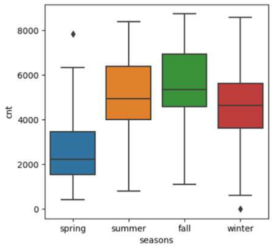
Assignment-based Subjective Question

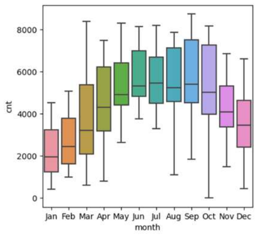
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer 1:

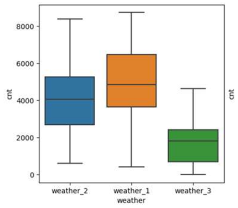
Analysing the categorical variables against the target variable ‘cnt’ we see that for some seasons like Summer & fall the demand has increased & during spring there is substantial decrease in demand of bikes .



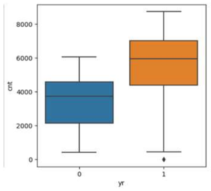
The ‘month’ variables show that there high demands in July , Sept where as very low demand in Jan ,Feb



The ‘weathersit ‘ variable shows there is no demand for bikes where value is 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog and there is a substantial decrease in demand where weathersit is 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds .



The ‘yr’ column shows demand has increased in 2019 .



Other categorical variables like ‘workingday’ or ‘day’ , don’t show much effect on target variable .

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

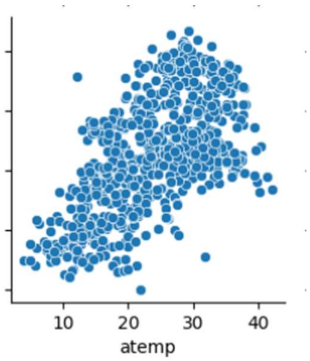
Answer 2:

If we do not use drop_first = True, then n dummy variables will be created for n predictors .These variales are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap .

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

Answer 3:

Looking at the pair plot of the numerical variables the highest co relation is visible between the target variable cnt and atemp .



Note : atemp & temp has high high co relation , hence I have dropped temp & not considered for model building . In pair plot co relation between cnt & atemp and cnt & tepm show similar pattern .

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

Answer 4:

For validating the model built on the training set we firstly do the Residual analysis . We plot a histogram plot with the error terms . The expected distribution should be a Normal distribution .Then will do an evaluation on the TEST set of data . We will do prediction of the target variable of the test dataset & compute the R2 score . the R2 score for the train dataset & test dataset should be very similar (variance of 5% is acceptable)

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

Answer 5:

The top 3 features contributing to the demand of shared bikes are –

- atemp (feeling temperature in Celsius)

- yr (year)
- weathersit – 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

General Subjective Questions

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

Answer 1:

Linear regression is that which explains the relationship between a dependent variable and one or more independent variable/s (Simple Linear regression and Multiple linear regression) .

The standard equation that represents simple linear regression is $Y = β_0 + β_1X$ and the equation that represents multiple linear regression is $Y = β_0 + β_1X1 + β2X2+$ ……..+ $βnXn+c$

The Linear regression algorithm will be as below :

- First we read , understand and analyse the data .
- 2nd we prepare the data for modelling – . Drop features that are least relevant for prediction – having high p-values . Check for correlations (heatmap) , multicollinearity (high VIF) of data and drop redundant features . Handle categorical variables – add dummies , binary values
- 3rd we train the model – i.e. try to fit the line . Least square estimation method is used to minimize the residuals .
- 4th Residual analysis is done by plotting the error terms . The values are expected to show a Normal distribution .
- Finally step is to test the significance or usefulness of the model . Evaluation is done on test set & final prediction is done .

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

Answer 2:

Anscombe's quartet was created by statistan Fransis Anscombe in 1973 to demonstrate the importance of graphing data when analysing it . It also explained the effects of outliers and other influential observations in statistical properties . Anscombe's quartet comprises of 4 datasets that have nearly identical statistical values (example – Mean , variance , correlation , linear regression line , R2 etc ..) but have very different distribution and appears to be very different when graphed .

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

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

Answer 3:

Pearson correlation coefficient is also known as Pearson's R. It was developed by Karl Pearson .

Pearson’s R is also known as PPMCC , Bivariate corelation coefficient or correlation coefficient . It is a measure of linear co relations between two sets of data. It is the ratio between the co variance of two variables and the product of their standard deviations . It is a normalized measurement of the covariance and the value is always between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations.

For a given pair of random variables (X,Y) the formula of

Pearson’s R = Covariance of X,Y /(Standard deviation of X)x(Standard Deviation of Y)

A key mathematical property of the Pearson correlation coefficient is that it is invariant under separate changes in location and scale in the two variables. Meaning - , *X* can be written as *a* + *bX* and *Y* as *c* + *dY*, where *a*, *b*, *c*, and *d* are constants with *b*, *d* > 0, without changing the correlation coefficient.

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

Answer 4 :

Scaling is a data pre-processing step which is applied to independent variables to normalize the data within a particular range . Normally the data collected is using different scales & units .Unless scaling is done the model will be incorrect .

Scaling is performed for 2 reasons :

a) Ease of interpretation
b) Faster convergence for gradient descent methods

Standardization and normalization are two most common techniques for feature scaling. In Normalization scaling the variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data . Standardization is about transforming the feature values to fall around mean as 0 with standard deviation as 1.

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

Answer 5

VIF is full form of Variance Inflation Factor . High VIF means multicollinearity .If there is perfect corelation between 2 independent variables the VIF value is infinity . The VIF value can range between 1 to infinity .

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

Answer 6

In statistics, a Q–Q plot i.e. quantile-quantile plot is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.

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. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions.

A Q–Q plot is generally more diagnostic than comparing the samples' histograms, but is less widely known. Q–Q plots are commonly used to compare a data set to a theoretical model. Q–Q plots are often arced, or "S" shaped, indicating that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other. This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic.