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? (3 marks)

- **Positive year to year growth: Bike rentals increased in 2019 compared to 2018**
- **Month wise, highest recorded in September month**
- **Season wise, fall recorded highest bike-rentals followed by summer.**
- **Clearly lower bike-rentals on a Holiday**
- **No significant difference in bike-rentals based on weekday.**
- **More bike-rentals recorded during Sunny days (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist)**

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

- **For a categorical variable with n values, we just need n-1 dummy variables as n^{th} dummy variable can be explained using remaining n-1 dummy variables. This will also reduce multicollinearity.**
- **For example, gender column has values 'Male', 'Female' and 'Other' can be explained with just 2 dummy variables instead of 3.**

gender_Male	gender_Female	Derived Value
1	0	Male
0	1	Female
0	0	Other

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

- **Temperature variables (temp & atemp) have the highest correlation with the target variable 'cnt'**

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- Validated **VIF** to confirm **No Multicollinearity** between final independent variables.
- Validated **non-linearity** in predicted values using scatter plot**.**

- Validated **Homoscedasticity** by visualizing funnel shape in distribution plot for error terms.
- Validated **Normality** of errors/residuals by using Q-Q plot
- Pair plots are used to see linear relationship between dependent and independent variables**.**

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

a) **Temperature** with coeff. **0.43**
b) **Year** with coeff. **0.23**
c) **Rainy weather** with coeff. **-0.23**

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fitting line to represent this relationship.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as temp, salary, humidity, windspeed, etc.

The linear regression model is expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p * x_p + \varepsilon$$

y is the dependent variable (or Target variable)

$x_1, x_2, \ldots, x_p$ are the independent variables (or predictors)

$\beta_0, \beta_1, \ldots, \beta_p$ are the coefficients (or weights) assigned to respective independent variable

ε is the error term, representing the difference between the predicted value and the actual value

The **goal of linear regression is to find the coefficients** ($\beta_0, \beta_1, \ldots, \beta_p$) that minimize the sum of squared residuals (the difference between the predicted and actual values). This is often achieved using the **OLS** (least squares method)

Linear regression makes several assumptions:

1. **Linearity:** The relationship between the dependent and independent variables is linear.
2. **Independence/Non-Linear:** The observations are independent of each other.
3. **Homoscedasticity:** The variance of the errors/residuals is constant.
4. **Normality:** The errors/residuals are normally distributed.
5. **No multicollinearity:** Independent variables are not perfectly correlated with each other.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a famous set of four datasets that highlight the importance of visualizing data before drawing conclusions solely based on statistical measures like correlation. Despite having identical summary statistics (mean, variance, correlation), these datasets exhibit vastly different patterns when visualized.

The four datasets share the same:

- Mean of x: 9
- Mean of y: 7.5
- Variance of x: 10
- Variance of y: 10
- Correlation coefficient: 0.816

However, when visualized, they reveal distinct patterns:

1. **Dataset 1:** A linear relationship with a positive slope.
2. **Dataset 2:** A quadratic relationship with a positive slope.
3. **Dataset 3:** A perfect linear relationship with one outlier.
4. **Dataset 4:** A horizontal line with no relationship between x and y.

I		II		III		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

3. What is Pearson's R? (3 marks)

Pearson's correlation coefficient (r) is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to 1:

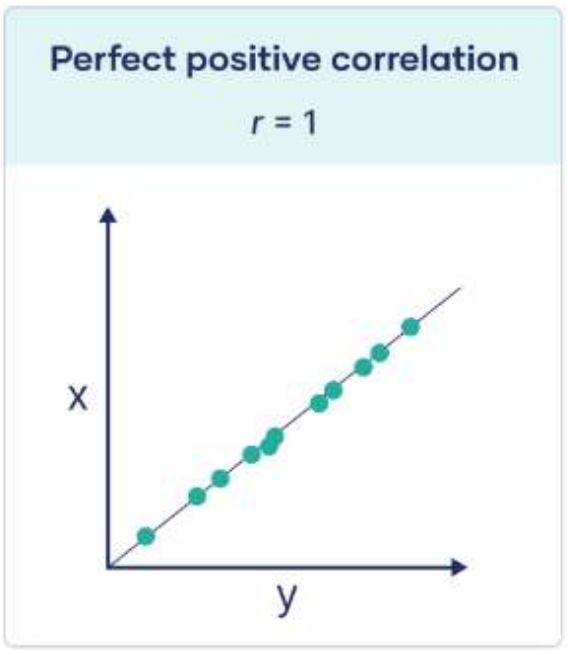
- **r = 1:** Perfect positive correlation, meaning the variables increase or decrease together perfectly.
- **r = -1:** Perfect negative correlation, meaning one variable increases as the other decreases perfectly.
- **r = 0:** No correlation between the variables.

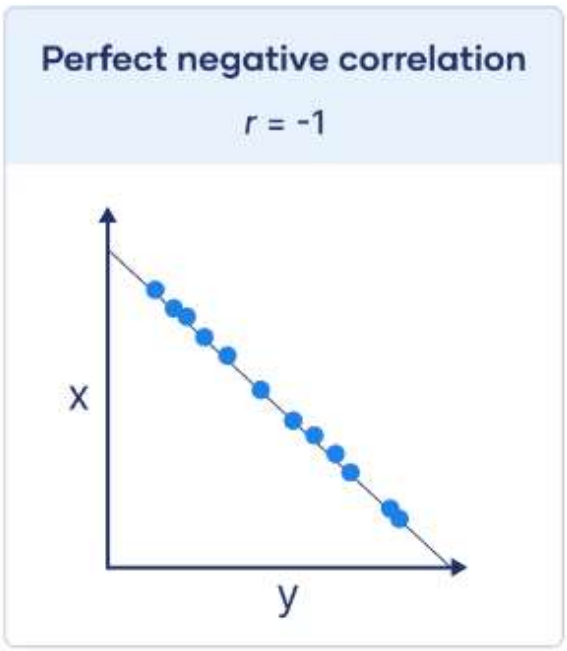
The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

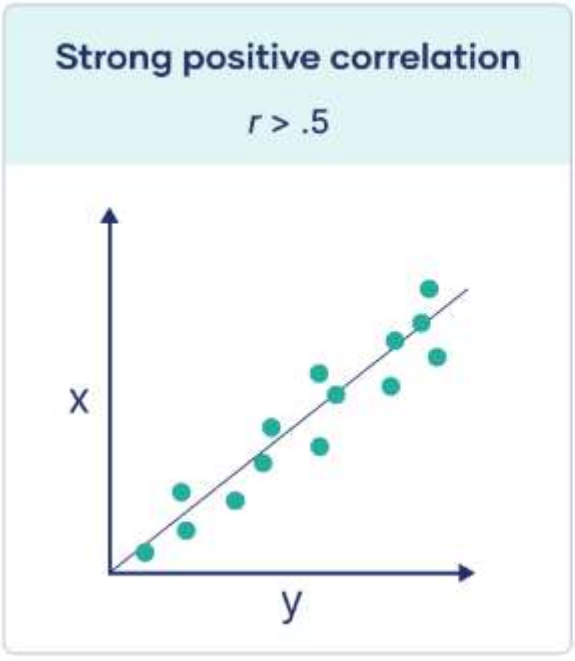
The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, *r* is negative. When the slope is positive, *r* is positive.

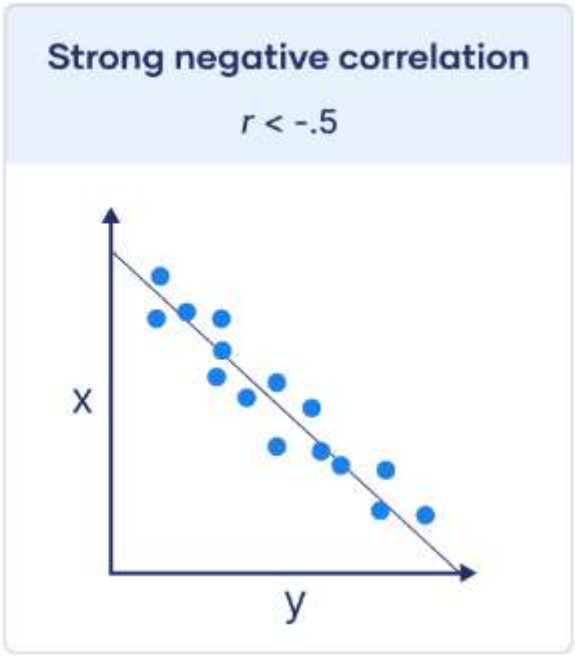
When *r* is 1 or –1, all the points fall exactly on the line of best fit:



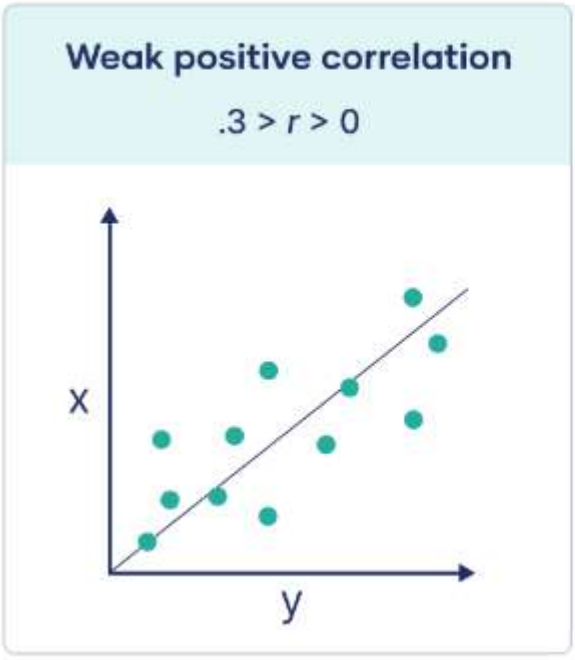


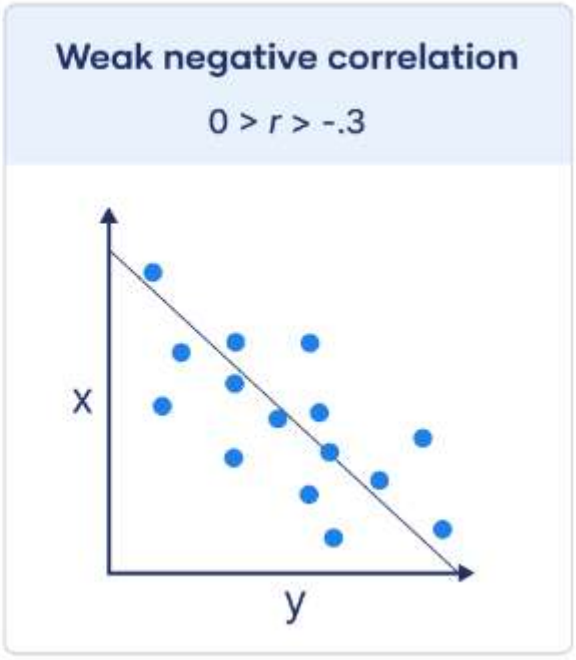
When *r* is greater than .5 or less than –.5, the points are close to the line of best fit:



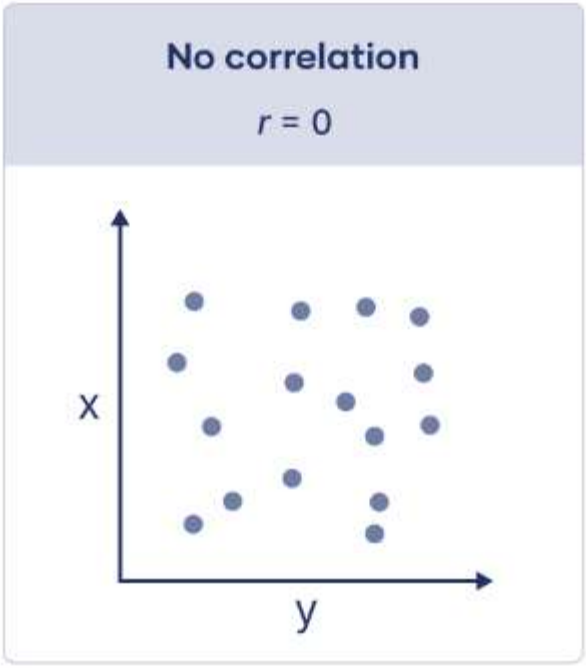


When *r* is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:





When *r* is 0, a line of best fit is not helpful in describing the relationship between the variables:



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a technique used in data preprocessing to transform numerical data to a common scale. This is often necessary for algorithms that are sensitive to the magnitude of features, such as machine learning models like linear regression

Benefits of Scaling:

- **Improved Algorithm Performance**: Many machine learning algorithms assume that features are on a similar scale. Scaling can help prevent features with larger magnitudes from dominating the learning process, leading to more accurate models.
- **Faster Convergence**: Scaling can accelerate the convergence of gradient-based optimization algorithms.
- **Regularization**: Some regularization techniques, like L1 and L2 regularization, benefit from scaled features.

Normalized Scaling (Min-Max Scaling)

- **Formula:** (x - min(x)) / (max(x) - min(x))
- **Range:** Scales data to the range of [0, 1].
- **Use Case:** Suitable when we want to preserve the relative differences between values and when the range is known and meaningful.

Standardized Scaling (Z-Score Standardization)

- **Formula:** (x - mean(x)) / std(x)
- **Range:** Transforms data to have a mean of 0 and a standard deviation of 1.
- **Use Case:** Commonly used when we want to center the data around zero and make it unit-less. It's useful when the data distribution is approximately normal or when we don't have prior knowledge about the range of values.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

When VIF becomes infinite, it indicates a perfect linear relationship between the predictor variable and the other predictors.

Formula for VIF = $1/(1-R^2)$

If the R^2 value comes out to be 1, the VIF will become infinite (as denominator becomes 0). This is quite possible when one of the independent variables is strongly correlated with many of the other independent variables.

VIF infinite represents either Perfect Multicollinearity or Near-Perfect Multicollinearity.

Consequences of Infinite VIF if not handled

- **Model Instability:** The regression coefficients can become highly unstable, fluctuating wildly with small changes in the data.
- **Inaccurate Inference:** It's difficult to draw meaningful conclusions about the individual effects of the predictor variables due to the high correlation.
- **Computational Issues:** Some algorithms might fail to converge or produce unreliable results.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A **Q-Q plot** (Quantile-Quantile plot) is a graphical technique used to compare the distribution of a sample data set against a theoretical distribution. It's particularly useful for assessing whether a data set follows a normal distribution.

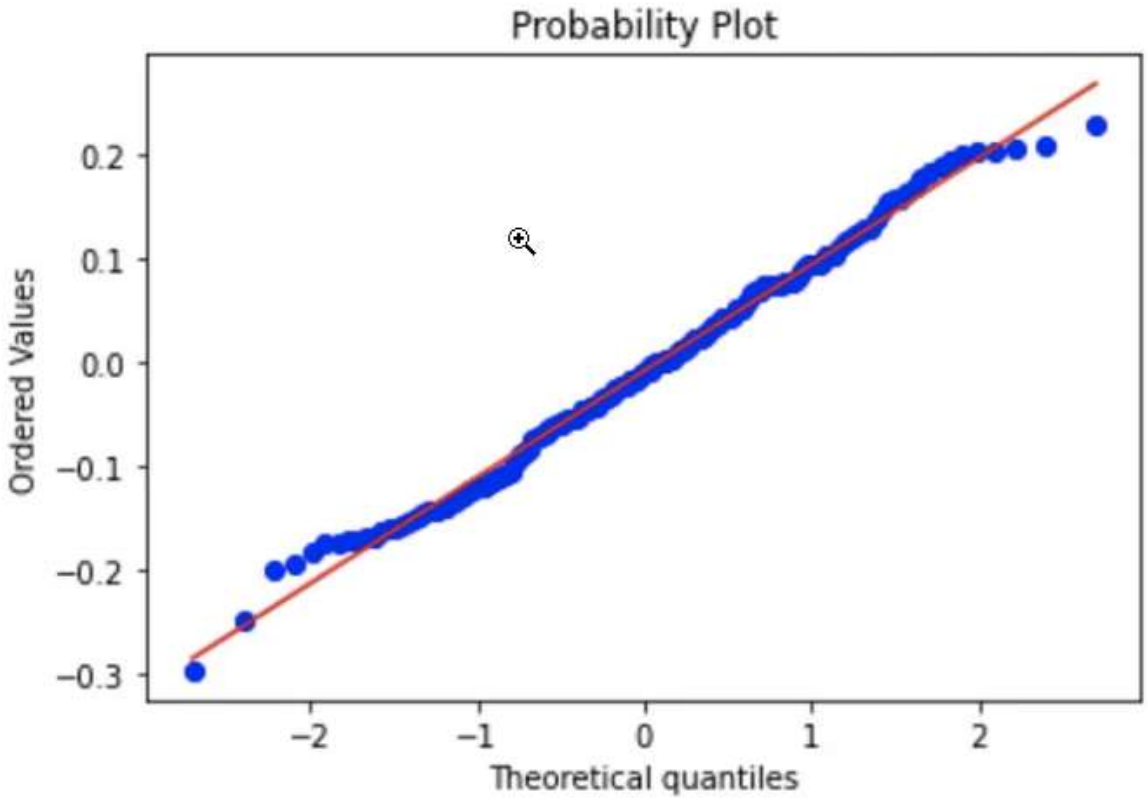
In linear regression, the assumption of normality is crucial for the validity of statistical inferences. Q-Q plots are used to assess whether the residuals (the differences between the actual and predicted values) follow a normal distribution.

Advantages of Q-Q plot

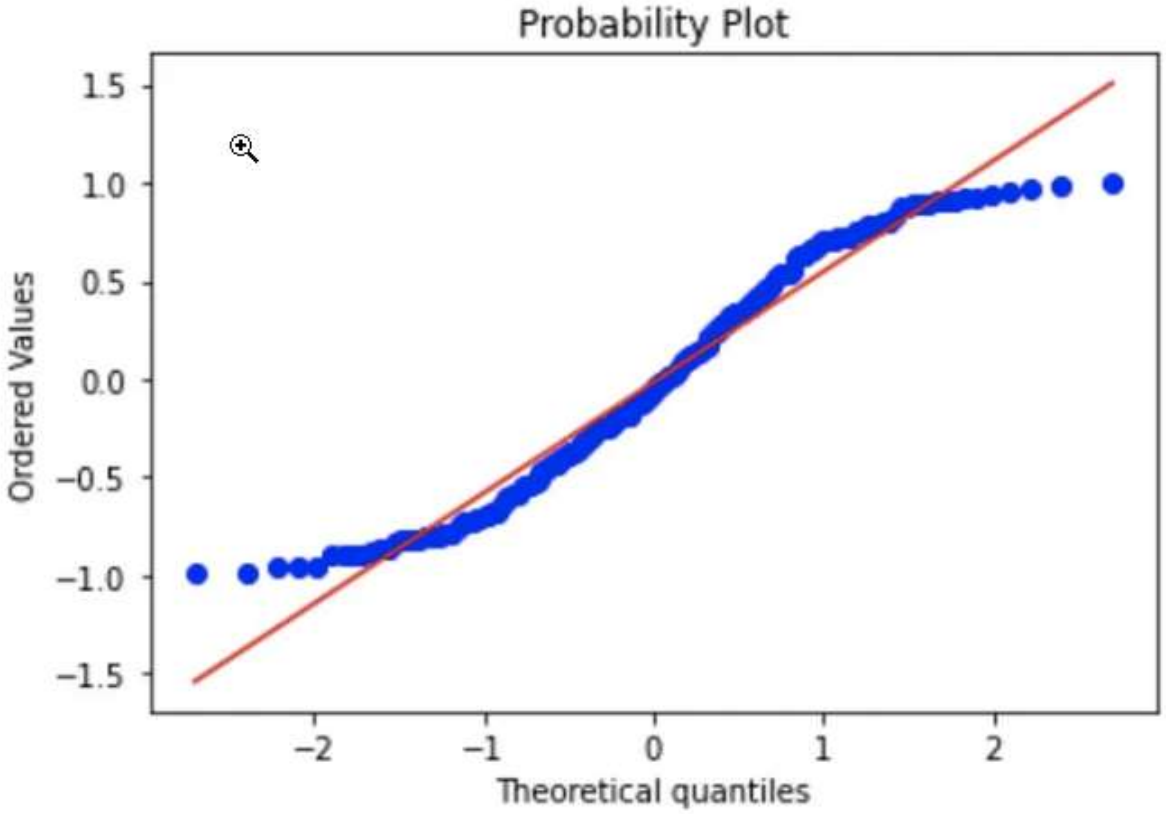
1. **Flexible Comparison**: Q-Q plots can compare datasets of different sizes without **requiring equal sample sizes.**
2. **Dimensionless Analysis**: They are dimensionless, making them suitable for comparing datasets with **different units or scales.**
3. **Visual Interpretation**: Provides a clear visual representation of data distribution compared to a theoretical distribution.
4. **Sensitive to Deviations**: Easily detects departures from assumed distributions, aiding in identifying data discrepancies.
5. **Diagnostic Tool**: Helps in assessing distributional assumptions, identifying outliers, and understanding data patterns.

Interpreting Q-Q plot

- **Straight Line:** If the points fall close to a straight line, it suggests that the residuals are normally distributed.
- **Deviations:** If the points deviate from the line, it indicates departures from normality. For example, a "S" shape might suggest skewness, while a "U" shape might suggest kurtosis/curved.



Above Q-Q plot confirms data points lie approximately in a straight line, this confirms data point normally distributed



Above Q-Q plot confirms most of the points do not lie in a straight line. Showing that the underlying distribution is not normal.