INDIVIDUAL ASSIGNMENT 1

TECHNOLOGY PARK MALAYSIA

EXPLORATORY DATA ANALYSIS AND DATA PRE-PROSSING

CT075-3-2-DTM

DATA MANAGEMENT

APD2F2011CS(DA)

NAME:	**MD NOWSHAD UL ALAM**
TP NUMBER:	**TP057903**
HAND OUT DATE:	**4TH DECEMBER 2020**
HAND IN DATE:	**22ND FEBRUARY 2021**
WEIGHTAGE:	**50%**

Contents

1.0 Introduction 4

2.0 Types of Attributes 4

3.0 Exploratory data Analysis(EDA) 5

3.1 Job Type 5

3.2 Marital Status 5

3.3 Job 6

3.4 Relationship 6

3.5 Race 7

3.6 Gender 7

3.7 Salary 8

3.8 Age,YearinEd,Cgain,Closs,Workperweek 8

3.9 Qualification 10

3.10 Country 11

4.0 Pre-processing Techniques 12

4.1 Incomplete data 12

4.1.1 10TH GRADE 12

4.1.2 12th GRADE 13

4.1.3 11TH GRADE 13

4.1.4 GRADE FROM 1st-8th 14

4.1.5 Assoc-acdm 14

4.1.6 ASSOC-VOC 15

4.1.7 Bachelors degree 15

4.1.8 Masters 16

4.1.9 Country 17

4.2 Noisy data 18

4.2.1 AGE 18

4.2.2 YearinEd 19

4.2.3 Cgain 20

4.2.4 CLOSS 21

4.2.5 WorkPerweek 22

4.3 INCONSISTENT DATA 23

4.3.1 Relationship 23

5.0 Conclusion 24

References 25

1.0 Introduction

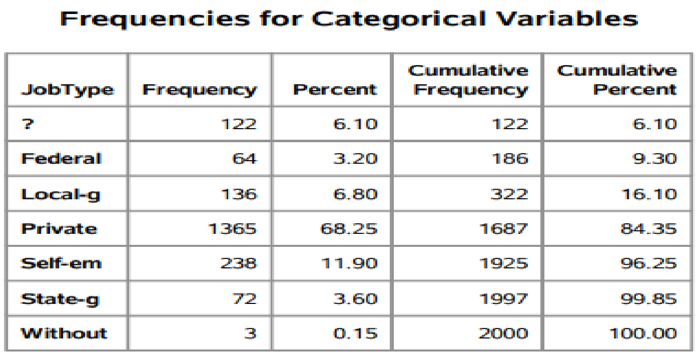
This is the documentation which is all about exploring dataset. There are 14 attributes in the given dataset including 2000 records of the people's data who lives in united states America. The information of the given dataset formatted and stored in Microsoft excel. The given dataset has many noisy, incomplete, and inconsistent data and to clean these data we will be using data pre-processing techniques. To explore the given data set we will use sas studio.

2.0 Types of Attributes

Number of Column	Name of Column	Value of data
1	Age	discrete, quantitative, Nominal
2	JobType	Qualitative, Nominal
3	Qualification	Qualitative, Nominal
4	YearinED	Qualitative ,Ordinal
5	MatritalStatus	Qualitative, Nominal
6	JobType	Qualitative, Nominal
7	Relationship	Qualitative, Nominal
8	Race	Qualitative, Nominal
9	Gender	Qualitative, Nominal
10	Cgain	Continuous, quantitative, nominal
11	Closs	Continuous, quantitative, ratio
12	WorkPerWeek	Continuous, quantitative, ratio
13	Country	Qualitative, Nominal
14	Salary	Descrete, quantitative, ordinal

3.0 Exploratory data Analysis(EDA)

3.1 Job Type



Frequencies for Categorical Variables

JobType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
?	122	6.10	122	6.10
Federal	64	3.20	186	9.30
Local-g	136	6.80	322	16.10
Private	1365	68.25	1687	84.35
Self-em	238	11.90	1925	96.25
State-g	72	3.60	1997	99.85
Without	3	0.15	2000	100.00

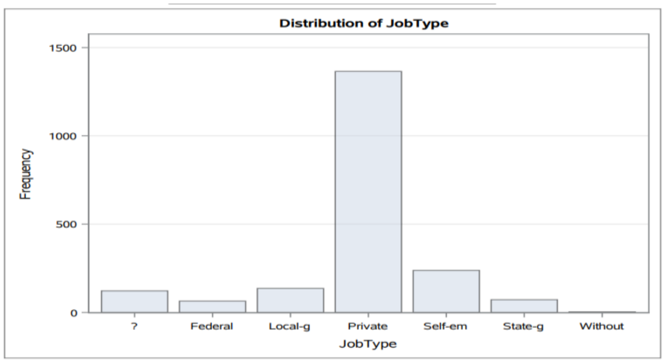
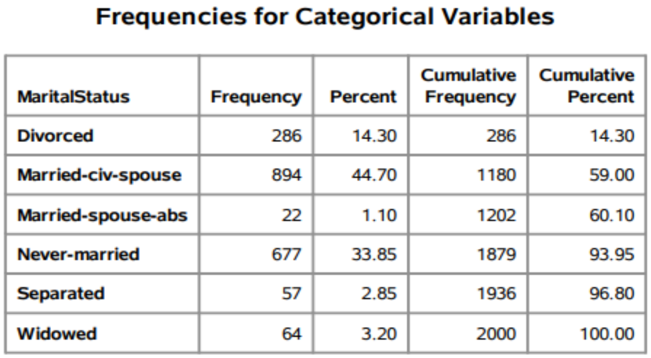


Figure- 1(frequency of jobType)

Figure- 2(Distribution of jobType)

Here, if we notice we can see figure-1 in the above, there are 7 rows of which 122 rows bears empty value in job type out of 2000 thousand rows. Here in the above figure-1 and figure -2, there are six type of job type and among these, most of the people are in private job.

3.2 Marital Status



Frequencies for Categorical Variables

MaritalStatus	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Divorced	286	14.30	286	14.30
Married-civ-spouse	894	44.70	1180	59.00
Married-spouse-abs	22	1.10	1202	60.10
Never-married	677	33.85	1879	93.95
Separated	57	2.85	1936	96.80
Widowed	64	3.20	2000	100.00

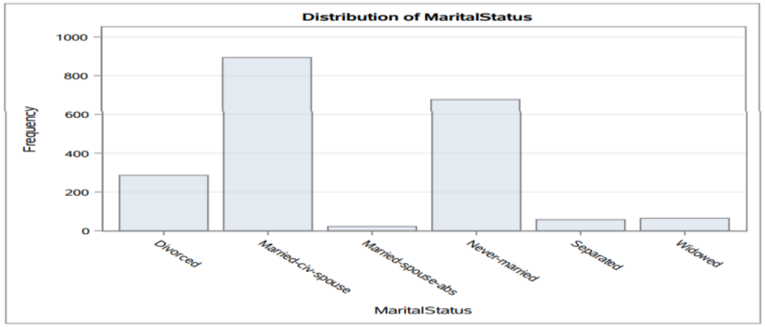
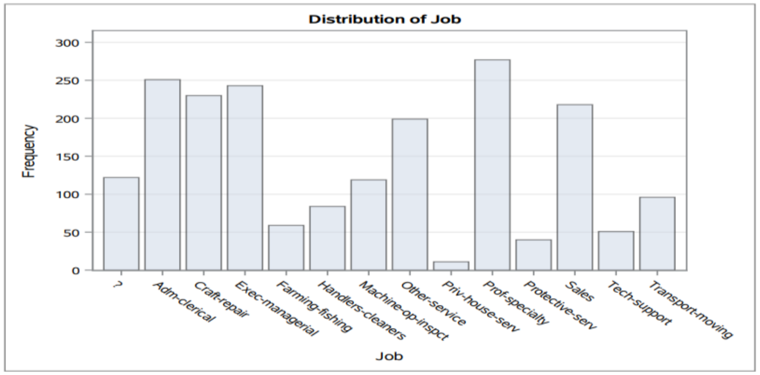


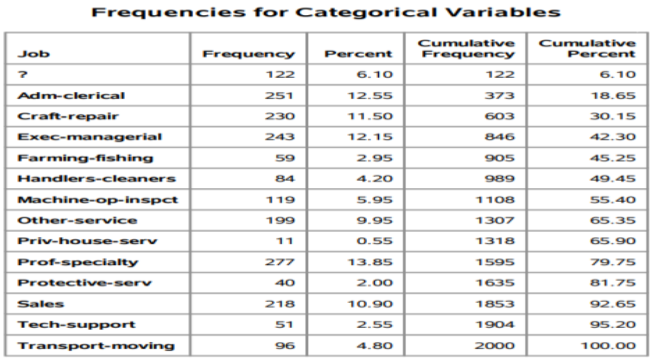
Figure- 3(frequency of MaritalStatus)

Figure- 4(Distribution of MaritalStatus)

As in the figure-3 and figure-4 shown in the above bears similar type of information which is about Marital Status. Currently there are six type of marital status in the above data where most people are in Married-civ-spouse that is about 894 and the lowest number of people are in Married-spouse-abs which is 22 out of 2000 people.

3.3 Job




Frequencies for Categorical Variables

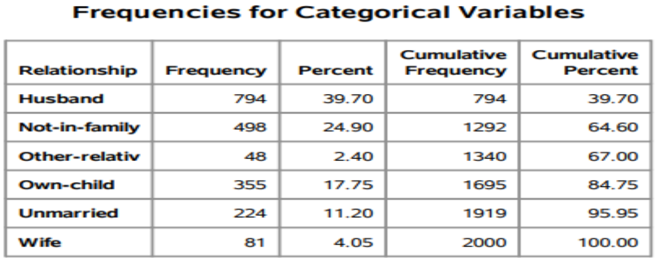
Job	Frequency	Percent	Cumulative Frequency	Cumulative Percent
?	122	6.10	122	6.10
Adm-clerical	251	12.55	373	18.65
Craft-repair	230	11.50	603	30.15
Exec-managerial	243	12.15	846	42.30
Farming-fishing	59	2.95	905	45.25
Handlers-cleaners	84	4.20	989	49.45
Machine-op-inspct	119	5.95	1108	55.40
Other-service	199	9.95	1307	65.35
Priv-house-serv	11	0.55	1318	65.90
Prof-specialty	277	13.85	1595	79.75
Protective-serv	40	2.00	1635	81.75
Sales	218	10.90	1853	92.65
Tech-support	51	2.55	1904	95.20
Transport-moving	96	4.80	2000	100.00

Figure- 5(Distribution of Job)

Figure- 6(frequency of Job)

According to the information given in the figure -5 and figure-6 respectively we can see, there are about 122 empty rows out of 2000 which is same as job type. Here in the above, there are few job which has more demand than other type of job these are- Adm-clerical, craft-repair, Exec-managerial, Machine-op-inspect, sales and prof-specialty but among them prof-specialty and Adm-clerical has high demand which are 277 and 251 respectively and the lowest demanded job are protective-serv and priv-house-serv which is 40 and 11 respectively.

3.4 Relationship


Frequencies for Categorical Variables

Relationship	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Husband	794	39.70	794	39.70
Not-in-family	498	24.90	1292	64.60
Other-relativ	48	2.40	1340	67.00
Own-child	355	17.75	1695	84.75
Unmarried	224	11.20	1919	95.95
Wife	81	4.05	2000	100.00

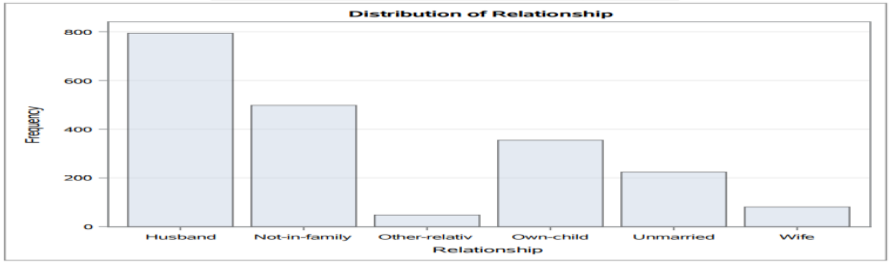
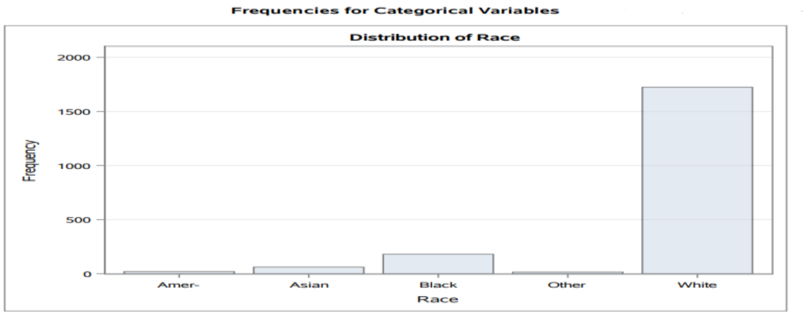


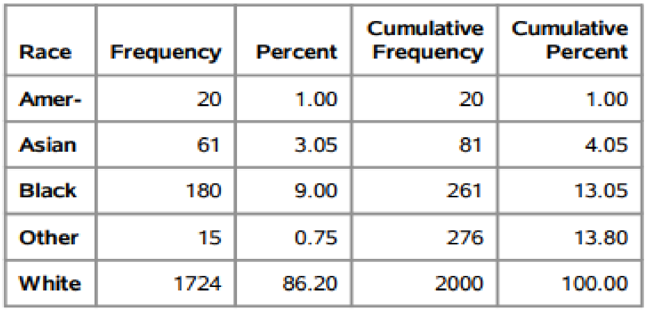
Figure- 7(frequency of Relationship)

Figure- 8(Distribution of Relationship)

Both figure – 7 and 8 shown in the above illustrate about relationship status and there are about six type of relationship status these days according to the given information in the table. From the above material, we can notice most of the employees are husband which means majority of them are married and the value is 794 and 48 people are in other other-relationship which is not mentioned.

3.5 Race





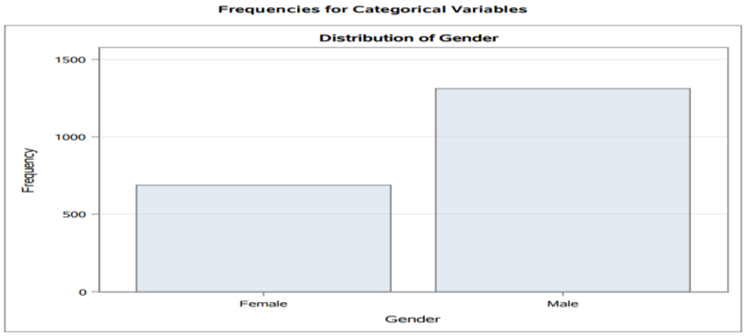
Race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Amer-	20	1.00	20	1.00
Asian	61	3.05	81	4.05
Black	180	9.00	261	13.05
Other	15	0.75	276	13.80
White	1724	86.20	2000	100.00

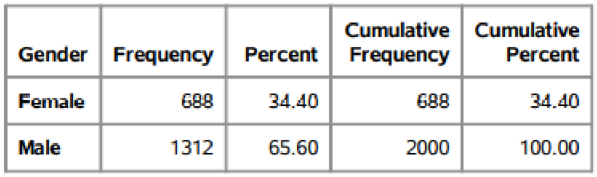
Figure- 9(Distribution of Race)

Figure- 10(Frequency of Race)

Based on the information mentioned in the above figure – 9 and 10, here most of the people belongs to the white race which is 1724 out of 2000 and almost 86.20 percent of the total population and 22 people from American race which is only 1.00 percent and less among all the race mentioned above. Again, there are about 15 people whose race is mentioned as other that is unknown.

3.6 Gender





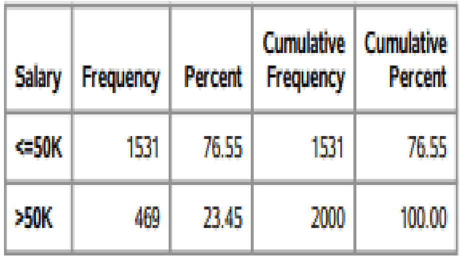
Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	688	34.40	688	34.40
Male	1312	65.60	2000	100.00

Figure- 11(Distribution of Gender)

Figure- 12(Frequency of Gender)

According to the distribution figure 11 and frequency figure-12 in the above, we can clearly see that most of the employees in the data given are male which is 1312 out of 2000 people and 65.60 percent of the total population which is almost twice than female.

3.7 Salary



Salary	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<=50K	1531	76.55	1531	76.55
>50K	469	23.45	2000	100.00

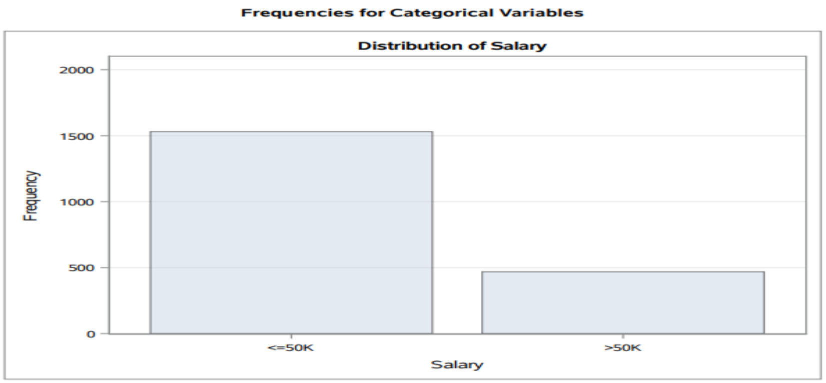
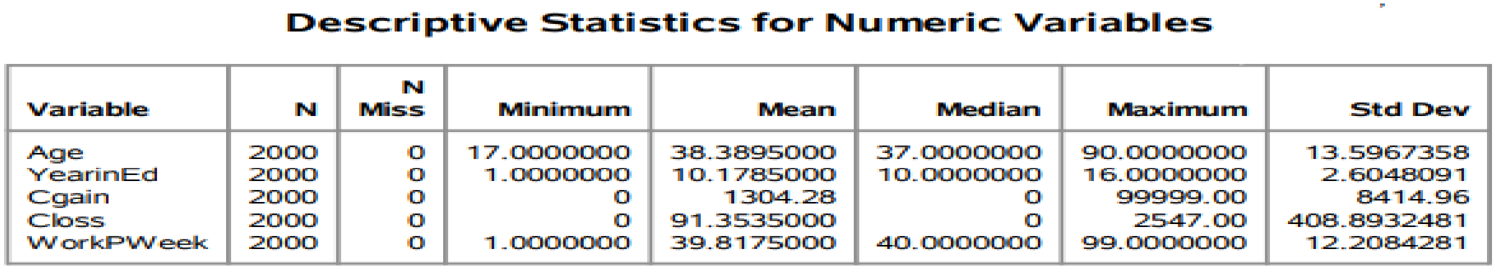


Figure- 13(frequency of Salary) *Figure- 14(Distribution of Salary)*

It is observed, both in the figure– 13 and14, majority of the people gets salary less than 50k which is 76.55 percent of the whole data and 469 people out of 2000 gets more than 50k which is almost ¼ of the total population or less than that.

3.8 Age, YearinEd, Cgain, Closs and Workperweek

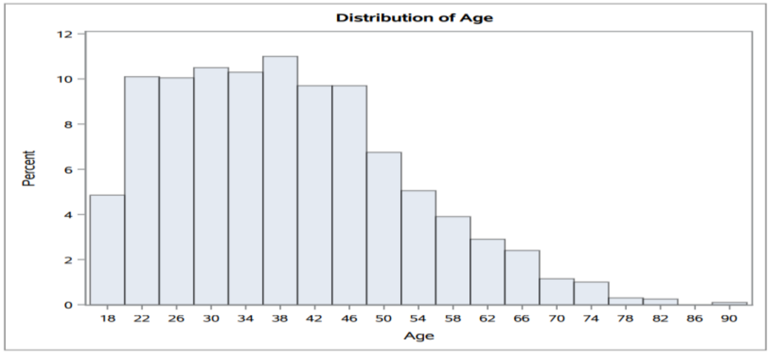


Descriptive Statistics for Numeric Variables

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
Age	2000	0	17.0000000	38.3895000	37.0000000	90.0000000	13.5967358
YearinEd	2000	0	1.0000000	10.1785000	10.0000000	16.0000000	2.6048091
Cgain	2000	0	0	1304.28	0	99999.00	8414.96
Closs	2000	0	0	91.3535000	0	2547.00	408.8932481
WorkPWeek	2000	0	1.0000000	39.8175000	40.0000000	99.0000000	12.2084281

Figure-15(mean,median,maximum,minimum and std Dev of Age,yearinEd,Cgain,Closs and workperweek)

In the above figure-15 we can see there are five different types of variable mentioned, these are: Age, yearinEd, cgain, closs, workperweek. There is no missing value and the average of these variable are 38.38, 10.17, 1304.28 , 91.35 and 39.81 respectively. Here minimum and maximum value of - age are 17.00 and 38.38 years, yearinEd are 1.00 and 10.00 years, cgain are 0 and 99999.00, closs are 0 and 2547.00 and workperWeek are 1.00 and 40.00 hours respectively.



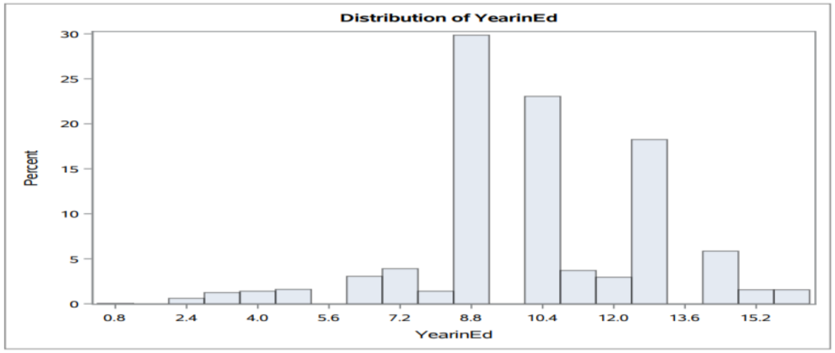
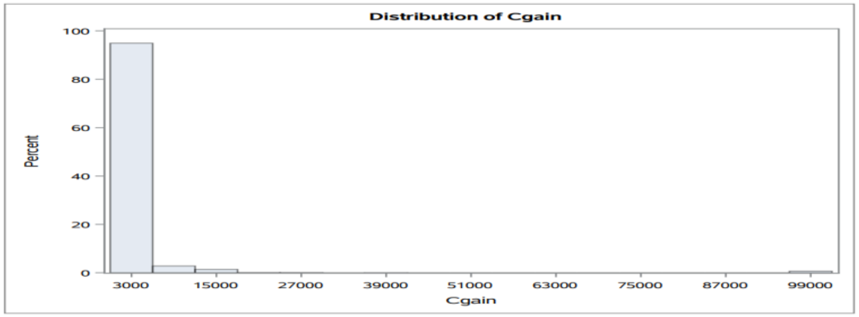


Figure- 16(distribution of Age)

Figure- 17(distribution of YearinED)

Here in the figure – 16 mentioned above , it gives data about distribution of age from which we can see, age of most of the employees are nearly 22,26,30,34,38,42 and 46 years which is together almost 69-71 percent of the whole employees and the rest of the people are from below 22 or above 46 years. Again in the figure- 17 there is a description about distribution of year in education or we can say it as time spent in doing study/ education. Here most of the people spent nearly 8.8 years which is 29 out of 100 and very few people has less than 3 years of spending time in education.



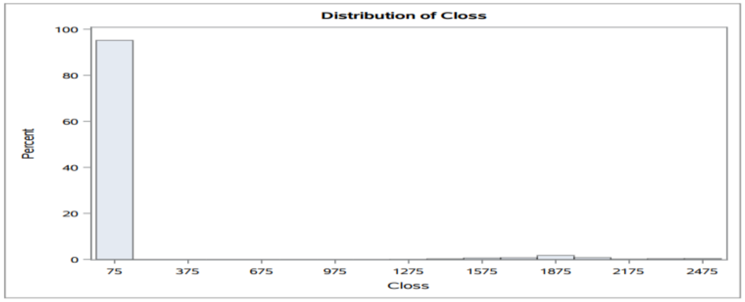


Figure- 18(distribution of Cgain)

Figure- 19(distribution of Closs)

In the above figure – 18 and 19 illustrate about capital gain and capital loss respectively. Here we can clearly see that both capital gain and capital loss is almost same that is above 90 percent respectively.

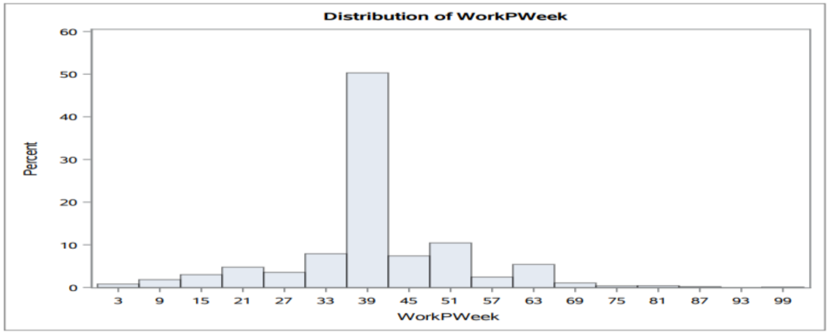
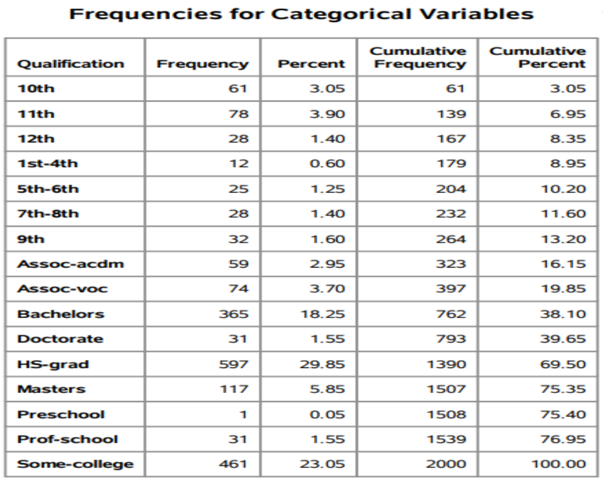


Figure- 20(distribution of Workperweek)

Figure – 20 in the above, mentioned about working hours per week. From that figure we can observe that majority of the people works more than 38 and less than 40 hours and the percentage of these people are nearly 50 percent of the total population.

3.9 Qualification



Frequencies for Categorical Variables

Qualification	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10th	61	3.05	61	3.05
11th	78	3.90	139	6.95
12th	28	1.40	167	8.35
1st-4th	12	0.60	179	8.95
5th-6th	25	1.25	204	10.20
7th-8th	28	1.40	232	11.60
9th	32	1.60	264	13.20
Assoc-acdm	59	2.95	323	16.15
Assoc-voc	74	3.70	397	19.85
Bachelors	365	18.25	762	38.10
Doctorate	31	1.55	793	39.65
HS-grad	597	29.85	1390	69.50
Masters	117	5.85	1507	75.35
Preschool	1	0.05	1508	75.40
Prof-school	31	1.55	1539	76.95
Some-college	461	23.05	2000	100.00

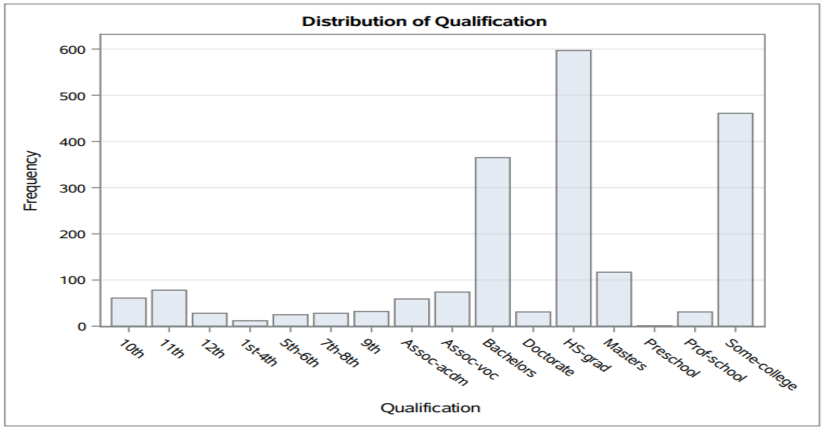
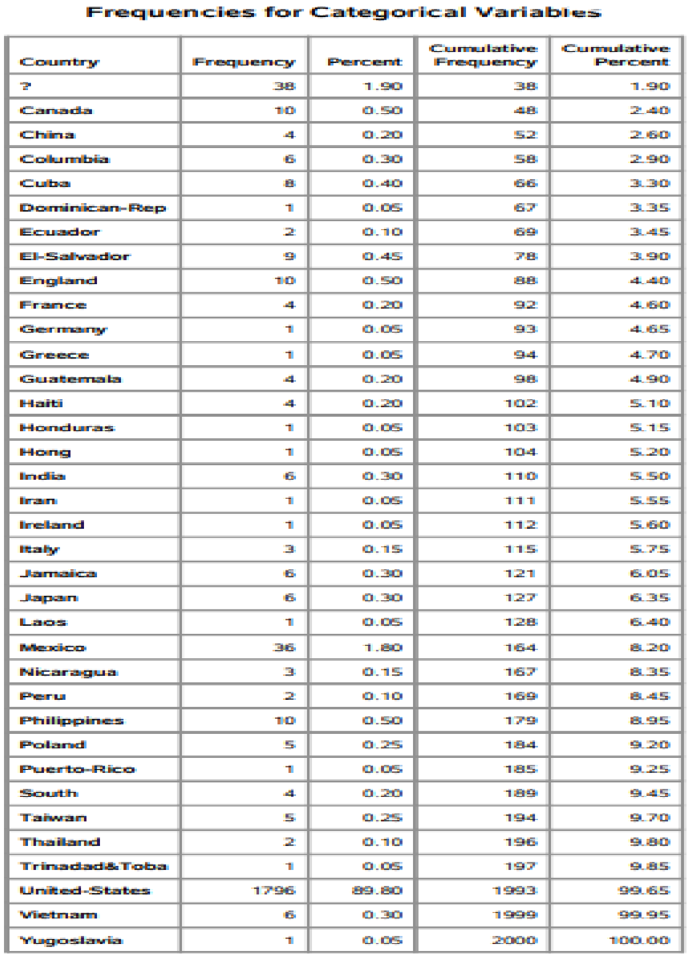


Figure- 21(frequency of Qualification)

Figure- 22(distribution of Qualification)

In the above figure 21 and figure-22 bears information about Qualification and its distribution. Here we can see lowest qualification is 1st-4th class and there are 12 people who belongs on that qualification and highest level of qualification is Doctorate and 31 people has this qualification. Here most of the people are from HS-grad qualification which 600 of the total population.

3.10 Country



Frequencies for Categorical Variables

Country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
?	38	1.90	38	1.90
Canada	10	0.50	48	2.40
China	4	0.20	52	2.60
Columbia	6	0.30	58	2.90
Cuba	8	0.40	66	3.30
Dominican-Rep	1	0.05	67	3.35
Ecuador	2	0.10	69	3.45
El-Salvador	9	0.45	78	3.90
England	10	0.50	88	4.40
France	4	0.20	92	4.60
Germany	1	0.05	93	4.65
Greece	1	0.05	94	4.70
Guatemala	4	0.20	98	4.90
Haiti	4	0.20	102	5.10
Honduras	1	0.05	103	5.15
Hong	1	0.05	104	5.20
India	6	0.30	110	5.50
Iran	1	0.05	111	5.55
Ireland	1	0.05	112	5.60
Italy	3	0.15	115	5.75
Jamaica	6	0.30	121	6.05
Japan	6	0.30	127	6.35
Laos	1	0.05	128	6.40
Mexico	36	1.80	164	8.20
Nicaragua	3	0.15	167	8.35
Peru	2	0.10	169	8.45
Philippines	10	0.50	179	8.95
Poland	5	0.25	184	9.20
Puerto-Rico	1	0.05	185	9.25
South	4	0.20	189	9.45
Taiwan	5	0.25	194	9.70
Thailand	2	0.10	196	9.80
Trinadad&Toba	1	0.05	197	9.85
United-States	1796	89.80	1993	99.65
Vietnam	6	0.30	1999	99.95
Yugoslavia	1	0.05	2000	100.00

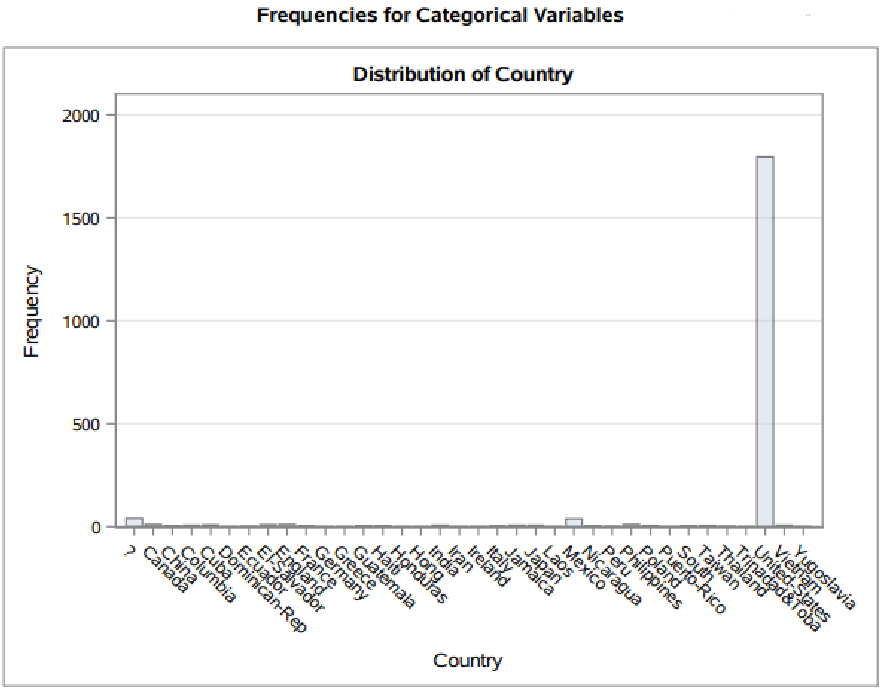


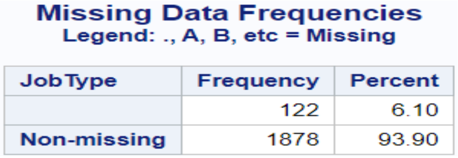
Figure- 23(frequency of Country)

Figure- 24(distribution of Country)

In the above, figure 23 and figure – 24 represent about the distribution of the country and it is clear that almost every people are from united states which about 89.80 percent of the total population and there 38 empty value in country which is unknown.

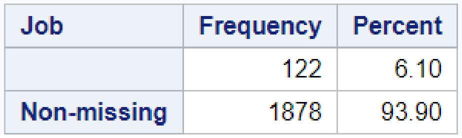
4.0 Pre-processing Techniques

4.1 Incomplete data

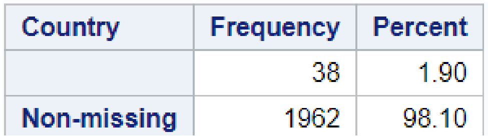


Missing Data Frequencies
Legend: ., A, B, etc = Missing

JobType	Frequency	Percent
	122	6.10
Non-missing	1878	93.90



Job	Frequency	Percent
	122	6.10
Non-missing	1878	93.90

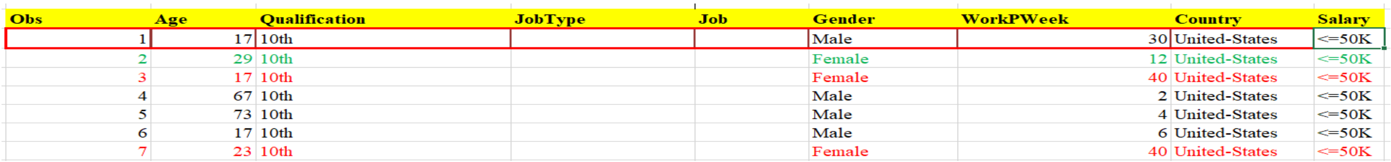


Country	Frequency	Percent
	38	1.90
Non-missing	1962	98.10

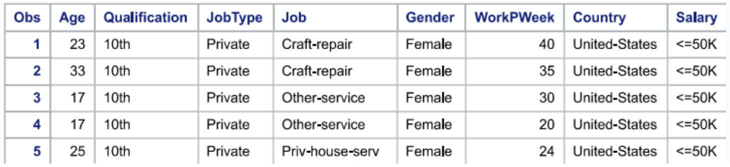
Figure – 25(Frequency of missing data)

Figure in the above shows us the number of missing records of the attributes of the given data set. According the dataset there are 38 missing records in country attributes and both jobtype and job have 122 missing records respectively.

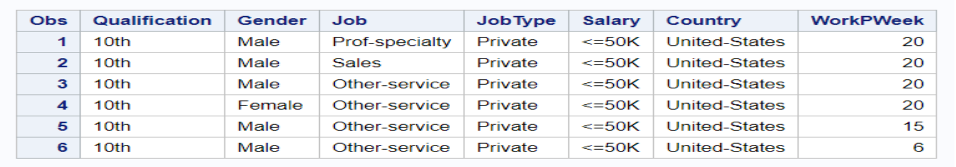
4.1.1 10TH grade



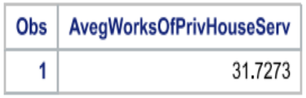
Obs	Age	Qualification	JobType	Job	Gender	WorkPWeek	Country	Salary
1	17	10th			Male	30	United-States	<=50K
2	29	10th			Female	12	United-States	<=50K
3	17	10th			Female	40	United-States	<=50K
4	67	10th			Male	2	United-States	<=50K
5	73	10th			Male	4	United-States	<=50K
6	17	10th			Male	6	United-States	<=50K
7	23	10th			Female	40	United-States	<=50K



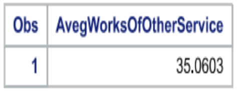
Obs	Age	Qualification	JobType	Job	Gender	WorkPWeek	Country	Salary
1	23	10th	Private	Craft-repair	Female	40	United-States	<=50K
2	33	10th	Private	Craft-repair	Female	35	United-States	<=50K
3	17	10th	Private	Other-service	Female	30	United-States	<=50K
4	17	10th	Private	Other-service	Female	20	United-States	<=50K
5	25	10th	Private	Priv-house-serv	Female	24	United-States	<=50K



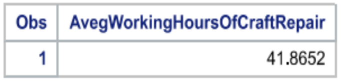
Obs	Qualification	Gender	Job	JobType	Salary	Country	WorkPWeek
1	10th	Male	Prof-specialty	Private	<=50K	United-States	20
2	10th	Male	Sales	Private	<=50K	United-States	20
3	10th	Male	Other-service	Private	<=50K	United-States	20
4	10th	Female	Other-service	Private	<=50K	United-States	20
5	10th	Male	Other-service	Private	<=50K	United-States	15
6	10th	Male	Other-service	Private	<=50K	United-States	6



Obs	AvegWorksOfPrivHouseServ
1	31.7273



Obs	AvegWorksOfOtherService
1	35.0603

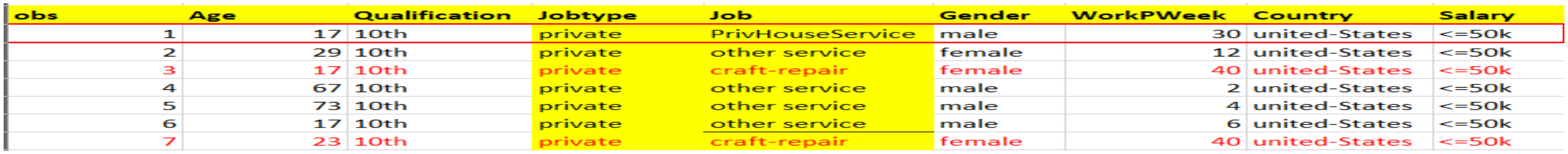


Obs	AvegWorkingHoursOfCraftRepair
1	41.8652

Figure – 26(missing records in jobtype and job of 10th grade)

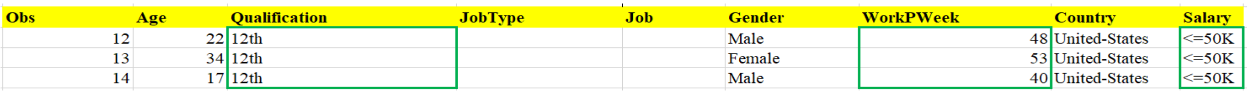
Here observation number 1 might have job of **PrivHouseService** as its average and salary are nearly same as this category and jobtype must be private as most of the people who works as **PrivHouseService** their job type is private . Obsevation number 3 and 7 will be **craft-repair** because people who study in 10th grade,who works nearly 40hours and salary is less than 50k are in carft-repair and average working hours of carft-repair is also nearly same as 40hours and similarly their job type is private. And all other coloumn are under category of **other service** because people who are in 10th grade and do working hourly rate less than 20 are in other service and their job type is private.

After modification

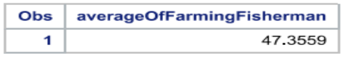


obs	Age	Qualification	Jobtype	Job	Gender	WorkPWeek	Country	Salary
1	17	10th	private	PrivHouseService	male	30	united-States	<=50k
2	29	10th	private	other service	female	12	united-States	<=50k
3	17	10th	private	craft-repair	female	40	united-States	<=50k
4	67	10th	private	other service	male	2	united-States	<=50k
5	73	10th	private	other service	male	4	united-States	<=50k
6	17	10th	private	other service	male	6	united-States	<=50k
7	23	10th	private	craft-repair	female	40	united-States	<=50k

4.1.2 12th grade



Obs	Age	Qualification	JobType	Job	Gender	WorkPWeek	Country	Salary
12	22	12th			Male	48	United-States	<=50K
13	34	12th			Female	53	United-States	<=50K
14	17	12th			Male	40	United-States	<=50K

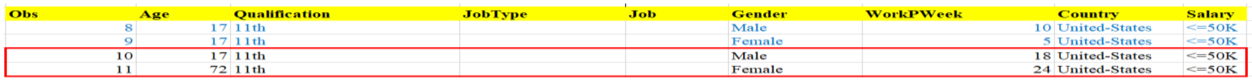


Obs	averageOfFarmingFisherman
1	47.3559

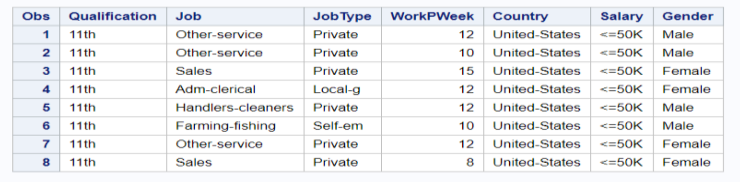
Figure – 27(missing records in jobtype and job of 12th grade)

In the above table, people under observation no 22nd and 23th works more than 45 hours and earn less than 50k which is nearly average and salary of the people under **framing and fisherman**(47.35Avg) and their jobtype is self-employee again people under row number 17th can be in **Craft-repair** as its average is 41.8652 and jobtype is private.

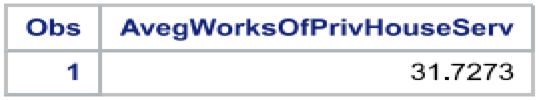
4.1.3 11TH grade



Obs	Age	Qualification	JobType	Job	Gender	WorkPWeek	Country	Salary
8	17	11th			Male	10	United-States	<=50K
9	17	11th			Female	5	United-States	<=50K
10	17	11th			Male	18	United-States	<=50K
11	72	11th			Female	24	United-States	<=50K



Obs	Qualification	Job	JobType	WorkPWeek	Country	Salary	Gender
1	11th	Other-service	Private	12	United-States	<=50K	Male
2	11th	Other-service	Private	10	United-States	<=50K	Male
3	11th	Sales	Private	15	United-States	<=50K	Female
4	11th	Adm-clerical	Local-g	12	United-States	<=50K	Female
5	11th	Handlers-cleaners	Private	12	United-States	<=50K	Male
6	11th	Farming-fishing	Self-em	10	United-States	<=50K	Male
7	11th	Other-service	Private	12	United-States	<=50K	Female
8	11th	Sales	Private	8	United-States	<=50K	Female

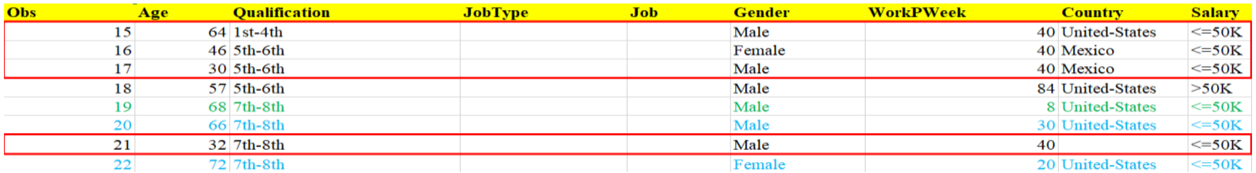


Obs	AvegWorksOfPrivHouseServ
1	31.7273

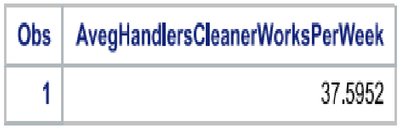
Figure – 28(missing records in jobtype and job of 11th grade)

If we notice we can see people who works under **PrivHouseService** their working average is nearly 30 so we can say that observation 10th and 11th both are under this category and again we can see female whose qualification is 11th grade and working rate is under 15 hours per week are basically in **sales** category and male are in **other-service** job.

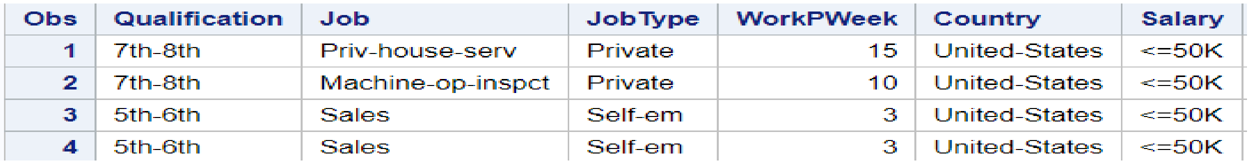
4.1.4 grade from 1st-8th



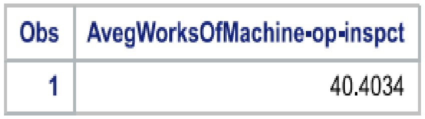
Obs	Age	Qualification	JobType	Job	Gender	WorkPWeek	Country	Salary
15	64	1st-4th			Male	40	United-States	<=50K
16	46	5th-6th			Female	40	Mexico	<=50K
17	30	5th-6th			Male	40	Mexico	<=50K
18	57	5th-6th			Male	84	United-States	>50K
19	68	7th-8th			Male	8	United-States	<=50K
20	66	7th-8th			Male	30	United-States	<=50K
21	32	7th-8th			Male	40		<=50K
22	72	7th-8th			Female	20	United-States	<=50K



Obs	AvegHandlersCleanerWorksPerWeek
1	37.5952



Obs	Qualification	Job	JobType	WorkPWeek	Country	Salary
1	7th-8th	Priv-house-serv	Private	15	United-States	<=50K
2	7th-8th	Machine-op-inspct	Private	10	United-States	<=50K
3	5th-6th	Sales	Self-em	3	United-States	<=50K
4	5th-6th	Sales	Self-em	3	United-States	<=50K

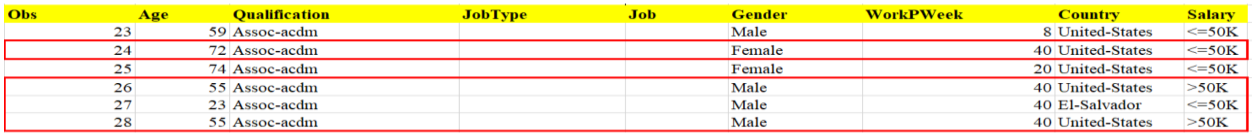


Obs	AvegWorksOfMachine-op-inspct
1	40.4034

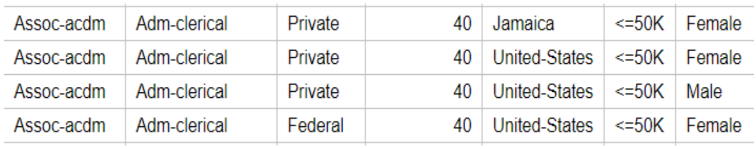
Figure – 29(missing records in jobtype and job of 1st - 8th grade)

As most of the people whose grade is from 1st – 8th and who works average 40 hours pe week and get less or equal salary of 50k are in **Machine-op-inspct** (40.4034Avg) so that we can assume that observaton 15,16,17 and 21 are under that category of job. And people who works nealy 30 hours per week we can assume that they are in **Priv-House-Service**(31.7273Avg) which are row number 20th and 22nd . Again observation number 18th can be **farm-fisherman(**47.35Avg) as its average is nearly closer to it compare to other coloumn value and observation number 19 are under **Machine-op-inspct** as it table we can see people whose grade are 7th – 8th and earn less than 50k are in Machine-op-inspct actegory.

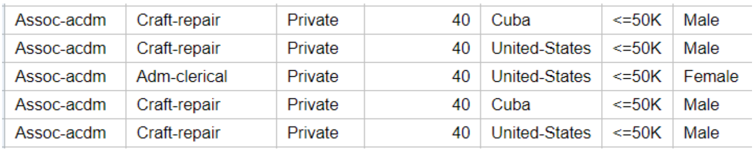
4.1.5 Assoc-acdm



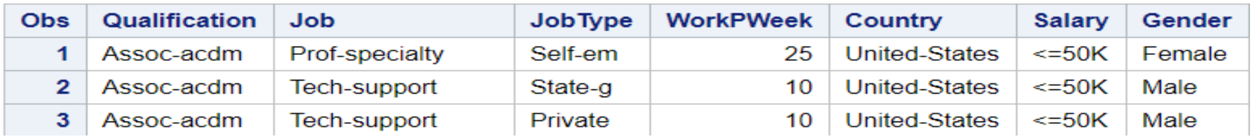
Obs	Age	Qualification	JobType	Job	Gender	WorkPWeek	Country	Salary
23	59	Assoc-acdm			Male	8	United-States	<=50K
24	72	Assoc-acdm			Female	40	United-States	<=50K
25	74	Assoc-acdm			Female	20	United-States	<=50K
26	55	Assoc-acdm			Male	40	United-States	>50K
27	23	Assoc-acdm			Male	40	El-Salvador	<=50K
28	55	Assoc-acdm			Male	40	United-States	>50K



Assoc-acdm	Adm-clerical	Private	40	Jamaica	<=50K	Female
Assoc-acdm	Adm-clerical	Private	40	United-States	<=50K	Female
Assoc-acdm	Adm-clerical	Private	40	United-States	<=50K	Male
Assoc-acdm	Adm-clerical	Federal	40	United-States	<=50K	Female



Assoc-acdm	Craft-repair	Private	40	Cuba	<=50K	Male
Assoc-acdm	Craft-repair	Private	40	United-States	<=50K	Male
Assoc-acdm	Adm-clerical	Private	40	United-States	<=50K	Female
Assoc-acdm	Craft-repair	Private	40	Cuba	<=50K	Male
Assoc-acdm	Craft-repair	Private	40	United-States	<=50K	Male

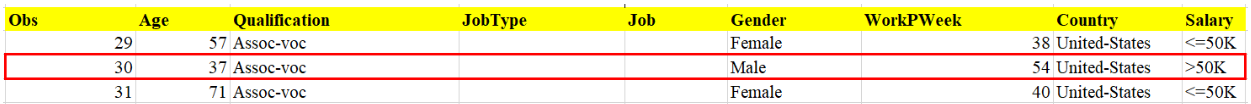


Obs	Qualification	Job	JobType	WorkPWeek	Country	Salary	Gender
1	Assoc-acdm	Prof-specialty	Self-em	25	United-States	<=50K	Female
2	Assoc-acdm	Tech-support	State-g	10	United-States	<=50K	Male
3	Assoc-acdm	Tech-support	Private	10	United-States	<=50K	Male

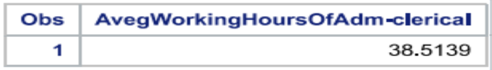
Figure – 30(missing records in jobtype and job of Assoc-acdm)

Here from above table we can see male who hava qualification of Assoc-acdm and who works 40 hours are in carft-repair and female with same qualification and working hours fall under adm-clerical so that we can assume row number 26th-27th are under **carft-repair** and row 24th is in **adm-clerical** category. Again in the same way row 24 is in **pro-specialty as** female wo works less or equal to 25hours and gets less or equal to 50k are in this category and in the same way row 23th is in **tech-support** job type.

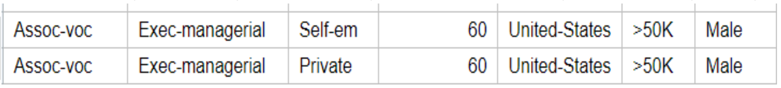
4.1.6 Assoc-voc



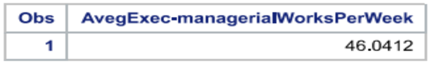
Obs	Age	Qualification	JobType	Job	Gender	WorkPWeek	Country	Salary
29	57	Assoc-voc			Female	38	United-States	<=50K
30	37	Assoc-voc			Male	54	United-States	>50K
31	71	Assoc-voc			Female	40	United-States	<=50K



Obs	AvegWorkingHoursOfAdm-clerical
1	38.5139



Assoc-voc	Exec-managerial	Self-em	60	United-States	>50K	Male
Assoc-voc	Exec-managerial	Private	60	United-States	>50K	Male

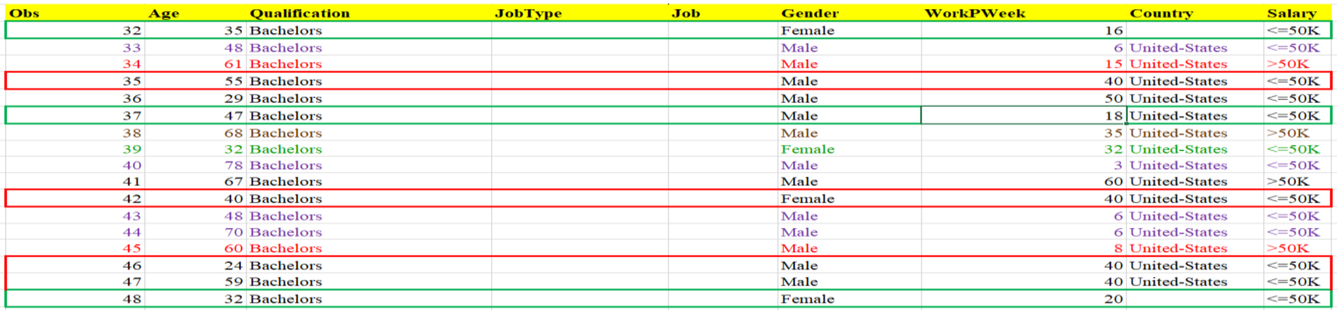


Obs	AvegExec-managerialWorksPerWeek
1	46.0412

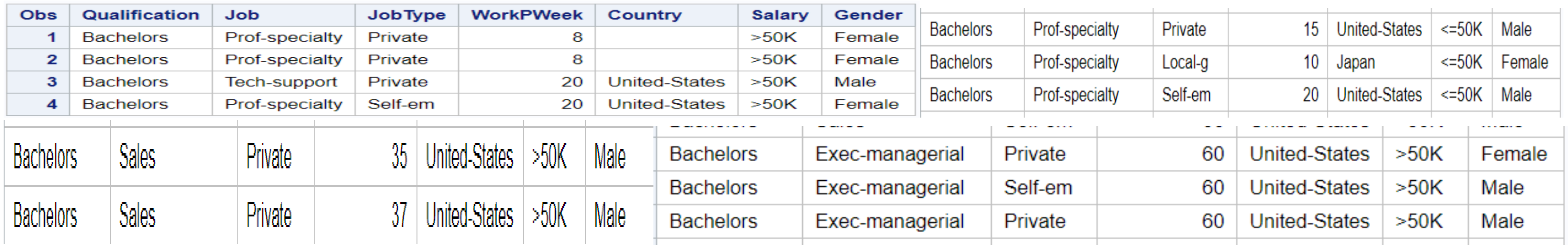
Figure – 31(missing records in jobtype and job of Assoc-voc)

Most of the people who studied assoc-voc, works more than 45hours and gets a salary of more than 50k are in **Exec-managerial** job category so we can assume row 30th is in Exec-managerial category. On the other hand people who has same degree but works less or equal to 40hours and gets a salary of equal or less than 50k are in **Adm-clercial** (38.5139 Avg working hours) so we can say that row 29th and 30th are in that job.

4.1.7 Bachelors degree



Obs	Age	Qualification	JobType	Job	Gender	WorkPWeek	Country	Salary
32	35	Bachelors			Female	16		<=50K
33	48	Bachelors			Male	6	United-States	<=50K
34	61	Bachelors			Male	15	United-States	>50K
35	55	Bachelors			Male	40	United-States	<=50K
36	29	Bachelors			Male	50	United-States	<=50K
37	47	Bachelors			Male	18	United-States	<=50K
38	68	Bachelors			Male	35	United-States	>50K
39	32	Bachelors			Female	32	United-States	<=50K
40	78	Bachelors			Male	3	United-States	<=50K
41	67	Bachelors			Male	60	United-States	>50K
42	40	Bachelors			Female	40	United-States	<=50K
43	48	Bachelors			Male	6	United-States	<=50K
44	70	Bachelors			Male	6	United-States	<=50K
45	60	Bachelors			Male	8	United-States	>50K
46	24	Bachelors			Male	40	United-States	<=50K
47	59	Bachelors			Male	40	United-States	<=50K
48	32	Bachelors			Female	20		<=50K



Obs	Qualification	Job	JobType	WorkPWeek	Country	Salary	Gender
1	Bachelors	Prof-specialty	Private	8		>50K	Female
2	Bachelors	Prof-specialty	Private	8		>50K	Female
3	Bachelors	Tech-support	Private	20	United-States	>50K	Male
4	Bachelors	Prof-specialty	Self-em	20	United-States	>50K	Female

Bachelors	Prof-specialty	Private	15	United-States	<=50K	Male
Bachelors	Prof-specialty	Local-g	10	Japan	<=50K	Female
Bachelors	Prof-specialty	Self-em	20	United-States	<=50K	Male

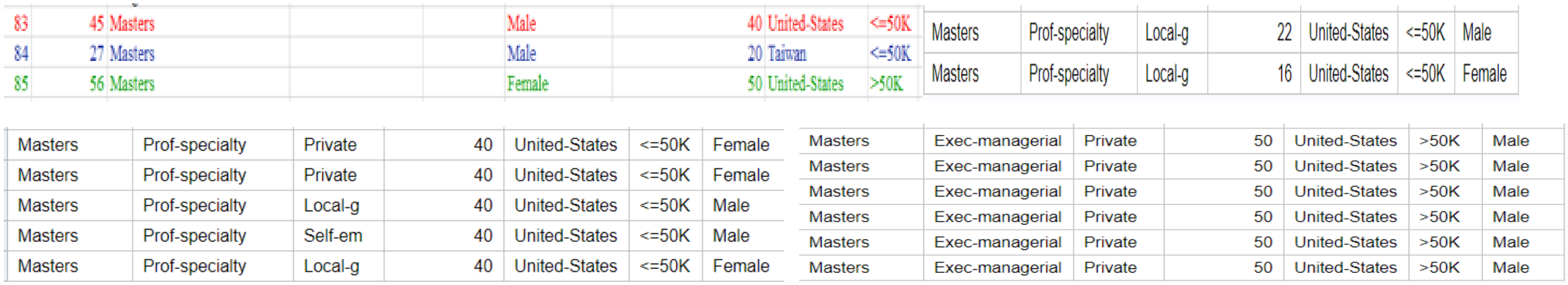
Bachelors	Sales	Private	35	United-States	>50K	Male
Bachelors	Sales	Private	37	United-States	>50K	Male

Bachelors	Exec-managerial	Private	60	United-States	>50K	Female
Bachelors	Exec-managerial	Self-em	60	United-States	>50K	Male
Bachelors	Exec-managerial	Private	60	United-States	>50K	Male

Figure – 32(missing records in jobtype and job of Bachelors)

From above table and information given here we can see people who has Bachelor degree and works less or equal to 20 hours per week and gets less ,equal or greater than 50k their job is **prof-speialty** so row 32,33,34, 37,40,43,44,45,48 are in that job category. Again with same degree and salary but working hours 40 per week also falls under pro-specialty job category as its average is 40.6245 so we can assume row 35,42,46,47 are in **pro-specialty** job. Again row 41 falls in **Exec-managerial** category as we can see most of the people who get more than 50k and works 60 hours are in that category. And in the same way row 38 falls under sales type category as we can see people who earn more than 50k and work nearly 35hours per week are in that category.

4.1.8 Masters



83	45	Masters			Male	40	United-States	<=50K
84	27	Masters			Male	20	Taiwan	<=50K
85	56	Masters			Female	50	United-States	>50K

Masters	Prof-specialty	Local-g	22	United-States	<=50K	Male
Masters	Prof-specialty	Local-g	16	United-States	<=50K	Female

Masters	Prof-specialty	Private	40	United-States	<=50K	Female
Masters	Prof-specialty	Private	40	United-States	<=50K	Female
Masters	Prof-specialty	Local-g	40	United-States	<=50K	Male
Masters	Prof-specialty	Self-em	40	United-States	<=50K	Male
Masters	Prof-specialty	Local-g	40	United-States	<=50K	Female

Masters	Exec-managerial	Private	50	United-States	>50K	Male
Masters	Exec-managerial	Private	50	United-States	>50K	Male
Masters	Exec-managerial	Private	50	United-States	>50K	Male
Masters	Exec-managerial	Private	50	United-States	>50K	Male
Masters	Exec-managerial	Private	50	United-States	>50K	Male
Masters	Exec-managerial	Private	50	United-States	>50K	Male

Figure – 33(missing records in jobtype and job of Masters)

In the above table we can see people having a master degree and who works 50 hours per week and makes more than 50k are in Exec-managerial(46.0412) works and jobtype is private so job and jobtype of row 85 is **exec-managerial** and **private**. Again we see people working average 40hours and makes less or equal to 50k are in prof-specialty(40.6245) so row 83,84 job and jobtype are **pro-specialty** and **local-g** as most of them jobtype is **local-g**. Here we can also fill up hs-grade by the following ways discussed above.

4.1.9 Country

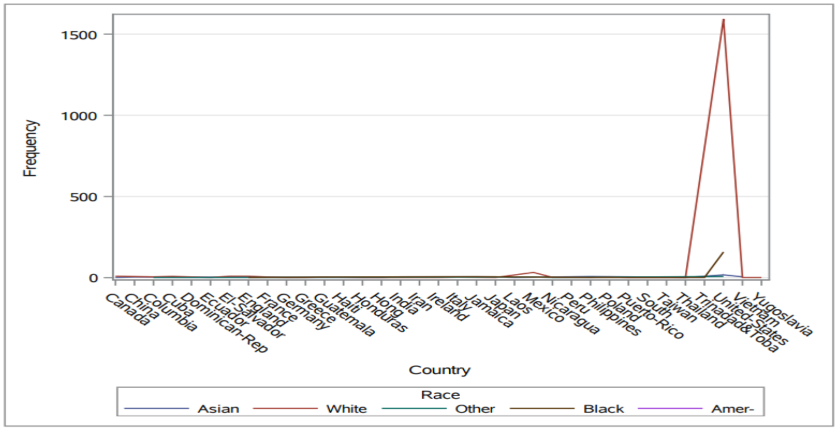
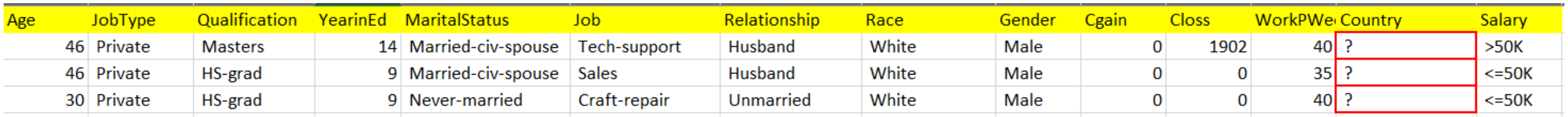


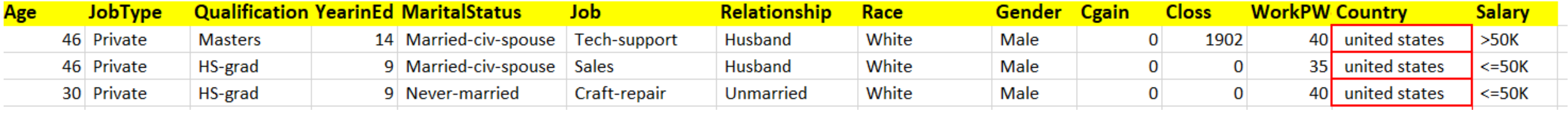
Figure – 34(missing records in country)

According to the sample data set most of the people whose country are missing are from white race and if we notice that majority of the people who lives in United states are white. So I am assuming that people who has missing value in country are from united states. Here below figure 35 and 36 shows before and after modification of the missing value of country.



Age	JobType	Qualification	YearinEd	MaritalStatus	Job	Relationship	Race	Gender	Cgain	Closs	WorkPWe	Country	Salary
46	Private	Masters	14	Married-civ-spouse	Tech-support	Husband	White	Male	0	1902	40	?	>50K
46	Private	HS-grad	9	Married-civ-spouse	Sales	Husband	White	Male	0	0	35	?	<=50K
30	Private	HS-grad	9	Never-married	Craft-repair	Unmarried	White	Male	0	0	40	?	<=50K

Figure – 35(Before modification)



Age	JobType	Qualification	YearinEd	MaritalStatus	Job	Relationship	Race	Gender	Cgain	Closs	WorkPW	Country	Salary
46	Private	Masters	14	Married-civ-spouse	Tech-support	Husband	White	Male	0	1902	40	united states	>50K
46	Private	HS-grad	9	Married-civ-spouse	Sales	Husband	White	Male	0	0	35	united states	<=50K
30	Private	HS-grad	9	Never-married	Craft-repair	Unmarried	White	Male	0	0	40	united states	<=50K

Figure – 36(After modification)

4.2 Noisy data

4.2.1 Age

In the given data set there are five outliers in age these are 78,79,80,81,90.

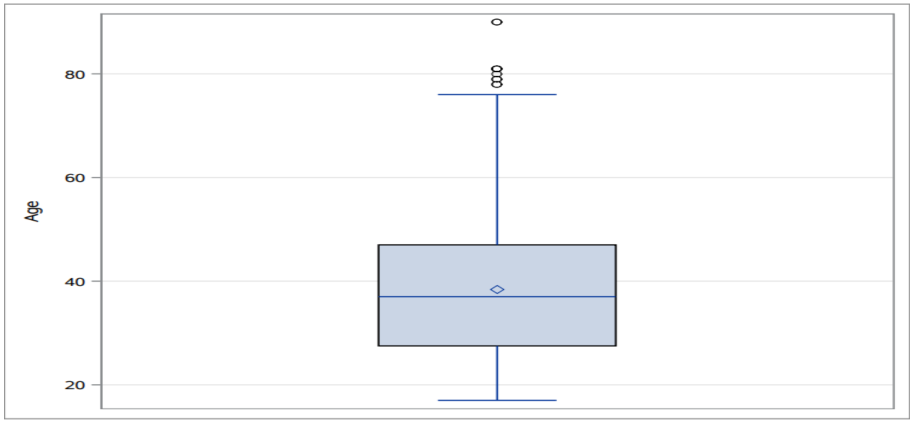
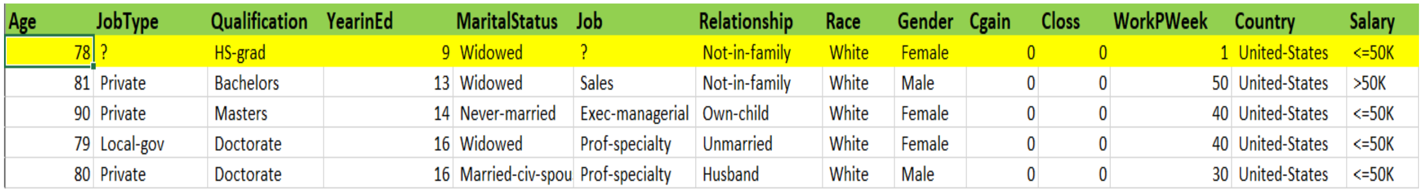


Figure – 37(Noisy data in Age)

In USA people at the age of 70 get retire from the job so I assume that people whose age is 78,79,80,81,90 and still doing job other than self-employment are outlier and it will be replaced by mean(38). Again, here in first row people who age is 78 does not match requirement in working hours so it should be removed (Caplinger, 2020).

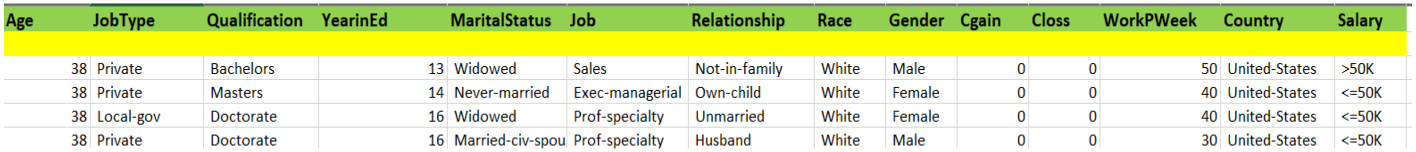
Before pre-processing



Age	JobType	Qualification	YearinEd	MaritalStatus	Job	Relationship	Race	Gender	Cgain	Closs	WorkPWeek	Country	Salary
78	?	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	0	1	United-States	<=50K
81	Private	Bachelors	13	Widowed	Sales	Not-in-family	White	Male	0	0	50	United-States	>50K
90	Private	Masters	14	Never-married	Exec-managerial	Own-child	White	Female	0	0	40	United-States	<=50K
79	Local-gov	Doctorate	16	Widowed	Prof-specialty	Unmarried	White	Female	0	0	40	United-States	<=50K
80	Private	Doctorate	16	Married-civ-spou	Prof-specialty	Husband	White	Male	0	0	30	United-States	<=50K

Figure – 38

After pre-processing



Age	JobType	Qualification	YearinEd	MaritalStatus	Job	Relationship	Race	Gender	Cgain	Closs	WorkPWeek	Country	Salary
38	Private	Bachelors	13	Widowed	Sales	Not-in-family	White	Male	0	0	50	United-States	>50K
38	Private	Masters	14	Never-married	Exec-managerial	Own-child	White	Female	0	0	40	United-States	<=50K
38	Local-gov	Doctorate	16	Widowed	Prof-specialty	Unmarried	White	Female	0	0	40	United-States	<=50K
38	Private	Doctorate	16	Married-civ-spou	Prof-specialty	Husband	White	Male	0	0	30	United-States	<=50K

Figure – 39

4.2.2 YearinEd

YearinEd means duration taken by a person in education. Here there are two outliers and these are 1 and 2.

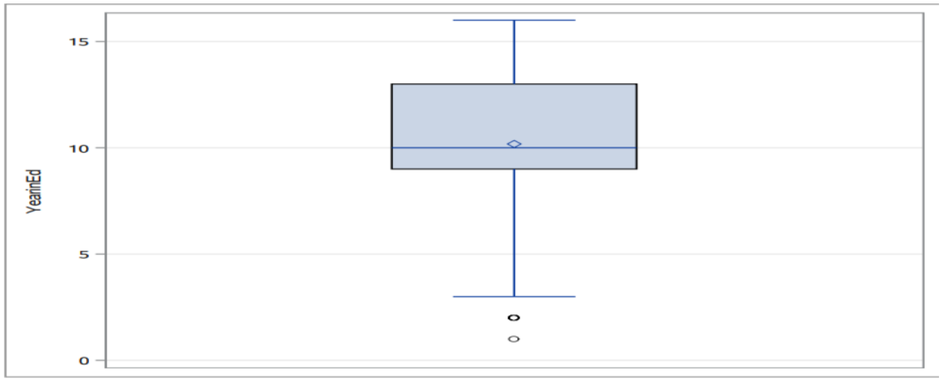
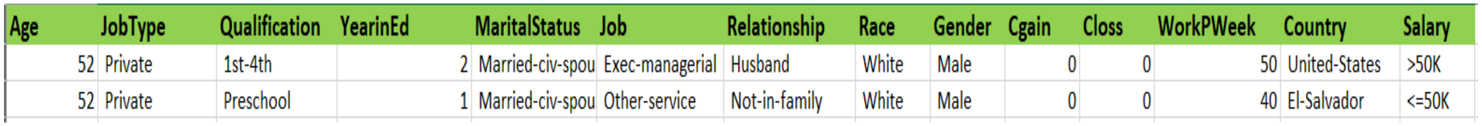


Figure – 40(Noisy data in YearinEd)

According to the sample dataset, qualification of people whose jo is exec-managerial are mostly from Bachelors degree and people whose year in education is 2 years and qualification is 1st-4th they mostly do fishing-farming and other types of job so this information doesn't match the requirement.On the other hand people cannot have pre-school within one years so that it should also be removed.

Before pre-processing



Age	JobType	Qualification	YearinEd	MaritalStatus	Job	Relationship	Race	Gender	Cgain	Closs	WorkPWeek	Country	Salary
52	Private	1st-4th	2	Married-civ-spou	Exec-managerial	Husband	White	Male	0	0	50	United-States	>50K
52	Private	Preschool	1	Married-civ-spou	Other-service	Not-in-family	White	Male	0	0	40	El-Salvador	<=50K

Figure – 41

After pre-processing

Two rows of the outlier from year in education are removed as it doesn’t match with most of the information given in data set.

4.2.3 Cgain

Here cgain means capital loss.

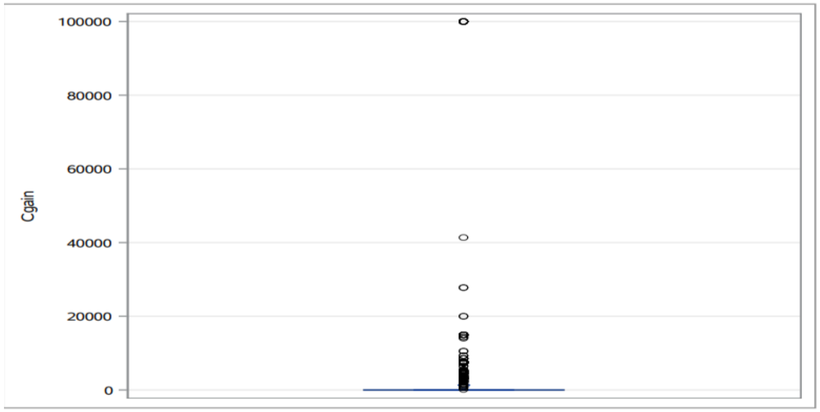
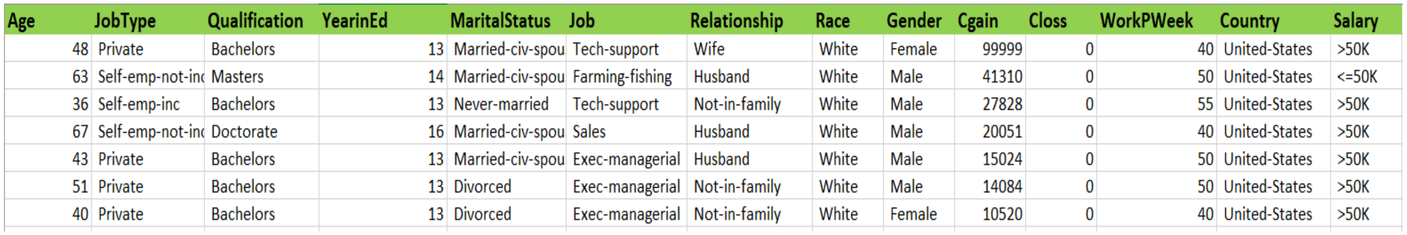


Figure – 42(Noisy data in Cgain)

In USA people whose salary is more than 40k should pay 15% of his salary as capital gain. Here capital gain of 50k salary is 7500(Frankel, 2020).

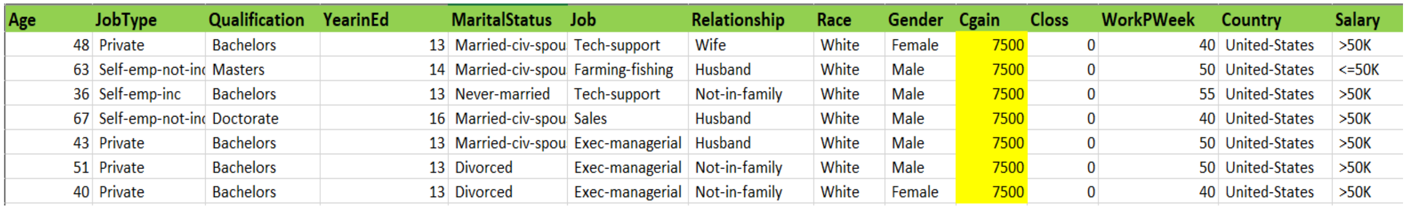
Before pre-processing



Age	JobType	Qualification	YearinEd	MaritalStatus	Job	Relationship	Race	Gender	Cgain	Closs	WorkPWeek	Country	Salary
48	Private	Bachelors	13	Married-civ-spou	Tech-support	Wife	White	Female	99999	0	40	United-States	>50K
63	Self-emp-not-inc	Masters	14	Married-civ-spou	Farming-fishing	Husband	White	Male	41310	0	50	United-States	<=50K
36	Self-emp-inc	Bachelors	13	Never-married	Tech-support	Not-in-family	White	Male	27828	0	55	United-States	>50K
67	Self-emp-not-inc	Doctorate	16	Married-civ-spou	Sales	Husband	White	Male	20051	0	40	United-States	>50K
43	Private	Bachelors	13	Married-civ-spou	Exec-managerial	Husband	White	Male	15024	0	50	United-States	>50K
51	Private	Bachelors	13	Divorced	Exec-managerial	Not-in-family	White	Male	14084	0	50	United-States	>50K
40	Private	Bachelors	13	Divorced	Exec-managerial	Not-in-family	White	Female	10520	0	40	United-States	>50K

Figure – 43

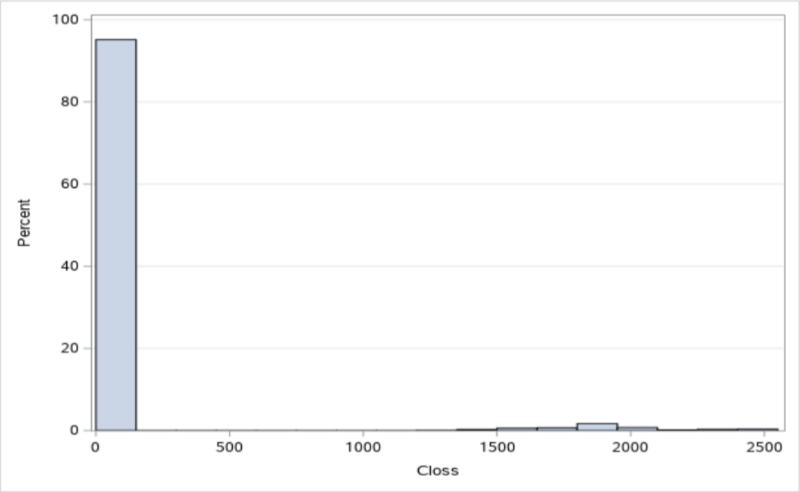
After pre-processing



Age	JobType	Qualification	YearinEd	MaritalStatus	Job	Relationship	Race	Gender	Cgain	Closs	WorkPWeek	Country	Salary
48	Private	Bachelors	13	Married-civ-spou	Tech-support	Wife	White	Female	7500	0	40	United-States	>50K
63	Self-emp-not-inc	Masters	14	Married-civ-spou	Farming-fishing	Husband	White	Male	7500	0	50	United-States	<=50K
36	Self-emp-inc	Bachelors	13	Never-married	Tech-support	Not-in-family	White	Male	7500	0	55	United-States	>50K
67	Self-emp-not-inc	Doctorate	16	Married-civ-spou	Sales	Husband	White	Male	7500	0	40	United-States	>50K
43	Private	Bachelors	13	Married-civ-spou	Exec-managerial	Husband	White	Male	7500	0	50	United-States	>50K
51	Private	Bachelors	13	Divorced	Exec-managerial	Not-in-family	White	Male	7500	0	50	United-States	>50K
40	Private	Bachelors	13	Divorced	Exec-managerial	Not-in-family	White	Female	7500	0	40	United-States	>50K

Figure – 44

4.2.4 CLOSS



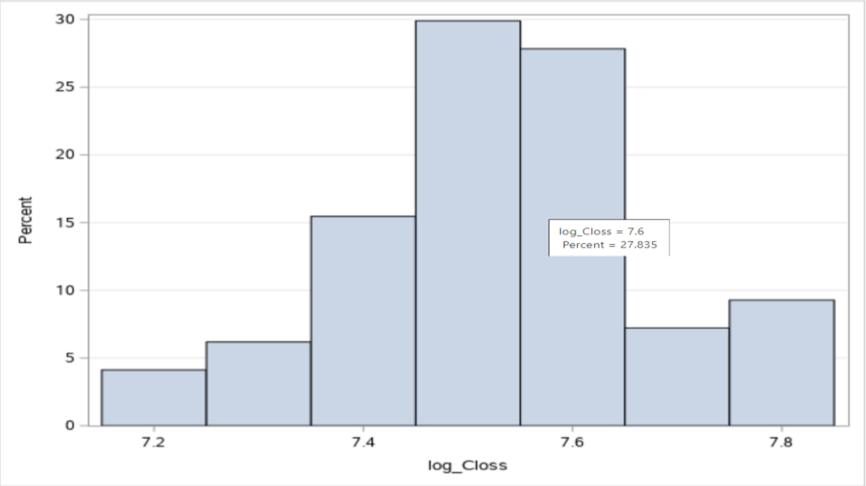


Figure – 45(Noisy data in Closs before pre-processing) **Figure – 46(data in Closs after pre-processing)**

Here I used transformation method(natural log) to remove noisy data which belongs to closs attributes which means I reduced the size of the value of closs attributes and here after doing log method in transformation there might be a noisy data which we should remove.

4.2.5 WorkPerweek

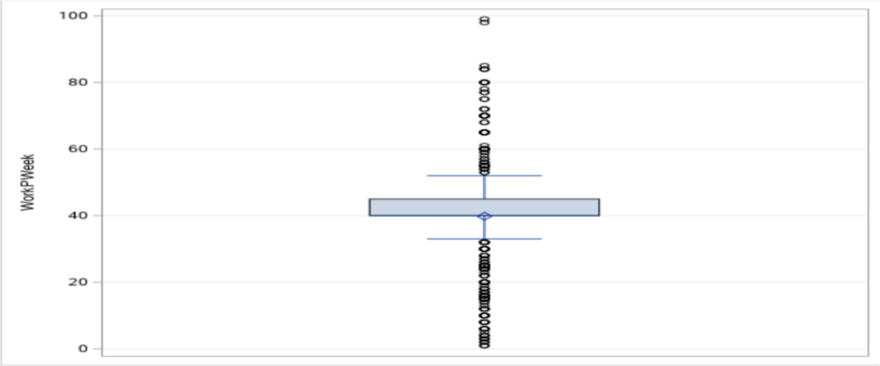


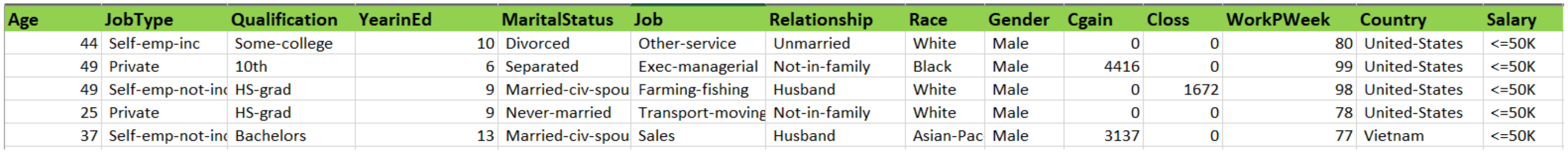
Figure – 47(Noisy data in Workperweek)

In usa average working hours of a person based on age are (https://www.facebook.com/thebalancecom, 2021).

AGE	WORKS PER WEEK(HOURS)
16-19 years	24.1 hours
20-24 years	34.8 hours
25-54	40.5 hours
50 or above	38.0 hours

The above table shows the working rate of the people in America based on their age. And according this table we remove noisy data and filled it.

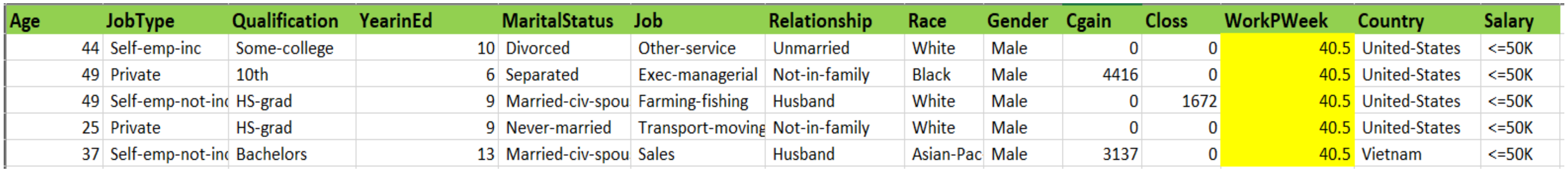
<u>BEFORE PRE-PROCESSING</u>



Age	JobType	Qualification	YearinEd	MaritalStatus	Job	Relationship	Race	Gender	Cgain	Closs	WorkPWeek	Country	Salary
44	Self-emp-inc	Some-college	10	Divorced	Other-service	Unmarried	White	Male	0	0	80	United-States	<=50K
49	Private	10th	6	Separated	Exec-managerial	Not-in-family	Black	Male	4416	0	99	United-States	<=50K
49	Self-emp-not-inc	HS-grad	9	Married-civ-spou	Farming-fishing	Husband	White	Male	0	1672	98	United-States	<=50K
25	Private	HS-grad	9	Never-married	Transport-moving	Not-in-family	White	Male	0	0	78	United-States	<=50K
37	Self-emp-not-inc	Bachelors	13	Married-civ-spou	Sales	Husband	Asian-Pac	Male	3137	0	77	Vietnam	<=50K

Figure – 48

AFTER PRE-PROCESSING



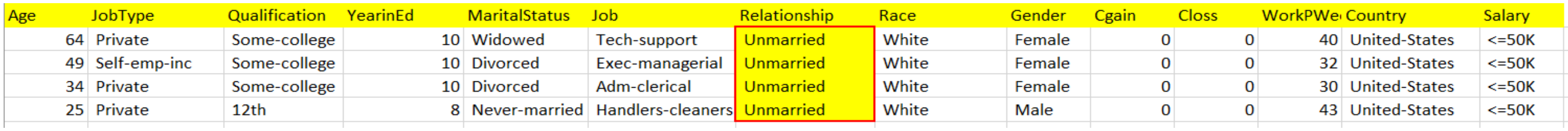
Age	JobType	Qualification	YearinEd	MaritalStatus	Job	Relationship	Race	Gender	Cgain	Closs	WorkPWeek	Country	Salary
44	Self-emp-inc	Some-college	10	Divorced	Other-service	Unmarried	White	Male	0	0	40.5	United-States	<=50K
49	Private	10th	6	Separated	Exec-managerial	Not-in-family	Black	Male	4416	0	40.5	United-States	<=50K
49	Self-emp-not-inc	HS-grad	9	Married-civ-spou	Farming-fishing	Husband	White	Male	0	1672	40.5	United-States	<=50K
25	Private	HS-grad	9	Never-married	Transport-moving	Not-in-family	White	Male	0	0	40.5	United-States	<=50K
37	Self-emp-not-inc	Bachelors	13	Married-civ-spou	Sales	Husband	Asian-Pac	Male	3137	0	40.5	Vietnam	<=50K

Figure – 49

4.3 INCONSISTENT DATA

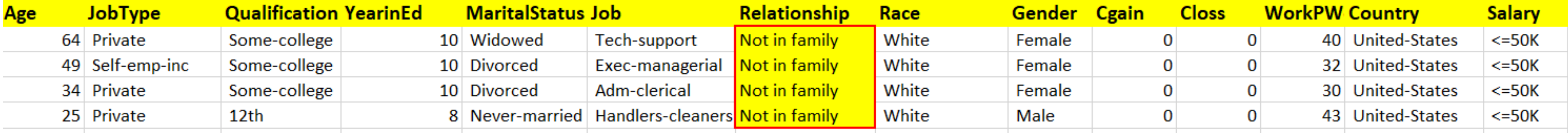
After observing all the above data set given, if we notice we can only find one inconsistant data which in Relationship attributes and that is **Unmarried**. Here **Unmarried** data should have placed in **Marital status** so we consider it as inconsistant data and intead of that value we have given **not-in-family** because if we observed in the given data set most of the people who belongs to divorced,never-married and widowed their relationship status is **not-in-family**.

4.3.1 Relationship



Age	JobType	Qualification	YearinEd	MaritalStatus	Job	Relationship	Race	Gender	Cgain	Closs	WorkPWe	Country	Salary
64	Private	Some-college	10	Widowed	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K
49	Self-emp-inc	Some-college	10	Divorced	Exec-managerial	Unmarried	White	Female	0	0	32	United-States	<=50K
34	Private	Some-college	10	Divorced	Adm-clerical	Unmarried	White	Female	0	0	30	United-States	<=50K
25	Private	12th	8	Never-married	Handlers-cleaners	Unmarried	White	Male	0	0	43	United-States	<=50K

Figure – 50(before pre-proscessing)



Age	JobType	Qualification	YearinEd	MaritalStatus	Job	Relationship	Race	Gender	Cgain	Closs	WorkPW	Country	Salary
64	Private	Some-college	10	Widowed	Tech-support	Not in family	White	Female	0	0	40	United-States	<=50K
49	Self-emp-inc	Some-college	10	Divorced	Exec-managerial	Not in family	White	Female	0	0	32	United-States	<=50K
34	Private	Some-college	10	Divorced	Adm-clerical	Not in family	White	Female	0	0	30	United-States	<=50K
25	Private	12th	8	Never-married	Handlers-cleaners	Not in family	White	Male	0	0	43	United-States	<=50K

Figure – 51(After pre-proscessing)

5.0 Conclusion

There are four types of attributes (measurement) which are -nominal, ordinal, interval and ratio again which we can further divide by – qualitative, quantitative and Continuous. And in this documentation, in the beginning these have been discussed and afterward we have data exploration which we have done by using SAS studio and where some figure like – graph, histogram and frequency table have shown. Finally, in the last part we have incomplete, noisy and inconsistent data which is solved by data pre-processing techniques.

The main goal of these given steps is to make data set more accurate and acceptable.

6.0 References

Caplinger, D., 2020. *The Motly Fool.* [Online]
Available at: https://www.fool.com/retirement/2020/01/20/2020-is-the-last-year-before-this-social-security.aspx#:~:text=2020%20is%20the%20last%20year%20in%20which%20someone%20turning%2066,full%20years'%20worth%20of%20credits.
[Accessed friday February 2021].

Elgabry, O., 2019. *towards data science.* [Online]
Available at: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4
[Accessed sunday february 2021].

Library, B. E. H. B., 2020. *Understand the levels of measurement.* [Online]
Available at: https://library.buffalostate.edu/measurements/overview
[Accessed sunday february 2021].

SASCRUNCH TRAINING, 2020. *How to deal with missing data.* [Online]
Available at: https://www.sascrunch.com/dealing-with-missing-values.html
[Accessed sunday february 2021].

Uusitalo, J., 2020. *data Direct.* [Online]
Available at: https://www.sciencedirect.com/topics/computer-science/preprocessing
[Accessed tuesday february 2021].

Frankel, M. (2020). *How Much Is Capital Gains Tax in 2020?* [online] Millionacres. Available at: https://www.fool.com/millionacres/taxes/capital-gains/how-much-capital-gains-tax-2020/ [Accessed 17 Feb. 2021].

https://www.facebook.com/thebalancecom (2021). *What is the Average Hours Per Week Worked in the US?* [online] The Balance Careers. Available at: https://www.thebalancecareers.com/what-is-the-average-hours-per-week-worked-in-the-us-2060631 [Accessed 17 Feb. 2021].