SUBJECTIVE QUESTIONS AND ANSWERS

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)
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

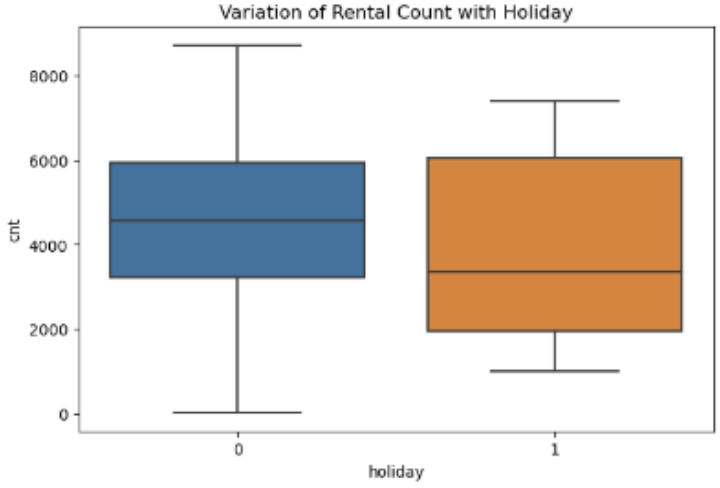
General Subjective Questions

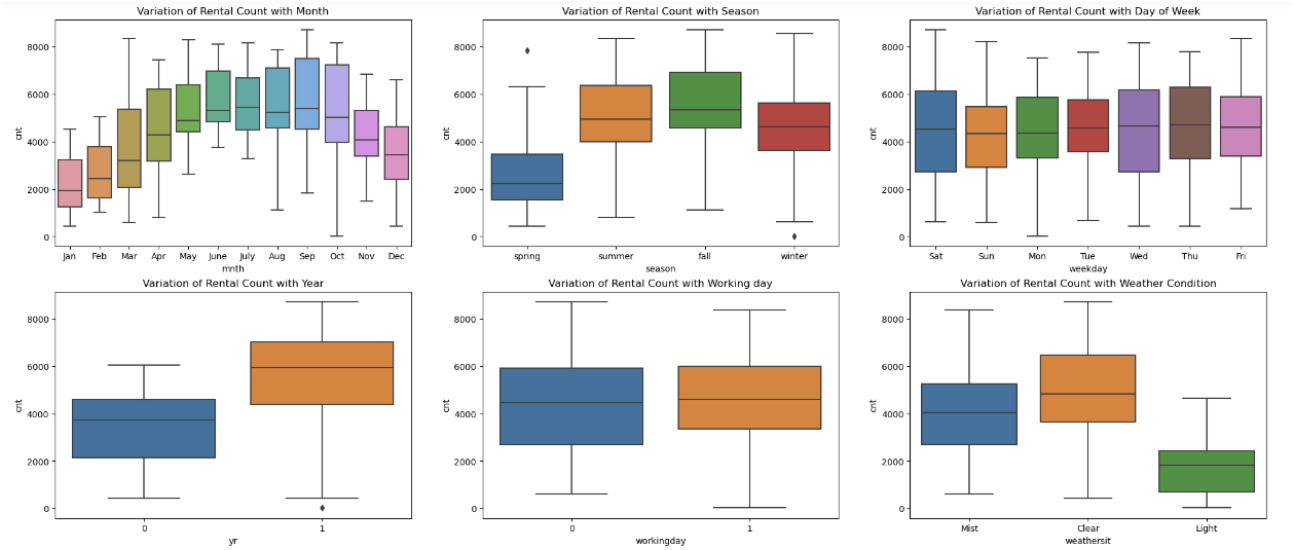
1. Explain the linear regression algorithm in detail. (4 marks)
2. Explain the Anscombe's quartet in detail. (3 marks)
3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

ANSWERS:

Assignment-based subjective Questions

1. Based on our analysis of Categorical variables we have generated the following plots with respect to **'cnt'** as our target variable.





According to these plots, Median no of rental counts are significantly higher during

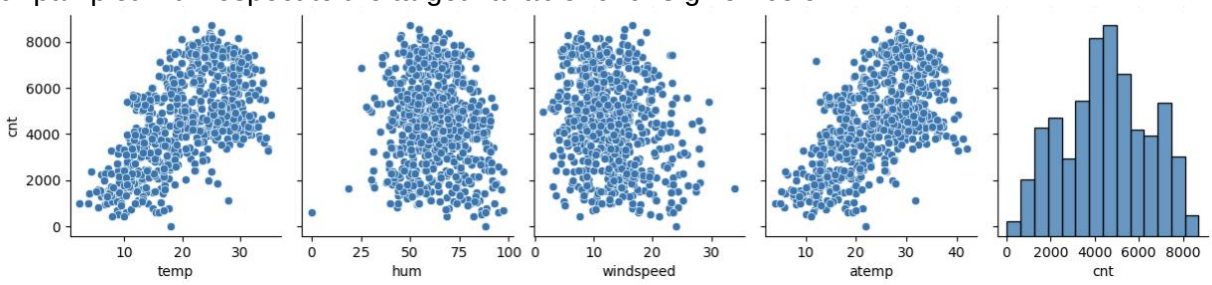
1. 2019 as compared to 2018
2. Regular days as compared to holidays, although the variance on holidays is higher
3. Fall as compared to other seasons
4. Clear, Few clouds, Partly cloudy, Partly cloudy days as compared to other weather conditions
5. July and September, in comparison to all other months, with September having a relatively higher variance in rental counts

2. During dummy variable creation from a categorical column having n levels, **drop_first**=True enables removal of the first One Hot Encoded (OHE) column, from the n different OHE columns created from the categorical column. Not doing so, leads to the Dummy Variable Trap in which variables are highly correlated to each other leading to multicollinearity. After dropping, we will thus be left with n-1 OHE columns, that can perfectly help in explaining all the n levels. The syntax in Pandas will be:

   pd.get_dummies(df, drop_first=True)

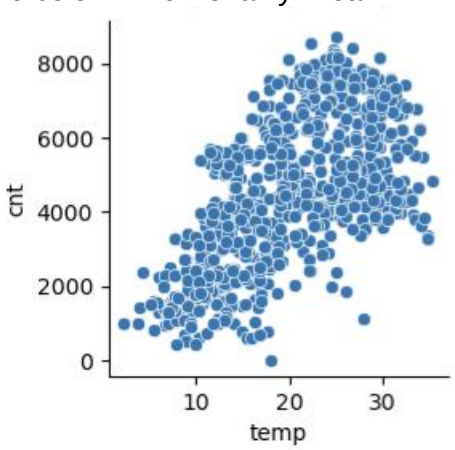
   E.g., If we have a categorical variable **'opinion'** that has 3 levels **'Yes', 'No'** and **'Maybe'**, then it will suffice having only 2 OHE columns **'Yes'** and **'No'**. When both **'Yes'** and **'No'** are 0, it can represent the **'Maybe'** status

3. Our pair plot with respect to the target variable ***'cnt'*** is given below

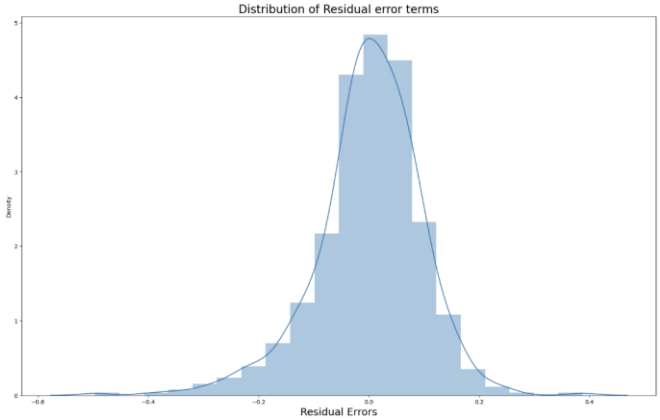


According to the plot, **'cnt'** has the highest correlation with **'temp'** and **'atemp'**.

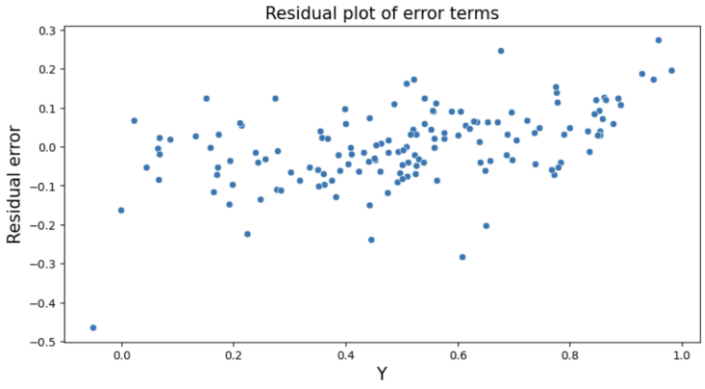
4. We have built the Linear regression model based on the following assumptions:
    a. The relationship between the independent and dependent variables are linear. For instance, ‘cnt’ and ‘temp’ are plotted along a scatter plot and looks like below which is fairly linear



    b. The residuals are normally distributed about 0, as evident in the Dist plot below



    c. Homoscedasticity signifies that the variance of the residual error terms are fairly constant across the zero error line, especially when plotted across the y-values. Barring a few outliers, we see a fairly constant variance of the residuals about the zero error



d. Little or no multicollinearity between the predictors is corroborated by the VIF table below where each of the final list of predictors has a VIF below 5, thereby indicating a low degree of multicollinearity

	Features	VIF
1	temp	4.86
2	windspeed	4.80
4	winter	2.33
0	yr	1.98
3	spring	1.85
7	Nov	1.69
10	Mist	1.53
6	July	1.42
5	Dec	1.37
8	Sun	1.16
9	Light	1.07

5. The demand variable 'cnt' or count has been modelled as per the below equation

$$count = 0.2573 - (0.1215 \times windspeed) + (0.4353 \times temp) + (0.2476 \times yr) - (0.1290 \times spring) - (0.0414 \times Sun) - (0.2798 \times Light) + (0.0774 \times winter) - (0.0781 \times Nov) - (0.0851 \times Mist) - (0.0853 \times July) - (0.0639 \times Dec)$$

As per the equation, the top 3 variables contributing significantly towards explaining the demand are:

a. 'temp' – Temperature of the day. It increases with an increase in temperature
b. 'yr'- The year of booking. Demands are higher on 2019 than 2018
c. 'Light' – Demands tend to fall with the onset of weather conditions such as Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

General Subjective Questions

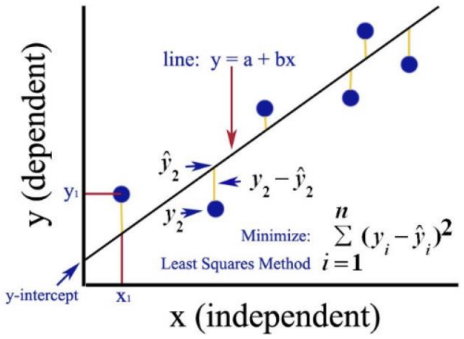
1. Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the outcome or response variable) and one or more independent variables (also known as predictor or explanatory variables). The goal of linear regression is to find the best-fitting line/hyperplane through the data points, which can be used to predict the value of the dependent variable based on the values of the independent variables.

The algorithm for linear regression is as follows:

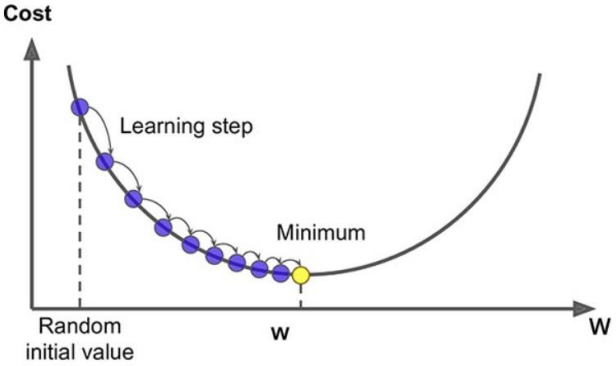
1. Collect a dataset consisting of (x, y) pairs, where x represents the independent variable(s) and y represents the dependent variable

2. The first step is to choose the best-fitting line that describes the relationship between x and y. The equation of the line is represented by:
   $y = \beta_0 + \beta_1.x$ (for Simple Linear Regression)
   $y = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + \beta_3.x_3 + \ldots + \beta_n.x_n$ (for Multiple Linear Regression)
   where β_i is the slope/gradient of the line/hyperplane and β_0 is the y-intercept

3. The next step is to determine the values of β_0 and β_i that minimize the sum of the squared differences between the predicted values of y (based on the equation of the line/hyperplane) and the actual values of y. This is done using a technique called "Ordinary Least Squares" (OLS). OLS is a method to find the parameters that minimize the sum of the squared residuals. Below is an example of a fitted line with minimal OLS



4. Parameter tuning can be done using a variety of methods. One such method is the Gradient Descent Method as shown below.



5. Once the best-fitting line/hyperplane is determined, it can be used to make predictions about the value of the dependent variable based on the value of the independent variable(s).
6. The coefficients of the model are then evaluated with some goodness of fit measures such as R-squared, Mean squared error, and adjusted R-squared. A very high value of any of these metrics for e.g., R-squared of 0.99 may signify overfitting which can give a bad performance on the test dataset

It's important to notice that linear regression assumes the following

a. The relationship between the independent and dependent variables is linear

b. The residuals are normally distributed about 0

c. The distribution of residuals is homoscedastic, which means that they have a distribution of almost constant variance about the zero error line

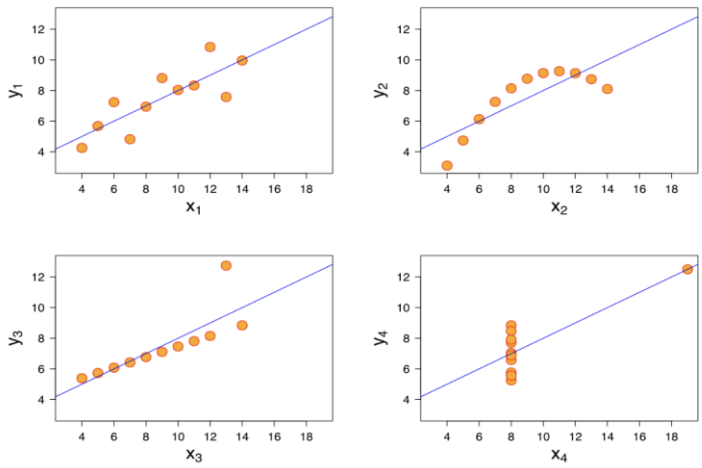
d. No multicollinearity between the predictors. VIF and p-values of predictors from model summary can be used here for the same

2. Anscombe's quartet is a set of four datasets, each consisting of 11 (x, y) points, that have the same summary statistics (mean, variance, and correlation) but have vastly different looks when plotted. The quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analysing it.

The four datasets are:

- Dataset I: A linear relationship between x and y, with a correlation coefficient of 0.816, fitted with a regression line of equation $y = 3.00 + 0.500x$
- Dataset II: A quadratic relationship between x and y, with a correlation coefficient of 0.816, fitted with a regression line of equation $y = 3.00 + 0.500x$
- Dataset III: A non-linear relationship between x and y, with a correlation coefficient of 0.816, fitted with a regression line of equation $y = 3.00 + 0.500x$
- Dataset IV: No visual correlation between x and y, yet with a correlation coefficient of 0.816 and fitted with a regression line of equation $y = 3.00 + 0.500x$

The quartet is given below



The purpose of Anscombe's quartet is to demonstrate that summary statistics, such as the mean and correlation coefficient, can be misleading if not accompanied by a visual representation of the data.

Anscombe's quartet is a classic example of the importance of data visualization and the need to look beyond summary statistics when analysing data. It shows that data can be very misleading if we only rely on summary statistics and not visualize the data.

3. Pearson's R, which is also known as Pearson's Correlation Coefficient is a quantity that signifies the correlation between two variables x and y, based on how they vary together.

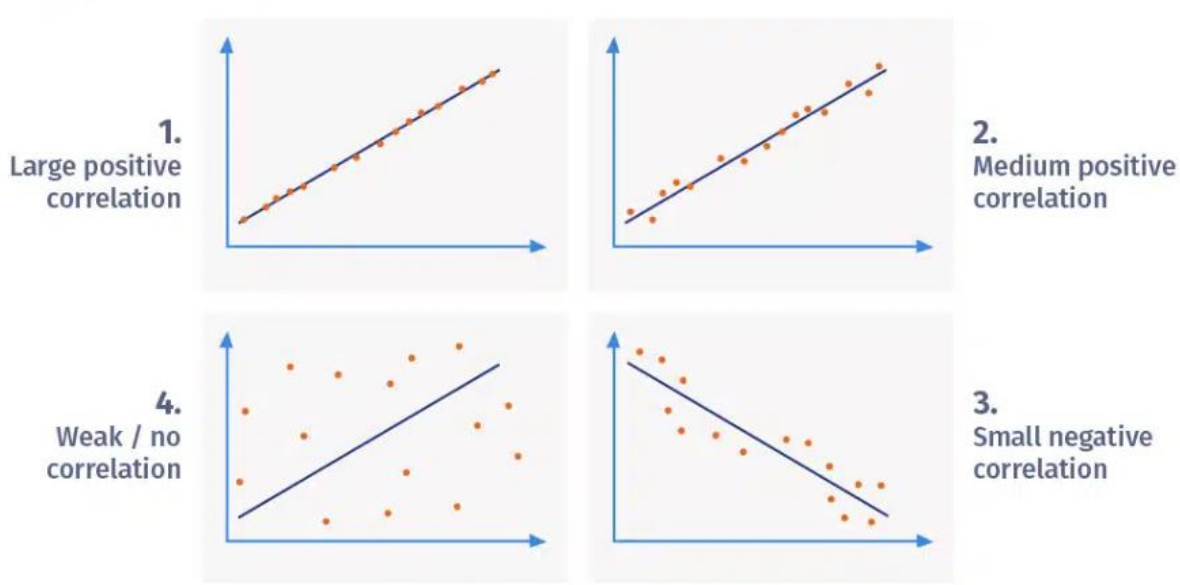
- $(-1 \leq r \leq +1)$ It lies between -1(perfect negative correlation) and +1(perfect positive correlation)
- Positive correlation signifies that if one variable X increases, the other variable Y also increases in value
- Negative correlation signifies that if one variable X increases, the other variable Y decreases in value
- The closer the r value is to 0, the weaker the correlation
- An r value of zero signifies no correlation at all

Thus, it is given by the following formula.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

The effect of the coefficient can be understood by the following diagrams, where X values are plotted across the x-axis and Y values across the y-axis



4. Scaling in Linear Regression is the method by which a numerical variable existing within a particular numerical range (typically higher) is transformed such that it now exists in a standardized/normalized numerical range as all other numerical variables

The predictor variables may have different units (e.g., seconds, meters and hours) or magnitudes (in the range of hundreds, thousands or more) that, in turn, may mean the variables have different scales. If scaling is not done then the algorithm makes the model learn coefficients with greater weight values, hence causing greater bias and leading to lower performance. Scaling helps in quicker fitting and smoother learning of models; otherwise, the learning speed is impacted for larger coefficients.

Normalization or Min-Max Scaling is used to transform X feature to a scaled X_new feature by subtracting the minimum value of X and dividing by the difference between the maximum and the minimum value of X. It is based on the following transformation.

```
X_new = (X - X_min)/(X_max - X_min)
```

Standardization is used to transform X feature to a scaled X_new feature, by subtracting from mean and dividing by standard deviation of the feature values. It is based on the following transformation.

```
X_new = (X - mean)/Std
```

5. VIF or Variance Inflation Factor is a significant metric to measure multicollinearity among predictors and thereby helps us in feature selection. It is described by the following formula

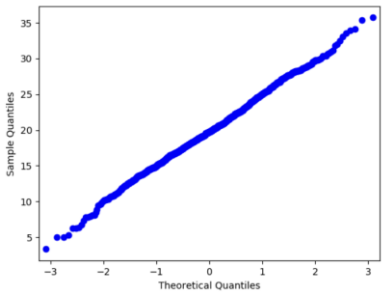
$$VIF_i = \frac{1}{(1-R_i^2)}$$

Therefore, if VIF is infinite, it signifies that the R-squared value of the particular variable is 1. This signifies that this predictor can be perfectly explained by a linear combination of all other predictor variables present in the model. For better models, such a variable should be immediately dropped.

6. A Q-Q plot or a Quantile-Quantile plot is a plot where quantile values of one distribution are plotted against the quantile values of another distribution. (A quantile of percentage q is a measure of the value within which q% of the data resides in a distribution). If the values come from the same normal distribution, then the plotted points will fall on a straight line of 45 degrees, known as the reference line.

Linear regression assumes a normal distribution of residual terms. Therefore, in a Linear Regression, a Q-Q plot can be used in the following:

a. Checking whether the residuals are normally distributed or not. Greater the plot fit along a straight line of 45 degrees, the more normally distributed it is. For e.g.,



b. Checking whether the distribution of the actual and predicted target variables come from the same distribution or not. Greater the plot fit along a straight line of 45 degrees, the more similar they are. For e.g.,