# 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) Following is the inference of my analysis on categorical variables:

1) Season: The demand for bicycles is highest in fall followed by summer and least in spring
2) Year: Bicycle demand has increased by almost 65% in 2019 as compared to 2018
3) Month: The demand for bicycles increases constantly from January till September and then gradually decreases till December. The bicycle demand is more in April till December as compared to months from January to March. The demand is highest in months of June till September.
4) Holiday: The median demand is much higher on no holiday days (4563) as compared to holiday ones (3351).
5) Weekday: All the weekday's have almost same amount of median demand of bicycles
6) Working Day: The demand for bicycles is more on working days. 25% of the bicycle demand on working days is more than the 25% on non-working days.
7) Weather: The demand for bicycle is maximum on clear/partly cloudy days followed by mist/cloudy days and least on light rain days.

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

Ans) It is important to use drop_first=True during dummy variable creation because all the information about the categorical variable will be contained in rest of categories. So, if we don't use drop_first=True so unnecessarily we will passing redundant info to the model which will create the problem of multicollinearity with categorical variables and hence will lead to overfitting of model

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

Ans) Temp has the highest correlation with the target variable of approx. 63%.

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

Ans) I validated the assumptions via following methods:

1) Predictors should have a linear relationship: Via scatterplot of continuous variables with target and calculating pearson correlation with target
2) No Multicollinearity among predictors: Removed multicollinearity by removing predictors with VIF > 5
3) Observations must be independent: Via Durbin-watson test. The value of the test statistic is ~2 which shows no autocorrelation
4) Errors should have a normal distribution: Via QQ plot of residuals and histogram of residuals

5) Residuals should have a constant variance: Via scatterplot of predicted count and residuals. The spread of residuals showed no pattern

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

Ans) The top 3 features are:

1) Year
2) Month
3) Temperature; increase in these factors increases the demand for bicycles

## General Subjective Questions

**1. Explain the linear regression algorithm in detail.**

Linear regression may be defined as the statistical model that analyses thelinear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increaseor decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help offollowing equation −

$$Y = mX + c$$

X - is the independent variable.
m - is the slope of the regression line.
c - is a constant, known as the Y-intercept.

It is one of the easiest and most popular Machine Learning algorithms. It isa 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.

Types of Linear Regression

1) 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 calledSimple Linear Regression.

2) 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 iscalled Multiple Linear Regression.

Mathematically the relationship of multiple linear regression can berepresented with the help of following equation –

$$Y = \beta 0+\beta 1*x1+ \beta 2*x2 + \ldots . + \beta i*xi + \varepsilon$$

Y= Dependent Variable.
Xi= Independent Variable.
β0= intercept of the line.
βi = Linear regression coefficient.
ε = random error

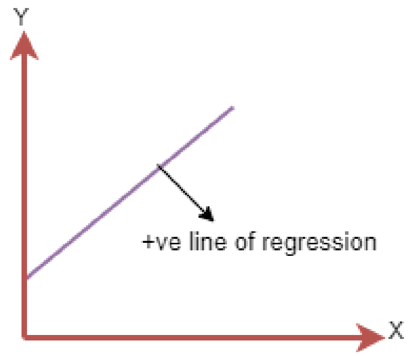
Linear Regression Line

1. Positive Linear Relationship:

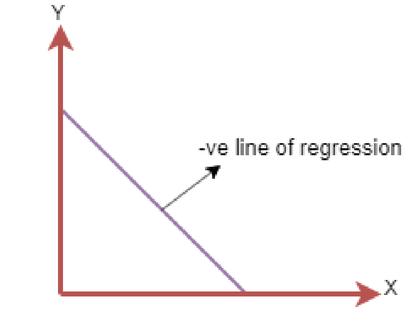
If the dependent variable increases on the Y-axis and independent variableincreases on X-axis, then such a relationship is termed as a Positive linearrelationship.

2. 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 equation will be: **$Y= a_0+a_1x$**



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

Assumptions -

The following are some assumptions about dataset that is made by LinearRegression model –

1) Multi-collinearity

Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when theindependent variables or features have dependency in them.

2) Linear relationship

Linear regression model assumes that the relationship between responseand feature variables must be linear.

3) Homoscedasticity

Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

4) Normal distribution of error terms

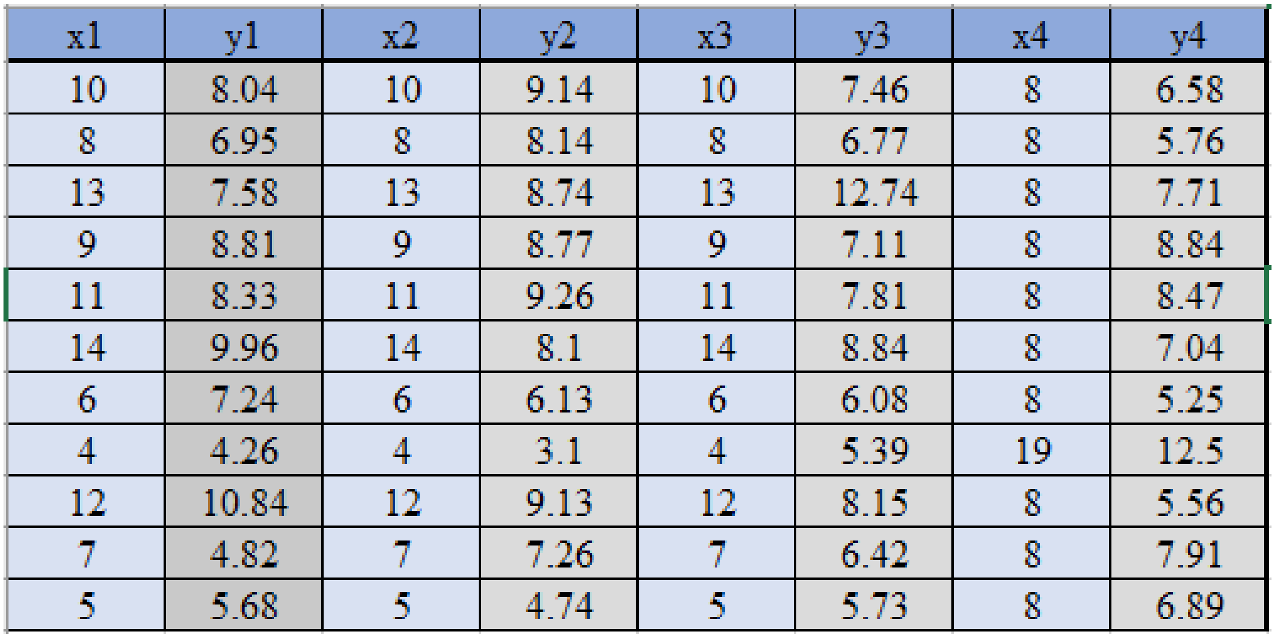
Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.

5) No autocorrelations

The linear regression model assumes no autocorrelation in error terms. Ifthere will be any correlation in the error term, then it will drastically reducethe accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

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

Anscombe's Quartet is the modal example to demonstrate the importanceof data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties.
It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics (mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

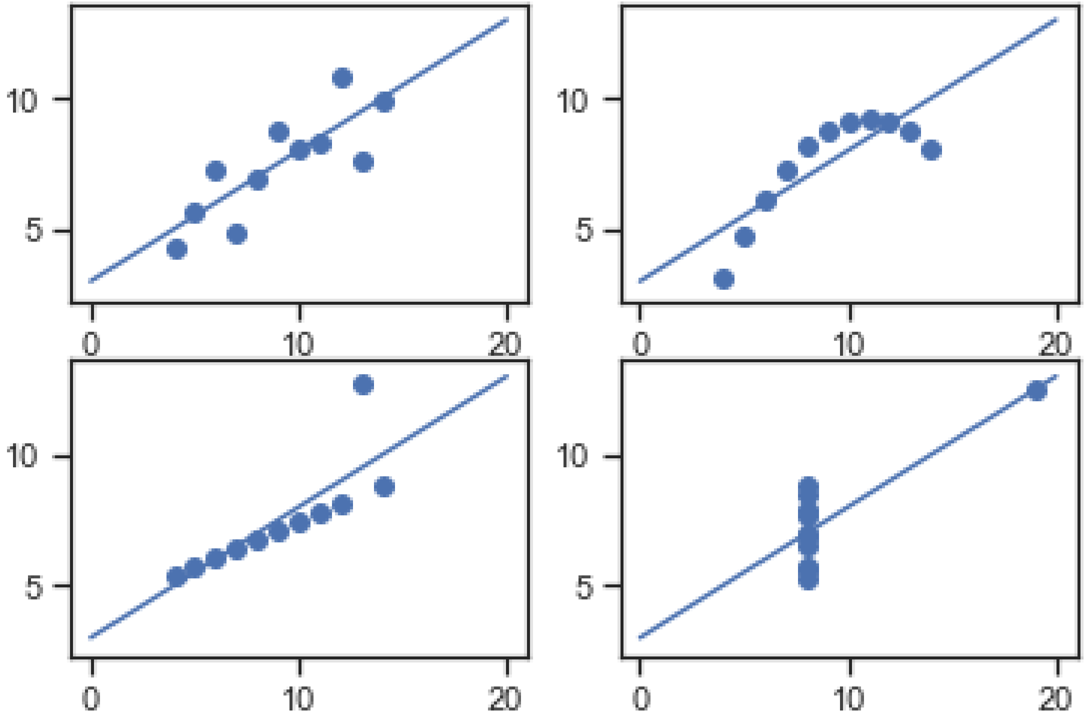


| x1 | y1 | x2 | y2 | x3 | y3 | x4 | y4 |
|---|---|---|---|---|---|---|---|
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

Apply the statistical formula on the above data-set:

Average Value of x = 9
Average Value of y = 7.50
Variance of x = 11
Variance of y=4.12
Correlation Coefficient = 0.816
Linear Regression Equation: y = 0.5 x + 3

However, the statistical analysis of these four data-sets is pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represents the different behavior.



Data-set I — consists of a set of (x,y) points that represent a linearrelationship with some variance.

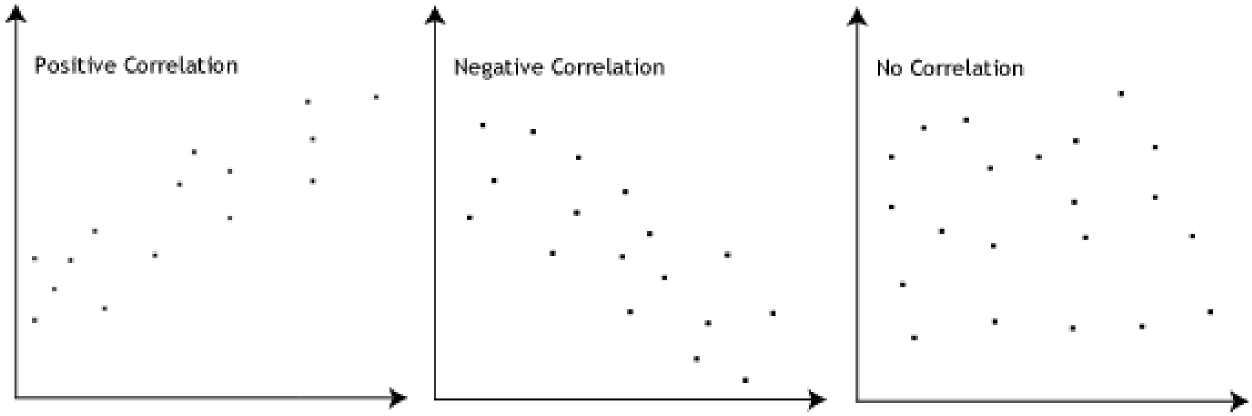
Data-set II — shows a curve shape but doesn’t show a linear relationship(might be quadratic?).
Data-set III — looks like a tight linear relationship between x and y, exceptfor one large outlier.
Data-set IV — looks like the value of x remains constant, except for one outlier as well.

Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data

### 3. What is Pearson’s R?

Pearson's r, also known as the Pearson correlation coefficient, is a statistical measure that describes the linear relationship between two continuous variables. It is a value between -1 and 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship.

Pearson's r measures the degree to which the variables are related by calculating the ratio of the covariance between the variables to the product of their standard deviations. In other words, it measures how much the variables vary together relative to how much they vary independently.



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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method, then it can consider the value 3000 meter to be greater than 5 km but that’s actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

| **Normalization** | **Standardization** |
|---|---|

| | |
|---|---|
| It is used when features are ofdifferent scales. | It is used when features are ofdifferent scales |
| Minimum and maximum values are used for scaling | Mean and standard deviation isused for scaling. |
| Scales values between [0, 1] or [-1,1]. | It is not bounded to a certain range. |
| It is really affected by outliers. | It is much less affected by outliers. |
| It is a often called as Scaling Normalization | It is a often called as Z-Score Normalization. |

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factorof 4 due to the presence of multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

If there is perfect correlation, then VIF = infinity. This 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. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. 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.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. QQ plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests