

## 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:**

- The demand of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non working day.
- The bike demand is high when weather is clear and Few clouds however .

2. Why is it important to use drop\_first=True during dummy variable creation?

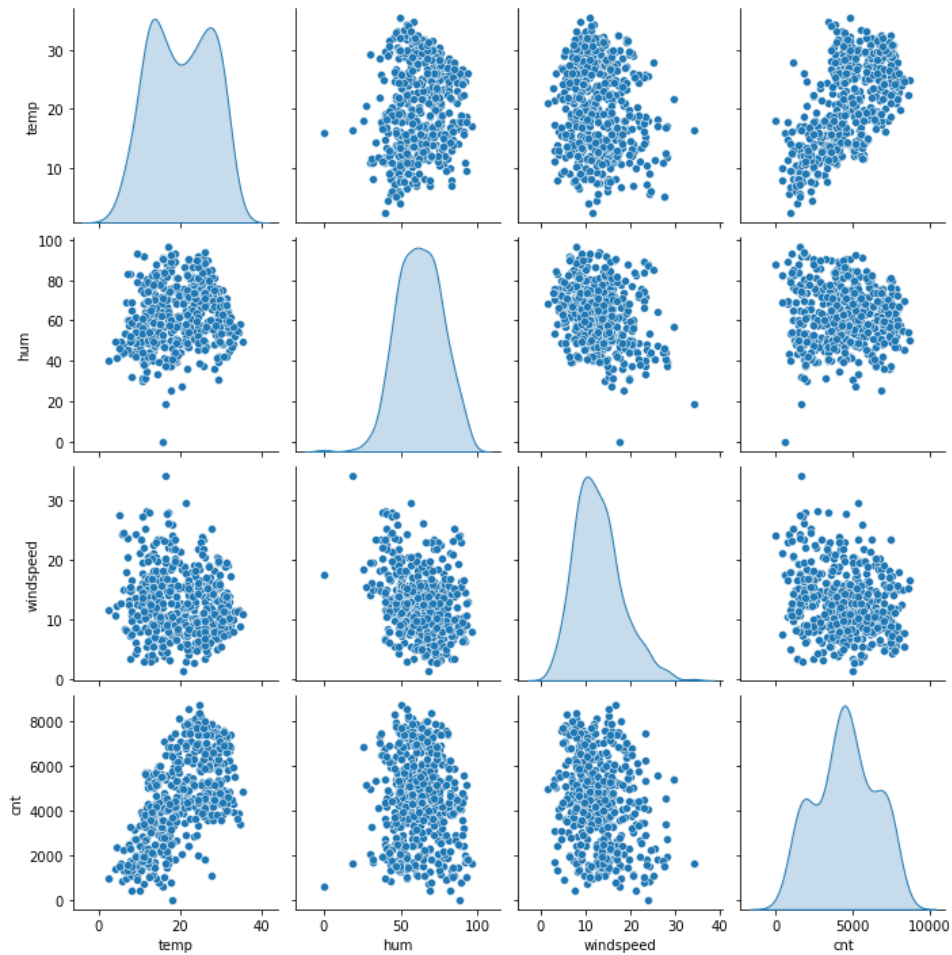
**Answer:**

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

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

**Answer:**

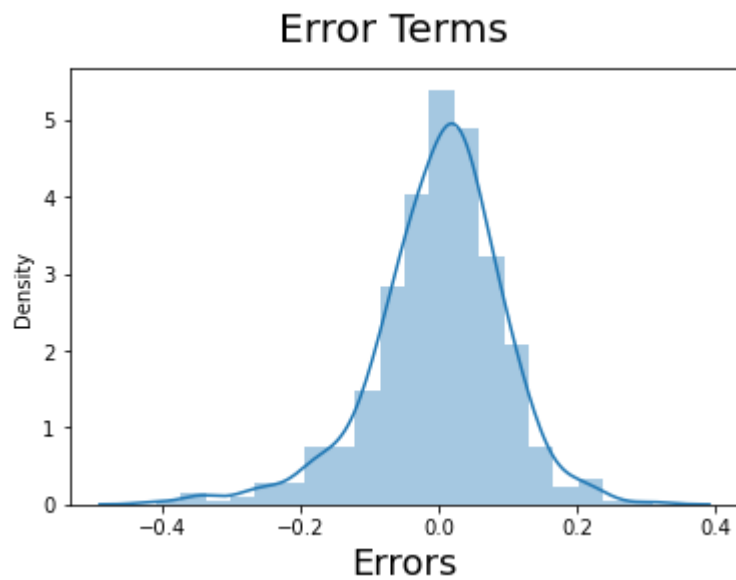
- Target variable (count) has significantly high correlation with temperature (temp).



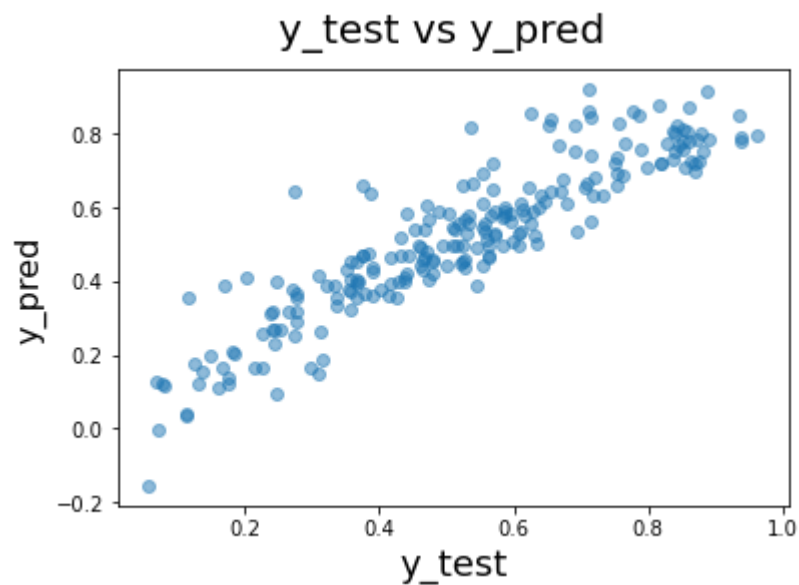
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

**Answer:**

- Residual errors follow Normal distribution.



- Maintains Linear relation between dependent variable (Test and Predicted)



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

**Answer:**

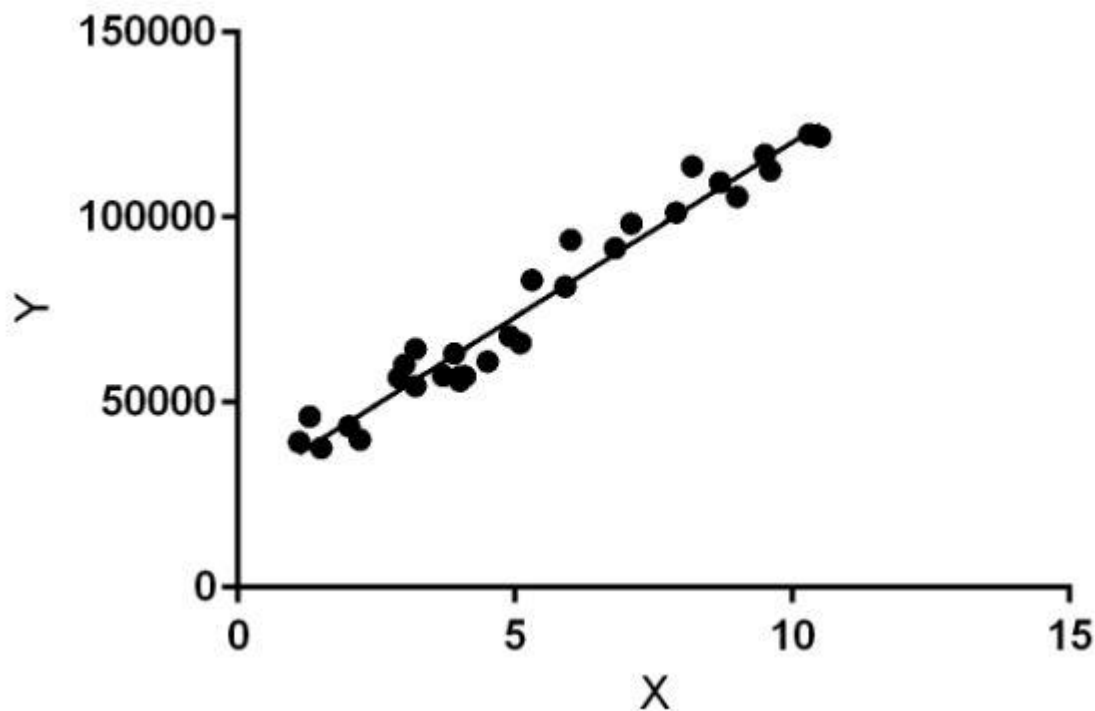
1. Temperature (0.5636)
2. season: winter (0.128744)
3. year (0.230846)

## General Subjective Questions

1. Explain the linear regression algorithm in detail.

**Answer:**

**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.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

## Hypothesis function for Linear Regression :

$$Y = \theta_1 + \theta_2 \cdot x$$

While training the model we are given :

x: input training data (univariate – one input variable(parameter))

y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best  $\theta_1$  and  $\theta_2$  values.

$\theta_1$ : intercept

$\theta_2$ : coefficient of x

Once we find the best  $\theta_1$  and  $\theta_2$  values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

**We need to update  $\theta_1$  and  $\theta_2$  values to get the best fit line. It is as follows:-**

### Cost Function (J):

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the  $\theta_1$  and  $\theta_2$  values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

$$\text{minimize } \frac{1}{n} \sum_{i=1}^n (\text{pred}_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^n (\text{pred}_i - y_i)^2$$

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

### Gradient Descent:

To update  $\theta_1$  and  $\theta_2$  values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random  $\theta_1$  and  $\theta_2$  values and then iteratively updating the values, reaching minimum cost.

## 2. Explain the Anscombe's quartet in detail.

**Answer:**

According to the definition given in [Wikipedia](#), **Anscombe's quartet** comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

**Simple understanding:**

**Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.**

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

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

Please download the csv file [here](#).

**Code: Python program to find mean, standard deviation, and the correlation between x and y**

```
# Import the required libraries
import pandas as pd
import statistics
from scipy.stats import pearsonr

# Import the csv file
df = pd.read_csv("anscombe.csv")

# Convert pandas dataframe into pandas series
list1 = df['x1']
list2 = df['y1']

# Calculating mean for x1
```

```

print('%0.1f' % statistics.mean(list1))

# Calculating standard deviation for x1
print('%0.2f' % statistics.stdev(list1))

# Calculating mean for y1
print('%0.1f' % statistics.mean(list2))

# Calculating standard deviation for y1
print('%0.2f' % statistics.stdev(list2))

# Calculating pearson correlation
corr, _ = pearsonr(list1, list2)
print('%0.3f' % corr)

# Similarly calculate for the other 3 samples

# This code is contributed by Amiya Rout

```

**Output:**

```

9.0
3.32
7.5
2.03
0.816

```

So let me show you the result in a tabular fashion for better understanding.

Summary						
Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)	
1	9	3.32	7.5	2.03	0.816	
2	9	3.32	7.5	2.03	0.816	
3	9	3.32	7.5	2.03	0.816	
4	9	3.32	7.5	2.03	0.817	

**Code: Python program to plot scatter plot**

```

# Import the required libraries
from matplotlib import pyplot as plt
import pandas as pd

# Import the csv file
df = pd.read_csv("anscombe.csv")

# Convert pandas dataframe into pandas series
list1 = df['x1']
list2 = df['y1']

# Function to plot scatter

```



```
plt.scatter(list1, list2)

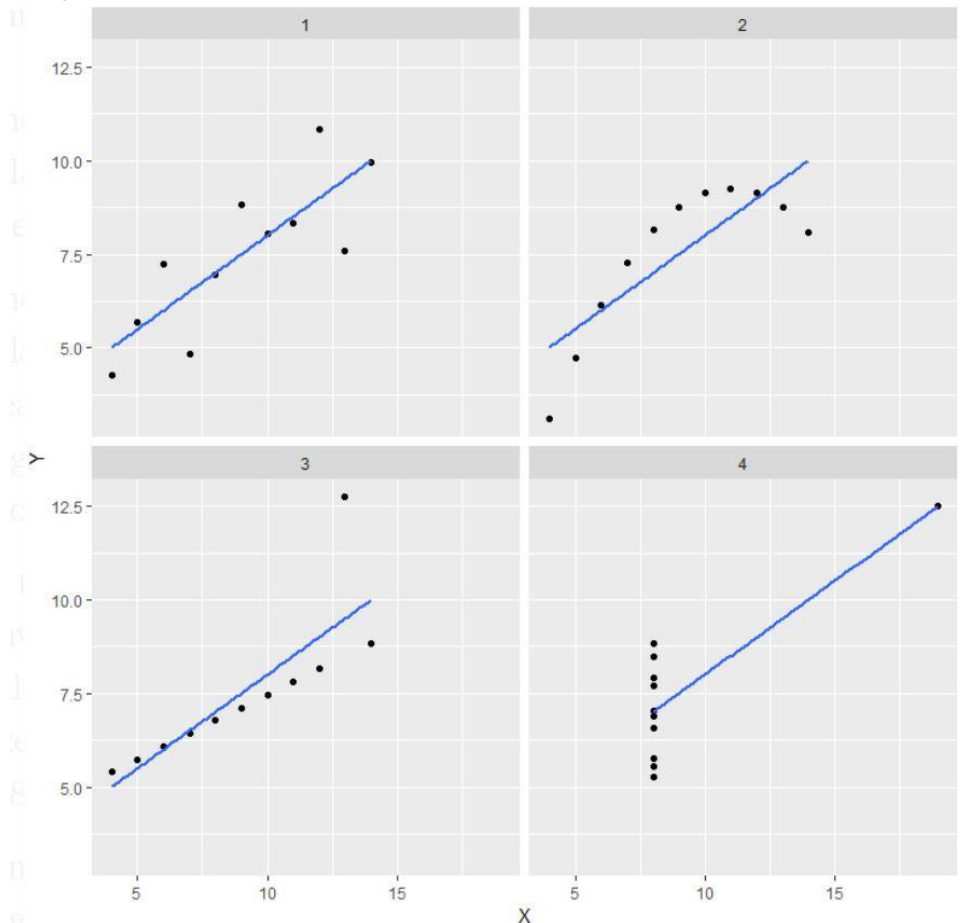
# Function to show the plot
plt.show()

# Similarly plot scatter plot for other 3 data sets

# This code is contributed by Amiya Rout
```

For regression line refer [this](#).

Output:



Note: It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

### Explanation of this output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.

- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

**Application:**

**The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.**

### 3. What is Pearson's R?

**Answer:**

According to Wikipedia, in statistics, the **Pearson's  $r$**  is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between  $-1$  and  $1$ . As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than  $0$ , but less than  $1$  (as  $1$  would represent an unrealistically perfect correlation).

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

Answer:

**What?**

- *It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.*

**Why?**

- *Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.*
- *It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.*

**Normalization/Min-Max Scaling:**

- *It brings all of the data in the range of 0 and 1.  
**sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.*

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

### Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean ( $\mu$ ) zero and standard deviation one ( $\sigma$ ).

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

- **`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**.

### Example:

Below shows example of Standardized and Normalized scaling on original values.



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

**Answer:**

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

**Answer:**

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
- A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

## **Links referred:-**

[www.geeksforgeeks.org](http://www.geeksforgeeks.org)

[www.wikipedia.org](http://www.wikipedia.org)

<https://medium.com>

<https://www.itl.nist.gov/>

**Thank You**