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)

Answer:

1. Spring season has the least demand for the bikes. Whereas, fall has the highest demand for the bikes.
2. September has the highest demand of bikes. And, demand decreases quickly as the winter approaches. Once the winter is done the demand increases gradually.
3. The demand of bike stays almost similar during weekdays.
4. The bike demand is not impacted by weekday or weekend.
5. The demand is higher when there is clear weather.

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

Answer: It helps in reducing the extra column in the data set, which gets generated during dummy variable creation. If there are n parameters the total dummy variables will be create is n-1. This will reduce the correlation between dummy variables.

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

Answer: Temp has highest correlation with the target variable cnt.

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

Answer:

Assumption 1 - there is a linear relation between independent variable and dependent variable. This assumption was validated by plotting scatter plot between dependent variable cnt and independent variables like temp, humidity, windspeed.

Assumption 2 - Error terms are normally distributed. This assumption was validated by plotting the distplot for the residual of train and predicted variables. The graphs plotted shows normal distribution and the mean of residual will come 0 or close to 0.

Assumption 3 - the error terms should not be dependent on one another. This was verified by plotting the scatter plot on train and residual and there is no visible pattern in the plot.

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

Answer:

1. Temperature is the most contributing factor with coefficient value of 0.5477.
2. Storm is the second most impacting factor with coefficient value of -0.2884.
3. Windspeed is the third most impacting factor with coefficient value of -0.1542.

Yr which has coefficient of 0.2330 is not considered because it is year detail whether the data is of 2018 to 2019. I don't believe it is any kind of factor which will have any impact on bike hiring.

General Subjective Questions

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

Answer: It is the process of estimating relationship among variables. It explains how the value of dependent variable changes with the change in the value of predictors. It is used for forecasting and

predictions. Linear regression shows correlation between variables. It is a form of parametric regression. In parametric regression, it is assumed the sample data is coming from the population which follows a particular probability distribution on a fixed set of parameters. Regression predicts on the basis of continuous variable datas. Output variable to be predicted is continuous / numeric variable. It is a supervised method of predicting. The linear regression model provide a sloped straight line representing the relationship between variables.
Mathematically it can be represented as **Y = mX + C + e**.
Y is dependent variable, X is independent variable, m is linear regression coefficient, C is intercept of the line and e is random error. Linear regression can be further divided into Simple Linear Regression and Multiple Linear Regression.
Simple regression is used when single independent variable is used to predict dependent variable. Multiple regression is used when there are more than one independent variables to predict dependent variable.

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

Answer: Anscombe's is a collection of four data sets that have same identical simple descriptive statistics, but they appear very different when they are plotted on scatter plots. This tells us about the importance of visualising the data set before applying various algorithms. The data distribution must be plotted to see the distribution of the samples that can help you identify outliers present in the data. For this type of cases only linear regression can be considered to handle these type of datas.

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

Answer: The Pearson's R is the most common way of measuring the linear correlation. It is a descriptive statistic meaning that it summarises the characteristic of the dataset. It describes the strength and direction of the linear relationship between two variables. It also tells whether the slope of the line of negative or positive. It is good to choose Pearson correlation when the variables are normally distributed, when the relationship is linear and the data has not outliers.

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

Answer: It is the process of modifying and bringing the dataset to one scale. The range dataset of each variables could in hundreds or thousands or millions. In that case the coefficients of independent variables will have huge difference. This could also produce the wrong predictions as well. It will bring the ease of interpretation when the variables are brought to one scale.

Normalisaiton	Standardisation
This process converts or compresses the data between 0 and 1.	This process converts the data so that mean becomes 0 and standard deviation becomes 1.
Normalisation handles outliers, because it has to bring the whole data between 0 and 1.	There is nothing specific to outliers in standardisation.
Formula: (X - Xmin) / (Xmax -Xmin)	Formula: (X - Xmean) / sigma
Normalisation do not have any effect on dummy variable as the dummy variables are already in 0 or 1.	Standardisation could have impact on dummy variables. It could scale dummy variable in such a way that the mean should become 0 and standard deviation becomes 1. So this will clearly distort the dummy variables because some of the variables will become negative.

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

Answer: If there is a perfect correlation of an independent variable with other variables in that particular dataset then that independent will have infinite VIF. Infinite VIF means, that variable can be perfectly predicted by other variables in the model. In this can we need to drop one of the independent variable and keep on dropping the variables one after the other which are causing complete multicollinearity.

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

Answer: It is a graphical technique of determining if two datasets come from populations with a common distribution. It is created by plotting two sets of quantiles against one another. If both sets of quantiles came from same distribution, we could see the point forming a line that's almost straight. It is just a visual check. It allows to have a visual confirmation on our assumptions. Q-Q plots take sample data, sort it in ascending order and then plot them verses quantiles calculated from a theoretical distribution.