mapdata\_finalization\_notebooks

* main\_label\_pipe23.ipynb
  + INPUT: “../data/raw/>150k/publick\_150k\_plus\_230630.csv”
  + Borrower State Cleanup for Null Listings
  + Borrower Address Cleanup for Null Listings
  + Temporary Replacement of Address Details (missing address,city,state) with “Not Available”.
  + Write to New CSV
  + Columns Overview
    - Names
    - Count of Nulls
    - Dtypes
  + Addition of Estimated Lender Percent based on Current Approval Amount of Loan
    - 150k – 350k = %.05
    - 350k – 2m = %.03
    - >2m = %.01
  + Addition of Estimated Lender Profit based on Lender Percent of Current Approval
    - Current Approval \* Lender Percent
  + Timestamp Conversions mm/dd/yyyy > yyyymmdd
    - DateApproved > DateApprovedTS
    - ForgivenessDate > ForgivenessDateTS
    - LoanStatusDate > LoanStatusDateTS
    - Replacement of NaT > NaN, fill of NaN with “Not Available”
  + Write to new CSV
  + ForgivenessAmount / UndisbursedAmount fill of NaN for “Not Available”, dtype > str
  + NAICS merging
    - INPUT: “../data/raw/NAICS/NAICSind.csv”
      * NA (2-digit NAICS field), Industry (Name)
    - INPUT: “../data/raw/NAICS/NAICSfull.csv”
      * NAICSCode (6-digit NAICS), LongName, TriIndustrySector, IndustrySubsector (4-digit NAICS field + Subsector Name), 4DigitNAICS, NA (2-digit NAICS field)
    - Input files merged on “NA”
      * Null Check
      * Dtypes
    - NAICSCode column in main DataFrame
      * fill of NaN with 0
      * dtype > int
    - Main DataFrame merges with final NAICS DataFrame on NAICSCode
      * Null Check for NAICS fields
      * NA, 4DigitNAICS, fill of NaN with 0, type > np.int64 > str
      * Industry, IndustrySubsector, LongName, fill of NaN with “Not Available”
      * NAICSCode = 999990, Industry, IndustrySubsector, Long Name = “Nonclassifiable Establishments”
      * Final Null Check for NAICS fields = 0
  + Addition of SBAOfficeLabel (District Office) based on SBAOfficeCode
  + Addition of Loan\_Range Labels column based on Current Approval Amount (5 Ranges)
    - 150k – 350k
    - 350k – 1m
    - 1m – 2m
    - 2m – 5m
    - 5m – 10m
  + Addition of job\_Range Labels column based on JobsReported (8 Ranges)
    - change of JobsReported with “0” string > 0 int
    - Fill of NaN with 0
    - <5
    - 5 – 9
    - 10 – 19
    - 20 – 49
    - 50 – 99
    - 100 – 124
    - 250 – 499
    - > 500
  + Addition of full\_add column - combining BorrowerAddress + BorrowerCity + BorrowerState
  + Addition of ServicingLenderFullAddress column - combining ServicingLenderAddress + ServicingLenderCity + ServicingLenderState
  + Determining Business Age & Type Naming Consistency
    - Replacing BusinessAgeDescription = Unanswered > “Not Available”
    - Formatting Limited Liability Company (LLC)
    - Fill of NaN with “Not Available” for BusinessAgeDescription, BusinessType, FranchiseName
    - Fill of NaN with “N” for NonProfit
  + Borrower Name Cleanup
    - Replacement of &AMP; > &
    - Removal of SBA SMALL 7A TERM > ‘’
  + Dropped Columns Not used in final pipe file:
    - 'BorrowerAddress', 'BorrowerCity', 'BorrowerZip', 'SBAGuarantyPercentage', 'InitialApprovalAmount', 'BusinessAgeDescription', 'ProjectCity', 'ProjectCountyName','ProjectState', 'ProjectZip','Race', 'Ethnicity', 'ServicingLenderAddress', 'ServicingLenderCity', 'ServicingLenderState', 'ServicingLenderZip', 'UTILITIES\_PROCEED', 'PAYROLL\_PROCEED', 'MORTGAGE\_INTEREST\_PROCEED', 'RENT\_PROCEED', 'REFINANCE\_EIDL\_PROCEED', 'DateApproved', 'ForgivenessDate', 'HEALTH\_CARE\_PROCEED', 'DEBT\_INTEREST\_PROCEED', 'Gender', 'Veteran'
  + Write to New CSV
    - “../data/ppp\_pipe23.csv”
* main\_hashing23.ipynb
  + INPUT: “../data/ppp\_pipe23.csv”
  + Industry Sector Review – View Groupings to determine best Hash combinations
    - Groupby Industry, Industry Subsector, Aggregate by Count of Loans
    - Groupby Industry, Industry Subsector, Loan\_Range, BusinessType, Aggregate by Count of Loans
    - Groupby Industry Subsector, 4DigitNAICS, Aggregate by LoanNumber Count
    - Groupby Industry, NA, 4DigitNAICS, Aggregate by Count of Loans
    - Groupby Industry, NA, Industry Subsector, 4DigitNAICS, Aggregate by Count of Loans
  + Hashing Code Values
    - NA, 4DigitNAICS Naming Consistency
      * NA = 0 > 00
      * 4DigitNAICS = 0 > 0000
    - Creating New Columns for Hash Variables
      * INDcode = “IND + NA” e.g., IND00
      * SScode = “SS + 4DigitNAICS.str[-2:]
      * LRcode based on Loan\_Range (5) conditions (LR1, LR2, etc.)
      * JRcode based on job\_Range (8) conditions (JR1, JR2, etc.)
      * BTcode based on BusinessType (26) (BT01, BT02, etc.)
      * BAcode based on BusinessAgeDescription (5) (BA1, BA2, etc.)
      * PMcode based on ProcessingMethog (PPP/PPS)
      * NPcode (NonProfit), LMIcode (LMIIndicator), HZcode (HubzoneIndicator), RUcode (RuralUrbanIndicator) - created (not used in final)
      * HASHcodePM based on INDcode + SScode + LRcode + JRcode + BTcode + BAcode + PMcode
        + Creating New DataFrame using LoanNumber, INDSScode, INDcode, SScode, LRcode, JRcode, BTcode, BAcode, PMcode, HASHcodePM, Industry, Industry Subsector
        + Verifying length of all rows in HASHcodePM = 25
      * Writing to new CSV
        + “../data/ppp\_hash23.csv”
      * Sorting Hash Arrays
        + Sorting DataFrame based on INDcode, SScode, LRcode, JRcode, BTcode, BAcode, PMcode values
        + Reducing columns in final DataFrame to LoanNumber, HASHcodePM, Industry, Industry Subsector
        + Writing to new CSV

“../data/ppp\_hashSort23.csv”

* main\_concat23.ipynb
  + INPUT: “../data/ppp\_pipe23.csv”
    - Dropping Columns NA, 4DigitNAICS, NAICScode, Industry, IndustrySubsector, TRIIndustrySubsector, Loan\_Range, job\_Range (information provided in subsequent Hash file)
  + INPUT: “../data/ppp\_hashSort23.csv”
  + INPUT: “../data/state\_data/geo/ppp\_geo\_fips23.csv”
  + Merging Pipe file with Hash file based on LoanNumber
  + Merging combined file with Geo Fips file based on LoanNumber
  + Null Checks
  + Final Dtype Check
  + Write to new CSV
    - “../data/pipe\_hpgf23.csv”