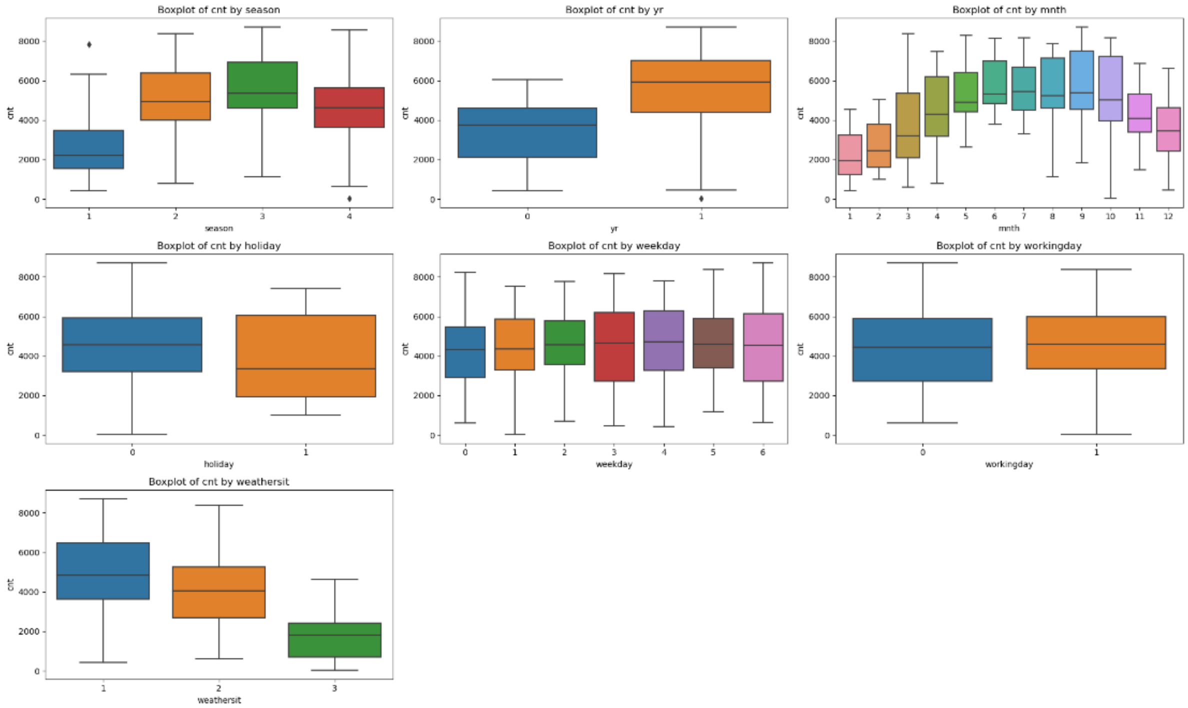
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



a) Season 3 has the highest impact on cnt
   a. On the same lines, month 6-9 sees more bookings
b) cnt was higher in 2019 as compared to 2018 on all the parameters (Median, max, percentile etc)
c) cnt is more on holidays
d) Clear weather has a significant positive impact on cnt
e) Working, non-working days doesn’t have a major impact

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

Using drop_first=True during dummy variable creation is important for several reasons:

a) Avoiding the Dummy Variable Trap:
   - It prevents perfect multicollinearity, which occurs when the dummy variables are perfectly correlated. This happens because the sum of all dummy variables is always equal to 1 for each observation.
   - Dropping one dummy variable leaves k−1 dummy variables for k categories, resolving the multicollinearity issue.
b) Simplifying the Model:
   - Dropping the first dummy variable simplifies the model without losing any information. The dropped category serves as the reference group.
c) Interpretability of Coefficients:
   - The coefficients of the dummy variables indicate the effect of each category relative to the reference category (the dropped category).
d) Model Efficiency:
   - Reducing the number of dummy variables makes the model more efficient to run and easier to interpret.

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

temp has the highest correlation with the target variable

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

Below is how I validated the assumptions:

S. No.	Assumption	Method of Validation
1	Normality of Error terms	Normal distribution curve, Q-Q plot
2	Multicollinearity Check	Checked through VIF. All features have a VIF < 10
3	Linear relationship validation	We can see Linearity in the visualizations
4	Homoscedasticity	No visible pattern observed from above plot for residuals.
5	No auto-correlation	Durbin-Watson value of final model lr_4 is 2.032, which signifies there is no autocorrelation.

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

Here's the final model:

cnt = 0.0856 + 0.2331 x yr + 0.0555 x workingday + 0.5116 x temp - 0.1542 x windspeed + 0.1015 x season_2 + 0.1260 x season_4 + 0.0546 x mnth_8 + 0.1172 x mnth_9 + 0.0384 x mnth_10 + 0.0662 x weekday_6 - 0.0849 x weathersit_2 - 0.2932 x weathersit_3

Top 3 features are

a) temp: 0.5116
b) weathersit_3: -0.2932
c) yr: 0.2331

General Subjective Questions

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

Linear regression is a method used to understand the relationship between one outcome (dependent variable) and one or more predictors (independent variables). It finds the best straight-line relationship between variables, helping us make predictions and understand how changes in predictors affect the outcome. Here's a simple breakdown:

1. Model Equation

- The relationship is represented by a straight-line equation: outcome = intercept + (coefficient1 * predictor1) + (coefficient2 * predictor2) + ... + (coefficientN * predictorN).

2. Goal

- The goal is to find the best values for the intercept and coefficients that make the predicted outcomes as close as possible to the actual outcomes.

3. Best Fit Line

- We find the best fit line by minimizing the differences between the actual and predicted outcomes. This is done using a method called Ordinary Least Squares (OLS).

4. Making Predictions
   - Once we have the best coefficients, we can use the equation to predict outcomes for new predictor values.

5. Assumptions
   - Linear regression works best when:
     - The relationship between variables is linear.
     - The differences between actual and predicted outcomes (residuals) are normally distributed.
     - The residuals have constant variance.

6. Evaluating the Model
   - We check how well the model works using:
     - R-squared: Tells us how much of the variation in the outcome is explained by the predictors.
     - Mean Squared Error (MSE): The average squared difference between actual and predicted outcomes.
     - Mean Absolute Error (MAE): The average absolute difference between actual and predicted outcomes.

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

Anscombe's quartet comprises four datasets that appear to have nearly identical simple descriptive statistics and yet have very different distributions when graphed. It is created by the statistician Francis Anscombe in 1973 with the purpose of showing the importance of graphing data prior to analysing it and the limitations of relying on summary statistics.

Each data set in Anscombe's quartet consists of 11 (x, y) pairs. When considered individually, each data set seems to have similar information pertaining to mean, variance, correlation coefficient, and linear regression line. However, when the data is plotted, the different patterns become very apparent.

The four datasets generally take the following form:

1. Dataset I: This data set represents a simple linear relation, where members of this set appear to cluster around a straight line.
2. Dataset II: It is a linear relation but has one outlier that seriously affects the correlation coefficient and linear regression line.
3. Dataset III: This dataset represents a non-linear relation, and a single outlier drastically influences the linear regression line.
4. Dataset IV: This dataset comprises two groups of data, for which summary statistics may suggest a relationship. However, plotting the data provides no such pattern.

Anscombe's quartet illustrates the importance of graphic visualization in understanding structure within data, as well as a caution against using summary statistics for datasets of different patterns, which can be identical. This illustrates the requirement of exploratory data analysis and graphical methods for a full understanding of data.

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

Pearson's correlation coefficient, often denoted by R, measures the strength and direction of a linear relationship between two variables. It indicates the relationship between two continuous variables.

The value of Pearson's R has a range from -1 to 1:

1. R = 1: It will be a situation of perfect positive linear correlation, as one variable increases, the other also increases proportionally.
2. R = -1: Perfect negative linear correlation, meaning that as one variable increases, the other decreases proportionally.
3. R = 0: This implies no linear correlation; thus, there is no systematic relationship between the variables.

The Pearson correlation coefficient is sensitive to outliers and assumes that the relationship between the variables is linear. It is a widely used test in different disciplines: statistics, economics, psychology, and social science in general, to test the strength and direction of relationships between variables.

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

Scaling is the process of changing data to a common scale without distortion in the differences in value ranges. It is done so that variables with different scales (like cnt, temp, humidity etc) contribute equally toward analysis and modelling.

Normalized scaling rescales the data to lie in the range between 0 and 1; it maintains relative relationships of data but not the original distribution. Standardized scaling describes the transformation that data is put through to have a mean of 0 and a standard deviation of 1 while maintaining its shape.

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

Yes, I found many features with infinite VIF when I first created the model without RFE. I had to remove it as somehow that code was messing up with the rest of the code. VIF is infinite when there is perfect multicollinearity among predictor variables, i.e., one or more independent variables can be perfectly predicted from others.

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

A Q-Q (Quantile-Quantile) plot is a graphical method that compares a dataset to a theoretical distribution, usually the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution.

The principal check for this regression pertains to the normality of residuals. The Q-Q plot works by plotting the quantiles of residuals of the model against the corresponding quantiles that the observed residuals would be expected to have if they came from a normally distributed population. One can thereby visually check if the residuals are normally distributed. This is important because linear regression assumes the residuals to be normally distributed, and any violation would thereafter compromise the regression analysis.