# Assignment-based Subjective Questions Answers

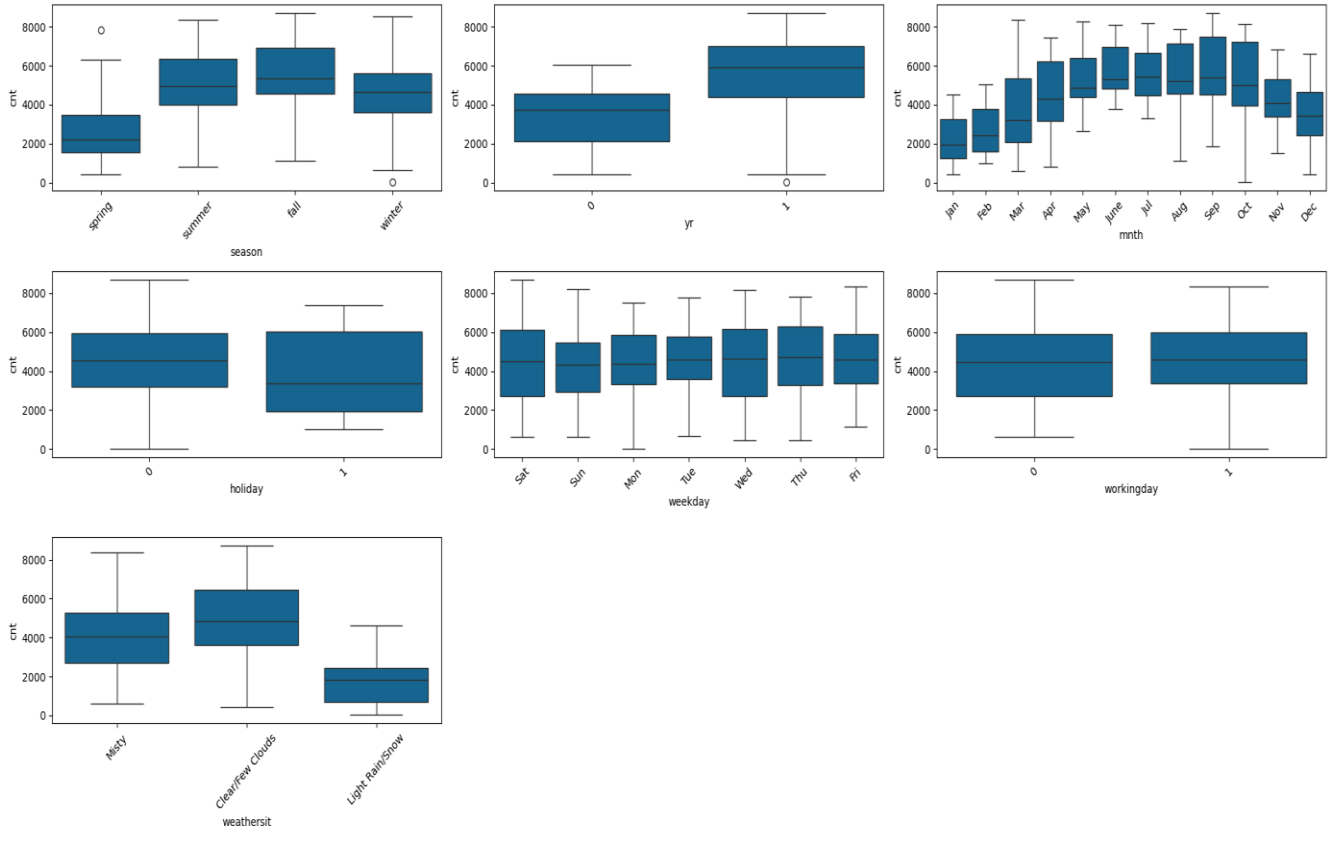
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans:

There are many categorical variables in the dataset like year, weather situation, weekday, month, holiday, workingday, season etc. and demand(cnt) is dependent variable.

Out of these, season, month, weather situation show a good variation across their values as observed in chart below.

Values of holiday and year also show significant variation in demand across their values.



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

Ans:

Dropping first category while creating dummies creates n-1 features for n unique values of categories for that variable.

This reduces the number of features in the overall dataset for training decreasing the complexity while retaining the necessary information.

We can also interpret the coefficients of the remaining dummy variables relative to this baseline category. This makes the results more intuitive and easier to understand.

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

Ans:

Looking at the pair-plot among the numerical variables, **temperature** has the highest correlation with the target variable indicated by a correlation coefficient value of 0.65

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

Ans:

To validate the assumptions of Linear Regression I plotted following 2 graphs which were used to validate the corresponding assumptions as mentioned below

1. Distribution of Residual Errors in Histogram
   - Helped to validate the normality of distribution of errors
2. Scatter plot for y_train vs. y_train_pred
   - Helped to validate linear relationship between actual and predicted values
   - Helped to validate homoscedasticity by checking variance of errors

Also I checked VIF scores for all variables to ensure that there is minimal multicollinearity between features.

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

Ans:

Based on final model following are the top three features contributing significantly towards explaining the demand of the shared bikes

1. Temperature (Coeff Value: 0.51) :
   Increase in temperature increases the demand

2. Weather Situation Particularly light rain or snow (Coeff Value: -0.3):
   Light rain or snow weather conditions tend to reduce bike demand

3. Year (Coeff Value: 0.23):
   If year is 2019 there is high chance that the demand is more

## General Subjective Questions

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

Ans:

Linear regression algorithm is used to create a model that describes the relationship between a dependent variable and one or more independent variables. Depending on whether there are one or more independent variables, a distinction is made between simple and multiple linear regression analysis.

There are 2 types of Linear Regression

**Simple Linear Regression**: Involves one dependent variable and one independent variable.

**Multiple Linear Regression**: Involves one dependent variable and multiple independent variables.

**Simple** Linear Regression

$$\hat{y} = b \cdot x + a$$

**Multiple** Linear Regression

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

Here **a** is constant term, **b1, b2, ..., bk** are coefficients which are found during model training and **x1, x2, ..., xk** are independent variables and **y-cap** is the dependent variable.

Linear Regression Algorithm can be defined mathematically as follows

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

**Main objective of Linear Regression Algorithm** is to minimize the total squared error between actual and predicted value across all the points in training dataset.

Following are the key assumptions of Linear Regression Algorithm since in order to interpret the results of the regression analysis meaningfully, certain conditions must be met.

1. **Linearity**: The relationship between independent and dependent variables is linear.
2. **Independence**: Observations are independent of each other.
3. **Homoscedasticity**: The variance of residuals is constant across all levels of the independent variables.
4. **Normality of Errors**: The residuals (errors) are normally distributed.
5. **No Multicollinearity**: Independent variables should not be highly correlated.

Example of Linear Regression

Finding out which factors have an influence on the cholesterol level of patients. For this purpose, you analyze a patient data set with cholesterol level, age, hours of sport per week and so on.

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

Ans:

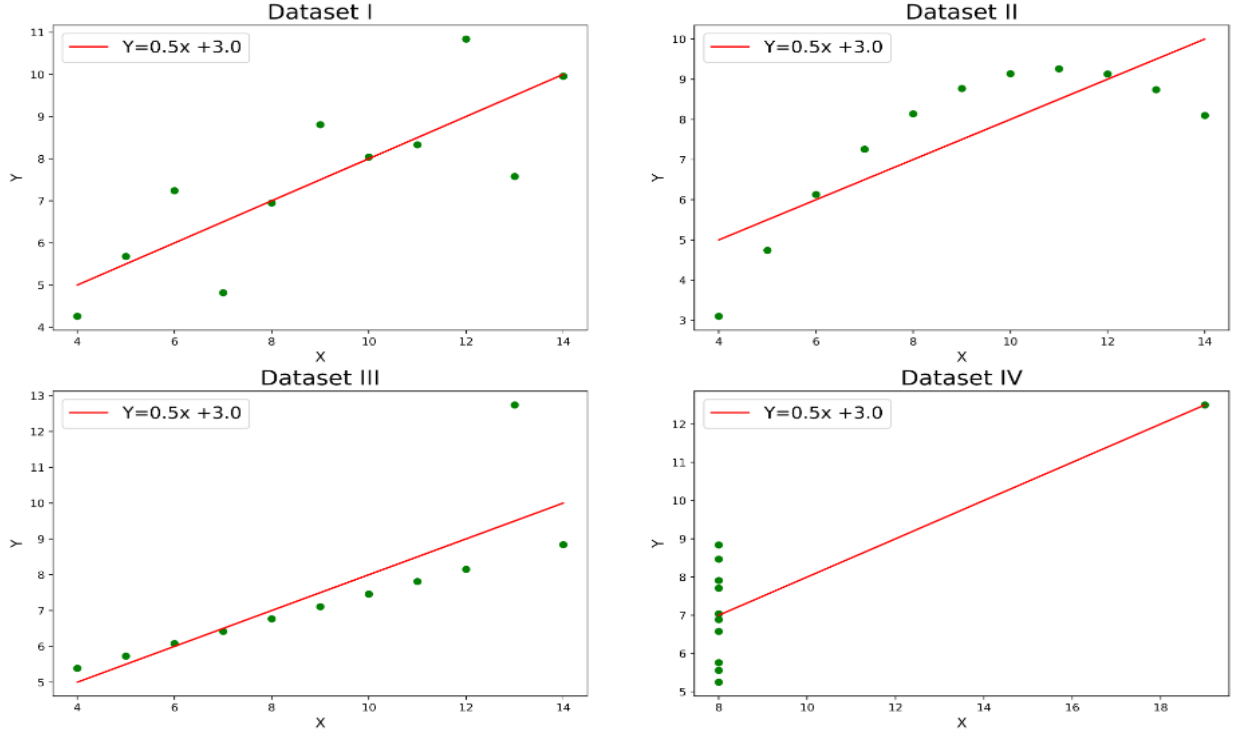
Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Summary Stats for Anscombe's quarter

| | I | II | III | IV |
|---|---|---|---|---|
| Mean_x | 9.000000 | 9.000000 | 9.000000 | 9.000000 |
| Variance_x | 11.000000 | 11.000000 | 11.000000 | 11.000000 |
| Mean_y | 7.500909 | 7.500909 | 7.500000 | 7.500909 |
| Variance_y | 4.127269 | 4.127629 | 4.122620 | 4.123249 |
| Correlation | 0.816421 | 0.816237 | 0.816287 | 0.816521 |
| Linear Regression slope | 0.500091 | 0.500000 | 0.499727 | 0.499909 |
| Linear Regression intercept | 3.000091 | 3.000909 | 3.002455 | 3.001727 |

Visualization of Anscombe's quarter




Observations in above images

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

To conclude, While the descriptive statistics of Anscombe's Quartet may appear uniform, the accompanying visualizations reveal distinct patterns, showcasing the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

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

Ans:

The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.

The Pearson coefficient is a mathematical correlation coefficient representing the relationship between two variables, denoted as X and Y.

Pearson coefficients range from +1 to -1, with +1 representing a positive correlation, -1 representing a negative correlation, and 0 representing no relationship.

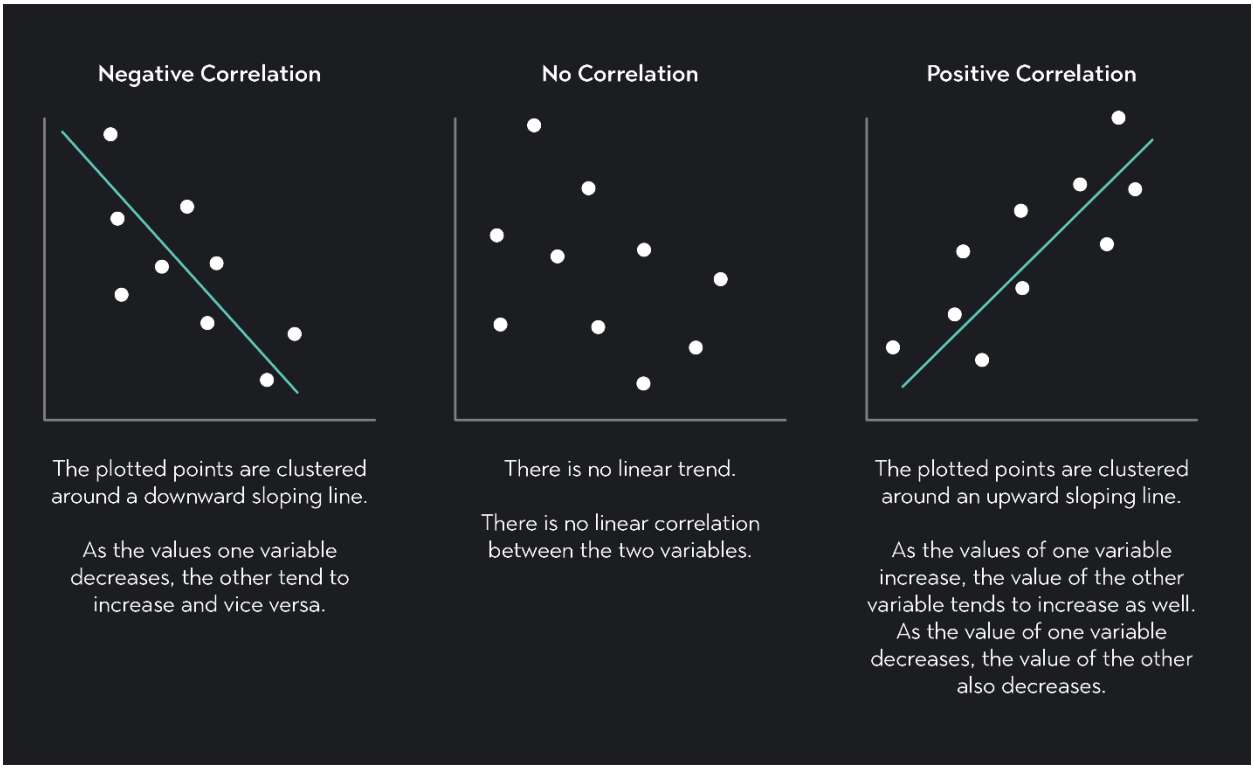
The Pearson coefficient shows correlation, not causation.

Formula for Pearson's r is given by

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

$r$ = correlation coefficient
$x_i$ = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
$y_i$ = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Variations of Pearson Correlation Coefficient and its meaning



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

Ans:

What is scaling?

Scaling is a preprocessing technique used to standardize the range of independent variables (features) in a dataset. It transforms the data so that it fits within a specific scale, which can improve the performance of machine learning algorithms.

Why scaling?

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.

## Difference Between Normalized Scaling and Standardized Scaling

1. **Normalized Scaling (Min-Max Scaling)**:
   - **Definition**: Transforms the data to fit within a specified range, usually [0, 1].
   - **Formula**:

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)}$$

   - **Usage**: Useful when you want to maintain the relationships in the data and when the scale is important for algorithms like neural networks.
2. **Standardized Scaling (Z-score Scaling)**:
   - **Definition**: Centers the data around the mean with a unit standard deviation.
   - **Formula**:

$$Z = \frac{X - \mu}{\sigma}$$

     where $\mu$ is the mean and $\sigma$ is the standard deviation.
   - **Usage**: Useful when the data follows a Gaussian distribution, and it allows for comparison between different datasets.

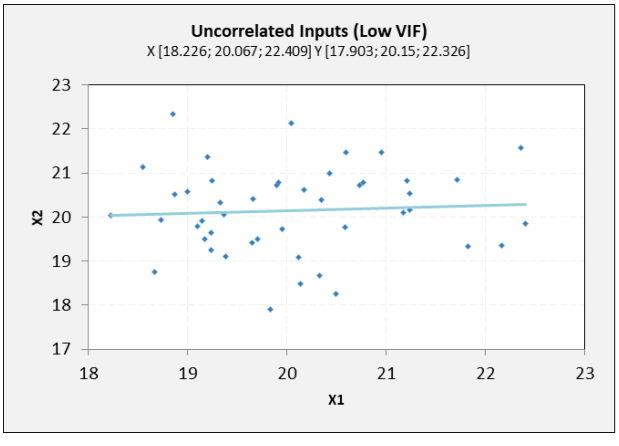
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

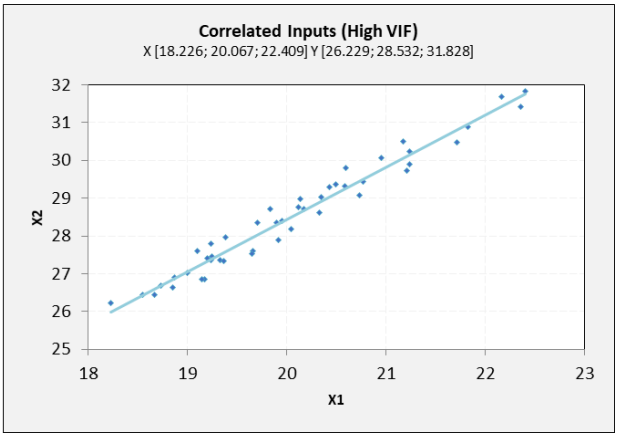
Ans:

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A VIF value of 1 indicates no correlation among the predictor variables, while values above 1 suggest increasing multicollinearity.

Infinite VIF occurs when one predictor variable is a perfect linear combination of one or more other predictor variables. This means there is a perfect correlation (correlation of 1 or -1) between the variables, leading to redundancy.

For example, if you have two variables X1 and X2 such that X2 = C * X1 (for some constant C), the regression model cannot uniquely estimate the coefficients for both X1 and X2.




When VIF is infinite, it signals a serious multicollinearity problem. It's essential to address this by:

- Removing or combining highly correlated predictors.
- Re-evaluating the model to ensure relevant variables are included without redundancy.

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

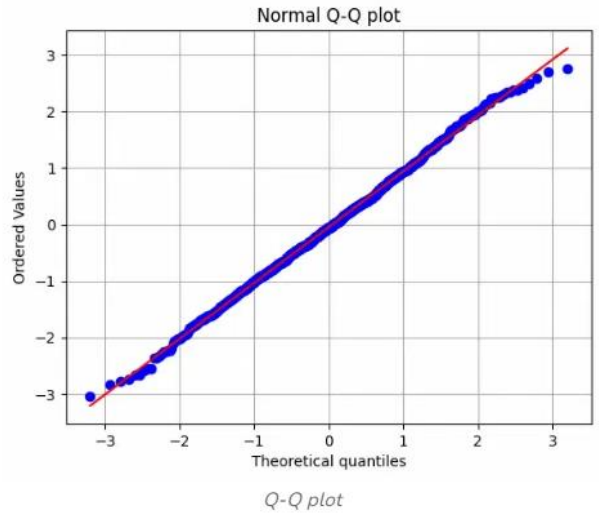
(3 marks)

Ans:

The quantile-quantile (q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

**Sample Q-Q Plot**




*Q-Q plot*

**Interpretation of Q-Q plot**

If the points on the plot fall approximately along a straight line, it suggests that your dataset follows the assumed distribution.

Deviations from the straight line indicate departures from the assumed distribution, requiring further investigation.

**Importance of Q-Q Plots in Linear Regression**

1. Validation of Assumptions:

   Validating the assumption of normality is crucial for making valid statistical inferences from the regression model, such as hypothesis tests and confidence intervals.

2. Improving Model Performance:

By identifying non-normal residuals, you can take corrective measures (e.g., transformations, removing outliers), ultimately leading to a more reliable and robust model.

3. Visual Interpretation:

Q-Q plots provide a straightforward visual representation, making it easier to communicate the findings regarding normality to stakeholders.