**Report: Data Mining Assignment 1**

**Task 1:**

This task deals with data understanding, visualization and pre-processing of data.

**Data Understanding:**

This means information about data which we cannot get just by looking at the data such as measure of central tendency i.e. mean, median, mode and mid-range, the interquartile ranges, maximum value of a certain attribute in data, minimum value of a certain attribute in data, the standard deviations i.e. how close the data is to the mean value and the variance i.e. the square of the standard deviation.

**Data Visualization:**

This describes the effort to help people to understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization techniques. Some data visualization methods include box plot i.e. the five-number summary, the quantile plot, the quantile-quantile plot, data cube aggregation, Chernoff faces and many more.

**Data Preprocessing:**

This technique of data mining involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, and lacking certain behaviors or trends and is likely to contain many errors, can contain missing values, can also contain falsified data, can also contain noise. Data preprocessing is a technique to resolve such issues so that we can work with this data easily and without worrying about the above-mentioned hurdles in our data mining tasks.

**Task 1a:**

In this task we are asked to fill in the missing values for the highest and the lowest price column using the appropriate filter. The filter I used for this task is replace missing values with the average of that column and the following statistical tendencies are observed for both columns:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Min | Max | Average | Deviation | Variance |
| Lowest-Price | 0 | 300 | 0.470 | 4.567 | 20.857 |
| Highest-Price | 0 | 3050 | 2.043 | 18.980 | 356.832 |

In case of lowest price, the data is more centered around mean, hence the standard

deviation is very low as compared to that of highest price. So, data for highest price has

more spread and its values are therefore away from mean as compared to lowest price.

**File Name:** The file name for this task is named as **task1a**.

**Task 1b:**

The dishes name is sorted using the sort operator with the sorting direction as decreasing. This task is further divided into the following sub tasks:

**Task1b-1:**

For unique entries, I have used the remove duplicate operator on a single attribute: name. The result is that only 6 entries are duplicate and the remaining 99,993 are unique.

**Task1b-2:**

In this task we are asked to remove the entries which matches with numbers. For this purpose, I have used filter examples operator with invert filter using regular expression [0-9]+ which means that if name matches with any of the following numbers then remove it from the data set. The result is that from the 99,999 entries 99,988 examples are left. This means that 11 entries are removed.

**Task1b-3:**

In this task we are asked to remove the null entries i.e. the entries containing missing values. For this purpose, I have used the filter examples with a custom invert filter that if name =? values then remove those entries. The result is that 99,986 entries are left meaning that only tow entries are filtered out in this process.

**Task1b-4:**

In this task we are asked to remove the punctuation from the name column

using the appropriate filter. The filter I used is the replace filter with the regular expression: [-!"#$%&'()\*+,./:;<=>?@\[\\\]\_`{|}~]+ that if any of the punctuation is found replace it with a whitespace. This operation includes trailing and leading whitespaces in the name column. This operator is applied to the whole data set and 99,986 entries are left in the data set after this operation.

**Task1b-5:**

In this task, we are asked to remove the leading and trailing white spaces in the name column. For this purpose, I have used that operator trim to remove the leading and trailing whitespaces. The data is still 99,986 but in a more understandable format.

**Task1b-6:**

In this task we are asked to convert the uppercase letters to lower case letters. For this purpose, I have used 26 replace filters to replace every uppercase letter to lower case letter. The data is still 99,986 but much more understandable now.

**Task1b-7+1b-8:**

In this take we are asked to remove duplicate entries from the data set. The operator I have used is replace duplicated with attribute filter type as single. After this operation only 90,110 examples are left out in the data set.

**Task1b-9:**

In this task after identifying and replacing the ten close neighbors of English breakfast tea and then filtering them out, the size of data set changes from 90,110 to 90,100.

**Task1c:**

This task is already discussed in task 1a and task1b

**Task1d:**

The ten most popular dishes are recognized by the number of times they appeared on a menu. They are as follows in order of decreasing popularity:

|  |  |
| --- | --- |
| 1 | Boiled Potatoes |
| 2 | Milk |
| 3 | Chicken Salad |
| 4 | Assorted Cakes |
| 5 | French fried potatoes |
| 6 | Strewed Prunes |
| 7 | Stewed Tomatoes |
| 8 | Lima Beans |
| 9 | Cigars |
| 10 | Ice-cream |

**Task 1e:**

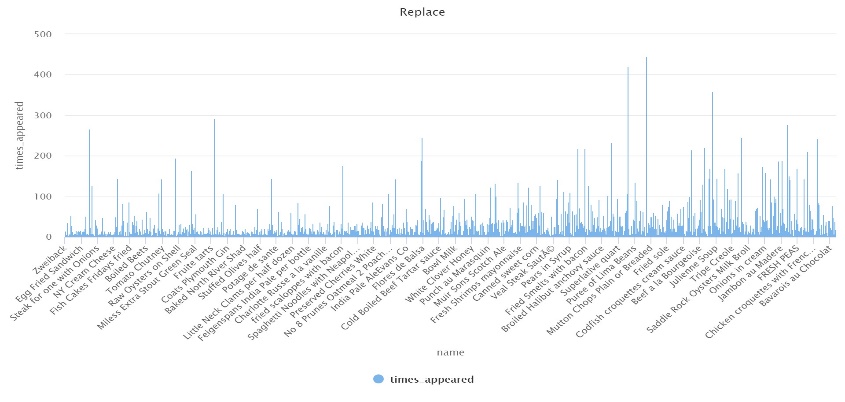
The ten most expensive dishes are as follows in order of decreasing highest cost:

|  |  |
| --- | --- |
| 1 | Cream Cheese with Barleduc jelly |
| 2 | Grape Fruit |
| 3 | Pommery Granno ex Dry |
| 4 | Coffee Pot |
| 5 | Pol roger |
| 6 | Canvasback duck |
| 7 | Veuve Clicquot Yellow label |
| 8 | Spring Turkey |
| 9 | Moet Chandon |
| 10 | Mallard duck |

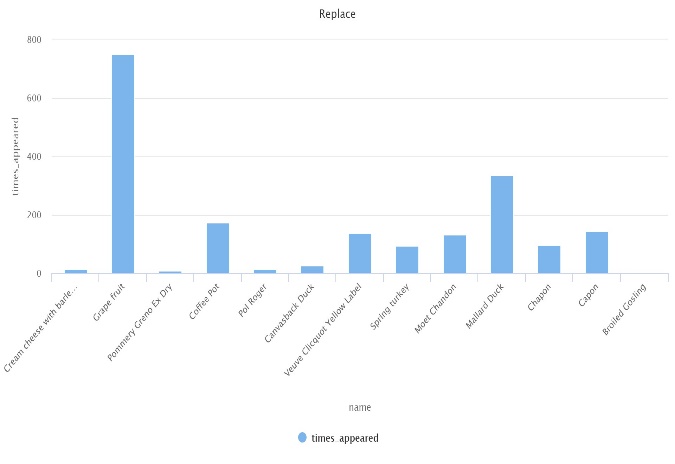
The ten most cheap dishes in order of decreasing cheapness are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | |  | | --- | | zweiback | |  | |
| 2 | vienna horns |
| 3 | rock cakes |
| 4 | iced cakes vanilla |
| 5 | biscuit sandwiches |
| 6 | |  | | --- | | wm youngers scotch ale | |
| 7 | salmon roll sandwich |
| 8 | salmon roll |
| 9 | manilla cigars |
| 10 | ice cold glass of milk |

Bar Plot of cheap items with the number of times they appeared. Dishes with same popularity are included for a reference. The graphs are also submitted in the zip file or can be expanded.



Bar Plot of most expensive items



**Task 2:**

The task is divided into following tasks:

**Task 2a:**

**The following are numerical attributes:**

Age

Industry Code

Occupation Code

Education

Full time or part time employment status

Capital gains

Capital Loses

Live in this house for more than one year

Country of birth mother

Citizen ship

Total Person Income

Own business or self employed

Detailed household and family and family income

**The following are categorical attributes**:

Class of worker

|  |  |  |
| --- | --- | --- |
| 1 | Not in universe | 4881 |
| 2 | Private | 3731 |
| 3 | Self-employed-not incorporated | 409 |
| 4 | Local government | 407 |
| 5 | State government | 245 |
| 6 | Federal government | 153 |
| 7 | Self-employed-incorporated | 148 |
| 8 | Never worked | 17 |
| 9 | Without pay | 9 |

**Adjusted gross income**

|  |  |  |
| --- | --- | --- |
| 1 | High school graduate | 2454 |
| 2 | Children | 2338 |
| 3 | Some college but no degree | 1418 |
| 4 | Bachelors degree(BA AB BS) | 973 |
| 5 | 7th and 8th grade | 456 |
| 6 | 10th grade | 384 |
| 7 | Masters degree(MA MS MEng MEd MSW MBA) | 347 |
| 8 | 9th grade | 322 |
| 9 | 11th grade | 318 |
| 10 | Associates degree-occup /vocational | 261 |
| 11 | Associates degree-academic program | 218 |
| 12 | 5th or 6th grade | 135 |
| 13 | 1st 2nd 3rd or 4th grade | 96 |
| 14 | Prof school degree (MD DDS DVM LLB JD) | 88 |
| 15 | 12th grade no diploma | 84 |
| 16 | Doctorate degree(PhD EdD) | 69 |

**Wage per hour**

|  |  |  |
| --- | --- | --- |
| 1 | Not in universe | 9356 |
| 2 | High school | 358 |
| 3 | College or university | 286 |

**Enrolled in edu last week**

|  |  |  |
| --- | --- | --- |
| 1 | Never married | 4307 |
| 2 | Married-civilian spouse present | 4274 |
| 3 | Divorced | 653 |
| 4 | Widowed | 510 |
| 5 | Separated | 148 |
| 6 | Married-spouse absent | 80 |
| 7 | Married-A F spouse present | 28 |

**Marital status**

|  |  |  |
| --- | --- | --- |
| 1 | Not in universe or children | 4898 |
| 2 | Retail trade | 891 |
| 3 | Education | 451 |
| 4 | Manufacturing-durable goods | 451 |
| 5 | Manufacturing-nondurable goods | 385 |
| 6 | Finance insurance and real estate | 331 |
| 7 | Construction | 308 |
| 8 | Business and repair services | 268 |
| 9 | Medical except hospital | 246 |
| 10 | Public administration | 235 |
| 11 | Other professional services | 218 |
| 12 | Transportation | 218 |
| 13 | Wholesale trade | 192 |
| 14 | Hospital services | 187 |
| 15 | Agriculture | 154 |
| 16 | Personal services except private HH | 147 |
| 17 | Social services | 125 |
| 18 | Entertainment | 90 |
| 19 | Private household services | 59 |
| 20 | Utilities and sanitary services | 56 |
| 21 | Communications | 55 |
| 22 | Mining | 25 |
| 23 | Forestry and fisheries | 8 |
| 24 | Armed Forces | 2 |

**Major industry code**

|  |  |  |
| --- | --- | --- |
| 1 | Not in universe | 4898 |
| 2 | Adm support including clerical | 784 |
| 3 | Professional specialty | 700 |
| 4 | Executive admin and managerial | 638 |
| 5 | Other service | 628 |
| 6 | Sales | 627 |
| 7 | Precision production craft & repair | 524 |
| 8 | Machine operators assmblrs & inspctrs | 330 |
| 9 | Handlers equip cleaners etc | 220 |
| 10 | Transportation and material moving | 202 |
| 11 | Technicians and related support | 166 |
| 12 | Farming forestry and fishing | 165 |
| 13 | Protective services | 65 |
| 14 | Private household services | 51 |
| 15 | Armed Forces | 2 |

**Major occupation code**

|  |  |  |
| --- | --- | --- |
| 1 | White | 8363 |
| 2 | Black | 1027 |
| 3 | Asian or Pacific Islander | 309 |
| 4 | Other | 183 |
| 5 | Amer Indian Aleut or Eskimo | 118 |

**Mace**

|  |  |  |
| --- | --- | --- |
| 1 | All other | 8672 |
| 2 | Mexican-American | 401 |
| 3 | Mexican (Mexicano) | 362 |
| 4 | Central or South American | 195 |
| 5 | Puerto Rican | 144 |
| 6 | Other Spanish | 106 |
| 7 | Cuban | 61 |
| 8 | NA | 31 |
| 9 | Do not know | 16 |
| 10 | Chicano | 12 |

**Hispanic origin**

|  |  |  |
| --- | --- | --- |
| 1 | Female | 5312 |
| 2 | Male | 4688 |

**Sex**

|  |  |  |
| --- | --- | --- |
| 1 | Not in universe | 8996 |
| 2 | No | 857 |
| 3 | Yes | 147 |

**Member of labor union**

|  |  |  |
| --- | --- | --- |
| 1 | Not in universe | 9728 |
| 2 | Other job loser | 97 |
| 3 | Re-entrant | 86 |
| 4 | Job loser - on layoff | 42 |
| 5 | Job leaver | 30 |
| 6 | New entrant | 17 |

**Reason for unemployment**

|  |  |  |
| --- | --- | --- |
| 1 | Children or Armed Forces | 6156 |
| 2 | Full-time schedules | 2097 |
| 3 | Not in labor force | 1314 |
| 4 | PT for non-econ reasons usually FT | 184 |
| 5 | Unemployed full-time | 119 |
| 6 | PT for econ reasons usually PT | 72 |
| 7 | Unemployed part- time | 35 |
| 8 | PT for econ reasons usually FT | 23 |

**Dividend from stocks**

|  |  |  |
| --- | --- | --- |
| 1 | Nonfiler | 3676 |
| 2 | Joint both under 65 | 3478 |
| 3 | Single | 1888 |
| 4 | Joint both 65+ | 397 |
| 5 | Head of household | 388 |
| 6 | Joint one under 65 & one 65+ | 173 |

**Federal income tax liability**

|  |  |  |
| --- | --- | --- |
| 1 | Not in universe | 9218 |
| 2 | South | 225 |
| 3 | Midwest | 200 |
| 4 | West | 187 |
| 5 | Northeast | 147 |
| 6 | Abroad | 23 |

**Tax filer status**

|  |  |  |
| --- | --- | --- |
| 1 | Not in universe | 9218 |
| 2 | California | 68 |
| 3 | Utah | 62 |
| 4 | ? | 46 |
| 5 | North Carolina | 46 |
| 6 | Minnesota | 36 |
| 7 | Oklahoma | 34 |
| 8 | Abroad | 33 |
| 9 | Florida | 31 |
| 10 | North Dakota | 26 |
| 11 | Kentucky | 24 |
| 12 | New Mexico | 22 |
| 13 | Indiana | 19 |
| 14 | Michigan | 19 |
| 15 | Alaska | 18 |
| 16 | Colorado | 14 |
| 17 | Oregon | 14 |
| 18 | Arizona | 13 |
| 19 | Iowa | 13 |
| 20 | Montana | 13 |
| 21 | Kansas | 12 |
| 22 | New Hampshire | 12 |
| 23 | Ohio | 12 |
| 24 | Pennsylvania | 11 |
| 25 | Tennessee | 11 |
| 26 | Nevada | 10 |
| 27 | Vermont | 10 |
| 28 | Virginia | 10 |
| 29 | Arkansas | 9 |
| 30 | Illinois | 9 |
| 31 | Louisiana | 9 |
| 32 | Missouri | 9 |
| 33 | New York | 9 |
| 34 | Alabama | 8 |
| 35 | Massachusetts | 8 |
| 36 | Mississippi | 8 |
| 37 | West Virginia | 8 |
| 38 | Wyoming | 8 |
| 39 | Georgia | 7 |
| 40 | Connecticut | 6 |
| 41 | South Carolina | 6 |
| 42 | Texas | 6 |
| 43 | Maryland | 5 |
| 44 | District of Columbia | 4 |
| 45 | Maine | 4 |
| 46 | Nebraska | 4 |
| 47 | New Jersey | 4 |
| 48 | South Dakota | 4 |
| 49 | Wisconsin | 4 |

**Region of previous residence**

|  |  |  |
| --- | --- | --- |
| 1 | Householder | 2689 |
| 2 | Child | 2511 |
| 3 | Spouse of householder | 2107 |
| 4 | Nonfamily householder | 1104 |
| 5 | Child 18+ never marr Not in a subfamily | 575 |
| 6 | Secondary individual | 317 |
| 7 | Other Rel 18+ ever marr not in subfamily | 95 |
| 8 | Grandchild | 82 |
| 9 | Other Rel 18+ never marr not in subfamily | 81 |
| 10 | Child 18+ ever marr Not in a subfamily | 60 |
| 11 | Grandchild | 53 |
| 12 | Child under 18 of RP of unrel subfamily | 44 |
| 13 | Other Rel | 40 |
| 14 | Child 18+ never marr RP of subfamily | 34 |
| 15 | Other Rel 18+ spouse of subfamily RP | 33 |
| 16 | Child 18+ ever marr RP of subfamily | 31 |
| 17 | RP of unrelated subfamily | 31 |
| 18 | Other Rel | 30 |
| 19 | Other Rel 18+ ever marr RP of subfamily | 29 |
| 20 | Grandchild 18+ never marr not in subfamily | 23 |
| 21 | Child 18+ spouse of subfamily RP | 8 |
| 22 | In group quarters | 8 |
| 23 | Child | 7 |
| 24 | Other Rel 18+ never marr RP of subfamily | 3 |
| 25 | Child | 2 |
| 26 | Grandchild 18+ spouse of subfamily RP | 2 |
| 27 | Other Rel | 1 |

**State of previous residence**

|  |  |  |
| --- | --- | --- |
| 1 | Householder | 3795 |
| 2 | Child under 18 never married | 2518 |
| 3 | Spouse of householder | 2107 |
| 4 | Child 18 or older | 708 |
| 5 | Other relative of householder | 472 |
| 6 | Nonrelative of householder | 393 |
| 7 | Group Quarters- Secondary individual | 5 |
| 8 | Child under 18 ever married | 2 |

**Detailed household summary**

|  |  |  |
| --- | --- | --- |
| 1 | ? | 5043 |
| 2 | Nonmover | 4088 |
| 3 | MSA to MSA | 534 |
| 4 | NonMSA to nonMSA | 143 |
| 5 | Not in universe | 87 |
| 6 | MSA to nonMSA | 39 |
| 7 | NonMSA to MSA | 23 |
| 8 | Not identifiable | 20 |
| 9 | Abroad to MSA | 19 |
| 10 | Abroad to nonMSA | 4 |

**Instance Weight**

|  |  |  |
| --- | --- | --- |
| 1 | ? | 5043 |
| 2 | Nonmover | 4088 |
| 3 | Same county | 506 |
| 4 | Different county same state | 133 |
| 5 | Not in universe | 87 |
| 6 | Different region | 61 |
| 7 | Different state same division | 37 |
| 8 | Abroad | 23 |
| 9 | Different division same region | 22 |

**Migration code change in msa**

|  |  |  |
| --- | --- | --- |
| 1 | ? | 5043 |
| 2 | Nonmover | 4088 |
| 3 | Same county | 506 |
| 4 | Different county same state | 133 |
| 5 | Not in universe | 87 |
| 6 | Different state in South | 45 |
| 7 | Different state in West | 29 |
| 8 | Different state in Midwest | 24 |
| 9 | Abroad | 23 |
| 10 | Different state in Northeast | 22 |

**Migration code change in reg**

|  |  |  |
| --- | --- | --- |
| 1 | Not in universe under 1 year old | 5130 |
| 2 | Yes | 4088 |
| 3 | No | 782 |

**Migration code change within reg**

|  |  |  |
| --- | --- | --- |
| 1 | ? | 5043 |
| 2 | Not in universe | 4175 |
| 3 | No | 519 |
| 4 | Yes | 263 |

**Migration prev res in subelt**

|  |  |  |
| --- | --- | --- |
| 1 | Not in universe | 7238 |
| 2 | Both parents present | 1921 |
| 3 | Mother only present | 676 |
| 4 | Neither parent present | 83 |
| 5 | Father only present | 82 |

**Num persons worked for employment**

|  |  |  |
| --- | --- | --- |
| 1 | United-States | 8011 |
| 2 | Mexico | 484 |
| 3 | ? | 299 |
| 4 | Puerto-Rico | 122 |
| 5 | Italy | 103 |
| 6 | Germany | 77 |
| 7 | Dominican-Republic | 70 |
| 8 | Poland | 70 |
| 9 | Canada | 66 |
| 10 | Cuba | 65 |
| 11 | Philippines | 64 |
| 12 | China | 50 |
| 13 | El-Salvador | 43 |
| 14 | England | 41 |
| 15 | Ireland | 33 |
| 16 | Columbia | 32 |
| 17 | Vietnam | 27 |
| 18 | South Korea | 24 |
| 19 | Haiti | 22 |
| 20 | India | 21 |
| 21 | Nicaragua | 20 |
| 22 | Jamaica | 19 |
| 23 | Guatemala | 18 |
| 24 | Ecuador | 17 |
| 25 | Greece | 17 |
| 26 | Peru | 17 |
| 27 | Japan | 16 |
| 28 | France | 15 |
| 29 | Honduras | 14 |
| 30 | Hungary | 14 |
| 31 | Scotland | 14 |
| 32 | Cambodia | 13 |
| 33 | Portugal | 12 |
| 34 | Iran | 11 |
| 35 | Yugoslavia | 11 |
| 36 | Outlying-U S (Guam USVI etc) | 8 |
| 37 | Hong Kong | 7 |
| 38 | Laos | 7 |
| 39 | Taiwan | 7 |
| 40 | Thailand | 7 |
| 41 | Holand-Netherlands | 6 |
| 42 | Trinadad&Tobago | 5 |
| 43 | Panama | 1 |

**Family members under 18**

|  |  |  |
| --- | --- | --- |
| 1 | United-States | 8065 |
| 2 | Mexico | 488 |
| 3 | ? | 283 |
| 4 | Puerto-Rico | 117 |
| 5 | Italy | 87 |
| 6 | Germany | 74 |
| 7 | Philippines | 70 |
| 8 | Cuba | 61 |
| 9 | England | 61 |
| 10 | Dominican-Republic | 57 |
| 11 | Poland | 57 |
| 12 | Canada | 53 |
| 13 | China | 46 |
| 14 | El-Salvador | 46 |
| 15 | Ireland | 37 |
| 16 | Columbia | 31 |
| 17 | Vietnam | 27 |
| 18 | South Korea | 26 |
| 19 | Nicaragua | 22 |
| 20 | Haiti | 20 |
| 21 | India | 20 |
| 22 | Jamaica | 20 |
| 23 | Japan | 20 |
| 24 | Guatemala | 19 |
| 25 | Hungary | 19 |
| 26 | Peru | 18 |
| 27 | Ecuador | 17 |
| 28 | Greece | 16 |
| 29 | Honduras | 15 |
| 30 | Scotland | 13 |
| 31 | France | 12 |
| 32 | Portugal | 12 |
| 33 | Iran | 9 |
| 34 | Yugoslavia | 9 |
| 35 | Cambodia | 8 |
| 36 | Taiwan | 8 |
| 37 | Laos | 7 |
| 38 | Outlying-U S (Guam USVI etc) | 7 |
| 39 | Thailand | 7 |
| 40 | Trinadad&Tobago | 6 |
| 41 | Holand-Netherlands | 5 |

**Total person earnings**

|  |  |  |
| --- | --- | --- |
| 1 | United-States | 8873 |
| 2 | Mexico | 278 |
| 3 | ? | 153 |
| 4 | Puerto-Rico | 62 |
| 5 | Germany | 51 |
| 6 | Cuba | 46 |
| 7 | Philippines | 43 |
| 8 | Dominican-Republic | 39 |
| 9 | El-Salvador | 33 |
| 10 | Poland | 31 |
| 11 | England | 30 |
| 12 | China | 29 |
| 13 | Italy | 25 |
| 14 | Vietnam | 24 |
| 15 | Canada | 21 |
| 16 | Columbia | 20 |
| 17 | India | 17 |
| 18 | Jamaica | 17 |
| 19 | Nicaragua | 17 |
| 20 | South Korea | 17 |
| 21 | Haiti | 16 |
| 22 | Ecuador | 13 |
| 23 | Guatemala | 12 |
| 24 | Greece | 11 |
| 25 | Honduras | 11 |
| 26 | Japan | 11 |
| 27 | Peru | 11 |
| 28 | Taiwan | 11 |
| 29 | Ireland | 10 |
| 30 | Portugal | 8 |
| 31 | Thailand | 7 |
| 32 | Cambodia | 6 |
| 33 | Hungary | 6 |
| 34 | Iran | 6 |
| 35 | France | 5 |
| 36 | Hong Kong | 5 |
| 37 | Scotland | 5 |
| 38 | Trinadad&Tobago | 5 |
| 39 | Laos | 4 |
| 40 | Outlying-U S (Guam USVI etc) | 4 |
| 41 | Panama | 3 |
| 42 | Holand-Netherlands | 2 |

**Country of birth father**

|  |  |  |
| --- | --- | --- |
| 1 | Native- Born in the United States | 8873 |
| 2 | Foreign born- Not a citizen of U S | 675 |
| 3 | Foreign born- U S citizen by naturalization | 298 |
| 4 | Native- Born abroad of American Parent(s) | 88 |
| 5 | Native- Born in Puerto Rico or U S Outlying | 66 |

**Country of birth self**

|  |  |  |
| --- | --- | --- |
| 1 | Not in universe | 9895 |
| 2 | No | 84 |
| 3 | Yes | 21 |

**Taxable income amount**

|  |  |  |
| --- | --- | --- |
| 1 | - 50000. | 9413 |
| 2 | 50000+. | 587 |

**Fill inc questionnaire for vetran**

Missing values

**Veteran Benefits**

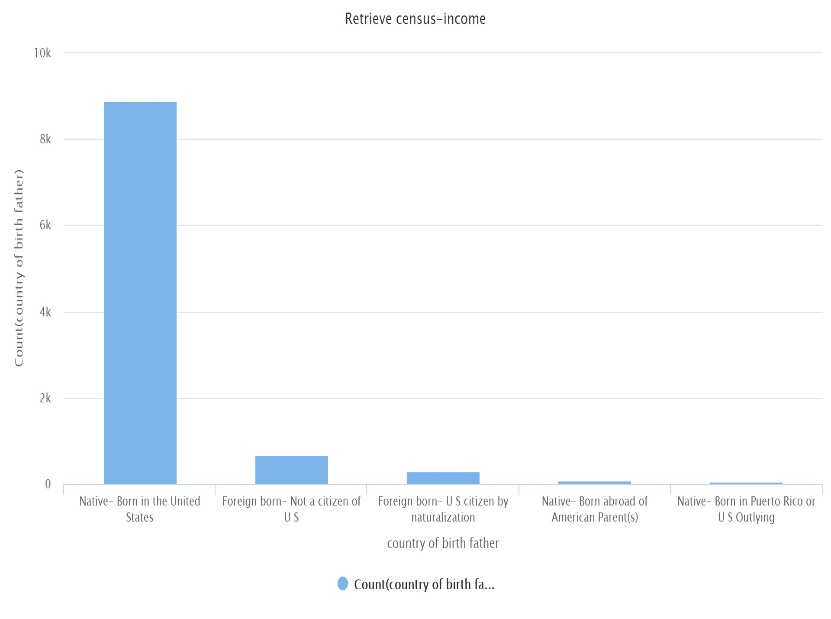
Missing values

**Weeks workd in a year**

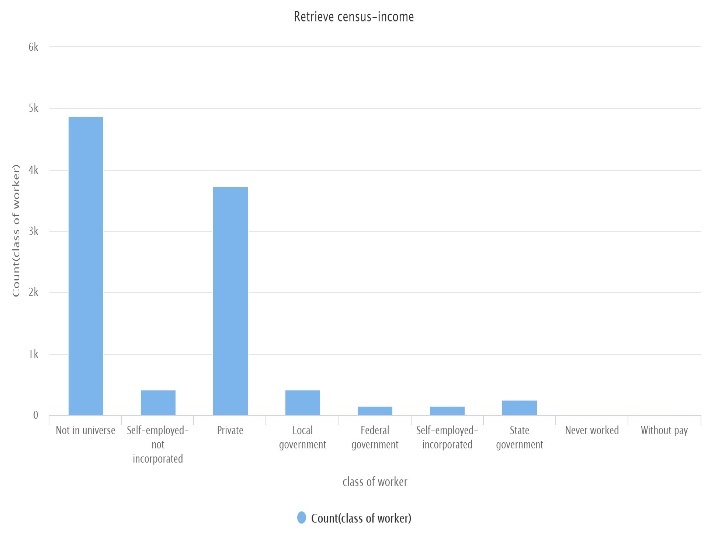
Missing values

**Task 2b:**

**Bar chart for Country of birth father**



**Bar Chart for class of workers**



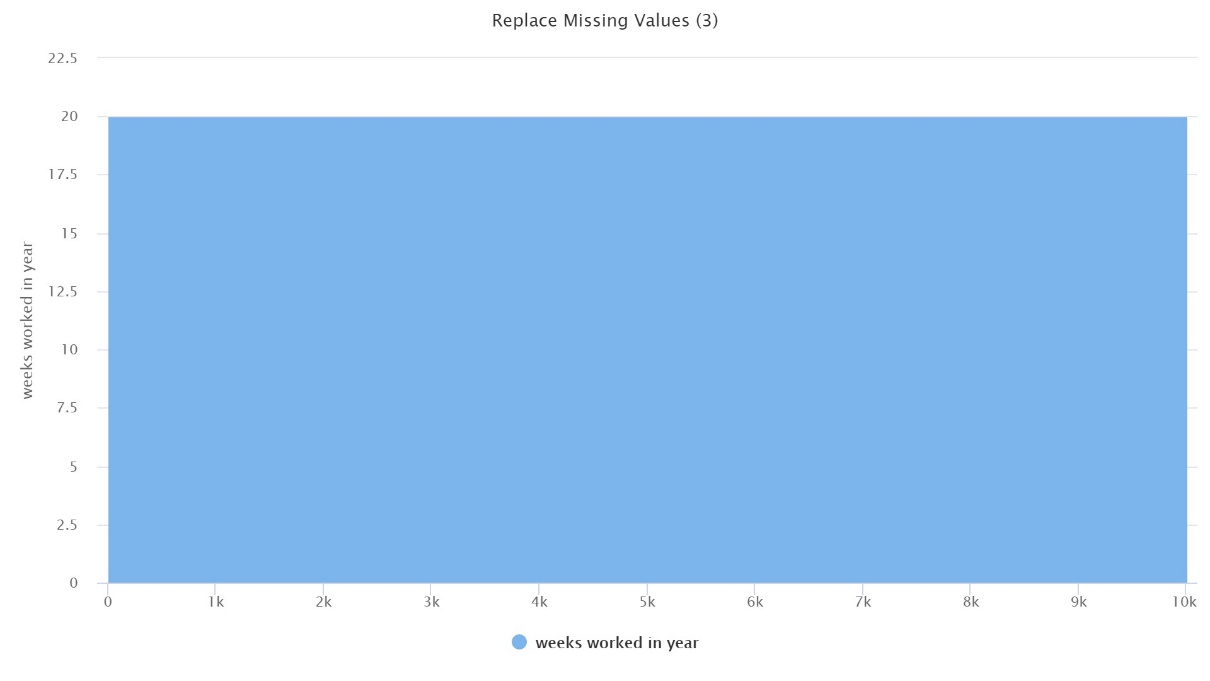
**Task 2c:**

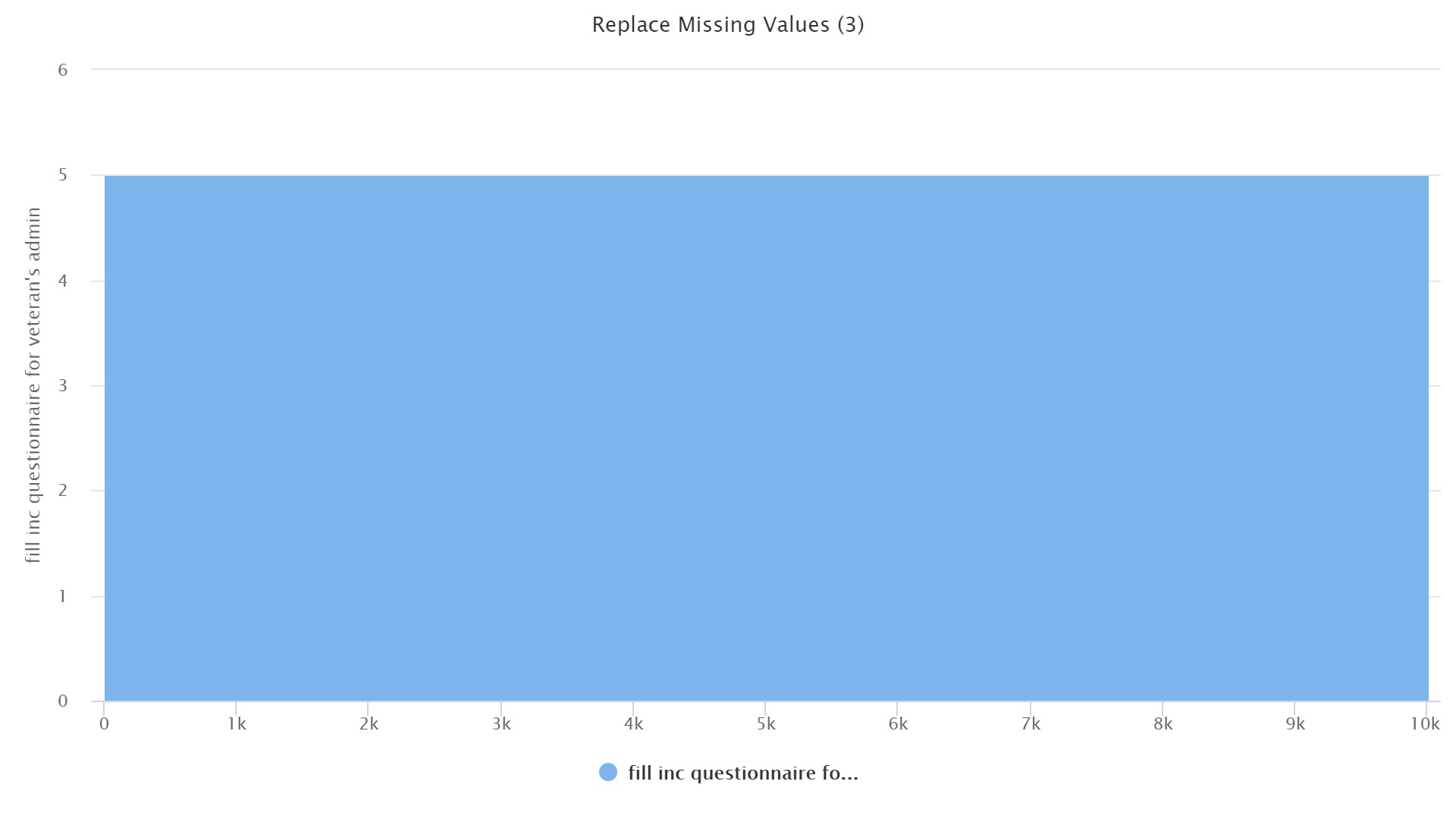
After replacing and converting the values to numeric the values of three attributes changed only:

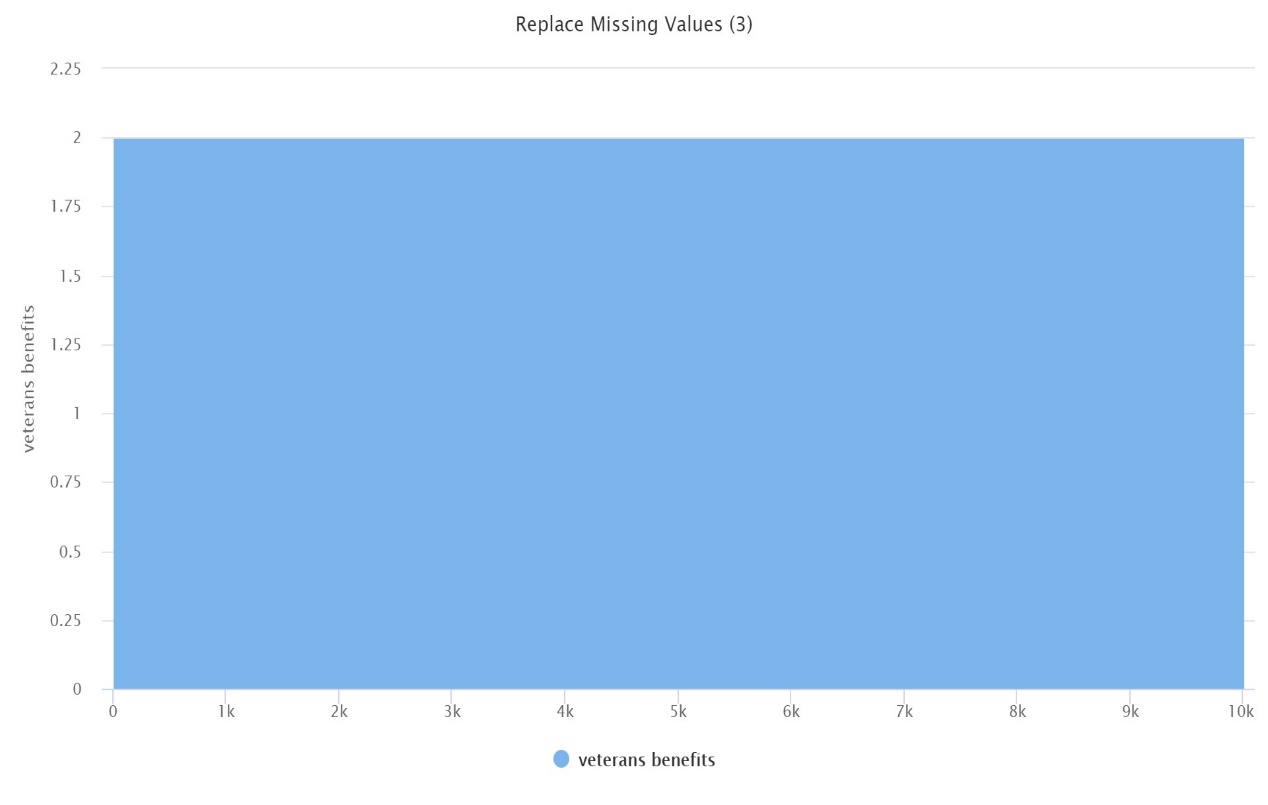
Vetran Benefits: 2

Weeks works per year:20

Fill inc questionnaire:5



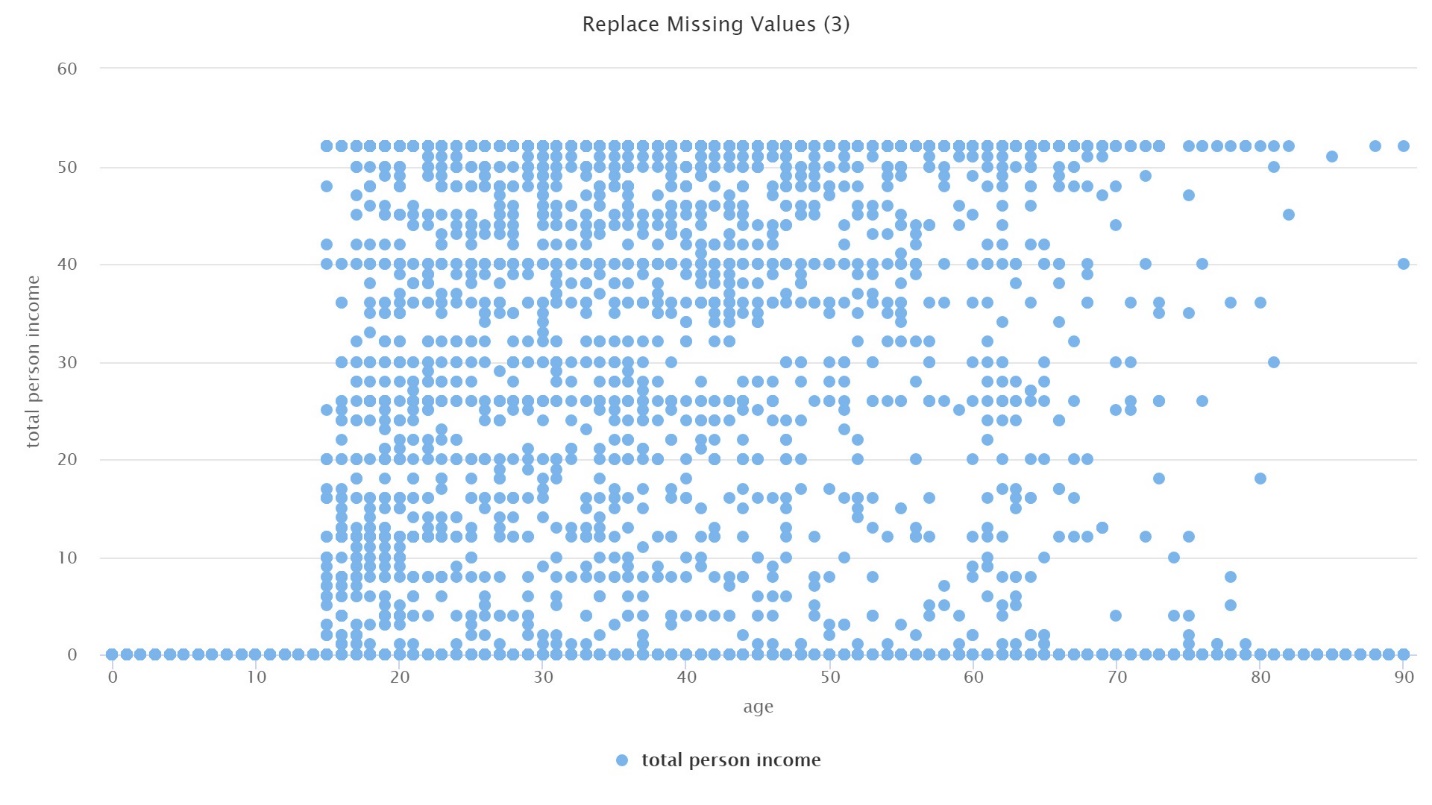


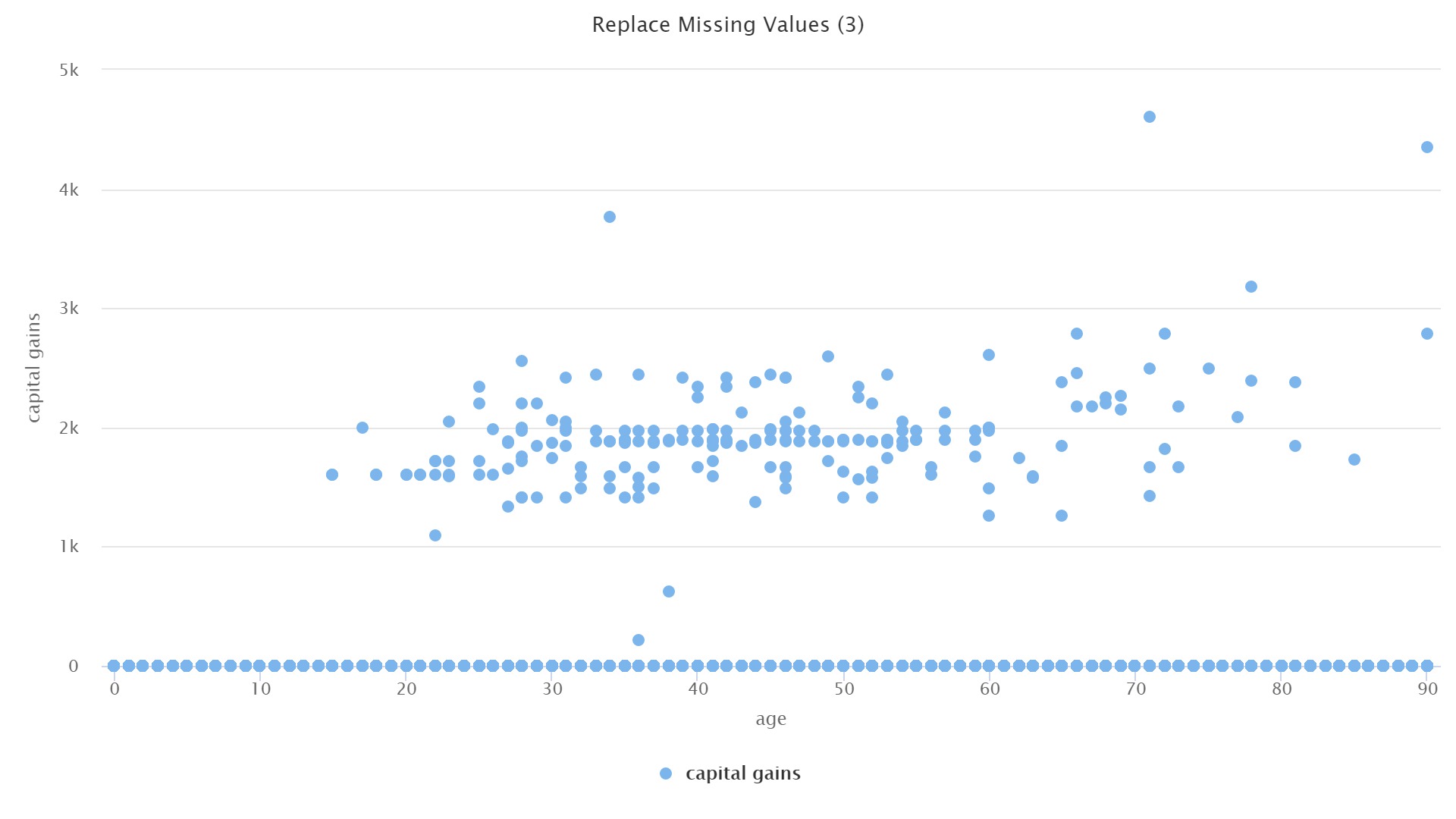
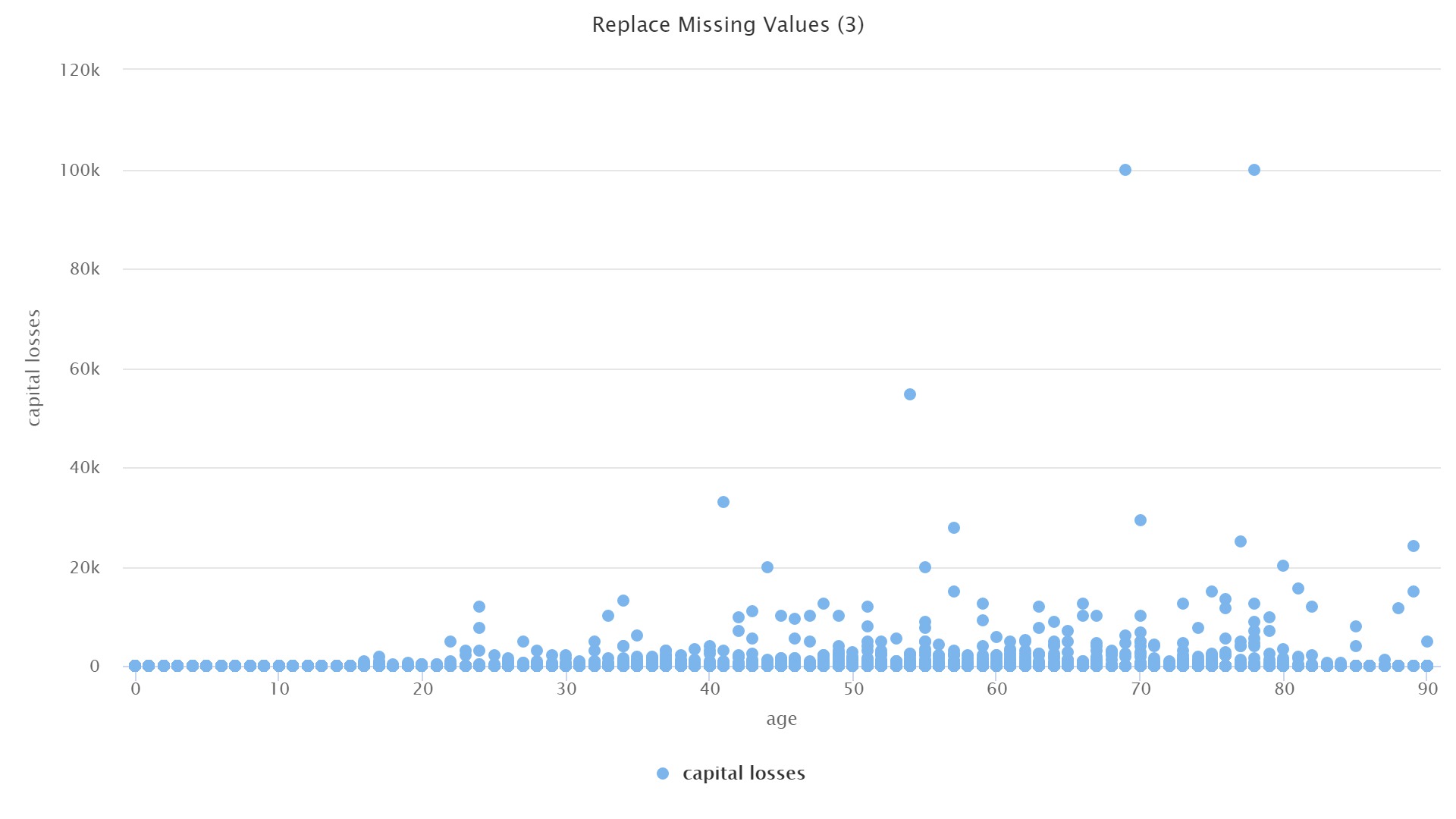
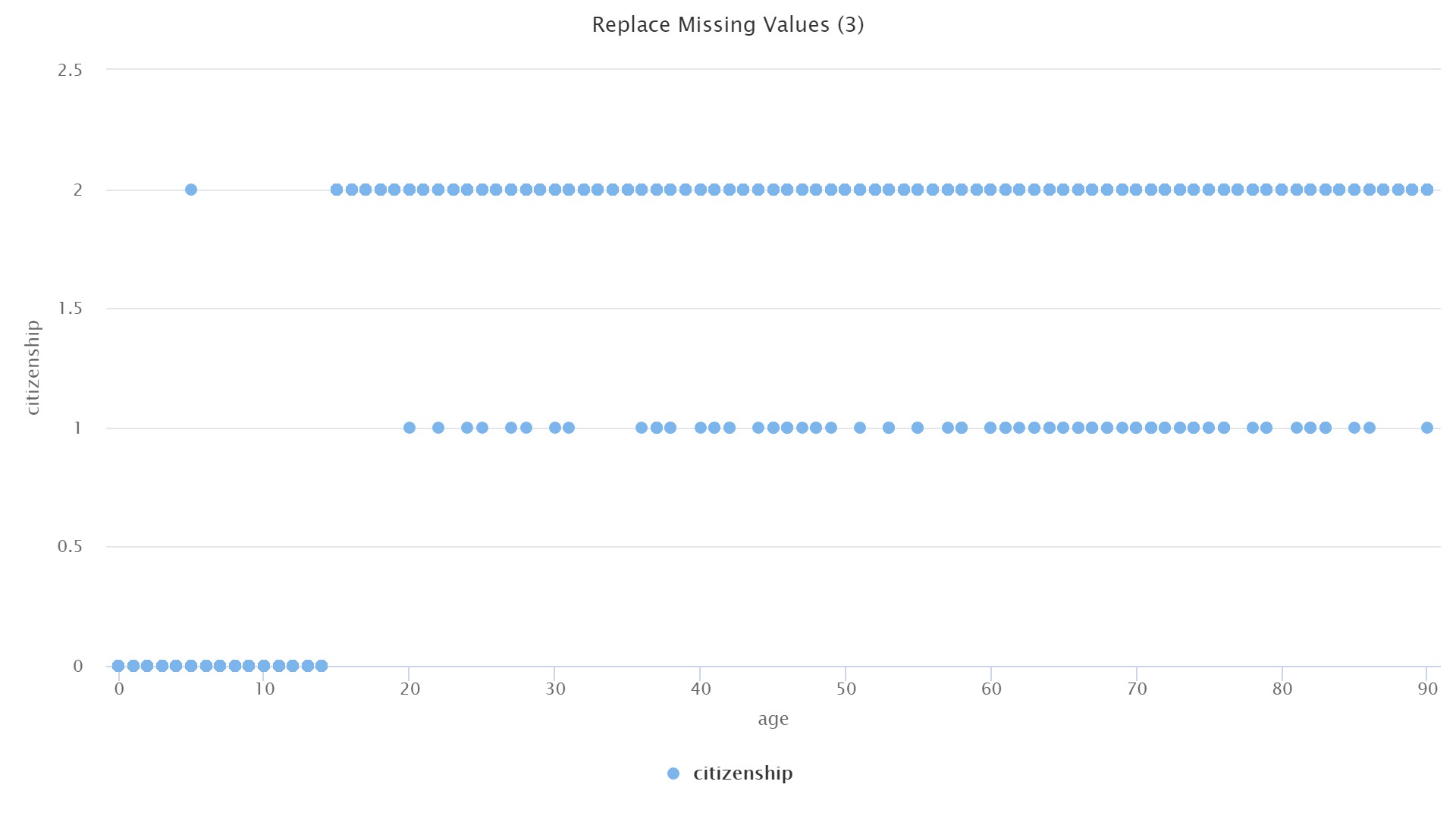
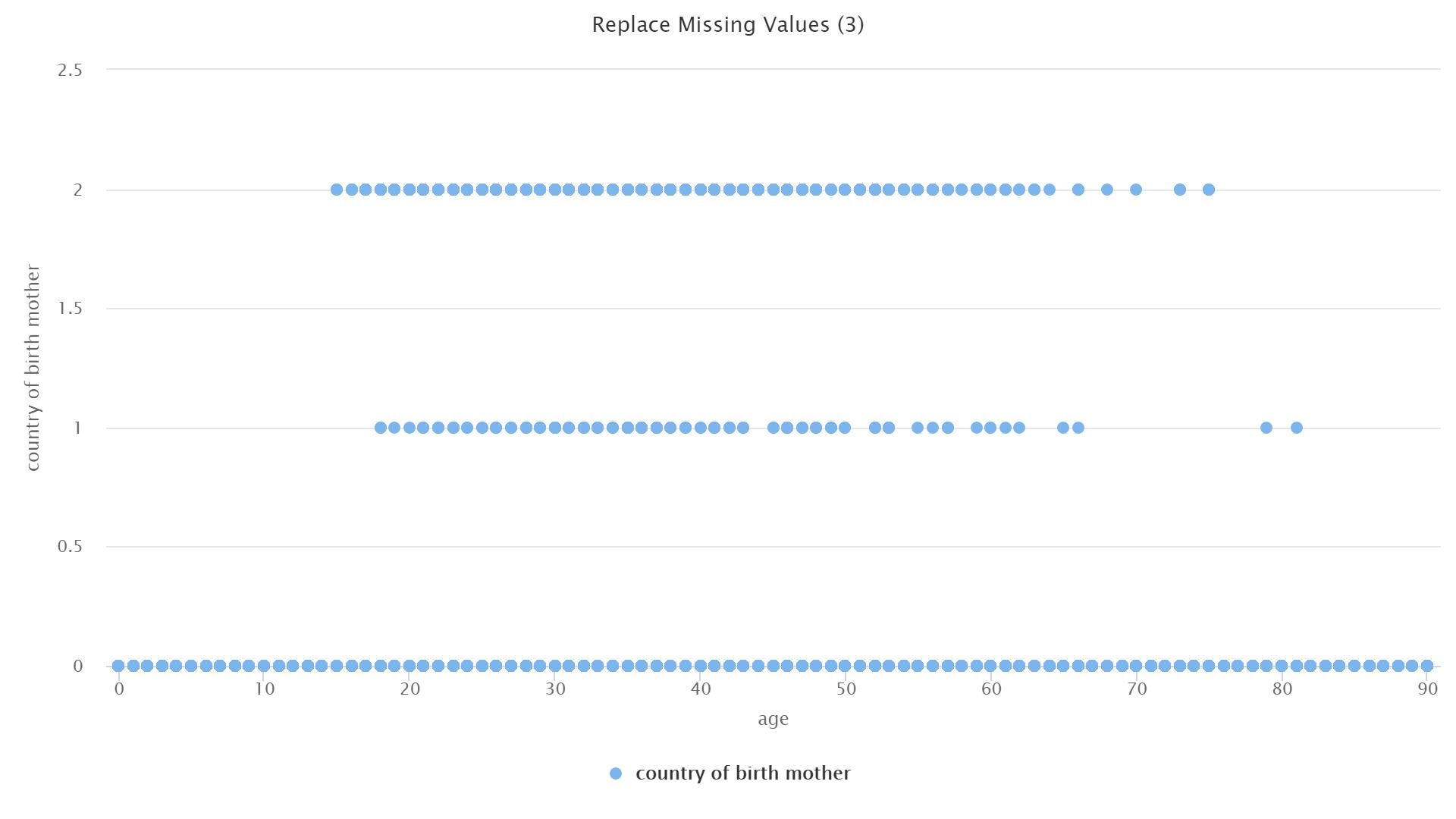
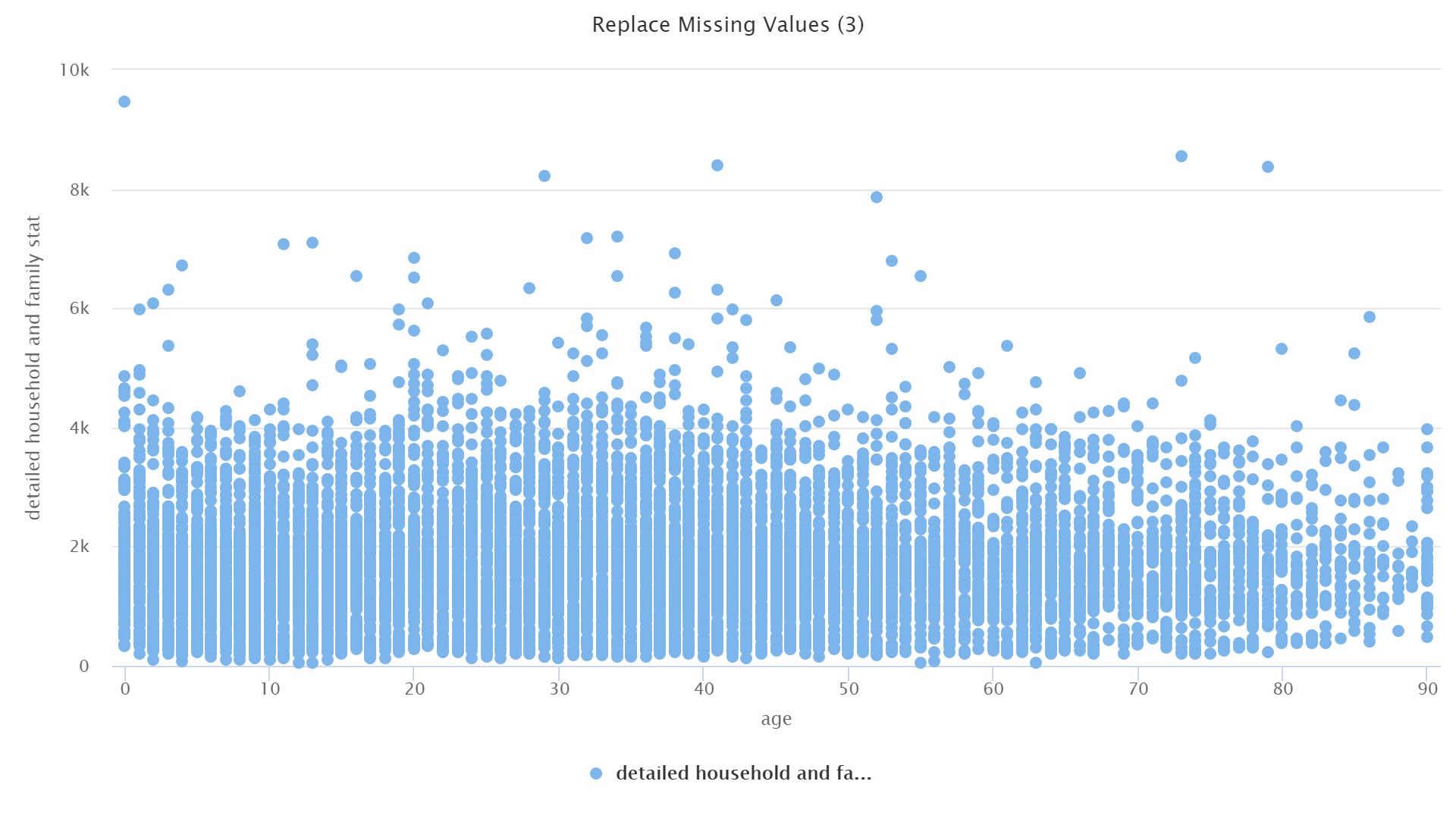
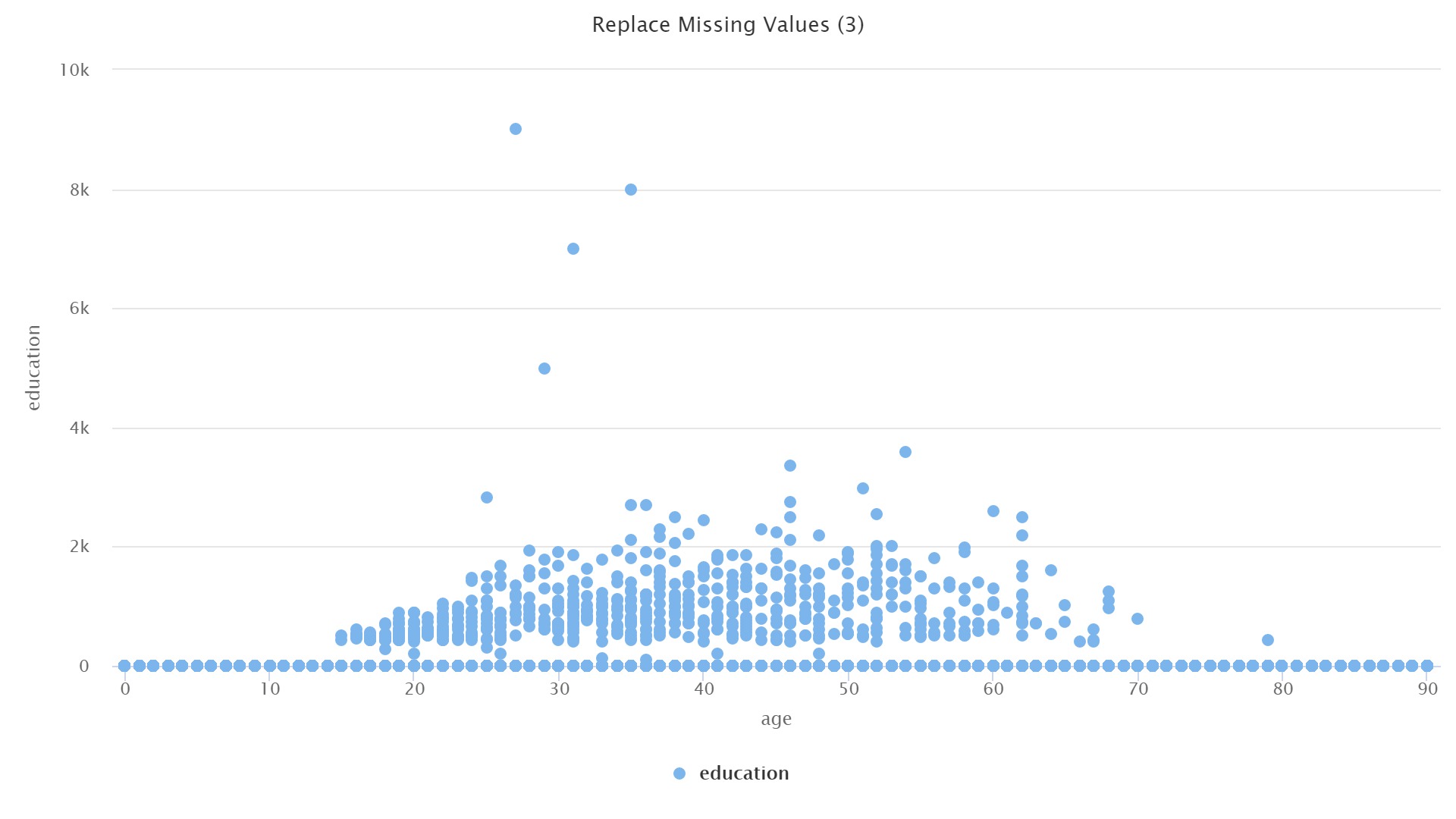
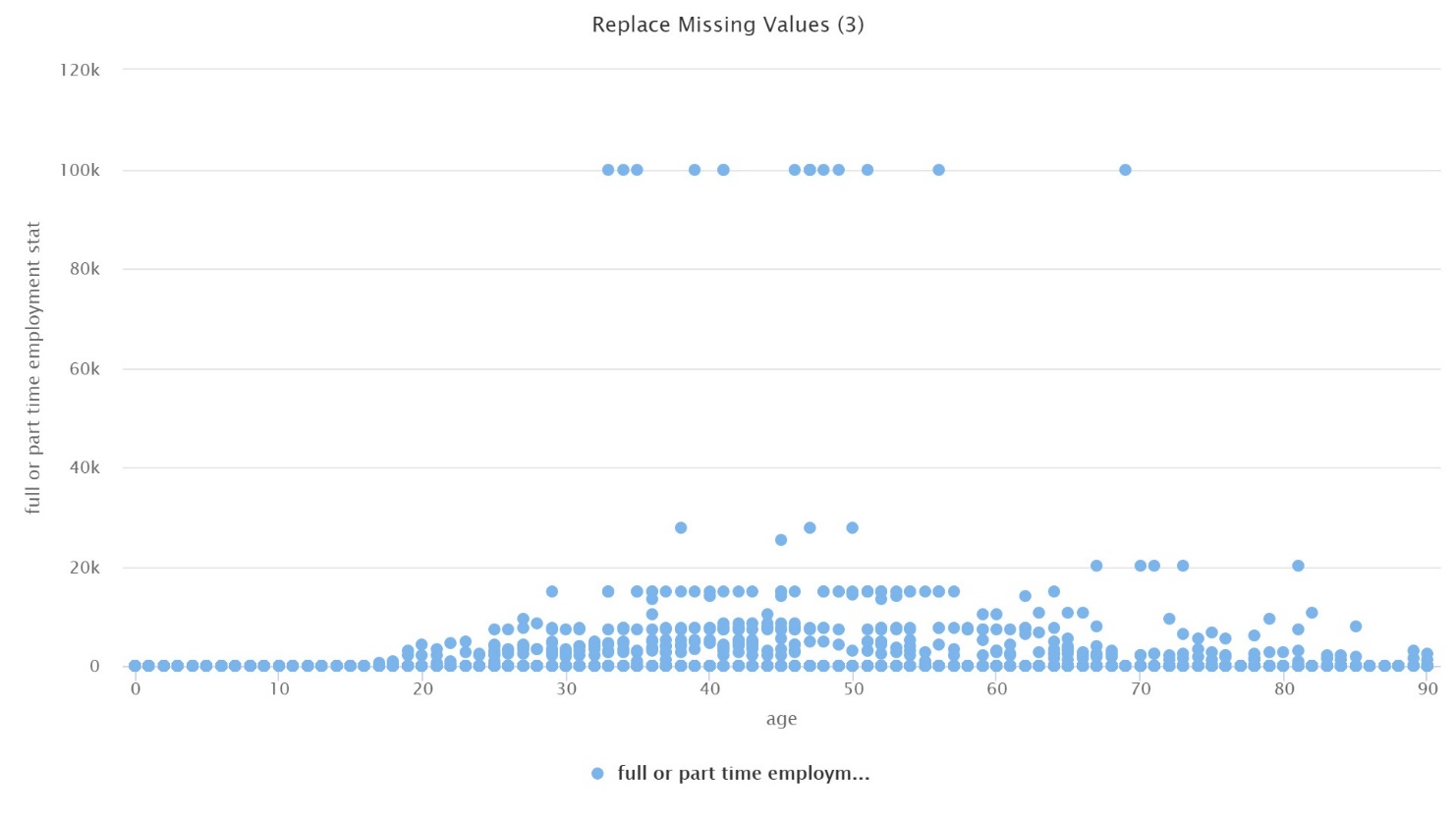
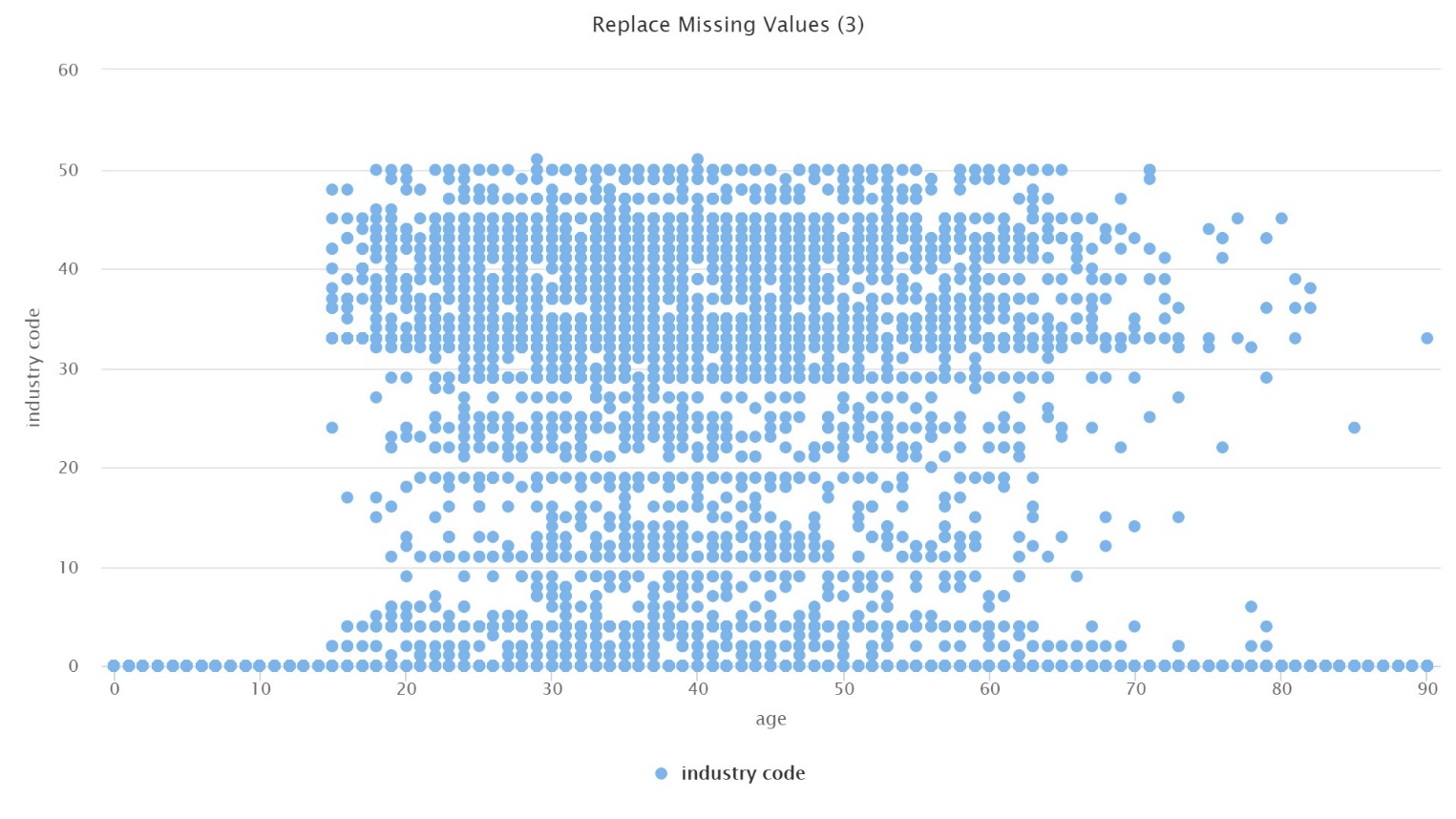
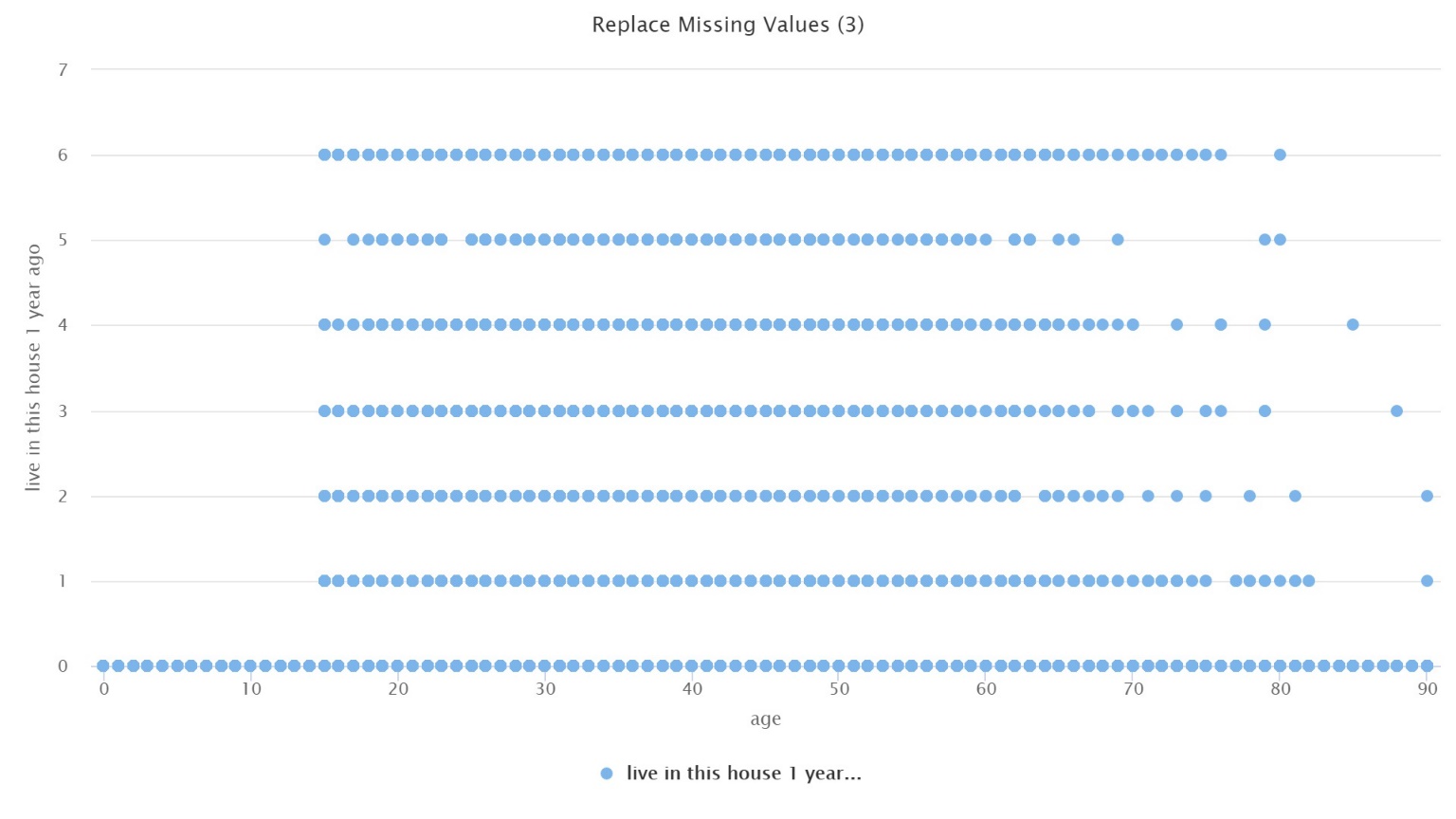
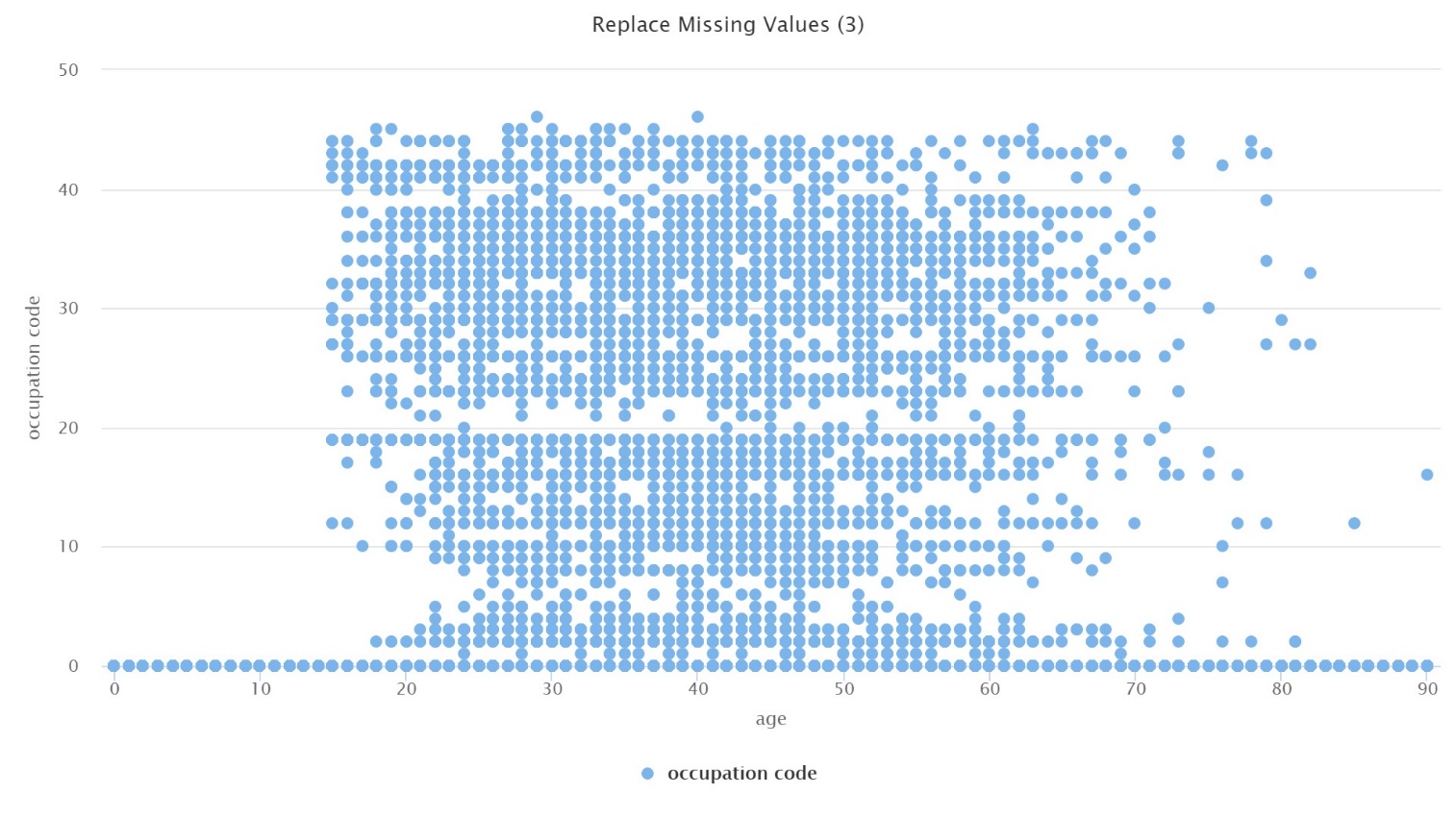
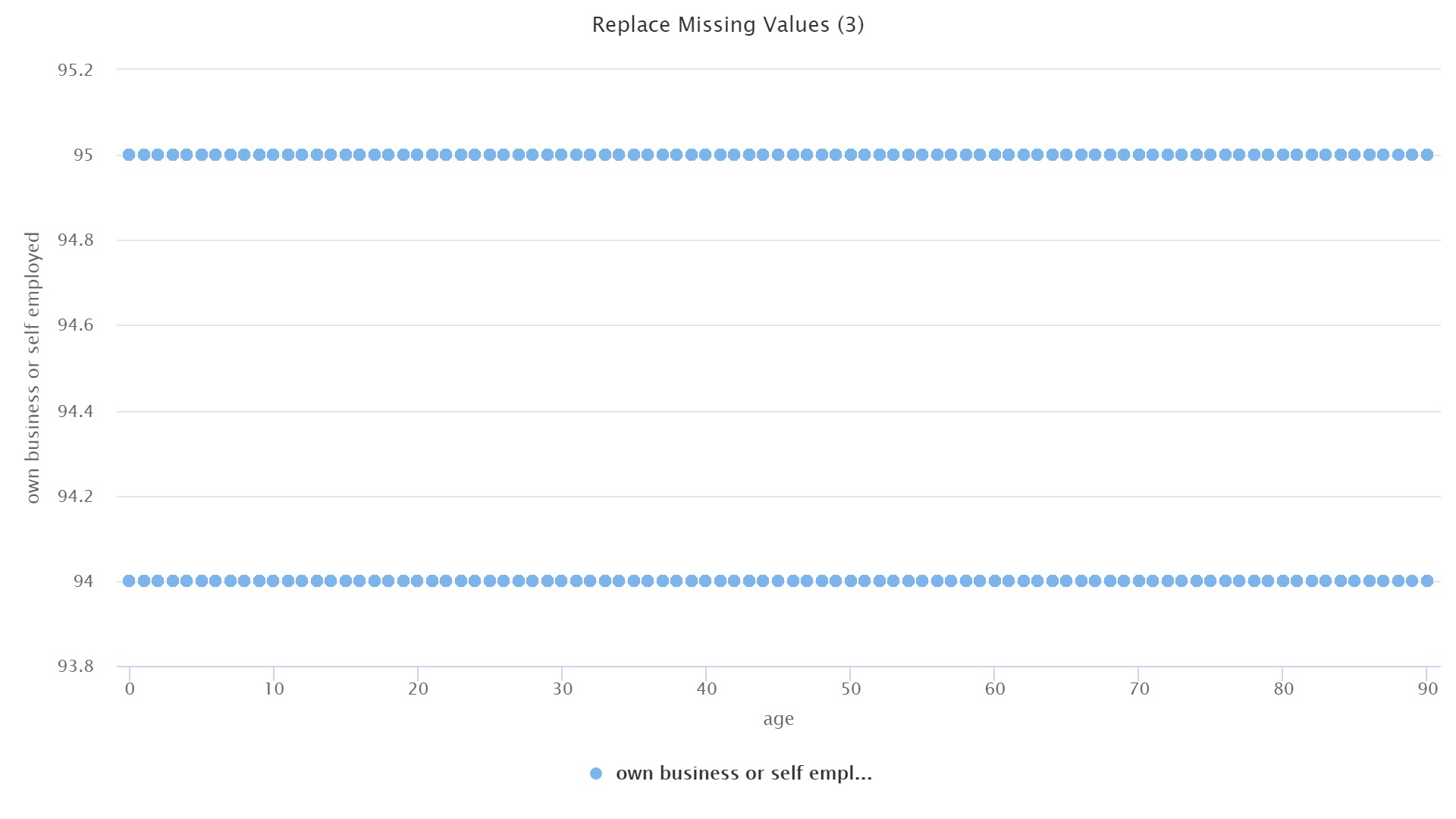


**Task 2d:**

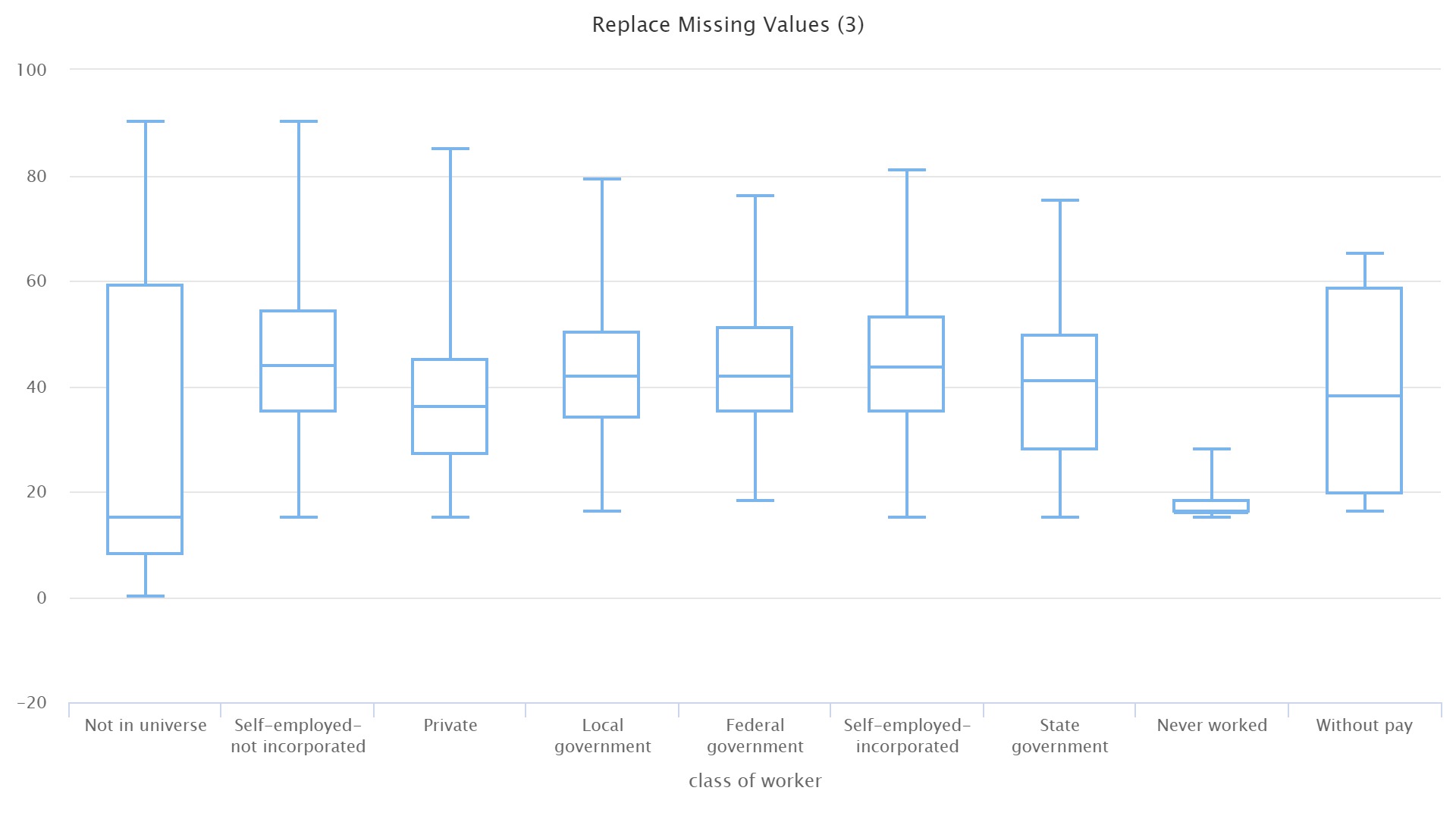
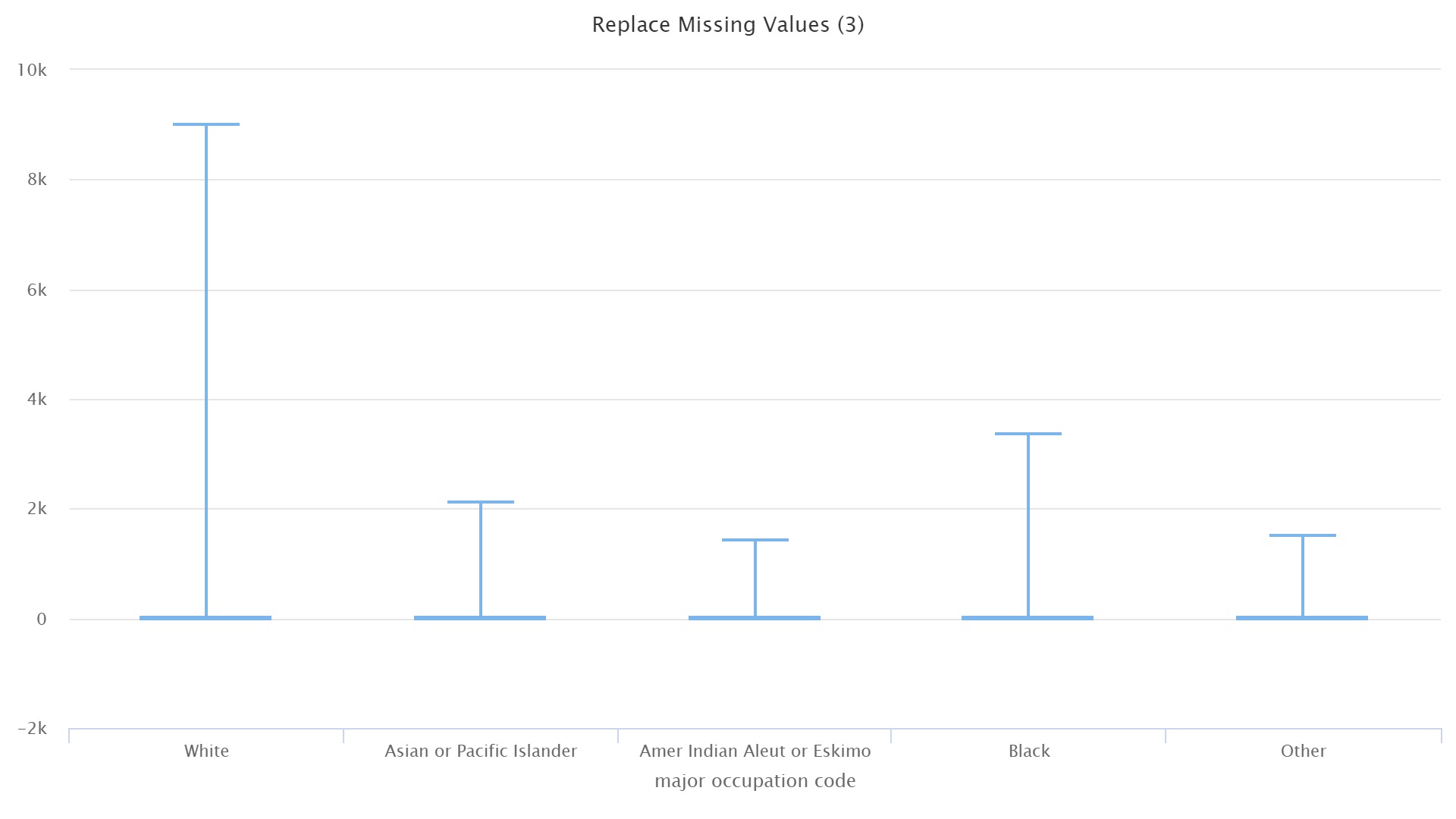
The covariance matrix is in covariance matrix in the zip file.

**Task 2e:**



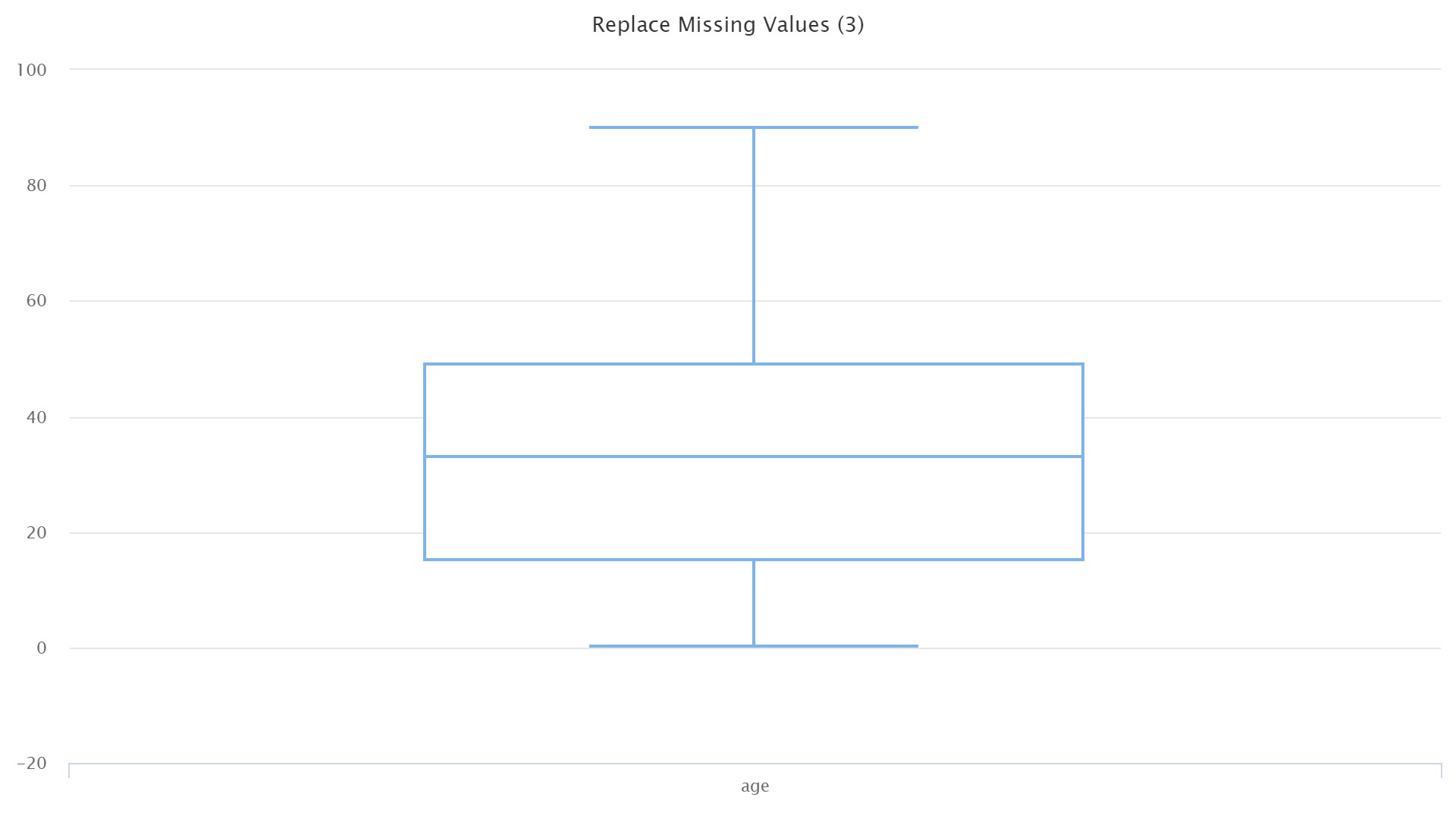


**Task 2f:**

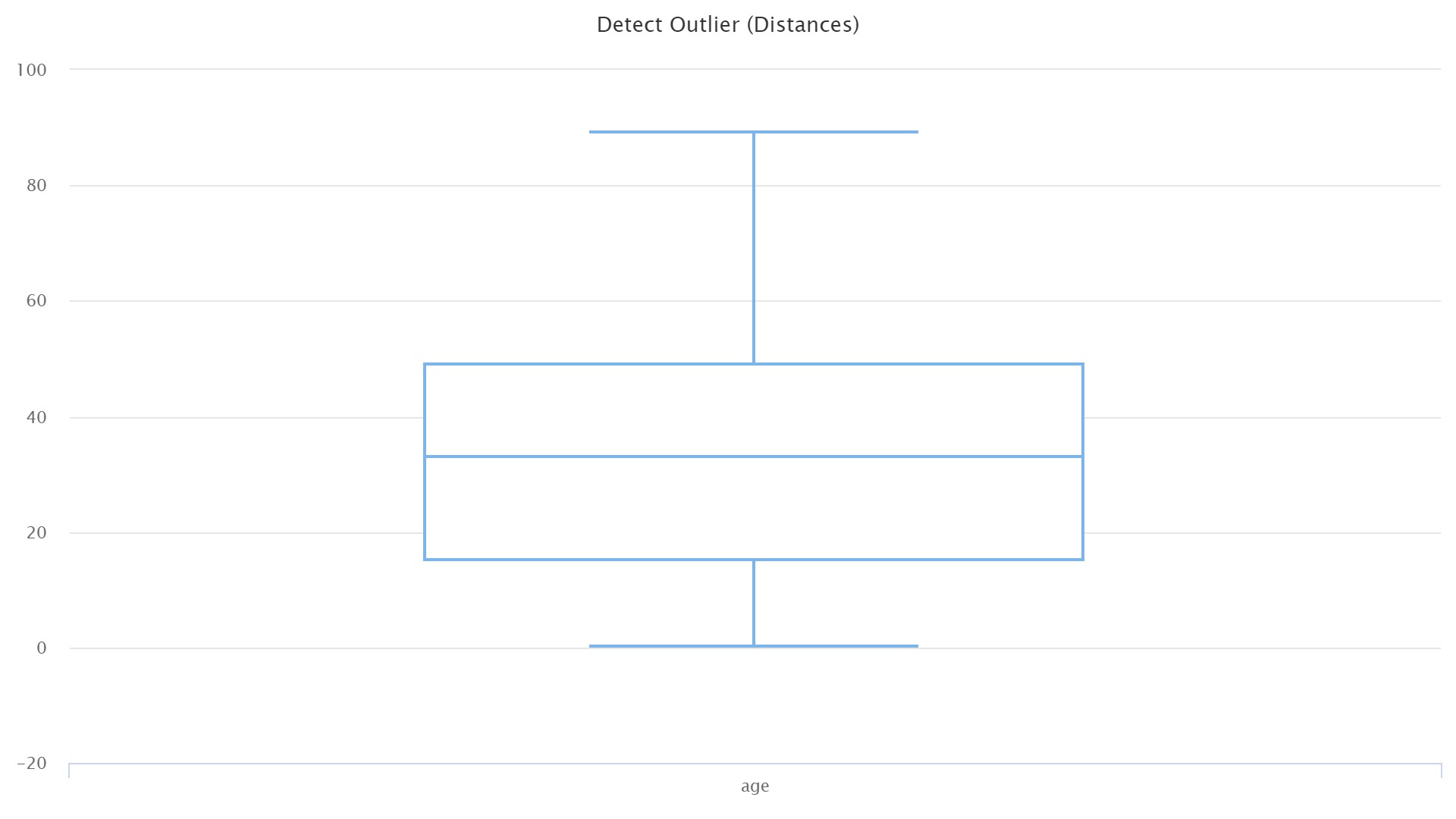


**Task 2g:**

Without removing missing values



After removing outliers:



**Task 3**

The negative values in correlation matrix generated by R code represent an inverse relationship while positive values in the matrix represent a direct relationship. Zero value means that there is no correlation in the variables being studied. The more the value is positive the clearer the relationship can be seen and vice versa. This is also true for inverse relations.

**Plots:**

The plots can be generated by running the R code.

In the first plot the number of one’s is higher than 150 while the number of zeroes is less than 150. This code is run by generating the second block of code in R script.

In the second plot it can be seen easily that the distribution is quite skewed. The number of zeroes is quite higher than that of number of zeroes in the plot of sex.

Due to skewed distribution, the mean, median and mode are also not same as compared to that of a standard normal distribution.