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

Answer:

We have below categorical variable:

(1) **Season: ‘**summer’ and ‘fall’ has higher number of bookings comparing to other season.

(2) **Weekday:** does not have great impact on bike booking as its boxplot showing overlapping area**.**

(3) **Weathersit:** clear weather has higher number of booking. light rain has low no of booking.

(4) **Month:** Jan to oct bookings either increasing or constant in range but Nov and dec and Jan saw down trend in booking.

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

Answer: For categorical We need to create dummy variable for example below:

season: - summer, winter, rainy

to use those, we have to create dummy variable like below:

Season
Summer
Winter
rainy

summer	Winter	Rainy
1	0	0
0	1	0
0	0	1

Now this can be also be same if we drop first column

Winter	Rainy
0	0
1	0
0	1

0 0 => gives same meaning as 1,0,0 . by this we are able to reduce columns and complexity of model .

So **drop_first=True does same thing and drop first column.**

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

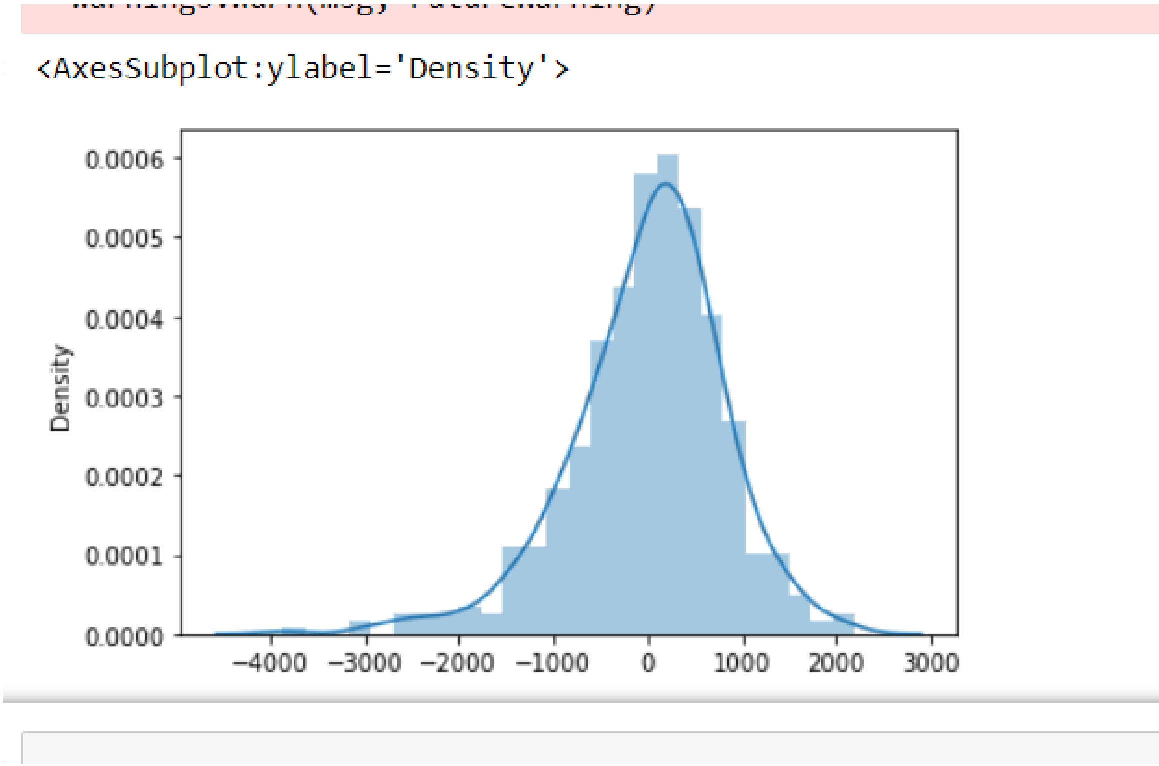
   Answer:
   By looking pair plot of numerical variables, we found **atemp** => feeling temperature in Celsius and **temp=> temperature**   has highest correlation between **cnt=>** count of total rental bike.

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

   Assumptions:

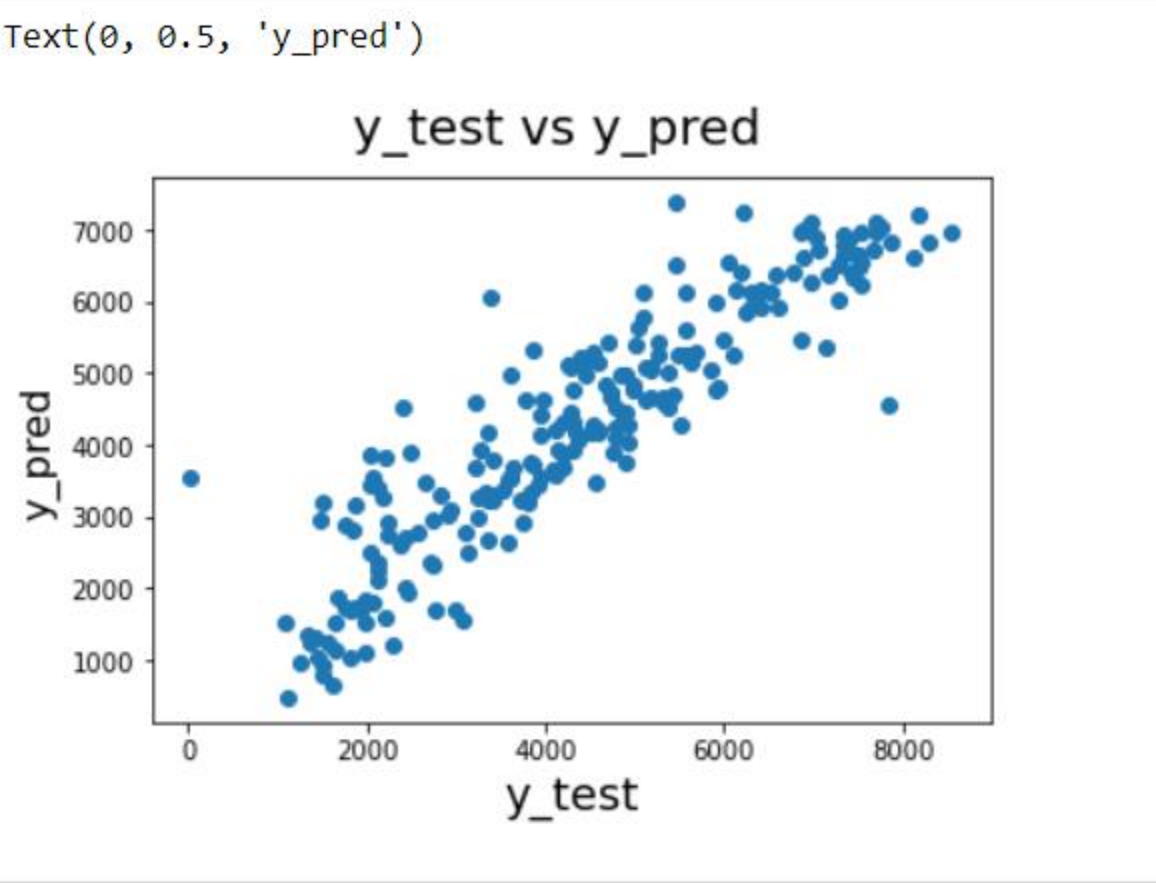
   (i) Dependent variable and independent variable should have liner relationship.
   We checked it with creating pair plot and verified some columns has liner relationship.

   (ii) Error term are normally distributed with mean zero.
   We checked it by creating distribution plot of residual error and found its normally distributed.



```
  warnings.warn(msg, FutureWarning)
<AxesSubplot:ylabel='Density'>
```

   (iii) Plot between actual Y and predicted Y should be constant variance.
   We plotted graph between **y** and **y predicted** and its constant variance.



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

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:

   (1) feel like temperature
   (2) light rain
   (3) windspeed
   (4) spring season
   (5) December month

* **General Subjective Questions**

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Answer: Linear regression is approach for modelling relationship between dependent and independent variables. case of one dependent variable called simple linear regression and multiple dependent variables called multiple linear regression.

Linear Predictor function: used for modelling relationship between dependent and independent variable.

The basic form of linear predictor function:
F(i)=b0 +b1X1+b2X2+Bi Xi…

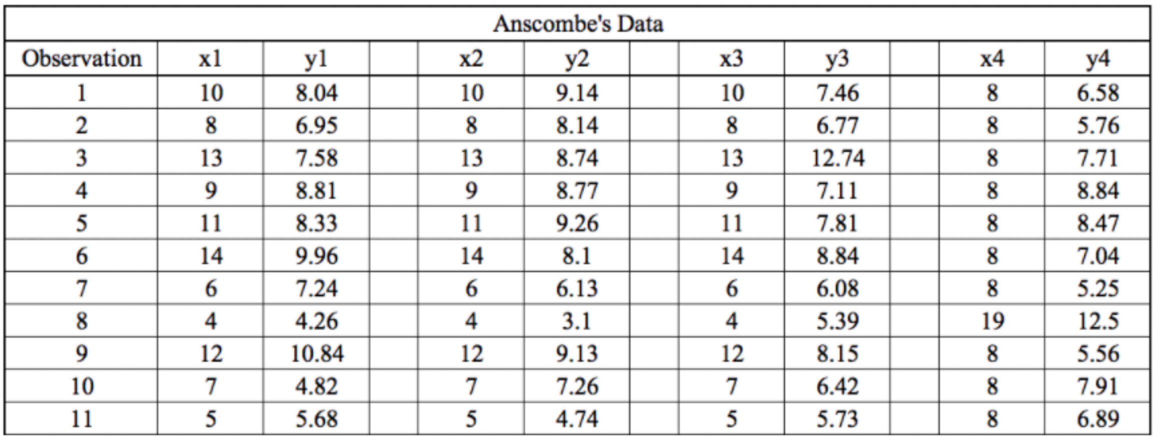
Linear regression model representation:
Ordinary least squares: this representation seeks to minimize the sum of residual squared. means that given a representation line   through data we calculate the distance from each data point to the regression line, square it and sum all squared error together. this should be minimized.

Gradient distance:
In this method we start by random values for each coefficient. This sum of squared errors calculated for each pair of input and output values. A learning rate is used to control convergence. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

Prediction on trained model:
After training of model with above method we get coefficient predicted value, which can be used to predict on independent variables

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   Answer: Anscombe's quartet can be defined as a group of four datasets which are nearly identical in simple descriptive statistics, but there are some peculiarities in dataset that fool the regression model if built. They have different distribution and appear differently when plotted on scatter plot.

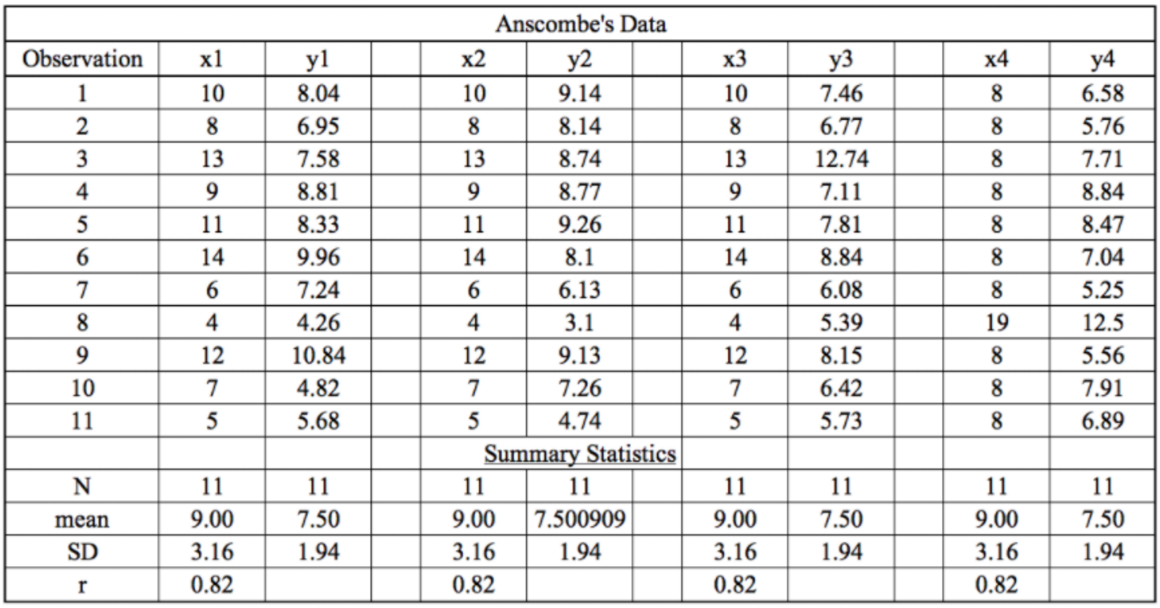
   This tells us importance of visualising data before applying any algorithm.



Anscombe's Data

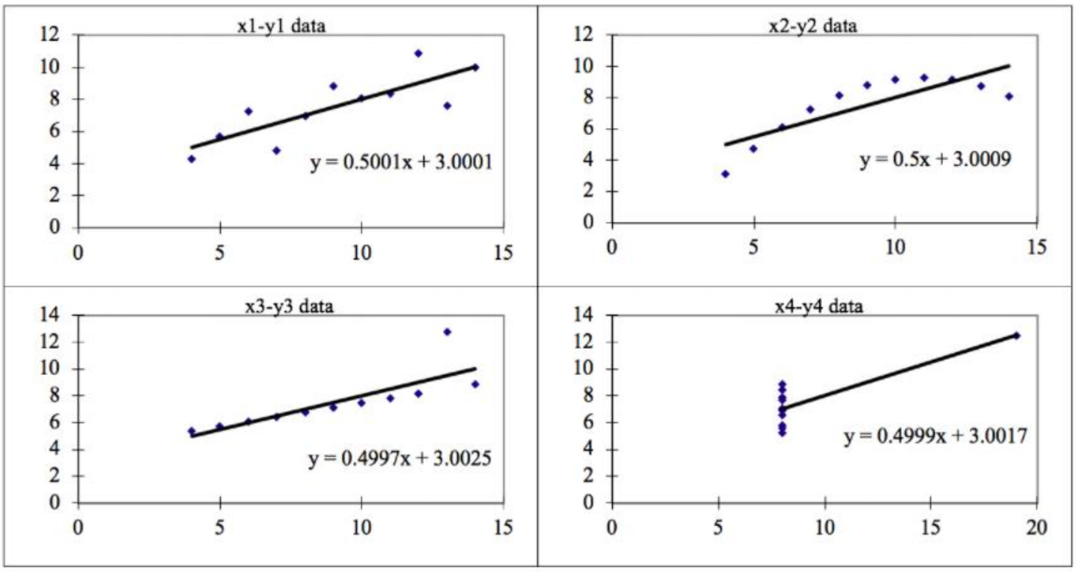
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	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

Statistical information



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	



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

Answer: Pearson's R is a correlation coefficient commonly used in linear regression. Pearson 's r is commonly used in "linear regression ". It is a measure of of linear correlation between two set of data. It is ratio between covariance of two variables and the product of their standard deviations it is normalized measurement of the covariance such that value always has a value -1 to 1.

Its formula is: **r(x,y) =cov(x,y)/ sigma(x) * sigma(y)**

Where r(x,y) is correlation coefficient
Cov(x,y)=> covariance .

Sigma(x) is standard deviation of x.
Sigma(y) is standard deviation of y.

where **Cov(x,y)=>E(X-ux)(y-uy)]**

4**. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
Answer: It is one step to applied on independent variable to normalize data into within particular range. It helps to speed up calculation in learning of model.

independent variables can be in different range with high magnitude to low, units and range also. If we don't do scaling, algorithm learn on magnitude not on values.

Normalized scaling: It bring all data from range of 0 to 1.
Minmax scaling: x- min(x)/max(x)-min(x)

standardized scaling: it replaces the value to Z score .it bring all data into a standard normal distribution with mean(mu) and standard deviation (sigma).
Standardization x:= 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: In case of perfect correlation the VIF =infinity. because we get R2=1 in case of perfect correlation:
So vif=1/1/r2 = 1/1-1 = infinity

To solve this issue, we should drop one of the variables causing 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: Q-Q plot is a graphical tool to help to do analysis if set of data came from same theoretical distribution for example Normal, exponential etc. we can also use this to check if two datasets came from population of common distribution.

A Q-Q plot is a plot of quantile of first dataset against second dataset.
Possible interpretation:
(a) If all point of quantile lies or close to straight line at angle of 45 degree.
(b) If y values <x values
(c) If y values >x values

(d) Different distribution: if all point is away from the straight line of 45 degree. It means it is different distribution.

Usage in linear regression:
It can be used in linear regression to check if our training and test set are from same population distribution or not.

Many distributional aspects like shift in location, shift in scale and change symmetry and presence of outlier can be detected by this.