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)

- 'yr': The bike bookings are more in yr value 1 (2019) than yr value 0 (2018). This means the bike sharing company is making good progress in their business over time.
- 'season' : season_3 (Fall season) attract most bike bookings as compared to other seasons.
- 'mnth' : Bike bookings rise from the starting months , tend to peak in the middle months of the year, with particularly high demand during the months mnth_6 (june), mnth_7 (july), mnth_8 (august), mnth_9 (september), mnth_10 (october) and then gradually decrease.
- 'weekday' : weekday_5, weekday_6 attract slightly higher bike bookings than rest of the days.
- 'weathersit' : weathersit 1 (clear/partly cloudy) attracted highest bike bookings followed by weathersit 2 (misty/cloudy) with the lowest bike bookings in weathersit 3 (snow/rain/thunderstorm). It is natural that people prefer clear weather conditions for outdoor activities like biking.
- 'workingday' : The bike bookings on working days and non working days are quite similar overall. However, working days tend to have higher bookings at the lower end of the distribution than non working days.
- 'holiday' : The median bike bookings on holidays appears to be lower than on non-holidays.Bike bookings on holidays have wide interquartile range with few extreme values.

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)

drop_first = True is important to use because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

For example, let us suppose we have 3 variables undergraduate, graduate, postgraduate.

Undergraduate	Graduate	Postgraduate
1	0	0
0	1	0
0	0	1

If we did not have spring-summer column, then Graduate=0 and Postgraduate=0 would naturally imply that education level is neither graduate, nor postgraduate so it is the remaining one which is undergraduate.

So we can easily drop the first column.

Graduate	Postgraduate
0	0
1	0

0	1

Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

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)

'temp' variable has the highest correlation with the target variable 'cnt'. They are positively correlated.

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)

- Normality of Residuals : by plotting the distribution of residuals I validated that they are normally distributed.
- Homoscedasticity : By plotting the scatter plot of residuals I validated that there is no significant pattern in them.
- Independence of Residuals : By noting the Durbin-Watson value 2 (actually 2.072) of final model, I validated that there is no dependency on the residuals.
- Multicollinearity : By ensuring that VIF (Variance Inflation Factor) values of all the predictors in final model are below the permissible value 5, I validated that there is no multicollinearity.
- Linearity : By noting that the CCPR (Component and component plus residual) plots of all the features in the final model are producing straight lines, I validated that the predictors have linear relationship with target variable.

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)

As per the absolute values of the coefficients or weights of the linear regression model, the top 3 features contributing significantly towards the demand of shared bikes are as follows:

- temp
- yr
- weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

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 is a supervised machine learning algorithm that models the relationship between a dependent variable y and one or more independent variables (also known as predictors or features) x1, x2, x3….xn . The goal is to find the best-fitting line that predicts the dependent variable using the independent variables. Linear relationship means that as the independent variable increases/decreases, the dependent variable increases/decreases linearly.

Simple Linear Regression:
Involves only one independent variable x to predict the dependent variable y.

$$\boldsymbol{y = \beta 0 + \beta 1 . x + \varepsilon}$$

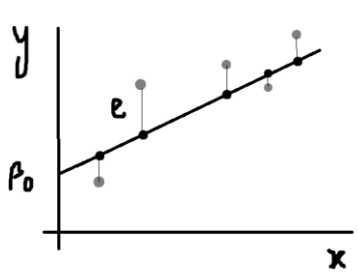
Where,
y = dependent variable (the value we are trying to predict)
x = independent variable used to predict y
β0 = the intercept term i.e the value of y when x=0
β1 = slope coefficient (the change in y for a one-unit change in x)
ε = error term (captures the deviation from the predicted value)

Goal is to find the values of β0 and β1 that best describe the relationship between x and y.

Multiple Linear Regression:
Involves two or more independent variables to predict the dependent variable.

$$\boldsymbol{y = \beta 0 + \beta 1 . x 1 + \beta 2 . x 2 + \cdots + \beta n . x n + \varepsilon}$$

Where,
y = dependent variable (the value we are trying to predict)
x = independent variable used to predict y
β0 = the intercept term i.e the value of y when x=0
β1, β2 … βn = coefficients for the independent variables
ε = error term (captures the deviation from the predicted value)



$y = \beta_0 + \beta_1 x + \varepsilon$

The difference between the predicted y value (ypred) and the actual y value yi is called the residual (denoted as e in above figure).
We need to minimize the residual sum of squares to find the best fit line.

$$\boldsymbol{Rss = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2}$$

Where,
y_i= actual value of the dependent variable for the i-th observation.
$\hat{y}_i$= predicted value of the dependent variable for the i-th observation.
m = total number of data points in the dataset.

But since the RSS is sensitive to change of units of the independent variables we use R^2 to measure the how well the independent variables explain the variability of the dependent variable. It's often referred to as the coefficient of determination.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where TSS is the Total Sum of Squares given by

$$TSS = \sum_{i=1}^{m} (y_i - \bar{y})^2$$

Where,
y_i= actual value of the dependent variable for the i-th observation.
$\bar{y}$= mean of the actual values of the dependent variable.

Assumptions of Linear Regression

- Linearity: The relationship between the dependent and independent variables is linear.
- Independence: The residuals (errors) are independent.
- Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables.
- Normality of Residuals: The residuals should be normally distributed.

Example Use Cases for Linear Regression:

a. Predicting housing prices based on features like size, number of rooms, and location.
b. Estimating sales based on advertising spend.

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 were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of data visualization and how different datasets with identical descriptive statistics (such as mean, variance, correlation, etc.) can have very different distributions and relationships. It shows that summary statistics alone (like the mean or correlation) are not sufficient to fully understand the nature of the data, and that visualization is essential for proper data analysis.
The data features must be plotted to see the distribution of the samples that can help us identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

Anscombe's Quartet Four Datasets

- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to deceive a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

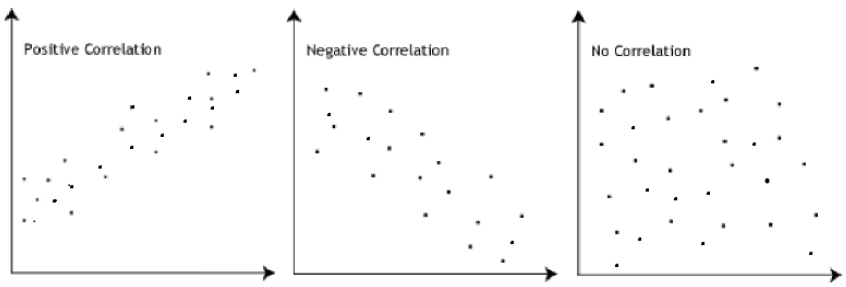
Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R, also called the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It indicates both the direction and strength of the relationship. It is the numerical summary of the linear association between the variables.

Formula to calculate Pearson's R:

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



Interpretation of Pearson's rrr:

- r=1: Perfect positive linear correlation. As x increases, y increases in perfect proportion (i.e.,

the points lie exactly on an upward-sloping straight line).

- r=−1: Perfect negative linear correlation. As x increases, y decreases in perfect proportion (i.e., the points lie exactly on a downward-sloping straight line).
- r=0: No linear relationship between the variables. The variables do not have any discernible linear pattern, though there might still be some other kind of relationship (e.g., non-linear).
- 0<r<1 : A positive linear relationship exists. As x increases, y tends to increase as well. The closer r is to 1, the stronger the positive relationship.
- −1<r<0: A negative linear relationship exists. As x increases, y tends to decrease. The closer r is to -1, the stronger the negative relationship.

Strength of the Relationship:

- Strong positive correlation: r between 0.7 to 1.
- Weak positive correlation: r between 0.3 and 0.7.
- Very weak / almost no correlation: r between -0.3 and 0.3.
- Weak negative correlation: r between -0.3 and -0.7.
- Strong negative correlation: r between -0.7 and -1.

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>

Feature scaling is the pre-processing technique of transforming feature values to a similar scale either by standardizing or normalizing the range of independent variables (features) in a dataset, so that each feature contributes equally to the model. Common methods of feature scaling include min-max scaling, which rescales the data to a specific range (usually 0 to 1), and standardization (z-score normalization), which centers the data around zero with a unit variance.

Scaling is especially important for algorithms that rely on distance metrics (like k-NN or SVM) or gradient-based optimization (like linear regression, logistic regression or neural networks), where features with larger values can dominate the model.

For gradient based algorithms: The difference in the ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

For distance based algorithms: Let's say we have data containing high school CGPA scores of students (ranging from 0 to 5) and their future incomes (in thousands Rupees). Since both the features have different scales, there is a chance that higher weightage is given to features with higher magnitudes. This will impact the performance of the machine learning algorithm making it biased towards one feature. Therefore, we scale our data before employing a distance based algorithm so that all the features contribute equally to the result.

Normalization	Standardization
Rescales values to a range between 0 and 1	Centers data around the mean 0 and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/standard deviation
Scikit-Learn provided MinMaxScaler for normalization.	Scikit-Learn provided StandardScaler for standardization.

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 is given by:

$$\boldsymbol{VIF = \frac{1}{1-R^2}}$$

VIF infinite means that $\boldsymbol{R^2}$ is 1. This indicates that there is a perfect correlation of a variable with one or more other variables.
Variance Inflation Factor (VIF) specifically measures the degree of multicollinearity among the independent variables in a regression model. It tells how much the variance of a regression coefficient is inflated due to collinearity with other predictors.

Suppose we have 4 feature variables $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \boldsymbol{x_4}$. If VIF of variable x1 is infinite, it means that

$$\boldsymbol{x_1 = a_2x_2 + a_3x_3 + a_4x_4}$$

Where $\boldsymbol{a_2, a_3, a_4}$ are constants.

This means that x1 can be perfectly predicted from the other variables, and no additional unique information is being provided by x1 in relation to the outcome variable.

In such cases of infinite VIF, we have to either remove the perfectly correlated variable or do

Principal Component Analysis.

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>
The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. For example, if the points are curved, it means that the data are skewed or have heavy tails. If the points are scattered or have gaps, it means that the data have outliers or are multimodal.

Use of Q-Q plot in linear regression:
A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals. You can also use a Q-Q plot to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To do this, you need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution.

Importance of Q-Q plot:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.
