Assignment-based Subjective Questions

Questions 1. From your analysis of the categorical variables from **the** dataset, what could you infer about their effect on the dependent variable?

Answer 1: Based on the boxplots we can infer the below

- The demand i,e the count is high in the "fall" season
- The demand for renting bike has increased in 2019 as compared to 2018
- If we analyze, we can notice that in Summer season during Weekend the demand is high
- During "Excellent" weathersit which refers to "Clear, Few clouds, Partly cloudy, Partly cloudy" weather conditions the cnt is high.
- We can notice that the demand continuously increases from Jan till Jun and then there is a decrease in Jul again increasing the month of Sep and then decreasing till end of year.

Question 2. Why is it important to use drop_first=True during dummy variable creation?

Answer 2: drop_first=True is important to use, as it helps in reducing the extra columns that gets created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Dropping the first columns as (p-1) dummies can explain p categories .

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

Answer 3- Temp and atemp have the highest correlation with the target variable(cnt). The correlation is 0.63.

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

Answer 4: Checked for the assumptions of Linear Regression

1) Independence of Residuals – Residuals should be independent of each other
   Validated this by using Durbin-Watson test. ACF should not show any significant patterns, and Durbin-Watson statistic should be close to 2.In the model finalized the value is 2.02
2) Linearity: The relationship between the independent variables and the dependent variable is linear.
   Validated linearity by plotting the actual vs. predicted values and looking for a pattern or curvature in the residuals. A scatterplot of residuals against fitted values should shows no clear pattern.
3) No Multicollinearity: Independent variables should not be highly correlated with each other. Calculated the correlation matrix or use variance inflation factors (VIF). High correlations or VIF values above a 5 threshold indicate multicollinearity. Removed the correlated variables

4) Homoscedasticity: Residuals should have constant variance across all levels of the independent variables.
   Used scatterplots of residuals against predicted values

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

Answer 5: The top 3 features contributing significantly towards the demand of the shared bikes are

1) Year(+ve correlation of 2051.84)
2) Temp(+ve correlation of 3677.87)
3) Weathersit_Bad(-ve correlation of -2286.17) – Weathersit value of 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

General Subjective Questions

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

Answer 1: Linear regression is a statistical and machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors) by fitting a linear equation to the observed data. It is widely used for tasks such as prediction, forecasting, and understanding the relationships between variables.

Assumptions: Linear regression is based on several key assumptions, including linearity (the relationship between variables is linear), independence of errors (residuals are not correlated), homoscedasticity (constant variance of residuals), and normally distributed residuals.

Simple Linear Regression: Simple linear regression is used when there's a single independent variable. The linear relationship between the target variable (Y) and a single predictor (X) is represented by the equation:

$Y=\beta 0+\beta 1*X+\varepsilon$

Y is the dependent variable.

X is the independent variable.

β0 is the y-intercept (the value of Y when X is 0).

β1 is the slope (the change in Y for a one-unit change in X).

ε represents the error term or residuals.

Multiple Linear Regression: Multiple linear regression extends the model to multiple independent variables. It can be represented as:

$Y=\beta0+\beta1*X1+\beta2*X2+...+\beta n*Xn+\varepsilon$

Y is the dependent variable.

X1, X2, ..., Xn are the independent variables.

β0 is the y-intercept.

β1, β2, ..., βn are the coefficients for each independent variable.

ε represents the error term.

Model Training: The goal is to find the values of the coefficients (β0, β1, β2, ..., βn) that minimize the sum of the squared residuals (the vertical distance between observed data points and the regression line). This is typically done using the method of least squares.

Predictions: Once the model is trained, you can make predictions by plugging new values of the independent variables into the equation.

Evaluation: Linear regression models are evaluated using various metrics, including:

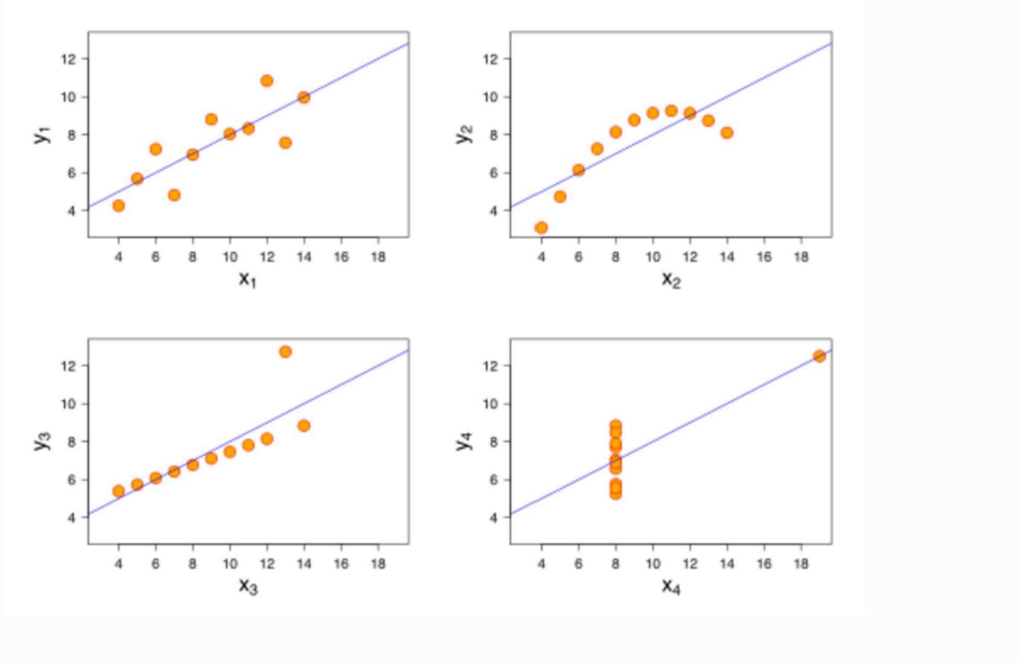
Mean Squared Error (MSE): The average of the squared differences between predicted and actual values.

Root Mean Squared Error (RMSE): The square root of MSE, which is in the same unit as the target variable.

R-squared (R^2): A measure of how well the independent variables explain the variation in the dependent variable. It ranges from 0 to 1, with higher values indicating a better fit.

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

Answer 2: Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that 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 like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.



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

Answer 3: Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

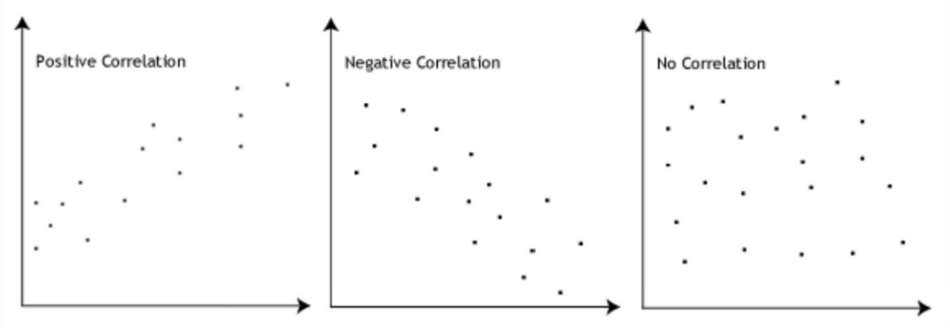
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

- r = 0 means there is no linear association

- r > 0 < 5 means there is a weak association

- r > 5 < 8 means there is a moderate association

- r >8 means there is a strong association



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here, • r =correlation coefficient

• xi =values of the x-variable in a sample

• x̅=mean of the values of the x-variable

• yi =values of the y-variable in a sample

• ȳ =mean of the values of the y-variable

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

Answer 4: 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. When we collect data set it 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 must do scaling to bring all the variables to the same level of magnitude. Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.

MinMaxScaling :x = x-min(x)/ max(x) – min(x)

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).

Standardisation:x = x-mean(x)/ sd(x)

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer 5: VIFi=1/1−Ri^2

Where:

VIFi is the Variance Inflation Factor for the i-th independent variable.

Ri^2 is the coefficient of determination (R-squared) when the i-th variable is regressed against all the other independent variables.

A high VIF indicates a strong multicollinearity problem for the corresponding variable. Typically, a VIF of 1 indicates no multicollinearity, and VIF values above 1 suggest increasing degrees of multicollinearity. A commonly used threshold is a VIF of 5, above which multicollinearity is considered problematic.

If "infinite VIF" is encountered for a variable, it means that the multicollinearity is so severe that the VIF calculation approaches infinity. This situation arises when a variable can be perfectly predicted by a linear combination of other independent variables in the model. In practical terms, this means that the variable doesn't contribute any unique information to the regression model because its effects are completely captured by other variables. In such cases, it may be necessary to reconsider the model and possibly remove one or more of the highly correlated variables to address the multicollinearity issue and obtain meaningful results from the regression analysis.

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

Answer 6: A Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics and data analysis to assess whether a dataset follows a specific theoretical distribution, such as a normal distribution. It compares the quantiles of the dataset to the quantiles of the theoretical distribution, typically a standard normal distribution.

Q-Q plots are important in linear regression for assessing the distributional assumptions of the model, primarily the assumption of normally distributed residuals. By visually inspecting the Q-Q plot, you can determine whether your data conforms to these assumptions or if there are issues that need to be addressed in your regression analysis.