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?

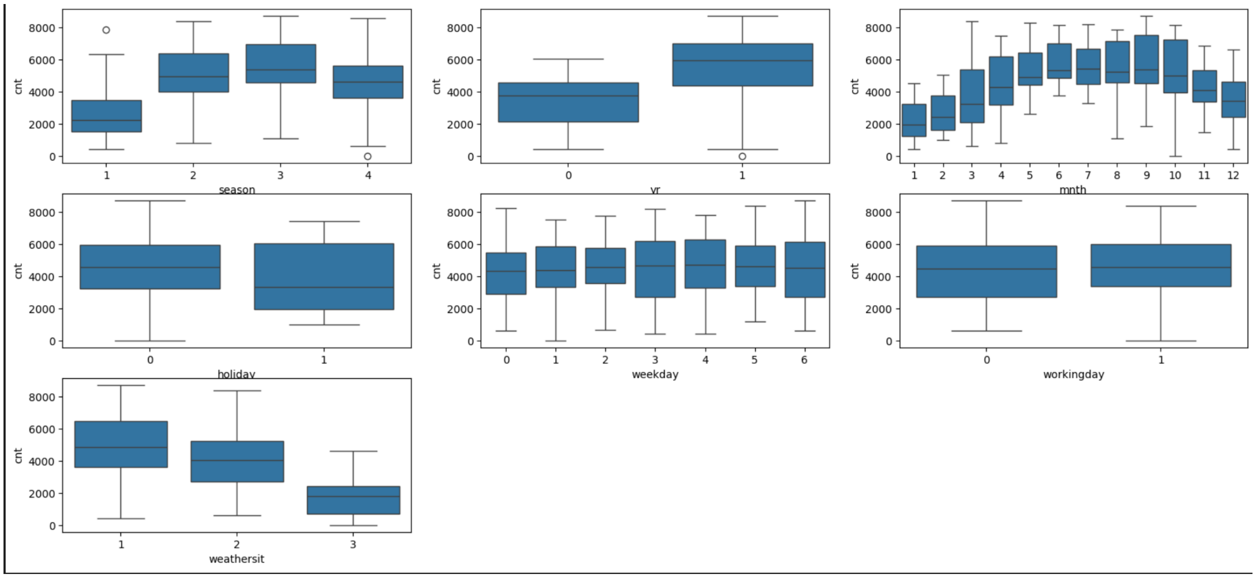
Ans:

I have done the analysis on both numerical and categorical data using different plots and diagrams

Analysis:

Booking Timelines:

- #3 season(Fall) which is the fall seasons where maximum booking was done
- 2019 year had more bookings compared to 2018
- Sep and Oct(9 and 10) months had the max bookings
- Usually 3ed day(Wednesday) during the week had good bookings
- Clear days had the good bookings compared to heavy snow/rain days
-



2- Why is it important to use drop_first=True during dummy variable creation?

Ans: When creating dummy variables for categorical features, one category is typically the baseline category. By including a dummy variable for every category, you introduce perfect multicollinearity into your model, which means that one variable can be perfectly predicted using the other variables. This is problematic for many statistical models, such as linear regression, where multicollinearity can cause issues with the estimation of coefficients. By setting drop_first=True, you drop one of the dummy variables (typically the first one) and use it as the reference category, thus avoiding multicollinearity.

Dropping the first dummy variable makes the model coefficients easier to interpret.

Including all dummy variables creates redundant information because the sum of all dummy variables for a given categorical feature is always one. By dropping the first dummy variable, you reduce this redundancy, leading to a more efficient model.

Q3- Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable

Temp/atemp

Q-4 How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: - Error terms should be normalized
The histogram of residuals resembles a normal distribution.
There should be insignificant multicollinearity among variables
No auto-correlation

Q5 - Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- Temp/atemp
- Winter
- Clear weather

General Subjective Questions

Q1- Explain the linear regression algorithm in detail.

Linear regression is a supervised learning algorithm that compares input (X) and output (Y) variables based on labeled data. It's used for finding the relationship between the two variables and predicting future results based on past relationships.

Simple Linear Regression

Completing a simple linear regression on a set of data results in a line on a plot representing the relationship between the independent variable X and the dependent variable Y. The simple linear regression predicts the value of the dependent variable based on the independent variable.

Linear Regression Formula

$Y=mx+c$ where m is slope and c is Intercept and Y is predicted value

The correlation coefficient or R-squared value helps in determining if the model is fit properly. The R-squared value ranges from 0 to 1.0, denoting zero correlation at the low end (0) and a 100% correlation at the high end (1.0).

Q2- Explain the Anscombe’s quartet in detail.

Anscombe’s Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets
is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

Key Statistics

For all four datasets, the following statistics are almost identical:

- Mean of xxx values
- Mean of yyy values
- Variance of xxx values
- Variance of yyy values
- Correlation between xxx and yyy
- Linear regression line (yyy = mx+cmx + cmx+c)

Q3 What is Pearson’s R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables.

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

As the correlation coefficient increases in magnitude, the points become more tightly concentrated about a straight line through the data.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope ( i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope ( i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

---

Q4- What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the process of transforming the features of your data to fall within a specific range or to have specific properties. This is often done to ensure that different features contribute equally to the model, especially when the features have different units or vastly different scales.

Why is Scaling Performed?

- Helps to improve model performance
- Coverages Faster
- Interpretability

Normalized Scaling vs. Standardized Scaling
Normalization scales the data to a fixed range, usually [0, 1] or [-1, 1]. It is often used to maintain the relationship between the data points.

Standardization scales the data to have a mean of 0 and a standard deviation of 1. This is useful for algorithms that assume the data is normally distributed.

---

Q5- You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors in the model.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

---

Q6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to assess whether a set of data follows a particular distribution, typically a normal distribution.

**Axes**:

- **X-axis**: Theoretical quantiles from the specified distribution.
- **Y-axis**: Sample quantiles from the dataset.

In linear regression, several assumptions need to be validated for the model to be considered reliable. One of these assumptions is that the residuals (differences between observed and predicted values) are normally distributed. A Q-Q plot helps in assessing this assumption.

Importance of Q-Q Plots in Linear Regression

Ensuring that the residuals are normally distributed is crucial for the validity of hypothesis tests and confidence intervals in linear regression.

Q-Q plots are a diagnostic tool to identify potential problems with the model, such as outliers or non-normality.