A2. Exploratory Data Analysis
Yang Yang Qian
W209 – Section 3

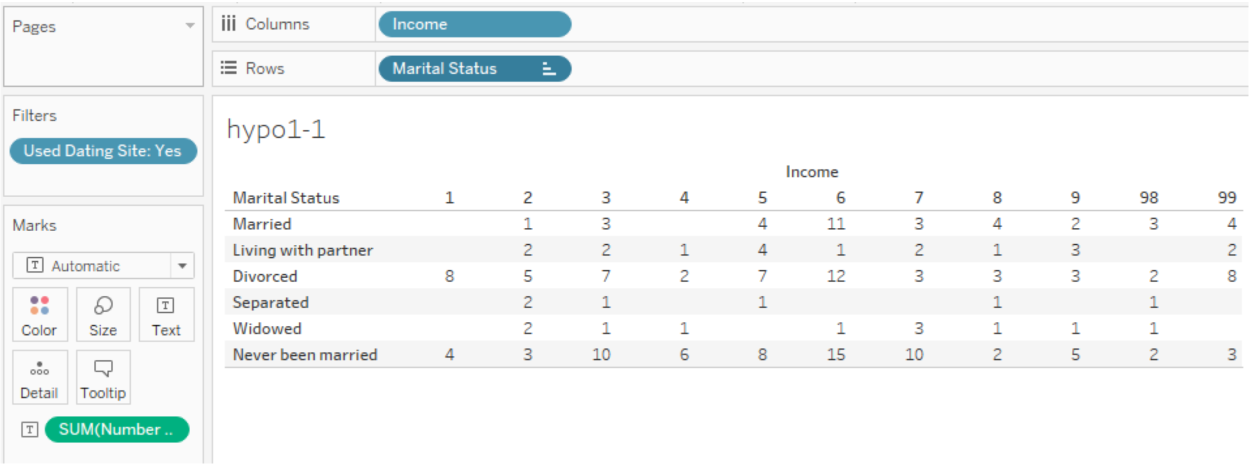
1. Online Dating and Relationships

http://annettegreiner.com/vizcomm/Dating.csv
From the Pew Research Center's Internet and American Life Spring Tracking Survey, April 17-May 19, 2013. N=2,252 adults ages 18+. Interviews were conducted in English and Spanish and on landline and cell phones. This is a national survey of dating and relationships in the digital era, the first dedicated study of this subject by the Pew Research Center's Internet Project since 2005. A list of the Survey questions is available for download as a PDF.
For original, see
http://www.pewinternet.org/dataset/may-2013-online-dating/

2. Hypothesis 1

Among those who have used online dating services, those with higher income are more likely use be single.

2.1. View 1



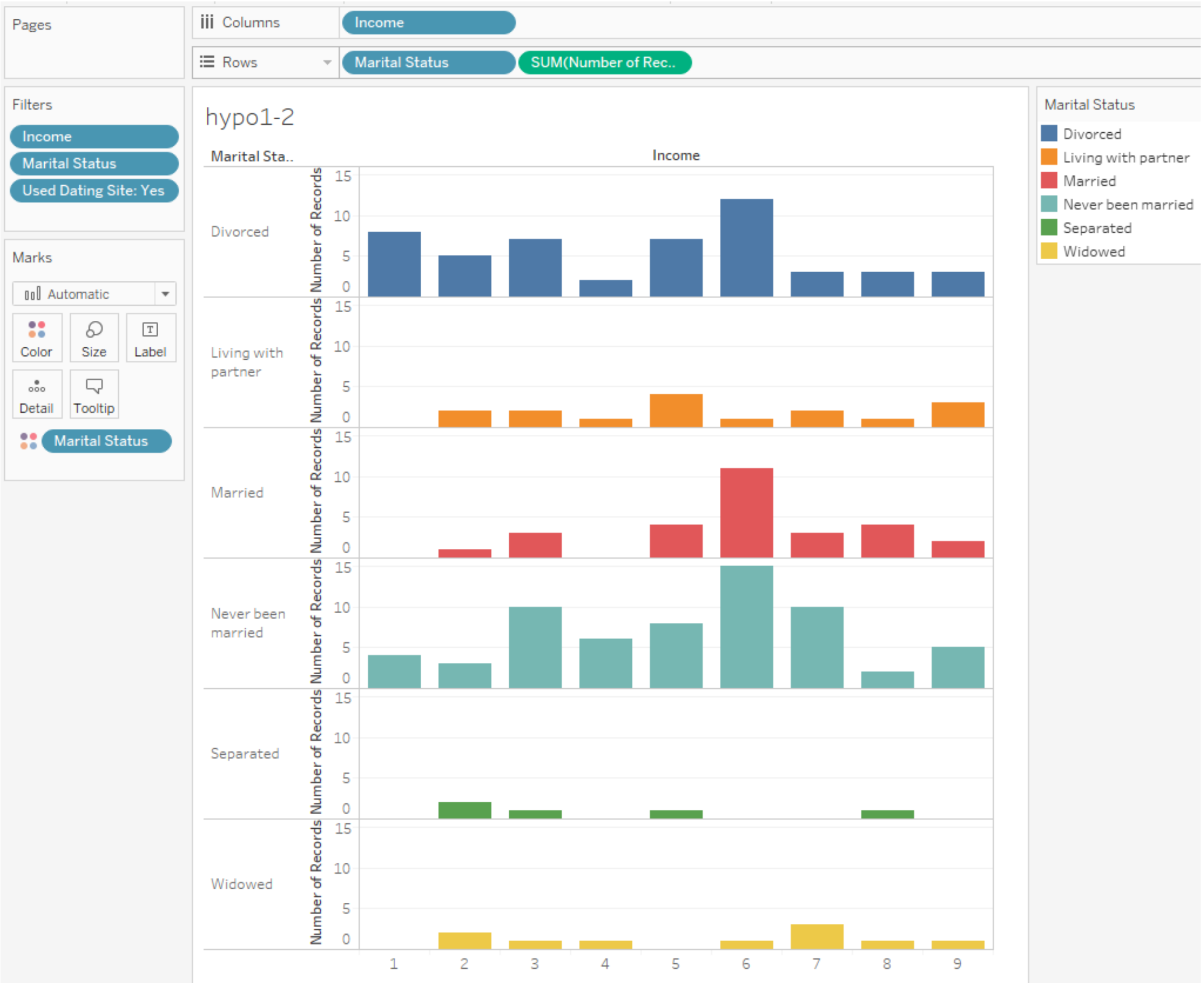
hypo1-1

	Income										
Marital Status	1	2	3	4	5	6	7	8	9	98	99
Married		1	3		4	11	3	4	2	3	4
Living with partner		2	2	1	4	1	2	1	3		2
Divorced	8	5	7	2	7	12	3	3	3	2	8
Separated		2	1		1			1		1	
Widowed		2	1	1		1	3	1	1	1	
Never been married	4	3	10	6	8	15	10	2	5	2	3

What's informative about this view: This table shows the exact number of respondents for each of the Marital Status categories and Income brackets, for those who have used online dating services. This view is useful for a first level exploration of the data. It shows that certain groupings may have too view samples to draw strong statistical conclusions from, so some re-binning might be needed. For example, we only have 1 respondent for those who were married and had Income level of 2 (10 to under $20,000). And it also shows some respondents Refused to answer the income question. We might want to cut down on the noise by excluding that subset of respondents.

What could be improved about this view: This view is not very good at showing patterns among the categories; it only uses text to convey all the data. A different visual idiom, such as a bar chart, might be more appropriate.

2.2. View 2



Pages

Columns: Income

Rows: Marital Status, SUM(Number of Rec..

Filters: Income, Marital Status, Used Dating Site: Yes

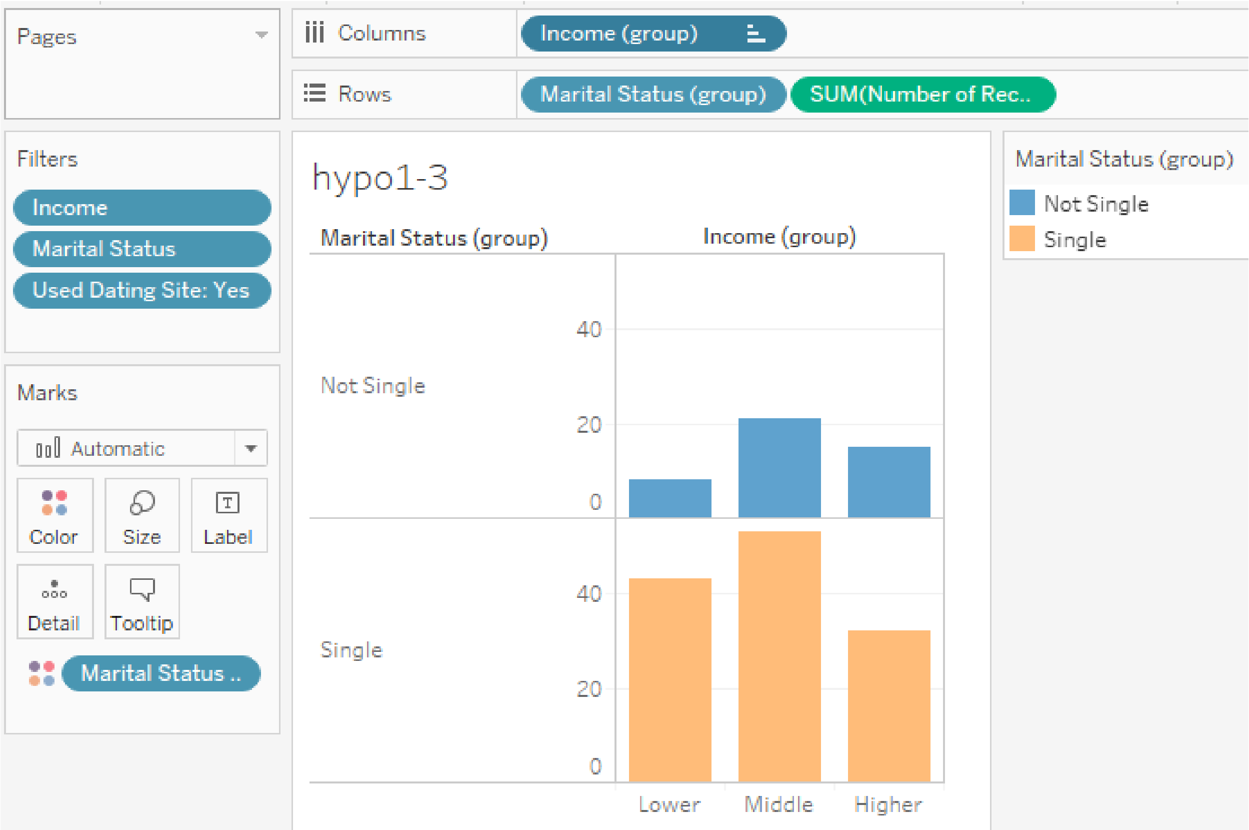
Marks: Automatic

Color: Marital Status

What's informative about this view: This view uses the color and length to shows the distribution of counts of respondents per Marital Status and Income. It improves on the previous view by filtering out any irrelevant categories. i.e. those who did not refuse to answer their household income & marital status.

What could be improved about this view: However, it is very difficult to see any overall trends, because there are too many Marital Status and Income bins.

2.3. View 3



Pages

Columns: Income (group)

Rows: Marital Status (group) SUM(Number of Rec..

Filters

Income

Marital Status

Used Dating Site: Yes

Marks

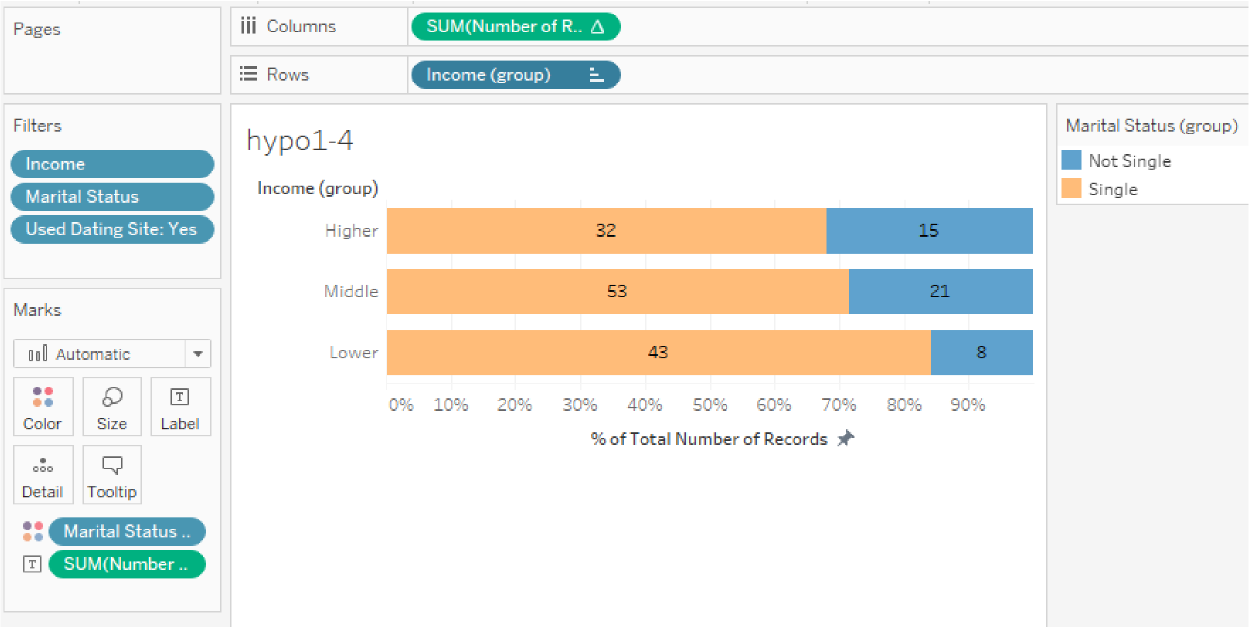
Automatic

Marital Status ..

What's informative about this view: This view re-bins the Marital Status and Income values, so that the overall trends are clearer.

What could be improved about this view: However, it is still very difficult to see any meaningful conclusion. Each of the different combinations of Marital Status and Income groups have different subtotals. Therefore, it is tough to compare the relative likelihoods.

2.4. Final view


Pages

Columns: SUM(Number of R.. Δ

Rows: Income (group)

Filters: Income, Marital Status, Used Dating Site: Yes

Marks: Automatic

Color: Marital Status ..

Label: SUM(Number ..

What's informative about this view: This bar chart shows the ratios of Marital Status groups, per Income group. Because it shows the counts as percentage of totals, it more clearly shows the overall trend of likelihoods. It is more compact and efficient at showing the data vs some of the previous views.

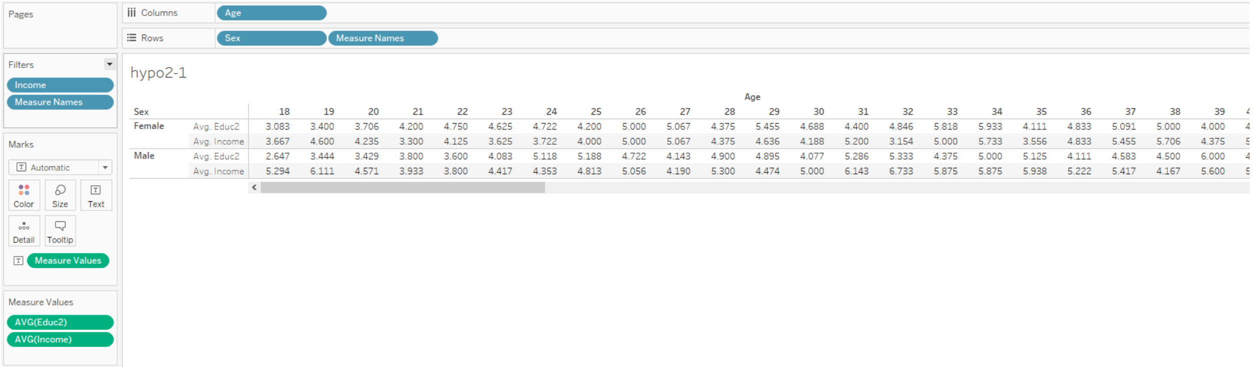
What could be improved about this view: The choice of bins for Income and Marital Status is somewhat arbitrary. Different bins can affect the shape of the visuals.

Conclusion This view does not support the hypothesis. Instead, it suggests that for those who have used online dating services, the lower income groups are more likely to be single.

3. Hypothesis 2

As our male and female respondents age, they tend to have more education and more income.

3.1. View 1



hypo2-1

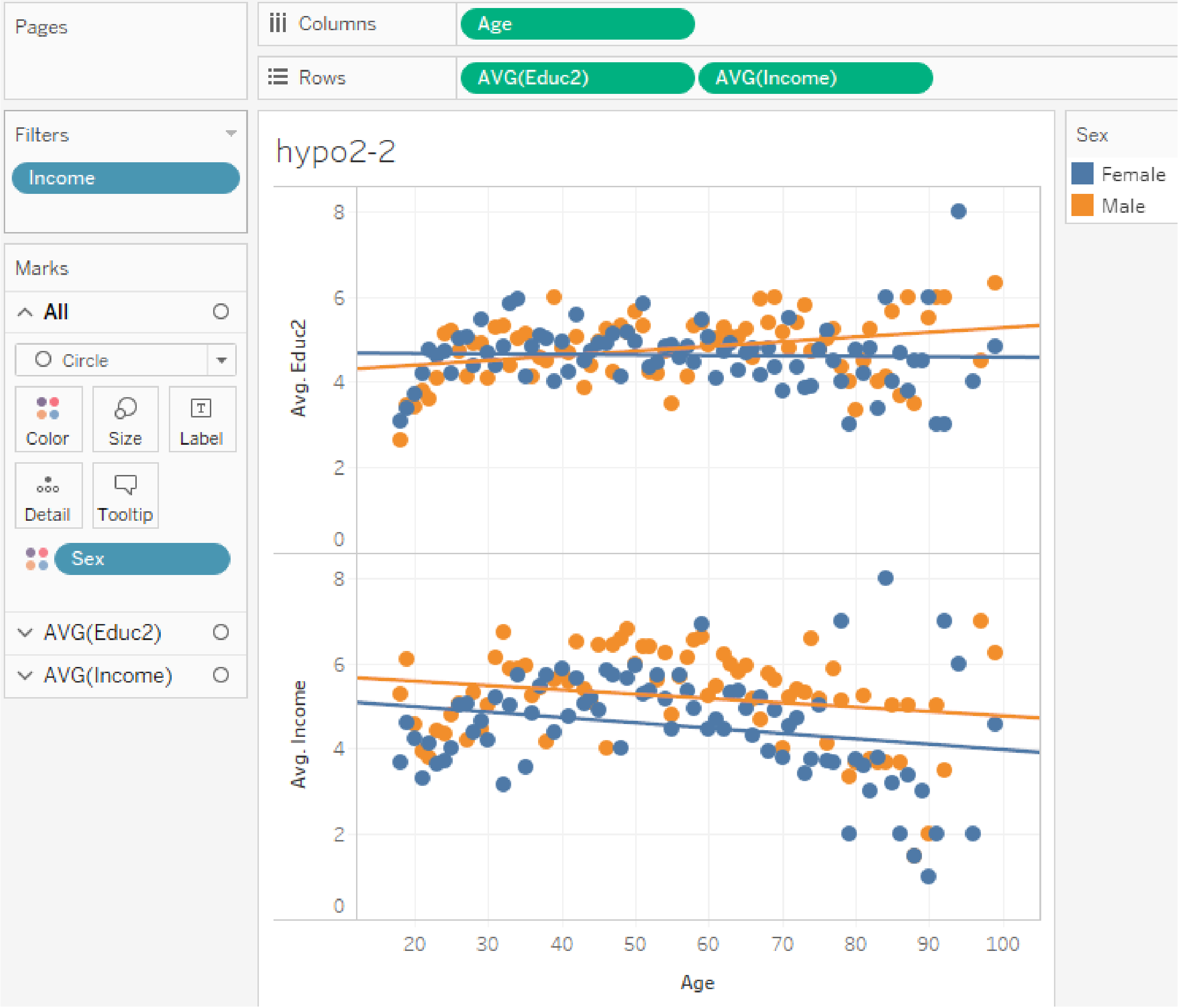
		Age									
Sex		18	19	20	21	22	23	24	25	26	27
Female	Avg. Educ2	3.083	3.400	3.706	4.200	4.750	4.625	4.722	4.200	5.000	5.067
	Avg. Income	3.667	4.600	4.235	3.300	4.125	3.625	3.722	4.000	5.000	5.067
Male	Avg. Educ2	2.647	3.444	3.429	3.800	3.600	4.083	5.118	5.188	4.722	4.143
	Avg. Income	5.294	6.111	4.571	3.933	3.800	4.417	4.353	4.813	5.056	4.190

		Age											
Sex		28	29	30	31	32	33	34	35	36	37	38	39
Female	Avg. Educ2	4.375	5.455	4.688	4.400	4.846	5.818	5.933	4.111	4.833	5.091	5.000	4.000
	Avg. Income	4.375	4.636	4.188	5.200	3.154	5.000	5.733	3.556	4.833	5.455	5.706	4.375
Male	Avg. Educ2	4.900	4.895	4.077	5.286	5.333	4.375	5.000	5.125	4.111	4.583	4.500	6.000
	Avg. Income	5.300	4.474	5.000	6.143	6.733	5.875	5.875	5.938	5.222	5.417	4.167	5.600

What's informative about this view: This table shows that there are data points for both average income and education levels at every age group, for both males and females.

What could be improved about this view: However, it is very difficult to see the overall trend. And there is so much text that on a smaller screen, the user will have to scroll to see all the data.

3.2. View 2



Columns: Age

Rows: AVG(Educ2) AVG(Income)

Filters: Income

Marks: All, Circle, Color: Sex

What's informative about this view: This scatterplot compactly shows trendlines for each of the measures, and for each of our categories. The slope of the trendlines clearly show the direction of the estimated association between age and our measure of interest.

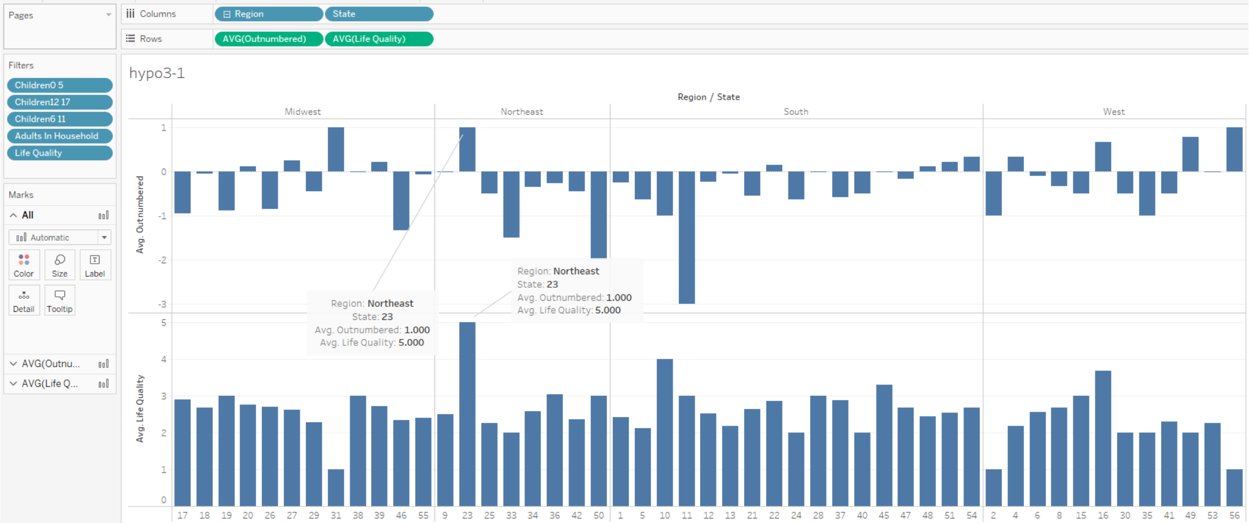
What could be improved about this view: However, this view is only good for linear models with 2 or 3 features. More than that, and the line chart or scatter plot idiom does not work. Also, our choice of just Sex, and Age as dependent variables in the models means our estimates are probably affected by omitted variable biases.

Conclusion This view suggests that as our male respondents age, they tend to have more average education, but less average income. And that as our female respondents age, they tend to have roughly the same average education, and less average income. The view also suggests that both male and female respondents tended to have less income at similar rates as they aged (though males started at a higher income level).

4. Hypothesis 3

As the number of children exceed the number of adults in the family, life quality tends to suffer.

4.1. View 1



Pages

Columns: Region, State

Rows: AVG(Outnumbered), AVG(Life Quality)

Filters: Children0 5, Children12 17, Children6 11, Adults In Household, Life Quality

Marks: All, Automatic

AVG(Outnu...

AVG(Life Q...

hypo3-1

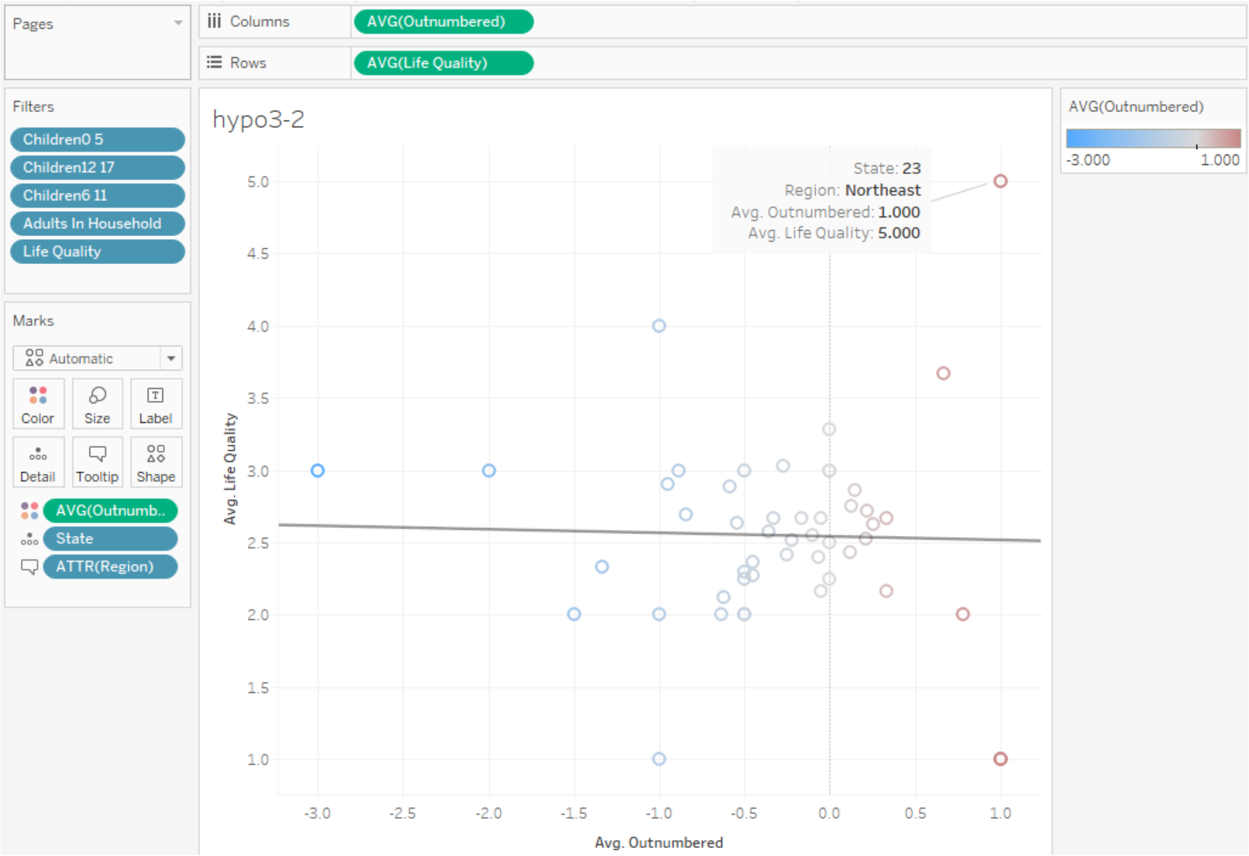
Region: Northeast
State: 23
Avg. Outnumbered: 1.000
Avg. Life Quality: 5.000

Region: Northeast
State: 23
Avg. Outnumbered: 1.000
Avg. Life Quality: 5.000

What's informative about this view: This bar chart shows which states averages both the highest outnumber scores, and life quality scores. The bar charts show the shape of the data. State 23 is clearly has the “hardiest parents” since it has both one of highest average outnumber score, and the highest average life quality score. Note, outnumbered is defined as “(Children 0 to 5 + 6 to 11 + 12 to 17) - Number of Adults”

What could be improved about this view: However, it is very difficult to combine both the outnumbered and life quality scores for visual comparison amongst states. A different visual idiom is needed to show the trend.

4.2. View 2



Columns: AVG(Outnumbered)

Rows: AVG(Life Quality)

Filters: Children0 5, Children12 17, Children6 11, Adults In Household, Life Quality

Marks: Automatic

AVG(Outnumb..

State

ATTR(Region)

State: 23
Region: Northeast
Avg. Outnumbered: 1.000
Avg. Life Quality: 5.000

What's informative about this view: This scatterplot shows the relationship between average outnumbered and average life quality per state. It also uses a divergent color scale to show which states have children outnumber parents on average. The trendline also shows a slight negative slope. Which state has the “hardiest parents” is more apparent in the scatterplot, because it is using position to directly encode both life quality and outnumbered, instead of relying on users to mentally calculate based on compared lengths.

What could be improved about this view: However, this visual assumes at that a simple linear model using just the Outnumbered calculated field can sufficiently explain the variation in Life Quality. If a different linear model with more features is used, the scatterplot idiom will not work.

Conclusion This scatterplot suggests that as the children outnumber the adults, there is a slight negative association with life quality. But it is unclear if the slope estimator suffers from omitted variable biases, or other biases. State 23 in the Northeast region appears to have the “hardiest parents”.