Introduction

Hypertension (high blood pressure) is a condition where the pressure of the blood pumping against the walls of the arteries is elevated beyond normal levels. The goal of this analysis is to predict the history of hypertension in pregnant women with no previous pregnancy lasting 20 weeks or more (nulliparas). Specifically, there is an importance to identifying those who have a history of hypertension (high blood pressure). A history of hypertension might be an indication of a higher risk for developing gestational hypertension. For some women the development of gestational hypertension may increase their risk for more serious complications later in pregnancy, like preeclampsia.

Our preliminary results indicate that predicting history of hypertension is a difficult, if not impossible, problem to solve, but our results highlight what approaches are possible. Adaboost (which is an ensemble technique further discussed in the methods section) is the best performing model using variables like systolic, diastolic, and race, which were deemed important, but even this comes with its set of limitations and nuances which are discussed in this report.

Methods

The goal of the analysis is to predict the history of hypertension in pregnant women. The information provided to tackle this problem includes many items from the medical history of the patient. We want to train a statistical model that can use this information to predict whether or not the person has a history of hypertension, which could potentially serve as a proxy for predicting whether the person will develop hypertension during pregnancy.

Since our data is imbalanced (only 3% of individuals in the data have a history of hypertension), we upsampled the training set in an attempt to have our model learn from more examples of hypertension cases. Synthetic Minority Over-sampling Technique (SMOTE) was used to upsample the training data so that the number of observations with hypertension matched the number of observations without hypertension. The algorithm is similar to K-Nearest Neighbors, where the algorithm generates examples from the predictors which are neighbors of the predictors from the minority class. Since SMOTE requires the computation of numeric distances between neighboring points, it is necessary for the data to be numeric.

One-hot encoding was used to create numeric representations of categorical variables. One-hot encoding transforms a categorical variable into several columns where each categorical level has a column with entries of 1 or 0, with 1 indicating that the column has the given property. For example, the race variable was split into multiple columns like race.white, race.black, race.hispanic etc where exactly one of these columns had a 1 and others were 0. Several categories had very few responses, so columns with low variance were dropped from the training of some of our models. However, some variables with low variance were retained in the model due to the underlying social factors associated with hypertension, such as including native american women in the dataset.

Before the models were fitted with the preprocessed data, we had to identify appropriate evaluation metrics to enable comparison of different models. A confusion matrix allows us to understand how well the model is doing at predicting different kinds of outcomes i.e true positives, true negatives, false positives and false negatives. We use metrics such as Recall, Precision, and F1 score to provide a more holistic view of model performance. Recall represents the percentage of women with history of hypertension correctly classified by the model as having a history of hypertension, while Precision represents the percentage of women who actually had a history of hypertension amongst all women classified as having a history of hypertension. Finally, the F1 score is the weighted combination of recall and precision, generally described as the harmonic mean of two. These metrics are clearly more powerful in describing the usefulness of our models, as simply looking at accuracy (percentage classified correctly) can be misleading due to the data being imbalanced.

Another aspect of model training was to be cautious about overfitting. To tackle this, we used 5-fold cross validation for all models that were fitted. Cross-validation is a resampling procedure used to evaluate models on a limited data sample. The k-fold cross validation splits data into k groups, where k-1 groups are trained and the last is used for validation. This helps models utilize all the information in a limited data set and avoid overfitting. We use this approach to tune parameters which cannot be done strictly using the training set because otherwise we risk overfitting.

Given the combination of categorical and numerical variables, and our goal of predicting a binary outcome (presence or absence of history of hypertension), we settled on the following models that are suitable for binary classification problems:

Logistic regression

Building a logistic regression model will allow us to determine how strong the association is between two or more factors and the patient’s condition. Logistic regression can be used to predict whether a patient will have hypertension. Advantages to using logistic regression include that it is easy to interpret. It not only measures the association between variables, but also measures the direction of the association (positive or negative). A limitation is the assumption of linearity between the outcome and independent variables. With logistic regression, it is difficult to model complex relationships.

Since there are many variables in the data set, we will consider using backward and forward selection to obtain the best model.

Multi-Layer Perceptron

Multi-Layer Perceptron is a relatively simple form of Neural Network, where the information is feedforward from input to hidden and then to output layer. It utilizes backpropagation, an algorithm through which one feeds forward the values, calculates the errors and propagates it back to the earlier layers. Advantage comes that its multiple layers and non-linear activation function can help to classify data that is not linearly separable. A limitation for this approach is the shortage of interpretation, that it relies on complex co-adaptations of weights during the training phase instead of measuring and comparing quality of splits, and thus we are not able to easily access information other than prediction scores.

K-nearest neighbors (KNN)

This technique uses an integer k and a similarity measure (typically euclidean distance) to find the k nearest points determined by the similarity measure, and makes the prediction based on the classification of the majority of the neighbors. If the variables have very different ranges and units, all variables will be standardized. Generally, a smaller value of k gives a more flexible model. For example, when k = 1 each new point would be classified by only its nearest neighbor, thus, a small k can lead to overfitting when training the model. Higher dimensions (more predictors) mean that points are generally farther apart which means that neighbors cover a larger area which increases the model's bias. Reducing the number of neighbors reduces the bias, but the predictions become noisier because there are not as many neighbors to balance leading to noisier predictions and a higher variance. Cross-validation will be used to choose the value of k, the number of neighbors. The results of the model are not very interpretable and the method does not offer a form of variable selection. Further, it cannot use categorical variables as predictors.

AdaBoost

AdaBoost is an ensemble type method where a collection of "weak learners" (e.g. stumps of decision trees) are fit in a sequential manner, and each of these learners makes a prediction. Points which are misclassified are upweighted to penalize the misclassification, and points which are correctly classified are down weighted. The final classification is decided by aggregating the decision of those weighted classifiers. AdaBoost was chosen as boosting type models are typically comparable to Random Forest models, and allows us to access variable importance, so some of the mysteries of black box models are avoided. Limitations of the model are twofold: hyperparameter tuning is computationally expensive, and there is still some inaccessibility as to why the final classifier is making certain decisions. As the model is fit in a sequential order, parallel processing is not possible. Completing a small grid search for optimal hyperparameters took between 12 to 30 minutes, and larger grid searches seem to yield diminishing returns in terms of increases in precision.

Random Forest

Similar to AdaBoost, Random Forest is another ensemble technique which aggregates multiple learners in order to make a prediction. The key difference between AdaBoost and a Random Forest is in the sequentiality during the model training phase. A Random Forest is just a collection of decision trees (or stumps), and each tree's prediction can be obtained in parallel. For a classification problem like this one, the final prediction of the Random Forest is whatever the majority of trees predicted as the outcome (hypertension, True or False). Similar to AdaBoost, Random Forests also provide information about variable importance so the model is not entirely a black box. Furthermore, an arbitrarily picked decision tree from the forest is presented in an attempt to aid interpretability. However, a Random Forest is still not inherently interpretable in terms of inference of causal relationships which is one of its primary relationships. It cannot provide information about how each variable of the dataset affects the outcome (like Logistic Regression for example). Since the goal was to optimize on the prediction accuracy and a compromise on inference and interpretation was warranted, this limitation did not necessarily need to be addressed.

Results

The provided data is a subset of data from the NuMoM2b study, collected by interviews, questionnaires, clinical measurements, and so on. Each row in the dataset contains a patient's medical records and information including demographic, psychosocial, dietary, physiological, health, and pregnancy outcomes. The outcome variable (dv.hypertension1) our analysis focuses on is the history of hypertension recorded on the first visit. The data contains 7,934 observations, with 26 variables (25 predictors), and was split into a testing and training set of 2,380 and 5,554 observations, respectively. The imbalance of the data, where only 3% patients are recorded to have hypertension history, is the main issue.

Several variables had a level that encompassed all of the reasons that someone would not have an answer to a question (i.e. they did not know or refused to answer or the value is just missing). For example, in the age column, there are values that are zero when the patients in this dataset should all be older than zero.

All of the data variables are further summarized in the baseline table presented in the Appendix as requested by the client.

We present the results of each of the models along with some interpretation of them where necessary. Adaboost was the best performing model, but it is useful to look at the results from the other models as well.

Logistic regression

There is not a notable difference between the test error of the full model and the models chosen using forward and backward selection. Considering AIC as another model metric, the model chosen using backward selection is the best model as it has a lower AIC score. It is also a simpler model with less variables. Variables included in the model are age, race, systolic blood pressure, diastolic blood pressure, worry symptoms score, pre-pregnancy weight, history of kidney disease, history of PCOS, whether a patient has been discriminated against and whether the mother was born early. Results indicate that age, systolic blood pressure, diastolic blood pressure, pre-pregnancy weight, the mother being born early, history of kidney disease and history of PCOS are all significantly associated with having hypertension at a $\alpha = 0.05$ significance level. Further, the odds of having hypotension is .43 times lower if you are white compared to other races. On the other hand, the odds of having hypertension are 1.76 times higher among those with a history of PCOS.

Further, results indicate the model is good at predicting cases when individuals do not have a history of hypertension. However, the focus on our analysis is to be able to correctly predict those who do have a history of hypertension. Out of the 71 patients in the data set who do have a history of hypertension, only 2 of these individuals were correctly classified. The precision of this model is only 2.82%.

Multi-Layer Perceptron

All variables were included in the model, with One Hot encoding on variable race, MinMaxScalar (normalization) on numeric features, and dropping duplicate samples. After that, the training data size was 5516 samples and 29 feature columns. Income, age and pre-pregnancy weight with zero entries that were considered to be missing value, and were replaced with their mean values. Through grid search and 5-fold cross validation, the model was tuned to find optimal hidden layer size, activation function, solver, learning rate type, and l2 penalty parameter based on recall score criteria. MLP was tuned in two scenarios: the one fitted on the original imbalanced data only had 9 patients to be errorly predicted hypertension yet only 6 out of 71 history hypertension cases were found. Its precision score was up to 0.4 while poor behavior in recall score (0.0845) and f1 score (0.1395). The other one fitted on the upsampled data, which contains the same size for each class. It turned out to correctly predict 46 out of 71 hypertension cases, yet 187 patients were errorly predicted to have history of hypertension, and this led to a higher recall score (0.3521) and lower precision score (0.1179). However, the f1 score increased to 0.1767 after upsampled.

KNN

After trying different versions of KNN, including a weighted version using different kernels and one that was trained on the upsampled data, it was determined that the regular KNN performed just as well as the other versions in terms of overall accuracy and performed very close to the weighted model when comparing recall scores (difference of about .01). Choosing the simpler model provides a more straightforward methodology. These models were tuned using 5 fold cross validation and both the overall accuracy and recall were calculated. Recall was a determinant in the model selection because correctly identifying the women with hypertension is important. The most accurate models did not correspond to the best recall models. After tuning, the optimal parameter value found was k = 2. This result feels as though it is overfitting because of the very small value of k. However, considering the KNN methodology, classifying based on the majority of the neighbors' classifications, given that by far most of the data is patients without hypertension, adding more neighbors will lead to nearly always predicting not having hypertension which undermines the most important goal, accurately identify as many people with hypertension as possible. The overall accuracy of the model was about .95, which was about .02 lower than the most accurate regular KNN models, but the much higher recall of .0845 for this model made this the best KNN model overall.

AdaBoost

Similar to what was done in the MLP pre-processing, predictors which were factors were converted to one-hot encoding, and numeric columns were centered and scaled, and numeric columns which had NA entries (indicated by 0s), were imputed to be the mean of the non-zero entries of that column (e.g. age, pre-pregnancy weight). Of the 82 predictors in the one-hot encoded predictor matrix, 37 of those predictors had nearly zero variance. Consideration was taken as to not remove variables that had low variance, but might be important to include for equity reasons, such as the race variable. The ten most important variables as determined gain in Gini index is given in Figure xxx. Both diastolic and systolic blood pressure taken at the time of first visit were the two most important by variables in predicting history of hypertension, this make some sense as the definition of hypertension is a systolic pressure greater than or equal to 130, or a diastolic pressure greater than or equal to 80.

The optimal models were found using a grid search over the number of trees used, and max depth of those trees where the target metric was to maximize the area under the precision-recall curve (PRAUC). Two models were fit: one trained on the one-hot encoded training data as described in the previous paragraph, and one on the upsampled SMOTE training set, which contained the same number instances for each class.

The model trained on the original unbalanced data yielded correctly predicted a woman having a history of hypertension 0.0845 of the time (precision), and the model was only able to identify about a third of women predicted to have had hypertension correctly (recall). The F1 score is 0.1348, which is the harmonic mean of the precision and recall. In contrast the model trained on balanced upsampled data correctly predicted a woman having a history of hypertension 0.1408 of the time, and similarly, the model was only able to correctly predict if a woman had hypertension sometime in the past correctly about a third of the time. The F1 score increased to 0.1923 is the harmonic mean of the precision and recall. The overall test accuracy of both models was nearly 97%, which is mostly the byproduct of the unbalanced classes. These results demonstrate how upsampling can be used to train a model to increase the precision of its prediction, although the gains are marginal at best.

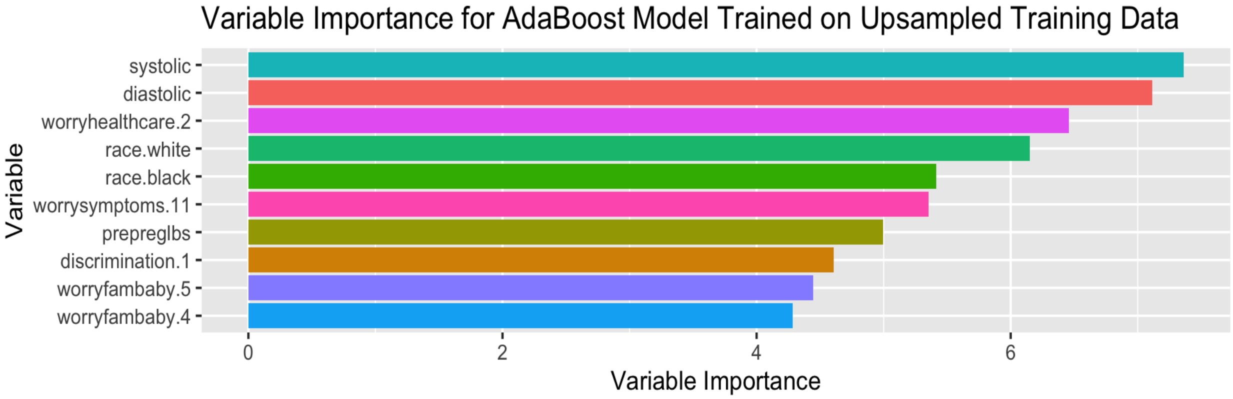


Figure 1: Variable Importance of AdaBoost Model

Random Forest

The data preprocessing was performed similarly to that of the other models with categorical variables being one-hot-encoded. Variables for the model were selected based on the variable-importance that Random Forests provide. Once the variables were selected, models were fitted on both the original imbalanced data as well as the SMOTE-balanced data. The model that was trained on the imbalanced data was quite unusable as it predicted 100% of test samples to not have hypertension. While this is still a high accuracy model (97%) it is not ideal when comparing more meaningful metrics like the F1 Score, Precision, and Recall, all of which were 0. However, the random forest that was fitted on the balanced data performed much better. While the overall accuracy of predictions dropped a little, this was acceptable as other metrics improved. This model had a really high recall score (0.59) and a precision score of 0.09, to give an F1 metric of 0.15.

As for the details about the training itself, a Cross Validation approach with Grid Search was used, with varying the values of two hyperparameters (number of estimators, and maximum depth of each tree). The best model selected as per Grid Search had 100 estimators and a maximum depth of 3. It must be noted that the model prefers high values of tree depth, but this would only lead to more overfitting and less interpretability (each tree is very dense) so the possible values were constrained.

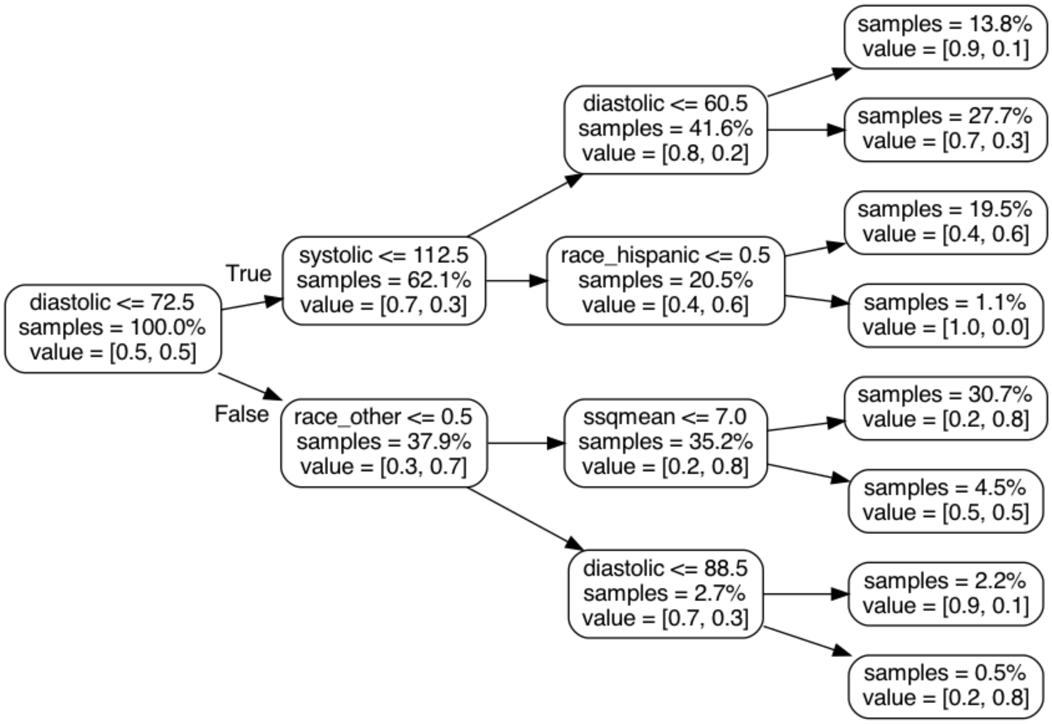


Figure 2: One of the 100 decision trees used by the Random Forest model. Sample represents percentage of data present in that node, and values is the False/True split of having history of hypertension.

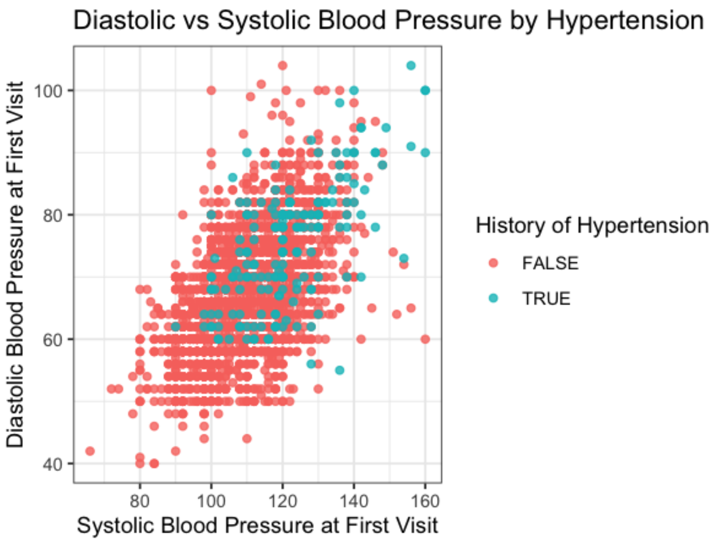


Figure 3: This figure plots the diastolic versus systolic blood pressure of each patient at the time of their first visit, it also displays whether or not the patient has a history of hypertension.

Finally, we present in Table 1, the performance of each model for easy comparison. While there is no single model that holistically outperforms the rest, Adaboost, Random Forest, and MultiLayer Perceptron are more promising than the others. Adaboost has the best F1 Score which is the metric of interest, and as described in Figure 1, deems the diastolic, systolic, and race variables among others to be most important in the prediction. This is reiterated by the Random Forest model as well, as Figure 2 shows decisions being based on these same variables. Moreover, Figure 3 depicts the relationships between the systolic and diastolic variables while also depicting the history of hypertension. Finally, we also learned that upsampling the data is particularly helpful in building a better predictive model.

Table 1: Model Metrics on the Test Data

Model	Recall	Precision	F1 Score=2*Precision*Recall/ (Precision+Recall)
AdaBoost	0.3333	0.0845	0.1348
AdaBoost w/ minority class upsampled to match majority	0.3030	0.1408	0.1923
Random Forest w/ upsampling	0.59	0.09	0.15
MLP	0.0845	0.4000	0.1395
MLP with minority class upsampled to match majority	0.3521	0.1179	0.1767
Logistic Regression (Full Model)	0.25	0.0281	0.0506
Logistic Regression (Forward/Backward Selection)	0.2222	0.0282	0.05
KNN	0.08451	0.10	0.09160
KNN with upsampling	0.09859	.09836	0.09091

Conclusion

The goal of this analysis was to be able to predict the history of hypertension in pregnant women with no previous pregnancy lasting 20 weeks-0 days or more estimated gestational age (nulliparas). While trying to predict the history of hypertension, we provided how various models and approaches perform in different evaluations, and were able to get an understanding of what factors may contribute to a history of hypertension.

Results from the logistic regression model (which we treat as baseline) indicate that age, systolic blood pressure, diastolic blood pressure, pre-pregnancy weight, the mother being born early, history of kidney disease and history of PCOS are all associated with having a history of hypertension. Further, the odds of having hypertension is lower if you are white or hispanic compared to other races. The best performing Adaboost model is mostly in agreement with these findings, as it picks systolic, diastolic, and race as some of its most important variables in predicting the outcome.

A limitation of this analysis is the class imbalance of our data. The majority of the patients in this data set do not have a history of hypertension. This leads to issues as our primary focus is to predict whether a patient has a history of hypertension. While this limitation is addressed by the upsampling techniques, it is definitely not comparable to having balanced data in the first place.

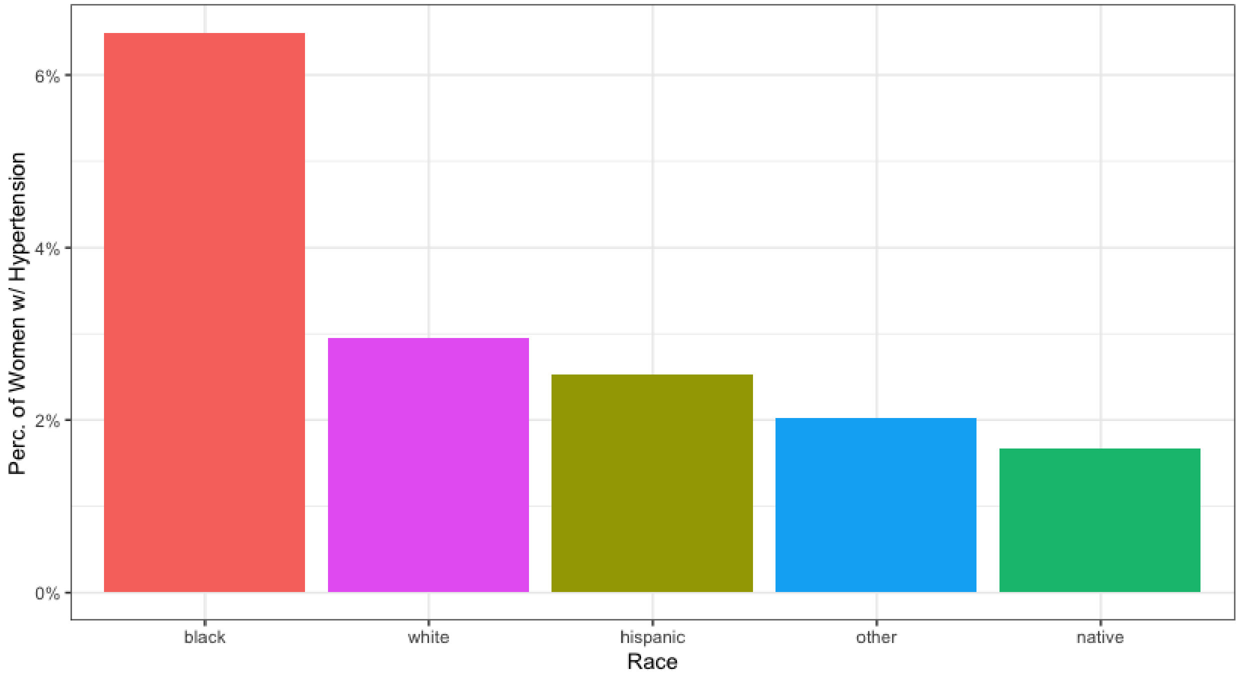
Appendix:

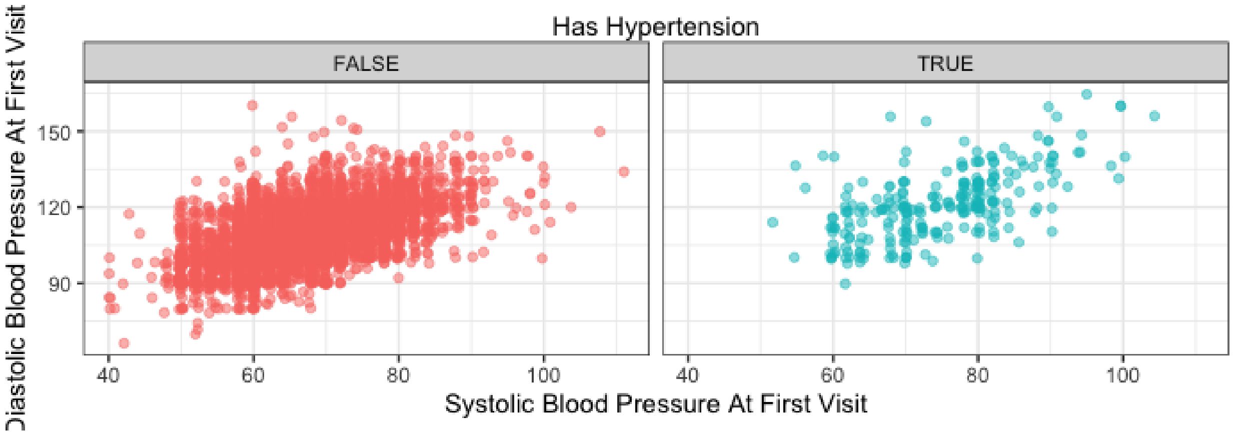
Table 2: Baseline Table of Descriptives

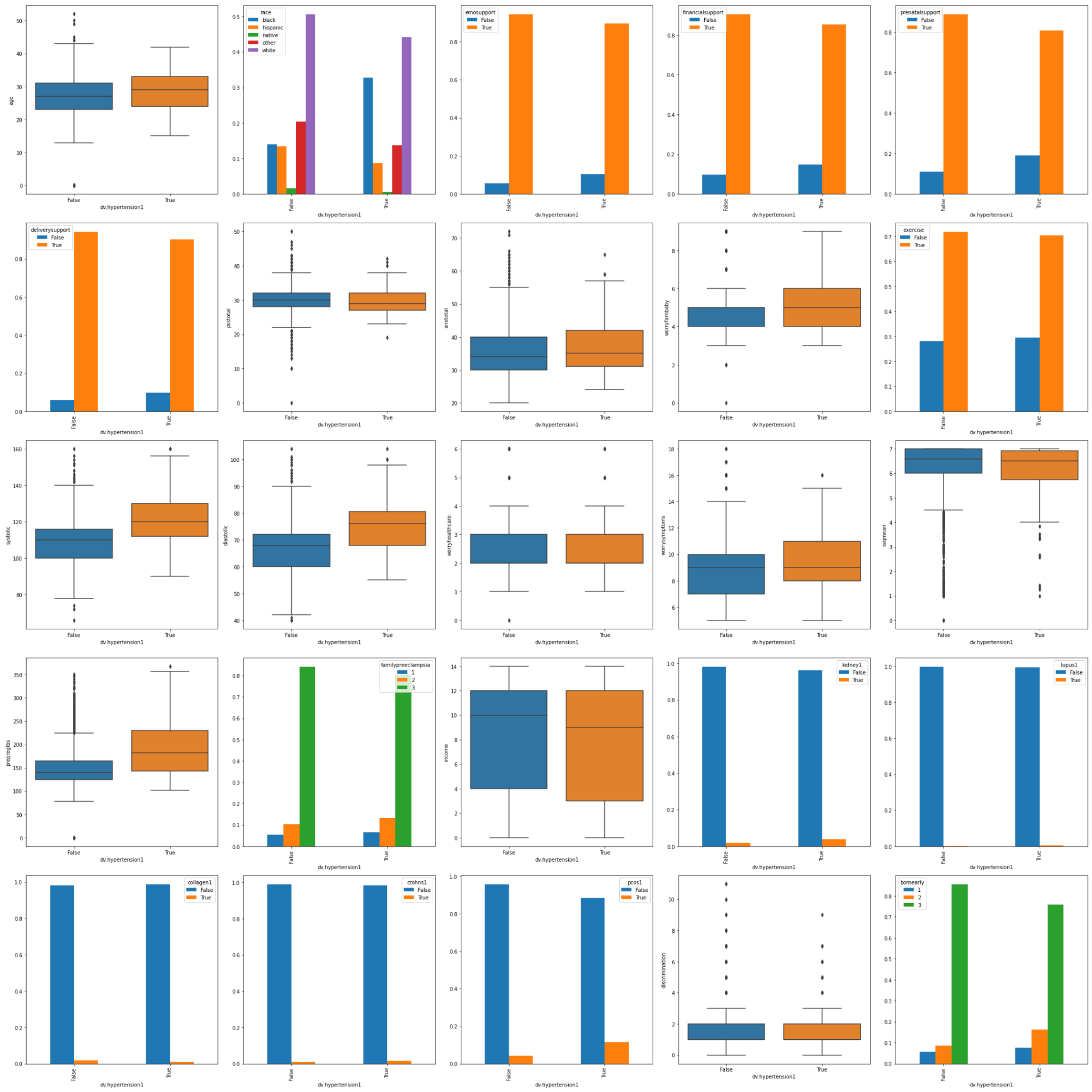
Variable	**Number of Unique Factors**	**Top Factor Counts or IQR**	**Mean or Proportion**
race	5	whi: 4003, oth: 1588, bla: 1155, his: 1069	NA
worryfambaby	9	4: 2663, 5: 2224, 3: 1208, 6: 1117	NA
worryhealthcare	7	2: 4464, 3: 1983, 4: 986, 5: 342	NA
worrysymptoms	15	8: 1516, 9: 1416, 7: 1209, 10: 1102	NA
familypreeclampsia	3	3: 6681, 2: 838, 1: 415	NA
discrimination	12	1: 4493, 2: 2036, 3: 524, 0: 303	NA
bornearly	3	3: 6791, 2: 699, 1: 444	NA
emosupport	2	TRU: 7470, FAL: 464	0.9415
financialsupport	2	TRU: 7137, FAL: 797	0.8995
prenatalsupport	2	TRU: 7026, FAL: 908	0.8856
deliverysupport	2	TRU: 7461, FAL: 473	0.9404
exercise	2	TRU: 5703, FAL: 2231	0.7188
dv.hypertension1	2	FAL: 7680, TRU: 254	0.032
kidney1	2	FAL: 7798, TRU: 136	0.0171
lupus1	2	FAL: 7917, TRU: 17	0.0021
collagen1	2	FAL: 7800, TRU: 134	0.0169
crohns1	2	FAL: 7865, TRU: 69	0.0087
pcos1	2	FAL: 7575, TRU: 359	0.0452
age	NA	(23,31)	27.2328
psstotal	NA	(28,32)	29.7545

anxtotal	NA	(30,40)	35.4239
systolic	NA	(100,118)	109.1274
diastolic	NA	(60,72)	67.1273
ssqmean	NA	(6,7)	6.2057
prepreglbs	NA	(125,168)	151.7918
income	NA	(4,12)	7.95

Exploratory Data Analysis:






age
dv.hypertension1
False
True
race
black
hispanic
native
other
white
emosupport
financialsupport
prenatalsupport
deliverysupport
pgtotal
anxtotal
worryfambaby
exercise
systolic
diastolic
worryhealthcare
worrysymptoms
ssgmean
prepregbs
familypreeclampsia
1
2
3
income
kidney1
lupus1
collagen1
crohns1
pcos1
discrimination
bornearly