Data Exploration

Task 1.1(a): The Data_Set_B overall contains the 2001 observations and 14 variable. Below all variable will categorized by attributes Types by definition of each scales.

Data Types	Nominal	Ordinal	Interval	Ratio
	Job type	Qualification	Years in Education	Age
	Marital status		Works per week	Capital Gain
	Job			Capital Loss
	Relationship status			
	Race			
	Gender			
	Country			
	Salary			

By definitions of each scales. Such as **Nominal**: Nominal scales are used for labeling variables, without any quantitative value. "Nominal" scales could simply be called "labels.
Ordinal: it is the order of the values is what's important and significant, but the differences between each one is not really known.
Interval: Interval scales are numeric scales in which we know not only the order, but also the exact differences between the values
Ratio: Ratio scales are the ultimate nirvana when it comes to measurement scales because they tell us about the order, they tell us the exact value between units,

```
NOTE: 2001 records were read from the infile "/home/ameerkhoso470/workshop/DataSet_B.csv".
      The minimum record length was 75.
      The maximum record length was 149.
NOTE: The data set DM.PROFILE_DATA has 2001 observations and 14 variables.
NOTE: DATA statement used (Total process time):
      real time           0.03 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              1329.46k
      OS Memory           41128.00k
```

Fig:1 SAS Screenshot of observation and variables

Task1.1 (b):

Table 1: Analysis of Age Attribute: The Maximum age of 2001 respondent was 90 and the minimum age was 17 years. However the average mean of Age was 38.3 following by median 37. Below table 1 shows the summary of Age variable.

Summarizing Properties of of Age

The MEANS Procedure

Analysis Variable : age							
N	Minimum	Maximum	Mean	Median	Variance	Lower Quartile	Upper Quartile
2000	17.0000000	90.0000000	38.3895000	37.0000000	184.8712254	27.5000000	47.0000000

Fig: 2 Distribution of Age By Age Category: Figure 2 illustrates the overall spread of Age according to their age category where majority of respondent was 38.8 (776) are from Middle-age. Following by young age category was 26.32% (527) respondent. However mature respondent ratio was 22.35% (447). And the lowest was Teenager and old which is illustrated by below fig 2

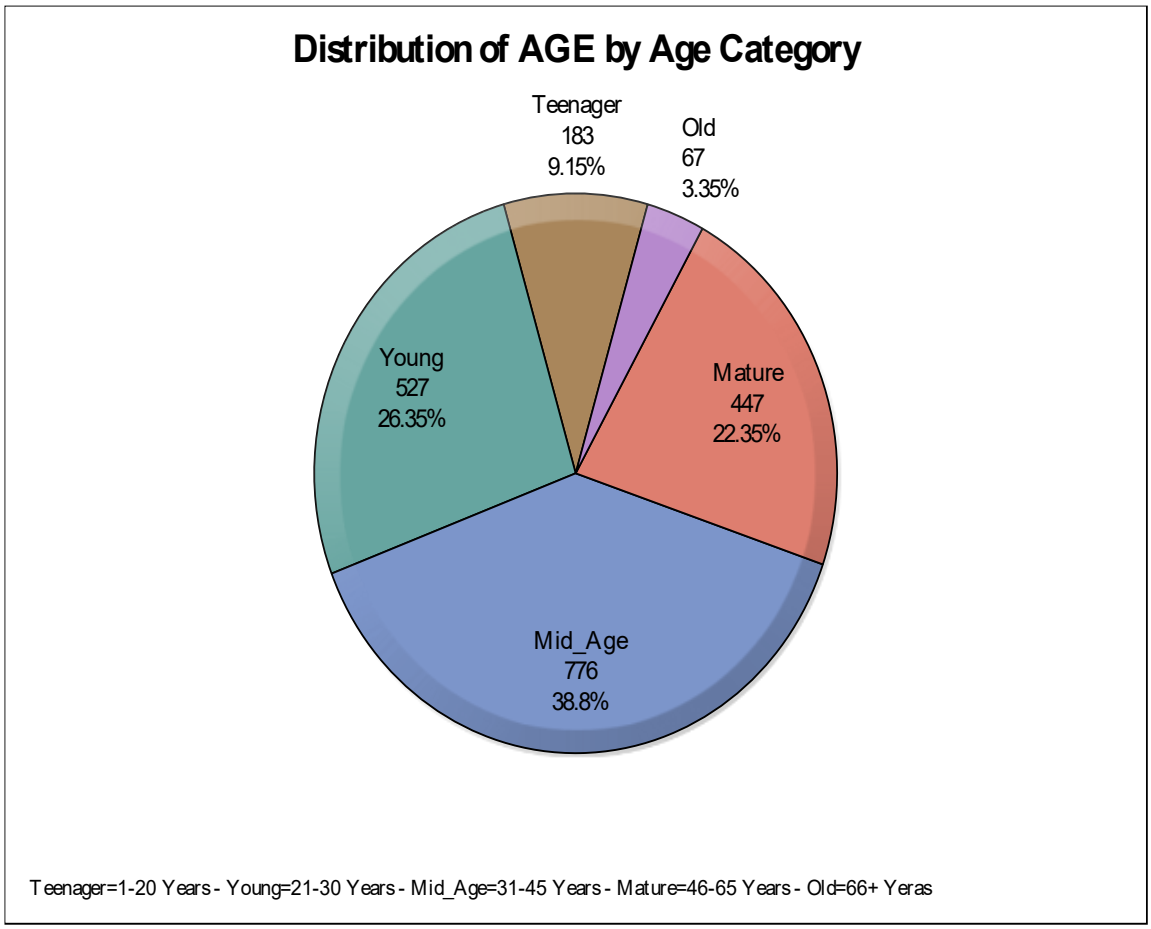


Table 2: Analysis of Years in education: From the above respondent. The maximum years in education was recorded 16 and minimum was 1. However the Average mean was 10.17 followed by median 10. Below table 2 shows the summary of Years in Education.

Report as of &currentdate &currenttime

Summarizing Properties of Years in Education

The MEANS Procedure

Analysis Variable : years_in_edu							
N	**Minimum**	**Maximum**	**Mean**	**Median**	**Variance**	**Lower Quartile**	**Upper Quartile**
2000	1.0000000	16.0000000	10.1785000	10.0000000	6.7850303	9.0000000	13.0000000

Table 3: Analysis of Capital Gain: From 2001 respondent the capital gain was recorded. The maximum capital gain was 99999 and minimum was 0. On the other side the Average mean was 1304.28 recorded and median was 0. Below table 3 illustrates the summary of capital gain.

Report as of &currentdate &currenttime

Summarizing Properties of Capital Gain

The MEANS Procedure

Analysis Variable : Capital_gain							
N	**Minimum**	**Maximum**	**Mean**	**Median**	**Variance**	**Lower Quartile**	**Upper Quartile**
2000	0	99999.00	1304.28	0	70811469.38	0	0

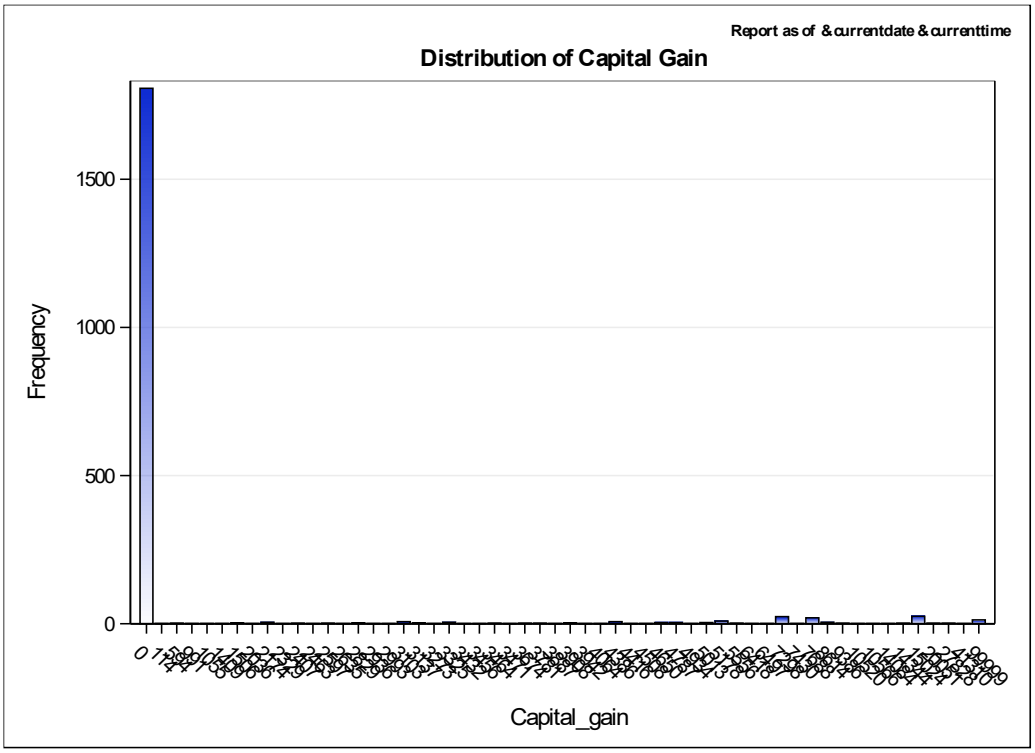


Fig 3: Distribution of Capital Gain

Table 4: Analysis of Capital Loss: From 2001 respondent the capital loss was recorded. The maximum capital loss was 2547 and minimum was 0. On the other side the Average mean was 91.35 recorded and median was 0. Below table 3 illustrates the summary of capital gain.

Report as of &currentdate &currenttime

Summarizing Properties of Capital Loss

The MEANS Procedure

Analysis Variable : Capital_loss							
N	Minimum	Maximum	Mean	Median	Variance	Lower Quartile	Upper Quartile
2000	0	2547.00	91.3535000	0	167193.69	0	0

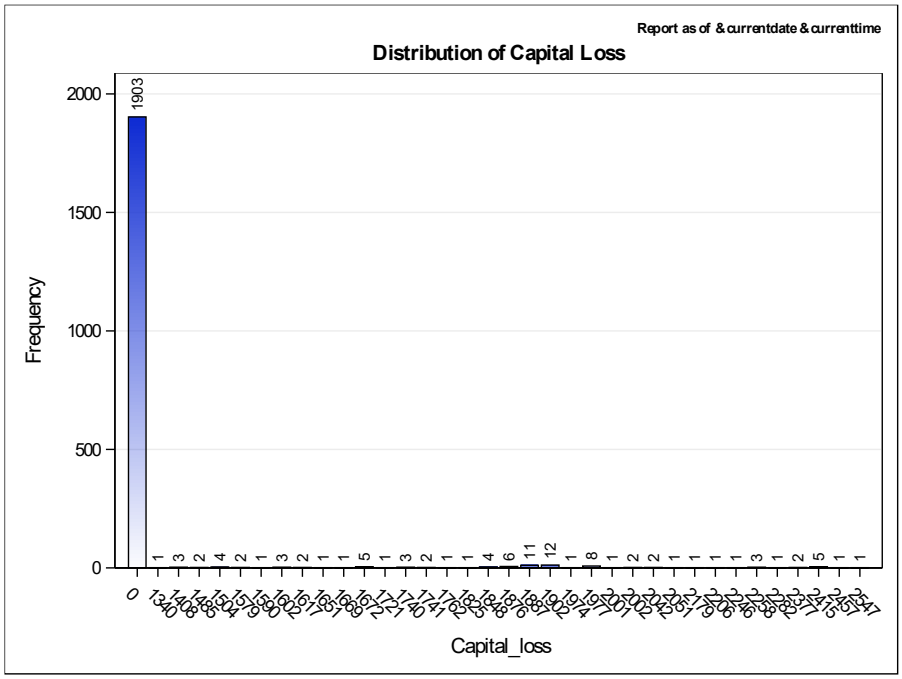


Fig: 4 Distribution Of capital Loss

Table 5: Work per Week: From the respondent the work per week. The maximum respondent was 99 and minimum was 1. On the other side the Average mean was 39.81 recorded and median was 40. Below table 3 illustrates the summary of capital gain.

Report as of &currentdate &currenttime

Summarizing Properties of Works Per Week

The MEANS Procedure

Analysis Variable : Work_per_week							
N	**Minimum**	**Maximum**	**Mean**	**Median**	**Variance**	**Lower Quartile**	**Upper Quartile**
2000	1.0000000	99.0000000	39.8175000	40.0000000	149.0457166	40.0000000	45.0000000

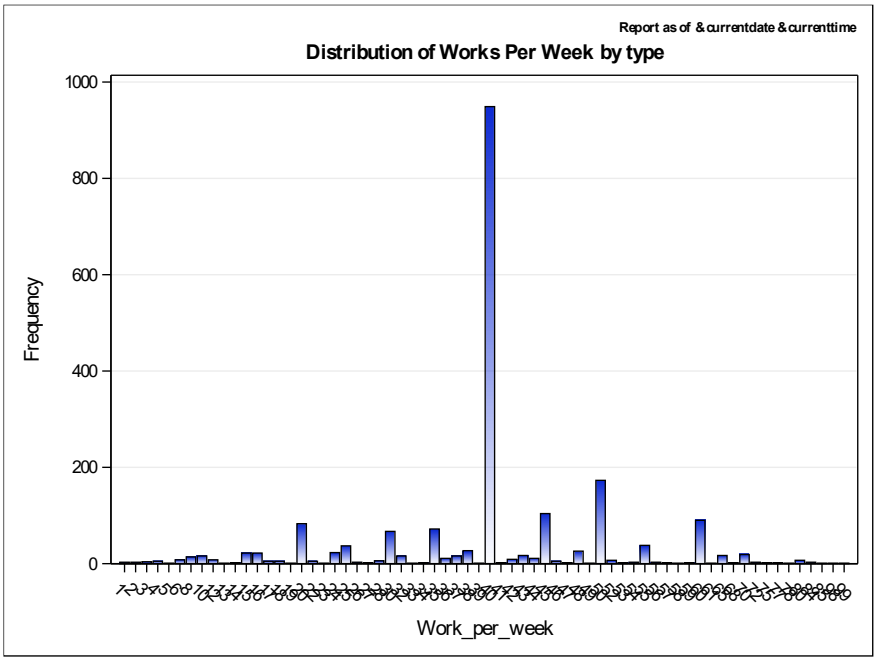


Fig 4: Distribution of work per week

Fig 5: Distribution of jobs: this figure illustrates that the highest number of responded recorded by prof-specialty by 227 respondents followed by Adm clerical however the lowest was private house servant. Below figure illustrating the overall distribution of job

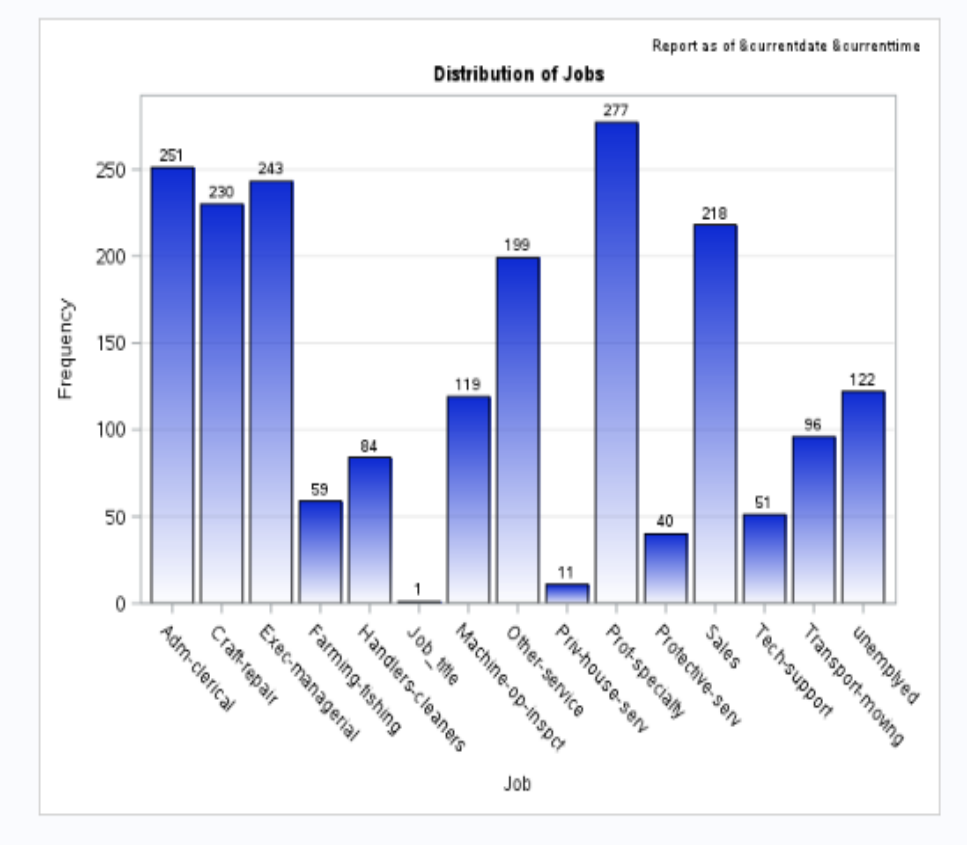


Fig 6: % of jobs and Un-employed Ratio: this table illustrated that highest number of respondent was 68.22% (1365) doing private job followed by other 10.14% (203). The least number of respondent was un-employed which is 6.1 %

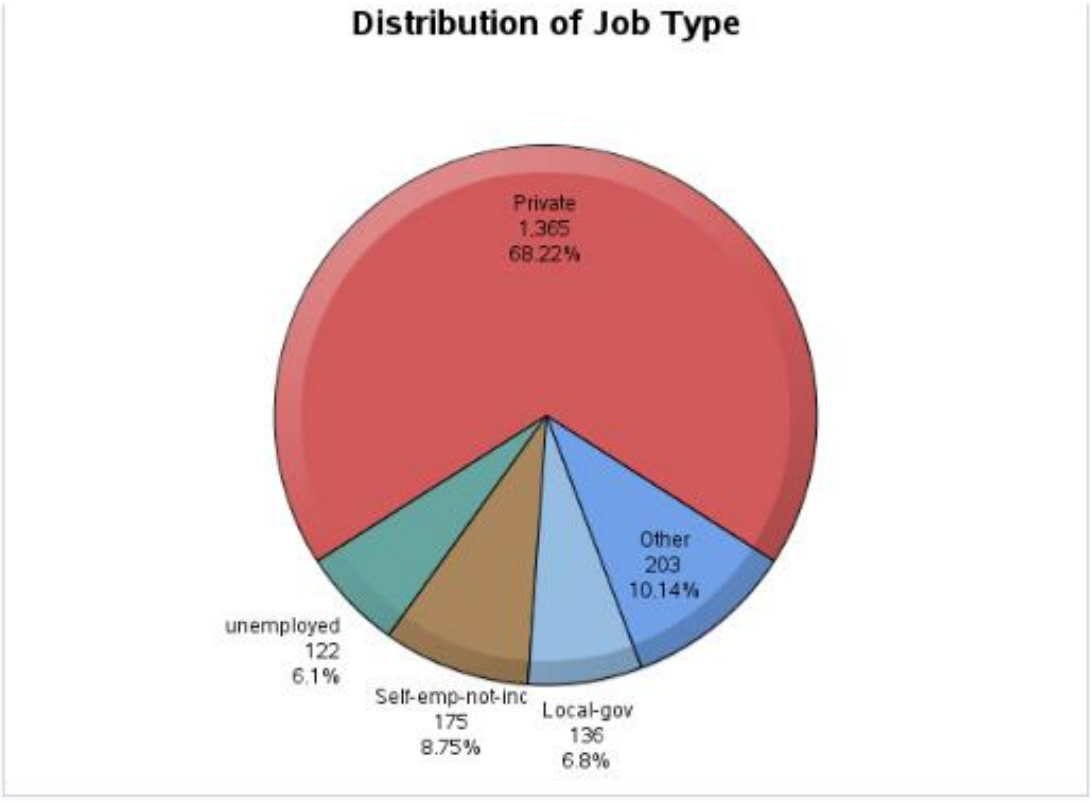


Fig 6: Distribution of Job by category

Fig 7: Distribution of Race: this table showed that highest number of respondent was white people which is 1724 followed by black peoples 180. The least number of respondent was in other groups

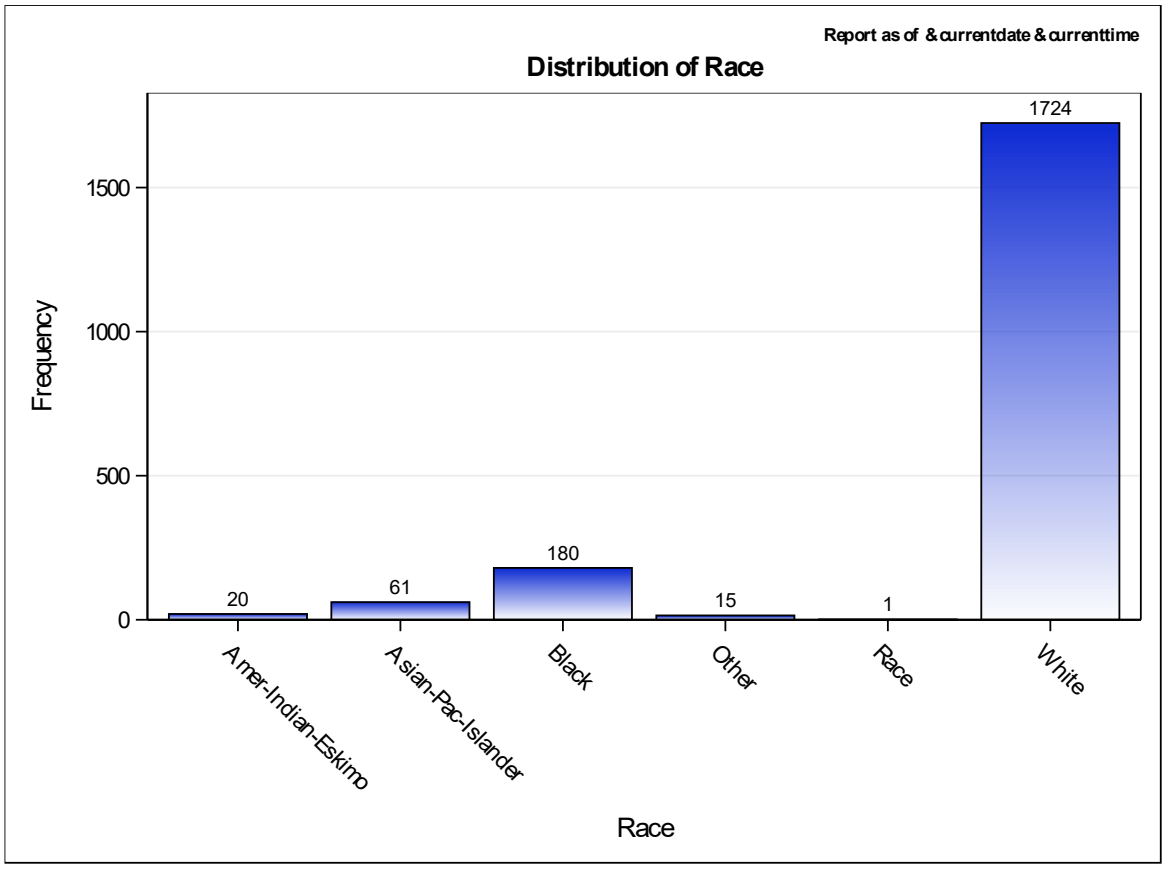


Fig 8: Distribution of Nationality: this table showed that highest number of respondent was United States which is followed by others. Below table is illustrating

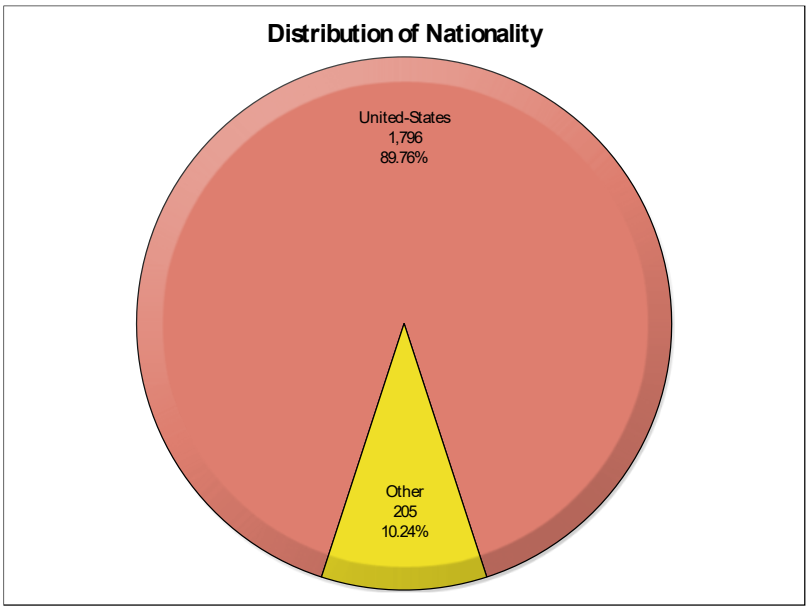


Fig 8: Distribution of Nationality

Fig 9: Distribution of Relationship status: this table showed that highest number of respondent was husbands which is followed by Not in family and lowest was others relative.

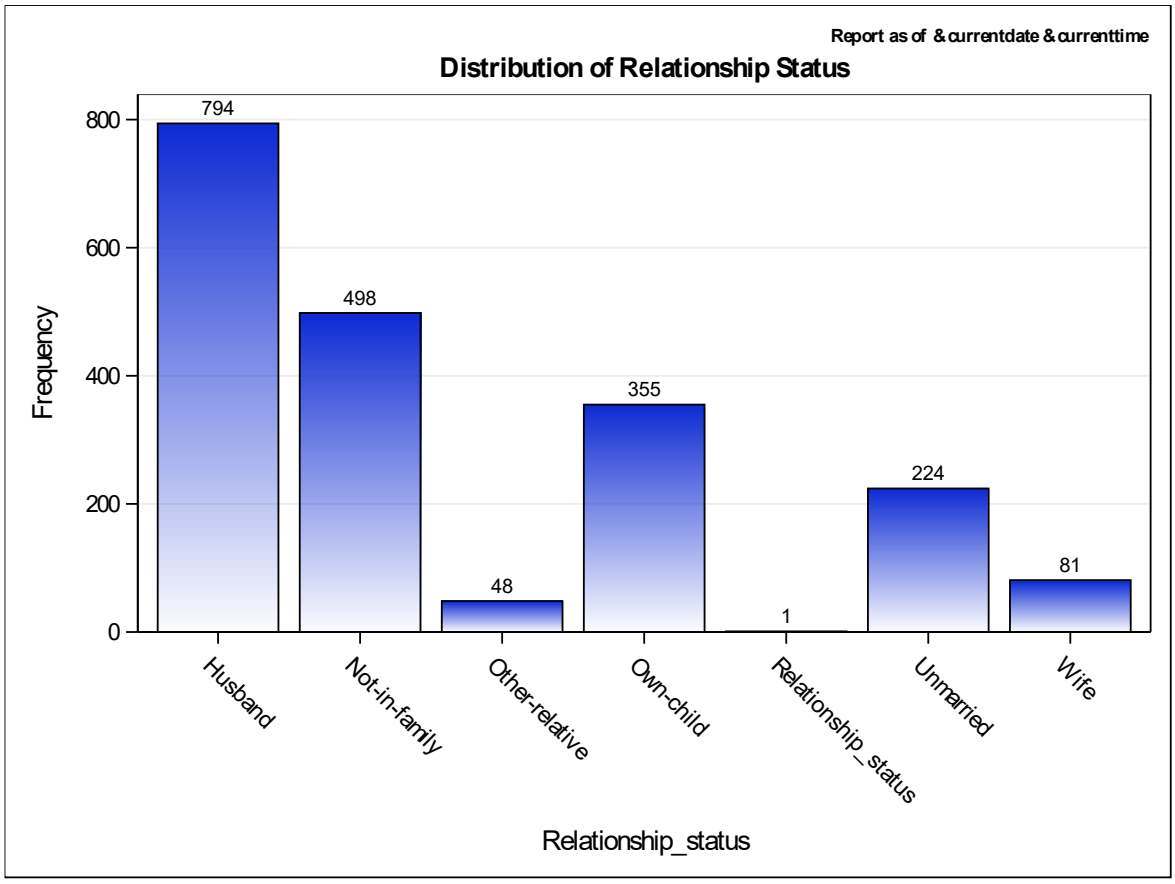


Fig 10: Distribution of Martial Status: this table showed that highest number of respondent was married which is 44.68% and followed by Never Married 33.83. Below table is illustrating

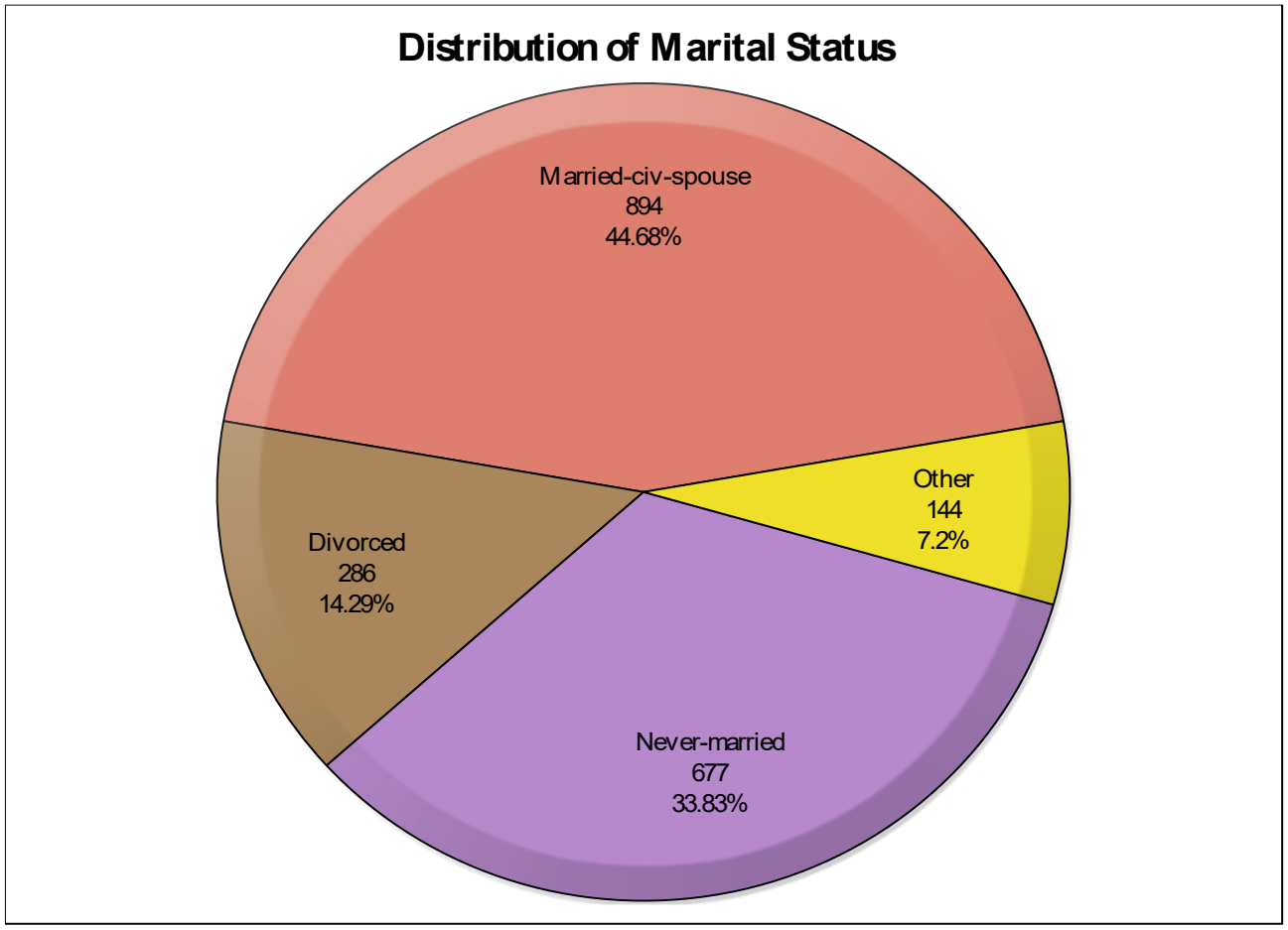


Fig 11: Distribution of Country: this table showed that highest number of respondent was United States

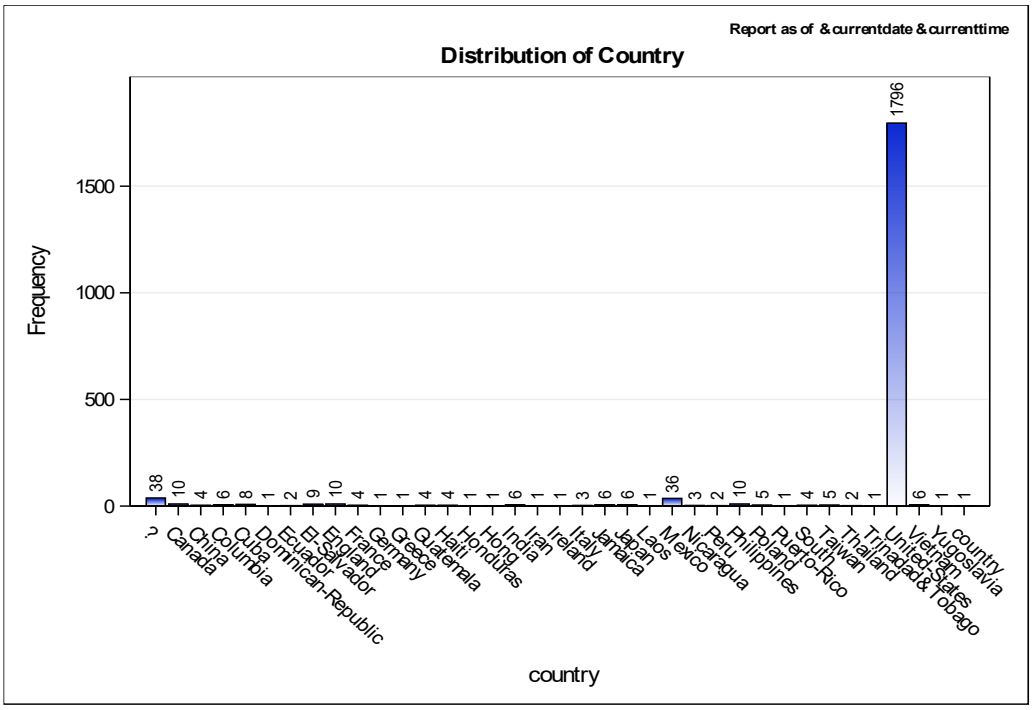
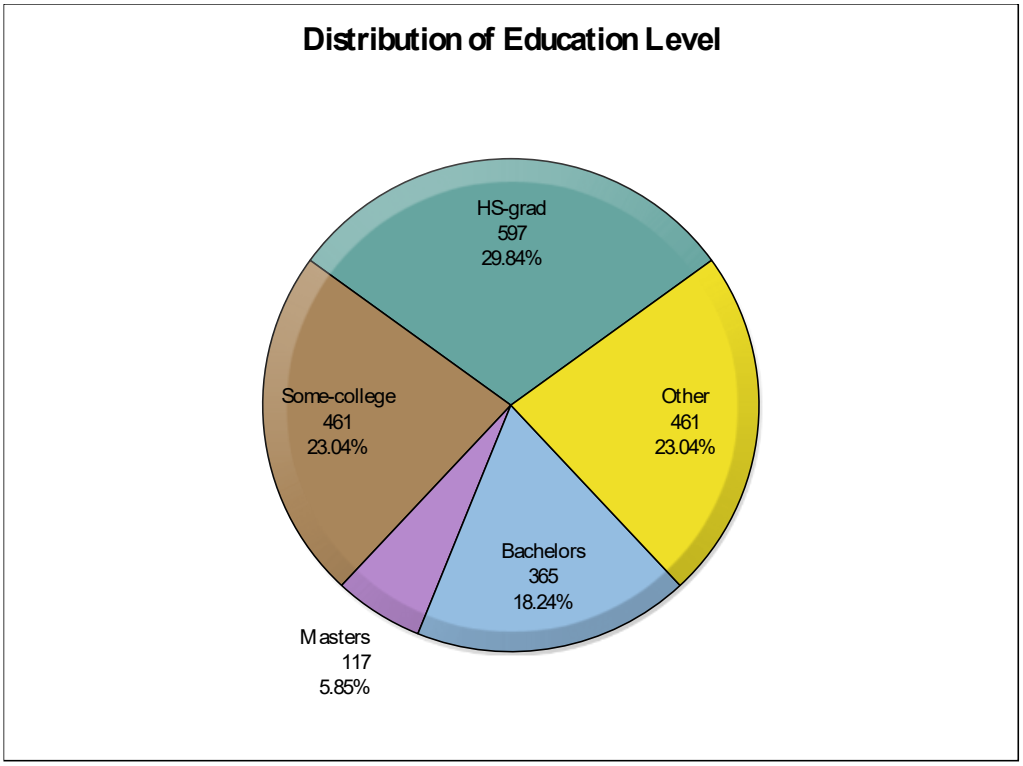


Fig 12: Distribution of Education level: this table showed that the highest of respondent was High grad which is 29.84% and other and some college sharing same number of respondent which both has 23.04 which is followed by bachelors 18.24%. the lowest respondent was masters



Task 1.1(c) Outliers:

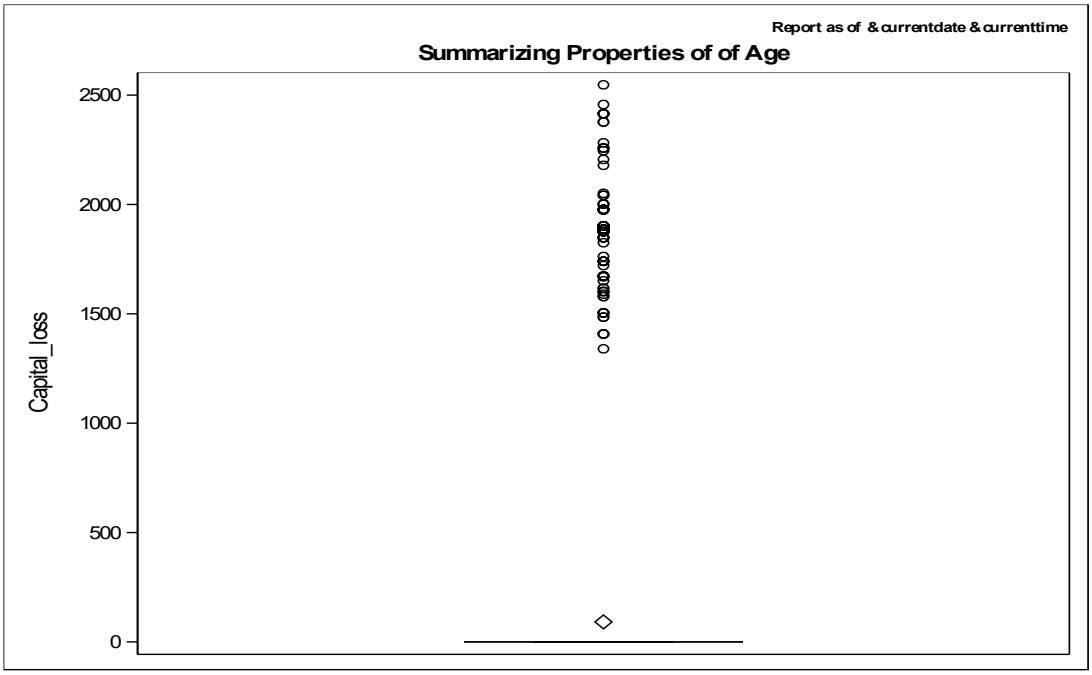
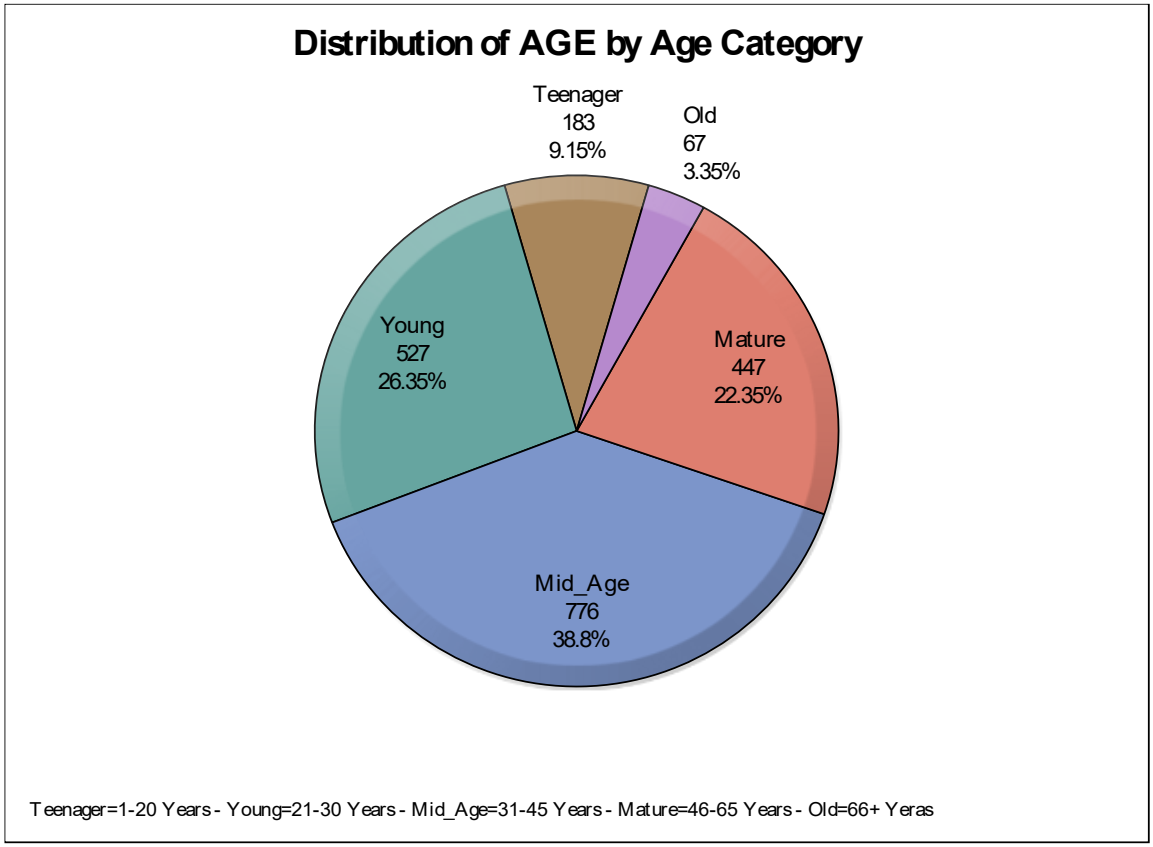


Fig 13: showing the Outliers

Data Preprocessing

Task 2.1(a) Binning Method: in this data sets we used binning method for categorize the Age in Different scales.

1. **Teenager: 1-20**
2. **Young: 21-30**
3. **Mid-Age: 31-45**
4. **Mature: 46-65**
5. **Old: 66+**



```
proc format;
    value age_tiers
        low-21='Teenager'
        21<-31='Young'
        31<-46='Mid_Age'
        46<-66='Mature'
        66<-high='Old';
run;

/*to assign format to attributes*/
data dm.profile_data;
    set dm.profile_data;
        format
        age age_tiers.;
run;
```

Fig: 14: Binning code snap shot

Task 1.2 (b):
In this data sets the job type and job attributes were need to transform them, the data was difficult to understand them. Following code snaps are illustrating the data transformation step.

```
47          Years_in_Edu 2.
48          Capital_gain 12.
49          Capital_loss 12.
50          Work_per_week 12.;
51  run;
52      proc sql noprint;
53      update dm.profile_data set Job_Type = 'unemployed' where Job_Type = '?';
54      update dm.profile_data set Job = 'unemplyed' where Job = '?';
55  quit;
```

Fig: 15: Data Transformation code snap shot

Task 1.2 (c) : Below code are representing the age attribute based on category.

```
57  proc format;
58      value age_tiers
59          low-21='Teenager'
60          21<-31='Young'
61          31<-46='Mid_Age'
62          46<-66='Mature'
63          66<-high='Old';
64  run;
65
66  /*to assign format to attributes*/
67  data dm.profile_data;
68      set dm.profile_data;
69          format
70          age age_tiers.;
71  run;
```

Fig: 16: Age category code snap shot

Task 2 Research Report on Privacy Preserving Data Publications