Recommender System Using the Amazon Dataset

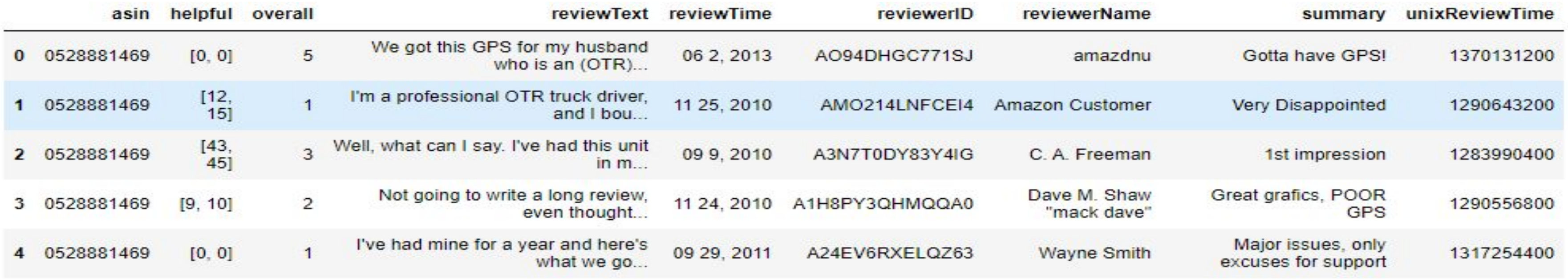
By: Thompson Pham

Data Wrangling

- The Dataset came from a USCD Professor named Julian McAuley
  - She provides an entire Amazon dataset as well as various smaller subsets of the data.
- The dataset was stored locally and imported into python
  - .read_json() method
- Dropped redundant columns
  - reviewerID vs reviewerName
  - reviewTime vs unixReviewTime
- Changed data type of some fields
  - reviewTime from object data type to a DateTime data type
- Split up the helpful column
  - Separate columns for found helpful and total helpful
- Changed names of columns
  - Asin to itemID
  - Overall to rating

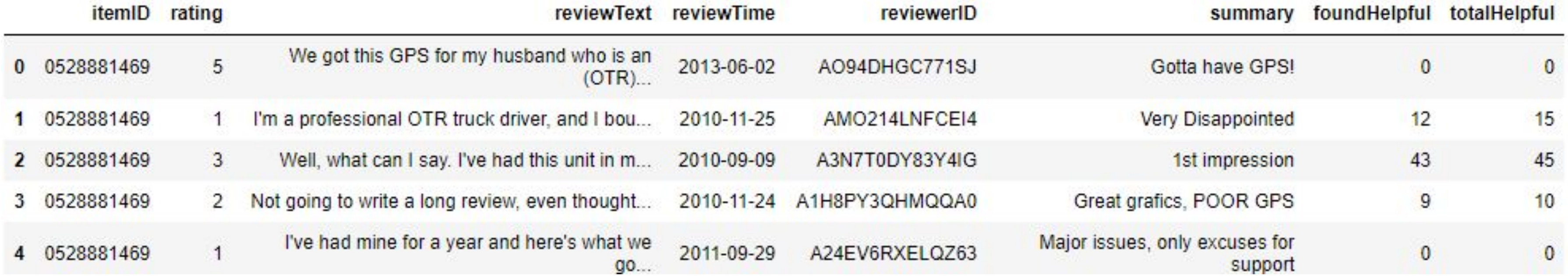
Data Wrangling

Before Cleaning:



	asin	helpful	overall	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime
0	0528881469	[0, 0]	5	We got this GPS for my husband who is an (OTR)...	06 2, 2013	AO94DHGC771SJ	amazdnu	Gotta have GPS!	1370131200
1	0528881469	[12, 15]	1	I'm a professional OTR truck driver, and I bou...	11 25, 2010	AMO214LNFCEI4	Amazon Customer	Very Disappointed	1290643200
2	0528881469	[43, 45]	3	Well, what can I say. I've had this unit in m...	09 9, 2010	A3N7T0DY83Y4IG	C. A. Freeman	1st impression	1283990400
3	0528881469	[9, 10]	2	Not going to write a long review, even thought...	11 24, 2010	A1H8PY3QHMQQA0	Dave M. Shaw "mack dave"	Great grafics, POOR GPS	1290556800
4	0528881469	[0, 0]	1	I've had mine for a year and here's what we go...	09 29, 2011	A24EV6RXELQZ63	Wayne Smith	Major issues, only excuses for support	1317254400

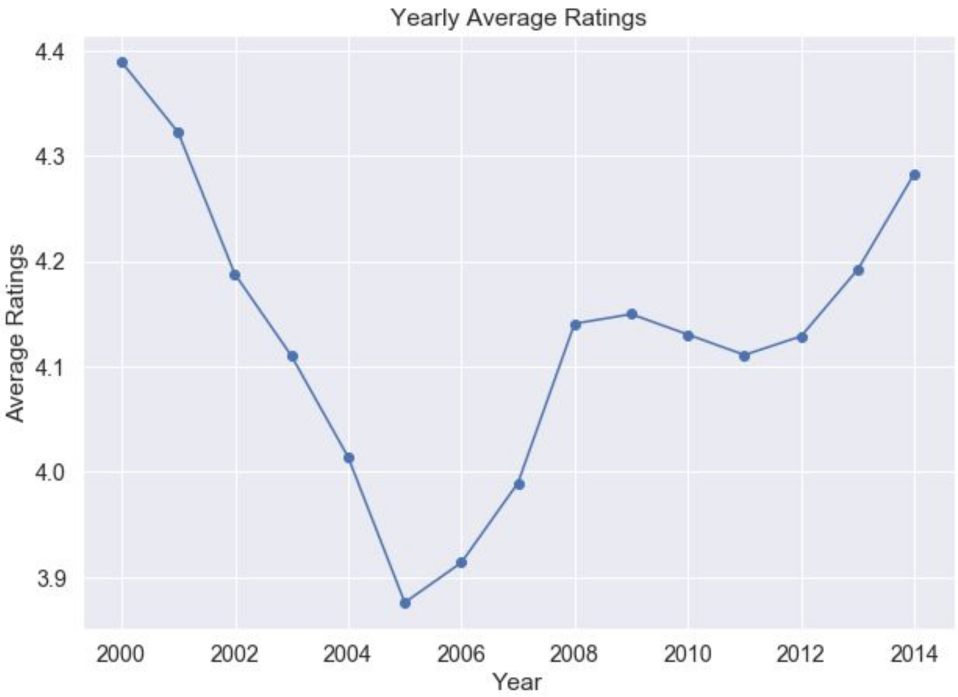
After Cleaning:



	itemID	rating	reviewText	reviewTime	reviewerID	summary	foundHelpful	totalHelpful
0	0528881469	5	We got this GPS for my husband who is an (OTR)...	2013-06-02	AO94DHGC771SJ	Gotta have GPS!	0	0
1	0528881469	1	I'm a professional OTR truck driver, and I bou...	2010-11-25	AMO214LNFCEI4	Very Disappointed	12	15
2	0528881469	3	Well, what can I say. I've had this unit in m...	2010-09-09	A3N7T0DY83Y4IG	1st impression	43	45
3	0528881469	2	Not going to write a long review, even thought...	2010-11-24	A1H8PY3QHMQQA0	Great grafics, POOR GPS	9	10
4	0528881469	1	I've had mine for a year and here's what we go...	2011-09-29	A24EV6RXELQZ63	Major issues, only excuses for support	0	0

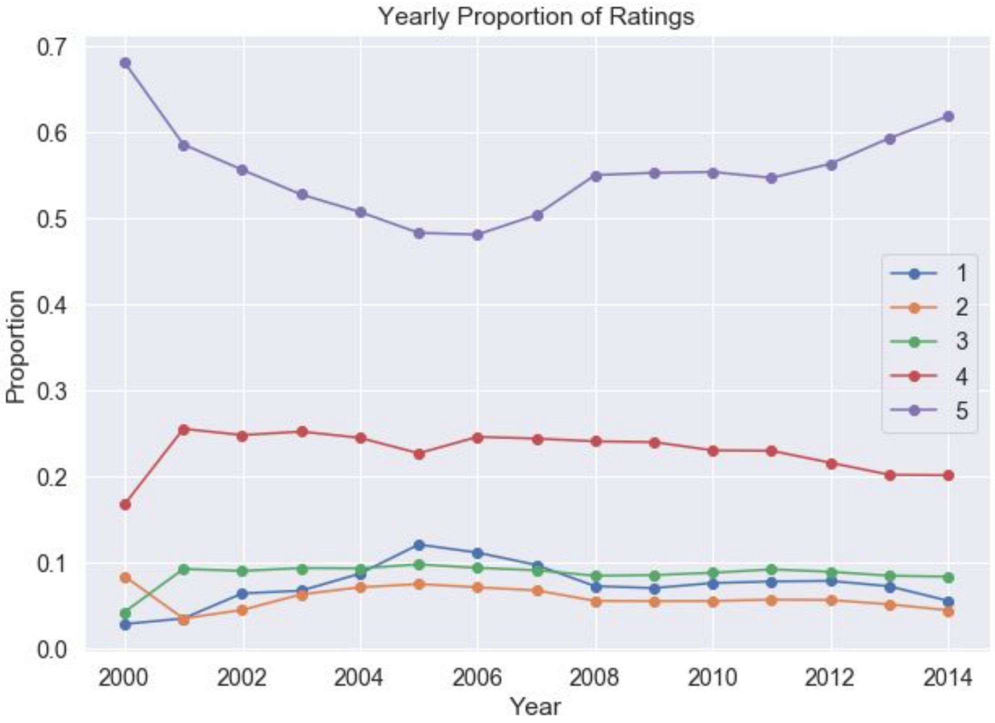
Exploratory Data Analysis

- Plot of the average ratings over the years
- Very dynamic movement
- The lowest rating occurred in 2005 with an average value of about 3.88



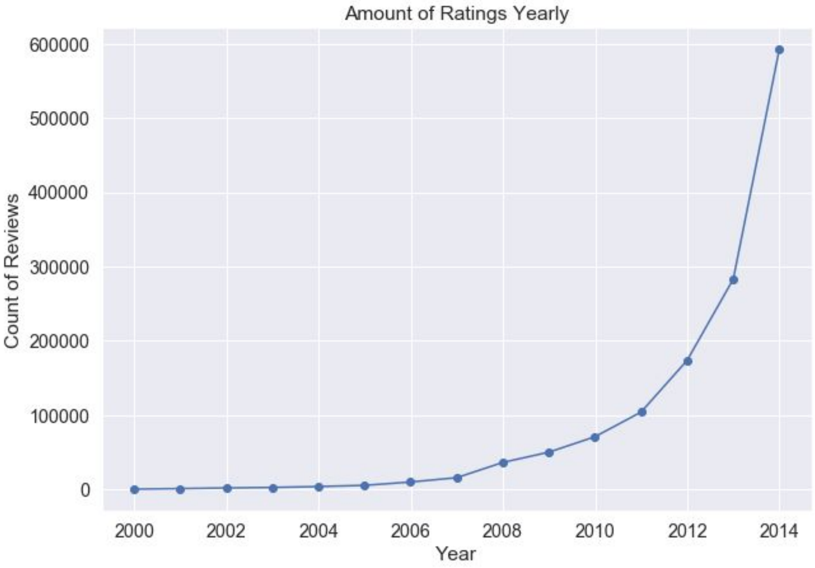
Deeper Analysis

- Split the average ratings up into proportion of ratings over the years
- The rating of 5 is always the majority proportion
- The ratings of 5 and 1 are the most dynamic
- The lowest average rating in 2005 can be seen from the lowest proportion for 5 and the highest proportion for 1
- Relative to the other ratings, the ratings of 2, 3, and 4 do not change much in proportions over the years



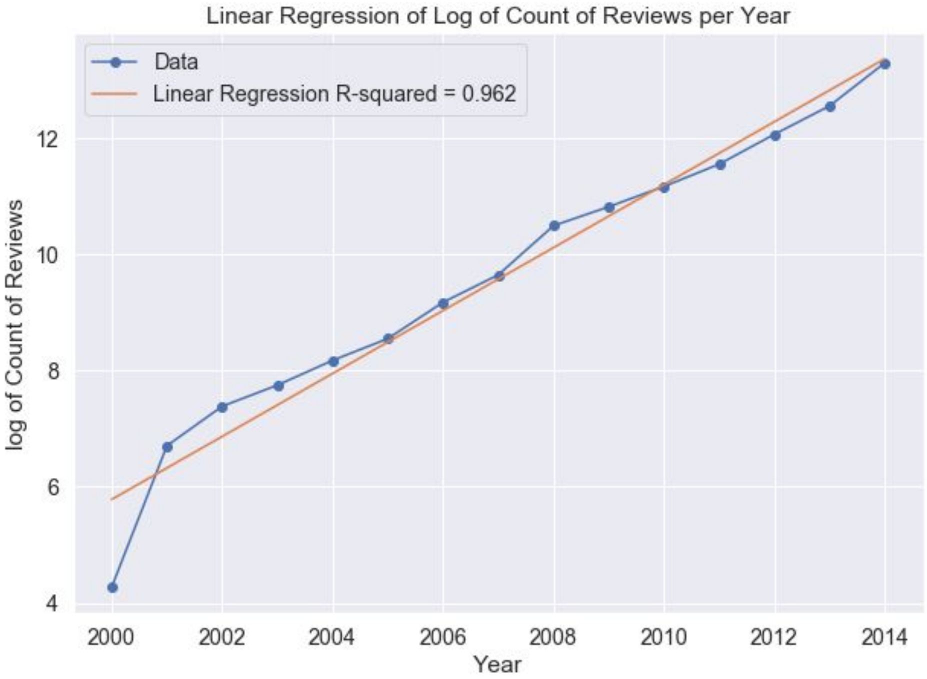
Exploratory Data Analysis

- Plot of the count of reviews given over the years
- Follows an exponential curve
- Would like to make predictions on the plot
- Performing a log transformation on the count of reviews will change it into a linear plot
- Allows for linear regression to fit a line to the plot



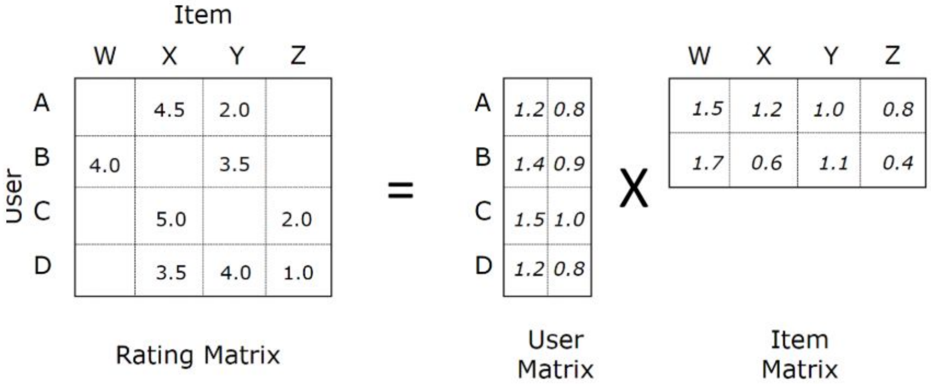
Deeper Analysis

- The fitted line had an R-squared value of 0.962
  - ~96% of the variance in the log of Count of Reviews could be explained by the years value
- Linear regression line allows for predictions on the growth of the count of reviews over the years



Recommender System

- Matrix Factorization
- Factorizes a large User x Item matrix into two smaller, separate Users and Items matrix
- Performing the dot product on the separate matrices will recreate the original matrix
- The separate users and items matrix has latent feature values that can be updated to best predict the ratings based on the original matrix
- Can predict user/item pairs of ratings that did not originally exist

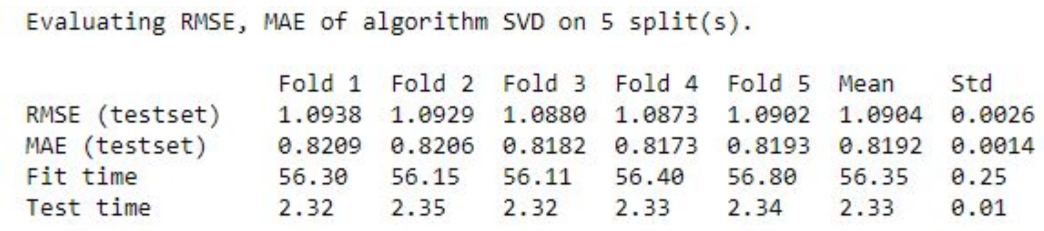


$$\begin{array}{c|cccc} & W & X & Y & Z \\ \hline A & & 4.5 & 2.0 & \\ B & 4.0 & & 3.5 & \\ C & & 5.0 & & 2.0 \\ D & & 3.5 & 4.0 & 1.0 \end{array} = \begin{array}{c|cc} A & 1.2 & 0.8 \\ B & 1.4 & 0.9 \\ C & 1.5 & 1.0 \\ D & 1.2 & 0.8 \end{array} \text{X} \begin{array}{cccc} W & X & Y & Z \\ \hline 1.5 & 1.2 & 1.0 & 0.8 \\ 1.7 & 0.6 & 1.1 & 0.4 \end{array}$$

https://medium.com/@connectwithghosh/simple-matrix-factorization-example-on-the-movielens-dataset-using-pyspark-9b7e3f567536

Recommender System

- Used the scikit Surprise package to implement the recommender system
- The Surprise SVD algorithm closely resembles matrix factorization
- Fitted the model on the Amazon dataset
  - 5-fold cross-validation
- Results of cross-validation shows model performed equally well on the 5 splits of the dataset

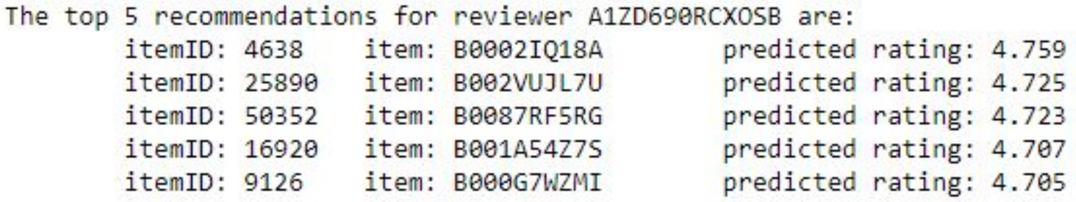


```
Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    1.0938  1.0929  1.0880  1.0873  1.0902  1.0904  0.0026
MAE (testset)     0.8209  0.8206  0.8182  0.8173  0.8193  0.8192  0.0014
Fit time          56.30   56.15   56.11   56.40   56.80   56.35   0.25
Test time         2.32    2.35    2.32    2.33    2.34    2.33    0.01
```

Recommender System

- The fitted model allows for predictions of ratings for any reviewer/item pair
- For any specific reviewer, all the unrated items can have their ratings predicted
- The top 5 items with the highest predicted ratings can be recommended for that specific reviewer



```
The top 5 recommendations for reviewer A1ZD690RCXOSB are:
        itemID: 4638    item: B0002IQ18A        predicted rating: 4.759
        itemID: 25890   item: B002VUJL7U        predicted rating: 4.725
        itemID: 50352   item: B0087RF5RG        predicted rating: 4.723
        itemID: 16920   item: B001A54Z7S        predicted rating: 4.707
        itemID: 9126    item: B000G7WZMI        predicted rating: 4.705
```

Summary

- Saw that the average ratings over the years could be mostly explained by the proportion of ratings that are 5 and 1

- The count of reviews over the years followed an exponential curve. Linear regression was performed on the log transformed plot and a fitted line was able to explain about 96% of the variance

- Matrix factorization was applied to the Amazon dataset using the scikit Surprise package SVD algorithm.
    - The model performed equally well on different subsets of the data
    - For any arbitrary reviewer, the top 5 items with the highest predicted ratings can be recommended