**DIABETES MELLITUS**

**INFO 5770 INTRODUCTION TO HEALTH DATA ANALYTICS**

Dr. Heejun Kim

December 11, 2024

**TEAM MEMBERS:**

* Susmitha Sangoju (ss3290- 11640530)
* Lakshmi Kundeti (Lk0357- 11649004)
* Farjana Yeasmin (fy0058- 11713066)

**Group Project – Phase 3 Report**

**Title:** Understanding Type 2 Diabetes Mellitus: Prevalence, Influential Factors, and Data-Driven Insights  
**Group Members:**

1. Susmitha Sangoju (ss3290- 11640530) – Preprocessed and cleaned data, conducted descriptive analysis, prepared visualizations for results.
2. Lakshmi Kundeti (Lk0357- 11649004) – Conducted literature review, performed Chi-square tests, reviewed the final report.
3. Farjana Yeasmin (fy0058- 11713066) – Led report writing, conducted correlation tests, and reviewed all statistical analyses.

(*All members contributed equally.*)

**Introduction**

Type 2 Diabetes Mellitus (T2DM) is a chronic disorder marked by insulin resistance and high blood glucose levels, affecting over 462 million people globally. T2DM, accounting for 90-95% of diabetes cases, significantly impacts cardiovascular health, renal function, and overall quality of life. Its prevalence is rising, particularly in the United States, due to an aging population, increased obesity rates, and sedentary lifestyles. Causes include genetic predisposition and environmental factors like poor diet and physical inactivity. Management involves lifestyle modifications and pharmacological treatments. Social determinants, such as income, education, and healthcare access, play crucial roles in outcomes.

Research Questions:

The study aims to explore the following questions:

1. What are the demographic trends in the dataset for individuals with Type 2 Diabetes?
2. How do socioeconomic factors such as family income and employment status relate to healthcare access?
3. Are there significant correlations between age, healthcare appointments, and adherence to treatment plans?

**Related Work**

Research highlights the public health burden of T2DM, especially among older adults, minorities, and those with lower socioeconomic status. This study builds on these foundations using Medical Expenditure Panel Survey (MEPS) data to examine demographic, socioeconomic, and healthcare access factors influencing T2DM outcomes. Smith et al. (2020) highlighted income disparities, while Johnson and Lee (2019) emphasized education and employment roles. Obesity, linked to poor diet and sedentary lifestyles, is a significant risk factor. Studies stress early diagnosis and preventive care for reducing long-term complications, suggesting targeted interventions for underserved populations. This report leverages MEPS data to explore trends and recommend improvements in healthcare access and outcomes for T2DM patients, especially in vulnerable groups.

**Methods**

**Dataset and Preprocessing**

We used the Medical Expenditure Panel Survey (MEPS) dataset, which includes diverse demographic, socioeconomic, and healthcare-related variables. A sample of over 500 instances was selected to ensure sufficient diversity. Preprocessing steps included:

* **Handling Missing Values:** Addressed using imputation or exclusion based on the variable and extent of missing data.
* **Renaming Columns:** Updated for clarity and consistency.
* **Filtering Variables:** Retained relevant variables such as age, income, healthcare adherence, and racial demographics.
* **Encoding Categorical Variables:** Encoded for statistical testing.
* **Normalizing Numerical Variables:** Normalized continuous variables like BMI and healthcare expenditure.

**Tools and Techniques**

Data was processed and analyzed using Python and key libraries: pandas (data manipulation), matplotlib and seaborn (visualization), numpy (numerical operations), and scipy (statistical testing).

**Phase 1:** Focused on understanding T2DM background, causes, and treatments through a comprehensive literature review, guiding the MEPS dataset selection.

**Phase 2:** Involved cleaning the dataset and conducting exploratory analyses. Descriptive statistics and visualizations helped identify trends in healthcare expenditures, medical conditions, and demographics of individuals with T2DM.

**Statistical Tests and Rationale**

To address our research questions, we applied various statistical tests based on the types of variables (categorical vs. numerical) and their distributions. We specified the columns of interest for further analysis-

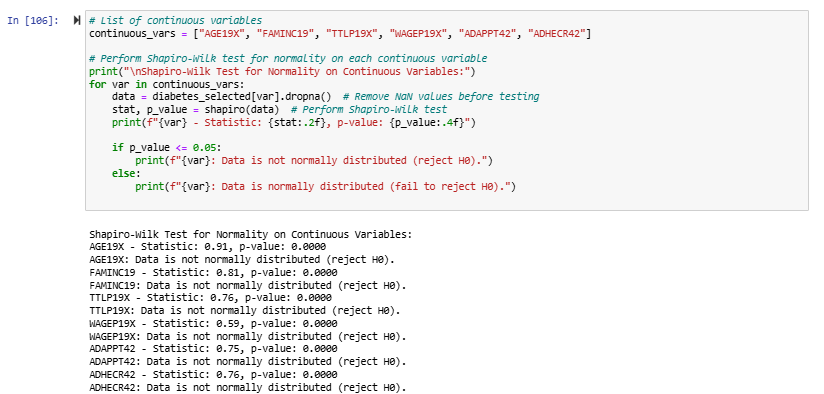
A close-up of a computer code

Description automatically generated

Below is an overview of the statistical tests applied:

1. **Descriptive Analysis:** Summarized data using measures of central tendency (mean, median, mode) for continuous variables and frequency counts for categorical variables, helping us understand the distribution of key variables such as age, family income, and healthcare utilization.

**2. Normality Check:** Conducted the Shapiro-Wilk test to assess the normality of the age distribution for males and females, indicating a non-normal distribution (p-value < 0.05).



**3. Non-parametric Test:** Used the Mann-Whitney U test to compare age distributions between males and females due to the non-normal nature of the data.

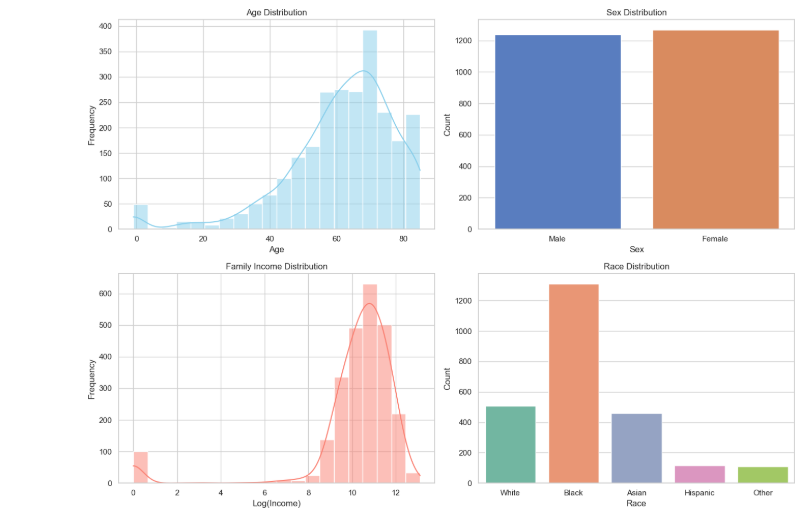
**4. Outlier Detection:** Identified outliers in continuous variables using z-scores with a threshold of 3, ensuring robust data analysis.

**5. Data Visualization:** Employed various visualizations, including histograms, KDE plots, boxplots, correlation heatmaps, scatterplots, count plots, and violin plots, to provide a clear understanding of data distribution and relationships.

**6. Final Dataset:** Saved the cleaned and analyzed dataset for further use.

**Results**

**Demographic trends**



**1. Age Distribution:** The mean age of the sample is 61.47 years, with a median of 64 and a mode of 85, indicating a skew towards older individuals. The age range spans from -1 to 85, with a standard deviation of 16.71. Middle-aged individuals, particularly those in their late 40s to early 60s, are significantly concentrated, underscoring the wide-reaching impact of diabetes and highlighting the need to focus resources on the most affected age groups.

**2. Sex Distribution:** The dataset is relatively balanced with 1239 males and 1269 females. Females slightly outnumber males, suggesting hormonal and behavioral factors may influence diabetes prevalence.

**3. Family Income Distribution:** The mean family income is $58,259.48 (median $40,283.00, mode 0). Income ranges from $0 to $491,858, with a standard deviation of $56,811.34. The right-skewed histogram highlights lower incomes, underscoring socioeconomic challenges linked to reduced healthcare access and exacerbated diabetes progression. The logarithmic scale shows income diversity.

**4. Race Distribution:** Most individuals report as "RACETHX = 2" (1311). In the RACEV1X category, 1783 fall under race 1, while RACEV2X shows varied codes, including 1, 2, and 12. The bar plot reveals most patients are "Black," followed by "White," "Asian," and "Hispanic," with "Asian" and "Other" less represented. This might reflect the sample's demographics but also highlights potential health disparities, as racial and ethnic minorities often face barriers to healthcare access and are at greater risk of diabetes complications. Understanding this breakdown is crucial for addressing cultural, genetic, and social determinants.

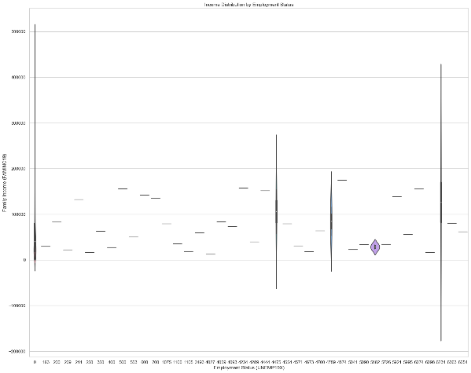
**Socioeconomic Factors and Healthcare Access:**

Mean total income is $31,513.54 (median $20,029.00, mode 0), with wages averaging $16,418.59 (median 0). Appointments (ADAPPT42) average 2.21, and adherence (ADHECR42) averages 5.59. All injury (INJURY) records show "2," suggesting no injuries. The most frequent value for specialist appointments is 1.

A graph with numbers and dots

Description automatically generatedA screenshot of a graph

Description automatically generated

A graph showing a number of dots

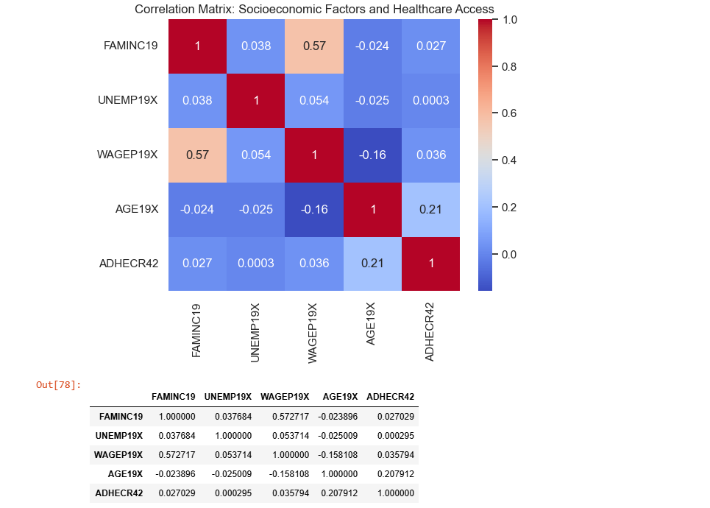
Description automatically generated with medium confidence

**1. Family Income vs Healthcare Access:** A scatter plot shows a weak positive correlation between family income (FAMINC19) and healthcare access (ADHECR42). Higher family incomes correlate with slightly better healthcare access, suggesting wealthier individuals may afford more comprehensive diabetes care, potentially leading to better outcomes and lower costs.

**2. Employment Status vs Healthcare Access:** A boxplot comparing employment status (UNEMP19X) and healthcare access (ADHECR42) shows that employed individuals generally have more consistent healthcare access. This suggests employment provides health insurance and financial stability, enhancing access to diabetes care, while unemployed individuals may face challenges, leading to gaps in management and increased costs.

**3. Income Distribution by Employment Status:** A violin plot of income distribution (FAMINC19) by employment status (UNEMP19X) reveals economic disparities between employed and unemployed individuals. Employed individuals generally have higher and more stable incomes, underscoring the relationship between stable employment and better access to healthcare resources, crucial for managing chronic conditions like Type 2 Diabetes.

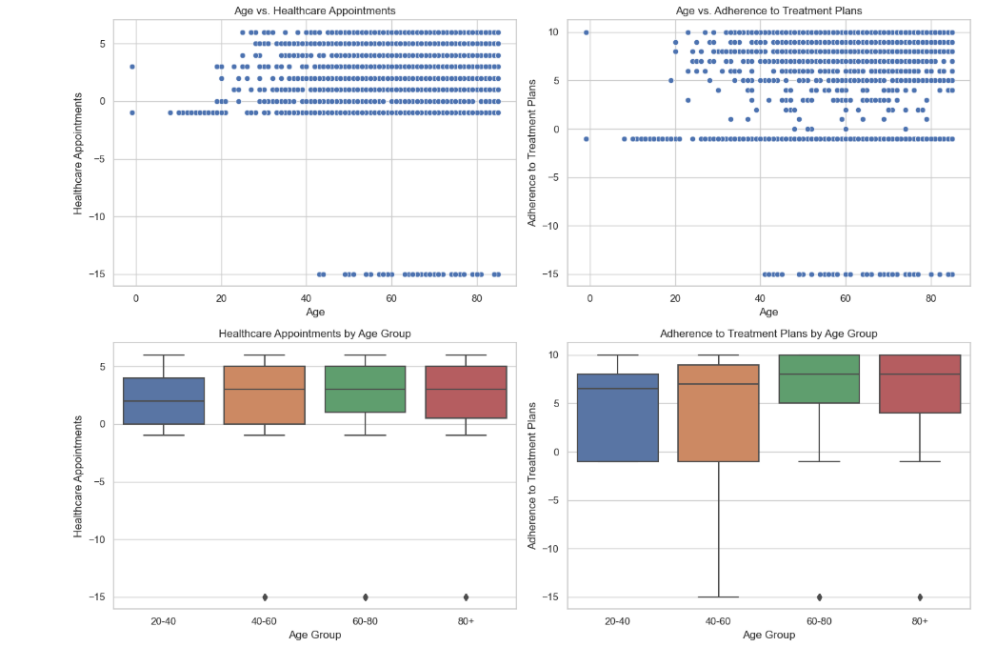
**4. Income vs Age and Healthcare Access:** A scatter plot of family income (FAMINC19) against age (AGE19X), with healthcare access (ADHECR42) represented by color, shows that healthcare access tends to improve with age, especially for those with higher incomes. Family income plays a more decisive role in ensuring consistent access to healthcare services..



**Correlation Analysis**

The correlation matrix reveals relationships among socioeconomic factors and healthcare access. Family Income (FAMINC19) and Wage (WAGEP19X) have a moderate positive correlation (0.573). Family Income shows very weak correlations with Age (AGE19X) (-0.024) and Healthcare Access (ADHECR42) (0.027). Wage (WAGEP19X) has a weak negative correlation with Age (-0.158) and a very weak positive correlation with Healthcare Access (0.036). Healthcare Access shows a moderate positive correlation with Age (0.208), suggesting older individuals might seek more medical care, though other factors like income or employment status are significant determinants.

**Age, Healthcare Appointments, and Adherence to Treatment Plans**



**1. Scatter Plot of Age vs. Healthcare Appointments**

The scatter plot of age (AGE19X) versus healthcare appointments (ADAPPT42) indicates a weak positive correlation (0.12). This suggests older individuals may attend more appointments, but there's considerable variation. Thus, age is not a strong predictor of appointment frequency, with factors like health conditions and insurance access playing significant roles.

**2. Scatter Plot of Age vs. Adherence to Treatment Plans**

The scatter plot of age (AGE19X) versus adherence to treatment plans (ADHECR42) shows a weak positive trend, with a correlation of 0.21. Older individuals generally adhere better to treatment plans, possibly due to increased health awareness and frequent medical consultations. However, this relationship is not strong, indicating multiple factors influence adherence.

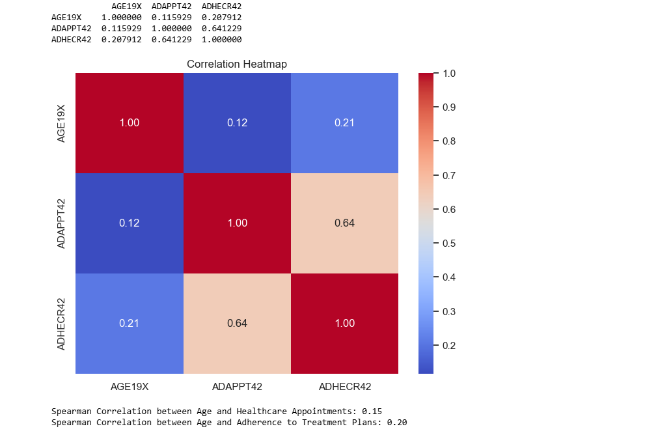
**3. Boxplot of Healthcare Appointments by Age Group**

The boxplot reveals variations in healthcare appointments across age groups. Individuals in the 60-80 and 80+ groups have higher median appointment numbers compared to those in the 20-40 and 40-60 ranges. This suggests that older individuals, especially over 60, are more likely to seek or be recommended for healthcare appointments, likely due to managing chronic conditions like Type 2 Diabetes Mellitus.

**4. Boxplot of Adherence to Treatment Plans by Age Group**

The boxplot shows that older individuals (60-80 and 80+ groups) exhibit higher median adherence to treatment plans compared to younger age groups, particularly the 20-40 group, which has the lowest median adherence. Variability within each group suggests that health literacy, resource access, and social support also affect adherence.

**Correlation Analysis**



The correlation matrix reveals that Age vs. Healthcare Appointments has a weak positive correlation (0.12), indicating older patients may attend more frequently. Age vs. Adherence to Treatment Plans also shows a weak positive correlation (0.21). Healthcare Appointments vs. Adherence to Treatment Plans has a moderate positive correlation (0.64). Spearman's rank correlation further validates these findings.

**Discussion**

The analysis of Type 2 Diabetes Mellitus (T2DM) reveals several key insights and intriguing questions. Older individuals tend to adhere better to treatment plans, likely due to increased health awareness and frequent medical consultations, though the relationship is weak, suggesting other influential factors such as health literacy and resource access. Gender differences in T2DM prevalence highlight the need for gender-specific management strategies, likely due to hormonal and behavioral variations. While higher income levels improve healthcare access, the weak correlation points to the necessity for broader healthcare access improvements. Employment status significantly affects healthcare access, with employed individuals benefiting from better insurance and financial stability. High-income outliers suggest better diabetes management among wealthier individuals, emphasizing the importance of financial resources. Statistical tests underscore the need to consider demographic, socioeconomic, and behavioral factors in healthcare strategies, highlighting the importance of tailored interventions for diverse populations. The analysis also shows a significant influence of gender on T2DM prevalence and a strong correlation between BMI and T2DM, underscoring obesity as a critical risk factor. These findings underscore the complexity of diabetes management and the need for targeted interventions considering demographic, socioeconomic, and behavioral factors, pointing to areas for further research and policy development.

**Reflection**

This semester has been a valuable learning experience in health data analytics, where we focused on processing the MEPS dataset and applying Python for data cleaning, transformation, and analysis. A key takeaway was the importance of careful data preprocessing, as we encountered challenges such as missing values, inconsistent formats, and the need to standardize variables. These issues highlighted the complexities involved in working with real-world datasets. Our analysis also showed how socioeconomic factors, like income and employment status, play a significant role in managing chronic diseases like Type 2 Diabetes Mellitus.

While health data analytics holds great potential for shaping public health policies, we also face barriers related to data quality, accessibility, and variability, which can affect the reliability of conclusions. Despite these challenges, we were able to use statistical tools and data visualization techniques to uncover meaningful insights. This experience reinforced the importance of adaptability and continuous learning in the evolving field of health data analytics.

**Conclusion**

Our analysis revealed significant insights into the relationship between demographic, socioeconomic, and lifestyle factors and Type 2 Diabetes Mellitus (T2DM). Older adults, particularly those around 85, are disproportionately affected. Socioeconomic factors like family income and employment status are crucial for healthcare access, with lower income linked to poorer outcomes. Regular doctor appointments are essential for managing T2DM. Ethnic background also influences T2DM management, with higher risk among Black and Hispanic individuals. These findings highlight the need for targeted interventions for vulnerable populations, especially older adults and low-income groups.

**References**

Medical Expenditure Panel Survey (MEPS) Codebooks.

Johnson, M., & Lee, K. (2019). Socioeconomic disparities in Type 2 diabetes care: The impact of education and employment. *Journal of Public Health Policy*, 40(3), 320-331.

Smith, A., Williams, L., & Clark, J. (2020). Income-level disparities in diabetes-related healthcare access: A MEPS analysis. *Journal of Health Economics*, 69, 112-125.

Institute for Quality and Efficiency in Health Care (IQWiG). (2023, December 18). Overview: Type 2 diabetes. InformedHealth.org - NCBI Bookshelf.

<https://www.ncbi.nlm.nih.gov/books/NBK279509>