## Coursework

**You are required to analyse a large data set of your choice, which has been agreed with your module tutor:**

Your project may use any combination of data analysis techniques, data-mining algorithms and software that has been covered in the module. You may also apply them to any aspect(s) of the dataset for knowledge discovery.

You should cover the areas indicated below and your findings should be presented in the form of a report no more than **2400 word (absolute limit 3000 words)**. You will also be expected to give a **5-minute** oral presentation (see information below).

**The report will vary depending on your data set, so the details below are for guidance. The final report should cover all aspects highlighted below. Remember you are looking for trends and patterns which could be used by managers for some aspect of Business Intelligence.**

Please see below the aspects that you should consider:

**Individual Report**

- Introduction (10 marks) (240 – 300 words)
  - Brief overview of your dataset
  - Discussion of any cleaning undertaken on the dataset.

For this section you need to

1. Give a brief introduction to your dataset including:
   a. What your dataset represents
   b. Where the dataset came from
   c. The type a variable available
2. Discuss any changes/cleaning you have made to the data set ready for use in Tableau.
3. If your dataset is clean, then discuss how you could have cleaned the dataset

**NOTE: If you do not need to clean your data, you must discuss what you could have done to get the marks.**

- Data Analysis and Visualisation (35 marks) (approx. 840 to 1050 words)
  - Initial analysis of the data using visualisation techniques within Tableau (use diagrams/graphs to highlight important patterns/findings).
  - Discussion and interpretation of result.
  - Discussion of overall trends and patterns observed.

For this section you need to

1. Create about 10 graphs using Tableau. These can be any type of graph but should show some interesting information about your data set.
2. You should try to start with general graphs and then use more detailed graphs to investigate interesting findings.
3. For each graph you need to discuss the graph and the information identified, remember you are looking for overall trends and patterns.
4. This section could be concluded with a summary of the main finding (this could also be included within the discussion of each graph).

- Selection of Data Mining Algorithm and Data Pre-processing (10 marks) (240 – 300 words)
  - Select one data mining algorithm suitable for further analysis of your data.
  - Clearly justify your choice, with reference to the visualisation analysis carried out.
  - Identify and resolve any anomalies in the data (i.e. missing values, outliers etc.).
  - Carry out any appropriate pre-processing/transformations to the data set.

For this section you need to:

1. Discuss your chosen data mining algorithm
2. Briefly discuss how the data mining algorithm works
3. Explain why this algorithm is suitable for your dataset. This should refer to the visualisation you created in the previous section. For example, if you are using a classifier, you could refer to visualisation which indicate a relationship with your class variable.
4. Identify the main variables you will be using in your algorithm, including the class variable if appropriate.
5. If your dataset is clean, then define the anomalies and discuss how you could have dealt with them
6. Carry out any other changes to your dataset.
7. If you do not need to make any other changes, then discuss one transformation you could have carried out.

**NOTE: If you do not need to pre-process your data, you must discuss what you could have done to get the marks.**

- Data Mining (25 marks) (about 600 to 750)
  - Use the chosen data mining algorithm for further analysis of your pre-processed data set.
  - Clearly discuss the implementation of the data mining algorithm.
  - Discuss and interpret the results.

This section should use 1 data mining algorithms. Data mining algorithms include Time Series in Tableau, In Weka: Association Rules- Apriori algorithm, Clustering – K means, Decision Trees – J48 and Neural Networks.

For each Weka algorithm, you should:

1. Run 2/3 iterations of the algorithm using different subset of your data set. The following is a suggested format
    a. First iteration; use the whole data set, interpret the results.
    b. Second iterations; based on the results of the first iteration remove some variables from the dataset (justify the choice of variables to remove), interpret the results
    c. Third iterations; based on the results of the second iteration remove some other variables from the dataset (justify the choice of variables removed), interpret the results
    d. Choose the best iteration to discuss further.
2. Discuss the results/findings and comment on the fit of the model to the data for the best iteration. The strength of the fit of the data model can be judged by the following:
    a. Association Rule: use the confidence of each rule
    b. K-means clusters: use 1. The "within cluster squared errors", the smaller the better; 2. The comparison of the percentage in each cluster between the training set and the test set when using the percentage split option.
    c. Decision Trees and Neural Networks; Use the "percentage of correctly predicted instances" and the confusion matrix.

3. For Time series analysis, you should
    a. Produce several different time series (approx. 8 to 10) graphs making use of trend lines and forecasts.
    b. Discuss the results/findings and comment on the fit of the regression lines to the data. The strength of the fit of the lines can be judged by the r-value'. Tableau gives the r-squared value, from which you can calculate the 'r-value'. The r value lies between 0 and +/-1, the closer to 1 the better the fit. See notes in the timeseries folder on unihub.

- Data Ethics (10 marks) (240 to 300 words)
    - A discussion of data ethical issues related to the analysis and use of business data.

For this section you need to:

1. Discuss the ethical consideration relating to data analysis
2. Discuss the legal consideration relating to data analysis
3. Discuss the professional consideration relating to data analysis
4. Refer to the data ethic lecture.

- Conclusion (10 marks) (240 to 300 words)
    - A discussion of the overall visualisation results (e.g. What were the important findings? Summary of overall trends and patterns).

This section should summaries the overall trends and patterns found from your visualisation analysis.

- A discussion of the data mining results (e.g. How well did the model fit your data?).

This section should discuss the data mining algorithms used. The discussion should include:

1. The main findings, trends and patterns of the data mining
2. How well the final data model fitted the data.

- A discussion of the business intelligence that can be obtained from these results.

This section is a discussion of how the findings/results discussed above can be used by the managers (or other relevant parties) to improve their strategic decision making. The questions you should be asking are:

1. Who could use the finding/results?
2. How could they use the findings/results?
3. For what can they use the findings/results?
4. If possible give examples.

- **Pre-recorded Oral Presentation** (100 marks)
  - This **5-minute** oral presentation will allow you to discuss your analysis and results. (More details will be given in the lab sessions).

Details of the pre-recorded oral presentation will be given on Unihub.