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 : There are some categorical variables in the data set – such as yr (year of rental sales), season, weather situation, month of the year , day of the week , holiday or not , working day or not etc..

- There is definite effect of season on rental count, summer and spring seasons seems to have more takes for bike compared to winter months. This correlates a lot with month of the year since seasonal months are failing into those months. During model building we might have dropped season to negate correlation effect with month or other factors and easy to put down month as a factor in the model for planning.
- Year of rental count sales shows there is definite increased adoption year on year, which of course is supported by 2 years data only but health awareness and other socio-economic factors and eco-friendly approach, may have been contributed to the fact, people do tend to use bikes more and more and may continue to do so especially registered users and working class.
- Weather situation has an impact on rental count since rains or snow or storms may result in lesser adoption due to bad weather. Any planning in future may be taken based on forecast of the month and seasonal rains or snow into consideration.
- There is lot of casual rent counts during holidays compared to working days and in general working days have higher total count which supports the fact, there is sizeable target population in the form of registered working or students' class who are may be constant users. We can elaborate further, whether we can target more sales on holidays and casual adopters, while increasing/sustaining the registered working class.

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

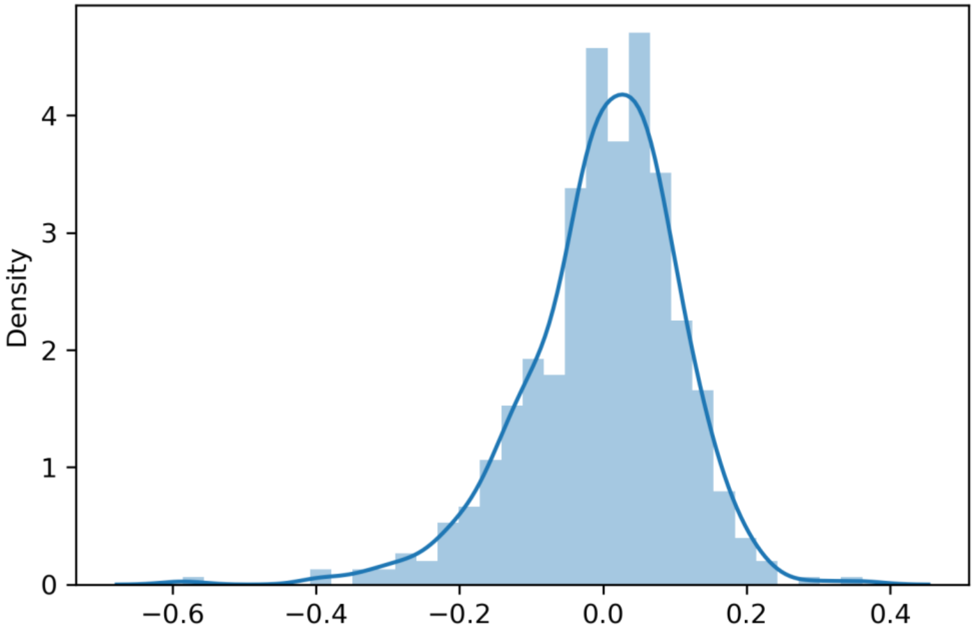
Ans : Number of dummy variables needed to encode the categorical variables and categories can be n-1 or n-2 where n is the count of the different categorical values. we can use drop_first=True to remove 1st of them and use remaining variables to define encoding. it is better to keep the independent variables as minimal as possible as it will reduce correlation also.

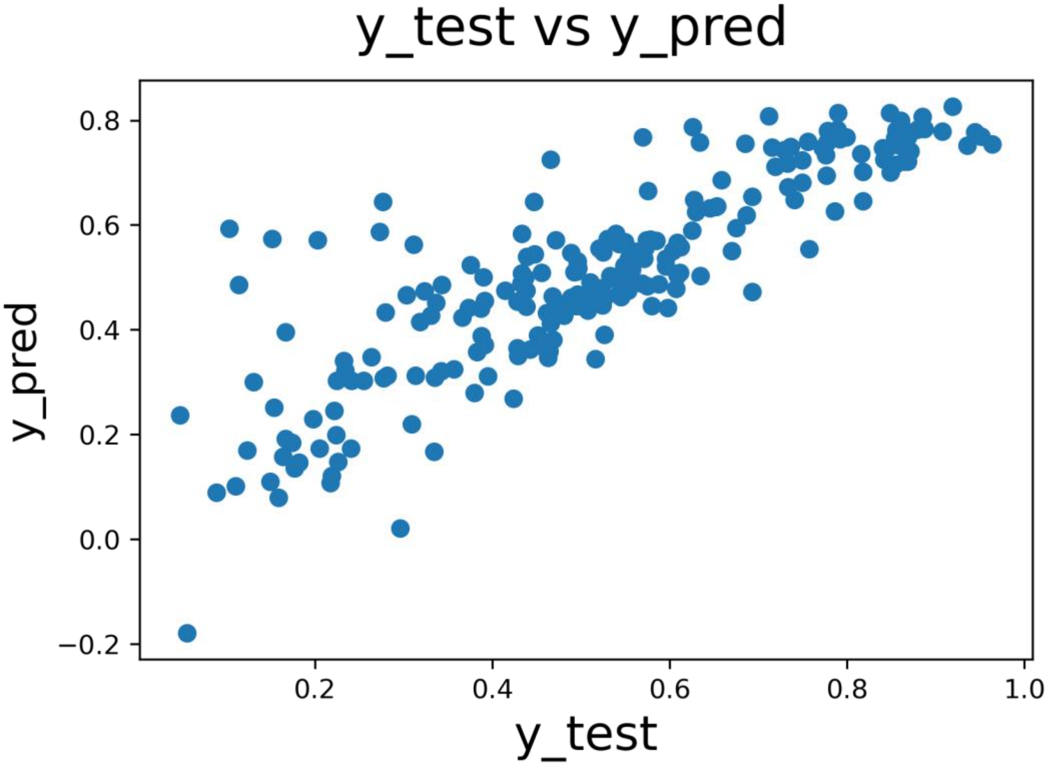
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

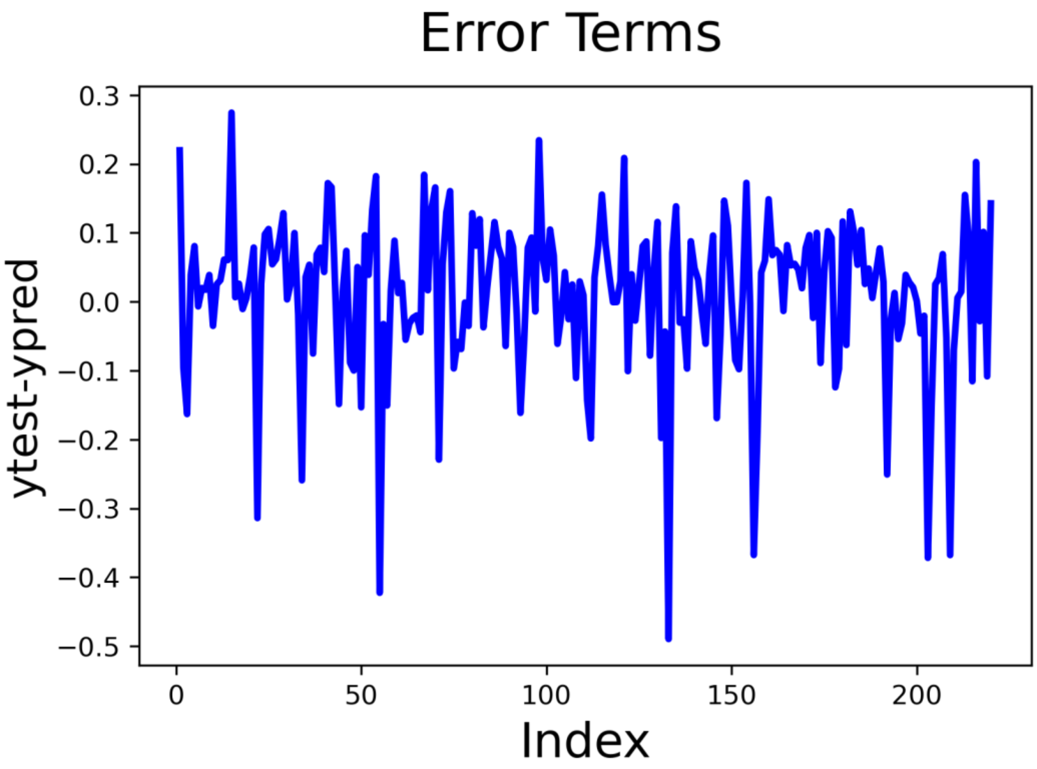
Ans: The temp or atemp variable has highest correlation with the target cnt variable.

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

Ans: After the model is built and final model is selected, we can validate the assumptions of Linear regression by plotting the distplot on the residual error values on prediction and check if follows a normalized curve. Also, we can do a scatter plot on the errors to see if there is no specific pattern.







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

General Subjective Questions

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

   Ans:
   Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.
   Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.

   Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.
   The linear regression model provides a sloped straight line representing the relationship between the variables.

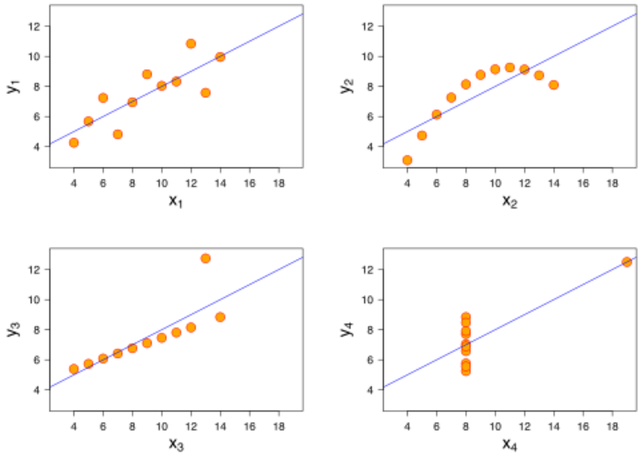
   The best-fit line is found by minimizing the expression of RSS(Residual Sum of Squares) which is equal to sum of squares of the residual for each data point . Residual for any data point is found by subtracting predicted value of dependent variable from actual value of the dependent variable.
   Simple Linear regression algorithm and model is built on a straight line where as Multiple

regression is formulated in a similar way for multiple independent variables.
Minimization of the residual errors can be done in various algorithms , the most prominent one being gradient descent algorithm , as a method of updating the coefficients iteratively until the cost function is minimal or lower tip of the valley.

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

Ans : We need a reminder to us all - graphing data really matters. Graphing data is as important as computation when doing initial data inspection.

Anscombe’s Quartet was devised by the statistician Francis Anscombe to illustrate how important it was to not just rely on statistical measures when analyzing data. To do this he created 4 data sets which would produce nearly identical statistical measures.
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. It was described and being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."



3. What is Pearson’s R? (3 marks)

Ans: In statistics, the Pearson correlation coefficient (PCC) — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient is a measure of the linear coefficient between two sets of data.
Pearson’s r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.
In simple words, Pearson’s correlation coefficient calculates the effect of change in one

variable when the other variable changes.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where n is sample size
xi,yi are individual sample points indexed with i
xbar and ybar are mean of the data respectively.

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

   Ans: Datasets used for ML has multiple features spanning varying degrees of magnitude, range, and units. This is a significant obstacle as a few machine learning algorithms are highly sensitive to these features. For Eg one feature is entirely in kilograms while other in grams, another one is litres and so on. We cannot use these features when they vary vastly hence concept of feature scaling is required. It is a crucial part of the data preprocessing stage before learning model.

   There are 2 types of feature scaling – normalization and standardized feature scaling.

   Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

   Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

   Here Xmax and Xmin are the maximum and minimum values of feature respectively.

   Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

   Here's the formula for standardization:

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

Ans : This happens when the variables have perfect correlation btw them and in case of perfect correlation R2=1 which leads to 1/1-R2*2 as infinite. Best way to solve is to drop those variables which such strong correlation.

6. 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 plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if 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.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

The power of Q-Q plots lies in their ability to summarize any distribution visually.
QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
- Skewness of distribution