Data Cleaning

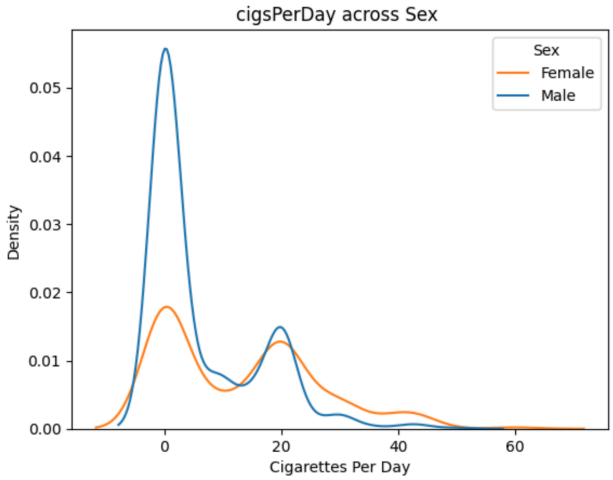
When cleaning the Coronary Heart Disease (CHD) data set, the first thing we did was look at the value counts for each variable to get a better idea of the data types and values. Second, we checked for missing variables by looking at the percentage of missing values in each column compared to the total rows in the dataset. Overall, the data did not have many missing values. Education, cigsPerDay, BPMeds, totChol, and BMI had between 0% and 2.5% missing values; however, glucose had almost 9% missing values. We used the KNNImputer to impute missing values. The imputer requires an 'n_neighbors', or k, argument. To get the optimal k, we fit a kNN model that clustered the ten year risk of developing heart diseases on five variables with no missing values. These five variables were age, sex, current smoker, stroke, and diabetes. The optimal number of neighbors to use was 6. The imputer worked pretty well, except when imputing values for the 'BPMeds' variable, or blood pressure medicine. BPMeds consists of binary 0 or 1 values, and after imputation, it contained values of 0.33 and 0.167. We thought about dropping missing values only in BPMeds before the imputation, but this didn't work because it messed up the dimensions of our dataframe, causing the KNN Imputer to create missing values where there were originally none. To get around this error, we decided to round values of 0.33 and 0.167 to the binary 0 (because they are closer to 0 than to 1). This is not a foolproof method, but it is the best solution we came up with given the data we were working with.

In the next stage of our data cleaning process, we addressed the issue of outliers that could potentially skew our model's performance, because they might have influenced it in a way that doesn't accurately reflect the trends within the data. Winsorization is a method of limiting extreme values in the data set to reduce the influence of outliers. This is done by replacing the extreme values with the closest values at the specified percentiles. For our data set, we selected different percentile thresholds for winsorization based on the variable's distribution and the extent of outliers observed. We applied winsorization to the variables for totChol, diaBP, heart rate, BMI, sysBP, and glucose levels. Specifically, we replaced the bottom and top 2% of data for totChol, diaBP, and heart rate with the values at the 2nd and 98th percentiles, respectively. For BMI, we adjusted the thresholds to the 3rd and 97th percentiles; for sysBP, to the 4th and 96th percentiles; and for glucose levels, to the 6th and 94th percentiles. These adjustments were made with the goal of removing all outliers while altering the original data as minimally as possible. After winsorization, we created boxplots again, and they showed a significant reduction in outliers, with the data distributions appearing more normalized and representative of the underlying trends.

Visualization

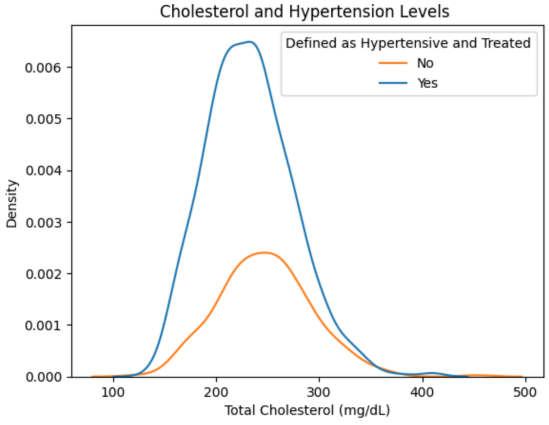
Potential variables of interest were examined through the creation of kernel density plots. While more extensive plots can be found in the repository, graphs of note are discussed below. Graph 1 correlates with the first predictive model examined, Cigarettes Per Day and Sex. Both genders show bimodal peaks, centered around 0 and 20 cigarettes per day. The increase in density of men compared to women at 0 per day is due to the greater number of men relative to women in the study. However, towards the upper end of the range of Cigarettes per day it does

appear than women have a higher density of more “extreme” smokers than men. It will be interesting when modeling to see if a higher number of cigarettes smoked in women is predicted of higher CHD risk.



Graph 1: Cigarettes per Day Across Sex

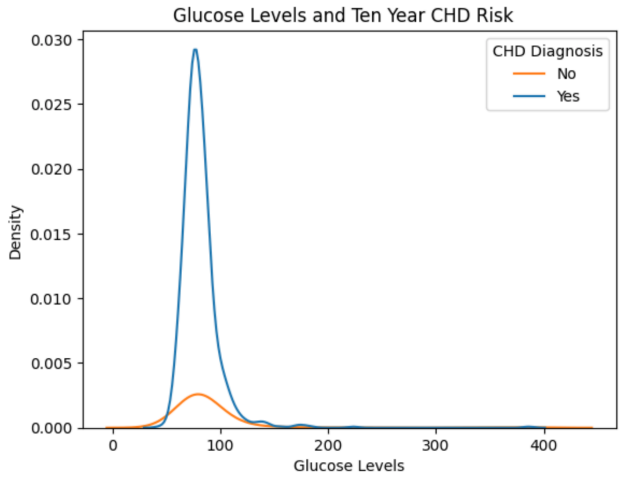
Our next plot, Graph 2, examines the relationship between Cholesterol (mg/dL) and Hypertension levels, where patients were defined as hypertensive (1) if they were treated. Those treated for hypertension show a unimodal peak of cholesterol levels around 220 mg/dL, while those not treated show a peak closer to 250 mg/dL. Thus based on this density plot, we wouldn’t expect the regression of these two variables to be extremely indicative of CHD risk. However, the cholesterol variable is total cholesterol, and does not distinguish between LDL and HDL the “bad” and “good” types of cholesterol. Thus, those not diagnosed with CHD may have higher levels of HDL, contributing to higher total cholesterol, but helping to reduce their risk of CHD development. Therefore, an additional variable for investigation could look at the relationship between HDL and CHD risk in comparison to just total cholesterol.



Graph 2: Cholesterol (mg/dL) and Hypertension

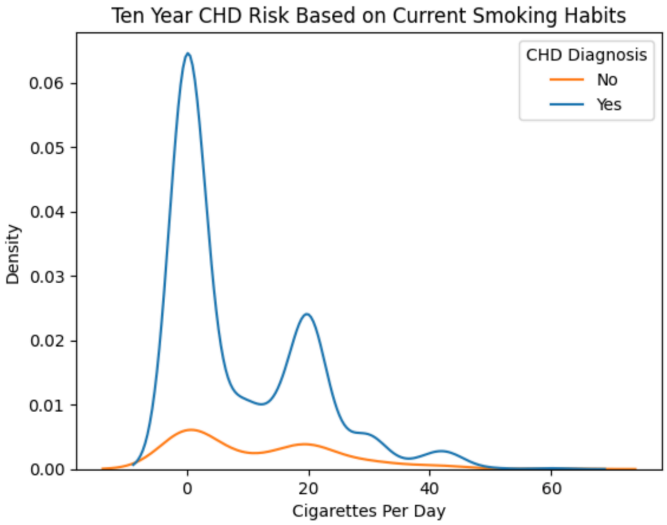
Graph 3 examines the relationship between Glucose levels (mg/dL) and Ten Year Risk of CHD. Neither those diagnosed with CHD or those free of the disease show significant variance in the

glucose levels, which would make sense given the body's need to maintain homeostasis. The graph for those diagnosed does show a more concentrated and higher density distribution than those who didn't develop CHD. Therefore, there may be some relationship between glucose levels and predicting CHD development. However, again glucose levels can relate to many lifestyle factors such as diet and exercise that may also confound CHD development.



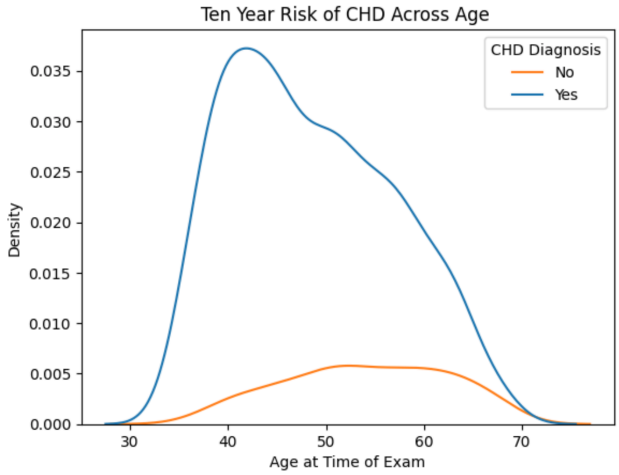
Graph 3: Glucose Levels (mg/dL) and Ten Year Risk.

Graph 4 examines the relationship between Cigarettes per Day and Ten Year Risk of developing CHD. The plot demonstrates that as the number of cigarettes per day increased, so did the density of those diagnosed with CHD. Thus in isolation, the number of cigarettes smoked per day would appear to have some predictive relationship with CHD development. However, it's also possible that people who smoke, especially upwards of several cigarettes a day, are likely to have other lifestyle factors that are conducive to CHD development. Given that the bimodal distribution for those who develop CHD also has its highest density at 0 cigarettes per day, since most participants do not smoke, the data clearly shows that other factors are involved in CHD risk.



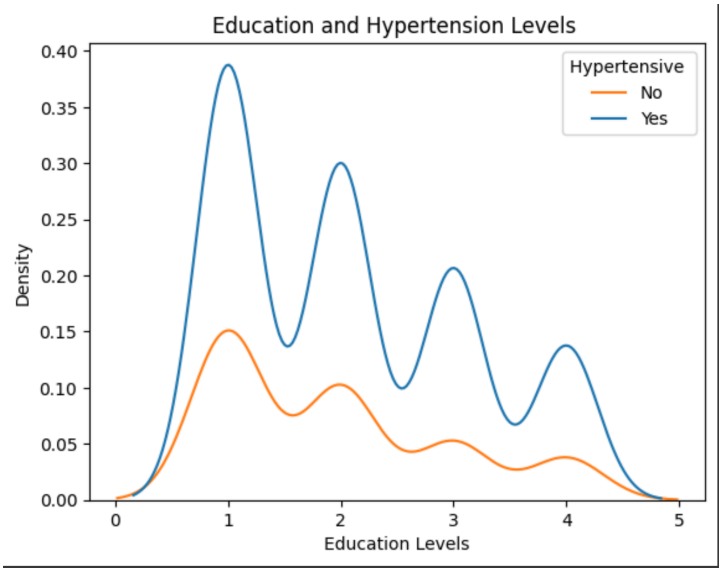
Graph 4: Cigarettes Smoked per Day and Ten-Year Risk of CHD

Our next plot, Graph 5, examines the data between age at time of exam and relative ten-year risk of CHD. People who were later diagnosed with CHD (categorized as 1 in the data set) had a higher risk of diagnosis at an earlier age. This may be because CHD is a disease that manifests early in several different physical symptoms and also has a genetic component. People with a history of CHD may be more likely to get an early diagnosis or take actions towards preventative care. The graph for 'Yes' to later diagnosis is also likely skewed to peak around 40 because that is the age when most people likely begin testing or treatment for CHD. The low R^2 found via the subsequent regression of these variables is likely explained by the confounding impact of lifestyle and genes on age.



Graph 5: CHD Risk and Age

Our final graph examines the relationship between education and hypertension levels. Education levels examined include some high school, a high school diploma, some college, and a college degree (1-4 respectively). Both the curve for those diagnosed with hypertension and for those without show that the risk of hypertension tends to decrease as educational level increases. Specifically for those with hypertension there is a steep drop off in diagnosis as education increases. Again the education variable is likely confounded with socio-economic status which influences factors such as diet, stress, substance use, exercise, and medical care. All of these are likely to have various impacts on hypertension and CHD risk as well.



Graph 6: Education Level and Hypertension

Results

The next step was creating our models using the cleaned training data and cleaned testing data. A linear regression was chosen based on its simplicity, interpretability, and predictive power. This tool is helpful for understanding a clear relationship between variables and making accurate predictions. The first model we tried was CigsPerDay and Sex, which yielded a R^2 result of 0.106 and RMSE of 11.321. This R^2 is okay, but small, and suggests the model's predictive power is pretty limited when regressing CigsPerDay by Sex. RMSE, or Root Mean Squared Error, is a measure of the difference between the predicted values of the model versus the observed ones. For the sake of this analysis, we are going to focus on the R^2 value. The second model we tried was Total Cholesterol and Hypertension, which yielded a R^2 result of 0.027 and RMSE of 41.045. This R^2 is very low and suggests the model's predictive power is very limited when regressing Total Cholesterol by Hypertension. The third model we tried was Glucose and TenYearRisk, which yielded a R^2 result of -0.058 and RMSE of 11.71. This R^2 is negative, indicating the model does not fit the data at all. The fourth model we tried was CigsPerDay and TenYearRisk, which yielded a R^2 result of -0.009 and RMSE of 11.445. Again, the R^2 value indicates the model does not fit the data. The fifth model we tried was Age and TenYearRisk, which yielded a R^2 result of 0.059 and RMSE of 0.349. This R^2 is very low and suggests the model's predictive power is very limited when regressing Age by TenYearRisk. The final model we tried was Education and Hypertension, which yielded a R^2 result of 0.012 and RMSE of 0.464. Again, this is extremely low and suggests the model's predictive power is very limited when regressing Education by Hypertension.

The most predictive model was found to be the first one tested, which regressed CigsPerDay by Sex and yielded a R^2 result of 0.106. Certain models were not able to be run due to the variables selected. For example, HeartRate, sysBP, and BMI outputted an error message stating “X has 50 features, but LinearRegression is expecting 70 features as input”. These variables were therefore excluded from the model.