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?**

   “season”, “weathersit”, “holiday” ,”month” , “workingday”, and “weekday” were the categorical variables in the dataset. A boxplot was used to visualize these. These variables influenced our dependent variable in the following ways

- season: The boxplot revealed that the spring season had the lowest value of cnt, while the fall season had the highest value of cnt.
- weathersit: When there is heavy rain/snow, there are no users, indicating that the weather is extremely unfavourable. The highest count was observed when the weather forecast was clear and cloudy.'
- holiday: Rentals were found to be lower during the holiday
- month: September had the most rentals and Jan had the fewest.
- weekday : Weekends saw a significant increase in book hiring compared to weekdays
- working day : It had little effect on the dependent variable

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

   If we create dummy variables the end result column should be (n-1).
   Eg: If there is a column which has 10 unique categorical values or labels, using pd.getdummies() we convert them into a binary vector which makes 10 columns, one column for each unique value of our original column and wherever this value is true for a row it is indicated as 1 else 0. if drop first is true it removes the first column which is created for the first unique value of a column. In our case, it will be 9 columns, not 10 columns. It is useful because it reduces the number of columns, here is how, when all the other columns are zero that means the first column is 1.

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

   Temp and atemp are linearly correlated (positive correlation) with cnt (target variable)

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

- The Dependent variable and Independent variable must have a linear relationship.
  We can check this by using a pair plot and if it is not a linear one then we can fix this with fit_transform using the scaling method.
- No Perfect Multicollinearity
  In the case of very fewer variables, one could use a heatmap, but that isn’t so feasible in the case of a large number of columns.
  Another common way to check would be by calculating VIF (Variance Inflation

Factor) values.
If VIF=1, Very Less Multicollinearity
VIF<5, Moderate Multicollinearity
VIF>5 , Extreme Multicollinearity

We can fix this by dropping high VIF value variables one by one.

- ➢ Residuals must be normally distributed.
  Use a Distribution plot or histogram on the residuals and see if it is normally distributed.
  If the Residuals are not normally distributed, non–linear transformation of the dependent or independent variables can be tried.

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

As per our final Model, the top 3 predictor variables that influence the bike booking are:

1. Year (yr): The coefficient value of 0.2316 indicated that a unit increase in yr v ariable increases the bike hire numbers by 0.2308 units.

2. Temperature (temp): The coefficient of 0.5673 indicates that a unit increase in the temp variable increases the bike hire number by 0.5673 units.

3. windspeed: The coefficient of -0.1515 that a unit increase in windspeed variabl e decreases the bike hire numbers by 0.1515 units.

General Subjective Questions

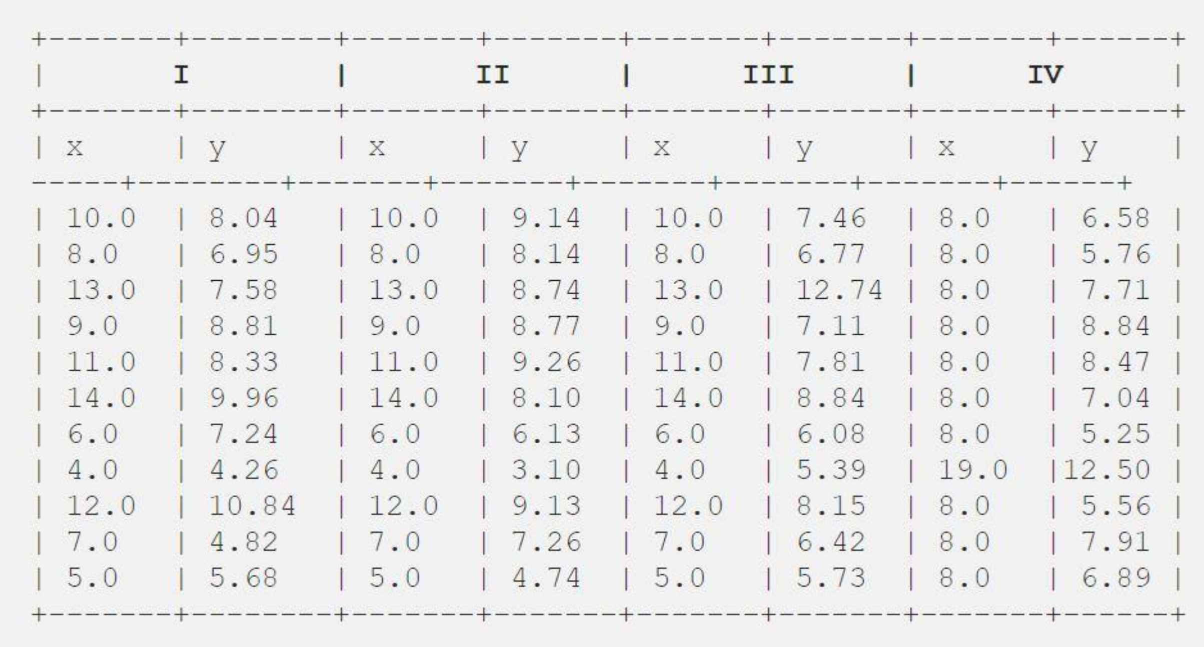
1. **Explain the linear regression algorithm in detail.**

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables.
y=mx+c
Here, x and y are two variables on the regression line.
m -slop of the line
c - intercept of the line
x – independent variable
y – dependent variable

2. **Explain the Anscombe’s quartet in detail.**

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

Once Francis John “Frank” Ans, a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points.



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

After that, the council analysed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

3. What is Pearson’s R?

- In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r.
- Bivariate correlation is a measure of linear correlation between two sets of data.
- It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.
- The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope

r = -1 means the data is perfectly linear with a negative slope

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

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

r- correlation coefficient

xi=values of the x-variable in a sample

(xbar) = mean of the values of the x-variable

yi =values of the y-variable in a sample

(ybar) =mean of the values of the y-variable

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

It is a step of data Pre-Processing that 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 time, the collected data set contains features highly varying in magnitudes, units, and ranges. If scaling is not done then the algorithm only takes magnitude into 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.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data falls below that point and 50% lies above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.