BAN 110 – ZAA Data Preparation and Handling

Final project report – group 6

2022

Table of Contents

[Dataset and task description 2](#_Toc100931108)

[Load Dataset 3](#_Toc100931109)

[Dataset Characteristics 4](#_Toc100931110)

[Categorical Variables 4](#_Toc100931111)

[Interpretation: 11](#_Toc100931112)

[Numerical Variables 12](#_Toc100931113)

[Summary Statistics 12](#_Toc100931114)

[Shape of Distribution: Histograms 12](#_Toc100931115)

[Data Handling: Categorical Variables 16](#_Toc100931116)

[Step 1: Separating categorical variables 16](#_Toc100931117)

[Step 2: Running the freq procedure on the categorical variables 16](#_Toc100931118)

[Step3: Data Cleaning 22](#_Toc100931119)

[Step 4: Creating derived variables 23](#_Toc100931120)

[Data Handling: Numerical Variables 25](#_Toc100931121)

[Error Checking: Out of Range and Missing Values 25](#_Toc100931122)

[Checking for errors using pre-determined range 25](#_Toc100931123)

[Correcting Out of Range Values by Deletion 26](#_Toc100931124)

[Deleting Missing Values for Age 26](#_Toc100931125)

[Detecting Outliers 27](#_Toc100931126)

[Tests for Normality and Distribution 31](#_Toc100931127)

# Dataset and task description

This data was extracted from the census bureau database found at:

<http://www.census.gov/ftp/pub/DES/www/welcome.html>

The dataset is to Predict whether income exceeds $50K/yr based on census data. It is also known as "Adult" dataset.

The dataset contains the age, sex, race, relationship, marital status, Education, Workclass, Occupation, Capital\_gain, Capital\_Loss, Hour\_Week, Native\_Country and Annual Income of each adult.

The data Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions:

((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 48842 | **Area:** | Social |
| **Attribute Characteristics:** | Categorical, Integer | **Number of Attributes:** | 14 | **Date Donated** | 1996-05-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | Yes | **Number of Web Hits:** | 673712 |

The dataset contains 13 attributes and 32562 instances out of which 2399 is missing or unknown. The description and value of each variable is given below. The dependent variable is annual income.

|  |  |  |
| --- | --- | --- |
| **VARIABLE NAME** | **DESCRIPTION** | **TYPE** |
| Age | Age of each adult | Numerical |
| Workclass | The job sectors of each person | Characteristic |
| fnlwgt | Final weight refers to number of people with similar demographic characteristics. | Numerical |
| Education | The education level of each adult | Characteristic |
| Marital Status | The marital status of each person | Characteristic |
| Occupation | The type of occupation of each adult | Characteristic |
| Relationship | The relationship status of each person | Characteristic |
| Race | The race of each person | Characteristic |
| Sex | The gender of each adult | Characteristic |
| Capital\_gain | The increase in a capital asset's value and is realized when the asset is sold. | Numerical |
| Capital\_Loss | A capital property for less than its adjusted cost base plus the outlays and expenses involved in selling the property. | Numerical |
| Hour\_Week | The number of hour that group is working per week | Numerical |
| Native\_Country | The native of each person | Characteristic |
| Income | The annual income | Numerical |

# Load Dataset

SAS Code:

Graphical user interface, text, application, email

Description automatically generated

Result

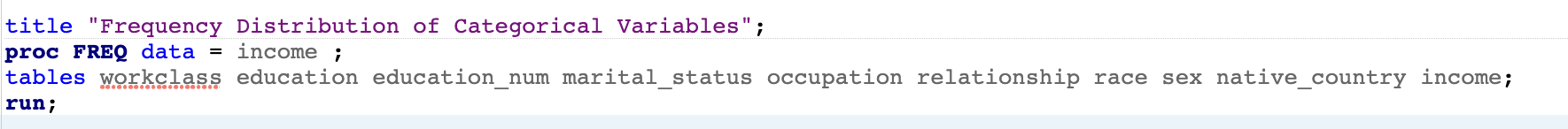
Graphical user interface, application, email

Description automatically generated

# Dataset Characteristics

## Categorical Variables

SAS Code:



Result:

**Frequency Distribution of Categorical Variables**

**The FREQ Procedure**

| **workclass** | | | | |
| --- | --- | --- | --- | --- |
| **workclass** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Frequency Missing = 1836** | | | | |
| **Federal-gov** | 960 | 3.12 | 960 | 3.12 |
| **Local-gov** | 2093 | 6.81 | 3053 | 9.94 |
| **Never-worked** | 7 | 0.02 | 3060 | 9.96 |
| **Private** | 22696 | 73.87 | 25756 | 83.83 |
| **Self-emp-inc** | 1116 | 3.63 | 26872 | 87.46 |
| **Self-emp-not-inc** | 2541 | 8.27 | 29413 | 95.73 |
| **State-gov** | 1298 | 4.22 | 30711 | 99.95 |
| **Without-pay** | 14 | 0.05 | 30725 | 100.00 |

| **education** | | | | |
| --- | --- | --- | --- | --- |
| **education** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **10th** | 933 | 2.87 | 933 | 2.87 |
| **11th** | 1175 | 3.61 | 2108 | 6.47 |
| **12th** | 433 | 1.33 | 2541 | 7.80 |
| **1st-4th** | 168 | 0.52 | 2709 | 8.32 |
| **5th-6th** | 333 | 1.02 | 3042 | 9.34 |
| **7th-8th** | 646 | 1.98 | 3688 | 11.33 |
| **9th** | 514 | 1.58 | 4202 | 12.91 |
| **Assoc-acdm** | 1067 | 3.28 | 5269 | 16.18 |
| **Assoc-voc** | 1382 | 4.24 | 6651 | 20.43 |
| **Bachelors** | 5355 | 16.45 | 12006 | 36.87 |
| **Doctorate** | 413 | 1.27 | 12419 | 38.14 |
| **HS-grad** | 10501 | 32.25 | 22920 | 70.39 |
| **Masters** | 1723 | 5.29 | 24643 | 75.68 |
| **Preschool** | 51 | 0.16 | 24694 | 75.84 |
| **Prof-school** | 576 | 1.77 | 25270 | 77.61 |
| **Some-college** | 7291 | 22.39 | 32561 | 100.00 |

| **education\_num** | | | | |
| --- | --- | --- | --- | --- |
| **education\_num** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **1** | 51 | 0.16 | 51 | 0.16 |
| **2** | 168 | 0.52 | 219 | 0.67 |
| **3** | 333 | 1.02 | 552 | 1.70 |
| **4** | 646 | 1.98 | 1198 | 3.68 |
| **5** | 514 | 1.58 | 1712 | 5.26 |
| **6** | 933 | 2.87 | 2645 | 8.12 |
| **7** | 1175 | 3.61 | 3820 | 11.73 |
| **8** | 433 | 1.33 | 4253 | 13.06 |
| **9** | 10501 | 32.25 | 14754 | 45.31 |
| **10** | 7291 | 22.39 | 22045 | 67.70 |
| **11** | 1382 | 4.24 | 23427 | 71.95 |
| **12** | 1067 | 3.28 | 24494 | 75.22 |
| **13** | 5355 | 16.45 | 29849 | 91.67 |
| **14** | 1723 | 5.29 | 31572 | 96.96 |
| **15** | 576 | 1.77 | 32148 | 98.73 |
| **16** | 413 | 1.27 | 32561 | 100.00 |

| **marital\_status** | | | | |
| --- | --- | --- | --- | --- |
| **marital\_status** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Divorced** | 4443 | 13.65 | 4443 | 13.65 |
| **Married-AF-spouse** | 23 | 0.07 | 4466 | 13.72 |
| **Married-civ-spouse** | 14976 | 45.99 | 19442 | 59.71 |
| **Married-spouse-absent** | 418 | 1.28 | 19860 | 60.99 |
| **Never-married** | 10683 | 32.81 | 30543 | 93.80 |
| **Separated** | 1025 | 3.15 | 31568 | 96.95 |
| **Widowed** | 993 | 3.05 | 32561 | 100.00 |

| **occupation** | | | | |
| --- | --- | --- | --- | --- |
| **occupation** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Frequency Missing = 1843** | | | | |
| **Adm-clerical** | 3770 | 12.27 | 3770 | 12.27 |
| **Armed-Forces** | 9 | 0.03 | 3779 | 12.30 |
| **Craft-repair** | 4099 | 13.34 | 7878 | 25.65 |
| **Exec-managerial** | 4066 | 13.24 | 11944 | 38.88 |
| **Farming-fishing** | 994 | 3.24 | 12938 | 42.12 |
| **Handlers-cleaners** | 1370 | 4.46 | 14308 | 46.58 |
| **Machine-op-inspct** | 2002 | 6.52 | 16310 | 53.10 |
| **Other-service** | 3295 | 10.73 | 19605 | 63.82 |
| **Priv-house-serv** | 149 | 0.49 | 19754 | 64.31 |
| **Prof-specialty** | 4140 | 13.48 | 23894 | 77.79 |
| **Protective-serv** | 649 | 2.11 | 24543 | 79.90 |
| **Sales** | 3650 | 11.88 | 28193 | 91.78 |
| **Tech-support** | 928 | 3.02 | 29121 | 94.80 |
| **Transport-moving** | 1597 | 5.20 | 30718 | 100.00 |

| **relationship** | | | | |
| --- | --- | --- | --- | --- |
| **relationship** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Husband** | 13193 | 40.52 | 13193 | 40.52 |
| **Not-in-family** | 8305 | 25.51 | 21498 | 66.02 |
| **Other-relative** | 981 | 3.01 | 22479 | 69.04 |
| **Own-child** | 5068 | 15.56 | 27547 | 84.60 |
| **Unmarried** | 3446 | 10.58 | 30993 | 95.18 |
| **Wife** | 1568 | 4.82 | 32561 | 100.00 |

| **race** | | | | |
| --- | --- | --- | --- | --- |
| **race** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Amer-Indian-Eskimo** | 311 | 0.96 | 311 | 0.96 |
| **Asian-Pac-Islander** | 1039 | 3.19 | 1350 | 4.15 |
| **Black** | 3124 | 9.59 | 4474 | 13.74 |
| **Other** | 271 | 0.83 | 4745 | 14.57 |
| **White** | 27816 | 85.43 | 32561 | 100.00 |

| **sex** | | | | |
| --- | --- | --- | --- | --- |
| **sex** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Female** | 10771 | 33.08 | 10771 | 33.08 |
| **Male** | 21790 | 66.92 | 32561 | 100.00 |

| **native\_country** | | | | |
| --- | --- | --- | --- | --- |
| **native\_country** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Frequency Missing = 583** | | | | |
| **Cambodia** | 19 | 0.06 | 19 | 0.06 |
| **Canada** | 121 | 0.38 | 140 | 0.44 |
| **China** | 75 | 0.23 | 215 | 0.67 |
| **Columbia** | 59 | 0.18 | 274 | 0.86 |
| **Cuba** | 95 | 0.30 | 369 | 1.15 |
| **Dominican-Republic** | 70 | 0.22 | 439 | 1.37 |
| **Ecuador** | 28 | 0.09 | 467 | 1.46 |
| **El-Salvador** | 106 | 0.33 | 573 | 1.79 |
| **England** | 90 | 0.28 | 663 | 2.07 |
| **France** | 29 | 0.09 | 692 | 2.16 |
| **Germany** | 137 | 0.43 | 829 | 2.59 |
| **Greece** | 29 | 0.09 | 858 | 2.68 |
| **Guatemala** | 64 | 0.20 | 922 | 2.88 |
| **Haiti** | 44 | 0.14 | 966 | 3.02 |
| **Holand-Netherlands** | 1 | 0.00 | 967 | 3.02 |
| **Honduras** | 13 | 0.04 | 980 | 3.06 |
| **Hong** | 20 | 0.06 | 1000 | 3.13 |
| **Hungary** | 13 | 0.04 | 1013 | 3.17 |
| **India** | 100 | 0.31 | 1113 | 3.48 |
| **Iran** | 43 | 0.13 | 1156 | 3.61 |
| **Ireland** | 24 | 0.08 | 1180 | 3.69 |
| **Italy** | 73 | 0.23 | 1253 | 3.92 |
| **Jamaica** | 81 | 0.25 | 1334 | 4.17 |
| **Japan** | 62 | 0.19 | 1396 | 4.37 |
| **Laos** | 18 | 0.06 | 1414 | 4.42 |
| **Mexico** | 643 | 2.01 | 2057 | 6.43 |
| **Nicaragua** | 34 | 0.11 | 2091 | 6.54 |
| **Outlying-US(Guam-USVI-etc)** | 14 | 0.04 | 2105 | 6.58 |
| **Peru** | 31 | 0.10 | 2136 | 6.68 |
| **Philippines** | 198 | 0.62 | 2334 | 7.30 |
| **Poland** | 60 | 0.19 | 2394 | 7.49 |
| **Portugal** | 37 | 0.12 | 2431 | 7.60 |
| **Puerto-Rico** | 114 | 0.36 | 2545 | 7.96 |
| **Scotland** | 12 | 0.04 | 2557 | 8.00 |
| **South** | 80 | 0.25 | 2637 | 8.25 |
| **Taiwan** | 51 | 0.16 | 2688 | 8.41 |
| **Thailand** | 18 | 0.06 | 2706 | 8.46 |
| **Trinadad&Tobago** | 19 | 0.06 | 2725 | 8.52 |
| **United-States** | 29170 | 91.22 | 31895 | 99.74 |
| **Vietnam** | 67 | 0.21 | 31962 | 99.95 |
| **Yugoslavia** | 16 | 0.05 | 31978 | 100.00 |

| **income** | | | | |
| --- | --- | --- | --- | --- |
| **income** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **<=50K** | 24720 | 75.92 | 24720 | 75.92 |
| **>50K** | 7841 | 24.08 | 32561 | 100.00 |

### Interpretation:

The dataset is not balanced because of the following reasons:

1. About three quarters of the population falls in the lower income category i.e., less than-equal to 50K.
2. 91% of the population is a native of United States of America. The rest 9% belong to rest of the world.
3. Distribution of Sex is not equitable. 2/3rd of the population is Male.
4. The population is White dominant as the constitution of whites is 85%. About 10% are Blacks.

## Numerical Variables

### Summary Statistics

SAS CODE:

A picture containing text

Description automatically generated

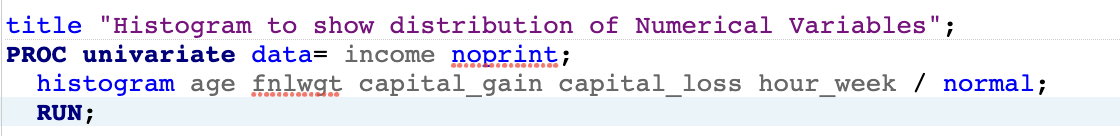
Results:

Table

Description automatically generated

### Shape of Distribution: Histograms

SAS Code:



Result:

**Histogram to show distribution of Numerical Variables**

**The UNIVARIATE Procedure**

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Graphical user interface

Description automatically generated

Graphical user interface

Description automatically generated

Chart, histogram

Description automatically generated

# Data Handling: Categorical Variables

## Step 1: Separating categorical variables

**data adults\_categorical;**

**set adults\_data (keep = WorkClass Education MaritalStatus Occupation Relationship Race**

**Sex NativeCountry);**

**run;**

## Step 2: Running the freq procedure on the categorical variables

**Proc freq data = aduls\_categorical;**

**run;**

Results

**The FREQ Procedure**

| **WorkClass** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Frequency Missing = 1** | | | | |
| **?** | 1836 | 5.64 | 1836 | 5.64 |
| **Federal-gov** | 960 | 2.95 | 2796 | 8.59 |
| **Local-gov** | 2093 | 6.43 | 4889 | 15.01 |
| **Never-worked** | 7 | 0.02 | 4896 | 15.04 |
| **Private** | 22696 | 69.70 | 27592 | 84.74 |
| **Self-emp-inc** | 1116 | 3.43 | 28708 | 88.17 |
| **Self-emp-not-inc** | 2541 | 7.80 | 31249 | 95.97 |
| **State-gov** | 1298 | 3.99 | 32547 | 99.96 |
| **Without-pay** | 14 | 0.04 | 32561 | 100.00 |

| **Education** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Frequency Missing = 1** | | | | |
| **10th** | 933 | 2.87 | 933 | 2.87 |
| **11th** | 1175 | 3.61 | 2108 | 6.47 |
| **12th** | 433 | 1.33 | 2541 | 7.80 |
| **1st-4th** | 168 | 0.52 | 2709 | 8.32 |
| **5th-6th** | 333 | 1.02 | 3042 | 9.34 |
| **7th-8th** | 646 | 1.98 | 3688 | 11.33 |
| **9th** | 514 | 1.58 | 4202 | 12.91 |
| **Assoc-acdm** | 1067 | 3.28 | 5269 | 16.18 |
| **Assoc-voc** | 1382 | 4.24 | 6651 | 20.43 |
| **Bachelors** | 5355 | 16.45 | 12006 | 36.87 |
| **Doctorate** | 413 | 1.27 | 12419 | 38.14 |
| **HS-grad** | 10501 | 32.25 | 22920 | 70.39 |
| **Masters** | 1723 | 5.29 | 24643 | 75.68 |
| **Preschool** | 51 | 0.16 | 24694 | 75.84 |
| **Prof-school** | 576 | 1.77 | 25270 | 77.61 |
| **Some-college** | 7291 | 22.39 | 32561 | 100.00 |

| **MaritalStatus** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Frequency Missing = 1** | | | | |
| **Divorced** | 4443 | 13.65 | 4443 | 13.65 |
| **Married-AF-spouse** | 23 | 0.07 | 4466 | 13.72 |
| **Married-civ-spouse** | 14976 | 45.99 | 19442 | 59.71 |
| **Married-spouse-absent** | 418 | 1.28 | 19860 | 60.99 |
| **Never-married** | 10683 | 32.81 | 30543 | 93.80 |
| **Separated** | 1025 | 3.15 | 31568 | 96.95 |
| **Widowed** | 993 | 3.05 | 32561 | 100.00 |

| **Occupation** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Frequency Missing = 1** | | | | |
| **?** | 1843 | 5.66 | 1843 | 5.66 |
| **Adm-clerical** | 3770 | 11.58 | 5613 | 17.24 |
| **Armed-Forces** | 9 | 0.03 | 5622 | 17.27 |
| **Craft-repair** | 4099 | 12.59 | 9721 | 29.85 |
| **Exec-managerial** | 4066 | 12.49 | 13787 | 42.34 |
| **Farming-fishing** | 994 | 3.05 | 14781 | 45.39 |
| **Handlers-cleaners** | 1370 | 4.21 | 16151 | 49.60 |
| **Machine-op-inspct** | 2002 | 6.15 | 18153 | 55.75 |
| **Other-service** | 3295 | 10.12 | 21448 | 65.87 |
| **Priv-house-serv** | 149 | 0.46 | 21597 | 66.33 |
| **Prof-specialty** | 4140 | 12.71 | 25737 | 79.04 |
| **Protective-serv** | 649 | 1.99 | 26386 | 81.04 |
| **Sales** | 3650 | 11.21 | 30036 | 92.25 |
| **Tech-support** | 928 | 2.85 | 30964 | 95.10 |
| **Transport-moving** | 1597 | 4.90 | 32561 | 100.00 |

| **Relationship** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Frequency Missing = 1** | | | | |
| **Husband** | 13193 | 40.52 | 13193 | 40.52 |
| **Not-in-family** | 8305 | 25.51 | 21498 | 66.02 |
| **Other-relativ** | 981 | 3.01 | 22479 | 69.04 |
| **Own-child** | 5068 | 15.56 | 27547 | 84.60 |
| **Unmarried** | 3446 | 10.58 | 30993 | 95.18 |
| **Wife** | 1568 | 4.82 | 32561 | 100.00 |

| **Race** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Frequency Missing = 1** | | | | |
| **Amer-Indian-Eskimo** | 311 | 0.96 | 311 | 0.96 |
| **Asian-Pac-Islander** | 1039 | 3.19 | 1350 | 4.15 |
| **Black** | 3124 | 9.59 | 4474 | 13.74 |
| **Other** | 271 | 0.83 | 4745 | 14.57 |
| **White** | 27816 | 85.43 | 32561 | 100.00 |

| **Sex** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Frequency Missing = 1** | | | | |
| **Female** | 10771 | 33.08 | 10771 | 33.08 |
| **Male** | 21790 | 66.92 | 32561 | 100.00 |

| **NativeCountry** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **Frequency Missing = 1** | | | | |
| **?** | 583 | 1.79 | 583 | 1.79 |
| **Cambodia** | 19 | 0.06 | 602 | 1.85 |
| **Canada** | 121 | 0.37 | 723 | 2.22 |
| **China** | 75 | 0.23 | 798 | 2.45 |
| **Columbia** | 59 | 0.18 | 857 | 2.63 |
| **Cuba** | 95 | 0.29 | 952 | 2.92 |
| **Dominican-Rep** | 70 | 0.21 | 1022 | 3.14 |
| **Ecuador** | 28 | 0.09 | 1050 | 3.22 |
| **El-Salvador** | 106 | 0.33 | 1156 | 3.55 |
| **England** | 90 | 0.28 | 1246 | 3.83 |
| **France** | 29 | 0.09 | 1275 | 3.92 |
| **Germany** | 137 | 0.42 | 1412 | 4.34 |
| **Greece** | 29 | 0.09 | 1441 | 4.43 |
| **Guatemala** | 64 | 0.20 | 1505 | 4.62 |
| **Haiti** | 44 | 0.14 | 1549 | 4.76 |
| **Holand-Nether** | 1 | 0.00 | 1550 | 4.76 |
| **Honduras** | 13 | 0.04 | 1563 | 4.80 |
| **Hong** | 20 | 0.06 | 1583 | 4.86 |
| **Hungary** | 13 | 0.04 | 1596 | 4.90 |
| **India** | 100 | 0.31 | 1696 | 5.21 |
| **Iran** | 43 | 0.13 | 1739 | 5.34 |
| **Ireland** | 24 | 0.07 | 1763 | 5.41 |
| **Italy** | 73 | 0.22 | 1836 | 5.64 |
| **Jamaica** | 81 | 0.25 | 1917 | 5.89 |
| **Japan** | 62 | 0.19 | 1979 | 6.08 |
| **Laos** | 18 | 0.06 | 1997 | 6.13 |
| **Mexico** | 643 | 1.97 | 2640 | 8.11 |
| **Nicaragua** | 34 | 0.10 | 2674 | 8.21 |
| **Outlying-US(G** | 14 | 0.04 | 2688 | 8.26 |
| **Peru** | 31 | 0.10 | 2719 | 8.35 |
| **Philippines** | 198 | 0.61 | 2917 | 8.96 |
| **Poland** | 60 | 0.18 | 2977 | 9.14 |
| **Portugal** | 37 | 0.11 | 3014 | 9.26 |
| **Puerto-Rico** | 114 | 0.35 | 3128 | 9.61 |
| **Scotland** | 12 | 0.04 | 3140 | 9.64 |
| **South** | 80 | 0.25 | 3220 | 9.89 |
| **Taiwan** | 51 | 0.16 | 3271 | 10.05 |
| **Thailand** | 18 | 0.06 | 3289 | 10.10 |
| **Trinadad&Toba** | 19 | 0.06 | 3308 | 10.16 |
| **United-States** | 29170 | 89.59 | 32478 | 99.75 |
| **Vietnam** | 67 | 0.21 | 32545 | 99.95 |
| **Yugoslavia** | 16 | 0.05 | 32561 | 100.00 |

The results shows that the variables NativeCountry, Occupation and Workclass have missing values.

## Step3: Data Cleaning

1. Checking for records where all three variables are missing

**proc print data = adults\_categorical;**

**where workclass = "?" and occupation = "?" and NativeCountry="?";**

**run;**

1. Deleting records where all three variables are missing as the number of records are insignificant.

**data missing\_correction;**

**set adults\_categorical;**

**if workclass="?" and occupation="?" and NativeCountry="?" then delete;**

**run;**

1. Replacing missing values for workclass and occupation with ‘Other’

**data missing\_other;**

**set missing\_correction;**

**if workclass = "?" then workclass="Other";**

**if Occupation = "?" then Occupation = "Other";**

**run;**

1. Deleting records where NativeCountry is missing since there are only 58 records.

**data missing\_country;**

**set missing\_other;**

**if NativeCountry="?" then delete;**

**run;**

## Step 4: Creating derived variables

Two new derived variables are created to get more insight on the data. The variables are Min wage and Location. The dependency of the derived variables is as following:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Hour Week** | **Annual Income** | **Min Wage** |
| Values | >40 | <50k | Yes |
|  | <40 | >50k | No |

Code:

**DATA AnnualIncome\_Clean;**

**set AnnualIncome;**

**if hour\_week>40 and income='<=50K'**

**then Min\_Wage='Yes';**

**if hour\_week<=40 and income='>50K'**

**then Min\_Wage='No';**

**if hour\_week>40 and income='>50K'**

**then Min\_Wage='NW';\*NW represents Normal Wage;**

**if hour\_week<=40 and income='<=50K'**

**then Min\_Wage='NW';\*NW represents Normal Wage;**

**run;**

Output:

Table, calendar

Description automatically generated

Table

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **native\_country** | **Location** |
| Values | United-States | US |
| (All other countries) | Non-US |

Code:

**DATA AnnualIncome\_Clean;**

**set AnnualIncome;**

**if native\_country='United-States'**

**then Location='US';**

**if native\_country^='United-States'**

**then Location='non US';**

**run;**

Output:

Table

Description automatically generated

Table

Description automatically generated

# Data Handling: Numerical Variables

## Error Checking: Out of Range and Missing Values

Here are all the numerical variables with some basic statistical information.

**proc means data=mylib.adulterror n nmiss min max mean;**

**run;**

Table

Description automatically generated

## Checking for errors using pre-determined range

**data \_null\_;**

**set mylib.adulterror;**

**file print;**

**if age lt 16 or age gt 120 then**

**put "Out of range AGE value: " age " for ID " ID;**

**if hour\_week le 0 then**

**put "Out of range HOUR\_WEEK value: " hour\_week " for ID " ID;**

**if education\_num lt 1 or education\_num gt 16 then**

**put "Out of range EDUCATION\_NUM value: " education\_num " for ID " ID;**

**run;**

* Two Out of Range Values
* 6 Missing Values

Graphical user interface, text, application, email

Description automatically generated

## Correcting Out of Range Values by Deletion

Out of range values were deleted because it was an obvious error considering human lifespan. Anything over age of 120 was deleted.

**data error\_correction;**

**set mylib.adulterror;**

**if age ge 120 or hour\_week eq 0 then**

**delete;**

**run;**

## Deleting Missing Values for Age

**data deleted\_values;**

**set error\_correction;**

**if missing(age) then**

**delete;**

**run;**

The number of missing values were only 6 and wouldn’t have any impact on integrity of the dataset.

## Detecting Outliers

Distribution of variables: AGE, HOUR\_WEEK

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

**We’ll be using Interquartile range to detect outliers for age because it’s positively skewed.**

**Age**

**proc means data=deleted\_values noprint;**

**var age;**

**output out=work.Tmp Q1=Q3=QRange= / autoname;**

**run;**

**data \_null\_;**

**file print;**

**set deleted\_values(keep=id age);**

**if \_n\_=1 then**

**set Tmp;**

**if age le age\_Q1 - 2\*age\_QRange and not missing(age) or age ge**

**age\_Q3 + 2\*age\_QRange then**

**put "Possible Outlier for index " ID "Value of Age is " age;**

**run;**

Table

Description automatically generated

Outliers for Age were deleted because it represented a small number of observations. Outliers for age were just 0.14% of values in the dataset.

**data age\_Out;**

**set deleted\_values;**

**if \_n\_=1 then**

**set Tmp;**

**if age le age\_Q1 - 2\*age\_QRange and not missing(age) or age ge**

**age\_Q3 + 2\*age\_QRange then**

**output;**

**run;**

**title "Isolated Outliers for age";**

**proc means data=age\_Out;**

**var age;**

**run;**

Table

Description automatically generated

**Hour\_Week**

**proc means data=outlier\_deleted noprint;**

**var hour\_week;**

**output out=work.Tmp2 Q1=Q3=QRange= / autoname;**

**run;**

**data \_null\_;**

**file print;**

**set outlier\_deleted(keep=id hour\_week);**

**if \_n\_=1 then**

**set tmp2;**

**if hour\_week le hour\_week\_Q1 - 3\*hour\_week\_QRange and not**

**missing(hour\_week) or hour\_week ge hour\_week\_Q3 +**

**3\*hour\_week\_QRange then**

**put "Possible Outlier for index " ID "Value of Hour\_Week is"**

**hour\_week;**

**run;**

There were 6513 outliers for hour\_week even though extreme range values were used. The reason for this is similar mean, median and mode. No outliers were removed to preserve the integrity of the data because this represents about 20% of the data.

**data hour\_weekOut;**

**set outlier\_deleted(keep=id hour\_week);**

**if \_n\_=1 then**

**set tmp2;**

**if hour\_week le hour\_week\_Q1 - 3\*hour\_week\_QRange and not missing(hour\_week) or hour\_week ge hour\_week\_Q3 + 3\*hour\_week\_QRange then output;**

**run;**

**proc means data=hour\_weekOut;**

**var hour\_week;**

**run;**

Table

Description automatically generated

## Tests for Normality and Distribution

**Fnlwgt**

Chart, histogram

Description automatically generated

**QQ Plot forms a curve which means that the data is highlight skewed.**

Chart

Description automatically generated

**Applying Log Transformation on fnlwgt**

**data log\_transform;**

**SET outlier\_deleted;**

**log\_fnlwgt=log(fnlwgt);**

**root4\_fnlwgt=(fnlwgt) \*\* 0.25;**

**run;**

Chart, histogram

Description automatically generated

Now the distribution looks more normal and linear.

Chart, line chart

Description automatically generated