Task 6.1 - Sourcing the Right Data

Dataset: 01 Obesity BMI Table_BRFSS				DF = obesity_bmi	
Data Summary					
Data Source	This is an external data source, owned by the Centers for Disease Control and Prevention. High level of trustworthiness.				
Data Collection	Data collected via Interview and Survey by the CDC's Behavioral Risk Factor Surveillance System. Conducted via telephone; responses are calculated based on self-reported height and weight. The table is a calculation of the BMI results based on participants height and weight response. Details are linked.				
Data Contents	The table is a calculation of the BMI results based on participants height and weight response in survey, from 2011 - 2021. This will be a primary table in analysis.				
Timeliness	Timely - the data is recent and updated yearly.				
Limitations	Survey data poses a risk of collection bias, however, this division of the CDC is experienced and dedicated to surveying the American population on health habits and conditions. This dataset also contains more information than necessary for this particular analysis, there is potential for additional data to become useful later in analysis. There are very few limitations with this dataset, it comes from a reliable source, and is relevant to the analysis.				
Data Profile					
Data Grain: Year > State > Category > Overall Rate %				Starting Shape: 58538, 27	
Variable	Description	Time-Variant/ -Invariant	Structured/ Unstructured	Qualitative/ Quantitative	Nominal/Ordinal Discrete/Cont.
Year	Data collection year	Variant	Structured	Quantitative	Discrete
State	Data collection location	Invariant	Structured	Qualitative	Nominal
Category	BMI category: Overweight = BMI 24.0-29.9 Obese = BMI 30+	Invariant	Structured	Qualitative	Ordinal
Rate_%	Percent of sample size with calculated BMI in overweight or obese category	Variant	Structured	Quantitative	Continuous
Data Quality					
Accuracy and Consistency- Descriptive Statistics					
Variable	Year	Rate_%	Sample Size		

Minimum	2011	20.2	646		
Maximum	2021	40.8	11820		
Mean	2016	32.7	2705		
Median	2016	34.1	2302		
Mode	2016	35.9	1810		
Data Integrity - Complete and Unique					
Value Count	Year - 110	Year - 110	State - 22	Category - 559	
Missing	FL missing entry from 2021, Identified adult obesity rate of FL in 2021 and input the average overweight value in FL across time. Sample_Size is average of Category sample_size.	NJ missing entry from 2019. Identified adult obesity rate of NJ in 2019 and input the average overweight value in NJ across time. Sample_Size is average of Category sample_size.	Virgin Islands only have values for two years. Records for Virgin Islands will be removed.	No missing values.	
Duplicates	No duplicates found.				

Data Wrangling	
Issue	Resolution
Column Name: Locationdesc	Changed to 'State' for location clarification.
Column Name: Data Value	Changed to 'Rate_%' for value clarification.
Column Name: Response	Changed to Category for classification clarification.
Values in 'State' column	22 entries listed a median rate for the US and US + Territories. Because we are only analyzing data for US States, the median entries and all territories have been removed from the dataset with the exception of the District of Columbia.
Values in 'Response' column	Removed the '(BMI range)', leaving just 'Overweight' and 'Obese', in order to avoid mixed-type data.
Values in 'Sample_Size' contained commas	Removed commas within values to avoid data type issues.
Missing 'Category' entries for NJ in 2019 and FL in 2021	Identified adult obesity rate for NJ in 2019 and FL in 2021, input into dataset. 'Overweight' missing entries were input with calculated average of state's 'overweight' rate across time.
20 columns dropped	Removed unnecessary data:

	Locationabbr, Class, Topic, Question, Confidence_limit_Low, Confidence_limit_High, Display_order, Data_value_unit, Data_value_type, Data_Value_Footnote_symbol, Data_Value_Footnote, DataSource, ClassId, TopicId, LocationID, BreakoutID, BreakOutCategoryID, QuestionID, ResponseID, GeoLocation	
End Shape: 1122, 5	Data cleaned using Python: GitHub Repository	df = _clean
Additional Questions: What states are associated with the highest obesity rates? Are there any other demographics associated with obesity? How do obesity rates change over time?		

Dataset: 02 Obesity Rates_2022	Shape: 52, 12
Data Summary	
Data Source	This is an external data source, owned by the Centers for Disease Control and Prevention. High level of trustworthiness. Data compiled by stateofchildhoodobesity.org and downloaded from worldpopulationreview.com. Acceptable level of trustworthiness.
Data Collection	Data collected via Interview and Survey by the CDC's Behavioral Risk Factor Surveillance System. Conducted via telephone; responses are calculated based on self-reported height and weight. The table is a calculation of the BMI results based on participants height and weight response. Details are linked.
Data Contents	The table is a calculation of the BMI results based on participants height and weight response in survey, from 2022. This will be used as supplemental data, providing 2022 state obesity rates, to be merged with primary set '01 Obesity BMI Table_BRFSS'
Timeliness	Timely - the data is recent as 2022 ended less than 30 days from analysis.
Limitations	Survey data poses a risk of collection bias, however, this division of the CDC is experienced and dedicated to surveying the American population on health habits and conditions. This dataset also contains more information than necessary for this particular analysis, there is potential for additional data to become useful later in analysis. There are very few limitations with this dataset, it comes from a reliable source, and is relevant to the analysis.

Data Profile					
Data Grain: State > Population Year > Obesity Rate %					
Variable	Description	Time-Variant/ -Invariant	Structured/ Unstructured		
State	Data collection location	Invariant	Structured		
Population 2022	Data collection year	Variant	Structured		
Obesity Rate %	Percent of adults in considered "Obese"	Invariant	Unstructured		

Data Quality				
Accuracy and Consistency- Descriptive Statistics				
Variable	Obesity Rate			
Minimum				
Maximum				
Mean				
Median				
Mode				
Data Integrity - Complete and Unique				
Value Count	State	Population 2022	Obesity Rate %	
Missing	No missing values.			
Duplicates	No duplicates found.			
Data Wrangling				
Issue		Resolution		
Dropped Columns		Dropped 9 columns of unnecessary data ranging from previous years' population, density values, and growth rates.		
Column Name: pop2022		Clarity, to match other sets.		
Column Name: obesityRate		Changed to 'Obesity Rate_%' for value clarification.		
End Shape: 52, 3		Data cleaned manually.		Data merged

Dataset: 03 CDC_COVID Deaths by State_Age		DF = covid
Data Summary		
Data Source	This is an external data source, owned by the Centers for Disease Control and Prevention. High level of trustworthiness.	
Data Collection	Administrative data collected by the National Center for Health Statistics, based on death records received from state vital offices.	
Data Contents	Deaths involving COVID-19 reported to NCHS by jurisdiction of occurrence, place of death, and age group. Technical Notes linked. This will be a primary dataset.	
Timeliness	This data is recent as of January 11th, 2023 and includes frequently updated COVID death data, beginning 01/01/2020.	

<table>
<tr><td>Limitations</td><td colspan="5">Counting the exact number of deaths related to COVID-19 is not possible due to a number of reasons; it's possible that death resulting from COVID-19 may be under-reported due to it being a voluntary system (collection bias) and patient may have not been tested for disease at time of death (exclusion bias). Death counts are suppressed if less than 9 to protect patient confidentiality. As a result, this analysis will focus on COVID-19 deaths of all ages. Age_Group '0-17' has a majority of suppressed data, so youth rates should not impact total rates as the total death count for 0-17 only accounts for 0.15% of the COVID-19 deaths in the US.</td></tr>
<tr><td colspan="6">Data Profile</td></tr>
<tr><td colspan="4">Data Grain: Group > Year > State> Place of Death > Age Group > COVID-19 Deaths</td><td colspan="2">Starting Shape: 170586,17</td></tr>
<tr><td>Variable</td><td>Description</td><td>Time-Variant/
-Invariant</td><td>Structured/
Unstructured</td><td>Qualitative/
Quantitative</td><td>Nominal/Ordinal
Discrete/Cont.</td></tr>
<tr><td><b>Group</b></td><td>Allows the user to view the data by Total (2020-present), by Year, or by Month.</td><td>Time-Invariant</td><td>Structured</td><td>Qualitative</td><td>Ordinal</td></tr>
<tr><td><b>Year</b></td><td>Year can range from 2020 - 2022.</td><td>Time-Variant</td><td>Structured</td><td>Quantitative</td><td>Discrete</td></tr>
<tr><td><b>State</b></td><td>State counting the deaths.</td><td>Time-Invariant</td><td>Structured</td><td>Qualitative</td><td>Nominal</td></tr>
<tr><td><b>Place of Death</b></td><td>Allows the user to group data by place of death; ex) hospital</td><td>Time-Invariant</td><td>Structured</td><td>Qualitative</td><td>Nominal</td></tr>
<tr><td><b>Age Group</b></td><td>Allows the user to group data by Age group, beginning with 0-17 and every 10 years thereafter until 85+.</td><td>Time-Invariant</td><td>Structured</td><td>Qualitative</td><td>Ordinal</td></tr>
<tr><td><b>COVID-19 Deaths</b></td><td>Total cumulative count of death based on filtered variables.</td><td>Time-Variant</td><td>Unstructured</td><td>Quantitative</td><td>Discrete</td></tr>
<tr><td colspan="6">Data Quality</td></tr>
<tr><td colspan="6">Accuracy and Consistency- Descriptive Statistics</td></tr>
<tr><td>Type</td><td>Year</td><td>State</td><td>Place of Death</td><td>Age Group</td><td>COVID-19</td></tr>
</table>

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td>Deaths</td></tr>
<tr><td>Minimum</td><td>2020</td><td></td><td></td><td></td><td>145</td></tr>
<tr><td>Maximum</td><td>2022</td><td></td><td></td><td></td><td>463195</td></tr>
<tr><td>Mean</td><td>2021</td><td></td><td></td><td></td><td>13512</td></tr>
<tr><td>Median</td><td>2021</td><td></td><td></td><td></td><td>4694</td></tr>
<tr><td>Mode</td><td>Equal</td><td></td><td></td><td></td><td>308</td></tr>
<tr><td colspan="6">Data Integrity - Complete and Unique</td></tr>
<tr><td>Value Count</td><td>52</td><td>3159</td><td>18252</td><td>18252</td><td></td></tr>
<tr><td>Missing</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>Duplicates</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td colspan="6">Data Wrangling</td></tr>
<tr><td colspan="2">Column Update</td><td colspan="4">Resolution</td></tr>
<tr><td colspan="2">Drop Columns</td><td colspan="4">Dropped due to data being unnecessary for analysis: Data as of, Start Date, HHS Region, Month, Total Deaths, Pneumonia Deaths, Pneumonia and COVID-19 Deaths, Influenza Deaths, Footnote</td></tr>
<tr><td colspan="2">Column 'Group'</td><td colspan="4">Filtered data to include only 'By Year' as the analysis requires deaths over time. Renamed to 'Group By' for clarity.</td></tr>
<tr><td colspan="2">Column 'State'</td><td colspan="4">Filtered data to remove US Territories since this analysis is interested in US state deaths only.</td></tr>
<tr><td colspan="2">Column 'Place of Death'</td><td colspan="4">Filtered data to include only 'Total - All Place of Death' since this analysis is interested in all deaths, regardless of place.</td></tr>
<tr><td colspan="2">Column 'Age Group'</td><td colspan="4">Filtered to include 'All Ages' since this analysis is not looking specifically at age groups. Many of the data points have been suppressed for confidentiality. This may come in useful later.</td></tr>
<tr><td colspan="2">Drop Column 'Group By'</td><td colspan="4">No longer needed, analyzing yearly counts.</td></tr>
<tr><td colspan="2">Drop Column 'Place of Death'</td><td colspan="4">No longer needed, data has been filtered to include only this information. Entries are the same throughout the new df.</td></tr>
<tr><td colspan="2">Drop Column 'Age Group'</td><td colspan="4">No longer needed, data has been filtered to include only this information. Entries are the same throughout the new df.</td></tr>
<tr><td colspan="2">End Shape: 153, 4</td><td colspan="2">Data cleaned using Python
GitHub Repository</td><td colspan="2">df = _clean (unfiltered)
df = _FILTERED (basic)</td></tr>
<tr><td colspan="6">Additional Questions:
What states saw the highest COVID-19 death rates?
How have COVID-19 death rates changed over time?</td></tr>
</table>

Is there a relationship between obesity rates and COVID-19 death rates?

<table>
<tr><td colspan="4">Dataset: 04 CENSUS_2010_2019 Population</td><td colspan="2">DF = population</td></tr>
<tr><td colspan="6">Data Summary</td></tr>
<tr><td>Data Source</td><td colspan="5">This is an external data source, owned by the US Census Bureau. High level of trustworthiness.</td></tr>
<tr><td>Data Collection</td><td colspan="5">Data is collected through survey and administrative data collection. Estimates are calculated from base population + births - deaths + immigration = estimated population.</td></tr>
<tr><td>Data Contents</td><td colspan="5">2010 census data and estimated population count by state and age group, along with several other demographics. This data may be used to normalize obesity rates and COVID-19 death rates by state and age group.</td></tr>
<tr><td>Timeliness</td><td colspan="5">US Census data is collected every 10 years. This dataset lists the 2010 Census results for each state, and population estimates for each state through 2019.</td></tr>
<tr><td>Limitations</td><td colspan="5">Because census data has decades worth of population data, projections can be considered relatively reliable, regardless of an interruption in collection. This data will be merged with 2020 Census data to show population over time.</td></tr>
<tr><td colspan="6">Data Profile</td></tr>
<tr><td colspan="4">Data Grain: State > 2010 Population</td><td colspan="2">Starting Shape: 58, 151</td></tr>
<tr><td>Variable</td><td>Description</td><td>Time-Variant/ -Invariant</td><td>Structured/ Unstructured</td><td>Qualitative/ Quantitative</td><td></td></tr>
<tr><td><b>State</b></td><td>State of data collection</td><td>Time-Invariant</td><td>Structured</td><td>Qualitative</td><td></td></tr>
<tr><td><b>2010 Population</b></td><td>2010 total population based on state</td><td>Time-Variant</td><td>Unstructured</td><td>Quantitative</td><td></td></tr>
<tr><td><b>2011 Population - 2019 Population</b></td><td>State population estimate based on 2010 census, with births, deaths, and immigration taken into account</td><td>Time-Variant</td><td>Unstructured</td><td>Quantitative</td><td></td></tr>
<tr><td colspan="6">Data Quality</td></tr>
<tr><td colspan="6">Accuracy and Consistency- Descriptive Statistics</td></tr>
<tr><td>Type</td><td>State</td><td>2010 Population</td><td>2011 Population</td><td>2019 Population</td><td rowspan="2"></td></tr>
<tr><td>Minimum</td><td></td><td>563626</td><td>567299</td><td>578759</td></tr>
</table>

Maximum		37253956	37638369	39512223	
Mean		6053834	6108958	6436069	
Data Integrity - Complete and Unique					
Value Count	51	51	51	51	
Missing	0	0	0	0	
Duplicates	0	0	0	0	
Data Wrangling					
Issue		Resolution			
Drop unnecessary columns		Columns remaining: Name, Census 2010 Pop, Pop Estimate 2011-2019			
Renamed Columns		Name : State, Census 2010 Pop : 2010 Pop			
Filtered State column		Removed rows with data broken out by region instead of state: 'midwest region', 'northwest region', 'west region', 'south region', United States totals, and Puerto Rico.			
Population columns data type		Removed commas from population estimates to avoid data type errors in Jupyter.			
End Shape: 52, 11		Data cleaned using Excel, Descriptive Statistics and Frequency Counts using Python		df = population_clean	

.

Dataset: 05 CENSUS_2020_2022 Population	DF = census
Data Summary	
Data Source	This is an external data source, owned by the US Census Bureau. High level of trustworthiness.
Data Collection	Due to the COVID-19 pandemic, this data is a projection of previously collected data for 2020 and 2021 results. Previous data collected by way of the American Community Survey. This data may be collected directly from respondents via survey. Additional administrative data is collected from federal, state, and local governments, and even some commercial entities.
Data Contents	Estimate population for 2020 through 2022 count by state and age group, along with several other demographics. This data may be used to normalize obesity rates and COVID-19 death rates by state and age group.
Timeliness	US Census data is collected every 10 years, this dataset is recent as of 2022
Limitations	Because census data has decades worth of population data, projections can be considered relatively reliable, regardless of an interruption in collection.
Data Profile	

Data Grain: State > Estimated > Total Population				Starting Shape: 96, 213	
Variable	Description	Time-Variant/ -Invariant	Structured/ Unstructured	Qualitative/ Quantitative	Nominal/Ordinal Discrete/Cont.
State	State of data collection	Time-Invariant	Structured	Qualitative	Nominal
Total Population	2020 Estimate of states total population	Time-Variant	Unstructured	Quantitative	Discrete
Youth Population	2020 Estimate of states youth population (0-17)	Time-Variant	Unstructured	Quantitative	Discrete
Adult Population	2020 Estimate of states adult population (18+)	Time-Variant	Unstructured	Quantitative	Discrete
Data Quality					
Accuracy and Consistency- Descriptive Statistics					
Type	State	Total Population	Youth Population	Adult Population	
Minimum		58134	115632	446548	
Maximum		39346023	8956641	30389382	
Mean		6403320	1437191	4966129	
Data Integrity - Complete and Unique					
Value Count	52	52	52	52	
Missing	0	0	0	0	
Duplicates	0	0	0	0	
Data Wrangling					
Issue		Resolution			
Transpose data		The top row of this dataset was divided by State with columns of each state's population estimate based on corresponding row conditions. Data transposed so all unique state population data can be found within a single row.			
Dropped columns corresponding with each state		Each state lists an estimate, margin of error, percent, and percent margin of error. This information would clog the data. Confirmed all estimates are within +/-0.1%.			
Dropped fine grain population data columns		These columns included breakdowns of gender, 10 year age groups, race, and voting statistics. Remaining columns include State, Total Population, Youth Population, and Adult Population.			

Population columns data type	Removed commas from population estimates to avoid data type errors in Jupyter.	
Renamed Columns	Total Population : 2020 Population	
Drop Columns	Youth Population and Adult population will be deleted as the analysis will not account for age groups.	
Column: State > Selected Values	Dropped all entries that were not a US state or DC.	
End Shape: 52, 3	Data cleaned using Excel, Descriptive Statistics and Frequency Counts using Python	df = census_clean

					Nominal/Ordinal Discrete/Cont.
					Nominal
					Discrete
					Discrete