# Assignment-based Subjective Questions

## **Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit). Total Marks: 3 marks (Do not edit)**

**Answer:** Output from the analysis of categorical variables shows a strong relationship with the dependent variable `cnt`, explaining 83.4% of its variability. Key findings include:

|  |  |
| --- | --- |
| Year | Counts increase significantly over time, reflecting growth. |
| Seasonal Effects | Spring has a negative effect, while winter shows a positive effect, indicating seasonal fluctuations in demand. |
| Monthly Patterns | December, January, and November have lower counts, while September shows a positive impact, likely due to favourable conditions. |
| Weekly Trends | Counts are generally lower on Sundays. |
| Weather Impact | Light weather conditions negatively affect counts, indicating that certain weather types reduce demand. |

These categorical factors significantly explain seasonal and temporal patterns in `cnt`, confirming their relevance in the model.

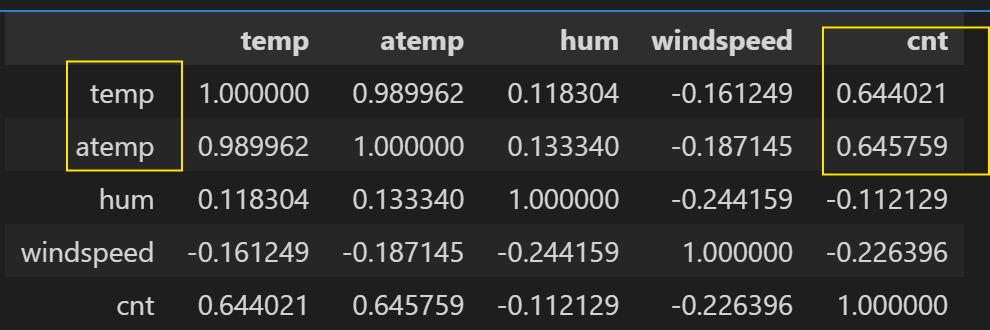
## **Question 2. Why is it important to use drop\_first=True during dummy variable creation? (Do not edit). Total Marks: 2 marks (Do not edit)**

**Answer:** Using `drop\_first=True` in dummy variable creation avoids multicollinearity by dropping one category and setting it as a baseline. This prevents redundant information and ensures each dummy variable compares its category to a reference, allowing the model to produce accurate and interpretable coefficients.

For example, with the `season` variable (spring, summer, fall, winter), using `drop\_first=True` drops `summer` as the baseline. This way, the dummy variables `season\_spring`, `season\_fall`, and `season\_winter` represent each season’s effect relative to summer. This avoids redundancy and ensures accurate coefficient estimates.

## **Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit) Total Marks: 1 mark (Do not edit)**

**Answer:** atemp and temp both have near to same correlation



## **Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit) Total Marks: 3 marks (Do not edit)**

**Answer:** To validate the assumptions of linear regression after building the model on the training set, I examined several key aspects. First, I checked for linearity between predictor variables and the response by analyzing residual plots; a random scatter pattern around zero suggested a linear relationship. Next, I assessed the normality of the error terms by reviewing a Q-Q plot and histogram of residuals, confirming that the errors followed a roughly normal distribution, which is essential for accurate inference. I also verified homoscedasticity by plotting residuals against fitted values, ensuring that the variance of the errors remained constant across all levels of the predictors. Finally, I addressed multicollinearity by calculating the Variance Inflation Factor (VIF) for each predictor, aiming for values below 5 to confirm that predictors were relatively independent. Together, these checks validated that the model met the core assumptions of linear regression, supporting the reliability and accuracy of its predictions.

## **Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit) Total Marks: 2 marks (Do not edit)**

**Answer**: Based on the final model, the top three features contributing significantly to explaining the demand for shared bikes (i.e., the `cnt` variable) are:

|  |  |
| --- | --- |
| Year (yr) | with a high coefficient (0.2438) and a t-statistic of 29.213, indicating a strong positive impact on bike demand over time. |
| Temperature (temp) | with the largest coefficient (0.3754) and a t-statistic of 12.190, showing that higher temperatures are strongly associated with increased bike demand. |
| Weather Situation (weathersit\_light) | with a large negative coefficient (-0.2612) and a t-statistic of -10.028, indicating that light weather conditions (likely less favourable) reduce bike demand significantly. |

These features exhibit high statistical significance and substantial impact on demand.

# General Subjective Questions

## **Question 6. Explain the linear regression algorithm in detail. (Do not edit) . Total Marks: 4 marks (Do not edit)**

**Answer: *1. Linear Regression*** is a key data science tool for predicting continuous outcomes. Linear regression is a fundamental algorithm in statistics and machine learning that models the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to observed data. Its simplicity and interpretability make it widely used for predictive modelling and inferential statistics.

The basic form of a linear regression equation is *y = b0 + b1\*x* where "y" is the dependent variable, "x" is the independent variable, "b0" is the intercept (where the line crosses the y-axis), and "b1" is the slope (how much "y" changes for every unit change in "x")

The goal of linear regression is to find the line that best fits the data points, minimizing the overall error between the predicted values and the actual values. This is often achieved using the "least squares" method, which calculates the line that minimizes the sum of squared errors.

Example - Predicting future sales based on historical sales data and other factors like marketing spend

***2. Assumptions of Linear Regression:*** For linear regression to be valid and reliable, certain assumptions should ideally hold:

Linearity: The relationship between each independent variable and the dependent variable should be linear.

Independence: Observations should be independent of each other.

Homoscedasticity: The residuals (errors) should have constant variance across all levels of the independent variables.

Normality of Errors: The residuals should follow a normal distribution, which is particularly important for small sample sizes.

Low Multicollinearity: Independent variables should not be highly correlated with each other, as multicollinearity can make estimates unstable.

***3.Fitting the Model: Ordinary Least Squares (OLS):*** The most common method for estimating the coefficients β in linear regression is Ordinary Least Squares (OLS). The goal of OLS is to minimize the sum of squared residuals (the squared differences between the observed and predicted values).

***4. Model Evaluation:*** After fitting the model, we evaluate its performance and validate the assumptions:

R-squared: Measures the proportion of variance in the dependent variable explained by the independent variables. Higher values indicate a better fit.

Adjusted R-squared: Adjusts R-squared for the number of predictors, penalizing for adding irrelevant predictors.

Residual Plots: Plotting residuals vs. fitted values helps detect non-linearity, heteroscedasticity, and outliers.

Q-Q Plot: Helps verify the normality of residuals by plotting them against a normal distribution.

Variance Inflation Factor (VIF): Checks multicollinearity, with VIF values above 5 (or 10) indicating high multicollinearity.

Once the model is validated, it can be used for prediction.

A diagram of a linear regression

Description automatically generated

Useful link - <https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/>

<https://datatab.net/tutorial/linear-regression>

<https://towardsdatascience.com/linear-regression-explained-1b36f97b7572>

## **Question 7. Explain the Anscombe’s quartet in detail. (Do not edit) Total Marks: 3 marks (Do not edit)**

**Answer:** Anscombe’s Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

A graph of a function

Description automatically generated with medium confidence

* Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
* Data-set II — shows a curve shape but doesn’t show a linear relationship (might be quadratic?).
* Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
* Data-set IV — looks like the value of x remains constant, except for one outlier as well.

Useful link - <https://medium.com/analytics-vidhya/anscombes-quartet-an-importance-of-data-visualization-856b3d1bd403>

<https://www.geckoboard.com/blog/anscombes-quartet-why-summary-metrics-lie/>

<https://www.research.autodesk.com/publications/same-stats-different-graphs/>

## **Question 8. What is Pearson’s R? (Do not edit). Total Marks: 3 marks (Do not edit)**

**Answer:** The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Key Assumptions:

Independence: Each case should be independent of others.

Linearity: There must be a linear relationship between the variables, which can be verified through a scatterplot. If the plot forms a straight line, the criterion is met.

Homoscedasticity: The scatterplot of residuals should approximate a rectangular shape.

Characteristics: Range: The coefficient’s value ranges from +1 (perfect positive correlation) to -1 (perfect negative correlation), with 0 indicating no correlation.

Unit Independence: The coefficient is unaffected by the units of measurement, ensuring comparability across different scales.

Symmetry: The correlation between two variables remains consistent, regardless of the variable order (X with Y or Y with X).

A white grid with blue text

Description automatically generated

A screenshot of a graph

Description automatically generated

Useful links: <https://www.scribbr.com/statistics/pearson-correlation-coefficient/>

<https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/>

<https://www.sciencedirect.com/topics/computer-science/pearson-correlation>

## **Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit). Total Marks: 3 marks (Do not edit)**

**Answer:** Feature Scaling is a critical step in building accurate and effective machine learning models. One key aspect of feature engineering is scaling, normalization, and standardization, which involves transforming the data to make it more suitable for modelling. These techniques can help to improve model performance, reduce the impact of outliers, and ensure that the data is on the same scale.

**A diagram of a normal distribution

Description automatically generated**

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.

**A math equation with a number and a symbol

Description automatically generated with medium confidence**

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).

**A black text on a white background

Description automatically generated**

* sklearn.preprocessing.scale helps to implement standardization in python.
* One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Useful links - <https://medium.com/@premal.matalia/what-is-scaling-why-is-scaling-performed-normalized-vs-standardized-scaling-5113c86688f8>

<https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/>

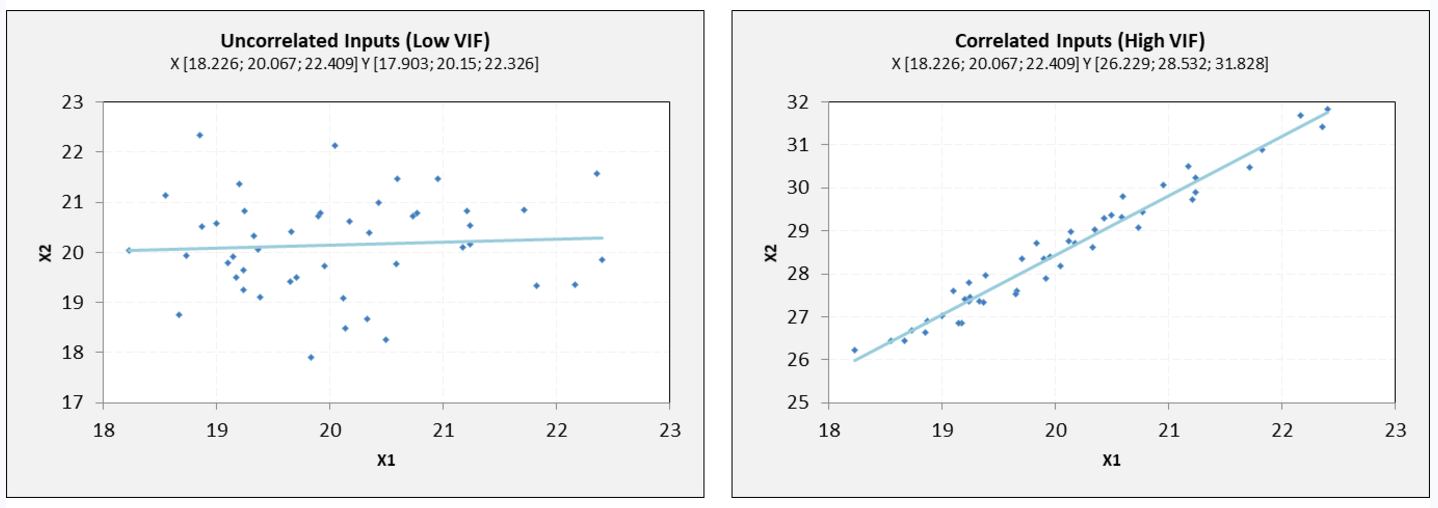
<https://www.simplilearn.com/normalization-vs-standardization-article>

## **Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit) Total Marks: 3 marks (Do not edit)**

**Answer:** A VIF value becomes infinite when there is perfect multicollinearity, meaning one independent variable in your regression model can be perfectly predicted by a linear combination of the other independent variables, essentially creating an exact duplicate and causing the VIF calculation to become undefined and reach infinity.

A math problem with black text

Description automatically generated

****

VIF equal to 1 = variables are not correlated

VIF between 1 and 5 = variables are moderately correlated

VIF greater than 5 = variables are highly correlated

The higher the VIF, the higher the possibility that multicollinearity exists, and further research is required. When VIF is higher than 10, there is significant multicollinearity that needs to be corrected.

**Dummy Variable Trap:** The dummy variable trap occurs when all categories of a categorical variable are included as dummy variables in a regression model, leading to multicollinearity. This creates a situation where one or more predictors are perfectly correlated, making it impossible to accurately estimate their individual effects. To avoid this issue, we drop one dummy variable to serve as the reference category. By doing so, the model estimates the effect of each remaining category relative to the reference, ensuring no redundancy and allowing for reliable coefficient estimates.

Useful links: <https://www.sigmamagic.com/blogs/what-is-variance-inflation-factor/>

<https://www.investopedia.com/terms/v/variance-inflation-factor.asp>

<https://www.geeksforgeeks.org/ml-dummy-variable-trap-in-regression-models/>

## **Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit) Total Marks: 3 marks (Do not edit)**

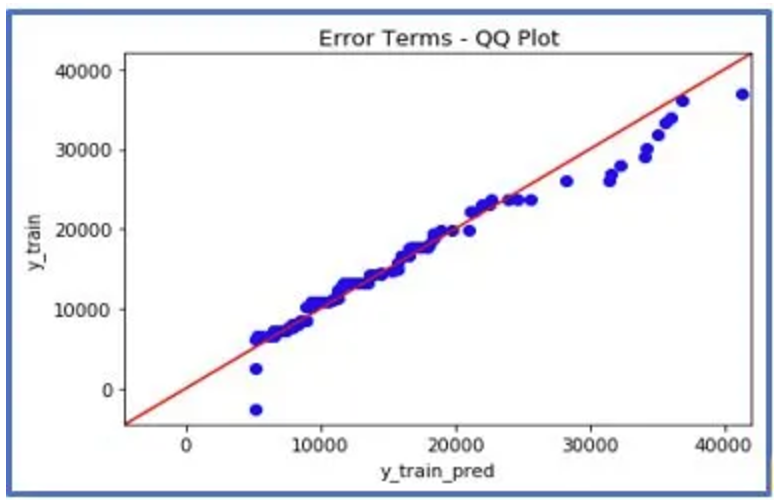
**Answer:** A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to visually compare the distribution of a dataset to a theoretical distribution, like the normal distribution, by plotting the quantiles of your data against the quantiles of the theoretical distribution; in linear regression, it's particularly important for assessing whether the residuals (errors) from the model follow a normal distribution, a key assumption for reliable inference.

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

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

****

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

**A graph with blue dots

Description automatically generated**

d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

statsmodels.api provide qqplot and qqplot\_2samples to plot Q-Q graph for single and two different data sets respectively.

Useful links: <https://medium.com/@premal.matalia/q-q-plot-in-linear-regression-explained-ab040567d86f>

<https://www.geeksforgeeks.org/quantile-quantile-plots/>

<https://towardsdatascience.com/significance-of-q-q-plots-6f0c6e31c626>

<https://www.analyticsvidhya.com/blog/2021/09/q-q-plot-ensure-your-ml-model-is-based-on-the-right-distributions/>