ETL Report

Dean Papadopoulos, Philip Waymeyer, Juliann Groglio, Savion Ponce

7/18/22

Introduction

This ETL report showcases the process required to extract and transform the Annual Business Survey 2019 dataset used in this analysis. This survey covers all demographic, financial, and technological aspects of business owners in the United States. Utilizing the Company Summary and Technology Characteristics of Businesses datasets, we cleaned and merged the data to eventually produce telling visuals.

Data Sources

https://www.census.gov/data/developers/data-sets/abs.2019.html

Extraction

In order to access our data we had to request an API Key from the census website. Listed below are the two links that provided our “Company Summary” and “Technology Characteristics of Businesses” datasets. Once we obtained our API key we used *request.get()* to import both of the datasets as listed in our steps for transformation.

https://api.census.gov/data/2018/abscs?get=GEO_ID,NAME,NAICS2017,NAICS2017_LABEL,SEX,SEX_LABEL,ETH_GROUP,ETH_GROUP_LABEL,RACE_GROUP,RACE_GROUP_LABEL,VET_GROUP,VET_GROUP_LABEL,EMPSZFI,EMPSZFI_LABEL,YEAR,FIRMPDEMP,FIRMPDEMP_F,RCPPDEMP,RCPPDEMP_F,EMP,EMP_F,PAYANN,PAYANN_F,FIRMPDEMP_S,FIRMPDEMP_S_F,RCPPDEMP_S,RCPPDEMP_S_F,EMP_S,EMP_S_F,PAYANN_S,PAYANN_S_F&for=us:*&NAICS2017=00&key=YOURKEYGOESHERE

https://api.census.gov/data/2018/abstcb?get=GEO_ID,NAME,NAICS2017,NAICS2017_LABEL,FACTORS_P,FACTORS_P_LABEL,FACTORS_U,FACTORS_U_LABEL,IMPACTWF_P,IMPACTWF_P_LABEL,IMPACTWF_U,IMPACTWF_U_LABEL,MOTPRODTECH,MOTPRODTECH_LABEL,MOTUSETECH,MOTUSETECH_LABEL,TECHSELL,TECHSELL_LABEL,TECHUSE,TECH

USE_LABEL,YEAR,RCPPDEMP_S,RCPPDEMP_S_F,RCPPDEMP_PCT_S,RCPPDEMP_PCT_S_F,EMP_S,EMP_S_F,EMP_PCT_S,EMP_PCT_S_F,PAYANN_S,PAYANN_S_F,PAYANN_PCT_S,PAYANN_PCT_S_F&for=us:*&key=YOURKEYGOESHERE

Transformation

1. Import the following libraries
    a. Pandas
    b. Matplotlib
    c. Plotly
    d. Seaborn
2. Read in API for company summary with the following lines:
    a. response = requests.get('https://api.census.gov/data/2018/abscs?get=GEO_ID,NAME,NAICS2017,NAICS2017_LABEL,SEX,SEX_LABEL,ETH_GROUP,ETH_GROUP_LABEL,RACE_GROUP,RACE_GROUP_LABEL,VET_GROUP,VET_GROUP_LABEL,EMPSZFI,EMPSZFI_LABEL,YEAR,FIRMPDEMP,FIRMPDEMP_F,RCPPDEMP,RCPPDEMP_F,EMP,EMP_F,PAYANN,PAYANN_F,FIRMPDEMP_S,FIRMPDEMP_S_F,RCPPDEMP_S,RCPPDEMP_S_F,EMP_S,EMP_S_F,PAYANN_S,PAYANN_S_F&for=us:*&NAICS2017=00&key=YOUR_KEY_HERE’)
    b. data = response.json()
    c. df = pd.DataFrame(data[1:], columns=data[0])

3. Overwrite company summary dataframe to only include columns with useful data:
    a. df = df[['SEX_LABEL','ETH_GROUP_LABEL','RACE_GROUP_LABEL','VET_GROUP_LABEL','EMPSZFI_LABEL','FIRMPDEMP','EMP','PAYANN','RCPPDEMP']]
4. Rename column labels:
    a. df.columns=['Sex','Ethnicity','Race','Vet Status','Firm Size','Number of Firms', 'Number of Employees','Payroll','Revenue']
5. Read in API for technology characteristics for business with the following lines
    a. response2 = requests.get('https://api.census.gov/data/2018/abstcb?get=GEO_ID,NAME,NAIC

S2017,NAICS2017_LABEL,FACTORS_P,FACTORS_P_LABEL,FACTORS_U,FACTORS_U_LABEL,IMPACTWF_P,IMPACTWF_P_LABEL,IMPACTWF_U,IMPACTWF_U_LABEL,MOTPRODTECH,MOTPRODTECH_LABEL,MOTUSETECH,MOTUSETECH_LABEL,TECHSELL,TECHSELL_LABEL,TECHUSE,TECHUSE_LABEL,YEAR,RCPPDEMP_S,RCPPDEMP_S_F,RCPPDEMP_PCT_S,RCPPDEMP_PCT_S_F,EMP_S,EMP_S_F,EMP_PCT_S,EMP_PCT_S_F,PAYANN_S,PAYANN_S_F,PAYANN_PCT_S,PAYANN_PCT_S_F&for=us:*&key=YOUR_KEY_HERE')

b. data2 = response2.json()

c. df2 = pd.DataFrame(data2[1:], columns=data2[0])

d. df2 = df2[['NAICS2017_LABEL','FACTORS_P_LABEL','FACTORS_U_LABEL','IMPACTWF_P_LABEL','IMPACTWF_U_LABEL','MOTPRODTECH_LABEL','MOTUSETECH_LABEL','TECHSELL_LABEL','TECHUSE_LABEL']]

6. Merge the two data frames with the following line of code:
    a. merged = pd.merge(left=df, right=df2, right_index=True, left_index=True,how='inner')
7. Changed the data type to int for the following columns: “Number of Employees”, “Number of Firms”, “Payroll”, and “Revenue”

Conclusion

After following the process above step by step any user should be able to replicate the dataset which was used for data analysis in this project. The dataframe named *merged* is the cleaned dataset in its final form.