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?

Initial Observations:

- "yr": The demand for bikes from "BoomBikes" has increased in 2019, as compared to 2018
- "season": The demand for bikes is high during the "fall", as compared to the demand during the other seasons
- "holiday": The volatility in demand is a bit more during holidays, as compared to the other days
- "mnth": The demand is high for each of the months from May to September as compared to the first 4 months and last two months of the year
- "weekday": The volatility in demand is high during Wednesday and Saturday, as compared to the other days
- "weathersit": The demand for bikes from "BoomBikes" is high when the weather is "1: Clear, Few clouds, Partly cloudy, Partly cloudy", as compared to the demand during the other weather patterns

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

When we create dummy variables/columns (also known as one-hot encoding), we convert the categorical variables into binary columns, to represent different categories. Using "drop_first=True" is important in this case.

Let us take a look at why this is important.

To reduce extra columns:

We avoid creating an extra column for the first category.

Correlation reduction:

Dummy variables include correlations among themselves. When we include all dummy columns, they can be linearly dependent leading to multicollinearity. By dropping the first column, we mitigate this issue and maintain independence among the dummy variables.

General Rule:

As a general rule, we should use n-1 columns to represent the dummy variables, while dealing with a categorical variable having n levels. The dropped column serves as the reference category, and the remaining columns indicate the presence of other categories.

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

From the pair plot, it looks like the numerical variable's "temp" and "atemp" have the highest correlation with the target variable ("cnt").

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

The assumptions of linear regression are as follows:

1. Linearity

This has been checked even before developing the model. Have tried a scatter plot which is "cnt" versus each of the following: "temp", "atemp", "hum" and "windspeed"

2. Independence

The residuals must be independent. Have plotted a scatter plot comparing the residuals and the predictions for y (for the train set) and there are no patterns among consecutive residuals.

3. Homoscedasticity

The residuals should have a constant variance at every level of x. The scatter plot comparing the residuals and the predictions for y (for the train set) show constant variance.

4. Normality

The residuals of the model should follow a normal distribution. Have come up with a distribution plot for residuals and the residuals are normally distributed.

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

As per the final model, the top three features contributing significantly towards explaining the demand of shared bikes are,

- "temp"
- "light snow"
- "yr"

Light-snow is a column created from the initial column in the dataset "weathersit".

General Subjective Questions

1. Explain the linear regression algorithm in detail.

At a high level, the linear regression algorithm can be explained as follows:

- It is a supervised learning algorithm (i.e., the target data is labelled and the algorithm learns from it)
- It models the relationship between one or many input variable(s) and an output variable
- Mainly used for predicting continuous outcomes

Let us look at the linear regression algorithm a little more in detail.

Linear regression is a technique used to predict the value of unknown data by using other related/known data. Mathematically, we model the unknown or dependent variable (response) and the known or independent variable (predictor) as a linear equation.

The linear equation, in this case, is the line or surface that best fits the data. The important parts of the linear equation are

- Intercept, and
- The coefficient of independent variable(s)

The primary objective in a linear regression algorithm is to arrive at the best-fit line (i.e. the line or plane in which the error between the predicted and actual values is kept to a minimum).

The Mean Squared Error (MSE) cost function is employed to determine the optimal values for the intercept and the coefficient(s) of input variable(s), thereby providing the best-fit line for the data points.

Using the MSE cost function, the iterative process of gradient descent is applied to update the value of intercepts and coefficients. The idea is to ensure that the MSE value converges to a “global minima” satisfying the most accurate fit of the linear regression to the dataset.

The final result is a linear regression line that minimizes the overall squared differences between the predicted and actual values, providing an optimal representation of the underlying relationship in the data.

Assumptions of linear regression:

1. Linear Relationship: The relationship between the independent and dependent variables is linear. The linearity assumption can be tested with the scatter plots.
2. No or little multicollinearity: The linear regression assumes that there is little or no multicollinearity in the data. It can be tested with a correlation matrix, tolerance, VIF (Variance Inflation Factor)

3. No auto-correlation: Linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent of each other.
4. Normality: The errors follow a normal distribution
5. No homoscedasticity: The variance of the errors/spread should be constant across all levels of the independent variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet has four distinct datasets that have identical descriptive statistical properties but have different representations when we visualize them using a scatter plot. The datasets were created by the statistician Francis Anscombe in 1973 to highlight the importance of visualizing data/exploratory data analysis and also to bring out the point that summary statistics alone can be misleading at times.

Each of the four datasets includes 11 x-y pairs of data. The data sets share identical summary statistics,

- Same means for both x and y
- Same variances for both x and y
- same correlation coefficients, and
- same linear regression lines

However, when the data was plotted, each dataset seemed to have a unique connection between x and y, with unique variability patterns and different correlation strengths.

Anscombe's Quartet emphasizes the importance of combining statistical analysis with exploratory data analysis/visualization for good data interpretation.

3. What is Pearson's R?

"Pearson's r" also referred to as "Pearson Correlation Coefficient", is a measure of the strength and direction of the linear relationship between two variables. It takes a value between -1 and +1.

- A value of 0 indicates there is no relationship/correlation between the two variables
- A value between 0 and 1 indicates there is a positive correlation between the two variables (i.e. when one variable changes, the other one also tends to change in the same direction). If the value is close to 1, it indicates a strong positive relationship.
- A value between 0 and -1 indicates there is a negative correlation between the two variables (i.e. when one variable changes, the other one tends to change in the opposite direction). If the value is close to -1, it indicates a strong negative relationship.

Note: The Pearson coefficient shows correlation and not causation.

One of the areas in which the Pearson Coefficient can be useful is in investments to hedge for risks and also in portfolio diversification.

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

Scaling:

The data set that is used to create a model would contain features that vary in magnitude, units, and range. Scaling is a data pre-processing technique that is used to transform the values/features in a data set to a similar scale/magnitude. It is the process of transforming the values of the features of a dataset till they are within a specific range, e.g., 0 to 1 or -1 to 1.

Reasons for performing scaling:

Since the algorithm will consider only the magnitude and not the units, we will have an incorrect model, if scaling is not performed. Scaling will ensure that each of the features has a comparable impact on the model and that features with larger values do not dominate the model.

Also, algorithms like Gradient Descent converge faster when the features are scaled.

Therefore, for the machine learning model to interpret the features on the same scale and also to improve the performance of the algorithm, we need to perform feature scaling.

Difference between normalized scaling and standardized scaling:

Normalized Scaling:

- This method scales the model using minimum and maximum values.
- It is a scaling method in which the numbers are scaled and moved between (0, 1) or (-1, 1).
- Normalization is helpful when we are not sure about the feature distribution or when the feature distribution is not “Gaussian Distribution”.
- Outliers in the data will be impacted by normalized scaling as it needs a wide range to function correctly.

Standardized Scaling:

- It is a scaling method in which the numbers are scaled using mean and standard deviation.
- Values on a scale are not constrained to a particular range.
- Standardization presupposes that the distribution of your data is Gaussian
- In contrast to normalization, Standardization does not always have a bounding range. So, the outliers in your data won’t be impacted in this scaling technique.

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

VIF (Variance Inflation Factor) is used to measure the degree of multicollinearity among independent variables in a regression model. Multicollinearity occurs when an independent variable in the model is highly correlated with one or more of the other independent variables in the model.

A high VIF value implies there is a strong multicollinearity. Mostly, a VIF value of 5 or more is considered to be high. When the value of VIF is infinite for an independent variable, it means that the particular independent variable can be predicted perfectly by the other variables in the model.

Mathematically, when the coefficient of determination is 1, the VIF tends to infinity. To explain this further, in the case of perfect correlation, we get R^2 =1, which leads to 1/(1-R^2) to be infinity.

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

Q-Q plots are known as Quantile-Quantile plots.

- They can be used to plot the quantiles of a sample distribution against the quantiles of a theoretical distribution (or)
- the quantiles of the first data set against the quantiles of the second data set.

Use/Importance of Q-Q plots:

Q-Q plot is a statistical tool that helps us to know whether a sample comes from a specified distribution and also to visualize how the sample deviates from the original distribution.

Also, it helps us to determine whether two datasets come from populations with a common distribution.

A few more advantages of Q-Q plots are as follows,

- To understand the distributional fit and other aspects related to distribution - like shifts in location, shifts in scale, changes in symmetry
- To detect departures from normality by looking at whether the points are on the 45-degree line or not
- It is robust and can be used even with small sample sizes
- The plot reveals outliers and skewness by showing deviations from the expected line
- Multiple datasets can be compared using the Q-Q plot. If the datasets have similar Q-Q plots, then they fall under the same distribution
- It helps to validate whether the transformed data aligns better with a theoretical distribution