**DATA 511 Intro to Data Science**

**Course Project 1: Data Prep and EDA**

**Prof Larose**

**Name:**

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.
* There is no need to post your R code in this document. That is what the R script is for.
* 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.)

**Good luck!**

**Prof Larose**

1. **Insert your Executive Summary here. (A strategy for this is given at the end.)**
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*.**
   1. **Set these 99999 values to missing using something like the following code:**

proj1$capital.gain[proj1.imp$capital.gain == 99999] <- NA

* 1. **Find the mean and standard deviation of *capital.gain*, after Step (a). Fill in the following equations.   
     Something like the following code might be helpful.**

cgm <- mean(proj1$capital.gain, na.rm = TRUE)

cgsd <- sd(proj1$capital.gain, na.rm = TRUE)

* 1. **Set your seed to 12345 (set.seed(12345)). 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)

* 1. **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)

1. **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.**
2. **Add a new field, *ID*. To show it is working, select only Record #2001 to display here.**
3. **Consider *marital-status*.** 
   1. **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.**
   2. **Rename *marital-status* as *marital-status-old.***

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

* 1. **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.**

1. **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*.**
2. **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!** 
   1. **What is the proportion of records with high income, over all records in the data set?**
   2. **Consider *capital.loss*. The upper cutoff point for identifying outliers is the mean plus three times the standard deviation (*s*), or . Write it here by completing the equation begun in Equation Editor.**
   3. **Select only the records with values of *capital.loss* greater than this cut-off value. How many are there?**
   4. **What is the proportion of records with high income, among these outlier records?**
   5. **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.**
3. **We would like to bin *education* based on predictive value.** 
   1. **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.**
   2. **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.**

* 1. **Redo the contingency table from (b), this time with the column proportions, rounded to two decimal places.**
  2. **Discuss your results from (b) and (c).**

**Exploratory Data Analysis**

1. **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.** 
   1. **Non-normalized stacked bar graph.**
   2. **Normalized stacked bar graph.**
   3. **In one sentence, describe the distribution of *educ.bin* regardless of *income*.**
   4. **In one sentence, describe the distribution of *educ.bin*, with respect to *income*.**
   5. **Briefly state the benefit of using the non-normalized version and of using the normalized version.**

|  |  |
| --- | --- |
|  |  |

1. **Provide the following contingency tables of *sex* (columns) and *income* (rows).**
   1. **Table of counts, with row and column totals.**
   2. **Table of column percentages, rounded to two decimal places, with totals provided for the columns only (not the rows).**
   3. **In one sentence, how is the table in (a) preferable to the table in (b)?**
   4. **In one sentence, how is the table in (b) preferable to the table in (a)?**
   5. **In one sentence, describe your results from (a).**
   6. **In one sentence, describe the effect of *sex* on *income* from (b).**

**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:**
   1. **Problem 3**
   2. **Problem 5c**
   3. **Problem 6**
   4. **Problem 8d**
   5. **Problem 9a and 9b**
   6. **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!**

**Deliverables:**

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.**