

## Assignment #2

Student Name: Ruturajsinh Solanki	Section: 004
Student ID: 0827884	Date: 2023-03-20

### 1. Steps of building a decision tree, based on the data provided by the CEO.

City Size	Average Income	Local Investor	LOHAS Awareness	Decision
Big	High	Yes	High	Yes
Med	Med	No	Med	No
Small	Low	Yes	Low	No
Big	High	No	High	Yes
Small	Med	Yes	High	No
Med	High	Yes	Med	Yes
Med	Med	Yes	Med	No
Big	Med	No	Med	No
Med	High	Yes	Low	No
Small	High	No	High	Yes
Small	Med	No	High	No
Med	High	No	Med	No

- First to select the root node: Average Income (as its error rate is lowest)

#### 1) Error table for City Size:

Attribute	Rules	Error	Total Error
City Size	Big → Yes	1/3	5/12
	Med → No	1/5	
	Small → Yes	3/4	

#### 2) Error table for Average Income:

Attribute	Rules	Error	Total Error
Average Income	High → Yes	2/6	3/12
	Med → No	0/5	
	Low → Yes	1/1	

#### 3) Error table for Local Investors:

Attribute	Rules	Error	Total Error
Local Investors	Yes → Yes	4/6	6/12
	No → No	2/6	

#### 4) Error table for LOHAS Awareness:

Attribute	Rules	Error	Total Error
LOHAS Awareness	High → Yes	2/5	5/10
	Med → No	1/5	
	Low → Yes	2/2	

## Assignment #2

- According to these error tables, we can say that if Average Income is **MED** then the decision is also **NO** for every instance. On the other hand, for Average Income's **LOW** values the decision is **NO** for every instance. Thus, now we will create next level of decision tree using **HIGH** values of Average Income.
- Subbranches using Average Income
- High: use this branch data to build the subtree further.
- Med: no branch. (Decision = NO)
- Low: no branch. (Decision = NO)

City Size	Local Investor	LOHAS Awareness	Decision
Big	Yes	High	Yes
Big	No	High	Yes
Med	Yes	Med	Yes
Med	Yes	Low	No
Small	No	High	Yes
Med	No	Med	No

- For first subbranch of the tree:

1) Error table for City Size with Average Income HIGH:

Attribute	Rules	Error	Total Error
City Size	Big → Yes	0/2	1/6
	Med → No	1/3	
	Small → Yes	0/1	

2) Error table for Local Investor with Average Income HIGH:

Attribute	Rules	Error	Total Error
Local Investor	Yes → Yes	1/3	3/6
	No → No	2/3	

3) Error table for LOHAS Awareness with Average Income as HIGH:

Attribute	Rules	Error	Total Error
LOHAS Awareness	High → Yes	0/3	2/6
	Med → No	1/2	
	Low → Yes	1/1	

- As it can be clearly seen from the tables that City Size attribute has lowest total error. Thus, we will take next City Size as next branch.

## Assignment #2

City Size	Local Investor	LOHAS Awareness	Decision
Big	Yes	High	Yes
Big	No	High	Yes
Med	Yes	Med	Yes
Med	Yes	Low	No
Small	No	High	Yes
Med	No	Med	No

- As it is clear from the above table if the City Size is SMALL or BIG then the decision is YES. Thus, we will create subbranches using City Size values of Med.
- Subbranches using City Size:
  - Big: no branch. (Decision = YES)
  - Med: use this branch data to build the subtree further.
  - Small: no branch. (Decision = YES)
- For second Subbranch of the tree:
- Now, we will create using City Size = Med

1) Error table for Local Investor where City Size = MED:

Attribute	Rules	Error	Total Error
Local Investor	Yes $\rightarrow$ Yes	1/2	1/3
	No $\rightarrow$ No	0/1	

2) Error table for LOHAS Awareness where City Size = MED:

Attribute	Rules	Error	Total Error
LOHAS Awareness	High $\rightarrow$ Yes	0/0	2/3a
	Med $\rightarrow$ No	1/2	
	Low $\rightarrow$ Yes	1/1	

- For subbranch City Size = Med, we will create its new subbranch using Local Investor as it has least error.

Local Investor	LOHAS Awareness	Decision
Yes	Med	Yes
Yes	Low	No
No	Med	No

## Assignment #2

---

- Subbranches using Local Investor:
- Yes: use this branch data to build the subtree further.
- No: no branch. (Decision = NO)

- For third Subbranch of the tree:

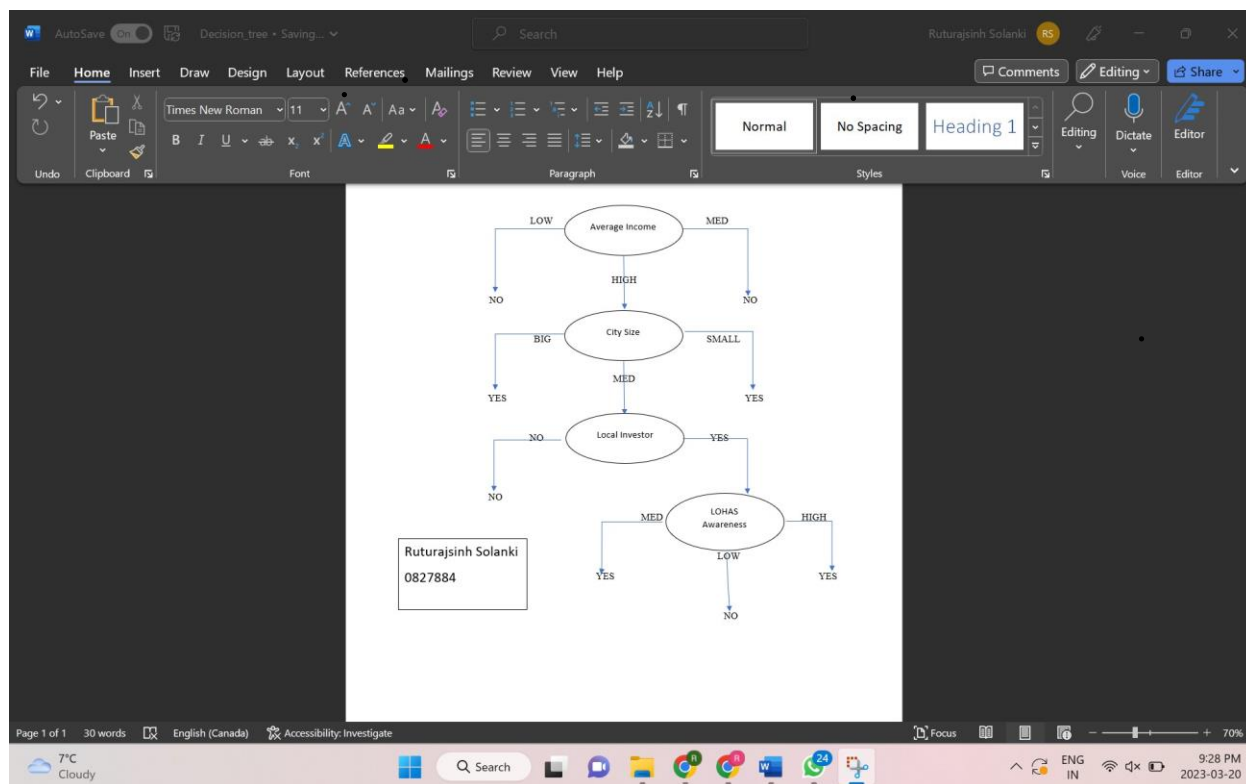
1) Error table for LOHAS Awareness where Local Investor = YES:

Attribute	Rules	Error	Total Error
LOHAS Awareness	High → Yes	0/0	2/2
	Med → No	1/1	
	Low → Yes	1/1	

- Thus, when the value of Local Investor = YES, we have just one variable left so we have no need of further dividing the tree.

## Assignment #2

### 2. Draw the decision tree in PowerPoint or Word.

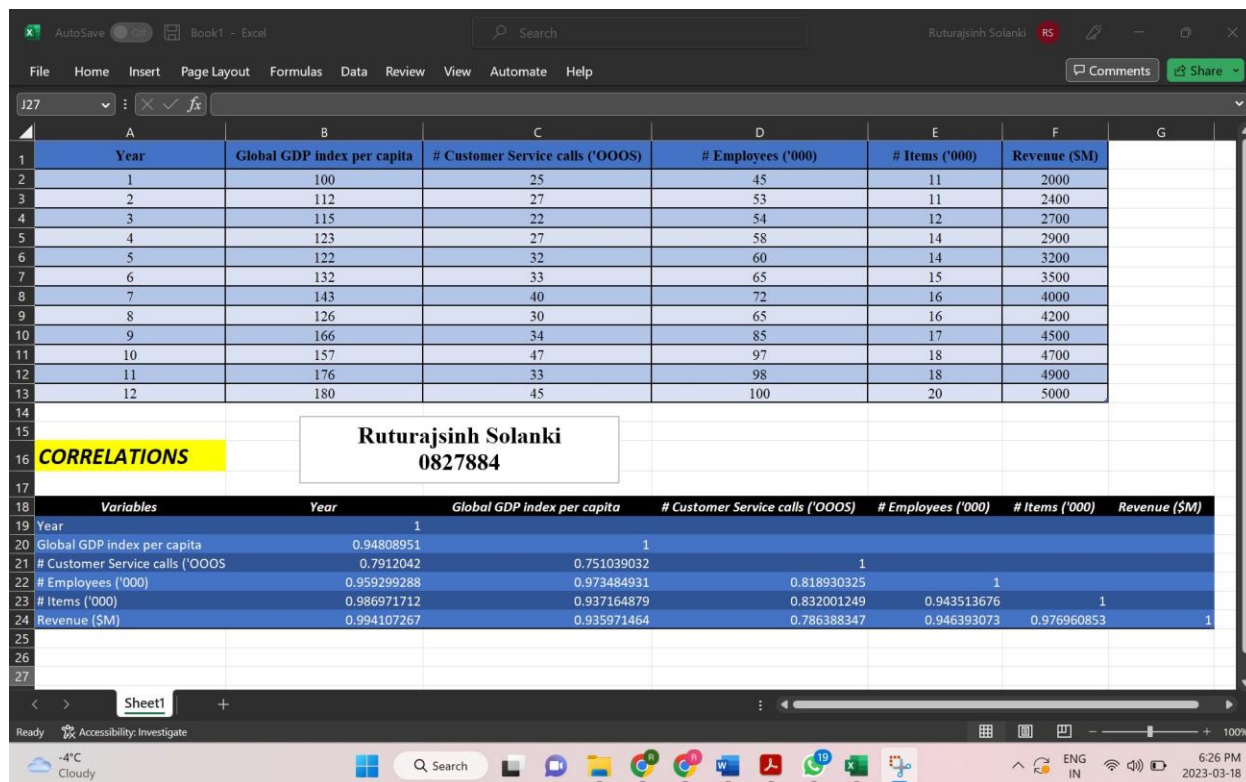


### 3. Your advice to the CEO about the new application.

Based on the available information, it is recommended that we refrain from opening a new store at this location. The reason for this is that the data within the new application appears to already be present in the provided dataset, and the decision based on this data is negative. This can be further confirmed by consulting the decision tree provided above.

## Assignment #2

### 4. Compute the correlations.



The screenshot shows an Excel spreadsheet with two sheets. The first sheet, 'Book1 - Excel', contains a dataset with 7 columns: Year, Global GDP index per capita, # Customer Service calls ('OOOS), # Employees ('000), # Items ('000), and Revenue (\$M). The data spans 12 years (rows 2-13). The second sheet, 'Sheet1', displays a correlation matrix for the same variables. The matrix shows that 'Year' is highly correlated with all other variables, with the strongest correlation being with 'Revenue (\$M)' at 0.994107267. The correlation between 'Global GDP index per capita' and '# Employees ('000)' is 0.973484931, and between '# Items ('000)' and 'Revenue (\$M)' is 0.976960853. The correlation between '# Customer Service calls ('OOOS)' and the other variables is significantly lower, all below 0.8.

Variables	Year	Global GDP index per capita	# Customer Service calls ('OOOS)	# Employees ('000)	# Items ('000)	Revenue (\$M)
Year	1					
Global GDP index per capita	0.94808951	1				
# Customer Service calls ('OOOS)	0.7912042	0.751039032	1			
# Employees ('000)	0.959299288	0.973484931	0.818930325	1		
# Items ('000)	0.986971712	0.937164879	0.832001249	0.943513676	1	
Revenue (\$M)	0.994107267	0.935971464	0.786388347	0.946393073	0.976960853	1

### 5. Explain which variables are strongly correlated.

Based on the correlation matrix, we can see that there are several variables that are strongly correlated with each other. The following pairs of variables have a correlation coefficient of above 0.95: (year is tightly correlated from all the variables)

- Year and Revenue - 0.994107267
- Year and Number of Items - 0.986971712
- Year and No. of Employees - 0.959299288
- Global GDP and Employees - 0.973484931
- Number of Items and Revenue - 0.976960853

Other variables are also strongly related with almost every pair of variables having correlation coefficient of above 0.8.

Just the pairs with Number of Customer Service Calls variable have coefficient less than 0.8.

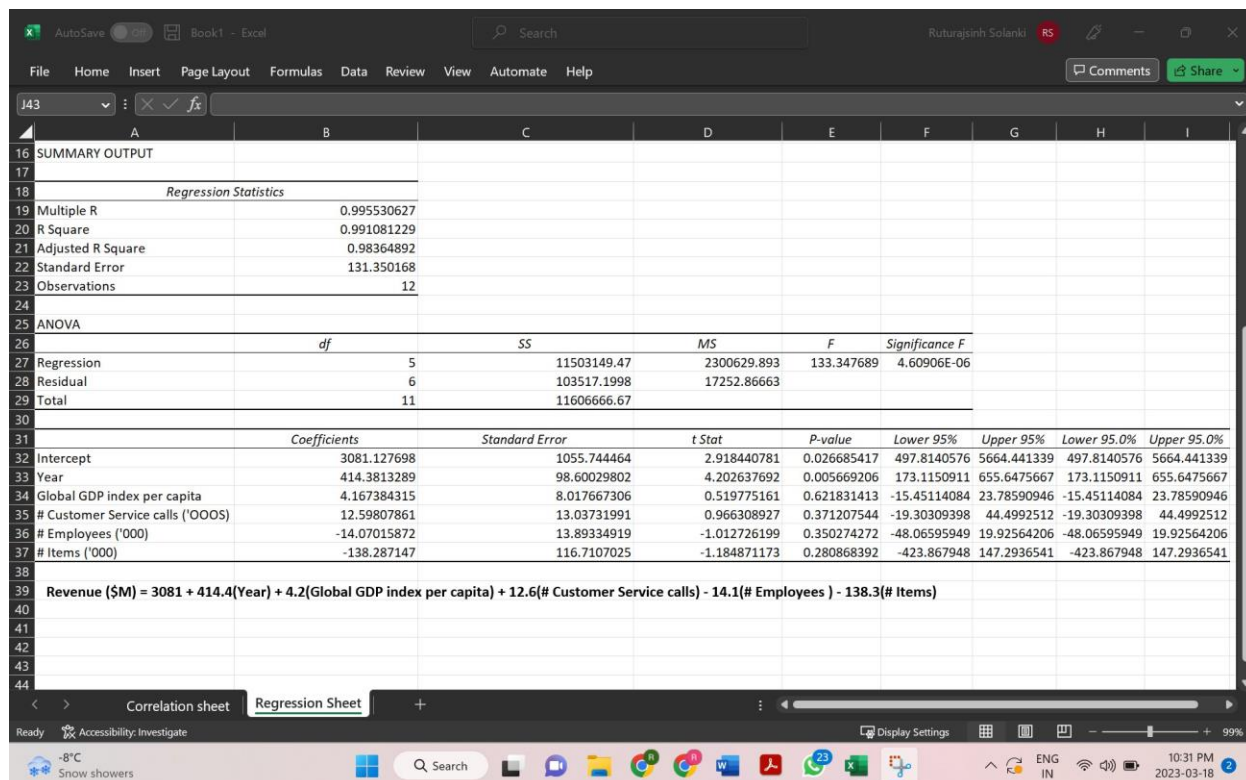
# DAB100 Introduction to Data Analytic

7 / 9

## Assignment #2

### 6. Your regression model for predicting the revenue.

Regression Equation: **Revenue (\$M) = 3081 + 414.4(Year) + 4.2(Global GDP index per capita) + 12.6(# Customer Service calls) - 14.1(# Employees) - 138.3(# Items)**



The screenshot shows an Excel spreadsheet with the following data:

SUMMARY OUTPUT								
<b>Regression Statistics</b>								
Multiple R		0.995530627						
R Square		0.991081229						
Adjusted R Square		0.98364892						
Standard Error		131.350168						
Observations		12						
<b>ANOVA</b>								
	df	SS	MS	F	Significance F			
Regression	5	11503149.47	2300629.893	133.347689	4.60906E-06			
Residual	6	103517.1998	17252.86663					
Total	11	11606666.67						
<b>Coefficients</b>								
		Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3081.127698	1055.744464	2.918440781	0.026685417	497.8140576	5664.441339	497.8140576	5664.441339
Year	414.3813289	98.60029802	4.202637692	0.005669206	173.1150911	655.6475667	173.1150911	655.6475667
Global GDP index per capita	4.167384315	8.017667306	0.519775161	0.621831413	-15.45114084	23.78590946	-15.45114084	23.78590946
# Customer Service calls ('OOOS)	12.59807861	13.03731991	0.966308927	0.371207544	-19.30309398	44.4992512	-19.30309398	44.4992512
# Employees ('000)	-14.07015872	13.89334919	-1.012726199	0.350274272	-48.06595949	19.92564206	-48.06595949	19.92564206
# Items ('000)	-138.287147	116.7107025	-1.184871173	0.280868392	-423.867948	147.2936541	-423.867948	147.2936541
Revenue (\$M) = 3081 + 414.4(Year) + 4.2(Global GDP index per capita) + 12.6(# Customer Service calls) - 14.1(# Employees) - 138.3(# Items)								

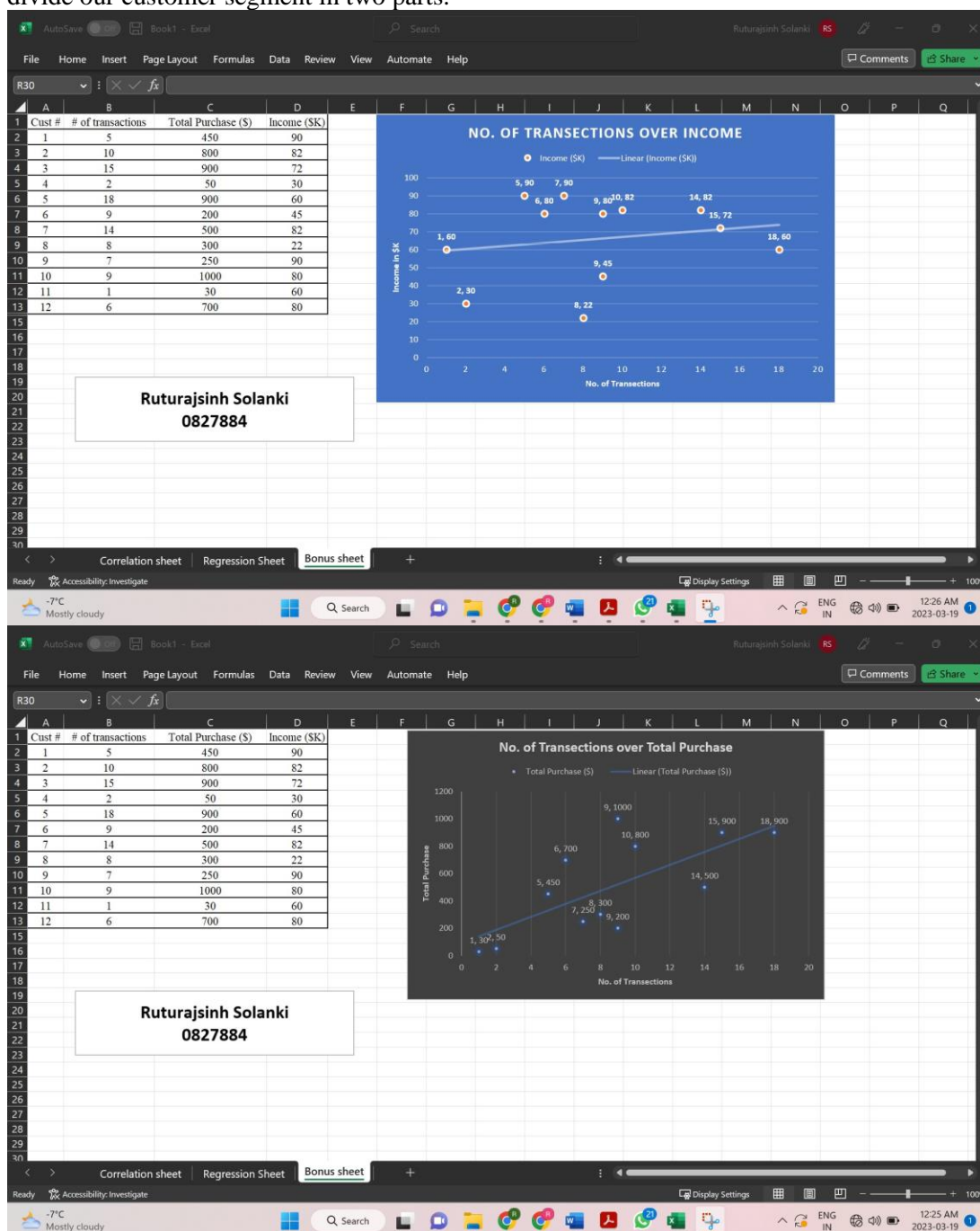


## Assignment #2

Below are bonus questions:

### 7. Determine the right number of customer segments. (30 points)

- First, we will plot scatter plots for our data one using No. of transactions vs. Income and another using No. of transaction and Total Purchase.
- Then, we can use gridline to see where we can see clusters and by adding a trendline we can divide our customer segment in two parts.





# DAB100 Introduction to Data Analytic

9 / 9

## Assignment #2

### 8. Centroids of each customer segments. (30 points)

- To create Centroid first we will need to find average of our features. Then, we will add this average to our both graphs using select data function of graph and then using add option we can add values respective to our graphs.

