How to improve the public health services of Victorian residents?

GitHub repository:

https://github.com/COMP20008/assignment-2-comp20008-assignment-2-group-99

Yuanhao Zhuo 1161918
Siyu Tian 928307
Haoze Xia 1131343
Fuhan Sun 1131339

COMP20008 Final Report

The University of Melbourne
14/10/2021

1. What is the research question and how is it related to the theme of understanding the liveability, inclusiveness, health, and sustainability of communities in Victoria?

Our research questions focus on how to improve the public health services of Victorian residents by analyzing the number of beds per thousand people through economic adjustments and other methods.This project will give some suggestions about how to optimize of Australian long-term medical resources in the future so this topic is related to the liveability and health of Australia.

It specifically includes the following tasks:

- Data crawling
- Data preprocessing
- Visualization
- Cluster analysis
- Principal component analysis and logistics regression
- Give some suggestions based on the model analysis.

2. What are the datasets you've used and how have you linked them together?

After data pre-processing, the following data files has been generated:

Table 2.1: Dataset information

VARIABLE TYPE	NAME	FORMAT	ROW	COLUMN	TIMELINE
INDEPENDENT VARIABLE	Public Hospital	csv	9	17	2004-2020
	Private Hospital	csv	9	13	2004-2016
	Beds Number	csv	3	17	2004-2020
DEPENDENT VARIABLE	PM2.5	csv	9	21	2000-2019
	Income	csv	21	9	2001-2020
	Population Density	csv	21	9	2001-2020
	Population	csv	20	9	2001-2019
	Life expectancy	csv	20	9	2001-2019
	Unemployment Rate	csv	21	9	2001-2020
	Gross value added	csv	15	9	2004-2017
	Employment 15 years old	csv	20	9	2001-2019
	Hospital Budget	xlsx	10	7	2004-2020

Datasets included in this project can be considered as these two parts:
Independent variable: Health services in Australia dataset, such as the number of public hospitals and the number of beds.

Dependent variables: Australia’s population quality dataset, Australia’s air quality dataset.
The independent variable and the dependent variable are linked according to the timeline to observe the relationship between Australian health services, population quality, and air quality.

Through the above analysis, we can then predict the changing trend of Australia’s health services. And to suggest how the state government should adjust health resources to respond to public emergencies.

3. What wrangling and analysis methods have you applied? Why have you chosen these methods over other alternatives?

3.1 Wrangling

3.1.1 Crawling

In previous work, the question to be researched and the impact variables involved in this question have been confirmed. Thus, the team members consulted a large number of secure and authoritative websites and found enough databases available according to the central question.

The datasets are mainly obtained by web crawling because of crawling’s quickness and accuracy. The process of crawling uses two packages in Python which are “urllib” and “BeautifulSoup” packages. And codes for datasets are divided into two functions which contain downloading documents from URLs the team found and saving the documents locally. For the “downloading documents” part, the permission and response from URLs should be gained firstly, and inspecting the links for documents, then, using the anchors and “end with” contents in the new link to download the documents. In the meantime, the “saving” part needs to write codes for the local location of storage the team preferred.

3.1.2 Data Preprocessing

The data obtained through the web crawler is incomplete, inconsistent, and mixed with many noises. Therefore, data pre-processing is needed to solve such problems.

For PDF format, the steps for pre-processing is as follows:

- Convert useful pdf pages into image format
- Use OCR technology to recognize the text in the picture
- Use ‘Regex’ to capture the required data and then summarize the data
- Save csv file

For CSV format, the steps for pre-processing is as follows:

- Read csv, select the key column and save it as list
- Extract the required data by state(NSW, VIC,...)
- Sorted the data by timeline
- Combine 8 states’ data and save to csv file

For Excel format, the steps for pre-processing is as follows:

- Use ‘pandas’ package to read excel files
- Use double for loop to traverse each cell
- Find the position of the ‘NaN’ value
- Use column mean or average of front and back cells to replace ‘NaN’
- Save Dataframe to csv file.

We have tried to use the ‘pyPDF2’ package in python to process pdf data. However, the outputs are unsatisfactory, unreadable, and meaningless. Therefore we choose OCR technology in python to process pdf data.

3.2 Analysis methods

When we first got the data, we wanted to build multiple regression models for each area (NSW, VIC, QLD, SA, WA, TAS, NT, ACT) based on independent variables gross value added, income, labour, life expectancy, environment(pm2.5), population, unemployment rate.

We checked the assumption by starting with scatter plots between independent variables for each area.

This is the plot of
VIC :

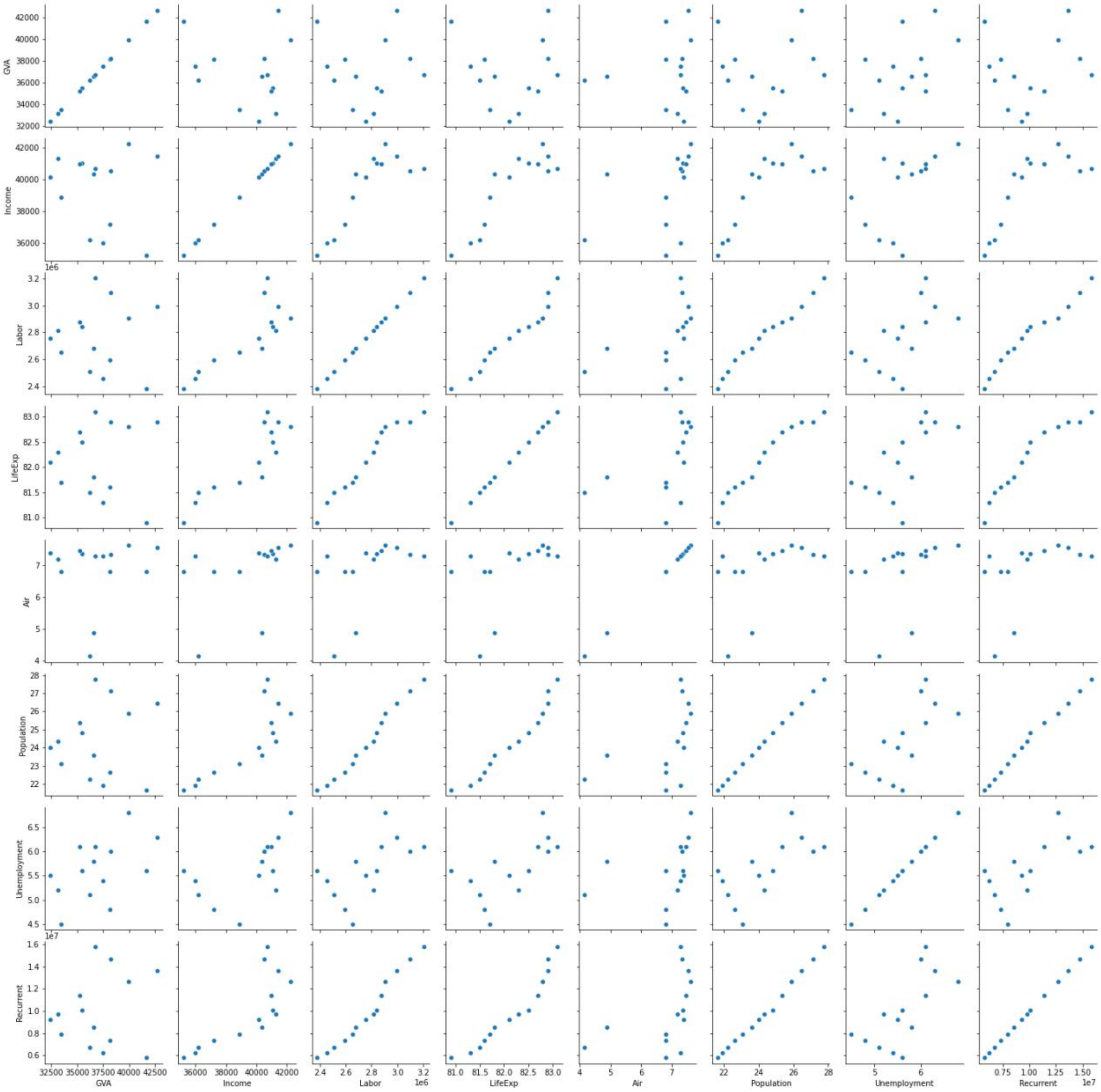


Figure 3.2: scatter plot of each factors for VIC
(The visualization results of the other areas can be found in the appendix)

It is obvious that some independent variables are correlated to others. we double-checked it by computing the Pearson coefficient.

The Pearson coefficient between labor and the population is about 0.993(NSW), 0.990(VIC), 0.967(QLD), 0.911(SA), 0.979(WA), 0.983(TAS), 0.981(NT), 0.940(ACT).

The Pearson coefficients are extremely high. Therefore, it is fair to conclude that these two variables are not independent. Besides population and labor, it seems like there also exists a linear relationship between labor and income. Therefore the assumption of multiple regression is not satisfied in this situation. There are two methods to fix this problem:

Method 1: list all the combinations of independent variables and select the combinations that satisfy the independent assumption. Then create models based on all combinations and choose the one that gives the smallest total sum of squares.

Method 2: Use forward selection to avoid using dependent variables.

However, the calculation requirements of the first method are extremely large. And for the second method, python does not have packages to do the forward selections Automatically. Besides this, there is still one more problem: Comparing with the sample size, the number of features is quite large. This may lead to imprecise parameter estimation. In order to solve all these problems, we decided to use the principal component analysis. It can avoid inappropriate independent variables selection and reduce the dimension at the same time.

We built the model by the following steps:

3.2.1 Cluster analysis

Since the data granularity provided in the original data set is annual, the amount of data in Victoria is not enough to support regression analysis. Therefore, we collected relevant data across Australia and performed a cluster analysis with the state as the granularity to find out states similar to Victoria, and also use the data of these states to perform regression analysis.

3.2.1.1 Normalization

Before cluster analysis, we must first normalize the data. For a specific set of column vectors x, the normalized expression can be written as:

$$x_{normalized} = \frac{x - min(x)}{max(x) - min(x)}$$

In order to eliminate the influence of dimensions, for some data, it is necessary to take the logarithm first and then perform the normalization operation, this expression can be written as:

$$x'_{normalized} = \frac{log(x) - min(log(x))}{max(log(x)) - min(log(x))}$$

3.2.1.2 K-means and Hierarchical Clustering

In order to ensure correctness and rationality, we use both k-means and hierarchical clustering methods at the same time.

For k-means clustering, we initialize 2 different clusters, randomly 2 center points, and use the center of each cluster to update the cluster center with the mean value of the value contained in the cluster over and over again, until it is stable and generate 2 categories.

For hierarchical clustering, we treat each object as a cluster, merge close clusters from bottom to top, and repeat until all objects are merged.

3.2.2 Principal Component Analysis

Because there are many factors in our model, the relationship between the factors is not clear, so the correlation analysis between factors is difficult, thus we choose to bypass the correlation analysis and directly use the principal component analysis (PCA) method to reduce the dimensionality of the data.

3.2.2.1 Standardization

Because PCA is susceptible to outliers, it is necessary to standardize the data before performing PCA. For a column vector x of a factor, the specific standardization method is as follows:

$$x_{standardized} = \frac{x - mean(x)}{\sqrt{Var(x)}}$$

Since the result is close to the standard normal distribution, there is no need to take the logarithmic.

3.2.2.2 PCA

Principal component analysis obtains new variables that are orthogonal to each other by linearly combining the original variables. Since the new variables are orthogonal to each other, which means there is no correlation, so that they can be subsequently applied to the regression model. This can be achieved by calling the sklearn library in python.

3.2.3 Logistic Regression Model

After getting the PCA results, we used logistic regression for regression analysis. Since the value of y is not in the range of 0 to 1, re-scale y first. According to the data provided by the World Bank, as of 2016, the highest number of hospital beds per thousand people in the world is 13 (Japan), so we set y=13 as the upper bound of logistic regression.

In addition, on the basis of PCA, we also need to add a 1 vectors to the X matrix obtained by PCA to represent the intercept.

After the above operations, we can use the re-scale y and x matrices to perform the following regression analysis:

$$y_{rescale} = sigmoid(\eta)$$

Where we have:

$$\eta = X \times \vec{w}$$

$$sigmoid(\eta) = \frac{1}{1 + e^{\eta}}$$

And $\vec{w}$ is the regression parameter vector we need to get.

3.2.4 Analysis of the influence of each factor

After obtaining the relevant parameters of the PCA and the logistic regression model, we can analyze each factor by controlling variates. By changing the raw value of each factor (±10%, ±5%, ±1%, unchanged) and observing the changes in the final predicted value, certain suggestions are made for each influencing factor. The specific steps are as follows:

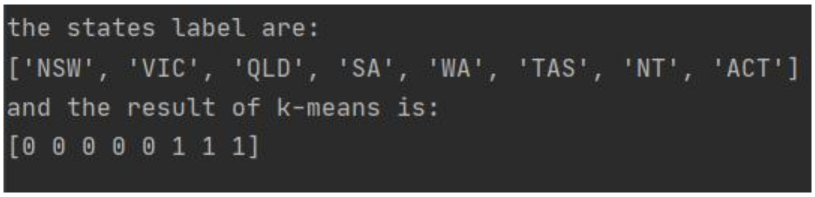
- Change the raw value of a factor, standardize them, and then use PCA transform
- Add 1 factor representing the intercept in the obtained data
- Fit the y value with the previously obtained logistic model
- Compare and draw conclusions

4. What are the key results your research has obtained?

4.1 Result of Clustering

In order to use data from other states with similar conditions to enhance the generalization of the model, we conducted a cluster analysis of the Australian states based on eight factors including economy, health service expenditure, environment, and population.

Firstly, the result by using k-means clustering is (k=2):



```
the states label are:
['NSW', 'VIC', 'QLD', 'SA', 'WA', 'TAS', 'NT', 'ACT']
and the result of k-means is:
[0 0 0 0 0 1 1 1]
```

Figure 4.1.1: Result of k-means clustering (k=2)

And the result by using hierarchical clustering is:

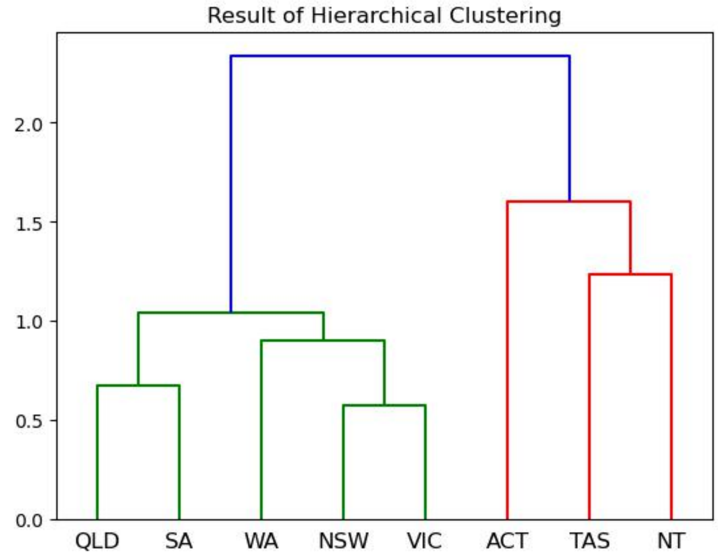


Figure 4.1.2: Result of hierarchical clustering

From the above results, it is not difficult to find that when only the last two categories are retained, the results of k-means and hierarchical clustering are exactly the same. Therefore, we can think that QLD, NSW, SA, WA, and VIC are similar. In the regression model, data from five states are used for analysis instead of just a single state of Victoria.

4.2 Result of Logistic Model

PCA reduces our input to 9 dimensions. Since each factor of the new x is a linear combination of the original x and has no specific meaning, the specific results are not shown here.

According to the results of the gradient descent method, we can get a logistic regression model and make the following diagnostic plot:

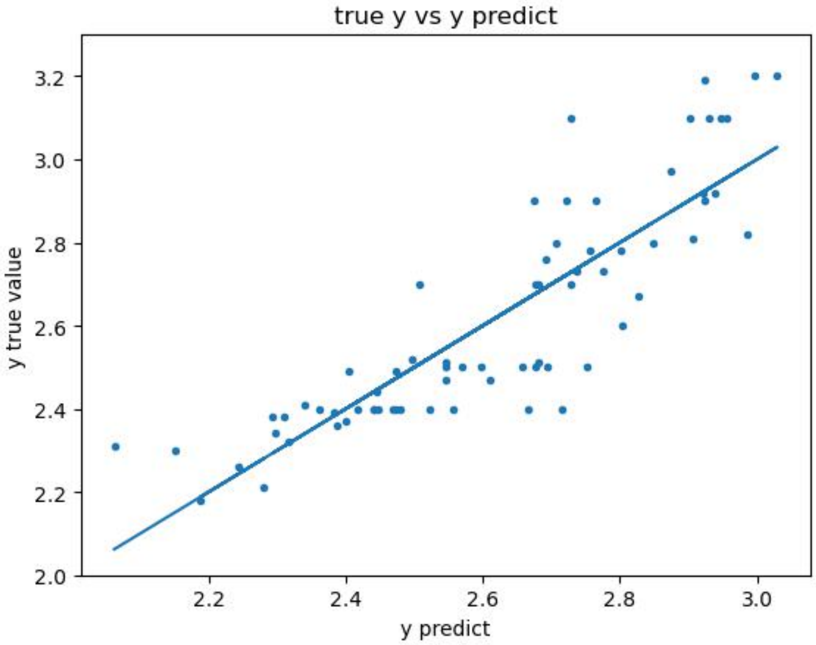


Figure 4.2.1: Diagnostic plot

Where the dot is the regression result versus the true value, and the line is for predict value equals true value. The distance between dots and lines are the residuals. It is not difficult to find from the figure that the points are evenly distributed above and below the line, so this is an acceptable model (the specific rationality quantitative analysis will be shown later).

4.3 Sensitivity Analysis and Suggestions to the Government

Omit the factors that cannot draw valid conclusions, the following table shows some of the results (whole table will be shown in the appendix):

Table 4.3.1: Sensitivity analysis of the influence

	GVA	unemployment	income	labour	life exp
+10%	0.08(meaningless)	1.34	4.25	11.70	13.00(meaningless)
+5%	2.49	2.50	2.49	2.49	2.49
+1%	2.56	2.52	2.49	2.48	2.45
unchanged	2.49	2.48	2.46	2.47	2.48
-1%	2.50	2.77	2.51	2.49	2.46
-5%	2.49	2.49	2.48	2.47	2.48
-10%	2.49	2.49	2.48	2.48	2.49

And this gives the graphic result as follows:

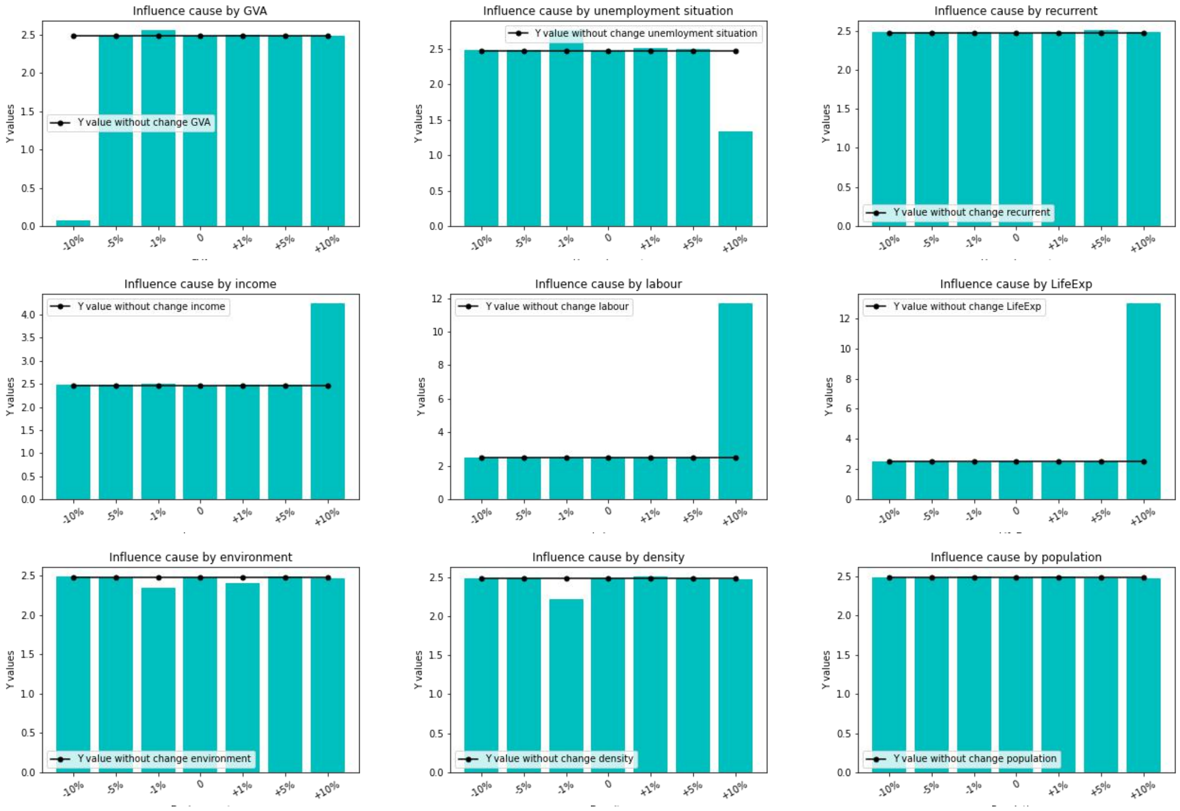


Figure 4.3.2: Sensitive analysis graphic result

4.4 final conclusion

All in all, based on our model, the following suggestions can be made to the state government: boost the economy, increase investment in health services, and avoid an aging population (labour).

5. Why are your results significant and valuable?

5.1 Model Rationality

In order to test whether our model is reasonable, based on the knowledge of mathematical statistics, we conducted an F test on the model. The F value can be calculated with the following formula:

$$F = \frac{(SS_{mean} - SS_{fit})/df1}{SS_{fit}/df2}$$

Where:

$$\mathrm{SS_{fit}} = \sum (\hat{y} - y)^2$$

$$\mathrm{SS_{mean}} = \sum (\bar{y} - y)^2$$

This give F=20.41236. Since for 95% confidence interval, F value is 2.787, which is much smaller than 20.41, this means we reject the null hypothesis, and this model is a reasonable model.

5.2 Value of the Entire Project

Our research is valuable for these two reasons. The first reason is that no one has done this kind of research recently. Secondly, as the epidemic is not over yet, we need to be always prepared for the recurrence of the epidemic. Whether there are enough beds has a big impact on the control of the epidemic. Once medical resources are in short supply, the situation will be difficult to control. Our research can give some suggestions about how Australian medical resources can improve the ability to deal with sudden largeterm public safety events.

6. What are the limitations of your results and how can the project be improved for the future?

6.1 Limitation

- The timeliness of the data is not strong enough, and the amount of data is not large enough.
- The correlation between the various factors is not clear. Although the correlation is bypassed with PCA, the dimensionality reduction effect of PCA is not satisfactory enough.
- Because PCA bypasses the correlation, the correlation between variables is not clear.

6.2 Improvement

- After collecting more sufficient data, analyze the correlation between variables to optimize the model.
- Find a better way to deal with pdf files.
- Upgrade the computer’s configuration to be able to cope with a larger number of data sets.

APPENDIX A

Table A: whole result of sensitivity analysis of the influence

	Year	GVA	unemployment	recurrent	income	labour	life exp	environment	density	pop
+10%	0.00	0.08	1.34	2.49	4.25	11.70	13.0	2.46	2.47	2.48
+5%	2.49	2.49	2.50	2.51	2.49	2.49	2.49	2.49	2.48	2.48
+1%	2.48	2.56	2.52	2.49	2.49	2.48	2.45	2.41	2.51	2.50
0	2.49	2.49	2.48	2.47	2.46	2.47	2.48	2.48	2.49	2.49
-1%	2.49	2.50	2.77	2.48	2.51	2.49	2.46	2.35	2.22	2.50
-5%	2.49	2.49	2.49	2.48	2.48	2.47	2.48	2.48	2.48	2.49
-10%	2.49	2.49	2.49	2.48	2.48	2.48	2.49	2.49	2.49	2.49

APPENDIX B

Table B: URL links

URLs
● https://www.aihw.gov.au/reports/hospitals/ahs-2004-05/contents/table-of-contents
● https://www.aihw.gov.au/reports/hospitals/ahs-2005-06/contents/table-of-contents
● https://www.aihw.gov.au/reports/hospitals/ahs-2006-07/contents/table-of-contents
● https://www.aihw.gov.au/reports/hospitals/ahs-2007-08/contents/table-of-contents
● https://www.aihw.gov.au/reports/hospitals/ahs-2008-09/contents/table-of-contents
● https://www.aihw.gov.au/reports/hospitals/australian-hospital-statistics-2009-10/contents/table-of-contents
● https://www.aihw.gov.au/reports/hospitals/australian-hospital-statistics-2010-11/contents/table-of-contents
● https://www.aihw.gov.au/reports/hospitals/hospital-resources-ahs-2015-16/contents/summary
● https://www.aihw.gov.au/reports/hospitals/ahs-2016-17-hospital-resources/contents/table-of-contents
● https://www.aihw.gov.au/reports-data/myhospitals/content/data-downloads
● https://stats.oecd.org/index.aspx?DataSetCode=REGION_EDUCAT#
● https://covidlive.com.au/states-and-territories

APPENDIX C: Scatter plots for the independent variables of each area.

NSW

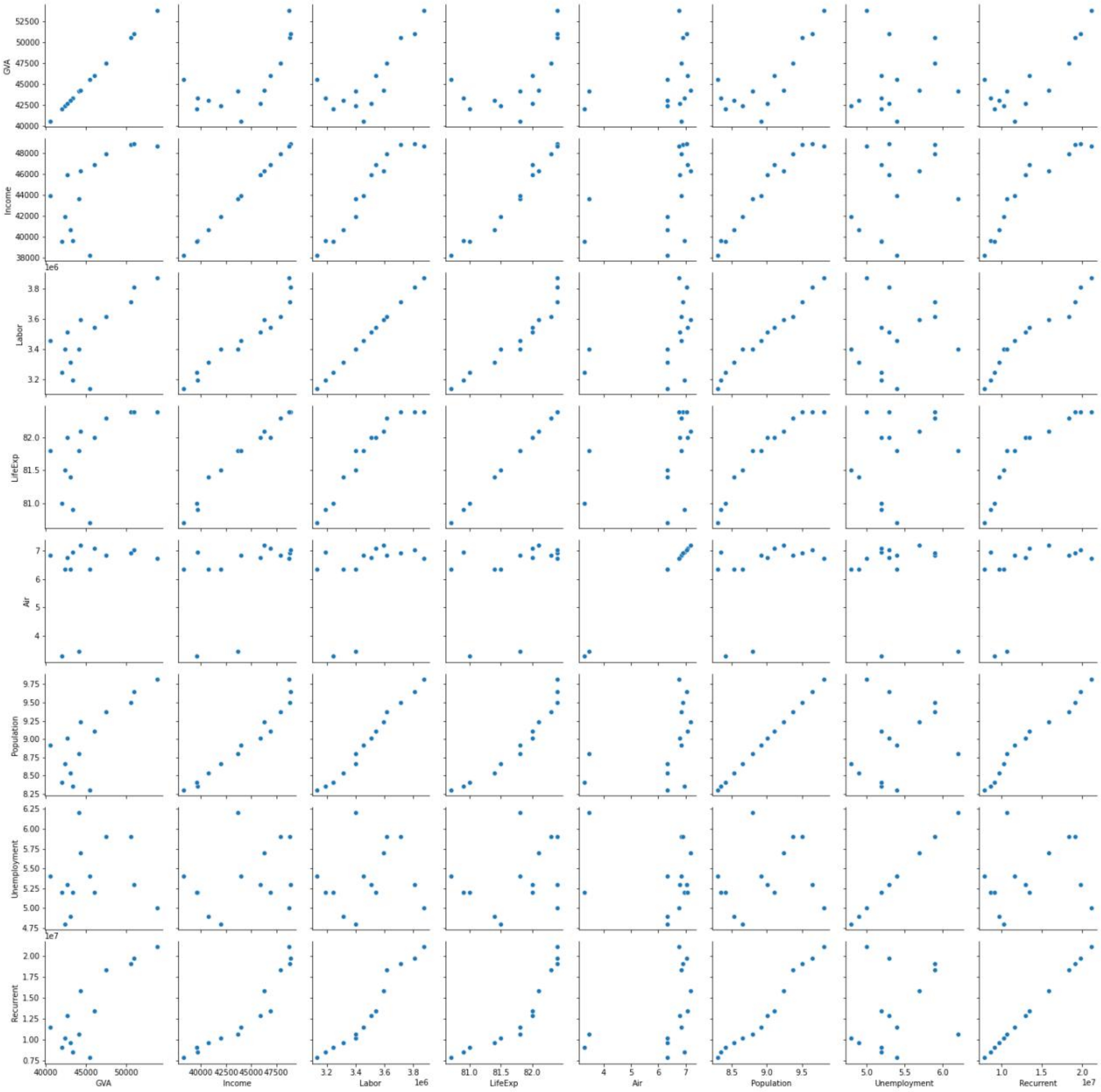


Figure C.1: Scatter plots for the independent variables for NSW

QLD

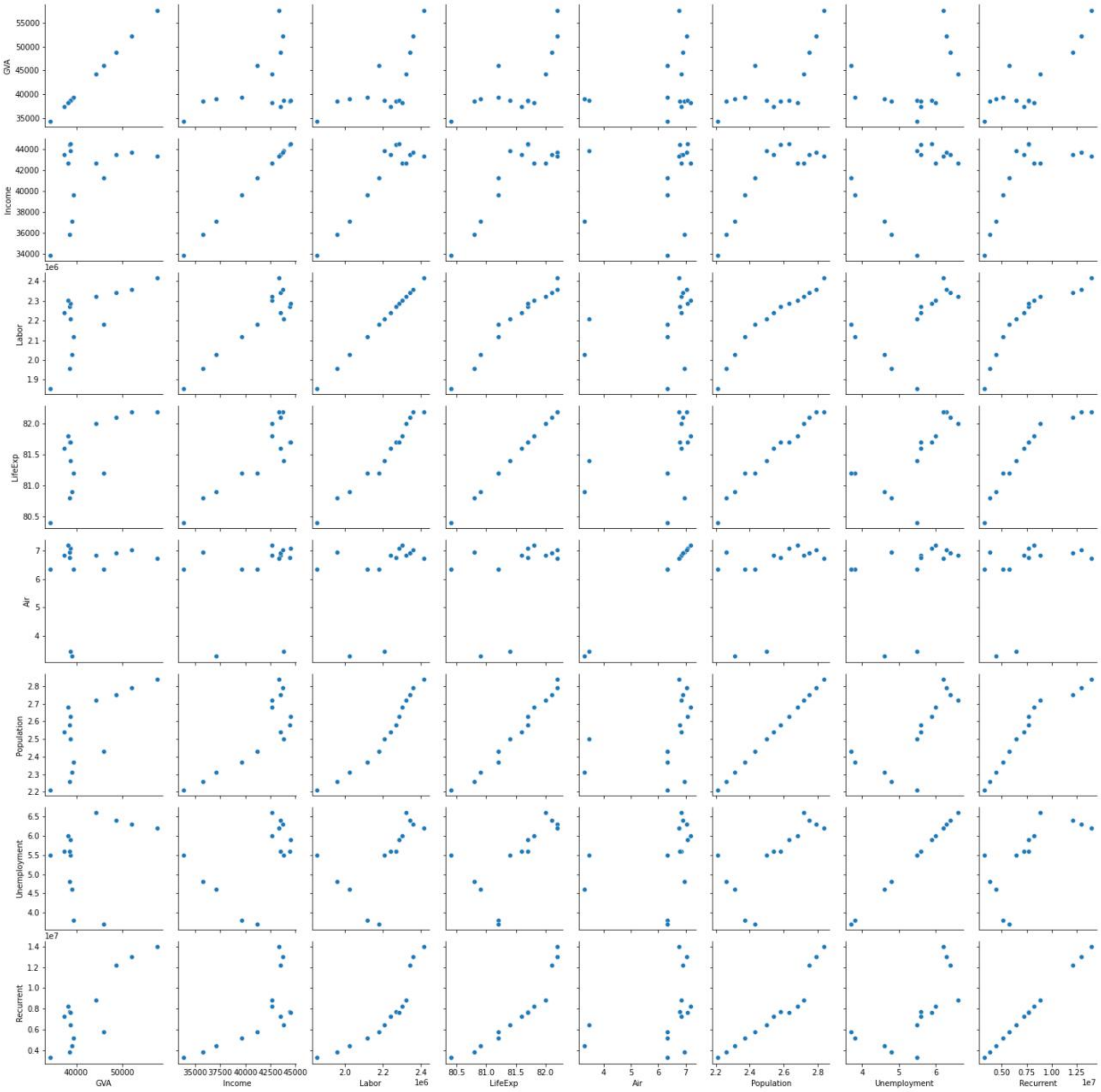


Figure C.2: Scatter plots for the independent variables for QLD

SA

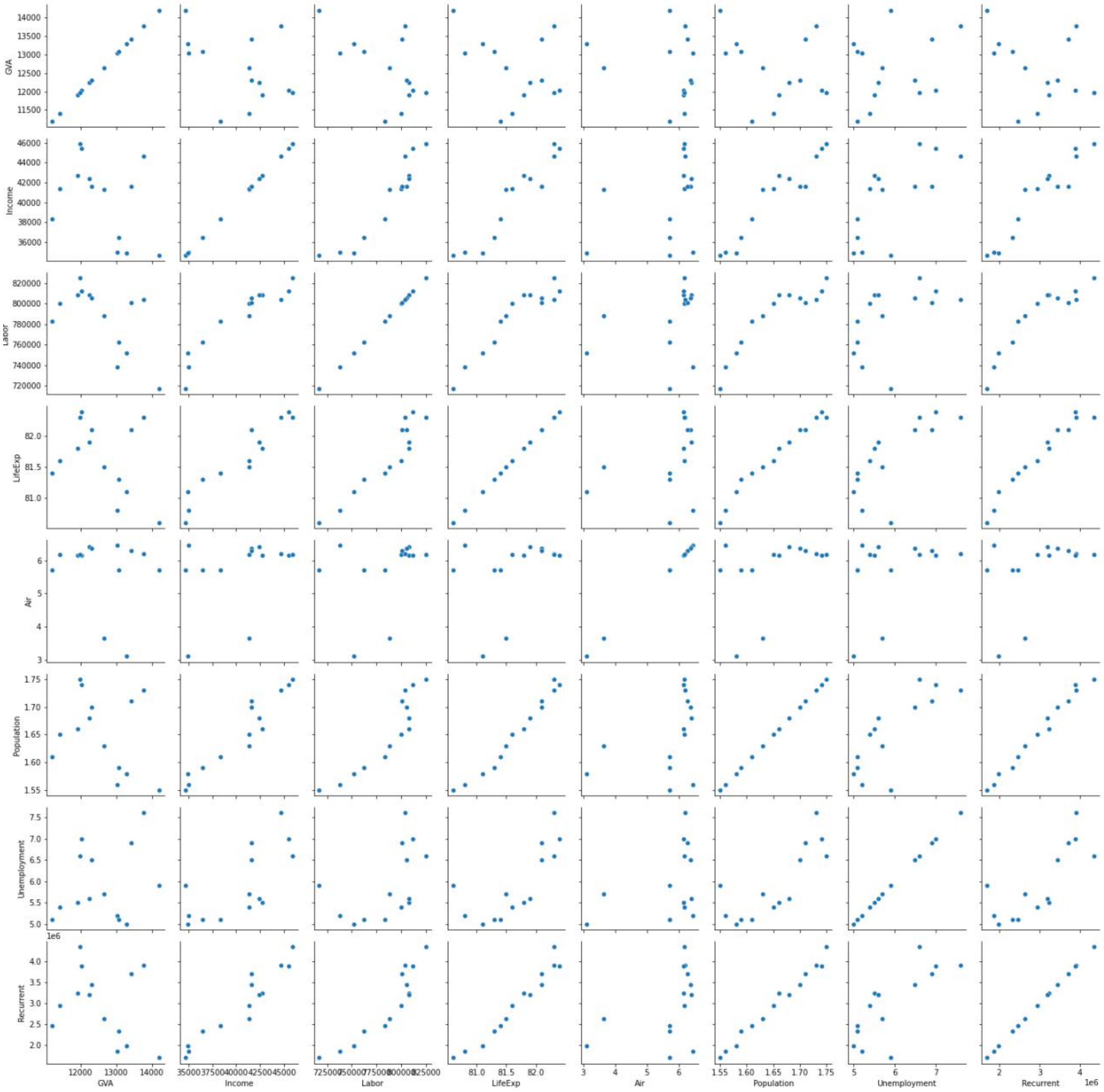


Figure C.3: Scatter plots for the independent variables for SA

WA

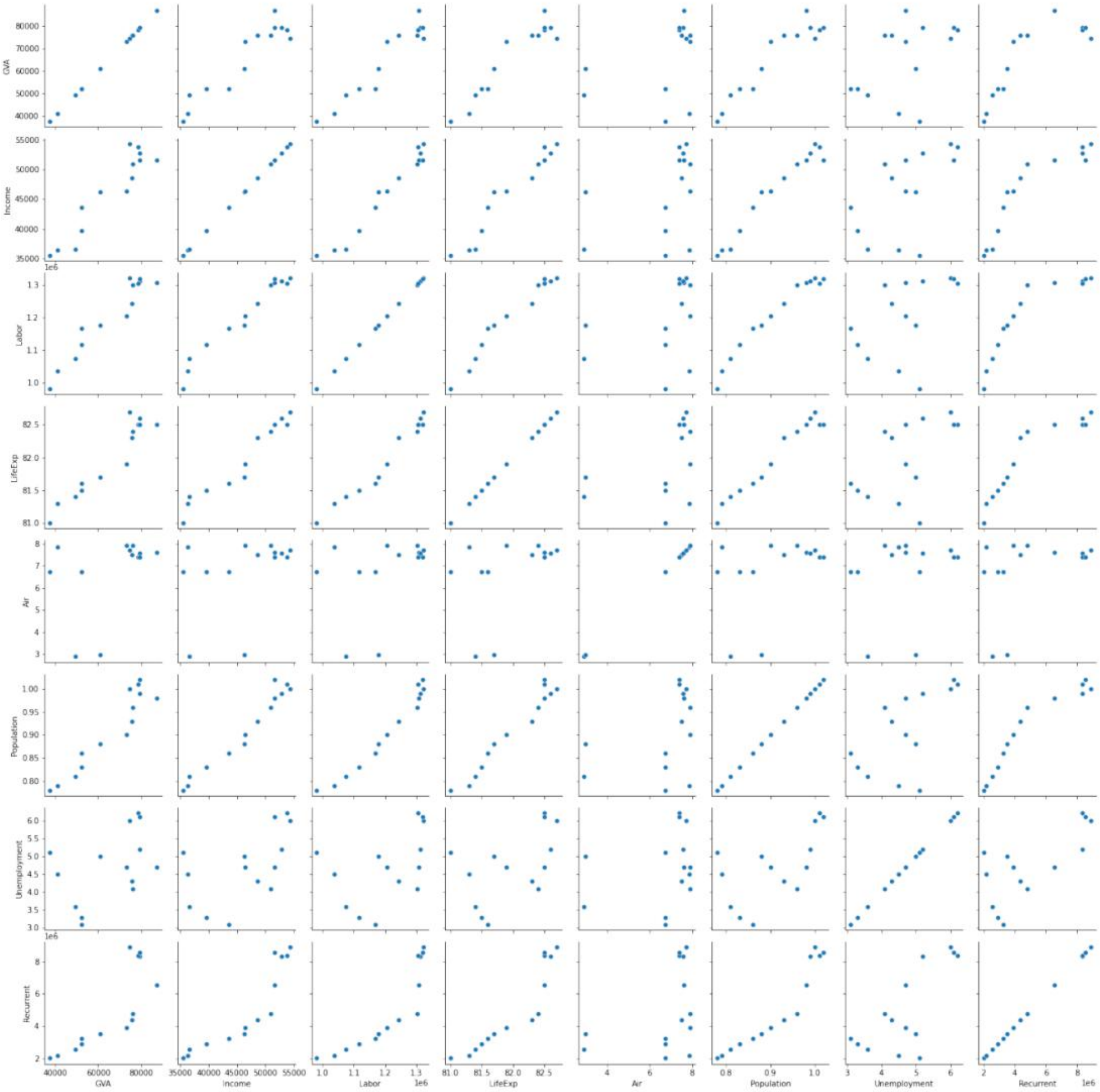


Figure C.4: Scatter plots for the independent variables for WA

TAS

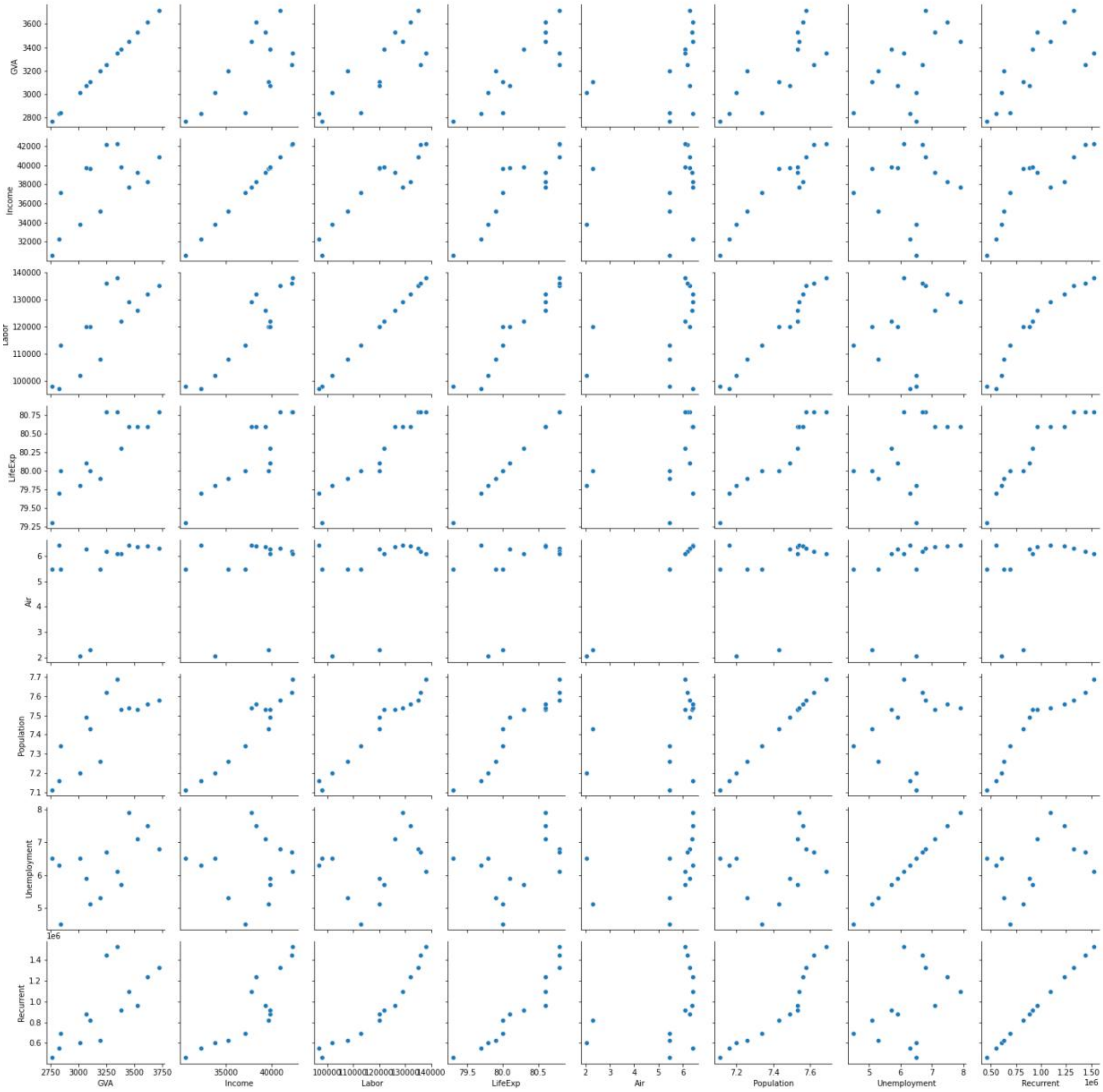


Figure C.5: Scatter plots for the independent variables for TAS

NT

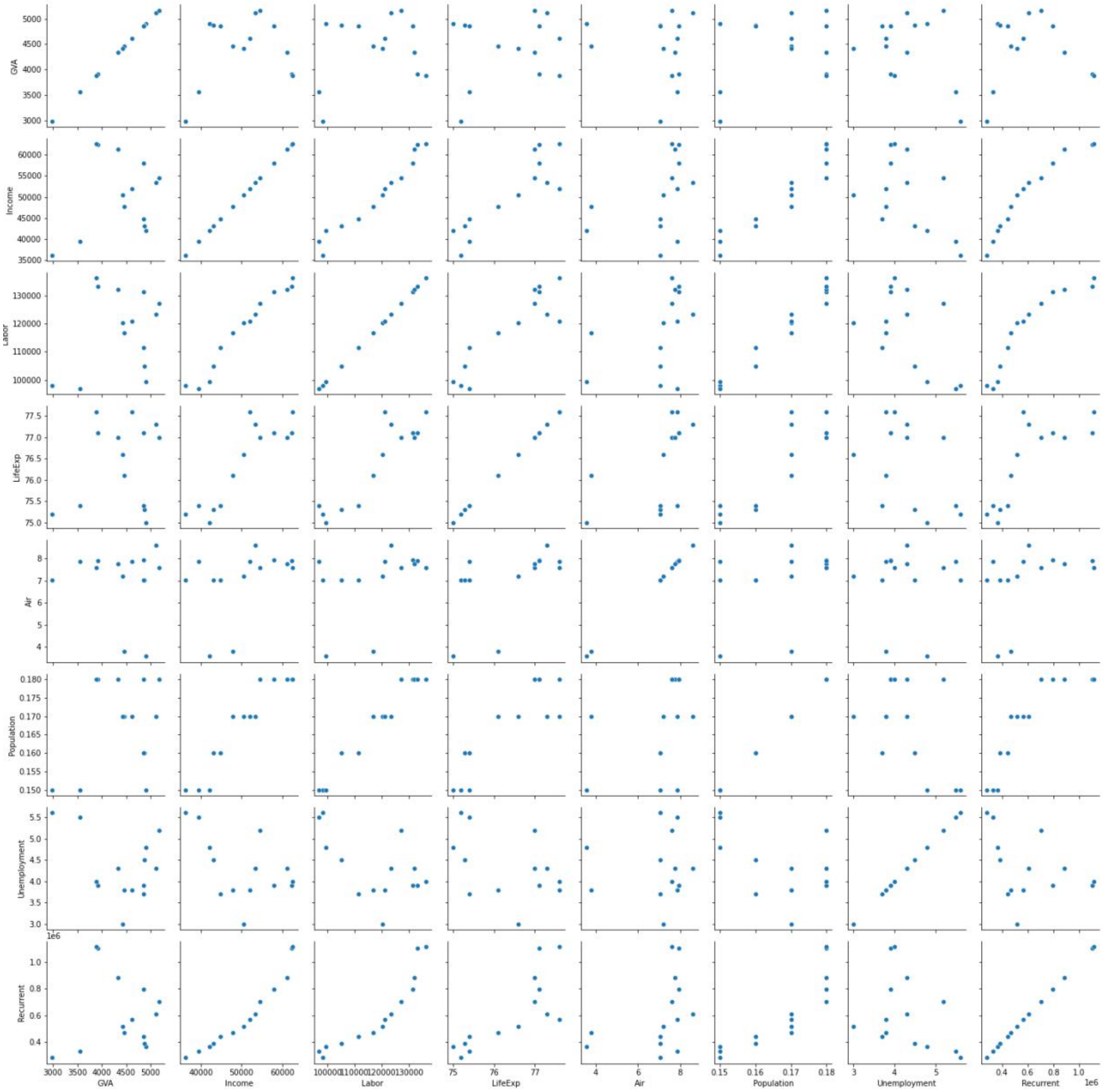


Figure C.6: Scatter plots for the independent variables for NT

ACT

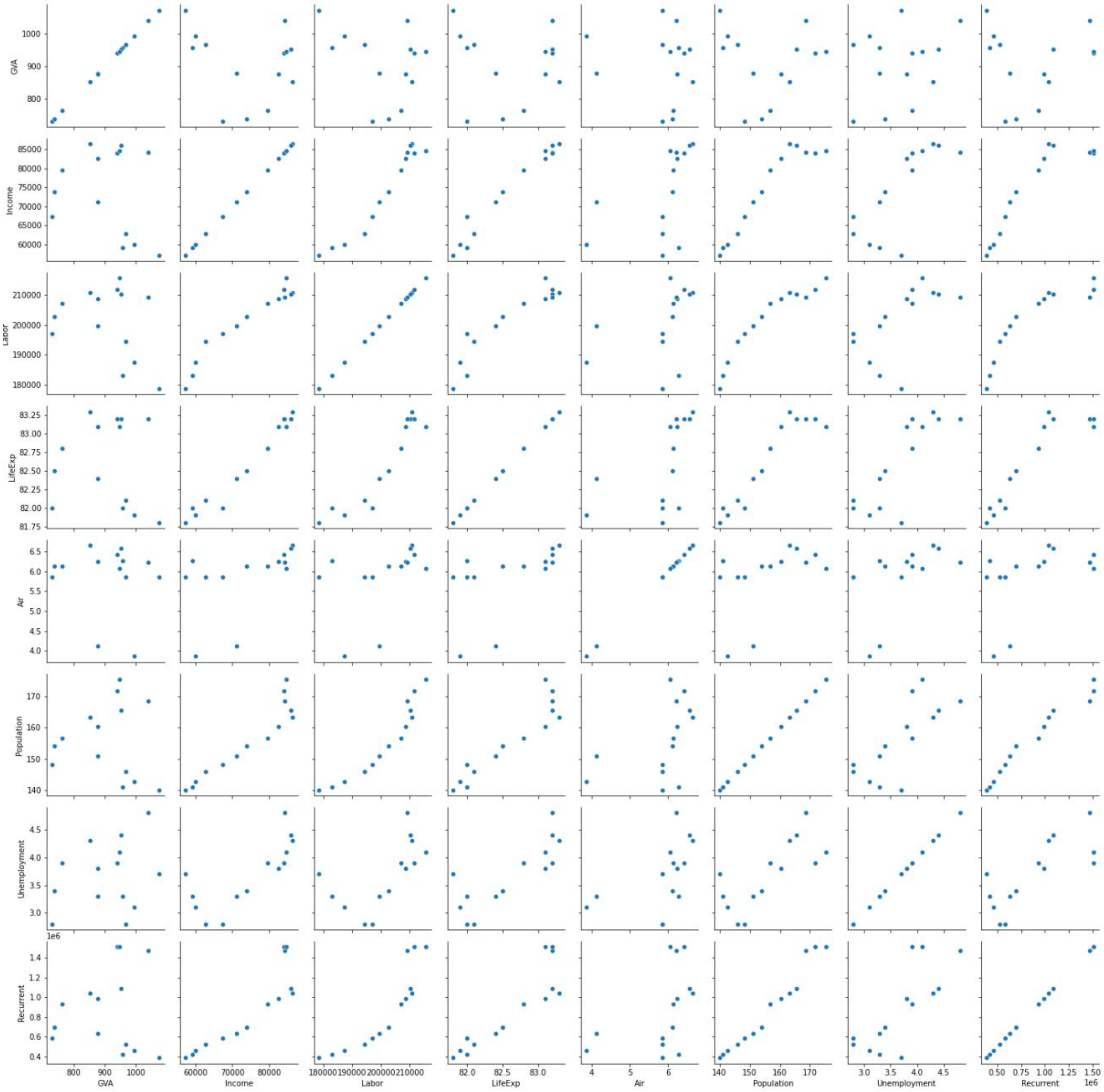


Figure C.7: Scatter plots for the independent variables for ACT