**Programming for Data Science CSL225**

**Project Report**



Faculty name: **Ms. Akanksha Kaushik**

Students name: **Sanya Manchanda(23csu282)**

**Sarthak Sabharwal(23csu284)**

**Shobhit Niraj Anadkat(23csu290)**

Semester: 3rd

Group: DS-B-III

**Department of Computer Science and Engineering**

**The NorthCap University, Gurugram- 122001, India** (**Session 2024-25 )**

**Topic: Data Cleaning and Exploration of a Health Survey Dataset Using Python**

**Table of Contents**

|  |  |
| --- | --- |
| S.No | Topics |
| 1. | Project Description |
| 2. | Problem Statement |
| 3. | Techniques and Tools Used |
| 4. | Challenges Addressed and Feature of Project |
| 5. | Code |
| 6. | Output (Screenshots) |
| 7. | Conclusion and Future Scope |

# Project Description

This project focuses on cleaning and exploring a health survey dataset using Python to derive meaningful insights and understand health trends. The cleaning process includes handling missing values, correcting inconsistencies, removing duplicates, and standardizing data formats to enhance its quality and usability. Exploratory Data Analysis (EDA) will be conducted to identify patterns, relationships, and trends through descriptive statistics, correlation analysis, and data visualization. Tools like Pandas, NumPy, Matplotlib, and Seaborn will be utilized for efficient data manipulation and visualization. The project aims to create a clean, analysis-ready dataset and uncover insights that can aid in healthcare research and decision-making.

# Problem Statement

Health survey datasets often contain raw, unstructured data with missing values, inconsistencies, and redundancies, making them unsuitable for analysis. Without proper cleaning and exploration, critical insights into health patterns, trends, and relationships remain hidden, limiting their potential for informed decision-making in healthcare. The challenge lies in transforming such raw data into a clean, standardized, and analysis-ready format while uncovering meaningful insights

through exploratory data analysis. This project aims to address these challenges by utilizing Python for data cleaning and EDA, ensuring data quality and deriving actionable insights to support healthcare research and decision-making.

## Techniques and Tools Used

**Python**: A versatile programming language chosen for its strong data analysis and visualization capabilities.

**Libraries**:

* **Pandas**: Utilized for loading and cleaning the dataset, manipulating data into DataFrames, and conducting statistical analyses.
* **Matplotlib**: Used for creating detailed charts and plots to visualize insights, such as bar charts for delay reasons and pie charts for cancellation rates.

**Development Environment**: The project was developed using Jupyter Notebook, an environment ideal for iterative coding and visualization.

## Challenges Addressed

* **Handling Missing Values:** Missing data is managed through imputation or removal techniques to ensure the dataset remains reliable and comprehensive for analysis, preventing biases or inaccuracies caused by incomplete information.
* **Resolving Data Inconsistencies**:  
  Formatting errors, invalid entries, and inconsistent units are corrected, ensuring uniformity and enabling seamless comparisons and analyses across all dataset variables.
* **Dealing with Outliers**:  
  Outliers are detected and appropriately addressed to avoid distortion of statistical insights, ensuring genuine patterns and trends are accurately represented.
* **Removing Duplicate Records**:  
  Duplicate entries are identified and removed to prevent skewed results and maintain the integrity of the dataset for accurate analysis.
* **Standardizing the Dataset**:  
  All data is converted into consistent formats, scales, and units, making it analysis-ready and improving the efficiency of subsequent exploration processes.

**Features of Project**

 **Data Retrieval**: The project fetches health-related data in real-time from a CDC API, which provides valuable demographic information. The data is downloaded in CSV format and processed using Pandas for further analysis and visualization.

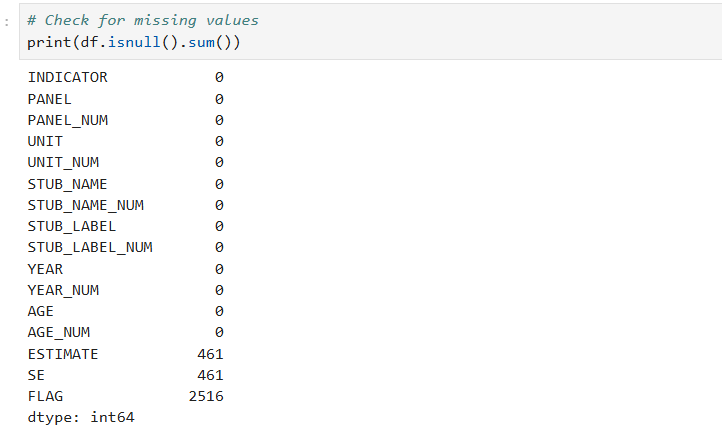
 **Data Analysis**: The notebook performs a comprehensive exploration of the dataset, summarizing key statistics, demographics (e.g., age, sex, ethnicity), and health-related behaviors. Filters are applied to segment data, allowing focused insights on specific categories for deeper understanding.

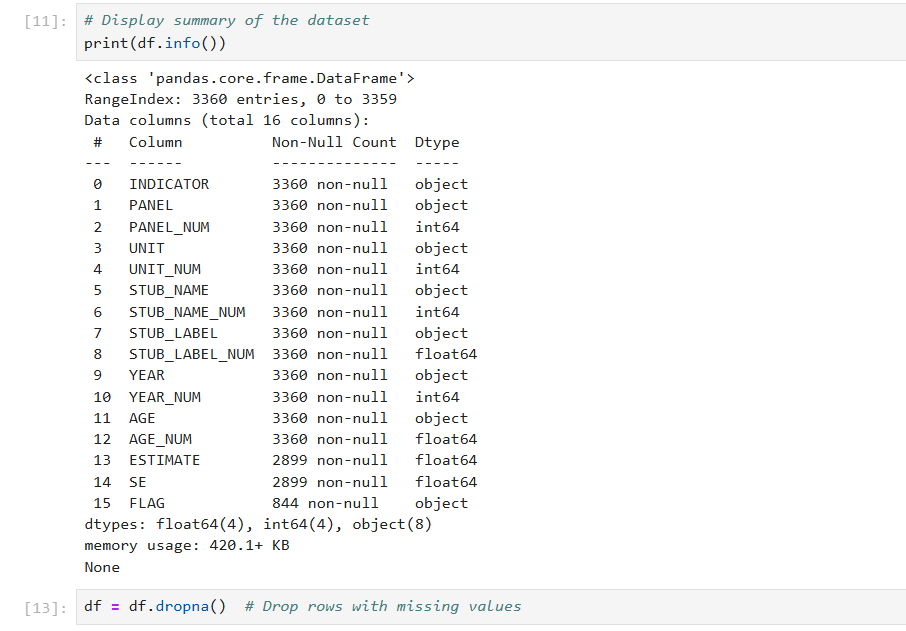
 **Visualizations**: Various visualizations are generated using Seaborn and Matplotlib, primarily bar plots, to showcase trends and distributions in the dataset. These visualizations help identify patterns and provide a clear understanding of the data's structure and trends.

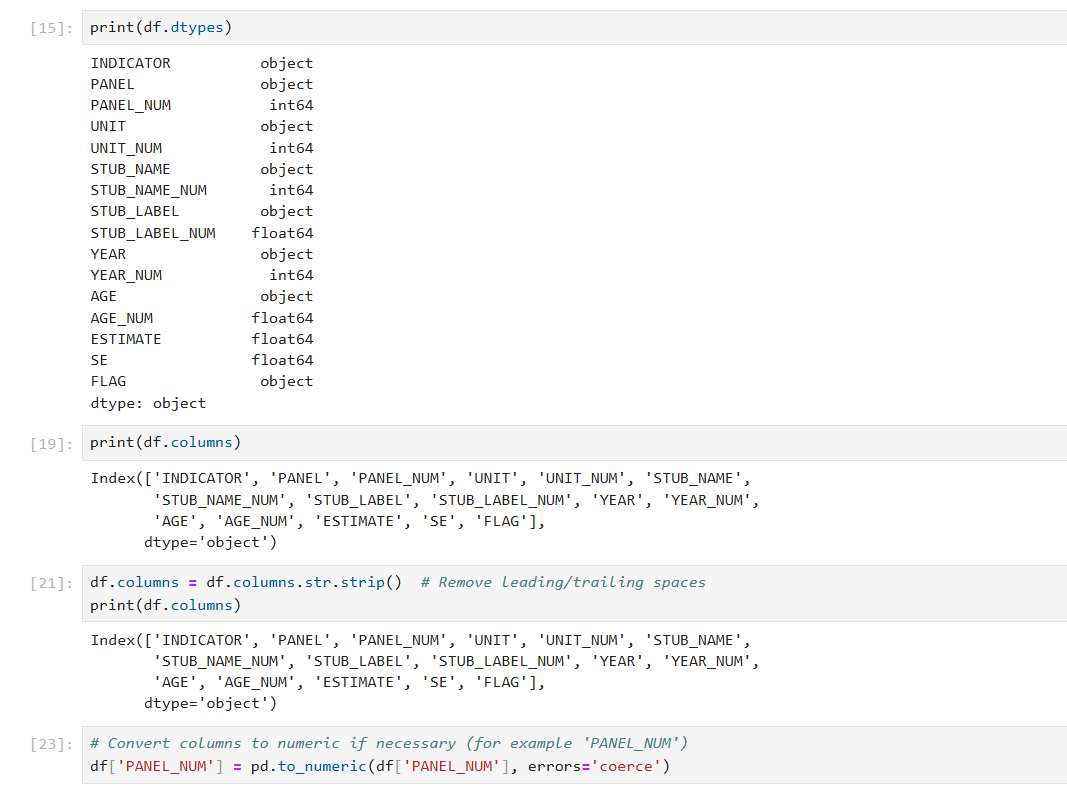
 **Insights Export**: The project allows users to export insights and analysis results in different formats, such as CSV or PDF. This feature is useful for sharing reports or generating summaries that can be used for further research or presentations.

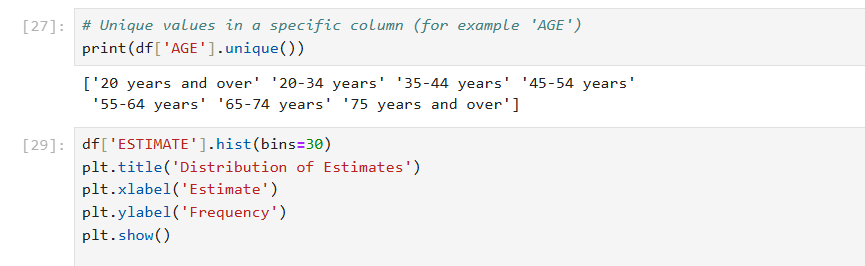
**Code**

****

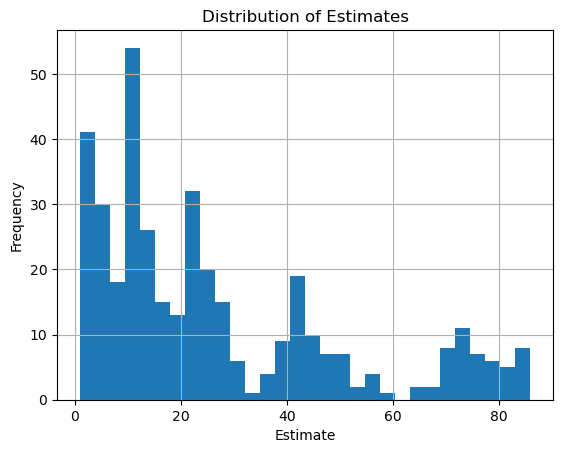


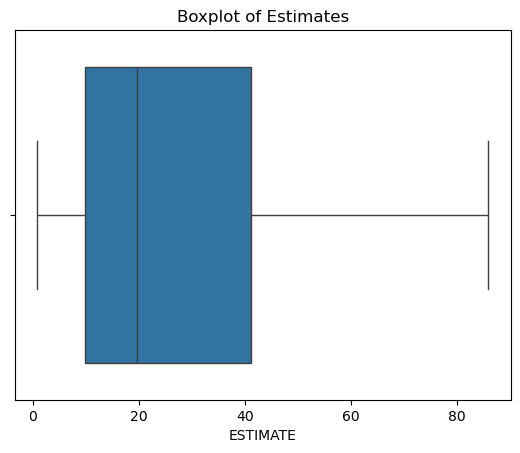


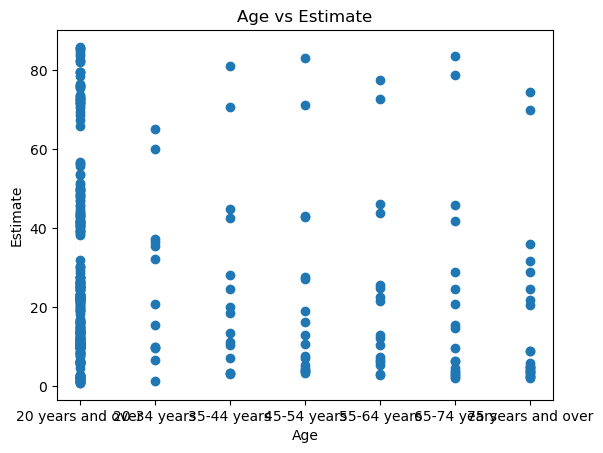


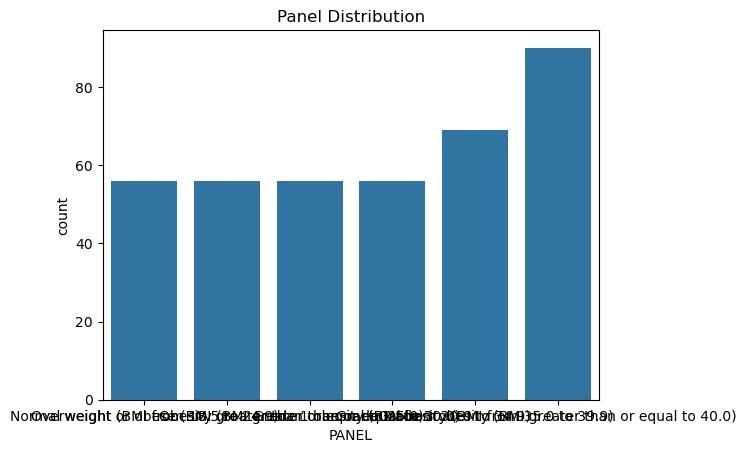


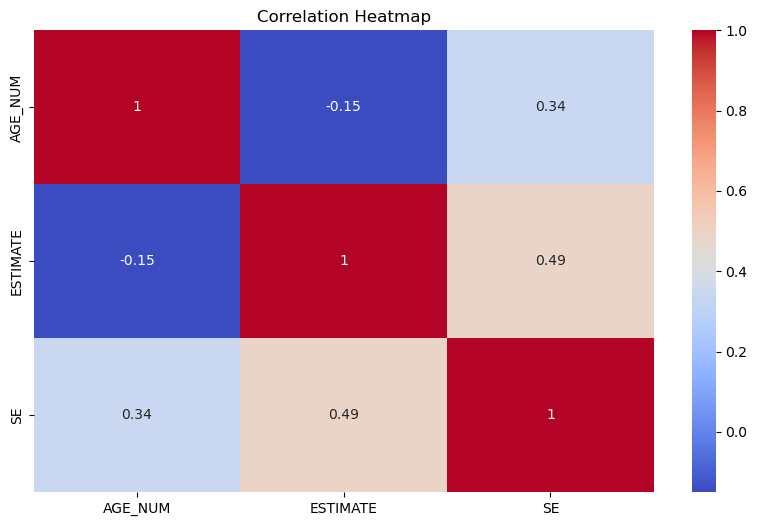
**Output**

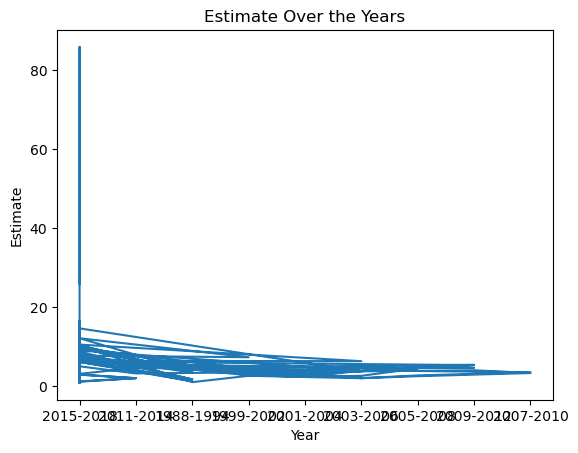












**Conclusion and Future Scope**

## Conclusion

This project cleaned and explored a health survey dataset using Python, addressing issues like missing values and data inconsistencies. The analysis revealed key health trends and patterns, demonstrating the importance of data cleaning for reliable insights and informed healthcare decision-making.

## Future Scope

1.Advanced analytics such as machine learning models could be implemented to predict health outcomes and identify risk factors based on survey data.

2.Integrating additional health datasets for more comprehensive insights, enabling analysis across larger populations and regions.

3.Developing real-time data pipelines and interactive dashboards to facilitate dynamic data visualization and improve decision-making in healthcare settings.