Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependentvariable? (3 marks)

Answer:
Some of the insights drawn from the plots

- Season “Fall” has the highest demand for rental bikes.
- Count has increased for the year 2019 compared to 2018.
- There is continuous growth till the month of June. September month has the highest demand. After September, demand decreased.
- Whenever there is a holiday, demand decreases.
- Weekday doesn’t give a clear picture of demand.
- The clear_cloudy weather situation (weathrsit) has highest demand.

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

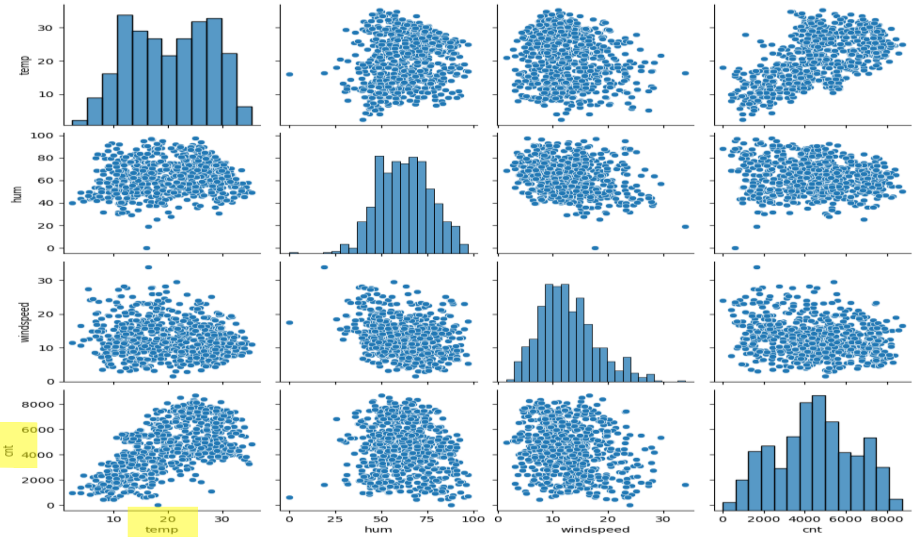
Answer:

Drop_first= True or dummy encoding basically reduces the number of dummy variables by 1 without losing any information. Hence it reduces the correlations created among dummy variables. Dropping the first columns as (p-1) dummies can explain p categories. Another advantage is we reduce the number of variablesthat machine learning algorithm needs to learn

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

Answer:

From the pair plot, we can derive that cnt has highest correlation with temp.



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

Answer:

- Residual Analysis:
  Errors are normally distributed with a mean of 0.
  Actual and predicted results follow the same pattern.
  The error terms are independent of each other.

- R2 value for test predictions:
  R2 value for predictions on test data (0.801) is almost same as R2 value of train data(0.774).
  It is a good R-squared value, and we can see our model is performing good even on test data

- Homoscedasticity:
  The variance of the residuals (error terms) is constant across predictions. i.e., error term does not vary much as the value of the predictor variable changes. The variance here is equally spread.

- Plot Test vs Predicted value test:
  The prediction for test data is very close to actuals.

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

   The top 3 features are:

   - yr (positive correlation)
   - temp (positive correlation)
   - Winter season (negative correlation)

General Subjective Questions

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

Answer:

Linear regression is a statistical regression method used for predictive analysis that shows the relationship between the continuous variables. It shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis. 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.
The linear regression model can be represented by the following equation:
y= a0+a1x+ ε

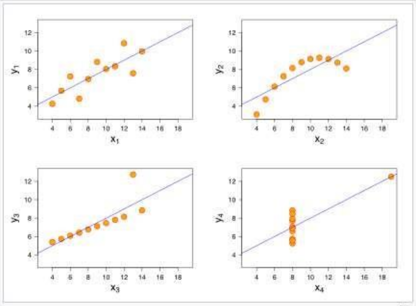
The linear regression model provides a sloped straight line representing the relationship between the variables.
y= Dependent Variable (Target Variable) x= Independent Variable (predictor Variable) a0= intercept of the line (Gives an additional degree of freedom) a1 = Linear regression coefficient (scale factor to each input value). ε = random error.
The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.
The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable. This mapping function is also known as the Hypothesis function.

In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

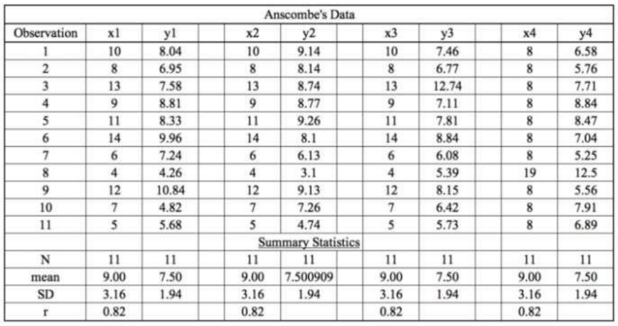
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed



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	

From the above data, we can observe that the summary stats are nearly identical for all datasets, but their plots are varying. Below are some of the derivations from it

- Data can be linear or non-linear
- There can be outliers which can or cannot be handled by the linear regression model

Important thing is that, plotting of data is required before the right model is picked for a given dataset

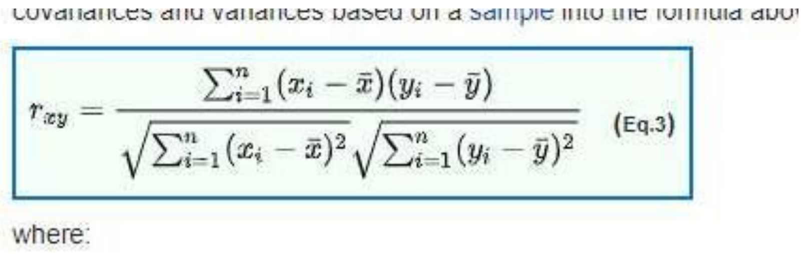
3. What is Pearson's R? (3 marks)

Answer:

Pearson correlation coefficient (PCC), also referred to as Pearson's R. The Pearson product-moment correlation coefficient, or the bivariate correlation, is a measure of linear correlation between the two sets of data. It is the covariance of two variables, divided by the product of their standard deviations. Thus, it is fundamentally a standardized measurement of the covariance, so that the result will always has a value between −1 and 1.

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

- 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



covariances and variances based on a sample into the formula abo

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{(Eq.3)}$$

where:

Here,
- r =correlation coefficient
- xi =values of the x-variable in a sample
- x̄=mean of the values of the x-variable
- yi =values of the y-variable in a sample
- ȳ =mean of the values of the y-variable

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

Answer:

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. This helps in speeding up Beta derivation using gradient descent.
If scaling is not done, then algorithm only takes magnitude into account and not the units leading to incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

Normalized scaling or Min Max scaling tries to fit data in [0 and 1] scale by doing

$(x-x_{min})/(x_{max}-x_{min})$

Standardized scaling scales value in such a way that mean lies at 0. It is computed by

(x-mean(x))/standard deviation(x)

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

Answer:

If VIF is infinite, then it means that there is a perfect correlation. This shows a perfect correlation between two independent variables. If we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this multicollinearity.
If variables are highly corelated R2 becomes 1. This causes the denominator to become 0 and hence infinite.

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 (Quantile-Quantile) is a graphical tool which helps us evaluate if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. This helps in a scenario of linear regression where we have training and test data sets obtained individually and then we can check using Q-Q plot if both the data sets are from populations with same distributions. Q-Q Plots are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. The purpose of using Q-Q plot is to find out if two sets of data come from the same distribution.

A 45-degree angle is plotted on the Q-Q plot, if the two data sets come from a common distribution, the points will fall on that reference line. If the two distributions that are being compared are similar, then the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly associated, the points in the Q–Q plot will roughly lie on a line, but not essentially on the line y = x. Q–Q plots can also be used as a graphical method of estimating parameters in a location-scale family of distributions.