**Data science programming of Taobao Commodity Based on Python**

Group W

MB95510 Ren Xuguang and MB95525 Bao Rong

1. **Introduction**

1.1 Object

(1) Taobao commodity: sofa

(2) Quantity: 4400 items in 100 pages

(3) Screening Conditions: TMall, Sales Volume from High to Low, Price above 500 yuan.

1.2 Purpose

(1) Text analysis of commodity title, word cloud visualization; (2)Statistical analysis of sales corresponding to different keywords word; (3)Analysis of commodity price distribution; (4)Analysis of sales distribution of goods; (5)Average sales distribution of commodities in different price ranges; (6)Analysis of the impact of commodity price on sales volume; (7)Analysis of the impact of commodity price on sales; (8)Average sales distribution of goods in different provinces.

1.3 Steps and codes

(1) Data collection: Python crawls commodity data of taobao.com (2) Data cleaning and processing (3) Text analysis: Jieba segmentation, wordcloud visualization (4) Visualization of data histogram (5) Histogram visualization hist (6) Scatter visualization (7) Data regression analysis visual regplotv

1.4 Libraries

Requests, Retrying, Re, Missingno, Jieba, Matplotlib, Wordcloud, Imread, Seaborn, Pyecharts, etc

1. **Collect the Data by Web Crawler**

2.1 Basic process: send requests; get the corresponding content; interpret HTML content: use Re library or third-party library; export data as Excel document.

* 1. Detailed process:
     1. Import libraries
     2. Requests (cookie, user-agent)



图片包含 物体

描述已自动生成

Unfortunately, according to the Robots Exclusion Standard, the search pages of Taobao does not allow crawling. Our project just discuss technology for academic purpose and ensure non-commercial use. With many strategies for coping with anti-spider, we successfully anti anti-scraping and got the data. How to deal with the anti-scraping mechanisms was the biggest challenge for our project.

(1) Time our program using time.clock() to find out how long it takes to execute our solution. (2) We observed that there were 44 products shown in every webpage and then we decided to crawl 1-100 pages. (3) Retrying is a general-purpose retrying library, so we use it to simplify the task of adding retry behavior. (4) Using requests library: make requests using HTTP methods such as GET. (5) Map: similar to map in python but the iterables are collected immediately rather than lazily, func is executed asynchronously and several calls to func may be made concurrently. With a lot of strategies

图片包含 屏幕截图

描述已自动生成

图片包含 屏幕截图

描述已自动生成

(6) Customize our requests by modifying headers: Because there may be anti-crawlers on Taobao, it is necessary to use cookies to build the header to pretend to be a browser.



* + 1. Regular expression (Re library)

(1) Regular expression can be used to check if a string contains the specified search pattern and findall() function returns a list containing all matches.

图片包含 屏幕截图

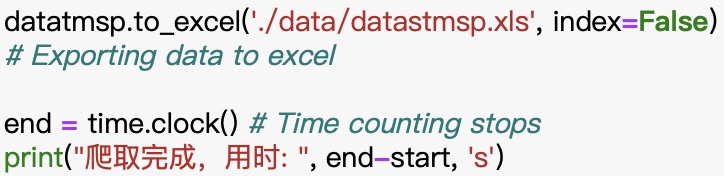
描述已自动生成

图片包含 屏幕截图, 文字

描述已自动生成(2) Turning the page numbers of successfully crawled webpages into the num value in the URL. (3) Recording the page numbers of unsuccessfully crawled webpages in the list for cyclic crawling. (4) Stop cycling when the number of unsuccessfully crawled webpages is 0.

* + 1. Export

Finally, we export the data as Excel document.



**3. Data Cleaning and Data Processing**

3.1 Reading data and general understanding.

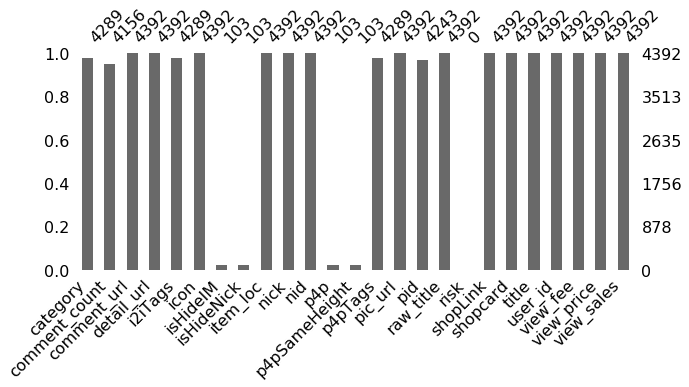
3.2 Using visualization tool (Missingno library) to observe data loss.

3.3 Using data.dropna() to delete columns with more than half of the missing values.

3.4 Deleting duplicated rows by pandas.DataFrame.

图片包含 屏幕截图

描述已自动生成



3.5 Check the first five data

图片包含 屏幕截图

描述已自动生成

3.6 According to the purpose of this project, the following attributes are needed for this analysis: item\_loc, raw\_title, view\_price, view\_sales

图片包含 屏幕截图

描述已自动生成

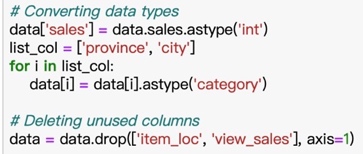
3.7 Organizing these four kinds of data.

Firstly, split the provinces and cities listed in item\_loc to get two columns: 'province' and 'city'.

Special case: Since the provinces and cities of the municipality directly under the central government are the same, it was judged according to the character length. Secondly, extract the numbers in the 'view\_sales' column to get the 'sales' column. Thirdly, view data type.

图片包含 屏幕截图

描述已自动生成

3.8 Convert data types as category data type in order to save time and space as well as improve computing efficiency. Then delete unused data columns.

3.9 After that, we successfully cleaned and organized the data via Web crawler. The demo was as following:

图片包含 屏幕截图

描述已自动生成

**4. Data Mining and Analysis**

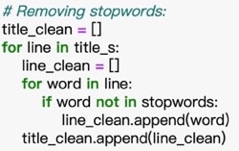
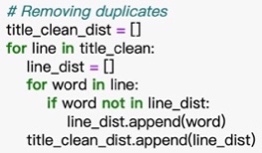
4.1 Use word breaker jieba library and achieve Word segmentation for each title by lcut function.

4.2 Filter every element (str) in title\_s（list of lists form） to remove unnecessary words from the stopwords list then import the stopwords list.

图片包含 屏幕截图

描述已自动生成

4.3 Remove stopwords and duplicates.



4.4 Converting title\_clean\_dist to only one list; Turning the list allwords\_clean\_dist to DataFrame; Classifying and summarizing the words filtered and de duplicated.

 图片包含 屏幕截图

描述已自动生成

**5. Word Cloud Visualization**

5.1 Use libraries: Matplotlib, wordcloud and imread, to visualize word cloud as following.

图片包含 屏幕截图

描述已自动生成

5.2 Conclusion: (1)Words such as '沙发', '组合', '客厅', '小户型' and '简约' are frequently used. (2)Among the words describing sofa material, '布艺沙发' appears the most frequently. (3)Among the words describing sofa style, '简约' appears the most frequently. (4)Among the words used to describe the type of house the sofa is suitable for, '小户型' appears the most frequently.

5.3 Prediction: (1) As for sofa material, due to the large sales volume of fabric sofa, that is, the demand is large, it is expected that the enterprise will increase the supply in response to the purchase demand. (2) About sofa style, simple style is very popular. More enterprises will produce simple style sofas, and more businesses will sell simple style sofas. (3) About the type of houses the sofa is suitable for, it is very popular for small family and more businesses will sell sofa suitable for small family.

## 6. the Impact of Different Keywords on Sales

Using Numpy to process the data we need.

图片包含 屏幕截图

描述已自动生成图片包含 文字

描述已自动生成

Combining data tables (concatenation) using Pandas

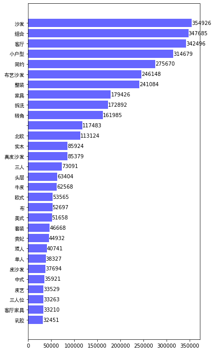


 图片包含 屏幕截图

描述已自动生成

Visualizing data by Numpy and Matplotlib.

图片包含 屏幕截图

描述已自动生成

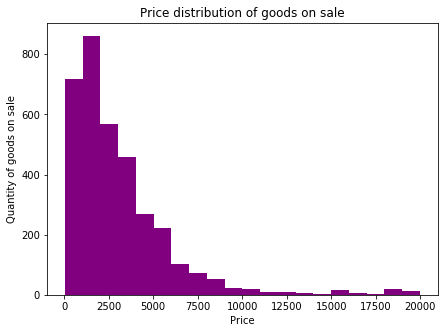
## 7. the Price Distribution of Commodities

7.1 Analyze and visualize the price distribution of goods. In order to visualizing more intuitive, we only considering products with price less than 20,000.

图片包含 屏幕截图

描述已自动生成图片包含 屏幕截图

描述已自动生成

7.2 Conclusion: (1)The largest number of sofas on sale with a price between ￥1,000-2,000. (2)The prices of sofas are mainly between ￥0-4,000. (3)The quantity of goods shows a downward trend with the price rising. The higher the price, the less goods are on sale.

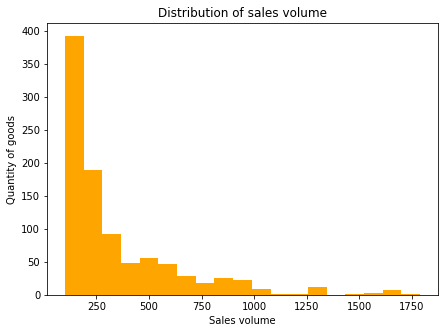
7.3 Prediction: As a necessity of daily life, the price elasticity of demand is relatively small. Most middle-class consumers will tend to buy low-end sofas, and businesses will also focus on selling sofas with a price range of ￥ 1000-2000. In the future, there will still be businesses selling high-quality and high-grade expensive sofas to high-end consumers, but the number of sales is limited.

## 8. the Sales Distribution of Commodities

## 8.1 In order to make the visualization more intuitive, only considering products with sales more than 100, less than 20,000.

图片包含 屏幕截图

描述已自动生成



8.2 Conclusion: (1) Goods with a sales volume in the range of 100 to 2000 accounted for only 27.1%, of which the sales of 100-200 is the most, followed by 200-300. (2) When the sales volume is between 100 and 500, the quantity of goods shows a downward trend with the sales volume, and the trend is steep, sales of most goods are very low. (3) There are few goods with sales volume of more than 1000.

8.3 Prediction: Our conclusion is in line with the 28 law of Pareto, an Italian economist. That is to say, 20% of all kinds of goods bring 80% of the sales volume, so Pareto believes that only this part should be retained, and the rest should be discarded. According to this Law and the actual data we analyze, we predict that in the future, the sales of those small amount of goods will be more, and some of the goods with low sales will be eliminated by the market.

## 9. Average Sales Volume Distribution of Commodities in Different Price Ranges

## 图片包含 屏幕截图 描述已自动生成9.1 According to the prices of goods, the price range is divided into 10 equal parts, and match each commodity with the corresponding price range.

## 图片包含 屏幕截图 描述已自动生成

## Then, calculate the average sales volume of goods belonging to the same price range.

## 图片包含 屏幕截图 描述已自动生成

## After calculating, we use matplotlib to visualize average sales volume distribution of commodities in different price ranges.

图片包含 屏幕截图

描述已自动生成9.2 Conclusion: (1)The average sales volume of goods with price between ￥40-588 is the highest, followed by ￥998-1380, and the lowest with price over ￥6600; (2) Overall, it shows a downward trend.

9.3 Prediction: In the future, the commodities with price range of ￥ 40-588 will continue to dominate, and most of the middle and low-end sofas will dominate the market. Businesses can find their own position and choose whether to benefit from the sales of middle-class consumers or provide expensive high-end sofas for a few people.

## 10. the Influence of Price on Sales Volume

## 图片包含 屏幕截图 描述已自动生成图片包含 屏幕截图 描述已自动生成10.1 In order to make the visualization more intuitive, only considering products with price less than 20,000.

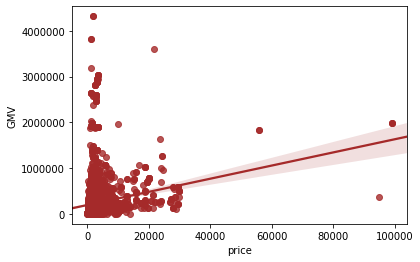
## /var/folders/r9/gy6x9k3s03x7vxp9wxmlq3100000gn/T/com.microsoft.Word/Content.MSO/E2C947E0.tmp10.2 Conclusion: (1) General trend: the lower the price of a commodity, the higher its sales volume. (2) Commodity with a price range of ￥0-4000 are easier to sell.

## 10.3 Prediction: Most businesses will choose to provide middle and low-end sofas to middle-class consumers, of course, there will be a relatively small number of businesses dedicated to selling expensive high-end sofas for specific consumers

## 11. the Influence of Price on Sales

11.1 Use Seaborn library to visualize and analyze the influence of price on sales.





11.2 Conclusion: (1) General trend: as can be seen from the linear regression fitting line, with the increase of price, the total sales shows an upward trend. (2) The prices of most commodities are low, and the total sales is also low.

11.3 Prediction: low and medium-end sofas with relatively low prices are favored by consumers. According to the conclusion, we can see that most consumers are not very sensitive to the price, and businesses can get benefits by increasing the price appropriately.

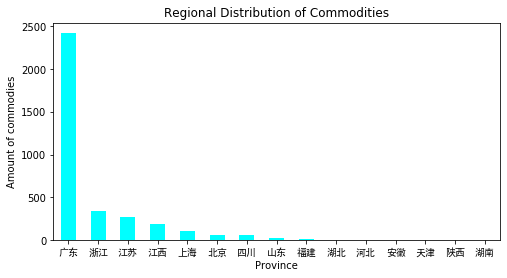
## 12. the Regional Distribution of Commodities

## 12.1 Use Numpy and Matplotlib to visualize and analyze the regional distribution of commodities.

图片包含 屏幕截图

描述已自动生成 图片包含 屏幕截图

描述已自动生成



12.2 We realize visualization of geographic information and commodities information by import map from pyecharts library. It has the dynamic characteristic and interactive feature. So that can make our visualization more interesting attractive and efficient.

图片包含 屏幕截图

描述已自动生成

图片包含 文字, 地图

描述已自动生成

## 12.3 Conclusion: (1) The number of commodities from Guangdong Province is the largest, followed by Zhejiang and Jiangsu. (2) In particular, the number of goods from Guangdong far exceeds that from Jiangsu, Zhejiang, Shanghai and other places, indicating that for the sub-category of sofa, stores in Guangdong are dominant. (3) The number of commodities from Jiangsu, Zhejiang and Shanghai is almost the same.

12.4. Prediction: the enterprises of sofa production and sales in mainland China will still be mainly concentrated in the southeast coastal areas, and Guangdong Province is dominant with obvious advantages. According to the theory of industrial cluster in economics, more sofa manufacturers will invest factories and sell sofa in southeast coastal areas.