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:

i. Most bookings during `Fall` season and second most in `Summer` season
ii. Most bookings when `weathersit` is `clear` and second most when `Cloudy`
iii. There is no booking on a `HeavyRain`

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

Ans: drop_first=true is used to drop the first dummy variable when categorical variable is converted in to dummy variables. It is important because n-1 variables can correspond to n numbers of categories. When all other variables are having value 0 then it can be translated as first category.

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

Ans: **"temp"**: by looking at the pair-plot. "temp" has the highest correlation with target variable "users".

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

Ans: If the model on the training set is already built, we can validate the assumption by checking the p-value and adj R-Squrred values

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: In final model "temp", "season" and "months" are 3 feature that contributes significantly towards the demand of shared bikes.

General Subjective Questions

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

Ans: Linear Regression: It is used to predict the value of a dependent variable based on a value of an independent variable. There are two types of linear regression models:

a. Simple linear regression
b. Multiple linear regression

Simple Linear Regression Model can be used when the number of independent variables is 1. The simple linear regression equation provides an estimate of the population regression line. Relationship between X and Y is described by a linear function.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Multiple linear regression: Model can be used when the number of independent variables is more than 1. This model is used when single input variable is not enough to predict the output variable.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

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

Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

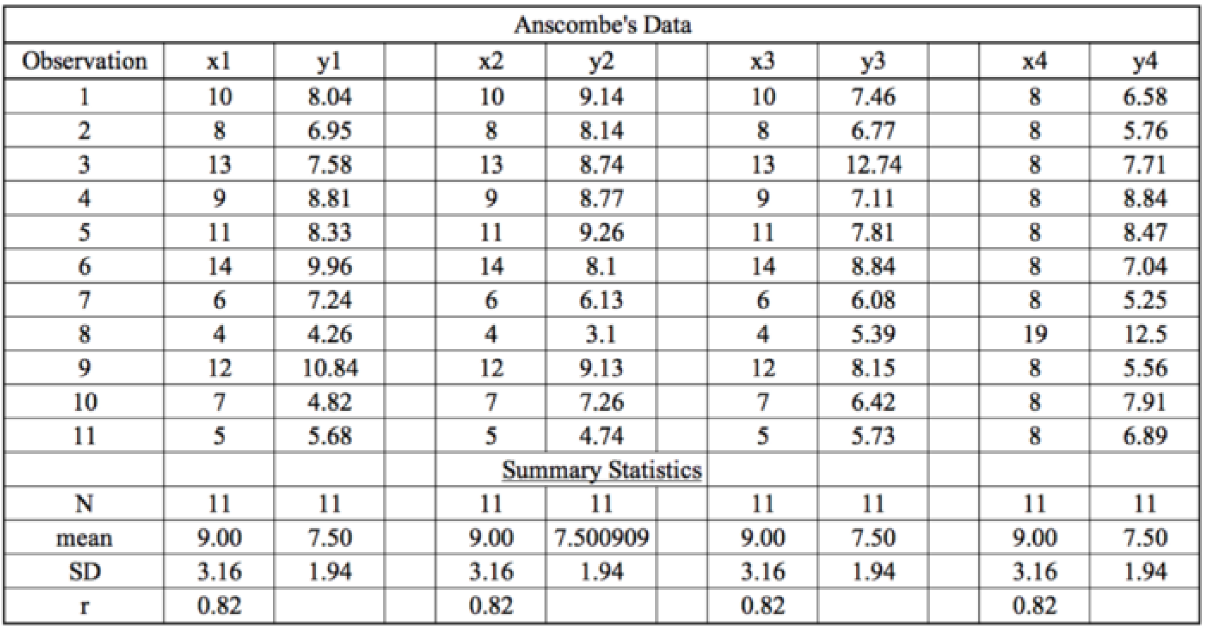
The four datasets can be described as:

i. Dataset 1: this fits the linear regression model pretty well.

ii. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

iii. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

iv. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model



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	

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

Ans: Pearson's R is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

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

Ans: Scaling: It 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.

Why scaling: Most times, collected data set contains features highly varying in units, magnitudes and range. If scaling is not performed, then algorithm only takes magnitude in account and not units hence incorrect modelling. To overcome 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.

The difference between normalised scaling and standardized scaling:

Normalized scaling brings all data in the range between 0 and 1 Normalized scaling $x = x\text{-}\min(x) / (\max(x) – \min(x))$	Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ). Standardization $x = x – \text{mean}(x) / \text{sd}(x)$

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

Ans: 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. (3 marks)

Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Importance of Q-Q Plot:

i. It can be used with sample sizes also

ii. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

iii. It is used to check following scenarios:

iv. If two data sets —

    a. come from populations with a common distribution
    b. have common location and scale
    c. have similar distributional shapes
    d. have similar tail behaviour