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)**

   Below is my analysis from the categorical variables present in the dataset:

   - Season 3 i.e. "fall" has highest demand for the bikes.
   - On holidays demand has decreased
   - When the weather is clear, bikes demand rises.
   - Demand of bike sharing has increased over the years (2019>2018)
   - During weekend there is small hike in bike demand as compared to other weekdays
   - Bike demand is continuously growing each month till June. September has highest demand. After September demand decreases

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

   Using *drop_first=True* when creating dummy variables is important to avoid **multicollinearity**.

   When we create dummy variables for a categorical variable with n categories by using *drop_first=True* , n-1 dummy variables are created. It is because the information contained in the nth dummy variable is redundant, as it can be inferred from the values of the other n-1 dummy variables.

   If we do not use *drop_first=True*, then all n dummy variables will be included in the model, which will lead to multicollinearity. It can cause problems with the interpretation of the model coefficients and can also make the model more difficult to fit.

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

   Temperature i.e. ‘temp’ variable has the highest correlation with the target variable.

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

Assumptions of Linear Regression after building the model on the training set has been validated using below points:

a. Linear Relationship exist between X and Y:
    i. Linear relationships is visible between predictive variable and target variables
b. Residual Analysis:
    i. Normal Distribution: Error Terms follows normal distribution.
    ii. No Visible pattern found in residual plots.
    iii. Error term has constant variance.
c. Multicollinearity Analysis:
    i. Final Built model has very low multicollinearity between the predictive 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)

Top 3 feature variable contributing significantly towards explaining the demand of the shared bikes are:

a. *Temperature:* Coeff "0.403895" indicates increase in temperature will increase in bike demand
b. *Weather (thunderstorm or situation 3):* Coeff "-0.249302" indicates increase in situation 3 will decrease bike demand.
c. *Year:* Coeff "0.242173" indicates increase in year will increase in bike demand

General Subjective Questions

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

Definition: A linear regression model is a statistical model that describes the relationship between a dependent variable and one or more independent variables using a linear equation.

Equation: The equation for a simple linear regression model with one independent variable is:

$y = mx + b$

where:

- y is the dependent variable
- x is the independent variable
- m is the slope of the line
- b is the y-intercept

Assumptions:

- The relationship between the dependent and independent variables is linear.
- The errors are independent and normally distributed.
- The variance of the errors is constant.

Steps in Building a Linear Regression Model:

1. **Collect Data:** Gather data on the dependent and independent variables.
2. **Prepare Data:** Clean the data and ensure there are no missing values.
3. **Choose a Model:** Select the appropriate linear regression model based on the number of independent variables.
4. **Estimate Model Parameters:** Use a statistical method like ordinary least squares (OLS) to estimate the slope and intercept of the line.
5. **Evaluate Model Performance:** Assess the model's performance using metrics like R-squared and adjusted R-squared.
6. **Make Predictions:** Use the model to make predictions about the dependent variable for new values of the independent variable.

Advantages:

- Linear regression models are easy to understand and interpret.
- It can be used to make predictions about the future.
- It can be used to identify the relationship between different variables.

Disadvantages:

- Linear regression models assume a linear relationship between the dependent and independent variables.
- LR model can be sensitive to outliers.
- It may not be appropriate for complex datasets.

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

Anscombe's quartet is a set of four datasets that have the same mean, variance, correlation coefficient, and regression line, but which look very different when plotted. This demonstrates the importance of visual inspection of data before making any conclusions based on summary statistics alone.

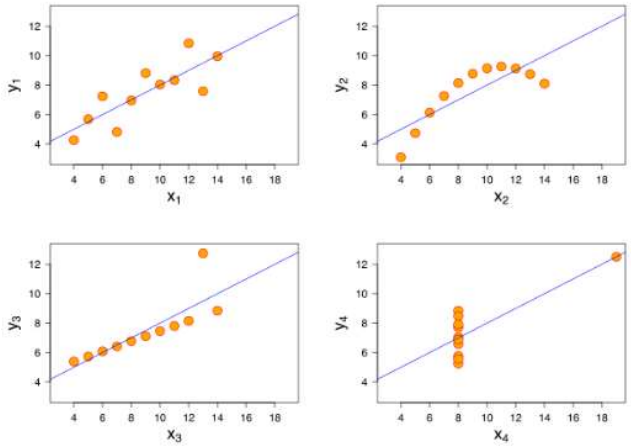
Please see below example to understand Anscombe's quartet.

Anscombe's quartet

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

Property	Value
Mean of x	9
Sample variance of x: s_x^2	11
Mean of y	7.50
Sample variance of y: s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression: R^2	0.67

Tough the statistical summary is same for all the dataset but when plotted the graph shows a completely different picture.



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

Pearson correlation coefficient (also known as Pearson's r or simply r) is a measure of the linear correlation between two variables. It is a standardized measure that ranges from -1 to 1, where:

- **-1:** Perfect negative correlation
- **0:** No correlation
- **1:** Perfect positive correlation

Pearson correlation coefficient is calculated using the following formula:

$$r = (\Sigma(x - \bar{x})(y - \bar{y})) / \sqrt{(\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2)}$$

where:

- x and y are the two variables
- $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

Pearson correlation coefficient measures the strength and direction of the linear relationship between two variables. It is important to note that correlation does not imply causation.

Assumptions of Pearson Correlation Coefficient:

- The data is normally distributed.
- The relationship between the two variables is linear.
- There are no outliers.

Applications of Pearson Correlation Coefficient:

- Identifying relationships between variables
- Predicting the value of one variable based on another.
- Selecting variables for inclusion in a statistical model.

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

Scaling

Scaling is the process of transforming numerical data so that it has a specific range or distribution. It is performed to ensure that all features in a dataset have a similar scale and to improve the performance of machine learning algorithms.

Normalized Scaling:

Normalized scaling, also known as min-max scaling, transforms the data so that it lies between a specified minimum and maximum value, typically 0 and 1. The formula for normalized scaling is:

$$x\_scaled = (x - \min(x)) / (\max(x) - \min(x))$$

where:

- x is the original value
- x_scaled is the scaled value
- min(x) is the minimum value in the dataset
- max(x) is the maximum value in the dataset

Standardized Scaling:

Standardized scaling, also known as z-score scaling, transforms the data so that it has a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

x_scaled = (x - mean(x)) / std(x)

where:

- x is the original value
- x_scaled is the scaled value.
- mean(x) is the mean of the dataset.
- std(x) is the standard deviation of the dataset.
-

Difference between Normalized Scaling and Standardized Scaling:

- **Normalized scaling:**
  - Scales the data to a specific range (e.g., 0 to 1)
  - Useful when the distribution of the data is unknown or when the data has outliers.
- **Standardized scaling:**
  - Scales the data to have a mean of 0 and a standard deviation of 1
  - Useful when the distribution of the data is normal and when there are no outliers.

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

The value of VIF (Variance Inflation Factor) can become infinite in the following scenarios:

- **Perfect multicollinearity:** This occurs when two or more independent variables in a regression model are perfectly correlated. In this case, the determinant of the correlation matrix becomes zero, and VIF becomes infinite.
- **High multicollinearity:** Even if the independent variables are not perfectly correlated, high multicollinearity can still cause VIF to become very large. This can happen when two or more independent variables are highly correlated with each other.

When does VIF become infinite?

VIF is calculated using the following formula:

VIF = 1 / (1 - R^2)

where R^2 is the coefficient of determination from a regression model.

If there is perfect multicollinearity, then R^2 will be equal to 1, and the denominator of the VIF formula will become zero, resulting in an infinite VIF.

What to do if VIF is infinite?

If VIF is infinite, it means that there is a problem with multicollinearity in the regression model. To address this issue, we can:

- **Remove one or more of the highly correlated variables:** This is the most straightforward solution, but it can lead to a loss of information.
- **Combine two or more of the highly correlated variables:** This can be done by creating a new variable that is a linear combination of the original variables.
- **Use a different regression technique:** Some regression techniques, such as ridge regression and LASSO regression, are more robust to multicollinearity than ordinary least squares regression.

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

Q-Q Plot:

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess the normality of a dataset. It compares the quantiles of the data to the quantiles of a normal distribution.

Use and Importance of Q-Q Plot in Linear Regression:

In linear regression, the normality of the residuals (errors) is one of the key assumptions. A Q-Q plot can be used to check this assumption.

- **Steps to create a Q-Q plot:**
    1. Calculate the residuals from the linear regression model.
    2. Calculate the quantiles of the residuals.
    3. Calculate the quantiles of a normal distribution with the same mean and standard deviation as the residuals.
    4. Plot the quantiles of the residuals against the quantiles of the normal distribution.

- **Interpreting a Q-Q plot:**
    - If the points on the Q-Q plot fall approximately on a straight line, then the residuals are normally distributed.
    - If the points deviate significantly from a straight line, then the residuals are not normally distributed.

Importance of Q-Q Plot in Linear Regression:

- **Assessing the normality assumption:** The Q-Q plot is a valuable tool for assessing the normality assumption of the residuals in a linear regression model.
- **Identifying outliers:** Outliers can be easily identified on a Q-Q plot as points that deviate significantly from the straight line.
- **Transforming the data:** If the residuals are not normally distributed, a Q-Q plot can be used to identify a suitable data transformation that can normalize the residuals.

References:

1. Wikipedia
2. Stack overflow