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?

Answer - The demand of bike is less in the month of spring when compared with other seasons. The demand bike increased in the year 2019 when compared with year 2018.

The demand is equally distributed among the weekdays

The demand is lower when the weather is bad

Data inferred Atemp and Temp are more or less the same

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

Answer - Assuming you have a column with 3 unique values dropping your first categorical variable is possible because if every other dummy column is 0, then this means your first value would have been 1. What you remove in redundancy.

For example a house is furnished semi-furnished or non furnished in the column called Furnished, we make use of drop_first to create dummy variables with only 2 columns (semi furnished and Furnished) if a house is either of these two then it is automatically assumed that it is not unfurnished hence an unfurnished value would be 0 for semi furnished and 0 for furnished as well hence we can remove the column called unfurnished.

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

Answer – Looking at the pair plot it is evident that Temp are the highest correlated to the target variable (cnt) - 0.63

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

we then performed the prediction on the training data using a regression model and found the y train pred value, subtracting the y train value to the y train pred value we found that the Errors are normally distributed here with mean 0. So everything seems to be fine

We also found the r_sqared value using this test analysis and noticed it was very close to the value from our final OLS model

We then created a y test pred and found that the r2_square value was almost similar to the values from the training data which suggested a training set of 70-30 with random state set to 42 percent was an ideal fit

R2 value for predictions on test data (0.7987479140295531) is almost same as R2 value of train data(0.7909321666618335). This is a good R-squared value, hence we can see our model is performing good even on unseen data (train data)

we lastly checked for homoscedasticity and found that the model passed the requirement.

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

weathersit_bad -1579.615781holiday -733.95825 and mnth_mar 476.054416 are the top three features contibuting to this model with the parameters.

General Subjective Questions

1. Explain the linear regression algorithm in detail

Answer – Linear regression is a machine learning algorithm that makes use of supervised learning. In linear regression we first find out how data is corelated with the target variables and then make use of a line graph (called as a regplot) when we see that there is a line passing through majority of the values that are plotted it becomes evident that there is a correlation between data and hence we make use of the concepts of linear regression. We then find multiple parameters that can be used to create a well defined linear regression model based on the data that we are provided with, we once we have a line drawn we calculate the distance between the line and the points that do not pass through the line we call this as R_sqare values.

Using ordinary least squares we make use of training data and test data by telling the system how much training data we want in a proportion to test data, this then creates training data with a random state

We also make use of Variable inflation factors calculation to find out how related and dependent are variables in our model hence describing multicollinearity. Anything above 5 is generally not prefered

We also check for P values in the system and verify if it is above 0.05, if it is then we try to remove them.

Lastly we check the for hemoscaatttered plotst hat proves that there are no constant trends in the error.

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

Answer - Anscombe Quartet can be described as a group of four identical data sets with simple descriptive statistics, but there are certain features in the database that trick a retrospective model when built. They have very different distribution and appear differently when arranged on scatter plots.

The data may seem the same when we use different formulas of statistic but in reality this data is very different and this is verified by a scattered plot

3. What is Pearson’s R?

Pearson's r is a numerical summary of the power of the linear relationship between variables. If the variable tends to go up and down together, the correlation coefficient will be positive. If a variable tends to go up and down against the lower values of one variation associated with the higher values of another, the correlation coefficient will be negative.

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

Scaling is a process to fit numbers in right, e.g. between zero and one, or one and a hundred. An example can be when we have to take the percentage of blood glucose levels you may write 0.12g rather than 12mg to have a neater looking number

It is ideally done in programming to determine a better understanding of data while corelating it to other parameters.

Normalization can be called standardization it is about the measurement of external 'normal' - local practice - such as subtracting the average value and separating with a standard deviation sample, e.g. so that your filtered data can be compared to standard accumulation. An example of this can be when the highest score in the class is 70 out of 100 then the marks of every student is normalized to consider 70 as the new 100%

The difference between the two can be considered as one where the values are not subtracted or added which is called as scaleing and the other where we perform a new calculation. In different places we tend to use different processes.

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

Answer when the correlation between two different indepentant variables is calculated we find that the VIF= Infinity, this takes places because the value of R2 tends to be 1 and by the formula of VIF we find that 1/1-R2 leads to 1/0 which gives us infinity, to fix this we need to drop on of the of variables from the parameters.

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Based on the slope of the graph is 45 and the plots lie on a Stright line we consider this as a similar distribution.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.