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 data analysis shows a strong relationship of people choosing to rent a bike with year of renting, season, month of the year & weather. The bikes renting shows a **yearly** growth of more than 50%**.** On an avg, lesser number of people rent bikes in months of **Jan, Feb, Nov and December**. Similarly, the renting reduces to the lowest in **Spring (by more than 55%)** from Fall when it is at peak**.** The renting reduces to the lowest (**more than 60%) in rainy weather** than a clear weather day when it is at its best. The weekday, holiday, humidity and windspeed doesn't indicate a clear pattern against the number of bikes rented.

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)
Using the parameter drop_first = True while creating dummy variables for categorical features helps in reducing the number of dummy_variable created for the feature. If a categorical variable has n levels (possible values) & we use drop_first= True, it will create n-1 variables (with last value of the variable represented by all n-1 00000). Else, n dummy variables will be created each representing one of values of the variable. All these n variables are corelated and will introduce multi collinearity in the model.

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 scatter plots show that variable "temp" has the best positive linear relation with target variable count(of number of bikes getting rented).

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)
Residual analysis. The distribution of residuals(predicted target "count" for training data – actual value of count of training data) shows it almost a normal distribution centered around 0. Also, the scatter plot shows that the individual errors are independent and shows no pattern. This proves the model is an unbiased representation of the training data.

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 coefficients in summary shows variables ‘atemp’, ‘yr’, ‘humidity’ have highest values(approx. 46%, 27%, 25% respectively). That indicates that these variables cause highest contribution to demand of bikes getting rented.

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)

The liner regression algorithm aims to estimate the linear relationship of variables (numerical as well as categorical). Here we define the target variable in terms of sums of all independent variables multiplied by coefficients. E.g.

Y=β0+β1x1+ β2x2, β3x3…. Bnxn

In here, βi represent how much the value of Y will change with one unit of xi. The higher value of βi represents higher correlation. A high -ve value of βi represent a direct negative correlation (e.g Y decreases as xi increases). β0 represent a constant value intercept which takes when all dependent variables are 0.

The linear regression is a supervised model which we use for predictive analysis of Y. This helps us to select the variables which are related to target variable along with how much. The equation drawn can then be used to predict Y.

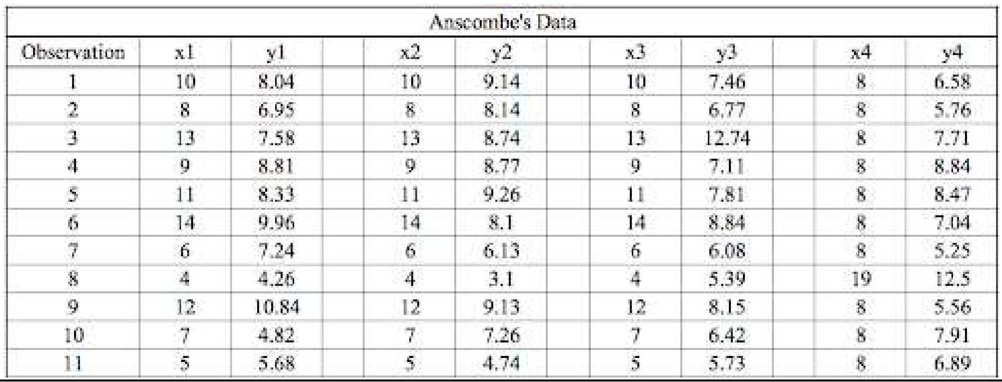
Couple of points to note:

1. Linear regression provides only correlation and not causation.
2. The predictions are limited to range of dependent variables used in training. So the extrapolations may not work.

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 was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of EDA and visualizing data and how depending solely on summary statistics can go wrong.

It consists of a set of four datasets which have same descriptive statistical properties like mean, variance, R-squared, correlations, and linear regression lines but still show difference when visualized by scatter plots. Despite the same summary stats, each of the four datasets include 11 x-y pairs of data. When these are plotted, each dataset shows unique variability patterns between x and y and distinctive correlation strengths.



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

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>

The Pearson correlation coefficient (R) is the most common way of measuring a linear correlation between two variables. It is a number between –1 and 1 that summarises the characterstic of a dataset. It is used to measure the strength and direction of the relationship between two variables.
A value closer to 1 & -1 show stronger relation. A -ve value indicate an inverse strong relation.
It can also be used for inference in hypothesis testing

$r = n(\sum xy) – (\sum x)(\sum y) / \surd[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2$

This indicates the correlation between the variables but can't be used to drive causation. It assumes:
1. a linear relation between the variables.
2. close to normal distribution of variables with minimal outliers
3. The dependent variables are fairly independent of each other

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>

During multilinear regression modelling, the multiple independent variables may be different scalese.e.g. distance of number of ppl in a country may be in crores but their age will be in smaller range ~0-100. Due to this difference, these variables may be different range for coefficients like 0.00003 and 900000. This creates a problem in interpretation of the model for business and further inferences.
To avoid this, the independent variables are scaled before fitting in the model. This helps in setting all coefficients in a common range and hence makes it easier to interpret.

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(Variable inflation factor) is a measure of multicollinearity in various independent variables in multilinear regression models. A higher VIF value of a variable suggests stronger collinearity with other variables which results in coefficient inflation of model. A value of inf for an independent variable represents its perfect multi colinear relationship with other independent variables. Such variables don't contribute to the goodness of the model and may cause deviation in coefficients for other variables. These variables should be removed (one by one) without causing an impact to R2 and will help improve 'adjusted R2'.

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 quantile-quantile (Q-Q) plot is a graphical tool to compare two sets of data to check if they come from the same distribution. It's a scatter plot that plots the quantiles of one set of data against the quantiles of the other set. The Q-Q plots are useful in linear regression because they can help ensure that the assumptions made about the data are valid. For example, if a statistical analysis assumes that the residuals are normally distributed, a normal Q-Q plot can be used to check that assumption. If the points on the plot fall approximately along a straight line, it suggests that your dataset follows the assumed distribution. However, the departure from the straight line indicate deviation from the assumed distribution.
