42047: Data Processing with Python

Assignment (Part C)

Report: Data Analysis and Visualization - Australian Bureau of Statistics (ASB) Datasets

Student Name: Monali Patil
ID: 14370946
Date: 28/10/2022
(Word Count: 1529)

Table of Contents

Abstract

Introduction: Gather creative insights around Australia.

Problem statement: Present a snapshot of Australia and how it's changing.

Process/Approach: Over the period of 10 years, would like to analyse significant aspects of Australia such as employment and income within different states and age groups and the trend of immigration from various countries. To aid this, will use (all learnings from 1st, 2nd, and future workshops) NumPy and Pandas packages in Python for exploratory data analysis and data cleaning. Further, will utilise statistical techniques of Data Science, Machine Learning to develop meaningful results. Will visualise datasets (in scatterplots, cluster maps etc.) with the help of different libraries from the Matplotlib package.

Conclusion: Display trend of features representing Australia like immigration, employment, housing patterns etc. over a decade.

1. Introduction and Background

The Australian Bureau of Statistics (ASB) is a National Organization of Australia which collects and provides authentic, dependable data-facts every 5 years also known as the Census about numerous aspects like Population, Immigration, Economics, Housing, and Social-Cultural constitute of Australia (Australian Bureau of Statistics, n.d.). The Census is an accumulated snapshot of data that informs and assists government departments, public planners, and stakeholders in planning and strategizing abundant services and infrastructures such as health care, schools, housing, transport etc. while encouraging them to make informed decisions (Australian Bureau of Statistics, n.d.).

To facilitate data analysis and visualization learning, the datasets are downloaded from Census year 2011 until 2021 and are available at ASB Data packs: https://www.abs.gov.au/census/find-censusdata/datapacks?release=2021&product=GCP&geography=ALL&header=S

1.1 The Business Problem

Analyse the evolution of Australia as a nation relating to features like families/person's income, and migration (foreign born) from different countries from 2011 until 2021. Also, investigate if any trends highlighting the significance of the country and impacting how it has changed.

1.2 Business Question

The analysis will cover below (the period of 3 censuses 2011, 2016 and 2021):

Income:

- Which are the states in Australia to which top income earners belong?
- How did families (couple and single parent) perform during this census period?
- Which income bracket shows the best performance?
- How have income levels changed throughout this period?

Migration:

- From which countries people migrated the most to Australia?
- Which are the top states in Australia where migrants prefer to settle?
- Is there any noticeable trend in immigration from various countries around the world?

1.3 Dataset

The ASB dataset consists of comprehensive (aggregated) data in multiple files about numerous features. The below figure shows the original files downloaded from the ASB source data packs.

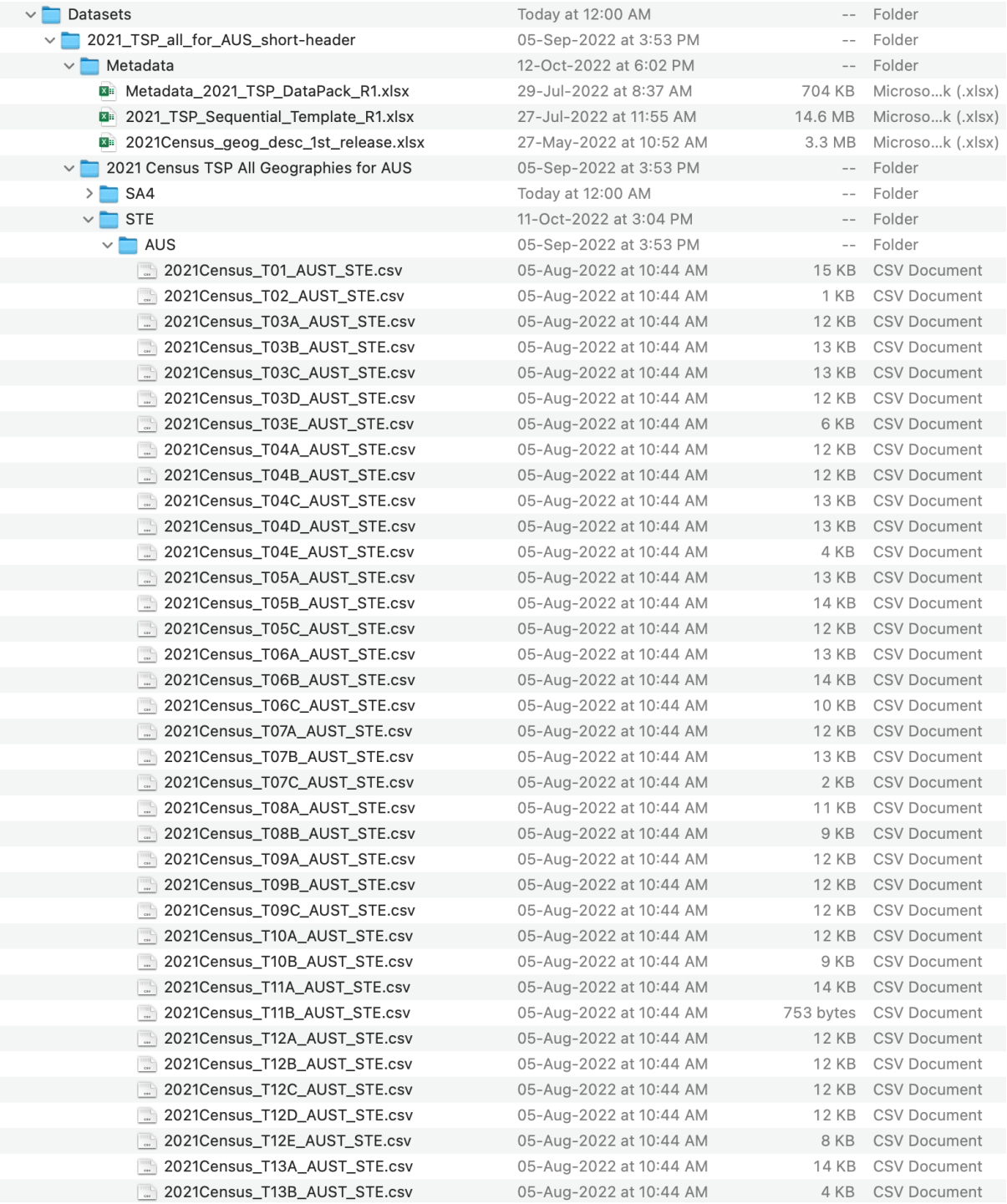

Figure 1: Original files downloaded from the ASB data pack.

For analysis purposes, only the data about required features of income and immigration files are read and combined in separate single datasets for income and migration. The below table provides information on various characteristics of the 1) income and 2) migration datasets.

Dataset/Feature	Final Dataset Name	Attributes and their Data Type	Dataset Summary
Income	df_income_details Rows: 390 Columns: 14	1. Type_Value: object 2. New South Wales: int64 3. Victoria: int64 4. Queensland: int64 5. South Australia: int64 6. Western Australia: int64 7. Tasmania: int64 8. Northern Territory: int64 9. Australian Capital Territory: int64 10. Other Territories: int64 11. census_year: int64 12. total_all_states: int64 13. income_bracket: int64 14. family_status: object	This aggregated data is about the number of families(single parent or couple) receiving remuneration in various salary brackets from different states of Australia for 3 censuses 2011, 2016 and 2021.
Migration	df_migration_details Rows: 342 Columns: 15	1. Type_Value: object 2. New South Wales: int64 3. Victoria: int64 4. Queensland: int64 5. South Australia: int64 6. Western Australia: int64 7. Tasmania: int64 8. Northern Territory: int64 9. Australian Capital Territory: int64 10. Other Territories: int64 11. census_year: object 12. sex: object 13. country: object 14. total_all_states: int64 15. iso_code: object	This comprehensive information is regarding the migration of people born overseas who participated in the census living in different states of Australia from 2011 until 2021, 3 censuses.

Table 1: Details of various characteristics of the income and migration datasets.

The below shows the sampling of records from the datasets.

	Type_Value	New South Wales	Victoria	Queensland	South Australia	Western Australia	Tasmania	Northern Territory	Australian Capital Territory	Other Territories	total_all_states	income_bracket	census_year	family_status
20	C21_2000_2499_1PF_1C	20342	15155	12376	4186	6050	1193	571	1567	4	61444	2499	2021	SingleParent Family
278	C11_400_499_1PF_4mC	1603	985	1063	330	429	139	74	54	0	4677	499	2011	SingleParent Family
145	C16_1_149_CF_1C	660	700	421	173	225	48	79	26	0	2332	149	2016	Couple Family
382	C16_1000_1499_1PF_3C	4839	3588	3482	1114	1464	368	134	176	0	15165	1499	2016	SingleParent Family
454	C21_500_649_Tot	21968	16831	17025	6013	7628	2366	606	697	4	73138	649	2021	Other

Table 2: Sample of 5 rows from the income dataset.

	Type_Value	New South Wales	Victoria	Queensland	South Australia	Western Australia	Tasmania	Northern Territory	Australian Capital Territory	Other Territories	census_year	sex	country	total_all_states	iso_code
227	Philippines_C11_P	70387	38003	29462	8860	17234	1267	3581	2420	10	2011	P	Philippines	171224	PHL
322	Born_elsewhere_C21_F	206679	181184	102567	32395	69067	5583	4361	12144	44	2021	F	Born elsewhere	614024	NaN
240	Poland_C21_M	5914	6006	2406	2079	2368	234	60	430	0	2021	M	Poland	19497	POL
222	Pakistan_C21_M	19736	18279	3592	3212	4685	1055	450	1671	12	2021	M	Pakistan	52692	PAK
320	Born_elsewhere_C16_P	349172	306044	158743	54517	114069	8172	7305	18898	129	2016	P	Born elsewhere	1017049	NaN

Table 3: Sample of 5 rows from the migration dataset.

2. Overview of the Data Analysis Pipeline

2.1 Data Preparation

The data read from the ASB source is in aggregated form to deidentify people for privacy purposes, so it was essential to prepare the income and migration data before analysis. Below is the sample of original data read from the ASB source.

STE_CODE_2021	index	1	2	3	4	5	6	7	8	9
0	C21_650_799_CF_1C	5935	5004	3250	1334	1526	491	277	143	3
1	C21_650_799_CF_2C	4385	3949	2292	943	1156	283	213	90	0
2	C21_650_799_CF_3C	2186	1643	954	420	488	126	126	54	0
3	C21_650_799_CF_4mC	1158	901	623	250	278	65	74	24	0
4	C21_650_799_Tot	13662	11498	7122	2950	3452	967	682	313	9

Table 4: Sample of 5 rows from the original income data.

STE_CODE_2021	index	1	2	3	4	5	6	7	8	9
0	Netherlands_C11_P	18214	21636	14988	7282	9978	2378	449	1120	0
1	Netherlands_C16_M	8538	9982	7237	3247	4585	1106	220	491	0
2	Netherlands_C16_F	8360	9834	6832	3353	4552	1091	202	550	3
3	Netherlands_C16_P	16898	19816	14066	6597	9130	2196	421	1041	3
4	Netherlands_C21_M	7865	9268	7022	3005	4361	1064	165	473	3

Table 4: Sample of 5 rows from the original migration data

To process the above-read income and migration aggregated data, the following three functions are created which will assist in preparing and arranging the data in a meaningful shape for analysis.

- add_year(value):

  The census year information in the data is in form of the attribute value of the index/Type_Value column, for example, if the value is “C11_” or “C21_” it means the census year is 2011 and 2021 respectively. To extract this census year information, the add_year() function is created.

- add_family_status(value):

  The wage earners are considered to be of two family status/types: couple families and single-parent families and it is present in the index/Type_Value column as “_1PF_” or “_CF_” informing Single Parent and Couple families. Thus, to derive this data add_family_status() function is applied.

- add_income_brackets(value):

  There are various income brackets starting from 149 to 4000 and again are present in the index/Type_Value column. Except for 4000, all the salary brackets are described as “up to 149” and “up to 299” whereas it’s “4000 and more” for the last income bracket. So, to draw this information add_income_brackets() function is generated.

Techniques	Attributes/Columns Involved	Summary
Preparation of **Income** Data		
Renaming Attribute	index -> Index_Value	Renamed the column named "index" to "Index_Value" which involves information on census years, family status and income brackets.
Renaming Attributes	1 – 9 -> Associated specific state names. For instance, 1 -> "New South Wales"	Renamed columns from 1 – 9 as particular states of Australia.
Adding New Attributes	New columns are added namely: total_all_states, income_bracket, census_year and family_status	Using the 3 functions the specific information from the index/Type_Value column is extracted and accordingly, new columns are added to the dataset
Casting Datatype	income_bracket, census_year to int	The data type of two columns income_bracket and census_year is type casted to 'int'.
Removing Incomplete Datapoints	income_bracket	Some people who have not shared their income details while data was being collected by ASB, those families are part of 0/Zero income bracket so removing these values.
Sorting Dataset	All the attributes of dataset	Finally, all the attributes are sorted by income_bracket and census_year
Preparation of **Migration** Data		
Renaming Attribute	index -> Index_Value	Renamed the column named "index" to "Index_Value" which consists of details of overseas birth country name, census years and gender of people.
Renaming Attributes	1 – 9 -> Associated specific state names. For instance, 1 -> "New South Wales"	Renamed columns from 1 – 9 as particular states of Australia.
Adding New Attributes	New columns are added namely: census_year, sex, country, total_all_states	Using the 1 function add_year() and string indexing methods the specific information from the index/Type_Value column is extracted and accordingly, new columns are added to the dataset
Cleaning Attribute Values	country	Cleaning heading/tailing tabs and correcting characters like "Unitd Kingdom" to "United Kingdom" and "Unit Sts Amer" to "United States" etc.

Table 2: Details of various methods used for the preparation of the income and migration datasets.

Additionally, to generate a world map to show the birth country of migrants, the country's ISO code is necessary and this data was downloaded from https://plotly.com/python/choropleth-maps/and is merged as a new column "iso_code" with the migration dataset.

Further, different functions (head, tail, sample, info, shape) and statistical methods (describe) are used to understand the data points information of both datasets. Please check the 1st, 2nd and 6th sections of the iPython notebook.

2.2 Missing Value Exploration

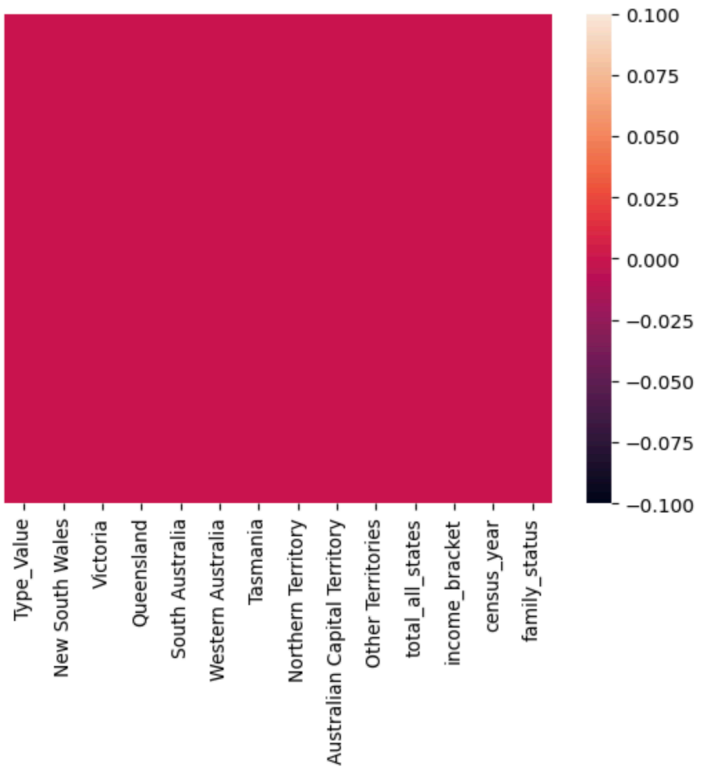

Figure: Income dataset's heatmap.

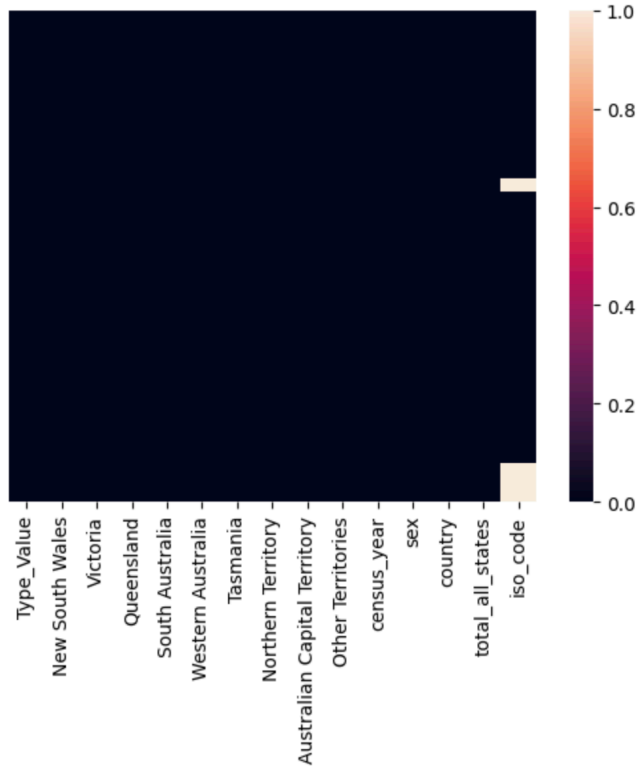

Figure: Migration dataset's heatmap.

The above Heatmaps, used in the 4th section of the iPython notebook, demonstrate if any missing values are present in the datasets. The income dataset doesn't have any missing values, however, there are missing values in the iso_code (abbreviations of countries) attribute of the migration dataset and would only be used to generate a world map.

2.3 Outlier Identification

The outliers in the numeric attributes of the income and migration dataset involve the number of families' earnings (in various salary brackets) and the migration of people from around the world to different states of Australia as represented and explained in section 7 of the iPython notebook (or pdf result) using box plots. For instance, outliers from numeric attributes 'New South Wales' of income and 'Victoria' of migration datasets are represented below.

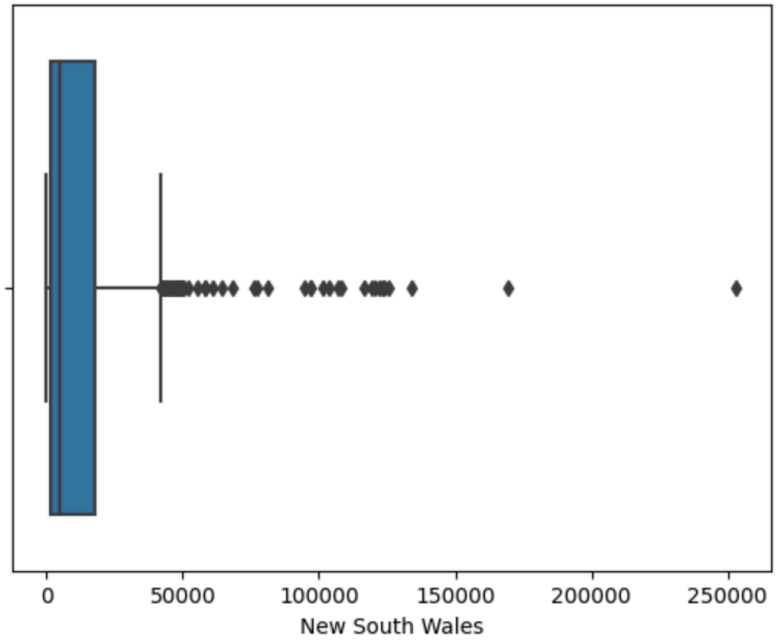

Figure 1: Outliers in the income dataset's 'New South Wales' attribute.

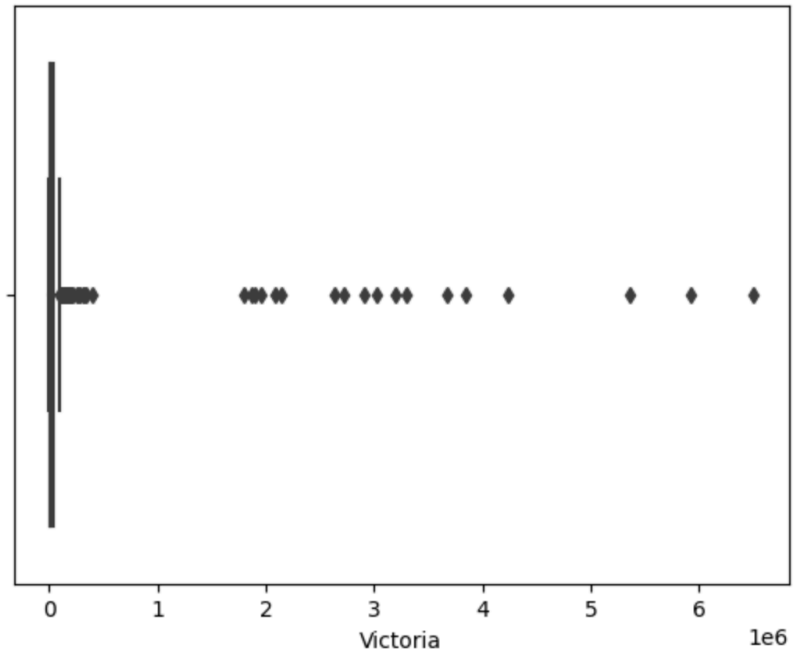

Figure 2: Outliers in the 'Victoria' attribute of the migration dataset.

Reserving the extreme values in both datasets as they explain an elevated number of families' earnings and a rise in immigration to different states of Australia from 2011 until 2021.

2.4 Data Visualization

Section 8 from the iPython notebook (or pdf result) describes and explains different data visualization graphs for various purposes. Below are some examples of graphs used in the income and migration dataset analysis.

a) Bar graphs to understand the distribution of values of categorical and numeric attributes.

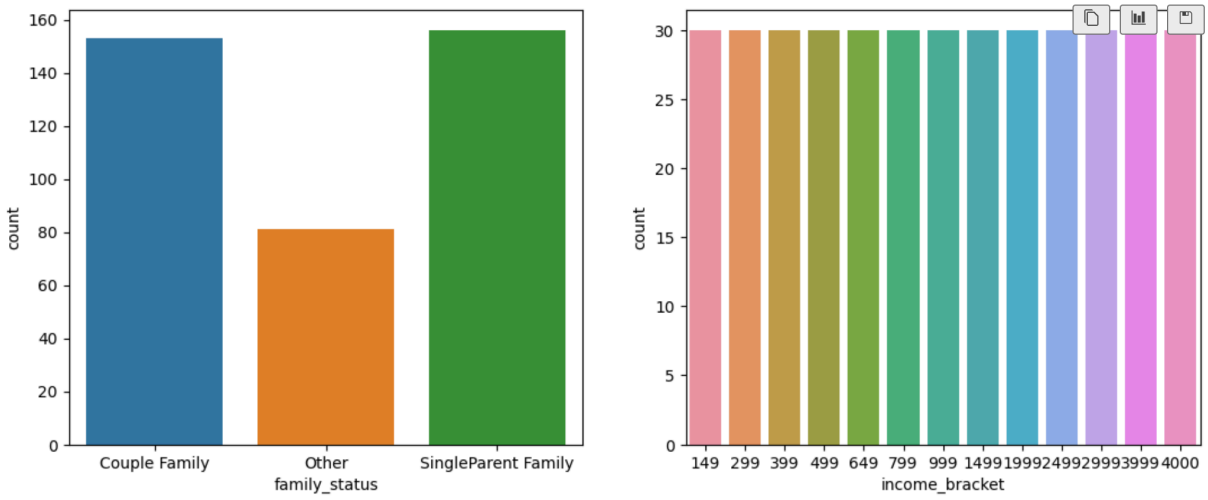

Figure: Distribution of data points using bar plots of family_status and income_bracket attributes of income dataset.

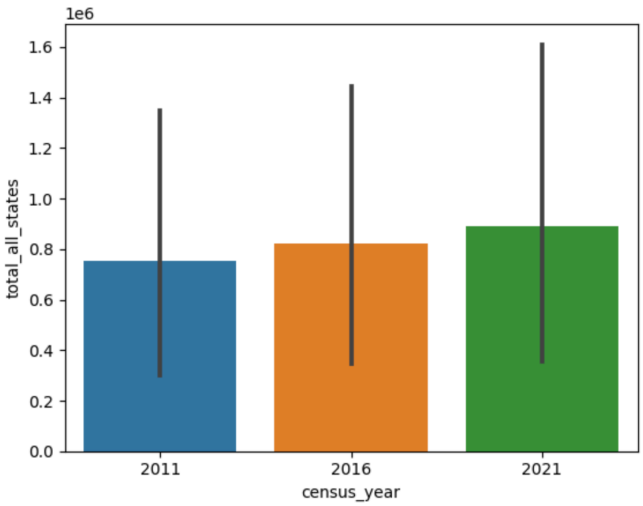

Figure: Distribution of the average number of migrants for 3 censuses from 2011 until 2021 of migration dataset.

b) Pie graphs to show the composition of values of the attributes.

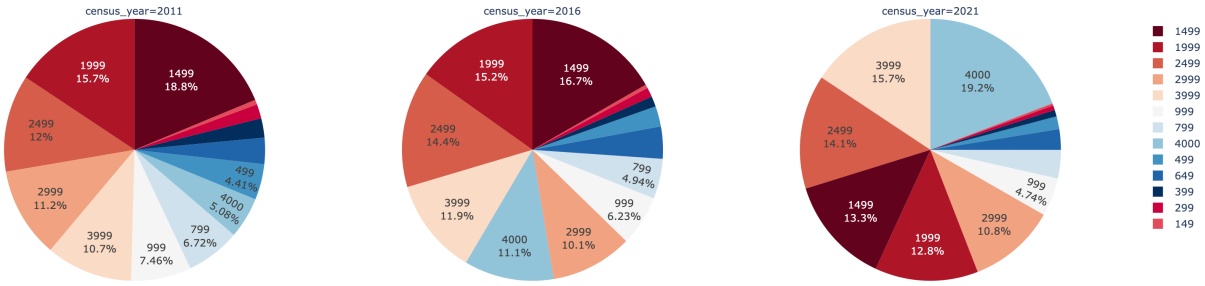

Figure: Composition of data points using pie plots of different income_brackets for 3 censuses of the income dataset.

c) Scatter graphs for identifying the relationship between various numeric and categorical attributes.

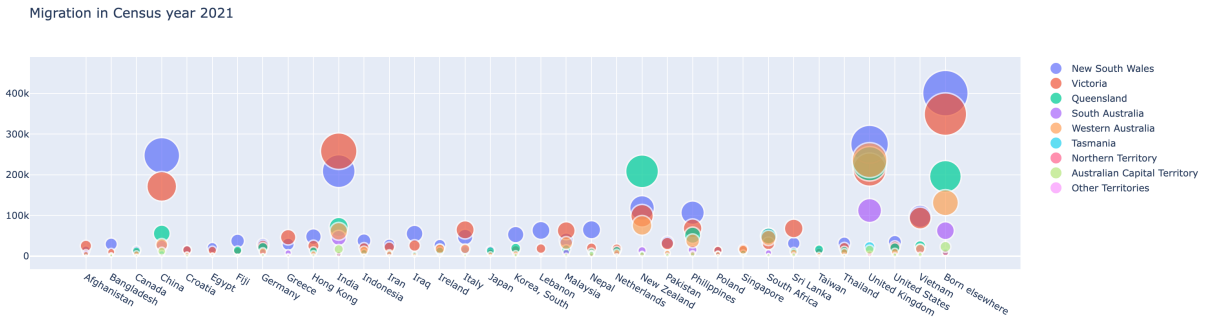

Figure: Scatter plot showing immigrants settling in different states of Australia from across the globe during census 2021.

d) Pair plots, associating best attributes while describing a relationship.

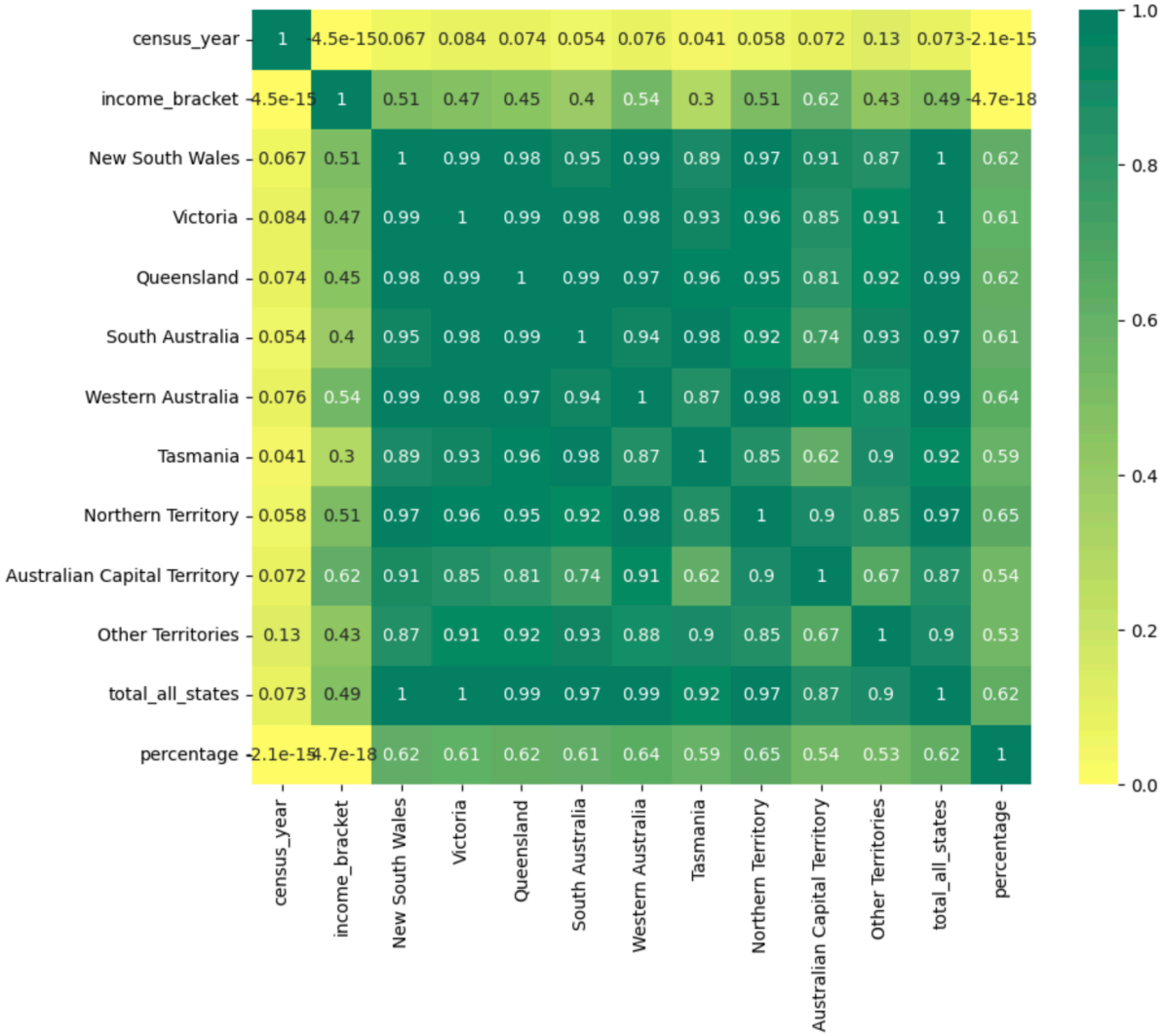

Figure: Pair plot identifying best columns for explaining a relationship of attributes of income dataset.

Please refer to sections 8.1 and 8.2 of the iPython notebook (or pdf result) for detailed data visualization presenting information and insights about income and migration datasets respectively.

3. Discussion and Conclusions

Income:

NSW, Victoria and Queensland are the top states where the highest number of salaried workers reside. While the ratio of single-parent families' income has improved compared to couple families. In 2011 and 2016, the best performing two salary brackets were 1499 and 1999, however, in 2021 it jumped and shifted to 4000(and more) and 3999 income brackets. The number of families' earnings in various income groups was almost similar in the 2011 and 2016 censuses as compared, in 2021 majority of the families' income earners are from higher income brackets.

Migration:

During the period from 2011 until 2021, a large number of migrants are from China and India. NSW, Victoria and Queensland are the top choices among all migrants for 3 censuses, where Indians prefer Victoria, in contrast, Chinese and British people preferred NSW. Queensland is the favourite destination for people migrating from New Zealand.

Please check section 8th of the iPython notebook (or pdf result) for detailed observations using various data visualization plots.

4. References

Australian Bureau of Statistics. (n.d.). About. https://www.abs.gov.au/about

Australian Bureau of Statistics. (n.d.). The Australian Census. https://www.abs.gov.au/census/about-census/australian-census

Australian Bureau of Statistics. (n.d.). INFORMING A NATION THE EVOLUTION OF THE AUSTRALIAN BUREAU OF STATISTICS 1905-2005 (Catalogue Number: 1382.0). ABS.

https://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/A8B7911F73578F1ACA2570AA00750101/$File/13820_2005.pdf