# Background

Our team was motivated to look at sleep disorders through research done through Stanford on devices that record sleeping patterns.

The project employed a comprehensive health dataset from the CDC, made up of seven (7) separate sections detailing information from demographics to financial status to health and wellness statistics. Credit for the original study goes to the National Center for Health Statistics (NHANES).

# Abstract

Our team was interested in discovering what lifestyle choices are correlated to sleep disorders. Leveraging available CDC data from 2013-2014, we were able to build a model that is a non-scientific predictor of a user’s potential to be diagnosed with a sleep disorder given certain health and lifestyle selections.

**How does your lifestyle affect your sleep?**

# Building the Model

The team leveraged functions from the scikit-learn and SciPy libraries in Python. In addition, for our feature selection we used SelectKBest, which conducted an ANOVA-like analysis on the data to aid in selecting the top fifteen (15) features. These features were then transformed and input into a sparse matrix. We used the logistic regression model from the scikit-learn linear model library. We fit our model to our sparse data matrix using the fifteen (15) features.

This CSV file was now ready to feed both the Tableau data visualizations as well as build the model.

# ETL Approach

We approached the dataset with one goal in mind: determine which of the chosen factors we selected correlated to a patient’s diagnosis of an unspecified sleep disorder.

Within those seven (7) files, there were 3,588 distinct measures. The team systematically reviewed each measure to reduce the list to 47 key indicators across six (6) categories:

* Demographics
* Diet & Exercise
* Drug and Alcohol Use
* Financial Status
* Sleep Habits
* Social-Emotional

Subsequently, those 47 key factors were used to find the appropriate subset that created a statistically viable model. Likewise, the data from those seven (7) disparate files were combined into a master CSV file.

# Analysis and Results

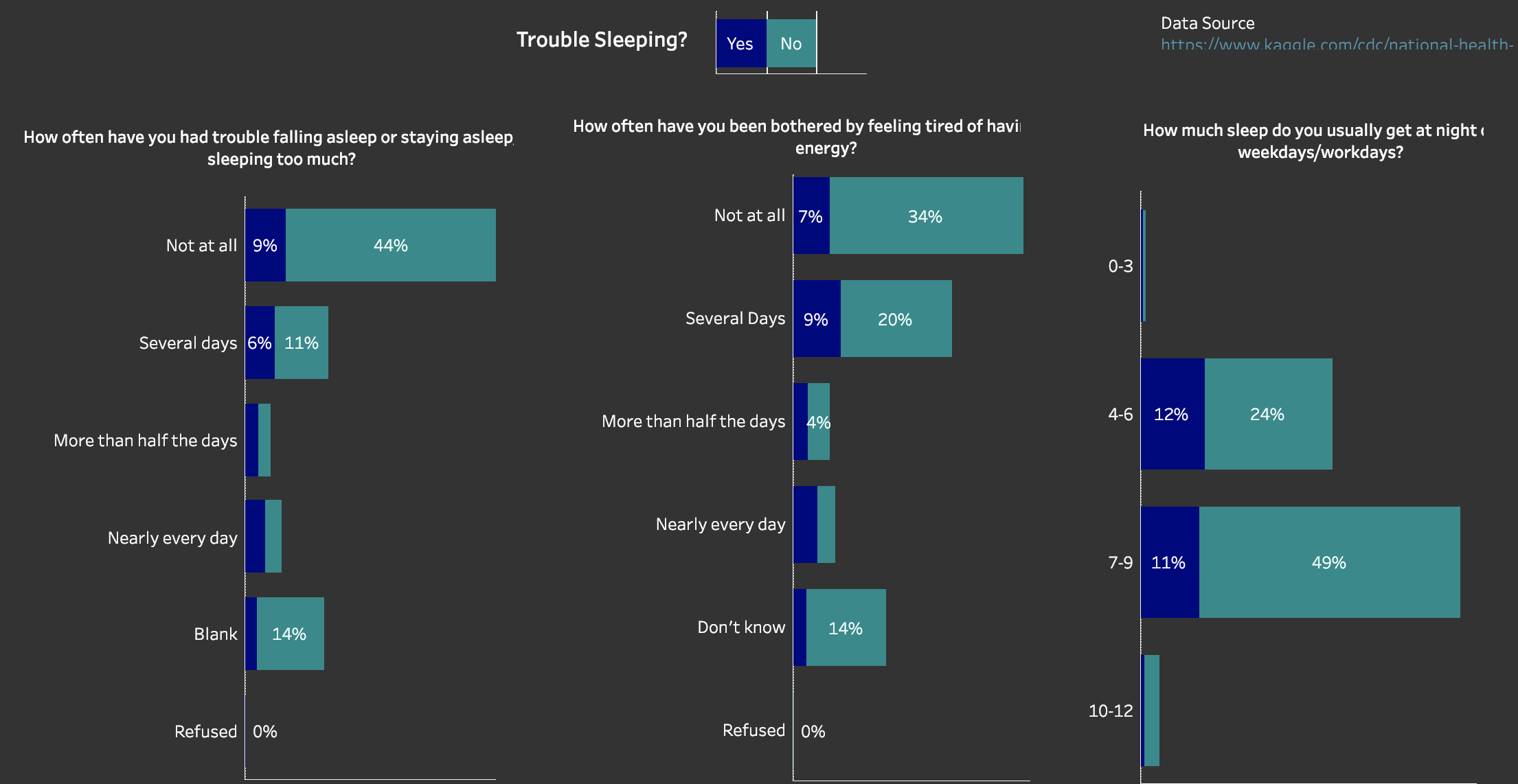
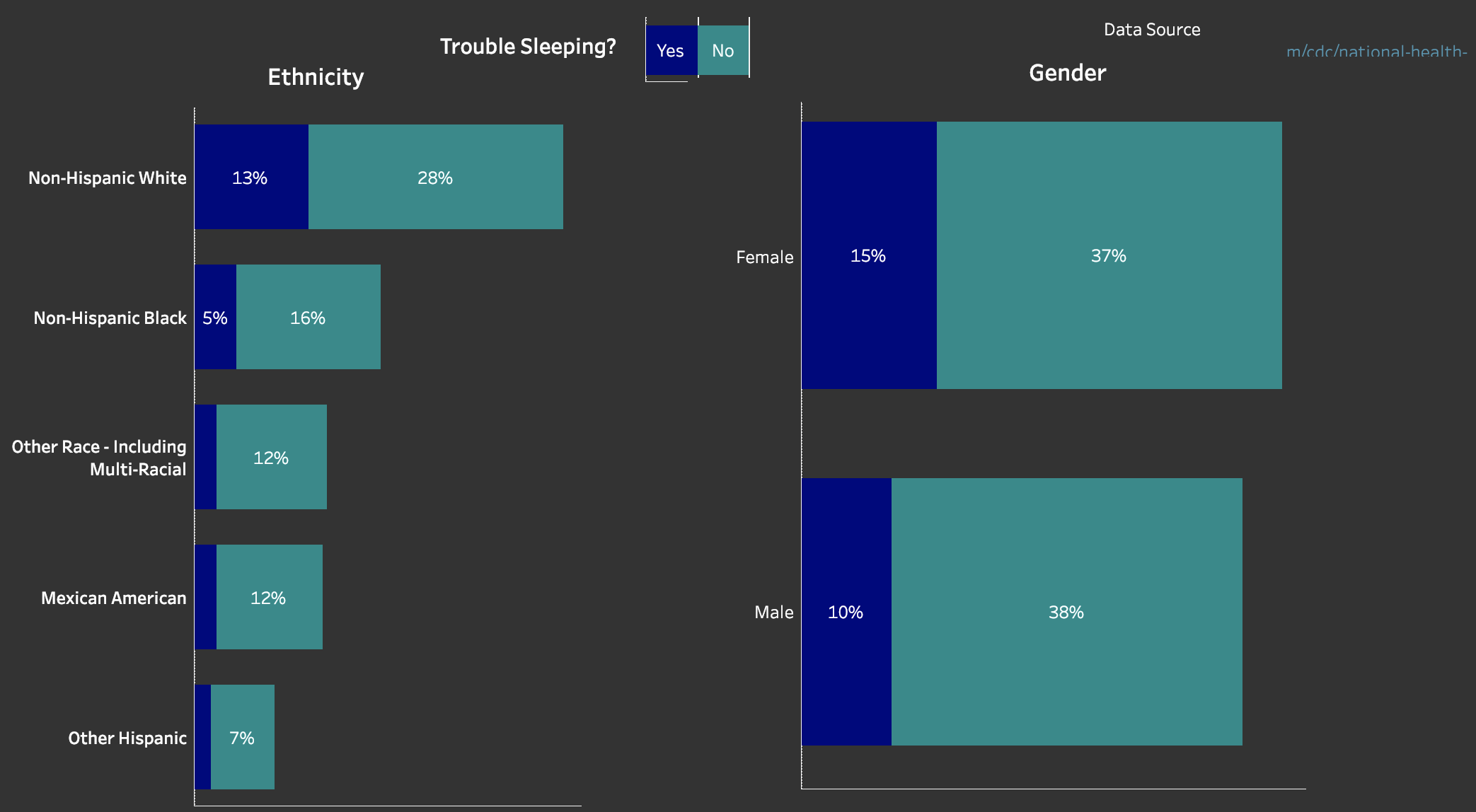
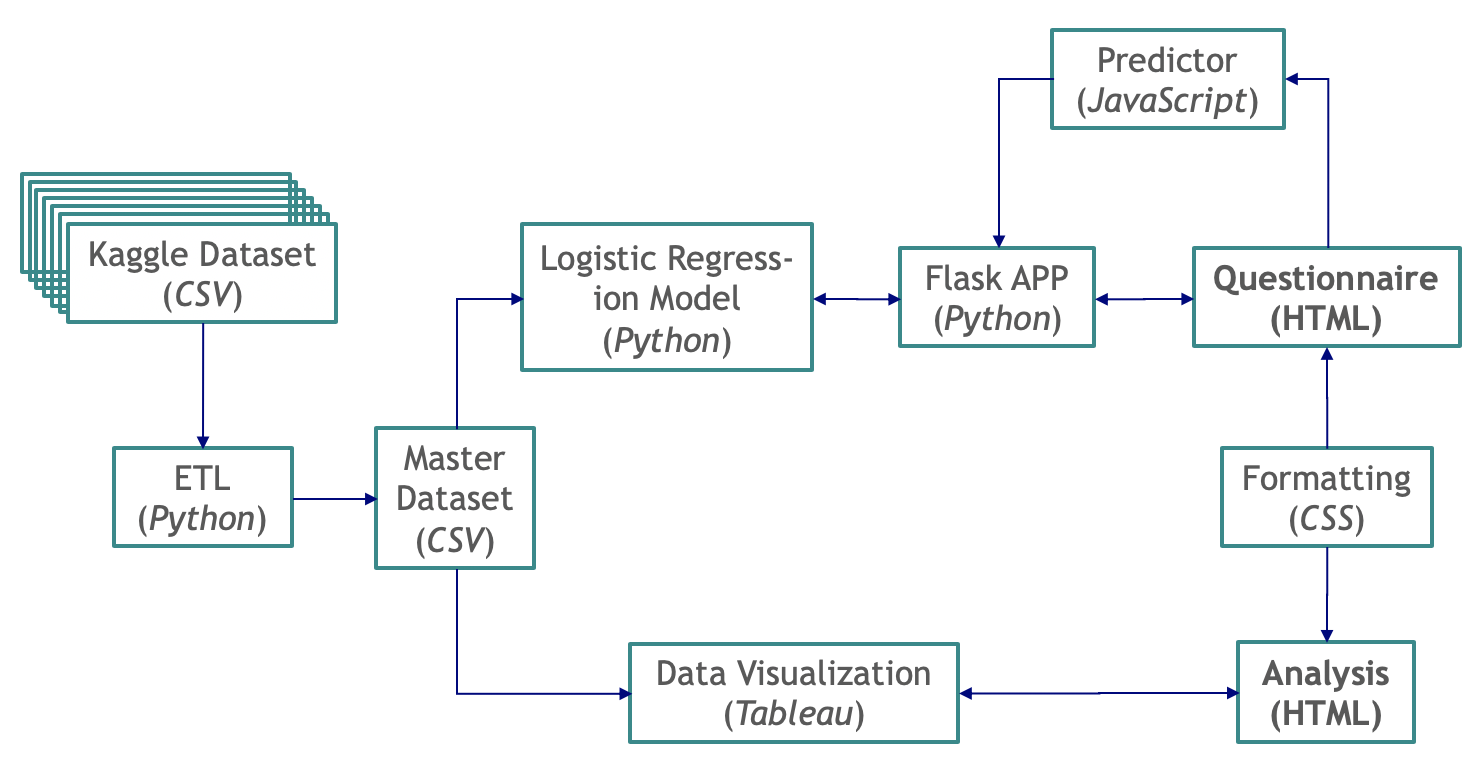
Using our model, we determined fifteen (15) critical factors used to determine a user’s propensity to be diagnosed with a sleep disorder. These fifteen (15) factors come from the areas of demographics, financial status, sleep habits, and social-emotional information.

As a result, a user can complete the questionnaire and receive a diagnosis prediction with a 77% precision value and a recall of 78%. An interesting aspect of our model is that the recall for a positive response is 25%, meaning that the model is not as good at predicting a positive diagnosis as it is a negative diagnosis (recall of 96%).

One additional feature of the model that is worth mentioning: it is scalable and could be easily modified for any other set of variables from the large dataset.

# Data Visualization

Examples of the data analysis using demographic data (top left) and sleep habits data (bottom right).



# Architecture Diagram