PROJECT JOURNAL

Topic :Tax Data Analysis Using Python and Impact of Tax on Alzheimer's Disease

Team member : Naveen Mummana

Student ID: X23254777

WEEK 1

Date: 7/01/25

Task: Data Collection and Processing

In the first week, I focused on acquiring and processing a dataset on income tax components from 1999 onwards. The task involved importing raw data in CSV format, normalising it for consistency, and structuring it for analysis. MongoDB was selected as the database platform because of its ability to handle semi-structured data efficiently.

Steps Completed:

- Compile a comprehensive dataset of income tax returns by income size and residence.

- Using Pandas and other Python libraries to ensure cleanliness and consistency of raw data.

- Structure clean data and upload it to MongoDB collections that can be easily queried and analyzed.

Time Spent:10 hours

Challenges Faced:

To manage incomplete or inconsistent records in raw data - Normalize data in multiple formats into a single structure for integration with MongoDB.

Solutions:

- Use data cleaning scripts to fill in missing values and eliminate redundancies.

- Create validation checks to ensure post-import data integrity

Learning Outcomes:

- Learned the intricacies of using MongoDB for managing semi-structured datasets.

- Gained hands-on experience in cleaning and normalizing tax-related data.

WEEK 2

Date: 13/01/25

Task: Exploratory Data Analysis (EDA)

During this week, I conducted EDA to uncover trends and spatial patterns in income tax burdens. Visualizations were key in interpreting data and highlighting insights.

Steps Completed:

- Analyzed temporal patterns, identifying fluctuations in tax burdens over the years.

- Explored regional variations, categorizing areas by high and low tax burdens.

- Generated static visualizations using Matplotlib and Seaborn for initial insights.

Time Spent: 8 hours

Challenges Faced:

- Distinguishing between geographic factors and economic influences in tax trends.

- Identifying patterns while avoiding overfitting conclusions to preliminary data.

Solutions:

- Applied statistical filters to ensure robust trend identification.

- Focused on deriving insights from consistent patterns across different regions.

Learning Outcomes:

- Improved skills in interpreting regional disparities through data visualizations.

- Gained a deeper understanding of tax trends and their socio-economic impacts.

WEEK 3

Date:22/01/25

Task: Feature Engineering and Advanced Visualizations

In the third week, I concentrated on creating new features and developing advanced, interactive visualizations. This involved grouping regions and assessing income group tax burdens.

Steps Completed:

- Generated derived features, such as the percentage change in tax components over time.

- Grouped regions by income levels for comparative analysis.

- Created interactive dashboards using Plotly for enhanced accessibility.

Time Spent:10 hours

Challenges Faced:

- Ensuring derived features added meaningful insights without introducing bias.

- Balancing the complexity of interactive dashboards with user accessibility.

Solutions:

- Validated new features through statistical tests to confirm relevance.

- Iteratively improved dashboard design based on user feedback.

Learning Outcomes:

- Developed expertise in creating meaningful features for tax trend analysis.

- Strengthened proficiency in building interactive visualizations with Plotly.

WEEK 4

Date: 27/01/25

Task:Reporting and Future Directions

In the final week, I compiled research findings and proposed directions for future work. This included summarising temporal and spatial tax trends and identifying limitations.

Steps Completed:

- Highlighted key findings, such as notable fluctuations and regional disparities.

- Proposed integrating socio-economic indicators into future analyses.

- Suggested using machine learning models to predict tax trends and evaluate policy impacts.

Time Spent: 8 hours

Challenges Faced:

- Ensuring comprehensive reporting while maintaining clarity.

- Accounting for limitations, such as missing data and lack of external variable modeling.

Solutions:

- Structured reports for clarity and included actionable insights for improvement.

- Emphasized the need for future data augmentation and modeling approaches.

Learning Outcomes:

- Improved reporting skills for summarizing complex analytical findings.

- Recognized the importance of contextualizing results within socio-economic frameworks.

SUMMARY

Total Time Spent: 36+ hours

Key Accomplishments:

- Structured and managed large tax datasets using MongoDB.

- Derived meaningful insights through EDA and advanced visualizations.

- Proposed future work integrating machine learning for predictive analysis.