Theme Extraction of Marketplace Reviews

By: Vaastav Anand
Supervisor: Stanislaw Nowak

VANCOUVER INSTITUTE FOR VISUAL ANALYTICS

BUCKMEUP

AIM : What Am I Doing?

- I am trying to classify the reviews that people leave for the services into 3 major categories:
  1. Quality
  2. Money
  3. Time
- Analyzing the categories of reviews to see variation in them with different factors like cost, location etc

AIM: Why Am I Doing This?

- Classifying and analysing reviews will provide a general idea as to what aspect of the service does the majority of the users care about
- This will allow BMU to better provide to the needs of the users and will be beneficial to the users as they will be able to find what they are looking for faster
- It will also allow BMU to rank the hourlies based on different qualities
- It will result in essentially a form of optimized matching

Project Pipeline

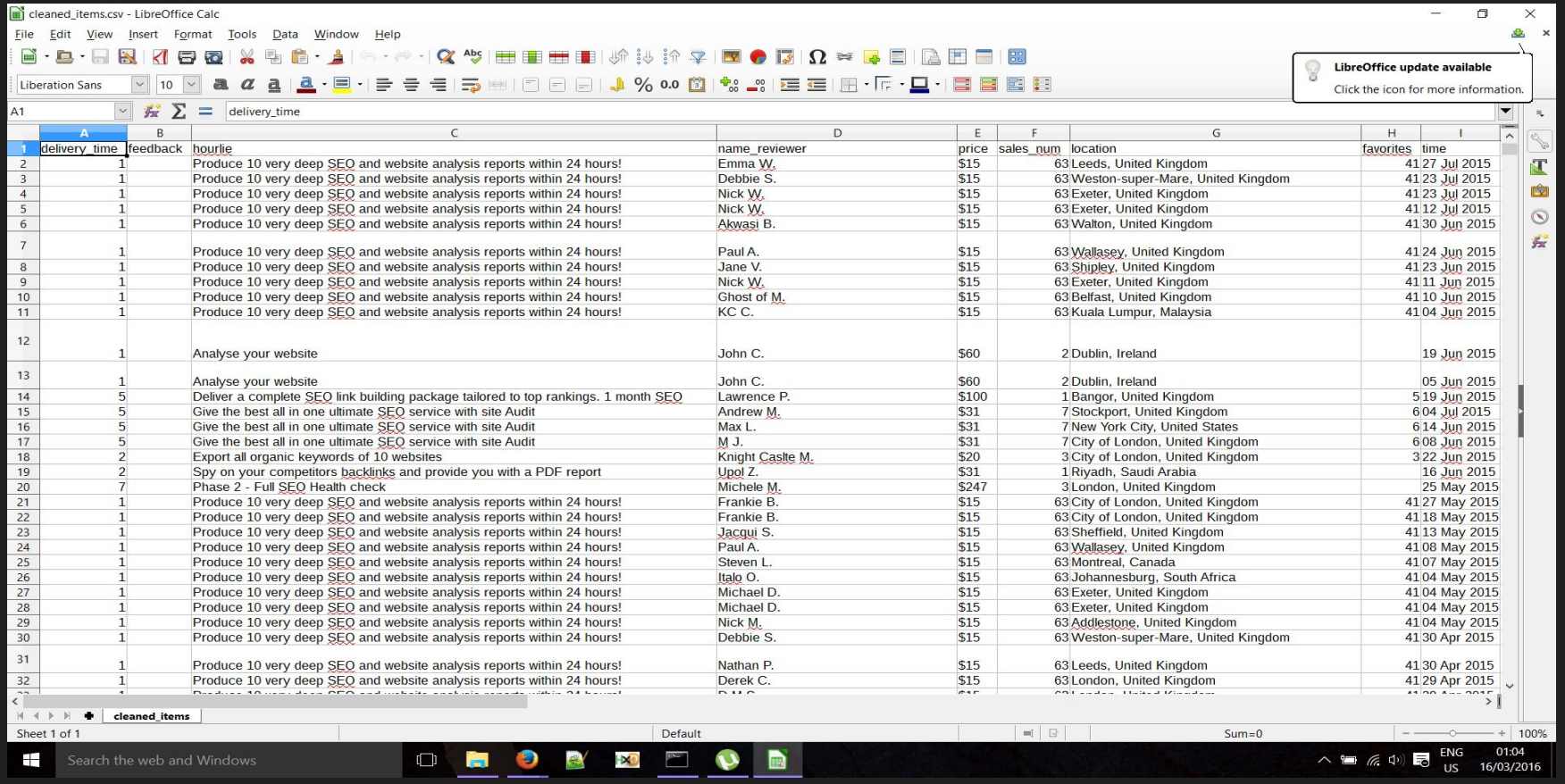
Data Gathering

- All the data used for the analysis and classification was scraped from peopleperhour.com SEO Analysis sub theme
- The reason why this website was chosen was because it has the reviews in the exact domain that we were looking for; reviews for hourlies based on single jobs.
- Software Used: Python (scrapy, re and urllib packages)

Data Gathering: Problems Faced

- It was difficult to automate the scraping as the urls keep redirecting to incorrect websites but it was managed in the end

What the Uncleaned Data looks like?



cleaned_items.csv - LibreOffice Calc

File Edit View Insert Format Tools Data Window Help

LibreOffice update available
Click the icon for more information.

Liberation Sans 10

A1 delivery_time

	A	B	C	D	E	F	G	H	I
1	delivery_time	feedback	hourlie	name_reviewer	price	sales_num	location	favorites	time
2	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Emma W.	$15	63	Leeds, United Kingdom	41	27 Jul 2015
3	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Debbie S.	$15	63	Weston-super-Mare, United Kingdom	41	23 Jul 2015
4	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Nick W.	$15	63	Exeter, United Kingdom	41	23 Jul 2015
5	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Nick W.	$15	63	Exeter, United Kingdom	41	12 Jul 2015
6	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Akwasi B.	$15	63	Walton, United Kingdom	41	30 Jun 2015
7	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Paul A.	$15	63	Wallasey, United Kingdom	41	24 Jun 2015
8	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Jane V.	$15	63	Shipley, United Kingdom	41	23 Jun 2015
9	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Nick W.	$15	63	Exeter, United Kingdom	41	11 Jun 2015
10	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Ghost of M.	$15	63	Belfast, United Kingdom	41	10 Jun 2015
11	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	KC C.	$15	63	Kuala Lumpur, Malaysia	41	04 Jun 2015
12	1		Analyse your website	John C.	$60	2	Dublin, Ireland		19 Jun 2015
13	1		Analyse your website	John C.	$60	2	Dublin, Ireland		05 Jun 2015
14	5		Deliver a complete SEO link building package tailored to top rankings. 1 month SEO	Lawrence P.	$100	1	Bangor, United Kingdom	5	19 Jun 2015
15	5		Give the best all in one ultimate SEO service with site Audit	Andrew M.	$31	7	Stockport, United Kingdom	6	04 Jul 2015
16	5		Give the best all in one ultimate SEO service with site Audit	Max L.	$31	7	New York City, United States	6	14 Jun 2015
17	5		Give the best all in one ultimate SEO service with site Audit	M J.	$31	7	City of London, United Kingdom	6	08 Jun 2015
18	2		Export all organic keywords of 10 websites	Knight Caslte M.	$20	3	City of London, United Kingdom	3	22 Jun 2015
19	2		Spy on your competitors backlinks and provide you with a PDF report	Upol Z.	$31	1	Riyadh, Saudi Arabia		16 Jun 2015
20	7		Phase 2 - Full SEO Health check	Michele M.	$247	3	London, United Kingdom		25 May 2015
21	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Frankie B.	$15	63	City of London, United Kingdom	41	27 May 2015
22	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Frankie B.	$15	63	City of London, United Kingdom	41	18 May 2015
23	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Jacqui S.	$15	63	Sheffield, United Kingdom	41	13 May 2015
24	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Paul A.	$15	63	Wallasey, United Kingdom	41	08 May 2015
25	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Steven L.	$15	63	Montreal, Canada	41	07 May 2015
26	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Italo O.	$15	63	Johannesburg, South Africa	41	04 May 2015
27	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Michael D.	$15	63	Exeter, United Kingdom	41	04 May 2015
28	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Michael D.	$15	63	Exeter, United Kingdom	41	04 May 2015
29	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Nick M.	$15	63	Addlestone, United Kingdom	41	04 May 2015
30	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Debbie S.	$15	63	Weston-super-Mare, United Kingdom	41	30 Apr 2015
31	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Nathan P.	$15	63	Leeds, United Kingdom	41	30 Apr 2015
32	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Derek C.	$15	63	London, United Kingdom	41	29 Apr 2015

cleaned_items

Sheet 1 of 1 Default Sum=0 100%

Search the web and Windows ENG US 01:04 16/03/2016

Data Cleaning: Pipeline

- **Cleaning the Review**
  This part included removing html tags from the scraped reviews and encoding the reviews in utf-8 encoding

- **Finding Keywords in a Reviews**
  This part uses Python’s nltk package to find keywords in the reviews

- All the word clouds are made with the wordcloud and matplotlib packages in python

Data Cleaning: Cleaned Text

Data Cleaning: Cleaned Text

- The wordcloud made from just the cleaned review has a lot of pronouns and conjunctions and other parts of speech which we really don’t care about unnecessarily showing up in the wordcloud
- So, I decided that I should only be looking at certain keywords
- So, I extracted all the keywords from a review and made a wordcloud out of that

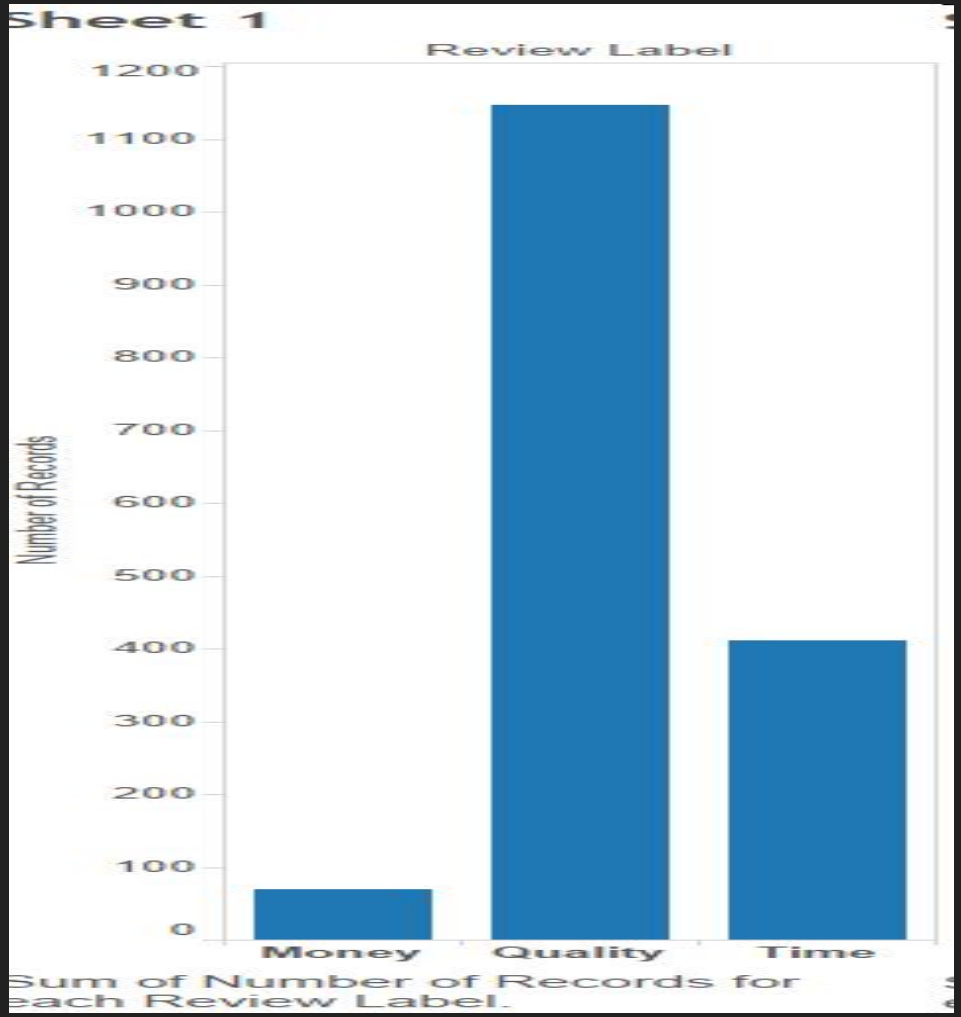
Data Cleaning: Keyworded Text

Building a Classifier

- So, just to recap. We want to classify the reviews into 3 categories; money, time and quality
- The classifier used was a NaiveBayesClassifier from textblob package in python
- So, I decided to use 80% of the extracted data as the train data for training the classifier and 20% as the test data
- The classifier was 60% accurate; and it takes about 5-6 minutes on my computer.

Data Analysis: Trend in themes selection

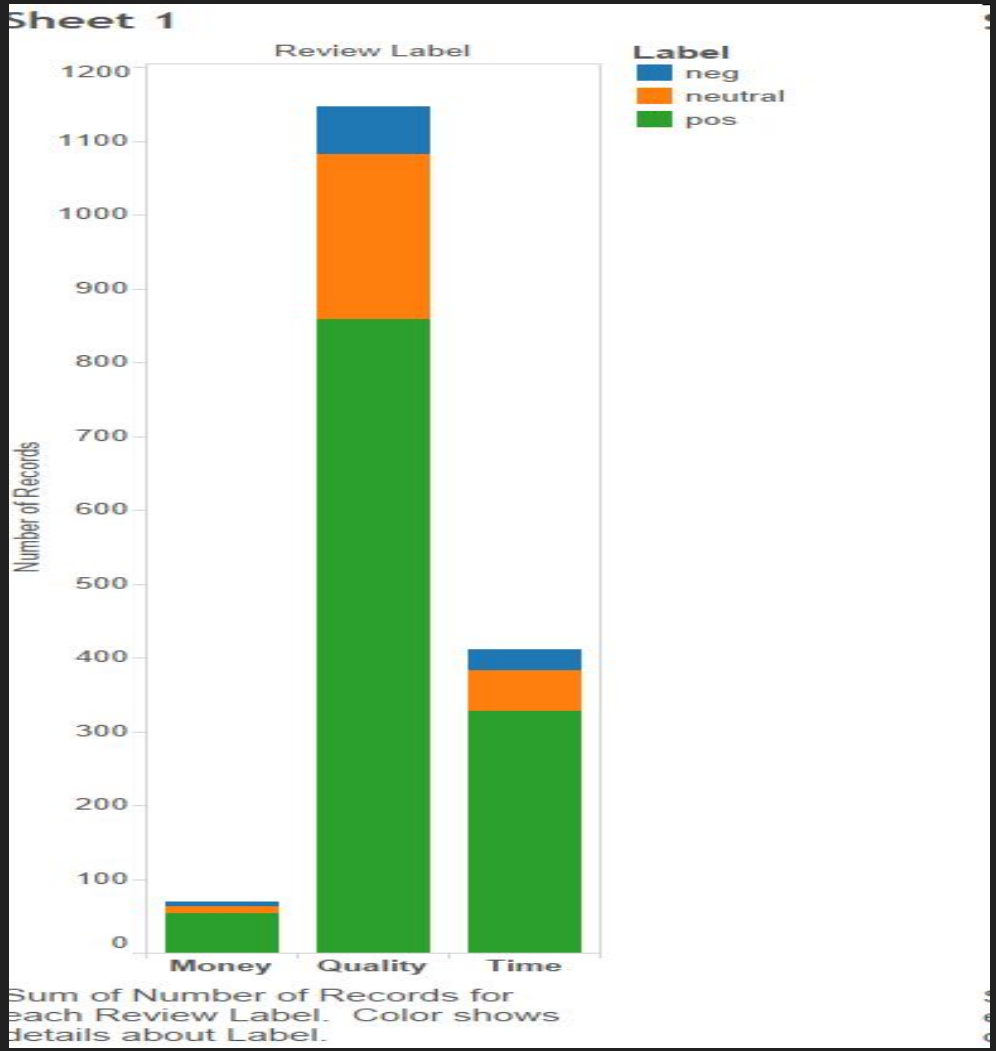
- The software I used for data analysis was Tableau
- The surprising thing that I found was the fact that people who reviewed cared more about the quality of the product than the cost and time of execution


Sheet 1
Review Label
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
Number of Records
Money
Quality
Time
Sum of Number of Records for
each Review Label.

Data Analysis: Variation with sentiment score

- The number of positive reviews dominated the number of neutral and negative reviews which resulted in every theme having a higher number of positive reviews
- The reviews tagged with the theme of Time had the highest relative proportion of negative reviews. This suggests that people tend to get more annoyed when the job isn’t done within the expected time
- While, the reviews tagged with the theme of Money had the highest relative proportion of positive reviews. This suggests that people tend to only write money-related reviews if there is something special about the cost of the service.

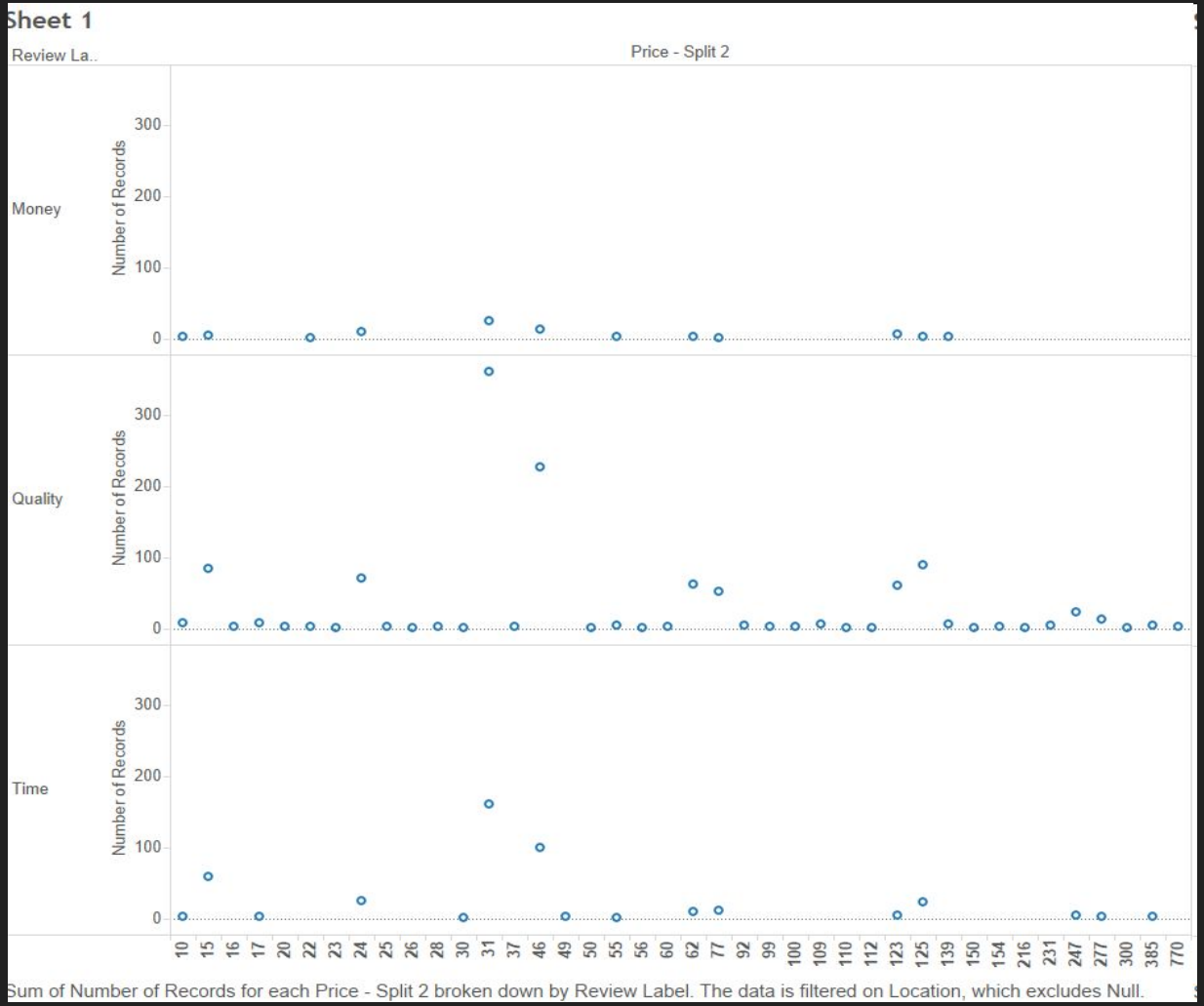
Sheet 1



Sum of Number of Records for each Review Label. Color shows details about Label.

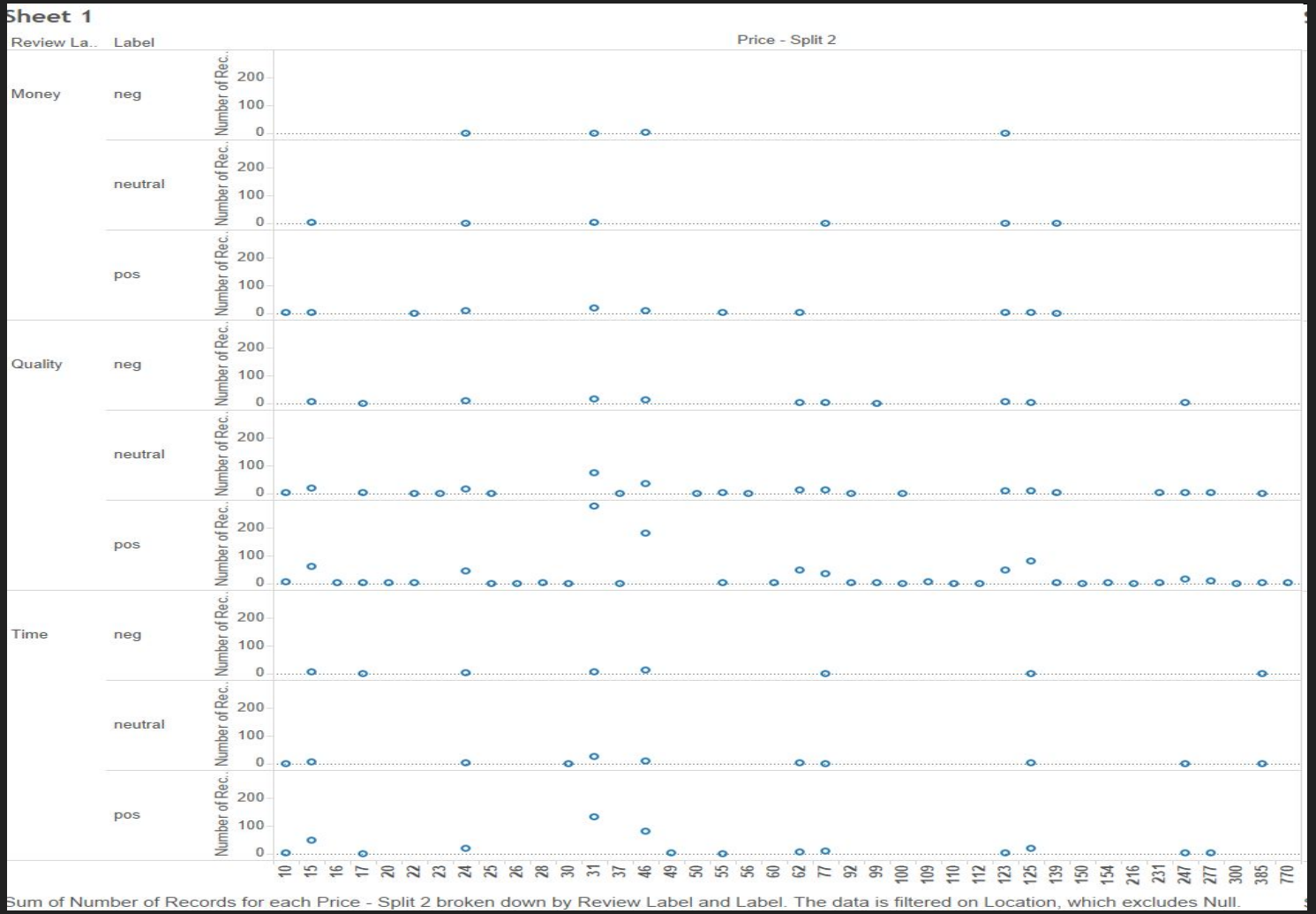
Data Analysis: Variation with Cost

- The majority of the money-related reviews were for lower cost jobs
- While, the majority of the reviews for time-based reviews and quality-based reviews were for jobs costing 30-60 dollars


Sheet 1
Review La..
Price - Split 2
Money
Quality
Time
Number of Records
0
100
200
300
10
15
16
17
20
22
23
24
25
26
28
30
31
37
46
49
50
55
56
60
62
77
92
99
100
109
110
112
123
125
139
150
154
216
231
247
277
300
385
770
Sum of Number of Records for each Price - Split 2 broken down by Review Label. The data is filtered on Location, which excludes Null.

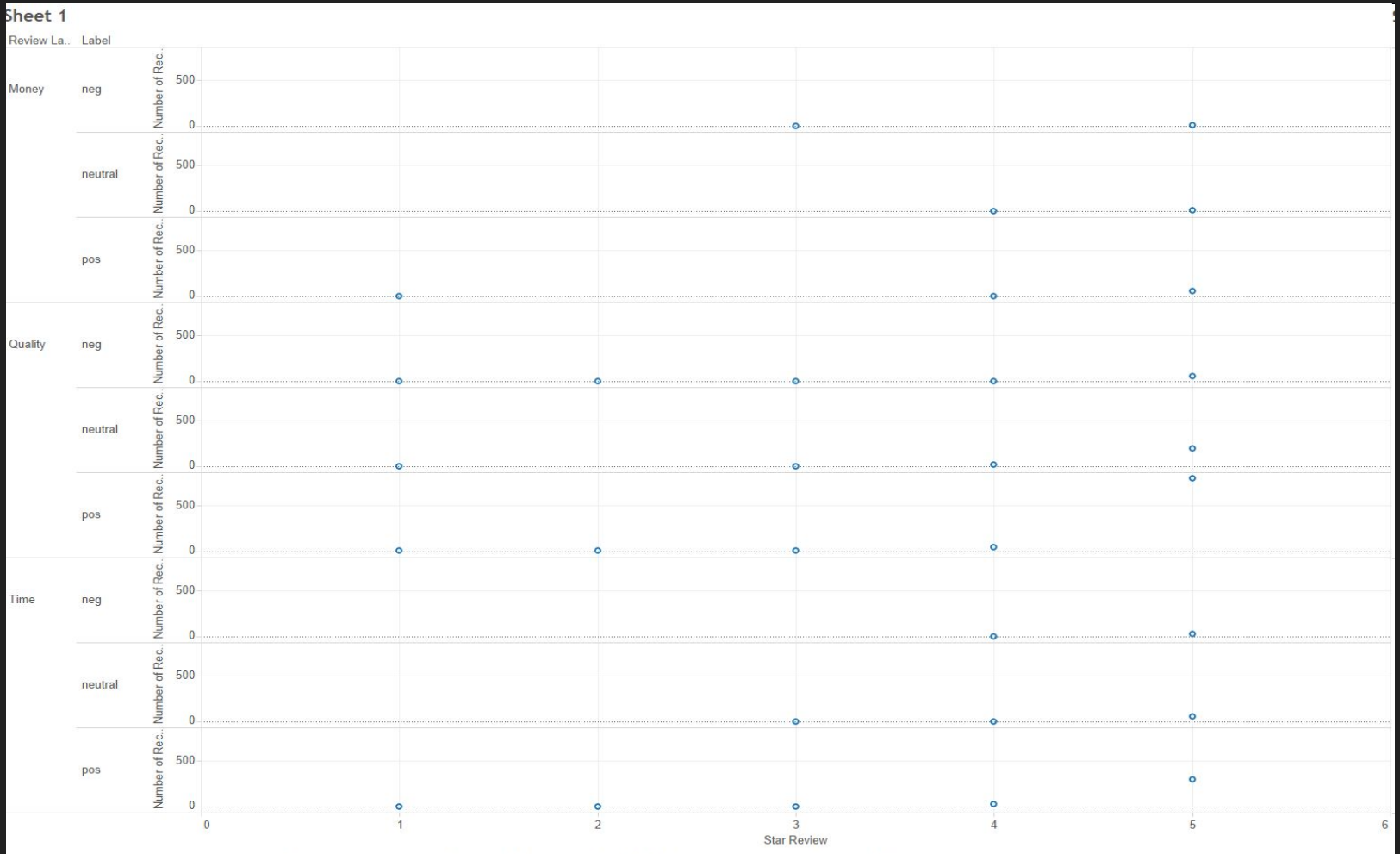
Data Analysis: Variation in Sentiment and Cost

- The positive money related reviews were mostly for lower-priced jobs whereas the negative money related reviews were for higher-priced jobs
- The positive time-related reviews were high in number for the price range 30-40 dollars
- The positive quality-related reviews had a very high number for the price range 40-50 dollars whilst there was also a high number of such reviews in the 120-130 price range.


Sheet 1
Review La..
Label
Price - Split 2
Money
Quality
Time
neg
neutral
pos
Number of Rec..
200
100
0
10
15
16
17
20
22
23
24
25
26
28
30
31
37
46
49
50
55
56
60
62
77
92
99
100
109
110
112
123
125
139
150
154
216
231
247
277
300
385
770
Sum of Number of Records for each Price - Split 2 broken down by Review Label and Label. The data is filtered on Location, which excludes Null.

Data Analysis: Variation with Star Reviews

- Surprisingly, none of the negative money-related reviews got a rating below 3 stars
- Similarly, all of the negative time-related reviews scored a rating above or equal to 4
- The negative quality-related reviews were evenly distributed evenly across all the 5 reviews
- This all leads to suggest that the star reviews approach is not consistent with the feelings the reviewer has

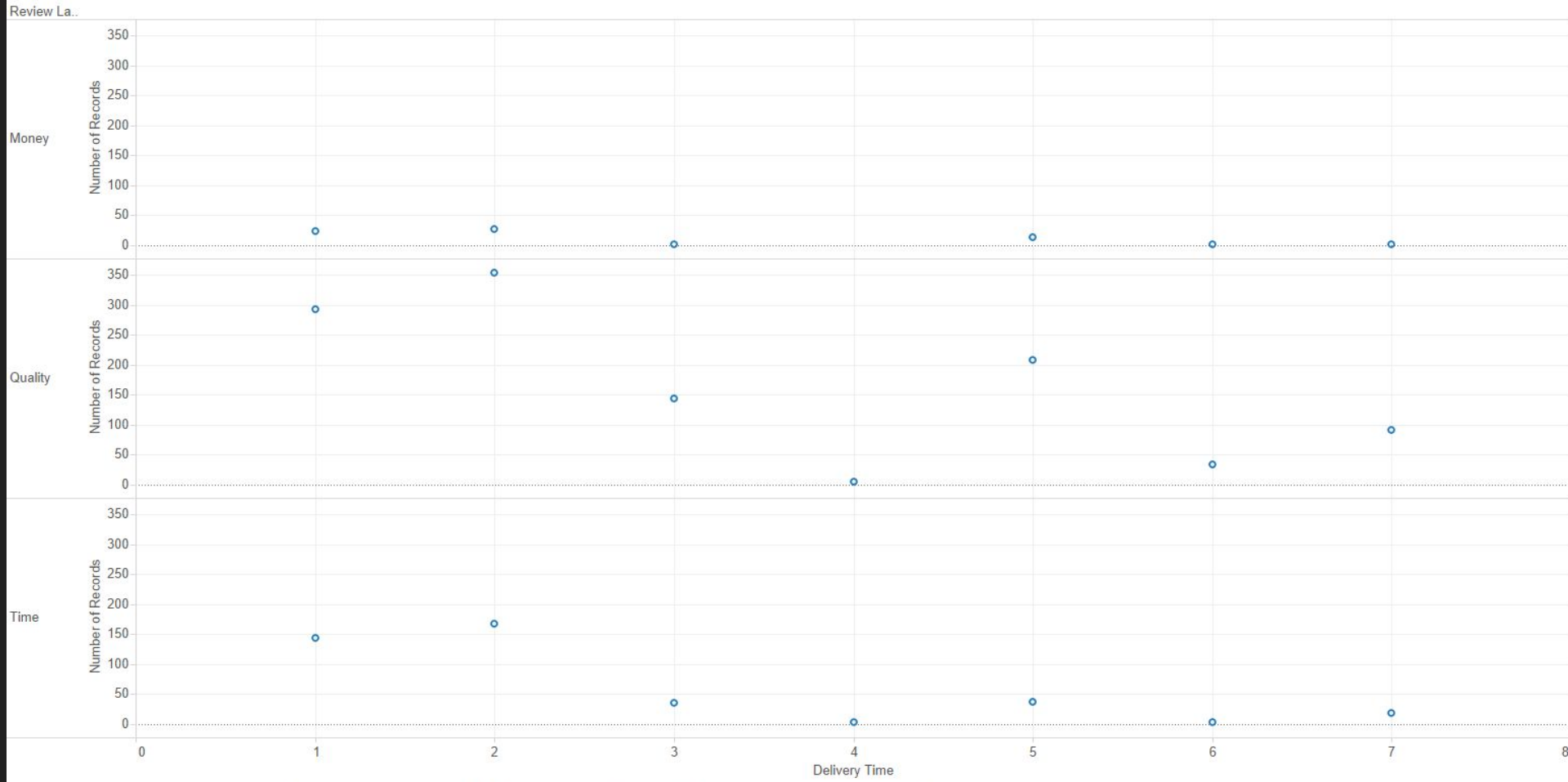


The plot of sum of Number of Records for Star Review broken down by Review Label and Label. The data is filtered on Location, which excludes Null.

Data Analysis: Variation with Delivery Time

- Most of the time-related reviews were for jobs that had 1 or 2 days as the delivery time. This suggests that people only comment about the speed of the service if the service is really fast

Sheet 1

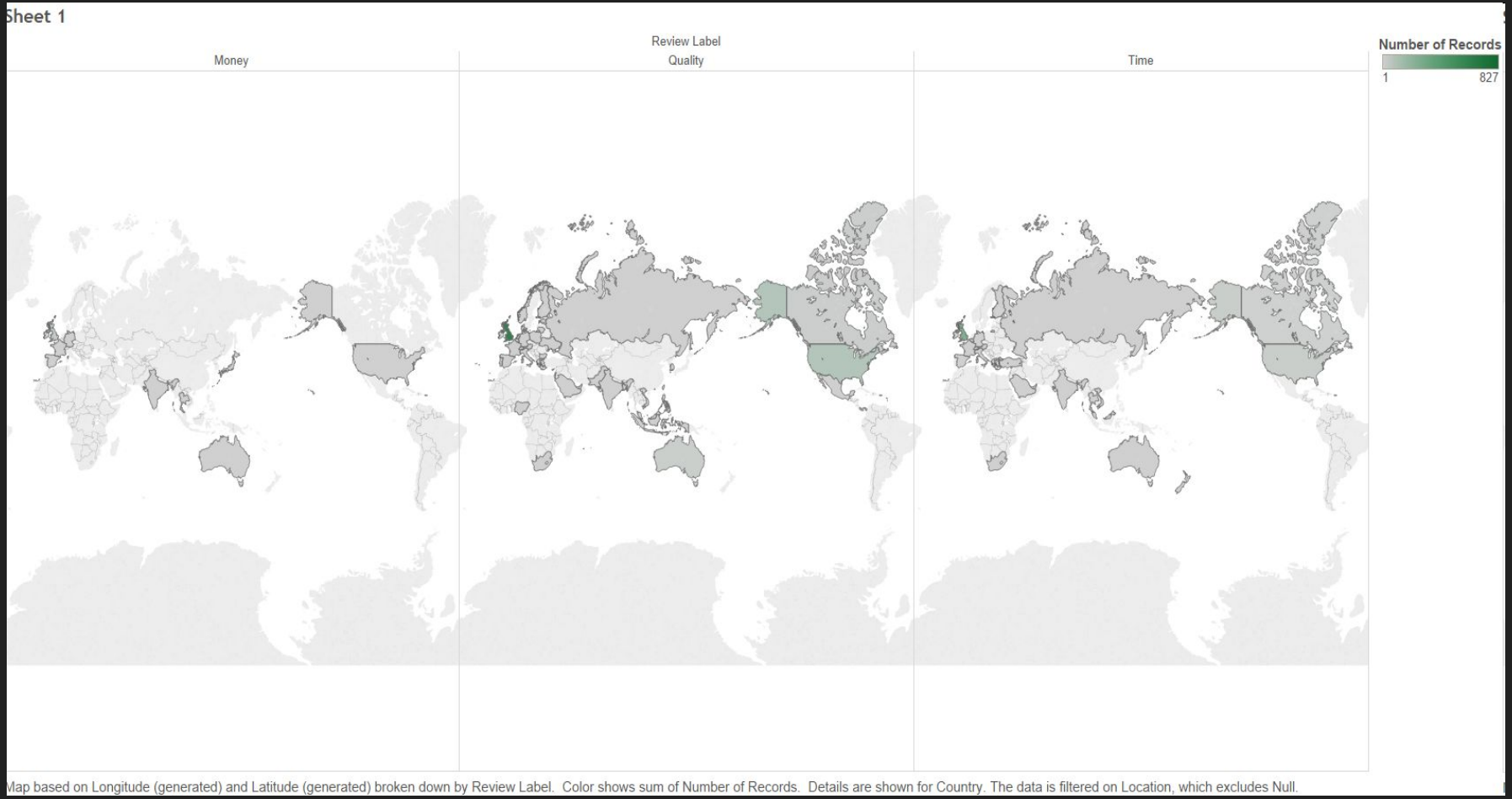


The plot of sum of Number of Records for Delivery Time broken down by Review Label. The data is filtered on Location, which excludes Null.

Data Analysis: Location

- There was no significant variation noticed in terms of the themes of the reviews
- All locations seemed to follow the trend which the whole world collectively follows

Sheet 1



Map based on Longitude (generated) and Latitude (generated) broken down by Review Label. Color shows sum of Number of Records. Details are shown for Country. The data is filtered on Location, which excludes Null.

Takeaways

- To get a better understanding, we need to get more data
- The location data was more biased towards United Kingdom as that’s where most of the reviews were from

Future Possibilities for BuckMeUp

- They could refine the review metrics they offer so that people get more detailed reviews
- They could build an automated text theme classifier into their web app for automated review classification

Challenges Faced

- Scraping the data from the website
- Writing an accurate yet fast classifier for classifying the reviews
- Classifying the test data for the classifier. It was done manually so I had to make sure I was being consistent. (Some amount of human error remains)