**Project 4**

*Quantitative Methods for Business Analytics and Supply Chain – Fall 2018*

Due Thursday, December 13th at noon.

For this project, you will work on a realistic “found data” setting – namely, using data from Lending Club and from the IRS’s Statement of Income database. Relative to the other projects, this one is more open ended (a feature meant to resemble more realistic data analysis environments).

***Deliverables***.

1. Write a **self-contained and insightful typed report** on the data, statistical methodology applied, and findings interpreted in the context of the data.
2. As an appendix to the report, include the relevant R code and output (script files, log files) you used to perform the calculations in your report.
3. In the main text of your report, you should have typed tables and figures (not just copied from R output, but formatted nicely) that help you summarize nicely the relevant calculations and graphics you discuss in the main text of the report.

Note: Do not structure your report as a point-by-point answering of the questions that I pose below as guidance. The report should contain an introduction that states the broad objectives of the project, a data section that describes the data sources and provides useful summary information, a methods discussion to describe any data merging and processing as well as specifications employed, a results section to describe the main findings, and a brief conclusions section to summarize the lessons learned from the exercise.

There is no page limit, but in my experience, reports that are five pages long do not provide enough detail and 30-page reports provide too much detail (lacking structure).

***Project Guidance***.

*Part A. What is in a credit score?*

1. Download the Lending Club data ([https://www.lendingclub.com/info/download-data.action)](https://www.lendingclub.com/info/download-data.action) for the year 2014, and read the data into a data frame in R.
2. Process the data file to be more useful. In particular, the data contain many variables that we will not examine in this project.
   1. Restrict the data frame to the following variables: grade, sub\_grade, zip\_code, term, loan\_amnt, annual\_inc, verification\_status, purpose, tax\_liens, pct\_tl\_nvr\_dlq, int\_rate, and loan\_status. Download the data dictionary from Lending Club to learn about what each of these variables means.
   2. Inspect the characteristics of each of the variables in (a). This is a necessary step, but one that won’t enter much into the final report.
      1. How are each of the variables stored within R? Is this the format that is most useful? For any variables stored in the “wrong” format, devise a strategy to convert these to a more useful format.
      2. For continuous variables, are there any outliers or skew? Think about the likely consequences of these characteristics and any practical solutions. You should address these issues in how you aggregate and analyze the data, and describe any relevant issues in your report.
3. The Lending Club data are at the loan level. Produce two alternative “aggregated” data sets, which we’ll examine at different points throughout the analysis.
   * 1. **ZIP x Verification Status**. A data set in which each observation is an aggregation of all of the loans with the same combination of (zip\_code, verification\_status).
     2. **ZIP x Verification Status x Term x Subgrade.** A data set in which each observation is an aggregation of all of the loans with the same combination of (zip\_code, verification\_status, term, sub\_grade).

*Some guidance on how to aggregate within each group*. For the continuous variables, compute the average within each group. For loan\_status, compute two fractions: the fraction of loans that are current, and the fraction of loans that are fully paid. For home\_ownership, compute the fraction of loans with a homeowner. For purpose, compute fraction of debt consolidations.

1. *Numerical and Graphical Summaries*. For the **ZIP-Verification Status-Term-Subgrade**-level data set, summarize the key variables graphically and numerically. For this part, you should focus on what is the most informative way to describe the distributions of the potential outcome variables (int\_rate, loan\_status, and loan\_amnt), and their relation to annual income (and its verification), subgrade of the loan, the fraction of debt consolidations, previous delinquencies, and tax liens.
2. *Questions of Interest.* Use the **ZIP-Verification Status-Term-Subgrade**-level data to answer the following questions of interest:
   1. Does the term of the loan predict the status of the loan? Think about what this comparison tells you about the structure of the data.
   2. Does annual income matter for the loan amount and/or its interest rate? Is the relationship any different if the income is verified, source verified, or not verified?
   3. Is the subgrade assigned to the loan useful for predicting the likelihood of defaulting on the loan (“Charged Off” means the borrower defaulted)? As you think about this question, think about whether examining 36-month term versus 60-month term would tell you different information.
   4. Conditional on the loan’s subgrade, are borrower credit characteristics (i.e., tax liens, previous delinquencies) important for whether the borrower will default on the loan? As you think about this question, think about whether examining 36-month term versus 60-month term would tell you different information.
   5. Conditional on the loan’s subgrade, are borrower credit characteristics (i.e., tax liens, previous delinquencies) important for the interest rate or the loan amount?
   6. Does the purpose of the loan or the borrower’s homeownership status matter for the likelihood of defaulting on the loan?

*Part B. Trust but Verify?* The goal of this part is to relate the income measure from Lending Club (which is sometimes verified, but sometimes not) to the income measure from the IRS (which is verified via employer reporting and withholding of taxes at the time of payment).

1. For this part, you should read up on how Lending Club verifies income (<https://www.lendingclub.com/public/income-verification.action)>. As you read this, ask yourself what features of the verification process could be relevant for comparing verified individuals to individuals that Lending Club chooses not to verify.
2. Retain the 2014 Lending Club data you worked with in Part A. In addition, read into your R environment the file “14zpallnoagi.csv,” which you can download from the IRS tax statistics website (<https://www.irs.gov/uac/soi-tax-stats-individual-income-tax-statistics-2014-zip-code-data-soi> -- please download “all States, does not include AGI”), which provides tons of tax information from tax filings to the Internal Revenue Service.
   1. The variable names are not so informative, but we will concern ourselves with only three of them: ZIPCODE, N02650, and A02650. ZIPCODE is the 5-digit ZIP code (unit of observation for the data set), N02650 is the number of tax returns from the ZIP code that report a total income, and A02650 is total amount of total income reported by taxpayers in that ZIP.
   2. Discard the rest of the variables (aside from the 3 that interest us), and use this information to compute a variable representing the average total income of taxpayers in the ZIP code.
   3. Summarize the average total income of taxpayers by ZIP code, both graphically and numerically. As you inspect the variable, think about whether outliers or skew are important for this distribution. Your subsequent analysis should pay mind to this.
3. Merge the **ZIP x Verification Status** data from Lending Club from Part A with the IRS ZIP code data.
   1. Note: Lending Club is available at ZIP3, but IRS produces ZIP5. Before merging, think about the most appropriate transformation, aggregation, etc. to merge the two data sets [Your report should describe how you do this]. In particular, you should think about how the choice of what to do in this part interacts with your analysis in 1c.
4. Using the merged data, produce informative summary statistics and well-chosen plots that examine the relationship between Lending Club income (annual\_inc) to IRS income (average total income), and whether this relationship is stronger or weaker for Lending Club borrowers whose income was verified.
5. *Questions of interest* – to be answered with multiple regression analysis and useful graphical evidence.
   1. Is there evidence that the incomes of Lending Club (LC) borrowers and IRS taxpayers are drawn from distributions with the same average income? Do LC borrower income and IRS borrower income have a similar spread? How are these conclusions affected by the choices you made when you aggregated the data and merged?
   2. Does verification strengthen the relationship between Lending Club income and verified IRS reported income?
   3. Construct a binned IRS income variable to help with this one. Does IRS income exhibit a stronger relationship with LC income for different parts of the distribution of income (e.g., the top quartile versus the bottom quartile)?

Other ways to distinguish your project. This part is a little less structured.

* The IRS data are actually available at the ZIP x Adjusted Gross Income Category level (at the link above). You could download those data, and think about how to produce an informative merge to answer question 4c. How would that analysis be different and/or more informative?
* Lending Club claims that they verify income that could come from more uncertain sources. You could dig into the Lending Club data dictionary to learn more about this question using some of the personal characteristics of the borrowers (particularly their employment information).
* Digging further, you could examine whether the employment information provided on Lending Club is more or less useful than Income Verification status to explain the strength of the relationship between IRS income and LC income.
* Finally, you could provide evidence about what characteristics lead Lending Club to verify income, and think about how these results influence how you interpret the correlations from the main part of the analysis.

I believe there is plenty to think about with this project, and it ought to give you a taste for the kind of analysis you would do in a real business context. I hope you enjoy doing the project as much as I enjoyed putting it together. Let me know as you encounter questions with the project.