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 categorical variables like

- season
- weathersit
- Mnth
- Holiday

The above variables have high significance towards the target variable. A point to be noted,

- weekday variable has no significant difference between their sub-groups and target variable.
- Workingday and holiday is correlated – we should only either of these two

---

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)

For categorical variables having more than two levels, we need to generate dummy variables for model building.

For categorical variables with levels (m>2); when m=4 then the number of dummy variables created should be m-1=3. Hence drop_first=True is important

This is done to remove perfectly multi-collinear variables among independent variables

---

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)

- The highest correlated variables were casual and registered with 67% and 95% respectively. However, these two variables are out of the input features concern. Therefore, the next highest correlated variables with cnt are temp and atemp with 63%.
- Temp and atemp are highly correlated among themselves – multi-collinear

---

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)
Checked for the following,

- Less or no multi-collinearity and statistical significance of variables towards the fit of the model
- Residual Analysis – homoscedasticity (equal variance in the error term)
- Residual Analysis – Error term following standard normal distribution with 0 and 1

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)
Positively contributing features towards the demand of the shared bikes:

- Temperature
- Year
- Winter season

Highest +ve contributing features

	index	features	coefficients
0	14	temp	0.416215
1	5	yr	0.239976
2	2	season_winter	0.112853
3	12	mnth_September	0.051725
4	1	season_summer	0.027800

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 finds the optimal gradient of predictors to predict the output variables with minimal error. LR is a method of fitting a straight through the linearly correlated variables. It is a supervised learning algorithm used for predicting continuous variable

The target variable and independent variables should have linear realationship

Simple linear Regression: Let's consider we have single independent variable (X) and a target variable (Y)
Y=C+mX +e

C is the intercept, when there's x=0 (no feature value)
X is the feature that explains Y
e is the error term
Multiple Linear Regression: The regression with multiple predictors (1,2..n)
Y=b0+b1x1+b2x2+..+bnxn +e
Y is the dependent/ target variable/ predicted variable
b0 is the intercept
b1, b2 ...bn are the gradient/coefficients of features/independent variables or predictors x1,x2...,xn
e is the error term following standard normal distribution

The predicted values will be compared with the actual values to evaluate the fit of the model. The best model equation is the one that has the minimal errors (difference of actual values and predicted values)

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 is a set of four datasets that appear nearly identical when analyzed using basic statistical methods but are strikingly different when visualized graphically

Similar dataset, exhibiting vastly different patterns when plotted. For an example, it is good example for showing that summary statistics are misleading quite often.

Summary statistics provides an overview of data. But cannot reveal relationships and patterns. This stays as the motivation to EDA.

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>

- It's an approach to measure the relationship between two continuous variables.
- It ranges from –1 to +1.
- +1 indicates, positively correlated variables. -ve indicates, negatively correlated variables. 0, there's no linear relationship between two 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: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is performed to keep the consistency of independent variables to make a reasonable comparison during the model evaluation and selection part.

Why scaling performed?

- Improved comparability
- Smooth convergence during optimization methods

Normalization is rescaling the features to fixed range of [0,1]. Formula: (x-xmin)/(xmax-xmin)

Standardization is centering the features around the mean 0 and scales them by unit variance. Formula: (x-mean)/std.error

---

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>

VIF(Xi)=1/1−Ri2 = 1/1-1 = inf

Inf signifies perfect multicollinearity among the independent varaiables.

Those variables with inf are advised to remove. The usual threshold is >5 or >10(sometimes). Variables with VIF are a threat to model performance.

High p-value +high VIF – Remove those variables

Low p-value +high VIP – Keep but treat them with regularization method like ridhe, lasso and elasticnet

Low p-value +low VIF – Keep them

High p-value + low VIF – If it did not have any compounding effect on other variables, remove those variables.

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 (commonly the normal distribution). It helps determine whether the data follows the assumed distribution by plotting the quantiles of the dataset against the quantiles of the theoretical distribution.

The importance of Q-Q plot:

To assess normality of residuals
To identify outliers
Validate model fit

Overall, for improved decision making.
