Analysis of Comcast Complaint Data: A Brief Data Science Investigation

Michael Runyan
University of Georgia
msr88812@uga.edu

Abstract

This short investigation explores the uses of complaint data and proposes a relationship between the type of complaint, location of complaint, and frequency of complaints. Comcast was chosen as a company to investigate, because of the availably of data. Comcast released a sample of customers complaints which totalled to over 5000 complaints throughout 2500 different cities in the United State. The data was categorized by instances of key words. I propose that the content of these complaints and their location can be used to expose cities that Comcast is vulnerable to lose. Possible uses of this data include prioritizing resources or targeting weak areas with advertising.

Resources: Tableau, Python, Excel

1 Introduction

The world is constantly changing, and capitalism is at the forefront of change. With the emergence of computers, it did not take long for companies to take advantage of their computational power to gain an edge over competition. In today's market, every interaction between a customer and business is recorded and saved. This data has unlimited potential to be tapped and improve those interactions.

Most people assume their complaints to a company go unheard. While this proposed process does not involve each complaint being read individually and addressed, it does give it some power, much like a vote in an election. With enough complaints, a company may target aid to the customers who need it. On the other hand, this data can be used by competitors to target dissatisfied customers with personalized advertising campaigns.

Other researchers have taken on the task of analysing customer complaints. Many, such as Amir Karami, have taken the same approach, to profile the data to better address the needs [3]. This analysis does that and also considers the location of these complaints.

2 Methods

2.1 Source

The data was obtained from Kaggle where Comcast has provided the complaints for educational purposes (The copyright is displayed below in references) [1]. The data was downloaded as an Excel sheet and structured with three columns "Author, Date, Rating, and Text". The Author category contained the first name of the customer, their city, and state (Example: "Sarah of Atlanta, GA). Using Excel, these were split into three separate columns to more easily use and sort the data.

2.2 Data Distribution and Analysis

These Excel sheets were then visualized with Tableau to search for trends. It was found that cities were an excellent tool to group the complaints.

Using Python, the complaints per city were grouped together with Python's powerful Pandas package. However, the data was skewed, because more complaints come in for cities with larger customer bases. Because Comcast does not release the number of customers per city, the data was normalized by dividing the total complaints of each city by the population and multiplying it by a constant. The population was acquired from the 2015 US census (year the latest complainants were gathered). The population data only included cities exceeding 50,000 and so all cities smaller then that were assumed to have a population of 25,000.

$$C = K \times \frac{\Sigma\, complaints}{population}$$

Equation 1: Equation used to normalize the complaints per city. C represents the normalized complaints, and K is a constant to account for the large difference in population and complaints. It transforms the data so it is more easily interpreted.

2.3 Profiling Complaints

Up to this point, the complaints were just summed, and the content of the complaints had not been considered. To evaluate the content of the complaints, key words were searched for in order to determine what the complaints were filed for. The words were chosen by reading through a sample of post and picking words which occurred frequently and indicate a specific issue.
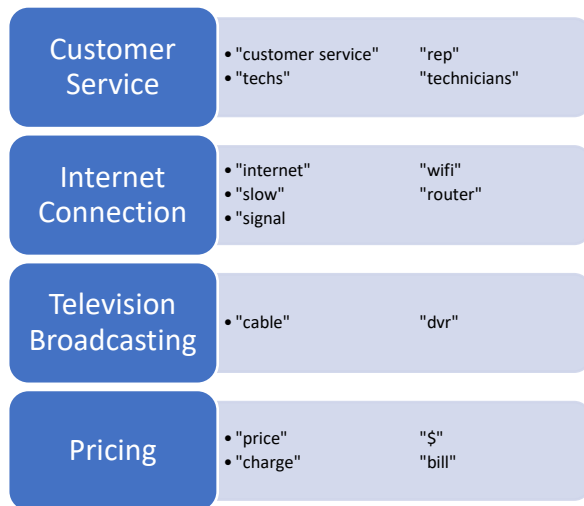
Figure 1: Categories of complaint issues, and the words searched for to indicate placements in these categories (complaints can be put into more than one category).

These categories were then visualized in Tableau to search for relationships.

3   Results

When the complaints were summed up by city, they had an uneven distribution, centered heavily in large urban areas. This is seen in figure 2.

Figure 2: The 10 cities which had the highest number of total complaints.

Once the data was normalized, the order of this list was changed drastically.
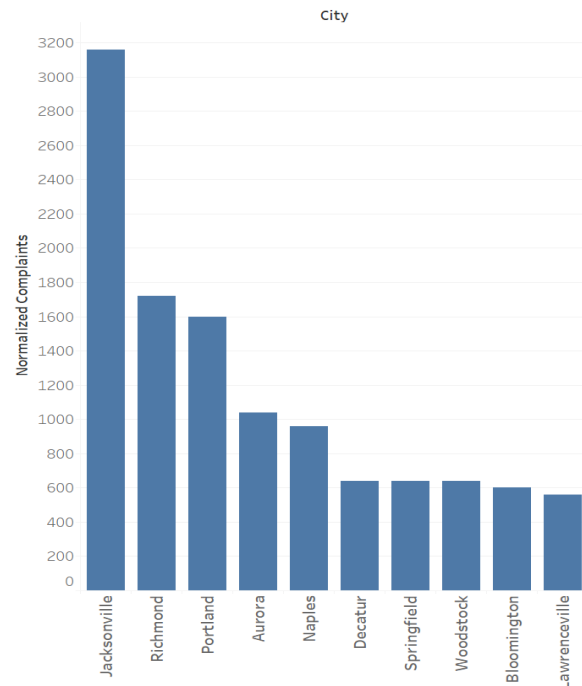

Figure 3: Graphical representation of the ten cities with the highest normalized complaints.

Diving into the content of the of the complaints, it was found each city had a unique profile of issues. These profiles are represented in the graph below. It is clear to see the unique profile in each city, like Jacksonville for example, receives many more complaints revolving around price and customer service, while Richmond also seems to have higher connectivity issues.
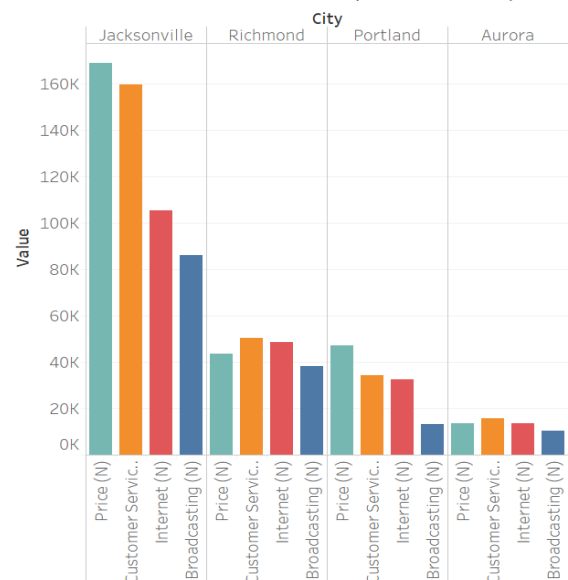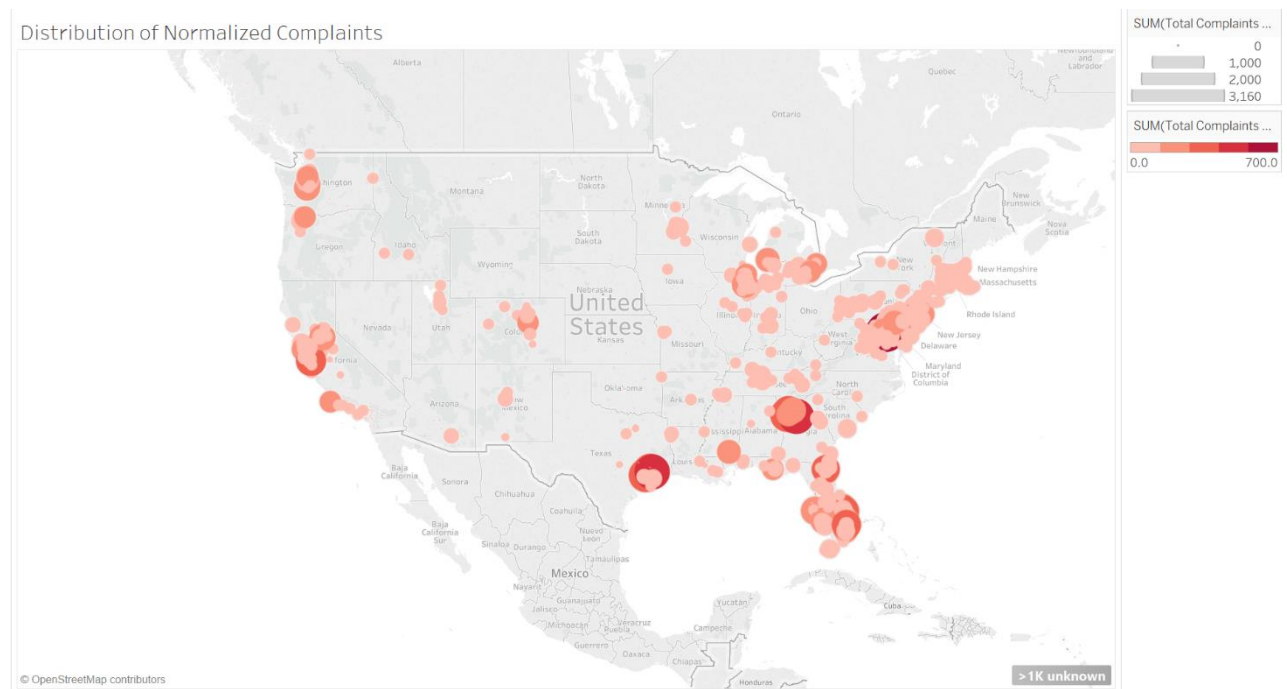

Figure 4: Represents the complaints per city, and the content of those complaints. The cities with the highest 3 normalized complaints are shown.

In order to better understand the distribution of complaints, the normalized complaints were plotted by city on a map.

Figure 5: Describes the cities with the highest normalized complaints. The complaints are described by colour and size of the indicator.

The table at the end of this document contains a list of the top hundred cities, ordered by total complaints normalized.

4 Conclusion

The total complaints were clearly skewed by the population. This is seen in the large shift in order when the data was normalized. Many of the complaints however were still made in larger cities. The complaints were centered around the coast. Most cities in the Midwest had no complaints, suggesting this is an incomplete data set, or Comcast has no infrastructure there.

The normalized data suggest a clear top 4 dissatisfied cities: Jacksonville, Richmond, Portland, and Aurora. After these cities, there is a drop off normalized complaint.

Each city showed a unique profile of issues. These profiles varied drastically, which suggest that target advertising can be applied with likely success.

It is worth noting that not many cities had more then 20 complaints. With such small numbers of complaints, the data is easily skewed. In the future, a more complete data set should be found.

5 Future Reasearch

This data can be improved in multiple ways and associated with many more factors.

The text may be analyzed with sentiment and be used to influence the urgency of each situation. The value may be multiplied into the normalized complaints.

Not all cities contribute to Comcast evenly, and cities which contribute to larger revenue of sales should be prioritized. This can be done by including the amount of revenue from each city in the calculation of importance.

The significance of complaints should also be analyzed to determine the real importance of this data. This can be done by comparing the frequency of complaints to the stock price of Comcast, or revenue by a time basis.

6 Acknowledgements

This serves as just an introductory study into complaint data. It helped me personally learn some new powerful software such as Tableau. It is my hope, that by prodding into the potential of this data may lead others to explore its usefulness. I'd also like to thank Dr. Michael Cotterell for mentoring me through the research process in the past.

7 References

[1] Charlie H., (2016, November 28). Comcast Consumer Complaints. https://www.kaggle.com/archaeocharlie/comcast complaints/version/2#comcast_consumeraffairs_

complaints.csv. Copyright © 2016 Consumers Unified LLC. All Rights Reserved.

[2]   2015 City Population Census Data Table. Retrieved from http://www.governing.com/gov-data/census/2014-city-population-estimates-data-table.html

[3]   Karami, A., & Pendergraft, N. (2018). Computational Analysis of Insurance Complaints: GEICO Case Study. Computer Science Statistics. Retrieved February 21, 2019, from http://arxiv.org/abs/1806.09736

Distribution Data

City	Total Complaints	Total Complaints (N)	Internet (N)	Price (N)	Broadcasting (N)	Customer Service (N)	Population
Jacksonville	79	3,160	105,260	169,054	86,122	159,485	2,109,714
Richmond	43	1,720	48,670	43,456	38,241	50,409	1,294,416
Portland	40	1,600	32,598	47,267	13,039	34,228	1,133,762
Aurora	26	1,040	13,650	13,650	10,500	15,750	1,051,322
Naples	24	960	520	640	280	560	25,000
Decatur	16	640	4,733	7,437	3,380	4,733	510,874
Springfield	16	640	5,790	6,514	1,448	3,619	698,400
Woodstock	16	640	400	360	200	320	25,000
Bloomington	15	600	5,107	5,745	2,553	2,553	609,876
Lawrenceville	14	560	360	280	160	320	25,000
Silver Spring	14	560	160	440	240	320	25,000
Marietta	33	559	237	339	237	423	59,067
Cumming	13	520	320	360	120	400	25,000
Jackson	13	520	4,399	3,849	3,849	3,849	458,950
Westminster	12	480	2,007	4,014	3,512	3,512	484,228
Wilmington	12	480	3,047	4,570	2,031	4,062	443,896
Woodbridge	12	480	200	400	120	280	25,000
Conyers	11	440	200	280	120	320	25,000
Humble	11	440	280	200	320	240	25,000
Stafford	11	440	200	280	240	320	25,000
Bellevue	10	400	3,488	2,616	872	2,616	361,020
Douglasville	10	400	160	280	80	240	25,000
Katy	10	400	240	320	120	280	25,000
Powder Springs	10	400	280	280	320	320	25,000
Spring	10	400	200	280	160	280	25,000
Cordova	9	360	160	320	200	200	25,000
Littleton	9	360	120	200	160	200	25,000
Milton	9	360	200	240	40	200	25,000
Port St Lucie	9	360	120	280	200	200	25,000
Concord	8	320	1,028	1,714	1,028	1,714	375,392
Danville	8	320	120	80	80	80	25,000
Fredericksburg	8	320	200	200	200	120	25,000
Kennesaw	8	320	160	160	160	160	25,000
Lake Worth	8	320	120	120	120	280	25,000
Lakewood	8	320	1,197	2,394	1,995	1,596	402,624
Los Gatos	8	320	80	280	80	120	25,000
Somerset	8	320	280	80	160	80	25,000
Boca Raton	28	300	139	193	107	182	93,235
Acworth	7	280	80	280	120	160	25,000
Auburn	7	280	1,249	1,873	624	624	299,118
Englewood	7	280	120	200	120	80	25,000
Flemington	7	280	160	120	160	200	25,000
Goose Creek	7	280	120	160	40	200	25,000
Hermitage	7	280	200	240	200	280	25,000
Lancaster	7	280	1,569	941	1,255	941	293,678
Orange Park	7	280	160	160	80	200	25,000
Sicklerville	7	280	80	120	40	160	25,000
Snellville	7	280	200	160	40	160	25,000
The Woodlands	7	280	160	200	80	160	25,000
West Chester	7	280	120	160	120	200	25,000
Sarasota	15	272	200	200	109	91	55,118
Fort Myers	18	243	68	135	95	108	74,013
Bedford	6	240	120	80	40	120	25,000
Bethesda	6	240	80	120	160	120	25,000
Brandon	6	240	160	80	80	120	25,000
Brewster	6	240	80	120	120	120	25,000
Brighton	6	240	160	160	120	200	25,000
Clinton	6	240	160	80	80	160	25,000
Columbia	6	240	514	514	514	770	388,216
Dickinson	6	240	200	120	120	160	25,000
Downers Grove	6	240	120	160	120	200	25,000
El Dorado Hills	6	240	160	240	0	240	25,000
Florence	6	240	80	160	80	80	25,000
Hudson	6	240	160	200	80	160	25,000
Lebanon	6	240	80	160	160	0	25,000
Lilburn	6	240	120	40	160	200	25,000
Loganville	6	240	120	240	160	200	25,000
Manassas	6	240	200	200	80	240	25,000
New Castle	6	240	160	80	40	200	25,000
Parker	6	240	80	160	120	160	25,000

Puyallup	6	240	40	160	80	120	25,000
Saint Paul	6	240	160	160	80	160	25,000
Stuart	6	240	160	40	80	40	25,000
Vero Beach	6	240	80	200	80	120	25,000
West Hartford	6	240	160	80	40	80	25,000
Alpharetta	15	236	141	110	79	110	63,693
Boynton Beach	17	230	149	176	135	122	73,966
Augusta	5	200	40	120	160	120	25,000
Austell	5	200	80	160	80	120	25,000
Bel Air	5	200	120	160	200	120	25,000
Belleville	5	200	80	80	40	200	25,000
Belmont	5	200	120	40	120	80	25,000
Bloomfield	5	200	40	160	40	160	25,000
Branford	5	200	80	80	80	160	25,000
Brick	5	200	40	120	40	120	25,000
Brookline	5	200	120	120	120	160	25,000
Cherry Hill	5	200	40	120	80	80	25,000
Chesterfield	5	200	120	160	160	80	25,000
Dover	5	200	120	160	120	120	25,000
Glen Burnie	5	200	0	160	0	80	25,000
Hattiesburg	5	200	40	200	80	80	25,000
Hillsborough	5	200	40	120	120	40	25,000
Mill Valley	5	200	160	120	120	160	25,000
Montgomery Vill..	5	200	80	80	80	120	25,000
Muskegon	5	200	80	80	80	120	25,000
North Fort Myers	5	200	120	40	120	80	25,000
North Miami Bea..	5	200	200	160	120	80	25,000
Owings Mills	5	200	40	80	40	80	25,000
Paducah	5	200	200	160	80	40	25,000
Palm Beach	5	200	120	200	160	160	25,000