Final Project

SOCIOECONOMIC FACTORS AND OBESITY

A US NATION/STATE-LEVEL ANALYSIS

Elaine da Silva (w0445885) & Tiankun Deng (w0453048)

DBAS3018 | November, 2024

Contents

[Project Overview 3](#_Toc184321252)

[Datasets Overview 4](#_Toc184321253)

[Nutrition, Physical Activity, and Obesity Dataset 4](#_Toc184321254)

[Yellow Taxi Trip Dataset 4](#_Toc184321255)

[Activity 1 ERD and Data Dictionary 5](#_Toc184321256)

[Item 1 – ERD Image 5](#_Toc184321257)

[Item 2 – Data Dictionary Image 6](#_Toc184321258)

[Activity 2 Data Wrangling 7](#_Toc184321259)

[Item 1 Exploring data wrangling 8](#_Toc184321260)

[Part A – Most used columns 8](#_Toc184321261)

[Part B - Dealing with missing values 8](#_Toc184321262)

[Activity 3 Exploring NumPy and Pandas 10](#_Toc184321263)

[Item 1 – Function, Loop, Conditionals, Read from & Write to CSV files 10](#_Toc184321264)

[PART A –Read from CSV files 10](#_Toc184321265)

[PART B –Write to CSV files 10](#_Toc184321266)

[PART C – Conditionals 10](#_Toc184321267)

[PART D – Loops 11](#_Toc184321268)

[Item 2 – Exploring 2+ plots (visualization) 11](#_Toc184321269)

[PART A - Visualization 1 11](#_Toc184321270)

[PART B - Visualization 2 12](#_Toc184321271)

[Item 3 – Exploring descriptive statistics 12](#_Toc184321272)

[Item 4 – Exploring Numpy Array and 2+ related functions 12](#_Toc184321273)

[Item 5 – Exploring Pandas DataFrame and 2+ related functions 13](#_Toc184321274)

[Activity 4 Time & Memory comparison Pandas vs Spark 14](#_Toc184321275)

[Item 1 – Datasets information 14](#_Toc184321276)

[Item 2 – Small dataset statistics 14](#_Toc184321277)

[Item 3 – Big dataset statistics 14](#_Toc184321278)

[Item 4 – Results analysis 15](#_Toc184321279)

[Item 5 – How to decide between Pandas & Spark 15](#_Toc184321280)

[Activity 5 Conclusion (findings) 16](#_Toc184321281)

[Activity 6 References 17](#_Toc184321282)

[Figure 1 – ERD implementation 5](#_Toc184321283)

[Figure 2 - Data Dictionary 6](#_Toc184321284)

[Figure 3 - “Nutrition\_Physical\_Activity\_and\_Obesity.csv” dataset 7](#_Toc184321285)

[Figure 4 - Dataset filtered by "Gender" category 8](#_Toc184321286)

[Figure 5 – dealing with missing data (Nulls) 9](#_Toc184321287)

[Figure 6 - dealing with missing data (NA) 9](#_Toc184321288)

[Figure 7 - Read from CSV 10](#_Toc184321289)

[Figure 8 - Write to CSV 10](#_Toc184321290)

[Figure 9 - Conditionals 10](#_Toc184321291)

[Figure 10 - Loops 11](#_Toc184321292)

[Figure 11 - Obesity rate over the years 11](#_Toc184321293)

[Figure 12 - Obesity rate by Education level 12](#_Toc184321294)

[Figure 13 - Descriptive statistics 12](#_Toc184321295)

[Figure 14 - Numpy arrays and functions 12](#_Toc184321296)

[Figure 15 - Pandas dataframe and functions 13](#_Toc184321297)

[Figure 16 - Small and Big dataset information 14](#_Toc184321298)

# Project Overview

This project focuses on conducting an exploratory analysis of the **"Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System"** dataset, titled **“Nutrition\_Physical\_Activity\_and\_Obesity.csv”**, available on [DATA.GOV](https://www.data.gov). The analysis aims to uncover meaningful insights from the dataset while demonstrating proficiency in Python programming, emphasizing Object-Oriented Programming (OOP) concepts and the use of key libraries, such as **NumPy**, **Pandas**, and **Spark**.

Additionally, a larger dataset, **“yellow\_tripdata\_2015-01.csv”**, available on [KAGGLE](https://www.kaggle.com/datasets/elemento/nyc-yellow-taxi-trip-data), will be utilized to compare time and memory processing performance between Pandas and Spark.

The project addresses five significant research questions, leveraging the following programming features:

* **Python**: Implementation of functions, loops, and conditionals.
* **Pandas**: Reading and writing CSV files, handling and manipulating DataFrames and Series, and utilizing associated functions.
* **NumPy**: Managing and manipulating arrays and applying relevant functions.
* **Data Wrangling**: Managing missing data effectively.
* **Visualization**: Creating and interpreting various types of plots.
* **Performance Monitoring**: Evaluating time and memory usage of Pandas and Apache Spark during data processing.

# Datasets Overview

## Nutrition, Physical Activity, and Obesity Dataset

The “**Nutrition\_Physical\_Activity\_and\_Obesity.csv**” dataset comprises **33 columns** and **104,272 rows** and is based on a survey conducted among adults aged 18 years and older, covering the period from **2011 to 2023**. It focuses on three key areas?

1. **Nutrition**: Represented by data related to the "Fruits & Vegetables" topic.
2. **Obesity**: Represented by data from the "Obesity and Weight Status" topic.
3. **Physical Activity**: Represented by data related to the "Physical Activity" topic.

The dataset presents its questions in percentage format, grouping responses into demographic categories such as **gender**, **income**, **education**, **age**, **race**, and **overall** averages. Additionally, the data is separated by state-level results and national averages, with the latter representing overall trends across the United States. This structure enables detailed analysis across various demographic and geographic segments, providing valuable insights into patterns and trends related to nutrition, physical activity, and obesity.

## Yellow Taxi Trip Dataset

The **“yellow\_tripdata\_2015-01.csv”** dataset contains **19 columns and 12,748,986 rows**, representing detailed information about taxi trips in New York City during January 2015. This dataset includes features such as:

* **Trip start and end times**.
* **Pick-up and drop-off locations**, recorded as longitude and latitude.
* **Passenger count**.
* **Trip distance**, in miles.
* **Fare amount** and other monetary metrics, such as tolls and tips.

This dataset provides a rich resource for analyzing large-scale transportation data, making it ideal for comparing the performance of Pandas and Spark in processing big datasets.

# Activity 1 ERD and Data Dictionary

## Item 1 – ERD Image

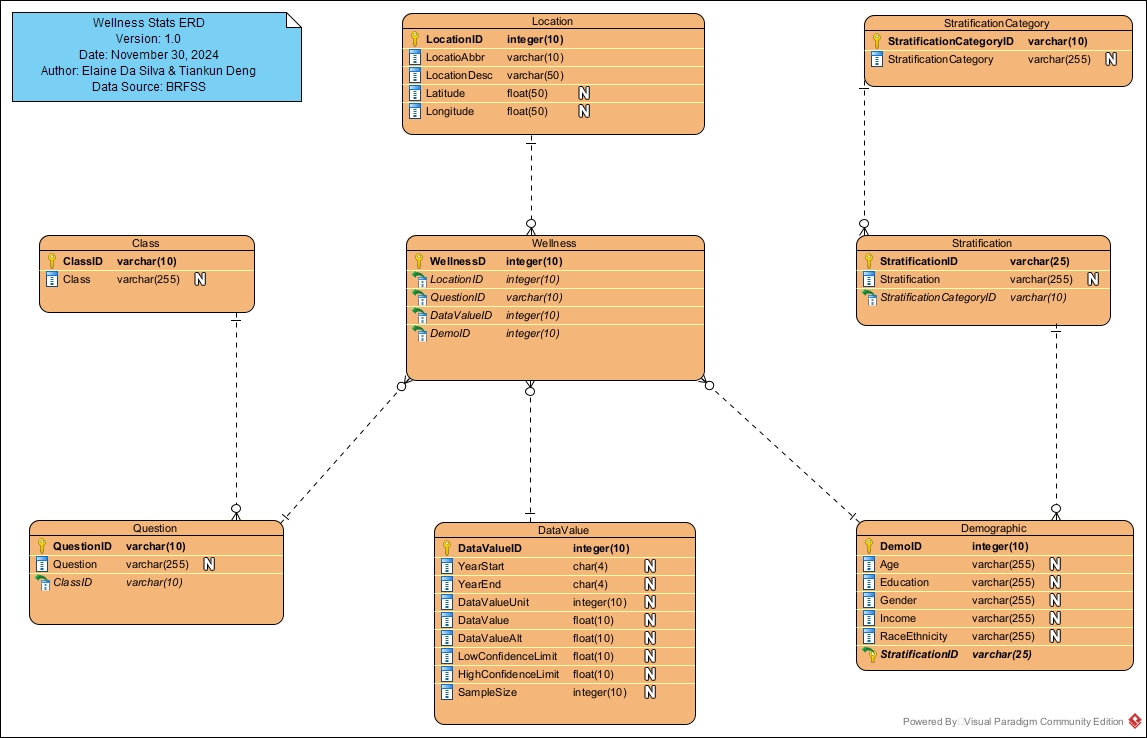


Figure 1 – ERD implementation

## Item 2 – Data Dictionary Image

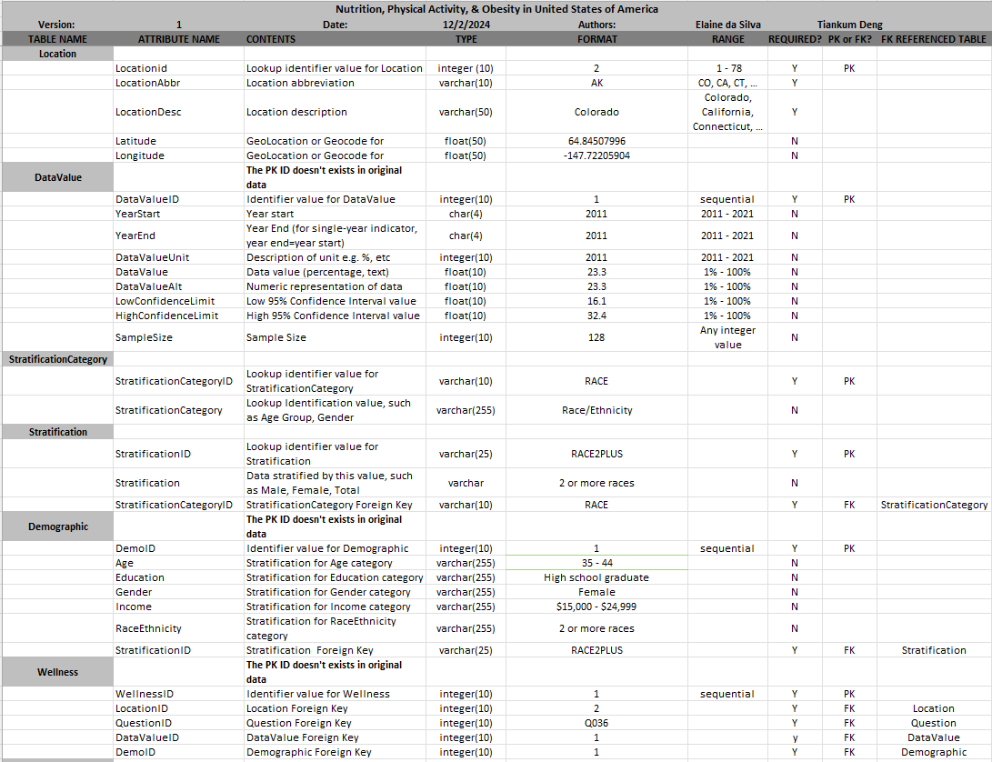


Figure 2 - Data Dictionary

# Activity 2 Data Wrangling

The dataset **“Nutrition\_Physical\_Activity\_and\_Obesity.csv”** has a unique characteristic: when one demographic category is selected, the other categories remain unpopulated or contain no values. This design ensures that the data is mutually exclusive for each demographic group, simplifying the analysis of individual categories but requiring careful handling during data processing to avoid misinterpretation of empty fields.

The figure below illustrates the demographic columns in their raw, unfiltered state, highlighting the structure of the data before any transformations or filtering are applied.

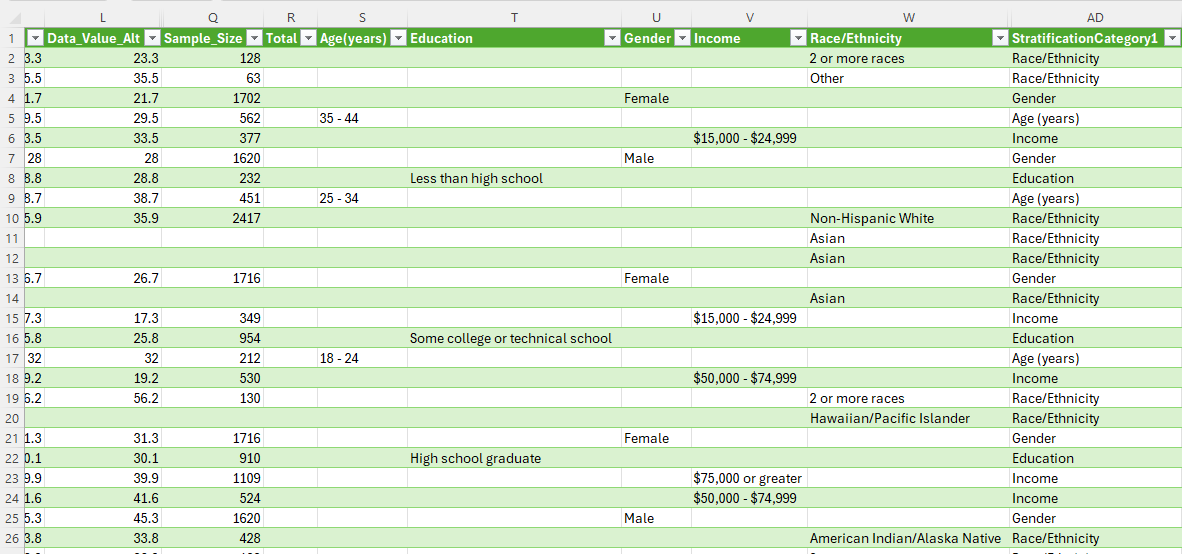


Figure 3 - “Nutrition\_Physical\_Activity\_and\_Obesity.csv” dataset

The figure below illustrates the demographic columns after filtering the “*StratificationCategory1*” column by “*Gender*”. This filtering isolates data specific to the selected demographic category, ensuring that only relevant values are retained while other categories remain unpopulated. This approach simplifies the analysis by focusing on the specific subset of interest, such as gender-based trends

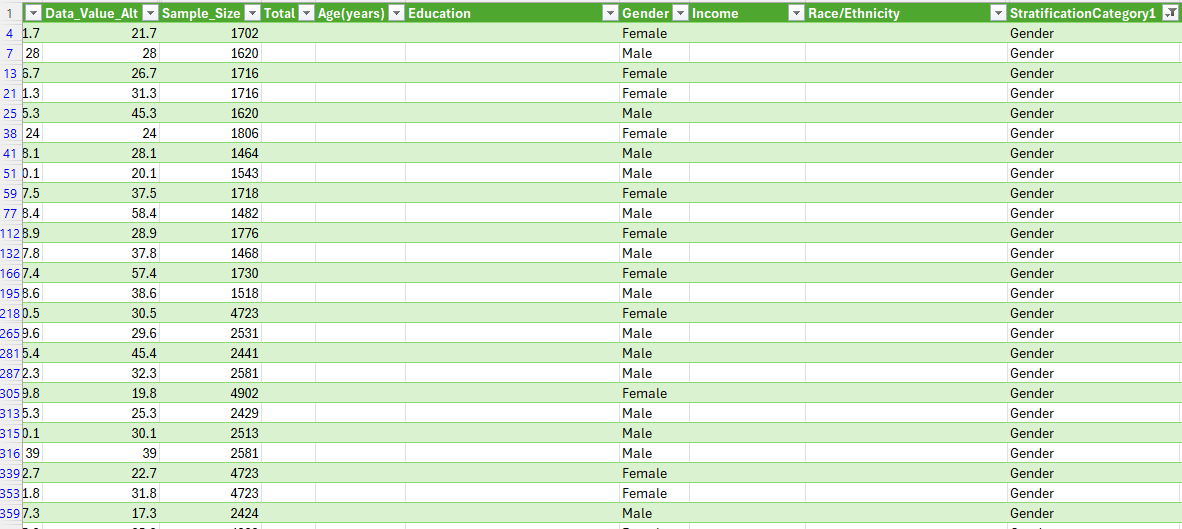


Figure 4 - Dataset filtered by "Gender" category

## Item 1 Exploring data wrangling

## Part A – Most used columns

*'Question', 'YearStart', 'LocationAbbr', 'LocationDesc'*: label columns

*'Data\_Value\_Alt'*: percentual column

*‘Income’, ‘Race/Ethnicity’, ‘Education’, ‘Age(years)’, ‘Gender’:* categorical columns

## Part B - Dealing with missing values

Due to the structure of the dataset, where selecting one demographic category results in the absence of values for the other categories, it was necessary to remove missing values when filtering by a specific category. This step ensures that only complete and relevant data is included in the analysis, allowing for more accurate insights and avoiding potential bias from incomplete records. To handle this issue, both null values and "data not reported" cases were removed, as demonstrated in the figures below.

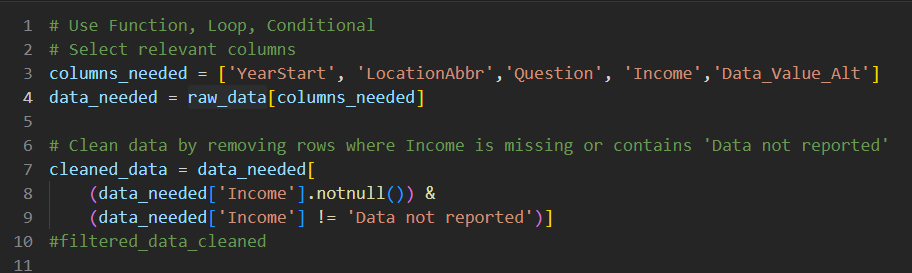


Figure 5 – dealing with missing data (Nulls)

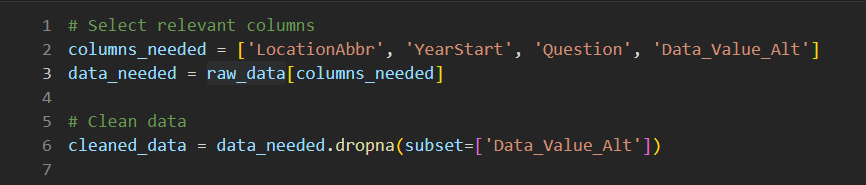


Figure 6 - dealing with missing data (NA)

# Activity 3 Exploring NumPy and Pandas

## Item 1 – Function, Loop, Conditionals, Read from & Write to CSV files

## PART A –Read from CSV files

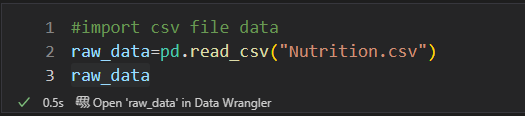


Figure 7 - Read from CSV

## PART B –Write to CSV files

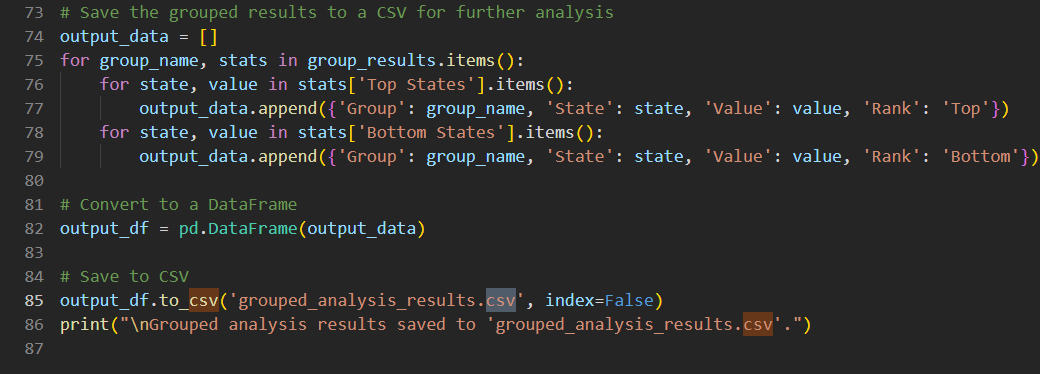


Figure 8 - Write to CSV

## PART C – Conditionals

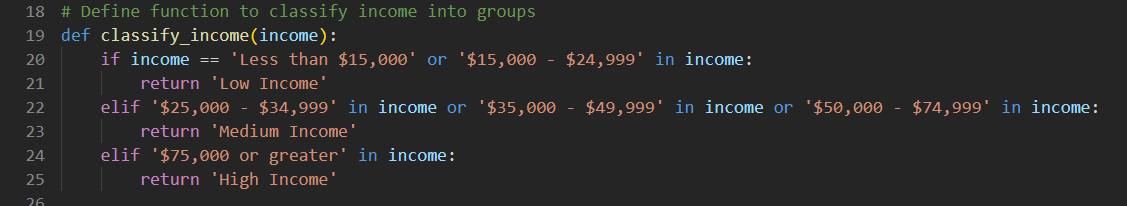


Figure 9 - Conditionals

## PART D – Loops

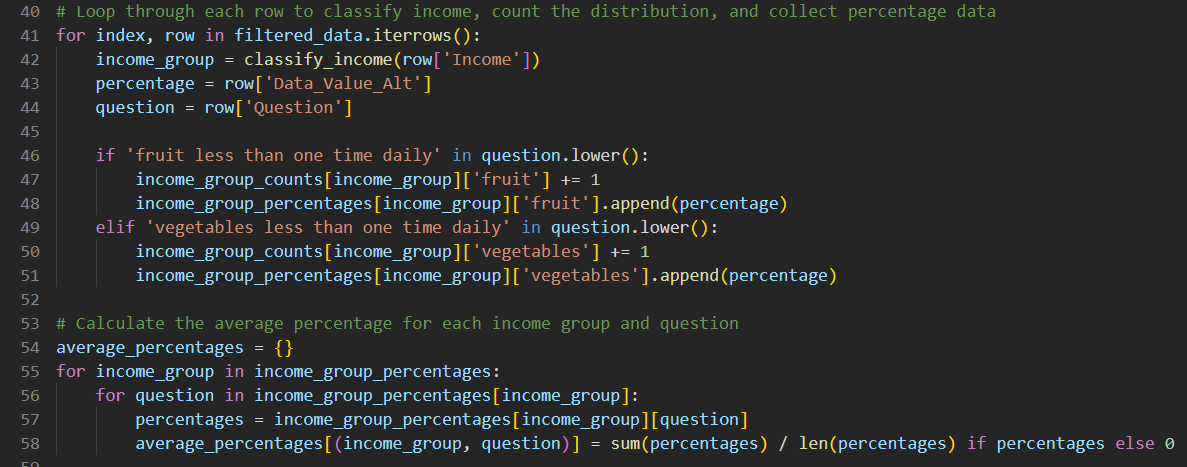


Figure 10 - Loops

## Item 2 – Exploring 2+ plots (visualization)

## PART A - Visualization 1

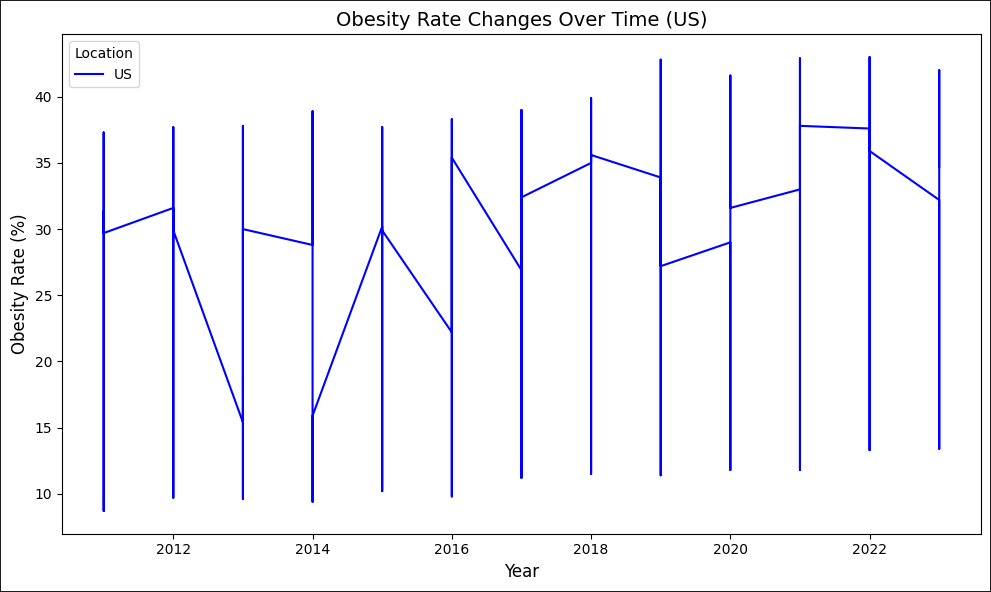


Figure 11 - Obesity rate over the years

## PART B - Visualization 2

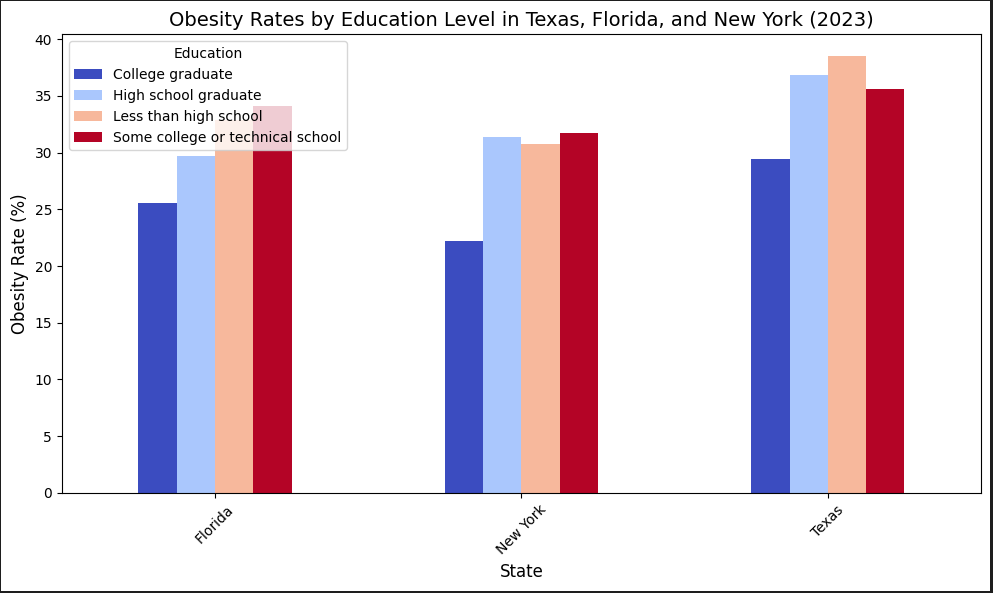


Figure 12 - Obesity rate by Education level

## Item 3 – Exploring descriptive statistics

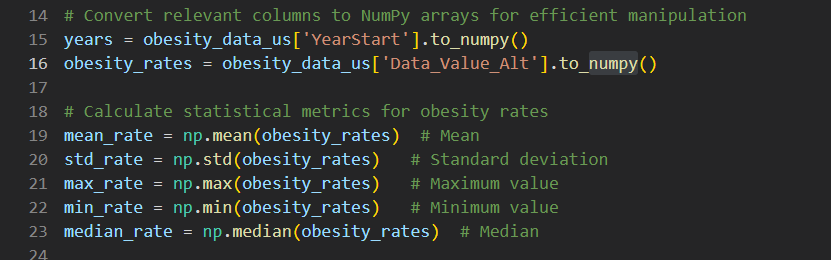


Figure 13 - Descriptive statistics

## Item 4 – Exploring Numpy Array and 2+ related functions

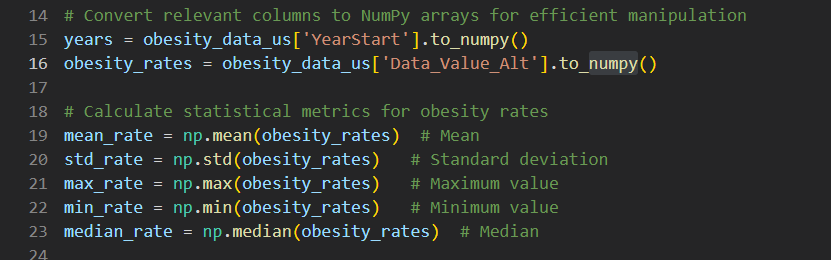


Figure 14 - Numpy arrays and functions

## Item 5 – Exploring Pandas DataFrame and 2+ related functions

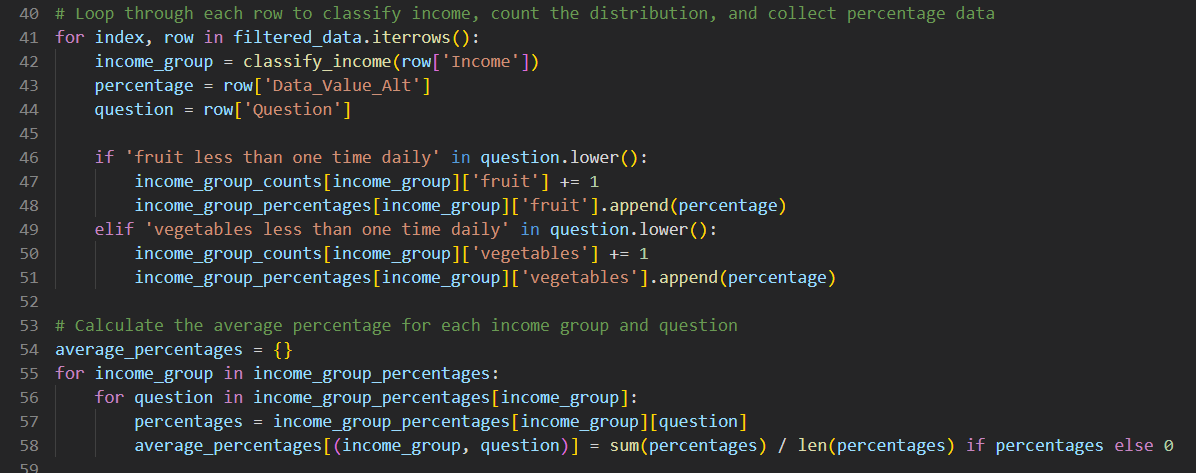


Figure 15 - Pandas dataframe and functions

# Activity 4 Time & Memory comparison Pandas vs Spark

The code used for this comparison is included as an attachment to this project on Brightspace.

## Item 1 – Datasets information

|  |  |  |
| --- | --- | --- |
|  | SMALL DATASET | BIG DATASET |
| Name | 'Nutrition\_Physical\_Activity\_and\_Obesity.csv' | yellow\_tripdata\_2015-01.csv |
| Columns | 33 | 19 |
| Rows | 104,272 | 12,748,986 |
| Size | 34,934 | 1,939,986 |
|  |  |  |

Figure 16 - Small and Big dataset information

## Item 2 – Small dataset statistics

PANDAS TIME AND MEMORY STATS: START => (2024-12-04 - 09:22:46)

Pandas: Data loaded: 0.53 seconds, Memory: 32.95 MB

Pandas: Filter operation: 0.13 seconds, Memory: 6.57 MB

Pandas: Aggregation: 0.00 seconds, Memory: 0.41 MB

PANDAS TIME AND MEMORY STATS: END => (2024-12-04 - 09:22:47)

SPARK TIME AND MEMORY STATS: START => (2024-12-04 - 09:23:45)

Spark: Data loaded: 0.39 seconds, Memory: 0.00 MB

Spark: Filter operation: 0.15 seconds, Memory: 0.08 MB

Spark: Aggregation: 0.34 seconds, Memory: 0.00 MB

SPARK TIME AND MEMORY STATS: END => (2024-12-04 - 09:23:46)

## Item 3 – Big dataset statistics

PANDAS TIME AND MEMORY STATS: START => (2024-12-04 - 09:24:12)

Pandas: Data loaded: 47.67 seconds, Memory: 3104.45 MB

Pandas: Filter operation: 2.63 seconds, Memory: 397.64 MB

Pandas: Aggregation: 0.08 seconds, Memory: 2.20 MB

PANDAS TIME AND MEMORY STATS: END => (2024-12-04 - 09:25:03)

SPARK TIME AND MEMORY STATS: START => (2024-12-04 - 09:36:32)

Spark: Data loaded: 17.05 seconds, Memory: 0.00 MB

Spark: Filter operation: 0.14 seconds, Memory: -0.07 MB

Spark: Aggregation: 5.57 seconds, Memory: 0.05 MB

SPARK TIME AND MEMORY STATS: END => (2024-12-04 - 09:36:55)

## Item 4 – Results analysis

**Small dataset​**

* Overall, Spark is more memory-efficient in all operations.​
* Spark is more optimized for loading.​
* Pandas is more efficient at aggregating data.​

​**Big dataset**​

* Spark is more memory-efficient for loading and filtering.
* Spark is more optimized for loading and filtering.​
* Pandas is more efficient at aggregating data

## Item 5 – How to decide between Pandas & Spark

In summary, Pandas and Spark are highly effective tools for data manipulation and analysis, each tailored to different needs and scenarios. The decision to use one over the other hinges on factors like dataset size, available computational power, and the specific demands of the analysis.

Pandas is ideal for processing small to medium datasets within a single-machine setup, offering simplicity and efficiency. On the other hand, Spark is designed to handle extensive datasets by utilizing distributed computing, making it the optimal choice for managing large-scale data and complex dataflows.

# Activity 5 Conclusion (findings)

* Higher-income groups tend to have healthier dietary habits, though this varies regionally.
* The average obesity rate in the U.S. is 30.64%, with a standard deviation of 6.3%, indicating significant state-level variation.
* Certain racial groups exhibit higher obesity rates, highlighting the influence of socioeconomic and cultural factors.
* Individuals with higher education levels are more likely to engage in physical activities.
* Obesity and physical activity rates have shown slight changes over the past decade (2011 vs 2021) with state-level trends varying.
* In California (2023), men slightly outperform women in physical activity participation, but with higher obesity rates.
* In Texas, Florida, and New York, higher educational attainment is inversely correlated with obesity rates, with notable differences among the states.
* Strength Training Across Texas, Florida, and New York varies significantly among these states.
* The top and bottom states in obesity, physical activity, and dietary habits reveal strong correlations between lifestyle choices and health outcomes.

# Activity 6 References

*Data.gov. (2023, December 8). U.S. Department of Health & Human Services - Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System. Retrieved December 1, 2024, from* [*https://catalog.data.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system*](https://catalog.data.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system)

*Nutrition, physical activity, and obesity - Behavioral Risk Factor Surveillance System | Data | Centers for Disease Control and Prevention. (2024, October 7). Retrieved December 1, 2024, from* [*https://chronicdata.cdc.gov/Nutrition-Physical-Activity-and-Obesity/Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7/about\_data*](https://chronicdata.cdc.gov/Nutrition-Physical-Activity-and-Obesity/Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7/about_data)

*NYC yellow taxi trip data. (2021, December 9). Kaggle. Retrieved December 4, 2024, from* [*https://www.kaggle.com/datasets/elemento/nyc-yellow-taxi-trip-data*](https://www.kaggle.com/datasets/elemento/nyc-yellow-taxi-trip-data)

*Functions — PySpark 3.5.3 documentation. (n.d.). Retrieved December 3, 2024, from* [*https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/functions.html*](https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/functions.html)

*Nelamali, N. (2024, October 1). Pandas vs PySpark DataFrame With Examples. Spark by {Examples}.* [*https://sparkbyexamples.com/pyspark/pandas-vs-pyspark-dataframe-with-examples/*](https://sparkbyexamples.com/pyspark/pandas-vs-pyspark-dataframe-with-examples/)