Predicting the Success of Second-Generation Immigrants in the United States

Linh Tang & Yesheng Chen

STA-230 - Grinnell College

Presentation Overview

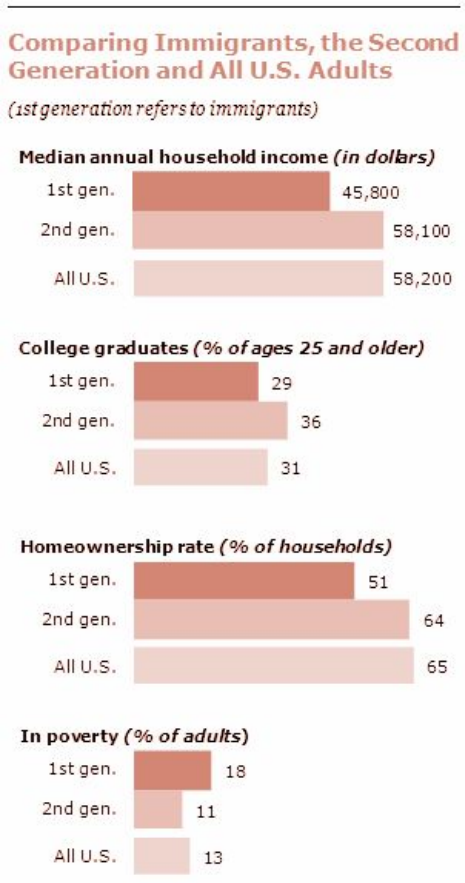
Introduction

Analysis Goals

- examine how to estimate the "success" of second generation immigrants in the United States in the period 2001-2003
- analyze the factors that contribute to the success of second generation immigrants during their adaptation process

Project Idea:

build a model to predict the success index of second-generation immigrants based on family background, education attainment and adaptation process.



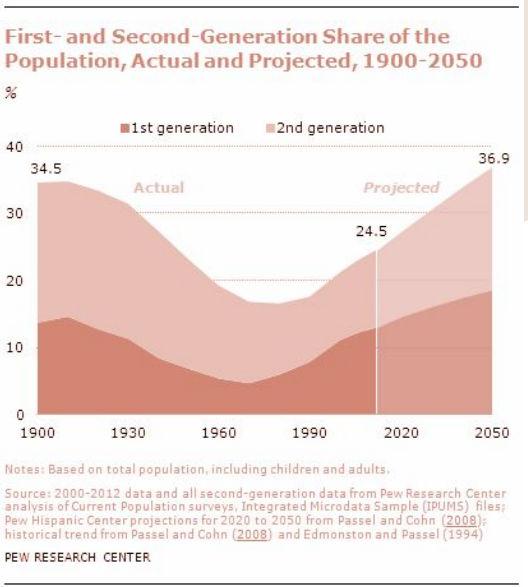
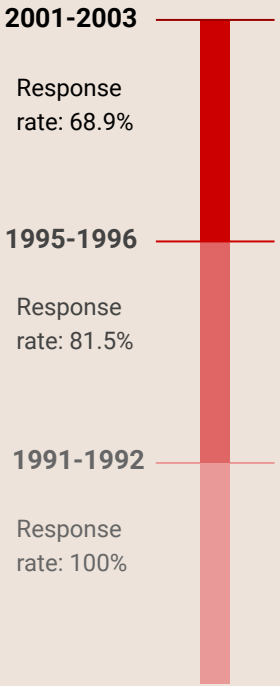


Image source: Pew Research Center
https://www.pewsocialtrends.org/2013/02/07/second-generation-americans/

About the Dataset

Children of Immigrants Longitudinal Study

- Sample size: **5262** secondary school students from 77 (original) nationalities
- Variables (questions): **665**
- Length of study: 15 years, 1991-2006, from **early adolescence to early adulthood**
- Where: metropolitan areas of Miami/Ft. Lauderdale in Florida and San Diego, California
- Focus of survey: family background, schooling situation, living condition, education achievement and expectation, adaptation process, self evaluation, etc.



Final Follow-up

3613 respondents at the beginning of their work careers at the average age of 24. The survey focus on outcome of adaptation process: education, socioeconomic status, income, etc…

First follow-up

4288 students prior to high school graduation at the average age of 17. The survey focus on examine the changes of the factors in previous survey over adolescence years.

Initial survey

5262 students in their junior high school at the average age of 14, focus on families, demographic characteristics, language use, self-identities, and academic attainment of the attendees

How was the data collected ?

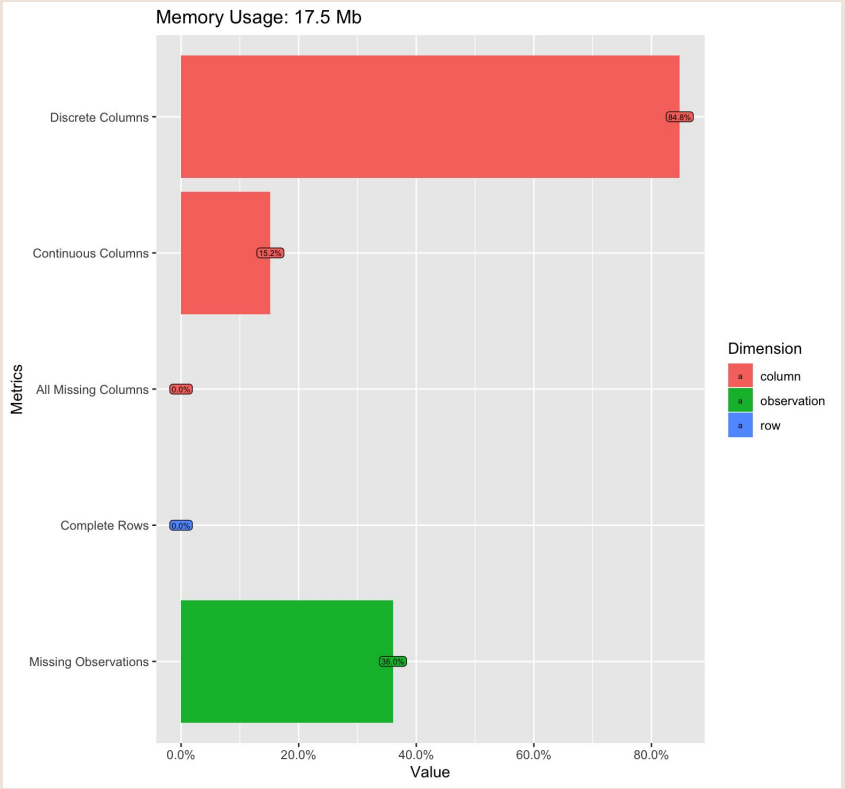
Methodology

01 Exploring data (EDA): Tidy meta dataset based on how data was collected, examine and select potentially effective predictors (subjectively), visualize correlations between variables and detect possible patterns

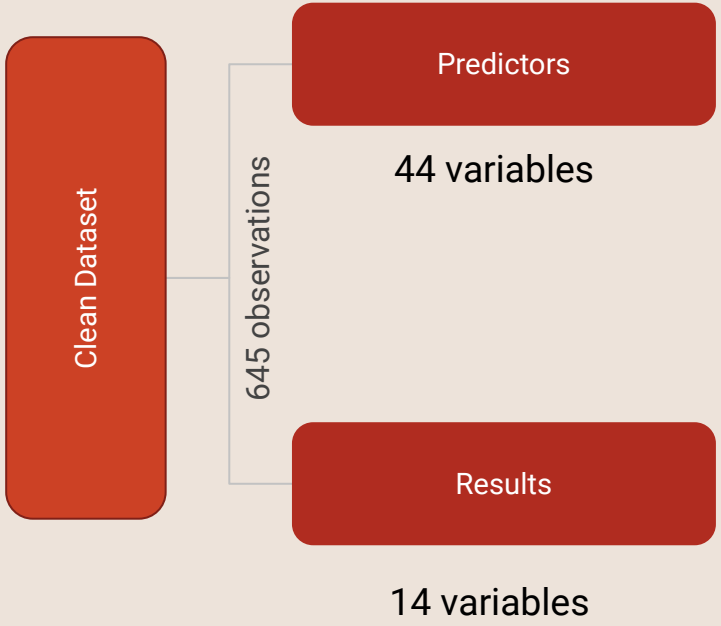
02 Constructing outcome: research published papers on success measurement, develop a formula namely “success index” to measure success based on research of Populace, in partnership with Gallup.

03 Building model: conduct Principal Component Analysis to derive important variables, run different models (LM, LASSO, GAM, PCR) and perform out-of-sample cross validation to choose best predictive model (LASSO)

01 Data Exploration

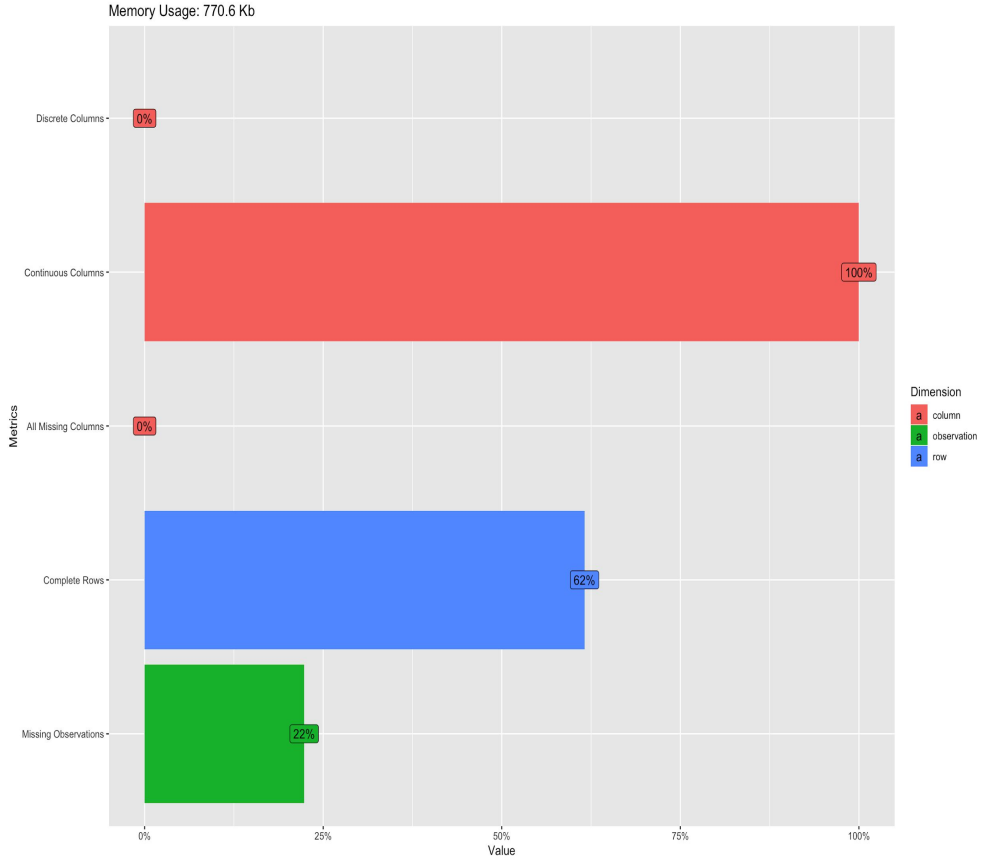


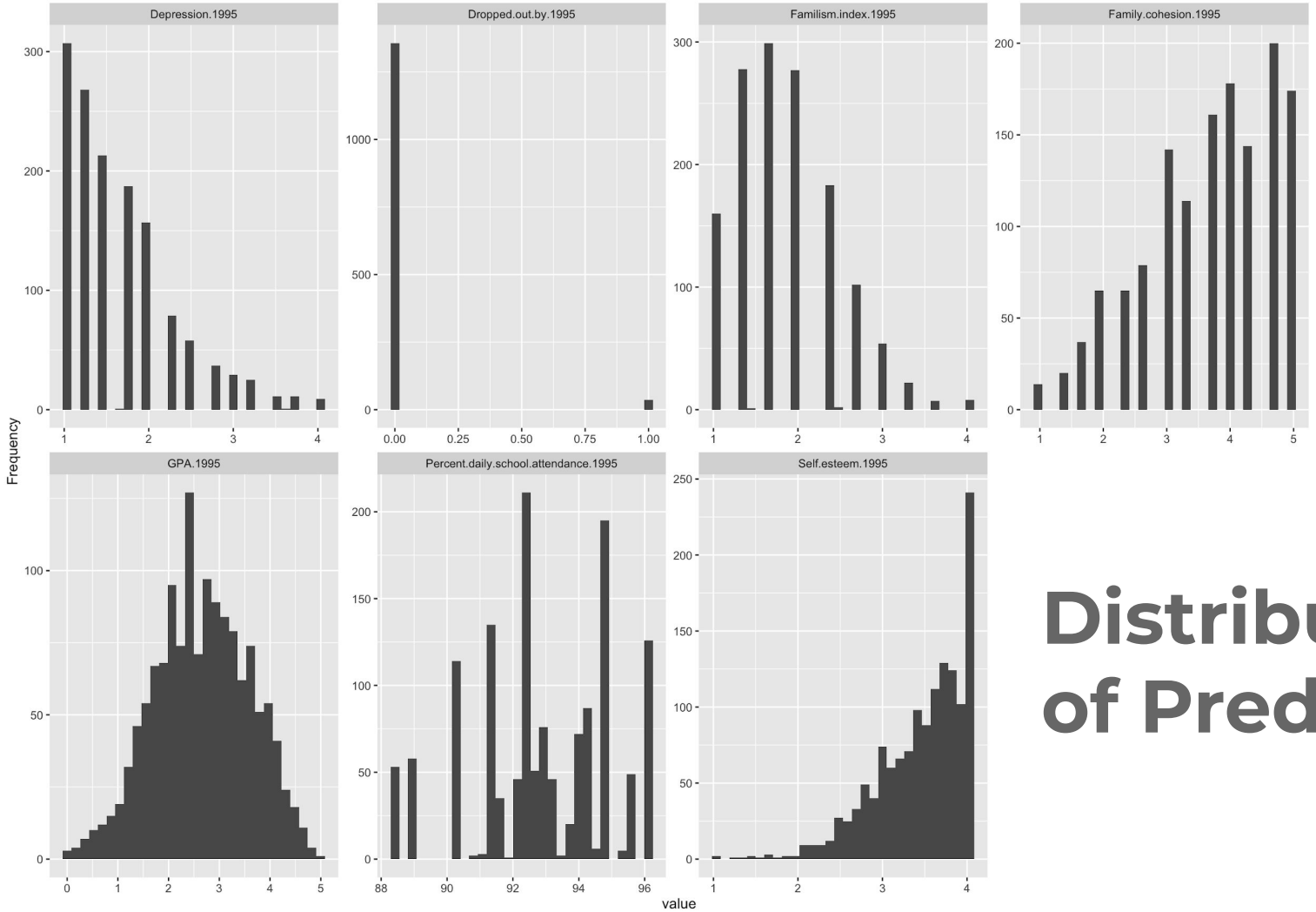
- ★ 36% of the meta dataset is incomplete
- ★ After tidying the dataset, we proceed with analysis on



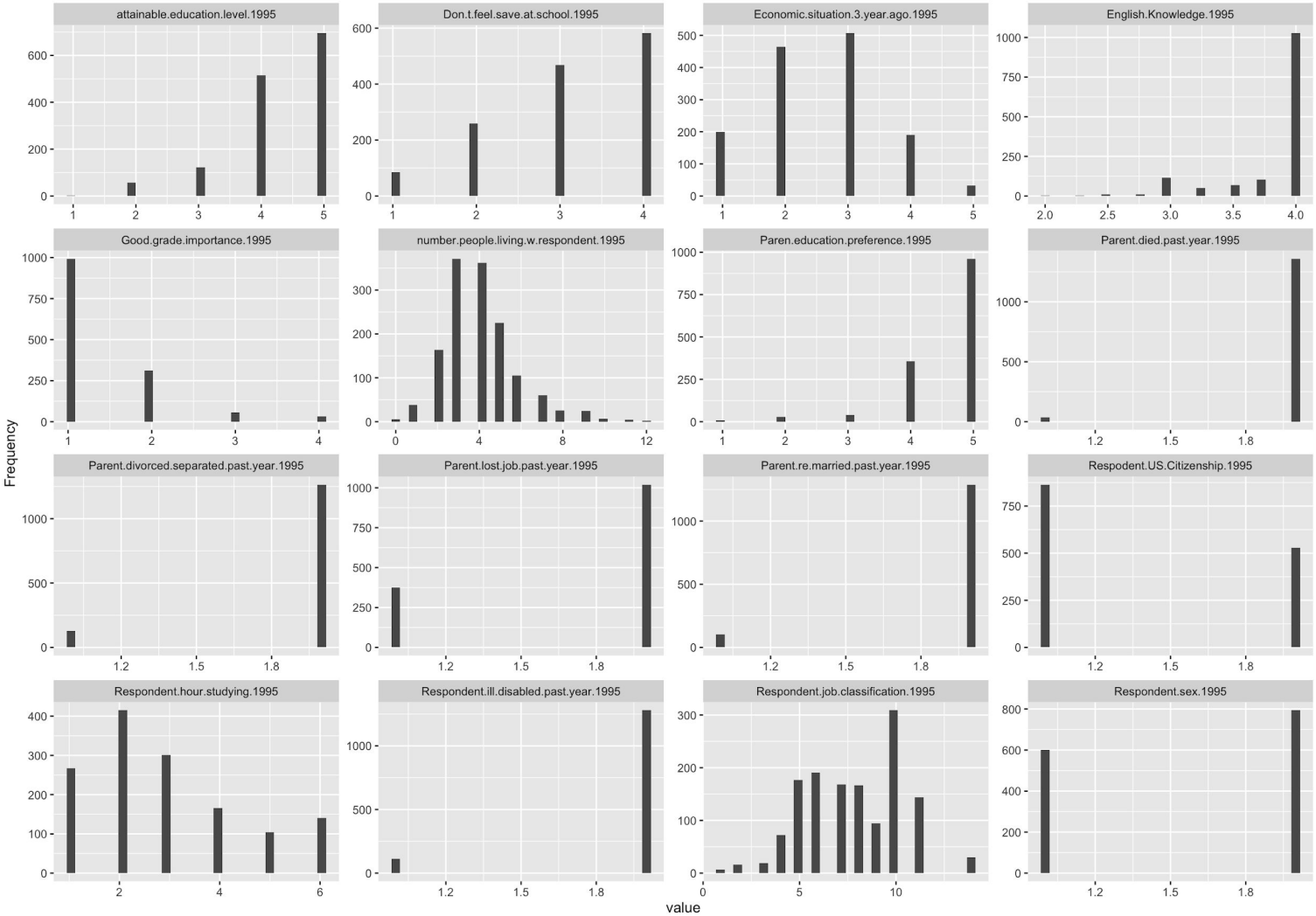
Summary of Predictors dataset

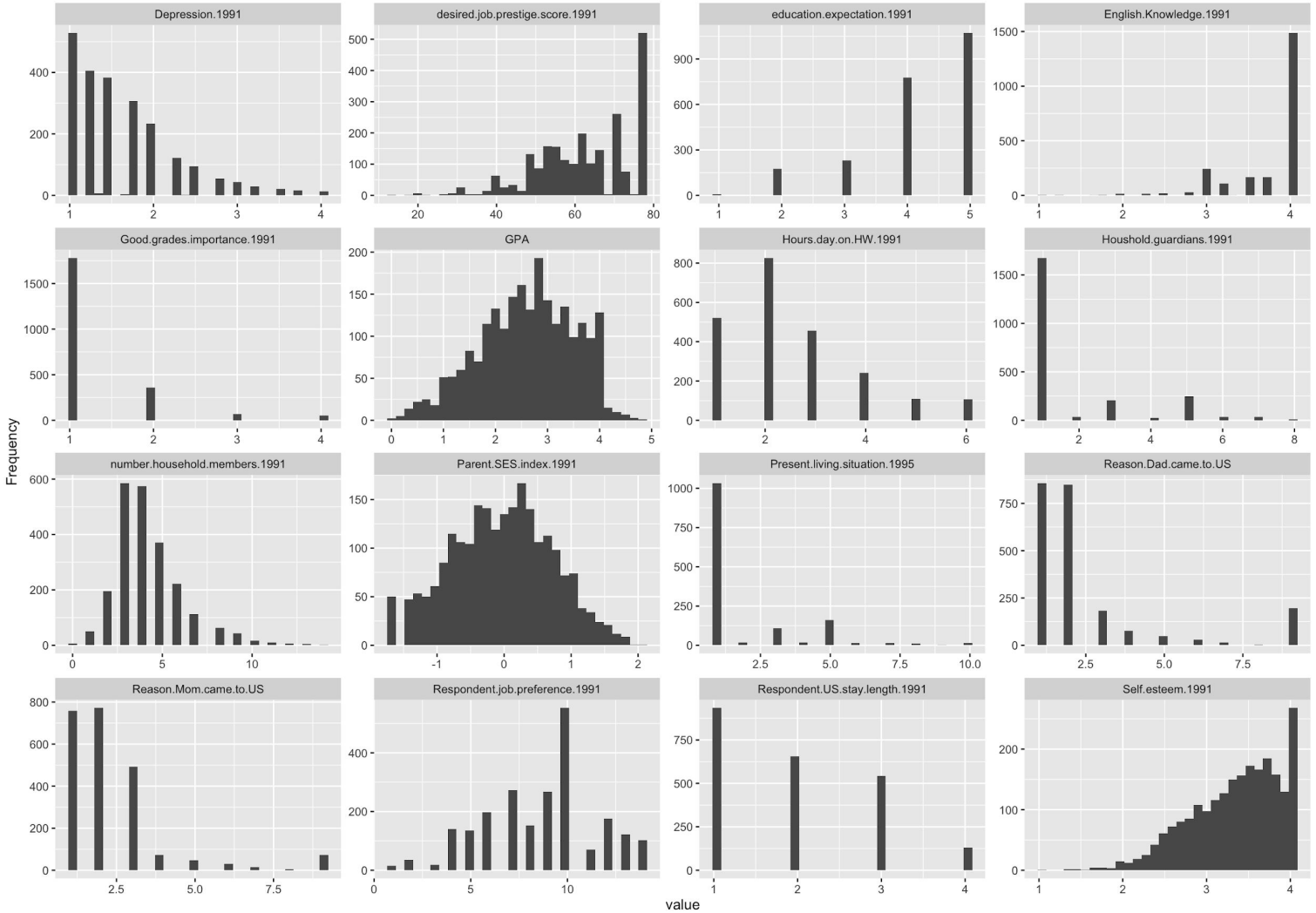
	mean	median	var	sd	valid.n
desired.job.prestige.score-1991	62.66	62.00	154.27	12.42	2262
GPA	2.60	2.67	0.83	0.91	2262
Parent.SES.index-1991	-0.03	-0.01	0.57	0.75	2262
English.Knowledge-1991	3.74	4.00	0.20	0.45	2262
Private.school-1991	0.05	0.00	0.05	0.23	2262
Houshold.guardians-1991	1.86	1.00	2.57	1.60	2262
number.household.members-1991	4.33	4.00	3.53	1.88	2262
Sex	1.51	2.00	0.25	0.50	2262
Respondent.US.stay.length-1991	1.94	2.00	0.88	0.94	2262
Respondent job preference-1991	8.65	9.00	8.31	2.88	2262
felt discriminated-1991	1.44	1.00	0.25	0.50	2262
Depression-1991	1.64	1.50	0.39	0.62	2262
Self-esteem-1991	3.33	3.40	0.26	0.51	2262
education expectation-1991	4.21	4.00	0.87	0.93	2262
Hours/day on HW-1991	2.52	2.00	1.78	1.34	2262
Good grades importance-1991	1.29	1.00	0.40	0.64	2262
Reason Dad came to US	2.54	2.00	5.21	2.28	2262
Reason Mom came to US	2.33	2.00	2.78	1.67	2262
Present living situation-1995	1.94	1.00	3.35	1.83	1393
number people living w/respondent-1995	4.07	4.00	3.03	1.74	1393
Economic situation/3 year ago-1995	2.56	3.00	0.94	0.97	1393
Parent divorced/separated past year-1995	1.91	2.00	0.08	0.29	1393
Parent re/married past year-1995	1.93	2.00	0.07	0.26	1393
Parent lost job/past year-1995	1.73	2.00	0.20	0.44	1393
Respondent ill/disabled past year-1995	1.92	2.00	0.07	0.27	1393
Parent died past year-1995	1.97	2.00	0.02	0.16	1393
Respondent sex-1995	1.57	2.00	0.25	0.50	1393
Respodent US Citizenship-1995	1.38	1.00	0.24	0.49	1393
Respondent job classification-1995	7.80	8.00	6.26	2.50	1393
Don't feel save at school-1995	3.11	3.00	0.83	0.91	1393
attainable education level-1995	4.32	4.00	0.66	0.81	1393
Paren education preference-1995	4.60	5.00	0.49	0.70	1393
Respondent hour studying-1995	2.89	3.00	2.38	1.54	1393
Good grade importance-1995	1.37	1.00	0.45	0.67	1393
English Knowledge-1995	3.82	4.00	0.13	0.36	1393
Depression-1995	1.66	1.50	0.41	0.64	1393



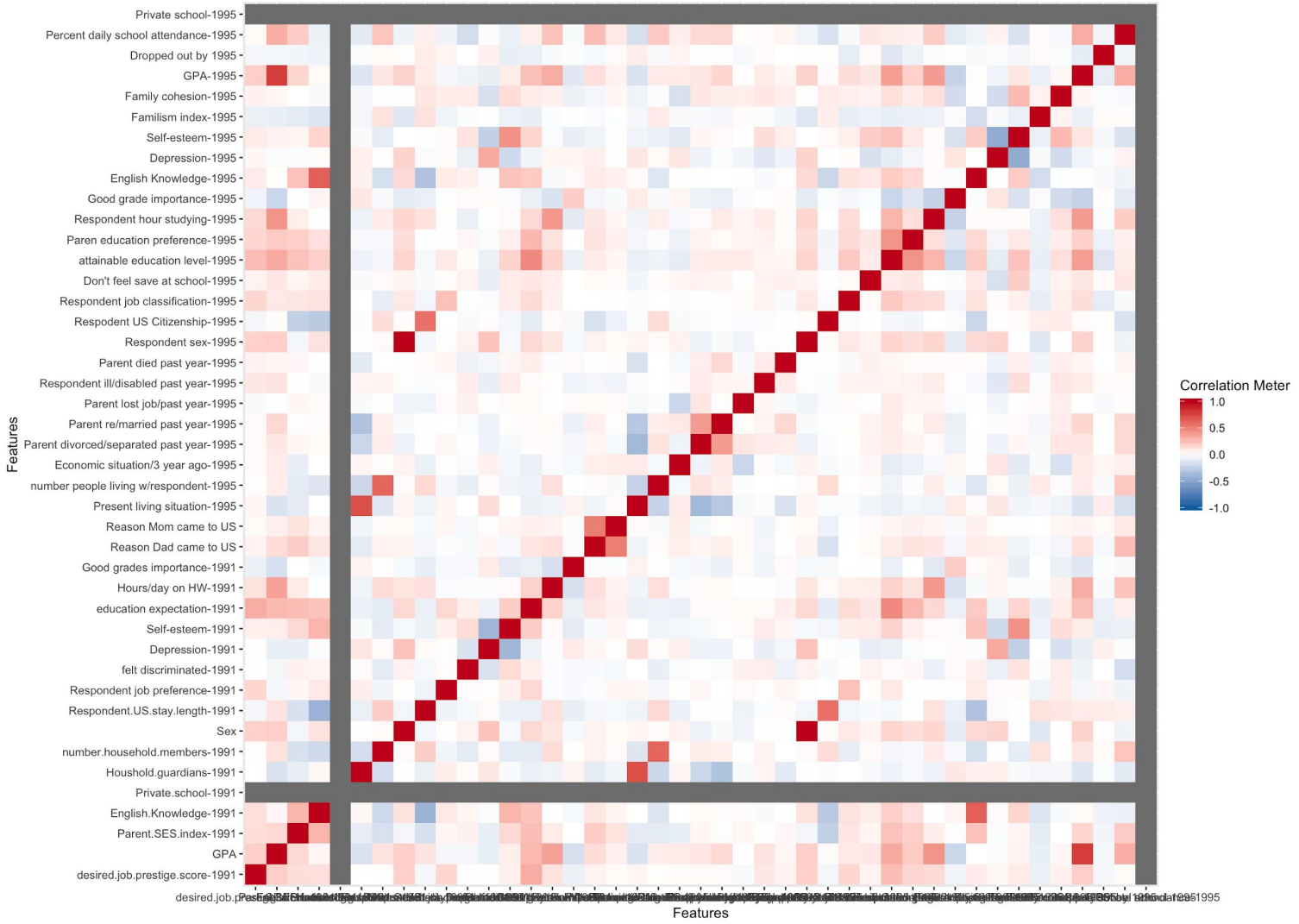


Distribution of Predictors


attainable.education.level.1995
Don.t.feel.save.at.school.1995
Economic.situation.3.year.ago.1995
English.Knowledge.1995
Good.grade.importance.1995
number.people.living.w.respondent.1995
Paren.education.preference.1995
Parent.died.past.year.1995
Parent.divorced.separated.past.year.1995
Parent.lost.job.past.year.1995
Parent.re.married.past.year.1995
Respodent.US.Citizenship.1995
Respondent.hour.studying.1995
Respondent.ill.disabled.past.year.1995
Respondent.job.classification.1995
Respondent.sex.1995
Frequency
value


Depression.1991
desired.job.prestige.score.1991
education.expectation.1991
English.Knowledge.1991
Good.grades.importance.1991
GPA
Hours.day.on.HW.1991
Houshold.guardians.1991
number.household.members.1991
Parent.SES.index.1991
Present.living.situation.1995
Reason.Dad.came.to.US
Reason.Mom.came.to.US
Respondent.job.preference.1991
Respondent.US.stay.length.1991
Self.esteem.1991
Frequency
value

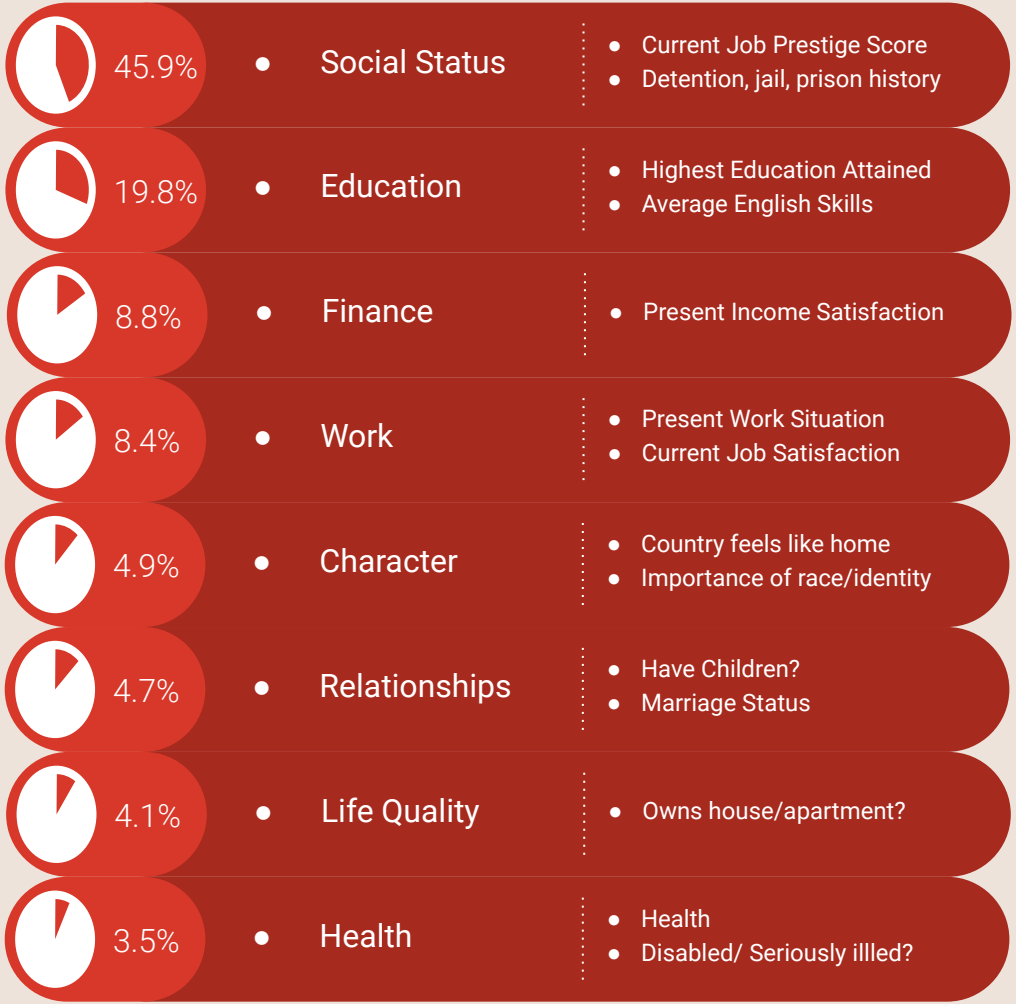
Pairwise correlation between potential predictors



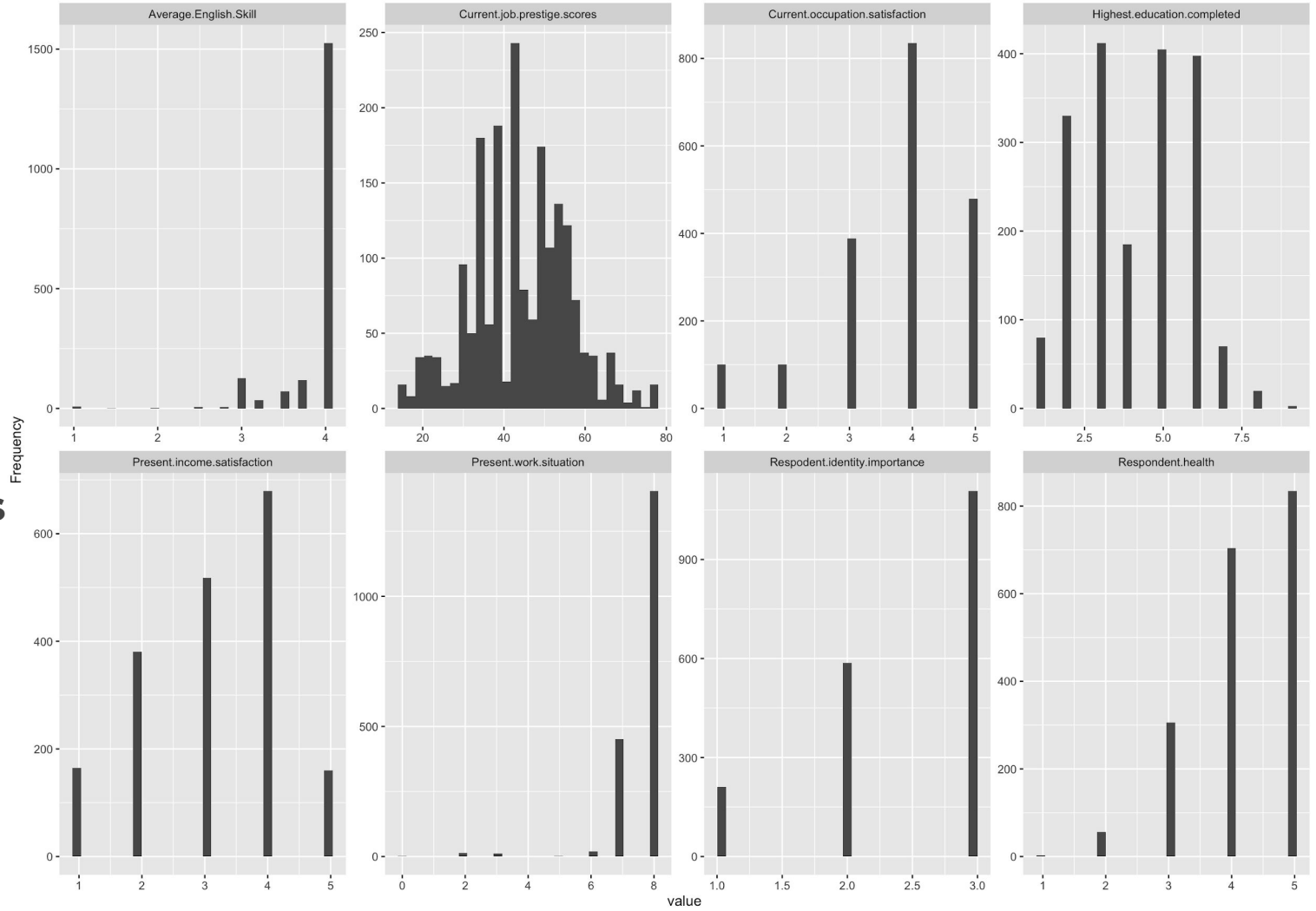
02

Success Index

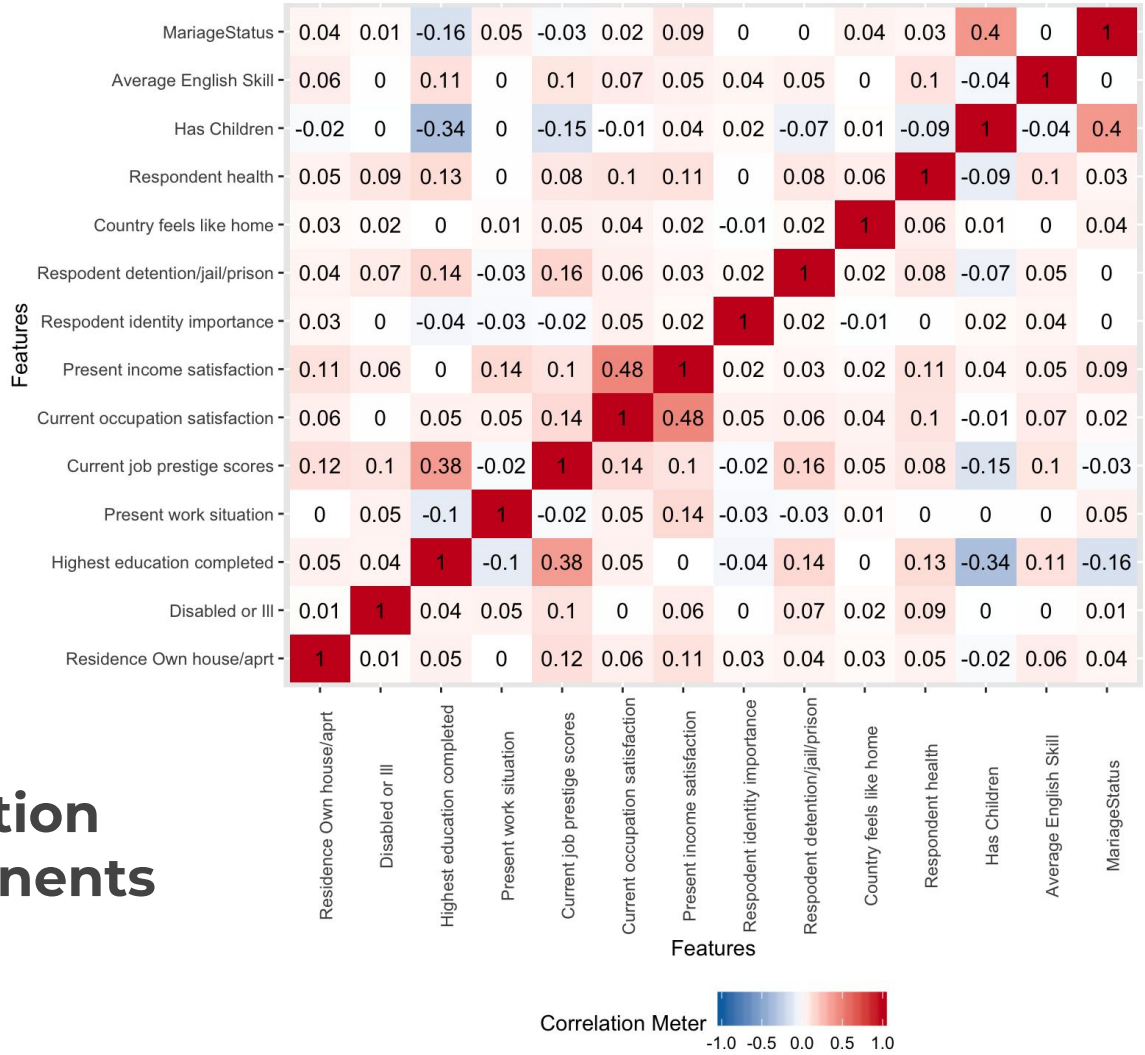
- ★ A study conducted by Populace in July 2019, with a random sample of 5,242 U.S. adults age 18 and older, who are members of Gallup Panel, through web survey using the choice-based conjoint approach. Each attribute is measured based on their perception of how society defines success.
- ★ We categorized the variables in last survey into different attributes and assigned them weights according to the Populace's study to calculate the success index for each observation.



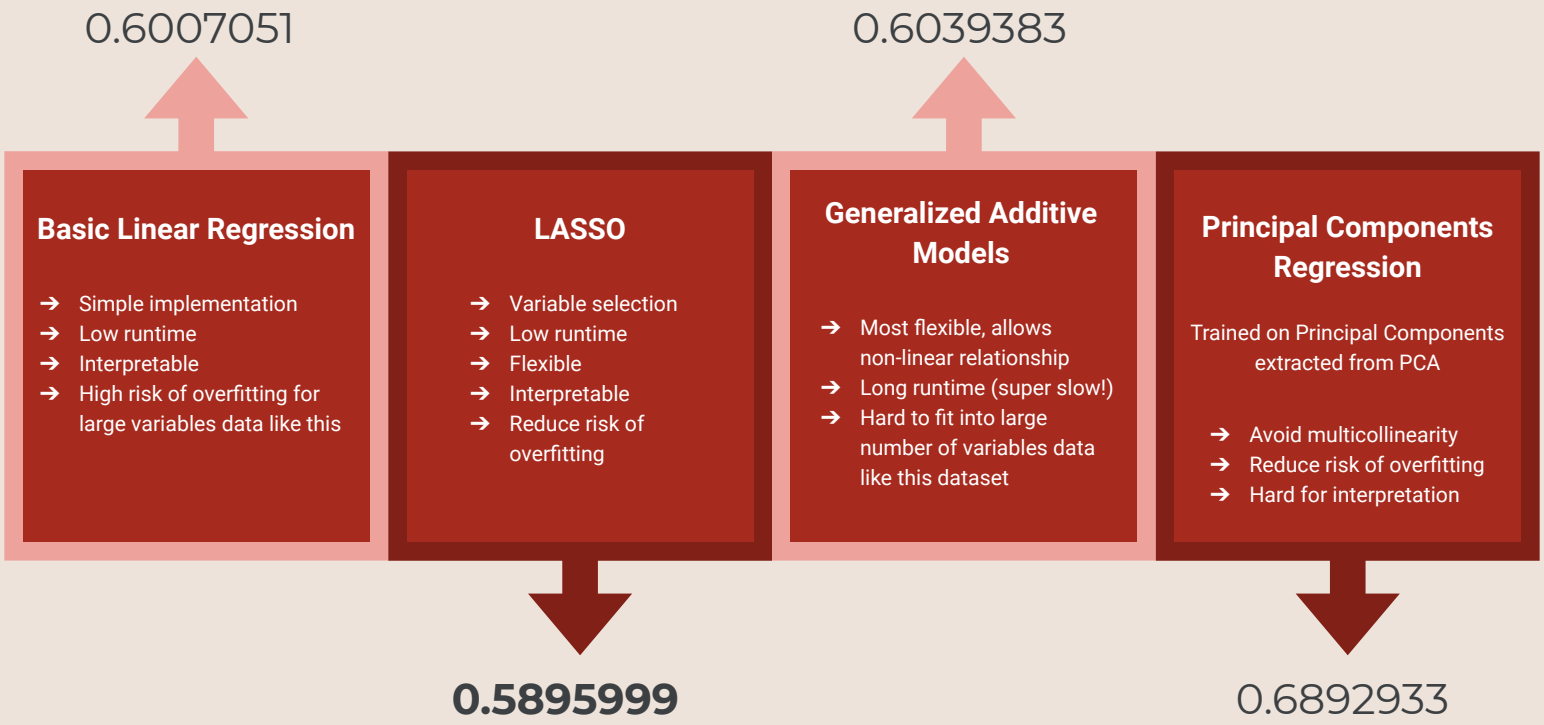
Distribution of components of success index



Pairwise correlation between components of success index

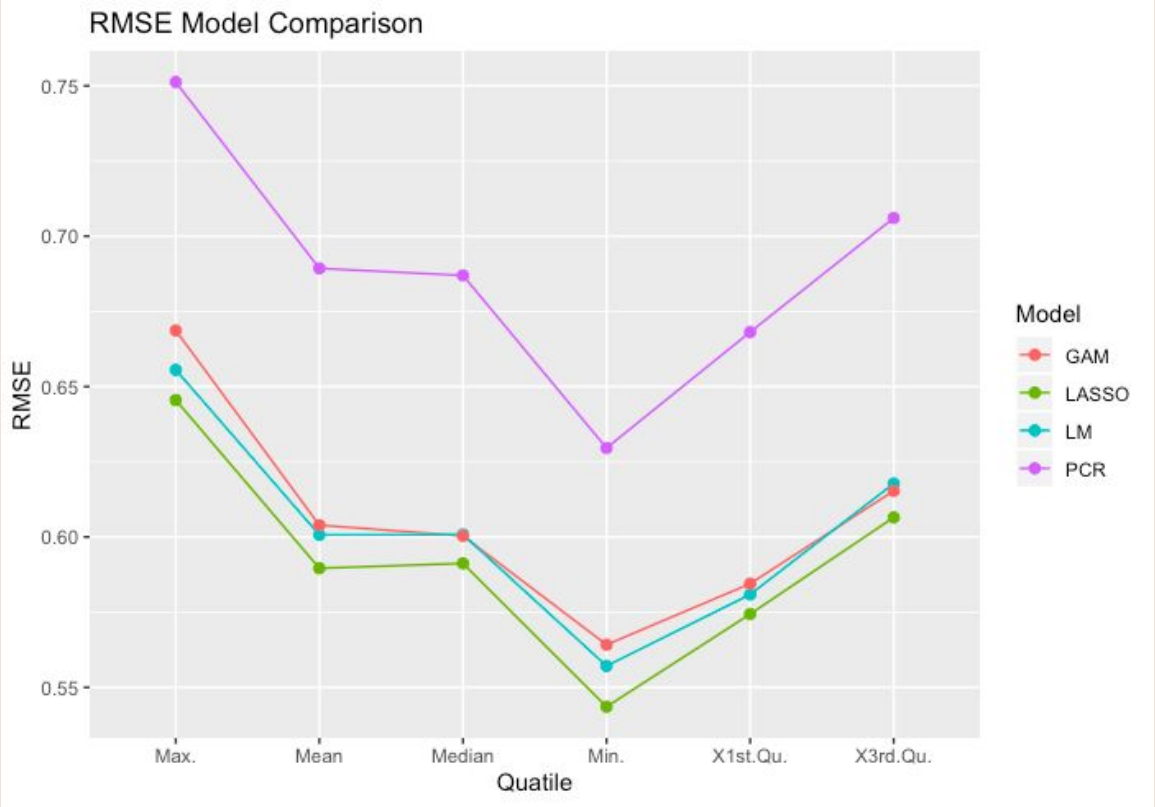


03 Model Building: comparing RMSE



Performance Evaluation

- ★ Out-of-sample cross-validation (10 repeats, 5 fold) on four models
  - LASSO stands out with lowest RMSE
- ★ Evaluation Metric: RMSE and MAE
  - best metric for numeric outcomes prediction



Results & Discussions

- ★ Final model: **LASSO**
- ★ Variables to train on: **24** out of 44 variables (and 665 original variables)
  -> model is efficient

44 x 1 sparse Matrix of class "dgCMatrix"	1
(Intercept)	3.322773e+00
desired.job.prestige.score.1991	9.400377e-04
GPA	9.969484e-02
Parent.SES.index.1991	7.669943e-02
English.Knowledge.1991	6.926964e-02
Private.school.1991	.
Houshold.guardians.1991	-2.997830e-03
number.household.members.1991	-2.857505e-02
Sex	2.901105e-02
Respondent.US.stay.length.1991	.
Respondent.job.preference.1991	.
felt.discriminated.1991	.
Depression.1991	.
Self.esteem.1991	1.181811e-02
education.expectation.1991	7.722124e-02
Hours.day.on.HW.1991	.
Good.grades.importance.1991	.
Reason.Dad.came.to.US	.
Reason.Mom.came.to.US	.
Present.living.situation.1995	.
number.people.living.w.respondent.1995	.
Economic.situation.3.year.ago.1995	.
Parent.divorced.separated.past.year.1995	.
Parent.re.married.past.year.1995	.
Parent.lost.job.past.year.1995	-2.778579e-04
Respondent.ill.disabled.past.year.1995	1.467390e-01
Parent.died.past.year.1995	2.294185e-01
Respondent.sex.1995	4.944564e-05
Respodent.US.Citizenship.1995	5.078880e-03
Respondent.job.classification.1995	6.880424e-03
Don.t.feel.save.at.school.1995	.
attainable.education.level.1995	9.896779e-02
Paren.education.preference.1995	1.860510e-02
Respondent.hour.studying.1995	-7.730633e-03
Good.grade.importance.1995	.
English.Knowledge.1995	1.898666e-01
Depression.1995	-7.680899e-02
Self.esteem.1995	2.085149e-02
Familism.index.1995	.
Family.cohesion.1995	2.519425e-02
GPA.1995	1.856483e-01
Dropped.out.by.1995	.
Percent.daily.school.attendance.1995	-3.282556e-02
Private.school.1995	.

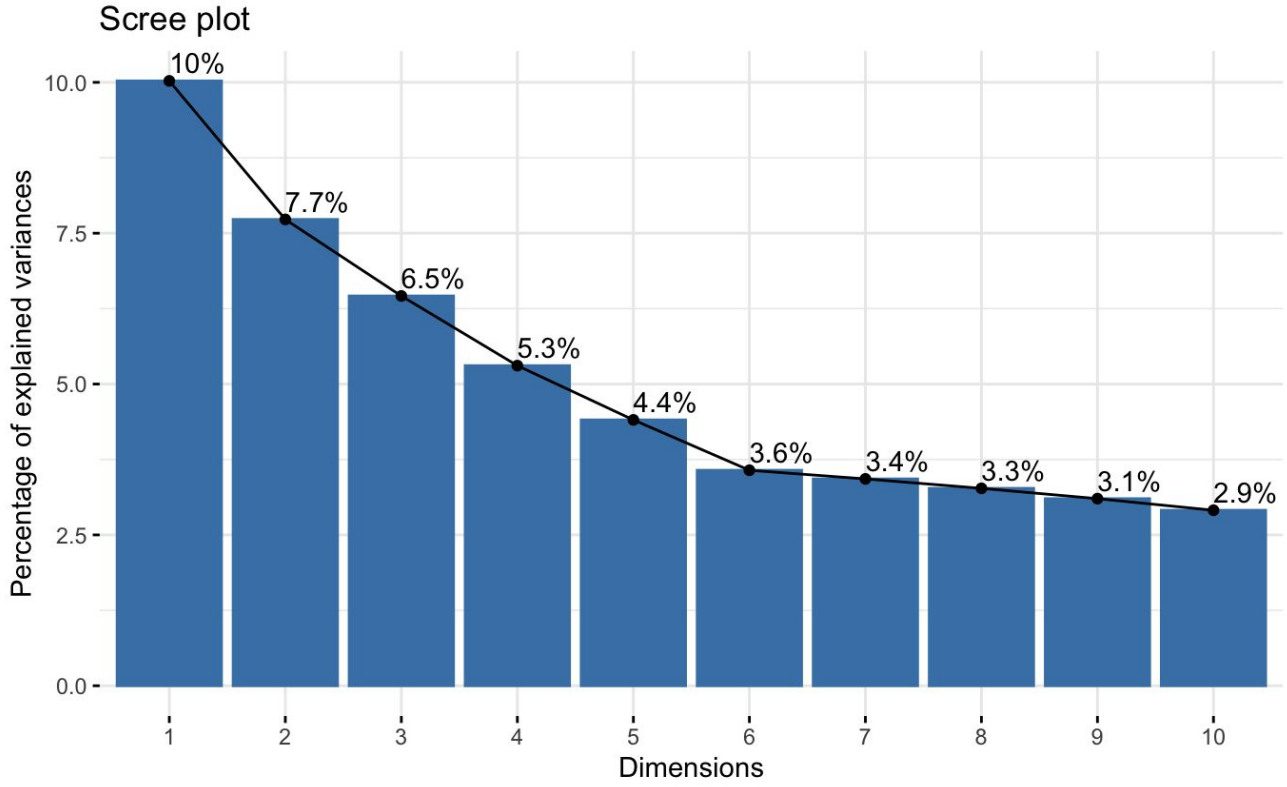
- ★ Mostly **ruled-out** variables related to **family**
- ★ Highly weighted variables are collected in second survey
  -> development during **high school** time matters most
- ★ **Female** students are more likely to have higher success index
- ★ **School Attendance** and **Studying Time** slightly **negatively** impact success **!!!**

Limitations & Future Work

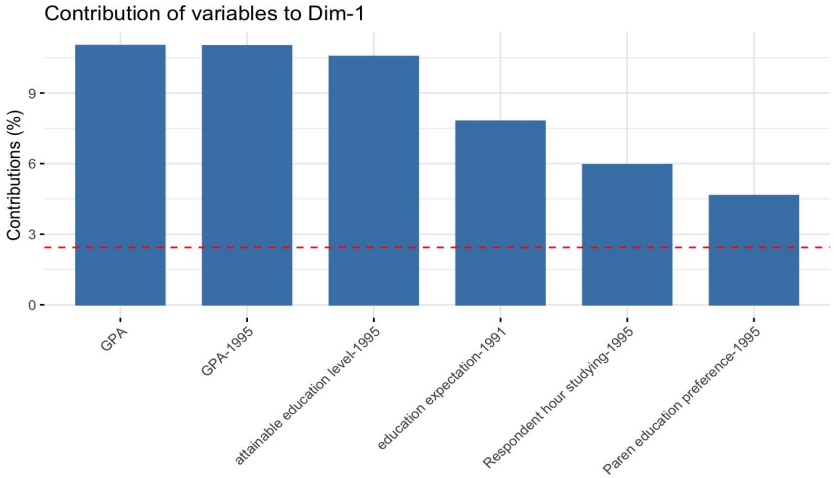
- ★ The dataset is too noisy and not up-to-date
- ★ Time Constraint
- ★ Subjective formula for success index

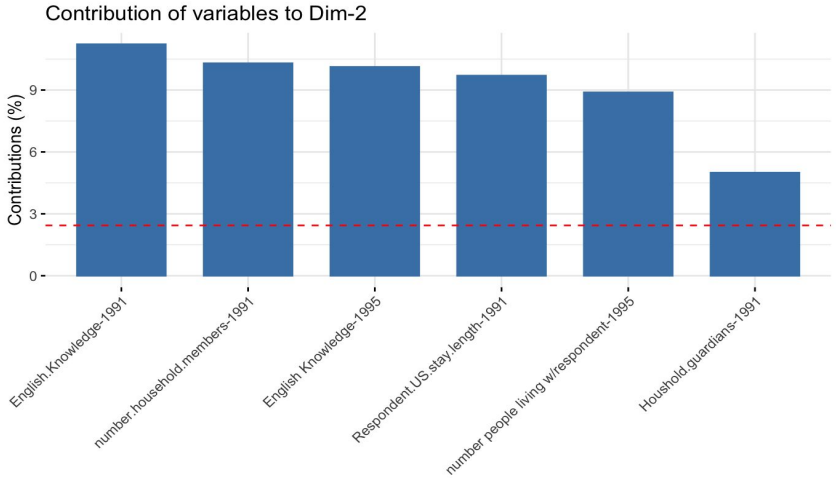
- ➔ Find better, tidier datasets for training
- ➔ Conduct pilot study
- ➔ Conduct survey to test hypotheses based on our results and discussion
- ➔ Use different evaluation metrics
- ➔ Compare with the average of all Americans
- ➔ Compare between different races/ ethnicities.

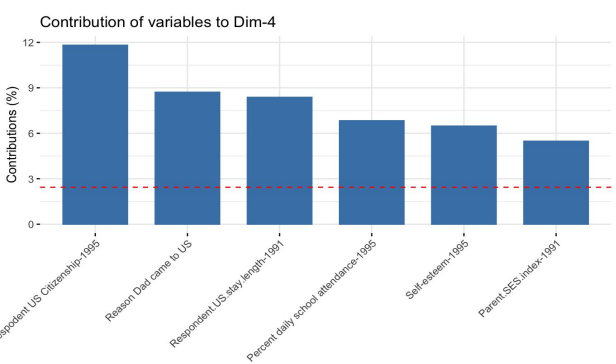
PCA on Predictors

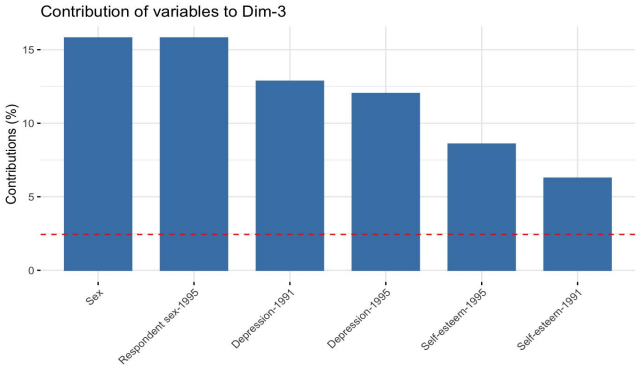


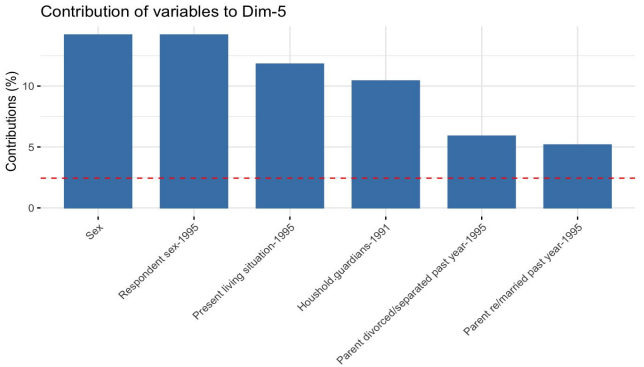
Contributions of Variables on PCs



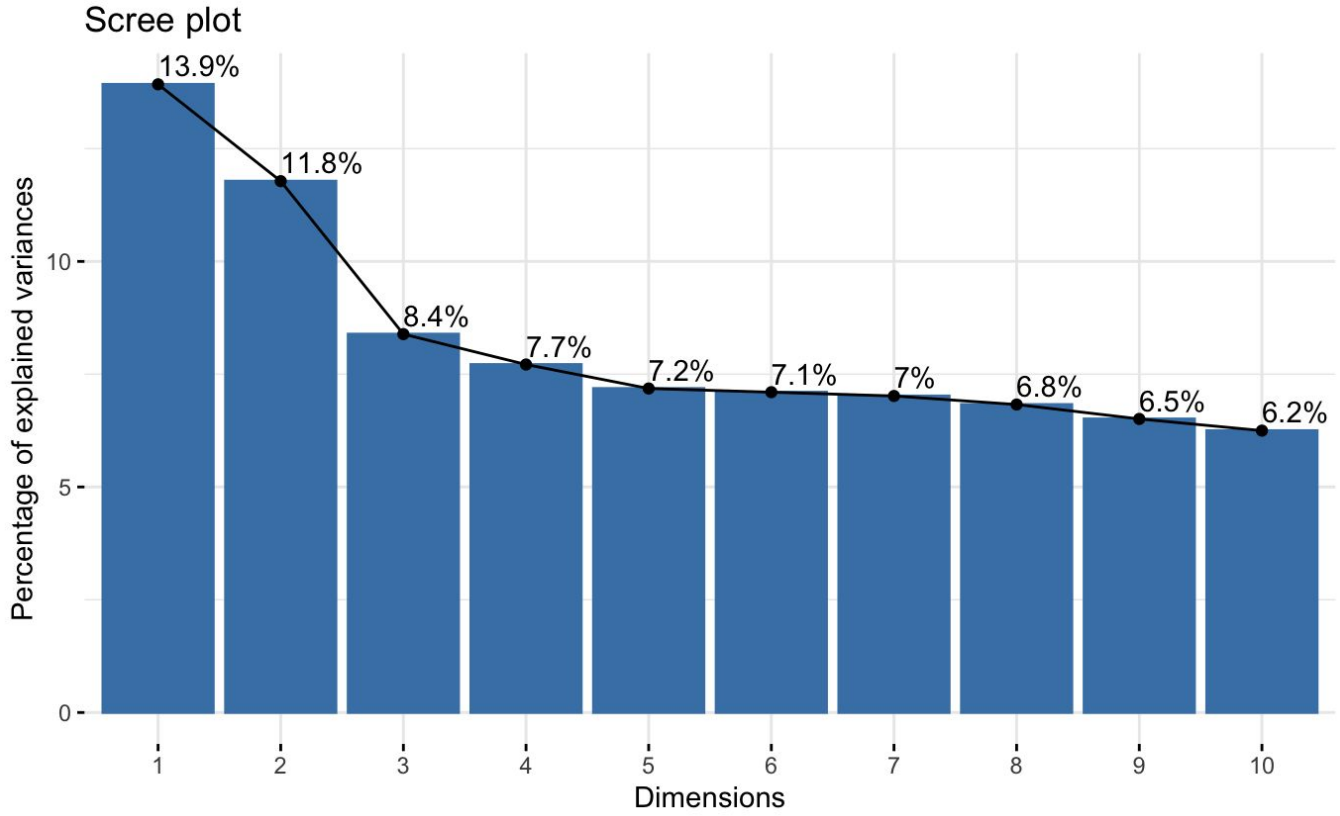




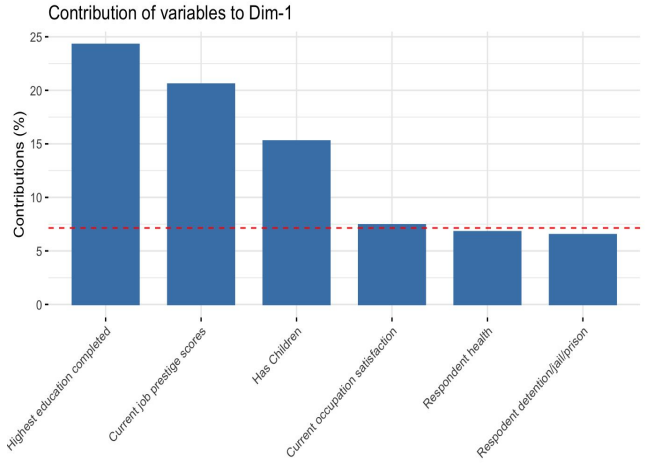


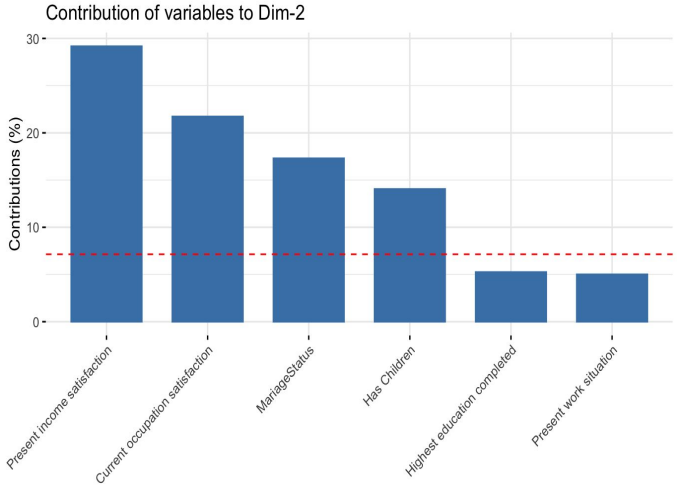


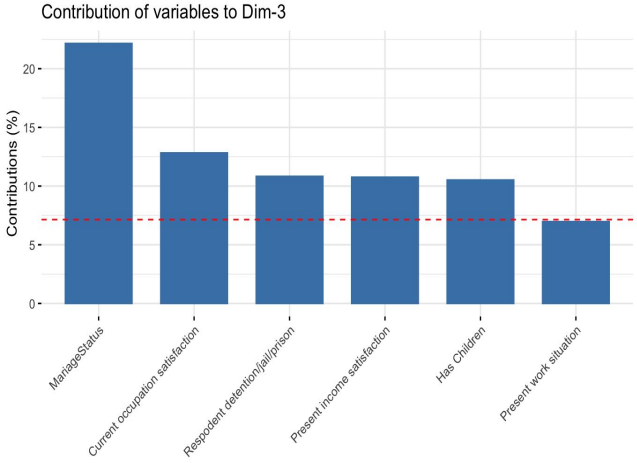
PCA on Results

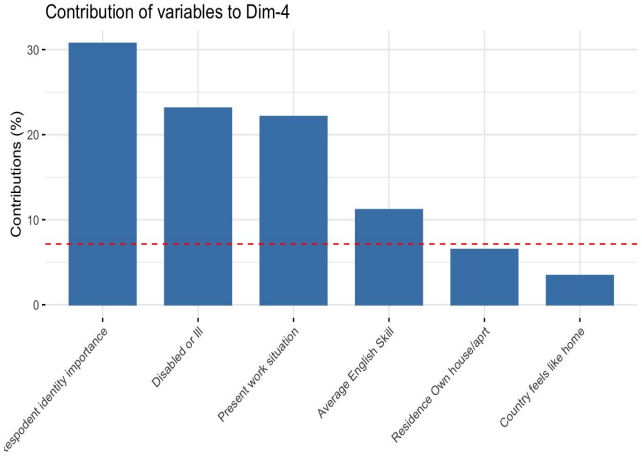


Contributions of Variables on PCs









References:

- Portes, Alejandro, and Rumbaut, Rubén G. Children of Immigrants Longitudinal Study (CILS), San Diego, California, Ft. Lauderdale and Miami, Florida, 1991-2006. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2018-12-12. https://doi.org/10.3886/ICPSR20520.v3

- https://www.pewsocialtrends.org/2013/02/07/second-generation-americans/

- https://static1.squarespace.com/static/59153bc0e6f2e109b2a85cbc/t/5d939cc86670c5214abe4b50/1569955251457/Populace+Success+Index.pdf

- https://www.r-bloggers.com/performing-principal-components-regression-pcr-in-r/