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)

An1: From the analysis we can see that season and weather sit have impact on the dependent variable cnt.

Weather situation has –ve coefficient hence increase in weather situation would decrese the count

Season too have impact ots a +ve coefficient hence incease in value should increase the bike hire counts

Year has just 2 values so not much difference although its +ve coefficient

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

A2 : Drop_first= true setting helps us to get it till n-1 levels , actually we don't need it till n level as its redundant here its not needed since one of the combination will be uniquely representing this redundant column

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

A3:Temp is having a strong correlation with cnt

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

A4 doing a scatter plot on test and predicted y values and seein the R^2 and adj R^2 values

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

A5 : Top 3 features impacting the demand are

Temperature (coefficient of approx. 0.5)

Year (coeeff of +0.2) and weather_sit(coeff of -0.3)

General Subjective Questions

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

Ans: The algorithm of Linear regression model is involves various steps , the various steps outlined are namely :

a.) **Data Validation and Correction** : This step involves reading the data , understand the structure , look for duplicates ,null values , outliers etc.

b.) **Visualize the data and Transform data** : Draw scatter, box plots to determine multicollinierity and understand the relationships among the features it will help us to determine whether the relationships are strong or weak, here we also convert the values of categorical variables using dummy variables

c.) **Split the data set in Test and Train** : Using standard Python library sklearn we split the data set in 2 parts in ratio of 70:30 normally.

d.) **Scaling** : In any data set every feature will have different types and scales of values some may be small some big numbers, Scaling does not impact the model but it helps to bring the features to 'comparable' scale , if we don't do scaling it will result in disproportionate coefficients , this helps in getting the coefficients to same scale. We normally use out of 2 ways either Min Max scaler or Standardization

e.) **Build the model** : You can use standard python libraries like statsmodel to build the model

f.) **Fine tune the model** : Getting the summarized results and comparing P values and VIF you can further drop the features and fine tune it , this is a iterative process till we get a acceptable values for vif for defined features , this can be out fit model for predictions.

g.) **Residual analysis of Training data** : Look for error terms and validate the assumption of linear regression

h.) Make predictions on chosen model : Using the model now apply it on test data for predictions

i.) Evaluate the Model : Draw scatter plots in test nad predicted data to confirm the linearity and validate the R^2 and Adjusted R^2 values , based on the same draw the inferences.

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

Ans2: It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs it helps in visualization of data , it consists of 4 pairs of graphs or data sets which are having same mean , std and regression line but are different qualitatively.

Idea in terms of model building is that inspite of having same linearity the 11 data points generate insights in to pattern, correlation , trends, outliers that normally we don't get in summary statistics

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

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.The Pearson correlation coefficient can also be used to test whether the relationship between two variables is significant.

So for eg.Positive correlation When one variable changes, the other variable changes in the same direction.(r between 0 and 1)

No correlation There is no relationship between the variables.(r is zero)

Negative correlation When one variable changes, the other variable changes in the opposite direction.(r is negative ie. 0 and -1)

Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables It helps in determine of how close the observations are to a line of best fit. It also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive. When r is 1 or –1, all the points fall exactly on the line of best fit:

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

A4: Scaling : In any data set every feature will have different types and scales of values some may be small some big numbers, Scaling does not impact the model but it helps to bring the features to 'comparable' scale , if we don't do scaling it will result in disproportionate coefficients , this helps in getting the coefficients to same scale.

Normalization

Instead of using the min() value , in this case, we will be using the average() value.

X(new) = X- X(mean) / (X(max) - X(min))

Is normalization formula

in normalization you arechanging the shape of the distribution of your data

Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k-nearest neighbours and artificial neural networks.

Standardization

In standardization, we calculate the z-value for each of the data points and replaces those with these values.

This will make sure that all the features are centred around the mean value with a standard deviation value of 1. This is the best to use if your feature is normally distributed like salary or age

Standardization is useful when your data has varying scales and the algorithm you are using does make assumptions about your data having a Gaussian distribution, such as linear regression, logistic regression, and linear discriminant analysis

X(new) = X- X(mean) / Sigma

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

Ans 5 : Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other

VIF = 1 / 1-R^2

Now if due to multicolleniartiy between features and strong correlation R^2 value is very high and almost 1 or 1 then VIF becomes 1/1-1 ie. 1/0 which is infinity we need to drop those features from model

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans 6

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically\Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s)

Q-Q plot can also be used to test distribution amongst 2 different datasets. For example, if dataset 1, the age variable has 200 records and dataset 2, the age variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be particularly helpful in machine learning, where we split data into train-validation-test to see if the distribution is indeed the same.