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)

   The following inference can be made from the categorical variables present in the dataset:
   1) The number of bikes rented are more during 'fall' season and less during 'spring' season
   2) The bike count is more in 2019 over 2018
   3) August, September and October had a greater number of bike hires
   4) The count is more when there is a clear sky

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

   While creating dummy variables, an extra column gets added which is not necessary for data analysis. The drop_first= True removes that extra column. This would reduce the column redundancy which would reduce multicollinearity ultimately.

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

   From the pair plot obtained, it is observed that the variables ‘temp’ and ‘atemp’ have the highest correlation with the target variable ‘cnt’.

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

   Validation was done by:

   1) Checking the error terms by plotting the residuals. They were binomially distributed with a mean at 0
   2) Checking homoscedasticity
   3) Checking collinearity

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?** (2 marks)

   The top 3 features contributing significantly towards explaining the demand of shared bikes are:

   1) Temperature (0.54)
   2) Year (0.23)
   3) Light snow (-0.23)

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm. Supervised means that it is dependent on the past data. In Linear regression, we have a past data. We train the model with the past data and then predict future values by testing the model. It provides linear relationship between independent variables and the target variable to predict values for continuous variables.
Equation of Linear Regression- y=mX + C (y- dependent variable, X- independent variable, C – intercept, m- slope).
There are 2 types of linear regression- Simple Linear Regression and Multiple Linear Regression
There are a few assumptions while doing Linear Regression:

1) Target variable and input variable are Linearly dependent
2) The error terms are normally distributed with a mean at 0
3) Error terms are independent of each other
4) Error terms have constant variance (homoscedasticity)

Steps:

1) Reading, understanding and visualizing data
2) Preparing the data for modelling (train-test split, rescaling)
3) Training the model
4) Residual analysis
5) Predictions and evaluation on test set

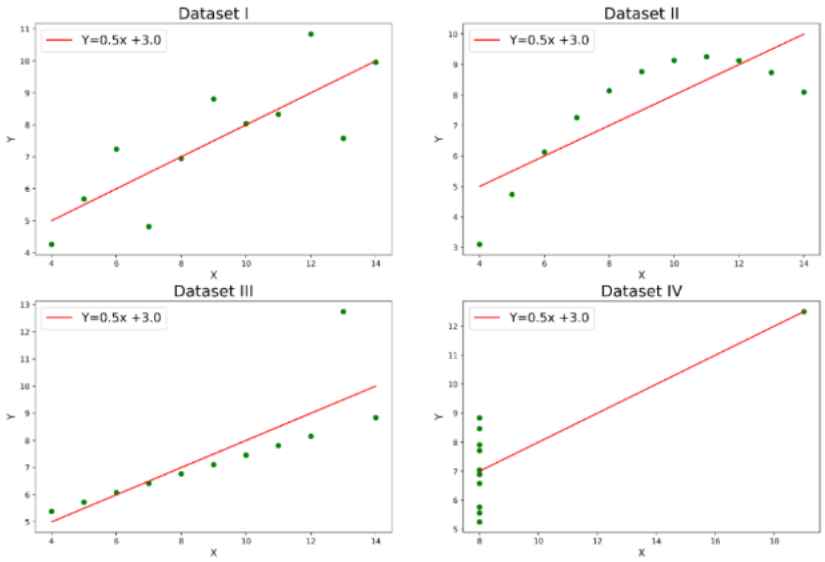
Our main goal at last is to find the best fit line with the minimum value of coefficients.

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

As the name stands, it is a set of four datasets created by Francis Anscombe in 1973. With the help of this dataset, we can demonstrate the importance of visualizing data and to present that summary statistics alone can be deceptive.

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

They have identical mean, variance, R-square, correlations.



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

Pearson's correlation coefficient R is a number which is between -1 and 1 which deals with linearity. It gives the strength and direction of relationship between 2 variables.

A R value > 0 denotes positive correlation
A R value = 0 denotes no correlation
A R value less than 0 denotes negative correlation

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

Scaling is the process of having all the attributes in the data of the same range. This is very useful for data analysis. For e.g., one column with values in lakhs and another column with values in single digit values would cause a lot of problems while analyzing. So, it is always recommended to match all the values in the same range. This process is called as scaling.

In normalized scaling, values range from 0 to 1. In standardization, mean is 0 and standard deviation is 1.

Normalized scaling is highly affected by outliers whereas there is no effect due to outliers in standardization.

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

VIF is called as Variance Inflation Factor.
It is given by:

$$VIF=1/(1-R2)$$

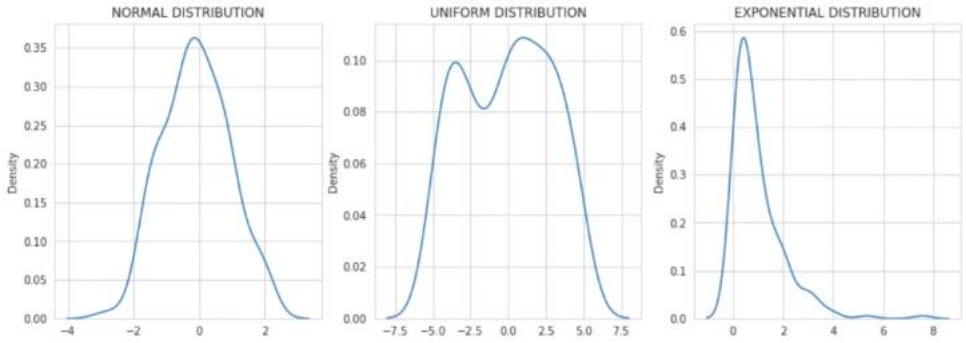
VIF becomes infinite when R2 is 1.

R2= 1 means that the variables are perfectly correlated (multicollinearity).
To solve this, a few columns have to be dropped.

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

(3 marks)

Q- Q plot – Quantile-Quantile plots



They comprise of normal, uniform and exponential distributions. They plot the quantiles of sample distribution against theoretical distribution.

In linear regression, it is always important to do residual analysis. There is an assumption that the error terms are normally distributed. The Q-Q plot will be useful in this case. It is also useful in determining the skewness of the distribution.