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? (3 marks)
Answer:

I have done analysis on categorical variable using boxplot and barplot, the observations are as follows:

- Fall season attracted more people to use shared bikes among all seasons.
- User count seems to be high from May to October.
- Most of the people use shared bikes on weekends and holidays which is practically true since many people wanted to go out on holidays.
- User count seems to be less in the start of the week and it's gradually increasing towards the end of the week.
- Clear weather attracts more people to use shared bikes.
- The count has been increased in the year of 2019 rather than 2018.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Answer:

The get_dummies() in pandas library by default creates dummy variables as many as labels of the categorical variable. But using the (n-1) number of variables is enough to understand those categorical variables where n is the number of labels. To reduce the dummy variable count, we are using one of the parameters of get_dummies() which is drop_first=True. The default value of this parameter is False.

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

The variable ‘temp’ has the highest correlation with the target variable among others.

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

The validation has been done with the following assumptions of linear regression:

- Linear relationship: Observed linear relationship between the variables
- Normality in error: The residuals are normally distributed with mean zero.

- Multicollinearity: The VIF value of the feature variable confirms there is no multicollinearity between the variables.
- Homoscedasticity: There is no defined pattern.
- No endogeneity: Independent variables should not be correlated with residuals.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes? (2 marks)
Answer:

The top three features that defines the demand for the shared bikes are:

1. temp
2. Windspeed
3. Year

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Answer:

Linear Regression is a type of supervised machine learning algorithm that explains the linear relationship between the dependent variable and one or more independent variables. In simple words this algorithm predicts the dependent variable with the independent variable where there is a linear relationship.

Mathematical equation for Linear regression can be presented as

$$y = mX+c$$

where , y is the dependent variable, X is the independent variable, m is the slope of the linear line and c is the y-intercept.

The linear relationship can be defined in two ways,

- Positive linear relationship, where the dependent variable increases when the independent variable increases.
- Negative linear relationship, where the dependent variable decreases when the independent variable increases.

Linear regression is of two types:

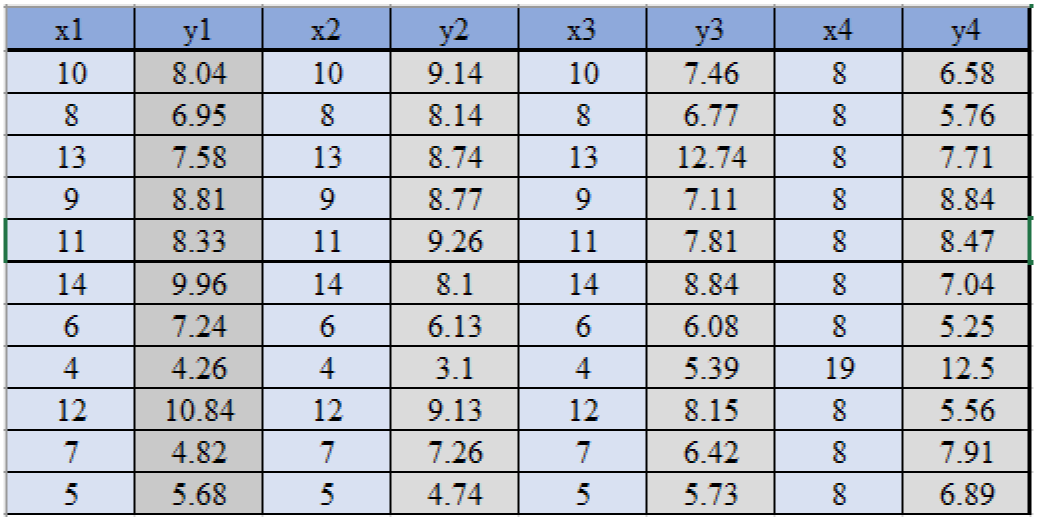
- Simple linear regression - has only one independent variable
- Multiple linear regression - has more than one independent variables

Assumptions of Linear Regression:

- Linear relationship between variable: Linear regression assumes that there is linear relationship between the independent and predictor variable.
- Multicollinearity: There should be very less or no multicollinearity between the independent variables. This will occur if there is some dependency between variables.
- Auto-corelation: There should be any auto- correlation in the data which means there is no correlation between residuals and independent variables.
- Normality of residuals: Residuals should be normally distributed with mean zero.
- Homoscedasticity: There should be no visual patterns in the residuals.

2. Explain the Anscombe's quartet in detail. (3 marks)
Answer:

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombein 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises 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.

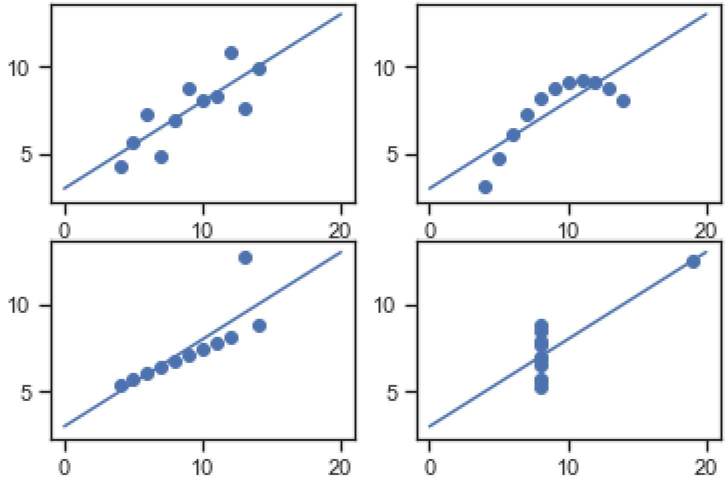


x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7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



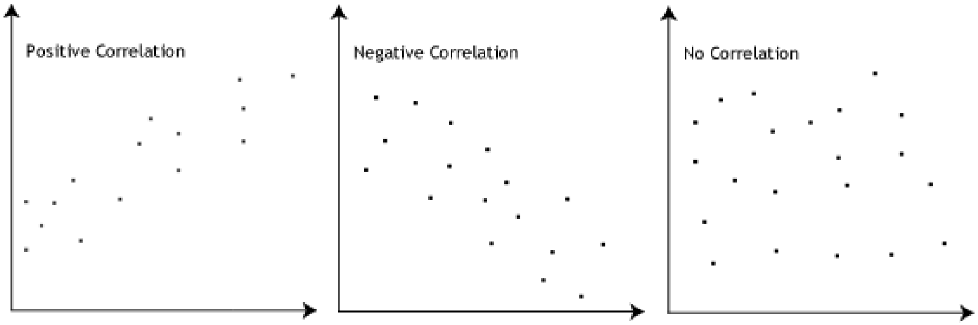
- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn’t show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.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

3. What is Pearson’s R? (3 marks)

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer:

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.

difference between normalized scaling and standardized scaling:

Normalized Scaling	Standardized Scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bound to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

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 factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which leads to 1/ (1-R2) infinity. To solve this 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 in linear regression. (3 marks)

Answer:

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. The q-q 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.