PREDICTING HEALTHCARE COST

Do demographics predict healthcare spending?

BACKGROUND

- Healthcare spending will be 20% of the USA economy by 2025[1]
- Where is that spending happening?
- Who is that spending on?
- What demographic factors are most likely to influence spending in a county?

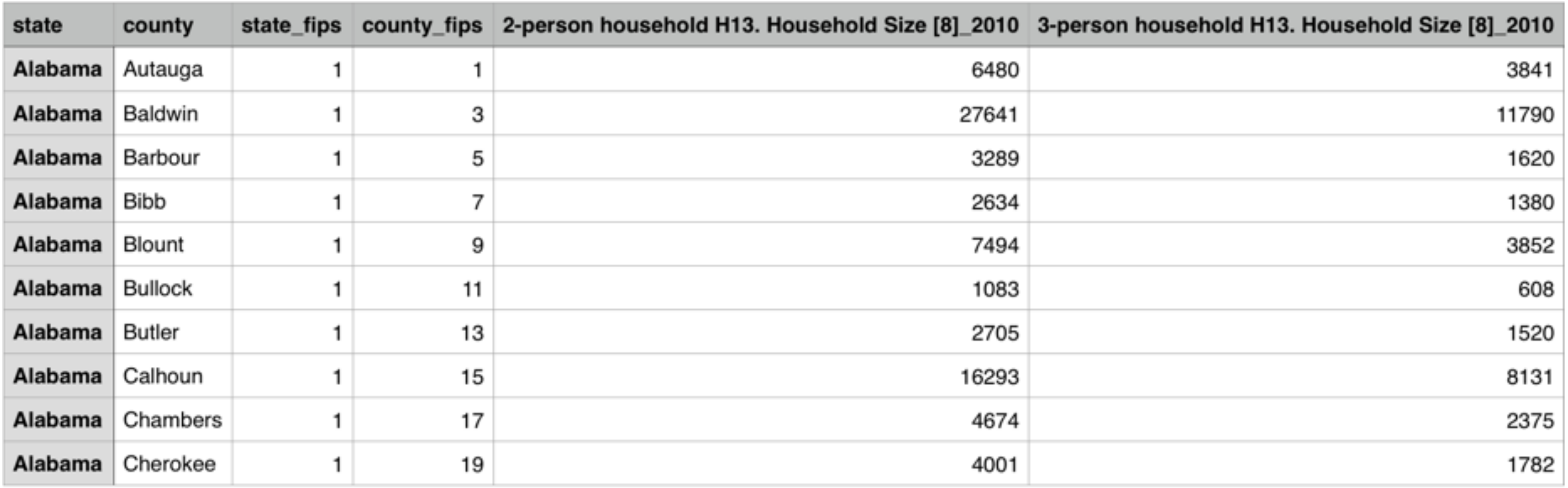
[1] *Sean P. Keehan, John A. Poisal, Gigi A. Cuckler, Andrea M. Sisko, Sheila D. Smith, Andrew J. Madison, Devin A. Stone, Christian J. Wolfe, and Joseph M. Lizonitz National Health Expenditure Projections, 2015–25: Economy, Prices, And Aging Expected To Shape Spending And Enrollment Health Affairs Vol 35 No. 7*

DATASET

- USA Census bureau 2010 decennial census data
- Centers for Medicaid and Medicare outpatient spending data (2011-2014)
- US Dept of Housing and and Urban Development HUD USPS ZIP Code crosswalk Q2 2016

US CENSUS BUREAU DECENNIAL CENSUS DATA (2010)

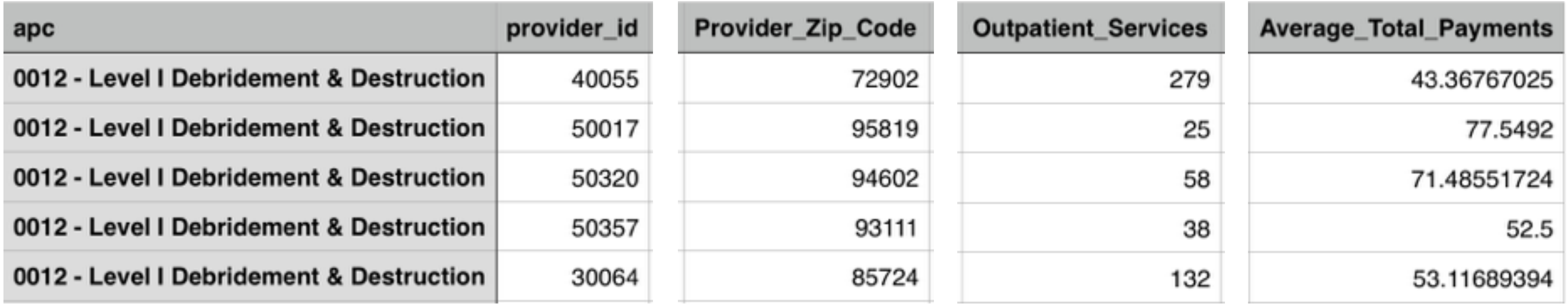
- Data available through API only



state	county	state_fips	county_fips	2-person household H13. Household Size [8]_2010	3-person household H13. Household Size [8]_2010
Alabama	Autauga	1	1	6480	3841
Alabama	Baldwin	1	3	27641	11790
Alabama	Barbour	1	5	3289	1620
Alabama	Bibb	1	7	2634	1380
Alabama	Blount	1	9	7494	3852
Alabama	Bullock	1	11	1083	608
Alabama	Butler	1	13	2705	1520
Alabama	Calhoun	1	15	16293	8131
Alabama	Chambers	1	17	4674	2375
Alabama	Cherokee	1	19	4001	1782

CMS OUTPATIENT PAYMENT DATA

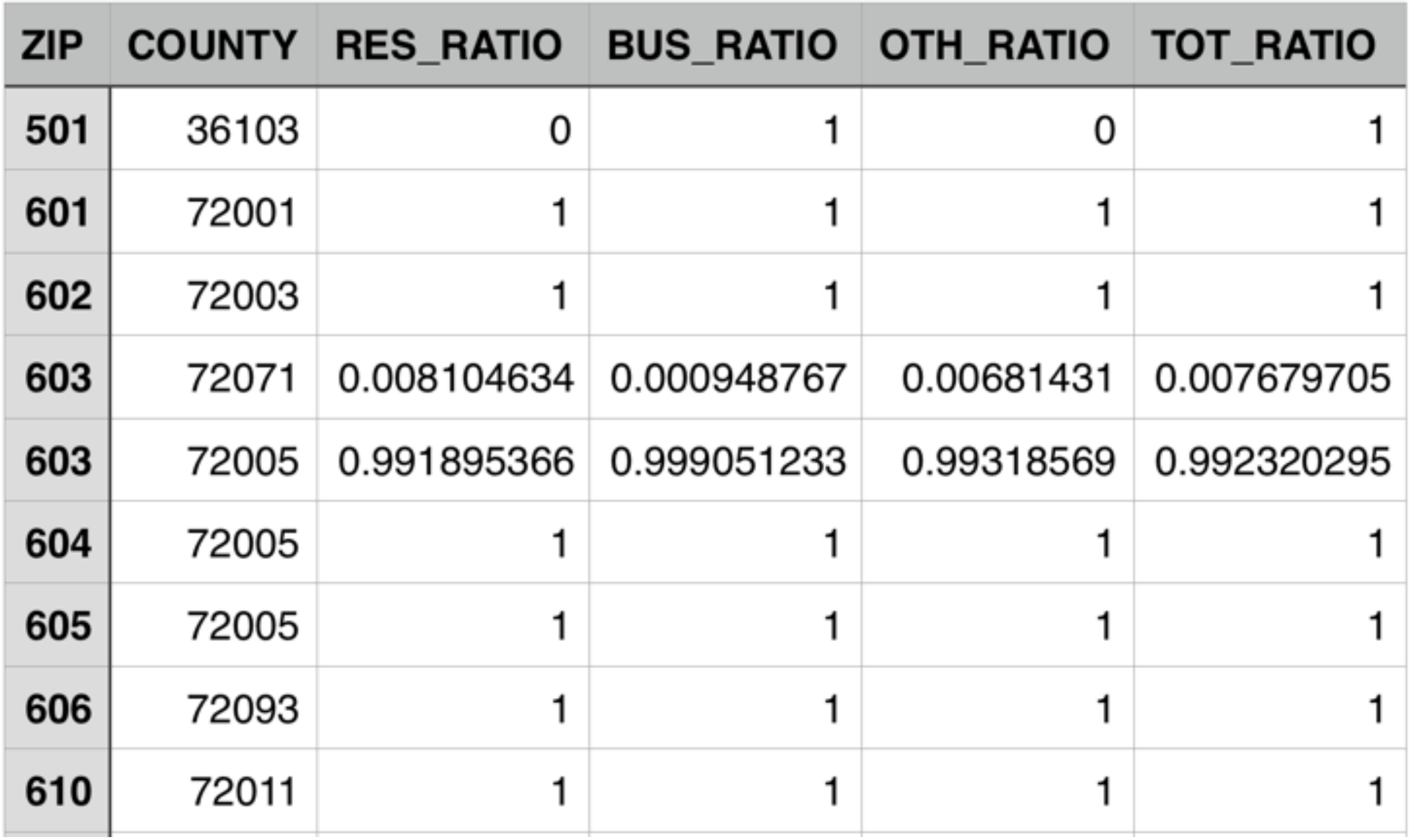
- 2014 data available as download from CMS website
    - 4.5 mb CSV



apc	provider_id	Provider_Zip_Code	Outpatient_Services	Average_Total_Payments
0012 - Level I Debridement & Destruction	40055	72902	279	43.36767025
0012 - Level I Debridement & Destruction	50017	95819	25	77.5492
0012 - Level I Debridement & Destruction	50320	94602	58	71.48551724
0012 - Level I Debridement & Destruction	50357	93111	38	52.5
0012 - Level I Debridement & Destruction	30064	85724	132	53.11689394

HUD ZIP CROSSWALK

- Available as direct download from HUD website
  - Download CSV - 3.2 mb

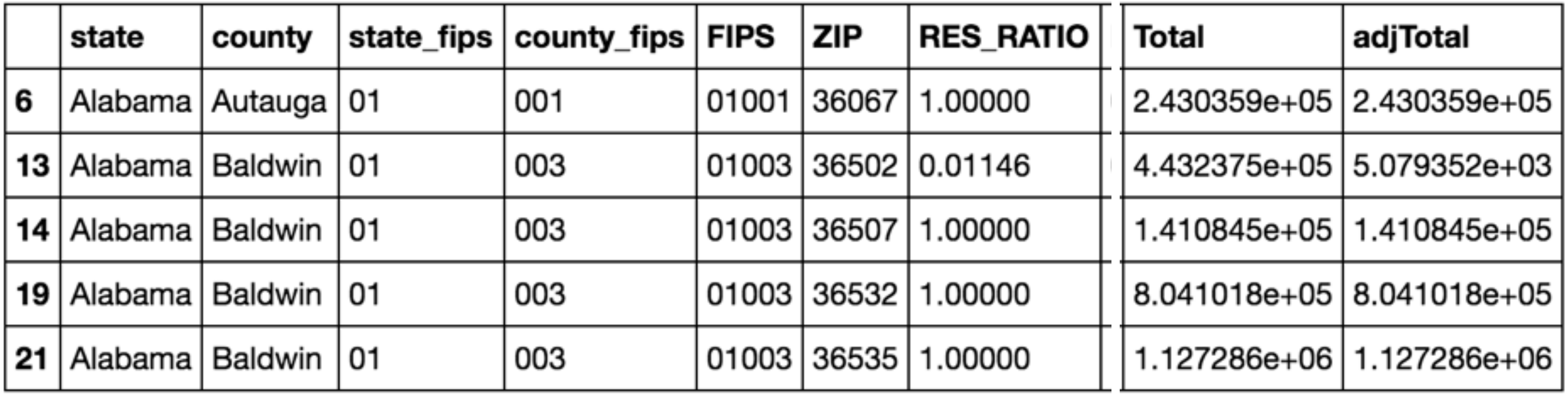


ZIP	COUNTY	RES_RATIO	BUS_RATIO	OTH_RATIO	TOT_RATIO
501	36103	0	1	0	1
601	72001	1	1	1	1
602	72003	1	1	1	1
603	72071	0.008104634	0.000948767	0.00681431	0.007679705
603	72005	0.991895366	0.999051233	0.99318569	0.992320295
604	72005	1	1	1	1
605	72005	1	1	1	1
606	72093	1	1	1	1
610	72011	1	1	1	1

DATA CLEANING

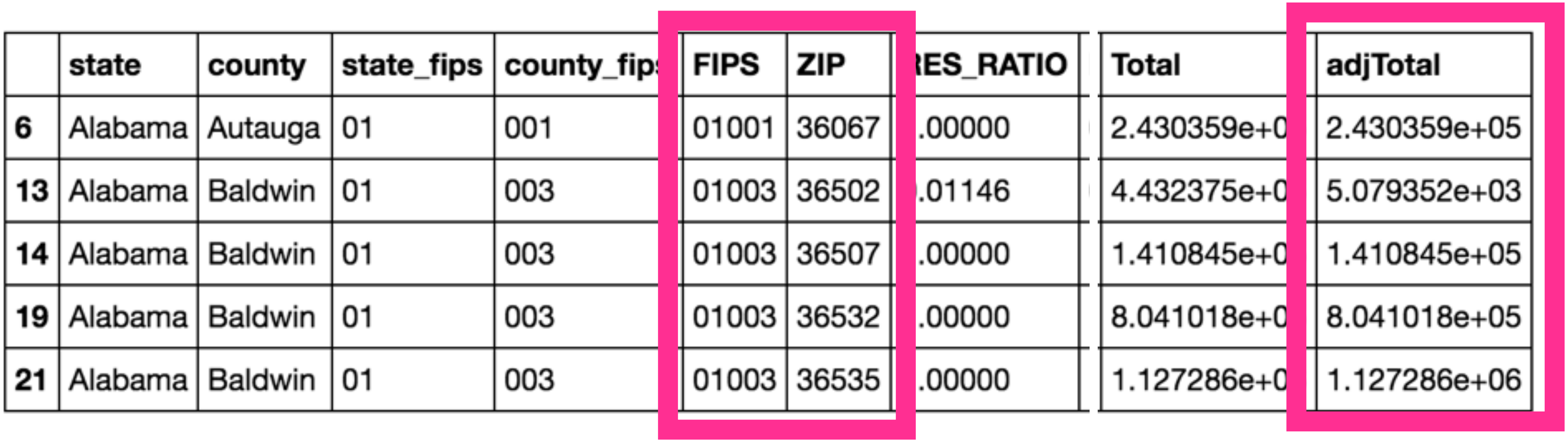
- Ensure that all the zip codes still have leading zeros
- Payment info is grouped and summed into one payment value per zip code (rather than several payment types)
- create combined FIPS codes from component state and county FIPS to match counties between dataframes
- change column names that will be used as keys with the merge command. The dataframes to be merged must each have a column with the same name which will be used to line up each row from each dataframe.
- merge the zip to county cross walk and census data frames in a new dataframe
- merge zip-census data frame with the payment dataframe
- Drop columns that are 100 % NaN values
- convert columns we want to treat as numeric to numeric types
- drop rows with NaN values
- calculate adjusted payment for zip codes not entirely within counties

CLEAN DATA (RESPONSE)



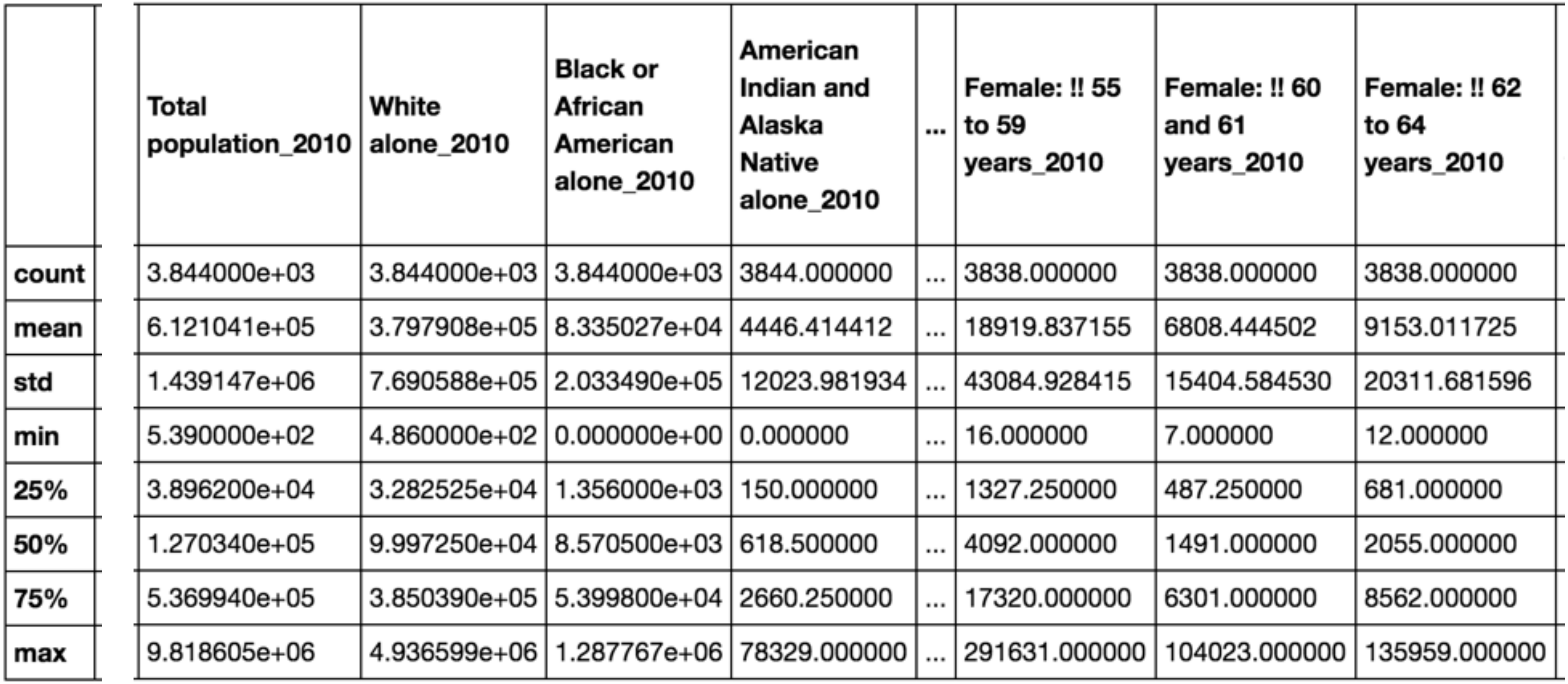
	state	county	state_fips	county_fips	FIPS	ZIP	RES_RATIO	Total	adjTotal
6	Alabama	Autauga	01	001	01001	36067	1.00000	2.430359e+05	2.430359e+05
13	Alabama	Baldwin	01	003	01003	36502	0.01146	4.432375e+05	5.079352e+03
14	Alabama	Baldwin	01	003	01003	36507	1.00000	1.410845e+05	1.410845e+05
19	Alabama	Baldwin	01	003	01003	36532	1.00000	8.041018e+05	8.041018e+05
21	Alabama	Baldwin	01	003	01003	36535	1.00000	1.127286e+06	1.127286e+06

CLEAN DATA (RESPONSE)



	state	county	state_fips	county_fip	FIPS	ZIP	RES_RATIO	Total	adjTotal
6	Alabama	Autauga	01	001	01001	36067	.00000	2.430359e+0	2.430359e+05
13	Alabama	Baldwin	01	003	01003	36502	.01146	4.432375e+0	5.079352e+03
14	Alabama	Baldwin	01	003	01003	36507	.00000	1.410845e+0	1.410845e+05
19	Alabama	Baldwin	01	003	01003	36532	.00000	8.041018e+0	8.041018e+05
21	Alabama	Baldwin	01	003	01003	36535	.00000	1.127286e+0	1.127286e+06

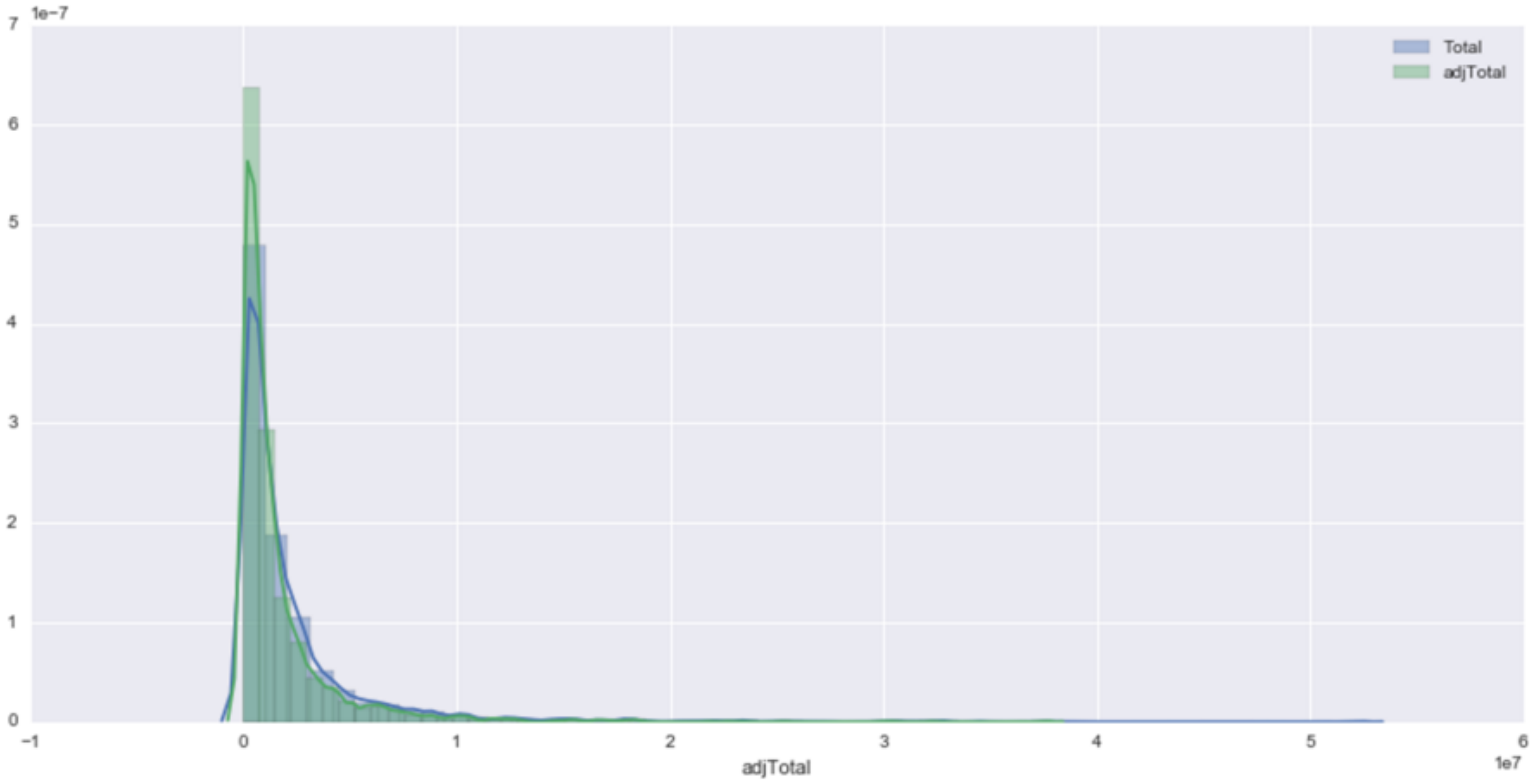
CLEAN DATA (FEATURES - SAMPLE)



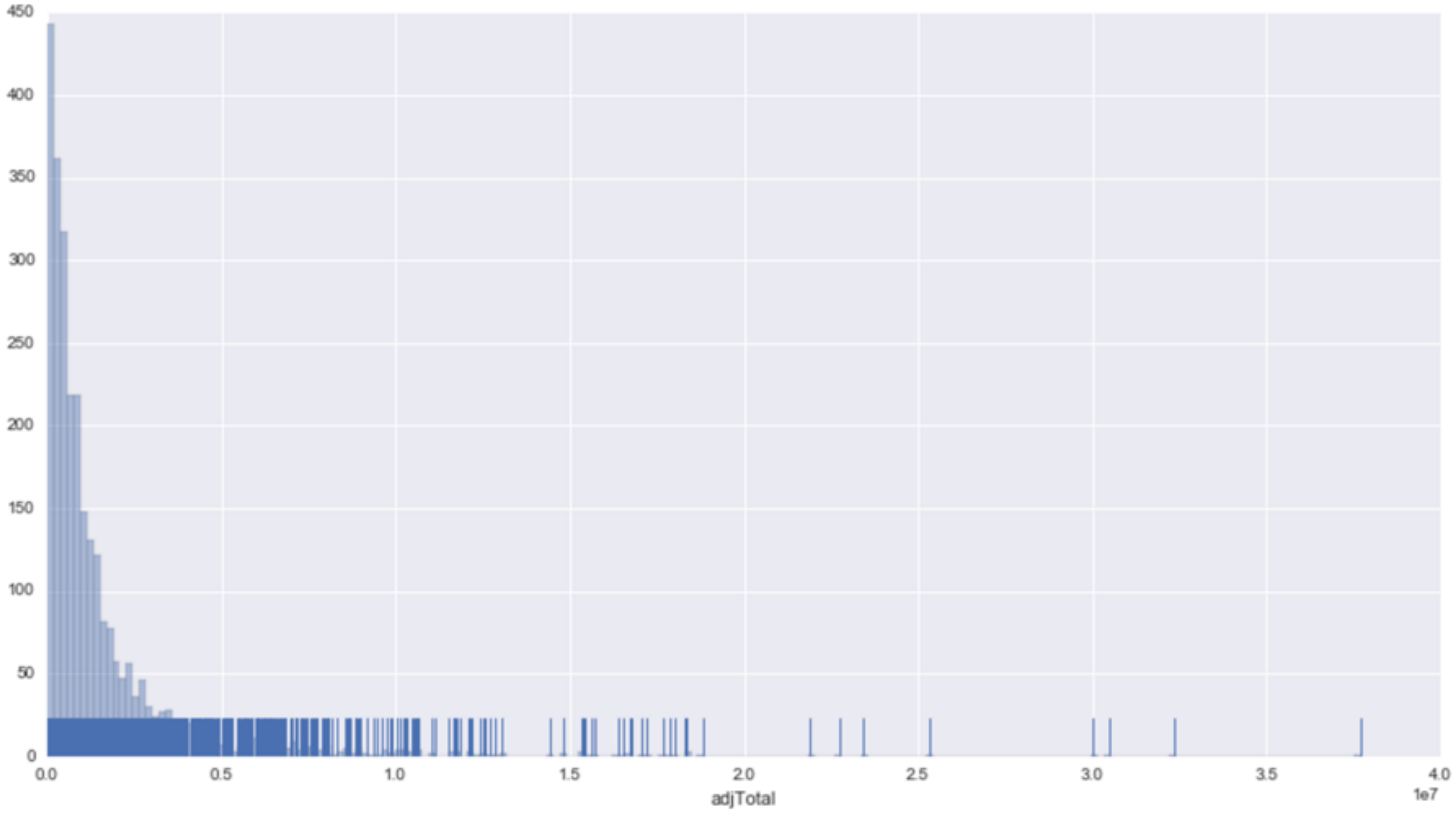
	Total population_2010	White alone_2010	Black or African American alone_2010	American Indian and Alaska Native alone_2010	...	Female: !! 55 to 59 years_2010	Female: !! 60 and 61 years_2010	Female: !! 62 to 64 years_2010
count	3.844000e+03	3.844000e+03	3.844000e+03	3844.000000	...	3838.000000	3838.000000	3838.000000
mean	6.121041e+05	3.797908e+05	8.335027e+04	4446.414412	...	18919.837155	6808.444502	9153.011725
std	1.439147e+06	7.690588e+05	2.033490e+05	12023.981934	...	43084.928415	15404.584530	20311.681596
min	5.390000e+02	4.860000e+02	0.000000e+00	0.000000	...	16.000000	7.000000	12.000000
25%	3.896200e+04	3.282525e+04	1.356000e+03	150.000000	...	1327.250000	487.250000	681.000000
50%	1.270340e+05	9.997250e+04	8.570500e+03	618.500000	...	4092.000000	1491.000000	2055.000000
75%	5.369940e+05	3.850390e+05	5.399800e+04	2660.250000	...	17320.000000	6301.000000	8562.000000
max	9.818605e+06	4.936599e+06	1.287767e+06	78329.000000	...	291631.000000	104023.000000	135959.000000

EXPLORATORY DATA ANALYSIS

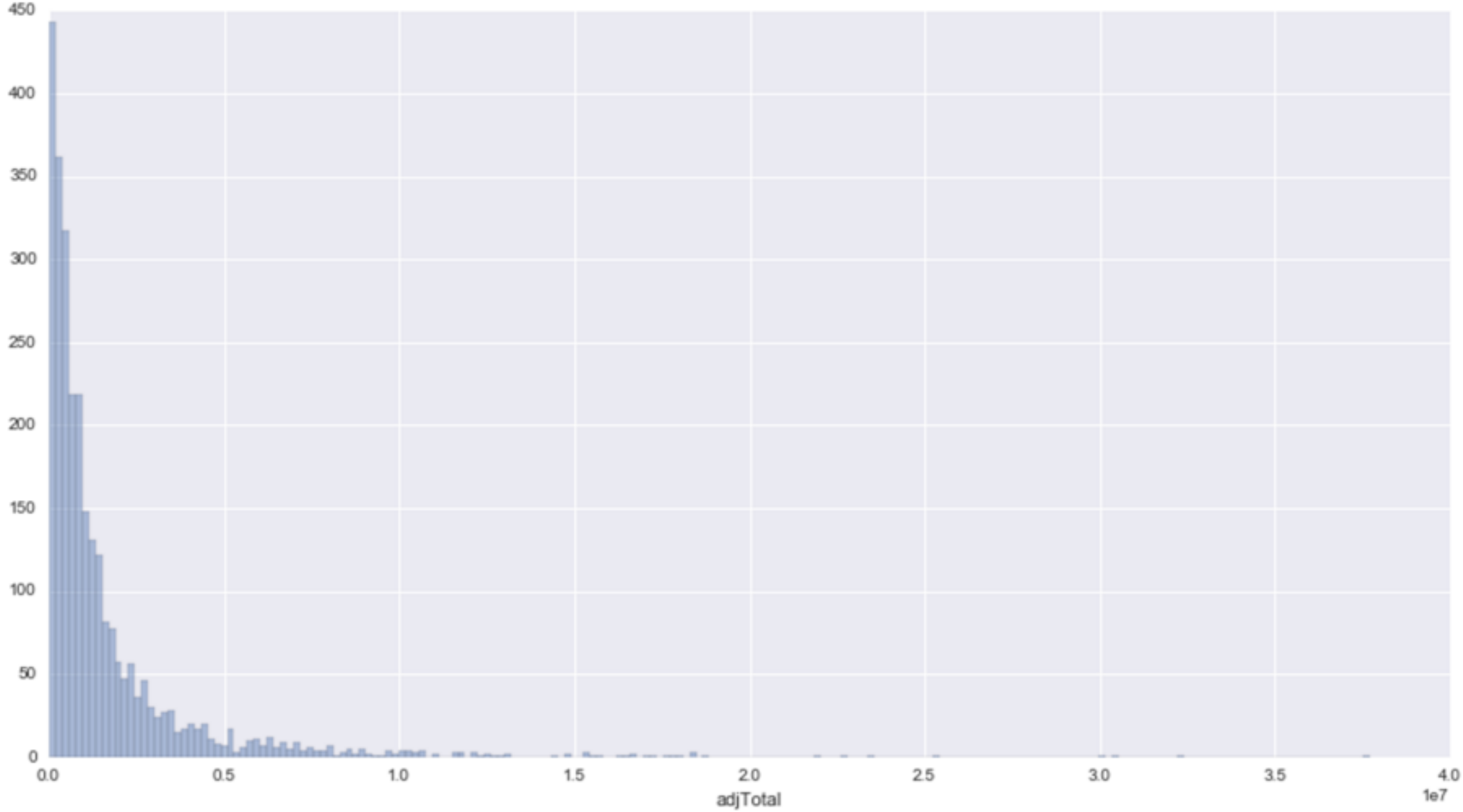
- response before and after ratio adjustment



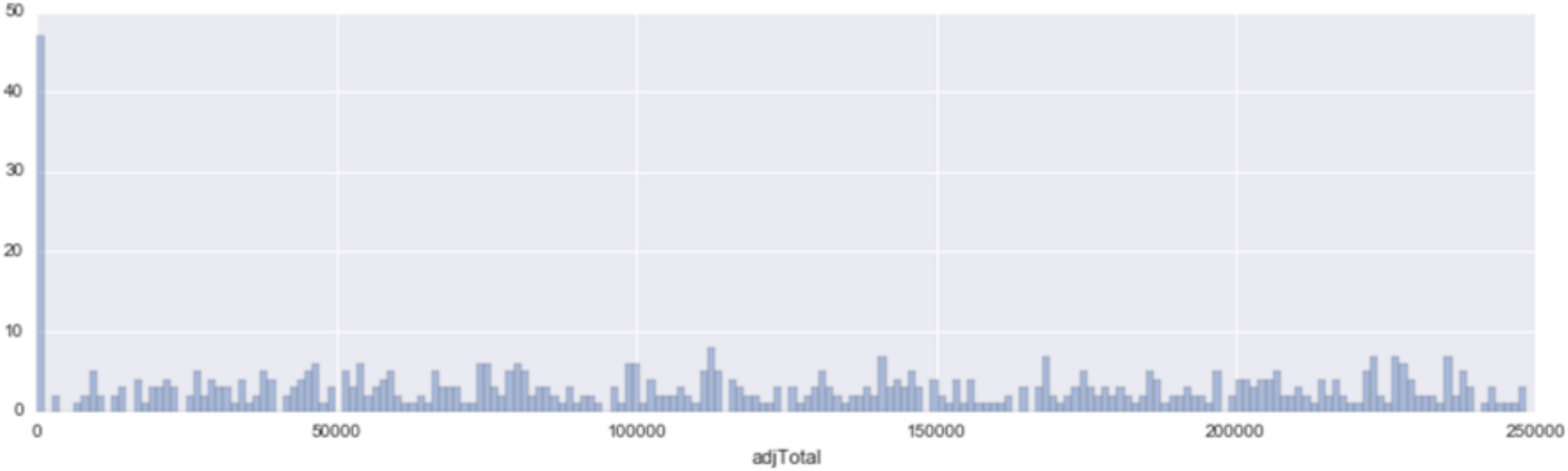
DISTRIBUTION OF PAYMENTS PER ZIP CODE

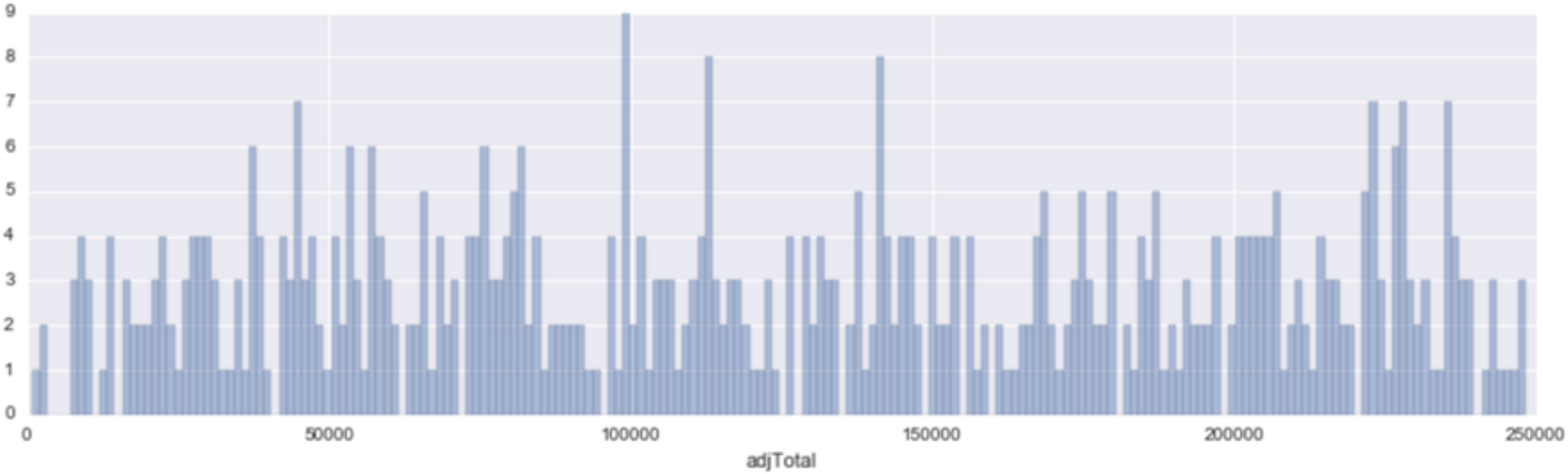


DISTRIBUTION OF PAYMENTS PER ZIP CODE

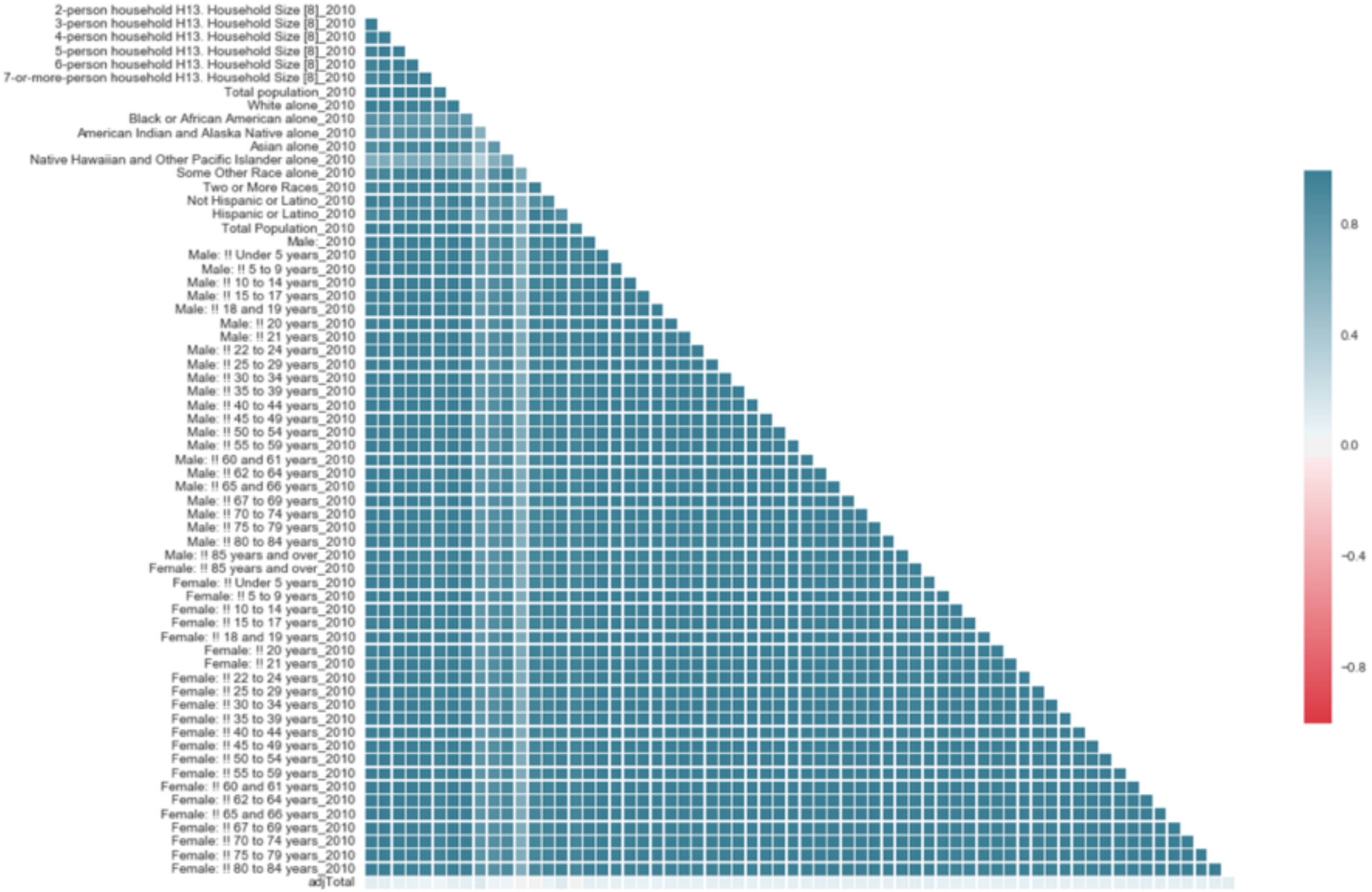


COUNTS OF ZIP CODES WITH PAYMENTS < 250,001

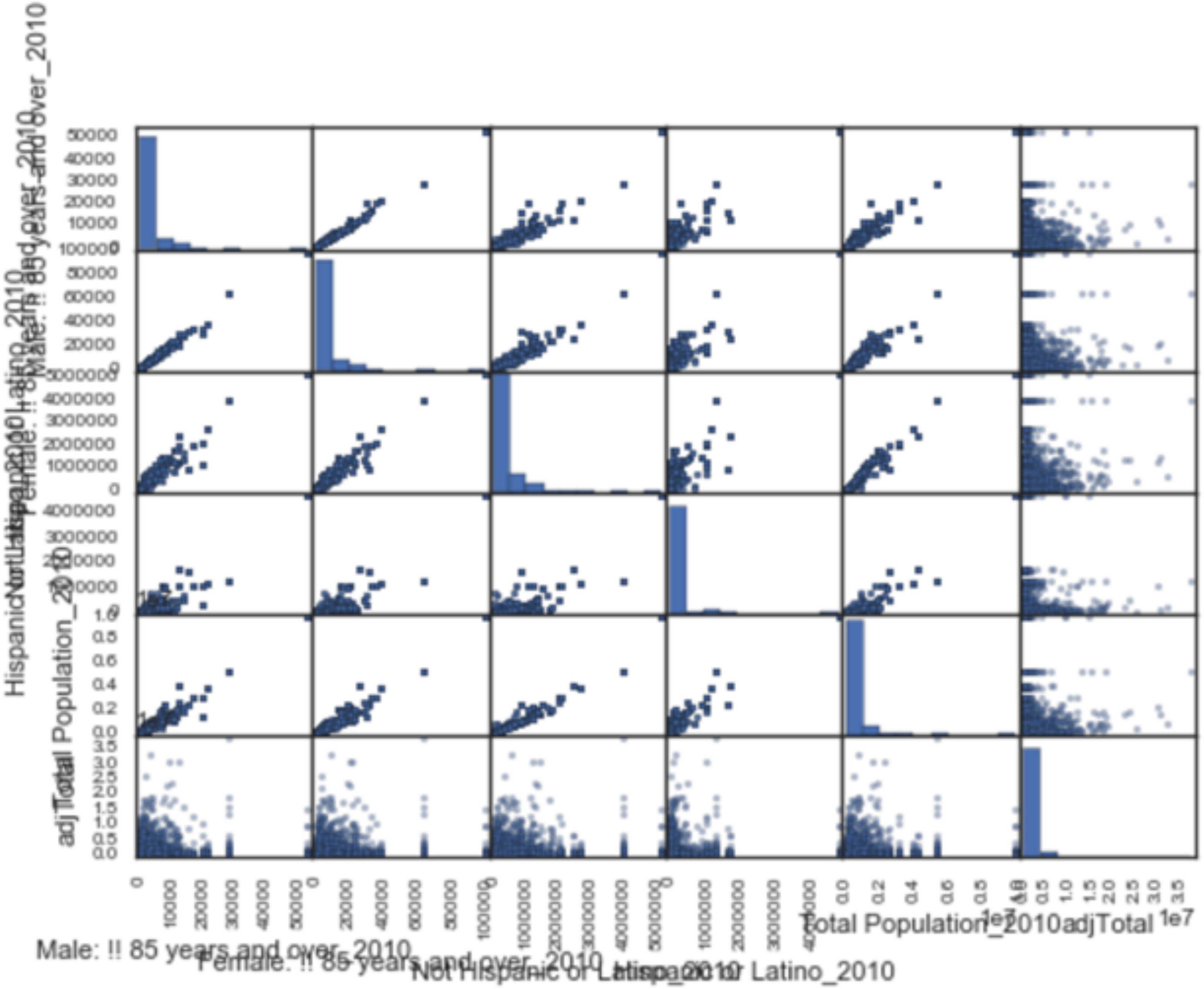




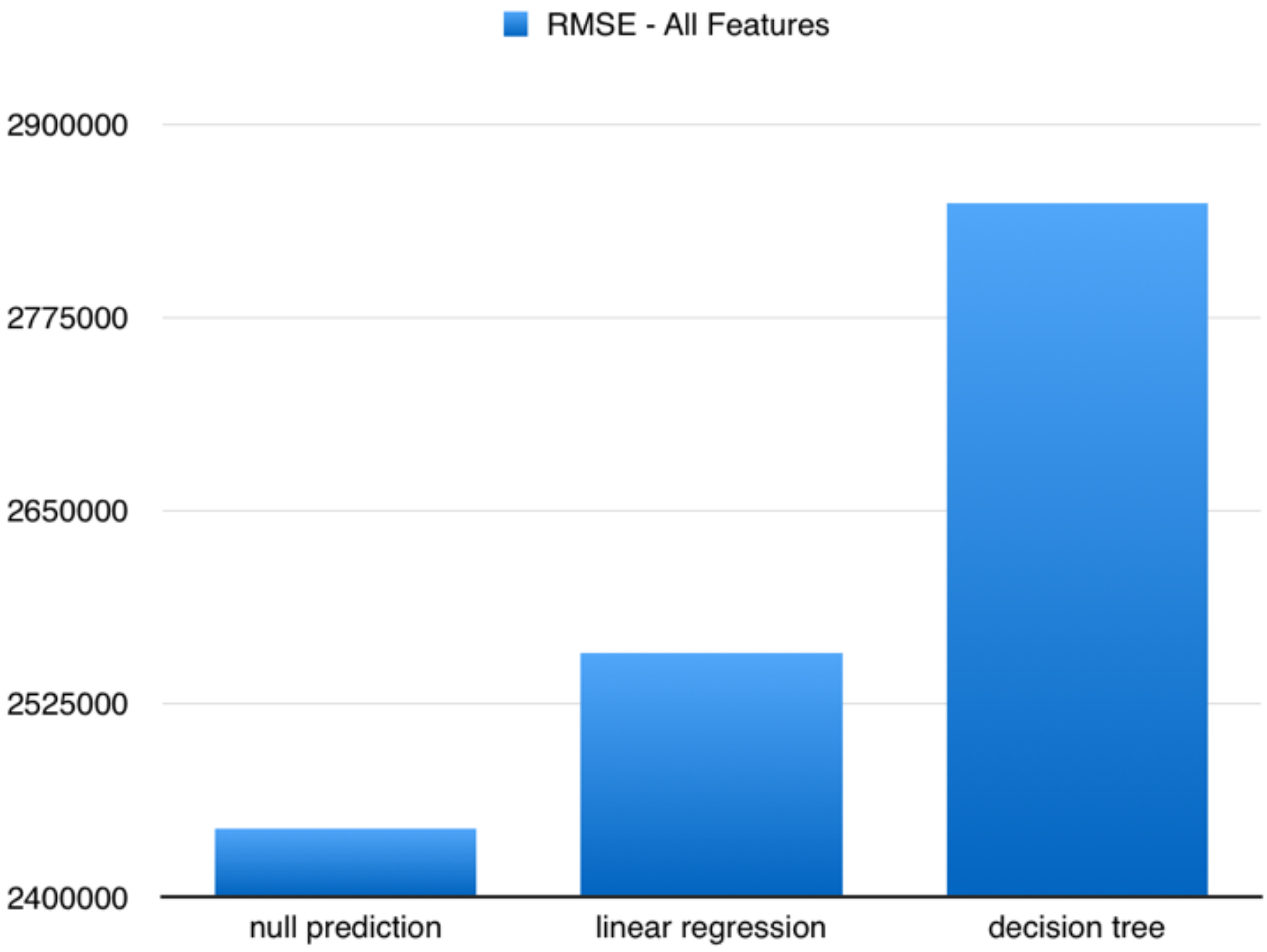
CORRELATION MATRIX



SCATTER MATRIX



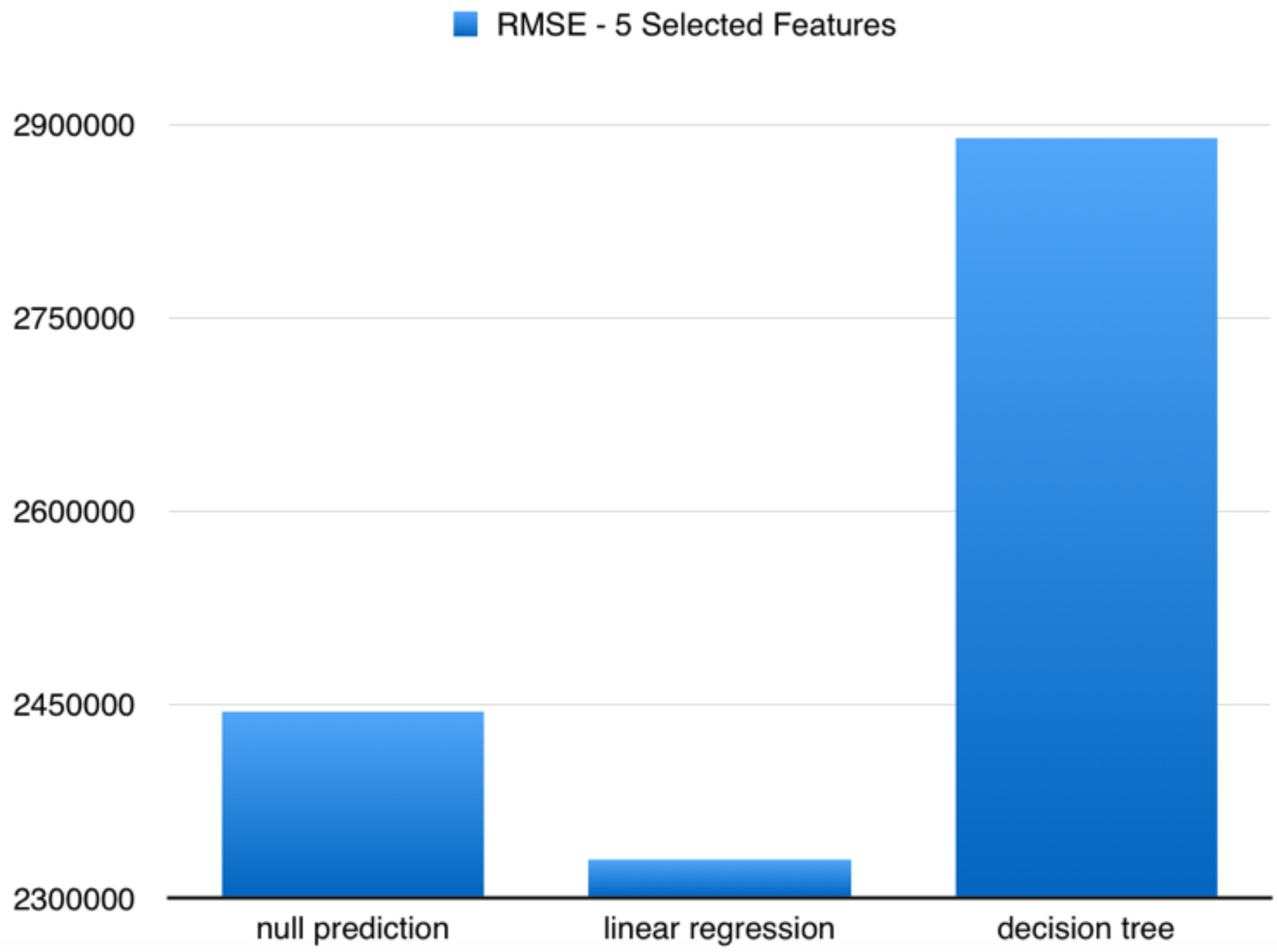
REGRESSION – ALL 64 FEATURES



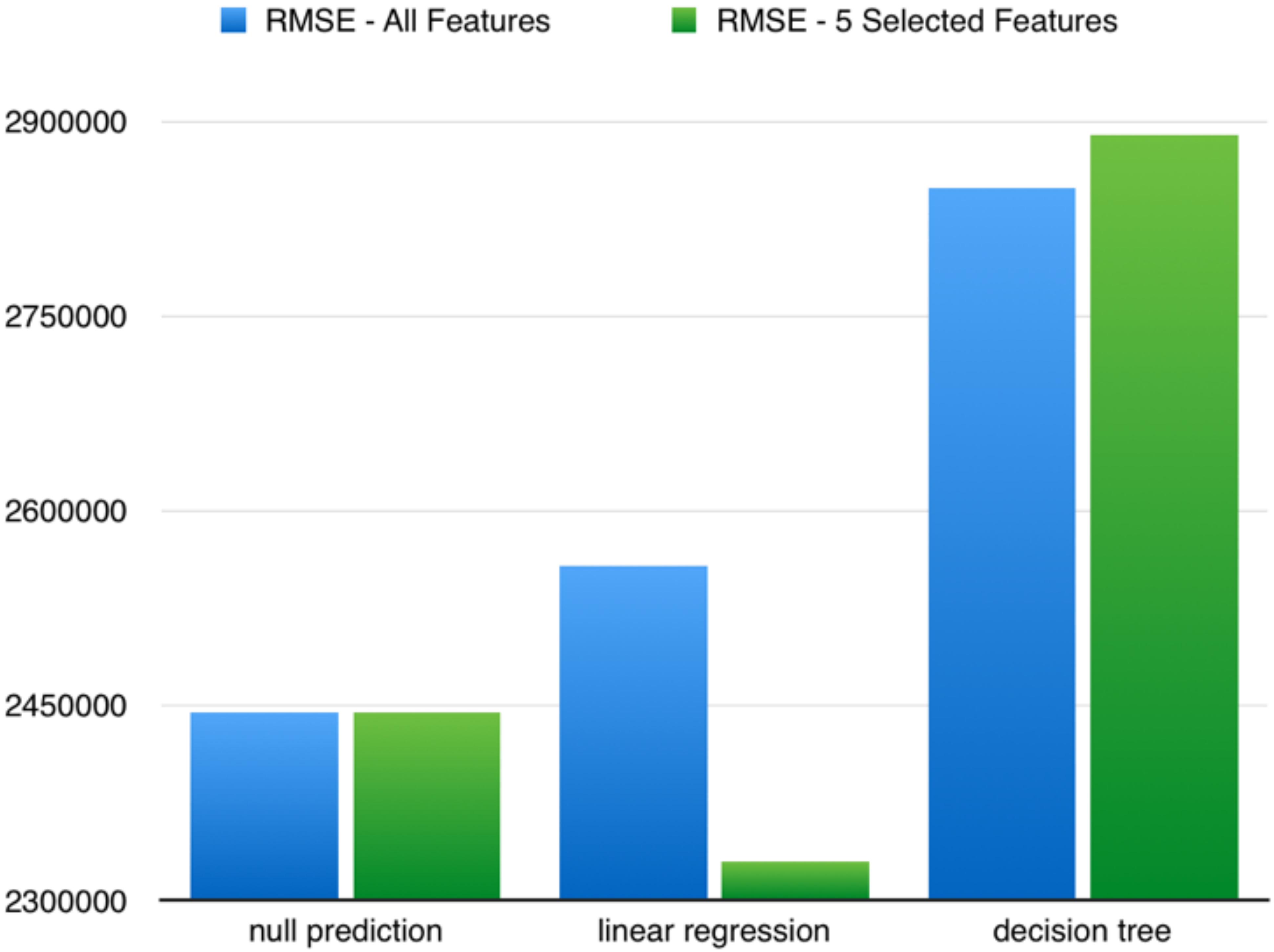
REGRESSION – 5 SELECTED FEATURES

- 'Male: !! 85 years and over_2010',
- 'Female: !! 85 years and over_2010',
- 'Not Hispanic or Latino_2010',
- 'Hispanic or Latino_2010',
- 'Total Population_2010’

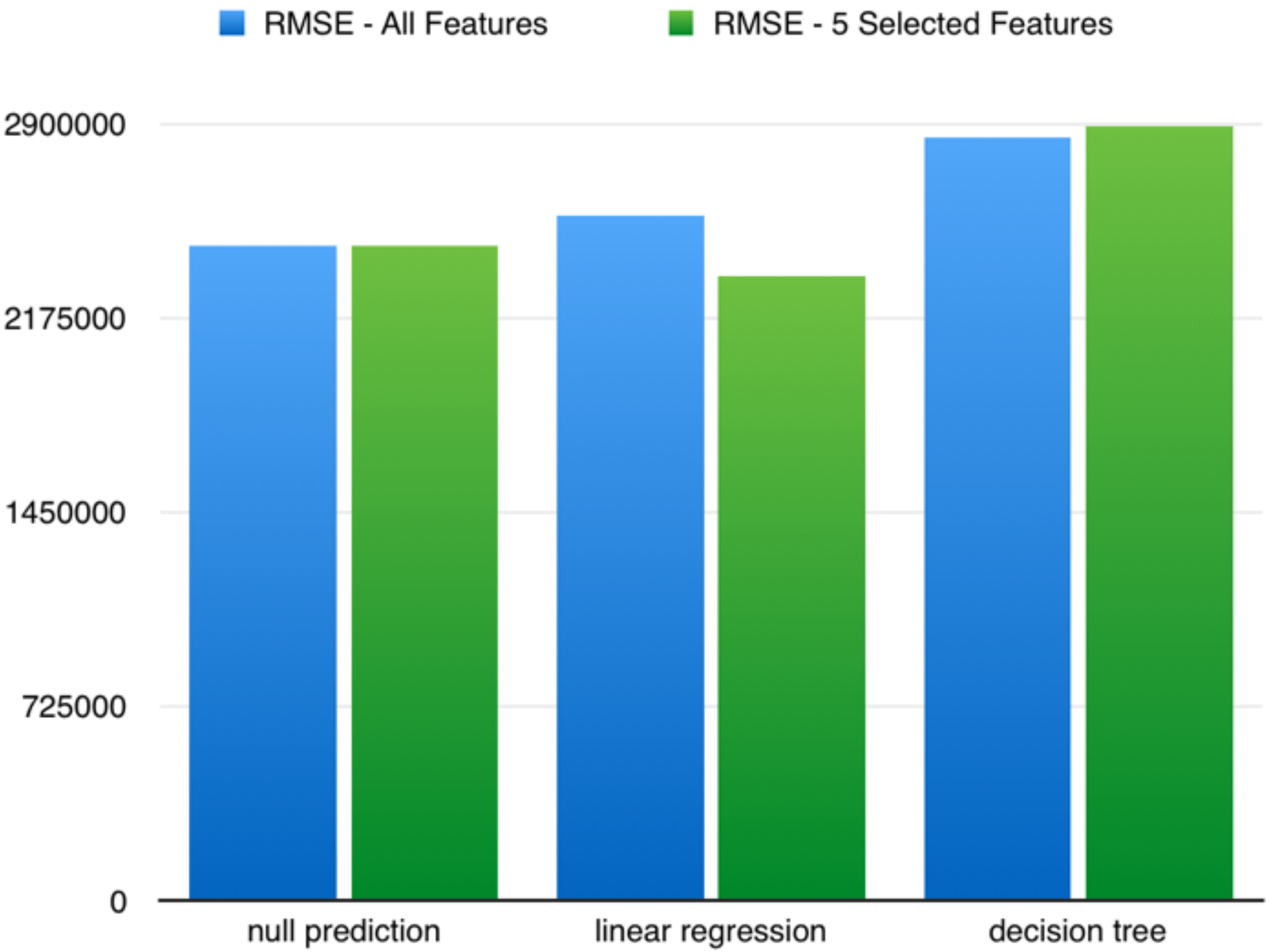
REGRESSION – 5 SELECTED FEATURES



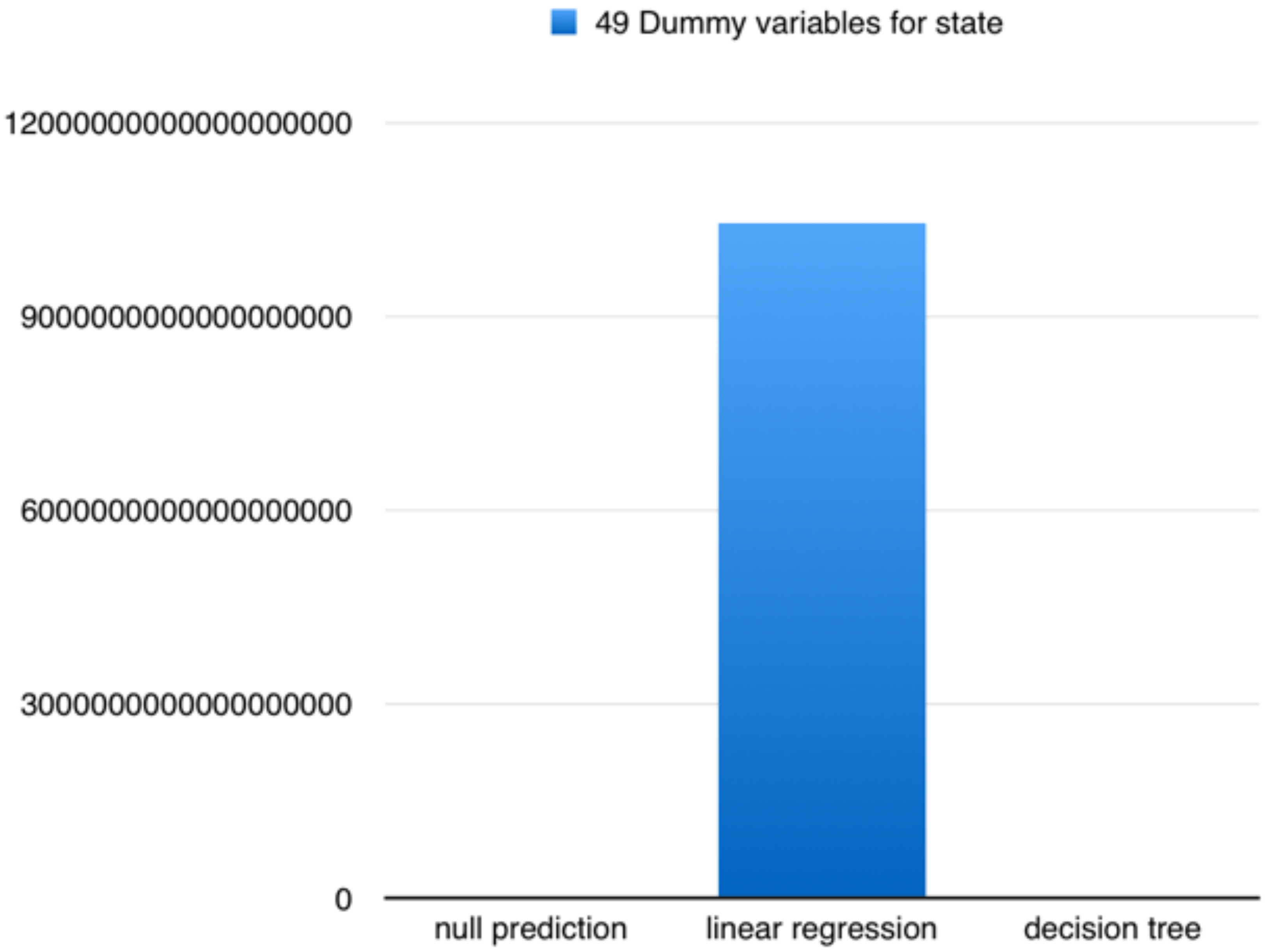
REGRESSION COMPARISON



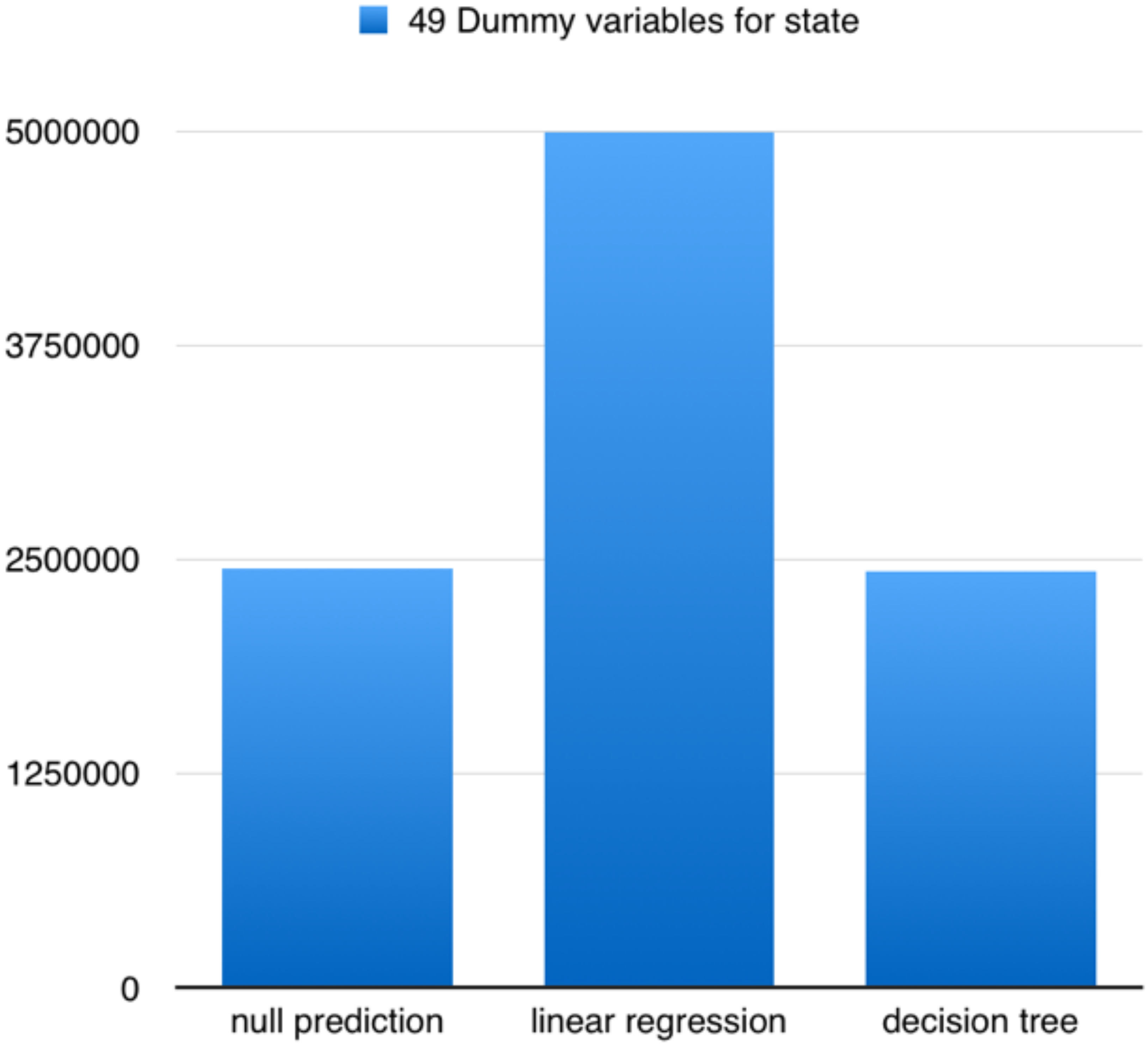
REGRESSION COMPARISON



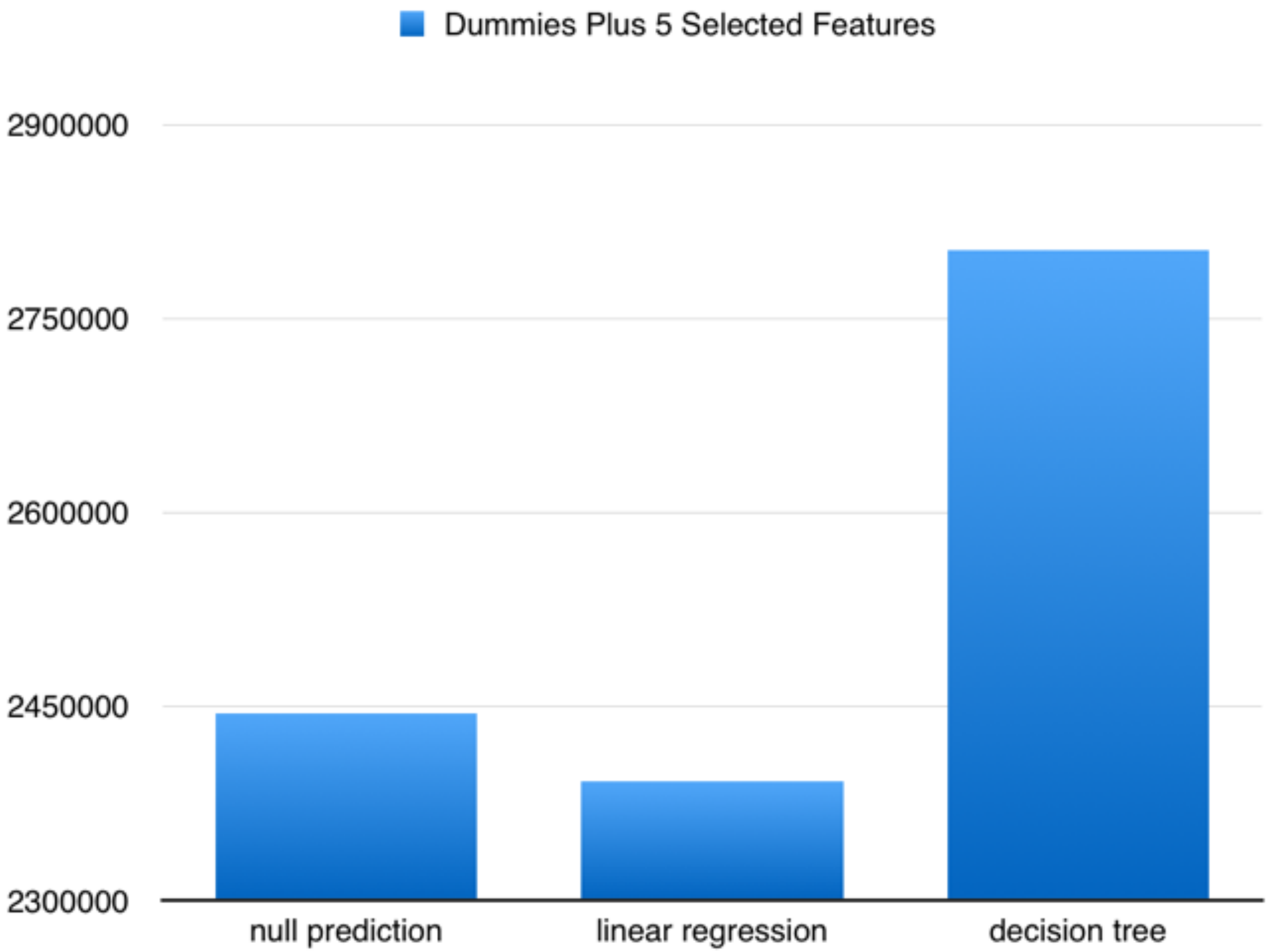
FEATURE ENGINEERING – STATE DUMMY VARIABLES



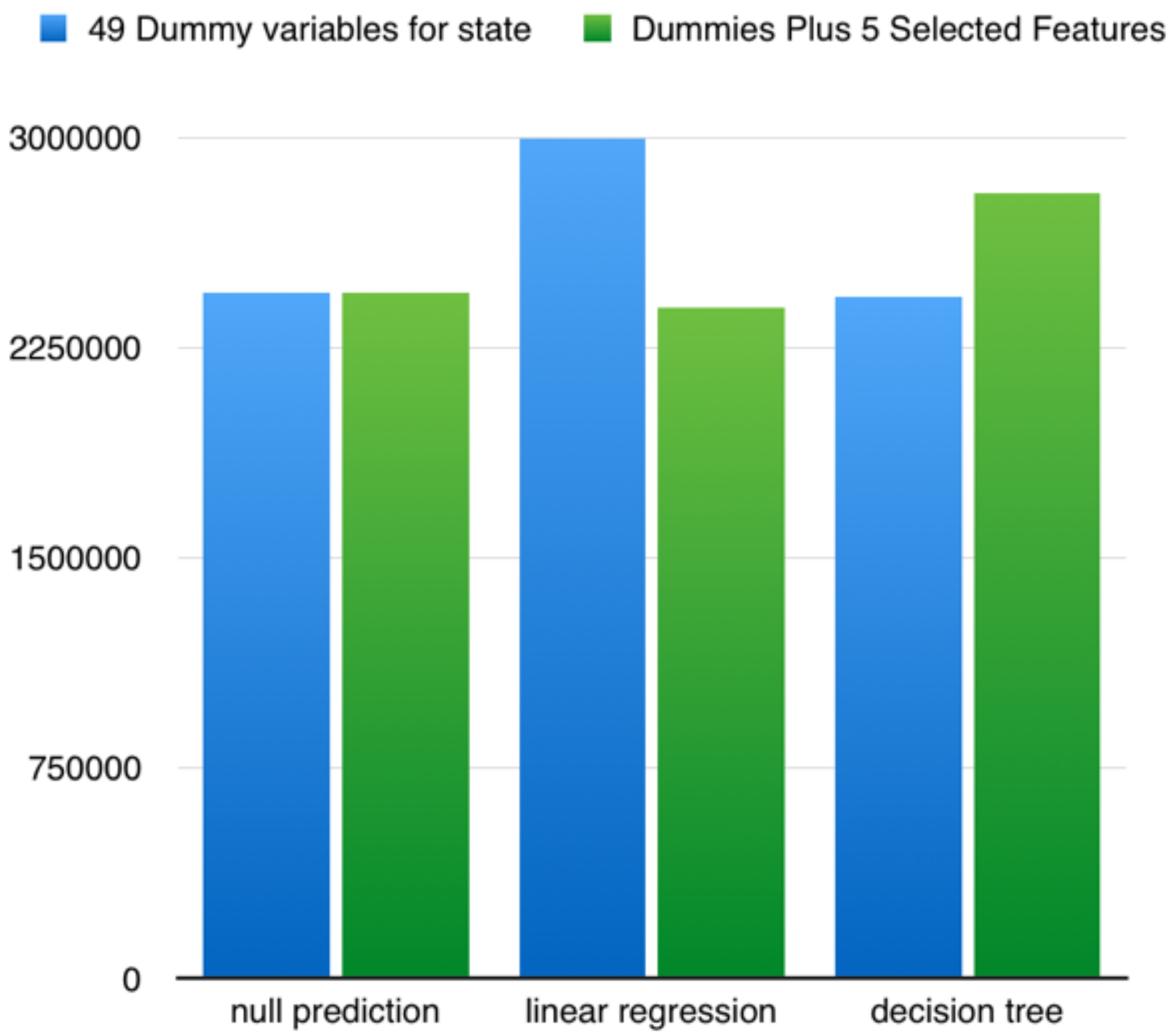
FEATURE ENGINEERING - STATE DUMMY VARIABLES



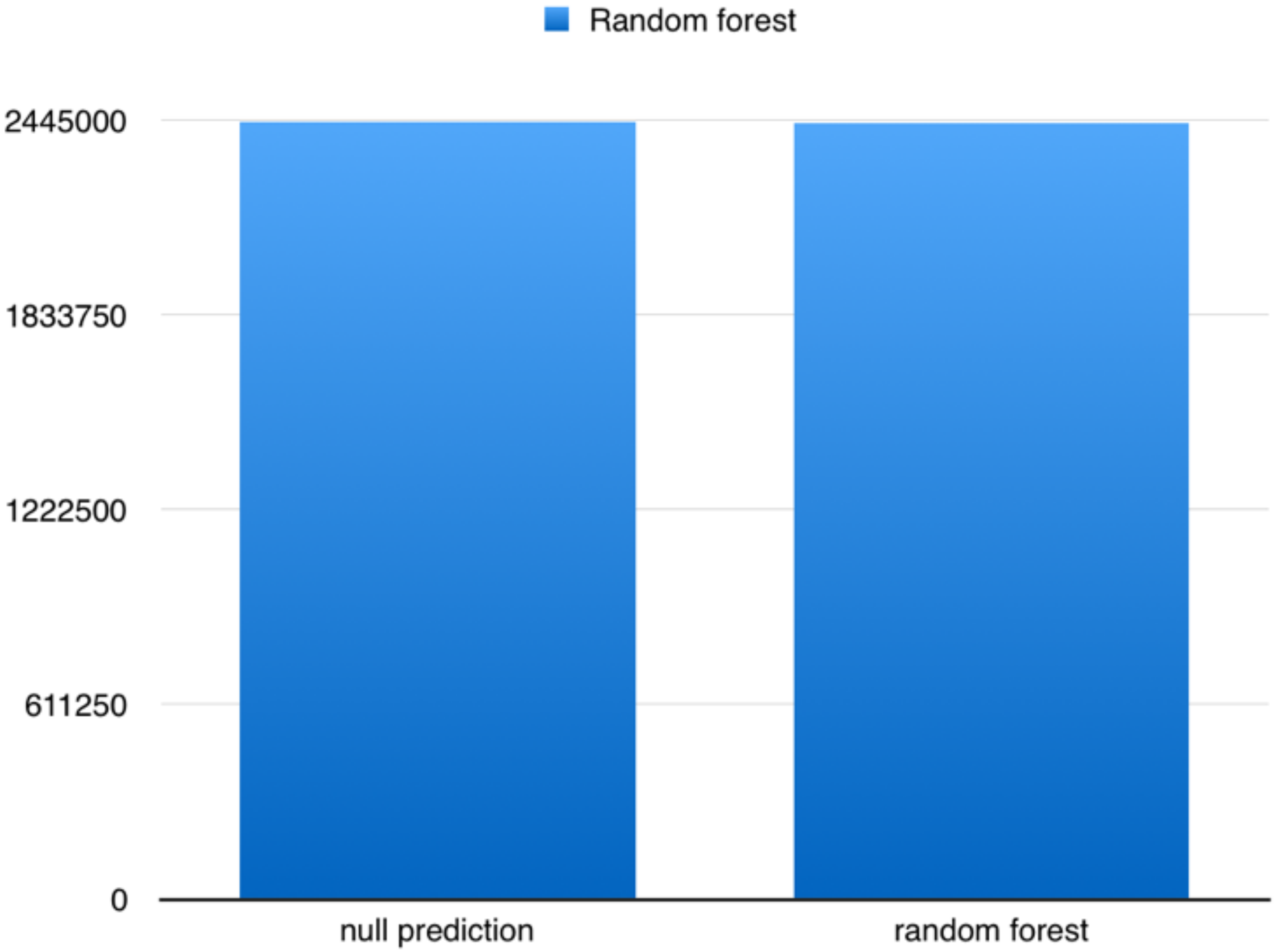
FEATURE ENGINEERING – DUMMIES PLUS 5



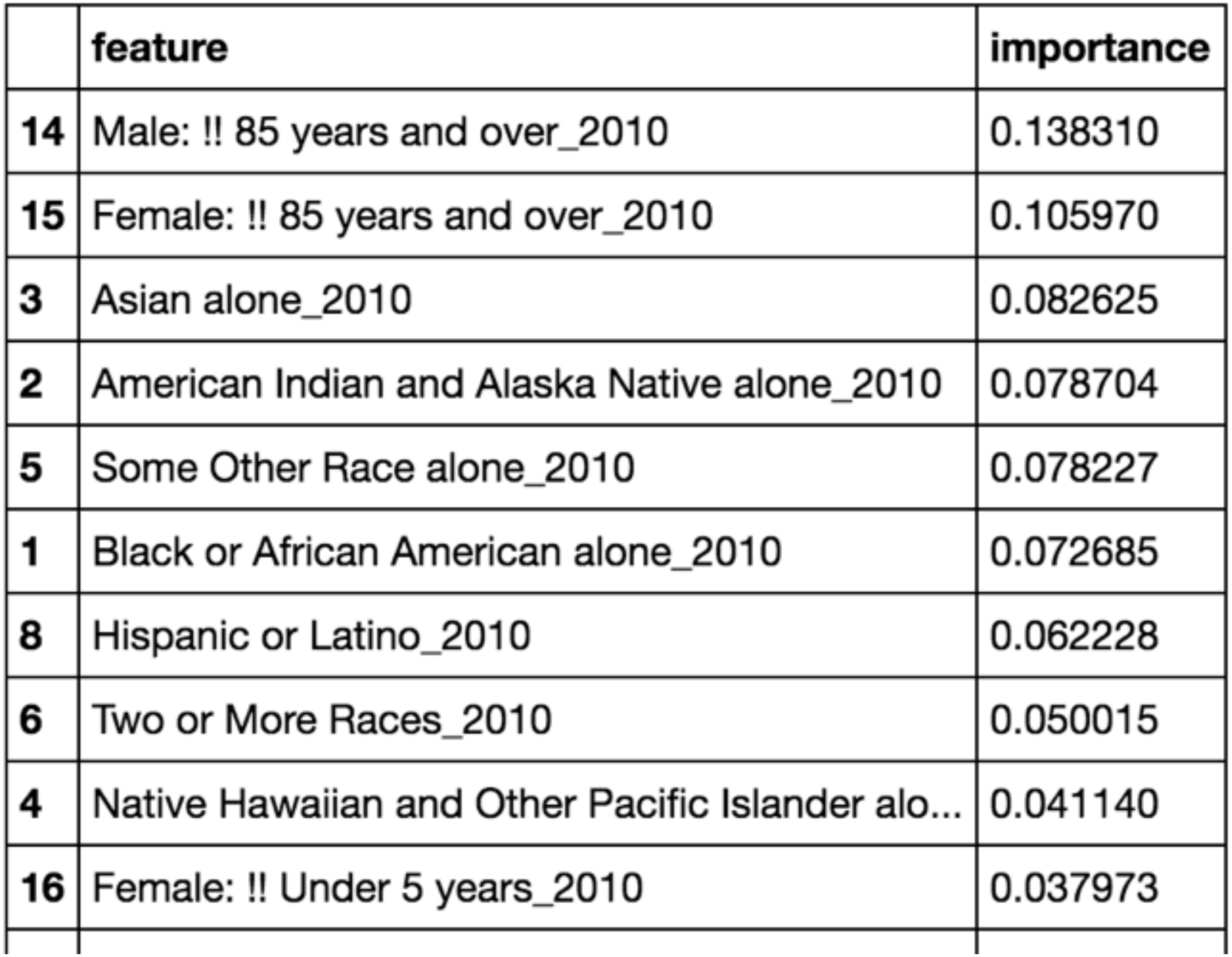
FEATURE ENGINEERING - STATE DUMMY PLUS 5 COMPARISON



RANDOM FOREST

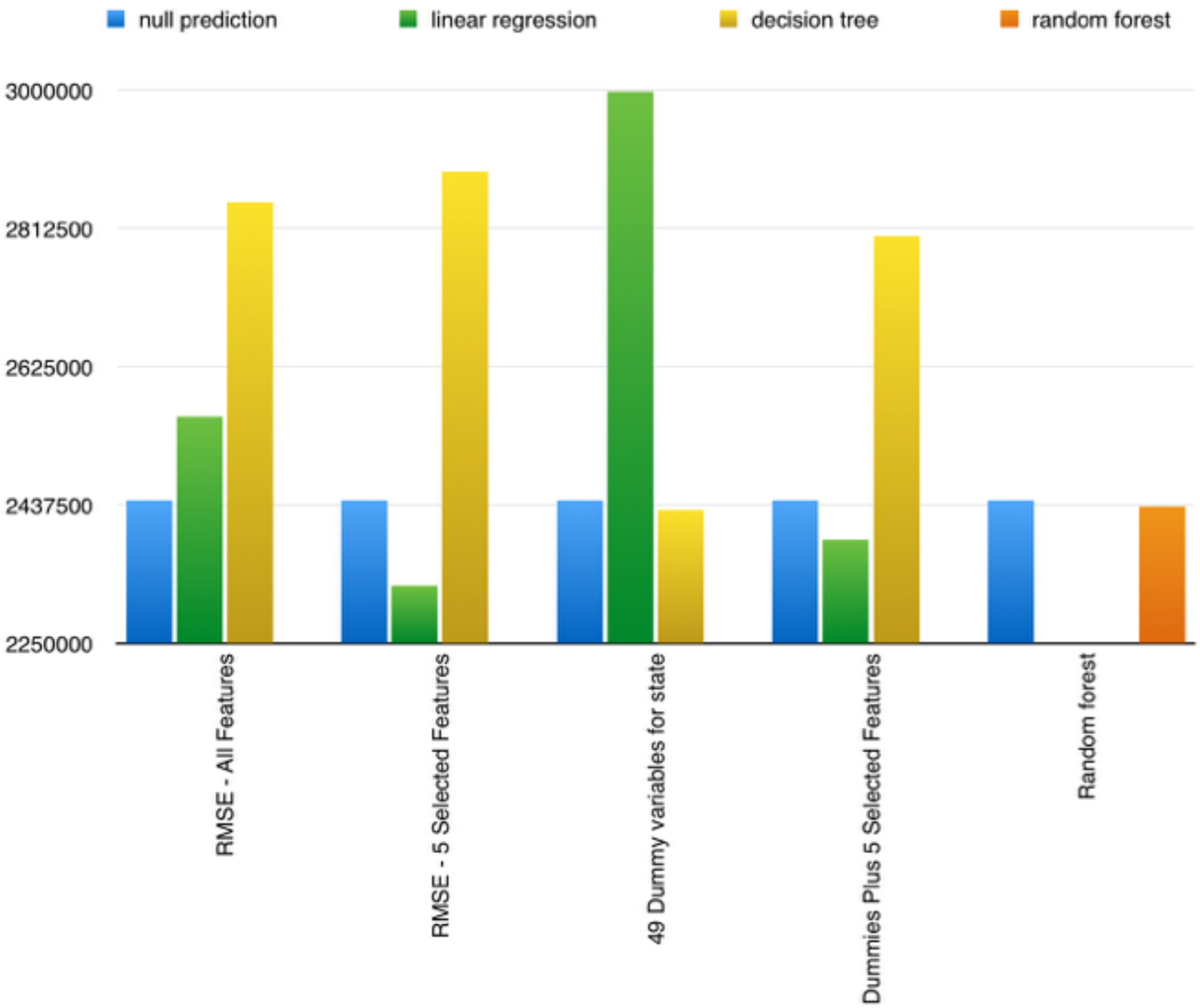


RANDOM FOREST- IMPORTANCE



	feature	importance
14	Male: !! 85 years and over_2010	0.138310
15	Female: !! 85 years and over_2010	0.105970
3	Asian alone_2010	0.082625
2	American Indian and Alaska Native alone_2010	0.078704
5	Some Other Race alone_2010	0.078227
1	Black or African American alone_2010	0.072685
8	Hispanic or Latino_2010	0.062228
6	Two or More Races_2010	0.050015
4	Native Hawaiian and Other Pacific Islander alo...	0.041140
16	Female: !! Under 5 years_2010	0.037973

COMPARISON OF ALL MODELS



EXTENSIONS

- get more data - try adding employment and population density per zip code as features
- try PCA feature reduction
- Attempt support vector regression modeling
- Engineer more features
- Switch problem to classification - high/low cost

fin