Assignment-based Subjective Questions

Q1. 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 and Weathersit are the two categorical variables. In seasons Fall has the highest impact on dependent variables, while in weathersit, type 2 which is *Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist* has the biggest impact.

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Answer: For a categorical variable with ‘n’ levels, you create ‘n-1’ new columns each indicating whether that level exists or not using a zero or one. Therefore, when all are zero, it represents the excluded level. Keeping that level will be redundant, hence we drop it.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer: Registered has the highest correlation (0.95) with the target variable which is cnt.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Answer: By Plotting the histogram of the error terms.

sns.distplot((y_train - y_train_predict), bins = 20)

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer: (1) temp
(2) weathersit (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
(3) windspeed

General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is one of the simplest and most commonly used algorithms in machine learning and statistics. It is used to model the relationship between one or more independent variables (features) and a dependent variable (target) by fitting a linear equation to the observed data.

The goal of linear regression is to predict the value of the dependent variable based on the values of the independent variables. It assumes a linear relationship between the input features and the target output.

There are two types of Linear Regressions: Simple Linear Regression which involves only one independent variable and Multiple Linear Regression which involves two or more independent variables.

Assumptions of Linear Regression:

Linear regression relies on several key assumptions:

1. **Linearity**: The relationship between the independent and dependent variables is linear.
2. **Independence**: The observations in the dataset are independent of each other.
3. **Homoscedasticity**: The variance of the error terms ϵ\epsilonϵ is constant (i.e., no matter the value of the independent variables, the errors should have the same variance).
4. **Normality**: The error terms are normally distributed.
5. **No Multicollinearity (for Multiple Regression)**: The independent variables should not be highly correlated with each other.

Limitations of Linear Regression

- Assumes a linear relationship, which may not be suitable for complex, non-linear data.
- Sensitive to outliers, which can skew the model.
- Assumes no or little multicollinearity between the independent variables.

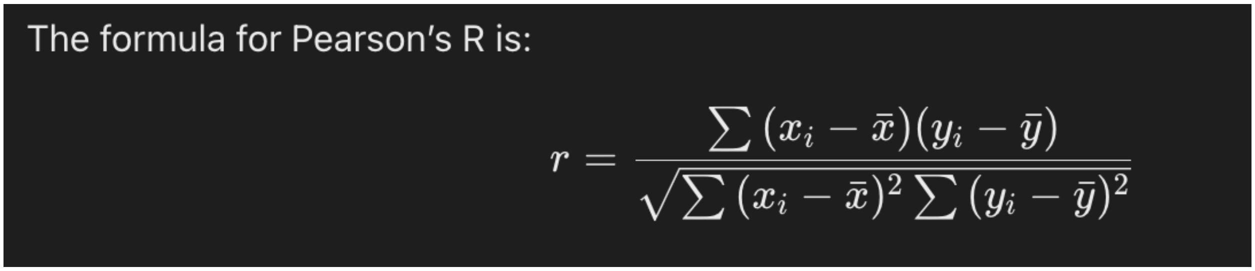
Q2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe's Quartet is a set of four different datasets created by statistician **Francis Anscombe** in 1973. The purpose of Anscombe’s Quartet is to demonstrate the importance of **visualizing data** before performing statistical analyses. All four datasets have nearly identical simple statistical properties, but they differ dramatically when visualized in a scatter plot.

Anscombe’s Quartet shows that relying solely on summary statistics such as the mean, variance, correlation, and regression line can be misleading, as these statistics can be nearly identical for datasets that are visually and structurally very different.

Q3. What is Pearson’s R? (3 marks)

Answer: **Pearson’s R**, also known as the **Pearson correlation coefficient** or **Pearson product-moment correlation coefficient**, is a statistical measure that calculates the strength and direction of the **linear relationship** between two continuous variables. Named after Karl Pearson, who developed it in the early 20th century, this metric is widely used in statistics, data analysis, and machine learning to understand the degree of association between two variables.



The formula for Pearson’s R is:

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

Where:

- rrr is the Pearson correlation coefficient.
- xix_ixi and yiy_iyi are the individual data points for the two variables xxx and yyy, respectively.
- xˉ\bar{x}xˉ and yˉ\bar{y}yˉ are the means of the variables xxx and yyy.
- The numerator is the **covariance** between xxx and yyy, which measures how the variables vary together.
- The denominator is the product of the **standard deviations** of xxx and yyy, which standardizes the result.

Interpretation of Pearson’s R

Pearson’s R produces a value between **-1** and **+1**, where:

- r=1r = 1r=1: A perfect **positive linear correlation**. As xxx increases, yyy increases in a perfectly straight line.
- r=−1r = -1r=−1: A perfect **negative linear correlation**. As xxx increases, yyy decreases in a perfectly straight line.
- r=0r = 0r=0: **No linear correlation**. There is no linear relationship between the variables, although there could still be a non-linear relationship.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead to a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

You can scale the features using two very popular method:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

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-square, etc.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: A high VIF indicates that a predictor is highly collinear with other predictors, which can make coefficient estimates unstable and unreliable. VIF can take an **infinite value** when there is **perfect multicollinearity** in the data. This means that one predictor variable is an **exact linear combination** of one or more other predictor variables. In this situation, the linear regression model is unable to calculate the regression coefficients because the design matrix (which represents the predictor variables) becomes **singular** or **non-invertible**.

To address perfect multicollinearity and avoid infinite VIF values, you can:

- **Remove redundant variables**: Eliminate one of the collinear variables if they provide the same information.
- **Combine variables**: If the variables are linearly dependent, create a new variable by combining them (e.g., adding them together).
- **Regularization**: Techniques like **Ridge regression** (which adds a penalty term to the coefficients) can help reduce the impact of multicollinearity.

Q6. 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** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, usually the **normal distribution**. It plots the quantiles of the observed data against the quantiles of a theoretical distribution (like the normal distribution) to assess how closely the data follow that distribution.

A Q-Q plot is a scatter plot where:

- The **x-axis** represents the theoretical quantiles (the expected values if the data were perfectly normally distributed).
- The **y-axis** represents the sample quantiles (the actual data points from the dataset).

In a Q-Q plot:

- **Quantiles** are cut points dividing a probability distribution into intervals with equal probabilities.
- The goal of the plot is to see whether the observed data follows a particular theoretical distribution.

Interpreting a Q-Q Plot

- **If the points lie on or near the 45-degree reference line**, the sample data comes from the theoretical distribution (e.g., normal distribution).
- **If the points deviate from the reference line**:
  - **Upward or downward curvature** suggests a **skewed** distribution.
  - **S-shaped curve** indicates **heavier tails** or a **light-tailed** distribution.
  - **Flattened ends** suggest the data has less variance than expected.

Use of a Q-Q Plot in Linear Regression

In the context of **linear regression**, the Q-Q plot is used primarily to assess whether the **residuals** (the difference between the observed and predicted values) follow a **normal distribution**.

Why is Normality Important in Linear Regression?

In linear regression, certain assumptions are made about the error terms (residuals), one of which is that the residuals are normally distributed. The normality of residuals is crucial for:

1. **Validating Statistical Inference**: Many hypothesis tests (such as t-tests for coefficients) and confidence intervals in linear regression rely on the assumption that residuals are normally distributed.
2. **Predictive Accuracy**: If residuals are not normally distributed, predictions may not be accurate, especially for outliers or extreme values.
3. **Homoskedasticity**: While homoskedasticity refers to constant variance of residuals, normality of residuals often complements this by suggesting that errors are symmetrically distributed around zero.