**Data Exploration of Indian Matrimonial Data**

**Abstract**

India has a long tradition of arranged marriage, and in recent years online matrimonial websites have become a popular destination where Indians seek out a partner. The author downloaded 325,000 profiles from a popular matrimonial website with the aim of seeing whether aggregate analysis of the dataset could be useful. The null hypothesis, i.e. the data would not be useful, was conclusively proven false. Several trends were clearly visible, indicating various sociological, cultural, economic and political biases, especially when the dataset was broken down by gender, location and economic class. Thus data from matrimonial websites can be a rich source of information about Indian society and culture.

**Introduction**

India has a long tradition of arranged marriage, in which parents or extended family help pick a spouse for eligible children. With the advent of the Internet in India, Internet entrepreneurs founded matrimonial services that are similar to American dating sites but focus on educational and family background rather than personality-based matching criteria[[1]](#endnote-1). These sites are used by parents, as well as by prospective brides and grooms. SimplyMarry.com is one such site that targets the urban, educated middle class in large cities and towns.

As finding a partner is a deeply personal and significant decision with long-lasting repercussions, users are often forthright in stating their preferences regarding their desired partner, and are unlikely to hide any deeply held cultural or sociological biases in order to be perceived as being politically correct. Further, these sites have fields for sensitive personal attributes such as caste and even skin tone.

Yet in the broader Indian context discrimination based on the caste system, which has existed for millennia, was constitutionally outlawed at the founding of the modern Indian republic[[2]](#endnote-2). Models and actors from films, television or advertisements almost always have very fair skin, despite the fact that very few of their real-world compatriots do[[3]](#endnote-3). Patriarchy and sexism are common. Recent cases of violence against women led to citywide protests that had political repercussions, as political parties finally began engaging with women’s issues as part of their campaign manifestos[[4]](#endnote-4).

The author conjectured that downloading and analyzing the aggregate data from SimplyMarry.com could provide insights into the abovementioned sociological biases. The data could potentially provide answers to interesting questions with associated policy ramifications, such as do these biases reduce among more educated users, or more urban users? Matrimonial data would probably be far more accurate a depiction of social biases than analysis of data obtained from surveys, for instance. The null hypothesis was that the data would not be useful at all.

**Methodology**  
SimplyMarry.com is one of the leading matrimonial websites, containing approximately 325,000 profiles, which were all downloaded using a Python-based web scraper. iPython notebooks were created to analyze different facets of the data. The data had to be cleaned, for which the Pandas and numPy frameworks were used, following which it was analyzed and graphed using MatPlotLib.

Total profiles available/downloaded: 325688

Number of males: 248893 (76.42 %)

Number of females: 76795 (23.58 %)

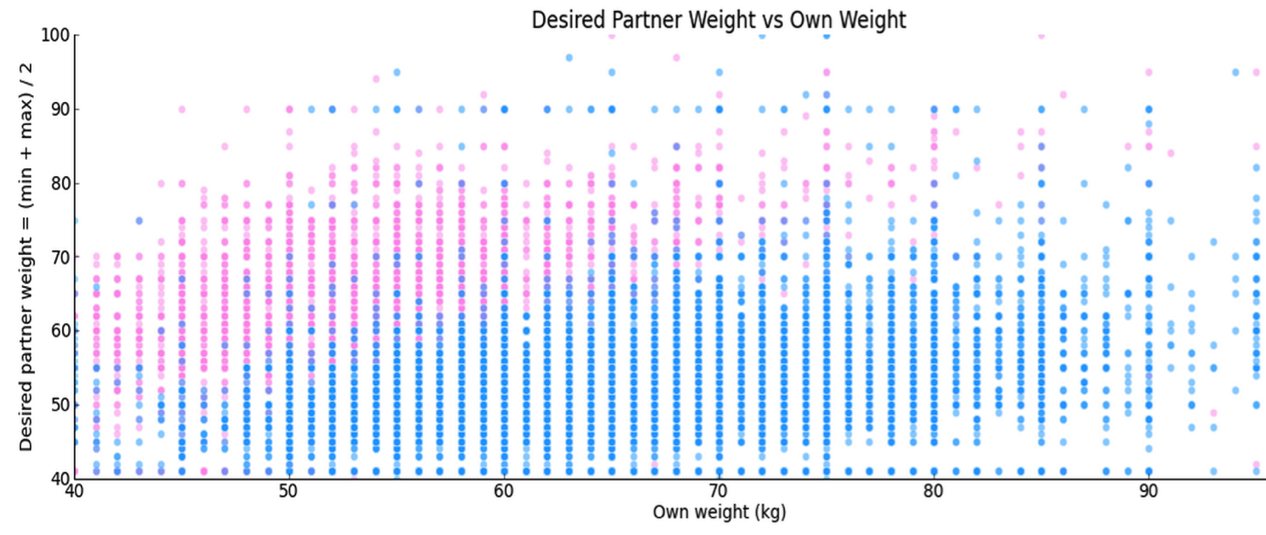
**Results & Discussion**

Due to the richness of the data, a very large number of graphs were generated. The knowledge and insights from almost every single graph could lead to its own individual conclusion. However as those insights are open to analysis and debate by experts in domains such as sociology and economics, only a few broad trends will be highlighted here.

Textual analysis- the *about me* textual portion of each profile proved very insightful. A naïve Bayes classifier, which comes with the SciKit learn package, was trained and tested on the data. In the training portion the package was told whether a profile was male or female, and for the test data the algorithm was made to determine the profile gender based on a bag-of-words representation of the *about me* section. Surprisingly, the classifier could be optimized for up to approximately 92% accuracy. The word clouds showing the words that most strongly determine male or female gender are shown in appendix A.

Graphical visualization

Gender (appendix B)- breaking up the data by gender proved very useful. It showed that although women are better educated on average, they have lower incomes. Men tend to create their profiles by themselves whereas for women it is more often created for them by parents or siblings. As a man’s weight increases, the weight he reports for his desired partner tends to stay fairly constant, whereas as a woman’s weight increases the weight she reports for her desired partner increases:



Economic class (appendix C)- As Indians move from middle to upper class, horoscope matching becomes less important. High class & middle class Indians tend to live in joint families, whereas the rich/affluent & upper middle class live independently. Upper class Indians are less traditional and more liberal, modern and international. Upper class Indians are more comfortable with drinking (or admitting they drink). Body type matters less for upper class Indians. Upper class Indians are also healthier, with fewer of them in the heavy category and more of them in the average category. Self-selection of skin tone doesn’t change much by class. In fact more upper class Indians categorize themselves as fair:



Location (appendix D)- For Indians residing in the US, skin tone doesn’t matter as much, and horoscope matching isn’t as important. However this variable doesn’t change much between large cities and small towns. Indians in small towns tend to live in joint (combined) households more often. Indians in the US are more modern and liberal. Indians in big cities are more moderate than the traditional Indians from small towns. Indians in small towns are least likely to drink (or admit they drink). This doesn’t matter as much for Indians in the US. Indians seem to get healthier as they move from small towns to big cities. For Indians residing in the US, matching based on caste isn’t as important. Interestingly, this doesn’t change much between big cities and small towns:



**Conclusions**

The data sheds light upon numerous sociological biases that exist among the Indian community, both inside India and abroad. Several of these biases became clearer after the data was broken up based on gender, economic class and location. A separate conclusion could be reached from practically every single graph that was generated.

However, it is more important to note that the null hypothesis, i.e. analyzing matrimonial data will reveal nothing of interest, was convincingly proven false. Numerous political, social, cultural and economic conclusions could be reached which can affect policy going forward.

This shows the simplicity with which matrimonial data can reveal insights. Apart from the web scraping and graphing, the statistical tools used were fairly rudimentary. Broad inferences could be made using very basic statistics. The data proved to be very rich and insightful, probably more so than a simple survey could have been. Yet, as far as the author is aware, this data source had not been analyzed in detail previously.

Due to the richness of the dataset there is room for more detailed analysis. For instance instead of breaking down the data by self-reported class, trends across different income groups could be analyzed. Similarly instead of breaking down the location into just three categories (Indian cities, Indian small towns/villages, US cities), more detailed GIS-type analysis could be done given the vast number of localities present in the dataset. No analysis was done on profile photos, which a small subset of profiles contained. Additionally the data could be analyzed from a privacy standpoint, as the dataset contains full names and other identifying information.

**Appendix A**

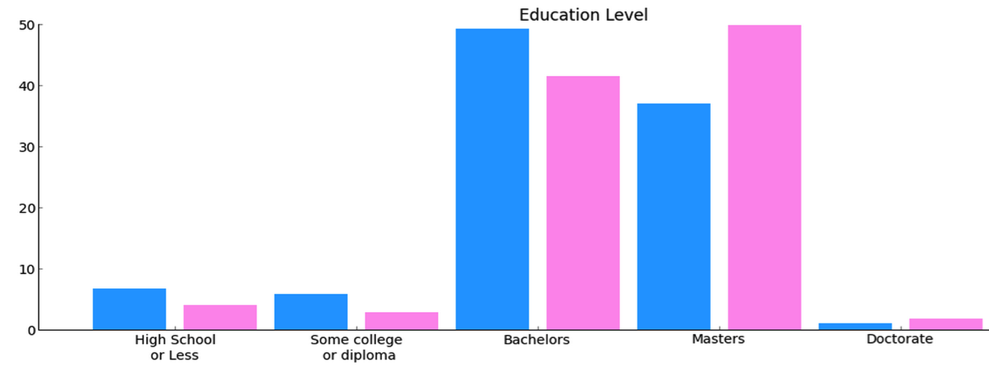
Words that most strongly determine a male profile

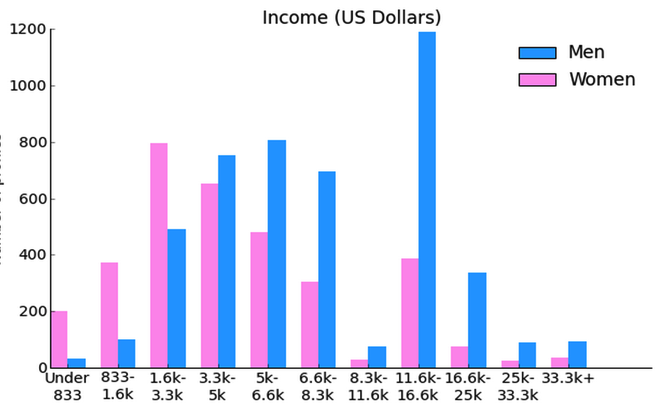


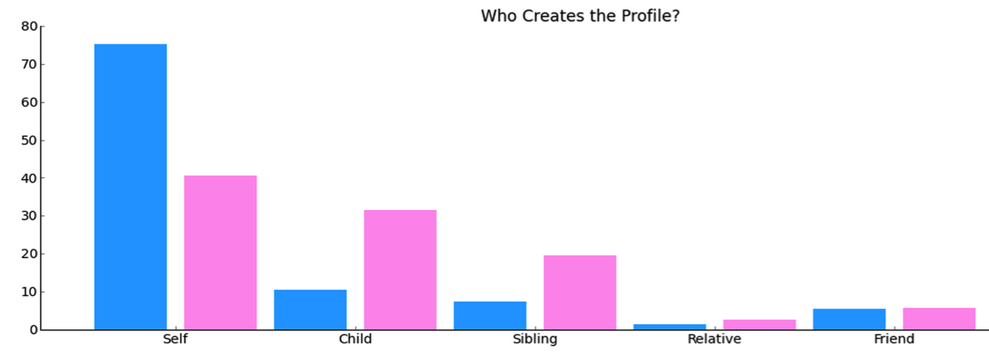
Words that most strongly determine a female profile



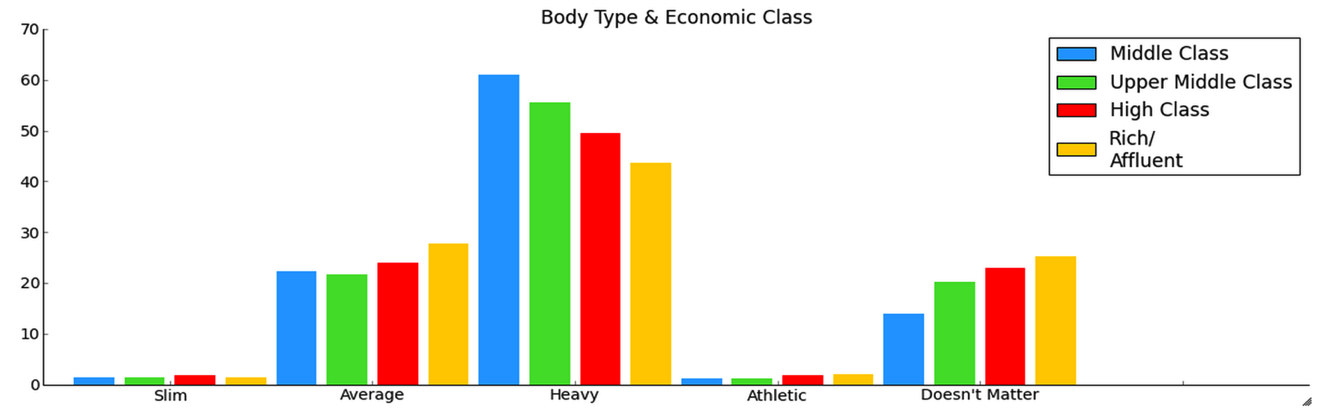
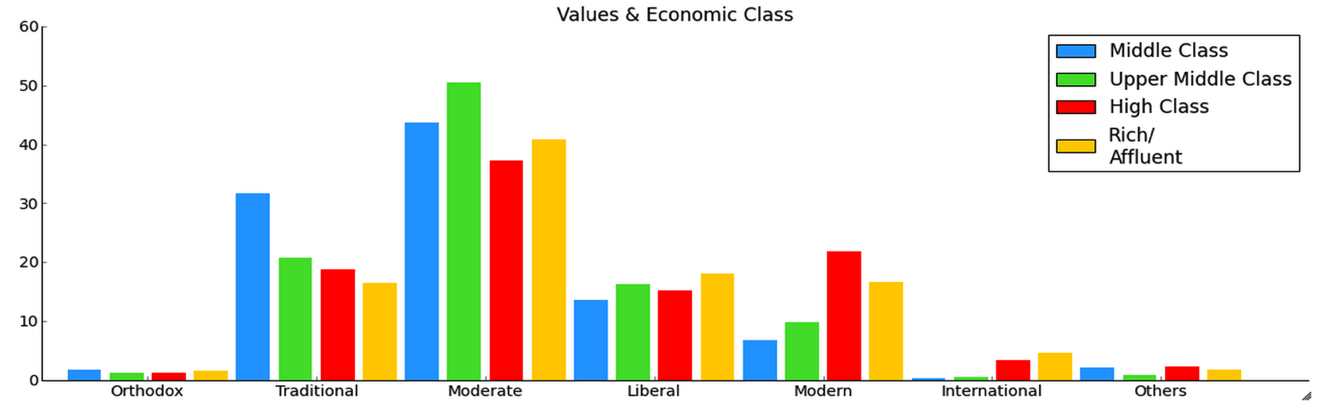
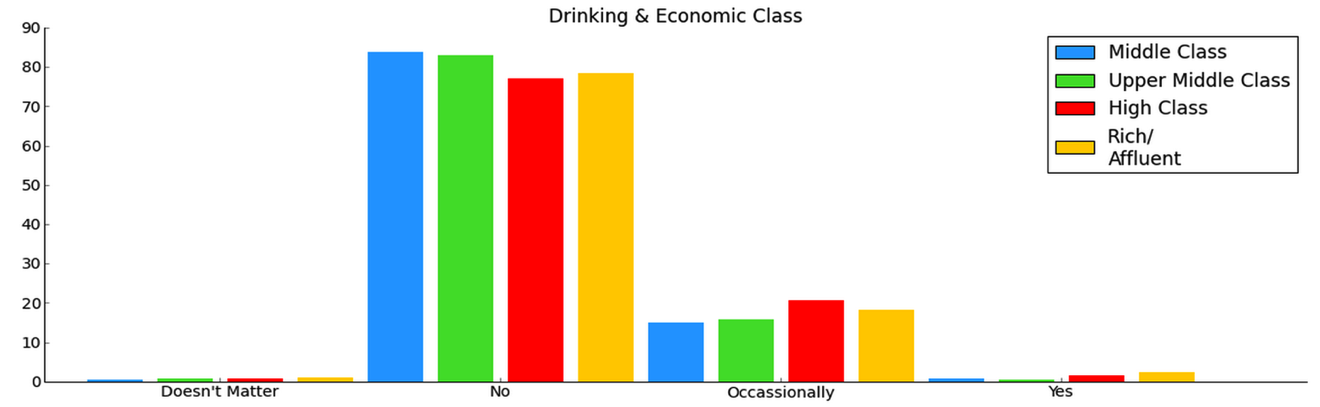
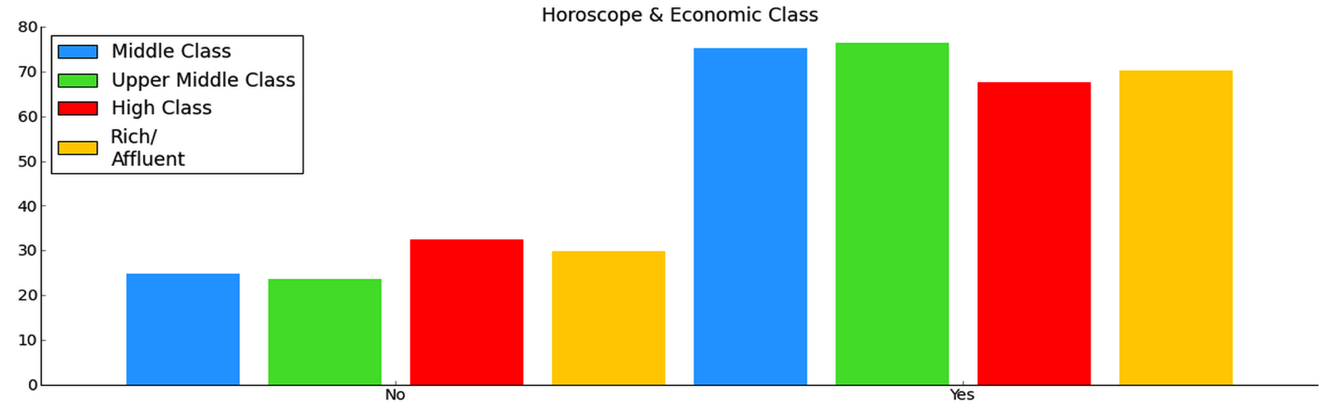
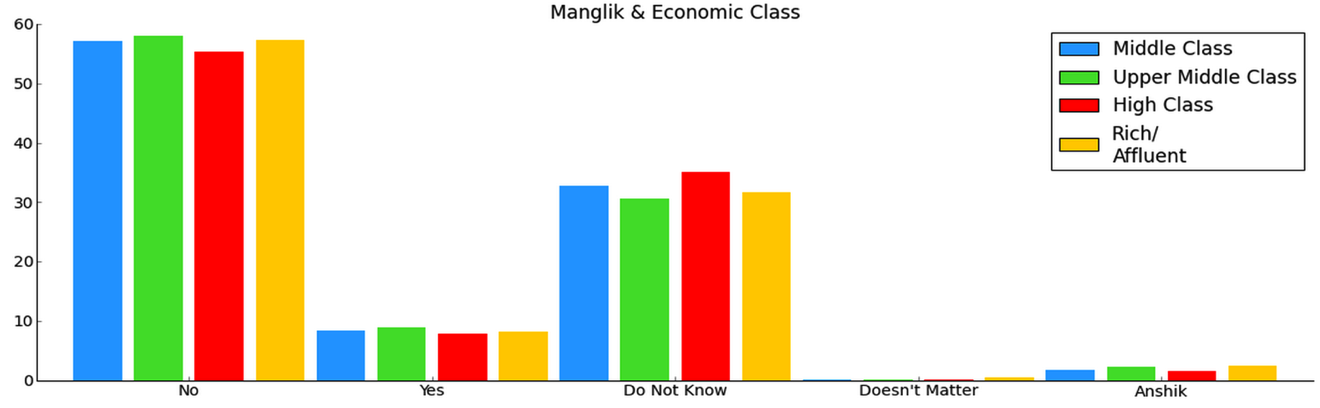
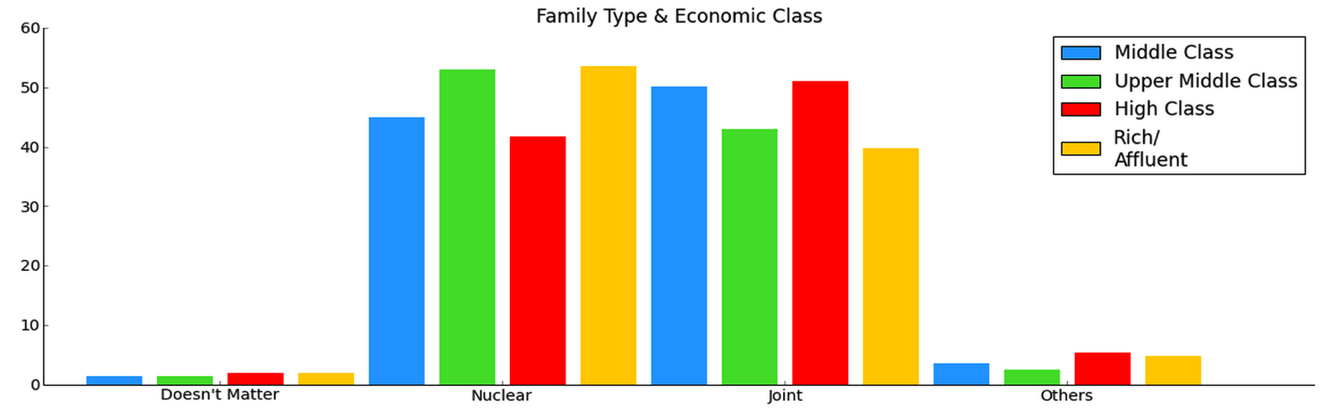
**Appendix B (gender: blue = male, pink = female)**



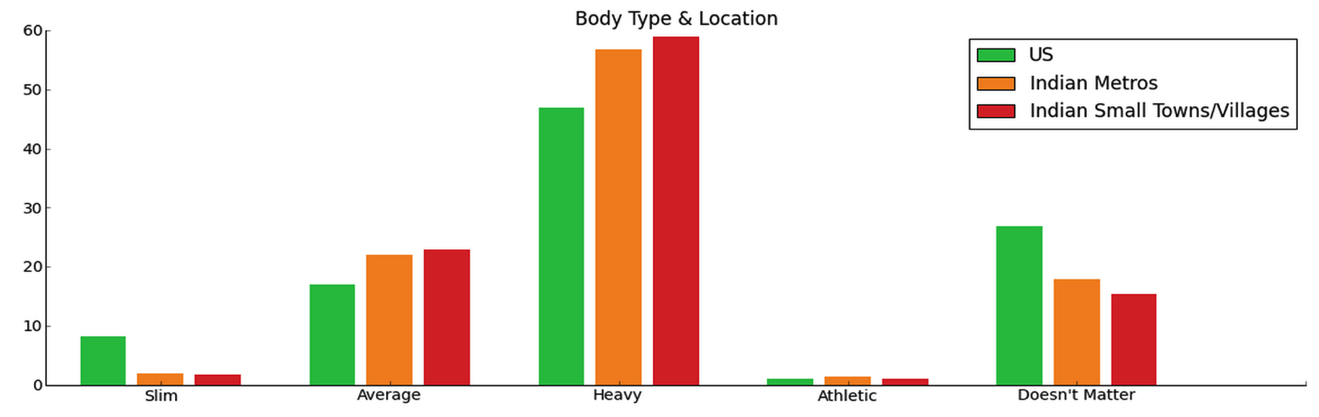
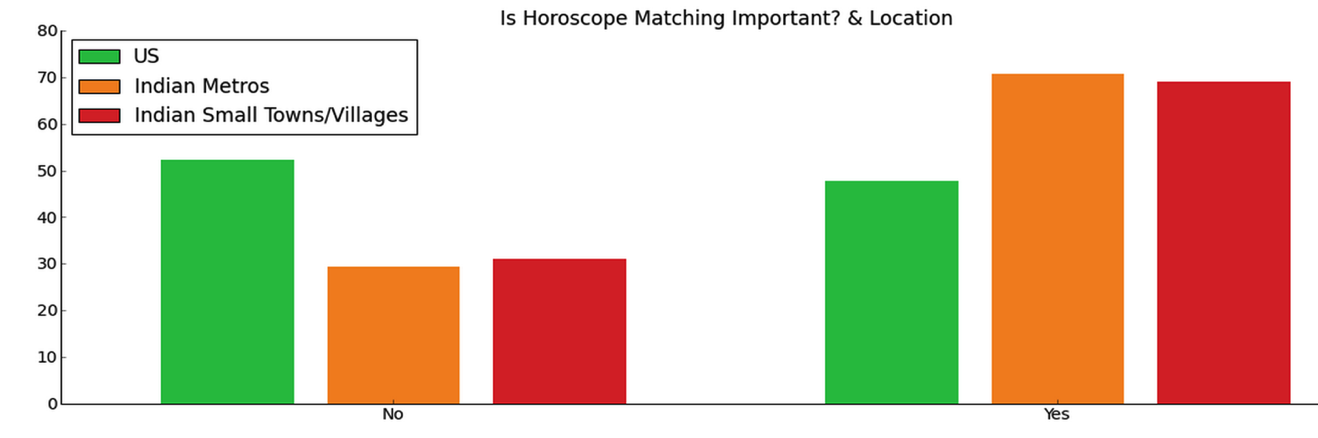
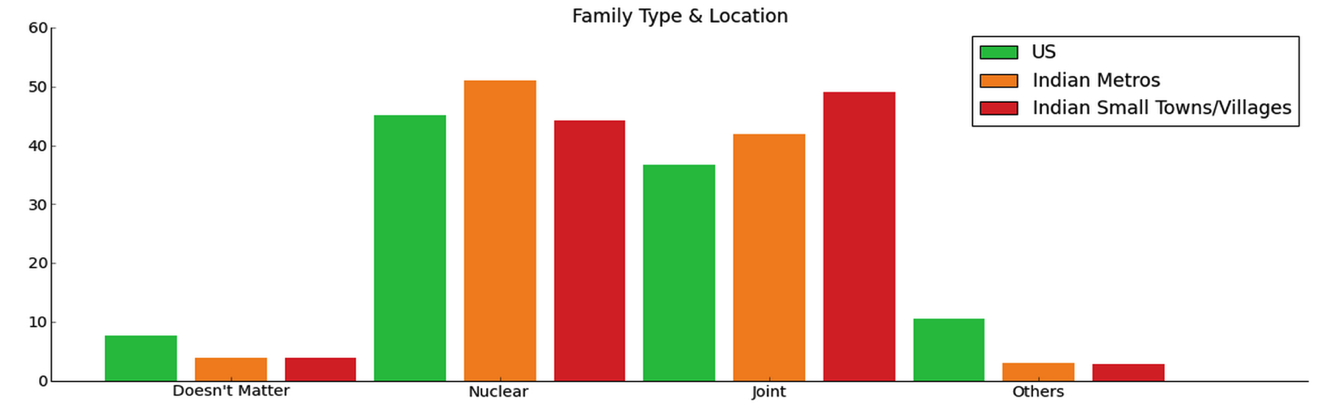
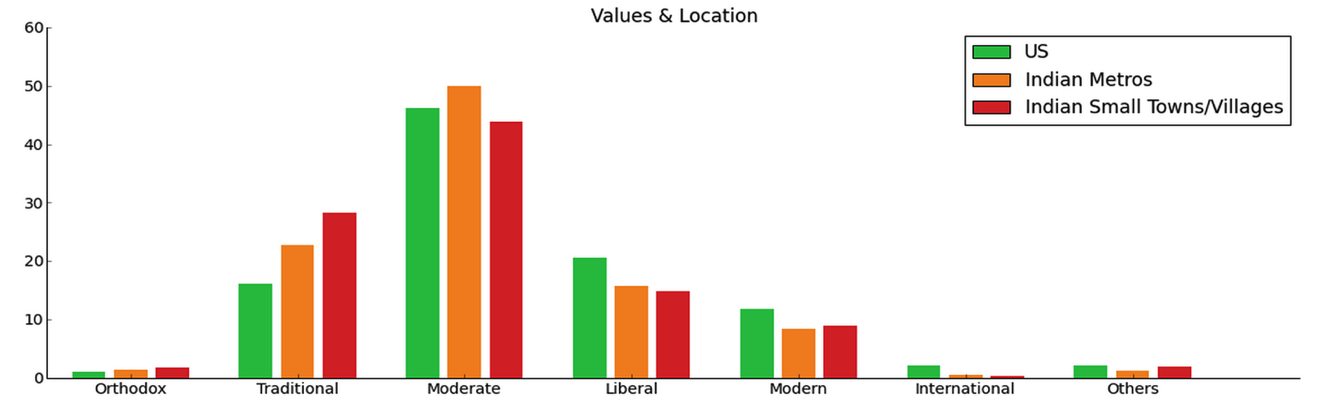




**Appendix C (economic class)**

****

**Appendix D (location)**

****

**References**

1. Seth, N., & Patnayakuni, R. (2009). Online Matrimonial Sites and the Transformation of Arranged Marriage in India. [↑](#endnote-ref-1)
2. Lakshmanasamy, T. (2013). How deep is Caste Discrimination and Social Exclusion? Methodologies for. *American Economic Review*, *84*, 23-28. [↑](#endnote-ref-2)
3. Jha, S., & Adelman, M. (2009). Looking for love in all the white places: A study of skin color preferences on Indian matrimonial and mate-seeking websites. Studies in South Asian Film & Media, 1(1), 65-83. [↑](#endnote-ref-3)
4. Dahlerup, D. (Ed.). (2013). *Women, quotas and politics*. Routledge.

   [↑](#endnote-ref-4)