The U.S. Census

A socioeconomic look at factors indicative of wealth



*"Any society, any nation, is judged on the basis of how it treats its weakest members -- the last, the least, the littlest." – (Cardinal Roger Mahoney, 1998, Creating a Culture of Life)*

A group effort by:

John Benischeck

Tommy Baw

Shachi Parikh

Xiaoqin Helen Yi

Xiaomeng Blair Chen

Table of Contents

**Abstract**3

**Introduction**3

**The Data Collection process**3

**The Data Cleaning process**4

**Data Exploration**5-7

**Predictive Modeling**8-17

GlmNet Regression8-10

Logistic Regression10-13

Logistic Regression14-17

**Conclusion and Remarks**18

**Bibliography and References**19

**Appendix**20-26

**Individual Contribution and Addressing Comments**27

**Abstract**

The purpose of this project is to identify key social/economic factors that are statistically proven to have strong influence on an individual’s level of income. The task is to build predictive models which determine whether or not a person makes more than $50,000 per year. While the predictive algorithm will be the main focus of the work, we will also use data visualization and correlation analysis to take a look at other potential correlations, such as ethnicity to education levels, and gender to income, to extract further information beyond the predictive models.

**Introduction**

What is the American Dream? How do I achieve it? For many, that answer, and the path to it, will vary. Inevitably, those dreams will require money to pursue and maintain. News and media outlets regularly remind us that the paths for many Americans are impeded, and even blocked, due to a host of reasons. How can these roadblocks be overcome, and even removed, for future dream-seekers?

Once per decade, the United States government collects a wealth of information from and about its citizens. To the government, lobbyists, special interest and advocacy groups, this information is a goldmine; it is a reflection on the successes and, more importantly, the failures of the past decade’s governance and social policy on the population.

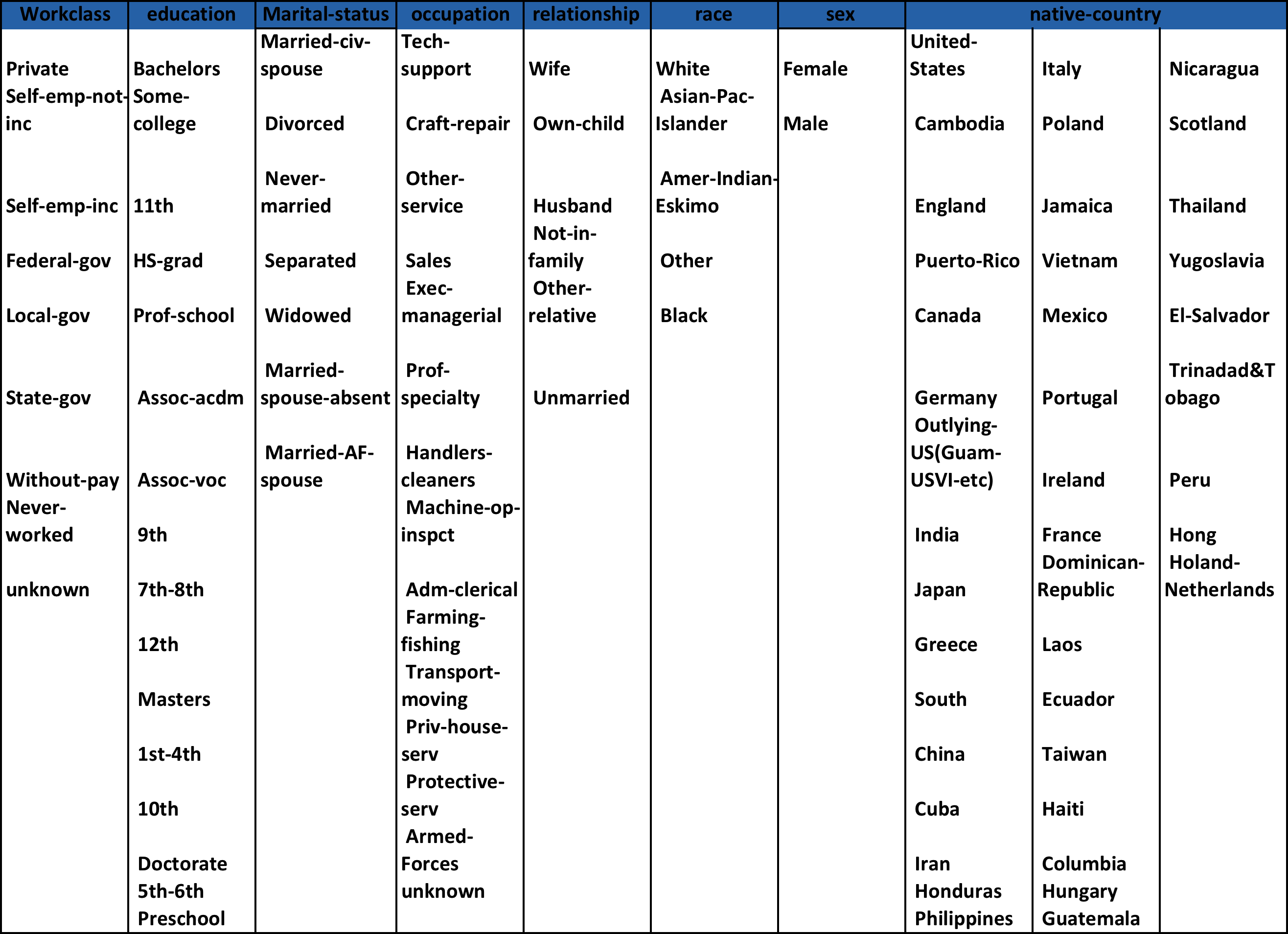
Many questions can be answered with this trove of data. The aim of this paper, however, is to explore commonly cited reasons for income inequality in the United States, and test their validity.

**Data Collection**

The dataset used in this project was pulled from the UCI Machine Learning Repository[[1]](#footnote-1), which further cites: *“Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))” (University of California Irvine, 1994*)*.* We concatenated their test and train sets, for a total of 48,842 observations and 15 attributes. In the modeling process, the variable “salary” is treated as the response, with string values of either “over 50k” or “below 50k”, indicating personal level of wealth.

**Data Cleaning**

Because the response variable, along with 8 predictor variables, are categorical, dummy variables were created in order to perform predictive algorithms and modeling. For the dependent variable “salary”, the dummy labels were assigned as follows: 1 for observations of earning “over $50k” per year and 0 for earning “below $50k” per year. Similarly, the same was done with the independent variables. For example, “work class” was expanded to 9 dummy, binary variables, “education” to 17 variables, “marital status” to 7 variables, and so on. This expanded our dataset to 109 predictors, and 1 response variable, ready to enter the models.



*The sub-categories of each categorical variable.*

Because of the categorical nature of the data, outlier detection was not a concern and, in the data exploration, we found that the sample was generally representative of the population. Additionally, it is worth noting that we observed missing values in the original dataset, and replaced them with “unknown” values. All of the missing data points were categorical and we believed that it would be more appropriate to treat “unknown” as a single dummy variable, rather than deleting an entire instance, when other information in that observation was provided.

**Data Exploration**

As laid forth above, we sought to use our data to find if certain socioeconomic and demographic factors could predict whether or not an individual is wealthy (defined as making more than $50,000 per year). We looked for early indications through visualization of the data of what a regression model might include as important factors of wealth. Additionally, we used visualization to check for anomalies in the data. There were, intuitively, three factors which we felt would be major contributors to wealth: Education, Sex (gender), and Race. Good education, as the trope and mantra goes, is the key to a good job, which implies wealth. Sex and Race are frequently touted as reasons for income inequality as well, and we wished to test these notions.

First, we wanted to know if our sample was representative of the population, and Race is perhaps one of the easiest ways to measure this. Per the below, “White” is nearly the entire sample population of our dataset. However, these results are generally consistent with the 1990 Census (Appendix A.2). Dissimilar from the percentages in Appendix A.2, the UCI dataset does not include the Race category “Hispanic”. UCI did not provide a reason for dropping this observation. However, and while this cannot be confirmed, it appears that the percentage overweight in the Race category “white” is roughly equal to the missing Hispanic value. Again referring to the Race “White”, Appendix A.1 shows that the sample population is lopsided in a 2:1 gender split favoring males. Normally, genders are relatively equal in number, typically favoring females by a small percent *(Resident Population Estimates of the United States by Sex, Race, and Hispanic Origin, 2001)*. These two quirks aside, and given the overall sample size, we accepted this dataset as generally random, and as a representative sample of the population.



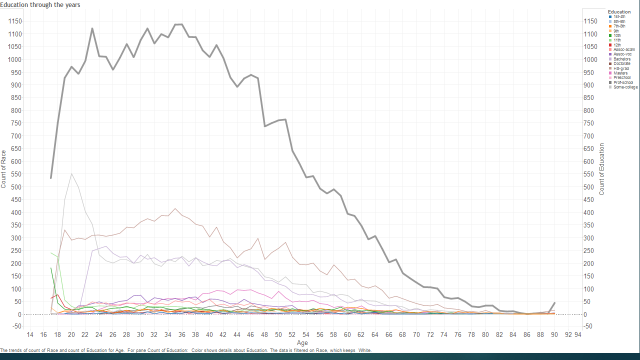
*Figure 2 and Appendix A.3*

After finding no critical issues with our dataset, we looked at the percentage of individuals at each education level making over $50,000 per year. As expected, the results show a positive trend where the higher the education level, the higher chances of earning more money. Yet there is some noise in this graph. For instance, it does not make sense for a pre-school student to be making any money given labor and education laws, and competency levels, among other issues.



*Figure 3 and Appendix A.4*

To account for this, we look at education levels across all age groups sampled. Seen below is a graph of total white population by education levels[[2]](#footnote-2) , which intuitively shows that certain percentages of the population achieve a particular level of education, and stop there until they die. This, despite the noise at the bottom of the graph, can most clearly be seen by the variable “high school grad”, which is the first line under the thick, total population line. In other words, certain segments of the population may have not received a full education, yet have managed to do well for themselves in their adult years. As will be discussed later, this success may also be accounted for by other rationale.



*Figure 4 and Appendix A.5*

If education could be indicative of wealth, does this apply equally across races and sex? The answer may not be so clear-cut when visualizing. Figures A.6 through A.9 in the appendix show the dispersion of higher education trends higher for the Race “Asian” while “White” and “Black” fall somewhere in the middle, and “Indian/Eskimo” lag behind. However, in the chart below, we see that the average of Black and White earning more than $50,000 per year are very far apart, and further see the gross inequality of the same between males and females.[[3]](#footnote-3)



*Figure 5 and Appendix A.10*

With the added complexity of gender inequality, we looked at available potential explanatory factors, such as average hours worked. Below, we see that, on average, women worked about 4 hours less than men per week. While this may have some role in the results shown in figure 5, it is clearly not enough to explain the chasm between genders.



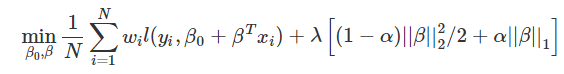
*Figure 6 and Appendix A.11*

By now, visualization has done its job in demonstrating that inequality does exist, and is perhaps explained by more than one factor. While these charts and figures may be enough for advocacy groups to fight for change, and government to review particular policies, we turned our focus to predictive modeling for a deeper analysis and potentially multiple causal factors.

**Method 1 – GlmNet**

The first and least meaningful method we tested was Glmnet, based on its ability to work with a binary predictor variable. GLMNet differs from Glm in that the method fits a generalized linear model via penalized maximum likelihood, with either lasso or elastic-net regularization, while Glm does not. This model tries to explain as much variance in the data as possible while keeping the model coefficients small. In addition to being fast, it can deal with the problem of sparsity in the input matrix Xi.

The algorithm of GlmNet is as follows:

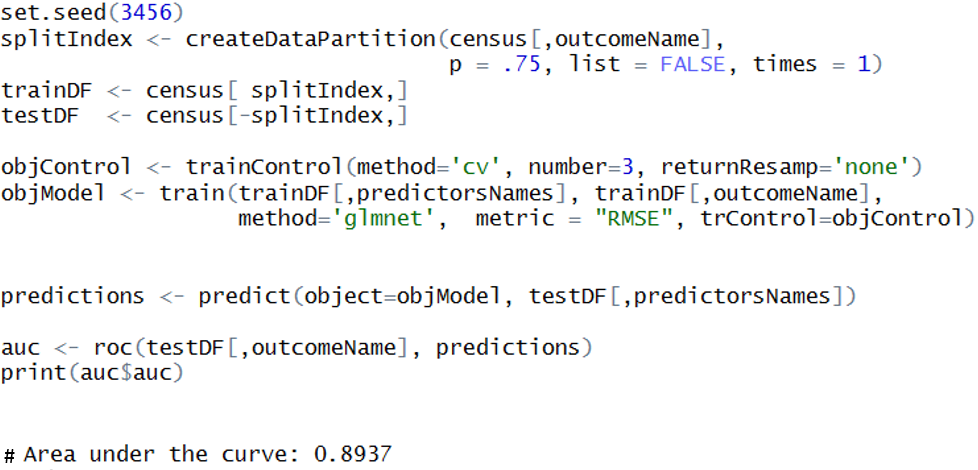


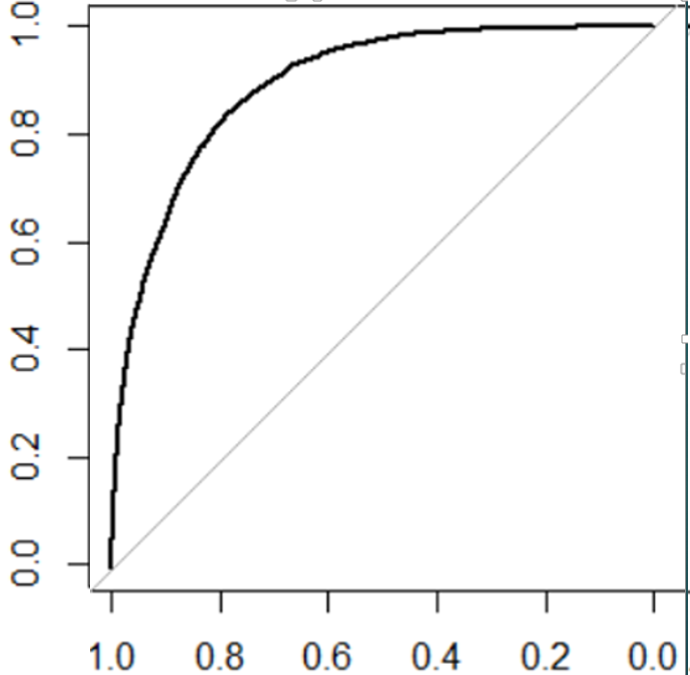
where variable is the weight, is the negative log-likelihood contribution for observation Xi, and α controls the penalty. Regularization parameter lambda covers the entire range, and penalizes the size of estimated coefficients *(Hastie & Qian, 2014).*

Using the following code, GlmNet suggested using 2 types of models:

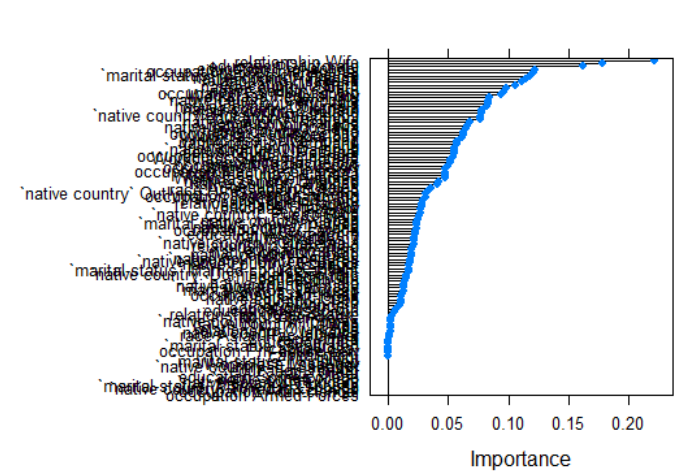


To test the predictive power of the model, we choose to split the data as 75% to train and 25% to test. Further, to show the accuracy of the model, we chose to use the Receiver Operating Characteristic (ROC) curve, further known as Area Under the Curve (AUC), as follows:

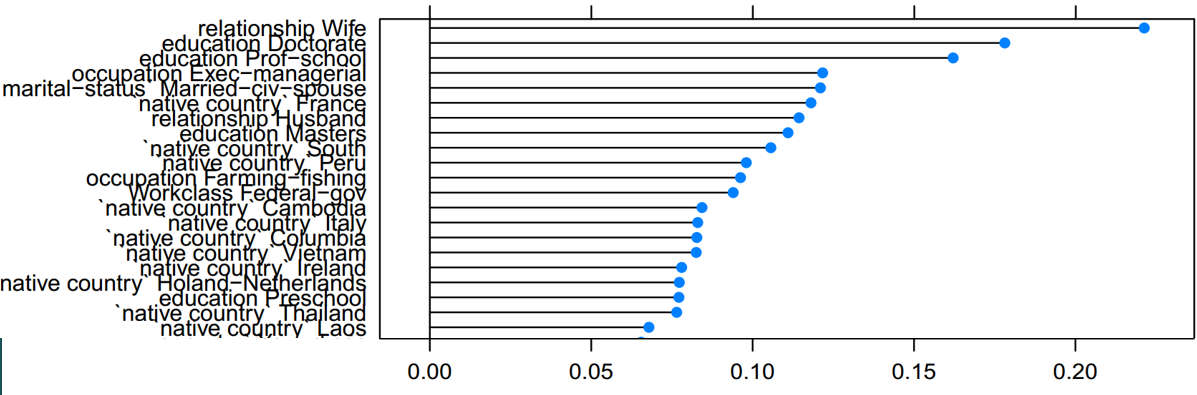




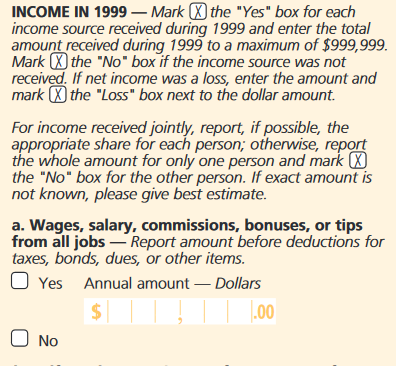
There is a lot of information to be derived from the AUC test, of which the most relevant to us is that the area under the curve is a measure of accuracy.  An area of 1 represents a perfect test, and an area of .5 represents a worthless test. *(Kutner, Nachtsheim, & Neter, 2004)* For GlmNet, the AUC value (accuracy) is 89.75%.

To help us understand which variables were important in the model derived by GlmNet, we used the function “varlmp” (abbreviation for Variance Importance) in the caret package. The resulting bar chart may have values on both the positive and negative side, which translates to a positive or negative impact on our model:

Naturally, this output can be messy when dealing with so many variables. To make a clearer output, the top 22 variables were retained:



Per the cleaner output, the top influential variable is “Relationship\_Wife”. There may be a few explanations as to why this variable stood out, not limited to a good education, or even reporting dual income, per the below Census snapshot. The second most important variable is “education\_Doctorate”. This echoes the result from the data exploration; the better the education, the more likely he/she will have a high paying job.



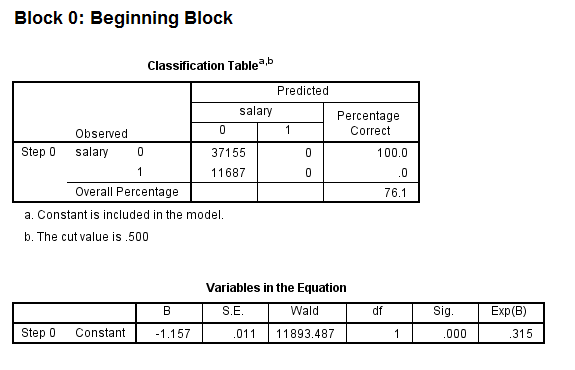
*Figure 7 – sample Census Form[[4]](#footnote-4)*

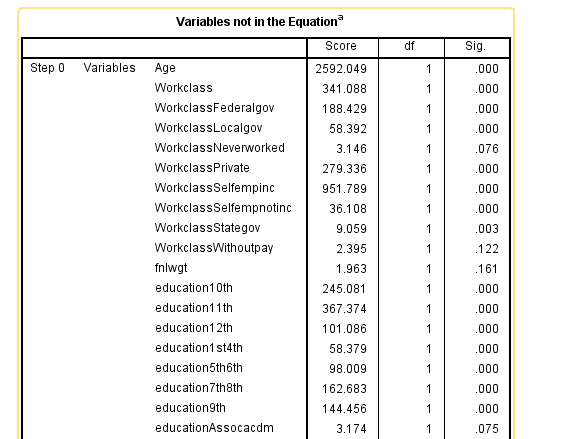
**Method 2 - Logistic Regression**

Quite often, a research question will contain in the dataset a Y-variable that is binary, or yes/no. Historically, when such a question was being tested, methods such as Linear discriminant function analysis(LDFA) and ordinary least squares (OLS) would be used. However, both methods have strict assumptions surrounding the data, and often these assumptions are violated. For example, OLS assumes linearity and continuity in the dataset. Because of our dummy variables, this would violate the continuity assumption  *(Peng, Lee, & Ingersoll, 2002)*.

To work around these strict assumptions, we turned to the logistic regression model. Logistic models are more apt to work with data which contains both continuous and categorical data, where the decision variable is dichotomous. *(UCLA, 2016).*

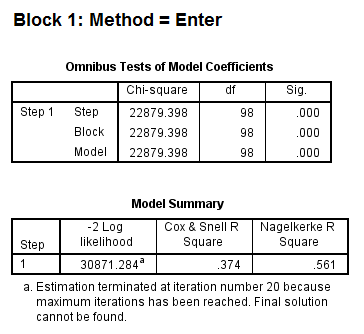
Using this regression method in SPSS yielded the following results:



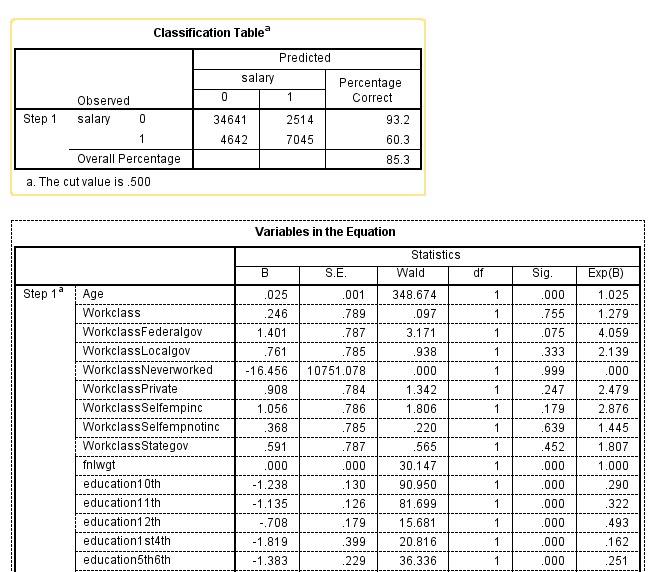


The first model of the output is a null model, which is to say SPSS fit a model with no predictors. The constant in the table labeled “Variables in the Equation” gives the unconditional log odds of variable admission to the model (i.e., admit=1).

This unfitted model is accompanied by a table labeled “Variables not in the Equation”, which gave the results of a score test known as the Lagrange multiplier test. The column labeled “Score” gave the estimated change in model fit if a term is added to the model, and the other two columns gave the degrees of freedom, and p-value (labeled “Sig.”) for the estimated change. Based on the p-value, almost all of the predictors shown were expected to improve the fit of the model.



The second test (output shown above) gave the overall test and score for a model that includes the predictor variables. The chi-square value of 22879.398, with a p-value of less than 0.0005, suggested that the model as a whole fit significantly better than the null model.

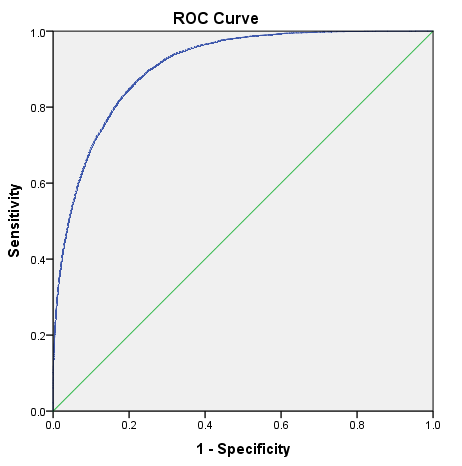


Additionally, this output included a classification table, which showed the overall usefulness of the model. As seen above, it accurately predicted/classified about 85% of the observations. In the table labeled “Variables in the Equation”, further diagnostics were given, such as the coefficients, their standard errors, the Wald test statistic with associated degrees of freedom and p-values, and the exponentiated coefficient (also known as an odds ratio).

Based on the p-values, there were 36 variables shown to be significant, some of which being age, education, native country relationship status, and occupation. This was a large model, so to reduce the dimensionality, we identified the most significant variables by looking at their odds ratio. Based on odds ratio, we found 18 significant variables:

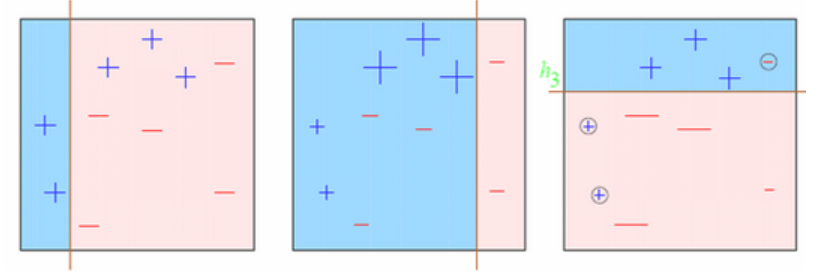
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Work Class** | **Education** | **Marital Status** | **Occupation** | **Native Country** |
| Federal Gov't | Doctorate | Married Armed Forces Spouse | Executive / Managerial | Columbia |
| Never Worked | Masters | Married Civilian Spouse |  | Holand / Netherlands |
| Private | Preschool |  |  | Laos |
| Self Employed INC | Professional School |  |  | Trinidad & Tobago |
|  |  |  |  | Vietnam |

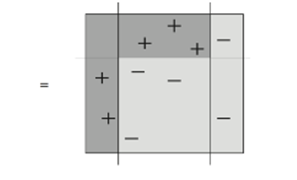
The final output below of interest was the Receiver Operating Characteristic curve. This model produces an ROC Curve accuracy rate of 90.80%, which means that the Logistic Regression model performed slightly better than GlmNet.



**Method 3 – Gradient Boosting Machine**

The last method we used was the Gradient Boosting Machine Model (GBM). With the versatility in the uses of GBM, a decision was made to use this as a classification method. The way this method works is similar to support vector machines, where it attempts to build a model by splitting the data through a hyperplane. The difference is that GBM is a type of ensemble method. That means that it is an additive model which runs several iterations to build a learning algorithm in a stage-wise manner. The end result is a model that finds a linear association between the past iterations. (Friedman, 2001) The figure below shows a step by step example of this process:





*Figure 8 – the GBM Model procedure[[5]](#footnote-5)*

There are many benefits to using this type of machine learning algorithm. The first is that it prevents over fitting by choosing the optimal number of iterations to run through monitoring of prediction error. Gradient boosting is also typically used with decision trees, so there are strategies to overcome over fitting by changing a combination of parameters like the learning rate (shrinkage), depth of tree, and number of trees. With correct tuning parameters, GBM generally provides better results than other ensemble methods such as random forests (Friedman, 2001).

The Gradient Boosted Method was run in R using the gbm and caret packages. The first task was in creating a new variable to force GBM into using the classification mode instead of the original binary dependent variable:

#use gbm by first creating a new classification variable

census$salary2 <- ifelse(salary==1,'yes','no')

census$salary2 <- as.factor(census$salary2)

outcomeName <- 'salary2'

From there, the data was split into a test and train set to evaluate the model. Setting a seed guaranteed that we would get the same split in subsequent runs:

#splitting into train and test data

set.seed(1234)

splitIndex <- createDataPartition(census[,outcomeName], p = .75, list = FALSE, times = 1)

trainDF <- census[ splitIndex,]

testDF <- census[-splitIndex,]

Using the caret package, we were able to get the most out of our model by controlling the resampling of the data through cross validation. In this case, we decided to cross-validate the data 10 times, therefore training it 10 times on different portions of the data to make sure we settled on the best tuning parameters.

Next we created the model and taught it how to recognize whether or not someone made above $50,000 per year:

objControl <- trainControl(method='cv', number=10, returnResamp='none', summaryFunction = twoClassSummary, classProbs = TRUE)

objModel <- train(trainDF[,predictorsNames], trainDF[,outcomeName],

method='gbm',

**trControl=objControl,**

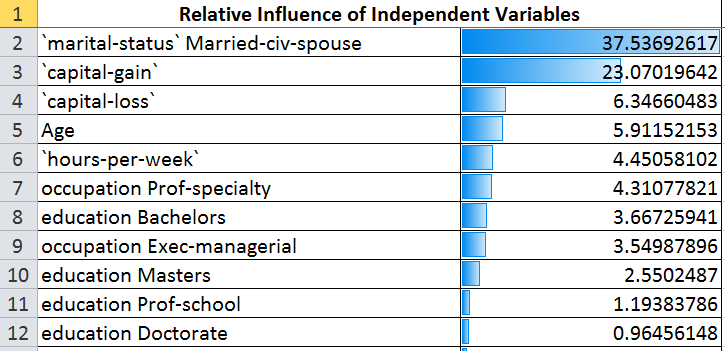
metric = "ROC",

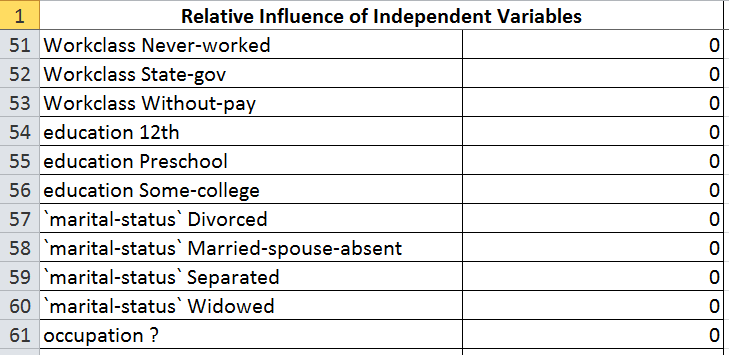
preProc = c("center", "scale"))

Below is a sample output of the learning process as the model improved through each iteration. The “TrainDeviance” measures the error, which the model tries to minimize it after each new iteration:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Iter** | **TrainDeviance** | **ValidDeviance** | **StepSize** | **Improve** |
| 1 | 1.0638 | nan | 0.1 | 0.0185 |
| 2 | 1.0339 | nan | 0.1 | 0.0149 |
| 3 | 1.0099 | nan | 0.1 | 0.0117 |
| 4 | 0.9869 | nan | 0.1 | 0.0118 |
| 5 | 0.9707 | nan | 0.1 | 0.008 |
| 6 | 0.9516 | nan | 0.1 | 0.0095 |
| 7 | 0.9357 | nan | 0.1 | 0.0078 |
| 8 | 0.9242 | nan | 0.1 | 0.006 |
| 9 | 0.915 | nan | 0.1 | 0.0046 |
| 10 | 0.9024 | nan | 0.1 | 0.0063 |
| 20 | 0.8335 | nan | 0.1 | 0.0025 |
| 40 | 0.758 | nan | 0.1 | 0.0009 |
| 60 | 0.717 | nan | 0.1 | 0.001 |
| 80 | 0.6925 | nan | 0.1 | 0.0004 |
| 100 | 0.6751 | nan | 0.1 | 0.0004 |
| 120 | 0.6634 | nan | 0.1 | 0.0003 |
| 140 | 0.6544 | nan | 0.1 | 0.0001 |
| 150 | 0.6508 | nan | 0.1 | 0.0001 |

By running a summary of the model, we were given the relative importance of each independent variable, and were able to place weighted importance. It also minimized more than half of our original variables to zero, where we could conclude that they did not have an impact on the boosted method. The first figure below shows the most important factors, which are marital status, capital gain, and capital loss. The second figure indicates that more than half of the variables are classified as unimportant.





With the model completed, we were interested in how well it performed. The model was able to classify salary with 86.5% accuracy, by observing the type 1 and type 2 errors highlighted in the confusion matrix below:

> #find out accuracy of model

> predictions <- predict(object=objModel, testDF[,predictorsNames], type='raw')

> print(postResample(pred=predictions, obs=as.factor(testDF[,outcomeName])))

Accuracy Kappa

0.8649357 0.5932111

> confusionMatrix(objModel)

Cross-Validated (10 fold) Confusion Matrix

(entries are percentages of table totals)

Reference

Prediction no yes

no 72.4 9.9

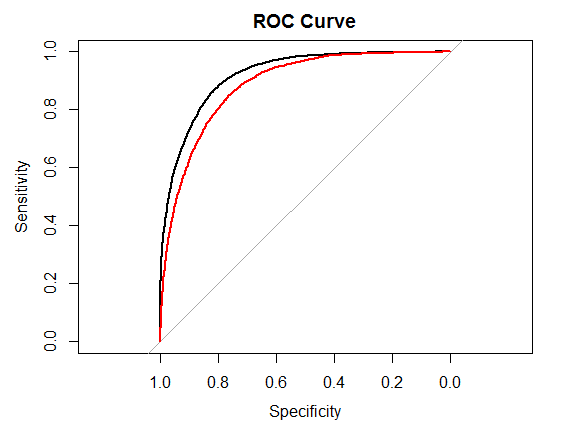
yes 3.7 14.0

Calculating the ROC, we observed that this method performed the best out of the three models used. The graph of the ROC curve also compares GBM in black to the GLMnet method in red. Clearly, GBM has outperformed since it has a larger area:

> auc <- roc(ifelse(testDF[,outcomeName]=="yes",1,0), predictions[[2]])

> print(auc$auc)

Area under the curve: 0.9193



**Conclusions and Remarks**

Of the three types of models we ran, they all performed very well. Despite a close performance by each model, the Gradient Boosting Machine performed the best. Depending on what is at stake, every additional percentage and even hundredth of a percentage of accuracy can make a big difference.

|  |  |
| --- | --- |
| **The Model Used** | **Area Under the Curve (AUC) Score** |
| GlmNet | .8975 |
| Logistic Regression | .9080 |
| Gradient Boosting Machine (GBM) model | .9193 |

While there were few similarities between the variables deemed important by each model, the ones which were similar dealt with education levels and marital status. Education is rather intuitive to interpret, as one would expect a higher level of education to be correlated with more specialized fields, which are typically highly paid.

The marriage level is more difficult to interpret. As seen from the snippet of the Census, there is the possibility that a family who filled out the census reported their income as it was reported on their joint tax return. That is to say that they may not have delineated who earned how much. This creates the several possible scenarios. For example, two spouses could work and jointly earn over $50,000 if the census respondents reported their joint income, where they technically both earn less than $50,000 per year.

The dataset was not without its shortfalls. Not including the “Hispanic” category of race deviated from normal Census reporting, and would have been an interesting feature to measure. Further, the sample size was perhaps too low to more smoothly and accurately look at categories of Race outside of “White”. Perhaps more interesting is what we didn’t observe. For instance, none of the three models placed a high importance on race, nor on gender, despite visually significant differences between them. This is possibly due to having a dichotomous dependent variable, the effects of marriage and joint income, a combination of the two, or perhaps even other variations which we do not have the expertise to discern.

As a final note, future analysis of the same points (time-series) may be of additional use in terms of tracking trends in the areas we found to matter, such as levels of education achieved by the general population, and the health of the “institution of marriage”. Further, use of this particular model on a future Census dataset can show whether or not there has been a shift in how wealth can be predicted, and if certain social issues still exist.

Bibliography and References

# Bibliography

Friedman, J. H. (2001). 1999 Reitz Lecture Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, Vol. 29, No.5, 1189-1232.

Hastie, T., & Qian, J. (2014, June 26). *Glmnet Vignette*. Retrieved from Stanford: https://web.stanford.edu/~hastie/glmnet/glmnet\_alpha.html

Kutner, M., Nachtsheim, C., & Neter, J. (2004). *Applied Linear Regression Models - 4th Edition.* McGraw-Hill Education.

Mahoney, C. R. (1998). Creating a Culture of Life.

Peng, C.-Y. J., Lee, K. L., & Ingersoll, M. G. (2002). *An Introduction to Logistic Regression.* Retrieved from Indiana University-Bloomington: http://www-psychology.concordia.ca/fac/kline/734/peng.pdf

*Resident Population Estimates of the United States by Sex, Race, and Hispanic Origin*. (2001, January 2). Retrieved from U.S. Census Bureau: https://www.census.gov/population/estimates/nation/intfile3-1.txt

Srivastava, T. (2015, September 11). *Learn Gradient Boosting Algorithm for better predictions (with codes in R)*. Retrieved from Analytics Vidhya: http://www.analyticsvidhya.com/blog/2015/09/complete-guide-boosting-methods/

U.S. Department of Commerce: Bureau of the Census. (2000, April 1). *United States Census 2000.* Retrieved from Census.gov: https://www.census.gov/dmd/www/pdf/d-61b.pdf

UCLA. (2016). *SPSS Data Analysis Examples Logit Regression*. Retrieved from Institute for Digital Research and Education: http://www.ats.ucla.edu/stat/spss/dae/logit.htm

University of California Irvine. (1994). *Census Income Data Set*. Retrieved from UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Census+Income

# Reference

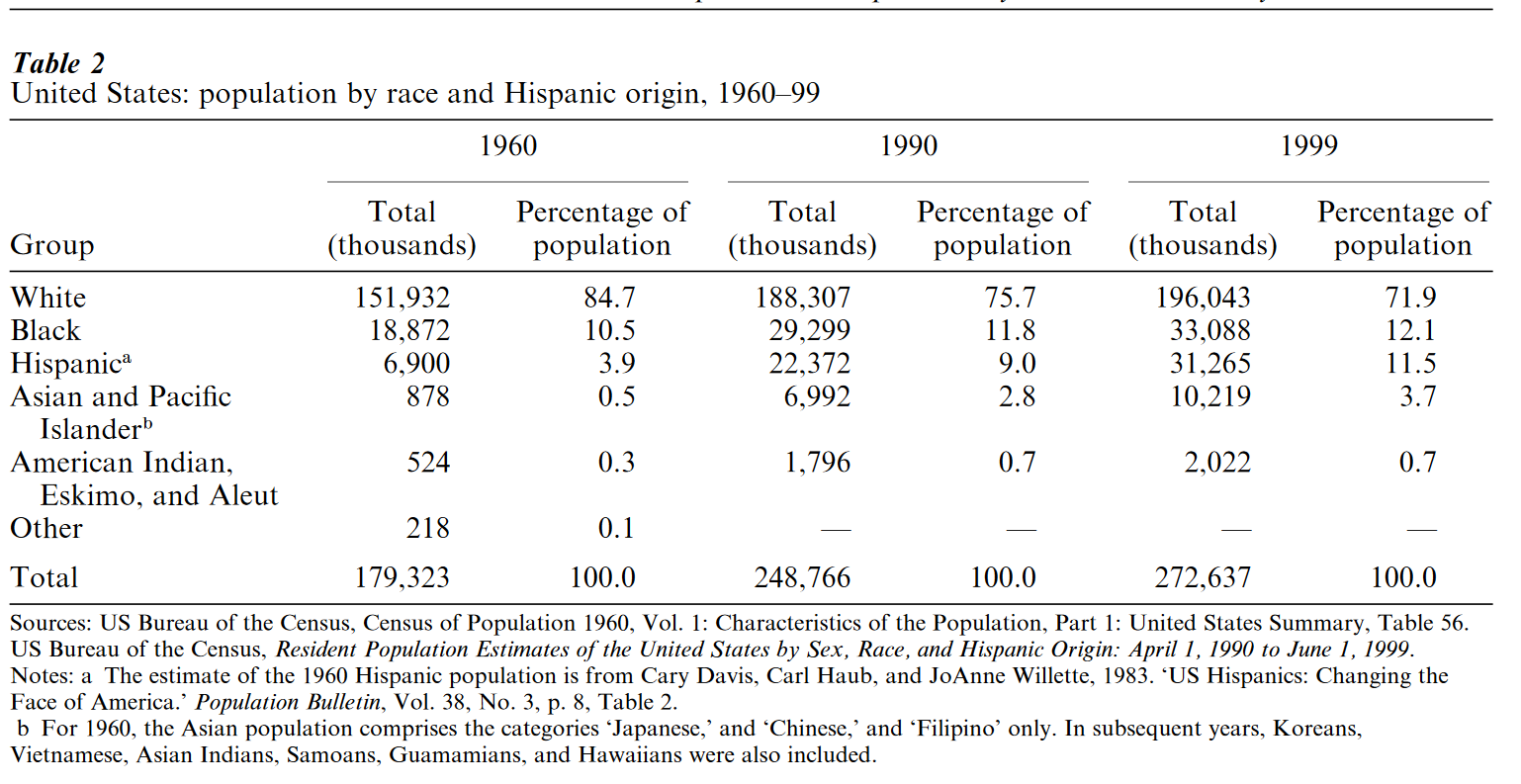
Cover Image of Uncle Sam and the Census book was pulled from <http://backstoryradio.org/files/2010/10/answersready.jpg>. The full image can be found in the government archives at <http://loc.gov/pictures/resource/cph.3g08370/>.

Appendix

**Figure A.1**



**Figure A.2**



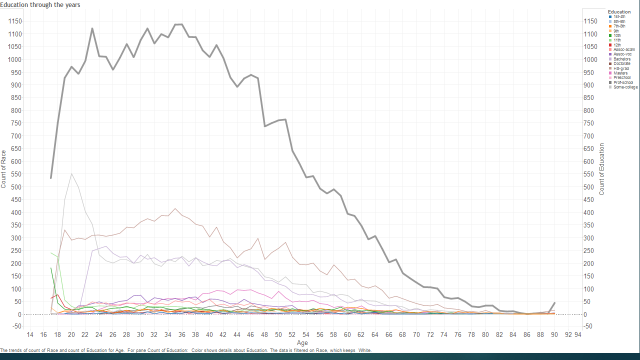
**Figure A.3**

****

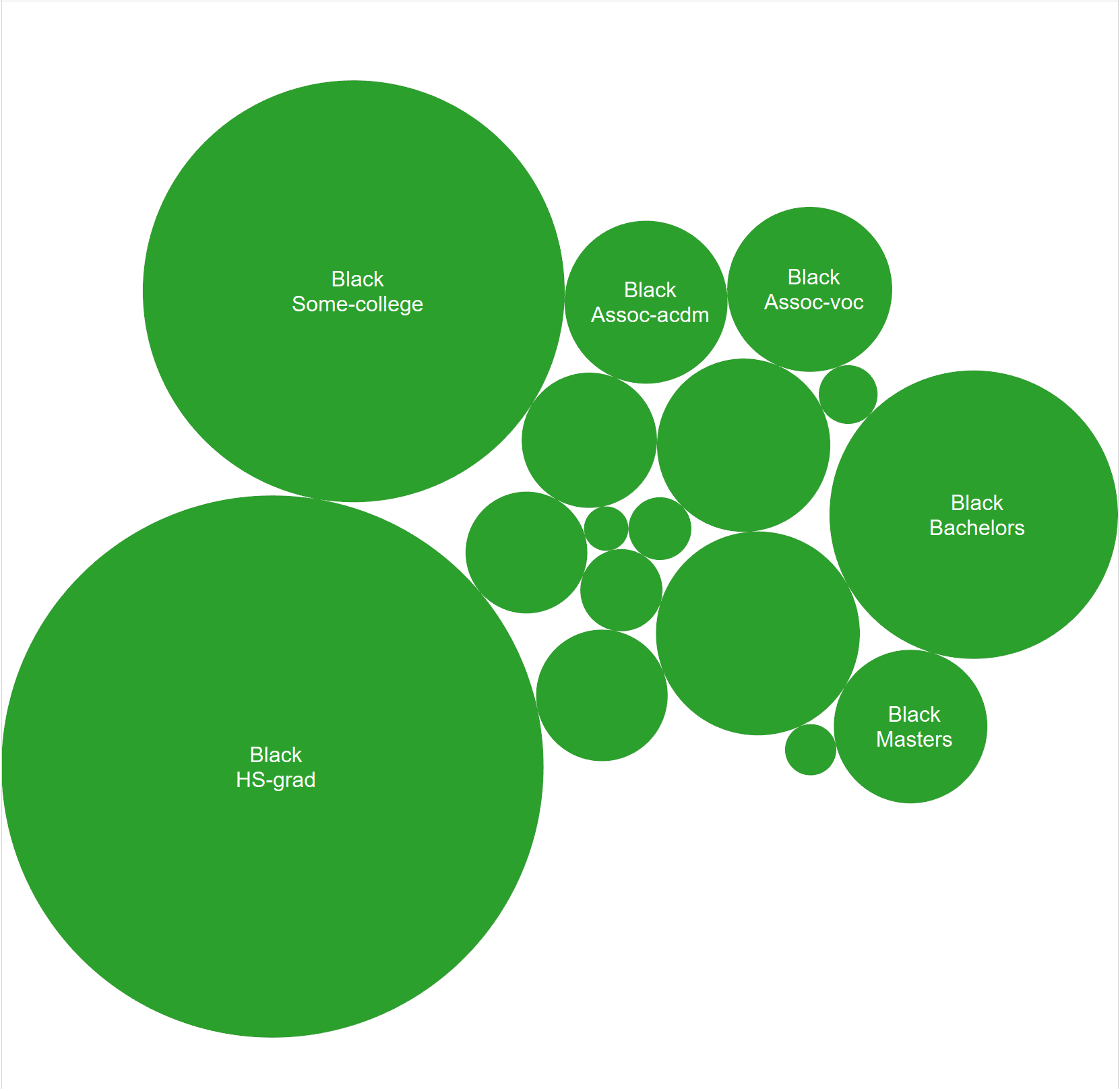
**Figure A.4**



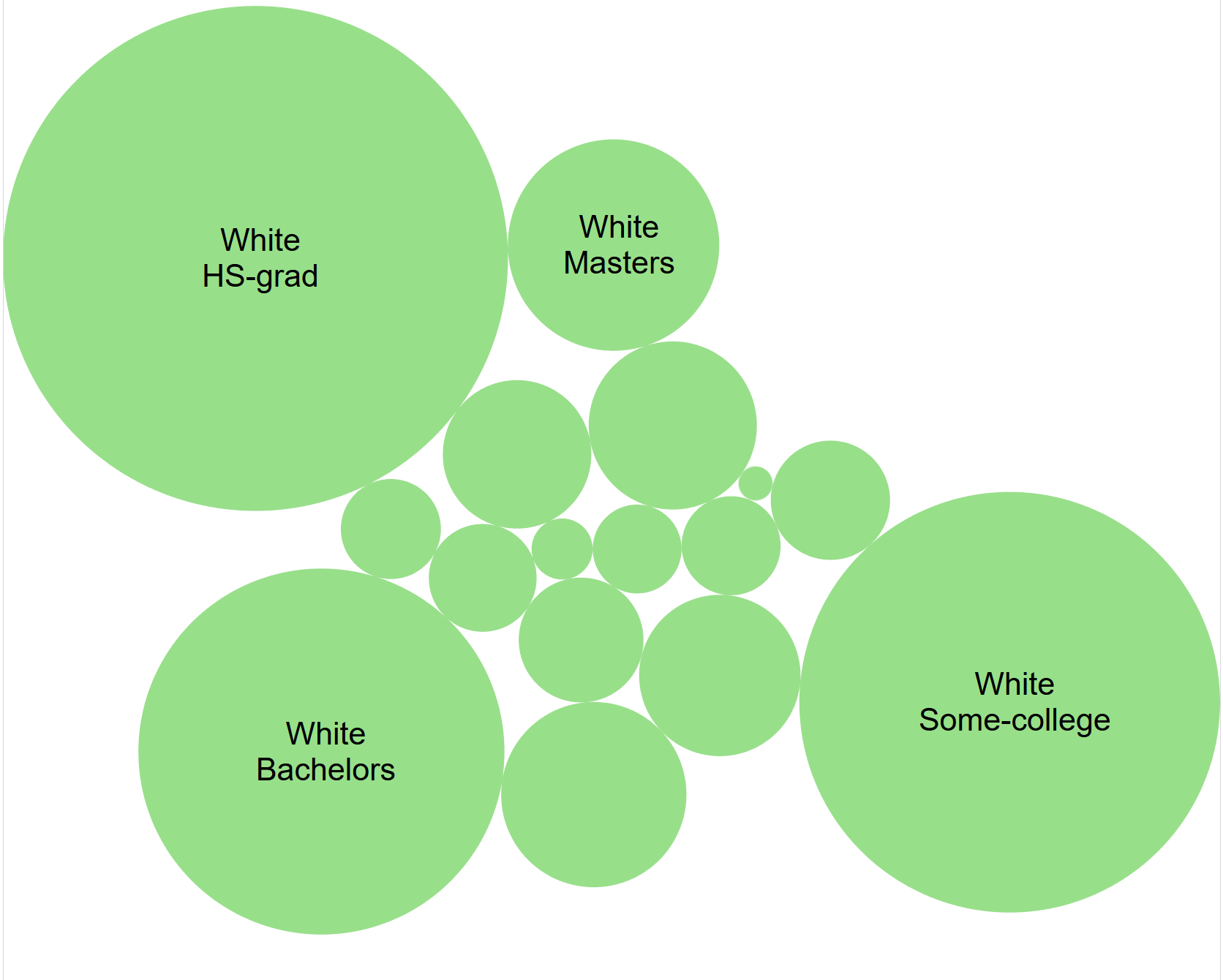
**Figure A.5**

****

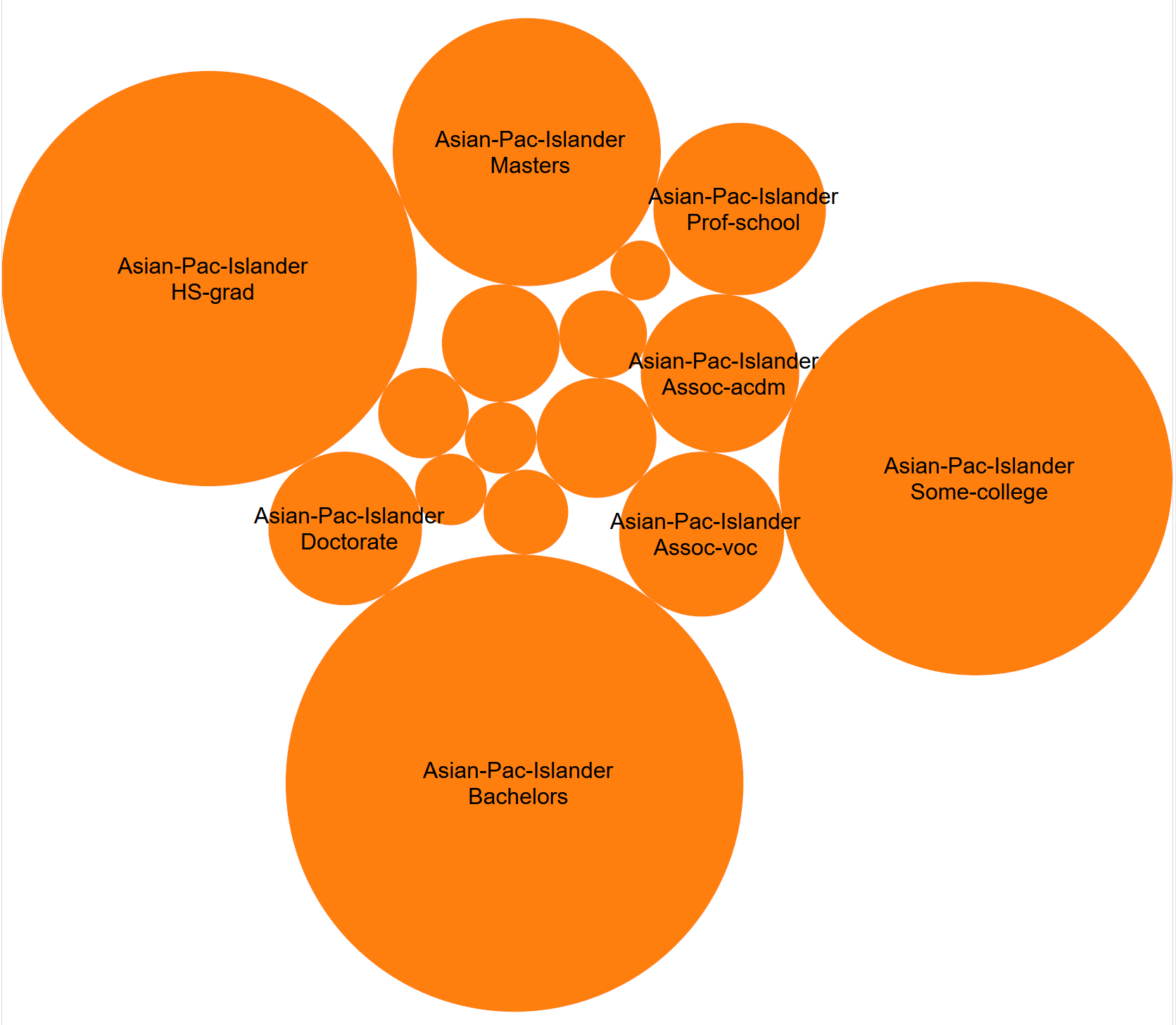
**Figure A.6**

****

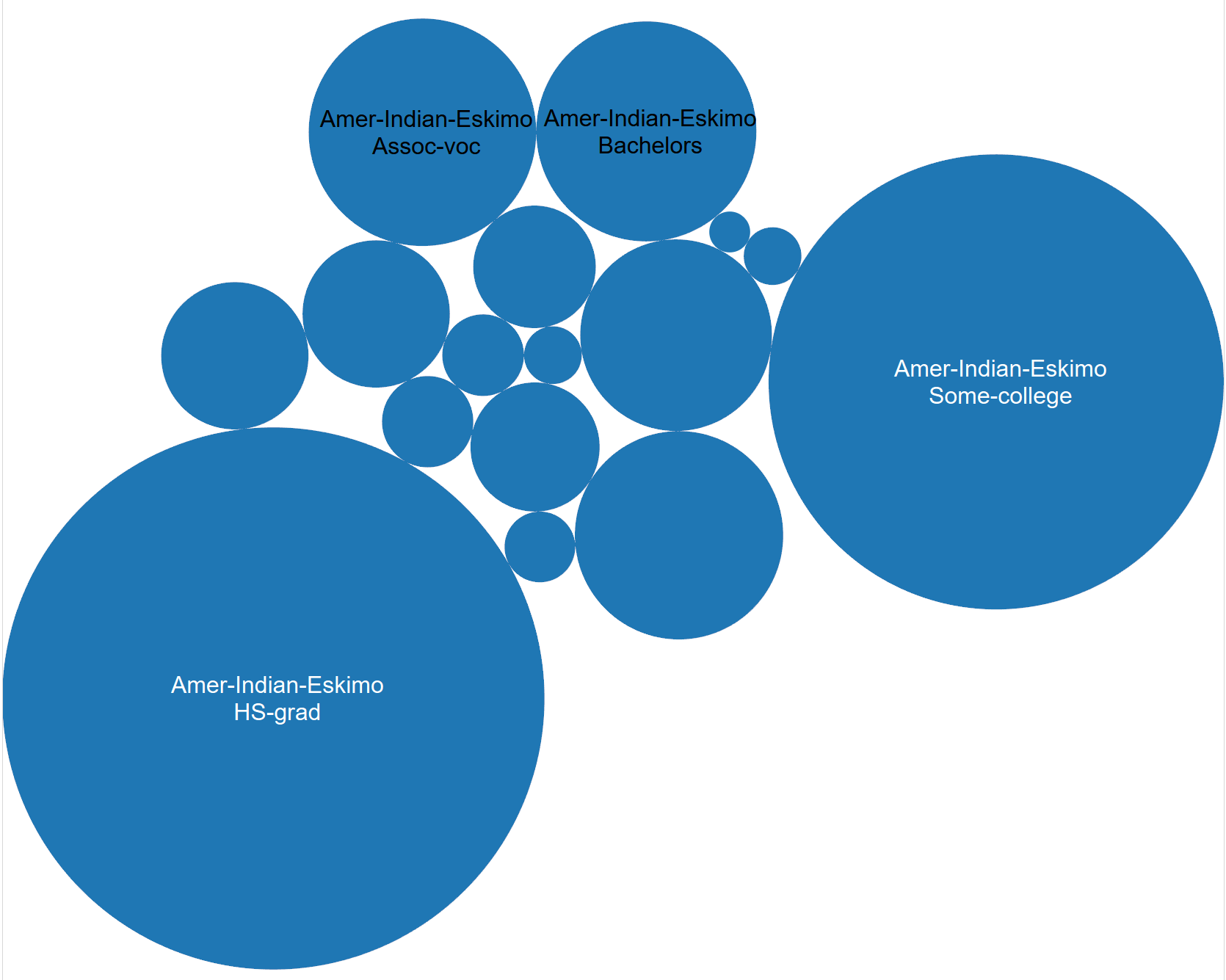
**Figure A.7**

****

**Figure A.8**

****

**Figure A.9**

****

**Figure A.10**

****

**Appendix A.11**

****

**Individual Contributions**

**John Benischeck** – Data visualization section and graphs. Compiled and edited paper.

**Tommy Baw** – GBM Model research, implementation, and writeup.

**Shachi Parikh** – Logistic Regression research, implementation, and writeup.

**Xiaoqin Helen Yi** – GlmNet research, implementation, and writeup.

**Xiaomeng Blair Chen** – data gathering and cleaning, introduction, and writeup.

**Addressing Comments**

We received no questions about our project.

1. The dataset comes from USI Machine Learning website <http://archive.ics.uci.edu/ml/datasets/Census+Income>, which was extracted from the Census Bureau database found at http://www.census.gov/ftp/pub/DES/www/welcome.html. [↑](#footnote-ref-1)
2. Using the category of Race = “White” made visualization of this point easier, because of the large sample size. While other Race categories more or less follow the same trend, their graphs are much more jagged and noisy due to less data. [↑](#footnote-ref-2)
3. The graph shows each category compared to itself. For example, the blue bar categorized as “Asian Female” is the percent of Asian women who make over $50,000 per year. Thus, each category is out of a total possibility of 100%. The black horizontal line is the average of the whole sample, most likely buoyed by White Males. [↑](#footnote-ref-3)
4. **(U.S. Department of Commerce: Bureau of the Census, 2000)** [↑](#footnote-ref-4)
5. Image source(Srivastava, 2015) [↑](#footnote-ref-5)