Semester Final Project Summary

Fall-2021

**ECON 8320 -Tools for Data Analysis**

* Piyush Basia

**“Analysis of ‘Clothes and Shoes’ products for Men and Women listed on etsy.com”**

**Data Source:**

The data source for this project is: <https://www.etsy.com/>

**Introduction:**

As Part of the final project for ECON-8320\_Tools for Data Analysis class, I have conducted my proposed research based on my project proposal. I have narrowed the item category to 'clothes and shoes' to avoid handling extremely large and unnecessary data. The code snippet works perfectly and can be used to scrape other pages data. I have tried to utilize almost everything I learned from this class. The resulting output and graphs satisfy my research purpose aptly.

**Data Collection:**

I have used BeautifulSoup, which is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping. I utilized the web scraping technique that I learnt in class to create a data frame where I scraped through the etsy.com website to find out the cloth and shoe products listed as suitable for her and for him.

**Methods:**

* **(Scraping and Cleaning)**

I have used ‘mimir notebook’ to do all my analysis for this project.

At first, I have imported all the required python libraries into my notebook. Then I have created a function named ‘collectProductData’ to collect product details by scraping Etsy.com web pages. This Function and further analysis use all the important methods I learned in the class like loops, functions, classes, regular expression, NumPy, pandas, web scraping, Plotly etc.

Next, I have used this function to scrape the site for men and women’s products listed in site under ‘clothes and shoes’ section and store the data in different data frames named ‘MenCS’ and ‘WomenCS’.

After that, I have dropped any duplicate records and merged Men and Women data into one data frame named ‘Alldata\_CS’.

* **(Plotting and Analysis)**

Once I have gathered all the required data, I have created a plot to display price variation of products for men and women and saving the figure in fig1. The Results are satisfying.

Next, I have found the Product/s from the dataset which has/have the most and least number of customer ratings.

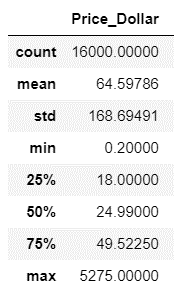
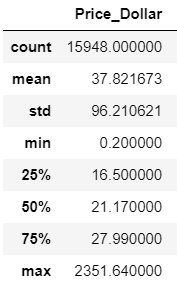
After this, I grouped the men’s and women’s data set separately on basis of ‘Rating\_out\_of\_5’ values. Using describe() function found that rating min value is 3.0 and highest value is 5.0. So, interval distributed from 3.0 to 5.0 unevenly. so that we get a proper view of the data.

Finally, using the describe() function, I found that ‘Discount\_Percentage’ ranges from 10 to 70 in the data frames. So, I created groups from the Men/Women data on basis of grouped Discount\_Percentage. I plotted Discount\_Percentage data for men and women to find any trendline is Discount Percentages at etsy.com

**Results:**

1. The first analysis of variation of products prices for men and women in etsy.com is shown below fig1. The below tables indicates that the average price for Items listed for Women is $64.59786 while for men it’s just $37.82163. Though the minimum price for listed products is same for both, there is a huge difference between the costliest item for them.

**Women Men**

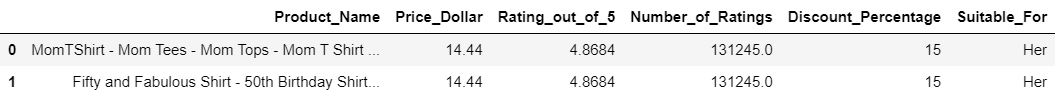
 

This plot describes, how there are much more products listed for women lies in a higher price range than the products listed for men on their site.



fig1

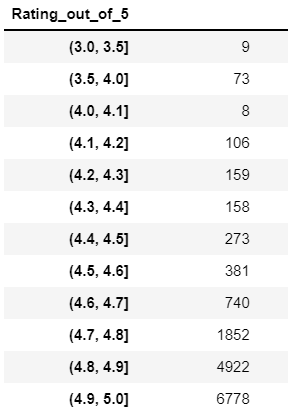
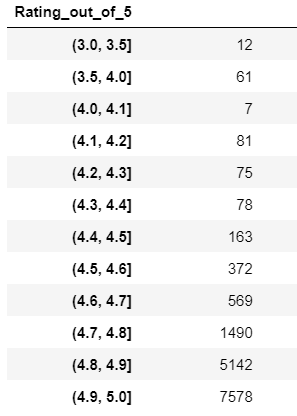
1. Second analysis is around finding the Product/s from the dataset which has/have the most/least number of customer ratings. While there are only two records for the most numbers of rating (131245), I found there are 12 records with least number of ratings (1). The below figures show the results.





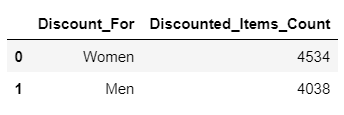
1. Next, I explored the customer ratings for each product listed for men and women. It shows that for both the cases highest rating is between 4.9 and 5.0 interval.

Women Men

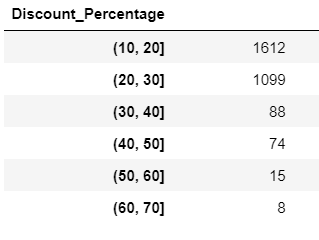
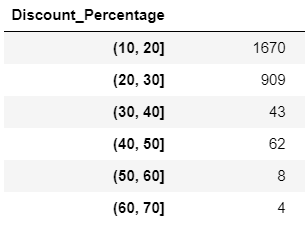
1. Lastly, I performed a comparison analysis for ‘Discount\_Percentage’. The below table shows the total number of discounted items listed in etsy.com website. The discount for both data set ranges from 10% to 70%.

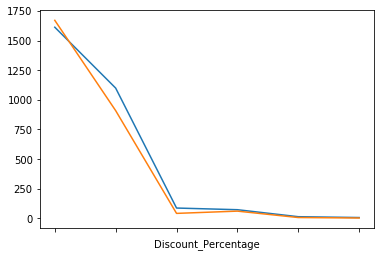
There’s not much difference in the numbers for men and women.



I plotted the Discount\_Percentage data for men and women to find any trendline is Discount Percentages at etsy.com. It shows a similar trend for both datasets. The mostly used discount percentages are 10 to 20% and there are only 8 products for women and 4 products for men which has a discount rate of 60 to 70%.

Women(orange) Men(blue)



**Conclusion**: I would like to thank Dr. White for all his efforts to teach us all these tools and techniques for data analysis. I’m very happy with this final project and satisfied with the analysis.