Demand Prediction: EDA

10.13.2017

•••

Nirmal Budhathoki
Garrett Cheung
Jillian Jarrett
Toby Moreno
Orysya Stus

Project Objective

Providing valuable statistics, identifying data issues, and creating visualizations with the end goal of predicting book sale demand.

General Questions

- How will this data help us predict demand?
- What exactly is our target? By which metric are we going to predict book demand?
  - Region
  - Time
  - Genre
  - All of the above?
- How are our features related to one another?
- Which features are indicators in predicting our target? Which aren’t?
- What features can we engineer to better predict our target?

Overall Approach

Milestone I	Milestone II	Milestone III	Milestone IV
Data dictionary Preliminary statistics Histograms for columns distributions	Visualize trends, correlations, and important relations Text analysis	Provide directions for feature engineering Machine learning recommendations	Determine if machine learning predictions and findings are in agreement with EDA

Note: This is tentative, other group feedback may alter process.

Tools

Data Dictionary

Table	Data Source	Description
calendar	PostGreSQL	Master calendar from 1/1/1950 - 12/31/2050 including holidays and DoW
campaigns		Discounts/Free Shipping promotions
customers		Customers and which household they belong to
orders		Purchases by each customer
orderlines		How Amazon distributed/shipped the purchased items
products		Books, prices, ASIN, category, in stock
reviews		Reviews, reviewer name, score, time
subscribers		Subscribers (dealer, mail, store, chain), monthly fee, start/stop dates
zipcensus		Comprehensive census broken down by zip code
zipcounty		Zip codes and their Geographic/Demographic data

Data Dictionary

Table	Data Source	Description
reviews	AsteriskDB	Json based repository for reviews by category
Reviews	Solr	Searchable reviews in text format.

PostGreSQL Tables - Preliminary Stats

For each table access the following documents:

- **Data.txt**: data shape, data type of each column, count of nominal, numeric, and datetime attributes
- **nominal_stats.csv**: unique and null count
- **numeric_stats.csv**: mean, min, max, std, percentiles
- **datetime_stats.csv**: min, max, most frequent date
- **columnname_val_counts.csv**: for a specific nominal column name a count of the unique values
- **columnname.png/columnname_logscale.png**: normal and log scale histograms for numeric attributes

Find preliminary stats here: https://github.com/mas-dse-jejarret/DSE203_Demand_EDA/blob/master/PostGreSQL_Tables_PreliminaryStats.zip

Example: campaign table

```
Initial data has 239 rows and 5 columns
The datatype for column: campaignid is <class 'numpy.int64'>.
The datatype for column: campaignname is <class 'str'>.
The datatype for column: channel is <class 'str'>.
The datatype for column: discount is <class 'numpy.int64'>.
The datatype for column: freeshppingflag is <class 'str'>.
Nominal attribute count: 3
Numeric attribute count: 2
Datetime attribute count: 0
```

Data.txt

	unique_val_count	null_val_count	null_%_from_total
campaignname	1	0	0
channel	13	0	0
freeshppingflag	2	0	0

nominal_stats.csv

	count	mean	std	min	25%	50%	75%	max	null_count	median
campaignid	239	2120	69.13754	2001	2060.5	2120	2179.5	2239	0	2120
discount	239	5.887029	11.31407	0	0	0	10	50	0	0

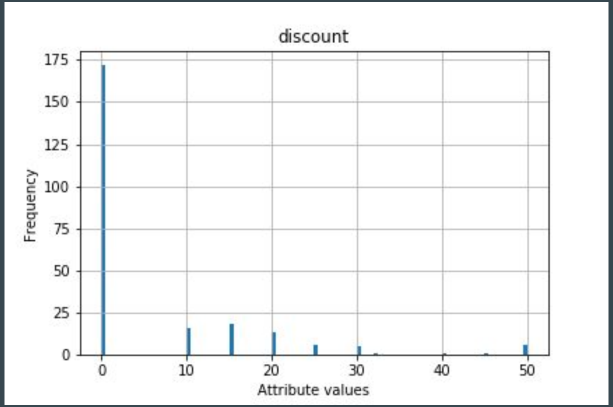
numeric_stats.csv

Find preliminary stats here: https://github.com/mas-dse-jejarret/DSE203_Demand_EDA/blob/master/PostGreSQL_Tables_PreliminaryStats.zip

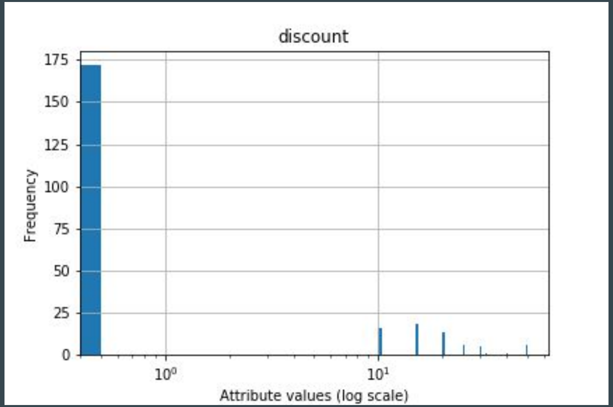
Example: campaign table

	attribute_values	value_count	value_count_%_from_total
2	AD	80	33.4728
3	BULK	3	1.25523
4	CATALOG	4	1.67364
5	CONFERENCE	1	0.41841
6	EMAIL	14	5.857741
7	EMPLOYEE	9	3.76569
8	INSERT	18	7.531381
9	INTERNAL	1	0.41841
10	MAIL	46	19.24686
11	PARTNER	39	16.31799
12	REFERRAL	1	0.41841
13	SURVEY	1	0.41841
14	WEB	22	9.205021
15	Nulls	0	0

channel_val_counts.csv



discount.png



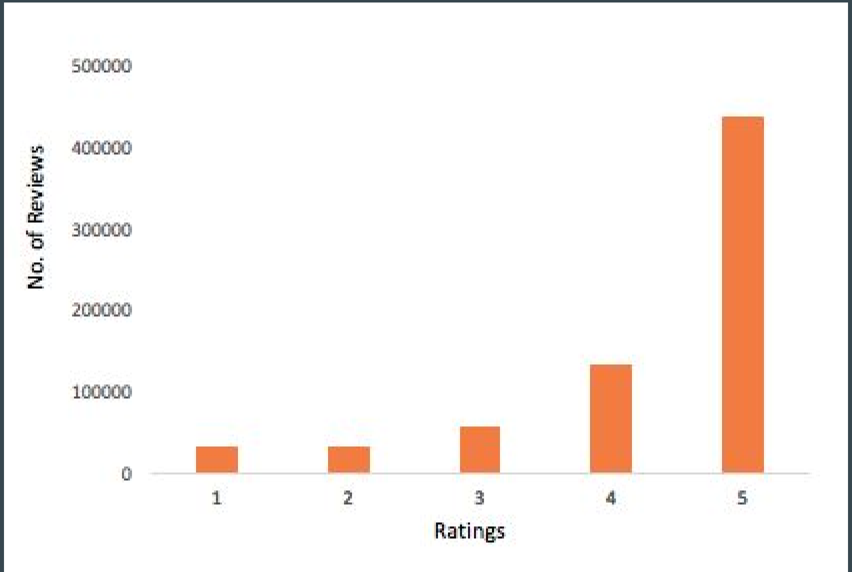
discount_logscale.png

Find preliminary stats here: https://github.com/mas-dse-jejarret/DSE203_Demand_EDA/blob/master/PostGreSQL_Tables_PreliminaryStats.zip

Asterix Data - Preliminary Stats

Reviews Data

- Total Reviews: 77,164
- Total Unique Reviewers: 69,729
- Missing values: 21 (Reviewer's Name)
- Average Ratings from Reviewers= 4.3
- Total unique products: 4040
- Each product has 19 reviews



How can we use Reviews data ?

- Historical reviews are very important for predictive models (e.g. Predicting demand for a new book can be compared against similar books sold in past)
- Negative reviews can also be valuable (e.g. price adjustment)
- Sentiment analysis

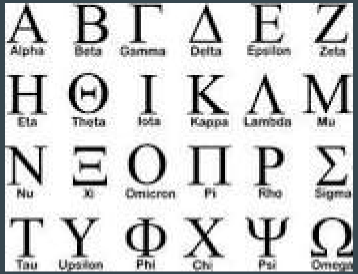
Challenges:

- The review data is not yet linked to the product data or customer data
- Finding the product attributes to compare and find similarity against books

Classification/Category Data

[{ "uid": "Children's Books", "count": 793 }
, { "uid": "Christian Books & Bibles", "count": 288 }
, { "uid": "Computers & Technology", "count": 435 }
, { "uid": "Crafts, Hobbies & Home", "count": 347 }
, { "uid": "Engineering & Transportation", "count": 272 }
, { "uid": "Gay & Lesbian", "count": 39 }
, { "uid": "History", "count": 384 }
, { "uid": "Law", "count": 144 }
, { "uid": "Literature & Fiction", "count": 428 }
, { "uid": "Medical Books", "count": 288 }
, { "uid": "Mystery, Thriller & Suspense", "count": 59 }
, { "uid": "Religion & Spirituality", "count": 534 }
, { "uid": "Science & Math", "count": 432 }
, { "uid": "Sports & Outdoors", "count": 233 }
, { "uid": "Travel", "count": 651 }
, { "uid": "Arts & Photography", "count": 389 }]

Solr Data - Preliminary Stats



Solr Data Review

- Could be a useful platform for exploring large sets of text data
  - Host Target Location should be a cloud infrastructure
  - Manual data integration requires converting data to XML first
  - Is there Real-time Twitter Feed Configuration?
- Data Access requires simple HTTP protocol (built-in Java client support)
  - May require multiple transformations
    - Tokenization / TFIDF
    - POS tagging / Stemming and Lemmatization
- Derived sentiment analysis could be used to project demand
- Still a Working Progress
  - Too Early to tell if it would be useful
  - Collaborating with other stakeholders is critical

Next Steps

- Work with Integrated Schema and Justification Team to review preliminary statistics.
- Work with Query capability and Learning Team to create the necessary views/tables for us to use.
- Discuss and create meaningful visualizations of trends with the suggestions from Machine Learning Team.