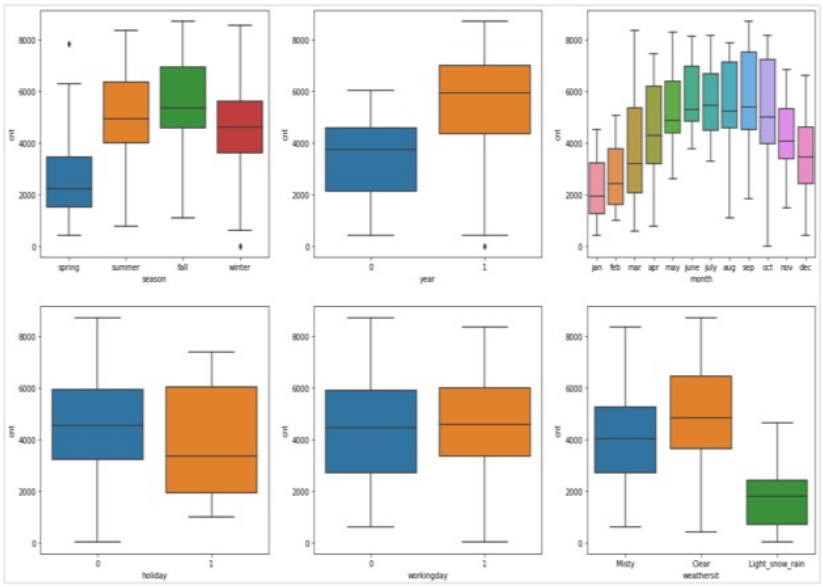
Assignment-based Subjective

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:
Columns with categorical data are:

- "season"
- "year"
- "month"
- "holiday"
- "workingday"
- "weathersit"

These categorical variables have major effect on the dependent variable “cnt”. We used boxplot to conclude on this.



We could see some patterns here:

- "fall" has on an average higher "cnt" than other seasons. 2019 has an increase over 2018.
- Bookings during the month of may, june, july, aug, sep and oct is more as compared to other months.
- Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Months in 2019 have more bookings as each months in 2018.
- "2019" year has on an average higher "cnt" when compared to "2018".
- When weather situation is "Clear", "cnt" is a little higher as compared to other weather situations. "2019" has an increase over "2018".
- “workingday” seem to be having no significant pattern wrt. “cnt” overall.

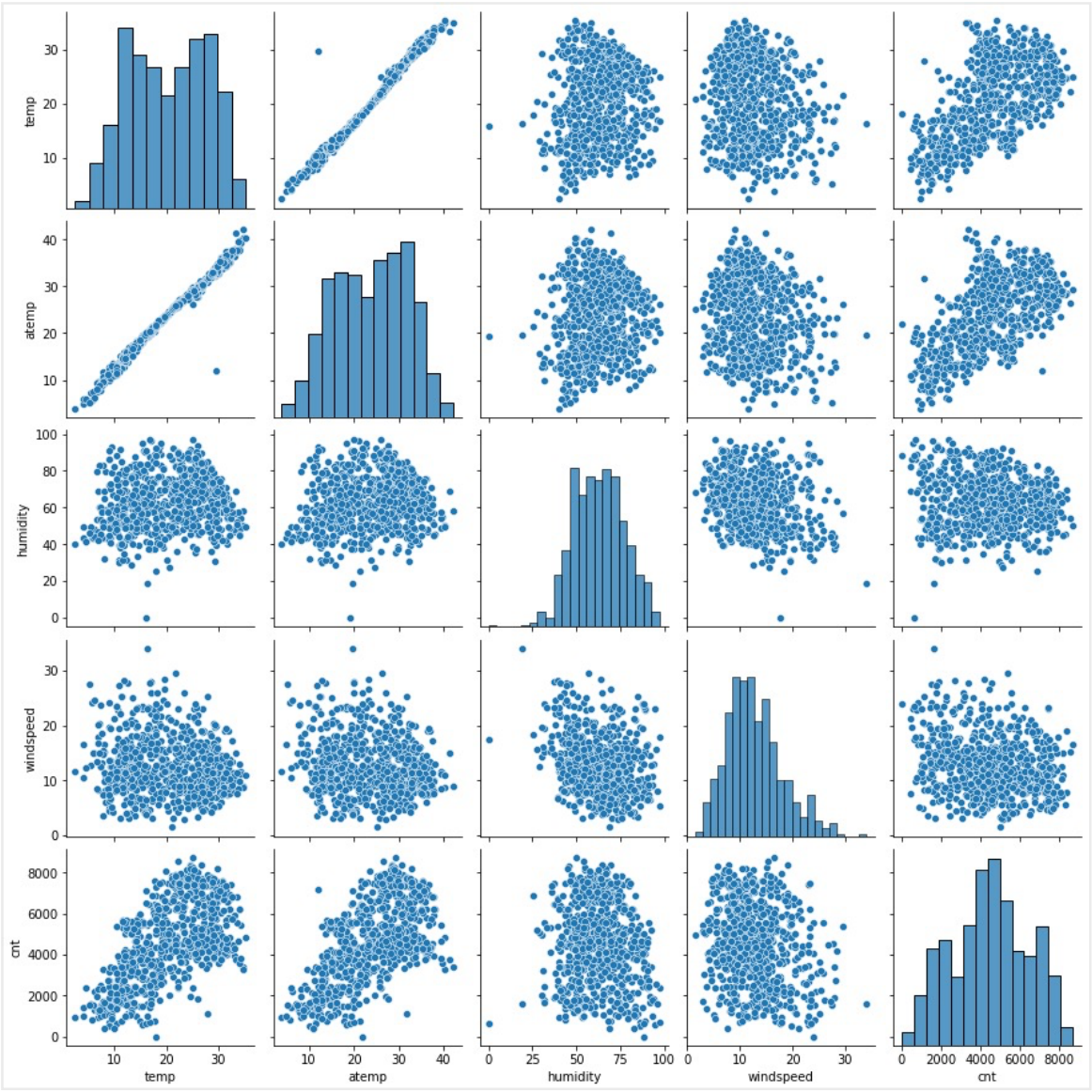
2. Why is it important to use drop_first=True during dummy variable creation?
(2 mark)

Answer:
drop_first=True is important to use. It helps in reducing the extra column created during dummy variable creation.
Hence it reduces the correlations created among dummy variables.
Eg:-
Suppose, we have a categorical column(/feature) that contains three types of values- "A", "B", "C".
So if the variable is neither of “A” or “B”, then it will be “C”.
So we don’t need the third variable to identify “C”.

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

Answer:

Here is the pair-lot among numerical variables.



We could see that “temp” and “atemp” have the highest correlation with target variable (“cnt”).

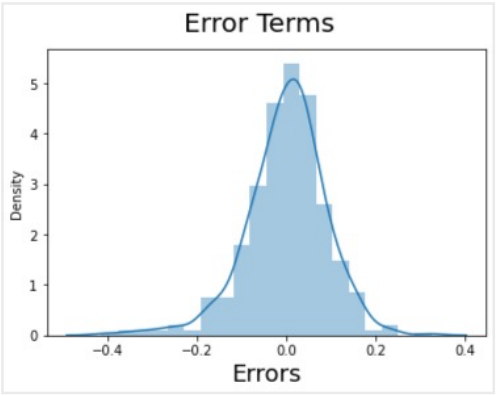
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Answer:

I have validated the assumptions of linear regression after building the model on the training set based on:

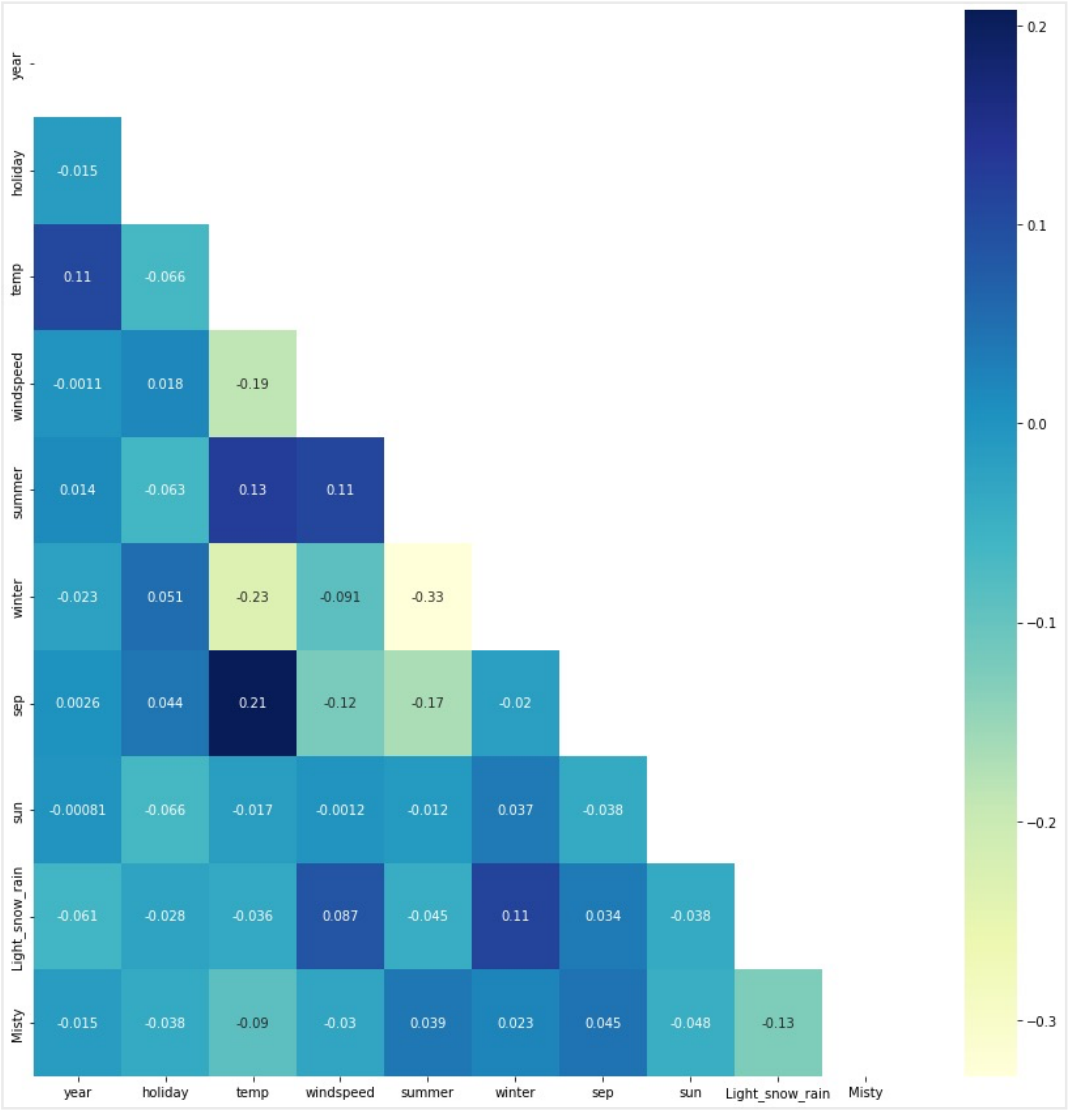
Normality of error terms

- Errors are normally distributed for any given value of X

Multicollinearity check



Ø There is no significant correlation between features based on the heatmap

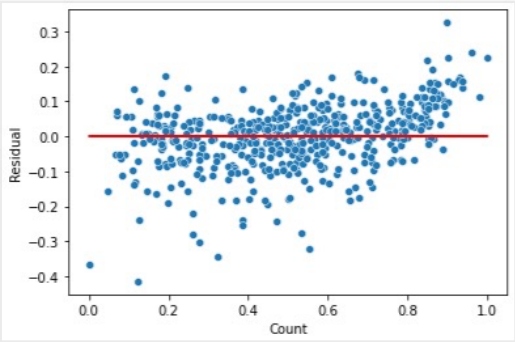


Linear Relationship Validation

- Linearity should be visible between variables.

Homoscedasticity

- Ø The variance of the residual is more or less constant for this model.



Independence of errors

- Error values are statistically independent.
- Durbin-Watson value of final model is 2.103 [value ranging between o and 4], which signifies there is no autocorrelation.

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 top three features contributing significantly towards explaining the demand of the shared bikes:

- temp
- weathersit: Light_snow_rain
- year

General Subjective Questions

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

Linear regression is process of estimating relation between variables where focus is on establishing relationship between dependent and independent variables.
Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task**.**
Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.
Different regression models differ based on the kind of relationship between dependent and independent variables they are considering, and the number of

independent variables getting used.

Regression Line:

- $y = mx + c$
- Or $y = \beta_0 + \beta_1 x$, where

- y = how far up
- x = how far along
- m = Slope or Gradient (how steep the line is) or β_1
- c = value of y when x=0; c or β_0

Note:
If there are several variables used to predict the outcome of a response variable , it is called as multiple linear regression.
This is an extension of linear regression.

Assumptions on linear regression-

- Linear Regression assumes that there is no or very little multicollinearity in the data.
- There is a linear relationship between X and Y. X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.
- Error terms are normally distributed with mean zero(not X, Y).
- Error terms are independent of each other.
- Error terms have constant variance (homoscedasticity).

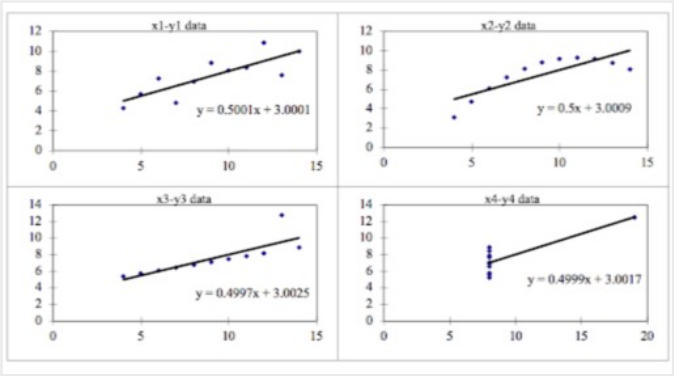
2. Explain the Anscombe's quartet in detail. (3 marks)
Answer:

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before we could analyse it and build your model.
Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics.
They provide the same information (involving variance and mean) for each x and y point in all four data sets.
However, when we plot these data sets, they look very different from one another.

Example:
The statistical information for these four data sets are approximately similar. We can compute them as follows:

Anscombe's Data									
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

When these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

In summary, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm.
So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

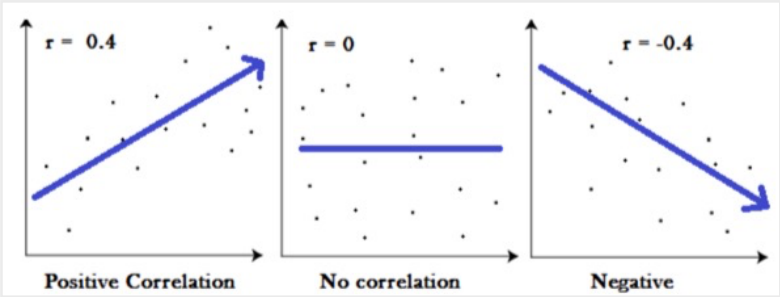
3. What is Pearson's R? (3 marks)
Answer:

Pearson's r is a statistic that measures the linear correlation between two variables.
Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient.
However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson correlation coefficient (r):

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

Diagrams Depicting correlations:



In summary:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

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

What is scaling -
Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

Why is scaling performed -
Scaling helps in speeding up the calculations in an algorithm.
In general, collected data set contains features varying in magnitudes, units and range.
If scaling is not done, then algorithm only takes magnitude in account and not unit. This results in incorrect modelling.
To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Difference between normalized scaling and standardized scaling-
Normalization Scaling -
It brings all of the data in the range of 0 and 1.
It is also called as MinMax Scaling.

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

sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardised 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**).

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

sklearn.preprocessing.scale helps to implement standardization in python.

- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

If there is perfect correlation, then VIF = infinity.
This 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.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which thereby show an infinite VIF as well).

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.
It helps us assess whether a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
Also, it helps us determine if two data sets come from populations with a common distribution.

Use Case :
It is used to check following scenarios, if two data sets:

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour

Few advantages:

- It can be used with sample sizes also.
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

Interpretations:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Possible interpretations for two data sets:

- 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