Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   Ans:
   Effect of Categorical Variables on dependent variable(*cnt*)—

   - **season -** 'cnt' increases from spring to summer with further increase in fall and slight decrease in winter which is obviously due to the weather conditions in the seasons.
     Spring has the lowest median count and a smaller IQR, indicating lower and less variable booking.
     Summer and **fall** have higher medians and larger IQRs, suggesting higher and more variable booking.
     Winter shows a lower median than summer and fall but higher than spring, with a slightly smaller IQR

   - **yr -** there is increase in cnt from 2018 to 2019 indicating that bookings were higher and more variable in 2019 (as suggested in the problem statement - 'bike-sharing systems are slowly gaining popularity, the demand for these bikes is increasing').

   - **mnth:** There is a noticeable increase in 'cnt' from January to around July or August, with the highest median and IQR around these months, followed by a decline toward December. This suggests seasonal variation in the bookings.

   - **weekday:** The median counts appear relatively consistent across the weekdays, with some variations in the IQRs. No significant difference is apparent, indicating that counts are similar regardless of the day of the week

   - **workingday:** There is a slight increase in the median count on working days, but the difference in the IQR is not substantial, suggesting only a minor variation between working and non-working

   - **weathersit:** 'cnt' is high for 'clear' weather(highest median and IQR, indicating higher and more variable bookings) and very low

for 'Light Rain' weather(lowest median and IQR), which is obvious, and None in 'Heavy Rain'

So overall,

- In the US, the peak tourist season typically starts in late spring or early summer. Tourist-related bike rentals may not pick up until these months.
- Spring is often a busy time for students with exams and the end of the academic year approaching which may reduce the time available for leisure activities like biking.
- During early spring, the days are shorter, which might reduce the time available for outdoor activities like biking, especially for those who prefer riding in daylight.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans:
By setting *drop_first=True*, one of the categories is dropped, and the remaining dummy variables represent the presence or absence of the other categories relative to this reference.
using *drop_first=True* is important as it avoids multicollinearity. For example, if you have a categorical variable with three categories (A, B, and C), creating dummy variables for all three categories will result in a perfect linear relationship: knowing the values for two categories automatically determines the third. This leads to multicollinearity, which can distort the model coefficients and make the model unstable.
Including fewer dummy variables reduces redundancy and reduces the number of parameters in the model, which can help with model efficiency and reduce the risk of overfitting, especially in cases where the number of categories is large. In doing so there is no loss of information as absence in all dummy variables of a feature means presence in the dropped variable.
So to summarize, *drop_first=True* is used during dummy variable creation to avoid the dummy variable trap, improve interpretability, and enhance model efficiency. This approach ensures that the model is well-specified and free from issues related to multicollinearity.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: temp

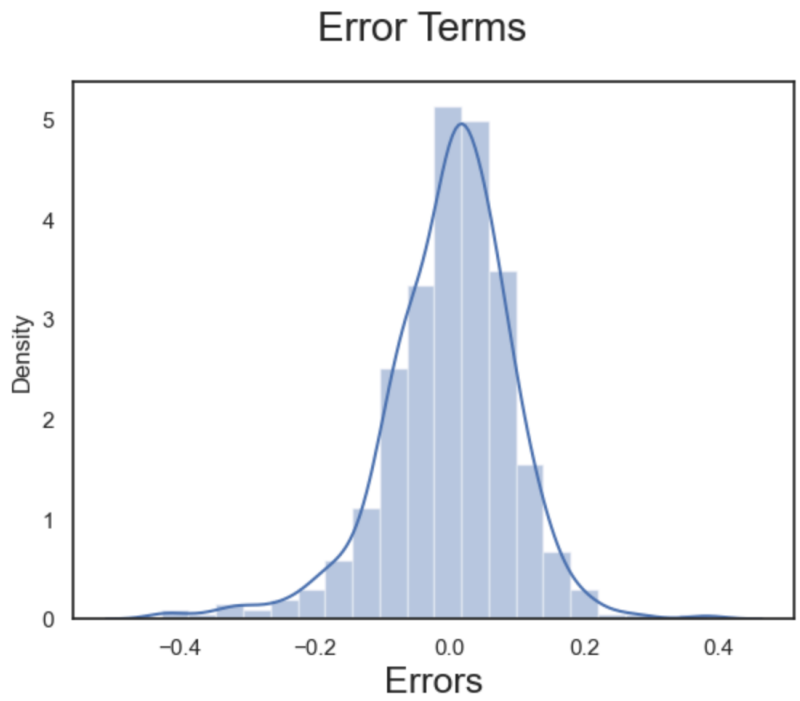
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

Assumption 1: Linearity - If the p-values for the predictors are small (usually < 0.05), it suggests that the predictors have a statistically significant linear relationship with the dependent variable. In the final model, all p-values < 0.05

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1883	0.024	7.879	0.000	0.141	0.235
2019	0.2357	0.009	27.655	0.000	0.219	0.252
temp	0.4090	0.030	13.839	0.000	0.351	0.467
windspeed	-0.1372	0.026	-5.311	0.000	-0.188	-0.086
spring	-0.1189	0.016	-7.614	0.000	-0.150	-0.088
winter	0.0451	0.013	3.536	0.000	0.020	0.070
September	0.0662	0.016	4.134	0.000	0.035	0.098
Clear	0.0769	0.009	8.498	0.000	0.059	0.095
Light Rain	-0.2074	0.026	-7.978	0.000	-0.258	-0.156

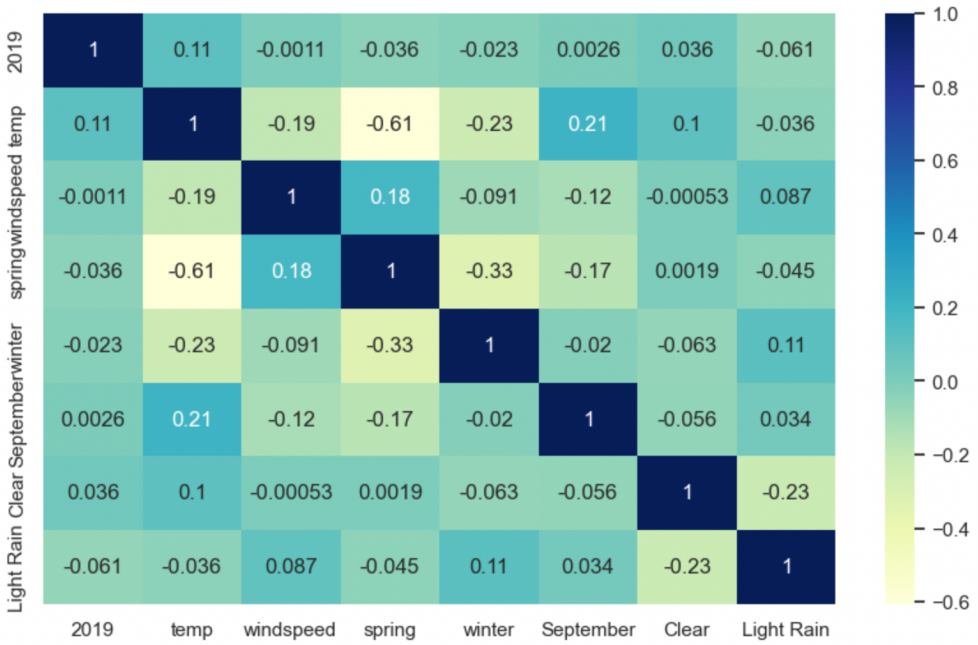
Assumption 2: Normality of Residuals - Plot a histogram of the residuals. Residuals should roughly follow normal distribution pattern.



Assumption 3: No Multicollinearity - Calculate VIF for each predictor. A VIF value above 5 or 10 indicates high multicollinearity.

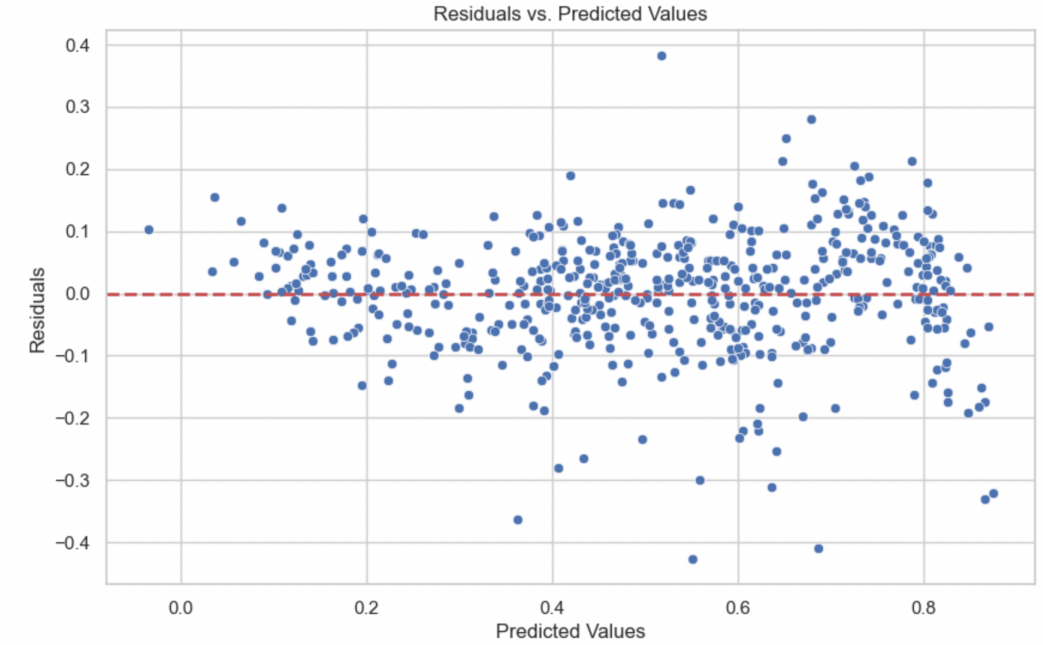
	Features	VIF
1	temp	4.60
2	windspeed	3.95
6	Clear	2.82
0	2019	2.05
3	spring	1.69
4	winter	1.36
5	September	1.15
7	Light Rain	1.11

Check the correlation matrix to identify highly correlated predictors.



Assumption 4: Error terms have *constant variance* (homoscedasticity):

Residuals vs. Fitted Plot: Plot the residuals against the fitted (predicted) values. The plot should show a constant spread of residuals across the range of predicted values.



Assumption 5: Error terms are *independent* of each other, Durbin-Watson Test: A value near 2 indicates no autocorrelation.

```
from statsmodels.stats.stattools import durbin_watson

dw = durbin_watson(y_train - y_train_cnt)
print(f'Durbin-Watson statistic: {dw}')
```

```
Durbin-Watson statistic: 1.9999932733055643
```

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: To identify the most important features contributing significantly towards explaining the demand of the shared bikes, we should examine the absolute values of the regression coefficients. Higher absolute values indicate more significant features. So, in our final model, **temp - 0.409, 2019(yr) – 0.2357, Light Rain(weathersit) – 0.2074**

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable (target) which is continuous and one or more independent variables (features). The goal is to find the best-fitting linear relationship that describes how changes in the independent variables affect the dependent variable. The algorithm is as follows:

Step 1: Data Loading and Understanding the Data

Step 2: Data Cleaning and handling missing values
Step 3: Removing columns which are not features
Step 4: Visualising and Understanding the Data (EDA)
Step 5: Data Preparation

in order to fit a regression line, we would need numerical values and not string. So we should create dummies for the categorical variables.

Step 6: Splitting the Data into Training and Testing Sets (it could be 70%-30% or 80%-20%)
Step 7: Feature Scaling (MinMax scaling/ Standardisation)
Step 8: Dividing into X(Predictor variable) and Y(Target variable) sets for the model building
Step 9: Feature Selection(Automatic + Manual) and building model
Step 10: Model Assessment and Comparison(reselecting features and reassessment on the basis of p-value and VIFs)
Step 11: Residual Analysis of the final train data
Step 12: Making Predictions Using the Final Model
Step 13: Model Evaluation (r2-score) – difference in r2 score of test and train data should be less(more difference means overfitted model)

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans:

Anscombe's quartet is a collection of four datasets that are commonly used to demonstrate the importance of graphing data before analyzing it and to show how descriptive statistics alone can be misleading. Each dataset in Anscombe's quartet has nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression line. However, when the data is visualized, each dataset reveals very different patterns and characteristics. This highlights the importance of visual data exploration and understanding the underlying data distribution and relationships.

The Four Datasets are:

1. Dataset I: Linear Relationship with a Small Error
2. Dataset II: A Nonlinear Pattern
3. Dataset III: Outlier Influences the Correlation
4. Dataset IV: Vertical Line with an Outlier

Anscombe's quartet illustrates several important lessons for data analysis:

1. Visual Inspection is Crucial
2. Understand the Data's Context
3. Check for Outliers and Influential Points
4. Use Multiple Tools and Approaches

Anscombe's quartet serves as a powerful reminder that data visualization and careful examination of data are essential components of any data analysis process. By going beyond simple statistical summaries, analysts can avoid misinterpretations and better understand the true nature of the data.

3. What is Pearson's R? (3 marks)

Ans:

Pearson's R, also known as Pearson's correlation coefficient, is a measure of the linear

relationship between two variables. It quantifies the strength and direction of this relationship.
Key Characteristics –

- Pearson's R ranges from -1 to 1.
- A value of 1 indicates a perfect positive linear relationship, meaning as one variable increases, the other variable also increases proportionally.
- Similarly, A value of -1 indicates a perfect negative linear relationship, meaning as one variable increases, the other variable decreases proportionally.
- A value of 0 indicates no linear relationship between the variables.
- Pearson's R is calculated using the formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- x_i and y_i are the individual sample points.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y values, respectively.

- Assumptions:
  Linearity: Pearson's R assumes that the relationship between the variables is linear.
  Homoscedasticity: The spread of the data points around the line of best fit should be roughly constant across the range of the variables.
  Normality: It assumes that the variables are normally distributed, particularly when inferring population parameters from sample data.

Pearson's R is a useful statistic for understanding the strength and direction of a linear relationship between two variables. However, it can be misleading in cases of non-linearity, outliers, or other data irregularities.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:

Scaling refers to the process of adjusting the range of feature values in a dataset so that they fall within a specific range or have specific statistical properties. This is often done before applying machine learning algorithms.
Scaling is performed for –

- Improving Algorithm Performance
- Ensuring Features Contribute Equally
- Avoiding Numerical Instability
- Consistent Interpretation of Coefficients

Normalization (or Min-Max Scaling) vs. Standardization (or Z-Score Normalization) –

1. Range:

Normalization rescales the data to a fixed range, usually [0, 1] while, Standardization rescales the data to have a mean of 0 and a standard deviation of 1.
2. Sensitivity to Outliers:
Normalization can be heavily influenced by outliers since it depends on the min and max values of the feature, while, Standardization is less sensitive to outliers, but they can still affect the mean and standard deviation.
3. Interpretation:
In normalization the transformed data points are within the same range, but not necessarily centered around 0 or with unit variance, while, in standardization the transformed data points are centered around 0 with unit variance, allowing for interpretation in terms of standard deviations.
4. Data distribution:
Normalization is generally recommended when the data has different ranges and the algorithm we are using does not assume normally distributed features (e.g., k-nearest neighbours, neural networks), while, Standardization is generally recommended when the data follows a normal distribution, and the algorithm assumes or benefits from normally distributed features (e.g., linear regression, logistic regression, principal component analysis).

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans:

The VIF for a predictor X_i is defined as:

$$\text{VIF}(X_i) = \frac{1}{1-R_i^2}$$

An infinite VIF occurs when $R_i^2 = 1$, indicating perfect multicollinearity. This occurs when one predictor variable is an exact linear combination of other predictor variables or the same variable is included more than once (Duplicate Variables)
When VIF is infinite, it indicates that the regression coefficients cannot be uniquely determined. This is because the design matrix X\mathbf{X}X (containing the predictors) becomes singular or near-singular, and the matrix inversion needed to compute the coefficients does not exist or is numerically unstable. Even if the VIF is not infinite but very large, it indicates that the standard errors of the coefficients are inflated. This means the model's estimates are less reliable, and it becomes challenging to determine the individual effect of each predictor.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Ans:

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to assess whether a dataset follows a specific distribution, often a normal distribution. In the context of linear regression, Q-Q plots are particularly important for checking the assumption that the residuals (errors) of the model are normally distributed.
X-axis: Represents the theoretical quantiles from a specified distribution (often the standard normal distribution).

Y-axis: Represents the observed quantiles from the dataset (often the residuals in the context of regression).
A 45-degree line (or line of identity) is typically drawn on the plot. If the data follows the specified distribution, the points on the Q-Q plot will lie approximately along this line.

Importance of Q-Q Plots in Linear Regression :

1. Checking Normality Assumption - A Q-Q plot can help assess if normality assumption holds. If the residuals are not normally distributed, the estimates and inferences made from the model may not be reliable.
2. Identifying Outliers - Q-Q plots can help identify outliers or extreme values that do not fit the expected distribution. Outliers can significantly affect the model, including the estimates of coefficients and the overall fit.
3. Assessing Model Fit - By examining the residuals through a Q-Q plot, you can determine if there are any systematic deviations from normality that might suggest a poor fit.
4. Guiding Model Adjustments - If the Q-Q plot indicates non-normality, you might consider transforming the response variable, applying a different model, or adding additional predictors to better capture the data's structure.

In short, Q-Q plots are a valuable diagnostic tool in linear regression for checking the normality assumption of residuals, identifying outliers, and assessing overall model fit. They provide a visual means to ensure that the underlying assumptions of linear regression are reasonably met, thereby helping to ensure the validity and reliability of the model's inferences.