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)

1. In fall season, there is a significant increase in bike hires
2. In september, october and november there is a significant increase in bike hires which also explains previous point
3. Bike hire is more in weekdays
4. Clear weather attracts more bike hires
5. Bike hire decreases on holiday
6. In 2019, there are more hires which says that there is much more possibility of increase in next year

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

1. It is important to do drop_first = True because it removes the extra column created for each categorical variable during dummy variable creation.
2. By doing drop_first = True, we are reducing the number of redundant columns which inturn reduces collinearity.

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

1. Temp variable has the highest correlation with the target variable which means that the unit increase in temperature, bike hires will significantly increase.
2. Clear weather has higher chance of increase in bike hires.

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

1. Error terms are normally distributed with 0 mean
2. Homoscedasticity by plotting a graph of residuals and analyse variance
3. Checking for collinearity by plotting heatmap
4. Checking for linear relationship between independent variables and 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)

1. Temp with coefficient 0.54
2. Yr with coefficient 0.23
3. Winter with coefficient 0.15

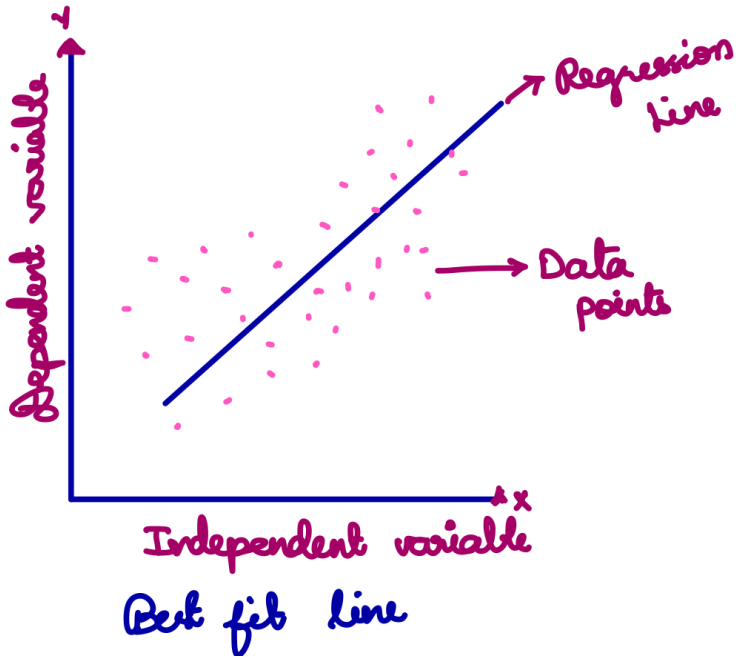
General Subjective Questions

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

Linear Regression provides linear relationship between independent variables and target variables to predict the outcome of future.

Its a statistical method used in datascience and machine learning

Linear Regression is a supervised machine learning algorithm which explains mathematical relationship between independent variables and predicts values for continuous variable which is also called as target variable



Linear regression equation: y = mX + c

X - independent variable

Y - Dependent variable

C - Intercept

For one unit increase in X, y increases by m times

When X = 0 , y value is c

Types of regression:

1. Simple linear regression

   Single independent variable is used to predict a dependent continuous variable

2. Multiple linear regression

   More than one independent variable is used to predict a dependent continuous variable.

Types of Linear Regression Line:

1.Positive linear relationship:

Increase in X - independent variable constitutes to increase in Y - dependent variable.

2.Negative linear relationship:

Increase in X - independent variable constitutes to decrease in Y - dependent variable.

Best-Fit line:

1. Our main goal is to find a best fit line which has the minimum value for co-efficients m and c.
2. For finding values for co-efficients we use **cost-function**
3. We find the best possible value for co-efficients using optimization algorithm called **Gradient descent** which uses cost function

Assumptions of Linear Regression:

1. Normal distribution of error terms with mean 0
2. Linear relationship between independent variables and dependent variable
3. No visible pattern of distribution of error terms - homoscedasticity
4. Less or no multicollinearity between independent variables
5. No autocorrelations between the error terms

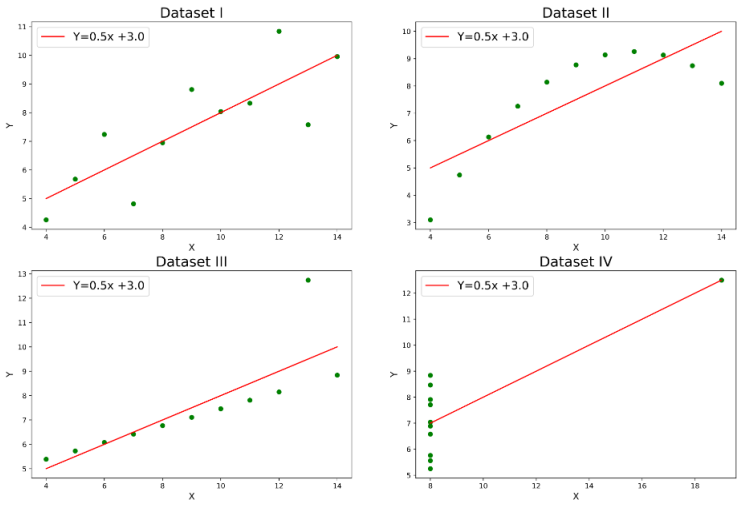
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of four datasets created by statistician Francis Anscombe in 1973.

All four datasets have identical mean, variance, R-squared, correlations.

Its created to demonstrate the importance of visualizing the 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



This explains the importance of EDA and disadvantages of depending only on the summary statistics.

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

Pearson's correlation co-efficient r is the most common way of quantitatively measuring the linear correlation.

Its a number between -1 and 1 which measures the strength and direction of relationship between two variables.

a. 0 to 1 - Positive correlation
b. 0 - No correlation
c. 0 to -1 Negative correlation

We can choose Pearson's co-efficient when all the following points are true:

1.Both variables are quantitative
2.Variables are normally distributed
3. Linear relationship between the variables exists
4. No outliers in data

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

Scaling is a process of converting all the variable to fall in the same range .

Real world data often contains data in different scales.

For example for an MNC, profits earned in different countries will be quantified using the respective currency. Ex: 100 Crore INR, 10 million dollars etc.

For our machine learning model to interpret the data correctly in the same scale, scaling is significant.

Normalised Scaling	Standardised scaling
Value ranges from 0 to 1	Mean is 0 and standard deviation is 1
Highly affected by outliers	Outliers does not have much effect
Original distribution's shape is preserved	Changes the original distribution
Upper limit for values is 1	No upper limit
Scikit provides MinMaxScaler transformer	Scikit provides StandardScaler transformer

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

1.VIF is calculated by following formula

VIF = 1/(1-R^2)

2.VIF becomed infinite when R^2 becomes 1 (1-1 = 0 ; 1/0 = infinite)

3. R^2 becomes one when there is a perfect correlation between variables

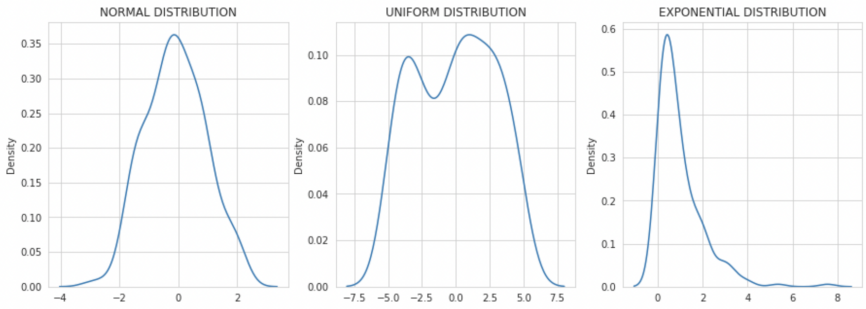
4.To solve this, we have to drop the variable which gives perfect multicollinearity

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 is also known as Quantile-Quantile plots.

They plot the quantiles of sample distribution against theoretical distribution

This helps us in understanding if our dataset follows any particular probability like Normal, exponential, uniform distribution.



Q-Q plots is very useful in determining

If two populations follows same distribution

If residual's follow normal distribution

Skewness of the distribution

Q-Q plots is important to check assumptions of linear regression.