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)

   I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization:

   - Rental bike demand increases in fall season.
   - Rental bike demand is more in 2nd year.
   - Bike Demand is highest in Aug,Sep and Oct month
   - Less demand on Holiday.
   - Demand on Thursday is highest however not much variation in demand by weekdays.
   - Good weather has highest demand.

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

   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

   '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)

I have validated the assumption of Linear Regression Model based on below 5 assumptions –

- Normality of error terms
- Error terms should be normally distributed
- Multicollinearity check
- There should be insignificant multicollinearity among variables.
- Linear relationship validation o Linearity should be visible among 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 features contributing significantly towards explaining the demand of the shared bikes –

- temp
- Sep
- winter

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 (target) and one or more independent variables (predictors). It is one of the simplest and most widely used algorithms in machine learning and statistics. The goal of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that predicts the dependent variable based on the values of the independent variables.

   Types of Linear Regression

   1. Simple Linear Regression: Involves a single independent variable.

      For simple linear regression, the relationship between the dependent variable yyy and the independent variable xxx is modelled as a straight line:

      y=β0+β1x+ϵy = \beta_0 + \beta_1 x + \epsilony=β0+β1x+ϵ

   - yyy: Dependent variable (target)
   - xxx: Independent variable (predictor)
   - β0\beta_0β0: Intercept (the value of yyy when x=0x = 0x=0)
   - β1\beta_1β1: Slope (the change in yyy for a unit change in xxx)
   - ϵ\epsilonϵ: Error term (captures the noise or any other variability not explained by the linear model)

   2. Multiple Linear Regression: Involves two or more independent variables.

      For multiple linear regression with nnn independent variables, the model is represented as:

      y=β0+β1x1+β2x2+⋯+βnxn+ϵy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilony=β0+β1x1+β2x2+⋯+βnxn+ϵ

      or in matrix form:

      y=Xβ+ϵ\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}y=Xβ+ϵ

      where:

   - y\mathbf{y}y is an m×1m \times 1m×1 vector of the dependent variable for mmm observations.
   - X\mathbf{X}X is an m×(n+1)m \times (n+1)m×(n+1) matrix of input features (including a column of ones for the intercept).
   - β\boldsymbol{\beta}β is an (n+1)×1(n+1) \times 1(n+1)×1 vector of coefficients (including the intercept).
   - ϵ\boldsymbol{\epsilon}ϵ is an m×1m \times 1m×1 vector of errors.

   Assumptions of Linear Regression

   1. Linearity: The relationship between the dependent and independent variables should be linear.
   2. Independence: The residuals (errors) should be independent. This means there should be no correlation between consecutive errors.

3. Homoscedasticity: The residuals should have constant variance at every level of xxx.
4. Normality: The residuals should be normally distributed (this is particularly important for constructing confidence intervals and hypothesis testing).

Assumptions Checking

- Linearity: Need to check if the relationship between the independent and dependent variables appears linear. This can be verified using scatter plots or by plotting the residuals.
- Normality: Usage of a Q-Q plot to check is required if residuals are normally distributed.
- Homoscedasticity: Should plot residuals against fitted values to see if the variance is constant. If not, it indicates heteroscedasticity.
- Independence: Required the usage of Durbin-Watson test to check for autocorrelation in the residuals.

Limitations of Linear Regression

- Linearity Assumption: Linear regression assumes a linear relationship between the independent and dependent variables, which might not always hold in real-life scenarios.
- Outliers: Linear regression is sensitive to outliers, which can skew the results.
- Multicollinearity: In multiple regression, when independent variables are highly correlated, it can be difficult to estimate individual regression coefficients accurately.
- Overfitting: Adding too many variables can lead to overfitting, where the model performs well on the training data but poorly on unseen data.

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

Anscombe's quartet tells us about the importance of data visualization before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

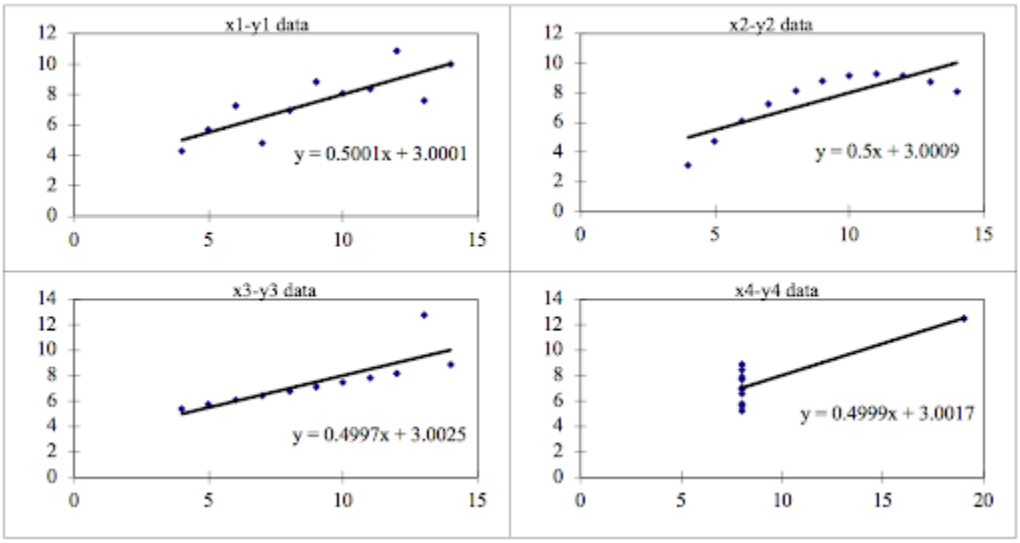
We can define these four plots as follows:

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

The statistical information for these four data sets are approximately similar. We can compute them as follows:

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

Anscombe's Quartet Four Datasets

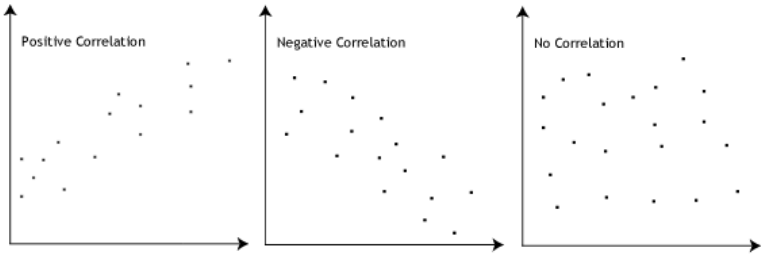
- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

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

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is one of the most widely used methods for assessing correlations in statistics.

Key Characteristics of Pearson's R:

1. Range: Pearson's R ranges from -1 to +1.
    - +1 indicates a perfect positive linear correlation, meaning as one variable increases, the other variable also increases proportionally.
    - -1 indicates a perfect negative linear correlation, meaning as one variable increases, the other variable decreases proportionally.
    - 0 indicates no linear correlation 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 step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- *It brings all of the data in the range of 0 and 1.*
- sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling:

- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).*

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

VIF is infinite, shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

(3 marks)

Q–Q plots are commonly used to compare a data set to a theoretical model. This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic. Q–Q plots are also used to compare two theoretical distributions to each other.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Importance of a Q-Q Plot in Linear Regression

1. Validation of Model Assumptions:
    - The normality of residuals is one of the key assumptions in linear regression, especially for hypothesis testing and constructing confidence intervals. Violations of this assumption can lead to incorrect conclusions.
    - By using a Q-Q plot, analysts can validate this assumption and ensure that the model's results (e.g., p-values, confidence intervals) are reliable.
2. Guiding Model Improvements:
    - If the Q-Q plot indicates that residuals are not normally distributed, it suggests that the linear regression model may not be the best fit for the data.
    - This can guide the analyst to consider alternative models or transformations (such as a logarithmic transformation of the dependent variable) to better satisfy the model assumptions.
3. Identifying Outliers and Leverage Points:
    - Q-Q plots can help identify outliers that may disproportionately affect the model. Points that fall far from the line in a Q-Q plot might be outliers or leverage points.
    - Identifying these points can prompt further investigation to determine if they represent data entry errors, unusual but valid observations, or if they require some form of treatment (e.g., robust regression).