
Analyzing Public Health Data

Faye Ip, Chalenge Masekera, Ramit Malhotra

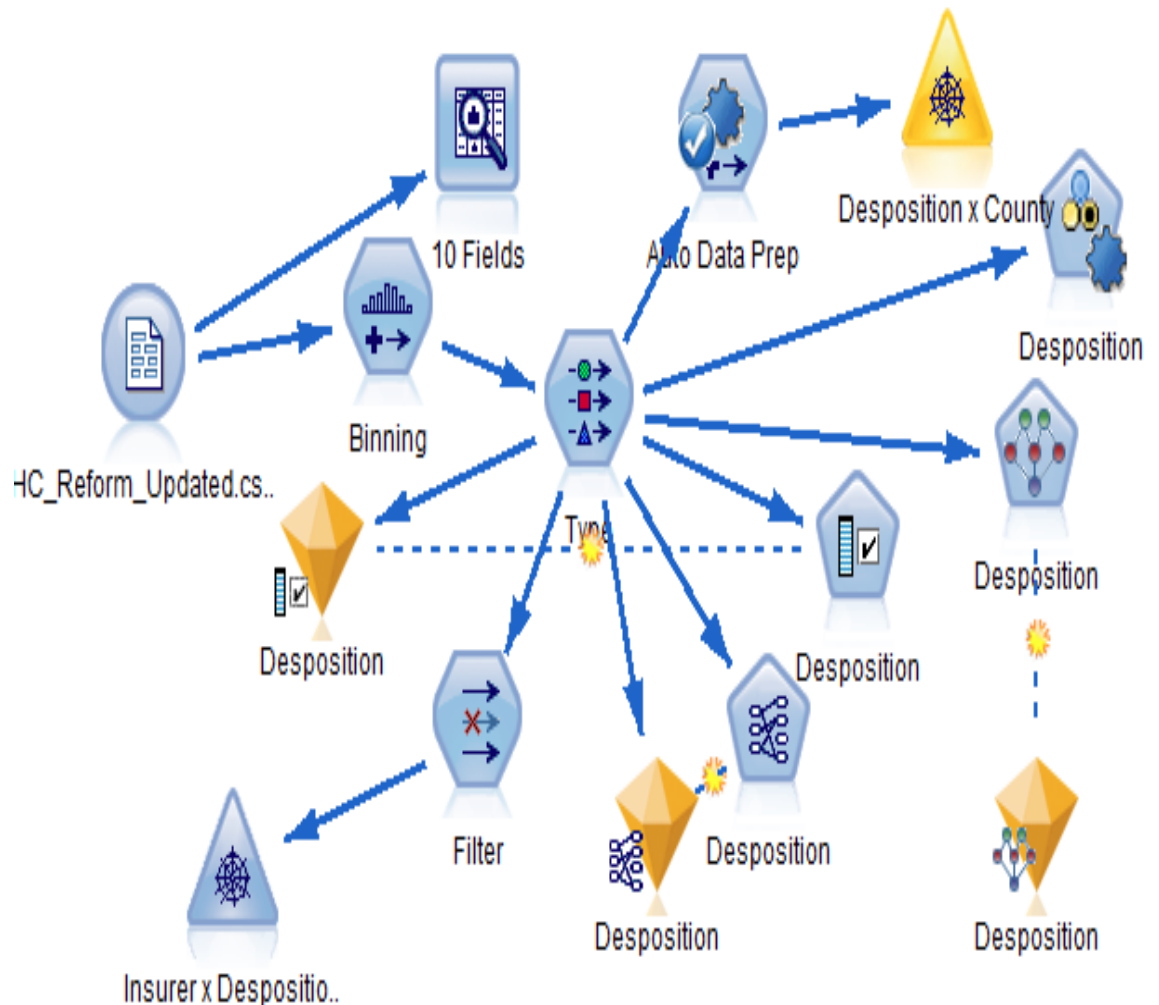
About the dataset

- Contains 330,035 records
- Each record represents one hospital visit in Massachusetts between 2006 and 2008

ICD_9_CODE	DISPO	AGE	SEX	RACE	ZIP	Insur	INSURANCE	DOS	YEAR
455.4	1	37	M	White	01013	MEDICAID MASS	6	17-Sep-08	3
592.1	0	47	M	Puerto Rican	01013	MEDICAID MASS	6	17-Sep-08	3
724.2	1	33	F	Unknown	01013	MEDICAID MASS	6	17-Sep-08	3
723.1	1	20	M	White	01028	MEDICAID MASS	6	17-Sep-08	3
784.0	1	31	M	Puerto Rican	01108	MEDICAID MASS	6	17-Sep-08	3
782.1	1	39	F	American	01108	MEDICAID MASS	6	17-Sep-08	3
560.1	0	65	M	American	01104	MEDICAID MASS	6	17-Sep-08	3
276.51	1	19	F	American	01109	MEDICAID MASS	6	17-Sep-08	3
922.31	1	63	M	Puerto Rican	01108	MEDICAID MASS	6	17-Sep-08	3
300.01	0	45	F	American	01030	MEDICAID MASS	6	17-Sep-08	3
783.6	0	11	M	Unknown/Not Specified	01108	MEDICAID MASS	6	17-Sep-08	3
923.21	1	16	M	White	01082	MEDICAID MASS	6	17-Sep-08	3
345.90	1	18	M	Puerto Rican	01109	MEDICAID MASS	6	17-Sep-08	3
847.0	0	23	M	White	05301	MEDICAID VERMONT	6	17-Sep-08	3
924.8	1	32	F	White	01001	AUTO COMMERCE INS	4	18-Sep-08	3
924.8	1	5	F	White	01001	AUTO COMMERCE INS	4	18-Sep-08	3
812.01	0	16	M	White	01095	AUTO COMMERCE INS	4	18-Sep-08	3
847.0	1	38	M	Vietnamese	01013	AUTO COMMERCE INS	4	18-Sep-08	3
959.01	1	20	M	White	01119	AUTO OTHER	4	18-Sep-08	3
847.0	1	46	F	Unknown/Not Specified	01020	AUTO OTHER	4	18-Sep-08	3
813.43	1	88	M	White	01013	AUTO OTHER	4	18-Sep-08	3


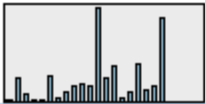





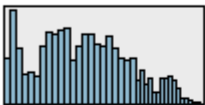


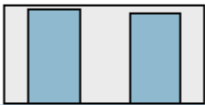




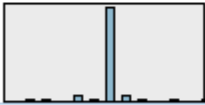




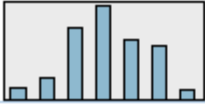

Tools

1. SPSS
2. Tableau
3. SciKit
4. WEKA
5. Python (lol)

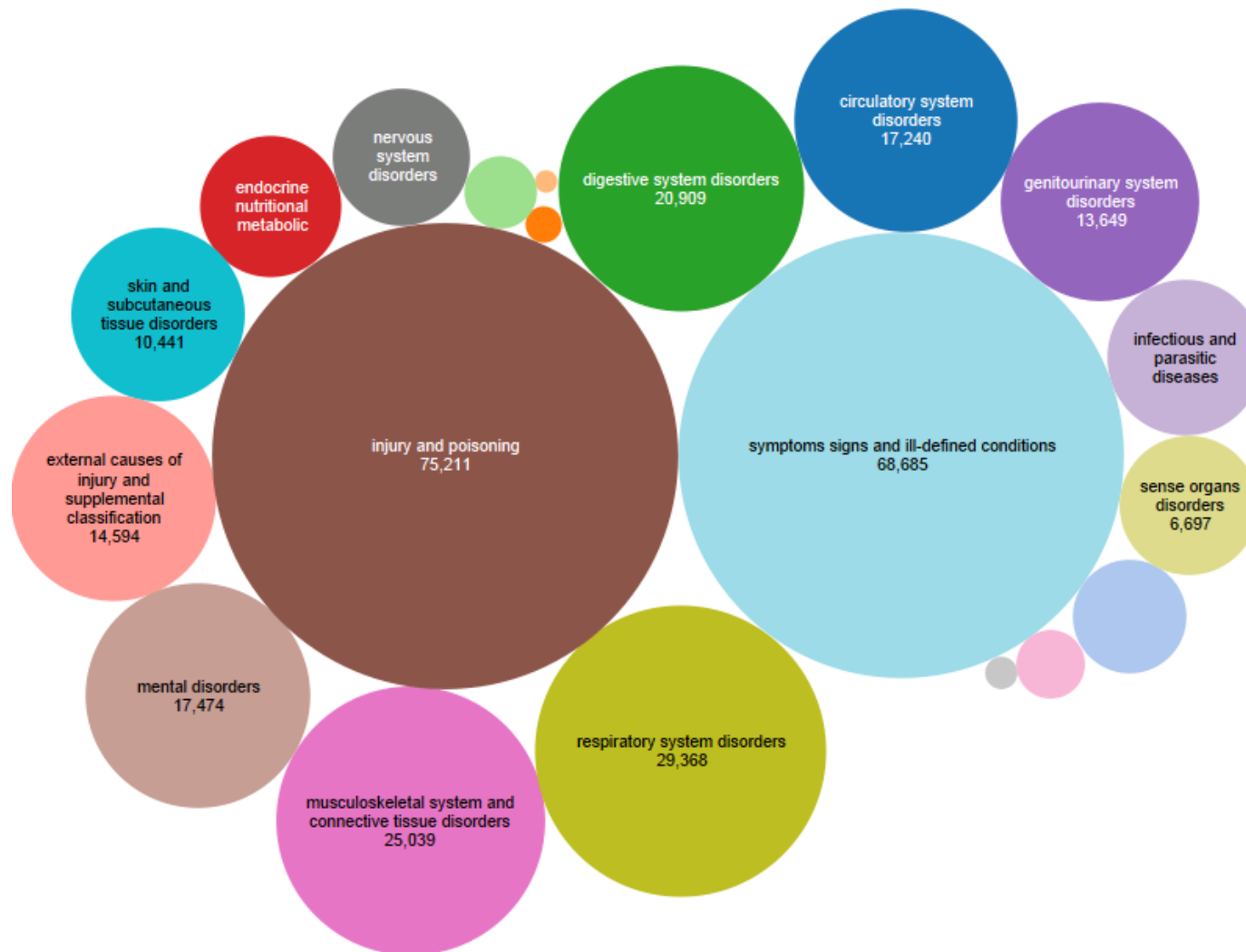


Preliminary Data Analysis with SPSS

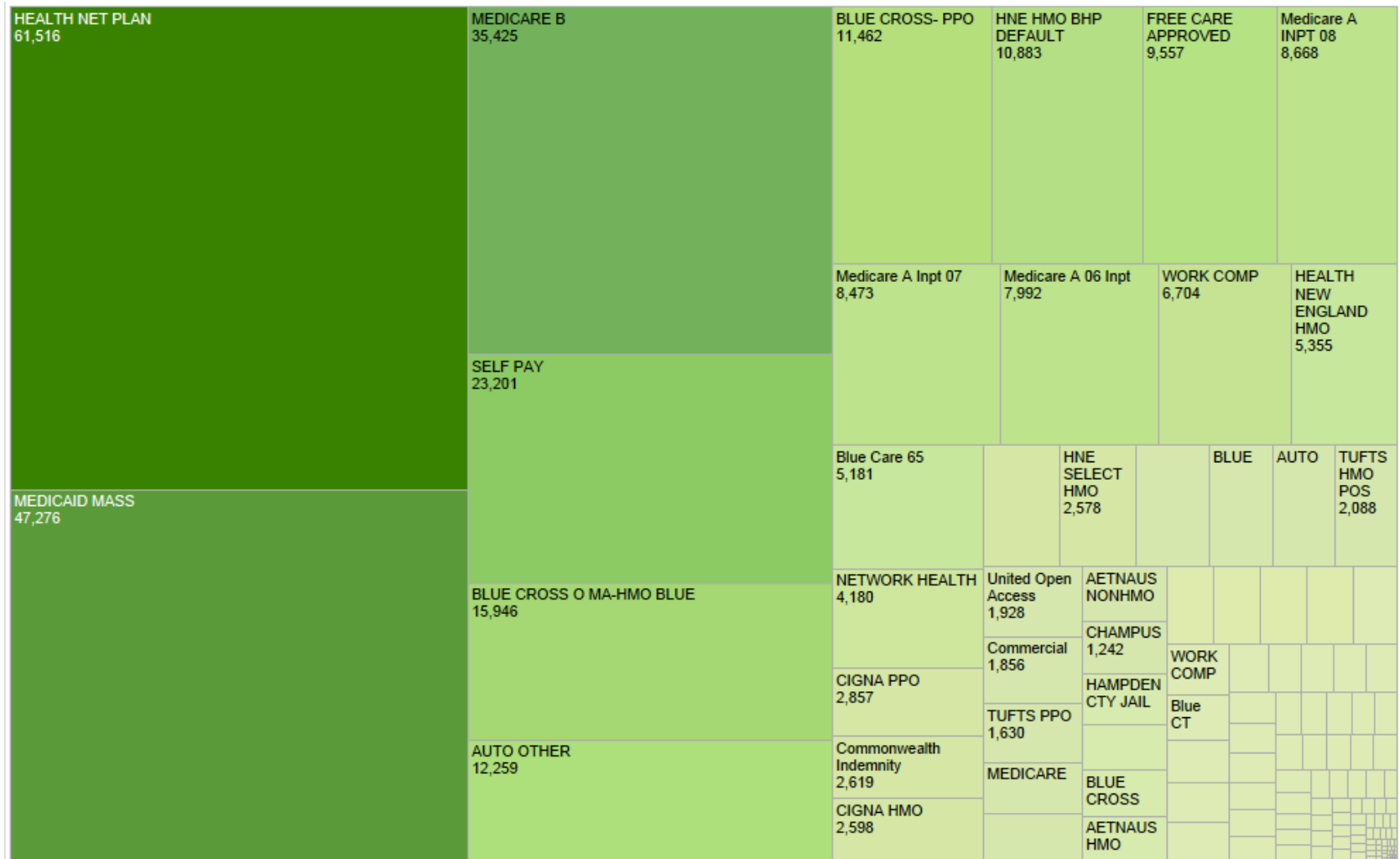
Features: NOMINAL, CONTINUOUS:

 Disease_Category		 Nominal	--	--	--	--	--	20	330035
 Desposition		 Nominal	--	--	--	--	--	3	330035
 Age		 Continuous	0	110	39.360	23.824	0.370	--	330035
 Gender		 Nominal	--	--	--	--	--	2	330035
 Race		 Nominal	--	--	--	--	--	120	330013
 County		 Nominal	--	--	--	--	--	14	330035
 Insurer		 Nominal	--	--	--	--	--	140	330035
 Insurance_Categ...		 Nominal	1	7	--	--	--	7	330035

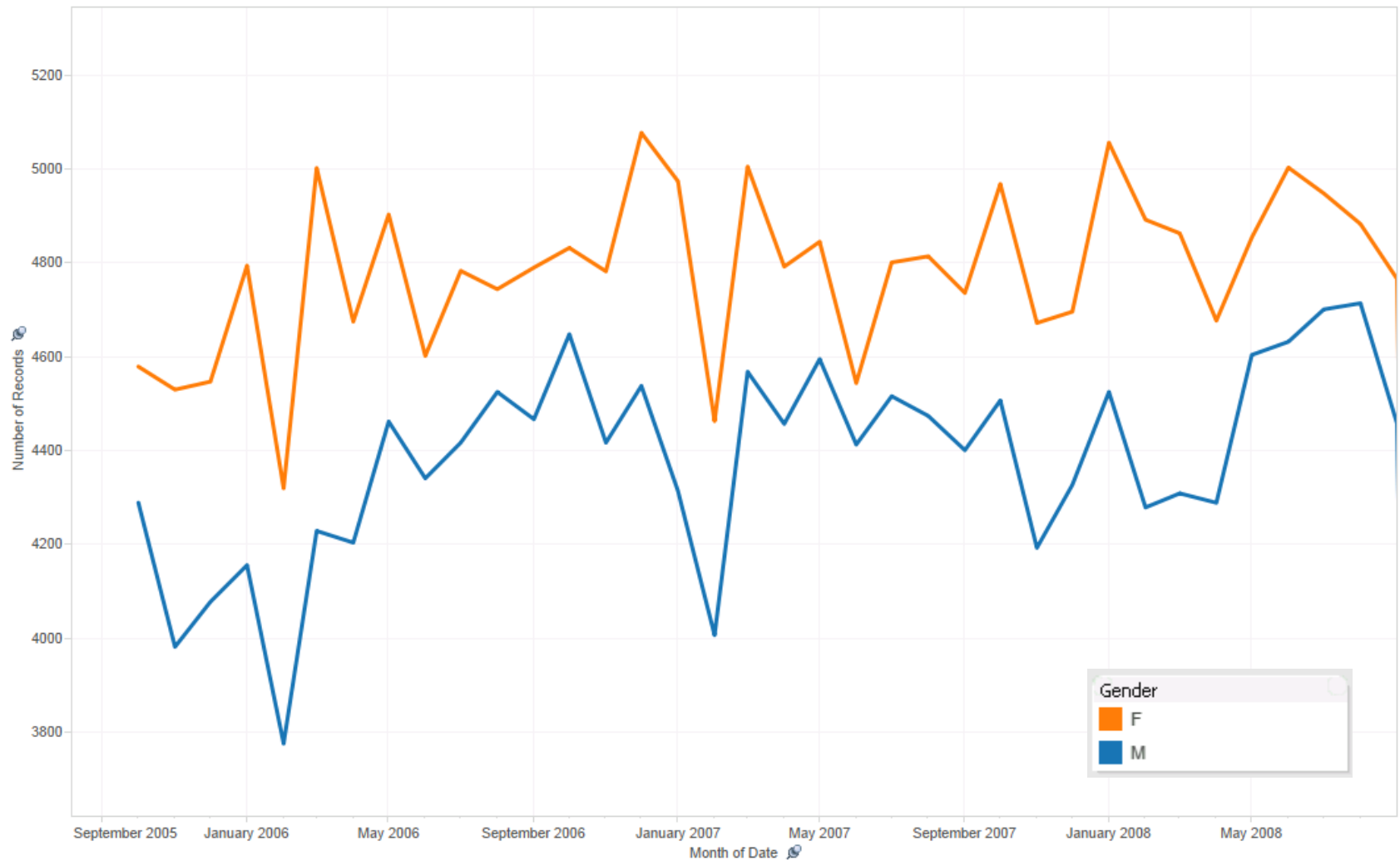
Data Exploration with Tableau



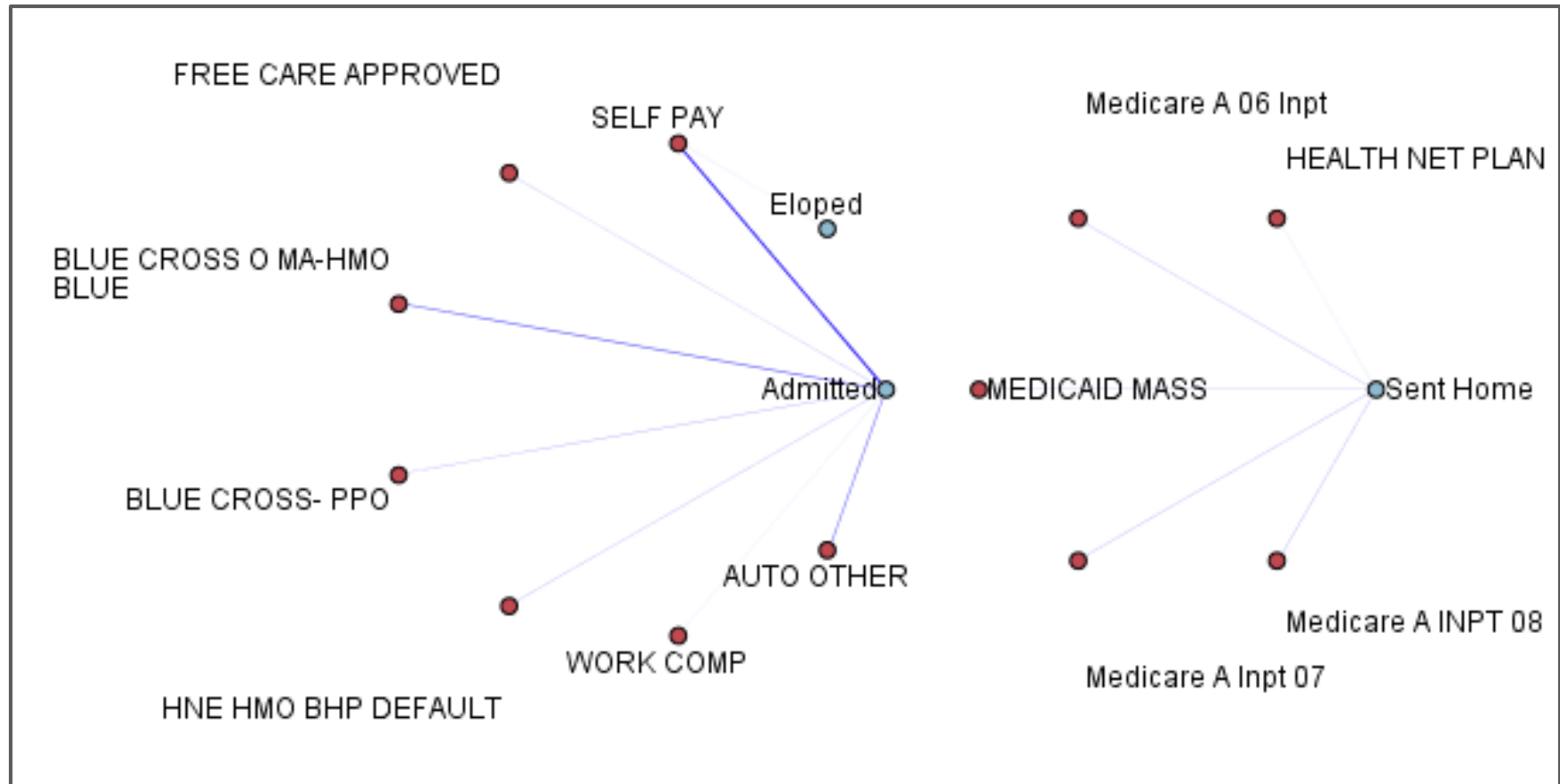
Data Exploration with Tableau



Data Exploration with Tableau



Preliminary Data Analysis with SPSS



Preliminary Data Analysis with SPSS

Disease_Category x Disposition

Strong Links

Links	Field 1	Field 2
64,335	Disease_Category = "injury and poisoning"	Disposition = "Admitted"
52,865	Disease_Category = "symptoms signs and ill-defined conditions"	Disposition = "Admitted"
22,608	Disease_Category = "musculoskeletal system and connective tissue disorders"	Disposition = "Admitted"
21,067	Disease_Category = "respiratory system disorders"	Disposition = "Admitted"
13,840	Disease_Category = "mental disorders"	Disposition = "Admitted"
13,152	Disease_Category = "circulatory system disorders"	Disposition = "Sent Home"
13,013	Disease_Category = "digestive system disorders"	Disposition = "Admitted"
11,484	Disease_Category = "external causes of injury and supplemental classification"	Disposition = "Eloped"
10,655	Disease_Category = "symptoms signs and ill-defined conditions"	Disposition = "Sent Home"
10,392	Disease_Category = "injury and poisoning"	Disposition = "Sent Home"
9,750	Disease_Category = "genitourinary system disorders"	Disposition = "Admitted"
8,935	Disease_Category = "skin and subcutaneous tissue disorders"	Disposition = "Admitted"
8,118	Disease_Category = "respiratory system disorders"	Disposition = "Sent Home"
7,676	Disease_Category = "digestive system disorders"	Disposition = "Sent Home"

Frequent
Patterns:

Disease
Category to
Patient
Disposition:

Results:
Injury and
Poisoning →
Admitted

External Injury
and Others →
Eloped

Decision Tree Classification

Our Decision Tree analyzes the dataset in 5 steps:

1) Preprocess:

Initial cleaning up of data and feature formats

2) Prepare: Vectorises nominal features and separates class labels for training and testing.

3) Train: Uses 95% of the data to train our model, and cross validates it with 10% of the training data

Test: We use the remainder of the dataset to test our model for accuracy.

Output: We then display the class predictions along with their accuracy and probability for each test row.

We save this information in a results file.

ATTRIBUTE SELECTION

WEKA helped us with feature selections:

Attribute selection output

=== Attribute Selection on all input data ===

Search Method:
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 Date):
Information Gain Ranking Filter

Ranked attributes:

0.84841	1	Disease_Code
0.40362	7	Zip_Code
0.3406	8	Insurer
0.13375	6	Race
0.07492	2	Disease_Category
0.01305	3	Desposition
0.00246	5	Gender
0	4	Age
0	9	Insurance_Category

Selected attributes: 1,7,8,6,2,3,5,4,9 : 9

=== Attribute Selection on all input data ===

Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 64
Merit of best subset found: 0.017

Attribute Subset Evaluator (supervised, Class (nominal): 5 Gender)
CFS Subset Evaluator
Including locally predictive attributes

Selected attributes: 1,2,6,7,10 : 5
Disease_Code
Disease_Category
Race
Zip_Code
Insurance_Category

Attribute Subset Evaluator (supervised, Class (nominal): 10 Date)
CFS Subset Evaluator
Including locally predictive attributes

Selected attributes: 1,7,8 : 3
Disease_Code
Zip_Code
Insurer

Decision Tree Classification

```
Cross Validation Score
0.98
Mean Accuracy is: 0.725498883278
Classification Report:
```

	precision	recall	f1-score	support
0	0.89	0.50	0.64	19867
1	0.43	0.40	0.42	71271
2	0.79	0.84	0.82	238398
avg / total	0.72	0.73	0.72	329536

Test Prediction Results

The top row is the actual record: ICD_9_CODE, DISEASE CATEGORY, DISPO, AGE, SEX, RACE, ZIP, COUNTY, INSURER, INSURANCE, DATE, YEAR.

The percentage-wise breakdown of the predicted disposition is below each of the actual records.

Record #0 : V64,external causes of injury and supplemental classification,Eloped,42,M,White,1105,Hampden County,SELF PAY,2,02-Oct-05,2006

['Eloped: 100.0%']

Record #1 : 784,symptoms signs and ill-defined conditions,Admitted,44,F,African American,1109,Hampden County,SELF PAY,2,02-Oct-05,2006

['Admitted: 100.0%']

Record #2 : 802,injury and poisoning,Admitted,36,M,White,12180,Essex County,SELF PAY,2,02-Oct-05,2006

['Admitted: 100.0%']

Record #3 : 786,symptoms signs and ill-defined conditions,Admitted,43,F,Puerto Rican,1151,Hampden County,SELF PAY,2,02-Oct-05,2006

['Admitted: 100.0%']

Record #4 : 625,genitourinary system disorders,Admitted,31,F,American,1057,Hampden County,SELF PAY,2,02-Oct-05,2006

['Admitted: 100.0%']

Record #5 : V64,external causes of injury and supplemental classification,Eloped,25,M,Irish,1082,Hampshire County,SELF PAY,2,02-Oct-05,2006

['Eloped: 100.0%']

Record #6 : V64,external causes of injury and supplemental classification,Eloped,46,F,White,1151,Hampden County,SELF PAY,2,02-Oct-05,2006

['Eloped: 100.0%']

Future work

- Validate our analysis by using data from other hospitals
 - Anomaly Detection - for medical aid payments
 - Use other mining techniques in an exploratory fashion to find more patterns
-