ISQS 6339 – Fall 2023 – Group 6

Final Project

October 10, 2023

INTRODUCTION

In this technical report, we present a comprehensive overview of a Business Intelligence project aimed to investigate and analyze the potential correlation between specific disease prevalence and key demographic factors within a specified geographical area. Raw data was sourced from various governmental agencies. Using Python scripting, it was extracted, transformed, and loaded according to ETL best practices. The objective of this project is to gather meaningful data and use it to provide valuable insights that will inform the selection of optimal locations for new healthcare clinics. By understanding how these diseases correlate with factors such as the number of liquor establishments, median household income, population density, internet access, percentage of males, and percentage of females, we can make data-driven decisions to improve healthcare access and address community needs.

DATA DOMAIN

The foundation of our data analysis is built upon three distinct datasets related to the healthcare industry. The team elected to go with healthcare data as Dr. Dwayne Hoelscher has 28 years of hospital based healthcare experience including risk management, utilization review, and the last 15 years in informatics. In his experience, there are areas across the United States that lack healthcare access. Additionally, research is beginning to show alcohol consumption is related to increased rates of cancer even with minimal weekly consumption.

According to Shahzad et al (2020, June 26), movement in the healthcare industry is shifting toward an emphasis on population health management and health equity. Shahzad and his colleagues gathered evidence from over 21 empirical examples that demonstrated a shift toward using data to enhance clinical decision making. With the Social Determinants of Health dataset, the healthcare industry can identify areas of improvement in healthcare delivery and can inform policy decisions. The datasets they provide can be used to identify disadvantaged and vulnerable populations, and also help identify communities who would benefit the most from interventions. In conjunction with the Diagnosis by County dataset, populations vulnerable specifically to cancer, depression, and diabetes can easily be identified and targeted for areas of improvement and growth.

The healthcare industry is also shifting toward the use of artificial intelligence and machine learning including personalized medicine and prescriptive healthcare. Bohr & Memarzadeh (2020) stated that “[the massive amount of data we accumulate] can be used for very detailed personal profiling, which may be of great value for behavioral understanding and targeting but also has potential for predicting healthcare trends.” When combined, our datasets show trends in populations and specific human behaviors that may make certain populations more susceptible to specific diseases.

1. **Centers for Disease Control and Prevention (CDC)**
   One of our datasets (Diagnosis_State_County.csv) is from the Centers for Disease Control and Prevention, a public health and epidemiology domain. The CDC’s national center for health statistics is a tool used by researchers, public health officials, and even policymakers to make informed decisions regarding public health. The CDC’s data can provide valuable information in this effort by identifying vulnerable populations, healthcare gaps, and regions where specific

health conditions are most prevalent, thus guiding resource allocation and the deployment of specialized healthcare services.

2. **U.S. Census Bureau**
   The "Liquor_Establishments.csv" dataset is from the County Business Patterns (CBP) dataset in the U.S. Census Bureau's data platform. The U.S. Census Bureau is a federal agency responsible for gathering comprehensive demographic information about the U.S. and its population. The CBP specifically focuses on gathering information related to business establishments. It is a useful platform for researchers, economists, businesses, and entrepreneurs seeking to understand economic trends across periods of time and geographic locations in the U.S. Data from this domain can be used to make well-informed decisions about where to establish or expand their operations. In conjunction with healthcare data, this can be used to pinpoint business development opportunities within the healthcare sector.

3. **Agency for Healthcare Research and Quality (AHRQ)**
   Our third dataset, "SDOH.xlsx," is associated with the AHRQ. It is a government agency that focuses on improving America's healthcare system. As stated on the AHRQ's website, its mission is to produce evidence to make health care safer, higher quality, more connected, and more accessible. They provide data so that healthcare professionals, policymakers, and the public can make informed decisions about their healthcare outcomes.

ANALYSIS OF DATA

1. **Diagnosis by State and County (PLACES: Local Data for Better Health, County Data 2022 Release | Data | Centers for Disease Control and Prevention, 2023)**
   This dataset, denoted as "Diagnosis_State_County.csv," constitutes a comprehensive snapshot of health conditions across the United States in the year 2020. It includes 21 variables that offer a granular understanding of disease prevalence patterns at the county level. In an attempt to explore the relationship between key demographics and disease prevalence, we chose to primarily utilize the following columns: LocationID (county FIPS code), Data_Value (prevalence of the respective disease), and MeasureID (diagnosis). The county FIPS code, represented as integers, was chosen in lieu of the corresponding string-type county and state names in string format. This choice was made to mitigate potential issues related to data consistency associated with string-type variables.

2. **Liquor Establishments by State and County (U.S. Census Bureau, n.d.)**
   The "Liquor_Establishments.csv" dataset presents a unique perspective by providing the presence of beer, wine, and liquor stores at the county level in 2020. This dataset assumes significance in the context of public health, particularly in potential correlations between the density of liquor establishments and disease prevalence within counties. There are a total of 25 variables in this dataset, only three of which we chose to use: geographic identifier code, employment size of establishments code, and number of establishments. The geographic identifier code is an object type variable containing numbers and letters. The county FIPS code was contained inside each geographic identifier code, which was extracted and used to identify the counties. The employment size of establishments code was used to ensure we gathered data from all establishments, and the number of establishments is the key variable we used to explore correlations between liquor establishments and disease prevalence.

3. **Social Determinants of Health (Social Determinants of Health Database (Beta Version), n.d.)**
   Our third dataset, "SDOH.xlsx," from the Agency for Healthcare Research and Quality, provides various demographic factors by county. There were numerous variables associated with this dataset, most of which were determined to be unnecessary to our research. We chose to keep the following variables: county FIPS code (integer), state FIPS code (integer), state name (string), county name (string), total civilian population above 18 (float), median household income (float), number of households without access to internet (float), female population percentage (float), and male population percentage (float).

   The FIPS codes were used to merge datasets, and the corresponding state and county names were used for readability purposes in our visualizations. The population variable was used in conjunction with the liquor establishments dataset to standardize the value of liquor establishments. Instead of using the raw count of establishments, we divided the population by the number of establishments. This calculation allowed us to determine the number of individuals per liquor store, providing a more meaningful comparison across various populations. The remaining variables were the key factors used to explore the correlation between demographic factors and disease prevalence.

DATA CLEANING

The overall quality of the data in this report can be considered good. They are well-structured datasets from reputable sources as previously discussed.

The following is a list of considerations made and limitations found when cleaning the data for analysis:

1. **Temporal limitation:** The cross-sectional datasets shared one common year, 2020, limiting our analysis to a singular snapshot of the examined metrics. It is imperative to note that the public health landscape in 2020 encompassed many factors that may have greatly influenced the outcomes in ways that are unaccounted for.

2. **Data completeness:** The Social Determinants of Health (SDOH) dataset was only missing data from territories outside the United States, which we excluded. Due to the fact that we possessed alcohol related information for approximately half of the counties, we imputed the missing value based on population and number of establishments using the Python scikit module K-Nearest Neighbors (KNN) imputation. KNN will return a mean of the number of liquor establishments from a given population for the closest 3 population counts by county.

3. **Derived data:** Given that the Liquor Establishments dataset contained information from counties of differing sizes, our team concluded that utilizing imputation with the raw count of liquor establishments would yield valuable insights. Consequently, we derived the establishment density for each county by dividing the number of liquor establishments by the respective total population after KNN imputation, found in the Social Determinants of Health demographic dataset.

4. **Missing data:** As a follow-up to the above mentioned derived data, the computation of liquor establishment density utilized the Social Determinants of Health (SDOH) dataset. To mitigate potential bias stemming from the inclusion of children in certain communities, our team opted not to utilize the total population figure. Instead, we made the decision to employ an alternative variable found within the SDOH dataset, namely, the total population aged 18 and above. Although this does not precisely capture the population, it represents the nearest available approximation.

5. **Data transformation:** The 'Geographic identifier code' field in the Liquor Establishments dataset contained the STATEFIPS and COUNTYFIPS in a combined field.To provide a useable key to join, the three datasets, the 'Geographic identifier code' was split and only the COUNTYFIPS was retained.

6. **Data validation:** The data quality overall was good. Minimal cleaning was required as all fields were imported into Python in the expected format (objects, integers, floats, etc.). We checked our key variables for outliers using descriptive statistics- comparing the mean, median, min, and max for any outliers or anomalies (see table I). Additionally, we checked for consistency throughout the merging process by continually counting the number of records to ensure we were not losing data or creating duplicates.

DATA MERGING

Following data selection and cleaning, we relied on the countyfips code to merge the three data sets. The SDOH was used as a primary key for countyfips as it was the most complete. We then combined all datasets employing a left join. The main issue we encountered was splitting the Geographic identifier code on the ETOH dataset. When we split at the US delimiter, the countyfips contained a leading zero when the other two data sets did not have a leading zero, thus requiring striping said zero.

The original liquor establishments data set is missing a significant number of counties compared to our other data sets. We used KNN imputation to fill in the missing county data.

As previously mentioned, we created a variable called population per establishment. Our liquor store dataset initially included a raw count of the number of liquor establishments in each county. Recognizing the variation in county populations, we believe it would be more informative to calculate the population per liquor store.

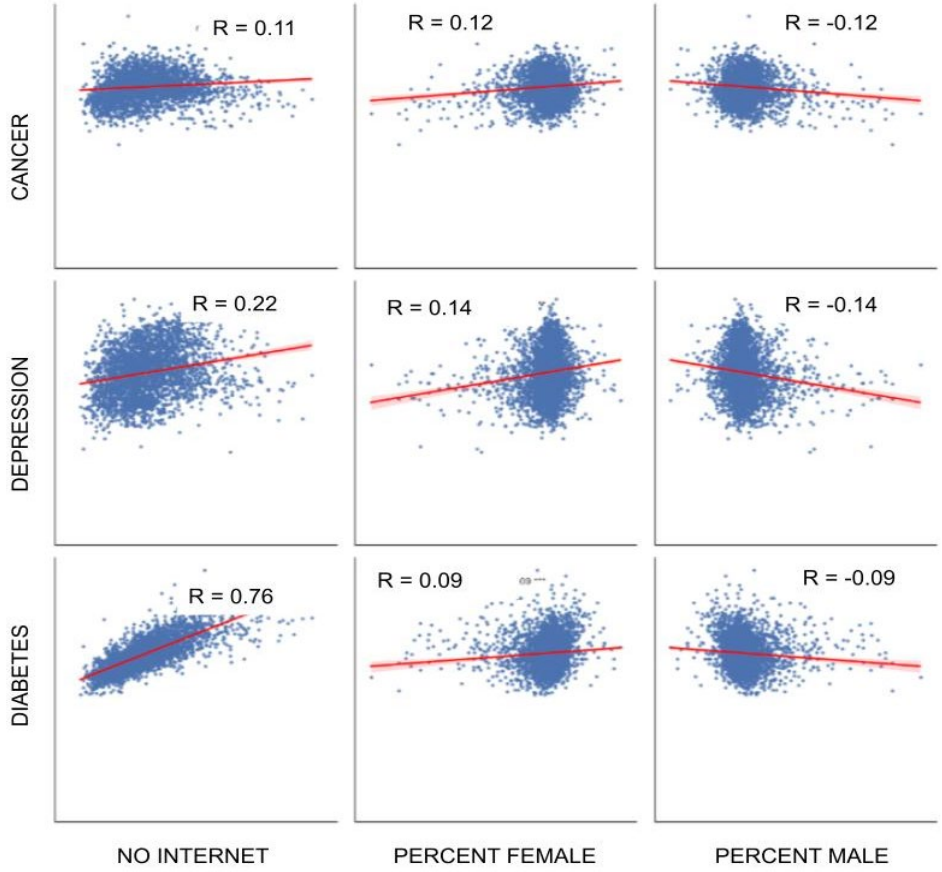
VISUALIZATION OF DATA

Graphs A and B, below, and TABLE II (in the appendix) illustrate correlation between our selected demographic variables and the prevalence of disease. Notably, the combination most strongly correlated, with an R-value of 0.76, is the lack of internet access and diabetes. Our data shows that counties with limited internet access exhibit higher rates of diabetes. This aligns with the correlation observed between median household income and diabetes, which also exhibits a significant (negative) correlation with an R-value of -0.71. As median household income rises, there is a decline in diabetes rates.

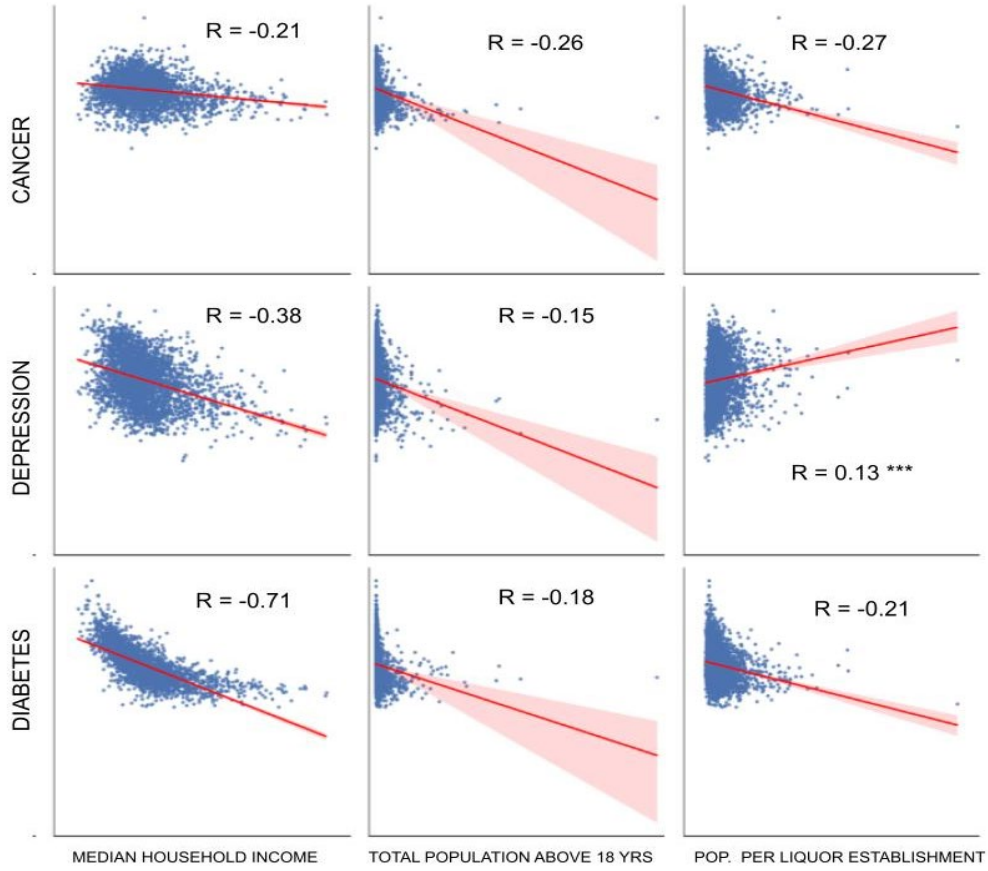
Additionally, our team made an unexpected discovery: there is a weak negative correlation between the density of liquor establishments and diabetes rates. Additionally, we were surprised to find a moderate negative correlation, with an R-value of -0.38, between depression and household income. Our data suggests that counties with higher median household incomes may have higher rates of depression.

It is important to acknowledge that there could be confounding variables influencing these findings. For instance, it is possible that higher median household incomes could lead to increased access to mental health treatment, and consequently, a greater likelihood of diagnosing depression.

GRAPH A



GRAPH B



ADDITIONAL DOCUMENTATION

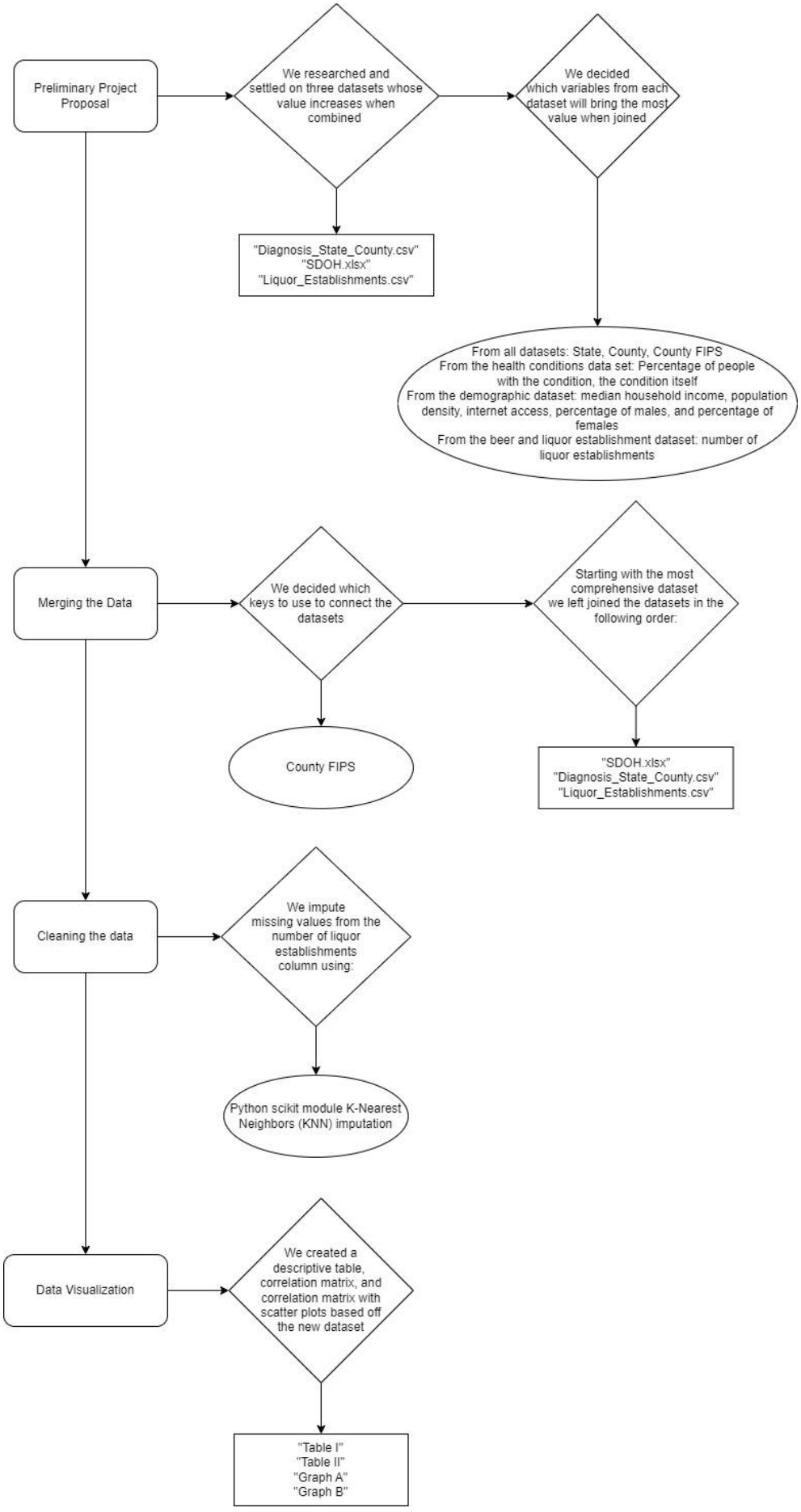
In ensuring the quality of our dataset for future use in MSDS courses, we have carefully addressed specific metrics outlined for this purpose. Firstly, to meet the requirement of a minimum of 500 records, we reviewed our data collection strategy. While expanding the time frame wasn't feasible, we optimized the utilization of the existing dataset, making sure that each record contributes meaningfully to the analyses conducted.

We identified key variables to meet the standard of at least 5 predictors and 3 target variables. Our selection included demographic factors such as the number of liquor establishments, median household income, population density, internet access, percentage of males, and percentage of females. For target variables, we chose diseases with distinct prevalence patterns, including a range from cancer to depression and diabetes. This diversified approach allows for a subtle exploration of the correlations between demographic factors and various health outcomes.

To meet the requirement of having at least 5 number things that connect meaningfully, we looked at possible predictors and targets. We did some math, matrix correlation and showed a chart that reveals important connections in our project. These charts don't just make our study deeper, but it's also a helpful thing for the future to understand the complex links in the data.

In summary, these metrics make our dataset more flexible and useful for upcoming MSDS courses. Every choice we've made in selection, conversion, and correlation of variables adds depth to the dataset, making it a strong base for ongoing learning and research in the field of data science.

FLOW DIAGRAM



INSTRUCTIONS FOR CODE

The following is a list of libraries required for our code: pandas, KNNImputer from the sklearn.impute module, matplotlib.pyplot, seaborn, and stats from scipy.

Actual library import code:

```
import pandas as pd
from sklearn.impute import KNNImputer
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
```

We utilized three datasets for our analysis. We chose to download and save the datasets locally. In order to run our code, the user will need to save the datasets and direct the following filepaths to the corresponding files:

```
VAR_FILE_PATH = 'filepath to folder with all datasets'

VAR_SDOH = 'SDOH.xlsx'
    #From https://www.ahrq.gov/downloads/sdoh/sdoh_2020_tract_1_0.xlsx

VAR_DX = 'Diagnosis_State_County.csv'
    # From https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-
    County-Data-20/duw2-7jbt

VAR_ETOH = 'Liquor_Establishments.csv'
    #From
    https://data.census.gov/table?q=CBP2020.CB2000CBP&n=4453&tid=CBP2020.CB2000C
    BP
```

REFERENCES

Bohr, A., & Memarzadeh, K. (2020, June 26). *The rise of artificial intelligence in healthcare applications*. National Library of Medicine. Retrieved October 4, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/

PLACES: Local Data for Better Health, County Data 2022 release | Data | Centers for Disease Control and Prevention. (2023, June 15). https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-County-Data-20/duw2-7jbt

Shahzad, M., Upshur, R., Donnelly, P., Bharmal, A., Wei, X., Feng, P., & Brown, A. (2020, June 26). *A population-based approach to integrated healthcare delivery: A scoping review of clinical care and public health collaboration*. BMC Public Health. Retrieved October 1, 2023, from https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-7002-z

Social Determinants of Health Database (Beta Version). (n.d.). https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html

U.S. Census Bureau. (n.d.). Explore Census data. https://data.census.gov/table?q=CBP2020.CB2000CBP&n=4453&tid=CBP2020.CB2000CBP

APPENDIX A

TABLE I

Index	count	mean	std	min	25%	50%	75%	max	median	var
CANCER	3143	7.54750239	1.16116311	2.7	6.8	7.6	8.3	13.6	7.6	1.34829978
DEPRESSION	3143	21.0914731	3.14970983	10.7	18.8	21.1	23.3	30.7	21.1	9.92067202
DIABETES	3143	12.7316577	2.61941413	6.2	10.9	12.4	14.4	25.9	12.4	6.86133038
ACS_PCT_HH_NO_INTERNET	3143	18.1756538	7.67033892	2.16875	12.8546429	17.1584615	22.3416667	57.6211111	17.1584615	58.8340992
ACS_PCT_FEMALE	3143	49.9239834	2.41135299	29.06	49.2645833	50.3098413	51.1027083	58.845	50.3098413	5.81462323
ACS_PCT_MALE	3143	50.0760197	2.41135187	41.155	48.8972917	49.6901587	50.7354167	70.94	49.6901587	5.81461783
ACS_MEDIAN_HH_INC	3142	56213.1665	15409.2408	22216	46337.8125	53869	62729.8458	148781.365	53869	237444703
ACS_TOT_CIVIL_POP_ABOVE18	3143	80219.6872	256586.843	91	8437.5	20167	52590	7856756	20167	6.58368081…
ESTABLISHMENTS_PER_POPULATION	3143	5575.61194	5776.09049	30.3333333	2320.40476	3888.66667	6918.63986	103429.75	3888.66667	33363221.4

TABLE II

Index	CANCER	DEPRESSION	DIABETES	ACS_PCT_HH_NO_INTERNET	ACS_PCT_FEMALE	ACS_PCT_MALE	ACS_MEDIAN_HH_INC	ACS_TOT_CIVIL_POP_ABOVE18	POPULATION_PER_ESTABLISHMENTS
CANCER	1	-0.02520919…	0.162085146	0.113165846	0.11530136	-0.115302211	-0.206447031	-0.261382293	-0.266694743
DEPRESSION	-0.0252…	1	0.310979805	0.22068027	0.142663641	-0.142663835	-0.375336884	-0.145295051	0.126860985
DIABETES	0.16208…	0.310979805	1	0.762561637	0.0900162835	-0.0900167447	-0.707814771	-0.177893587	-0.211594569
ACS_PCT_HH_NO_INTERNET	0.11316…	0.22068027	0.762561637	1	-0.0462244657	0.0462239272	-0.705537568	-0.245034429	-0.312922445
ACS_PCT_FEMALE	0.11530136	0.142663641	0.090016…	-0.0462244657	1	-1	0.00422759271	0.104045133	0.139897986
ACS_PCT_MALE	-0.1153…	-0.142663835	-0.09001…	0.0462239272	-1	1	-0.00422715372	-0.10404481	-0.139897587
ACS_MEDIAN_HH_INC	-0.2064…	-0.375336884	-0.70781…	-0.705537568	0.00422759271	-0.00422715372	1	0.319432678	0.241770158
ACS_TOT_CIVIL_POP_ABOVE18	-0.2613…	-0.145295051	-0.17789…	-0.245034429	0.104045133	-0.10404481	0.319432678	1	0.220150814
POPULATION_PER_ESTABLISHMENTS	-0.2666…	0.126860985	-0.21159…	-0.312922445	0.139897986	-0.139897587	0.241770158	0.220150814	1

Crosswalk of data:

Report Name	Data name imported Header Name	Description	Source
CANCER	CANCER	Average crude prevalence for Cancer (excluding skin cancer) among adults aged >=18 years	Diagnosis_State_County.csv
DEPRESSION	DEPRESSION	Average crude prevalence for Depression among adults aged >=18 years	Diagnosis_State_County.csv
DIABETES	DIABETES	Average crude prevalence for Diabetes among adults aged >=18 years	Diagnosis_State_County.csv
TOTAL POPULATION ABOVE 18 YRS	ACS_TOT_CIVIL_POP_ABOVE18	Total civilian population (ages 18 and over)	SDOH.xlsx
PERCENT FEMALE	ACS_PCT_FEMALE	Percentage of population that is female	SDOH.xlsx
PERCENT MALE	ACS_PCT_MALE	Percentage of population that is male	SDOH.xlsx
STATE	STATE	State name as listed by the US Census Bureau	SDOH.xlsx
COUNTY	COUNTY	County name as listed by the US Census Bureau	SDOH.xlsx
STATEFIPS	STATEFIPS	State ID as listed by the US Census Bureau	SDOH.xlsx
COUNTYFIPS	COUNTYFIPS	County ID as listed by the US Census Bureau	SDOH.xlsx
NO INTERNET	ACS_PCT_HH_NO_INTERNET	Percent of households without Internet access	SDOH.xlsx
MEDIAN HOUSEHOLD INCOME	ACS_MEDIAN_HH_INC	Median household income	SDOH.xlsx
Number of establishments	Number of establishments	Only used for calculating the POP. PER LIQUOR ESTABLISHMENTS, then removed for reporting	Liquor_Establishments.csv
POP. PER LIQUOR ESTABLISHMENTS	establishment_per_population	Calculation of ACS_TOT_CIVIL_POP_ABOVE18/Number of establishments	