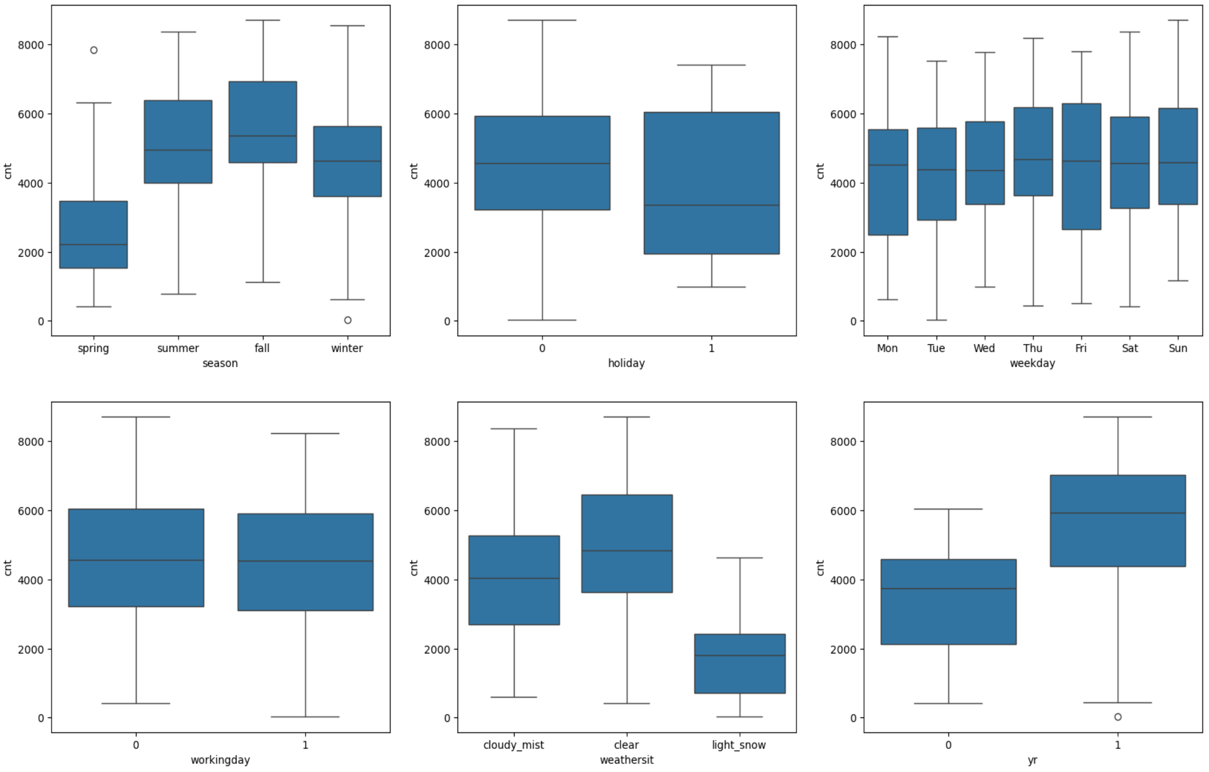
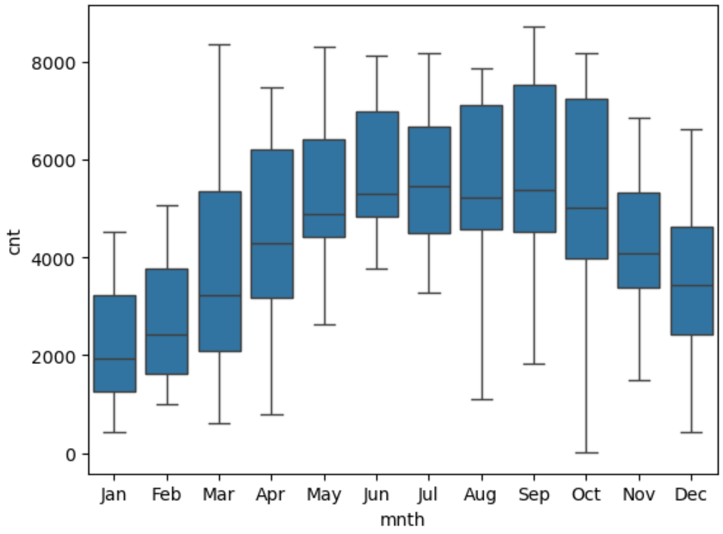
Arpan Das – Subjective Questions

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

   Ans: From the graphs we can infer that

   a. Seasonal demand- Spring is the worst season for bikes as per the data
   b. A clear sky leads to higher count of bike rentals followed by cloudy, snow results in least usage of bikes
   c. People use bikes less during holidays
   d. There not much difference in bike rental count in working vs weekdays
   e. 2019 there was spike in bike usage when compared with 2018, indicating a growing trend
   f. Months resonate the impact of weather, with cold and winter/spring months (Nov -Mar) resulting in least bike rentals





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

   Ans: Reduces multicollinearity by dropping the first category for each categorical variable. This helps reduce multi collenirarity among the varibales and efficient regression. As the effect of dropped category is captured in the other category variables no impact is missed. The dropped category serves as the reference, and the remaining dummies indicate how much they differ from this reference. Instead of interpreting each category in isolation, we can interpret them in terms of differences from the reference category.

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

   Ans : Temperature (temp/temp) has highest collineraity (0.63) with the target variable – cnt.

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

   Ans: Please find below the assumptions and how they were validated

   a. Linear relationship between predictor and target variables was identified through scatter plots before model building.
   b. Multi collinearity was identified through scatter plots, correlation matrix heatmap visualization before model building. Post the linear modelling multi collinearity was checked using VIF to ensure VIF < 5
   c. Multivariate Normality i.e. Normal distribution of the residuals was checked using residual plots namely displot. For quantitative evaluation Shapiro-Wilk Test or Anderson Darling test can be used.
   d. Homoscedasticity: checked using residual plots
   e. Independence of errors though not required in this case as its not timeseries checked through ACF plot

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

   Ans: Following are the the top 3 features contributing significantly towards explaining the demand of the shared bikes

   i. Temp – coefficient: 0.57
   ii. Year – coefficient: 0.23
   iii. Weathersit_light_snow - coefficient: (-0.24)

6. Explain the linear regression algorithm in detail.

   Ans: Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable and one or more independent variables. The goal is to fit a linear equation to the observed data such that it minimizes the error between the predicted and actual value.
   It can be simple linear regression if there is one predictor varibale or multiple linear regression if more than one predictor variables

   Simple Linear regression equation :

   $$y = \beta_0 + \beta_1 x + \epsilon$$

   Multiple Linear regression equation:

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

   where y is independent variable and x1… xn are dependent variables. β – slope and coefficients and ε Error

   Objective of the algorithm is to find the right combination of constant and the coefficients to get the minimum sum of errors/ residual. Residual defined y actual – y predicted.

   Loss function which is minimized to find the best-fit line Linear regression - *Mean Squared Error (MSE)*

Assumptions:

i. Linearity: The relationship between independent (x) and dependent (y) should be linear.
ii. Independence: Observations are independent of each other.
iii. Homoscedasticity: The variance of residuals is constant across predictor values
iv. Normality: Residuals are normally distributed.
v. No multicollinearity: Independent variables are not highly correlated with each other.

Steps of Linear Regression Algorithm

a. Define the hypothesis (linear equation) to model the relationship(h(x)=β0+β1x1+β2x2 +⋯+βnxn)
b. Initialize Parameters Start with initial guesses for β0,β1,…,βn (often set to 0)
c. Compute Predictions - Use the current parameter values from (b) to compute the predicted values for all observations
d. Calculate the Loss Function - Compute the MSE to quantify the error between predicted and actual values
e. Optimize Parameters using Gradient Descent to minimize MSE using a Learning rate
f. Evaluate the model using R2, Adj R2 and MSE

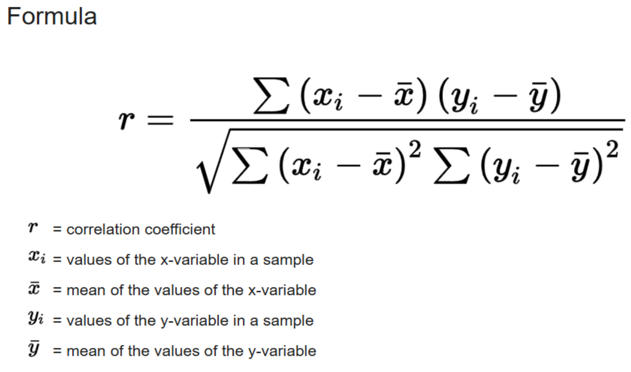
7. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet is a set of four datasets created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it statistically. Anscombe's Quartet emphasizes that relying solely on summary statistics (or even regression analysis) can be misleading without visualizing the data. It highlights how very different data distributions to have the same statistical measures signifying the importance of visulaization.

8. What is Pearson's R?

Ans: Pearson correlation coefficient**,** is a statistical measure that quantifies the strength and direction of a linear relationship between two variables. Its values range from -1 to +1.
-1 represents strong negative linear relationship
+1 represents strong positive linear relationship
0 represents no relationship between the variables

Formula of Person's R –



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling refers to the process of transforming the features (variables) of a dataset into a common scale without distorting the differences in the ranges of values. It is eequried to esnure that predictors without high numerical values don't dominate the model.
While P value and model accuracy doesn't change with scaling, it helps in

a. Making the interpretation easier by making Coefficients more comparable
b. Improves performance by making gradient descent algorithm faster

Typically, there are two methods to scale one is Normalized (or) Min Max Scaler and other standardization scaling.

Difference between normalized scaling and standardized scaling

Normalized Scaling (Min Max Scaler)

Rescales data to a specific range (usually [0, 1]). Has a fixed range 0 to 1. Handles outliers.

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Standardized scaling

Centers the data around the mean with unit variance. No fixed range; mean = 0, variance = 1.Its less less sensitive to outliers.

$$X_{std} = \frac{X - \mu}{\sigma}$$

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

Ans: A VIF becomes infinite or extremely large when the correlation between one predictor and the other predictor variables is perfect (i.e. R square =1 or tends to 1). This means that one predictor variable can be exactly predicted by a linear combination of the other predictors- perfect multicollinearity.
High VIF is observed if there is high multicollinearity among variables, duplicate variables or linear dependency within the variables (e.g. x3 = x1+ 3 * x2).

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

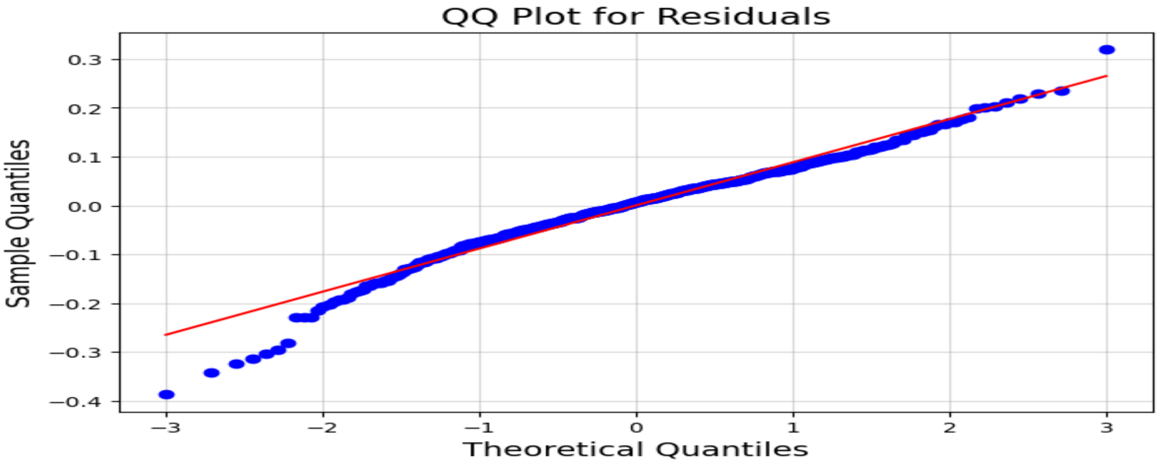
Ans: Quartile – Qurtile plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It is constructed by plotting the quantiles of the data against the corresponding quantiles of the reference distribution.

X-axis: Theoretical quantiles e.g., normal distribution
Y-axis: Sample quantiles from the data

Insights from QQ Plot for Linear regression

a. Q-Q plot used to indetify the normality of a distribution. If normal they will appear as strainght line.Can be used in linear regression to test the Residuals if they are normally distributed or not
b. Points deviating from the straight line at the ends suggest the presence of heavy tails (outliers)
c. Systematic deviation like S shape from the line on one side indicates skewness in the residuals.

QQ plot of residuals from my linear regression assignment.


QQ Plot for Residuals
Sample Quantiles
Theoretical Quantiles