Categorical and Numerical Analysis of the Dataset "Adults"

Ravi Kailasam Rajendran, Department of Computer Science,Boston University, MA

Abstract: This report is based on the numerical and categorical analysis of the dataset called "Adult". This dataset is used again to study the performance of the various sampling techniques analyzed over one particular attribute in the dataset. Along with this, the different confidence intervals are also calculated and analyzed with the population mean of the dataset.

Introduction: Dataset is about the survey of adults whose age range is from 17-90 and focuses mainly on the earning of the people based on many factors which is taken from SGI database[1]. This dataset has two main classes of earnings which brings it under the binary classification. The dataset has nearly 48,841 entries and 15 attributes in total including the class attribute. This report is based on the discussion of the relation of the attributes with the class labels and about the important of each attributes in defining their each classes.

Adults: This dataset is about the survey of people and it is taken to analyze the income range of the people and finding out the important attributes that contribute the most. The two classes are <=50K and >50K earnings and the number of entries of each classes are 37,154 (76.07%) and 11,687 (23.93%). The dataset is totally imbalanced and it has a huge amount of missing values indicated with '?'. The dataset must be cleaned before doing the analysis. So the preprocessing is done on the dataset. The preprocessing is done in two steps. The first one is trying to impute the missing values and the second one is removing the entries which can't be imputed. The dataset is comprised of both the numerical and categorical attributes with latter the most. The continuous numerical attributes are age, number of years of education, capital gain, capital loss, number of hours of work per week and final weight. The categorical attributes are occupation, relationships, race, sex, native, work class, education and marital status. The main labeled class is called Earnings.

Preprocessing: The dataset is analyzed and found out that there exists more than 5000 missing values in the dataset and that needs to be cleaned. After analyzing, it is found out that there exists three attributes that all the missing values come from and they are WorkClass, Occupation and Native. The missing value counts of each of the attributes are 2799, 2809 and 857 respectively. Since the missing values matter so much in the attribute as they are the maximum or close to the maximum

[1]https://www.sgi.com/tech/mlc/db/adult.data

values, imputing the values makes it more biased towards the edge with which it is being replaced. Hence, imputing the missing values with the mean or mode of numerical and categorical attributes respectively of the corresponding attributes does not seem to be a better option in this particular dataset, it is safe to remove the missing values of attributes. Since the unknown values are represented by the '?' instead of using NA (not available), it needs to be processed as the COMPLETE.CASES function in the R library does not work. So, all the '?' are being replaced by NA and then complete.cases function is applied on the entire dataset which bring the dataset down to 45,521(7.4% fall) entries with the same attributes count. The new classes's sizes are changed in a way that the low class size is brought up to 0.86% and the dominant class is brought down by 0.86%. The fall of 0.86% in the class (Earning <+50K) makes the size down to 34,013 and the class (Earning >50K) size is now 11,208. The proportion of the classes are now 75.21% and 24.79% respectively. After the preprocessing is done, the dataset is now used for the analysis of the relation of the attributes with the classes.

Attributes with Classes: The first attribute that is analyzed with the class is the age attribute. The figure.1 shows the histogram of the age of the people in the survey. Curve shown in the blue, is the line for the people earning <=50K (class1) and the red curve shows the (class 2) which is earning >50K.

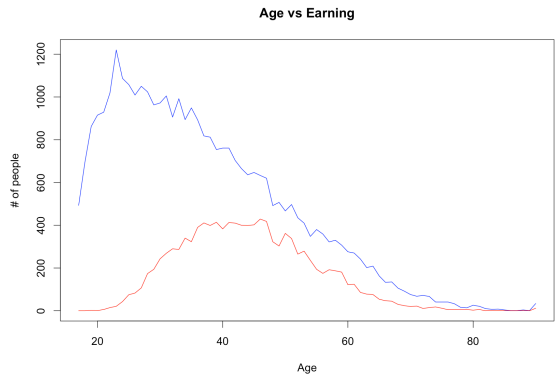


Figure1 : Age vs Earnings , the blue curve shows the earning in the range less than <=50K and the red curve show the earnings >50K.

The curve shows that the income of the people in the low range <=50K is decreasing after the main peak at age close to 20 which is a good indicator. There are some people less than age 20 who falls in the class >50K. But majority of the people in the high income falls in the age range between 30 and 50 approx.

The next attribute taken for analysis is gender in which each gender is compared with the classes according to their earnings range. Figure.2 shows the plot for the gender vs the earnings. The low income in both the gender are relatively equal. In total, there are nearly 30,526 males and 14,695 females.

MALE		FEMALE	
30,526		14,695	
<=50K	>50K	<=50K	>50K
20,987	9,539	13,026	1,669

Table1 : Gender vs Earnings distribution of people.

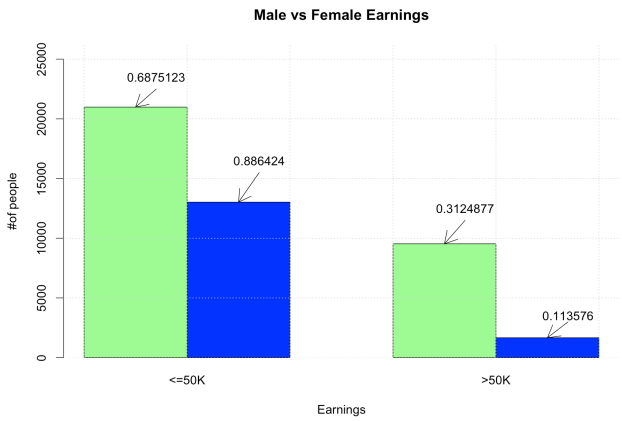


Figure2 : Gender vs Earnings of the dataset. Green shows the male category and the blue is for the female category.

But closely looking at the results, the proportions of each income group in each gender are quite relatively same. The table.1 give the idea of the distribution of the people in each gender with the earning classes.

The third attribute taken is the education and this is one of the important attributes which tells more about the people's earnings and it is a deciding factor as well. It is analyzed in two ways. First one is to find the male and female counts per category each class and the second one is to analyze the most occurred event in the attribute. The results are shown here in the figure.3. It is clearly shown in the figure. 3.a that no women who went to school up to fourth grade earns more than men do in that category. The other thing to notice here is the people earning more who has education no more than High school is very low as. From figure 3.3.b, it's safe to tell more number of people are falling under this category.

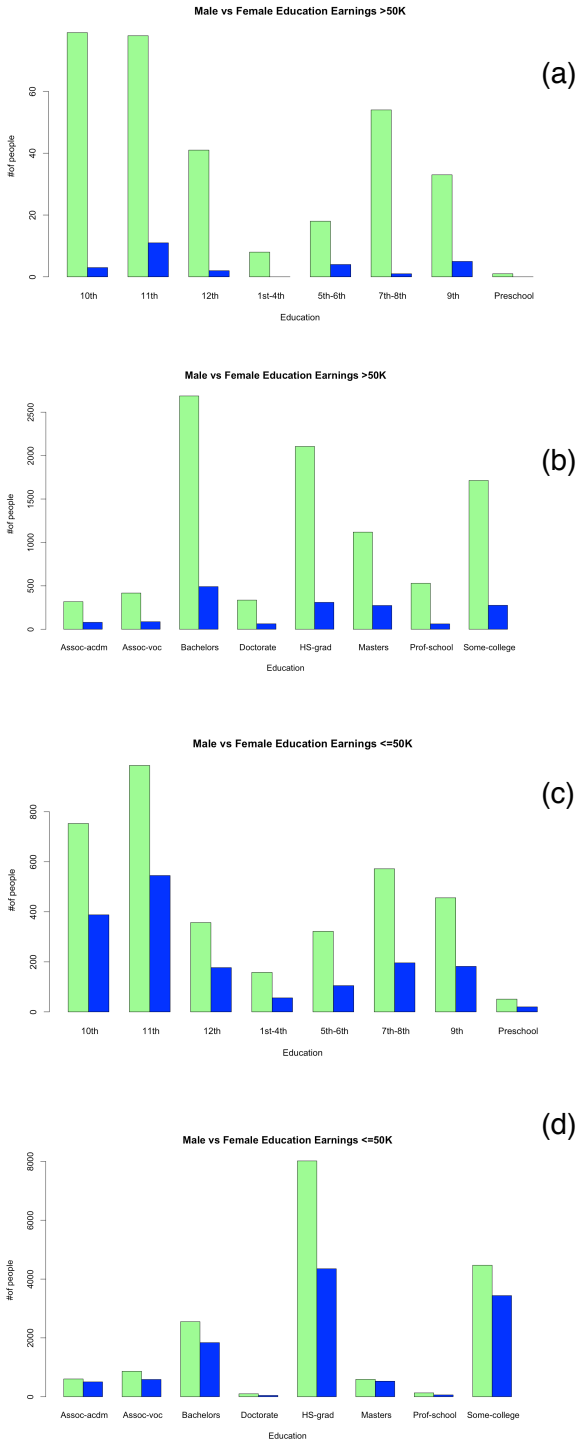


Figure3 : Education vs Earnings of the dataset.
(a) Education(up to High school) Vs Earnings >50K
(b) Education(after High school) Vs Earnings >50K
(c) Education(up to High school) Vs Earnings <=50K
(d) Education(after High school) Vs Earnings <=50K.
Green represents men and Blue represents women

From table 3.3.d, we can say that the number of people who has masters' degree and earns <=50K are same in both gender and they are relatively very low compared to other category in this comparison.

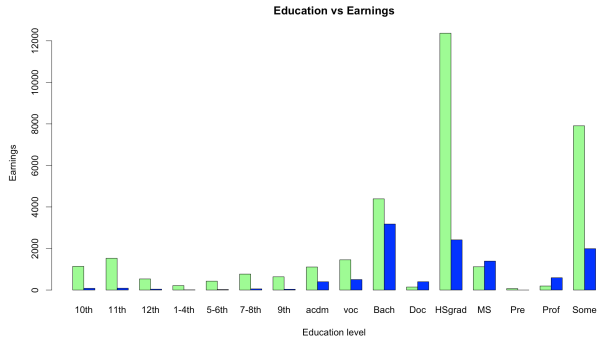


Figure4 : Education vs Gender summary of all categories in the dataset. Green represents male category and Blue represents the female category.

Seeing the cumulative graph of the above individual graphs, the high school grads are really high and analyzing that particular HS-graduate alone gives the following chart.

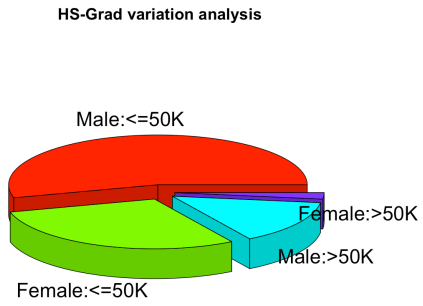
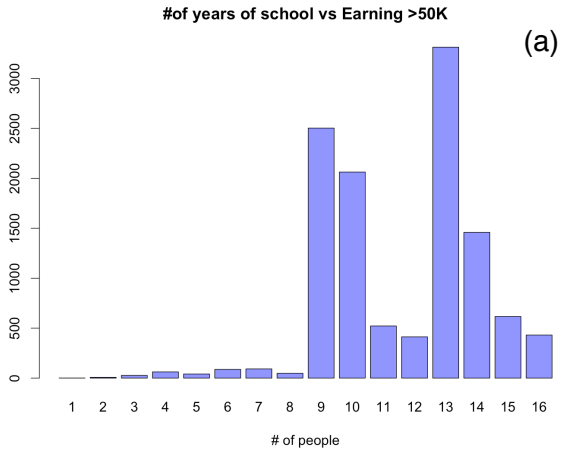


Figure5 : Analysis of the high school graduates in the entire dataset.

The above pie chart shows that more number of males have got the high school level education.

The next attribute taken for analysis is years of education. This is also another important attribute which defines the earnings. Figure 6.a shows that earnings



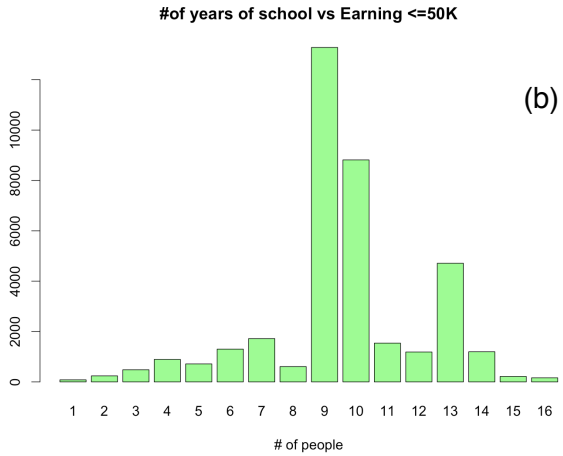


Figure6 : Number of education years vs Earnings
(a) Number of years of education for earnings >50K
(a) Number of years of education for earnings <=50K

of the people is high who went to school for more number of years. People who have gone to school for 13 or more years earns more than anyone else. The figure. 6.b shows that the normal number of school years of people in the low income earnings category is much higher than the rest of the years.

Putting the last two graph together says more details than analyzing individually.

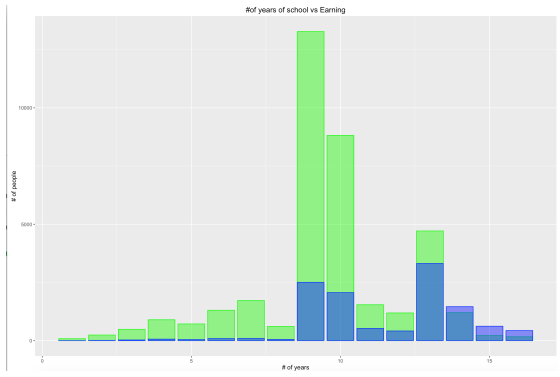


Figure6 : Number of education years vs Earnings. Green represents the low income and blue represents the high income.

The overlapping graph says that the women who went to school for 14 or more years are earning more than men in that respective education number ranges. Even though the number of tuples for the women are low, it gives the significant importance to the women.

The next attribute to analyze is the native which is the nativity of the people. This attribute is widely spread and very imbalanced. Since the survey is taken in the US, the most of the people involved in the survey are US natives. So, removing the first maximum gives us the remaining entries. But the entry Mexico being the second highest doesn't give any clear view of the other entries in the attribute. Removing both the entries gives a view about the incomes of the people from other natives. The red curve shows the people with the low income and the blue curve shows the people with the high income. The entries are numbered from 1 to 40 and they natives are sorted in alphabetical order. The last column graph clearly shows that many natives are performing good compared to the Mexico and United States natives shown in the first and second column. The difference between the two peaks at each corresponding entry is huge in both the cases and it is low in the third column graph.

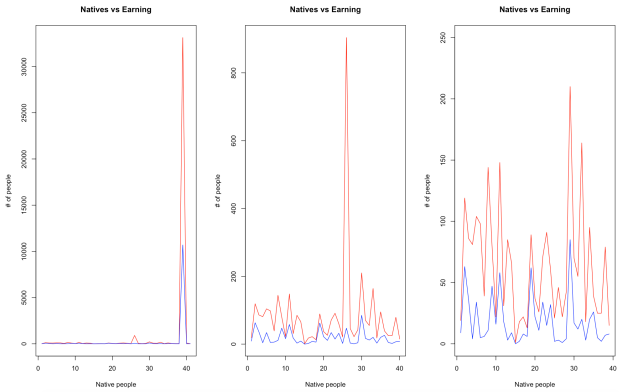


Figure7 : Nativity vs Earnings. The first column is the graph for entire attribute with no truncation in the entries. The second column shows the graph after the truncation of the dominating entry. The third column is for other entries other than first two maximums. Blue represents the high income and red represents the low income.

No people from any country mentioned here, has a higher income range than the lower income range. But more people from India and Iran are earning more income close to the people earning less income.

The next attribute is the number of hours of work per week and their reaction with the earnings. It is obviously shown here that there exists numerous peaks in the graphs and most of them are around 40 hours of work per week. Another interesting fact in this attribute is that the number of people working more than 40

hours a week are very low and more than 15,000 people who had taken this survey work for 40 hours and they are less paid.

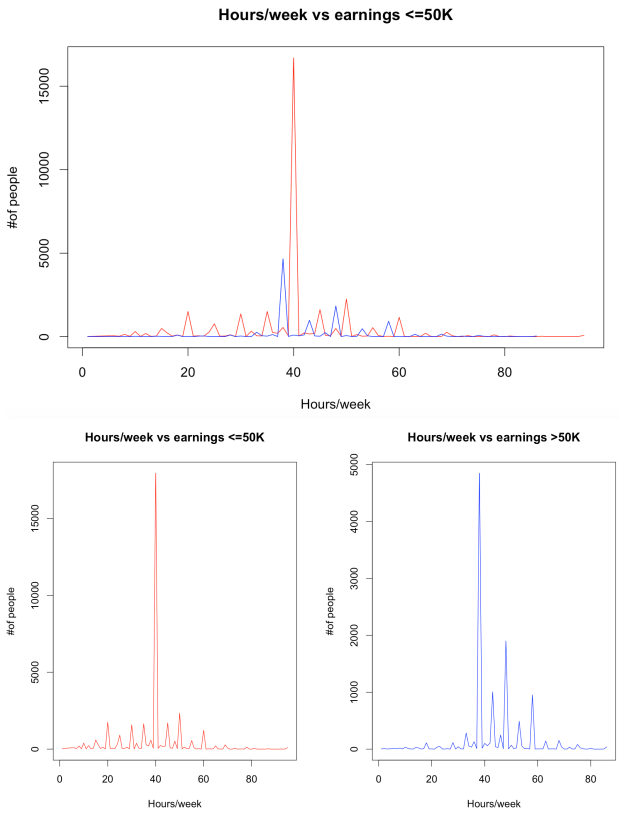


Figure8 : Number of hours of work vs Earnings. Red curve shows the low income people and the blue curve shows the high income people.

The number of people who work more than 40 hours a week are earning high.

The attribute relationship is taken for analysis with the earnings and the figure.9 shows that the people who are married are earning more and the people who are not in family and unmarried are earning very low. Especially wife, the number of the people whose income are high are quite equal to people who have low income.

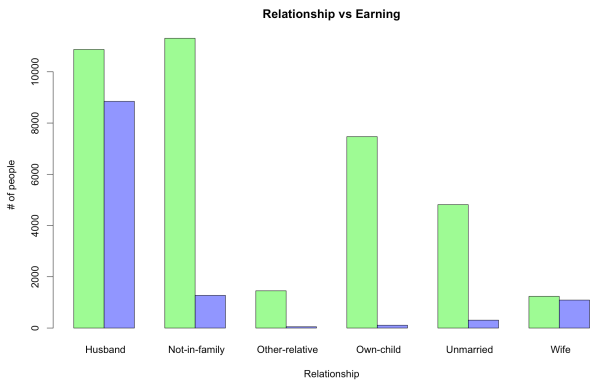


Figure9 : Relationship vs Earnings. Green represents the low income and blue represents the high income.

The next attribute is a numerical continuous attribute called FinalWeight which is a main predictor of the class.
The graphs are shown in the figure.11. This attribute has a maximum range of values in the lower end of the spectrum. So splitting the whole set into upper and lower gives some idea about the distribution of the data. There exists two clear distinctive peaks in each spectrum which has the main weightage.

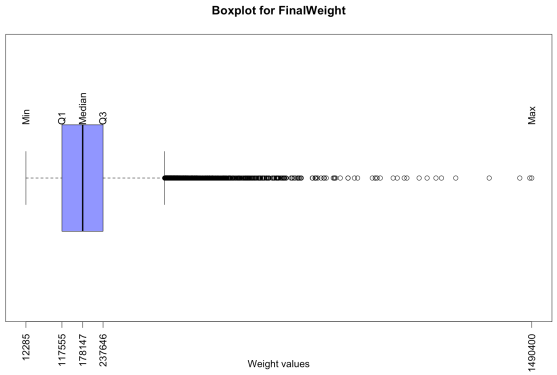


Figure10 : Boxplot of the attribute Final_Weight

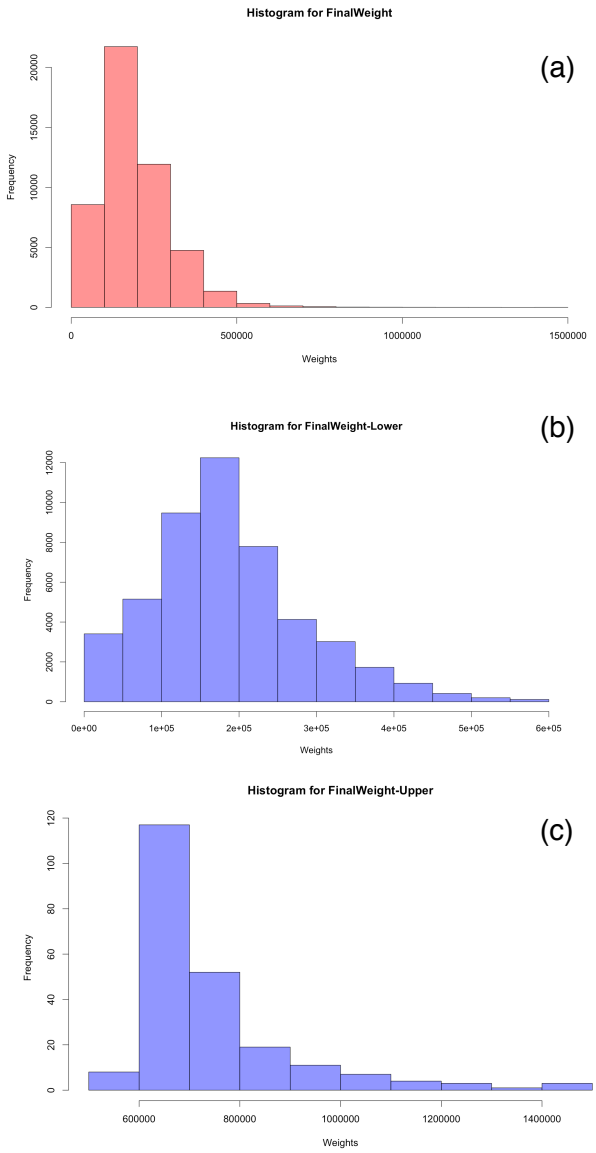


Figure10 : Distribution of the attribute Final_Weight (a) shows the whole distribution (b) shows the lower half of the spectrum and (c) shows the upper half of the spectrum.

Central Limit Theorem: The attribute final weight (numerical attribute) is taken for proving the central limit theorem. 5,000 samples are taken randomly with different sample sizes say 5, 10, 50, 100, 200, 500 and the histograms are plotted here. It is clearly shown here that the mean of the data in all samples are same and the standard deviation falls when the sample size of the sampling increases and the mean of the each samples are around the population mean.

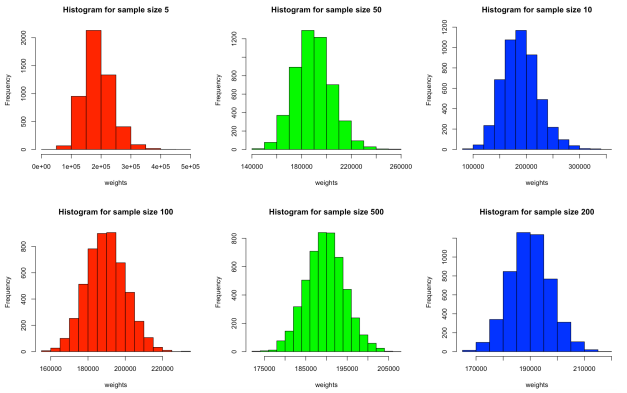


Figure11 : Histograms of the samples with different sample sizes.

The mean of the samples and the standard deviation are also shown in the graph below.

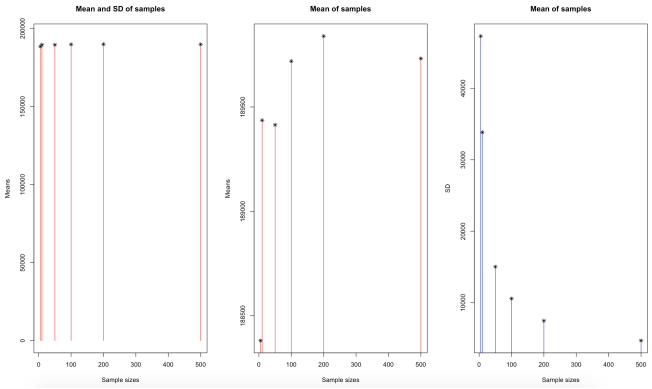


Figure12 : The mean and the standard deviation of the different samples with different sample sizes. The first one is the mean of the samples and the second one shows the variation in the means of the samples. The third one is the standard deviation of the samples.

It is observed from the graph in figure.12 that the deviation is decreasing when the sample sizes are increasing exponentially for this particular dataset.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	TYPES	Cambodia	Canada	China	Columbia	England	France	Holand	Honduras	Hong	India	Italy	Japan	Mexico	Philippines	United State	Yugoslavia
2	SRS-WR	3	43	20	17	27	6	0	3	3	28	25	16	223	61	9068	6
3	SRS-WOR	5	37	24	15	25	7	0	5	9	31	23	10	212	69	9167	2
4	SYS-EQUAL	9	29	26	15	19	8	0	4	6	31	11	18	178	62	8249	6
5	SYS-UNEQUA	0	3	2	2	6	0	0	1	1	15	14	12	148	52	9534	4
6	STRATA	6	36	25	18	26	8	1	4	6	33	22	20	199	63	9130	5

Figure13 : The Native attribute entries count after various sampling techniques

Sampling: The analysis of the attributes are done one by one and now some sampling techniques are going to be used over the data and here the best sampling methods for the dataset is going to be chosen. The sampling techniques used here on the dataset are simple random sampling with and without replacement, systematic sampling with equal and unequal probabilities. It is observed that the sampling methods are not doing the perfect sampling for the entries in the attributes which has the least probability.

In the figure.13, we can find that the entry Holland , none of the sampling techniques picked it after sampling. Even when doing systematic sampling with both the equal and unequal probability, it's not taken. Strata sampling does a good job in this dataset.

```
Sample.size :  5
80% confidence intervals = 124835.4  -  245737.8
80% confidence intervals = 114983.2  -  235885.6
80% confidence intervals = 82337.81  -  203240.2
Sample.size :  10
80% confidence intervals = 178474.8  -  263965.6
80% confidence intervals = 114302.8  -  199793.6
80% confidence intervals = 142237.1  -  227727.9
Sample.size :  50
80% confidence intervals = 176338.4  -  214571.1
80% confidence intervals = 180163.4  -  218396.1
80% confidence intervals = 187108.9  -  225341.6
```

Figure14 : Confidence intervals for the dataset of 80% for different sample sizes 5,10,50.

Confidence Intervals: The confidence intervals are found for 80% and 90% and it is finally shown in the graph that the population mean falls between the upper and lower end of the intervals. Figure.16 shows that when the sample size is increased, the range comes close to population mean.

```
Sample.size :  100
90% confidence intervals = 189594.1  -  224337.8
90% confidence intervals = 174309.2  -  209052.8
90% confidence intervals = 163396.6  -  198140.3
Sample.size :  200
90% confidence intervals = 166835.6  -  191403.1
90% confidence intervals = 174147.9  -  198715.4
90% confidence intervals = 172323.1  -  196890.6
Sample.size :  500
90% confidence intervals = 180658.8  -  196196.7
90% confidence intervals = 186675.8  -  202213.7
90% confidence intervals = 178751.4  -  194289.2
```

Figure15 : Confidence intervals for the dataset of 90% for different sample sizes 100, 200, 500.

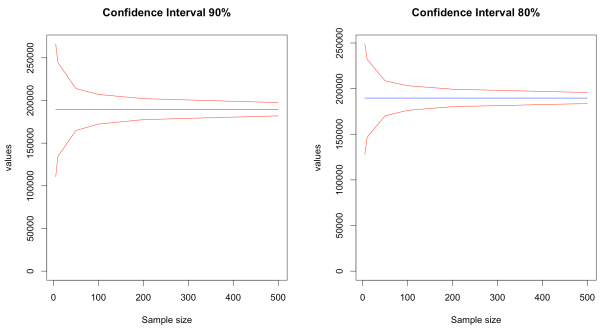


Figure16 : Confidence intervals for the dataset of 80% and 90 %. The red curves are the upper and the lower limits and the blue line is the population mean.

Figure.16 shows that when the sample size is increased, the range comes close to population mean.

Results: All the attributes in the dataset are analyzed and the relation of attributes with classes are discussed and the respective graphs are plotted. The replacement of the missing values in the dataset was a time taking process.

Sampling techniques are found to be not helpful in this dataset except the strata sampling which has the ordering of the attributes issue. The categorical values are to be converted to the numerical data for the systematic sampling with unequal probability.