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. We saw that seasons like summer and winter have high demand whereas spring had a bit less demand, also the month of September contributed to higher demand, we also noticed that the year 2019 witnessed more demand than year 2018

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

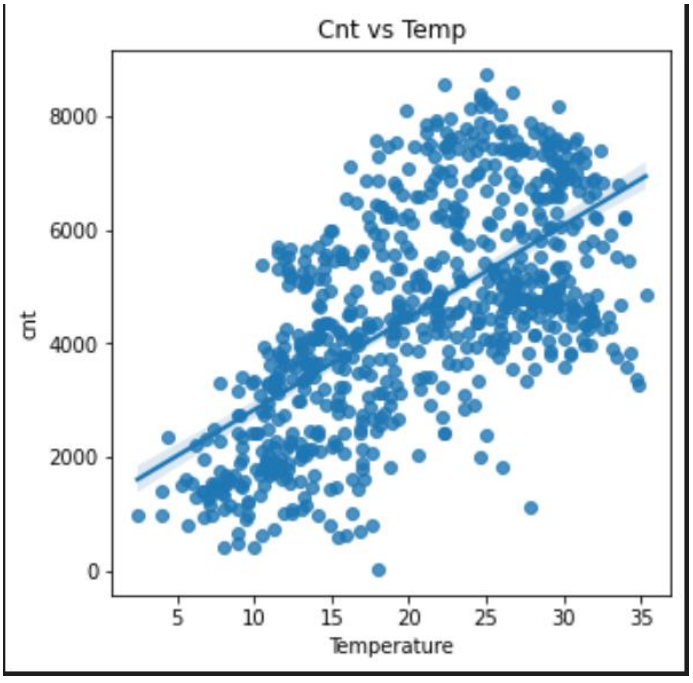
Ans. Whenever there are n values we can express/encode it into n-1 values, this helps us reduce the computations for an extra unnecessary column.
For example we have three categories A,B,C
These three categories can be encoded into 2 values as follows:

	A	B
A	1	0
B	0	1
C	0	0

So drop_first = True, helps us drop the first unnecessary column, without which the encoding can be done correctly.

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

Ans. Looking at the pair plot, temp has the highest correlation with the target variable i.e. cnt



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

Ans. The Assumptions of Linear Regression were validated as follows:

- **Multicollinearity** was validated by ensuring the VIF(Variance Inflation Factor) of all the attributes in the model are below 5, ensuring there are no closely inter related variables in the final model.

- **Homoscedasticity** was validated by ensuring that the variance of the residuals is having constant variance throughout.
- **Normality of Error** was validated by ensuring the Error values are distributed for any given value of X.
- **Independence of Errors** was validated by ensuring that the error values are statistically independent
- **We also validated the linear dependency of some variables with the target variable.**

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. Based on the final model we see that the Temperature, Year, and Season contribute significantly towards the demand of shared bikes.

General Subjective Questions

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

Ans. Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. It's used as a method for predictive modelling in machine learning, in which an algorithm is used to predict continuous outcomes.

Linear regression performs the task of predicting a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

A Linear Equation is of the form:
$y = c0 + c1*X1 + c2*X2 + \dots cn*Xn$

where is the dependent variable, and the X1,X2,X3 are the known attributes/predictor variables.

And c0 is the intercept, together this equation defines the linear relationship of the attributes to the dependent variables.

The coefficients c1,c2,etc can be defined as the number of units of increase in y a unit increase in Xi contributes to when all other predictor variables are kept constant.

The underlying algorithm behind finding the correct/best values of these coefficient is as follows:

While training the model we provide the training data as X(independent predictor variables) and y (target variable)

The model's error is calculated as such: Residual = y_pred – y_actual

And the cost function is as follows: minimize 1/n * Summation of (y_pred – y_actual)^2

This is also called the RMSE(Root Mean Squared Error) the goal is to minimize this to get the best fit line.

The model then uses the technique of Gradient Descent to reach to the local minima, the idea is to start with random variables and then iteratively reach to the minimum cost.

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

Ans. Anscombe's quarter comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed.

This shows us the importance of visualizing data before applying various algorithms to build models. It also shows that linear regression can only be considered for a data with linear relationships and is incapable of handling any other kind of data set.

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

Ans. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

The coefficient is calculated as follows:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,][\, n\sum y^2 - (\sum y)^2 \,]}}$$

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

Ans. Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling brings all of the data in the range of 0 and 1. The formula for this is as follows: x = (x-min(x))/(max(x) – min(x))

However, Standardization replaces the values by their Z scores, it brings all of the data into a standard normal distribution which has a mean of 0 and standard deviation as 1.
The formula for this is as follows: x = (x – mean(x))/sd(x)

It is used when the feature distribution is normal/gaussian. It is least impacted by outliers.

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

Ans. Variance Inflation factor(VIF) tells us to what extent the variables are correlated to each other.

It is computed as VIF = 1/(1-R2)

R2 values are within a range of [0,1] so when R2 is ~1, denominator would be close to 0 which would result in an infinite VIF

Which in turn means that this variable can be explained perfectly by other predictor variables, and hence can be easily dropped.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans. Quantile-Quantile plot is a graphical method for comparing two probability distributions. It is a plot of quantiles of two distributions against each other.

For a uniformly distributed data, q-q- plot would be perfect straight line through many points.

If we compare 2 different sample groups, it explains how one is different to other.