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 values do impact target variables. They need to be thoroughly analyzed via dummy variable method. In my analysis 7 independent variables are being used and 4 are categorical variables, approx. 57% variables are categorical variables. So, they have big impact.

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)

It reduces one unnecessary variable. If M levels are output of categorical variable then M-1 suffice.

---

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)

if one of categorical variable level is not dropped then VIF calculation doesn't gives result. Its value is infinite (INF)

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)

Two ways:

1. Residual analysis: distribution should be centered around zero and it should be normal
2. R2 comparison : R2 of training data and R2 of test data (using prediction model) should be very similar.

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)

1. Apparent temperature
2. Working day (not weekened)
3. Weather should be clear

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 type of machine-learning algorithm more specifically a** supervised machine-learning algorithm **that learns from the labelled datasets and maps the data points to the most optimized linear functions, which can be used for prediction on new datasets.**
It computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.

Important assumptions of linear regression:

- Linear relationship between the feature and target:
- Small or no multicollinearity between the features
- Homoscedasticity: error term is the same for all the values of independent variables.
- Normal distribution of the error terms
- No autocorrelations: The linear regression model assumes no autocorrelation in error terms.

Linear Regression equation:
y(x) = p0 + p1 * x
where,
y = output variable. Variable y represents the continuous value that the model tries to predict.
X = x is the feature, while it is termed the independent variable in statistics. Variable x represents the input information provided to the model at any given time.
p0 = y-axis intercept
p1 = the regression coefficient

Multiple Linear regression equation:
y(x) = p0 + p1x1 + p2x2 + … + p(n)x(n)
The equation for multiple linear regression is similar to the equation for a simple linear equation, i.e., y(x) = p0 + p1x1 plus the additional weights and inputs for the different features which are represented by p(n)x(n).

Cost function for Linear Regression
The cost function is the difference between the predicted value (Y') and true value (Y)
In Linear Regression, the **Mean Squared Error (MSE)** cost function is employed, which calculates the average of the squared errors between the predicted values and the actual value. The purpose is to determine the optimal values for the intercept and the coefficient of the input feature providing the best-fit line for the give data points.
MSE can be calculated as:

Cost function(*J*)=*n*1∑*ni*(*yi*^−*yi*)2

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 comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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)

Correlation coefficients are used to measure how strong a relationship is between two variables. There are different types of formulas to get a correlation coefficient, one of the most popular is Pearson's correlation (also known as Pearson's r) which is commonly used for linear regression.

The Pearson correlation coefficient, often symbolized as (r), is a widely used metric for assessing linear relationships between two variables. It yields a value ranging from –1 to 1, indicating both the magnitude and direction of the correlation. A change in one variable is mirrored by a corresponding change in the other variable in the same direction.
Pearson's correlation helps in measuring the correlation strength (it's given by coefficient r-value between -1 and +1) and the existence (given by p-value) of a linear correlation relationship between the two variables and if the outcome is significant we conclude that the correlation exists.

Pearson Correlation Coefficient (*r*) Range	**Type of Correlation**	**Description of Relationship**	**New Illustrative Example**
$0 < r \leq 1$	Positive	An increase in one variable associates with an increase in the other.	**Study Time vs. Test Scores:** More hours spent studying tends to lead to higher test scores.
$r = 0$	None	No discernible relationship between	**Shoe Size vs. Reading Skill:** A person's shoe size doesn't predict their ability to read.

Pearson Correlation Coefficient (*r*) Range	Type of Correlation	Description of Relationship	New Illustrative Example
		the changes in both variables.	
-1 ≤ *r* < 0	Negative	An increase in one variable associates with a decrease in the other.	**Outdoor Temperature vs. Home Heating Cost:** As the outdoor temperature decreases, heating costs in the home increase.

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)

Feature scaling is one of the most important data preprocessing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled.

Feature scaling helps machine learning, and deep learning algorithms train and converge faster.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. This scales the range to [0, 1] or sometimes [-1, 1]. Geometrically speaking, transformation squishes the n-dimensional data into an n-dimensional unit hypercube. **Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes but the age is close to uniform.**

```
X_new = (X - X_min)/(X_max - X_min)
```

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score. Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively. We can see that we are just changing mean and standard deviation to a standard normal distribution which is still normal thus the shape of the distribution is not affected.

Standardization does not get affected by outliers because there is no predefined range of transformed features.

```
X_new = (X - mean)/Std
```

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)

Variance Inflation Factor (VIF) is used for detecting multicollinearity in regression models. It measures how much the variance of a regression coefficient is inflated due to multicollinearity with other independent variables in the model.

You can get inf values for VIF due to the **perfect multicollinearity**. This happens when two or more independent variables in a model are perfectly linearly dependent. That is, one independent variable in the model can be entirely predicted by another independent variable.

If you have multiple identical columns in the input dataset, there will be perfect multicollinearity.

In addition, high correlation (correlation coefficients close to 1 or -1) between the independent variables can also give very high VIF values that could lead to inf values.

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)

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.
Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s).
Importance of Q-Q plots are:

- Check for normality: Q-Q plots are used to check if the dependent variable and residuals from a linear regression model are normally distributed. This is important because normality is an assumption for many parametric tests and confidence intervals.
- Check for constant variance: Q-Q plots can be used to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model.
- Compare data sets: Q-Q plots can be used to compare two data sets to see if they come from populations with the same distribution. This is useful when training and test data sets are received separately.
