**Homework Assignment 1**

**STA 141A A02**

**Submitted By:**

**Question1:** How many observation are recorded in the dataset? How many colleges are recorded?

**Solution 1:**

**a)** There are total **3312** rows and **51** columns which are recorded from the given dataset. This can be found out using *dim()* command on the dataset.

**b)**Total number of colleges recorded are **2431** as based on the main campuses count.

**Question2:**  How many features are there? How many of these are categorical? How many are discrete? Are there any other kinds of features in this dataset?

**Solution 2:**

**a)**The total number of features is equivalent to number of columns in the given dataset that is 51. Class of data as present in the dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Type** | Character | Factor | Integer | Logical | Numeric |
| **Frequency** | 4 | 4 | 15 | 3 | 25 |

Table 1: Different Types of Data Distribution

We obtain the above result using *sapply* function on class.

**b)**Total number of categorical values include all values of character and all values from logical variable so total 7 are categorical. As categorical variables are those whose variable value are in category.

**c)**There are total 25 discrete feature are there that is of numeric class.

**d)**There is another type called continuous feature means value of variable will vary as a continuous pattern.

**Question 3:** How many missing values are in the dataset? Which feature has the most missing values?Are there any patterns?

**Solution3**:   
**a)**There are total **23197** many NA in the dataset which can be obtained by using sum function *on is.na()* function which is used to find total number of NA in the data-frame.

**b)**The feature ***avg\_sat*** has most missing value which can be obtained by finding maximum column on is.na() function.

**c)** The data supposed to have more private details which have more numbers of NA then data which has public details. For example: the mean SAT score is private to a college based on which student selection into the college takes place whereas average faculty income or spend\_per\_student data is more or less made public by the college.

**Question 4**: Are there more public colleges or private colleges recorded? For each of these, what are the proportions of highest degree awarded? Display this information in one graph and comment on what you see.

**Solution 4:** **a)** There are total 716 total public colleges recorded and 2596 private colleges

**b)** Below is the table for proportion of highest degree awarded

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | Public | Nonprofit | For Profit |
| **Frequency** | 0.2161836 | 0.5163043 | 0.2675121 |

Table 2: Proportions of Public, Non Profit and Profit Organizations

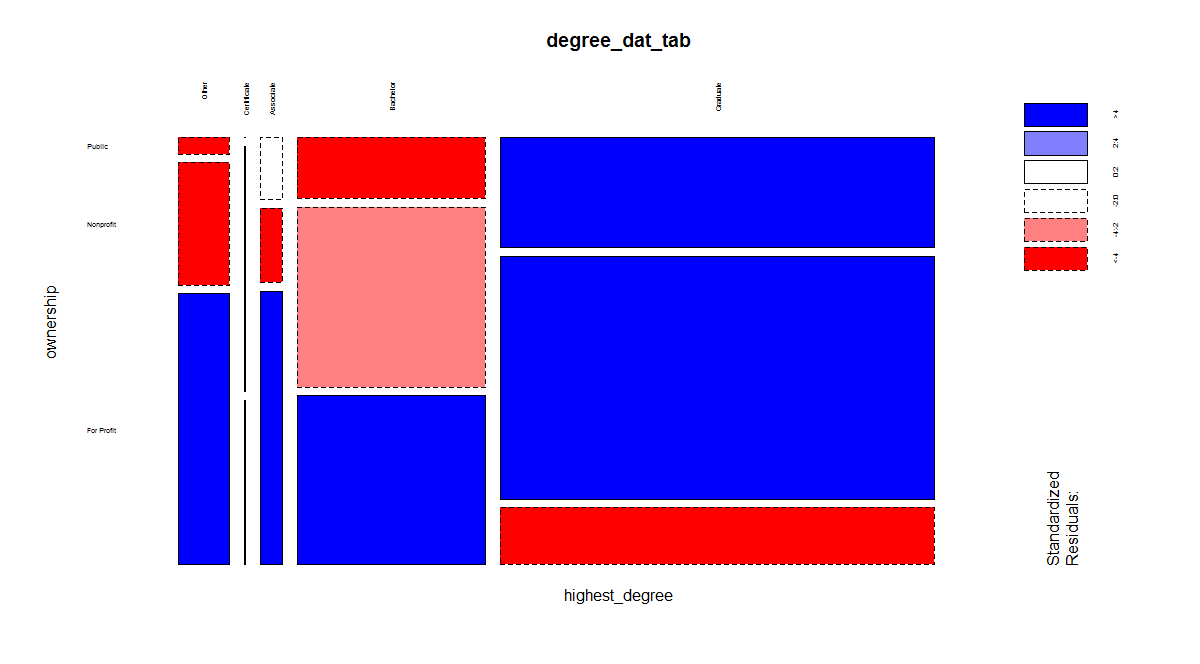


Figure 1: ownership v/s highest\_degree

**c)** More proportion of Graduate as compared Bachelor and other degree's as clearly seen from the graph.

**Question5:** What is the average undergraduate population? What is the median? What are the deciles? Display these statistics and the distribution graphically. Do you notice anything unusual?

**Solution 5:**

**a)**Average undergrad population is mean of the following undergrad distribution. On executing command summary() on the undergrad population we obtain the following table . Mean of this distribution is **3600.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Features | Min | 1st Qu | Median | Mean | 3rd Qu. | Max. | NA's |
| Values | 0 | 428 | 1295 | 3600 | 3372 | 166800 | 490 |

Table 3: Summary of Undergrad Population

**b)** Median for the distribution is **1295.**

**c)** Decile iseach of ten equal groups into which a population can be divided according to the distribution of values of a particular variable. The deciles can be found out using this following command

*quantile(hwdata1$undergrad\_pop[length(which(!is.na(hwdata1$undergrad\_pop)))], probs = seq(0, 1, length=11))*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Percentages | 0% | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
| Values | 0.0 | 153.0 | 319.2 | 536.0 | 847.6 | 1295.0 | 1811.8 | 2674.5 | 4550.8 | 9629.8 | 166816.0 |

Table 4: Table for deciles

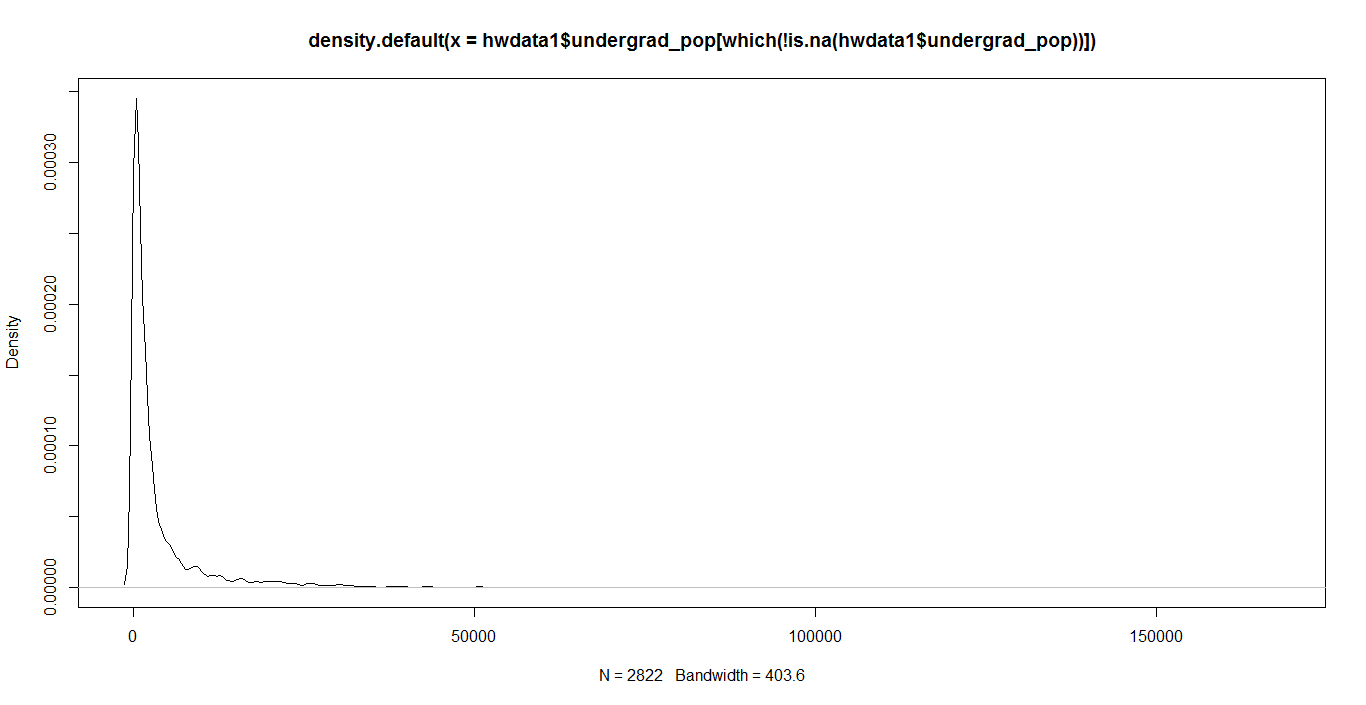
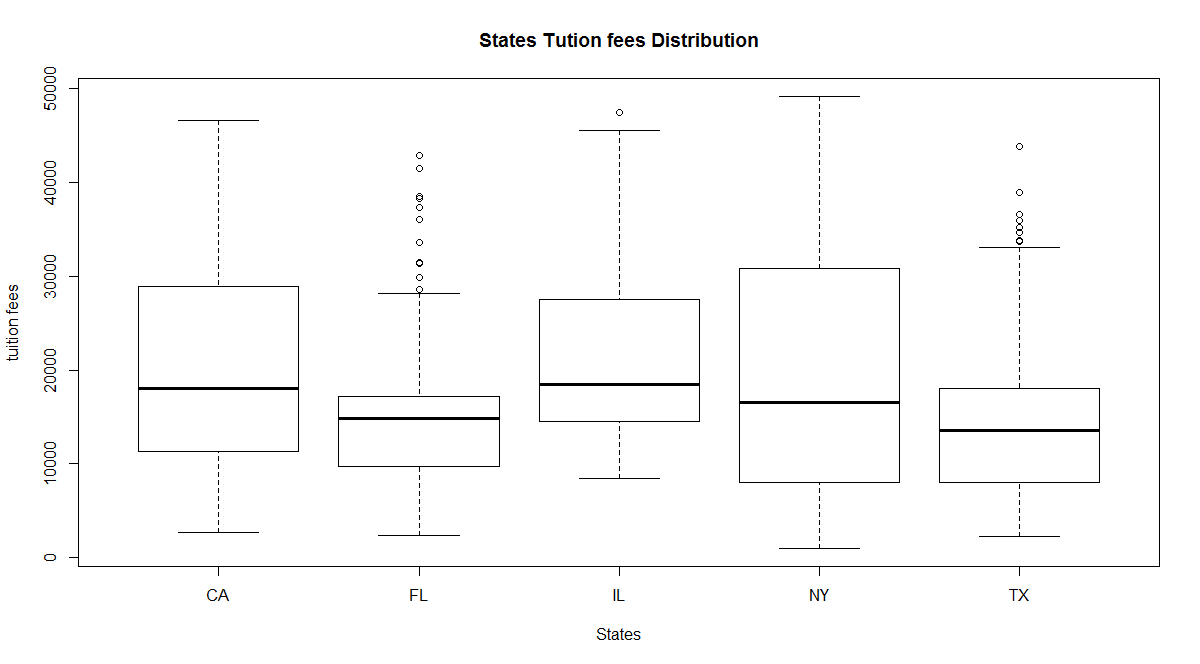
**d)** **

Figure 2: Density Plot for undergrad population based on frequency

**e)** The unusual thing about this data is only one college contains undergrad population above 150000 and all other college contain undergrad population below 70000.

**Question 6:** Compare tuition graphically in the 5 most populous states. Discuss conclusions you can draw from your results.

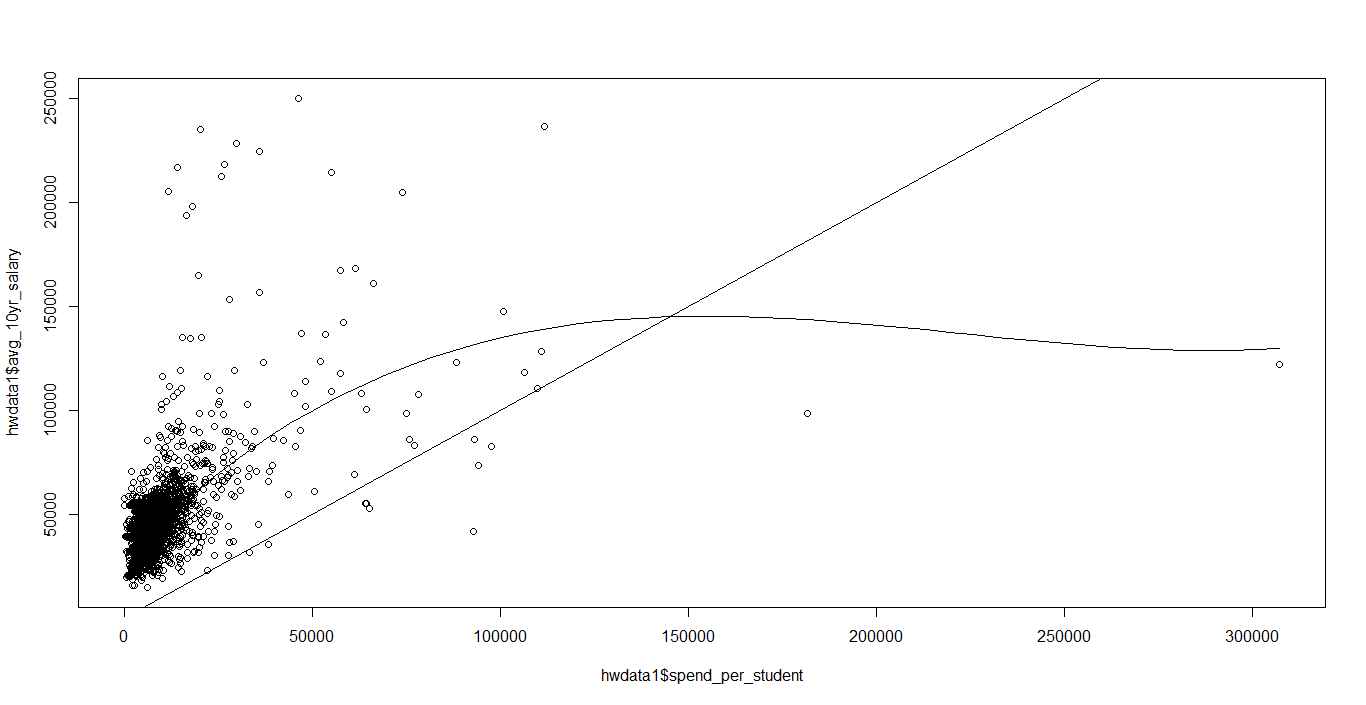
**Solution 6: a)**

 **Figure3: Above Sketch is Box-plot for 5 countries**

**b)** Based on the data as displayed in the Box plot above the average tuition fees for IL and California has high tuition fee as compared to Florida, Texas and New-York. So IL and California colleges are more expensive as far as tuition fee is concerned.

**Question 7.** Display and comment on how spending per student (by the college) and students’ 10-year earnings are related. Is this relationship affected by whether a college is public, nonprofit, or for profit?

**Solution 7: a)**



**Figure 4:Smooth Scatter graph between *spending per student* and *avg 10yrs salary.***

Based on the above graph the conclusion can be drawn that the for avg. student population the avg. earning is more than spending.

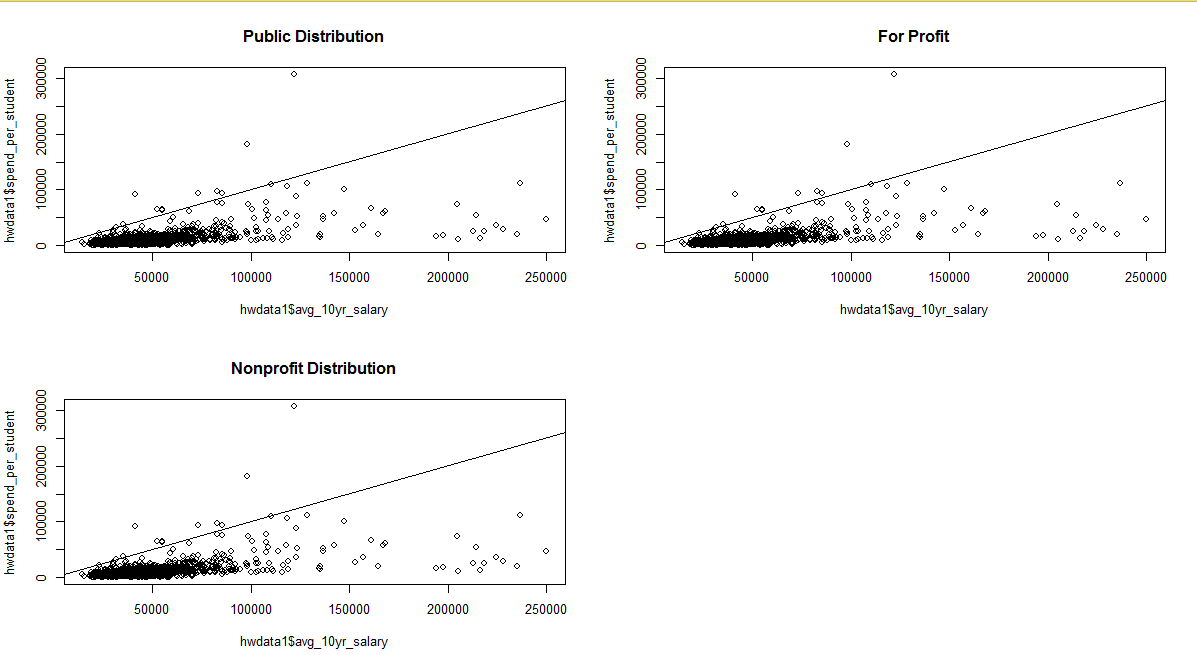


Figure 5:Smooth Scatter graph between spending per student and avg 10yrs salary for Public and Private ownership

**b)** The relationship is not affected by weather the college is public or private as clearly seen by data.

**Question 8:** Which colleges give the best earnings for the cost? Explain how you determined this. Discuss limitations of your result and features2 you did not examine that could confound your result.

**Solution 8:**

**a)**"United States Merchant Marine Academy","Augusta University", "South Texas University" etc. are the university gives highest ratio of avg. salary to cost. For getting best earning by cost for top colleges we need to make a new column having this ratio value and by making data frame of it..Then ordering the data-frame by ratio values.

**b)** Limitation of considering avg\_salary salary is that it cannot determine for a particular person as the data is scattered between high earning as well as low earning students so by seeing the avg. ratio we cannot determine exactness in our conclusion. This data can also be misinterpreted if a person has moderate earning and low spending cost.

**Question 9:** Which colleges are the most racially diverse? Explain the strategy you used to determine this.

**Solution 9:**

**a)** American Conservatory Theater, Southern California University of Health Sciences, Lincoln College of New England-Hartford are some of the college which is most racially diverse.

**b)**I determined this by extracting data related to all race i.e.(race\_white, race\_black...etc.) making data frame of it. Finding mean for all data containing races. Then making data frame of college names and their respective mean. Then order them to find most diverse college list.

Question 10. How does UC Davis compare to other colleges in the nation? Use statistical summaries and graphicsto examine at least 3 characteristics that students might be interested in.

Solution 10. **a)** UC Davis can be compared to other colleges based on the avg student earning, student population in various disciplines ,tuition fees and many more factors.

**b)**

|  |  |  |  |
| --- | --- | --- | --- |
| Colleges | Avg\_10yrssal | Undergrad pop | Tuition |
| UC-DAVIS | Mean: 65200  Median: 65200 | Mean: 26483  Median: 26483 | Median: 13895  Mean: 13895 |
| Other Colleges | Mean: 46654  Median: 43100 | Median: 1295  Mean: 3600 | Mean: 17611  Median: 15495 |

Table 5 Comparison between UC Davis and other colleges features

Based on the data provided in the above table the comparison is done based on 1. tuition fee for college 2. Undergrad. pop 3. Avg\_10yrs salary for students so we can clearly able to see the all factors are more for UC Davis except the tuition fee. Based on this we can call UC Davis as an above average or a good university in United States with more undergrad students, more salary for the student employed and less tuition fee.

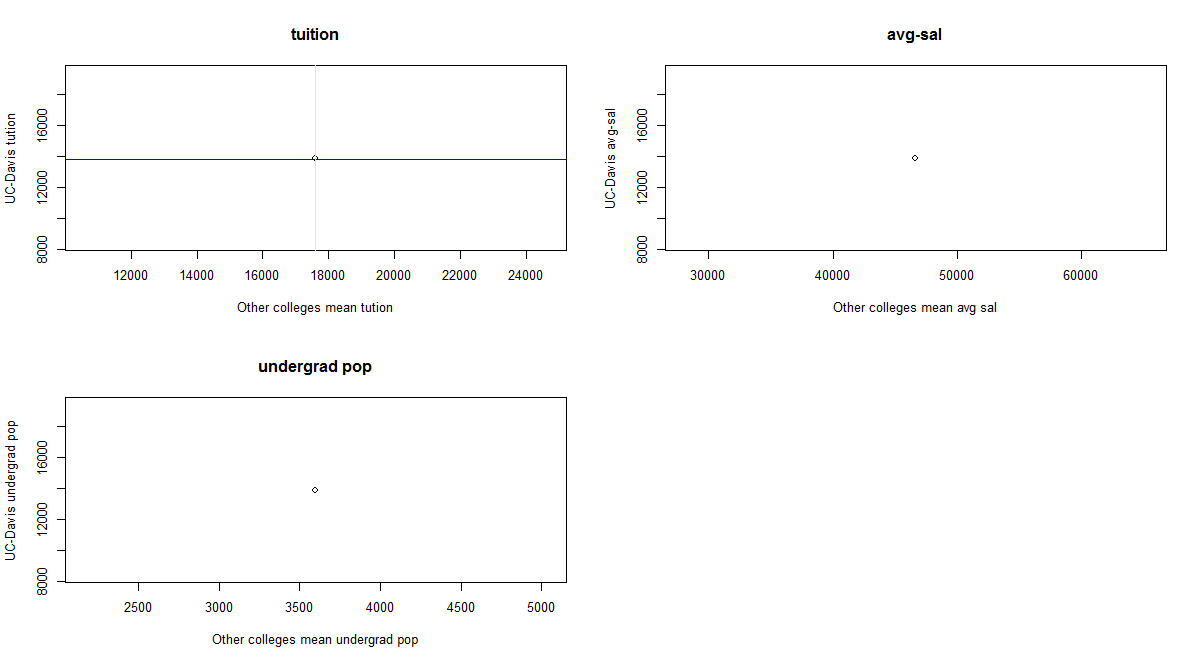


Figure 6: UC-Davis features v/s other colleges features mean value

**References:**

1.) **From Wikipedia**: https://en.wikipedia.org/wiki/List\_of\_U.S.\_states\_and\_territories\_by\_population

2.) **From Stack overflow**: https://stackoverflow.com/questions/18304672/sapply-in-r-how-to-use

: https://stackoverflow.com/questions/tagged/ggplot2?sort=faq

3.)**Student reference**: Mik.

4.)**Piazza:** Instructor and T.A's help.