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)

Categorical variables do have an influence on the target variable.
E.g. The **yr 2019** has had higher counts of usage.
The **fall and summer season** shows higher usage.
The months from **June to Sept** show higher usage, which correspond to the season inferences.
The **weathersit** being **clear** increases the usage count.

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 when created using the **pd.get_dummies()** function creates **N** columns for each of the **N** categories.
We only need **N-1** columns to avoid multicollinearity. If there are N columns, they will be linearly dependent. One column can be fully predicted by the other.
This causes issues in model building as we want the variables to be as independent of each other as possible.

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)

Looking that the pair-plot, **temp** has the highest correlation with the target variable.

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)

1. Checked the residual distribution using distplot in seaborn. Calculate the residuals by subtracting the Actual training set target value from Predicted target value of the training set from the model. (**I got a normal curve centered at 0**).
2. Check the R-Squared and the Adjusted R-Squared value. R-Squared and Adjusted R-squared must be within around 1 – 5% of each other. (**The difference is 0.4% for my model**).
3. Construct a regression plot (regplot) between the **predicted values** (**Y_train_pred**) and the actual values (**Y_train**) to see if it is a straight line. (**It is a straight line across the regression plot**)

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)

The top three features are:

Year (coefficient = 0.234)
Temperature (coefficient = 0.472)
Light Snow / Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (coefficient = -0.289)

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>

The linear regression algorithm is mathematically represented as an equation of a line

Y = ßo + X1xß1 + X2xß2 +…+ XnßnX + ε

ßo -> Contant (Intercept)

ß1 …. ßn -> Coefficients associated with the features/ independent variables (X1 … Xn) where there are N features.

ε -> Error term (variations)

It is called simple linear regression when 1 feature (Independent variable) is involved which is the equation of a simple line: Y = ßo + X1xß1 + ε corresponding to y = mx + c + ε.

Multiple linear regression is when multiple features are involved which explain the trend of the target Y variable.

In essence it predics the outcome based on the values of input features or variables and it assumes a linear relationship between the input variables and the target variable.

It has 5 assumptions:

It assumes a linear relationship between X and Y (Input and target variable)
It assumes that the input variables are independent of each other.
Homoscedasticity: The variance in the error is constant
Residuals (Predicted value – Actual value) is normally distributed.
Independent variables are not highly correlated. (No multicollinearity)

Models basically find the best fit line that minimizes the RSS (Residual sum of Squares) and the RMSE (Root mean square error)

How to evaluate Linear regression models

RSS = sum (square (Y - Ypred)) → Residual sum of Squares (Should be as low as possible)
TSS = sum (square (Y – mean(Y))) → Total sum of squares
MSE = (1/n) x sum (square (Y - Ypred)) → Mean square error (Should be as low as possible)
RMSE = sqrt (MSE) → RMSE (Root mean square error)
R-squared = 1 – (RSS/TSS) → Higher it is the better (Ranges from 0 to 1)
[But not too high else it will overfit the data]
Adjusted R-Squared = 1 – ((1 - R-squared)/ (1 – p – N)) → Takes into influence the number of features and the number of records.

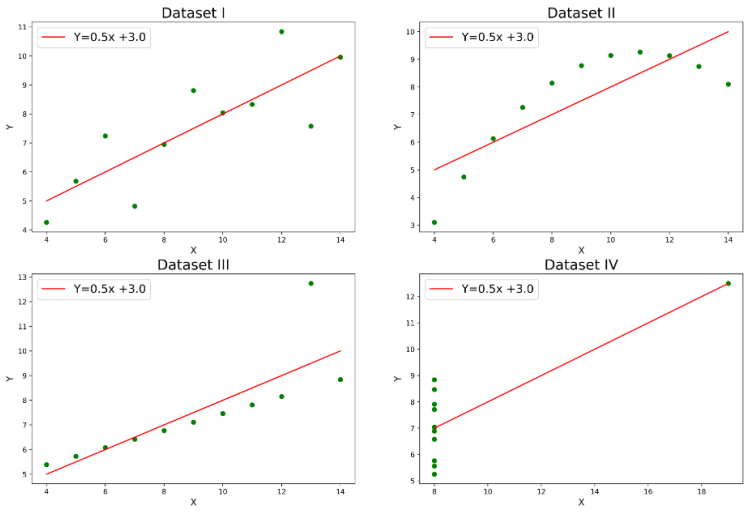
P -> Number of features
N -> Number of records
Ypred -> Predicted value of the target variable

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 comprises of 4 datasets with identical / nearly identical summary statistics, but with different representations on the scatter plot.

1. In the first dataset we see a linear relationship between x and y.
2. In the second dataset we see a nonlinear relationship between x and y.
3. In the third dataset we see a linear relationship with outliers.
4. In the fourth dataset we see that the data is showing no linear relationship but one point can misleadingly show a linear relationship.



This emphasizes the limitations of analyzing datasets solely on the summary statistics and not using Exploratory data analysis to check trends and spot outliers.

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>
Pearson's R or Pearson's correlation coefficient is a way to measure linear correlation. The values range from -1 to +1.
Where -1 shows perfect negative correlation and +1 shows perfect positive correlation. 0 shows no correlation.
Values between 0 to 1 show positive correlation e.g. Higher the Area of a house, higher the price.
0 Value means no correlation at all e.g. The number of tile squares in a house and the price of the house.
Values between -1 to 0 indicate negative correlation. E.g. Higher the altitude, lower is the atmospheric pressure and temperature.

Values between 0 and 0.3 in both directions (Negative and Positive) show weak correlation.
Values between 0.3 and 0.5 in both directions (Negative and Positive) show medium correlation
Values above 0.5 in both directions (Negative and Positive) show strong correlation

---

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 is used to standardize data features to make the dataset more comparable with each other. i.e. It brings all the variables / feature to one scale
(Between -1 and +1 or between -σ and +σ with the mean in between where σ is the standard deviation).
This is performed so that the one variable with a higher range of values does not dominate the analysis.

Normalized scaling : The scales range from [-1 to +1]. It is calculated by

$$X_{new} = (X - X_{min})/(X_{max} - X_{min})$$

X_{max}: Max value of X in the column,
X_{min}: Minimum value of X in the column.

Standardization : The scales range from [-σ to +σ]

Xnew = (X - Xmean)/stddev

Xmean: Mean of the rows in the column

Stddev: Standard deviation of the column

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 formula for VIF is 1/ (1 – R-squared). This means that the R-squared of the column would need to be 1.0.

R-squared of the variable will be 1.0 if the data is overfitting on the variable or the variable is well explained by the rest of the features. This shows high multicollinearity for the variable and hence the variable needs to be eliminated.

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>

Quantile Quantile Plot (Q-Q Plot) is a graphical way of determining if a dataset follows a certain probability distribution or whether two samples come from the same population or not. Used in quality control to check assumptions and identify departures from expected values.

Quantiles are of different types:

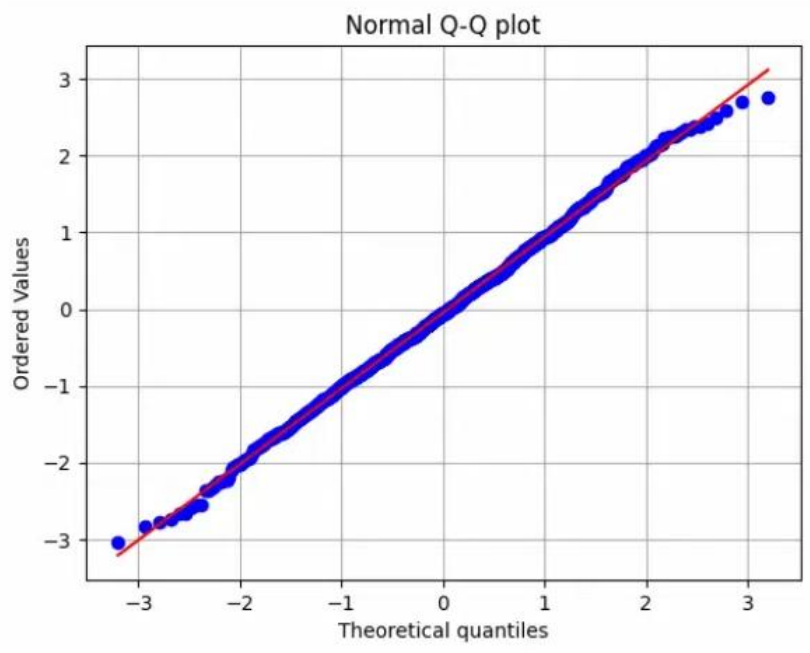
Median (50^{th} Percentile): Middle value dividing the dataset into 2 parts

Quartiles (25^{th} 50^{th} and the 75^{th} Percentiles): Divides the dataset into 4 parts or quartiles.

Percentiles: Divide the dataset into 100 parts.

To make a Q-Q plot we first

1. Get the data
2. Sort the data
3. Chose the distribution to compare against (e.g. Normal distribution)
4. Calculate the theoretical quantiles for the distribution (For normal distribution we would use the CDF (Cumulative distribution function) to find the expected quantities).
5. Plot the theoretical values on the Y axis and the sorted values on the X axis
6. Connect the data points to visually inspect the relationship.



In this case it follows a straight line, hence the assumption of a normal distribution is correct.
