Data Warehouse Design & ETL Implementation Using Informatica - Capstone Project

+ | - | +/-

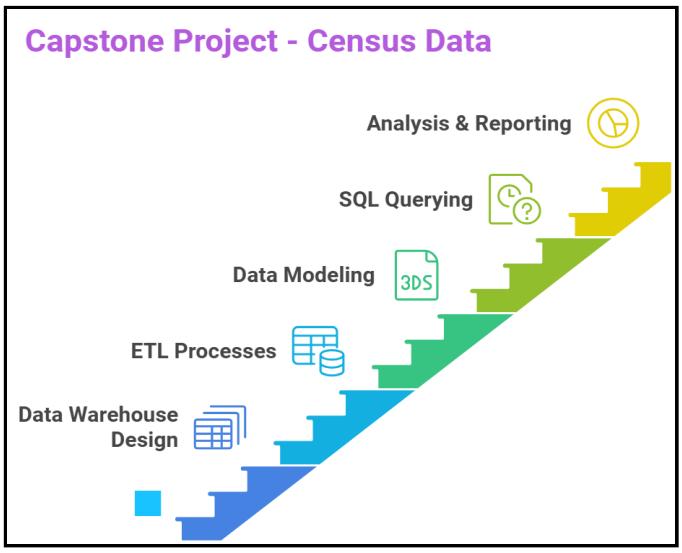
Objective

. To design and implement a data warehouse

. Extracting insights from the Census datasets provided

. ETL Processes, Data Modelling, SQL querying using Informatica

. Analysis & reporting

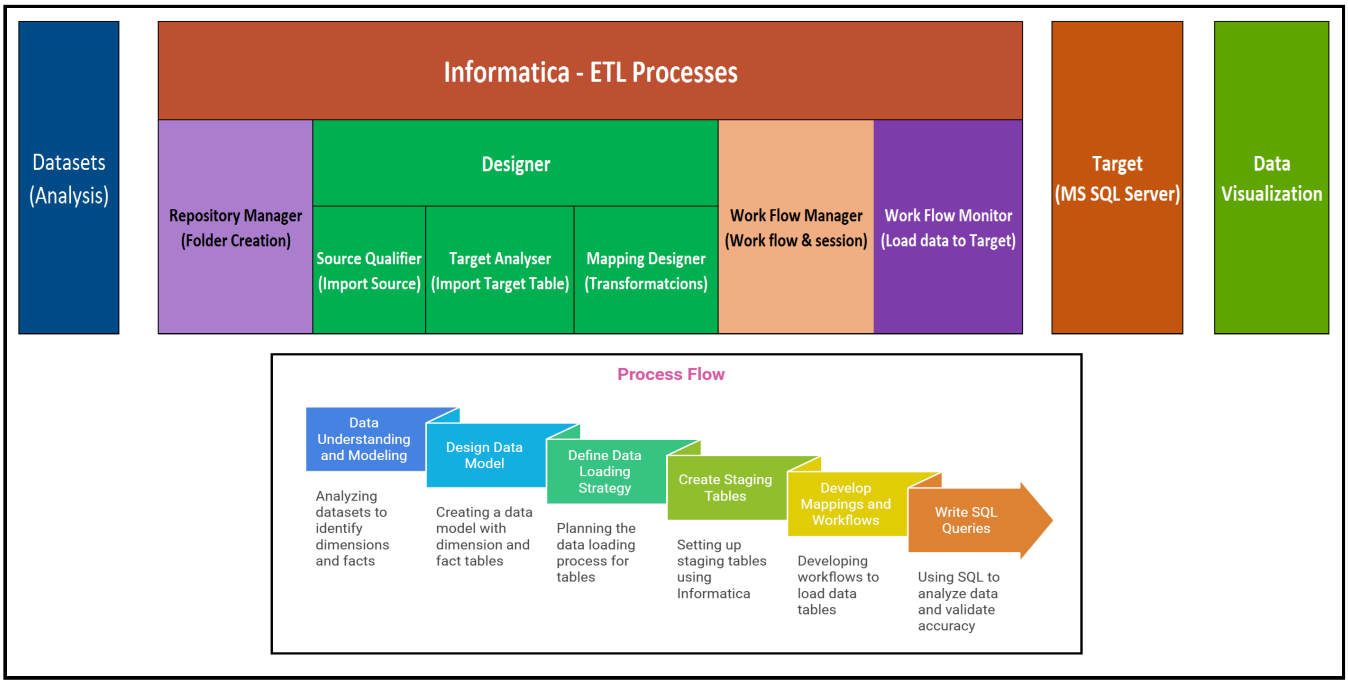


Project Scope & Process flow

. Data Understanding & Modelling (Analyze Datasets, identify dimensions & facts, Design the data model)

. ETL Implementation (Data Loading strategy, Staging Tables, Mappings & workflows in Informatica)

. Data Analysis (SQL querying and Analyzing the Data)



Datasets - Summary

. There are 6 Data Sets (1 main dataset and 5 dimension lookup datasets for mapping)

. Summary is given in the below table

S_No	Dataset	Filename	Records	Columns	Col_1	Col_2	Col_3	Col_4	Col_5	Col_6	Remarks
1	Main Dataset	Data8277.csv	34959672	6	Year	Age	Ethnic	Sex	Area	count	Mapping of data to be done in with Lookup dataset. Datasets are having lookup data in different dimensioning groups.
2	Year reference dataset	DimenLookupYear8277.csv	3	3	Code	Description	Sort Order				Year and Year code - both are same
3	Age reference dataset	DimenLookupAge8277.csv	148	3	Code	Description	Sort Order				Age lookup data having different dimensioned grouped data.
4	Ethnic reference dataset	DimenLookupEthnic8277.csv	11	3	Code	Description	Sort Order				Dimensioning is vailable for value '6"
5	Sex reference dataset	DimenLookupSex8277.csv	3	3	Code	Description	Sort Order				Male, Female, Total Total is Male + Female
6	Area reference dataset	DimenLookupArea8277.csv	2386	3	Code	Description	Sort Order				Area lookup data having different dimensioned grouped data.

Mapping with Lookup

. Mapping the columns in the main data with Lookup data as given in the following table

Dimension Lookup mapping with Main Data									
Type	Dataset	Records	Column-1	Column-2	Column-3	Column-4	Column-5	Column-6	Mapping Column
Main	Data8277.csv	34959672	Year_Code	Age_Code	Ethnic_Code	Sex_Code	Area_Code	count	
Lookup	DimenLookupYear8277.csv	3	Year_Code						Year_Description
Lookup	DimenLookupAge8277.csv	148		Age_Code					Age_Description
Lookup	DimenLookupEthnic8277.csv	11			Ethnic_Code				Ethnic_Description
Lookup	DimenLookupSex8277.csv	3				Sex_Code			Sex_Description
Lookup	DimenLookupArea8277.csv	2386					Area_Code		Area_Description

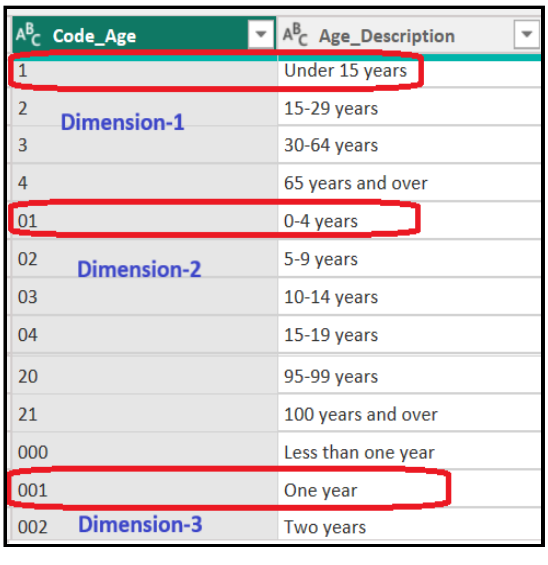
Important points to note

There are few codes which are in **Numerical format** with different sizes. (Ex: Age Lookup dataset)

Actually, these codes are designed for mapping in **different DIMENSIONS** (Ex: Age Lookup dataset)

This leads to discrepancy in the data values, if we consider these values as numericals.

Hence all the codes are to be considered as **"Strings" to treat them as different dimensions in the lookup.**



Datasets - Transformation

. Main Dataset (Data8277.csv)

. The count of Population should be in "Whole Number"

. Non-Numerical values are to be treated carefully in the "Count" Column. (Ex: "..C" values)

1. Year (DimenLookupYear8277.csv)

. Year_Code and Year_Description - both are same.

. No complexity observed

2. Age (DimenLookupAge8277.csv)

. Grouping of values (Dimensions) are present in the AGE Lookup data.

. Proper mapping is to be done with necessary lookup for accurate results.

. The following "AGE" Lookup values are present in different Dimensions:

Dimension-1:

Age_Code	Description
1	Under 15 years
2	15-29 years
3	30-64 years
4	65 years and over

Dimension-2:

Age_Code	Description
01	0-4 years
02	5-9 years
03	10-14 years
04	15-19 years
.	
.	
.	
20	95-99 years
21	100 years and over

Dimension-3:

Age_Code	Description
000	Less than one year
001	One year
002	Two years
003	Three years
.	
.	
.	
119	119 years
120	120 years and over

3. Ethnic (DimenLookupEthnic8277.csv)

. Grouping of values (Dimensions) are present in the ETHNIC Lookup data for value '6' only.

. Proper mapping is to be done with necessary lookup for accurate results.

. The following "ETHNIC" Lookup values are considered:

Ethnic_Code	Description
1	European
2	Maori
3	Pacific Peoples
4	Asian
5	Middle Eastern/Latin American/African
6	Other ethnicity(61-New Zealander,69-Other ethnicity)
9	Not elsewhere included

4. Sex (DimenLookupSex8277.csv)

. The following are the "SEX" Lookup values:

Sex_Code	Description
1	Male
2	Female

5. Area (DimenLookupArea8277.csv)

. Grouping of values (Dimensions) are present in the AREA Lookup data.

. Proper mapping is to be done with necessary lookup for accurate results.

. The following "AREA" Lookup values are present in different Dimensions:

. The Census details are provided for "Region wise", "Territorial Authority" and "District Health Board".

AREA - Categories:

. New Zealand by Regional Council

. New Zealand by Territorial Authority

. New Zealand by District Health Board

Dimension-1:
New Zealand by Regional Council:

Area_Code	Description
01	Northland Region
02	Auckland Region
03	Waikato Region
04	Bay of Plenty Region
05	Gisborne Region
06	Hawke's Bay Region
07	Taranaki Region
08	Manawatu-Wanganui Region
09	Wellington Region
16	Tasman Region
17	Nelson Region
18	Marlborough Region
12	West Coast Region
13	Canterbury Region
14	Otago Region
15	Southland Region

Dimension-2:
New Zealand by Territorial Authority:

```
-----------------------------------
Area_Code       Description
-----------------------------------
001             Far North District
002             Whangarei District
003             Kaipara District
.
.
.
073             Southland District
074             Gore District
075             Invercargill City
-----------------------------------
```

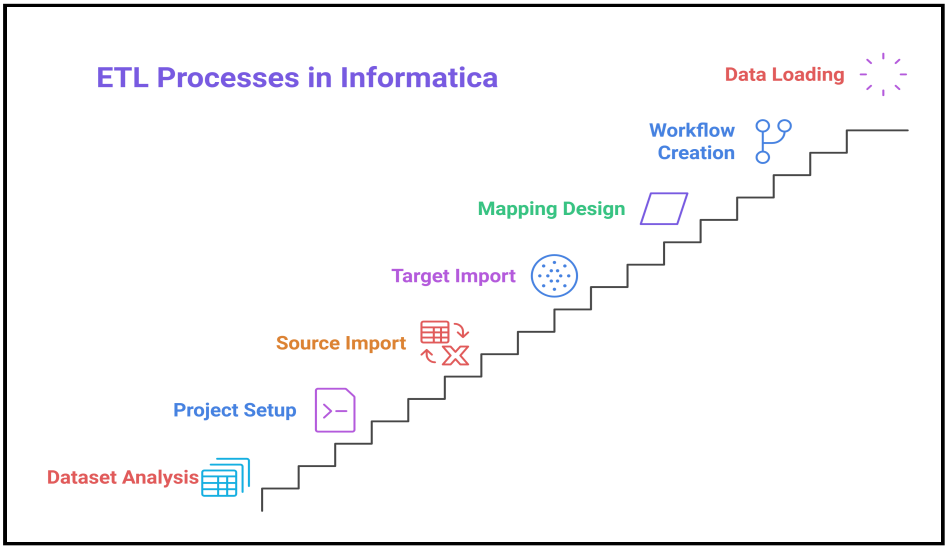
Dimension-3:

New Zealand by District Health Board

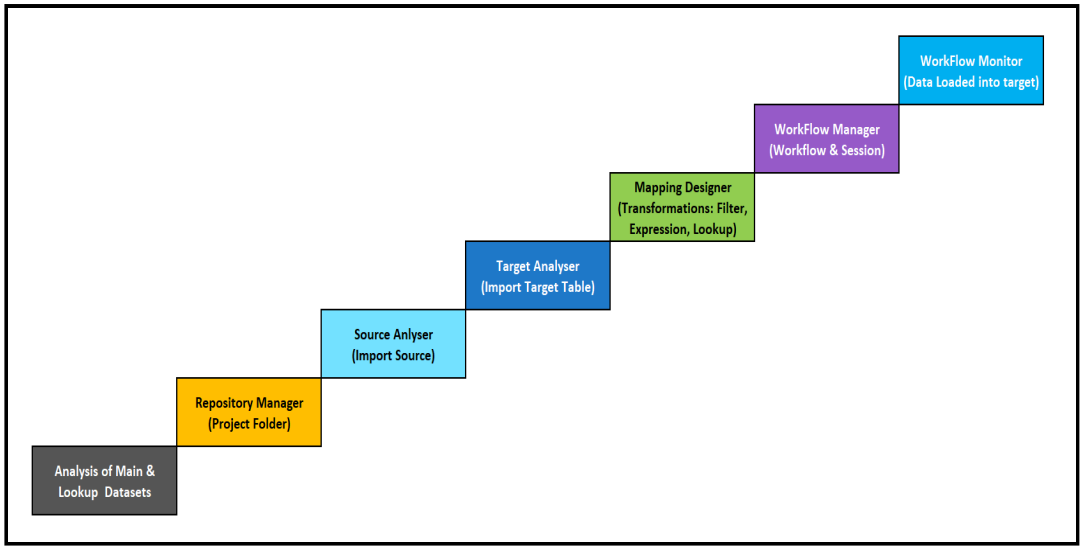
```
-----------------------------------
Area_Code       Description
-----------------------------------
DHB01           Northland
DHB02           Waitemata
DHB03           Auckland
.
.
.
DHB18           Canterbury
DHB19           South Canterbury
DHB22           Southern
-----------------------------------
```

ETL Process in Informatica

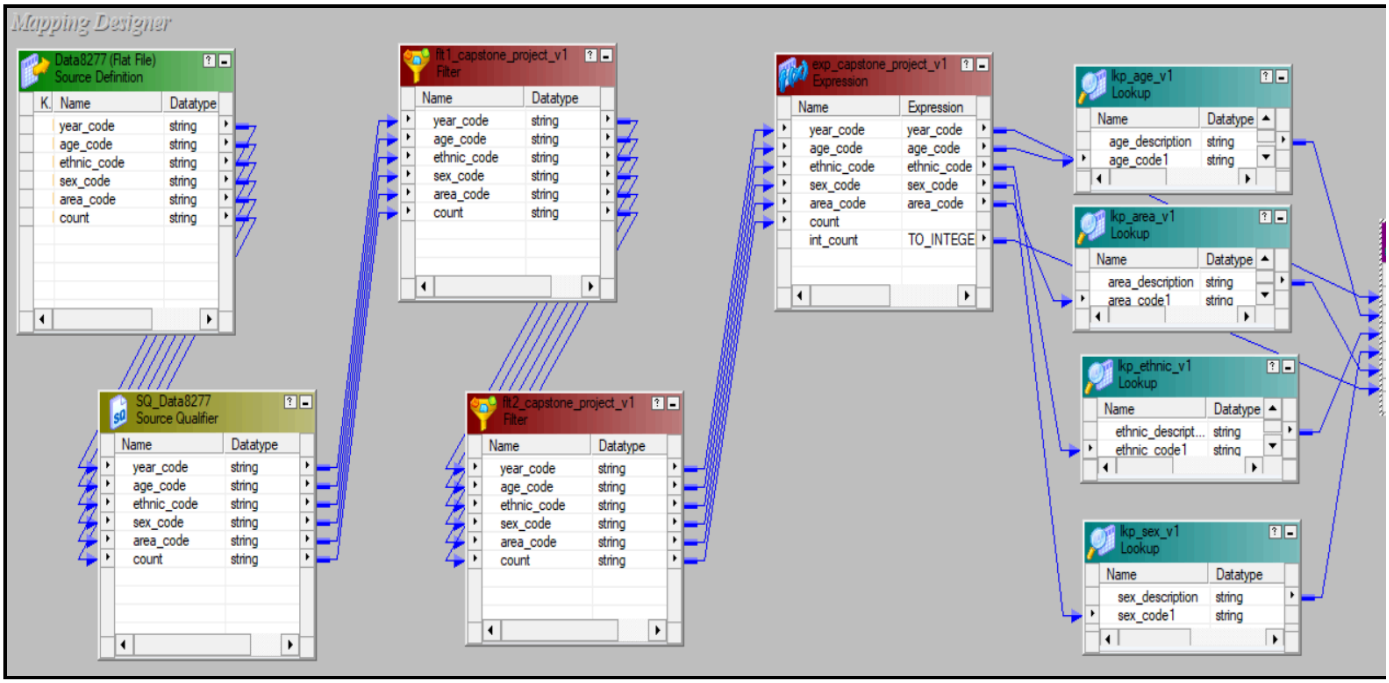
Once Data Analysis is done => find out the key observations => Finalized the Solution => ETL Processes(Informatica)



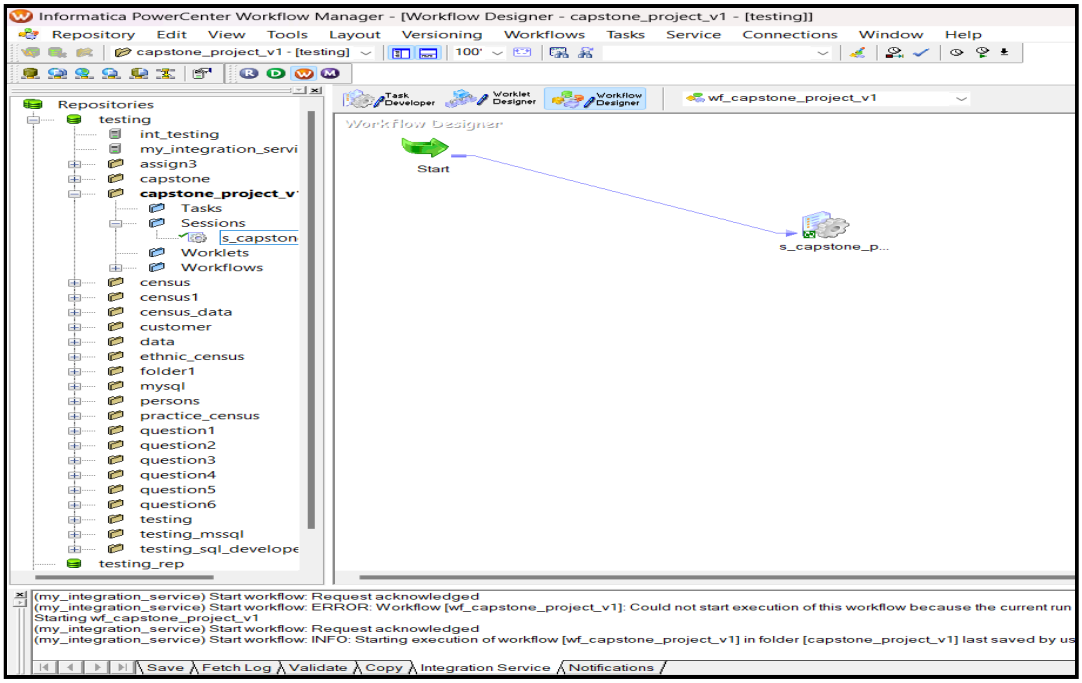
The list of High level Tasks are given in the following diagram:



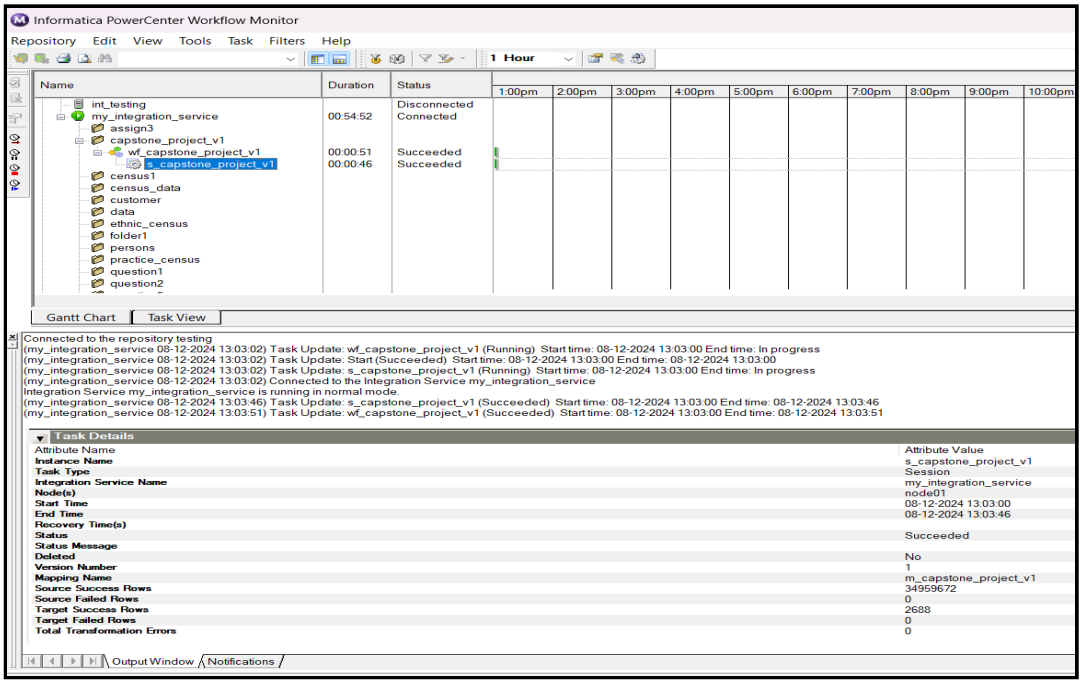
Mapping Designer



Work Flow Manager

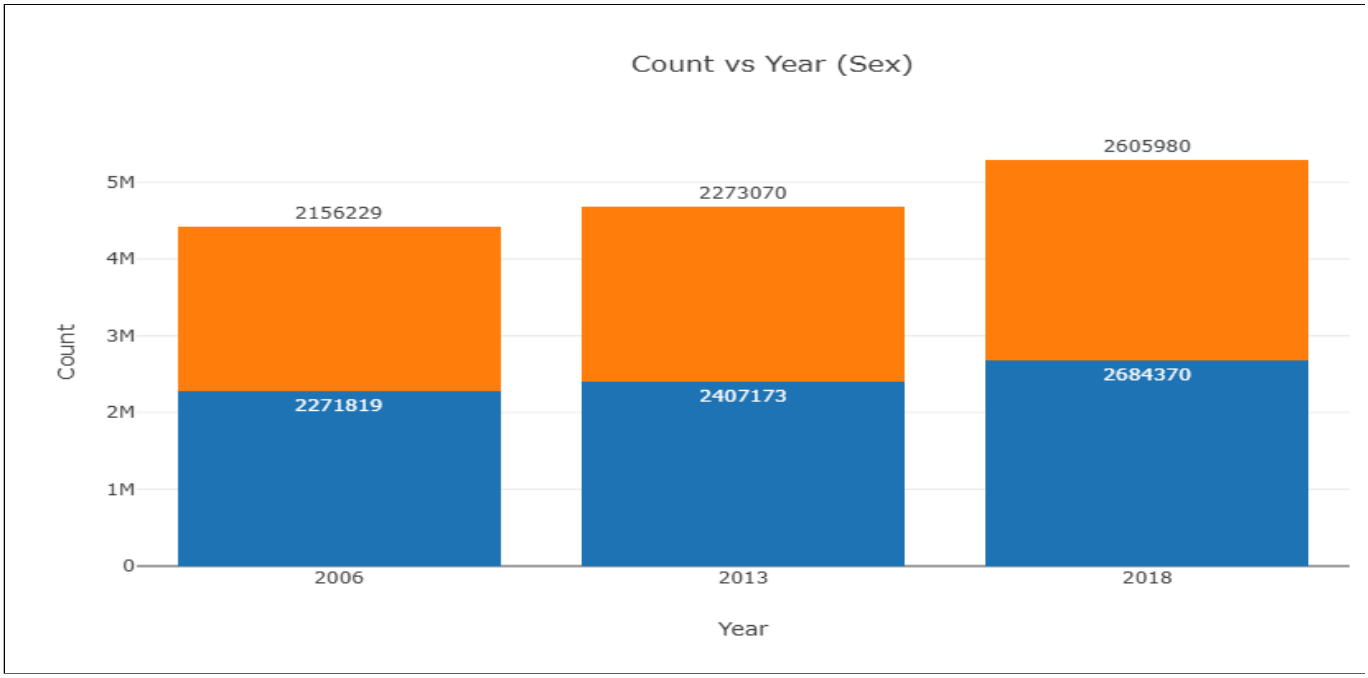


Work Flow Monitor

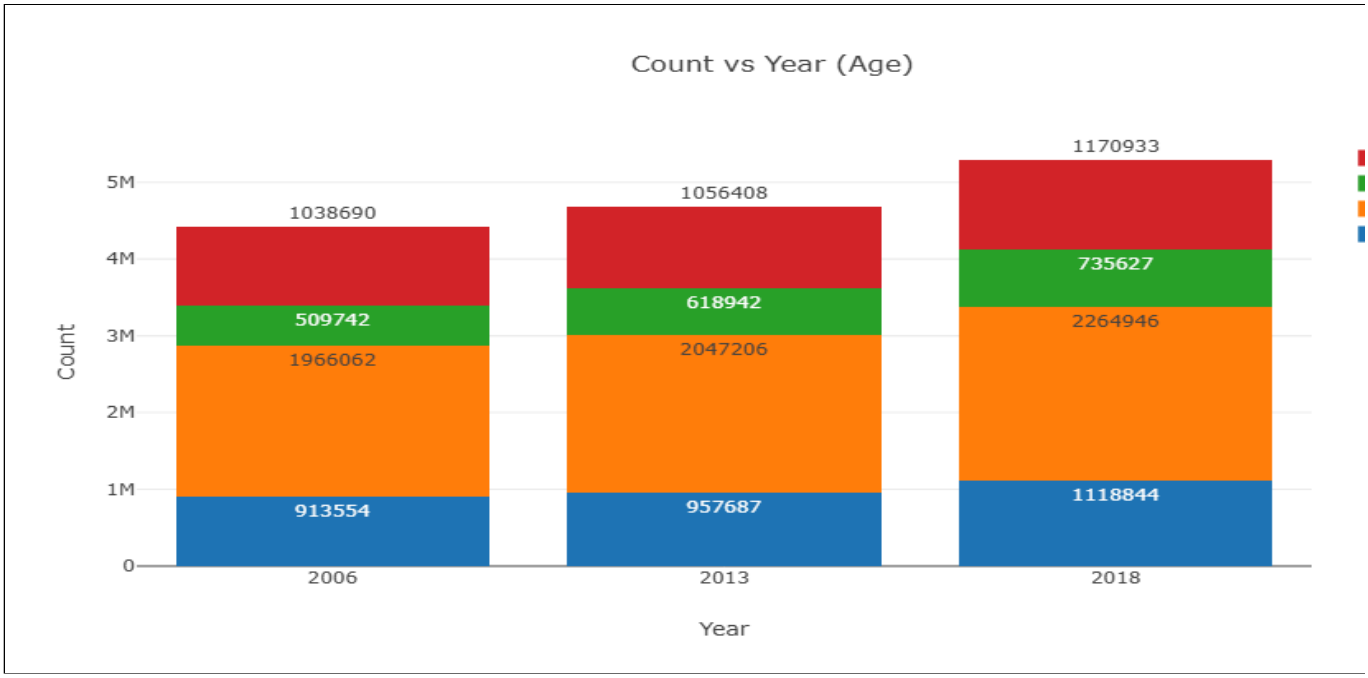


Visual Analytical Reports

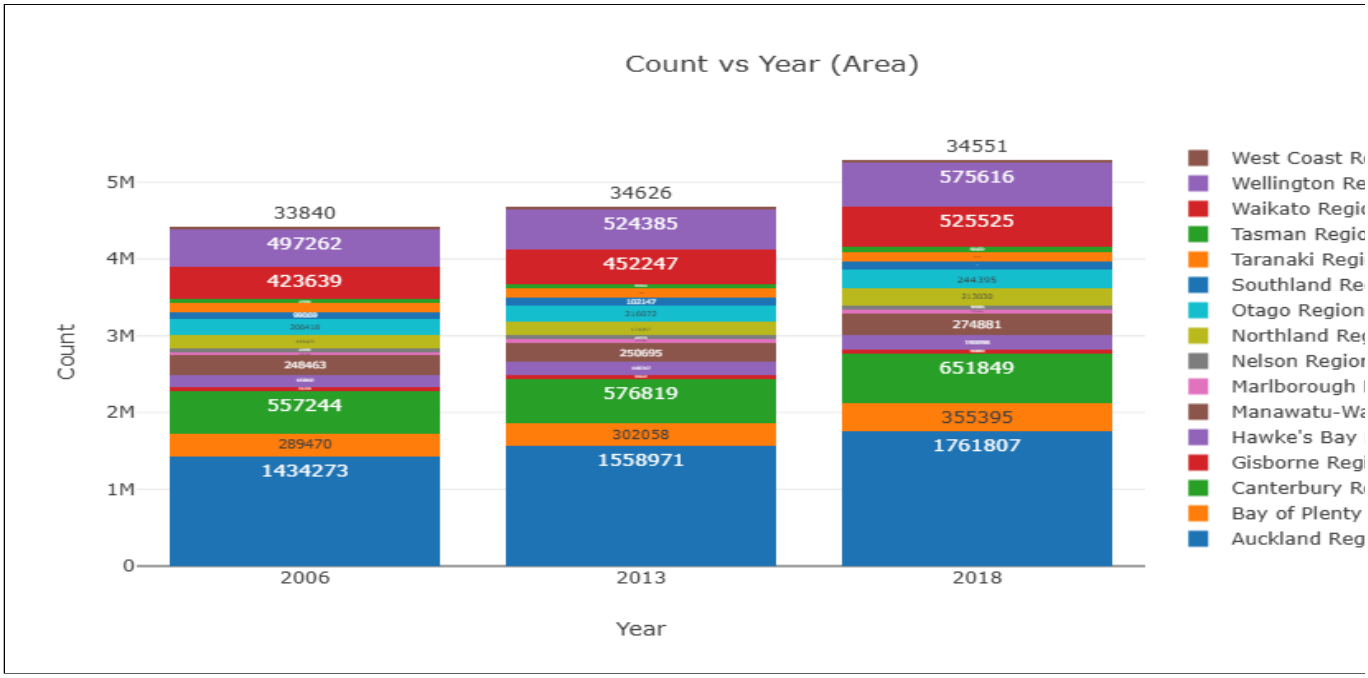
Census - Year Vs Male/Female Population



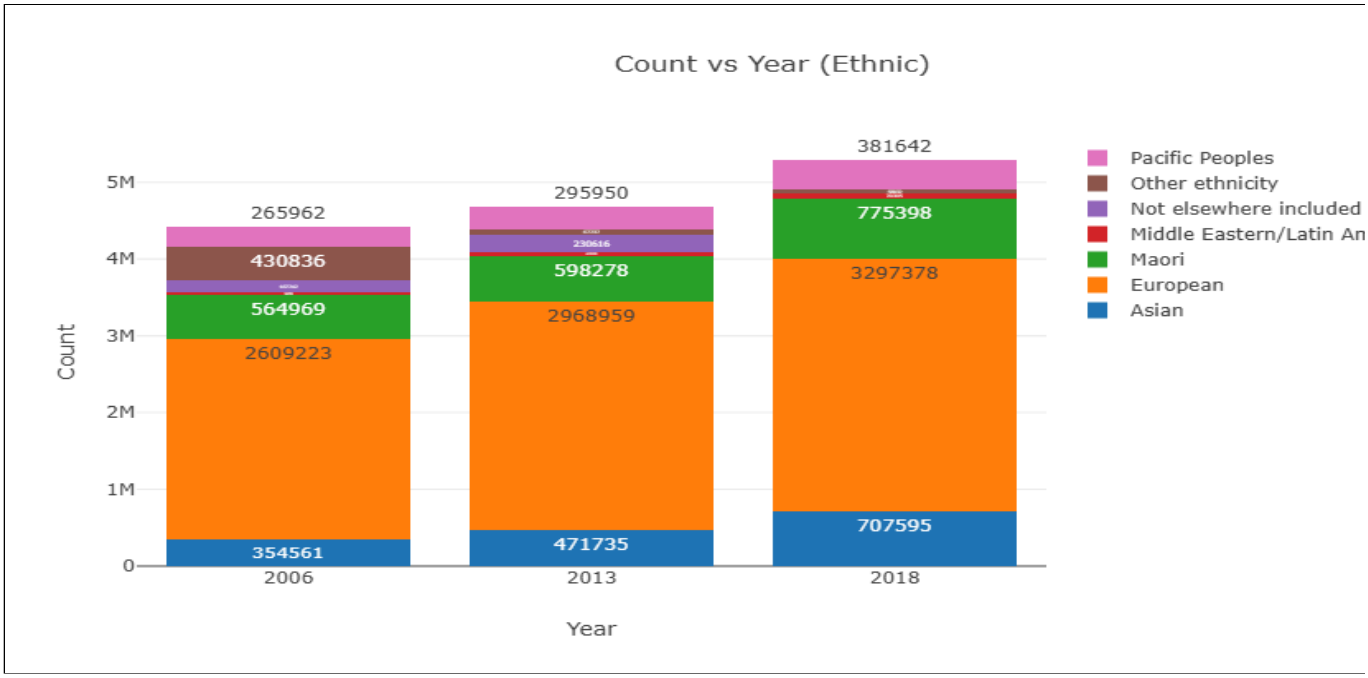
Census - Year Vs AgeGroup



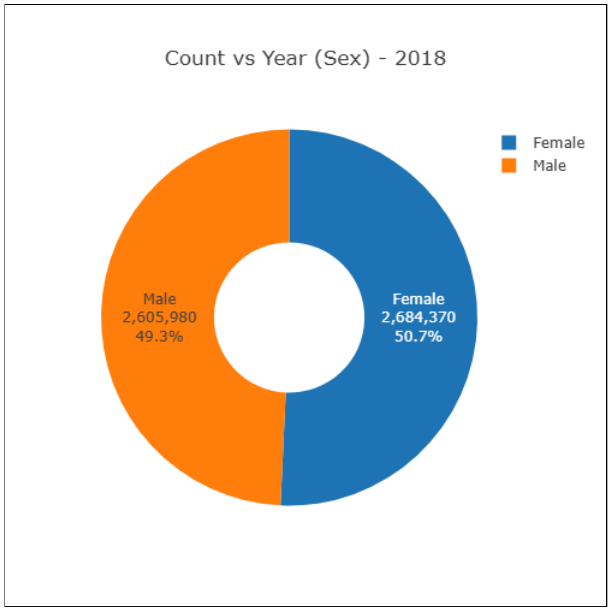
Census - Year Vs Area



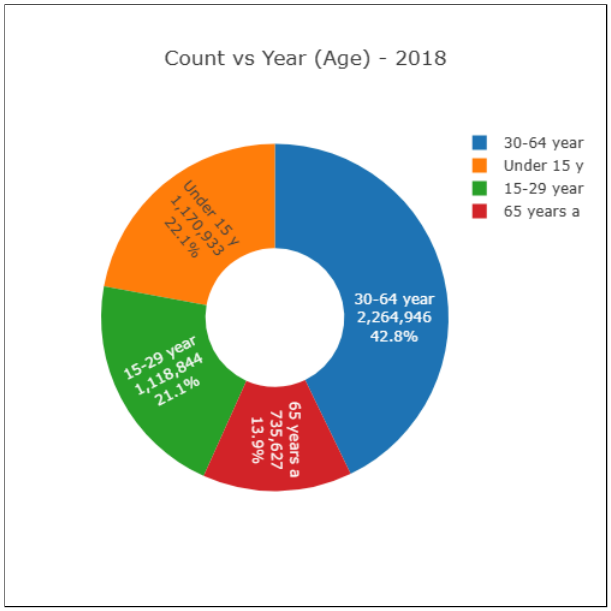
Census - Year Vs Ethnic Groups



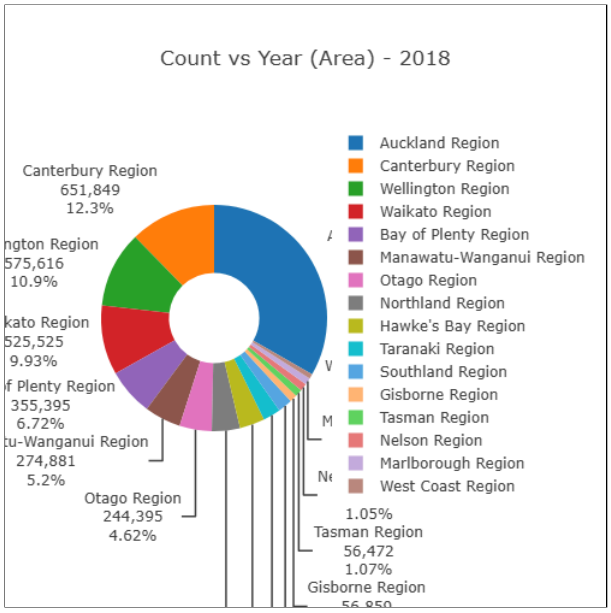
Census - Male/Female Ratio (Donut Pie Chart)



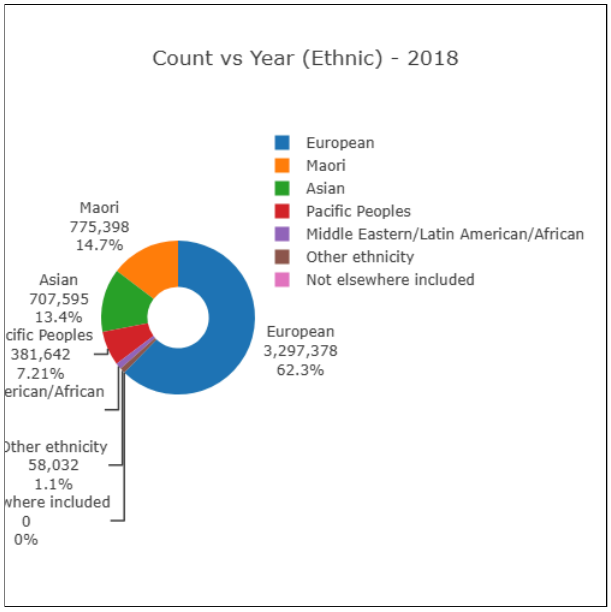
Census - Age Group Ratio (Donut Pie Chart)



Census - Area Wise (Donut Pie Chart)



Census - Ethnic Group Wise (Donut Pie Chart)


Count vs Year (Ethnic) - 2018
European
Maori
Asian
Pacific Peoples
Middle Eastern/Latin American/African
Other ethnicity
Not elsewhere included
Maori
775,398
14.7%
Asian
707,595
13.4%
cific Peoples
381,642
7.21%
erican/African
European
3,297,378
62.3%
)ther ethnicity
58,032
1.1%
where included
0
0%