DATA DRIVEN\_PROJECT TEAM\_1

WHITE PAPER

Introduction:

Team\_1 decided to use and analyze Lending Club data with the goal to identify trends within the data set. Lending Club is a lending organization that specializes in debt-consolidation loans and credit cards. They cater to all of the U.S. Population and provides their services primarily online.

Business Understanding:

It was highly important to not only understand the data, but to understand the business model. Lending Club offers Fixed rates and allows customers to borrow a minimum of $1K to $40K in loans. Additionally, lending Club rates tend to be a bit lower, because the decisions to get loans are made online creating lower overhead.

Data Understanding

The core data used by Team\_1 was the Lending Club Personal Loan Origination data for Q1 - Q4 2018. The data was sourced from Lending Club Data website and downloaded by Quarters. The team looked at the data set in its entirety with the help of a data dictionary to decide on where each would focus on and split the assignments.

* [Lending Club Data Download](https://www.lendingclub.com/info/download-data.action) -- data source link

The additional data used to complete the project had to also come from external sites to determine trends with disaster and stock data in 2018.

* [FEMA Disaster Data Download](https://www.fema.gov/openfema-dataset-disaster-declarations-summaries-v1) -- data source link
* [FRED Delinquency Data Download](https://fred.stlouisfed.org/series/DRSFRMACBS) -- data source link

Data Importing:

The .csv files that we used were pulled in using a pd.read\_csv code with an output of the entire data frame, and then segmented by each of the team members depending upon the columns we would primarily establish our focus.

Data Cleaning & Manipulation:

After the data was pulled in and segmented, each of us had some data cleaning to do; however, it was not easily recognizable that cleaning would be necessary. The grade data when compared to interest rate showed some outliers in grades D, E, F & G. The grades are set up by Lending Club and with further research found out that they have grades with subgrades, but they are consistent with the credit score. Grades D, E, F & G showed lower interest rates in many instances and had to be deleted.

Statistical Modeling & Machine Learning:

Throughout the exercise, statistical modeling was used to show correlations with increased borrowing driving by disasters within the U.S., and then used to determine loan volume when looking at stock prices. The modeling used showed no correlation to either. We did not use Machine Learning, but it will be used in future projects as we get more knowledge in this area.

Reporting and Visualization:

Python using Jupyter Notebook were our primary tools used to build code that would ultimately get to the plotting of the data. It was found that debt consolidation loans were the highest type of loans taken out with credit cards being second which is consistent with the type of organization that Lending Club is. The data was used to ask questions about when were loans the highest in 2018, what type of borrower took out loans and why.

Conclusion:

It is very important as a data analyst to not only know how to pull data, but to learn the business model in order to determine what can be reported and what can’t. The expectation is to not have pre-conceived notions that the data will be clean and easy to analyze, but that where there are challenges to continue to “dig in” and keep driving through until a “meaningful” story can be told. But that should only be the beginning of the data journey.