C.Satish Kumar

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

Inside the motorbike sharing dataset, let’s bear in mind the impact of the explicit variable ‘weathersit’ at the target variable ‘cnt’. at the same time as acting EDA, I visualized the relationship among the specific variables and the target variable. It became seen that during the weathersit_3 (light Snow, light Rain + Thunderstorm + Scattered clouds, mild Rain + Scattered), decreases inside the bike hires numbers by zero.333164 devices had been seen. approximately. further, sure inferences could be made by using season_Spring while season_Winter indicates reverse fashion.

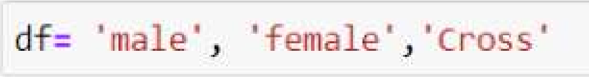
additionally, in the course of version building on inclusion of specific functions including yr,season and many others we noticed a

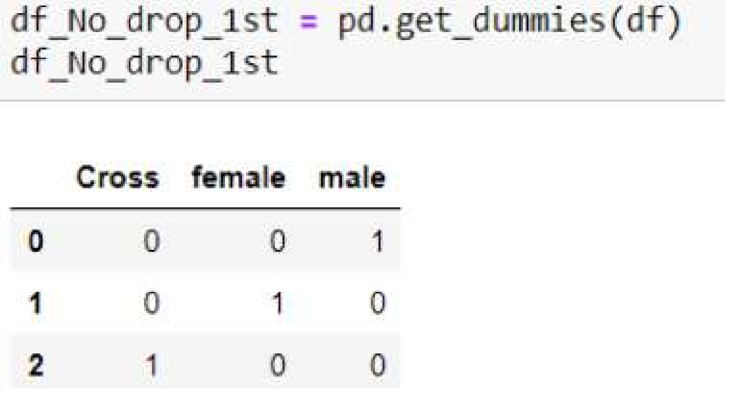
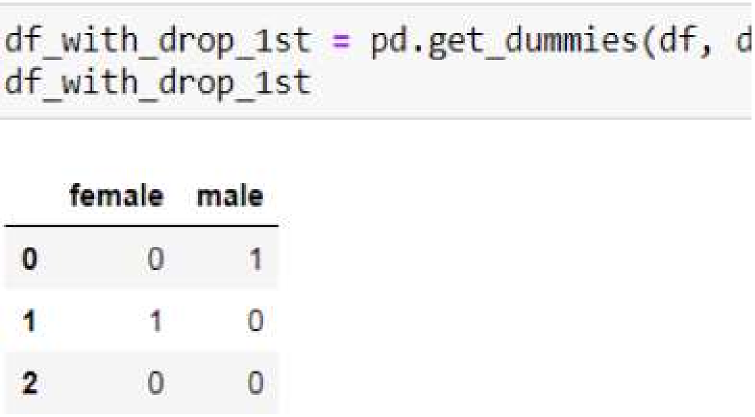
enormous change inside the fee of R-squared and changed R-squared. this means that the explicit capabilities had been useful in explaining a extra share of variances inside the statistics sets

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

Answer

To avoid multicollinearity in dummy variable encoding, it's best to set **drop_first=True**. This removes the first dummy variable, treating it as the reference category. This ensures that the model can accurately interpret the impact of each category without redundancy.

For example:	
suppose we have a categorical feature ‘male’, ‘Female’ and ‘Cross’ as df as shown in fig(a)	 `df= 'male', 'female','Cross'`

<table>
<tr>
<td>If we not use <b>drop_first=True</b> we will get all 3 variables as dummy that is as shown in the fig.(b)</td>
<td>
df_No_drop_1st = pd.get_dummies(df)
df_No_drop_1st
<table>
<tr><th></th><th>Cross</th><th>female</th><th>male</th></tr>
<tr><td><b>0</b></td><td>0</td><td>0</td><td>1</td></tr>
<tr><td><b>1</b></td><td>0</td><td>1</td><td>0</td></tr>
<tr><td><b>2</b></td><td>1</td><td>0</td><td>0</td></tr>
</table>
</td>
</tr>
<tr>
<td>If we use drop_first=True we will get all 2 variables as dummy that is sufficient to serve our purpose as shown in the fig(c):
If it is Female: Female shows the value as 01.
If it is Male: Male shows the value as 10.
If it is not Female nor Male its Cross 00.</td>
<td>
df_with_drop_1st = pd.get_dummies(df, d
df_with_drop_1st
<table>
<tr><th></th><th>female</th><th>male</th></tr>
<tr><td><b>0</b></td><td>0</td><td>1</td></tr>
<tr><td><b>1</b></td><td>1</td><td>0</td></tr>
<tr><td><b>2</b></td><td>0</td><td>0</td></tr>
</table>
</td>
</tr>
</table>

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

Answer

Before data preparation, the numerical variable 'registered' showed the highest correlation (0.95) with the target variable 'cnt'. However, after removing 'registered' from the analysis due to its high correlation with other features (multicollinearity), the numerical variable 'atemp' emerged as the strongest predictor of 'cnt' with a correlation of 0.63.

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

Answer

After training the linear regression model, I assessed the following assumptions:

1. Linear relationship between independent and dependent variables
2. Normally distributed error terms
3. Normally distributed residuals in the training data

To address multicollinearity, I employed variable selection techniques based on VIF and p-values.

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

Answer

Our final model indicates that the following three variables significantly influence bike bookings:

1. **Temperature (temp):** A positive coefficient of 0.375922 suggests that a one-unit increase in temperature is associated with a 0.375922 increase in the number of bike hires, all else being equal.
2. **Weather Situation 3 (weathersit_3):** A negative coefficient of -0.333164 implies that a one-unit increase in this weather category (e.g., light snow, light rain) is associated with a 0.333164 decrease in bike hires, holding other factors constant.
3. **Year (yr):** A positive coefficient of 0.232965 suggests that a one-unit increase in year is associated with a 0.232965 increase in bike hires, controlling for other variables.

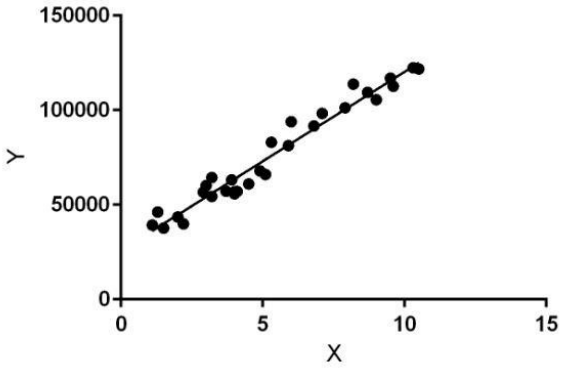
To optimize bike bookings, it's essential to prioritize these factors in planning and operational strategies.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer

Linear Regression is a supervised machine learning technique that employs a linear equation to model the relationship between a dependent variable and one or more independent variables. This algorithm is widely used for both prediction and forecasting tasks. The choice of regression model depends on factors such as the type of relationship between variables (linear or non-linear) and the number of independent variables involved.



Linear Regression is a statistical technique that models the linear relationship between a dependent variable (y) and one or more independent variables (x). It aims to find the best-fitting line that minimizes the distance between the predicted values and the actual values. The general equation for a multiple linear regression model is:

$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$

where:

y: predicted value

θ_0: intercept

$\theta_1, \theta_2, ..., \theta_n$: coefficients

$x_1, x_2, ..., x_n$: independent variables

By analyzing the coefficients of the model, we can understand the impact of each independent variable on the dependent variable. For example, in the context of weight prediction, we can determine the influence of factors like sleep, stress, and exercise, in addition to calorie intake.

How to update θ1 and θ2 values to get the best fit line?

Cost Function (J): By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2 \qquad J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

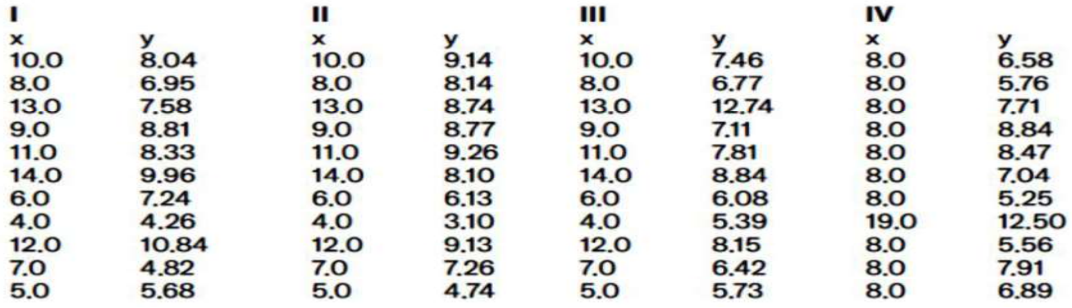
Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).

Gradient Descent: To update θ1 and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values and then iteratively updating the values, reaching minimum cost.

2. Explain the Anscombe’s quartet in detail.

Answer

It is a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset contains of eleven (x, y) pairs as follows:-

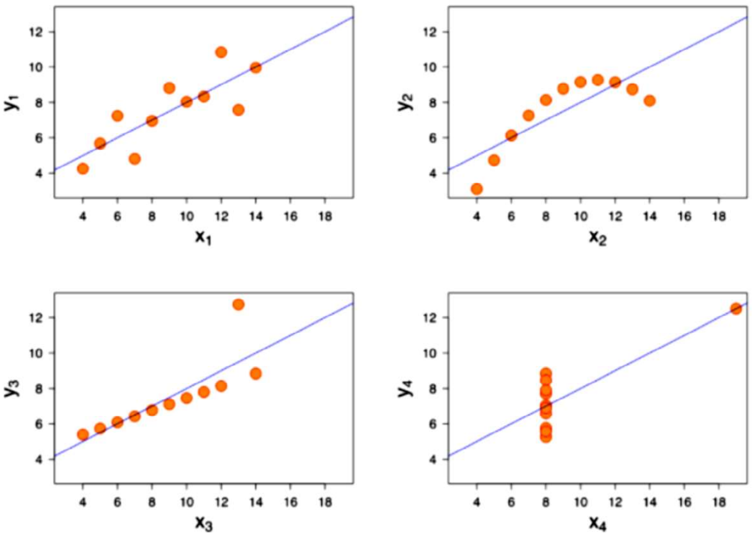


I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

All the summary statistics for each dataset are identical
1. The average value of x is 9.
2. The average value of y is 7.5.
3. The variance for x is 11 and y is 4.12
4. The correlation between x and y is 0.816
5. The line of best for is y = 0.5x + 3.
But the plots tell a different and unique story for each dataset.

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.

The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.



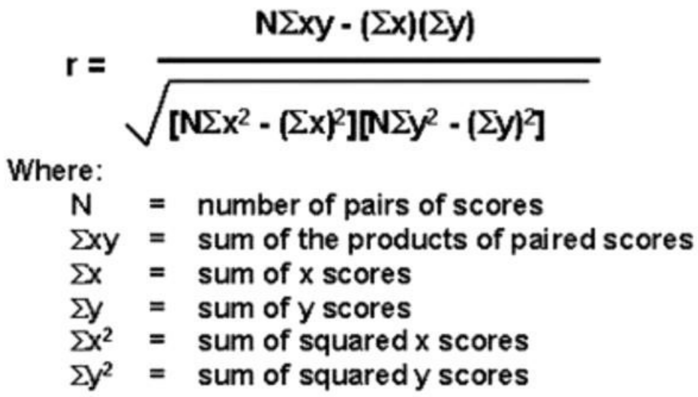
In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R?
Answer

Pearson's R is a numerical summary of the strength of the linear association between the variables. It varies between -1 and +1. 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. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
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



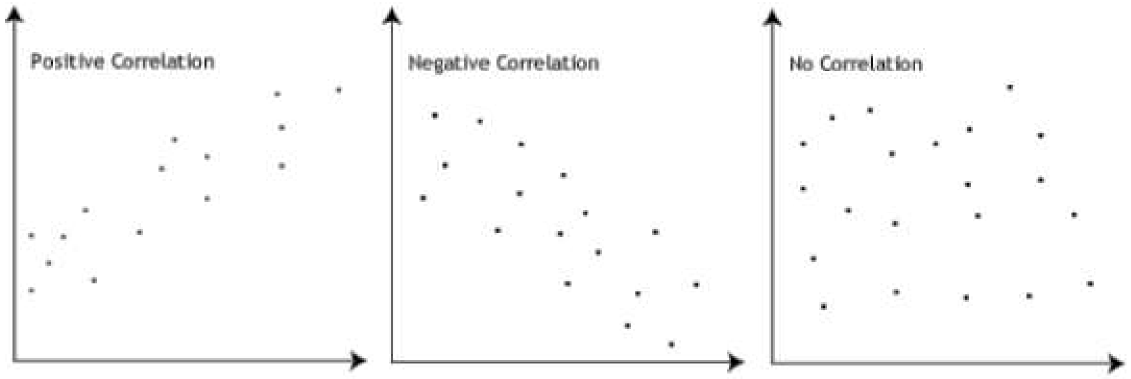
$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

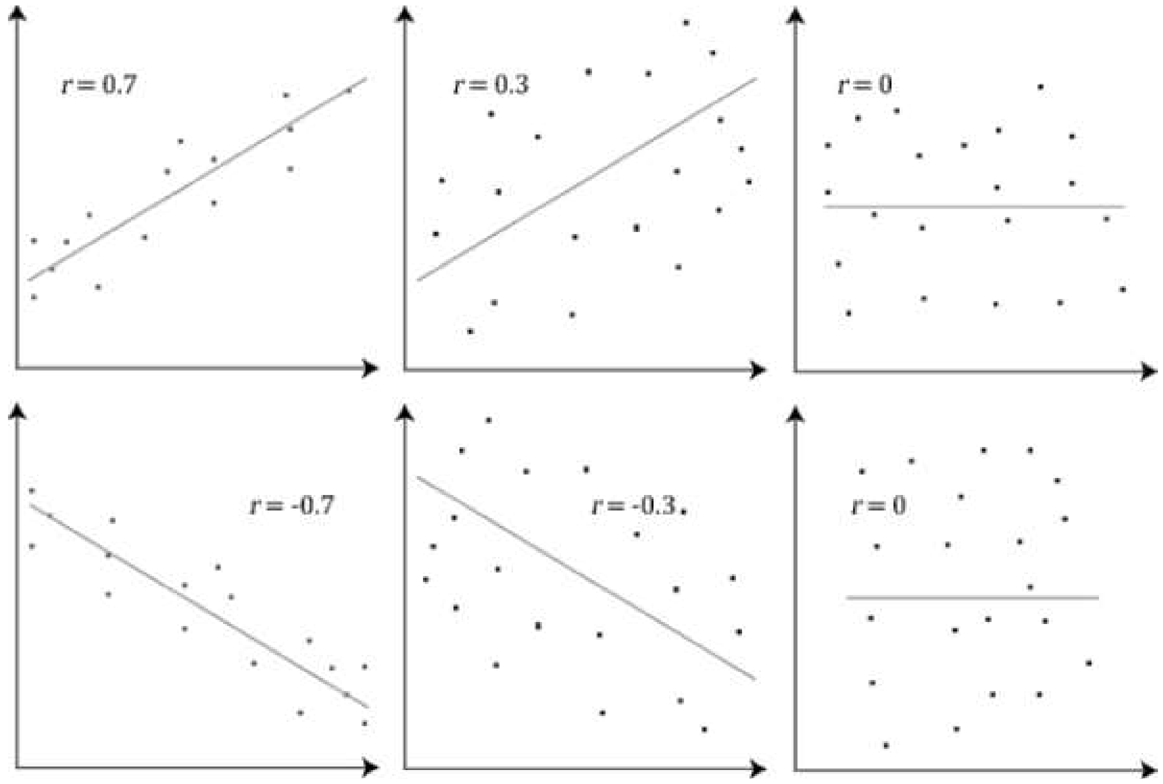
N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

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:



The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of *r* to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



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

Answer

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. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. Normalized scaling means to scale a variable to have values between 0 and 1, while standardized scaling refers to transform data to have a mean of zero and a standard deviation of 1

What?

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.
- sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

The Variance Inflation Factor (VIF) is a measure of colinearity among predictor variables within a multiple regression. It is calculated by taking the the ratio of the variance of all a given model's betas divide by the variane of a single beta if it were fit alone.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. 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).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios: If two data sets.

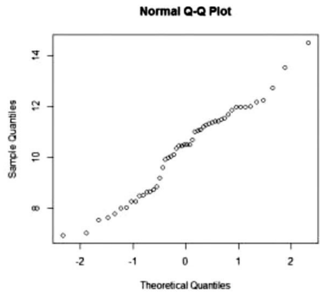
i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight i.e.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles