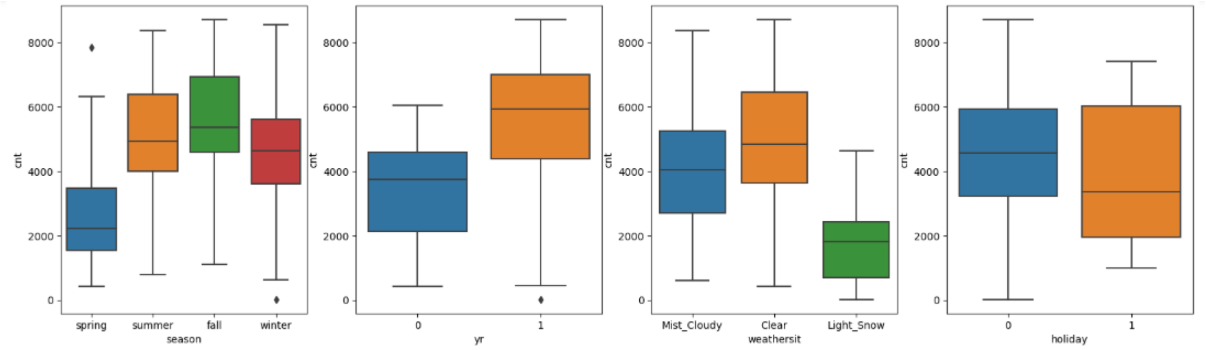
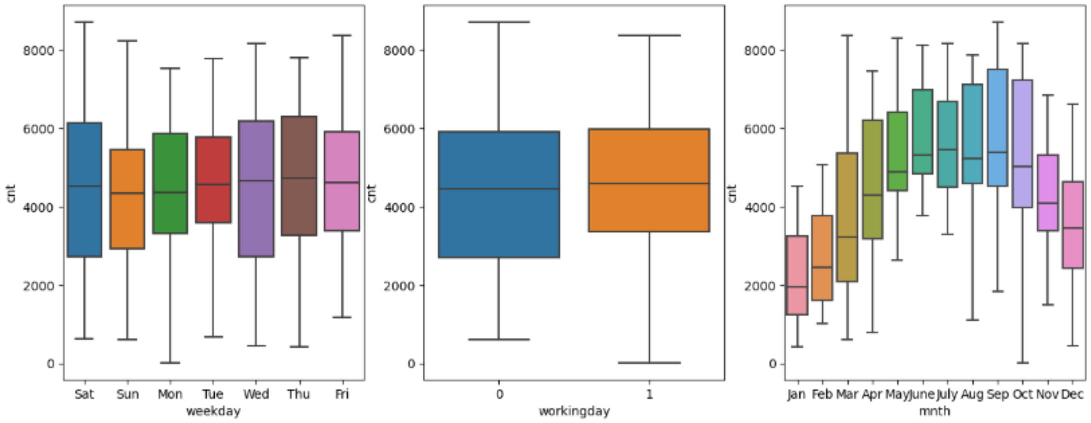
Assignment-based Subjective Questions

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





Observations

- Highest number of bike rentals are seen in Fall season followed by Summer season and winter. Spring sees lesser number of bike rentals in comparison.
- Number of bike rentals started picking up in 2019 compared to that of 2018
- Highest number of bike rentals are seen when the weather situation is clear followed by Misty_Cloudy weather situation. There is also bike rentals happening Light_Snow or Light_Rain but no bike rentals happening during heavy rain, snow + fog.
- Bike rentals are seen more during May, June, July, August, September, October months

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

It is important to use drop_first=True during dummy variables to avoid multi-collinearity. If we include all dummy variables, it introduces multi-collinearity because the sum of dummy variables for each observation will always be 1.

It also simplifies the model by reducing one predictive variable while making sure that the information is still captured and interpreted with the remaining dummy variables.

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

The temp (Temperature) variable has the highest correlation with the bike rental count target variable.

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

The assumptions of the linear regression were validated after building the model on the training set are as follows:

- P-value – Checked if the correlation between the variables to the target variable is significant. P-value should be less than 0.05
- VIF – Calculated the VIF of the variables in the model and dropped some of the variables with high VIF to eventually get VIF of all the variables under 5
- Error distribution of residuals – Verified that the error distribution of residuals is following a normal distribution.

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

Top 3 feature that are significantly contributing towards explaining the demand of the shared bikes are as follows:

- Temperature (positively contributing) – coefficient value of 0.3593
- Weather situation of (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) is a negatively contributing factor – coefficient value of -0.2975
- Year is 2019 (1) or 2018 (0) (positive contributing) – coefficient value of 0.2360

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used for predicting a (target) dependent variable based on the given independent variables (predictors). This regression technique tries to establish a linear relationship between target variable and other predictor variables.

There are two types of linear regression:

- Simple Linear Regression – Single independent variable or predictor is used to predict a target variable.

- Multiple Linear Regression - Multiple independent variable or predictors are used to predict a target variable.

A linear line (best fit line) showing relationship between predictors and target variable is called a regression line. A best fit line is calculated using R squared value which is based on residual sum of squares (RSS) and total sum of squares (TSS).

R squared = 1 – (RSS/TSS)

A perfect R square value means overfitting due to multi-collinearity in a regression model. Variance Inflation Factor (VIF) is used to identify multi-collinearity and p-value is used for calculating the significance of the predictors to further tune the linear regression model once we have the R squared values.

The linear regression model is the further validated by checking if error residuals are normally distributed. The linear regression model is then tested against the test data to see if the R squared value is coming inside 5% variance for making sure that the model is working as expected.

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

Anscombe's quartet consists of 4 datasets that are having seemingly identical descriptive statistics but have very different graphical distribution once we visualize the same graphically underlining the importance of graphical visualization. Each dataset in the quartet has distinct characteristics highlighting the importance of looking at a set of data graphically before starting the data analysis and the inadequacy of basic statistic properties for describing a dataset.

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

Pearson's R or Pearson's Correlation Coefficient is used to establish linear relationship between two variables. It ranges from -1 to 1.

- 1 indicates a perfect positive linear relationship.
- 0 indicates no linear relationship.
- -1 indicates a perfect negative linear relationship.

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

Feature scaling is a technique used to standardize the independent variables or the predictors in the dataset in a fixed range. It is performed as a pre-processing step to building a linear regression model.

Feature scaling is performed due to following reasons:

- To prevent some features from having disproportionately large coefficients
- Faster convergence while using gradient descent algorithms

Standardized Scaling Vs Normalized Scaling:

- Normalized Scaling, also known as min-max scaling - scales features between 0 and 1
- Standardized Scaling – scales features to have a mean of 0 and standard deviation of 1

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

Value of VIF becomes infinite when there is a perfect correlation between two independent variables or predictors. The R squared value is 1 in this case and since VIF is calculated as 1/(1-R squared), the value becomes infinite. This would mean that there is multi-collinearity between two of the predictor variables and we have to drop one of them to form a working linear regression model.

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

Q-Q plot or Quantile-Quantile plot is a graphical tool used in statistics to analyse if a given dataset follows a particular theoretical distribution like normal distribution etc.

Use & Importance of Q-Q plot in linear regression:

- Used to evaluate if residuals from a linear regression model are normally distributed.
- Can help in identifying outliers in residuals of a linear regression model