Product Intern Project

Project Description

Every observation on the demographic data represents a connection between two people and the data continues to describe the gender, year of birth and the race of the connected persons. The aim is to analyze variables that shape the relationship between these connections. The impacting variables will help understand the social graph better.

Dataset

The dataset has 34037 observations with following variables:

AC_Sink (a unique ID representing a person)

AC_Source (a unique ID representing a person)

Sink_gender (the gender of AC_Sink)

Sink_YOB (the year of birth of AC_Sink)

Sink_race (the race of AC_Sink)

Source_gender (the gender of AC_Source)

Source_YOB (the year of birth of AC_Source)

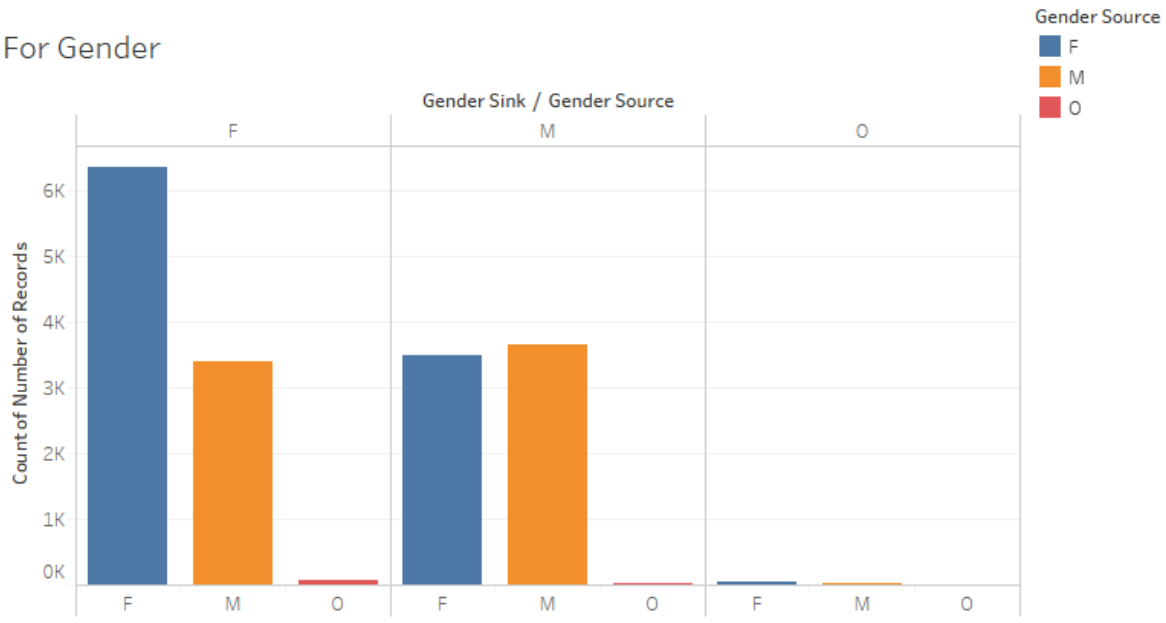
Source_race (the race of AC_Source)

Since each row represents relation between two individuals, one would assume the observations to be unidirectional eg: Ac_Sink knows Ac_Source. Since we have records that display the reverse relation eg: Ac_Sink knows Ac_Source and Ac_source knows Ac_sink ; we consider the observations to be bidirectional as suggested.

Analysis and Findings

Relationship between same Gender:

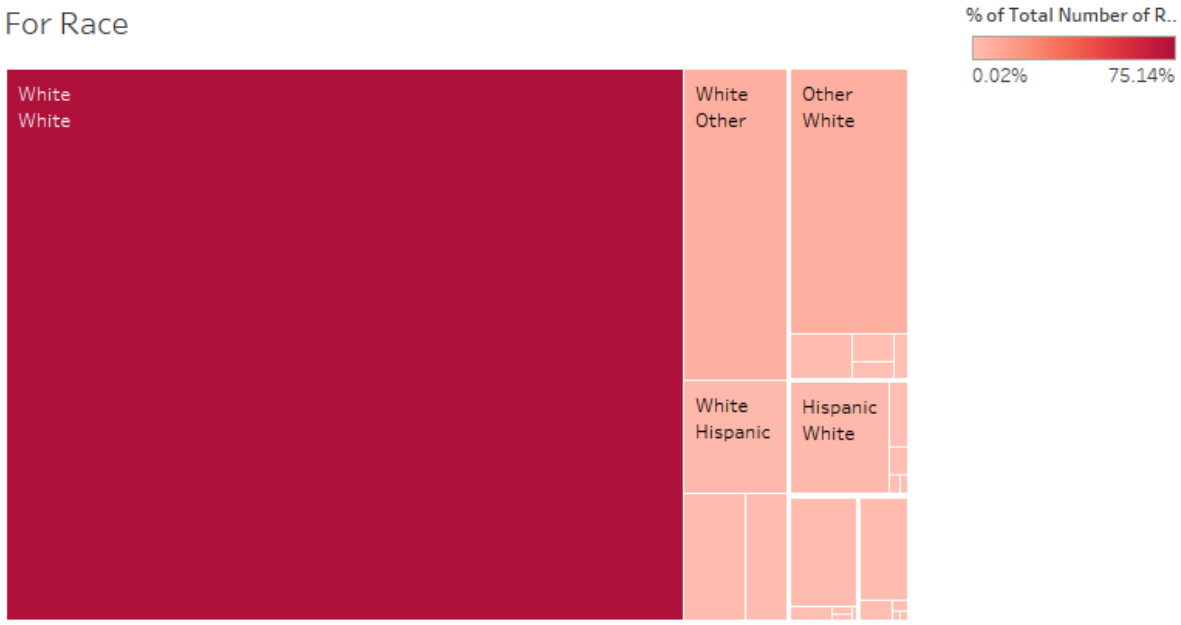
Count of gender_sink	Column Labels			
Row Labels	**F**	**M**	**O**	**Grand Total**
F	37.28%	20.41%	0.26%	57.95%
M	19.96%	21.42%	0.13%	41.51%
O	0.36%	0.17%	0.01%	0.54%
Grand Total	**57.60%**	**42.00%**	**0.40%**	**100.00%**



The numbers and the graph suggest that people from the same gender are related. 37 % of the data represents female to female relations and 21 % of the data represent male to male relations.

Relationship between same Race:

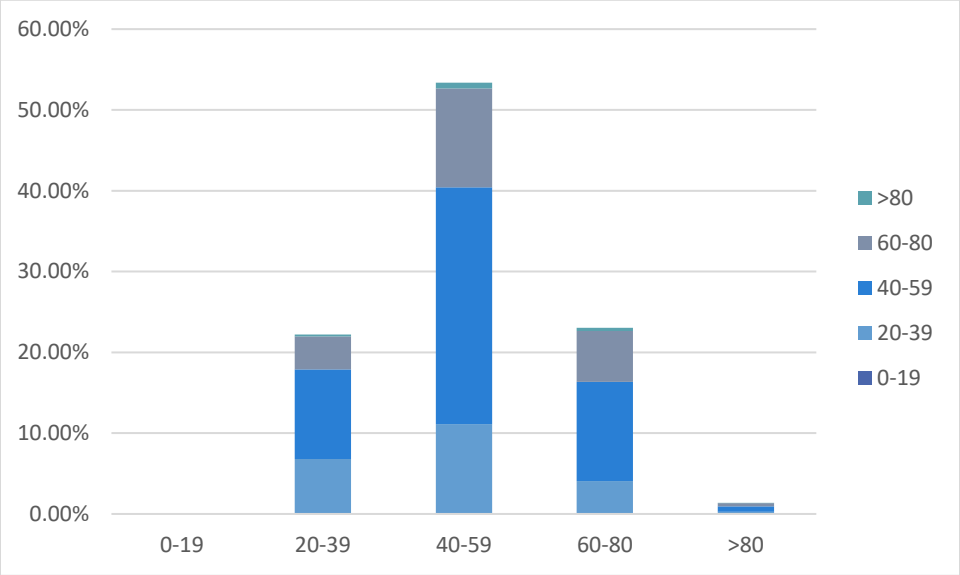
Count of sink_race	Column Labels					
Row Labels	**Asian**	**Black**	**Hispanic**	**Other**	**White**	**Grand Total**
Asian	0.02%	0.02%	0.04%	0.14%	1.07%	1.28%
Black	0.02%	0.03%	0.03%	0.13%	1.57%	1.77%
Hispanic	0.04%	0.03%	0.11%	0.24%	2.37%	2.79%
Other	0.14%	0.13%	0.24%	0.56%	6.47%	7.55%
White	1.07%	1.57%	2.37%	6.47%	75.14%	86.61%
Grand Total	**1.28%**	**1.77%**	**2.79%**	**7.54%**	**86.61%**	**100.00%**



The above results support the hypotheses of similar races to have more connections. Around 20148 same race relations are observed which makes up 75 % of the data.

Relationship between same Age:

Count of sink_age	Column Labels					
Row Labels	**0-19**	**20-39**	**40-59**	**60-80**	**>80**	**Grand Total**
0-19	0.00%	0.00%	0.01%	0.00%	0.00%	0.01%
20-39	0.00%	6.80%	11.09%	4.07%	0.26%	22.23%
40-59	0.01%	11.09%	29.32%	12.26%	0.70%	53.37%
60-80	0.00%	4.07%	12.26%	6.33%	0.37%	23.03%
>80	0.00%	0.26%	0.70%	0.37%	0.03%	1.36%
Grand Total	**0.01%**	**22.23%**	**53.37%**	**23.03%**	**1.36%**	**100.00%**



Out of the 5 groups, only one relation stands out. The group that represents individuals between 40-59 seem to be connected to individuals in the similar age range. For age ranges 20-39 and 60-80 around 6 % of the records show relation between same ages respectively. So around 42 % of the data represents connection between similar age groups.

Conclusion

While relations between individuals of same race and gender seems strong we cannot say the same about the relation between people of the same age group.