Importing research-specific data files

**Workflow Chart and Process Design**

import pandas as pd

Data1='C:/Users/Dell/Desktop/Data Files/IPEDS Data.xlsx'

Data2='C:/Users/Dell/Desktop/Data Files/Poverty Rate by Zipcode.xlsx'

Data3='C:/Users/Dell/Desktop/Data Files/Educational Attainment.csv'

IPEDSData=pd.read\_excel(Data1)

PovData=pd.read\_excel(Data2)

EduData=pd.read\_csv(Data3)

Python Commands

Action/Result

from uszipcode import SearchEngine

def retzip(lat, lng, rad):

search = SearchEngine()

result = search.by\_coordinates(lat, lng, radius=rad, returns=30)

ls\_zip = []

for zip in result:

ls\_zip.append(zip.values()[0])

return ls\_zip

Calling the “uszipcode” programmable database which includes up-to-date census and geometry information, specifically the latitudes and longitudes of zip codes within the U.S.

Creating a loop to generate a list of zip codes which fall within 6 miles of Texas higher education institutions’ latitudes and longitudes.

uni\_rad = pd.DataFrame()

uni\_rad['uni'] = IPEDSData['Institution']

uni\_rad['lat'] = IPEDSData['Latitude']

uni\_rad['lng'] = IPEDSData['Longitude']

uni\_rad['zips'] = pd.Series()

for i in range(0, len(uni\_rad)):

res = retzip(uni\_rad.loc[i]['lat'], uni\_rad.loc[i]['lng'], 6)

uni\_rad['zips'][i] = res

uni\_rad

Creating a loop to generate the average poverty rates of zip codes that fall within 6 miles of a higher education institution in Texas.

PovData = PovData.replace("-", 0)

PovData = PovData.replace(np.nan, 0)

uni\_rad['avg\_povrate'] = pd.Series()

for i in range(0, len(uni\_rad)):

ls = uni\_rad['zips'][i]

num = 0

for zip in ls:

rate1 = float(PovData[PovData['ZIP\_NEW']==int(zip)].PovRate12.item())

rate2 =float(PovData[PovData['ZIP\_NEW']==int(zip)].PovRate13.item())

rate3 = float(PovData[PovData['ZIP\_NEW']==int(zip)].PovRate14.item())

rate4 = float(PovData[PovData['ZIP\_NEW']==int(zip)].PovRate15.item())

rate5 = float(PovData[PovData['ZIP\_NEW']==int(zip)].PovRate16.item())

rate6 = float(PovData[PovData['ZIP\_NEW']==int(zip)].PovRate17.item())

totalrate=rate1 + rate2 + rate3 + rate4 + rate5 + rate6

num = num + totalrate

if len(ls)>0:

avg = num/(6\*len(ls))

else:

avg = np.nan

uni\_rad['avg\_povrate'][i] = avg

print(ls, avg, totalrate)

uni\_rad

import statsmodels.api as smmodel = sm.OLS(y4, x4, missing ='drop').fit()

reg = model.predict(X) # make the predictions by the model

model.summary()

x = uni\_rad["avg\_povrate"]

y = IPEDSData["Avg\_enroll\_12-17"]

model = sm.OLS(y, x, missing ='drop').fit()

reg = model.predict(X) # make the predictions by the model

model.summary()

Running an ordinary least squares regression to analyze the effect of the average poverty rate of zip codes that fall within 6 miles of a higher education institution on the average enrollment of students at a Texas higher education institutions between 2012 and 2017. The output showed an R-squared value of 0.884 and a p-value of 0.000 which shows statistical significance.