DSC-540 Final Project

Chitramoy Mukherjee-DSC540-T304

Date : 09/22/2023

3 data sources and it's descriptions

1. acs2017_county_data.csv : This data file contains US county level census data for year-2017.This dataset is downloaded from kaggle.
2. Wikipedia List of states table contains US state information. Thsi table contains US 50 states information.
3. US Government Data: The US government provides a wide range of public APIs, including data on demographics, economics, and crime. (https://www.census.gov/data/developers/data-sets.html)

Relationship between 3 data sources

US Census Data(acs2017_county_data.csv) and the List of US States table can be linked by geographic location ans state name. We could use the Google Maps API to determine the latitude and longitude for each state in the List of US States table, and then use this information to link the state data to the demographic and economic data in the US Census Data dataset.

Interpretation and operations on dataset to accomplish future milestones

```
    Based on the state name and it's geographic information, we can
merge this 3 datasets after removing the headers from    those. After
the first step will remove the unwanted columns from the datasets
andf then merge those three into one dataset and that dataset could
be used to inform policy makers and economic developers about the
factors that contribute to population growth. It could also be used
to identify states that are at risk of population decline, and to
develop targeted interventions to promote population growth in these
states.

    As a data wrangling project using these datasets would be to
create a dataset that maps the demographics and economic factors of
each US state to the state's population growth rate. This could be
done by linking the US Census Data dataset and the List of US States
table, as described above. Once the datasets are linked, we could use
statistical analysis to calculate the population growth rate for each
state, and then identify correlations between the population growth
rate and demographic and economic factors, such as median income,
poverty rate, and education levels.
```

Data Disctionary for acs2017_county_data.csv :

Data columns (total 37 columns):

Column No. Column Data type Description

---

0 Countyid int64 County identification #
1 State object Name of the state 2 County object Name of the county 3 TotalPop int64 Total population 4 Men int64 Men count 5 Women int64 Women count 6 Hispanic float64 % of population that is Hispanic/Latino 7 White float64 % of population that is white 8 Black float64 % of population that is black 9 Native float64 % of population that is Native American or Native Alaskan 10 Asian float64 % of population that is Asian 11 Pacific float64 % of population that is Native Hawaiian or Pacific Islander 12 VotingAge int64 Voting age in days
13 Income float64 Median household income ( )$14 IncomeErr float64 Medianhouseholdincomeerror()$ 15 IncomePerCap float64 Income per capita ()$16 IncomePerCapErr float64 Incomepercapitaerror()$ 17 Poverty float64 % under poverty level 18 ChildPoverty float64 % of children under poverty level 19 Professional float64 % employed in management, business, science, and arts 20 Service float64 % employed in service jobs 21 Office float64 % employed in sales and office jobs 22 Construction float64 % employed in natural resources, construction, and maintenance 23 Production float64 % employed in production, transportation, and material movement 24 Drive float64 % commuting alone in a car, van, or truck 25 Carpool float64 % carpooling in a car, van, or truck 26 Transit float64 % commuting on public transportation 27 Walk float64 % walking to work 28 OtherTransp float64 % commuting via other means 29 WorkAtHome float64 % working at home 30 MeanCommute float64 Mean commute time (minutes) 31 Employed int64 Number of employed (16+) 32 PrivateWork float64 % employed in private industry 33 PublicWork float64 % employed in public jobs
34 SelfEmployed float64 % self-employed 35 FamilyWork float64 % in unpaid family work 36 Unemployment float64 Unemployment rate (%)

List of states Wikipedia Table data dictionary :

Column No. Column Data type Description

---

1 Postal abbrevation object State Name 2 Cities object Major City bypopulation/state capital 3 Established Date Year state formed 4 Population int64 total state population 5 Total area int64 Total area 6 Land area int64 Total land 7 Water area int64 Total water area

Project subject area

Will apply different data wragling techniques on the source data and merge it to perform the analysis.

As a part of this project we will be merging 3 different dataset of differnt type using a common key(state name) and will perform statistical analysis to identify correlations between crime rates

and demographic and economic factors, such as poverty, unemployment, and education levels.

Data Sources:

1. acs2017_county_data.csv (https://www.kaggle.com/code/alawdisoft/us-census-demographic-data/input?select=acs2017_county_data.csv)
2. The US government provides a wide range of public APIs, including data on demographics, economics, and crime. US Census Bureau provides an API for accessing census data. (https://www.census.gov/data/developers/data-sets.html)
3. his Wikipedia table contains a list of all 50 US states, along with their capitals and population.(https://simple.wikipedia.org/wiki/List_of_U.S._states)

Relationships :

All 3 datasets contain data based on state. The lowest granularity of this 3 dataset data is state name.

Ethical implications and Challenges :

Ethical implications of using US Census Data for a data wrangling project include:

Privacy: The US Census Data contains personal information about individuals and households. It is important to take steps to protect the privacy of this data, such as anonymizing the data or using differential privacy techniques.

Bias: The US Census Data may be biased in certain ways. For example, it may be more difficult to reach certain populations, such as low-income households or immigrant communities. It is important to be aware of these potential biases and to take steps to mitigate them.

Discrimination: The US Census Data could be used to discriminate against certain groups of people. For example, it could be used to target certain groups with marketing messages or to deny them access to services or opportunities. It is important to use the data in a responsible and ethical way to avoid discrimination.

Use differential privacy techniques: Differential privacy is a set of techniques that can be used to protect the privacy of individuals in a dataset while still allowing for accurate analysis.

Some of the challenges that you might face in a US Census Data project include:

Data quality: The US Census Data is a large and complex dataset. It is important to carefully clean and prepare the data before using it for analysis.

Data complexity: The US Census Data contains a wide range of variables. It is important to understand the meaning of the variables and how they can be used for analysis.

Ethical considerations: As discussed above, there are a number of ethical considerations that must be taken into account when using US Census Data. It is important to design your project in a way that respects the privacy of the data and avoids bias and discrimination.

Milestone-2

Apply 5 transformations to acs2017_county_data.csv dataset

In [142…

```
import pandas as pd #Linear Algebra
import numpy as np #Data Processing
import seaborn as sns #Visualization
import matplotlib.pyplot as plt #Visualization
import pandasql as psql
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_17608\2747350820.py in <module>
      3 import seaborn as sns #Visualization
      4 import matplotlib.pyplot as plt #Visualization
----> 5 import pandasql as psql

ModuleNotFoundError: No module named 'pandasql'
```

In [123…

```
import pandas as pd

# Load the CSV file
file_path = 'C:\\Users\\14024\\OneDrive\\Desktop\\MS-DSC\\DSC-540\\DSC-540 Project\\Mi
data = pd.read_csv(file_path)

# Optionally, you can also display the first few rows to verify the new headers
print(df.head())
```

```
   CountyId    State          County  TotalPop        Men      Women  \
0      1001  Alabama  Autauga County     55036  48.875282  51.124718
1      1003  Alabama  Baldwin County    203360  48.941286  51.058714
2      1005  Alabama  Barbour County     26201  53.341476  46.658524
3      1007  Alabama     Bibb County     22580  54.255979  45.744021
4      1009  Alabama   Blount County     57667  49.404339  50.595661

   Hispanic  White  Black  Asian  ...  OtherTransp  WorkAtHome  MeanCommute  \
0       2.7   75.4   18.9    0.9  ...          1.3         2.5         25.8
1       4.4   83.1    9.5    0.7  ...          1.1         5.6         27.0
2       4.2   45.7   47.8    0.6  ...          1.7         1.3         23.4
3       2.4   74.6   22.0    0.0  ...          1.7         1.5         30.0
4       9.0   87.4    1.5    0.1  ...          0.4         2.1         35.0

    Employed  PrivateWork  PublicWork  SelfEmployed  FamilyWork  Unemployment  \
0  43.811323         74.1        20.2           5.6         0.1           5.2
1  44.023899         80.7        12.9           6.3         0.1           5.5
2  33.884203         74.1        19.1           6.5         0.3          12.4
3  36.186891         76.0        17.4           6.3         0.3           8.2
4  37.074930         83.9        11.9           4.0         0.1           4.9

   OtherRace
0        0.3
1        0.8
2        0.2
3        0.4
4        0.3

[5 rows x 36 columns]
```

In [124…

```
# Modify the column headers with prefix "US_2017_"
data_census_2017 =  data.add_prefix('US_2017_')
data_census_2017.head()

# Modifying header/column name with US_2017_ to identify the data corresponds to US ar
```

Out[124]:

	US_2017_CountyId	US_2017_State	US_2017_County	US_2017_TotalPop	US_2017_Men	US_2017_Wo
0	1001	Alabama	Autauga County	55036	26899	2
1	1003	Alabama	Baldwin County	203360	99527	10
2	1005	Alabama	Barbour County	26201	13976	1
3	1007	Alabama	Bibb County	22580	12251	1
4	1009	Alabama	Blount County	57667	28490	2

5 rows × 37 columns

In [125…

```
# Grouping the data based on US_2017_state and US_2017_county column.

vars_to_merge = [x for x in data_census_2017.columns if x not in ['US_2017_CountyId',

data_census_agg = pd.DataFrame(data_census_2017.groupby(['US_2017_State', 'US_2017_Cou
data_census_agg.head(5)
```

Out[125]:

US_2017_State	US_2017_County	US_2017_TotalPop	US_2017_Men	US_2017_Women	US_2017_Hispanic
Alabama	**Autauga County**	55036	26899	28137	2.7
	Baldwin County	203360	99527	103833	4.4
	Barbour County	26201	13976	12225	4.2
	Bibb County	22580	12251	10329	2.4
	Blount County	57667	28490	29177	9.0

5 rows × 34 columns

In [129…

```
# check duplicates if any in the dataset
data_census_agg.duplicated().sum()

# There is no duplicate rows in data based on the columns present in the dataset.
```

Out[129]:

```
0
```

In [130…

```
# summary of zero values
zero_values = (data_census_agg == 0).sum()
print("Zero values:\n", zero_values)

# There are several columns in the dataset wth 0 values. However, when considering the
# will be good drop the rows that have 0 values. This because the TotalPop column repr
# and including rows with 0 values would not contribute meaningful insights to the and
```

```
Zero values:
 US_2017_TotalPop               0
US_2017_Men                     0
US_2017_Women                   0
US_2017_Hispanic               13
US_2017_White                   3
US_2017_Black                 189
US_2017_Native                481
US_2017_Asian                 388
US_2017_Pacific              2393
US_2017_VotingAgeCitizen        0
US_2017_Income                  0
US_2017_IncomeErr               0
US_2017_IncomePerCap            0
US_2017_IncomePerCapErr         0
US_2017_Poverty                 0
US_2017_ChildPoverty            9
US_2017_Professional            0
US_2017_Service                 1
US_2017_Office                  0
US_2017_Construction            1
US_2017_Production              1
US_2017_Drive                   0
US_2017_Carpool                 1
US_2017_Transit               640
US_2017_Walk                   14
US_2017_OtherTransp            72
US_2017_WorkAtHome              6
US_2017_MeanCommute             0
US_2017_Employed                0
US_2017_PrivateWork             0
US_2017_PublicWork              0
US_2017_SelfEmployed            1
US_2017_FamilyWork            610
US_2017_Unemployment           11
dtype: int64
```

In [131…

```
# since population percentage of 'Native' and 'Pacific' is very less, we can merge the
data_census_agg['OtherRace'] = data_census_agg['US_2017_Native'] + data_census_agg['US
data_census_agg.drop(['US_2017_Native', 'US_2017_Pacific'], axis=1, inplace=True)
data_census_agg.head()
```

Out[131]:

US_2017_State	US_2017_County	US_2017_TotalPop	US_2017_Men	US_2017_Women	US_2017_Hispanic
Alabama	**Autauga County**	55036	26899	28137	2.7
	Baldwin County	203360	99527	103833	4.4
	Barbour County	26201	13976	12225	4.2
	Bibb County	22580	12251	10329	2.4
	Blount County	57667	28490	29177	9.0

5 rows × 33 columns

In [137…
```
# change absolute columns to percentage
absolutes = ['US_2017_Men','US_2017_Women','US_2017_VotingAgeCitizen','US_2017_Employe
data_census_agg[absolutes] = data_census_agg[absolutes].div(data_census_agg["US_2017_T
data_census_agg.head()
```

Out[137]:

US_2017_State	US_2017_County	US_2017_TotalPop	US_2017_Men	US_2017_Women	US_2017_Hispanic
Alabama	**Autauga County**	55036	0.088806	0.092893	2.7
	Baldwin County	203360	0.024066	0.025108	4.4
	Barbour County	26201	0.203586	0.178079	4.2
	Bibb County	22580	0.240283	0.202586	2.4
	Blount County	57667	0.085672	0.087738	9.0

5 rows × 33 columns

In [140…
```
# The plot shows how all these variables related each other.We can use this plot to ur
# between demographic variables and find outliers. We can inspect the supected data pc
fig, axes = plt.subplots(nrows=2, ncols=5)
fig.set_figheight(5)
fig.set_figwidth(20)

data_census_agg.plot(ax=axes[0,0],x='US_2017_TotalPop', y='US_2017_Hispanic', kind = '
data_census_agg.plot(ax=axes[0,1],x='US_2017_Employed', y='US_2017_IncomePerCap', kinc
data_census_agg.plot(ax=axes[0,2],x='US_2017_Hispanic', y='US_2017_IncomePerCap', kinc
data_census_agg.plot(ax=axes[0,3], x='US_2017_IncomePerCap', y='US_2017_Poverty', kinc
data_census_agg.plot(ax=axes[0,4], x='US_2017_IncomePerCap', y='US_2017_ChildPoverty',
data_census_agg.plot(ax=axes[1,0], x='US_2017_Unemployment', y='US_2017_IncomePerCap',
data_census_agg.plot(ax=axes[1,1], x='US_2017_Hispanic', y='US_2017_Poverty', kind = '
data_census_agg.plot(ax=axes[1,2], x='US_2017_Hispanic', y='US_2017_ChildPoverty', kir
data_census_agg.plot(ax=axes[1,3], x='US_2017_Unemployment', y='US_2017_Poverty', kinc
data_census_agg.plot(ax=axes[1,4], x='US_2017_Unemployment', y='US_2017_ChildPoverty',

plt.tight_layout()
```