



# Boston Marathon

Could you be the next winner?



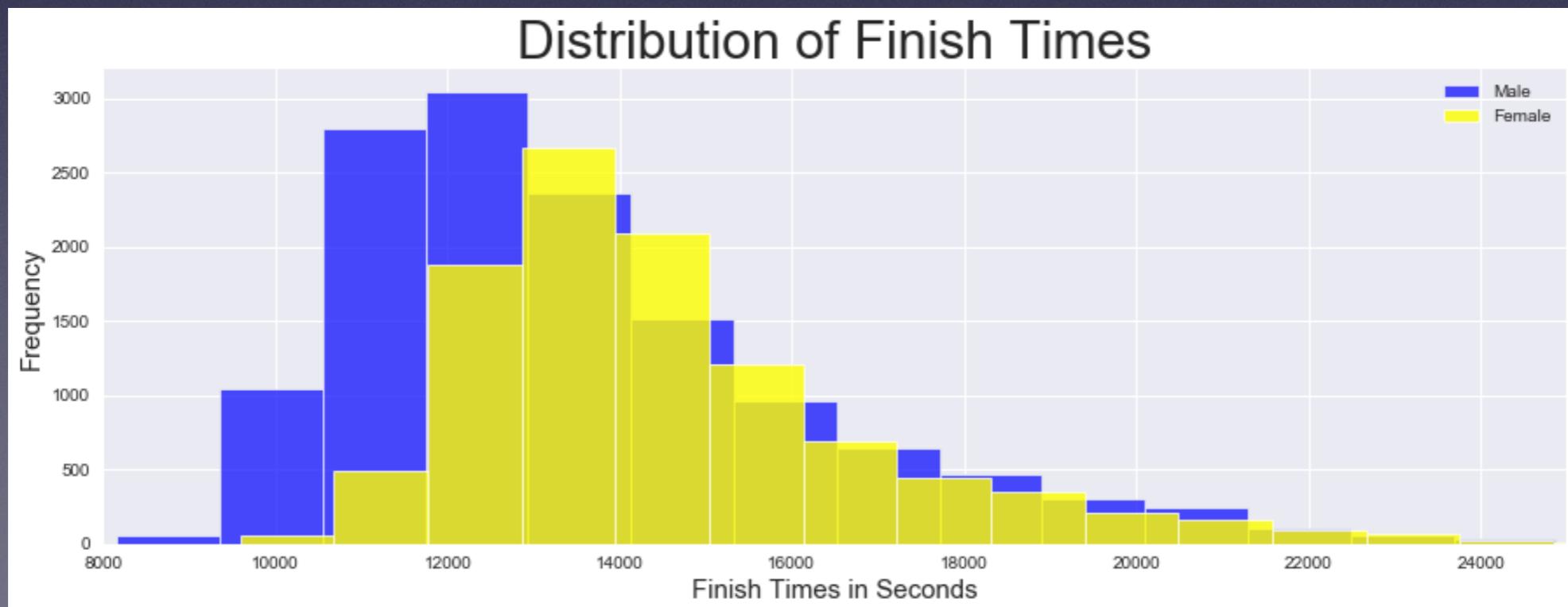
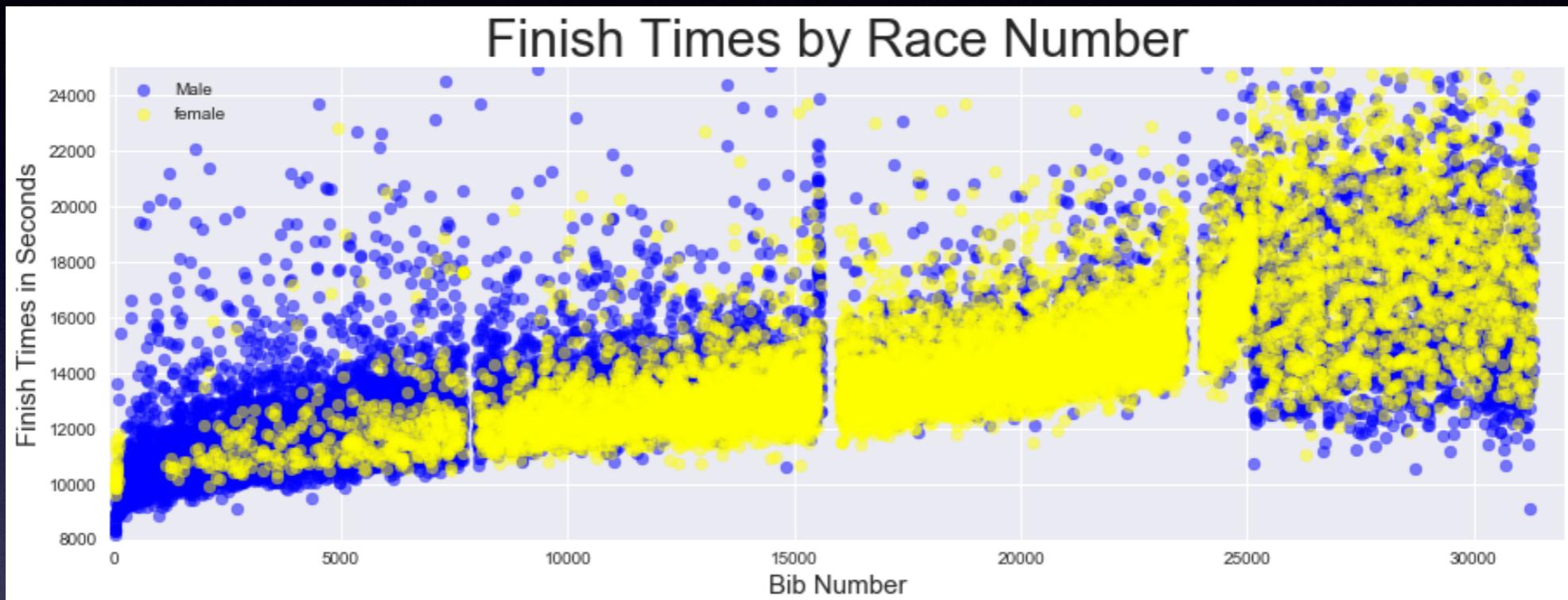
by Mailei Vargas

# Why Predict Race Times?

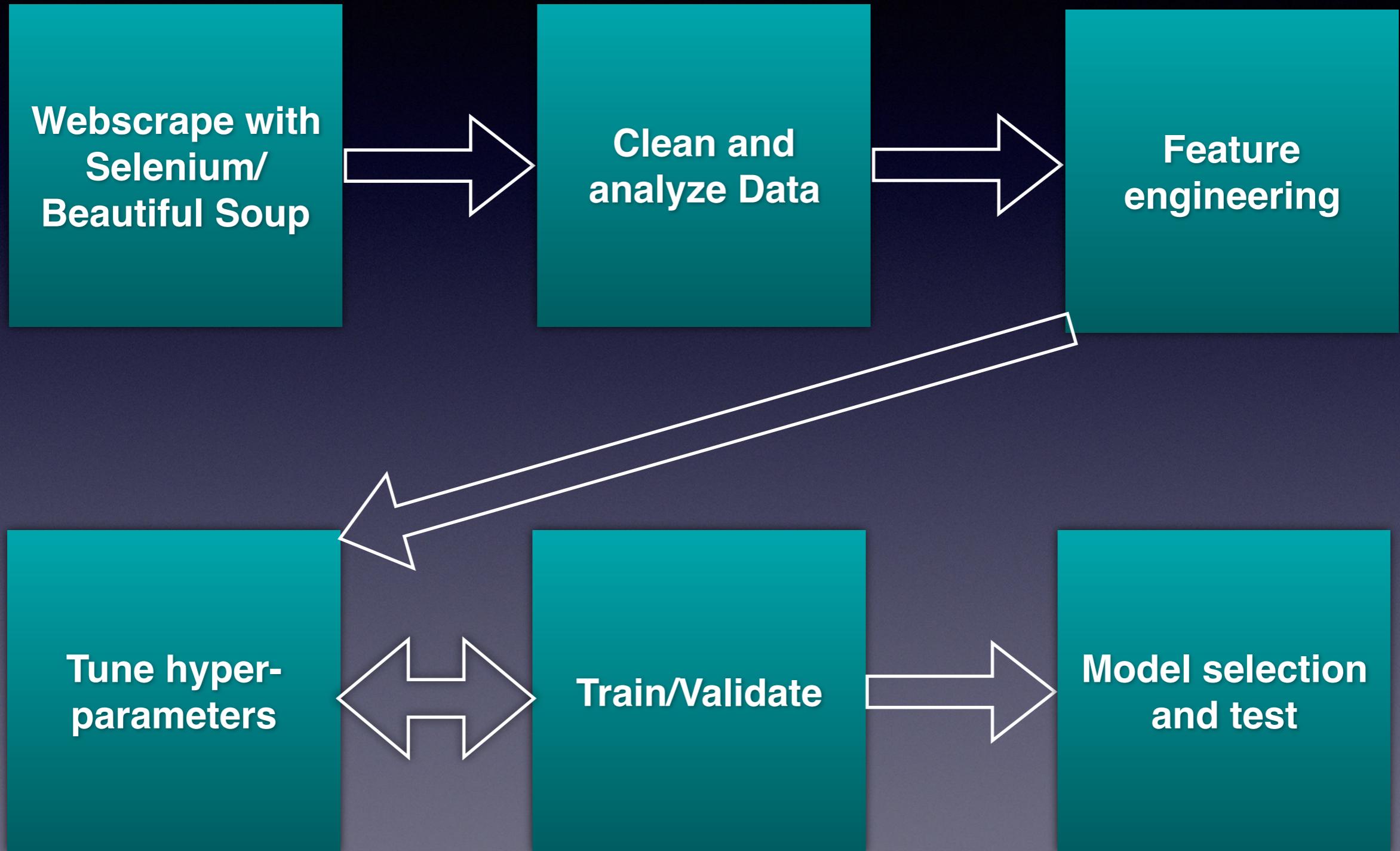
- Road running ~\$1.4 billion in 2015 -> millions run!
- Running improves physical/mental health
- ~20,000 Boston qualifiers, ~5000 charity runners
- \$830,500 in purse money for elites in 2018
- Can future years be predicted using prior results?

# 2015-2018 ~100k > ~13k legacy runners

BIB	NAME	AGE	M/F	CITY	ST	CTRY	CTZ
8341	De La Via, Claudia	32	F	Seattle	WA	USA	
	5k	10k	15k	20k	Half	25k	30k
	0:23:36	0:47:24	1:11:23	1:36:21	1:41:41	2:01:20	2:26:43
	Finish:	Pace		Proj. Time	Offl. Time	Overall	Gender
			0:07:56	-	3:27:54	6495	1245
							Division
							1016

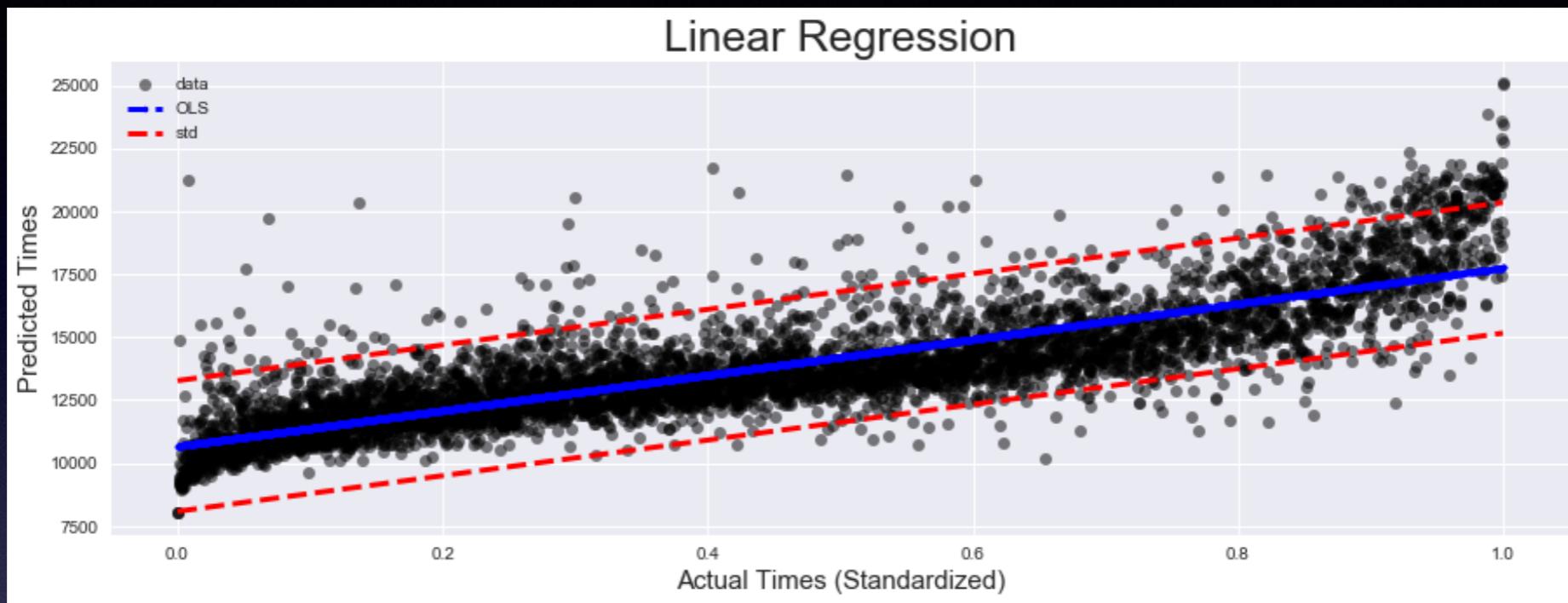


# Workflow and Tools

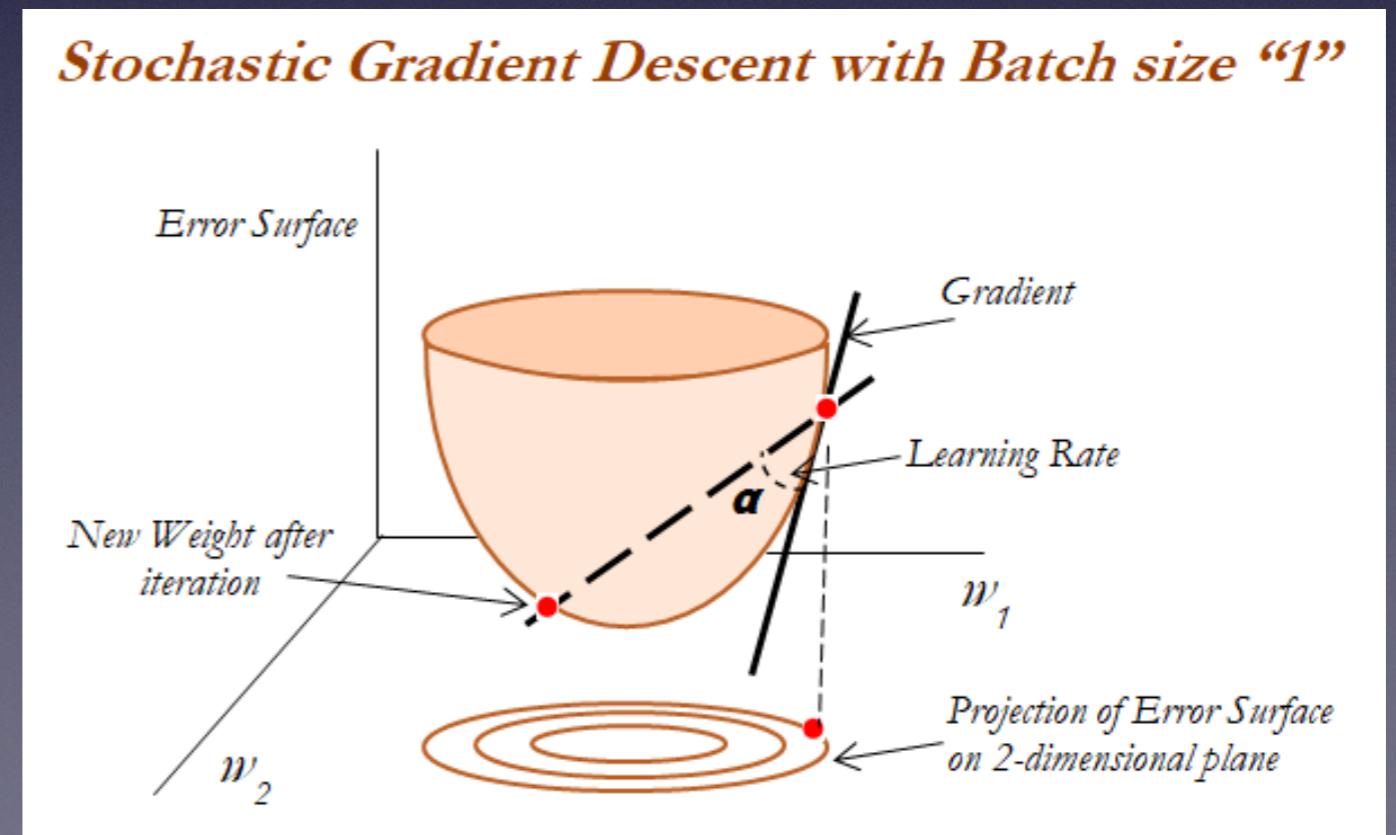


# Models

## Linear Regression



## Stochastic Gradient Descent



# Metrics and Results

Model	R <sup>2</sup>	RMSE	Minutes
Linear Regression	0.66	1426.35	23.77
<b>Linear Regression CV</b>	<b>0.69</b>	<b>1353.81</b>	<b>22.56</b>
SDGRegressor (No penalty)	-1.83x10 <sup>16</sup>	1.02x10 <sup>16</sup>	1.75x10 <sup>15</sup>
SDGRegressor (Limit X)	0.63	1486.79	24.78
SDGRegressor (Std)	0.65	1458.81	24.31
SDGRegressor (LASSO)	0.65	1463.54	24.39
SDGRegressor (RIDGE)	0.31	2035.24	33.92



# Conclusion

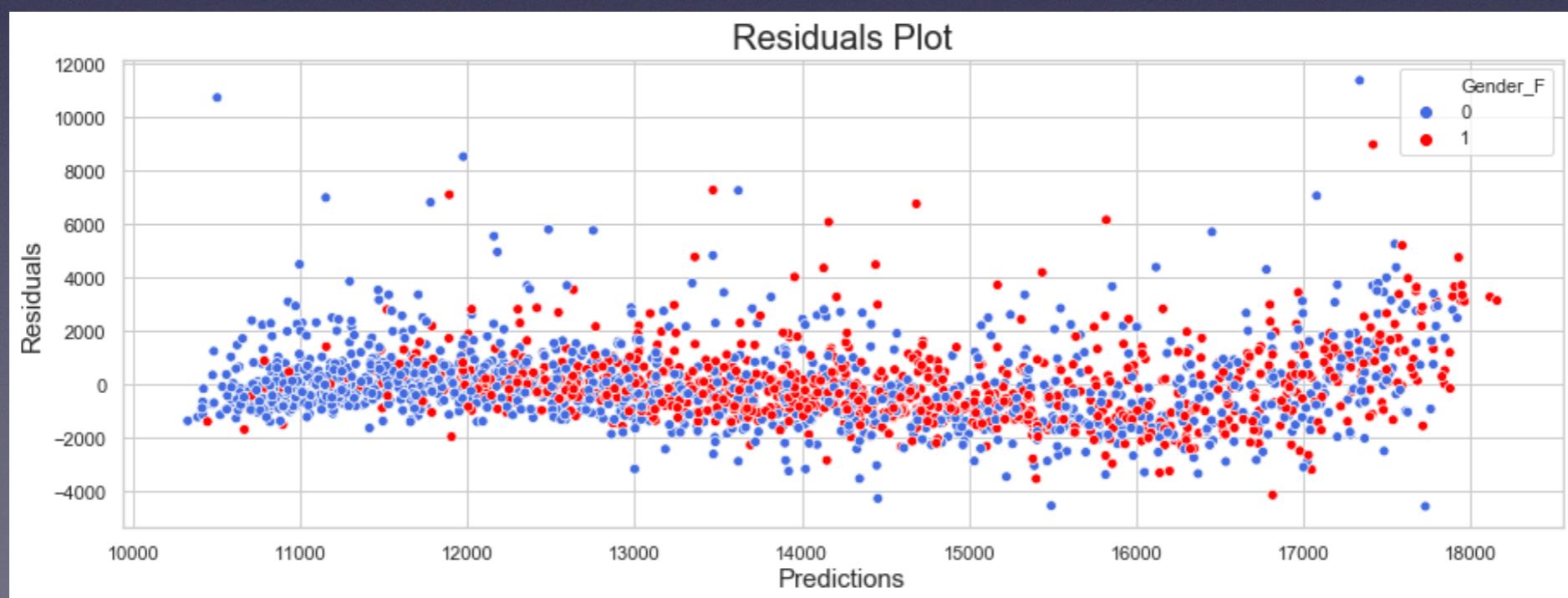


- Linear Regression CV, KFolds = 5
- How did Claudia do?
- 3:20:57 (prediction) -> 3:43:52 (actual)
- Interesting -> Outliers could be cheaters!
- Improvements: City data, interactive features, additional models.
- Questions?

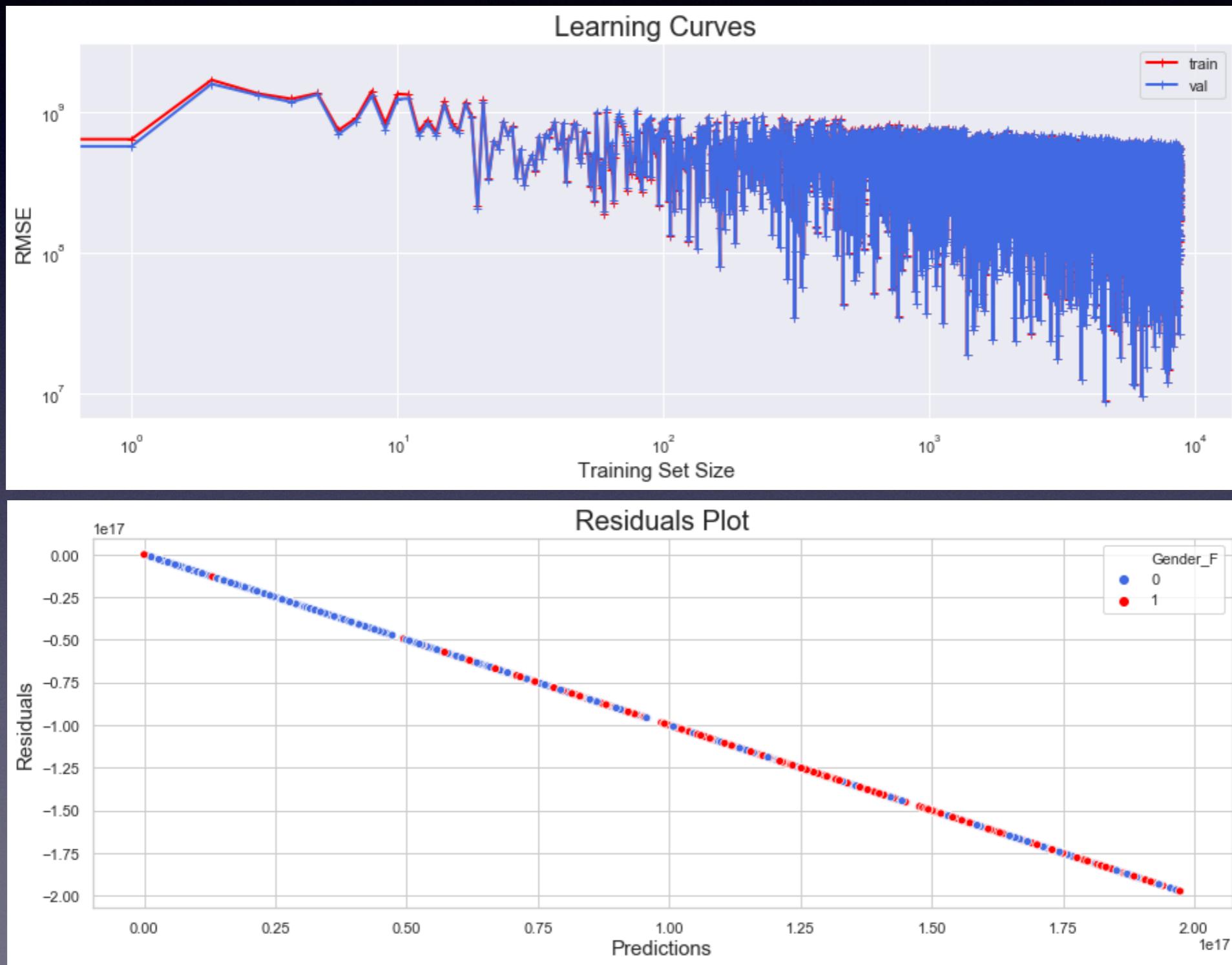
# Appendix



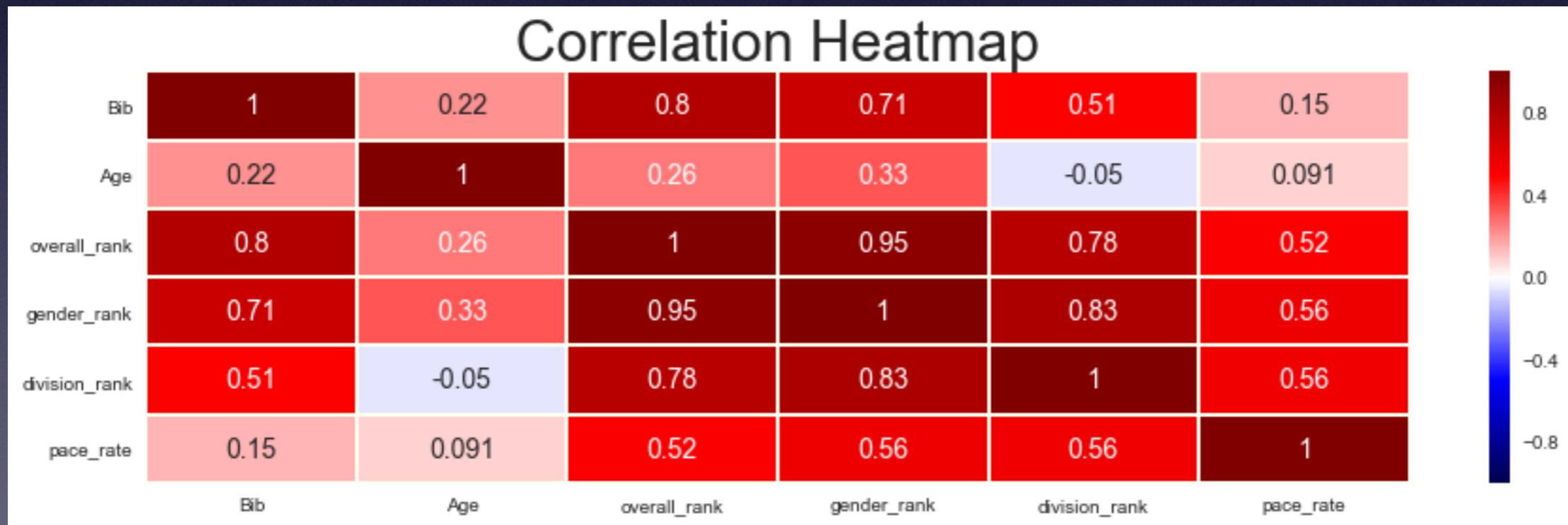
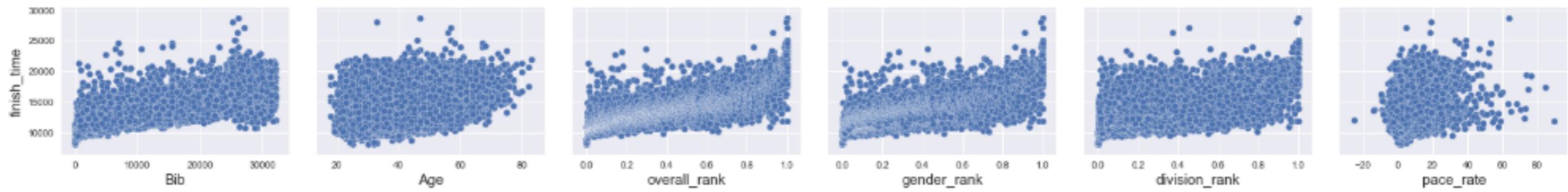
# Linear Model Learning Curve and Residual Plot



# SDGRegressor with No Convergence



# Correlation of a Few Predictors



# More on Metrics...

