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?

Ans:

The dataset 'day.csv' includes predictor variables that can be differentiated int numerical and Categorical.

The numerical variables include temp, atemp, hum, windspeed, etc.

The categorical variables include season, yr, month, holiday, weatherlist, etc.

From the analysis on categorical variables, I have concluded that: -

- there is increase in demand of bikes in fall and summer season.
- There is an increase in demand of bikes from month of May which is supposed to be a holiday for children and teenagers.
- Week day is not showing and inconsistency over the week.
- In 'cnt' vs 'weather' situation analysis we can clearly see a drop in demand in the snow weather.

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

Ans: -

The get_dummies() function is used to convert categorical variable into dummy/indicator variables. Data of which to get dummy indicators.

After creating the dummy variables for the categorical variable, we use

Drop_first=True as it helps in reducing the extra column created during dummy variable creation. This also reduces the correlations created among dummy variables.

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

Ans: -

The dataset has numerical variables that include temp, atemp, windspeed, hum, etc.
From the pairplot of these variables with cnt we can see there is linear relation between cnt with temp and atemp.
We can use these corelation for predction.

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

Ans: -

For the model we have created, the NULL hypothesis is H0:B1=B2=...=Bn=0 where, Bi is every predictor variable's numeric value multiplies by slope.

The Alternative hypothesis is H1: at least one Bi! = 0 .

From the model we got expression as:

cnt = 0.0336 + yr 0.2363 + temp0.3452 + windspeed (-0.1312)+ season_spring *(-0.1661) + mnth_jul -0.0604 + weekday_sat* 0.195 + weekday_sun *0.0393 + weathersit_mist *-0.2667

We can see there is no 0 values that's why 'We Reject the NULL hypothesis'.

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

Ans: -

From the overall analysis of model, after seeing the p -values and VIF values and other summary we got out top three variables that affects most to the booking of the bike sharing booking are:

- Temperature (temp) - A coefficient value of '0.5636' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of '-0.3070' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.
- Year (yr) - A coefficient value of '0.2308' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: -

In supervised learning the data to be processes is labelled, linear Regression is a machine learning algorithm based on supervised learning.

Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis). To calculate best-fit line linear regression uses a traditional slope-intercept form(y = mX + c).

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship

The algorithm performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

While creating the linear Regression model we first analyse independent variables who has linear relation with output variable.

Eg. The price of house to be calculated from the dataset where independent variables are area, bedrooms, terrace, etc.

2. Explain the Anscombe's quartet in detail.

Ans: -

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

3.What is Pearson's R?

Ans: -

The Pearson's correlation coefficient (PCC, pronounced /ˈpɪərsən/) — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of linear correlation between two sets of data.

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive.

The bivariate Pearson Correlation produces a sample correlation coefficient, r, which measures the strength and direction of linear relationships between pairs of continuous variables

The Pearson correlation coefficient (r) is used to identify patterns in things whereas the coefficient of determination (R^2) is used to identify the strength of a model.

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

Ans: -

The most common techniques of feature scaling are Normalization and Standardization. Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation.

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step. Normalization adjusts the values of your numeric data to a common scale without changing the range whereas scaling shrinks or stretches the data to fit within a specific range. Scaling is useful when you want to compare two different variables on equal grounds

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

Ans: -

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. Variance inflation factor (VIF) is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis. Multicollinearity inflates the variance and type II error. It makes the coefficient of a variable consistent but unreliable

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

Ans: -

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution