

**Recommendation System**

Dipeeka Bastola

Jayson Faulds

Jina Kim

**I . Introduction: Overview and Motivation**  
With online retail growing rapidly, many firms have been adapting recommendation systems to suggest products to users to generate higher sales. Among many companies, Amazon is the largest online retail platform, which is valued highly for its product and service recommendations to its customers through heavy data-driven marketing approaches over the past decades. Amazon currently uses item-to-item collaborative filtering that produces recommendations for its users. Since Amazon possesses large amounts of data on its customers and products, performing network analysis on Amazon’s customer and rating network would provide opportunities for improvements and enhancements to its current recommendation system. Through analysis of ratings in the consumer and product networks, valuable insights might be uncovered that might go unnoticed otherwise.

**Research Questions**

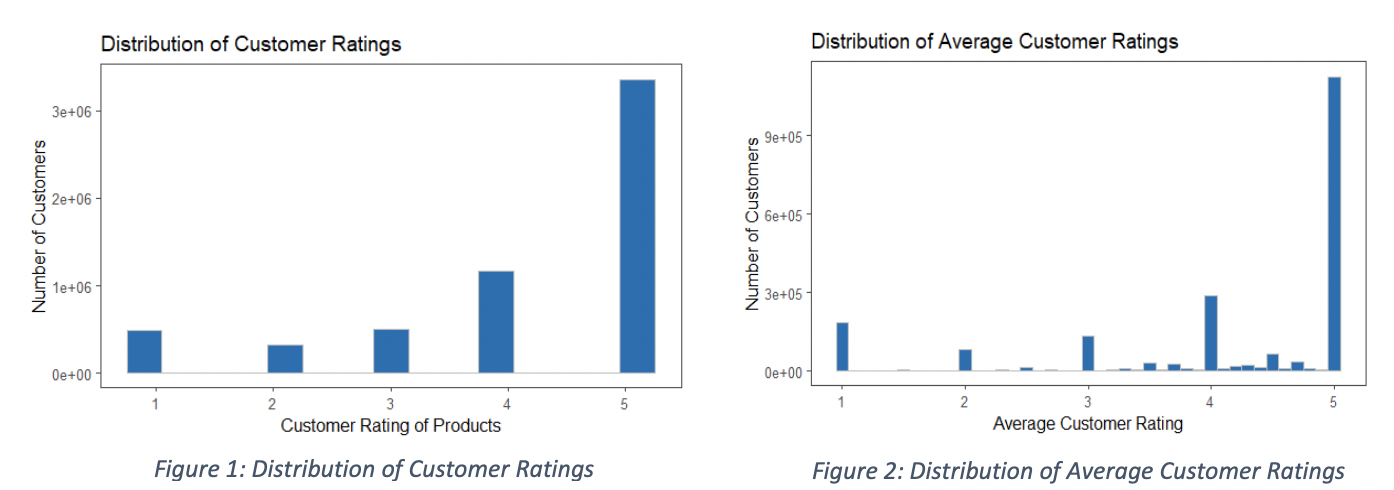
This research focuses on examining the relationship between customers’ purchasing behaviors to their preferences and future purchasing habits. Our first research question is whether the customers who bought and rated similar products in the past are more likely to buy similar products in the future as well. The second question is whether customers with co-purchasing history of similar items are likely to give out similar ratings, indicating similar preferences. These questions will be answered through dyad level analysis of the customer network using purchasing history, behavioral changes of customers and time of the purchase. Along with predictive analytics to answer these questions, network analysis is performed to answer a third question of how ratings are impacted by centrality measures in the user and product networks. Identifying customer behaviors through these research questions and network analysis enables Amazon to personalize recommendations based on ties in the network and recognize any potential marketing opportunities.

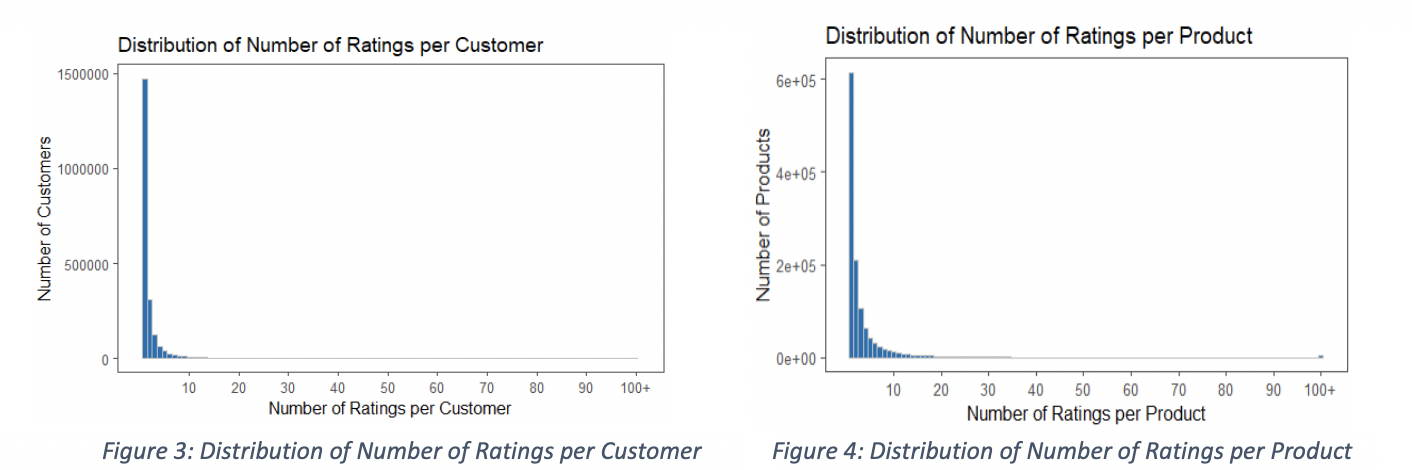
**II. Background and Data Overview**

To pursue this research, Amazon Rating dataset was acquired from Network Repository. The original dataset spans 10 years from 1996 to 2006, with over 5.8 million ratings of 1.2 million products by 2.1 million Amazon users. It is in bipartite network format, which contains an edgelist with nodes as customers and products, ratings of products (edge weights), and timestamp of when the purchase and rating was made. Since the original data has unix timestamp for each rating, it had to be converted into dates. A summary of the dataset used for this research is shown below.

*Table 1: Data Overview*

|  |
| --- |
| Data Overview |
| Total Observations 5,838,041  Number of Users who rated 2,146,057  Number of Products 1,230,915 |

***Exploratory Data Analysis:*** Taking a closer look at the dataset, some interesting insights were discovered through Exploratory Data Analysis. In *figure 1* which shows distribution of customer ratings in the dataset, we can clearly see that the distribution is skewed towards higher ratings. The average rating that users give out is 4.15 and the median is 5 (Out of 5). Most customers who rate products tend to give out very high ratings, which can be interpreted that customers mostly rate items that they were satisfied with. 

Through further exploration on the distribution of average customer ratings in *figure 2*, it is apparent that average ratings also skew towards higher ratings. The distribution barely has any variation from *figure 1*, which indicates that most users do not give out many ratings in general. Similarly, there are very few bars between the whole numbers indicating that most customers rate only a single product or tend to give similar ratings. Taking a closer look at rating distribution on customer and product level in *figures 3* and *4*, it can be seen that the number of ratings per customer and product are both very skewed to the right. 

On average, users rate 2.7 products with median number of ratings as 1. This being said, the inference made earlier from *figure 2* is further supported. *Figure 4* also exhibits a similar trend, but on the product level. Most products have less than 10 ratings, with an average of 4.7 ratings per product. This may be due to the numerous amount of products Amazon carries and the tendency of customers to not rate products in general. Such discoveries of user rating behaviors informed us that the users in the Amazon rating dataset may be disconnected to some degree, considering the low number of ratings each user gives and each product receives. This implies that finding the influence of a node in the network, using measures such as closeness may not be effective. These disconnected graphs, however, provide opportunities to analyze the missing ties using the property of transitivity.

***Data Transformation:***  Some new variables were created to be used for the regressions. The first variable created was the number of products two customers rated in a year, which is the sum of all products every pair of customers bought in a particular year. The second variable created was total number of products two customers have purchased and rated in common as of the prior year. This calculation was made by lagging the cumulative number of products rated till the current year. Finally, a weight variable for recent co-purchases was created. The most recent co-purchase made by a customer-customer pair is represented by weights, scaling from 1 to 10. The indexing process was done by finding the difference between the year of latest co-purchase and the second to latest co-purchase history. If the last co-purchase was from the prior year, a weight of 10 was assigned, and as the difference gets bigger, which represents an older purchase history, lower weight was assigned.

*Table 2: New Variables Overview*

|  |
| --- |
| New Variables Overview |
| Number of co-purchased/rated products  Cumulative number of co-purchased/rated products till prior year  Weight variable for recent co-purchases |

**III. Analysis and Results**  
***Regression predicting number of products co-purchased on dyad level***: Customers rating data on a dyad level was used to analyze whether customers that bought similar products in the past are more likely to buy similar products in the future. A regression was performed to analyze this relationship. The dependent variable in the regression was total number of products two customers bought until prior year. This variable would capture the customers’ co-purchase and rating history. However, a customer’s preferences might change over time, or customers might have new life events that may change their purchase behavior (e.g: having a baby). In order to capture this effect, a new variable was created to weigh purchases from 1 to 10 where 10 is assigned to the most recent purchase/rating while 1 is assigned to the oldest purchase. Finally, a linear control for year was added to the regression to control for time effects. The dataset used in this analysis consisted of ties between customers that had made a co-purchase in the past 3 years (network with tie decay). This was done to ensure that the regression is not capturing co-purchases that were made many years ago. Generalized Linear Model (GLM) regression with negative binomial was used because our dependent variable is a count variable that is heavily right skewed.

*Table 3: Regression output predicting number of products co-purchased on a dyad level (Appendix 1)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Formula: glm.nb ( Number of co-purchased products ~ Cumulative co-purchases till prior year + Weight variable for recent purchases + Year)** | | | | |
| Estimate Std. Error t value Pr(>| t |) | | | | |
| (Intercept) | 4.956e+01 | 1.472e+00 | 33.66 | <2e-16 \*\*\* |
| Cumulative co-purchases till prior year | 3.893e-02 | 3.616e-05 | 1076.54 | <2e-16 \*\*\* |
| Weight variable for recent purchases | 1.118e-01 | 1.719e-03 | 65.03 | <2e-16 \*\*\* |
| Year | -2.506e-02 | 7.339e-04 | -34.15 | <2e-16 \*\*\* |

As seen from the regression output above, all variables are statistically significant. The positive coefficient for cumulative co-purchases till prior year signifies that customers who have co-purchase and rating history tend to buy more of the same products in the future. Similarly, the positive coefficient on weight variable shows that most recent purchases are more likely to predict than older purchases. The coefficient for year variable is negative. This is most likely due to the fact that Amazon’s product portfolio has massively increased over the years; even in the dataset there were 10K unique products in 1997, which increased to 230K unique products in 2006. Variety of products in the same category means that two customers might buy a similar product (e.g: razors) but different brands or SKUs within same brands. Overall, this regression shows that two customers are more likely to buy and rate products if they have bought more similar products in the past, with most recent purchases affecting the co-purchase more than older co-purchases.

***Network Analysis on both product-user level and user-user level***:

|  |  |
| --- | --- |
| Figure 5: Network Graph of Bipartite Network  (Product - Customer Network) | Figure 6: Network Graph of Co-Affiliation Network  (Customer - Customer Network) |

Product - Customer Network: The original data came in the format of a bipartite network where customers and products were the nodes and an edge was rating from a customer to a product. A node level analysis was performed on products to better understand how centrality measures such as degree and eigen centrality affect the ratings received by a product. In this network, degree of a product represents the number of reviews it received and eigen centrality is measured based on how many ratings the neighbors of a product receives. The **correlation** between the degree and eigen centrality is very low (**0.012).** This is most likely due to lack of common customers among products, which results in a more disconnected graph with less neighbors for each node. This can be seen in *Appendix 4*, where most eigen centrality values are close to zero, with very few dispersed throughout the graph. The distribution of degree and average rating were then analyzed which were both heavily skewed. Taking log of degrees made values somewhat more normal, while data transformation did not really affect the distribution of average ratings.

*Table 4: Regression output predicting average rating based on degree of products (Appendix 5)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Formula: lm (Average product rating ~ Degree of products)** | | | | |
| Estimate Std. Error t value Pr(>| t |) | | | | |
| (Intercept) | 4.232e+00 | 9.199e-04 | 4599.96 | <2e-16 \*\*\* |
| Degree of products | -4.953e-04 | 2.689e-05 | -18.42 | <2e-16 \*\*\* |

As seen from the regression output above, the higher the degree of a product, the lower the average ratings. This implies that as a product receives many ratings, the customer sentiments differ widely. More number of ratings could also mean product appealing to masses vs. niche products that might have higher ratings. Another regression was run to predict average rating of product based on log of degrees after transformation. The results were similar to the table above, where degree was statistically significant with a negative coefficient (refer to *Appendix 6* for output).

Customer - Customer Network: Customer to customer is a co-affiliation network created by transforming the original dataset, where nodes are customers and an edge connects two customers based on a product co-purchased and rated. This network allows for a dyad level analysis on customers that forms the basis of our recommender system proposed later. **Tie decay** was performed for customer relationships if two customers had not purchased a common product in over 3 years. This was done to capture customers’ changing purchasing habits and preferences. On node level, centrality measures such as degree and eigen centrality were calculated. The degree represents the number of ratings given by a particular user; eigen centrality represents influence in the form of neighbors that give out more ratings. The **correlation** between degree and eigen centrality in the network is very low **(0.0015**), which is due to the sparse relations between customers in the network. Before predicting average customer ratings, eigen centrality was transformed by taking a log to make the distribution more normal.

|  |  |
| --- | --- |
| Figure 7: Distribution of customer eigen centrality (original scale) |  |

*Table 5: Regression output predicting average rating by user (Appendix 7)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Formula: lm (Average rating by customer ~ Degree of customers + Log (Eigen Centrality)** | | | | |
| Estimate Std. Error t value Pr(>| t |) | | | | |
| (Intercept) | 3.902e+00 | 1.479e-01 | 26.383 | <2e-16 \*\*\* |
| Degree of customers | 2.704e-05 | 2.304e-06 | 11.738 | <2e-16 \*\*\* |
| Log (Eigen centrality of customers) | -1.730e-02 | 1.285e-02 | -1.347 | 0.178 |

As seen from the table, only degree of customers is statistically significant and positive, meaning customers that give out more ratings tend to give high ratings. This could be an indication that customers that give one or two ratings might only do so when they dislike a product.

In addition to linear regression, loess curves were plotted to predict average rating based on degree of both users and products. As seen from the charts below, average ratings generally stay within 3 to 5 with an increase in average ratings after a tiny dip as degree increases.

|  |  |
| --- | --- |
|  |  |

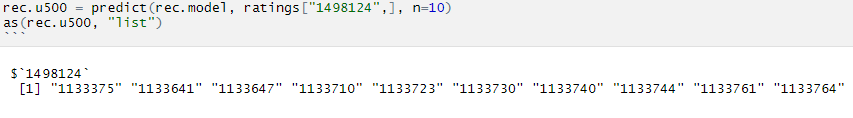
***Tie Closures***: For a network level analysis, identifying missing ties in the customer network could represent opportunity for Amazon’s recommender system. The measure used to analyze tie closures is transitivity. Transitivity is used because most customers rate one or two products and this tie is important to predict possibilities of future ties. Transitivity looks for closed triangular relationships; for instance, if two customers A and B are both connected to C (bought and rated the same products as C) but are not connected to each other, this might represent an opportunity to market products to both A and B. **Transitivity** for the customer network with decay is **0.339** meaning around **34%** of all customers form triadic closures while **66% of the network triads lack closure** . High proportion of missing ties among customers also motivated the development of a recommender system, as covered in conclusions and implications section.

***Regression predicting difference between customer ratings***: The research question for this regression is whether customers who have co-purchase and rating history are more likely to give products the same rating, meaning they both like or dislike the product. The hypothesis here is that a co-purchase relationship leads to similar ratings between customers. For example, we would expect two people that have purchased a lot of similar products in the past to give similar ratings on a new product today. The dependent variable is the difference in rating between two customers (Customer A giving a 5 and Customer B giving a 4 would yield a difference of 1). The independent variables are the total number of products rated by two customers up to the previous year and time since the most-recent co-purchase, with year as a control variable. Note: these are the same predictors as the regression fromour first research question above*.* A linear least squares model was used for the regression. A square root transformation to the dependent variable was applied, in an effort to make our dependent variable more normally-distributed for more accurate analysis. Relative strength and direction of our predictor coefficients remain similar to the untransformed version, so those results are displayed here. Regression output for the transformed version can be found in *Appendix 3.*

*Table 6: Regression output: difference in customer ratings*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Formula: lm ( Difference in ratings ~ Cumulative co-purchases till prior year + Weight variable for recent purchases + Year)** | | | | |
| Estimate Std. Error t value Pr(>| t |) | | | | |
| (Intercept) | -6.764e+01 | 1.079e+00 | -62.70 | <2e-16 \*\*\* |
| Cumulative co-purchases till prior year | -8.860e-04 | 7.063e-05 | -12.54 | <2e-16 \*\*\* |
| Weight variable for recent purchases | -3.697e-02 | 1.116e-03 | -33.12 | <2e-16 \*\*\* |
| Year | 3.427e-02 | 5.377e-04 | 63.73 | <2e-16 \*\*\* |

From the regression output table above, it can be seen that all of our predictor variables are statistically significant. There is a negative effect on ratings differences when customers have co-purchased a larger number of products. Put more simply, customers tend to rate things similarly today when they buy similar products in the past. This is a significant effect, as it provides evidence that customers who buy similar products will tend to rate them similarly as well. If two customers have co-purchased a large amount of products, and one customer gives a product a 5 rating, we suspect that the other customer will like that product as well. Knowing this, we feel more comfortable creating a recommendation system based on customer ratings, which we will do in the next section. Our weight variable for recent purchases holds a negative coefficient as well. This implies that customers who have *recently* made co-purchases are more likely to give similar ratings. Customers that last made a co-purchase 8 years ago might not be very similar today, which could lead to different customer experiences and ratings. For a recommender system, this means we value more-recent customer-customer ties, as these pairs are more likely to rate products the same way. Finally, the year predictor has a positive coefficient, which implies that in later years, customers will rate products more differently. We interpret this as the following: Amazon’s product portfolio is expanding over time, and with this expansion, there are more products available to customers. There may even be many different options for a desired product. The addition of these options may have an impact on how individuals rate different products.

**IV. Conclusions and Implications**  
***Recommendation System:*** With the results from the research questions above, we can feel confident in building a recommender system, despite having very little information regarding products. For our recommender system, we have filtered the data so that it only contains information regarding products purchased and rated in the Winter of 2016. By limiting our products to the winter season, we can be more confident that our product recommendations are seasonal in nature (i.e. winter coats, Christmas presents, etc.). This means our recommender system is more likely to propose new products to customers that are appropriate for the current season. For the system to work, the data needed to be coerced into an affiliation matrix, with the customer as rows and product as labels. Every time a customer rated a product, and a match is made in the matrix, the cell value is filled with the rating given by the customer. Every time there isn’t a match (the customer did not buy and rate a product), the cell is filled with NAs. The recommender system will use a mixture of UBCF and Jaccard similarities to recommend products to different customers. Ultimately, we pass a specific customer into the function, and it will tell us the top 10 products to recommend to that customer (Seen below is an example for customer 1498124). These recommendations are largely determined by the ratings given out by this customer, as well as the ratings given out by similar customers. *Figure 1: Product recommendations from the model for a customer (ID 1498124)*

Finally, we applied the concept of data mining to assess which similarity methods are best to use. To do this, we used 90% of the affiliation matrix to train our recommender system, and used the remaining 10% to test how accurate it was. The model predicts what ratings were given out by customers, and we compare these predictions to what the actual ratings were. As can be seen by the error matrix below, the Jaccard method had the lowest error rates in a few different error measures.

*Table 7: Comparison of errors (Appendix 8)*

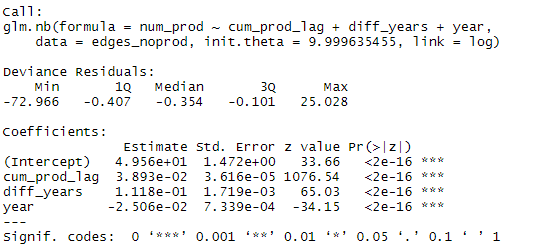
|  |  |  |  |
| --- | --- | --- | --- |
| RMSE MSE MAE | | | |
| Cosine | 1.096 | 1.201 | 0.775 |
| Jaccard | 0.932 | 0.869 | 0.659 |
| Pearson | 1.313 | 1.725 | 1.313 |

***Limitations:*** Since the dataset obtained had only four columns with numbers for unique customer and product IDs, the scope of analysis was limited. First, the lack of names for products limits the recommendations to product IDs, which is not very meaningful for analysis. However, for Amazon that has this product information, the prediction on product ID level might be meaningful. Similarly, lack of categories for products prevents the recommendation system or regression from grouping products instead of SKU (product ID level). Further, there is no information on price or sales volume for a product that could provide information on product sales, profitability and ROI. Having sales volume and profitability information could further enhance the recommender algorithm to recommend items to customers that are profitable but are not meeting volume targets. Finally, a limitation of this dataset was its size that significantly increased the computation time and power required for analysis. While this might be a limitation for this analysis, this is an opportunity for Amazon as they have the infrastructure and ability to process big data and discover insights.

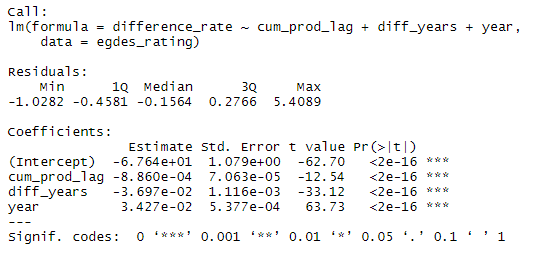
***Conclusion and Business Implication***: Despite the limitations, network data can be used to further enhance Amazon’s recommender system. Capturing behaviors and connections between users and similarity in their behavior can enhance Amazon’s recommender algorithm. We believe applying network analyses as above, coupled with Amazon’s more extensive product information from their databases, can lead to greater sales overall for Amazon. Additionally, the benefits from recommending a good product to a customer extend beyond just profits; it increases the likelihood that they trust Amazon recommendations in the future. This helps maintain a positive feedback loop, where Amazon makes better recommendations over time, and customers in turn will act on these recommendations by purchasing the product.

**Appendices:**

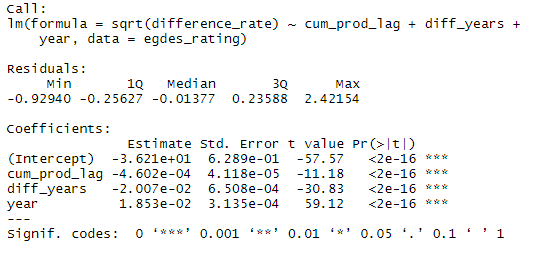
**Appendix 1: Regression predicting number of co-purchased products**



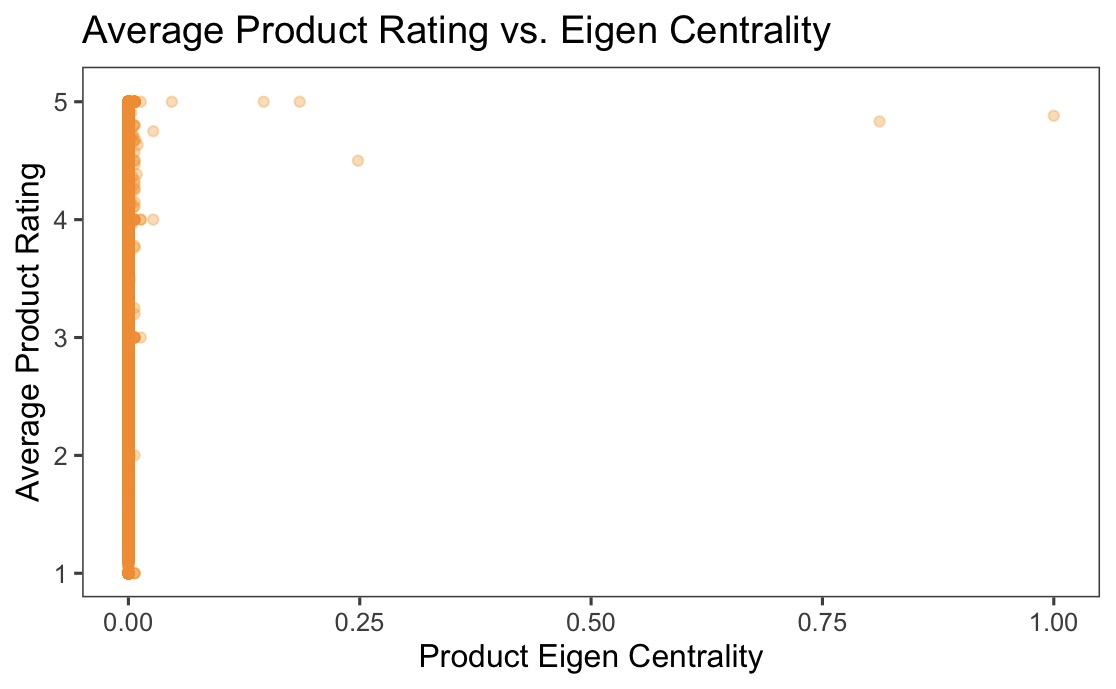
**Appendix 2: Regression predicting difference in rating between customers (original scale)**



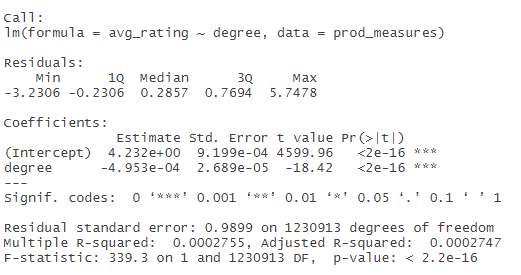
**Appendix 3: Regression predicting difference in rating between customers (square rooted Y)**



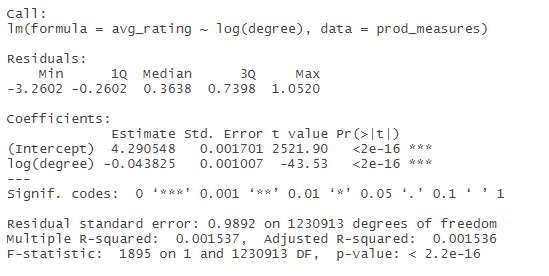
**Appendix 4: Product Rating as a function of eigen centrality of products**

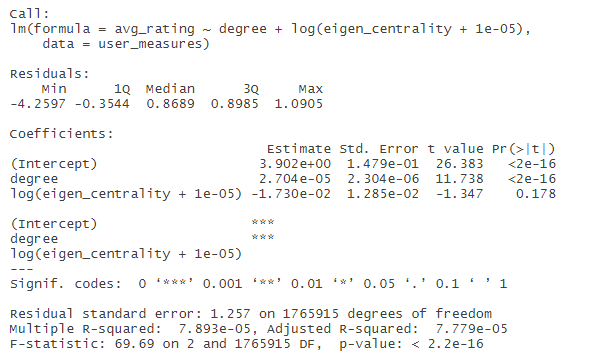


**Appendix 5: Regression output predicting average product rating using degree of products**



**Appendix 6: Regression output predicting average product rating using degree of products**



**Appendix 7: Regression output predicting average rating by user** 

**Appendix 8: Comparison of errors for the recommender system**

