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)**

Ans: The categorical variables are **"season", "workingday", "weathersit", "weekday", "yr", "holiday",** and **"mnth".**

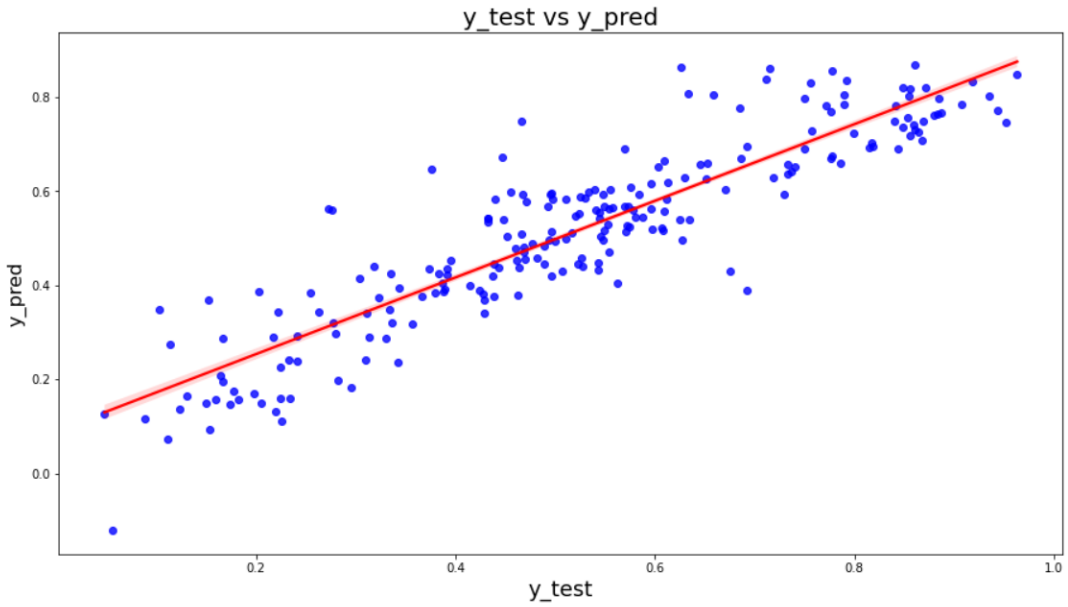
- **"season":**
    - We have observed that the most fitted season for bike rent is "summer" and Fall.
    - Spring season have lowest consuming ration for bike rents.
- **"workingday":**
    - Subscribed user preferred the bike rent on weekday but guest user prefer the bike rent on weekends.
    - With the help of registered and guest user, we can identify working day and non-working day.
- **"weathersit":**
    - The most favorable weather is clean/few clouds days.
    - The subscribed users count is high when the light rainy days, so that mean registered user go to the office more often with rental bike
- **"weekday":**
    - For count variable, we couldn't see much changes with weekday column.
    - We have observed that bike usage is higher on working days which comes under registered user.
- **"yr":**
    - We have 2 year of data and observed that in 2019 the broom bike has more business compare to 2018
- **"holiday":**
    - Compare registered user, guest user taking more bike on rent in holiday.
- **"mnth":**
    - In June, July, August, September and October the bike rent is in higher ratio.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
Ans: Using dummy variable creation, we are change the variable into categorical variable. Every dummy variable, we derive into 0 and 1 values.

Here **drop_first=True,** we are using while creating dummy variable and drop the base/ reference category. This is helpful to provide multi-collinearity and getting added into the model if all dummy variable.

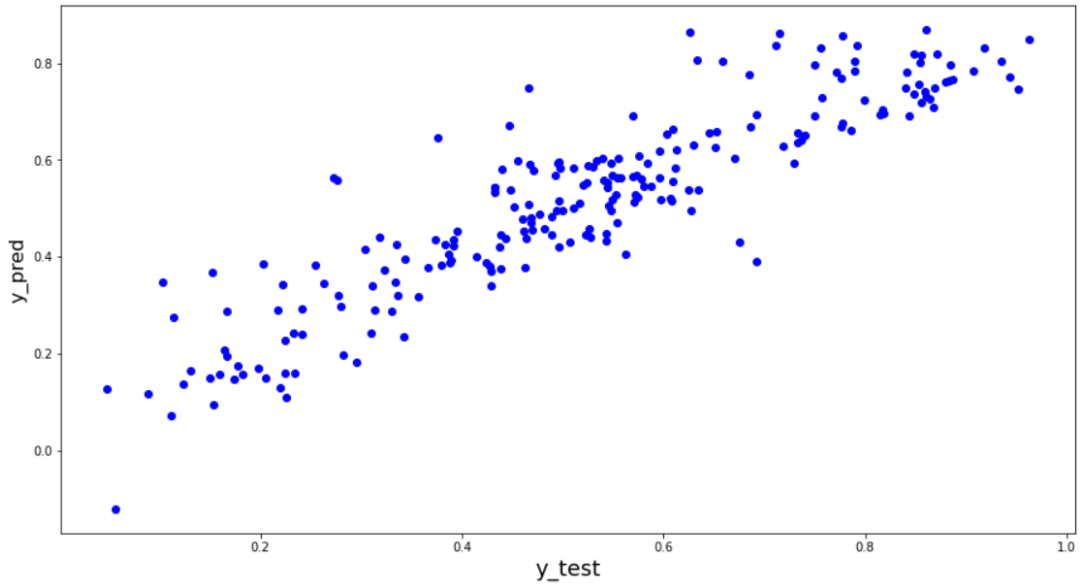
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
Ans: "Temp" is the highest correlation with the target variable.
The sum of the registered and guest user highest correlation with target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
1. We have validated the Linear Regression with e actual vs predicted plot as shown in the below figure.



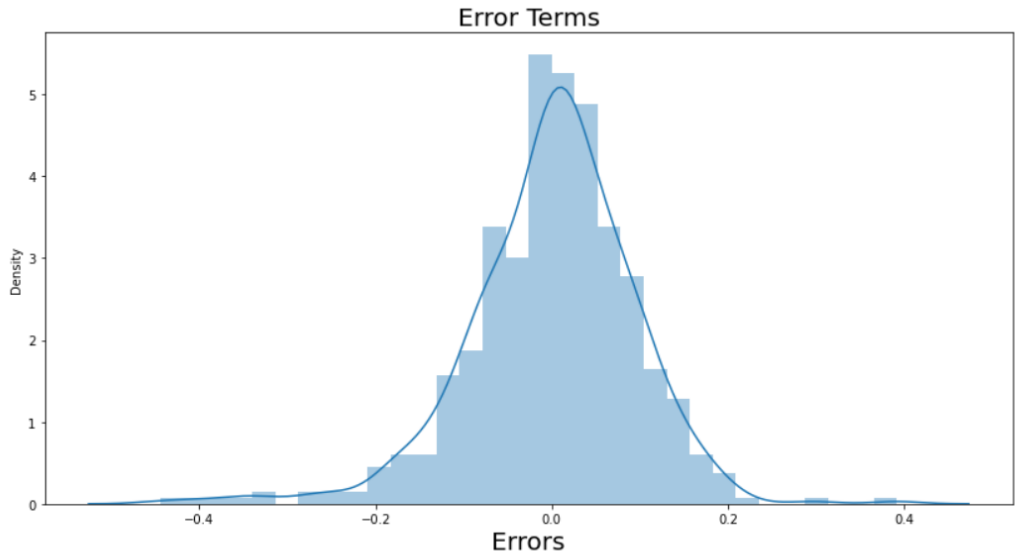
2. We can see no specific indication for errors terms. So, we can say error terms are depends on each other.

<Figure size 432x288 with 0 Axes>



3. In Histogram and distribution plot the normal distribution of error terms along mean is 0. Refer the below figure.

<Figure size 432x288 with 0 Axes>



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

Ans: A/c to data set, we can say that the top most features are:

1. Weathersit :
   Temp. is most significate for business, because we can see that Raining, Humidity, Windspeed and Cloudy weather effect the business in the negative manner.
2. 'Yr':
   The growth of the business is determined by the model whether in which year business is more or less.

3. ‘season’:
   Winter session took measure role demand of the bikes.

General Subjective Questions

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

- Linear regression helps to find best linear relationship within the independent variables and dependent variables.
- This algorithm gives the best fitted line between n independent variables with dependent variable.
- There are 2 types of linear regression algorithms
    - Simple Linear Regression → We have single independent variable.

      $Y = \beta 0 + \beta 1X$
    - Multiple Linear Regression – In Multiple regression, we have multiple variables.

      $Y = \beta 0 + \beta 1X1 + \cdots + \beta pXp + \in$

      Here

      $\beta 0 \rightarrow intercept$

      $\beta 1 \rightarrow$ Slope
- The prediction line equation would be $Ypred = \beta 0 + \beta 1xi$ and the actual Y is as Yi
- RSS = = $\sum (yi - ypred)$ 2

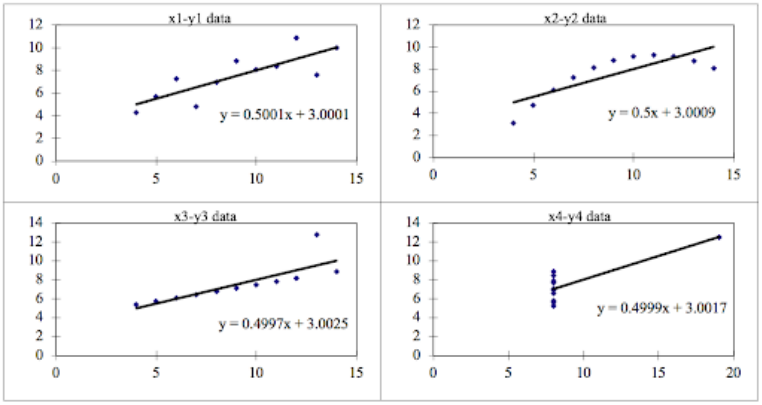
  The best fit is obtained by minimizing a quantity called RSS.

  Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.
- Cost Function
    - Differentiation
    - Gradient Desent Approach.

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

Ans: Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some difference in the dataset that fools the regression model if built.

They have very different distributions and appear differently when plotted on scatter plots



This show the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all (x,y) points in all four datasets. 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 in order 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

The four datasets can be described as

Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

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

- 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:

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

This is a Pre-Processing step 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, data set contains features which are 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.
1- Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1
2- 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).

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

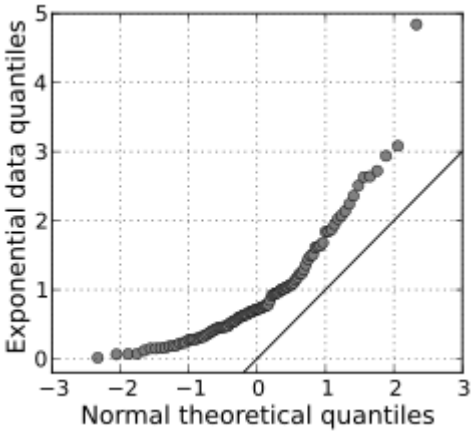
- The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model.
- It is used for find collinearity/multicollinearity.
- Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.
  VIF = 1/1-R^2

If there is perfect correlation, then VIF = infinity.
A large value of VIF indicates that there is a correlation between the variables.

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

Ans: Q-Q Plots (Quantile-Quantile plots) are of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. Q Q plots purpose is to find out whether 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.

A Q Q plot showing the 45-degree reference line:


Exponential data quantiles
Normal theoretical quantiles
0
1
2
3
4
5
−3
−2
−1
0
1
2
3