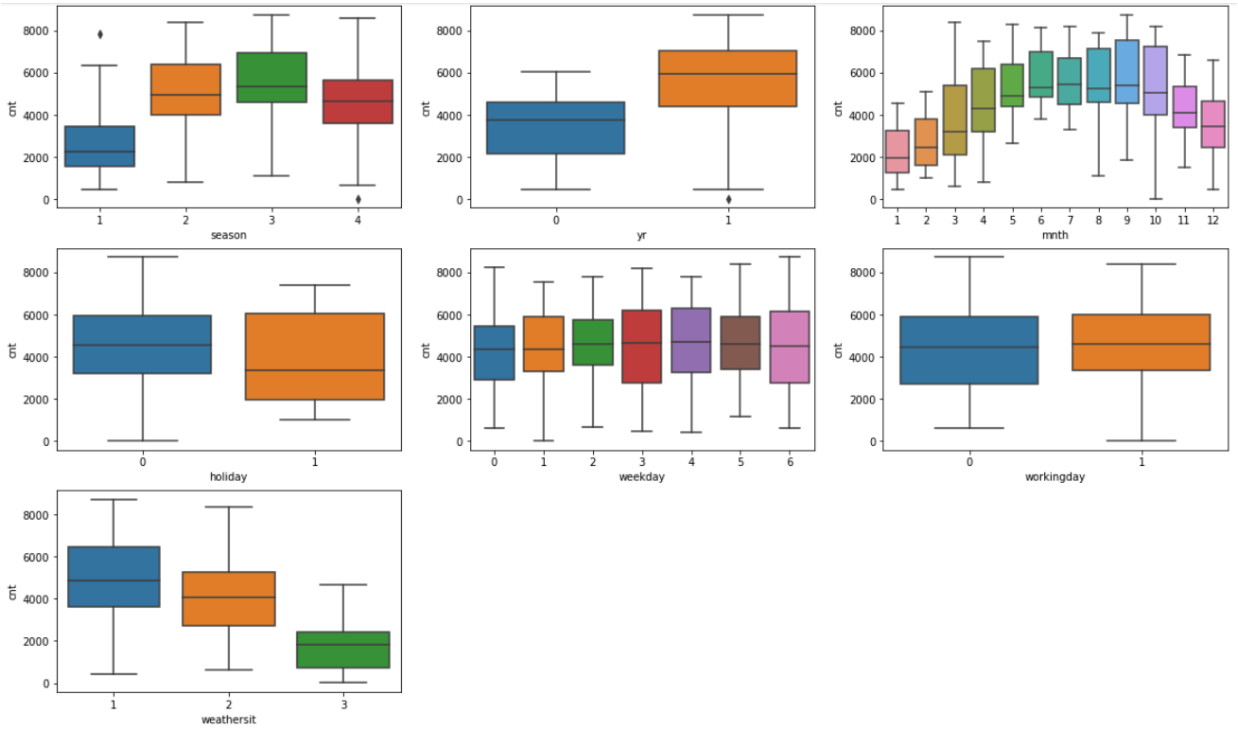
Assignment-based Subjective Questions

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

List of categorical variables in the dataset are: season, yr, mnth, holiday, weekday, workingday, weathersit.

1) Seasons seem to have a significant impact on the count of rides booked.
    a) In Season 3 - Fall, the count of bike ride is significantly higher than the seasons 1 & 2 (spring & winter)
    b) whereas, the 75th quantile is comparatively higher than season 2 (summer)
2) Year has significant impact as the number of rides in 2019 is very high (as shown by the median line in the below diagram)
3) Workingday is a very critical variable which explains a lot of data variability. From the visualization exercise it was not quite clear but it does show that the lower hinge gets shifted up on a working day, meaning more demand. RFE clearly shows that it is an impactful variable to consider.
4) Similarly with holiday as a variable is very much similar to working day and provides opposite but similar information, thus can be ignored.
5) By visualization exercise, the median for each class of weekdays remains the same but the distribution varies. Later, in the modeling exercise it became clear that it is not an impactful variable to capture demand.
6) 'Mnth' variable clearly shows a trend and strong correlation with demand. Though, in modeling exercise this came out to be a less impactful variable as the majority of the predictive power provided by this variable is easily captured by season.
7) Weathersit has strong correlation with demand both shown by visualization below and the RFE rank



5	season_3
10	registered_ratio
2	workingday
3	temp
4	season_2
6	season_4
9	weekday_6
8	weathersit_3
7	weathersit_2
1	yr

---

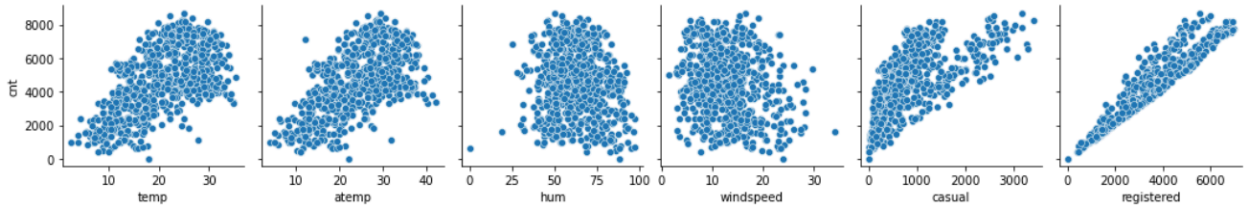
2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

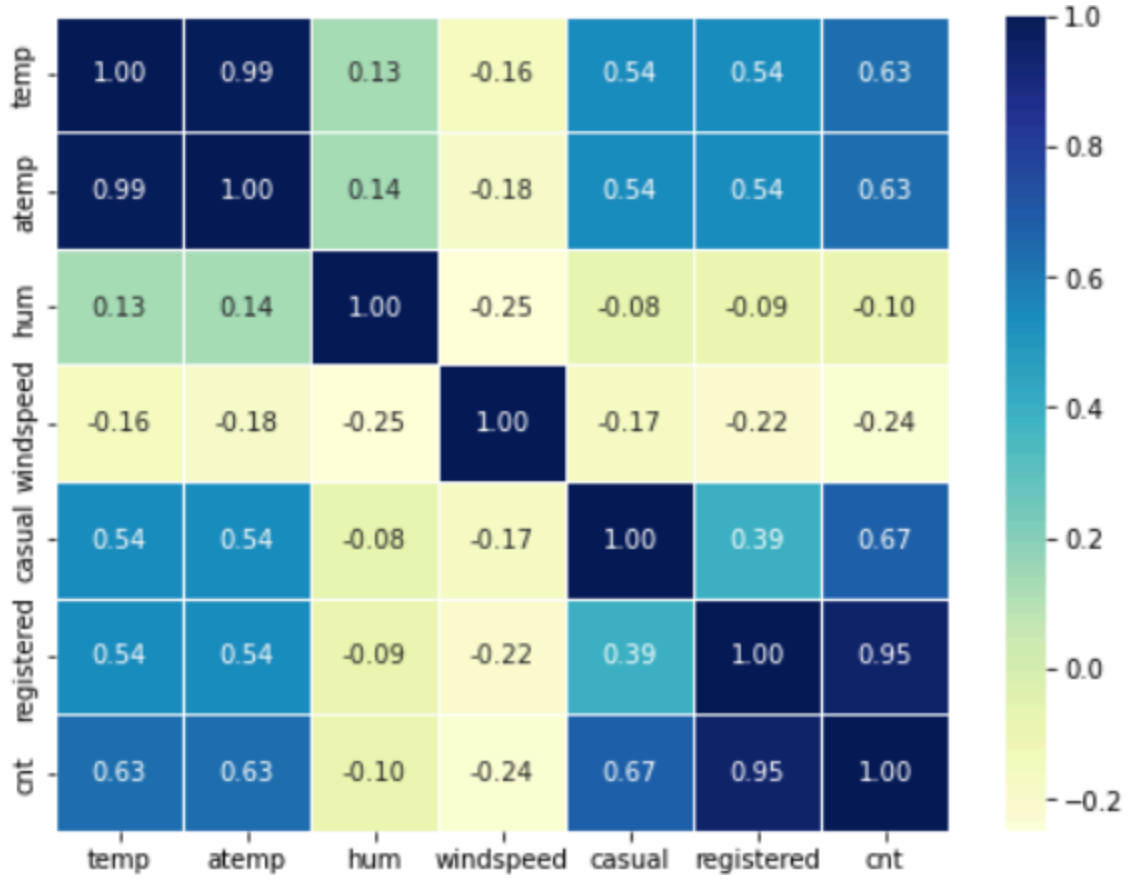
For a categorical variable with k levels, we need only k-1 dummy variables. Thus, to eliminate the redundancies in the modeling process one column from k dummy variables needs to be removed. Thus, we opt for the condition drop_first = True.

---

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

'Registered' has the highest correlation with the 'cnt' target variable.

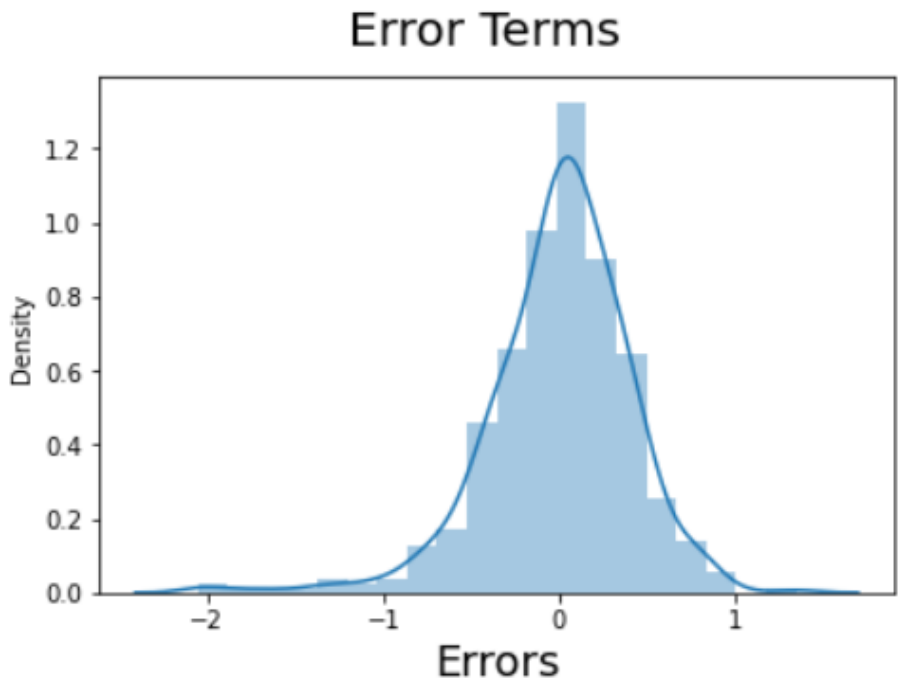




---

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

1) Validated that at least few of the independent variables has a linear relationship with the target variable and shows some correlation.
2) Checked the distribution of the error terms as they should not show any pattern, thus should be independent of each other
3) The error terms should follow a normal distribution and has constant variance.



---

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

1) Season_3 - Fall
2) Registered ratio - registered / cnt - share of bookings from registered users
3) Working day
4) temp

General Subjective Questions

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

Linear regression is a fundamental supervised learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). It is widely used for prediction and forecasting tasks in various fields such as statistics, economics, finance, and machine learning.
The goal is to find the best-fitting linear model that describes how the target variable varies as the input variables change.

Assumptions of Linear Regression:
Linear regression makes several assumptions about the data and model:

1. **Linearity:** Assumes a linear relationship between the independent and dependent variables.

2. **Independence:** Assumes that the residuals are independent of each other.
3. **Homoscedasticity:** Assumes constant variance of residuals across all levels of predictors (homogeneity of variance).
4. **Normality:** Assumes that the residuals follow a normal distribution.
5. **No Multicollinearity**: Assumes that independent variables are not highly correlated with each other.

Linear Regression models are of two types:

1. **Simple Linear Regression:** In simple linear regression, there is only one independent variable (feature) that is used to predict the dependent variable (target). The relationship between the feature (X) and the target (Y) is modeled using a linear equation of the form:

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon$$

2. **Multiple Linear Regression:** In multiple linear regression, there are multiple independent variables (features) used to predict the dependent variable (target). The relationship is modeled using a linear equation of the form:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n + \epsilon$$

---

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

Anscombe's quartet is a compelling demonstration of how diverse datasets with different distributions & characteristics can have similar statistical properties, highlighting the limitations of summary statistics and the necessity of visual exploratory analysis in data science and statistical modeling. The quartet is often used in statistics education to teach concepts such as descriptive statistics, regression analysis, outliers, and the importance of data visualization. It serves as a cautionary example against over-reliance on summary metrics and the importance of critically examining data through multiple perspectives.

---

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

Pearson's correlation coefficient, often denoted as r or Pearson's r, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is named after Karl Pearson, a prominent mathematician and statistician.

Assumptions: Pearson's correlation assumes that the relationship between variables is linear, both variables are approximately normally distributed, and there is homoscedasticity (constant variance of residuals).

Pearson's correlation coefficient can take values between -1 and 1, where:

- r=1 indicates a perfect positive linear relationship (as one variable increases, the other variable also increases proportionally).

- r=−1 indicates a perfect negative linear relationship (as one variable increases, the other variable decreases proportionally).
- r=0 indicates no linear relationship (the variables are not linearly related).

---

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

Scaling is a preprocessing technique used in data analysis and machine learning to standardize the range of independent variables or features of a dataset. The goal of scaling is to transform the data so that all variables are on a similar scale, which helps in improving the performance of certain algorithms and models.

Why Scaling is Performed:

1. Algorithm Performance: Some machine learning algorithms, such as gradient descent-based algorithms (e.g., linear regression, logistic regression, neural networks), support vector machines (SVMs), and k-nearest neighbors (KNN), are sensitive to the scale of features. Scaling helps these algorithms converge faster and perform better.
2. Feature Interpretation: Scaling ensures that features contribute equally to the analysis, preventing features with larger scales from dominating the learning process or model outcomes.
3. Distance-based Algorithms: Algorithms that rely on distance metrics (e.g., KNN, clustering algorithms) can be influenced by the scale of features. Scaling ensures that distances are calculated accurately and fairly across features.

Types of Scaling:

1. **Normalized Scaling (Min-Max Scaling):**
   Formula: Xnorm = (X - Xmin) / (Xmax - Xmin)
   Normalized scaling (also known as Min-Max scaling) rescales the feature values to a fixed range, typically between 0 and 1.
   This scaling method preserves the relative relationships between the original data points.
   It is sensitive to outliers, as outliers can affect the range of the normalized data.
2. **Standardized Scaling (Z-score Scaling):**
   Formula: $X_{std} = (X-\mu)/\sigma$
   - Standardized scaling (Z-score scaling) transforms the data to have a mean (μ) of 0 and a standard deviation (σ) of 1.
   - It centers the data around zero and scales it based on the variance.
   - Standardized scaling does not bound the data to a specific range, and it can handle outliers better than normalized scaling.
   - It preserves the shape of the original distribution but does not necessarily preserve the exact relationships between data points.

Differences between Normalized Scaling and Standardized Scaling:

1. Range: Normalized scaling bounds the data to a specific range (e.g., 0 to 1), while standardized scaling does not impose a specific range.
2. Mean and Standard Deviation: Standardized scaling centers the data around a mean of 0 and a standard deviation of 1, while normalized scaling does not alter the mean and standard deviation.
3. Handling Outliers: Standardized scaling is more robust to outliers compared to normalized scaling because it uses the standard deviation for scaling instead of the range.

---

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

Formula for VIF = 1 / (1-R^2)
For VIF to be infinite, R-squared needs to be 1.
If R-squared for any linear model is 1, this indicates that the predictor variable may be highly correlated with other predictor variables in the model, leading to unstable and unreliable coefficient estimates.

---

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 technique used to assess whether a given dataset follows a particular distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution, typically a normal distribution.
The Q-Q plot is a powerful tool in statistical analysis and is particularly useful for assessing the assumptions of linear regression models, such as the normality of residuals.

Use and Importance of Q-Q Plot in Linear Regression:

1. Assessing Normality of Residuals:
    - In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed.
    - A Q-Q plot of the residuals can visually show whether the residuals follow a normal distribution. If the points in the Q-Q plot roughly follow a straight line, it indicates that the residuals are approximately normally distributed. Deviations from the line suggest departures from normality.
2. Identifying Outliers and Skewness:
    - Apart from normality, a Q-Q plot can also help identify outliers and skewness in the distribution of residuals.
    - Outliers or skewness may manifest as points that deviate significantly from the expected linear pattern in the Q-Q plot, indicating potential issues with the model assumptions or data quality
3. Model Validation and Assumption Checking:

- Q-Q plots are part of the model validation and diagnostic process in linear regression.
- By examining the Q-Q plot, analysts can gain insights into whether the linear regression model is appropriate for the data and whether the assumptions of the model, such as normality of residuals, are met.

4. Interpretation of Model Results:
    - A Q-Q plot can influence the interpretation of regression results. If the Q-Q plot shows significant deviations from normality, it may warrant further investigation or consideration of alternative modeling approaches.
    - Normality of residuals is important for valid statistical inference, such as hypothesis testing and confidence interval estimation, in linear regression.