Case Study

James Wang

Individual Project: Interim 1 – Analysis Plan

Project Title: Probing the Implicit Metrics of Hidden Desires – A Study in Pink

Project Overview:

Sexual desires and preferences have been a hard topic to study. If the study of public opinion in politics is difficult due to the survey constrains such as voluntary response bias and social desirability bias, the study of sexual preference, being a sensitive and private topic, is even more challenging. Yet, while people might falsify their preferences during surveys, they are by and large honest during private internet activities. In this project, I aim to identify the drivers of sexual desirability, at least in masturbation. Scraping porn-star data from pornhub.com, I will try to study both the biological, social, and behavior factors that contribute to the ranking of porn-stars, which serves as a benchmark for people’s sexual preference.

Data Collection:

The constructed dataset will consist of the 12k registered and ranked porn-stars on pornhub.com, the world’s largest adult video site. I will harvest the variables that I believe contribute to their ranking on the site.

*Biological data*: Dimension, Weight, Height, Age, (no ethnicity available)

*Marketing data*: number of views, number of videos, number of subscribers, average rating of videos

*Behavior data*: sex, sexual orientation, most prevalent video category, average number of actors in video

Modelling:

I plan to fit a regression (probably log-linear) to predict a star’s ranking base on these variables. Due to the foreseeable possibility of multi-collinearity between variables, I will either perform variable selection and shrinkage, or switch to regression tree models.

Limitations:

The star ranking is an arbitrary term by the website. Using this term as the response might not be the best metric.

Current Progress and Challenges:

I’ve already scrape the bio-page url of 12k stars from its 127 pages of the list. The website kicks me out when it identifies the scraper. In response, I put a time-out between each loop -- a random number sampled from a truncated normal distribution with a mean of 10 sec – to mimic natural browsing behavior. I believe this procedure will also work when I next scrape the 12k individual bio pages.

Note:

Traditionally, when collecting data on socially controversial topics, there is the concern that the over-exposure of these topics on the researcher would be detrimental to his wellbeing. However, the method of webscraping, which automates the data collection process, minimizes the risk of harm associated. This method enables researchers to answer these sensitive questions such as sexuality with objectivity and clarity. And similarly, the findings of the study will be presented in a socially acceptable, purely quantitative manner, with no display of unwholesome content.