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?

Ans -

- The count of rental bikes is higher in the fall season followed by summer and winter but it's lower in spring .
- The count of rental bikes is higher in the months of June , July ,August and September .
- The count of rental bikes is lower on holidays .
- The count of rental bikes is higher on thursday followed by saturday and sunday .
- The count of rental bikes is slightly lower on working days .
- The count of rental bikes is higher if the weather situation is Clear or Few clouds and lower in case of light snow .

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

Ans-

- Suppose, we have a column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".
- So We don’t need another column for "Unknown".
- That's why when we make a dummy variable for a categorical column we always use drop_first=True .

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

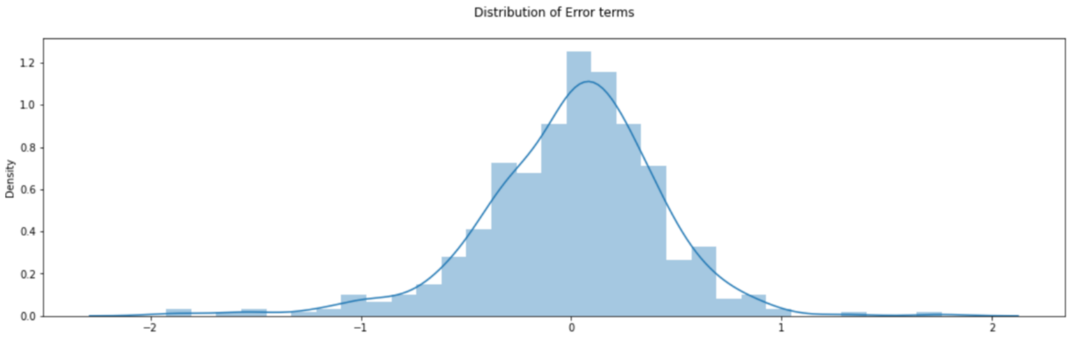
Ans -

- Column ‘registered’ is highly correlated with the target variable .

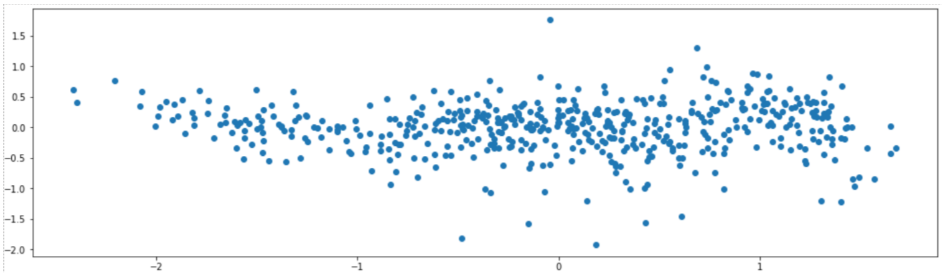
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans-

- First we have to find the residuals and then plot a histogram of the residuals
  Like this → y_train_pred = lm.predict(X_train)
  Residuals = y_train - y_train_pred
  sns.distplot(Residuals)
  plt.show()



- And for checking the variance of error terms we plot a scatter plot
  Like this → plt.scatter(y_train_pred, res)
  plt.show()



- If there is no such pattern in the scatter plot then we can conclude that the error terms have constant variance or homoscedasticity .
- If error terms are normally distributed with mean zero and having constant variance then we successfully validate the assumptions of Linear Regression .

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

Ans-

- The top3 features are
  - I. Temperature
  - II. Light snow weather
  - III. Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

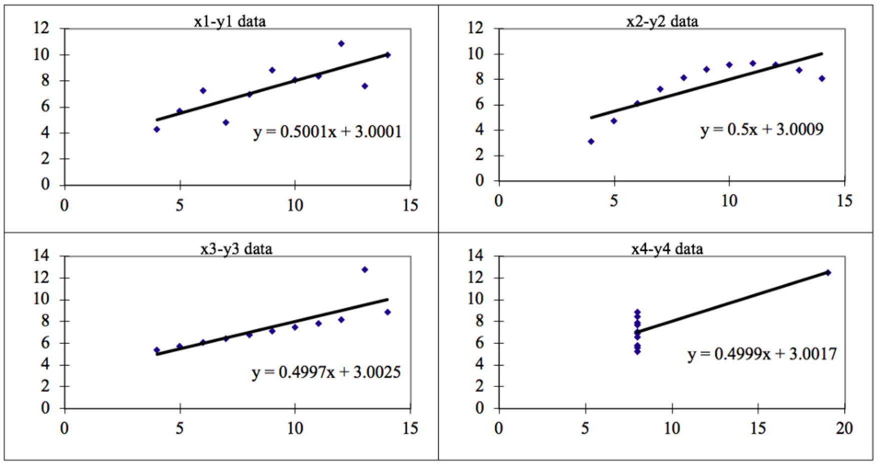
Ans-

- Linear Regression is a machine learning algorithm based on supervised learning ,it performs the task to predict a dependent variable value (y) based on a given independent variable (x) .So ,this regression technique finds out a linear relationship between x(input) and y(output) . Hence ,the name is Linear Regression .
- The equation of a linear equation is y = ß0 + ß1*X1 + ß2*X2 + ….+ ßn*Xn
  Where ß0 = intercept and ß1,ß2,...,ßn are coefficients of respective predictor variables .
- When training the model it fits the best line to predict the value of y for a given value of x . The model gets the best regression fit line by finding the best intercept(ß0) and coefficients(ßn) values .
- 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 intercept(ß0) and coefficient values, to reach the best value that minimizes the error between predicted y value and true y value .
- In Linear Regression we have to minimize the Root Mean Squared Error (RMSE) also called cost function between predicted y value and true y value .
- So to update intercept(ß0) and coefficient values in order to reduce the cost function and achieve the best fit line the model uses Gradient Descent . The idea is to start with random intercept(ß0) and coefficient values and then iteratively updating the values, reaching minimum cost.

2. Explain the Anscombe’s quartet in detail.

Ans -

- Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



- Conclusion -
  All the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model .

3. What is Pearson's R?

Ans-

- Pearson's r is a numerical summary of the strength of the linear association between the variables. 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.
- The Pearson's correlation coefficient varies between -1 and +1 where:
- Formula is $$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$
- 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

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

Ans -

- 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.
- Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling occurs . To solve this issue, we have to 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.

- from sklearn.preprocessing import MinMaxScaler helps to implement normalization in python.
- Formula is $\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$
- **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).
- from sklearn.preprocessing import StandardScaler helps to implement standardization in python.
- Formula is $\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$
- Disadvantage :
  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?

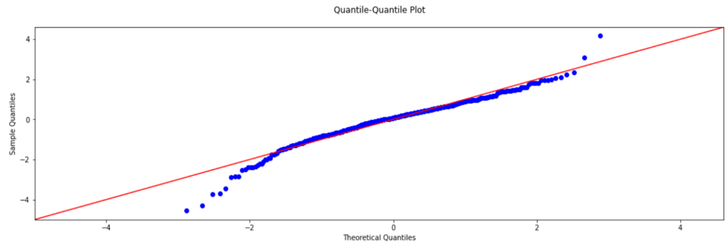
Ans-

- 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 leads 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).

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

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 the same distributions.
- 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.
- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.
- For example, if a given distribution needs to be verified if it is a normal distribution or not, we run statistical analysis and compare the unknown distribution with a known normal distribution. Then by observing the results of the Q-Q plot, we can confirm if the given distribution is normally distributed or not.
- statsmodels.api provides qqplot and qqplot_2samples to plot Q-Q graphs for single and two different data sets respectively.


Quantile-Quantile Plot
Sample Quantiles
Theoretical Quantiles