Project Midterm Report

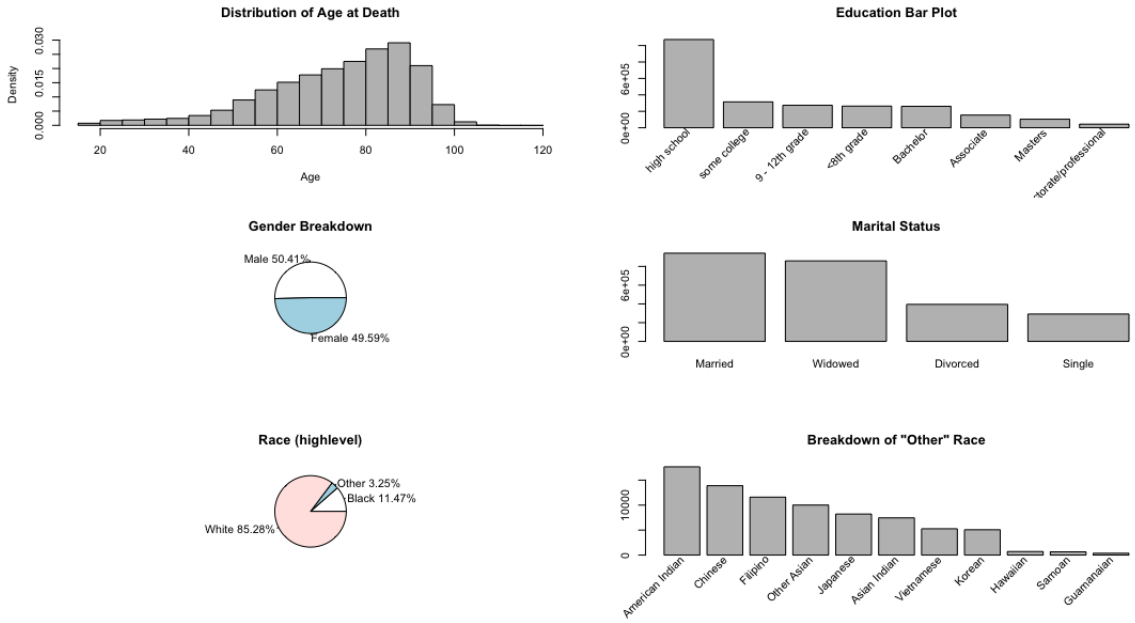
Kenneth Lipke (kel89) and Charlotte Wang (xw476)

With mortality data released by Centers for Disease Control and Prevention, we are trying to explore the underlying patterns and answer the question of "when someone will die". In this project, we aim to develop a prediction tool for insurance companies to calculate premiums and individuals to plan asset management ahead of time.

To start, lets take a high level look at our data. We will describe only the data for 2015. Before cleaning, this includes 2,718,198 observations, each corresponding to a different person. The dataset also contains 77 variables, however, many of them are useless—re-codes of the same quantity or values that are irrelevant for our analysis, such as method of disposal of the body (e.g., cremation of burial). We plan to focus on the following key variables: the age of death, education, sex, place of death, marital status, manner of death, race (with an extra breakout for Hispanic origin), and "entity condition," or the health conditions of the individual. In cleaning the data, we removed all the observations with ages below 18, as we want to focus on adults, as well as errors in age coding (which presented themselves as 999)–there were 34,764. Next, we removed the education levels that were missing or coded as "unknown," of which there were 192,874. We removed the observations with unknown places of death (1,290 of them), as well as the 20,104 with unknown marital status.

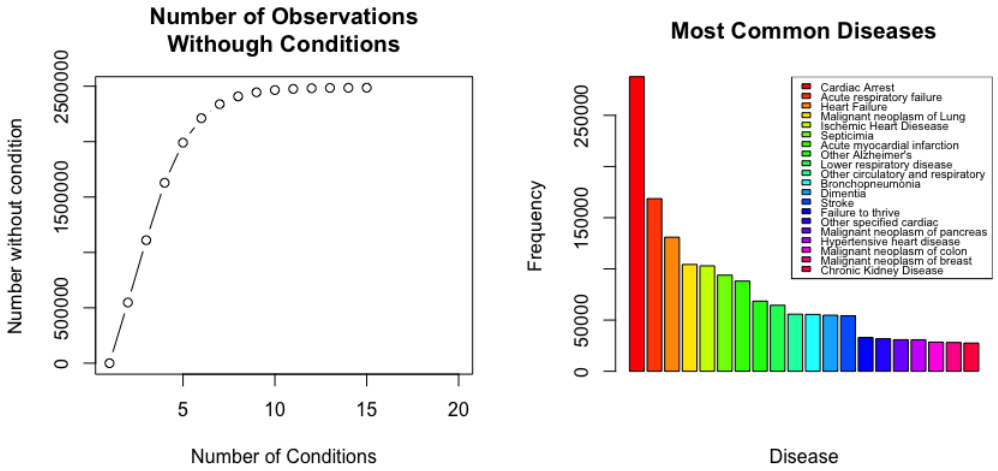
Before we dive into the medical data, lets get a feel for the distribution of the above discussed variables with a few histograms:

Figure 1: Distribution graphics



Look at the above, we see that there is significant variation in all the key variable. This is important as it will make it less likely that we over-fit the data. As there is large variation, it is unlikely that many (if any) of the observations are the same. This, in conjunction with the very large size of the data (the millions of observations) gives us huge latitude to apply complex models without having to worry too much about over fitting. On the other hand, this means we need to fit a complicated model to combat under-fitting. It is difficult to speculate on what exactly "complicated" means here, therefore, we will largely rely on model

Figure 2: Disease Graphics



cross validation to determine the optimal complexity.

Lets now turn our attention to the medical data. This is the most interesting dimension of the data, as we have information on not only the condition that is the direct cause of death, but also other conditions (up to 16 of them) that afflicted the individual. There are 16 variables for each observation that code these conditions. Most of the individuals in our data did not suffer from that many conditions, in fact, most only suffered from three of four. Looking at the left panel of the figure 2 we see the number of observations that have no condition entry for a given number of conditions. Based on this, we do not plan on using every column in our analysis. It would also serve to add complexity, we plan, therefore to only use the first four conditions.

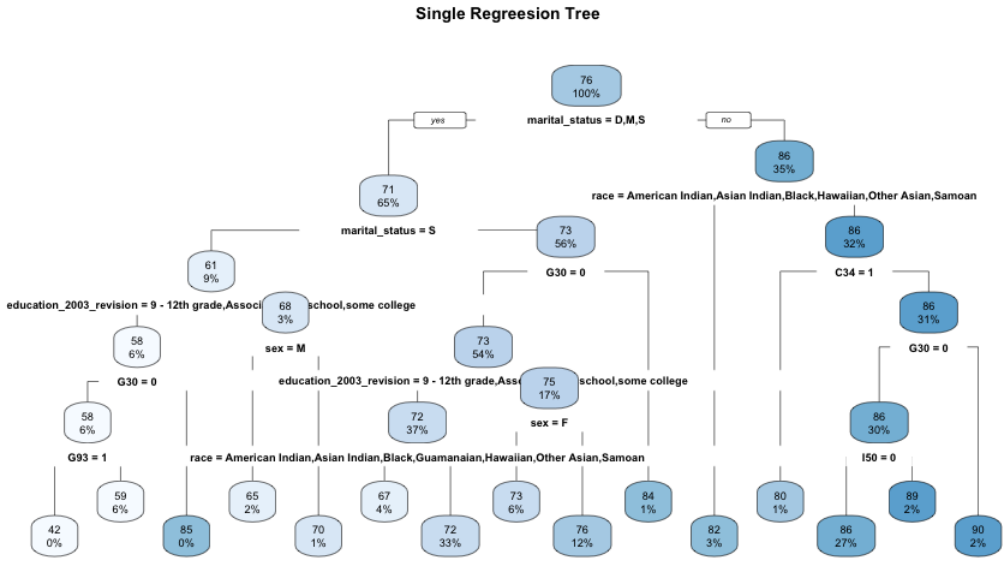
The condition data is coded according to the World Health Organization ICD 10 protocol. Each condition is coded as a letter, followed by two numbers (then possibly two more numbers, however we ignore those). In total, our data has 1,470 different conditions present. The right panel of figure 2 shows the top number of occurrences of the 20 most common conditions in our data set. We remove the observations corresponding to death from non-natural causes, as these, we feel, are too random to be predicted. Finally, we are left with a dataset of 2,486,372 observations.

We split the data into a training and test set, in order to test how well our current model does on out-of-sample data. In our analysis, we use 60% of the 2015 year dataset as our training set, this was used to build up our prediction algorithms. With the training set, we created multiple algorithms in order to compare their performances during the cross-validation phase. In the cross-validation phase, we use 20% of the 2015 year data as our validation set. We test multiple algorithms created based on the training set and will select the preferred prediction algorithm that has the best performance. Next, we will apply the chosen model on our test data set, which is 20% of the 2015 year data, to see how the model performs on new unseen data set.

We decided to start with a single tree. We think tree based methods will preform very well one this data as they succinctly capture interactions between variables. For analysis, the diseases are each coded as individual binary columns, therefore, if we were to attempt to handle the interactions in a linear regression setting, we would need to include a huge number of interaction terms, which would be cumbersome, and possibly lead to over fitting. Our first tree is shown in figure 3.

This tree has a mean squared error of 146.69 on the validation set. This, though, really has no meaning in the absence of other models to compare it to–so we will keep that number in mind for when we have

Figure 3: Preliminary Analysis: Single Regression Tree. The numbers in each node indicate the number of observations, the % indicates the percent of the data in that node, and the intensity of the color corresponds to the value of the prediction.



more models. However, we can still derive some insights from this single tree that will inform later analysis. Looking at the tree, the majority of the splits are made with the non-medical variables. The most important split, at the top is based on marital status, for example. The only medical conditions that did indeed make the cut are G30 (Alzheimer's), G93 (Other brain disorders associated with Alzheimer's), C34 (lung cancer), and I50 (heart conditions and hypertension). This is not what we would have expected; apriori we thought the medical conditions would by far be the best predictors of mortality, when it instead appears that other factors are more important.

Looking deeper into the fitting procedure for the tree also serves to inform our next step in analysis. Clearly, a singular tree is not ideal—we see that some of the terminal nodes contain no observations, so there is more work to be done in refining the fitting procedure. Beyond that, however, the way we fit this tree keeps track of the next best variables for each split. At the first split for instance, the top variables considered are shown in the below table.

Variable	Marital Status	Sex	G30	Race	I50
Potential Improvement	0.22744710	0.02637056	0.02018779	0.01889058	0.01069751

We see that marital status is no doubt the best choice for this first split, however, when we look at this table for different splits, the choice is less obvious. This leads us to believe that a random forest approach, which randomly limits the number of variables one can choose at each step, may dramatically increase prediction accuracy.

Moving forward, we first plan to apply more tree based models to our dataset. Mainly, we plan to focus on the random forest approach for the reasons described above. We will then refine the models using various cross validation methods, and variable selection methods to eliminate irrelevant predictors. Finally, once we have our models established, we will unleash them on our test set to get a final sense of prediction accuracy.