A2. Exploratory Data Analysis

Matthew Prout

W209 - Section 1

Introduction

The purpose of this assignment is to research three hypotheses with a data set, using visualizations to find evidence for or against each hypothesis.

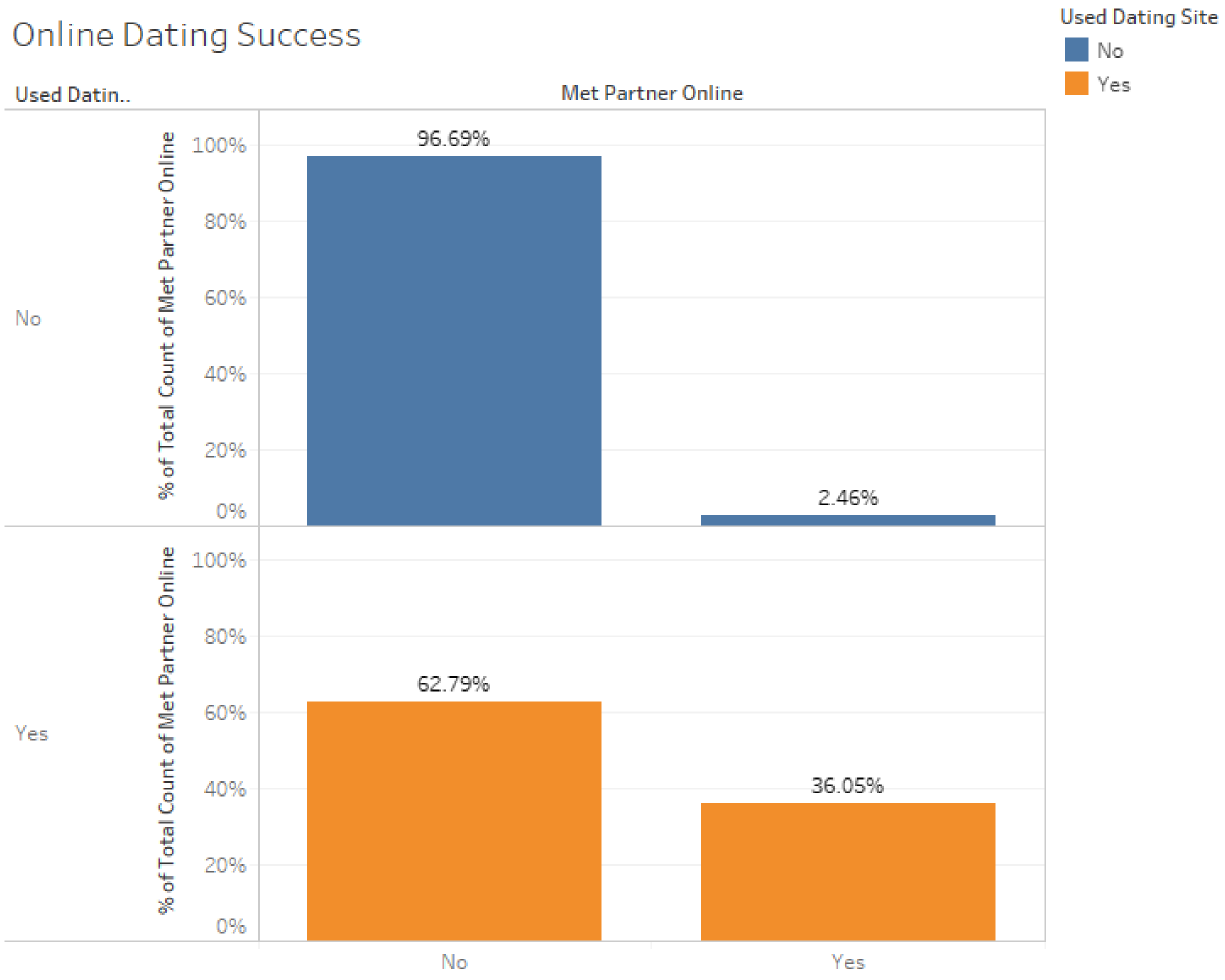
For this assignment, I chose the data set that comes from the Online Dating & Relationships survey from the Pew Research Center. This study was a national survey that was conducted in 2013, and the purpose was to learn about people's opinions of online dating. The data was a .csv file with observations from 2252 adults.

The data set had missing data (NULLs) for questions that were not answered by the survey respondents. Those values were left alone (not removed or imputed), as it was observed that data correlated with the missing data often had different distributions.

During the course of this assignment, it was determined that additional fields needed to be added to the dataset. These additional fields were added using R (refer to Appendix A), and I will note when the new fields were used.

Exploratory Data Analysis

Hypothesis 1: Online daters have limited success (less than 50%) at finding a partner.



What's informative about this view:

One of the things that I am interested in knowing about online dating is whether it actually “works” for people. To try to directly answer this question, I created a bar chart showing the proportion of those who met their partner online for those who use online dating vs. those who do not use online dating. This was done by aggregating the ‘Met Partner Online’ field and plotting it with the ‘Used Dating Site’ field.

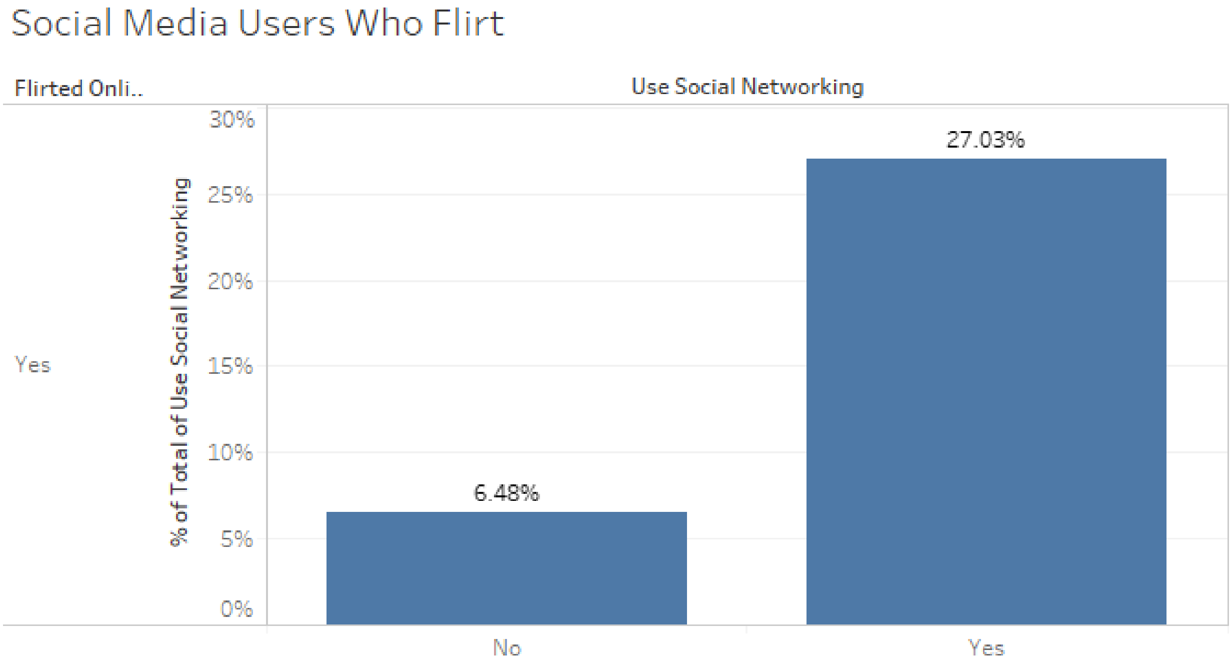
36% of those who used online dating met their partner online. Because this is below the stated hypothesis (50%), the hypothesis cannot be rejected.

One thing that is interesting to note is that according to the respondents, 2.5% met their partners online, but they did not use online dating. This raises an interesting question of how else they met their partner online. One possible way that people meet online is through social networking.

What could be improved about this view:

Some people might find the top row confusing, where respondents answer “No” to using online dating, and this is compared to whether the user found their partner online. Actually it turns out that people can find their partner through other online platforms such as social networking.

Another improvement to this graph would be to show the bar charts side by side.



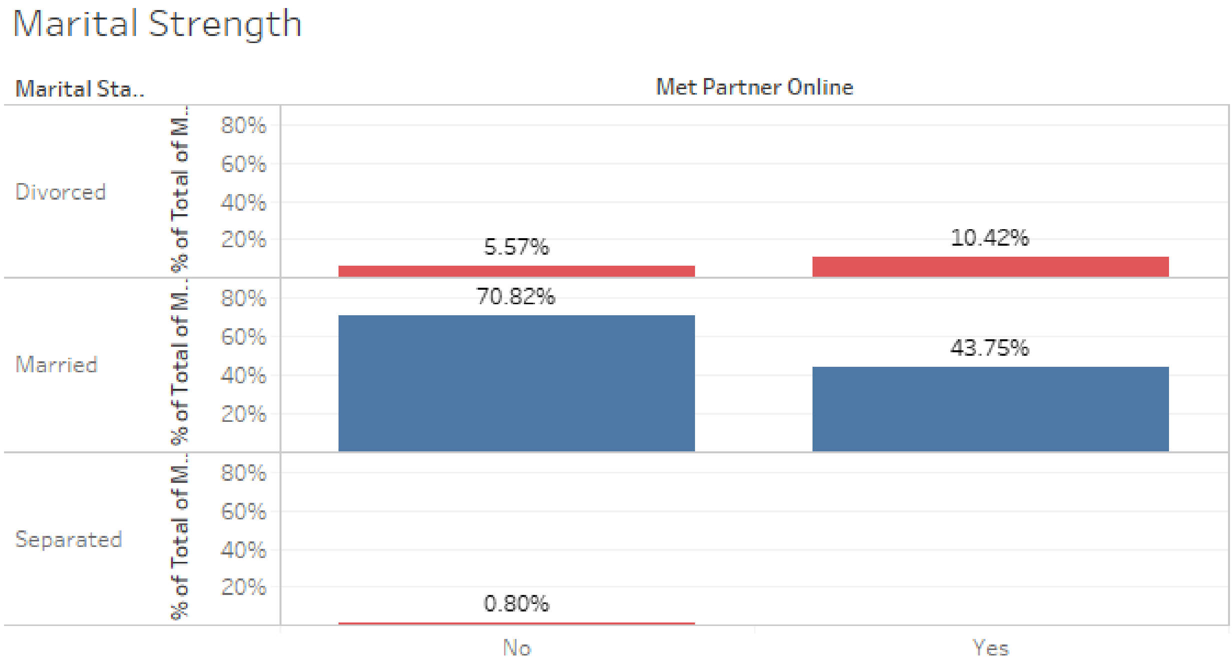
What's informative about this view:

To explore the feasibility of people meeting their partners on social media, I created a bar chart comparing those who used social networking with those who flirt online. This chart shows that 27% of those who use social networking flirt online, so this offers evidence that people could meet their partners through social networking.

What could be improved about this view:

The 6.5% of the users in the ‘No’ column are those who do not use social networks, but flirt online. This refers to some other platform where these users flirt online, which is not apparent.

In order to create the proportions, I had to create a chart comparing ‘Use Social Networking’ and ‘Flirted Online’, and hide the row where ‘Flirted Online’ is ‘No’. However, this leaves the ‘Yes’ in the margin under the heading ‘Flirted Online’, so this is extra text that does not help the view.



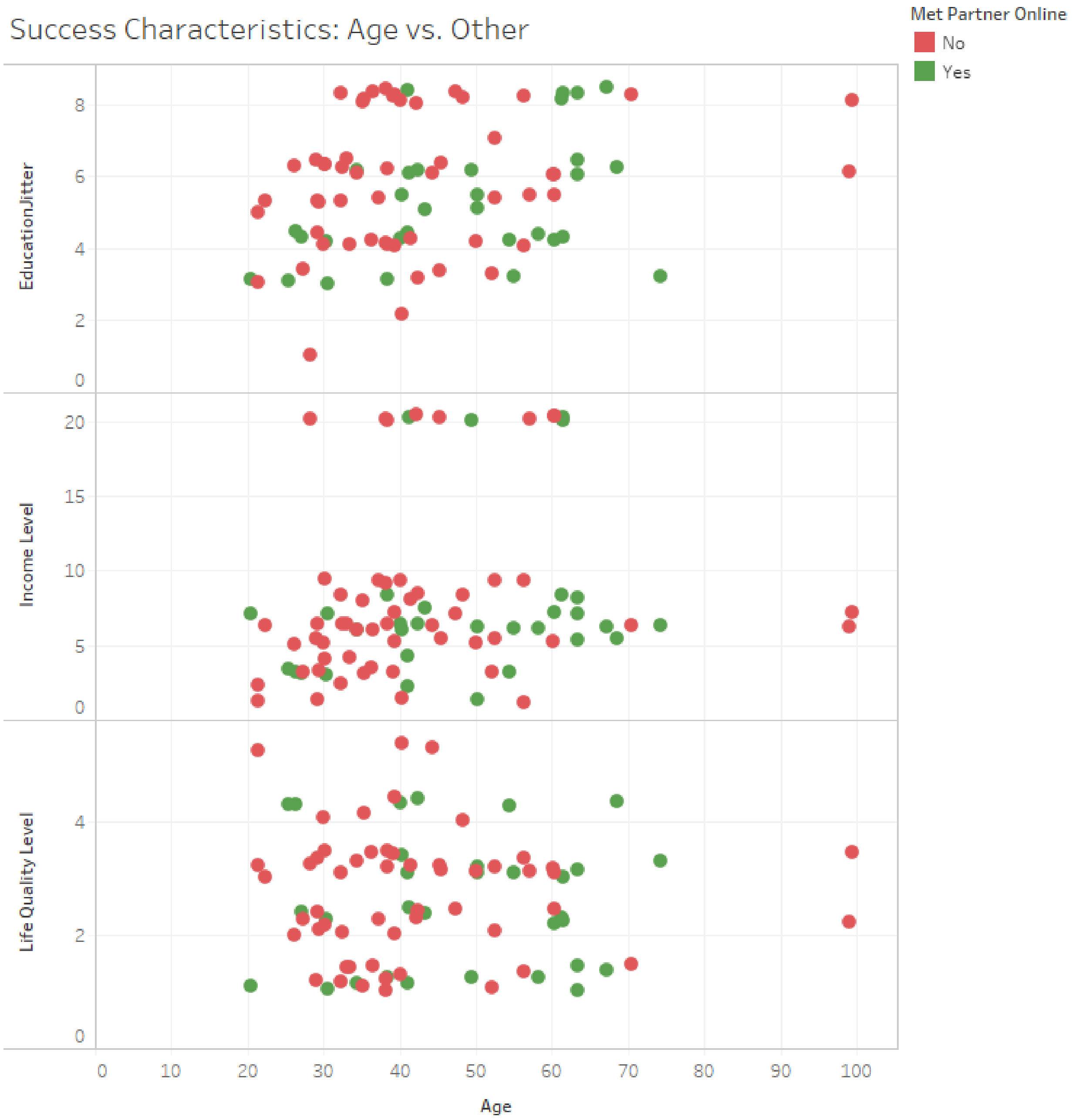
What's informative about this view:

A related question that I have about online dating is whether these kind of relationships end up enduring. To measure the strength of these relationships, I created a visualization of marital status with whether the respondent met their partner online.

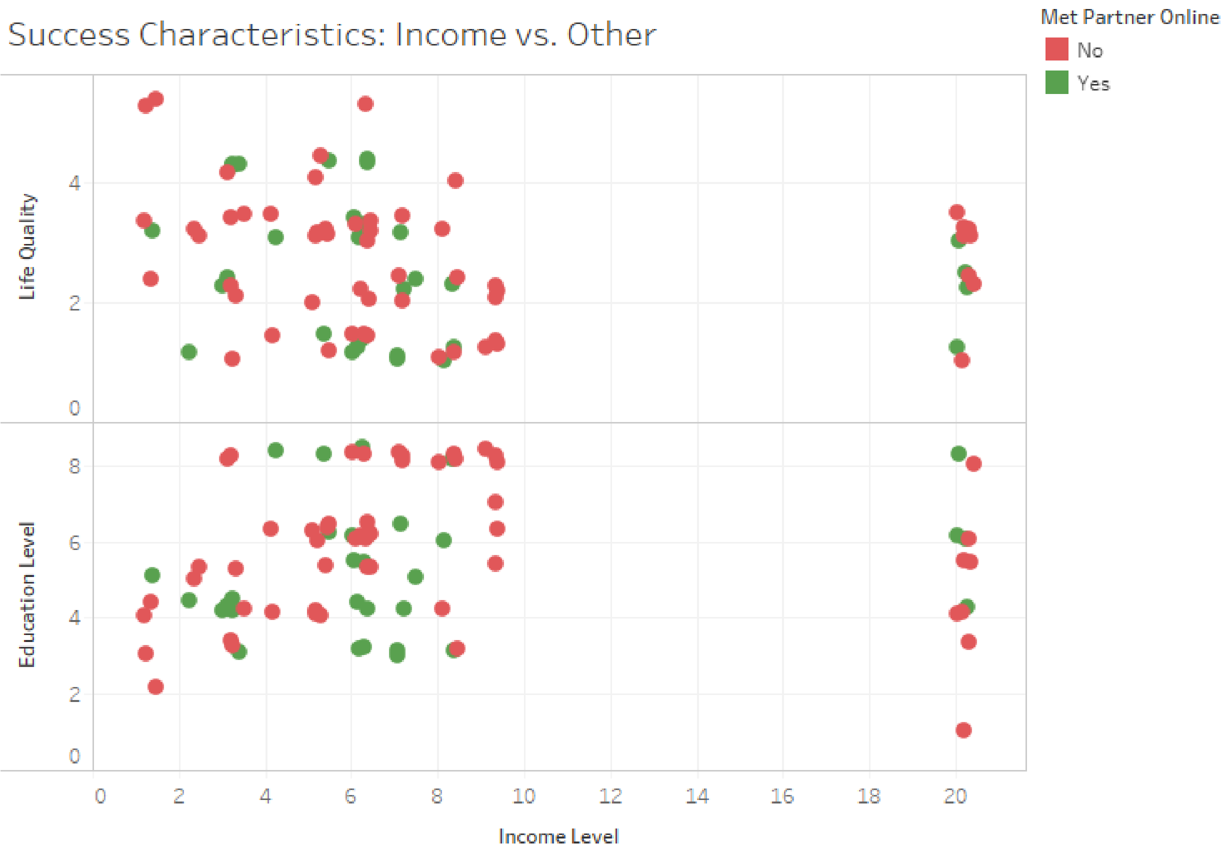
This figure shows that of those who met their partner online, 10.4% are divorced and 0% are separated, compared to 5.6% divorced and 1% separated for those who did not meet their partners online. So it appears that there is evidence that those who meet online have a higher divorce rate.

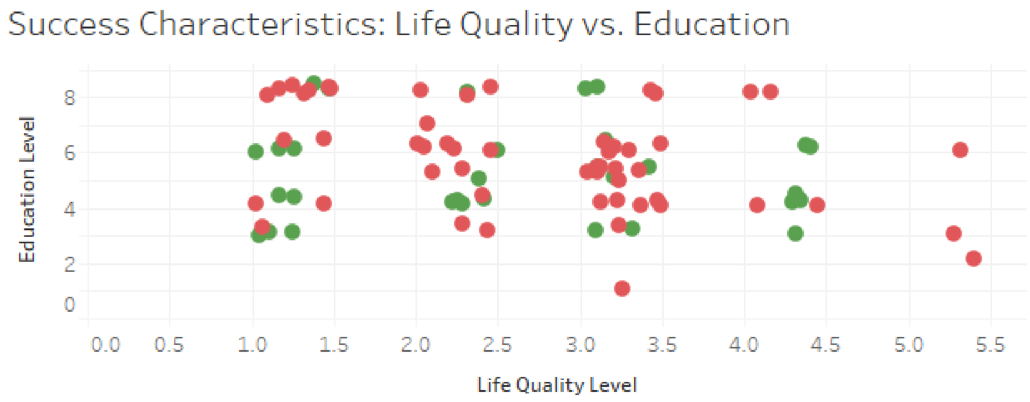
What could be improved about this view:

Some particular demographic may be pushing the divorce rate up for those who met their partner online, so splitting that out (if it exists) might be helpful. Also showing a time series of divorce over time for the two groups may be more interesting.



Success Characteristics: Age vs. Other





What's informative about this view:

Scatter plot charts were made to compare characteristics for those who do online dating to see if clusters emerge for those who were successful at meeting their partner online. All combinations of characteristics were made in the scatter plots, for a total of $\binom{4}{2}$= 6 plots:

- age vs. income, life quality, education level
- income vs. life quality, education level
- life quality vs. education level

Groups of green circles represent clustering of characteristics for people who met their partners online, with jittering applied to the data to allow multiple data point to be seen that would otherwise be occluded.

From these scatter plots, I can see no clustering of characteristics for those who successfully found their partner online.

What could be improved about this view:

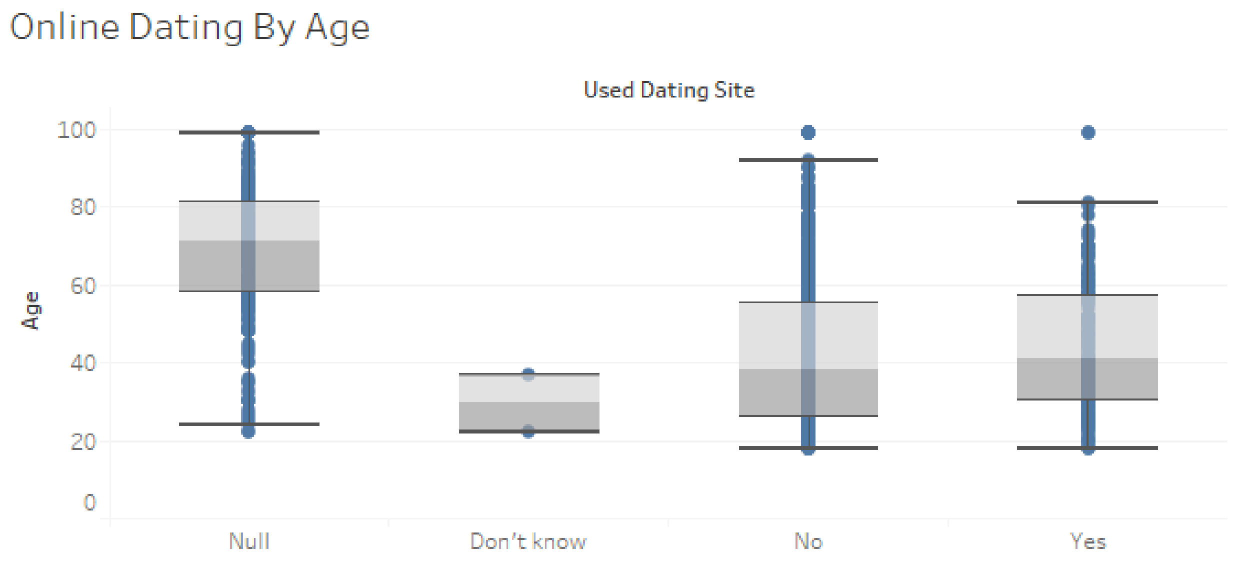
For income, I had to create a new field called 'incomelimit', as the original 'income' field had data points such as 98 and 99 that were outliers and made it hard to view the main distribution. The values in this new field are limited to the value 20, which can be seen by the column of data points at value 20. These data points could possibly be removed.

Conclusion

36% of those who used online dating met their partner online. Because this is below the stated hypothesis (50%), the hypothesis cannot be rejected.

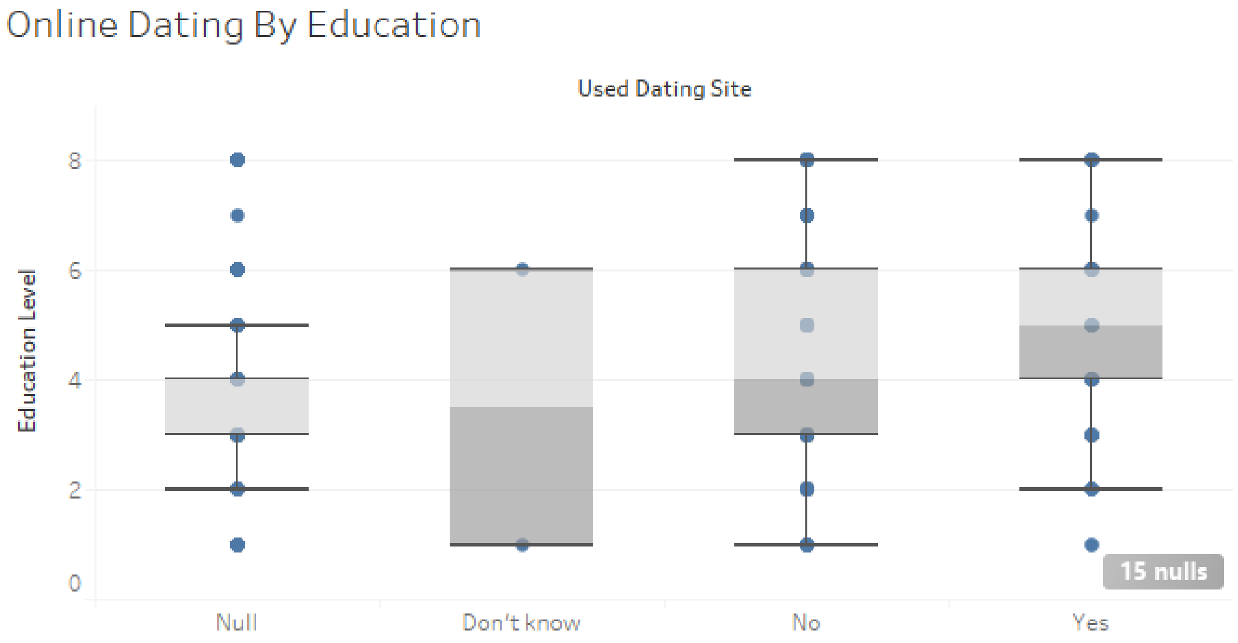
Hypothesis 2: People who use online dating to find a partner are less likely to have children.

First, I wanted to see if there was a difference in the age distribution for those who use online dating and those who do not. A box plot is a nice way to show the median, interquartile range, and outliers of the data, and is also good at showing distributions side by side for comparison. A filter was added on the data to exclude people who have been in a relationship more than 15 years (15 years prior to 2013 was 1998, approximately when online dating became popular).



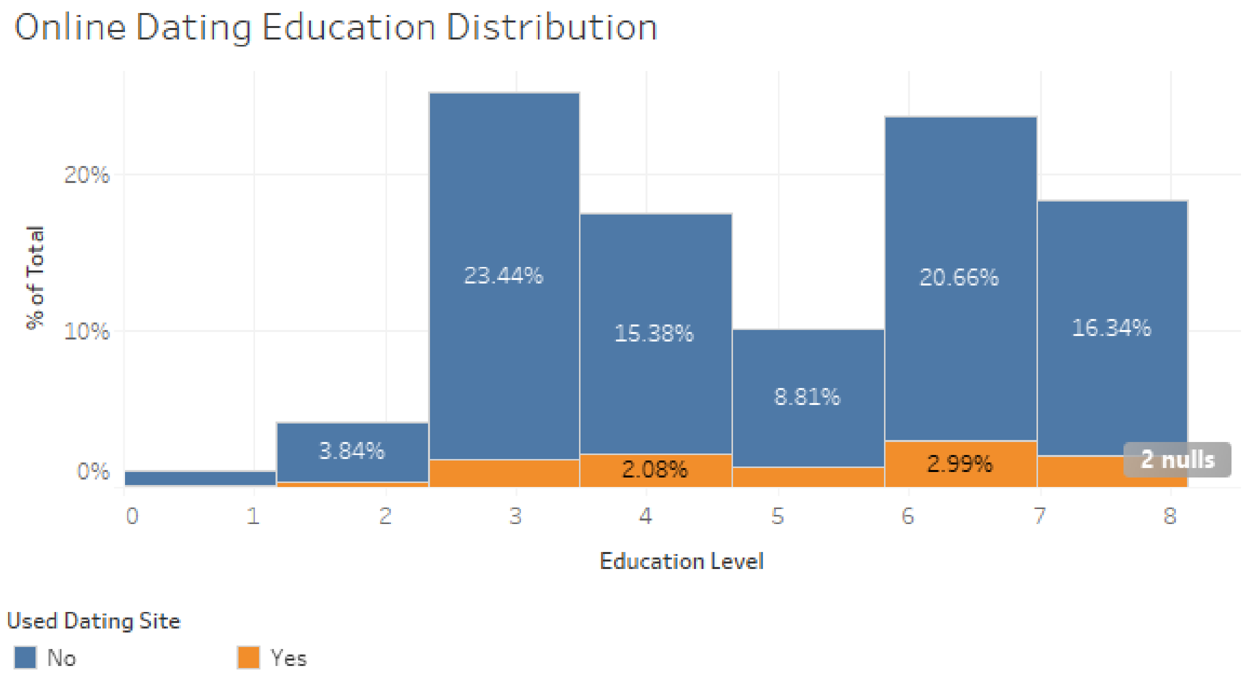
The boxplot shows that those who answered ‘Yes’ is slightly higher median (41 years) compared to those who answered ‘No’ (38 years). A significant portion of the survey respondents did not respond to this question, and their median age is 71. It is possible that many of these people have used online dating, and including these non-compliers would change the results of this comparison. Therefore it remains unclear whether those who use online dating tend to be older.

Next, I was interested in seeing if there was a difference in the level of education distribution for those who use online dating. A box plot was used again, and the years in a relationship filter was again used to compare only the group of people who could have use online dating.



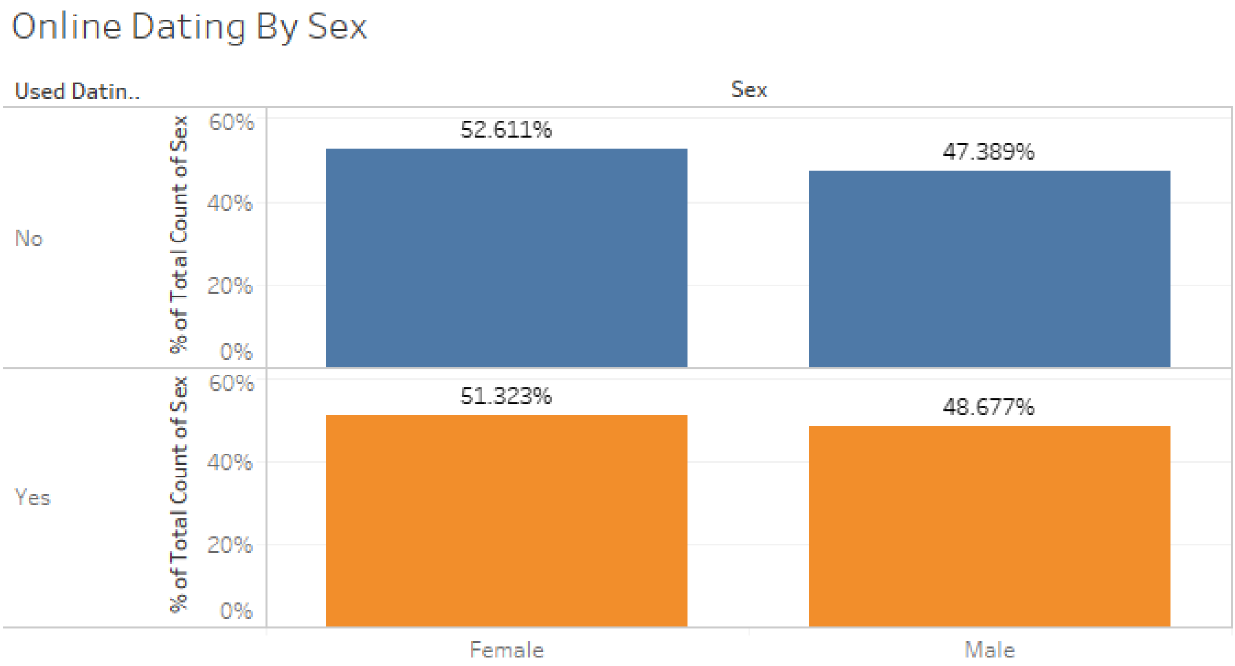
The boxplot shows that those who answered ‘Yes’ is slightly higher median (level 5) compared to those who answered ‘No’ (level 4). A significant portion of the survey respondents did not respond to this question, and their median education level is 3. Assuming that the non-compliers have equal chance of using / not using online dating, then the education level distribution for ‘Yes’ would be lowered, but still remain above ‘No’. So there is slight evidence that those who use online dating have, on average, a higher education level.

As a refinement to the above chart, I also made a overlapping histogram to better compare the education distributions of those who did and did not use online dating:



The histogram does not seem to provide any better clarity in the distribution of the education levels.

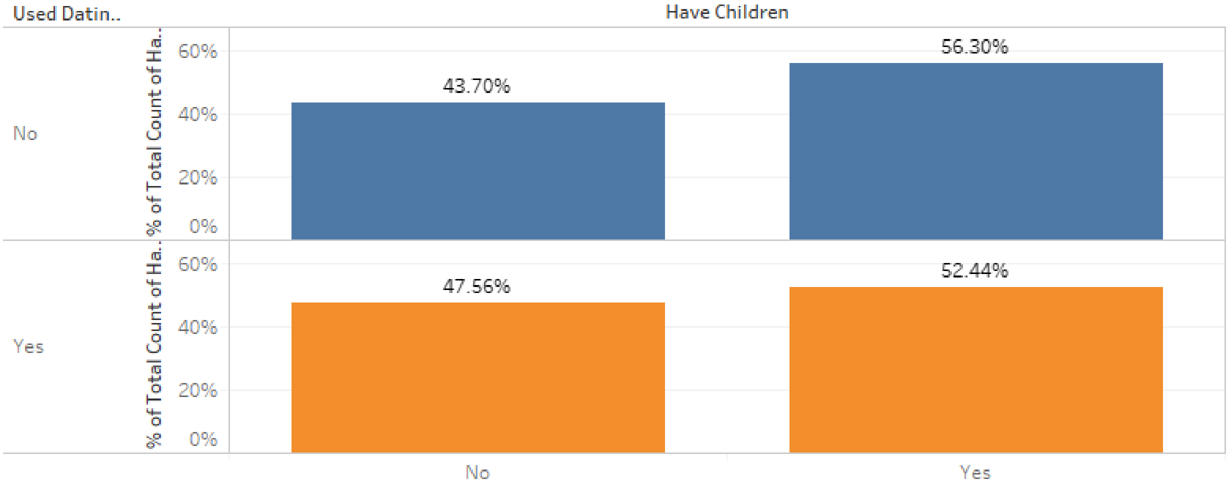
Next, I was interested in determining if one sex uses online dating more than another. Since Sex is a Dimension in Tableau, I created a bar chart comparing the proportion of those who used online dating between males and females, with the same filter on the years in a relationship.



The graph shows that of those who use online dating, 51.3% are female and 48.7% are male. Therefore it appears that there is a slight difference in favor of males for online dating.

In order to directly test the hypothesis, I created a bar chart of ‘Have Children’ plotted as a bar chart along with ‘Used Dating Site’. It was aggregated to show the proportion of those who have children for those who did and did not use online dating. An age filter was applied to capture those who were 40 years of age at the time that online dating became popular (the upper end of the child rearing years) down to 20 years of age (at the low end in age of those who start having children). Another filter was used to only compare those who lived together at some point.

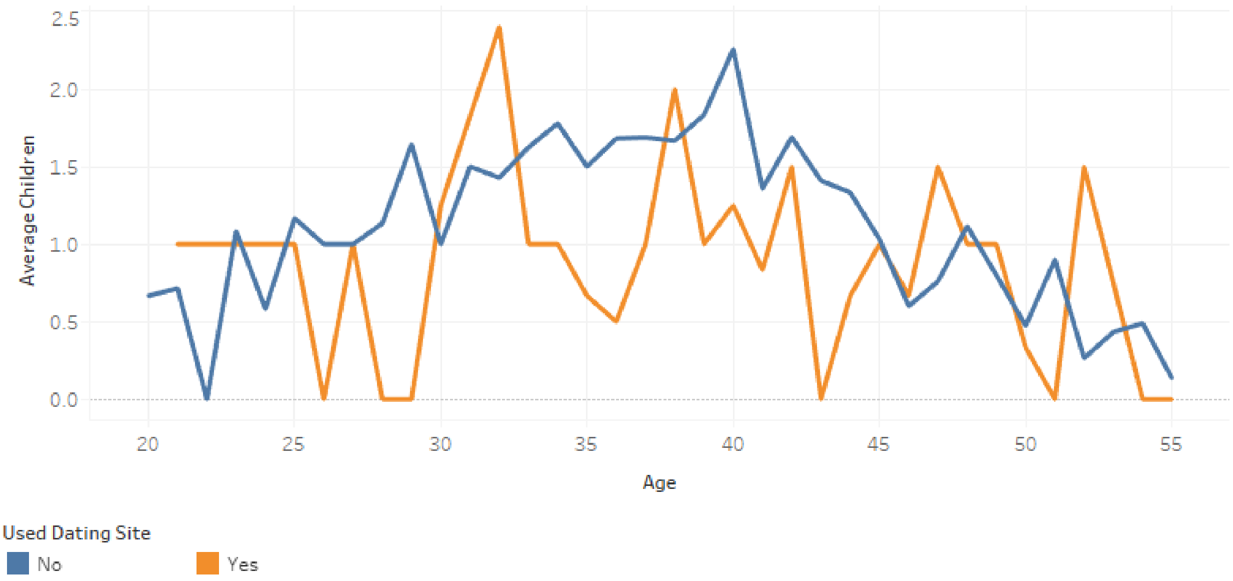
Online Dating By Have Children



This figure shows that the proportion of those who use online dating who have children is 52.4%, compared to 56.3% of those who do not use online dating, so based on this data, the hypothesis cannot be rejected. The answer is not entirely clear, however, as there are a number of non-compliers, which could influence the results.

A final time series graph was created to show the number of children for those who used online dating with those who did not, per age range. This required adding a new field to the data set with the total number of children per respondent.

Online Dating By Total Children



This graph shows that those who use online dating seem to mostly have fewer children during the childrearing years. There are also a couple late spikes in the later years that suggests that those who used online dating started their families later in life, because there are still children ages 0-17 in their household.

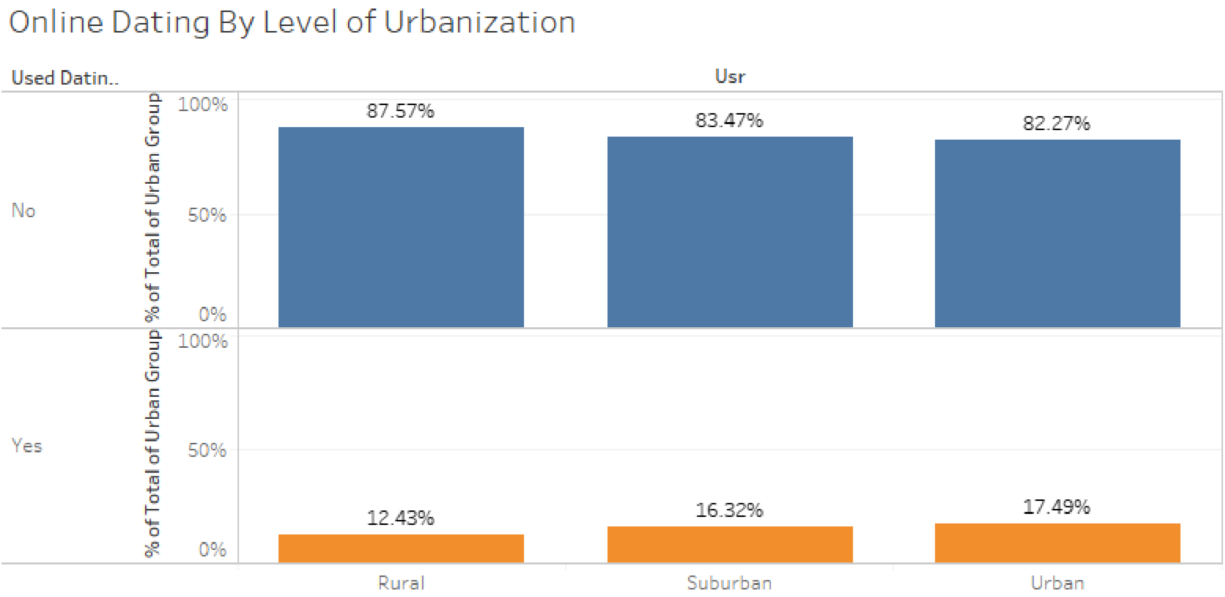
Conclusion

The results show that respondents who use online dating have fewer children than those who do not. However, there were a significant number of non-compliers, so it is not entirely clear if there is no difference between the groups.

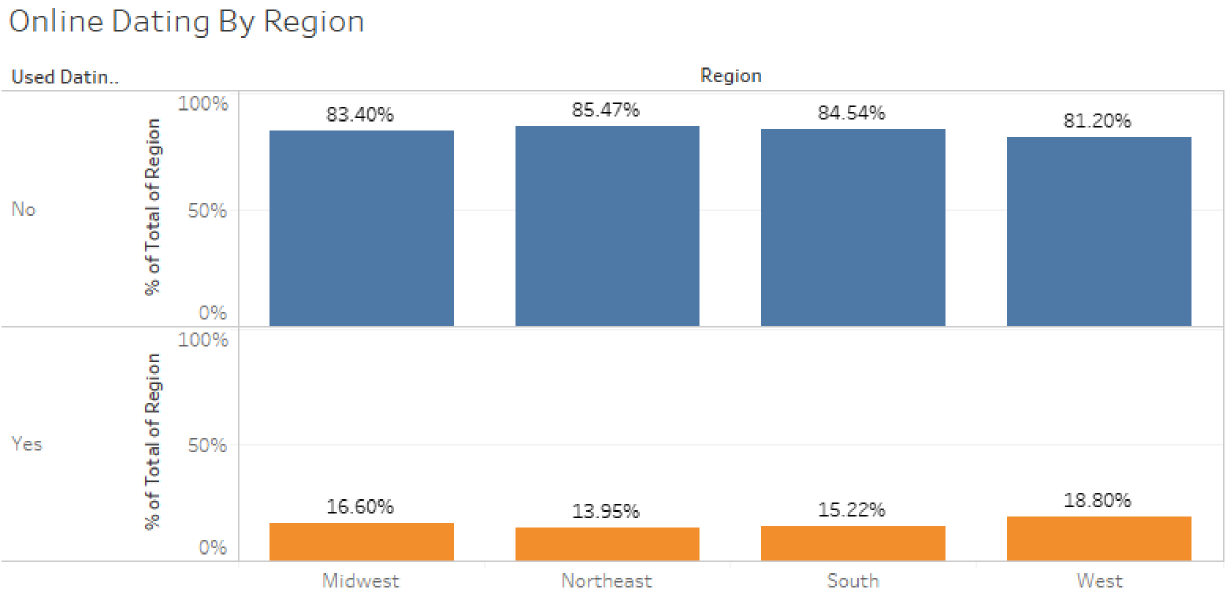
Hypothesis 3: People in both urban and rural environments use online dating at the same rate.

First, I wanted to visualize the proportion of those who use online dating based on the level of urbanization that the survey respondent is from. This was done by aggregating the 'Usr' field and plotting it against 'Used Dating Online', and then displaying the proportion as a percent.

A filter was created to only compare respondents who were 'Online'. This was a field I added to the data set that is true if the user responded true if they had Internet, email, or used social media. This field was necessary as 'Use Internet' was not a reliable field to determine if the respondent had an Internet connection. A second filter was used to only include respondents who were in a relationship less than 15 years. This excludes respondents who started a relationship prior to when online dating became popular in 1998.

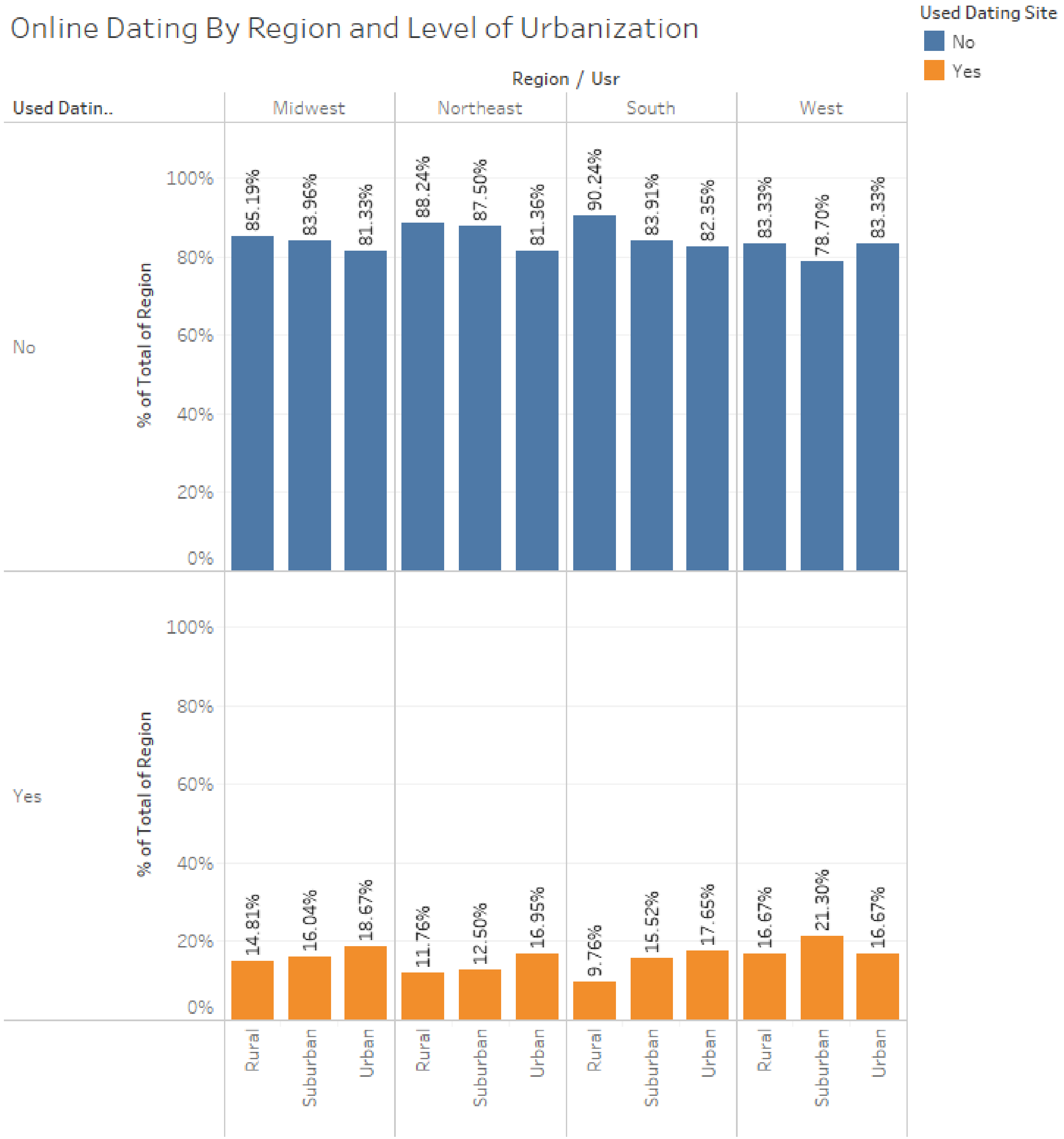


Next I wanted to see if people used online dating the same across different regions of the US. Note that the 'Northeast' is considered to be the most urban region, whereas the 'West' is considered to be the most rural region. The same filters were used to select respondents as in the previous chart.



From this visualization it appears that the highest online dating percentage is in the West (18.8%), and the lowest is in the Northeast (14%), which is a surprising result.

Creating a refinement of the above graphs that include comparisons of both region and urbanization, we have:



For all regions of the country except the west, urban users have a higher usage of online dating. However in the west, suburban users (21.3%) use online dating more than urban users (16.7%). This then explains why the west may have the highest online dating usage even though urban users use online dating (on the whole of the US) more than rural and suburban users.

Conclusion

Based on the graph of ‘Online Dating By Level of Urbanization’ , there appears to be evidence against the hypothesis that rural and urban areas use online dating at the same rate. In this case, 12.4% of the respondents in rural areas use online dating, compared to 17.5% of respondents in urban areas.

Appendix A: R Code For New Fields

The following R code was used to create additional features for the visualizations in this report:

```
wd <- './Homework2'
setwd(wd)

dating.df <- read.csv('Dating_saved.csv')

#
# Determine if a person can access the internet. This means answering true to
# any of the following: 'use internet', 'use email', 'use social networking'
#
dating.df$online <- dating.df$use_internet == "Yes" |
                    dating.df$use_email == "Yes" |
                    dating.df$use_social_networking == "Yes"

#
# Assume data is wrong if person has more than 10 children. Set to 0 children.
#
dating.df$children0_5 <- ifelse(dating.df$children0_5 > 10, 0, dating.df$children0_5)
dating.df$children6_11 <- ifelse(dating.df$children6_11 > 10, 0, dating.df$children6_11)
dating.df$children12_17 <- ifelse(dating.df$children12_17 > 10, 0,
dating.df$children12_17)

#
# Calculate total children
#
dating.df$totalchildren <- ifelse(!is.na(dating.df$children0_5) &
!is.na(dating.df$children6_11) & !is.na(dating.df$children12_17),
dating.df$children0_5+dating.df$children6_11+dating.df$children12_17, 0)

#
# Calculate income with outliers set to a limit of 20
#
dating.df$incomelimit <- ifelse(dating.df$income > 20, 20, dating.df$income)

write.table(dating.df, file ='Dating2.csv', quote = FALSE, sep=',', row.names = FALSE)
```