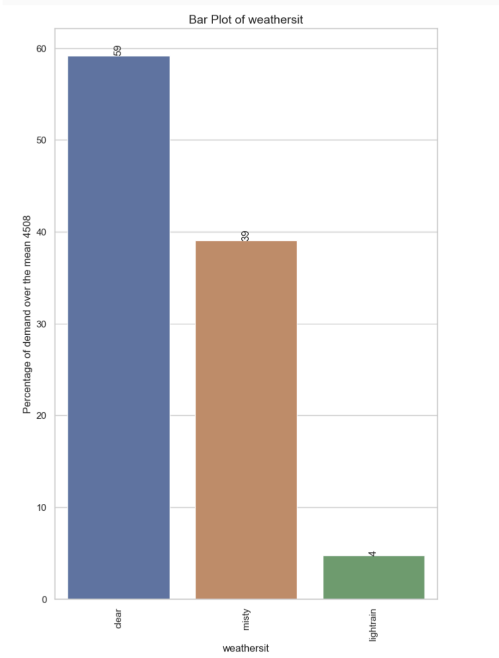
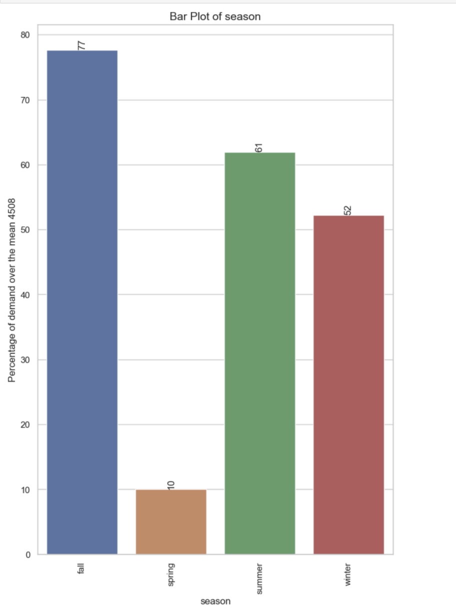
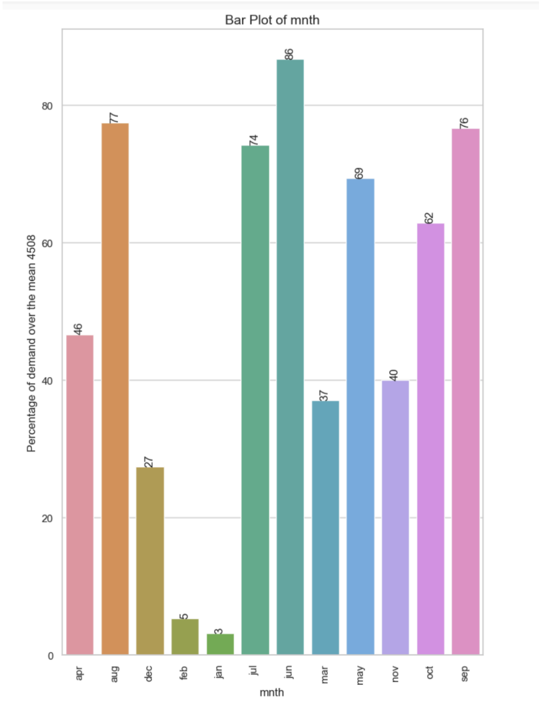
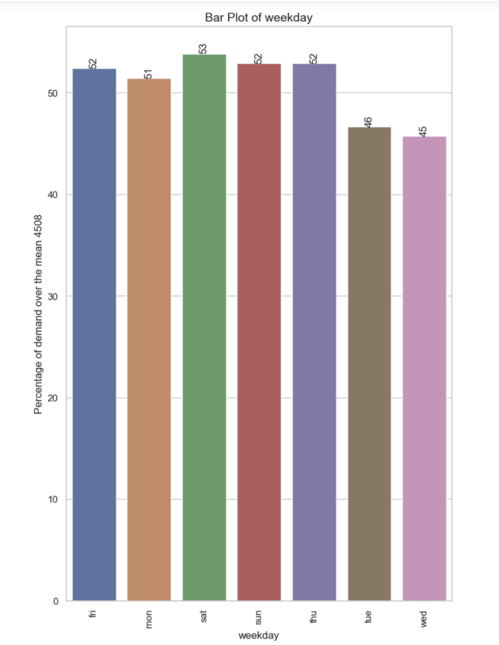
Subjective Questions

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







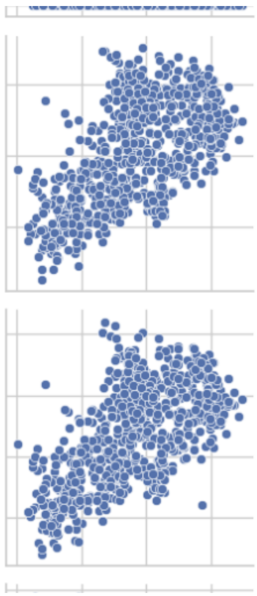


- When the sky is clear 59% of the demand is over the mean
- Fall is the best season for the demand with 77% of days on which the demand is over the mean

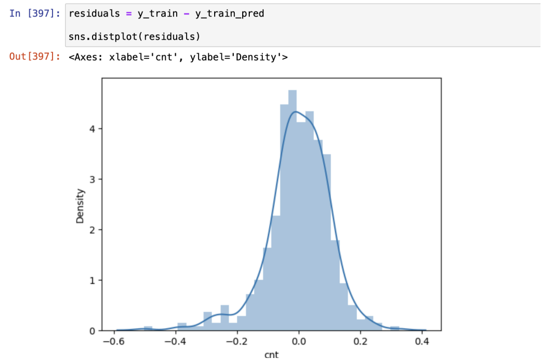
- Fall is the best season for the demand with 86% of days on which the demand is over the mean
- Apart from Tuesday and Wednesday, all the other days seem to have similar demand.

2. Why is it important to use drop_first=True during dummy variable creation?
Using drop_first=True during dummy variable creation in linear regression is important to avoid the dummy variable trap. The dummy variable trap occurs when there is perfect multicollinearity among the dummy variables, which means that one of the dummy variables can be perfectly predicted from the others. This perfect multicollinearity can cause problems in estimating the regression coefficients.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Temp and atemp have 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?
After conducting residual analysis, I observed that the median is approximately 0.



```
In [397]: residuals = y_train - y_train_pred

          sns.distplot(residuals)
Out[397]: <Axes: xlabel='cnt', ylabel='Density'>
```

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

1. Temperature explains it positively
2. Year explains it positively
3. Windspeed explains it negatively

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable (response) and one or more independent variables (predictors). The goal is to find the linear equation that best predicts the dependent variable from the independent variables. The equation of a simple linear regression model (with one predictor) is "y = b0 + b1x", where:

"y" is the dependent variable.
"x" is the independent variable.
"b0" is the intercept.
"b1" is the slope (coefficient).
For multiple linear regression (with multiple predictors), the equation is "y = b0 + b1x1 + b2x2 + ... + bnxn".

The steps of the linear regression algorithm are:

- Data Collection: Gather data containing the dependent variable and independent variables.
- Data Preparation: Clean and preprocess the data, handle missing values, and convert categorical variables into numerical ones using dummy variables.

- Feature Selection: Select the right set of features using automated plus manual process.
- Model Training: Use the training data to estimate the coefficients "b0, b1, ..., bn" that minimize the sum of squared residuals (differences between observed and predicted values). This is typically done using the Ordinary Least Squares (OLS) method.
- Model Evaluation: Assess the model's performance using metrics like R-squared, Mean Squared Error (MSE), or Root Mean Squared Error (RMSE). Plotting residuals can also help check for assumptions like homoscedasticity and normality.
- Prediction: Use the model to predict the dependent variable for new data.

2. Explain the Anscombe’s quartet in detail. (3 marks)
Anscombe’s quartet is a collection of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and regression line) but appear very different when graphed. This was created by statistician Francis Anscombe to illustrate the importance of graphical analysis of data.

The datasets have:

- The same mean of "x" and "y".
- The same variance of "x" and "y".
- The same correlation coefficient between "x" and "y".
- The same linear regression line "y = 3 + 0.5x".

However, their scatter plots reveal:

- Dataset 1 shows a typical linear relationship.
- Dataset 2 shows a nonlinear relationship (a curve).
- Dataset 3 has a single outlier that influences the regression line.
- Dataset 4 shows a vertical line with almost all points having the same "x" value except one outlier.

Anscombe’s quartet emphasizes that relying solely on statistical summaries can be misleading and underscores the necessity of visualizing data to understand its structure and detect anomalies.

3. What is Pearson’s R?
Pearson’s R, also known as the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. It is a dimensionless value that ranges from -1 to 1, where:

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

The formula for Pearson’s R is:

$$r = (\Sigma(xi - \bar{x})(yi - \bar{y})) / \sqrt{(\Sigma(xi - \bar{x})^2 * \Sigma(yi - \bar{y})^2)}$$

Where "xi" and "yi" are the data points, and "$\bar{x}$" and "$\bar{y}$" are the means of "x" and "y", respectively. Pearson’s R is widely used in statistics to infer correlation and causation between variables.

4. 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. Scaling is crucial in machine learning, especially for algorithms that are sensitive to the magnitudes of the input features (like gradient descent optimization and distance-based algorithms).

Reasons for Scaling:

- Improves Convergence Speed: In optimization algorithms, scaling can lead to faster convergence.
- Prevents Dominance of One Feature: Ensures that no single feature dominates due to its magnitude.
- Reduces Bias: Helps models treat all features equally.

Normalized Scaling (Min-Max Scaling):

- Transforms data to a fixed range, usually [0, 1].
- Formula: $X' = (X - Xmin) / (Xmax - Xmin)$
- Sensitive to outliers as it scales based on min and max values.

Standardized Scaling (Z-Score Normalization):

- Centers the data around the mean with a standard deviation of 1.
- Formula: $X' = (X - \mu) / \sigma$
- Less sensitive to outliers but assumes data follows a Gaussian distribution.

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

Reasons for Infinite VIF:

- Exact Linear Relationship: If one predictor variable can be exactly predicted from others, the regression model cannot distinguish their individual effects.
- Perfect Multicollinearity: When the design matrix X is singular or non-invertible.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a particular distribution, typically the normal distribution. It compares the quantiles of the dataset against the quantiles of a specified theoretical distribution.

Here's how a Q-Q plot works:

- The data points are sorted in ascending order.
- The quantiles of the data are plotted against the quantiles of the specified theoretical distribution (usually the standard normal distribution).
- If the data points fall approximately along a straight line, it suggests that the data closely follows the specified distribution.

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

- Check Normality of Residuals: One of the assumptions in linear regression is that the residuals (errors) are normally distributed. A Q-Q plot of the residuals can visually assess this assumption.
- Detect Deviations from Normality: If the residuals follow a normal distribution, the points on the Q-Q plot will fall approximately along a straight line. Deviations from this line indicate departures from normality, such as skewness or kurtosis.
- Validation of Model Assumptions: Ensuring that residuals are normally distributed helps validate the assumptions of the linear regression model. Violations of this assumption can lead to biased parameter estimates, incorrect inference, and unreliable predictions.
- Model Improvement: Identifying non-normality in the residuals through Q-Q plots can guide model improvement strategies. For instance, transforming the response variable or introducing additional explanatory variables may help improve the normality of residuals.

In summary, Q-Q plots are valuable diagnostic tools in linear regression analysis. They provide insights into the distributional properties of residuals and help ensure that the assumptions underlying the regression model are met, ultimately leading to more reliable and accurate inference and predictions.