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 final Multiple Linear regression model contains many predictor variables that are categorical in nature and some of them have been encoded to dummy variables

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2059      0.028      7.423      0.000       0.151       0.260
yr             0.2341      0.008     27.985      0.000       0.218       0.251
workingday     0.0219      0.009      2.460      0.014       0.004       0.039
temp           0.4495      0.032     13.968      0.000       0.386       0.513
windspeed     -0.1487      0.026     -5.775      0.000      -0.199      -0.098
spring        -0.0928      0.018     -5.047      0.000      -0.129      -0.057
weathersit_2  -0.0799      0.009     -8.988      0.000      -0.097      -0.062
weathersit_3  -0.2878      0.025    -11.495      0.000      -0.337      -0.239
month_3        0.0517      0.015      3.440      0.001       0.022       0.081
month_4        0.0430      0.019      2.264      0.024       0.006       0.080
month_5        0.0542      0.017      3.258      0.001       0.022       0.087
month_9        0.0800      0.016      5.001      0.000       0.049       0.111
winter         0.0771      0.015      5.149      0.000       0.048       0.107
==============================================================================
Omnibus:                       68.165   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              181.841
Skew:                          -0.664   Prob(JB):                     3.26e-40
Kurtosis:                       5.607   Cond. No.                         17.8
==============================================================================
```

spring, winter falls under season category and have been dummy encoded. weathersit_2 and weathersit_3 falls under weathersit category and have been dummy encoded. Similarly, months variables fall under mnth category and have been dummy encoded. We can infer from above image that these variables are statistically significant and explain the variance in model very well.

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)

Using drop_first=True during dummy variable creation is important to avoid the dummy variable trap. The dummy variable trap occurs when you include all dummy variables for a categorical feature, leading to multicollinearity. This is because the sum of all dummy variables for a categorical feature will always be equal to 1 (for a given observation). This creates a situation where one dummy variable can be perfectly predicted by the others, causing redundancy in the model.

By setting drop_first=True, you drop one of the dummy variables, which effectively removes this redundancy. The dropped variable becomes the and the other dummy variables represent how the

feature differs from this reference category. This prevents multicollinearity and ensures that the model can estimate the effect of each category relative to the reference one.

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)

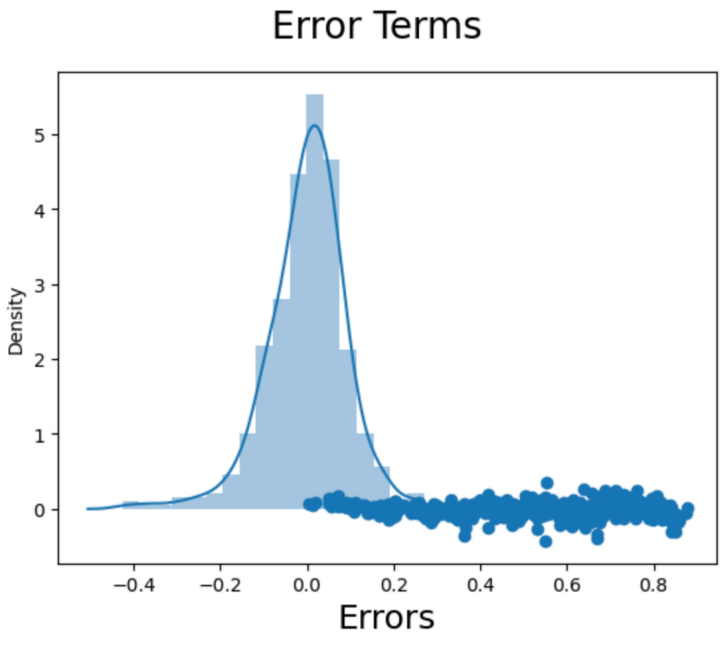
So, before model building and training, the pair plot shows highest correlation for registered variable having correlation 0.945. But we are not using casual and registered in our pre-processed training data for model training. casual + registered = cnt. This might leak out the crucial information and model might get overfit.

So, excluding these two variables atemp is having highest correlation with target variable cnt which is followed by temp. As per the correlation heatmap, correlation coefficient between atemp and cnt is 0.631. And correlation coefficient between temp and cnt is 0.627.

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)

To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

• ***Residual Analysis:***
We need to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression). I have plotted the histogram of the error terms and this is what it look like:



residuals are following the normally distribution with a mean 0. All good!

• ***Linear relationship between predictor variables and target variable:***
This is happening because all the predictor variables are statistically significant (p-values are less than 0.05). Also, R-Squared value on training set is 0.832 and adjusted R-Squared value on training set is 0.828. This means that variance in data is being explained by all these predictor variables.

• ***Error terms are independent of each other:*** *Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.*

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)

Top 3 features significantly contributing towards demand of shared bikes are:
1) ***temp*** *(coef:* ***0.4495****)*
2) ***yr*** *(coef:* ***0.2341****)*
3) ***month_9*** *(coef:* ***0.0800****)*

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)

Linear regression is a statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors) by fitting a linear equation to observed data.

1. Simple Linear Regression:

In simple linear regression, we model the relationship between a single independent variable (X) and a dependent variable (Y) using a linear equation of the form:

$$Y=mX+b$$

- *Y is the dependent variable (target).*
- *X is the independent variable (predictor).*
- *m is the slope of the line (coefficient of X).*
- *b is the y-intercept (constant).*

2. Multiple Linear Regression:

In multiple linear regression, we extend the idea to handle multiple independent variables. The equation becomes:

$$Y=b+m1\ X1\ +m2\ X2\ +\cdots+mn\ Xn$$

- *X1 ,X2 ,…,Xn are the independent variables.*
- *m1 ,m2 ,…,mn are the coefficients for the respective predictors.*

The algorithm will compute the best values for these coefficients that minimize the difference between the predicted and actual Y values.

3. Objective:

The primary objective is to minimize the cost function or loss function (commonly the Mean Squared Error or MSE) which measures how far the model's predictions are from the actual values. The formula for MSE is:

$$MSE=n1\ i=1\sum n\ (yi\ -y\hat{}\ i\ )2$$

Where:
- *yi is the actual value.*
- *y^ i is the predicted value.*
- *n is the number of data points.*

4. Finding the Best-Fitting Line:

Linear regression uses optimization techniques such as ***Gradient Descent*** *or* ***Ordinary Least Squares (OLS)*** *to minimize the cost function and find the optimal values of the coefficients m (for simple) or m1 ,m2 ,…,mn (for multiple).*

5. Assumptions in Linear Regression:

- *The relationship between the dependent and independent variables is linear.*
- *The residuals (errors) are normally distributed.*
- *The variance of the errors is constant (homoscedasticity).*
- *There is no multicollinearity in the predictors (in case of multiple linear regression).*

6. Applications: *Linear regression is widely used in various fields such as economics (predicting costs), healthcare (predicting patient outcomes), and engineering (modeling physical systems).*

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)

Anscombe's Quartet consists of four datasets that have nearly identical simple descriptive statistics (such as mean, variance, and correlation) but appear very different when graphed.

The Four Datasets: *Each dataset in Anscombe's Quartet consists of 11 data points, and for each, the following statistics are the same:*

- *Mean of X: 9*
- *Mean of Y: 7.5*
- *Variance of X: 11*

- *Variance of Y: 4.12*
- *Correlation between X and Y: 0.82*
- *Regression line (Y = aX + b) has the same slope and intercept for all four datasets.*

1. Dataset 1:

This dataset represents a standard linear relationship where the data points roughly follow a straight line. The relationship between X and Y is clear, and a linear regression model fits well.

2. Dataset 2:

In this dataset, most points follow a linear trend similar to Dataset 1, but one point is an outlier with a much higher Y-value. Despite the outlier, the linear regression line still fits well, but the visual representation shows the influence of the outlier.

3. Dataset 3:

This dataset shows a perfect quadratic relationship. The data points follow a curve rather than a straight line. The linear regression model would fail to capture the non-linear pattern, but the summary statistics are similar to the other datasets, which can be misleading without a plot.

4. Dataset 4:

In this dataset, most data points are clustered in a tight horizontal line with a single outlier point. The linear regression line appears almost flat, but the presence of the outlier distorts the analysis. Like the other datasets, the summary statistics are similar to the others.

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)

*Pearson's R, also known as the **Pearson correlation coefficient**, is a measure of the strength and direction of the linear relationship between two variables. It is denoted by **r** and ranges from **-1 to 1**, where:*

- ***r = 1**: Perfect positive correlation, meaning as one variable increases, the other variable increases in a perfectly linear manner.*
- ***r = -1**: Perfect negative correlation, meaning as one variable increases, the other decreases in a perfectly linear manner.*
- ***r = 0**: No linear correlation, meaning there is no linear relationship between the two variables.*
- ***r > 0**: Positive correlation, meaning that as one variable increases, the other tends to also increase.*
- ***r < 0**: Negative correlation, meaning that as one variable increases, the other tends to decrease.*

The formula for Pearson's R is:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

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)

Scaling refers to the process of transforming the features of a dataset so that they are on a similar scale. This is done to ensure that all features contribute equally to the model, particularly for algorithms that are sensitive to the magnitude of the features, such as ***K-nearest neighbors (KNN)****,* ***support vector machines (SVM)****, and* ***gradient descent-based algorithms*** *like* ***linear regression*** *and* ***neural networks****.*

Why Scaling is Performed:

1. ***Improve model performance:*** *Scaling ensures that no single feature dominates the model due to differences in units or magnitudes.*
2. ***Faster convergence in optimization algorithms:*** *In algorithms like gradient descent, if features have very different scales, the model might converge slowly or even fail to converge. Scaling helps speed up convergence.*
3. ***Ensure equal weighting:*** *In machine learning models, features with larger ranges can disproportionately influence the model if they are not scaled.*

Difference between Normalized Scaling and Standardized Scaling:

Aspect	Normalized Scaling (Min-Max Scaling)	Standardized Scaling (Z-score Scaling)
Definition	Scales data to a fixed range, usually between 0 and 1.	Scales data to have a mean of 0 and a standard deviation of 1.
Output Range	[0, 1] or any specific range.	No specific range (mean = 0, standard deviation = 1).
When to Use	When data needs to be within a specific range (e.g., neural networks, KNN).	When data needs to be zero-centered with unit variance (e.g., regression, PCA).
Example Use Cases	Neural networks, KNN, when features have different units.	Linear regression, logistic regression, PCA, when normality assumption holds.

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)

*VIF becomes **infinite** when one of the **predictor variables** is perfectly correlated with one or more of the other predictor variables in the model. This situation occurs when there is **perfect multicollinearity** (i.e., when a predictor variable can be exactly predicted by a linear combination of other predictors).*

*Where R2 is the coefficient of determination from regressing the predictor variable on the other predictors. If **R2=1**, meaning there is perfect correlation, the denominator becomes **0**, resulting in an **infinite VIF**.*

Cause of Infinite VIF:

- ***Perfect Collinearity:** If two or more predictors are linearly dependent (e.g., one variable is a constant multiple or a direct combination of others), the regression model cannot distinguish their individual effects, leading to infinite variance inflation.*

- ***Linear Dependence:** In practical terms, this often happens if one of the predictor variables is derived from others in the dataset or there's redundancy among the predictors.*

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)

*A **Q-Q plot** is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. The plot compares the quantiles of the sample data to the quantiles of a specified reference distribution (usually normal). If the data is from the reference distribution, the points in the Q-Q plot will lie approximately along a straight line.*

How a Q-Q Plot Works:

- ***X-axis:** Represents the quantiles of the reference distribution (e.g., normal distribution).*
- ***Y-axis:** Represents the quantiles of the sample data.*
- ***Interpretation:** If the data follows the reference distribution, the points will lie approximately on a straight line. Deviations from this line suggest differences between the observed data and the expected distribution.*

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

*In the context of **linear regression**, the Q-Q plot is used to check whether the **residuals** (the differences between observed and predicted values) follow a normal distribution, which is one of the key assumptions of linear regression.*

Key Points:

1. ***Normality of Residuals:***
    - *A linear regression model assumes that the residuals are normally distributed. The Q-Q plot helps visually assess this assumption. If the residuals are normally distributed, the points in the Q-Q plot should lie along a straight line. If they deviate significantly from this line, it may indicate that the normality assumption is violated.*
2. ***Detection of Non-Normality:***
    - *If the residuals are skewed or have heavy tails, the Q-Q plot will show deviations from the straight line. This suggests that the model might not be appropriate or that some transformations of the data are required (e.g., logarithmic transformation for skewed data).*
3. ***Model Diagnostics:***
    - *A Q-Q plot is a useful diagnostic tool to assess the appropriateness of the linear regression model. Significant deviations from normality can indicate issues like* ***heteroscedasticity*** *(non-constant variance of residuals),* ***outliers****, or other violations of linear regression assumptions.*
