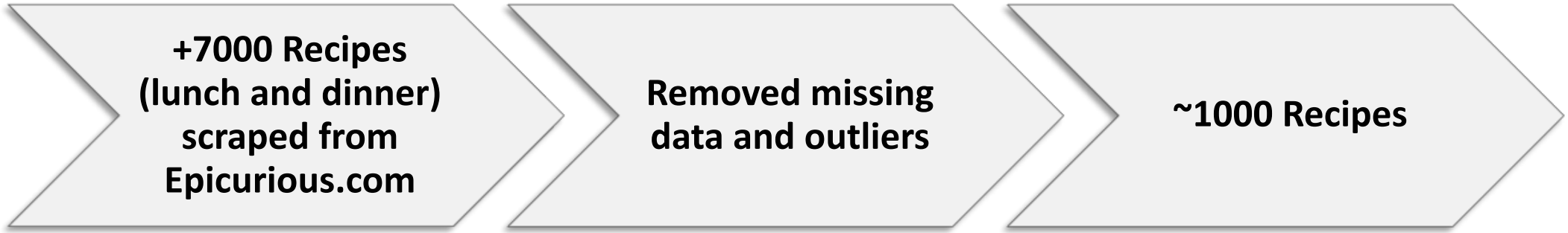


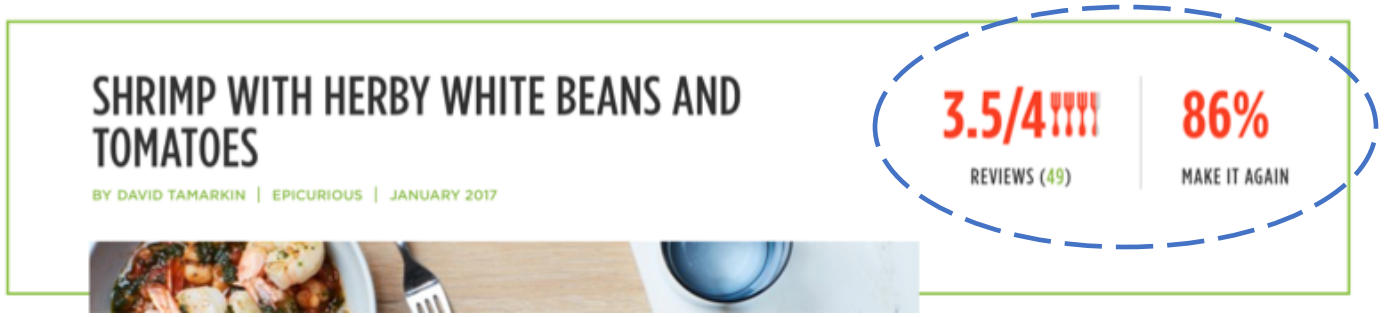
Predicting the Success of a Recipe

What can we conclude from the ratings in Epicurious.com?

Web Scraping and Data Cleaning



Choosing the Target Variable



Success Score = Rating * ”Make it again”

Web Scraping and Data Cleaning

Independent Variables

Nutrition Data

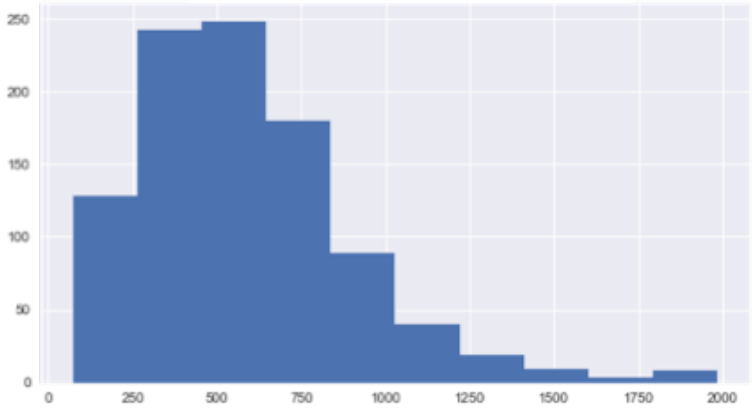
- **Calories**
- **Protein**
- **Cholesterol**
- **Etc.**

Ingredients (dummy)

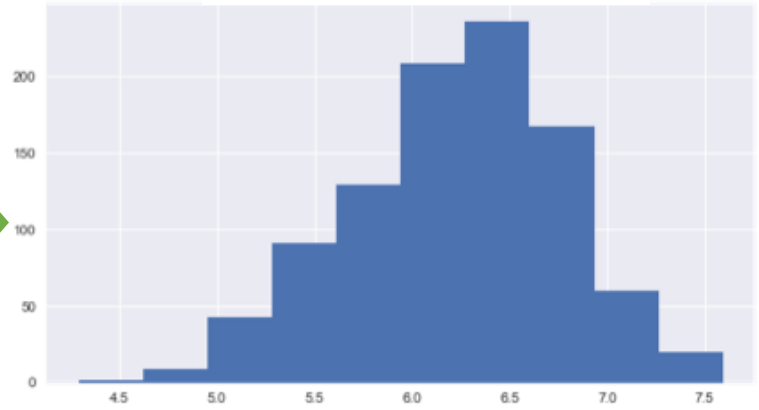
Data Wrangling

- **Log transformation in Nutrition Data**

Calories distribution



Log Calories distribution



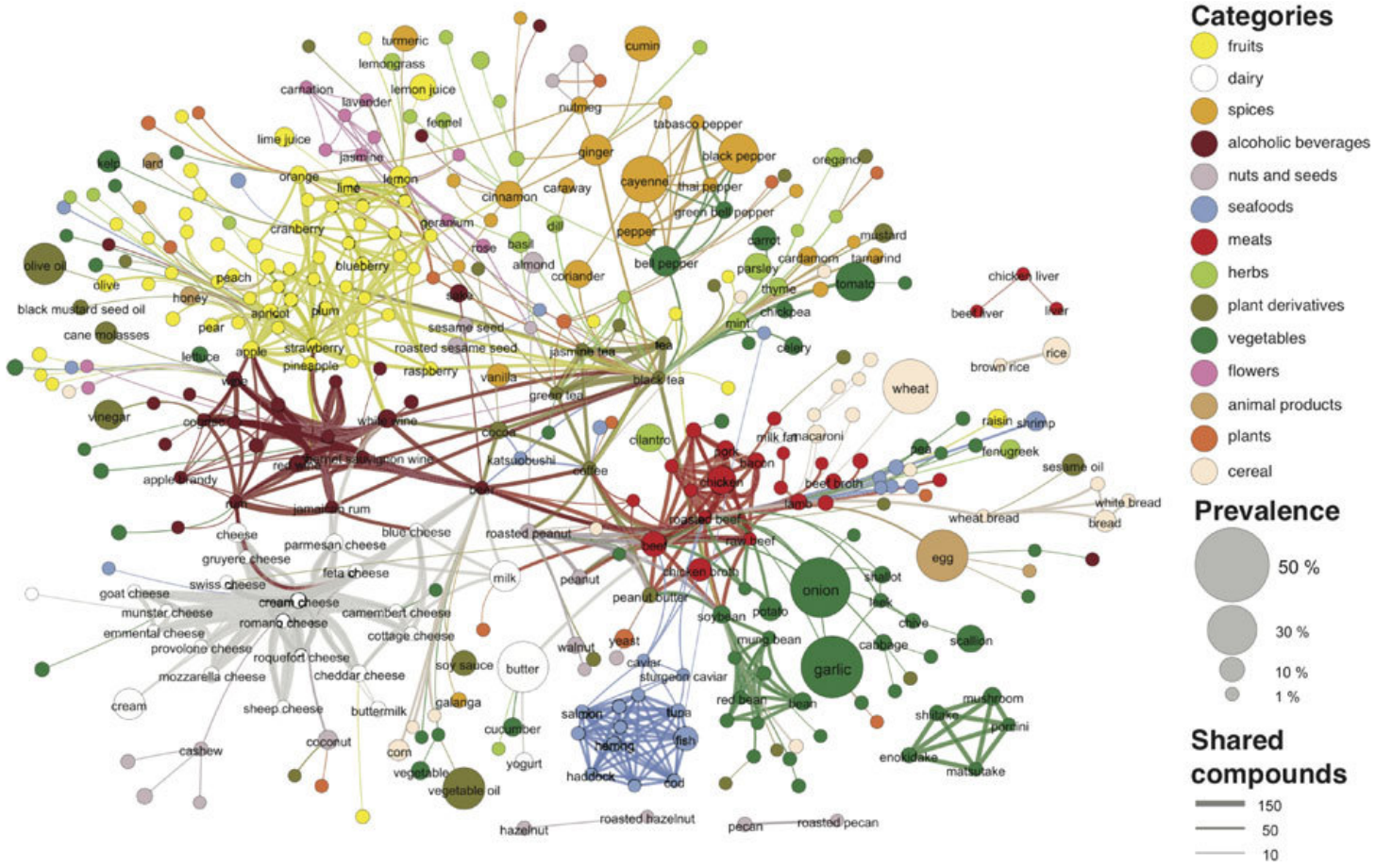
- **Feature Selection for choosing Ingredients**
  - Recursive Feature Elimination (sklearn.feature_selection)
    - Ranks the predictors based on their contribution to the model.
  - Selected 50 most important features

Preliminary Analysis

index	success_score	rating	calories	carbohydrates	protein	cholesterol	sodium
Top Recipes	4.0000	4.00	703.30	44.96	39.31	149.61	978.71
Bottom Recipes	1.4108	2.64	607.45	53.51	26.64	116.66	712.30
Comparison	2.8400	1.52	1.16	0.84	1.48	1.28	1.37

On average, the 100 best rated recipes have more calories, protein, cholesterol and sodium, and less carbohydrates than the 100 worse rated recipes.

Preliminary Analysis



Source: Article "What A Global Flavor Map Can Tell Us About How We Pair Foods (Nature.com)

The success of a recipe is a result of the combination of flavors and preparation methods

Exploratory Data Analysis
Linear Regression

- OLS Regression on training set (70% of data) beginning with the selected features.

OLS Regression Results

Dep. Variable:	success_score	**R-squared:**	0.950
Model:	OLS	**Adj. R-squared:**	0.946
Method:	Least Squares	**F-statistic:**	216.8
Date:	Thu, 01 Feb 2018	**Prob (F-statistic):**	0.00
Time:	18:44:32	**Log-Likelihood:**	-693.53
No. Observations:	676	**AIC:**	1497.
Df Residuals:	621	**BIC:**	1745.
Df Model:	55		
Covariance Type:	nonrobust		

MSE: 0.589

Removed features when p-value high

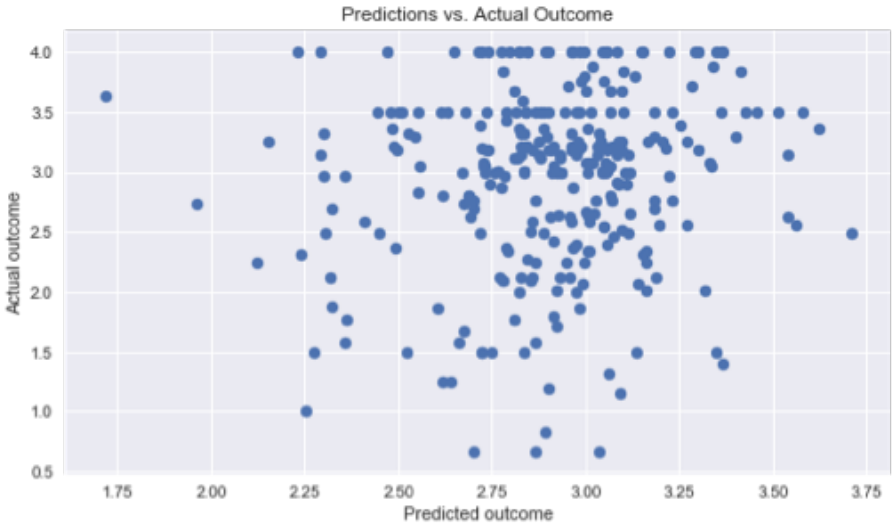
OLS Regression Results

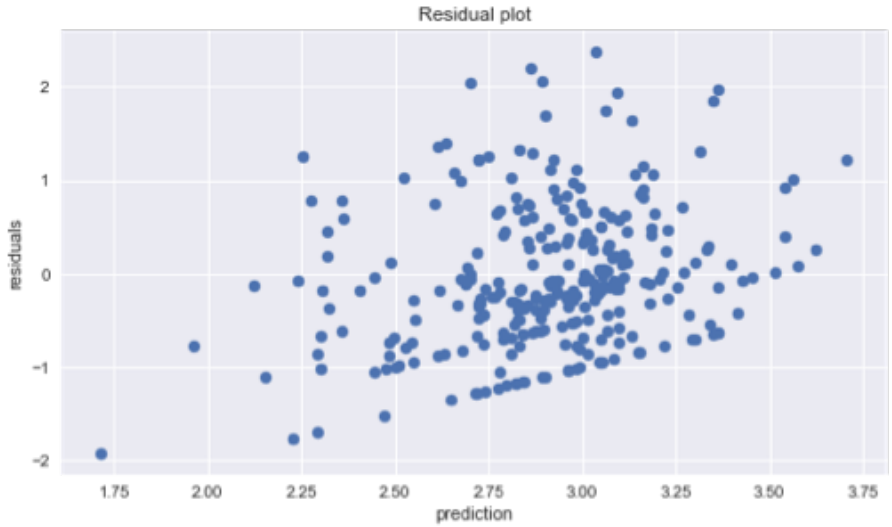
Dep. Variable:	success_score	**R-squared:**	0.951
Model:	OLS	**Adj. R-squared:**	0.949
Method:	Least Squares	**F-statistic:**	406.0
Date:	Thu, 01 Feb 2018	**Prob (F-statistic):**	0.00
Time:	18:47:23	**Log-Likelihood:**	-685.96
No. Observations:	676	**AIC:**	1434.
Df Residuals:	645	**BIC:**	1574.
Df Model:	31		
Covariance Type:	nonrobust		

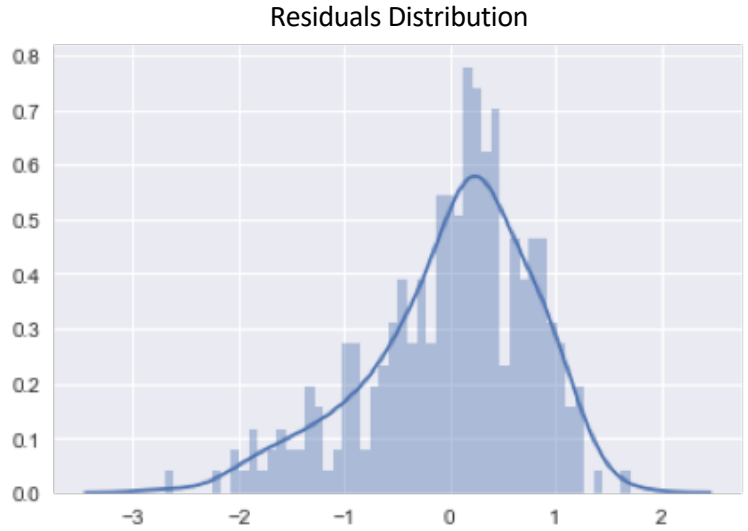
MSE: 0.576

Exploratory Data Analysis

Results



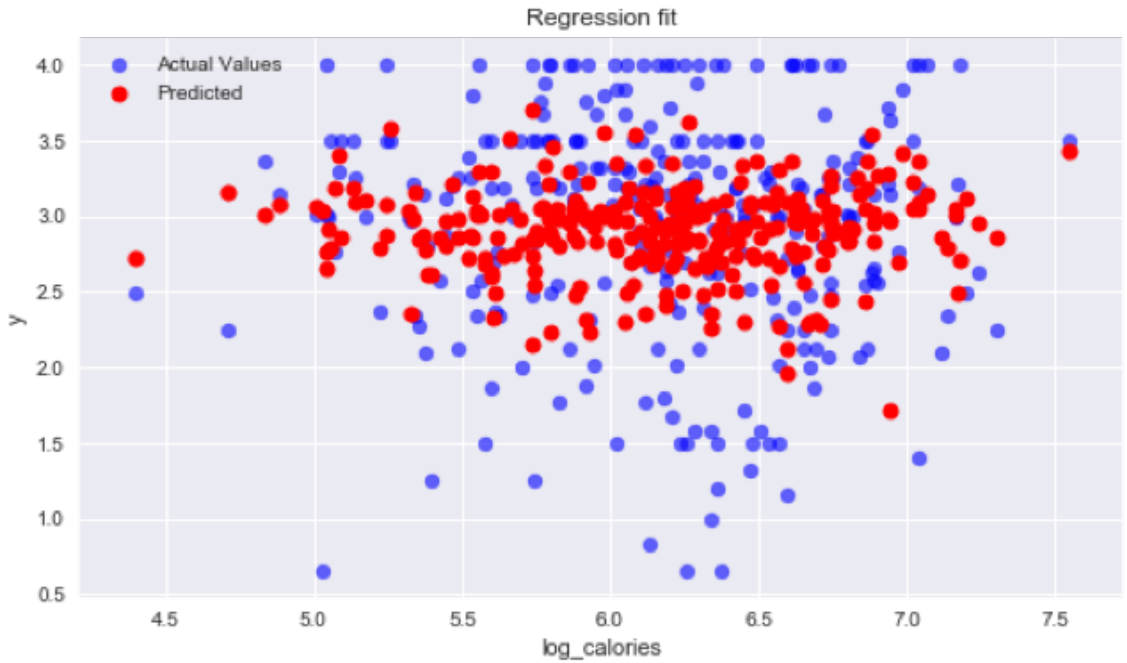


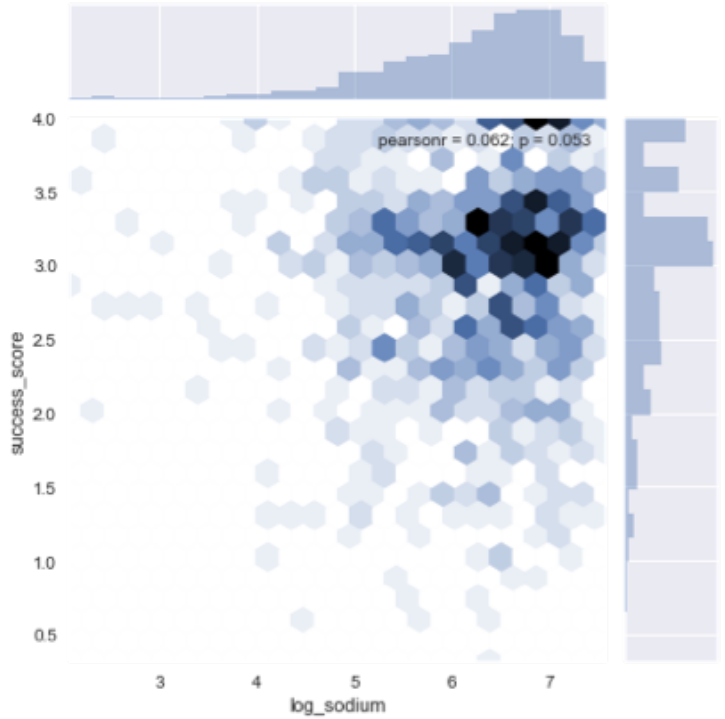


Exploratory Data Analysis
Results

	coef	std err	t	P>|t|	[0.025	0.975]
log_calories	0.5571	0.051	10.922	0.000	0.457	0.657

1% increase in log_calories, the success score increases 0.56





Exploratory Data Analysis

Features selected are mostly not main ingredients, but flavor additions to the recipes

- Prune
- Grains
- Ground Lamb
- Nutmeg
- Pear
- Orange Juice
- Okra
- Trout
- log_calories
- log_carbohydrates
- log_protein
- log_sodium
- Shallot
- Pine Nut
- Goat Cheese
- Coconut
- Oat
- Celery
- Cranberry
- Monterey Jack
- Pistachio
- Pineapple
- Lemon
- Cheddar
- Chile Pepper
- Mint
- Squash
- Walnut
- Feta
- Lentil
- Breadcrumbs

Conclusions

- People prefer more caloric and salty meal
- Main ingredients can not be individually analyzed in a recipe
- “Fancy” ingredients may increase (or decrease) the rating

Next Steps

- Use a classification method
- Extract proportions of ingredients
- Use NLP for analyzing the reviews