PREDICTING EMERGENCY ROOM PAYER MIX

EMILY RINALDI

SPRINGBOARD FOUNDATIONS OF DATA SCIENCE WORKSHOP CAPSTONE

OBJECTIVE

- Determine whether demographic and economic characteristics of a particular zip code can predict the proportion of commercial patients at emergency department facilities located in that zip code
- The client is an emergency department staffing provider, whose revenue depends on the health care coverage of the patients they treat
- Predicting which facilities will see more patients with private health insurance and fewer patients with no insurance or governmental insurance will allow the client to focus its business development efforts on facilities with favorable payer mixes

DATA SOURCES

- Emergency Department Data by Ex

- Contains the distribution of emerge
hospitals years 2010-2014

2010-2014 American Community Survey

- Collects information such as age, race, income, commute time to work, home value, veteran status, and other important economics-related data

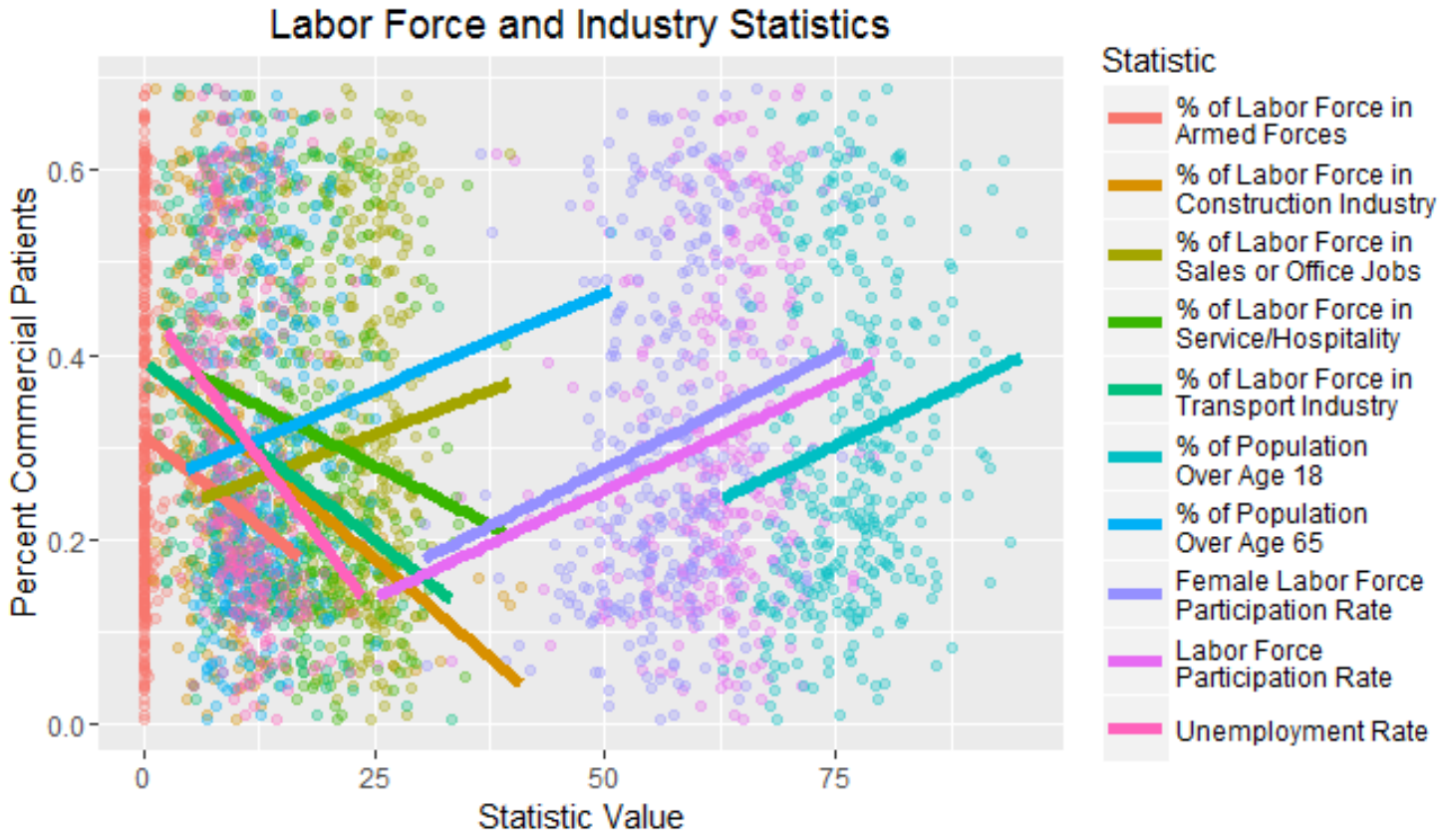
2010 Census Demographic Profile

- Contains data on population characteristics including sex, age, race, household relationship, household type, and housing characteristics including occupancy and tenure

We joined the data sources above into a single dataset with 346 observations, each representing one emergency room facility and the demographic and economic characteristics of the zip code in which it is located.

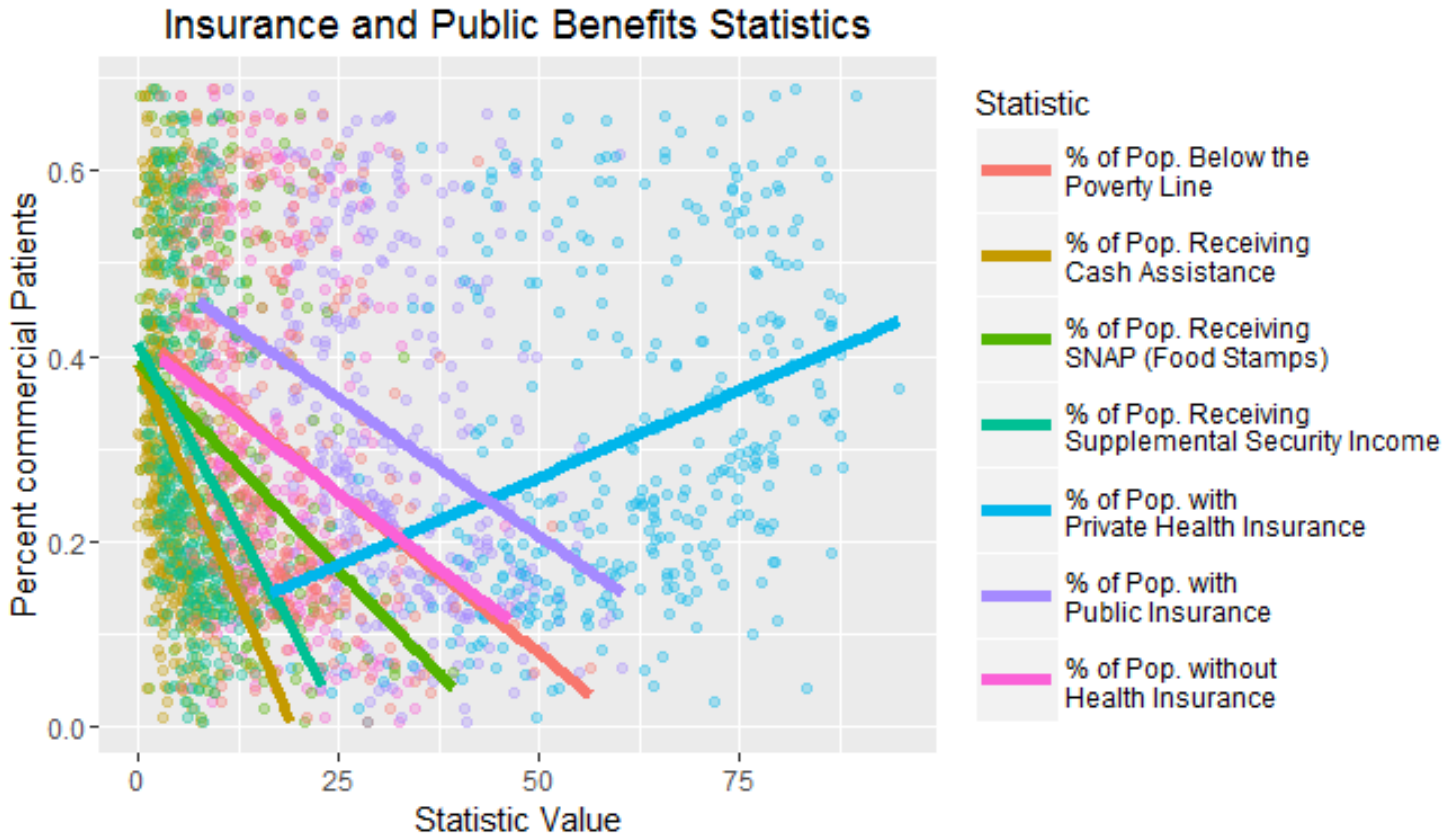
DATA EXPLORATION

- Higher labor force participation typically indicates more commercial patients
- More workers in sales or office jobs indicates more commercial patients, while armed forces, construction, transport, and service industries indicate fewer commercial patients
- Both more people over age 18 and over 65 correspond with more commercial patients



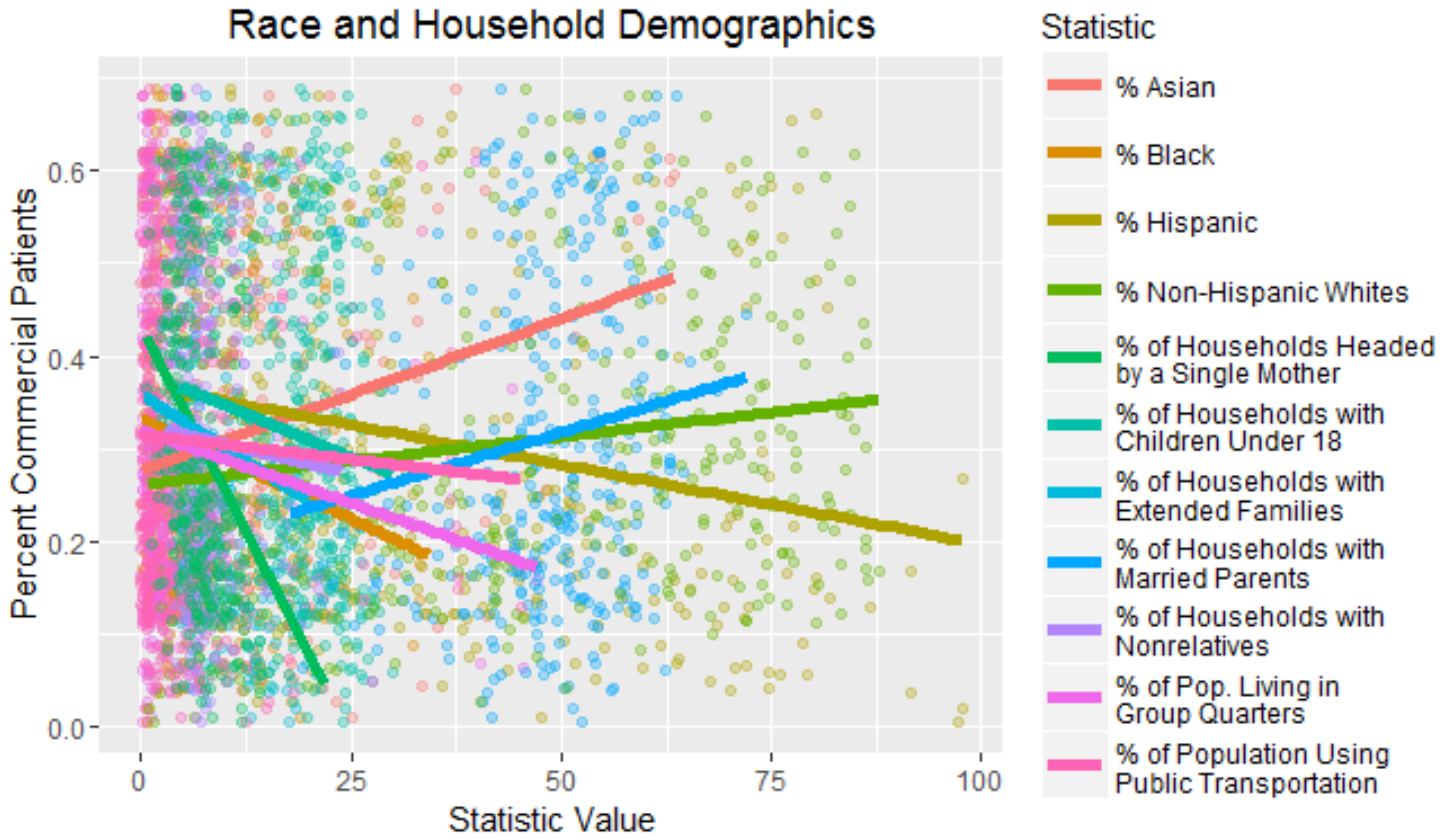
DATA EXPLORATION

- More people in a zip code receiving the following types of public assistance corresponds with fewer commercial patients
  - Cash Assistance
  - SNAP (Food Stamps)
  - Supplemental Security Income
  - Medicaid or Medicare
- Not surprisingly, more people in a zip code with private health insurance indicates more commercial insurance patients



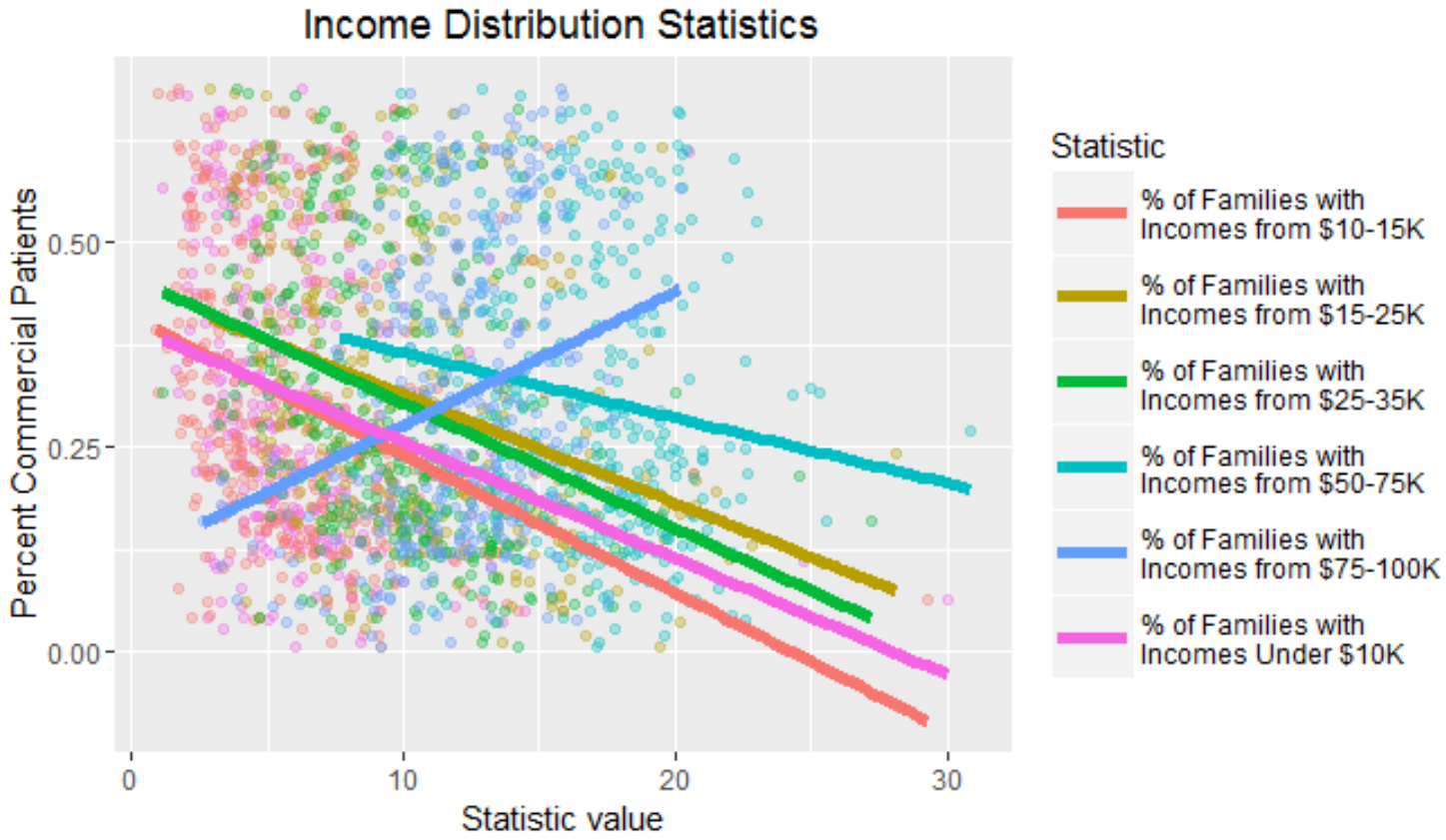
DATA EXPLORATION

- More commercial patients seen in areas with higher Asian and non-Hispanic white populations
- More households with married parents corresponds with more commercial patients
- More households headed by a single mother indicates fewer commercial patients at ERs



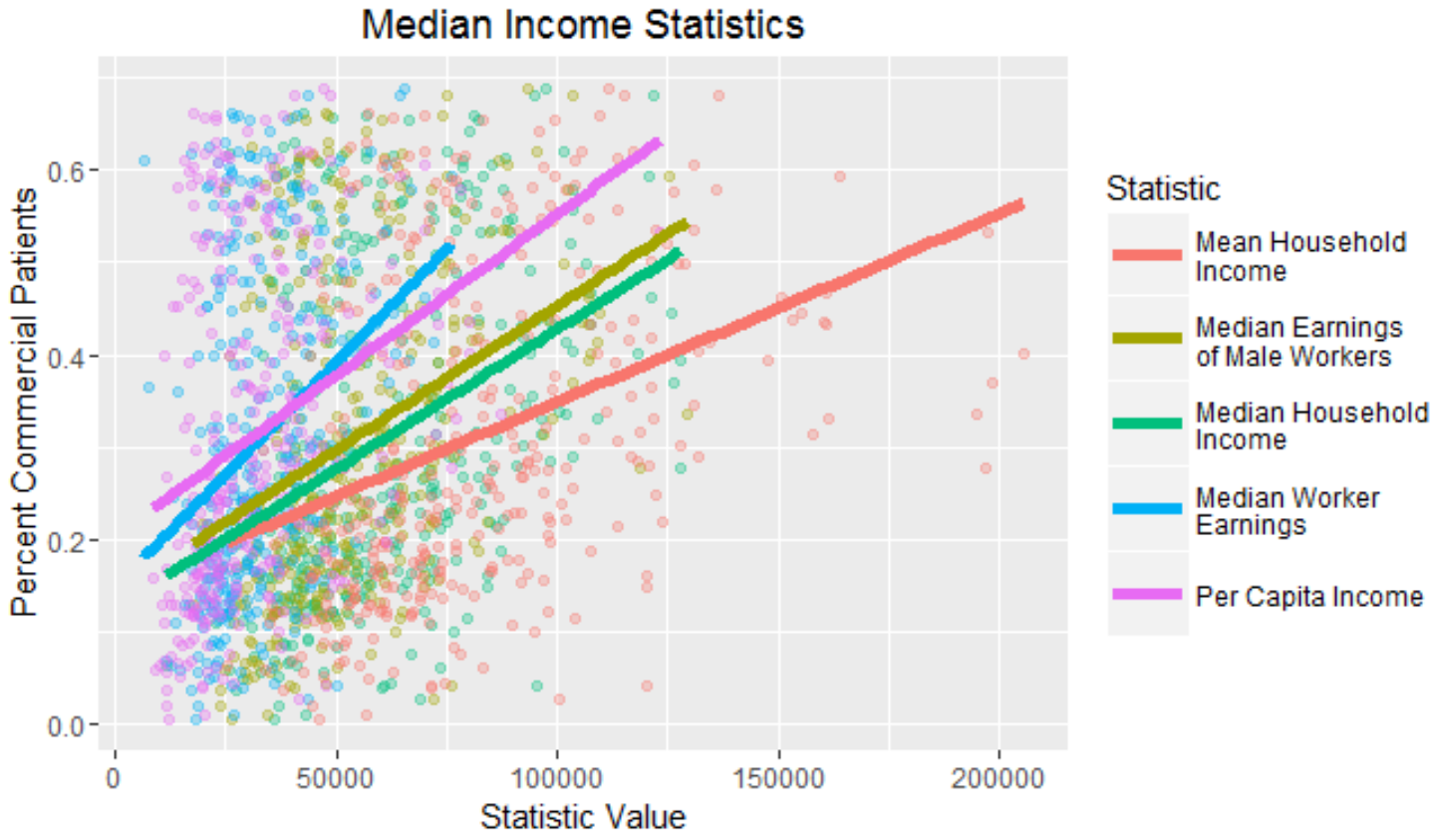
DATA EXPLORATION

- Of the incomes included, more people with incomes in any range below $75,000 indicates fewer commercial patients



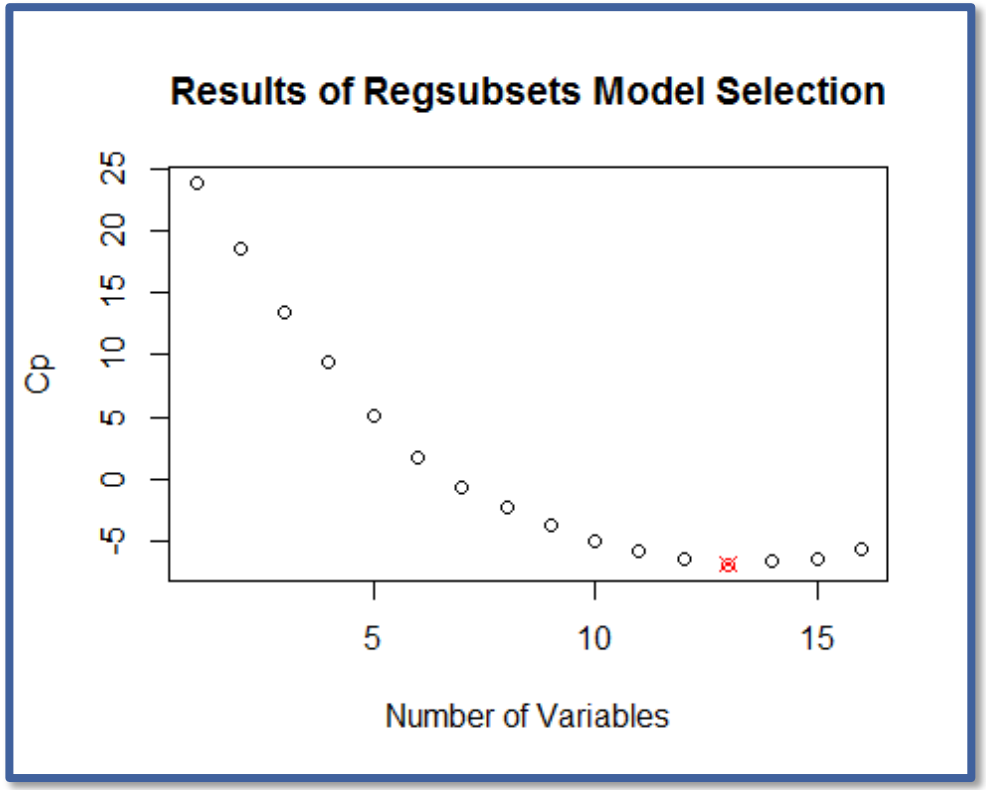
DATA EXPLORATION

- All income statistics are positively correlated with more commercial patients
- Higher wages likely means more people that can afford health benefits offered by their employers and few workers eligible for Medicaid



FEATURE SELECTION

To narrow down the variables in our dataset, we used the regsubsets function to perform an exhaustive search for the best model of various sizes and selected the variables that appeared in the model that minimized Mallow's Cp



Variables Selected by regsubsets Search:

- Facility Ownership: investor, nonprofit, or public
- % of Population in Service Industry
- % of Population in Sales or Office Jobs
- % of Workers Earning $75-100K
- Median Earnings of Male Workers
- % with Public Insurance
- Total Population
- % Black
- % Asian
- % Hispanic
- % Non-Hispanic White
- % of Houses that are Vacant

MODEL TRAINING RESULTS

Model ID	Method	Independent Variables	Training Formula	Cross-validated R2	Test Set R2
Fit1	**Linear Regression**	**All 13 variables**	**pct_comm ~ .**	**.226**	**.259**
Fit2	Linear Regression	Significant variables from Fit1	pct_comm ~ ownerPublic + pct_service_ind + pct_75to100K + med_male_earnings + pct_asian	.217	.159
FitGBM1	Stochastic Gradient Boosting	All 13 variables	pct_comm ~ .	.214	.017
FitGBM2	Stochastic Gradient Boosting	Top 5 variables from FitGBM1 by varImp	pct_comm ~ med_male_earnings + pct_public_ins + pct_black + pct_service_ind + pct_hisp	.227	.103

BEST PREDICTIVE MODEL – LINEAR REGRESSION WITH ALL PREDICTORS

Linear Model Coefficient Summary:

	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	-0.0689347	0.1603153	-0.4299944	0.6676006	
ownerNonprofit	-0.0478559	0.0272140	-1.7585038	0.0799860	.
ownerPublic	-0.0795775	0.0363282	-2.1905143	0.0294871	*
pct_service_ind	0.0056271	0.0024576	2.2896805	0.0229426	*
pct_sales_office	0.0023551	0.0027397	0.8596160	0.3908920	
pct_75to100K	0.0105072	0.0043663	2.4064234	0.0168961	*
med_male_earnings	0.0000036	0.0000010	3.7537581	0.0002204	***
pct_public_ins	-0.0019233	0.0016176	-1.1889821	0.2356675	
tot_pop	-0.0000012	0.0000006	-1.9249413	0.0554648	.
pct_black	-0.0018589	0.0018368	-1.0120660	0.3125654	
pct_asian	0.0020633	0.0010383	1.9871357	0.0480887	*
pct_hisp	0.0003600	0.0007148	0.5036645	0.6149771	
pct_vacant_houses	0.0016752	0.0013119	1.2769683	0.2028955	

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- According to the model, facilities with the most commercial patients are likely privately owned or nonprofit facilities located in zip codes with
  - higher wages
  - more Asian/Asian-American residents
  - a large service and hospitality industry

MODEL LIMITATIONS

- Small sample size and highly variable data
  - Final dataset of only 346 observations
  - Very noisy data - model only explains fraction of variance
- Limited scope of dataset, which includes only California emergency departments
  - California is unique –early adoption of Medicaid expansion, high Asian/Asian-American population, travel and entertainment industry, etc.
  - May not be able to generalize model to the entire country
- Major assumptions
  - The population residing in a facility's zip code may not be best representation of its patient base
  - Patients with non-emergent issues may self-select the facility that is best for their insurance or that has a generous charity policy rather than visiting the one closest to their home

RECOMMENDATIONS

- Purchase or create a new dataset from internal data that is more robust and representative of a larger number of facilities in various geographic areas
  - Test the current model's predictive power on the larger dataset
  - Train a new predictive model if necessary
- Consider other possible explanations for variation in commercial patient proportion
  - Medicaid expansion
  - Patient satisfaction ratings of facilities
- Study the validity of the assumption that the population residing in a zip code is an accurate representation of the patients at a facility in that zip code

ACKNOWLEDGEMENTS

Thanks to Chris Bishop, my mentor, for all of the guidance, advice, and feedback throughout this project!