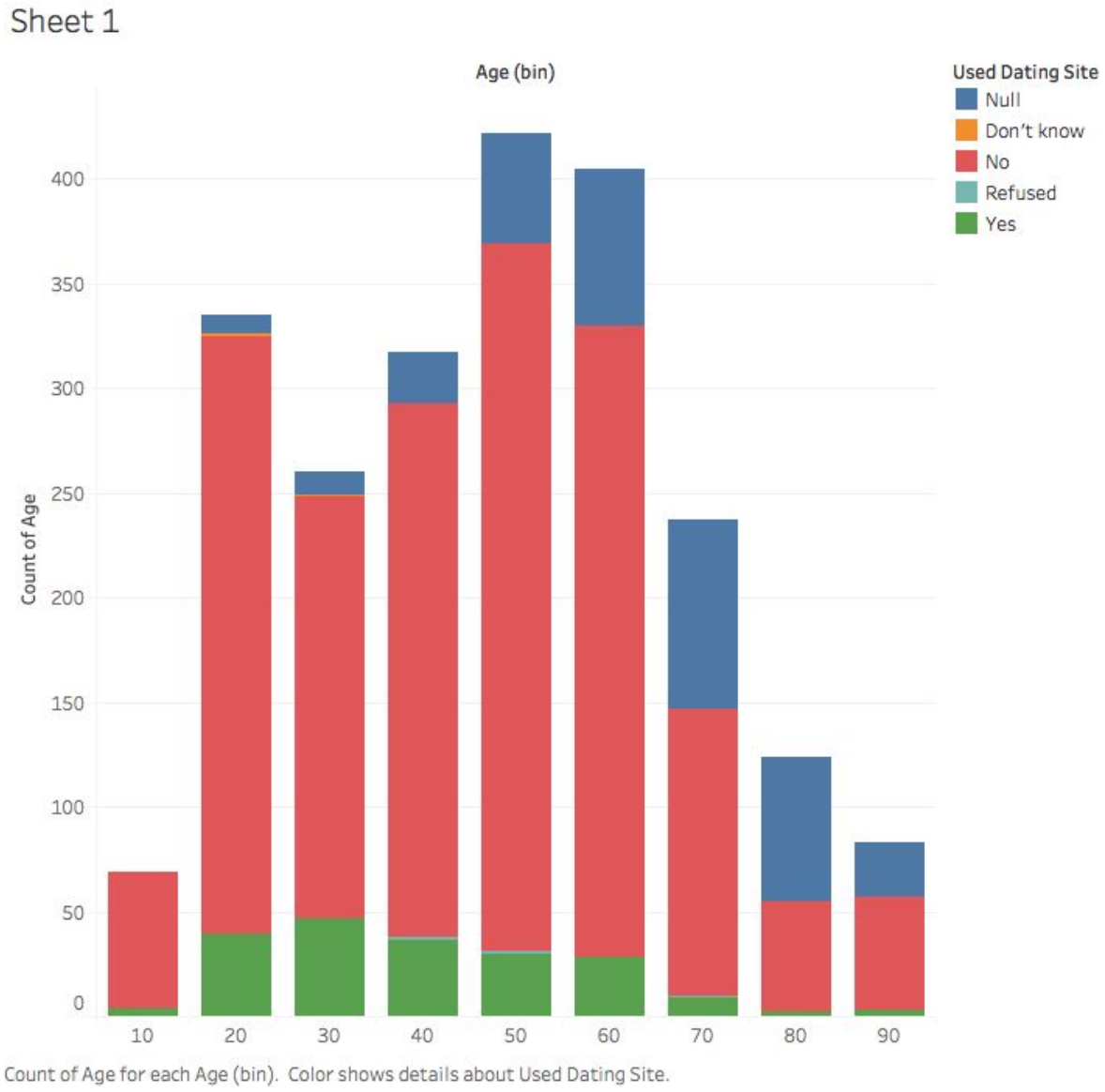
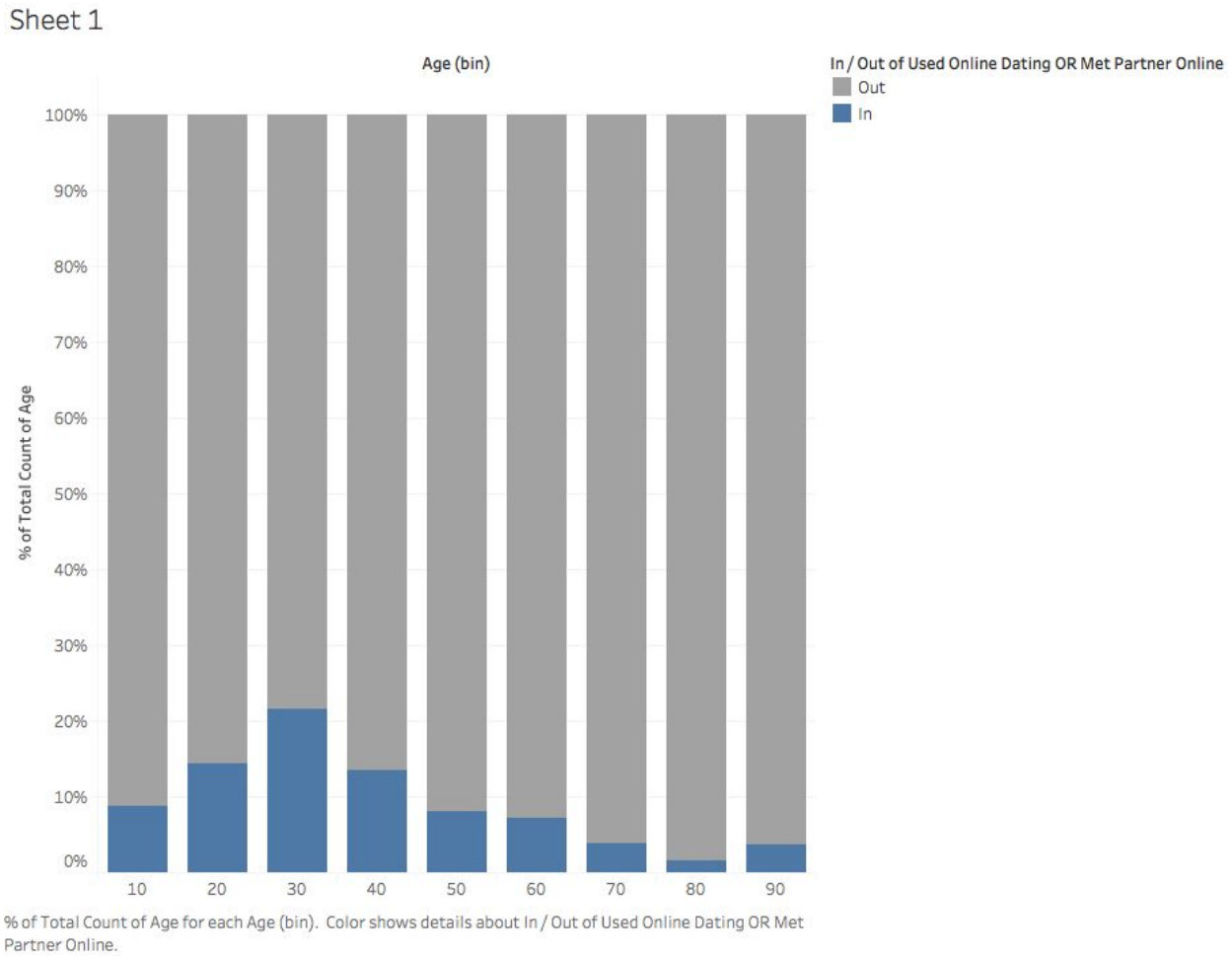
A2 : Exploratory Data Analysis with Tableau
Daniel Olmstead
W209 Section 1 - Andy Reagan

Hypothesis 1 (demographics): Those who met their partner online or have used dating sites will tend to be younger, urban, and from the West or Northeast.



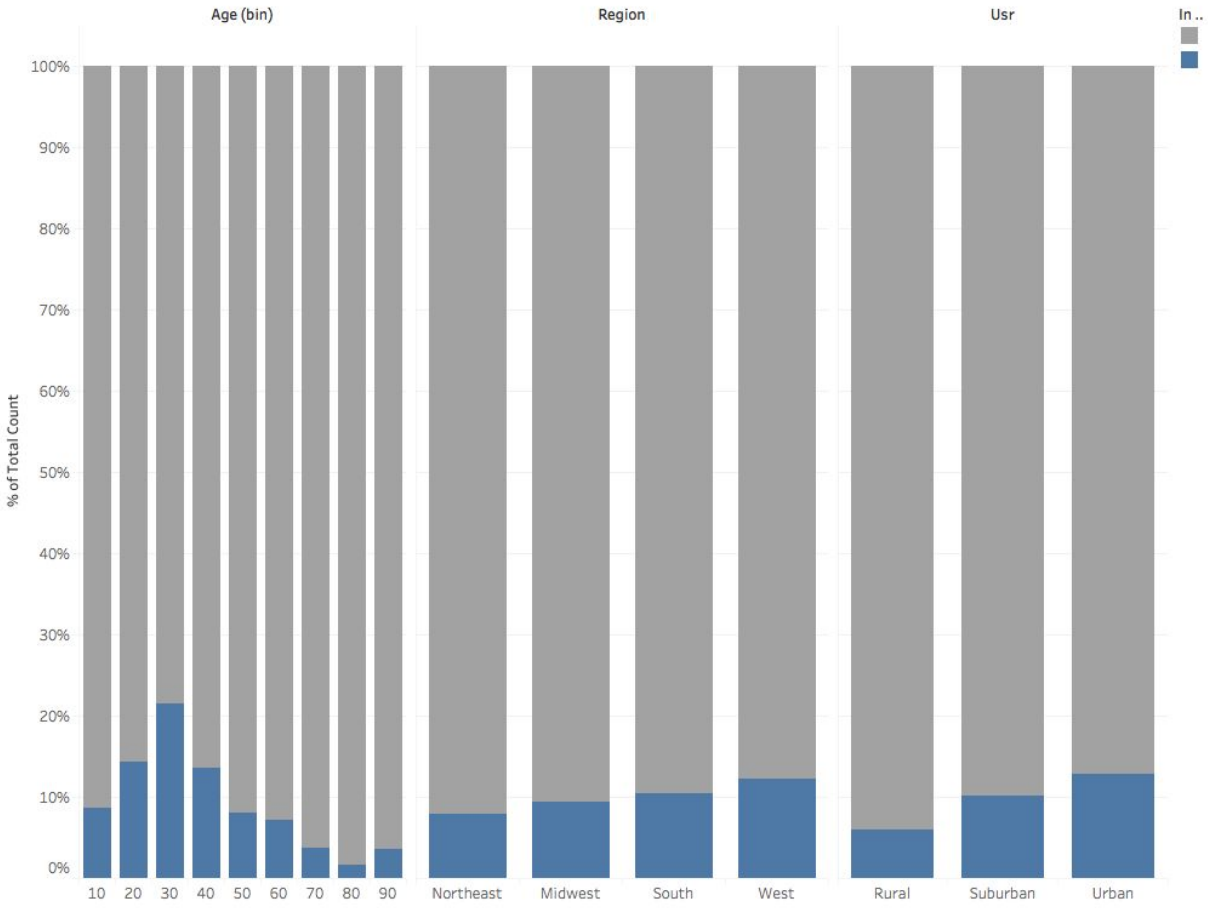
What's informative about this view: This is a distribution of ages for the entire dataset, and shows us a few things: the vast majority of respondents have not used dating sites, but for those who have, the average age looks to be in the 30s, compared to the overall average of maybe around 50. So we're on the right track.

What could be improved about this view: It doesn't show us anything about people who met their current partner online (we want one OR the other), nor does it say anything about whether the respondent is urban or coastal. It also gives all the response categories, when we're really only interested in the people who said "Yes."



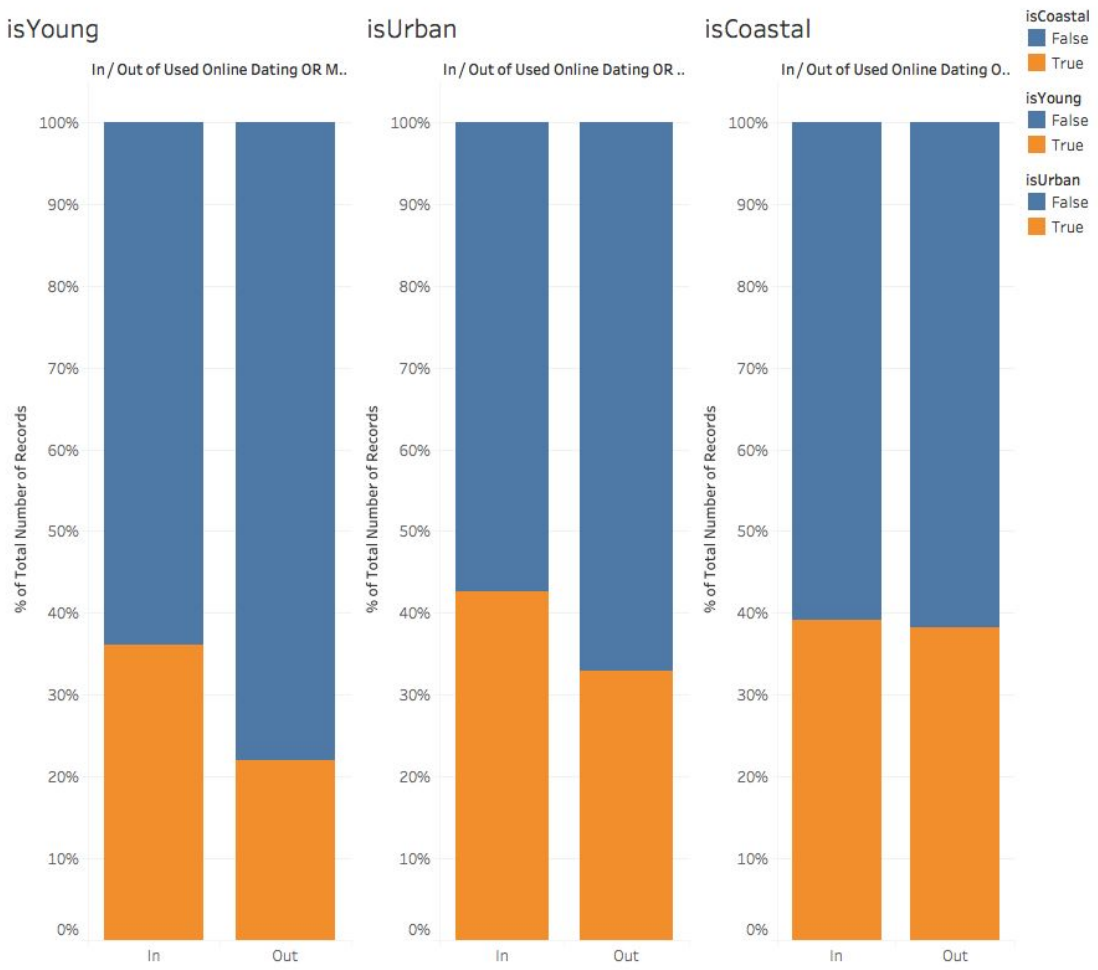
What's informative about this view: This view combines respondents who either used online dating OR met their current partner online, and shows their proportion across age groups in true binary, rather an also including all response categories. This definitely shows that people who look for/find partners 1) are a much smaller group and 2) tend to be younger than the overall population. The slight bump up in proportion for people over 90 is interesting, but is mostly just a function of proportions at small number sizes.

What could be improved about this view: This is good for the age dimension, but doesn't tell us anything about the other demographic categories. It also doesn't give us a sense of the size of the correlation.



What's informative about this view: Functionally the same as the previous graph, but this time incorporates regionality and urbanity. While the hypothesis of urban usage looks good, and there is higher usage in the West, the Northeast shows the lowest adoption of all the regions.

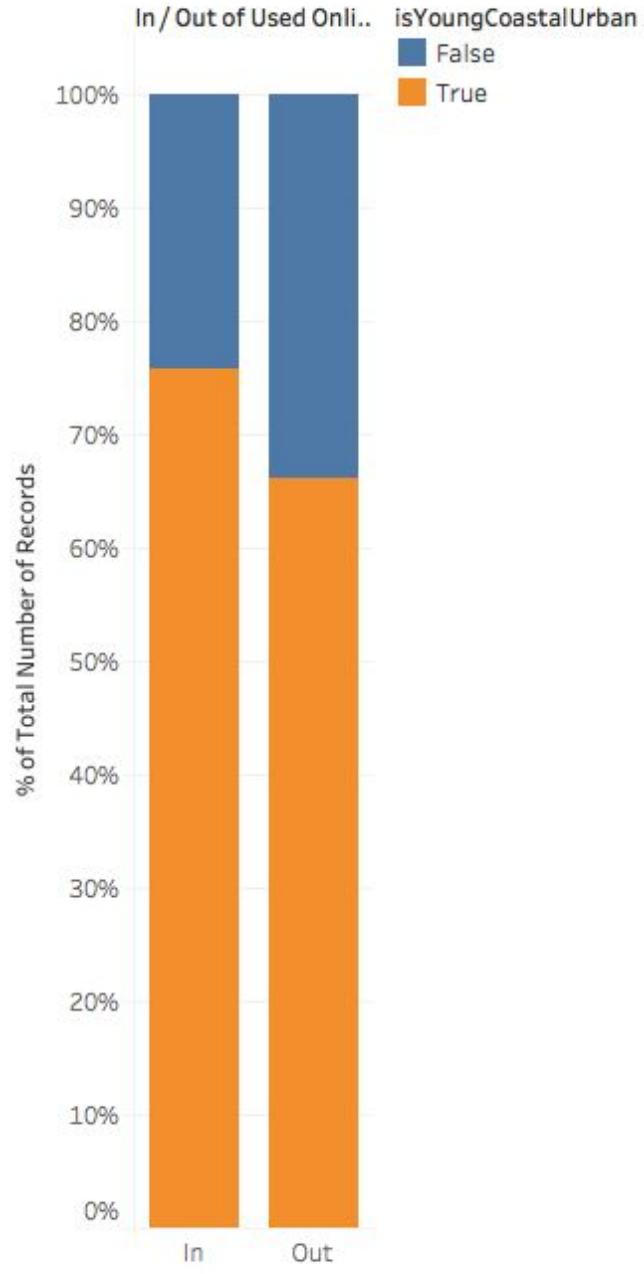
What could be improved about this view: This is just three isolated views side-by-side. It gives us a sense that the hypothesis is true from eyeballing it, but really doesn't give us an idea of the intersection of the three categories. Also, the view overweights the size discrepancy between the populations, when the hypothesis is really just looking to compare their demographic makeup irrespective of size.



What's informative about this view: This is the clearest answer to the hypothesis, broken out by the separate categories. By breaking everything into boolean variables, we can compare the populations side by side and see the probability of a respondent being either young (under 35), urban or coastal (living in West or Northeast regions).

What could be improved about this view: This answers the hypothesis, but could be simplified further to combine all three categories.

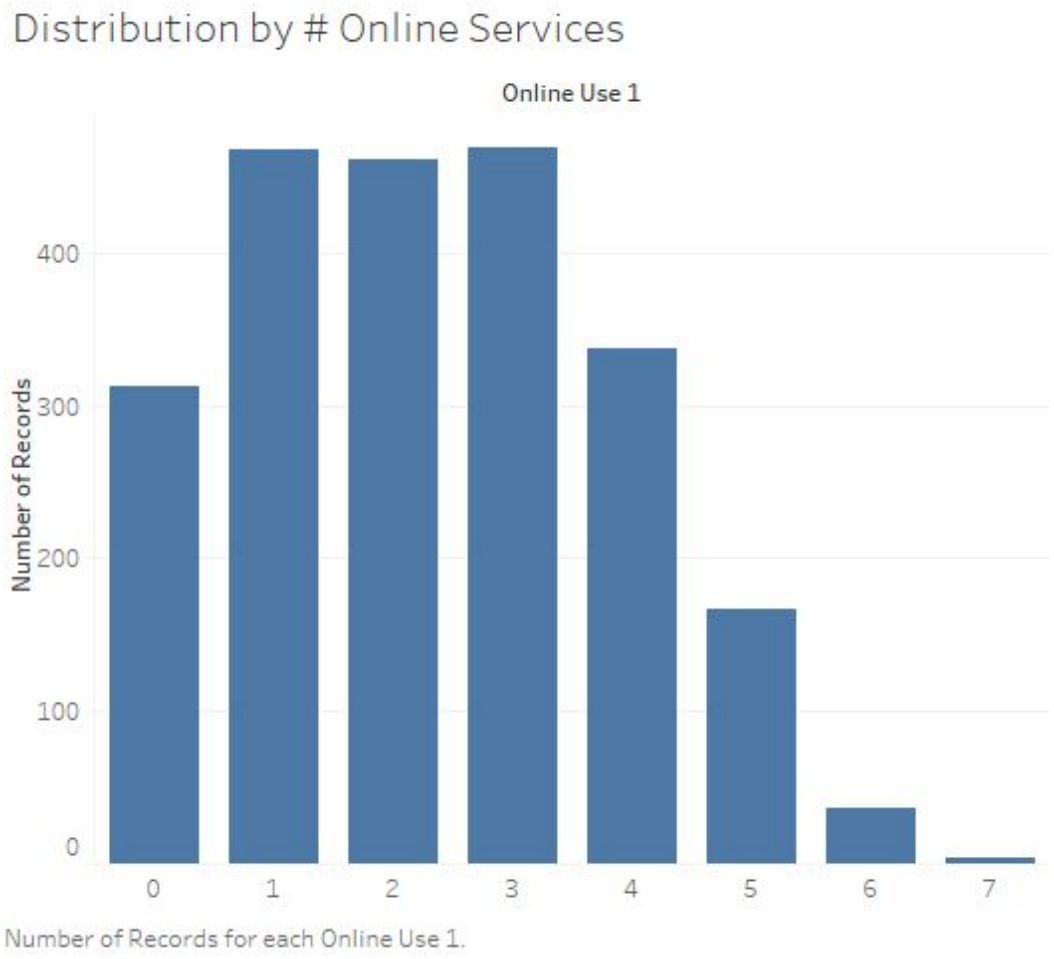
Sheet 7



% of Total Number of Records for each In / Out of Used Online Dating OR Met Partner Online. Color shows details about isYoungCoastalUrban. The data is filtered on isUrban, which keeps False and True.

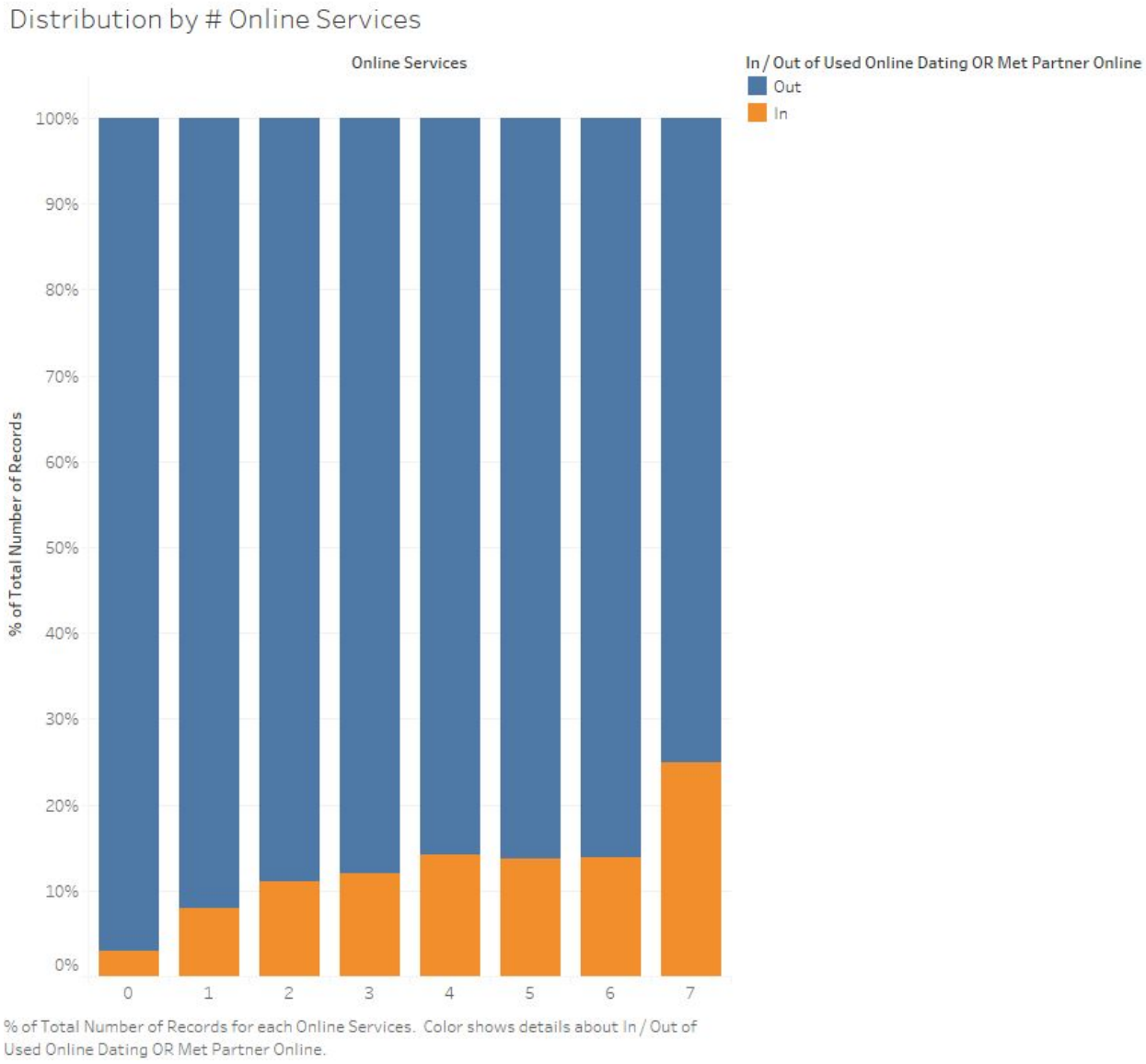
We can see here that people who have used dating site or found their partner online are ~10% more likely to be young, urban or live in a coastal region than those who have not.

Hypothesis 2 (usage): The more online services a person uses, the more likely they are to use dating sites.



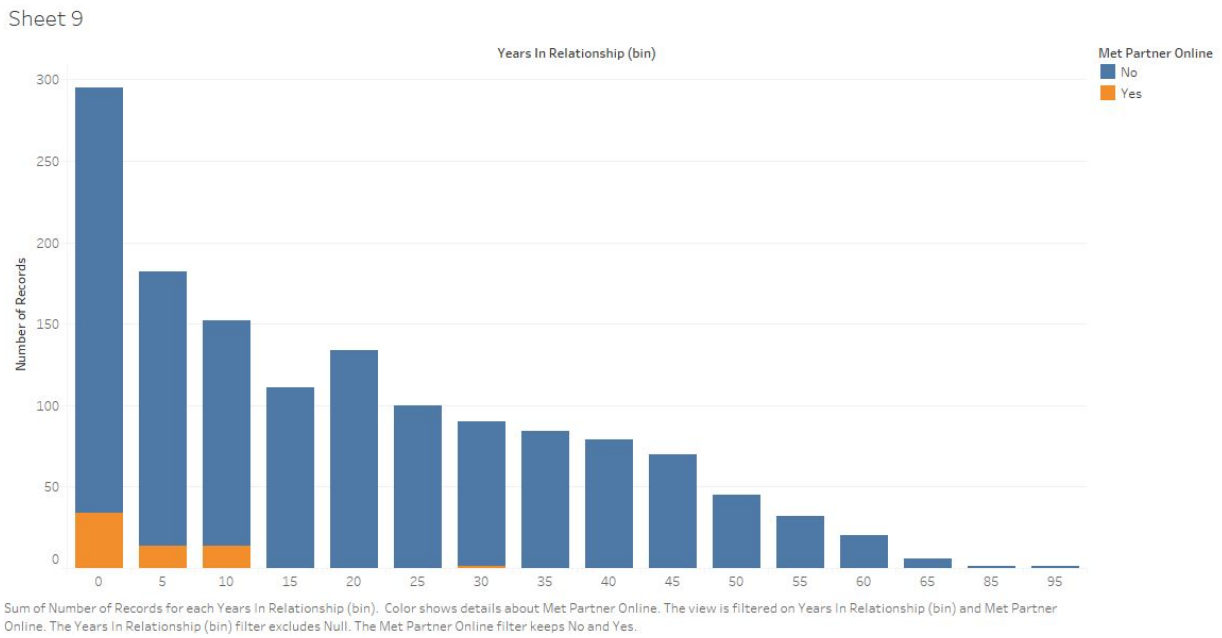
What's informative about this view: Here we see a breakdown of the respondents according to the number of online services/technologies that they use. A solid proportion of the study group uses 1-3 services, but it drops off pretty quickly after that. Only a very small number admits to using everything.

What could be improved about this view: It doesn't tell us anything about the relationship between tech/adoption and online dating.



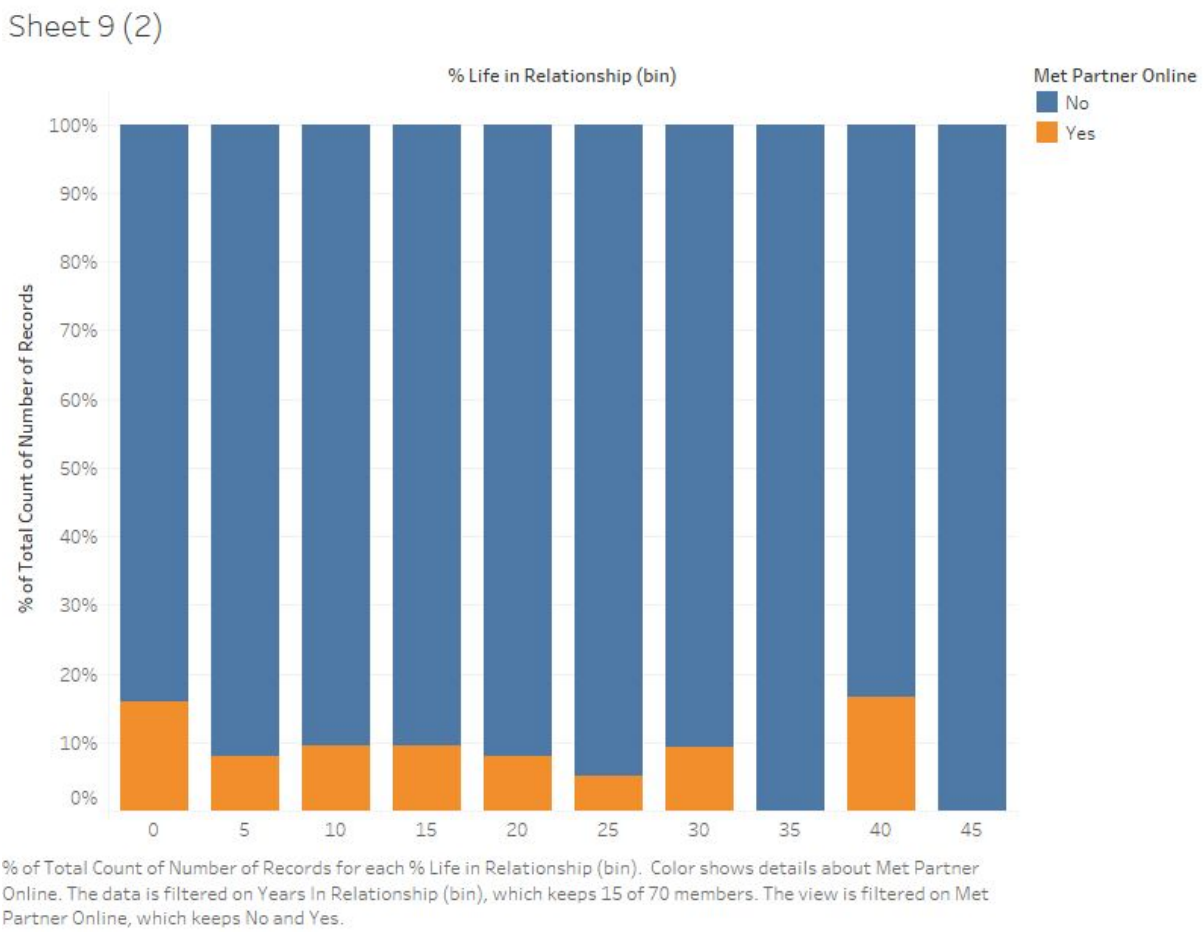
What's informative about this view: Once we normalize and include the proportion of people who use online dating, we see that by and large the hypothesis is correct: the more online/technology services a person uses, the more likely they are to also use online dating.

Hypothesis 3 (longevity): People who meet online tend to stay together longer than those who don't, normalized both by maximum relationship time (to account for advent of online dating) and by percentage of lifespan (to correct for age discrepancy).



What's informative about this view: A histogram of relationship duration shows us the largest bulk of relationships are under five years, and then there's a long tail. As expected, the proportion of people who met their partner online is small, and only for durations < 15 years (except one couple that met online in the 80s?).

What could be improved about this view: Without adjusting for the timescale of online dating (let's say 15 years) and the relative age of the participants, this doesn't tell us whether online relationships have any correlation with longer relationship duration.

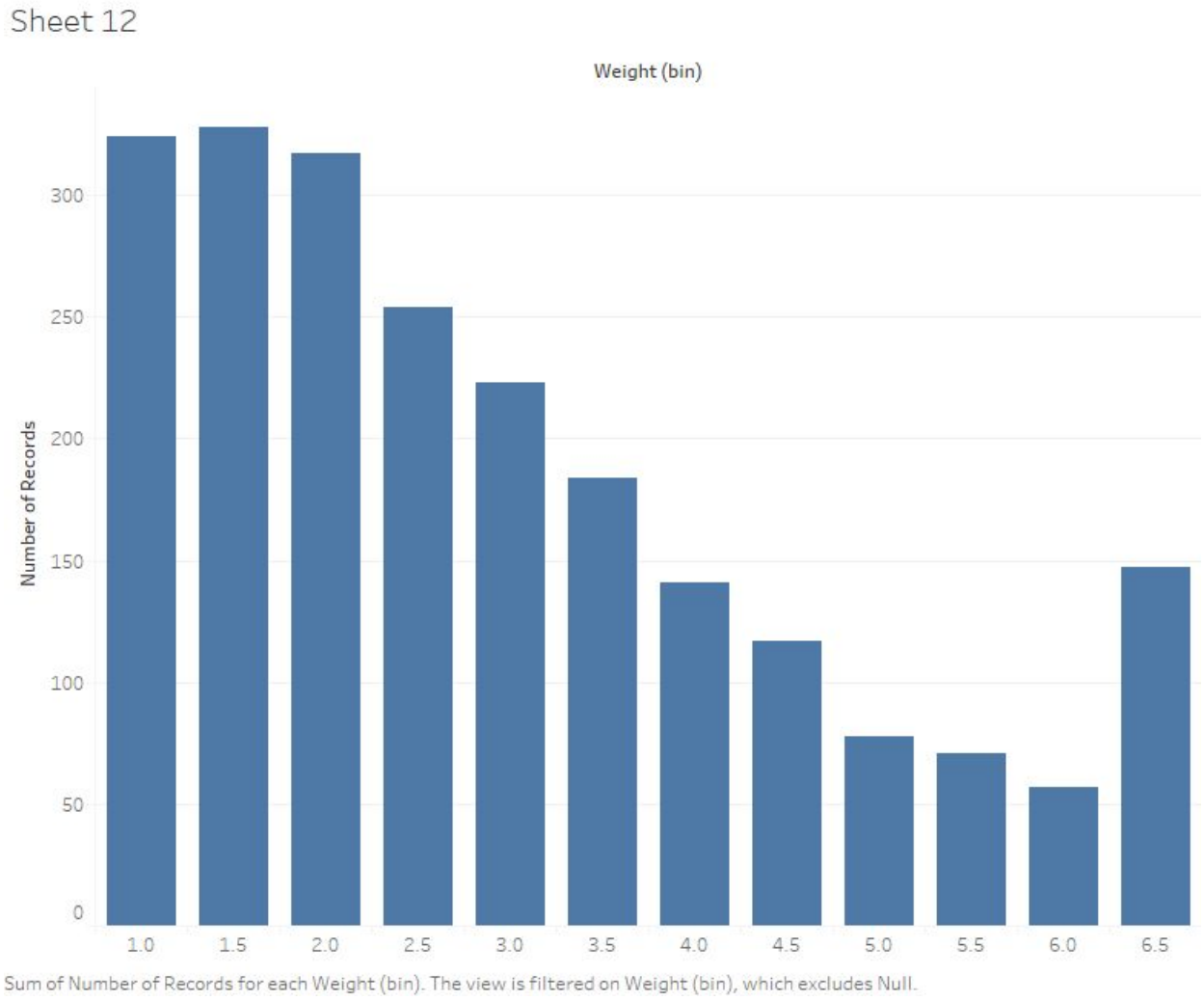


What's informative about this view: Once we make the necessary adjustments, there doesn't appear to be any trend that shows that people who meet online have spent a longer portion of their lives in that relationship than those who didn't. If anything, the opposite looks to be more likely, but at that end of the chart there is probably insufficient data (it would require people who adopted online dating early in both its lifespan and theirs).

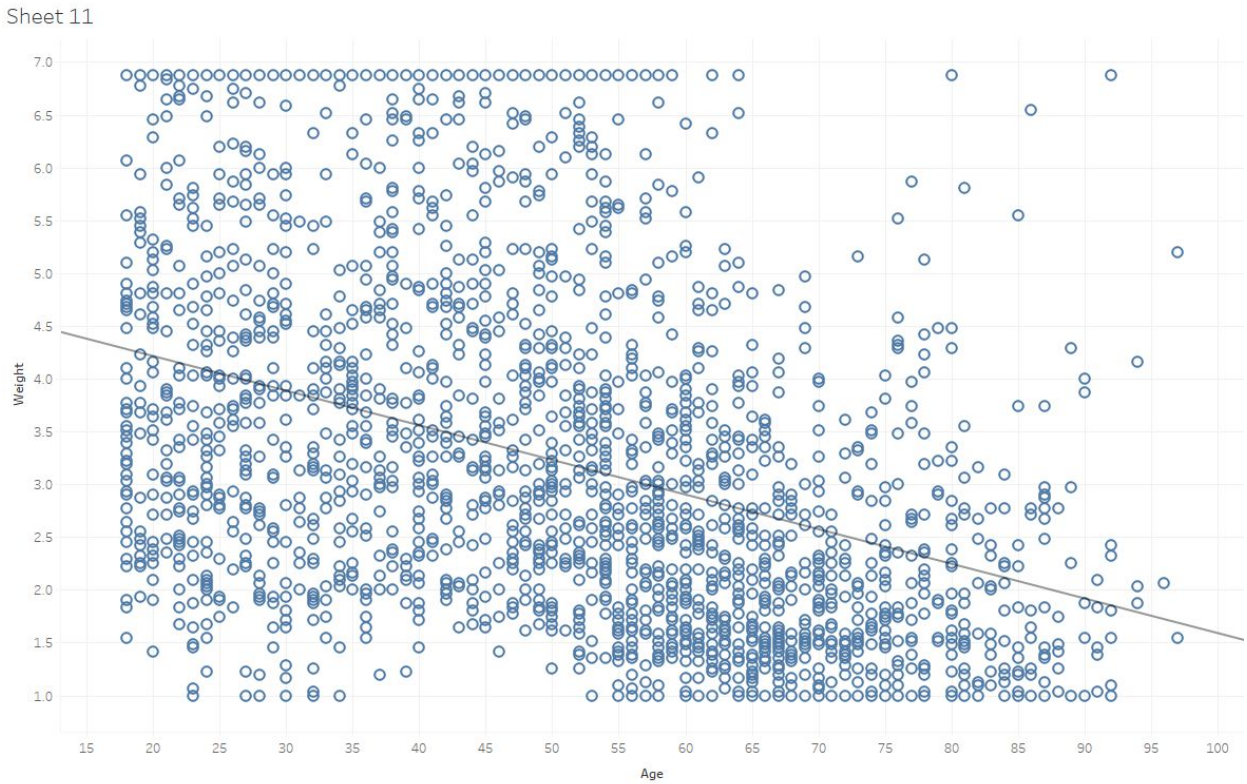
Just Because I'm Tired of Bar Charts

The dataset includes a "Weight" column, which seems straightforward, except the values don't correspond with any sensible weight. The survey itself doesn't mention the word "Weight." What is this?

A quick look at the data itself shows values between approximately 1 and 7, which might be an ordinal index like 'Income' or ' Quality of Life," except the values appear continuous rather than discrete.



A histogram shows us what looks like half a normal distribution, with mean slightly over 1 and then a max a little over 6 that collects all of the tail > 6. Weight distribution should be at least normal-ish over a population this size. Maybe this doesn't even refer to what a person weighs?



Age vs. Weight. The view is filtered on Age, which ranges from 18 to 98.

A scatterplot shows a clear negative correlation with Age, with almost all the upper-bound participants coming under age 60 and almost all the lower-bounds coming after age 55.

Maybe “weight” doesn’t refer to a person’s physical weight, but rather their statistical weight in some equation? And for some reason younger people tend to get assigned a heavier weight?