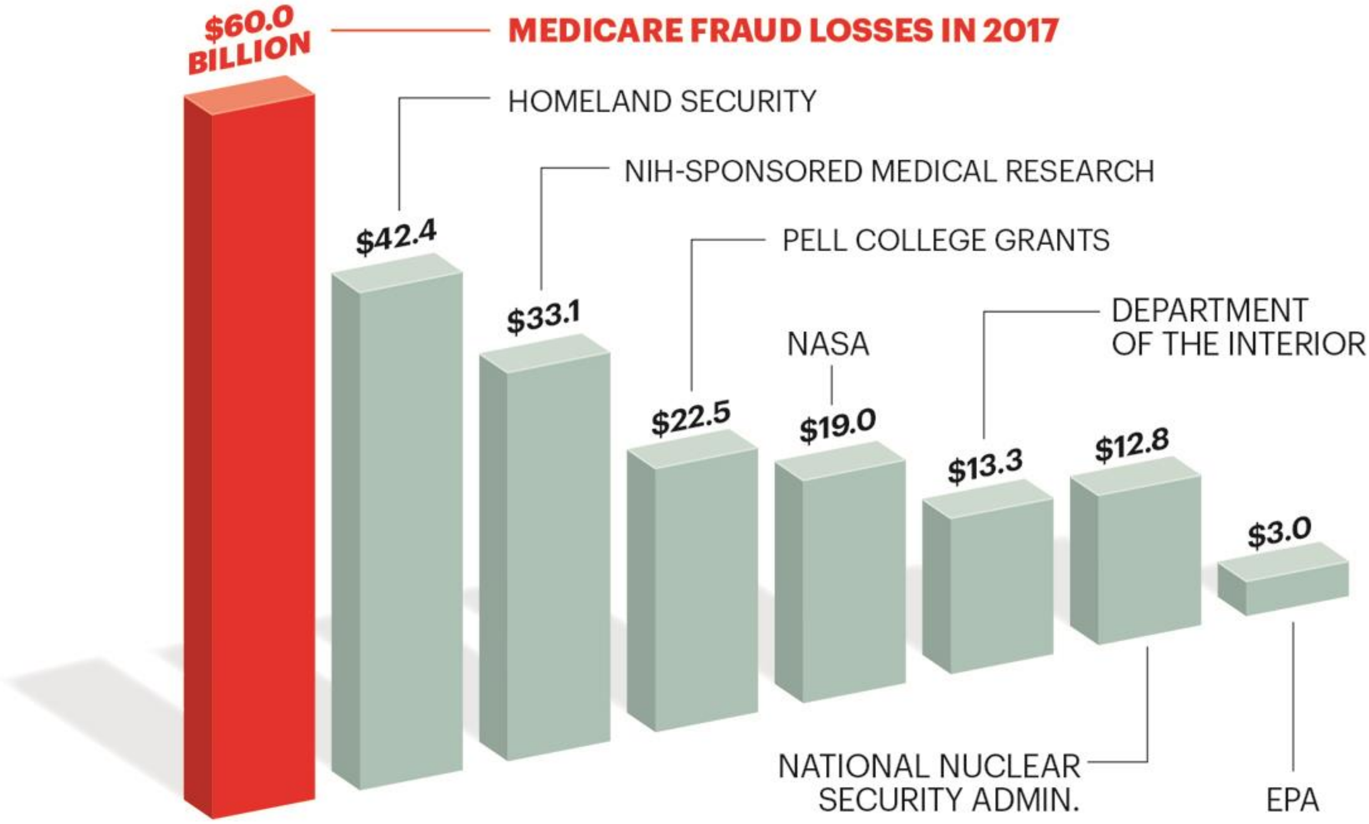
Exploratory analysis of US Medicare Fraud

Actl3142 Assignment
Z5194905



[1]

[1] https://www.aarp.org/money/scams-fraud/info-2018/medicare-scams-fraud-identity-theft.html

Executive Summary:

This report is an exploratory analysis of two Medicare datasets: one complete, and one aggregated per provider. It investigates how fraudulent claims and providers differ to their honest counterparts and makes modelling suggestions to predict future bad providers.

Individual Fraud:

In the complete data, there was little differentiation between the fraudulent and non-fraudulent providers, however, some variables were better measures of bad claims. The best indicator of individual fraud were the location specific variables; county and state (See Figures 1.1-1.4). In authentic claims, patients are relatively evenly distributed across counties and states, whereas in dishonest claims, patients are much more likely to come from certain areas, especially from state 31 and county codes 590, 160 and 130.

Fraudulent claims also share the following trends –

Table 1 *(figures in appendix)*

- Outpatient annual deductible amount is less variable for fraudulent claims (fig A1.1 – A1.2)
- Some physicians were much more common amongst dishonest providers (figure 1.5)
- There is more chronic diagnosis per patient on fraudulent claims (figure A1.3 – A1.4)
- Fraudulent claims last longer (figure A1.5 – A1.6)
- Higher amounts of claims in some months (figure A1.7 – A1.8)

Figure 1.1

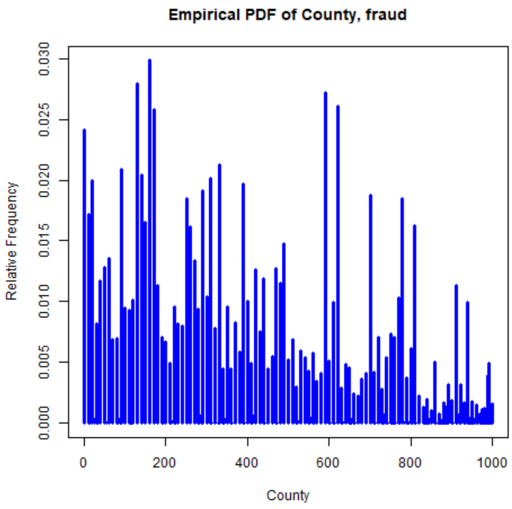


Figure 1.2

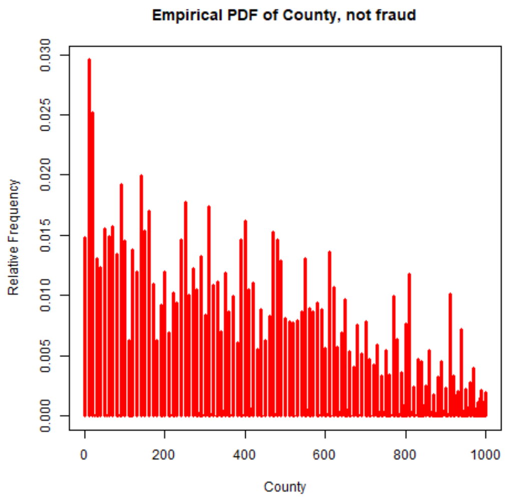


Figure 1.4

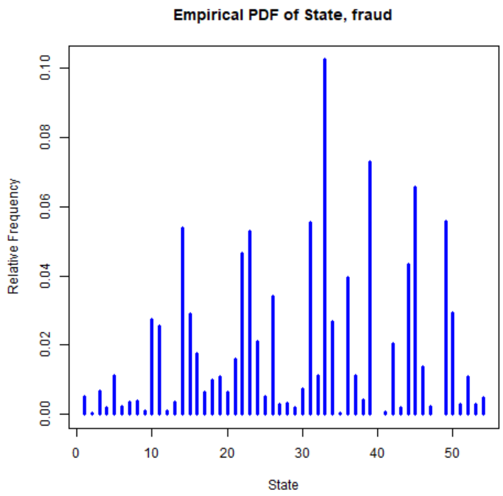


Figure 1.4

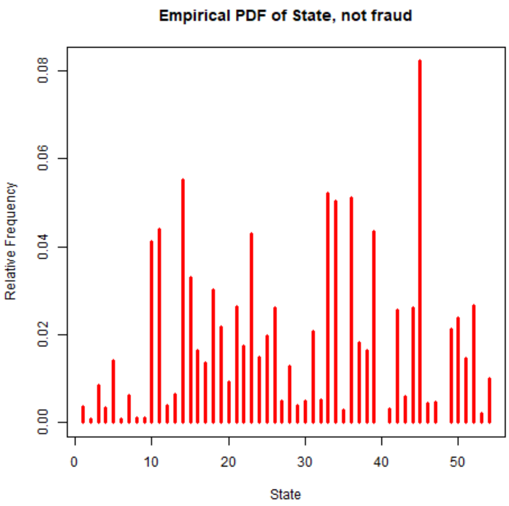
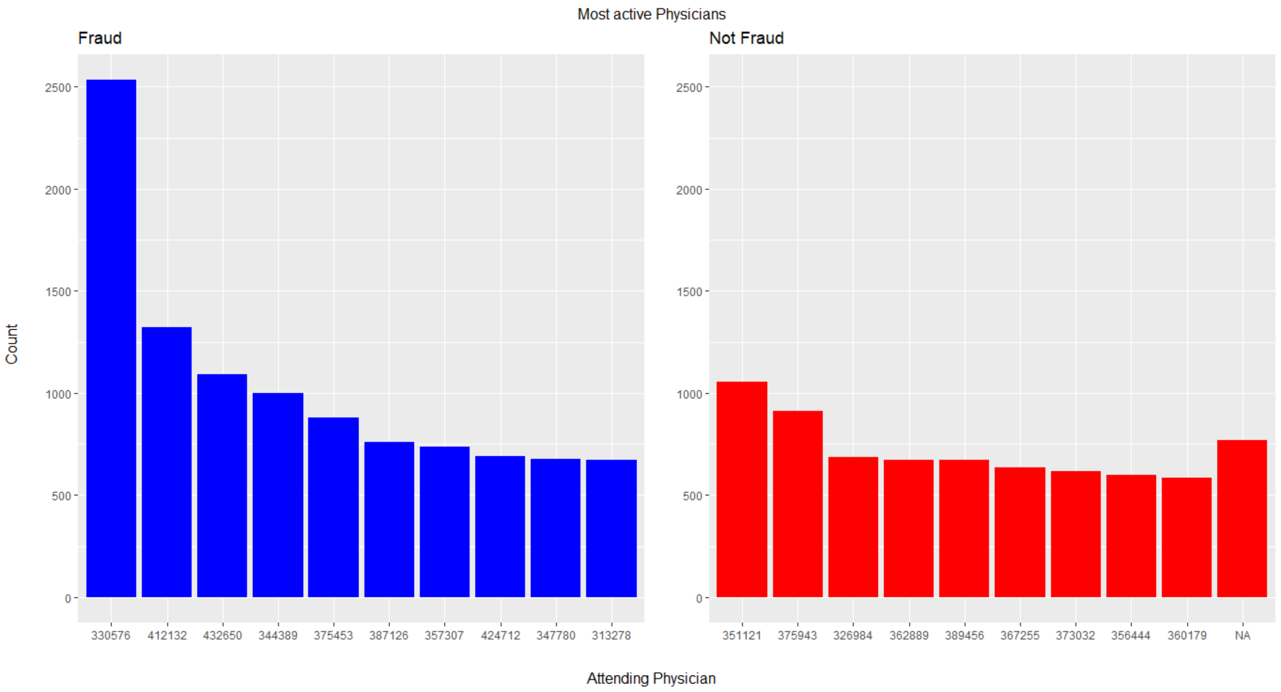


Figure 1.5



Aggregating by Provider:

Age: Average age for both providers is similar, however variance for non-fraudulent providers is much higher

Figure 2.3

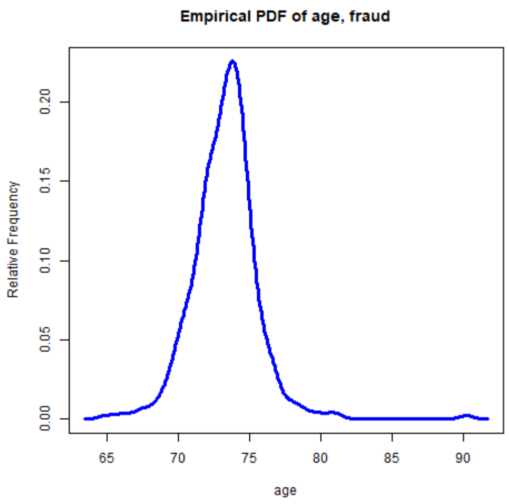
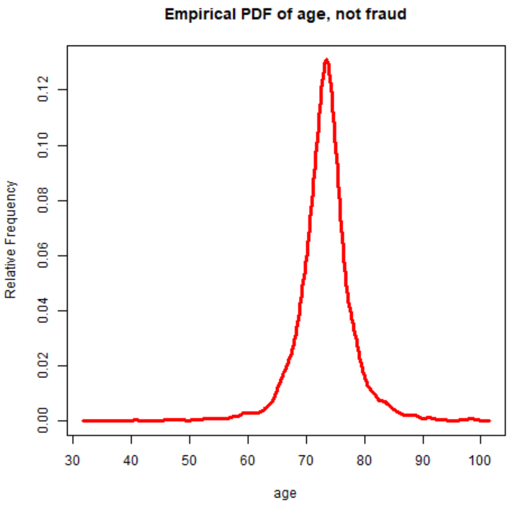


Figure 2.4



Averaged Binary Variables:

The following variables take one of two values in the complete dataset. The average of these can, at first, appear to produce quite interesting plots where fraudulent and not-fraudulent providers are distinct from one another. However, the variability of the average largely depends on the number of claims per provider. This issue will be considered at the end of this segment.

Number of Claims: Providers with more claims are generally fraudulent as most with at least 1000 claims are bad. However, this increase in fraudulence may not indicate a systematic increase in bad practice by the provider, but instead higher probability of a single instance of “opportunistic” fraud, for example, a dishonest physician. This single illegitimate claim could then label the whole provider as fraudulent.

Figure 2.1

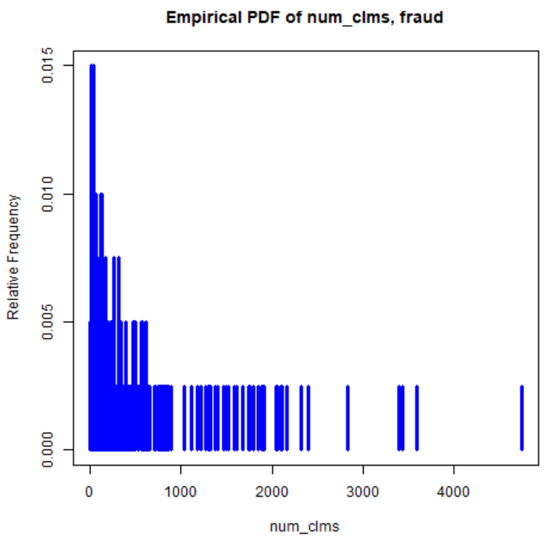
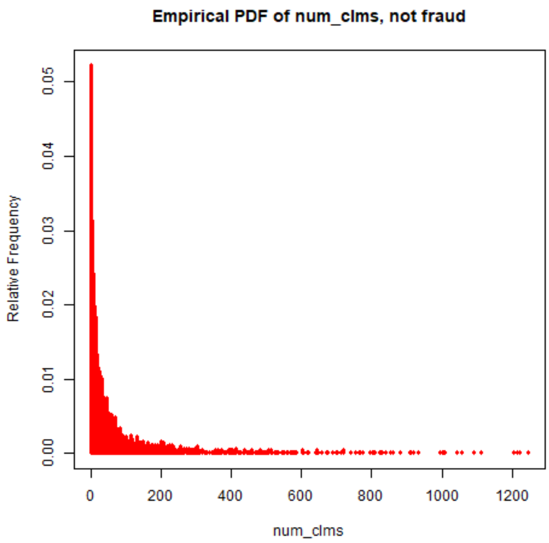


Figure 2.2



Chronic Conditions: The average diagnosis of a condition per patient is relatively consistent among fraudulent providers. For non-fraudulent providers, the curve is multimodal with modes at 1 and 2, suggesting that providers who either; specialize in treating the disease, or are unable to provide treatment for it, are less likely to be fraudulent. Similarly, fraudulent providers number of diagnosed chronic conditions were less variable (figure 2.5 – 2.6 and A2.1 – A2.4).

Gender: Like chronic conditions, providers that are gender specific are less fraudulent (figure 2.7 – 2.8).

Inpatient: Again, providers who specialize in insuring either; those admitted to hospital, or those who are not, are much less fraudulent (figure A2.5 – A2.6). Bad providers have claims which are much more uniform between inpatients and outpatients shown by figure 2.9 below-

Figure 2.5

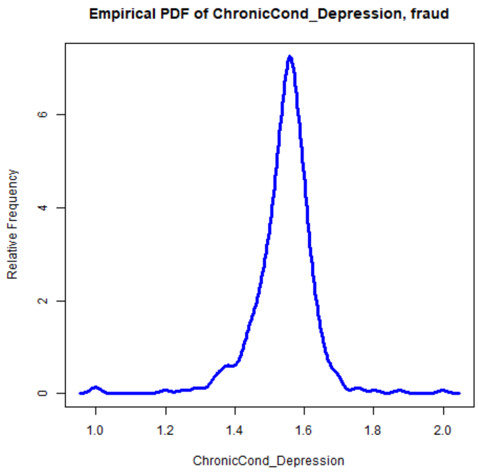


Figure 2.6

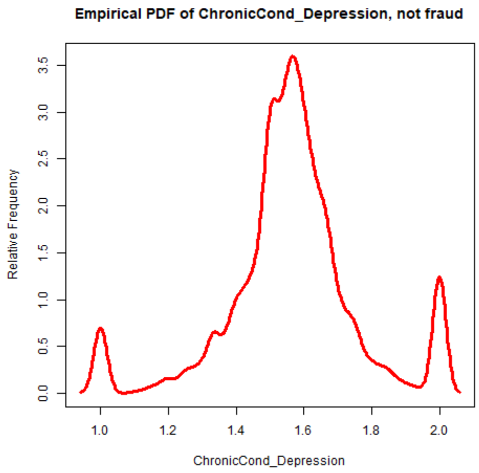


Figure 2.7

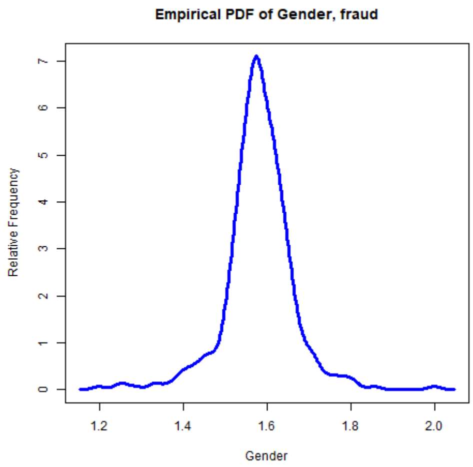
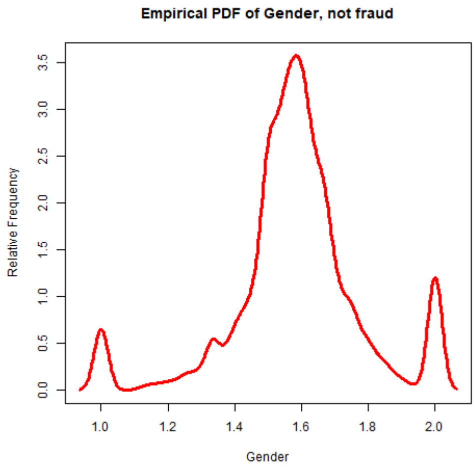


Figure 2.8



Deductible Amount Paid: For non-fraudulent providers, the deductible is either paid in all claim instances or none of them, with an average of either 0 or 1168. However, amongst fraudulent providers it is a continuous range between these values (figure 2.10 – figure 2.11), indicating again that fraudulent providers have less homogenous beneficiaries.

Figure 2.9

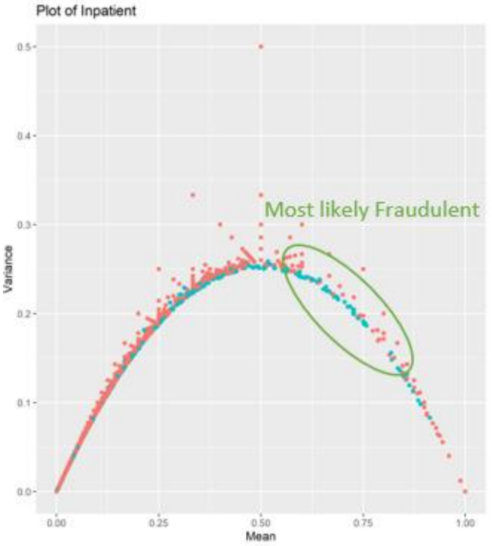


Figure 2.10

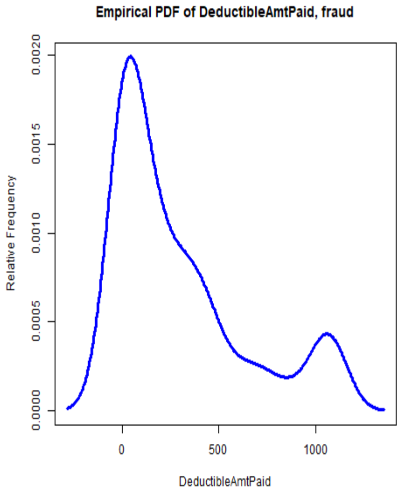
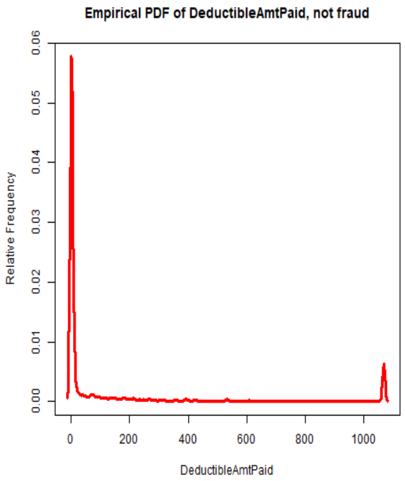


Figure 2.11



Averaged Binary Variables: The issue

Providers which seem to have homogenous beneficiaries (all males/ all Alzheimer's, etc), tend to be less fraudulent, however, it is unclear whether this is solely due to the fact that providers with only a few claims are much less likely to be fraudulent and will also appear to be more homogenous. From figure 2.13 the latter appears likely as only providers with claims close to 0 are homogenous. However, on further inspection from figure 2.14, there may be some "specialization effect" as the chances of a single provider having 10 claims of depression is minimal at 0.00027. In the data, 80 providers have 10 claims and so the probability that there is a provider with 10 claims where all are diagnosed with depression is 0.0215, thus suggesting the provider in figure 2.12 with the single claim is specialized. Further analysis should be conducted on this effect to isolate the contribution of the number of claims to the variability of the average of these binary variables.

Figure 2.12

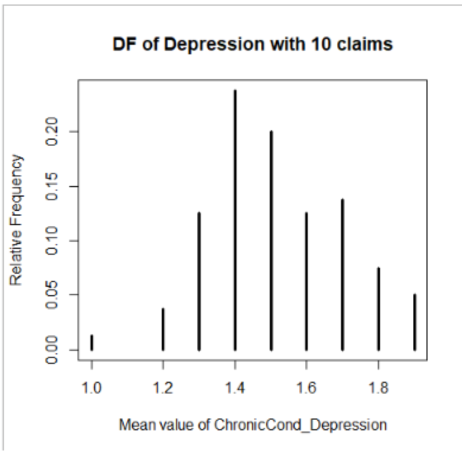


Figure 2.13

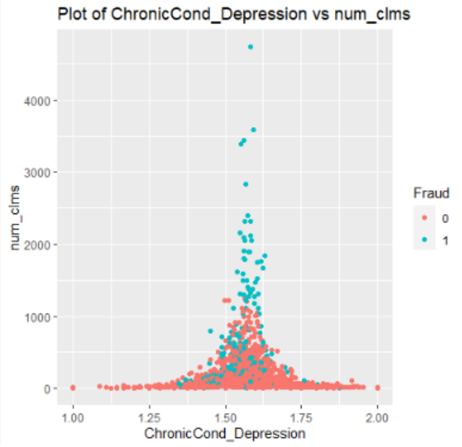
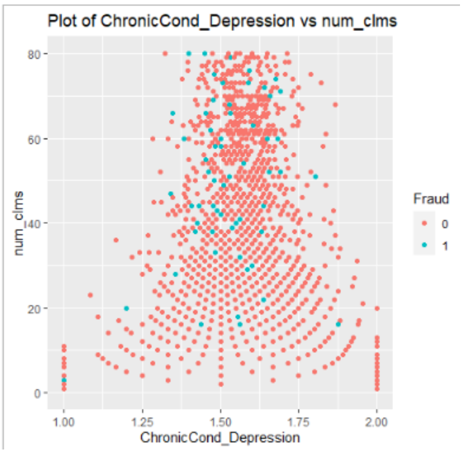


Figure 2.14



Interactions between variables

We can observe the interactions between two variables and classify areas that could produce interesting results through KNN. By plotting the number of claims against both inpatient and the claim amount reimbursed, we can see the fraud rates have much fatter tails. Therefore, claims in this area can be an indicator of illegitimacy (figure 3.1 – 3.2). Figure 3.3 shows a lack of variability in forged claims by the dense datapoints in the center. Again, we see how specialization is a good indicator of a trustworthy provider, as the red lines represent those who treat a fixed number of Chronic conditions.

Figure 3.1

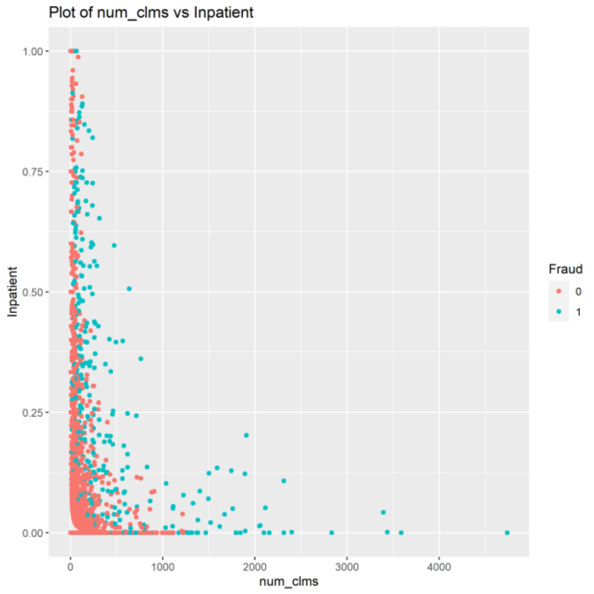
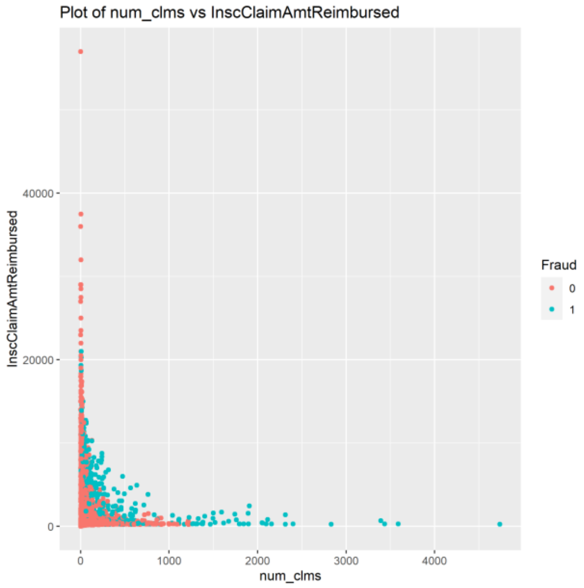


Figure 3.2



In figure 3.4, there is a linear relationship between the claim amount reimbursed and the number of chronic conditions diagnosed by a bad provider, whereas for good providers, there is no relationship. This suggests that dishonest providers falsely diagnose a larger amount of chronic conditions for the sole purpose of increasing profits. Unlike other trends, this does not rely on averaged binary variables and is therefore a good measure of fraud for all provider sizes.

Figure 3.3

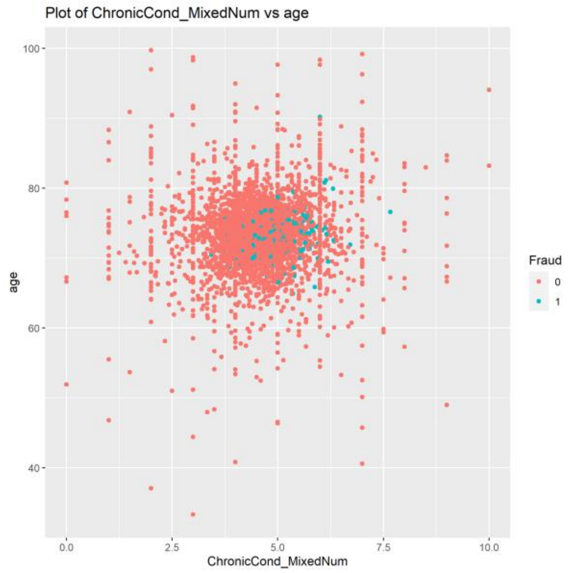
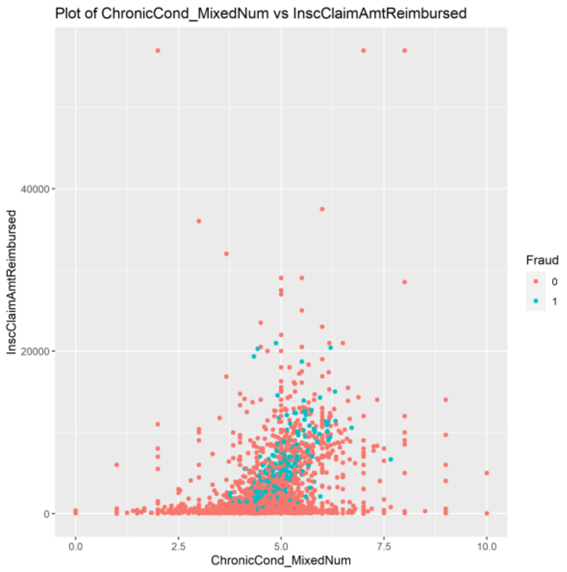


Figure 3.4



Recommendation for further analysis

This exploratory analysis has shown which variables from the data can indicate fraud, and so, modelling is required to predict future bad providers. This report recommends building a KNN model based on the interactions above as well as a generalized linear model of provider fraud based on age, chronic conditions, gender, admission to hospital, deductible amount paid and number of claims. A model of individual fraud should also be made based on location, and the variables in table 1. Table 2 in the appendix shows the high correlation between chronic conditions and these should be considered when constructing linear models.

Conclusion

The data shows that many of the variables can indicate fraudulent providers, and so, further modelling should be done to predict whether a provider has illegitimate claims.

Appendix:

Assumptions:

- The count of 54 states was not an error
- Data retrieval did not include day of birth since all individuals are born on the first day of the month
- Negative OPAnnualReimbursementAmt and IPAnnualReimbursementAmt values were not removed and treated as normal since their fraudulent rates did not vary greatly from the total average.
- A single known fraudulent claim labels the whole provider as fraudulent

Complete data plots-

Figure A1.1

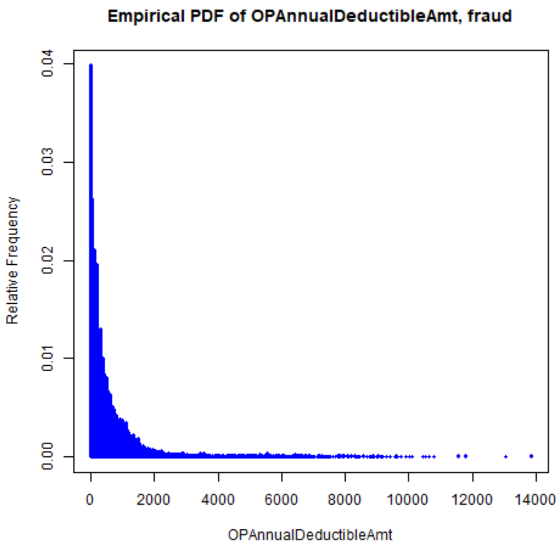


Figure A1.2

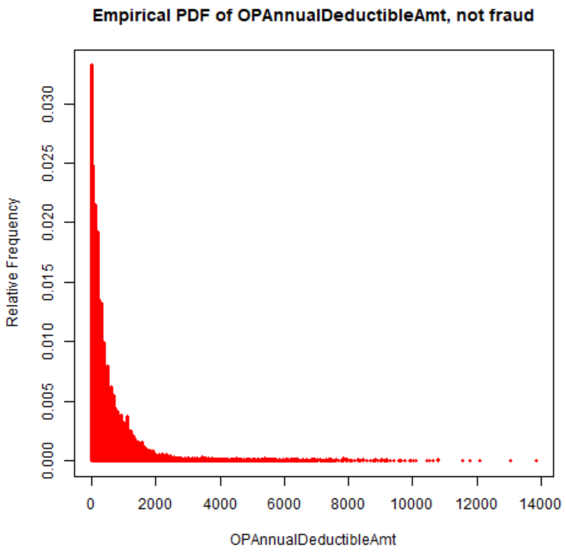


Figure A1.3

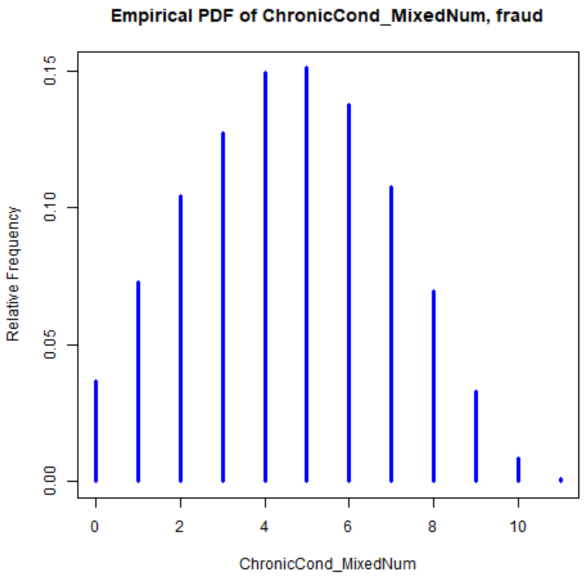


Figure A1.4

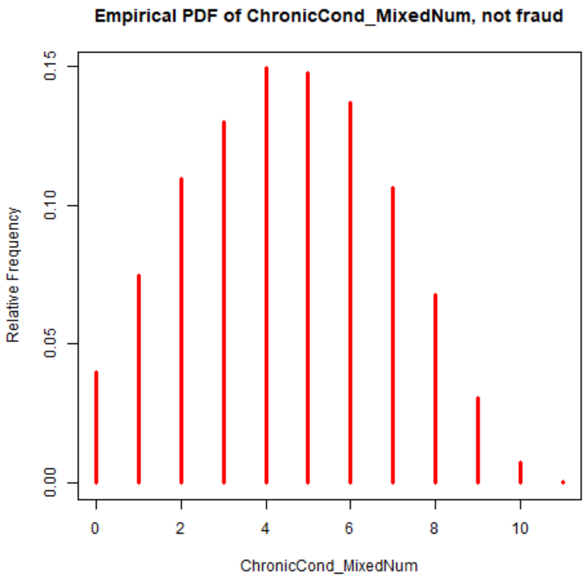


Figure A1.5

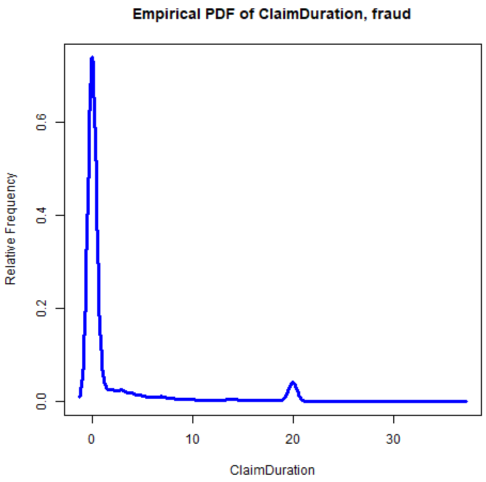


Figure A1.6

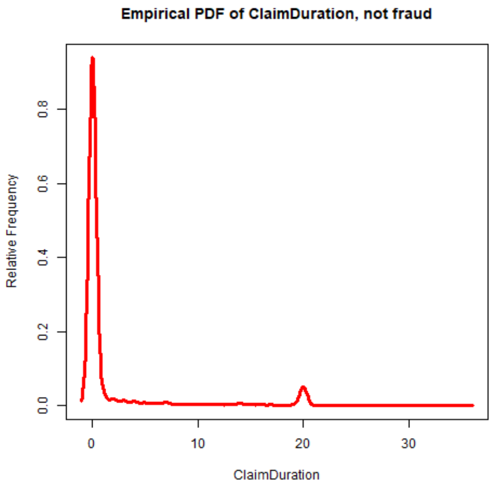


Figure A1.7: *Proportion of fraud per month*

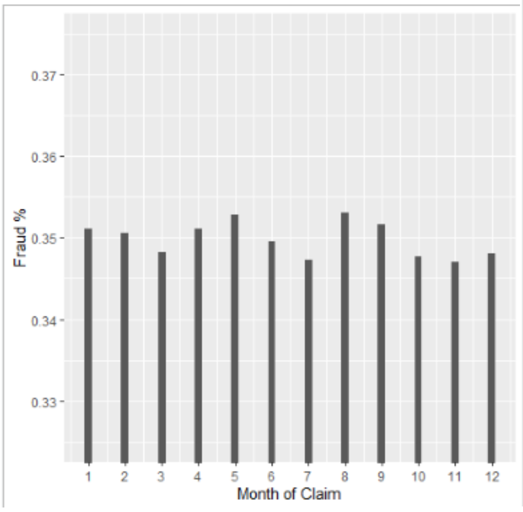


Figure A1.8: *Deductibles per patient each month*

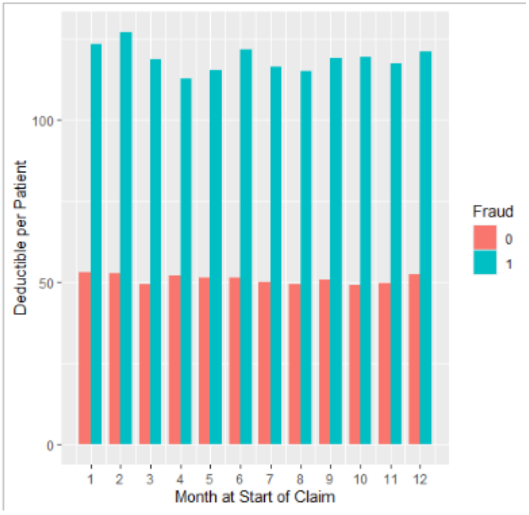
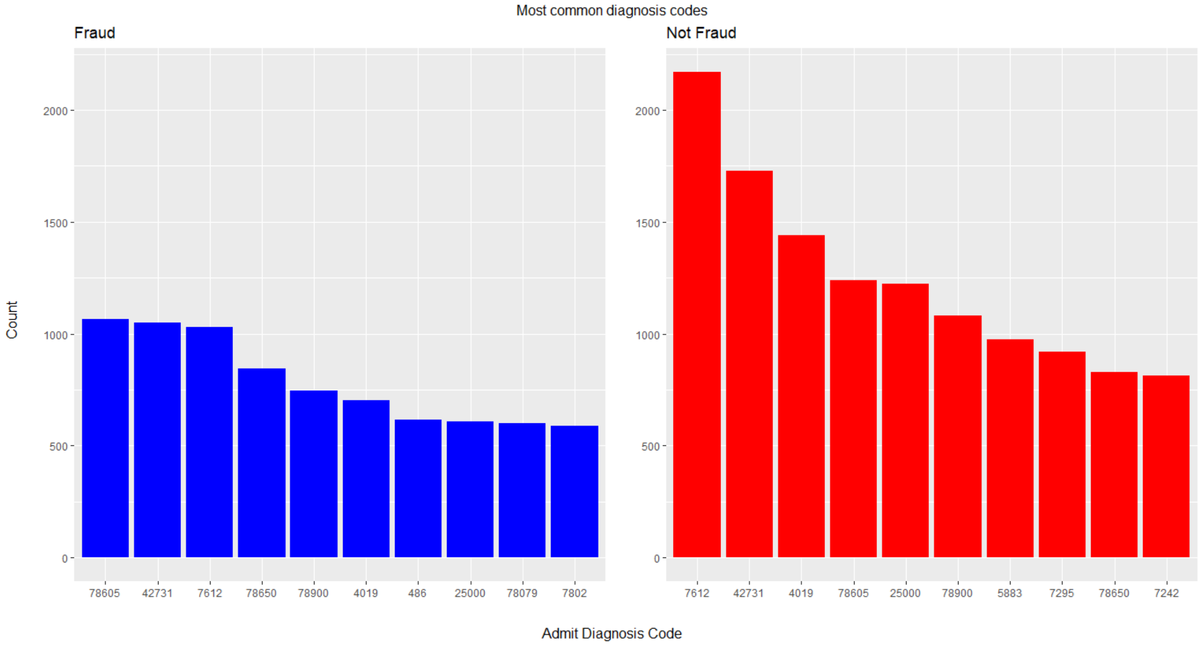


Figure A1.9



Data aggregated by provider plots

Figure A2.1

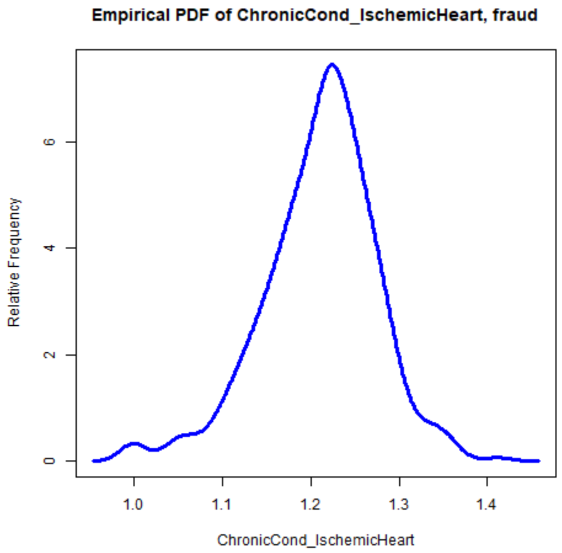

Empirical PDF of ChronicCond_IschemicHeart, fraud

Figure A2.2

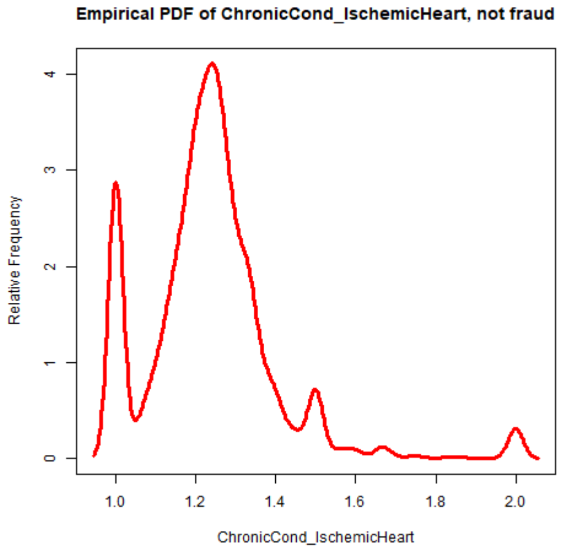

Empirical PDF of ChronicCond_IschemicHeart, not fraud

Figure A2.3

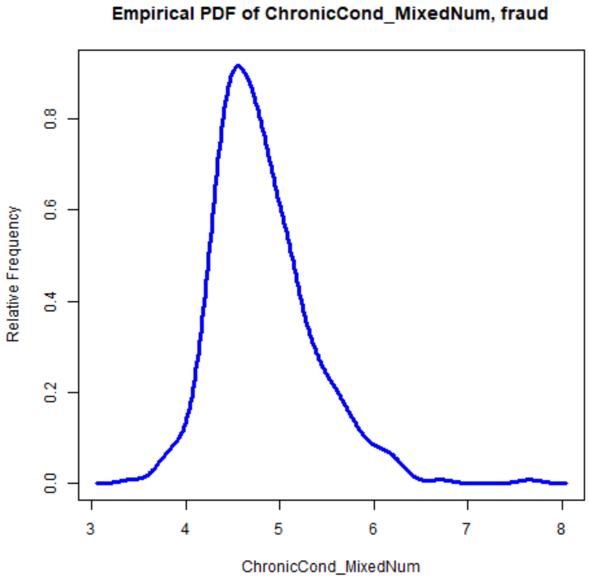

Empirical PDF of ChronicCond_MixedNum, fraud

Figure A2.4

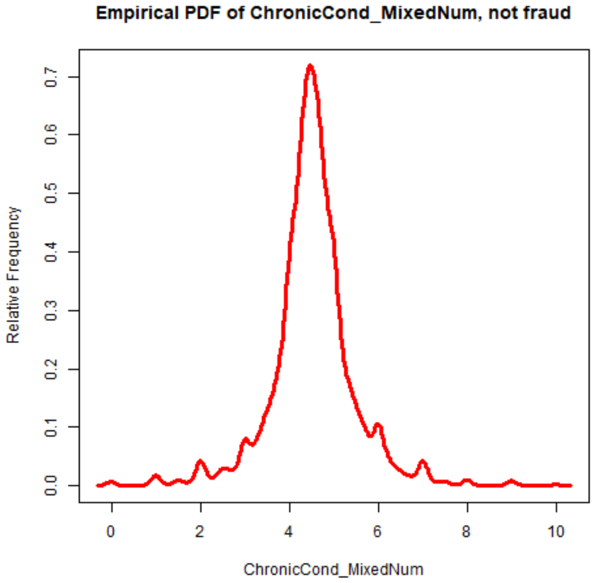

Empirical PDF of ChronicCond_MixedNum, not fraud

Figure A2.5

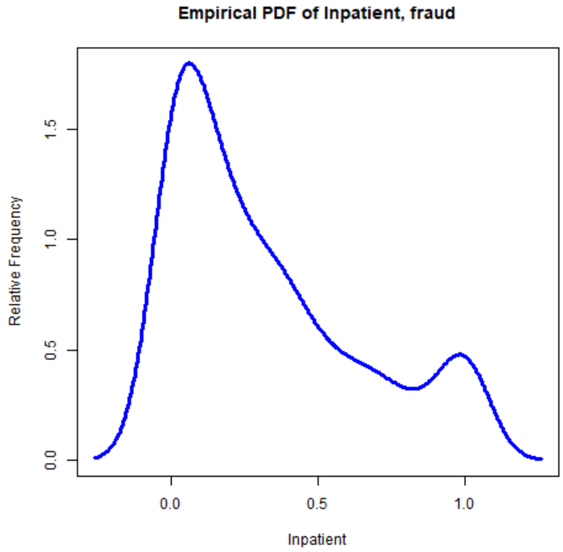
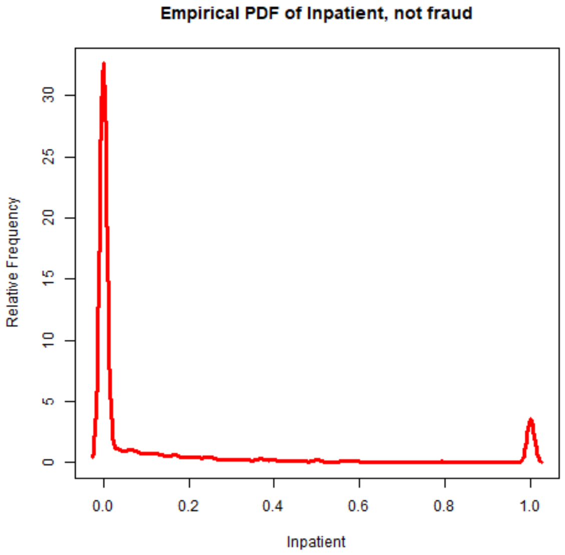

Empirical PDF of Inpatient, fraud

Figure A2.6


Empirical PDF of Inpatient, not fraud

Interaction plots

Figure A3.1

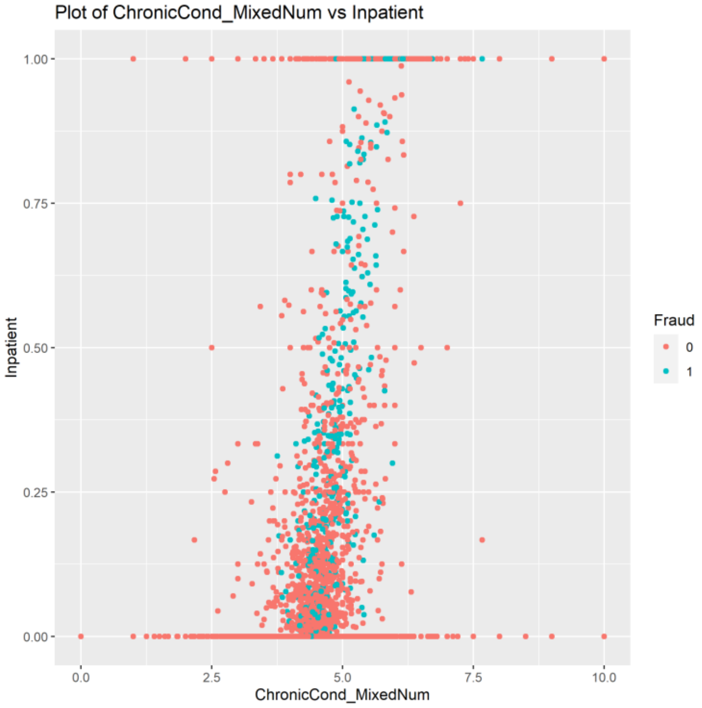


Table 2: *Highly correlated variables*

(Filtered where absolute value(cor) > 0.5)

	Row	Column	cor	p
1	ChronicCond_Alzheimer	ChronicCond_Heartfailure	0.5497501	2.44E-03
2	ChronicCond_Alzheimer	ChronicCond_KidneyDisease	0.5453139	2.69E-03
3	ChronicCond_Heartfailure	ChronicCond_KidneyDisease	0.6910783	4.67E-05
4	ChronicCond_Alzheimer	ChronicCond_ObstrPulmonary	0.506687	5.93E-03
5	ChronicCond_Heartfailure	ChronicCond_ObstrPulmonary	0.6228657	4.00E-04
6	ChronicCond_KidneyDisease	ChronicCond_ObstrPulmonary	0.609342	5.78E-04
7	ChronicCond_Alzheimer	ChronicCond_Diabetes	0.5218536	4.40E-03
8	ChronicCond_Heartfailure	ChronicCond_Diabetes	0.6237593	3.90E-04
9	ChronicCond_KidneyDisease	ChronicCond_Diabetes	0.6439966	2.17E-04
10	ChronicCond_ObstrPulmonary	ChronicCond_Diabetes	0.5468578	2.60E-03
11	ChronicCond_Heartfailure	ChronicCond_IschemicHeart	0.5719079	1.47E-03
12	ChronicCond_KidneyDisease	ChronicCond_IschemicHeart	0.5475998	2.56E-03
13	ChronicCond_Diabetes	ChronicCond_IschemicHeart	0.5842772	1.10E-03
14	InscClaimAmtReimbursed	IPAnnualReimbursementAmt	0.8162662	1.19E-07
15	ChronicCond_Heartfailure	IPAnnualReimbursementAmt	-0.5111644	5.44E-03
16	ChronicCond_KidneyDisease	IPAnnualReimbursementAmt	-0.5895699	9.61E-04
17	ChronicCond_Alzheimer	IPAnnualDeductibleAmt	-0.5197055	4.59E-03
18	ChronicCond_Heartfailure	IPAnnualDeductibleAmt	-0.570772	1.51E-03
19	ChronicCond_KidneyDisease	IPAnnualDeductibleAmt	-0.6134876	5.17E-04
20	ChronicCond_ObstrPulmonary	IPAnnualDeductibleAmt	-0.5961225	8.15E-04
21	ChronicCond_Diabetes	IPAnnualDeductibleAmt	-0.5125133	5.30E-03
22	IPAnnualReimbursementAmt	IPAnnualDeductibleAmt	0.694706	4.10E-05
23	ChronicCond_KidneyDisease	OPAnnualReimbursementAmt	-0.5386162	3.11E-03
24	ChronicCond_KidneyDisease	OPAnnualDeductibleAmt	-0.5491211	2.48E-03
25	OPAnnualReimbursementAmt	OPAnnualDeductibleAmt	0.9838157	0.00E+00

26	ChronicCond_Alzheimer	ChronicCond_MixedNum	-0.7329814	9.16E-06
27	ChronicCond_Heartfailure	ChronicCond_MixedNum	-0.80486	2.43E-07
28	ChronicCond_KidneyDisease	ChronicCond_MixedNum	-0.8104717	1.72E-07
29	ChronicCond_ObstrPulmonary	ChronicCond_MixedNum	-0.77592	1.23E-06
30	ChronicCond_Depression	ChronicCond_MixedNum	-0.6705744	9.43E-05
31	ChronicCond_Diabetes	ChronicCond_MixedNum	-0.7812871	9.25E-07
32	ChronicCond_IschemicHeart	ChronicCond_MixedNum	-0.6920967	4.50E-05
33	ChronicCond_rheumatoidarthritis	ChronicCond_MixedNum	-0.5777044	1.29E-03
34	ChronicCond_stroke	ChronicCond_MixedNum	-0.5638034	1.78E-03
35	IPAnnualReimbursementAmt	ChronicCond_MixedNum	0.6003802	7.31E-04
36	IPAnnualDeductibleAmt	ChronicCond_MixedNum	0.6896903	4.91E-05
37	OPAnnualReimbursementAmt	ChronicCond_MixedNum	0.5126441	5.28E-03
38	OPAnnualDeductibleAmt	ChronicCond_MixedNum	0.5305121	3.68E-03
39	InscClaimAmtReimbursed	ClaimDuration	0.568701	1.59E-03
40	IPAnnualReimbursementAmt	ClaimDuration	0.5524956	2.30E-03