Data Science Industry Project – Part 1 (MAST90106)
Written Report

THE UNIVERSITY OF MELBOURNE

Profiling Electricians on Airtasker to Improve Public Safety

Group 9

Chen Zhou (987776)

Meghna Panda (1128829)

Vignesh Lakshminarayanan (1032043)

Xinyu Mao (1091647)

Contents

1 Introduction

Energy Safe Victoria (ESV) is the state's safety regulator for electricity, gas and pipelines. Under the Energy Safety Victoria Act 2005 ESV is responsible for monitoring, auditing, and enforcing compliance with the requirements as well as administering licensing, registration and approval systems that maintain safety standards and skills [1]. Over the several years statistics show that people are most tempted to do small electrical DIY jobs around the house, such as changing power points or light switches this has either resulted in fatal injuries or deaths. Two Victorians were killed in early 2016, when both attempted to carry out DIY electrical work on farms, about 30% of workplace deaths occur on farms due to work carried out by unlicensed electrical worker or handyman [2]. It is always essential to use a registered electrical contractor to ensure the safety of everyone, Like other profession electricians have varying license levels based on the work they are qualified to do.

Airtasker website is a Australian based company which enable users to outsource everyday tasks from an online and mobile marketplace. Users post their task and quote a budget to complete the work, community members then bid to complete the task.

The project aims to look at the tasker profiles and task listings on Airtasker website to find out their demographics, licensing and registration details of electricians or handymen who offer doing electrical jobs.

The methodology procedure in our project can be seen in the figure below.

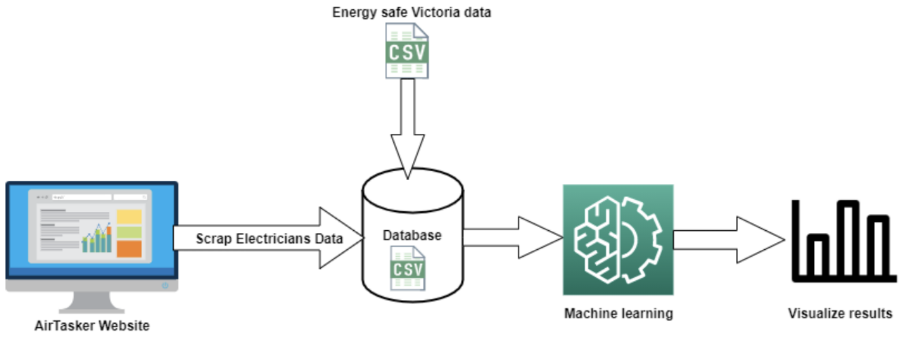


Figure 1: Process

As stated by our client, the main objectives from this project can be summarized as:

1) Web scrapping electricians tasker profiles and task listings on Airtasker website.

2) Extend the worker list by web scrapping other platform such as Gumtree or other home improvement website. Then combine and scrapped data with the client data to get the final dataset.

3) Cleaning and preprocessing of the data to get a proper electrician dataset.

4) Apply machine learning algorithm and text analysis to process, classify and model the harvested data.

5) Present the clear visualization and reasonable report for this industry project.

2 Related Work

An electrician license in Victoria, guarantees the safety of the electrical installation. Since, for each licence holder, they could carry out all types of electrical installation work in Victoria without supervision [3]. Also, there are lots requirements for the license, like completed Licensed Electrician's Assessment (LEA) and with at least 12 months electrical installation experience and so on.Thus, we could based on the requirement of the electrical license to distinguish whether electrician could guarantee the safety while doing the electrical work.Thus, from our electrician list, we could extract them out.From the airtasker website, on electrician page, there are lots of handyman doing the electrical work. They definitely have an understanding of electrical systems and repairs. Whereas, Handymen are not licensed meaning they are not permitted to work on home's electrical system.They could not be insured for damages that may occur as a result if a mistake is made while working with your homes electrical wiring. And the same time, electrician is a professional who is licensed specifically in installing, managing, and repairing electrical systems. [4] Thus, for the project, we could consider the handyman as a factor of non-incensed electrician, but we would mainly focus on the performance of the formal electrician.

One of the main tasks was the scrapping of data. Web scraping, also known as web data extraction, is a type of data scraping that is used to collect information from websites. Web scraping software can use the Hypertext Transfer Protocol (HTTP) or a web browser to access the World Wide Web. While a software user can perform web scraping manually, the word usually refers to automated procedures carried out by a bot or web crawler. It's a type of copying in which specific data is acquired and copied from the internet, usually into a central local database or spreadsheet for retrieval or analysis later [5] [6].

The methodology is divided into three parts: the web scraper extracts the desired connections from the web, the data is extracted from the source links, and finally the data is saved in a csv file [5] [6]. Various programming languages like SQL or Python can be used to carry out the task.

There are various techniques to scrape the data. We can use tools like Scrapy [6] which is a powerful Web scraping framework for Python where web pages are automatically parsed and Web contents are extracted using XPath expressions, Visual Web Ripper [7] which offers more functionality and allows users to scrape data from any website. After finishing data scraping task data can be exported to structured CSV, Excel, or XML format, Beautiful Soup [8] is another Python package for parsing HTML and XML files and extracting data. It integrates with your preferred parser to provide idiomatic navigation, search, and modification of the parse tree. It is common for programmers to save hours or even days of work.

3 Data analysis / preliminary model development

In the analysis part, we will study the relationship between worker profile information and the percentage of worker won an offer. In other words, the client wants to know that what attributes are most valuable for a worker to get a job offer. Besides, another analysis is to check whether electricians on the Airtasker have expired licenses.

There are two sources we can get the data. One is web crawling for the Airtasker. Another one is the license information of registered electricians from Energy Safe VICTORIA (ESV), which we can get from the client and compare with the personal information of workers in the Airtasker. Thus, the first step we should do is crawling the worker page.

After surfing the Airtasker website, the main steps we get into the worker page are using the ‘browse Task’ function on the main page, choosing the area we want, and then inputting the task keywords in the search bar. On each task page, we can get into the personal page of workers who want to participate in this work. Though it is a bit complex, so far, this approach is the only way to get into the personal page of workers directly by pointing the name, which is easy for us to crawl the worker page list. Therefore, the primary process of data collecting from the website is:

1) Crawl the electrical task list page.

2) Crawl the worker list from each task URL. During doing this step, the number of times each worker participates in the task assignment competition and the number of times each worker won the offer will also be counted and collected to calculate the percentage of worker won an offer.

3) Crawl the worker profile information from their URL.

3.1 Task URL List Crawling Process

In the task crawling step, the location and keywords of the task are two critical elements we should define.

As for the location part, we were first using the ‘Victoria, Australia’ as the option. However, after several grabs, we found a maximum number of 600 can show on the page even if we have been logged in the Airtasker (The website specifies that only when you are logged in you can see more tasks). So except for the ‘Victoria, Australia’ option, we also use ‘Bendigo’, ‘Shepparton’, ‘Traralgon’, ‘Harsham’, ‘Warrnambod’, ‘Wangaratta’ options as task searching area, because they are some distance away from each other. Besides, there are many ‘Melbourne’ and ‘Geelong’ tasks, so adding these two locations to the searching area can prevent missing tasks. This process will produce many duplicate values, so duplicate value eliminating has been added in the code when crawling the task page list.

Concerning the keywords selection in the search bar, for now, we only use the ‘electrician’ and ‘electric’. Expanding the keywords will be the further step we can improve our analysis.

Now we have crawled 1100 tasks. There is an overview statistical of tasks geographic distribution. Task location can get from the postcode information on the task page or poster page using the Regular Expression Matching (matching pattern with r'VIC [0-9]4' and r'[0-9]4'). However, not all tasks contain the text in postcode format. After crawling and matching, there are only 786 tasks that enable the postcode. According to the correspondence table between postcode and Statistical Area Level 4 main code, which has been well supported and illustrated by the data (Commonwealth of Australia, 2020) [9], we assigned each task to their related SA4 area.

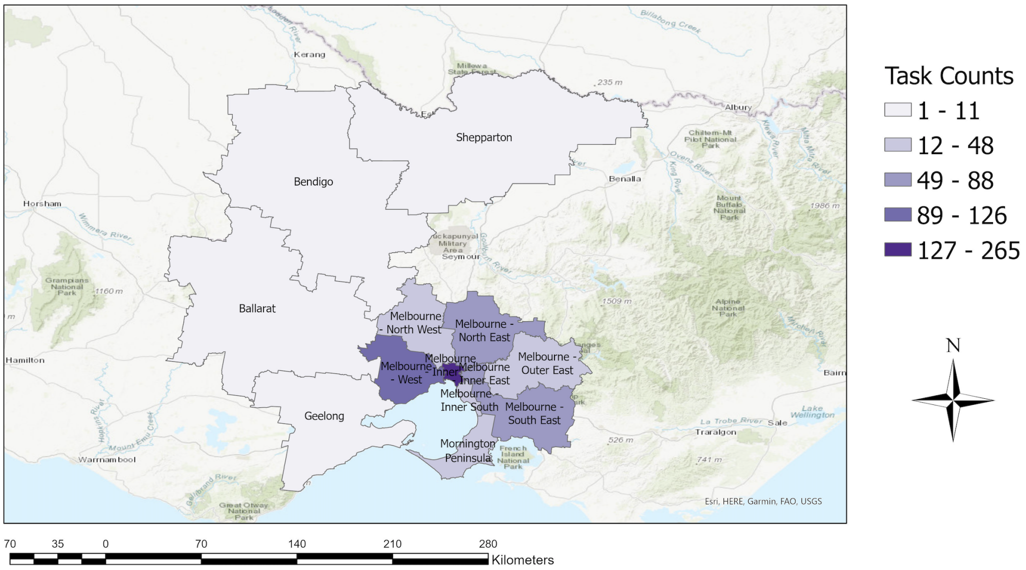


Figure 2: Spatial Distribution Map

As can be seen in the Spatial Distribution Map, We can conclude that the closer they are to the CBD, tasks appear more frequently. The inner Melbourne has the most tasks with 265, far away from other counts. This is reasonable because we only chose the in-person task.

3.2 Worker URL List Crawling Process

After determining what type of task we want, we can crawl the list of all task page URL. Next will be to crawl the worker page list on each task page. There are two sections we can get the worker page link. One is the 'OFFERS' section, showing the workers who have given the poster quotation. Another one is the 'QUESTIONS' section, consisting of the workers interested in this work and asked poster work-related questions. In general, we only need to crawl the worker list in the 'OFFERS' section because they are the poster's option.

Nevertheless, during this part, we found that only the task assigned worker remained in the 'OFFERS' area when a task has been assigned. At the same time, others in the competition will be moved to the question. To avoid missing workers, we have to determine the current status of each task when crawling the task page list.

3.2.1 Task Status Determining

'Open', 'Assigned', 'Expired', 'Cancelled' and 'Completed' are five kinds of status of the task. We only need to deal with the first four status since the last status is the same as the 'Assigned' status in our analysis. These statuses can be identified through the 'pending task state' section at the top of each task page. The specific judgment methods are shown in the table below.

Status	**Judgment**
Open	section.text = ASSIGNED & element.style = dlBcVh (section color is grey)
Assigned	section.text = ASSIGNED & element.style = kRdOrS (section color is green)
Expired	section.text = EXPIRED
Cancelled	section.text = CANCELLED

Table 1: Judgement Methods

3.2.2 Worker page URL Crawling

The only task status that needs to be treated differently is 'Assigned'. Because not only do we have to deal with the problem data in this kind of tasks, but we also have to count the number of times that workers have participated in the competition and the number of times that workers won offers. There is one thing to note in this step. The poster name also would appear in the 'QUESTIONS' section in the case of answering questions. Thus, we crawled the poster name firstly and then deleted this name when returning the worker list.

During the process of crawling assigned tasks, the worker of the 'OFFERS' section would be referred to as the won worker, then both the counts of times he won the offer and the counts of times he participates in the competitions will be increased. In comparison, other workers in the 'QUESTIONS' section only increased in the counts of times they participated in the competition. As a result, we will get two statistics about the times workers attend and won. Then, the proportion of worker won an offer, which will be the response variable in our analysis, can be calculated by using the formula:

$$\frac{counts \quad of \quad times \quad worker \quad won \quad the \quad offers}{counts \quad of \quad times \quad worker \quad participated \quad in \quad the \quad competition} \times 100\%$$

As for the tasks in another status, we only crawl the workers in the ‘OFFERS’ section. We only need their URL list when they have never been crawled before.

3.3 Extracting Data from URL’s

The URLs collected by crawling in the previous processes are used to scrap the worker and task data. There are a lot of methods to scrap the data but we have used the best method available. We scrapped all our data using the python page Beautiful Soup [5]. Beautiful Soup is best suited towards tiny and straightforward undertakings. It can collect data from websites that use JavaScript, but the web scraping methodology it employs is effective.

3.3.1 Web Scrapping for Tasks

The tasks were extracted to find a relationship between workers and tasks. The task lists indicate the person who has posted an ad on Airtasker. Approximately, 1100 tasks were extracted. The following task variables were extracted from Airtasker.

1) Task Name - The task to be done or "task advertisement" posted by someone who needs something done in their homes.

2) Address – The person’s suburb where the task needs to be done. The customers share their full address to only that worker that they choose to do repairs. In this way, data privacy is ensured by Airtasker.

3) Due Date – The deadline by which a person wants the task to be completed.

4) Availability – The time that person is present from the worker to come and work.

5) Budget – The fees they are offering for the worker to complete the job.

6) Description – The details about the task i.e., what needs to be repaired, where it needs to be repaired and so on.

3.3.2 Web Scrapping of Worker Profiles

In total, a 1000 worker profile URLs were scrapped and from all those websites, the information about the workers as mentioned in their profiles were extracted.

1) Name – The name of the worker.

2) Address – The suburb name and state from where the worker comes from.

3) Member Since – The date and year of that person joining Airtasker.

4) License Number – All professional workers have a licence number. A licence number indicates that they have been verified by the government to do the tasks. The license numbers extracted were of two types:

   - Numbers starting with REC (e.g., REC 11224) - Registered Electrical Contractors (REC) are required by Energy Safe Victoria to operate in a safe and compliant way. Applicants must have the necessary experience, technical and business knowledge, and proven ability to perform electrical installation work in order to be registered as a REC.

   - Numbers starting with A (e.g., A14352) - An A grade electrician's licence allows the holder to perform all forms of electrical installation work without supervision in Victoria. The holder of this license is not permitted to do electrical installation services for profit or reward For this, the electrician should also hold a Registered Electrical Contractors (REC) license.

   - For all other workers (if present), the licence number was just a set of 5 digit number.

5) Description – This free-form text field allow the workers to bring all of the elements of their profile together to express themselves where they can describe their personality and how they can contribute in the best way possible. Every worker has a short description about themselves where they introduce themselves stating their qualifications, experience and skills.

6) Ratings - Number of stars each electrician has been given by the customer out of 5.

7) Tasks Completed – The total tasks completed by the electrician.

8) Reviews – Number of reviews that electricians have received from all the customers they served.

9) Completion Rate – The completion rate is the percentage of tasks that have been completed out of the total number of tasks that have been allocated to someone. This is calculated for any worker based on the number of jobs performed minus the number of offers accepted.

10) Education – The education qualifications of the worker.

11) Specialties – Specialties include all the skills that a worker is proficient at. Some workers at Airtasker are skilled at more than one category. For example, one worker can be an electrician as well as a handyman doing jobs like fitting a television set etc.

12) Language – All languages by which a worker can communicate efficiently.

13) Experience – the professional work gained via full-time employment in a field connected to their work or in which the person is or intends to be licensed.

14) Transportation – The ways in which a worker can travel to work.

3.4 Data Cleaning

The task dataset was fine, with no missing values and duplicate data. Therefore, no cleaning was required for that. For the worker list a lot of cleaning had to be done to get the data ready for the next steps.

1) Irrelevant Data – The data extracted from Airtasker was all mixed up. The workers were not all electricians but a mixture of handymen, industrial workers, window cleaners etc. All these were filtered out and removed. (Exception - Some workers have combined lists of tasks for example, the worker can do electrical as well as handyman jobs. Therefore, we have kept these cases in our data).

2) Missing Data – Not all workers had all information present, therefore the missing texts were left blank and the missing numbers were marked as 0. Also, around 30 workers had no information present in their profiles. Just their names and address were mentioned. We were not able to identify what type of tasks exactly do these workers do. Therefore, these were removed.

3) Duplicate Observations - There were some duplicates present which were removed.

4) Remove white spaces - Extra spaces at the beginning or end of the strings were removed.

5) Formatting of Numerical Data - The numbers extracted were in string format by default and that was converted back to numeric form.

6) License Number Extraction - License Number is one of the the most important features in our dataset. But, extracting this was not an easy task. This is because the number was added to the HTML dynamically from a Boostrap JSON. So, some extra steps was required to get the required data.

3.5 Exploratory Data Analysis

After cleaning, out of 1100 data totally extracted, we had 723 worker data left. After this, we performed some basic exploratory data analysis to understand our worker dataset.
We first saw the almost half of the workers have more than a 4.5 star rating and had a number of positive reviews. In the graph below we can see that 455 workers out of a total of 723 workers have a 5 star rating indicating that the customers were satisfied with the job done.

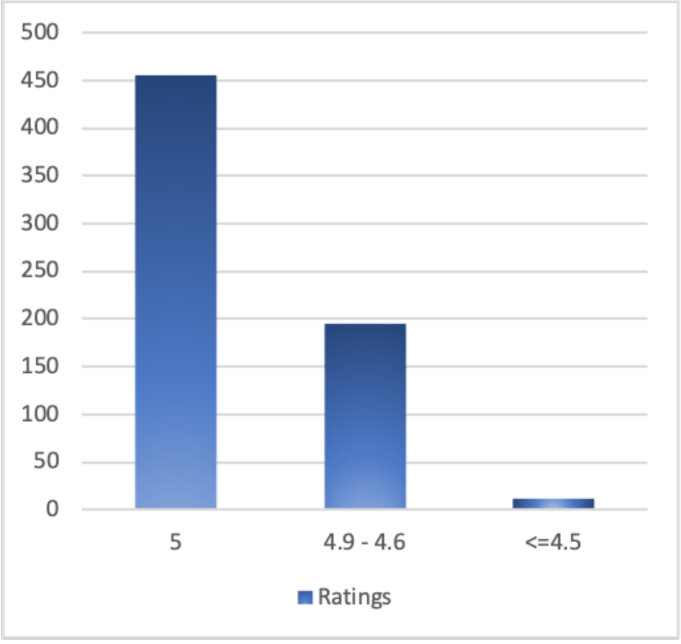


Figure 3 : Worker Rating by Customers.

We also observed that the workers have a good completion rate of tasks with about 507 workers having a percentage between 80 to 100.

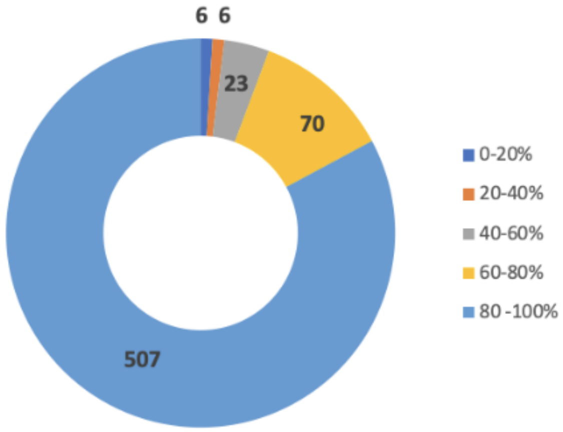


Figure 4 : Task Completion Rate for Workers

From all the workers, we had to filter the workers who do some sort of **electric** work. They maybe be professional electricians or handymen who have the ability to do basic electrical work. From all of the worker data, only 153 such electric workers were present.

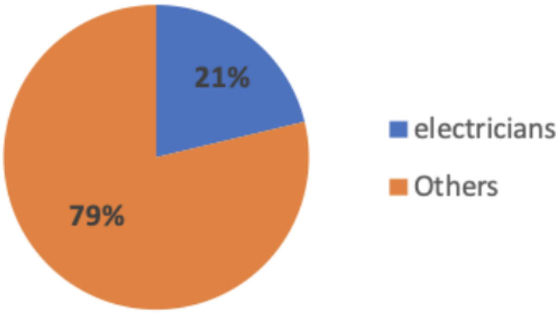


Figure 5: Total vs Electrical Workers Ratio.

The licence number gives an idea of whether all electrical workers are verified or not. But, not all electrical workers have a licence number in their profiles. Out of 153 electrical workers only 68 of them have a licence number present.

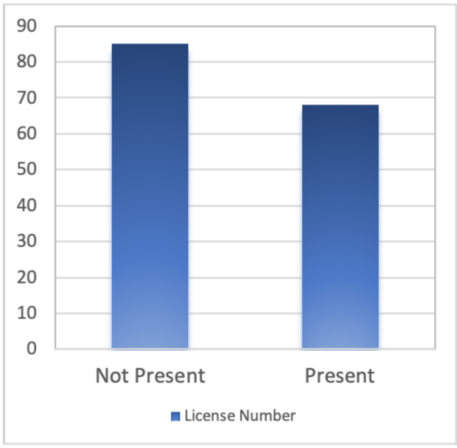


Figure 6: License Numbers

4 Proposal for semester two

4.1 Expand the data

We have only crawled 1100 tasks and 723 workers in the Airtasker because of the lack of searching keywords. Next semester, we intend to overcome this problem through Natural Language Process (NLP) in two main ways.

Firstly, expand our searching keywords according to the task name we have crawled. These task name would include the task list we have crawled and the historic tasks we can crawl in each electrician personal URL. Then, we will get a task name list related to the electrical works. The initial plan for processing this task name list is as follow:

1) Use the Part-of-speech tagging analysis in the NLP to tag and extract the noun of each task name.

2) Use the Stemming analysis in the NLP to change words to a base form. This step is mainly used to solve the problem of the plural form being different words because we only have the noun.

3) Count up the number of occurrences of the same word to analyze the electrician doing the most work. For example, 'socket' and 'light' will be the reasonably expected output in this step.

Using the right keywords will reduce the filtering effort, but there will still be many non-electrical works. These electrical works cannot be filtered by using the task description because they should have been crawled using the current keywords (electrician and electric). The only thing we can do is to crawl all the workers' URL through the task list and then filter by the workers' profile. The initial plan for processing this workers' profile is as follow:

(1) Crawl the section of the individual resume of each worker, consisting of the 'ABOUT' part (self-introduction) and 'SKILLS' part (EDUCATION, SPECIALITIES and WORK).

(2) Use the Part-of-speech tagging analysis in the NLP to tag and extract the noun list and adjective list.

(3) Use the regular expression to match each noun word and adjective word. We can use the r'electri?' as the match pattern. Once a matching word appears, we can assume that the person is an electrician.

(4) Crawl all the matched electricians' historical task list and add them to our electrical task list.

In our imagination, the two approaches are interlaced. This is our temporary improvement plan, and we will continue to optimize our methods, such as the match pattern and NLP usage, in the coming days.

Even after doing text analysis we do not get sufficient data we would have to expand our search to other websites for example Gumtree to get more electrician data.

4.2 Modelling

After getting our entire dataset, we need to find a relationship between the workers and the tasks. For this we will run our data through various models and co-relation tests.

4.2.1 Supervised Learning

We need to develop a classification model between the workers, the number of offers they receive and the total tasks completed by the worker. This can be done using logistic regression. [10]

For the classification problem, we will utilise the Logistic Regression model, which is basic yet effective. When the dependent response variable is binary, the model is utilised (potential or not). It uses a logit link function to connect the explanatory characteristics to the output class, as seen below:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 * X$$

To find a relationship we can use various tests between features in Machine Learning. We can use the covariance or correlation tests to determine the relationships.

Using the features like age, education, skills, and taking the name as the response variable,etc we can also determine the demographic of the workers using supervised learning models [11].

4.3 Unsupervised Learning

We can use clustering techniques to classify the electrical workers into 3 groups:

1) Licensed Electricians - The electric professionals who had a valid license ID present in their profiles.

2) Unlicensed Electricians - The electric professionals with no licence numbers present.

3) Combined Workers - The workers who can do electrical as well as other handyman tasks.

The idea is to see which group is doing the most tasks and getting the most number of offers and which type of workers do the customers prefer to do their tasks [12].

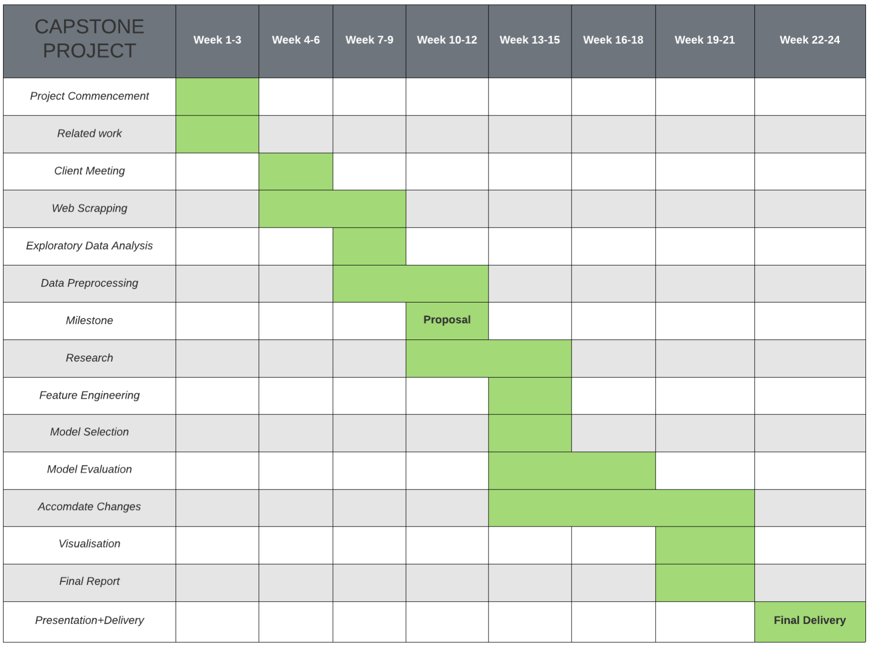
A geographic analysis could then also be determined to see where does each group mostly located and where do they do most of their tasks.

4.4 Visualization

Finally, when the model is made, we have to create a dashboard using either PowerBI or Tableau to display our work. The dashboard would display the demographics and the results we find after developing our supervised and unsupervised models.

5 Timeline

The Gantt chart represents the timeline of our project, the entire project is divided into two stages,



CAPSTONE PROJECT	Week 1-3	Week 4-6	Week 7-9	Week 10-12	Week 13-15	Week 16-18	Week 19-21	Week 22-24
Project Commencement	🟩							
Related work	🟩							
Client Meeting		🟩						
Web Scrapping		🟩	🟩					
Exploratory Data Analysis			🟩					
Data Preprocessing			🟩	🟩				
Milestone				Proposal				
Research				🟩	🟩			
Feature Engineering					🟩			
Model Selection					🟩			
Model Evaluation					🟩	🟩		
Accomdate Changes					🟩	🟩	🟩	
Visualisation							🟩	
Final Report							🟩	
Presentation+Delivery								Final Delivery

Figure 7: Gantt Chart

1. **Stage 1:** Planning, define requirements and initial exploration of data, web scrapping data from Airtasker website, merging scrapped data with ESV data, exploratory data analysis.

2. **Stage 2**: Feature engineering, model selection, model evaluation, visualization using Tableau or Power BI, testing and deployment, the final deliverable will delivered to the client by end of semester two.

6 Team Contribution

Name	*Roles
Chen Zhou	Feature engineering, Visualisation
Meghna Panda	Web Scrapping, Data Processing and Visualisation
Vignesh Lakshminarayanan	Model selection, Visualisation
Xinyu Mao	Web Scrapping,Data Processing, Model Selection

*This can change with the changing project requirements.

References

[1] S. Kiriaki. Victorian man convicted after carrying out electrical work while unlicensed. [Online]. Available: https://esv.vic.gov.au/news/victorian-man-convicted-after-carrying-out-electrical-work-while-unlicensed/ 1

[2] D. Ranger. Electrical ddiy - don't do it yourself. [Online]. Available: https://esv.vic.gov.au/campaigns/ddiy/ 1

[3] A. Robertson. Electrician's licence (a/ a/e). [Online]. Available: https://esv.vic.gov.au/licensing-coes/electrical-licences/electricians-licence/ 2

[4] Electrician vs. handyman: What is the difference? [Online]. Available: https://prolineelectric.ca/difference-between-electrician-vs-handyman/ 2

[5] D. M. Thomas and S. Mathur, "Data analysis by web scraping using python," in *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2019, pp. 450–454. 2

[6] M. E. Asikri, S. Krit, and H. Chaib, "Using web scraping in a knowledge environment to build ontologies using python and scrapy," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 3, pp. 433–442, 2020. 2

[7] A. V. Saurkar, K. G. Pathare, and S. A. Gode, "An overview on web scraping techniques and tools," *International Journal on Future Revolution in Computer Science & Communication Engineering*, vol. 4, no. 4, pp. 363–367, 2018. 2

[8] C. Steven, "Web scraping wikipedia using python and beautifulsoup," 11 2019. 2

[9] AustraliaGovernment, "Asgs geographic correspondence," 2018. [Online]. Available: https://data.gov.au/data/dataset/23fe168c-09a7-42d2-a2f9-fd08fbd0a4ce 3.1

[10] A. Linden and P. Yarnold, "Some machine learning algorithms find relationships between variables when none exist – cta doesn't," 03 2019. 4.2.1

[11] S. Idowu, "Analysis of population data using multiple and logistic regression model," 04 2019. 4.2.1

[12] K. Chen, W. Zhang, H. Zhao, and H. Mei, "An approach to constructing feature models based on requirements clustering," in *13th IEEE International Conference on Requirements Engineering (RE'05)*, 2005, pp. 31–40. 4.3