Post Score in US State and City Subreddits

Kemengtian Ma

Stat 222 Capstone Project

Research Concerns

- 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

Data Description and Manipulation

- A Subset of the 2.5 million top reddit posts
- The original data set:

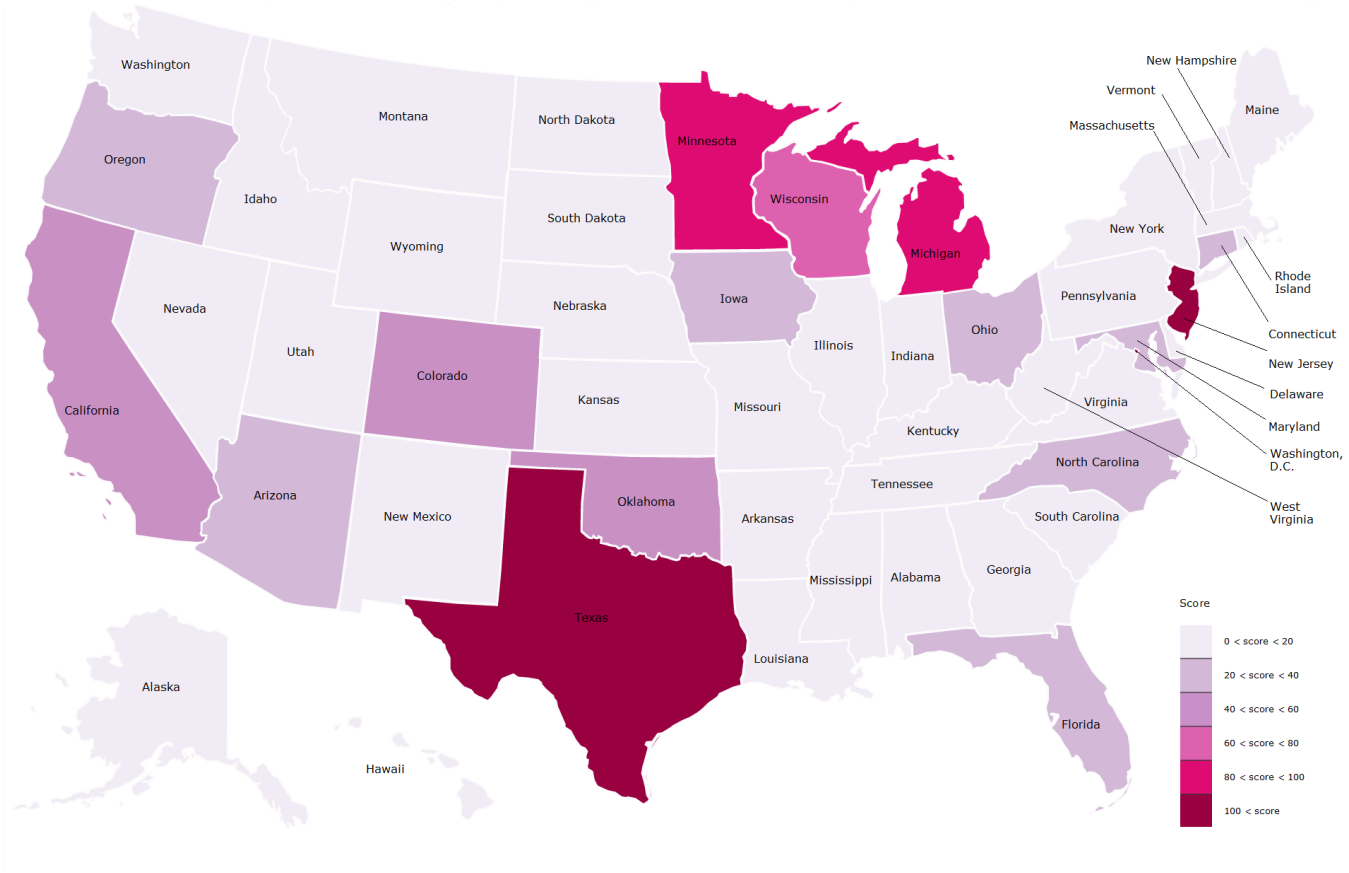
  2500 csv files

  top 1000 posts in the subreddit, ranked by post score

- Choose files with state names and US city names
- Delete useless variables
- Calculate averages of the numeric statistics of posts,

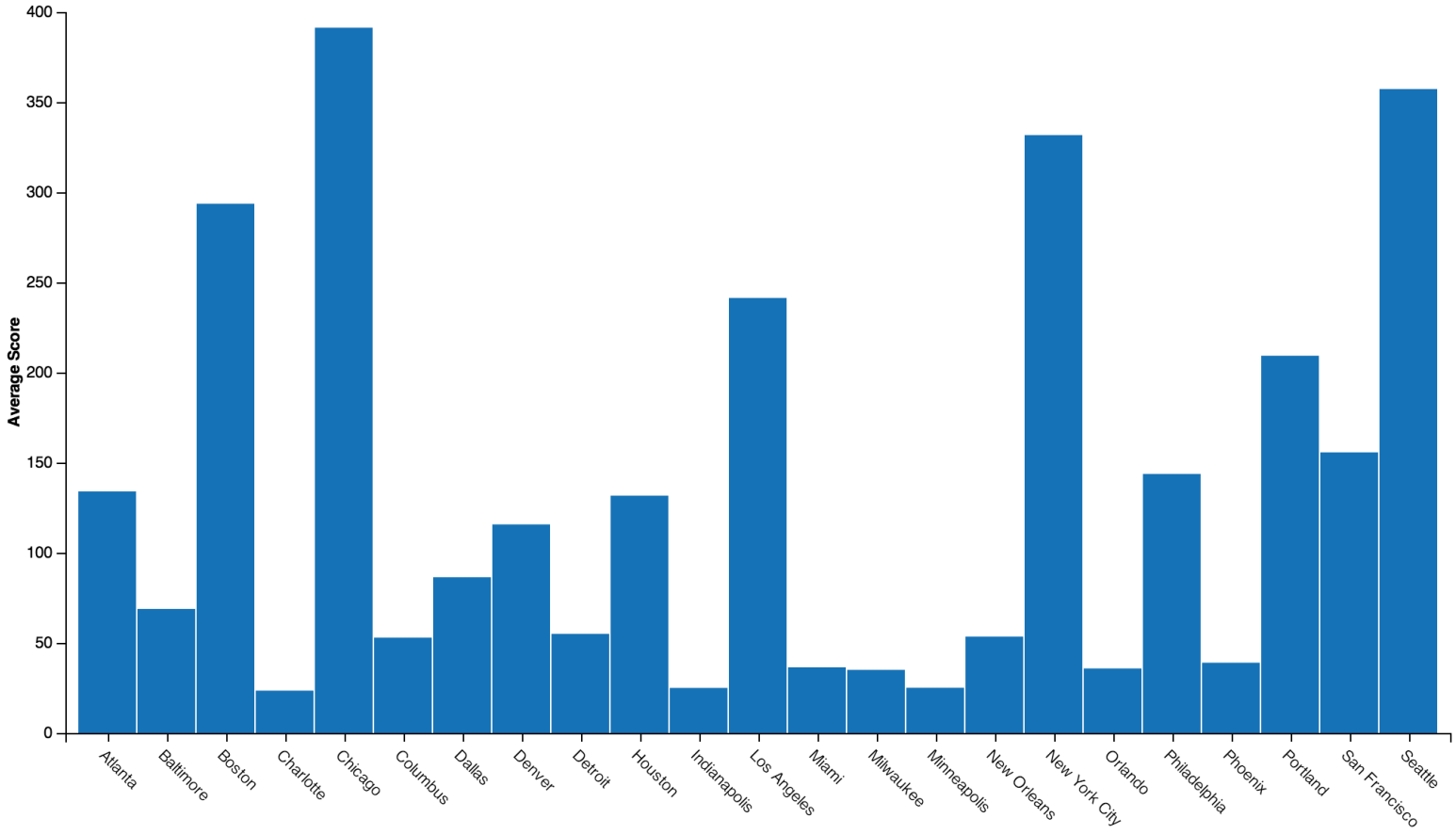
  E.g number of comments

US States in Top Subreddits

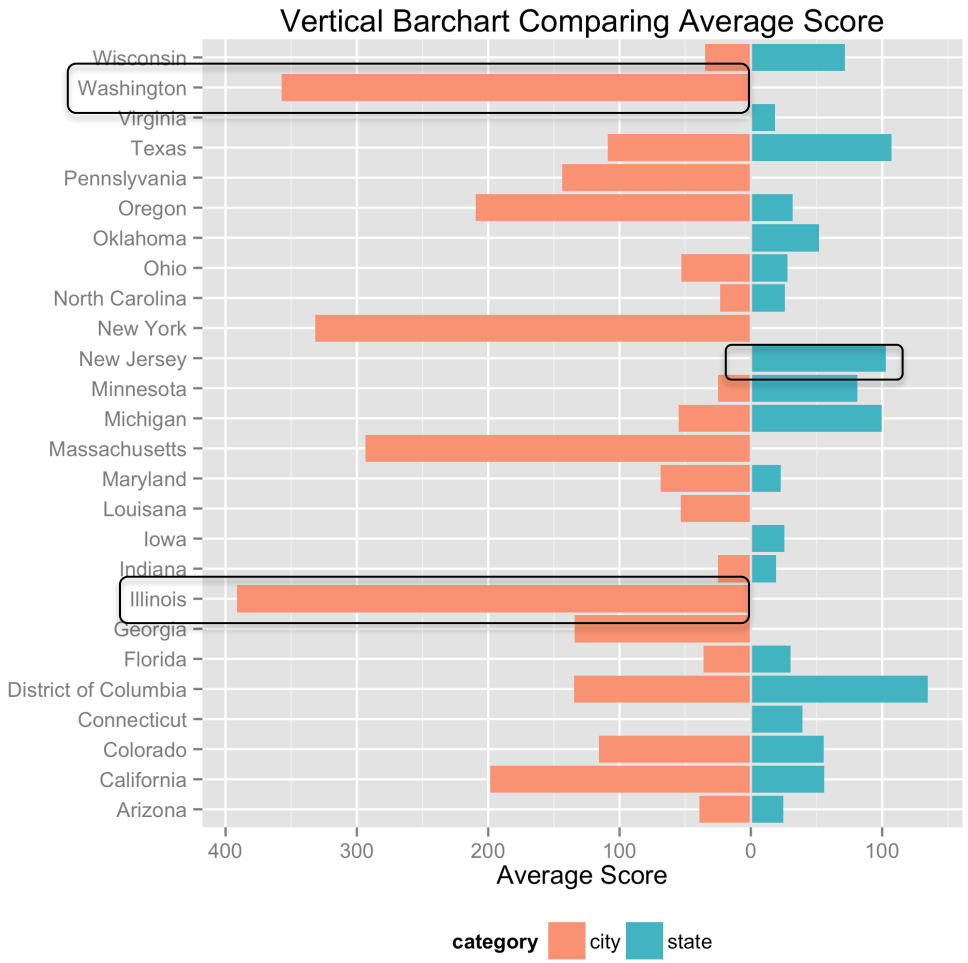


- Shaded by the different levels of average scores
- Score 0 means not in the top subreddits
- Texas and New Jersey
- States around lake regions

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

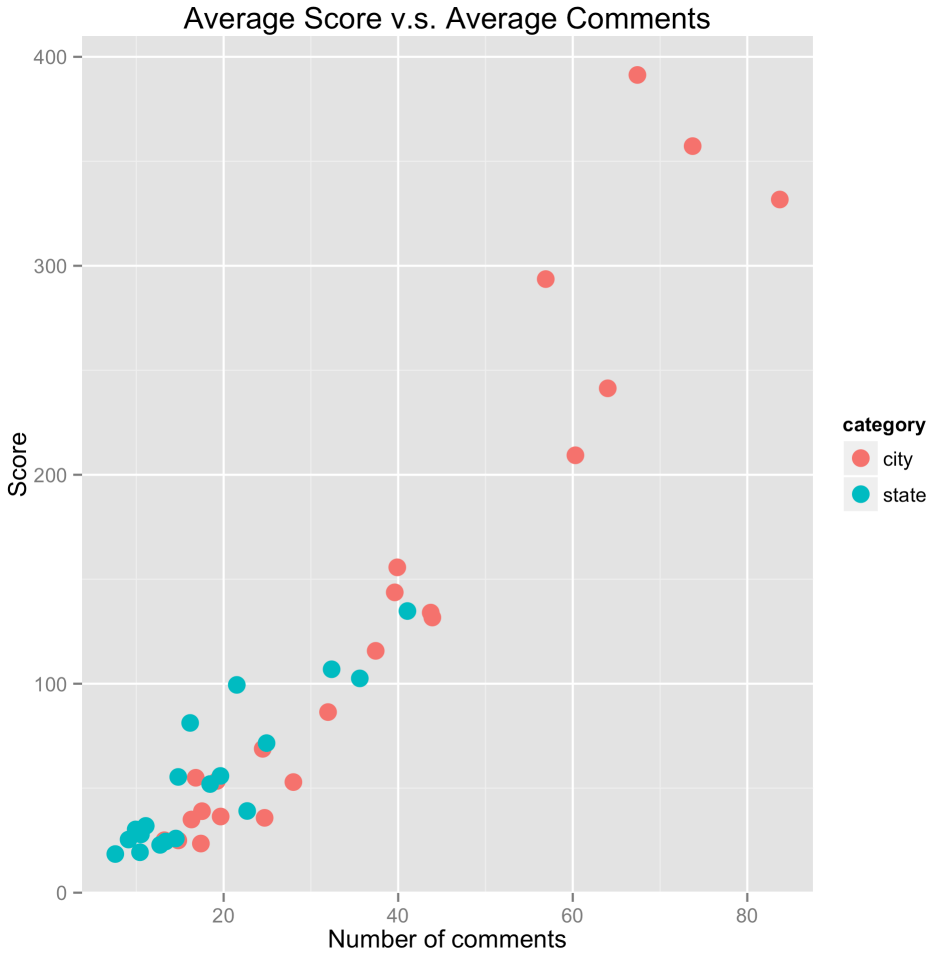
One Tail Hypothesis Test

```
        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

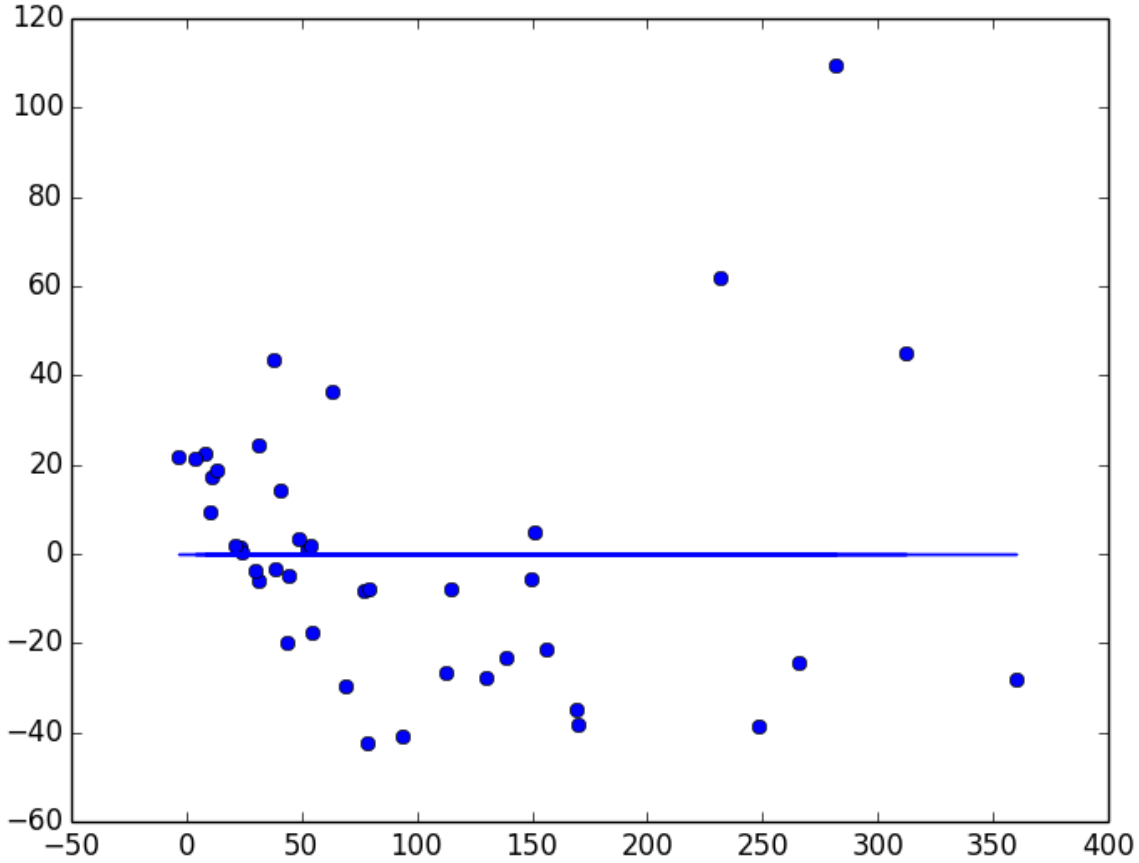
- Model:

  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



- 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

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
- Check the word frequency in content or comments of the post

Thank You