```
In [ ]:
```

Module_1: Alzheimer's Data Analysis

Team Members:

Jaya Kinley (nzh2yy), William Crouch (ken8gq)

Project Title:

Investigating the Relationship Between the Presence of the APOE4 gene and MMSE Score in Alzheimer's Patients

Project Goal:

This project seeks to investigate the relationship between the presence of the APOE4 gene and MMSE scores in patients with Alzheimer's disease.

Disease Background:

Fill in information about 11 bullets:

- Prevalence & incidence (https://www.alz.org/alzheimers-dementia/facts-figures#:~:text=in%20each%20state-,Quick%20facts,1%20in%2010%20for%20men.)
    - Domestically, Alzheimer's has a prevalence of seven million Americans, with one in nine Americans over 65 diagnosed.
- Economic burden (https://www.nature.com/articles/s41514-024-00136-6)
    - The health care costs of treating these patients is $450bntodayandprojectedtogrowtoover$3.3 tn by 2060.
- Risk factors (genetic, lifestyle) (https://www.nia.nih.gov/health/alzheimers-causes-and-risk-factors/what-causes-alzheimers-disease)
    - Alzheimer's has been correlated with some genetic markers, such as APOE, a gene that is involved in the production of proteins that transport fat and cholesterol in the bloodstream. Other factors such as TBIs, poor vascular health, physical inactivity, smoking, and lack of mental stimulation have been correlated with an increased incidence of AD. The following mutations also significantly increase the likelihood of developing AD.
        - Amyloid precursor protein (APP) on chromosome 21
        - Presenilin 1 (PSEN1) on chromosome 14
        - Presenilin 2 (PSEN2) on chromosome 1

- Societal determinants (https://www.cdc.gov/alzheimers-dementia/php/sdoh/index.html)
  - AD is seen at a higher incidence in adults 45 and older who have a lower level of education, with twice the incidence rate for adults without a high school diploma. People without consistent access to healthcare are also at an elevated risk, as other treatable issues, such as those with the circulatory system, remain untreated and increase the risk of developing AD.
- Symptoms
  - Memory loss, confusion, difficulty thinking or concentrating, poor planning, and frequently losing things
- Diagnosis
  - Diagnosing Alzheimer’s generally involves cognitive tests, as well as lab work and brain imaging to measure levels of tau and investigate potential brain shrinkage. The diagnosis process starts with simple mental status testing of memory and critical thinking, and will progress to more advanced and specialized methods such as functional MRIs if these tests indicate a likelihood of AD.
- Standard of care treatments (& reimbursement) (https://www.nhs.uk/conditions/alzheimers-disease/treatment/#:~:text=Cognitive%20stimulation%20therapy%20(CST)%20involves
  - Cognitive stimulation therapy (CST): involves taking part in group activities and exercises designed to improve memory and problem-solving skills
  - Cognitive rehabilitation: working with a profession (e.g. occupational therapist) and a relative or friend to achieve personal goals and everyday tasks. Prompts you to use parts of your brain that are working to help the parts that aren't.
  - Reminiscence and life story work: talking about things and events from your past using props like photos, possessions, music, notes, keepsakes, etc.
  - Acetylcholinesterase (AChE) inhibitors: increase levels of acetylcholine, a substance in brain that helps nerve cells communicate. Currently being prescribed by specialists such as psychiatrists or neurologists, may be prescrbed by GP on advice of specialist. Medicines include Donepezil, galantamine, and rivastigmine (prescriped for people with early- to mid-stage Alzheimer's)
  - Memantine: blocks effects of excessive amount of glutamine. Used for moderate to severe Alzheimers.
- Disease progression & prognosis (https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-stages/art-20048448)
  - Preclinical Alzheimer's: asymptomatic, can last for years, amyloid plaques and neurofibrillary tangles developing (biomarkers)
  - Mild cognitive impairment (MCI) due to Alzheimer's: mild changes in memory and thinking ability, aren't enough to affect work or relationships, just forgetting info that isn't easily remembered (e.g. convos or recent events), may have

    - trouble judging amount of time needed for a task, number of steps needed to complete a task, trouble making decisions.
    - Mild demntia due to Alzheimer's: often diagnosed here, when it becomes clear to family and doctors that a person has significant trouble with memory and thinking, affects daily functioning, people may experience memory loss of recent events, trouble with problem-solving and complex tasks, changes in personality, trouble organizing or expressing thoughts, getting lost, and misplacing belongings.
    - Moderate dementia due to Alzheimer's: more confused and forgetful, begin to need help with daily activities and self care. May show increasingly poor judgement and deepening confusion (losing track of where they are, day, week, season, people), experience greater memory loss, and undergo significant changes in personality and behavior (developing unfound suspicions, restless, agitated, aggressive).
    - Severe dementia due to Alzheimer's disease: mental function continues to decline, more issues with movement, lose ability to communicate, require daily assistance with personal care, experience decline in physical abilities (may become unable to walk without assistance, sit up, hold their head, muscles may become rigid and reflexes don't respond in usual way, eventual loss of ability to swallow, control bladder and bowel functions).
    - Rate of progression through stages: varies - on average, people live between 3 and 11 years after diagnosis, how advanced it was by time of diagnosis can impact life expectancy; pneumonia is a common cause of death (trouble swallowing due to pneumonia allows food or beverages to enter lungs where an infection can begin), others include dehydration, malnutrition, falls, and other infections.
- Continuum of care providers (https://www.alzheimers.gov/professionals/health-care-providers)
    - primary care providers
    - neurologists
    - geriatricians
    - geriatric psychiatrists
    - neuropsychologists
    - clinical psychologists, social workers, and general psychiatrists
    - speech, physical, and occupational therapists
    - nurses
    - home health aids
- Biological mechanisms (anatomy, organ physiology, cell & molecular physiology) (https://www.nia.nih.gov/health/alzheimers-causes-and-risk-factors/what-happens-brain-alzheimers-disease)
    - Communication: when neurons receive signals from other neurons, generates charge that travels down axon and releases neurotransmitter chemicals across

synapses. Each neurotransmitter molecule binds to specific receptor sites on a dendrite of nearby neuron, which triggers chemical or electrical signals that stimulate or inhibit activity in neuron receiving signal.

    - metabolism: in people with Alzheimer's, reduction in glucose entering brain and decrease in energy production that can affect neurons due to their high energy needs.
    - glial cells: types of brain cells that provide physical and chemical support to neurons (microglia, astrocytes, oligodendrocytes).
        - microglia protect neurons from physical and chemical damage and are responsible for clearing foreign substances and cellular debris from the brain.
        - astrocytes are glial cells with important metabolic, structural, regulatory, and protective functions.
        - both microglia and astrocytes are involved in immune response in the brain.
    - In Alzheimer's damage is widespread, many neurons stop functioning properly, lose connections with other neurons, and eventually die. Disrupts processes vital to communication, metabolism, and repair.
        - at first, Alzheimer's damages connections among neurons in parts of brain involved with memory, including entorhinal cortex and hippocampus. Later affects areas in cerebral cortex responsible for language, reasoning, social behavior. Eventually many other areas of brain and surrounding neurons are dmaaged and stop working normally.
    - Amyloid plaques: beta0amyloid formed from breakdown of larger protein called amyloid precursor. Comes in several different molecular forms that collect between neurons. Abnormal levels of naturally occurring protein clumb together to form plaques that disrupt cell function.
    - Neurofibrillary tangles: abnormal accumulations of a protein called tau that collect inside neurons. In healthy neurons, tau binds to and stabilizes microtubules (which help guide nutrients and molecules from cell body to axon and dendrites). In Alzheimer;s, abnormal chemical changes cause tau to detach from microtubules and stick to other tau molecules, forming tangles inside neurons that block its transport system and harms the synaptic communication between neurons.
    - as neurons are injured and stop working, connections among neurons break down and many regions of the brain begin to shrink. By final stages of Alzheimer's, this process (brain atrophy) is widespread, resulting in significant cell death and causing loss of brain volume.
    - chronic inflammation may be caused by buildup of harmful secretions of malfunctioning glial cells. When microglia fail to clear away waste, debris, protein collections (e.g. beta-amyloid plaques), Alzheimer's can develop.
- Clinical Trials/next-gen therapies (https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-treatments/art-

20047780#:~:text=Keeping%20tau%20from%20tangling,Researching%20insulin%20re

Targeting clumps of beta-amyloid (plaques):

- Recruiting the immune system:
  - Medicines known as monoclonal antibodies may prevent beta-amyloid from clumping into plaques and remove plaques that have formed by helping the body clear them from the brain
  - Mimic antibodies your body naturally produces as part of the immune system's response to foreign invaders or vaccines. FDA has approved lecanemab (Leqembi) and donanemab (Kisunla) for people with mild Alzheimer's and mild cognitive impairment.
  - Clinical trials found that the medicines slowed declines in thinking/functioning in people with early Alzheimer's and prevented amyloid plaques in brain from climping.
  - Given as infusions
- Preventing distruction:
  - medicine initially developed as possible cancer treatment (saracatinib)
  - in mice, it turned off a protein that allowed synapses (small spaces between brain cells through which cells communicate) to start working again --> animals experienced reversal of some memory loss
  - human trials are underway
- Production blockers:
  - reduce amount of beta-amyloid formed in brain
  - beta-amyloid produced from "parent protein" in brain in two steps performed by different enzymes
  - experimental medicines aim to block activity of parent enzymes (beta- and gamma-secretase inhibitors)
  - studies showed that beta-secretase inhibitors didn't slow cognitive decline and were associated with significant side effects for those with mild or moderate Alzheimer's.
- Keeping tau from tangling:
  - Tau aggregation inhibitors and tau vaccines being studied in clinical trials
- Reducing Inflammation:
  - Alzheimer's causes chromic, low-level brain cell inflammation
  - medicine sargrramostim (Leukine) is currently in research
  - may stimulate immune system to protect the brain from harmful proteins
- Researching insulin resistance:
  - researchers studying how insulin changes in brain may be related to Alzheimer's.
  - trial testing of an insulin nasal spray determined medicine wasn't effective in slowing progression of Alzheimer's.
- Studying heart-head connection:

- risk of developing dementia appears to increase as result of many conditions that damage heart or arteries (high blood pressure, heart disease, stroke, diabetes, and high cholesterol)
- studies exploring how best to build on connection: strategies include...
- research about whether blood pressure medicines benefit people with Alzheimer's/may reduce risk of dementia
- studies looking at how connection between heart disease and Alzheimer's works at molecular level to find new potential medicines
- lifestyle choices with known heart benefits may help prevent Alzheimer's or delay onset (exercising and eating heart-healthy diet)

Data-Set:

Data was acquired from the study Integrated multimodal cell atlas of Alzheimer’s disease (Gabitto et al., 2024).

Quantitative neuropathological assays were performed on postmorten donors obtained from the Adult Changes in Thought (ACT) Study and the University of Washington Alzheimer’s Disease Research Center (ADRC). The study cohort includes all ACT precision rapid autopsies and UW ADRC Clinical Core autopsies, with exclusion of those with a diagnosis of frontotemporal dementia (FTD), frontotemporal lobar degeneration (FTLD), Down's syndrome, amyotrophic lateral sclerosis (ALS) or other confounding degenerative disorder (not including Lewy Body Disease or uVBI). The cohort also excludes individuals that tested positive for COVID-19. The cohort represents the full spectrum of Alzheimer’s disease severity (Gibitto et al., 2024).

The MetaData file specifies the APOE genotype for 84 subjects, expressed as allele combinations (3/3, 3/4, 2/3, 4/4, 2/4, 2/2) with the mode being 3/3 (ε3/ε3), and the least frequent being 2/2 (ε2/ε2). We investigated specifically patients that had the APOE4 gene, which is a significant genetic risk factor for Alzheimer's disease.

The MetaData file also contains MMSE score (a score ranging from 0-30 in which higher numbers indicate better cognitive function) that details the last MMSE the subject received, and the interval from the last MMSE (months). The last MMSE score data has a mean of 24.65 across 80 subjects, with a range of 27. The data for the length of time between death and last MMSE has a mean of 36.63 months, and a range of 115.8 months.

Source: Gabitto, M.I., Travaglini, K.J., Rachleff, V.M. et al. Integrated multimodal cell atlas of Alzheimer’s disease. Nat Neurosci 27, 2366–2383 (2024). https://doi.org/10.1038/s41593-024-01774-5

Data Analyis:

We began our data analysis by examining several key questions, including the link between education and AD onset, as well as its severity. Our final outputs from this model, however, diverged significantly from the current scientific consensus, showing an increase in AD in higher-educated populations, such as those with Ph. D.s. From here, we sought to leverage the unique attributes of this dataset, namely the brain processing for tau and beta levels, and plot that against CASI scores. However, this data was too noisy and lacked sufficient quantity to draw any meaningful conclusions. Finally, we examined the genotypes and plotted whether patients had the APOE4 allele against their MMSE scores, finding a marked decrease in MMSE scores among carriers of the allele.

The code below imports patient-level data from two CSV files (**Luminex data** and **metadata**), creates a `Patient` class to store relevant attributes (MMSE, APOE genotype, tau and beta levels, demographics), and ensures that donor IDs are matched correctly across datasets.

It filters out rows with missing **MMSE** or **APOE** data to maintain clean comparisons. The analysis then:

- **Separates patients** into APOE4+ and APOE4– groups
- **Performs descriptive statistics** (means, standard deviations)
- **Runs an independent t-test** to assess whether group means differ significantly
- **Generates visualizations** including:
  - Bar plots comparing MMSE means between groups
  - Jittered scatterplots with transparent boxplot overlays to show distribution and individual variability
  - Regression plots of MMSE vs. Age at Death with the line of best fit and (R^2) value

`patient.py`

Defines a `Patient` class to:

- Load patient biomarker data (`ABeta40` , `ABeta42` , `tTau` , `pTau`) from a main CSV.
- Merge additional metadata (sex, death age, CASI, MMSE, APOE genotype, etc.) from a second CSV by **Donor ID**.
- Store all patients in `Patient.all_patients` for easy iteration and analysis.
- Provide helper methods for filtering, sorting, and grouping by attributes (e.g., education level, Thal score).

In [2]:

```
import csv

class Patient:
    all_patients = []

    def __init__(self, donor_id, abeta40=None, abeta42=None, ttau=None, ptau
                 apoe=None, mmse=None, casi=None, sex=None, death_age=None):
        self.DonorID = donor_id
        self.ABeta40 = abeta40
        self.ABeta42 = abeta42
        self.tTAU = ttau
        self.pTAU = ptau
        self.APOE = apoe
        self.MMSE = mmse
        self.CASI = casi
        self.sex = sex
        self.death_age = death_age

    def __repr__(self):
        return (f"Patient(DonorID={self.DonorID}, APOE={self.APOE}, "
                f"MMSE={self.MMSE}, CASI={self.CASI}, "
                f"ABeta42={self.ABeta42}, tTAU={self.tTAU}, pTAU={self.pTAU}

    @classmethod
    def instantiate_from_csv(cls, luminex_path, metadata_path):
        """Load patients from Luminex + Metadata CSV, matching by Donor ID a
        cls.all_patients.clear()

        # --- Load Luminex data ---
        luminex_data = {}
        with open(luminex_path, newline='', encoding="utf-8-sig") as csvfile
            reader = csv.DictReader(csvfile)
            # Normalize header names by stripping BOM/whitespace
            reader.fieldnames = [h.strip() for h in reader.fieldnames]
            for row in reader:
                donor_id = row.get('Donor ID')
                if donor_id and donor_id.strip():
                    luminex_data[donor_id.strip()] = {
                        "ABeta40": cls._safe_float(row.get('ABeta40 pg/ug'))
                        "ABeta42": cls._safe_float(row.get('ABeta42 pg/ug'))
                        "tTAU": cls._safe_float(row.get('tTAU pg/ug')),
                        "pTAU": cls._safe_float(row.get('pTAU pg/ug')),
                    }

        # --- Load Metadata and merge ---
        with open(metadata_path, newline='', encoding="utf-8-sig") as csvfil
            reader = csv.DictReader(csvfile)
            reader.fieldnames = [h.strip() for h in reader.fieldnames]
            for row in reader:
                donor_id = row.get('Donor ID')
                if not donor_id or not donor_id.strip():
                    continue

                donor_id = donor_id.strip()
                lum_data = luminex_data.get(donor_id, {})

                patient = cls(
                    donor_id=donor_id,
                    abeta40=lum_data.get("ABeta40"),
                    abeta42=lum_data.get("ABeta42"),
                    ttau=lum_data.get("tTAU"),
                    ptau=lum_data.get("pTAU"),
                    apoe=row.get('APOE Genotype', '').strip() or None,
                    mmse=cls._safe_float(row.get('Last MMSE Score')),
                    casi=cls._safe_float(row.get('Last CASI Score')),
                    sex=row.get('Sex', '').strip() or None,
                    death_age=cls._safe_int(row.get('Age at Death'))
                )

                cls.all_patients.append(patient)

    @staticmethod
    def _safe_float(value):
        """Convert to float if possible, else None."""
        try:
            return float(value)
        except (TypeError, ValueError):
            return None

    @staticmethod
    def _safe_int(value):
        """Convert to int if possible, else None."""
        try:
            return int(value)
        except (TypeError, ValueError):
            return None
```

Bar Plot and T-Test Analysis

This section visualizes and statistically analyzes the relationship between **MMSE scores** and **APOE4 status** (presence or absence of the APOE4 allele) in patients using **bar plots**.

Purpose

- **Bar Plot:** Displays the **mean MMSE score** for each group, with error bars representing the variability (standard deviation).
- **T-Test:** Compares the mean MMSE scores between the two groups to determine whether there is a statistically significant difference.

Key Steps

1. **Data Preparation:**
    - Split patients into two groups based on APOE4 presence.
    - Remove any patients with missing MMSE scores or non-numeric entries.
2. **Visualization:**

   - Create a bar plot showing the mean MMSE score for APOE4+ and APOE4- patients.
   - Add error bars representing standard deviation to indicate the spread of the data.

3. **Statistical Testing:**

   - Perform an independent t-test (`scipy.stats.ttest_ind`) on the MMSE scores of APOE4+ vs APOE4- patients.
   - Record and display the **t-statistic** and **p-value** in the notebook output.

Plot Details

- **X-axis:** APOE4 Status (Present vs Absent)
- **Y-axis:** Mean MMSE Score
- **Bars:** Show group mean scores
- **Error Bars:** Represent group standard deviations

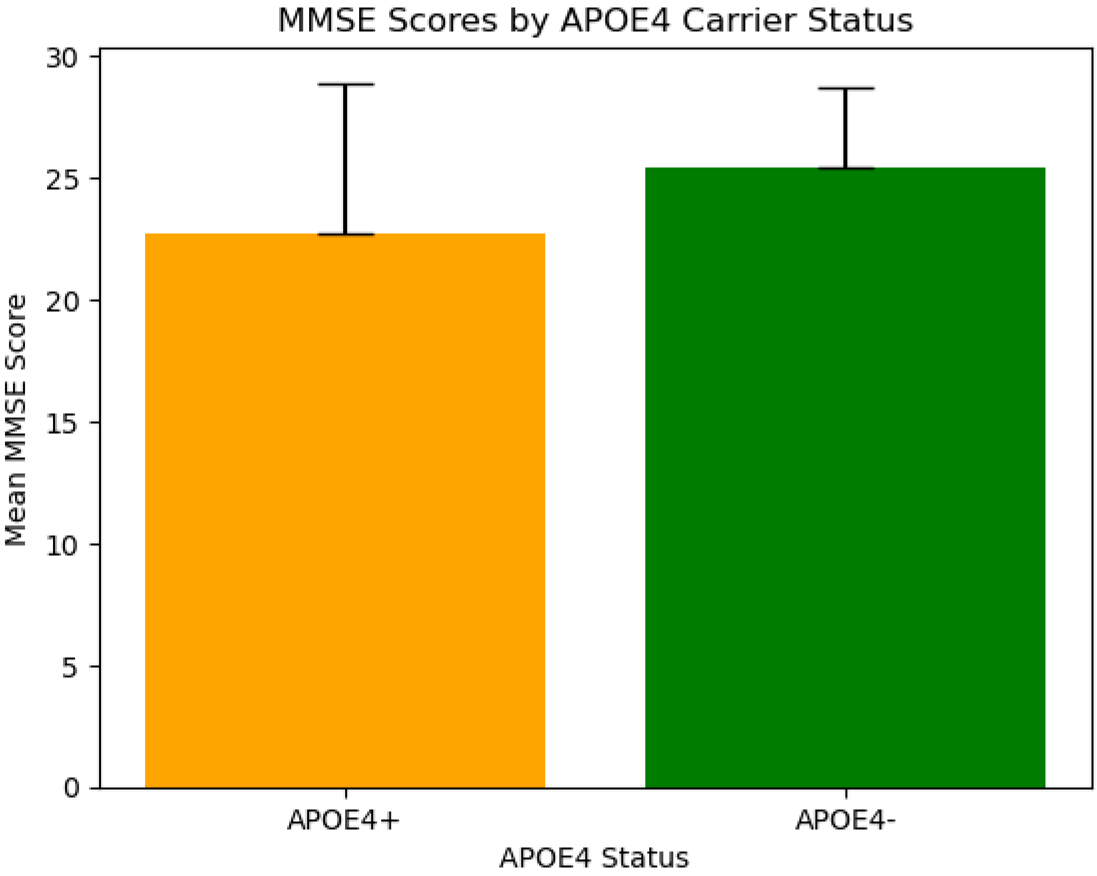
Interpretation

- The **bar plot** makes it easy to compare the average MMSE scores between APOE4 carriers and non-carriers.
- The **t-test** indicates whether the observed difference is likely due to chance or represents a significant association.
- Together, this approach provides a clear summary of central tendencies and supports statistical inference about the effect of APOE4 status on cognitive scores.

In [3]:

```
from patient import Patient
from matplotlib import pyplot as plt
from scipy import stats
import numpy as np
import statistics

# Load patients (if not already loaded)
Patient.instantiate_from_csv('UpdatedLuminex.csv', 'NODATEGENOTYPEUpdatedMet

# Collect MMSE values by APOE4 carrier status, ignoring None
MMSE_apoe4_vals = []
MMSE_no_apoe4_vals = []

for patient in Patient.all_patients:
    if patient.MMSE is not None and patient.APOE is not None:
        try:
            mmse_val = float(patient.MMSE)
        except ValueError:
            continue  # skip non-numeric MMSE values
        if "4" in str(patient.APOE):  # APOE4 carrier
            MMSE_apoe4_vals.append(mmse_val)
        else:
            MMSE_no_apoe4_vals.append(mmse_val)

# Safety check: make sure both groups have data
if not MMSE_apoe4_vals or not MMSE_no_apoe4_vals:
    print("Error: One of the groups has no valid MMSE data.")
else:
    # Compute means and standard deviations
    mean_apoe4 = statistics.mean(MMSE_apoe4_vals)
    mean_no_apoe4 = statistics.mean(MMSE_no_apoe4_vals)

    stdev_apoe4 = statistics.stdev(MMSE_apoe4_vals) if len(MMSE_apoe4_vals)
    stdev_no_apoe4 = statistics.stdev(MMSE_no_apoe4_vals) if len(MMSE_no_apo

    print(f"Mean MMSE (APOE4+): {mean_apoe4}, StdDev: {stdev_apoe4}")
    print(f"Mean MMSE (APOE4-): {mean_no_apoe4}, StdDev: {stdev_no_apoe4}")
    print(f"Number of APOE4+ patients: {len(MMSE_apoe4_vals)}")
    print(f"Number of APOE4- patients: {len(MMSE_no_apoe4_vals)}")

    # Perform independent t-test
    t_stat, p_val = stats.ttest_ind(MMSE_apoe4_vals, MMSE_no_apoe4_vals, equ
    print("t-statistic:", t_stat)
    print("p-value:", p_val)

    # Plot MMSE by APOE4 status
    group_labels = ["APOE4+", "APOE4-"]
    mean_vals = [mean_apoe4, mean_no_apoe4]
    stdev_vals = [stdev_apoe4, stdev_no_apoe4]
    yerr = [np.zeros(len(mean_vals)), stdev_vals]

    plt.bar(group_labels, mean_vals, yerr=yerr, capsize=10, color=["orange",
    plt.title("MMSE Scores by APOE4 Carrier Status")
    plt.xlabel("APOE4 Status")
    plt.ylabel("Mean MMSE Score")
    plt.show()
```

```
Mean MMSE (APOE4+): 22.695652173913043, StdDev: 6.197028485922585
Mean MMSE (APOE4-): 25.43859649122807, StdDev: 3.2896892159966438
Number of APOE4+ patients: 23
Number of APOE4- patients: 57
t-statistic: -2.011460490327938
p-value: 0.054299900702373616
```



The p-value for the comparison between APOE4 presence and MMSE scores in Alzheimer’s patients was calculated to be 0.0543. Since this p-value is greater than the significance level (α = 0.05), we fail to reject the null hypothesis. This suggests that, at the 5% significance level, there is not enough statistical evidence to conclude that the presence of the APOE4 allele is significantly associated with MMSE scores in this sample of Alzheimer’s patients.

Scatterplot with Overlaid Transparent Boxplot

Visualizes the relationship between **MMSE scores** and **APOE4 status** (present/absent) by combining:

- **Scatterplot:** Shows individual patient MMSE scores with jitter for visibility.
- **Boxplot (transparent):** Summarizes each group's distribution (median, quartiles, outliers).

Key Steps

1. **Data Prep:**

   - Split patients into APOE4+ and APOE4- groups.
   - Remove patients missing MMSE scores.
2. **Plot:**

   - Create jittered scatter points for each group.
   - Overlay transparent boxplots (`facecolor='none'`) for distribution context.
   - Label axes and add title.

Output

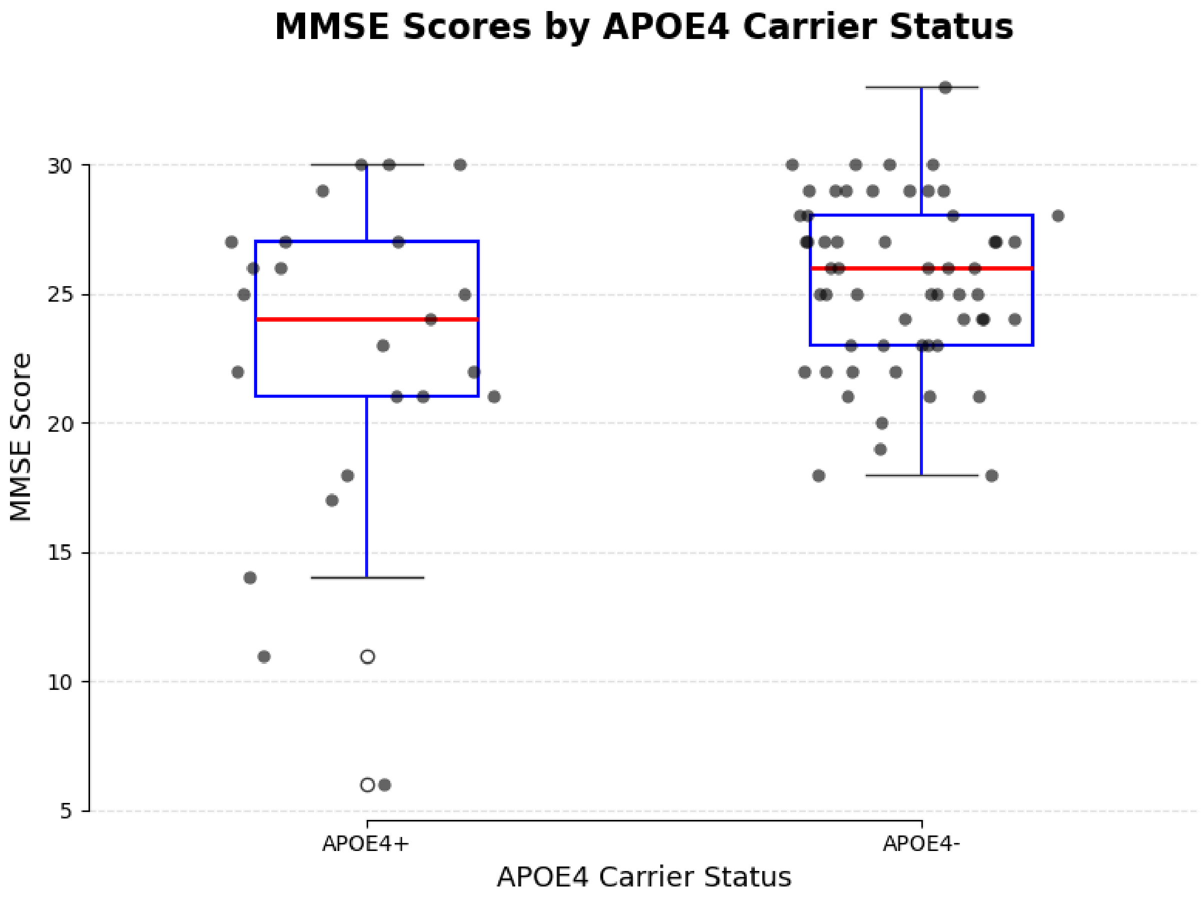
- **X-axis:** APOE4 Status (– vs +)
- **Y-axis:** MMSE Score
- **Dots:** Individual patients
- **Boxplots:** Distribution summary for each group

This plot makes it easy to compare cognitive performance (MMSE) between APOE4 carriers and non-carriers.

In [4]:

```
from patient import Patient
from matplotlib import pyplot as plt
import seaborn as sns

# Load patients
Patient.instantiate_from_csv('UpdatedLuminex.csv', 'NODATEGENOTYPEUpdatedMet

# Collect MMSE values by APOE4 status
MMSE_apoe4_vals = []
MMSE_no_apoe4_vals = []

for patient in Patient.all_patients:
    if patient.MMSE is not None and patient.APOE is not None:
        try:
            mmse_val = float(patient.MMSE)
        except ValueError:
            continue
        if "4" in str(patient.APOE):
            MMSE_apoe4_vals.append(mmse_val)
        else:
            MMSE_no_apoe4_vals.append(mmse_val)

# Combine into one dataset for plotting
x_labels = ["APOE4+"] * len(MMSE_apoe4_vals) + ["APOE4-"] * len(MMSE_no_apoe
mmse_scores = MMSE_apoe4_vals + MMSE_no_apoe4_vals

plt.figure(figsize=(8, 6))

# Jittered scatter points (stripplot)
sns.stripplot(
    x=x_labels, y=mmse_scores,
    jitter=0.25, alpha=0.6, color="black", size=6
)
```

```
# Overlay boxplot
sns.boxplot(
    x=x_labels, y=mmse_scores,
    showcaps=True,
    boxprops=dict(facecolor='none', edgecolor='blue', linewidth=1.5),
    whiskerprops=dict(color='blue', linewidth=1.3),
    medianprops=dict(color='red', linewidth=2),
    width=0.4
)

# Styling
plt.title("MMSE Scores by APOE4 Carrier Status", fontsize=16, weight="bold")
plt.xlabel("APOE4 Carrier Status", fontsize=13)
plt.ylabel("MMSE Score", fontsize=13)
plt.grid(axis="y", linestyle="--", alpha=0.35)
sns.despine(trim=True)  # removes top/right spines for a cleaner look

plt.tight_layout()
plt.show()
```



Scatterplot of MMSE vs Age at Death

This plot visualizes the relationship between **MMSE scores** and **Age at Death** for patients.

Purpose

This is mainly for **practice with visualization and regression analysis**, not to test a specific hypothesis. It allows us to:

- Plot individual patient MMSE scores against age at death.
- Fit and display a **linear regression line** to explore potential trends.
- Annotate the figure with the **regression equation** and **R² value**.

Plot Details

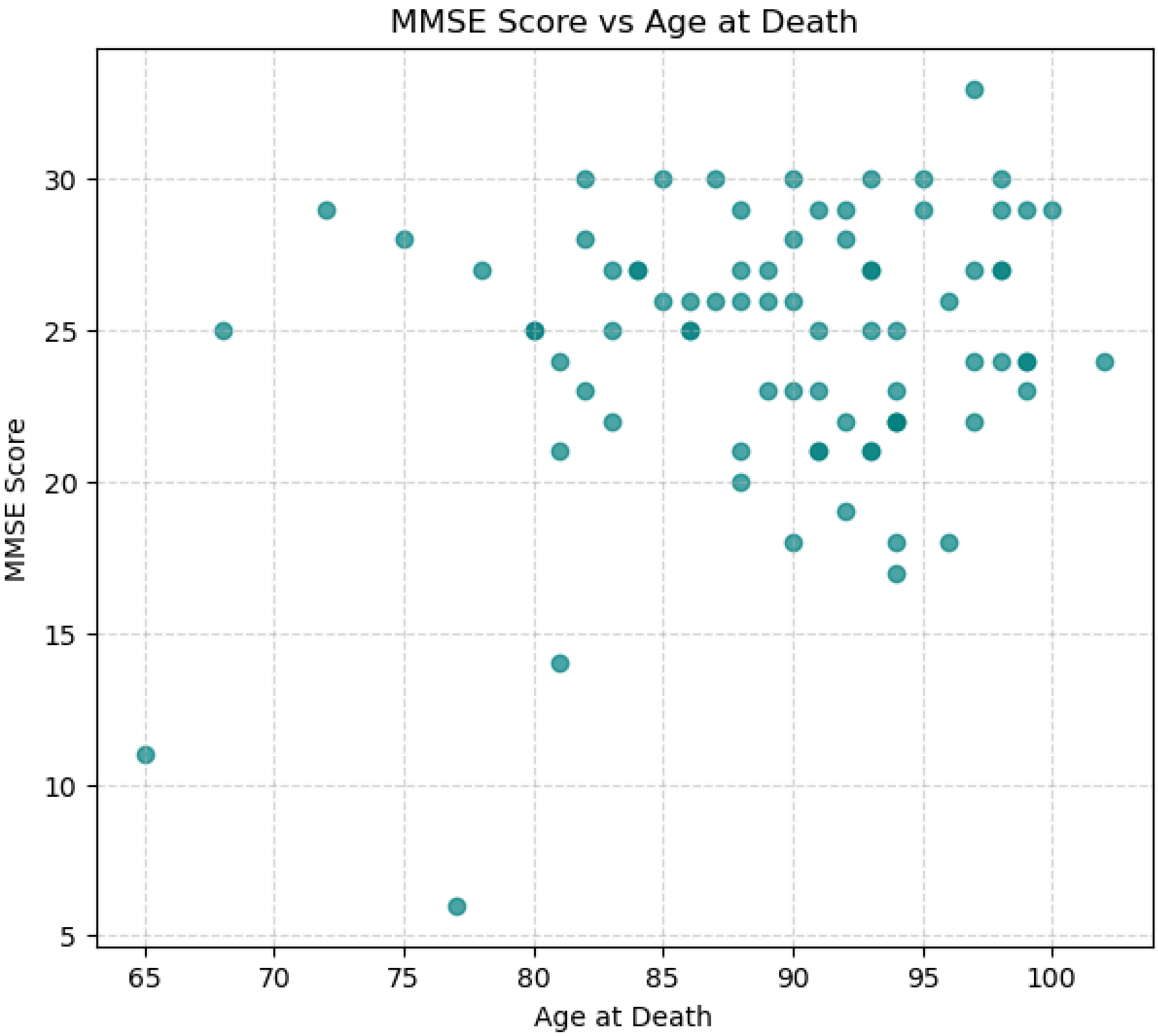
- **X-axis:** Age at Death
- **Y-axis:** MMSE Score
- Points represent patients; a fitted regression line shows potential association.

Learning Outcome

- Practice creating **scatterplots** and running **linear regression** in Python.
- Learn to **interpret and annotate** regression results visually.

In [5]:

```
from patient import Patient
from matplotlib import pyplot as plt

# Load patients (if not already loaded)
Patient.instantiate_from_csv('UpdatedLuminex.csv', 'NODATEGENOTYPEUpdatedMet

# Collect data for scatterplot
age_death_vals = []
mmse_vals = []

for patient in Patient.all_patients:
    if patient.death_age is not None and patient.MMSE is not None:
        try:
            age_val = int(patient.death_age)
            mmse_val = float(patient.MMSE)
        except ValueError:
            continue  # skip non-numeric entries
        age_death_vals.append(age_val)
        mmse_vals.append(mmse_val)

# Create scatterplot
plt.figure(figsize=(7, 6))
plt.scatter(age_death_vals, mmse_vals, alpha=0.7, color='teal')
plt.xlabel("Age at Death")
plt.ylabel("MMSE Score")
plt.title("MMSE Score vs Age at Death")
plt.grid(True, linestyle='--', alpha=0.5)
plt.show()
```



Exporting Data to .csv File

In [6]:

```
import csv
from patient import Patient

# Load patients
Patient.instantiate_from_csv('UpdatedLuminex.csv', 'NODATEGENOTYPEUpdatedMet

# Collect data
age_death_vals = []
mmse_vals = []

for patient in Patient.all_patients:
    if patient.death_age is not None and patient.MMSE is not None:
        try:
            age_val = int(patient.death_age)
            mmse_val = float(patient.MMSE)
        except ValueError:
            continue
        age_death_vals.append(age_val)
        mmse_vals.append(mmse_val)
```

```python

# Print the data
print("Age_at_Death | MMSE")
for age, mmse in zip(age_death_vals, mmse_vals):
    print(f"{age} | {mmse}")

# Export to CSV
with open("MMSE_patient_data.csv", mode="w", newline="") as f:
    writer = csv.writer(f)
    writer.writerow(["Age_at_Death", "MMSE"])  # header
    writer.writerows(zip(age_death_vals, mmse_vals))

print("Data exported to 'MMSE_patient_data.csv'.")
```

```
Age_at_Death | MMSE
77 | 6.0
81 | 24.0
94 | 17.0
94 | 18.0
97 | 22.0
90 | 26.0
93 | 21.0
80 | 25.0
95 | 30.0
80 | 25.0
90 | 28.0
86 | 26.0
91 | 29.0
94 | 22.0
86 | 25.0
94 | 25.0
88 | 27.0
81 | 21.0
92 | 29.0
99 | 29.0
91 | 25.0
99 | 23.0
91 | 21.0
87 | 30.0
82 | 28.0
97 | 33.0
94 | 22.0
97 | 24.0
91 | 23.0
86 | 25.0
87 | 26.0
81 | 14.0
98 | 27.0
68 | 25.0
85 | 30.0
99 | 24.0
100 | 29.0
96 | 18.0
85 | 26.0
93 | 25.0
93 | 30.0
83 | 25.0
90 | 23.0
93 | 27.0
96 | 26.0
65 | 11.0
92 | 19.0
94 | 22.0
98 | 24.0
98 | 27.0
78 | 27.0
92 | 28.0
94 | 23.0
99 | 24.0
82 | 23.0
```

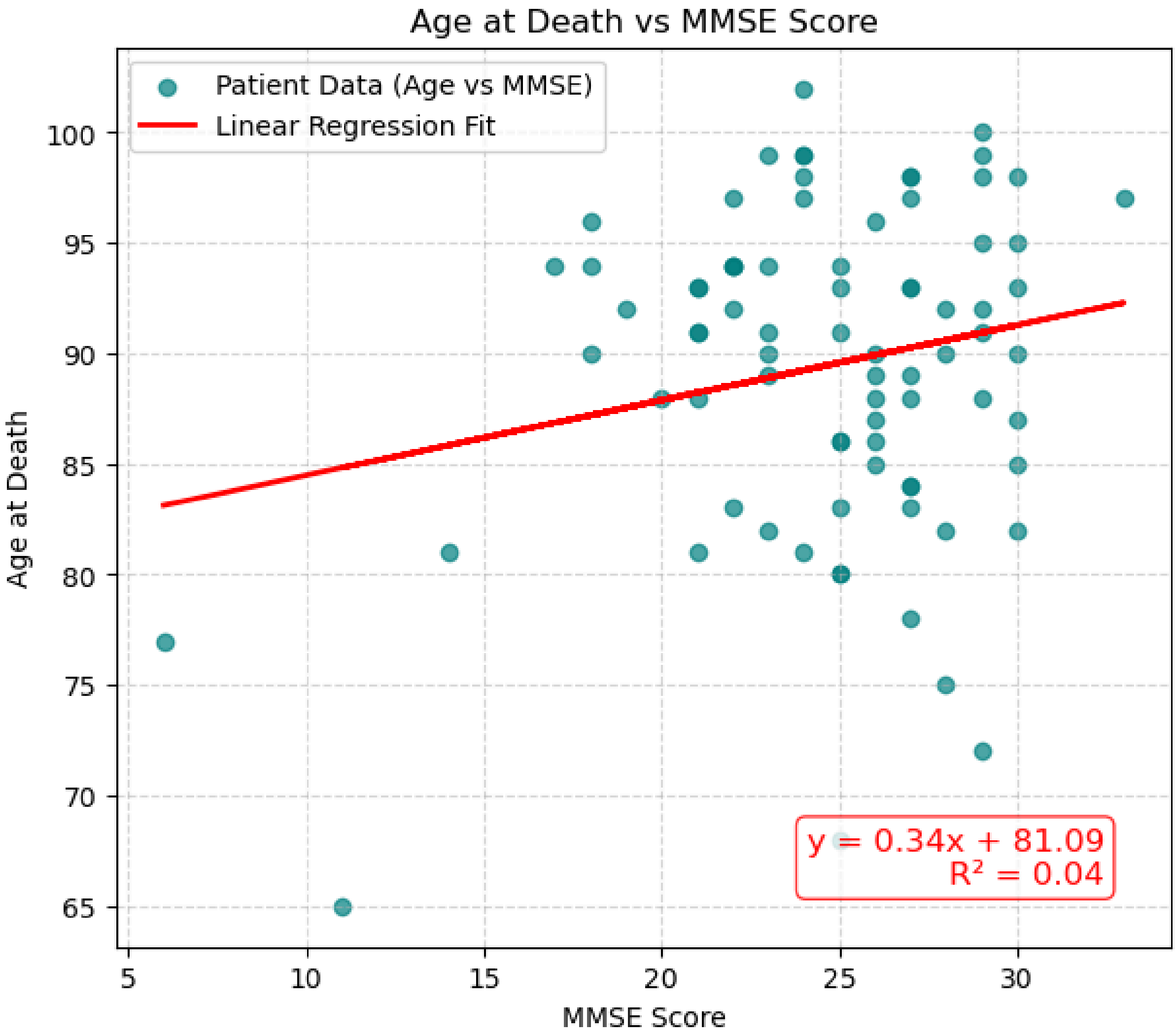
```
93 | 27.0
82 | 30.0
75 | 28.0
89 | 23.0
102 | 24.0
88 | 29.0
88 | 21.0
90 | 30.0
92 | 22.0
72 | 29.0
89 | 26.0
89 | 27.0
84 | 27.0
98 | 30.0
83 | 27.0
90 | 18.0
97 | 27.0
93 | 21.0
88 | 26.0
84 | 27.0
88 | 20.0
83 | 22.0
98 | 29.0
91 | 21.0
95 | 29.0
Data exported to 'MMSE_patient_data.csv'.
```

Performing Linear Regression

In [7]:

```
from sklearn.linear_model import LinearRegression
import pandas as pd
import matplotlib.pyplot as plt

# Load data
df = pd.read_csv("MMSE_patient_data.csv")

# Swap predictor/response
x = df["MMSE"].values.reshape(-1, 1)  # predictor (MMSE)
y = df["Age_at_Death"].values        # response (Age at Death)

# Fit linear regression
model = LinearRegression()
model.fit(x, y)

slope = model.coef_[0]
intercept = model.intercept_
r2 = model.score(x, y)

# Scatter plot
plt.figure(figsize=(7, 6))
plt.scatter(x, y, alpha=0.7, color='teal', label="Patient Data (Age vs MMSE)
plt.plot(x, model.predict(x), color="red", linewidth=2, label="Linear Regres

# Annotate equation (bottom-right corner)
```

```
x_pos = x.max() - (x.max() - x.min()) * 0.02
y_pos = y.min() + (y.max() - y.min()) * 0.02
equation = f"y = {slope:.2f}x + {intercept:.2f}\nR² = {r2:.2f}"
plt.text(x_pos, y_pos, equation, color="red", fontsize=12,
         verticalalignment='bottom', horizontalalignment='right',
         bbox=dict(facecolor='white', edgecolor='red', boxstyle='round,pad=0

# Labels and title
plt.xlabel("MMSE Score")
plt.ylabel("Age at Death")
plt.title("Age at Death vs MMSE Score")
plt.legend()
plt.grid(True, linestyle='--', alpha=0.5)
plt.show()
```



Linear Regression Analysis

The regression produced the following equation:
Age at Death = 0.34 × MMSE + 81.09
with an **R^2 value of 0.04**.

Interpretation

Line of Best Fit

- **Slope (0.34):** For each additional point in MMSE, Age at Death increases by about **0.34 years** on average.
- **Intercept (81.09):** When MMSE = 0, the model predicts an Age at Death of **81.09 years**.
- **Practical Implication:** The positive slope suggests a **slight association** between better MMSE scores and later Age at Death, but the effect size is very small.

R^2 Value (0.04)

- **Low explanatory power:** Only **4% of the variability** in Age at Death is explained by MMSE.
- **Conclusion:** MMSE has very limited predictive value for Age at Death — most of the variation is due to other factors.

Regression Takeaways

Although there is a slight upward trend, the very low R^2 indicates that this relationship is weak and likely not clinically meaningful.

Verify and validate your analysis:

During our analysis, we calculated a p-value of 0.0543 for the association between APOE4 presence and MMSE scores in Alzheimer's patients. Since the threshold for significance was $\alpha = 0.05$, this p-value indicates that the relationship was not statistically significant (0.0543 is just above the cutoff).

However, literature suggests that a relationship between the two variables exists. We investigated two studies on the relationship between APOE4 and MMSE decline: one developed by researchers from Massachusetts General Hospital, the University of Antioquia in Colombia, and the Banner Alzheimer's Institute, and another conducted by researchers from National Cheng Kung University, China Medical University and Hospital, and the National Health Research Institutes in Taiwan.

Both studies show a clear relationship between the APOE4 gene and MMSE scores: carrying APOE4 is linked to faster cognitive decline. In the Nature Communications study, people with a genetic form of Alzheimer's who also carried APOE4 experienced earlier onset of symptoms and a quicker drop in MMSE scores, while those with the protective APOE2 version declined more slowly. In the American Journal of Geriatric Psychiatry study of older adults in Taiwan, APOE4 carriers—especially those with two copies—were several times more likely to see their MMSE scores fall over time compared to non-carriers. Overall, APOE4 consistently increases the risk and speed of MMSE decline, making it a strong genetic factor in memory and thinking loss.

Sources: Langella, Stephanie, et al. “Effect of Apolipoprotein Genotype and Educational Attainment on Cognitive Function in Autosomal Dominant Alzheimer’s Disease.” Nature News, Nature Publishing Group, 23 Aug. 2023, www.nature.com/articles/s41467-023-40775-z.

Hsiao, Po-Jen, et al. “APOE-ε4 Alleles Modify the Decline of MMSE Scores Associated With Time-Dependent PM2.5 Exposure: Findings From a Community-Based Longitudinal Cohort Study.” The American Journal of Geriatric Psychiatry, vol. 32, no. 9, Sept. 2024, pp. 1080–1092.

Conclusions and Ethical Implications:

Our analysis yielded a p-value of 0.0543 for the association between APOE4 presence and MMSE scores in Alzheimer’s patients. While this result narrowly misses the threshold for statistical significance ($\alpha = 0.05$), it does not necessarily indicate the absence of a relationship. Supporting evidence from the literature confirms that APOE4 is a strong genetic factor influencing cognitive deterioration in Alzheimer's patients, and our results mostly align with this evidence despite not reaching statistical significance within the data set.

From an ethical perspective, this relationship has several implications. First, it speaks to the potential utility of genetic testing, as knowledge of APOE4 carrier status could inform personal and family planning (e.g. lifestyle changes/preventative strategies, long-term care planning, financial planning, clinical support networks, etc.). Additionally, it raises questions about course of action once a patient is informed of their carrier status. For example, do people have the responsibility to inform relatives of their carrier status, as family members may also carry the gene and be at risk for Alzheimer's and a higher degree of genetic deterioration? Will communicating carrier status to patients simply cause anxiety about increased genetic risk?

Limitations and Future Work:

One key limitation of this research was the sample size. The relatively small cohort provided in the data set may have limited the ability to find a true association between APOE4 presence and MMSE score in Alzheimer's patients. Additionally, variability between patients (demographics, disease stage, etc.) may have influenced MMSE outcomes, confounding the effects of the APOE4 genotype on the score. Future work should focus on larger cohorts to improve statistical power. Additionally, it could employ statistical techniques to limit confounding such as utilizing a matched pair design in which participants with similar confounding characteristics (age, sex, education, etc.) could be paired across comparison groups. Finally, we would have liked to conduct longitudinal analysis of donor MMSE score progression as opposed to simply comparing APOE4 presence with the Last MMSE Score data. This could suggest if the presence of APOE4 impacted not just

Notes from Team:

- Original project goal: This project seeks to investigate the relationship between tau levels (pg/µg) and MMSE scores in patients with Alzheimer’s disease.
- New project goal: This project seeks to investigate the relationship between different APOE combinations and MMSE scores in patients with Alzheimer’s disease. We pivoted away from this question because we were having trouble making sense of any of the graphs that we generated to analyze the data. There seemed to be very few conclusions to make regarding any correlations or trends. Because of this, we decided to study the APOE genotype and how certain allele combinations suggest susceptibility to developing Alzheimer's disease. We were more interested in the results of this research goal, and so far, we've been able to make better sense of the data and pick out some vague trends.

Questions for Instructors:

N/A