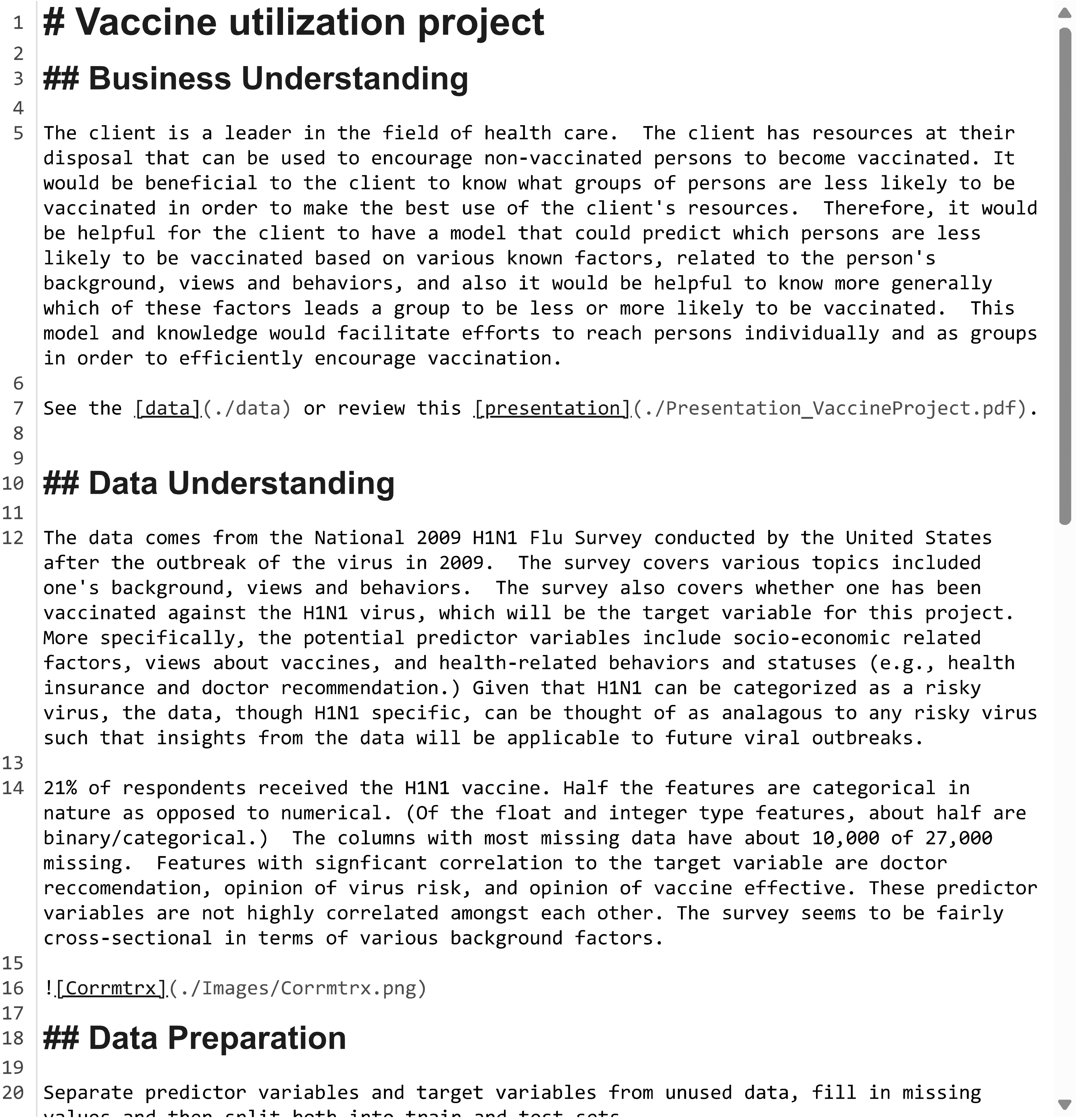

Vaccine utilization project

Business Understanding

The client is a leader in the field of health care. The client has resources at their disposal that can be used to encourage non-vaccinated persons to become vaccinated. It would be beneficial to the client to know what groups of persons are less likely to be vaccinated in order to make the best use of the client's resources. Therefore, it would be helpful for the client to have a model that could predict which persons are less likely to be vaccinated based on various known factors, related to the person's background, views and behaviors, and also it would be helpful to know more generally which of these factors leads a group to be less or more likely to be vaccinated. This model and knowledge would facilitate efforts to reach persons individually and as groups in order to efficiently encourage vaccination.

See the [data](./data) or review this [presentation](./Presentation_VaccineProject.pdf).

Data Understanding

The data comes from the National 2009 H1N1 Flu Survey conducted by the United States after the outbreak of the virus in 2009. The survey covers various topics included one's background, views and behaviors. The survey also covers whether one has been vaccinated against the H1N1 virus, which will be the target variable for this project. More specifically, the potential predictor variables include socio-economic related factors, views about vaccines, and health-related behaviors and statuses (e.g., health insurance and doctor recommendation.) Given that H1N1 can be categorized as a risky virus, the data, though H1N1 specific, can be thought of as analagous to any risky virus such that insights from the data will be applicable to future viral outbreaks.

21% of respondents received the H1N1 vaccine. Half the features are categorical in nature as opposed to numerical. (Of the float and integer type features, about half are binary/categorical.) The columns with most missing data have about 10,000 of 27,000 missing. Features with signficant correlation to the target variable are doctor reccomendation, opinion of virus risk, and opinion of vaccine effective. These predictor variables are not highly correlated amongst each other. The survey seems to be fairly cross-sectional in terms of various background factors.

![Corrmtrx](./Images/Corrmtrx.png)

Data Preparation

Separate predictor variables and target variables from unused data, fill in missing values and then split both into train and test sets.