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:

1. Seasonality plays a significant role in bike rentals. Summer and Fall (baseline) months showed the highest demand for bikes. Spring had a relatively lower demand, as indicated by its negative coefficient in the model. Winter showed reduced demand compared to fall but was higher than spring. This shows that Warmer months encourage more bike rentals, while colder months reduce demand.
2. Weather conditions greatly influence bike rentals too. Clear weather (the baseline category) led to the highest bike usage. Mist/Cloudy weather showed a moderate decline in rentals. Light Snow/Rain resulted in a significant drop in demand. Hence it is clear that Poor weather conditions discourage people from renting bikes.
3. The holiday variable had a minimal or negative impact on rentals. Fewer bikes were rented on holidays compared to regular days. This suggests that most users rent bikes for commuting rather than leisure, leading to decreased demand on holidays.
4. Both weekday and workingday had minor effects on bike rentals. There was no significant difference in rentals between weekdays and weekends. However, working days showed a slight increase in rentals, reinforcing the commuting pattern. Hence it can be concluded that Bike usage is relatively consistent throughout the week, but slightly higher on working days due to commuting.

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:

Using drop_first=True during dummy variable creation is essential to prevent multicollinearity in linear regression models. This technique addresses the dummy variable trap — a situation where one dummy variable can be perfectly predicted from the others, leading to perfect multicollinearity and causing issues in model estimation.

Dummy Variable Trap - When converting a categorical variable with k categories into dummy variables, you get k binary columns. Including all k dummies in a regression model along with the intercept leads to perfect multicollinearity because the sum of the dummy variables will always equal 1 (i.e., the intercept becomes redundant).

drop_first=True drops the first category during dummy encoding, reducing k dummies to k-1. This avoids perfect multicollinearity by making the dropped category the reference/baseline

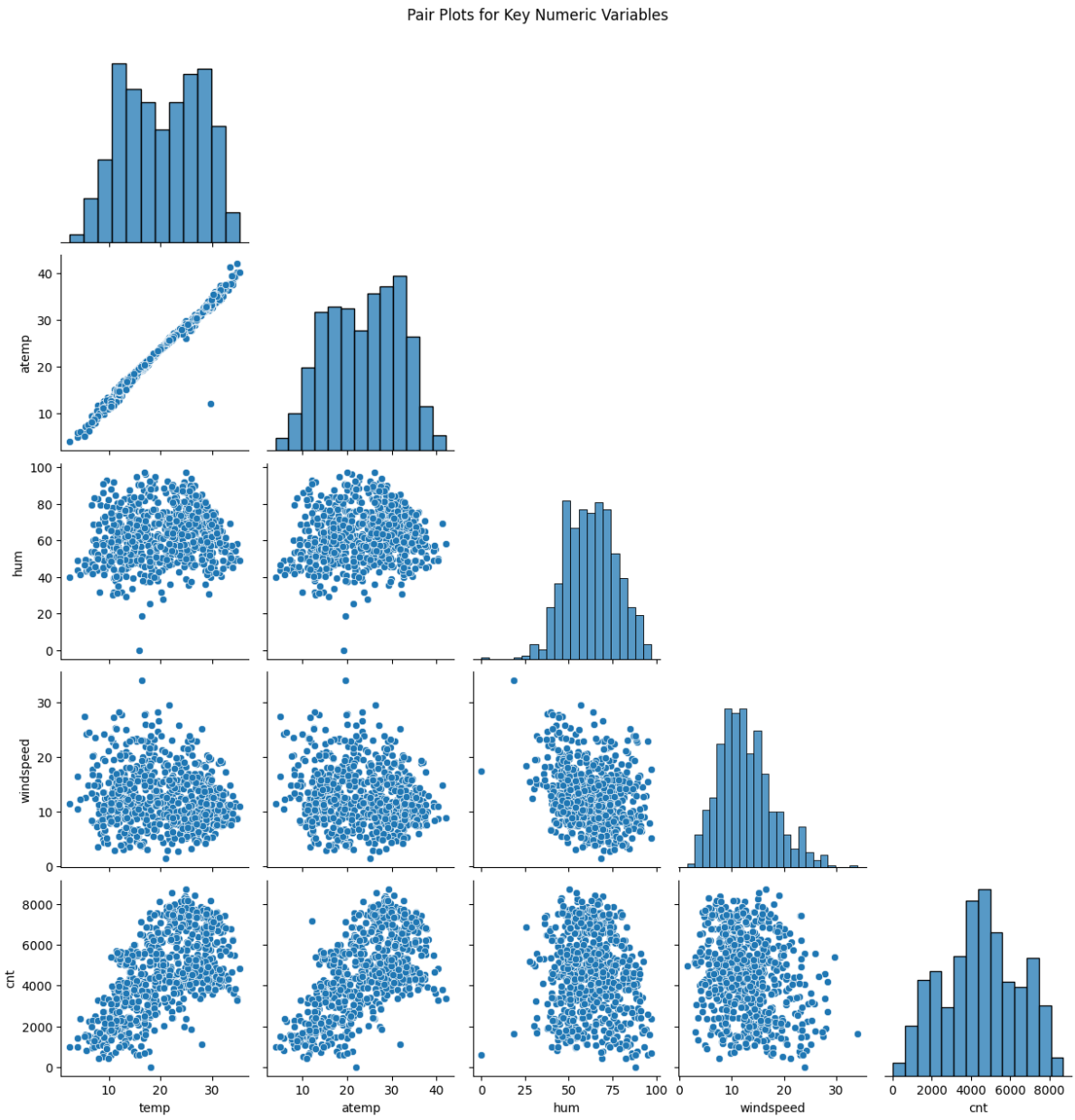
---

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:

Looking at the pair plot, the numerical variable that shows the highest correlation with the target variable (cnt) is temp (temperature).

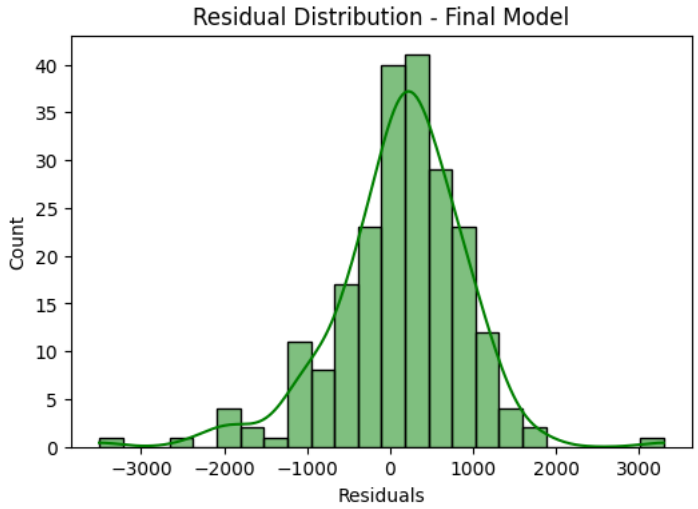


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

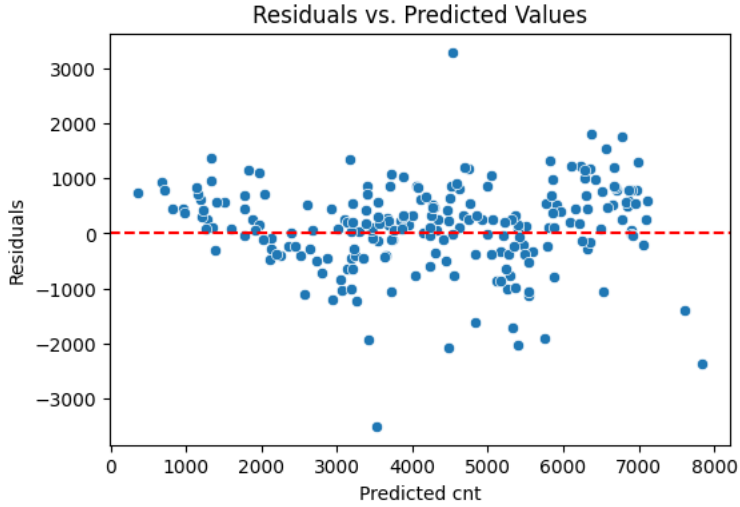
Total Marks: 3 marks (Do not edit)

Answer:

Plotted a residual distribution to check if Residuals are following a normal distribution to ensure valid hypothesis testing and confidence intervals.



Plotted residuals vs. predicted values to validate that the relationship between the independent variables and the dependent variable (cnt) should be linear.



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:
Based on the final linear regression model and the statistical significance of the predictors, the following are the top 3 features contributing significantly to explaining the demand for shared bikes (cnt):

- Season – Spring (season_spring), Coefficient: -1094.90, P-value: 0.000 (Highly significant)
- Weather Situation – Light Snow/Rain (weathersit_light_snow_rain), Coefficient: -1873.90, P-value: 0.000 (Highly significant)
- Year (yr) - Coefficient: +1975.22, P-value: 0.000 (Highly significant)

```
OLS Model Summary (Initial):
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:                    cnt   R-squared:
0.825
Model:                            OLS   Adj. R-squared:
0.820
Method:                 Least Squares   F-statistic:
166.5
Date:                Sat, 22 Feb 2025   Prob (F-statistic):
6.09e-177
Time:                        14:52:35   Log-Likelihood:
-4139.8
No. Observations:                 510   AIC:
8310.
Df Residuals:                     495   BIC:
8373.
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
=======================
                                 coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
-----------------------
const                        2236.7186    407.526      5.489      0.000
1436.024    3037.413
yr                           1975.2234     74.009     26.689      0.000
1829.812    2120.635
mnth                          -34.5323     19.771     -1.747      0.081
-73.377       4.312
holiday                      -473.7855    233.032     -2.033      0.043
-931.639     -15.932
weekday                        60.0138     18.159      3.305      0.001
24.337      95.691
workingday                    139.8062     80.561      1.735      0.083
-18.477     298.090
temp                           67.0049     65.749      1.019      0.309
-62.177     196.186
atemp                          51.4286     60.200      0.854      0.393
-66.850     169.708
hum                           -10.7343      3.566     -3.010      0.003
-17.741      -3.728
windspeed                     -35.8820      8.401     -4.271      0.000
-52.389     -19.375
season_spring               -1094.8999    205.883     -5.318      0.000
-1499.413    -690.387
season_summer                 183.8390    132.010      1.393      0.164
-75.530     443.208
season_winter                 800.9673    156.847      5.107      0.000
492.799    1109.135
weathersit_light_snow_rain -1873.9015    246.112     -7.614      0.000
-2357.455   -1390.348
weathersit_mist_cloudy      -468.6400     96.165     -4.873      0.000
-657.582    -279.697
==============================================================================
=======
```

```
Omnibus:                       68.977   Durbin-Watson:
2.024
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
153.972
Skew:                          -0.731   Prob(JB):
3.68e-34
Kurtosis:                       5.260   Cond. No.
897.
==============================================================================
=======

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer:

Linear Regression is one of the most fundamental and widely used algorithms in statistics and machine learning for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors). It aims to find the best-fitting linear relationship between variables.

Linear Regression predicts a continuous outcome by modeling a linear relationship between the dependent variable (Y) and independent variable(s) (X). Two main types of Linear Regression:

- Simple Linear Regression: Involves one independent variable.

  $Y=\beta0+\beta1X+\epsilon$ where

  - Y → Dependent variable (target)
  - X → Independent variable (predictor)
  - β_0 → Intercept (value of Y when X = 0)
  - β_1 → Slope coefficient (how much Y changes for a unit change in X)
  - ε → Error term (accounts for variability not explained by the model)

- Multiple Linear Regression: Involves two or more independent variables.

  $Y=\beta0+\beta1X1+\beta2X2+\cdots+\beta nXn+\epsilon$

  - $X_1, X_2, ..., X_n$ → Multiple predictors
  - $\beta_1, \beta_2, ..., \beta_n$ → Coefficients representing the impact of each predictor

The goal of Linear regression is to find the best-fitting line (or hyperplane) that minimizes the difference between the actual and predicted values. This is done by minimizing the Residual Sum of Squares (RSS).

Residual Sum of Squares (RSS):

RSS=∑i=1n(Yi−Y^i)2RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2RSS=i=1∑n(Yi−Y^i)2

- $Y_i \rightarrow$ Actual value
- $\hat{Y}_i \rightarrow$ Predicted value
- $n \rightarrow$ Number of data points

The process used to minimize RSS is called Ordinary Least Squares (OLS). OLS is the most common technique for estimating the coefficients (β) in linear regression.

For linear regression to produce valid and reliable results, the following assumptions must hold:

- Linearity: The relationship between independent variables and the target is linear.
- Independence of Errors: Residuals (errors) are independent of each other (no autocorrelation).
- Homoscedasticity: The residuals have constant variance at every level of X.
- Normality of Residuals: Residuals are normally distributed.
- No Multicollinearity: Independent variables are not highly correlated with each other.

To assess the goodness of fit, the following metrics are commonly used:

- R-squared (R^2):
    - Measures how well the model explains the variance in the dependent variable.
    - Ranges from 0 to 1, where higher values indicate a better fit.

      R2=1−(RSS/TSS)

      (TSS = Total Sum of Squares)

- Root Mean Squared Error (RMSE): Measures the average magnitude of prediction errors.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

Anscombe's Quartet is a set of four datasets created by statistician Francis Anscombe to highlight the importance of data visualization in statistical analysis. Though all four

datasets share nearly identical statistical properties — including the same mean, variance, correlation coefficient, and linear regression line — they exhibit vastly different patterns when plotted. One dataset follows a simple linear relationship, another shows a clear non-linear trend, the third has a linear pattern distorted by an outlier, and the fourth consists mostly of identical points with a single outlier affecting the regression line. The key lesson from Anscombe's Quartet is that relying solely on summary statistics can be misleading, and visualizing data is crucial to uncover underlying patterns, outliers, and relationships. It underscores the importance of combining both numerical and graphical analyses to ensure accurate interpretations and insights in data analysis.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It evaluates both the strength and direction of the association, with values ranging from -1 to +1. A value of +1 indicates a perfect positive linear correlation, meaning as one variable increases, the other also increases proportionally. A value of -1 represents a perfect negative linear correlation, where one variable increases as the other decreases. A value of 0 implies no linear relationship between the variables.

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:

Scaling is a data preprocessing technique used to adjust the range or distribution of numerical features, ensuring that they are on a comparable scale. It is crucial in machine learning algorithms that rely on distance measurements or assume certain data distributions, such as linear regression, k-nearest neighbors, and support vector machines.
Scaling is performed to improve model performance and convergence. Features with vastly different ranges can disproportionately influence the model, leading to biased results. For example, a feature ranging from 1 to 1000 can overshadow another feature ranging from 0 to 1, skewing the learning process. Scaling ensures that all features contribute equally to the model.
There are two common types of scaling:

- Normalized Scaling (Min-Max Scaling): This technique rescales data to a fixed range, usually between 0 and 1. Normalization is useful when the data distribution does not follow a Gaussian (normal) distribution and is beneficial for algorithms that require bounded input.

- Standardized Scaling (Z-score Normalization): This method transforms data to have a mean of 0 and a standard deviation of 1.

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:

The Variance Inflation Factor (VIF) measures the degree of multicollinearity among independent variables in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to the correlation with other predictors. A VIF value becomes infinite (or extremely large) when there is perfect multicollinearity — meaning one predictor variable is a perfect linear combination of one or more other predictors.
This situation typically occurs when:

1. Duplicate or Redundant Features: If two or more variables carry identical or nearly identical information (e.g., including both "age in years" and "age in months"), they will be perfectly correlated, leading to an infinite VIF.
2. Dummy Variable Trap: When creating dummy variables for categorical data without dropping one category (i.e., not using drop_first=True), the sum of the dummy variables becomes perfectly collinear with the intercept, causing infinite VIF.
3. Mathematical Dependencies: If the model includes variables that have a deterministic relationship (e.g., total sales and the sum of product category sales), it leads to perfect multicollinearity.

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:

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset against a theoretical distribution, most commonly the normal distribution. It helps assess whether a set of data follows a particular distribution by plotting the quantiles of the data against the quantiles of the theoretical distribution.
In a Q-Q plot, if the data follows the theoretical distribution, the points will lie approximately along a 45-degree reference line. Deviations from this line indicate departures from the expected distribution, such as skewness, kurtosis, or the presence of outliers.

In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. This is crucial for the validity of statistical tests (like t-tests for coefficients) and confidence intervals.
A Q-Q plot is used to:

---

1. Check Normality of Residuals: By plotting the residuals against a normal distribution, the Q-Q plot reveals whether the residuals follow a normal pattern.
    - If the points lie along the line, the residuals are approximately normal.
    - Systematic deviations suggest issues like skewness or heavy tails.
2. Identify Outliers: Extreme deviations from the line can indicate outliers that may unduly influence the regression model.
3. Validate Model Assumptions: Ensuring normally distributed residuals supports the reliability of hypothesis tests and the overall validity of the linear regression model.