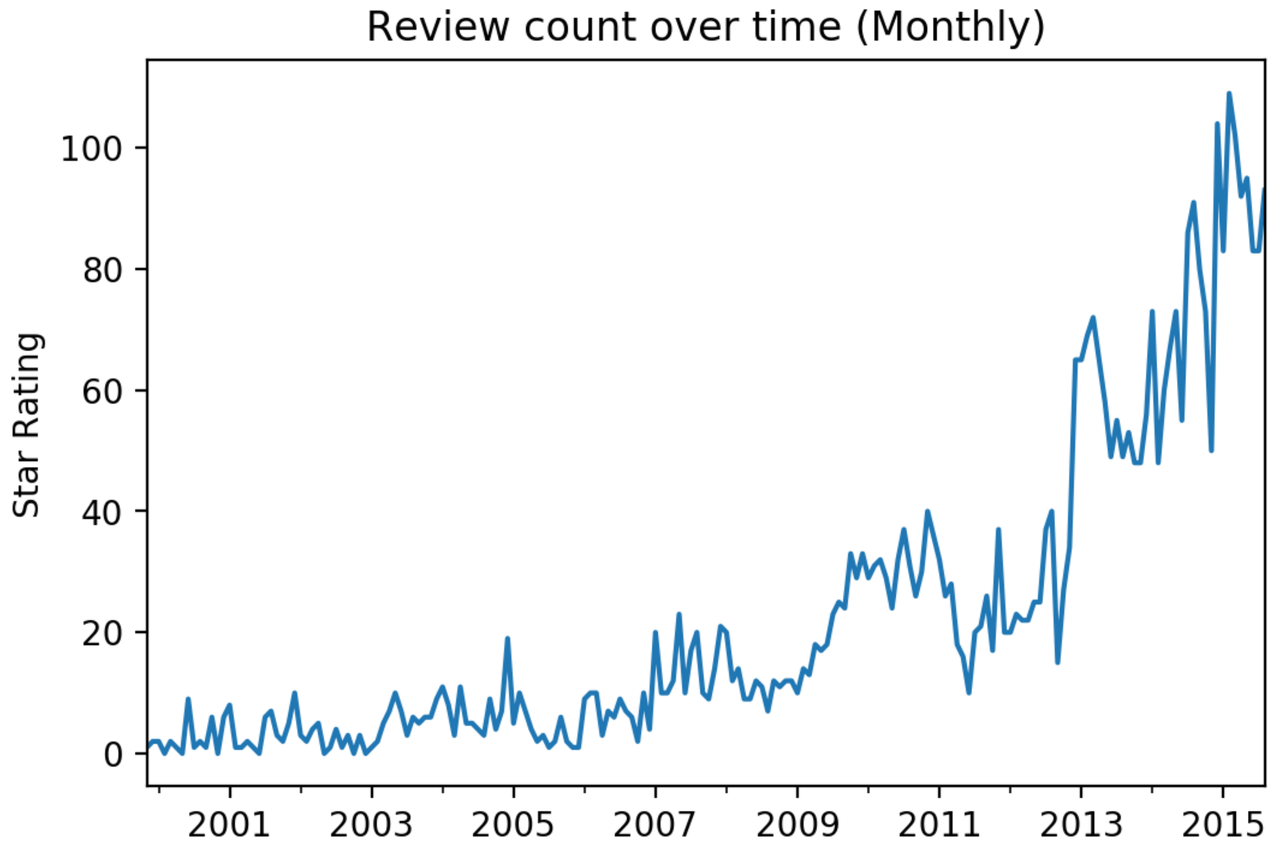
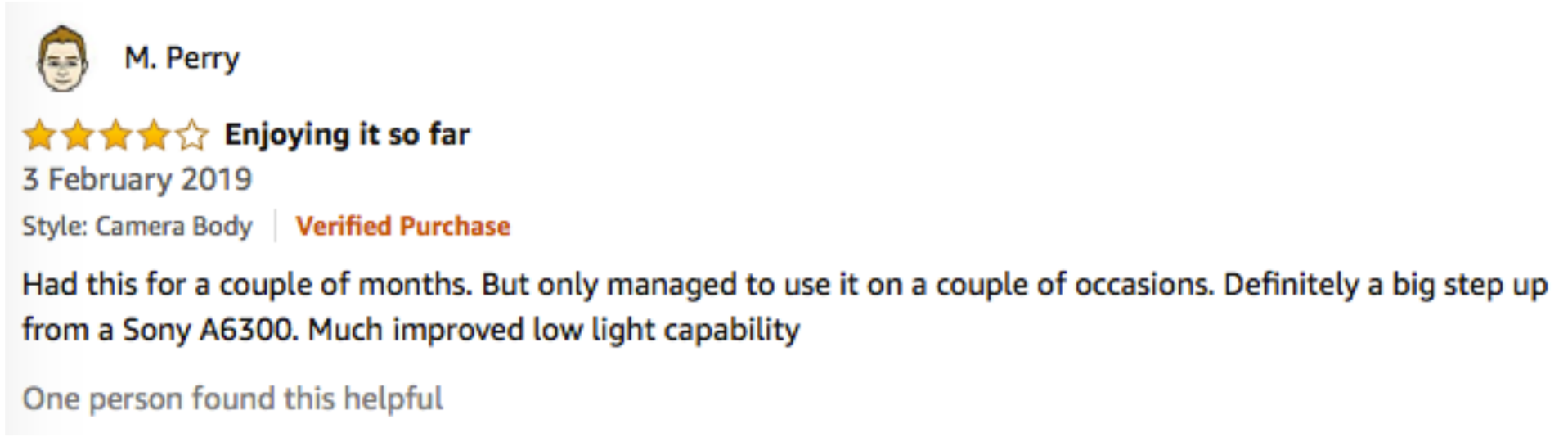
Predicting Camera review Ratings on amazon.com

Please leave a review..



Review's provide customer insight



M. Perry

★★★★☆ **Enjoying it so far**

3 February 2019

Style: Camera Body | **Verified Purchase**

Had this for a couple of months. But only managed to use it on a couple of occasions. Definitely a big step up from a Sony A6300. Much improved low light capability

One person found this helpful

Reviews drive purchase decisions

1 Contextual usage information

2 Reduce risk: validate manufacturer claims

3 **Understand consumer needs (met and unmet)**

Potential Audience

Product developers and marketeers seeking understanding of the digital camera market through reviewer behaviour and identifcation of unmet needs.

Goals

Identify camera features that lead to good or bad experiences

Predict the product rating better than baseline score
0.713

Dataset

A collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015.

(130M+ customer reviews)

Tab Separated .TSV
Web Scraping

https://registry.opendata.aws/amazon-reviews/

Dataset Features

Features

Marketplace

customer_id

review_id

product_id

product_parent

product_title

product_category

star_rating

helpful_votes

total votes

vine

verified_purchase

review_headline

review_body

review_date


36 of 39 people found the following review helpful

★★★★★ **Cute and comfortable? Perfect**, January 29, 2017

By **Zora F.**

Edit review Delete review

Verified Purchase (What's this?)

This review is from: Netgo Ombre Wig Long Wavy Blonde Ombre Wig Dark Roots Heat Resistant Sythentic Full Wigs for Women (Misc.)

Love, love love. This is going into my repertoire of casual wigs.

Appearance: Great. I had several people ask me when I'd dyed my hair (my father asked what my coworkers thought about the new look) and it's a beautiful and flattering ombre. The curls are pretty but don't look 'too perfect'. Yes, if you look closely towards the hairline there's a thin but obvious line, but for the price of the wig I will gladly accept that.

Fit: Wonderful! I have long, thick hair (it goes past my lower back) and run into the issue of 1. tucking it without looking absurd, and 2. being able to wear a wig without getting a splitting headache within a few hours due to tightness. I wore this for 8+ hours and didn't have any complaints.

Features

Marketplace

customer_id

review_id

product_id

product_parent

product_title

product_category

star_rating

helpful_votes

total votes

vine

verified_purchase

review_headline

review_body

review_date

Process (pt 1)

Import
and cleaning

Feature Engineering

EDA

Features

Marketplace

customer_id

review_id

product_id

product_parent

product_title

product_category

star_rating

helpful_votes

total votes

vine

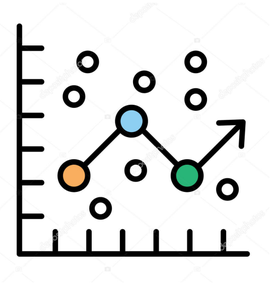
verified_purchase

review_headline

review_body

review_date

Process (pt 2)



Modelling

Results

Key takeaways & improvements

Technology List

Python
This project using Python-3.6 environment

Numpy
The fundamental package for scientific computing with Python

Sci-kit Learn
Simple and efficient package for data mining and data analysis.
Feature extraction using Count Vectoriser and Tfidf vectoriser.

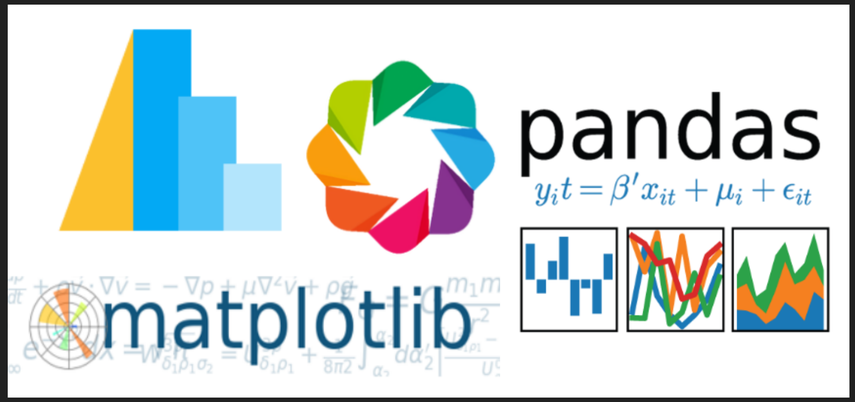
Pandas
Easy-to-use data structures and data analysis tools for Python

Selenium & Beastiful Soup
Automated browser to scrape information. Convert web page to xml so that all infomation on that page can be accessed easily

Vader Sentiment
lexicon and rule-based sentiment analysis tool

Imbalanced Learn
A package offering a number of re-sampling techniques commonly used in datasets showing strong between-class imbalance

Textacy
Python library for performing a variety of Natural Language Processing (NLP tasks)



Importing

```
b'Skipping line 85458: expected 15 fields, saw 22\nSkipping line 91161: e
xpected 15 fields, saw 22\n'
b'Skipping line 166123: expected 15 fields, saw 22\n'
b'Skipping line 225458: expected 15 fields, saw 22\nSkipping line 229936:
expected 15 fields, saw 22\nSkipping line 259297: expected 15 fields, saw
22\n'
b'Skipping line 284728: expected 15 fields, saw 22\nSkipping line 286334:
expected 15 fields, saw 22\nSkipping line 293400: expected 15 fields, saw
22\nSkipping line 294415: expected 15 fields, saw 22\nSkipping line 30815
0: expected 15 fields, saw 22\nSkipping line 315022: expected 15 fields,
saw 22\nSkipping line 315730: expected 15 fields, saw 22\nSkipping line 3
16071: expected 15 fields, saw 22\nSkipping line 326729: expected 15 fiel
ds, saw 22\n'
b'Skipping line 329101: expected 15 fields, saw 22\nSkipping line 333077:
expected 15 fields, saw 22\nSkipping line 377031: expected 15 fields, saw
22\nSkipping line 389496: expected 15 fields, saw 22\nSkipping line 39048
6: expected 15 fields, saw 22\n'
b'Skipping line 418308: expected 15 fields, saw 22\nSkipping line 454332:
expected 15 fields, saw 22\nSkipping line 458342: expected 15 fields, saw
22\n'
b'Skipping line 460704: expected 15 fields, saw 22\nSkipping line 466250:
expected 15 fields, saw 22\nSkipping line 486023: expected 15 fields, saw
22\nSkipping line 492819: expected 15 fields, saw 22\nSkipping line 51746
8: expected 15 fields, saw 22\nSkipping line 520963: expected 15 fields,
saw 22\n'
```

Data cleaning:

- Finding only cameras

```
9  body_only_us = df_us[df_us['product_title'].str.contains('[Bb]ody [Oo]nly', regex=True) ]
10 len(body_only_us)
11
```

```
4260
```

- Removing columns: [Marketplace, Product Category, Vine]
- No missing values
- No duplicated reviews
- Data types (Review Date) to datetime
- Investigated outliers for numerical variables

Most helpful review (outlier)

```
#Review with most helpful votes

body_only_us[body_only_us['helpful_votes']>1000]
print(body_only_us[body_only_us['review_id'] == 'R21UDHF662K69V']['review_headline'].values
print(body_only_us[body_only_us['review_id'] == 'R21UDHF662K69V']['review_body'].values)
```

```
['Does the 7D beat full frame cameras?']
["No, but it's so good that one starts to contemplate this question, which was never the case
before the 7D was introduced. Both systems, crop and full frame, have their pros and cons and
place in photography. But before I get into that let me say I have not been as excited about
a camera since the introduction of the 5D MK I four years ago. That's because the 7D raises t
he crop camera bar to the point where crop users will not feel at a disadvantage to full fram
e camera users, especially if coupled with awesome ef-s lenses such as the 17-55 f2.8.<br /><
br />How so? The 7D sets a new standard in four major ways.<br /><br />1. It produces whoppin
g 18MP pictures, which are just 3MP shy of the current top of the line full frame Canon camer
as. Just few years ago most pros were producing stellar results using the 1Ds MKII 16MP camer
a. Now you have more MPs in a crop sensor, that's a major achievement. This achievement trans
lates into bigger prints and, perhaps more importantly, cropping power. Out shooting wildlife
with a 300mm instead of 400mm? You can crop the 7D files down to 50% of their original file s
ize and still obtain sharp pictures. It's just not that easy with the 1D MK III 10MP files.<b
r /><br />2. Many worried that extra MPs in small crop sensors would translate into nosier pi
ctures, but the amazing thing is that this camera produces images with what seems to be less
noise than the 1Ds MKII. The noise level is very good. At ISO 1600 I still prefer pictures co
ming from my 5D MKII, but below ISO1600 they are very close. Frankly, I can go with either ca
mera because most of my professionally shot portraits and product pictures are shot at ISO100
. At ISO100 both produce very clean files and are practically indistinguishable.<br /><br />3
. Focus is the one area that was lacking on the previous 1.6 crop Canon cameras and this came
ra changes that. It's not a 1D in focus speed and accuracy, but it's the next best thing comp
ared to them. It's faster than the Canon 5D MKII, which is known to be slightly faster or aro
und the focus performance range of the 50D and 40D.<br /><br />4. The drive chain is fast, so
fast it's beyond anything I needed in my professional work in portrait, commercial, and produ
ct photography. Going through pictures taken at 8fps produces very little difference from fra
me to frame. One probably has to shoot a very fast moving subject/object to see the advantage
of such fast drive system.<br /><br />There are obviously many other things that I have not c
overed in this review. But based on the above, all I can say is that this camera has really r
aised the bar for all cameras and made it much more affordable to obtain a professional level
camera for all types of photography. If you were considering buying the 5D MKII as an upgrade
give this camera a test because it might be all you need.<br /><br />As for the advantages of
crop cameras I always find it odd that casual users who shoot many things but focus on landsc
```

Remaining:

4260 reviews (headline and body)

397 Total unique products.

	customer_id	review_id	product_id	product_parent	product_title	star_rating	h
0	30739283	R190J2PDOZ5GVK	B00ZDWGFR2	390090468	Sony a7R II Full-Frame Mirrorless Interchangea...	3	
1	15760475	R3SGZ5G1GJAWVU	B00TSR7YPK	515216474	Nikon D750 DSLR Camera (Body Only) + 32GB Extr...	5	
2	10861723	R3BWM499VCMGS7	B00ZDWGFR2	390090468	Sony a7R II Full-Frame Mirrorless Interchangea...	5	
3	22343417	R3SMKIWNMR55UB	B00O29LKN6	474362814	Canon EOS 7D Mark II Digital SLR Camera (Body ...	5	
4	106593	RNBM8M0T11BV0	B009F2OUOQ	968361935	Olympus OM-G OMG OM-20 Manual Focus Film Camer...	1	

Feature engineering

helpful_ratio - Helpful votes : Total votes (fillna: mean)

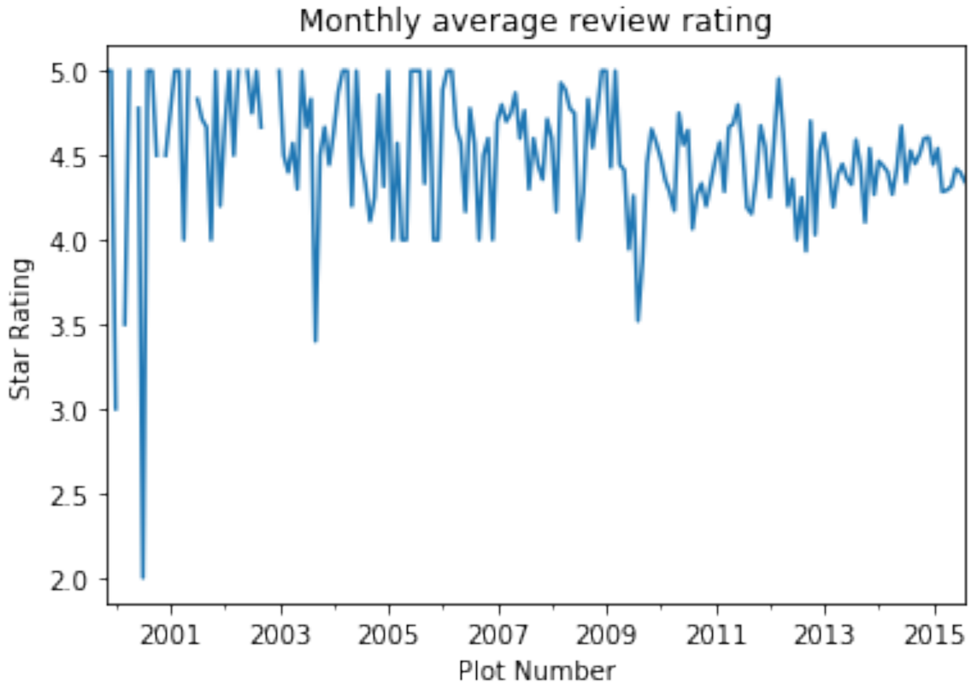
review_length - Word count of review body

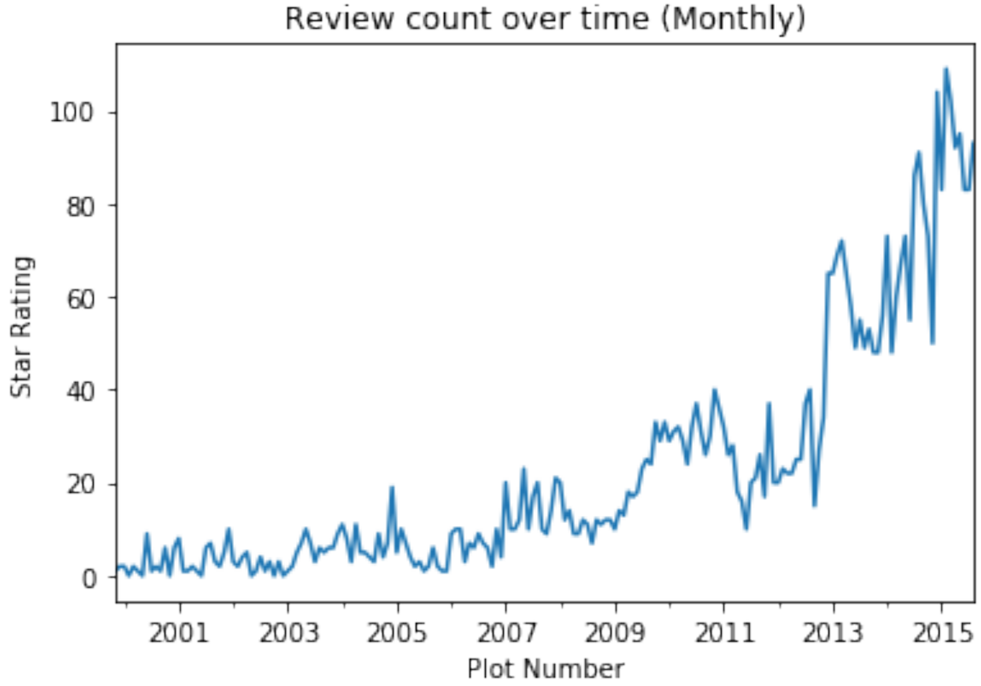
camera_brand - Extracted from product title. Regrouped.

VADER scores – Sentiment scores or each review

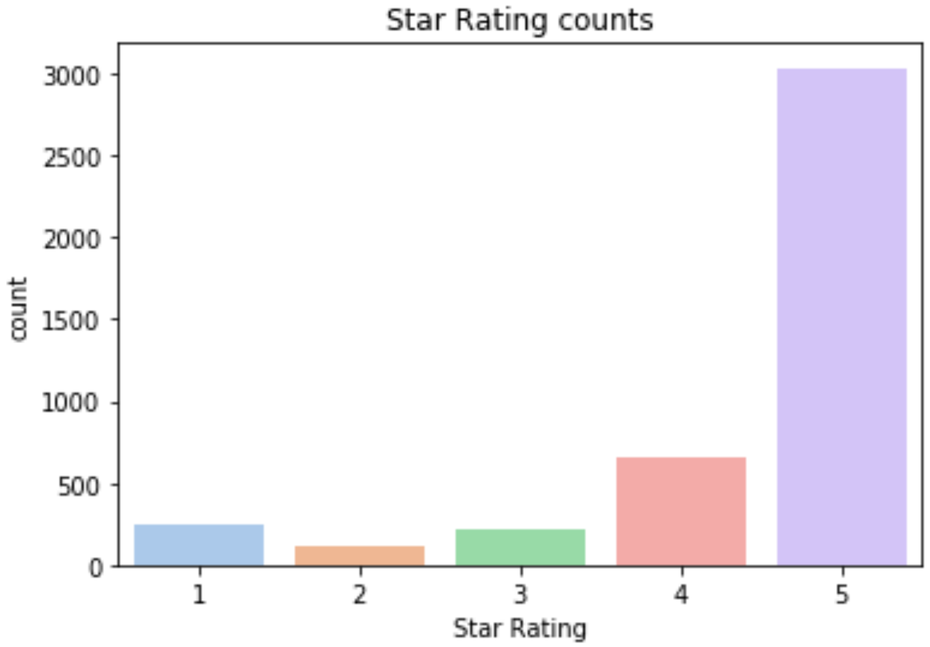
helpful_ratio	review_length	camera_brand	vader_compound	vader_neg	vader_neu	vader_pos
0.705882	199	sony	0.8964	0.103	0.768	0.129
1.000000	21	nikon	0.5100	0.000	0.788	0.212
0.923077	212	sony	0.9967	0.022	0.707	0.271

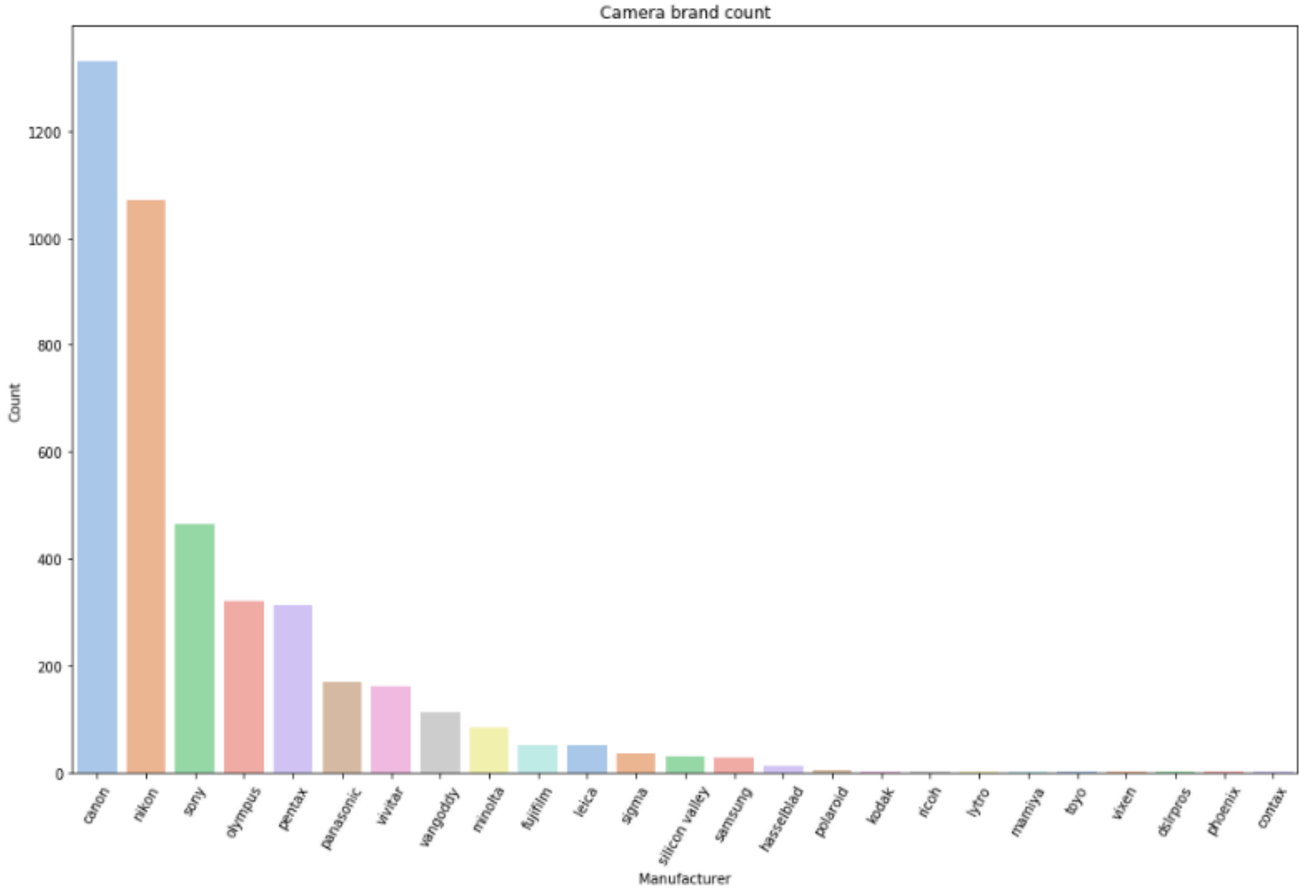
EDA: Time trends





EDA: ratings, and brand counts





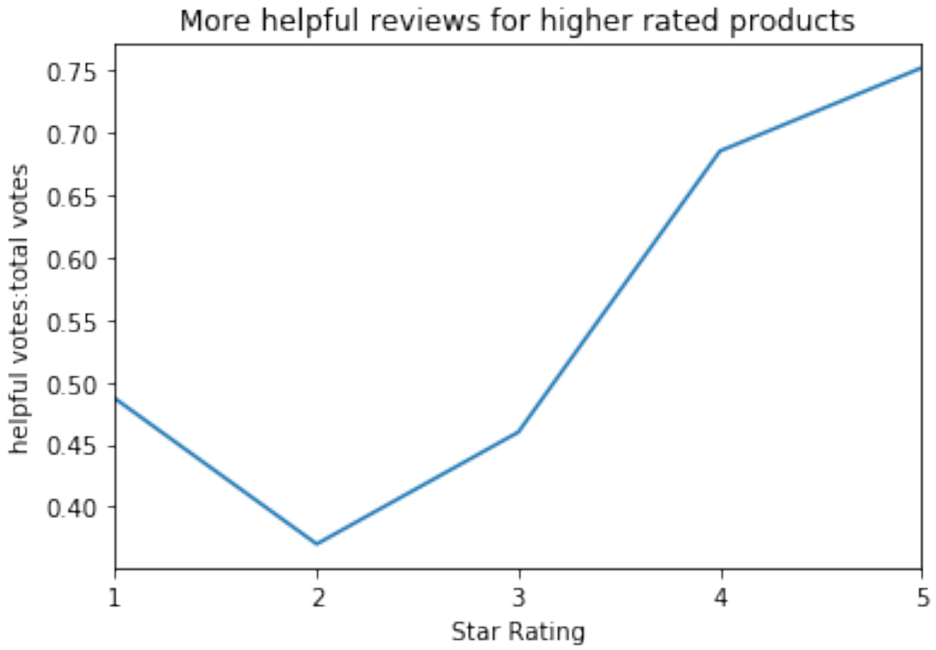
Brand average ratings
Canon 4.534586
Nikon 4.545709
Sony 4.496774
Vivitar 2.493827
Polaroid 2.200000

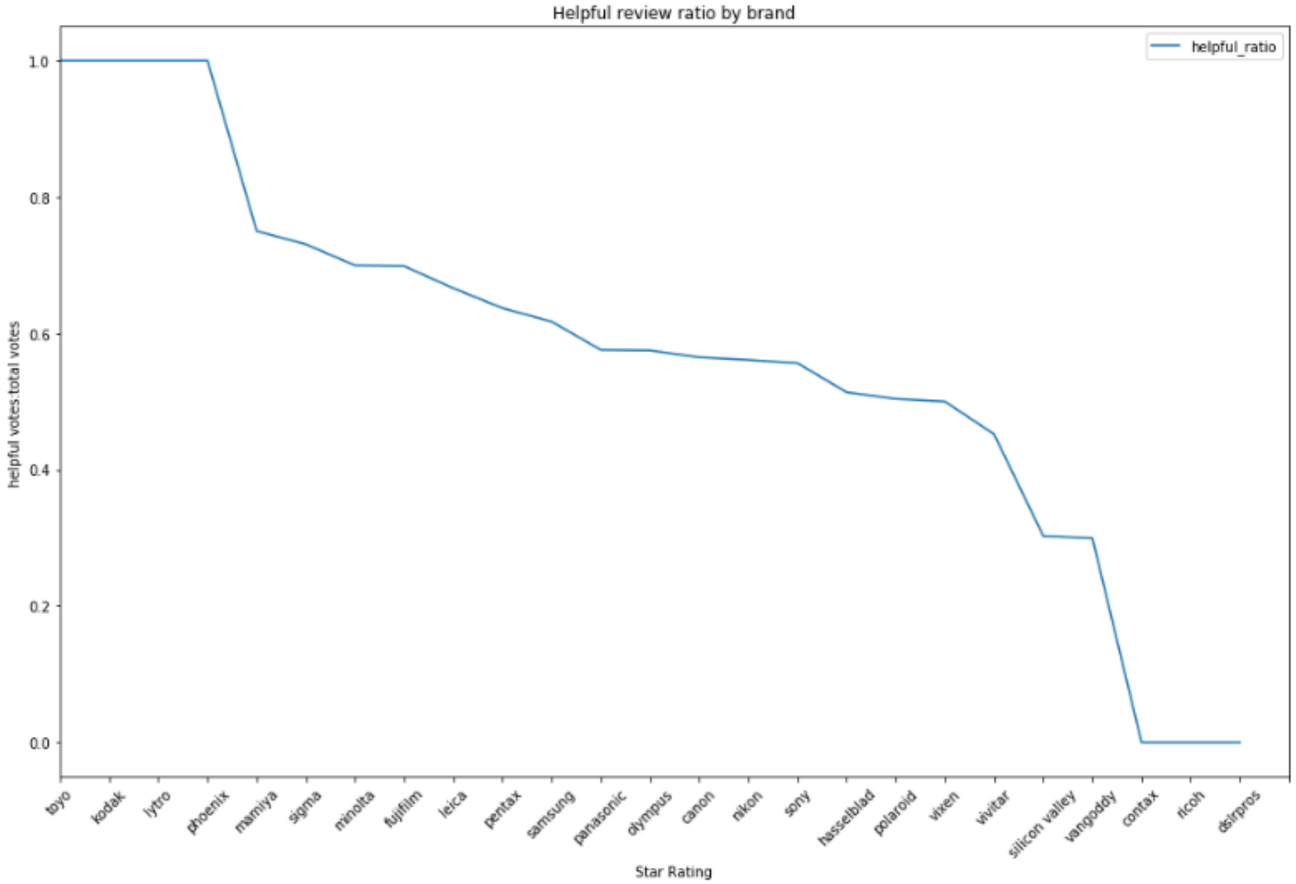
EDA: Most and least reviewed

```
Top 10 most reviewed cameras with review count

609 ['Canon EOS 7D 18 MP CMOS Digital SLR Camera Body Only discontinued by manufacturer']
207 ['Nikon D200 10.2MP Digital SLR Camera (Body Only)']
152 ['Olympus 16MP Mirrorless Digital Camera with 3-Inch LCD - Body Only']
112 ['Nikon D800E 36.3 MP CMOS FX-Format Digital SLR Camera (Body Only) (OLD MODEL)']
100 ['Panasonic Lumix DMC-GH3K 16.05 MP Digital Single Lens Mirrorless Camera with 3-Inch OLE
D - Body Only']
95 ['Vivitar 8 MP Compact System Camera with 2.4-Inch LCD Body Only']
```

```
Bottom 10 most reviewed products with review count

1 ['Nikon D5200 24.1 MP CMOS Digital SLR Camera Body Only (Bronze) + EN-EL14 Replacement Lith
ium Ion Battery\xa0w/ External Rapid Charger + 32GB SDHC Class 10 Memory Card + 52mm Wide Ang
le / Telephoto Lens + 52mm 3 Piece Filter Kit + Mini HDMI Cable + Carrying Case + Full Size T
ripod + External Flash + SDHC Card USB Reader + Memory Card Wallet + Deluxe Starter Kit\xa0Da
visMAX Bundle']
1 ['Fujifilm X-T1 16 MP Compact System Camera with 3.0-Inch LCD (Body Only)']
```

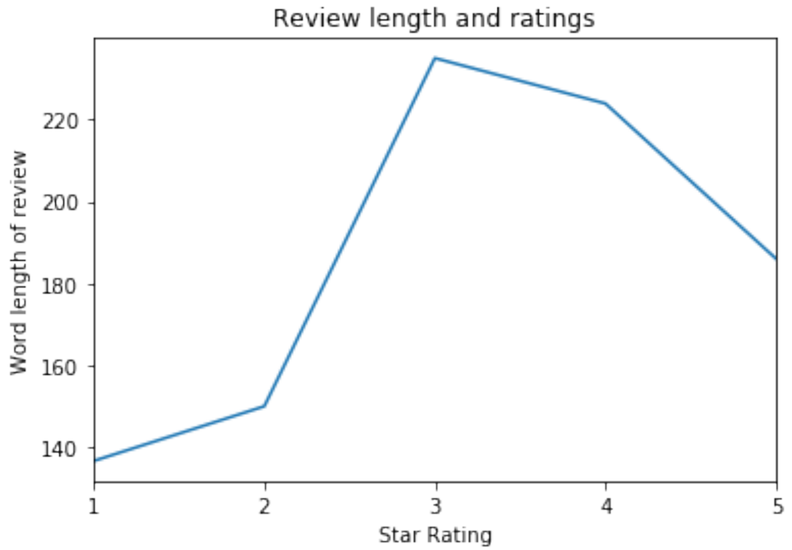
EDA: Helpful ratio





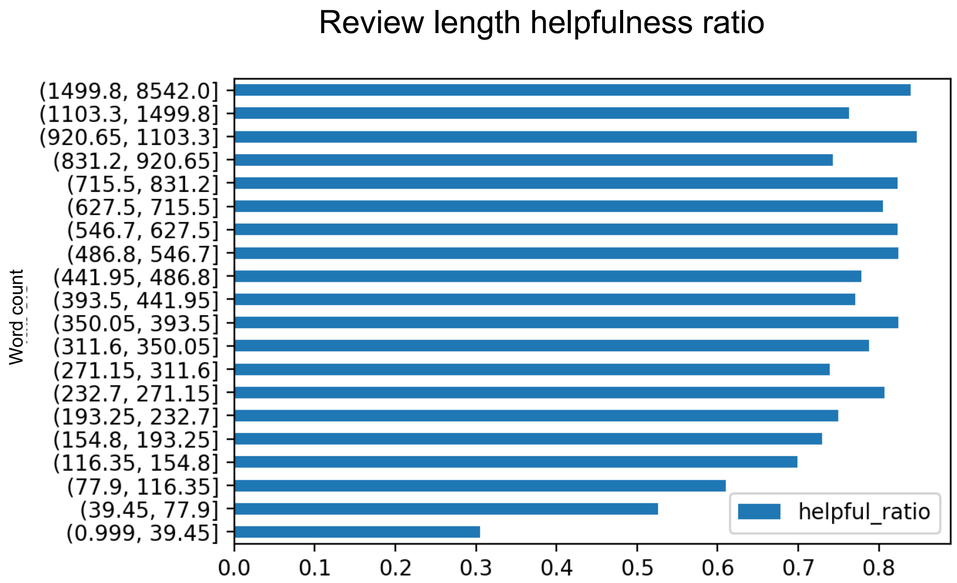
EDA: Review length

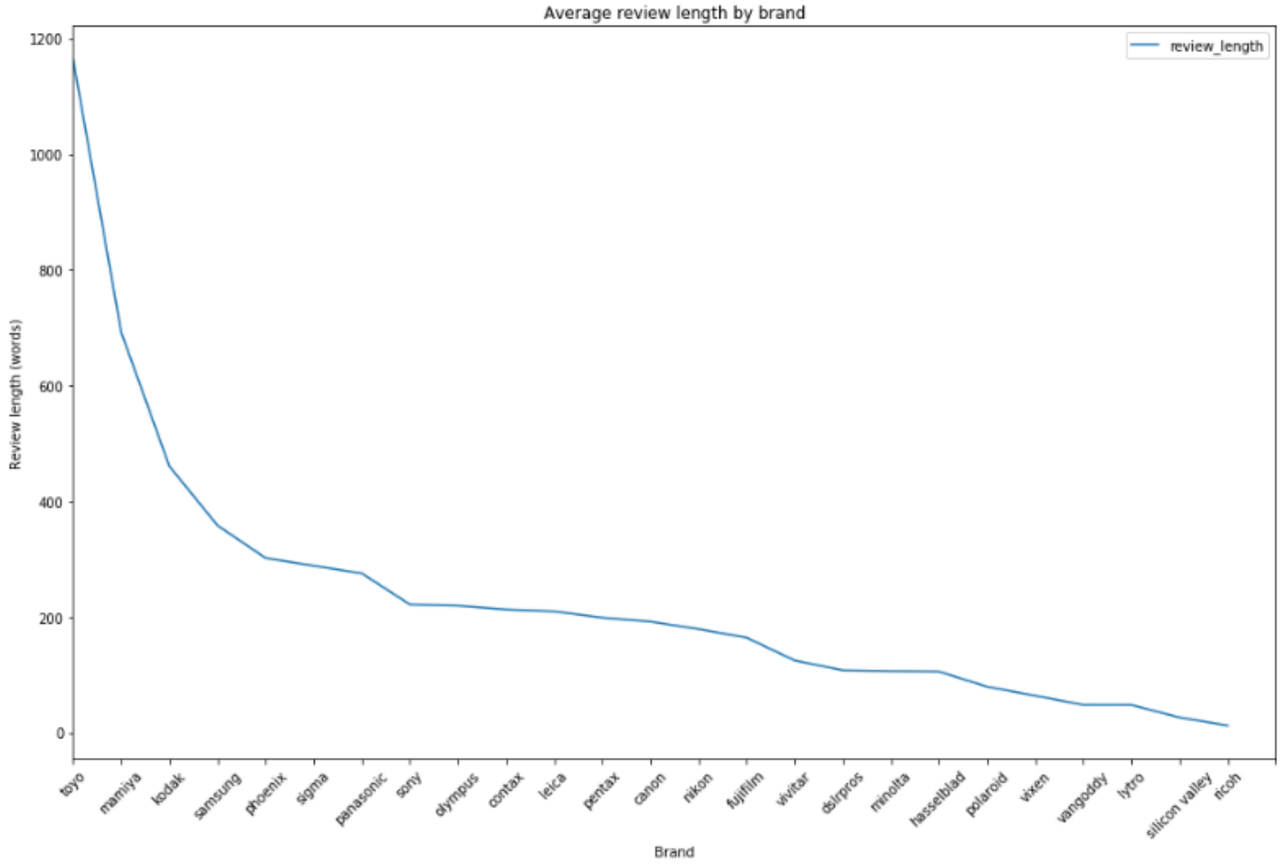
```
count     4252.000000
mean       190.542803
std        343.538724
min          1.000000
25%         32.000000
50%         85.000000
75%        218.250000
max       8542.000000
Name: review_length,
```



	product_title	star_rating	helpful_votes	total_votes	helpful_ratio	review_length
3946	Canon EOS 20D DSLR Camera (Body Only) (OLD MODEL)	5	96	96	1.0	1507
2546	Nikon D4 16.2 MP CMOS FX Digital SLR with Full...	5	69	69	1.0	579
3506	Nikon D40 6.1MP Digital SLR Camera (Body Only)	5	58	58	1.0	1800
4228	Canon EOS Elan IIe Date 35mm SLR Camera (Body ...	5	47	47	1.0	301
3104	Canon EOS Rebel XS 10.1-Megapixel Digital SLR ...	5	35	35	1.0	147

EDA: Review length





EDA: Reviewer behaviour

```
45664110    6
40109303    6
9115336     5
19541636    4
27140716    4
10572690    4
46160224    4
17237889    3
44250386    3
49410262    3
Name: customer_id,
```

```
array(['For the first-timer or family photographer, the Pentax K-S2 is a nice dSLR, but it dr
ops the ball for video and you should pass on the new 18-50mm kit lens when compared to somet
hing like this Nikon here: http://amzn.to/1LnoPdl<br /><br />THE GOOD:<br />+ The Pentax K-S2
offers excellent photo quality, a solid weather-sealed build, a big viewfinder and a broad fe
ature set.<br /><br />THE BAD:<br />- Its video quality is middling to poor.<br /><br />An al
l-round good Pentax camera offering good quality you would expect but should only be purchase
d for the image quality aspect as the video shooting leaves some to be desired. If video will
be important then maybe try something like this Canon here: http://amzn.to/1dPisGK',
       'For the first-timer or family photographer, the Pentax K-S2 is a nice dSLR, but it dr
ops the ball for video and you should pass on the new 18-50mm kit lens when compared to somet
hing like this Nikon here: http://amzn.to/1bA7YcP<br /><br />THE GOOD:<br />+ The Pentax K-S2
offers excellent photo quality, a solid weather-sealed build, a big viewfinder and a broad fe
ature set.<br /><br />THE BAD:<br />- Its video quality is middling to poor.<br /><br />An al
l-round good Pentax camera offering good quality you would expect but should only be purchase
d for the image quality aspect as the video shooting leaves some to be desired. If video will
be important then maybe try something like this Canon here: http://amzn.to/1dPisGK',
       "Like most Leica camara's you will encounter you can be sure they have good build qual
ity and this 10760 is no exception and is only peaked by the version on Amazon seen here:  ht
tp://amzn.to/1GIikkq<br /><br />FOR:<br />+ Dedicated to monochrome shooting<br />+ No filter
```

EDA: Verified purchases

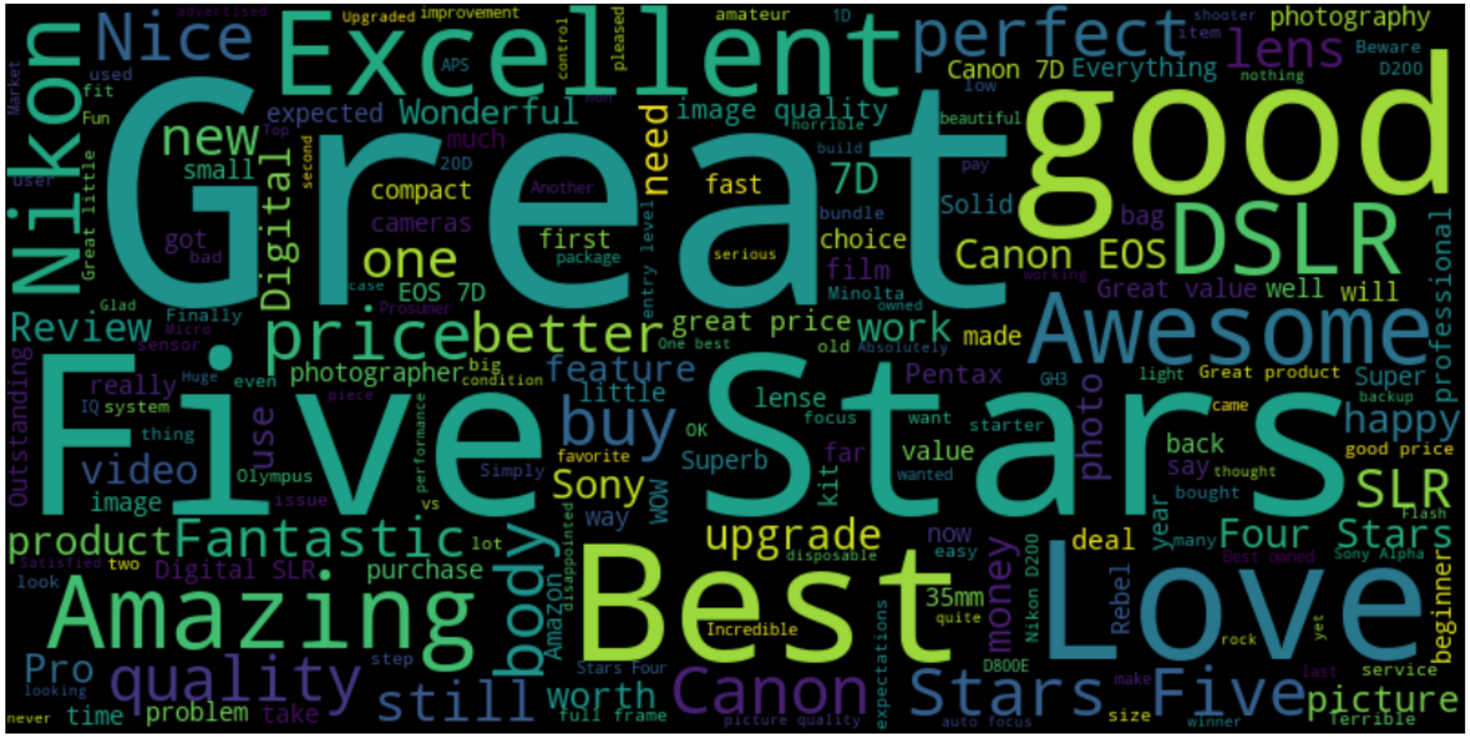
EDA: text features

One star reviews

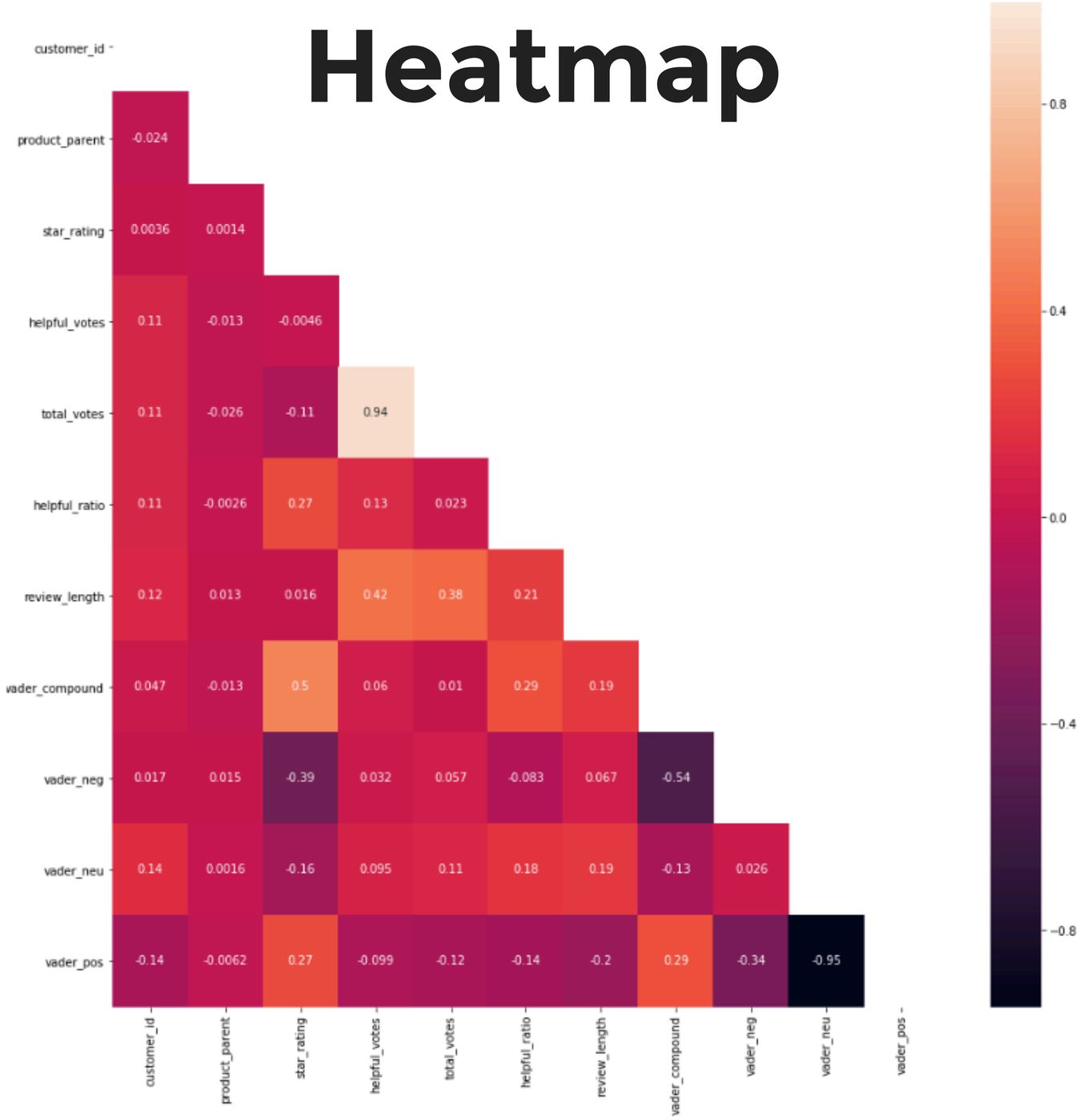
memory card	22
customer service	16
camera body	15
image quality	13
high iso	13
fn iii	13
used camera	12
picture quality	12
hot pixels	12
low light	11
piece junk	9
brand new	9
buy camera	9
auto focus	9
purchased camera	9
take pictures	8
camera one	8
camera work	8
digital camera	8
shutter speed	8
could get	8
take picture	8
camera would	8
received camera	7
white balance	7
sd card	7
got camera	7
use camera	7
low iso	7
camera case	7
dtype: int64	

Five star reviews

low light	457
image quality	416
full frame	331
great camera	328
point shoot	244
auto focus	221
high iso	218
love camera	192
live view	188
kit lens	182
easy use	174
ed af	163
camera body	158
picture quality	154
digital camera	153
much better	152
mark ii	147
white balance	137
dynamic range	136
digital slr	135
battery life	134
shutter speed	133
camera great	123
af dx	120
use camera	120
highly recommend	120
canon eos	119
per second	114
build quality	113
depth field	109
dtype: int64	



Heatmap


customer_id
product_parent
star_rating
helpful_votes
total_votes
helpful_ratio
review_length
vader_compound
vader_neg
vader_neu
vader_pos
-0.024
0.0036 0.0014
0.11 -0.013 -0.0046
0.11 -0.026 -0.11 0.94
0.11 -0.0026 0.27 0.13 0.023
0.12 0.013 0.016 0.42 0.38 0.21
0.047 -0.013 0.5 0.06 0.01 0.29 0.19
0.017 0.015 -0.39 0.032 0.057 -0.083 0.067 -0.54
0.14 0.0016 -0.16 0.095 0.11 0.18 0.19 -0.13 0.026
-0.14 -0.0062 0.27 -0.099 -0.12 -0.14 -0.2 0.29 -0.34 -0.95
0.8
0.4
0.0
−0.4
−0.8

Success Metrics

- Model that performs above the baseline accuracy 0.713.
- F1 score, Precision, Recall, F1.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textit{Recall} = \frac{TP}{TP + FN}$$

$$\textit{Precision} = \frac{TP}{TP + FP}$$

$$\textit{F - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

Star_rating	Value_counts normalized	count
5	0.713547	3034
4	0.153104	651
1	0.057385	244
3	0.050094	213
2	0.025870	110

Model: single feature [review body]

	Logreg CVEC	Logreg TFIDF	MN NaiveBayes	Random Forest	Gradient Boost	KNN	SVM
precision	0.647411	0.652059	0.618355	0.588417	0.548345	0.587785	0.687722
recall	0.710815	0.700627	0.713950	0.702978	0.713950	0.710031	0.651254
f1-score	0.673710	0.674320	0.597667	0.615236	0.597588	0.608114	0.664229
support	1276.000000	1276.000000	1276.000000	1276.000000	1276.000000	1276.000000	1276.000000
Cross_val training score	0.707660	0.714046	0.714048	0.699272	0.713374	0.704303	0.643136
test score	0.710815	0.700627	0.713950	0.702978	0.713950	0.710031	0.651254

F1 vs accuracy

Although the multinomial Naïve Bayes Classifier had the highest cross validation and test accuracy scores. You can see that it predicted 5* nearly every time. This is a problem in imbalanced datasets.

Logistic regression on the other hand had the greatest F1 score which is a weighted average of precision and recall. You can see it has a greater spread of predictions and more correct classifications of the minority classes.

Going forward F1 will be the primary metric of comparison.

Logistic regression Tfidf

	p1	p2	p3	p4	p5
True 1	45	0	3	8	17
True 2	12	0	2	12	7
True 3	11	3	1	23	26
True 4	9	2	10	44	130
True 5	15	0	1	91	804

Multinomial Naive Bayes CVEC

	p1	p2	p3	p4	p5
True 1	1	0	0	0	72
True 2	0	0	0	0	33
True 3	0	0	0	0	64
True 4	0	0	0	1	194
True 5	0	0	0	2	909

Model: multiple feature’s

Features used :

helpful_ratio
review_length
vader_pos
vader_neg
camera_brand
verified_purchase
review_body

Model: multiple feature's TFIDF

	KNN	SVC	Logreg	RandForest	GradBoost	MultinomialNB
precision	0.610031	0.691934	0.649611	0.576510	0.652488	0.509724
recall	0.695925	0.706113	0.719436	0.710031	0.728056	0.713950
f1-score	0.643642	0.697820	0.669140	0.617243	0.631536	0.594795
support	1276.000000	1276.000000	1276.000000	1276.000000	1276.000000	1276.000000
cross_val training score	0.704979	0.694901	0.722777	0.709691	0.733201	0.714720
test score	0.695925	0.706113	0.719436	0.710031	0.728056	0.713950

F1 vs accuracy: again

Similarly to the single feature model. The Gradient Boosting classifier has a higher accuracy score. Predicting primarily 5* though also some 1* predictions.

SVC on the other hand has a greater F1 score and at a cost of accuracy correctly predicts across 2,3 4* classes.

SVC TFIDF

	p1	p2	p3	p4	p5
True 1	44	4	6	8	11
True 2	14	3	5	9	2
True 3	11	5	9	23	16
True 4	10	3	5	60	117
True 5	13	3	16	94	785

Gradient Boost TFIDF

	p1	p2	p3	p4	p5
True 1	20	0	0	2	51
True 2	8	0	0	0	25
True 3	7	1	1	2	53
True 4	2	0	0	3	190
True 5	3	0	0	3	905

Model: multiple features w/ Count Vectoriser

	KNN	SVC	Logreg	RandForest	GradBoost	Multinomial NB
precision	0.596557	0.629360	0.658725	0.588333	0.570702	0.613339
recall	0.705329	0.666144	0.735110	0.709248	0.724138	0.727273
f1-score	0.623883	0.645660	0.686371	0.609829	0.626816	0.642532
support	1276.000000	1276.000000	1276.000000	1276.000000	1276.000000	1276.000000
cross_val training score	0.710002	0.689562	0.735233	0.709342	0.734883	0.723452
test score	0.705329	0.666144	0.735110	0.709248	0.724138	0.727273

Models with highest F1 score

- SVC Tfidf Vectoriser (0.697820)
- Logistic Regression CVEC (0.686371)

SVC TFIDF

	p1	p2	p3	p4	p5
True 1	44	4	6	8	11
True 2	14	3	5	9	2
True 3	11	5	9	23	16
True 4	10	3	5	60	117
True 5	13	3	16	94	785

Logistic regression is making larger errors at prediction of 1* for true 5* reviews and vice versa.

Logistic Regression CVEC

	p1	p2	p3	p4	p5
True 1	44	0	3	6	20
True 2	15	0	3	6	9
True 3	10	1	4	15	34
True 4	5	0	6	30	154
True 5	8	0	2	41	860

Spread of predictions is limited. Let’s try high or low.

Predicting high or low star rating

Creating a new column in the dataset 'binary rating' with the following assigned depending on the star rating.
Low: 1 and 2
High: 4 and 5

The same class imbalance is present which has so far been accounted for using class_weights = balanced. As a hyperparameter and using F1 as the performance metric.

Baseline accuracy for multi classification: **0.912**

Star_rating	Value_counts normalized	count
1.0	0.912355	3685
0.0	0.087645	354

Model: Binary problem w/ Count Vectoriser

	KNN	SVC	Logreg	RandForest	GradBoost	Multinomial NB
precision	0.888478	0.937927	0.952023	0.920469	0.931145	0.934784
recall	0.914191	0.938944	0.952970	0.920792	0.931518	0.938119
f1-score	0.888662	0.938412	0.952455	0.890858	0.912360	0.925321
support	1212.000000	1212.000000	1212.000000	1212.000000	1212.000000	1212.000000
cross_val training score	0.921825	0.934553	0.944815	0.918288	0.926066	0.928197
test score	0.914191	0.938944	0.952970	0.920792	0.931518	0.938119

Model: Binary problem w/ TFIDF

	KNN	SVC	Logreg	RandForest	GradBoost	Multinomial NB
precision	0.925095	0.946223	0.937891	0.921289	0.932601	0.832732
recall	0.932343	0.948845	0.915017	0.929868	0.930693	0.912541
f1-score	0.927220	0.947197	0.923103	0.914009	0.910073	0.870812
support	1212.000000	1212.000000	1212.000000	1212.000000	1212.000000	1212.000000
cross_val training score	0.927490	0.943758	0.916166	0.920409	0.927127	0.912275
test score	0.932343	0.948845	0.915017	0.929868	0.930693	0.912541

Binary Classification: Highest F1 score

- Again the same two models as the multi classification model had the highest F1 scores.
-
- SVC Tfidf Vectoriser (0.947197)
- Logistic Regression CVEC (0.952455)

SVC TFIDF

	pred Low	pred High
True Low	68	38
True High	24	1082

Overall the Logistic Regression model made fewer errors. With fewer False Positives, Though two more False negatives.

Logistic Regression CVEC

	pred Low	pred High
True Low	75	31
True High	26	1080

Resampling

- In an imbalanced dataset there is more likelihood of predicting the majority class to improve accuracy.
- Resampling methods such as under sampling, over sampling or generating synthetic samples were trialled.
- The best results were from removing Tomeklinks with the logistic regression model with a small improvement in both accuracy and F1 scores.

Logistic Regression CVEC

	pred Low	pred High
True Low	75	31
True High	26	1080

Logistic Regression CVEC With Tomeklinks resampling

	pred Low	pred High
True Low	77	29
True High	25	1081

Misclassified results

Themes emerging from misclassified reviews:

True low / predicted high
- Excited at first but then disappointed.
- Satisfied with camera but an important downside (price, warranty or customer service)

True high / predicted low
- Big problem that was resolved. e.g replacement camera
- Similarly to above, satisfied with results but experienced a difficulty
- 'No complaints' terms like this.

Some words that may have give a negative sentiment emotionally but had low reviews.

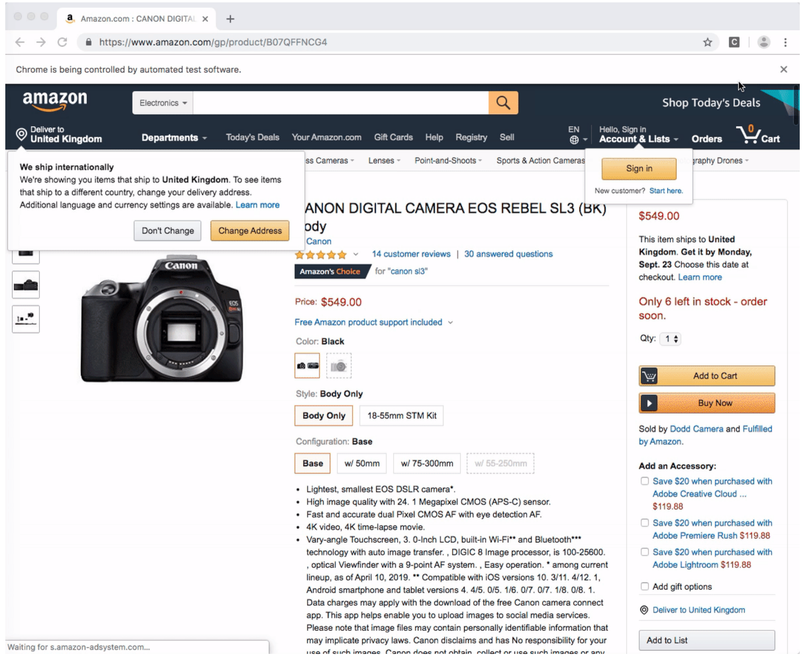
- Technical errors e.g motion stutter
- Product experience: Unusable, defect, unreliable
- Return e.g sent back to seller

	actual	predicted	mismatch	TEXT
138	1.0	1.0	False	Amazing upgrade to my T1i. I'm able to grab fo...
3500	1.0	1.0	False	The Canon 10D is a good camera as far as it's ...
3521	1.0	1.0	False	Impressive does not begin to contend for the i...
2192	1.0	1.0	False	Being a Minolta fan for some years, I one day ...
219	1.0	1.0	False	awesome camera. my uncle let me borrow his wit...
1111	1.0	1.0	False	excellent quality and meets all my expectation...
557	1.0	1.0	False	What a simply amazing camera. My love of photo...
3927	1.0	1.0	False	This is a great 35MM SLR. It provides a very g...
2242	1.0	1.0	False	It us a good deal for the meager price you pay...
3802	1.0	1.0	False	I'm very happy with this camera and with the S...
79	1.0	1.0	False	Takes amazing photos, video's are shaky but ve...
1815	1.0	1.0	False	I picked my Rebel T3 / 1100D around Christmas ...
1476	1.0	1.0	False	I got this as a gift for my husband. He's very...
616	1.0	1.0	False	epic camera !
1191	1.0	1.0	False	I've owned this camera for about one month and...
1581	1.0	1.0	False	I initially had a Canon rebel XT and was wanti...
1947	1.0	1.0	False	I love my bag, especially the color! My bag is...

Scraping current reviews (investigating class imbalance)

```
def product_scraper(base_url, page):

    asin = []
    url = []
    product_name = []
    review_count = []
    avg_rating = []
    prices = []
    products = {}

    #FIRST PAGE


    driver = webdriver.Chrome(executab
    driver.get(base_url)
    #Save page source into Beautiful S
    Soup = BeautifulSoup(driver.page_s
    #select each product listing
    first_page_results = Soup.select('
```

```
def review_scraper(asin, reviewpages):

    #Get URL and goto page
    driver = webdriver.Chrome(executable_path='./ch
    driver.get('https://www.amazon.com/gp/product/{


    #In review section scrape all details
    profile_name = []
    user_url_id = []
    review_date = []
    star_rating = []
    review_title = []
    helpful_votes = []
    total_comments = []
    verified_purchase = []
    review_body = []
    review_id = []

    def review_scraper_inner():

        #Scrape page source
        Soup = BeautifulSoup(driver.page_s




    for review in Soup.find_all(attrs=
```

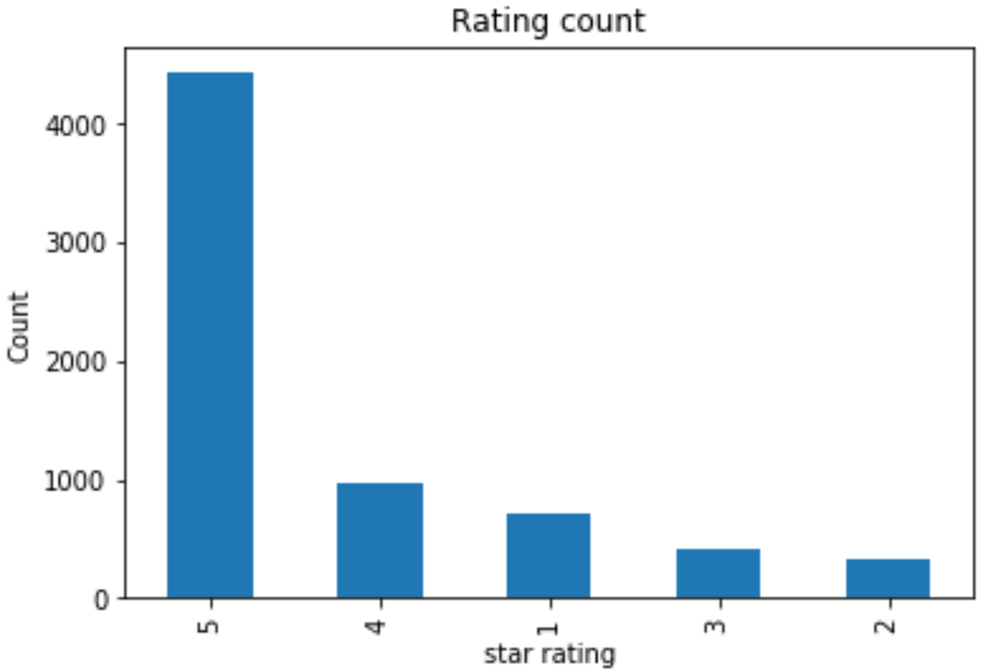


Web Scraping (6864 recent reviews)

```
dslr = pd.read_csv('Scraped_data/DSLR.csv')
mirrorless = pd.read_csv('Scraped_data/MIRRORLESS.csv')
action = pd.read_csv('Scraped_data/ACTION.csv')
pos = pd.read_csv('Scraped_data/POINTANDSHOOT.csv')
```

```
In [91]:
1 df_scraped_combined = df_scraped_combined[df_scraped_combined.star_rat
2 df_scraped_combined.shape
```

```
Out[91]: (6864, 11)
```



Key takeaways:

- Prioritise features: low light, image quality, full frame, auto focus in new cameras
- Focus on fixing: memory card, customer service, camera body, image quality, high iso
- Amazon reviews are skewed towards high ratings.
- F1 is a more suitable metric in multi class problem due to the class imbalance.
- Binary predictor performs fairly well: 0.9551 F1 score and 0.9554 accuracy score. Greater than the baseline of **0.912**.

Limitations:

- Further parameter tuning required
- Text features are noisy. Further preprocessing like sentence stemming possible.
- Class Imbalance in data
- How legitamate are non-verified reviews? Did they actually purchase?

Next Steps:

- Further feature engineering
- Consider multiple camera types and price points on scraped dataset

Thank you.