Capstone Project

Background

Different from conventional physical delivery systems that are characterized by a scarcity of resources, on-line stores can make anything that exists available to the customer. For example, a physical bookstore may only have several thousand books on its shelves due to limited shelf space, but Amazon can offer millions of books to their customer. Due to overwhelming number of products, it is almost impossible for customers to go through all of them and find the products they are looking for. Recommendation in the online world is becoming more and more important. I will build an online retailer recommendation system, which offering customers suggestions about what they might like to buy. This system is mainly based on Amazon product reviews and metadata (e.g. price, related products, sales rank, and brand).

Dataset

Amazon product (Clothing, Shoes and Jewelry) dataset, which contains product reviews and metadata from Amazon, spanning May 1996 - July 2014.

See http://jmcauley.ucsd.edu/data/amazon/links.html

Date Collection

The data collection process has been divided into several steps:

- Download reviews_Clothing_Shoes_and_Jewelry_5.json.gz and meta_Clothing_Shoes_and_Jewelry.json.gz from the following links:
  - http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Clothing_Shoes_and_Jewelry_5.json.gz
  - http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/meta_Clothing_Shoes_and_Jewelry.json.gz
- Convert the data to strict json using script strictJson.py
- Parse json file and import the data into mysql database using script metaJsonToMysql.py and reviewJsonToMySql.py

Exploratory Data Analysis

Star Category vs The Number of Reviews

- Target: Observing star category distribution over the number of reviews.
- Steps:
  - Database inquering: SELECT overall, count(*) AS count FROM reviews_clothing_shoes_and_jewelry GROUP BY overall;
  - Exporting the result into overall_numberofreviews.csv
  - Reading overall_numberofreviews.csv into R:

```
sc <- read.csv("overall_numberofreviews.csv", stringsAsFactors = FALSE)
```

– Ploting star category distribution graph:

```
install.packages("ggplot2")
```

```
library(ggplot2)
h1 <- ggplot(sc, aes(x = overall, y = count, fill = overall))
h1 <- h1 + geom_bar(position = "stack", stat = "identity")
h1 <- h1 + labs(list(title = "Star Category Distribution for Reviews",
     x = "Star Category", y = "Number of Reviews", fill = "Star Category"))
h1
```

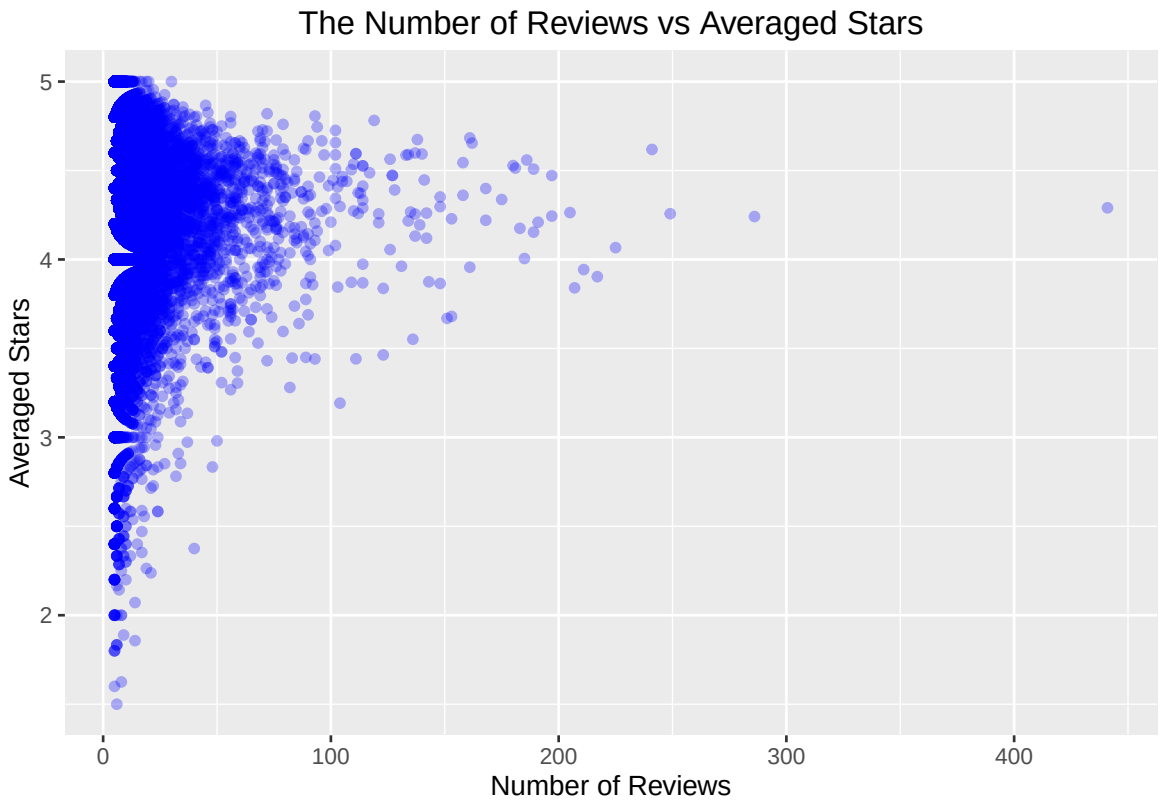
The Number of Reviews vs Averaged Stars

- Target: Observing the relationship between number of reviews and averaged stars
- Steps:
  - Database inquering: SELECT asin, count(*) AS count, avg(overall) AS star FROM reviews_clothing_shoes_and_jewelry GROUP BY asin;
  - Exporting the result into product_review_star.csv
  - Reading product_review_star.csv into R:

```
prs <- read.csv("product_review_star.csv", stringsAsFactors = FALSE)
```

  - Ploting star category distribution graph:

```
h2 <- ggplot(prs, aes(x = count, y = star))
h2 <- h2 + geom_point(col = "blue", alpha = 0.3)
h2 <- h2 + labs(list(title = "The Number of Reviews vs Averaged Stars",
      x = "Number of Reviews", y = "Averaged Stars"))
h2
```



The Accumulated Number of Reviews Changing Over Time

- Target: First find the most popular product(having the most reviews) and then observe its accumulated number of reviews chaning over time.
- Steps:
  - Database inquering:
    SELECT asin, count(*) AS count FROM reviews_clothing_shoes_and_jewelry GROUP BY asin ORDER BY count DESC LIMIT 1;
    SELECT * FROM reviews_clothing_shoes_and_jewelry WHERE asin = 'B005LERHD8';
  - Exporting the result into product_with_most_reviews.csv
  - Reading product_with_most_reviews.csv into R and operate the data:

```
df <- read.csv("product_with_most_reviews.csv", stringsAsFactors = FALSE)
df <- df %>% mutate(reviewTime = format(as.Date(reviewTime, "%m %d, %Y"), "%Y%m"))
         %>%  arrange(reviewTime)
         %>% group_by(reviewTime)
         %>% summarise(count = n())
df <- mutate(df, cum = cumsum(count))
df <- mutate(df, cum = as.numeric(cum))
```

– Writing df into df.csv and modifing df.csv (solving the problem that there is no reviews in some months):

– Ploting star category distribution graph:

```
df <- read.csv("df.csv", colClasses = c("factor", "integer", "integer"))
h3 <- ggplot(df, aes(as.numeric(reviewTime)))
h3 <- h3 + geom_ribbon(aes(x = as.numeric(reviewTime),
      ymin = cum - 20, ymax = cum + 20), fill = "grey")
h3 <- h3 + geom_line(aes(y = cum),group = 1)
h3 <- h3 + scale_x_continuous(breaks = 1:30, labels = levels(df$reviewTime))
h3 <- h3 + theme(axis.text.x = element_text(angle = 90, hjust = 1))
h3 <- h3 + labs(list(title = "The Accumulated Number of Reviews Changing Over Time",
      x = "Review Time", y = "The Accumulated Number of Reviews"))
h3
```