Data cleansing and pre-processing

The data is downloaded and stored in folders. These aggregated files are then combined into one files for each source category. all single consolidated files are in “CSV” format.

Finding missing values :

In this part data will enter the cleansing phase, the first step is checking the data consistency .i.e. look for missing values or irrelative values. The result was that only one column has missing values of 100% which is the **Gender** Column in both tables of **Omani’s in private sector** and **Non-Omani’s in private sector**. But the column is added in the combined file by simply indicates whether the row has Total Male or Female number and update the **Gender** column with the one who has a value greater than zero based on the following formula :

$$if\ Total\ Male > 0\ then\ Gender = Male$$

$$else\ Gender = Female$$

Renaming columns :

Since there are four data categories for the data, it is found that the column names are not the same among the data sets. The columns where renamed in order to make the linking easier between the data sets based on the following universal naming in **Table 12**:

Universal Column Name	Data sets renamed
SECTION_DESC_ARB	Omani’s in private sector Non-Omani’s in private sector Personal labors
OCCUPATION_DESC_ARB	Omani’s in private sector Non-Omani’s in private sector Personal labors
GOVERNORATE_NAME_AR	Employment by region Omani’s in private sector Non-Omani’s in private sector Personal labors
GOVERNORATE_NAME_EN	Omani’s in private sector

Table 12- renamed columns in data sets

Data translation

One of the issues assigned with the current datasets in use that the data value are in Arabic language. Thus , data were translated by taking the unique values of the targeted columns from the four data categories and grouped into one data frame, then exported into a file and translated every slot in order to create a dictionary to be used for all the records in the dataset. the following Table 13 shows the summary of the translation process :

Data Column	Process of Translation	Description of translation	Number Of Records translated
Governates	Dictionaries	Dictionary was created manually because of small number of records .	11
Section Description	Dictionaries	Dictionary was created manually because of small number of records .	31
Occupation Description	Dictionaries	Dictionary was created from a translated file due to large number of records .	2069

Table 13- Data translation process summary

After creating the data dictionaries, they were applied to the data sets in order to translate all the records by comparing the Arabic columns value to the corresponding English columns. As a result, three new columns ware created in all the data sets as an English translation column.

These new columns will allow the application to show the results in both Arabic and English language. It is possible that the English translation might not be perfect but it's understandable for most educated users.

Skipped columns

After taking a look at the data and trying to understand the use and the questions that needs to be answered, some columns are isolated and ignored because they are considered as not useful to be used in this projects. and they are the following Table 14:

Data Source	Category	Columns
The Sultanate Of Oman Ministry of Manpower	Personal Labor	Unnamed: 0 VILLAGE OCCUPATION LEVEL
	Omani's in private sector	Unnamed: 0 VILLAGE OCCUPATION LEVEL AGE
	Non-Omani's in private sector	VILLAGE OCCUPATION LEVEL ACTIVITY
National Centre of Statics and Information	Population	AGE GROUP UNITS
	Town and Wilayat and Governates	TOWN_ID WILAYAT_ID GOVERNATE_ID

Table 14 - Data sources categories skipped columns

Data consistency

In this section I will try to understand if some data sources are derived from a combination or aggregation of some other sources ,a hypothesis based on the following :

H_0 : *the source* Employment by Region *is dependent and derived from adding the totals of Omanis , Non Omani's and Personal Labor sources.*

H_1 : *the source* Employment By Region *is independent and created based on other sources.*

To test the previous hypothesis, a python code as in **APPENDIX: Code 1** is created to group the data rows by the Governate/Region and sum all the Male and Female totals column in a new column called **Total** and the following table shows the results found for every source.

GOVERNORATE_NAME_EN	TOTAL	SOURCE
AD DAKHLIYAH GOVERNORATE	51353	Omanis in Private Sector
	94161	Non Omanis in Private Sector
	20027	Personal Labor
ADH DHAHIRAH GOVERNORATE	20862	Omanis in Private Sector
	49689	Non Omanis in Private Sector
	10445	Personal Labor
AL BATINAH NORTH GOVERNORATE	79956	Omanis in Private Sector
	223806	Non Omanis in Private Sector
	40856	Personal Labor

GOVERNORATE_NAME_EN	TOTAL	SOURCE
AL BATINAH SOUTH GOVERNORATE	42551	Omanis in Private Sector
	99001	Non Omanis in Private Sector
	18568	Personal Labor
AL BURAYMI GOVERNORATE	10077	Omanis in Private Sector
	46017	Non Omanis in Private Sector
	8532	Personal Labor
AL WUSTA GOVERNORATE	4050	Omanis in Private Sector
	20515	Non Omanis in Private Sector
	3457	Personal Labor
ASH SHARQIYAH NORTH GOVERNORATE	28077	Omanis in Private Sector
	92354	Non Omanis in Private Sector
	14322	Personal Labor

GOVERNORATE_NAME_EN	TOTAL	SOURCE
ASH SHARQIYAH SOUTH GOVERNORATE	28165	Omanis in Private Sector
	90671	Non Omanis in Private Sector
	22617	Personal Labor
DHOFAR GOVERNORATE	31312	Omanis in Private Sector
	189058	Non Omanis in Private Sector
	31430	Personal Labor
MUSANDAM GOVERNORATE	2418	Omanis in Private Sector
	11685	Non Omanis in Private Sector
	1669	Personal Labor
MUSCAT GOVERNORATE	111717	Omanis in Private Sector
	734103	Non Omanis in Private Sector
	84399	Personal Labor

Table 15 - Total Labors by Governate / Region for three sources

And from the above table we create another table to compare the total of the three (Total of Three) data sources with the “Employment by Region“ source (Total of EBR) as in the following :

GOVERNORATE_NAME_EN	Total of Three	Total Of EBR	Difference %
AD DAKHLIYAH GOVERNORATE	165,541	12,984	1275%
ADH DHAHIRAH GOVERNORATE	80,996	5,570	1454%
AL BATINAH NORTH GOVERNORATE	344,618	42,920	803%
AL BATINAH SOUTH GOVERNORATE	160,120	18,822	851%
AL BURAYMI GOVERNORATE	64,626	6,638	974%
AL WUSTA GOVERNORATE	28,022	4,051	692%
ASH SHARQIYAH NORTH GOVERNORATE	134,753	7,207	1870%
ASH SHARQIYAH SOUTH GOVERNORATE	141,453	9,641	1467%
DHOFAR GOVERNORATE	251,800	29,026	867%
MUSANDAM GOVERNORATE	15,772	1,188	1328%
MUSCAT GOVERNORATE	930,219	453,648	205%

Table 16 - Comparison between the totals of three sources and the "Employment by Region " Source

From the above **Table 16** results we can see the difference between the total is significantly big – it ranges from 1870 % to 205% - which indicates that we reject our null hypothesis and conclude that the **Employment by region** source in calculated based on another factors. One of the possible reasons for the inconsistency is that the source covers only the time period between the years 2010 – and 2019.

The application

This part of the report is concerned about the design and the implementation of the application dedicated to visualize the data in the previously mentioned data sources.

Technologies used :

This application will be build using python 3.7 [16] scripts for both the user interface and the analysis functionalities. The integrated development environment (IDE) used is anaconda 3 [17] which is a famous open-source IDE for python development . In addition, some open-source python libraries for data analysis and visualization are used such as : HoloView and Bokeh [18] [19].

The main application prototype is developed using Tkinter which is a standard python interface development toolkit , it provides basic user interface tools and widgets such as : list boxes , buttons ,check boxes .. etc[20].

The python code is written using two famous IDE's which are : JupyterLab and Spyder 3, they come as part of Anaconda IDE. The script is written using one important practice which is the re-use of functions. Therefore, more than 30 functions are written to utilize the resources and enhance the performance of the application in general.

Importing and transforming the data :

Data is imported by executing a piece of code that will mainly read the combined data files and assign them into data frames using pythons Pandas packages. These data frames will hold data from four data sources by executing the function ReadDataSources( ) in Code 2.

The imported data frames will be passed into the function RenameColumns ( ) as in Code 3 to rename all the columns to a universal column naming format to make it easier to be manipulated and accessed later on in the application.

After the renaming process, some data need to be modified to maintain consistency among all sources , it is found that some columns has empty spaces and some of them has extra letters within then. For example, the Governate_Name_Arabic Column. This problems are solved by executing the function ModifyData( ) as in Code 4.

After the columns renaming process the data frames will be passed into the Translation( ) function as in Code 5, to translate some targeted columns based on pre-defined dictionaries - as mentioned previously in the report.

The translation dictionaries are created by aggregating the targeted columns unique values and then exporting them into external files in order to translate the data using google translation online tools [21] and saving the results into the exported file to be later used as a local dictionary in the application. This process is Done by executing the AggUniqueDataValues( ) function in Code 6.

At the end of the importing and transformation process a single data frame is created to hold data from the three sources of Manpower data ; the Omanis in private sector, Non-Omanis in private sector and the Personal labor. A new column is introduced as a flag column to determine if the records are for Omanis or Non-Omanis employees. This process is done by executing the function GroupAllManPower( ) as in Code 7.

The Design :

This part of the report will explore the product (application) prototype developed to visualize the data.

The product consists of two main tab windows , the first one to show the preview of user queries , the queries are built up by the user from the second tab which contains the filters and options to represent the data in the desired way chosen.

The application shows filtration and results in two languages ; English and Arabic and the user can choose the preference from a language tab in the filter window.

The Preview window

This part of the application mainly works as a preview window to show user queries execution results as the following Figures 7 -14.

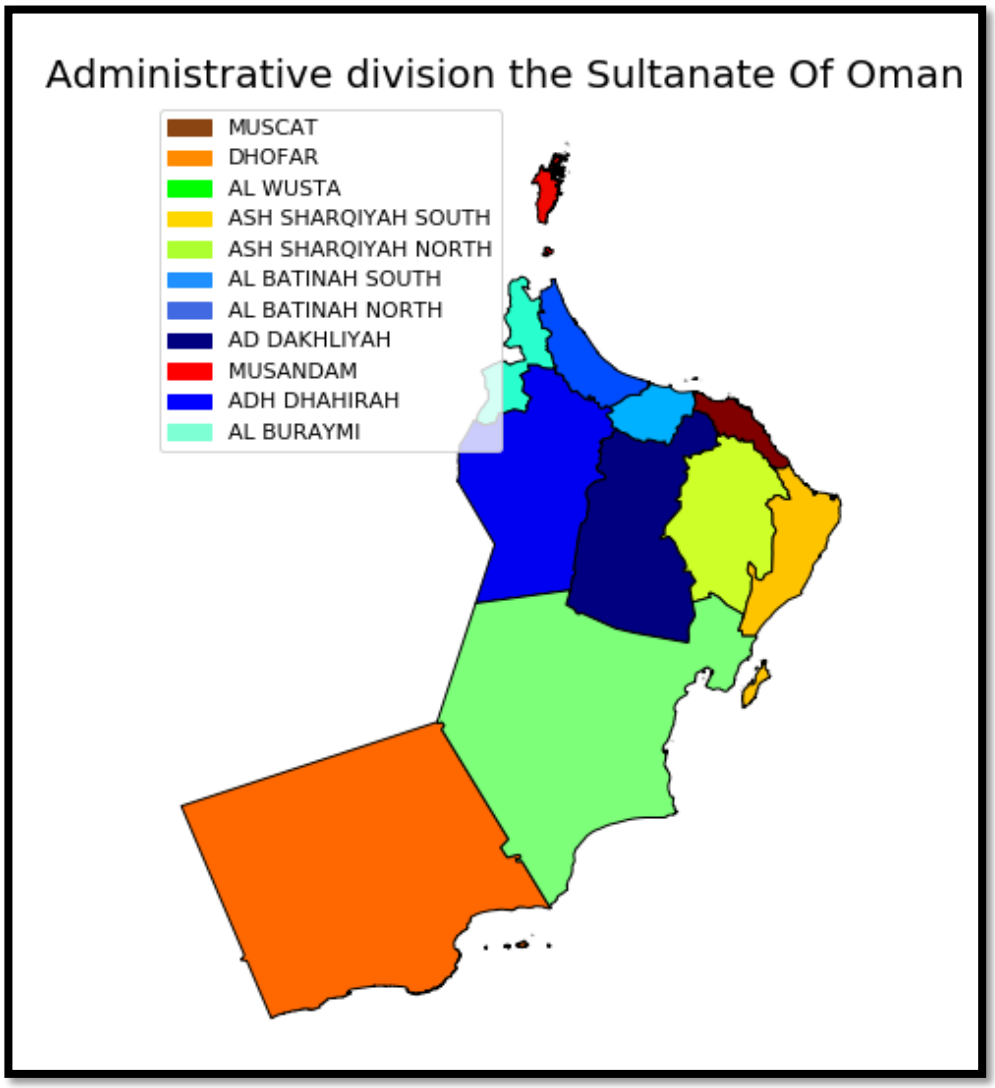

Figure 7-Adminstrative Division of the sultanate of Oman by Governates

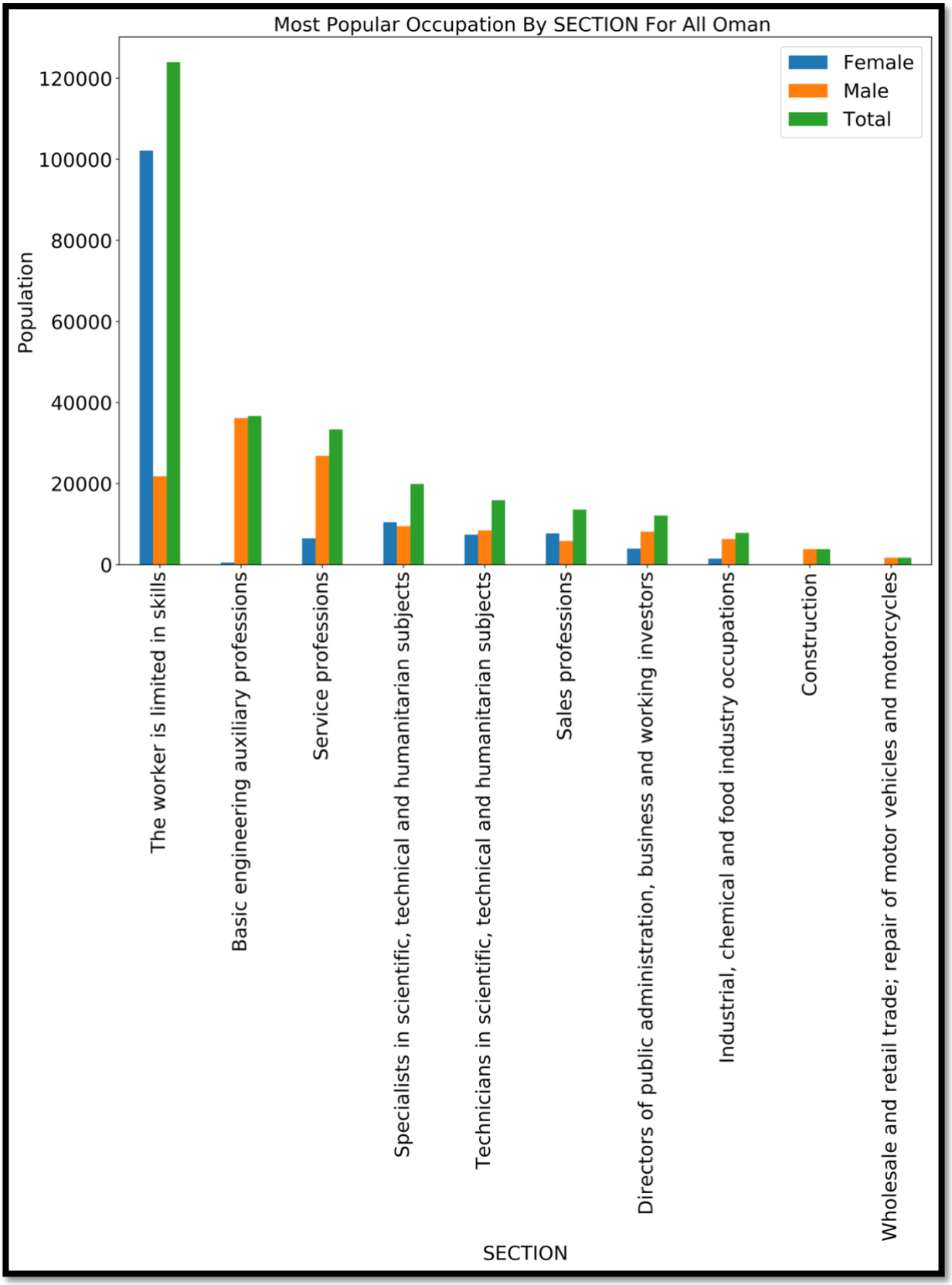

Figure 8 - Most Popular occupation by section for all Oman bar chart

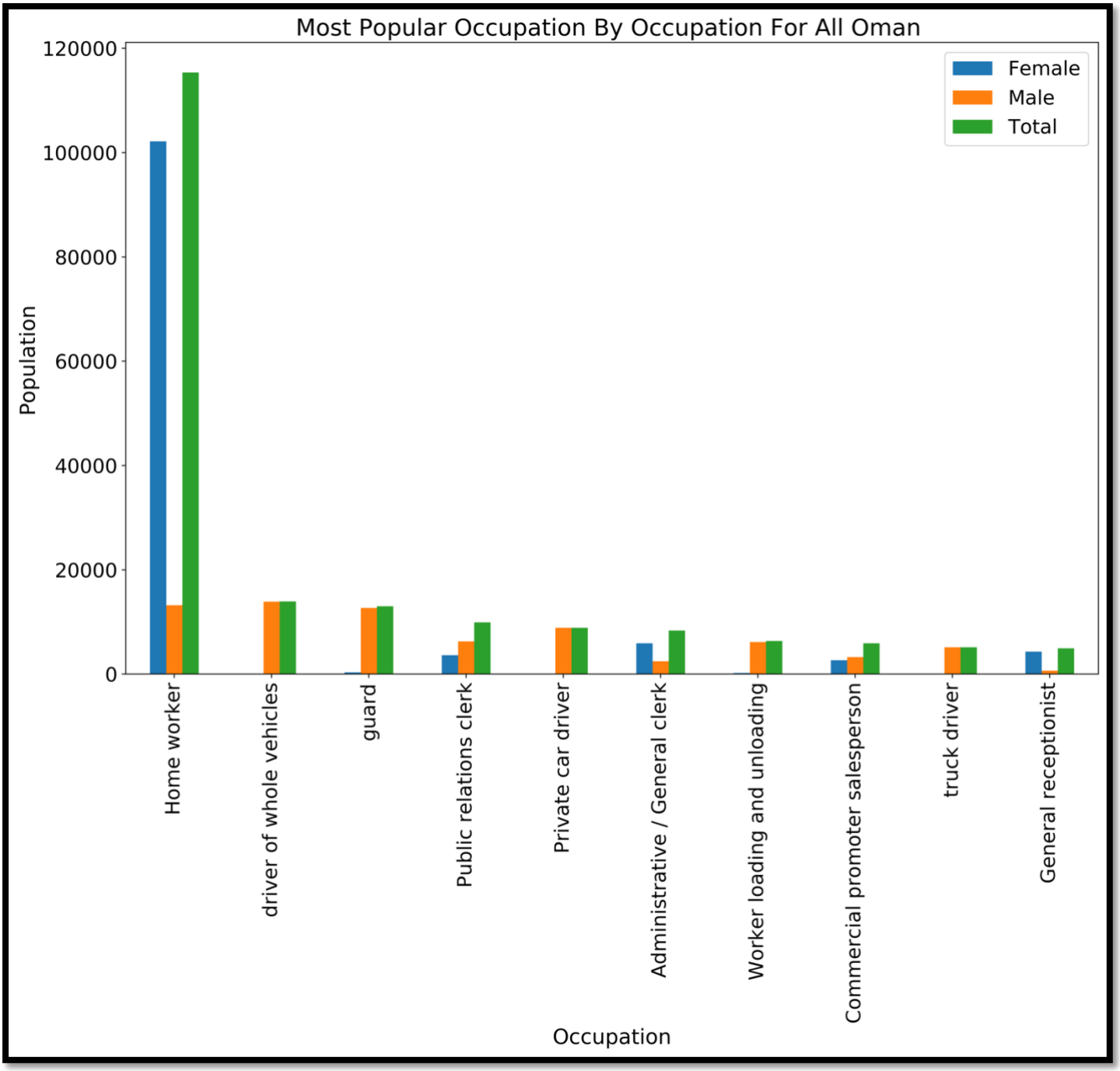

Figure 9- Most Popular occupation by Occupation for all Oman bar chart

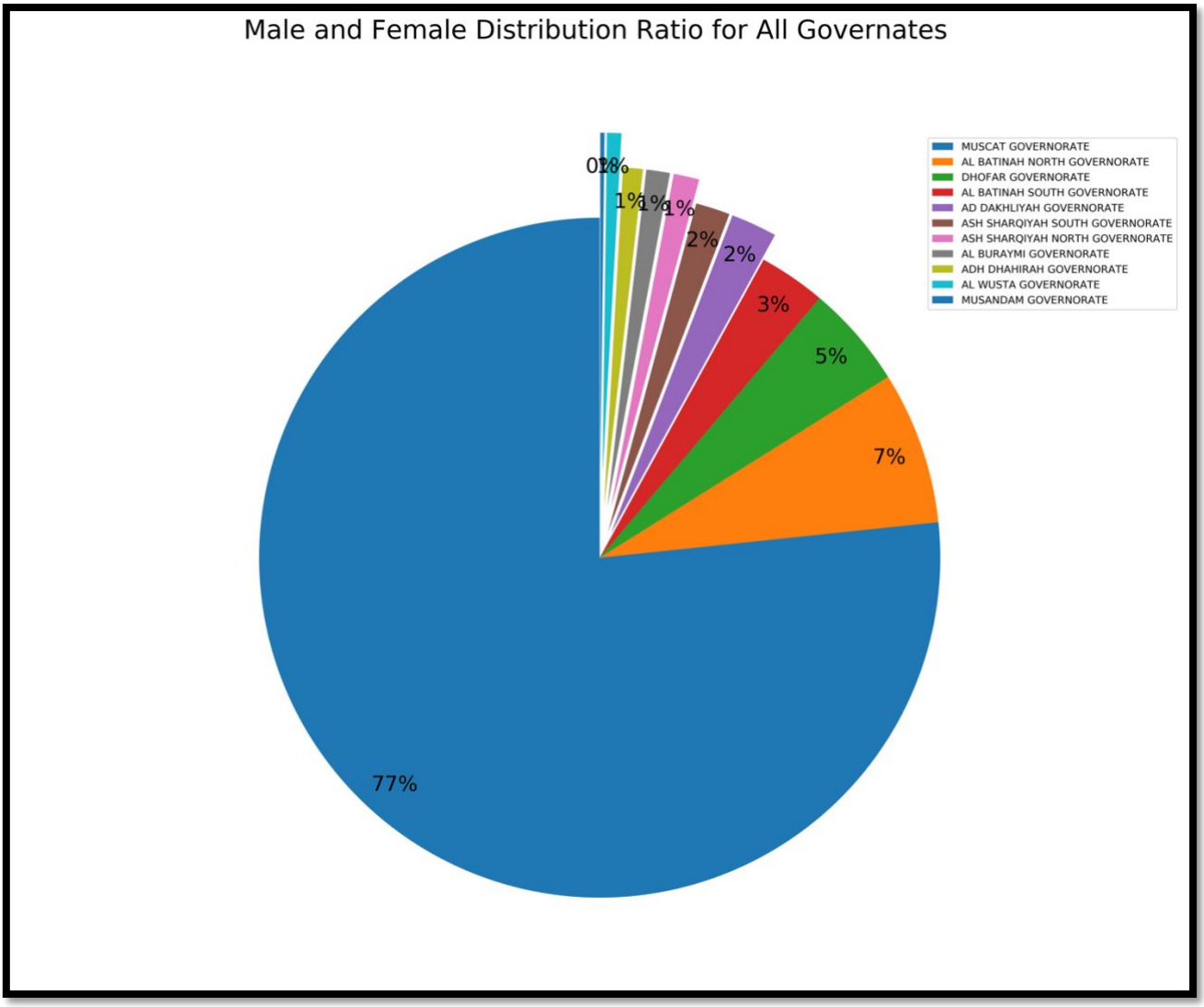

Figure 10 - Females and Males Ratio for all Governates

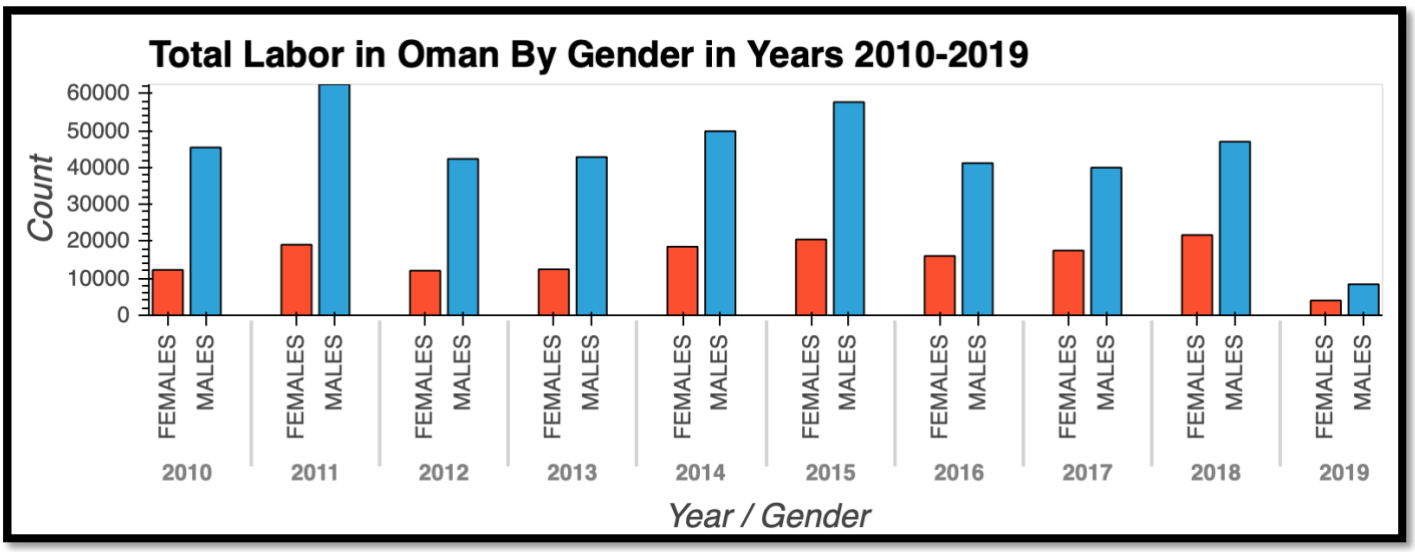

Figure 11 - Gender Distribution for all Oman by Year

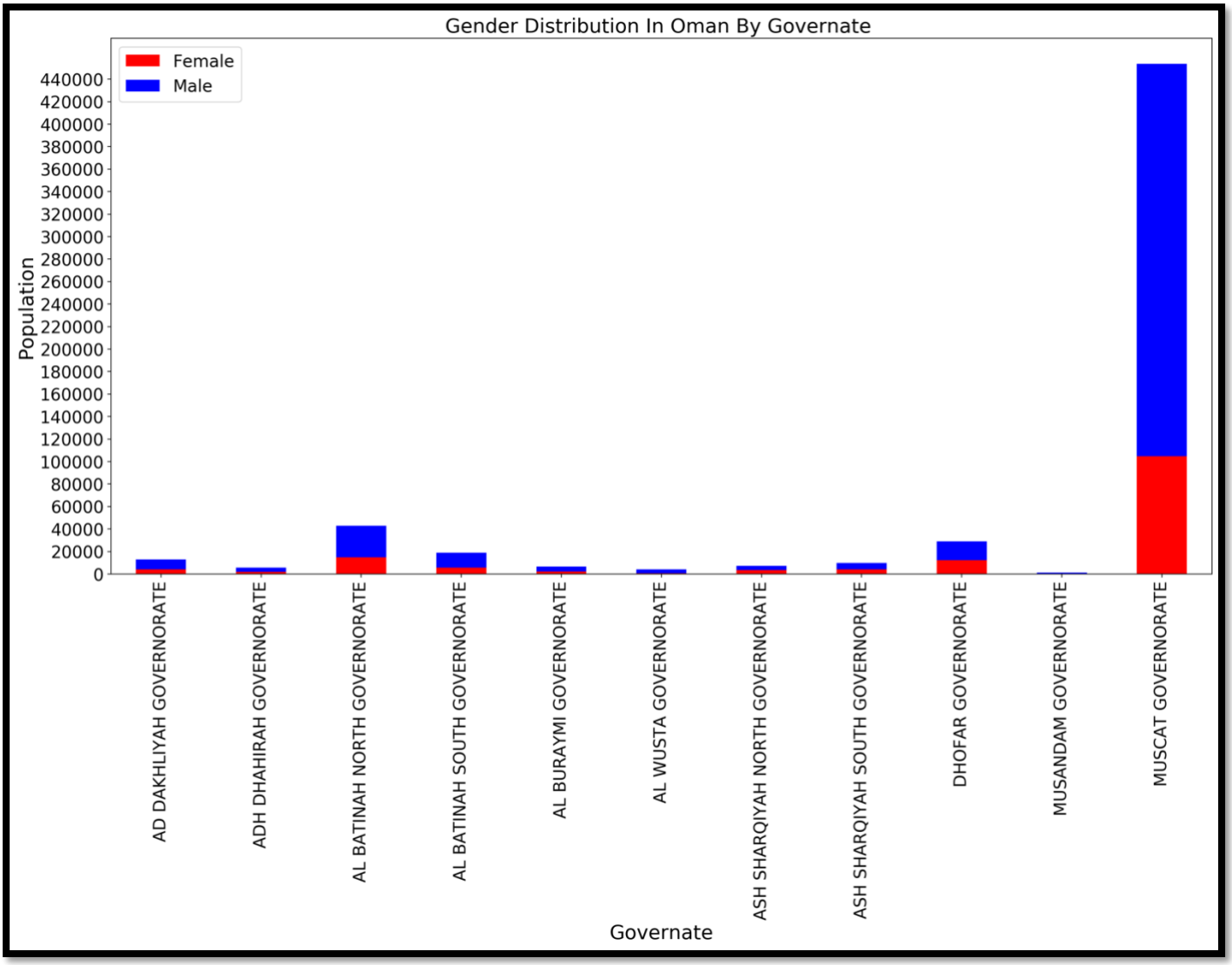

Figure 12- Gender Distribution By Governate

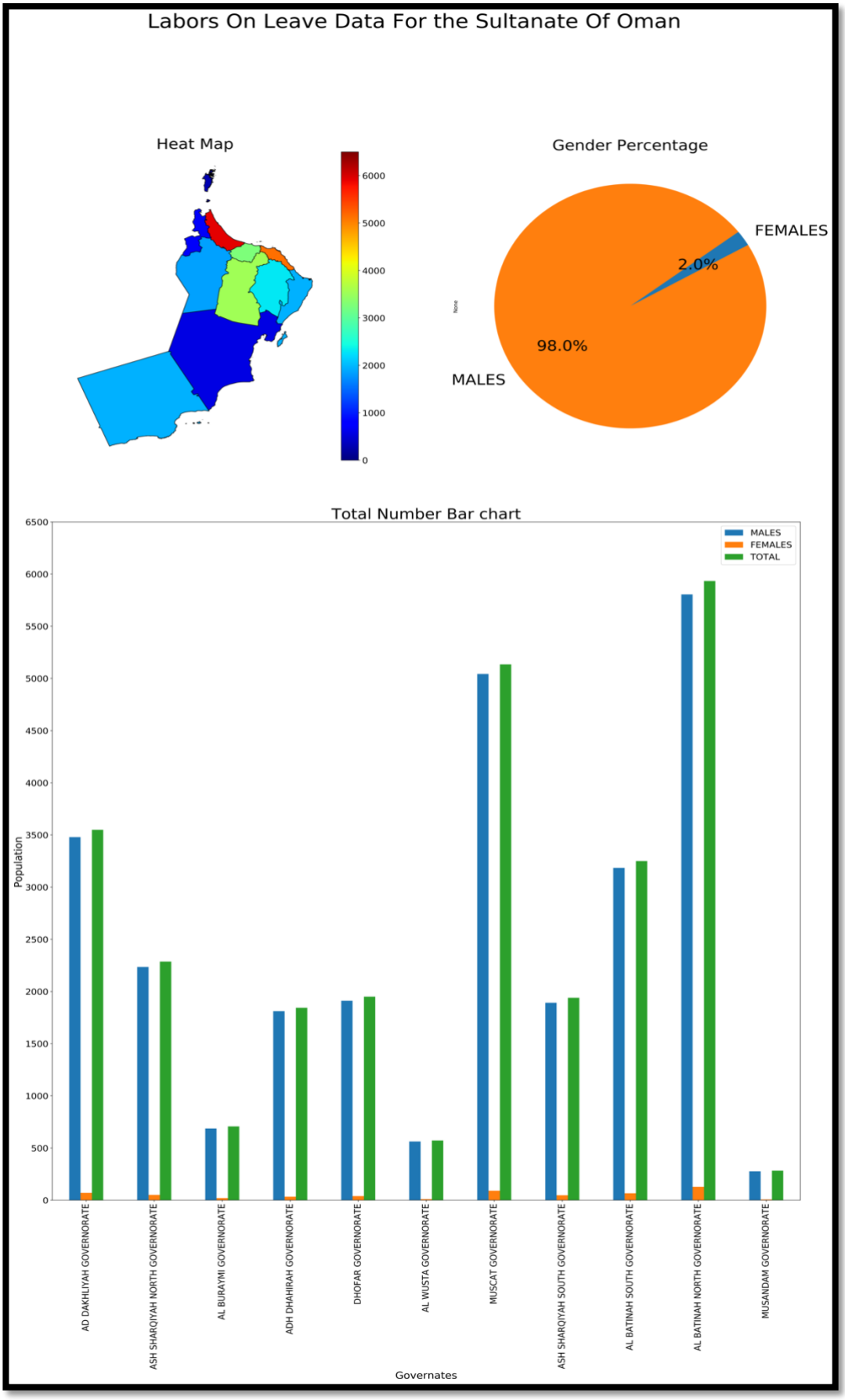

Figure 13- Labor on Leave data for all governates

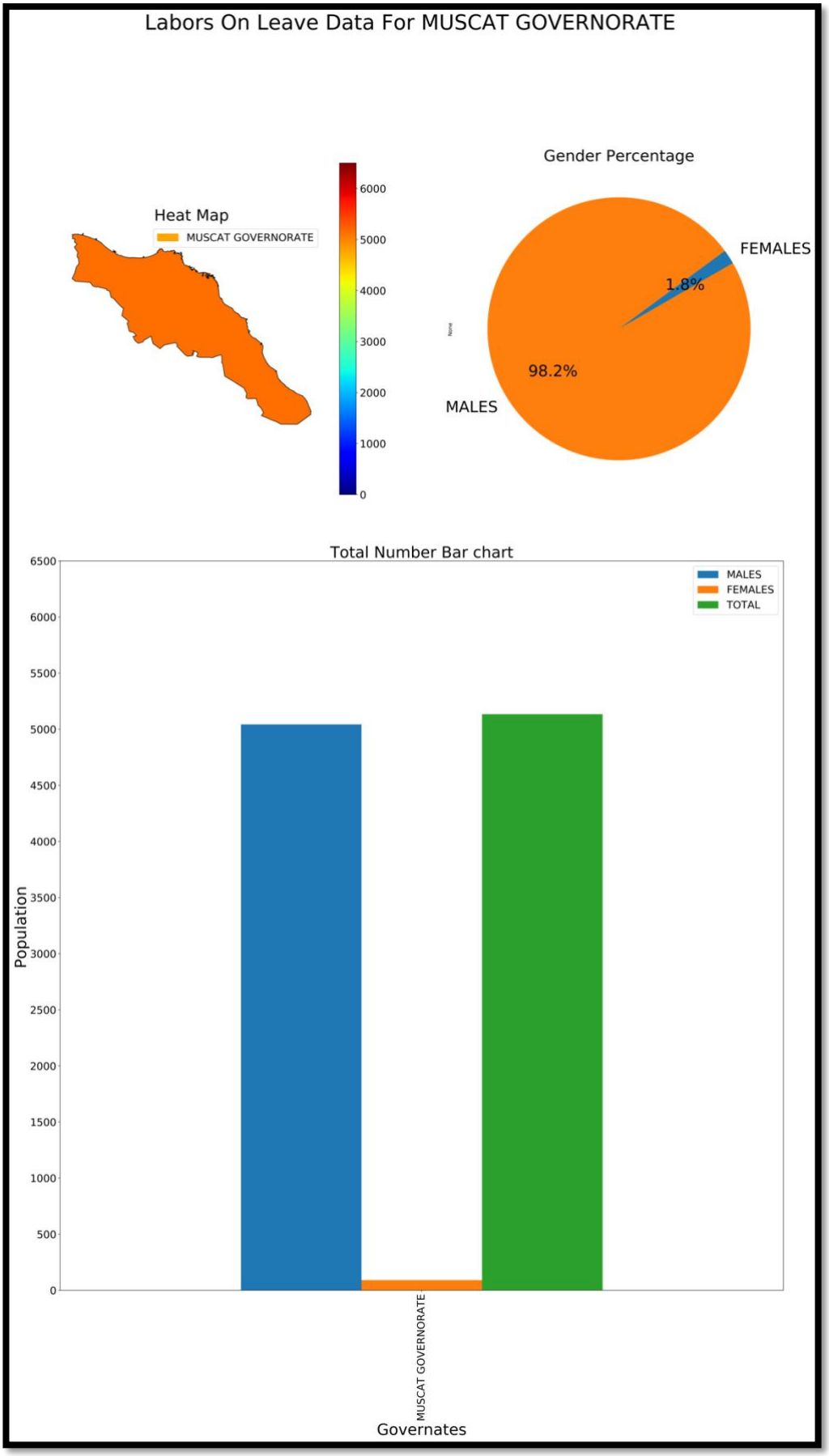

Figure 14 - Labor on Leave for specific governate

The previous figures represent some sample output representation generated by user customized queries. Some visualization are general to all governates of the Sultanate of Oman and some are specific to certain governate and analysis option. Most of these representations are filtered by gender.

The examples of general visualization are the Figures 7 – 13, where the Figure 7 is a map visualization to point the locations of the governates of Oman using “geopandas” python packages to draw the map.

Figures 8-9 represent the top 10 most popular occupations filtered by job section or occupation title based on the females, males and total count. On the other hand, Figures 10-12 are concerned with the gender distribution of work force , starting with the ratio pie chart , the distribution by year bar chart and the distribution by governate.

The Figures 13-14 are a dashboard like representation of the on-leave labors , they are made of a heat map to show the volume of labor in a certain governate , a pie chart to specify the gender ratios and a general total and gender count bar chart to represent a view of total number of labors in all governates.

The Filter windows

This window has multiple widgets that creates the user queries to show the data. These widgets are the following :

Widget name	Widget type	The purpose	Number of choices	Always showing
Language list	Drop down list	To define which language preference to be user .	One	yes
Governates	List box	To list the governates names with a highlight on the chosen values.	multiple	yes
Analysis options	Drop down list	To choose which analysis functionality to execute .	one	yes
Gender	Radio button	The gender option to be chosen by user with three values available : Males , Females or Both.	one	No
Occupation	Radio button	To choose to represent occupation values by two option ; Section or Occupation	one	no
Get selection	Button	To execute user query after the desired filters and chosen.	-	yes

Table 17- Filter window widgets description

As shown in Table 17 some widgets appears only by choosing certain options from the analysis options list , For example the gender options will appear to user for all analysis options except the administrative division of the Sultanate of Oman option.

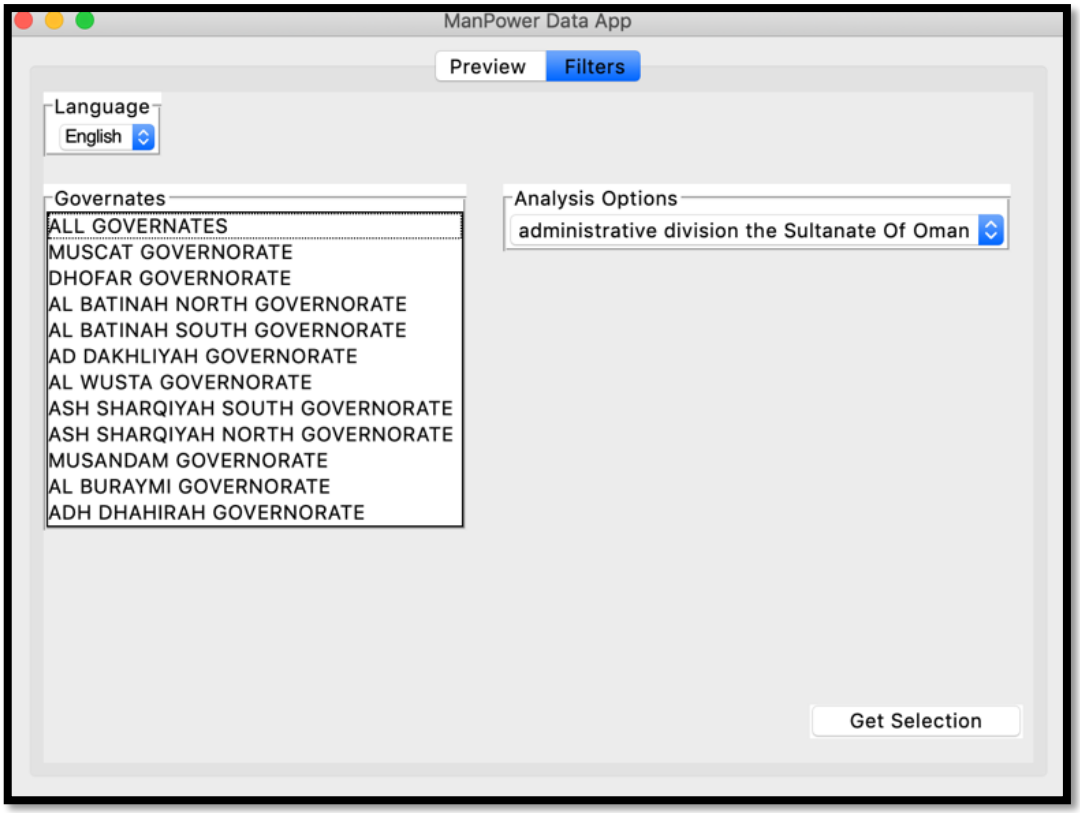

Figure 15- Filter Window Main Window in English

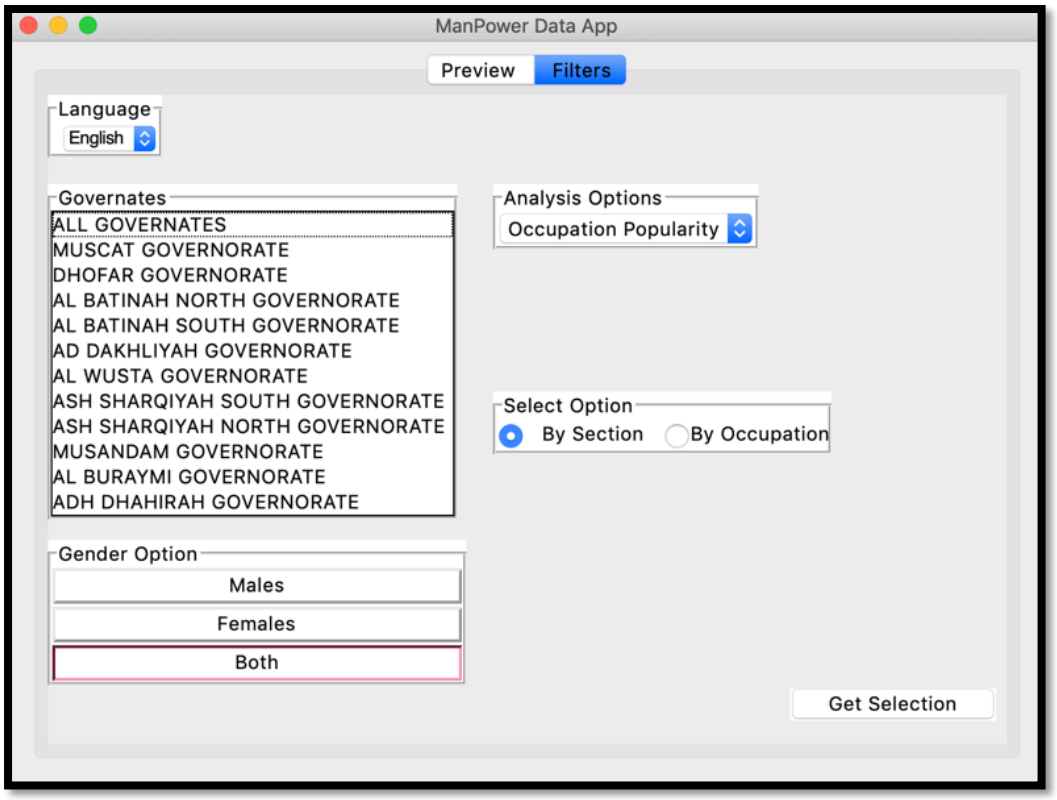

Figure 16 - Filter Window Gender and Occupation Filter options in English

Figure 17 - Filter Window Main Window in Arabic

Figure 18- Filter Window Gender and Occupation Filter options in Arabic

An essential component of the application is the Analysis options list which actually defines the main functionality to be executed by the application .i.e. The actual question to be answered by data. The list contains four optional analysis to be presented as the following :

1. Administrative division of the Sultanate Of Oman : will execute a plot of the governates of the Sultanate of Oman.
2. Gender distribution : will show a dashboard of the gender distribution for all workforce either by all country or specific governates. user can specify also to show results for a specific gender or both.
3. Occupation Popularity : will show the top 10 popular occupations filtered either by job Section or Occupation title .
4. On Leave Labor : will create a map view dashboard of the most Occupations or Governates with highest number of workforce on leave (ran away before completing their contract period).

After choosing the specific analysis option, the application will execute the user query and show the results in a semi dashboard like visualization as shown in the Figures above.

The code for the application product is provided in Appendix Code 8.