Name: Praveen Rathi
Email: rathipra@gmail.com

Submission of 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:

Following inferences are seen in categorical variables:

Season variable: Spring season has lowest median, max counts that means spring season customers do not prefer to rent bike.

Year ("yr") variable: 2019 year has better response than 2018, it is due to increase in popularity of the company.

Holiday: When it is holiday response is lower as the median and max value of counts are lower than when it is not holiday, its due to that on working day customers commute to their offices, or working areas

Weathersit: Its obvious that weather is clear than customer response is highest, than it falls as the weather becomes worse or worst

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

Answer:

Dummy variables are created using n-1 approach, where total number distinct values in a categorical variable are n and we create dummy variables n-1. For an example season variable has four category i.e. season (1:spring, 2:summer, 3:fall, 4:winter) if summer, fall, winter has 0 value then it is spring season, hence we need not to mention spring season in a separate variable, all other three are if 0 then its should be spring only. Below table explains the season dummy variables:

Season	Summer (Dummy)	Fall (Dummy)	Winter (Dummy)
Spring	0	0	0
Summer	1	0	0
Fall	0	1	0
Winter	0	0	1

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

Answer:

After visualising the pair plots of all numerical variables *"temp"* variable is very well in correlation with target variable.

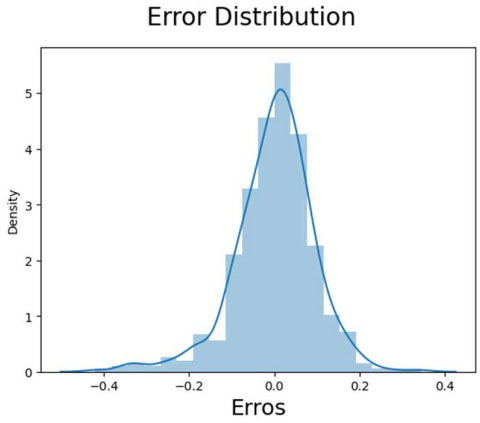
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

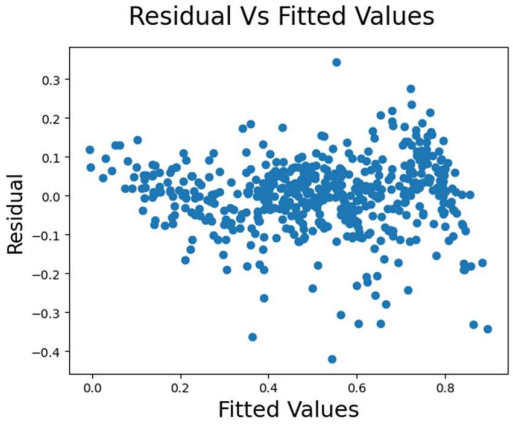
The assumptions of Linear Regression after building the model are validated as following:

Linear relationship: We have visualised through pair plots and heatmap that there is linear relationship found with target variable.
Multicollinearity: This is important check as predictors should not be corelated to each other as it will tend to biasness in predictions, and model will unnecessarily have more variables. Through VIF values we checked and all predictors having VIF values less than 5, hence there is no multicollinearity among all variables in the model.

Error terms are normally distributed: We performed residual analysis to validate that error terms are normally distributed or not, and we plotted a histogram and found that it is normally distributed as below graph explains:



Error terms are independent of each other and have constant variance (homoscedasticity): We plotted the fitted values with residual values to check whether error terms have constant variance or not as per below scatter plot, and it shows that there is no such pattern found and we can say that it is not heteroscedastic in nature. Also there seems no relationship between residual and fitted values, hence error terms are 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? (2 marks)
Answer:
Top 3 features based on final model and sort by coefficient values are as below:

Top Features	Coefficient Value

Temp	0.465
light snow	-0.279
yr	0.234

General Subjective Questions

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

Answer:

Equation: Linear regression algorithm is a supervised machine learning technique to predict y value based on single or multiple X predictors. Its equation is y = mX + b , where y is the dependent variable and X is an independent variable (predictor) and m is the slop of linear regression and b is the inception i.e. at the zero point of X what is the value of y.

Relationship: Linear regression algorithm is suitable only when there is a linear relationship between y as dependent variable or target and X as independent variables or predictors (X can be multiple variables), here the basic thing is that there should be a linear relationship, however causation is not explained by linear relationship alone. Linear relationship can be positive or negative.

Types: There are two types of linear regression, simple and multi linear regression. In simple linear regression there is only one predictor x and one target y but in multi linear regression there are more than one predictors X and one target y.

Assumptions: There are some assumptions for a linear regression to be hold validated.

A.) There should be a linear relationship between X and y variables

B.) There is very little or no multicollinearity between X predictors/variables.

C.) Residual (errors terms) i.e. difference between independent and dependent variables remain same for all independent variables that means there should be a constant variance, no upward or downward pattern that may lead to unbalanced scatter of residuals

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

Answer:

Anscombe's quartet is about visualisation of data set before attempting to build the linear model. There are four dataset of X and Y in Anscombe's quartet and all have same mean, standard deviation, correlation coefficient and their linear regression equation is also same, but still when we plot them all are different in terms of linear relationship. Hence we can be fooled by just merely seeing the statistical data of a dataset. Only after visualising the features we can if the feature is capable of linear relationship of not. This is always a good practice to visualise first.

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

Answer:

Pearson's R is a important parameter to check a strength of linear relationship between two variables. It is always between -1 to 1 where -1 is perfect negative relation and 1 is perfect positive relation and any other value between -1 to 1 is likely strength of relation between two variables.

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

Answer:

A.) Scaling a process of normalizing the feature with in a range. In other words we always get the data set having various numbers in various ranges, and we normalizes these features value to a particular range and that is called scaling.

B.) It is always good to scale the features to a particular range so that our model build is more correct and also faster in terms of speed.

C.) Normalized scaling is to scale the feature between range 0 to 1 by calculating every data point to min and max value proportion. Whereas in standardized scaling we scale all data points to have mean zero and standard deviation to one always.

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

Answer:

VIF is useful to find the multicollinearity between independent variables and when a variable is perfectly corelated to any other then VIF will be infinite provided the model R2 is perfect 1 as VIF is calculated by using formula VIF = 1/1 – R2 , hence when R2 reaches to perfect one then VIF becomes infinite.

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

Answer:

Q-Q plot is quantile to quantile, we plot two variables quantile to visualise if they are same. Quantile is obtained after sorting the values. Sometimes we train our model and evaluate it on test data but what if the variable values are almost same as train data. After plotting the two distributions if they remains on a linear line then it means they are either from same dataset or they are random but similar. So Q-Q plot helps to compare our sample distribution of a variable against any other possible distribution visually.