

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 columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –

- ✓ Fall season seems to have attracted more booking. 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 year.
- ✓ Clear weather attracted more booking which seems obvious.
- ✓ Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
- ✓ When it's not 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 booking from 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?** (2 mark)

Answer:

The intention behind the dummy variable is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.

Eg: If there are 3 levels, the drop_first will drop the first column..

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

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?** (3 marks)

Answer:

I have validated the assumption of the Linear Regression Model based on the below 5 assumptions -

- ✓ Normality of error terms
 - Error terms should be normally distributed

- ✓ Multicollinearity check
 - There should be insignificant multicollinearity among variables.
- ✓ Linear relationship validation
 - Linearity should be visible among variables
- ✓ Homoscedasticity
 - There should be no visible pattern in residual values.
- ✓ 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? (2 marks)**

Answer:

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

- temp
- winter
- sep

General Subjective Questions

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

Answer:

Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear

regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a_0 and a_1 to find the best fit line and the best fit line should have the least error.

In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a_0 and a_1 , which provides the best fit line for the data points.

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

Answer:

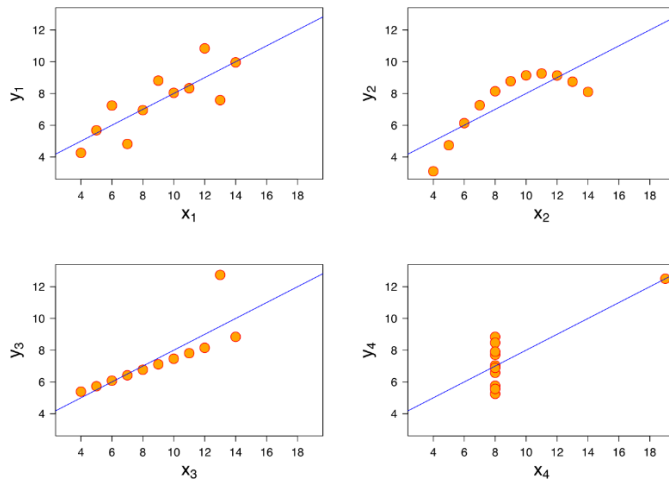
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

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 the mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and the variance of y is 4.13 for each dataset
- The correlation coefficient (how strong 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 coordinate plane, we can observe that they show the same regression lines as well, but each dataset tells a different story:



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV 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 the structure and a clear picture of the dataset.

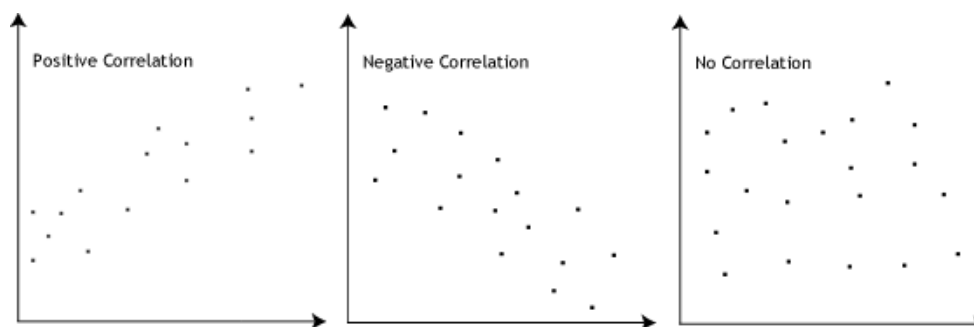
3. What is Pearson's R?

(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. 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?

(3 marks)

Answer:

Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in a specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units, and ranges. If scaling is not performed, then the algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features are used whereas in standardized scaling mean and the standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized

scaling is used to ensure zero mean and unit standard deviation.

3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when the distribution is normal.
6. Normalized scaling is called scaling normalization whereas standardized scaling is called as Z Score Normalization.

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 $VIF = \infty$. 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 $R^2 = 1$, which leads to $1/(1-R^2)$ 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:

Q-Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression :

In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also
- If both datasets came from a population with common distribution
- If both datasets have a common location and common scale
- If both datasets have a similar type of distribution shape
- If both datasets have tail behavior