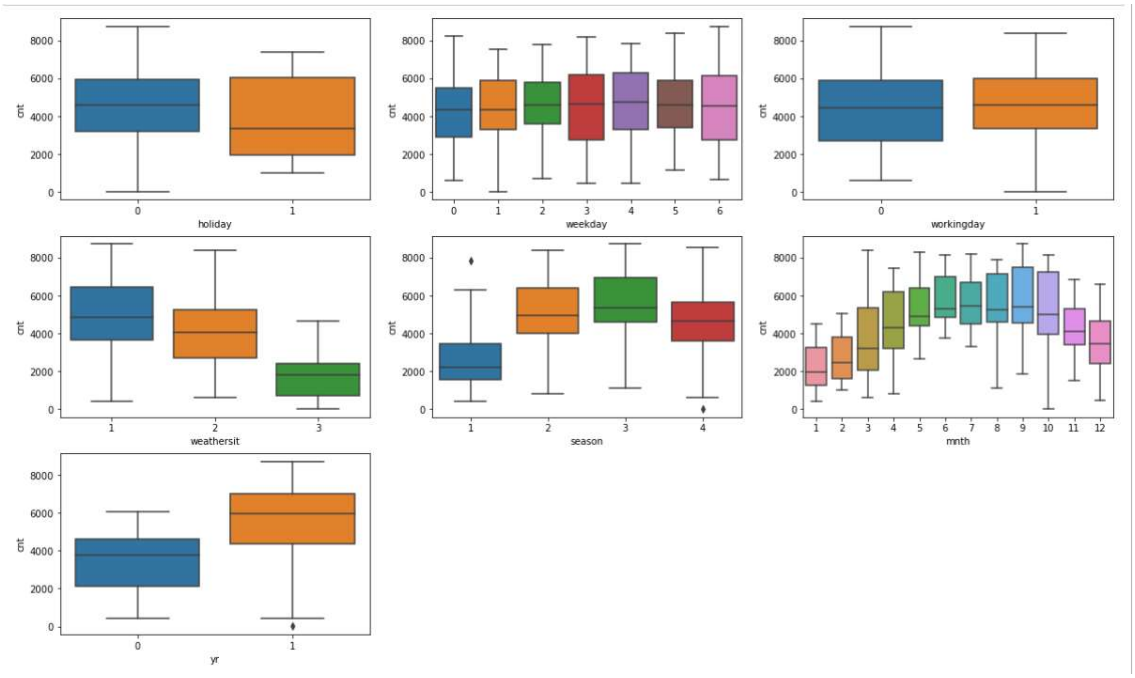
Assignment based subjective questions

1. From your analysis of the categorical variables from the dataset , what could you infer about their effect on dependent variable?
   Dependent variable is count which tells count of total rental bikes including both casual and registered
   Effect of categorical variables on dependent variable(count) can be analyzed based on box plot



   a. **Holiday** : mean count of non holiday were more than holidays
   b. **Weekday**: On weekday there is no significant difference in mean count
   c. **Working day**: Mean count of working days count were equal to non working days (Hence we can infer people were using bikes for spending non working days as well). Lower quartile (25th percentile of working days count is more than count of non working days)
   d. **Weathersit**:
      a. When the weather is 4 (heavy rain , thunderstorm , mist , fog) , there is no count observed . Hence people would not have used bike for transport when weather is 4.
      b. There is significant dip in mean count observed when we move from weather 1(clear) to 3(light snow, light rain) . Reason being when the weather is clear more people opted to travel in bike whereas when the weather moves from 1 to 2(mist, cloud) to 3 , there is less people who prefer to travel in bike due to climate change and increase in mist, light snow etc
   e. **Season**:
      a. There is significant increase in mean count when the season is moving from 1, 2, 3 (1:spring, 2:summer, 3:fall, 4:winter)
      b. After 3 , there is dip in mean count when the season moves towards 4 - winter where the climate would be mist and snowy difficult to travel using bike
   f. **Month**:
      a. From Month 1 , mean count started to increase gradually
      b. There is significant increase in mean count from month 1 to month 7

c. From Month 7 to 9 , mean count is steady
d. From month 9 to 12, mean count started to dip and gradually lowering by December

g. **Yr**:
   a. 2018 vs 2019 comparision: There is remarkable increase in count of rental bikes used if we compare 2018 count vs 2019 count
   b. It is notable to mention that upper quartiles (75th percentile count) of 2018 is same as lower quartile of (25th percentile count) of 2019 which is a great progress
   c. This makes us to infer that rental bikes company were performing more better year by year

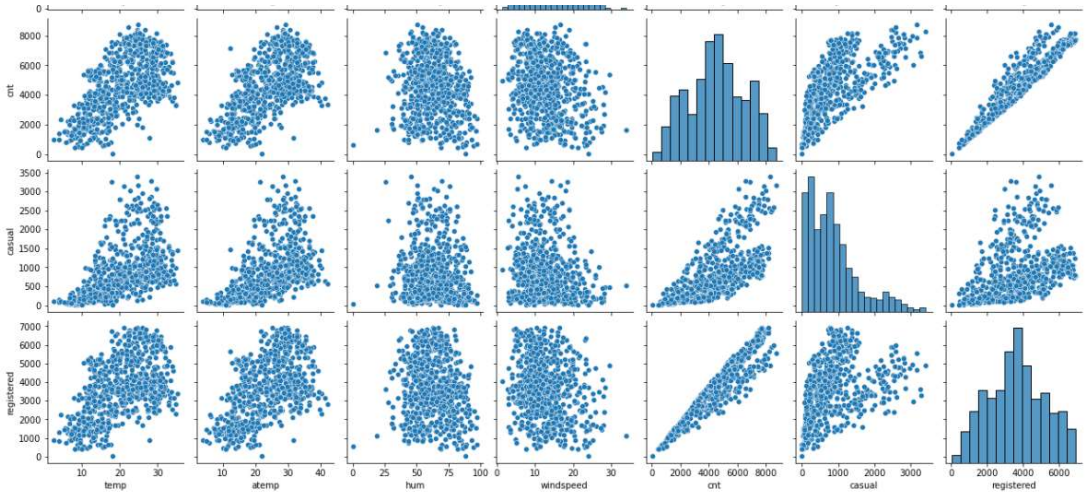
2. Why is it important to use the drop_first = True during dummy variable creation?
   a. It is important to use drop_first=True as it would end up in helping to avoid creation of 1 columns which is not required for our analysis.
   b. Without that additional 1 column we could infer the presence of other column by seeing the binary values of rest other columns
      Eg for categorical variable Gender , it can be referred with only 1 column That is Male 1 means Male, Male 0 means it is Female .

Gender
Male
Female

Male
1
0

   c. With that 1 columns not being added for each categorical variable, model performance would be better with reduced number of columns . easier for analysis etc
   d. Assume if we have 5 categorical variables, with this drop_first = True would help us in unnecessary 5 columns being added in features, its prediction time etc .
   e. Also with the binary values even without this additional columns we can infer the values based on values of other columns

3. Looking at the pair-plot among numerical variables which one has highest co-relation with target variable
   By looking at the pair-plot 'registered' has highest co-relation with target variable. See the first row 7th columns where the graph shows linearity between count and registered



4. How did you validate the assumptions of Linear regression after building the model on training set?
   We are validating the assumptions of Linear regression after building the model on training set by using doing residual Analysis of the train data, by checking the below
   a. error terms are normally distributed with mean = 0
   b. Plot this histogram of error terms to check the normal distribution
   c. Error terms should be independent of each other

   Qualitative analysis of below plt: Centre should be zero and shape should be like a bell curve

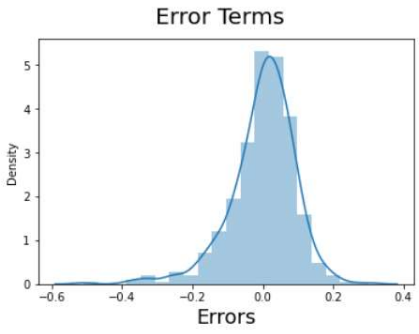
Residual Analysis of the train data

So, now to check if the error terms are also normally distributed (which is infact, one of the major assumptions of linear regression), let us plot the histogram of the error terms and see what it looks like.

```
y_train_pred = lm6.predict(X_train_new1)
```

```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_pred), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)                         # X-Label
```

```
Text(0.5, 0, 'Errors')
```



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

Temperature – temperature in Celsius
Year – Year 2019
September/october – When the month is September/October or winter season

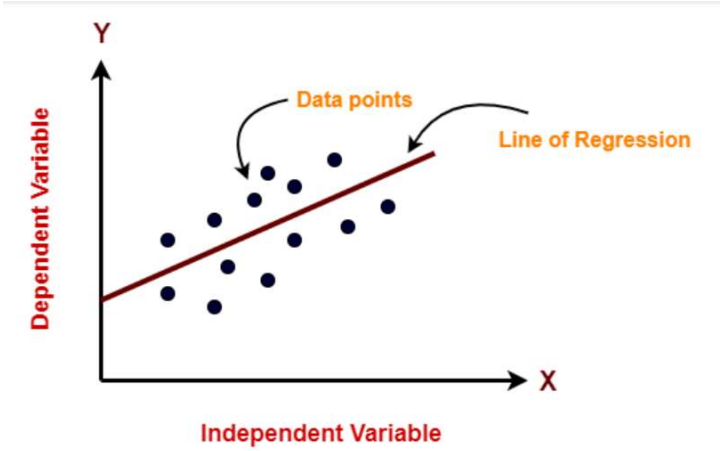
General subjective questions:

1. Explain the linear regression algorithm in detail

Linear regression algorithm is a machine learning algorithm used for supervised learning .
It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematical representation of linear regression

$y= a_0+a_1x+ \varepsilon$

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error
The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression:

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- **Multiple Linear regression:**

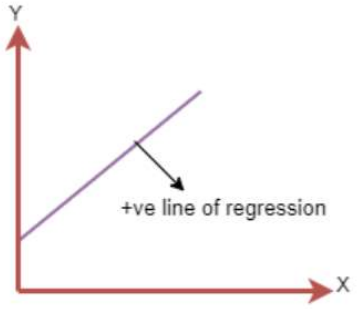
If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

A linear line showing the relationship between the dependent and independent variables is called a regression line. A regression line can show two types of relationship:

Positive Linear Relationship:

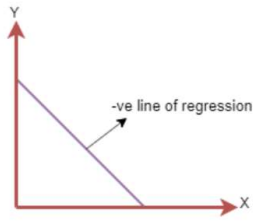
If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship



The line of equation will be: $Y= -a_0+a_1x$

Simple Linear regression equation :

$$Y = \beta_0 + \beta_1.X$$

Multiple Linear regression equation

- Ideal Equation of MLR

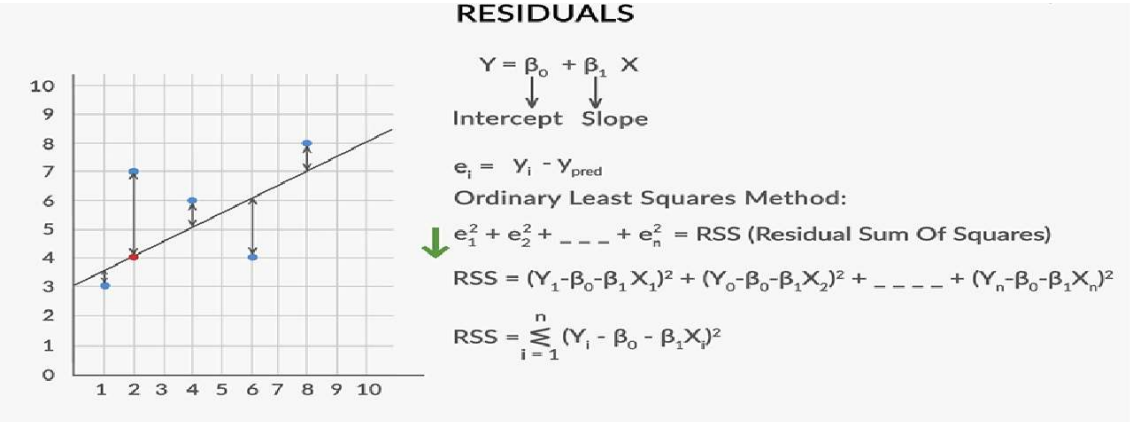
$$\hat{Y} = \hat{\beta}_0+\hat{\beta}_1\times x_1+\hat{\beta}_2\times x_2+\hat{\beta}_3\times x_n .... \hat{\beta}_n\times x_n$$

Finding the best fit line:

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (B1, B2) gives a different line of regression, so we need to calculate the best values for B1 and B2 to find the best fit line, so to calculate this we using various different approaches one of which is ordinary least squares method

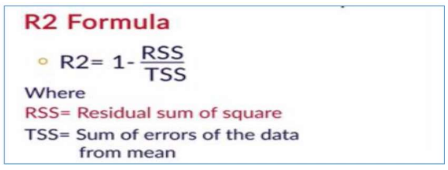
We need to find B0 and B1 in a way which has reduced error as mentioned below



Strength of linear regression can be assessed using 2 ways

1. R squared
2. Residual standard error

After we build the model , we will assess by above 2 metrics



Where RSS is total sum of error across the whole sample It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

TSS is TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is

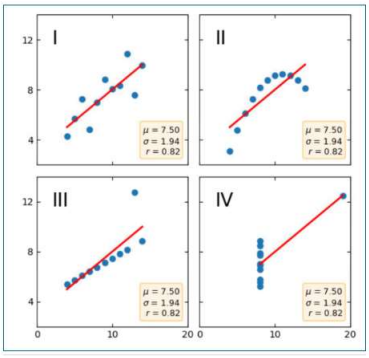
$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

You can now check the mean square error and r square value of your model.

Your model is getting a mean square error of XX which means the model is not able to match X% of the values only, which is good.
The r square value is about Y% which means our model is able to explain Y% of the variance which is also good

2. What is the Anscombe's quartet in details

   Anscombe's Quartet was devised by the statistician Francis Anscombe to illustrate how important it was to not just rely on statistical measures when analyzing data. To do this he created 4 data sets which would produce nearly identical statistical measures.



Statistical measures

1) Mean of x values in each data set = 9.00
2) Standard deviation of x values in each data set = 3.32
3) Mean of y values in each data set = 7.50
4) Standard deviation of x values in each data set = 2.03
5) Pearson's Correlation coefficient for each paired data set = 0.82
6) Linear regression line for each paired data set: y = 0.500x + 3.00

When looking at this data we would be forgiven for concluding that these data sets must be very similar – but really they are quite different.

Data Set A:

x = [10,8,13,9,11,14,6,4,12,7,5]

y = [8.04, 6.95,7.58,8.81,8.33, 9.96,7.24,4.26,10.84,4.82,5.68]

Data Set B:

x = [10,8,13,9,11,14,6,4,12,7,5]

y = [9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74]

Data Set C: (few outliers were there)

x = [10,8,13,9,11,14,6,4,12,7,5]

y = [7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73]

Data Set D: (not fit for linear regression)

x = [8,8,8,8,8,8,8,19,8,8,8]

y = [6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.50,5.56,7.91,6.89]

The moral of the story

So – the moral here is always use graphical analysis alongside statistical measures. A very common mistake would be rely on Pearson's Product coefficient without really looking at the scatter graph to decide whether a linear fit is appropriate. If you do this then you could end up trying to fit a regression line for data point which is not a proper data point for fitting linear regression

3. What is Pearson's R?

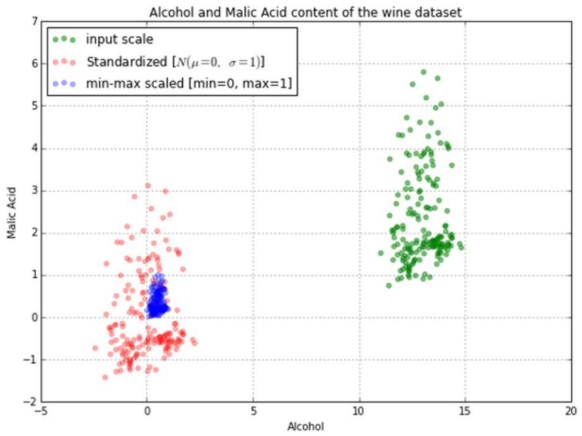
   The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (r)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Weight and obesity When the persons weight increases obesity increases
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	While we travel if our travel speed increases , time to destination will decrease

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

   Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.
   Eg: if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range, for example- centered around 0 or in the range (0,1) depending on the scaling technique

   In order to visualize the above, let us take an example of the independent variables of alcohol and Malic Acid content in the wine dataset from the "Wine Dataset" that is deposited on the UCI machine learning repository. Below you can see the impact of the two most common scaling techniques (Normalization and Standardization) on the dataset.



Why is scaling performed:

It helps to putting the feature values into the same range which make more easier for understanding in common scale .

What is the difference between normalized scaling and standardized scaling:

Normalized scaling : it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1].

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

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

Standardized scaling:
Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and $\bar{x}$ is the average of the feature vector

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

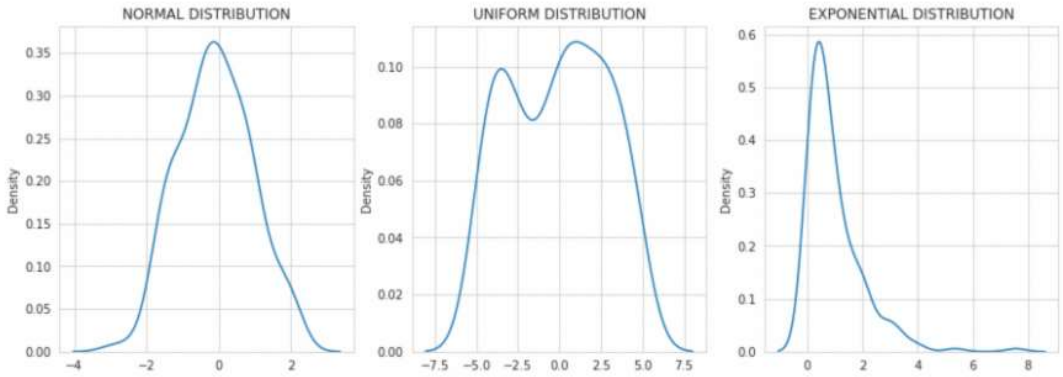
   If there is perfect correlation between variables , then VIF = infinity

   Infinity 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

VIF	Conclusion
1	No multicollinearity
4 – 5	Moderate
10 or greater	Severe

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

   Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. Before we dive into the Q-Q plot, let's discuss some of the probability distributions.



   In probability distributions, we represent data using charts where the x-axis represents the possible values of the sample and the y-axis represents the probability of occurrence.

   There are various probability distribution types like Gaussian or Normal Distribution, Uniform distribution, Exponential distribution, Binomial distribution, etc.

   Normal distributions are the most popular ones. They are a probability distribution that peaks at the middle and decreases at the end of the axis. It is also known as a bell curve or Gaussian Distribution. As normal distributions are central to most algorithms, we will discuss this in detail below.

   Uniform distribution is a probability distribution type where the probability of occurrence of x is constant. For instance, if you throw a dice, the probability of any number is uniform.

   Exponential distributions are the ones in which an event occurs continuously and independently at a constant rate. It is commonly used to measure the expected time for an event to occur