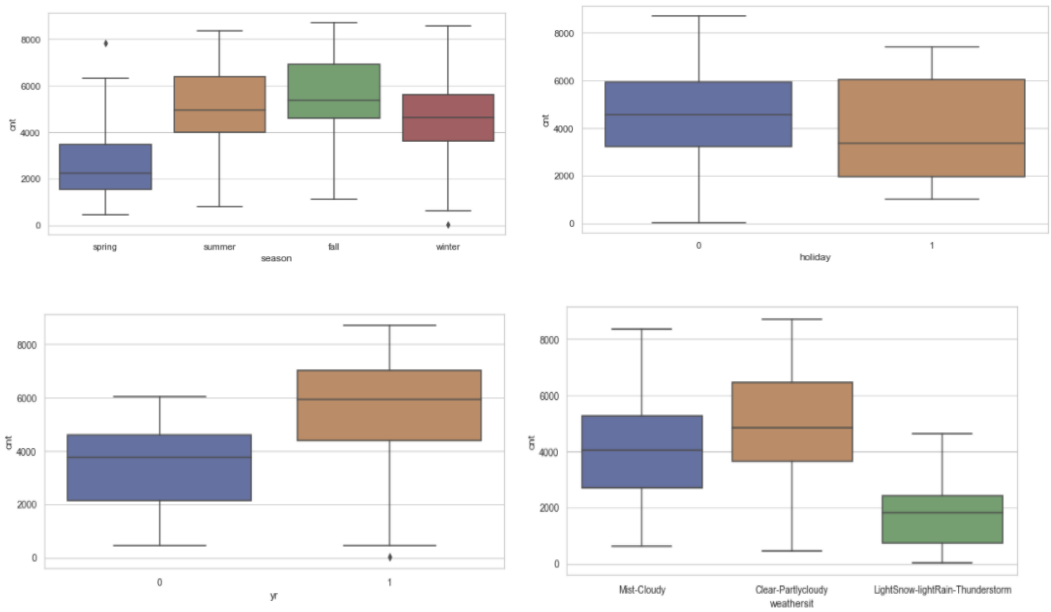
Assignment-based Subjective Questions

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

Answer:



1. Bike booking more in Summer, Fall and Winter with median around 5000. This indicates, season can be a good predictor for the target variable.
2. There is increase in Bike booking on holiday, hence holiday can be a good predictor.
3. Bike bookings were increased in year with median of greater than previous year booking, hence year looks good predictor.
4. There is increase in bike booking with weathersit Mist-Cloudy/clear-Partlycloudy. Looks good predictor.

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

Answer:

If we don't use "drop_first" we will get a redundant feature/column, with drop_first=True it reduces the extra column created during dummy variable creation and hence reduces the correlations created among dummy variables.

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

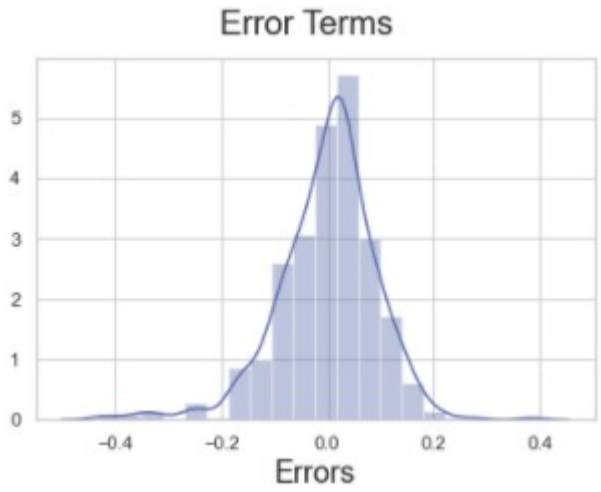
Answer:

temp and atemp has highest correlation with the target variable.

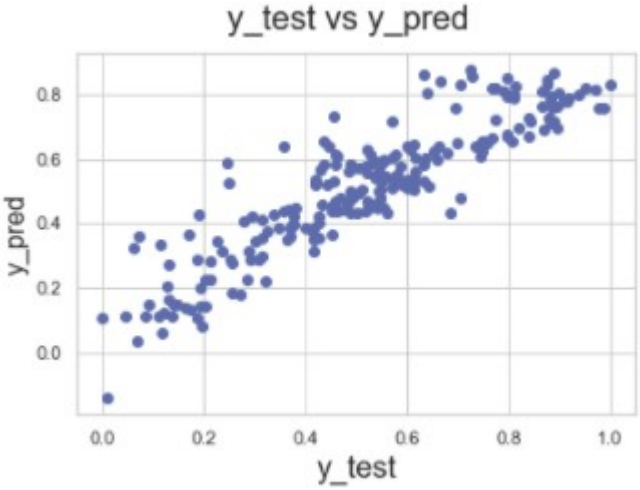
4. How did you validate the assumptions of Linear Regression aWer building the model on the training set?

   Answer:

   1. Check if the error terms are also normally distributed (which is infact, one of the major assumptions of linear regression)



   2. Graph for actual versus predicted values. Which should show linear relation between test and predictions.



   3. Check the r2_score which should be less than 1 and around computed R-squared.

```
[78]: from sklearn.metrics import r2_score
      r2_score(y_test, y_pred_m21)

[78]: 0.8054539524712321
```

Conclusion: Selected `model 21` looks significant as per the r2_score `0.8054` and computed R-squared is `0.822`.

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

   Answer:

   1. temp
   2. yr
   3. winter

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer:

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

The equation has the form y= a + bx, where y is the dependent variable (that's the variable that goes on the Y axis), x is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are multiple input variables, literature from statistics oWen refers to the method as **multiple linear regression**.

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called **Ordinary Least Squares**.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet tells us about the importance of visualising the data before applying various algorithms out there to build models.

It was constructed by statistician **Francis Anscombe** to illustrate the **importance** of **ploBng the graphs** before analysing and model building, and the effect of other **observations on statistical properties**. There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x, y points in all four datasets. They have very different distributions and appear differently when plotted on scatter plots.

3. What is Pearson's R?

   Answer:

Pearson's r is usually used to express the correlation between two quantities. Pearson's r value is used to evaluate whether the two quantities are correlated. R^2 is usually used to evaluate the quality of fit of a model on data.

Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.

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

   Answer:

   It 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.

   Normalized scaling:

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

   $$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

   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). klearn.preprocessing.scale helps to implement standardization in python.

   $$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

   Answer:

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q p

   Answer:

   A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles.

   This is a graphical tool to help in assessment if a set of data plausibly came from some theoretical distribution such as a normal or exponential.