Data Exploration and Visualization

COSC 4610

Charlie Irmiger

11/5/2020

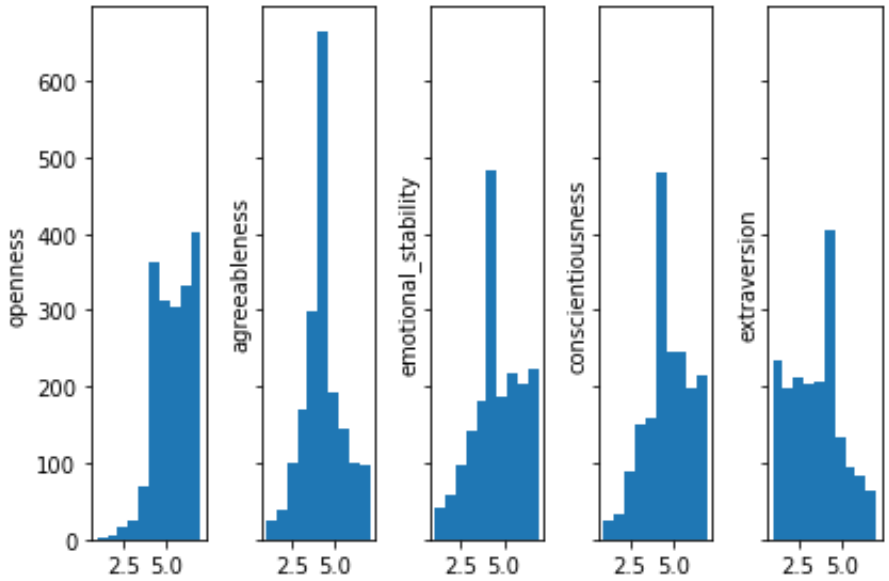
1. Cleaning

Before I began any visualization and exploration of my dataset, I needed to clean the data. This was a relatively straight forward process since my data was fairly clean already. The biggest problem was duplicate rows, of which there were 14. After removing the duplicates, my main focus was to scale down the dataset into something that was easier for me to process, understand, and make inferences from. Therefore, I dropped each of the movie columns and the predicted ratings columns since I was mainly interested in the 'enjoy_watching' column as my explanatory variable, which is the rating on a scale from 1-5 of the entire list of 12 movies. My clean data frame was then complete. I have 8 total columns, five of which are dedicated to the "Big 5" personality traits.

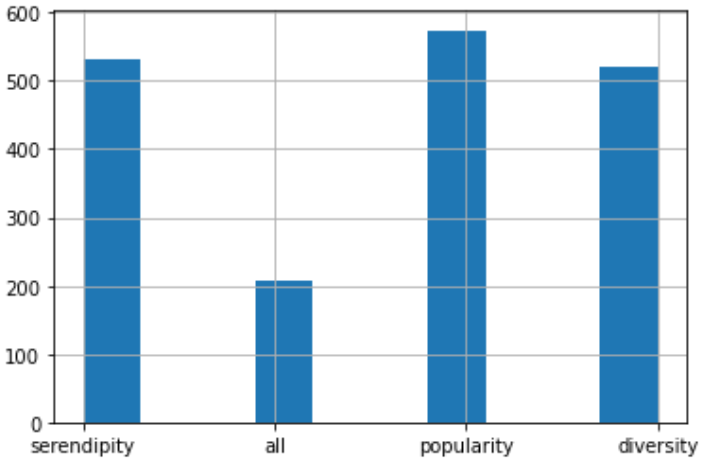
	userid	openness	agreeableness	stability	conscientiousness	extraversion	is_personalized	enjoy_watching
0	8e7cebf9a234c064b75016249f2ac65e	5.0	2.0	3.0	2.5	6.5	4	4
1	77c7d756a093150d4377720abeaeef76	7.0	4.0	6.0	5.5	4.0	2	3
2	b7e8a92987a530cc368719a0e60e26a3	4.0	3.0	4.5	2.0	2.5	2	2
3	92561f21446e017dd6b68b94b23ad5b7	5.5	5.5	4.0	4.5	4.0	3	3
4	030001ac2145a938b07e686a35a2d638	5.5	5.5	3.5	4.5	2.5	2	3

2. Initial Exploration

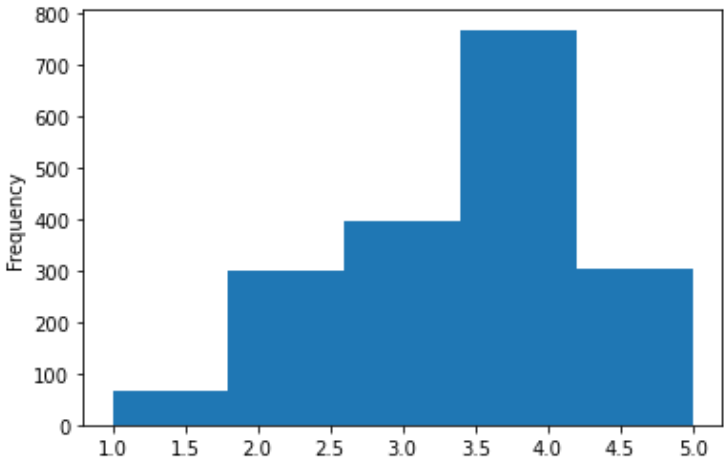
The first task in exploring my data was to get a general understanding of the various column distributions. The "Big 5" personality model is a widely accepted and popular model in the world of psychology and has been used to explain variance in numerous statistical environments [1][2]. The first step was to create simple visualizations for the distributions of the 5 traits. Since most of the column data consists of categorical data, I used histograms for many of the visualizations.



Agreeableness was the only metric that was close to normally distributed. Emotional stability, conscientiousness, and openness were all skewed right, whereas extraversion was skewed was skewed left.



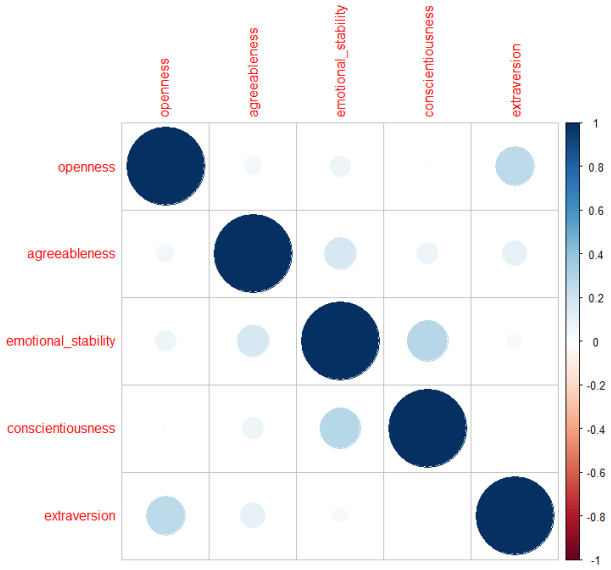
Next was the distribution of the “Assigned metric” column, which is the genre of the movie shown to the subject. All three genres are similar in size. “All” means that the movie contains all three genres.



Score given to movie (1-5)

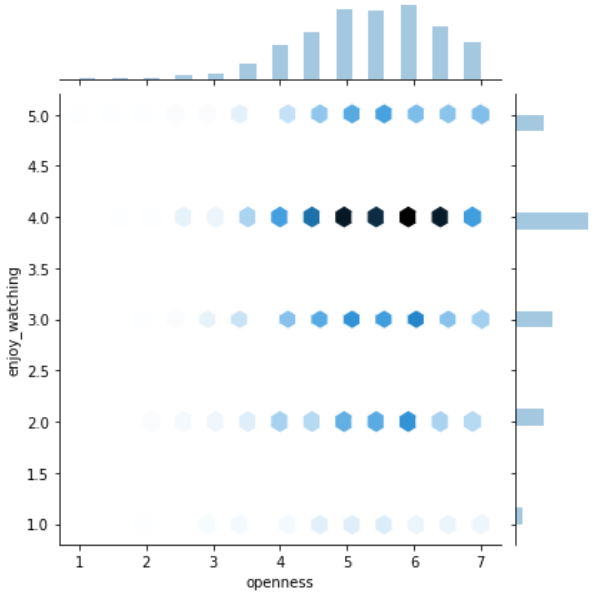
Next, I wanted to find the distribution for the overall scores given to the movies. Users were asked to rate a list of 12 movies of a selected genre on a scale from 1 to 5. The scores were skewed right, with a most common score given of 4.

3. Correlations

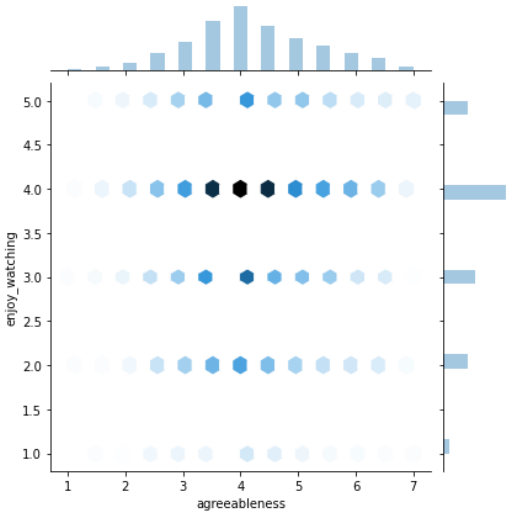


My initial curiosity caused me to look for any obvious correlations between any of the 5 personality traits. There were not any significant correlations, which allowed me to isolate each personality trait against the score given.

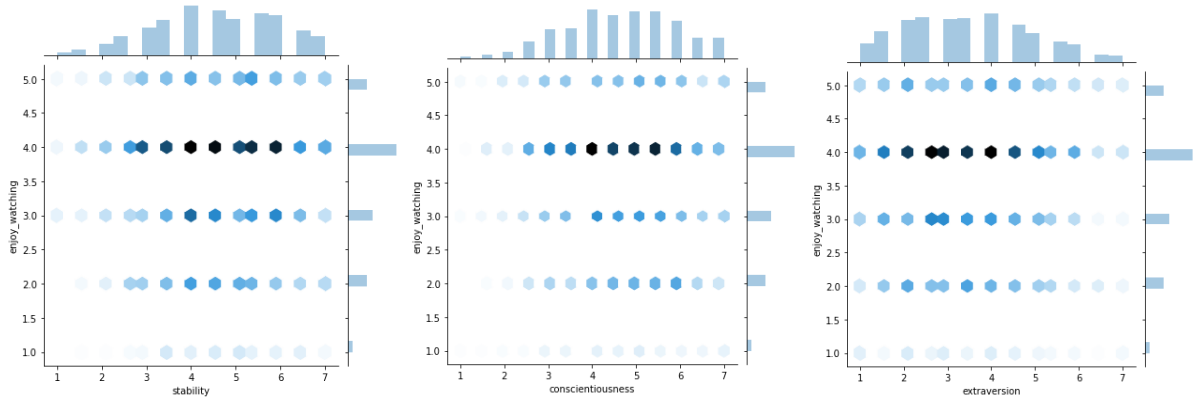
My next task was to try and find correlations between each of the five personality characteristics and the score given to the list of movies. I used *jointplot* from the seaborn package in python, which plots the relationship between two variables as well as the distributions for each on their respective axis. The more data around an intersection, the darker the color.



Openness was the first trait I plotted, as it was the trait I expected to have the largest impact on the score given. What I found was that there was no clear correlation between openness and the score given to the 12 movies.



Next was agreeableness, which again showed no clear correlation between the level of agreeableness and the score given.



The remaining three personality traits were also disappointments, as I saw no clear correlation between the score of the trait versus the score given to the movie. My examinations were confirmed by correlation calculations performed on each of the personality columns versus the "Enjoy watching" column.

	Openness	Agreeableness	Stability	Conscientiousness	Extraversion
Correlation to "Enjoy watching"	0.0639	0.0368	-0.0022	-0.0435	0.0272

References:

[1] Jonathan Haidt. 2012. The righteous mind: Why good people are divided by politics and religion. Vintage.

[2] Rothmann S, Coetzer. 2003. The big five personality dimensions and job performance. *SA Journal of Industrial Psychology*. 29. doi:10.4102/sajip.v29i1.88.

[3] Nguyen, T.T., Maxwell Harper, F., Terveen, L. et al. User Personality and User Satisfaction with Recommender Systems. *Inf Syst Front* 20, 1173–1189 (2018). https://doi.org/10.1007/s10796-017-9782-y

[4] Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A Very Brief Measure of the Big Five Personality Domains. *Journal of Research in Personality*, 37, 504-528.