Page 1: Model Performance Analysis

UI & Data Controls:

- **Title & Layout: The page is set up with a centered, wide layout. The left column contains controls and dropdowns while the right column displays the visualization and summary table.**
- **Controls:**
  - **NLP Approach Dropdown: Options include LDA, GPT, COMBINED, and LIWC.**
  - **Idiographic/Nomothetic Dropdown: Changes based on the NLP approach (for GPT, defaults to "Nomothetic" for data availability).**
  - **Outcome Dropdown: Selects the outcome (e.g., Negative Affect, Angry, Nervous, Sad). For GPT, outcome values are standardized (e.g., "negative affect" becomes "na") and filtered from the file.**

Graph Details:

- **Graph Type: Violin Plot**
- **Axes:**
  - **X-Axis: Represents the "ML Model." In GPT mode, it shows a single category ("GPT"), while in other cases it shows two categories: "Elastic Net (en)" and "Random Forest (rf)."**
  - **Y-Axis: Represents the R^2 values, showing the distribution of model performance.**
- **Data & Filtering:**
  - **GPT: Loads from modelfit_gpt_all.csv filtered by outcome and (if available) the nomothetic/idiographic value.**
  - **Non-GPT Models: Loads separate CSV files for Elastic Net and Random Forest; these datasets are concatenated.**
- **Colors:**
  - **GPT Mode: The violin is rendered in blue.**
  - **Non-GPT: Elastic Net is shown in red, and Random Forest is shown in blue. The colors help differentiate the model types and make it easier to visually compare their R^2 distributions.**
- **Additional Features: The violin plot includes boxplots and all data points are shown, offering insight into both central tendency and spread.**

Summary Table:

- **Content: A table summarizing key statistics per model including the mean (with standard deviation), count (N), range (min to max R^2), and the count of p-values below 0.05.**
- **Calculation: Each metric is computed on the filtered dataset and then presented in a neat tabular format.**

Page 2: Data Table View – Model Performance per Participant

UI & Data Controls:

- **Layout: The page uses a centered, wide layout with controls in the left column and a data table in the right column.**
- **Controls:**
  - **NLP Approach, Idiographic/Nomothetic, ML Model, and Outcome Dropdowns: These determine which CSV file is loaded. For GPT, the outcome is chosen from unique values in the file; otherwise, the outcome is standardized (e.g., "negative affect" to "na").**

Data Table:

- **Table Type: Interactive Data Table**
- **Displayed Columns: Typically includes id, participant, r, r2, rmse, and p_value.**
- **Purpose: Presents participant-level performance metrics from the selected file.**
- **Axes Explanation:**
  - **Unlike a graph, the table shows rows for each participant and columns for each metric.**
- **Color and Formatting:**

- **No explicit colors are used in the table; however, column headers and data formatting help in quick reading.**
- **Interaction: The table is scrollable with a fixed height (600px) and width (1200px).**

Page 3: True vs Predicted NA Levels
UI & Data Controls:

- **Title & Refresh: The title "True vs Predicted NA Levels" is displayed along with a "Clear All" button that resets the session state.**
- **Dynamic Graph Sections: The user can add multiple graph sections; each section has its own controls for outcome, ML model, and participant selection.**

Graph Details:

- **Graph Type: Line Chart**
- **Axes:**
  - **X-Axis: Represents time (the "time" column from the dataset).**
  - **Y-Axis: Represents NA levels; two lines are plotted—one for the actual (true) NA levels and one for the predicted estimates.**
- **Data & Filtering:**
  - **Data is loaded from a CSV file whose filename is constructed based on the selected ML model, outcome, and nomothetic/idiographic value.**
  - **After selecting a participant, the data is filtered to include only rows for that participant.**
- **Colors:**
  - **Actual NA Levels: Plotted in teal with a solid line.**
  - **Predicted NA Levels: Plotted in red using a dashed line.**
  - **Correlation Calculation: The correlation coefficient between the actual and predicted NA values is computed and displayed in the title.**
- **Additional Features:**
  - **The graph has fixed dimensions (900 px wide, 400 px high) to ensure consistency, and tooltips show detailed information when hovering.**

Page 4: Best Model Performance per Participant
UI & Data Controls:

- **Layout: The page features controls on the left (including checkboxes and dropdowns) and a results section on the right.**
- **Controls:**
  - **Checkbox: "Include both Elastic Net and Random Forest" – when checked, both models are compared; when unchecked, a dropdown allows selection of one model.**
  - **Outcome Dropdown: Allows selection of an outcome (e.g., Negative Affect, Angry, Nervous, Sad).**

Data Processing:

- **Aggregation: Data from multiple CSV files (covering different NLP approaches and both idiographic and nomothetic cases) are merged and filtered.**
- **Selection: For each participant, the record with the highest R^2 is selected.**

Displayed Elements:

- **Data Table:**
  - **Content: Shows the best performance for each participant, including Participant, ML Model (if both are included), Nomothetic/Idiographic, NLP Approach, R^2, RMSE, P Value, and Counts.**
- **Pie Charts:**
  - **Graph Types: Pie Charts**
  - **Pie Chart 1 (Nomothetic vs. Idiographic):**
    - **Slices: Represent counts of participants per category.**

        - **Colors: Use Plotly's default discrete colors to differentiate the groups.**
    - **Pie Chart 2 (NLP Approaches):**
        - **Slices: Represent the frequency of each NLP approach; "comb" is re-labeled as "All text features combined."**
        - **Colors: Also use discrete color mapping.**
    - **Pie Chart 3 (ML Models):**
        - **Shown only if both models are included.**
        - **Slices: Represent the counts for Elastic Net and Random Forest.**
        - **Colors: Defined via a discrete map (Elastic Net in red, Random Forest in blue).**
- **Axes Explanation for Pie Charts:**
    - **Pie charts do not have traditional axes; the slices' sizes represent proportions or counts.**
- **Legend & Layout:**
    - **The charts are arranged in columns with legends and centered titles for clarity.**

Page 5: Feature Importance Heatmap (SHAP Values)
UI & Data Controls:

- **Layout: The page uses a two-column layout with controls in the right column and the heatmap visualization in the left column.**
- **Controls:**
    - **Outcome and Model Dropdowns: Users select the outcome (e.g., Negative Affect, Angry, etc.) and the ML model (Elastic Net (EN) or Random Forest (RF)).**
    - **Participant Multi-select: Allows selection of specific participants, or an "All" option can select every participant.**
    - **Dynamic Symmetric Slider: A slider allows users to set a symmetric threshold for feature importance values. The slider is linked to a callback that forces the two handles to be opposites (e.g., if one is set to –0.005, the other becomes 0.005).**
    - **Info Box: A message box below the slider explains which SHAP importance values will be included (those outside the threshold) and which will be filtered out.**

Graph Details:

- **Graph Type: Heatmap**
- **Axes:**
    - **X-Axis: Represents the selected participants. When "All" is chosen, the x-axis shows a single label ("All Participants"); otherwise, it shows individual participant names.**
    - **Y-Axis: Represents the features (variables) that passed the importance threshold.**
- **Data Processing:**
    - **Data is loaded from a CSV (named by model and outcome). After grouping and averaging, the data is pivoted so that features form the rows and participants the columns.**
    - **The pivoted data is filtered based on the importance threshold.**
- **Colors:**
    - **Color Scale: A custom color scale is used:**
        - **Deep Blue represents strong negative SHAP values.**
        - **White represents neutral importance.**
        - **Deep Red represents strong positive SHAP values.**
    - **Feature Label Colors: Each feature label on the y-axis is colored according to its NLP method (using a predefined color map such as LIWC in red, GPT in blue, etc.).**
- **Layout Adjustments:**
    - **The heatmap's height and width are dynamically set based on the number of features and participants.**
    - **A legend below the graph (built using HTML) explains the color mapping for the NLP methods.**

Page 6: Feature Importance per Participant (SHAP Value)

UI & Data Controls:

- **Layout: Similar to Page 5, this page uses a two-column layout with controls in the right column and the visualization in the left column.**
- **Controls:**
  - **Outcome & Model Dropdowns: Allow the user to select the outcome and ML model (with outcome normalized).**
  - **Participant Dropdown: Enables the selection of a single participant from the loaded dataset.**
  - **Performance Metrics Loading: A secondary CSV is loaded to extract performance metrics (R^2 and RMSE) for the selected participant, which are later shown in the graph title.**
  - **Dynamic Symmetric Slider: As on Page 5, a slider is provided for setting the threshold for feature importance. It uses a symmetric range and displays an info box.**

Graph Details:

- **Graph Type: SHAP Summary Scatter Plot with Vertical Lines**
- **Axes:**
  - **X-Axis: Represents the SHAP (feature importance) values.**
  - **Y-Axis: Represents features. The y-axis values are evenly spaced for each feature, and feature names are used as tick labels.**
- **Data Processing:**
  - **Data is filtered for the selected participant and then filtered based on the chosen importance threshold.**
  - **The data is sorted by NLP method (and importance) and assigned evenly spaced y-axis positions.**
- **Colors:**
  - **Vertical Lines & Markers: Each feature's importance is shown with a line extending from 0 to the SHAP value, with a dot at the end. The color of the line and dot corresponds to the feature's NLP method (e.g., LIWC is red, GPT is blue, etc.).**
  - **Legend: Additional dummy traces ensure every NLP method appears in the legend.**
- **Additional Features:**
  - **The plot title includes the participant's name along with their performance metrics (R^2 and RMSE).**
  - **A dashed vertical line at x=0 helps delineate positive from negative importance.**

Page 7: Feature Importance Analysis

UI & Data Controls:

- **Layout: The page is divided into a left column for controls and a right column for the graphs.**
- **Controls:**
  - **Dropdowns:**
    - **ML Model & Outcome: Users select the model and outcome (with outcome standardized).**
  - **Checkbox: "Include the variable 'Time'" (default is checked).**
  - **Participant Filtering: A slider lets users select the percentage of participants in each group based on performance (R^2).**
  - **Minimum Variable Occurrence Slider: Sets a threshold for how many participants must have a feature for it to be included.**
- **Data Processing:**
  - **Performance Data: Loaded from a performance CSV, participants are split into high and low R^2 groups.**
  - **Feature Importance Data: Loaded from another CSV, with an option to exclude the "Time" variable.**
  - **Aggregation: The absolute mean importance for each feature is computed separately for high and low R^2 groups. The absolute mean difference is also calculated.**

Graph Details:

- **Graph Type: Bar Charts (two separate charts)**
- **First Bar Chart – High vs. Low R^2 Groups:**
  - **Axes:**
    - **X-Axis: Represents the absolute mean value for each group.**
    - **Y-Axis: Lists the features (sorted by the high R^2 group's values) with labels color-coded by NLP method.**
  - **Bars:**
    - **Two sets of bars for each feature: one for the high R^2 group (colored red) and one for the low R^2 group (colored turquoise).**
- **Second Bar Chart – Absolute Mean Difference:**
  - **Axes:**
    - **X-Axis: Represents the absolute mean difference between the high and low R^2 groups.**
    - **Y-Axis: Lists the top features (sorted by difference) with color-coded labels.**
  - **Bars:**
    - **A single set of gray bars represents the difference.**
- **Colors:**
  - **The bar colors (red, turquoise, and gray) are chosen to clearly contrast the groups, and the y-axis labels are enhanced with HTML to include NLP method colors.**

Legend:

- **An HTML-based legend at the bottom explains the NLP method color coding.**

Page 8: Common Top Predictive Features
UI & Data Controls:

- **Layout: This page uses a sidebar layout.**
  - **Left Sidebar (Controls):**
    - **Model Selection Dropdown: Allows the choice between Elastic Net (EN) and Random Forest (RF).**
    - **Sliders:**
      - **Number of Features per Participant (determines how many top features per participant are selected).**
      - **Number of Variables in Figure (sets how many variables appear in the aggregate view).**
      - **SHAP Value Threshold (filters features based on a minimum absolute SHAP value).**
    - **Checkbox: "Use ABS values only" toggles the data source and indicates that only absolute values are considered.**
- **Data Processing:**
  - **For each emotion (na, sad, angry, nervous), a CSV file is loaded based on the selected model and whether ABS values are used.**
  - **Each participant's top features (based on the SHAP threshold) are aggregated. The count of participants selecting each feature is calculated, and the SHAP sign (Positive or Negative) is determined.**
- **Graph Details:**
  - **Graph Type: Stacked Horizontal Bar Charts**
  - **Axes:**
    - **X-Axis: Represents the "count" (or percentage of participants) in which the feature is selected.**
    - **Y-Axis: Represents the feature names, with labels color-coded based on their associated NLP method.**
  - **Bar Colors:**
    - **Bars are split by SHAP sign:**

        - **Positive Values: Shown in a teal-like color (rgb(0,182,185)).**
        - **Negative Values: Shown in a red hue (rgb(255,79,82)).**
- **Layout of Graphs:**
    - **The four emotion-specific bar charts are arranged in a 2×2 grid in the main area.**
- **Legend:**
    - **A legend below the charts (rendered in HTML) explains the NLP method colors.**