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?

Based on the analysis categorical variable year (2019) has positive on the dependent variable. Holiday, spring and Light_Snow_Rain_Thunderstorm_Scattered_clouds_Light_Rain_Scattered_clouds has negative effect.

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

Drop_first drops the one of the dummy variables after creation as the dropped variable can be inferred from the remaining variables in the below example of the columns can be dropped.

As the type of furnishing can be identified with just the last two columns where —

- 00 will correspond to furnished
- 01 will correspond to unfurnished
- 10 will correspond to semi-furnished

	furnished	semi-furnished	unfurnished
0	1	0	0
1	1	0	0
2	0	1	0
3	1	0	0
4	1	0	0

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

Temp variable has highest correlation. Also season and month also show some correlation.

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

- Zero mean, independent, Normally distributed error terms that have constant variance

- p-values for variables are in acceptable range

- R-squared and adj R-Squared are close

- Prob F- Statistic value is in the acceptable range

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

Temp, weathersit and year are most significant features

General Subjective Questions

1. Explain the linear regression algorithm in detail

Ans. Linear regression is a supervised machine learning algorithm that predicts the outcome of an event based on independent variable data points. It is a statistical method used in data science and machine learning for predictive analysis.

Linear regression assumes a linear relationship between the features and the target variable. The model learns the coefficents that best fit the data and can make predictions for new inputs.

Linear regression can be used to predict continuous or numeric variables. When independent features is one, then its known as Univariate Linear regression and in case if more than one feature its known as Multivariate linear regression.

2. Explain the Anscombe's quartet in detail

Ans. Anscombe's quartet is used to illustrate the imporatnce of exporatory data analysis and the drawbacks depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outlineners, and other crucial details that might not be obvious from the summary statstics alone.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statstics, yet have very different distributions and appear veary different when graphed. The pecularities in the dataset can fool the regression model.

It was constructed by statistician Francis Anscombe in 1973 to illustrate the imporatnce of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

3. What is Pearson’s R?

Pearson's correlation coefficient, also known as Pearson's r, is a statistic that measures the linear relationship between two variables. It is the most common way of measuring a linear correlation.

Pearson's r is denoted as "r" and ranges from -1 to +1. The values indicate the following

- -1: A perfect negative linear relationship
- +1: A perfect positive linear relationship
- 0: No linear relationship
- Greater than 0: A positive association
- -1: Data is perfectly linear with a negative slope
- 1: Data is perfectly linear with a positive slope

Pearson's r can also be referred to as:

- Pearson product-moment correlation coefficient (PPMCC)
- Bivariate correlation

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

Feature scaling refers to putting the feature values into the same range. In regression, it is often recommended to scale the features so that the predictors have a mean of 0. This makes it easier to interpret the intercept term as the expected value of Y when the predictor values are set to their means.

The two most discussed scaling methods are Normalization and Standardization. Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

t seems that a some variables are able to create perfect multiple regressions on other variables (which would explain why all the VIF are infinity).

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

A Q-Q plot, or quantile-quantile plot, is a visual tool that compares the quantiles of two data sets. A quantile is the fraction of points that fall below a given value.

Q-Q plots can be used to:

- Determine if a dataset follows a specific probability distribution
- Determine if two samples of data came from the same population
- Assess the similarity between the distribution of one numeric variable and a normal distribution