**Data Analysis – LendingClubData**

**Solutions to Q1-Q9 are enclosed on .xlsx titled “LendingClubData Analyzed with Solutions”**

**PART I**

LendingClub is a US peer-to-peer lending company, headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. LendingClub operates an online lending platform that enables borrowers to obtain a loan, and investors to purchase notes backed by payments made on loans. LendingClub is the world's largest peer-to-peer lending platform. The company claims that $15.98 billion in loans had been originated through its platform up to December 31, 2015.

*From Wikipedia: https://en.wikipedia.org/wiki/Lending\_Club*

The dataset provided is a random sample of 2015 loan data from LendingClub. The file contains complete loan data for loans issued through the time period stated, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information.

*Data source: https://www.lendingclub.com/info/download-data.action*

**Please put final answers in this Word document, and Save As "Set Final Exam – YourName.docx"**

Answer the questions using the following files:

LendingClubDataDictionary.xlsx

LendingClubData.xlsx

Save the data as: "LendingClubData – YourName.xlsx"

**Follow the instructions of each question, showing your work in the Excel spreadsheet "LendingClubData – YourName.xlsx"**

**Question 0 (0 points)**

Inspect the two Excel files. The LendingClub Data is a random sample of 10000 loans from 2015.

**Question 1**

Calculate and report the following descriptive statistics for the loan amounts in this sample, rounding each value to the nearest cent:

* Mean:
* Median:
* Sample standard deviation:
* Minimum:
* Maximum:

Show your work in a new sheet of the data file, named **Q1\_Stats**.

**Question 2**

Create a frequency histogram of loan amount in Excel using between 8 and 12 bins. Neatly label the histogram.

Save the frequency table and resulting chart to a sheet named **Q2\_Histogram**.

**Question 3**

Compute the average loan amount by purpose using a PivotTable. Round the average loan amounts to the nearest cent. Add two more columns to this PivotTable, (1) count of loans for each type of purpose; (2) average interest rate of loans for each type of purpose. Round average interest rate to four decimal places.

Save the table to a sheet named **Q3Q4\_Pivot\_Chart**.

**Question 4**

Create a bar or column chart to display the number of loans according to the **Purpose** category. Clearly label all axes.

Save the resulting chart to the sheet named **Q3Q4\_Pivot\_Chart**.

**Question 5**

Create a new column in the dataset named **Income\_Annual**. For loans with **application\_type** = **Individual**, this new column should contain the self-reported annual income provided by the borrower during registration. For loans with **application\_type** = **Joint App**, this new column should contain the combined self-reported annual income provided by the co-borrowers during registration for loans.

Create this column using **IF()**. This column should be in the **loan10000** sheet.

**Question 6**

Create a scatter plot using the new column, **Income\_Annual**, and the user's total high credit/credit limit. **Income\_Annual** should be on the x-axis. Clean up and format the chart, labeling all axes.

Save the resulting chart to the sheet named **Q6\_Scatter\_Plot**.

**Question 7**

Determine the cutoff for outliers for **Income\_Annual**, and remove these records (rows) from the dataset. Use the 1.5\*IQR definition.

Save the new data in a sheet called **Q7\_NoOutliers** – including all columns, removing only the rows with outlier values for **Income\_Annual**.

**Question 8**

Perform linear regression using the data from **Q7\_NoOutliers**. Use the **Income\_Annual** column and the column that stores the user's total high credit/credit limit. Use **Income\_Annual** as the independent variable. Be sure to output the scatter plot with the linear regression model on top of it. Format the plot to show the model as a line.

Report the following values below:

* Slope: 2.57924567430189
* y-intercept: -16766.6856692963
* R-squared: 0.291739730230152
* Correlation coefficient: 0.540129364347238

Briefly discuss the linear relationship here: The linear relationship or association is either not strong or is moderate because the Mult-R coefficient is only .5

Save your regression output and the plot to a sheet named **Q8\_Regression**.

**Question 9**

We want to figure out if people who own their homes (OWN for **homeownership**) tend to have lower values for loan amount than the population. Assume the mean loan amount for the population is $15500. The population standard deviation is unknown. Use all of the relevant records from the original **loan10000** dataset as your sample to perform the appropriate hypothesis test.

Report the following values:

* Statements of the two hypotheses:

Null Hypothesis: >=, =

Alternative: mu\_own < Mu\_population

* Name of test performed:

One same t test

* Test statistic: -2281733629
* p-value: 0.011348154

Briefly discuss the outcome of your test here: My test was significant, the loan amounts for home ownership was lower than the mean.

Show your work in Excel, and save to a sheet named **Q9\_Test**.