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)

Ans: The categorical variables that I have used in my projects are Season, Year, Month, Holiday, Weekday, Workingday, Weathersit. The categorical variables holds correlation on the dependent variables as seen from the different plots. As the categorical variables are changing it is having an impact on the dependent variable parameters like r squared when measured through OLS(ordinary least squared method). To find the best fit value we have to keep adding and removing the categorical variables to get the best fit model on which we can predict.

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

It helps in reducing the extra column created during the dummy variable creation. In my case while I was working on dummy variable for season, whethersit, weekdays, month. I was basically categorizing there values into different columns to get a clear understanding of the variables to use.

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

Count has the highest correlation with temp, atemp, Year

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. I validated the assumptions based on the p-value and the r-squared value while training my dataset on the train model. The r-squared was mapped for all the variables then one by one the under-utilized variables were removed and the best fir were calculated with the help of Vif value and the p-value. The imp scenario of p-value and vif which is used it Lower p-value remove it. Higher Vif value remove it.
6. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

1. We can see the demand for bikes depends mainly on below variables:
2. year, holiday, Spring, Mist Cloudy, Light rain Light snow_Thunderstorm,3 ,5 ,6, 8, 9, Sunday, 7, 10
3. Demands increases in the month of 3, 5, 6, 8 ,9, 7, 10 and year
4. Demand decreases if it is holiday, Spring, Light rain_Light snow Thunderstorm, Mist_cloudy, Sunday

General Subjective Questions

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

- Linear regression algorithm can be defined as:
- A model that performs regression task and based on supervised learning.
- It is mostly used for finding relationship between variables and forecasting.
- It performs the operation of predicting a dependent variable value (y) based on independent variable (x).
- Regression shows line or curve that passes through all the data points on a target predictor

graph.

- To define the algorithm we import some imp libraries like numpy, panda, matplotlib, sklearn
- Define the dataset plot the datapoint.
- Then we perform the operation on the variables by assigning the suitable parameter and finding the r squared and p value for each variable.
- We also obtain a dummy variable from a categorized data set value.
- We then plot the various plots like box plot or correlation diagram to see which variable is more correlated with each other.
- Add the variable or remove the variables to calculate the vif values and obtain the correct graph.
- We use sklearn library to calculate the imp parameters and plot them to see the final curve if its centered around zero to find the parameters that are responsible to make the model a correct fit.

2. Explain the Anscombe's quartet in detail. (3 marks)
   - It Explains four dataset that have nearly identical simple descriptive statistics yet have very different distribution and appear very different when graphed.
   - Each dataset consist of 11 (x,y) point.
   - The Anscombe quartet basically tells us about the same mean, variance and standard deviation of a dataset but there graphs are dofferent.
   - The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship and the inadequacy of basic statistics property for describing the realistics dataset.
3. What is Pearson's R? (3 marks)
   - It is a measure of linear correlation between two sets of data.
   - It is a ratio between the covariance of two variable and the product of their standard deviation.
   - Pearson correlation coefficient when applied to the population is commonly represented by the Greek letter rho and may be referred as population correlation or the population pearson correlation coefficient
   - The correlation coefficient ranges from -1 to 1
   - An absolute value of 1 says the relationship between x and y perfect.
   - A value of 0 implies there is no linear dependency between variables.
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   - It is a step of data pre processing which is applied to the independent variable to normalize the data within a particular range . It also helps in the speeding of the calculation.
   - Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
   - It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared.
   - Normalization/Min-Max Scaling:
   - It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
   - Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

- If all the independent variable are orthogonal to each other then VIF= 1.0 If there is perfect correlation then VIF= infinity. A large value of vif indicate that there is a correlation between variables.

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

- Quantile – Quantile plot is a graphical tool to help assess if a set of data plausibly came from theoretical distribution such as normal, exponential, or uniform 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.