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)

From my analysis of the categorical variables, I drew following insights on their effect on the target:

- year: year 2019 have much visibly higher demand as compared to 2018, the demand is increasing over the years
- month: there's a pattern in months as well, summer months have higher demand as compared to winter months
- season: There's visible pattern in seasons, spring has very low demand, followed by winter and summer and falling have high demands
- weekday: weekday has no significant effect on target
- holiday: days with holidays have slightly lower demand as compared to non-holidays
- workingday: No significant effect on the target
- weather: There's 0 demand if there's heavy rain or thunderstorm, also clear weather sees high demand followed by mist and low demand for light rain

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

If there are n cardinal values in a categorical variable, it can be represented using n-1 dummy variables. It is because if the value of all the dummy variables is 0, it automatically infers it is the remaining category.

That is the purpose of drop_first =True, it removes the first dummy variable and creates n-1 dummy variables.

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

Looking at the scatter plot between numeric and target variable, we can see that temp and atemp columns have the highest correlation with the target variable and they are also highly correlated with each other.

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

We have validated the assumptions as follows:

- After building the model, we did the residual analysis to check for homoscedasticity. The residuals are normally distributed with mean of 0 with no clear patterns. Thus it passed the test
- We verified the vif factor for the final 15 features, all the features have vif value below 5. Hence, no multicollinearity
- In the final list of features, the continuous variable (atemp) do not have any outliers, have close to normal distribution and have linear relationship with the target.

Hence, I validated the assumptions of Linear Regression.

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

From my final model I can conclude that the following top features are contributing significantly towards predicting the shared bike demand:

a) Atemp: The feeling temperature is significantly contributing in the demand and is positively affecting the target
b) Year-season: The derived feature after combining year and season proves to be a very significant variable. The dummy features created from it (year-2019 season-summer, fall and winter) are highly contributing in the demand
c) Weathersit: The weather situation variable also comes out to be useful one. It's dummy variables light_rain and mist are contributing in the final model

General Subjective Questions

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

Linear regression is a supervised learning machine learning algorithm where we try to predict a continuous target variable by establishing a linear relationship with other independent variables.

We try to find the best fit line/ plane depending on the number of independent variables that most correctly gives the relationship with the target variable. The equation of a linear regression is given by:

$y = a_1x_1 + a_2x_2 + ... + a_nx_n + e$
Where y is the target variable, x1,x2… xn are the independent variables and a1,a2.. an are their corresponding coefficients and e is the error term.

We identify the best fit line by minimizing the Mean Squared Error (MSE) which is the mean of sum of squares of differences between the predicted and actual values of the target variable.

<u>Assumptions of linear regression</u>
Before applying linear regression algorithm on a data, there are some assumptions that must be satisfied:

- Linear relationship: There must be a linear relationship between the independent and the target variable else we can't fit a regression line.
- Normality: The data must be normally distributed. If there's a skewness in the data, the outliers will affect the regression line.
- Multicollinearity: The independent variables must not be correlated to each other.
- Homoscedasticity: The residual terms must be normally distributed around the mean and there should not be any pattern in it.

- No autocorrelation: The residuals of different observations should be independent of each other

Once all the assumptions are satisfied, we can apply linear regression algorithm on the data.

After the linear regression model is fit, we can test it's performance/ accuracy looking at the R-squared value.

R-squared value basically tells us how much variance in y can be explained using the independent variables. It measures how well our model is performing against a baseline model if all the predictions were made as mean of y. It's value varies between 0-1, the higher the r-squared value, the better the model.

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

Anscombe's quartet is a group of 4 datasets which are almost identical in terms of simple descriptive statistics including mean and standard deviation yet have very different distributions and appear differently when plotted.

The four datasets can be described as:

a) Dataset 1: Data fitting the linear regression line well
b) Dataset 2: Non-linear data that does not fit the linear regression well
c) Dataset 3: Data with outliers that cannot be handled by the linear regression model
d) Dataset 4: Data with outliers and clustered values that cannot be handled by linear regression model

All these four datasets, though when fed into the linear regression model produce very similar regression lines.
It was built to illustrate the importance of data visualization to identify different anomalies present in the data and how any regression algorithm can be easily fooled.

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

Pearson's R is a correlation coefficient to identify how strongly two variables are linearly related to each other. The value of Pearson's R ranges between -1 to 1 where,

- 1 refers to strong negative correlation, i.e. when one variable increases, other decreases and vice versa

0 refers to no correlation at all

1 refers to strong positive correlation, i.e. when one variable increases, other increases as well and vice versa

One major limitation of the Pearson's R is that it's only useful to identify linear relationships in the data and cannot identify monotonic or nonparametric correlations between the variables.

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

Feature scaling is a method of standardizing the range of independent variables in the data.

The different continuous variables in the data might have different order of magnitude and can cause issues in the algorithms where distance between the values is taken into consideration. To solve this problem, scaling is performed to bring all the variables within same level of magnitude.

Two major ways of scaling is Normalization and Standardization.

1) Normalization: It scales down the variables and bring them in the range of 0-1.
   Formula to normalize: $x_{scaled} = (x- x_{min})/x_{max}- x_{min}$

   Normalization removes the effect of outliers in the data and hence loses some information.

2) Standardization: It scales down the variable such that the mean of variable is 0 and standard deviation is 1
   Formula to standardize: $x_{scaled} =(x- x_{mean})/$std deviation

   It keeps the distribution of feature intact, hence no value is lost.

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

(3 marks)

VIF factor gives us the value of how an independent variable is correlated with a group of other independent variables and how much variance in the variable can be explained using other independent variables.

The VIF value for a feature i can be given by:

$VIF_i = 1/(1- Rsqaured_i)$

Where $Rsquared_i$ gives the value of how much variance in i can be explained using other independent variables.

The Value of VIF can be infinite when the R-squared value is 1, meaning if 100% of the variance in a feature can be explained using other features.

In this case, the variable i should be removed.

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

(3 marks)

Q-Q plot (quantile-quantile plot) is a graphical method of comparing two distributions by plotting their quantiles against each other. Q-Q plots are used to find the type of distribution for a random variable whether it be a gaussian, uniform or exponential distribution, etc. We can tell the type of distribution using the power of the Q-Q plot just by looking at the plot.

It is very useful in linear regression to test it's assumption of normality. It helps to determine very easily if a feature is normally distributed or not. To test the normality, we plot a standard normal distribution having mean =0 and standard deviation =1 on the x-axis and the ordered values of the feature we want to test on the y-axis.

If the data points at the ends of the curve fall perfectly on a straight line, then the feature is normally distributed otherwise it is not.

Also, q-q plots helps to identify the skewness in the data or even it is left -skewed or right-skewed.

We need enough data points to use the q-q plots else the results might not be significant.