COMP5310 Project Stage 1

Explore, clean, summarise and analyse the data

THE UNIVERSITY OF SYDNEY

Due: 11:59PM on 3rd of April 2025 (Week 6)

*This assignment is worth **15%** of the final mark of the unit of study.*

DATASETS

For this assignment, each member of a group needs to work on a different dataset. We have provided three datasets:

Dataset A:

- Auto Prices

Dataset B:

- Diabetes Diagnosis

Dataset C:

- Forest Covers

In Stage 1, each member will work on a different dataset, and then by the end of Stage 1, your group needs to agree on a single dataset to use for Stage 2 of the project which entails creating a predictive model to help answer your chosen research question.

GROUPS

This assignment must be done in **groups of 2 or 3**.

***Note:** there is work required from each member separately, but the project is handed in as a*

combined effort, and it is marked as a whole: there will be individual and group components to the marks, all based on the ***single submitted document****.*

Group formation procedure

Please head to Canvas and follow the following procedure to add yourself to a group: **COMP5310 Unit on Canvas Dashboard > Click on People > Click on the Project Stage 1 Groups Tab > then join a group** along with your group members. The minimum number of members in a group are 2 and the maximum number of members in a group are 3. Groups can be formed across tutorial sessions. **Groups cannot be changed after Friday Week 4**

Important Notes:

- **Within each group, each member must use a different dataset. If repeated datasets are used, marks will be deducted.**
- Exchange names and contact information (e.g., which social media platforms you prefer for coordinating).
- Arrange when to get together: At least one meeting per week is vital, but more frequent coordination is even better.
- For students who **cannot** form a group before **Friday Week 4**, they will be allocated to a group by the teaching team.

Dispute resolution

If during the course of the assignment work there is a dispute among group members that you can't resolve or that will impact your group's capacity to complete the task well, you need to inform the unit coordinator maryam.khanjanajafabadi@sydney.edu.au or one of the TAs: ssri4213@uni.sydney.edu.au or weiyi.wang@sydney.edu.au. Make sure that your email specifies the group name and is explicit about the difficulty; also, make sure this email is copied to all group members (including anyone you are complaining about).

COMP5310 Project Stage 1

Explore, clean, summarise and analyse the data

We need to know about problems in time to help fix them, so set early deadlines for group members, and deal with non-performance promptly (don't wait till a few days before the work is due to complain that someone is not delivering on their tasks). If necessary, the coordinator will split a group and leave anyone who didn't participate effectively in a group by themselves (they will need to achieve all the outcomes on their own). **This option is only available up until Friday Week 4**, which is the last day with time to resolve the issue before the due date. For any group issues that arise after this time, you will need to try to resolve the problem on your own, and you will continue to be treated as a single group which all get the same mark for this stage, based on whatever is submitted (though you should still let the coordinator and TAs know about them). If this is the case, groups may be changed after Stage 1 is finished.

PROJECT

Overview

The objective of Stage 1 of the project is to acquire and meticulously clean the dataset, followed by a comprehensive analysis to derive meaningful insights. Additionally, you will define a research question based on a research or business requirement, which you aim to answer in Stage 2.

DELIVERABLES

Report

The report must have **a maximum of 3 pages for each individual section**, **and either a maximum of 2 pages for the group section for a group of 2 or 3 pages for the group section for a group of 3**. You must use the high-level headings provided below to indicate the different sections and sub-sections of the report. You must use line spacing of at least 1.15pt, margins of at least 1.8cm, and body font size of at least 10pt. The goal is to convey the problem clearly and concisely.

The report should be in PDF format, named **"GroupX_A1_Report.pdf"**. Your **report MUST HAVE a front page** that gives the **group number**, and the **list of members** involved, giving their **SIDs** AND **Unikeys**. **DO NOT MENTION their names**. The body of the report must have a structure as follows:

COMP5310 Project Stage 1

Explore, clean, summarise and analyse the data

THE UNIVERSITY OF SYDNEY

Individual Component

The report must begin with a section per group member **mentioning the student's UNIKEY (DO NOT WRITE Student ID OR Student Name** to identify each individual section**)**. Each individual section must include:

1. **Topic and research question:** Provide a comprehensive and insightful description of the problem, highlighting the business/research need, clearly state your research question, and indicate some groups of stakeholders and how they could be helped by answering the research question.
2. **Data description:** Provide a description of the data, indicating the number of attributes and instances, and state the relevant metadata about this dataset, including a data dictionary which indicates the attributes on your dataset, a description of each attribute, and the data type of each attribute (int, float, string, date, etc.).

   Note: *The data dictionary should be included as an appendix and will not be counted towards the page limit.*
3. **Data ingestion and cleaning:** Describe the data ingestion and data quality assurance and cleaning process, including:

   3.1. Data ingestion: Describe any data ingestion steps, indicating if you used a Pandas data frame or a database in PostgreSQL, and briefly describe the data structure or schema.

   3.2. Data quality assurance and cleaning: Describe how you ensured data quality, if there were any quality problems, describe what they were and how you cleaned the data. Remember to **justify your decisions**, for example, if you decide to remove any rows with missing data, explain why you decided to do this and how your decision might impact data quality. Indicate which tools you used to ingest and clean the data, for example, indicate which Python functions you used to clean your data.

   Note: *You don't have to include the code on the report, as you will submit it separately.*

4. **Exploratory data analysis (EDA):** Describe in detail any exploratory data analysis you performed which provided you relevant information to answer your research question. This analysis must include **TWO supporting figures** and a detailed discussion of the results obtained, indicating what they tell you about your data and how these results could impact the modelling results in the next stage of the project. **Do not include a matrix of**

figures of multiple analysis of all attributes, you need to select and highlight the TWO most important results from your analysis. **Do not utilise any predictive models at this stage.** Please note that the **Two Supporting Figures** of the EDA section **must be provided in the EDA section**. **DO NOT put them in the appendix otherwise they will not be marked**.

Group Component

Finally, you will need to include a group section at the end of the report, including:

1. **Discussion:** Discuss your thoughts on the strengths and limitations of each dataset, for the purpose of investigating the question of interest. Discuss and critically analyse the exploratory data analysis performed in each individual section, highlighting the strengths and limitations of each approach.

2. **Conclusion:** Summarise the most important outcomes from the exploratory data analysis performed by all members of the group. Discuss how the cleaned dataset & outcome of the EDA will help you answer your research question. Finally, mention the final agreed upon dataset that your group will choose to work on in Project Stage 2 and justify your decision/choice.

Note: *The different report sections and sub-sections are aligned with the marking rubric. Therefore, please include only the requested contents and do not mix or merge the sections, as this will interfere with the marking process. For example, don't include the data cleaning steps under the exploratory data analysis section; this must be included in the data cleaning section. If you fail to do so, this won't be considered for the marking.*

Appendix Section in Report

Each report must have an appendix section. The appendix section must be placed at the very end of the full report, that is after all the individual and group components. The appendix section must include the data dictionary for each dataset and each data dictionary must be properly labelled. The appendix section can also include your references, if you have used any. Please note that the two supporting figures of the EDA section must be provided in the EDA section. DO NOT put them in the appendix. Any information in the appendix, apart from the data dictionaries and the references, will not be counted towards marking. The appendix will not be counted towards the page limit of the report.

COMP5310 Project Stage 1

Explore, clean, summarise and analyse the data

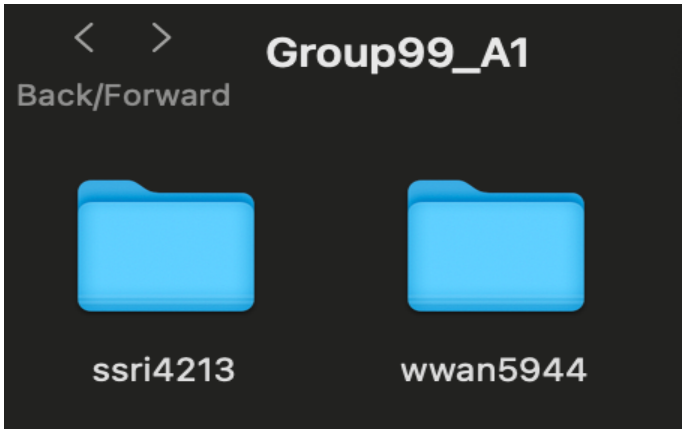
Code and Dataset

Along with the report, you must also submit the Python code used in this assignment as a **single zip or tar.gz folder** which **MUST BE** named using the following convention: "***GroupX_A1***", where X is your group number. **DO NOT SUBMIT A FOLDER THAT IS NAMED GroupX_A1.** This compressed folder MUST contain **one subfolder for each member of the group, where each subfolder must be named using each student's respective unikey** (***DO NOT PROVIDE STUDENT ID OR STUDENT NAME AS THE SUBFOLDER NAME***). Each subfolder must include the following:

1. The Jupyter notebook with the code each member used to perform their work, named using the following convention: "**unikey_A1_Code.ipynb**".
2. The final clean dataset in CSV format, named "**unikey_A1_CleanDataset.csv**".

Therefore, each group will have 1 main folder (named using the following convention: ***GroupX_A1***) that will be submitted on Canvas. This group folder MUST CONTAIN the following items:

- 2 subfolders for a group of 2 members or 3 subfolders for a group of 3 members. Each subfolder must be named after the unikey of the student whom the subfolder belongs to. DO NOT PROVIDE STUDENT ID OR STUDENT NAME.
- Each subfolder will have 2 items: 1 Python Code File (named using the following convention: **unikey_A1_Code.ipynb**) and 1 Cleaned Dataset in CSV format (named using the following convention: **unikey_A1_CleanDataset.csv**).
- Look at the screenshot below for reference:



Please note that the Group Report, which must be submitted as a PDF File, will be submitted on a different submission portal than the Code and Dataset submission portal. **The report should be in PDF format**, named using the following convention: ***"GroupX_A1_Report.pdf"***

COMP5310 Project Stage 1

Explore, clean, summarise and analyse the data

Submission Portals

1. Please upload the report in PDF format, named "***GroupX_A1_Report.pdf***", in the Report submission portal.

2. Please upload the main group folder, named "***GroupX_A1***", containing the code and final clean datasets in the Code and Dataset submission portal.

KEY POINTS OF INFORMATION

1. ***Each member MUST work on a different dataset.***

2. ***In any part of either the report or the code file, each student must identify their own section(s) or component(s) using their Unikey (THIS IS A UNIKEY: ABCD1234). DO NOT provide Student ID or Student Name.***

3. ***Your report MUST HAVE a front page. The front page of the report MUST contain Assignment Title which is Project Stage 1, Group Number, the Student ID (SID) of each member and the UNIKEY of each member. DO NOT provide names of any group member. DO NOT PROVIDE any additional information.***

4. ***DO NOT include a Content's Page at any part of your report. It is not required. Adding a content's page will be counted towards the total page count and marks will be deducted if any report section goes beyond the permissible page count limit set forth in the assignment guide.***

5. ***ONLY 1 SUBMISSION PER GROUP is required. In other words, only 1 member from the group must submit the assignment on Canvas.***

6. ***Each report must have an appendix section. The appendix section must be placed at the very end of the full report, that is after all the individual and group components. Any information in the appendix, apart from the data dictionaries and the references, will not be counted towards marking.***

7. ***The different report sections and sub-sections are aligned with the marking rubric. Therefore, please INCLUDE ONLY the requested contents and DO NOT MIX OR MERGE THE SECTIONS, as this will interfere with the marking process. If you fail to do so, this won't be considered for marking. DO NOT RENAME ANY SECTION HEADINGS.***

8. ***You MUST ONLY USE Jupyter Notebook for your code. You are NOT PERMITTED to use any other IDEs, such as Google Colab, Spyder, etc for your code file. MARKS WILL BE DEDUCTED if students require the markers to run the code file on any IDE other than Jupyter Notebook.***

9. ***Students must follow the report format exactly as given in the assignment guide. DO NOT add your own sections or sub-sections to the report. DO NOT RENAME ANY SECTION HEADINGS. DO NOT REMOVE ANY SECTIONS. Simply follow the report format mentioned in this assignment guide. Providing your own section headings or following a format different than what is***

mentioned in the assignment guide will lead to those sections being ignored by the marker and your marks being reduced. No further appeals will be entertained.

10. ***In the case of misnamed sections, sections being absent, or sections not following the right order as mentioned in the assignment guide, that specific section will be completely ignored by the marker.***

11. ***Ethical AI usage, critical thinking, and originality are fundamental expectations in all assignments. Ensure that your work is independently produced and reflects your own understanding.***

COMP5310 Project Stage 1

Explore, clean, summarise and analyse the data

MARKING

Marking Criteria	Marks
Individual Component	
Topic and research question	1
Data description	1
Data quality and cleaning	3
Exploratory data analysis	3
Supporting figures	1
Code quality	1
Group Component	
Discussion	3
Conclusion	1
Report format and presentation	1
TOTAL	**15**

Deductions

- 1 mark will be deducted if your section of the report exceeds the maximum number of pages. If the group section exceeds the maximum number of pages, the deduction will apply to all group members.

- 1 mark will be deducted from the team member whose unikey is not mentioned at the commencement of their respective code section and / or report section.

- 5% of the maximum awardable mark will be deducted per day of late submission. Zero marks will be awarded after 5 calendar days from the due date.