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: These are the effects of the categorical variables on the bike rental demand.

- *year* affects bike demand positively.
- People bike more during summers and winters.
- When the weather is either “Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist” or “Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds”, bike demand drops (drastically so) during the latter.
- Demand for bike rentals is less on holidays.
- Demand for bike rentals is slightly more on weekdays.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans: It is important to have **drop_first=True** during dummy variable creation. This is because we usually need *n-1* dummy variables if there are *n* levels. Having **drop_first=True** ensures that one of the dummy variables is eliminated.

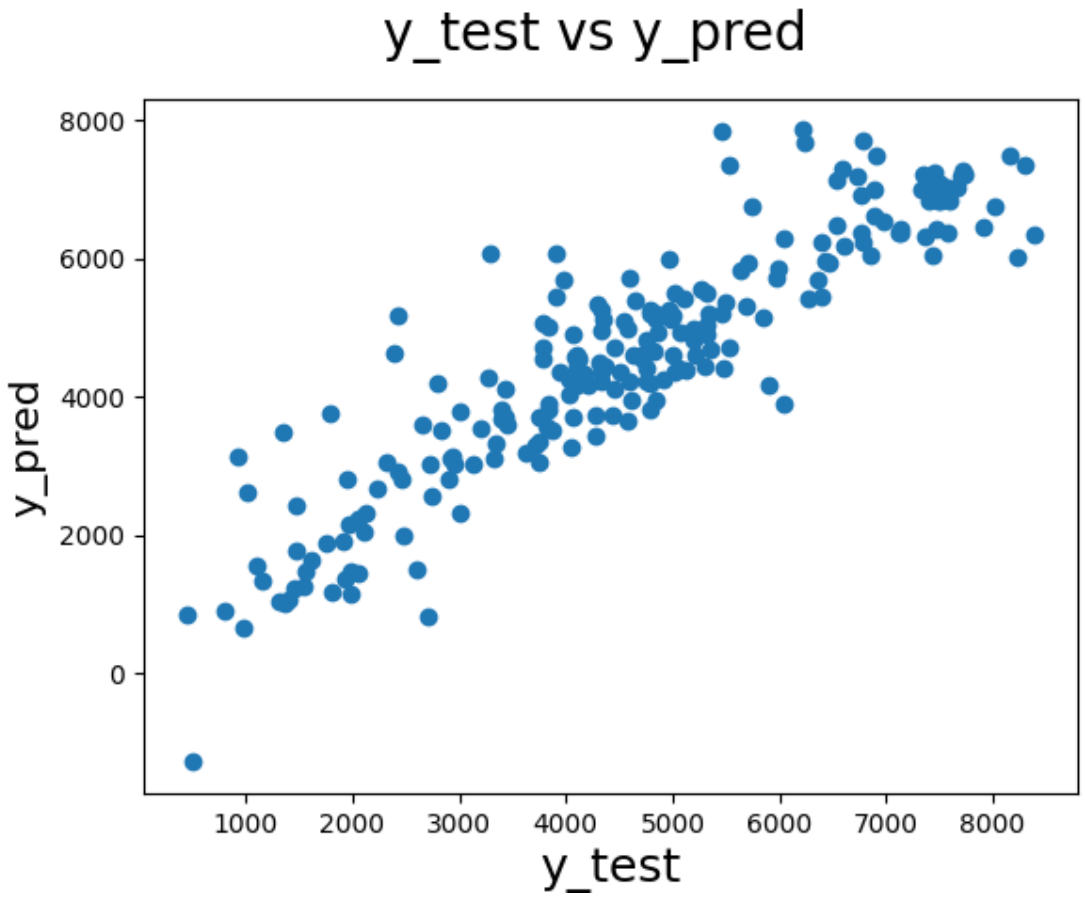
Also, keeping all generated dummy variables also results in Variance Inflation Factors for those **dummy variables hit Infinity**, which can lead to potentially wrong decisions when dropping columns.

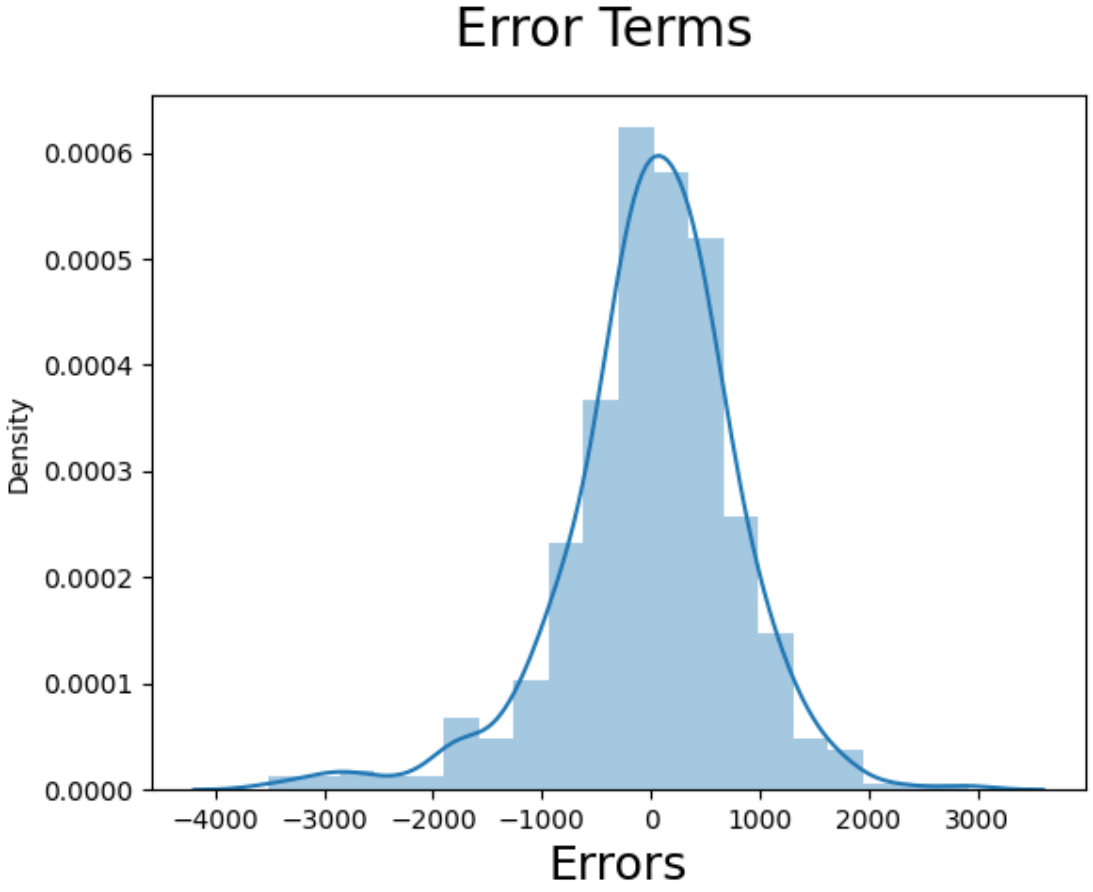
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Ans: *registered* has the highest correlation with the target variable, but since we believe *registered* and *casual* are separate target variables, the next useful independent predictor would be *temp* and *atemp*.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Ans: These were the steps taken to verify the assumptions of Linear Regression.

- The distribution of errors (residuals) was plotted to see if it conformed to a Normal distribution.
- The model was used to predict the values of the target variable using the **Test** data set, and the predicted values were graphed against the actual values of the target variable.

As can be seen below, the error terms conform to a Normal Distribution, and the model seems to generalise well to the **Test** set.


y_test vs y_pred
y_pred
y_test
8000
6000
4000
2000
0
1000
2000
3000
4000
5000
6000
7000
8000


Error Terms
Density
Errors
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
−4000
−3000
−2000
−1000
0
1000
2000
3000

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: The top 3 features contributing to the demand of shared bikes are:

- Temperature: positive, +4298.769
- Weather corresponding to “Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds”: negative, -2150.311
- Year: positive, +2019.478

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Ans: Assuming the matrix form of a linear relationship:

$$\mathbf{XB=Y}$$

We are given a vector X, and Y as the output, we are asked to find B which satisfy this equality. X will usually be a tall matrix, so we simply cannot take its inverse, however we can take the inverse of a square matrix, which we can obtain if we multiply both sides with X^T.

Multiplying both sides with $\mathbf{X^T}$, we get**:**

$$\mathbf{X^TXB=X^TY}$$

Multiplying again both sides with $\mathbf{(X^TX)^{-1}}$, we get**:**

$$\mathbf{(X^TX)^{-1}(X^TX)B=(X^TX)^{-1}X^TY}$$
$$\mathbf{=> B=(X^TX)^{-1}X^TY}$$

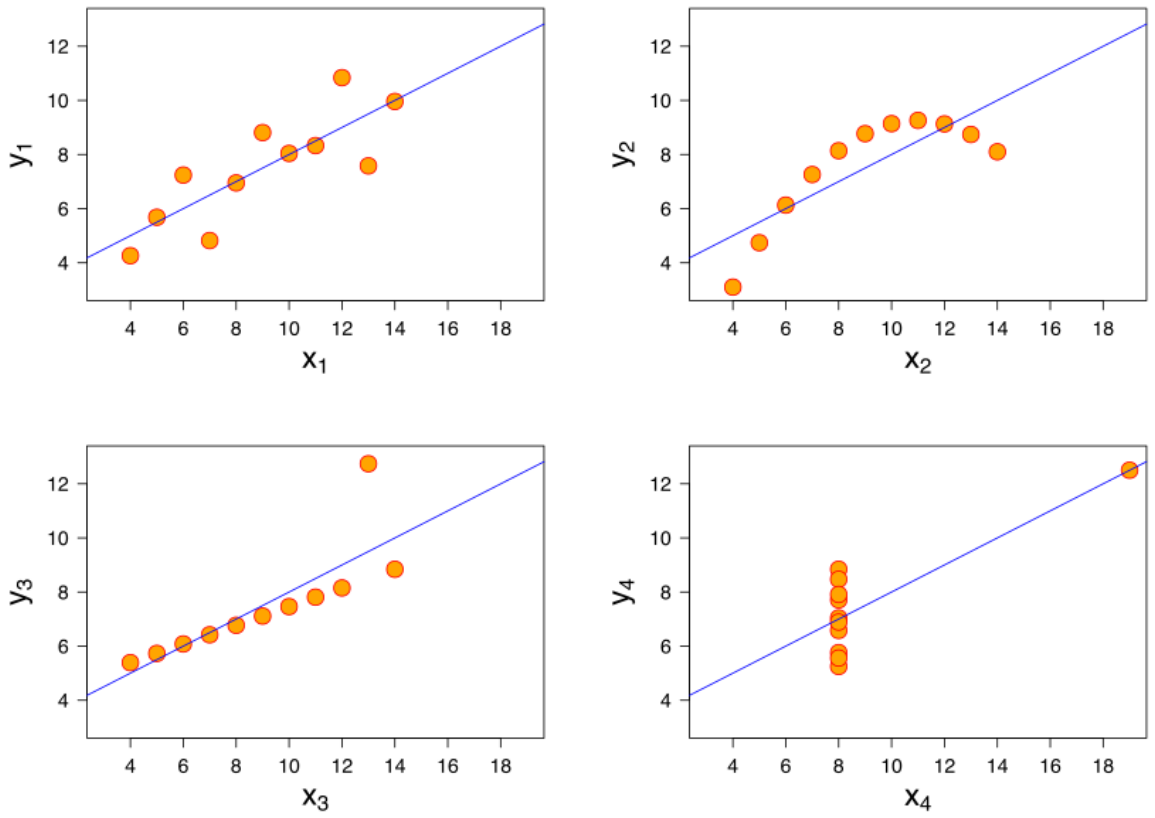
This can be solved using easy **matrix algebra** (finding inverses through **Gauss-Jordan**, **LU Decomposition**, etc.)
The same formula above can be arrived at by **minimising the square error.** Assuming that the error between XB and Y is E, the square error is:

$$\mathbf{|E|^2=(XB-Y)^T(XB-Y)}$$

Expanding and **differentiating** to allow for minimisation of the squared error, you can arrive at the same expression as above, which in fact, is where the minimum least squares technique gets its name.

2. Explain the Anscombe’s quartet in detail. (3 marks)
Ans: The **Anscombe’s Quartet** is a set of **four data sets** which are designed to demonstrate that descriptive statistics can be misleading at times, because all the data sets give the same metrics, but they look very different when graphed.

This set of data sets is intended to demonstrate the danger of blindly depending upon metrics when building a model, and highlights the importance of doing **Exploratory Data Analysis**.



- The first plot can be fit with Linear Regression, with a **linear relationship** between the dependent and independent variables.
- The second plot shows a **nonlinear relationship** between the dependent and independent variables. A Linear Regression model would not generalise well to this data set.
- The third plot shows the **effect of an outlier**. This regression line in this case has been pulled up out of line with the rest of the data set because of this outlier. This can be mitigated by adding some sort of **regularisation** term to the Linear Regression model, similar to the adjusted R-Squared formula (which penalises too many predictor variables).
- The fourth plot shows **how an outlier can cause a relationship to be deduced even though the rest of the points clearly indicate that there is no relationship between the predictor and the target variables**.

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

Ans: Pearson's R measures the correlation between two data sets (assuming they are mean-centered). In linear algebra terms, Pearson's R is nothing but the cosine of the angle between two vectors, if each data set is represented as a mean-centered vector. In this way, it provides an intuitive measure of the similarity between two data sets (vectors). Mathematically:

$R=x^Ty/(!x||y|)=cos(theta)$

Pearson's Correlation Coefficient is also the most common measure of correlation used.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Ans: Scaling is a technique used to bring variables to the same scale.
Scaling reduces or scales the range of the variables to be roughly the same. This results in any analysis done on them (like regression analysis) to give coefficients which can be compared in terms of their effects.

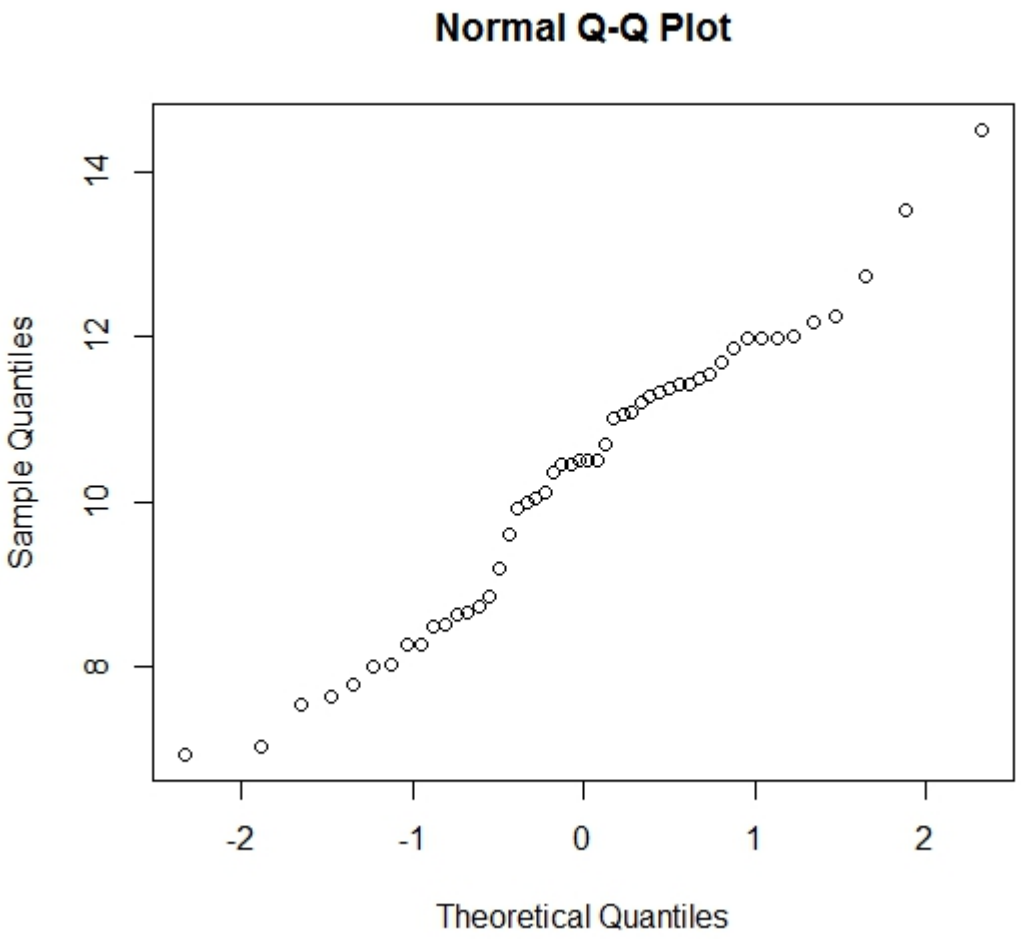
Since regression coefficients depend upon the units of the predictor variables, wildly varying ranges of variables can give coefficients which lead to wrong conclusions about the explanatory power of a particular predictor variable.

Standardised Scaling centers the data around zero, i.e., the mean of the data is zero.
Normalised Scaling "squishes" or expands the range of the data set to exist between 0 and 1, i.e., the highest value in the data set becomes one, and the lowest one, zero.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Ans: The value of VIF can become infinite if two predictor variables are perfectly correlated. Since VIF depends upon R, a perfect correlation implies R=1, leading to VIF=Infinity. This can also happen if all dummy categorical variables are retained, without dropping one of them.

An infinite VIF implies multicollinearity, which means one of the variables need to be dropped.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans: A Q-Q plot (or a Quantile-Quantile plot) is generally used to **compare two distributions**. It can be used to determine if two sets of data arise from the same distribution. For example, if one of the distributions is a normal distribution and we suspect that our data is normally distributed, we can do a Q-Q plot of our data with that normal distribution.



The plot plots all the quantiles (1-100) of each distribution, and puts them on the graph. If the quantiles of both distributions are mostly the same, they should fall close to the 45 degree line.
In Linear Regression, we can use Q-Q plots as a way to determine normality of residuals using the approach above.