Report Development Documentation – Pre-Model Phase

From Initial Dataset Creation to report_style.pdf

CSS Styling System Overview

What is CSS and How It Works

CSS (Cascading Style Sheets) is a styling language that controls the visual appearance of HTML documents. In this project, we use a `style.css` file to make our markdown reports look professional and consistent.

Key Concepts for Beginners:

- **CSS File**: Contains rules that define colors, fonts, spacing, and layout
- **HTML Classes**: Tags in the markdown that tell the CSS which styles to apply
- **Linking**: The markdown file connects to the CSS file using a special link tag

How the CSS File is Imported

In `all_visual_data_report.md`, the CSS is imported on the very first line:

```html
<link rel="stylesheet" href="style.css">
```

This line tells the browser (or PDF converter) to:

1. Look for a file named `style.css` in the same folder
2. Apply all the styling rules from that file to the current document
3. Use those styles to format the HTML elements in the markdown

How CSS Classes Work in the Report

Throughout the markdown file, you'll see HTML `<div>` tags with `class` attributes like:

```html
<div class="executive-summary">
<div class="key-findings-box">
<div class="dashboard-table">
```

These classes correspond to rules in the CSS file that define:

- **Colors**: Background colors, text colors, borders
- **Typography**: Font sizes, weights, families
- **Layout**: Spacing, margins, padding, positioning
- **Special Effects**: Hover effects, shadows, borders

Key CSS Classes Used in This Report

Class Name	Purpose	What It Styles
`executive-summary`	Main content sections	Headers, paragraphs, lists
`key-findings-box`	Highlighted information boxes	Important findings with colored backgrounds
`dashboard-table`	Data tables	Statistical results with borders and spacing
`page-break`	Page breaks for PDF	Forces new pages in printed output
`highlight`	Emphasized text	Important terms in different colors
`badge`	Small colored labels	Statistical significance indicators

CSS File Structure

The `style.css` file contains:

1. **Global Styles**: Body text, basic formatting
2. **Section-Specific Styles**: Different rules for different parts of the report
3. **Component Styles**: Tables, boxes, badges, etc.
4. **Print Styles**: Special formatting for PDF generation
5. **Responsive Design**: How the report looks on different screen sizes

How to Modify the Styling

To change the report's appearance:

1. **Edit `style.css`**: Modify colors, fonts, or spacing
2. **Add new classes**: Create new styling rules for custom elements
3. **Update existing classes**: Change how current elements look
4. **Test changes**: Convert to PDF to see the results

Common CSS Properties Used

- `color`: Text color
- `background-color`: Background color
- `font-size`: Text size
- `margin`/`padding`: Spacing around elements
- `border`: Lines around elements
- `text-align`: How text is positioned

Why This System Works

- **Consistency**: All reports use the same styling rules
- **Maintainability**: Change the CSS once, affects all reports
- **Professional Appearance**: Clean, readable formatting
- **Print-Ready**: Optimized for PDF generation

---

1. Purpose

Before running statistical models in `multi_model.ipynb`, we created a **clean, standardized dataset** and produced a **pre/post outcome descriptive report** (`report_style.pdf`).
This phase ensures that:

- All raw NYTD data is harmonized and filtered for the Tennessee foster care population.
- Variables are consistently coded across years and waves.
- The output dataset can be directly reused in later modeling.

If you are replicating this, complete **every step in this section before moving on**.

---

2. Data Acquisition

Primary Source

Use the NDACAN-provided **codebook** (link) to understand variable meanings and valid values.

Cohort Structure

You need **four longitudinal cohorts**, each surveyed at:

- **Wave 1** – Age 17 (baseline, in care)
- **Wave 2** – Age 19 (post-transition)
- **Wave 3** – Age 21 (post-transition)

Cohort Years:

- Cohort 1: 2011–2015
- Cohort 2: 2014–2018
- Cohort 3: 2017–2021
- Cohort 4: 2020–2024

---

3. Data Preparation & Integration

You must replicate this **in order**:

Step 1 – Load and Filter

1. Import each NYTD file into Pandas with `dtype=str` to prevent unwanted type conversion.
2. Filter `StFIPS` to Tennessee only.
3. Keep only youth identified as foster care cases in Wave 1.
4. Drop any youth with no data in either wave 2 or wave 3 (they only need to have data in one).

Step 2 – Harmonize Variable Names

- Across cohorts and waves, standardize column names to match a single naming convention (e.g., `Homeless_w1`, `Homeless_w2`, `Homeless_w3`).
- Use the NDACAN codebook to ensure codes mean the same thing in all years.

Step 3 – Merge Waves

- Merge Wave 1, Wave 2, and Wave 3 datasets **by PersonalID and Cohort**.
- Save the merged dataset as `all_waves_merged.csv`.

Step 4 – Convert to Binary

- For all outcome variables final data should be:
  - “Yes” responses → `1`
  - “No” responses → `0`
  - Missing → blank
    - Use 'One-hot encoding' for any multiple answer columns
    - Drop or convert any variable that is not 1, 0 or blank
- Save the binary output as `all_binary_counts_merged.csv`.

Step 5 – Create Post-Transition Variables

- Combine Wave 2 and Wave 3 for each outcome:
  - If Wave 2 or Wave 3 is `1` → combined = `1`
  - If both are `0` → combined = `0`
  - If one is blank, use the other’s value
- Save as `combined_w1_w23_binary_counts.csv`.

---

4. Outcome Measurement

note: You may need to convert to int64 in order to calculate correctly For replication, use **exactly these domains and variables**:

Housing Stability

- Homeless_w1, Homeless_w23, PublicHousing_w1, PublicHousing_w23

Employment

- CurrFTE_w1, CurrFTE_w23, CurrPTE_w1, CurrPTE_w23, JobTrain_w1, JobTrain_w23

Financial Independence

- FoodStamps_w1, FoodStamps_w23, SocSecrty_w1, SocSecrty_w23

Education & Training

- GED_w1, GED_w23, HighSchool_w1, HighSchool_w23, HigherEd_w1, HigherEd_w23

Family & Children

- HasChild_w1, HasChild_w23, MarriedAtBirth_w1, MarriedAtBirth_w23

Adult Support

- AdultConnect_w1, AdultConnect_w23

Justice System Involvement

- Incarc_w1, Incarc_w23

For each:

1. Calculate Wave 1 % (`count of 1s / valid responses`).
2. Calculate Post-transition % (combined Waves 2 & 3).
3. Calculate Lifetime % (yes in any wave).

---

5. Analysis Approach

When replicating:

- Use the cleaned binary dataset.
- Calculate:
  - Prevalence change = (Post % – Wave 1 %)
  - Identify trends in increase/decrease.
- Flag notable patterns:
  - Challenges increasing (e.g., homelessness, food stamps).
  - Positive changes (e.g., higher education).
  - Stable protective factors (e.g., adult support).

---

6. Report Creation – `report_style.md`

Structure to Recreate

- **Executive Summary** – One paragraph per domain summarizing major trends.
- **Methods** – Include data source, sample, cohort structure, and coding decisions.
- **Domain Sections** – Each:
  - Defines variables in plain language.
  - Shows a table of Wave 1 %, Post-transition %, Lifetime %.
  - Provides a short interpretation.
- **Appendix** – All variable definitions and mapping to NYTD codes. **Note:** You should be able to tell cursor/vs code to replace the data in the report with the new data.

Design

- Use Markdown with embedded HTML for tables.
- Link to a `style.css` for consistent colors, fonts, and spacing.
- Insert navigation anchors and “Back to Table” links.

---

7. PDF Generation – `report_style.pdf`

To replicate:

1. Convert `report_style.md` → HTML with `style.css` applied.
2. Export HTML → PDF

- a.using Pandoc or wkhtmltopdf with:
  - `--enable-local-file-access`
  - CSS print styles for page breaks.
- b. using extension: Markdown PDF
  - Install
  - Right Click on .md file and click 'Markdown PDF: exptort(pdf)'

3. Check:
   - All section headings start on a new page.
   - All links and tables render correctly.

---

8. Role in Later Modeling

- The final `call_waves_merged.csv` from this phase is the **input** for adding AFCARS and services data in the model phase.
- Trends from `report_style.pdf` guide **which predictors to test** in logistic regression.

Replication Summary Workflow

Download NYTD raw files → Filter TN youth → Harmonize columns → Merge waves → Convert to binary → Combine W2 & W3 → Calculate outcome metrics → Write report_style.md → Export to report_style.pdf

---

Report Development Documentation – Model Phase

From multi_model.ipynb to all_visual_data_report.pdf

---

1. Purpose

To run **logistic regression** on the cleaned dataset merged with AFCARS and services data to identify predictors of post-transition outcomes.

---

2. Data Sources

- `wave_service_afcars_final.csv` – NYTD Outcomes + Services + AFCARS placement data.
- `predictor_stats_detailed_df.csv` – Predictor distributions.
- `outcome_stats_detailed_df.csv` – Outcome distributions.
- `significant_model_results.csv` – Final significant predictor list.

---

3. Replication Steps

Step 1 – Load & Prepare

- Load `wave_service_afcars_final.csv`.
- Drop rows with missing outcome values for the model being run.
- Convert categorical predictors to dummies (e.g., `pd.get_dummies` for PlacementType).
- Keep all binary flags from NYTD services and AFCARS.

Step 2 – Define Predictor Groups

Group predictors logically:

- Demographics (Sex, Race/Ethnicity)
- Placement history (PlacementType, PlacementInstability, MonthsInCare)
- Disabilities (DiagDis, SpecEdSv)
- Services received before 18

Step 3 – Model Execution

For each outcome:

- Use `statsmodels.Logit(y, X)` where y is the outcome and X is predictors + controls.
- Fit model with `.fit(disp=False)`.
- Extract:
  - `params` (coefficients)
  - `pvalues`
  - `conf_int()`

Step 4 – Filter for Significance

- Keep predictors with **$p < 0.05$**.
- Save to `significant_model_results.csv`.

---

4. Verification

Run `comprehensive_data_verification.md` to:

- Cross-check percentages in the report with CSVs.
- Confirm model counts and p-values.
- Correct any misleading interpretations (e.g., odds ratios).

---

5. Report Assembly

- Write `all_visual_data_report.md` with:
  - Introduction
  - Dashboard
  - Data and Methods
  - Significant results (tables and charts)
  - Appendices with all variables
- Link `style.css` and insert `<div class="page-break">` where needed.

---

6. PDF Export

- Convert Markdown → HTML → PDF using same settings as pre-model report.
- Verify:
  - Charts render correctly.
  - Tables not split across pages.
  - TOC links work.

Replication Summary Workflow

Load final dataset → Prepare predictors/outcomes → Run logistic regression → Filter significant results → Verify statistics → Build Markdown report → Convert to PDF