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:
   Here are my inferences about the effects of the categorical variables on the dependent variable:

   - January onwards booking keeps increasing month on month till September, then starts to decrease from October up to December.

   - Booking is lowest on Sundays and gradually increases towards the weekend.

   - Working days has more bookings than non-working days.

   - Holidays seems to be having less booking than non-holidays.

   - Booking increased spring season onwards, reached its peak during the fall and declined during the winters. The trend increased over the years.

   - Understandably, booking of bikes increased on those days when weather was clear.

   - Bike rental business has really increased significantly year on year.

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

   Syntax -
   drop_first: bool, default False, which implies whether to get k-1 dummies out of k
   categorical levels by removing the first level.

   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 A and B, then It is obvious C. So we do not need
   3rd variable to identify the C.

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

   I have validated the assumption of Linear Regression Model based on 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 auto-correlation

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

   Answer:

   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes
   - atemp
   - sep
   - winter

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a statistical technique used to model the linear relationship between a dependent variable y and one or more independent variables (also called predictor or explanatory variables) X. It assumes that there is a linear relationship between the independent variables and the dependent variable. The goal of linear regression is to find the best-fit line that represents this relationship in the simplest possible way. It aims to predict the value of the dependent variable based on the values of the independent variables.

The following are some assumptions about dataset that is made by Linear Regression model :

- Multi-collinearity – Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

- Auto-correlation –Linear regression model assumes that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

- Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.

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

- Homoscedasticity – There should be no visible pattern in residual values.

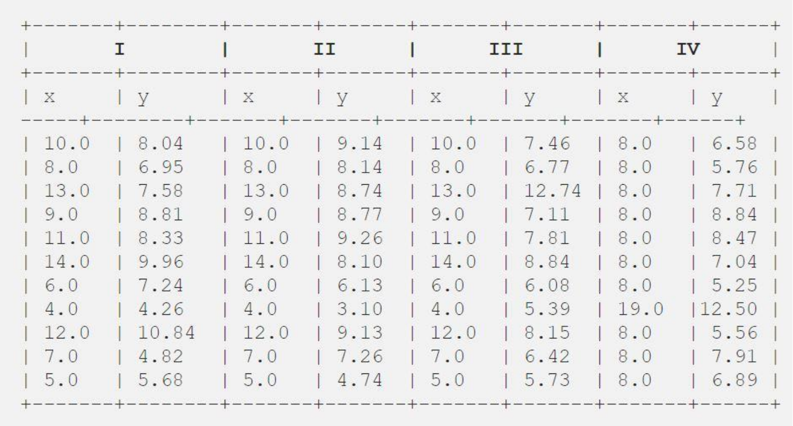
The linear regression algorithm consists of the following steps:

- Data collection: Obtain a dataset containing the dependent variable and at least one independent variable.
- Data preparation: Clean, transform, and normalize the data to make it suitable for analysis. Split the data into training and testing sets.
- Model selection: Select the type of linear regression model to use, such as simple linear regression (with one independent variable) or multiple linear regression (with multiple independent variables).

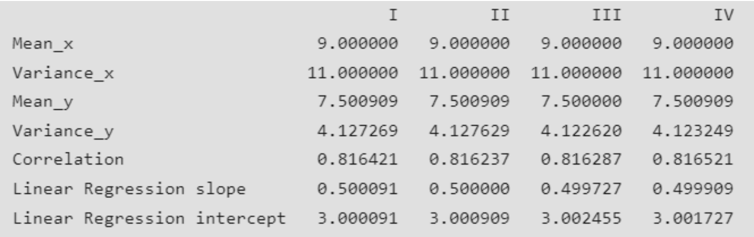
- Model training: Use the training set to estimate the model parameters, such as the intercept and the coefficients of the independent variables, using methods such as Ordinary Least Squares (OLS) or Maximum Likelihood Estimation (MLE).
- Model evaluation: Evaluate the model's performance on the testing set using metrics such as Mean Squared Error (MSE) or R-squared, which measure the goodness-of-fit between the predicted values and the actual values of the dependent variable.
- Model refinement: If the model performance is not satisfactory, refine the model by adding or removing independent variables, checking for collinearity among the independent variables or applying other techniques to improve the model's predictive power.
- Model deployment: Once the model is deemed satisfactory, the final model can be applied to new data for prediction.

2. Explain the Anscombe’s quartet in detail.
   Answer:

Anscombe’s quartet is a collection of four datasets that have nearly identical statistical properties, but look very different when plotted. The quartet was created in 1973 by the statistician Francis Anscombe as a demonstration of the importance of graphical data exploratory tools. Each dataset consists of 11 data points and contains two continuous variables: x and y.

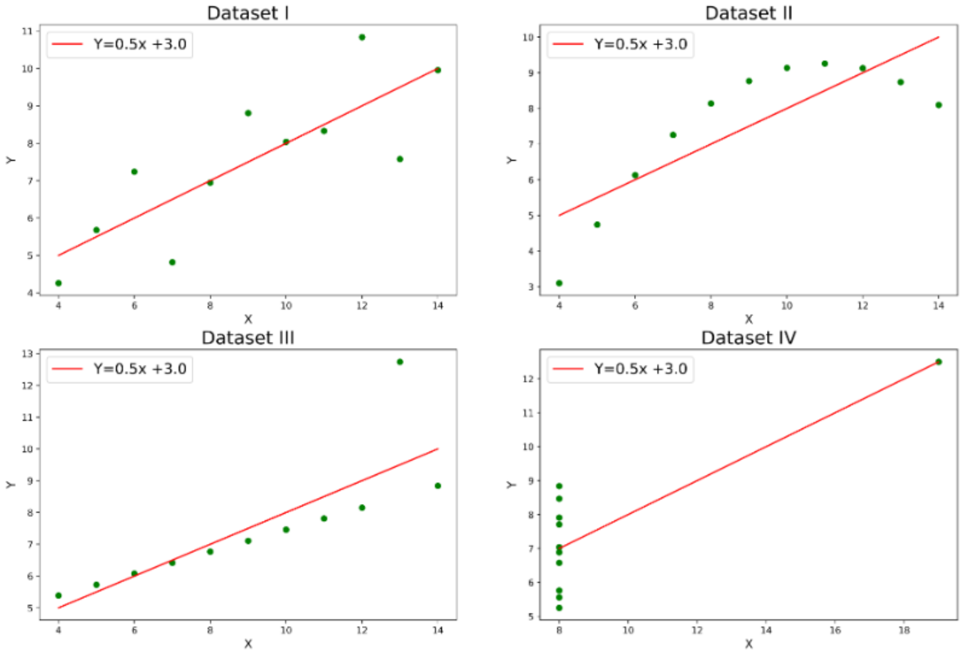


```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+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



```
                                   I         II        III         IV
Mean_x                      9.000000   9.000000   9.000000   9.000000
Variance_x                 11.000000  11.000000  11.000000  11.000000
Mean_y                      7.500909   7.500909   7.500000   7.500909
Variance_y                  4.127269   4.127629   4.122620   4.123249
Correlation                 0.816421   0.816237   0.816287   0.816521
Linear Regression slope     0.500091   0.500000   0.499727   0.499909
Linear Regression intercept 3.000091   3.000909   3.002455   3.001727
```

Despite having the same mean, variance, correlation, and linear regression model, the four datasets in Anscombe’s quartet exhibit very distinct patterns:



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

Through Anscombe's quartet, we can understand that the tendency to rely solely on summary statistics and not visually inspecting the data can lead to erroneous conclusions. Therefore, Anscombe's quartet serves as a powerful reminder that data exploration is essential in Data Science and that data visualization is a fundamental tool for understanding and communicating insights in a clear and meaningful way.

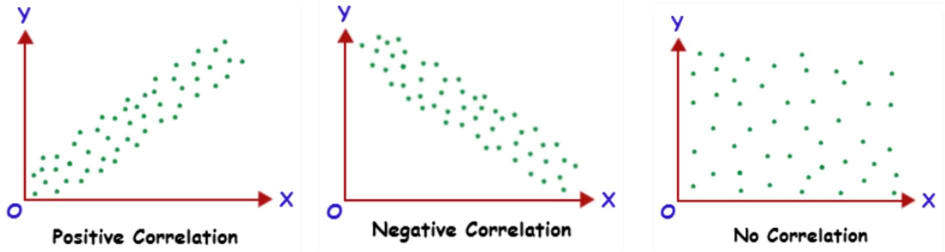
3. What is Pearson's R?

Answer:

Pearson's R (commonly referred to as Pearson correlation coefficient) is a measure of the linear correlation between two continuous variables x and y. It gives us an indication of the strength and direction of the relationship between the two variables. This correlation coefficient is named after the English mathematician and statistician Karl Pearson who developed the statistic in 1896.

Pearson's R ranges in value from -1 to 1. When the correlation is positive and strong, the value of R is closer to 1; when the correlation is negative and strong, the value of R is closer to -1; and when there is no correlation between the two variables, the value of R is 0.

The graphical representation of positive, negative and no correlation is shown below:



The formula for Pearson's R is:
$R = (\Sigma(x - \mu_x)(y - \mu_y)) / (\sqrt{}(\Sigma(x - \mu_x)^2 \ (\Sigma(y - \mu_y)^2))$
Where:
x and y are the values of the two variables we want to measure the correlation between
μ_x and μ_y are the mean values of x and y, respectively
The numerator represents the sum of the products of the deviations (how far each value is from its mean) of x and y from their respective means
The denominator represents the product of the standard deviations of x and y

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer:
Scaling is a technique in data preprocessing that transforms the data by making it more comparable or appropriate for use in certain types of analyses or models. The purpose of scaling is to bring different variables to the same scale or range, so that they can be compared on equal footing.

If feature scaling is not done, then different variables are compared based on their values even if their units are different, thus leading to incorrect inferences.

Difference between Normalization and Standardization:

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.

S.NO.	Normalization	Standardization
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

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 multicollinearity in multiple linear regression models. It indicates how much the variance of the estimated coefficient for a particular predictor variable is increased due to multicollinearity among the other predictor variables.

   The VIF value can be infinite when there is perfect multicollinearity among the predictor variables due to linear combinations of the data or the variables. Perfect multicollinearity can occur due to a variety of reasons, such as having duplicate or highly correlated predictor variables in the dataset, or using a dataset that is too small or not representative.
   In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2)
   infinity. To solve this we need to remove 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.

   Answer:
   A Q-Q plot (Quantile-Quantile plot) is a graphical technique used to assess whether a set of data follows a certain distribution, such as a normal distribution. The Q-Q plot compares the sample data to the theoretical distribution by plotting the quantiles of the sample data against the quantiles of the theoretical distribution. If the data perfectly follows the specified distribution, the points on the Q-Q plot will fall onto a straight line.

   In the context of linear regression, Q-Q plot is used to check the normality assumption of the errors or residuals. Linear regression models assume that the errors or residuals are normally distributed with mean zero and constant variance. The Q-Q plot can be used to visually assess whether or not the residuals are normally distributed.

   If the residuals are normally distributed in a linear regression model, the Q-Q plot will appear as a straight line. However, if the residuals are not normally distributed, the Q-Q plot will indicate curves or bends in the line, suggesting non-normality.

   The Q-Q plot is important in linear regression because normally distributed errors are a critical assumption for valid statistical inference of the model results. Violation of this assumption can lead to inaccurate parameter estimates, standard errors, and confidence intervals. Therefore, the Q-Q plot is a valuable tool for checking the normality assumption and ensuring that the linear regression model results are reliable.