EXECUTIVE SUMMARY

The main objective of this case study is to create a screening tool which allows us to easily and efficiently identify those individuals who are at a higher risk of suffering from CKD. Those with a high enough probability of having CDK should be screened for it. The data used in this analysis was gathered by the National Center for Health Statistics of the Centers for Disease Control and Prevention. The data contains information on 33 different variables from 8,819 adults of 20 years of age or older, and was collected between 1999 and 2002. Of the data, 6,000 observations contain information pertaining to a CKD diagnosis, and 2,819 observations do not contain CKD diagnosis information.

Chronic Kidney Disease (CKD) is a progressive condition that results in significant morbidity and mortality due to a gradual loss in kidney function. There are two main causes of CKD: diabetes and high blood pressure. These two causes are responsible for up to two-thirds of CKD cases. Risk factors associated with CKD include diabetes, hypertension, cardiovascular disease, family history of kidney disease, age, and race (Black and Hispanic Americans are more likely to have CKD). Additionally, “almost half of individuals with CKD also have diabetes and self-reported CVD”[1] Similarly, “Approximately 1 in 3 adults with diabetes has CKD.[4]” Anemia is associated with CKD, and increases as CKD progresses[3]. Additionally, women are more likely to have CKD than men.In preparation of this case study, we contacted Dr. Suma Raju to ask her about the risk factors of CKD: "Early phases of kidney disease usually do not show any symptoms and anyone who has the risk symptoms like Diabetes and High Blood Pressure should undergo a screening test as these are the main markers of kidney disease.

Some more additional symptoms include any heart disease or family history of kidney disease.In case of a patient having these symptoms, we recommend the patient to undergo a routine urine test, blood test and if required an ultrasound of the abdomen. In a urine test, we need to check for protein, blood and serum creatinine which signifies that there is some damage in the kidney which is causing any leakage. In ultrasound, we check for kidney size and if there are any blockages.[5]``

We begin our report by examining the distribution, descriptive statistics, and missingness of our data. Both our initial 6000 row “training” data, and our 2,819 “prediction” data appear to be similar, indicating we can create and use a model from these datasets without worry of too much distortion. We impute our missing data using MICE, and consider that some data we impute may be missing not at random (MNAR). Using our “training” data, we run a logistic regression using all variables. To counteract overfitting, we use the feature selection technique stepAIC to identify the model with the lowest AIC. We likewise examine the medical literature.

We observe that our backwards selection model has the lowest AIC, and we examine it using cross validation. Using this model, we predict the CKD probabilities for our data with CKD information. We examine where to set the threshold for CKD classification, beginning at .077 (the mean). In exploring how this threshold relates to our profit margins, we discover that a threshold of .077 gives us an expected profit of $292,200. We therefore set our threshold for our predicted probabilities at this number. To develop our **screening** tool, we utilize a OLS using the variables selected from the process above, and eliminate several variables that showed a negative logistic regression coefficient (associated with lower CKD risk). Using the coefficients from our OLS, we develop weights for each question. We then test our screening tool on our data with CKD to observe at what threshold to set the score for testing.

[1] https://www.niddk.nih.gov/health-information/health-statistics/kidney-disease#:~:text=The%20overall%20prevalence%20of%20CKD%20increased%20from%2012%20percent%20to,to%206.0%20percent%2C%20since%201988.

[3] https://jasn.asnjournals.org/content/23/10/1631

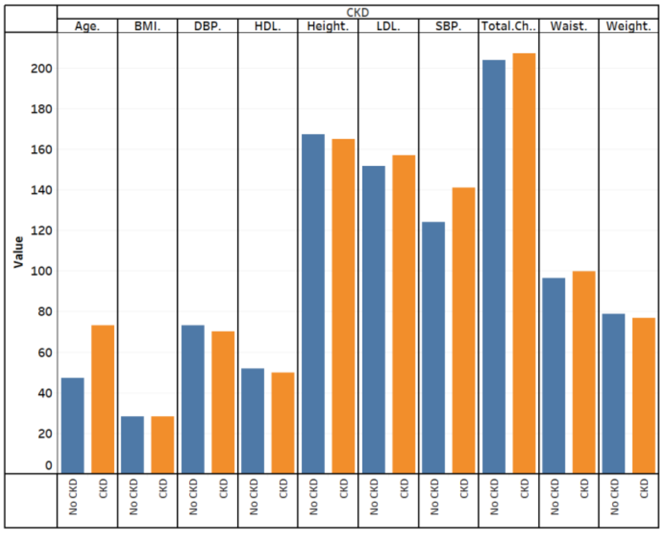
[4] https://www.cdc.gov/kidneydisease/prevention-risk/make-the-connection.html#:~:text=CKD%20is%20common%20in%20people,diabetes%20can%20cause%20kidney%20disease.

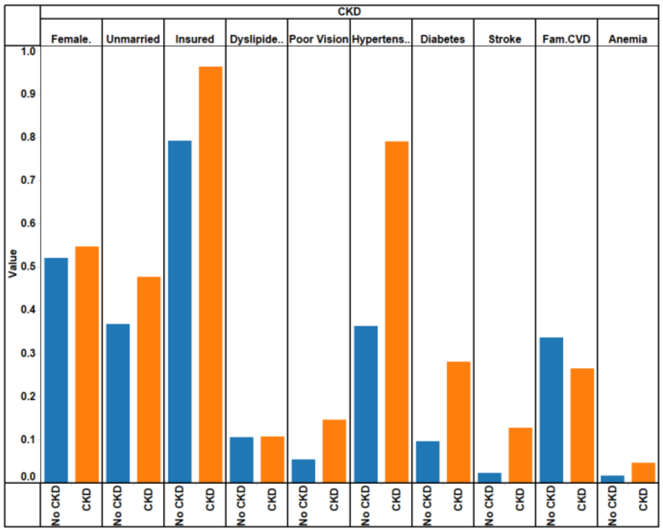
[5] https://youtu.be/zHS9gnlr5X4

DATA EXPLORATION

DATA DISTRIBUTIONS

First, we must ensure that the data we are basing our predictive model on is similar enough to the data we are predicting. We therefore compare the distributions of each variable in our respective predictive and predicted data sets, with missing values removed. We find that the distributions are roughly similar across all variables. Next, we compare the variable means of those with CKD and those without, in order to scan for any significant differences in those subjects with CKD. We observe that Unmarried, Insured, Poor vision, Hypertension, Diabetes, Age, Stroke, and Anemia all appear to have differences in means, hinting at variables important for CKD distinction.





EXPLORATION

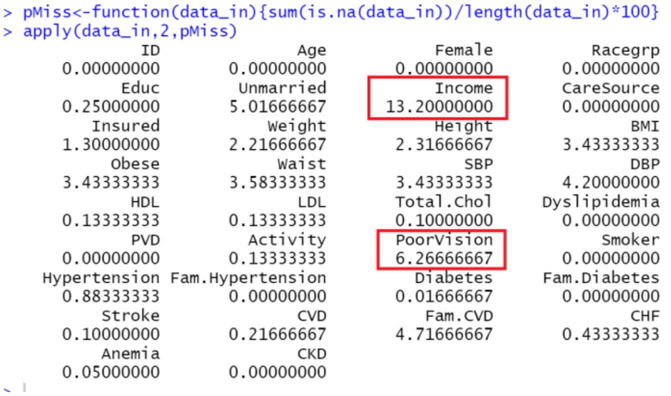
Approximately 7.73% of all patients have CKD. Not that this is slightly lower than the national prevalence of U.S. CKD between 1999-2004, which was approximately 15.2%[2]. Therefore, it is entirely possible that this dataset sample is not representative of the national population. Perhaps random selection was not used in acquiring participants.

MISSING DATA

Missing data is categorized into three categories: missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR). Each category has different recommendations and implications for conducting imputation. We can conduct imputation on MCAR without leading to bias, and we can also conduct imputation on MAR, though this may lead to some bias. However, it is not recommended to impute MNAR values, as this will lead to bias in the data set.

TRAINING DATA

We examine the missing data in our training data. Note that Income and PoorVision have the largest percentage of missing data.

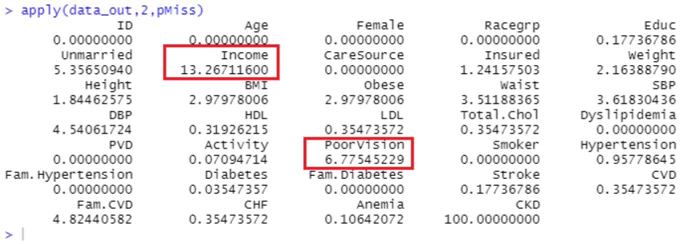


```
> pMiss<-function(data_in){sum(is.na(data_in))/length(data_in)*100}
> apply(data_in,2,pMiss)
            ID              Age           Female          Racegrp 
    0.00000000       0.00000000       0.00000000       0.00000000 
          Educ        Unmarried           Income       CareSource 
    0.25000000       5.01666667      13.20000000       0.00000000 
       Insured           Weight           Height              BMI 
    1.30000000       2.21666667       2.31666667       3.43333333 
         Obese            Waist              SBP              DBP 
    3.43333333       3.58333333       3.43333333       4.20000000 
           HDL              LDL       Total.Chol     Dyslipidemia 
    0.13333333       0.13333333       0.10000000       0.00000000 
           PVD         Activity       PoorVision           Smoker 
    0.00000000       0.13333333       6.26666667       0.00000000 
  Hypertension Fam.Hypertension         Diabetes    Fam.Diabetes 
    0.88333333       0.00000000       0.01666667       0.00000000 
        Stroke              CVD          Fam.CVD              CHF 
    0.10000000       0.21666667       4.71666667       0.43333333 
        Anemia              CKD 
    0.05000000       0.00000000 
```

Within this dataset, 1,864 rows contain at least one instance of missing data, while 4,136 rows contain no missing data. This is a substantial percent of our data. If we simply throw away these rows, we might be systematically changing the sample we base our predictive model on. In order to ensure our sample is representative of the original sample, we should strive to keep the rows with missing data by using imputation.

PREDICTING DATA

Note that in our predicting data Income and PoorVision also have the largest percentage of missing data.



```
> apply(data_out,2,pMiss)
              ID              Age           Female          Racegrp             Educ
      0.00000000       0.00000000       0.00000000       0.00000000       0.17736786
       Unmarried           Income       CareSource          Insured           Weight
      5.35650940      13.26711600       0.00000000       1.24157503       2.16388790
          Height              BMI            Obese            Waist              SBP
      1.84462575       2.97978006       2.97978006       3.51188365       3.61830436
             DBP              HDL              LDL       Total.Chol     Dyslipidemia
      4.54061724       0.31926215       0.35473572       0.35473572       0.00000000
             PVD         Activity       PoorVision           Smoker     Hypertension
      0.00000000       0.07094714       6.77545229       0.00000000       0.95778645
Fam.Hypertension         Diabetes     Fam.Diabetes           Stroke              CVD
      0.00000000       0.03547357       0.00000000       0.17736786       0.35473572
         Fam.CVD              CHF           Anemia              CKD
      4.82440582       0.35473572       0.10642072     100.00000000
> |
```

IMPUTATION

Univariate calculates missing values by using only that particular column and not looking at any other column. An example of a univariate imputation would be to insert the medium or mean of the same column into all missing data cells. Though this technique would allow the data's distribution to remain the same, it would affect the data's variance.

Multivariate imputation works by factoring in other variables to make better predictions about the potential missing value. The Multivariate Imputation by Chained Equation, or MICE, in R studio will be utilized in analysis of this dataset. MICE assumes the data is missing at random. For categorical values, we can predict missing values using logistic regression, because logistic regression can be used for categorical data. For continuous values, we can predict their missing values using predictive mean matching (pmm).

THRESHOLDS

Safe max threshold for missing data is 5% of the total data set. If missing data for a certain feature is more than 5%, we may need to drop the variable from our analysis. As we have 6000 observations, 5% of this is 300. We see that both Income (missing 13%) and PoorVision (missing 6.3%) have percent missing thresholds above 5% in both our training and testing data. Unmarried is just over 5%. It is likely that missing variables in Income are MNAR, as individuals may be adverse to disclosing financial information. It is therefore possible that imputing these variables will lead to bias in the dataset. Prior literature has established an association between these variables and CKD. However, upon running an initial regression, we do see that both Income and PoorVision have non significant relations with CKD. Therefore we drop these variables from our future screening tool.

DATA DISTRIBUTION

In order to know how "good" our data imputation was, we need to examine the data's distribution both before and after imputation. Ideally, the distributions will look similar. We examine if our imputed data distributions are similar to our non-imputed data distributions. We find that the distributions are comparable across nearly all variables.

LOGISTIC REGRESSION

Our model's intercept is negative, meaning that when all coefficients are zero, the estimated probability of CKD is much less than fifty. Given that CKD sample prevalence is around 7%, this makes intuitive sense. Upon running our regression (See *Figure 1*), we observe several variables have significant positive coefficients, and are therefore associated with an increase in CKD risk: Age, Female, Racegrpwhite, Unmarried, Hypertension, Anemia, etc. These coefficients are the log odds of having CKD, for instance, the coefficient for age can be interpreted "for every one year increase in age, the log odds of having CKD increases by .094". To choose our variables, we will rely on a (a) prior medical research (b) stepAIC and (c) cross validation, also called out-of-sample testing. According to prior research, risk factors associated with CKD include diabetes, hypertension, cardiovascular disease, family history of kidney disease, age, anemia, and race. We will ensure that at minimum, our model includes variables related to this.

AKAIKE INFORMATION CRITERIA

AIC estimates prediction error and the "relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model"[4]. A low AIC indicates a comparatively better model. Feature selection techniques are used to counteract overfitting and to "determine the smallest set of features that are needed to predict the response variable with high accuracy"[5]. StepAIC is a feature selection technique used in model selection. StepAIC works to find the model with the lowest AIC value. In this analysis, we utilize forward,

[2] https://nccd.cdc.gov/ckd/detail.aspx?Qnum=Q633

backwards and forward/backwards stepAIC. Forward selection starts from a null model and adds variables sequentially, starting with the highest correlation to the predicted variable. Backward selection begins with all predictors and subsequently removes those with the highest p-value until a give threshold is reached (in our case at P=.15). Note that AIC is a measure of relative model quality- it selects the comparatively best model. To determine the relationship between CKD and our variables, we need to conduct a hypothesis test.

MODEL SELECTION

In running our stepAICs, we observe that the backwards selection results in the lowest AIC value of all three attempted models. We also note that the backwards selection model has the lowest residual deviance, therefore it is comparatively better.

Method	**Original Logistic Reg**	**Forward**	**Backward**	**Both**
AIC	2233.6	2213.1	2211.8	2213.1
Residual Deviance	2161.6	2183.1	2171.8	2183.1

We therefore examine the model's coefficients (Figure 2). We observe both the p-values and coefficient magnitudes. Given that the variable Unmarried does not relate directly to CKD in the medical literature (though there may be some latent, indirect relation with CKD), we eliminate them it our model. Though Racegrupwhite also has a p-value of less than 0.1, we retain the race variable based on substantial medical literature affirming its relation to CKD. Though BMI had a low p-value, we retain it since weight has high significance to CKD. We therefore eliminate Weight from our model to avoid linearly dependent regression variables. Accordingly, we are left with the following model for consideration:

CKD ~ Age + Female + Racegrp + BMI + Waist + HDL + PVD + Activity + Hypertension + Fam.Hypertension + Diabetes + CVD + Fam.CVD + CHF + Anemia

Upon running this new model, we get an AIC of 2222.7 with a residual deviance of 2184.7. Note that this AIC is a slight increase from the model run using all variables, as well as the model run using AIC backwards selection. This means the model fits slightly less well, as expected since we are eliminating several variables to increase ease of screening. We run several codes where we eliminate variables with the highest p value or a smaller magnitude. We compared the AIC of these varying models, however they were not lower then the above model.

CROSS VALIDATION TESTING

To further examine our variable selection, we randomly divide our predictive dataset into two separate datasets- training and testing- in a randomized 75:25 split. We re-run our Backwards Selection stepAIC on the "training" data split. We find that the variables selected are the same as our original backwards stepAIC .We re-randomize our dataset split. Our second iteration has added the variable "insured" and taken away "unmarried". We re-randomize and find this third iteration has added 'Insured', but is the same as our original selection. Despite this flux, we can be sure we are choosing variables that all these models had in commonality. As backwards selection picks off the highest p values, it makes sense that there is a stable "core" of variables with the highest significance. We therefore draw our model variables from them.

CONFIDENCE INTERVALS

Our Null Hypothesis is that there is no relation between a variable and CKD, meaning the coefficient is zero. In order to reject the null hypothesis, we examine our confidence intervals. If 0 is within our confidence interval, we fail to reject our null hypothesis. We find that "female", "BMI", and "waist", have zero within their confidence interval. We therefore cannot be sure that the relation between the variable and CKD isn't null within this dataset. Note that medical research has established a clear relationship between female and CKD, however, our dataset differs from this.

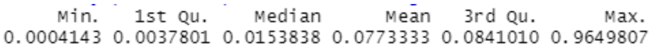
We find a p-value incredibly close to zero (2.663721e-217), indicating that, as expected, we can reject the null hypothesis. Our model is indeed related to CKD. We calculate the long likelihood of our model at 2186.29.

[4] https://en.wikipedia.org/wiki/Akaike_information_criterion
[5] https://ashutoshtripathi.com/2019/06/07/feature-selection-techniques-in-regression-model/

IN-SAMPLE PREDICTION

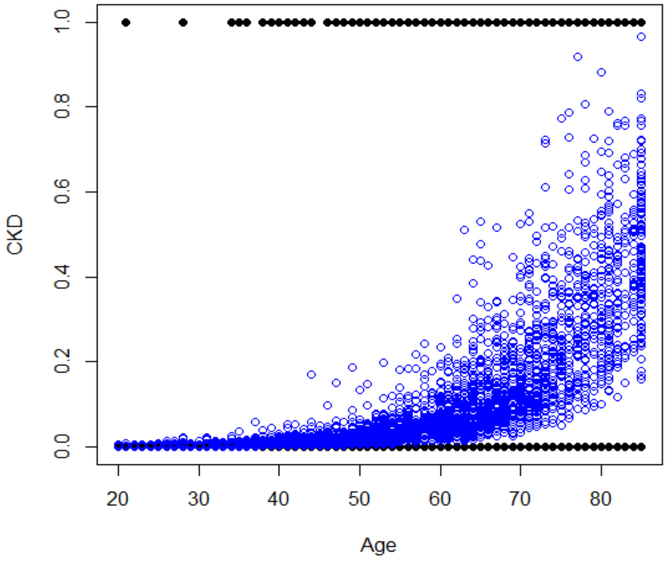
Using the command predict, we estimate the probability for CKD using our data which includes CKD data. This runs the following equation, utilizing all coefficients:

$$(\text{Probability that } Y_i = 1) = \widehat{p_i} = \frac{1}{1+\exp(-(\widehat{\beta_0}+\widehat{\beta_1}X_i))}$$



```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
0.0004143 0.0037801 0.0153838 0.0773333 0.0841010 0.9649807 
```

We observe that the median is .015, meaning half the patients have less than a 1.5% chance of getting CKD. We plot these probabilities according to subject age, and can see a general trend towards an increase in CKD risk, particularly above 60 years of age. As the CDC states that CKD is more common in those 65 or older, this corresponds with our visualization, as we see some subject probabilities of CKD begin to become significantly larger starting at around 65 years of age. We observe a great deal of variance and scatter within this graph, since CKD is dependent on multiple factors.



CLASSIFICATION

In order to classify someone as CKD or not, we need to establish the threshold probability cut off in which we classify them as CKD. This will be greatly influenced by the amount of money we receive for a true positive (1300) and a false positive(-100), as we are looking for the highest profit. Because this ratio is so high, we are able to allow for many false positives, so long as we increase our true positives. We begin by setting the threshold to the mean of the probabilities (.077) and calculate our expected profit. We examine thresholds above and below .077, and find that .077 gives us the greatest profit at $292,200. Therefore, we set our threshold at this point.

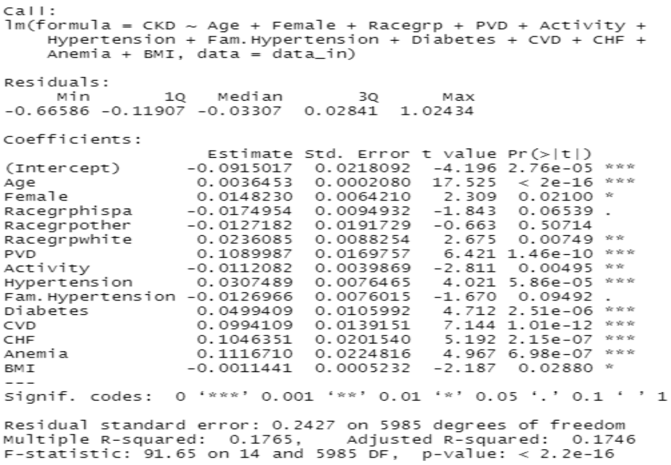
Threshold	.077	.0813
Accuracy	.78	.80
True P Rate	.85	.84
False P Rate	.23	.20
Profit	$292,200	$54,100

SCREENING TOOL

SCREENING TOOL WEIGHTS

Though logistic regression is ideally suited for determining the probability of a binary variable, its coefficient interpretability is low. Each unit change in a variable is associated with a log odds change, and therefore this changes depending on the location in the S curve. In order to assign weights to responses on our screening tool, we therefore utilize a simple linear regression. We employ this method with the full awareness that we are violating two linear regression assumptions, (1) that variables are normally distributed, as in this data they are not, and (2) that our dependent variable is continuous. Therefore, we would expect that our coefficients are much smaller than we would expect in a typical linear regression, since our DV is between 0 and 1.

We also expect that R squared will not have a meaningful interpretation, "since the regression line can never fit the data perfectly if the dependent variable is binary"[6]. Despite this, there is growing research supporting the usefulness of linear regression in approximating the value of a binary variable[7]. Therefore, we run an OLS on our chosen variables using the first 6000 observations (the data with CKD information). However, we eliminate the variables Waist, HDL, Family Hypertension, Fam CVD. Some of these variables were negatively related to CKD, and we tried to make the tool as short as possible. We observe the following model:

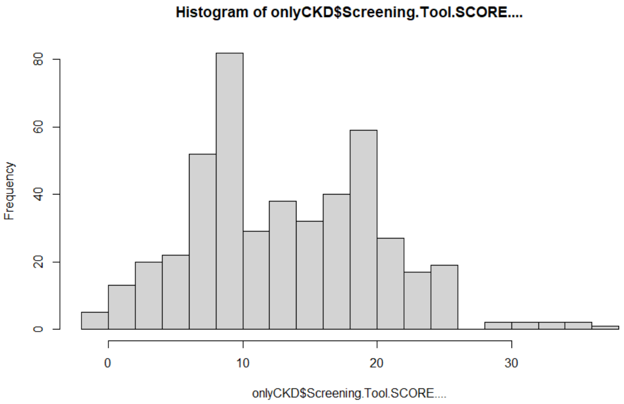


```
Call:
lm(formula = CKD ~ Age + Female + Racegrp + PVD + Activity + 
    Hypertension + Fam.Hypertension + Diabetes + CVD + CHF + 
    Anemia + BMI, data = data_in)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.66586 -0.11907 -0.03307  0.02841  1.02434 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -0.0915017  0.0218092  -4.196 2.76e-05 ***
Age               0.0036453  0.0002080  17.525  < 2e-16 ***
Female            0.0148230  0.0064210   2.309  0.02100 *  
Racegrphispa     -0.0174954  0.0094932  -1.843  0.06539 .  
Racegrpother     -0.0127182  0.0191729  -0.663  0.50714    
Racegrpwhite      0.0236085  0.0088254   2.675  0.00749 ** 
PVD               0.1089987  0.0169757   6.421 1.46e-10 ***
Activity         -0.0112082  0.0039869  -2.811  0.00495 ** 
Hypertension      0.0307489  0.0076465   4.021 5.86e-05 ***
Fam.Hypertension -0.0126966  0.0076015  -1.670  0.09492 .  
Diabetes          0.0499409  0.0105992   4.712 2.51e-06 ***
CVD               0.0994109  0.0139151   7.144 1.01e-12 ***
CHF               0.1046351  0.0201540   5.192 2.15e-07 ***
Anemia            0.1116710  0.0224816   4.967 6.98e-07 ***
BMI              -0.0011441  0.0005232  -2.187  0.02880 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2427 on 5985 degrees of freedom
Multiple R-squared:  0.1765,	Adjusted R-squared:  0.1746 
F-statistic: 91.65 on 14 and 5985 DF,  p-value: < 2.2e-16
```

individual who is both diabetic and 65 years or older should receive a score that indicates high risk. If we were to assign points based on the coefficient (.04) for each increase in age, we would end up distorting the model. Consider that the risk of CKD for age is particularly not linear in comparison with our other variables- we observed a curved (presumably) exponent-based line when we plotted our predicted probabilities on the previous page. Since those over 65 years of age carry a comparable prevalence of CKD as those with Diabetes (38% vs 33%) in the medical literature, we will group age into two variables: 0 if under 65, and 1 if over 65, and will assign the same amount of points as given to Diabetes.

all CKD cases. We plot a histogram distribution, and find there is only the slightest hint of being right skewed. Ideally, we would have had a left skewed histogram.

Variable	**Weight**
Age	5 points if over 65
Gender	1 point added if female
Race	2 points added if white
PVD, CVD, CHF	10 points added if "yes" to having one or more
Activity	1 point subtracted for each answer above 1; -1 point for “stand or walk a lot” (2) -2 points for “lift light loads or climb stairs” (3) -3 points for “heavy work and heavy loads” (4)
Hypertension	3 points added if yes
Diabetes	5 points added if yes
Anemia	11 points added if yes
	Total Points Possible: 37 Get screened if score >= 6

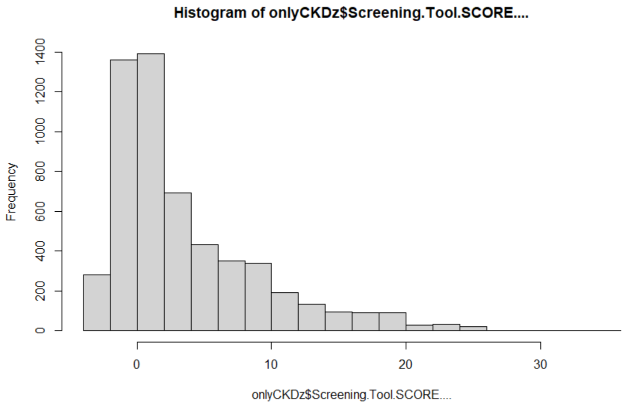


Additionally, CHF, CVD, and PVD all have common underlying pathologies- fatty build ups within blood vessels, which end up restricting circulation. Therefore, we combine these variables into one single question, as to not overweight our screening tool in favor of cardiovascular disorders. As all three of these variables are around .11 and .10, we average them to be .11, or 11 points. As the interpretations of the coefficients for Family Hypertension and BMI do not make sense in our screening tool context, we drop them from our tool. We therefore are left with the variables as shown on the table.

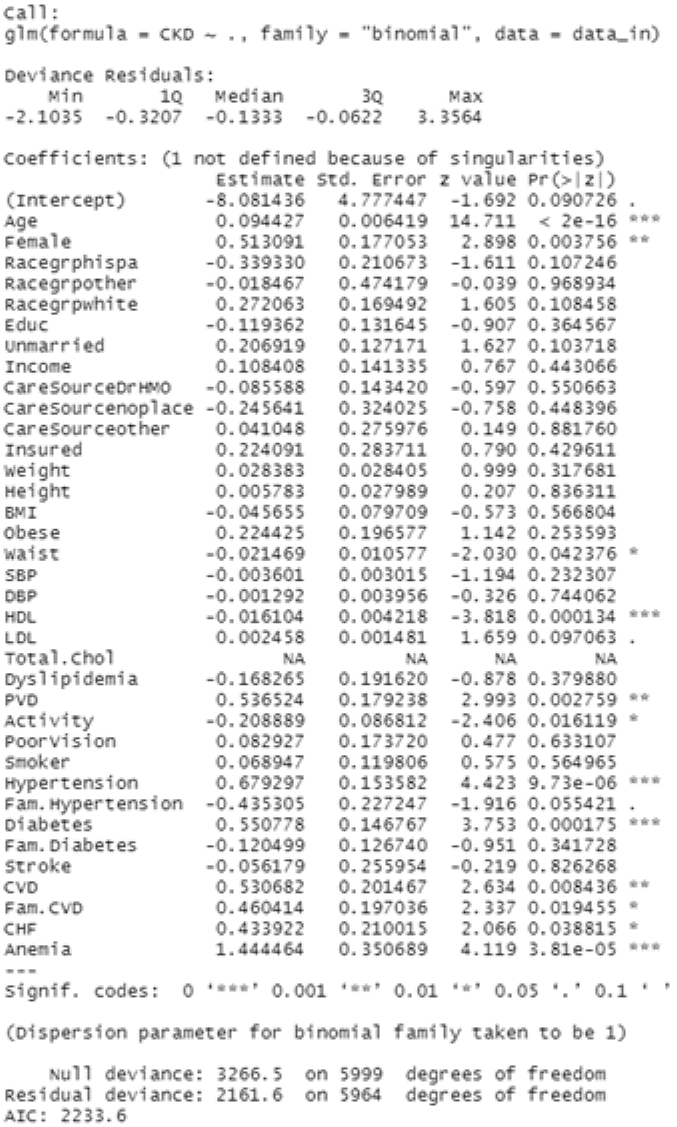
SCREENING TOOL THRESHOLDS

Next, we must determine at what point threshold will we screen for CKD. We compare the first 6000 rows to observe their screening tool scores and CKD information. We find that on average, those with CKD have a mean score of 13.62 and a median of 13 (indicating low outliers). Therefore, a screening point threshold of 13.62 would capture half of the individuals with CKD.
Conversely, those who did not have CKD have a mean screening tool score of 4 and median of 2. We plot the histogram of those without CKD, and see that it is right skewed, which is promising. Based on these histograms, we would set our point threshold at 6 in order to maximize the amount of CKD true positives and minimize the number of CKD false positives.

[6] https://www.econometrics-with-r.org/11-1-binary-dependent-variables-and-the-linear-probability-model.html
[7] https://jeffbloem.com/2019/09/06/binary-dependent-variable-just-use-ols/

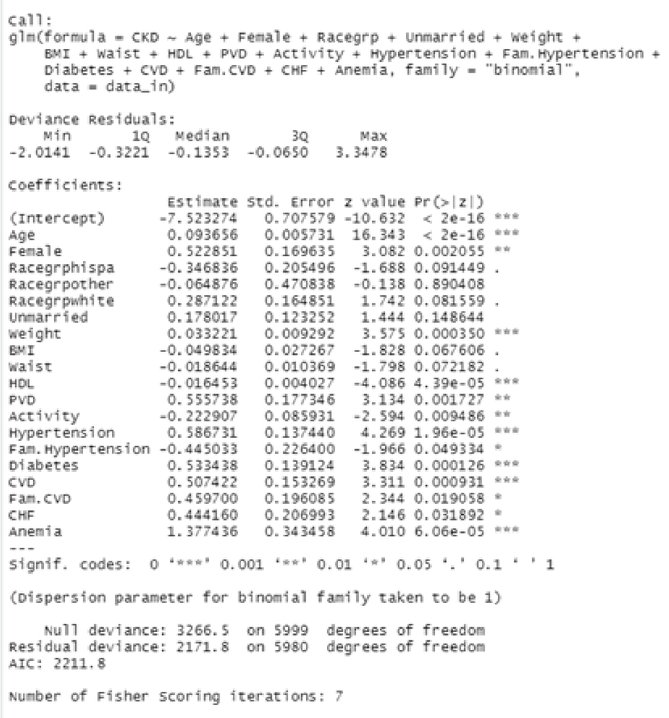

Histogram of onlyCKDz$Screening.Tool.SCORE....
Frequency
0
200
400
600
800
1000
1200
1400
0
10
20
30
onlyCKDz$Screening.Tool.SCORE....

APPENDIX



```
Call:
glm(formula = CKD ~ ., family = "binomial", data = data_in)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.1035  -0.3207  -0.1333  -0.0622   3.3564  

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -8.081436   4.777447  -1.692 0.090726 .  
Age                0.094427   0.006419  14.711  < 2e-16 ***
Female             0.513091   0.177053   2.898 0.003756 ** 
Racegrphispa      -0.339330   0.210673  -1.611 0.107246    
Racegrpother      -0.018467   0.474179  -0.039 0.968934    
Racegrpwhite       0.272063   0.169492   1.605 0.108458    
Educ              -0.119362   0.131645  -0.907 0.364567    
Unmarried          0.206919   0.127171   1.627 0.103718    
Income             0.108408   0.141335   0.767 0.443066    
CareSourceDrHMO   -0.085588   0.143420  -0.597 0.550663    
CareSourcenoplace -0.245641   0.324025  -0.758 0.448396    
CareSourceother    0.041048   0.275976   0.149 0.881760    
Insured            0.224091   0.283711   0.790 0.429611    
Weight             0.028383   0.028405   0.999 0.317681    
Height             0.005783   0.027989   0.207 0.836311    
BMI               -0.045655   0.079709  -0.573 0.566804    
Obese              0.224425   0.196577   1.142 0.253593    
Waist             -0.021469   0.010577  -2.030 0.042376 *  
SBP               -0.003601   0.003015  -1.194 0.232307    
DBP               -0.001292   0.003956  -0.326 0.744062    
HDL               -0.016104   0.004218  -3.818 0.000134 ***
LDL                0.002458   0.001481   1.659 0.097063 .  
Total.Chol               NA         NA      NA       NA    
Dyslipidemia      -0.168265   0.191620  -0.878 0.379880    
PVD                0.536524   0.179238   2.993 0.002759 ** 
Activity          -0.208889   0.086812  -2.406 0.016119 *  
PoorVision         0.082927   0.173720   0.477 0.633107    
Smoker             0.068947   0.119806   0.575 0.564965    
Hypertension       0.679297   0.153582   4.423 9.73e-06 ***
Fam.Hypertension  -0.435305   0.227247  -1.916 0.055421 .  
Diabetes           0.550778   0.146767   3.753 0.000175 ***
Fam.Diabetes      -0.120499   0.126740  -0.951 0.341728    
Stroke            -0.056179   0.255954  -0.219 0.826268    
CVD                0.530682   0.201467   2.634 0.008436 ** 
Fam.CVD            0.460414   0.197036   2.337 0.019455 *  
CHF                0.433922   0.210015   2.066 0.038815 *  
Anemia             1.444464   0.350689   4.119 3.81e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3266.5  on 5999  degrees of freedom
Residual deviance: 2161.6  on 5964  degrees of freedom
AIC: 2233.6
```

Figure 1. Logistic regression using all variables



```
Call:
glm(formula = CKD ~ Age + Female + Racegrp + Unmarried + Weight + 
    BMI + Waist + HDL + PVD + Activity + Hypertension + Fam.Hypertension + 
    Diabetes + CVD + Fam.CVD + CHF + Anemia, family = "binomial", 
    data = data_in)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.0141  -0.3221  -0.1353  -0.0650   3.3478  

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -7.523274   0.707579 -10.632  < 2e-16 ***
Age               0.093656   0.005731  16.343  < 2e-16 ***
Female            0.522851   0.169635   3.082 0.002055 ** 
Racegrphispa     -0.346836   0.205496  -1.688 0.091449 .  
Racegrpother     -0.064876   0.470838  -0.138 0.890408    
Racegrpwhite      0.287122   0.164851   1.742 0.081559 .  
Unmarried         0.178017   0.123252   1.444 0.148644    
Weight            0.033221   0.009292   3.575 0.000350 ***
BMI              -0.049834   0.027267  -1.828 0.067606 .  
Waist            -0.018644   0.010369  -1.798 0.072182 .  
HDL              -0.016453   0.004027  -4.086 4.39e-05 ***
PVD               0.555738   0.177346   3.134 0.001727 ** 
Activity         -0.222907   0.085931  -2.594 0.009486 ** 
Hypertension      0.586731   0.137440   4.269 1.96e-05 ***
Fam.Hypertension -0.445033   0.226400  -1.966 0.049334 *  
Diabetes          0.533438   0.139124   3.834 0.000126 ***
CVD               0.507422   0.153269   3.311 0.000931 ***
Fam.CVD           0.459700   0.196085   2.344 0.019058 *  
CHF               0.444160   0.206993   2.146 0.031892 *  
Anemia            1.377436   0.343458   4.010 6.06e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3266.5  on 5999  degrees of freedom
Residual deviance: 2171.8  on 5980  degrees of freedom
AIC: 2211.8

Number of Fisher Scoring iterations: 7
```

Figures 2. Logistic regression using stepAIC backwards