A2. Exploratory Data Analysis

Thong Bui

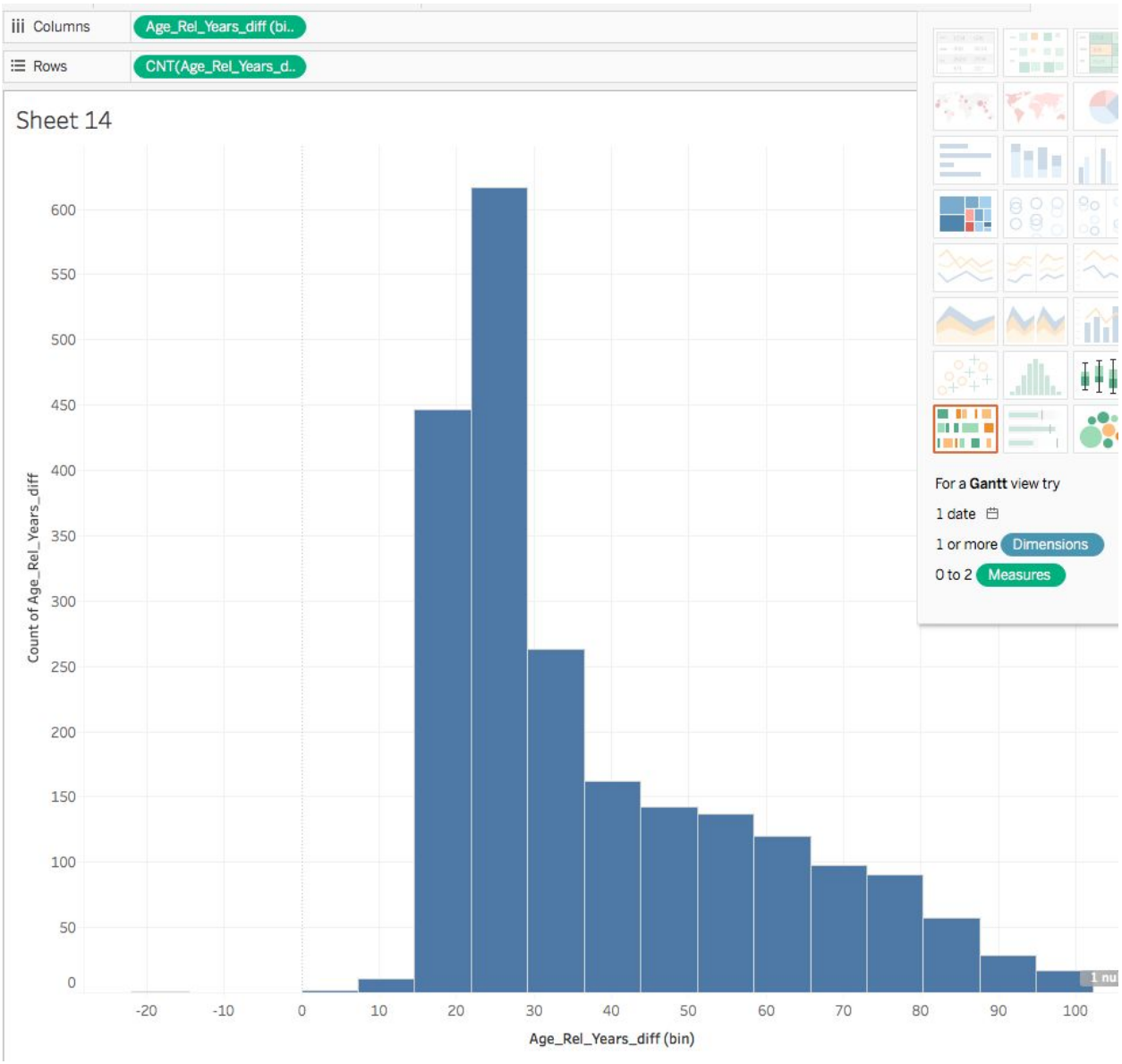
W209 - 3

In this assignment you am selecting to explore data from Online Dating and Relationships (Survey questions) using Tableau Desktop.

From inspecting the csv, I am noticing that the sample size is 2252 which is a good size to explore, with some interesting attributes such as age, sex, years in relationship, flirt on line, adults in household which will be used to explore with Tableau to attempt to support my three hypotheses. My hypotheses are simply built based on the stereotypes and common social generalizations about people in relationships. It will be interesting to see how well the data will support these hypotheses.

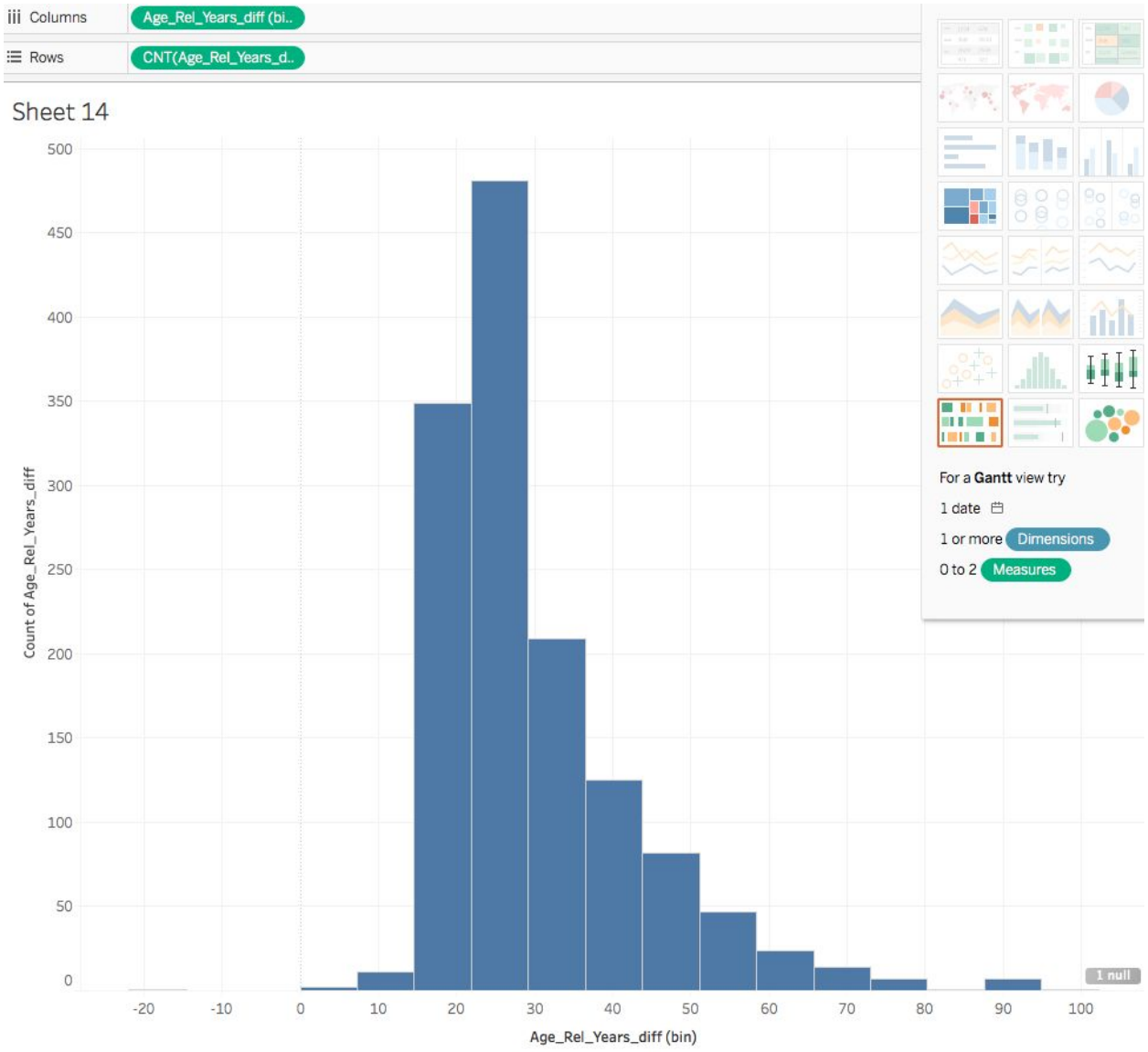
Hypothesis 1: Generally people start to get in relationship around 20 years old

Here I created a derived variable age_rel_years_diff = age - years_in_relationship from which I created the following histogram



What's informative about this view: This view shows a good summary of this data set with the difference starting at 20 years as predicted. The majority is from 20 to 40 years. There are some anomalies or outliers on the x axis at -20 which definitely is an error. Also, another anomaly is at 100 which implies that years in relationship = 0 when age is at its max 100

What could be improved about this view: because of the anomaly at 100, we should only include years_in_relationship greater 0 which is showing in this next histogram



What's informative about this view: After filtering the data with years_in_relationship > 0, the histogram now is showing the right skew is much less, with no data for x = 100

What is missing in this view: this histogram only shows summary of the data yet we don't have a detailed view of how the data look like for these two variables. In order to see more details of the data, I create the scatter plot below

What's informative about this view: In this scatter plot, it shows in great details of number of years of relationship vs age:

- There is large number of people are not in relationship (years_in_relationship = 0)
- The 2 outliers where years_in_relationship > 80 are the same errors that showed up in the previous histogram (< 0)
- The majority shows that the difference between age and years in relationship start at least 20 years

What could be improved about this view: it would be more interesting to see this difference for men vs women so I am adding Sex as another parameter on this graph to show that:

What's informative about this view: In this scatter plot, it still shows the same data as the previous plot but now it also shows that women general starts relationship earlier than men as the blue data (women) appears lower than the orange data (men)

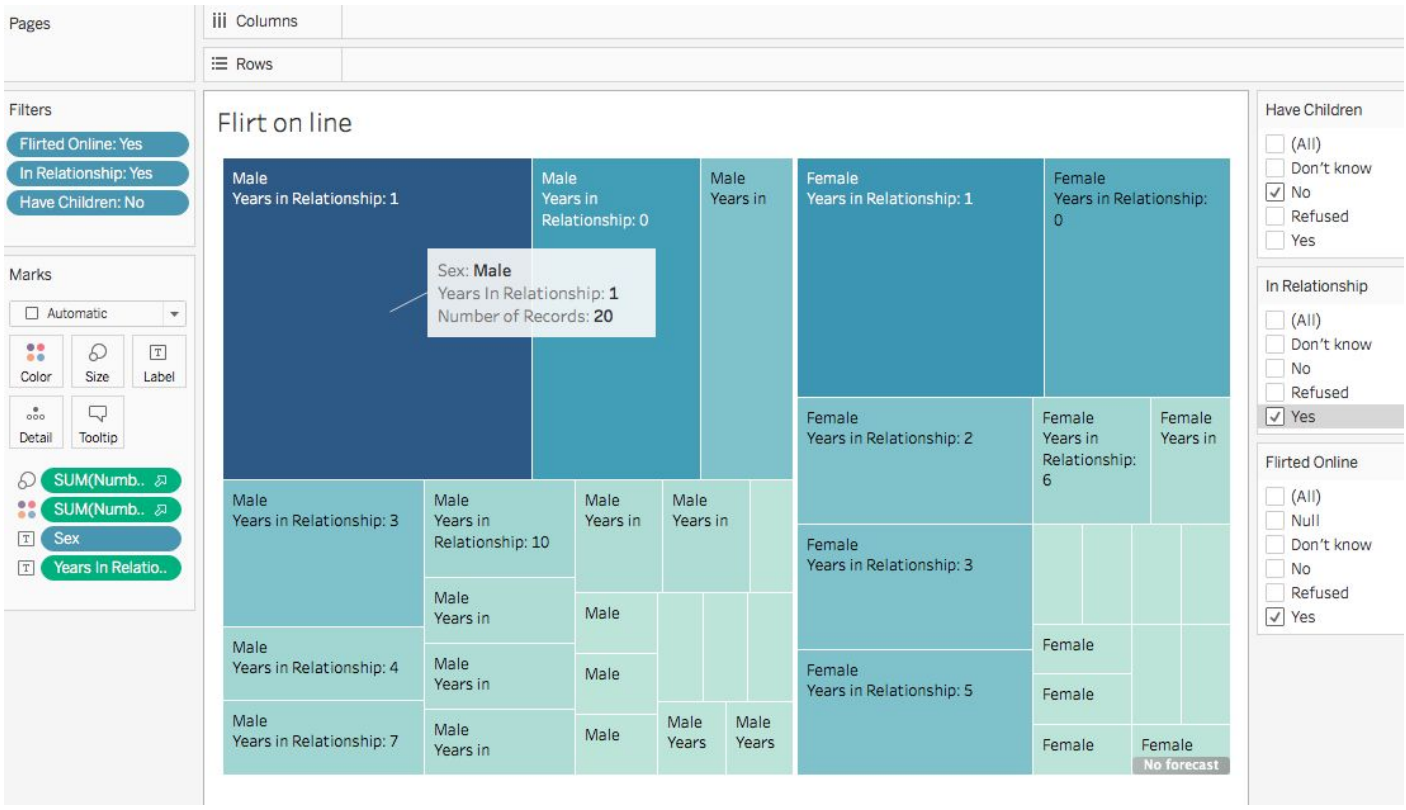
Now let's look at the same data partitioned by race

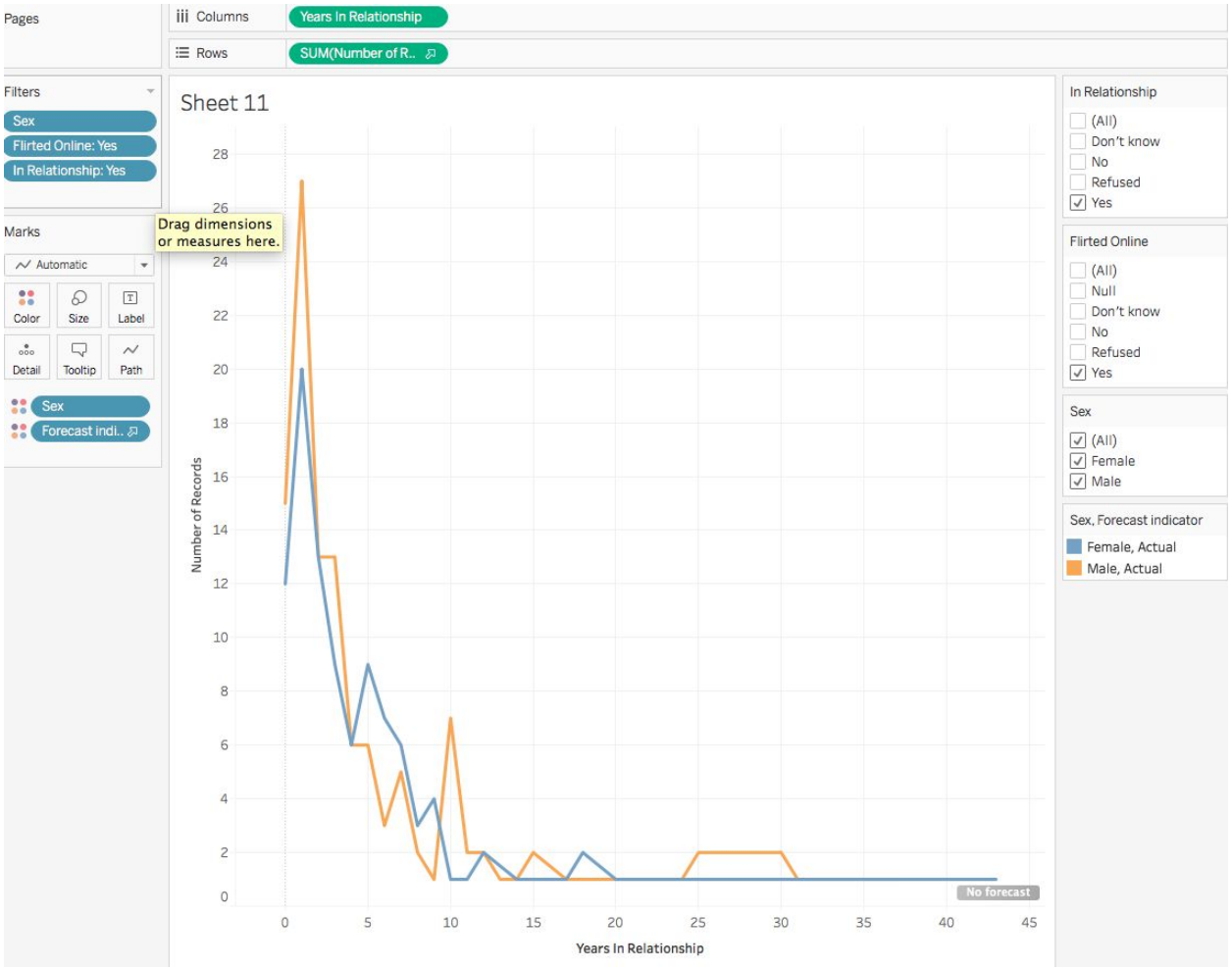
What's informative about this view: In this scatter plot, it shows that the majority of this sample are white (in turquoise). There much less data for black (in red) and asian (blue) and native american (green), etc.

What could be improved about this view: While there is significant data for white, there seems to be much less data for other races hence it's probably a good idea not to draw a generalized statement about this subject based based on race

Conclusion: As shown above, the data and visualizations do appear to support this hypothesis that people generally starts their relationship around 20 years old in the US. It also shows that women tend to start relationships a few years earlier than men

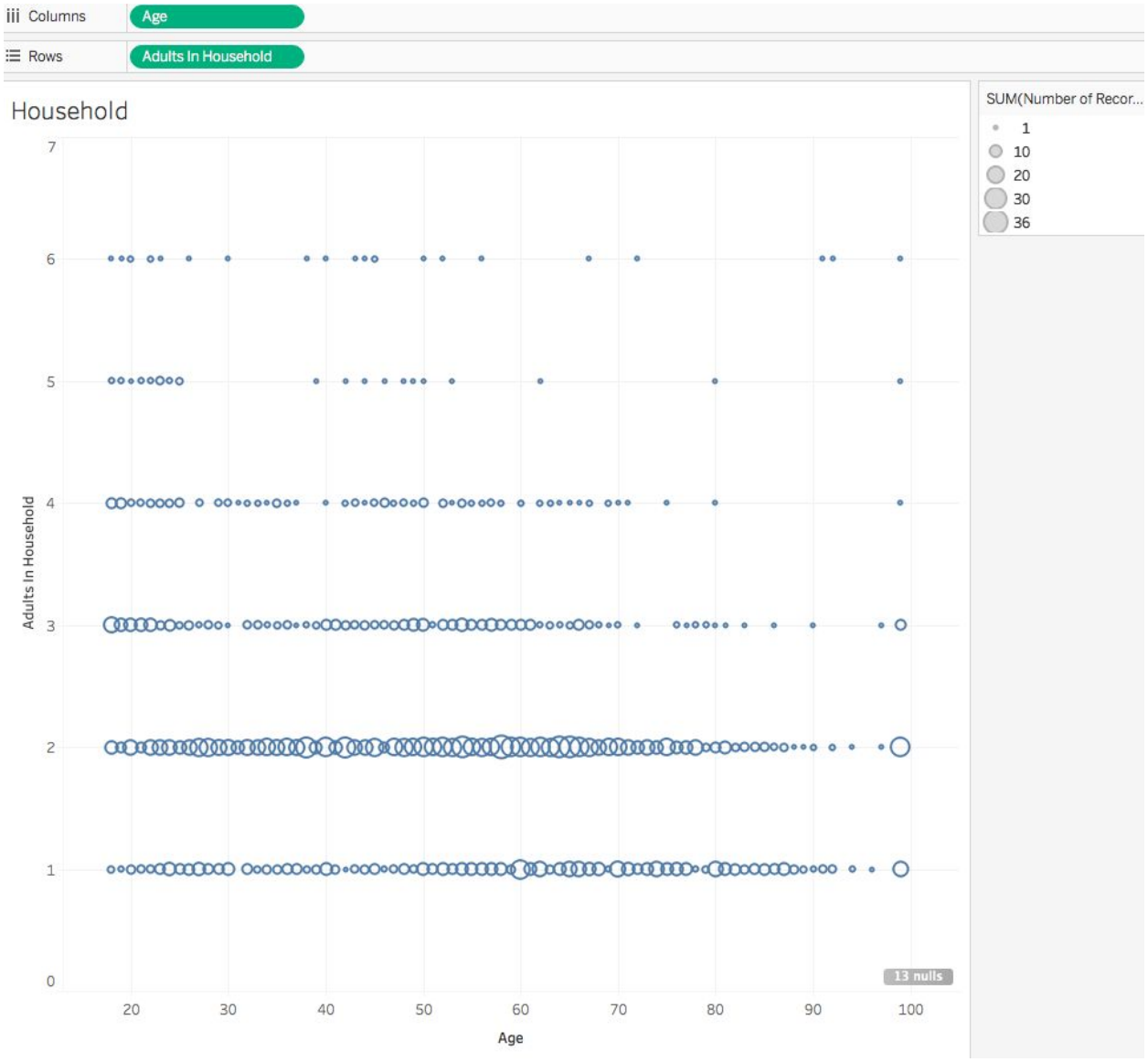
Hypothesis 2: Men in a relationship are more likely to flirt online than women in a relationship





This graph shows from this data set men in a relationship are more likely to flirt online than women, especially at the very early part of the relationship (0-5 years). The flirt-online trend drops significantly after that for both men and women, though interestingly women do flirt more online than men during 5 to 10 years during their relationship. After that, the flirt almost diminishes for both men and women

Hypothesis 3: The majority of the households have 2 adults


Columns
Age
Rows
Adults In Household
Household
SUM(Number of Recor...
1
10
20
30
36
Adults In Household
0
1
2
3
4
5
6
7
13 nulls
20
30
40
50
60
70
80
90
100
Age

Household