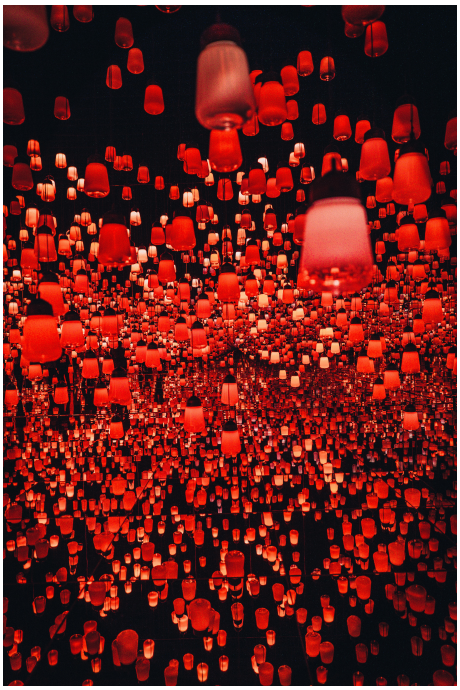


Report: Australian Census Data

Mohammed Faizan
MBAt

Adarsh More
MBAt

Yanhui LI
MBAt

Report for
Monash University

2 June 2021

Our consultancy - Star Wars
Mohammed Faizan &
Adarsh More&
Yanhui LI

(03) 9905 2478
questions@company.com

ABN: 12 377 614 630

Contents

1 Introduction

The Australian Bureau of Statistics(ABS) conducts the census for Australia every 5 years which includes all people present in Australia on the census night irrespective of their nationality. Wikipedia contributors (2021) defines census as "systematically calculating, acquiring and recording information about the members of a given population. This term is used mostly in connection with national population and housing censuses." A census aims to include the entire population as supposed to sampling and therefore data is recorded for every individual. However, when this data is released in public for interested institutions such as businesses, other government organizations, NGOs and other researchers it is only ethical to de-identify the data. Ethics has always been argued for risk versus benefit. With census data capturing personal details it must be de-identified and therefore ABS makes it available after perturbation as aggregated dataConfidentiality.

Census being a population data is able to capture insights about small geographic boundaries and demographics precisely. The 2016 Census data was output using the 2016 Australian Statistical Geography Standard (ASGS).The ABS Structures are a hierarchy of areas developed for the release of ABS statistical information. This statistical information represents data for all census geographies from Australia down to Statistical Area Level 1.@CensusDefinition say "Data can be represented

visually or analyzed in complex statistical models, to show the difference between certain areas, or to understand the association between different personal characteristics."

Our report is based on 2016 census data from the Australian Bureau of Statistics(ABS). In 2016, Census collected data for 10 million dwellings and approximately 24 million people, the largest number counted to date. The report dwells on the SA4 regions of Victoria and the topics for analysis are the Field of Study, Education Qualifications, Industry of Employment and Occupation. We try to determine the association between these topics based on age and gender. To further support these association insights, data from the Victorian Public Sector Commission(VPSC) is included.

2 Data Preparation

The census data was not in accordance to the tidy data definition and was spread across multiple files. Datapacks are provided in CSV format. Geopacks include comprehensive data files and associated Geographic Information System (GIS) boundary files in a format suitable for loading into proprietary software and/or client custom-built systems. Hence appropriate cleaning was performed and cell values were renamed in a more human readable context.

3 Victoria Population: Overview

The map 1 is population density map which shows the population concentration in each of the SA4 regions. The ABS divides the geographical areas on the basis of the population density such that each region has comparable densities irrespective of their area/size of the region. There are nineteen SA4 regions in Victoria each represented on the map with a number starting with the state code of *Victoria: 2*. Ten regions out of nineteen exist in Melbourne city suggesting that most people in Victoria reside in Melbourne and the country side of Victoria is sparsely populated. Precise numbers are present in Table 1. Regions in Melbourne have higher population with exception of region 297 and 299 having a population of 9 and 1994 respectively. Male and female population is comparable in all regions, however male population was higher in all regions.

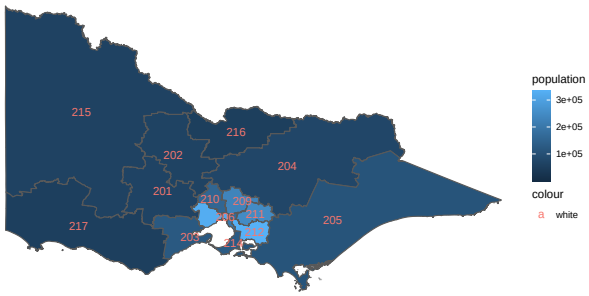


Figure 1: *Map: Victoria Population*

Age distribution for Victoria is represent by the density plot, Figure 2.

- Most population is Middle Aged, 20 to 50 years.
- Old people are vulnerable with a low population.
- Age distribution is similar for both male and female population.

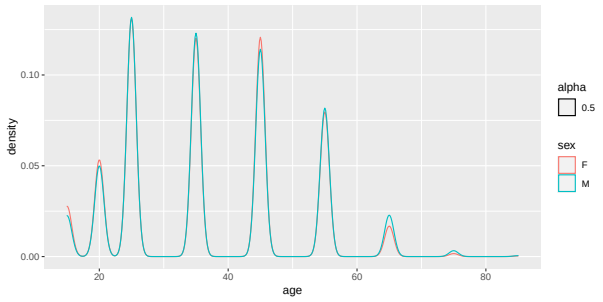


Figure 2: *Age Distribution*

Table 1: *Victoriqn Population*

SA4_CODE_2016	femalepopulation	malepopulation	population
201	32726	34691	67417
202	32396	34054	66450
203	60660	64307	124967
204	35934	39614	75548
205	52929	57572	110501
206	159362	160819	320181
207	81814	86786	168600
208	96482	101671	198153
209	109370	122195	231565
210	71224	85167	156391
211	118179	129501	247680
212	151481	184164	335645
213	147830	178340	326170
214	62731	68190	130921
215	29867	33492	63359
216	25915	28796	54711
217	26236	29297	55533
297	0	9	9
299	765	1229	1994

4 Victoria Population: Gender

To study the association between the topics of study, the figure 3 is a bar chart that shows the population in the sub-divisions of each topic arranged in decreasing order representing both male and female population. The following inferences were made by comparing the individual plots that each represent a topic of study.

- Highest people are are Health Care Professionals and the ratio between men to women is less than one.
- Similarly, in construction more men are employed as laborers.
- The population of women in the education sector is far exceeds that of men.
- Management & Commerce is the field that the most population have studied.
- More men have studied Engineering and Technology as compared to females. However, more people are employed in Health Care than in industries relating to Engineering.
- More women have studied Management and Commerce, however more men are employed as managers.

- Victorian population is educated upto level 7 and most are employed as professionals.
- However, a large population is employed as labourers when the population share of people who studied below high school is very less.
- Most of the residents achieved the level 7, which refers to the bachelor degree, and there are almost twice as many female as male.
- Majority of male residents achieved at the level 3 and 4.
- [GenderLinearModel] shows the relationship between male and female populations

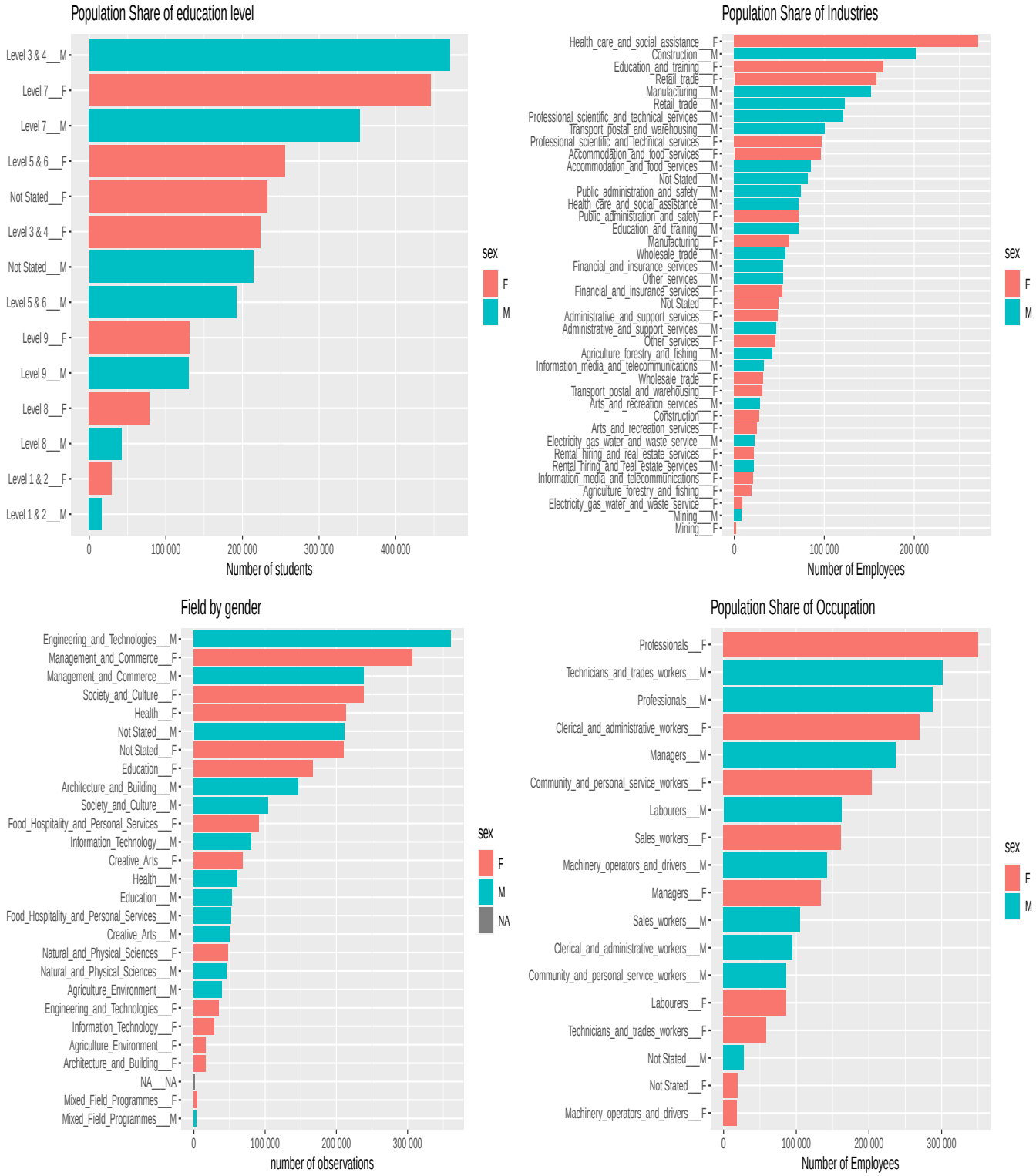


Figure 3: *Population: Gender*

The Figure 4 shows the total populations for sub-divisions within each topic.

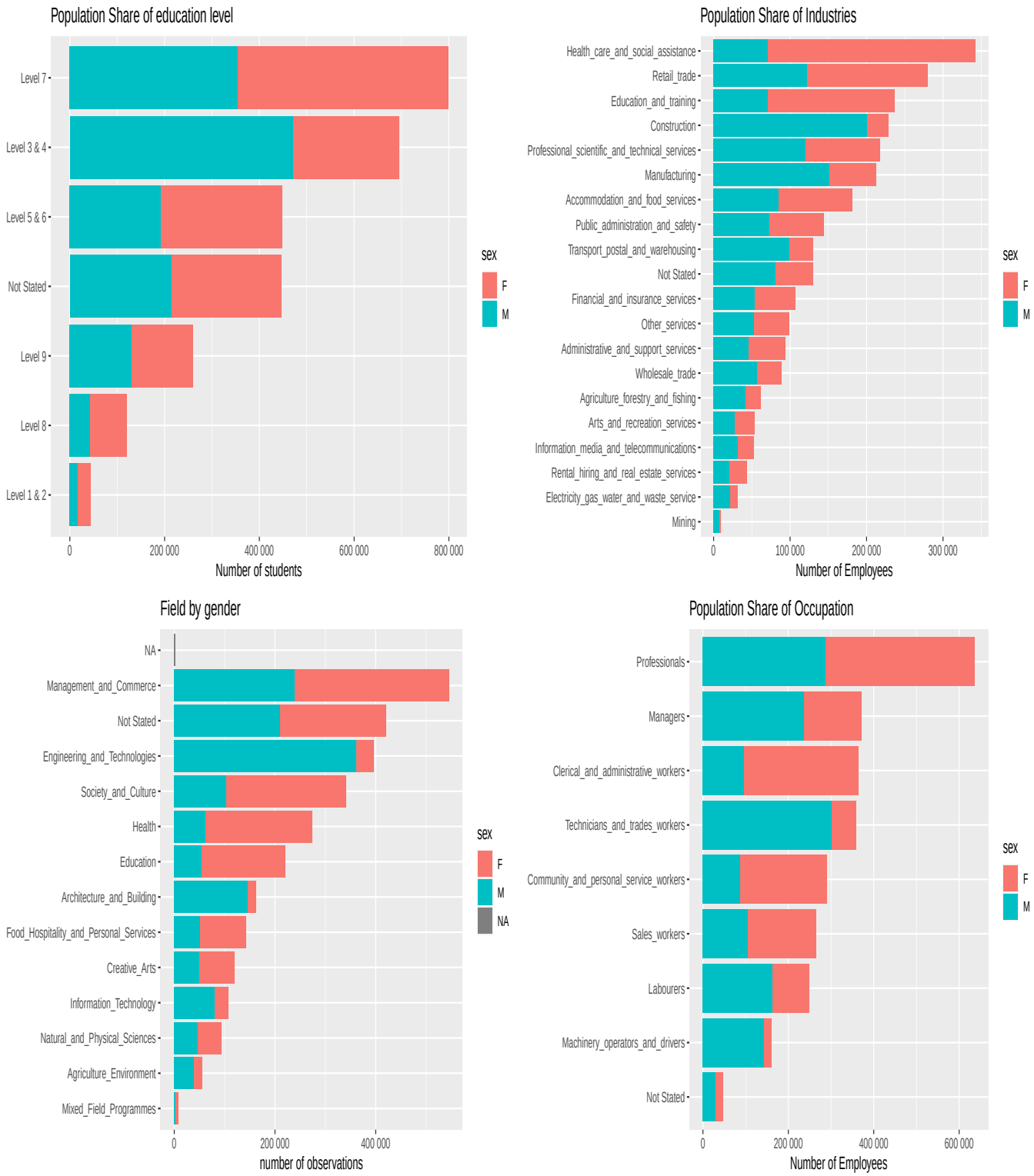


Figure 4: *Population: Topic*

Male vs Female

A linear model to determine the male to female population ratio is fit for the male and female population with respect to occupations and industries. A line(slope=1, black line) decides the value of this ratio. The models above this line have a higher female population and the models below this line have a higher male population. These models are shown in Figure 6 and Figure 8.

In Victoria, for any particular education level, more women have achieved it than men. More women are educated(ratio F:M):

- Graduate diploma and graduate certificate(18:10),
- Advanced Diploma and other diploma(13:10),
- Post graduate(11:10)
- Under graduate(12:10)

Whereas, for people having a qualification of certificate level 3 and 4, men far exceed women.

Male and female population is comparable with respect to professionals(a major occupation), however females have a higher ratio.

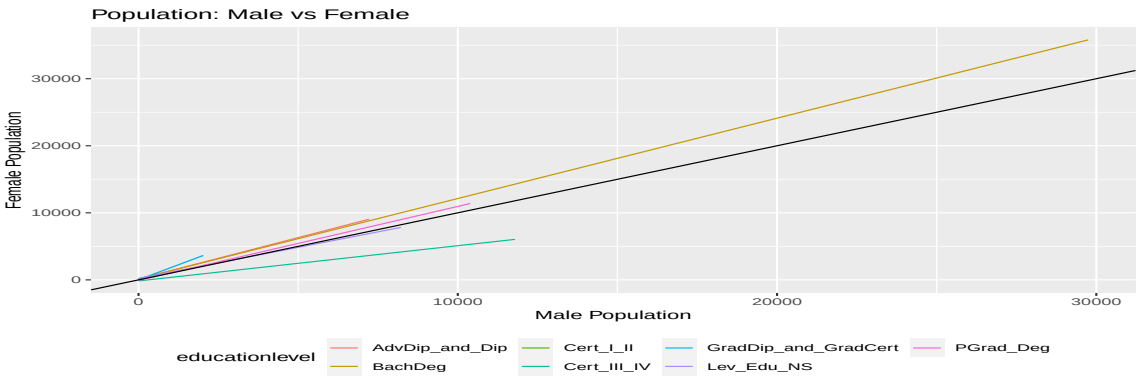


Figure 5: *Education Level: Male vs Female(linear model)*

More women are employed as(ratio F:M):

- Clerks(24:10),
- Community and personal service workers(21:10),
- Sales workers(14:10)

Whereas, more men are employed as(ratio M:F):

- Managers(17:10)
- Probable reason is the low education levels of men as seen above and their field of study.
- Labourers(19:10)
- Technicians and trades worker(52:10)
- Machinery operators and drivers(70:10)

Male and female population is comparable with respect to professionals(a major occupation), however females have a higher ratio. Probable reason is the low education levels of men as seen above.

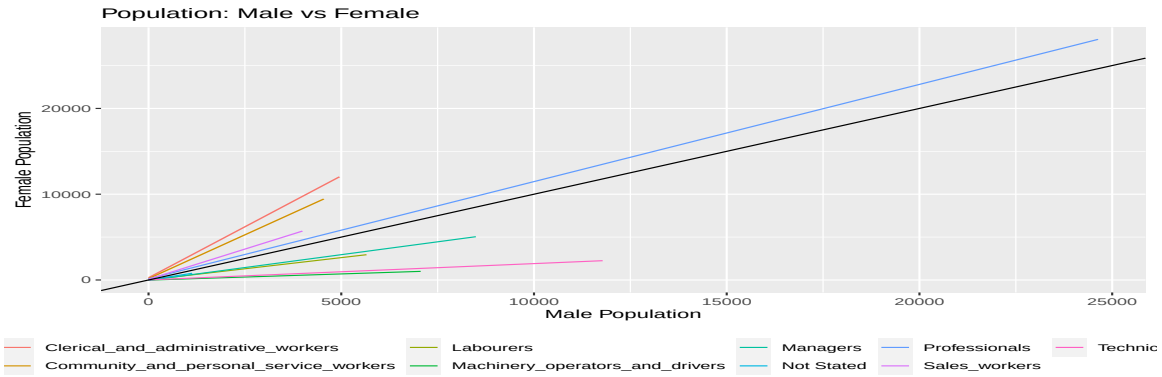


Figure 6: *Occupation: Male vs Female(linear model)*

More women studied in many fields of which most important are(ratio F:M):

- Education(32:10),
- Health(29:10),
- Society and Culture(21:10)
- Food, Hospitality and Personal Services(15:10)
- Creative Arts(14:10)
- Management and Commerce(13:10)

Whereas, more men studied(ratio M:F):

- Engineering and Technologies(94:10)
- Information Technology(29:10)
- Agriculture and Environment(23:10)
- Architecture and Building(67:10)

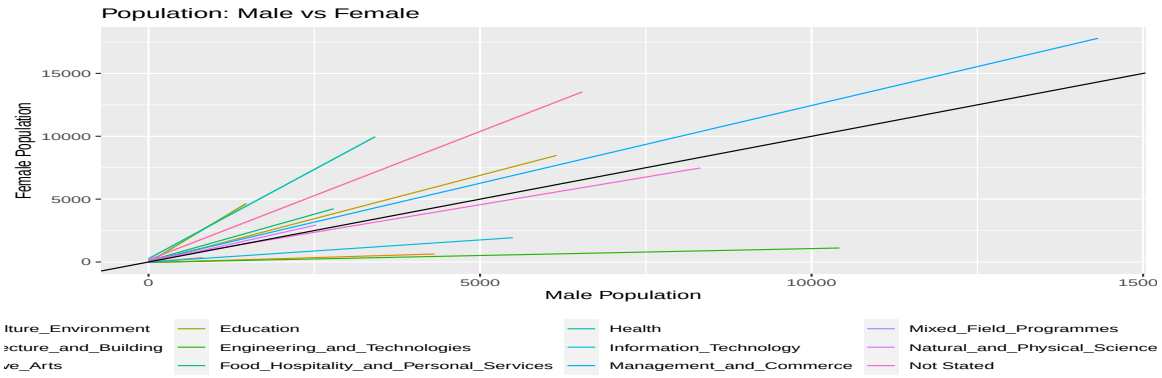


Figure 7: *Field: Male vs Female(linear model)*

More women are employed in(ratio F:M): Probable reason is their field of study.

- Health care and social assistance(30:10),
- Education and training(22:10),
- Retail trade(50:40)

Whereas, more men are employed in other industries of which most significant are(ratio M:F): Probable reason is their field of study.

- Construction(17:10)

- Manufacturing(19:10)
- Transport postal and warehousing(52:10)
- Machinery operators and drivers(7:1)

Male and female population is comparable in Administrative and support services, Financial and insurance services, Accommodation and food services and Professional scientific and technical services, however males have a higher ratio.

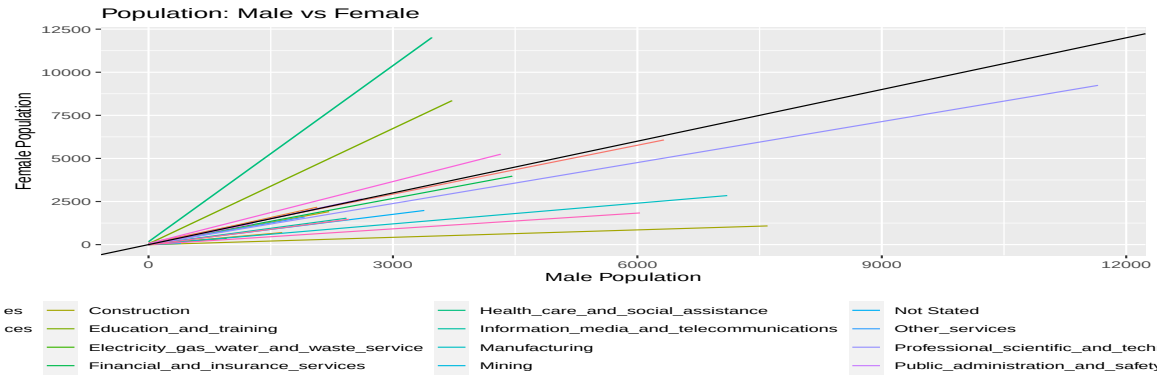


Figure 8: *Industry: Male vs Female(linear model)*

5 Victoria Population: Age

A network plot is a representation of the relationship between vertices and the strength of this relationship is determined by the edge weight(width, opacity, length, etc). In order to understand the relation between the topics and the age, the Figure 9 is a network plot representing the relationships between the sub-divisions within each topic and the age of the people. The edge weight is determined by the population size that belongs to that connection.The network graphs are based on the population and distribution can be compared only within each age group since different age groups have different populations.

First inferences from these networks are:

- Highest population within each sub-division for
  - Education level is Bachelor degree
  - Education field is Management and commerce
  - Industry is Health care and social assistance
  - Occupation is Professionals
- As seen from the age distribution, all sectors have people in the age group 25 to 45.
- The age group, 25-35 shares the highest population in every sector.
- A key observation is that some people aged over 75 are still working.
- The size of not stated educational level for all age group are similar.
- Other popular fields of study are engineering and technology, health and society and culture and education.
- Industries
  - In industries, a large part of residents who work in Health care and social assistance are in the age group of 25-35 as it has the thickest link on the graph.
  - Also, the manufacturing, professional scientific and technical services , retail trade and education and training have a similar size.
- Occupation:

- In occupation, for young people in the age group of 15-24, most of them are working as community and personal service workers, sales workers and some of them are laborers. This might because most of them are still looking for part time jobs in this age and in relatively speaking, requirements of part-time jobs are not as strict as the full-time jobs.
- A key observation is also founded that some seniors who aged over 75 are still working from this plot.

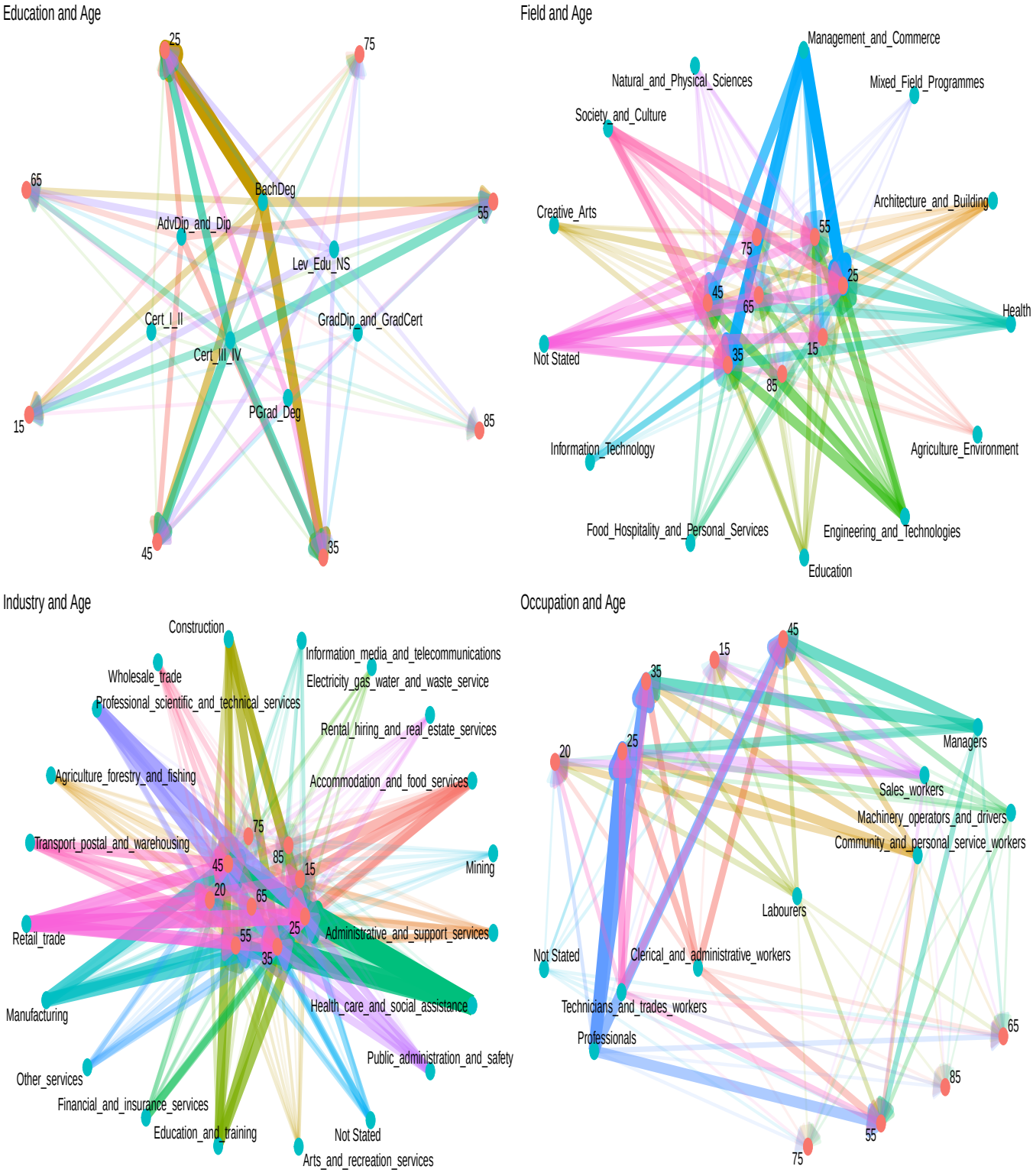


Figure 9: *Network: Topic and Age*

The following tables presents the age group with highest population for every sector and the age group [25-35) is found to be dominating every sector owing to the fact that this age group is the highest population of Victoria.

Table 2: *Education: Population*

afq_level	age_min	population
Level 1 & 2	15	9402
Level 3 & 4	25	146297
Level 5 & 6	25	96920
Level 7	25	245613
Level 9	25	83204
Not Stated	25	70455
Level 8	35	28908

Table 3: *Industry: Population*

industry	age_min	population
Accommodation_and_food_services	25	42103
Administrative_and_support_services	25	23086
Arts_and_recreation_services	25	13149
Construction	25	61959
Electricity_gas_water_and_waste_service	25	8039
Financial_and_insurance_services	25	32021
Health_care_and_social_assistance	25	80994
Information_media_and_telecommunications	25	14702
Not Stated	25	29901
Other_services	25	24089
Professional_scientific_and_technical_services	25	64125
Rental_hiring_and_real_estate_services	25	11796
Retail_trade	25	61803
Mining	35	2441
Wholesale_trade	35	22199
Education_and_training	45	56125
Manufacturing	45	55206
Public_administration_and_safety	45	37747
Transport_postal_and_warehousing	45	32663
Agriculture_forestry_and_fishing	55	12733

Table 4: *Field: Population*

field	age_min	population
Mixed_Field_Programmes	15	1813
Architecture_and_Building	25	42510
Creative_Arts	25	40334
Food_Hospitality_and_Personal_Services	25	42938
Health	25	67630
Information_Technology	25	37535
Management_and_Commerce	25	150571
Natural_and_Physical_Sciences	25	22171
Not Stated	25	71440
Society_and_Culture	25	80932
Agriculture_Environment	35	13016
Engineering_and_Technologies	45	77524
Education	55	44696
NA	NA	896

Table 5: *Occupation: Population*

occupation	age_min	population
Community_and_personal_service_workers	25	67104
Not Stated	25	11075
Professionals	25	190449
Sales_workers	25	51772
Technicians_and_trades_workers	25	99110
Managers	35	100601
Clerical_and_administrative_workers	45	89021
Labourers	45	49653
Machinery_operators_and_drivers	45	40922

6 Victoria: SA4 Sectors

The bar plots represent the SA4 regions and its working population with respect to their education levels, field of study, industry of employment and occupations. Each plot shows the region which had the highest population belonging to that subdivision.

- It can be observed that the region 206 had the most number of people with highest education levels which justifies that highest number of people in region 2016 were employed as professionals in their respective industries.
- Management and commerce, engineering and technology were the fields of study for most population and agriculture, environment and mixed field programs had the least population share.

- Health care, manufacturing and retail trade were the industries with most population while people were employed most for occupations of Professionals and Managers.

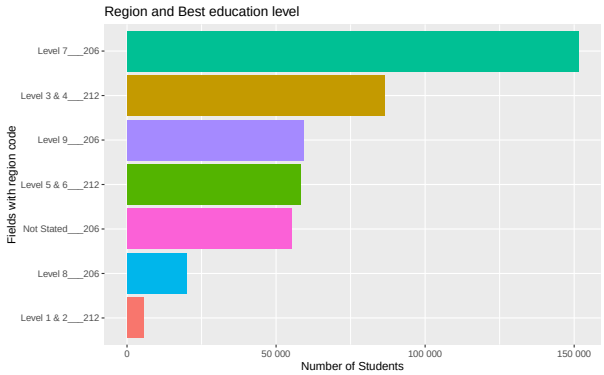


Figure 10: *Education Level: Region*

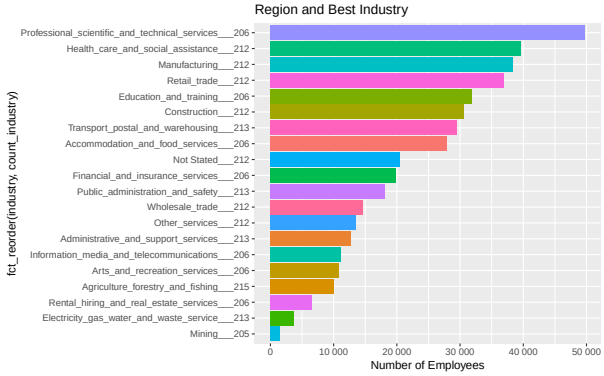


Figure 11: *Industry: Region*

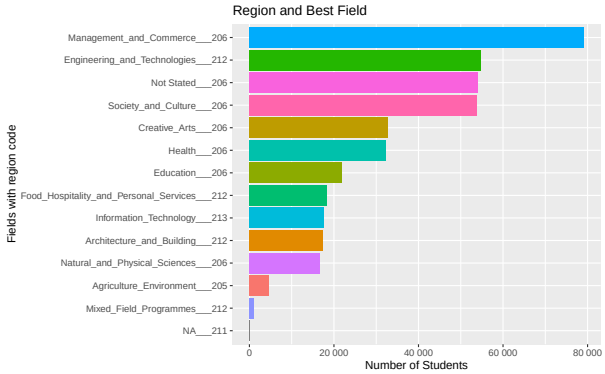


Figure 12: *Field: Region*

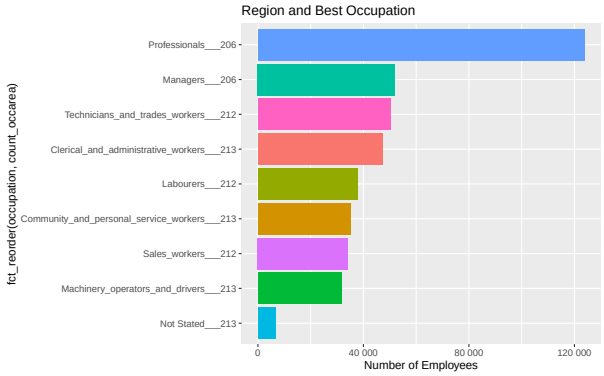


Figure 13: *Occupation: Region*

Maps

The maps represent the SA4 regions and the distribution of population by their education levels, industries, field of study and occupations respectively.

- Most population has completed education level 7 with management and commerce as their respective fields of study.
- It can be observed that the highest number of people are employed in the occupations: Professionals, Managers and Technicians and trade workers.
- Major industry in the city side is health care and the country regions are more operational in agricultural activities.

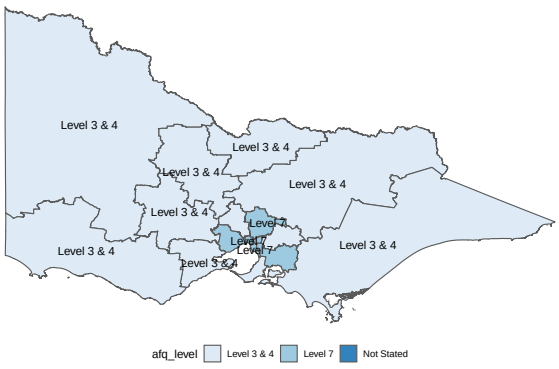


Figure 14: *Spatial Education Level Distribution*

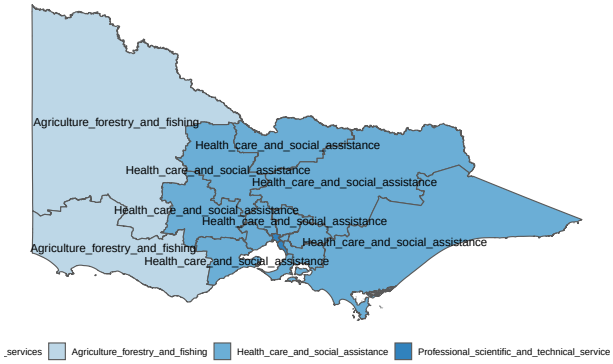


Figure 15: *Spatial Industry Distribution*

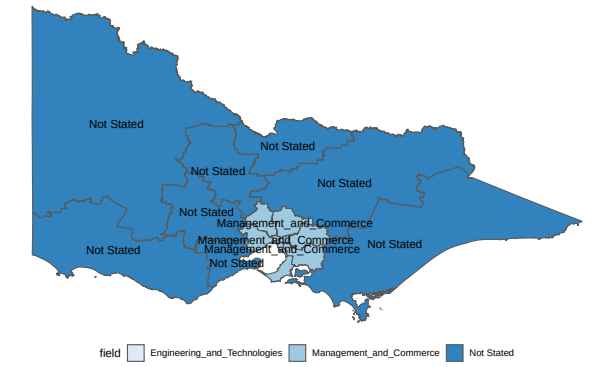


Figure 16: *Spatial Study Field Distribution*

Networks

The network graphs represent the relationship between the SA4 regions and the population with respect to the education levels, Industry, Field of Study and Occupations. For analysis, we grouped the SA4 region codes with education levels, industry, occupation and fields of study, then created a data frame for nodes and joined both data frames to create the network graphs.

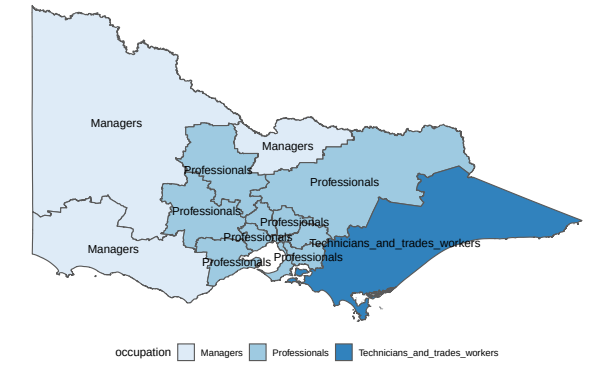
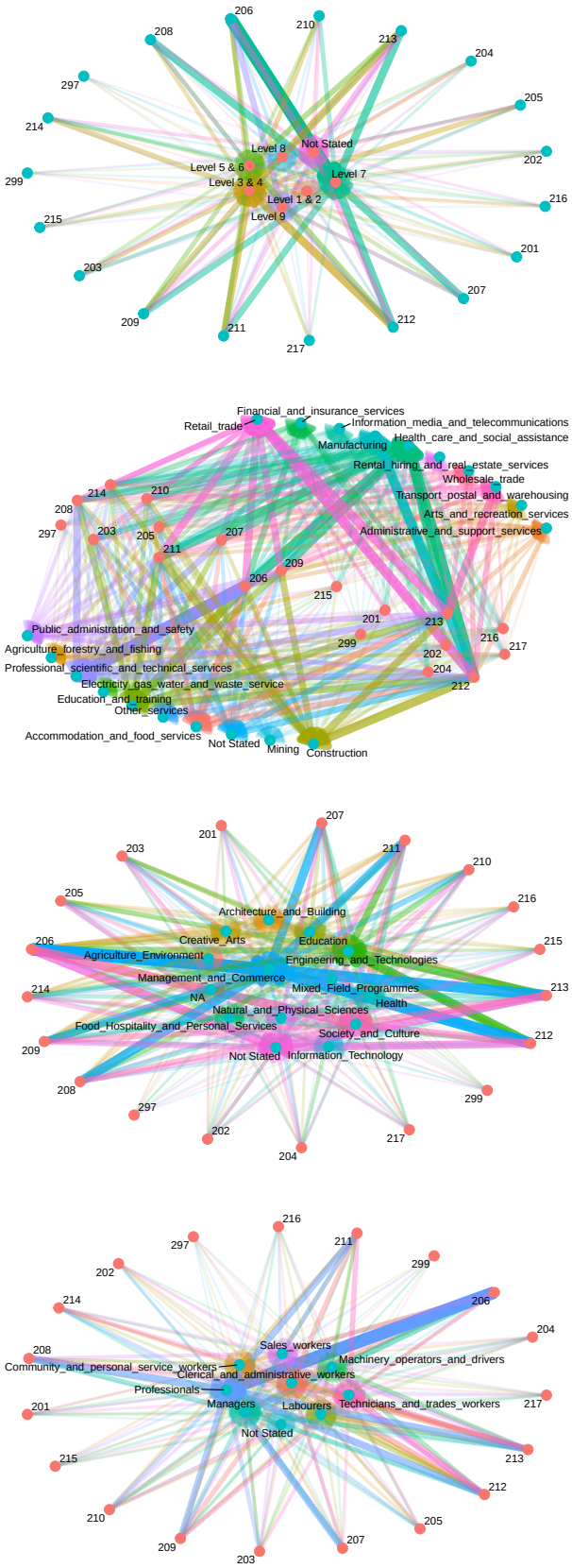


Figure 17: *Spatial Occupation Distribution*



From the network graphs it can be observed that - Region 206 has most population for all sectors throughout the graphs. - Highest number of people have studied Management and Commerce. - For

Industries, Health care and Professional and scientific services accounted for highest population. - Highest number of employees were employed as Professionals and Managers.

7 G52 and G58

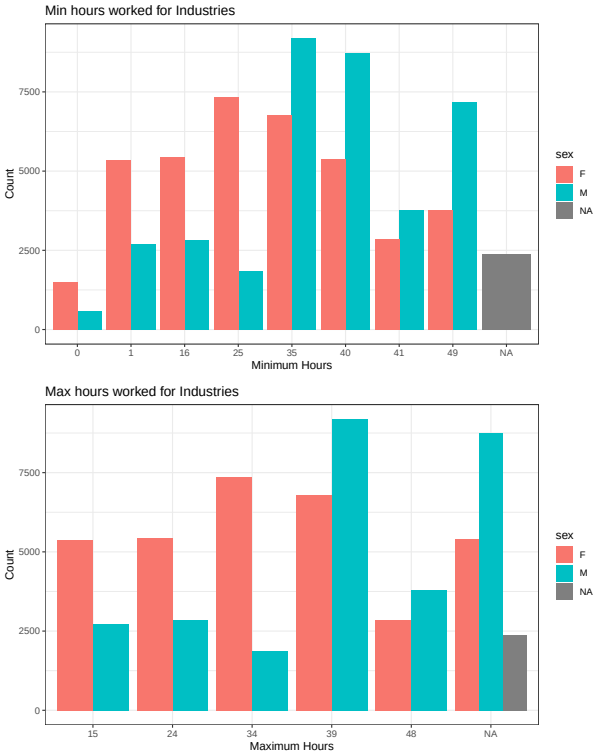
This section of the report represents the population of the SA4 regions aged 15 years and above. The population is represented by their Occupations and the Industries they are employed in and also by Sex and weekly working hours. The purpose is to determine whether males or females have worked for more hours, which industries have highest working population and region-wise which occupations had most number of employees.

For the analysis, the tables G52 and G58 were cleaned and relevant variables were renamed. For the cleaning process we excluded the total and population variables, converted the data frame into longer format and joined it with the geopath for SA4 geomap. We then summarized the minimum and maximum hours of the population working for industries as well as employees in occupations. The summarized data was plotted to compare the minimum and maximum working hours by gender.

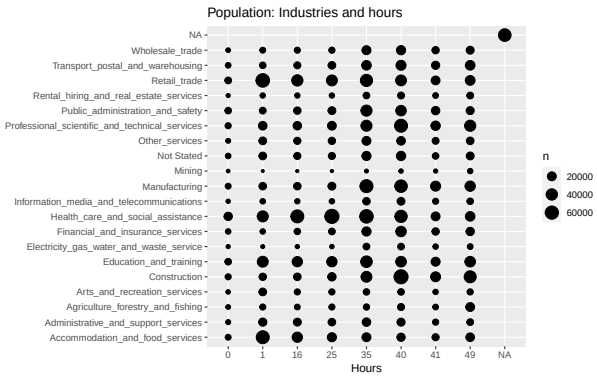
We used the summarized data to plot the minimum hours worked by people with respect to each industry in a count plot. Furthermore, we summarized the number of people and grouped them by the SA4 codes for each region and arranged the data sets in descending order to plot region-wise bar plots representing the regions, education level, field of study, industries and occupations.

Lastly, we plotted the distribution of population by their education levels, industries, field of study and occupations on 4 maps respectively. The analysis involved mapping the geometric values of the regions on the sa4 geomap with respect to total population and the respective industries and occupations.

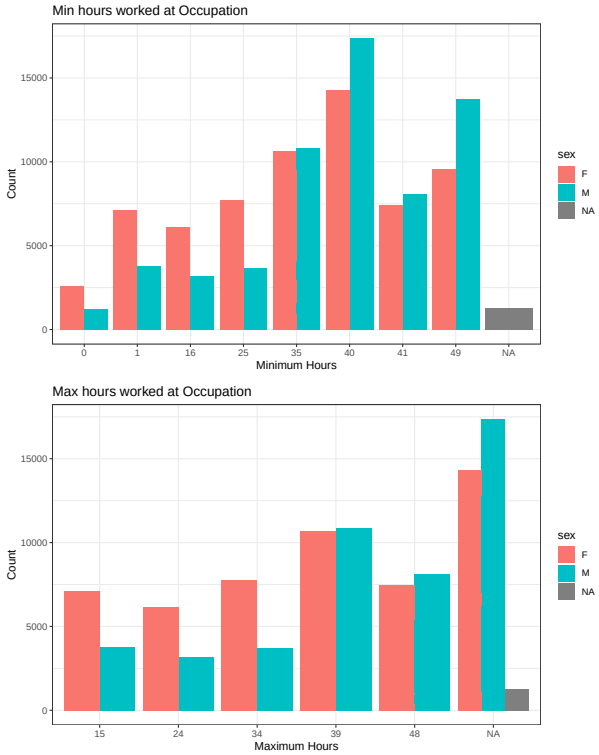
- It can be observed from both figures that overall females worked more than men. However, as the number of work-hours increased men have worked more than women.



- It can be observed from figure that industries like health care, education and training, construction and Professional and technical services have more working population as the working hours increased. Mining, electricity, gas, water showed low working population irrespective of work hours.



- It can be observed from figure that overall females worked more than men at all occupations. Although, for maximum hours worked, as number of working-hours increased, the number of men and women remained the same.



Conclusion

Conclusion

The education levels, field of study, industry of employment and occupation was studied for the Victorian SA4 level populations for the distributions according to gender and sex. The tables and plots were compared to mark the covariations between the population distributions.For example, the population trend between the field of study and industry of employment. Networks were drawn based on the population weights to analyze these trends. Some of the trends like more men were employed as managers when more women had studied management were found to be interesting. Cholropeth maps were made to analyze these trends spatially.

The goal of this report is to create a data story from these statistical summaries to enumerate the facts from the data and link them to the real world. The data provided by the Australian Bureau of Statistics is an aggregated open data and in no form identifies individuals who participated in the census. The ABS aims to integrate the census data with other datasets to make this census data more interesting. Thus, we aim to do the same and bring some interesting data stories as we progress building this report.

R Packages

R Core Team (2021)

Xie (2021a) Dietrich (2020) Wickham et al. (2021),

Wickham (2021a),

Wickham et al. (2020),

Zhu (2021),

Xie (2021b),

Tierney et al. (2020),

Pedersen (2020),

Henry and Wickham (2020),

Wickham and Hester (2020),

Wickham and Seidel (2020),

Wickham (2019),

Müller and Wickham (2021),

Wickham (2021b),

Wickham (2021c),

Xie (2021c),

Tierney (2019),

Xie (2016),

Wickham (2016),

Xie (2015),

Xie (2014),

Wickham et al. (2019),

Xie (2019),

Tierney (2017)

References

Dietrich, JP (2020). *citation: Software Citation Tools*. R package version 0.4.1. https://CRAN.R-project.org/package=citation.

Henry, L and H Wickham (2020). *purrr: Functional Programming Tools*. R package version 0.3.4. https://CRAN.R-project.org/package=purrr.

Müller, K and H Wickham (2021). *tibble: Simple Data Frames*. R package version 3.1.2. https://CRAN.R-project.org/package=tibble.

Pedersen, TL (2020). *patchwork: The Composer of Plots*. R package version 1.1.1. https://CRAN.R-project.org/package=patchwork.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Tierney, N (2017). visdat: Visualising Whole Data Frames. *JOSS* **2**(16), 355.

Tierney, N (2019). *visdat: Preliminary Visualisation of Data*. R package version 0.5.3. https://CRAN.R-project.org/package=visdat.

Tierney, N, D Cook, M McBain, and C Fay (2020). *naniar: Data Structures, Summaries, and Visualisations for Missing Data*. R package version 0.6.0. https://github.com/njtierney/naniar.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, H (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0. https://CRAN.R-project.org/package=stringr.

Wickham, H (2021a). *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 0.5.1. https://CRAN.R-project.org/package=forcats.

Wickham, H (2021b). *tidyr: Tidy Messy Data*. R package version 1.1.3. https://CRAN.R-project.org/package=tidyr.

Wickham, H (2021c). *tidyverse: Easily Install and Load the Tidyverse*. R package version 1.3.1. https://CRAN.R-project.org/package=tidyverse.

Wickham, H, M Averick, J Bryan, W Chang, LD McGowan, R François, G Grolemund, A Hayes, L Henry, J Hester, M Kuhn, TL Pedersen, E Miller, SM Bache, K Müller, J Ooms, D Robinson, DP Seidel, V Spinu, K Takahashi, D Vaughan, C Wilke, K Woo, and H Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**(43), 1686.

Wickham, H, W Chang, L Henry, TL Pedersen, K Takahashi, C Wilke, K Woo, H Yutani, and D Dunnington (2020). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.3.3. https://CRAN.R-project.org/package=ggplot2.

Wickham, H, R François, L Henry, and K Müller (2021). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.5. https://CRAN.R-project.org/package=dplyr.

Wickham, H and J Hester (2020). *readr: Read Rectangular Text Data*. R package version 1.4.0. https://CRAN.R-project.org/package=readr.

Wickham, H and D Seidel (2020). *scales: Scale Functions for Visualization*. R package version 1.1.1. https://CRAN.R-project.org/package=scales.

Wikipedia contributors (2021). *Census — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Census&oldid=1023830734. [Online; accessed 2-June-2021].

Xie, Y (2014). "knitr: A Comprehensive Tool for Reproducible Research in R". In: *Implementing Reproducible Computational Research*. Ed. by V Stodden, F Leisch, and RD Peng. ISBN 978-1466561595. Chapman and Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Xie, Y (2015). *Dynamic Documents with R and knitr*. 2nd. ISBN 978-1498716963. Boca Raton, Florida: Chapman and Hall/CRC. https://yihui.org/knitr/.

Xie, Y (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. ISBN 978-1138700109. Boca Raton, Florida: Chapman and Hall/CRC. https://bookdown.org/yihui/bookdown.

Xie, Y (2019). TinyTeX: A lightweight, cross-platform, and easy-to-maintain LaTeX distribution based on TeX Live. *TUGboat* (1), 30–32.

Xie, Y (2021a). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.22. https://CRAN.R-project.org/package=bookdown.

Xie, Y (2021b). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.33. https://yihui.org/knitr/.

Xie, Y (2021c). *tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents*. R package version 0.31. https://github.com/yihui/tinytex.

Zhu, H (2021). *kableExtra: Construct Complex Table with kable and Pipe Syntax*. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra.