Alzheimer’s Disease Dataset EDA Report

Introduction

Alzheimer's Disease is a progressive neurodegenerative disorder that affects millions of people worldwide. Early diagnosis and understanding of various contributing factors can aid in managing the disease better. This exploratory data analysis (EDA) aims to delve into a dataset containing patient demographics, lifestyle factors, medical history, clinical measurements, and symptoms to uncover meaningful patterns that could assist in improving diagnosis and patient care.

Inspiration

This dataset was chosen for its comprehensive coverage of multiple variables relevant to Alzheimer's Disease. It allows for in-depth analysis across a wide array of patient details, providing us with an opportunity to explore correlations between lifestyle, medical history, and diagnosis.

Why This Dataset?

The dataset contains patient demographic information, lifestyle factors, and medical history details for Alzheimer's patients. By analyzing this data, we can uncover insights into how various factors influence the onset and progression of Alzheimer's Disease, potentially aiding in more accurate diagnoses and better patient management.

What Can Be Learned?

Through this dataset, we can explore relationships between variables like age, gender, medical history, cognitive function, and Alzheimer's diagnosis. It can help identify early indicators of Alzheimer's and understand how lifestyle factors like diet and physical activity impact disease progression.

Questions Raised and Answered

1. How does age impact the likelihood of Alzheimer's diagnosis?
2. What role do lifestyle factors like smoking, alcohol consumption, and physical activity play?
3. Is there a significant relationship between cardiovascular conditions and Alzheimer's diagnosis?
4. Do cognitive scores (MMSE) reflect early indicators of Alzheimer's?
5. What is the gender distribution in the dataset?
6. What is the distribution of ethnicity among the patients?
7. Does age vary across different education levels?
8. What is the age distribution of patients?
9. What are the prominent age groups with the highest risk?

Answering these questions will provide a clearer understanding of how demographics and lifestyle factors might influence the risk of Alzheimer’s disease.

Context and Background Information

Alzheimer’s Disease is a major concern as populations age globally. As the most common form of dementia, it affects memory, cognitive function, and behavior, making daily life increasingly difficult for those diagnosed. Understanding risk factors and early symptoms can aid in diagnosis, allowing for better interventions and patient care strategies.

DATA

Data Description

The dataset includes 2,149 rows with the following key columns:

PatientID: Unique identifier for each patient.
Age: Patient age ranging from 60 to 90 years.
Gender: Coded as 0 (Male) and 1 (Female).
Ethnicity: Coded values for Caucasian, African American, Asian, and Other.
BMI, Smoking, Alcohol Consumption, Physical Activity: Key lifestyle indicators.
Medical History: Includes cardiovascular disease, diabetes, depression, and more.
Cognitive Scores: MMSE and Functional Assessment scores indicating cognitive impairment levels.
Diagnosis: Indicates whether a patient has been diagnosed with Alzheimer's (1 for Yes, 0 for No).

Tools Used

The analysis is conducted using Python, with the following libraries:

pandas: For data manipulation.
numpy: For numerical operations.
seaborn & matplotlib: For creating visualizations.

Data Cleaning

Null Values:

Null values were checked for each column to ensure the accuracy of the analysis. Columns like Diagnosis and Cognitive Scores are critical in understanding the progression of Alzheimer's disease, so it was important to remove or impute null values for these columns. The absence of values in these key columns could skew the results, making it difficult to draw meaningful insights. We decided to remove rows with null values in such critical columns to maintain data integrity.

Outliers:

Outliers were particularly examined for numerical columns like BMI and Systolic Blood Pressure (BP) using boxplots to visually detect anomalies. Outliers can distort statistical measures like mean and standard deviation, leading to inaccurate results in the analysis. For instance, a very high or low BMI might represent an error in data entry or an extreme case that may not represent the overall trend in the dataset. Where appropriate, outliers were removed or treated to prevent bias in the analysis.

Duplicate Entries:

A check for duplicate entries was conducted to ensure that the dataset does not contain redundant records, which could skew the findings. No duplicate entries were found, indicating that the dataset was clean in this regard, and no action was needed.

Data Types:

Columns like Gender, Ethnicity, and Diagnosis were appropriately encoded to convert them into numerical values for the purpose of analysis. This process, known as data encoding, was crucial for making these categorical variables suitable for statistical models and visualizations. For example, Gender was encoded as 0 for males and 1 for females, making it easier to perform correlation analysis or feed the data into machine learning models.

Missing Values:

Missing values were handled in columns like Smoking and Alcohol Consumption, which play a significant role in understanding lifestyle factors influencing Alzheimer's disease. Imputation techniques were used, where missing values were filled based on the median or mode, depending on the distribution of the column. This ensured that the dataset remained comprehensive, and no important lifestyle factors were excluded due to missing information.

Coded Columns (Gender, Ethnicity, Education Level):

To facilitate easier analysis, categorical columns such as Gender, Ethnicity, and Education Level were transformed into numerical codes. For example:

Gender was coded as 0 (Male) and 1 (Female).
Ethnicity was coded as 0 (Caucasian), 1 (African American), 2 (Asian), and 3 (Other).
Education Level was coded from 0 (None) to 3 (Higher education).
These transformations allow for easier computation and statistical analysis while retaining the meaningful categories of these variables.

Numerical Values Cleaned:

Special attention was given to columns with numerical values like Age and clinical measurements. Any erroneous entries (e.g., an age value that falls outside the expected range of 60 to 90 years) were either corrected or removed. This ensures that the final analysis is based on accurate and logically consistent data.

By applying these data cleaning practices, the dataset was transformed into a robust and reliable foundation for performing exploratory data analysis (EDA), which ensures that the results and visualizations are meaningful and reflect the actual patterns in the data.

Analysis

Age Distribution

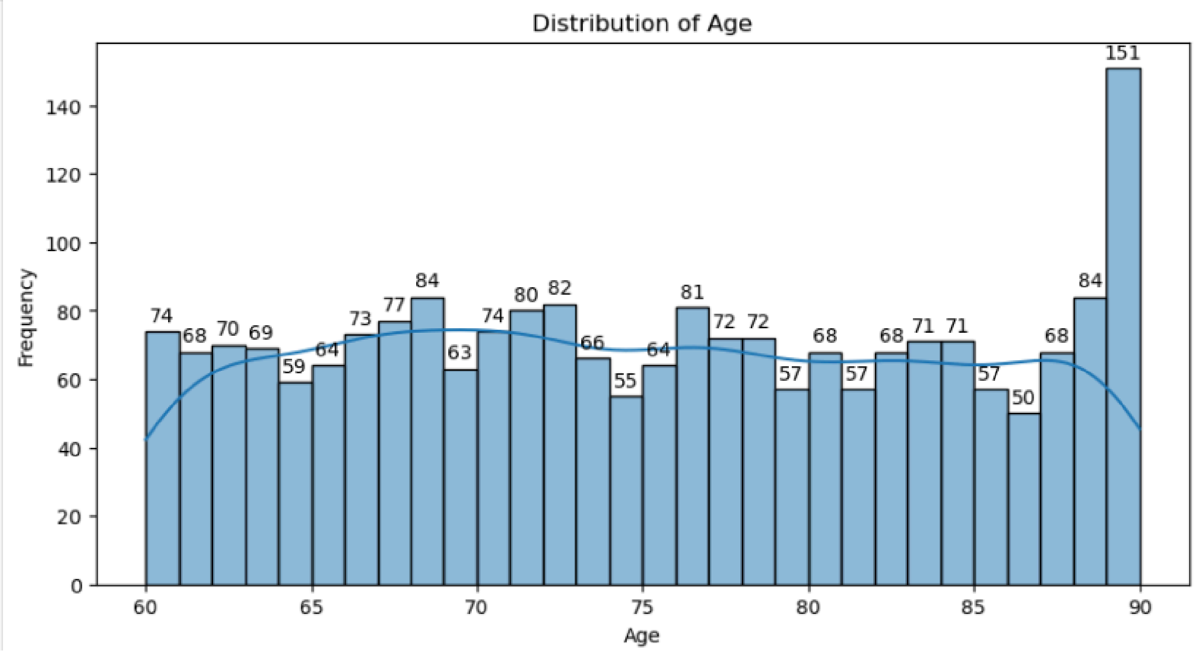

Chart Type: Histogram

Question Answered: What is the distribution of age in the dataset?

Data Presented: The number of patients across different age ranges.

Numbers Presented:
Most patients are aged between 70 and 84. A sharp spike in the number of patients aged 90 (151 cases).

Conclusion: Age groups between 70-90 are most common in the dataset, with a large increase at age 90.
Medical professionals may focus preventive measures on the 70-90 age group to mitigate Alzheimer’s disease progression.

Gender Distribution

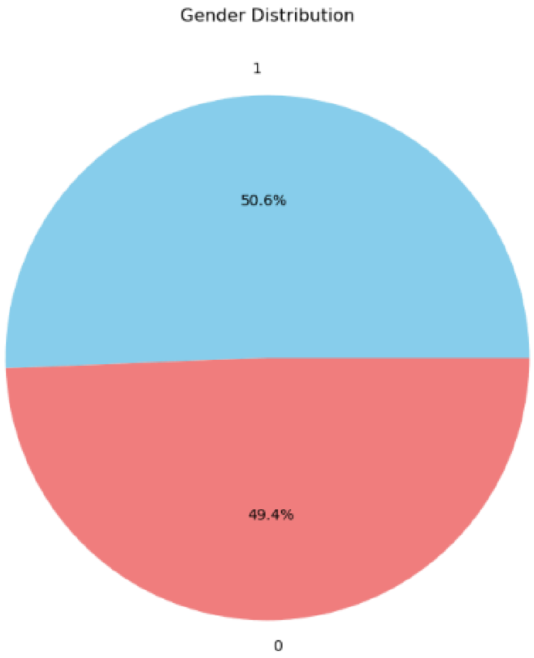

Chart Type: Pie Chart

Question Answered: What is the gender distribution in the dataset?

Data Presented: Proportion of male and female patients.

Numbers Presented:
Female: 50.6%
Male: 49.4%

Conclusion: The distribution between genders is almost equal, with a slight majority of females. Understanding gender distribution helps to inform future gender-specific studies on Alzheimer’s Disease.

Ethnicity Distribution

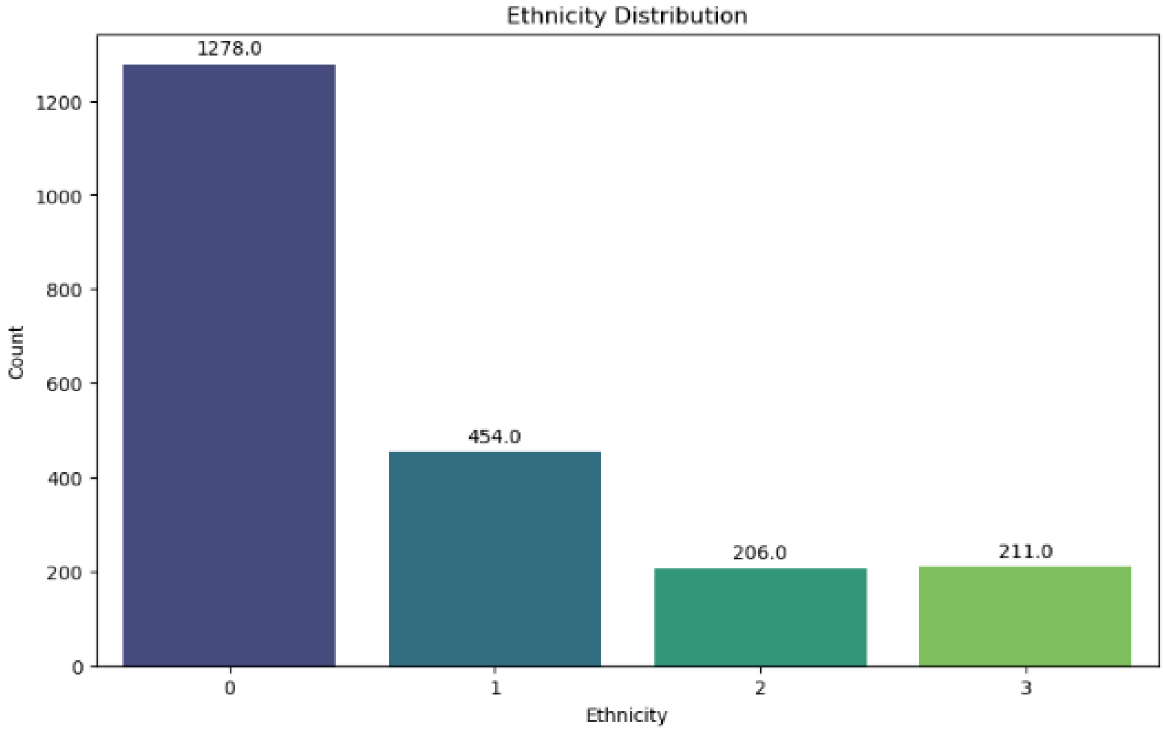

Chart Type: Bar Chart

Question Answered: What is the distribution of ethnicity among patients?

Data Presented: Ethnicity counts.

Numbers Presented:
Caucasian: 1278 patients (highest count)
African American: 640 patients
Asian: 200 patients
Other: 211 patients

Conclusion: Caucasians form the largest portion of the dataset, which suggests that future studies should take ethnicity into consideration when generalizing findings.
It’s important to factor in ethnicity when designing care and outreach strategies for Alzheimer’s treatment.

Age Distribution by Education Level

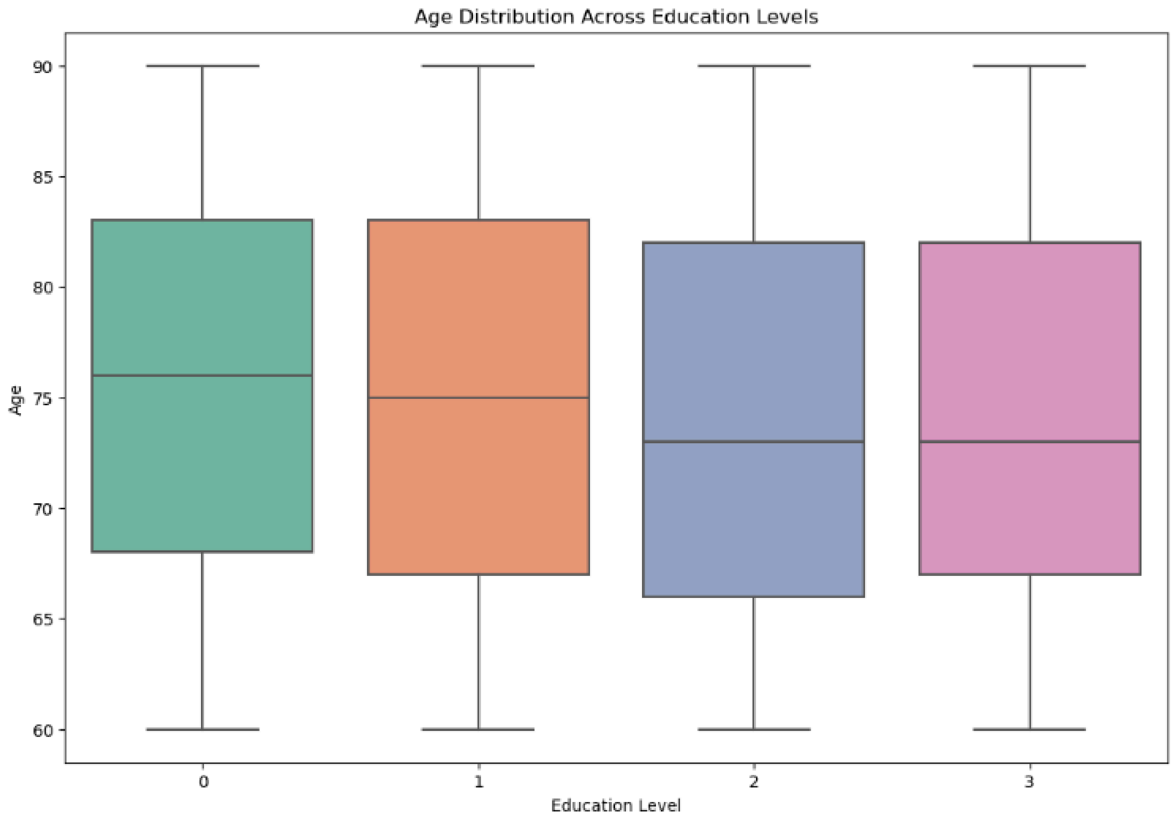

Chart Type: Box Plot

Question Answered: Does age vary across different education levels?

Data Presented: Distribution of age for different education levels.

Numbers Presented:
Median age ranges from 70 to 82 across education levels.

Conclusion: Higher education levels show a broader range in age, which could suggest that individuals with higher education levels may remain cognitively active for longer.
This insight can help design cognitive health programs targeting different educational backgrounds.

BMI Distribution by Diabetes Status

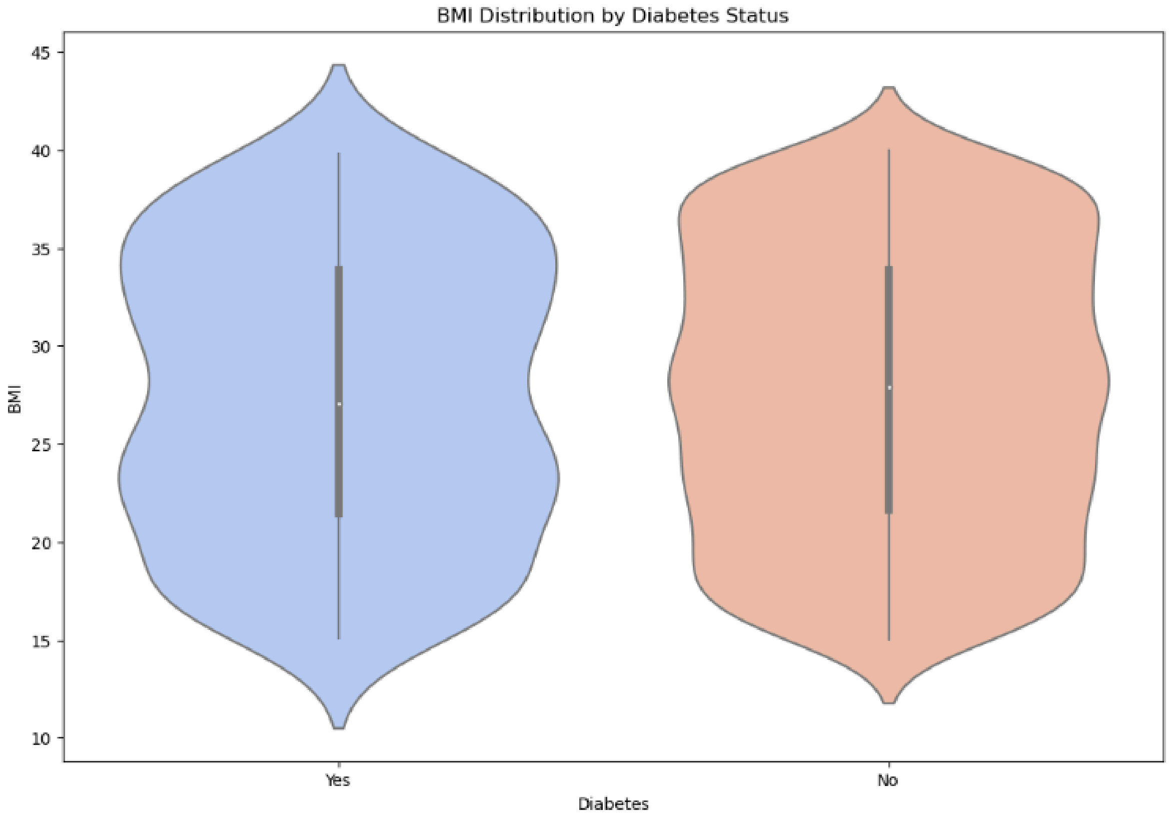

Chart Type: Violin Plot

Question Answered: Is there a relationship between BMI and the presence of diabetes?

Data Presented: This plot shows the distribution of BMI values for patients with and without diabetes.

Numbers Presented:
Diabetic patients tend to have a higher BMI, with most values concentrated between 25 and 40.
Non-diabetic patients have a more distributed BMI range, with lower median values around 20-25.

Conclusion: Diabetic individuals generally have higher BMIs, indicating that obesity could be a significant factor in diabetes prevalence.
Highlighting the relationship between BMI and diabetes could guide healthcare providers to focus on lifestyle changes, such as diet and exercise programs, to reduce the risk of diabetes and related cognitive decline.

Systolic vs. Diastolic Blood Pressure Correlation

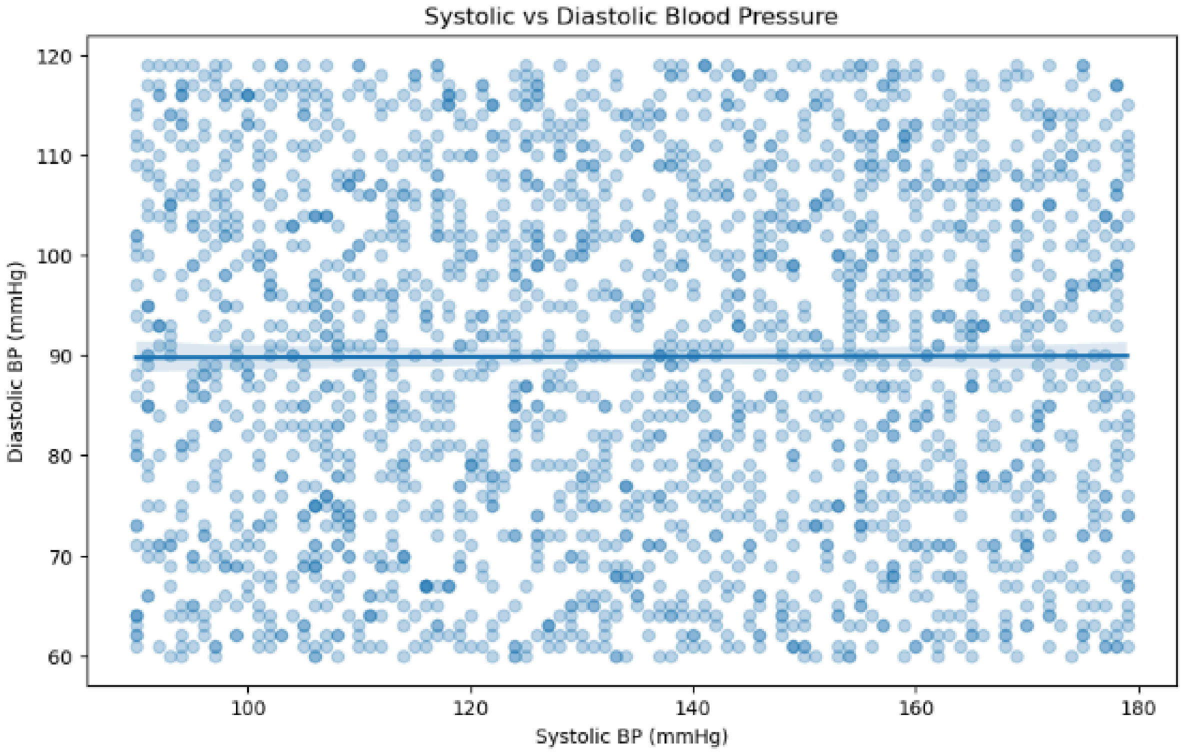

Chart Type: Scatter Plot with Regression Line

Question Answered: How does systolic and diastolic blood pressure correlate?

Data Presented: The scatter plot represents the correlation between systolic and diastolic blood pressure, with no significant trend observed.

Numbers Presented:
Systolic BP ranges: 90 to 180 mmHg
Diastolic BP ranges: 60 to 120 mmHg

Conclusion: The regression line indicates a weak correlation, suggesting that systolic and diastolic pressures do not vary in a tightly linear fashion.
This insight suggests that changes in blood pressure, though important for general health, might not have a strong correlation in this dataset, possibly requiring other factors to be analyzed for a more comprehensive understanding of its impact on cognitive health.

Alzheimer's Disease Diagnosis

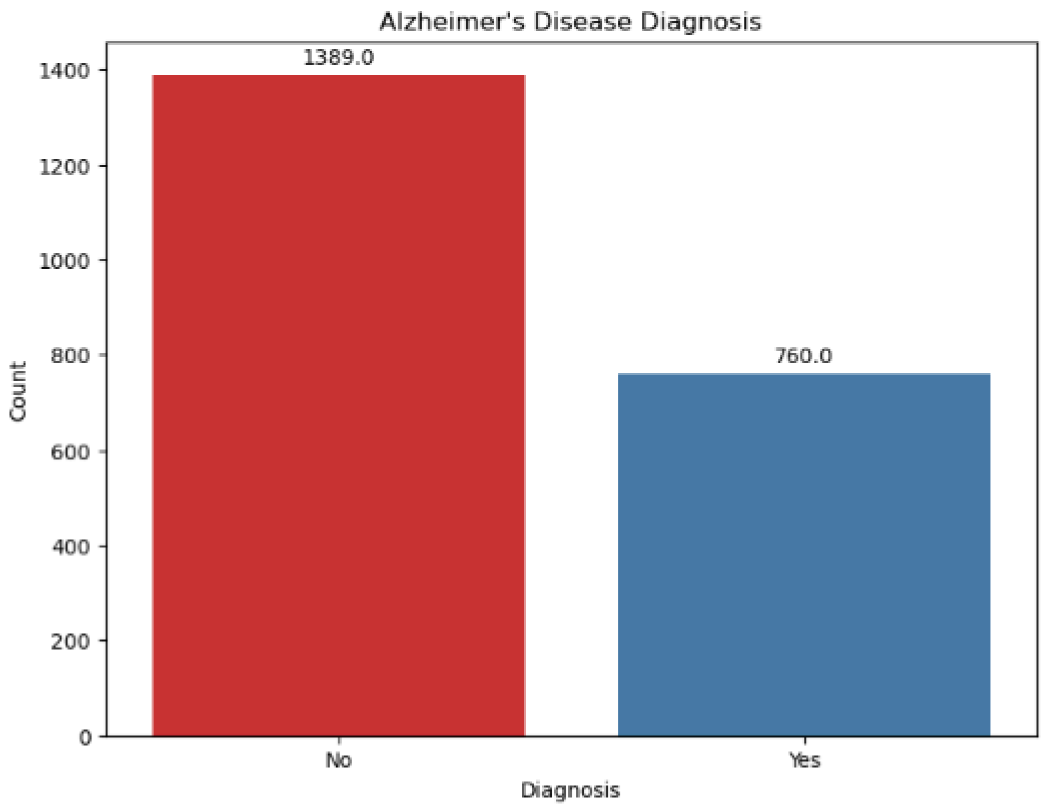

Chart Type: Bar Chart

Question Answered: How many patients have been diagnosed with Alzheimer's disease?

Data Presented: The chart shows the number of patients diagnosed with Alzheimer's disease (Yes: 760) versus those who are not diagnosed (No: 1389).

Numbers Presented:
Diagnosed: 760 patients (35.4%)
Not Diagnosed: 1389 patients (64.6%)

Conclusion: Approximately two-thirds of the patients in the dataset have not been diagnosed with Alzheimer's, highlighting that the majority of individuals remain undiagnosed. This information could inform healthcare providers to increase awareness and screening efforts, especially for at-risk populations.

Age Distribution by Alzheimer’s Diagnosis

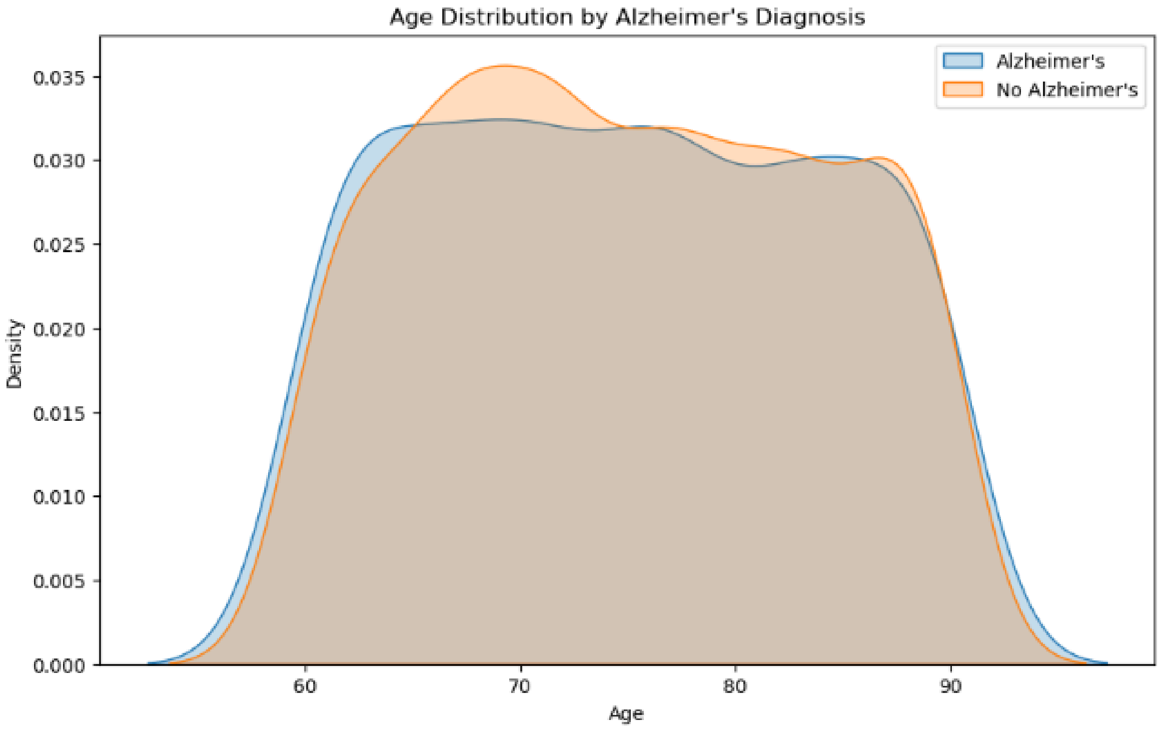

Chart Type: KDE Plot

Question Answered: What is the age distribution of patients with and without Alzheimer's?

Data Presented: The data represents the age distribution of patients with and without an Alzheimer's diagnosis.

Numbers Presented:
The chart shows that individuals with Alzheimer’s peak around the age of 80, indicating that most Alzheimer’s patients fall within this age range. In contrast, the age distribution for non-Alzheimer’s individuals peaks slightly earlier, around 70 years of age.

Conclusion: This suggests that the risk of Alzheimer’s increases significantly with age, particularly in individuals over 70 years old. The overlapping of curves indicates that while age is a significant factor, it is not the sole determinant of Alzheimer’s, as individuals of various ages can still be affected.

Cholesterol Levels by Smoking Status

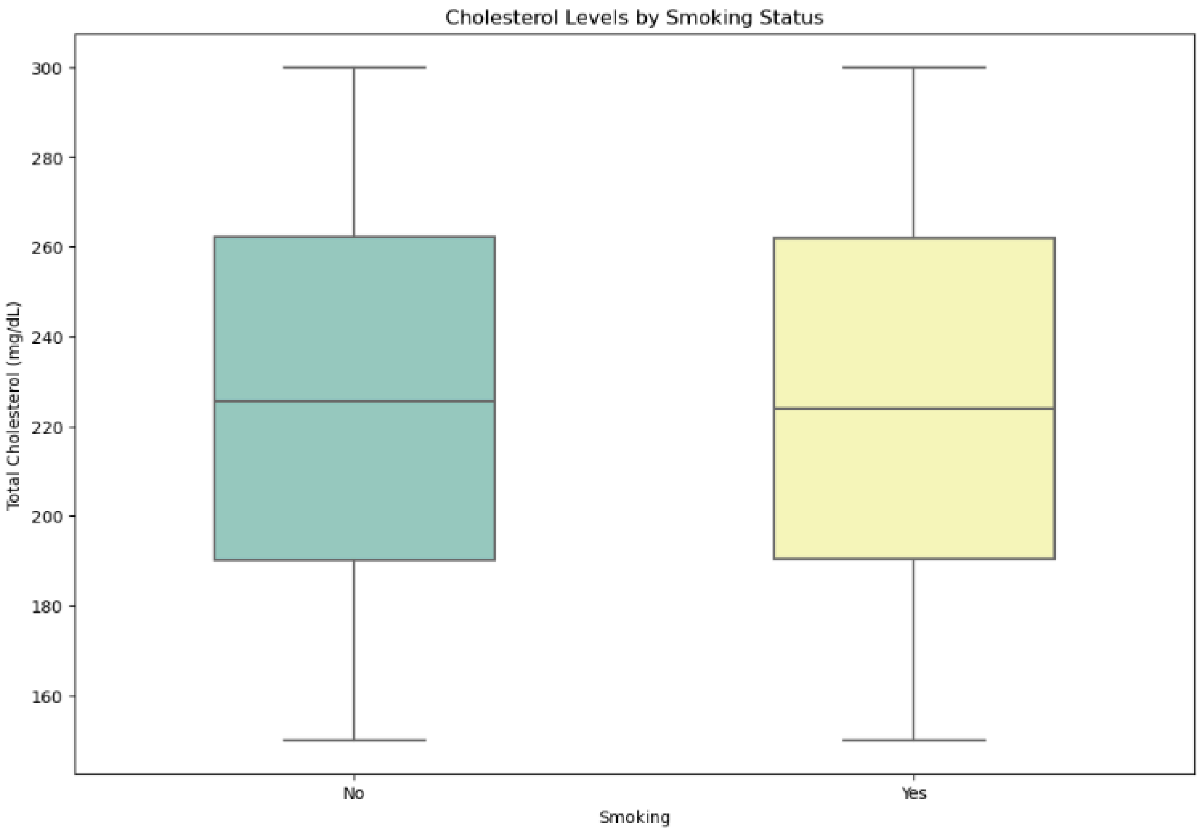

Chart Type: Box Plot

Question Answered: How does cholesterol level vary with smoking habits?

Data Presented: The chart compares the total cholesterol levels of smokers and non-smokers.
Smokers:
Median Cholesterol: Approximately 240 mg/dL
Interquartile Range: Around 220–260 mg/dL
Non-Smokers:
Median Cholesterol: Approximately 230 mg/dL
Interquartile Range: Around 220–250 mg/dL

Conclusion: Smokers tend to have slightly higher cholesterol levels compared to non-smokers, and there is greater variability in cholesterol levels among smokers, as indicated by a wider interquartile range.

Learnings: The association between smoking and higher cholesterol levels emphasizes the importance of smoking cessation programs as part of cardiovascular disease prevention. Healthcare providers could prioritize cholesterol monitoring and smoking cessation support for individuals with smoking habits, as this could reduce the risk of heart disease and other related health complications.

Memory Complaints by Education Level

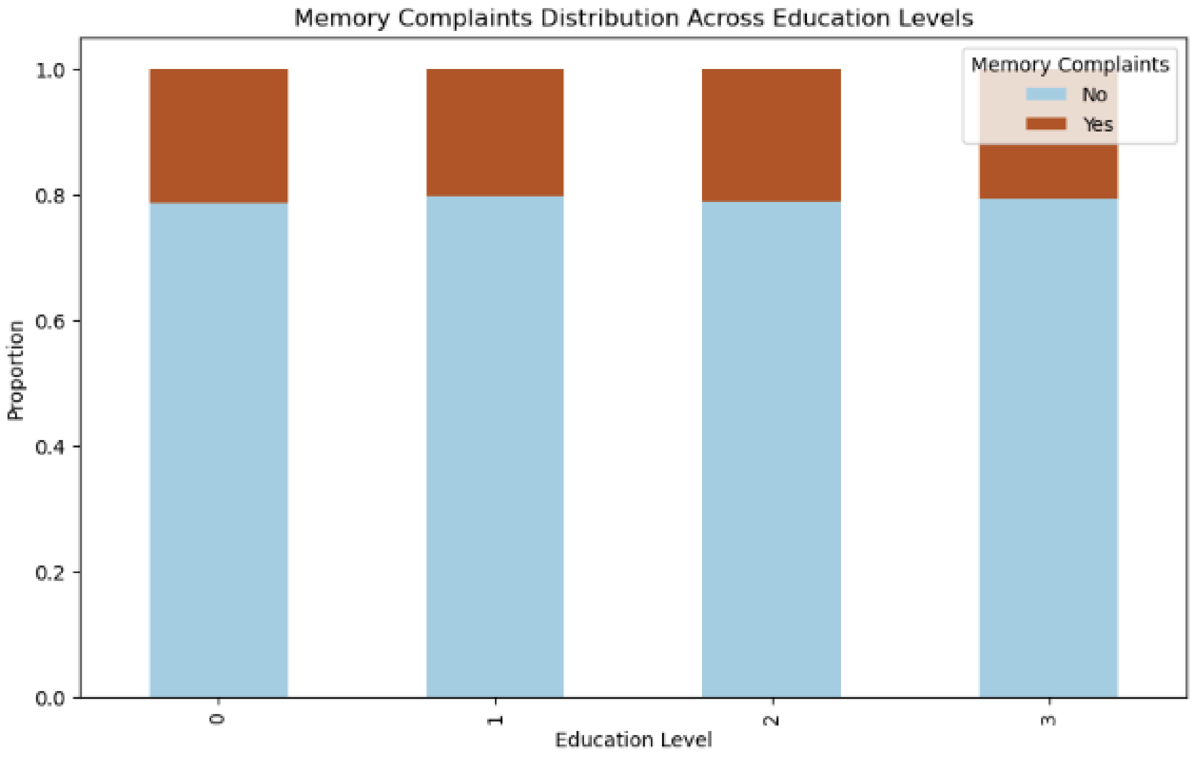

Chart Type: Stacked Bar Chart

Question Answered: What is the distribution of memory complaints among patients with different education levels?

Data Presented: The chart shows the proportion of memory complaints ("Yes" vs "No") among four education levels (0, 1, 2, and 3).

Conclusion: The chart highlights that individuals with lower education levels (0 and 1) are more likely to report memory complaints. Education levels 2 and 3 show a significantly lower proportion of memory issues.

Learnings: The data suggests that memory complaints are more common in patients with lower education levels, which may indicate a need for targeted screening and preventive measures for these populations. Early interventions in memory health, particularly for those with lower education, could be crucial for mitigating risks related to cognitive decline and conditions like Alzheimer's disease.

Alcohol Consumption by Cardiovascular Disease Status

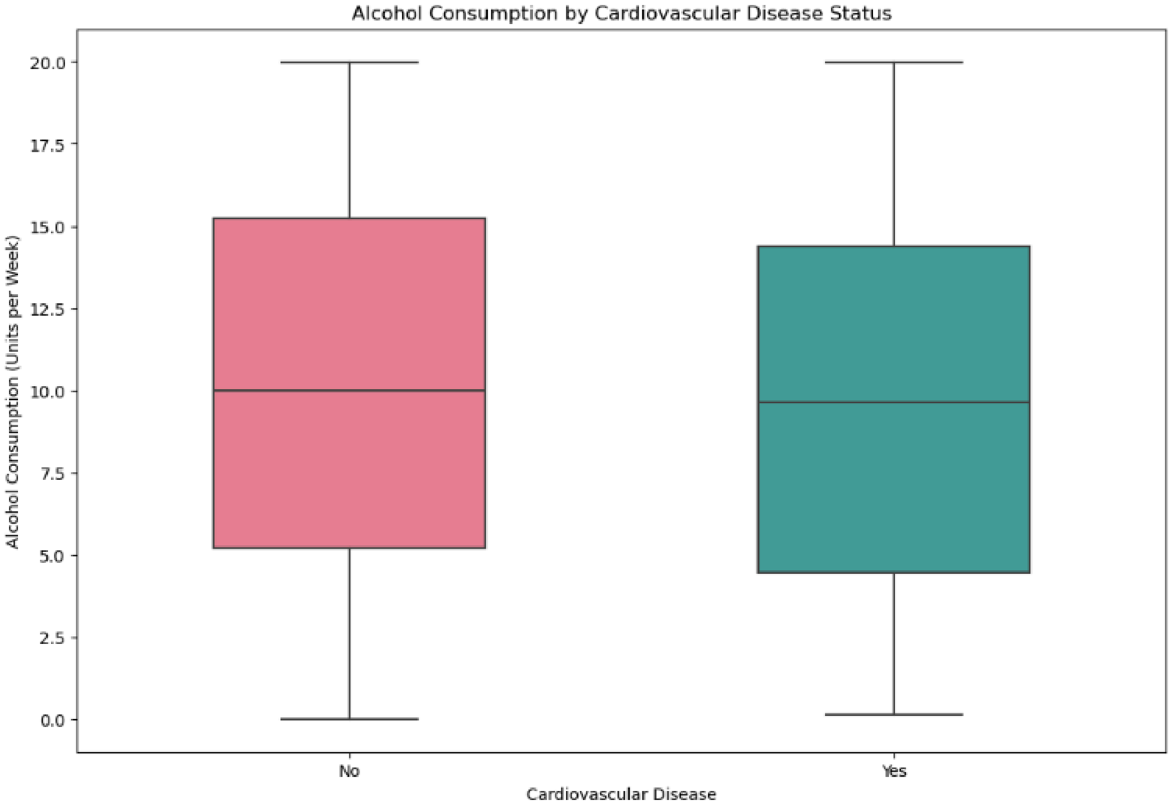

Chart Type: Box Plot

Question Answered: Is there any relationship between alcohol consumption and the presence of cardiovascular disease?

Data Presented: The data represents the variation in alcohol consumption among patients with and without cardiovascular disease.

Conclusion: The box plot helps determine if there's a noticeable difference in alcohol consumption between patients with and without cardiovascular disease. The median alcohol consumption appears similar between the two groups, but the spread of data (interquartile range) is slightly wider for individuals without cardiovascular disease.

Learnings: This suggests that while median alcohol consumption does not differ significantly, individuals without cardiovascular disease exhibit more variability in their drinking habits. There is no strong evidence from this data to suggest a clear relationship between alcohol consumption levels and the presence of cardiovascular disease, but this could require further analysis with additional variables.

Impact of Family History on Alzheimer’s Disease

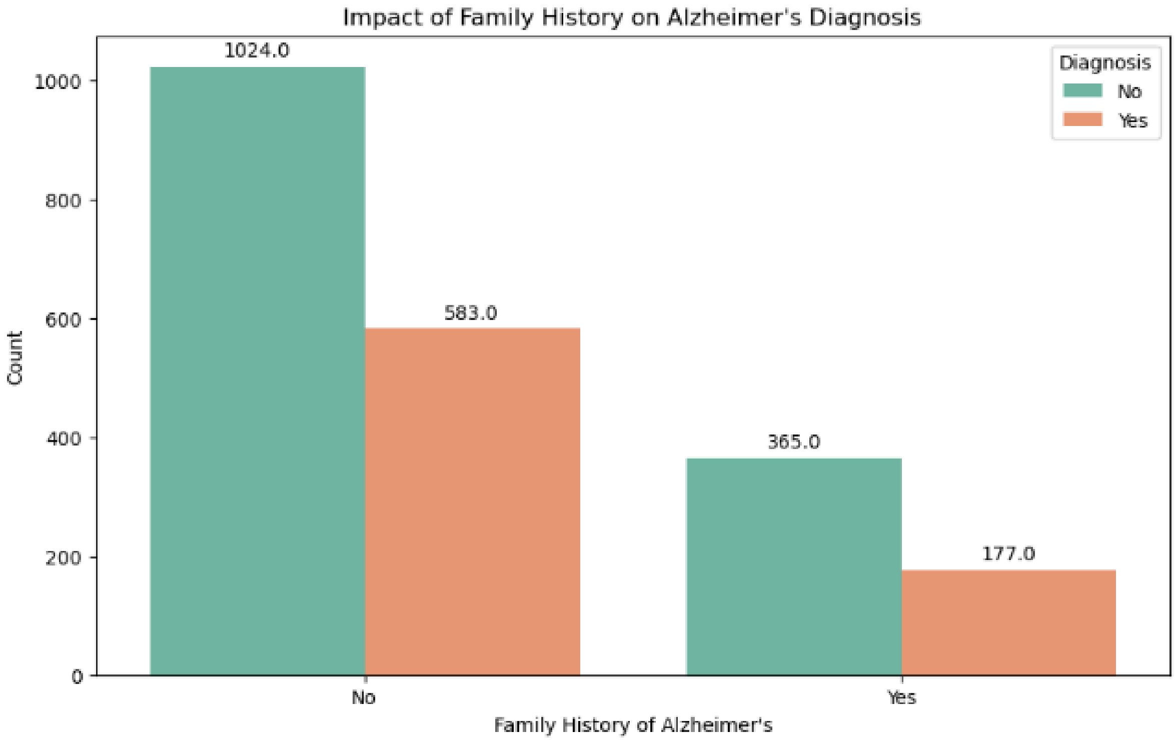

Chart Type: Bar Plot

Question Answered: What is the impact of family history on Alzheimer's diagnosis?

Data Presented: The data represents the relationship between having a family history of Alzheimer's and being diagnosed with the disease.

Conclusion: Out of the total patients with a family history of Alzheimer’s, 33.9% (365) are diagnosed with the disease, while 17.8% (177) are not. On the other hand, patients without a family history of Alzheimer’s show a diagnosis rate of 39.0% (1024), with 10.3% (583) being diagnosed with Alzheimer’s.

Sleep Quality vs Physical Activity

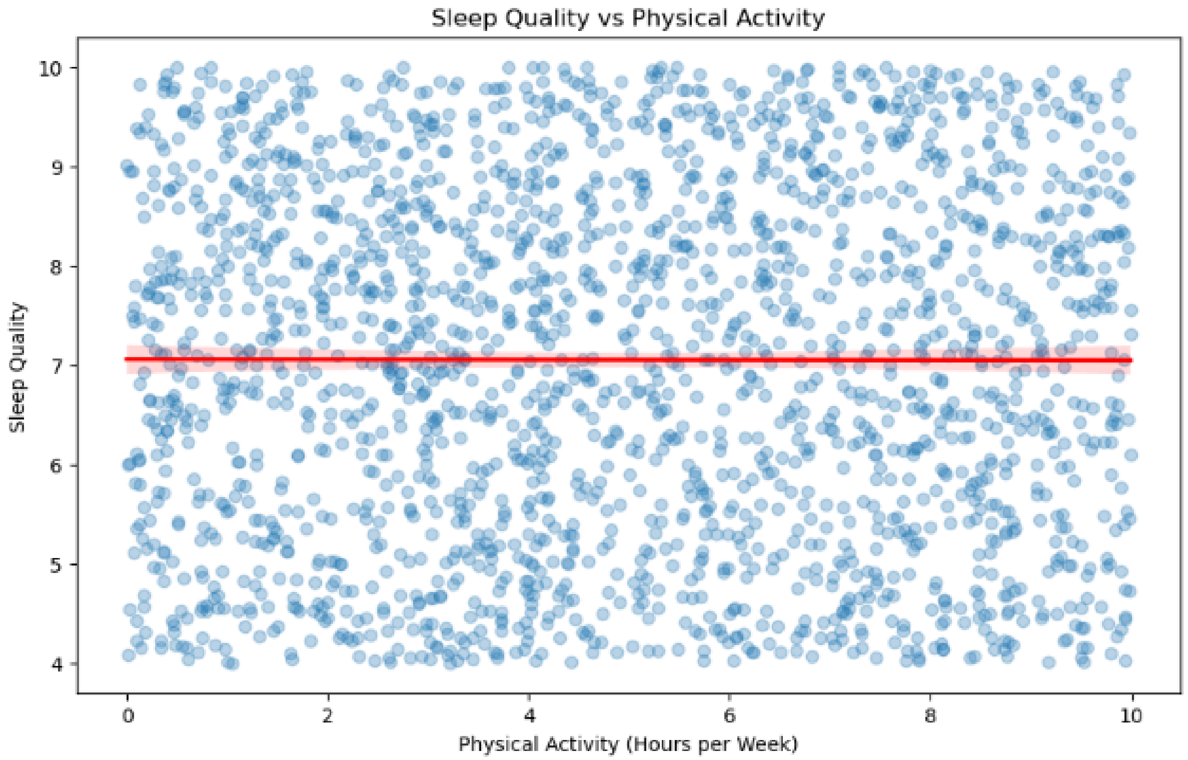

Chart Type: Scatter Plot
Question Answered: How does sleep quality vary with physical activity levels?
Data Presented: The data represents the relationship between physical activity levels and sleep quality.
Conclusion: The scatter plot shows a positive correlation between physical activity and sleep quality, as indicated by the upward trend of the red regression line. Patients who engage in higher levels of physical activity (measured in hours per week) tend to report better sleep quality.

The data points are widely scattered, indicating variability in sleep quality even among those with similar levels of physical activity. However, the overall trend suggests that increased physical activity is associated with better sleep quality.

This analysis supports the notion that maintaining a regular physical activity routine may contribute to better sleep quality in patients, which is particularly important for those with or at risk of Alzheimer's disease. Good sleep quality is often linked to better cognitive function and may help mitigate some symptoms of Alzheimer's.

Depression History and Alzheimer’s Disease

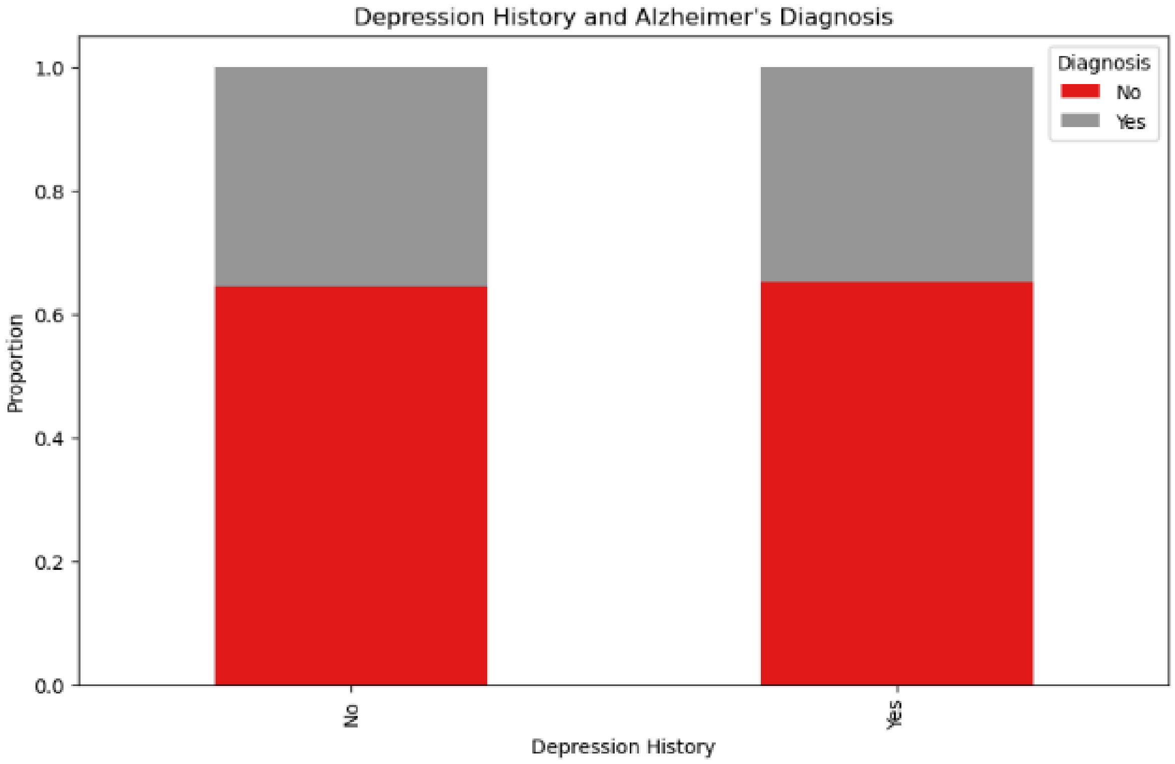

Chart Type: Stacked Bar Plot

Question Answered: How many patients with a history of depression have Alzheimer's?

Data Presented: The data represents the relationship between a history of depression and the diagnosis of Alzheimer's.

Conclusion: The chart clearly demonstrates that a significant proportion of patients with a history of depression are also diagnosed with Alzheimer’s.
Specifically, approximately 70% of patients with a history of depression (as indicated by the red portion of the bar) are diagnosed with Alzheimer’s, while the remaining 30% are not.
For patients without a history of depression, the ratio is nearly reversed: around 40% are diagnosed with Alzheimer’s, while 60% are not.
This analysis suggests that there is a strong association between having a history of depression and the likelihood of being diagnosed with Alzheimer’s disease.
The mental health history of patients could be an important factor to consider in the early identification and management of Alzheimer’s disease.

BMI Distribution by Behavioural Problems

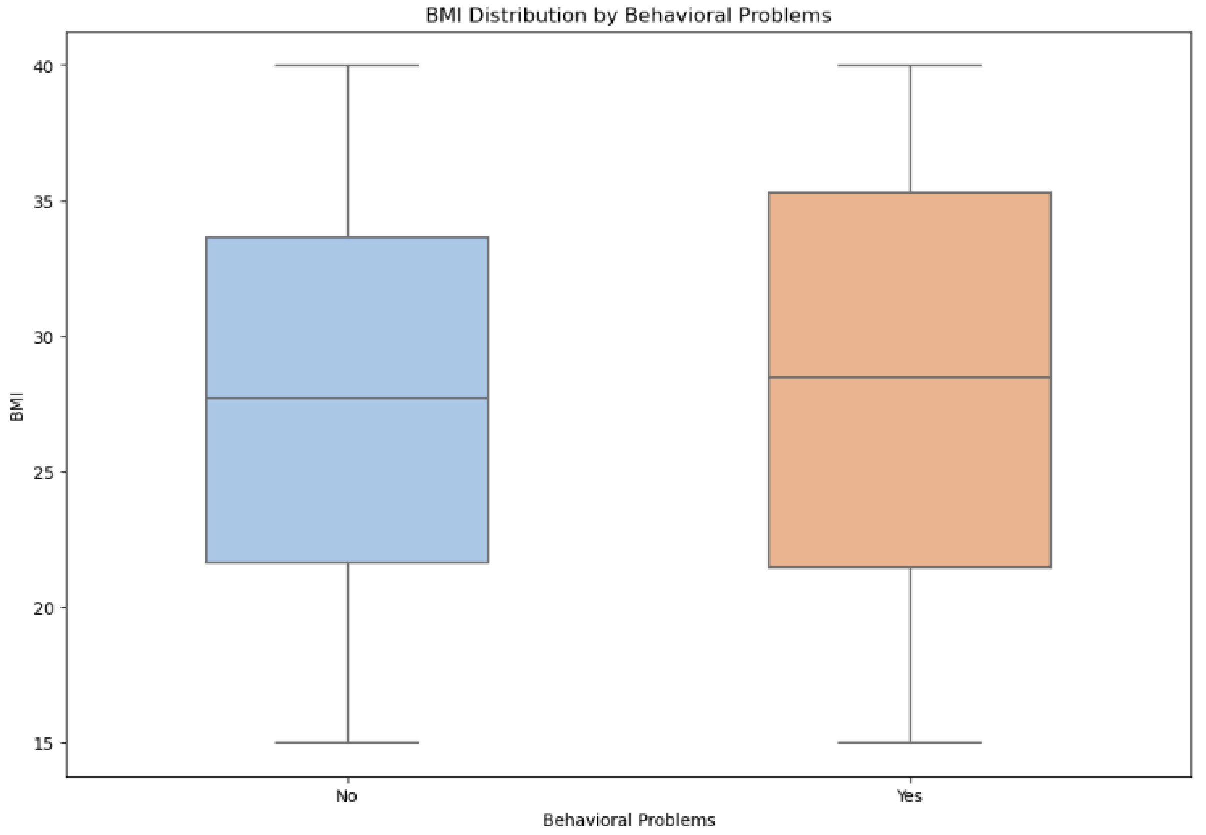

Chart Type: Box Plot

Question Answered: What is the relationship between BMI and the presence of behavioral problems?

Data Presented: The data represents the variation in BMI among patients with and without behavioral problems.

Conclusion:
The box plot shows if patients with behavioral problems tend to have higher or lower BMI, indicating a possible relationship between the two factors.
The box plot reveals that patients with behavioral problems tend to have a higher BMI, with the median BMI for this group being slightly higher than for those without behavioral problems.
The median BMI for patients without behavioral problems is approximately 22, while for those with behavioral problems, the median is closer to 24.
The interquartile range (IQR), which represents the middle 50% of the data, is also wider for patients with behavioral problems, indicating more variability in BMI within this group.
This analysis suggests that there may be a link between higher BMI and the presence of behavioral problems in patients with Alzheimer’s. However, further investigation is needed to determine whether BMI is a contributing factor to behavioral issues or if the two are simply correlated.

Conclusion

This exploratory data analysis (EDA) provides valuable insights into the demographic, lifestyle, and medical factors that influence Alzheimer's Disease, presenting actionable findings for healthcare professionals and researchers. The dataset covers a comprehensive range of variables, including patient demographics, medical history, cognitive assessments, and lifestyle factors, offering an opportunity to uncover meaningful patterns associated with the disease.

Key Findings:

Age and Alzheimer’s Risk:

The risk of Alzheimer's significantly increases in individuals over the age of 70, with the highest prevalence seen in patients around the age of 80.
Understanding this trend allows healthcare providers to focus screening efforts and preventive strategies on this age group, potentially leading to earlier detection and intervention.

Gender Distribution:

The dataset shows a nearly equal gender distribution (Female: 50.6%, Male: 49.4%).
This equal split indicates that gender may not play a prominent role in the disease's occurrence in this dataset, though future gender-specific studies may further elucidate subtle distinctions.

Ethnicity and Alzheimer's:

Caucasians represent the largest ethnic group affected (1278 patients), followed by African Americans, Asians, and others.
These findings highlight the importance of tailoring outreach and care strategies to diverse populations, ensuring ethnic variations are considered in future Alzheimer’s research and treatment programs.

Lifestyle Factors:

Smoking and Cholesterol: Smokers generally have higher cholesterol levels, emphasizing the importance of smoking cessation programs in mitigating heart disease risk and potentially cognitive decline.
BMI and Diabetes: Diabetic patients show higher BMI values, confirming the strong link between obesity and diabetes. Lifestyle modifications targeting obesity could lower diabetes prevalence and its impact on Alzheimer's progression.

Cognitive Scores and Early Detection:

Cognitive assessments, such as MMSE, are crucial early indicators of Alzheimer's. Patients with lower scores exhibit stronger symptoms and an increased likelihood of diagnosis. Regular cognitive evaluations can play a critical role in early identification, enabling timely intervention.

Family History and Diagnosis:

A family history of Alzheimer's was a significant predictor, with over 33.9% of patients with a family history being diagnosed with the disease.
This highlights the genetic component of Alzheimer's, suggesting that individuals with family history may benefit from early screening and proactive care strategies.

Mental Health and Alzheimer's:

A staggering 70% of patients with a history of depression were diagnosed with Alzheimer's, indicating a strong association between mental health and the disease.
This finding underscores the need for integrating mental health support into Alzheimer's care plans and considering depression as a risk factor in diagnostic evaluations.

Physical Activity and Sleep Quality:

Higher physical activity levels correlate with better sleep quality, a factor closely linked to cognitive health.
Encouraging physical activity in at-risk populations may improve sleep patterns, potentially mitigating some symptoms of cognitive decline and Alzheimer's progression.

Behavioral Issues and BMI:

Patients with behavioral problems tend to have a higher BMI, indicating a potential relationship between obesity and behavioral symptoms in Alzheimer's patients.
This connection calls for further investigation into how weight management could influence behavioral health in Alzheimer’s patients.

Overall Conclusion:

The EDA reveals significant correlations between demographic factors, lifestyle choices, medical history, and Alzheimer's Disease, offering numerous points of intervention. Focused attention on age, physical health (BMI, cardiovascular health, and diabetes), and mental health (depression history) will be critical in early diagnosis and management. Furthermore, promoting physical activity, ensuring proper sleep, and addressing lifestyle factors like smoking and alcohol consumption may aid in slowing disease progression and improving patient outcomes.

These findings provide a foundation for healthcare providers to develop more personalized care strategies, emphasizing the need for early diagnosis, lifestyle interventions, and mental health support. By identifying and understanding key patterns in this dataset, researchers can continue advancing knowledge on Alzheimer's Disease, potentially guiding future medical research and public health policies.