# 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?

Ans :- a) The Demand of the bike is dropped highest.

b) The Bike demand in year 2019 is higher than 2018.

c) The Demand of bike is the highest in clear weather and low in rainy season.

d) The bike demand is similar all the weekdays.

e) The Demand of the bike will be highest in Spring season.

f) The demand of the bike is in months from May to October.

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

Ans :- a) It is an important to achieve k-1 dummy variables .It can be used to delete extra column while creating dummy variables.

b) It can be used to reduce the collinearity among dummy variables.

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

Ans :- temp and both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

4.) How did you validate the assumptions of Linear Regression after building the model on the

training set?

Ans :- a) Less multi-collinearity among features with low VIF.

b) Constant variance of the errors.

c) Normal distribution of error terms.

d) Linearity of relationship between response and predictor variables.

5.) Based on the final model, which are the top 3 features contributing significantly towards

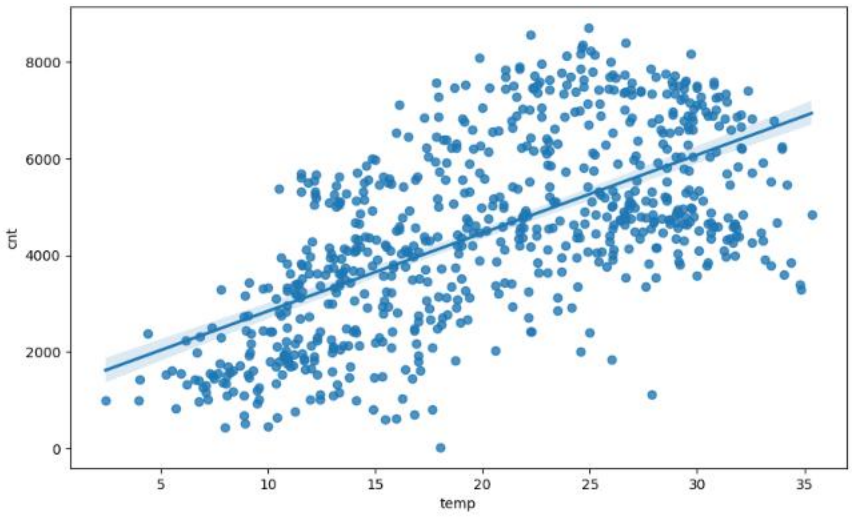
explaining the demand of the shared bikes?

Ans :- a) temp b) sept c)light_rain_snow

# General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Ans :- Linear Regression is one of the most fundamental algorithms in the Machine Learning world which comes under supervised learning. Regression models predict a dependent value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

```
In [48]: 1 sns.regplot(x="temp",y="cnt",data=df)
Out[48]: <Axes: xlabel='temp', ylabel='cnt'>
```




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 temp (temperature and Y (output) is cnt(count)of the bike. The regression line is the best fit line for our model.
Linear Regression may further divide into

1. Simple Linear Regression/ Univariate Linear regression
2. Multivariate Linear Regression

**Simple Linear Regression/ Univariate Linear Regression**

When we try to find out a relationship between a dependent variable (Y) and one independent (X) then it is known as Simple Linear Regression/ Univariate Linear regression.

The mathematical equation can be given as:

$Y = \beta 0 + \beta 1 * x$

Where

Y is the response or the target variable

x is the independent feature

β1 is the coefficient of x

β0 is the intercept

β0 and β1 are the model coefficients (or weights). To create a model, we must "learn" the values of these coefficients. And once we have the value of these coefficients, we can use the model to predict the target variable such as cnt!

NOTE: The main aim of the regression is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the points to the regression line.

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

Ans :- Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analysing it with statistical properties. There are four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when we plot these data sets, they look very different from one another. This suggests the data features must be plotted to see the distribution of the samples that can help us identify the various anomalies present in the data.

We can define these four plots as follows:-

| 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 | |

These models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm.

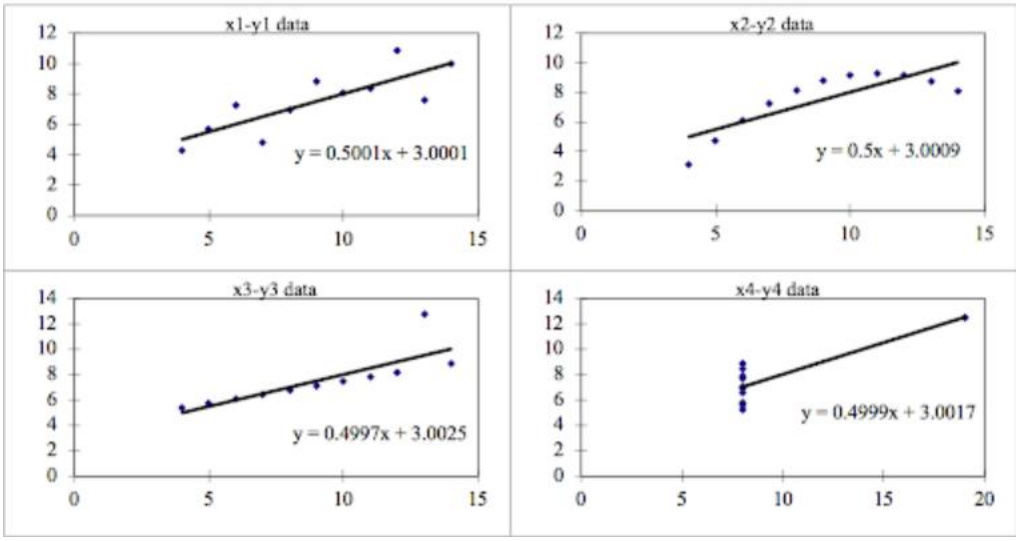
**ANSCOMBE'S Quartet four datasets**

Data_set_1: fits the linear regression model pretty well.

Data_set_2: cannot fit the linear regression model because the data is non-linear.

Data _set_3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear




regression model

---

3.) What is Pearson's R?

Ans :- The Pearson correlation coefficient is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope.

r = -1 means the data is perfectly linear with a negative slope.

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

➔ **When to use the Pearson correlation coefficient**

The Pearson correlation coefficient (r) is one of several correlation coefficients that we need to choose between when we want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:

a) Both variables are quantitative: we need to use a different method if either of the variables is qualitative.
b) The variables are normally distributed: we can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
c) The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data
d) the relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

**Pearson vs. Spearman's rank correlation coefficients**

Spearman's rank correlation coefficient is another widely used correlation coefficient. It's a better choice than the Pearson correlation coefficient when one or more of the following is true:

➔ The variables are ordinal.
➔ The variables aren't normally distributed.
➔ The data includes outliers.
➔ The relationship between the variables is non-linear and monotonic.

**Calculating the Pearson correlation coefficient**

Below is a formula for calculating the Pearson correlation coefficient (r):

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

The formula is easy to use when you follow the step-by-step guide below. We can also use software such as R or Excel to calculate the Pearson correlation coefficient for us.

---

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

**Ans** :- 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.

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 Scaling:-** It brings all of the data in the range of 0 and 1.

- ➔ sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

**Standardization Scaling**:- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

- ➔ 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

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

---

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

Ans :- A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.

VIF is infinite if there is a perfect corelation, and Ri value is 1.

VIF can be calculated by the formula below:-

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

Where Ri2 represents the unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones. The reciprocal of VIF is known as tolerance. Either VIF or tolerance can be used to detect multicollinearity, depending on personal preference.

Generally, a VIF above 4 or tolerance below 0.25 indicates that multicollinearity might exist, and further investigation is required. When VIF is higher than 10 or tolerance is lower than 0.1, there is significant multicollinearity that needs to be corrected.

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

Ans:- 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.

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.

**Advantages**:

-> It can be used with sample sizes also

-> Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.