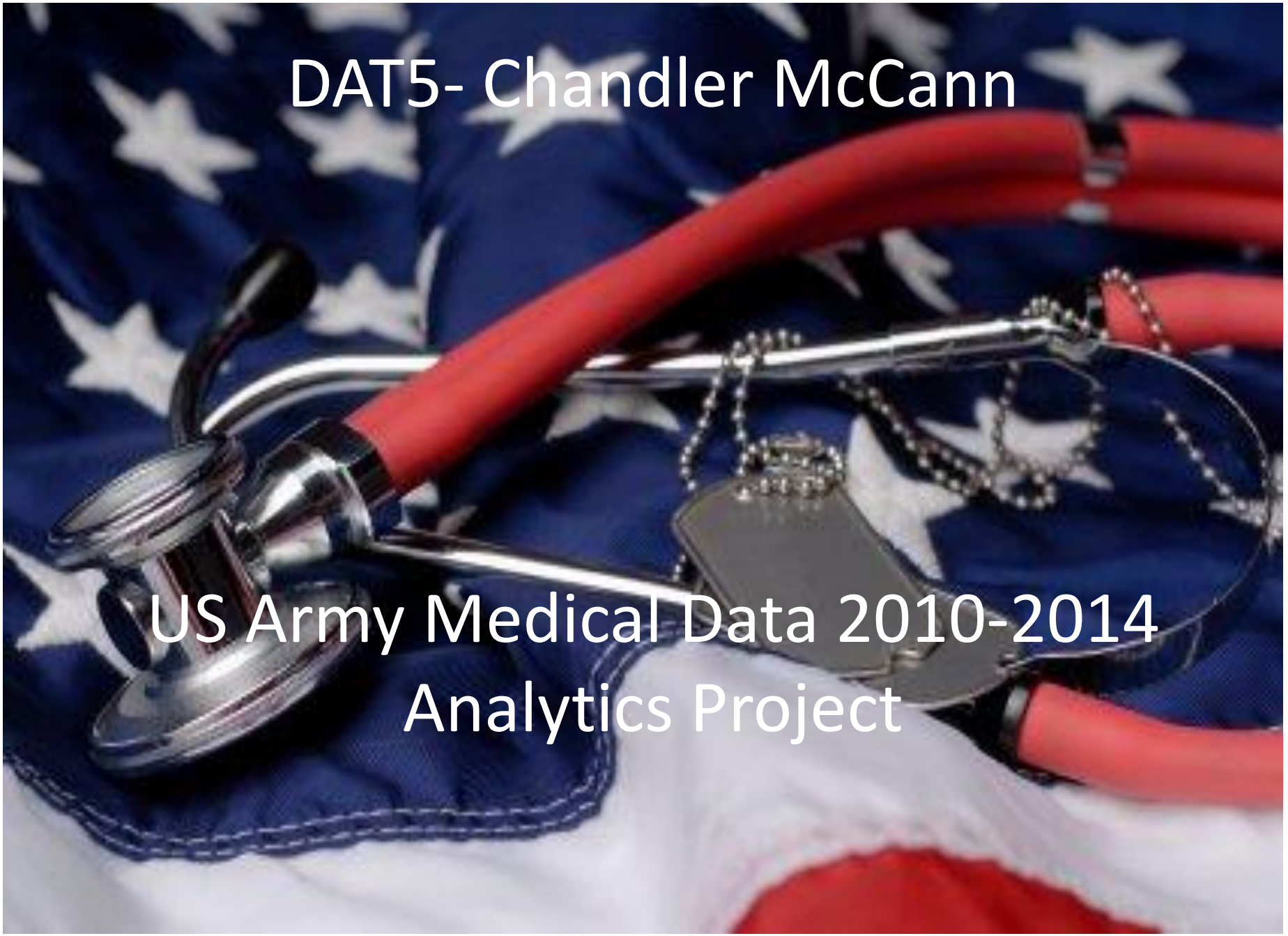


DAT5- Chandler McCann

US Army Medical Data 2010-2014 Analytics Project

The Problem

- U.S. Army medical expenses cost ~660M annually from work place illnesses and injuries*
- Similarly, the Army spends ~ 250M per year on Civilian workers compensation claims
- In parallel, this results in >100,000 days away from work, impacting the workforce and Army readiness.

- *Systems and tools for analytics are only recently growing*
- *Opportunities exist to reduce worker injuries through policy and programs, however we don't know where to apply "pressure"*

*not including medical equipment and prosthetics costs

The Question

- How do I reduce cost and time away from work due to injuries?
- Specifically
  - Can I create a model that identifies features that have the strongest correlation to time in the hospital or cost

The Data

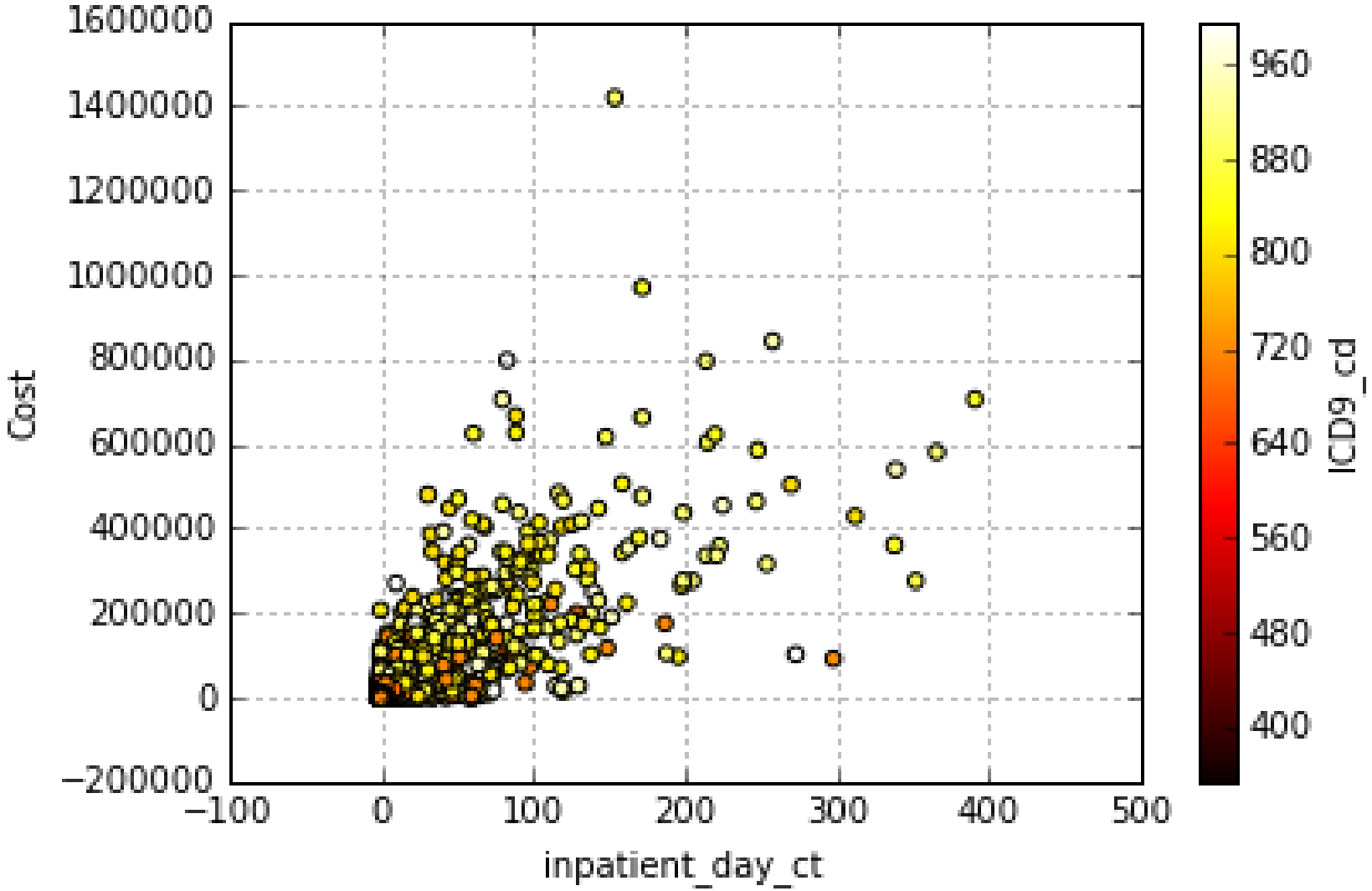
- 2010 – 2014 from DoD
- All military medical events, including combat (Iraq, Afghanistan) and non-combat (global bases eg Seoul, Germany, Italy, etc)
- Shape of 2010: (554737,37)
- 37 columns with two continuous responses, Cost and Inpatient Day count

	gender	occupation	month	FY	mtfcd	mtfstate	ICDL2	inpatient	LT	Cost	...	ICDL2_OPEN,WOUND OF LOWER LIMB (890-897)	ICDL2_OPEN,WOUND OF UPPER LIMB (880-887)	ICDL2_OST CHONDROI AND ACQU MUSCULOS DEFORMIT
0	1	AIR TRAFFIC CONTROL	1	2010	1599	TEXAS	Superficial Injury	0	0	494.52	...	0	0	0
1	1	AIRCRAFT STRUCTURES	1	2010	0108	TEXAS	SPRAINS,AND STRAINS OF JOINTS AND ADJACENT MUS...	0	0	783.45	...	0	0	0
		AIRCRAFT					SPRAINS,AND STRAINS OF							

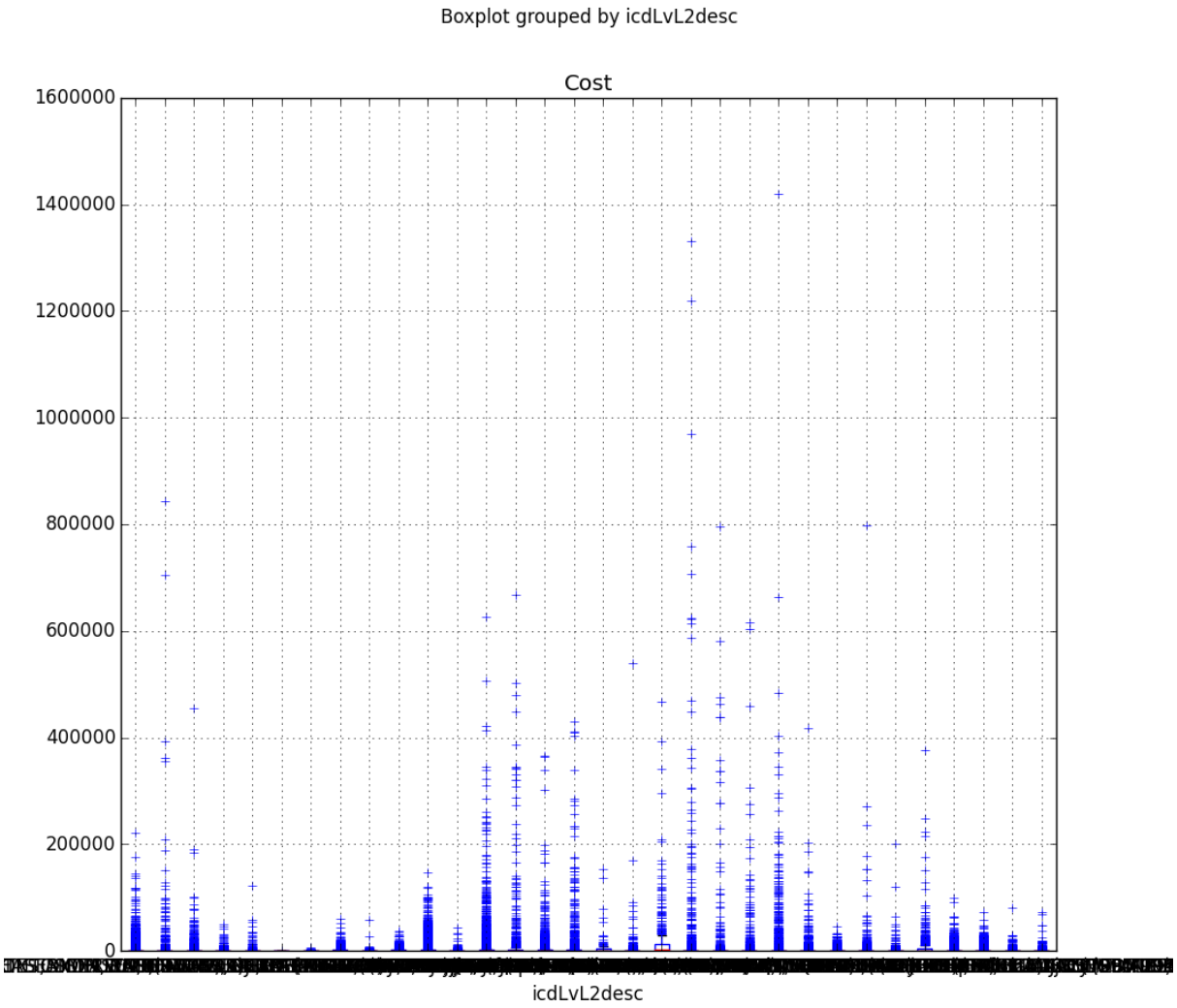
Data Exploration and Cleaning

- Highlights:
  - Many features I want as integers loaded as objects
  - 5 “types” of ICD9 codes, Level 2 has 32 values, Level 1 has 5, and Level 3 has > 200. Using L2
  - No null values on features I care about, however “not in dimension” for treatment facilities turned out to be non-medical (private) hospitals
  - Feature Set to model
    - Gender
    - Medical Treatment Facility State (region)
    - Occupation
    - ICD9 Level 2 Code (injury type)
    - Month of injury (time)
  - Responses
    - Inpatient Time (response A)
    - Cost (response B)
- Lowlights
  - Making a lot of visualizations and exploratory code prior to “cleaning” my feature names→ Duplication

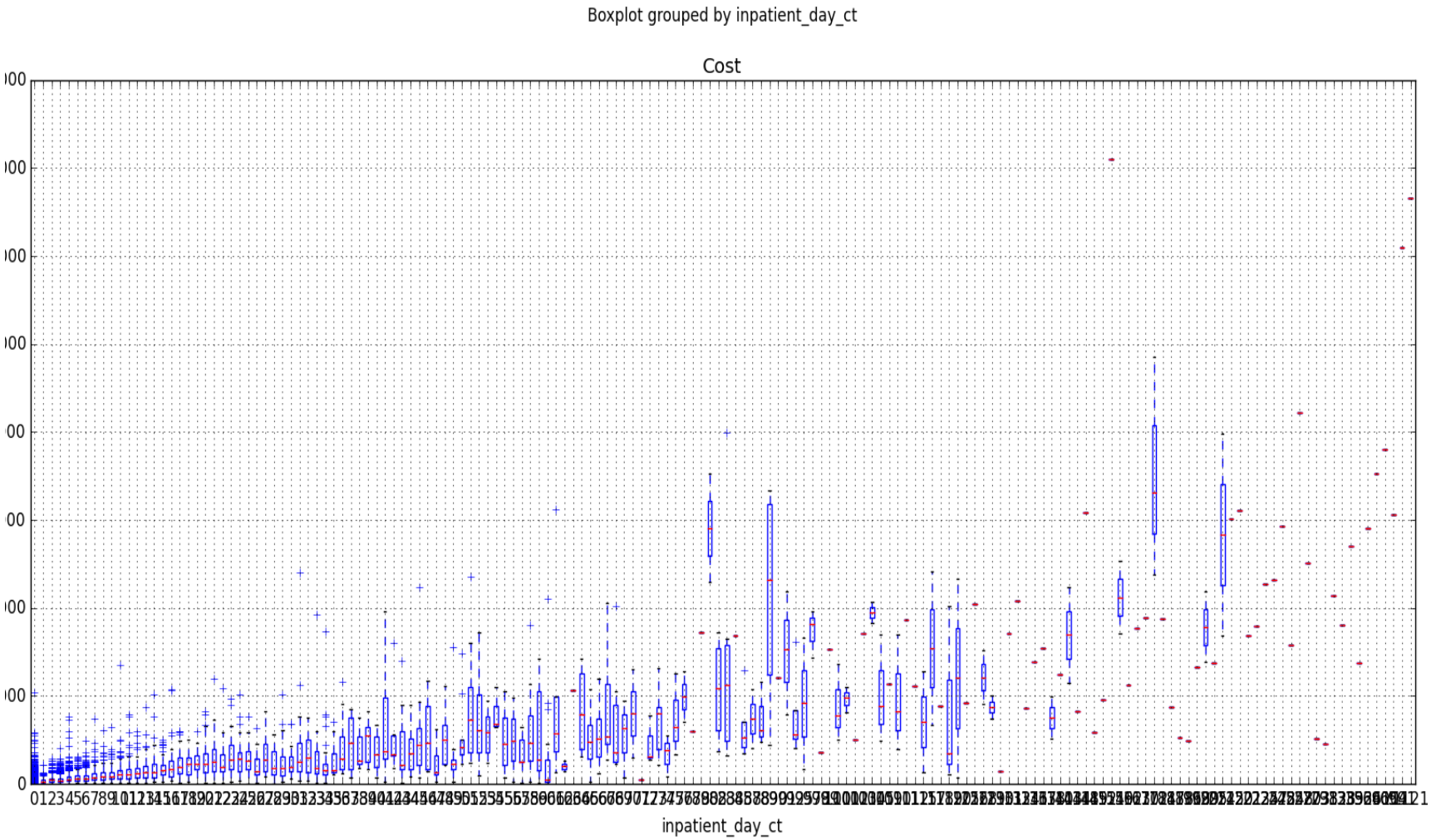
Cost, inpatient days and ICD9



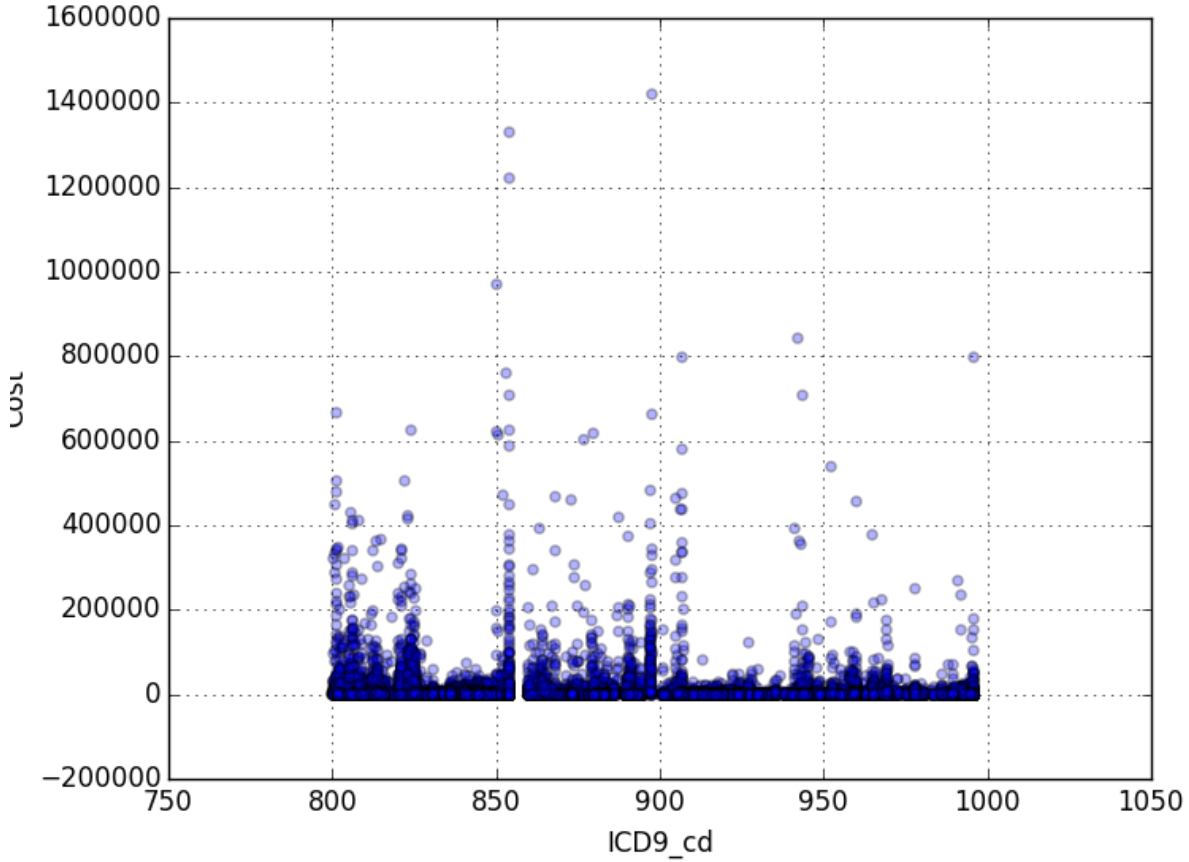
Cost vs ICDLVL2 Code



Cost vs Inpatient Days



ICD9 Code > 800 vs Cost



Models

- Developed 1 very bad model
  - Linear regresssion
    - RMSE 5874…

```
[('gender', -328.48064127217748),
 ('LT', 7448.9643367018853),
 ('month_2', 344891756585.09436),
 ('month_3', -344891756593.4613),
 ('month_4', 11841283658901.047),
 ('month_5', -11841283658830.91),
 ('month_6', 1179679091197.5239),
 ('month_7', -1179679091184.4863),
 ('month_8', -2470367766316.7529),
 ('month_9', 2470367766464.2466),
 ('month_10', -
3775259405522.1162),
 ('month_11', 3775259405571.1221),
 ('month_12', -
1565259754323.5764)]
```

Next Steps

- Filter out combat related injuries and make a separate DF for that
- Create region feature for treatment facility
- Create occupation “buckets” to reduce the number of values and group similar “risk level” jobs
- Filter out foreign injuries in
- Gather data that includes non-injuries to predict likelihood of injury by type
  - Injury rate data by occupation currently captured