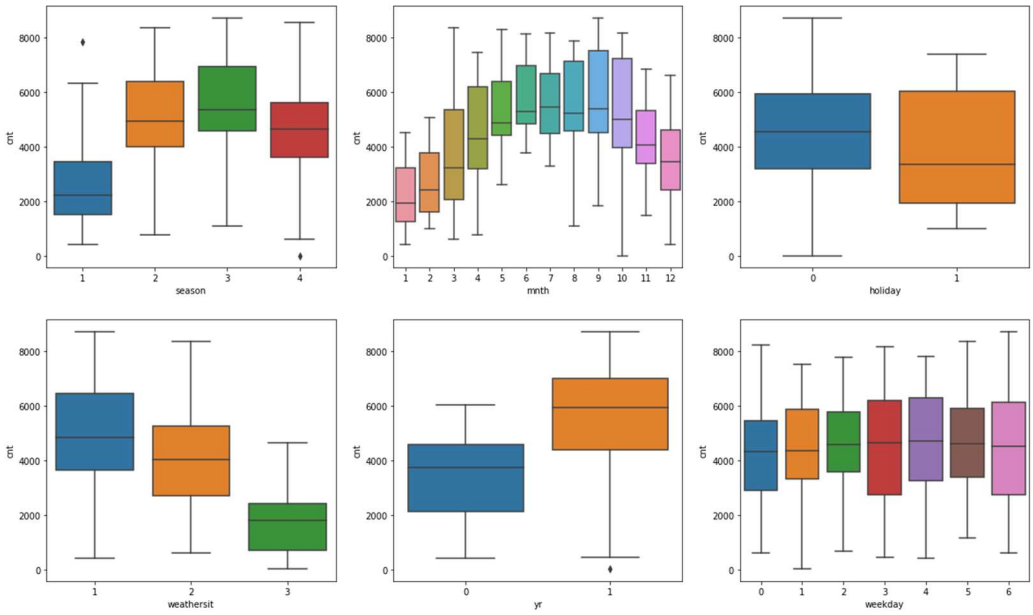
Assignment-based Subjective Questions

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

   Answer:

   Below are the categorical datasets and their effects on dependent variable cnt:



   a. season: During Spring very less bikes have been rented compared to other seasons. However, during fall the bike rental recorded highest counts.
   b. mnth : In January the count of rented bikes is less, however, in September its highest.
   c. Holiday: The median of rentals on holiday is significantly less than that during non-holiday. Hence rentals on holiday is less
   d. Weathersit: Less bikes are rented during light snow and rain whereas highest count has been seen during clear weather situations. No bikes rented during Heavy Snow & Rain.
   e. Yr : The number of rentals in 2019 is more than that of 2018
   f. Weekday:The average rentals on each day is similar with less variation has been observed. Mondays and Fridays has been seen large variance.

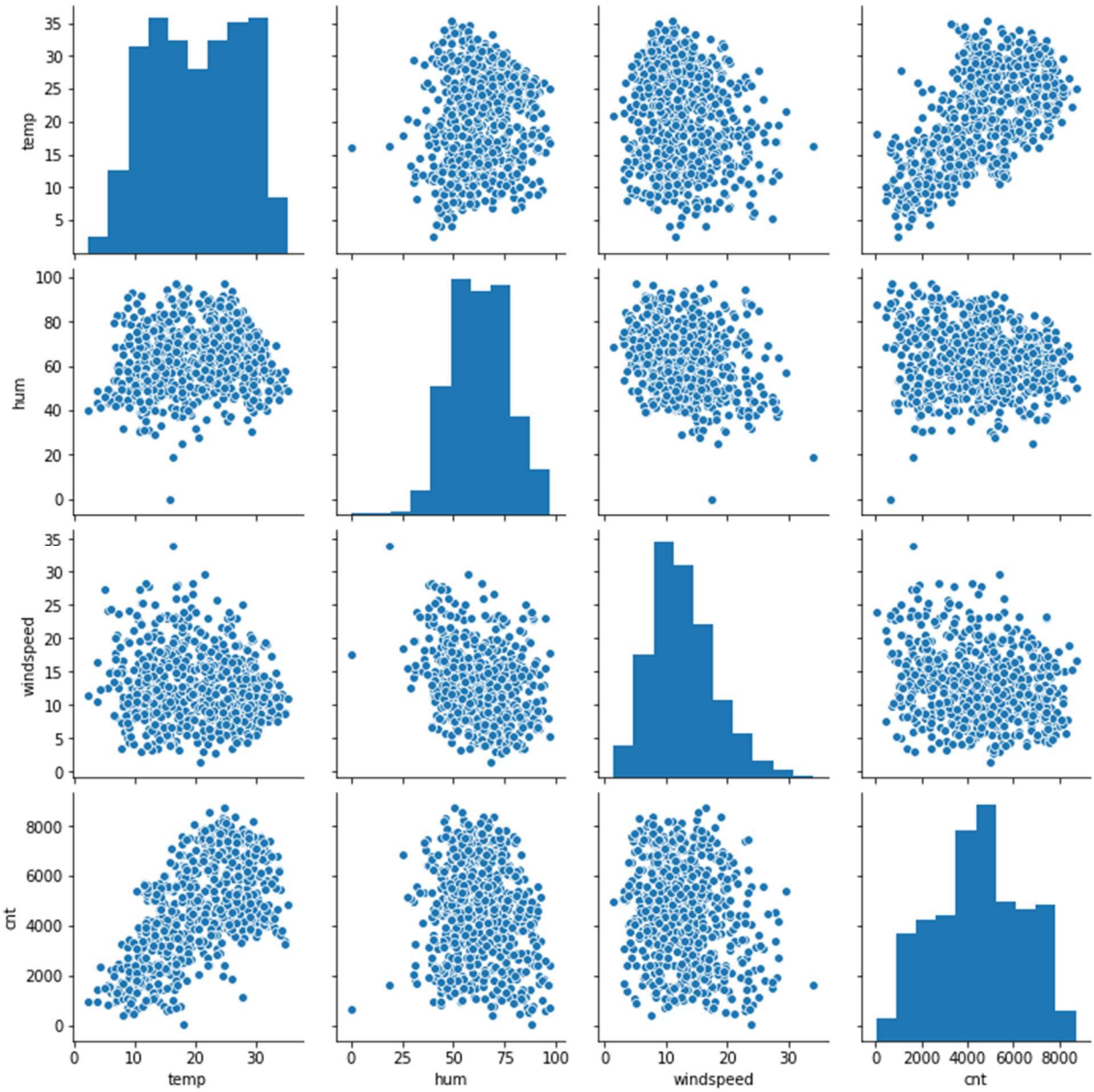
2. ***Why is it important to use drop_first=True during dummy variable creation?***

   When we convert the categorical variables to dummies, indirectly we are giving importance to each value in a categorical column by making each value as a column. why to drop one variable in regression is because the importance or value of that left-over variable can be found by remaining variables. So, to avoid redundancy dropping a column is preferred.

   For example, if there are Red, Blue and green are categorical variables. When we convert these into dummies Red ,blue and green will be the extra columns created. So, if the sample has green value, the Red, Blue will be indicated as Zero. automatically it indicates the sample belongs to green. If the value of green column can be explained by Red and blue column, then there is no need of green column.

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

   Answer: Looking at the pair-plot among numerical variables, **temp** has highest correlation with cnt



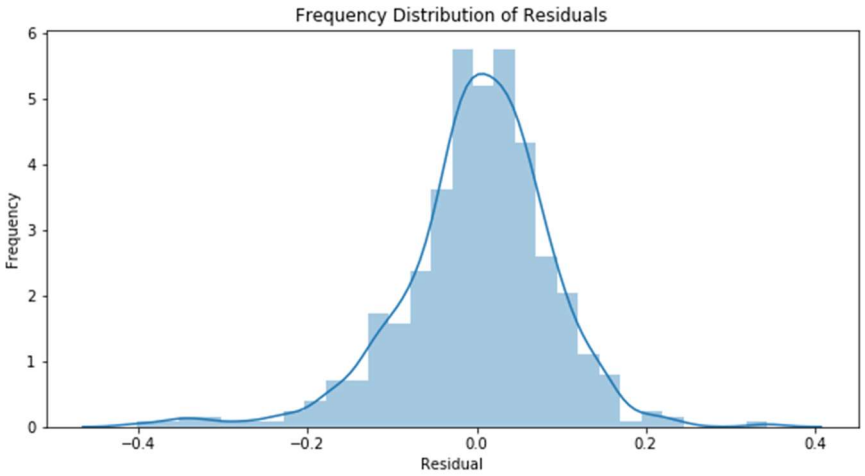
4. ***How did you validate the assumptions of Linear Regression after building the model on the training set?***

   Answer:

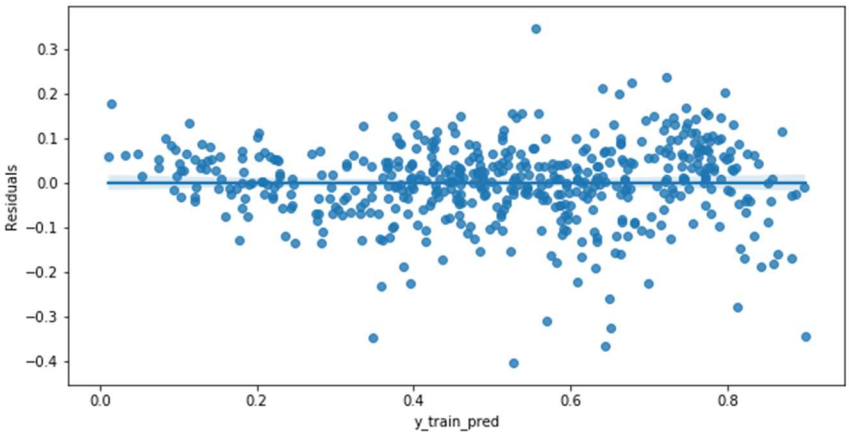
   Assumptions of Linear Regression are as follows:

   1. Error terms are normally distributed with mean zero.
   2. Error terms are independent of each other.
   3. Error terms have constant variance (homoscedasticity).

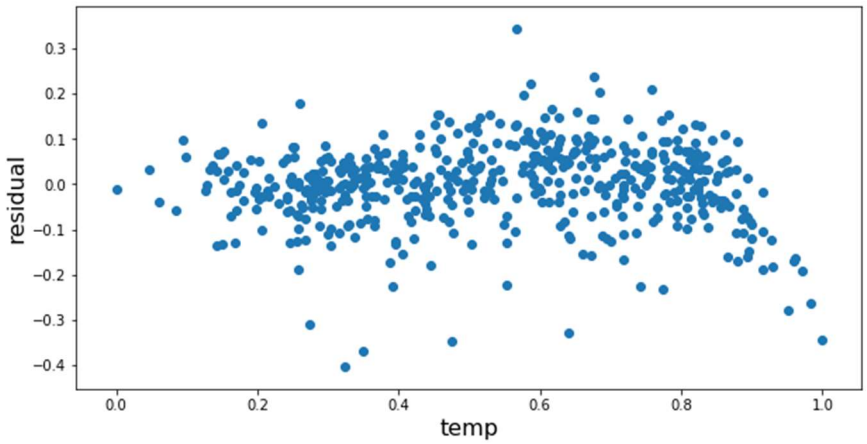
From the below residual distribution graph we can conclude that error terms are normally distributed and mean at zero.



From the below pairplot we can ascertain that error terms are independent of each other



From below plot its evident that Error terms have constant variance (homoscedasticity).



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

   Answer: Below are the coefficients values after building the final model. It's clear that "temp", "weathersit_Light Snow & Rain" and "yr" variable are significantly explaining the demand for the shared bikes.

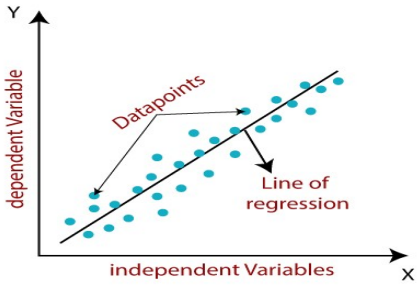
   Coefficients of the variables as below:

   - temp : 0.512675
   - weathersit_Light Snow & Rain : -0.248516
   - yr : 0.230700

index	Variables	Coefficient value
10	temp	0.512675
8	yr	0.230700
0	const	0.216679
3	season_Winter	0.111924
4	mnth_Sep	0.094023
2	season_Summer	0.068717
5	weekday_Monday	0.062701
9	workingday	0.053223
1	season_Spring	-0.044772
7	weathersit_Mist & Cloudy	-0.057310
11	hum	-0.150823
12	windspeed	-0.179875
6	weathersit_Light Snow & Rain	-0.248516

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Answer:
   Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). It is a statistical approach for modelling relationship between a dependent variable with a given set of independent variables



Regression is broadly divided into below types:

A. Simple Linear Regression:

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

Mathematically, we can represent a linear regression as:

$$y= B_0+B_1x+ \varepsilon$$

B. Multiple Linear regression

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.
Mathematically, we can represent a Multiple linear regression as:

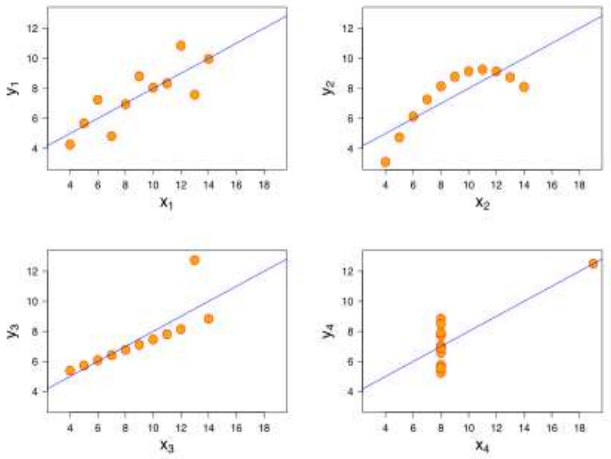
$$y= B_0+B_1x_1+B_2X_2+B_3X_3 + \ldots\ldots\ldots + B_pX_p + \varepsilon$$

Here,

y= Dependent Variable (Target Variable)
x= Independent Variable (predictor Variable)
B_0= intercept of the line (Gives an additional degree of freedom)
B_p = Linear regression coefficient (scale factor to each input value).
ε = random errors

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties



Anscombe's quartet

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

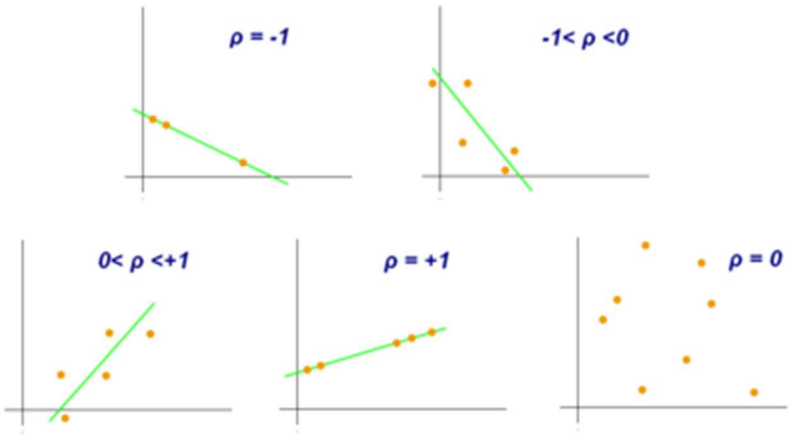
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : σ^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : σ^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R?

Pearson's R is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Below graph represents examples of scatter diagrams with different values of correlation coefficient



4. ***What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?***

What is it? - Scaling is a technique to standardize the independent features present in the data in a fixed range.

Why? - Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use distance between two data points in their computations, this is a problem. If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.
Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Techniques to perform Feature Scaling

a. **Min-Max Normalization**: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

b. **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

VIF is variance inflation factor.
The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.

$$(VIF) = 1/(1-R^2)$$

- If the independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.
  So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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

What is it? – Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly 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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

- It is used to check following scenarios:

If two data sets —

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behavior