# Bike Sharing Assignment
# Student Name: Debasish Mondal, Co-hort: May-2022
# Submission Date: 10th August, 2022

## 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)**

**Answer:** Based on the analysis performed on the categorical columns using the bar plot, we arrive at the following findings:

- The number of bookings is highest in the `fall` season.
- The month of `June` has the highest number of bookings.
- The majority of bookings are on `Friday`.
- The number of bookings is highest in the `clear (or few clouds or partly cloudy)` weather.
- The number of bookings is highest in the year `2019`.
- The number of bookings is greatest if it is not a `holiday`.
- The number of bookings is greatest if it is a `working day`.
- The most bookings are made when the `temperature` is between 22.0 and 25.0 degrees Celsius.
- The most bookings are made when the `feeling temperature` is between 27.0 and 31.0 degrees Celsius.
- The most bookings are made when the `humidity` is between 66.0 and 74.0 units.
- The most bookings are made when the `wind speed` is between 5.0 and 8.0 units.

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

**(2 mark)**

**Answer:** Use of 'drop first = True' reduces the correlations between dummy variables. During the process of transformation of categorical variable to dummy variables, many excess columns are produced, which required to be eliminated.

If we apply 'drop_first = True' on a k-level categorical variable, then the 1st dummy variable out of the k dummy variables created will be eliminated.

**Interpretation:** For a k-level categorical variable k-1 dummy variables are sufficient to build the model.
**Example:** Let us consider a categorical variable 'Gender' with two levels: 'Male' and 'Female'. For this, only 1 dummy variable is sufficient to describe: '0 for Male' and '1 for Female'.

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

**Answer:** The target variable ('cnt') and the variable temperature ('temp') have the highest correlation.

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

**Answer:** We validate the assumptions of Linear Regression after building the model on the training set in the following ways:

- Correlation Checking: By checking correlation between all the numerical variables, we first need to figure out which variables are highly correlated and then delete suitable variables from them.
- p-value Checking: If p-values of all variables are less than 0.5 , then model building is good.
- Multicollinearity Checking: If the Variance Inflation Factor (VIF) values of all variables are less than 5, then the model is free from multicollinearity.
- Residual Analysis: First, plot the distribution of the residual term (i.e., y_train_predicted y_train). Now, if the distribution is normally distributed with the zero mean, only then the model is good.

**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 three features that significantly contribute to explaining the demand for shared bikes are shown below.

- Feeling Temperature i.e. 'atemp' variable
- Light Weather (Snow/Rain/Thunderstorm/Scattered clouds etc.) i.e. 'weathersit_light' variable
- Year i.e. ‘yr’ variable

## General Subjective Questions

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

**Answer:** Linear regression model helps to identify the linear pattern between dependent and independent variables of a dataset. It is two types –

- Simple Linear Regression (SLR): In this case the number of dependent variable is 1.
- Multiple Linear Regression (MLR): In this case the number of dependent variable is more than 1.

Mathematically the model is described by the following equation –

Simple Linear Regression: $y = c + mx$

Multiple Linear Regression: $y = c + m_1x_1 + m_2x_2 + \cdots + m_kx_k$

where $y$ is the dependent or response variable and $X \equiv (x_1, x_2, \dots, x_k)$ is the independent or predictor variable.

**Algorithm:**

- **Data Reading:** Read the dataset and identify the response and predictor variables.

- **Data Standardization:** Remove null-valued data from the data set and present each of the variables in a standardized format so that they can be used in model analysis.
- **Data Cleaning:** Remove outliers from the numerical data.
- **Data Handling:** Remove multicollinearity from the dataset by checking correlation values.
- **Data Split:** Split the data into training and test parts.
- **Data Scaling:** Apply scaling to both the train and test data. But only apply fit to the test data.
- **Model Preparation:** Based on the training data, make an effective model by selecting appropriate features using the Recursive Feature Elimination (RFE) method.
- **Model Evaluation:** Check whether the error term is normally distributed with mean zero or not.
- **Model Prediction:** Predict for test data using the model learned from training data.
- **Final Decision:** If the difference between the $R^2$ (or Adjusted $R^2$) values of the training and test model is less than 5%, then the model is appropriate for business purposes.

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

**Answer:** Anscombe's quartet comprises four data sets with nearly identical mean and variance, though they have different distributions. This was developed by Francis Anscombe in 1973 to demonstrate the effect of outliers and other influential observations on statistical properties.

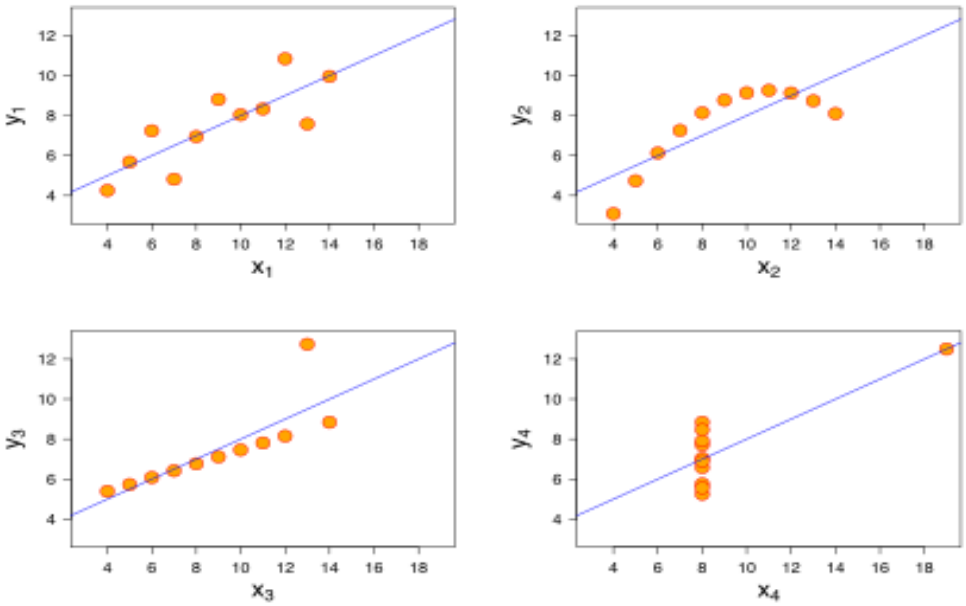


*Figure: Four different datasets with nearly identical mean and variance.*

*Image Credit: Wikipedia*

- The first scatter plot (top left) predicts a simple linear relationship between$x$ and $y$, as it appears.
- The second scatter plot (top right) predicts a simple linear relationship between $x$ and $y$, though it is non-linear.
- The third scatter plot (bottom left), predicts a simple linear relationship between $x$ and $y$, but should have a different regression line due to the presence of an outlier value.
- The fourth scatter plot (bottom right) predicts a simple linear relationship between $x$ and $y$, though data points do not indicate any specific relationship.

Thus, all four datasets are different in nature but have nearly identical mean and variance.

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

**Answer:** Pearson's R value is a measure of linear correlation between two numerical variables. Its values range from -1 to +1.

If Pearson's R > 0, then there is a positive linear relationship between the two variables, i.e., one will increase if the other increases.

If Pearson's R = 0, then there is no linear relationship between two variables, but a non-linear relationship may exist.

If Pearson's R < 0, then there is a negative linear relationship between the two variables, i.e., one will increase if the other decreases.

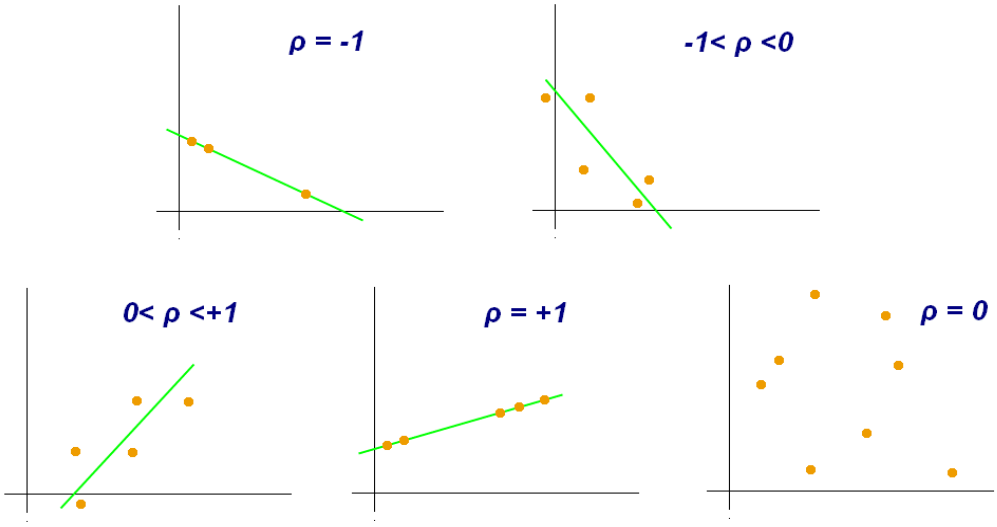


*Figure: Relationship between two variables for different Pearson's R values.*

*Image Credit: Wikipedia*

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

**Answer:**

- Scaling is a mathematical way to standardize the independent variables of the data in a specific range (e.g., say between 0 and 1, or between minimum and maximum).
- Scaling is performed to bring all data variables to the same magnitude, so that model prediction can be done in a more effective way regardless of the unit or magnitude of data.
- Difference between normalized scaling and standardized scaling:

| Normalized Scaling | Standardized Scaling |
|---|---|
| 1. Usually scaled down the values between 0 and 1. | 1. Usually scaled down the values between column's minimum and maximum value. |
| 2. It is affected by outliers. | 2. It does not much affected by outliers. |

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

**(3 marks)**

**Answer:**

- An infinite VIF value signifies the perfect mutual correlation between a variable and other variables in the dataset.
- If a variable is completely expressed as a linear combination of all other remaining variables, then only the VIF value of that variable becomes infinite.

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

**(3 marks)**

**Answer:**

- **Definition:** The Q-Q (or Quantile-Quantile) plot shows the distribution of the data against the expected normal distribution.
- **Use:** It checks whether two data sets come from a common population or not. A reference line is plotted for this. If the datasets come from the same population, then the points should fall along that reference line, else their chances to come from two different populations become greater.
- **Importance:** It provides more insight into the characteristics of the difference between two given datasets than traditional analytical methods like chi-square test etc.

**************************************END**************************************