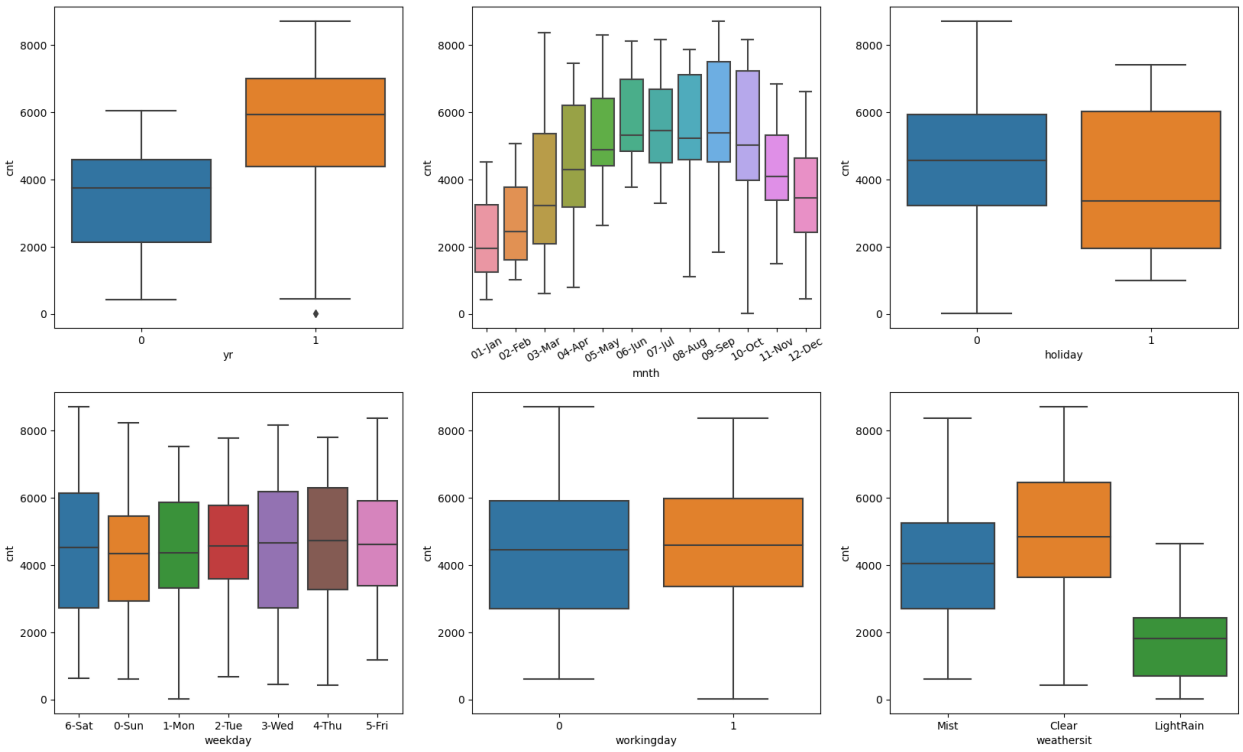
Linear Regression Subjective Questions

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? (3 marks)**

The analysis of categorical variables were done using a boxplot. The following variables were considered as categorical – yr, mnth, holiday, weekday, workingday, weathersit



From the above boxplots, we can infer the following

- Demand for bikes in 2019 has a significantly higher number of rentals compared to 2018
- Months June, July, August, September have higher demand compared to other months
- Demand is higher on a non-holiday days
- No significant difference in median values across the days of the week
- No significant difference in median values across working day or
- Higher demand for bikes on clear days

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

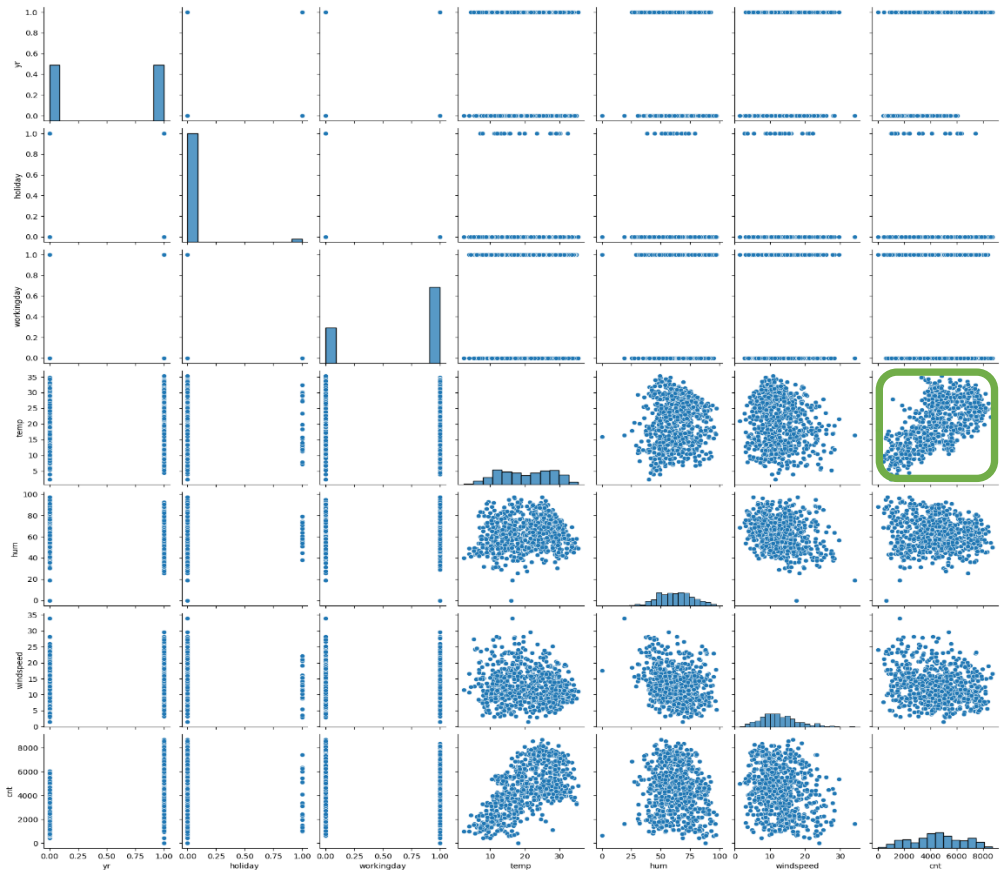
get_dummies method in pandas is used to convert categorical variables into dummy variables.

For a categorical variable that has 5 unique values in the dataset, get_dummies method generates 5 dummy variables by default (give default value of drop_first argument is false).

The key idea behind creating dummy variables is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one.

drop_first = True to obtain k-1 dummy variables.

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



Based on the above box plot, high correlation is noted between the target variable and **temp** independent variable.

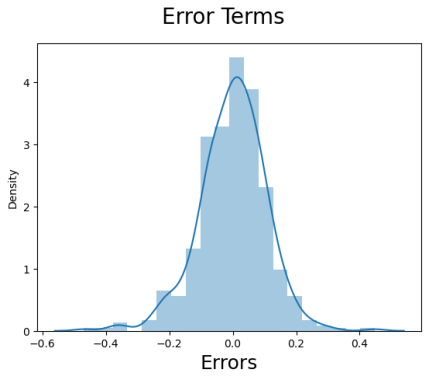
Based on correlation matrix that was done in an earlier step,

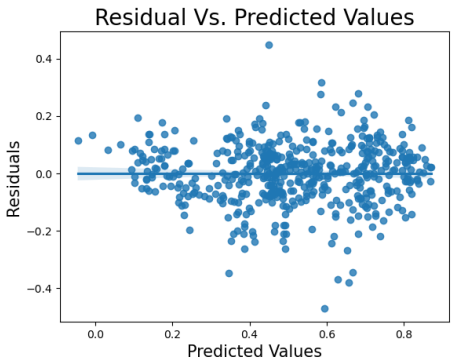
- We can see that **atemp** is highly correlated with **temp** and hence will be highly correlated to **cnt**

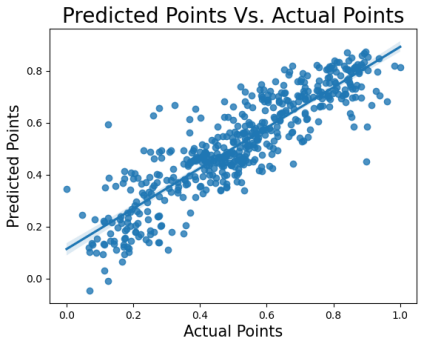
- We can also see that variables like **casual** and **registered** are also highly correlated with the target variable **cnt**

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

Residual analysis was performed to validate the assumptions of the Linear Regression Model. As part of residual analysis, the following visualisations were built







- By plotting predicted values against actual data, we see that the data points are fitted around the best fitted line except a very few exceptions. Hence, we infer that the model is a reasonably good model
- Based on residual analysis, we are able to validate some of the assumptions made for linear regression model
    - Error Terms distribution Plot - The error terms are normally distributed with a mean 0.
    - Residuals vs Predicted Values Scatter Plot -
        1) This visual helps ascertain the variance with the residual values. As can be seen from the visual, the residual values are distributed around the horizontal line indicating reasonably constant variance across all the residual values
        2) From the visual, it is noticed that the residual values are spread across the length of the x-axis suggesting that there is no visible pattern with respect to the residuals. This ascertains that the error terms are independent of each other

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

The top 3 features that explain the demand of shared bikes are

1. **Year** – Yr feature has a positive coefficient of 0.2477. This signifies that the demand is expected to increase by 0.2477 with a unit increase in the yr value.
2. **Weather Situation** - LightRain feature has a negative coefficient (-0.2893) suggesting that a demand is expected to decrease with an increase in unit value of these features. So, we can expect that the demand will decrease when there is light rain.
3. **Season** - Sprint feature - has a negative coefficient (-0.2324) suggesting that the demand is expected to decrease when it is spring.

General Subjective Questions

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

Regression is a method of modelling a target value based on independent predictors. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Linear Regression attempts to model the relationship between variables by fitting a linear equation / straight line.

There are 2 types of linear regression
1) Simple Linear Regression
2) Multiple Linear Regression

Simple Linear Regression is a statistical method for establishing the relationship between two variables (continuous dependent variable and independent variable) by using a straight line.

Simple Linear Regression can be represented by the formula

$$y = \beta 0 + \beta 1x + \varepsilon$$

where

- y is the predicted value of the dependent variable (y) for any given value of the independent variable (x).
- β0 is the intercept, the predicted value of y when the x is 0.
- β1 is the regression coefficient or slope – how much we expect y to change as x increases.
- x is the independent variable
- ε is the error of the estimate, or how much variation there is in our regression coefficient estimate.

Multiple Linear Regression helps understand relationship between a continuous dependent various and two or more independent variables.

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta ixi + \varepsilon$$

where

- y is the predicted value of the dependent variable (y)
- The regression coefficients β1 and β2 represent the change in y as a result of one-unit changes in x1 and x2.
- β refers to the slope coefficient of all independent variables
- ε term describes the random error (residual) in the model.

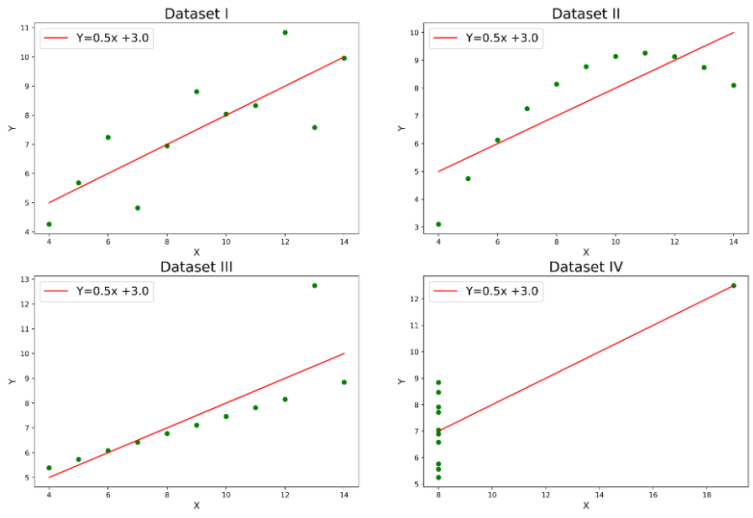
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe Quartet is a set of four datasets having identical properties such as mean, R-squared, variance, correlation and linear regression lines but having different representation we the data is represented in a scatter plot. The dataset was created by Francis Anscombe, a statistician to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets each contains 11 x-y pairs of data.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths.

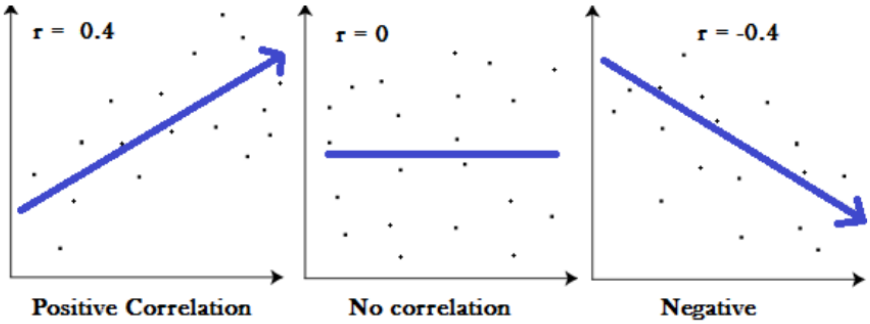


Anscombe's quartet Plot

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

Pearson's correlation, also called Pearson's R or Pearson Product Moment Correlation (PPMC) is a correlation coefficient commonly used in linear regression. Pearson coefficient measures the strength of the relationship between two variables. It is the ratio between the covariance of two variables and the product of their standard deviations

It is independent of the unit of measurement of variables and the coefficient values ranges between +1 and -1.



The formula to calculate Person's R is

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where

- r = Pearson Coefficient
- n= number of pairs of x, y values in the dataset
- ∑xy = sum of products of the x,y pairs
- ∑x = sum of the x scores
- ∑y= sum of the y scores
- ∑x2 = sum of the squared x scores
- ∑y2 = sum of the squared y scores

A higher absolute value of the correlation coefficient indicates a stronger relationship between variables. +1 indicates a strong positive relationship, -1 indicates a strong negative relationship and 0 indicates no relationship.

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

Scaling is a data preparation step that involves transforming the data points (continuous variables) to be in a similar scale. This is done to ensure all features contribute equally to the model and to avoid one variable dominating the model.

Without scaling features, the algorithm may be biased toward the feature with values higher in magnitude. Hence, we scale features so that all values are in the same range and the model uses every feature without any bias.

Normalized Scaling is a technique in which values are adjusted so that they range between 0 and 1. It is also known as Min-Max scaling.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization is a scaling method where the values are centered around the mean and the standard deviation is 1.

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

For categorical variables, scaling is done after creating dummy variables and converting them into numeric format.

---

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

VIF helps explaining the relationship between one independent variables with all other independent variables.

An infinite value of VIF for an independent variable indicates that it can be perfectly predicted by other variables in the model.

The formula to calculate VIF is

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

Based on the above formula, infinite value of VIF happens when R^2 value is 1.

R^2 represents a perfect correlation between two or more independent variables.

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

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. The power of Q-Q plots lies in their ability to summarize any distribution visually.

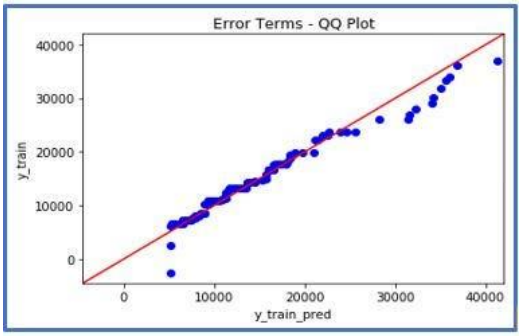
QQ plots is very useful to determine –

- If two populations are of the same distribution
- If residuals follow a normal distribution.
- Skewness of distribution

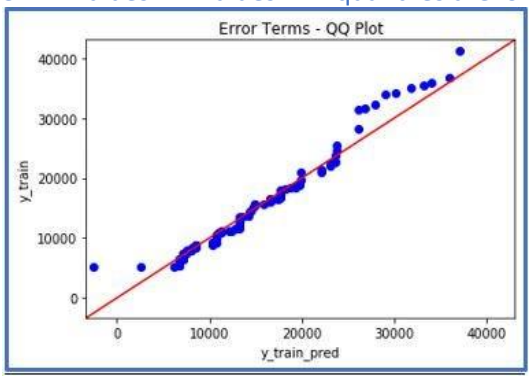
Below are the possible interpretations for two data sets.

1. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

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



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



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

In Linear Regression, when we have training and test data set received separately, plotting it using Q-Q plot will help determine if the two datasets follow similar distribution patterns.