***2012 Summary of Business Owners Census Survey Data***

Extract:

Transform:

1. Read data into Pandas data frame using .toPandas()
2. Drop top row to remove duplicate header values
3. Filter data frame to get firm totals only by industry, not demographic breakdowns by sex
   1. df = df[df['SEX'] == '001']
   2. Don’t forget to reset index after this step!
4. Drop columns unnecessary for analysis:
   1. ['GEO\_ID','SEX','SEX\_LABEL','RCPNOPD\_S','FIRMNOPD\_S','PAYANN\_S','EMP\_S','RCPPDEMP\_S','FIRMPDEMP\_S','RCPALL\_S','FIRMALL\_S','NAICS2012','NAICS2012\_F','RCPSZFI\_LABEL','RCPSZFI','PAYANN','index','GEO\_ID\_F']
5. Split [‘NAME’] column into [‘COUNTY’] and [‘STATE’] columns
   1. df.NAME.str.split(‘,’,expand=True)
   2. Drop [‘NAME’] and reorder columns so that [‘COUNTY’] and [‘STATE’] are first columns after index
   3. Order: [['COUNTY','STATE','YEAR','NAICS2012\_LABEL','FIRMALL','RCPALL','FIRMPDEMP','RCPPDEMP','EMP','FIRMNOPD','RCPNOPD']]
6. Drop duplicate and blank rows from data frame
7. Replace ‘S’ fields with value of 0 for columns [‘RCPALL’], [‘RCPPDEMP’], and [‘RCPNOPD’]
8. In [‘EMP’] column, replace non-castable int types to the median value of the respective range
   1. i.e. df['EMP'].replace('100 to 249 employees', '175', inplace=True)
9. Cast [‘FIRMALL’], [‘FIRMPDDEMP’], [‘FIRMNOPD’], [‘RCPALL’], [‘RCPPDEMP’], [‘RCPNOPD’], and [‘EMP’] .astype(‘int’) and then confirm proper data types
10. Replace fields equal to 0 with np.nan
11. Reset index
12. Iterate through [‘COUNTY’] column to remove the last 7 characters from each field (i.e. ‘ County’) so we are left with only the name of each county
13. Drop any extra index columns if necessary

Load: