Pankaj Ligade

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)**

Count is maximum for **fall season**.

Count is maximum **in weathersit 1 (Clear, Few clouds, Partly cloudy, Partly cloudy)** and is minimum **in weathersit 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds**, Light Rain + Scattered clouds)

Counts in middle **months** of the year is higher than in the beginning and the end.

Increase in number of counts in **year** 2019 from 2018.

Count median is almost same for all **days**

Count is higher for **working day**

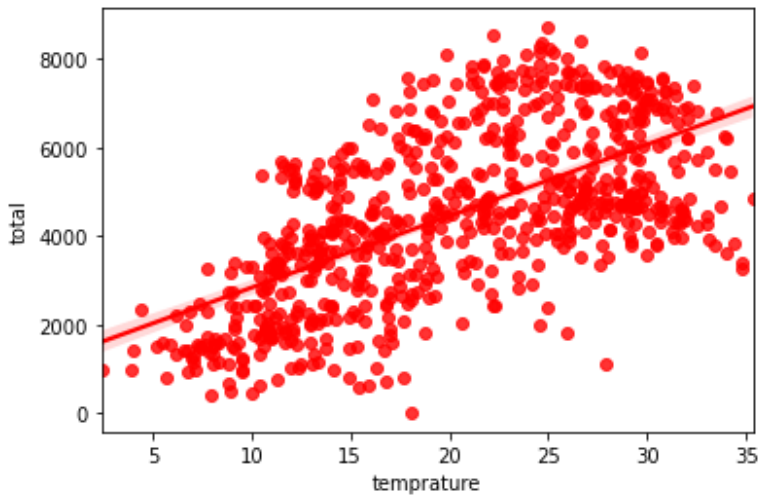
Count is lower on **holidays**.

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

Drop_first = true is important because it helps in reducing the extra column created during dummy variable creation. it reduces the correlations created among the dummy variables. it also reduces multicollinearity in the variables otherwise this would create unwanted and bulky data.

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

The feature **temp** has highest correlation. It is very well linearly related with the target cnt.



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

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

Top 3 features:

1. Temperature 2. Weather category 3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) 3. Mid-yearmonths. (May to September)

General Subjective Questions

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

Linear Regression algorithm is the supervised machine learning technique which is part of regression analysis. It's one of the simplest regression analysis techniques which is used to interpret how the independent input variables have influence over the dependent target variable

By the name linear, it means the variables on x and y axis are linearly correlated.

There are two types of linear regression mainly

• Simple Linear Regression: Is the type of linear regression where there is one target variable and one independent variable. Is represented by y = mx + c m: where m is slope c: is the intercept

• Multiple Linear Regression: Is the type of linear regression where there is one target variable and many independent variables. Is represented by $y = \beta0 + \beta1X1 + \beta2X2 + \beta3X3$ ----- βnXn. β0: is the intercept β1: is co-efficient of X1 β2: is co-efficient of X2 β3: is co-efficient of X3 and so on

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

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built.

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

Pearson's R which is also referred as Pearson correlation coefficient in statistics. It is the measure of linear correlation between two sets of data.

It is the ratio between covariance between two variables and product of their standard deviation which lies between the values -1 to +1 Say:

• If r is between 0 and 1 then the data is perfectly linear with positive slope.

• If r is between -1 and 0 then the data is perfectly linear with negative slope.

• If r =0 then there is no linear association between two sets of data

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

Scaling is a method to normalize the range of independent variables. It is performed to bring all the independent variables on a same scale in regression. If scaling is not done, then regression algorithm will consider greater values as higher and smaller values as lower values.

The difference between normalized and standardized scaling is

• Normalized Scaling: It brings all the data in range of 0 and 1. o For this we use sklearn.preprocessing.MinMaxScaler in python

• MinMax Scaling: x = x – min(x) / max(x) – min(x).

• Standardized Scaling: this is a method which replaces values by their Z scores. That is, it brings all of data into standard normal distribution which has mean zero and standard deviation as one.

• For standardized we use sklearn.preprocessing.scale in python. o Standardization: x = x -mean(x)/ sd(x)

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

If there is a perfect correlation, then VIF is infinity.in this case RSquare is one and

• VIF = 1 / 1 – Rsquare

• which will be 1/ 1-1 = 1/0 = infinity.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution

The purpose of Q Q plots is to find out 1. If two sets of data come from the same distribution. 2. To check if the two data sets come from a common distribution, 3. To check if the points will fall on that reference line.