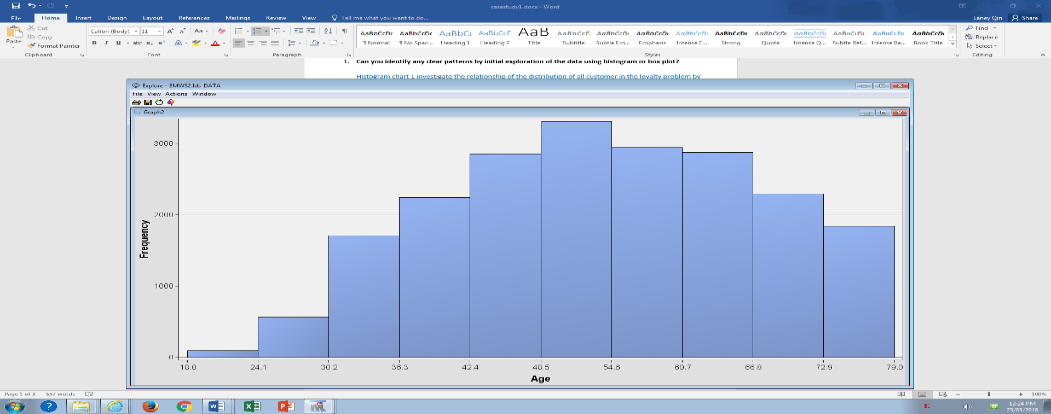
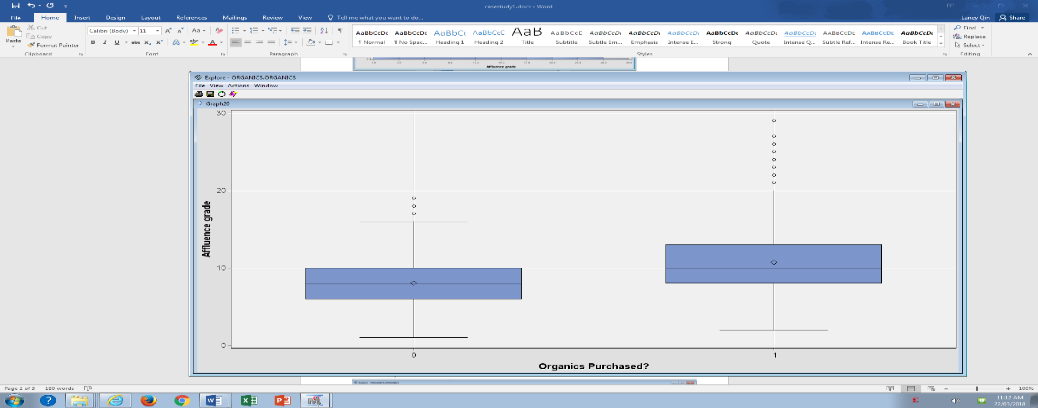
**Task 1. Data Selection and Distribution**

1. **Can you identify any clear patterns by initial exploration of the data using histogram or box plot?**  
     
   Histogram chart 1 investigate the distribution of customers in the loyalty problem by their ages. It tells that the customers whose ages are between 42.4 and 66.8 would be more likely to purchase organic products because of their significant proportion in the chart.

*Histogram chart 1: X = AGE*

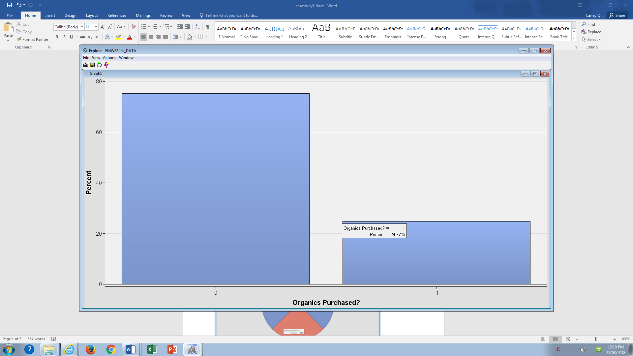
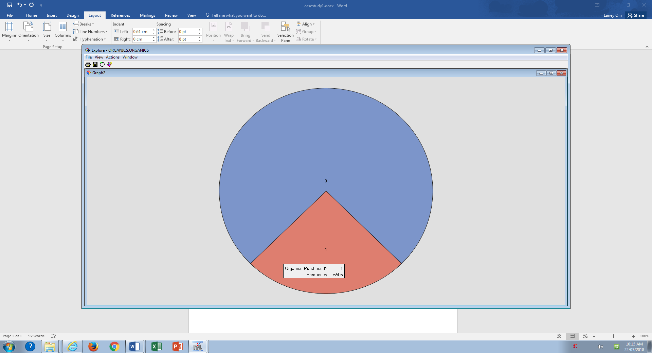
  
The box plot 1 explores the customer’s affluence grade with the relationship of whether they have bought organics or not. The plot tells that the customers with AFFL scaling from 8 to 13 (median grade: 10) have more chance to purchase organic products than the customers with scaling from 6-10 (median grade: 8).

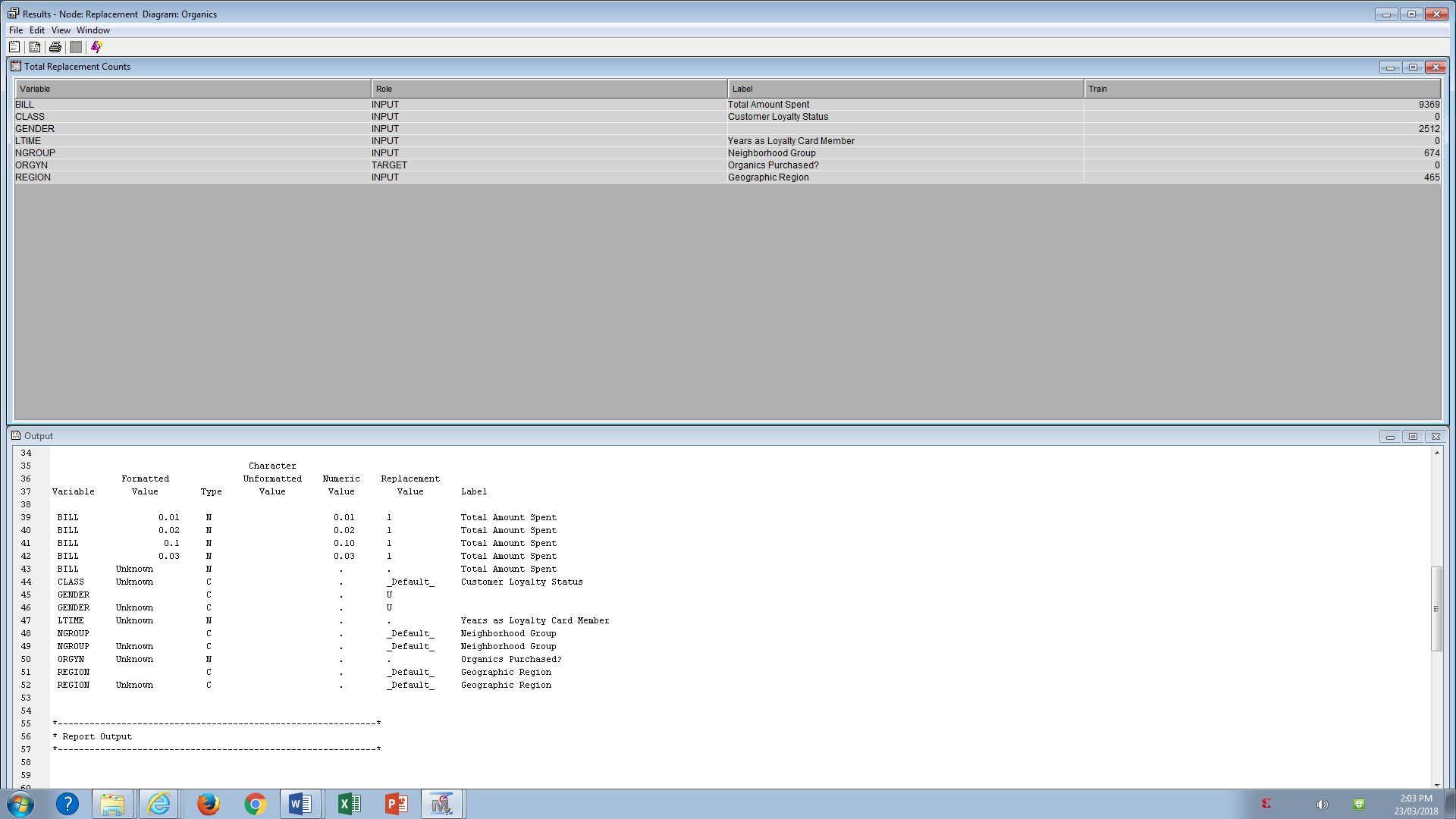
*Box Plot 1:*

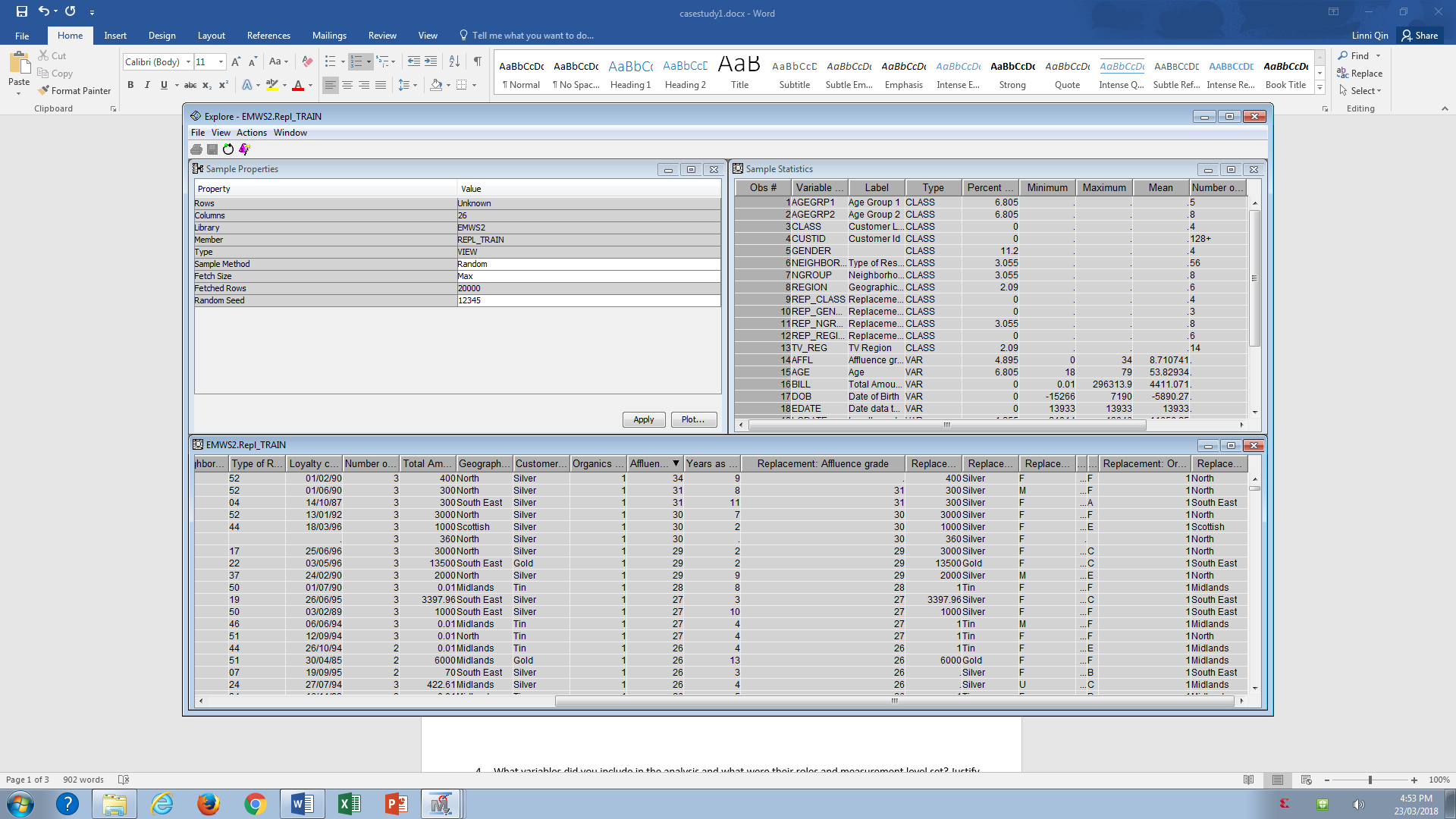
*X = ORGYN; Y = AFFL*

1. **What is the proportion of individuals who purchased organic products?**

As shown from the smaller sections on the following two charts, there are only 5,505 customers who have experience of purchasing organic products. This amount occupies around 24.77% of the total 22,223 customers.



1. **Did you have to fix any data qualify problems? Detail them.**  
     
   Yes, I made a certain of replacement for Class variables. For example, the percentage of missing value in the field of “GENDER” is significant, which is about 11.25%. “U”-unknown was used to replace the missing value so that the categories of GENDER reduce to three: “F, M, U”. Also, the value in the list of BILL less than $1 is replaced with 1.

  
Meanwhile, I made one change for Interval Variable – AFFL. After exploring the scale value under AFFL, it regularly scale from 0 to 31, however, there is one value without the equal distance “34”. I replaced it with missing value and the missing value is treated as a separate value.

1. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.  
     
   Based on total 18 variables, I chose 1 target and 8 input variables for the analysis for this case. The following table will list in detail their modelling role and measurement level respectively and the necessary justification.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Variable Name*** | ***Role*** | ***Reason*** | ***Level*** | ***Reason*** |
| AFFL | Input | Necessary to explore the relationship between AFFL and ORGYN. | Interval | Grade scales is equally from 1 to 30 |
| AGE | Input | Reject AGEGRP1, AGEGRP2, DOB, because they contains the same information. Also, the numeric value is easier to be handled by the software.  Necessary to explore the relationship between AGE and ORGYN. | Interval | Age range comes from 18 to 79 with equally gap. |
| BILL | Input | Necessary to explore the relationship between the amount of the customer spent and ORGYN. | Nominal | More distinct values without ordering. |
| CLASS | Input |  | Ordinal | Four status follows clear hierarchy. |
| CUSTID | Rejected | Unique character for each customer, it is not necessary for the analysis. | Nominal | More distinct values without ordering. |
| EDATE | Rejected | Not necessary for the analysis. | Unary |  |
| GENDER | Input | Necessary to explore the relationship between the GENDER and ORGYN. | Nominal | More distinct values without ordering. |
| LCDATE | Rejected | Repeated information of LTIME. | Interval |  |
| LTIME | Input | Necessary to explore the relationship between the time length of loyalty member and ORGYN. | Nominal | More distinct values without ordering. |
| NEIGHBORHOOD | Rejected | Information is collapsed in NGROUP | Nominal |  |
| NGROUP | Input | Not so necessary for the analysis. | Nominal | More distinct values without ordering. |
| ORGANICS | Rejected | Kind of repeated information of ORGYN. | Nominal |  |
| ORGYN | Target | It is relevant to customers who purchase or not the organic product and can be used for prediction. | Binary | 0 means has no experience of purchasing organic product. Conversely, 1 indicates yes. |
| REGION | Input | Necessary to explore the relationship between the living region of the customer and ORGYN. | Nominal | More distinct values without ordering. |
| TV REG | Rejected | Not necessary for the analysis. | Nominal |  |

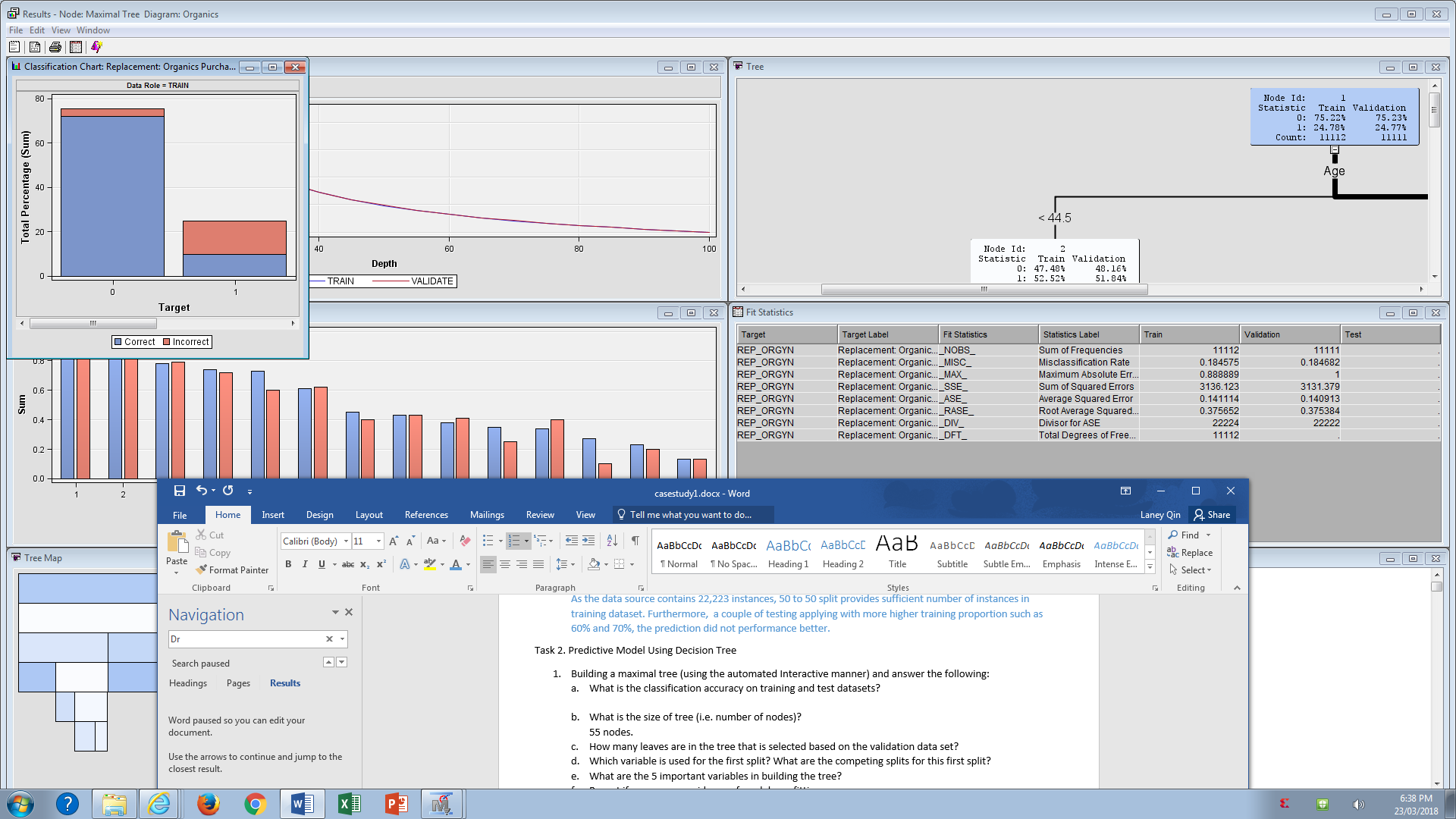
1. What distribution scheme did you use? What data partitioning allocation did you set? Explain your selection.

The distribution strategy for selecting the training and validation data sets is “Stratification”.

According to the lecture note of Dr Richi Nayak in week 2, stratification is able to derive an approximately equal samples of targeted from the data source by selecting the allocated percentage for each class (2018). In this case study, the targeted variable “ORGYN” contains two classes “0” and “1” and those classes should be equally sampled into training and validation data sets. Another reason to use “Stratification” is that the proportion of “0” is about 75% and “1” is solely around “25%”. The target amount is largely unbalanced.

As the data source contains 22,223 instances, 50 to 50 split provides sufficient number of instances in training dataset. Furthermore, a couple of testing applying with more higher training proportion such as 60% and 70%, the prediction did not performance better.

Task 2. Predictive Model Using Decision Tree

1. Building a maximal tree (using the automated Interactive manner) and answer the following:
2. What is the classification accuracy on training and test datasets?  
   
3. What is the size of tree (i.e. number of nodes)?  
   55 nodes.
4. How many leaves are in the tree that is selected based on the validation data set?
5. Which variable is used for the first split? What are the competing splits for this first split?
6. What are the 5 important variables in building the tree?
7. Report if you see any evidence of model overfitting.
8. Did changing the default setting (i.e. only focus on changing the setting of the number of splits to create a node) help improving the model? Answer the above question on the best performing tree.
9. Build an optimal tree and answer the following:
10. What is the classification accuracy on training and test datasets?
11. What is the size of tree (i.e. number of nodes)?
12. How many leaves are in the tree that is selected based on the validation data set?
13. Which variable is used for the first split? What are the competing splits for this first split?
14. What are the 5 important variables in building the tree?
15. Report if you see any evidence of model overfitting.
16. Did changing the default setting (i.e. only focus on changing the setting of the number of splits to create a node) help improving the model? Answer the above question on the best performing tree.
17. Choose the better performing tree amongst the maximal and optimal trees built in steps1 and 2. Using this model, can you identify which customers to garget for further marketing? Can you provide some descriptive summary of those customers?

Task 3. Predictive Modelling Using Regression

1. In preparation for regression, is any imputation of missing values needed for this data set? List the variables that needed this imputation and the process. Justify your choice of imputation if needed.
2. Build a regression model using the default regression method with the pre-processed data set. Answer the followings:
3. Name the regression function used.
4. Analyse the outcome and see whether the performance can be improved by using the selected variables (i.e. the subset of inputs selected either by stepwise or forward method).
5. Choose the best model and report the followings:
6. Which input selection method performed the best on this dataset or default was the best method?
7. What is classification accuracy/RMSE on training and test datasets on the best model?
8. How much was the difference in performance of various models build with each distinct method?
9. Show the set parameters for the best model.
10. Which variables are included in this regression model?
11. Provide the top-5 important variables in this model? Did the top-5 variables differ in other models? List them too.
12. Report any sign of over-fitting.
13. See whether you can further improve the performance by applying transformation to regularize input distributions. Choose the best model in previous step to apply transformation on variables. Report the variables that required transformation. What transformation function did you use and why? Does it improve the performance?
14. Using the best regression model, can you identify which customers to garget for further marketing? Can you provide some descriptive summary of those customers?

Task 4. Predictive Modelling Using Neural Networks

1. Build a Neural Network model using the default size setting. Answer the following:
2. What is the network architecture?
3. How much iteration is needed to train this network?
4. Do you see any sign of over-fitting?
5. Did the training process converge and resulted in the best model?
6. What is classification accuracy on training and test datasets?
7. Would feature selection help here? Build another Neural Network model with feature selection. Answer the followings:
8. Report the change in weight parameters by this setting.
9. What is the network architecture? What inputs are being used as the network input?
10. What is classification accuracy on training and test datasets? Is there any improvement in the outcome?
11. How much iteration is needed to train this network?
12. Do you see any sign of over-fitting?
13. Did the training process converge and resulted in the best model?
14. Would change in architecture help here? Build another Neural Network model by changing the number of hidden nodes of the network that performed better out of the previous two neural network models. Report the trained model.
15. What is the network architecture of the best performing network?
16. How much iteration is needed to train this network?
17. Sign of overfitting?
18. Did the training process converge and resulted in the best model?
19. What is classification accuracy on training and test datasets? Is there any improvement in the outcome?
20. Using the best Neural Network model, can you identify which customers to target for further marketing? Can you provide some descriptive summary of those customers? Is it easy to comprehend the performance of the best neural network model for decision making?

Reference

Nayak, R. (2018). Lecture02: Data Pre-processing for Data mining. Retrieved from https://blackboard.qut.edu.au/bbcswebdav/pid-7247916-dt-content-rid-10879791\_1/courses/IFN645\_18se1/2Lecturenotes%281%29.pdf