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?

Ans1. From my analysis here are few mentioned points, where we can infer about their effect on the dependent variable:

- Season 3 has highest demand for rental bikes.
- We can see that demand for next year has been increased.
- From month January to June, we can see that demand is getting increased, september month has highest demand for bike, after september month demand goes decreasing.
- During the year end and beginning, it is less, could be due to extreme weather conditions.
- Demand is decreased, whenever there is a holiday.
- Weekday is not giving clear picture about demand; therefore, both looks similar.
- If we look through Working days, then also there is no clear-cut picture.
- 1st weather list has highest demand that is clear weather.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans2. When creating dummy variables also known as one-hot encoding, it is important for avoiding the "dummy variable trap”.

It is encoding by 1 and 0. By dropping the first category, essentially create a reference category.

The coefficients for the remaining dummy variables represent the effect of each category relative to this reference. This simplifies interpretation.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans3. The highest correlation with the target variable is as follows:

(1) By this plot can see atemp and temp are highly co - related with each other.

(2) The target variable can’t have highest corelation with temp and atemp.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans4. Validating the assumptions of linear regression is crucial to ensure that model is reliable and that the results are interpretable.

(1) Relationship between independent and dependent variables should be linear.

Example — we have 'nt' target variable we plot graph between 'cat' and 'temp' there is good relationship between both of them. Residuals errors should be minimum.

(2). No Multicollinearity - like independent variables should not be highly correlated with each other.

Ex. — temp and atemp are highly correlated with each other.

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

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans5. the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

(1) holiday

(2) temp

(3) yr

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans1. Linear regression is a simple algorithm used to predict a target variable (the thing you want to predict) based on one or more input variables (called features). The goal is to find the best-fit line or plane that describes the relationship between the target and features.

In simple linear regression, we look at the relationship between one feature x and the target y. The relationship is described by the equation:

$y=\beta0+\beta1x+ \epsilon$

- y is the target variable.
- x is the feature (predictor).
- β0 is the y-intercept (the value of y when x=0).
- β1 is the slope (it shows how much y changes when x increases by 1).
- ϵ is the error term, the difference between the actual and predicted values of y.

The algorithm works by finding the values of β0 and β1 that minimize the error (the difference between predicted and actual y values). This is done using a method called Ordinary Least Squares (OLS), which calculates the line that best fits the data.

In multiple linear regression, more features are included, and the relationship is more complex, but the basic idea is the same: find the best combination of feature weights to predict the target accurately.

Q2. Explain the Anscombe's quartet in detail.

Ans2. Anscombe's Quartet is a set of four datasets created by the statistician Francis Anscombe in 1973 to show how summary statistics like mean, variance, and correlation can be misleading without looking at the data visually.

The four datasets in the quartet have nearly identical statistical properties:

- Same mean of x and y.
- Same variance for x and y.
- Same correlation between x and y.

Despite these similarities, the datasets look very different when plotted on a graph. Three of them show a clear linear relationship between x and y, while the fourth shows a pattern with one outlier that greatly affects the analysis.

Anscombe's Quartet teaches an important lesson: relying only on numbers like averages or correlation can hide the true nature of the data. It emphasizes the importance of visualizing data to understand its real structure before drawing conclusions.

Q3. What is Pearson's R?

Ans3. Pearson's R (also known as the Pearson correlation coefficient) is a number that tells you how strongly two things are related to each other. It's used to measure the *correlation* or relationship between two variables. The value of Pearson's R ranges from -1 to +1:

- A **+1** means a perfect positive relationship: as one thing goes up, the other also goes up.
- A **-1** means a perfect negative relationship: as one thing goes up, the other goes down.
- A **0** means no relationship: changes in one thing don't affect the other.

For example, if you were looking at the relationship between studying hours and test scores, a high positive Pearson's R (close to +1) would suggest that more study hours tend to lead to higher scores. Pearson's R helps you understand if two things move together in a predictable way and how strong that connection is.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans4 Scaling is the process of adjusting the values of data so that they fall within a similar range or have similar units. In real-world datasets, features (or columns) can have different ranges, units, or magnitudes. For example, in a dataset of people's information, the "age" feature might range from 0 to 100, while the "income" feature could range from a few thousand dollars to millions. If these features are used together in a model without scaling, the feature with larger numbers (income) could dominate, leading to incorrect results. Scaling helps to make sure that each feature contributes equally to the analysis.

Scaling is performed to ensure that machine learning models treat all features fairly. Many algorithms, like k-nearest neighbours (KNN), linear regression, and support vector machines (SVM), rely on the distance between data points. If one feature has a much larger scale than others, it will overpower the smaller features, making the model's predictions less accurate. Scaling also helps improve the speed and performance of models, especially those that involve optimization algorithms.

There are two common types of scaling:

1. **Normalized Scaling**: This technique transforms the data so that it fits within a specific range, usually between 0 and 1. It's done by subtracting the minimum value of each feature and then dividing by the range (maximum - minimum). This method is helpful when you need to maintain a bounded range, for example, when using neural networks.
2. **Standardized Scaling**: This method adjusts the data so that it has a mean of 0 and a standard deviation of 1. To standardize, you subtract the mean of the feature and divide by its standard deviation. This is useful when data has outliers or when the data distribution is not uniform.

Both methods help to ensure that the features are on a similar scale, but they serve different purposes based on the type of model and the distribution of the data.

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

Ans5. The **Variance Inflation Factor (VIF)** measures how much the variance (spread) of a regression coefficient is inflated due to multicollinearity — when two or more independent variables in a model are highly correlated with each other.

When the VIF value is **infinite**, it usually happens because one independent variable is perfectly or almost perfectly correlated with another variable in the model. In other words, one variable can be almost exactly predicted using another variable. This causes the regression model to become unstable, and the calculation of the VIF results in a value that cannot be defined or is extremely large.

For example, if you have two variables, **X1** and **X2**, that are almost the same (or perfectly correlated), the model can't distinguish between them, leading to an infinitely large VIF for one or both of them. This indicates that one variable is redundant and should be removed to improve the model.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Ans6. A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, such as a normal distribution. It plots the quantiles (specific data points, ordered by value) of your dataset against the quantiles of the reference distribution. If the dataset follows the theoretical distribution, the points on the plot will form a straight line.

In **linear regression**, the Q-Q plot is primarily used to check if the **residuals** (the differences between the observed and predicted values) are normally distributed. This is an important assumption in linear regression because many statistical tests, like confidence intervals and hypothesis tests, assume that residuals are normally distributed.

In summary, a Q-Q plot helps assess whether the residuals of a linear regression model meet the normality assumption, ensuring that statistical tests based on the model are valid and trustworthy.