

Predicting wikipedia article popularity

or: How to disobey robots.txt and scrape some stupidly clean wiki data

jon kislin
nyc18_ds14
project luther



Can we predict wikipedia article popularity?


Yes, and it's easy, because
wikipedia has already done
most of the work

Random article button
+ dynamic stats page
+ being a bad robot
= web-scraping success

Information for "New York City"

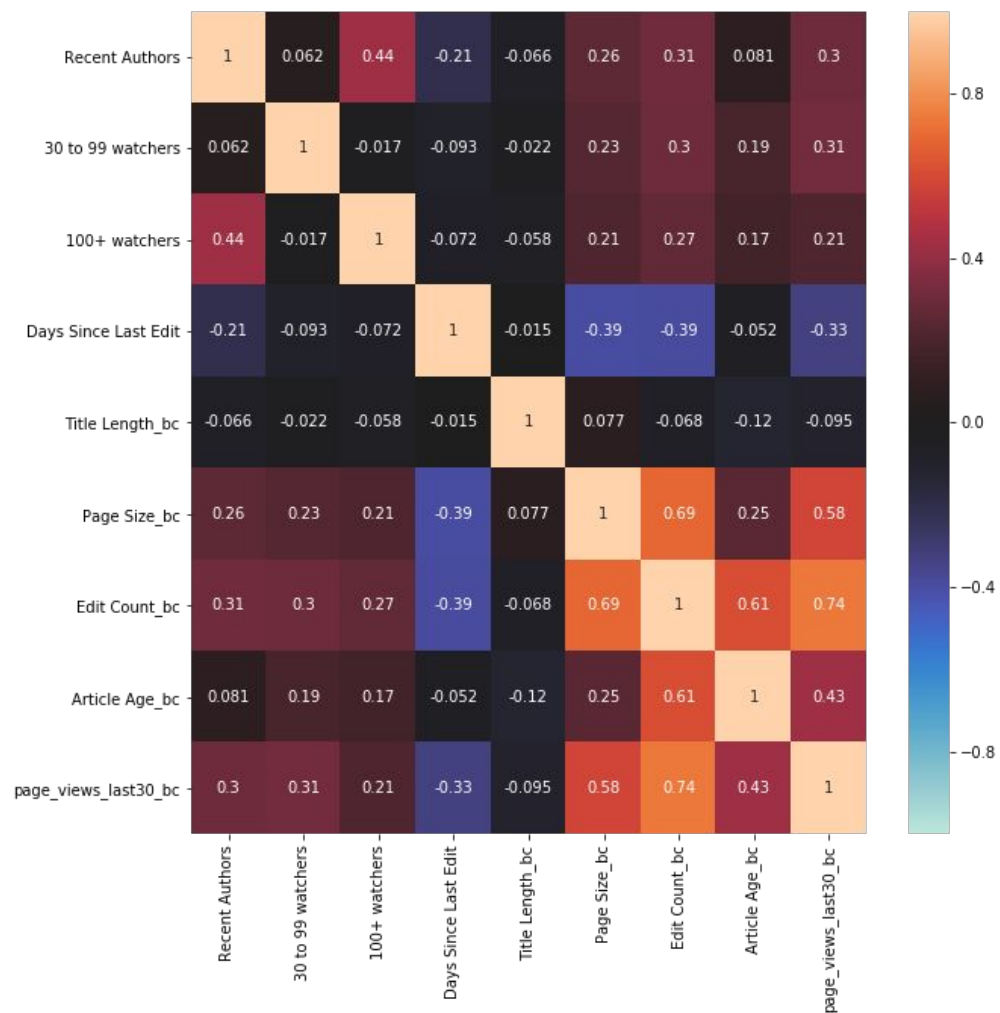
[Help:Page information](#)

Basic information

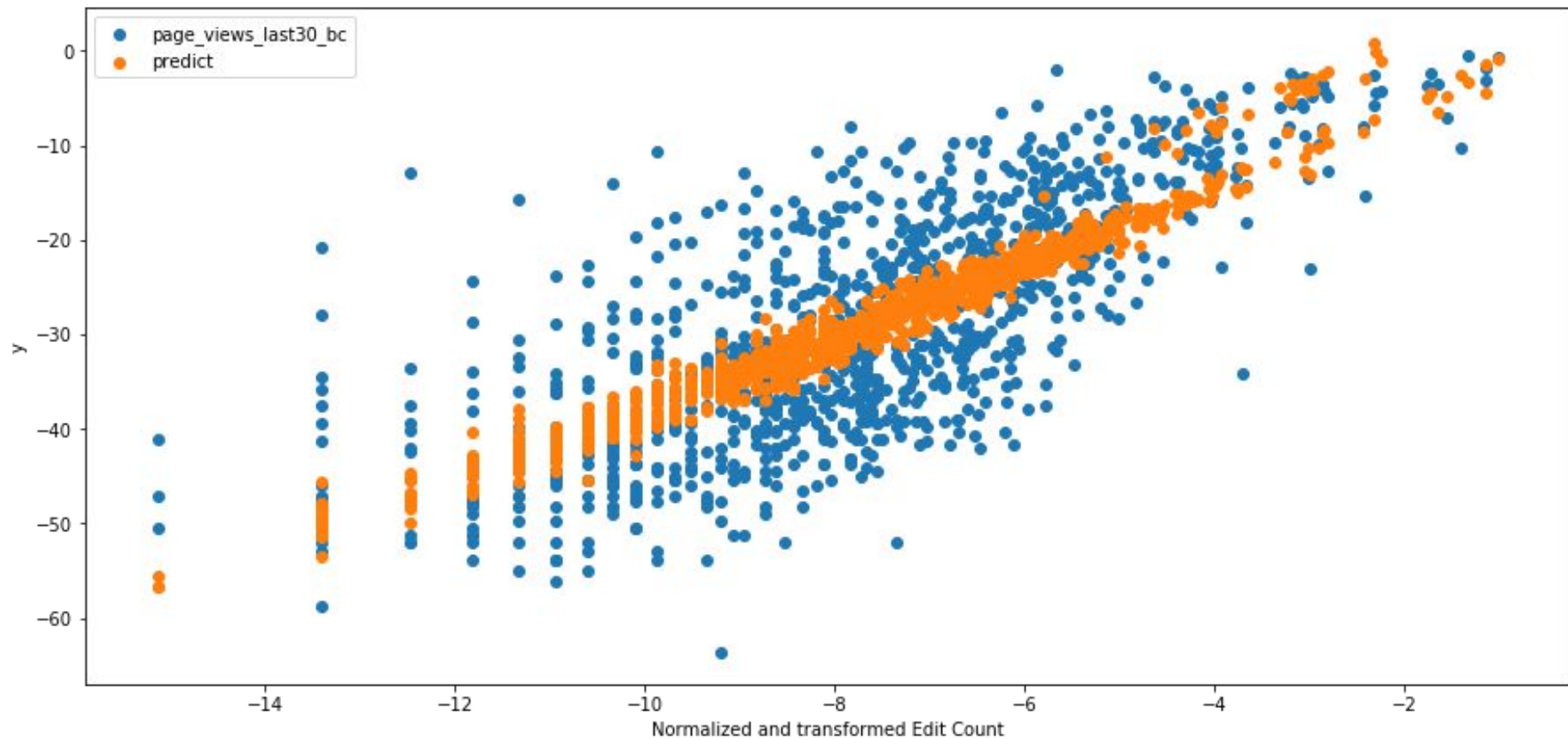
Display title	New York City
Default sort key	New York City
Page length (in bytes)	329,951
Page ID	645042
Page content language	en - English
Page content model	wikitext
Indexing by robots	Allowed
Number of page watchers	1,614
Number of page watchers who visited recent edits	173
Number of redirects to this page	81
Counted as a content page	Yes
Wikidata item ID	Q60
Page image	
Page views in the past 30 days	506,996

Structure

	Title	Page Size	Watchers	Edit Count	Creation Date	Last Edit Date	Recent Authors	page_views_last30	Title Length
0	La Seine no Hoshi	3,853	Fewer than 30 watchers	67	16:26, 19 June 2005	13:50, 20 January 2018	1	577	17
1	Gary Keating	4,391	Fewer than 30 watchers	18	20:05, 23 September 2014	20:26, 28 August 2017	0	50	12
2	Colton's Crossing Bridge	2,472	Fewer than 30 watchers	18	16:58, 14 July 2012	02:15, 15 October 2017	0	42	24
3	Philippe Echaroux	7,418	Fewer than 30 watchers	36	01:07, 1 December 2016	09:10, 1 July 2017	0	163	17



Edit counts vs. Page Views



OLS Regression Results

Dep. Variable:	page_views_last30_bc	R-squared:	0.577
Model:	OLS	Adj. R-squared:	0.574
Method:	Least Squares	F-statistic:	168.8
Date:	Thu, 25 Jan 2018	Prob (F-statistic):	4.29e-179
Time:	20:32:22	Log-Likelihood:	-3510.0
No. Observations:	999	AIC:	7038.
Df Residuals:	990	BIC:	7082.
Df Model:	8		
Covariance Type:	nonrobust		

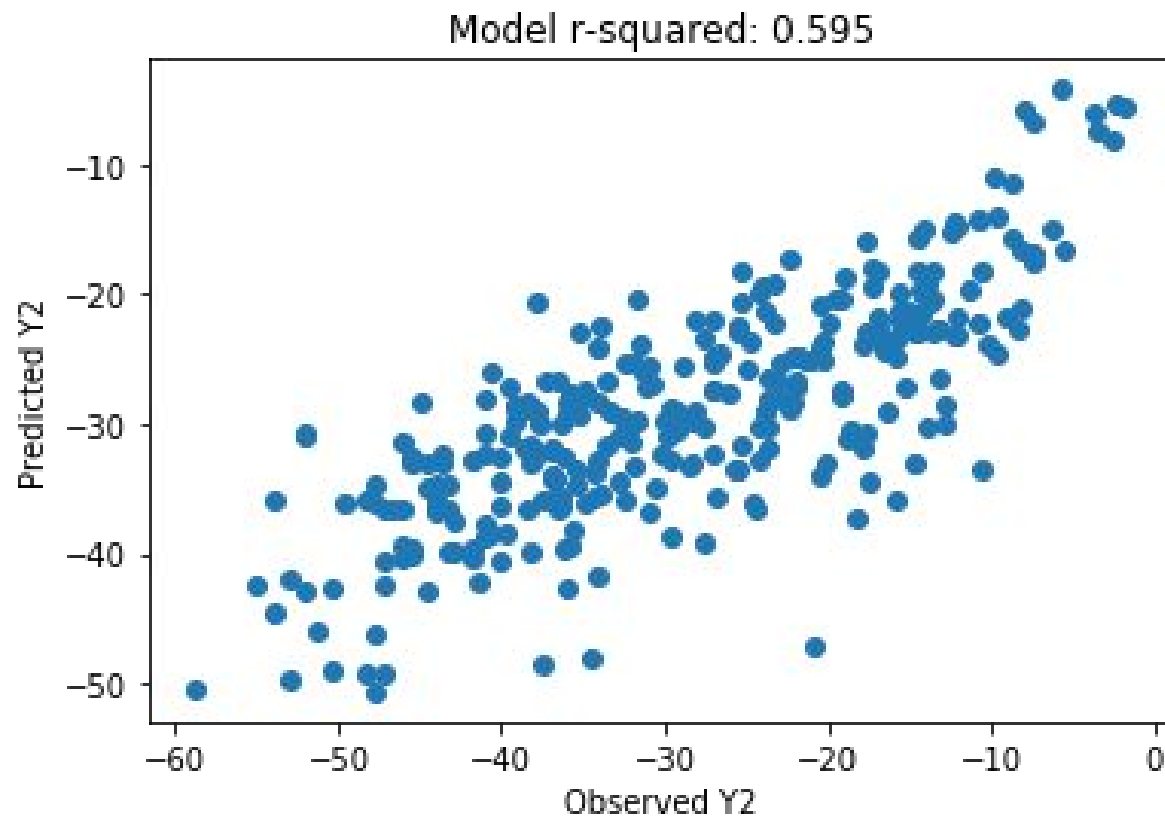
	coef	std err	t	P> t	[0.025	0.975]
const	1.3605	17.637	0.077	0.939	-33.249	35.970
Recent Authors	32.0884	9.844	3.260	0.001	12.770	51.406
30 to 99 watchers	39.0192	8.752	4.458	0.000	21.844	56.195
100+ watchers	-3.8296	9.329	-0.410	0.682	-22.137	14.478
Days Since Last Edit	-11.9709	11.607	-1.031	0.303	-34.748	10.806
Title Length_bc	-2.4494	0.880	-2.782	0.006	-4.177	-0.722
Page Size_bc	0.8293	0.187	4.445	0.000	0.463	1.195
Edit Count_bc	3.1202	0.217	14.385	0.000	2.695	3.546
Article Age_bc	7.0019	16.753	0.418	0.676	-25.873	39.877

Preliminary Model

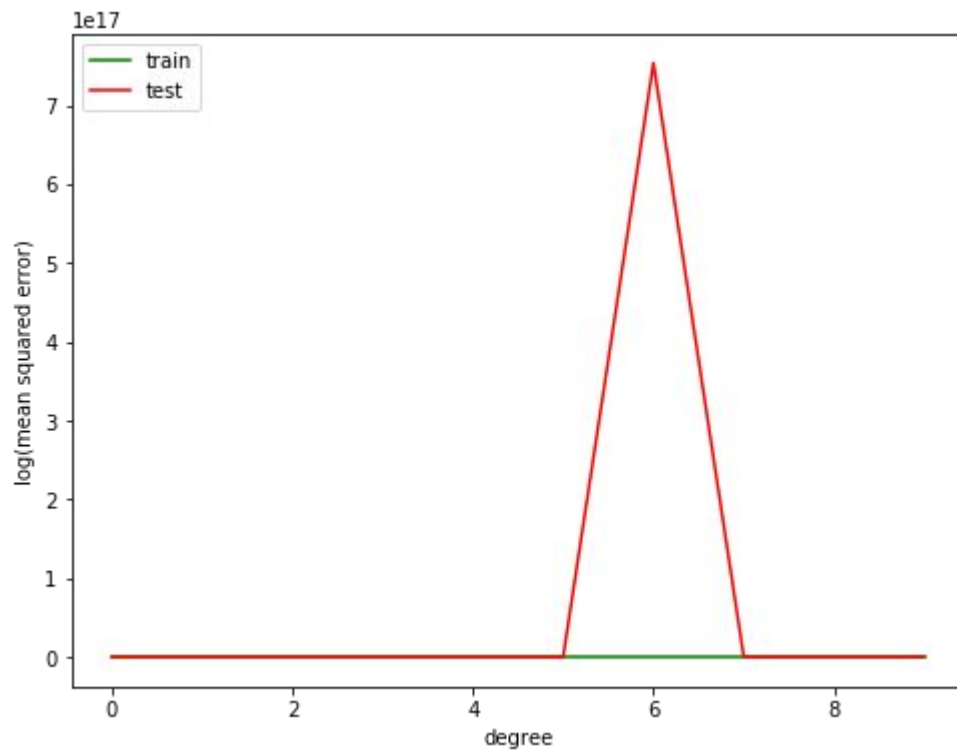
OLS Regression Results

Dep. Variable:	page_views_last30_bc	R-squared:	0.571
Model:	OLS	Adj. R-squared:	0.570
Method:	Least Squares	F-statistic:	331.2
Date:	Thu, 25 Jan 2018	Prob (F-statistic):	4.02e-181
Time:	20:32:26	Log-Likelihood:	-3516.6
No. Observations:	999	AIC:	7043.
Df Residuals:	994	BIC:	7068.
Df Model:	4		
Covariance Type:	nonrobust		

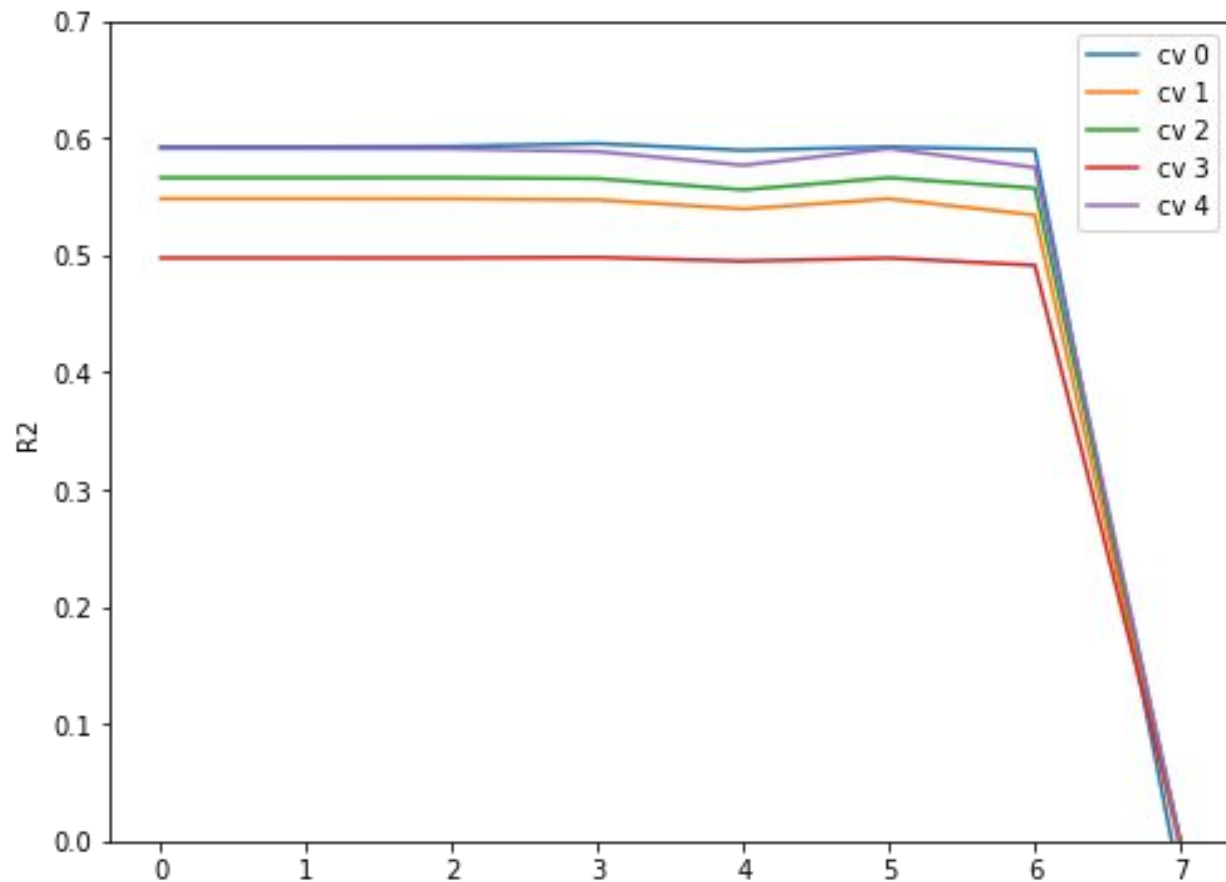
Test-train split



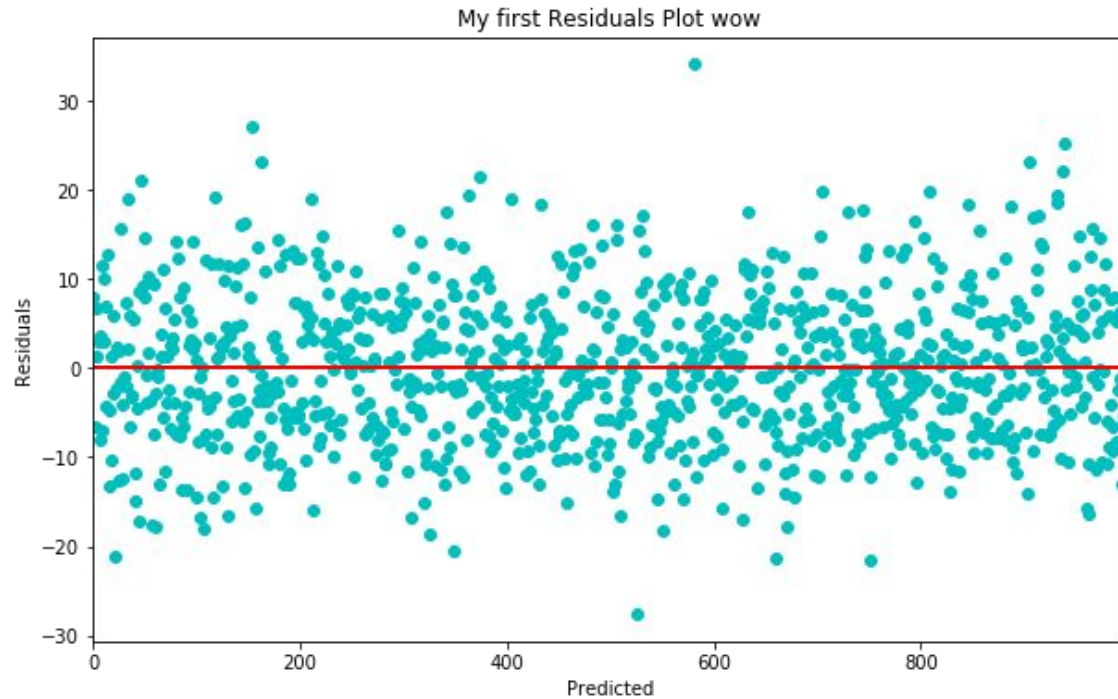
Polynomials?



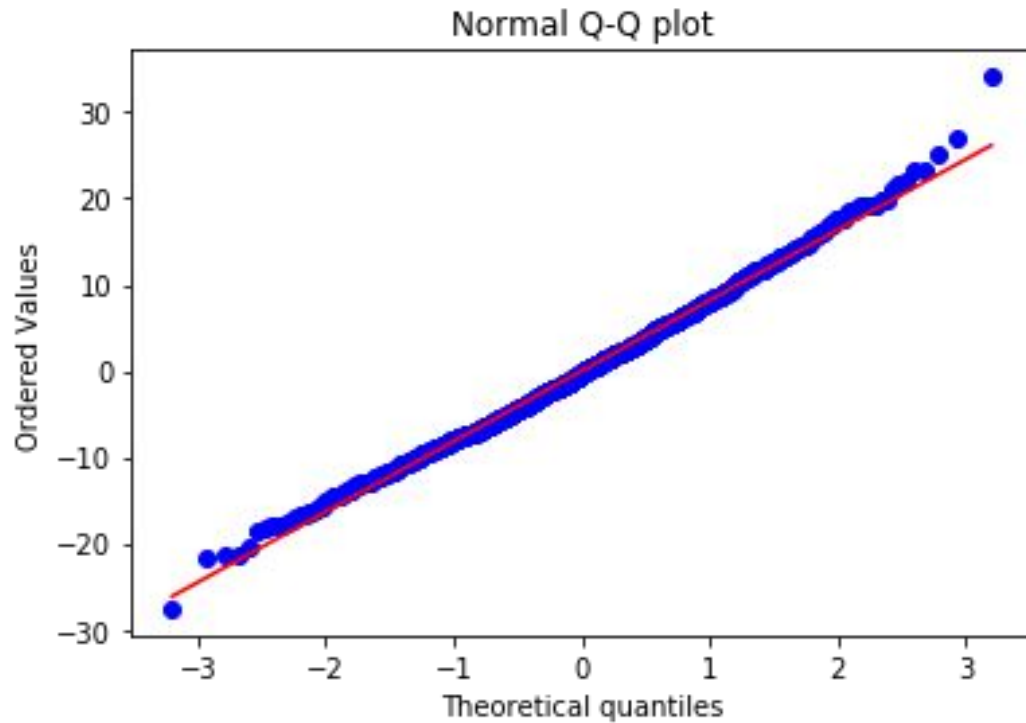
Finding the best lasso lambda

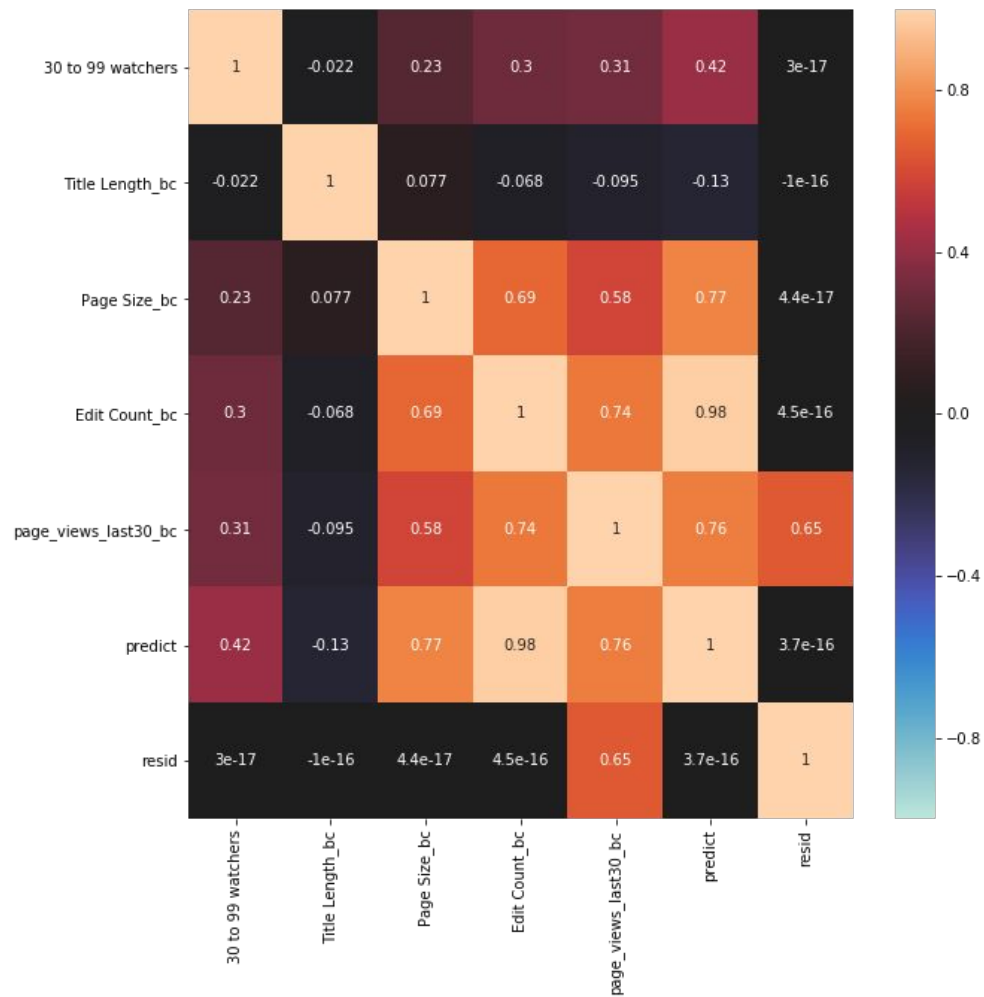


Violating assumptions?



Violating assumptions?





Didn't have time to include:

- Article *quality* as assessed by wiki
- Whether article was *featured* on main page or not

Arguably causal - do these arbitrary measures attract readers?

- From what I understand, featured/good articles are often pet-projects of domain experts

In short...

...Wikipedia is beautiful