Assignment-based Subjective

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

There are 4 categorical variables in the dataset:

a. Season: Has a negative impact on the dependant variable. Only spring season is found as a impactful season having a correlation coefficient of -0.1557
b. Weekday: Is not significant
c. Month: Is not significant
d. Weather Situation: Clear & Cloudy weather conditions have a strong impact on the total demand with coefficients of 0.247 and 0.1868

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

Dummy variables are created to convert categorical variables into numerical data. Let's say we have a categorical variable Gender with 3 values – Male, Female, Others. If we want to convert them to numerical form, we create 2 dummy variables, say D1 and D2

Male	Female	Others
1	0	0
0	1	0
0	0	1

We should delete anyone column (say: Others), because it is understood that if Male and Female both are 0, then it automatically means the value is '1' for others. This helps reduce multi-collinearity. If we do not remove one of the variables, it creates an additional variable that is redundant for model building and can create incorrect estimates.

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

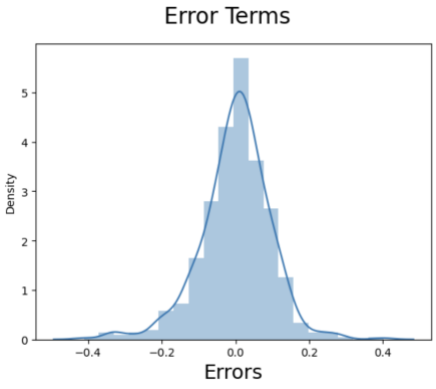
The best correlation is for:

a. Casual and Registered: We should not put too much focus on these parameters, because their sum is equal to the dependant variable
b. Atemp and temp: they have the same correlation (0.65) with cnt

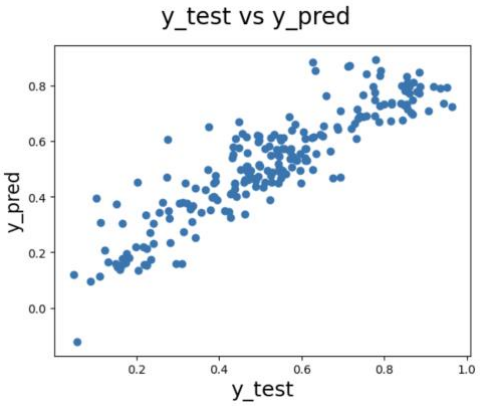
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Here are the steps to validate assumption of Linear Regression:

a. Residual Analysis: Scatter plot of actual values and predicted values are normally distributed and have a mean = 0



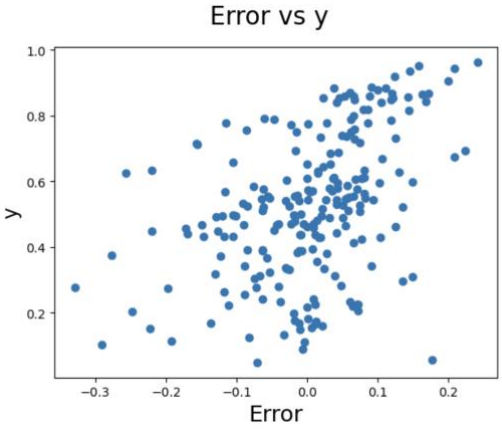
b. Constant Variance: The spread of variables is symmetric around the predicted line



c.

d. Low VIF: The independent variables do not have any dependency on each others. All parameters considered have VIF < 5.

e. Error Terms are independent of each other



f.

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

a. Temperature: has the highest positive correlation of 0.3672. It shows that higher the temperature, higher would be the demand. That means the summer months would be having a higher demand

b. Year: The demand is increasing greater with time showing that the later the year, the higher the demand

c. Clear: with a clear weather situation, the demand of bikes increases having a coefficient of 0.2470

General Subjective Questions

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

A linear regression model is a Machine Learning Model that tries to depict the dependent variable in terms of multiple dependant variables. It assumes a linear relationship between the target and independent variable. It is of two types:

- Simple Linear Regression: The target variable is represented in a linear relationship with one variable.

$$y = a_o + a_1X_1$$

- Multiple Linear Regression: The target variable is represented in a linear relationship with multiple variables. Assumption here is that if all other variables are kept constant, the target variable has a linear relationship with the final one

$$y = a_o + a_1X_1 + a_2X_2 + ... + a_kX_k$$

Here, y: dependent/Target variable
X_1, X_2:independent varaibles
a_o: constant term

How to build:

a. Convert categorical variables (if any) to numeric using dummy variables
b. Split data into train and test set (70-30, 80-20 ratio)
c. Choose training technique (forward, backward, automated, mixed)
d. Include terms which have p-value less than 0.05 and VIF < 5
e. The main goal is to minimise the error term (or the cost function)
f. Do not overfit or underfit the model
g. Evaluate the model: R squared/Root mean squared methods
h. Validate using Homoscedasticity, normally distributed error terms across mean = 0, etc.

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

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. Statistician Francis Anscombe constructed the data set to demonstrate both the importance of visualising data during analysis.
Each data set consisted of 11 data points as shown in table below:

	I		II		III		IV	
	x	y	x	y	x	y	x	y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89

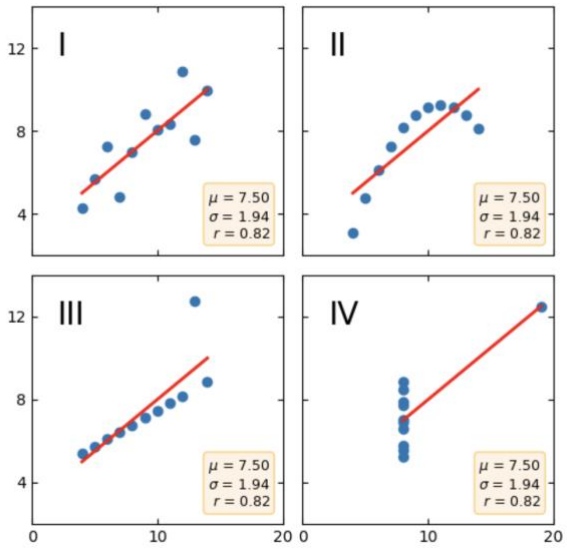
Statistical Summary (for all 4 datasets):

1. Mean of X: 9

2. Sample Variance of X: 11
3. Mean of Y: 7.5
4. Sample Variance of Y: 4.12
5. Correlation: 0.816
6. Linear Regression: y = 3.00+0.5x

While the statistical summary is the same, the graphs are widely different:

1. Graph 1: appears to be a simple linear relationship between x and y
2. Graph 2: The relationship is clearly not linear
3. Graph 3: While the data seems linear, one outlier changes the perfect correlation and reduces it from 1 to 0.82
4. Graph 4: just one outlier is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

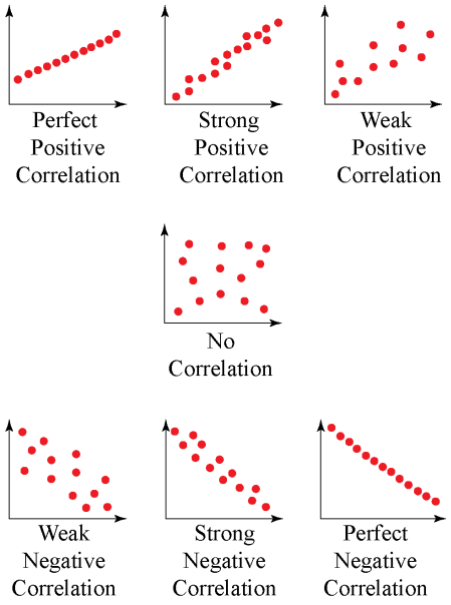


3. What is Pearson's R?

It is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It shows how much variance can happen in one variable if the other is changed by one unit. The value of 'r' can vary from -1 to 1. A positive correlation means a +ve linear regression (if one variable increases, so does the other) while a negative correlation means a -ve linear regression (if one variable increases, the other decreases). A value of 0 means no correlation between the variables.

There are a few drawbacks:

- Sensitive to outliers: As shown in Anscombe's quarter, it is highly sensitive to outliers and can manipulate the linear relationship
- Cannot capture non-linear relationships
- Correlation is not causation: If one variable increases, so does the other – it does not mean that one was the cause of the other. There could be many external reasons that can be responsible for the change



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

Scaling is an important pre-processing step. This step is to transform the numerical data to a specific scale. The goal here is to make all the features in a similar scale. It is performed because:

a. For Linear Regression: it becomes easier to interpret the terms. If the scale was different, the intercepts could be as low as 0.0001 (for area in sq ft) to as high as 1000 (for length in mm)
b. Many ML algorithms (gradient descent and distance-based algorithms) are sensitive to the scale of the features
c. Faster performance and convergence speed

This can be proven by an example: The cost of a house depends on its area (in sq ft) and number of bedrooms. The area in square feet would be in 1000s while number of bedrooms in single digit. The coefficient for both would be vastly different and makes it difficult to interpret.

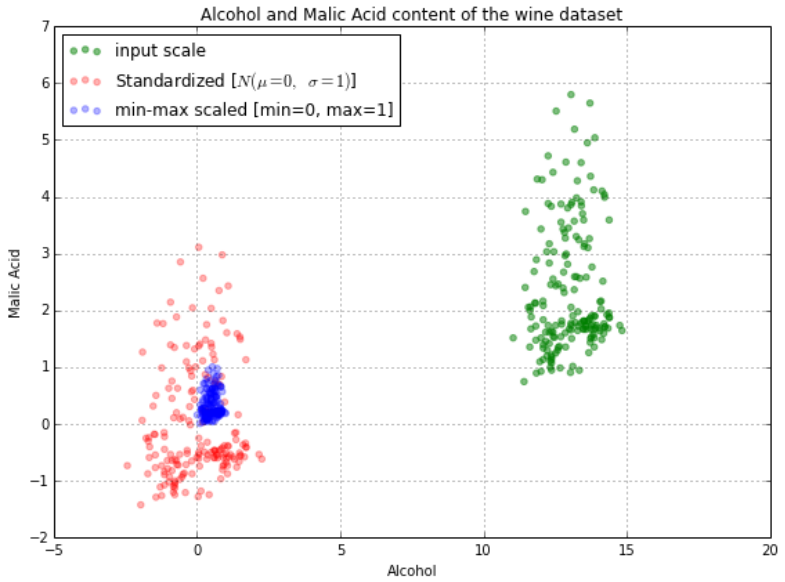
There are 2 major ways to scale data:

1. Normalised Scaling: It is also known as Min-Max scaling, it transforms the features to a fixed range, mostly between 0 and 1

$$X_{normalised} = \frac{X - Xmin}{Xmax - Xmin}$$

2. Standardized Scaling: It is also known as Z-score normalisation, it transforms the features to a mean of 0 and standard deviation of 1

$$X_{standardised} = \frac{X - Xmean}{Xstd}$$



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

Let's understand this through the formula of VIF:

$$VIF = \frac{1}{1 - Rsq}$$

VIF is used to decipher multi-collinearity. If there is a perfect correlation between 2 variables, the R value will be equal to 1. When that happens the denominator becomes 0 and VIF tends to infinite.

This might happen if the weight of the object is measures in kgs as well as lbs. They will have R=1. Alternatively, temperature in Celsius vs Fahrenheit will have the same effect.

How to resolve:

a. Remove one of the variables that is less significant. These terms will not add any value individually.
b. It can also be done via Principal Component Analysis (beyond scope right now)

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

Q-Q plot is a graphical tool that is used to 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.

For linear regression - when we divide the data into training and test data set - we can confirm using Q-Q plot that both the data sets are from populations with same distributions. It is used to check following scenarios. If two data sets:

a. Whether data comes from populations with a common distribution
b. Whether data has common location and scale
c. Whether data has similar distributional shapes
d. Whether data has similar tail behavior

Advantages of Q-Q plot

- While comparing two datasets the sample size need not to be equal.

- Since we need to normalize the dataset, so we don't need to care about the dimensions of values