Post Score in US State and City Subreddits

Kemengtian Ma

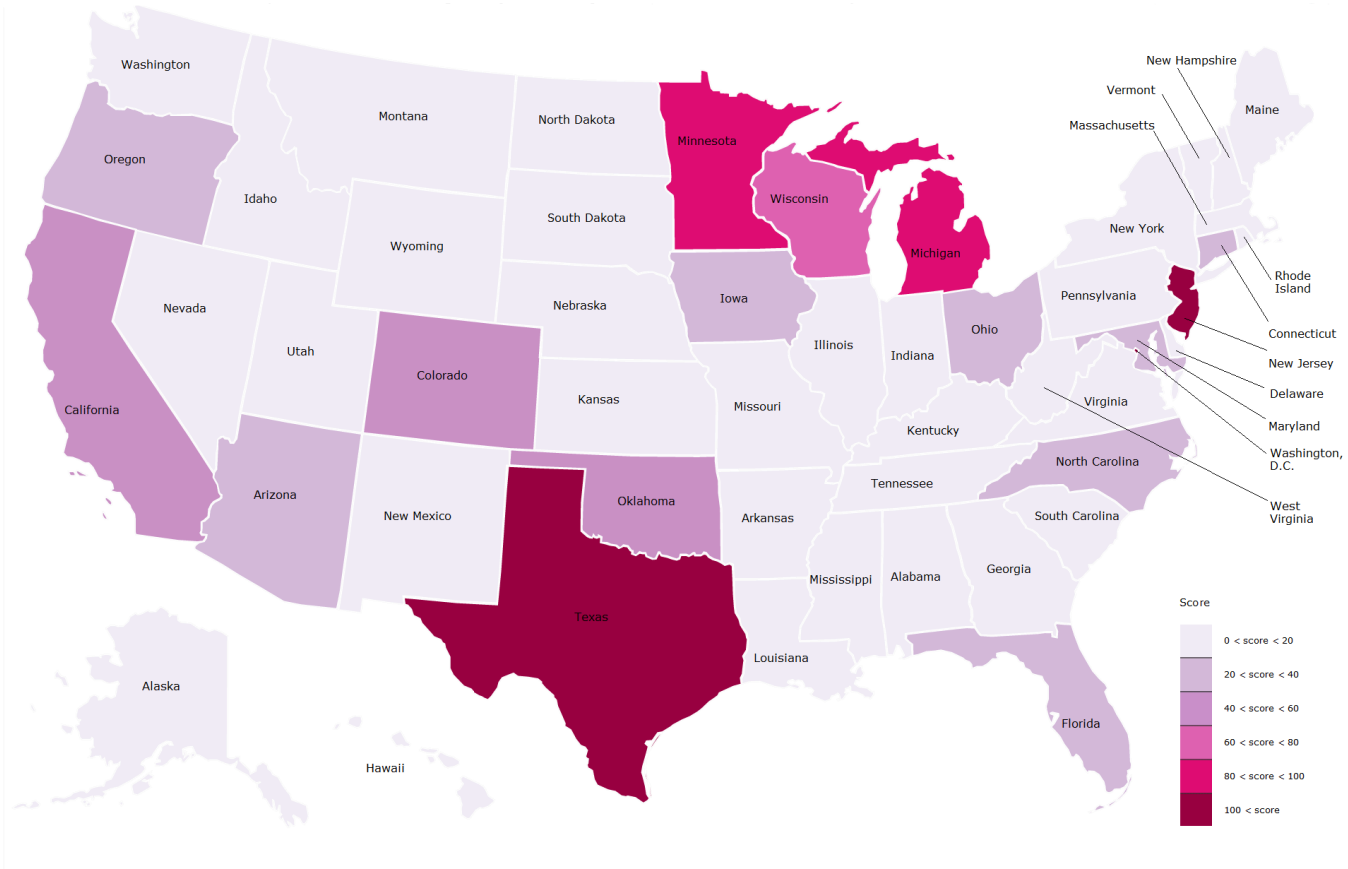
Stat 222 Capstone Project

Two main parts

- The comparisons of post scores among US states and cities.
  - Compare states
  - Compare cities
  - Compare states vs. cities

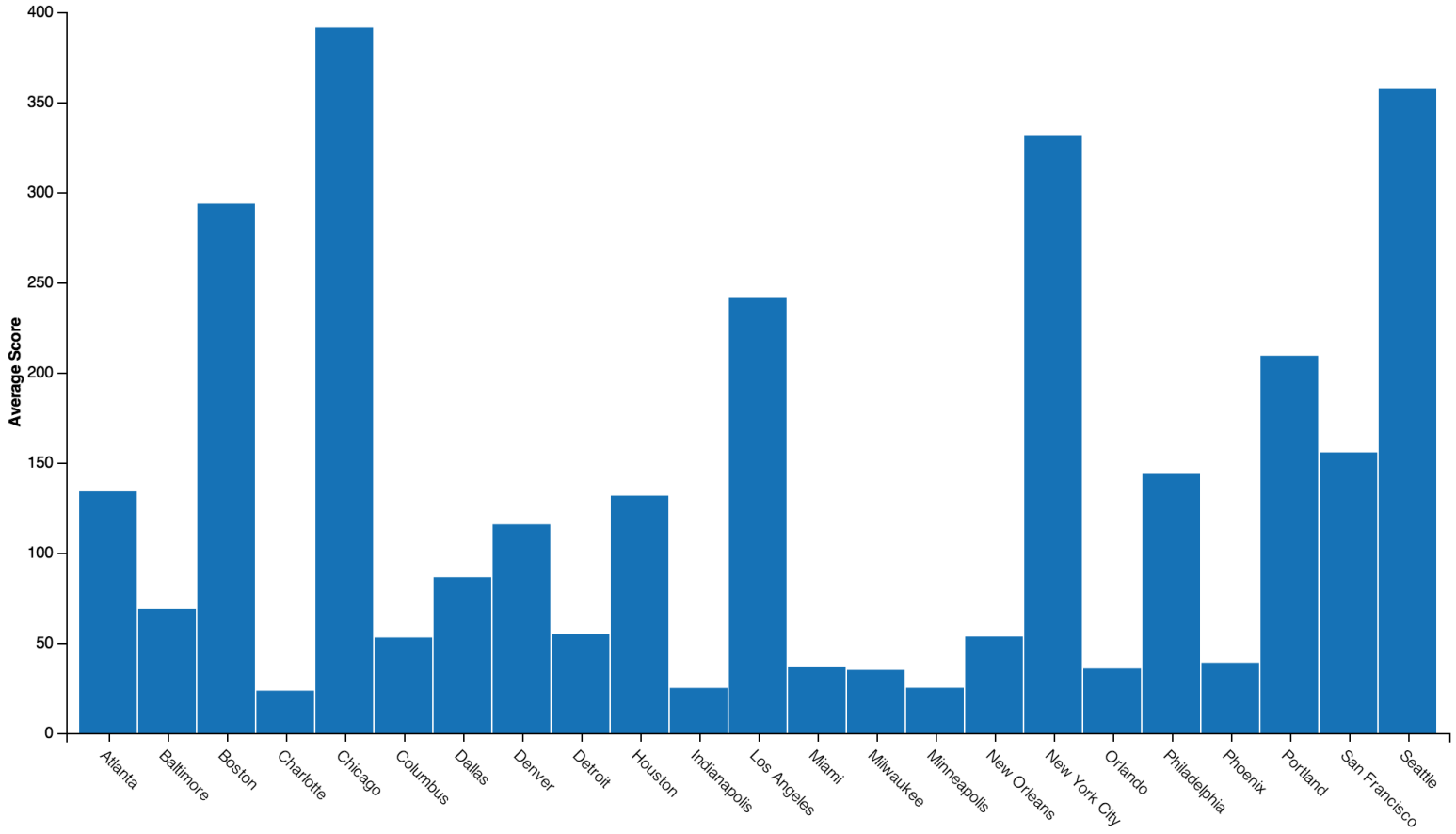
- Factors that may affect the post scores.
  - Number of comments
  - Title of the post

US States in Top Subreddits

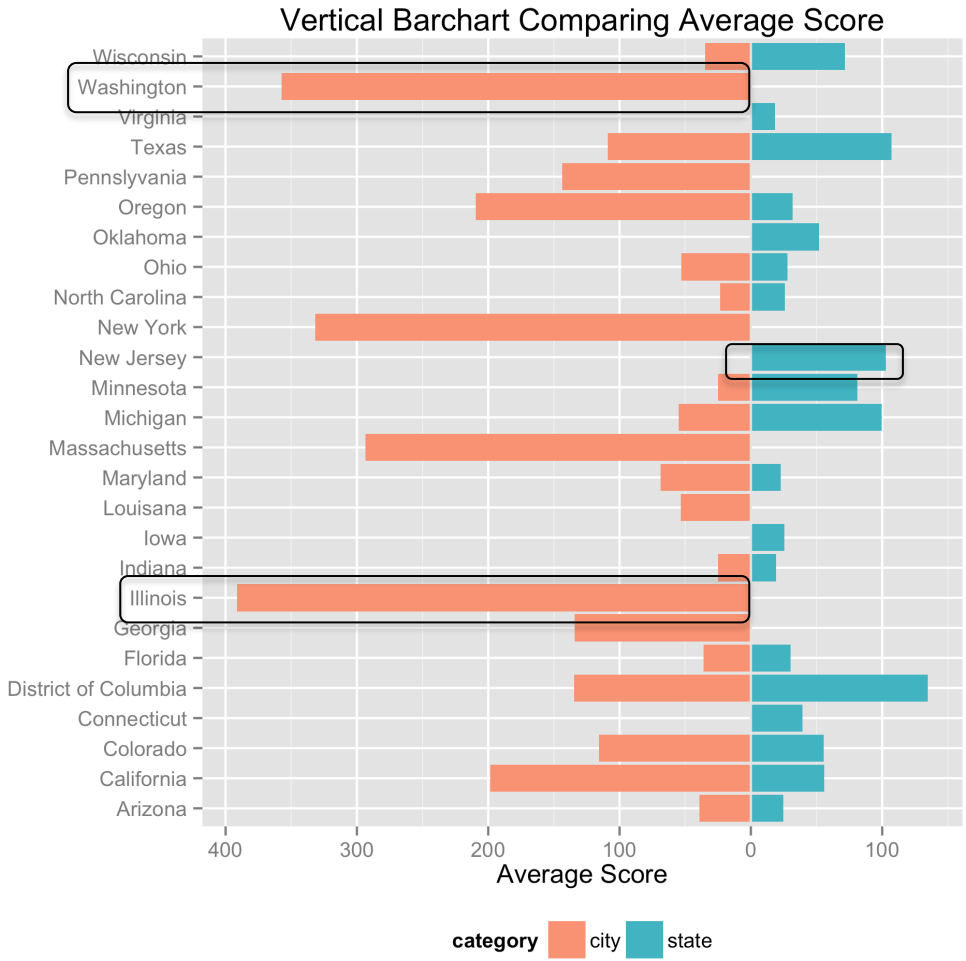


- Shaded by the different levels of average scores
- Score 0 means not in the top subreddits
- Texas and New Jersey
- States around lake regions

Average Scores of US Cities in Top Subreddits



Comparing Average Scores among States and Cities



- city scores > state scores
- Some states DO NOT have a state subreddit in the top but DO have a city subreddit to represent
- States with no large city do not have a city subreddit in the top

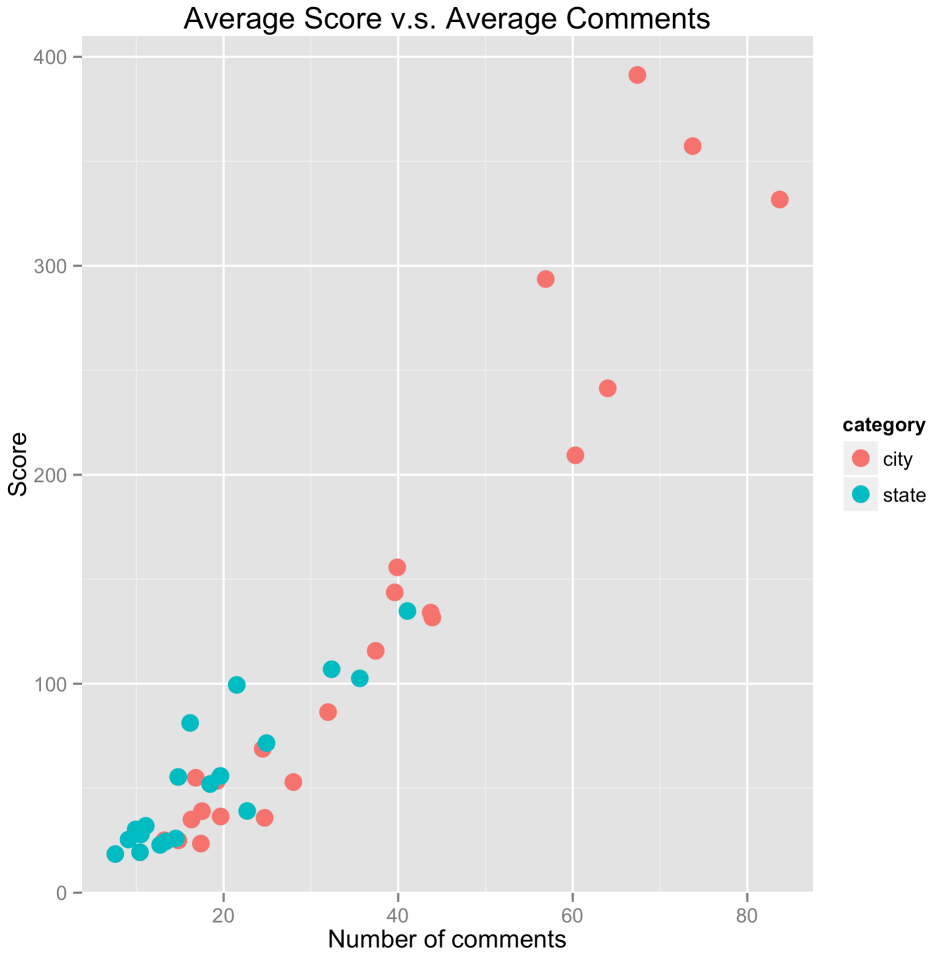
One Tail Hypothesis Test

```
        Welch Two Sample t-test

data:  cityscore and statescore
t = 2.8365, df = 30.263, p-value = 0.00403
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 27.93472      Inf
sample estimates:
mean of x mean of y
108.96744  39.44494
```

- Ho: average post score of city is greater than the average post score of corresponding state
- Small P-value = 0.00403
- Reject null hypothesis
- There is significant difference between these two average scores

Scatterplot



- Strong positive linear pattern
- High score posts usually have large number of comments
- Two categories
  city has a wider spread in both score and number of comment than state
- Try simple linear regression to estimate the relationship

Simple Linear Regression

OLS Regression Results

Dep. Variable:	score	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.900
Method:	Least Squares	F-statistic:	370.6
Date:	Mon, 17 Mar 2014	Prob (F-statistic):	7.81e-22
Time:	16:31:18	Log-Likelihood:	-202.27
No. Observations:	42	AIC:	408.5
Df Residuals:	40	BIC:	412.0
Df Model:	1		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	-39.5323	8.515	-4.643	0.000	-56.742	-22.322
num_comments	4.7689	0.248	19.250	0.000	4.268	5.270

Omnibus:	16.638	Durbin-Watson:	2.269
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21.719
Skew:	1.239	Prob(JB):	1.92e-05
Kurtosis:	5.503	Cond. No.	62.0

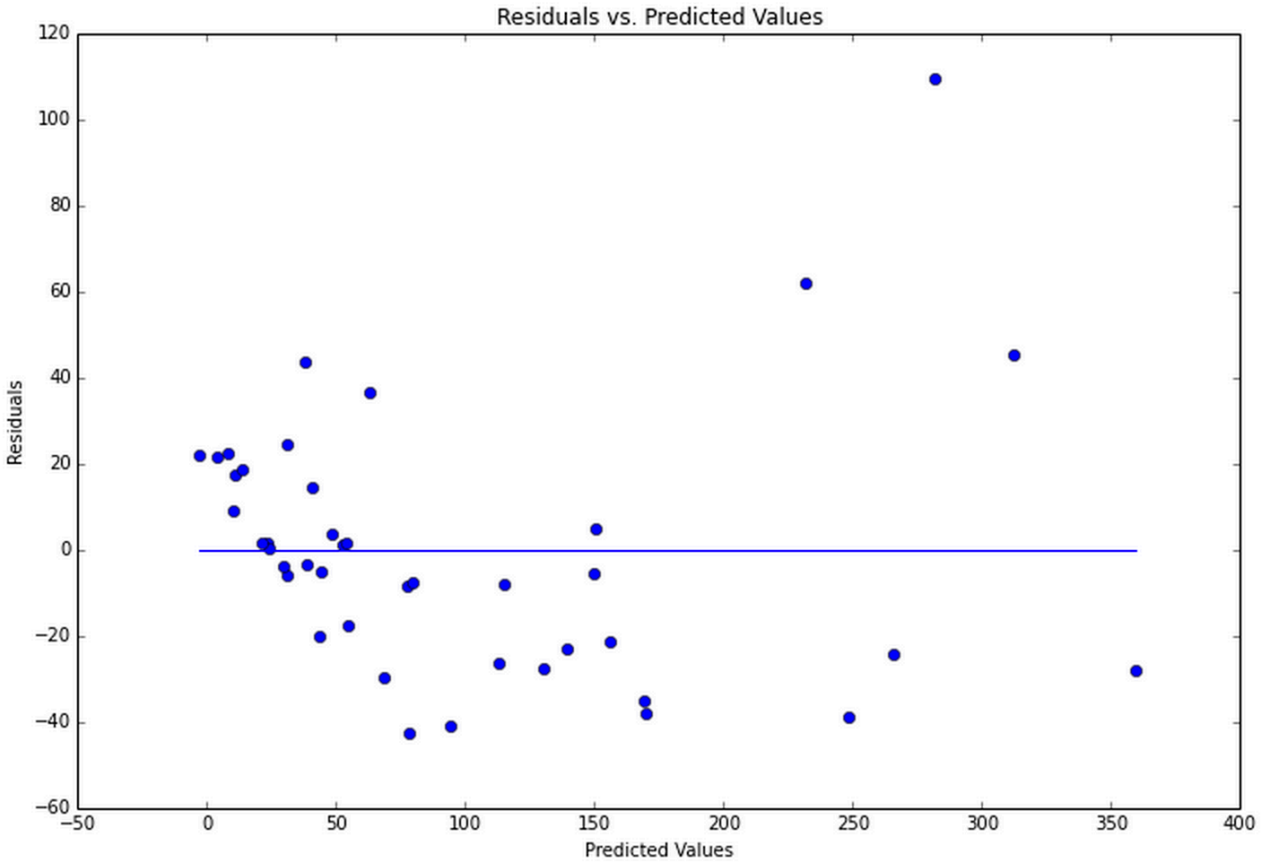
- Model:

  average post score = -39.5323 + 4.7689*average number of comments

- High R squared = 0.903

- P-values of intercept and slope are small

Residual vs. Fitted Plot



- Curved pattern. Do not look well
- Residuals are not scattered randomly around the zero line
- Normality of errors is violated
- There maybe non-linear relationships between the two variables

Quadratic Regression

OLS Regression Results

Dep. Variable:	score	R-squared:	0.915
Model:	OLS	Adj. R-squared:	0.911
Method:	Least Squares	F-statistic:	210.8
Date:	Sun, 27 Apr 2014	Prob (F-statistic):	1.24e-21
Time:	18:54:43	Log-Likelihood:	-199.33
No. Observations:	42	AIC:	404.7
Df Residuals:	39	BIC:	409.9
Df Model:	2		

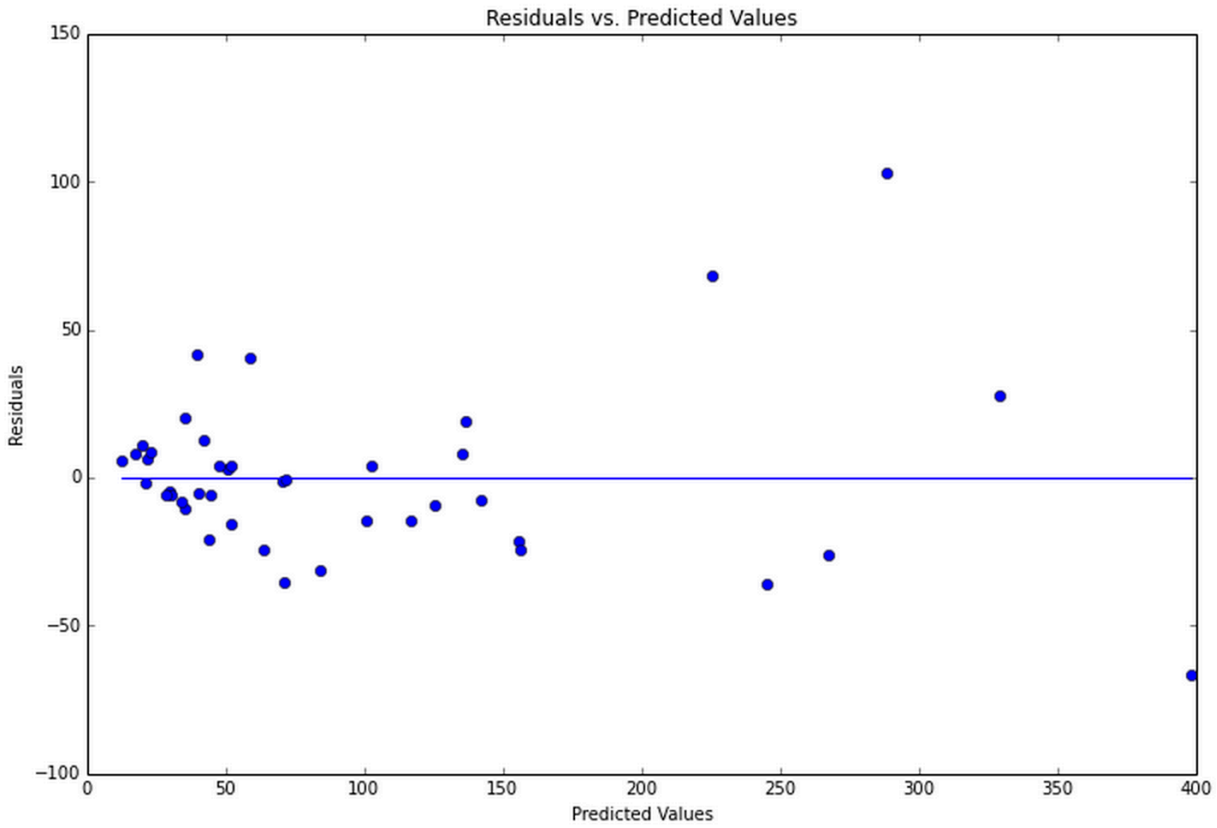
	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	-8.1809	15.240	-0.537	0.594	-39.006	22.644
np.power(num_comments, 2)	0.0279	0.012	2.422	0.020	0.005	0.051
num_comments	2.5205	0.957	2.633	0.012	0.584	4.457

Omnibus:	18.766	Durbin-Watson:	2.213
Prob(Omnibus):	0.000	Jarque-Bera (JB):	32.172
Skew:	1.200	Prob(JB):	1.03e-07
Kurtosis:	6.552	Cond. No.	6.77e+03

Warnings:
[1] The condition number is large, 6.77e+03. This might indicate that there are strong multicollinearity or other numerical problems.

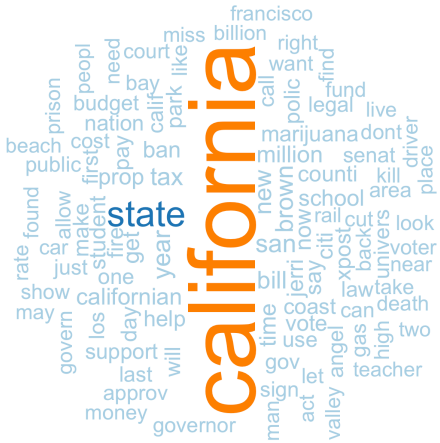
- Model:
  average post score = -8.1809 + 0.0279*average number of comments^2 + 2.5205*average number of comments
- Even higher Rsquared
- Strong multicollinearity

Residual vs. Fitted Plot



- Looks better
- Residuals spread as the predicted value increases

Word Clouds and High Frequency Words




```
[1] "ban"        "bill"       "brown"      "california" "californian"
[6] "get"        "million"    "new"        "now"        "prop"
[11] "san"       "state"      "tax"        "time"       "year"
```

```
[1] "angel"  "citi"   "downtown" "just"   "los"    "park"
[7] "time"   "today"  "xpost"    "year"
```

- Plot wordclouds and list the words by frequency
- California and Los Angel are the most frequent
- No similarity in words in post titles in these two subreddits

Comparing top 5 posts

	city	cityscore	title	state	statescore	title.1
0	Atlanta	2644	14 Year Old missing in Atlanta since Monday af...	Georgia	235	As someone who moved to Georgia last fall...
1	Atlanta	573	How it feels outside today	Georgia	190	How it feels living in GA lately.
2	Atlanta	545	Anyone interested in a former Homeless Atlanta...	Georgia	171	I think the GA subreddit has become too divisi...
3	Atlanta	514	Why I eat at the Vortex	Georgia	169	sometimes i miss living in the south
4	Atlanta	507	I saw one of you on my way to work	Georgia	146	How to be a better driver than 99% of people i...
5	Baltimore	924	Throwaway time... Have you ever been in the Ba...	Maryland	258	The most accurate weather map in Maryland history
6	Baltimore	364	I live in NY now, but I got this made awhile b...	Maryland	248	Finally received my MD flag bikini!
7	Baltimore	351	The go to pen of the Baltimore area	Maryland	225	Scumbag Maryland
8	Baltimore	332	Updated map! It got a little crowded, but it'...	Maryland	194	Most Maryland Photo I've Ever Taken Hon'
9	Baltimore	265	Was recently in Kansas City to see the Orioles...	Maryland	189	I made a logo for r/Maryland...What do you guy...

- Titles of top posts still look no similarity
- Also comparing titles in different score ranges e.g: 500 to 1000, they are still quite different.
- Important events or news lead to high score. But the events and news are varied, even in the same state.

Thank You