



Nuts&Dairy

5Nuts

Be Safe | Be Healthy

# Data Management Plan

Team No: TA20

Iteration 1

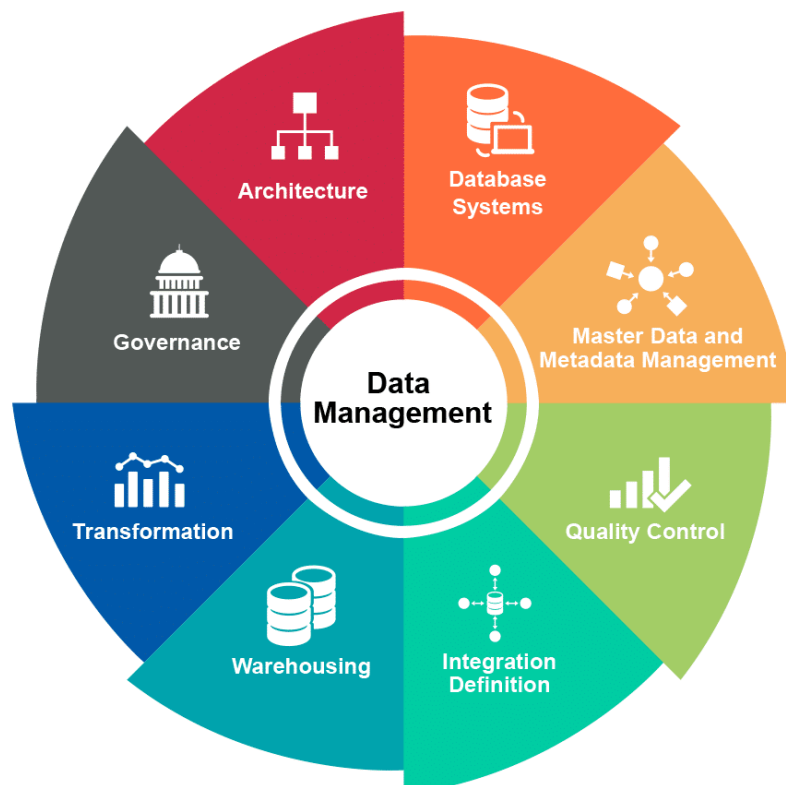


## Data Management Plan iteration 1

**Hindsight(Past):** With the kickstart of the IE this semester, we had formulated our problem statement along the lines of the food and the dairy industry. We were looking for datasets on effects on children, types of food for our databases and nutrients file. We were looking for a central database which contains all the food items and the nutrients so that we could build a recommendation platform. We were also searching for APIs, JSON data for google maps and hospitals data which could be integrated with our website.

**Insight(Present):** After a comprehensive research, we found datasets on the proportion of children suffering from Allergy in Australia, type of foods that could lead to allergies, various files on nutrients and ingredients. We also got a CSV list of food items which will act as a central database for our website. These files were found on websites like Australian Bureau of Statistics, Australian Institute of Health and Welfare, Allergy & Anaphylaxis Australia, Foodstandards.gov.au and Victorian Government Data. All these files are in the form of CSV and XLSX.

**Foresight(Future):** For iteration 1, we wish to do more research and gain knowledge on the APIs available for our website. APIs such as google maps, for games (quiz), for databases and hospital's location for our recommendations.



## Data Collection:

The Data is collected on the Australian government websites on the basis of surveys, SEHQ (School Entrant Health Questionnaire), and reviews including primary and secondary research conducted by Australian Bureau of Statistics. The Custodians of these data are the Department of Education and Training, Australian Bureau of Statistics, Victoria Public Data and Food Standards Australian Government.

## Data Storage:

For Iteration 1 the Data will be stored in CSV and XLSX files. As we move on to Iteration2, we will be handling more complex datasets which might need to be stored in SQL Databases.

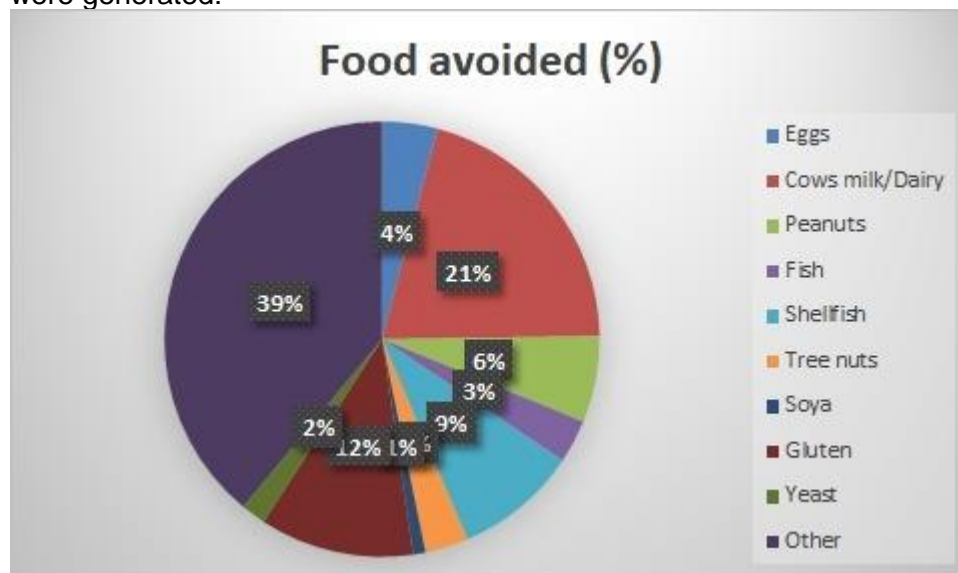
## Data Preprocessing steps, and Analysis

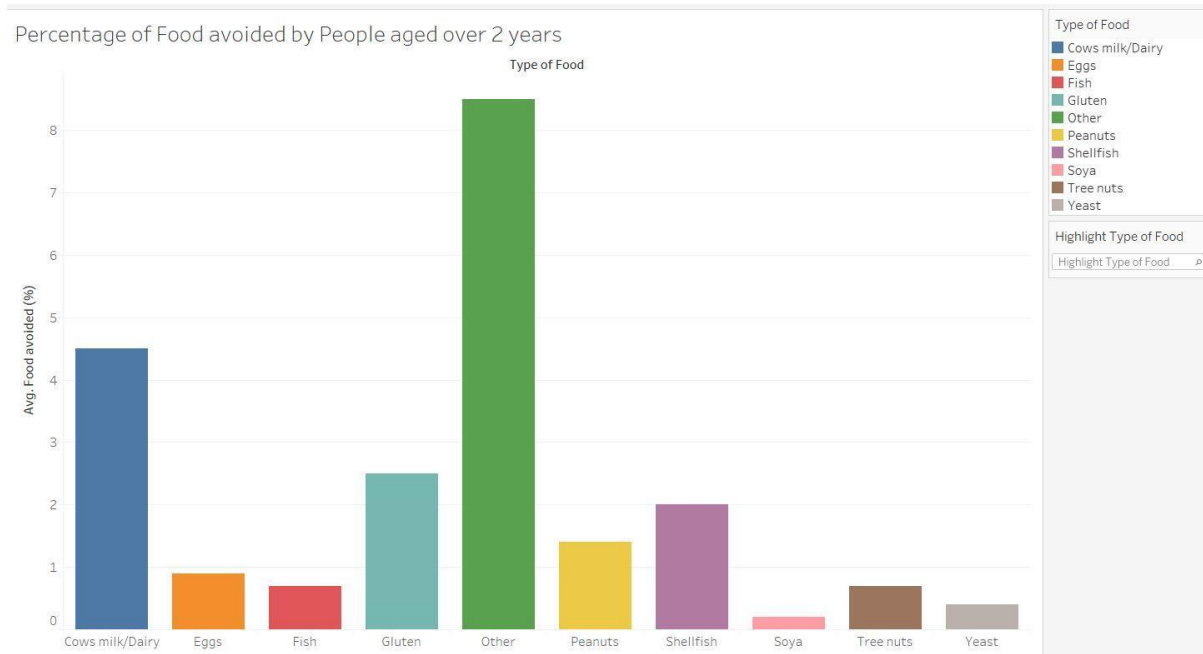
The following datasets are used for statistical analysis and visualization purposes  
Data Cleaning Steps done to clean the data:

1. Removing unnecessary data fields which is not required in the analysis  
Attributes to remove: Scientific Name, Food Profile ID, Classification ID, Classification Name
2. Checking for Null Values
3. Checking Data Consistency: The data type of the attributes were not in the proper format. The data was loaded in Excel and the format cells button was selected to change the format of the data.
4. Remove erroneous values which cause inaccurate analysis
5. Checking for missing values: The data did not have any missing value.
6. Outlier Analysis: No Outliers were recorded
7. Transforming the data of Victorian Population for ease of analysis

## Exploratory Data Analysis

The data was loaded into Excel and Tableau for initial analysis and the following insights were generated:

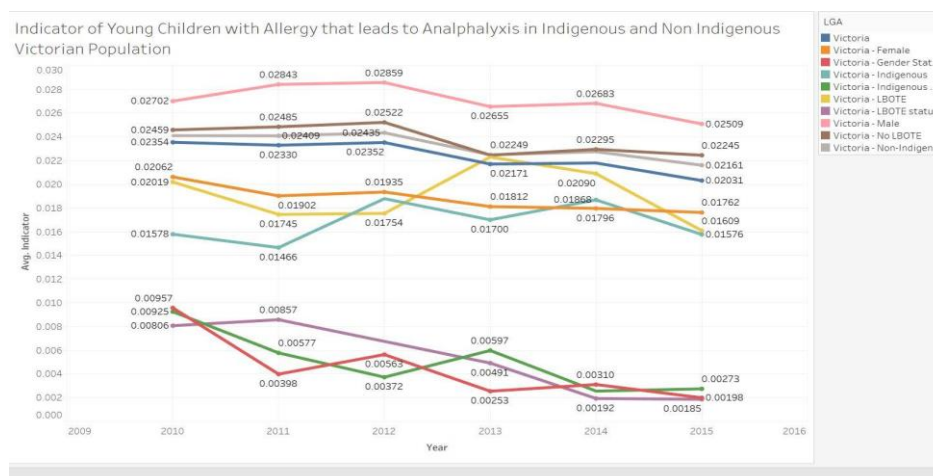




The above bar chart depicts the percentage of food avoided for people above age 2

From the above Analysis, Peanuts, Tree Nuts and Dairy Products account for 7.5% of the total.

The below chart represents the Indication of allergies which lead to Analphalyxis disease amongst Victorian Population



Nuts&Dairy

5Nuts

Be Safe | Be Healthy

# Data Management Plan

Team No: TA20

Iteration 2

# Introduction

The iteration 1 kickstarted with providing a comprehensive understanding to the problem of Nuts and Dairy allergies and the prevalence of the issue in Australia. The website captured user behaviour by providing the quiz where the user can understand the current level of knowledge.

Moving further, Iteration 2 emphasises on providing a descriptive statistic with a blend of preventative strategies where the user can learn the various treatment plans available for the issue. Along with it, the hero feature of the project includes creating a personalised lunch box plan according to the problem the user has.

## Datasets used

The Iteration 2 emphasises on creating a plan for the users to protect themselves and their families from food allergies. Therefore, the datasets used are:

1. Recipe File  
Data storage format: csv
  - a. Attributes:
    - i. Public Food key
    - ii. Food name
    - iii. Total Weight percentage change
    - iv. Ingredient Public Food Key
    - v. Ingredient Name
    - vi. Ingredient weight
    - vii. Retention Factor Id
2. Migration trends statistical package  
Data storage format: xlsx
  - a. Tables used: 56
  - b. Tables required: 2 (Table 1.6 and Table 1.5)
    - Attributes:
      1. Top 15 citizenship countries
3. Hospitals in Victoria  
Data format: csv  
Attributes
  1. Hospital ID
  2. Hospital Name
  3. Other name
  4. Emergency capability
  5. Agency type
  6. Suburb
  7. Postcode
  8. Location Address
  9. Public or Private hospital

# Data collection

**Hindsight:** Iteration 1 kickstarted with understanding user behaviour, while as we move on to Iteration2 we need to provide preventive strategies. We were able to get datasets on migration statistics and food recipes.

**Insight:** In the food recipes data, we wish to provide details on nutritional value on the foods selected by the user. The data is cleaned in the CSV format and sent to the database via a python script.

**Foresight:** We could not get contact details for some hospitals which might be a hinderance. While we do provide the location of the hospitals. We also wish to design a simple game for the children to play on our website.

## Data storage

After trying across various platforms, the final decision about data storage has been decided to use SQLite relational database. The main reason is that the project won't need a huge amount of database interaction, therefore, it is not necessary to have a MySQL database which needs to be hosted somewhere in a server.

SQLite gives the simple access from back-end technologies, Python3 in our case, and also can be expanded afterward when more features are required. The database has been designed to have 4 tables for current features, food, ingredient, nutrition and hospital.

The back-end side provides APIs for the front-end to access data when needed. Although the product doesn't support user insert into the database, parameterized query is mandatory to prevent second-order SQL injection.

## Data Usage: A Roadmap

In iteration1, the dataset is used for users to know more information. We display it as graphs to show the numbers and percentage. In iteration2, we will get the data of immigration in Australia and common lunch food for Australian kids. We will combine that information to let users build their own lunchbox and make some interactive games.

Dataset	Feature usage	Importance
Migration Trends	Depict a statistical analysis on the trends in migration from top 5 countries.	Ensuring a sense of safety and peace amongst the migrant people who are using this website. This builds a confidence in the parents that they are

		not alone in adapting to a new eating lifestyle in Australia.
Food Recipe	Provides recipes to the parents assisting in packing their lunch boxes or get innovative recipes for birthday parties or other events.	The busy and working parents of today have less time to plan the meals for their children. Amidst the global pandemic, and evolving tensions in work life, parents might forget to avoid the ingredients to which their children are allergic to. Therefore, this feature assists by notifying that the ingredient is a potential threat to the child.
Food Ingredient	Provides a list of ingredients in food items that the child might be prone to.	A lot of ingredients are food allergens such as milk products in cake. When we cook cakes, we mistakenly add ingredients which make the children prone. Therefore, this data interlinked with the food recipe acts as a central database for the project.

## Iteration planning requirements

Tools required:

- Wordpress
- Wordpress: Easy charts
- Elementor WP plugin: For creating drag and drop UI in Wordpress.
- HTML & CSS for frontend UI development
- Python or PHP: for backend development
- Database: To store the food recipes data
- Excel/CSV: To store the initial collected data from the website. Also useful for making initial data cleaning and statistical analysis.
- Python/Pandas/Numpy: Data clearing and wrangling.
- Tableau: Data Analysis and Visualizations for showing and articulating attractive storytelling.




- AWS: to host the website along with SSL certificate
- To get SSL certificate for website security
- Apache Charts: To integrate the visualization with HTML in frontend.
- Leankit & Google Drive: For Project Management and uploading and storing documents.

## Requirements

- Divide the datasets into different tables
- Clean the data using Python, Excel
- Design Database
- Write a python script to read csv file into database
- Find insights
- Normalise the database
- Plan for the hero feature and the UI required for it.

# Data Cleaning

## Before Cleaning (Migration Data)



Australian Government

Department of Home Affairs

Table 1.6: Skill stream outcome—top 15 citizenship countries, Regional visas, 2019–20<sup>1</sup>

Year	India	People's Republic of China	Philippines	Nepal	Pakistan	United Kingdom	South Africa	Sri Lanka	Vietnam	Republic of Korea	Brazil	Bangladesh	Iran	Egypt	Malaysia	Other	Total
2019–20	7,585	1,925	1,842	1,638	1,223	1,050	945	883	669	449	400	380	340	321	311	3,411	23,372

1. The Regional migration category commenced 1 July 2019.

Note: Includes both primary and secondary applicants. Top 15 citizenship countries are based on 2019–20.

[Click to return to contents](#)

Australian Migration Statistics—released November 2020

© Commonwealth of Australia 2020



Table 1.10: Skill stream outcome—top 15 citizenship countries, Skilled Independent visas, 2010–11 to 2019–20

Year	New Zealand <sup>1</sup>	India	People's Republic of China	Pakistan	Philippines	Nepal	United Kingdom	South Africa	Iran	Sri Lanka	Vietnam	Malaysia	United States of America	Canada	Singapore	Other <sup>2</sup>	Total
2010–11	n/a	7,326	9,823	727	789	839	2,868	1,508	1,178	1,529	314	1,850	203	155	488	6,570	36,167
2011–12	n/a	10,064	6,117	1,613	1,029	818	2,749	1,293	1,393	1,391	324	1,703	272	280	567	8,159	37,772
2012–13	n/a	14,970	6,032	672	1,095	1,635	3,440	1,120	734	1,409	336	1,811	386	322	721	9,568	44,251
2013–14	n/a	13,874	6,076	1,966	1,824	1,680	4,139	1,307	837	1,153	549	1,006	370	416	793	8,394	44,984
2014–15	n/a	11,165	5,922	4,530	2,107	825	3,721	1,051	1,419	1,093	542	1,232	406	343	760	8,874	43,990
2015–16	n/a	13,343	5,509	2,991	2,402	1,272	3,947	1,511	970	833	568	1,371	415	356	661	7,845	43,994
2016–17	n/a	13,781	5,991	2,844	2,596	1,032	3,074	1,828	1,019	658	375	1,121	404	297	553	6,849	42,422
2017–18	4,441	12,161	5,067	2,422	1,910	843	2,320	1,312	743	517	514	785	333	287	437	5,045	39,137
2018–19	5,517	11,741	4,160	1,353	1,515	1,089	1,811	852	678	510	425	473	433	257	256	3,177	34,247
2019–20	4,300	3,225	1,367	451	439	427	425	236	204	189	156	156	128	98	90	1,095	12,986

1. New Zealand citizen permanent visa grant numbers, before 2017–18, were not recorded against the Migration Program outcome.

2. Includes citizenship not specified.

Note: Includes both primary and secondary applicants. Top 15 citizenship countries are based on 2019–20.

## Family Stream outcome visas for top 15 countries



Table 1.17: Family stream outcome—top 15 citizenship countries, Partner visas, 2010–11 to 2019–20

Year	People's Republic of China	United Kingdom	India	Philippines	Vietnam	Afghanistan	United States of America	Thailand	Pakistan	Republic of Korea	Indonesia	Canada	Brazil	Taiwan	Germany	Other <sup>1</sup>	Total
2010–11	4,952	4,474	3,649	2,492	2,607	736	1,693	1,754	591	792	961	808	354	283	612	15,236	41,994
2011–12	5,140	4,545	4,468	3,287	2,807	640	1,807	1,740	840	845	923	865	446	363	715	15,719	45,150
2012–13	5,343	4,643	5,389	3,137	2,707	253	1,920	1,925	913	1,025	968	1,026	438	494	733	15,411	46,325
2013–14	5,366	4,339	5,175	3,331	2,832	557	1,966	1,816	1,360	1,058	930	910	425	488	660	16,539	47,752
2014–15	5,631	3,979	5,233	3,191	2,581	1,838	1,681	1,696	1,192	996	943	859	410	525	592	16,478	47,825
2015–16	5,865	4,055	5,503	3,354	2,654	1,349	1,823	1,941	950	884	890	914	445	627	667	15,904	47,825
2016–17	5,636	4,064	4,972	3,296	2,862	1,536	1,759	1,920	1,059	961	917	877	553	670	652	16,091	47,825
2017–18	4,249	3,175	3,625	3,160	2,455	2,200	1,065	1,775	1,247	604	1,086	520	538	571	367	13,162	39,799
2018–19	4,850	2,659	3,803	2,234	2,697	1,956	1,778	1,550	1,228	964	674	666	629	547	570	13,113	39,918
2019–20	3,553	2,788	2,684	2,338	2,245	2,220	1,758	1,340	1,044	876	856	773	769	673	605	12,596	37,118

1. Includes citizenship not specified.

## Family and Child Outcome Visas



Table 1.18: All other Family stream outcome and Child outcome visas—top 15 citizenship countries, Parent, Other Family and Child visas, 2010–11 to 2019–20

Year	People's Republic of China	India	Vietnam	United Kingdom	Philippines	Thailand	Sri Lanka	South Africa	Malaysia	Indonesia	Pakistan	United States of America	Cambodia	New Zealand <sup>1</sup>	Republic of Korea	Other <sup>2</sup>
2010–11	4,125	777	716	1,332	471	287	254	515	336	260	80	101	154	n/a	224	2,917
2011–12	4,563	1,021	878	1,201	572	326	265	456	278	175	125	89	142	n/a	263	3,100
2012–13	5,085	1,109	1,009	999	638	301	253	409	322	219	68	131	109	n/a	238	2,970
2013–14	4,961	1,120	851	928	574	282	244	313	393	259	86	122	184	n/a	218	2,825
2014–15	5,394	1,029	887	807	546	340	211	346	365	176	110	104	109	n/a	210	2,626
2015–16	5,561	1,087	882	690	537	302	217	217	363	173	154	96	112	n/a	185	2,511
2016–17	5,013	971	817	549	557	257	198	188	286	173	162	97	102	n/a	153	2,284
2017–18	5,013	1,000	752	447	572	268	208	209	254	126	227	116	99	118	113	1,761
2018–19	4,465	1,061	638	483	561	240	160	191	224	127	140	144	88	141	103	1,811
2019–20	2,439	839	582	433	400	218	159	156	151	130	126	125	103	89	75	1,299

1. New Zealand citizen permanent visa grant numbers, before 2017–18, were not recorded against the Migration Program.

2. Includes citizenship not specified.

Note: Includes both primary and secondary applicants. Top 15 citizenship countries are based on 2019–20.

## Dataset After cleaning

Year	New Zealand <sup>1</sup>	India	People's Republic of China	Pakistan	Philippines	Nepal	United Kingdom	South Africa	Iran	Sri Lanka	Vietnam	Malaysia	United States of America	Canada	Singapore	Other <sup>2</sup>	Total
2010-11	n/a	7,326	9,823	727	789	839	2,868	1,508	1,178	1,529	314	1,850	203	155	488	6,570	36,167
2011-12	n/a	10,064	6,117	1,613	1,029	818	2,749	1,293	1,393	1,391	324	1,703	272	280	567	8,159	37,772
2012-13	n/a	14,970	6,032	672	1,095	1,635	3,440	1,120	734	1,409	336	1,811	386	322	721	9,568	44,251
2013-14	n/a	13,874	6,076	1,966	1,824	1,680	4,139	1,307	837	1,153	549	1,606	370	416	793	8,394	44,984
2014-15	n/a	11,165	5,922	4,530	2,107	825	3,721	1,051	1,419	1,093	542	1,232	406	343	760	8,874	43,990
2015-16	n/a	13,343	5,509	2,991	2,402	1,272	3,947	1,511	970	833	568	1,371	415	356	661	7,845	43,994
2016-17	n/a	13,781	5,991	2,844	2,596	1,032	3,074	1,828	1,019	658	375	1,121	404	297	553	6,849	42,422
2017-18	4,441	12,161	5,067	2,422	1,910	843	2,320	1,312	743	517	514	785	333	287	437	5,045	39,137
2018-19	5,517	11,741	4,160	1,353	1,515	1,089	1,811	852	678	510	425	473	433	257	256	3,177	34,247
2019-20	4,300	3,225	1,367	451	439	427	425	236	204	189	156	156	128	98	90	1,095	12,986

## List of Victorian Hospitals Data Before cleaning

	A	B	C	D	E	F	G	H	I	J	K	L
1	Hospital ID	Formal Name	Other Name	Emergency	Location Address	Suburb	Postcode	Access Point	Category	Agency Type		
2	5488	Albert Road	Albert Road	NO	31-33 Albert Road	South Melbourne	3205		PRIVATE	Private Hospital		
3	3485	Albury Wood	Albury Wood	YES	69 Vermont Street	Wodonga	3690		PUBLIC	Public Hospital		
4	12990	Albury Wood	Albury Wood	YES	Borella	Albury	2640		PUBLIC	Public Hospital		
5	3491	Alexandra	Alexandra	NO	20 Cooper Street	Alexandra	3714		PUBLIC	Public Hospital		
6	11519	Alfred Heath	Alfred Heath	NO	Commercial Road	Melbourne	3004		PUBLIC	Public Hospital		
7	5618	Alpine Heath	Alpine Heath	NO	30 O'Donnell Avenue	Myrtleford	3737		PUBLIC	Multi Purpose Service		
8	3483	Angliss Hospital	Angliss Hospital	YES	Albert Street	Upper Ferntree Gully	3156		PUBLIC	Public Hospital		
9	12266	Appearance	Appearance	NO	57 Garsed Street	Bendigo	3550		PRIVATE	Day Procedure Centre		
10	3401	Austin Heath	Austin Heath	NO	145 Studley Road	Heidelberg	3084		PUBLIC	Public Hospital		
11	11733	Austin Heath	Austin Heath	YES	145 Studley Road	Heidelberg	3084		PUBLIC	Public Hospital		
12	11744	Austin Heath	Austin Heath	NO	300 Waterdale Road	Heidelberg	3081		PUBLIC	Public Hospital		
13	6272	Bairnsdale	Bairnsdale	YES	122 Day St	Bairnsdale	3875		PUBLIC	Public Hospital		
14	6057	Ballan District	Ballan District	NO	33 Cowie Street	Ballan	3342		PRIVATE	Bush Nursing Hospital		
15	11920	Ballarat District	Ballarat District	NO	1119/1123 Howitt Street	Ballarat	3355		PRIVATE	Day Procedure Centre		
16	7783	Ballarat Health	Ballarat Health	YES	Drummond Street North	Ballarat	3353		PUBLIC	Public Hospital		
17	10441	Barwon Health	Barwon Health	YES	272-322 Ryrie Street	Geelong	3220		PUBLIC	Public Hospital		
18	6884	Bass Coast	Bass Coast	NO	235 Graham Street	Wonthaggi	3995		PUBLIC	Public Hospital		
19	11533	Bayside District	Bayside District	NO	141 Cranbourne Road	Frankston	3199		PRIVATE	Day Procedure Centre		
20	10765	Bayside Emergency	Bayside Emergency	NO	441 South Road	Moorabbin	3189		PRIVATE	Day Procedure Centre		
21	12058	Bayswater	Bayswater	NO	664 Mountain Highway	Bayswater	3153		PRIVATE	Day Procedure Centre		
22	3609	Beaufort	Beaufort	NO	28 Havelock Street	Beaufort	3373		PUBLIC	Public Hospital		
23	3377	Beechworth	Beechworth	NO	52 Sydney Road	Beechworth	3747		PUBLIC	Public Hospital		
24	2416	Beleura Private	Beleura Private	NO	925 Nepean Highway	Mornington	3931		PRIVATE	Private Hospital		
25	2418	Bellbird Private	Bellbird Private	NO	198 Canterbury Road	Blackburn	3130		PRIVATE	Private Hospital		
26	3217	Benalla Health	Benalla Health	NO	45-63 Coster Street	Benalla	3672		PUBLIC	Public Hospital		
27	12695	Bendigo District	Bendigo District	NO	1-7 Chum Street	Bendigo	3550		PRIVATE	Day Procedure Centre		
28	3292	Bendigo Health	Bendigo Health	YES	62 Lucan St	Bendigo	3550		PUBLIC	Public Hospital		
29	12472	Bentleigh South	Bentleigh South	NO	155 Jasper Road	Bentleigh	3204		PRIVATE	Day Procedure Centre		

## Process required for data cleaning

- Check for null values
- Check for datatype of each column
- Apply filters
- Removing Non-Emergency hospitals
- Use of Excel → Apply Vlookup for checking if other names and hospital formal names are the same or not. If the same, then remove. In our case the formal name and the other name are the same. So, we remove the other name.
- Remove access points.
- Check for suburb
- check for postcode validation
- Check for data redundancies.

## After Cleaning Hospitals Data

Hospital ID	Formal Name	Emergency Capable	Location Address	Suburb	Postcode	Category	Agency Type
3485	Albury Wodonga Health	YES	69 Vermont Street	Wodonga	3690	PUBLIC	Public Hospital
12990	Albury Wodonga Health, Albury Ca	YES	Borella	Albury	2640	PUBLIC	Public Hospital
3483	Angliss Hospital	YES	Albert Street	Upper Ferntree Gully	3156	PUBLIC	Public Hospital
11733	Austin Health - Austin Hospital	YES	145 Studley Road	Heidelberg	3084	PUBLIC	Public Hospital
6272	Bairnsdale Regional Health Service	YES	122 Day St	Bairnsdale	3875	PUBLIC	Public Hospital
7783	Ballarat Health Services	YES	Drummond Street Nor	Ballarat	3353	PUBLIC	Public Hospital
10441	Barwon Health	YES	272-322 Ryrie Street	Geelong	3220	PUBLIC	Public Hospital
3292	Bendigo Health	YES	62 Lucan St	Bendigo	3550	PUBLIC	Public Hospital
3403	Box Hill Hospital	YES	51 Nelson Road	Box Hill	3128	PUBLIC	Public Hospital
12280	Casey Hospital	YES	52 Kangan Drive	Berwick	3806	PUBLIC	Public Hospital
11269	Central Gippsland Health Service	YES	155 Guthridge Parade	Sale	3850	PUBLIC	Public Hospital
3405	Dandenong Hospital	YES	105-135 David Street	Dandenong	3175	PUBLIC	Public Hospital
3052	Echuca Regional Health	YES	9-27 Frances Street	Echuca	3564	PUBLIC	Public Hospital
2447	Epworth Richmond	YES	89 Bridge Road	Richmond	3121	PRIVATE	Private Hospital
3407	Footscray Hospital	YES	160 Gordon Street	Footscray	3011	PUBLIC	Public Hospital
3465	Frankston Hospital	YES	Hastings Rd	Frankston	3199	PUBLIC	Public Hospital
6333	Goulburn Valley Health	YES	2 Graham Street	Shepparton	3630	PUBLIC	Public Hospital
6430	Latrobe Regional Hospital	YES	Cnr Princes Highway &	Traralgon West	3844	PUBLIC	Public Hospital
3416	Maroondah Hospital	YES	1-15 Mt Dandenong R	East Ringwood	3135	PUBLIC	Public Hospital
11607	Mercy Hospital for Women	YES	163 Studley Road	Heidelberg	3084	PUBLIC	Denominational Ho
3390	Monash Medical Centre, Clayton C	YES	246 Clayton Road	Clayton	3168	PUBLIC	Public Hospital
6802	Northeast Health Wangaratta	YES	35-47 Green Street	Wangaratta	3677	PUBLIC	Public Hospital
3466	Rosebud Hospital	YES	1527 Point Nepean Rc	Rosebud	9939	PUBLIC	Public Hospital
3424	Royal Melbourne Hospital - City C	YES	300 Grattan Street	Parkville	3050	PUBLIC	Public Hospital
3425	Sandringham Hosoiatal	YES	193 Bluff Road	Sandringham	3191	PUBLIC	Public Hospital

## Descriptive EDA Code Snippets

Checking null values

Removing unnecessary columns and renaming column names for easy of use

```
In [11]: 1 hospital_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 269 entries, 0 to 268
Data columns (total 10 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Hospital ID                          269 non-null   int64
1   Formal Name                          269 non-null   object
2   Other Name                           269 non-null   object
3   Emergency Capable                    269 non-null   object
4   Location Address                     269 non-null   object
5   Suburb                              269 non-null   object
6   Postcode                             269 non-null   int64
7   Access Point                         0 non-null     float64
8   Category                             269 non-null   object
9   Agency Type                          269 non-null   object
dtypes: float64(1), int64(2), object(7)
memory usage: 21.1+ KB
```

```
In [13]: 1 hospital_df.isnull().sum()
```

```
Out[13]: Hospital ID          0
Formal Name          0
Other Name           0
Emergency Capable    0
Location Address     0
Suburb               0
Postcode             0
Access Point         269
Category             0
```

```
In [18]: 1 hospital_df.drop('Other Name',
2 axis='columns', inplace=True)
```

```
In [20]: 1 hospital_df.head()
```

```
Out[20]:
```

	Hospital ID	Formal Name	Emergency Capable	Location Address	Suburb	Postcode	Access Point	Category	Agency Type
0	5488	Albert Road Clinic	NO	31-33 Albert Road	South Melbourne	3205	NaN	PRIVATE	Private Hospital
1	3485	Albury Wodonga Health	YES	69 Vermont Street	Wodonga	3690	NaN	PUBLIC	Public Hospital
2	12990	Albury Wodonga Health, Albury Campus	YES	Borella	Albury	2640	NaN	PUBLIC	Public Hospital
3	3491	Alexandra District Health	NO	20 Cooper Street	Alexandra	3714	NaN	PUBLIC	Public Hospital
4	11519	Alfred Health	NO	Commercial Road	Melbourne	3004	NaN	PUBLIC	Public Hospital

```
In [21]: 1 hospital_df = hospital_df.rename(columns={"Formal Name": "Hospital Name"})
2
```

```
In [23]: 1 hospital_df.head(10)
```

```
Out[23]:
```

	Hospital ID	Hospital Name	Emergency Capable	Location Address	Suburb	Postcode	Access Point	Category	Agency Type
0	5488	Albert Road Clinic	NO	31-33 Albert Road	South Melbourne	3205	NaN	PRIVATE	Private Hospital

## Checking Unique values of Hospitals

### Checking unique values of hospitals

```
In [30]: 1 n = len(pd.unique(hospital_df['Hospital Name']))
```

```
In [31]: 1 n
```

```
Out[31]: 269
```

# Data Governance: Milestones & Roadblocks

Data Governance is Imperative to ensure high quality data through a complete data lifecycle. Quality data is important for us for efficient analysis and to get valuable insights. Also since we are dealing with health related issues, it is seemingly more important that we provide accurate information and no bogus information is there.

## Milestones

- Additional information on nutritional value which is suited for children's health
- Regulatory requirements.
- Improved Data Quality.
- Resulting in Accurate representation of ER Diagram.
- Gives a 360-degree view of what might happen during the process of implementation for our project.

- Capture best insights
- 

#### Roadblocks

- No Contact information for emergency purposes.
- Might not capture Australian lifestyle.
- Very high granularity of the migration data might confuse the users.

## Data Exploration & Analysis

```
In [7]: 1 df_proportion=pd.read_csv("Persons aged 2 years and over - type of food avoided due to allergy or intolerance(a), 2011-12.csv")
<
In [8]: 1 df_proportion.head()
Out[8]:
```

	Type of Food	Food avoided (%)
0	Eggs	0.9
1	Cows milk/Dairy	4.5
2	Peanuts	1.4
3	Fish	0.7
4	Shellfish	2.0

```
In [14]: 1 df_proportion.columns
Out[14]: Index(['Type of Food', 'Food avoided (%)', dtype='object')

In [15]: 1 df_proportion.dtypes
Out[15]: Type of Food      object
Food avoided (%)    float64
dtype: object
```

We can see that the columns are taken as objects.

```
In [8]: 1 hospital_df.columns
Out[8]: Index(['Hospital ID', 'Formal Name', 'Other Name', 'Emergency Capable',
'Location Address', 'Suburb', 'Postcode', 'Access Point', 'Category',
'Agency Type'],
dtype='object')

In [10]: 1 hospital_df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 269 entries, 0 to 268
Data columns (total 10 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Hospital ID           269 non-null    int64
1   Formal Name           269 non-null    object
2   Other Name            269 non-null    object
3   Emergency Capable     269 non-null    object
4   Location Address      269 non-null    object
5   Suburb                269 non-null    object
6   Postcode              269 non-null    int64
7   Access Point          0 non-null      float64
8   Category              269 non-null    object
9   Agency Type           269 non-null    object
dtypes: float64(1), int64(2), object(7)
memory usage: 21.1+ KB
```

Hospital id is an artifact which is not useful for our analysis, nor required to be shown on the website, therefore we can remove it. Since there are no redundancies and no other cleaning required. We can export it into CSV.

Finally

## Hospitals Data After cleaning

Hospital ID	Formal Name	Emergency Capable	Location Address	Suburb	Postcode	Category	Agency Type
3485	Albury Wodonga Health	YES	69 Vermont Street	Wodonga	3690	PUBLIC	Public Hospital
12990	Albury Wodonga Health, Albury Ca	YES	Borella	Albury	2640	PUBLIC	Public Hospital
3483	Angliss Hospital	YES	Albert Street	Upper Ferntree Gully	3156	PUBLIC	Public Hospital
11733	Austin Health - Austin Hospital	YES	145 Studley Road	Heidelberg	3084	PUBLIC	Public Hospital
6272	Bairnsdale Regional Health Service	YES	122 Day St	Bairnsdale	3875	PUBLIC	Public Hospital
7783	Ballarat Health Services	YES	Drummond Street Noi	Ballarat	3353	PUBLIC	Public Hospital
10441	Barwon Health	YES	272-322 Ryrie Street	Geelong	3220	PUBLIC	Public Hospital
3292	Bendigo Health	YES	62 Lucan St	Bendigo	3550	PUBLIC	Public Hospital
3403	Box Hill Hospital	YES	51 Nelson Road	Box Hill	3128	PUBLIC	Public Hospital
12280	Casey Hospital	YES	52 Kangan Drive	Berwick	3806	PUBLIC	Public Hospital
11269	Central Gippsland Health Service	YES	155 Guthridge Parade	Sale	3850	PUBLIC	Public Hospital
3405	Dandenong Hospital	YES	105-135 David Street	Dandenong	3175	PUBLIC	Public Hospital
3052	Echuca Regional Health	YES	9-27 Frances Street	Echuca	3564	PUBLIC	Public Hospital
2447	Epworth Richmond	YES	89 Bridge Road	Richmond	3121	PRIVATE	Private Hospital
3407	Footscray Hospital	YES	160 Gordon Street	Footscray	3011	PUBLIC	Public Hospital
3465	Frankston Hospital	YES	Hastings Rd	Frankston	3199	PUBLIC	Public Hospital
6333	Goulburn Valley Health	YES	2 Graham Street	Shepparton	3630	PUBLIC	Public Hospital
6430	Latrobe Regional Hospital	YES	Cnr Princes Highway &	Traralgon West	3844	PUBLIC	Public Hospital
3416	Maroondah Hospital	YES	1-15 Mt Dandenong R	East Ringwood	3135	PUBLIC	Public Hospital
11607	Mercy Hospital for Women	YES	163 Studley Road	Heidelberg	3084	PUBLIC	Denominational Ho
3390	Monash Medical Centre, Clayton	YES	246 Clayton Road	Clayton	3168	PUBLIC	Public Hospital
6802	Northeast Health Wangaratta	YES	35-47 Green Street	Wangaratta	3677	PUBLIC	Public Hospital
3466	Rosebud Hospital	YES	1527 Point Nepean R	Rosebud	3939	PUBLIC	Public Hospital
3424	Royal Melbourne Hospital - City C	YES	300 Grattan Street	Parkville	3050	PUBLIC	Public Hospital
3425	Sandringham Hospital	YES	193 Bluff Road	Sandringham	3191	PUBLIC	Public Hosoiat

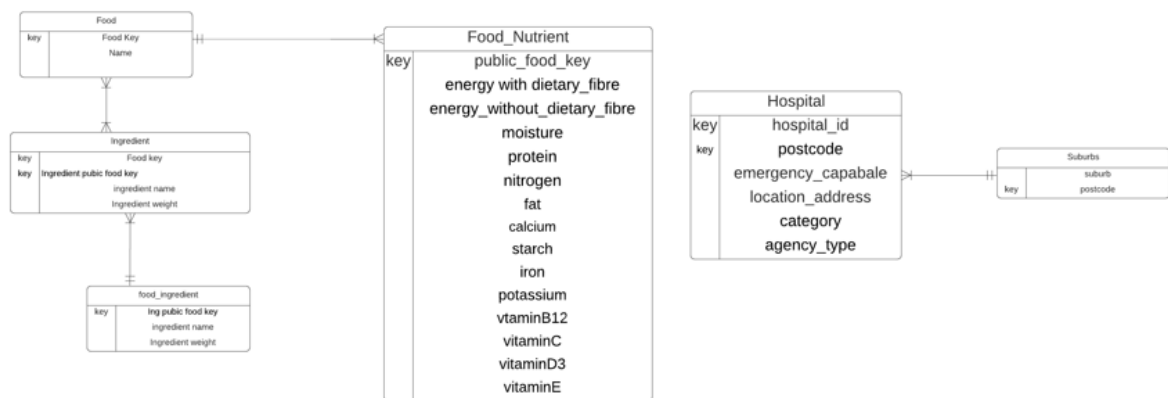
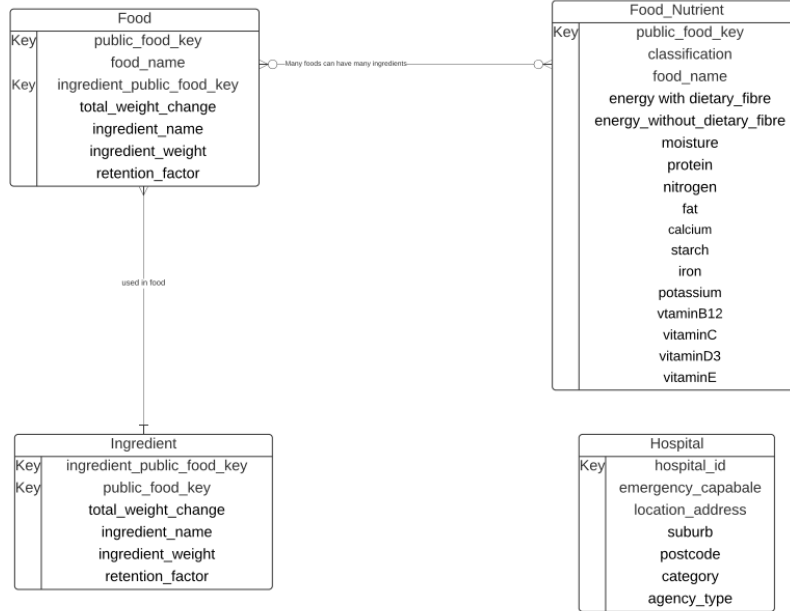
## Database creation

The food recipe file along with the Nutrition data and the hospitals data act as a central database for Iteration2. The user eats the food and if gets affected by allergies, he is admitted to the hospital.

The team TA20 wishes to take the common food recipes as the database and also ensures that the hospital's data is open data and does not violate any rights.

## ERDiagram







## Normalisation Database Schema

Food(public\_food\_key, food\_name)

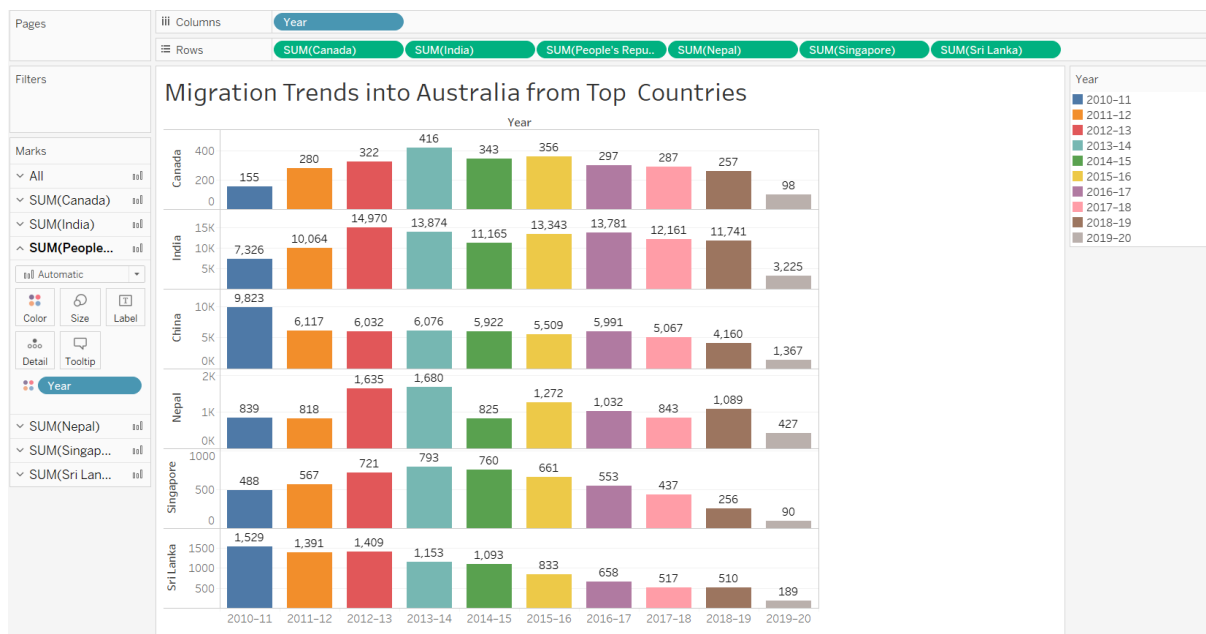
Ingredients(ingredient\_public\_food\_key, \*public\_food\_key, ingredient\_name, ingredient\_weight, retention\_factor\_ID)

Nutrient(public\_food\_key, classification, food\_name, energy\_with\_dietary\_fibre, energy\_without\_dietary\_fibre, moisture, protein, fat, dietary\_fibre, calcium, sodium, vitamin B12, vitamin C, vitamin D3 vitamin E, saturated\_fatty\_acid\_percent, saturated\_fatty\_acid\_gram)

Hospitals(hospital\_ID, formal\_name, other\_name, emergency\_capable, location\_address, suburb, postcode, access\_point, category, agency\_type)

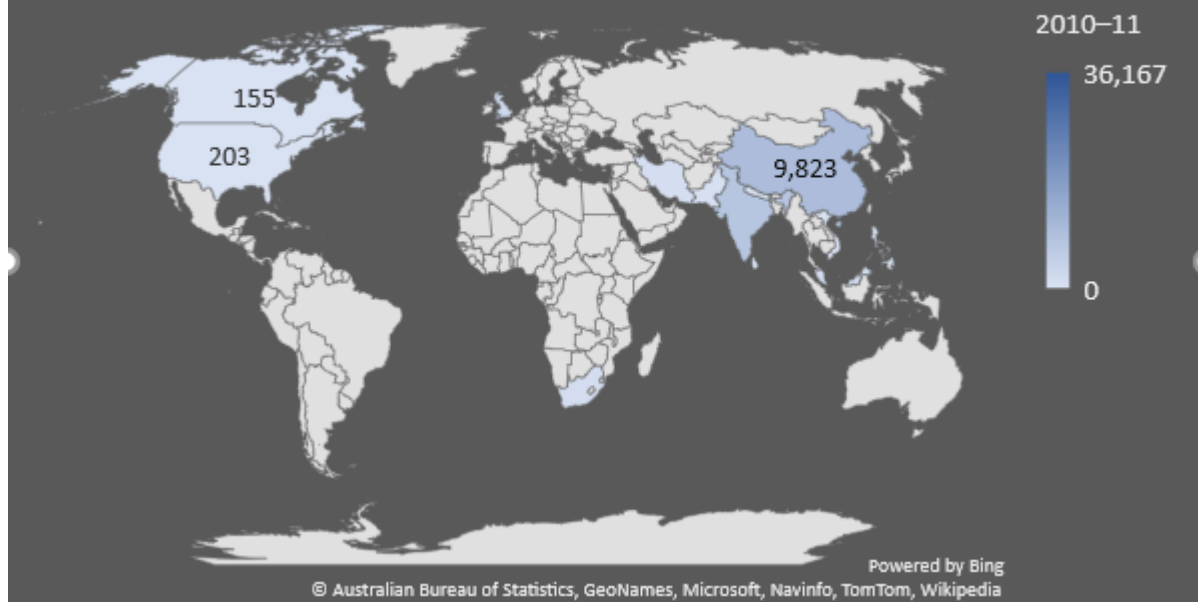
The above figure represents the database schema for the centralised database for our website.

## Data visualization

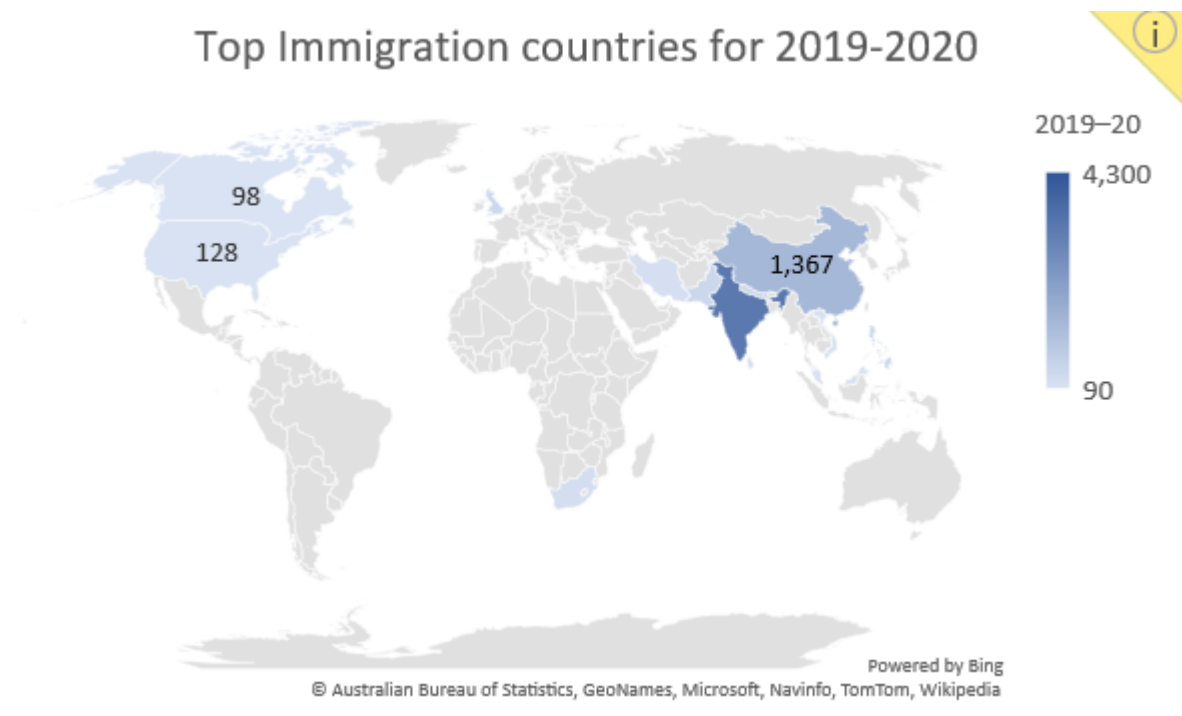


Comparison of with top most countries from which people migrate to Australia from 2010-2011 and 2019-2020

## Top Immigration Countries into Australia 2010-2011



## Top Immigration countries for 2019-2020



A dashboard depicting the migration trends from various APAC countries into Australia

### Migration Trends for Australia by various countries

