Assignment-based Subjective Questions

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

I have used bivariate analysis plots and box plots to analyze the categorical variables against the target variable. Following plots are obtained during the analysis.

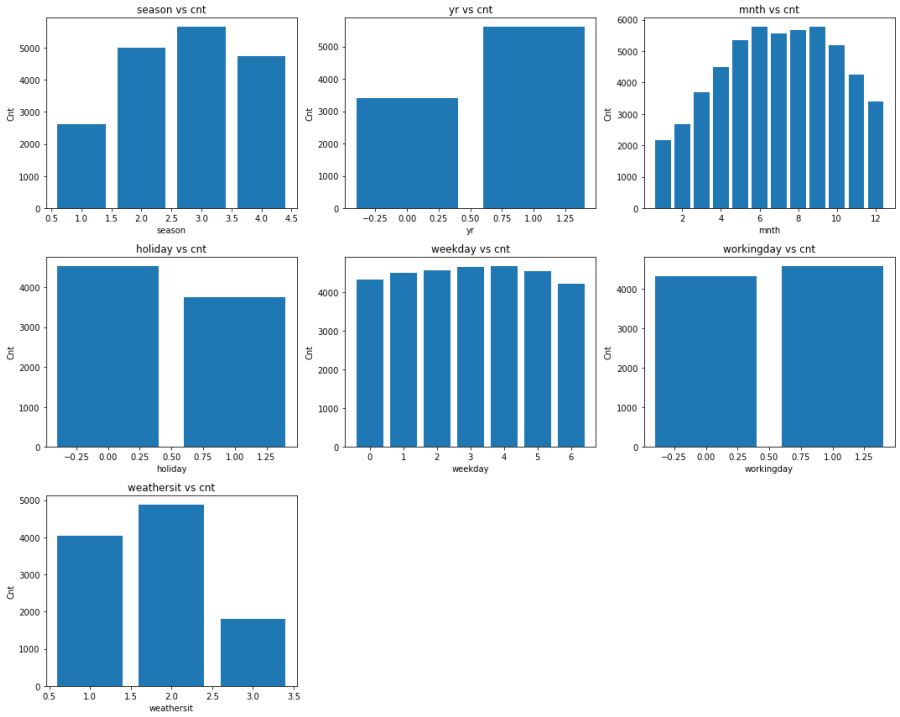

Fig. (1): Bivariate analysis showing the plots of categorical variable vs. the target variable.

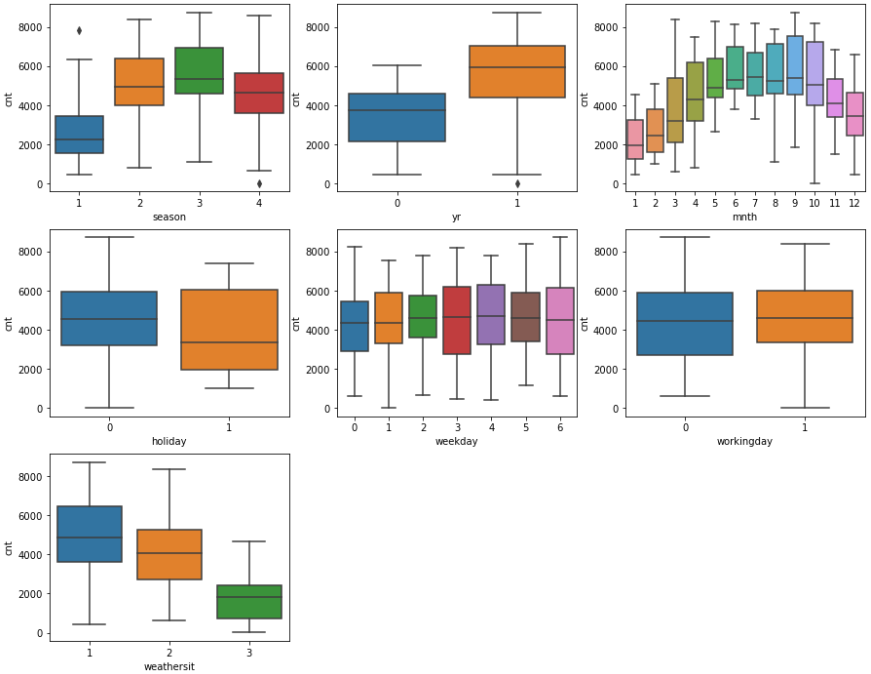

Fig. (2): Box plots of categorical variables against the target variable.

From the plots of categorical variables against the target variable, we can observe a definite trend in the average demand across few categorical variables. We can also observe that

'Workingday' & 'Weekday' variables are insignificant and may be dropped from the further analysis. Also, we can see from the Heat Map that Humidity is poorly corelated with the target variable, so it can also be dropped. Following the rule of Multi Collinearity we shall also drop 'atemp' from calculations.

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

drop_first=True is important to use because it helps in reducing the extra column created during the dummy variable creation. Hence it reduces the correlations created among the dummy variables. Hence if we have categorical variable with **n-levels**, then we will need to use only **(n-1)** columns to represent the dummy variables.

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

Temperature (temp) & feeling temperature (atemp) have the highest correlation with the target variable.

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

I have validated the assumptions by performing the residual analysis and check for homoscedasticity. In the residual analysis we observe that the errors are normally distributed and have zero mean. In homoscedasticity check we observe that the distribution of errors has constant variance. This is shown in the following figures Fig. (3a & 3b). And in the pair plot we observed the linear relationship with target variable. Hence there is no multicollinearity between the predictor variables.

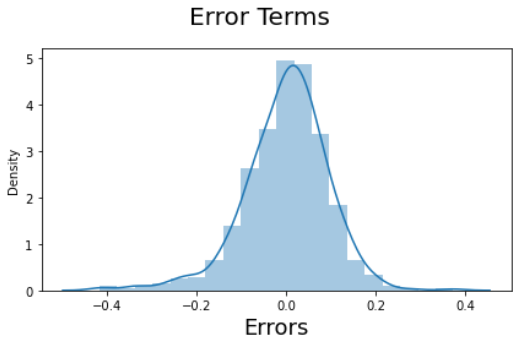

Fig. (3a): Normal distribution of errors.

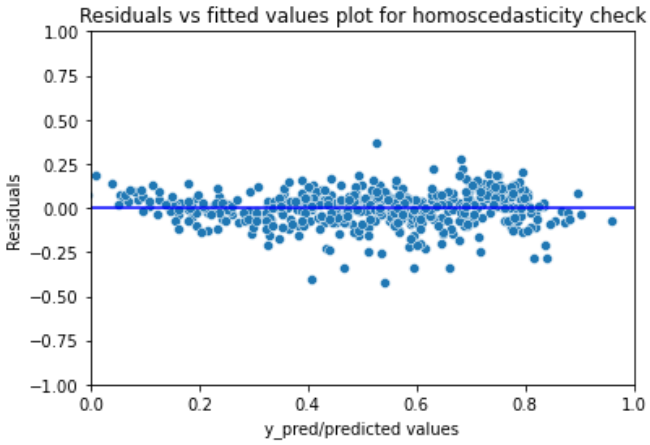

Fig. (3b): Homoscedasticity check

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

Following 3 features contribute significantly towards explaining the demand of the shared bikes:

(i) Months viz. 8^{th} and 9^{th} months.
(ii) Seasons viz. Summer and winter.
(iii) Temperature

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

There are two types of linear regression:

(i) Simple Linear Regression
(ii) Multiple Linear Regression

Following are the components of linear regression:

(i) Regression coefficient (β_1)
(ii) Intercept (β_0)
(iii) Residuals or Error terms

Objective of linear regression:

The objective of linear regression is to perform predictive analytics and it is done by making the machine learn the science of generating a trained (best fitted) line that will very well generalize how new and unknown data (test set or new dataset) will be evaluated, and how the fitted line will be able to accurately estimate new or unknown datasets.

Steps to be followed in linear regression algorithm are:

(i) Reading & understanding the data
(ii) Visualizing the data (Exploratory Data Analysis – EDA)
(iii) Data preparation
(iv) Splitting the data into training set & test set (usually in 70:30 or 80:20 ratio)
(v) Building a linear model
(vi) Residual analysis of the trained data
(vii) Making predictions using the final model & evaluation
(viii) Final conclusion or report.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is often used to illustrate the

importance of looking at a set of data graphically and not only relying on basic statistic properties. It comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. This is illustrated in the Fig. (4). 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.

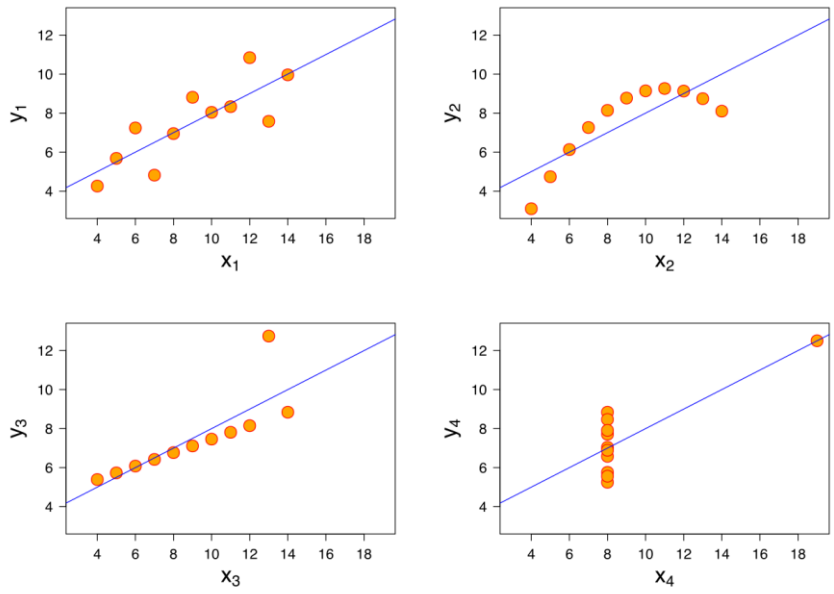

Fig. (4) All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.

3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

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, etc.

Normalized scaling:

- It brings all of the data in the range of 0 and 1.
- **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

Standardized 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?

VIF is given by the following formula,

$$VIF = \frac{1}{1 - R^2}$$

In the case of perfect correlation between two independent variables the value of $R^2 = 1$. This leads to 0 in the denominator of the above formula. Hence making the value of VIF as infinity.

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

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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

******

Answers submitted by

Sanjay M. Belgaonkar
sanjaymb@ieee.org