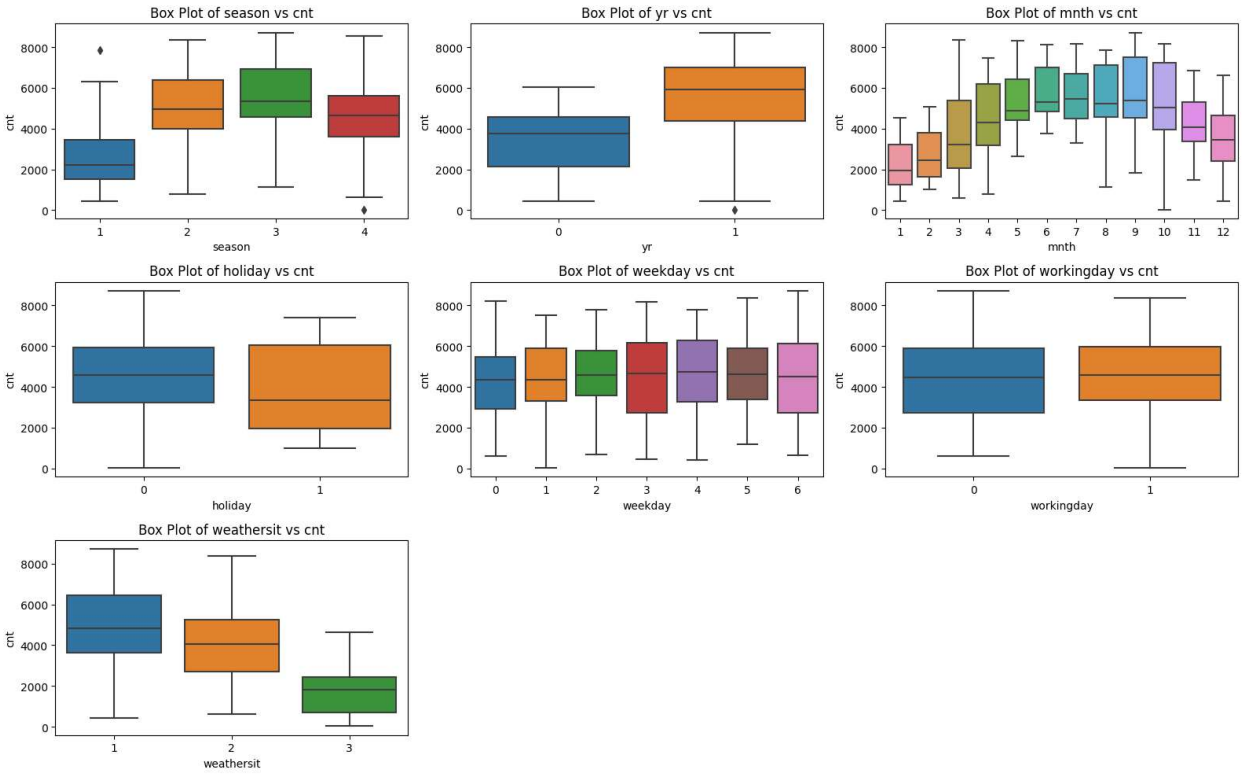
Assignment based subjective questions

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans:



From the analysis of above charts here are key observations:

- Seasonal Effects: Seasons have a strong influence on bike rentals, with Summer and Fall generally seeing higher demand.
- Weather Conditions: Clear weather promotes bike usage, while severe weather conditions (like heavy rain or snow) significantly reduce rentals.
- Temporal Trends: The increase in rentals from 2018 to 2019 suggests a growing adoption of the service.
- Day-Specific Effects: Holidays, working days, and weekdays all have varied impacts on bike rentals depending on the primary use case of the service (commuting vs. leisure).

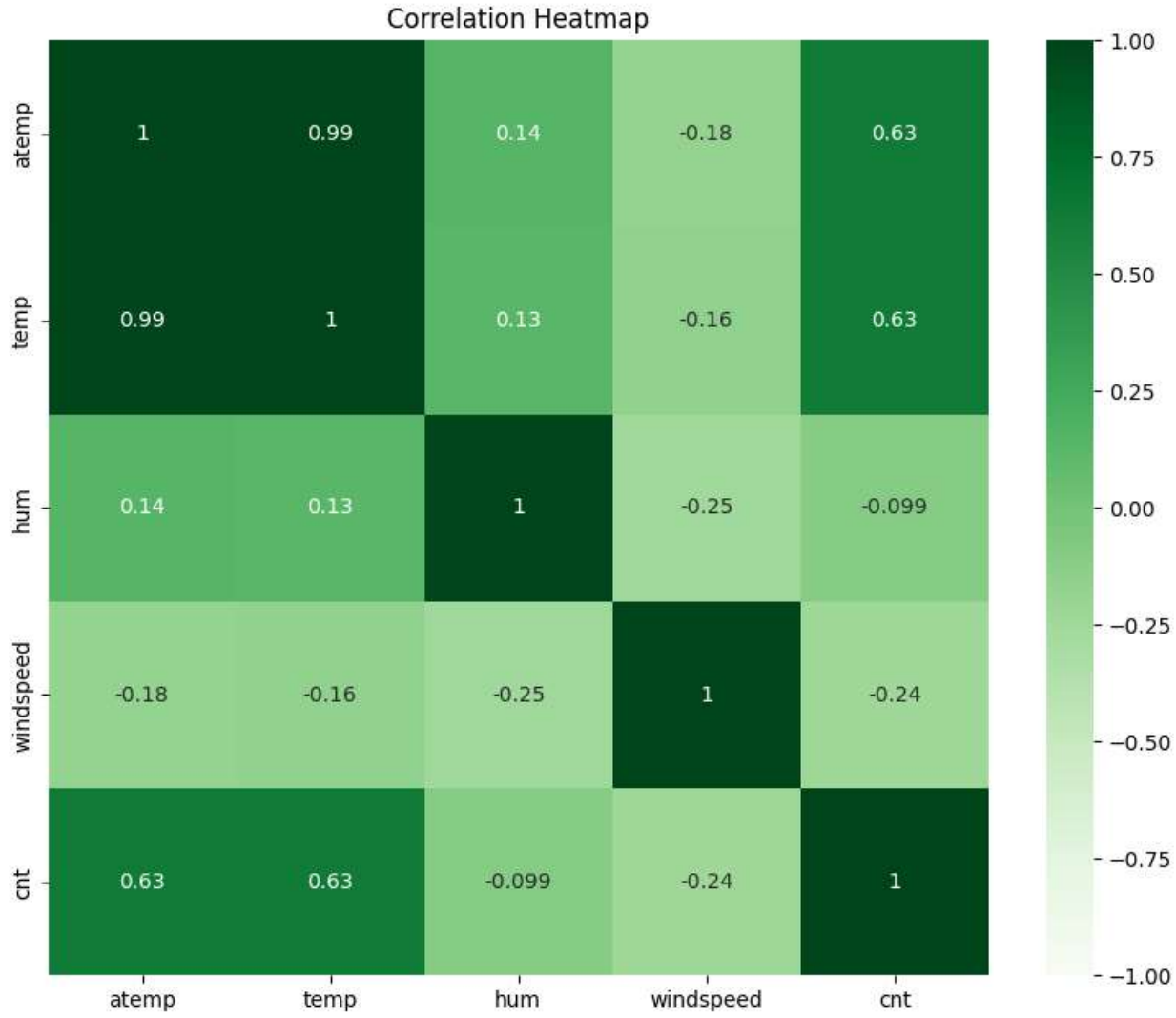
Question 2: Why is it important to use drop_first=True during dummy variable creation?

Ans: Using drop_first=True during dummy variable creation is important to avoid multicollinearity in the resulting dataset, particularly when you plan to use the dummy variables in a regression model. If all n dummy variables included in model, they become linearly dependent, meaning one dummy variable can be perfectly predicted using the others. This may make model unstable, less reliable.

Further this practice, reduces the number of features, making the model simpler without losing any information.

Question 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans:

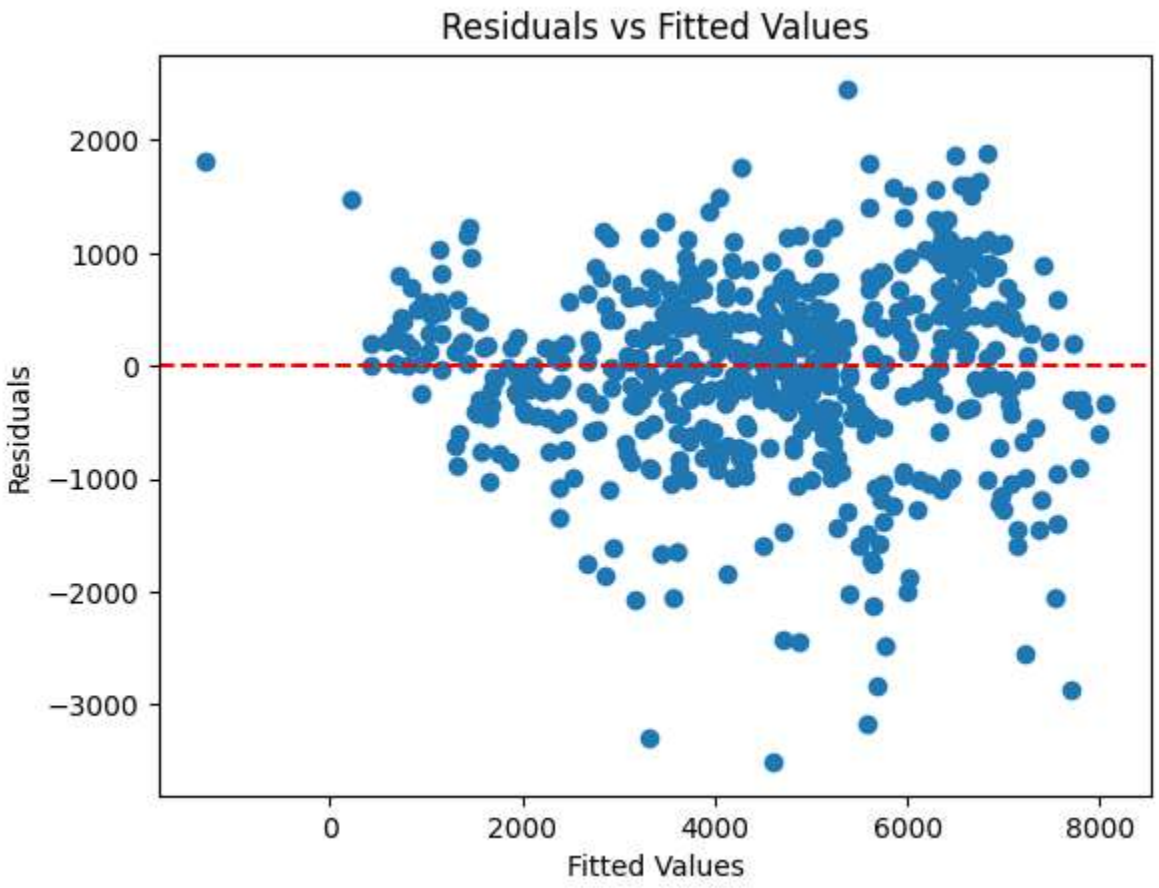


Based on heatmap above, temp & atemp are having highest correlation with 'cnt' which is target variable here.

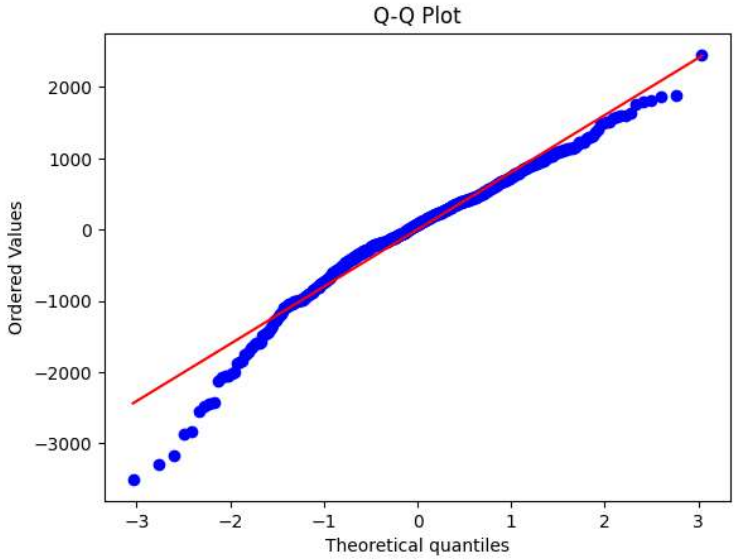
Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

- **Residuals vs. Fitted Values Plot:** Plotted the residuals (the differences between observed and predicted values) against the fitted values (predicted values). Since the residuals are randomly

scattered around zero line without any clear pattern, this indicates that the linear model is appropriate.



- Q-Q Plot: Q-Q plot compares the distribution of the residuals to a normal distribution. Since the points lie along the 45-degree line, the residuals seem to be approximately normally distributed.



- Variance Inflation Factor (VIF) and p-Values: VIF quantifies the degree of multicollinearity. Reviewed the VIF and p-values of all independent variables to ensure there are no multi-correlation.

	Feature	VIF	p-value
0	yr	1.033789	3.617596e-109
1	temp	1.810716	8.650553e-96
2	hum	1.959372	1.820193e-06
3	windspeed	1.188222	1.455213e-09
4	season_summer	1.833600	1.540312e-10
5	season_winter	1.906238	3.799809e-37
6	mnth_3	1.120222	1.540338e-02
7	mnth_5	1.504067	2.561812e-01
8	mnth_7	1.581694	1.370847e-01
9	mnth_9	1.250916	6.488314e-09
10	mnth_11	1.628914	3.619408e-03
11	mnth_12	1.312418	1.437777e-02
12	weekday_6	1.008684	2.387230e-02
13	weathersit_Mist	1.595088	1.017720e-05
14	weathersit_Light Snow/Rain	1.328363	3.555639e-17

Question 5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Considering snapshot provided above, VIF and P-Values - Top 3 significant features with acceptable VIF and p-values: 'yr', 'temp', 'season_winter'

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

- Linear regression assumes that the relationship between the dependent variable y and the independent variables x1,x2,……,xn is linear.

- The model attempts to find the best-fitting straight line (in the case of one predictor) or hyperplane (in the case of multiple predictors) that describes the relationship between the input features and the target variable.

The model attempts to derive the equation of straight line by calculating intercept and slop coefficients. The equation of best fitting line can be used to make the prediction.

- Model assumes – linear relationship, independence of observations, homoscedasticity and no multicollinearity amongst the independent variables

Question 2: Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (such as mean, variance, correlation, and linear regression line) but are graphically very different

Each of the four datasets in Anscombe's quartet consists of 11 (x, y) pairs. Despite having similar statistical properties, the datasets reveal very different relationships between the variables when graphed. The quartet emphasizes the importance of graphical analysis in addition to statistical analysis.

- Identical Statistics, Different Visuals: Anscombe's quartet consists of four datasets that share nearly identical summary statistics (mean, variance, correlation, and regression line) but differ significantly when visualized graphically.

- Demonstrates the Importance of Visualization: The quartet highlights how relying solely on summary statistics can be misleading, emphasizing the need for data visualization to fully understand the underlying relationships.

- Impact of Outliers: The datasets illustrate how outliers or influential points can drastically affect statistical measures and regression models, often masking or distorting the true data patterns.

- Linearity vs. Non-Linearity: While one dataset follows a clear linear pattern, another demonstrates a non-linear (quadratic) relationship that a simple linear regression fails to capture properly.

- Misleading Correlations: Despite similar correlations across the datasets, the actual relationships between variables differ widely, showing that correlation alone does not fully describe data relationships.

Question 3: What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables.

- **Measures Linear Relationship:** Pearson's R quantifies the strength and direction of a linear relationship between two continuous variables.
- **Range:** It ranges from **-1** (perfect negative correlation) to **+1** (perfect positive correlation), with **0** indicating no linear correlation.
- **Formula:** It is calculated using the covariance of the variables divided by the product of their standard deviations.
- **Interpretation:** Positive values indicate that as one variable increases, the other tends to increase. Negative values indicate that as one variable increases, the other tends to decrease.
- **Assumptions:** Pearson's R assumes linearity, normally distributed continuous variables, and no significant outliers.
- **Limitation:** It only captures linear relationships and is sensitive to outliers; it does not imply causation.

Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Scaling is used to bring all features onto a similar scale, which is important for many machine learning algorithms.
- Normalization vs. Standardization:
    - Normalization (Min-Max Scaling): Scales features to a specific range, typically [0, 1].
    - Standardization (Z-Score Scaling): Scales features to have a mean of 0 and a standard deviation of 1.
- Improves Model Performance: Scaling ensures that all features contribute equally, improving model performance, especially in distance-based algorithms like KNN and SVM.
- Faster Convergence: Gradient-based algorithms, like gradient descent, benefit from scaling as it leads to faster convergence during training.
- When to Use:
    - Normalization is preferred when the data needs to be within a specific range or for algorithms that require bounded input.
    - Standardization is ideal for algorithms that assume normally distributed data or require features with the same variance.

Question 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- Infinite VIF Indicates Perfect Multicollinearity: VIF becomes infinite when one variable is a perfect linear combination of other variables in the model.
- Caused by R-squared of 1: This happens because, in the VIF formula, if the R-squared from regressing one variable on the others equals 1, the denominator becomes zero, leading to an infinite VIF.
- Perfect Collinearity Example: An example is including both age and years since birth as predictors; they are perfectly collinear.
- Impact on Model: Infinite VIF signals that the model cannot uniquely estimate coefficients due to exact linear dependency among predictors.
- Solution: To resolve this, remove or combine the perfectly collinear variables.

Question 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- Q-Q Plot Overview: A Q-Q plot compares the quantiles of your data against the quantiles of a theoretical distribution, usually the normal distribution, to assess how well the data fits that distribution.
- Normality Check: In linear regression, a Q-Q plot is primarily used to check if the residuals (errors) are normally distributed, which is a key assumption of the model.
- Interpretation: If the points in the Q-Q plot lie on or near the 45-degree line, the residuals are likely normally distributed. Deviations from the line indicate non-normality.
- Detects Outliers and Non-Normality: The plot helps identify outliers, skewness, and heavy tails, which can suggest violations of model assumptions.
- Model Diagnostics: A Q-Q plot is an essential diagnostic tool used alongside other plots to evaluate the fit of a linear regression model and determine if any adjustments are needed.
- Improves Model Validity: By assessing the normality of residuals with a Q-Q plot, you can ensure the reliability of hypothesis tests, confidence intervals, and overall model performance in linear regression.