Assignment-based Subjective Questions

Question no 1:- rom your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer

Here many insights can be drawn from the plots:

1. Fall season has highest demand for rental bikes.
2. Demand for next year has grown.
3. Demand is showing continuous growth month on month till June. September month has highest demand. After September, demand is decreasing.
4. When there is a holiday, demand has decreased.
5. Weekday is not giving a clear picture about demand.
6. The clear weather situation (weathersit) has highest demand.
7. During September, bike sharing is more. During the end and beginning of year, it is less.

Question no 2:- . Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer

drop_first=True is important to use, as it helps in reducing the extra columns that gets created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Dropping the first columns as (p-1) dummies can explain p categories . In weathersit, first column was not dropped so as not to lose the info about severe weather situation.

Question no 3:-Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer

Looking at the pair-plot among the numerical variables, temp and atemp have the highest correlation (0.63) with the target variable (cnt)

Question no 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 result follow the same pattern. The error terms are independent of each other.

R2 value for test predictions: R2 value for predictions on test data (0.815) is almost same as R2 value of train data(0.818). This is a good R-squared value, hence we can see our model is performing good even on unseen data (test data)

Homoscedacity: We can observe that 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.

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

Question no 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 :

The top 3 features are:

1. yr (positive corelation)

2. temp (positive corelation)

3. weathersit_bad (negative corelation)

General Subjective Questions

Question no 1 :- Explain the linear regression algorithm in detail. (4 marks)

Answer :

Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. The Pearson's correlation coefficient varies between -1 and +1 where:

• 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

Question no 1 :- Explain the Anscombe’s quartet in detail. (3 marks)

Answer :-

Anscombe’s quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. The quartet was constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it and the effect of outliers and other influential observations on statistical properties

The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyze it, and the inadequacy of basic statistic properties for describing realistic datasets 12.

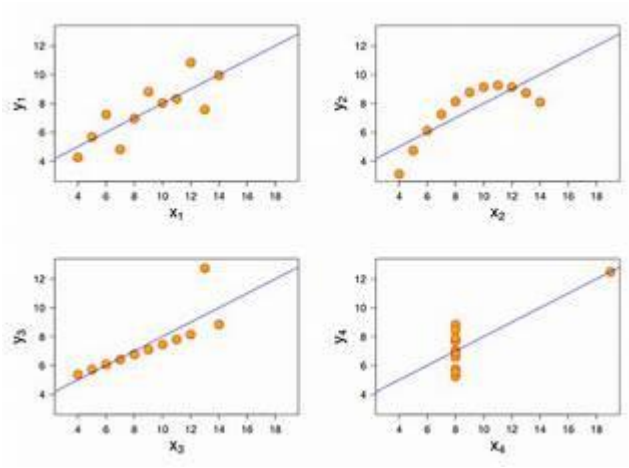
For instance, all four sets have the same mean, variance, correlation coefficient, and linear regression line. However, when plotted, they exhibit different patterns, such as linear, quadratic, or logarithmic

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

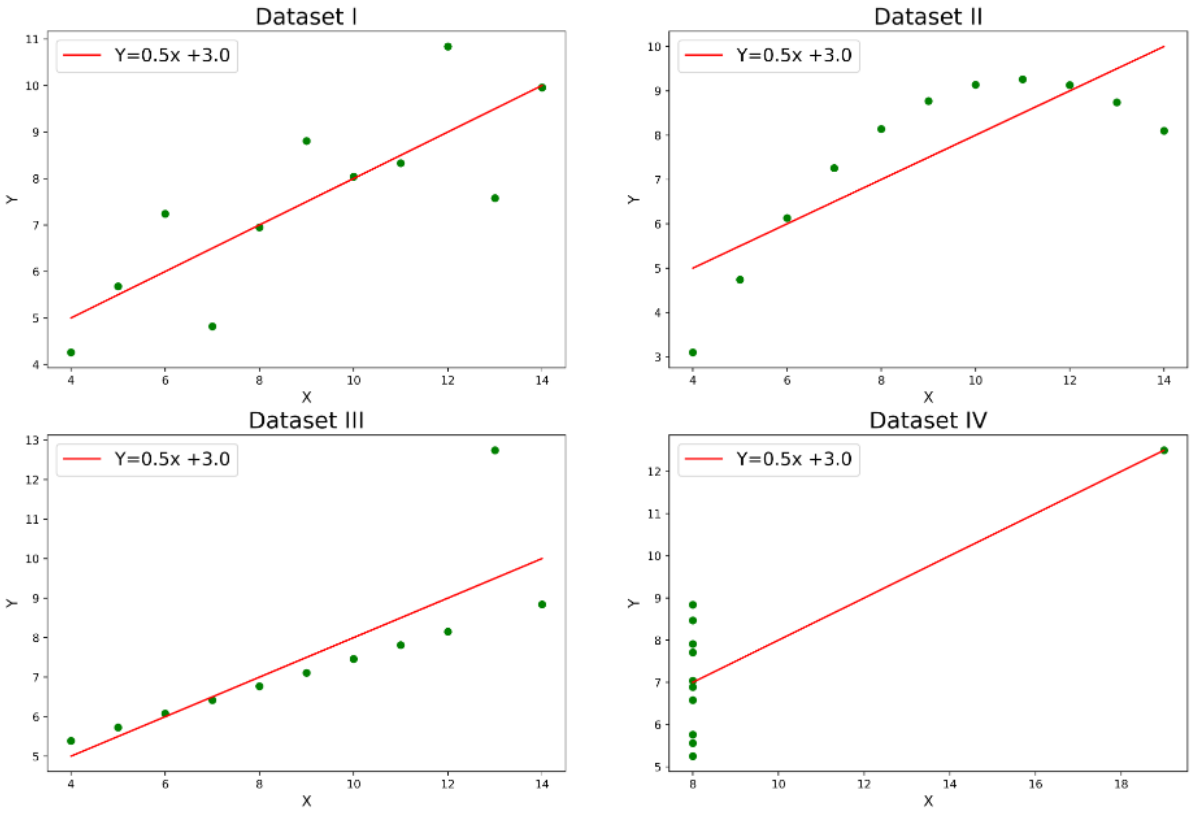
The four datasets of **Anscombe's quartet.**

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



How to implement the Anscombe's quartet Implementations

- Import the necessary libraries
- Load the dataset
- Find the descriptive statistical properties for the all four dataset
    - Find mean for x and y for all four datasets.
    - Find standard deviations for x and y for all four datasets.
    - Find correlations with their corresponding pair of each datasets.
    - Find slope and intercept for each datasets.
    - Find R-square for each datasets.
    - To find R-square first find residual sum of square error and Total sum of square error
    - Create a statistical summary by using all these data and print it.
- Find the descriptive statistical properties for the all four dataset
- Plot the scatter plot and linear regression line for each datasets



Note: It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

Explanation of this sample output, This sample graphs are just for understanding purpose:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

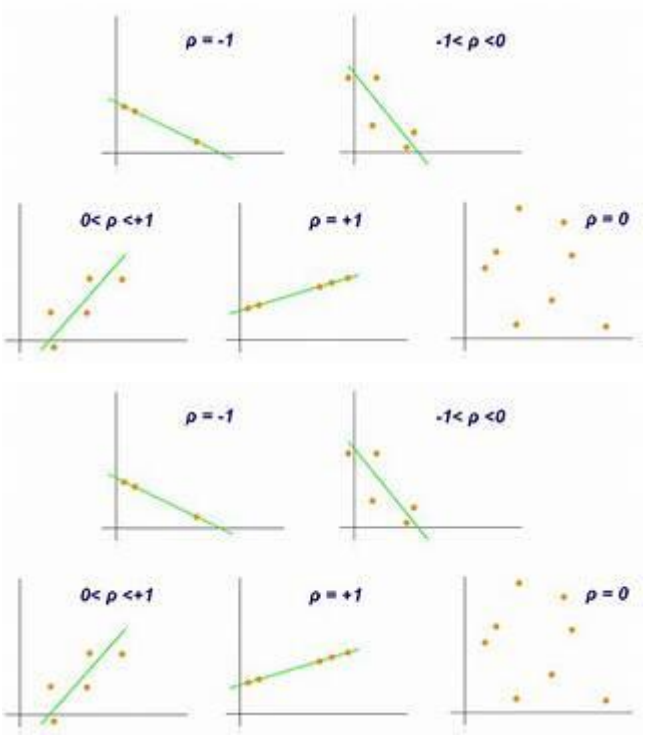
Question no 3:- What is Pearson's R

Answer :-

The Pearson correlation coefficient is a statistical measure of the strength and direction of a linear relationship between two variables. It is a number between -1 and 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation 1. The formula for calculating Pearson's r is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x and y are the two variables, n is the number of observations, and x_i and y_i are the i-th observations of x and y, respectivelyr can be calculated using a calculator or a spreadsheet program like Microsoft Excel



Like this we can see pearson Rs

This is the formulae in python which can be used to find pearson R

```
r = np.corrcoef(x, y)[0, 1]
```

```
import numpy as np

# x and y are two variables
x = [1, 2, 3, 4, 5]
y = [5, 4, 3, 2, 1]

# calculate Pearson's r
r = np.corrcoef(x, y)[0, 1]

print(f"Pearson's r: {r}")
```

Like this , This is just an example above to understand how we will do in python this Pearson R .

Question no 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. It also helps in speeding up the calculations in an algorithm. When we collect data set it 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 modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude. Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.

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

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).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Question no 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. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. 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).

Question no 6:- What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans. 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.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. The purpose of Q Q plots 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 being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.