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)

Below are the inferences:

- **2019 has seen more sales compared to 2018**
- **Weather is playing a key role in the Bikes Rented. "Clear, Few clouds, Partly cloudy" are seeing the highest numbers. There is NO data for " Heavy Rain + Ice Pallets" which could mean no Bikes are taken these days**
- **Temperature is playing a major role on the count of bikes being rented. During days when temperature is between 20 and 30, the rentals are highest**
- **Months from May to October see high number of bike rentals**
- **Fall and Summer see the highest Bike Rentals**
- **Spring sees the least bike rentals**

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

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. Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished

By discarding one dummy variable, we effectively remove the redundancy in the model caused by perfect multicollinearity. This allows the regression model to estimate the coefficients of the remaining dummy variables accurately and interpret their effects on the outcome variable independently..

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

Atemp and temp have the highest correlation with the target variable (0.63)

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

- **Validated the Variance Inflation Factor (VIF)**
- **Compared the R2 Score between the Training and Test Dataset**
- **Mean of Residuals is Zero**
- **Linear Relationship between Target and Predicted Varaible**

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

a) Temparature
b) Year
c) Season (Sunny and Winter)

General Subjective Questions

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

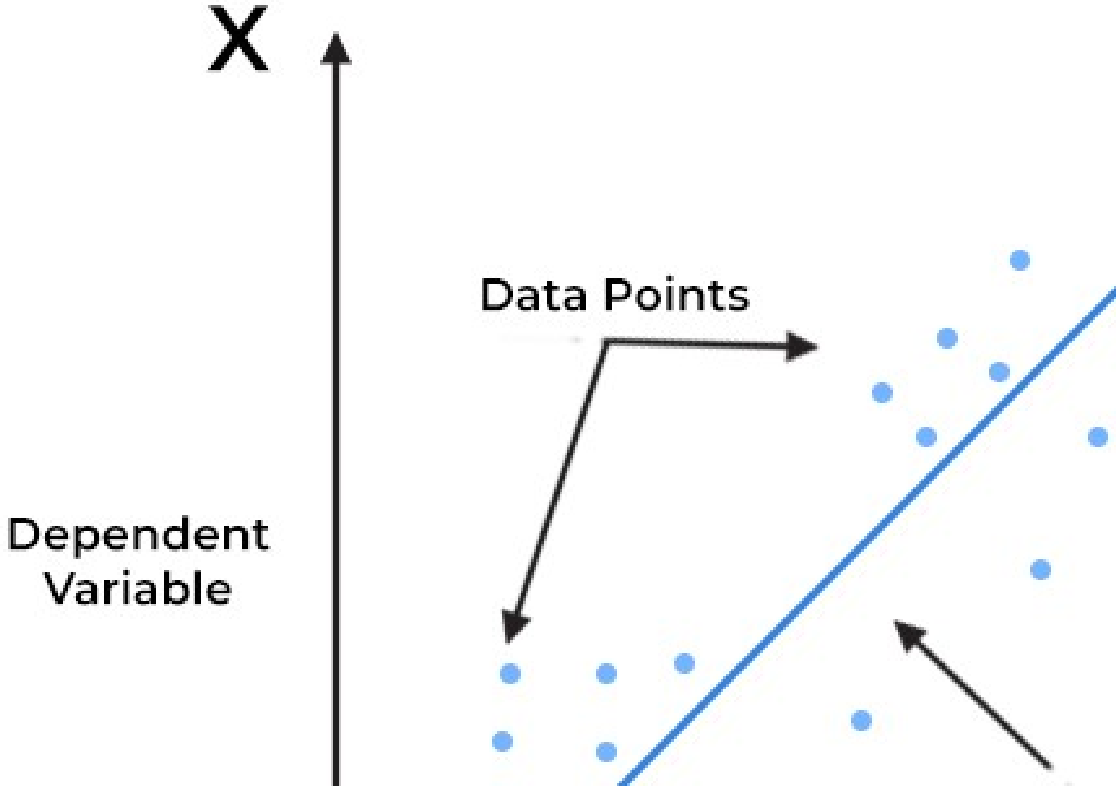
Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

This analysis method is advantageous when at least two variables are available in the data, as observed in stock market forecasting, portfolio management, scientific analysis, etc.

A sloped straight line represents the linear regression model.



In the above figure,

X-axis = Independent variable

Y-axis = Output / dependent variable

Line of regression = Best fit line for a model

Here, a line is plotted for the given data points that suitably fit all the issues. Hence, it is called the 'best fit line.' The goal of the linear regression algorithm is to find this best fit line seen in the above figure.

Key benefits of linear regression

Linear regression is a popular statistical tool used in data science, thanks to the several benefits it offers, such as:

1. Easy implementation

The linear regression model is computationally simple to implement as it does not demand a lot of engineering overheads, neither before the model launch nor during its maintenance.

2. Interpretability

Unlike other deep learning models (neural networks), linear regression is relatively straightforward. As a result, this algorithm stands ahead of black-box models that fall short in justifying which input variable causes the output variable to change.

3. Scalability

Linear regression is not computationally heavy and, therefore, fits well in cases where scaling is essential. For example, the model can scale well regarding increased data volume (big data).

4. Optimal for online settings

The ease of computation of these algorithms allows them to be used in online settings. The model can be trained and retrained with each new example to generate predictions in real-time, unlike the neural networks or support vector machines that are computationally heavy and require plenty of computing resources and substantial waiting time to retrain on a new dataset. All these factors make such compute-intensive models expensive and unsuitable for real-time applications.

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

Anscombe's Quartet, comprising four datasets with nearly identical summary statistics, underscores the limitations of relying solely on numerical metrics.

This article explores the quartet's datasets, emphasizing the importance of visualizing data for a comprehensive understanding.

What is Anscombe's Quartet?

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

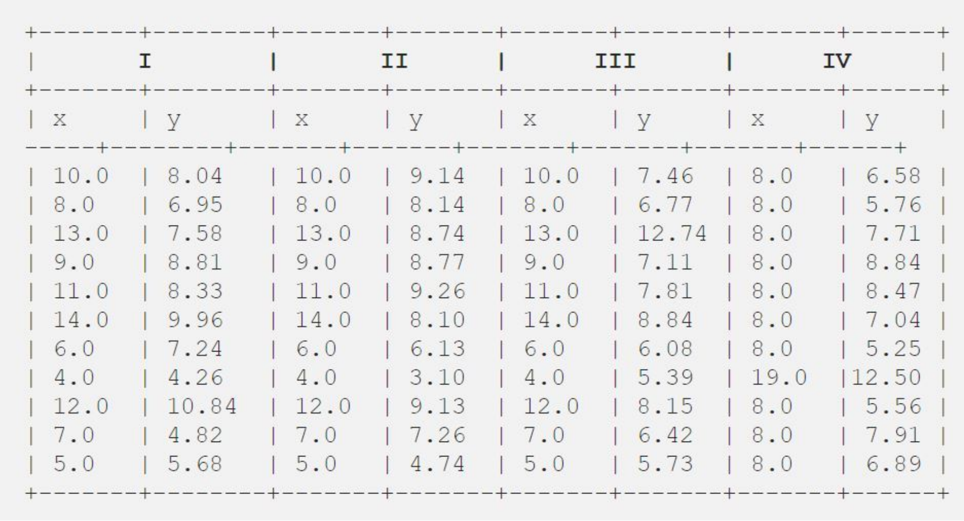
The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Purpose of Anscombe's Quartet

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.

Anscombe's Quartet Dataset

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

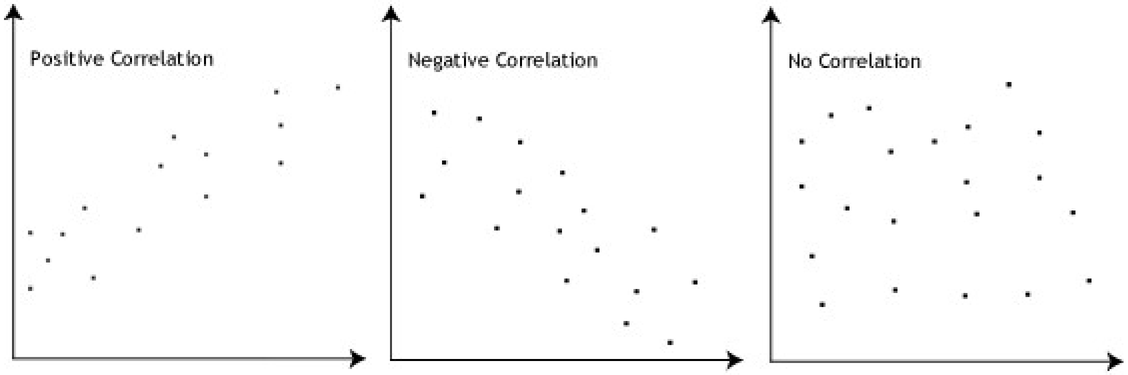


```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 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 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

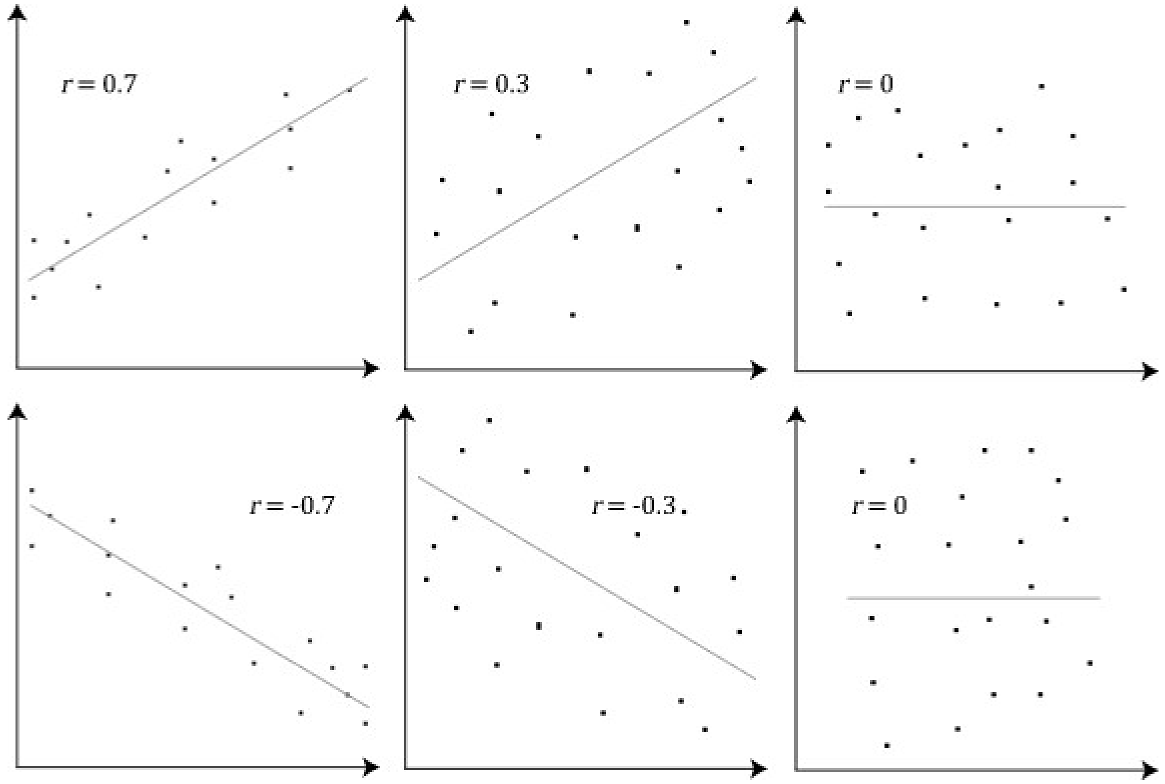
3. What is Pearson's R? (3 marks)

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? (3 marks)

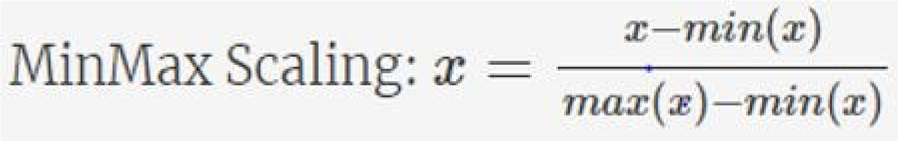
Scaling is step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

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 modelling. 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:

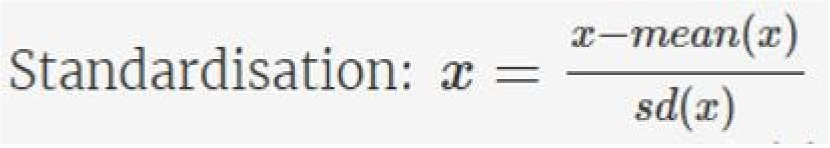
It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



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

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).



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

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 outlier

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

The Variance Inflation Factor (VIF) can be infinite when there is perfect correlation between independent variables. This can happen when one independent variable is strongly correlated with many other independent variables, or when the regressor is equal to a linear combination of other regressors. When this happens, the corresponding variable can be expressed exactly by a linear combination of other variables, which also have an infinite VIF.

Formula for VIF is given as:

VIF = $1/(1-R^2)$

Now, when you're calculating the VIF for one independent variable using all the other independent variables, if the R^2 value comes out to be 1, the VIF will become infinite. This is quite possible when one of the independent variables is strongly correlated with many of the other independent variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals

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

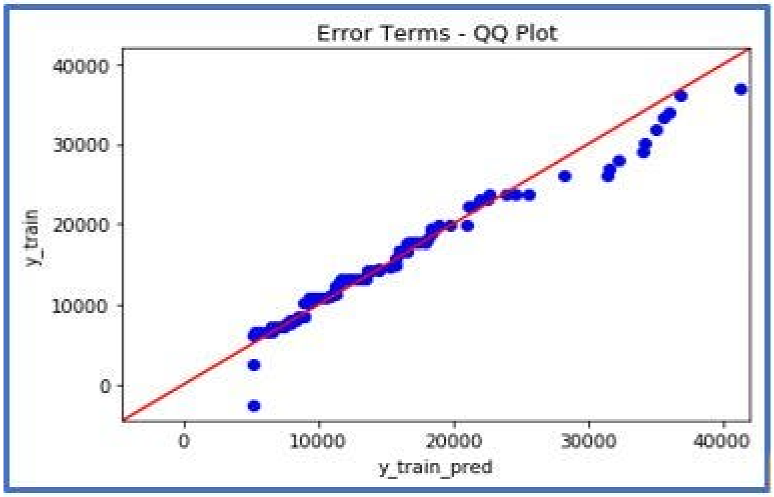
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

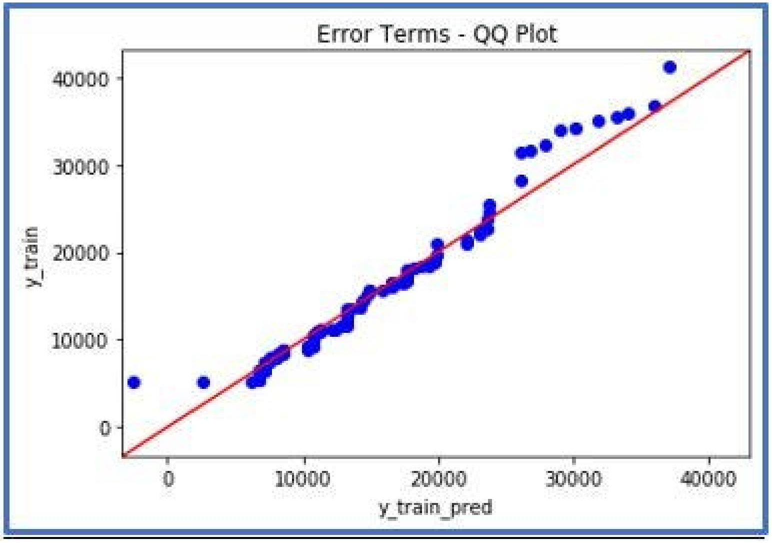
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Python:

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.