Data shown on this project has been downloaded from: https://catalog.data.gov/dataset/?tags=obesity.

The dataset is listed as: Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System.

The beginning of the dataset involves cleaning the data. The second portion of the dataset includes visualizations and report analysises.

Prior to analyzing the data, I want to look at the data to see what the values and columns look like.

Unknown integration DataFrame as Obesity_risk_factor

```
-- Explore the data in the table
SELECT *
FROM 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv'
LIMIT 10;
```

	Y.		Loca...	Loca...	Class	Topic	Question
0	2011	2011	AL	Alabama	Obesity / Weight Status	Obesity / Weight Status	Percent of adults
1	2011	2011	AL	Alabama	Obesity / Weight Status	Obesity / Weight Status	Percent of adults
2	2011	2011	AL	Alabama	Obesity / Weight Status	Obesity / Weight Status	Percent of adults
3	2011	2011	AL	Alabama	Obesity / Weight Status	Obesity / Weight Status	Percent of adults
4	2011	2011	AL	Alabama	Obesity / Weight Status	Obesity / Weight Status	Percent of adults
5	2011	2011	AL	Alabama	Obesity / Weight Status	Obesity / Weight Status	Percent of adults
6	2011	2011	AL	Alabama	Obesity / Weight Status	Obesity / Weight Status	Percent of adults
7	2011	2011	AL	Alabama	Obesity / Weight Status	Obesity / Weight Status	Percent of adults
8	2011	2011	AL	Alabama	Obesity / Weight Status	Obesity / Weight Status	Percent of adults
9	2011	2011	AL	Alabama	Obesity / Weight Status	Obesity / Weight Status	Percent of adults

Rows: 10 Expand

Determine the range of years for the dataset. This is important to determine how recent the data is. Newer data is more likely to relate to the current health status of the nation. This data can later be used to determine if the company wants to look at data from specific years that may be more relevant to the current population.

Unknown integration DataFrame as Obesity_risk_factor

```
--Determine the range of years for the dataset. This is important to determine how recent the data is. Newer data is more likely to
relate to the current health status of the nation.
SELECT YearStart, YearEnd
FROM 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv'
GROUP BY YearStart, YearEnd
ORDER BY YearStart DESC;
```

index	YearStart	YearEnd
0	2023	
1	2022	
2	2021	
3	2020	
4	2019	
5	2018	
6	2017	
7	2016	
8	2015	
9	2014	
10	2013	
11	2012	
12	2011	

Rows: 13 Expand

In sql I would drop columns from the table to alter the table to reflect data the is redundant or irrelevant to the data. However, this dataframe is not comprehending the code and instead the columns will be deleted in excel and transferred as a csv file.

-- Delete the redundant columns: ALTER TABLE 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv' -- Datasource as each value is Behavioral Risk Factor Surveillance System DROP COLUMN Datasource;

ALTER TABLE 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv' -- Datasource as each value is Behavioral Risk Factor Surveillance System DROP COLUMN LocationDesc;

ALTER TABLE 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv' -- Datasource as each value is Behavioral Risk Factor Surveillance System DROP COLUMN DataValue_Alt;

ALTER TABLE 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv' -- Data_Value_Unit as each cell is blank DROP COLUMN Data_Value_Unit;

ALTER TABLE 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv' -- Data_Value_Type as each value is Value DROP COLUMN Data_Value_Type;

ALTER TABLE 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv' -- Data_Value_footnote_Symbol as the cells either are blank or contains "-" DROP COLUMN Data_Value_Footnote_Symbol;

ALTER TABLE 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv' -- Total as each cell is either blank or contains "Total" DROP COLUMN Total;

ALTER TABLE 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv' -- Geolocation as the state is already set as a location DROP COLUMN GeoLocation;

ALTER TABLE 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv' -- DataValueTypeID as each cell contains "VALUE" DROP COLUMN DataValueTypeID;

ALTER TABLE 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv' -- Topic is a repeat category of Class DROP COLUMN Topic;

I would also rename the table to make the data easier to access but that will also have to be done from excel in this project.

ALTER TABLE 'Nutrition__Physical_Activity__and_Obesity_-_Behavioral_Risk_Factor_Surveillance_System.csv' RENAME TO Obesity_risk_factor;

Delete records that contain "Data not available because sample size is insufficient or data not reported." in Data_Value_Footnote. This indicates that the the data size is too small and should be considered an outlier. There are 12,755 records. Total records is 93, 506.

DELETE FROM main."Obesity_risk_factor.csv" WHERE Data_Value_Footnote = 'Data not available because sample size is insufficient or data not reported.';

Once this data has been deleted, I would drop the column Data_value_footnote as it is now full of blanks (null values). ALTER TABLE 'Obesity_risk_factor' DROP COLUMN Data_Value_Footnote;

Now that the data has been reviewed and reduced to include data that is more relevant and condensed for my research. I can now take a better look into the data to answer questions regarding the data.

Unknown integration DataFrame as d

```
-- Now that the data has been reviewed and shortened to include data that is more relevant and condensed for our research. We can
now take a better look into the data.

SELECT *
FROM Obesity_risk_factor.csv
LIMIT 10;
```

	Y.		Loca...	Class	Question	D...	Low_Conf
0	2011	2011	AL	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	34.8	
1	2011	2011	AL	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	35.8	
2	2011	2011	AL	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	32.3	
3	2011	2011	AL	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	34.1	
4	2011	2011	AL	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	28.8	
5	2011	2011	AL	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	16.3	
6	2011	2011	AL	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	27.8	
7	2011	2011	AL	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	35.2	
8	2011	2011	AL	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	35.5	
9	2011	2011	AL	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	38	

Rows: 10 Expand

I want to determine the different values in Class since this is mainy how the data is divided into categories.

Unknown integration DataFrame as

```
-- Let's take a look at the different values in Class are since this is mainy how the group is divided into categories.

SELECT Class AS Main_categories,
    COUNT(Class) AS participents_per_main_category
FROM Obesity_risk_factor.csv
GROUP BY Class;
```

	Main_categories	participents_per_main_category
0	Obesity / Weight Status	35388
1	Fruits and Vegetables	8155
2	Physical Activity	49962

Rows: 3 Expand

After seeing the three different types of main categories, I can now see how the participants were grouped based on the question regarding their health status.

Unknown integration DataFrame as

```
-- Now let's review the different types of questions.
SELECT Question AS question_category,
    COUNT(Question) AS participents_per_question
FROM Obesity_risk_factor.csv
GROUP BY Question;
```

	question_category	participents_per_question
0	Percent of adults who engage in muscle-stre...	8078
1	Percent of adults who achieve at least 150 m...	8052
2	Percent of adults who achieve more than 30...	8042
3	Percent of adults aged 18 years and older wh...	17694
4	Percent of adults who engage in no leisure-ti...	17748
5	Percent of adults who report consuming veg...	4075
6	Percent of adults aged 18 years and older wh...	17694
7	Percent of adults who report consuming fruit...	4080
8	Percent of adults who achieve at least 150 m...	8042

Rows: 9 Expand

Now I can see there are 3 main categories 1. Obesity / Weight Status, 2. Fruits and Vegetables, 3. Physical Activity.

I can also see that there are 8 question categories.

Let's see if there are questions that related to specific main categories.

Unknown integration DataFrame as

```
-- Now we can see there are 3 main categories 1. Obesity / Weight Status, 2. Fruits and Vegetables, 3. Physical Activity.
-- We can also see that there are 8 question categories. Let's see if there are questions that related to specific main categories.

SELECT Class AS main_category,
    Question AS question_category,
    COUNT(Question)
FROM Obesity_risk_factor.csv
GROUP BY Class, Question
ORDER BY Class;
```

	main_category	question_category	count(Que...
0	Fruits and Vegetables	Percent of adults who report consuming fruit...	4080
1	Fruits and Vegetables	Percent of adults who report consuming veg...	4075
2	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	17694
3	Obesity / Weight Status	Percent of adults aged 18 years and older wh...	17694
4	Physical Activity	Percent of adults who achieve more than 30...	8042
5	Physical Activity	Percent of adults who engage in muscle-stre...	8078
6	Physical Activity	Percent of adults who achieve at least 150 m...	8042
7	Physical Activity	Percent of adults who engage in no leisure-ti...	17748
8	Physical Activity	Percent of adults who achieve at least 150 m...	8052

Rows: 9 Expand

Let's find the top three states for any participant in the main category of Fruits and Vegetables.

Note In the data, US is listed as a Location. Due to this unknown location within the data, this inquiry will be left out of the query. This is because it is unsure if the location US is set to be indicated as territores that are included in the USA like Puerto Rico and the United Stated Virgin Islands or is mislabeled. Due to this, further investigation would be required if the company wanted US to be included in the location category since it is considered the top location for each of these next location questions.

Unknown integration DataFrame as

```
-- Let's find the top three states for any participant in the main category of Fruits and Vegetables or Obesity / Weight Status,
plus those labeled as engaging in no leisure-time physical activity.

SELECT LocationAbbr as State,
    COUNT(Class) AS Less_Than_1_Daily_Fruits_or_Vegetables
FROM Obesity_risk_factor.csv
WHERE Class = 'Fruits and Vegetables' AND LocationAbbr != 'US'
GROUP BY LocationAbbr
ORDER BY Less_Than_1_Daily_Fruits_or_Vegetables DESC
LIMIT 3;
```

••• ↑↓	••• ↑↓	Less_Than_1_Daily_Fruits_or_Veget... ••• ↑↓
0	WA	164
1	MD	162
2	AZ	162

Rows: 3 Expand

Let's find the top three states for any participant in the main category of Obesity / Weight Status.

Unknown integration DataFrame as

```
SELECT LocationAbbr AS State,
    COUNT(Class) AS Weight_Status_as_Obese_or_Overweight
FROM Obesity_risk_factor.csv
WHERE Class = 'Obesity / Weight Status' AND LocationAbbr != 'US'
GROUP BY LocationAbbr
ORDER BY Weight_Status_as_Obese_or_Overweight DESC
LIMIT 3;
```

••• ↑↓	••• ↑↓	Weight_Status_as_Obese_or_Overw... ••• ↑↓
0	WA	714
1	MD	698
2	CO	696

Rows: 3 Expand

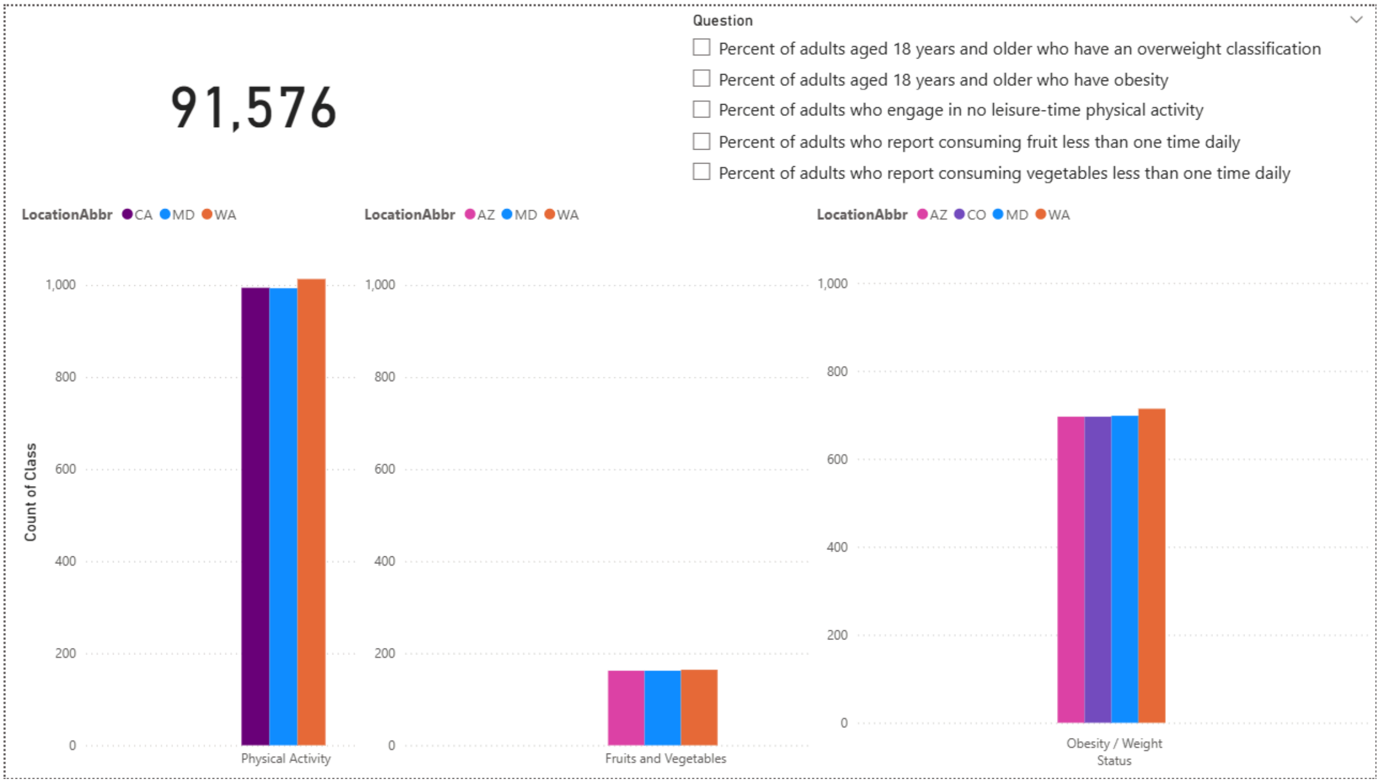
Let's find the top three states for any participant in the main category of Physical Activity that are labeled as engaging in no leisure-time physical activity.

Unknown integration DataFrame as

```
SELECT LocationAbbr AS State,
    COUNT(Question) AS No_Activity
FROM Obesity_risk_factor.csv
WHERE Question = 'Percent of adults who engage in no leisure-time physical activity' AND LocationAbbr != 'US'
GROUP BY LocationAbbr
ORDER BY No_Activity DESC
LIMIT 3;
```

••• ↑↓	••• ↑↓	No... ••• ↑↓
0	WA	357
1	MD	349
2	CA	349

Rows: 3 Expand

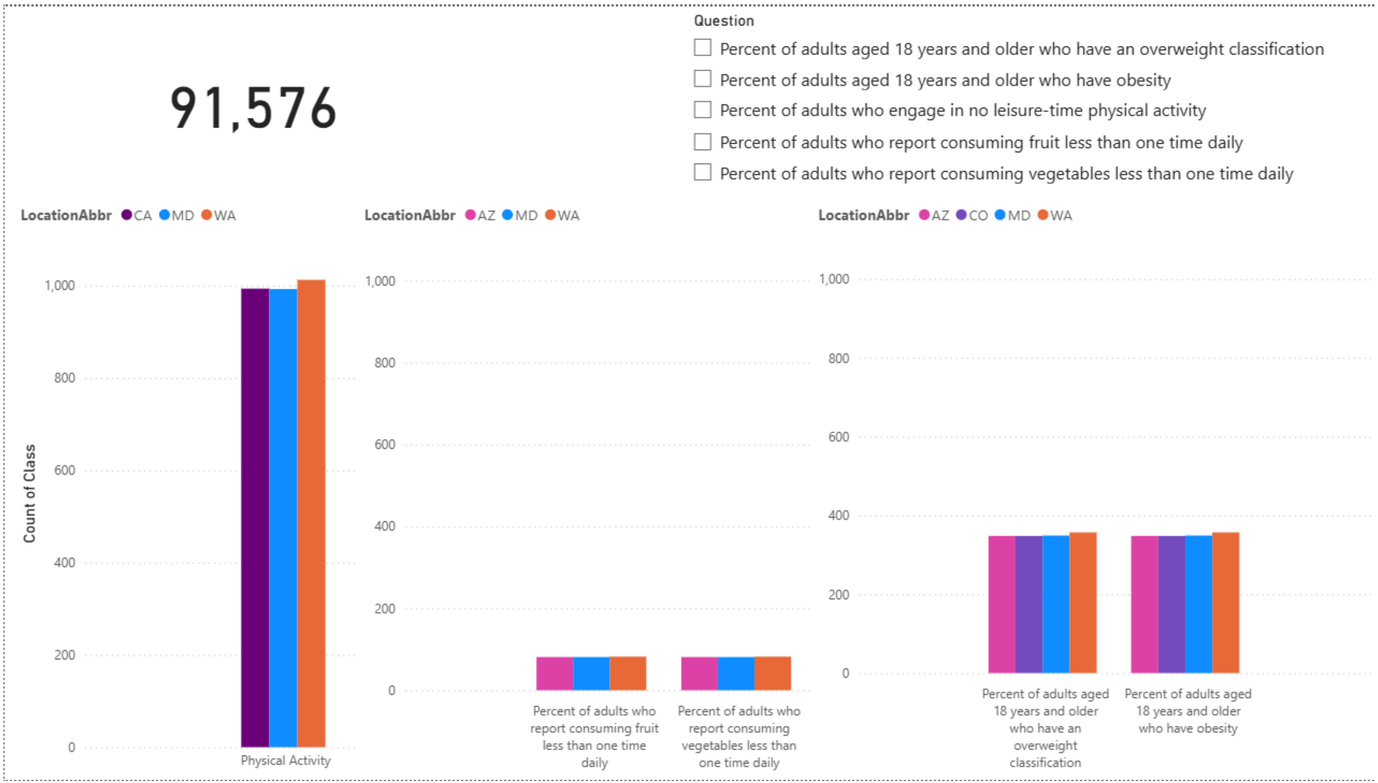


91,576

Question

□ Percent of adults aged 18 years and older who have an overweight classification
□ Percent of adults aged 18 years and older who have obesity
□ Percent of adults who engage in no leisure-time physical activity
□ Percent of adults who report consuming fruit less than one time daily
□ Percent of adults who report consuming vegetables less than one time daily

The graphic indicates the count for the top three states with the three health risk categories. Each state can be clicked on in the graphics to change the count on the card. The slicer can be used to visualize one specific question at a time.



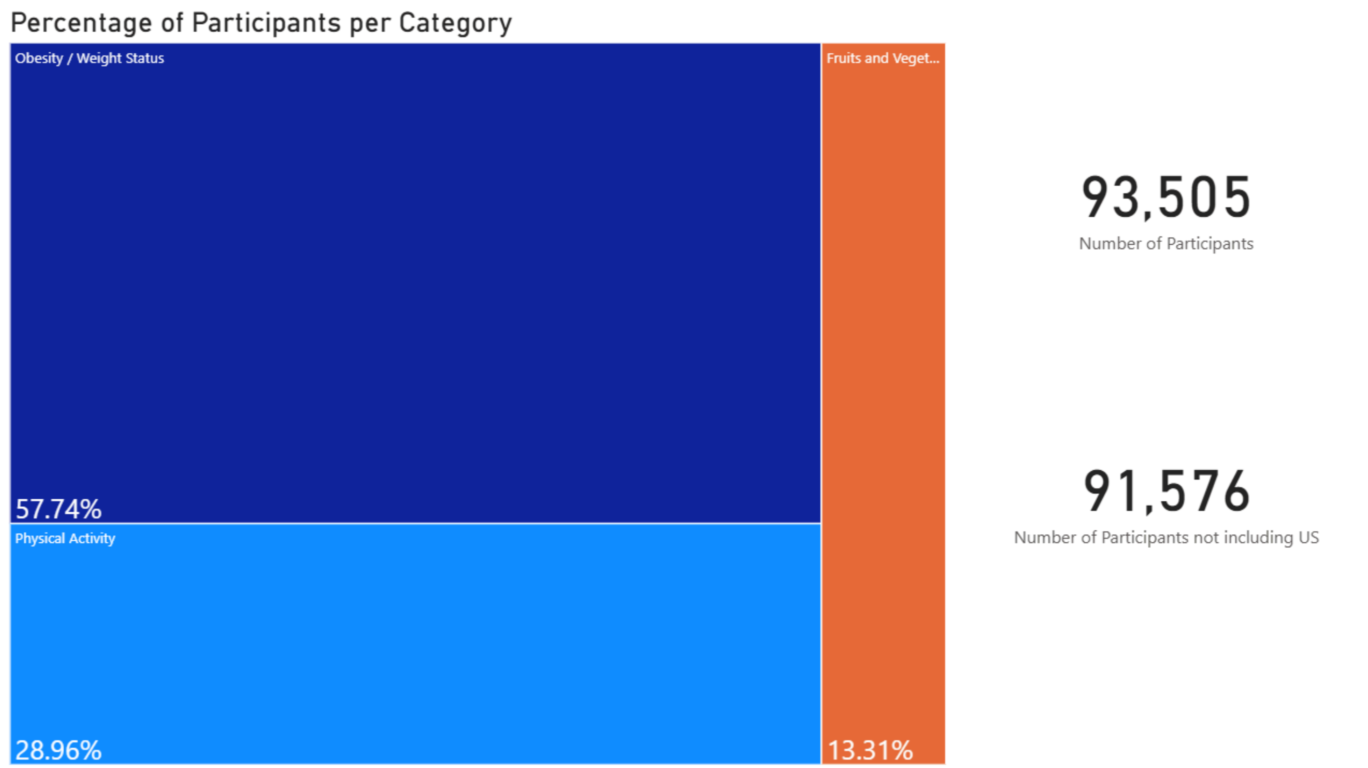
91,576

Question

□ Percent of adults aged 18 years and older who have an overweight classification
□ Percent of adults aged 18 years and older who have obesity
□ Percent of adults who engage in no leisure-time physical activity
□ Percent of adults who report consuming fruit less than one time daily
□ Percent of adults who report consuming vegetables less than one time daily

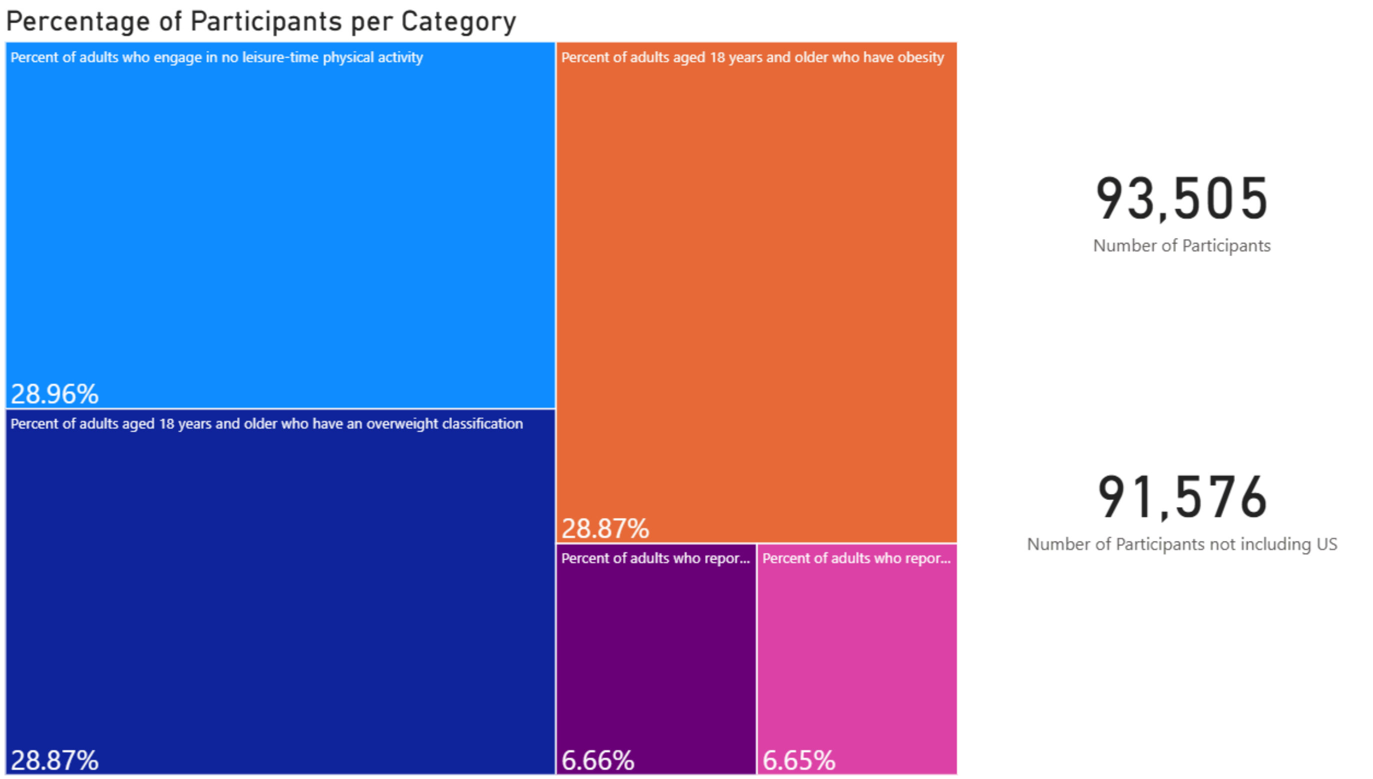
LocationAbbr ● CA ● MD ● WA
Count of Class
1,000
800
600
400
200
0
Physical Activity
LocationAbbr ● AZ ● MD ● WA
1,000
800
600
400
200
0
Percent of adults who report consuming fruit less than one time daily
Percent of adults who report consuming vegetables less than one time daily
LocationAbbr ● AZ ● CO ● MD ● WA
1,000
800
600
400
200
0
Percent of adults aged 18 years and older who have an overweight classification
Percent of adults aged 18 years and older who have obesity

The graphic indicates the top three states for each health risk question. This data indicates that these states see the highest risk for individuals for each of those categories. As you can see Maryland and Washington is included in each health risk category. This may pose the question if more health incentives/ advertisement should be completed in those states to gain a healthy population based on those health risk categories/ questions.



The graphic shows the percentage of risk factor based on the three categories. As shown, individuals that are classified as overweight or obese are of greatest concern. The card count can be used to select individua categories to determine the count for the specific category.

The two cards are included because the data includes US as a location. Due to this, an additional card was created as it is undetermined at this time if the location US is meant as US territories or if the data was mislabled and should not be included.



The graphic shows the percentage for each question that poses the greatest health risks for each category. The cards can be used to show the count for specific questions.

Now that we have seen some information regarding which states are at the highest risk factor for those health categories (1. Obesity / Weight Status, 2. Fruits and Vegetables, 3. Physical Activity.) We can take a look at the top healthiest states.

Unknown integration DataFrame as

```
SELECT LocationAbbr as State,
    COUNT(Class) AS Top_states_consuming_fruits_or_vegetables
FROM Obesity_risk_factor.csv
WHERE Class = 'Fruits and Vegetables' AND LocationAbbr != 'US'
GROUP BY LocationAbbr
ORDER BY Top_states_consuming_fruits_or_vegetables
LIMIT 3;
```

••• ↑↓	••• ↑↓	Top_states_consuming_fruits_or_ve... ••• ↑↓
0	VI	46
1	NJ	106
2	FL	108

Rows: 3 Expand

Unknown integration DataFrame as

```
SELECT LocationAbbr AS State,
    COUNT(Class) AS Top_states_not_labeled_overweight_or_obese
FROM Obesity_risk_factor.csv
WHERE Class = 'Obesity / Weight Status' AND LocationAbbr != 'US'
GROUP BY LocationAbbr
ORDER BY Top_states_not_labeled_overweight_or_obese
LIMIT 3;
```

••• ↑↓	••• ↑↓	Top_states_not_labeled_overweight... ••• ↑↓
0	VI	186
1	PR	462
2	GU	500

Rows: 3 Expand

For the category of physical fitness, to find the top three states, the data is going to contain information from the questions:

1. Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic physical activity (or an equivalent combination) and engage in muscle-strengthening activities on 2 or more days a week
2. Percent of adults who achieve more than 300 minutes a week of moderate-intensity aerobic physical activity or 150 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)

Unknown integration DataFrame as

```
SELECT LocationAbbr AS State,
    COUNT(Question) AS Top_states_for_minutes_of_physical_activity
FROM Obesity_risk_factor.csv
WHERE Question = 'Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic physical activity (or an equivalent combination) and engage in muscle-strengthening activities on 2 or more days a week' OR Question = 'Percent of adults who achieve more than 300 minutes a week of moderate-intensity aerobic physical activity or 150 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)'
GROUP BY LocationAbbr
ORDER BY Top_states_for_minutes_of_physical_activity
LIMIT 3;
```

••• ↑↓	••• ↑↓	Top_states_for_minutes_of_physical... ••• ↑↓
0	VI	47
1	PR	168
2	GU	200

Rows: 3 Expand

Based on the data, the top "healthiest" states based on the data are actually United States territories. The exception is that Florida and New Jersey are within the top three states for consuming the most fruits or vegeatbles. To see the top states, the data will need to exclude the US territories.

Unknown integration DataFrame as

```sql
SELECT LocationAbbr as State,
    COUNT(Class) AS Top_states_consuming_fruits_or_vegetables
FROM Obesity_risk_factor.csv
WHERE Class = 'Fruits and Vegetables' AND LocationAbbr != 'VI' AND LocationAbbr != 'PR' AND LocationAbbr != 'GU'
GROUP BY LocationAbbr
ORDER BY Top_states_consuming_fruits_or_vegetables
LIMIT 3;
```

...	...	Top_states_consuming_fruits_or_ve...
0	NJ	106
1	FL	108
2	MS	136

Rows: 3 Expand

Unknown integration DataFrame as

```sql
SELECT LocationAbbr AS State,
    COUNT(Class) AS Top_states_not_labeled_overweight_or_obese
FROM Obesity_risk_factor.csv
WHERE Class = 'Obesity / Weight Status' AND LocationAbbr != 'VI' AND LocationAbbr != 'PR' AND LocationAbbr != 'GU'
GROUP BY LocationAbbr
ORDER BY Top_states_not_labeled_overweight_or_obese
LIMIT 3;
```

...	...	Top_states_not_labeled_overweight...
0	MS	590
1	KY	594
2	PA	602

Rows: 3 Expand

Unknown integration DataFrame as

```sql
SELECT LocationAbbr AS State,
    COUNT(Question) AS Top_states_for_minutes_of_physical_activity
FROM Obesity_risk_factor.csv
WHERE Question IN
    ('Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a
week of vigorous-intensity aerobic physical activity (or an equivalent combination) and engage in muscle-strengthening activities
on 2 or more days a week', 'Percent of adults who achieve more than 300 minutes a week of moderate-intensity aerobic physical
activity or 150 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)')
    AND LocationAbbr != 'VI' AND LocationAbbr != 'PR' AND LocationAbbr != 'GU'
GROUP BY LocationAbbr
ORDER BY Top_states_for_minutes_of_physical_activity
LIMIT 3;
```

...	...	Top_states_for_minutes_of_physical...
0	PA	250
1	KY	250
2	NJ	264

Rows: 3 Expand

Seeing which states are considered the top healthiest and unhealthiest based on those categories is important. This information can later be used to see if there are reasons for why those states are in the top or bottom. The reasons could vary based on climate and top producers (food wise), number of fitness facilities, what types of fitness/ produce advertisements are made available, etc.

Since the top and bottom states have been recognized, let's take a look at the age categories given in the dataset. We will also need to look at how many participants did not include their age. If there are a numerous amount of participants that did not include their age, then age may not be a factor needed in the demographic information.

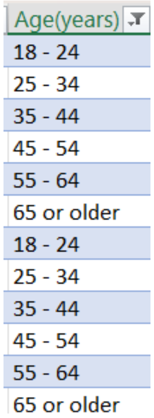
Unknown integration DataFrame as

```
SELECT "Age(years)", COUNT(*) AS Age_Count
FROM Obesity_risk_factor.csv
GROUP BY "Age(years)"
ORDER BY "Age(years)";
```

	Ag...	A
0	18 - 24	3684
1	25 - 34	3684
2	35 - 44	3684
3	45 - 54	3684
4	55 - 64	3684
5	65 or older	3684
6	null	71401

Rows: 7 Expand

Based on the data having 3684 records for each age category and 71,401 null records, the data is deemed inconclusive. This is because the data diplsays the categories continuouslu throughout the data inbetween blank values (table excludes blank values to display how the data is displayed.



Age(years)
18 - 24
25 - 34
35 - 44
45 - 54
55 - 64
65 or older
18 - 24
25 - 34
35 - 44
45 - 54
55 - 64
65 or older