# NHL PLAYER SALARIES

BY: CHARLES HESKETH

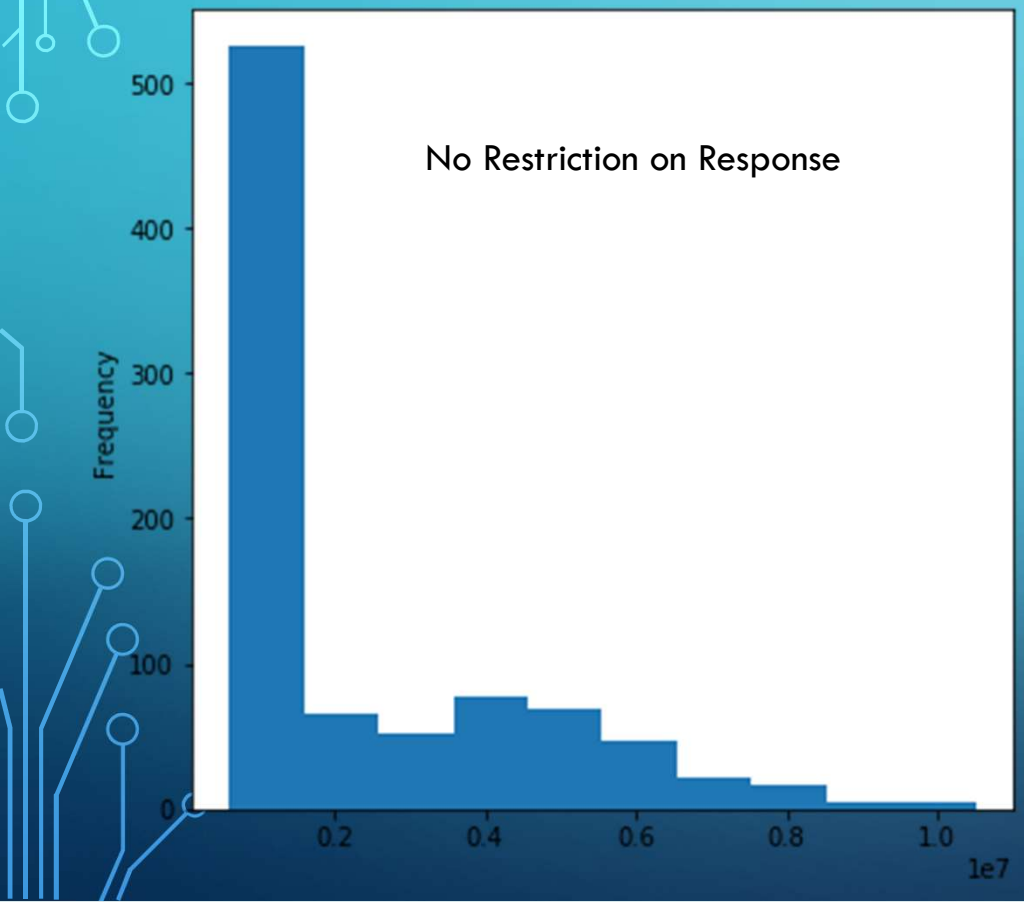
# DATA SET - PROS

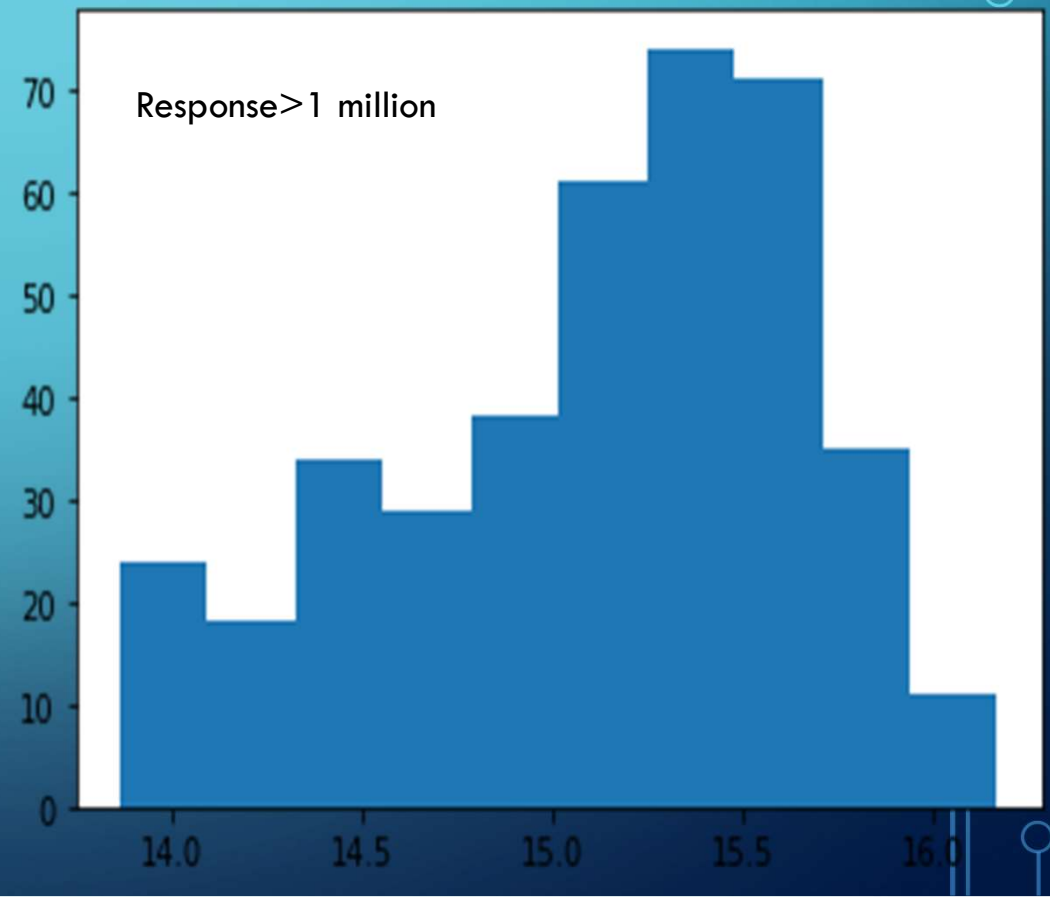
- Curated from Robert Vollman http://www.hockeyabstract.com/testimonials
- Several Years worth available.
- Each year has more data points than the year before.
- Contains “Advanced Analytics” i.e. Corsi
- Dense in metrics (over 150 metrics measured)

# DATA SET - CONS

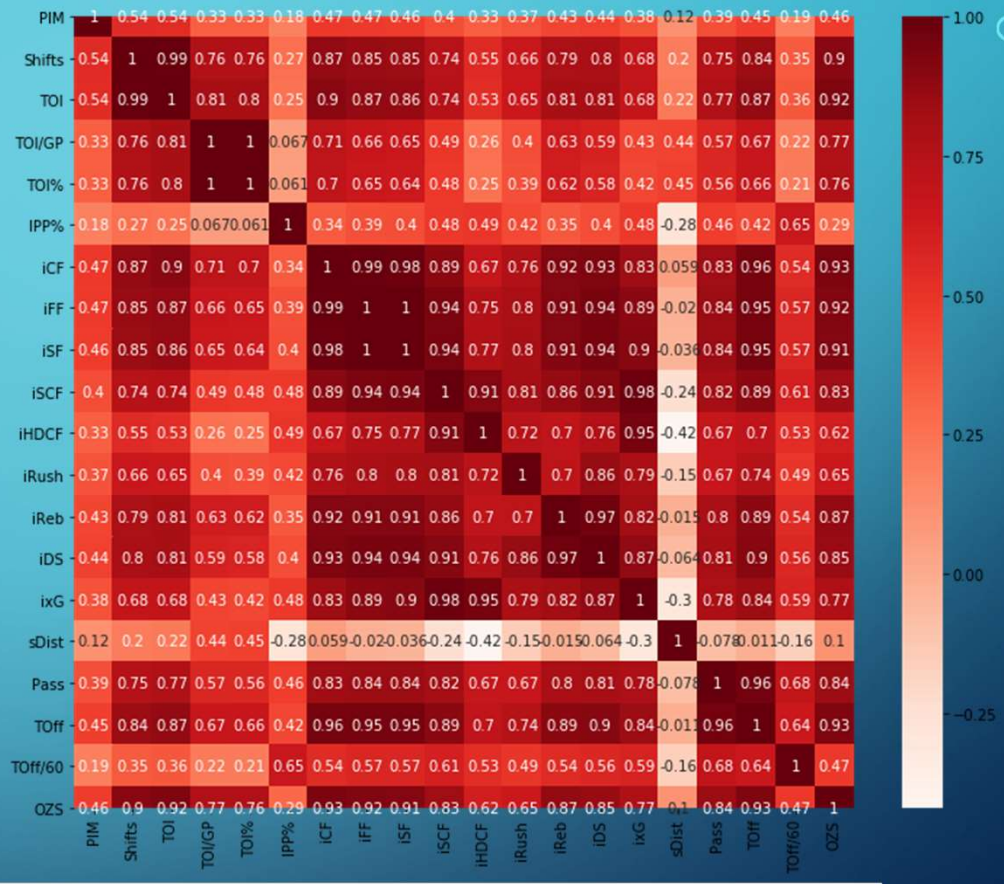
- Only two years with salary.
- Does not have length of contract
- Does not have what year in their contact they are (1st, 2nd, 3rd,etc.)
- Dense in metrics (over 150 metrics measured)
- Not many observations compared to metrics (700ish)
- Hyper Skewed response variable (Cap Hit).
- What is Independence?

# CAP HIT (RESPONSE VARIABLE) HISTOGRAM



# CORRELATION MAP AND FEATURE REDUCTION




```
season18[season18.columns[~
                season18.columns.isin(["TOI.1","TOI.2","TOI.3","iCF.1","iCF.2","iFF.1","iHF.1","iPENT2",
                                       "iPENT5","iPEND2","iPEND5",'CF.1','CA.1','FF.1','FA.1','SF.1','SA.1'
                                       ,'GF.1','GA.1','GP.1', 'G.1', 'A.1', 'PTS.1', '+/-.1','PIM.1',
                                       'TOI.4', 'OPS.1', 'DPS.1', 'PS.1','A1'])]]
```

# DATA PREP AND FEATURE SELECTION

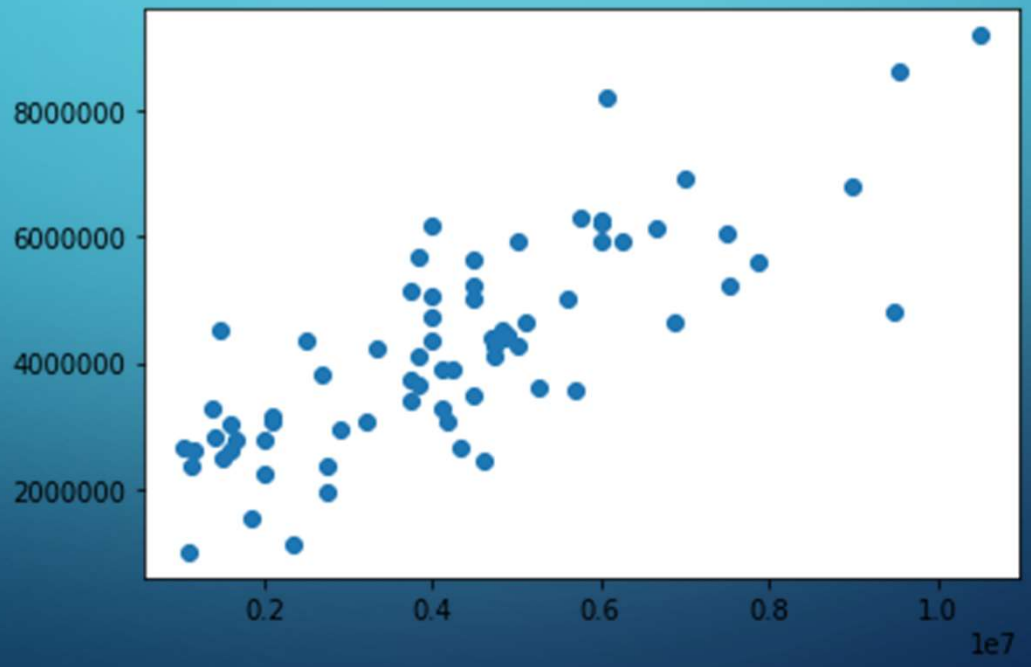
- Remove “duplicate” columns (i.e. TOI1,TOI2,TOI3 in favor of TOI)
- Remove columns with many NA’s (7 Non-Informative Columns)
- Remove Rows with NA’s (About 140 rows)
- Normalize all columns besides the response column
- Used Backward Elimination for one set of Feature Selection
- Used Lasso as another set of Feature selection
- Response variable as is and Log of Response variable

# HOW MODELS WERE IMPLEMENTED

- All manually trimmed variables used in the model w/ no transformation on Y
- All manually trimmed variables used in the model w/ log(Y)
- Backwards Elimination variables used in the model w/ no transformation on Y
- Backwards Elimination variables used in the model w/ log(Y)
- Lasso Regression variables used in the model w/ no transformation on Y
- Lasso Regression variables used in the model w/ log(Y)
- Bootstrap was used to create confidence intervals

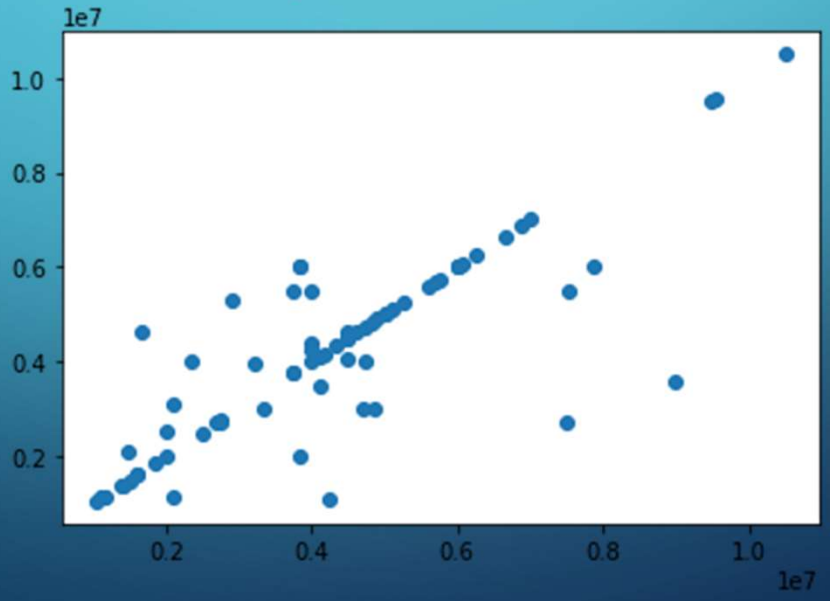
# LASSO

- Python Function LassoCV did not Converge after 1 million iterations.



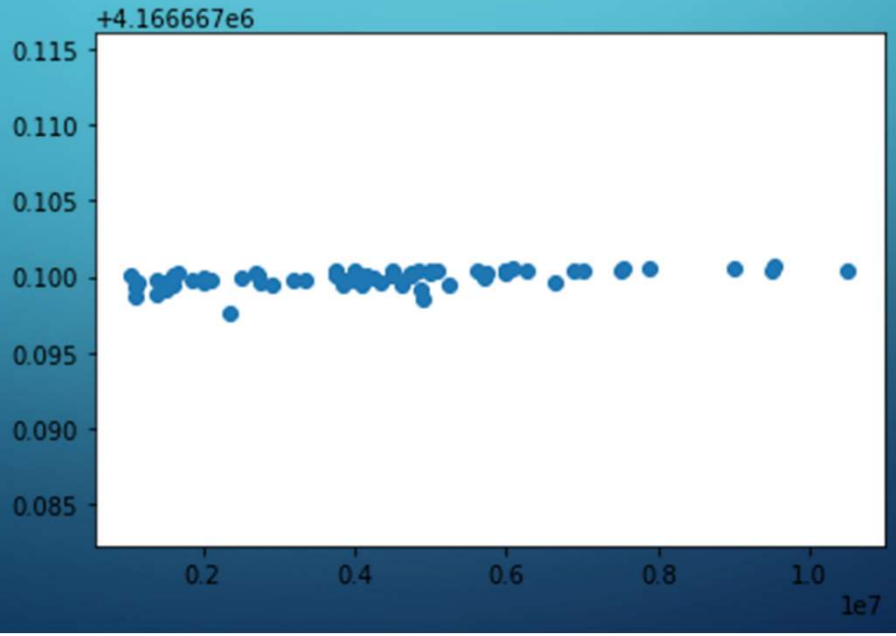
# DECISION TREE REGRESSION

- This produced a confidence interval of (1.8M,1.2M). Interestingly it appears it was fairly close at times, but way off at other times.



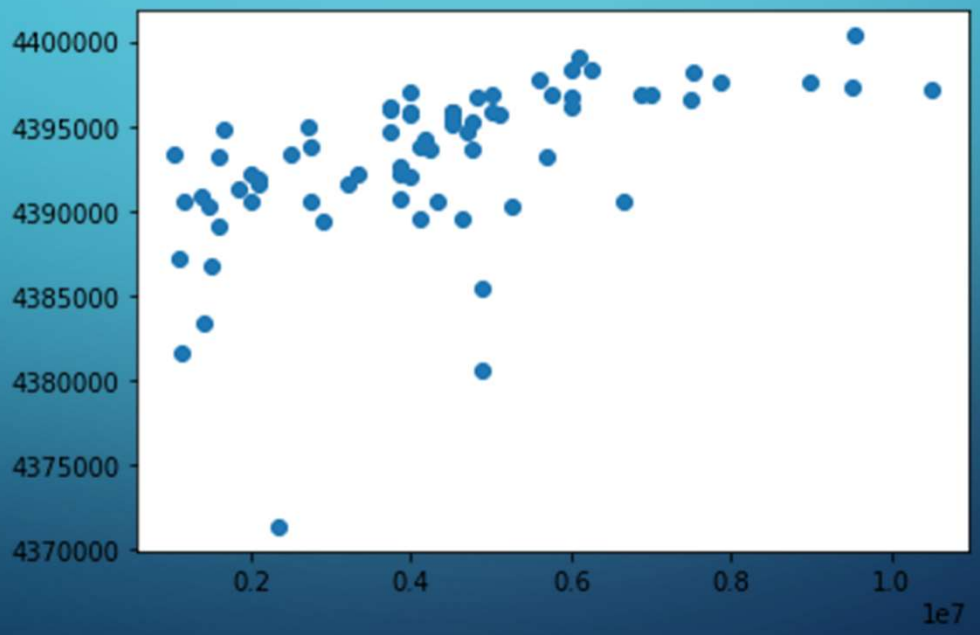
# SUPPORT VECTOR REGRESSION (SVR)

- This gave a confidence interval of (0,0), but this obviously ran into issues… it, also, did not converge:



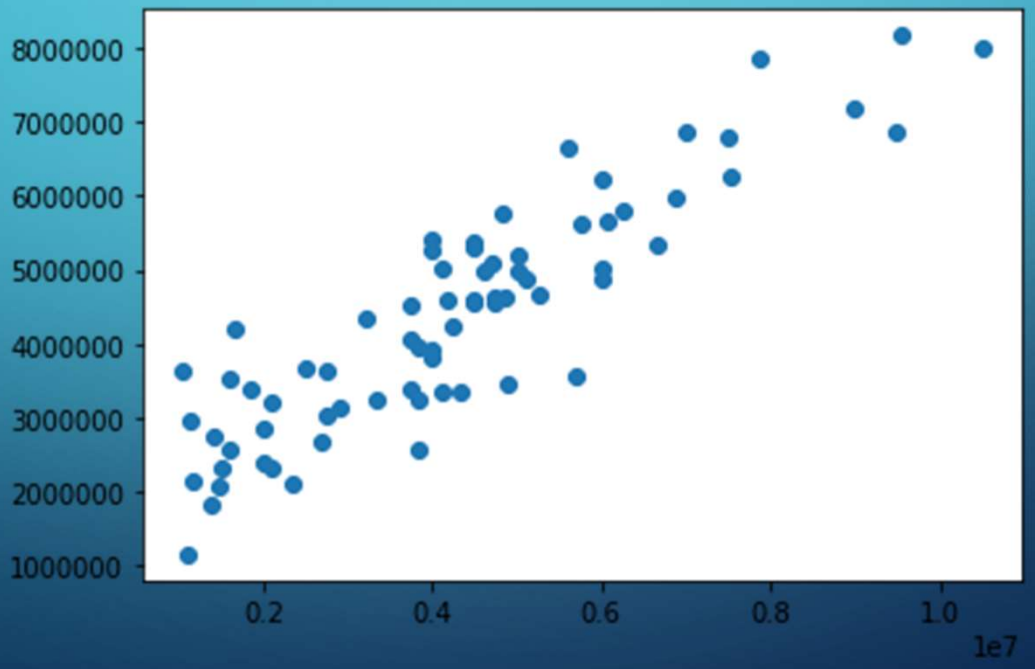
# STOCHASTIC GRADIENT DECENT (SGD)

- This gave a confidence interval of (1,874.4,35,123.7)



# BAGGING

- This produced a confidence interval of (1.4M,1.8M)



# MULTI-LAYER PERCEPTRON (MLP) NEURAL NETWORK

- This took way too long to realistically use in bootstrapping. One pass would converge somewhere above 1M iterations and less then 3M iterations. This did not make it realistic and the model was not competitive so no reason to put my computer through BS.

# BOOSTING

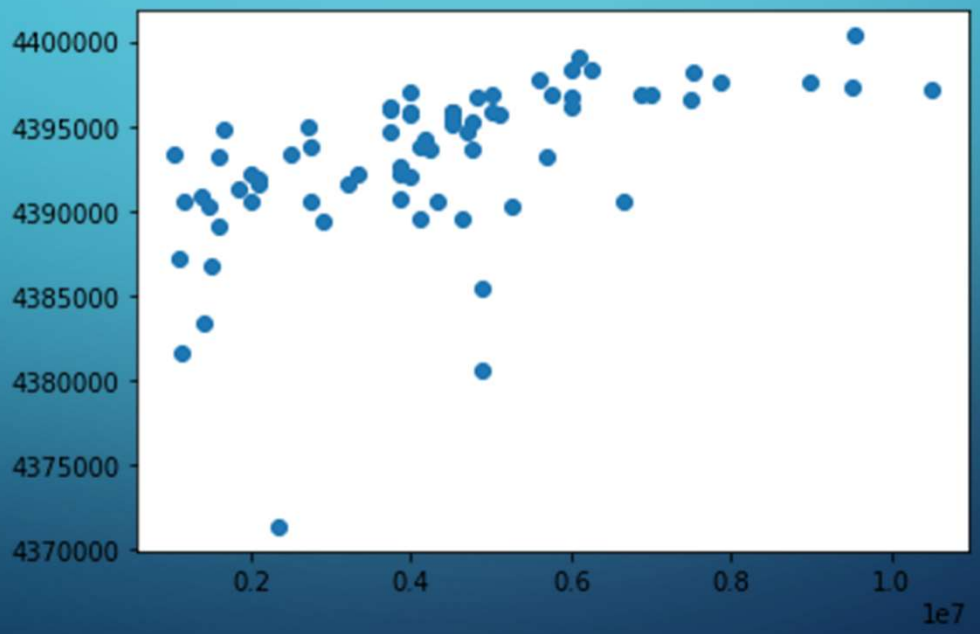
- Special set up needed to run this boot strap. Results will be found in the paper.

# PCA W/ LASSO

- Don't expect this to converge, but similarly needs special set up to do bootstrapping. Will be found in paper.

# BEST MODEL: SGD

- SGD was clearly the winner.



# POTENTIAL TO HELP CONVERGENCE

- I am hoping to implement a log transformation on my y to help with convergence.
- This was tested and proved to help LassoCV converge; however, I was having issues in python implementing exp() .
- Still didn't fix all convergence for SVR for sure and possibly others.

# WHY WE AREN'T GETTING BETTER RESULTS

- Contracts signed for this year are most indicative of results
- Long term trends are not available in the data set.
- Penalizing terms should be used on trending data, which is not available.
- Would like to use previous contracts to predict the new contracts to be given the next year as a test.
- Complexities within NHL CBA that need to be accounted for or protected against (i.e. minimum salaries and salary control measures)