BIKE SHARING ASSIGNMENT

BY RAHUL BATRA

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

Question 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

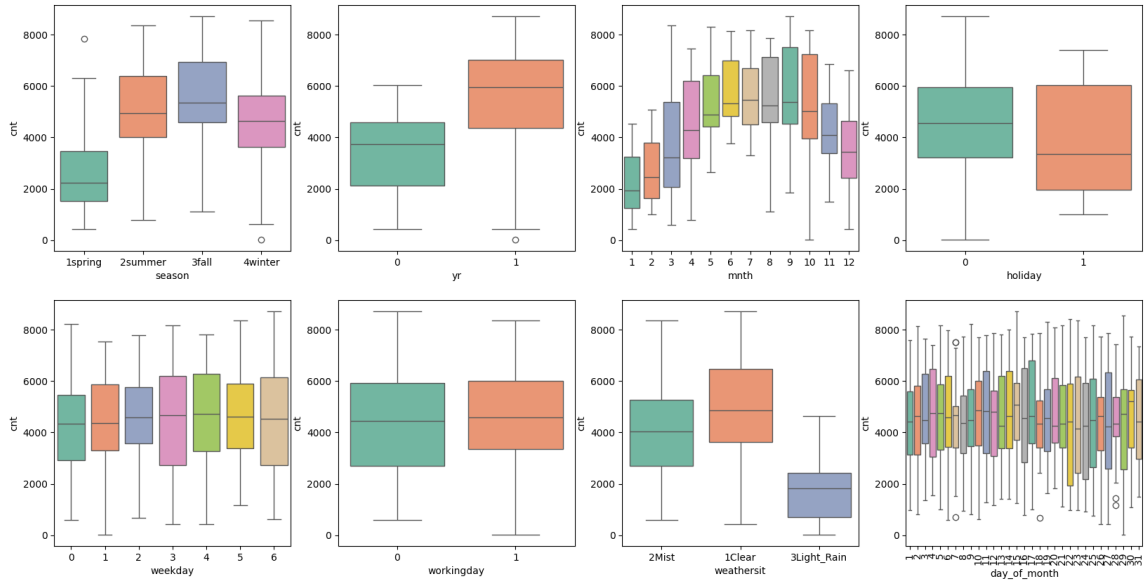

Figure 1: Box Chart of Categorical Variables

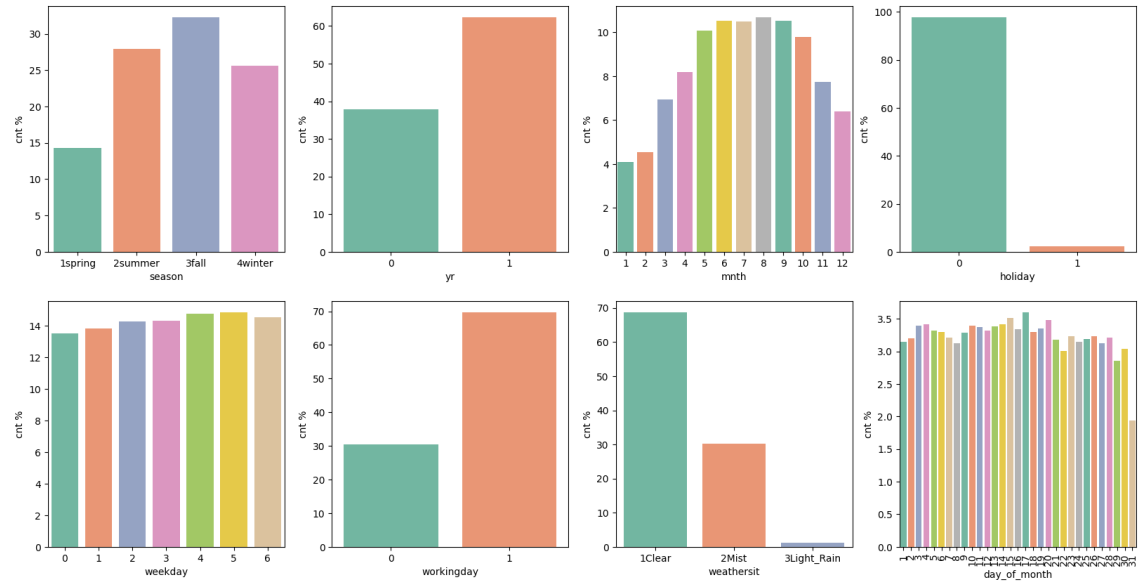

Figure 2: Bar Plot for bike counts % vs categorical variables

Below are the inferences from Categorical Variables:

- Bike counts % and median bike count are higher in Summer and Fall (season 2 and 3)
- Bike counts % and median bike count are higher in year 1 (2019) than year 0 (2018)
- Bike counts % are higher in month 2 to 12 months (lower in months 1 and 2).
- Median bike count are lower as compared to non-holiday

- Bike counts % are higher on a working day (which is logical as the working days are more). But weekdays are not making any difference on median bike counts (as reflected in box plots).
- Bike counts % and median bike are highest on clear day, while lowest for when weathersit = 3 raining
- Bike counts % drop in the last part of the month (though not significantly as reflected in box plot)

Question 2

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

Answer

If the categorical variable has "m" unique values and we create "m" dummy variables for each value, then 1 newly created dummy variable would be dependent on rest of the variable (e.g. we create 3 dummy variables for each of the unique values of a categorical variable, then the 3rd dummy variable can be derived if the first 2).

Hence, to avoid multi-collinearity, one variable needs to be dropped (total m-1 dummy variables for "m" unique values).

In this case, the function offers to drop the first variable as part of the input parameters and helps in avoiding multi-collinearity.

Question 3

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

Answer

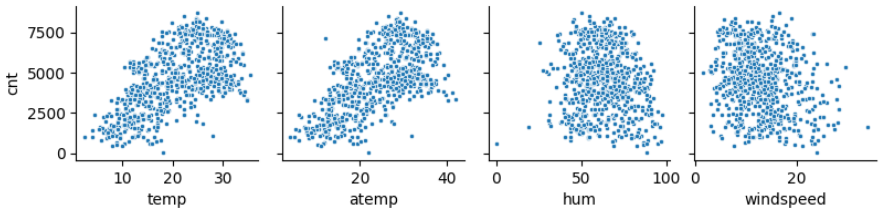

Figure 3: Scatter pair plot for numerical variables

We can see that temperature has the highest correlation to the target variable cnt. Though, atemp is equally highly correlated.

Question 4

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

Answer

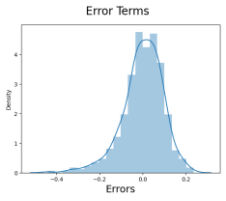

Figure 4: Error terms Residual Analysis

The error terms (y_train - y_train_pred) must follow a normal distribution. This is a one of most important assumptions of linear regression. This is reflected the above diagram that the error terms following a normal distribution.

Question 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

Below are the variables, which have the highest coefficients:

- **atemp** (coefficient 0.6376) – temperature of the day in Celsius (on colder days below 8 Celsius the bike counts are low and with rising temp, the bike counts rises)
- **yr** (coefficient 0.2413) – year, which is 1 for 2019 and 0 for 2018 (since bike counts in year 1 is higher)
- **weathersit_3Light_Rain** (coefficient -0.2163) – Derived dummy variable, which reflected the day with "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds" or a value of 3 for "weathersit". This is negative, as on such days, the bike counts is low.

GENERAL SUBJECTIVE QUESTIONS

Question 1

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

Answer

Liner regression tries to predict a **dependent variable based on other independent variables** (features). Meaning

Y = C + W1 x F1 + W2 x F2…

Y - dependent variable
W1, W2… – weights / coefficient for each feature
F1, F2… - Features / independent variables
C - constant which is the intercept

One of the methods in which the Linear regression model is trained (finding the optimal weights) is via **gradient descent algorithm**. The algorithm **iteratively** modifies the model weights / coefficients to minimize the error between actual and predicted values of the data. The gradient descent minimizes the error loss function (Mean Square Error MSE, etc.).

Gradient descent algorithm logic: The loss function is minimum when the gradient /slope of tangent (derivative of loss function) is 0. If the derivative of the loss function is -ve then the direction of the weight adjustment is +ve. In gradient decent algorithm, we adjust the weights based on this logic, till it reached local minimum.

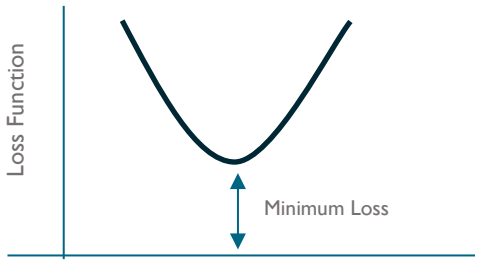

Figure 5: Loss function of weight vs weight plot

Once the model is trained and coefficients & constant determined, the data is run on the test data to predict the dependent variable Y_p.

Question 2

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

Answer

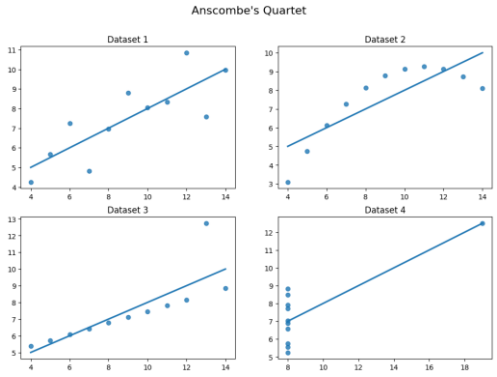

Figure 6: Reg Plot for Anscombe's quartet

Anscombe's quartet reflect 4 datasets, which present the same properties like Mean of x & y, Variance of x & y, Linear Regression line and Correction between x & y. Above is a Anscombe's quartet generated as part of the exercise in Appendix section of the python file.

Question 3

What is Pearson's R? (3 marks)

Answer

Pearson Correlation coefficient reflects the linear correlation between 2 set of variables. The calculation is covariance between these 2 variables divided by product of their standard deviations.
It ranges between -1 and 1 with 1 being perfectly correlated.

E.g. Pearson Correlation between temp and ateam for bike data set is 0.991 with pvalue=0.0. This would mean that these variables are highly correlated with very low p-value (high confidence level). This is generated as part of the exercise in Appendix section of the python file.

Question 4

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

Answer

Scaling is the process of bringing the variables to the same scales / range (e.g. could be between 0 and 1).
The advantages of doing scaling are that:

a. It addresses the issue of data skew on different variables (which mean some variables are way larger than the other variables). Without scaling, the **performance and accuracy** of the model reduces (as the convergence to local minimum / optimal point may not happen accurately or happen very slowly)

b. It ensures the same consideration is given to all variables. Without this large variables would dominate the learning and thus leading to **skewed outcomes**.

Normalization is the process of scaling, where it would try to transform the data by adjusting to the Max-Min of the feature.

- NormalScaled = (X - X_min)/(X_max - X_min)

Standardization is another process for scaling where transformation is done to ensure standard deviation equal 1 with mean as 0**.**

- StandardScaled = (X - mean)/Std

Question 5

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

Answer

Though practically very difficult to achieve in real life data but can happen in derived variables. Ff the variable is 100% dependent on other features), R2 of this variable would be 1 vis-à-vis other variables.

Considering the formulae of VIF = 1 / (1-R^2). This would be lead to infinite as a the answer for this variable's VIF.

Question 6

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

Answer

Quantile-Quantile plot helps in comparing 2 datasets follow the same distribution. It effectively plots the data vs normal distribution on a chart. If the 2 datasets are following the same distribution, the plots would appear to be the same. E.g. the "cnt" values in bike case study from train and test data.

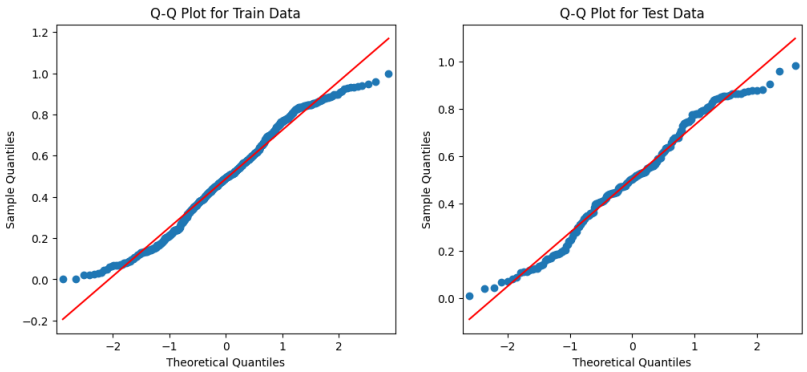

Figure 7: Q-Q plot for train and test data

In a linear regression, we need to ensure that train and test data are following the same distribution for regression to be valid. This is generated as part of the exercise in Appendix section of the python file.