**Lending Club Project**

Your mission is to build various models using the Lending Club data. For example, you would build a model to understand what factors determine the interest rate that borrowers pay. Lending Club is a platform that matches borrowers to lenders. Lending Club determines an interest rate for a loan based on a number of characteristics of the borrower.

The data will be posted here:

Here are some guidelines to help:

1. You are going to need to manipulate your data quite a bit. Be sure to create a copy of the data before you start manipulating. While manipulating data, often you would make a change that either has unintended consequences or has consequences that may prevent you from doing other analyses. You can always start again with the original dataset.
2. There are over 100 columns of variables. In the sheet loan\_stats, you would see all the columns that were originally in the dataset. This sheet also serves as a data dictionary for your column names. Many of these columns either had very little data or are not helpful to your analysis. This is the nature of a real datasets. I have extensively cleaned this dataset, but can use a little more cleaning, if you would like. You may consider deleting some more columns to make your data more manageable, again only if you would like—this is not a requirement.
3. Pay attention to your dependent variables. For the regression model, the dependent variable will be int\_rate. You want to make sure you carefully look at the descriptives of int\_rate prior to developing your regression model.
4. It is up to you to determine what factors affect your dependent variable in any given model. For example, you may want to start with the amount of the loan, credit grade, and other credit factors that your theoretically think will be important in determining interest rate.
5. When creating dummy variables, remember that SAS or SPSS does not know what is your bare group. For example, home ownership status might be important variable. You will need to create dummy variables for the different categories of home ownership. Remember that you need to leave one category out as that would become your base group.
6. To build valid models, your independent variables should not be highly correlated with one another. Be sure to run a correlation matrix on your independent variables to check for this.
7. When submitting, please include the code you ran (even if the SAS or SPSS generated this code), the output, and (most important) a bullet-point summary of what are your conclusions from your analyses.
8. Remember that a simple model is often the most useful.
9. **There is no one right answer.**

**Part 1: Comparing Populations**

Fifth third bank hired Louise Card as their new loan officer and charged her with increasing the loan portfolio of that bank. To help Louise with the task, bank provided her with data on all past customers, as well as an expense account for marketing costs. Louise quickly discovers that the expense account is not enough to market to all the customers. Louise decides she would target only target home owners (those customers who either own a home or have a mortgage on their home) instead of renters because she believes that home owners do not have much spare cash for unexpected expenses and are more likely to take a loan. Your job, as a friend of Louise, is to either confirm or disconfirm her belief. Here are some guidelines for this task:

1. You could take any data you would like for this task. For example, you might find inq\_last\_6mths helpful. You may run multiple models and see if they provide consistent information or not. You are only required to run a minimum of one analysis. **Additional analyses are** **optional**.
2. Very small subset of customers either did not provide their home-ownership data or they had multiple homes. They are coded as Any. You would want to remove those customers for your analysis.
3. You may want to move the data you want to use into a separate sheet for analysis.
4. You may also want to reorganize data such that home owners are in one column and renters are in other.
5. In your report:
   1. Explain which variables you chose for analyses and why.
   2. Write your hypotheses in words.
   3. Provide the output for each analysis.
   4. Summarize your findings. Clearly outline what Louis should do for her marketing campaign.

**Part 2: Manipulating Data**

For this exercise, we will practice coding the dataset to make it more usable for future modeling assignments. Remember to keep the original columns as it is. Whenever you manipulate data, just create new columns.

1. There are two possible loan terms. Code the loan term of 36 months as 0 and loan term of 60 months as 1.
2. The loan grade has multiple values A through F. Create six new columns. Each column would be named with the grade—i.e., Grade A as first column, Grade B as second and so on. Every time the original column grade is A, values in column Grade A are 1, otherwise it is 0. Same for the remaining columns.
3. What is the mean and standard deviation of the interest rate charged to the customers?
4. Run the correlations between variables you think might be important to predict the interest rate. Note down which variables are highly correlated with interest rate. Also note which of these variables that you think would predict interest rate are correlated with one another. Fun fact: if you are running a simple regression model, the square of the correlation is the regression coefficient.
5. In your report, submit briefly your correlation matrix from #4, and summarize your findings.

**Part 3: Regression Analysis**

Louise Card left her job at the fifth third bank to become the director of alumni relations with College of Business, her alma mater. She is looking for a $20,000 loan to build her dream outdoor kitchen so that she can entertain alums at her house on football game days. She would like to pay the loan off over 36 months. Her income is verified at $150,000. Her credit is very good (A) and credit history goes back to 2003. Having graduated not too long ago, she does have an outstanding mortgage on it. As of now, she has open credit cards with a revolving balance of $4000. You now took the job of the fifth third bank loan officer.

1. Develop a regression model to predict interest rate for Louise Card. To see how your model works, please estimate the interest rate of a loan Louise would like for her outdoor kitchen.
2. It may not be possible to incorporate all of the factors provided by Louise into the regression model. You just need to choose based on your regression model.
3. You may not have all the information you need for your regression model. What are you going to in that instance?
4. In addition to estimating the interest rate, you might want to address the following questions:
   1. If Louise decided to take out a $40,000 loan instead of $20,000, how would that effect the interest rate?
   2. If Louise instead had a B credit rating, how much more would she pay?
   3. What is the most important determinant of interest rate?
5. In your report:
   1. Explain which variables you chose for analyses and why.
   2. Provide the output for each analysis.
   3. Summarize your findings. Clearly outline what Louis should do for her outdoor kitchen.

Louise really wants to show her Cardinal spirit. She would like to spend as much money on the outdoor kitchen as she can get a loan for. She is not sure how much loan can she get approved and is looking at you for help.

1. Develop another regression model to predict the loan amount (variable loan\_amnt). You may notice that the data does not suggest how much loan a customer is eligible for. In other words, loan amount refers to the particular dollar amount loan customers applied for. You may, however, use this as the maximum amount because if the loan requested was higher than the amount customer was eligible for, the loan amount was reduced by Lending Club (from data disctionary: loan\_amnt~The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value).
2. Some of the predictors for this model will be continuous variables and others would be categorical. Include grade as one of the variables.
3. Treating Grade B as the baseline group, what is the difference between the loan amount for Grade A, C, D, E and F.
4. While conducting #3, would you include other variables in the model. Why? Why not?
5. Louise is wondering if her loan amount would be higher if she pays more money in each installment at a higher interest rate. It makes theoretical sense to you, but you have to provide proof that this is actually true given the historical data.
6. In your report:
   1. Explain which variables you chose for analyses and why for each model.
   2. Provide the output for each analysis.
   3. Summarize your findings. Clearly outline the impact of your findings for Louis.