Report

The Data Visualization Process

By Aisha Awudu

9/23/2022

Contents

[Introduction 2](#_Toc113448667)

[Wrangle data: 2](#_Toc113448668)

[Gathering Data: 2](#_Toc113448669)

[Assessing Data: 2](#_Toc113448670)

[Cleaning Data 2](#_Toc113448671)

[Store, analyze and visualize the wrangled data 3](#_Toc113448672)

[Write a report on wrangling and analyses process 3](#_Toc113448673)

# Introduction

This project was done as a part of Udacity's Data Analyst Nanodegree program. The objective of the project is to use Python visualization libraries to explore a dataset systematically. The analysis begins with exploring univariate variables followed by bivariate and multivariate analysis.

This analysis is followed by a short presentation to convey and highlight important findings using explanatory data analysis. A slide deck was prepared that followed the major path of exploration and a story was constructed for the readers to understand what was discovered.

# Problem statement

This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, and many others.

# Metrics

# Process

## Variables Definition

**Term**: The length of the loan expressed in months.

**LoanStatus**: The current status of the loan: Cancelled, Chargedoff, Completed, Current, Defaulted, FinalPaymentInProgress, PastDue. The PastDue status will be accompanied by a delinquency bucket

**BorrowerRate**: The Borrower's interest rate for this loan.

**LenderYield**: The Lender yield on the loan. Lender yield is equal to the interest rate on the loan less the servicing fee.

**EstimatedLoss**: Estimated loss is the estimated principal loss on charge-offs. Applicable for loans originated after July 2009.

**EstimatedReturn**: The estimated return assigned to the listing at the time it was created. Estimated return is the difference between the Estimated Effective Yield and the Estimated Loss Rate. Applicable for loans originated after July 2009.

**ProsperScore**: A custom risk score built using historical Prosper data. The score ranges from 1-10, with 10 being the best, or lowest risk score. Applicable for loans originated after July 2009.

**Occupation**: The Occupation selected by the Borrower at the time they created the listing.

**IncomeRange**: The income range of the borrower at the time the listing was created.

**EmploymentStatus**: The employment status of the borrower at the time they posted the listing.

**EmploymentStatusDuration**: The length in months of the employment status at the time the listing was created.

**IsBorrowerHomeowner**: A Borrower will be classified as a homeowner if they have a mortgage on their credit profile or provide documentation confirming they are a homeowner.

**CreditScoreRangeLower**: The lower value representing the range of the borrower's credit score as provided by a consumer credit rating agency.

**CreditScoreRangeUpper**: The upper value representing the range of the borrower's credit score as provided by a consumer credit rating agency.

**RevolvingCreditBalance**: Dollars of revolving credit at the time the credit profile was pulled.

**TotalProsperLoans**: Number of Prosper loans the borrower at the time they created this listing. This value will be null if the borrower had no prior loans

**TotalProsperPaymentsBilled**: Number of on time payments the borrower made on Prosper loans at the time they created this listing. This value will be null if the borrower had no prior loans.

**OnTimeProsperPayment**:s Number of on time payments the borrower had made on Prosper loans at the time they created this listing. This value will be null if the borrower has no prior loans.

**ListingCategory**: The category of the listing that the borrower selected when posting their listing: 0 - Not Available, 1 - Debt Consolidation, 2 – Home Improvement, 3 - Business, 4 - Personal Loan, 5 - Student Use, 6 - Auto, 7- Other, 8 - Baby&Adoption, 9 - Boat, 10 – Cosmetic Procedure, 11 - Engagement Ring, 12 - Green Loans, 13 - Household Expenses, 14 - Large Purchases, 15 - Medical/Dental, 16 -Motorcycle, 17 - RV, 18 - Taxes, 19 - Vacation, 20 - Wedding Loans

**OnTimeProsperPayments**: Number of on time payments the borrower had made on Prosper loans at the time they created this listing. This value will be null if the borrower has no prior loans.

Using a dataset from Twitter, WeRateDogs twitter archive resource the objectives of the project was:

* Wrangle data
* Store, analyze and visualize the wrangled data
* Write a report on wrangling and analyses process

## Wrangle data:

I wrangled the data using the following processes:

### Gathering Data:

I gathered data from three different sources. These were through a csv file provided on the Udacity platform, through url, also provided on the Udacity platform and using the Twitter API, via Tweepy library a Python package.

### Assessing Data:

After gathering the data from the three sources mentioned above, I assessed the data for quality and structure issues. I used visual screening and programmatic assessment using code. Bearing in mind, four quality principles and three tidiness principles. The quality principles being:

* + - Completeness: Screen for missing data (e.g. Nan row entries)
    - Validity: If data meets requirement (e.g. data type of int instead of string)
    - Accuracy: If wrong entries where made (e.g. name entries like such, a, an, etc)
    - Consistency: if entries are of equal standards (e.g. different time formats)

Structural principles being:

* + - Check if variable forms a column
    - Check if observation forms a row
    - Check if type of observation unit forms a table

### Cleaning Data

Next after assessment and identifying the issues, I cleaned the data using three steps:

* + - Defining the solution
    - Coding the solution
    - Testing the code to ensure right solution has been used

Below are examples of some of the cleaning I did:

* I created copies of al datasets.
* Joining all three clean datasets into one data frame.
* Deleting retweets.
* Dropping columns that I did not need.
* Changing **tweet\_id** from an integer to a string.
* Changing the timestamp to correct **datetime** format.
* Correcting name entry issues.
* Creating a new **dog\_breed** column using the image prediction data.

## Store, analyze and visualize the wrangled data

After cleaning the newly created dataset, I stored it to a csv master file. The data was analyzed and visualization drawn from insights unfolded.

## Write a report on wrangling and analyses process

Two reports were written, one focusing on the output and analysis of the project. The other report being the wrangle report gives a summary of all the steps used in the data wrangling in the project.