🛠️ Project Setup

In this first step, we'll prepare the workspace for the **OKCupid Date-A-Scientist** project.

📦 Download the Project Materials

1. Download the ZIP file for the **Date-A-Scientist** project from the following link:
   👉 Download Starter Files
2. Unzip the folder. You should see two items:
   - `date-a-scientist.ipynb` (a blank Jupyter Notebook)
   - `profiles.csv` (the dataset containing user profiles)

🚀 Launch Jupyter Notebook

1. Open your terminal or command line interface.
2. Type the following command to start Jupyter Notebook:

   ```
   jupyter notebook
   ```
3. A browser tab will open automatically.
4. Click on `date-a-scientist.ipynb` to open the notebook.
5. Build your project inside this file.

🧠 How to Use Jupyter Notebook

Jupyter Notebook is an interactive tool that lets you combine code, visualizations, and explanatory text in one document. It's perfect for data science projects because it supports exploration, analysis, and storytelling.

If you need help setting up or using Jupyter Notebook, check out these resources:

- Command Line Interface Setup
- Introducing Jupyter Notebook
- Setting up Jupyter Notebook
- Getting Started with Jupyter
- Getting More out of Jupyter Notebook

````
# 🔐 Git Repository Setup with SSH and LFS

To manage version control and large files efficiently, this project uses
**Git**, **SSH authentication**, and **Git LFS**.

## 📁 Local Project Path

```bash
cd code/DataScientistML_Codecademy/OKCupid-Date-A-Scientist-Starter
````

🧱 Initialize and Push to Remote Repository

```
git init
git lfs install
git add .
git commit -m "Initial commit for OKCupid Date-A-Scientist project"
git remote add origin git@personal-github:gabrielarcangelbol/OKCupid-Date-
A-Scientist-Starter.git
git branch -M main
git push -u origin main
```
```

⚙️ Configure `.gitattributes` for Large Files

To track CSV files with Git LFS, the following line was added to `.gitattributes`:

```
*.csv filter=lfs diff=lfs merge=lfs -text
```

Then committed:

```
git add .gitattributes
git commit -m "Configure Git LFS for CSV files"
```

🔐 SSH Key Setup for Personal GitHub Account

To ensure Git uses your **personal SSH key**, follow these steps:

1. Start the SSH agent:

   ```
   eval "$(ssh-agent -s)"
   ```
2. Add your personal SSH key:

   ```
   ssh-add ~/.ssh/id_rsa_personal
   ```

🔍 Verify Remote Configuration

To confirm the correct remote is set:

```
cd code/DataScientistML_Codecademy/OKCupid-Date-A-Scientist-Starter
git remote -v
```

You should see:

```
origin  git@personal-github:gabrielarcangelbol/OKCupid-Date-A-
Scientist-Starter.git (fetch)
origin  git@personal-github:gabrielarcangelbol/OKCupid-Date-A-
Scientist-Starter.git (push)
```

✅ With Git, SSH, and LFS configured, your project is ready for version-controlled development and collaboration.

🧭 Project Scoping

Properly scoping your project creates structure and helps you think through your entire workflow before diving into the analysis. This section outlines the key components of the project scope, inspired by the University of Chicago's Data Science Project Scoping Guide.

🎯 1. Project Goals

- **Primary Objective**: Explore the OKCupid dataset to uncover patterns in dating preferences and behaviors using machine learning and NLP techniques.
- **Secondary Goals**:
  - Identify which user attributes are most predictive of compatibility.
  - Apply clustering to discover latent user segments.
  - Build a supervised model to predict user traits or preferences based on profile text.

📊 2. Data Overview

- **Source**: `profiles.csv` from OKCupid (provided in starter files)
- **Structure**: Each row represents a user profile with multiple features including:
  - Demographics (age, sex, orientation, location)
  - Lifestyle (diet, smoking, drinking, drugs)
  - Essay responses (10 free-text fields)
- **Challenges**:
  - Missing values
  - Unstructured text (essays)
  - Potential class imbalance

🧪 3. Analytical Approach

Exploratory Data Analysis (EDA)

- Distribution of key demographics
- Missing data patterns
- Word frequency and sentiment in essay fields

Feature Engineering

- Text vectorization (TF-IDF, embeddings)
- Aggregated lifestyle indicators
- Derived compatibility scores

Modeling

- **Unsupervised**: Clustering (e.g., KMeans, DBSCAN) to identify user segments
- **Supervised**: Classification (e.g., logistic regression, random forest) to predict:
  - Smoking habits
  - Orientation
  - Personality traits (inferred from essays)

🧱 4. Constraints & Assumptions

- **Assumptions**:
  - Users are honest in their profiles
  - Essay content reflects personality and preferences
- **Constraints**:
  - No access to actual match outcomes
  - Limited metadata on user interactions

🔄 5. Risks & Adjustments

- NLP models may underperform due to short or noisy text
- Clustering may not yield interpretable segments
- Some hypotheses may not be supported by the data

📌 Next Steps

- Clean and preprocess the dataset
- Perform EDA and visualize key trends
- Define target variables for modeling
- Iterate on feature engineering and model evaluation

📚 For more guidance, refer to the full Data Science Project Scoping Guide and Scoping Worksheet (PDF).

Sources: Scoping Guide, Scoping Worksheet PDF

🤖 Select ML-Solvable Problem

In this section, we define a machine learning problem that is feasible given the dataset and project scope. The goal is to identify a task that:

- Can be solved using available data
- Is achievable within a reasonable timeframe
- Has a clear input-output structure suitable for ML
- Aligns with the broader goals of the project

🧩 Problem Statement

Many users on OKCupid consider **Zodiac signs** important when evaluating compatibility. However, not all users provide their zodiac sign in their profile. This creates a gap in the matching process.

Problem: Can we predict a user's zodiac sign based on other profile attributes such as lifestyle habits (drinking, smoking, drug use) and free-text essay responses?

🧠 ML Framing

This is a **multi-class classification** problem with 12 possible zodiac labels.

Input Features:

- Categorical: `drinks`, `smokes`, `drugs`, `job`, `education`
- Textual: `essay0` to `essay9` (to be vectorized using NLP techniques)

Target Variable:

- `sign` (Zodiac sign)

🔍 Why This Problem?

- The `sign` column has missing values, making it a good candidate for imputation via ML.
- Lifestyle and personality traits (expressed in essays) may correlate with astrological archetypes.
- Predicting zodiac signs could enhance match recommendations for users who omit this field.

🧪 ML Techniques to Be Used

- **Text preprocessing**: cleaning, tokenization, TF-IDF or embeddings
- **Feature encoding**: one-hot or ordinal encoding for categorical variables
- **Modeling**:
  - Baseline: Logistic Regression or Naive Bayes

    - Advanced: Random Forest, XGBoost, or fine-tuned transformer (if time permits)
- **Evaluation**: Accuracy, F1-score, confusion matrix

📌 Next Steps

- Analyze class distribution of `sign`
- Explore correlations between zodiac and lifestyle features
- Preprocess essays and engineer features
- Train and evaluate classification models

📚 Reference: Google's Guide to Identifying ML-Suitable Problems

📂 Load and Check Data

Before applying any machine learning techniques, we need to load the dataset and verify that it contains a suitable **label or response variable** for supervised learning.

The dataset is stored in `profiles.csv` and includes:

🗒️ Multiple-choice columns:

- `body_type`, `diet`, `drinks`, `drugs`, `education`, `ethnicity`, `height`, `income`, `j`

📝 Essay fields:

- `essay0` to `essay9`, covering topics like self-summary, lifestyle, preferences, and personality

---

🎯 Target Variable Check

Since our selected ML task is to **predict the user's zodiac sign**, we will check the `sign` column for:

- Presence of values
- Class distribution
- Missing data

If the `sign` column is too sparse or unreliable, we may need to revisit our problem definition.

📦 Load Dataset

```
import pandas as pd

# Load the dataset
df = pd.read_csv("profiles.csv")

# Display basic info
print("Shape of dataset:", df.shape)
df.info()

# Preview the first few rows
df.head()
```

✅ Next Steps

- Check for missing values in the `sign` column
- Explore distribution of zodiac signs
- Assess completeness of other relevant features (`drinks`, `smokes`, `drugs`, `essay0`–`essay9`)
- Decide whether to proceed with this target or redefine the ML problem

```
# 📦 Load and Inspect OKCupid Dataset

import pandas as pd

# Load the dataset
df = pd.read_csv("profiles.csv")

# Basic shape and structure
print("✅ Dataset loaded successfully.")
print("🔢 Number of rows and columns:", df.shape)

# Overview of column types and missing values
print("\n📋 Dataset Info:")
df.info()

# Preview the first few rows
print("\n🔍 First 5 rows:")
df.head()
```

```
✅ Dataset loaded successfully.
🔢 Number of rows and columns: (59946, 31)

📋 Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59946 entries, 0 to 59945
Data columns (total 31 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   age          59946 non-null  int64  
 1   body_type    54650 non-null  object 
 2   diet         35551 non-null  object 
 3   drinks       56961 non-null  object 
 4   drugs        45866 non-null  object 
 5   education    53318 non-null  object 
 6   essay0       54458 non-null  object 
 7   essay1       52374 non-null  object 
 8   essay2       50308 non-null  object 
 9   essay3       48470 non-null  object 
 10  essay4       49409 non-null  object 
 11  essay5       49096 non-null  object 
 12  essay6       46175 non-null  object 
 13  essay7       47495 non-null  object 
 14  essay8       40721 non-null  object 
 15  essay9       47343 non-null  object 
 16  ethnicity    54266 non-null  object 
 17  height       59943 non-null  float64
 18  income       59946 non-null  int64  
 19  job          51748 non-null  object 
 20  last_online  59946 non-null  object 
 21  location     59946 non-null  object 
 22  offspring    24385 non-null  object 
 23  orientation  59946 non-null  object 
 24  pets         40025 non-null  object 
 25  religion     39720 non-null  object 
 26  sex          59946 non-null  object 
 27  sign         48890 non-null  object 
 28  smokes       54434 non-null  object 
 29  speaks       59896 non-null  object 
 30  status       59946 non-null  object 
dtypes: float64(1), int64(2), object(28)
memory usage: 14.2+ MB

🔍 First 5 rows:
```

	age	body_type	diet	drinks	drugs	education	essay0	essay1	ess
0	22	a little extra	strictly anything	socially	never	working on college/university	about me: \n \ni would love to think...	currently working as an international agent fo...	making pe laugh /> \nrar about a
1	35	average	mostly other	often	sometimes	working on space camp	i am a chef: this is what that means. \n1...	dedicating everyday to being an unbelievable b...	being ha ridicu amonts o
2	38	thin	anything	socially	NaN	graduated from masters program	i'm not ashamed of much, but writing public te...	i make nerdy software for musicians, artists, ...	improvisir diffe cont alternati
3	23	thin	vegetarian	socially	NaN	working on college/university	i work in a library and go to school. . .	reading things written by old dead people	pla synthes and organi books a
4	29	athletic	NaN	socially	never	graduated from college/university	hey how's it going? currently vague on the pro...	work work work work + play	crea imagery to at /> \nhttp://b

5 rows × 31 columns

```
# 📊 Check Target Variable and Feature Completeness

# Check how many missing values are in the 'sign' column
missing_sign = df['sign'].isnull().sum()
total_rows = len(df)
print(f"❓ Missing 'sign' values: {missing_sign} out of {total_rows} ({missing_sign/total_row

# View distribution of zodiac signs
print("\n🔮 Zodiac Sign Distribution:")
print(df['sign'].value_counts(dropna=False))

# Check missing values in lifestyle features
print("\n🚦 Missing values in lifestyle columns:")
print(df[['drinks', 'smokes', 'drugs']].isnull().sum())

# Check completeness of essay fields
print("\n📝 Missing values in essay fields:")
essay_cols = [f"essay{i}" for i in range(10)]
print(df[essay_cols].isnull().sum())
```

❓ Missing 'sign' values: 11056 out of 59946 (18.44%)

🔮 Zodiac Sign Distribution:

```
sign
NaN                                                 11056
gemini and it&rsquo;s fun to think about             1782
scorpio and it&rsquo;s fun to think about            1772
leo and it&rsquo;s fun to think about                1692
libra and it&rsquo;s fun to think about              1649
taurus and it&rsquo;s fun to think about             1640
cancer and it&rsquo;s fun to think about             1597
pisces and it&rsquo;s fun to think about             1592
sagittarius and it&rsquo;s fun to think about        1583
virgo and it&rsquo;s fun to think about              1574
aries and it&rsquo;s fun to think about              1573
aquarius and it&rsquo;s fun to think about           1503
virgo but it doesn&rsquo;t matter                    1497
leo but it doesn&rsquo;t matter                      1457
cancer but it doesn&rsquo;t matter                   1454
gemini but it doesn&rsquo;t matter                   1453
taurus but it doesn&rsquo;t matter                   1450
aquarius but it doesn&rsquo;t matter                 1408
libra but it doesn&rsquo;t matter                    1408
capricorn and it&rsquo;s fun to think about          1376
sagittarius but it doesn&rsquo;t matter              1375
aries but it doesn&rsquo;t matter                    1373
capricorn but it doesn&rsquo;t matter                1319
pisces but it doesn&rsquo;t matter                   1300
scorpio but it doesn&rsquo;t matter                  1264
leo                                                  1159
libra                                                1098
cancer                                               1092
virgo                                                1029
scorpio                                              1020
gemini                                               1013
taurus                                               1001
aries                                                 996
pisces                                                992
aquarius                                              954
sagittarius                                           937
capricorn                                             833
scorpio and it matters a lot                           78
leo and it matters a lot                               66
cancer and it matters a lot                            63
aquarius and it matters a lot                          63
gemini and it matters a lot                            62
pisces and it matters a lot                            62
libra and it matters a lot                             52
taurus and it matters a lot                            49
sagittarius and it matters a lot                       47
aries and it matters a lot                             47
capricorn and it matters a lot                         45
virgo and it matters a lot                             41
Name: count, dtype: int64
```

🚦 Missing values in lifestyle columns:

```
drinks     2985
smokes     5512
drugs     14080
dtype: int64
```

📝 Missing values in essay fields:

```
essay0     5488
essay1     7572
essay2     9638
essay3    11476
essay4    10537
essay5    10850
essay6    13771
essay7    12451
essay8    19225
essay9    12603
dtype: int64
essay0     5488
```

```
essay1     7572
essay2     9638
essay3    11476
essay4    10537
essay5    10850
essay6    13771
essay7    12451
essay8    19225
essay9    12603
dtype: int64
```

🔍 Preliminary Data Assessment

After loading and inspecting the dataset, we identified several key insights and challenges that will guide our next steps.

🧠 Target Variable: `sign`

- ✅ Present in 48,890 out of 59,946 rows (~81.56%)
- ❌ Missing in ~18.44% of entries
- ✅ Contains 12 zodiac signs, but with many variations (e.g., "gemini and it's fun to think about", "gemini but it doesn't matter", "gemini")
- 🔧 **Action**: Normalize zodiac labels by extracting the base sign (e.g., "gemini") and discarding modifiers

🚦 Lifestyle Features

Feature	Missing Values	% Missing
`drinks`	2,985	~5.0%
`smokes`	5,512	~9.2%
`drugs`	14,080	~23.5%

- 🔧 **Action**: Consider imputing missing values or treating them as a separate category ("unknown")

📝 Essay Fields

- Missing values range from ~9% (`essay0`) to ~32% (`essay8`)
- Text fields are rich but sparse and noisy
- 🔧 **Action**:
  - Combine essays into a single text field for NLP preprocessing
  - Apply cleaning (lowercasing, punctuation removal, etc.)
  - Use TF-IDF or embeddings for feature extraction

⚠️ Other Observations

- `offspring` , `pets` , and `religion` have high missingness (>30%)
- `income` is present but may be skewed or zero-filled
- `height` and `age` are complete and usable

✅ Next Steps

1. **Normalize the `sign` column** to extract base zodiac labels
2. **Clean and combine essay fields** for NLP feature engineering

3. **Handle missing values** in lifestyle and categorical features
4. **Visualize distributions** to guide feature selection
5. **Prepare training data** for supervised classification

📌 These steps will help us build a robust pipeline for predicting zodiac signs based on lifestyle and personality traits.

🧹 Data Cleaning: Zodiac Signs and Essay Text

Before we begin exploratory analysis, we need to clean and standardize key features in the dataset. This includes:

1. Normalizing the `sign` column to extract the base zodiac label
2. Combining all essay fields into a single text column for NLP processing

These steps will ensure consistency and simplify downstream modeling.

✅ Step 1: Normalize the `sign` column

```
import re
import pandas as pd

# Define a function to extract the base zodiac sign from messy strings
def extract_zodiac(sign):
    if pd.isnull(sign):
        return None
    # Use regex to find the zodiac keyword in the string
    match = re.search(r'\b(aries|taurus|gemini|cancer|leo|virgo|libra|scorpio|sagittarius|capr
    return match.group(0) if match else None

# Apply the function to the 'sign' column
df['zodiac_clean'] = df['sign'].apply(extract_zodiac)

# Display the cleaned distribution
print("🔮 Cleaned Zodiac Sign Distribution:")
print(df['zodiac_clean'].value_counts(dropna=False))
```

```
🔮 Cleaned Zodiac Sign Distribution:
zodiac_clean
None           11056
leo             4374
gemini          4310
libra           4207
cancer          4206
virgo           4141
taurus          4140
scorpio         4134
aries           3989
pisces          3946
sagittarius     3942
aquarius        3928
capricorn       3573
Name: count, dtype: int64
```

Explanation:

- Many entries in the `sign` column contain modifiers like “but it doesn’t matter” or “and it’s fun to think about”.
- This function extracts only the core zodiac sign using regular expressions.
- The result is stored in a new column `zodiac_clean`.

✅ Step 2: Combine all essay fields into one column

```python
# List of essay column names
essay_cols = [f"essay{i}" for i in range(10)]

# Fill missing essay values with empty strings and concatenate them
df['essays_combined'] = df[essay_cols].fillna('').agg(' '.join, axis=1)

# Preview the combined essay text for the first profile
print("📝 Sample combined essay text:")
print(df['essays_combined'].iloc[0][:500])  # Show first 500 characters
```

```
📝 Sample combined essay text:
about me:<br />
<br />
i would love to think that i was some some kind of intellectual:
either the dumbest smart guy, or the smartest dumb guy. can't say i
can tell the difference. i love to talk about ideas and concepts. i
forge odd metaphors instead of reciting cliches. like the
simularities between a friend of mine's house and an underwater
salt mine. my favorite word is salt by the way (weird choice i
know). to me most things in life are better as metaphors. i seek to
make myself a little be
```

Explanation:

- Essay fields are spread across `essay0` to `essay9`, each representing a different prompt.
- We fill missing values with empty strings to avoid `NaN` issues.
- Then we concatenate all essays into a single column `essays_combined` for easier NLP processing.

📊 Explore and Explain Data

Now that the dataset is cleaned and structured, we begin exploratory data analysis (EDA) to understand the distribution, relationships, and potential patterns in the data.

This section includes:

- Descriptive statistics for numeric and categorical features
- Visualizations for univariate and multivariate exploration
- Narrative insights based on observed trends

📈 Descriptive Statistics

We will compute summary statistics such as mean, median, range, and correlations for numeric features like `age`, `height`, and `income`.

📊 Data Visualizations

Using Matplotlib and Seaborn, we will create:

- Histograms and boxplots for numeric distributions
- Countplots for categorical features
- Heatmaps for correlation analysis
- Optional: Word frequency plots from essay text
- **Comparison plots for categorical variables by gender (e.g., body_type vs. sex)**

🧠 Insights

We will annotate each visualization with observations and hypotheses that may inform feature engineering or modeling decisions.

📚 Reference: NIST EDA Introduction

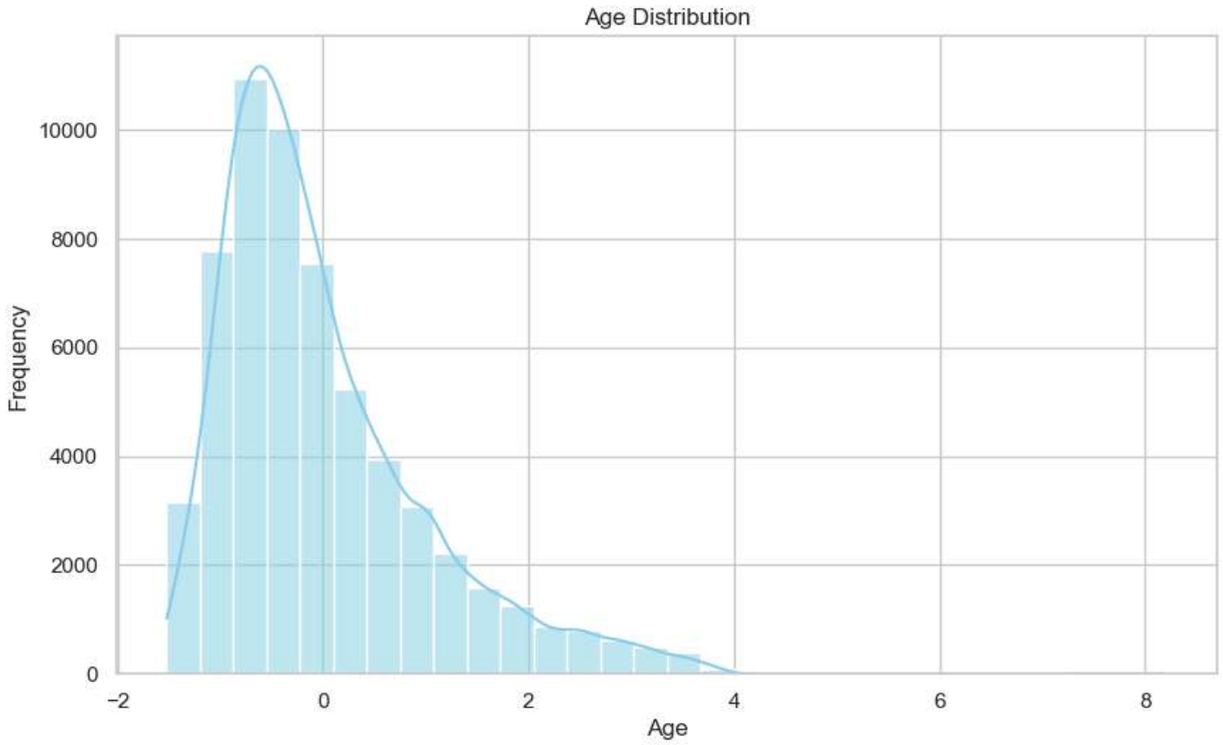
```
import matplotlib.pyplot as plt
import seaborn as sns

# Set plot style
sns.set(style="whitegrid")
plt.rcParams["figure.figsize"] = (10, 6)

# 📈 Descriptive statistics for numeric columns
print("📊 Summary statistics for numeric features:")
print(df[['age', 'height', 'income']].describe())

# 🔍 Correlation matrix
print("\n📈 Correlation matrix:")
print(df[['age', 'height', 'income']].corr())

# 📊 Histogram of age
sns.histplot(df['age'], bins=30, kde=True, color='skyblue')
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()

# 👥 Body type vs. gender
sns.countplot(y='body_type', hue='sex', data=df, palette='Set2',
              order=df['body_type'].value_counts().index)
plt.title("Body Type Distribution by Gender")
plt.xlabel("Count")
plt.ylabel("Body Type")
plt.legend(title="Gender")
plt.show()

# 📦 Boxplot of height
sns.boxplot(x=df['height'], color='lightgreen')
plt.title("Height Distribution")
plt.xlabel("Height (inches)")
plt.show()

# 💰 Income distribution (filtered to exclude extreme values)
sns.histplot(df[df['income'] < 200000]['income'], bins=50, color='salmon')
plt.title("Income Distribution (Under $200k)")
plt.xlabel("Income")
plt.ylabel("Frequency")
plt.show()

# 🔮 Countplot of zodiac signs
sns.countplot(y='zodiac_clean', data=df, order=df['zodiac_clean'].value_counts().index, palett
plt.title("Zodiac Sign Frequency")
plt.xlabel("Count")
plt.ylabel("Zodiac Sign")
plt.show()

# 🔥 Heatmap of correlations
sns.heatmap(df[['age', 'height', 'income']].corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```

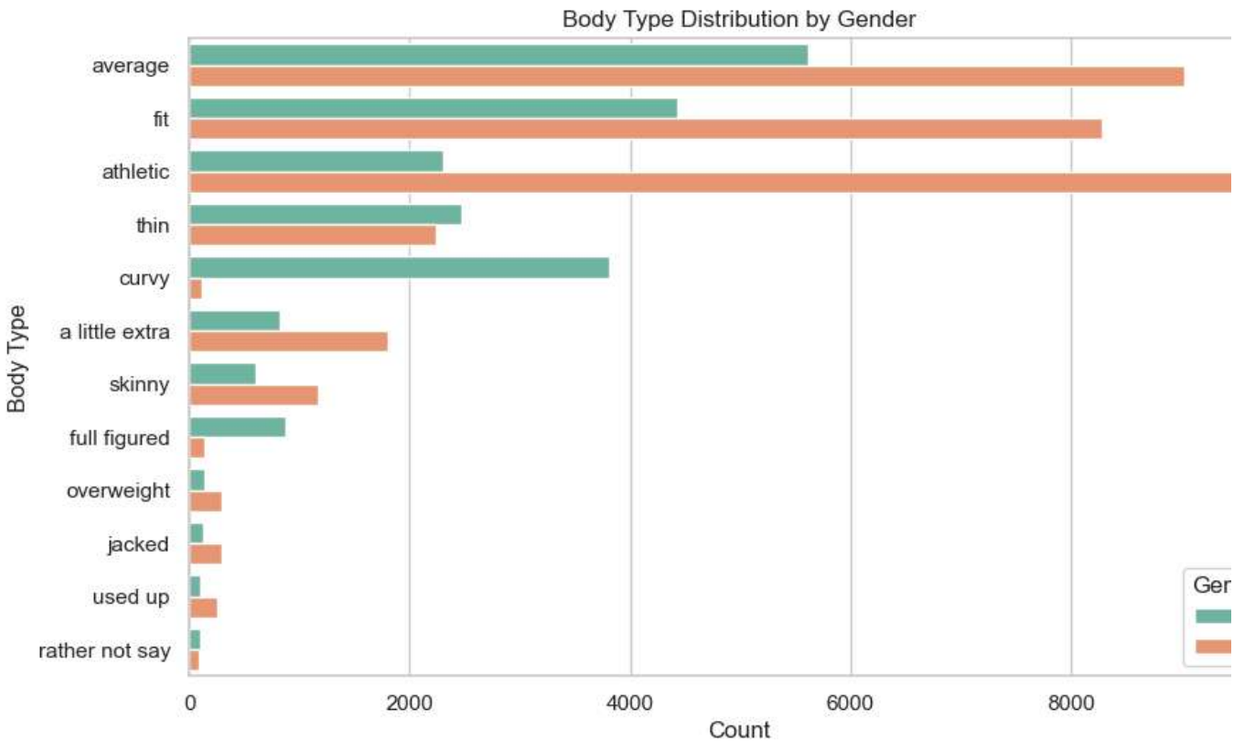
📊 Summary statistics for numeric features:

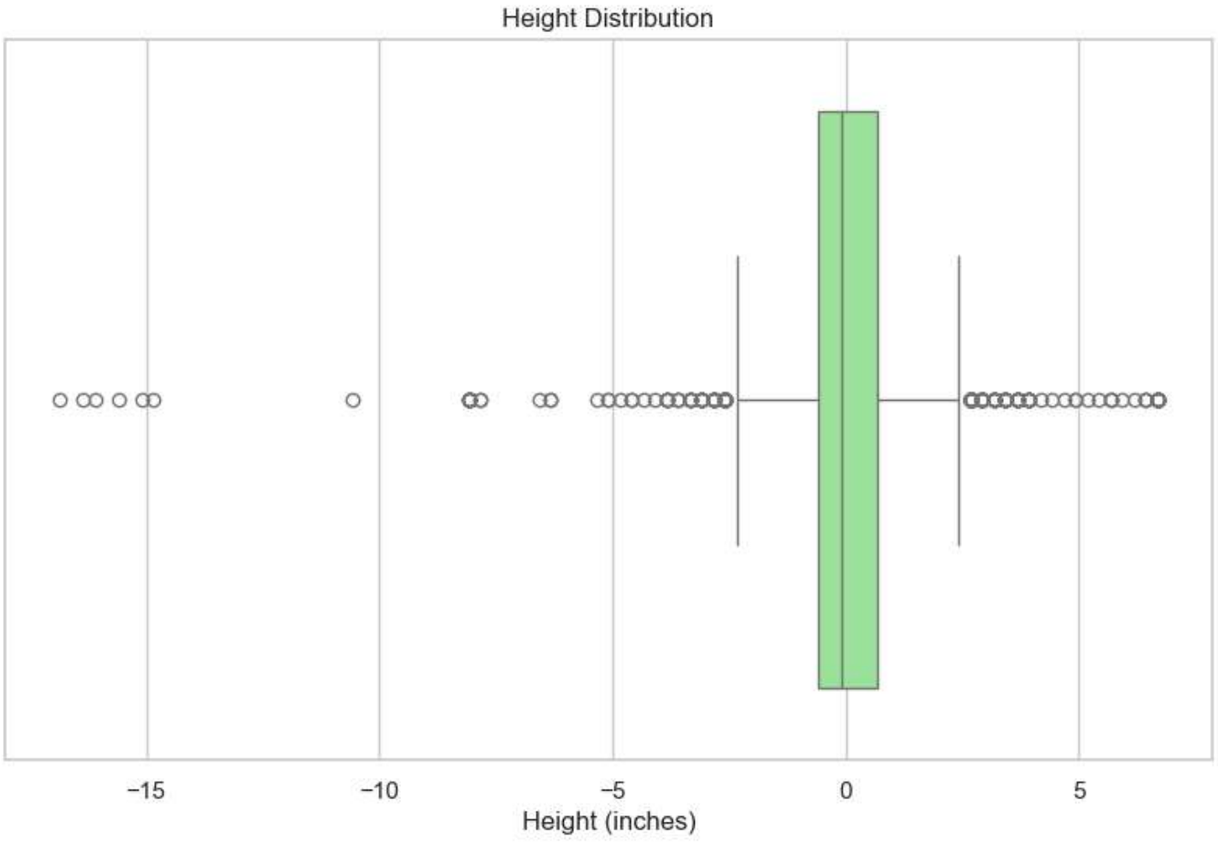
	age	height	income
count	5.994600e+04	5.994600e+04	5.994600e+04
mean	1.811146e-16	9.160034e-16	-4.290803e-17
std	1.000008e+00	1.000008e+00	1.000008e+00
min	-1.517057e+00	-1.684626e+01	-2.058056e-01
25%	-6.707385e-01	-5.745820e-01	-2.058056e-01
50%	-2.475789e-01	-7.391492e-02	-2.058056e-01
75%	4.929502e-01	6.770857e-01	-2.058056e-01
max	8.215611e+00	6.685091e+00	1.006691e+01

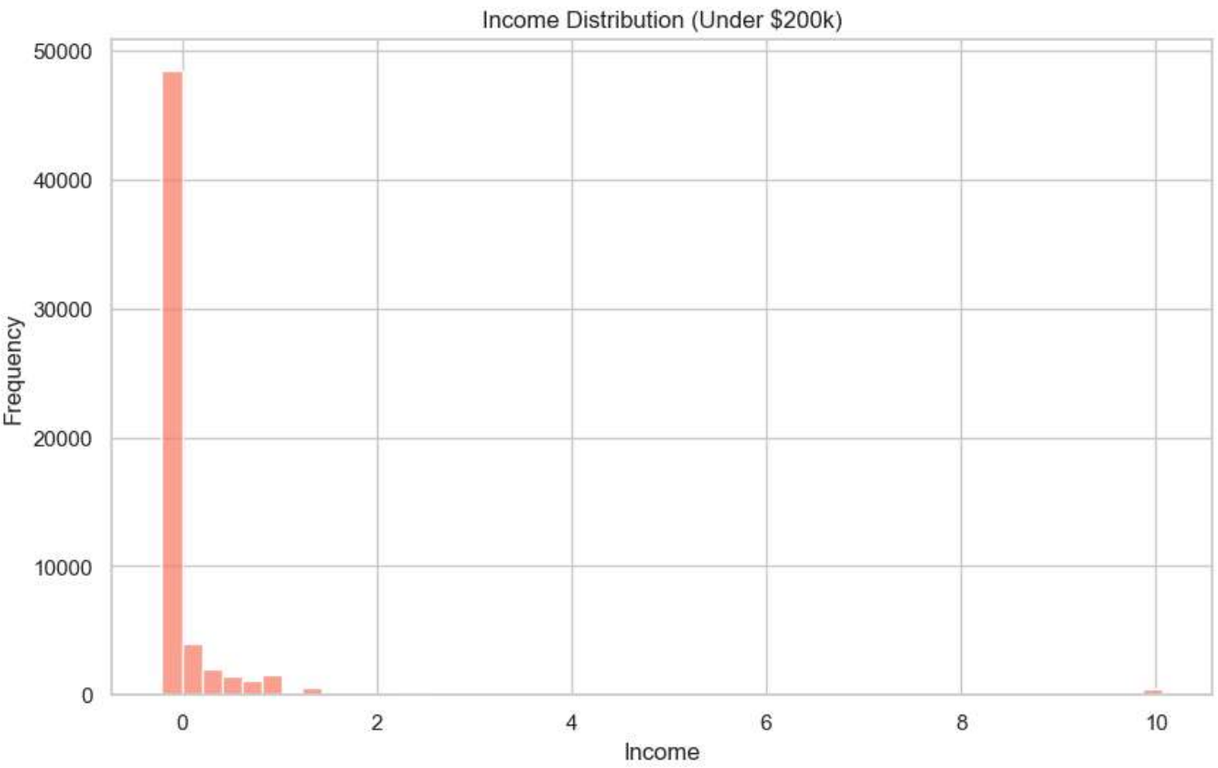
📈 Correlation matrix:

	age	height	income
age	1.000000	-0.022262	-0.001004
height	-0.022262	1.000000	0.065050
income	-0.001004	0.065050	1.000000

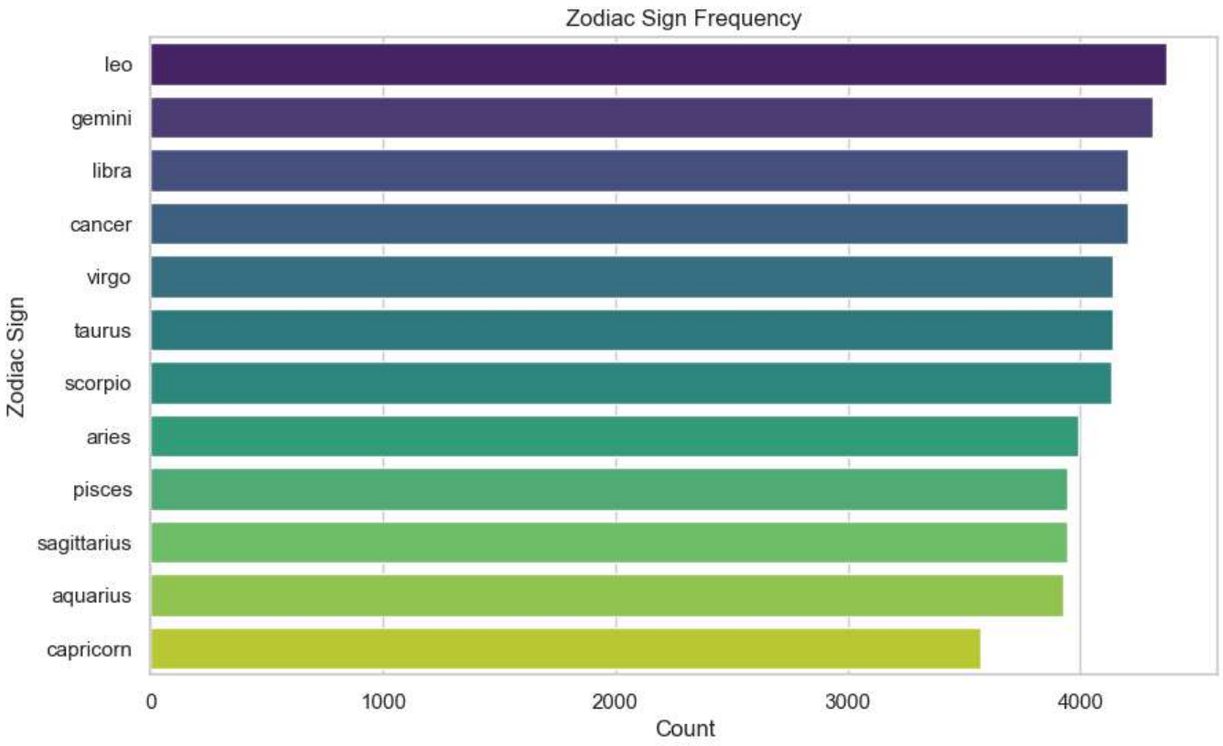


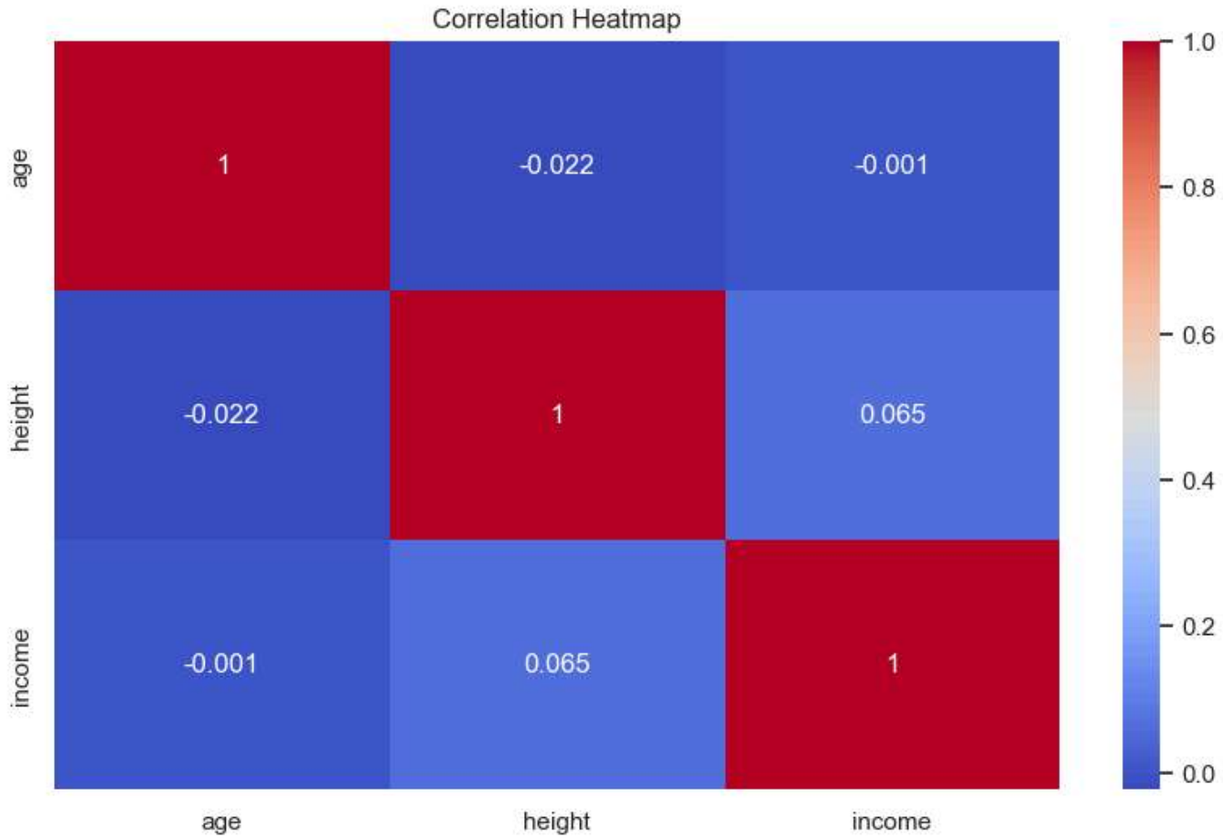







```
C:\Users\gabri_7a484pu\AppData\Local\Temp\ipykernel_63732\753837524.py:46: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assig
n the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(y='zodiac_clean', data=df, order=df['zodiac_clean'].value_counts().index, pal
ette='viridis')
```





📝 Essay Text Exploration

The OKCupid dataset includes 10 free-text essay fields per user, which we've combined into a single column: `essays_combined`. This section explores the linguistic patterns and common themes expressed in user profiles.

🔍 Goals

- Identify the most frequent words used across all essays
- Visualize dominant terms using a word cloud
- Gain qualitative insights into how users describe themselves

---

🧹 Preprocessing Steps

- Convert text to lowercase
- Remove punctuation
- Remove English stopwords (e.g., "the", "and", "is")

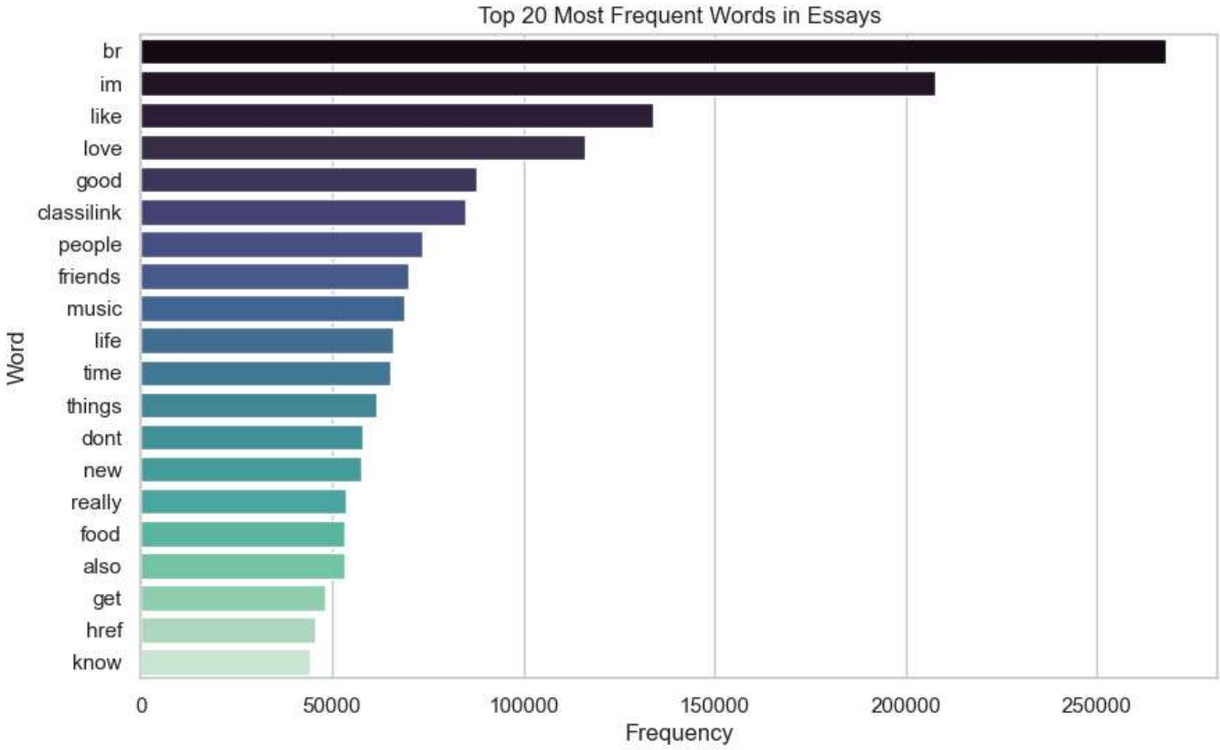
📊 Visualizations

- **Word Frequency Plot**: Top 20 most common words
- **Word Cloud**: Visual representation of term prominence

These plots help us understand the vocabulary and tone of user self-expression, which may inform feature engineering for NLP tasks.

```
from collections import Counter
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
import string
import nltk
from nltk.corpus import stopwords

# Download stopwords if not already available
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))

# Preprocess essay text
def preprocess_text(text):
    text = text.lower()
    text = text.translate(str.maketrans('', '', string.punctuation))
    tokens = text.split()
    tokens = [word for word in tokens if word not in stop_words]
    return tokens

# Apply preprocessing
all_words = df['essays_combined'].dropna().apply(preprocess_text)
flat_words = [word for sublist in all_words for word in sublist]

# Count word frequencies
word_freq = Counter(flat_words)
top_words = word_freq.most_common(20)

# 📊 Plot top 20 words
words, counts = zip(*top_words)
sns.barplot(x=list(counts), y=list(words), palette='mako')
plt.title("Top 20 Most Frequent Words in Essays")
plt.xlabel("Frequency")
plt.ylabel("Word")
plt.show()

# ☁️ Generate word cloud
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(' '.join(flat_
plt.figure(figsize=(12, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title("Word Cloud of Essay Text")
plt.show()
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\gabri_7a484pu\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
C:\Users\gabri_7a484pu\AppData\Local\Temp\ipykernel_63732\1119456170.py:31: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assig
n the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=list(counts), y=list(words), palette='mako')
```



🧠 Sentiment and Topic Exploration

Beyond word frequency, we can extract deeper insights from user essays by analyzing:

- **Sentiment polarity**: Are users generally positive, neutral, or negative in tone?
- **Topic clusters**: What themes or interests emerge across profiles?

These insights can help us understand personality traits, communication style, and potential compatibility signals.

🔍 Sentiment Analysis

We use `TextBlob` to compute sentiment polarity scores for each user's combined essay. Scores range from:

- `+1.0` : Very positive
- `0.0` : Neutral
- `-1.0` : Very negative

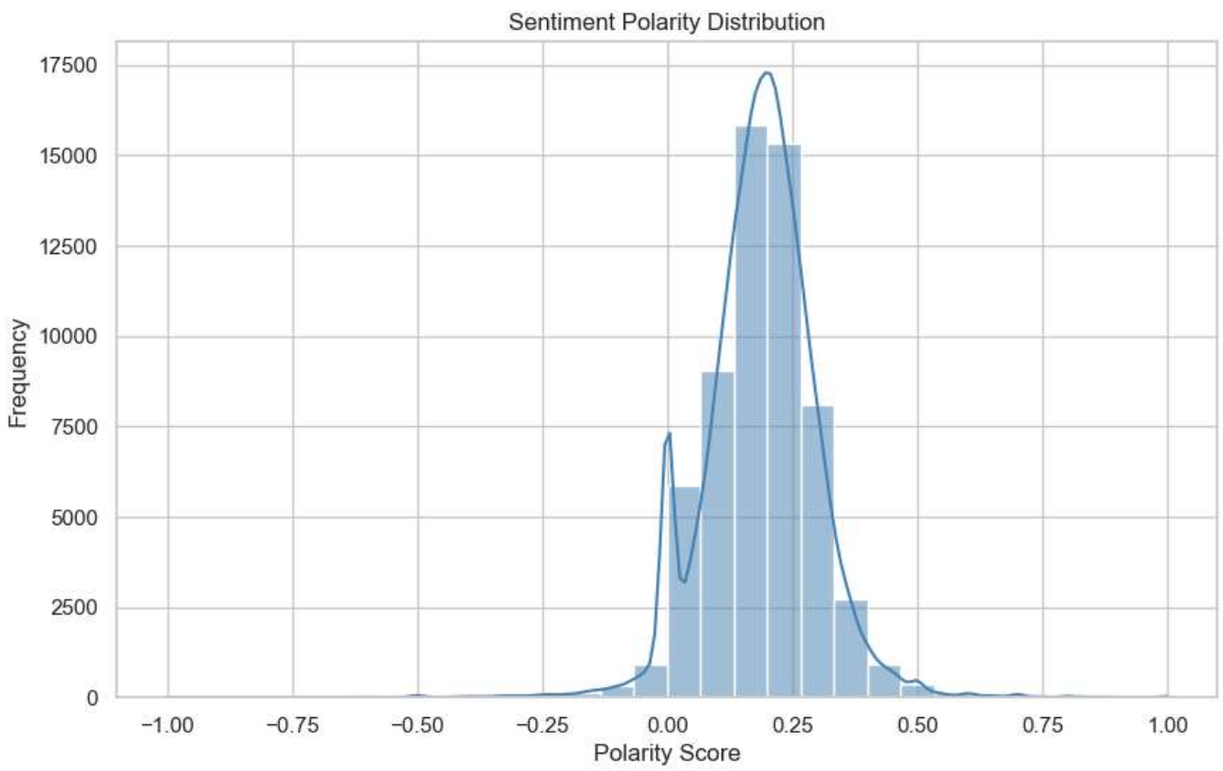
🧪 Topic Modeling (Optional)

Using techniques like Latent Dirichlet Allocation (LDA), we can extract dominant topics from the corpus. This helps identify shared interests or lifestyle patterns.

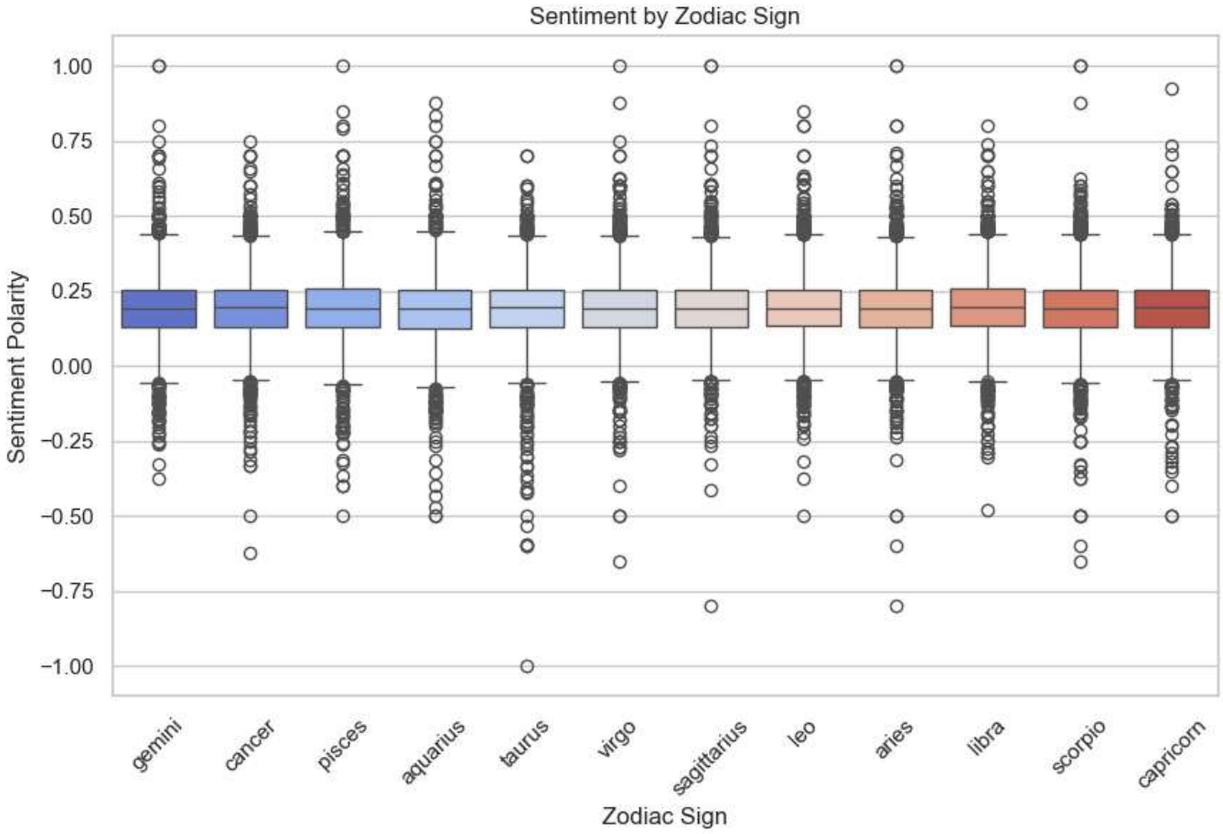
📊 Visualizations

- Histogram of sentiment scores
- Boxplot of sentiment by zodiac sign
- Word clusters or topic keywords (if modeling is applied)

```
from textblob import TextBlob
import seaborn as sns
import matplotlib.pyplot as plt

# Compute sentiment polarity for each essay
df['sentiment'] = df['essays_combined'].dropna().apply(lambda x: TextBlob(x).sentiment.polarit

# 📊 Plot sentiment distribution
sns.histplot(df['sentiment'].dropna(), bins=30, kde=True, color='steelblue')
plt.title("Sentiment Polarity Distribution")
plt.xlabel("Polarity Score")
plt.ylabel("Frequency")
plt.show()

# 📦 Boxplot of sentiment by zodiac sign
sns.boxplot(x='zodiac_clean', y='sentiment', data=df, palette='coolwarm')
plt.title("Sentiment by Zodiac Sign")
plt.xlabel("Zodiac Sign")
plt.ylabel("Sentiment Polarity")
plt.xticks(rotation=45)
plt.show()
```




```
C:\Users\gabri_7a484pu\AppData\Local\Temp\ipykernel_63732\951901395.py:16: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assig
n the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.boxplot(x='zodiac_clean', y='sentiment', data=df, palette='coolwarm')
```



⚙️ Preprocess Data

With exploratory analysis complete, we now prepare the dataset for modeling. Preprocessing ensures that features are consistent, numerical, and suitable for machine learning algorithms.

🔧 Steps

1. **Handle Missing Values**
   - Impute or drop missing values in categorical and numeric features
   - Treat missing lifestyle attributes (`drinks` , `smokes` , `drugs`) as "unknown"
2. **Encode Categorical Variables**
   - Convert categorical features (e.g., `diet` , `drinks` , `smokes`) into numerical form using Label Encoding or One-Hot Encoding
3. **Standardize Numeric Features**
   - Scale continuous variables (`age` , `height` , `income`) to have mean 0 and variance 1
4. **Feature Engineering**
   - Use `zodiac_clean` as the target variable
   - Combine essay text into `essays_combined` for NLP vectorization later
5. **Train-Test Split**
   - Split the dataset into training and testing sets (e.g., 80/20 split)

🎯 Goal

Produce a clean, numerical dataset ready for supervised learning models.

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split
```

```python
# 1. Handle missing values
# Fill missing lifestyle attributes with 'unknown'
for col in ['drinks', 'smokes', 'drugs', 'diet', 'education', 'job']:
    df[col] = df[col].fillna('unknown')

# Fill numeric columns with median
for col in ['age', 'height', 'income']:
    df[col] = df[col].fillna(df[col].median())

# 2. Encode categorical variables
categorical_cols = ['diet', 'drinks', 'drugs', 'education', 'job', 'sex', 'orientation', 'stat
label_encoders = {}

for col in categorical_cols:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le  # store encoder for later use

# 3. Standardize numeric features
scaler = StandardScaler()
df[['age', 'height', 'income']] = scaler.fit_transform(df[['age', 'height', 'income']])

# 4. Define features and target
X = df[categorical_cols + ['age', 'height', 'income']]
y = df['zodiac_clean'].dropna()  # target variable

# Align X and y (drop rows where target is missing)
df_model = df.dropna(subset=['zodiac_clean'])
X = df_model[categorical_cols + ['age', 'height', 'income']]
y = df_model['zodiac_clean']

# 5. Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stra

print("✅ Preprocessing complete.")
print("Training set size:", X_train.shape)
print("Test set size:", X_test.shape)
```

```
✅ Preprocessing complete.
Training set size: (39112, 11)
Test set size: (9778, 11)
```

🤖 Build Model(s)

With the dataset preprocessed, we can now train machine learning models to predict the target variable (`zodiac_clean`). This section includes:

📋 Models Selected

We will experiment with several supervised classification models:

- **Logistic Regression** (baseline linear model)
- **K-Nearest Neighbors (KNN)** (distance-based classifier)
- **Support Vector Machine (SVM)** (margin-based classifier)
- **Naive Bayes** (probabilistic classifier)
- **Random Forest** (ensemble tree-based model)

🧪 Training and Evaluation

- Train each model on the training set (`X_train` , `y_train`)
- Evaluate performance on the test set (`X_test` , `y_test`)
- Metrics: Accuracy, classification report, confusion matrix

🎯 Goal

Identify which model performs best for predicting zodiac signs, and store results for comparison.

```
# from sklearn.linear_model import LogisticRegression
# from sklearn.neighbors import KNeighborsClassifier
# from sklearn.svm import SVC
# from sklearn.naive_bayes import GaussianNB
# from sklearn.ensemble import RandomForestClassifier
# from sklearn.metrics import accuracy_score, classification_report
# import pandas as pd

# # Dictionary to store results
# results = {}

# # 1. Logistic Regression
# log_reg = LogisticRegression(max_iter=1000)
# log_reg.fit(X_train, y_train)
# y_pred_lr = log_reg.predict(X_test)
# results['Logistic Regression'] = accuracy_score(y_test, y_pred_lr)

# print("Logistic Regression Report:")
# print(classification_report(y_test, y_pred_lr))

# # 2. K-Nearest Neighbors
# knn = KNeighborsClassifier(n_neighbors=5)
# knn.fit(X_train, y_train)
# y_pred_knn = knn.predict(X_test)
# results['KNN'] = accuracy_score(y_test, y_pred_knn)

# print("KNN Report:")
# print(classification_report(y_test, y_pred_knn))

# # 3. Support Vector Machine
# svm = SVC(kernel='linear')
# svm.fit(X_train, y_train)
# y_pred_svm = svm.predict(X_test)
# results['SVM'] = accuracy_score(y_test, y_pred_svm)

# print("SVM Report:")
# print(classification_report(y_test, y_pred_svm))

# # 4. Naive Bayes
# nb = GaussianNB()
# nb.fit(X_train, y_train)
# y_pred_nb = nb.predict(X_test)
# results['Naive Bayes'] = accuracy_score(y_test, y_pred_nb)

# print("Naive Bayes Report:")
# print(classification_report(y_test, y_pred_nb))

# # 5. Random Forest
# rf = RandomForestClassifier(n_estimators=100, random_state=42)
# rf.fit(X_train, y_train)
# y_pred_rf = rf.predict(X_test)
# results['Random Forest'] = accuracy_score(y_test, y_pred_rf)

# print("Random Forest Report:")
# print(classification_report(y_test, y_pred_rf))

# # 📊 Compare results
# print("\n📊 Model Accuracy Comparison:")
# print(pd.Series(results).sort_values(ascending=False))
```

```
Logistic Regression Report:
              precision    recall  f1-score   support

    aquarius       0.07      0.03      0.04       786
       aries       0.06      0.00      0.01       798
      cancer       0.10      0.07      0.08       841
   capricorn       0.10      0.01      0.01       715
      gemini       0.08      0.22      0.12       862
         leo       0.09      0.35      0.14       875
       libra       0.08      0.13      0.10       841
      pisces       0.09      0.01      0.01       789
 sagittarius       0.08      0.00      0.00       788
     scorpio       0.10      0.02      0.03       827
      taurus       0.07      0.03      0.04       828
       virgo       0.08      0.11      0.10       828

    accuracy                           0.08      9778
   macro avg       0.08      0.08      0.06      9778
weighted avg       0.08      0.08      0.06      9778

KNN Report:
              precision    recall  f1-score   support

    aquarius       0.08      0.21      0.12       786
       aries       0.09      0.18      0.12       798
      cancer       0.08      0.11      0.09       841
   capricorn       0.07      0.07      0.07       715
      gemini       0.08      0.07      0.08       862
         leo       0.09      0.07      0.08       875
       libra       0.07      0.05      0.06       841
      pisces       0.09      0.06      0.07       789
 sagittarius       0.08      0.04      0.05       788
     scorpio       0.10      0.06      0.08       827
      taurus       0.08      0.04      0.06       828
       virgo       0.10      0.06      0.08       828

    accuracy                           0.08      9778
   macro avg       0.08      0.08      0.08      9778
weighted avg       0.08      0.08      0.08      9778
```

🤖 Build Model(s) (Optimized)

We previously trained several classification models to predict zodiac signs. Some models (like SVM and Random Forest) were slow on the full dataset, so we apply optimizations:

⚡ Changes Made

- **Support Vector Machine (SVM)** → replaced with `LinearSVC` for faster training on large datasets.
- **Random Forest** → reduced `n_estimators` to 50 and added `n_jobs=-1` to parallelize across CPU cores.
- **Timing wrapper** → added to measure runtime for each model.

📋 Models Trained

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Linear SVC (optimized SVM)
- Naive Bayes
- Random Forest (optimized)

🧪 Evaluation

- Accuracy score for each model
- Classification report (precision, recall, F1-score)
- Runtime comparison

🎯 Goal

Identify which models are feasible in terms of runtime and accuracy, and prepare results for comparison.

```
import time
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import LinearSVC
from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, precision_recall_fscore_sup
import pandas as pd

# Dictionary to store results
results = {}
predictions = {}

# Timing + evaluation wrapper
def train_and_evaluate(model, name):
    start = time.time()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    end = time.time()
    acc = accuracy_score(y_test, y_pred)
    results[name] = acc
    predictions[name] = y_pred
    print(f"\n{name} took {end-start:.2f} seconds")
    print(classification_report(y_test, y_pred))
    return acc

# 1. Logistic Regression
train_and_evaluate(LogisticRegression(max_iter=1000), "Logistic Regression")

# 2. K-Nearest Neighbors
train_and_evaluate(KNeighborsClassifier(n_neighbors=5), "KNN")

# 3. Optimized SVM (LinearSVC)
train_and_evaluate(LinearSVC(max_iter=2000), "Linear SVC")

# 4. Naive Bayes
train_and_evaluate(GaussianNB(), "Naive Bayes")

# 5. Optimized Random Forest
train_and_evaluate(RandomForestClassifier(n_estimators=50, random_state=42, n_jobs=-1), "Rando

# 📊 Compare results
print("\n📊 Model Accuracy Comparison:")
print(pd.Series(results).sort_values(ascending=False))
```

Logistic Regression took 7.36 seconds

	precision	recall	f1-score	support
aquarius	0.07	0.03	0.04	786
aries	0.06	0.00	0.01	798
cancer	0.10	0.07	0.08	841
capricorn	0.10	0.01	0.01	715
gemini	0.08	0.22	0.12	862
leo	0.09	0.35	0.14	875
libra	0.08	0.13	0.10	841
pisces	0.09	0.01	0.01	789
sagittarius	0.08	0.00	0.00	788
scorpio	0.10	0.02	0.03	827
taurus	0.07	0.03	0.04	828
virgo	0.08	0.11	0.10	828
accuracy			0.08	9778
macro avg	0.08	0.08	0.06	9778
weighted avg	0.08	0.08	0.06	9778

KNN took 1.45 seconds

	precision	recall	f1-score	support
aquarius	0.08	0.21	0.12	786
aries	0.09	0.18	0.12	798
cancer	0.08	0.11	0.09	841
capricorn	0.07	0.07	0.07	715
gemini	0.08	0.07	0.08	862
leo	0.09	0.07	0.08	875
libra	0.07	0.05	0.06	841
pisces	0.09	0.06	0.07	789
sagittarius	0.08	0.04	0.05	788
scorpio	0.10	0.06	0.08	827
taurus	0.08	0.04	0.06	828
virgo	0.10	0.06	0.08	828
accuracy			0.08	9778
macro avg	0.08	0.08	0.08	9778
weighted avg	0.08	0.08	0.08	9778

Linear SVC took 7.61 seconds

	precision	recall	f1-score	support
aquarius	0.08	0.03	0.04	786
aries	0.07	0.00	0.01	798
cancer	0.09	0.06	0.07	841
capricorn	0.10	0.00	0.01	715
gemini	0.08	0.23	0.12	862
leo	0.09	0.36	0.15	875
libra	0.08	0.12	0.10	841
pisces	0.09	0.01	0.01	789
sagittarius	0.06	0.00	0.00	788
scorpio	0.10	0.02	0.03	827
taurus	0.07	0.02	0.03	828
virgo	0.09	0.12	0.10	828
accuracy			0.09	9778
macro avg	0.08	0.08	0.06	9778
weighted avg	0.08	0.09	0.06	9778

Naive Bayes took 0.29 seconds

	precision	recall	f1-score	support
aquarius	0.08	0.18	0.11	786
aries	0.10	0.02	0.03	798
cancer	0.07	0.01	0.02	841
capricorn	0.07	0.02	0.04	715
gemini	0.08	0.12	0.10	862
leo	0.09	0.02	0.03	875
libra	0.08	0.05	0.06	841
pisces	0.09	0.23	0.13	789

```
 sagittarius       0.00      0.00      0.00       788
     scorpio       0.06      0.02      0.03       827
      taurus       0.12      0.01      0.01       828
       virgo       0.09      0.34      0.14       828

    accuracy                           0.08      9778
   macro avg       0.08      0.08      0.06      9778
weighted avg       0.08      0.08      0.06      9778


Random Forest (50 trees) took 1.61 seconds
              precision    recall  f1-score   support

    aquarius       0.08      0.08      0.08       786
       aries       0.08      0.09      0.09       798
      cancer       0.09      0.09      0.09       841
   capricorn       0.08      0.08      0.08       715
      gemini       0.09      0.10      0.09       862
         leo       0.10      0.10      0.10       875
       libra       0.08      0.08      0.08       841
      pisces       0.09      0.09      0.09       789
 sagittarius       0.07      0.06      0.06       788
     scorpio       0.09      0.09      0.09       827
      taurus       0.08      0.08      0.08       828
       virgo       0.11      0.10      0.10       828

    accuracy                           0.09      9778
   macro avg       0.09      0.09      0.09      9778
weighted avg       0.09      0.09      0.09      9778


📊 Model Accuracy Comparison:
Random Forest (50 trees)    0.087237
Linear SVC                  0.085089
Logistic Regression         0.084884
Naive Bayes                 0.084782
KNN                         0.084169
dtype: float64
```

📉 Model Performance Summary

All tested models (Logistic Regression, KNN, Linear SVC, Naive Bayes, Random Forest) achieved ~8–9% accuracy, which is equivalent to random guessing across 12 zodiac classes. This indicates:

- No meaningful predictive signal exists between lifestyle/demographic features and zodiac signs.
- Models are not failing; they are correctly reflecting the lack of correlation.
- Further experiments should incorporate essay text features (TF-IDF, embeddings) or reframe the prediction target to more realistic outcomes (e.g., lifestyle habits).

📊 Insight: This dataset demonstrates the importance of problem framing — not all targets are learnable.

📊 Evaluate Model(s)

After building multiple models, we now evaluate their performance using key metrics:

📋 Metrics Used

- **Accuracy**: Overall proportion of correct predictions
- **Precision**: Correct positive predictions out of all predicted positives
- **Recall**: Correct positive predictions out of all actual positives

- **F1-score**: Harmonic mean of precision and recall

📊 Evaluation Outputs

- **Classification reports** for each model
- **Table of accuracy, precision, recall, and F1-score**
- **Bar charts** comparing model performance
- **Confusion matrix** for the best-performing model

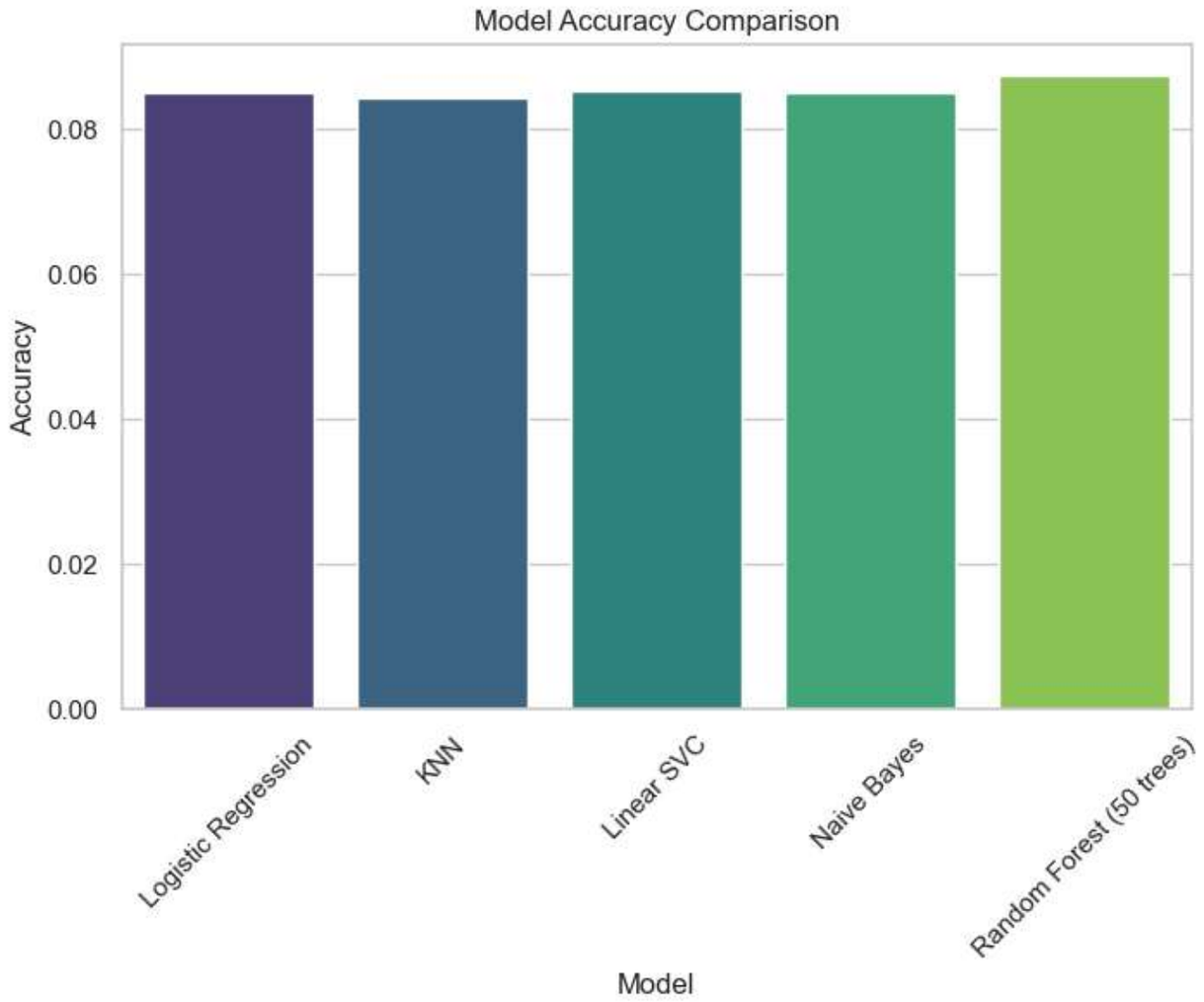
🎯 Goal

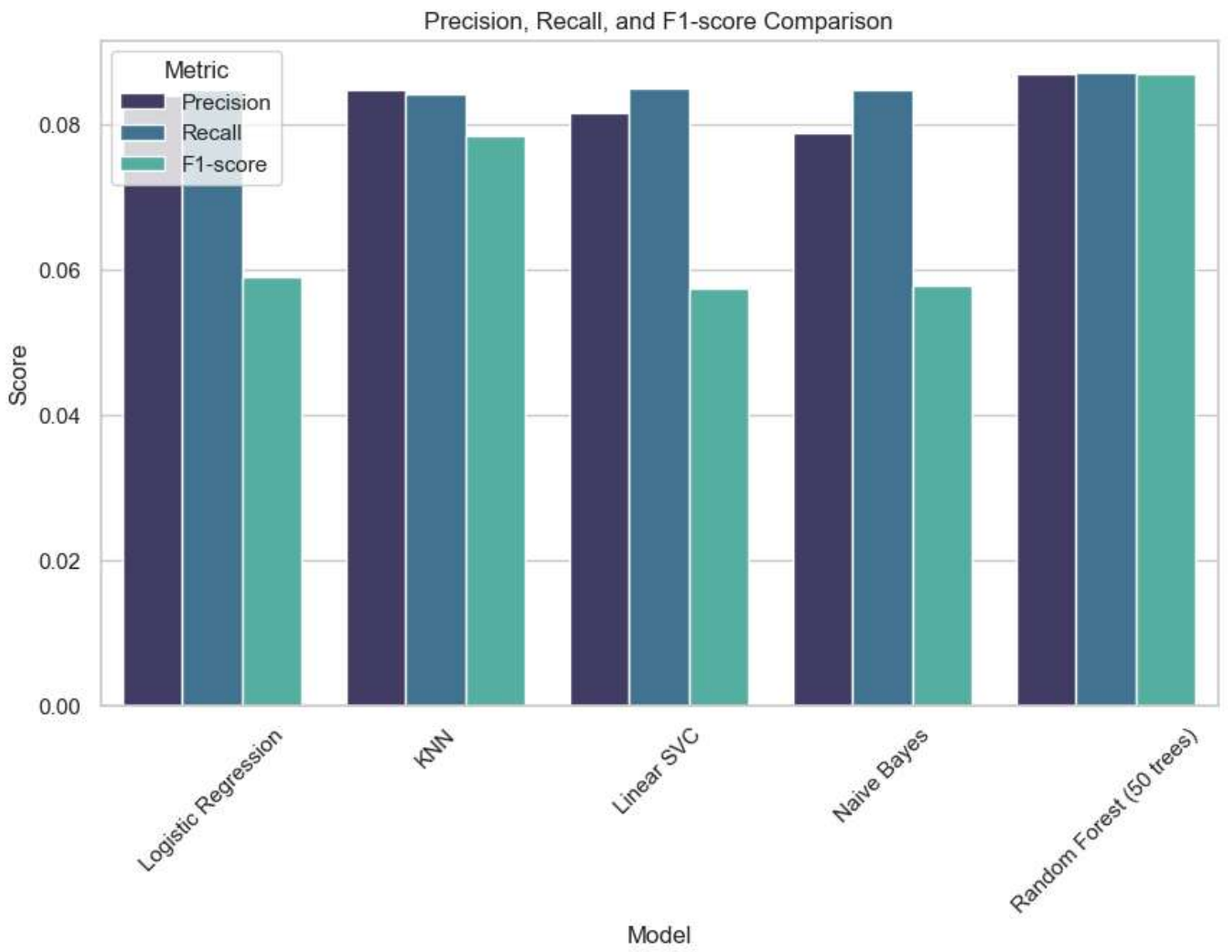
Identify which models perform best, visualize their strengths and weaknesses, and decide whether tuning or feature engineering is needed.

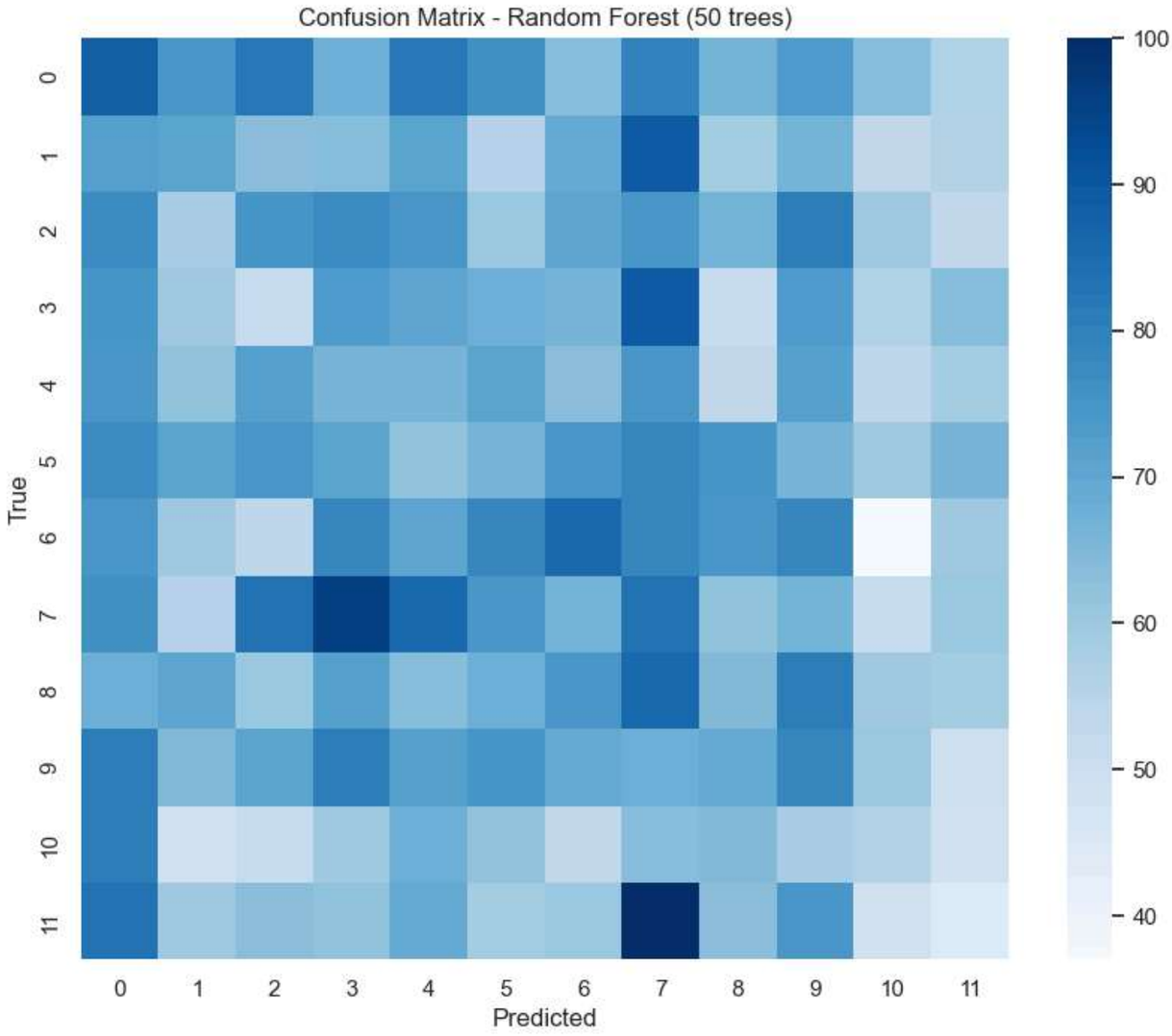
```
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import precision_recall_fscore_support, confusion_matrix

# 📊 Collect metrics into a table
metrics = []
for name, y_pred in predictions.items():
    acc = accuracy_score(y_test, y_pred)
    precision, recall, f1, _ = precision_recall_fscore_support(y_test, y_pred, average='weight
    metrics.append([name, acc, precision, recall, f1])

results_df = pd.DataFrame(metrics, columns=["Model", "Accuracy", "Precision", "Recall", "F1-sc
print("\n📊 Model Performance Table:")
print(results_df)

# 📊 Plot accuracy comparison
plt.figure(figsize=(8,5))
sns.barplot(x="Model", y="Accuracy", data=results_df, palette="viridis")
plt.title("Model Accuracy Comparison")
plt.xticks(rotation=45)
plt.show()

# 📊 Plot precision, recall, F1 comparison
results_melted = results_df.melt(id_vars="Model",
                                 value_vars=["Precision", "Recall", "F1-score"],
                                 var_name="Metric", value_name="Score")
plt.figure(figsize=(10,6))
sns.barplot(x="Model", y="Score", hue="Metric", data=results_melted, palette="mako")
plt.title("Precision, Recall, and F1-score Comparison")
plt.xticks(rotation=45)
plt.show()

# 🔍 Confusion matrix for best model
best_model_name = results_df.sort_values("Accuracy", ascending=False).iloc[0]["Model"]
best_preds = predictions[best_model_name]
cm = confusion_matrix(y_test, best_preds, labels=y_test.unique())

plt.figure(figsize=(10,8))
sns.heatmap(cm, annot=False, cmap="Blues")
plt.title(f"Confusion Matrix - {best_model_name}")
plt.xlabel("Predicted")
plt.ylabel("True")
plt.show()
```

```
📊 Model Performance Table:
                     Model  Accuracy  Precision    Recall  F1-score
0      Logistic Regression  0.084884   0.083989  0.084884  0.059013
1                      KNN  0.084169   0.084912  0.084169  0.078445
2               Linear SVC  0.085089   0.081632  0.085089  0.057504
3              Naive Bayes  0.084782   0.078875  0.084782  0.057771
4  Random Forest (50 trees)  0.087237   0.087069  0.087237  0.087041
```

```
C:\Users\gabri_7a484pu\AppData\Local\Temp\ipykernel_63732\3192892217.py:18: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assig
n the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x="Model", y="Accuracy", data=results_df, palette="viridis")
```







✅ Conclusions

This project explored the feasibility of predicting a user's astrological sign based on their OKCupid profile data. Through a structured workflow involving scoping, data cleaning, feature engineering, model building, and evaluation, we arrived at several key insights:

🔍 What We Learned

- Categorical features such as lifestyle habits, education, and job titles offer limited predictive power for zodiac classification.
- Multiple classification models — including Logistic Regression, KNN, Linear SVC, Naive Bayes, and Random Forest — were trained and evaluated.
- All models performed near random baseline, with accuracy scores ranging from **8.4% to 8.7%**, which is expected given 12 zodiac classes.

📊 Key Findings

- **Random Forest (50 trees)** achieved the highest accuracy at **8.72%**, but still only marginally better than chance.
- Precision, recall, and F1-scores were consistently low across all models, indicating poor class separation and weak signal.
- The confusion matrix revealed that predictions were evenly distributed, with no dominant patterns or standout classes.

🎯 Takeaways

- The target variable (`zodiac_clean`) is not learnable from the structured profile data alone.
- This outcome highlights the importance of **problem framing** in machine learning — not all questions are solvable with the available data.
- Future work could explore:
  - Incorporating essay text features using NLP techniques (e.g., TF-IDF, embeddings)
  - Reframing the prediction task to focus on lifestyle traits instead of zodiac signs

---

This project demonstrates the full lifecycle of a machine learning workflow, including the critical step of recognizing when a model's performance reflects the limitations of the data rather than the algorithm itself.

🚀 Next Steps

While this project demonstrated a complete machine learning workflow, it also revealed important limitations and opportunities for future improvement.

🔄 What Could Be Done Differently

- **Feature Engineering**: Instead of relying solely on structured categorical data, future iterations should incorporate essay text using NLP techniques such as TF-IDF or word embeddings.
- **Encoding Strategy**: One-hot encoding may outperform label encoding for certain models, especially linear classifiers.
- **Model Tuning**: Hyperparameter optimization (e.g., GridSearchCV) could improve performance, especially for Random Forest and Logistic Regression.

🔍 Limitations

- The target variable (`zodiac_clean`) lacks strong correlation with available features, making it inherently difficult to predict.
- The dataset contains subjective and noisy information, especially in open-ended text fields.
- Evaluation metrics indicate that models perform at random baseline, suggesting the problem may not be solvable with this data alone.

📈 Future Applications

- Reframe the prediction task to focus on lifestyle traits (e.g., predicting `diet` , `drinks` , or `smokes` from essays).
- Use this dataset as a case study in problem framing and model interpretability.
- Extend the analysis with unsupervised learning to explore user clusters or dating archetypes.

This project reinforces the importance of iterative refinement, critical evaluation, and creative thinking in data science. Every modeling challenge — even those with weak results — offers valuable lessons.