Linear Regression

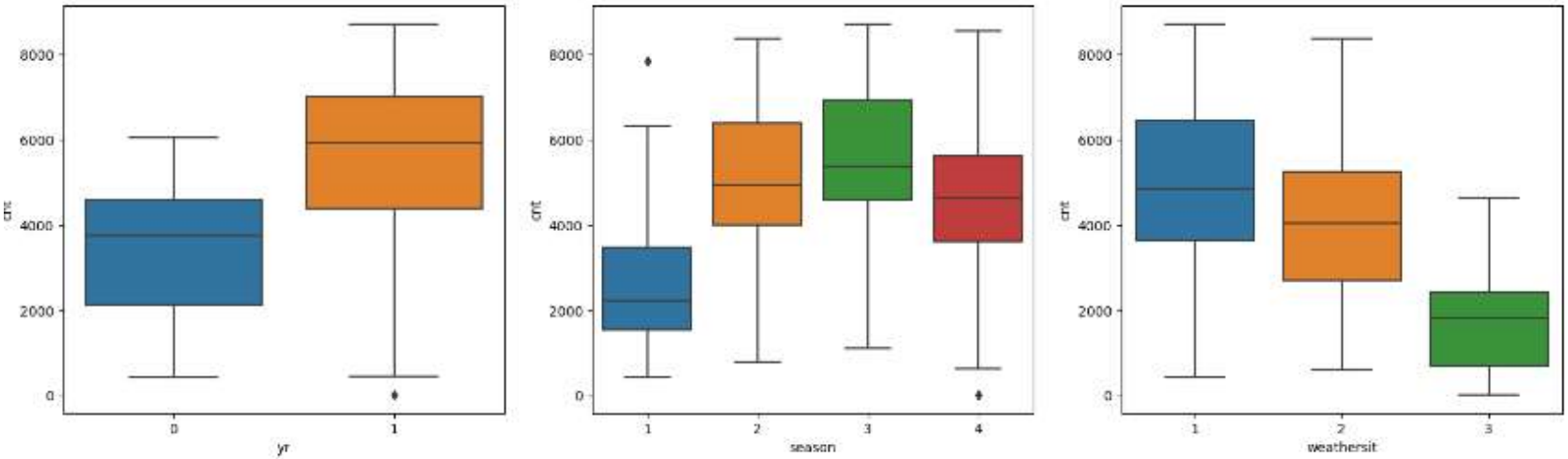
Rhishikesh Parkhi

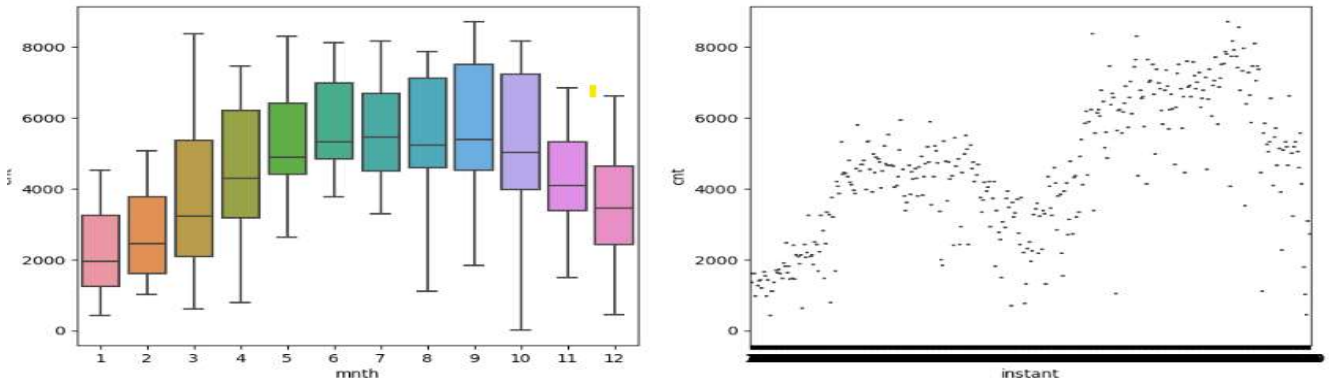
Assignment based Subjective Question

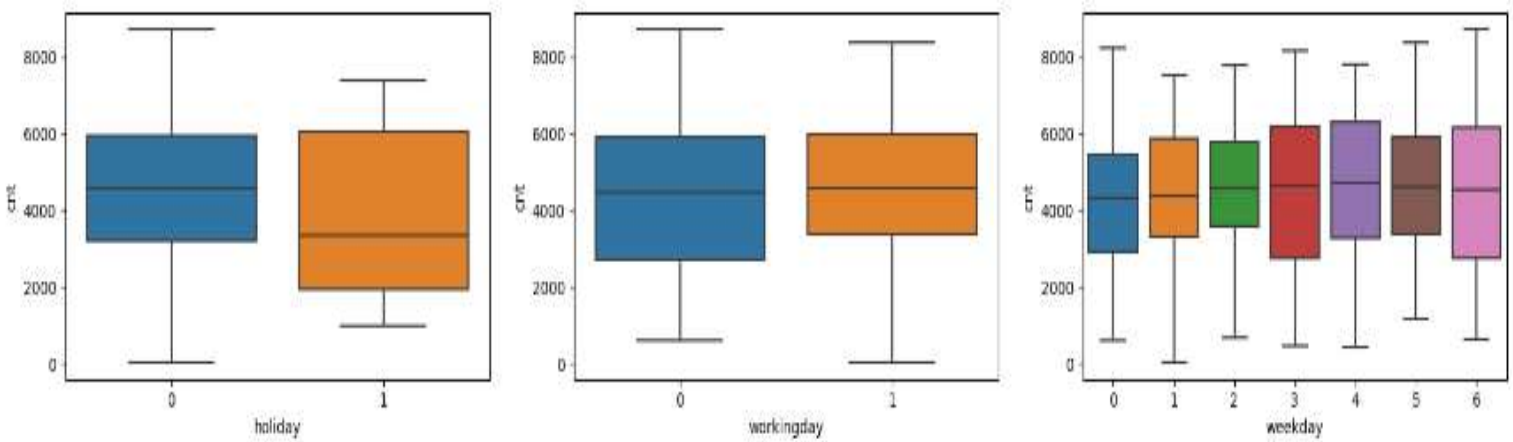
Q1-From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- yr –significant impact
- season- significant impact
- weathersit - significant impact
- mnth - significant impact
- holiday- has impact but not very significant.
- workingday – little to no impact
- weekday –little to no impact
- instant-significant impact

However there will be some overlap with variables season,month,year,instant.


cnt
8000
6000
4000
2000
0
0
1
yr
cnt
8000
6000
4000
2000
0
1
2
3
4
season
cnt
8000
6000
4000
2000
0
1
2
3
weathersit


8000
6000
4000
2000
0
1
2
3
4
5
6
7
8
9
10
11
12
mnth
cnt
8000
6000
4000
2000
0
instant

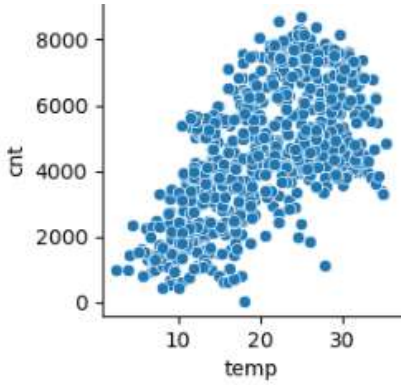

cnt
8000
6000
4000
2000
0
0
1
holiday
cnt
8000
6000
4000
2000
0
0
1
workingday
cnt
8000
6000
4000
2000
0
0
1
2
3
4
5
6
weekday

Q2-Why is it important to use drop_first=True during dummy variable creation?

- If a categorical variable has n values the number of dummy variables is n-1.to keep the count of dummy variables to n-1 we use **drop_first=True**
- N-1 values are enough to occupy all the info of n values.
- With n-1 values we reduce multicollinearity/redundancy within independent variables.
- Also the time for training the model and resources get reduced.
- Model becomes more interpretable.

Q3-Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- temp variable the highest correlation with target variable.



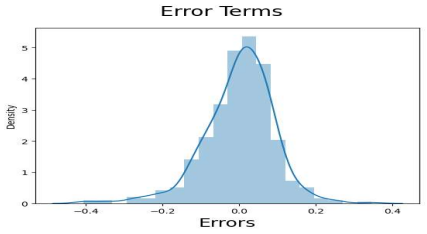
Q4-How did you validate the assumptions of Linear Regression after building the model on the training set?

- Residual Analysis

  Take diff of actual and predicted values from training set.

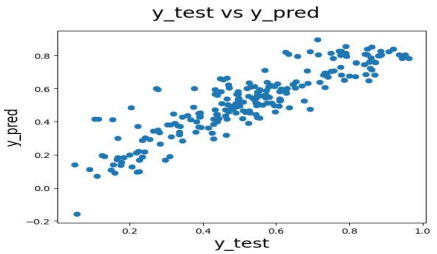
  Plot the histogram of the diff values

  Residual should have normalized distribution



- Linearity

  Plot actual values vs predicted values using scatter plot.It should be linear



VIF Variance inflation factor to remove multi collinearity.

Check RSME value should be near 0

```
In [50]: # Get the mean squared error
         rsme=np.sqrt(mean_squared_error(y_true=y_test, y_pred=y_pred))
         print('RSME=', rsme)

         RSME= 0.09946458878782669
```

R2 on test data should be near R2 on train data.

Training Set Value

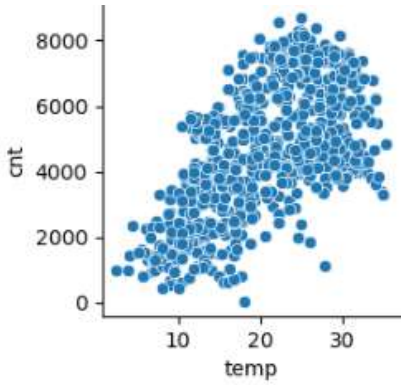
```
R-squared:                       0.831
```

Test Set Value

```
R-squared:                       0.813
```

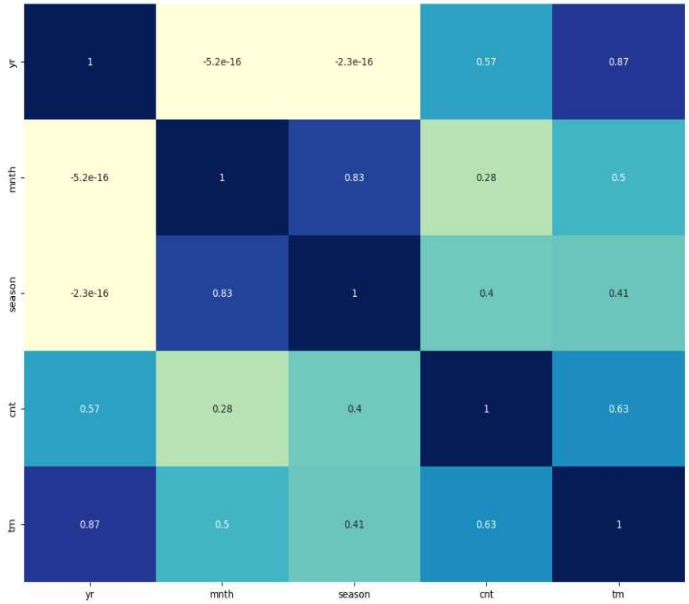
Q5-Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

1-temp



2-dtedate

Instant /dtedate (converted to tm and also cover yr,mnth,season to some extent as there is significant overlap within these also)



3-windspeed

General Subjective Questions

Q6-Explain the linear regression algorithm in detail.

- Regression means teaching a machine predict value of continuous variable using supervised learning method based on one or more inputs.
- Linear regression model tries to explain relation between dependent variable(output variable) and independent variable(predictor variable) using straight line. So linear dependence is primary assumption, also errors are normally distributed and no relation between predictor variables(no multicolinearirty).
- We have to find the best fit line that passes through the scatter plot of dependent and independent variable such that residual and RSS is minimized. This will give us the optimal value of m(slope) and c(intercept).$y=m(x) +c$ or $y=(\beta_1)x + \beta_0$

Q6-Explain the linear regression algorithm in detail.

Steps

5 The best fit regression line can be found by minimizing the cost function using differentiation or Gradient descent. The cost function used in linear regression is Mean Squared Error(MSE) which measures average squared difference between predicted and actual values.

6 Model strength is measured using R^2 where $R^2=1-(RSS/TSS)$

RSS is residual sum of Square , TSS is total sum of squares

al sum of squares

7 Coefficients β_o , $\beta 1 \ldots\ldots \beta n$ provide what relation lies between dependent and independent variables.

Q6-Explain the linear regression algorithm in detail.

Steps

1. The data set provided is divided in two parts training and test data using (70/30 or 80/20) or whichever combination desired.
2. Training data is used for teaching the model to learn during modelling
3. Test data is used for prediction based on trained model and hence model evaluation.
4. Simple Linear regression has only 1 independent variable. Multiple regression has more than one independent variable.

Q7-Explain the Anscombe's quartet in detail

Anscombe's quartet consist of 4 datasets having identical statistical properties like mean, variance, R-squared, correlations and linear regression lines but have very different scatter plot.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

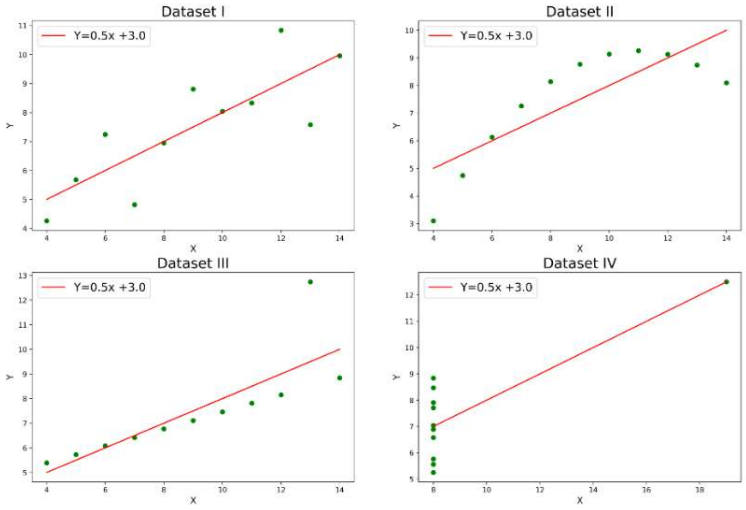
It tells us the importance of EDA and data visualization and not just depend on statistical summary.

Visualization and EDA can help identify outliers, trends and other critical details that summary statistics might not tell us.

Q7-Explain the Anscombe's quartet in detail

4 datasets in Anscombe's quartet each include 11 x-y pairs of data.

Despite significant variations all datasets have same summary(mean, variance, correlation coefficient and regression line)



DS1 -scatter plot you will see that there seems a linear relationship between x and y.
DS2- there is a non-linear relationship between x and y.
DS3- perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
DS4- when one high-leverage point is enough to produce a high correlation coefficient.

Q8-What is Pearson's R?

- Pearson's r is a measure of correlation between 2 continuous variables. It tells strength and direction of the correlation.
- Values range from -1 to +1
- +1 indicates perfect positive linear correlation. If one increases other increases proportionally.
- -1 indicates perfect negative linear correlation. If one increases other decreases proportionally.
- 0 indicates there is no linear correlation between the two.
- Pearson's r measures accurately linear relations, it might not capture accurately non linear relationship.
- Pearson's r does not imply one variable is causing other variable to change.
- $r = (\Sigma[(xi - \bar{x})(yi - \bar{y})]) / \sqrt{(\Sigma(xi - \bar{x})^2 * \Sigma(yi - \bar{y})^2)}$

  xi and **yi** with your data points for the two variables, and $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the mean values

Q9-What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Scaling is the method to bring values of different variables to common scale. This is essential as the algorithms used in learning are sensitive to scale.
- Gradient descent algorithm used to converge is very sensitive to scale. It works best when all variables are in the same scales. This will help reduce the resources and time to converge.
- Results can be biased/influenced by feature scale. Large scale will dominate the results.
- Two common type of scaling are
  - Min-Max Scaling/Normalised Scaling

    x_normalized = (x - min(x)) / (max(x) - min(x))
  - Z-score Scaling /Standardised Scaling)

    x_standardized = (x - mean(x)) / std(x)

Q9-What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Two common type of scaling are
  - Min-Max Scaling/Normalized Scaling

    x_normalized = (x - min(x)) / (max(x) - min(x))

    The values of feature are fit between 0 and 1

    Original distribution of data is maintained.
  - Z-score Scaling /Standardized Scaling)

    x_standardized = (x - mean(x)) / std(x)

    The values of feature are fit such that mean is 0 and standard deviation is 1

    This method transform data to have Gaussian distribution and helpful in case where algorithm is sensitive to outliers.

Q10-You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- VIF is predictor for multicolinearity between predictor variables.
- If predictor variables are highly correlated then it is difficult to predict which variable change is affecting the target variable.
- As variables tend towards perfect collinearity the value R_i^2 tends to 1 and hence VIF_i tends to infinity.
- $VIF_i = 1 / (1 - R_i^2)$
- If two variables are perfectly collinear VIF will be infinite as denominator is 0 if R_i^2 is 1.
- VIF should be used to remove redundancy from the model.

Q11-What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

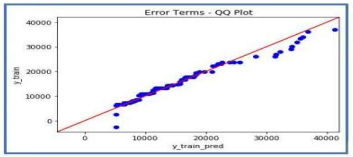
Q-Q plot is Quantile Quantile plot used to check if given dataset follows given distribution such as normal distribution. It is used to compare observed data distribution and expected distribution.

- Q-Q Plot tells if residuals are normally distributed.
- If Q-Q plot deviates from straight line it suggests that residuals are not normally distributed.
- Q-Q plot is also used for outlier detection
- If the points deviate from the line, especially at the tails, it indicates departures from normality.

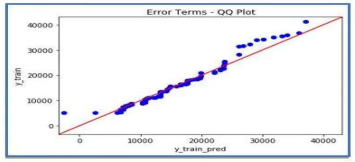
In linear regression when we have training and test data set received separately we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Q-11What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis
- Y-values < X-values: If y-quantiles are lower than the x-quantiles



- X-values < Y-values: If x-quantiles are lower than the y-quantiles.



- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis