

Modeling Social Determinants of Health

Analysis of Health and Retirement Study

Yiwen Shi, Data Science Intern at Analytiq Inc.





Summary

In this project, I explored potential relationships between different socioeconomic factors and health outcomes. My general hypothesis was that family background and financial status can have a significant impact on occurrences of health problems. By exploring a large health survey dataset and performing different types of statistical testing and modeling, I concluded that socioeconomic background does have a significant impact on health and some factors are more significant than the others.



Background

Health is extremely important for everyone, while certain health conditions are very good indicators for major diseases such as diabetes and cancer, it is also interesting and essential to explore social and economic factors that could potentially indicate the likelihood of a certain health problem or a particular disease. In my research project, I used KNIME Analytics Platform to build predictive models for different health outcomes, with the first two being modeling BMI and diabetes.



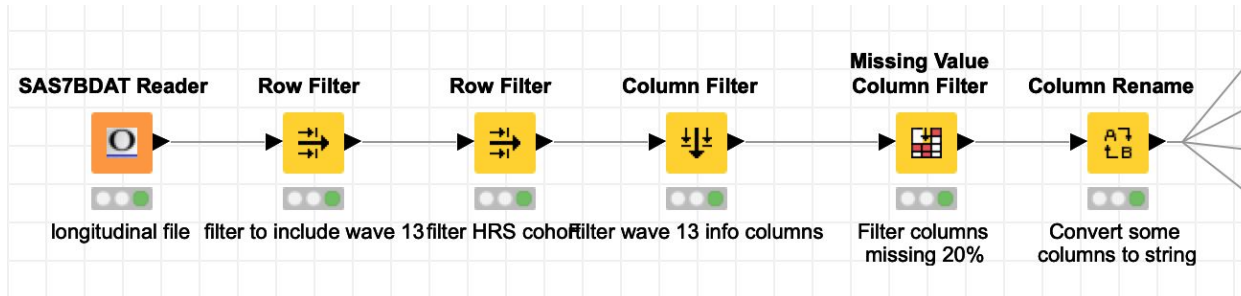
Source of Data

- Health and Retirement Study Longitudinal File 2016
- <https://www.rand.org/well-being/social-and-behavioral-policy/centers/aging/dataproducts/hrs-data.html>
- Biennial interviews since 1992
- Wave 13 (2016)
- Cohort HRS (born from 1931 to 1941)
- SAS version of the file read into KNIME



Data Prep Using KNIME

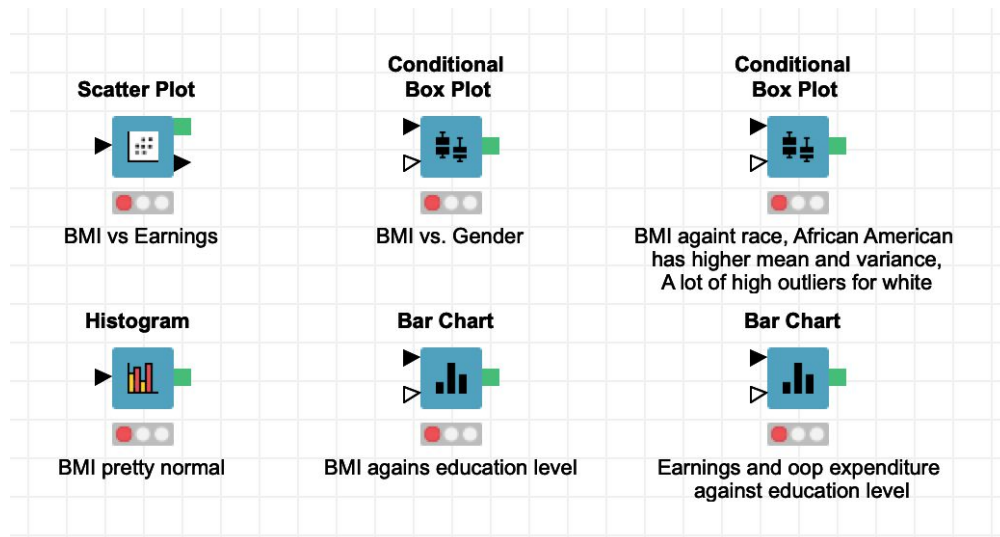
- SAS7BDAT Reader Node:
- Row Filters
- Column Filters
- Missing Value Filters
- Column Rename





Data Visualization: BMI

- BMI distribution: visualize using a histogram node, roughly normally distributed.
- BMI against groups: Race (RARACEM), Education (RAEDUC), Gender (RAGENDER)
- Conditional Boxplots, grouped bar charts and scatter plot were used.



Data Visualization: BMI (Cont.)

Below was the Conditional Boxplot Node configuration window, group variable was selected as the category column and BMI was the selected column (y-axis).

Category Column

☒ Manual Selection ☐ Wildcard/Regex Selection

Left Pane (Red Border):

- Filter
- ☒ HHIDPN
- ☒ S13HHIDPN
- ☒ R13MPART
- ☒ INW13
- ☒ RASPID1
- ☒ RASPCT
- ☒ R13MRCT
- ☒ R13MLN
- ☒ R13MLNM
- ☒ R13MDIV
- ☒ R13MWID

☐ Enforce exclusion

Right Pane (Green Border):

- Filter
- ☒ R13BMI

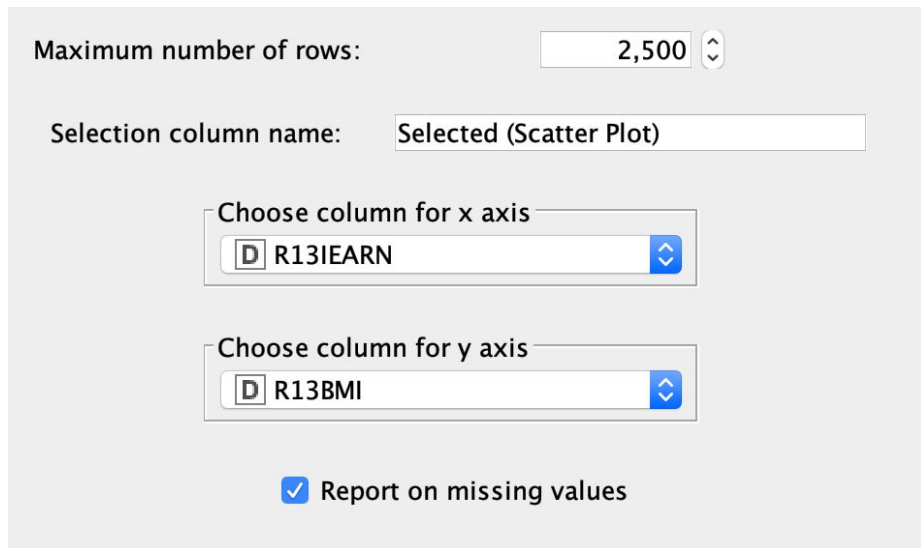
☒ Enforce inclusion

Selected Column



Data Visualization: BMI (Cont.)

Below was the Scatter Plot Node configuration window, independent variable was individual earning and BMI was the dependent variable (y-axis).



Maximum number of rows: 2,500

Selection column name: Selected (Scatter Plot)

Choose column for x axis

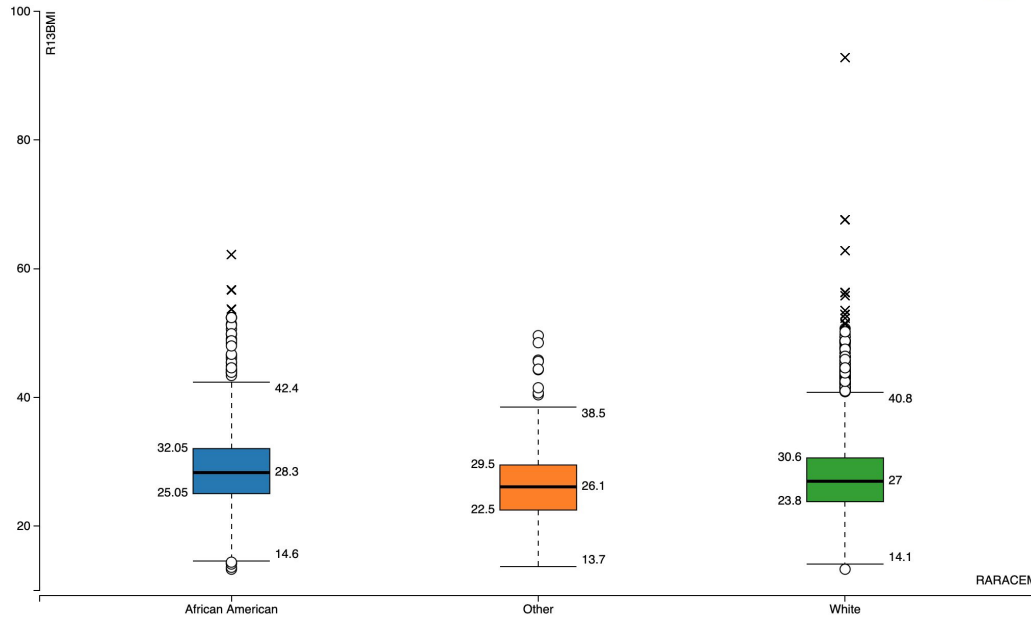
R13IEARN

Choose column for y axis

R13BMI

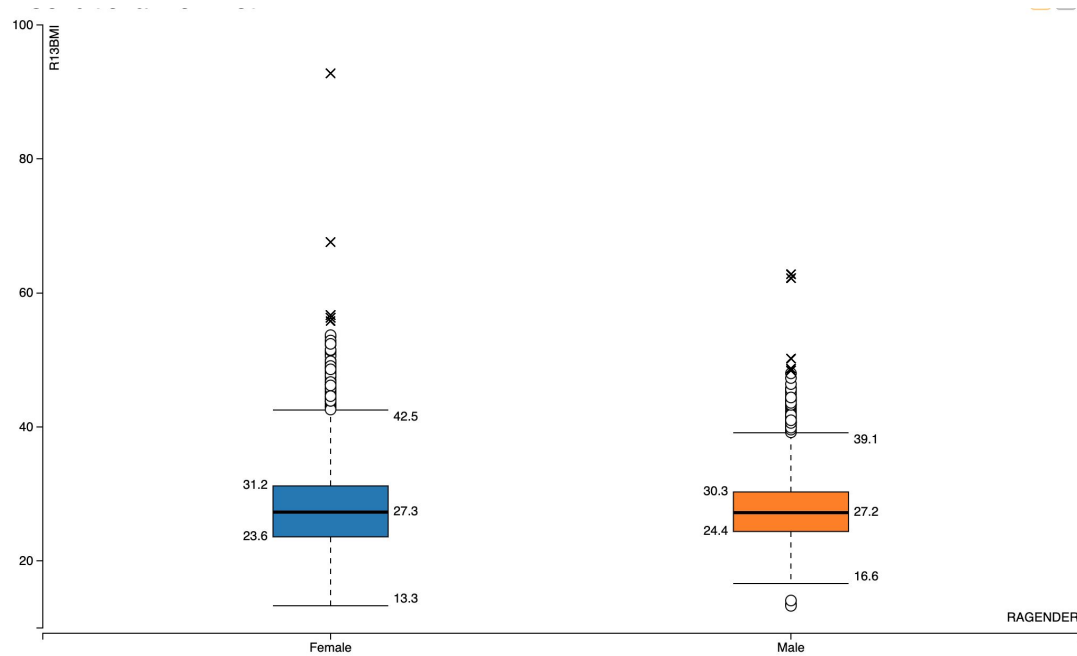
☒ Report on missing values

Data Visualization Results: BMI vs. Ethnicity



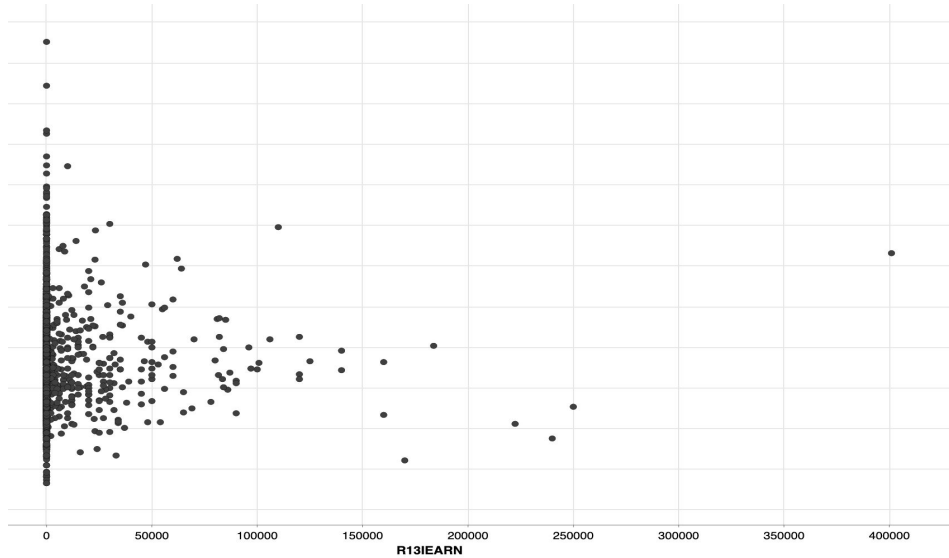
African Americans seem to have the highest mean BMI and largest variance, which indicates that ethnicity can be a good indicator of BMI. Whites seem to have a lot of outliers.

Data Visualization Results: BMI vs. Gender



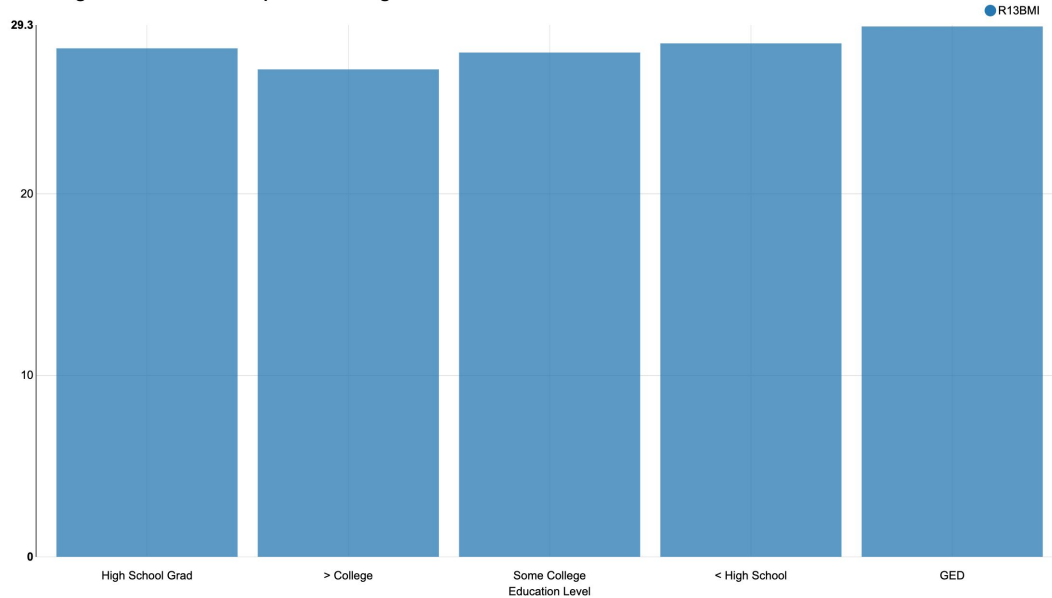
Male and Female
have similar
averages, in
addition, female
has larger variance
by looking at the
IQR

Data Visualization Results: BMI vs. Earnings



There is not a significant correlation between the two variables, a slightly negative line can be drawn. However, there are too many data points on the y-axis, which is not very representative of the trend.

Data Visualization Results: BMI vs. Education



There are differences in BMI among different education levels, with higher education levels corresponding to lower BMI and GED corresponding to highest average BMI.



Hypothesis Testing: BMI

- Quantitative vs. Categorical Variables: One-way ANOVA
- Assumption: BMI is approximately normal
- Group variables used: Gender, Race, Education, Religion, Poverty status
- General hypothesis: is there a significant relationship between BMI and the group variable?



Hypothesis Testing Results: BMI

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
R13BMI	Between Groups	25.5896	1	25.5896	0.7565	0.3845
R13BMI	Within Groups	190,777.9473	5640	33.8259		
R13BMI	Total	190,803.5369	5641			

The p-value of BMI vs. Gender is 0.3845, indicating that gender might not be a significant factor to distinguish BMI.



Hypothesis Testing Results: BMI

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
R13BMI	Between Groups	1,546.4819	2	773.2409	23.035	1.09E-10
R13BMI	Within Groups	189,256.8475	5638	33.5681		
R13BMI	Total	190,803.3294	5640			

The p-value of BMI vs. Race is very tiny, indicating that ethnicity is a significant factor affecting BMI.



Hypothesis Testing Results: BMI

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
R13BMI	Between Groups	388.8313	4	97.2078	2.8739	0.0216
R13BMI	Within Groups	189,790.8569	5611	33.8248		
R13BMI	Total	190,179.6883	5615			

The p-value for BMI vs religion is 0.02, which is also significant.



Hypothesis Testing Results: BMI

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
R13BMI	Between Groups	1,909.2695	4	477.3174	14.2403	1.44E-11
R13BMI	Within Groups	188,877.846	5635	33.5187		
R13BMI	Total	190,787.1155	5639			

The p-value for BMI vs education level is tiny, indicating significance.



Hypothesis Testing Results: BMI

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
R13BMI	Between Groups	139.4713	1	139.4713	4.1406	0.0419
R13BMI	Within Groups	184,082.8215	5465	33.684		
R13BMI	Total	184,222.2929	5466			

P-value for poverty status vs. BMI is smaller than 0.05, indicating that poverty status can be a significant predictor for BMI.

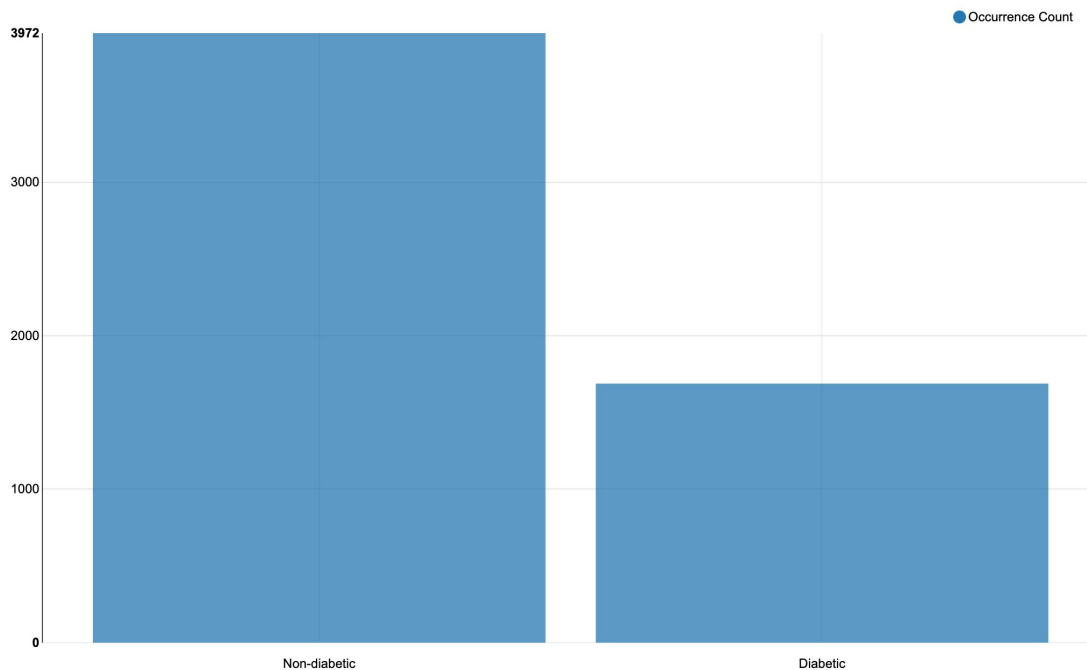


Data Visualization: Diabetes

- Used a bar chart to visualize the relative frequency of diabetes vs. non-diabetes.
- Used bar charts to visualize the relationship between number of children and diabetes



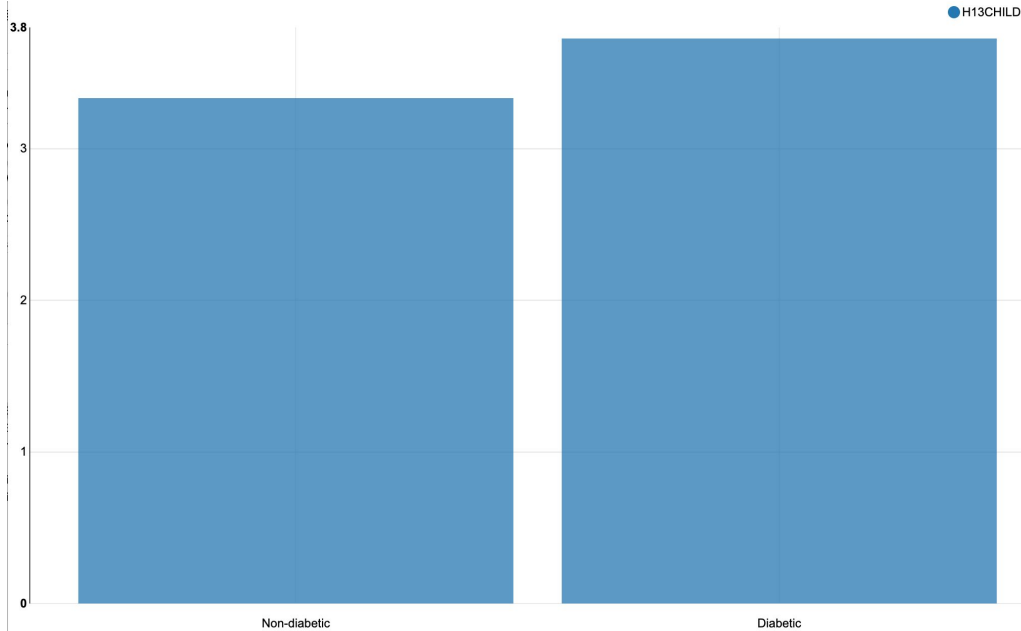
Data Visualization: Diabetes



Overall, there are about twice as many non-diabetic respondents as diabetic respondents.



Data Visualization: Diabetes vs. Num of Children



The bar on the right is higher than the one on the left, indicating possible association between number of children and diabetes



Hypothesis Testing: Diabetes

- Categorical vs. categorical variables
- Chi-square test of homogeneity
- Testing whether a grouping variable is affecting the distribution of diabetic people
- Using “Crosstab” node to derive a table consisting of counts, frequencies for each combination and a p-value using chi-square statistics.

Settings Flow Variables Memory Policy

Row variable:

Column variable:

Weight column:

☐ Enable hiliting

Diabetes vs. Education

- Chi-square statistic has a very tiny probability.
- Percentages among rows differ significantly.
- Indicating education level can be a significant predictor affecting diabetes.

Cross Tabulation of RAEDUC by R13DIAB

Frequency Row Percent	0.0	1.0	Total	<input checked="" type="checkbox"/> Frequency
?	2		2	<input type="checkbox"/> Expected
	100%			<input type="checkbox"/> Deviation
1.0	722	477	1,199	<input type="checkbox"/> Percent
	60.2168%	39.7832%		<input checked="" type="checkbox"/> Row Percent
2.0	180	89	269	<input type="checkbox"/> Column Percent
	66.9145%	33.0855%		<input type="checkbox"/> Cell Chi-Square
3.0	1,282	546	1,828	
	70.1313%	29.8687%		
4.0	875	342	1,217	
	71.8981%	28.1019%		
5.0	911	234	1,145	
	79.5633%	20.4367%		
Total	3,972	1,688	5,660	

Max rows:

Max columns:

Statistics for Table of RAEDUC by R13DIAB

Statistic	DF	Value	Prob
Chi-Square	5	108.3791	9.02E-22

Total sample size: 5660.0



Diabetes vs. Marital Status

- Chi-square statistic has a probability of 0.24.
- Percentages among rows differ but not that much.
- P-value indicates that marital status might not be a significant predictor.

Frequency Row Percent	0.0	1.0	Total
?	7 63.6364%	4 36.3636%	11
1.0	2,072 71.3744%	831 28.6256%	2,903
2.0	49 64.4737%	27 35.5263%	76
3.0	108 64.6707%	59 35.3293%	167
4.0	48 72.7273%	18 27.2727%	66
5.0	374 70.1689%	159 29.8311%	533
7.0	1,206 68.6007%	552 31.3993%	1,758
8.0	108 73.9726%	38 26.0274%	146
Total	3,972	1,688	5,660

☒ Frequency
☐ Expected
☐ Deviation
☐ Percent
☒ Row Percent
☐ Column Percent
☐ Cell Chi-Square

Max rows:

Max columns:

Statistics for Table of R13MSTAT by R13DIAB

Statistic	DF	Value	Prob
Chi-Square	7	9.1115	0.2447



Diabetes vs. Gender

- Chi-square statistic has a probability of 0.0195.
- Fisher's Exact Test has a probability of 0.0211.
- Both p-values are significant.
- Gender is correlated with diabetes.

Frequency Row Percent	0.0	1.0	Total
1.0	1,467 68.3597%	679 31.6403%	2,146
2.0	2,505 71.2863%	1,009 28.7137%	3,514
Total	3,972	1,688	5,660

- ☒ Frequency
- ☐ Expected
- ☐ Deviation
- ☐ Percent
- ☒ Row Percent
- ☐ Column Percent
- ☐ Cell Chi-Square

Max rows:

10

Max columns:

10

Statistics for Table of RAGENDER by R13DIAB

Statistic	DF	Value	Prob
Chi-Square	1	5.4523	0.0195
Fisher's Exact Test (2-tail)			0.0211

Total sample size: 5660.0



Diabetes vs. Ethnicity

- Chi-square statistic has a very tiny p-value.
- Percentages among rows differ a lot.
- There can be a strong association between ethnicity and diabetes.
- African American seems to have a higher percentage of diabetes (1.0).

Cross Tabulation of RARACEM by R13DIAB

Frequency Row Percent	0.0	1.0	Total
?		1	1
		100%	
1.0	3,337	1,228	4,565
	73.0997%	26.9003%	
2.0	485	357	842
	57.601%	42.399%	
3.0	150	102	252
	59.5238%	40.4762%	
Total	3,972	1,688	5,660

- ☒ Frequency
- ☐ Expected
- ☐ Deviation
- ☐ Percent
- ☒ Row Percent
- ☐ Column Percent
- ☐ Cell Chi-Square

Max rows:

10

Max columns:

10

Statistics for Table of RARACEM by R13DIAB

Statistic	DF	Value	Prob
Chi-Square	3	97.5766	5.16E-21

Total sample size: 5660.0



Diabetes vs. Poverty Status

- Chi-square statistic probability is tiny.
- Percentages among rows differ significantly.
- P-value indicates poverty status is a good indicator of diabetes, with a positive correlation.

Cross Tabulation of H13INPOV by R13DIAB

Frequency Row Percent	0.0	1.0	Total
?	116 63.7363%	66 36.2637%	182
0.0	3,524 71.3216%	1,417 28.6784%	4,941
1.0	332 61.825%	205 38.175%	537
Total	3,972	1,688	5,660

- ☒ Frequency
☐ Expected
☐ Deviation
☐ Percent
☒ Row Percent
☐ Column Percent
☐ Cell Chi-Square

Max rows:

10

Max columns:

10

Statistics for Table of H13INPOV by R13DIAB

Statistic	DF	Value	Prob
Chi-Square	2	24.5986	4.55E-6

Total sample size: 5660.0

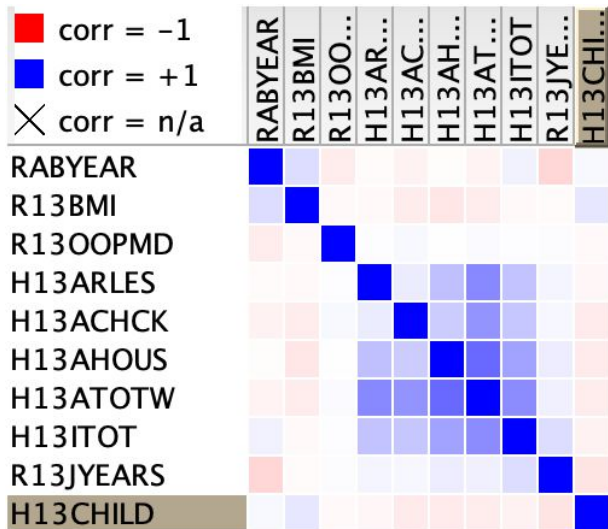


Modeling BMI: Multiple Linear Regression

- Independent variables (predictors): a chosen set of variables including those that were found significant using one-way anova tests.
- Include additional variables indicating other socioeconomic status such as assets, health insurance coverage and income.
- The correlation factor is small, but results can still reveal the relative importance of different predictor variables



Linear Correlation



First of all, I chose a set of variables and calculated their correlations with BMI

Variables	Correlation
Birth Year (RABYEAR)	0.1318
Assets on checking (H13ACHCK)	-0.07
Primary residence asset (H13AHOUS)	-0.096
Total wealth without IRA (H13TOTW)	-0.075
Number of children (H13CHILD)	0.0925

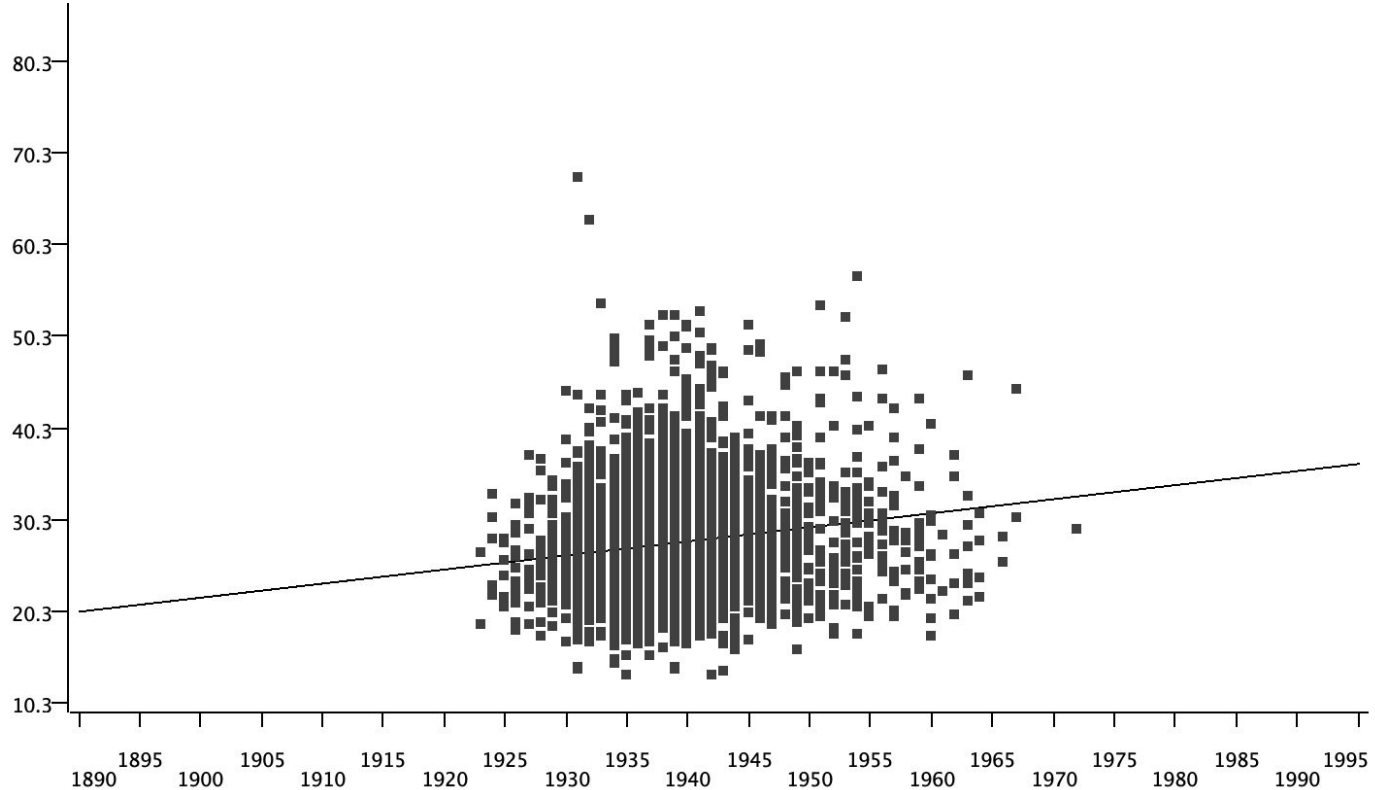
Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P> t
RABYEAR	0.1543	0.017	9.0932	0.0
H13ACHCK	-1.71E-6	9.46E-7	-1.8067	0.0709
H13AHOUS	-1.25E-6	4.20E-7	-2.9717	0.003
H13ATOTW	1.23E-7	1.15E-7	1.0675	0.2858
H13CHILD	0.1798	0.0421	4.2742	1.97E-5
1.0_R13MNEV	-0.8588	0.5437	-1.5797	0.1143
0.0_H13ANYFIN	-1.5377	1.4456	-1.0637	0.2875
1.0_H13ANYFAM	-0.1062	1.2437	-0.0854	0.932
0.0_H13CPL	-0.2258	0.2008	-1.1244	0.2609
2.0_RAGENDER	-0.973	0.2585	-3.7643	0.0002
3.0_RARACEM	-1.0957	0.4383	-2.4997	0.0125
2.0_RARACEM	0.9921	0.2568	3.8635	0.0001
3.0_RAEDUC	0.3566	0.2552	1.3976	0.1623
5.0_RAEDUC	-0.0551	0.3039	-0.1814	0.8561
4.0_RAEDUC	0.3012	0.284	1.0604	0.289
2.0_RAEDUC	1.1204	0.4555	2.4597	0.014
1.0_R13EFFORT	0.7441	0.2264	3.2867	0.001
2.0_R13VGACTX	-0.1019	0.5753	-0.177	0.8595
3.0_R13VGACTX	0.0398	0.6116	0.0651	0.9481
5.0_R13VGACTX	0.6501	0.5535	1.1745	0.2403
4.0_R13VGACTX	0.4885	0.6177	0.7908	0.4291
3.0_R13MDACTX	0.7525	0.3841	1.959	0.0502
2.0_R13MDACTX	0.6514	0.3551	1.8343	0.0667
5.0_R13MDACTX	2.0197	0.379	5.3287	1.05E-7
4.0_R13MDACTX	1.1722	0.414	2.8313	0.0047
3.0_R13LTACTX	0.4823	0.3655	1.3196	0.1871

2.0_R13LTACTX	-0.0371	0.3543	-0.1048	0.9165
4.0_R13LTACTX	-0.1939	0.4394	-0.4413	0.6591
5.0_R13LTACTX	-0.0876	0.4145	-0.2114	0.8326
1.0_R13MAMMOG	0.8917	0.2413	3.6947	0.0002
1.0_R13HLTHLM	0.8326	0.1885	4.4178	1.03E-5
3.0_H13NHMLIV	-1.736	0.6693	-2.5937	0.0095
2.0_H13NHMLIV	-0.097	1.562	-0.0621	0.9505
1.0_R13SSDI	0.2048	0.5186	0.3949	0.693
1.0_R13COVR	0.2287	0.2756	0.8297	0.4068
1.0_R13COVS	0.1714	0.3584	0.4784	0.6324
1.0_R13HIGOV	0.99	0.4228	2.3413	0.0193
1.0_R13PENINC	0.4793	0.1933	2.4799	0.0132
Intercept	-274.7857	33.019	-8.3221	2.22E-16
Multiple R-Squared: 0.0971 Adjusted R-Squared: 0.088				

For the social determinants, assets in checking account (ACHCK), asset for housing (AHOUS), number of children (H13CHILD), gender (RAGENDER) , education (RAEDUC), nursing home status(NHMLIV), government insurance (HIGOV), pension (PENINC) are all significant predictors (p-values < 0.05).

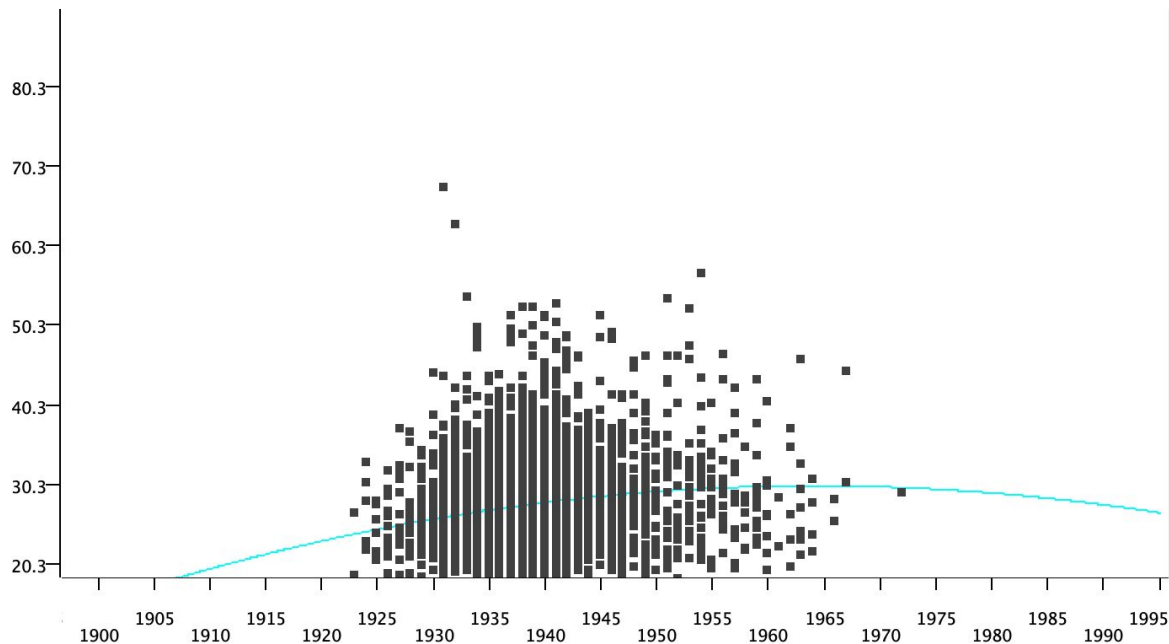
Multiple Linear Regression Visualization



Scatter plot of BMI vs. birth year (one of the predictor variables), there are a few extreme BMI data points.



Modeling BMI: Polynomial Regression

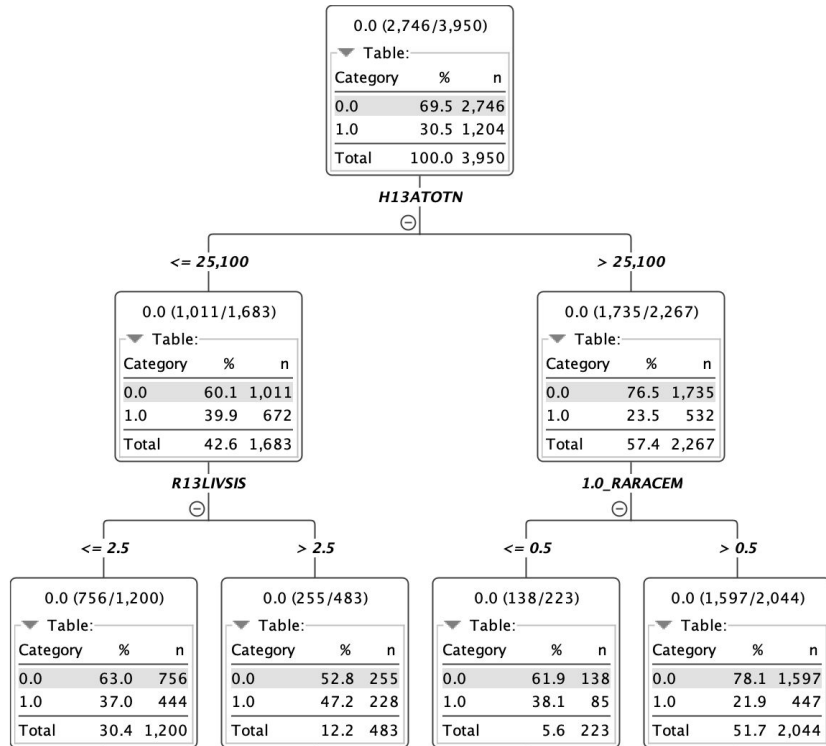


Scatterplot of BMI
vs Birth Year with
estimated 2nd
degree polynomial
regression line

Modeling Diabetes: Feature Selection

Row ID	I	Nr. of features	D	Accuracy	S	Added feature
4	4		0.705		1.0_H13CPL	whether couple household
2	2		0.703		R13IWCMP	worker compensation income
1	1		0.702		8.0_R13CENDIV	Census division
8	8		0.701		5.0_R13CENREG	Census region
3	3		0.7		1.0_H13ANYFAM	Any family respondent in household
10	10		0.696		2.0_R13MSTAT	Marital status indicator #2
5	5		0.694		R13HESRC3	Source of insurance plan #3
7	7		0.694		5.0_R13MSTAT	Marital status indicator #5
13	13		0.694		0.0_R13MNEV	R never married indicator
11	11		0.693		1.0_R13CENDIV	Census division
6	6		0.691		3.0_RARACEM	R race indicator
9	9		0.691		5.0_RARELIG	R religion indicator
12	12		0.69		0.0_R13HIGOV	R covered by government insurance
15	15		0.688		5.0_R13CENDIV	Census division
18	18		0.688		R13WORK2	R work at a 2nd job
14	14		0.687		2.0_RARACEM	R race indicator
20	20		0.685		0.0_R13HOSP	R hospital stay prev 2 yrs
16	16		0.68		4.0_RARELIG	R religion indicator
17	17		0.679		0.0_H13NHMLIV	R living in nursing home
19	19		0.676		1.0_R13MNEV	R never married indicator

Modeling Diabetes: Decision Tree



On the left are the first two levels of the decision tree. The first split was based on the column “H13ATOTN”, which represented total non-housing wealth, and the cutoff was 25100. Then, the left branch was first splitted by the number of living sisters with cutoff 2.5, and right branch splitted based on the race indicator #1, which is the indicator of being White/Caucasian.



Modeling Diabetes: Decision Tree

R13DIAB \ ...	0.0	1.0	4.0	3.0
0.0	997	213	0	0
1.0	354	130	0	0
4.0	0	0	0	0
3.0	0	0	0	0

The accuracy for the decision tree using the selected set of features was 66.529 %, which was expected to be lower than the other algorithms since it was consist of only one tree.



Modeling Diabetes: Random Forest

R13DIAB \ ...	0.0	1.0	4.0	3.0
0.0	1128	82	0	0
1.0	409	75	0	0
4.0	0	0	0	0
3.0	0	0	0	0

A total of 100 trees were modeled to create the random forest. The accuracy on the test set was 71.015%, which is much lower than using all variables, so the model is not overfitting, however, there are a lot of space to improve the model



Modeling Diabetes: Logistic Regression

R13DIAB \ ...	0.0	1.0	4.0	3.0
0.0	488	6	0	0
1.0	202	5	0	0
4.0	0	0	0	0
3.0	0	0	0	0

The accuracy for the logistic regression using the same set of features was 70.328 %, which is better than the decision tree model, however, there are much fewer observations due to missing values.



Modeling Diabetes: Gradient Boosting

R13DIAB \ ...	0.0	1.0	4.0	3.0
0.0	1089	121	0	0
1.0	382	102	0	0
4.0	0	0	0	0
3.0	0	0	0	0

The accuracy was 70.307%, similar to the logistic regression model. However, GB is a much more sophisticated method than logistic regression, therefore it might not be very reflective to use just one test set.



Evaluation: 10-fold Cross Validation

Row ID	D Error in %	I Size of Test Set	I Error Count
fold 0	37.522	565	212
fold 1	38.652	564	218
fold 2	43.186	565	244
fold 3	40.957	564	231
fold 4	38.652	564	218
fold 5	40	565	226
fold 6	40.78	564	230
fold 7	39.469	565	223
fold 8	40.957	564	231
fold 9	36.525	564	206

Decision Tree: Error rates for decision trees are higher than those generated from random forests, which was expected since random forest generalized the model much better.



Evaluation: 10-fold Cross Validation

Row ID	D Error in %	I Size of Test Set	I Error Count
fold 0	30.088	565	170
fold 1	30.496	564	172
fold 2	31.327	565	177
fold 3	31.383	564	177
fold 4	30.142	564	170
fold 5	29.204	565	165
fold 6	32.624	564	184
fold 7	27.434	565	155
fold 8	28.901	564	163
fold 9	29.078	564	164

Random Forest: Generally the error rates are around 30%, with some folds much lower than the other folds, and some folds much higher than the others.



Evaluation: 10-fold Cross Validation

Row ID	D Error in %	I Size of Test Set	I Error Count
fold 0	39.744	234	93
fold 1	38.889	234	91
fold 2	31.76	233	74
fold 3	46.581	234	109
fold 4	37.339	233	87
fold 5	31.624	234	74
fold 6	48.718	234	114
fold 7	48.498	233	113
fold 8	41.88	234	98
fold 9	34.335	233	80

Logistic Regression: Error rates varied a lot by folds, which indicated that data behaved quite differently across folds and logistic regression might not generalize well.



Evaluation: 10-fold Cross Validation

Row ID	D Error in %	I Size of Test Set	I Error Count
fold 0	32.743	565	185
fold 1	28.901	564	163
fold 2	29.912	565	169
fold 3	30.142	564	170
fold 4	28.546	564	161
fold 5	28.85	565	163
fold 6	29.965	564	169
fold 7	28.85	565	163
fold 8	29.433	564	166
fold 9	29.078	564	164

Gradient Boosting performed best out of the four categorical classifiers based on this specific cross validation. This indicated that although random forest had higher accuracy than boosting in the testing round, it might be simply due to the different testing data used.



Discussion

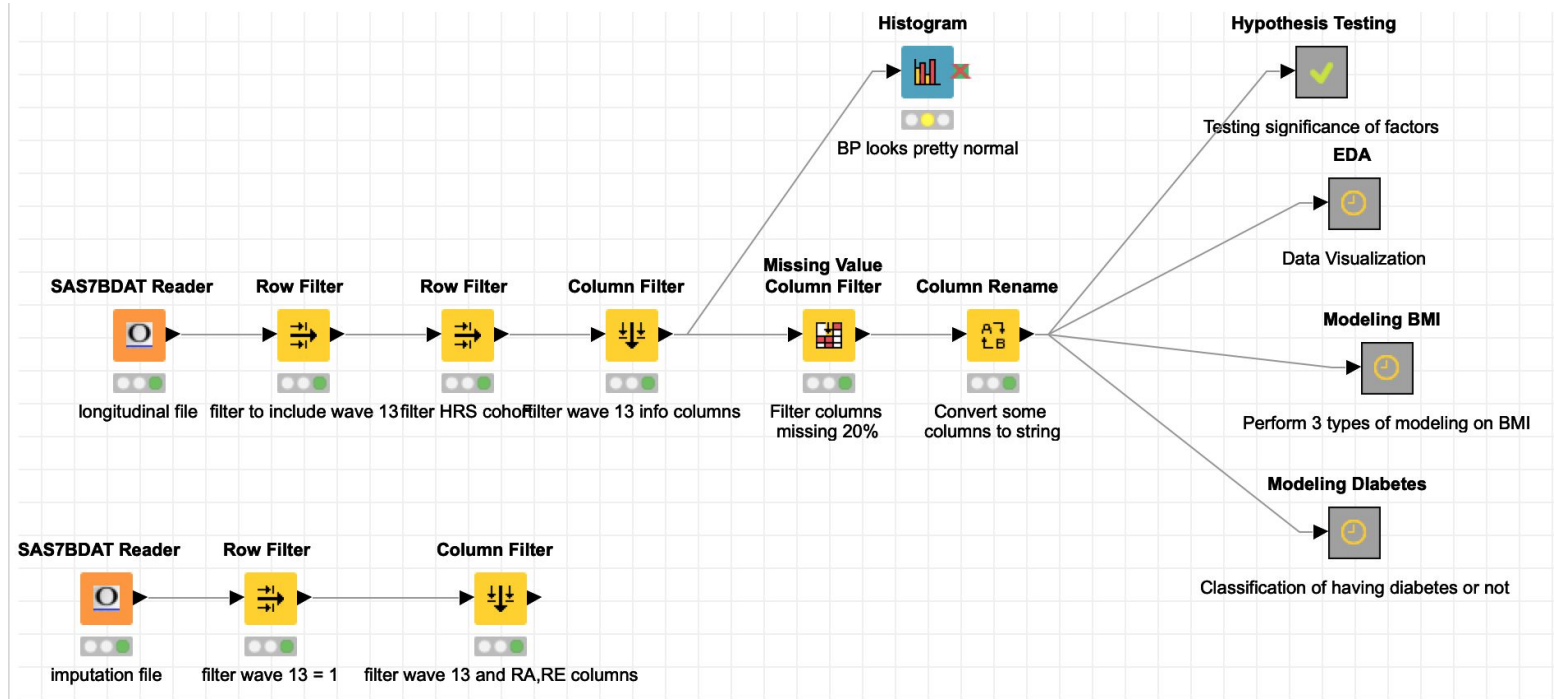
- Found several socioeconomic factors that are significantly related with health outcomes, specifically they are: religion, ethnicity, total wealth, different assets measurement, employment status and health coverage status.
- Particularly, asset variables are highly significant, although not all asset variables were included due to high correlation among those variables.
- Still need further exploration on different health indicators
- Some nodes like feature selection takes a long time, forward selection takes up to 20 minutes with 20 features, while backward selection takes more than several hours (and still did not end, force termination)
- Issue exists with covariance matrix for logistic regression and r-squared value for tree regression, need further exploration



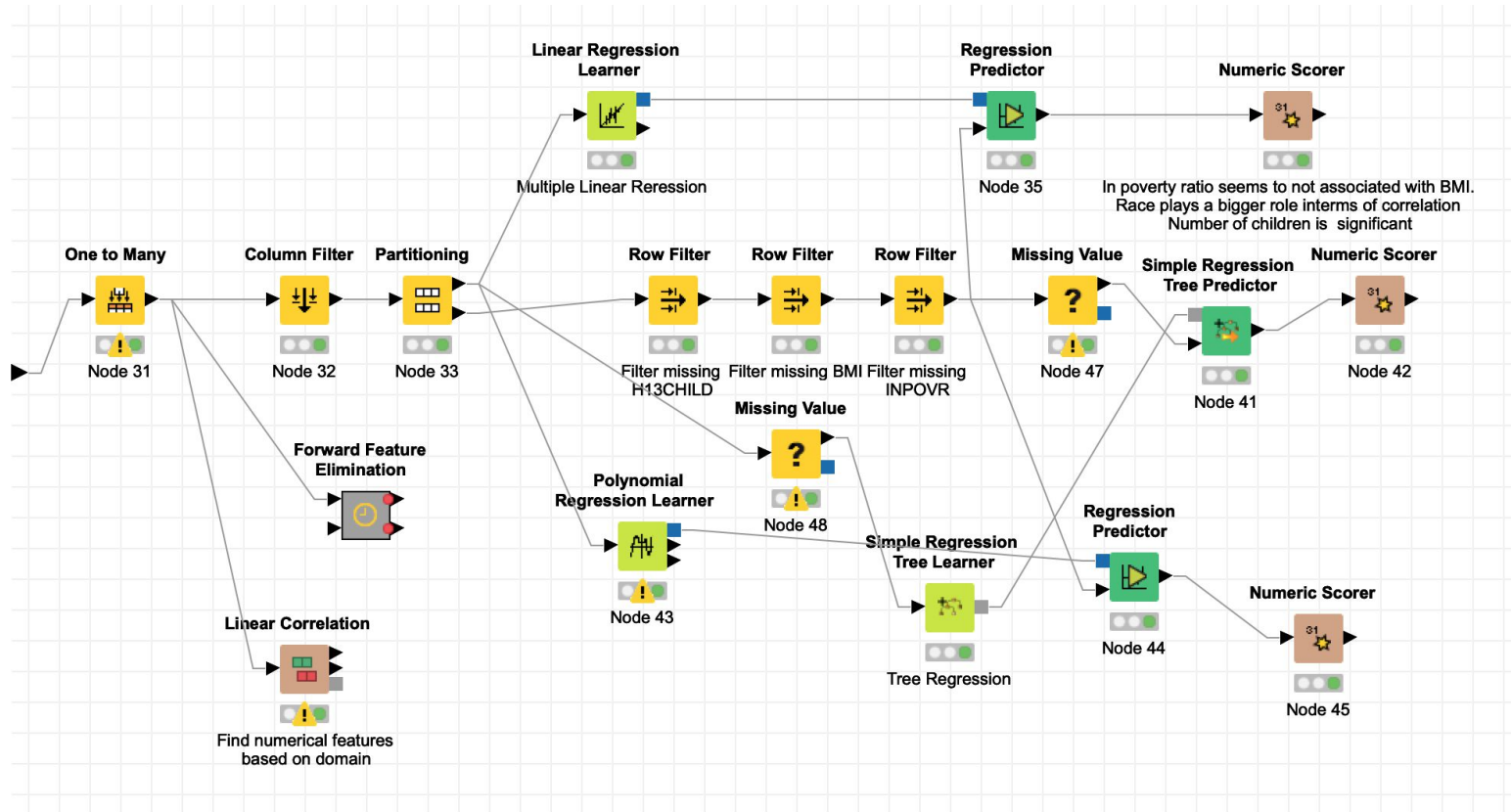
References

1. Bugliari, Delia, et al. "RAND HRS Longitudinal File 2016 (V1) Documentation." *Santa Monica, CA: RAND Center for the Study of Aging* (2019).
2. Seligman, Benjamin, Shripad Tuljapurkar, and David Rehkopf. "Machine learning approaches to the social determinants of health in the health and retirement study." *SSM-population health* 4 (2018): 95-99.

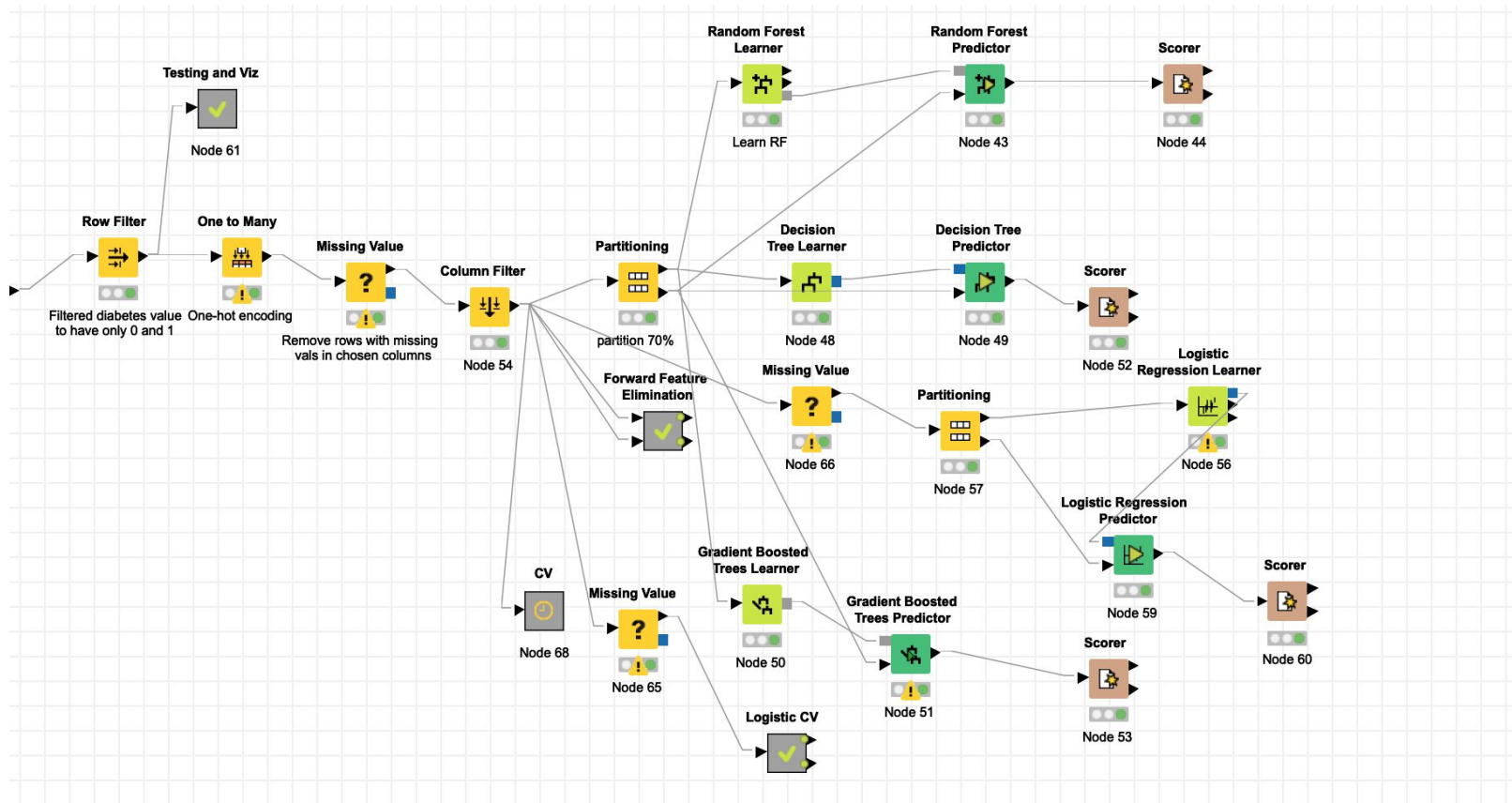
Appendix: KNIME Workflows



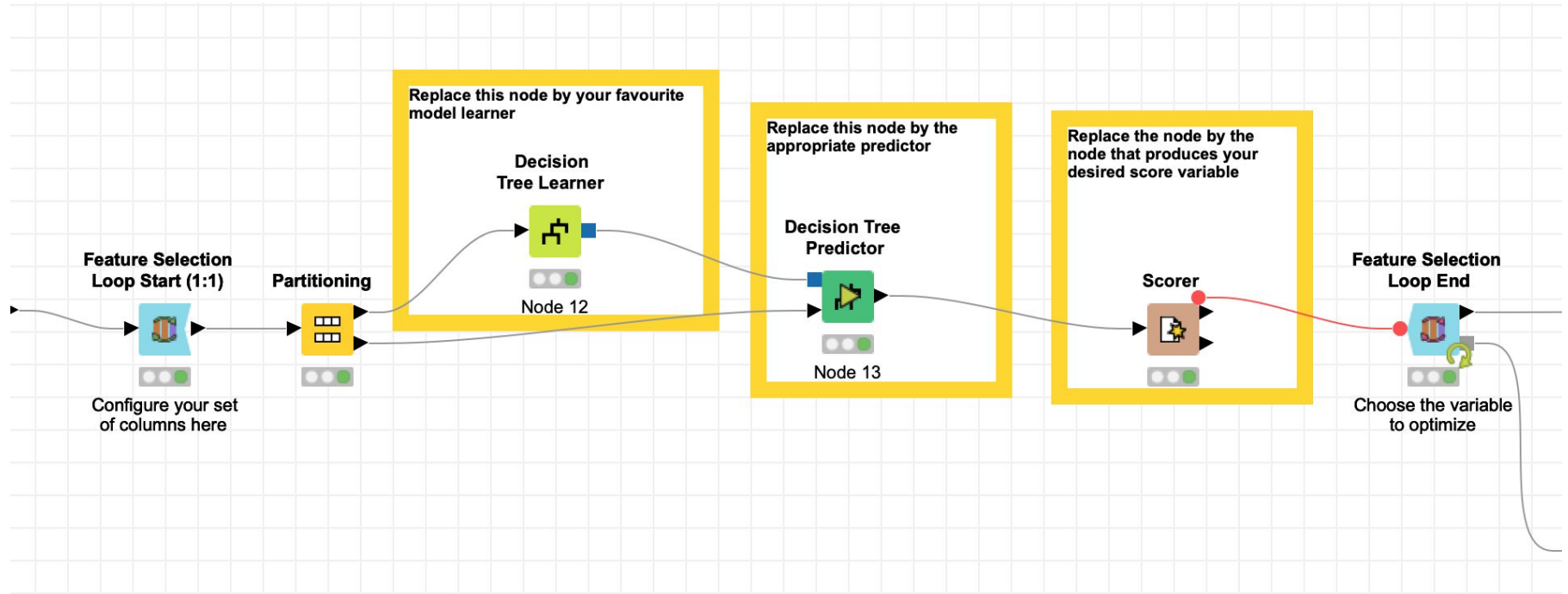
Overview of the whole workflow



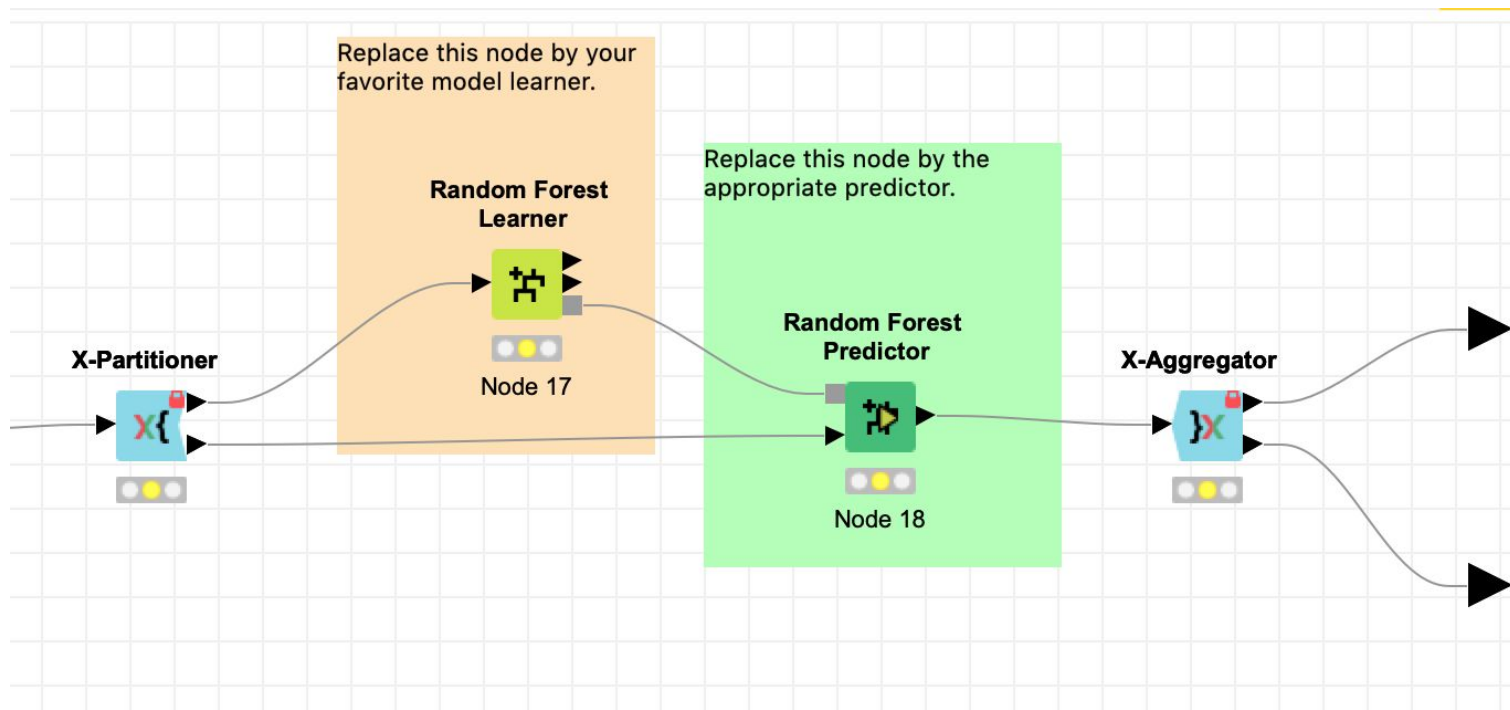
Modeling BMI workflow



Modeling diabetes workflow



Feature selection workflow



Cross validation workflow