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:

I have done analysis on categorical columns using the box plot and bar plot. Below are the few points we can infer from the visualization-

- Fall season seems to have attracted more bookings, and in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of the year.
- Clear weather attracted more bookings, which seems obvious.
- Thursday, Friday, Saturday and Sunday have more number of bookings as compared to the start of the week.
- When it's not a holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking seemed to be almost equal either on working day or non-working day.
- 2019 attracted more number of bookings than the previous year, which shows good progress in terms of business.

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

Answer:

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

Syntax -

Drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

Let's say we have 3 types of values in the Categorical column, and we want to create dummy variable for that column. If one variable is not A and B, it is obvious C. So, we do not need 3rd variable to identify the C.

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

Answer:

'temp' variable has the highest correlation with the target variable.

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

Answer:

I have validated the assumptions of Linear regression model based on below 5 assumptions-

1. Normality of errors terms: Error terms should be normally distributed
2. Multicollinearity: there should be insignificant multicollinearity among variables
3. Linear relationship validation: Linearity should be visible among variables
4. Homoscedasticity: There should be no visible among variables
5. Independence of residuals: No autocorrelation

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

Answer:

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes-

- Month of September
- Summer season
- Temperature n

General Subjective Questions

1) Explain the linear regression algorithm in detail?

Answer:

Linear regression algorithm is a type of supervised machine learning algorithm used to predict the value of a variable based on another variable or a set of variables. The variable you want to predict is called the dependent variable and the variables which we are using to predict the variable's value are called the independent variable.

The linear regression algorithm computes the relationship between the dependent variable and one or more independent features by fitting a linear equation to the observed data.

When there is only 1 independent variable, the model is called Simple Linear regression while if there is more than one, it is known as Multiple linear regression. Mathematically Simple and Multiple linear regression models can be given as

Simple Linear regression model:

$$Y=\beta_1 X + \beta_0$$

Where:
Y is the dependent variable
X is the independent variable
βo is the intercept
β_1 is the slope

Multiple linear regression model:

$$Y=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

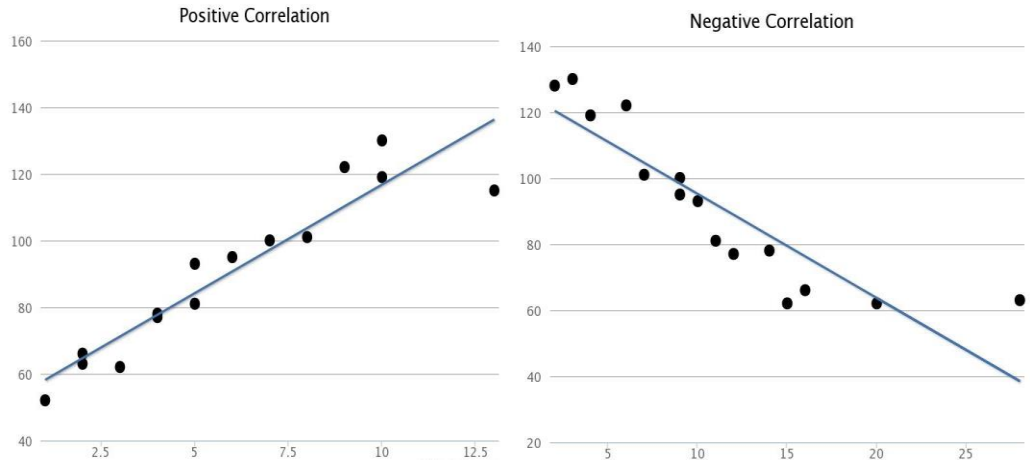
Where:
Y is the dependent variable
$X_1, X_2, X_3 .... X_n$ are the independent variables
βo is the intercept
$\beta_1, \beta_2, \beta_3 .... \beta_n$ are the slopes

The goal of the algorithm is to find the best fit line equation that can predict the values based on the independent variables.

Furthermore, the linear relationship can be positive or negative in nature as explained below:

- Positive linear relationship:
  A linear relationship will be called positive if both independent and dependent variable increases.

- Negative linear relationship:
  A linear relationship will be called negative if independent variable increases and dependent variable decreases.



Assumptions for linear relationship-

The following are some assumptions about datasets that is made by linear regression model-

- Multi- collinearity -
  Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi–collinearity occurs when the independent variables or features have dependency in them.
- Independence -
  Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Relationship between variables-
  The linear regression model assumes that the relationship between response and feature variables must be linear.
- Normality of error terms -
  Error terms should be normally distributed
- Homoscedasticity -
  There should be no visible pattern in residual values.

2) Explain the Anscombe's quartet in detail?

Answer:

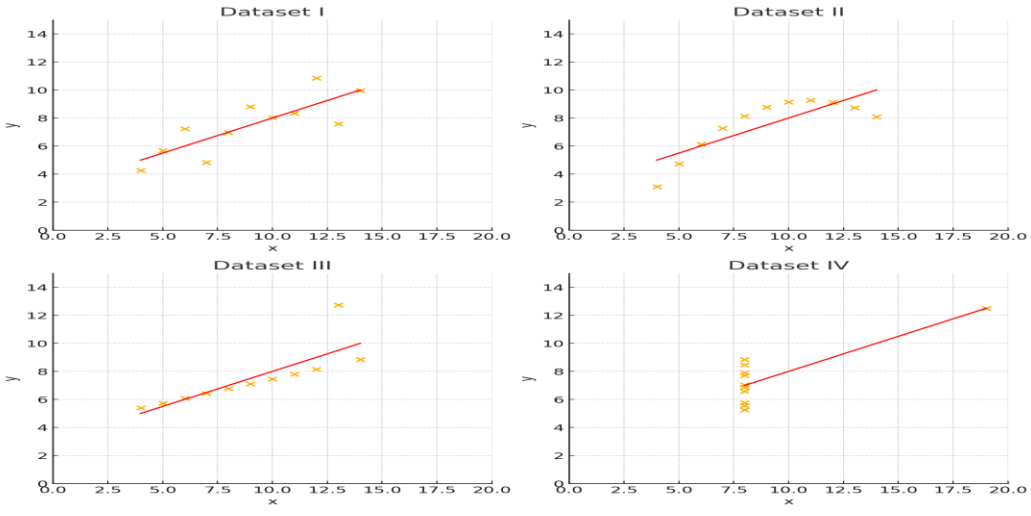
Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics yet appear very different when graphed. The quartet was constructed by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphical analysis of data, rather than relying solely on summary statistics.

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

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset.
- The correlation co-efficient (how string a relationship is between two variables) between x and y is 0.816 for each dataset.

When we plot these four datasets on an x/y co-ordinate plane, we can observe that they show the same regression lines as well, but each dataset tells a different story.



- Dataset 1 appears to have clean and well-fitting linear models.
- Dataset 2 is not distributed normally.
- In Dataset 3 the distribution is linear, but the calculated regression is thrown off by an outlier.

- Dataset 4 shows that one outlier is enough to produce a high correlation coefficient.

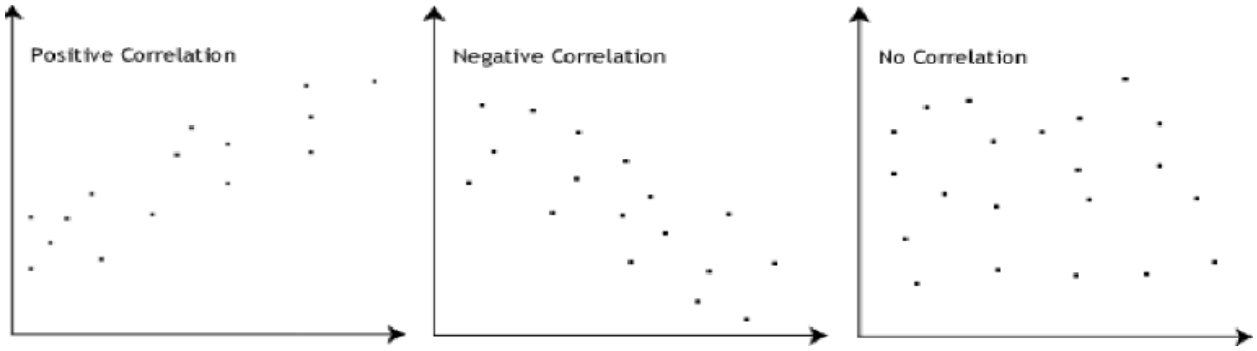
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of structure and a clear picture of the data.

3) What is Pearson's R?

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 the 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. This is shown in the diagram below:



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

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.

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

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	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?

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 below 5, this means that the variance of the model coefficient is inflated by a factor of 5 due to the presence of multi collinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we are R-squared(R2) =1, which lead to 1/(1-R2) infinity. To solve this, we need to drop one of the variables from the datasets which is causing this perfect multi collinearity.

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 the 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 data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of the 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 assumptions of a common distribution are 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 insights into the nature of the difference than analytical methods such as chi-square and Kolmogorov-Smirnov 2-sample tests.