Assignment 1: Exploring Data

Ricco Amezcua
CS422 Data Mining
Department of Computer Science
Illinois Institute of Technology

Feburary 11, 2013

Abstract

This is a report for the first assignment of CS422 Data Mining. In this report, two sets of data are looked at; one which contains informatin on properties of red and white wine and another which contains information on various adults. This report is intended to show the exploration of data using the R language. Various histograms and plots are created from the datasets of wine quality and of economics information of various persons.

1 Problem Statement

1.1 Wine Quality Set

This part involves observing the data set of a various wines. The data set contains various properties of the wines including a quality rating which was given by wine tasters.

First, the red wine data set is looked at. It is separated into high quality and low quality where high quality is a quality rating greater than 5 and low quality rating is a quality of 5 or lower. Then the mean and standard deviation of all of the properties of both high and low quality wine are looked at. From this a description of the differences of high quality and low quality wine is given. Then the correlation is plotted for *residual sugar*, *total sulfur dioxide*, and *alcohol* on a scatter plot matrix for both low and high quality wine.

Afterwards, the red and white data set is merged. From this two histograms are created which give the frequency of the quality of the wines. Then the first 50 red wines and first 50 white wines are sampled and merged. From this data, a parallel coordinates plot is created which contains the attributes: *citric acid*, *residual sugar*, *density*, and *quality*.

1.2 Adult Data Set

This part involves looking at a data set of various adults. The attributes include *age*, *education, occupation*, *capital-gain*, and various others.

The first part involved creating various graphs. The graphs created include: a histogram for *race* for all people who are native to *United-States*; a box plot of *education*, *capital-gain*, and *hours-per-week*; and a three dimensional plot that shows the relationship between *work class*, *race*, and *hours-per-week*.

Later, the *education levels* and *marital status* are looked at. Then the *hours per week, education* and *age* are compared.

2 Proposed Solution

The graphs used for exploring the data sets was done in R Studio using the R language.

3 Implementation details

The R scripts can be found inside the *code* folder. The scripts are separated in to four scrips: question 1 part a, question 1 part b, question 2 part a, and question 2 part b. They are ran by running all of the commands in order.

The code must be in the same file as the data files. Every script will create at least one graph. Question 1 part a will create four files in a folder called *data*. These four files are the mean and standard deviation of the low quality wines and high quality wines. The data for these files are found in table 1 and table 2.

There were no major problems with creating the graphs. Most of the difficulty was in learning the R language. One problem did occur in the creation of the document. There is a problem with importing the PDF images of the graphs into a LaTex on Mac. A solution was not found in a reasonable amount of time so JPEG images were substituted instead. This did not effect the quality of most graphs except for figure 5.10, because of its density and how RStudio exports images at a low resolution. It had to be exported as a large PDF and converted in to a JPEG by a third party program. PDF versions of the graphs can be found in the *data* folder.

4 Results and discussion

4.1 Wine Quality Set

The data shows that there is not much difference in the properties of the low quality and high quality red wines. Looking at the properties of the red wines in table 1 and table 2, it is apparent that low quality (quality rating 5 and below) and high quality (quality rating 5 and above) share similar mean values. The property *Total Sulfur Dioxide*, differs the most between the two, where low quality had a mean of 54.65 and high quality 39.35. The contents of *Total Sulfur Dioxide* also differs the most as it has the highest standard deviation of any wine property, and differs more so than low quality wines. This is confirmed in figure 4.1. In figure 4.1, where blue is high quality red wine and red is low quality red wine, *Total Sulfur Dioxide* is shown to differ between the two, while *Residual Sugar* does not. The *Alcohol* content also looks like it differs in figure 4.1, but looking at table 1 and 2, the means between low and high quality only differs by 0.93.

Looking at a sample of the first 50 white wines and first 50 red wines, it can be seen that the wine data sets does not offer a good representation of many different qualities of wines. From figure 4.2, many of the wines in the sample had a quality rating of 6. Separating the wines into a high quality (those with a quality rating above 5) and low quality (those with a quality rating 5 and below), it can be seen that there are much more high quality wines than low quality. Looking at the wine properties in figure 4.4, the *Residual Sugar* stayed the same for the majority of wines, while the *Citric Acid* and *Density* varied much between the wines.

4.2 Adult Data Set

There is not much variation between those of the *United* States. From figure 4.5, over 25,000 adults are white. The second largest race is black adults with a count of 4,000. The count for all other races was small.

The *education level* of all adults had a large spread in figure 4.6. The average is about 10th grade, but it deviates a lot; it goes from 5th grade all the way to college and above. However, looking at figure 4.7, most adults reported a capital gain of 0. There are many outliers to this, which means of the adult population, only a small few have some sort of capital gain, no matter their education level. Looking at the *hours worked per week* for all adults in figure 4.8, one can see that while the average for most adults is 40, the deviation is large as there are many adults who are working anywhere from no hours up to 100 hours per week. Furthermore, figure 4.9 shows that races and work class does not have an effect of how many hours are worked per week.

Looking at *marital status* and *education level* in figure 4.10, it seems that higher education (*some-college* and above) tends to be pursued by those that are divorced, widowed, in a civil union, or have never married. Also those adults that are divorced, in a civil union, or never married have the most deviation in their *education* level.

The figure 4.11 compares *hours per week, education* and *age*. Looking at the graph, adults between the ages of 30 and 60 tend to work 30 or more hours per week. However, adults 60 and to 80 tend to work 25 or less hours a week, unless they have a lower education, in which they can work up to anywhere close to 70 hours per week. Adults 80 and above and young adults around 20 tend to exclusively work 25 hours or less.

5 References

1. http://cran.r-project.org/manuals.html

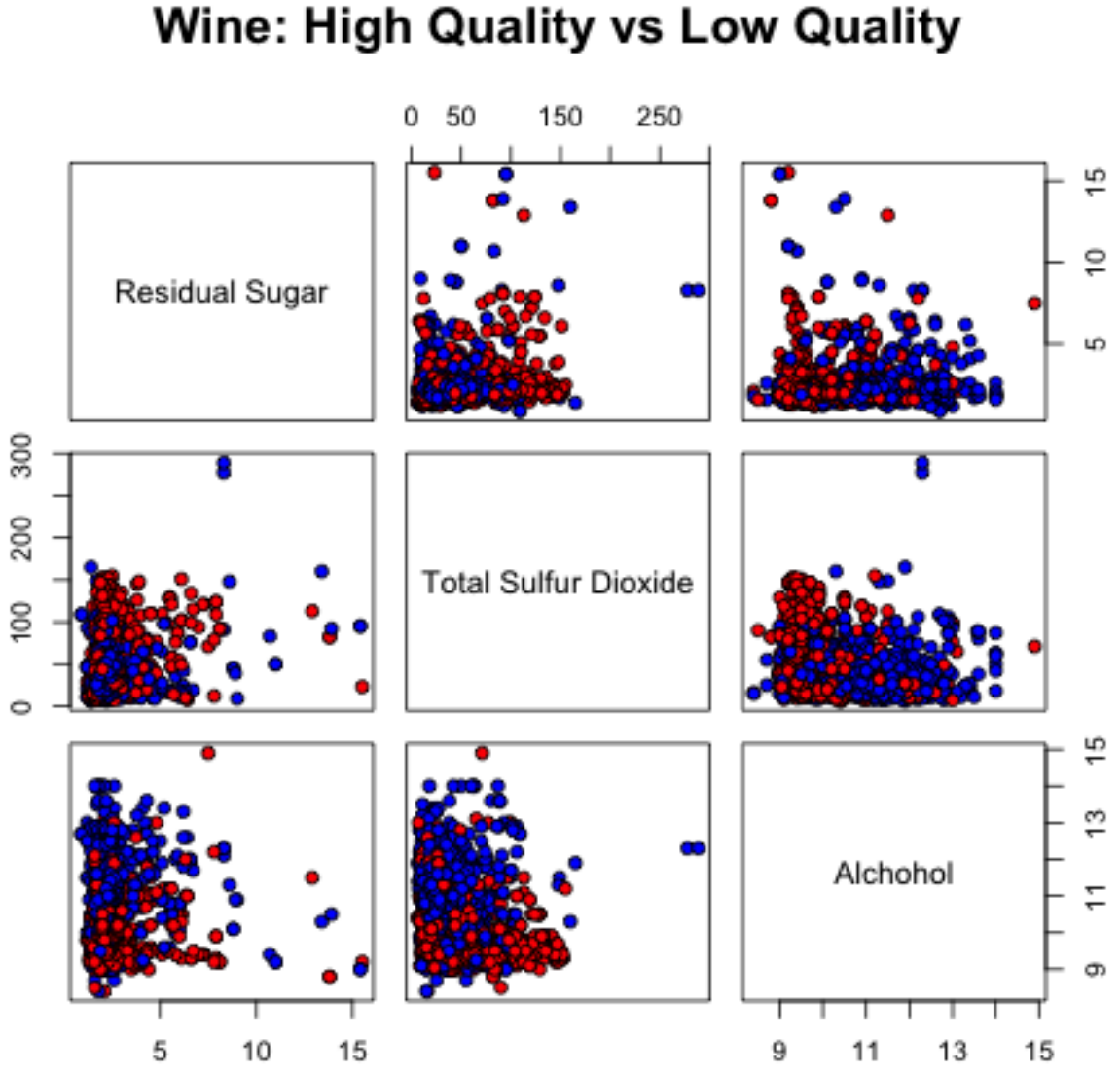
Wine: High Quality vs Low Quality
Residual Sugar
Total Sulfur Dioxide
Alchohol
0
50
150
250
5
10
15
0
100
200
300
9
11
13
15

Figure 4.1:

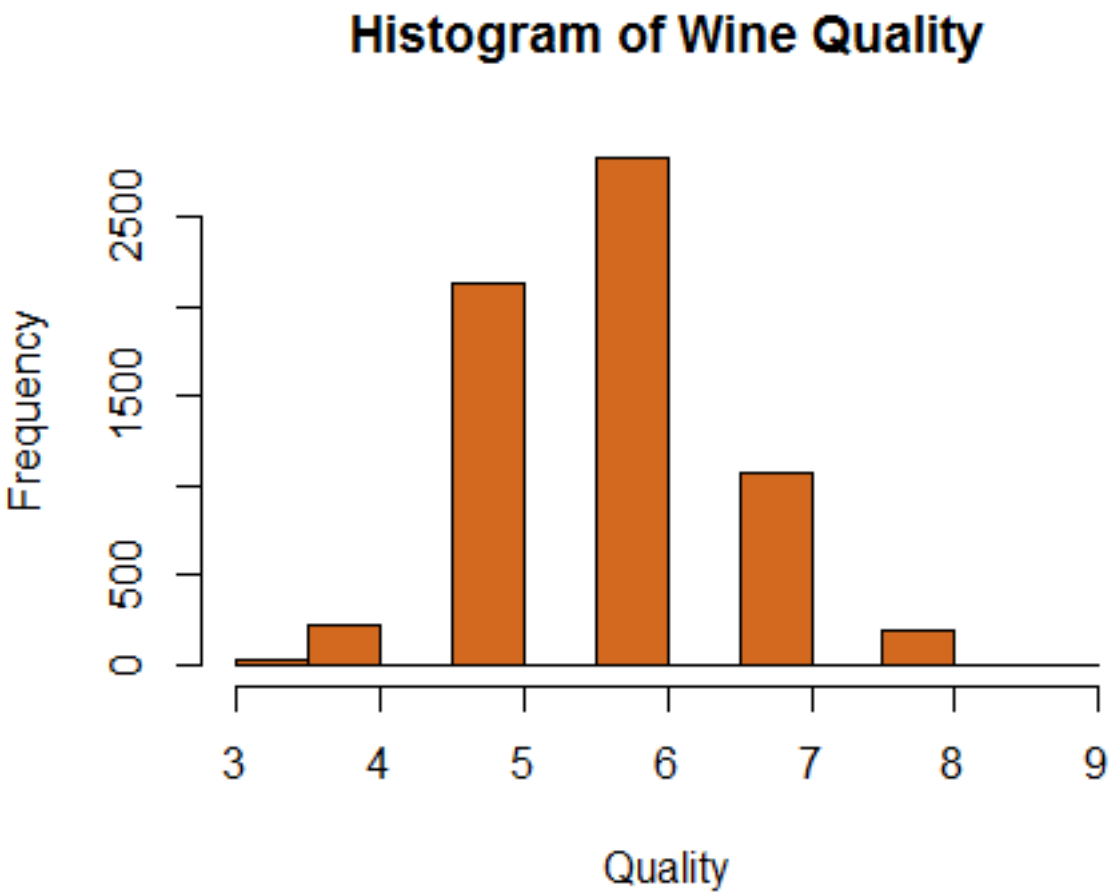
Histogram of Wine Quality
Frequency
0
500
1500
2500
3
4
5
6
7
8
9
Quality

Figure 4.2:


Histogram of Wine Quality (Two Bins)
Frequency
0
1000
3000
0
2
4
6
8
10
Quality

Figure 4.3:

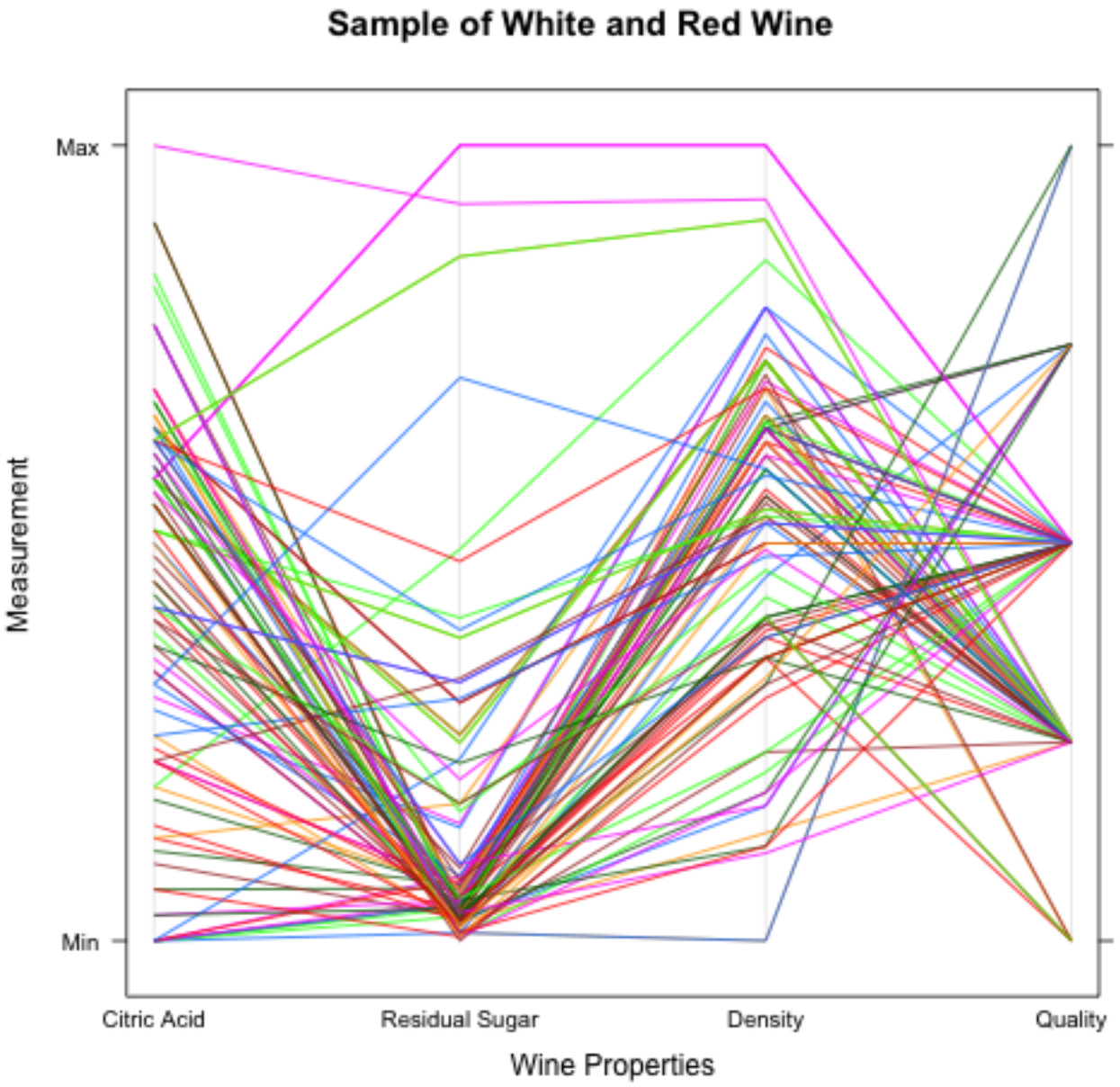
Sample of White and Red Wine
Max
Min
Measurement
Citric Acid
Residual Sugar
Density
Quality
Wine Properties

Figure 4.4:

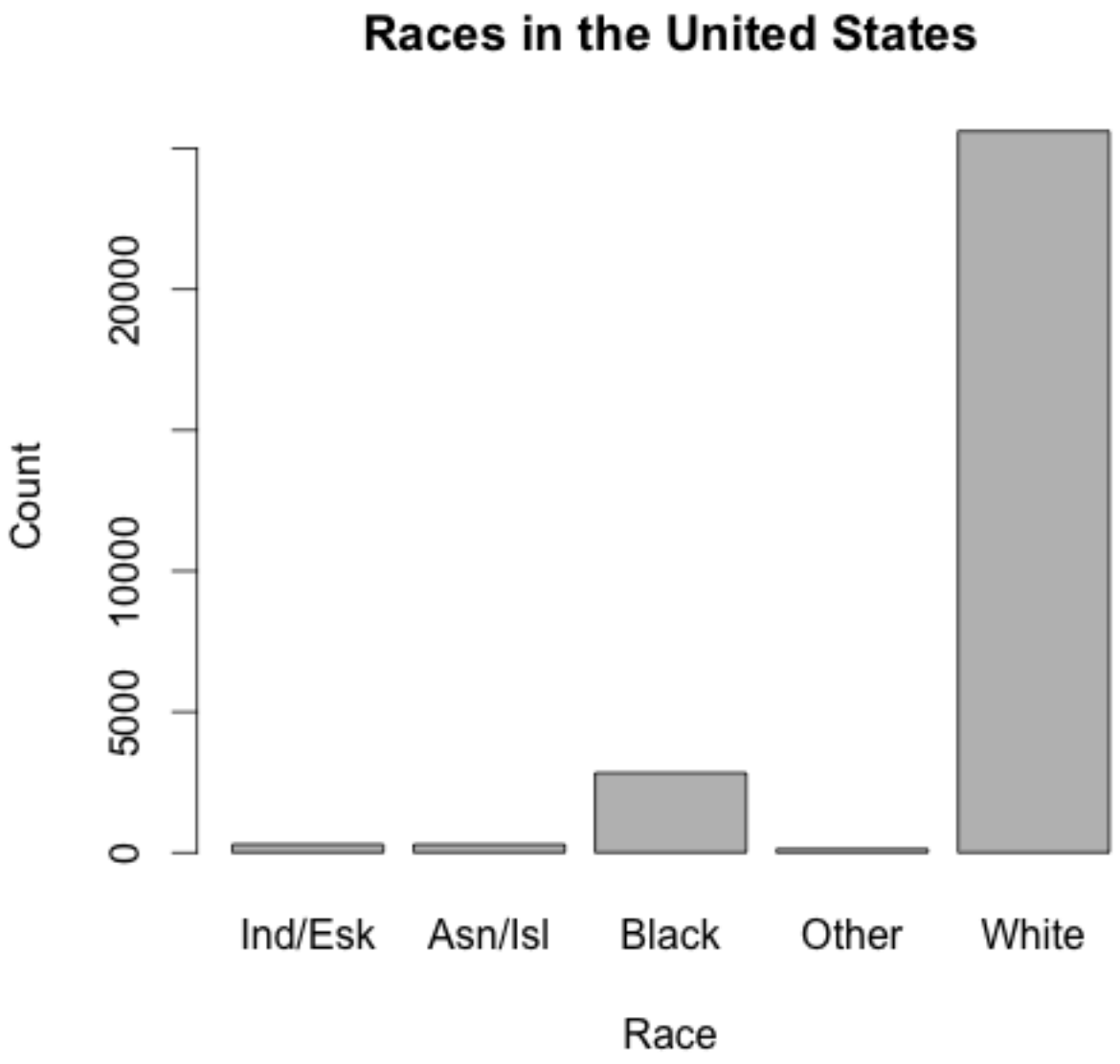
Races in the United States
Count
0
5000
10000
20000
Ind/Esk
Asn/Isl
Black
Other
White
Race

Figure 4.5:

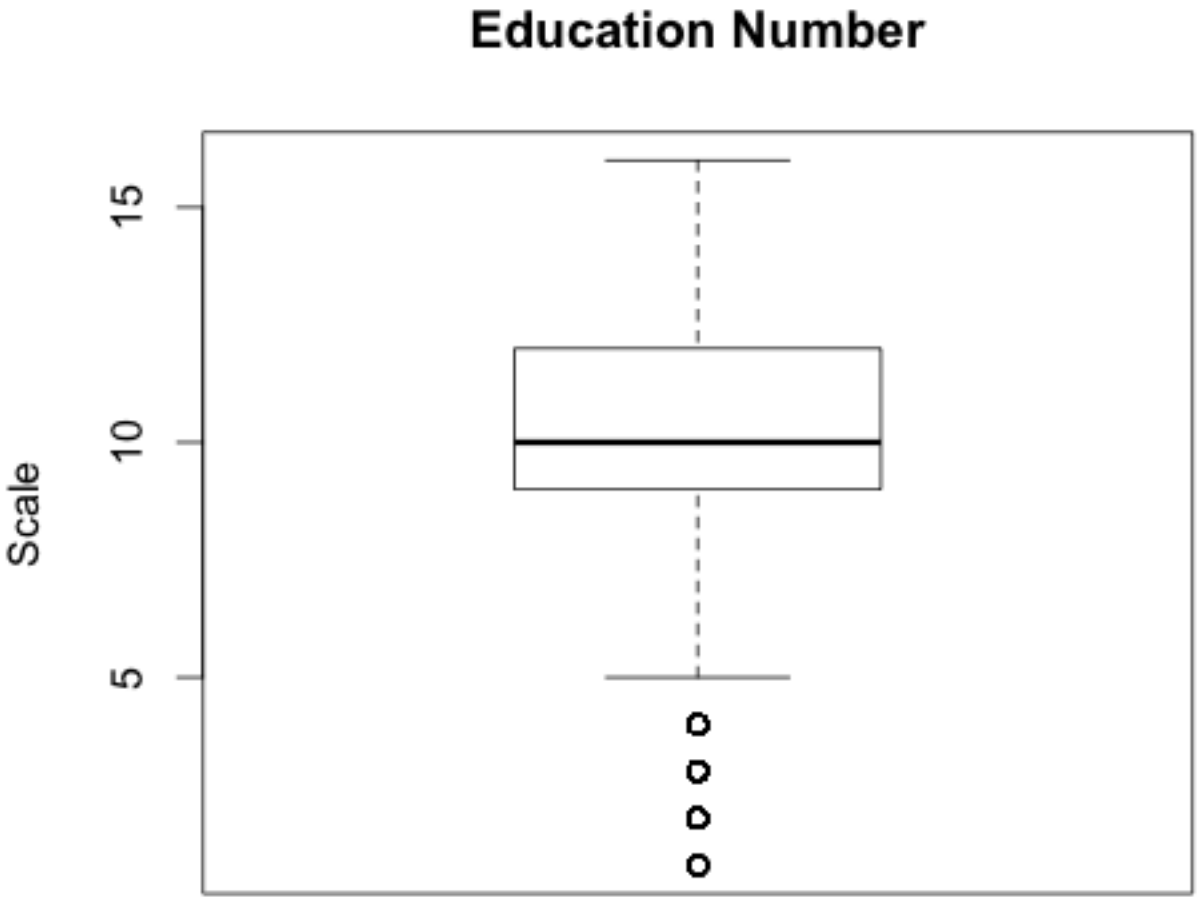

Figure 4.6:

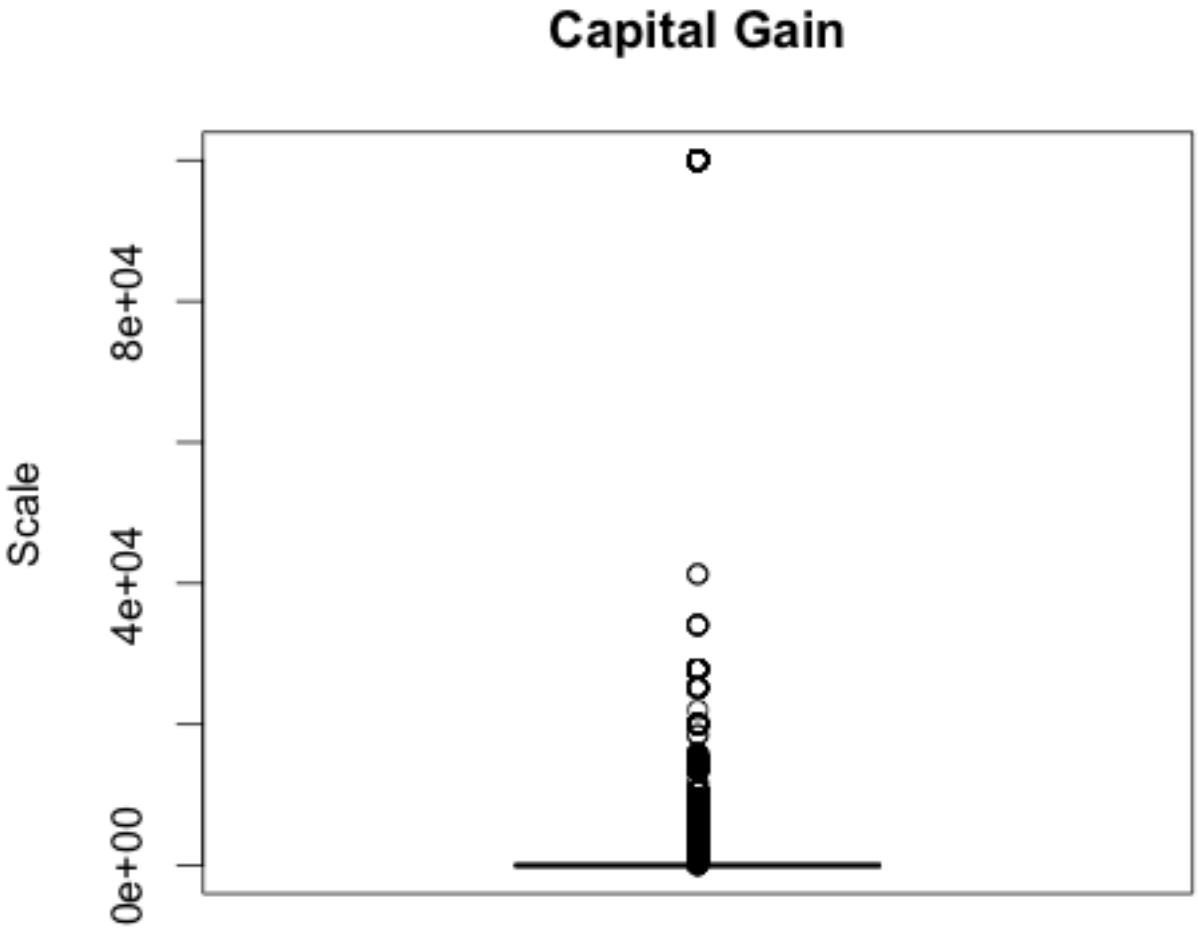

Figure 4.7:

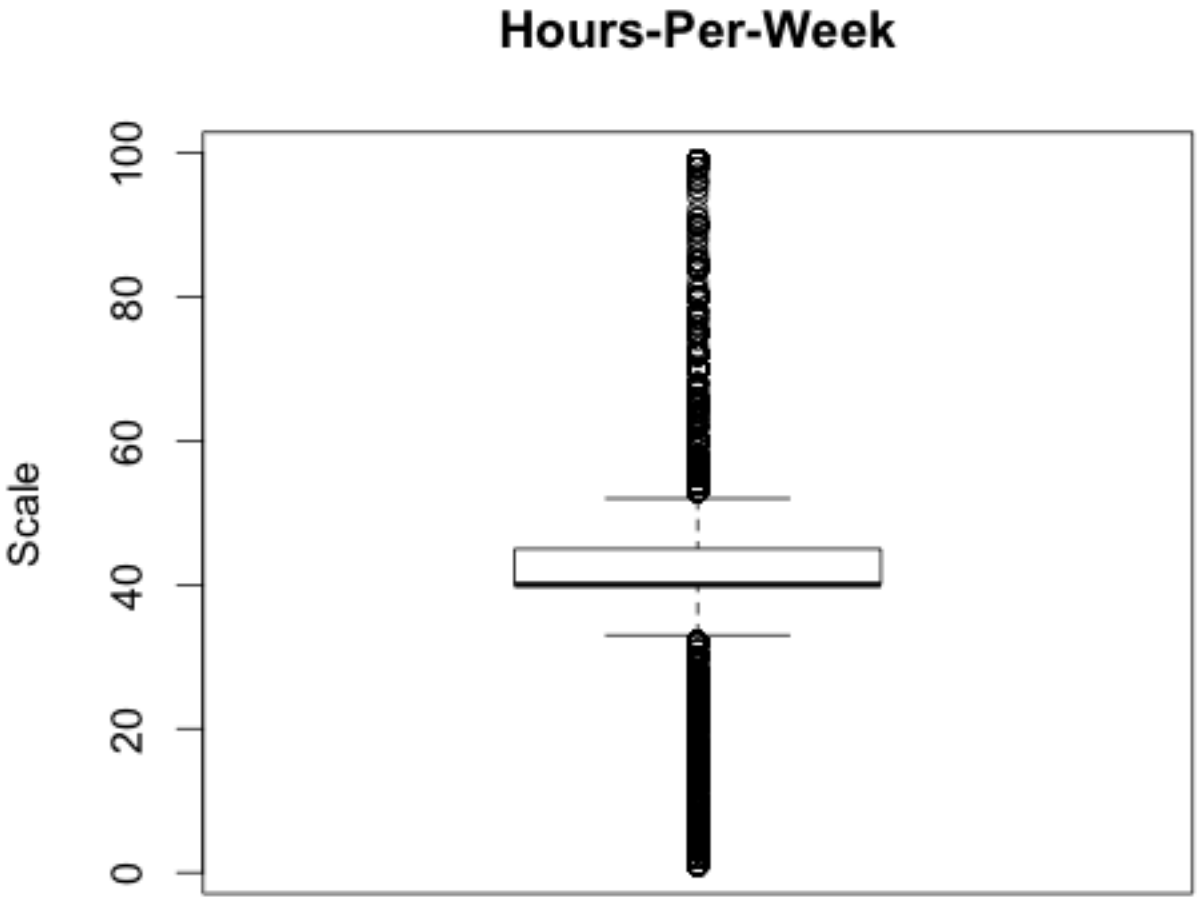

Figure 4.8:

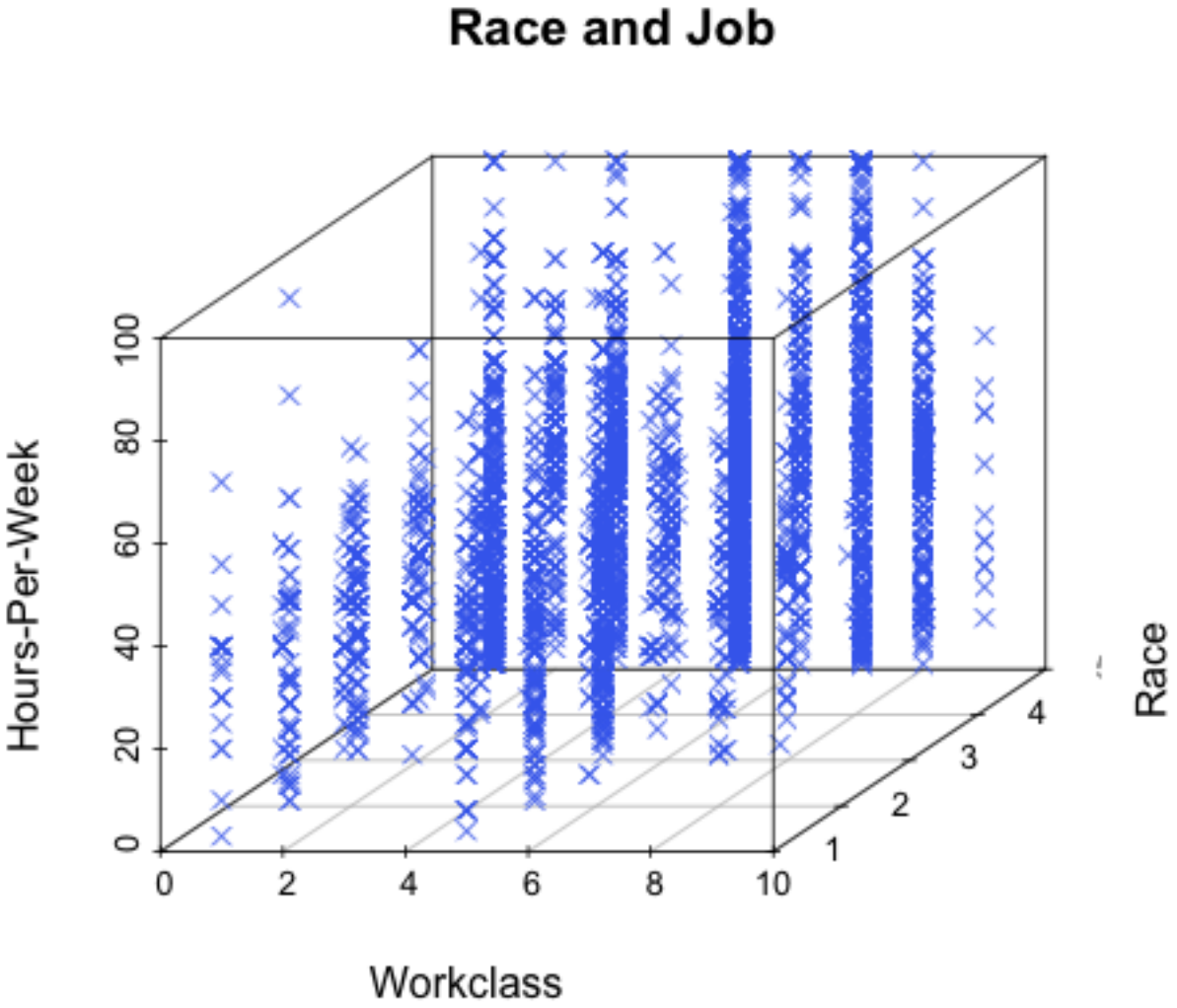
Race and Job
Hours-Per-Week
0
20
40
60
80
100
0
2
4
6
8
10
Workclass
1
2
3
4
Race

Figure 4.9:

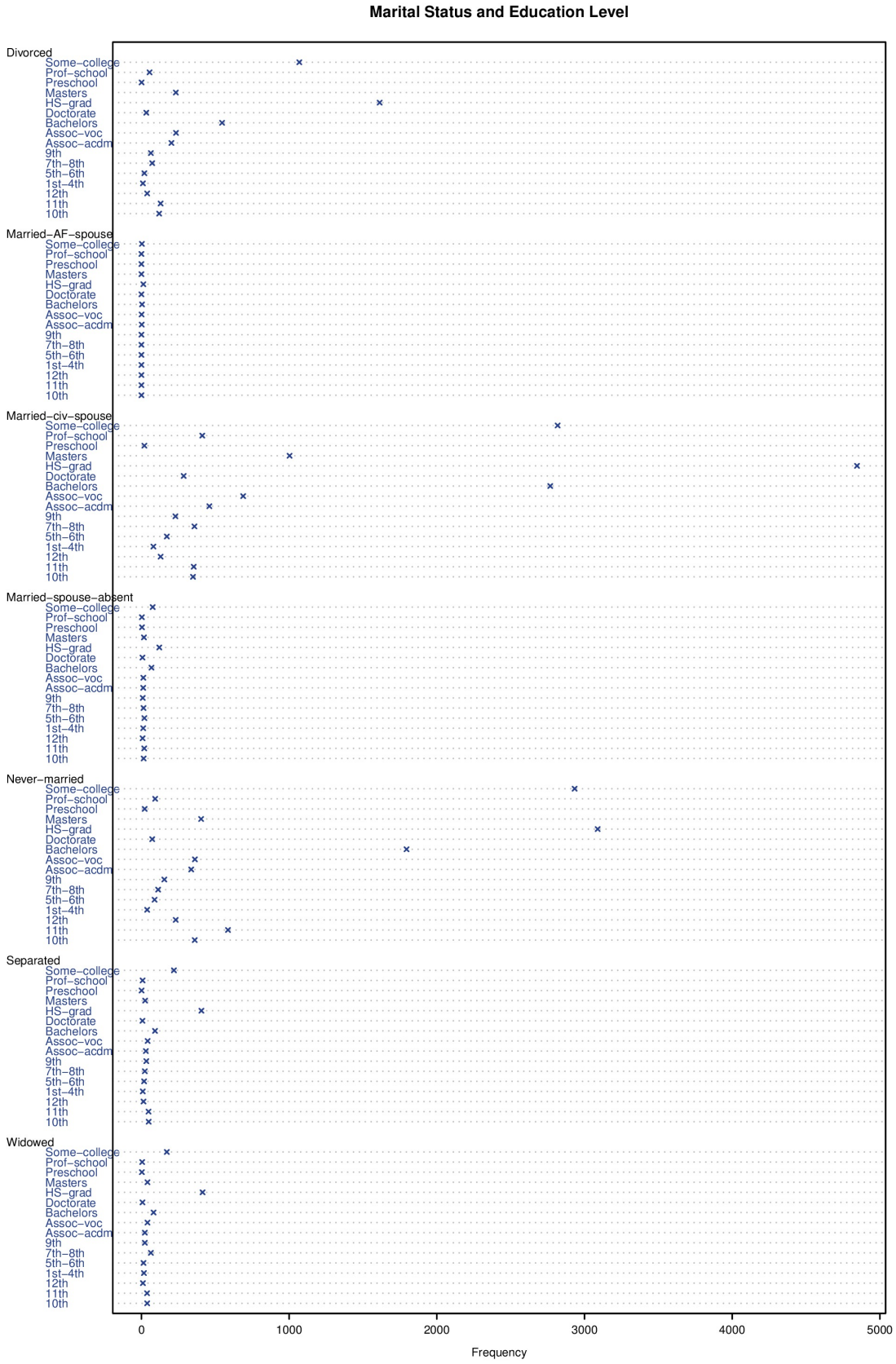
Marital Status and Education Level
Divorced
Married−AF−spouse
Married−civ−spouse
Married−spouse−absent
Never−married
Separated
Widowed
Some−college
Prof−school
Preschool
Masters
HS−grad
Doctorate
Bachelors
Assoc−voc
Assoc−acdm
9th
7th−8th
5th−6th
1st−4th
12th
11th
10th
0
1000
2000
3000
4000
5000
Frequency

Figure 4.10:

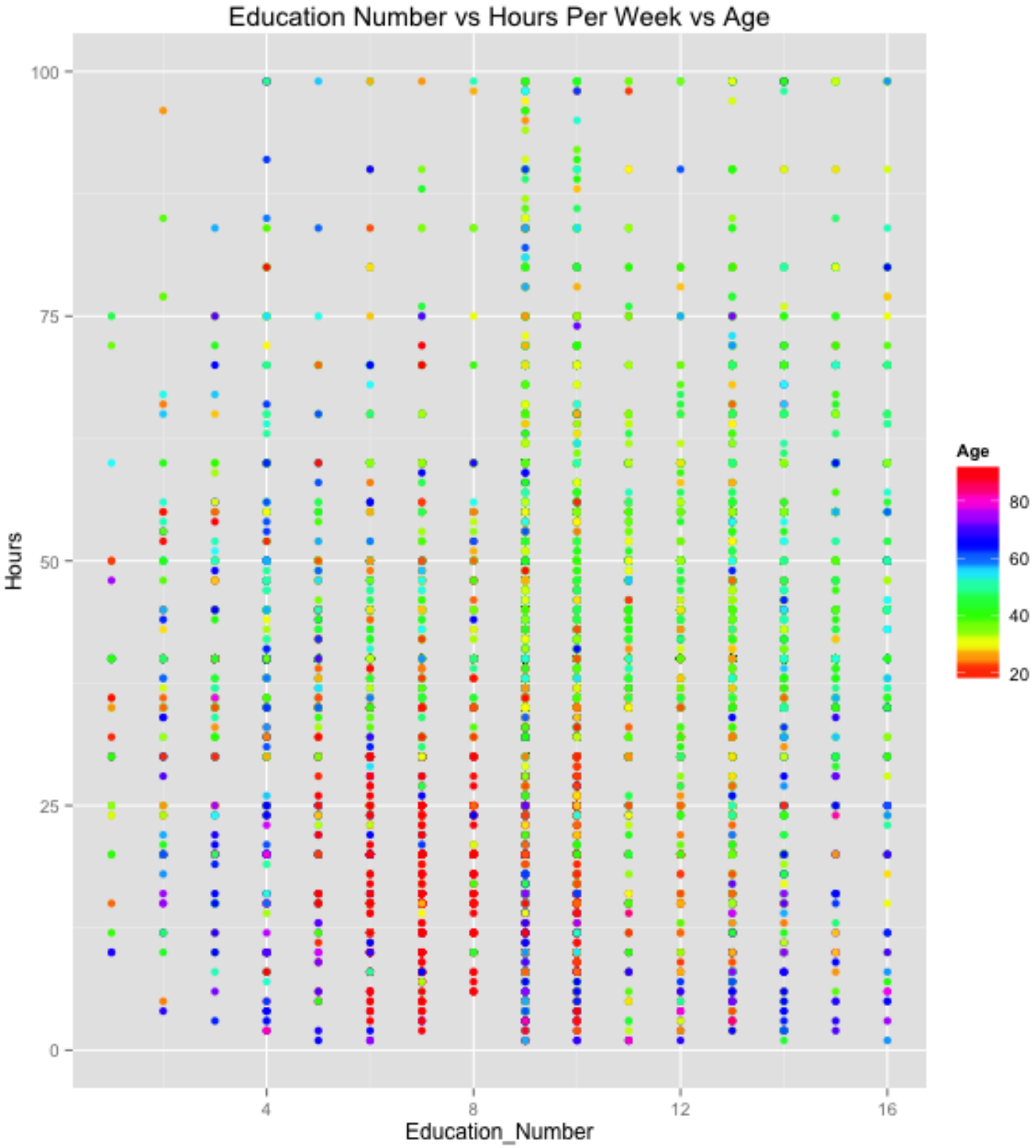
Education Number vs Hours Per Week vs Age
Hours
100
75
50
25
0
4
8
12
16
Education_Number
Age
80
60
40
20

Figure 4.11:

Table 1: Low Quality Red Wine Properties

	Mean	Standard Deviation
Fixed Acidity	8.14	1.57
Volatile Acidity	0.59	0.17
Citric Acid	0.24	0.18
Residual Sugar	2.54	1.39
Chlorides	0.09	0.05
Free Sulfur Dioxide	16.57	10.8
Total Sulfur Dioxide	54.65	36.72
Density	1.00	0.00
pH	3.31	0.15
Sulphates	0.62	0.17
Alcohol	9.93	0.75
Quality	4.90	0.33

Table 2: High Quality Red Wine Properties

	Mean	Standard Deviation
Fixed Acidity	8.47	1.86
Volatile Acidity	0.47	0.16
Citric Acid	0.29	0.19
Residual Sugar	2.53	1.42
Chlorides	0.03	0.03
Free Sulfur Dioxide	15.27	10.0
Total Sulfur Dioxide	39.35	27.25
Density	0.99	0.00
pH	3.31	0.15
Sulphates	0.69	0.15
Alcohol	10.85	1.10
Quality	6.27	0.49