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? (3 marks)

The categorical variables that has impact are

1. Few months: The count is expected to go slightly down if its Jul but expected to go slightly up if its September

2. Whether it is a holiday or not: . For ex: the count is expected to go down if it is a holiday

3. Few Season such as Summer and Winter: the count is expected to go slightly up in Summer but more up during winter season.

4. Weather condition such as Light rain: This is going to impact negatively.

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

drop_first is used to remove multicollinearity. The all other generated dummies with value as 0 explains the first variable. If we do not remove first, there would be a variable which is being explained by other variables.

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

The highest correlation is the temperature and atemp. However atemp is being explained by temp only.

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

1. The pair plot were showing that there is a linear relationship between target and few variables.

2. Created a histplot to see if error terms are normally distributed with mean at 0

3. scatter plot was created for error terms and no pattern found for error terms and hence concluded that error terms are independent of each other.

4. VIF values are fine and thus there is no multicollinearity.

5. Since the test set has good and close R2 value, we can conclude that there is no overfitting.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top 3 features are Temperature, humidity and windspeed. Yr is also significant contributing factor, however, that is debatable in this scenario with the way the year is presented in data.

General Subjective

Questions 1. Explain the linear regression algorithm in detail. (4 marks)

The objective of linear regression is to find the best fit line on the dataset that can predict the future outcomes.

1. First we need to find whether the target variable has a linear relationship with the independent variables provided or not. Variables can be of 2 types, continuous and categorical. The linear relationship of continuous variables can be determined using scatter plots. The same can be identified for the categorical variables using box plots across different categories and analyse for variability.
2. Once few variables are identified to have relations, correlations between independent variables can be found out using heat map. The variables having good correlations, one of them can be dropped since they both explain the same variability.
3. The categorical variables has to be converted into columns and a derived matrix has to be created.
4. Post these observations, we need to split the data in train and test data in order to build the model on train data and the built model can be tested on test data.
5. The test data then need to be scaled. Many methods can be used for scaling, however, MinMaxScaling is the best known method for the reason they capture outliers. The fit and transformation need to be done in order to do scaling.
6. The next step is to do the model building and feature selection. This can be achieved in multiple ways.
    a. Create model with 1 feature and then Feature addition seeing the impact on R squared and adjusted r squared value.
    b. Create model with all the variables and then drop the variables based on High VIF, high p-values. Rebuild the stats model and check for improvement in variables and keep looping the same until we have variable having VIF < 5 and p-value < 0.05.
    c. The method in above step is called recursive feature elimination can be done using sklearn library.
7. Once we arrived at a satisfactory model using above methods, the model has to be tested on the test data. The objective of testing is to get R-squared value of y_test and y__test_pred to be very close of R-squared value of model build using train data.
8. The assumptions of linear regression can then be validated by:
    a. Create a hist plot of error terms to verify if there is a normal distribution of error terms.
    b. Create a scatterplot of error terms against y and there should be no pattern proving they are independent.

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

Anscombe quartet consist of four dataset with similar statistical summary but appears very different when a visualized using scatter plot. The statistical summary like mean, variance, R-squared value etc for all the four datasets is very similar but their graphical representation is very different. Thus it lays the importance of data visualization.

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

Pearson's R is the correlation coefficient between two variables denoting the relationship between two variables. The value of Pearson's R lies between -1 and +1. The 0 value denotes that there is no correlation between 2 variables. +1 denotes that if the value of one variable increases, the value of other variable also increases. -1 denotes that increase in value of one variable cause the decrease in other variable.

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

The scaling means to treat independent variable and to bring all independent variables on same scale.

The scaling is performed so that the dependent variables can have the coefficients which are relative. For ex: the two columns have high numbers but one is scaled down to be represented in lakhs but another is absolute number, the number having high digit representation has high coefficients and thus will make wrong assumptions.

Normalized scaling where min and max column values are used. The column values become between 0 and 1 including outliers, hence it is very good to treat outliers as well.

Standardized scaling is when mean and std deviation is used for scaling. However the distribution of column values are more or less same. The outliers might have great impact on values and thus not very recommended.

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

VIF is calculated by treating one of the independent variable as dependent variable on the other independent variables and then calculating the R-square of such model. The higher the R-square means higher VIF means that the chosen variable is more clearly explained by other variables.

The infinite VIF means that the R-square of such arrangement of independent variables becomes 1. This means the chosen variable is absolutely explained by other variables since there are no errors in the chose straight line.

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

A Q-Q plot is a quantile-quantile scatter plot which is used to plot the distribution of one dataset against the distribution of another dataset. If the points lies at 45 degrees line that means that both datasets belong to the same population.

There can be 2 uses of Q-Q plots in linear regression:

1. It can be used to plot the errors against theoretical normal distribution. If the scatter points showcase a straight line at 45 degrees, then our assumption of normal distribution of error terms is correct.
2. The distribution of training dataset can be plot against test dataset which confirms that both the dataset belong to the same population.