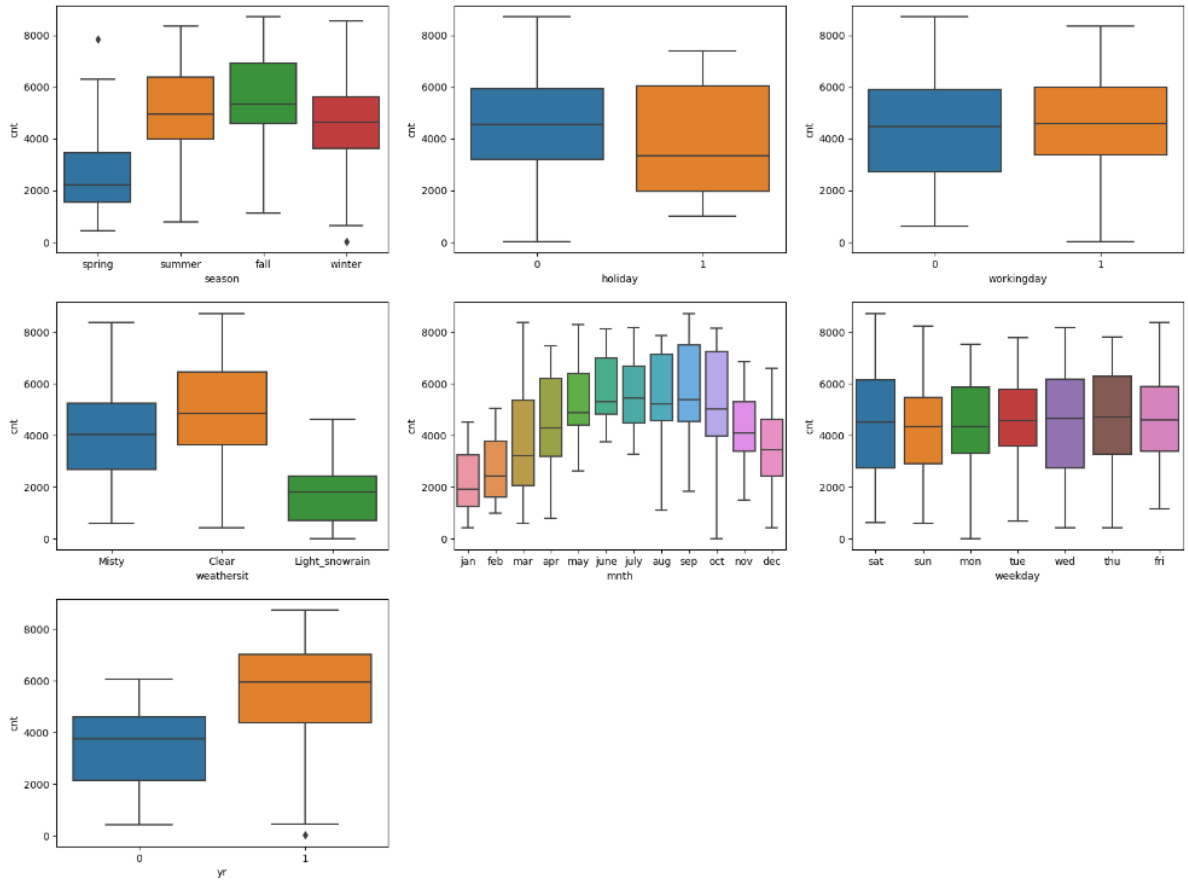
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

There are a couple of categorical variables namely season, holiday, mnth, yr, weekday, working day and weathersit. These categorical variables have a major effect on the dependent variable 'cnt'.



Inferences :-

- Fall, Summer and Winter season seems to have attracted more booking.
- When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking seemed to be almost equal either on working day or non-working day.
- Clear weather attracted more booking which seems obvious.
- Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Thu, Fri, Sat and Sun have slightly more number of bookings as compared to the start of the week.
- 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

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

Answer:

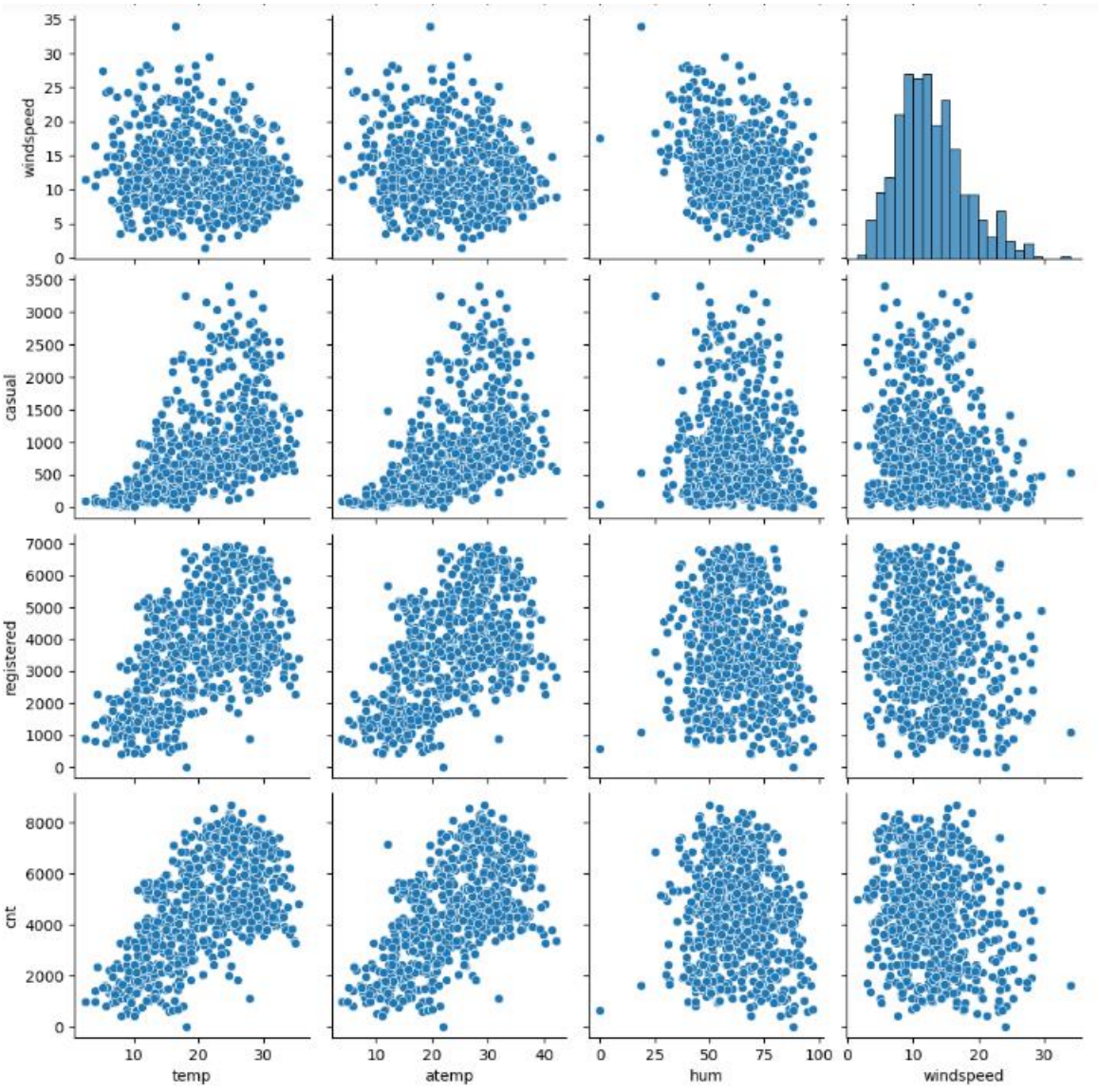
drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Dropping the column is important because the importance or value of that left over variable can be found by remaining variables. So to avoid redundancy we are dropping a column. This helps in the column to become linearly independent.

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

Answer:

The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.

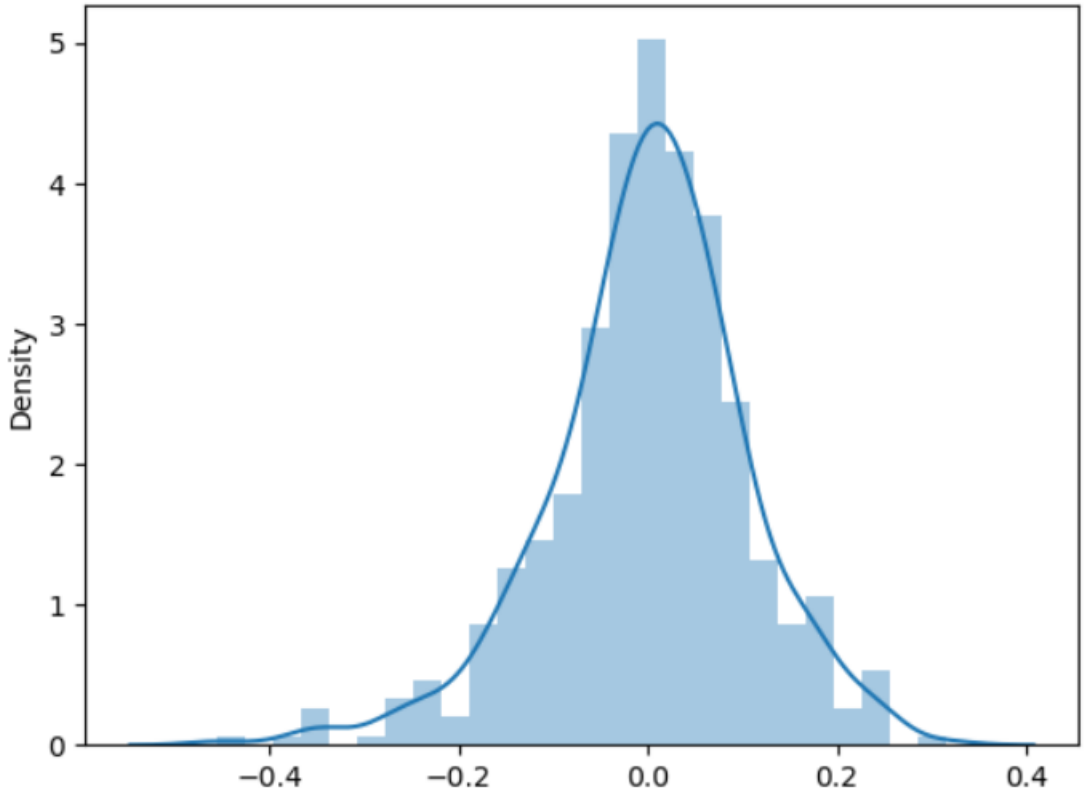


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

The assumption of Linear Regression Model based on below 5 assumptions : –

1. Normality of error - Error terms should be normally distributed



2. Multicollinearity check - There should be insignificant multicollinearity among variables.
3. Linear relationship validation - Linearity should be visible among variables.
4. Homoscedasticity - There should be no visible pattern in residual values.
5. Independence of residuals - No auto-correlation

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:

Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and season.

General Subjective Questions

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

Answer:

Linear Regression Algorithm is a machine learning algorithm based on supervised learning where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, rather than trying to classify them into categories. It is a part of regression analysis. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression.

Simple Linear Regression - Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction.

$$y = mX + b$$

where,

y = dependent variable

X = independent variable

m = intercept of the line

b = linear regression coefficient

When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

Assumptions of Simple Linear Regression –

There are four assumptions associated with a linear regression model:

- Linearity: The relationship between X and the mean of Y is linear.
- Homoscedasticity: The variance of residual is the same for any value of X.
- Independence: Observations are independent of each other.
- Normality: For any fixed value of X, Y is normally distributed.

Multiple Linear Regression - A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

$$y=w1X1 + w2X2 + w3X3$$

The variables X1, X2, X3 represent the attributes, or distinct pieces of information, we have about each observation.

Assumptions of Multiple Linear Regression –

Multiple linear regression analysis makes five key assumptions:

- Linear relationship: There exists a linear relationship between each predictor variable and the response variable.
- No Multicollinearity: None of the predictor variables are highly correlated with each other.
- Independence: The observations are independent.
- Homoscedasticity: The residuals have constant variance at every point in the linear model.
- Multivariate Normality: The residuals of the model are normally distributed.

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

Answer:

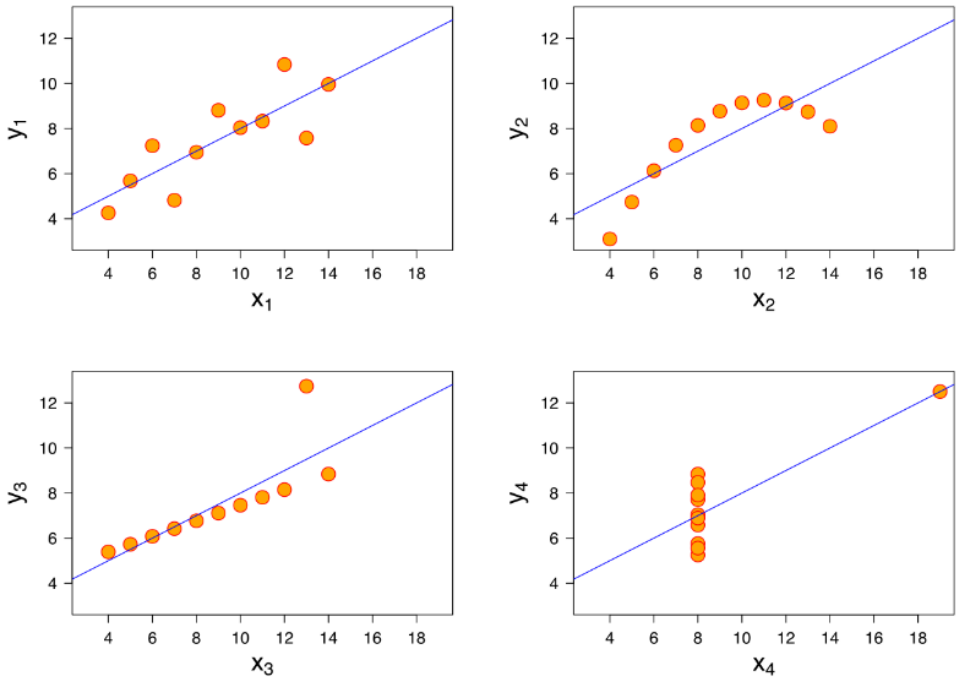
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

Answer:

The Pearson correlation method is the most common method used for numerical variables. It assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation. 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.

Pearson's R Formula is as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

1. r = correlation coefficient
2. xi = values of the x-variable in a sample
3. $\overline{x}$ = mean of the values of the x-variable
4. yi = values of the y-variable in a sample
5. $\overline{y}$ = mean of the values of the y-variable

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

Answer:

Scaling means transforming the data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.

6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

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

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

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

Answer:

VIF(Variance Inflation Factor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:
A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.
A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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 (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. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. The power of Q-Q plots lies in their ability to summarize any distribution visually.

The advantages of the Q-Q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested.

Q-Q plot is very useful to determine:

1. If two populations are of the same distribution
2. If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
3. Skewness of distribution