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: Categorical variables had an effect on dependent variables. Because dummy coding compares the mean of the dependent variable for each level of the categorical variable to the mean of the dependent variable at for the reference group, it makes sense with a nominal variable. However, it may not make as much sense to use a coding scheme that tests the linear effect of race.

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

Ans: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Ans: 'temp' and 'atemp' and 'registerd' and 'cnt' have the highest correlation as per the pair-plot.

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

Ans. 1. A Linear Relationship between the dependent variable and their predictors can be justified

using Pair Plots.

2. If error termsform a normal distribution, it validates the training model.

3. VIF (Variance Inflation Factor) value for all the feature should not be greater than 5.

4. P-values for all the features should be less than 0.05(assumed) significance level.

5. Error terms should be independent of each other .

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

Ans: According to my model top 3 features contributing significantly towards explaning the demand of the shared bikes are:

1.Tempreature

2.Year

3.Winter

4.September

For all these features value of coefficient is positive and VIF is less than 5.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Regression helps us to determine the strength of the Relationship between one dependent

Variable and one or more independent variables. ‘Linear’ in Linear Regression here doesn't mean a

perfect straight line but it means linearity in the parameters. Dependent Variable is that Variable on

which modal has to be build. Dependent variable is also known as the target variable. Independent

variable is known as the Predictor variable.

Some important things to note:

1. Target variable that has to be predicted must be a numerical/continues one.

2. Linear Regression comes under Supervised learning method because labels are present.

Formula: $Y= a + bX$

Types of Linear Regression:

1. Simple Linear Regression: Here predictor variable is one.

$y = B0 + B1*X + EB0$ – Intercept

2. Multiple Linear Regression: Here predictor variable can be more than one.

$Y = B0 + B1X1 + B2X2+ \ldots.. +BnXn +E$

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.
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.

3. What is Pearson’s R?

Ans: 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. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

$r = 1$ means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
$r = -1$ means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
$r = 0$ means there is no linear association
$r > 0 < 5$ means there is a weak association
$r > 5 < 8$ means there is a moderate association
$r > 8$ means there is a strong association

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

Ans: Scaling is a pre - processing stepwhich is taken to scale the value of the data within specified

range to improve the accuracy of the model. The scaling is optional in case of simple linear

regression while it must be compulsory for multiple linear regression modals.

The reason why we do scaling can be explained by an example: -

Let’s suppose, if an algorithm is not using the feature scaling method, then it canconsider the value

3000 meters to be greater than 5 km but that’s actually not true and, in this case, the algorithm will

give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes and

thus, tackle this issue.

Normalised scaling is also known as Min-Max scaling. Here scaling is done in such a way so that all

the values of all the variables lie between 0 to1. Minimum and maximum value of features are used

for scaling in normalization.

(X-Xmin)/(Xmax-Xmin)

StandardizedScaling is the type where Mean and standard deviation is used for scaling. In this mean

is 0 and sigma=1.

(X-Xmean)/std. deviation

Here, the difference between two the two techniques is first, Normalised scaling handles outliers

while standardized one can't handle. Second, normalizationis used when features are of different

scales and Standardization is used when we want to ensure zero mean and unit standard deviation.

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

Ans: Variance Inflation Factor(VIF) is the method that shows the relationship of one independent

variable with anotherindependent variable. For a particular variable, if VIF is high then it means it

has high association with other variable and vice versa.

When R=1, VIF = 1/(1-r^2) = Infinite, which means one independent variable is perfectly Correlated

with another variable. This leads to multicollinearity.

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

Ans: 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.
Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior