DATA 511 Intro to Data Science
Course Project 1: Data Prep and EDA
Prof Larose

Name: Danielle Senechal

Use the *proj1* data set for all items in this course project.

The project instructions are shown in bold. This is to distinguish the instructions from your work. Your work should be in not-bold.

- Work neatly. Aim for a professional-looking presentation. Report writing is part of the DATA 511 course description, so I will be grading your level of professionalism, as well as your English expression.
- Make sure all graphs and tables fit neatly on the page.
- Neither add nor delete pages.
- At the beginning of the project, set.seed to 12345.

For all text output, surround it with a text box. In Word, select the text output, then Insert > Text Box > Draw Text Box.

Apart from this document, which you will save as a pdf and submit, you must submit your R script, containing the code you used to solve the problems. The R script should be neat and easily understandable by people who are not you. It should be well-annotated, describing what you are doing so that anyone could understand it.

This Project is brand new, and may have typos, errors, etc, that I have missed. Please report these to me asap. For this and other reasons, this Project is subject to change at any time (though of course I will be reasonable.)

Note to experienced R programmers: I am looking for straightforward methods to impute missing data, especially categorical data. Extra credit is available for demonstrating your hotshot methods. :-)

Good luck!

Prof Larose

1. **Insert your Executive Summary here. (A strategy for this is given at the end.)**

   This project examines a set of 14,797 customers from our database, to determine which customer characteristics are associated with high income, defined as having income greater than $50,000. Income levels were compared to several variables including sex, marital status, education level, and capital gains/losses. The following was observed:

   - There are significantly more males (9,885) than females (4,912) in this data set.
   - Marital status was combined into two categories, married and other. Those who are married are relatively evenly split, with 43.80% having incomes greater than $50,000. Of those in the "other" category (divorced, never married, separated, widowed), 6.35% have incomes higher than $50,000.
   - When observing the capital gains variable, there are 69 missing values, all of which are in records that have individuals with incomes greater than $50,000.
   - Instead of looking at the amount of capital gains/losses, individuals were grouped into two categories, those who have any amount of either capital gains or losses, and those who have no gains or losses. It was observed that for those who have no gains or losses, 57.32% have incomes greater than $50,000. For those that have made capital gains or losses, 19.09% have incomes greater than $50,000.
   - Education was split into three categories, those with less than or equal to 13 years of education, those who have between 13 and 14 years of education, and those who have more than 14 years of education. The distribution between these three groups shows that the majority of individuals in the data set have less than or equal to 13 years of education.
   - When education level is compared to income, it can be seen that of all individuals who have less than or equal to 13 years of education, 20.34% have incomes higher than $50,000. 56.35% of individuals that have between 13 and 14 years of education have incomes higher than $50,000. Of all those with more than 14 years of education, 75.64% have incomes higher than $50,000.

   From the information above, the characteristics associated with higher incomes include being married, having more than 14 years of education, and having no capital gains or losses.

   This project was meant for exploratory data analysis only. No data models/modeling were developed or performed. Following steps could include developing a model to allow for predictions of which individuals may fall into the high-income category.

2. **Missing Data. Look at a histogram of *capital.gain*. (Don't insert here.) The extreme data values in the right tail all have the exact same value: 99999. It is unlikely that all these individuals have the exact same amount of capital gains. Thus it is likely that the 99999 entry is code for *missing*.**

   a. **Set these 99999 values to missing using something like the following code:**
   ```
   proj1$capital.gain[proj1.imp$capital.gain == 99999] <- NA
   ```

   b. **Find the mean and standard deviation of *capital.gain*, after Step (a).**
   Fill in the following equations. $\mu_{orig} = 603.0878$ $\sigma_{orig} = 2610.7$
   Something like the following code might be helpful.
   ```
   cgm <- mean(proj1$capital.gain, na.rm = TRUE)
   cgsd <- sd(proj1$capital.gain, na.rm = TRUE)
   ```

   c. **Set your seed to some constant. Use *knnImpute* to impute the missing values for *capital.gain*. Make sure the output data set is not the same as the input data set. In other words, the second step uses code something like this:**
   ```
   proj1.imp <- predict(imputation_model, proj1)
   ```

   d. **The *knnImpute* method standardizes all the variables. But we haven't done EDA yet, so this is inconvenient at this early stage. So, de-standardize the imputed *capital.gain*, so that it is on the original scale, with no missing values. Name this variable *cg.imp*, and make sure it belongs to the same data set that was input to the imputation algorithm (*proj1*). Provide the five-number summary, along with the standard deviation of *cg.imp*. Compare it to the original values, pre-imputation, and comment.**
   Something like the following code might be helpful.
   ```
   proj1$cg.imp <- round(proj1.imp$capital.gain * cgsd + cgm, 5)
   ```

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard Deviation
$0.00	$0.00	$0.00	$605.90	$0.00	$41,310.00	$2,606.44

The above table displays the computed 5 number summary (along with the mean and standard deviation) for capital gains with the imputed values. When compared to the 5-number summary pre-imputation, each statistic is identical, apart from the mean and standard deviation. The mean and standard deviation of the imputed values are close to those of the pre-imputed values. This demonstrates successful imputation of new values to replace the missing data. If the mean and standard deviation had changed dramatically, it would imply that the imputed values were done incorrectly, and that the data has been distorted.

3. **Construct a flag variable, *cg.miss*, that takes value 1 when *capital.gain* (pre-imputation) is missing, and 0 otherwise. Hint: Use an *ifelse* command. Construct a contingency table, with *income* as the rows, and *cg.miss* as the columns, showing the counts. Discuss.**

	Not Missing	Missing
<=50K	11,243	0
>50K	3,485	69

The above table displays the occurrences of missing and non-missing capital gain variables when compared to incomes. The counts show that there were 69 missing variables in capital gain. Each of the missing values occur in records where the individuals had incomes over $50,000.

4. **Add a new field, *ID*. To show it is working, select <u>only</u> Record #2001 to display here.**

	Education	marital.status	sex	capital.gain	capital.loss	income	cg.imp	cg.miss	ID
2001	10	Never-Married	Male	0	0	<=50K	0	0	2001

5. **Consider *marital-status*.**

   a. **Construct a contingency table with *income* for the rows and *marital.status* for the columns, asking for the column percentages, rounded to two decimal places. Use Lucinda Console font size 7, so that the table fits.**

```
        Divorced Married-AF-spouse Married-civ-spouse Married-spouse-absent Never-married Separated Widowed
<=50K    89.67            45.45             55.15                92.46         95.61     94.49   89.40
>50K     10.33            54.55             44.85                 7.54          4.39      5.51   10.60
```

   b. **Rename *marital-status* as *marital-status-old.***

```
(names(proj1)[names(proj1)=="marital.status"] <- "marital.status.old")
```

   c. **Make a new variable, *marital.status,* where the three married categories are combined into the new category *Married*, and the other statuses are combined into the new category *Other*. Construct a contingency table with *income* for the rows and *marital.status* for the columns, with column percentages, rounded to two decimal places. Compare the proportions of high income for the two categories.**

	Other	Married
<=50K	93.65%	56.20%
>50K	6.35%	43.80%

The above table shows the income proportions between the two categories of marital status. For those who are married, 56.20% have incomes less than or equal to $50,000, while 43.80% have incomes greater than $50, 000. Of those who fall into the other category (divorced, never married, separated, widowed), 93.65% have incomes less than or equal to $50,000, while 6.35% make incomes higher than $50,000.

6. **Derive a new flag variable, called *capgl*. This flag variable should equal 1 whenever a customer has *either* any (imputed) capital gains *or* any capital losses. It should equal 0 otherwise. Construct a contingency table, with *income* as the rows, and *capgl* as the columns, showing the column percentages. Clearly describe the effect of having any capital gains or losses on *Income*.**

	Has gains or losses	Has no gains or losses
<=50K	80.91%	42.68%
>50K	19.09%	57.32%

This table displays the income percentages between individuals who have capital gains/losses and individuals who don't have gains/losses. For those who have no gains or losses, 42.68% have incomes less than or equal to $50,000, and 57.32% have incomes greater than $50,000. For those that have made capital gains or losses, 80.91% have incomes less than or equal to $50,000, while 19.09% have incomes greater than $50,000. These percentages demonstrate that if an individual has either capital gains or capital losses, they are more likely to be in the lower income category. If an individual has no capital gains or losses, there is a slightly larger chance (57.32%) they will be in the upper income category, though the incomes of those with no gains or losses is relatively split evenly.

7. **Outliers. Your professor is not a fan of deleting outliers at the EDA stage, because it often results in changing the character of the data set. Let's demonstrate this!**

   a. **What is the proportion of records with high income, over all records in the data set?**

      The proportion of records with high income amongst all records is 21.02%.

   b. **Consider *capital-loss*. The upper cutoff point for identifying outliers is the mean ($\bar{x}$) plus three times the standard deviation (s), or $\bar{x} + 3s$. Write it here by completing the equation begun in Equation Editor.**

      $$\bar{x} + 3s = 1307.547$$

   c. **Select only the records with values of *capital-loss* greater than this cut-off value. How many are there?**

      There are 679 values greater than the cut off value of $1,307.55. This means there are 679 outliers in the upper cutoff.

   d. **What is the proportion of records with high income, among these outlier records?**

      The proportion of records with high income amongst the outlier records is 51.99%.

   e. **Describe the change to the character of the data set that will result if we delete these outlier records. State your conclusion regarding deleting outliers at the EDA stage.**

      If the outliers are deleted from the data set, the information in the data set will become distorted because the extreme values have been removed. When looking at higher income across all records, 24.02% of all individuals have incomes higher than $50,000. When looking at just the outlier records, 51.99% of these individuals have incomes higher than $50,000. Simplified, this means that of the 24.02% of individuals with higher incomes, 51.99% of those individuals would be considered outliers. If the 51.99% were to be removed, the overall percent of those with higher incomes would decrease dramatically, because 51.99% of the data on those with high incomes would have been removed. This could lead to incorrect conclusions being made about the data.

8. **We would like to bin *education* based on predictive value.**
   a. **Use the method and options shown in the notes to generate the decision tree for predicting *income* based on *education*. No need to copy it here.**

   b. **Use the *cut* function, along with the split thresholds from your tree in part (a), to construct a new variable *educ.bin*. Provide a contingency table here of the counts, with *income* for the rows and *educ.bin* for the columns.**

	(-21,13]	(13,14]	(14,21]
<=50K	10,784	344	115
>50K	2,753	444	357

   c. **Redo the contingency table from (b), this time with the column proportions, rounded to two decimal places.**

	(-21,13]	(13, 14]	(14, 21]
<=50K	79.66%	43.65%	24.36%
>50K	20.34%	56.35%	75.64%

   d. **Discuss your results from (b) and (c).**

   The table in (b) displays counts of the three categories of education relative to income. The category that has the most individuals, 10,784, is for those who have less than or equal to 13 years of education and have incomes less than $50,000. The table in (c) shows that of all individuals who have less than or equal to 13 years of education, 79.66% have incomes less than or equal to $50,000, and 20.34% (2,753 individuals) have incomes higher than $50,000. For those who have more than 13 but less than 14 years of education, 43.65% (344 individuals) have incomes less than or equal to $50,000 and 56.35% (444 individuals) have incomes higher than $50,000. Lastly, for individuals with more than 13 years of education, 24.36% (115 individuals) have incomes less than or equal to $50,000, and 75.64% (357 individuals) have incomes greater than $50,000. These two tables show that there is a greater chance of an individual having a higher income if they have had more than 14 years of education, and there is a lower chance of them having a higher income if they have had less than or equal to 13 years of education. For those with more than 13 years but less than 14, they are split almost evenly between lower incomes and higher incomes.

Exploratory Data Analysis

9. **Using *ggplot2,* construct the following stacked bar graphs of *educ.bin* with overlay of *income*. (The lingo here means that *educ.bin* will be on the horizontal axis, and *income* will represent the colors.) Insert them so that they both fit in the table I provided below.**
   a. **Non-normalized stacked bar graph.**
   b. **Normalized stacked bar graph.**
   c. **In one sentence, describe the distribution of *educ.bin* regardless of *income*.**

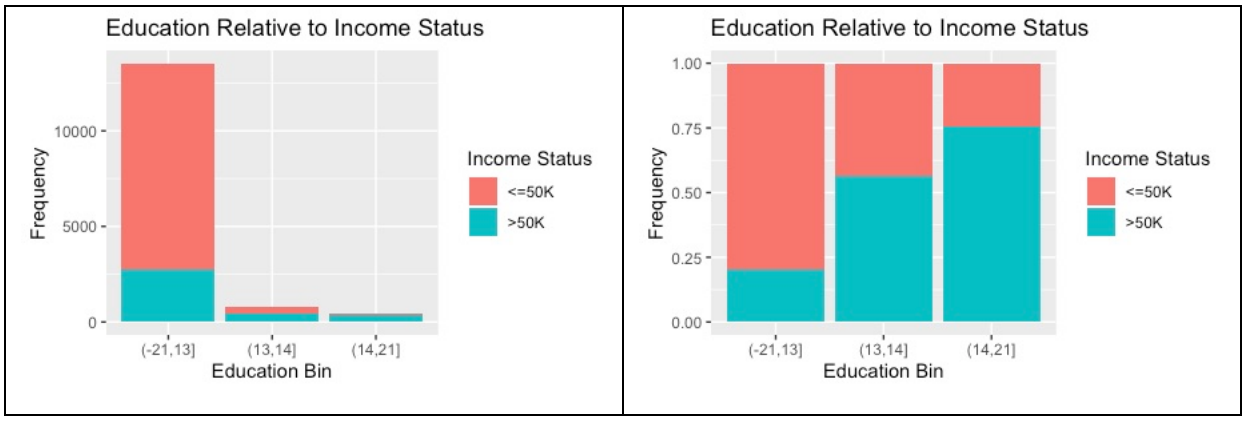
      The distribution between education bins shows that the majority of individuals in this data set have had less than or equal to 13 years of education, and significantly less individuals fall into the other two bins.

   d. **In one sentence, describe the distribution of *educ.bin*, with respect to *income*.**

      The distribution between education bin and income displays that individuals are more likely to have a higher income if they have more than 14 years of education and are less likely to have a higher income if they have less than or equal to 13 years of education.

   e. **Briefly state the benefit of using the non-normalized version and of using the normalized version.**

      The benefit of using the non-normalized version of the graph allows for the counts of individuals in each category to be easily seen. In this particular example, it is easy to see that the first bin contains the most individuals. The normalized version of the graph allows for the relation between the two variables to be easily seen as compared to the non-normalized version. In the normalized graph below, the distribution between education bin and income is easily seen.



10. Provide the following contingency tables of *sex* (columns) and *income* (rows).

a. Table of counts, with row and column totals.

	Female	Male	Sum
<=50K	4,348	6,895	11,243
>50K	546	2,990	3,554
Sum	4,912	9,885	14,797

b. Table of column percentages, rounded to two decimal places, with totals provided for the columns only (not the rows).

	Female	Male
<=50K	88.52%	69.75%
>50K	11.48%	30.25%
Sum	100.00%	100.00%

c. In one sentence, how is the table in (a) preferable to the table in (b)?

The table in (a) is preferable to the table in (b) because it gives actual counts, and if the exact number of how many individuals fall into one category is needed, it is easily found in this table.

d. In one sentence, how is the table in (b) preferable to the table in (a)?

The table in (b) is preferable to the table in (a) because it gives the percentage of individuals in each category, which makes the categories easier to compare to each other.

e. In one sentence, describe your results from (a).

It can be seen in the table from (a) that there are significantly more males (9,885) than females (4,912) in this data set.

f. In one sentence, describe the effect of *sex* on *income* from (b).

The table from (b) shows that females are less likely to have higher incomes than males, with 11.48%% (546) of females having incomes greater than $50,000 and 30.25% (3,554) of males having incomes greater than $50,000.

Craft your Executive Summary as follows.

Your boss makes more money than you. He or she has little time for the arcane details of all the data prep and other work you did to produce your report. Your boss is only interested in RESULTS.

A good executive summary should consist of the following.

1. **A quick summary of the *Objective* of the analysis. For this project, this would be something like the following: "This project examines a set of [this many] customers from our database, to determine which customer characteristics are associated with high income, defined as having income greater than $50,000." Feel free to copy and paste this sentence in your executive summary. Also include the original proportion of high-income customers.**

2. **Bullet points with explanations of your most salient results. I think you can make good bullet points with your results from the following problem numbers:**
    a. **Problem 3**
    b. **Problem 5c**
    c. **Problem 6**
    d. **Problem 8d**
    e. **Problem 9a and 9b**
    f. **Problem 10e and 10f**

3. **What NOT to include in your Executive Summary is anything about data prep, unless it affects managerial policy. I think it is safe to assume nothing in this project affects managerial policy. Problem 3 is EDA drawn from data prep, so deserves a mention.**

4. **Brief mentioning of next steps. This helps to delimit the scope of the project. For this project, you may say something like, "Note that this project is exploratory only. No actual data modeling has been performed. Rather, our next step should be to perform data modeling to predict which customers will be high-income."**

5. **And, whatever you do, do not exceed one page!** 😉

==========

Well done!

<u>Deliverables:</u>

1. **Save your completed Word document as a pdf file, named *Doe_Jane_Project1* (if your name is Jane Doe, with last name first!). Because of virus issues, no Word documents will be accepted.**
2. **Your well-annotated R script, named *Doe_Jane_Project1_RScript.***

Do NOT zip these two files together. Rather, make two separate submissions using the Project Submissions Tool.