Asia Pacific University of Technology & Innovation, Jalan Technology Park Malaysia

DATA MANAGEMENT ASSIGNMENT PART 2:
INCOME INEQUALITY IN UNITED STATES

INTAKE CODE: APDMF2112DSBA(DE)(PR)

MODULE CODE: CT051-3-M-DM

HAND OUT DATE: 01 March 2022

HAND IN DATE: 15 April 2022

LECTURE NAME: Assoc. Prof. Dr. Raja Rajeswari

STUDENT NAME: Linda Houmed Bililis

TP NUMBER: TP060864

TABLE OF CONTENT

I. INTRODUCTION

The US Census Bureau is chargeable for operating the national census once a minimum each ten years. The US Census Bureau is as well chargeable for manufacturing data concerning the American population and surveys regarding the economy and economic activity (*U.S. Census Bureau*, 2021). Moreover, it gathers the data and announces estimates on income and impoverishment to evaluate national economic trends and to grasp their impact on the well-being of household families, and people. The U.S. Census Bureau defined household income as the total gross cash income of all family members, defined as a group of people living together who are fifteen years or older. Pew Research Centre (2012) explain also that household financial gain is the addition of incomes earned by all members of the menage within the calendar year before the time of the survey. And, income inequality is how inconsistently financial gain is spread through a population. The minus equals the distribution, the upper financial gain difference is. Revenue inequality is commonly in the middle of wealth difference, that is the unequal distribution of wealth (How Income Inequality Works, 2021). Therefore, US revenue inequality is common in the middle of wealth difference, that is the unequal distribution of wealth. Omar & Inaba (2020) highlighted that because of insufficient revenue levels and market inequity in development provinces, there are yet numerous folks unwillingly eliminated from the economic system, which generates possible failure of investments, investable funds, and accumulation of capital. Additionally, it requires an increase in poverty level, which is defined as the economic condition rate which is a magnitude relation of the amount of the individual (in each group) whose financial gain fails below the economic condition line. Taken as 1\2 the median household income of the entire population. It is additionally on the market for a broad age group. Child poverty (0-17 years old), working-age economic condition, and old economic condition (66 years old or more) (Inequality - Poverty Rate - OECD Data, n.d.).

Census Demographic data is collected, in order to complete, this study. This report is concerned with analyses of this data, in aims to clarify the relationship between the income inequality based on the race, gender, poverty, job, child poverty, workplace, and employment. According to the Census Bureau, "there is a tendency in household surveys for the respondent to underreport their income. By an analysis of independently derived income assumption, salary and wage often seem to be much better reported than income type such as social security, net income, public assistance from concern, dividends, rent, and so on." (Sources et al., 2011).

The present paper is structured as follows: In the first section, related work has been developed to seek whether the current problem exists. Then followed by the methodology section to explore and evaluates the data. The present data is real-world data, which contains the missing values and outliers. Therefore, to deal with the noisy, and inadequate data several methods are used, such as missing values imputation, feature engineering, and so on. Additionally, the result has been discussed in the third section. And finally, the last section is dedicated to the conclusion.

II. RELATED WORK

This report presents information about income and poverty in United State. Based on US Census demographic data collected. US Census Bureau is chargeable for operating the national census once a minimum each ten years. The US Census Bureau is as well responsible for manufacturing data concerning the American population and surveys regarding the economy and economic activity (*U.S. Census Bureau*, 2021). It presents annual estimates of median menage income and financial condition by states and alternative smaller geographic units based mostly on information collected within the American Community Survey (ACS) (DeNavas-Walt & Proctor, 2014a). The ACS multi-year assessment program has deeply advanced prospects for finding out modifications within the demographic and characteristics of the US (Weden et al., 2015). US Census Bureau survey based on American Community Survey is a current yearly survey that reveals what the United States population feels like and the way it goes. The ACS assists society decide wherever to focus on services and resources. While demographic surveys determine education, housing equality, income, insurance coverage, laptop use, poverty, crime abuse, and plenty of alternative subjects (US Census Bureau, 2021).

Briefly, US revenue inequality is common in the middle of wealth difference, that is the unequal distribution of wealth. A quiet assessment of wealth within the US discovers proof of unexpected racial inequalities. Gaps in wealth between black and white households reveal the consequence of accrued difference and discrimination, as well as the difference in power and chance that may be derived back to the current nation's origination. The black-white wealth gap reflects a society that has not and does not afford equality of chance to any or all its voters (McIntosh et al., 2020). Omar & Inaba (2020) highlighted that because of insufficient revenue levels and market inequity in development provinces, there are yet numerous folks unwillingly eliminated from the economic system, which generates possible failure of investments, investable funds, and accumulation of capital.

DeNavas-Walt & Proctor summarized the finding of income and poverty in the US for the earlier years (2018, 2019, 2020) based on the previous recent Population Survey Annual Social Economic Supplements (CPSASE). DeNavas-Walt & Proctor (2014c) clarify in their report that the approved financial condition rate in 2018 was 11.8%, a reduction of 0.5% point from 2017. This is the 4th successive yearly decrease in the national poverty proportion. Moreover, the median family financial gain was 63179$, not completely changed from the 2017 median, subsequent three repeated years of annual will increase. They also added that the on lass to expertise an important growth in revenue rate from 2017 to 2018 is because of persons aged 25 or older with high school certification. In 2019, the median family revenue was 68703$, which is a rise of 6.8% from the 2018 median of 64324%. Though, the official economic condition proportion was 10.5%, down 1.3% point from 11.8 percent in 2018. It assumes that the median household income estimates in 2019 were higher or were not statistically distinct from the 2018 estimation. However, the revenue in 2018 the poverty rates were either lower or not statistically different (DeNavas-Walt & Proctor, 2014c). Additionally, in 2020 the median family financial gain declined from 2,9% It was 67,521$ compared to the 2019 median which is 69,560$. while, for the official financial condition percentage in 2020 was 11.4%, up 1.0 proportion from 10.5% in 2019 (DeNavas-Walt & Proctor, 2014a). Consequently, the fact of

income inequality is summarized by racial discrimination, growth of poverty rate, child poverty, and so on.

III. METHODOLOGY

Methods are used to get and evaluate inequality financial gain prediction data supported by the present population survey provided by the US Census Bureau. The dataset is known as "US Census Demographic Data" and is extracted from Kaggle, which is the 2005 United States Census Bureau database. And it contains 74001 observations and 21 features. The expanded data will provide way more fascinating analyses and can even be way more helpful to predict the income. It will facilitate searching for insight into how different attributes influence the income of an individual. Therefore, SAS software is employed to conduct the current study.

1. Data Exploration

a. Data Description

The dataset contains 20 independent variables and a response variable which is income. Below is the detail of the attributes.

Table 1: Data Description

Variables	Description
TotalPop	Total population
Men	Number of men
Women	Number of women
White	% of the population that is white
Black	% of population that is black
Citizen	Number of citizens
Poverty	% under poverty level
ChildPoverty	% of children under poverty level
Professional	% employed in management, business, science, and arts
Service	% employed in service jobs
Office	% employed in sales and office jobs
Construction	% employed in natural resources, construction, and maintenance
WorkAtHome	% working at home
MeanCommute	Mean commute time (minutes)
Employed	Number of employed (16+)
PrivateWork	% employed in private industry
PublicWork	% employed in public jobs
SelfEmployed	% self-employed
FamilyWork	% in unpaid family work
Unemployment	Unemployment rate (%)
Income	Median household income ($)

To start with, the dataset is imported and loaded on SAS under Work.Import file. And the result shows that the data have 74001 observations and 21 attributes.

```
1  /* Generated Code (IMPORT) */
2  /* Source File: assignment data.csv */
3  /* Source Path: /home/u60775558/sasuser.v94 */
4  /* Code generated on: 4/12/22, 2:12 PM */
5
6  %web_drop_table(WORK.IMPORT);
7
8
9  FILENAME REFFILE '/home/u60775558/sasuser.v94/assignment data.csv';
10
11 PROC IMPORT DATAFILE=REFFILE
12     DBMS=CSV
13     OUT=WORK.IMPORT;
14     GETNAMES=YES;
15 RUN;
16
17 PROC CONTENTS DATA=WORK.IMPORT; RUN;
18
```

The CONTENTS Procedure

Data Set Name	WORK.IMPORT	Observations	74001
Member Type	DATA	Variables	21
Engine	V9	Indexes	0
Created	04/12/2022 14:12:28	Observation Length	168
Last Modified	04/12/2022 14:12:28	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Figure 1: Data Content

b. Data identification

In this section the type and categories of data are identified, the figure 2 below demonstrates that the data is numerical, and all the variables are continuous. To take a closer look, the first 20 rows are selected and represented in figure 3.

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/13/22, 1:21 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS01-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

ods noproctitle;
ods select attributes position;

proc datasets;
    contents data=WORK.IMPORT order=varnum;
quit;
```

Variables in Creation Order

#	Variable	Type	Len	Format	Informat
1	TotalPop	Num	8	BEST12.	BEST32.
2	Men	Num	8	BEST12.	BEST32.
3	Women	Num	8	BEST12.	BEST32.
4	White	Num	8	BEST12.	BEST32.
5	Black	Num	8	BEST12.	BEST32.
6	Citizen	Num	8	BEST12.	BEST32.
7	Income	Num	8	BEST12.	BEST32.
8	Poverty	Num	8	BEST12.	BEST32.
9	ChildPoverty	Num	8	BEST12.	BEST32.
10	Professional	Num	8	BEST12.	BEST32.
11	Service	Num	8	BEST12.	BEST32.
12	Office	Num	8	BEST12.	BEST32.
13	Construction	Num	8	BEST12.	BEST32.
14	WorkAtHome	Num	8	BEST12.	BEST32.
15	MeanCommute	Num	8	BEST12.	BEST32.
16	Employed	Num	8	BEST12.	BEST32.
17	PrivateWork	Num	8	BEST12.	BEST32.
18	PublicWork	Num	8	BEST12.	BEST32.
19	SelfEmployed	Num	8	BEST12.	BEST32.
20	FamilyWork	Num	8	BEST12.	BEST32.
21	Unemployment	Num	8	BEST12.	BEST32.

Figure 2: Data Type

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/13/22, 1:15 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS01-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

title1 'List Data for WORK.IMPORT';

proc print data=WORK.IMPORT
    (obs=20) label;
    var TotalPop Men Women White Black Citizen Income Poverty ChildPoverty
        Professional Service Office Construction WorkAtHome MeanCommute Employed
        PrivateWork PublicWork SelfEmployed FamilyWork Unemployment;
run;

title1;
```

The list of 20 rows from data, clarifies the data type and variable categories.

List Data for WORK.IMPORT

Obs	TotalPop	Men	Women	White	Black	Citizen	Income	Poverty	ChildPoverty	Professional	Service	Office	Construction	WorkAtHome	MeanCommute	Employed	PrivateWork	PublicWork	SelfEmployed	FamilyWork
1	1948	940	1008	87.4	7.7	1503	61838	8.1	8.4	34.7	17	21.3	11.9	2.1	25	943	77.1	18.3	4.6	0
2	2156	1059	1097	40.4	53.3	1662	32303	25.5	40.3	22.3	24.7	21.5	9.4	0	23.4	753	77	16.9	6.1	0
3	2968	1364	1604	74.5	18.6	2335	44922	12.7	19.7	31.4	24.9	22.1	9.2	2.5	19.6	1373	64.1	23.6	12.3	0
4	4423	2172	2251	82.8	3.7	3306	54329	2.1	1.6	27	20.8	27	8.7	1.6	25.3	1782	75.7	21.2	3.1	0
5	10763	4922	5841	68.5	24.8	7666	51965	11.4	17.5	49.6	14.2	18.2	2.1	0.9	24.8	5037	67.1	27.6	5.3	0
6	3851	1787	2064	72.9	11.9	2642	63092	14.4	21.9	24.2	17.5	35.4	7.9	4.5	19.8	1560	79.4	14.7	5.8	0
7	2761	1210	1551	74.5	19.7	2060	34821	28.9	41.9	19.5	29.6	25.3	10.1	0	20	1166	82	14.6	3.4	0
8	3187	1502	1685	84	10.7	2391	73728	13	25.9	42.8	10.7	34.2	5.5	5.9	24.3	1502	78.1	14.8	7.1	0
9	10915	5486	5429	89.5	8.4	7778	60063	13.9	18.3	31.5	17.5	26.1	7.8	1.3	29.4	4348	73.3	22.1	4.6	0
10	5668	2897	2771	85.5	12.1	4217	41287	6.8	10	29.3	13.7	17.7	11	2.1	32.9	2485	77.9	15.2	6.9	0
11	3286	1740	1546	73.3	24.7	2470	49091	12.8	17.8	26.8	10.1	25	21.9	0.7	31.5	1541	76.9	19.5	3.6	0
12	3295	1666	1629	44.5	54.3	2695	32381	19.3	23.8	20.4	17.6	27.2	14	2.6	32.6	1496	70.7	24.7	4.6	0
13	3829	2157	1672	78.9	17.8	3170	39719	17	12.8	33.4	17.9	15.2	15.8	0.8	34.5	1406	73.9	21.5	4.6	0
14	2869	1324	1545	86	10.4	2182	31390	28.8	50.7	19.8	21.2	26.1	9.7	1.9	30.3	1221	78	13.9	8.1	0
15	7455	3462	3993	83.9	13.6	5596	51165	6.5	7	35.1	15.5	23.7	10.2	2.8	28.8	3033	79.4	17.5	2.4	0.8
16	4537	2311	2226	88.7	4.4	3483	44985	15	15.3	28	15.6	23.1	20.2	2.9	32.7	1790	80.6	11.4	8	0
17	5321	2711	2610	82.9	12.9	4129	41944	20.1	36.4	35.7	11.9	22.6	13.9	1	23.6	2042	69.3	23.8	6.9	0
18	3398	1821	1577	34	62.5	2262	27587	22.1	31.7	18.8	32.4	20.4	12.3	0.7	20.4	1392	71.7	23.9	4.5	0
19	7813	4098	3715	91.3	2.8	5679	74688	4.9	6.9	43.7	9.5	31.7	6.2	3.4	25.1	3838	81.1	11.5	6.4	0.9
20	16099	8213	7886	84.6	4.4	11287	82985	3.8	1.9	45.3	8.4	30.3	9.3	2.5	31.8	8843	80.1	16.1	3.8	0

Figure 3: List of Data

c. Descriptive statistic

Descriptive statistics will facilitate, describe, and perceive the variables of the working data by providing summaries regarding the sample and measures of the data. The summary statistic and univariate analysis are conducted in this section. They will provide an understanding of the central tendency and spread of the variables.

❖ Summary statistic

The following code used in SAS gives a summary statistic of the data, which is the median, mean, standard deviation, minimum, maximum, interquartile, and number of observations, as well as missing values of all the variables.

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/14/22, 1:54 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS01-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc means data=WORK.IMPORT chartype mean std min max median n nmiss vardef=df
        qrange qmethod=os;
    var TotalPop Men Women White Black Citizen Income Poverty ChildPoverty
        Professional Service Office Construction WorkAtHome MeanCommute Employed
        PrivateWork PublicWork SelfEmployed FamilyWork Unemployment;
run;
```

The result of the code shows in figure 4 that, the average household income in the US is 57225.56 $ in 2005, with the maximum household income being 248750$ and the minimum being 2611$. On average the percentage of the unemployed is 9.02%, the employed is 1983.91%, and the level of poverty is 16.96%. Moreover, there is a maximum of 27962 men, 27250 women, and a maximum of 53812 total population for that year. The median of the white variable is 71.40, whereas the mean is 62.03, this explains that the distribution is negatively skewed because the median is higher than the mean. For the childPoverty variable it shows the minimum is equal to 0, the median is 17.80 and the maximum is 100, this clarifies the presence of the outlier. Additionally, It also demonstrates the number of missing values for each variable. the variables White and Black contain 690 missing values. the variables Professional, Service, Office, Construction, PrivateWork, PublicWork, SelfEmployed, and FamilyWork include 807 missing values. income variable has 1100, ChilPoverty has 1118, poverty contains 835, and the variables WorkAtHome, MeanCommun, and Unemployment contain 797, 949, and 802 missing values respectively.

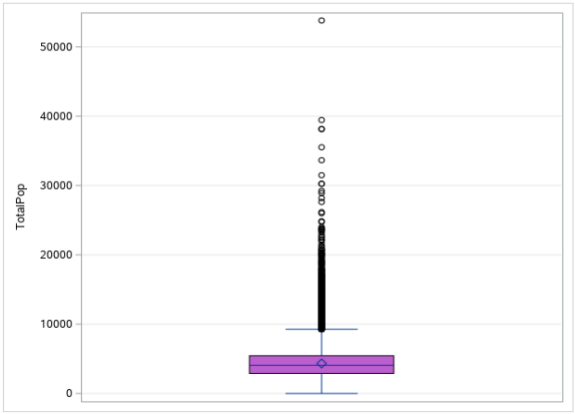
Variable	Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Quartile Range
TotalPop	4325.59	2129.31	0	53812.00	4063.00	74001	0	2551.00
Men	2127.65	1072.33	0	27962.00	1986.00	74001	0	1265.00
Women	2197.94	1095.73	0	27250.00	2066.00	74001	0	1313.00
White	62.0321057	30.6841523	0	100.0000000	71.4000000	73311	690	48.9000000
Black	13.2725812	21.7624828	0	100.0000000	3.7000000	73311	690	13.7000000
Citizen	3043.08	1475.49	0	37416.00	2863.00	74001	0	1801.00
Income	57225.56	28663.33	2611.00	248750.00	51094.00	72901	1100	32434.00
Poverty	16.9580201	13.1965176	0	100.0000000	13.4000000	73166	835	15.9000000
ChildPoverty	22.4868268	19.1909086	0	100.0000000	17.8000000	72883	1118	26.8000000
Professional	34.7988428	15.0070749	0	100.0000000	32.6000000	73194	807	19.8000000
Service	19.1013813	8.2791433	0	100.0000000	17.9000000	73194	807	10.2000000
Office	23.9515589	5.9572788	0	100.0000000	23.8000000	73194	807	7.4000000
Construction	9.2923368	6.0232895	0	100.0000000	8.4000000	73194	807	7.5000000
WorkAtHome	4.3680933	3.9049899	0	100.0000000	3.5000000	73204	797	4.1000000
MeanCommute	25.6673575	6.9648807	1.2000000	80.0000000	25.0000000	73052	949	9.0000000
Employed	1983.91	1073.43	0	24075.00	1846.00	74001	0	1304.00
PrivateWork	78.9752384	8.3457581	0	100.0000000	80.1000000	73194	807	10.0000000
PublicWork	14.6215660	7.5357861	0	100.0000000	13.4000000	73194	807	8.6000000
SelfEmployed	6.2338142	4.0429899	0	100.0000000	5.5000000	73194	807	4.6000000
FamilyWork	0.1697721	0.4582272	0	26.5000000	0	73194	807	0
Unemployment	9.0286630	5.9554415	0	100.0000000	7.7000000	73199	802	6.3000000

Figure 4: Summary Statistic

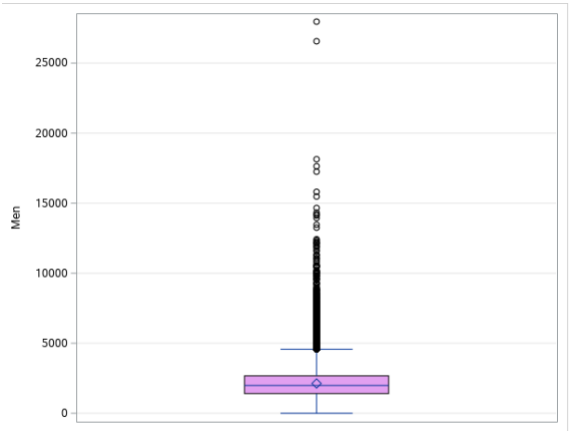
❖ Univariate Analysis

Univariate analysis assists in the understanding of the features one by one. Data visualization is also the best model to detect the outlier. As in the current data, the variables are all continuous, and boxplot and histogram are used.

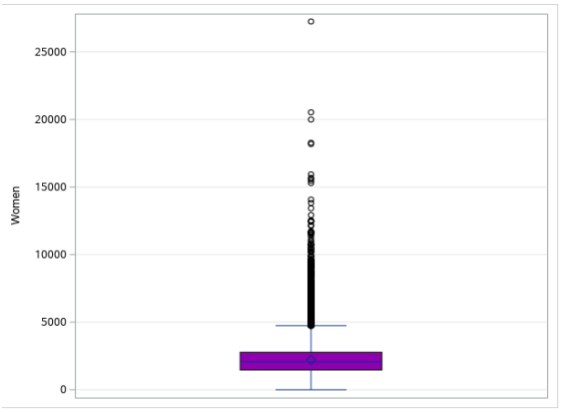
The following figure is the boxplot of the variables TotPop, Men, Women, and Professional. There is a presence of outliers, it is clear that the outlier is different from the typical data value.



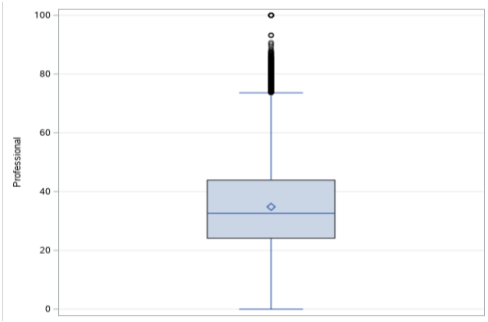

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/14/22, 3:06 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS01-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.IMPORT;
    vbox TotalPop / fillattrs=(color=CXbb5ece);
    yaxis grid;
run;

ods graphics / reset;
```




```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/14/22, 3:12 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS01-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.IMPORT;
    vbox Men / fillattrs=(color=CXe2a0f0);
    yaxis grid;
run;

ods graphics / reset;
```




```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/14/22, 3:16 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS01-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.IMPORT;
    vbox Women / fillattrs=(color=CX8e00af);
    yaxis grid;
run;

ods graphics / reset;
```




```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/14/22, 3:26 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS01-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.IMPORT;
    vbox Professional /;
    yaxis grid;
run;

ods graphics / reset;
```

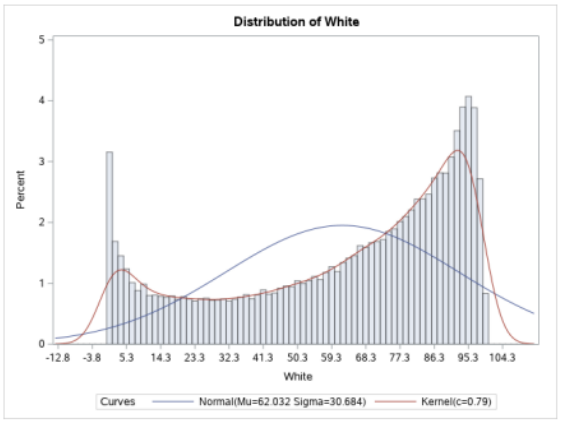
Figure 5: Boxplot for TotPop, Men, Women, and Professional

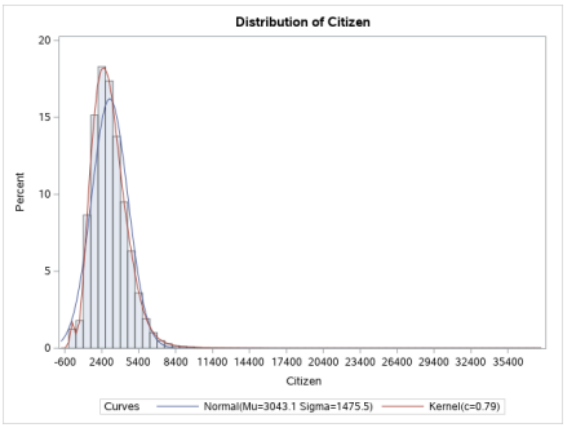
Then, by using a histogram, the outlier, and data distribution have been identified, and it also helps in understanding the center and spread of the data.

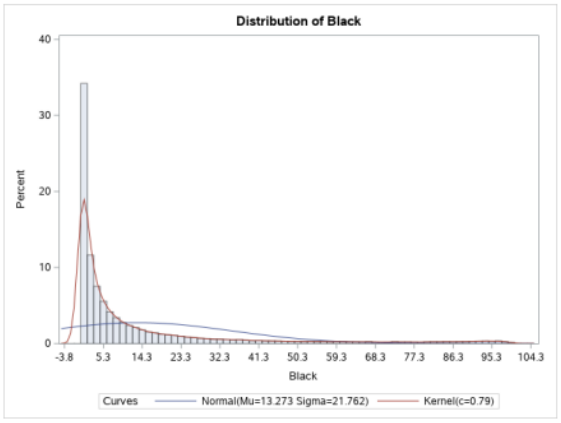
```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/14/22, 4:07 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS01-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

/* Exploring Data */
proc univariate data=WORK.IMPORT;
    ods select Histogram;
    var White Black Citizen Income Poverty ChildPoverty Service Office
        Construction WorkAtHome MeanCommute Employed PrivateWork PublicWork
        SelfEmployed FamilyWork Unemployment;
    histogram White Black Citizen Income Poverty ChildPoverty Service Office
        Construction WorkAtHome MeanCommute Employed PrivateWork PublicWork
        SelfEmployed FamilyWork Unemployment / normal kernel;
run;
```

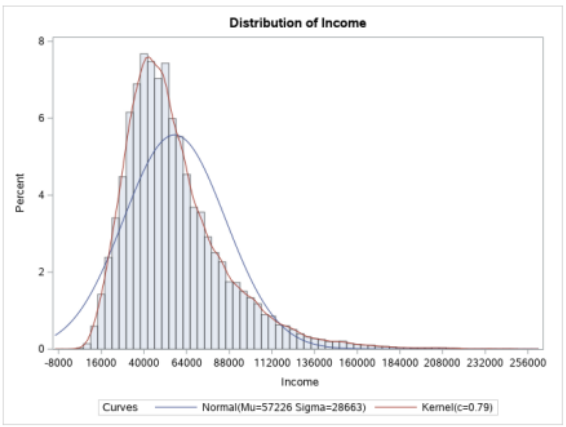
The histograms below show the distribution of the rest variables. The variables office and MeanCommute are normally distributed, and White and PrivateWork are negatively distributed, which means left-skewed distribution. For the rest variable, we have right-skewed, they are positively distributed. Based on the kernel density and normal curve the outlier is identified.

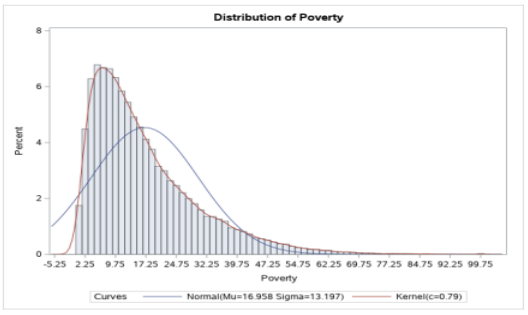
Based on the statistic descriptive, histogram, and boxplot, we visualized the data set comprising the missing values, as an outlier.

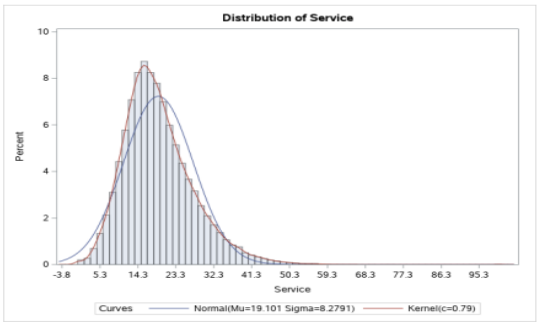

Distribution of White
Percent
White
Curves
Normal(Mu=62.032 Sigma=30.684)
Kernel(c=0.79)

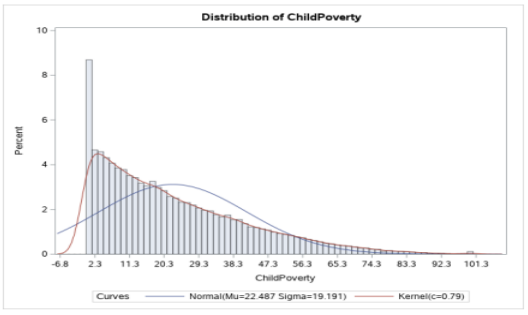

Distribution of Citizen
Percent
Citizen
Curves
Normal(Mu=3043.1 Sigma=1475.5)
Kernel(c=0.79)

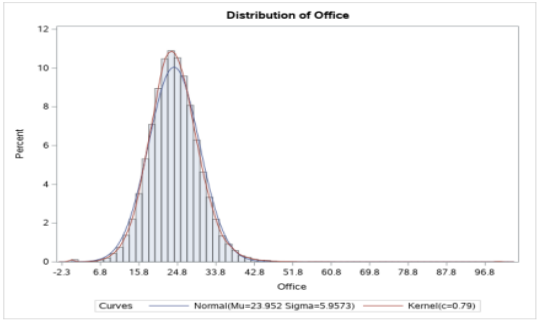

Distribution of Black
Percent
Black
Curves
Normal(Mu=13.273 Sigma=21.762)
Kernel(c=0.79)

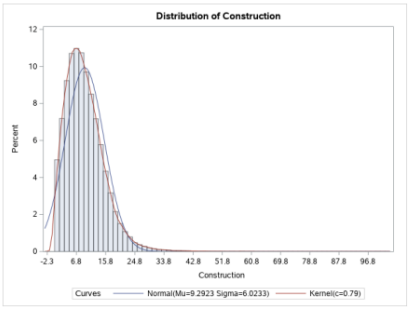

Distribution of Income
Percent
Income
Curves
Normal(Mu=57226 Sigma=28663)
Kernel(c=0.79)

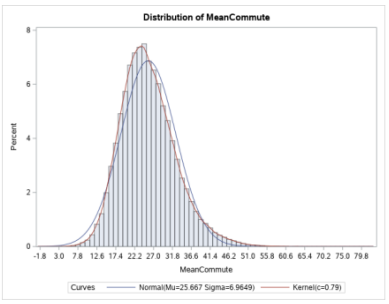

Distribution of Poverty
Percent
Poverty
Curves
Normal(Mu=16.958 Sigma=13.197)
Kernel(c=0.79)

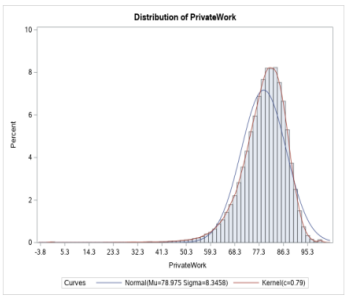

Distribution of Service
Percent
Service
Curves
Normal(Mu=19.101 Sigma=8.2791)
Kernel(c=0.79)

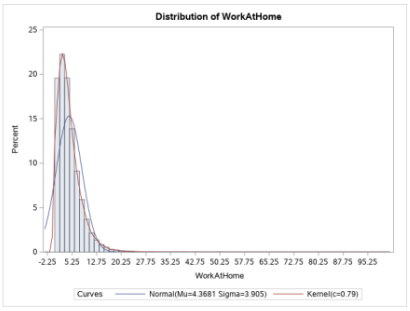

Distribution of ChildPoverty
Percent
ChildPoverty
Curves
Normal(Mu=22.487 Sigma=19.191)
Kernel(c=0.79)

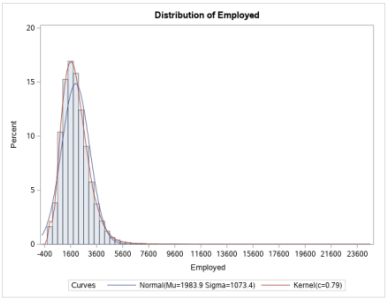

Distribution of Office
Percent
Office
Curves
Normal(Mu=23.952 Sigma=5.9573)
Kernel(c=0.79)

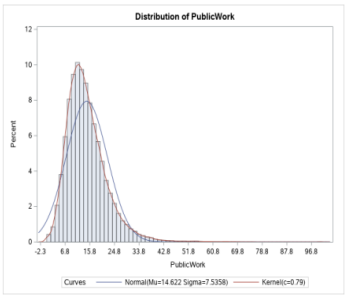

Distribution of Construction
Percent
Construction
Curves Normal(Mu=9.2923 Sigma=6.0233) Kernel(c=0.79)

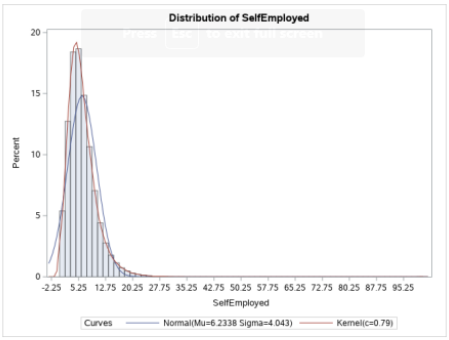

Distribution of MeanCommute
Percent
MeanCommute
Curves Normal(Mu=25.667 Sigma=6.9649) Kernel(c=0.79)

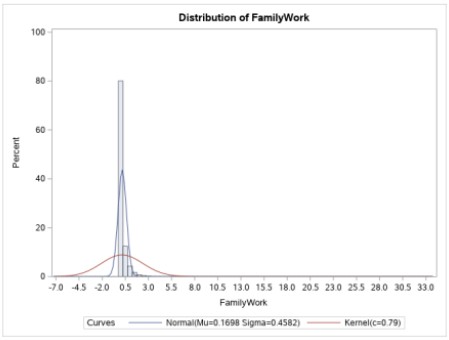

Distribution of PrivateWork
Percent
PrivateWork
Curves Normal(Mu=78.975 Sigma=8.3458) Kernel(c=0.79)

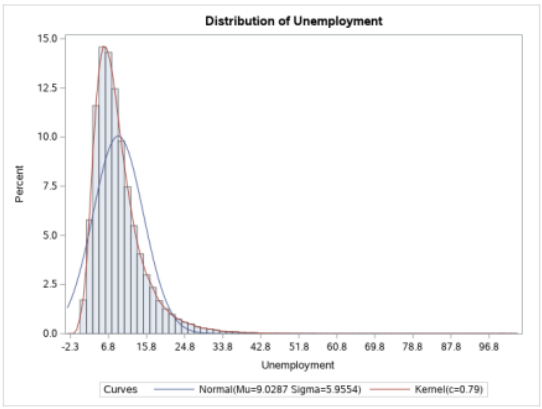

Distribution of WorkAtHome
Percent
WorkAtHome
Curves Normal(Mu=4.3681 Sigma=3.905) Kernel(c=0.79)


Distribution of Employed
Percent
Employed
Curves Normal(Mu=1983.9 Sigma=1073.4) Kernel(c=0.79)


Distribution of PublicWork
Percent
PublicWork
Curves Normal(Mu=14.622 Sigma=7.5358) Kernel(c=0.79)


Distribution of SelfEmployed
Percent
SelfEmployed
Curves Normal(Mu=6.2338 Sigma=4.043) Kernel(c=0.79)


Distribution of FamilyWork
Percent
FamilyWork
Curves Normal(Mu=0.1698 Sigma=0.4582) Kernel(c=0.79)


Distribution of Unemployment
Percent
Unemployment
Curves Normal(Mu=9.0287 Sigma=5.9554) Kernel(c=0.79)

2. Data preprocessing

Data preprocessing is a method of transforming row data into a comprehensible format. The real-world data is often content with the missing value. In the previous section, the outlier and missing value are detected in the data. This section provides how to handle the missing value and the outlier.

a. Missing Values Imputation

Missing values are contributed due to information flow interruptions, privacy considerations, and alternative factors. To complete the missing value in this dataset, the median imputation method is used. The missing value will be replaced by the column's median using the following codes in SAS.

```
data income_data;
set work.import;
if White =' ' then White= 71.40;
if Black= ' ' then Black = 3.70;
if Income= ' ' then Income= 51094;
if Poverty= ' ' then Poverty= 13.40;
if ChildPoverty= ' ' then ChildPoverty= 17.80;
if Professional=' ' then Professional= 32.60;
if Service=' ' then Service= 17.90;
if Office=' ' then Office= 23.80;
if Construction=' ' then Construction= 8.40;
if WorkAtHome=' ' then WorkAtHome= 3.50;
if MeanCommute=' ' then MeanCommute= 25;
IF PrivateWork=' ' then PrivateWork= 80.10;
if PublicWork=' ' then PublicWork= 13.40;
if SelfEmployed=' ' then SelfEmployed= 5.50;
if FamilyWork=' ' then FamilyWork=0;
if Unemployment=' ' then Unemployment= 7.70;
run;
```

Figure 6: Missing Value Imputation Code

A new data is created by imputing the missing values and all the missing values have been imputed. The result is shown in the figure below. It indicates that there is no more missing value in the dataset.

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/14/22, 5:49 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS01-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

ods noproctitle;

proc format;
    value _nmissprint low-high="Non-missing";
run;

proc freq data=WORK.INCOME_DATA;
    title3 "Missing Data Frequencies";
    title4 h=2 "Legend: ., A, B, etc = Missing";
    format TotalPop Men Women White Black Citizen Income Poverty ChildPoverty
        Professional Service Office Construction WorkAtHome MeanCommute Employed
        PrivateWork PublicWork SelfEmployed FamilyWork Unemployment _nmissprint.;
    tables TotalPop Men Women White Black Citizen Income Poverty ChildPoverty
        Professional Service Office Construction WorkAtHome MeanCommute Employed
        PrivateWork PublicWork SelfEmployed FamilyWork Unemployment / missing nocum;
run;

proc freq data=WORK.INCOME_DATA noprint;
    table TotalPop * Men * Women * White * Black * Citizen * Income * Poverty *
        ChildPoverty * Professional * Service * Office * Construction * WorkAtHome *
        MeanCommute * Employed * PrivateWork * PublicWork * SelfEmployed * FamilyWork
        * Unemployment / missing out=Work._MissingData_;
    format TotalPop Men Women White Black Citizen Income Poverty ChildPoverty
        Professional Service Office Construction WorkAtHome MeanCommute Employed
        PrivateWork PublicWork SelfEmployed FamilyWork Unemployment _nmissprint.;
run;
```

```
proc print data=Work._MissingData_ noobs label;
    title3 "Missing Data Patterns across Variables";
    title4 h=2 "Legend: ., A, B, etc = Missing";
    format TotalPop Men Women White Black Citizen Income Poverty ChildPoverty
        Professional Service Office Construction WorkAtHome MeanCommute Employed
        PrivateWork PublicWork SelfEmployed FamilyWork Unemployment _nmissprint.;
    label count="Frequency" percent="Percent";
run;

title3;

/* Clean up */
proc delete data=Work._MissingData_;
run;
```

Missing Data Patterns across Variables
Legend: ., A, B, etc = Missing

TotalPop	Men	Women	White	Black	Citizen	Income	Poverty	ChildPoverty	Professional	Service	Office	Construction	WorkAtHome	MeanCommute
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing

Employed	PrivateWork	PublicWork	SelfEmployed	FamilyWork	Unemployment	Frequency	Percent
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	74001	100

b. Handling Outlier

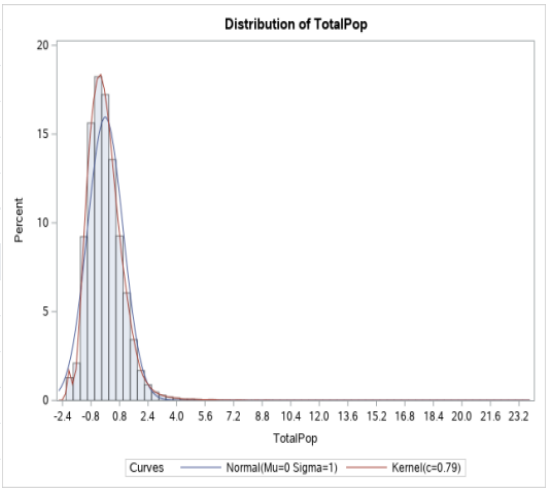
Handling outliers may be a method for eliminating outliers from the data. This methodology is often used on a spread of scales to supply an additional correct data illustration (Patel, 2022). Therefore, standardizing data is the best method. Because it scales the value while taking to the standard deviation. And so the outlier will be handled.

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/15/22, 2:41 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS02-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.4.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.4896.75 Safari/537.36 Edg/100.0.1185.39'
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST-199913-ypXIIxXc4bxHgfs1UmTv-cas'
 *
 */

ods noproctitle;

proc stdize data=WORK.INCOME_DATA method=std nomiss out=work.Stdizedata;
    var TotalPop Men Women White Black Citizen Income Poverty ChildPoverty
        Professional Service Office Construction WorkAtHome MeanCommute Employed
        PrivateWork PublicWork SelfEmployed FamilyWork Unemployment;
run;
```

Using the above code, the data is normalized. Here we presented three variables of them, and the following histogram shows that those variables are normally distributed, the cause means is equal to 0, and the standard deviation equal to 1.

Total rows: 74001 Total columns: 21 Rows 1-100

TotalPop	Men	Women
-1.116603465	-1.10753832	-1.085980705
-1.018919096	-0.996565229	-1.004756385
-0.63757435	-0.712138399	-0.54205155
0.0457465924	0.0413595632	0.048421879
3.023241284	2.6058637659	3.3247736714
-0.222885421	-0.317671025	-0.122240457
-0.734789082	-0.855750634	-0.590421089
-0.534723982	-0.583446915	-0.468128292
3.0946260148	3.1318202641	2.9487689532
0.6304438938	0.7174561257	0.5229909409
-0.488229979	-0.361500733	-0.594984253
-0.484003252	-0.43050921	-0.51923573



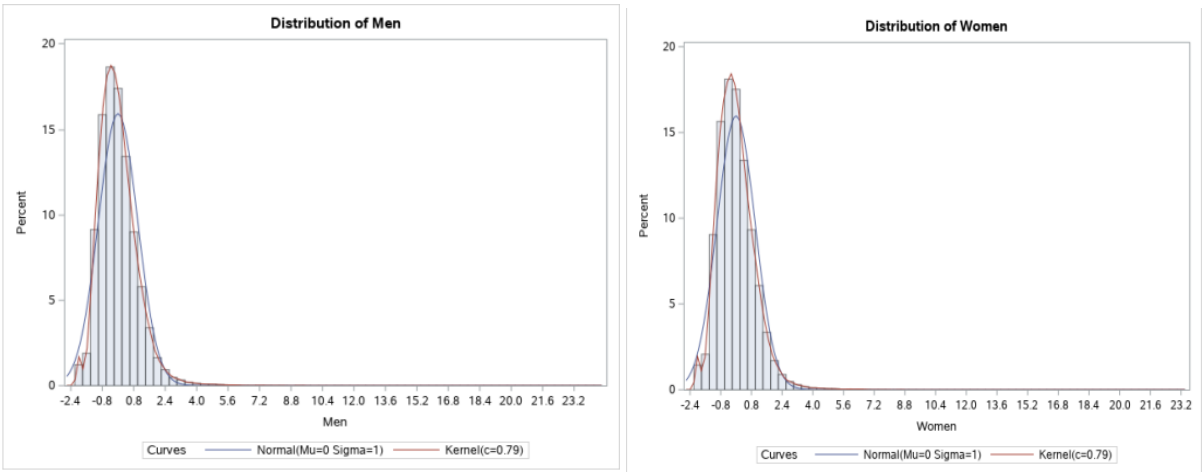


Figure 7: Standardization

3. Feature Engineering

Feature engineering is the method of choosing, manipulating, and transforming raw data into a feature that will be employed in supervised learning. To create machine learning work well on new tasks, it would be necessary to design and train better features (Patel, 2022). A generic feature engineering technique is mentioned during this sub-section: Feature Transformation.

a. Feature Transformation

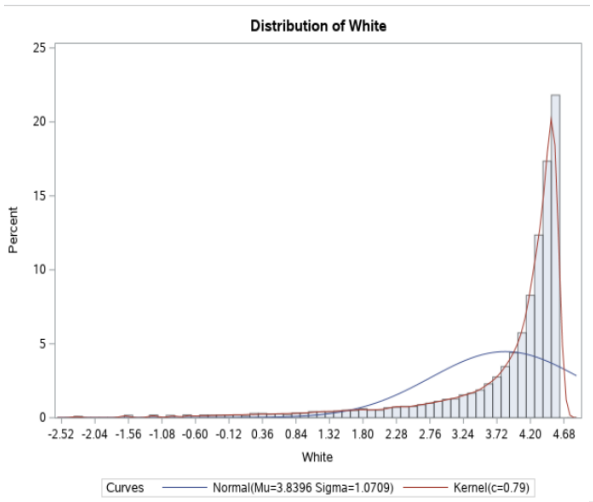
Once an imbalance is known, each attempt ought to be created to convert it into a standard distribution, so the study parametric tests are often applied for analysis. This could be accomplished by transformation (Manikandan, 2010).

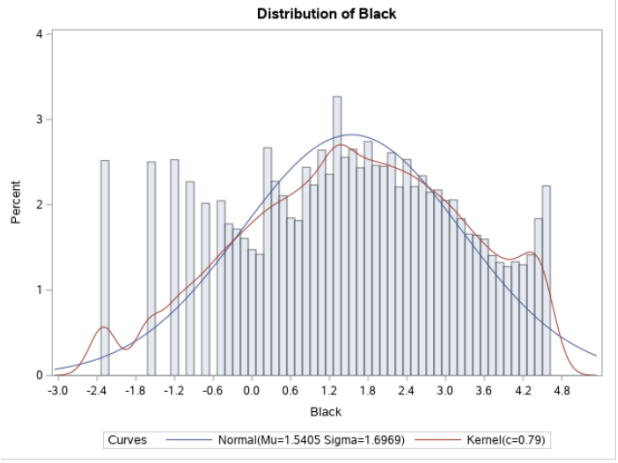
- Log transformation

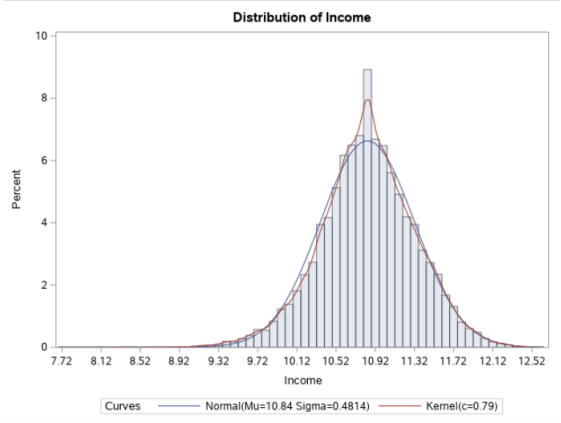
Log transformation is the most used method amongst data scientists. It is largely accustomed flip a skewed distribution into a standard or less skewed distribution. We get the log of the values in a column and use those values because of the column during this transform. It is accustomed carry out confusing data, and the information becomes a lot of approximate to the traditional application (Patel, 2022b). using the following code, data is less skewed.

```
data work.transform;
    set WORK.INCOME_DATA;
    TotalPop=log(TotalPop);
    Men=log(Men);
    Women=log(Women);
    White=log(White);
    Black=log(Black);
    Citezen=log(Citizen);
    Poverty=log(Poverty);
    ChildPoverty=log(ChildPoverty);
    Professional=log(Professional);
    Service=log(Service);
    Office=log(Office);
    Construction=log(Construction);
    WorkAtHome=log(WorkAtHome);
    MeanCommute=log(MeanCommute);
    Employed=log(Employed);
    PrivateWork=log(PrivateWork);
    PublicWork=log(PublicWork);
    SelfEmloyed=log(SelfEmployed);
    FamilyWork=log(FamilyWork);
    Unemployment=log(Unemployment);
    Income=log(Income);
run;
```

Figure 8: Code for Log Transformation

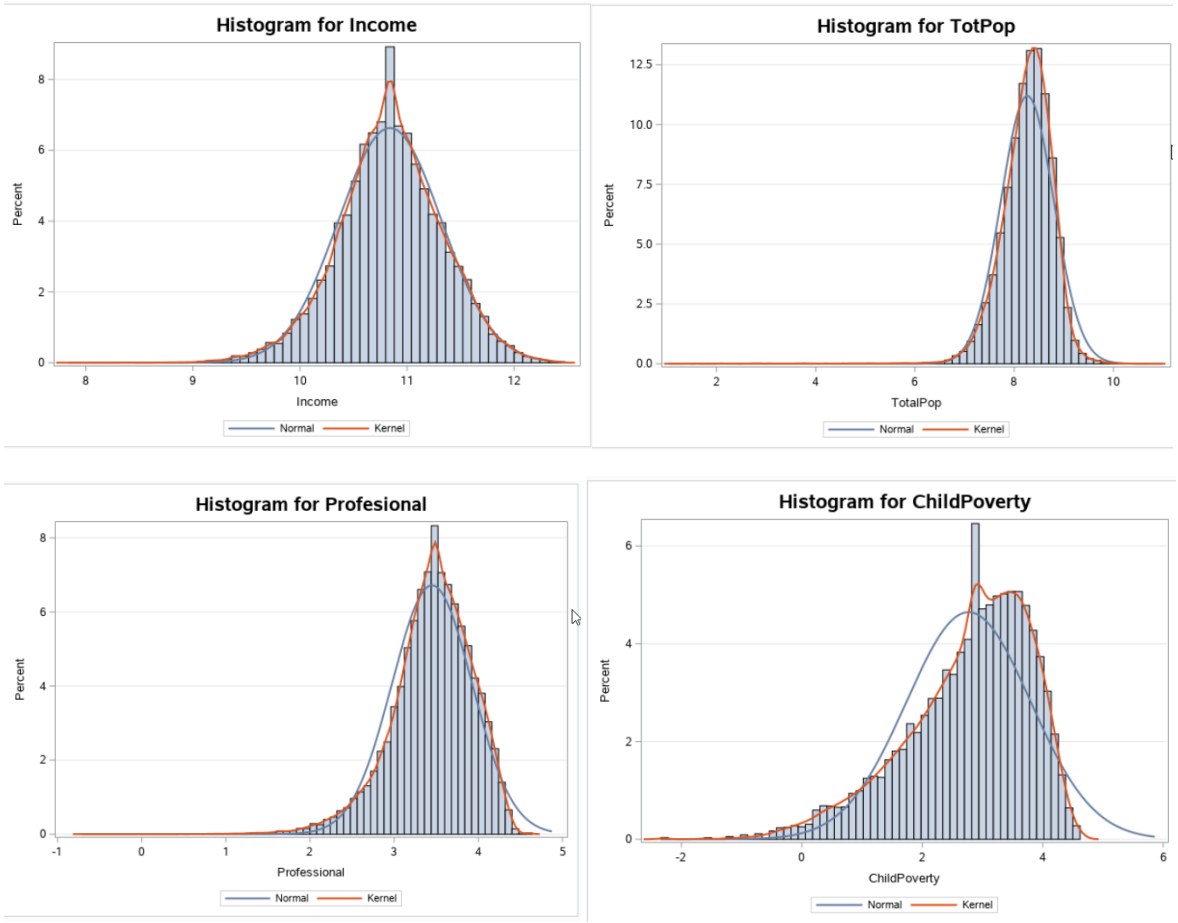






4. Exploratory Data Analysis

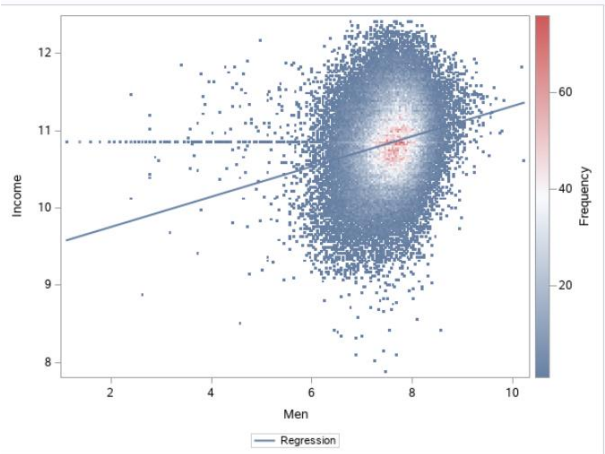
Exploratory data analysis (EDA) is defined as the crucial method of activity initial investigation of the dataset. To visualize the data, a clean dataset is used. Therefore, histogram and boxplot are used for univariate analysis, and scatter plot for bivariate and identify the relationship between income variables and the independent variable using the correlation



The histogram above explains that the data are normally distrusted.

Realizing the bivariate analysis with scatter plot, the heatmap, and correlation, the result of the heatmap with regression line shows that there is no relationship between income and men, which means the income rate does not depend on men.

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/15/22, 4:38 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS02-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.4.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.TRANSFORM;
    heatmap x=Men y=Income / name='HeatMap';
    reg x=Men y=Income / nomarkers;
    gradlegend 'HeatMap';
    keylegend / linelength=20 fillheight=2.5pct fillaspect=golden;
run;

ods graphics / reset;
```



The scatter plot with the regression line below demonstrates that poverty and income are related, it means there is a relation between income and poverty, in another word, it means income depends on the poverty rates.

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/15/22, 5:16 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS02-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.4.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.TRANSFORM;
    reg x=Income y=Poverty / nomarkers;
    scatter x=Income y=Poverty /;
    xaxis grid;
    yaxis grid;
run;

ods graphics / reset;
```

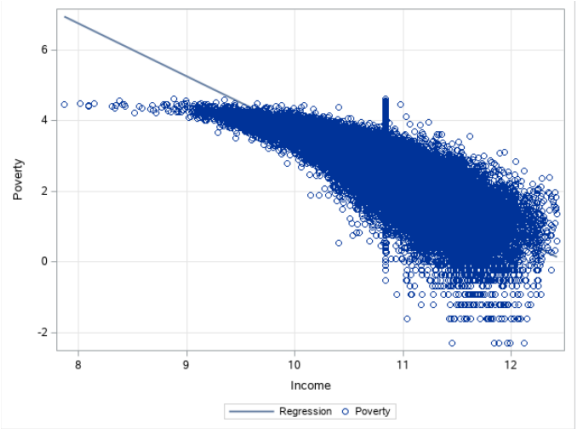


Figure 9: Scatter plot for income and poverty

Based on the correlation analysis between the target variable and independent variable, the result demonstrates there is a high negative correlation between income and child poverty, poverty, and service, and a high correlation positive between income with professionals. And no significant correlation between income and office, privateWork, publicWork, selfEmployed, FamilyWork, TotPop, Men, Women, Citezen.

20 With Variables:	TotalPop Men Women White Black Citizen Poverty ChildPoverty Professional Service Office Construction WorkAtHome MeanCommute Employed PrivateWork PublicWork SelfEmployed FamilyWork Unemployment
1 Variables:	Income

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/15/22, 4:47 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS02-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.4.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.48
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc corr data=WORK.INCOME_DATA pearson nosimple noprob plots=none;
    var Income;
    with TotalPop Men Women White Black Citizen Poverty ChildPoverty Professional
        Service Office Construction WorkAtHome MeanCommute Employed PrivateWork
        PublicWork SelfEmployed FamilyWork Unemployment;
run;
```

Pearson Correlation Coefficients, N = 74001	
	Income
TotalPop	0.17526
Men	0.17664
Women	0.16771
White	0.31266
Black	-0.30808
Citizen	0.20379
Poverty	-0.69026
ChildPoverty	-0.65862
Professional	0.72614
Service	-0.57671
Office	-0.06619
Construction	-0.32542
WorkAtHome	0.35054
MeanCommute	0.22960
Employed	0.31584
PrivateWork	-0.03426
PublicWork	-0.00803
SelfEmployed	0.08565
FamilyWork	0.00554
Unemployment	-0.47204

Figure 10: Correlation Analysis

IV. Discussion

The US census bureau, which is in charge of producing data concerning the American population and surveys, gives out a Census Demographic Data. This report is concerned with analyses of this data, it aims to clarify the relationship between the income inequality based on the race, gender, poverty, job, child poverty, and employment. So, the income variable is identified as a predictive variable and the rest variable as input features.

The descriptive analysis shows that the data contain 21 variables and 74000 observations with the presence of missing values. Then we identified the central tendency and spread of the data, for that we focused on the mean and standard deviation of each variable. To discover how the data is distributed, we first compared the mean and median values. In our result, the mean turning for the number of women is 2197.94, whereas the median is 2066. This suggests that the mean is higher than the median values, so the data appear to be skewed to the right. In other words, the data is not normally distributed. Based on the spread of data, we evaluate the standard deviation. The higher the standard deviation is, the better the data is spread. Hence, here the standard deviation of the percentage of Men is 1072.33, whereas, in the normal distribution, the observation is unfolded among 3 standard deviations on both sides of the mean, this is not the case with our data. To summarize, the present data is not normally distributed and well speeded, so the existence of an outlier is found. In addition to that, the histogram with normal curve and kernel curve is plotted, as well as the boxplot for all the variables to identify the outlier visually. From the output of the histogram, we observed that the data is whether skewed right or left for all the variables unless for the variables Office and MeanCommute which seems normally distributed. And with the boxplot of the number of Men, number of Women, also Total Population, we found that some data values are far from the rest data values. Therefore, the existence of an outlier is also discovered visually. However, we used several methods to eliminate the outlier and the missing values from the dataset. The missing values are imputed by the column's median using a SAS code (found on the SAS program file) and to handle the outlier, a feature engineering method is used which is log transformation and standardization. To eliminate the outlier, we normalized the data. Once the data is normalized, the data is scaled while taking to the standard deviation, so the outlier is treated. And for the distribution of data, log transformation distributed the data in a less skewed.

Once, the data is managed, we applied exploratory analysis to visualize the clarity of the data. The univariate analysis demonstrates that data is normally distributed. And the bivariate analysis clarifies the miss of outlier. Additionally, as, this study aims to recognize the relationship between the household income and the employment, poverty, gender, and so on. the relation between the variables is treated by displaying, the correlation analysis, scatterplot, and heatmap. The result of correlation analysis demonstrates that the income has a high positive correlation with the professional, which means the median household income depends on the percentage of the employed in management, business, science, and art. And while the percentage of employed in the science, management, and art increase, the median household income increases as well. Compared to the percentage of the public & private jobs, self-employed and those unpaid family work, who has any effect on the household median revenues. On the other hand, the income has a high negative correlation with childPoverty,

Poverty, and Service, which explain the that, the increase in the household income, will decrease the percentage of children under poverty level, the percentage of poverty, as well as the percentage of employed in the service job.

V. Conclusion

Analyzing US census bureau data which is based on the demographic and economic of the United States, to assess income inequality, which is a ways income is unfold through the population. Therefore, an understanding of real-world data was required as the information was missing due to interruption, privacy concerns, and alternative reasons. The existence of missing values and the outlier has been processed using median imputation and feature engineering methods respectively through this research. And finally, the research concludes that the household income depends on the percentage employed in management, science, and art. It also negatively affects the percentage of children under poverty levels, as well percentage of poverty levels. The household income has no effect on the gender, percentage total of population, and workplace.

REFERENCES

DeNavas-Walt, C. and, & Proctor, B. D. (2014a). Income and Poverty in the United States : 2013 Current Population Reports. *Current Population Reports, September*, P60-249.

DeNavas-Walt, C. and, & Proctor, B. D. (2014b). Income and Poverty in the United States : 2013 Current Population Reports. *Current Population Reports, 266*(September), P60-249.

DeNavas-Walt, C. and, & Proctor, B. D. (2014c). Income and Poverty in the United States : 2013 Current Population Reports. *Current Population Reports, 270*(September), P60-249.

How Income Inequality Works. (2021, November 2). Investopedia. https://www.investopedia.com/terms/i/income-inequality.asp#:%7E:text=Income%20inequality%20is%20how%20unevenly,the%20uneven%20distribution%20of%20wealth.

Inequality - Poverty rate - OECD Data. (n.d.). theOECD. https://data.oecd.org/inequality/poverty-rate.htm

Manikandan, S. (2010). Data transformation. *Journal of Pharmacology and Pharmacotherapeutics*, *1*(2), 126. https://doi.org/10.4103/0976-500x.72373

McIntosh, K., Moss, E., Nunn, R., & Shambaugh, J. (2020). Examining the Black-white wealth gap. *Brookings Instiution*, 1–9. https://www.brookings.edu/blog/up-front/2020/02/27/examining-the-black-white-wealth-gap/

Omar, M. A., & Inaba, K. (2020). Does financial inclusion reduce poverty and income inequality in developing countries? A panel data analysis. *Journal of Economic Structures*, *9*(1). https://doi.org/10.1186/s40008-020-00214-4

Patel, H. (2022a, January 5). *What is Feature Engineering — Importance, Tools and Techniques for Machine Learning*. Medium. https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10

Patel, H. (2022b, January 5). *What is Feature Engineering — Importance, Tools and Techniques for Machine Learning*. Medium. https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10

Pew Research Centre. (2012). *Appendix 2: methodology for income and wealth analysis*. 107–117. www.pewsocialtrends.org

Sources, D., Social, A., & Statistics, L. (2011). *Appendix 2 : Methodology for Income and Wealth*. 107–117.

U.S. Census Bureau. (2021, January 26). Investopedia. https://www.investopedia.com/terms/b/bureauofcensus.asp#:%7E:text=The%20U.S.%20Census%20Bureau%20is%20responsible%20for%20conducting%20the%20national,the%20economy%20and%20economic%20activity.

Weden, M. M., Peterson, C. E., Miles, J. N., & Shih, R. A. (2015). Evaluating Linearly Interpolated Intercensal Estimates of Demographic and Socioeconomic Characteristics of U.S. Counties and Census Tracts 2001–2009. *Population Research and Policy Review*, *34*(4), 541–559. https://doi.org/10.1007/s11113-015-9359-8

SAS CODE

Data importation

```
/* Generated Code (IMPORT) */
/* Source File: assignment data.csv */
/* Source Path: /home/u60775558/sasuser.v94 */
/* Code generated on: 4/15/22, 11:01 PM */
%web_drop_table(WORK.IMPORT);
FILENAME REFFILE '/home/u60775558/sasuser.v94/assignment data.csv';
PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=WORK.IMPORT;
        GETNAMES=YES;
RUN;
PROC CONTENTS DATA=WORK.IMPORT; RUN;
%web_open_table(WORK.IMPORT);
```

Data identification

```
*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/15/22, 11:07 PM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS02-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.4.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/100.0.4896.75 Safari/537.36 Edg/100.0.1185.39'
```

```
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST-50836-gr2YwITEegfeGg0lEtnP-cas'
 *
 */
ods noproctitle;
ods select attributes variables;
proc datasets;
	contents data=WORK.IMPORT order=collate;
quit;
```

list of tables

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/15/22, 11:10 PM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS02-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.4.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.4896.75 Safari/537.36 Edg/100.0.1185.39'
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST-50836-gr2YwITEegfeGg0lEtnP-cas'
 *
 */

title1 'List Data for WORK.IMPORT';
```

```
proc print data=WORK.IMPORT
        (obs=20) label;
        var TotalPop Men Women White Black Citizen Income Poverty ChildPoverty
                Professional Service Office Construction WorkAtHome MeanCommute Employed
                PrivateWork PublicWork SelfEmployed FamilyWork Unemployment;
run;
title1;
```

descriptive statistic

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/15/22, 11:13 PM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS02-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.4.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.4896.75 Safari/537.36 Edg/100.0.1185.39'
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST-50836-gr2YwITEegfeGg0lEtnP-cas'
 *
 */

ods noproctitle;
ods graphics / imagemap=on;
```

```
proc means data=WORK.IMPORT chartype mean std min max median n vardef=df qrange
                qmethod=os;
        var TotalPop Men Women White Black Citizen Income Poverty ChildPoverty
                Professional Service Office Construction WorkAtHome MeanCommute Employed
                PrivateWork PublicWork SelfEmployed FamilyWork Unemployment;
run;
```

Missing value imputation

```
data income_data;
set work.import;
if White =' ' then White= 71.40;
if Black= ' ' then Black = 3.70;
if Income= ' ' then Income= 51094;
if Poverty= ' ' then Poverty= 13.40;
if ChildPoverty= ' ' then ChildPoverty= 17.80;
if Professional=' ' then Professional= 32.60;
if Service=' ' then Service= 17.90;
if Office=' ' then Office= 23.80;
if Construction=' ' then Construction= 8.40;
if WorkAtHome=' ' then WorkAtHome= 3.50;
if MeanCommute=' ' then MeanCommute= 25;
IF PrivateWork=' ' then PrivateWork= 80.10;
if PublicWork=' ' then PublicWork= 13.40;
if SelfEmployed=' ' then SelfEmployed= 5.50;
if FamilyWork=' ' then FamilyWork=0;
if Unemployment=' ' then Unemployment= 7.70;
run;
```

Standardization

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/15/22, 3:30 AM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS02-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.4.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.4896.75 Safari/537.36 Edg/100.0.1185.39'
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=ST-77-wiWC0j6wsVVdWOKbL2Bo-cas'
 *
 */


ods noproctitle;
proc stdize data=WORK.INCOME_DATA method=std nomiss out=work.Stdizedata;
        var TotalPop Men Women White Black Citizen Income Poverty ChildPoverty
                Professional Service Office Construction WorkAtHome MeanCommute Employed
                PrivateWork PublicWork SelfEmployed FamilyWork Unemployment;
run;
```

Log transformation

```
data work.transform;
        set WORK.INCOME_DATA;
```

```
        TotalPop=log(TotalPop);
        Men=log(Men);
        Women=log(Women);
        White=log(White);
        Black=log(Black);
        Citezen=log(Citizen);
        Poverty=log(Poverty);
        ChildPoverty=log(ChildPoverty);
        Professional=log(Professional);
        Service=log(Service);
        Office=log(Office);
        Construction=log(Construction);
        WorkAtHome=log(WorkAtHome);
        MeanCommute=log(MeanCommute);
        Employed=log(Employed);
        PrivateWork=log(PrivateWork);
        PublicWork=log(PublicWork);
        SelfEmloyed=log(SelfEmployed);
        FamilyWork=log(FamilyWork);
        Unemployment=log(Unemployment);
        Income=log(Income);

run;
```

Heat Map

```
*
 *
 * Task code generated by SAS Studio 3.8
 *
```

```
 * Generated on '4/15/22, 11:32 PM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS02-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.4.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/100.0.4896.75 Safari/537.36 Edg/100.0.1185.39'
 * Generated on web client 'https://odamid-
apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=
ST-50836-gr2YwITEegfeGg0lEtnP-cas'
 *
 */
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=WORK.TRANSFORM;
        heatmap x=Men y=Income / name='HeatMap';
        reg x=Men y=Income / nomarkers;
        gradlegend 'HeatMap';
        keylegend / linelength=20 fillheight=2.5pct fillaspect=golden;
run;

ods graphics / reset;
```

Correlation Analysis

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '4/15/22, 11:26 PM'
 * Generated by 'u60775558'
 * Generated on server 'ODAWS02-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.4.1.el7.x86_64'
```

```
 * Generated on SAS version '9.04.01M6P11072018'

 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/100.0.4896.75 Safari/537.36 Edg/100.0.1185.39'

 * Generated on web client 'https://odamid-
apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&ticket=
ST-50836-gr2YwITEegfeGg0lEtnP-cas'

 *

 */


ods noproctitle;

ods graphics / imagemap=on;


proc corr data=WORK.TRANSFORM pearson nosimple noprob plots=none;

        var Income;

        with TotalPop Men Women White Black Citizen Poverty ChildPoverty Professional

                Service Office Construction WorkAtHome MeanCommute Employed
PrivateWork

                PublicWork SelfEmployed FamilyWork Unemployment Citezen SelfEmloyed;

run;


Recode Income Variable

/*

 *

 * Task code generated by SAS Studio 3.8

 *

 * Generated on '4/14/22, 9:16 PM'

 * Generated by 'u60775558'

 * Generated on server 'ODAWS02-APSE1.ODA.SAS.COM'

 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.4.1.el7.x86_64'

 * Generated on SAS version '9.04.01M6P11072018'
```

```
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.4896.75 Safari/537.36 Edg/100.0.1185.39'

 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B08%253A00&https%3A%2F%2Fodamid-apse1.oda.sas.com%2FSASStudio%2Findex='

 *

 */


data WORK.recoded_data;
	length Income_recode  $8;
	set WORK.INCOME_DATA;


	select;
		when (0 <=Income <=50000) Income_recode='<=50000';
		when (50001 <=Income <=100000) Income_recode='<=100000';
		otherwise Income_recode='>100000';
	end;
run;
```