Project 0: Exploratory Analysis

Chen Nan

Due date: Jan 25, 2023

1 Introduction

This project provides a test bed for basic data maniputation and exploratory analysis. The data was collected from a shopping App, where people can sell and buy new or used products. Columns of the data are organized in the following ways

1. train_id: the id of the product
2. name: the title of the product
3. item_condition_id: the condition of the product provided by the sellers
4. category_name: category of the product
5. brand_name: the product's brand name
6. price: the price that the product was sold for
7. shipping: 1 if shipping fee is paid by seller and 0 if shipping fee is paid by buyer
8. item_description: the full description of the product

The dataset is provided in **"P0_price.tsv"** which is a tab separated file with the header. Please read the data into your project (Jupyter Notebook), and perform the following analysis or exploration. Please take note of following points in your submission.

- Please using one block of codes to address each one of the questions below.
- Organize the questions and outputs in the same sequence as shown below.
- Attempt to use the simplist way/function to answer the questions.
- Only submit the ".ipynb" file. Do NOT submit the data file.

2 Exploratory Analysis

Please also quote the following questions in your notebook as the *Markdown cell*.

1. How many number of samples are included in the dataset?
2. List the number of missing values in each of the 8 columns.
3. Split the data in the column "categories" into 3 columns:"main category", "first subcategory", and "second subcategory"
4. List the unique "categories" in the data.
5. List the number of products each category have, ranked in descending order.
6. Plot the sorted frequncy of categories.
7. List the top 10 first subcatgories by product count.
8. List the top 10 first subcatgories by product average price.
9. List the top 10 popular brands by item counts.
10. Historgram of the product counts by item conditions.
11. Plot and compare the price distribution for products with shipping paid by sellers and buyers respectively.
12. Replace the numbers in "item_condition_id" by letters, i.e., "$1 \Rightarrow A, 2 \Rightarrow B$,...". List the first 5 samples after replacement.
13. Find out the "category" and "condition" combination such that it has highest percentage of shippings paid by sellers.
14. Find out the "brand" and "condition" combination such that it has highest percentage of shippings paid by buyers.
15. Generate the word cloud based on the "descriptions" of all items.
16. Generate the word cloud based on the "descriptions" of the top 10% most expenstive items.
17. For all items in the first subcategory "athletic apparel", plot and compare the price distributions, for different conditions.
18. Create a pie chart of main categories, proportional to sales revenue.
19. (Optional) Other exploratory analysis you feel relevant and informative.