Ordered Sets In Data Analysis - Lazy Learning

Frank Lawrence Nii Adoquaye Acquaye

December 15, 2019

1 Problem Statement

I was intrigued by the possibility of predicting a persons gender based on their choices hence I selected a binary classification task based on this [1] data from kaggle. The data was collected by asking 66 people, 33 male and 33 females about their preferences. The questions were quite generic hence may pose a challenge to classification.

2 Data Exploration

I performed initial exploration of the data in order to have a better understanding of the the data. After my exploration some duplicates were found in the data. The duplicates were removed since they would mostly be redundant for lazy FCA.

3 Scaling

After exploring the data it was found that all the attributes were categorical in nature hence the categories were used in the lazy FCA. The Categories for the various attributes are listed below

Favorite Beverage	Favorite Soft Drink	Favorite Color	Favorite Music Genre
Whiskey	Coca Cola/Pepsi	Cool	Rock
Beer	7UP/Sprite	Warm	Electronic
Other	Fanta	Neutral	Hip hop
Doesn't drink	Other		R and B / soul
Wine			Pop
Vodka			Jazz/Blues
			Folk/Traditional

[1] gender classification `https://www.kaggle.com/hb20007/gender-classification`.

4 Statistical Findings

	Favorite_Color	Favorite_Music_Genre	Favorite_Beverage	Favorite_Soft_Drink	Gender
0	Cool	Rock	Doesn't drink	Coca Cola/Pepsi	M

Figure 1: Most Frequent Attribute for Male

	Favorite_Color	Favorite_Music_Genre	Favorite_Beverage	Favorite_Soft_Drink	Gender
0	Cool	Pop	Beer	Coca Cola/Pepsi	F

Figure 2: Most Frequent Attribute for Female

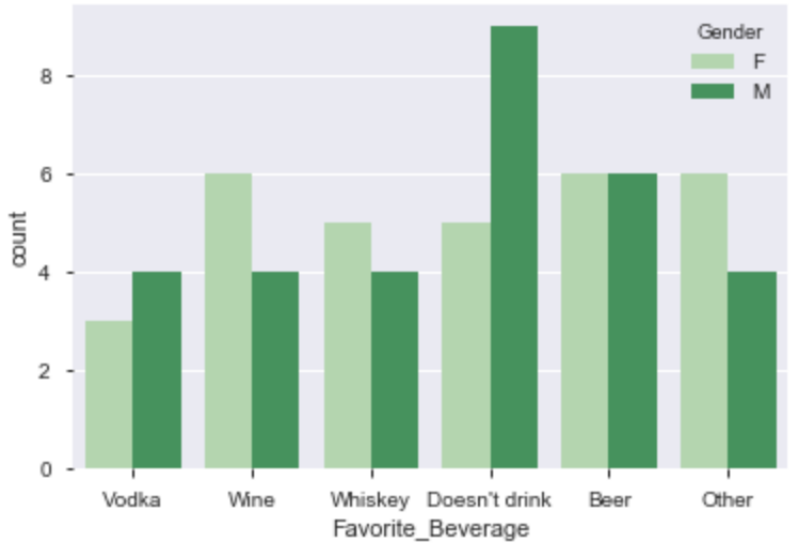

Figure 3: Bar Chart of Favourite Beverage By Gender

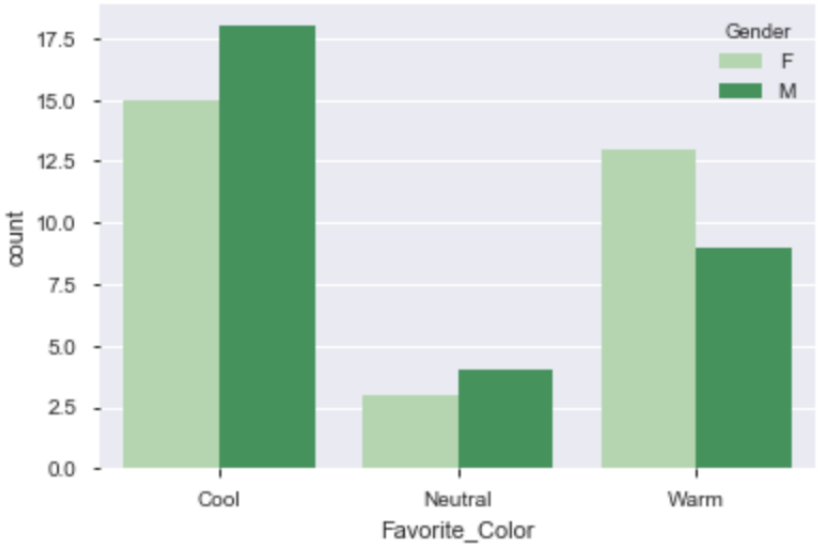

Figure 4: Bar Chart of Favourite Colour By Gender

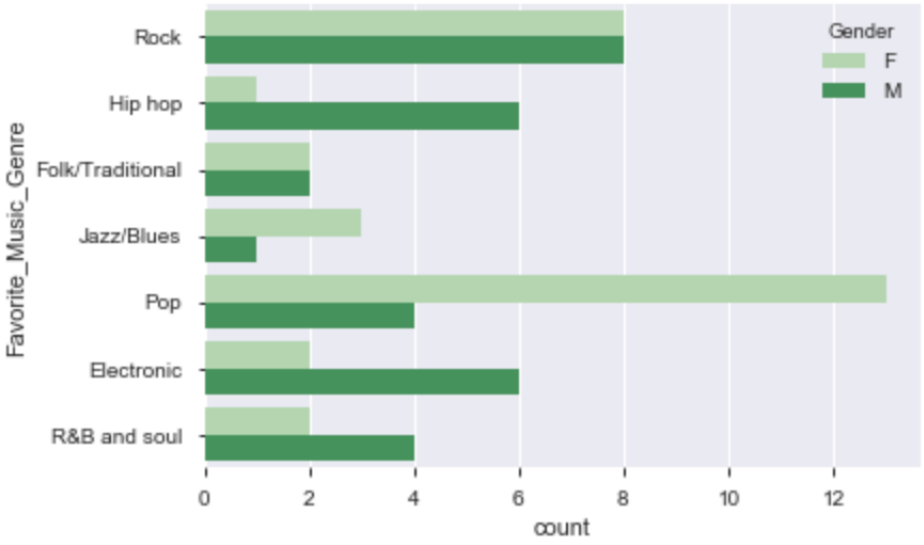

Figure 5: Bar Chart of Favourite Music Genre By Gender

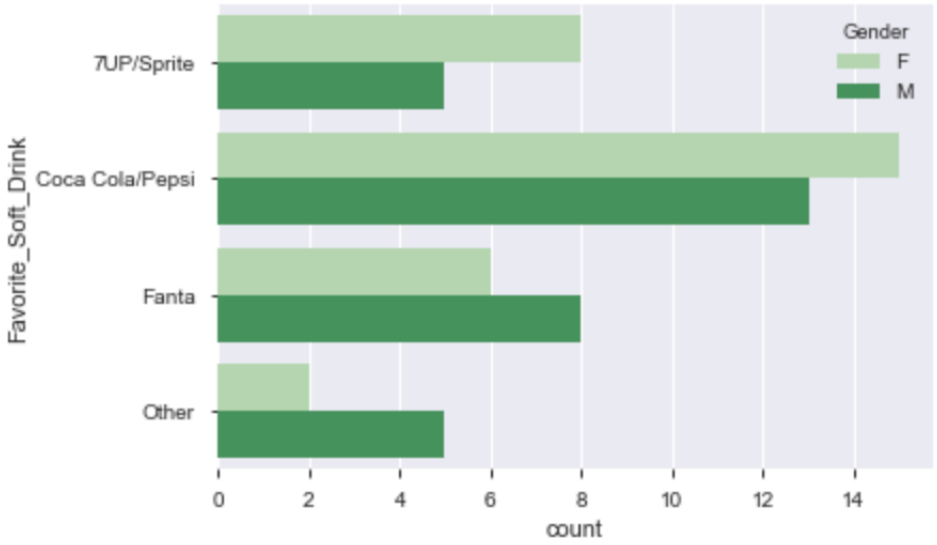

Figure 6: Bar Chart of Favourite Soft Drink By Gender

5 Algorithm

step1: Find support based on:

$$|\frac{Supp_F(P)}{|F|} - \frac{Supp_M(P)}{|M|}| \geq 0.2$$

step 2: All hypothesis that fulfill the requirement in **step 1** are accepted as a strong hypothesis

step 3: Prediction is done by assuming objects are negative(male) unless one of the hypothesis in **step 2** holds then data is classified as positive

6 Conclusion

After running some tests the following results were obtained:

```
accuracy = 0.58
recall = 0.5
precision = 0.6
F1 Score = 0.55
```

Figure 7: Results of Tests

The model's performance is not as great, this might primarily be due to the nature of the data. i.e The data exploration shows that **Pop Music** is the clear differentiator between male and female and this is supported in **Figure 5.** Also the model returns **Pop Music** as the clear hypothesis for classification but this is clearly contradictory according to the male class.

7 Future Work/Recommendations

1. It is observed that tweaking the threshold returns different test results. Perhaps an automated approach can be used to generate the best results.
2. Some Cross Validation will also be useful for the model
3. Possible tests can be run with other data sources to help improve the model or confirm the efficiency of the algorithm.