**Project Objective:**

* **What is the goal?** *Identify the most common business type(s), funds received and what funds were used for and how to be better prepared for future needs and to mitigate potential fraudulent activity.*
* **What problem are you trying to solve?** *How to minimize loan fraud*
* **Who will be your audience?** *Stakeholders and Small Business Banking Fraud and Investigations*

**Data Collection Source:** *Data was collected via public domain search.*

**website links:**

https://www.kaggle.com/datasets/danb91/covid-ppp-loan-data-with-fraud-examples

<https://www.arnoldporter.com/en/general/cares-act-fraud-tracker>

<https://www.pppdetective.com/>

<https://www.stlouisfed.org/publications/regional-economist/2022/jul/was-paycheck-protection-program-effective>

https://www.federalpay.org/paycheck-protection-program/tx

**Methodology:**

* *Describe your preprocessing techniques. Data was pulled into Python and sampled, then necessary columns were selected. Columns were rearranged and some of the names were changed and performed data quality checks on remaining data.*
* *How was the data organized? Data was organized by rearranging columns for a better flow of the data and filtered out all data pertaining to TX in order to condense the dataset and to have a specific area of focus.*
* *What was the workflow?*
* *What tools were used to organize and better understand the data? Python and Excel*
* *How was the data cleaned?*
* *Were there any typos? N/A*
* *How was this corrected? N/A*
* *Any data type conversions, e.g. string to date values? N/A*
* *Any other formatting, e.g. renaming columns or values? Some columns were renamed and formated columns referencing percentages and dollar amounts.*
* *Were any missing values replaced? If so, describe why? No*
* *How will replacing the values impact the analysis?*
* *Are there any exclusions (columns, rows, or values)? No*
* *Include any other methods in processing and organizing the data. Excel proper function for borrower names, city and bank name columns.*

**Findings:**

1. **Data Size:** *How many rows and columns?****Rows 76,184 Columns 28***
2. **Missing Values:** *How many?* ***No NULL values***

*Will these need to be replaced or deleted?* ***N/A***

1. **Duplicates:** *Describe how this was detected.* ***No duplicates / dataframe.duplicated()***
2. **Dates:***What is the date range?* ***April 2020 to June 2021***

*Will this need to be modified for the project?* ***No***

1. **Additional Findings:** *Identify any other findings or errors in the data. Any outliers? Any inconsistencies relative to the project?* ***N/A***

**Attach any SQL or Python code in a separate file**

http://localhost:8888/notebooks/Desktop/Intro\_To\_Python/Capstone%20Data%20Quality%20Check.ipynb