Matthew Peetz

Regis University

MSDS 621

Week 5 Lab: EDA and Visualizations

This week's assignment will be combining many of the concepts from our prior lectures.

Below, I've loaded the immunization dataset from week 3 (nispuf14.csv). You can feel free to use your output from week3 or the verse I have provided in the assign_wk5 folder.

Here is what I've demonstrated below with the immunization dataset.

- separated continuous and categorical columns
- minimally handled NaNs
- built a baseline RandomForestClassifier with a random column

I've selected FRSTBRN as my target variable target variable -- not a very interesting column, but it had no missing values. Do not worry, you will have a chance to improve on my work.

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# Makes graphs prettier
sns.set()
# Magic command for plots
%matplotlib inline
```

Note:: There are a couple of new parameters I'm using with read_csv. I encourage you to go look those up and understand what they do.

Load the data

In [2]:
```
df = pd.read_csv("assign_wk5/nispuf14.csv", na_values=['.'], low_memory=False)
```

In [3]:
```
df.head()
```

Out[3]:

	SEQNUMC	SEQNUMHH	PDAT	PROVWT_D	PROVWT_D_TERR	RDDWT_D	F
0	11	1	2	.	.	218.30024855484000	
1	21	2	1	806.84601169505000	806.84601169505000	454.86041741251200	
2	31	3	2	.	.	30.54542540283290	
3	41	4	1	63.44868567610260	63.44868567610260	36.96593137368630	
4	51	5	1	94.87263225744540	94.87263225744540	64.62020426239790	

5 rows × 461 columns

Pop Quiz:: What am I doing in the cell below?

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24897 entries, 0 to 24896
Columns: 461 entries, SEQNUMC to INS_11
dtypes: float64(220), int64(29), object(212)
memory usage: 87.6+ MB
```

Answer: All of the various formats that can be numeric are being made numeric, with the call to prevent it from telling you have the cells that can not be changed into numbers not throwing a warning.

In [5]:
```
for col in df.columns:
    df[col] = pd.to_numeric(df[col], errors='coerce')
```

In [6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24897 entries, 0 to 24896
Columns: 461 entries, SEQNUMC to INS_11
dtypes: float64(432), int64(29)
memory usage: 87.6 MB
```

In [7]:
```
df.head()
```

Out[7]:

	SEQNUMC	SEQNUMHH	PDAT	PROVWT_D	PROVWT_D_TERR	RDDWT_D	RDDWT_D_TERR	ST
0	11	1	2	NaN	NaN	218.300249	218.300249	
1	21	2	1	806.846012	806.846012	454.860417	454.860417	
2	31	3	2	NaN	NaN	30.545425	30.545425	
3	41	4	1	63.448686	63.448686	36.965931	36.965931	
4	51	5	1	94.872632	94.872632	64.620204	64.620204	

5 rows × 461 columns

I like to make a copy of my dataset before I start manipulating the data. This is a personal preference.

In [8]:
```
df_copy = df.copy()
```

In [9]:
```
!pip install PyPDF2
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: PyPDF2 in c:\users\matth\appdata\roaming\python\pyth
on310\site-packages (3.0.1)
```

Seperate Categorical vs Continuous Variables

In [10]:
```
from PyPDF2 import PdfReader
#from PyPdf2.pdf import Destination  # read the pdf file
```

In [11]:
```
pdf_file = 'assign_wk5/NISPUF14_CODEBOOK.PDF'
```

In [12]:
```
reader = PdfReader(pdf_file)
outlines = reader.outline
```

I notice the PDF's outline tells us which variables are continuous. I would like to be able to read that outline and get those variables. Unfortunately, the PDF outline format is not terribly conducive to this operation.

As you can see, in the outlne "tree", when we see the text "Continuous Statistics", we need the **previous** entry, but PDF outlines are weird mixes of dictionaries and lists. I finally found the following answer on StackOverflow that satisfied the need.

From https://stackoverflow.com/questions/1011938/python-previous-and-next-values-inside-a-loop

In [13]:

```
from itertools import tee, islice, chain

def previous_and_next(some_iterable):
    prevs, items, nexts = tee(some_iterable, 3)
    prevs = chain([None], prevs)
    nexts = chain(islice(nexts, 1, None), [None])
    return zip(prevs, items, nexts)
```

Testing if I could find the word "Contnuous" in the PDF outline.

In [14]:
```
'Continuous' in outlines[7][0].title
```

Out[14]:
```
True
```

In [15]:
```
cont_list = []
for prev, item, nxt in previous_and_next(outlines):
    if isinstance(item, list) :
        if 'Continuous' in str(item[0].title):
            cont_list.append(prev.title)
```

In [16]:
```
cont_list[:5]
```

Out[16]:
```
['PROVWT_D', 'PROVWT_D_TERR', 'RDDWT_D', 'RDDWT_D_TERR', 'BF_ENDR06']
```

In [17]:
```
cat_list = [c for c in df_copy.columns if c not in cont_list]
```

Cleaning up the Missing Values

I'm not going to do a lot to handle the missing values in this demo. **However**, I will expect you to do more than I do.

I'll simply:

- for categories, fill with -999
- for continuous, fill with mean

In [18]:
```
for col in cat_list:
    df_copy[col].fillna(-999, inplace=True)
```

In [19]:
```
for col in cont_list:
    df_copy[col].fillna(df[col].mean(), inplace=True)
```

In [20]:
```
df_copy.head()
```

Out[20]:

	SEQNUMC	SEQNUMHH	PDAT	PROVWT_D	PROVWT_D_TERR	RDDWT_D	RDDWT_D_TERR	ST
0	11	1	2	383.438950	382.821312	218.300249	218.300249	
1	21	2	1	806.846012	806.846012	454.860417	454.860417	
2	31	3	2	383.438950	382.821312	30.545425	30.545425	
3	41	4	1	63.448686	63.448686	36.965931	36.965931	
4	51	5	1	94.872632	94.872632	64.620204	64.620204	

5 rows × 461 columns

Random Forest Classifier Benchmark

My target is going to be FRSTBRN -- FIRST BORN STATUS OF CHILD.

FRSTBRN is a categorical variable (1 - No, 2 - Yes), so I will use the classifier version of RandomForest. If your target variable is continuous, you will need to use the regressor version of this algorithm.

Take a look at:
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

In [21]:
```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
```

Pop Quiz:: Do you know what the following does?

Answer: The computer uses a random seed for splitting the test and train set. This means that each time the program is run a different set is created. If we want to be able to replicated our results the seed mus be set. Which number should be used. 42, of course.

See Hitch Hikers Guide to the Galaxy, Douglas Adams

In [22]:
```
np.random.seed(42)
```

In [23]:
```
y = df_copy['FRSTBRN']
X = df_copy.drop('FRSTBRN', axis=1)
```

I'm going to get the min and max values from one of the continuous variables to use as the min/max for the random column.

In [24]:
```
the_min = X.PROVWT_D.min()
the_max = X.PROVWT_D.max()
X['random'] = np.random.normal(the_min, the_max, size=X.shape[0])
```

In [25]:
```
x_train,x_test,y_train,y_test=train_test_split(X,y,test_size=0.3)
```

In [26]:
```
clf=RandomForestClassifier(n_estimators=100)
```

In [27]:
```
clf.fit(x_train,y_train)
```

Out[27]:
```
▾ RandomForestClassifier
RandomForestClassifier()
```

I'm going to check the accuracy of my model. Ultimately we want to use this model to help us learn something about our data. So, if our model's accuracy is low, then our "learning" with have a low level of confidence too.

In [28]:
```
y_pred=clf.predict(x_test)
```

In [29]:
```
from sklearn import metrics
```

In [30]:
```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8832663989290496
```

A model with a 88% accuracy is really good. One technique you can use to increase your accuracy is scale/normalize your data. The reasoning behind this is that algorithms like Random Forest tend to "favor" columns with a large range of values. Scaling/Normalizing your data will reduce this bias effect in your algorithm.

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html

Feature Selection: Random Number Trick

In [31]:
```
features = x_train.columns
importances = clf.feature_importances_
std = np.std([tree.feature_importances_ for tree in clf.estimators_],
             axis=0)
indices = np.argsort(importances)[::-1]


feature_rank = []
# Print the feature ranking
print("Feature ranking:")

for f in range(x_train.shape[1]):
    feature = f"{f + 1}. feature {features[indices[f]]}   \t{importances[indices[f]
    if 'random' in features[indices[f]]:
        feature += " <=="
    print(feature)
    feature_rank.append([features[indices[f]], importances[indices[f]]] )
```

```
Feature ranking:
1. feature CHILDNM      27.04%
2. feature C1R          15.07%
3. feature random       1.72% <==
4. feature SEQNUMC      1.61%
5. feature RDDWT_D_TERR         1.59%
6. feature SEQNUMHH     1.58%
7. feature RDDWT_D      1.53%
8. feature STRATUM      1.46%
9. feature INCPORAR     1.41%
10. feature M_AGEGRP    1.32%
11. feature BF_ENDR06           1.27%
12. feature EST_GRANT           1.27%
13. feature ESTIAP14    1.24%
14. feature STATE       1.18%
15. feature BF_FORMR08          1.06%
16. feature INCQ298A    1.02%
17. feature BF_EXCLR06          0.99%
18. feature NUM_CELLS_HH        0.74%
19. feature AGEGRP      0.70%
20. feature EDUC1       0.66%
21. feature D6R         0.61%
22. feature PROVWT_D    0.60%
23. feature PROVWT_D_TERR       0.58%
24. feature INCPOV1     0.58%
25. feature C5R         0.56%
26. feature DDTP4       0.51%
27. feature CEN_REG     0.51%
28. feature INTRP       0.50%
29. feature DVRC1       0.49%
30. feature NUM_CELLS_PARENTS           0.49%
31. feature DHEPB2      0.48%
32. feature DHIB3       0.48%
33. feature DDTP3       0.48%
34. feature DFLU2       0.48%
35. feature DHIB4       0.48%
36. feature DPCV4       0.48%
37. feature RACEETHK    0.47%
38. feature DHEPB3      0.47%
39. feature DHEPA2      0.47%
40. feature DMMR1       0.46%
41. feature DHEPA1      0.45%
42. feature DFLU1       0.45%
43. feature DPOLIO3     0.45%
44. feature DDTP2       0.44%
45. feature DPOLIO2     0.44%
46. feature DPCV3       0.44%
47. feature RENT_OWN    0.44%
48. feature DDTP1       0.43%
49. feature DHIB1       0.42%
50. feature DHIB2       0.42%
51. feature DPCV2       0.42%
52. feature MARITAL2    0.41%
53. feature DROT2       0.41%
54. feature CWIC_02     0.40%
55. feature DROT1       0.39%
56. feature DFLU3       0.38%
57. feature DPOLIO1     0.38%
58. feature RACE_K      0.38%
```

```
59. feature DPCV1       0.37%
60. feature SEX         0.35%
61. feature NUM_PHONE           0.34%
62. feature DROT3       0.34%
63. feature FLU1_AGE    0.31%
64. feature FLU2_AGE    0.30%
65. feature FLU3_AGE    0.30%
66. feature DHEPB1      0.29%
67. feature HEA2_AGE    0.29%
68. feature INS_1       0.29%
69. feature DTP4_AGE    0.29%
70. feature CWIC_01     0.26%
71. feature HEP3_AGE    0.26%
72. feature INS_2       0.26%
73. feature HIB4_AGE    0.25%
74. feature HEA1_AGE    0.25%
75. feature INS_3       0.24%
76. feature PCV4_AGE    0.24%
77. feature P_NUMFLU    0.23%
78. feature INS_4_5     0.22%
79. feature INS_11      0.22%
80. feature CBF_01      0.22%
81. feature MOBIL_I     0.22%
82. feature INS_3A      0.21%
83. feature I_HISP_K    0.21%
84. feature MMR1_AGE    0.21%
85. feature INS_6       0.20%
86. feature HIB3_AGE    0.20%
87. feature P_NUMFLUN           0.20%
88. feature VRC1_AGE    0.20%
89. feature DHEPB4      0.19%
90. feature LANGUAGE    0.19%
91. feature HEP2_AGE    0.19%
92. feature P_NUHIBX    0.19%
93. feature DFLU4       0.18%
94. feature DTP3_AGE    0.18%
95. feature PROV_FAC    0.18%
96. feature P_NUMHS     0.17%
97. feature P_NUMDTA    0.17%
98. feature P_NUMDIH    0.16%
99. feature REGISTRY    0.16%
100. feature DPOLIO4    0.15%
101. feature D7         0.15%
102. feature POL3_AGE           0.15%
103. feature PCV3_AGE           0.15%
104. feature FLU4_AGE           0.15%
105. feature P_NUHEPX           0.14%
106. feature P_NUMHM    0.14%
107. feature P_NUMHEA           0.14%
108. feature POL4_AGE           0.13%
109. feature HIB2_AGE           0.13%
110. feature POL2_AGE           0.13%
111. feature DTP2_AGE           0.13%
112. feature VFC_ORDER          0.13%
113. feature ROT3_AGE           0.13%
114. feature P_NUMDHI           0.12%
115. feature ROT2_AGE           0.12%
116. feature P_NUMIPV           0.12%
117. feature PCV2_AGE           0.12%
```

```
118. feature XPOLTY3    0.12%
119. feature HEP4_AGE           0.11%
120. feature ROT1_AGE           0.11%
121. feature XPOLTY1    0.11%
122. feature P_NUMRM    0.11%
123. feature N_PRVR     0.11%
124. feature PCV1_AGE           0.11%
125. feature P_NUMHEP           0.10%
126. feature DTP1_AGE           0.10%
127. feature P_NUMROT           0.10%
128. feature XPOLTY2    0.10%
129. feature HEP1_AGE           0.10%
130. feature XDTPTY1    0.10%
131. feature POL1_AGE           0.10%
132. feature P_NUMRG    0.10%
133. feature XDTPTY2    0.09%
134. feature U1D_HEP    0.09%
135. feature HIB1_AGE           0.09%
136. feature XDTPTY4    0.09%
137. feature XDTPTY3    0.08%
138. feature P_NUMPCC13         0.08%
139. feature P_UTDHEPA2         0.08%
140. feature AGECPOXR           0.08%
141. feature HAD_CPOX           0.08%
142. feature P_NUMHIB           0.08%
143. feature P_NUMHIN           0.08%
144. feature XHEPTY2    0.08%
145. feature P_NUMPOL           0.08%
146. feature XHEPTY3    0.07%
147. feature HEP_BRTH           0.07%
148. feature P_NUMPCC           0.07%
149. feature U3D_HEP    0.07%
150. feature P_NUMFLUM          0.07%
151. feature U2D_HEP    0.07%
152. feature XHEPTY4    0.06%
153. feature P_NUMVRX           0.06%
154. feature P_NUMPCV           0.06%
155. feature P_NUMMMRX          0.06%
156. feature XHEPTY1    0.05%
157. feature P_UTDROT_S         0.05%
158. feature P_NUMDTP           0.05%
159. feature P_UTD431H314_ROUT_S        0.05%
160. feature P_UTD431H313_ROUT_S        0.04%
161. feature P_UTD431           0.04%
162. feature XMMRTY1    0.04%
163. feature P_UTD431H31_ROUT_S         0.04%
164. feature P_NUMVRC           0.04%
165. feature P_NUMRO    0.04%
166. feature P_NUMMMR           0.04%
167. feature P_UTD431H_ROUT_S           0.04%
168. feature PU431_314          0.04%
169. feature P_NUMMRV           0.04%
170. feature XPCVTY4    0.04%
171. feature P_UTDHEPA1         0.04%
172. feature XPCVTY1    0.04%
173. feature P_UTDHIB_ROUT_S    0.04%
174. feature P_NUMPCN           0.04%
175. feature P_UTDTP4           0.04%
176. feature PU4313314          0.04%
```

```
177. feature XPOLTY4    0.04%
178. feature P_UTD431H3_ROUT_S          0.04%
179. feature P_NUMMMX           0.04%
180. feature XPCVTY2    0.04%
181. feature PU431331           0.04%
182. feature P_NUMPCC7          0.04%
183. feature MMR2_AGE           0.03%
184. feature DMMR2      0.03%
185. feature PU4313313          0.03%
186. feature P_NUMFLUL          0.03%
187. feature VRC2_AGE           0.03%
188. feature PU431_31           0.03%
189. feature P_UTDHEP           0.03%
190. feature PUT43133           0.03%
191. feature XPCVTY3    0.03%
192. feature P_UTDPCV           0.03%
193. feature DVRC2      0.03%
194. feature P_NUMTPN           0.03%
195. feature P_NUMOLN           0.03%
196. feature PUTD4313           0.03%
197. feature P_U12VRC           0.03%
198. feature DHIB5      0.03%
199. feature DFLU5      0.02%
200. feature DTP5_AGE           0.02%
201. feature P_NUMHG    0.02%
202. feature P_UTDHIB           0.02%
203. feature P_NUMHION          0.02%
204. feature P_NUMHHY           0.02%
205. feature DDTP5      0.02%
206. feature P_UTDMMX           0.02%
207. feature HIB5_AGE           0.02%
208. feature P_UTDMCV           0.02%
209. feature P_NUMHEN           0.02%
210. feature PCV5_AGE           0.02%
211. feature P_NUMDAH           0.02%
212. feature P_UTDPOL           0.02%
213. feature P_UTD331           0.02%
214. feature FLU5_AGE           0.02%
215. feature P_UTDHIB_SHORT_S           0.02%
216. feature DPCV5      0.02%
217. feature XMMRTY2    0.02%
218. feature P_UTDPCVB13        0.02%
219. feature XHIBTY3    0.02%
220. feature XHIBTY4    0.02%
221. feature P_UTDPC3           0.02%
222. feature P_NUMOPV           0.02%
223. feature DHEPB5     0.01%
224. feature XDTPTY5    0.01%
225. feature XPCVTY5    0.01%
226. feature P_NUMMCN           0.01%
227. feature P_NUMVRN           0.01%
228. feature POL5_AGE           0.01%
229. feature P_NUHPHB           0.01%
230. feature DPOLIO5    0.01%
231. feature DHEPA3     0.01%
232. feature P_UTDTP3           0.01%
233. feature P_NUMPCP           0.01%
234. feature HEA3_AGE           0.01%
235. feature HEP5_AGE           0.01%
```

```
236. feature XHEPTY5    0.01%
237. feature XHIBTY1    0.01%
238. feature P_NUMMPR           0.01%
239. feature P_NUMPCCN          0.01%
240. feature P_NUMMS    0.01%
241. feature PDAT       0.01%
242. feature P_NUMMSM           0.01%
243. feature XHIBTY2    0.00%
244. feature P_NUMMSR           0.00%
245. feature P_NUMMP    0.00%
246. feature P_NUMRB    0.00%
247. feature XPOLTY5    0.00%
248. feature ROT4_AGE           0.00%
249. feature HEP_FLAG           0.00%
250. feature DROT4      0.00%
251. feature DPCV6      0.00%
252. feature DPOLIO6    0.00%
253. feature XPOLTY6    0.00%
254. feature XPCVTY6    0.00%
255. feature HIB6_AGE           0.00%
256. feature DHEPB6     0.00%
257. feature DDTP6      0.00%
258. feature DTP6_AGE           0.00%
259. feature DHIB6      0.00%
260. feature PCV6_AGE           0.00%
261. feature XDTPTY6    0.00%
262. feature MMR3_AGE           0.00%
263. feature XHEPTY6    0.00%
264. feature DHEPA4     0.00%
265. feature DFLU6      0.00%
266. feature DHEPA5     0.00%
267. feature DHEPB8     0.00%
268. feature DDTP7      0.00%
269. feature DDTP8      0.00%
270. feature YEAR       0.00%
271. feature SHOTCARD           0.00%
272. feature BFENDFL06          0.00%
273. feature BFFORMFL06         0.00%
274. feature DHEPB9     0.00%
275. feature DDTP9      0.00%
276. feature DHEPB7     0.00%
277. feature DHEPA6     0.00%
278. feature DFLU7      0.00%
279. feature DFLU8      0.00%
280. feature DHEPA9     0.00%
281. feature DHEPA8     0.00%
282. feature DHEPA7     0.00%
283. feature DFLU9      0.00%
284. feature DRB1       0.00%
285. feature DHIB7      0.00%
286. feature MPR2_AGE           0.00%
287. feature ROT5_AGE           0.00%
288. feature ROT6_AGE           0.00%
289. feature ROT7_AGE           0.00%
290. feature ROT8_AGE           0.00%
291. feature ROT9_AGE           0.00%
292. feature VRC3_AGE           0.00%
293. feature VRC4_AGE           0.00%
294. feature VRC5_AGE           0.00%
```

```
295. feature VRC6_AGE          0.00%
296. feature VRC7_AGE          0.00%
297. feature VRC8_AGE          0.00%
298. feature VRC9_AGE          0.00%
299. feature XDTPTY7    0.00%
300. feature XDTPTY8    0.00%
301. feature XDTPTY9    0.00%
302. feature XFLUTY1    0.00%
303. feature XFLUTY2    0.00%
304. feature XFLUTY3    0.00%
305. feature XFLUTY4    0.00%
306. feature RB9_AGE    0.00%
307. feature RB8_AGE    0.00%
308. feature RB7_AGE    0.00%
309. feature PCV9_AGE          0.00%
310. feature MPR4_AGE          0.00%
311. feature MPR5_AGE          0.00%
312. feature MPR6_AGE          0.00%
313. feature MPR7_AGE          0.00%
314. feature MPR8_AGE          0.00%
315. feature MPR9_AGE          0.00%
316. feature PCV7_AGE          0.00%
317. feature PCV8_AGE          0.00%
318. feature POL6_AGE          0.00%
319. feature RB6_AGE    0.00%
320. feature POL7_AGE          0.00%
321. feature POL8_AGE          0.00%
322. feature POL9_AGE          0.00%
323. feature RB1_AGE    0.00%
324. feature RB2_AGE    0.00%
325. feature RB3_AGE    0.00%
326. feature RB4_AGE    0.00%
327. feature RB5_AGE    0.00%
328. feature XFLUTY5    0.00%
329. feature XFLUTY6    0.00%
330. feature XFLUTY7    0.00%
331. feature XROTTY9    0.00%
332. feature XROTTY1    0.00%
333. feature XROTTY2    0.00%
334. feature XROTTY3    0.00%
335. feature XROTTY4    0.00%
336. feature XROTTY5    0.00%
337. feature XROTTY6    0.00%
338. feature XROTTY7    0.00%
339. feature XROTTY8    0.00%
340. feature XVRCTY1    0.00%
341. feature XPOLTY8    0.00%
342. feature XVRCTY2    0.00%
343. feature XVRCTY3    0.00%
344. feature XVRCTY4    0.00%
345. feature XVRCTY5    0.00%
346. feature XVRCTY6    0.00%
347. feature XVRCTY7    0.00%
348. feature XVRCTY8    0.00%
349. feature XVRCTY9    0.00%
350. feature XPOLTY9    0.00%
351. feature XPOLTY7    0.00%
352. feature XFLUTY8    0.00%
353. feature XHIBTY9    0.00%
```

```
354. feature XFLUTY9    0.00%
355. feature XHEPTY7    0.00%
356. feature XHEPTY8    0.00%
357. feature XHEPTY9    0.00%
358. feature XHIBTY5    0.00%
359. feature XHIBTY6    0.00%
360. feature XHIBTY7    0.00%
361. feature XHIBTY8    0.00%
362. feature XMMRTY3    0.00%
363. feature XPCVTY9    0.00%
364. feature XMMRTY4    0.00%
365. feature XMMRTY5    0.00%
366. feature XMMRTY6    0.00%
367. feature XMMRTY7    0.00%
368. feature XMMRTY8    0.00%
369. feature XMMRTY9    0.00%
370. feature XPCVTY7    0.00%
371. feature XPCVTY8    0.00%
372. feature MPR3_AGE           0.00%
373. feature MPR1_AGE           0.00%
374. feature DHIB8      0.00%
375. feature MP9_AGE    0.00%
376. feature DMPRB6     0.00%
377. feature DMPRB7     0.00%
378. feature DMPRB8     0.00%
379. feature DMPRB9     0.00%
380. feature DPCV7      0.00%
381. feature DPCV8      0.00%
382. feature DPCV9      0.00%
383. feature DPOLIO7    0.00%
384. feature DPOLIO8    0.00%
385. feature DPOLIO9    0.00%
386. feature DRB2       0.00%
387. feature DRB4       0.00%
388. feature DRB5       0.00%
389. feature DRB6       0.00%
390. feature DRB7       0.00%
391. feature DRB8       0.00%
392. feature DRB9       0.00%
393. feature DROT5      0.00%
394. feature DROT6      0.00%
395. feature DMPRB5     0.00%
396. feature DMPRB4     0.00%
397. feature DMPRB3     0.00%
398. feature DMP1       0.00%
399. feature DHIB9      0.00%
400. feature DMMR3      0.00%
401. feature DMMR4      0.00%
402. feature DMMR5      0.00%
403. feature DMMR6      0.00%
404. feature DMMR7      0.00%
405. feature DMMR8      0.00%
406. feature DMMR9      0.00%
407. feature DMP2       0.00%
408. feature DMPRB2     0.00%
409. feature DMP3       0.00%
410. feature DMP4       0.00%
411. feature DMP5       0.00%
412. feature DMP6       0.00%
```

```
413. feature DMP7       0.00%
414. feature DMP8       0.00%
415. feature DMP9       0.00%
416. feature DMPRB1     0.00%
417. feature DROT7      0.00%
418. feature DROT8      0.00%
419. feature DROT9      0.00%
420. feature MMR8_AGE           0.00%
421. feature HEP9_AGE           0.00%
422. feature HIB7_AGE           0.00%
423. feature HIB8_AGE           0.00%
424. feature HIB9_AGE           0.00%
425. feature MMR4_AGE           0.00%
426. feature MMR5_AGE           0.00%
427. feature MMR6_AGE           0.00%
428. feature MMR7_AGE           0.00%
429. feature MMR9_AGE           0.00%
430. feature HEP7_AGE           0.00%
431. feature MP1_AGE    0.00%
432. feature MP2_AGE    0.00%
433. feature MP3_AGE    0.00%
434. feature MP4_AGE    0.00%
435. feature MP5_AGE    0.00%
436. feature MP6_AGE    0.00%
437. feature MP7_AGE    0.00%
438. feature MP8_AGE    0.00%
439. feature HEP8_AGE           0.00%
440. feature HEP6_AGE           0.00%
441. feature DVRC3      0.00%
442. feature DTP9_AGE           0.00%
443. feature DVRC4      0.00%
444. feature DVRC5      0.00%
445. feature DVRC6      0.00%
446. feature DVRC7      0.00%
447. feature DVRC8      0.00%
448. feature DVRC9      0.00%
449. feature DTP7_AGE           0.00%
450. feature DTP8_AGE           0.00%
451. feature FLU6_AGE           0.00%
452. feature HEA9_AGE           0.00%
453. feature FLU7_AGE           0.00%
454. feature FLU8_AGE           0.00%
455. feature FLU9_AGE           0.00%
456. feature HEA4_AGE           0.00%
457. feature HEA5_AGE           0.00%
458. feature HEA6_AGE           0.00%
459. feature HEA7_AGE           0.00%
460. feature HEA8_AGE           0.00%
461. feature DRB3       0.00%
```

Hmmm... Our random number turns out to be the third most important feature for my target variable of FRSTBRN.

So my FRSTBRN investigation shows the top 2 variables for determining whether a child is firstborn are:

```
* CHILDNM -- Number of children in the household, and
* C1R  -- Number of people in the household.
```

That's kindda sad!

Lesson of the day, choose your variables wisely.

Assignment Requirements

Now it's your turn to perform a similar type of analysis.

Using the immunization dataset, complete the following:
1) Load the dataset and cleanup the missing values.

- As noted above, you need to do more than I did in my demonstration.
- Defend your handling of these values.
    - Tell me why are you doing what you are doing to the data, it matters! 2) Choose your own target variable (can be categorical or continuous, if you feel brave) 3) Determine if any of the variables are correlated to each other.
- Produce a correlation matrix at a minimum.
- If you decide to do a pairplot, remember the warning about wide datasets.
- Use the Random Number Trick to determine relevant variables.
    - Remember to check the accuracy of your model, before making any decisions.
        - If your accuracy is under 85%, you should improve your model or slect a new target variable. 4) Reduce the dataset to only include correlated variables from above. 5) Complete EDA using the lecture notebook as a baseline -- feel free to add your own tests.
            - Produce visualizations for each of your remaining columns.
            - Explain what your visuals are telling you about your data. Be specific! 6) Based on EDA results, either
            - Go back for more variables, or
            - Drop more variables 7) Summary/conclude of your findings. Be specific and detailed in your explanations.
            - What do all the visuals mean in your analysis?
            - Why did you handling the data the way you did?
            - Don't assume I'll understand based on your code.

Deliverables

1) Load the dataset and cleanup the missing values.

- As noted above, you need to do more than I did in my demonstration.
- Defend your handling of these values.
  - Tell me why are you doing what you are doing to the data, it matters!

In [32]:
```
# Loading the dataset
df = pd.read_csv("assign_wk5/nispuf14.csv", na_values=['.'], low_memory=False)
```

In [33]:
```
# Taking a look at the dataset
df.head()
```

Out[33]:

	SEQNUMC	SEQNUMHH	PDAT	PROVWT_D	PROVWT_D_TERR	RDDWT_D
0	11	1	2	.	.	218.30024855484000
1	21	2	1	806.84601169505000	806.84601169505000	454.86041741251200
2	31	3	2	.	.	30.54542540283290
3	41	4	1	63.44868567610260	63.44868567610260	36.96593137368630
4	51	5	1	94.87263225744540	94.87263225744540	64.62020426239790

5 rows × 461 columns

As was demonstrated above, a lot of things, like age, are being stored in object form. I am going to go ahead and change these to numbers.

In [34]:
```
for col in df.columns:
    df[col] = pd.to_numeric(df[col], errors='coerce')
```

In [35]:
```
# Making a copy of the data frame
df_new = df
```

After looking at a lot of the data, and trying other methods, filling the continous values with the mean is probably the lesser of all the evils. A lot of the missing data is things like weight of child, and there is just not a good way to impute it any other way. There are 122 continous variables, if you had all the time in the world you could come up with an individual solution for each, like imputing weight based on the average of that age catagory. This may give you a cleaner result.

One thing that I did notice is that the number of values missing in a lot of the rows is the same. After reading through the report I found the following:

To start, I want to limit the rows to only those that have "Adequate Provider Data" I am going to do that using the code below.

In [36]:
```
# Picked a random column that has 15,059 from the vaccine info.
df_adequate = df.dropna(subset=['P_NUHEPX'])
```

In [ ]:
```
df_adequate.info(verbose=True, show_counts=True)
```

It is going to be helpful to separate the values into catagorical values AND continous values. As such I am going to use the code that was shown above to kick off dealing with the missing values.

In [38]:
```
# Using the code from the example to split the data into continous and categorical

from itertools import tee, islice, chain

def previous_and_next(some_iterable):
    prevs, items, nexts = tee(some_iterable, 3)
    prevs = chain([None], prevs)
    nexts = chain(islice(nexts, 1, None), [None])
    return zip(prevs, items, nexts)

cont_list = []
for prev, item, nxt in previous_and_next(outlines):
    if isinstance(item, list) :
        if 'Continuous' in str(item[0].title):
            cont_list.append(prev.title)

cont_list[:5]

cat_list = [c for c in df_adequate.columns if c not in cont_list]
```

Catagorical values are going to get a 0, that way I can later use the column as a catagory and just know that 0 means "Unknown" or missing.

In [ ]:
```
for col in cat_list:
    df_adequate[col].fillna(0, inplace=True)
```

In [40]:
```
df_adequate.info(verbose=True, show_counts=True)
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 15059 entries, 1 to 24896
Data columns (total 461 columns):
 #    Column              Non-Null Count  Dtype  
---   ------              --------------  -----  
 0    SEQNUMC             15059 non-null  int64  
 1    SEQNUMHH            15059 non-null  int64  
 2    PDAT                15059 non-null  int64  
 3    PROVWT_D            14893 non-null  float64
 4    PROVWT_D_TERR       15059 non-null  float64
 5    RDDWT_D             14893 non-null  float64
 6    RDDWT_D_TERR        15059 non-null  float64
 7    STRATUM             15059 non-null  int64  
 8    YEAR                15059 non-null  int64  
 9    AGECPOXR            15059 non-null  float64
 10   HAD_CPOX            15059 non-null  int64  
 11   SHOTCARD            15059 non-null  int64  
 12   AGEGRP              15059 non-null  int64  
 13   BF_ENDR06           11724 non-null  float64
 14   BF_EXCLR06          12100 non-null  float64
 15   BF_FORMR08          10516 non-null  float64
 16   BFENDFL06           15059 non-null  float64
 17   BFFORMFL06          15059 non-null  float64
 18   C1R                 15059 non-null  int64  
 19   C5R                 15059 non-null  int64  
 20   CBF_01              15059 non-null  int64  
 21   CEN_REG             15059 non-null  float64
 22   CHILDNM             15059 non-null  int64  
 23   CWIC_01             15059 non-null  int64  
 24   CWIC_02             15059 non-null  float64
 25   EDUC1               15059 non-null  int64  
 26   FRSTBRN             15059 non-null  int64  
 27   I_HISP_K            15059 non-null  int64  
 28   INCPORAR            14218 non-null  float64
 29   INCPOV1             15059 non-null  int64  
 30   INCQ298A            15059 non-null  int64  
 31   INTRP               15059 non-null  float64
 32   LANGUAGE            15059 non-null  int64  
 33   M_AGEGRP            15059 non-null  int64  
 34   MARITAL2            15059 non-null  int64  
 35   MOBIL_I             15059 non-null  int64  
 36   NUM_PHONE           15059 non-null  float64
 37   NUM_CELLS_HH        15059 non-null  float64
 38   NUM_CELLS_PARENTS   15059 non-null  float64
 39   RACE_K              15059 non-null  int64  
 40   RACEETHK            15059 non-null  int64  
 41   RENT_OWN            15059 non-null  int64  
 42   SEX                 15059 non-null  int64  
 43   ESTIAP14            15059 non-null  int64  
 44   EST_GRANT           15059 non-null  float64
 45   STATE               15059 non-null  int64  
 46   D6R                 15059 non-null  float64
 47   D7                  15059 non-null  float64
 48   N_PRVR              15059 non-null  int64  
 49   PROV_FAC            15059 non-null  float64
 50   REGISTRY            15059 non-null  float64
 51   VFC_ORDER           15059 non-null  float64
 52   HEP_BRTH            15059 non-null  float64
 53   HEP_FLAG            15059 non-null  float64
```

```
 54   P_NUHEPX             15059 non-null  float64
 55   P_NUHIBX             15059 non-null  float64
 56   P_NUHPHB             15059 non-null  float64
 57   P_NUMDAH             15059 non-null  float64
 58   P_NUMDHI             15059 non-null  float64
 59   P_NUMDIH             15059 non-null  float64
 60   P_NUMDTA             15059 non-null  float64
 61   P_NUMDTP             15059 non-null  float64
 62   P_NUMFLU             15059 non-null  float64
 63   P_NUMFLUL            15059 non-null  float64
 64   P_NUMFLUM            15059 non-null  float64
 65   P_NUMFLUN            15059 non-null  float64
 66   P_NUMHEA             15059 non-null  float64
 67   P_NUMHEN             15059 non-null  float64
 68   P_NUMHEP             15059 non-null  float64
 69   P_NUMHG              15059 non-null  float64
 70   P_NUMHHY             15059 non-null  float64
 71   P_NUMHIB             15059 non-null  float64
 72   P_NUMHIN             15059 non-null  float64
 73   P_NUMHION            15059 non-null  float64
 74   P_NUMHM              15059 non-null  float64
 75   P_NUMHS              15059 non-null  float64
 76   P_NUMIPV             15059 non-null  float64
 77   P_NUMMCN             15059 non-null  float64
 78   P_NUMMMR             15059 non-null  float64
 79   P_NUMMMRX            15059 non-null  float64
 80   P_NUMMMX             15059 non-null  float64
 81   P_NUMMP              15059 non-null  float64
 82   P_NUMMPR             15059 non-null  float64
 83   P_NUMMRV             15059 non-null  float64
 84   P_NUMMS              15059 non-null  float64
 85   P_NUMMSM             15059 non-null  float64
 86   P_NUMMSR             15059 non-null  float64
 87   P_NUMOLN             15059 non-null  float64
 88   P_NUMOPV             15059 non-null  float64
 89   P_NUMPCV             15059 non-null  float64
 90   P_NUMPCP             15059 non-null  float64
 91   P_NUMPCC             15059 non-null  float64
 92   P_NUMPCC7            15059 non-null  float64
 93   P_NUMPCC13           15059 non-null  float64
 94   P_NUMPCCN            15059 non-null  float64
 95   P_NUMPCN             15059 non-null  float64
 96   P_NUMPOL             15059 non-null  float64
 97   P_NUMRB              15059 non-null  float64
 98   P_NUMRG              15059 non-null  float64
 99   P_NUMRM              15059 non-null  float64
 100  P_NUMRO              15059 non-null  float64
 101  P_NUMROT             15059 non-null  float64
 102  P_NUMTPN             15059 non-null  float64
 103  P_NUMVRC             15059 non-null  float64
 104  P_NUMVRN             15059 non-null  float64
 105  P_NUMVRX             15059 non-null  float64
 106  P_U12VRC             15059 non-null  float64
 107  P_UTD331             15059 non-null  float64
 108  P_UTD431             15059 non-null  float64
 109  P_UTDHEP             15059 non-null  float64
 110  P_UTDHEPA1           15059 non-null  float64
 111  P_UTDHEPA2           15059 non-null  float64
 112  P_UTDHIB             15059 non-null  float64
```

```
 113  P_UTDHIB_ROUT_S      15059 non-null  float64
 114  P_UTDHIB_SHORT_S     15059 non-null  float64
 115  P_UTDMCV             15059 non-null  float64
 116  P_UTDMMX             15059 non-null  float64
 117  P_UTDPC3             15059 non-null  float64
 118  P_UTDPCV             15059 non-null  float64
 119  P_UTDPCVB13          15059 non-null  float64
 120  P_UTDPOL             15059 non-null  float64
 121  P_UTDROT_S           15059 non-null  float64
 122  P_UTDTP3             15059 non-null  float64
 123  P_UTDTP4             15059 non-null  float64
 124  PU431331             15059 non-null  float64
 125  P_UTD431H31_ROUT_S   15059 non-null  float64
 126  PU431_31             15059 non-null  float64
 127  PU4313313            15059 non-null  float64
 128  P_UTD431H313_ROUT_S  15059 non-null  float64
 129  PU4313314            15059 non-null  float64
 130  P_UTD431H314_ROUT_S  15059 non-null  float64
 131  PU431_314            15059 non-null  float64
 132  PUT43133             15059 non-null  float64
 133  P_UTD431H3_ROUT_S    15059 non-null  float64
 134  PUTD4313             15059 non-null  float64
 135  P_UTD431H_ROUT_S     15059 non-null  float64
 136  U1D_HEP              15059 non-null  float64
 137  U2D_HEP              15059 non-null  float64
 138  U3D_HEP              15059 non-null  float64
 139  DDTP1                14688 non-null  float64
 140  DDTP2                14547 non-null  float64
 141  DDTP3                14345 non-null  float64
 142  DDTP4                13007 non-null  float64
 143  DDTP5                216 non-null    float64
 144  DDTP6                7 non-null      float64
 145  DDTP7                15059 non-null  float64
 146  DDTP8                15059 non-null  float64
 147  DDTP9                15059 non-null  float64
 148  DFLU1                11205 non-null  float64
 149  DFLU2                9292 non-null   float64
 150  DFLU3                6102 non-null   float64
 151  DFLU4                1755 non-null   float64
 152  DFLU5                103 non-null    float64
 153  DFLU6                4 non-null      float64
 154  DFLU7                15059 non-null  float64
 155  DFLU8                15059 non-null  float64
 156  DFLU9                15059 non-null  float64
 157  DHEPA1               12910 non-null  float64
 158  DHEPA2               9379 non-null   float64
 159  DHEPA3               63 non-null     float64
 160  DHEPA4               3 non-null      float64
 161  DHEPA5               15059 non-null  float64
 162  DHEPA6               15059 non-null  float64
 163  DHEPA7               15059 non-null  float64
 164  DHEPA8               15059 non-null  float64
 165  DHEPA9               15059 non-null  float64
 166  DHEPB1               14665 non-null  float64
 167  DHEPB2               14334 non-null  float64
 168  DHEPB3               13844 non-null  float64
 169  DHEPB4               3869 non-null   float64
 170  DHEPB5               154 non-null    float64
 171  DHEPB6               9 non-null      float64
```

```
172  DHEPB7              1 non-null      float64
173  DHEPB8              15059 non-null  float64
174  DHEPB9              15059 non-null  float64
175  DHIB1               14625 non-null  float64
176  DHIB2               14436 non-null  float64
177  DHIB3               14079 non-null  float64
178  DHIB4               10905 non-null  float64
179  DHIB5               170 non-null    float64
180  DHIB6               6 non-null      float64
181  DHIB7               1 non-null      float64
182  DHIB8               15059 non-null  float64
183  DHIB9               15059 non-null  float64
184  DMMR1               14047 non-null  float64
185  DMMR2               285 non-null    float64
186  DMMR3               2 non-null      float64
187  DMMR4               1 non-null      float64
188  DMMR5               15059 non-null  float64
189  DMMR6               15059 non-null  float64
190  DMMR7               15059 non-null  float64
191  DMMR8               15059 non-null  float64
192  DMMR9               15059 non-null  float64
193  DMP1                15059 non-null  float64
194  DMP2                15059 non-null  float64
195  DMP3                15059 non-null  float64
196  DMP4                15059 non-null  float64
197  DMP5                15059 non-null  float64
198  DMP6                15059 non-null  float64
199  DMP7                15059 non-null  float64
200  DMP8                15059 non-null  float64
201  DMP9                15059 non-null  float64
202  DMPRB1              1 non-null      float64
203  DMPRB2              15059 non-null  float64
204  DMPRB3              15059 non-null  float64
205  DMPRB4              15059 non-null  float64
206  DMPRB5              15059 non-null  float64
207  DMPRB6              15059 non-null  float64
208  DMPRB7              15059 non-null  float64
209  DMPRB8              15059 non-null  float64
210  DMPRB9              15059 non-null  float64
211  DPCV1               14578 non-null  float64
212  DPCV2               14381 non-null  float64
213  DPCV3               14091 non-null  float64
214  DPCV4               12862 non-null  float64
215  DPCV5               194 non-null    float64
216  DPCV6               6 non-null      float64
217  DPCV7               1 non-null      float64
218  DPCV8               15059 non-null  float64
219  DPCV9               15059 non-null  float64
220  DPOLIO1             14602 non-null  float64
221  DPOLIO2             14463 non-null  float64
222  DPOLIO3             14135 non-null  float64
223  DPOLIO4             2122 non-null   float64
224  DPOLIO5             58 non-null     float64
225  DPOLIO6             5 non-null      float64
226  DPOLIO7             15059 non-null  float64
227  DPOLIO8             15059 non-null  float64
228  DPOLIO9             15059 non-null  float64
229  DRB1                1 non-null      float64
230  DRB2                15059 non-null  float64
```

```
 231  DRB3                     15059 non-null  float64
 232  DRB4                     15059 non-null  float64
 233  DRB5                     15059 non-null  float64
 234  DRB6                     15059 non-null  float64
 235  DRB7                     15059 non-null  float64
 236  DRB8                     15059 non-null  float64
 237  DRB9                     15059 non-null  float64
 238  DROT1                    13141 non-null  float64
 239  DROT2                    12617 non-null  float64
 240  DROT3                    8643 non-null   float64
 241  DROT4                    21 non-null     float64
 242  DROT5                    1 non-null      float64
 243  DROT6                    15059 non-null  float64
 244  DROT7                    15059 non-null  float64
 245  DROT8                    15059 non-null  float64
 246  DROT9                    15059 non-null  float64
 247  DVRC1                    13886 non-null  float64
 248  DVRC2                    243 non-null    float64
 249  DVRC3                    2 non-null      float64
 250  DVRC4                    15059 non-null  float64
 251  DVRC5                    15059 non-null  float64
 252  DVRC6                    15059 non-null  float64
 253  DVRC7                    15059 non-null  float64
 254  DVRC8                    15059 non-null  float64
 255  DVRC9                    15059 non-null  float64
 256  DTP1_AGE                 14688 non-null  float64
 257  DTP2_AGE                 14547 non-null  float64
 258  DTP3_AGE                 14345 non-null  float64
 259  DTP4_AGE                 13007 non-null  float64
 260  DTP5_AGE                 216 non-null    float64
 261  DTP6_AGE                 7 non-null      float64
 262  DTP7_AGE                 15059 non-null  float64
 263  DTP8_AGE                 15059 non-null  float64
 264  DTP9_AGE                 15059 non-null  float64
 265  FLU1_AGE                 11205 non-null  float64
 266  FLU2_AGE                 9292 non-null   float64
 267  FLU3_AGE                 6102 non-null   float64
 268  FLU4_AGE                 1755 non-null   float64
 269  FLU5_AGE                 103 non-null    float64
 270  FLU6_AGE                 4 non-null      float64
 271  FLU7_AGE                 15059 non-null  float64
 272  FLU8_AGE                 15059 non-null  float64
 273  FLU9_AGE                 15059 non-null  float64
 274  HEA1_AGE                 12910 non-null  float64
 275  HEA2_AGE                 9379 non-null   float64
 276  HEA3_AGE                 63 non-null     float64
 277  HEA4_AGE                 3 non-null      float64
 278  HEA5_AGE                 15059 non-null  float64
 279  HEA6_AGE                 15059 non-null  float64
 280  HEA7_AGE                 15059 non-null  float64
 281  HEA8_AGE                 15059 non-null  float64
 282  HEA9_AGE                 15059 non-null  float64
 283  HEP1_AGE                 14665 non-null  float64
 284  HEP2_AGE                 14334 non-null  float64
 285  HEP3_AGE                 13844 non-null  float64
 286  HEP4_AGE                 3869 non-null   float64
 287  HEP5_AGE                 154 non-null    float64
 288  HEP6_AGE                 9 non-null      float64
 289  HEP7_AGE                 1 non-null      float64
```

```
 290  HEP8_AGE            15059 non-null  float64
 291  HEP9_AGE            15059 non-null  float64
 292  HIB1_AGE            14625 non-null  float64
 293  HIB2_AGE            14436 non-null  float64
 294  HIB3_AGE            14079 non-null  float64
 295  HIB4_AGE            10905 non-null  float64
 296  HIB5_AGE            170 non-null    float64
 297  HIB6_AGE            6 non-null      float64
 298  HIB7_AGE            1 non-null      float64
 299  HIB8_AGE            15059 non-null  float64
 300  HIB9_AGE            15059 non-null  float64
 301  MMR1_AGE            14047 non-null  float64
 302  MMR2_AGE            285 non-null    float64
 303  MMR3_AGE            2 non-null      float64
 304  MMR4_AGE            1 non-null      float64
 305  MMR5_AGE            15059 non-null  float64
 306  MMR6_AGE            15059 non-null  float64
 307  MMR7_AGE            15059 non-null  float64
 308  MMR8_AGE            15059 non-null  float64
 309  MMR9_AGE            15059 non-null  float64
 310  MP1_AGE             15059 non-null  float64
 311  MP2_AGE             15059 non-null  float64
 312  MP3_AGE             15059 non-null  float64
 313  MP4_AGE             15059 non-null  float64
 314  MP5_AGE             15059 non-null  float64
 315  MP6_AGE             15059 non-null  float64
 316  MP7_AGE             15059 non-null  float64
 317  MP8_AGE             15059 non-null  float64
 318  MP9_AGE             15059 non-null  float64
 319  MPR1_AGE            1 non-null      float64
 320  MPR2_AGE            15059 non-null  float64
 321  MPR3_AGE            15059 non-null  float64
 322  MPR4_AGE            15059 non-null  float64
 323  MPR5_AGE            15059 non-null  float64
 324  MPR6_AGE            15059 non-null  float64
 325  MPR7_AGE            15059 non-null  float64
 326  MPR8_AGE            15059 non-null  float64
 327  MPR9_AGE            15059 non-null  float64
 328  PCV1_AGE            14578 non-null  float64
 329  PCV2_AGE            14381 non-null  float64
 330  PCV3_AGE            14091 non-null  float64
 331  PCV4_AGE            12862 non-null  float64
 332  PCV5_AGE            194 non-null    float64
 333  PCV6_AGE            6 non-null      float64
 334  PCV7_AGE            1 non-null      float64
 335  PCV8_AGE            15059 non-null  float64
 336  PCV9_AGE            15059 non-null  float64
 337  POL1_AGE            14602 non-null  float64
 338  POL2_AGE            14463 non-null  float64
 339  POL3_AGE            14135 non-null  float64
 340  POL4_AGE            2122 non-null   float64
 341  POL5_AGE            58 non-null     float64
 342  POL6_AGE            5 non-null      float64
 343  POL7_AGE            15059 non-null  float64
 344  POL8_AGE            15059 non-null  float64
 345  POL9_AGE            15059 non-null  float64
 346  RB1_AGE             1 non-null      float64
 347  RB2_AGE             15059 non-null  float64
 348  RB3_AGE             15059 non-null  float64
```

```
 349  RB4_AGE             15059 non-null  float64
 350  RB5_AGE             15059 non-null  float64
 351  RB6_AGE             15059 non-null  float64
 352  RB7_AGE             15059 non-null  float64
 353  RB8_AGE             15059 non-null  float64
 354  RB9_AGE             15059 non-null  float64
 355  ROT1_AGE            13141 non-null  float64
 356  ROT2_AGE            12617 non-null  float64
 357  ROT3_AGE            8643 non-null   float64
 358  ROT4_AGE            21 non-null     float64
 359  ROT5_AGE            1 non-null      float64
 360  ROT6_AGE            15059 non-null  float64
 361  ROT7_AGE            15059 non-null  float64
 362  ROT8_AGE            15059 non-null  float64
 363  ROT9_AGE            15059 non-null  float64
 364  VRC1_AGE            13886 non-null  float64
 365  VRC2_AGE            243 non-null    float64
 366  VRC3_AGE            2 non-null      float64
 367  VRC4_AGE            15059 non-null  float64
 368  VRC5_AGE            15059 non-null  float64
 369  VRC6_AGE            15059 non-null  float64
 370  VRC7_AGE            15059 non-null  float64
 371  VRC8_AGE            15059 non-null  float64
 372  VRC9_AGE            15059 non-null  float64
 373  XDTPTY1             15059 non-null  float64
 374  XDTPTY2             15059 non-null  float64
 375  XDTPTY3             15059 non-null  float64
 376  XDTPTY4             15059 non-null  float64
 377  XDTPTY5             15059 non-null  float64
 378  XDTPTY6             15059 non-null  float64
 379  XDTPTY7             15059 non-null  float64
 380  XDTPTY8             15059 non-null  float64
 381  XDTPTY9             15059 non-null  float64
 382  XFLUTY1             15059 non-null  float64
 383  XFLUTY2             15059 non-null  float64
 384  XFLUTY3             15059 non-null  float64
 385  XFLUTY4             15059 non-null  float64
 386  XFLUTY5             15059 non-null  float64
 387  XFLUTY6             15059 non-null  float64
 388  XFLUTY7             15059 non-null  float64
 389  XFLUTY8             15059 non-null  float64
 390  XFLUTY9             15059 non-null  float64
 391  XHEPTY1             15059 non-null  float64
 392  XHEPTY2             15059 non-null  float64
 393  XHEPTY3             15059 non-null  float64
 394  XHEPTY4             15059 non-null  float64
 395  XHEPTY5             15059 non-null  float64
 396  XHEPTY6             15059 non-null  float64
 397  XHEPTY7             15059 non-null  float64
 398  XHEPTY8             15059 non-null  float64
 399  XHEPTY9             15059 non-null  float64
 400  XHIBTY1             15059 non-null  float64
 401  XHIBTY2             15059 non-null  float64
 402  XHIBTY3             15059 non-null  float64
 403  XHIBTY4             15059 non-null  float64
 404  XHIBTY5             15059 non-null  float64
 405  XHIBTY6             15059 non-null  float64
 406  XHIBTY7             15059 non-null  float64
 407  XHIBTY8             15059 non-null  float64
```

```
 408  XHIBTY9            15059 non-null  float64
 409  XMMRTY1            15059 non-null  float64
 410  XMMRTY2            15059 non-null  float64
 411  XMMRTY3            15059 non-null  float64
 412  XMMRTY4            15059 non-null  float64
 413  XMMRTY5            15059 non-null  float64
 414  XMMRTY6            15059 non-null  float64
 415  XMMRTY7            15059 non-null  float64
 416  XMMRTY8            15059 non-null  float64
 417  XMMRTY9            15059 non-null  float64
 418  XPCVTY1            15059 non-null  float64
 419  XPCVTY2            15059 non-null  float64
 420  XPCVTY3            15059 non-null  float64
 421  XPCVTY4            15059 non-null  float64
 422  XPCVTY5            15059 non-null  float64
 423  XPCVTY6            15059 non-null  float64
 424  XPCVTY7            15059 non-null  float64
 425  XPCVTY8            15059 non-null  float64
 426  XPCVTY9            15059 non-null  float64
 427  XPOLTY1            15059 non-null  float64
 428  XPOLTY2            15059 non-null  float64
 429  XPOLTY3            15059 non-null  float64
 430  XPOLTY4            15059 non-null  float64
 431  XPOLTY5            15059 non-null  float64
 432  XPOLTY6            15059 non-null  float64
 433  XPOLTY7            15059 non-null  float64
 434  XPOLTY8            15059 non-null  float64
 435  XPOLTY9            15059 non-null  float64
 436  XROTTY1            15059 non-null  float64
 437  XROTTY2            15059 non-null  float64
 438  XROTTY3            15059 non-null  float64
 439  XROTTY4            15059 non-null  float64
 440  XROTTY5            15059 non-null  float64
 441  XROTTY6            15059 non-null  float64
 442  XROTTY7            15059 non-null  float64
 443  XROTTY8            15059 non-null  float64
 444  XROTTY9            15059 non-null  float64
 445  XVRCTY1            15059 non-null  float64
 446  XVRCTY2            15059 non-null  float64
 447  XVRCTY3            15059 non-null  float64
 448  XVRCTY4            15059 non-null  float64
 449  XVRCTY5            15059 non-null  float64
 450  XVRCTY6            15059 non-null  float64
 451  XVRCTY7            15059 non-null  float64
 452  XVRCTY8            15059 non-null  float64
 453  XVRCTY9            15059 non-null  float64
 454  INS_1              15059 non-null  float64
 455  INS_2              15059 non-null  float64
 456  INS_3              15059 non-null  float64
 457  INS_3A             15059 non-null  float64
 458  INS_4_5            15059 non-null  float64
 459  INS_6              15059 non-null  float64
 460  INS_11             15059 non-null  float64
dtypes: float64(432), int64(29)
memory usage: 53.1 MB
```

Making some progress!! Columns that still have missing information:

- 3 this is child weight
- 5 this is child weight, but from a different source and time
- 13,14,15 are information on breastfeeding, going to be impossible to fill this with anything other than "unknown"
- 18 this is census region, some if just missing, it could be imputed from other geographic data
- 26 this is income to poverty ratio
- 34 the number of landlines in the home, more than 50% of that data it missing

It is going to be impossible to fill in things like the child's breastfeeding habit, so this information either needs a dummy value or to be removed from the data set. Same with the child's weight. You can not impute it based on age, as kids very so much in size. You could use a mean for those values, but that will be less than ideal as well. This is where cleaning up data is so much effort. I do not want to toss out any more data, I already am taking just the "Adequate" information from the data and have filled in a lot of the categorical information with "Unknown". At this point the lesser of evils is to fill in the continous data with the mean of the column.

In [41]:
```
# Looking at a column with a lot of missing continous data
df_adequate['DFLU3']
```

Out[41]:
```
1         421.0
3         553.0
4        1099.0
5           NaN
6           NaN
          ...
24889       NaN
24890       NaN
24891     756.0
24895       NaN
24896       NaN
Name: DFLU3, Length: 15059, dtype: float64
```

In [42]:
```
df_full = df_adequate.fillna(df_adequate.mean())
```

Checking to see if there are still any missing values

In [43]:
```
df_full.isnull().values.any()
```

Out[43]:
```
False
```

2) Choose your own target variable (can be categorical or continuous, if you feel brave)

I am curious to look at chicken pox infections, data is stored as either a 1 - child had chicken pox, a 2 - child never had chicken pox, or 77, unknown, or 99 missing.

In [44]:
```
df_pox = df_full
```

In [45]:
```
# Looking at the unique values in the "had chicken pox" column
df_pox['HAD_CPOX'].unique()
```

Out[45]:
```
array([ 2,  1, 77], dtype=int64)
```

According to the code book "77" means "Unknown" if child had chicken pox, those are going to need to be dropped for our classifier.

In [46]:
```
df_pox = df_pox[df_pox.HAD_CPOX != 77]
```

In [47]:
```
# Looking at the unique values in the "had chicken pox" column
df_pox['HAD_CPOX'].unique()
```

Out[47]:
```
array([2, 1], dtype=int64)
```

3) Determine if any of the variables are correlated to each other.

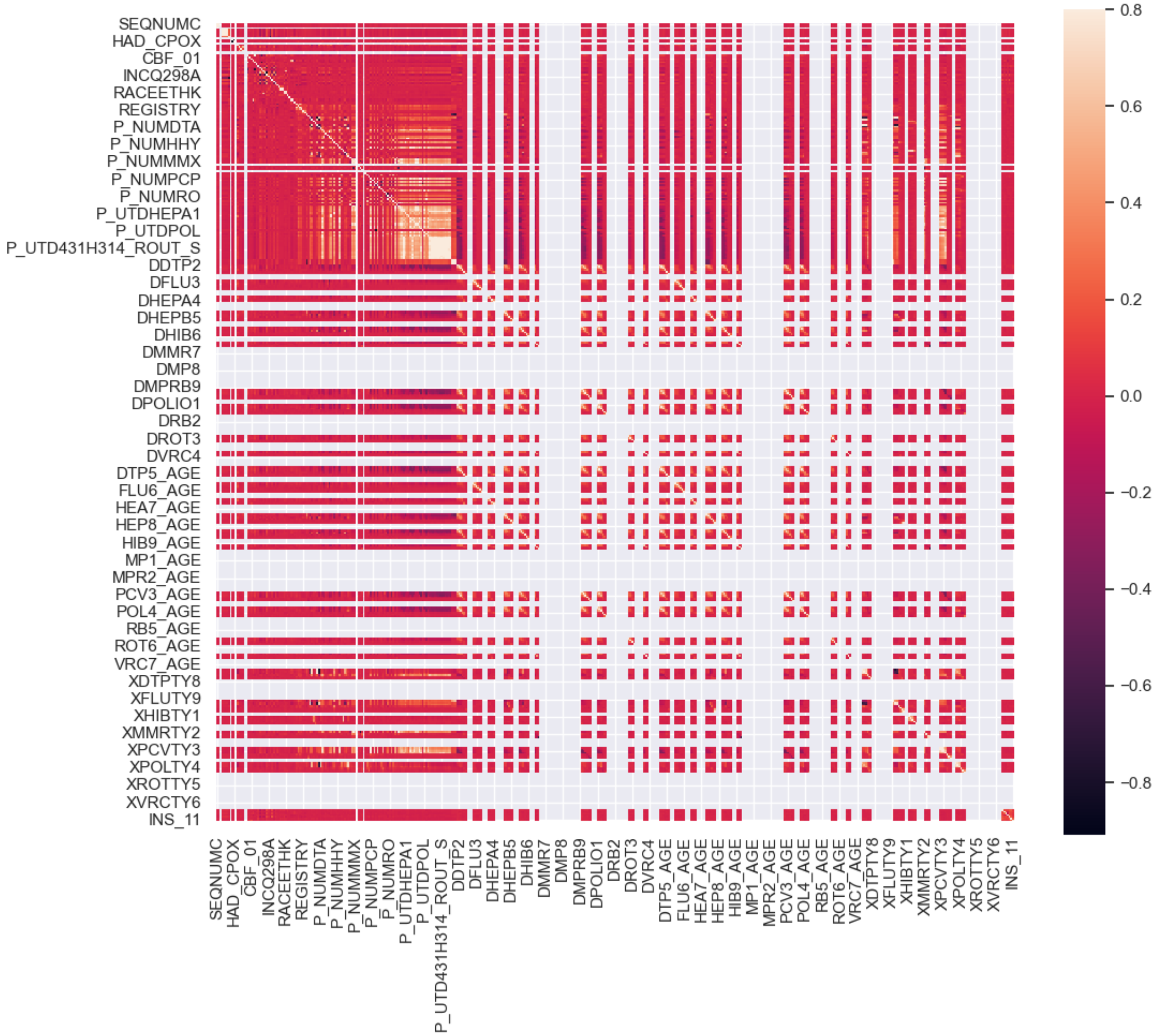
- Produce a correlation matrix at a minimum.
- If you decide to do a pairplot, remember the warning about wide datasets.
- Use the Random Number Trick to determine relevant variables.
    - Remember to check the accuracy of your model, before making any decisions.
        - If your accuracy is under 85%, you should improve your model or slect a new target variable.

Producing a correlation matrix

This is going to be a little bit of a challenge, considering there are 233 features. It is just going to be really large and hard to tell what is correlated or not. Below is a correlation matrix with all the data.

In [48]:
```
corrmat = df_pox.corr()
f, ax = plt.subplots(figsize=(12,10)) #setting some parameters of the plot to help
sns.heatmap(corrmat, vmax = .8, square=True)
```

Out[48]:
```
<Axes: >
```



It is going to be necessary to first pair down the features. This will be done following the method in the lecture material of creating a random number to use as a feature, and then selecting only features that perform better than the random number.

In [49]:
```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier

y = df_pox.HAD_CPOX
x = df_pox.drop(['HAD_CPOX'], axis=1)
```

In [50]:
```
# Creating the random number
np.random.seed(42)
x['random'] = np.random.normal(0.0, 1.0, size=x.shape[0])
```

```
C:\Users\matth\AppData\Local\Temp\ipykernel_18632\1480610971.py:3: PerformanceWarni
ng: DataFrame is highly fragmented.  This is usually the result of calling `frame.i
nsert` many times, which has poor performance.  Consider joining all columns at onc
e using pd.concat(axis=1) instead. To get a de-fragmented frame, use `newframe = fr
ame.copy()`
  x['random'] = np.random.normal(0.0, 1.0, size=x.shape[0])
```

In [51]:
```
# Checking that the random number was added
x.head(10)
```

Out[51]:

	SEQNUMC	SEQNUMHH	PDAT	PROVWT_D	PROVWT_D_TERR	RDDWT_D	RDDWT_D_TERR
1	21	2	1	806.846012	806.846012	454.860417	454.860417
3	41	4	1	63.448686	63.448686	36.965931	36.965931
4	51	5	1	94.872632	94.872632	64.620204	64.620204
5	52	5	1	152.273845	152.273845	85.219413	85.219413
6	61	6	1	210.186351	210.186351	112.170514	112.170514
7	71	7	1	204.953336	204.953336	142.607339	142.607339
8	81	8	1	1016.753531	1016.753531	499.775831	499.775831
11	111	11	1	390.532585	390.532585	177.088881	177.088881
12	121	12	1	248.745510	248.745510	171.865720	171.865720
13	131	13	1	489.064864	489.064864	396.329703	396.329703

10 rows × 461 columns

In [52]:
```
# Creating a 70/30 train test split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)

clf=RandomForestClassifier(n_estimators=100)
clf.fit(x_train,y_train)
```

Out[52]:

```
▾ RandomForestClassifier

RandomForestClassifier()
```

In [53]:
```
# Running the random forest
features = x_train.columns
importances = clf.feature_importances_
std = np.std([tree.feature_importances_ for tree in clf.estimators_],
             axis=0)
indices = np.argsort(importances)[::-1]

# Save the feature ranking to a list for later use
# and print it on the screen

feature_rank = []
print("Feature ranking:")

for f in range(x_train.shape[1]):
    feature = f"{f + 1}. feature {features[indices[f]]}   \t{importances[indices[f]
    if 'random' in features[indices[f]]:
        feature += " <=="
    print(feature)
    feature_rank.append([features[indices[f]], importances[indices[f]]] )
```

```
Feature ranking:
1. feature AGECPOXR     39.53%
2. feature SEQNUMC      1.00%
3. feature SEQNUMHH     0.98%
4. feature random       0.97% <==
5. feature STRATUM      0.93%
6. feature BF_ENDR06    0.85%
7. feature PROVWT_D     0.83%
8. feature PROVWT_D_TERR        0.80%
9. feature EST_GRANT    0.78%
10. feature DDTP4       0.78%
11. feature RDDWT_D_TERR        0.76%
12. feature DFLU1       0.73%
13. feature ESTIAP14    0.71%
14. feature DDTP2       0.69%
15. feature BF_FORMR08          0.69%
16. feature DHEPB2      0.67%
17. feature DHIB1       0.65%
18. feature DHEPA2      0.64%
19. feature RDDWT_D     0.63%
20. feature DPCV1       0.63%
21. feature DVRC1       0.62%
22. feature DPCV4       0.62%
23. feature DPCV2       0.62%
24. feature DDTP1       0.62%
25. feature BF_EXCLR06          0.61%
26. feature DHIB3       0.61%
27. feature STATE       0.60%
28. feature DHIB2       0.60%
29. feature NUM_CELLS_HH        0.60%
30. feature DPCV3       0.59%
31. feature DDTP3       0.59%
32. feature DPOLIO3     0.59%
33. feature INCPORAR    0.57%
34. feature DROT3       0.55%
35. feature DPOLIO1     0.55%
36. feature EDUC1       0.55%
37. feature DMMR1       0.55%
38. feature DHIB4       0.55%
39. feature DTP4_AGE    0.54%
40. feature DPOLIO2     0.54%
41. feature DFLU2       0.52%
42. feature INCQ298A    0.51%
43. feature NUM_CELLS_PARENTS           0.51%
44. feature FLU1_AGE    0.50%
45. feature DHEPB3      0.48%
46. feature HEP2_AGE    0.47%
47. feature DROT2       0.47%
48. feature DHEPB1      0.46%
49. feature DHEPA1      0.45%
50. feature VRC1_AGE    0.45%
51. feature C1R         0.44%
52. feature HIB4_AGE    0.43%
53. feature PCV4_AGE    0.43%
54. feature HIB3_AGE    0.42%
55. feature DFLU3       0.42%
56. feature HEA1_AGE    0.40%
57. feature PCV2_AGE    0.39%
58. feature DROT1       0.39%
```

```
59. feature D6R        0.38%
60. feature INS_11     0.38%
61. feature HEA2_AGE   0.38%
62. feature C5R        0.37%
63. feature FLU2_AGE   0.36%
64. feature POL3_AGE   0.36%
65. feature DTP3_AGE   0.36%
66. feature MMR1_AGE   0.35%
67. feature INCPOV1    0.35%
68. feature PROV_FAC   0.35%
69. feature INS_2      0.34%
70. feature CHILDNM    0.34%
71. feature HEP3_AGE   0.34%
72. feature CEN_REG    0.33%
73. feature DHEPB4     0.33%
74. feature P_NUMVRC   0.31%
75. feature CWIC_02    0.31%
76. feature P_NUMFLUN          0.31%
77. feature P_NUMDIH   0.30%
78. feature DTP2_AGE   0.29%
79. feature HIB1_AGE   0.28%
80. feature FLU3_AGE   0.28%
81. feature P_NUMDHI   0.28%
82. feature N_PRVR     0.27%
83. feature P_NUMFLU   0.26%
84. feature HIB2_AGE   0.26%
85. feature INS_3A     0.26%
86. feature MARITAL2   0.25%
87. feature ROT3_AGE   0.25%
88. feature PCV1_AGE   0.25%
89. feature P_NUMHS    0.25%
90. feature HEP1_AGE   0.25%
91. feature POL2_AGE   0.24%
92. feature AGEGRP     0.24%
93. feature INS_1      0.23%
94. feature P_NUMPCC13         0.23%
95. feature P_NUHIBX   0.23%
96. feature FLU4_AGE   0.23%
97. feature INS_3      0.23%
98. feature RENT_OWN   0.23%
99. feature P_NUHEPX   0.22%
100. feature POL4_AGE          0.22%
101. feature P_NUMFLUM         0.22%
102. feature P_NUMHIB          0.22%
103. feature DPOLIO4   0.21%
104. feature CBF_01    0.21%
105. feature P_NUMHM   0.20%
106. feature P_NUMDTA          0.20%
107. feature PCV3_AGE          0.20%
108. feature P_NUMHEA          0.20%
109. feature INS_6     0.20%
110. feature SEX       0.20%
111. feature DTP1_AGE          0.20%
112. feature P_NUMROT          0.20%
113. feature P_NUMHEP          0.20%
114. feature REGISTRY          0.19%
115. feature RACEETHK          0.19%
116. feature VRC2_AGE          0.19%
117. feature XPOLTY1   0.18%
```

```
118. feature P_NUMVRX           0.18%
119. feature NUM_PHONE          0.17%
120. feature VFC_ORDER          0.17%
121. feature DFLU4      0.17%
122. feature ROT1_AGE           0.17%
123. feature INTRP      0.17%
124. feature P_U12VRC           0.16%
125. feature P_NUMPOL           0.16%
126. feature XDTPTY2    0.16%
127. feature XPOLTY3    0.16%
128. feature FRSTBRN    0.16%
129. feature INS_4_5    0.16%
130. feature XPOLTY2    0.16%
131. feature RACE_K     0.16%
132. feature XHEPTY2    0.15%
133. feature ROT2_AGE           0.15%
134. feature XHEPTY3    0.15%
135. feature P_NUMPCV           0.15%
136. feature P_NUMDTP           0.15%
137. feature P_NUMMMRX          0.14%
138. feature P_NUMIPV           0.14%
139. feature POL1_AGE           0.13%
140. feature XDTPTY3    0.13%
141. feature M_AGEGRP           0.13%
142. feature P_NUMPCC           0.13%
143. feature DDTP5      0.13%
144. feature HEP_BRTH           0.13%
145. feature CWIC_01    0.13%
146. feature DTP5_AGE           0.13%
147. feature P_NUMHIN           0.13%
148. feature XPCVTY4    0.12%
149. feature HEP4_AGE           0.12%
150. feature I_HISP_K           0.12%
151. feature XHEPTY1    0.12%
152. feature DVRC2      0.12%
153. feature P_NUMHEN           0.12%
154. feature LANGUAGE           0.11%
155. feature XMMRTY1    0.11%
156. feature P_UTDHEPA2         0.11%
157. feature PU431331           0.11%
158. feature P_NUMHG    0.11%
159. feature MMR2_AGE           0.11%
160. feature P_UTDTP4           0.11%
161. feature P_NUMRG    0.10%
162. feature P_NUMRM    0.10%
163. feature U2D_HEP    0.10%
164. feature P_UTDMCV           0.10%
165. feature P_UTD431H_ROUT_S           0.10%
166. feature DMMR2      0.10%
167. feature XDTPTY1    0.10%
168. feature HEP5_AGE           0.10%
169. feature XHEPTY4    0.10%
170. feature MOBIL_I    0.10%
171. feature XPCVTY2    0.10%
172. feature PUT43133           0.09%
173. feature P_UTDPCVB13        0.09%
174. feature P_NUMMRV           0.09%
175. feature P_NUMMMX           0.09%
176. feature P_NUMHION          0.09%
```

```
177. feature P_NUMMMR           0.09%
178. feature XHIBTY3    0.08%
179. feature P_UTD431H31_ROUT_S         0.08%
180. feature XDTPTY4    0.08%
181. feature P_NUMTPN           0.08%
182. feature XPCVTY1    0.08%
183. feature DPOLIO5    0.07%
184. feature P_UTD431H3_ROUT_S          0.07%
185. feature P_NUMPCC7          0.07%
186. feature XPOLTY4    0.07%
187. feature P_UTDHIB_ROUT_S    0.07%
188. feature U3D_HEP    0.07%
189. feature P_UTD431H313_ROUT_S        0.07%
190. feature PU431_31           0.07%
191. feature P_UTD331           0.07%
192. feature P_NUMFLUL          0.07%
193. feature P_UTDHEPA1         0.06%
194. feature P_UTD431H314_ROUT_S        0.06%
195. feature FLU5_AGE           0.06%
196. feature XHIBTY1    0.06%
197. feature PU4313314          0.06%
198. feature XHEPTY5    0.06%
199. feature DHIB5      0.06%
200. feature PUTD4313           0.06%
201. feature PCV5_AGE           0.05%
202. feature P_NUMPCN           0.05%
203. feature P_UTD431           0.05%
204. feature D7         0.05%
205. feature U1D_HEP    0.05%
206. feature HIB5_AGE           0.05%
207. feature DPCV5      0.05%
208. feature P_UTDMMX           0.05%
209. feature DFLU6      0.04%
210. feature P_NUMRO    0.04%
211. feature PU4313313          0.04%
212. feature DFLU5      0.04%
213. feature XPCVTY3    0.04%
214. feature P_UTDHIB_SHORT_S           0.04%
215. feature PU431_314          0.04%
216. feature P_UTDPOL           0.03%
217. feature P_UTDROT_S         0.03%
218. feature P_UTDPCV           0.03%
219. feature DHEPA3     0.03%
220. feature XHIBTY2    0.03%
221. feature FLU6_AGE           0.03%
222. feature XMMRTY2    0.02%
223. feature P_NUHPHB           0.02%
224. feature P_UTDHEP           0.02%
225. feature P_UTDHIB           0.02%
226. feature P_NUMOLN           0.02%
227. feature HEA3_AGE           0.02%
228. feature POL5_AGE           0.01%
229. feature XHIBTY4    0.01%
230. feature XPOLTY5    0.01%
231. feature P_NUMHHY           0.01%
232. feature P_UTDPC3           0.01%
233. feature DHEPB5     0.01%
234. feature P_UTDTP3           0.00%
235. feature XPCVTY5    0.00%
```

```
236. feature DHEPB6     0.00%
237. feature BFENDFL06          0.00%
238. feature BFFORMFL06         0.00%
239. feature P_NUMMP    0.00%
240. feature P_NUMMPR           0.00%
241. feature PDAT       0.00%
242. feature DHEPB7     0.00%
243. feature DHEPB8     0.00%
244. feature DHEPB9     0.00%
245. feature SHOTCARD           0.00%
246. feature P_NUMMCN           0.00%
247. feature YEAR       0.00%
248. feature DHEPA8     0.00%
249. feature DHIB6      0.00%
250. feature DHIB7      0.00%
251. feature DHIB8      0.00%
252. feature DHIB9      0.00%
253. feature DHEPA9     0.00%
254. feature P_NUMOPV           0.00%
255. feature DHEPA7     0.00%
256. feature DMMR3      0.00%
257. feature HEP_FLAG           0.00%
258. feature DDTP6      0.00%
259. feature DDTP7      0.00%
260. feature DDTP8      0.00%
261. feature P_NUMVRN           0.00%
262. feature DDTP9      0.00%
263. feature P_NUMRB    0.00%
264. feature P_NUMPCCN          0.00%
265. feature DFLU7      0.00%
266. feature DHEPA6     0.00%
267. feature DFLU8      0.00%
268. feature DFLU9      0.00%
269. feature P_NUMDAH           0.00%
270. feature P_NUMMSR           0.00%
271. feature DHEPA4     0.00%
272. feature DHEPA5     0.00%
273. feature P_NUMMSM           0.00%
274. feature P_NUMMS    0.00%
275. feature P_NUMPCP           0.00%
276. feature DRB1       0.00%
277. feature DMMR4      0.00%
278. feature VRC9_AGE           0.00%
279. feature ROT7_AGE           0.00%
280. feature ROT8_AGE           0.00%
281. feature ROT9_AGE           0.00%
282. feature VRC3_AGE           0.00%
283. feature VRC4_AGE           0.00%
284. feature VRC5_AGE           0.00%
285. feature VRC6_AGE           0.00%
286. feature VRC7_AGE           0.00%
287. feature VRC8_AGE           0.00%
288. feature XDTPTY5    0.00%
289. feature MPR5_AGE           0.00%
290. feature XDTPTY6    0.00%
291. feature XDTPTY7    0.00%
292. feature XDTPTY8    0.00%
293. feature XDTPTY9    0.00%
294. feature XFLUTY1    0.00%
```

```
295. feature XFLUTY2    0.00%
296. feature XFLUTY3    0.00%
297. feature XFLUTY4    0.00%
298. feature XFLUTY5    0.00%
299. feature ROT6_AGE           0.00%
300. feature ROT5_AGE           0.00%
301. feature ROT4_AGE           0.00%
302. feature RB9_AGE    0.00%
303. feature MPR7_AGE           0.00%
304. feature MPR8_AGE           0.00%
305. feature MPR9_AGE           0.00%
306. feature PCV6_AGE           0.00%
307. feature PCV7_AGE           0.00%
308. feature PCV8_AGE           0.00%
309. feature PCV9_AGE           0.00%
310. feature POL6_AGE           0.00%
311. feature POL7_AGE           0.00%
312. feature POL8_AGE           0.00%
313. feature POL9_AGE           0.00%
314. feature RB1_AGE    0.00%
315. feature RB2_AGE    0.00%
316. feature RB3_AGE    0.00%
317. feature RB4_AGE    0.00%
318. feature RB5_AGE    0.00%
319. feature RB6_AGE    0.00%
320. feature RB7_AGE    0.00%
321. feature RB8_AGE    0.00%
322. feature XFLUTY6    0.00%
323. feature XFLUTY7    0.00%
324. feature XFLUTY8    0.00%
325. feature XPOLTY7    0.00%
326. feature XPOLTY9    0.00%
327. feature XROTTY1    0.00%
328. feature XROTTY2    0.00%
329. feature XROTTY3    0.00%
330. feature XROTTY4    0.00%
331. feature XROTTY5    0.00%
332. feature XROTTY6    0.00%
333. feature XROTTY7    0.00%
334. feature XROTTY8    0.00%
335. feature XROTTY9    0.00%
336. feature XVRCTY1    0.00%
337. feature XVRCTY2    0.00%
338. feature XVRCTY3    0.00%
339. feature XVRCTY4    0.00%
340. feature XVRCTY5    0.00%
341. feature XVRCTY6    0.00%
342. feature XVRCTY7    0.00%
343. feature XVRCTY8    0.00%
344. feature XVRCTY9    0.00%
345. feature XPOLTY8    0.00%
346. feature XPOLTY6    0.00%
347. feature XFLUTY9    0.00%
348. feature XPCVTY9    0.00%
349. feature XHEPTY6    0.00%
350. feature XHEPTY7    0.00%
351. feature XHEPTY8    0.00%
352. feature XHEPTY9    0.00%
353. feature XHIBTY5    0.00%
```

```
354. feature XHIBTY6    0.00%
355. feature XHIBTY7    0.00%
356. feature XHIBTY8    0.00%
357. feature XHIBTY9    0.00%
358. feature XMMRTY3    0.00%
359. feature XMMRTY4    0.00%
360. feature XMMRTY5    0.00%
361. feature XMMRTY6    0.00%
362. feature XMMRTY7    0.00%
363. feature XMMRTY8    0.00%
364. feature XMMRTY9    0.00%
365. feature XPCVTY6    0.00%
366. feature XPCVTY7    0.00%
367. feature XPCVTY8    0.00%
368. feature MPR6_AGE           0.00%
369. feature MPR4_AGE           0.00%
370. feature DMMR5      0.00%
371. feature DRB5       0.00%
372. feature DPCV7      0.00%
373. feature DPCV8      0.00%
374. feature DPCV9      0.00%
375. feature DPOLIO6    0.00%
376. feature DPOLIO7    0.00%
377. feature DPOLIO8    0.00%
378. feature DPOLIO9    0.00%
379. feature DRB2       0.00%
380. feature DRB4       0.00%
381. feature DRB6       0.00%
382. feature MPR3_AGE           0.00%
383. feature DRB7       0.00%
384. feature DRB8       0.00%
385. feature DRB9       0.00%
386. feature DROT4      0.00%
387. feature DROT5      0.00%
388. feature DROT6      0.00%
389. feature DROT7      0.00%
390. feature DROT8      0.00%
391. feature DROT9      0.00%
392. feature DPCV6      0.00%
393. feature DMPRB9     0.00%
394. feature DMPRB8     0.00%
395. feature DMPRB7     0.00%
396. feature DMMR6      0.00%
397. feature DMMR7      0.00%
398. feature DMMR8      0.00%
399. feature DMMR9      0.00%
400. feature DMP1       0.00%
401. feature DMP2       0.00%
402. feature DMP3       0.00%
403. feature DMP4       0.00%
404. feature DMP5       0.00%
405. feature DMP6       0.00%
406. feature DMP7       0.00%
407. feature DMP8       0.00%
408. feature DMP9       0.00%
409. feature DMPRB1     0.00%
410. feature DMPRB2     0.00%
411. feature DMPRB3     0.00%
412. feature DMPRB4     0.00%
```

```
413. feature DMPRB5     0.00%
414. feature DMPRB6     0.00%
415. feature DVRC3      0.00%
416. feature DVRC4      0.00%
417. feature DVRC5      0.00%
418. feature HIB7_AGE           0.00%
419. feature HIB9_AGE           0.00%
420. feature MMR3_AGE           0.00%
421. feature MMR4_AGE           0.00%
422. feature MMR5_AGE           0.00%
423. feature MMR6_AGE           0.00%
424. feature MMR7_AGE           0.00%
425. feature MMR8_AGE           0.00%
426. feature MMR9_AGE           0.00%
427. feature MP1_AGE    0.00%
428. feature MP2_AGE    0.00%
429. feature MP3_AGE    0.00%
430. feature MP4_AGE    0.00%
431. feature MP5_AGE    0.00%
432. feature MP6_AGE    0.00%
433. feature MP7_AGE    0.00%
434. feature MP8_AGE    0.00%
435. feature MP9_AGE    0.00%
436. feature MPR1_AGE           0.00%
437. feature MPR2_AGE           0.00%
438. feature HIB8_AGE           0.00%
439. feature HIB6_AGE           0.00%
440. feature DVRC6      0.00%
441. feature HEP9_AGE           0.00%
442. feature DVRC7      0.00%
443. feature DVRC8      0.00%
444. feature DVRC9      0.00%
445. feature DTP6_AGE           0.00%
446. feature DTP7_AGE           0.00%
447. feature DTP8_AGE           0.00%
448. feature DTP9_AGE           0.00%
449. feature FLU7_AGE           0.00%
450. feature FLU8_AGE           0.00%
451. feature FLU9_AGE           0.00%
452. feature HEA4_AGE           0.00%
453. feature HEA5_AGE           0.00%
454. feature HEA6_AGE           0.00%
455. feature HEA7_AGE           0.00%
456. feature HEA8_AGE           0.00%
457. feature HEA9_AGE           0.00%
458. feature HEP6_AGE           0.00%
459. feature HEP7_AGE           0.00%
460. feature HEP8_AGE           0.00%
461. feature DRB3       0.00%
```

Yikes!!! That is not good, there is only a single feature, Age at which child had chicken pox, that is a better indicator of if a child has had chicken pox, then a random number. That is just not going to work well for building a model. Lets try another possibility, Rent or Owning a home.

In [54]:
```
df_rentown = df_full
```

In [55]:
```
# Unique values in Rent or Own
df_rentown['RENT_OWN'].unique()
```

Out[55]:
```
array([ 2,  1,  3, 99, 77], dtype=int64)
```

The values in Rent / Own are either: 1 own (or buying) , 2 renting, 3 (other arangement 3% of the data), or missing / unknown. I am going to drop anything other than 1 or 2.

In [56]:
```
df_rentown = df_rentown[df_rentown.RENT_OWN != 3]
df_rentown = df_rentown[df_rentown.RENT_OWN != 77]
df_rentown = df_rentown[df_rentown.RENT_OWN != 99]
```

In [57]:
```
# Unique values in Rent or Own
df_rentown['RENT_OWN'].unique()
```

Out[57]:
```
array([2, 1], dtype=int64)
```

In [58]:
```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier

y = df_rentown.RENT_OWN
x = df_rentown.drop(['RENT_OWN'], axis=1)
```

In [59]:
```
# Creating the random number
np.random.seed(42)
x['random'] = np.random.normal(0.0, 1.0, size=x.shape[0])
```

```
C:\Users\matth\AppData\Local\Temp\ipykernel_18632\1480610971.py:3: PerformanceWarni
ng: DataFrame is highly fragmented.  This is usually the result of calling `frame.i
nsert` many times, which has poor performance.  Consider joining all columns at onc
e using pd.concat(axis=1) instead. To get a de-fragmented frame, use `newframe = fr
ame.copy()`
  x['random'] = np.random.normal(0.0, 1.0, size=x.shape[0])
```

In [60]:
```
# Checking that the random number was added
x.head(10)
```

Out[60]:

	SEQNUMC	SEQNUMHH	PDAT	PROVWT_D	PROVWT_D_TERR	RDDWT_D	RDDWT_D_TERR
1	21	2	1	806.846012	806.846012	454.860417	454.860417
3	41	4	1	63.448686	63.448686	36.965931	36.965931
4	51	5	1	94.872632	94.872632	64.620204	64.620204
5	52	5	1	152.273845	152.273845	85.219413	85.219413
6	61	6	1	210.186351	210.186351	112.170514	112.170514
7	71	7	1	204.953336	204.953336	142.607339	142.607339
8	81	8	1	1016.753531	1016.753531	499.775831	499.775831
11	111	11	1	390.532585	390.532585	177.088881	177.088881
12	121	12	1	248.745510	248.745510	171.865720	171.865720
13	131	13	1	489.064864	489.064864	396.329703	396.329703

10 rows × 461 columns

In [61]:

```
# Creating a 70/30 train test split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)

clf=RandomForestClassifier(n_estimators=100)
clf.fit(x_train,y_train)
```

Out[61]:

```
▾ RandomForestClassifier

RandomForestClassifier()
```

In [62]:

```
# Running the random forest
features = x_train.columns
importances = clf.feature_importances_
std = np.std([tree.feature_importances_ for tree in clf.estimators_],
             axis=0)
indices = np.argsort(importances)[::-1]

# Save the feature ranking to a list for later use
# and print it on the screen

feature_rank = []
print("Feature ranking:")

for f in range(x_train.shape[1]):
    feature = f"{f + 1}. feature {features[indices[f]]}   \t{importances[indices[f]
    if 'random' in features[indices[f]]:
        feature += " <=="
    print(feature)
    feature_rank.append([features[indices[f]], importances[indices[f]]] )
```

```
Feature ranking:
1. feature INCQ298A     3.91%
2. feature INCPORAR     3.56%
3. feature CWIC_02      2.87%
4. feature CWIC_01      2.32%
5. feature INS_1        2.04%
6. feature INCPOV1      1.99%
7. feature RDDWT_D      1.55%
8. feature PROVWT_D_TERR        1.54%
9. feature INTRP        1.51%
10. feature PROVWT_D    1.46%
11. feature RDDWT_D_TERR        1.41%
12. feature STRATUM     1.40%
13. feature EDUC1       1.37%
14. feature SEQNUMC     1.22%
15. feature random      1.21% <==
16. feature MARITAL2    1.20%
17. feature NUM_CELLS_HH        1.18%
18. feature SEQNUMHH    1.11%
19. feature ESTIAP14    1.11%
20. feature EST_GRANT           1.09%
21. feature STATE       1.04%
22. feature DPCV3       1.02%
23. feature DDTP4       0.97%
24. feature DHIB3       0.97%
25. feature DDTP3       0.96%
26. feature DHEPB3      0.96%
27. feature DPOLIO3     0.96%
28. feature DMMR1       0.96%
29. feature DPCV4       0.95%
30. feature BF_ENDR06           0.95%
31. feature DVRC1       0.94%
32. feature DHEPB2      0.92%
33. feature DFLU1       0.92%
34. feature DHEPA1      0.92%
35. feature RACEETHK    0.90%
36. feature DPCV2       0.88%
37. feature DPCV1       0.87%
38. feature DHIB4       0.87%
39. feature DHIB2       0.87%
40. feature DDTP2       0.85%
41. feature MOBIL_I     0.85%
42. feature NUM_CELLS_PARENTS           0.84%
43. feature DPOLIO2     0.84%
44. feature C1R         0.83%
45. feature DHEPA2      0.83%
46. feature DDTP1       0.82%
47. feature DROT2       0.81%
48. feature DFLU2       0.81%
49. feature DHIB1       0.79%
50. feature DROT1       0.78%
51. feature M_AGEGRP    0.78%
52. feature INS_2       0.77%
53. feature DPOLIO1     0.77%
54. feature BF_FORMR08          0.74%
55. feature RACE_K      0.71%
56. feature BF_EXCLR06          0.70%
57. feature DROT3       0.66%
58. feature LANGUAGE    0.64%
```

```
59. feature FLU1_AGE    0.63%
60. feature FLU2_AGE    0.61%
61. feature DFLU3       0.60%
62. feature DHEPB1      0.57%
63. feature DTP4_AGE    0.56%
64. feature PROV_FAC    0.53%
65. feature I_HISP_K    0.52%
66. feature INS_3A      0.51%
67. feature HIB4_AGE    0.51%
68. feature PCV4_AGE    0.50%
69. feature HEA2_AGE    0.49%
70. feature FLU3_AGE    0.48%
71. feature HEP3_AGE    0.47%
72. feature C5R         0.46%
73. feature MMR1_AGE    0.46%
74. feature HEA1_AGE    0.45%
75. feature CEN_REG     0.44%
76. feature DHEPB4      0.42%
77. feature HEP2_AGE    0.42%
78. feature INS_4_5     0.42%
79. feature VRC1_AGE    0.41%
80. feature P_NUMFLU    0.41%
81. feature HIB3_AGE    0.41%
82. feature NUM_PHONE           0.41%
83. feature P_NUMHS     0.40%
84. feature P_NUMFLUN           0.37%
85. feature D6R         0.37%
86. feature PCV3_AGE    0.36%
87. feature POL3_AGE    0.36%
88. feature P_NUMDIH    0.36%
89. feature P_NUHIBX    0.35%
90. feature INS_3       0.35%
91. feature DTP3_AGE    0.33%
92. feature REGISTRY    0.33%
93. feature CHILDNM     0.33%
94. feature P_NUHEPX    0.31%
95. feature VFC_ORDER           0.31%
96. feature AGEGRP      0.31%
97. feature P_NUMHM     0.31%
98. feature P_NUMDTA    0.29%
99. feature DPOLIO4     0.29%
100. feature POL2_AGE           0.27%
101. feature P_NUMRM    0.27%
102. feature PCV2_AGE           0.26%
103. feature P_NUMDHI           0.26%
104. feature XPOLTY2    0.25%
105. feature HIB2_AGE           0.25%
106. feature PCV1_AGE           0.24%
107. feature ROT3_AGE           0.24%
108. feature HEP4_AGE           0.24%
109. feature DFLU4      0.24%
110. feature DTP2_AGE           0.24%
111. feature P_NUMIPV           0.23%
112. feature XPOLTY3    0.23%
113. feature HEP1_AGE           0.23%
114. feature FLU4_AGE           0.23%
115. feature ROT1_AGE           0.23%
116. feature P_NUMROT           0.23%
117. feature ROT2_AGE           0.22%
```

```
118. feature INS_11     0.22%
119. feature XDTPTY3    0.22%
120. feature P_NUMHEA           0.22%
121. feature XDTPTY1    0.21%
122. feature N_PRVR     0.21%
123. feature XPOLTY1    0.21%
124. feature DTP1_AGE           0.21%
125. feature P_NUMHEP           0.21%
126. feature FRSTBRN    0.20%
127. feature P_NUMRG    0.20%
128. feature HIB1_AGE           0.20%
129. feature XDTPTY2    0.20%
130. feature POL4_AGE           0.20%
131. feature SEX        0.20%
132. feature P_NUMPCC13         0.20%
133. feature POL1_AGE           0.18%
134. feature P_NUMPCC           0.18%
135. feature P_NUMHIN           0.17%
136. feature XHEPTY3    0.17%
137. feature P_NUMHIB           0.17%
138. feature XDTPTY4    0.17%
139. feature XHEPTY4    0.17%
140. feature P_NUMPOL           0.16%
141. feature U1D_HEP    0.16%
142. feature CBF_01     0.15%
143. feature INS_6      0.15%
144. feature XHEPTY2    0.15%
145. feature HEP_BRTH           0.15%
146. feature P_UTDHEPA2         0.15%
147. feature U2D_HEP    0.14%
148. feature U3D_HEP    0.13%
149. feature P_NUMPCV           0.13%
150. feature P_NUMMMRX          0.13%
151. feature P_UTDROT_S         0.12%
152. feature P_NUMDTP           0.12%
153. feature XMMRTY1    0.12%
154. feature P_NUMVRX           0.12%
155. feature XPCVTY4    0.11%
156. feature XHEPTY1    0.11%
157. feature P_NUMFLUM          0.10%
158. feature XPCVTY1    0.10%
159. feature P_UTDHEPA1         0.10%
160. feature XPCVTY2    0.10%
161. feature XPCVTY3    0.09%
162. feature P_UTDHIB_ROUT_S    0.09%
163. feature P_UTD431H314_ROUT_S        0.09%
164. feature PU4313314          0.09%
165. feature P_NUMMMR           0.08%
166. feature P_NUMMMX           0.08%
167. feature P_UTD431H3_ROUT_S          0.08%
168. feature PU431_314          0.08%
169. feature P_UTD431H313_ROUT_S        0.08%
170. feature P_NUMMRV           0.08%
171. feature P_UTD431H31_ROUT_S         0.08%
172. feature XPOLTY4    0.07%
173. feature P_NUMVRC           0.07%
174. feature PU431331           0.07%
175. feature P_NUMPCC7          0.07%
176. feature P_UTD431           0.07%
```

```
177. feature PU4313313          0.07%
178. feature P_NUMRO    0.07%
179. feature P_UTD431H_ROUT_S           0.07%
180. feature P_UTDTP4           0.07%
181. feature PU431_31           0.06%
182. feature PUTD4313           0.06%
183. feature PUT43133           0.06%
184. feature P_UTDPCV           0.06%
185. feature P_NUMPCN           0.06%
186. feature P_UTDMMX           0.05%
187. feature P_NUMOLN           0.05%
188. feature P_NUMTPN           0.05%
189. feature DMMR2      0.05%
190. feature MMR2_AGE           0.05%
191. feature AGECPOXR           0.05%
192. feature P_U12VRC           0.05%
193. feature P_NUMFLUL          0.05%
194. feature DVRC2      0.05%
195. feature P_UTDMCV           0.05%
196. feature P_UTD331           0.05%
197. feature DTP5_AGE           0.04%
198. feature VRC2_AGE           0.04%
199. feature P_UTDHEP           0.04%
200. feature DHIB5      0.04%
201. feature P_UTDHIB           0.04%
202. feature HAD_CPOX           0.04%
203. feature P_NUMHION          0.04%
204. feature DPCV5      0.04%
205. feature DHEPB5     0.03%
206. feature XMMRTY2    0.03%
207. feature DDTP5      0.03%
208. feature XHIBTY4    0.03%
209. feature PCV5_AGE           0.03%
210. feature DHEPA3     0.03%
211. feature HIB5_AGE           0.03%
212. feature P_UTDPOL           0.03%
213. feature XHIBTY3    0.03%
214. feature HEP5_AGE           0.03%
215. feature HEA3_AGE           0.03%
216. feature P_UTDHIB_SHORT_S           0.03%
217. feature XHEPTY5    0.02%
218. feature P_UTDPC3           0.02%
219. feature XPCVTY5    0.02%
220. feature P_NUMHG    0.02%
221. feature XDTPTY5    0.02%
222. feature P_UTDTP3           0.02%
223. feature XHIBTY1    0.02%
224. feature P_NUMHHY           0.02%
225. feature P_UTDPCVB13        0.02%
226. feature D7 0.02%
227. feature DFLU5      0.01%
228. feature P_NUMMCN           0.01%
229. feature DPOLIO5    0.01%
230. feature P_NUMHEN           0.01%
231. feature P_NUMDAH           0.01%
232. feature P_NUHPHB           0.01%
233. feature FLU5_AGE           0.01%
234. feature XHIBTY2    0.01%
235. feature P_NUMPCP           0.01%
```

```
236. feature P_NUMVRN          0.01%
237. feature ROT4_AGE          0.01%
238. feature POL5_AGE          0.01%
239. feature P_NUMMS    0.01%
240. feature DROT4      0.01%
241. feature P_NUMOPV          0.01%
242. feature XPOLTY5    0.00%
243. feature DHIB6      0.00%
244. feature DPOLIO6    0.00%
245. feature FLU6_AGE          0.00%
246. feature XHEPTY6    0.00%
247. feature DFLU6      0.00%
248. feature PCV6_AGE          0.00%
249. feature DPCV6      0.00%
250. feature HEP_FLAG          0.00%
251. feature HIB6_AGE          0.00%
252. feature DVRC3      0.00%
253. feature HEP6_AGE          0.00%
254. feature XMMRTY3    0.00%
255. feature DDTP6      0.00%
256. feature HEA4_AGE          0.00%
257. feature XHEPTY7    0.00%
258. feature POL6_AGE          0.00%
259. feature MMR3_AGE          0.00%
260. feature PDAT       0.00%
261. feature DHEPA8     0.00%
262. feature XVRCTY8    0.00%
263. feature DHEPA9     0.00%
264. feature XVRCTY9    0.00%
265. feature YEAR       0.00%
266. feature DHEPA6     0.00%
267. feature P_NUMMSM          0.00%
268. feature BFFORMFL06        0.00%
269. feature BFENDFL06         0.00%
270. feature P_NUMMPR          0.00%
271. feature SHOTCARD          0.00%
272. feature DHEPA7     0.00%
273. feature XVRCTY3    0.00%
274. feature DHEPA5     0.00%
275. feature DDTP9      0.00%
276. feature P_NUMPCCN         0.00%
277. feature XVRCTY4    0.00%
278. feature XVRCTY5    0.00%
279. feature P_NUMRB    0.00%
280. feature DDTP7      0.00%
281. feature DDTP8      0.00%
282. feature DHEPB6     0.00%
283. feature DHEPA4     0.00%
284. feature P_NUMMP    0.00%
285. feature P_NUMMSR          0.00%
286. feature DFLU7      0.00%
287. feature DFLU8      0.00%
288. feature DFLU9      0.00%
289. feature XVRCTY6    0.00%
290. feature XVRCTY7    0.00%
291. feature XVRCTY2    0.00%
292. feature DHEPB7     0.00%
293. feature ROT7_AGE          0.00%
294. feature RB4_AGE    0.00%
```

```
295. feature RB5_AGE    0.00%
296. feature RB6_AGE    0.00%
297. feature RB7_AGE    0.00%
298. feature RB8_AGE    0.00%
299. feature RB9_AGE    0.00%
300. feature ROT5_AGE           0.00%
301. feature ROT6_AGE           0.00%
302. feature ROT8_AGE           0.00%
303. feature XDTPTY7    0.00%
304. feature ROT9_AGE           0.00%
305. feature VRC3_AGE           0.00%
306. feature VRC4_AGE           0.00%
307. feature VRC5_AGE           0.00%
308. feature VRC6_AGE           0.00%
309. feature VRC7_AGE           0.00%
310. feature VRC8_AGE           0.00%
311. feature VRC9_AGE           0.00%
312. feature RB3_AGE    0.00%
313. feature RB2_AGE    0.00%
314. feature RB1_AGE    0.00%
315. feature POL9_AGE           0.00%
316. feature MP7_AGE    0.00%
317. feature MP8_AGE    0.00%
318. feature MP9_AGE    0.00%
319. feature MPR1_AGE           0.00%
320. feature MPR2_AGE           0.00%
321. feature MPR3_AGE           0.00%
322. feature MPR4_AGE           0.00%
323. feature MPR5_AGE           0.00%
324. feature MPR6_AGE           0.00%
325. feature MPR7_AGE           0.00%
326. feature MPR8_AGE           0.00%
327. feature MPR9_AGE           0.00%
328. feature PCV7_AGE           0.00%
329. feature PCV8_AGE           0.00%
330. feature PCV9_AGE           0.00%
331. feature POL7_AGE           0.00%
332. feature POL8_AGE           0.00%
333. feature XDTPTY6    0.00%
334. feature XDTPTY8    0.00%
335. feature DHEPB8     0.00%
336. feature XPOLTY9    0.00%
337. feature XMMRTY9    0.00%
338. feature XPCVTY6    0.00%
339. feature XPCVTY7    0.00%
340. feature XPCVTY8    0.00%
341. feature XPCVTY9    0.00%
342. feature XPOLTY6    0.00%
343. feature XPOLTY7    0.00%
344. feature XPOLTY8    0.00%
345. feature XROTTY1    0.00%
346. feature XDTPTY9    0.00%
347. feature XROTTY2    0.00%
348. feature XROTTY3    0.00%
349. feature XROTTY4    0.00%
350. feature XROTTY5    0.00%
351. feature XROTTY6    0.00%
352. feature XROTTY7    0.00%
353. feature XROTTY8    0.00%
```

```
354. feature XROTTY9    0.00%
355. feature XMMRTY8    0.00%
356. feature XMMRTY7    0.00%
357. feature XMMRTY6    0.00%
358. feature XMMRTY5    0.00%
359. feature XFLUTY1    0.00%
360. feature XFLUTY2    0.00%
361. feature XFLUTY3    0.00%
362. feature XFLUTY4    0.00%
363. feature XFLUTY5    0.00%
364. feature XFLUTY6    0.00%
365. feature XFLUTY7    0.00%
366. feature XFLUTY8    0.00%
367. feature XFLUTY9    0.00%
368. feature XHEPTY8    0.00%
369. feature XHEPTY9    0.00%
370. feature XHIBTY5    0.00%
371. feature XHIBTY6    0.00%
372. feature XHIBTY7    0.00%
373. feature XHIBTY8    0.00%
374. feature XHIBTY9    0.00%
375. feature XMMRTY4    0.00%
376. feature MP6_AGE    0.00%
377. feature MP5_AGE    0.00%
378. feature MP4_AGE    0.00%
379. feature DPCV7      0.00%
380. feature DMPRB2     0.00%
381. feature DMPRB3     0.00%
382. feature DMPRB4     0.00%
383. feature DMPRB5     0.00%
384. feature DMPRB6     0.00%
385. feature DMPRB7     0.00%
386. feature DMPRB8     0.00%
387. feature DMPRB9     0.00%
388. feature DPCV8      0.00%
389. feature MP3_AGE    0.00%
390. feature DPCV9      0.00%
391. feature XVRCTY1    0.00%
392. feature DPOLIO7    0.00%
393. feature DPOLIO8    0.00%
394. feature DPOLIO9    0.00%
395. feature DRB1       0.00%
396. feature DRB2       0.00%
397. feature DRB4       0.00%
398. feature DMPRB1     0.00%
399. feature DMP9       0.00%
400. feature DMP8       0.00%
401. feature DMP7       0.00%
402. feature DHEPB9     0.00%
403. feature DHIB7      0.00%
404. feature DHIB8      0.00%
405. feature DHIB9      0.00%
406. feature DMMR3      0.00%
407. feature DMMR4      0.00%
408. feature DMMR5      0.00%
409. feature DMMR6      0.00%
410. feature DMMR7      0.00%
411. feature DMMR8      0.00%
412. feature DMMR9      0.00%
```

```
413. feature DMP1       0.00%
414. feature DMP2       0.00%
415. feature DMP3       0.00%
416. feature DMP4       0.00%
417. feature DMP5       0.00%
418. feature DMP6       0.00%
419. feature DRB5       0.00%
420. feature DRB6       0.00%
421. feature DRB7       0.00%
422. feature HEA5_AGE           0.00%
423. feature HEA7_AGE           0.00%
424. feature HEA8_AGE           0.00%
425. feature HEA9_AGE           0.00%
426. feature HEP7_AGE           0.00%
427. feature HEP8_AGE           0.00%
428. feature HEP9_AGE           0.00%
429. feature HIB7_AGE           0.00%
430. feature HIB8_AGE           0.00%
431. feature HIB9_AGE           0.00%
432. feature MMR4_AGE           0.00%
433. feature MMR5_AGE           0.00%
434. feature MMR6_AGE           0.00%
435. feature MMR7_AGE           0.00%
436. feature MMR8_AGE           0.00%
437. feature MMR9_AGE           0.00%
438. feature MP1_AGE    0.00%
439. feature MP2_AGE    0.00%
440. feature HEA6_AGE           0.00%
441. feature FLU9_AGE           0.00%
442. feature DRB8       0.00%
443. feature FLU8_AGE           0.00%
444. feature DRB9       0.00%
445. feature DROT5      0.00%
446. feature DROT6      0.00%
447. feature DROT7      0.00%
448. feature DROT8      0.00%
449. feature DROT9      0.00%
450. feature DVRC4      0.00%
451. feature DVRC5      0.00%
452. feature DVRC6      0.00%
453. feature DVRC7      0.00%
454. feature DVRC8      0.00%
455. feature DVRC9      0.00%
456. feature DTP6_AGE           0.00%
457. feature DTP7_AGE           0.00%
458. feature DTP8_AGE           0.00%
459. feature DTP9_AGE           0.00%
460. feature FLU7_AGE           0.00%
461. feature DRB3       0.00%
```

That looks much better! There are 14 features that are better than a random number. Those are going to be pulled out.

In [63]:
```
top_ranks = feature_rank[:14]
top_ranks
```

Out[63]:
```
[['INCQ298A', 0.0391078205871675],
 ['INCPORAR', 0.035646692927725536],
 ['CWIC_02', 0.028654993902786093],
 ['CWIC_01', 0.023230175100834492],
 ['INS_1', 0.02038651884983103],
 ['INCPOV1', 0.019937656568219],
 ['RDDWT_D', 0.015504673576702621],
 ['PROVWT_D_TERR', 0.015420602721063194],
 ['INTRP', 0.015110562235401708],
 ['PROVWT_D', 0.014569486403276415],
 ['RDDWT_D_TERR', 0.014149759071087026],
 ['STRATUM', 0.014039891742704431],
 ['EDUC1', 0.013715384235440805],
 ['SEQNUMC', 0.012218604653439855]]
```

Creating a dataset with just those features AND the target, Rent Own

In [64]:
```python
top_rank_cols = [s[0].split(',')[0] for s in top_ranks]
top_rank_cols.append('RENT_OWN')
top_rank_cols
```

Out[64]:
```
['INCQ298A',
 'INCPORAR',
 'CWIC_02',
 'CWIC_01',
 'INS_1',
 'INCPOV1',
 'RDDWT_D',
 'PROVWT_D_TERR',
 'INTRP',
 'PROVWT_D',
 'RDDWT_D_TERR',
 'STRATUM',
 'EDUC1',
 'SEQNUMC',
 'RENT_OWN']
```

In [65]:
```python
# Creating a clean copy of the dataset to prevent messing with the OG set
top_rank_df = df_rentown[top_rank_cols].copy()
```

In [66]:
```python
top_rank_df.head(10)
```

Out[66]:

	INCQ298A	INCPORAR	CWIC_02	CWIC_01	INS_1	INCPOV1	RDDWT_D	PROVWT_D_TERR	IN
1	4	0.500000	0.0	2	2.0	3	454.860417	806.846012	
3	14	3.000000	0.0	2	1.0	1	36.965931	63.448686	
4	3	0.500000	1.0	1	2.0	3	64.620204	94.872632	
5	3	0.500000	1.0	1	2.0	3	85.219413	152.273845	
6	5	1.089867	1.0	1	2.0	2	112.170514	210.186351	
7	14	3.000000	0.0	2	1.0	1	142.607339	204.953336	
8	14	3.000000	0.0	2	1.0	1	499.775831	1016.753531	
11	10	1.438797	0.0	2	2.0	2	177.088881	390.532585	
12	9	1.481544	1.0	1	2.0	2	171.865720	248.745510	
13	5	0.639352	1.0	1	2.0	3	396.329703	489.064864	

Pair Plot of the data set

In [67]:

```
sns.pairplot(data=top_rank_df)
```

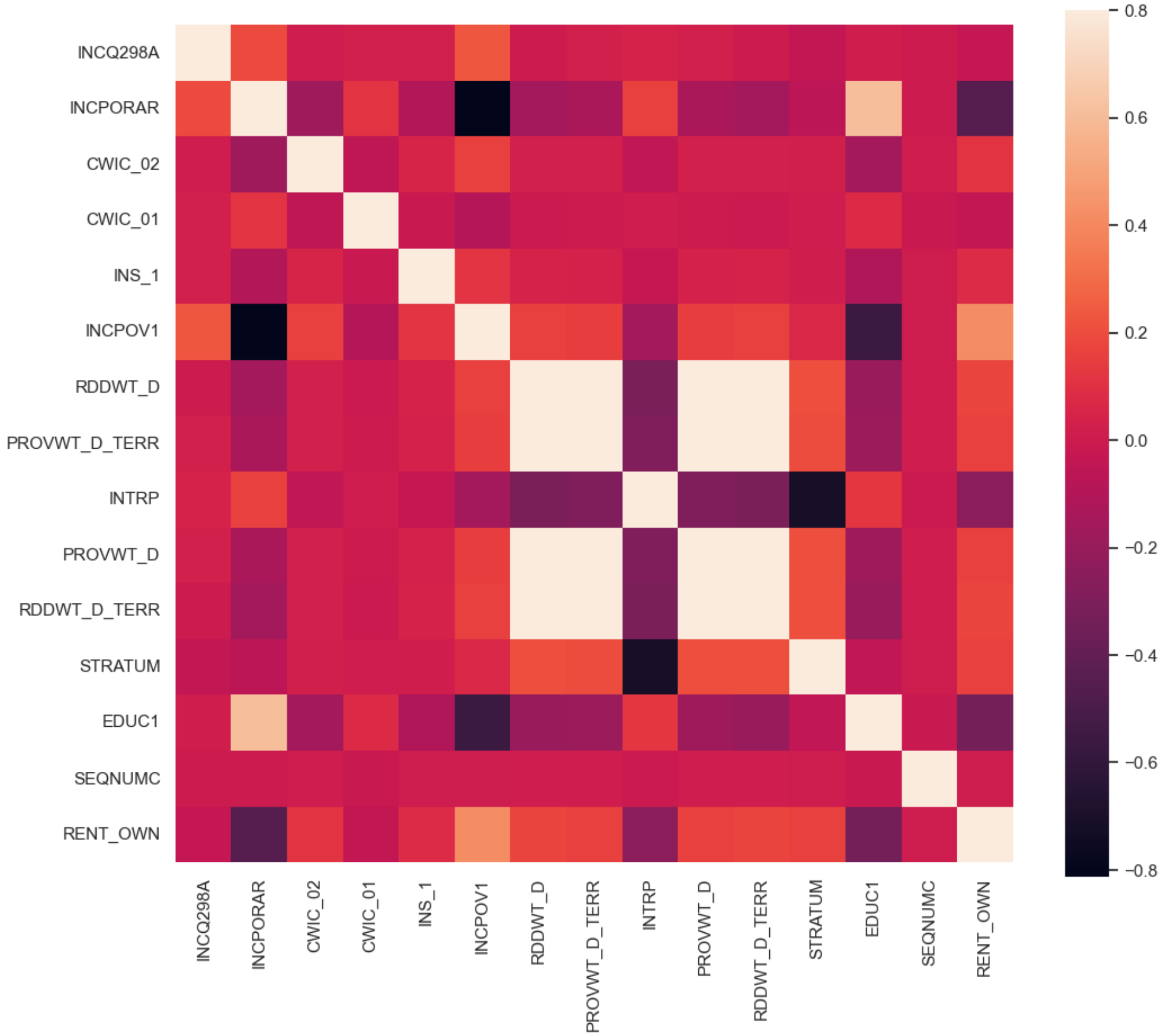
Out[67]:

```
<seaborn.axisgrid.PairGrid at 0x1bb44b7fbb0>
```

New Correlation Matrix of the data

In [68]:
```
cormat = top_rank_df.corr()
f, ax = plt.subplots(figsize=(12,10)) #setting some parameters of the plot to help
sns.heatmap(cormat, vmax = .8, square=True)
```

Out[68]: <Axes: >



Looking at the correlation plot, owning / renting a home are closely associated with income and education features.

In [69]:
```
y = top_rank_df.RENT_OWN
x = top_rank_df.drop(['RENT_OWN'], axis=1)
```

In [70]:
```
# Creating a 70/30 train test split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)

clf=RandomForestClassifier(n_estimators=100)
clf.fit(x_train,y_train)
```

Out[70]:
```
▾ RandomForestClassifier
RandomForestClassifier()
```

In [71]:
```
y_pred=clf.predict(x_test)
```

In [72]:
```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7555299539170507
```

Unfortunately the accuracy of the model is only 75%. I am going to take a look at a new variable, maritial status.

In [73]:
```
df_marital = df_full
```

In [74]:
```
y = df_marital.MARITAL2
x = df_marital.drop(['MARITAL2'], axis=1)
```

In [75]:
```
# Creating a 70/30 train test split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)

clf=RandomForestClassifier(n_estimators=100)
clf.fit(x_train,y_train)
```

Out[75]:
```
▾ RandomForestClassifier
RandomForestClassifier()
```

In [76]:
```
y_pred=clf.predict(x_test)
```

In [77]:
```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8488269145639663
```

All right, a model that gives an 85% accuracy. Let us see what the features are in the model.

In [78]:
```
# Creating the random number
np.random.seed(42)
x['random'] = np.random.normal(0.0, 1.0, size=x.shape[0])
```

```
C:\Users\matth\AppData\Local\Temp\ipykernel_18632\1480610971.py:3: PerformanceWarni
ng: DataFrame is highly fragmented.  This is usually the result of calling `frame.i
nsert` many times, which has poor performance.  Consider joining all columns at onc
e using pd.concat(axis=1) instead. To get a de-fragmented frame, use `newframe = fr
ame.copy()`
  x['random'] = np.random.normal(0.0, 1.0, size=x.shape[0])
```

In [79]:
```
# Checking that the random number was added
x.head(10)
```

Out[79]:

	SEQNUMC	SEQNUMHH	PDAT	PROVWT_D	PROVWT_D_TERR	RDDWT_D	RDDWT_D_TERR
1	21	2	1	806.846012	806.846012	454.860417	454.860417
3	41	4	1	63.448686	63.448686	36.965931	36.965931
4	51	5	1	94.872632	94.872632	64.620204	64.620204
5	52	5	1	152.273845	152.273845	85.219413	85.219413
6	61	6	1	210.186351	210.186351	112.170514	112.170514
7	71	7	1	204.953336	204.953336	142.607339	142.607339
8	81	8	1	1016.753531	1016.753531	499.775831	499.775831
11	111	11	1	390.532585	390.532585	177.088881	177.088881
12	121	12	1	248.745510	248.745510	171.865720	171.865720
13	131	13	1	489.064864	489.064864	396.329703	396.329703

10 rows × 461 columns

In [80]:

```
# Creating a 70/30 train test split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)

clf=RandomForestClassifier(n_estimators=100)
clf.fit(x_train,y_train)
```

Out[80]:

```
▾ RandomForestClassifier

RandomForestClassifier()
```

In [81]:

```
# Running the random forest
features = x_train.columns
importances = clf.feature_importances_
std = np.std([tree.feature_importances_ for tree in clf.estimators_],
             axis=0)
indices = np.argsort(importances)[::-1]

# Save the feature ranking to a list for later use
# and print it on the screen

feature_rank = []
print("Feature ranking:")

for f in range(x_train.shape[1]):
    feature = f"{f + 1}. feature {features[indices[f]]}   \t{importances[indices[f]
    if 'random' in features[indices[f]]:
        feature += " <=="
    print(feature)
    feature_rank.append([features[indices[f]], importances[indices[f]]] )
```

```
Feature ranking:
1. feature INCQ298A     4.49%
2. feature NUM_CELLS_PARENTS    4.21%
3. feature INCPORAR     3.56%
4. feature CWIC_02      3.41%
5. feature CWIC_01      3.28%
6. feature INS_1        2.66%
7. feature EDUC1        2.41%
8. feature C1R          1.85%
9. feature C5R          1.79%
10. feature M_AGEGRP    1.73%
11. feature NUM_CELLS_HH        1.69%
12. feature BF_ENDR06           1.65%
13. feature INCPOV1     1.47%
14. feature PROVWT_D_TERR       1.18%
15. feature BF_FORMR08          1.17%
16. feature RDDWT_D     1.14%
17. feature RENT_OWN    1.11%
18. feature PROVWT_D    1.11%
19. feature RDDWT_D_TERR        1.09%
20. feature INS_2       1.00%
21. feature random      1.00% <==
22. feature SEQNUMC     0.99%
23. feature STRATUM     0.99%
24. feature SEQNUMHH    0.97%
25. feature ESTIAP14    0.95%
26. feature INS_3A      0.86%
27. feature EST_GRANT           0.86%
28. feature DFLU1       0.85%
29. feature RACE_K      0.85%
30. feature DVRC1       0.84%
31. feature DHEPB2      0.84%
32. feature DPCV4       0.83%
33. feature DDTP4       0.83%
34. feature DHIB3       0.81%
35. feature DHIB4       0.81%
36. feature DHEPB3      0.81%
37. feature RACEETHK    0.81%
38. feature DPOLIO3     0.80%
39. feature DDTP3       0.80%
40. feature DHEPA1      0.80%
41. feature STATE       0.79%
42. feature DMMR1       0.78%
43. feature DPCV2       0.77%
44. feature DDTP1       0.76%
45. feature DPCV3       0.75%
46. feature DROT2       0.72%
47. feature DDTP2       0.71%
48. feature DHEPA2      0.70%
49. feature DHIB1       0.70%
50. feature DPOLIO2     0.70%
51. feature DHIB2       0.70%
52. feature BF_EXCLR06          0.69%
53. feature DROT1       0.68%
54. feature DPCV1       0.68%
55. feature DFLU2       0.67%
56. feature DPOLIO1     0.67%
57. feature CBF_01      0.65%
58. feature FLU1_AGE    0.57%
```

```
59. feature DROT3       0.56%
60. feature DTP4_AGE    0.52%
61. feature CHILDNM     0.52%
62. feature DHEPB1      0.51%
63. feature FLU2_AGE    0.50%
64. feature INS_3       0.49%
65. feature PCV4_AGE    0.47%
66. feature HEA2_AGE    0.47%
67. feature DFLU3       0.47%
68. feature HIB4_AGE    0.46%
69. feature LANGUAGE    0.44%
70. feature FLU3_AGE    0.44%
71. feature HEA1_AGE    0.44%
72. feature PROV_FAC    0.42%
73. feature HEP3_AGE    0.41%
74. feature I_HISP_K    0.39%
75. feature P_NUMFLUN           0.39%
76. feature HIB3_AGE    0.38%
77. feature MMR1_AGE    0.38%
78. feature DHEPB4      0.38%
79. feature VRC1_AGE    0.37%
80. feature P_NUMHS     0.36%
81. feature CEN_REG     0.35%
82. feature FRSTBRN     0.35%
83. feature P_NUMFLU    0.35%
84. feature P_NUHIBX    0.34%
85. feature PCV3_AGE    0.32%
86. feature HEP2_AGE    0.31%
87. feature POL3_AGE    0.30%
88. feature DTP3_AGE    0.29%
89. feature P_NUMDTA    0.29%
90. feature AGEGRP      0.28%
91. feature REGISTRY    0.28%
92. feature VFC_ORDER           0.28%
93. feature P_NUHEPX    0.27%
94. feature INTRP       0.27%
95. feature P_NUMDIH    0.27%
96. feature HIB2_AGE    0.26%
97. feature HEP4_AGE    0.25%
98. feature P_NUMRM     0.25%
99. feature P_NUMHM     0.24%
100. feature P_NUMROT           0.24%
101. feature INS_11     0.24%
102. feature P_NUMDHI           0.23%
103. feature XPOLTY2    0.23%
104. feature ROT2_AGE           0.23%
105. feature DTP2_AGE           0.23%
106. feature INS_4_5    0.22%
107. feature D6R        0.22%
108. feature P_NUMHEA           0.22%
109. feature ROT3_AGE           0.21%
110. feature POL2_AGE           0.21%
111. feature POL1_AGE           0.21%
112. feature XPOLTY1    0.21%
113. feature ROT1_AGE           0.20%
114. feature HEP1_AGE           0.20%
115. feature P_NUMIPV           0.19%
116. feature DPOLIO4    0.19%
117. feature PCV1_AGE           0.19%
```

```
118. feature PCV2_AGE           0.19%
119. feature NUM_PHONE          0.19%
120. feature XDTPTY3    0.19%
121. feature XPOLTY3    0.19%
122. feature DFLU4      0.18%
123. feature XDTPTY2    0.18%
124. feature P_NUMPCC13         0.18%
125. feature SEX        0.17%
126. feature XDTPTY1    0.17%
127. feature HIB1_AGE           0.17%
128. feature P_NUMHEP           0.17%
129. feature P_NUMRG    0.16%
130. feature POL4_AGE           0.16%
131. feature FLU4_AGE           0.16%
132. feature DTP1_AGE           0.16%
133. feature XHEPTY3    0.16%
134. feature P_NUMHIB           0.15%
135. feature XHEPTY4    0.15%
136. feature N_PRVR     0.15%
137. feature U2D_HEP    0.15%
138. feature XDTPTY4    0.14%
139. feature P_NUMPOL           0.14%
140. feature XHEPTY2    0.14%
141. feature P_NUMPCC           0.14%
142. feature U1D_HEP    0.14%
143. feature HEP_BRTH           0.13%
144. feature P_UTDHEPA2         0.13%
145. feature INS_6      0.13%
146. feature P_NUMPCV           0.12%
147. feature U3D_HEP    0.12%
148. feature P_NUMHIN           0.11%
149. feature P_NUMMMRX          0.11%
150. feature MOBIL_I    0.11%
151. feature P_NUMDTP           0.10%
152. feature P_UTDROT_S         0.10%
153. feature XMMRTY1    0.10%
154. feature P_UTDHEPA1         0.10%
155. feature P_NUMVRX           0.09%
156. feature XHEPTY1    0.09%
157. feature P_NUMMRV           0.08%
158. feature XPCVTY2    0.08%
159. feature XPCVTY4    0.08%
160. feature XPCVTY1    0.08%
161. feature P_NUMFLUM          0.08%
162. feature P_NUMVRC           0.08%
163. feature P_UTD431H313_ROUT_S        0.07%
164. feature P_UTD431H3_ROUT_S          0.07%
165. feature PU431_31           0.07%
166. feature PU431331           0.07%
167. feature XPCVTY3    0.07%
168. feature P_NUMMMR           0.07%
169. feature PU4313313          0.07%
170. feature PU4313314          0.07%
171. feature P_UTDHIB_ROUT_S    0.07%
172. feature PUT43133           0.06%
173. feature P_UTD431H_ROUT_S           0.06%
174. feature P_UTD431H31_ROUT_S         0.06%
175. feature PUTD4313           0.06%
176. feature PU431_314          0.06%
```

```
177. feature P_NUMMMX           0.06%
178. feature P_NUMPCC7          0.06%
179. feature P_UTDPCV           0.06%
180. feature P_UTD431H314_ROUT_S        0.06%
181. feature P_UTDHEP           0.05%
182. feature P_NUMPCN           0.05%
183. feature P_UTDTP4           0.05%
184. feature P_UTDMCV           0.05%
185. feature P_UTD331           0.05%
186. feature XPOLTY4    0.05%
187. feature P_UTD431           0.05%
188. feature P_NUMRO    0.05%
189. feature VRC2_AGE           0.04%
190. feature P_NUMOLN           0.04%
191. feature P_NUMFLUL          0.04%
192. feature P_UTDMMX           0.04%
193. feature P_NUMTPN           0.04%
194. feature P_U12VRC           0.04%
195. feature DVRC2      0.04%
196. feature P_UTDHIB           0.04%
197. feature AGECPOXR           0.04%
198. feature P_UTDPC3           0.03%
199. feature DDTP5      0.03%
200. feature DMMR2      0.03%
201. feature P_UTDPOL           0.03%
202. feature HAD_CPOX           0.03%
203. feature P_NUMHION          0.03%
204. feature DPCV5      0.03%
205. feature DTP5_AGE           0.03%
206. feature P_UTDTP3           0.03%
207. feature XPCVTY5    0.02%
208. feature P_UTDHIB_SHORT_S           0.02%
209. feature XHIBTY3    0.02%
210. feature P_NUMHG    0.02%
211. feature MMR2_AGE           0.02%
212. feature P_NUMHHY           0.02%
213. feature P_UTDPCVB13        0.02%
214. feature PCV5_AGE           0.02%
215. feature HEP5_AGE           0.02%
216. feature HIB5_AGE           0.02%
217. feature DHEPB5     0.02%
218. feature XHEPTY5    0.02%
219. feature DFLU5      0.02%
220. feature P_NUMOPV           0.02%
221. feature DHIB5      0.02%
222. feature XHIBTY2    0.02%
223. feature P_NUMHEN           0.02%
224. feature FLU5_AGE           0.02%
225. feature XDTPTY5    0.02%
226. feature XMMRTY2    0.01%
227. feature XHIBTY4    0.01%
228. feature XHIBTY1    0.01%
229. feature HEA3_AGE           0.01%
230. feature P_NUMPCP           0.01%
231. feature P_NUHPHB           0.01%
232. feature DHEPA3     0.01%
233. feature D7         0.01%
234. feature POL5_AGE           0.01%
235. feature P_NUMDAH           0.01%
```

```
236. feature P_NUMMCN           0.01%
237. feature P_NUMMS    0.01%
238. feature P_NUMVRN           0.01%
239. feature DPOLIO5    0.01%
240. feature PCV6_AGE           0.00%
241. feature DPCV6      0.00%
242. feature HEP_FLAG           0.00%
243. feature DHEPB6     0.00%
244. feature XPOLTY5    0.00%
245. feature ROT4_AGE           0.00%
246. feature XHIBTY5    0.00%
247. feature DROT4      0.00%
248. feature DTP6_AGE           0.00%
249. feature HIB6_AGE           0.00%
250. feature HEP6_AGE           0.00%
251. feature DHEPA4     0.00%
252. feature DHIB6      0.00%
253. feature VRC3_AGE           0.00%
254. feature DVRC3      0.00%
255. feature XPCVTY6    0.00%
256. feature P_NUMPCCN          0.00%
257. feature MMR3_AGE           0.00%
258. feature FLU6_AGE           0.00%
259. feature P_NUMRB    0.00%
260. feature XVRCTY8    0.00%
261. feature P_NUMMSR           0.00%
262. feature DHEPA5     0.00%
263. feature DHEPA6     0.00%
264. feature XVRCTY4    0.00%
265. feature XROTTY7    0.00%
266. feature XVRCTY9    0.00%
267. feature XVRCTY5    0.00%
268. feature XVRCTY7    0.00%
269. feature DHEPA7     0.00%
270. feature DHEPA8     0.00%
271. feature DHEPA9     0.00%
272. feature XROTTY9    0.00%
273. feature XROTTY8    0.00%
274. feature BFENDFL06          0.00%
275. feature PDAT       0.00%
276. feature P_NUMMSM           0.00%
277. feature P_NUMMPR           0.00%
278. feature P_NUMMP    0.00%
279. feature BFFORMFL06         0.00%
280. feature XVRCTY3    0.00%
281. feature DDTP6      0.00%
282. feature DDTP7      0.00%
283. feature DDTP8      0.00%
284. feature DDTP9      0.00%
285. feature YEAR       0.00%
286. feature XVRCTY2    0.00%
287. feature XVRCTY1    0.00%
288. feature XROTTY5    0.00%
289. feature SHOTCARD           0.00%
290. feature DFLU6      0.00%
291. feature DFLU7      0.00%
292. feature DFLU8      0.00%
293. feature DFLU9      0.00%
294. feature XROTTY6    0.00%
```

```
295. feature XVRCTY6    0.00%
296. feature XROTTY4    0.00%
297. feature DHEPB7     0.00%
298. feature POL7_AGE           0.00%
299. feature POL9_AGE           0.00%
300. feature RB1_AGE    0.00%
301. feature RB2_AGE    0.00%
302. feature RB3_AGE    0.00%
303. feature RB4_AGE    0.00%
304. feature RB5_AGE    0.00%
305. feature RB6_AGE    0.00%
306. feature RB7_AGE    0.00%
307. feature RB8_AGE    0.00%
308. feature RB9_AGE    0.00%
309. feature ROT5_AGE           0.00%
310. feature ROT6_AGE           0.00%
311. feature ROT7_AGE           0.00%
312. feature ROT8_AGE           0.00%
313. feature ROT9_AGE           0.00%
314. feature VRC4_AGE           0.00%
315. feature VRC5_AGE           0.00%
316. feature POL8_AGE           0.00%
317. feature POL6_AGE           0.00%
318. feature MP2_AGE    0.00%
319. feature PCV9_AGE           0.00%
320. feature MP4_AGE    0.00%
321. feature MP5_AGE    0.00%
322. feature MP6_AGE    0.00%
323. feature MP7_AGE    0.00%
324. feature MP8_AGE    0.00%
325. feature MP9_AGE    0.00%
326. feature MPR1_AGE           0.00%
327. feature MPR2_AGE           0.00%
328. feature MPR3_AGE           0.00%
329. feature MPR4_AGE           0.00%
330. feature MPR5_AGE           0.00%
331. feature MPR6_AGE           0.00%
332. feature MPR7_AGE           0.00%
333. feature MPR8_AGE           0.00%
334. feature MPR9_AGE           0.00%
335. feature PCV7_AGE           0.00%
336. feature PCV8_AGE           0.00%
337. feature VRC6_AGE           0.00%
338. feature VRC7_AGE           0.00%
339. feature VRC8_AGE           0.00%
340. feature VRC9_AGE           0.00%
341. feature XHIBTY9    0.00%
342. feature XMMRTY3    0.00%
343. feature XMMRTY4    0.00%
344. feature XMMRTY5    0.00%
345. feature XMMRTY6    0.00%
346. feature XMMRTY7    0.00%
347. feature XMMRTY8    0.00%
348. feature XMMRTY9    0.00%
349. feature XPCVTY7    0.00%
350. feature XPCVTY8    0.00%
351. feature XPCVTY9    0.00%
352. feature XPOLTY6    0.00%
353. feature XPOLTY7    0.00%
```

```
354. feature XPOLTY8    0.00%
355. feature XPOLTY9    0.00%
356. feature XROTTY1    0.00%
357. feature XROTTY2    0.00%
358. feature XHIBTY8    0.00%
359. feature XHIBTY7    0.00%
360. feature XHIBTY6    0.00%
361. feature XFLUTY4    0.00%
362. feature XDTPTY6    0.00%
363. feature XDTPTY7    0.00%
364. feature XDTPTY8    0.00%
365. feature XDTPTY9    0.00%
366. feature XFLUTY1    0.00%
367. feature XFLUTY2    0.00%
368. feature XFLUTY3    0.00%
369. feature XFLUTY5    0.00%
370. feature XHEPTY9    0.00%
371. feature XFLUTY6    0.00%
372. feature XFLUTY7    0.00%
373. feature XFLUTY8    0.00%
374. feature XFLUTY9    0.00%
375. feature XHEPTY6    0.00%
376. feature XHEPTY7    0.00%
377. feature XHEPTY8    0.00%
378. feature MP3_AGE    0.00%
379. feature MP1_AGE    0.00%
380. feature DHEPB8     0.00%
381. feature DMP9       0.00%
382. feature DMPRB2     0.00%
383. feature DMPRB3     0.00%
384. feature DMPRB4     0.00%
385. feature DMPRB5     0.00%
386. feature DMPRB6     0.00%
387. feature DMPRB7     0.00%
388. feature DMPRB8     0.00%
389. feature DMPRB9     0.00%
390. feature DPCV7      0.00%
391. feature DPCV8      0.00%
392. feature DPCV9      0.00%
393. feature XROTTY3    0.00%
394. feature DPOLIO6    0.00%
395. feature DPOLIO7    0.00%
396. feature DPOLIO8    0.00%
397. feature DPOLIO9    0.00%
398. feature DRB1       0.00%
399. feature DMPRB1     0.00%
400. feature DMP8       0.00%
401. feature MMR9_AGE           0.00%
402. feature DMP7       0.00%
403. feature DHEPB9     0.00%
404. feature DHIB7      0.00%
405. feature DHIB8      0.00%
406. feature DHIB9      0.00%
407. feature DMMR3      0.00%
408. feature DMMR4      0.00%
409. feature DMMR5      0.00%
410. feature DMMR6      0.00%
411. feature DMMR7      0.00%
412. feature DMMR8      0.00%
```

```
413. feature DMMR9      0.00%
414. feature DMP1       0.00%
415. feature DMP2       0.00%
416. feature DMP3       0.00%
417. feature DMP4       0.00%
418. feature DMP5       0.00%
419. feature DMP6       0.00%
420. feature DRB2       0.00%
421. feature DRB4       0.00%
422. feature DRB5       0.00%
423. feature DRB6       0.00%
424. feature HEA4_AGE           0.00%
425. feature HEA5_AGE           0.00%
426. feature HEA6_AGE           0.00%
427. feature HEA7_AGE           0.00%
428. feature HEA8_AGE           0.00%
429. feature HEA9_AGE           0.00%
430. feature HEP7_AGE           0.00%
431. feature HEP8_AGE           0.00%
432. feature HEP9_AGE           0.00%
433. feature HIB7_AGE           0.00%
434. feature HIB8_AGE           0.00%
435. feature HIB9_AGE           0.00%
436. feature MMR4_AGE           0.00%
437. feature MMR5_AGE           0.00%
438. feature MMR6_AGE           0.00%
439. feature MMR7_AGE           0.00%
440. feature MMR8_AGE           0.00%
441. feature FLU9_AGE           0.00%
442. feature FLU8_AGE           0.00%
443. feature FLU7_AGE           0.00%
444. feature DROT9      0.00%
445. feature DRB7       0.00%
446. feature DRB8       0.00%
447. feature DRB9       0.00%
448. feature DROT5      0.00%
449. feature DROT6      0.00%
450. feature DROT7      0.00%
451. feature DROT8      0.00%
452. feature DVRC4      0.00%
453. feature DTP9_AGE           0.00%
454. feature DVRC5      0.00%
455. feature DVRC6      0.00%
456. feature DVRC7      0.00%
457. feature DVRC8      0.00%
458. feature DVRC9      0.00%
459. feature DTP7_AGE           0.00%
460. feature DTP8_AGE           0.00%
461. feature DRB3       0.00%
```

There are now 17 features that are better at predicting maritial status than a randomw number. Unfortunately some of them are things like, "Number of cell phones in the house", and "Number of cell phones parents have".

In [82]:
```
top_ranks = feature_rank[:17]
top_ranks
```

Out[82]:
```
[['INCQ298A', 0.04493876505087042],
 ['NUM_CELLS_PARENTS', 0.04205578681194119],
 ['INCPORAR', 0.035614140113308386],
 ['CWIC_02', 0.034059552087169584],
 ['CWIC_01', 0.03281737474751377],
 ['INS_1', 0.026599281255561942],
 ['EDUC1', 0.024139091182115505],
 ['C1R', 0.018546259039010072],
 ['C5R', 0.017921462918803675],
 ['M_AGEGRP', 0.01726100006084509],
 ['NUM_CELLS_HH', 0.01690526401982556],
 ['BF_ENDR06', 0.016486299096405605],
 ['INCPOV1', 0.014730609307548475],
 ['PROVWT_D_TERR', 0.01182666975841087],
 ['BF_FORMR08', 0.011680061895478211],
 ['RDDWT_D', 0.01141982463766642],
 ['RENT_OWN', 0.011076082424101616]]
```

In [83]:
```
top_rank_cols = [s[0].split(',')[0] for s in top_ranks]
top_rank_cols.append('MARITAL2')
top_rank_cols
```

Out[83]:
```
['INCQ298A',
 'NUM_CELLS_PARENTS',
 'INCPORAR',
 'CWIC_02',
 'CWIC_01',
 'INS_1',
 'EDUC1',
 'C1R',
 'C5R',
 'M_AGEGRP',
 'NUM_CELLS_HH',
 'BF_ENDR06',
 'INCPOV1',
 'PROVWT_D_TERR',
 'BF_FORMR08',
 'RDDWT_D',
 'RENT_OWN',
 'MARITAL2']
```

In [84]:
```
top_rank_df = df_marital[top_rank_cols].copy()
```

In [85]:
```
top_rank_df.head(10)
```

Out[85]:

	INCQ298A	NUM_CELLS_PARENTS	INCPORAR	CWIC_02	CWIC_01	INS_1	EDUC1	C1R	C5R
1	4	2.0	0.500000	0.0	2	2.0	2	6	1
3	14	2.0	3.000000	0.0	2	1.0	4	4	1
4	3	3.0	0.500000	1.0	1	2.0	2	8	3
5	3	3.0	0.500000	1.0	1	2.0	2	8	3
6	5	1.0	1.089867	1.0	1	2.0	3	2	3
7	14	2.0	3.000000	0.0	2	1.0	4	3	2
8	14	2.0	3.000000	0.0	2	1.0	4	3	1
11	10	2.0	1.438797	0.0	2	2.0	3	5	4
12	9	1.0	1.481544	1.0	1	2.0	3	4	1
13	5	1.0	0.639352	1.0	1	2.0	2	3	1

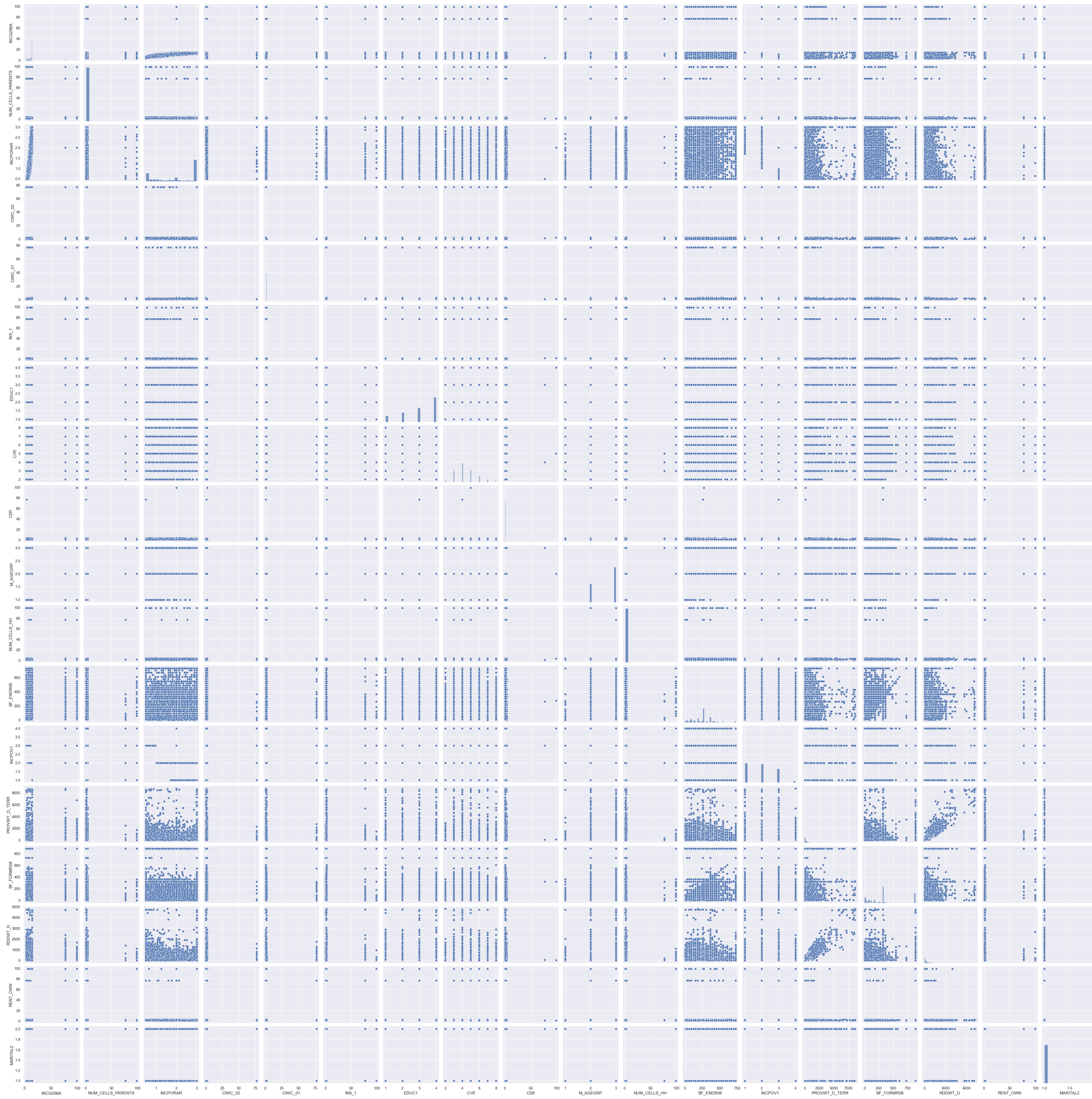
Marital Pair Plot

In [86]:

```
sns.pairplot(data=top_rank_df)
```

Out[86]:

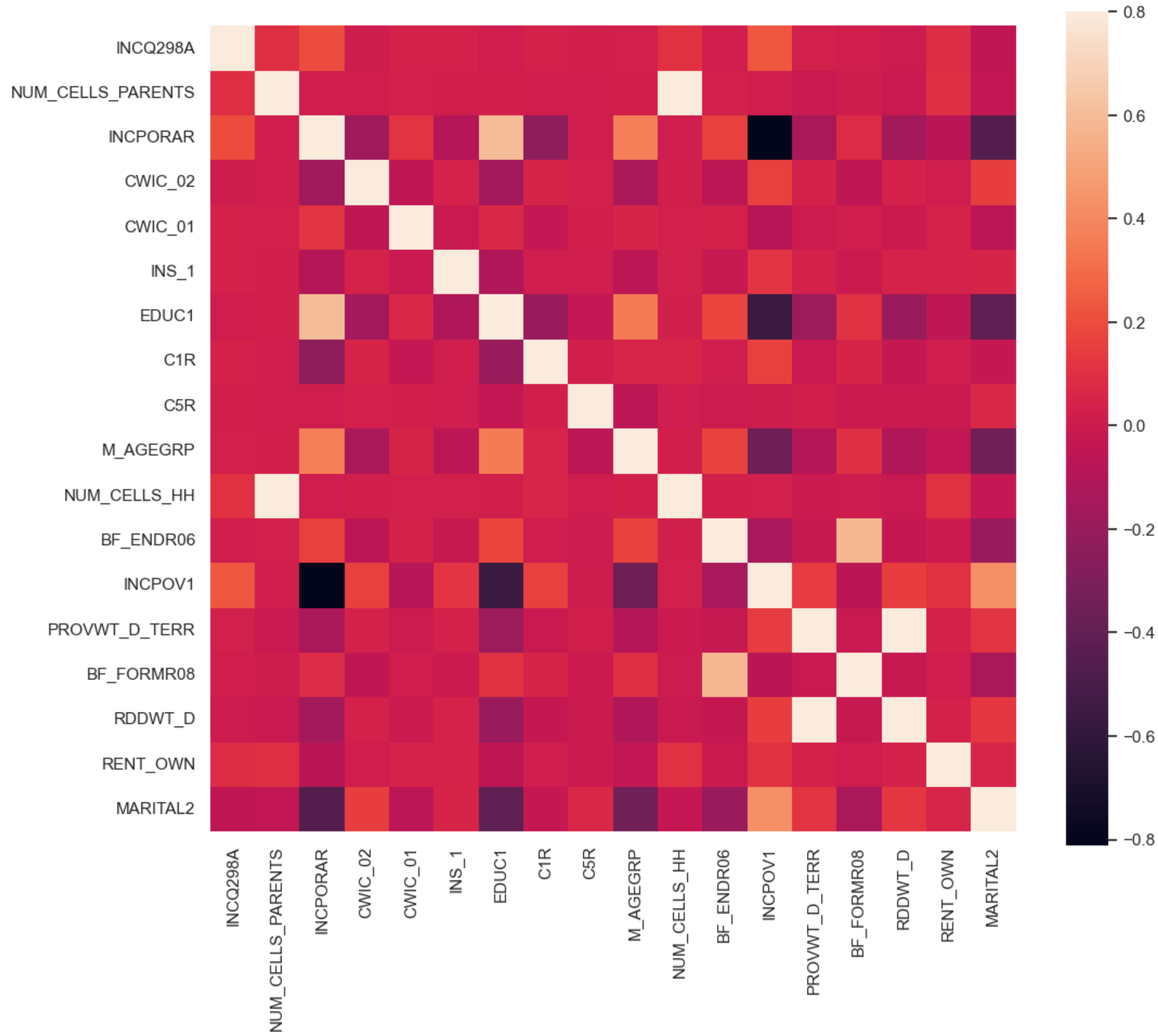
```
<seaborn.axisgrid.PairGrid at 0x1bb0d8ba1a0>
```



Marital Correlation Plot

In [87]:
```
corrmat = top_rank_df.corr()
f, ax = plt.subplots(figsize=(12,10)) #setting some parameters of the plot to help
sns.heatmap(corrmat, vmax = .8, square=True)
```

Out[87]: <Axes: >



It looks like the, wic benefits and poverty status, are positively correlated and the, income to poverty ratio, marital age group, and education, are negatively correlated with marital status.

In [89]:
```
marital = top_rank_df
marital.head(10)
```

Out[89]:

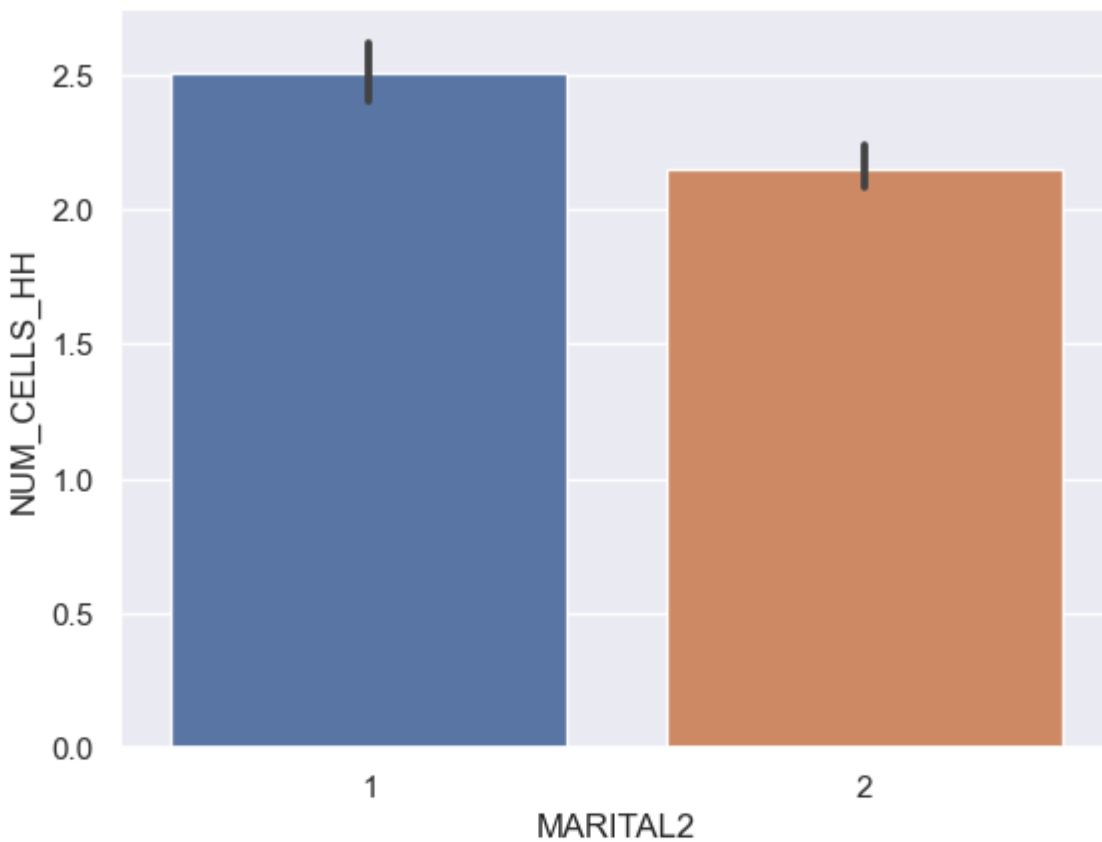
	INCQ298A	NUM_CELLS_PARENTS	INCPORAR	CWIC_02	CWIC_01	INS_1	EDUC1	C1R	C5R
1	4	2.0	0.500000	0.0	2	2.0	2	6	1
3	14	2.0	3.000000	0.0	2	1.0	4	4	1
4	3	3.0	0.500000	1.0	1	2.0	2	8	3
5	3	3.0	0.500000	1.0	1	2.0	2	8	3
6	5	1.0	1.089867	1.0	1	2.0	3	2	3
7	14	2.0	3.000000	0.0	2	1.0	4	3	2
8	14	2.0	3.000000	0.0	2	1.0	4	3	1
11	10	2.0	1.438797	0.0	2	2.0	3	5	4
12	9	1.0	1.481544	1.0	1	2.0	3	4	1
13	5	1.0	0.639352	1.0	1	2.0	2	3	1

In [93]:

```
sns.barplot(x="MARITAL2", y="NUM_CELLS_HH", data=marital)
```

Out[93]:

```
<Axes: xlabel='MARITAL2', ylabel='NUM_CELLS_HH'>
```



A look at number of cell phones compared to marital status. 1 is the married group, 2 is the not-married group. Bar in the confidence interval.

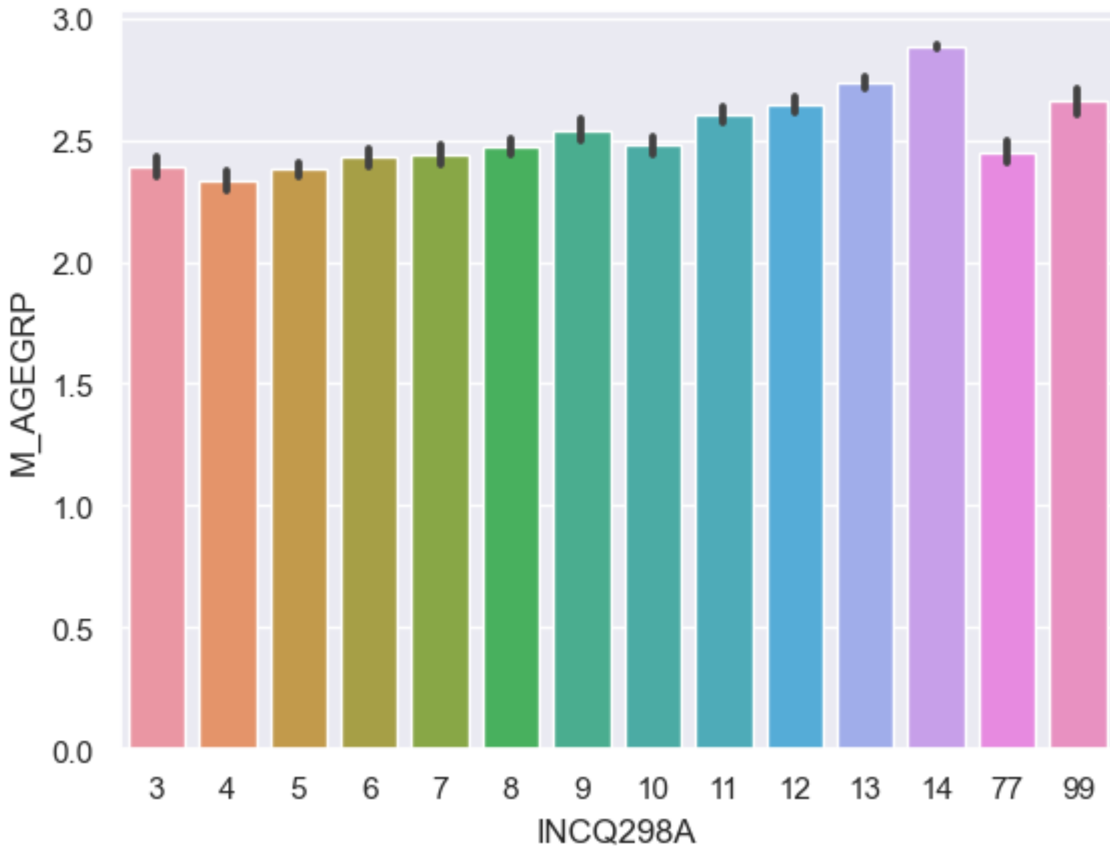
A look at income catagory compared to the years you have been married. This shows that the longer you have been married the higher your income is group.

In [115…

```
sns.barplot(x='INCQ298A', y='M_AGEGRP', data=marital)
```

Out[115]:

```
<Axes: xlabel='INCQ298A', ylabel='M_AGEGRP'>
```



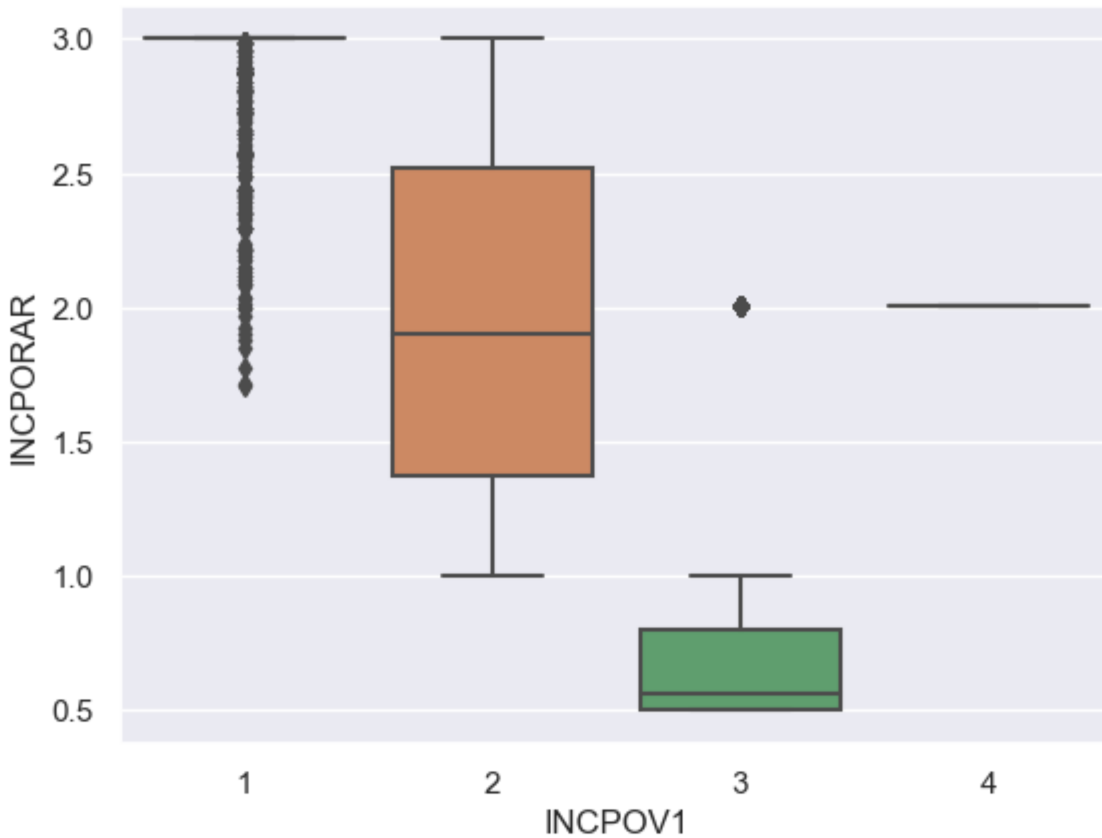
A box plot of income to poverty ratio compated to Poverty status. Those in group 3 are Below the poverty line and have a very low income to poverty ratio.

In [121…

```
sns.boxplot(x='INCPOV1', y='INCPORAR', data = marital)
```

Out[121]:

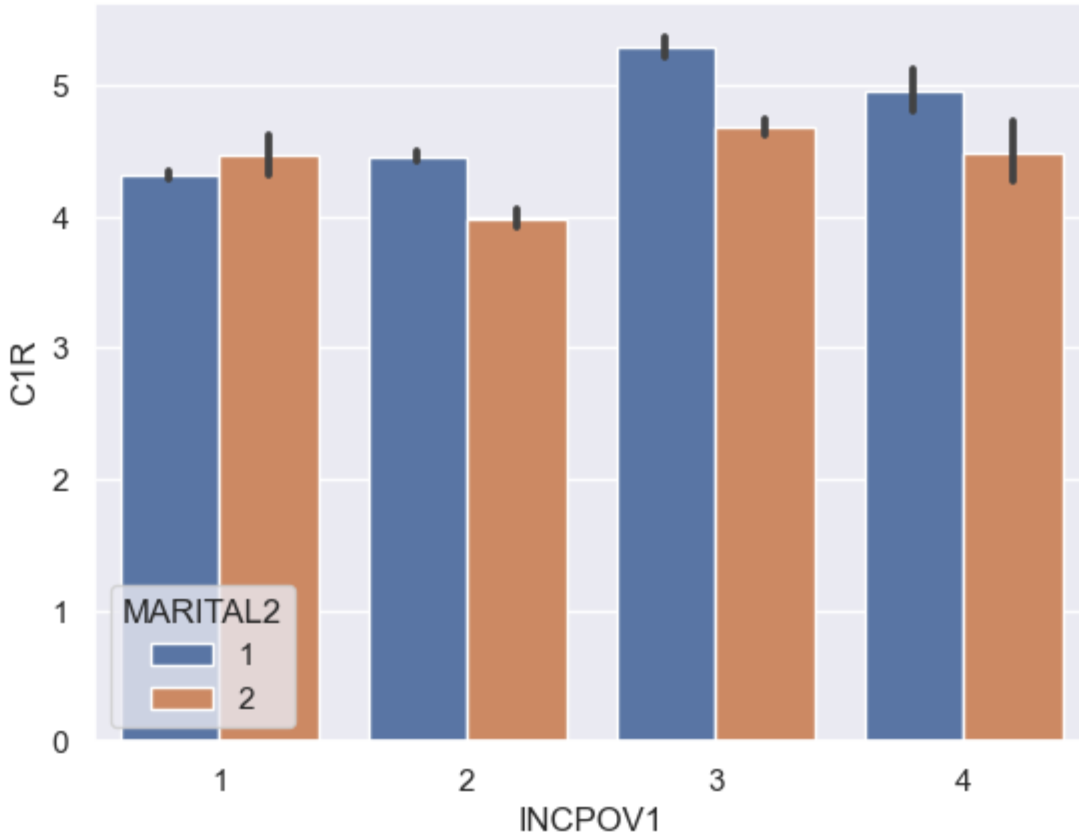
```
<Axes: xlabel='INCPOV1', ylabel='INCPORAR'>
```



A look at the number of people in the household compared to poverty and marital status. Those who are Below Poverty Level (3) and married have the most people living in their houses. Those that are unmarried and above the poverty level, but below $75k a year, have the least amount of people living in their houses.

In [123…
```
sns.barplot(x='INCPOV1',y='C1R', hue='MARITAL2', data=marital)
```

Out[123]:
```
<Axes: xlabel='INCPOV1', ylabel='C1R'>
```



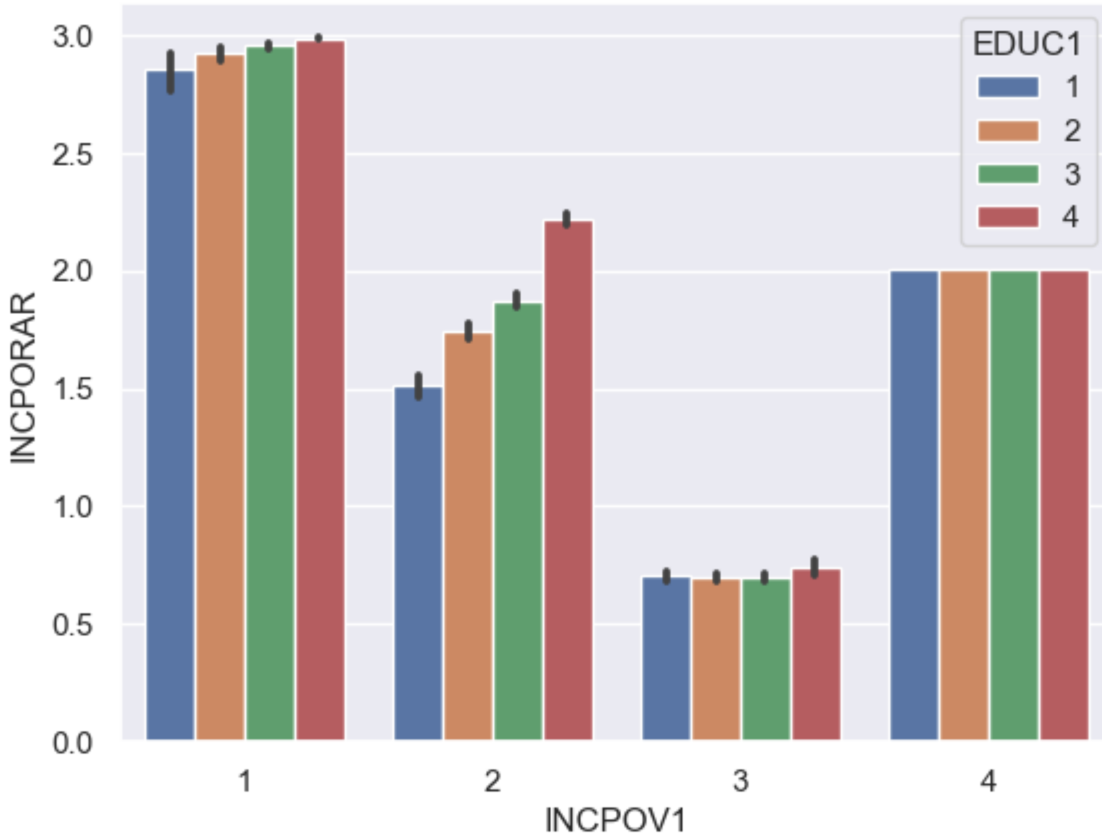
A look at education and poverty rate. Those with a 4 (College Grad) score the highest in each of the Income to poverty caragories.

In [127…

```
sns.barplot(x='INCPOV1',y='INCPORAR', hue='EDUC1', data=marital)
```

Out[127]:

```
<Axes: xlabel='INCPOV1', ylabel='INCPORAR'>
```



Dropping Varriables: Based on the sns plot and correlation martix several of the variables are going to be trimmed out due to having a low correlation. INCQ298A, the cell phone features, wic status, number of people in household C1R, and rent to own will all be dropped.

In [131…

```
small_marital = marital.drop(['INCQ298A', 'NUM_CELLS_PARENTS', 'CWIC_01', 'C1R', 'N
```

In [132…

```
small_marital.head(5)
```

Out[132]:

	INCPORAR	CWIC_02	INS_1	EDUC1	C5R	M_AGEGRP	BF_ENDR06	INCPOV1	PROVWT_D_TERI
1	0.500000	0.0	2.0	2	1	2	91.312500	3	806.84601
3	3.000000	0.0	1.0	4	1	3	334.812500	1	63.44868
4	0.500000	1.0	2.0	2	3	2	258.079564	3	94.87263
5	0.500000	1.0	2.0	2	3	2	258.079564	3	152.27384
6	1.089867	1.0	2.0	3	3	2	121.750000	2	210.18635

In [133…

```
y = marital.MARITAL2
x = marital.drop(['MARITAL2'], axis=1)
```

In [134…

```
# Creating a 70/30 train test split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)

clf=RandomForestClassifier(n_estimators=100)
clf.fit(x_train,y_train)
```

Out[134]:
```
▾ RandomForestClassifier
RandomForestClassifier()
```

In [135…
```
y_pred=clf.predict(x_test)
```

In [136…
```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8528109783089862
```

That slightly improved the model, by about 1%.

Discussion / Conclusion

The biggest challenge in working with this dataset is the number of unknown values that it contained. I decided to start off by pairing down the data to just the rows that the paper said contained "Adequate Data". In retrospect this may not have been a good approach. When I started working I wanted to look at Chicken Pox and if it was affected by immunizations. I would have needed a data set that contained all the immunization information to have done so.

After running the random forest classifier with "Had Chicken Pox" as the target it was found that the data was unable to provide a good model, and that the strongest feature was at what age a child had had chicken pox.

I then repeated the random forest classifier with Owning or Renting a house, again the accuracy of the model was too low.

Finally I performed the random forest classifier with Marital Status as the target. This produced a model with 84.5% accuract using the 17 features that performed better than a random number. I paried this feature set down to 11 features which improved the models accuracy to 85.3%.

Visualization of model features is shown and includes the relationship between marital status and income, number of people living in the home, education, and poverty ratio.