# 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)

**Answer.**

- Season: Fall, Summer and Winter seasons show comparatively higher demand than Spring season
- Year: 2019 shows an increase in demand as these bike-sharing systems are slowly gaining popularity
- Month: April to November months have pretty high and steady demand, Jan to Feb are the lowest.
- Weekday: No insight could be drawn as demand seems to be steady for all weekdays
- Weather: Demand grows with good weather (Bad weather < Normal weather < Good weather)

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

**Answer.**

Using drop_first=True helps to prevent multicollinearity, improve model stability, and enhance the interpretability of regression models.

Let's take an example with the variable "day of the week" which has categories like Monday, Tuesday, Wednesday, and so on. When we convert these categories into numbers, we create what we call dummy variables. Each category gets its own dummy variable.

Now, if we have all the dummy variables, we might think we need one for each day of the week: Monday, Tuesday, Wednesday, etc. But actually, we only need one less than the total number of categories. Because if we know the values of all the other dummy variables, we can figure out the value of the missing one. For example, if we know it's not Monday, Tuesday, Wednesday, or any other day, then it must be Thursday.

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

**Answer.**

The correlation for both ‘temp’ and ‘atemp’ with target variable ‘cnt’ is 0.63, the highest among all correlations.

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

**Answer.**

Assumptions of Linear Regression:

- Linear relationship between X and Y – Validated using Pair Plots between Independent and Target Variable.

- Error terms are normally distributed (not X, Y) – Validated error distribution using Dist. Plot (histogram) of the residual error.
- Error terms have constant variance (homoscedasticity) – Validated by maintaining p-values less than equal to 0.05 for all selected features. Other features were not considered for prediction.
- Error terms are independent of each other – Validated by maintaining VIFs lower than 5 for all selected features. Other features were not considered for prediction.

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

**Answer.**

Based on final model, below were the top 3 features contributing significant contributors:

- **Temperature** with a coefficient of **0.5013**
- **Year** with a coefficient of **0.2403**
- **Windspeed** with a negative coefficient of **-0.1792**

# General Subjective Questions

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

**Answer.**

Linear regression is a statistical method used to model the relationship between a dependent variable (often denoted as ‘*y*’ and one or more independent variables (often denoted as x1, x2, …, xn). The goal of linear regression is to find the "best-fitting" straight line, known as the regression line, that represents the relationship between the independent variables and the dependent variable.

The regression line is represented by the equation:

y = β0 + β1*x1 + β2*x2 + … + βn*xn + €

where:
- β0 is the intercept term,
- β1, β2, …, βn are the coefficients (also known as slopes) of the independent variables,
- x1, x2, …, xn are the values of the independent variables,
- € is the error term (representing the difference between the actual and predicted values).

The linear regression algorithm aims to estimate the values of the coefficients (slopes) β0, β1, …, βn that minimize the sum of squared differences between the observed and predicted values of the dependent variable. This is typically achieved using the method of least squares.

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

**Answer.**

Anscombe's quartet is a set of four datasets that have nearly identical descriptive statistics (such as mean, variance, correlation, and regression coefficients), yet have very different graphical representations. These datasets were created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics.

Each dataset in Anscombe's quartet consists of 11 paired observations of two variables, x and y. Despite having the same summary statistics, the datasets exhibit different patterns when plotted. The four datasets demonstrate the following characteristics:

- **Dataset I**: This dataset forms a perfect linear relationship between x and y, with no variability in the residuals. It highlights the importance of examining scatterplots to understand the relationship between variables.
- **Dataset II**: This dataset also exhibits a linear relationship between x and y, but with one outlier that significantly influences the regression line and correlation coefficient. It emphasizes the impact of outliers on regression analysis.

- **Dataset III**: This dataset shows a non-linear relationship between x and y, with a clear quadratic pattern. It underscores the necessity of considering alternative functional forms when modeling data.

- **Dataset IV**: This dataset has nearly identical summary statistics to Dataset I but contains an outlier that greatly affects the regression line. It highlights the importance of graphical inspection to identify influential observations.

  Overall, Anscombe's quartet serves as a cautionary example against relying solely on summary statistics and reinforces the value of data visualization in understanding and interpreting relationships within datasets. It underscores the importance of exploring data graphically to uncover patterns, trends, and potential anomalies that may not be evident from numerical summaries alone.

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

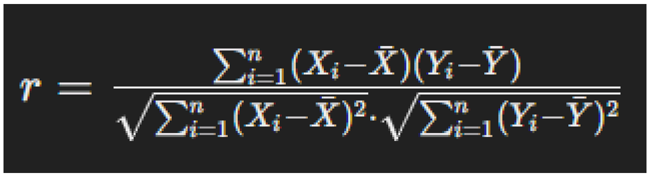
**Answer.**

Pearson's correlation coefficient, commonly denoted as r, is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the linear association between the variables.

Pearson's r can take values between -1 and 1:

- If r = 1, it indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable increases proportionally.
- If r = -1, it indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- If r = 0, it indicates no linear relationship between the variables.

The formula to calculate Pearson's correlation coefficient r between two variables X and Y is:



$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\cdot\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

Where:
- Xi and Yi are individual data points of variables X and Y,
- X(bar) and Y(bar) are the means of variables X and Y, respectively,
- n is the number of data points.

Pearson's correlation coefficient is widely used in various fields such as statistics, economics, psychology, and biology to measure the strength and direction of the linear relationship between variables. It helps in understanding the degree to which changes in one variable are associated with changes in another variable, aiding in hypothesis testing, predictive modeling, and data exploration.

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

**Answer.**

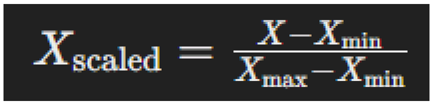
Scaling is the process of transforming numerical features in a dataset to a common scale. It involves adjusting the values of the features so that they fall within a specified range or distribution. Scaling is typically performed to ensure that all features contribute equally to the analysis and to improve the performance and interpretability of machine learning models.

The primary reasons for scaling are:

- Improving model performance: Many machine learning algorithms, such as gradient descent-based algorithms and distance-based algorithms (e.g., K-nearest neighbors), perform better when the features are on a similar scale. Scaling prevents features with large magnitudes from dominating the optimization process.
- Facilitating interpretation: Scaling ensures that the coefficients or weights assigned to the features in the model are comparable. This makes it easier to interpret the importance of each feature in influencing the outcome.

There are two common methods of scaling: normalized scaling and standardized scaling.

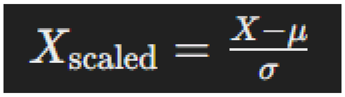
- Normalized scaling: In normalized scaling, also known as min-max scaling, the values of the features are scaled to fall within a specified range, usually between 0 and 1. The formula for normalized scaling is:



$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

where X is the original feature value X(min) is the minimum value of the feature, and X(max) is the maximum value of the feature.

- Standardized scaling: In standardized scaling, also known as z-score normalization, the values of the features are transformed to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:



$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

where X is the original feature value, μ is the mean of the feature, and σ is the standard deviation of the feature.

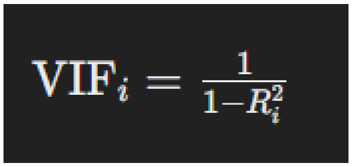
The main difference between normalized scaling and standardized scaling lies in the scale and distribution of the transformed values. Normalized scaling preserves the original range of the data, while standardized scaling centers the data around 0 and adjusts the spread of the data. The choice between the two scaling methods depends on the specific requirements of the machine learning algorithm and the characteristics of the dataset.

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

**Answer.**

The occurrence of infinite values for the Variance Inflation Factor (VIF) typically indicates perfect multicollinearity among the independent variables in the regression model. Perfect multicollinearity occurs when one or more independent variables can be exactly predicted by a linear combination of other independent variables in the model.

In the context of VIF, perfect multicollinearity leads to an infinite value because it results in the denominator of the VIF formula becoming zero. The VIF formula is:



$$\mathrm{VIF}_i = \frac{1}{1-R_i^2}$$

where Ri-square is the R-square value obtained when regressing the i-th independent variable on all the other independent variables in the model.

When perfect multicollinearity is present, Ri-square will be equal to 1, meaning that the i-th independent variable can be perfectly predicted by the other independent variables. As a result, the denominator (1 - Ri-square) becomes zero, leading to an undefined value for VIF.

Perfect multicollinearity can arise due to several reasons, such as:

- Including a variable that is a linear combination of other variables already present in the model.
- Including dummy variables for all categories of a categorical variable, leading to perfect collinearity among them.
- Data coding errors or data duplication.

To address infinite VIF values caused by perfect multicollinearity, it's essential to identify and resolve the underlying issues in the dataset or model specification. This may involve removing redundant variables, combining variables, or reevaluating the model's structure. Additionally, techniques such as ridge regression or principal component analysis (PCA) can be used to mitigate multicollinearity in regression models.

**6.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

**Answer.**

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a specified distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution, typically the normal distribution.

In a Q-Q plot:
- The x-axis represents the quantiles of the theoretical distribution (e.g., the standard normal distribution).
- The y-axis represents the quantiles of the dataset being analyzed.

If the dataset follows the specified distribution closely, the points in the Q-Q plot will fall approximately along a straight line. Deviations from this line indicate departures from the specified distribution.

In linear regression, Q-Q plots are often used to assess the assumption of normality of residuals. Residuals are the differences between the observed values and the values predicted by the regression model. The assumption of normality of residuals is crucial for linear regression models because it ensures that the statistical inference and hypothesis tests based on the model are valid.

The use and importance of a Q-Q plot in linear regression can be summarized as follows:

- Checking Normality Assumption: By examining the Q-Q plot of residuals, we can visually inspect whether the residuals are normally distributed. If the points in the plot deviate significantly from the straight line, it suggests that the assumption of normality may be violated.
- Detecting Skewness and Outliers: Q-Q plots can help identify skewness or heavy-tailedness in the distribution of residuals, as well as potential outliers. Skewed or heavy-tailed residuals can indicate violations of regression assumptions and may require further investigation or data transformation.
- Model Evaluation: Q-Q plots provide a quick and intuitive way to evaluate the adequacy of the regression model. A Q-Q plot with residuals that closely follow the diagonal line suggests that the model's assumptions are met, whereas departures from the line may indicate issues with the model's specification or data quality.

Overall, Q-Q plots serve as a valuable diagnostic tool in linear regression analysis, aiding in the assessment of model assumptions, the identification of potential problems, and the refinement of the regression model.