

## Why this Project?

When we think of the United States we think of one of the top and most developed countries in the world. We think of its culture and its greatness. Its beauty and its precious land. However, this country has several imperfections; its citizens are not very well educated and are not aware about this beauty and wonderful potential, how diverse it is, and how rich it is in culture. An article made by the New York Post in 2019 presented this unfortunate problem. It stated: “A new survey found that Americans have an abysmal knowledge of the nation’s history... Most disturbingly, the results show that only 27 percent of those under the age of 45 across the country demonstrate a basic knowledge of American history”. It presents the idea that to a great extent ignorance is present throughout the United States and a great amount of their citizens are illiterate in this topic. We think that to know the history of our country we must first know who are the people that compose it and its demographics. We consider that there is a large sector of the population that does not have access to this information or simply see the way to search the Census tables as a tedious and boring task. The purpose of our project is to educate and invite them to learn more about their country in a fun and interactive way. Knowing this valuable information will give them a clearer vision of the problems that occur in the United States and the circumstances in which they find themselves. Citizens will have the opportunity to understand that they live in a highly diverse country and that there are a great number of people of every social class, which will probably reduce racism in the country. If citizens know the country’s poverty rate, they will be more grateful for what they have and motivate them to help the less fortunate. With this program the citizens will understand the importance of education and the true value of their country.

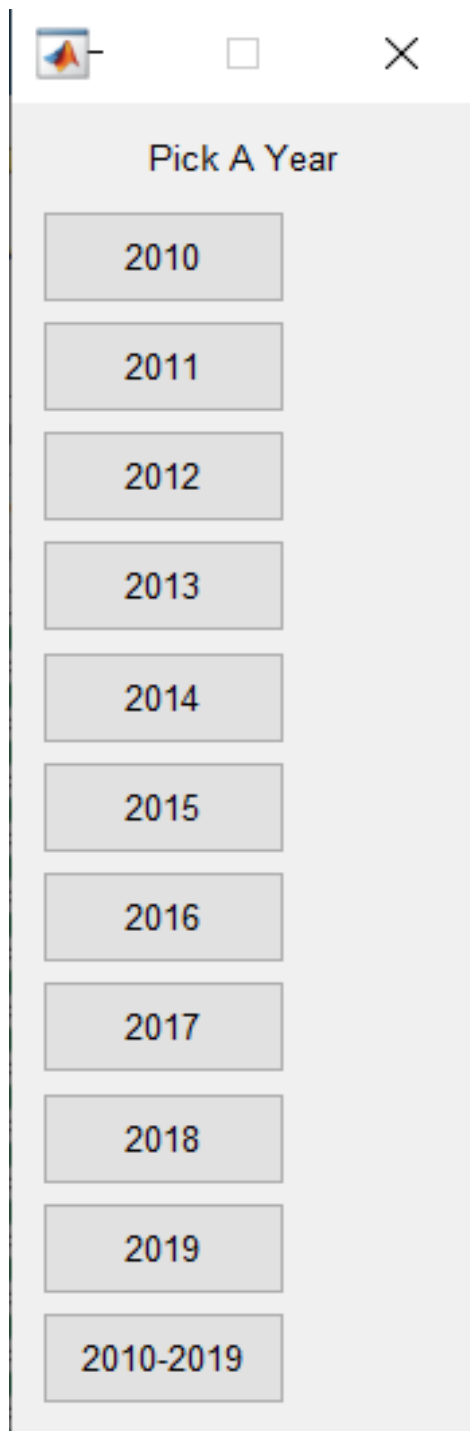
## How did we tackle the problem?

First we used and analyzed the different tables that are provided by the census. Among the tables are the following: ('People and Population', 'Families and Living arrangements', 'Health', 'Education', 'Business and Economy', 'Employment', 'Housing', 'Income and Poverty', 'Veterans'). We obtained the information of each data table from 2010 to 2019 in order to carry out our project. Then we created a code so that any kind of person has access to this valuable information. When we run the program it looks like this:



In the beginning of our program we created a menu with the following options. These options are the different information that is provided by the United States Census. We consider that the knowledge provided in each of these branches is crucial for a more educated society and a more complete system. Having knowledge of this data will make them have an open mind about the decisions that are made in the country. For example from this menu they can learn about the amount of persons living in poverty or the unemployment rate, something that will make them aware of their reality.




For the last case we used the Machine Learning Toolbox to predict the degree of a person based on their income.



Once the user selected the category that they felt that it was crucial for their learning process or that they felt more interested, we make the user select a specific year or the combination of all the years. It is important to know our country, not only in the present but also in previous years. In this way we will have a broader vision of our reality, of the things that occur in the country, and for what reason they occur.

These first two menus are the base menu that every user will receive. The next menus will depend on the selection of the user. We will present several of these cases and how they might look.

## Case 1: Age & Sex (in this example we use the information from 2015)



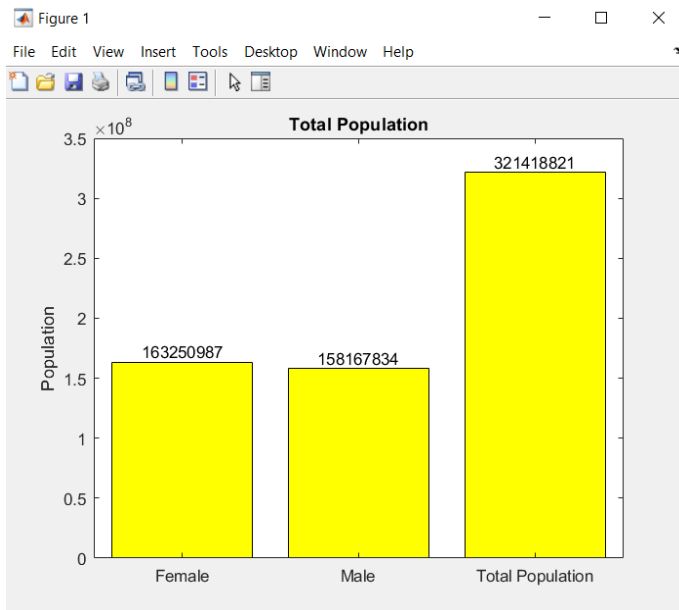
Pick a category

Total Population

Age

Race

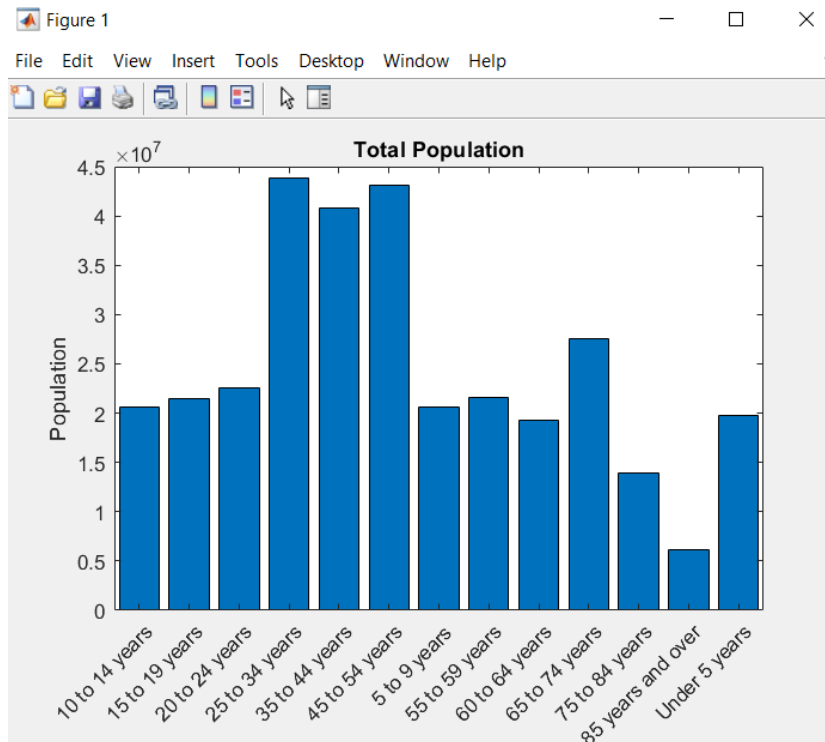
Assuming that the user selected Age & Sex and a specific year , then it will be prompted to select a topic about the category selected.



If the user selected Total Population then the following graph of that specific year will appear. It presents the population divided in three categories (Total Population, Male, and Female).

Total Population is 321418821  
Male is 158167834  
Female is 163250987

If the user, instead, selects the option of Age he/she will receive the following graph and information.



#### Command Window

```
Under 5 years is 19793807
5 to 9 years is 20582473
10 to 14 years is 20627389
15 to 19 years is 21426912
20 to 24 years is 22541077
25 to 34 years is 43897832
35 to 44 years is 40804130
45 to 54 years is 43135580
55 to 59 years is 21590716
60 to 64 years is 19286425
65 to 74 years is 27587267
75 to 84 years is 13984046
85 years and over is 6161167
>>
```

(This information will make the user have a clearer understanding of how young or how old are the people that compose their country and how the future years might look according to this data.)

The final option of the Sex & Age Graph is Race. Here the user will have the opportunity to learn the population of a specific race in the United States in a specific year or in the last 10 years.

MENU

Pick a Category

(Total Population)	(Samoan)	(Two races excluding Some other race, and Three or more races)
(One Race)	(Other Pacific Islander)	
(Two or more races)	(Some other race)	
(One race)	(Two or more races)	
(White)	(White and Black or African American)	
(Black or African American)	(White and American Indian and Alaska Native)	
(American Indian and Alaska)	(White and Asian)	
(Cherokee)	(Black or African American and American Indian and Alaska Native)	
(Chippewa)	(Hispanic or Latino of any race)	
(Navajo)	(Mexican)	
(Sioux)	(Puerto Rican)	
(Asian)	(Cuban)	
(Asian Indian)	(Other Hispanic or Latino)	
(Chinese)	(Not Hispanic or Latino)	
(Filipino)	(White alone)	
(Japanese)	(Black or African American alone)	
(Korean)	(American Indian and Alaska Native alone)	
(Vietnamese)	(Asian alone)	
(Other Asian)	(Native Hawaiian and Other Pacific Islander alone)	
(Native Hawaiian and Other Race)	(Some other race alone)	
(Native Hawaiian)	(Two or more races)	
(Guamanian or Chamorro)	(Two races including Some other race)	

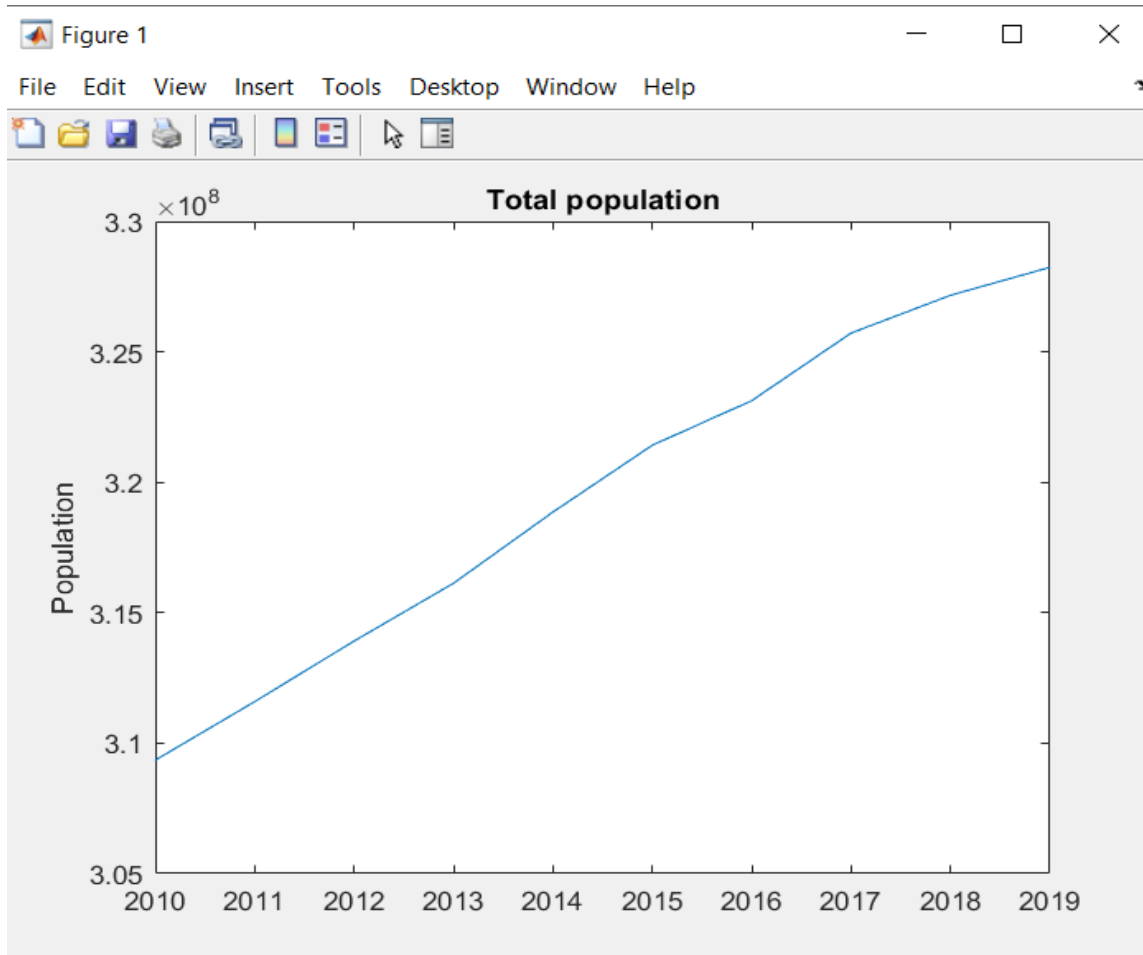
The population of (Mexican) people in the United States in 2015 was 35797080 .

fx

>>

If the user wants to learn the population of Mexican in 2015, he/she will receive exactly what they want as shown above. We believe that if the citizens know this information about race, it will reduce the amount of discrimination in the country because they will be aware of how diverse is the United States.

If instead of selecting a specific year, the user selected all years (2010-2019), he/she will receive a random fact of the Age & Sex table based on the last ten years, as shown below.



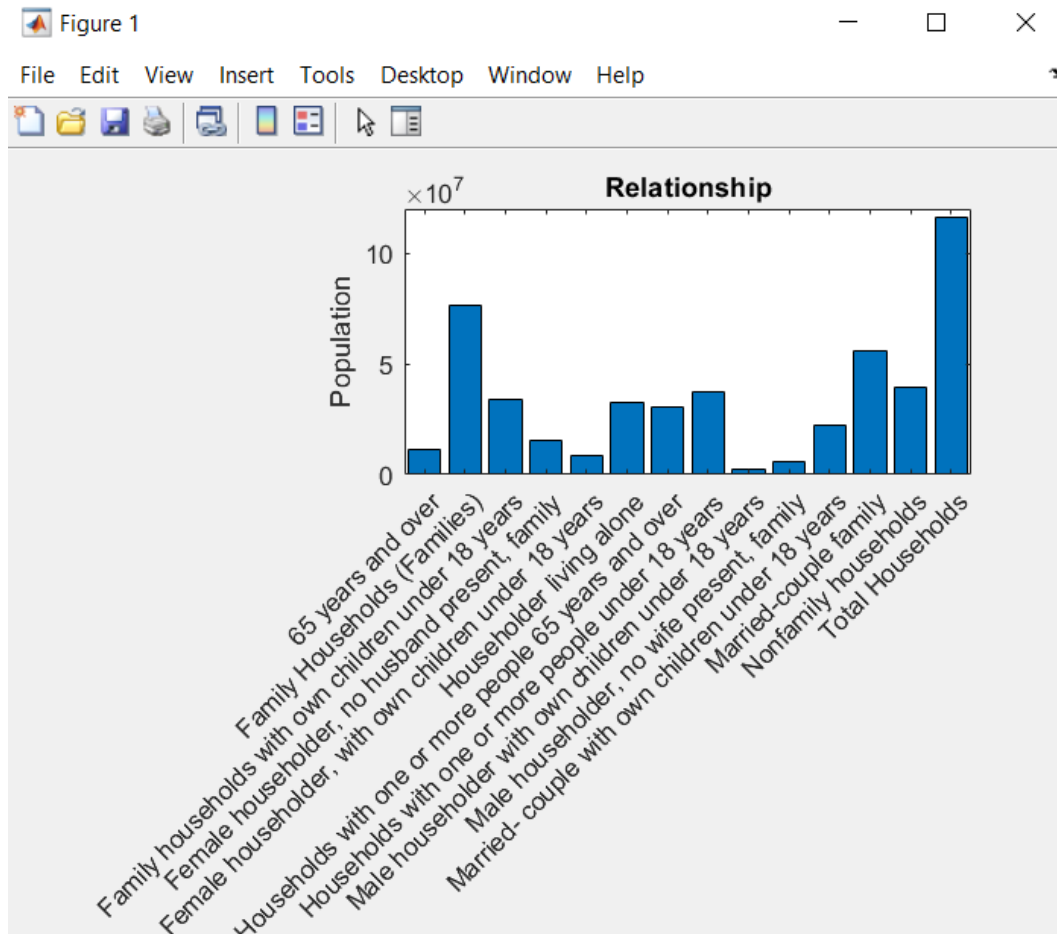
```
2010 is 309349689
2011 is 311591919
2012 is 313914040
2013 is 316128839
2014 is 318857056
2015 is 321418821
2016 is 323127515
2017 is 325719178
2018 is 327167439
2019 is 328239523
```

*fx* >>

## Case 2: Families and Living Arrangements

If the user picks this category it will have a several amount of options to choose from and one of the results might be:

(This information is based on 2012)

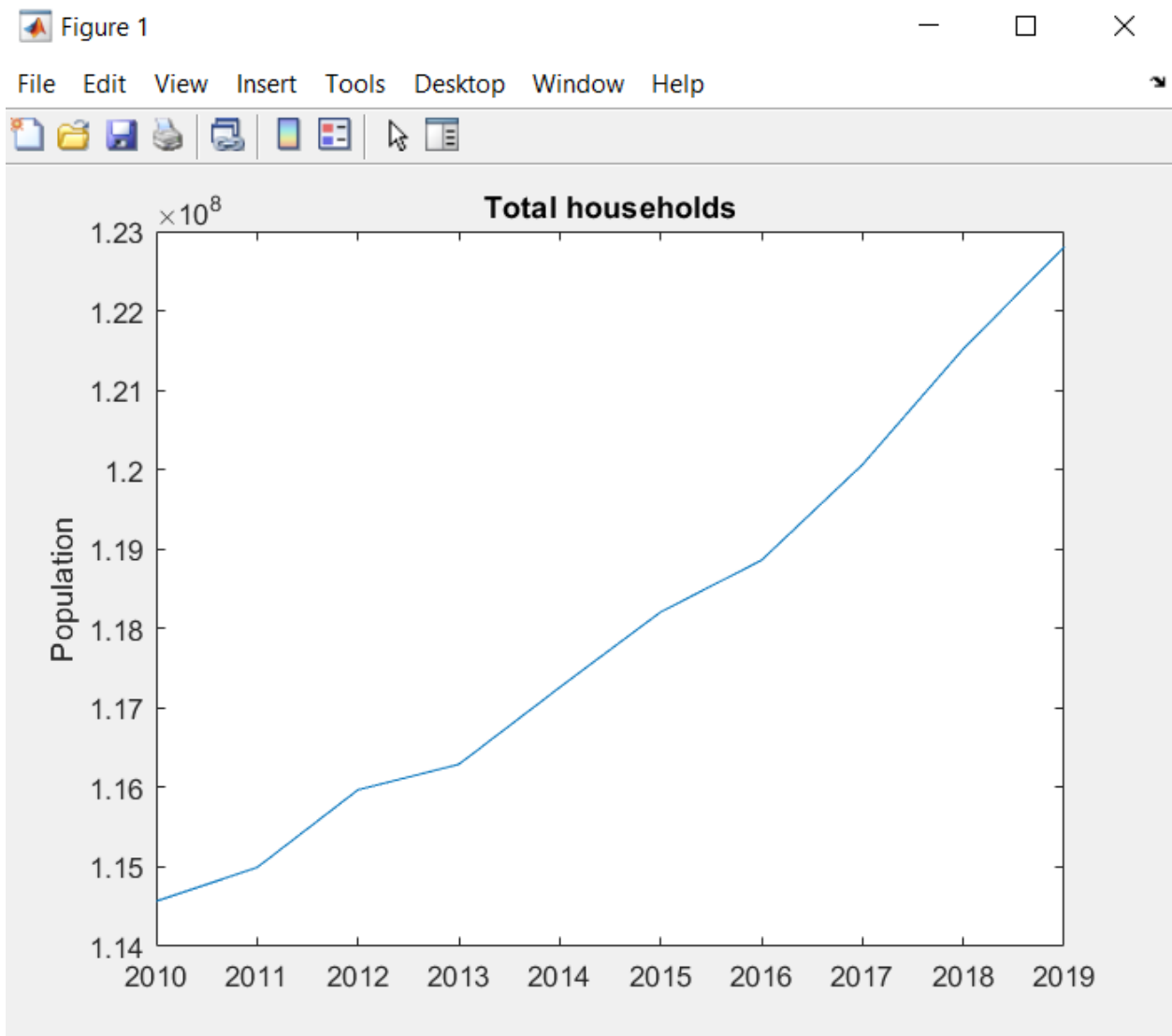


```
Total Households is 115969540
Family Households (Families) is 76509262
Family with own children under 18 years is 33612973
Married-couple family is 55754450
Married - couple with own children under 18 years is 22423949
Male householder, no wife present, family is 5578212
Male householder with own children under 18 years is 2697636
Female householder, no husband present, family is 15176600
Female householder, with own children under 18 years is 8491388
Nonfamily households is 39460278
Householder living alone is 32256217
65 years and over is 11513067
Households with one or more people under 18 years is 37555698
Households with one or more people 65 years and over is 30193187
```

*fx* >>



The option of (2010-2019) will produce the following graph:

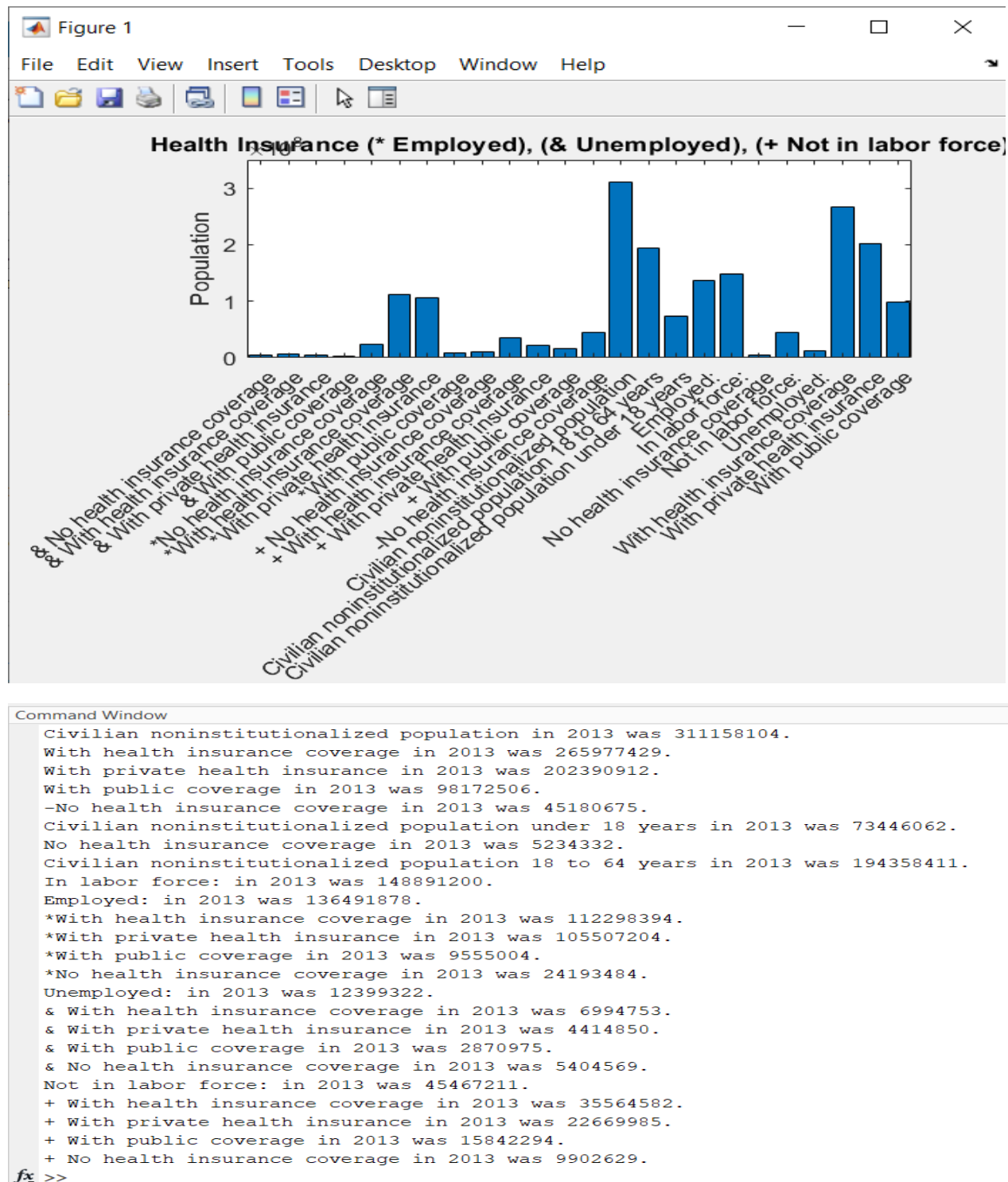


```
2010 is 114567419
2011 is 114991725
2012 is 115969540
2013 is 116291033
2014 is 117259427
2015 is 118208250
2016 is 118860065
2017 is 120062818
2018 is 121520180
2019 is 122802852
```

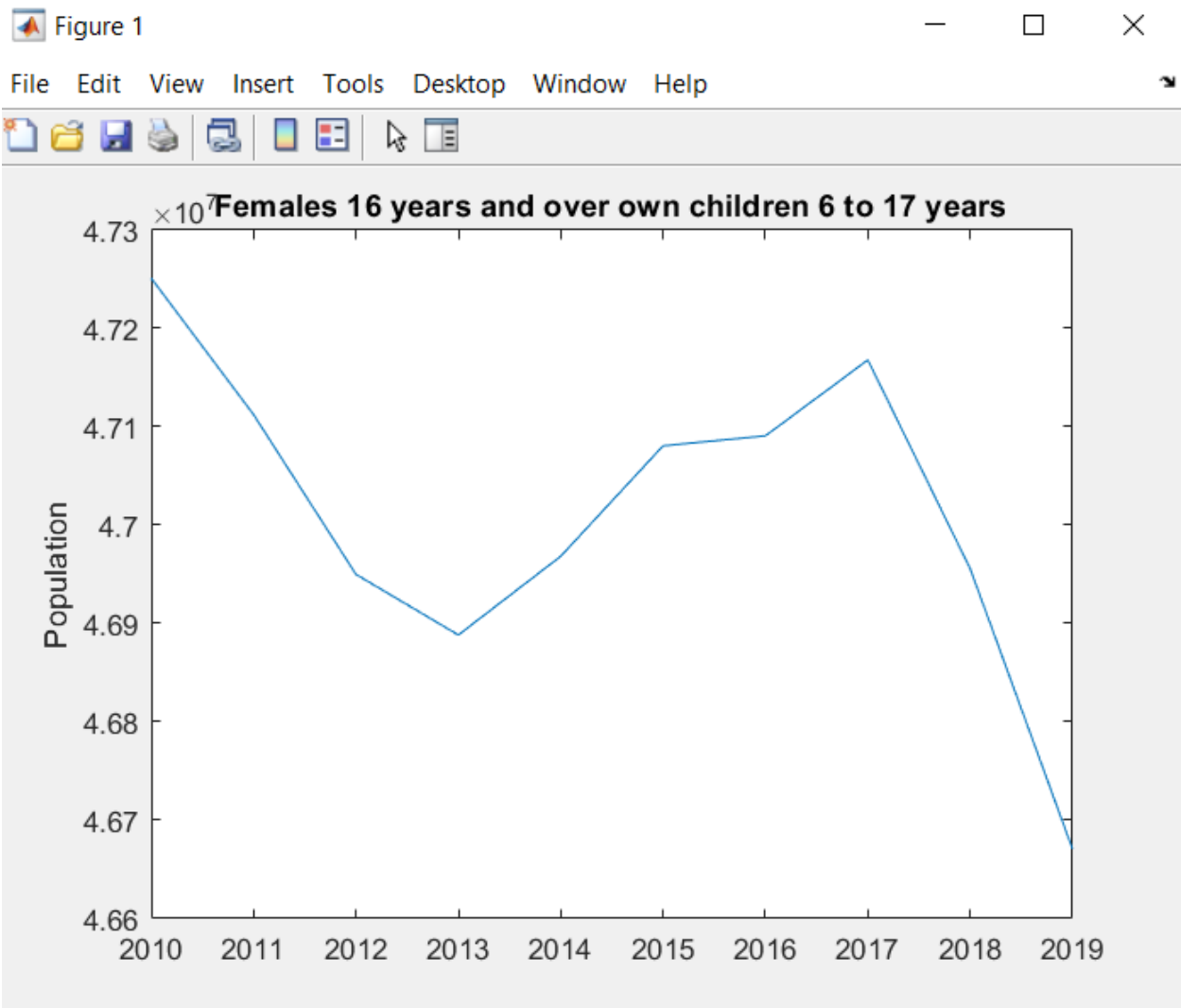
*fx* >>

### Case 3: Health

One of the subcategories of Health is Health Insurance. Here we show an example of what the user output might be. This graph is based on the year 2013.




The option of (2010-2019) will produce the following graph:



```
2010 is 47251112
2011 is 47112352
2012 is 46949993
2013 is 46888388
2014 is 46968394
2015 is 47080679
2016 is 47090847
2017 is 47167941
2018 is 46955632
2019 is 46670652
>>
```

## Case 4: Education

 MENU — □ ×

Pick a category

Household By Type

Relationship

Marital Status

Fertility

Grandparents

School Enrollment

Educational Attainment

Veteran Status

Disability

Residence one year ago

Place of Birth

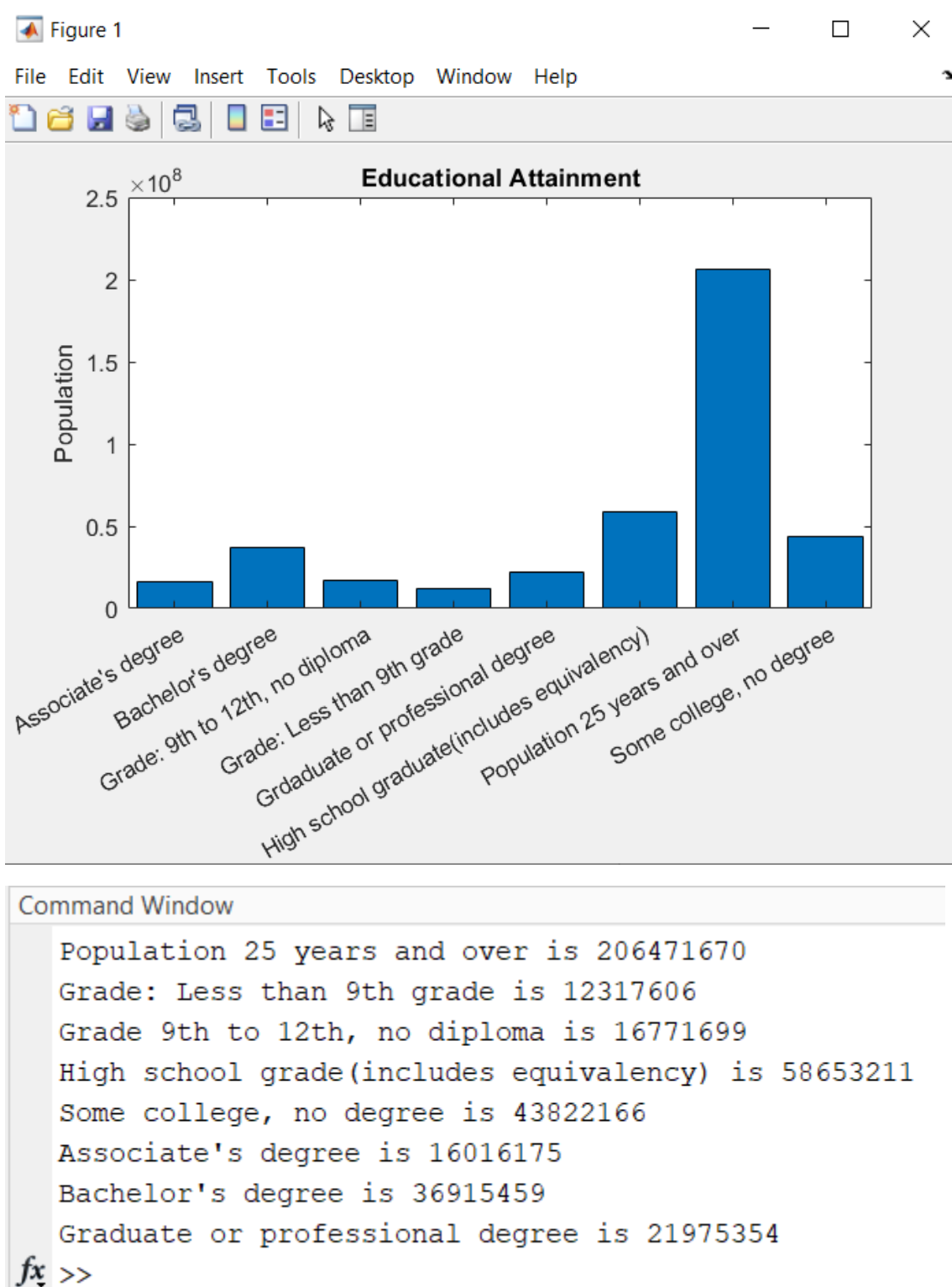
US Citizenship Status

World Region of Birth of Foreign Born

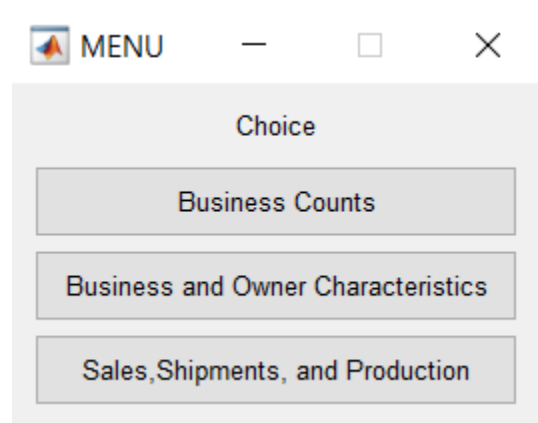
Ancestry

Under the Education table, the Census provide different subcategories. Here we created a menu such that the users can select their category of interest.

If, for example, the user picks the section of Education Attainment (2011), he would receive the desired information in a bar graph.



## Case 5: Business and Economy



A screenshot of a web application window titled "MENU". The window has a standard browser-like header with a small icon, the title "MENU", and three control buttons: a minus sign, a square, and an "X". Below the header, the main content area is titled "Choice" and contains three large, light-gray rectangular buttons stacked vertically. The buttons are labeled "Business Counts", "Business and Owner Characteristics", and "Sales, Shipments, and Production".

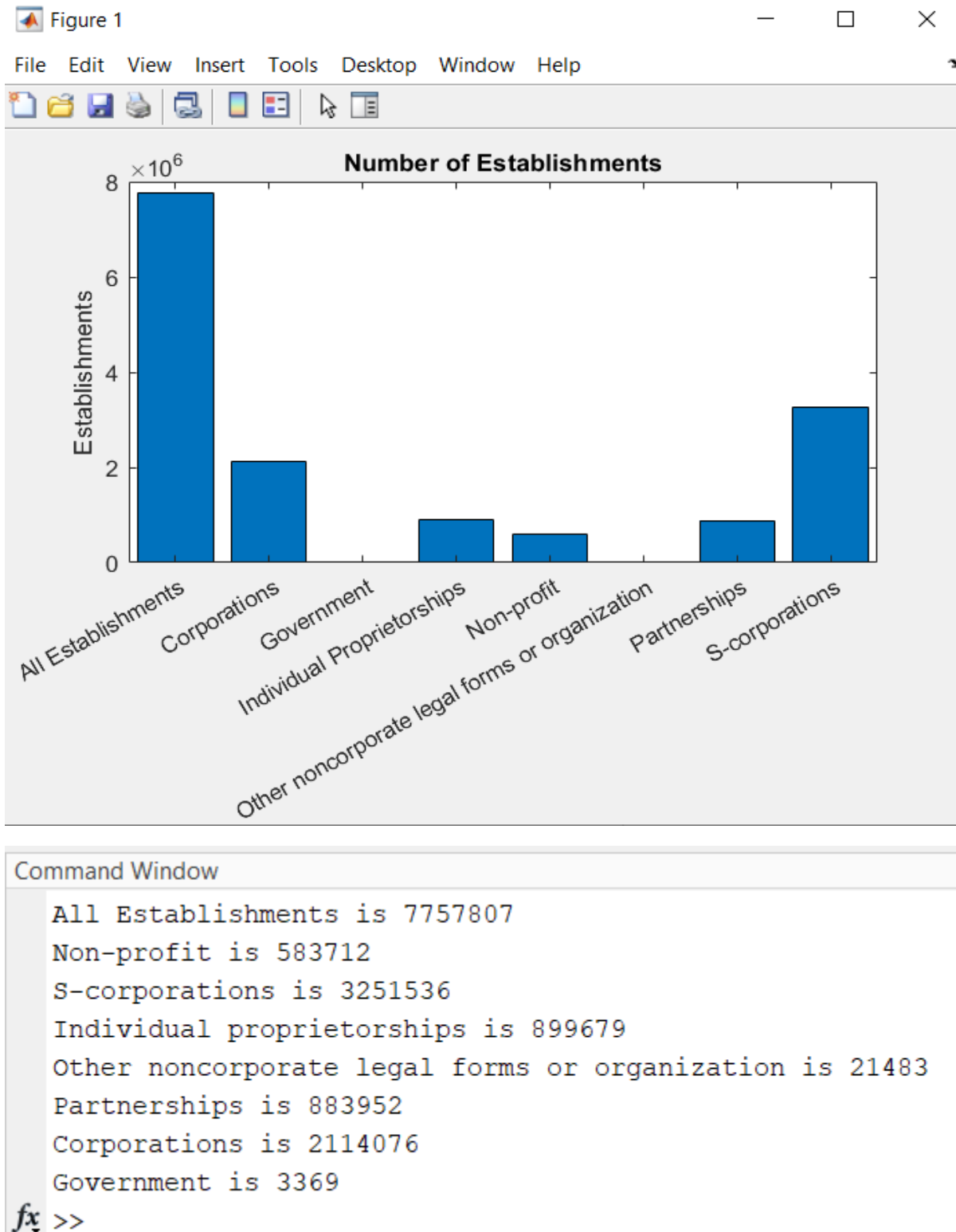
The table of Business and Economy is divided in three big categories as shown here. We created this menu so that the user receive specific and accurate data.



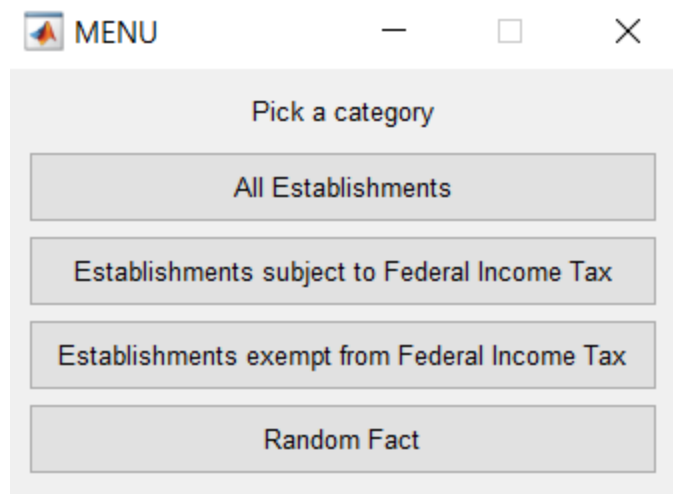
A screenshot of a web application window titled "MENU", similar to the one above. The header is identical. The main content area is titled "Pick a category" and contains a vertical list of ten light-gray rectangular buttons. The buttons are labeled: "All Establishments", "Establishments with 1 to 4 employees", "Establishments with 5 to 9 employees", "Establishments with 10 to 19 employees", "Establishments with 20 to 49 employees", "Establishments with 50 to 99 employees", "Establishments with 100 to 249 employees", "Establishments with 250 to 499 employees", "Establishments with 500 to 999 employees", and "Establishments with 1,000 employees or more". At the bottom of the list is a button labeled "Random fact".

If the user selects the first option (Business Counts) then another menu even more specific will appear.

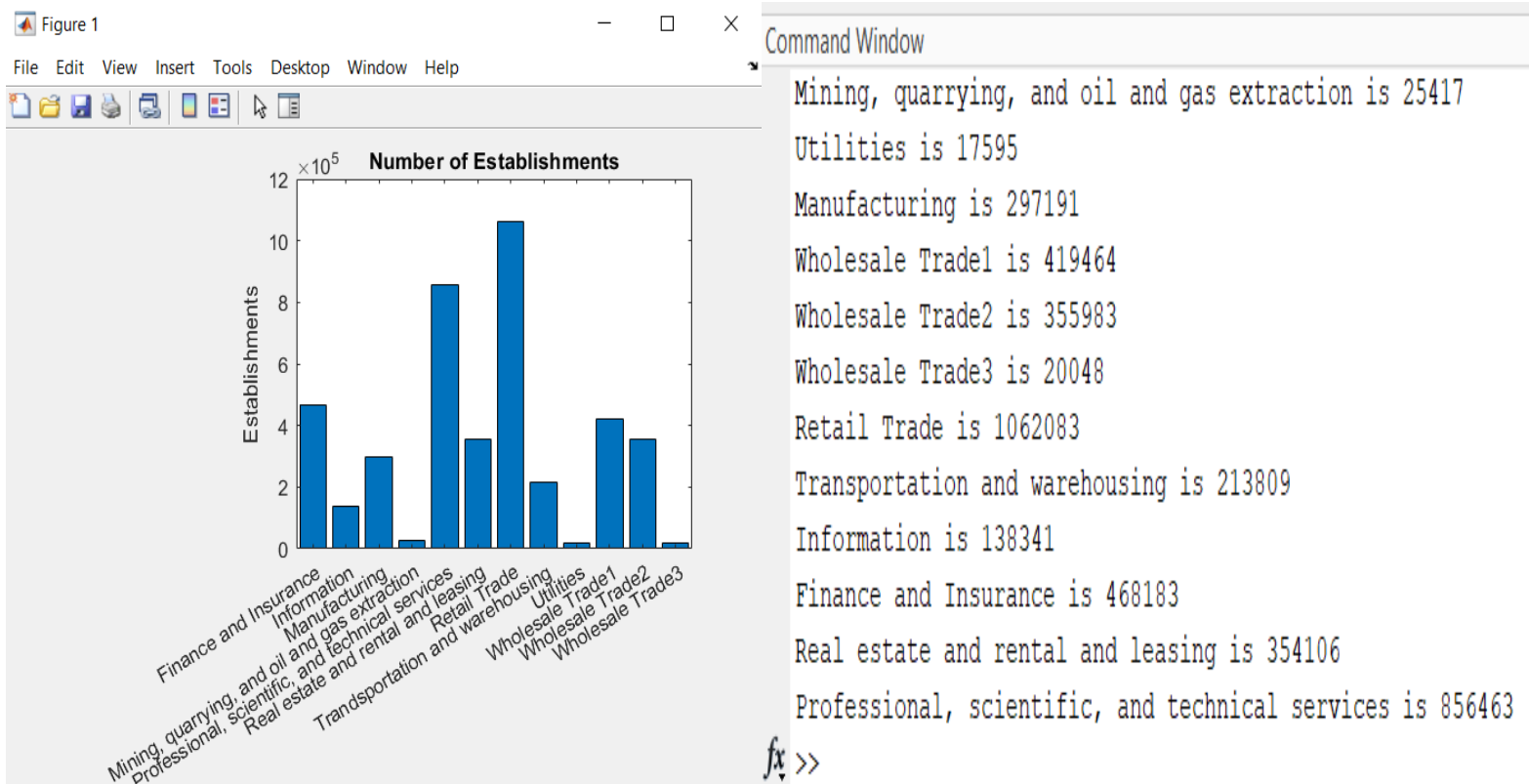
The option of all establishments will look like this. (The Census only provides information about the year 2016)



The third option of the first menu of Business and Economy will create another menu as shown below.



The output of All Establishments in 2012 will look like this:





## Case 6: Employment

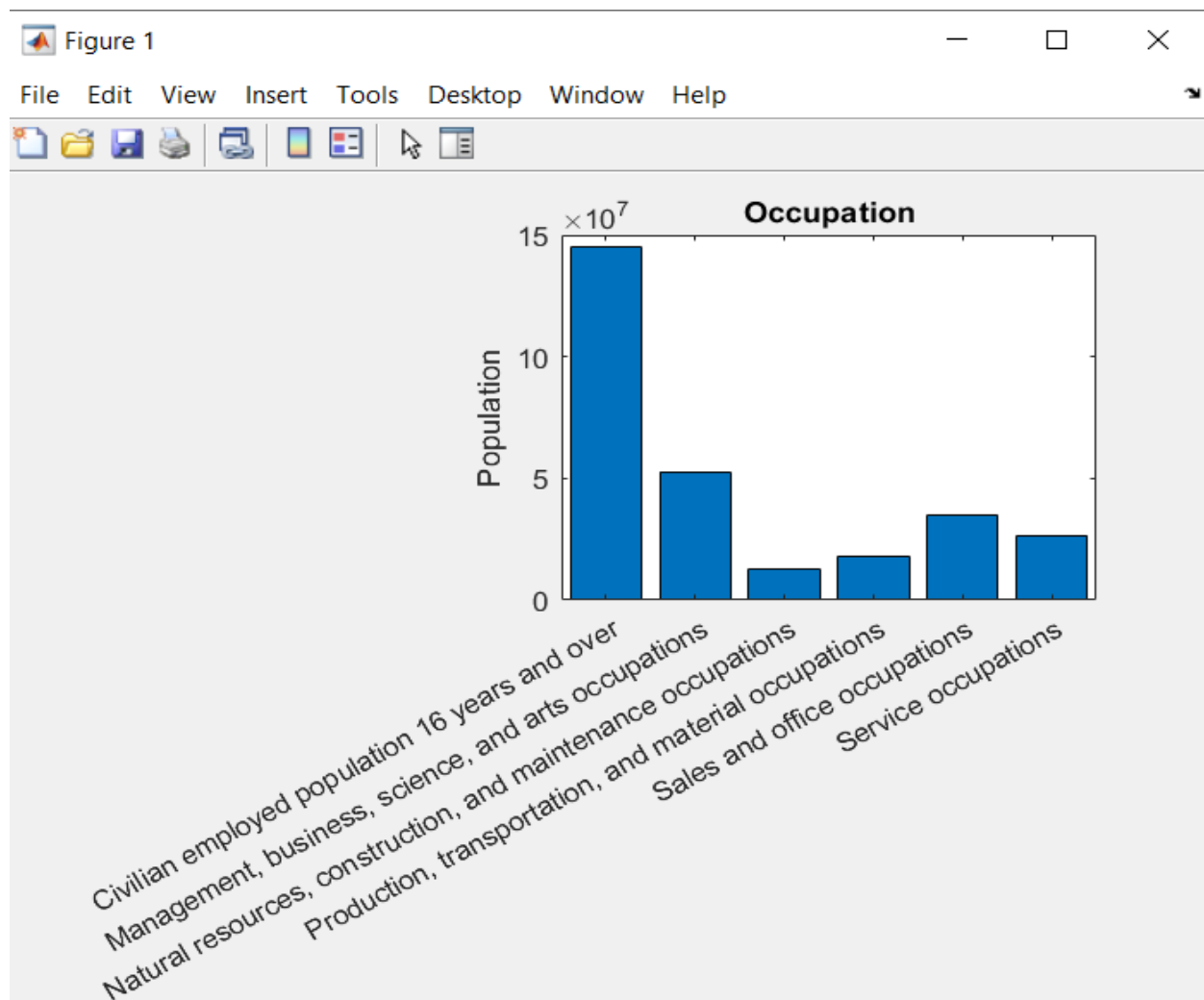


The screenshot shows a web application window with a title bar containing a logo, a minus sign, a square icon, and a close button. The main content area has a header "Pick a category" and a list of seven categories, each in a light gray button:

- Employment Status
- Commuting to Work
- Occupation
- Industry
- Class of Worker
- Income and Benefits
- Health Insurance coverage

The case of Employment is composed of different subcategories as shown here.

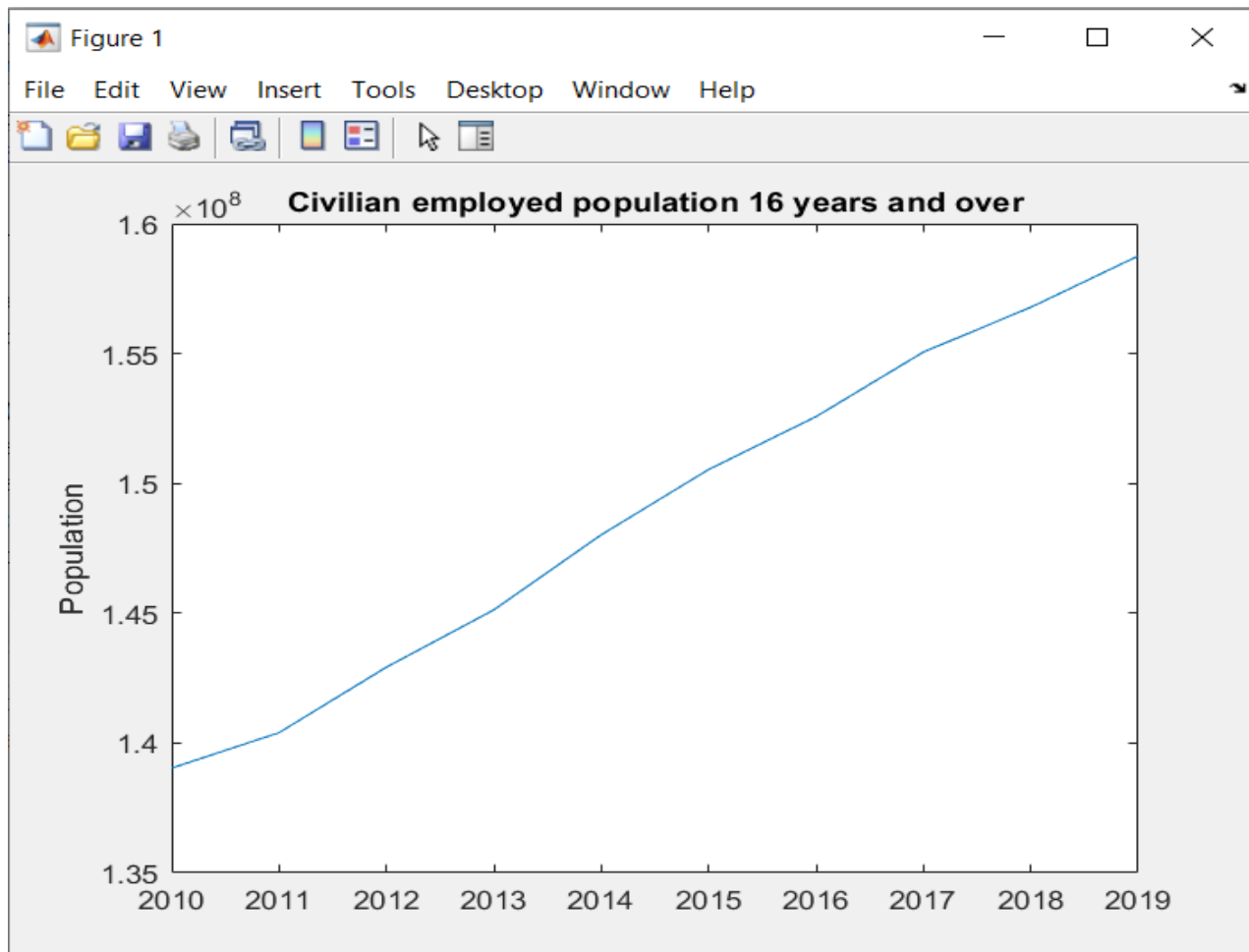
If the user selected Employment->2013->Occupation, then he/she will receive the following information:



#### Command Window

```
Civilian employed population 16 years and over is 145128676
Management, business, science, and arts occupations is 52753573
Service occupations is 26654335
Sales and office occupations is 35109334
Natural resources, construction, and maintenance occupations is 12924043
Production, transportation, and material occupations is 17687391
>>
```

If instead of selecting a specific year, he/she selects all years then the following graph might be one of the outputs. (The choice of 2010-2019 will return a graph of a random datapoint of that census table).

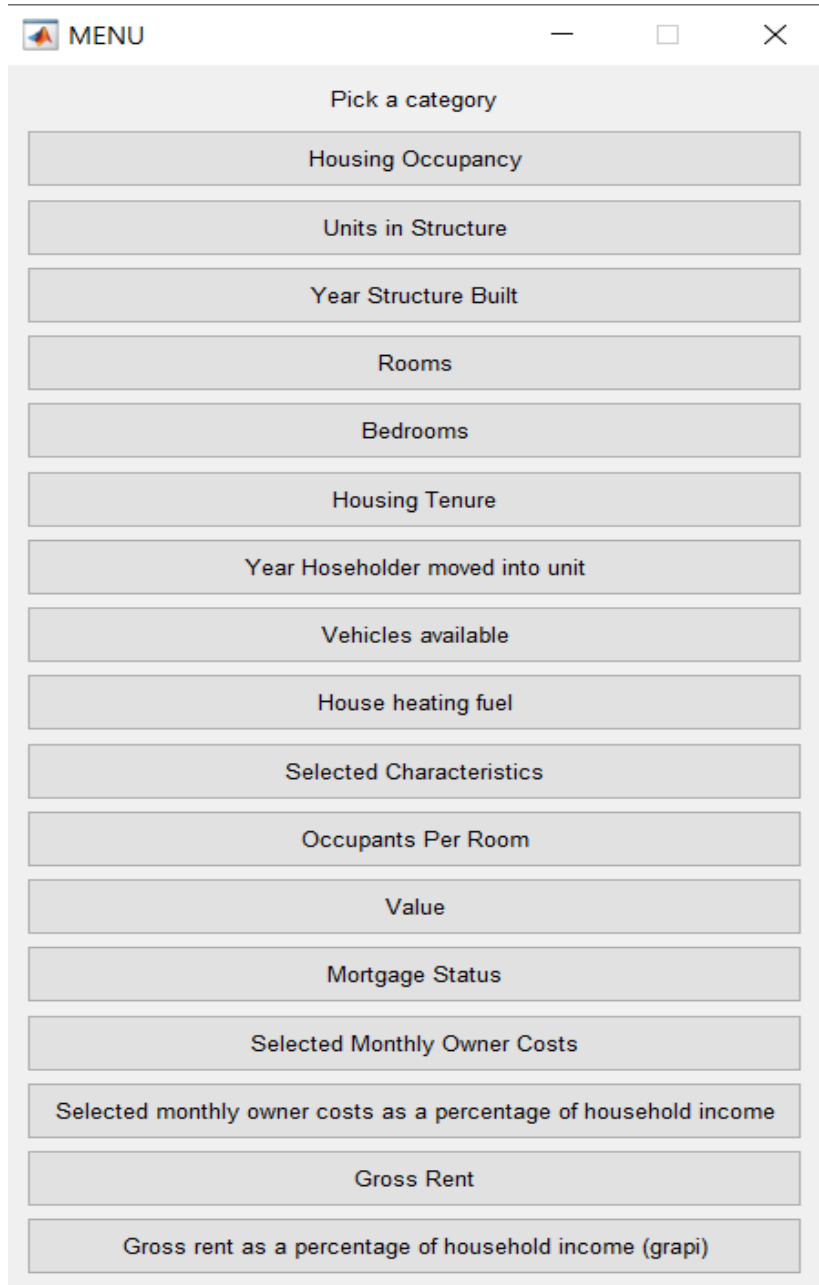


#### Command Window

```
2010 is 139033928
2011 is 140399548
2012 is 142921687
2013 is 145128676
2014 is 148019908
2015 is 150534773
2016 is 152571041
2017 is 155058331
2018 is 156783165
2019 is 158758794
>>
```

## Case 7: Housing

Once the user selected the option of Housing and the year of interest, it will be prompted with the following choices:

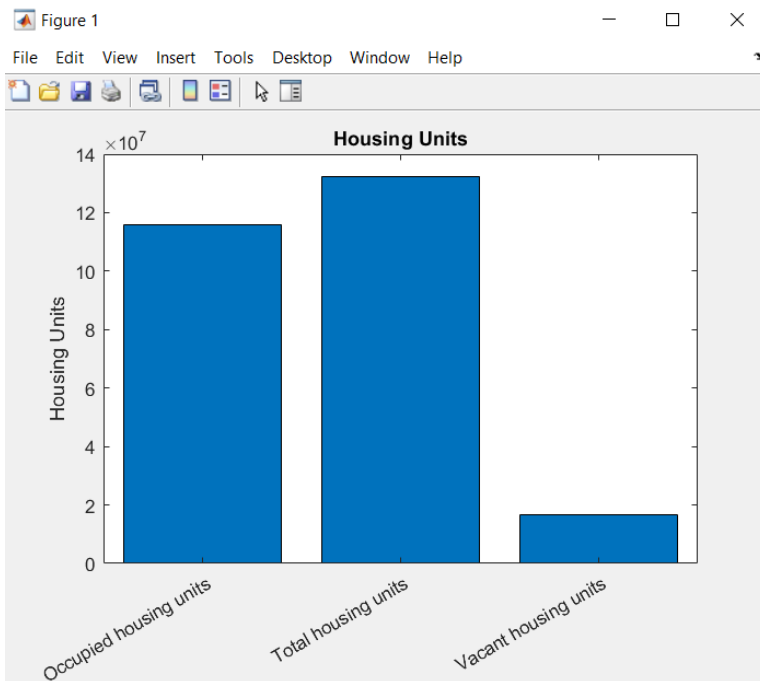


A screenshot of a software window titled "MENU". The window has a standard title bar with a minimize button, a maximize button, and a close button. Below the title bar, the text "Pick a category" is centered. A list of 18 categories is displayed, each in a separate rectangular button. The categories are: Housing Occupancy, Units in Structure, Year Structure Built, Rooms, Bedrooms, Housing Tenure, Year Hoseholder moved into unit, Vehicles available, House heating fuel, Selected Characteristics, Occupants Per Room, Value, Mortgage Status, Selected Monthly Owner Costs, Selected monthly owner costs as a percentage of household income, Gross Rent, and Gross rent as a percentage of household income (grapi).

Pick a category
Housing Occupancy
Units in Structure
Year Structure Built
Rooms
Bedrooms
Housing Tenure
Year Hoseholder moved into unit
Vehicles available
House heating fuel
Selected Characteristics
Occupants Per Room
Value
Mortgage Status
Selected Monthly Owner Costs
Selected monthly owner costs as a percentage of household income
Gross Rent
Gross rent as a percentage of household income (grapi)

This are all of the options that the Census provides under the topic of Housing.

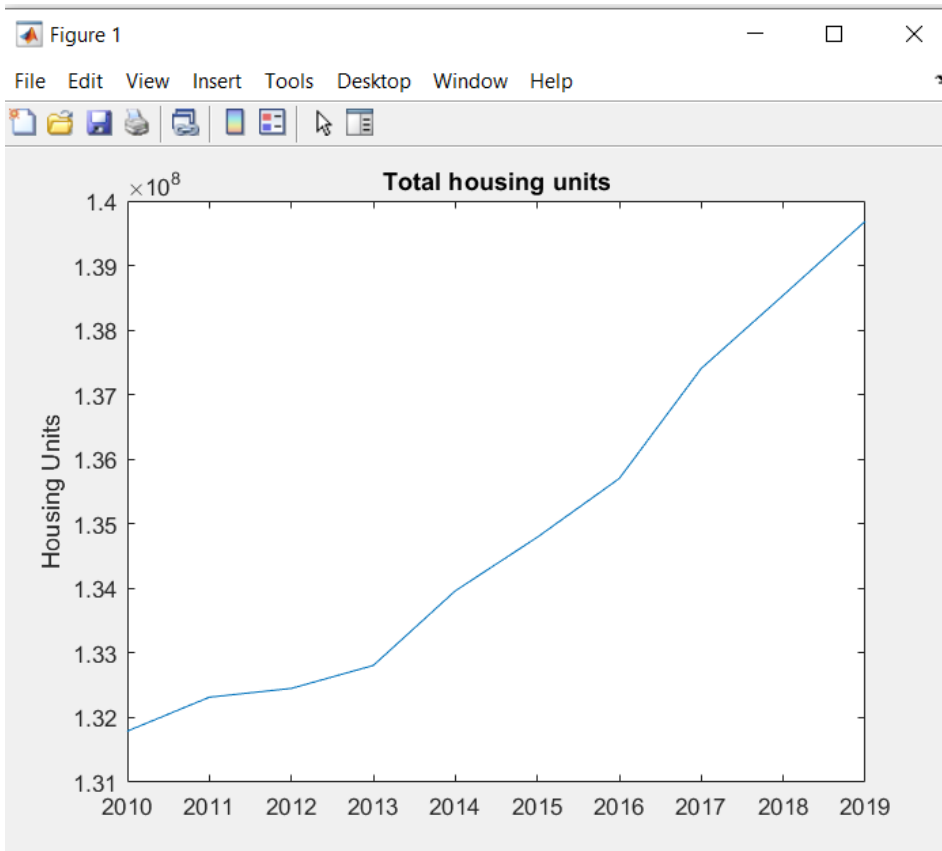
Some of this outputs might be the following (this are based in the year 2012):



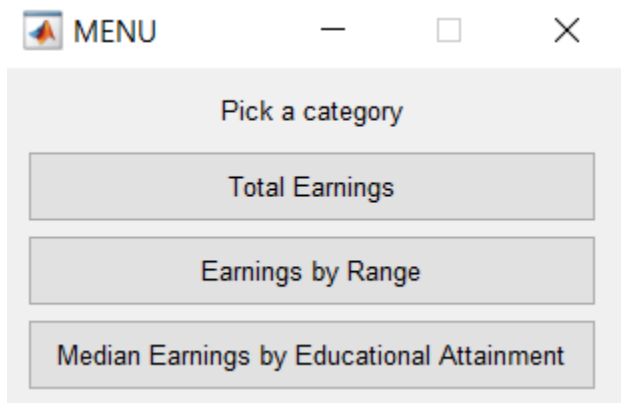
#### Command Window

```
Total housing units is 132452249  
Occupied housing units is 115969540  
Vacant housing units is 16482709  
>>
```

The output of all years might be the following:

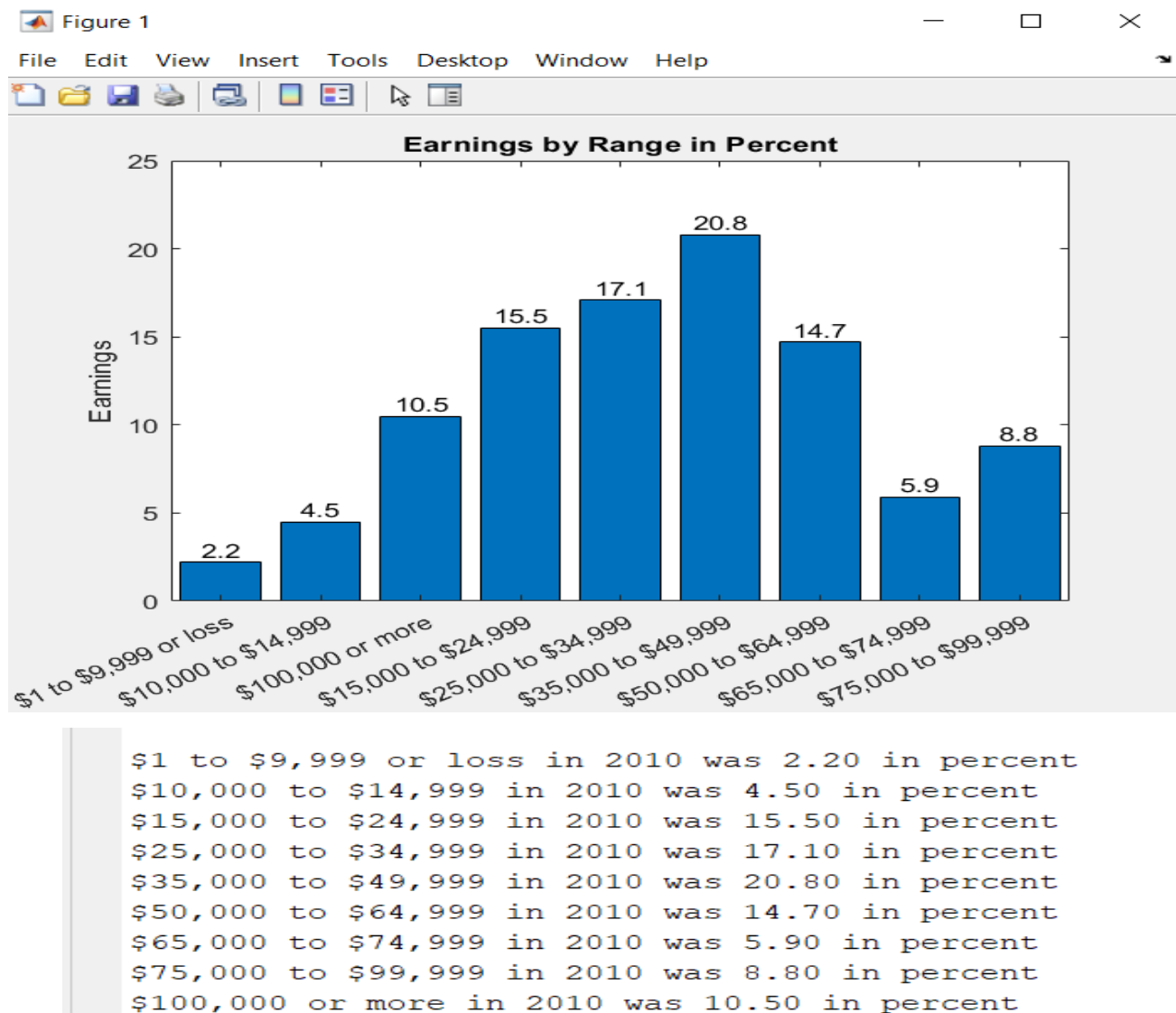


## Case 8: Income and Poverty

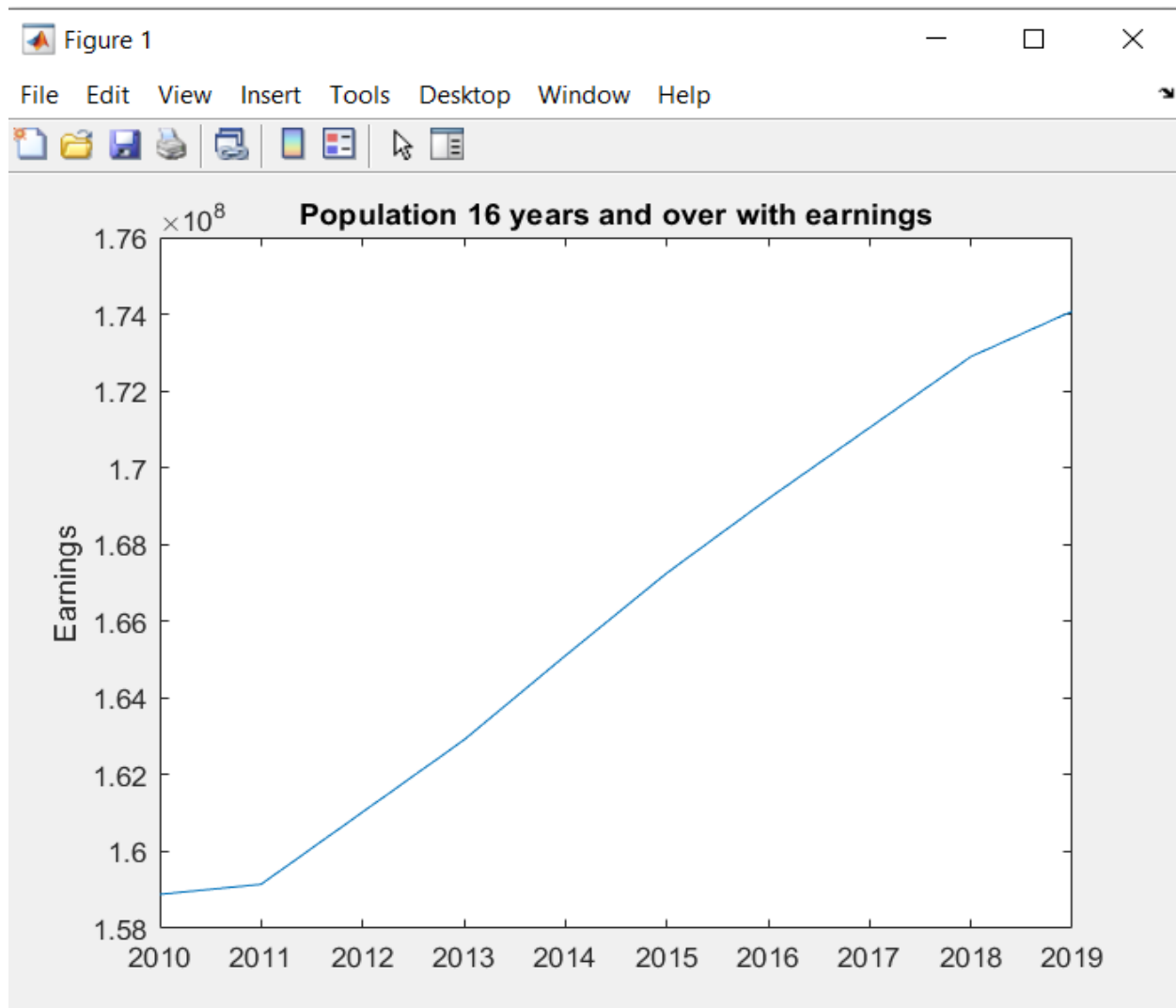


This are all the different options that the Census provides under the topic of Income and Poverty; so we decided to create a menu so that the user have the opportunity to learn about all these topics.

If, for example, the user selected the option of Earnings by range in the year 2019, he/she will receive the following output:




If instead, the option of all years is chosen, then one of the possible outputs might be the following graph.



```
2010 is 158884879
2011 is 159147308
2013 is 162908126
2014 is 165102809
2015 is 167254814
2016 is 169190685
2018 is 172894743
2019 is 174079762
>>
```

## Case 9: Veterans

If the user selected veterans, it will have the opportunity to select one of the following dataset related to veterans.

 MENU — □ ×

Pick a category

Total Veterans

Period of Service

Sex

Race

Median Income in the Past 12 Months

Educational Attainment

Employment Status

Poverty Status in the Past 12 Months

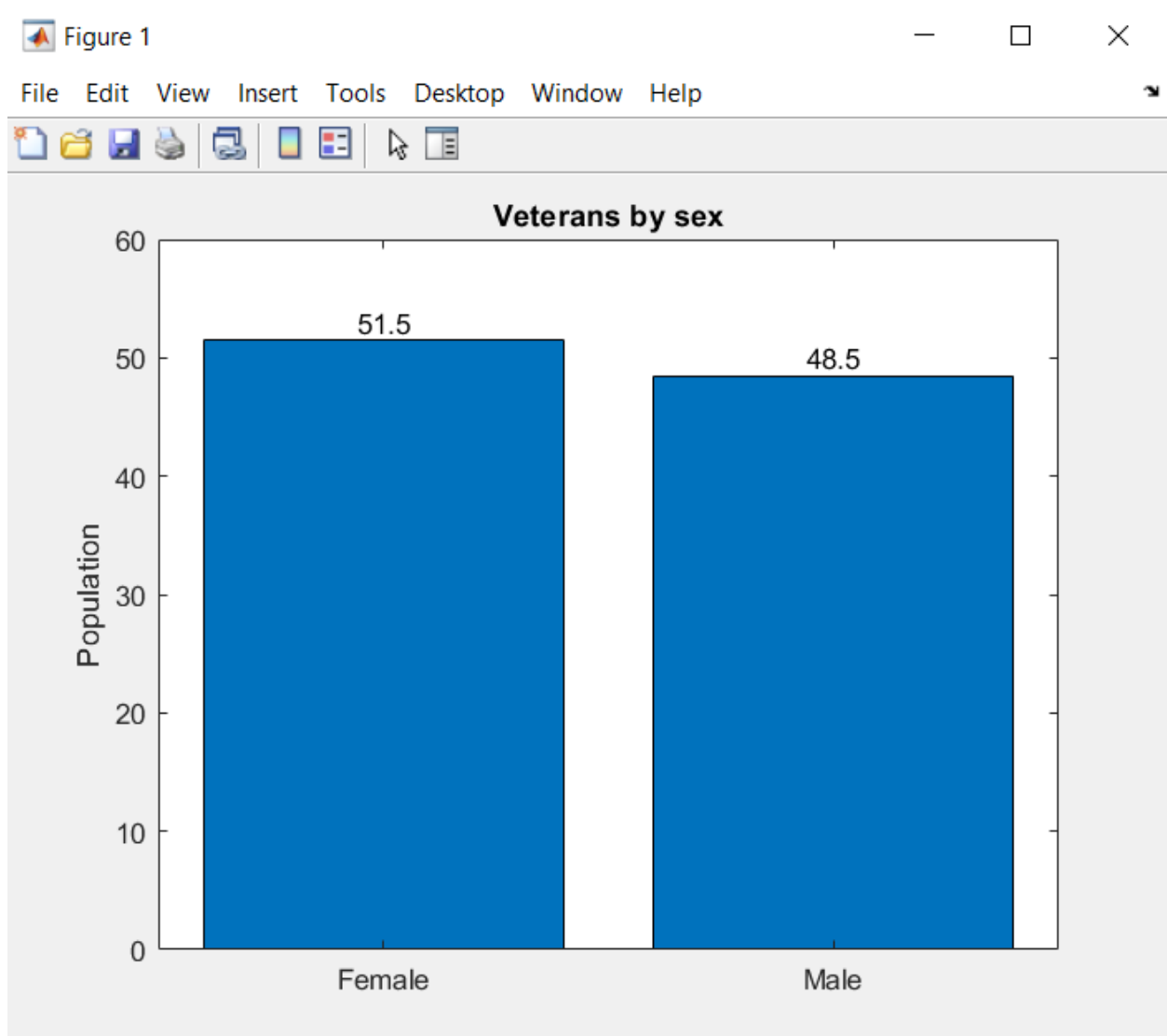
Disability Status

Age



The outputs that are about to be shown are based on the year 2019 (the selection made by the user).

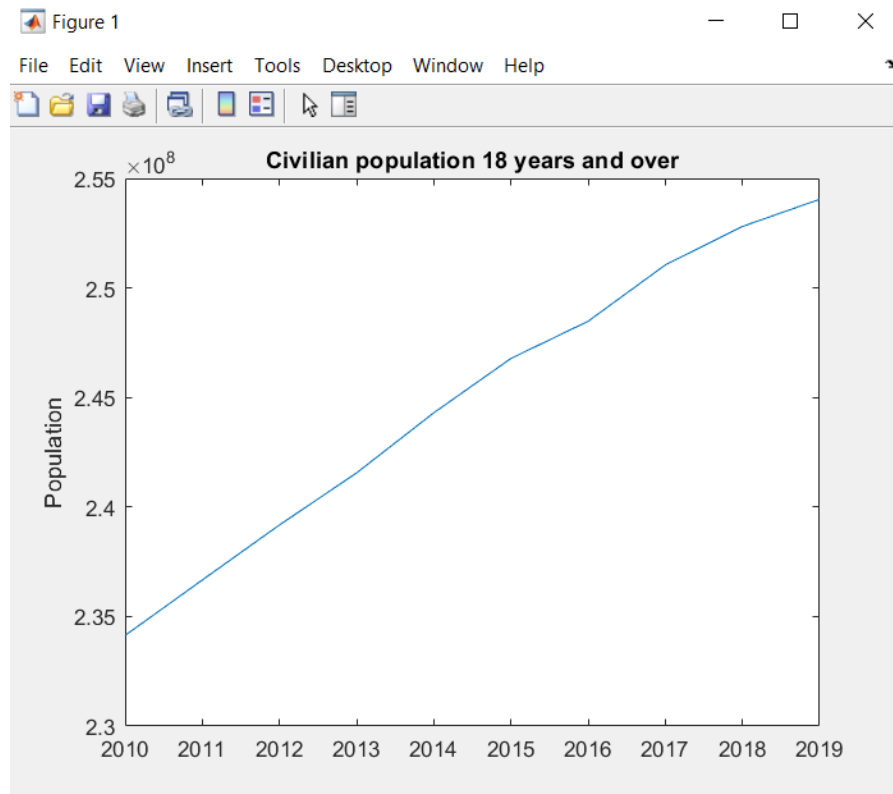
Here we show an example of Veterans with the option of Veterans by Sex.



There were 48.50 percent Male veterans in 2019.  
There were 51.50 percent Female veterans in 2019.

*fx* >>

If the option of all years is selected, then a possible output might be the following graph. The program will randomly pick a line from the table and will use the data from the same line from every year from 2010 and 2019 and will plot the data values. The graph below won't appear always, another graph showcasing another topic from the data table might appear if the user decides to pick the same topic again :



```
2010 is 234137287
2011 is 236665774
2012 is 239178768
2013 is 241556724
2014 is 244298660
2015 is 246780172
2016 is 248478651
2017 is 251047650
2018 is 252806449
2019 is 254046196
>>
```

## Case 10:

If the user selected the last option which was “predict your degree based on your income”, it will first prompt the user to enter their income from 2019 to 2015. And if the user enters this information, it will print out what our trained program (See below) thinks the user’s degree is.

```
>> projectek125
Please enter your income for 2019: 120000
Please enter your income for 2018: 115000
Please enter your income for 2017: 100000
Please enter your income for 2016: 90000
Please enter your income for 2015: 80000
Your degree is Graduate or professional degree
```

There are five different outputs, which are: less than high school, highschool graduate, some college or associate’s degree, bachelor degree, or graduate or professional degree

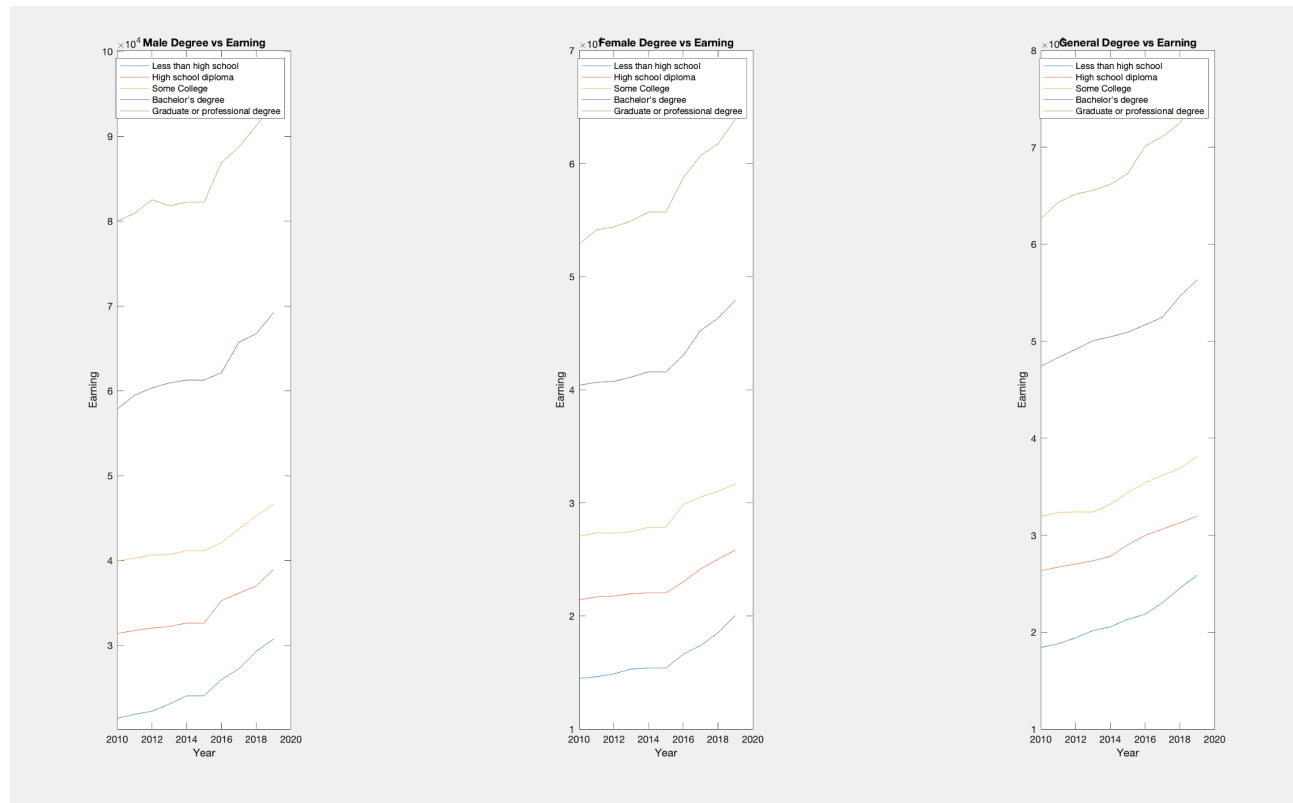
```
>> projectek125
Please enter your income for 2019: 19000
Please enter your income for 2018: 17000
Please enter your income for 2017: 12000
Please enter your income for 2016: 50000
Please enter your income for 2015: 0
Your degree is Less than high school graduate
```

```
>> projectek125
Please enter your income for 2019: 30000
Please enter your income for 2018: 30000
Please enter your income for 2017: 25000
Please enter your income for 2016: 24000
Please enter your income for 2015: 20000
Your degree is High school graduate (includes equivalency)
```

```
>> projectek125
Please enter your income for 2019: 42000
Please enter your income for 2018: 40000
Please enter your income for 2017: 37000
Please enter your income for 2016: 36000
Please enter your income for 2015: 35000
Your degree is Some college or associate's degree
```

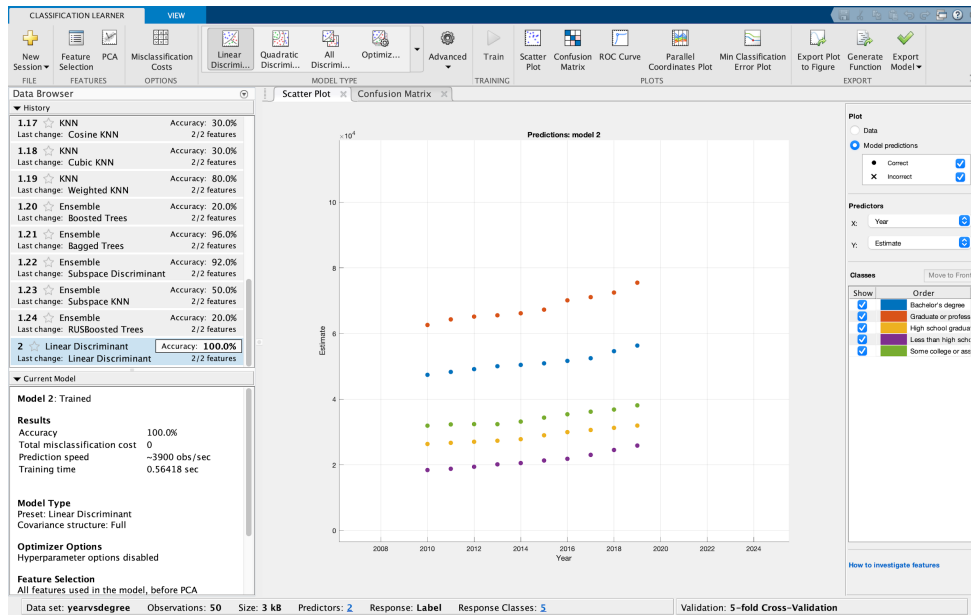
```
>> projectek125
Please enter your income for 2019: 50000
Please enter your income for 2018: 40000
Please enter your income for 2017: 30000
Please enter your income for 2016: 25000
Please enter your income for 2015: 0
Your degree is Bachelor's degree
```

After the system prints out its prediction, then a graph showcasing the earnings for male, female, and the general population based on their degree over 2010 to 2019



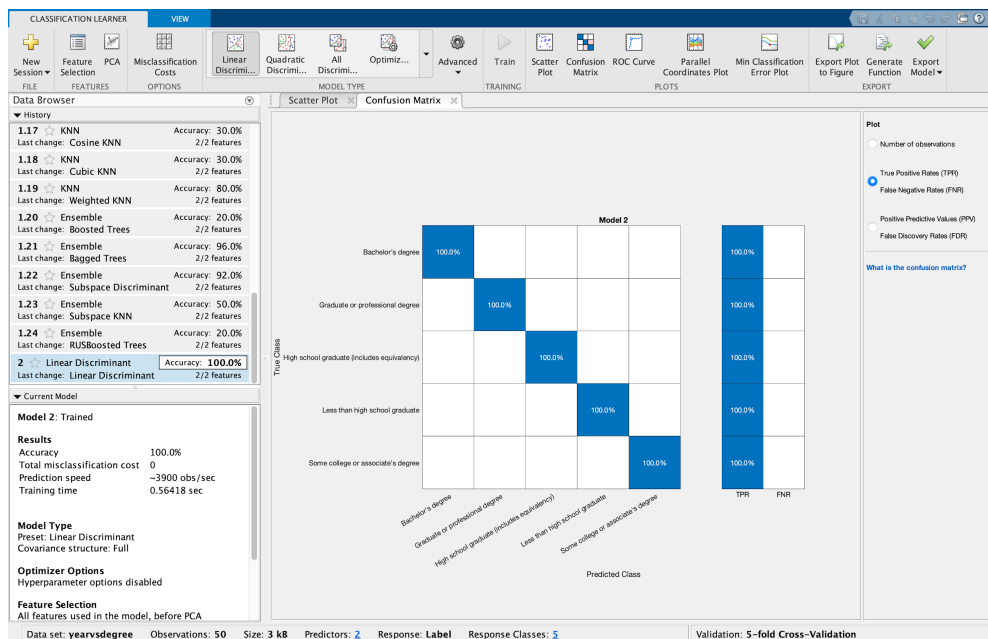
By then seeing these data graphs, the user might compare what another person with the same degree might be earning. Unfortunately, these earnings are just general, the census table didn't provide information of the earnings for different majors, and so we can't get into specific. How is this predicting your degree based on your earnings related to our project and the problem we are trying to solve? People will be able to learn more about other wages of people with different or similar degrees and earnings. Employers might look at data like these and could even give their workers a raise because they might have picked the household options and might've learned about the rent cost. Then the employer might think that an employee needs a raise because they're struggling with all these expenses.

## How did we use the Machine Learning Toolbox?



We use the classification learner tool and train it in order to predict a user's educational degree based on their income. We used data found in the census earnings table and plotted year vs earning for

different types of degree such as: less than high school, high school graduate, some college, bachelor's degree, and graduate or professional degree. We use the linear discriminant option in order to train and predict the degree type.



In this confusion matrix plot, you can see that the program predicted 100% of the degree. Unlike predicting numbers, this program predicts a category.

## **How did we accomplish our goal?**

Our main goal for this experiment was to create a program that is user friendly, where a user could pick any topic from the U.S. census and learn more about such topics. Overall, we wanted the user to realize how useful and valuable data is, not only that, but how we could use different tools to predict something based on a given data. We believe that we accomplish these goals that we had in mind, not only did we create an easy to use program, where it's engaging and easy to learn for the user, but we also used tools such as classification learner from Matlab to predict a person's degree based on their income. Overall, we believe that by using our program, citizens and even outsiders of the United States will learn more about different topics of the U.S, such topics are education, business, population, etc. And the more people know about these topics, the more progressive the United States will become, because more people will become more educated and education is the premise of progress.

## Article

The article written by Prabhakar Krishnamurthy, “Understanding Data Bias.” First gives us examples of some instances of biases caused by ML, such as Amazon shutting their ML program that scores an applicant’s application because it punished female applicants, and how an ad ranking system was accused of racial and gender profiling. Data bias, as Prabhakar stated, does not include variables that accurately help us predict something, or content produced by humans which contain bias. We see that data bias is broken down into 5 different categories: Response or Activity Bias (Which is content generated by humans), Selection bias due to feedback loops (Occurs when a model influences the generation of data that is used to train it), Bias due to system drift (changes over time to the system generating data), Omitted variable bias (some attributes are missing that influence the outcome), societal bias, (content produced by humans, such as gender or race stereotypes). Some solutions on identifying and stopping data bias are mapping a data generation process and design interventions to either pre-process data or get more data, perform exploratory data analysis, use data pre - processing before training, in-processing during training, and post-processing after training.

Krishnamurthy, Prabhakar. “Understanding Data Bias.” *Medium*, Towards Data Science, 22 Oct. 2019, [towardsdatascience.com/survey-d4f168791e57](https://towardsdatascience.com/survey-d4f168791e57).

### **Conclusion:**

In this project we applied our skills learned about Matlab in order to solve a societal problem. A problem that we feel that needs to be taken into account immediately. Our world needs a more educated society in order to promote change and progress. We consider that this project is a great start to achieve that change that will bring so many good things to this country. Through this project we invite citizens to appreciate the value of education and to become aware of the reality of our country because as we stated in our title: “Knowledge is Power. Education Is the Premise of Progress in Every Society”.



**Link of our code and data table we used:**

<https://drive.google.com/drive/folders/1U0otOWyxyy4shgBXrYLepOuOCQbVjW7B?usp=sharing>