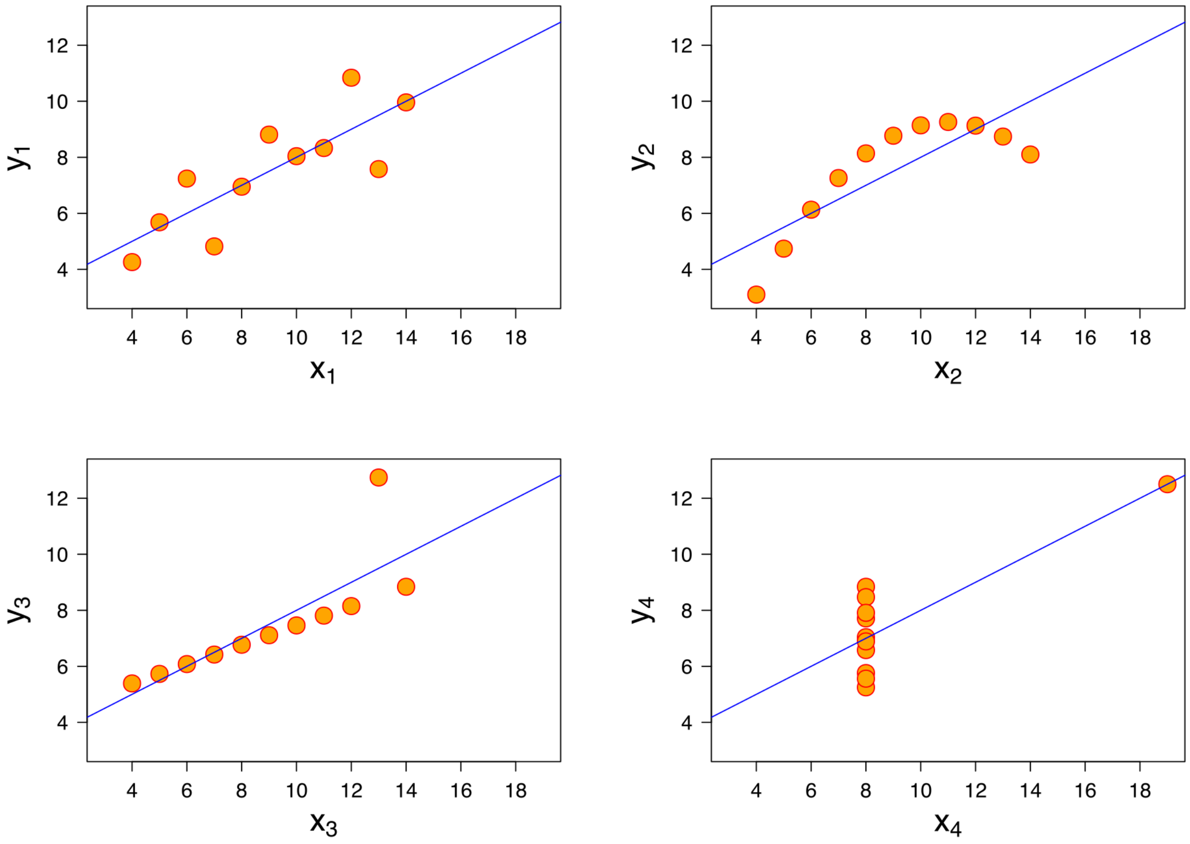
Data set I		Data set II		Data set III		Data set IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56

7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Here is the summary statistics of all 4 datasets

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*: *s*2 *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*: *s*2 *y*	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R2	0.67	to 2 decimal places

Here is how they look plotted on graph



Dataset 1 follows simple linear relationship
Dataset 2 follows a relationship but not linear
Dataset 3 contains an outlier
Dataset 4 does not follow any relationship between x and y but their correlation coefficient is very high but to one outlier

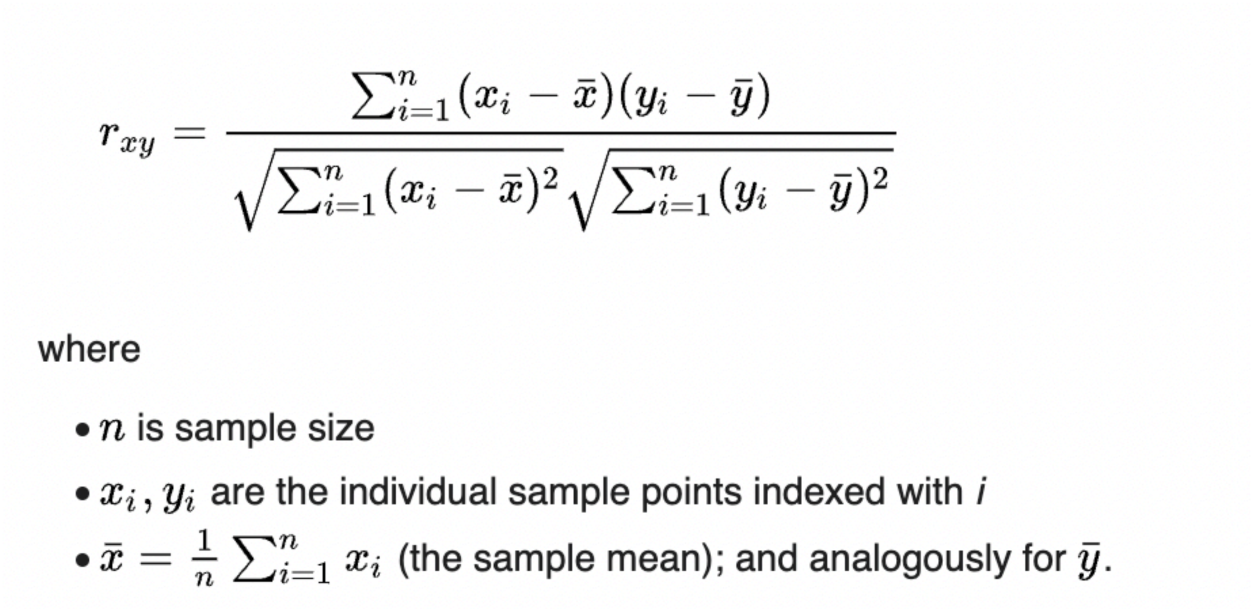
Above example shows why it is so important to study graphs of a dataset and why mere summary statistics does not give us a clear picture of it.

3. What is Pearson's R?
Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson, a prominent statistician.
Pearson's r takes on values between -1 and 1, where:

- 1 indicates a perfect positive linear relationship: As one variable increases, the other variable also increases proportionally.
- -1 indicates a perfect negative linear relationship: As one variable increases, the other variable decreases proportionally.

- 0 indicates no linear relationship: There is no systematic relationship between the variables.

Formula for any sample:



$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling refers to the process of transforming the values of variables into a specific range or distribution. It is performed to ensure that all variables have a comparable scale, which can be crucial for certain machine learning algorithms to perform effectively.

Advantages of scaling:

- Scaling can make the coefficients or feature importances in linear models more interpretable by making them directly comparable.
- Gradient-based optimization algorithms (e.g., gradient descent) converge faster when the features are on a similar scale.

Normalized Scaling:

Formula:

x norm = x−mean(x)/max(x)−min(x)

Normalized scaling centers the variables around 0 and scale them to range between -1 and 1

It preserves the shape of the original distribution

It is sensitive to outliers as the range is determined by the minimum and maximum values.

Standardized scaling:

Formula:

$$\boldsymbol{x}_{\textbf{std}} = \boldsymbol{x}\textbf{−mean(}\boldsymbol{x}\textbf{)/std(}\boldsymbol{x}\textbf{)}$$

It centers the variable around zero and scales it to have a standard deviation of 1.
May not preserve the shape of original distribution if it does not follow normal
Less sensitive to outliers compared to normalized scaling, as the range is based on the variability of the data.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Formula:

$$\mathbf{VIF = 1/1-Rj^2}$$

R_{j2} is the coefficient of determination (R-squared) obtained by regressing the predictor variable X_j on all other predictor variables in the model.

VIF becomes infinite when Rj square becomes 1 or when Xj has perfect collinearity with other predictor variables.

Perfect collinearity can occur in many cases:

- Perfect Fit: Xj follows perfectly linear relation with other variables
- Duplicate variables: There are Xj’s duplicate predictor variables present
- Linear relation: When one predictor variable is a linear combination of other predictor variables in the model. For example: Casual and registered variables in bike sharing assignment are linearly related with cnt variable

  Cnt = Casual + registered

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a set of data follows a particular probability distribution. It compares the quantiles of the data to the quantiles of a theoretical distribution, typically a normal distribution, but it can be used with other distributions as well.

- In linear regression analysis, one of the assumptions is that the residuals (the differences between observed and predicted values) are normally distributed.
- Q-Q plots are commonly used to visually assess whether this assumption holds. By comparing the quantiles of the residuals to the quantiles of the normal distribution, you can determine if the residuals follow a normal distribution pattern.
- A well-fitting linear regression model typically exhibits residuals that closely follow a straight line on the Q-Q plot, indicating that the residuals are normally distributed.