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

- When it is heavily raining/snowing there is no demand for shared bike
- During Spring season, demand for bike is less
- Demand for shared bike is high in working days
- Demand is high in the month of June, July, August and September
- There is no significant difference in demand for bike throughout the weekdays
- Demand for shared bike is increasing every year (though the data is for only 2 years)

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

- It reduces the extra column created while creating dummy variables
- For eg. In the bike sharing data, there are 4 seasons.

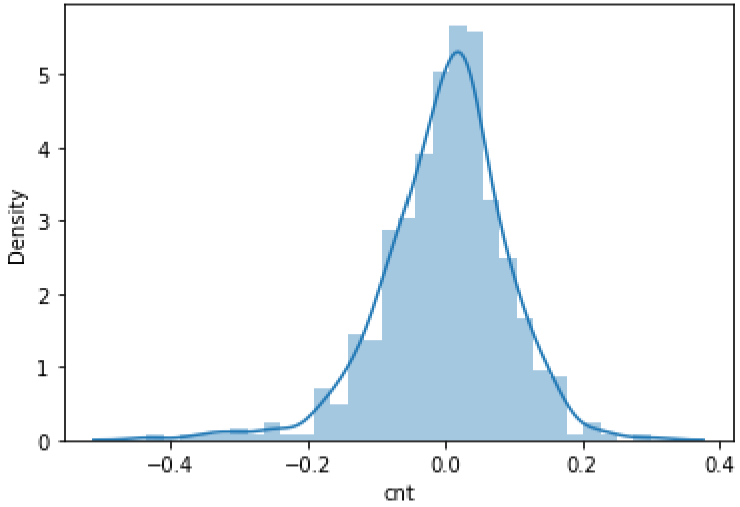
When we create dummy variables for season, it will create 3 dummy variables

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

- Temp/atemp has the highest correlation(0.63) with target variable(cnt)

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

- When we draw distplot for y_train and y_train_pred, we can see that errors are distributed normally around 0.0. This indicates that assumptions of linear regression holds true



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

- **Temp (Positive correlation)**
- **Yr (Positive correlation)**
- **Light Snow/Rain (Negative correlation)**

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical regression method used for predictive analysis and to show the relationship between the continuous variables. It shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis)

For eg. If CGPA is given and we need to predict the GRE score then, CGPA will be independent variable which will be represented in X axis and GRE score will be dependent variable which will be represented in Y axis.

In this case, since there is ONLY ONE independent variable, it will be called as simple linear regression. If there were multiple independent variables then it would be called as multiple linear regression.

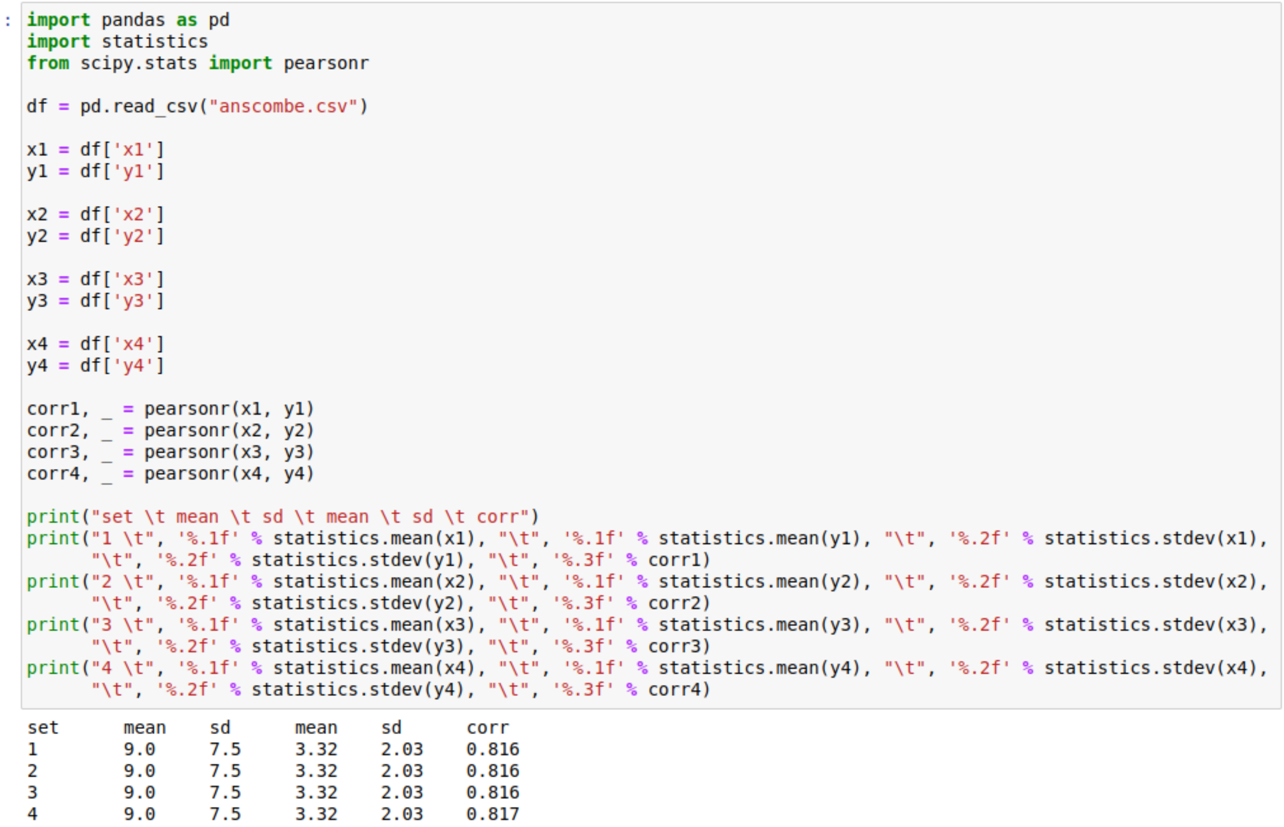
- It will be mathematically represented as

$$\boxed{Y = f(x)}$$

Where Y is dependent variable, that is GRE score and x is independent variable, that is CGPA

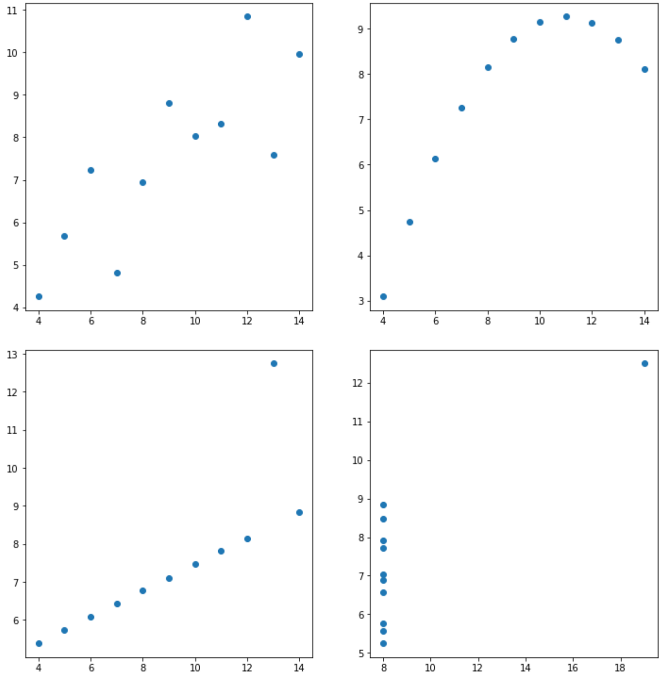
2. Explain the Anscombe's quartet in detail.

- 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.



```
import pandas as pd
import statistics
from scipy.stats import pearsonr

df = pd.read_csv("anscombe.csv")

x1 = df['x1']
y1 = df['y1']

x2 = df['x2']
y2 = df['y2']

x3 = df['x3']
y3 = df['y3']

x4 = df['x4']
y4 = df['y4']

corr1, _ = pearsonr(x1, y1)
corr2, _ = pearsonr(x2, y2)
corr3, _ = pearsonr(x3, y3)
corr4, _ = pearsonr(x4, y4)

print("set \t mean \t sd \t mean \t sd \t corr")
print("1 \t", '%.1f' % statistics.mean(x1), "\t", '%.1f' % statistics.mean(y1), "\t", '%.2f' % statistics.stdev(x1),
      "\t", '%.2f' % statistics.stdev(y1), "\t", '%.3f' % corr1)
print("2 \t", '%.1f' % statistics.mean(x2), "\t", '%.1f' % statistics.mean(y2), "\t", '%.2f' % statistics.stdev(x2),
      "\t", '%.2f' % statistics.stdev(y2), "\t", '%.3f' % corr2)
print("3 \t", '%.1f' % statistics.mean(x3), "\t", '%.1f' % statistics.mean(y3), "\t", '%.2f' % statistics.stdev(x3),
      "\t", '%.2f' % statistics.stdev(y3), "\t", '%.3f' % corr3)
print("4 \t", '%.1f' % statistics.mean(x4), "\t", '%.1f' % statistics.mean(y4), "\t", '%.2f' % statistics.stdev(x4),
      "\t", '%.2f' % statistics.stdev(y4), "\t", '%.3f' % corr4)
```

```
set      mean    sd      mean    sd      corr
1        9.0     7.5     3.32    2.03    0.816
2        9.0     7.5     3.32    2.03    0.816
3        9.0     7.5     3.32    2.03    0.816
4        9.0     7.5     3.32    2.03    0.817
```



3. What is Pearson’s R?

Pearson’s R measures the strength of the linear relationship between two variables.

It is always between -1 and 1

1 indicates strongest possible positive correlation

For e.g. if a person is paid depending on number of apples he pick, that is, more apples he pick more he get paid. Here the correlation is 1

-1 indicates strongest possible inverse correlation

In the above e.g. correlation between money pile and number of apple picked will be -1.

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

- Is the data pre processing step applied on independent variables to normalize the data within particular range.

- Normalized scaling brings all of the data in the range of 0 and 1.
- 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).

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

- It happened when there is a perfect correlation

That is, R2 = 1 when there is perfect correlation. Which leads to 1/1-1 which is infinity

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

- Q-Q Plot (Quantile-Quantile plots) is a plot of the quantiles of the first data set against the quantiles of the second data set.

Where quantile means, the fraction or percent of points below the given value.

That is 0.3 or 30% quantile is the point at which 30% of the data fall below and 70% fall above that value

It is important when we have to find out:

- if 2 data sets come from populations with common distribution
- if 2 data sets have common location and scale
- if 2 data sets have similar distributional shapes
- if 2 data sets have similar tail behavior