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? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

The dependent variable (Count of Bike rentals, or Demand for Bike rentals on a given day) is determined by the equation:

Demand of bike rentals (count of rentals on a given day**) Y =**
2040.728472 * *(yr)* **-856.298580*** *(holiday)* **+3915.657583 *** *(temp)*
-1211.631001 * *(windspeed)* **-973.775471*** *(season_spring)* **+ 401.913171*** *(season_winter)*
-636.310347* *(mnth_jul)* **+ 495.793864*** *(mnth_sept)*
-696.415150 *(weathersit_2)* **-2487.368512 *** *(weathersit_3)*
+ 2217.347907

From the categorical variables from the given data set: Holiday, Temperature, Windspeed, Spring season, Winter season, Month of July and Septerber and certain Weather situations best explain the expected demand for the bike rentals.

1. Temperature has a very high positive coefficient – indicating it is a strong predictor of the demand, suggesting people prefer to take bikes on rentals on warm or hot days over nike rentals on a cold day
2. Windspeed has a negative coefficient suggesting less interest from riders to take a bike rental on a windy day
3. If the weather situation is misty or indicative of rain, thunderstorm, or snowfall, the demand is likely to fall
4. Seasonality and months also explain the demand

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

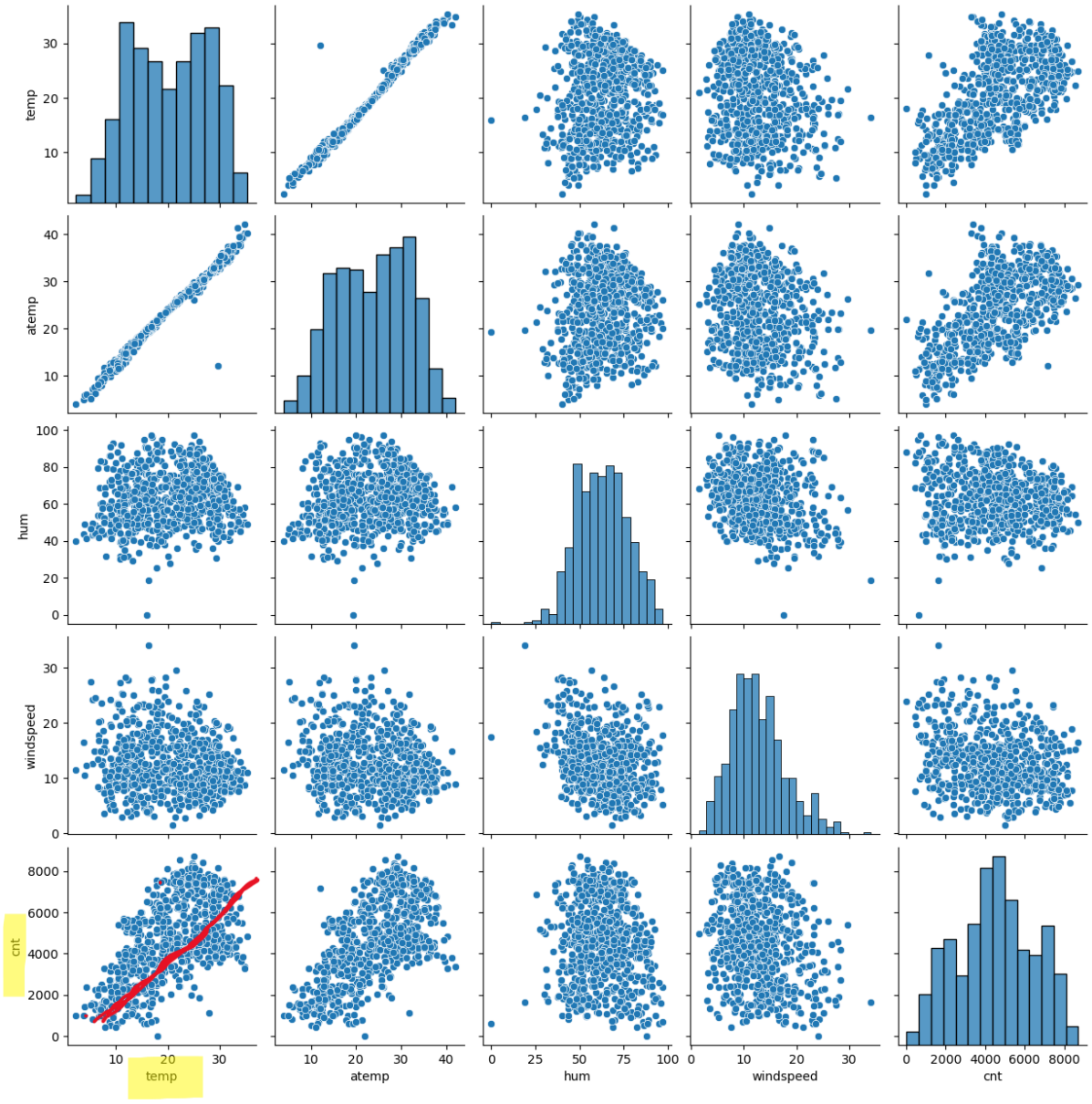
Dummy variables use 0/1 to represent the categorical variables. If all the variables are included, with all the dummy variables, it leads to perfect multicollinearity. One dummy variable can be linearly predicted from the others. This causes issues in linear regression models as it goes against the assumptions of linear regression models that all the variables are independent and cannot be explained by other variables. If this assumption fails due to inclusion of all categories of dummy variables, the model fails because they cannot handle perfectly collinear features.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

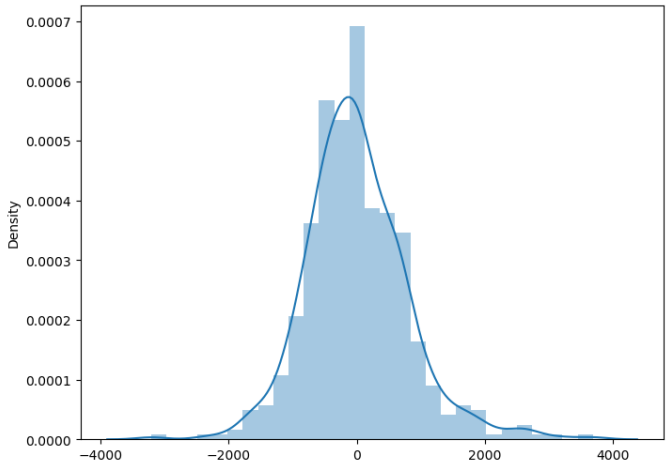
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

From the pair plots of numerical variables Temperature (temp) has the highest correlation with the target variable cnt (Demand for rental bikes).



Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)
Checking for assumptions of Linear Regression:

Normality: The error terms / residuals of the model are normally distributed:
Plotting a histogram of the error terms of the training data.

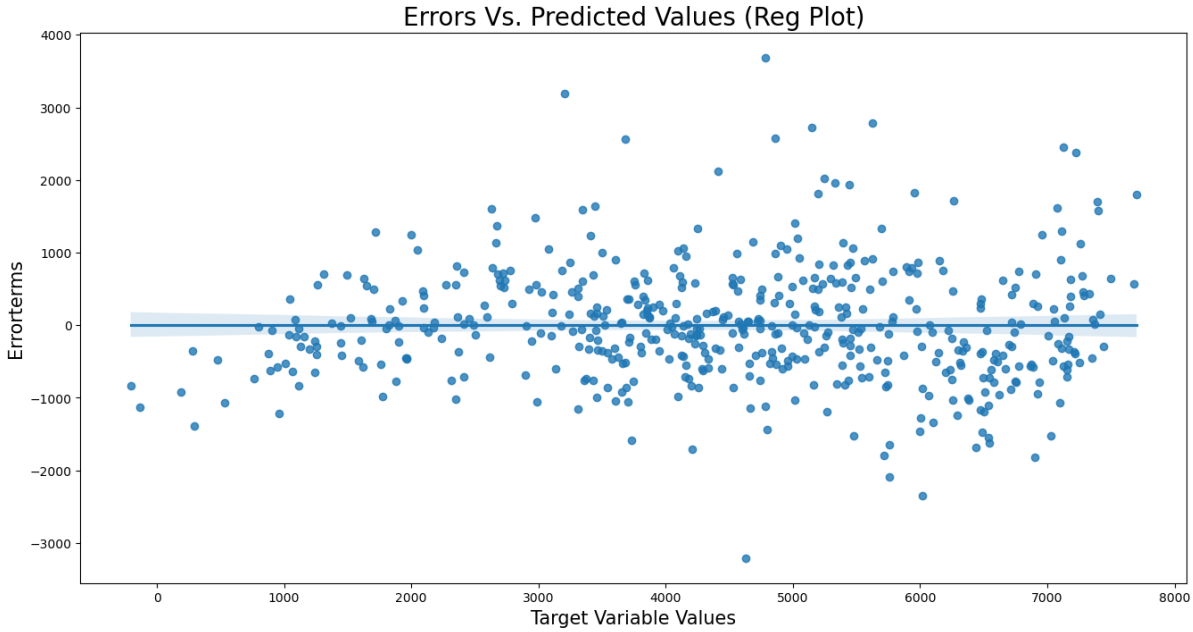


The histogram for residuals on training data is a normally distributed histogram with mean at zero.

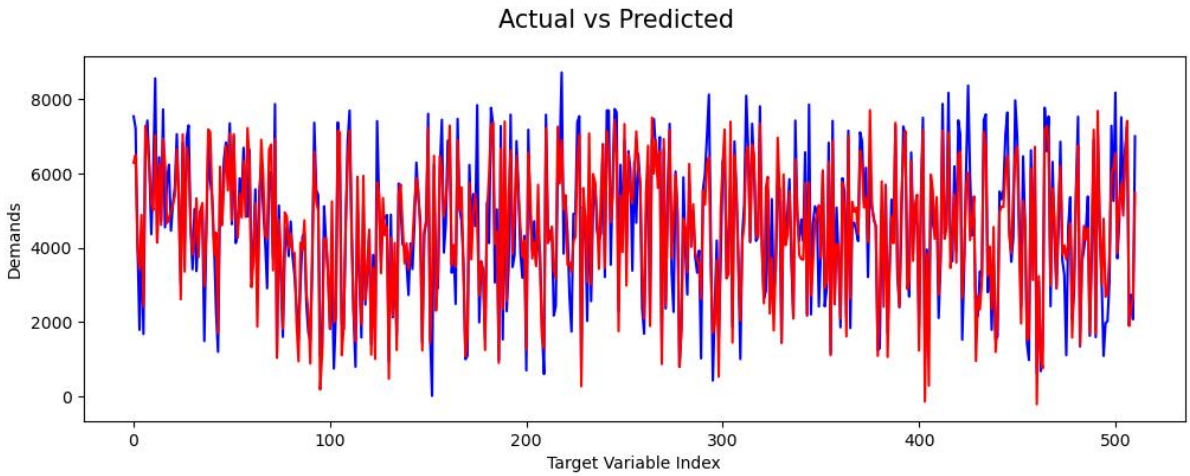
Independence: The error terms are with-in 2 standard deviations on either size of Zero.

Homoscedasticity: The target values have same variance at every level of X

The plot below showing uniform scatter of error terms for all values of Y validates Homoscedasticity of the model



Linear relationship: The Y predicted follows the same pattern as that of Y actual and this proves a linear relationship of the model on the training data



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Based on the final linear regression model developed, the features with most significant impact on the demand are following:

1. Temperature: Positive correlation – For increase in 1 degree of temperature the demand increases by 3915.65, indicating riders prefer to rent bikes on warmer days compared to colder days
2. Year: Positive Correlation – For increase in 1 year of operations, current model predicts an increase in demand of 2040.72. This could be due to successful operations and marketing over the base year which added this increase in demand. The reason for this is still unclear from the data, and this only holds true for current year, what is the degree of impact on future years and significance of year as a predictor can only be assessed with more data in future.
3. Weather situation : Negative Correlation – If the weather it can negatively impact the demand by following units -696.415150 (weathersit_2: Misty, Cloudy conditions) -2487.368512 (weathersit_3: rainy, hailstones, thunderstorms etc)

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear Regression is a statistical method used to predict a target variable (Y) based on given

features or predictors (X).
The method works by analyzing various categorical & numerical inputs that can impact the variable to be predicted and tries to identify the variables which explain the value of Y the most for a given value of X.

The method tries to arrive at a simple linear equation between Y and X , which can be stated as follows :

$$Y = \sum \beta_i X_i + C$$

for i in 0 to n,
where n is the number of predictors,
β is the coefficient of the respective predictor variable,
C is the constant

The method tries to determine the significance of predictors based on certain assumptions like values of X are independent of each other, one variable cannot be explained by another variable, there is no multicollinearity between the variables, and variables have constant coefficients, meaning the degree of increase in Y for a unit increase in X_i is fixed and can be explained by the model.

Using a training data the equation is created with appropriate values, and any variables which can lead to overfitting of the model or collinearity are removed from the model.

The trained model is used to predict values of test data and creates a feedback loop to correct the coefficients with increase in training data.

Certain statistical values like F-Statistic, R-Squared value, Adjusted R-Squared value, p-values of predictors are used to ensure the model follows basic assumptions of linear regression and avoids overfitting or memorizing the training data.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet, consists of four datasets specifically designed to demonstrate the importance of visualizing data before analyzing it.

All four datasets share identical statistical properties, such as:

- Mean
- Variance

- Identical correlation coefficient
- The same regression line (y = 3 + 0.5x),

However, once the data is plotted, they exhibit very different behavior.

The four data sets represent 4 different patters:

- Data set 1 has a linear relationship between X and Y
- Data set 2 plots a curved pattern
- Data set 3 has an outlier which skews the statistical properties for otherwise a straight line
- Data set 4 is a vertical line except one outlier

This highlights the potential pitfalls of relying solely on summary statistics without examining the data graphically.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearsson's R also known as correlation coefficient, explains the degree of association between two variables. The value of R ranges between -1 and +1.
A positive value of R indicates a positive correlation and a negative value indicates a negative correlation. Zero value of R indicates absolutely no correlation.

Higher the value of R, higher is the association of variables and lower value is associated with low correlation.

This statistic measures the association of two measurable variables, but the degree of association of the variables or the significance of association.

A high value of R does not necessarily mean that the variables are significant, and dependent on one another, change in one does not necessarily mean change in another variable.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling:

Scaling is a data pre-processing technique used in machine learning and statistical modeling to standardize the range of independent variables (features/predictors) in a dataset.
It ensures that the features have comparable magnitudes, which is important for algorithms that rely on distance metrics or gradient-based optimization.

In Linear regression model, which uses gradient-decent methodology to arrive at a model, standardization plays an important role in ensuring all the variables have comparable magnitude.

Why is scaling performed:
In absence of scaling, variables with higher numeric values can misguide the model by getting more gradient weightage.

Scaling also improves performance of the model. If all the features are on vastly different scales, the models can produce biased or suboptimal results due to uneven weightage given the the predictor variables.

In optimization-based algorithms, like linear regression scaled data can arrive at faster convergence and avoids oscillating between variable choices for model, which would have been the case if the variables are on different magnitude.

Scaling ensures that no feature dominates the model, just because it has higher weightage. It balances the effect of weights for all features.

Difference between Normalised Scaling and Standardized scaling:
Normalized Scaling: Normalization rescales the data to fit within a fixed range, usually [0, 1]. It transforms each value based on its minimum and maximum values.

Normalization uses the following formula:

$$X_{(norm)} = [X - \min(x)] / [\max(x) - \min(x)]$$

Normalized scaling is useful when features have no outliers, and the dataset has a uniform distribution.

Standardized Scaling: Standardization transforms the data to have a mean of 0 and a standard deviation of 1. It focuses on centering the data around 0 and scaling it based on the standard deviation.
Standardization uses the following formula:

$$X_{(std)} = X - \mu / \sigma$$

Where, μ is mean and σ is standard deviation
Standardization scaling is preferred when the dataset contains outliers or when features follow a Gaussian distribution

---

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>
The Variance Inflation Factor (VIF) measures the degree of multicollinearity in a given set of variables. If the value of VIF is infinite, it typically indicates perfect multicollinearity, which occurs when one or more independent variables are exact linear combinations of others.

An infinite VIF value must be resolved to ensure the stability and interpretability of the regression model. It usually indicates a redundant or improperly handled variable in the dataset.

VIF for a variable (X) is calculated using formula:

$$\mathbf{VIF = 1 / [1-R^2]}$$

Where R^2 (R-squared value) is the coefficient of determination obtained by regressing X on all other independent variables.

If there if perfect collinearity R-squared value becomes 1 and VIF value becomes infinite.
This indicates that the variable cannot be distinguished from others, making the regression unstable or even impossible.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>
A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, often the normal distribution. It plots the quantiles (data points at specific cumulative probabilities) of the observed data against the quantiles of the theoretical distribution.

If the data is normally distributed, the Q-Q plot closely follows a 45-degree straight line.
If the plot deviates from this line, then it indicates departure from normality in dataset

Q-Q plot helps in model validation by confirming whether the assumption of normally distributed residuals holds, which is critical for accurate p-values and confidence intervals.

The plot also helps in detecting model deviations and issues: if the plot curves away from the line, it indicates skewedness in the residual, heavy or long tails explaining kurtosis of the residuals can be explained by outliers on the plot.
If residuals are not normal, the Q-Q plot can indicate whether transformations (e.g., log, square root) are needed to normalize the data.
