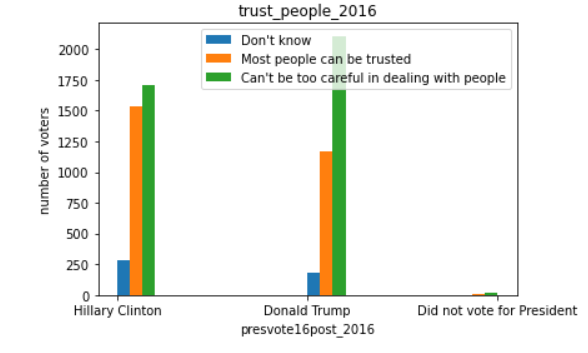
Description of dataset: Survey of 8000 US voters from 2016 that originally contains approximately 600 variables that describe their values, beliefs, demographics, voting record and everyday issues. For this project I selected 20 of these variables that include among others their support for political candidates, religious and smoking habits and basic demographics.

1. Describe data insight no. 1

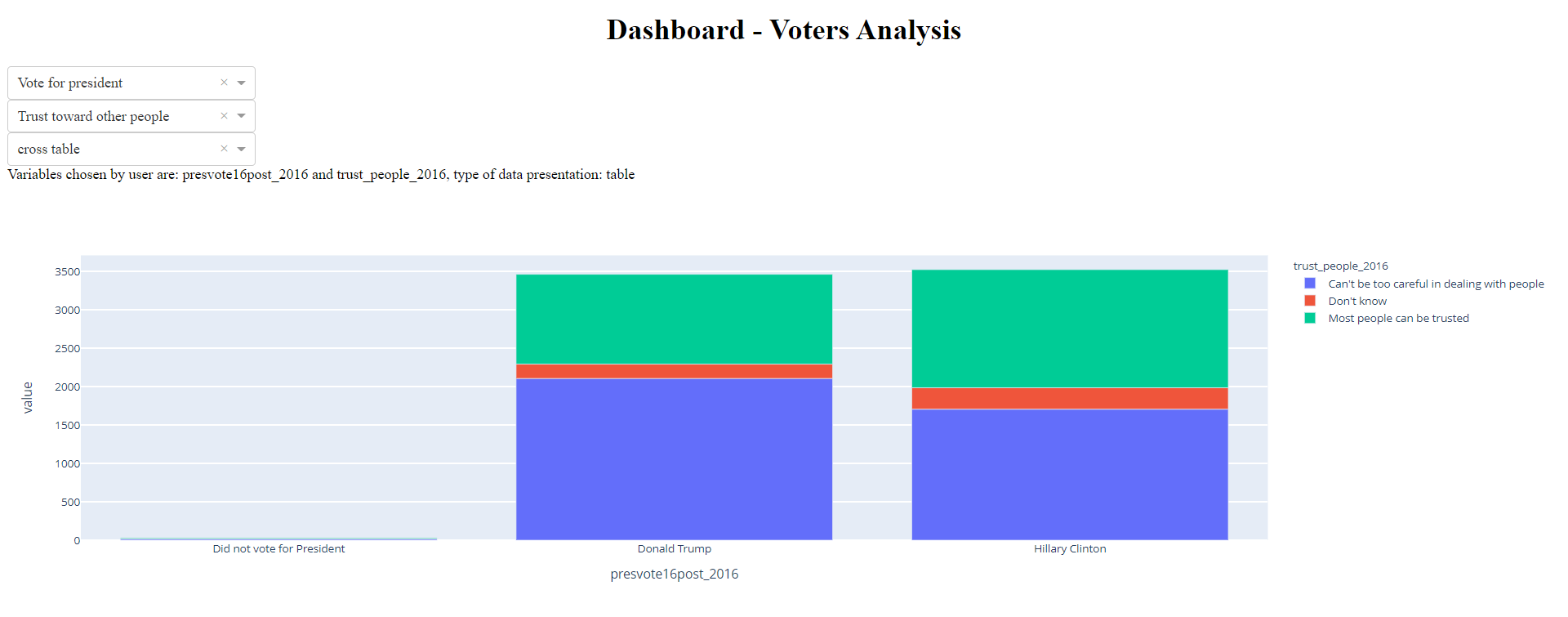
US voters which do not trust other people “Can’t be too careful in dealing with people” tended to vote for Donald Trump.

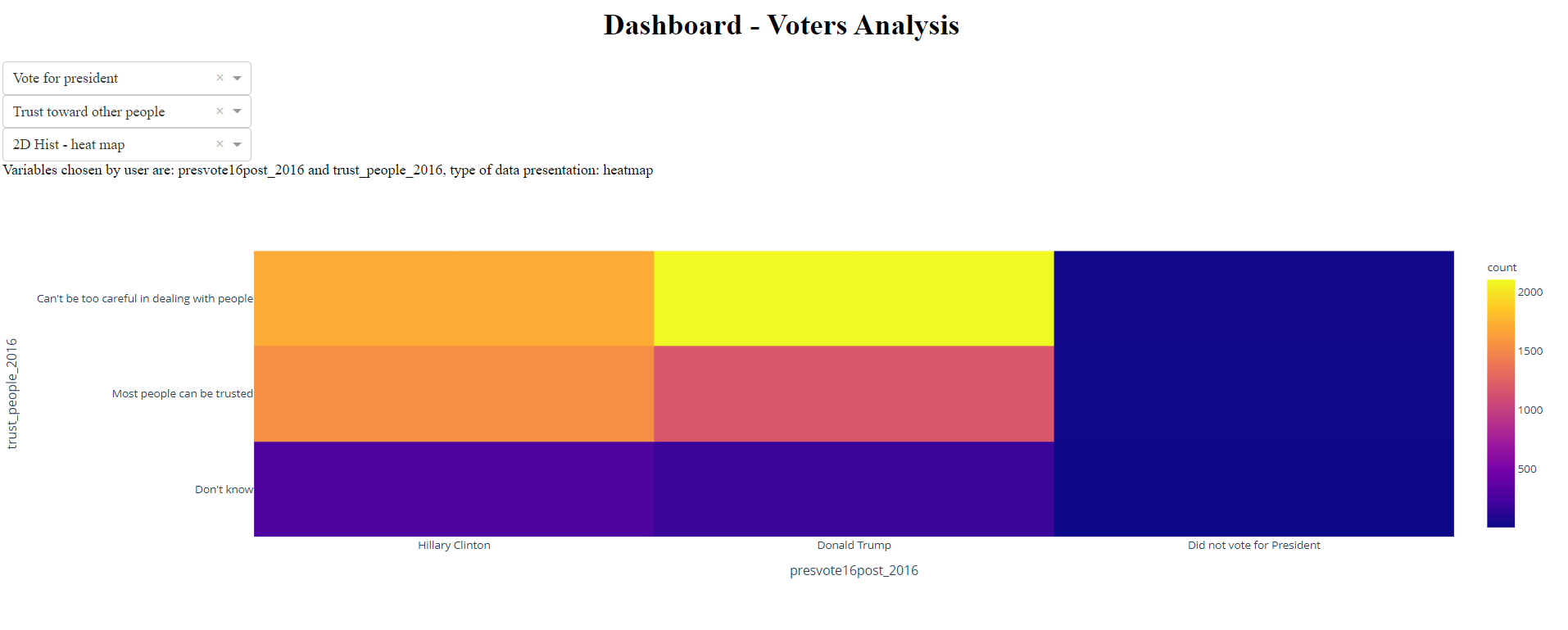
To find different candidates for statistically significant correlations between values in different variables describing surveyed voters I plotted mixed distributions from two variables and then I analyzed statistically significance of found hypotheses.

Distribution representing Trust vs Voting Record:



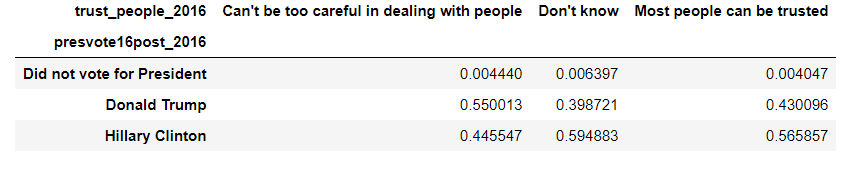
Visual analysis tool output:





Statistical analysis:

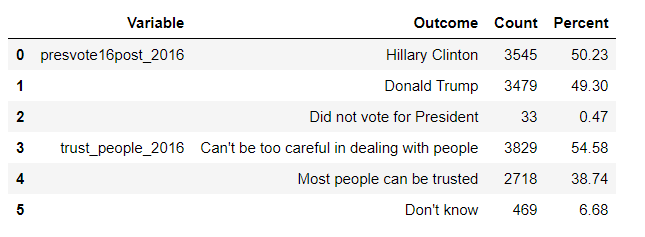
Crosstab which is normalized by columns (each value in column was divided by sum of values in column) shows this tendency in percentages. Among people who do not Trust other people 55 percent voted for Trump while among those who believe that “Most people can be trusted” 57% voted for Hillary Clinton.



Statistical significant of given correlation was tested by calculation of Pearson Chi-square which proved that differences for values in this dataset are not caused by random distribution.

|  |  |
| --- | --- |
| Pearson Chi-square ( 4.0) = | 110.3963 |

Amount of Clinton and Trump voters is evenly distributed in my data what we can see from chart below:

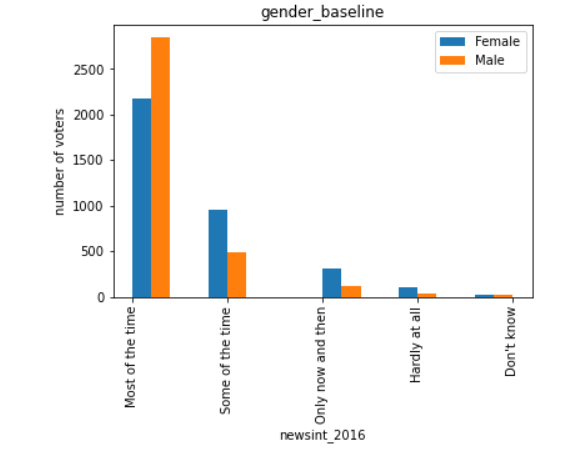


1. Describe data insight no. 2

Men tend to follow news more often than women.

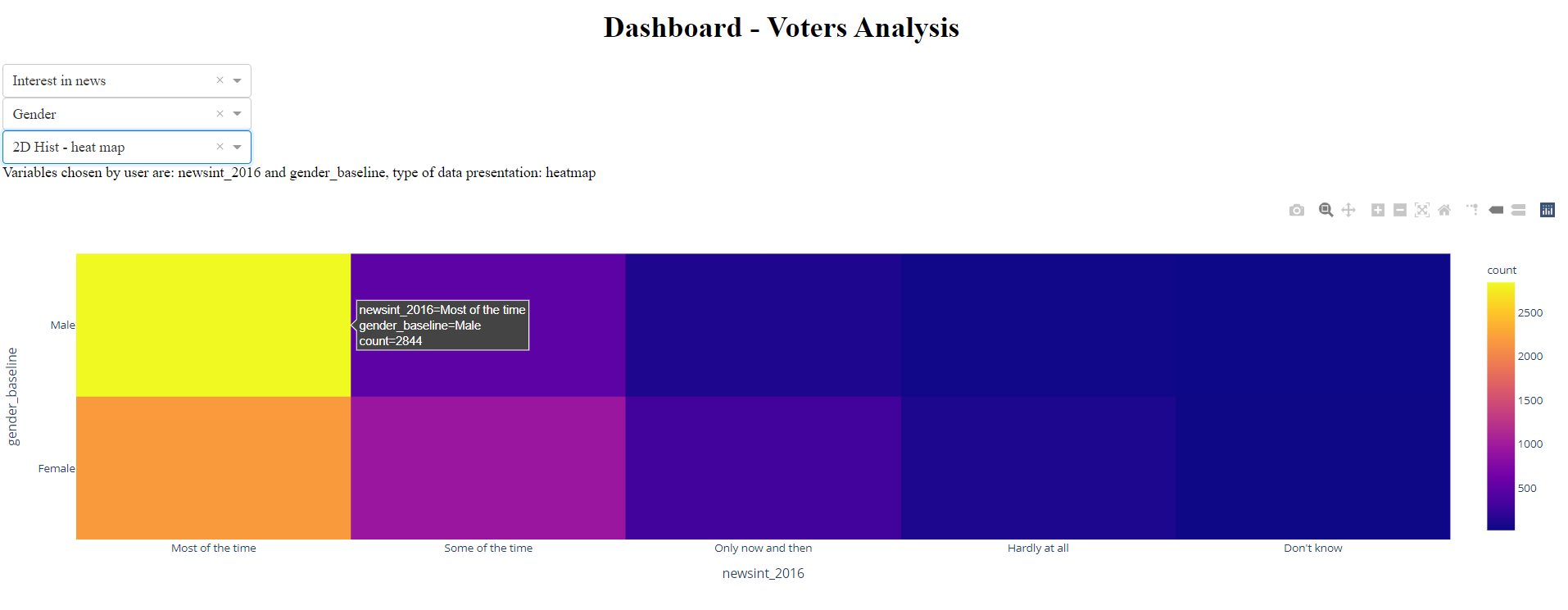
This hypothesis was found from set of plots generated for this dataset the same way as insight from first hypothesis. Initial distribution plot is presented below.

On X axis we have answer to question “How often do you follow news?”



Distribution clearly shows that for “Most of the time” answer clear majority is male and for other answers concerning following news less often dominate answers from women.

Visual analysis tool output:

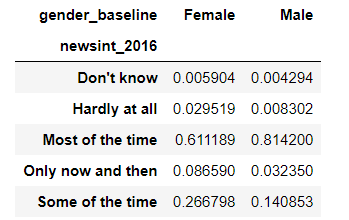


Statistical analysis:

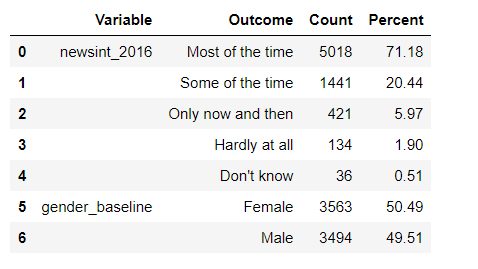
Pearon Chi-square coefficient is very high for this dataset what proves that differences for different values are not caused by random distribution and that there is relation between two variables.

|  |  |
| --- | --- |
| Pearson Chi-square ( 4.0) = | 368.2658 |

From chart below we can see that percentage of males that follow news “Most of the time” is 20 percentage points bigger than for corresponding Female group.



There is one remark worth mentioning about distribution of answers about frequency of news following that holds for both sexes. In set which is analyzed group of people that follow news often is strongly dominant:



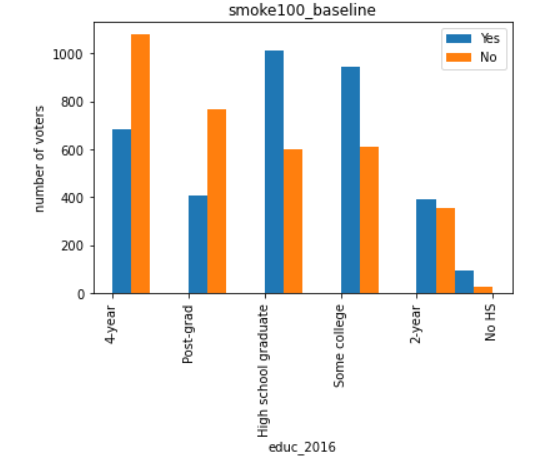
1. Describe data insight no. 3

People with higher education tended to smoke smaller amount of cigarettes in their whole life than people with lower education.

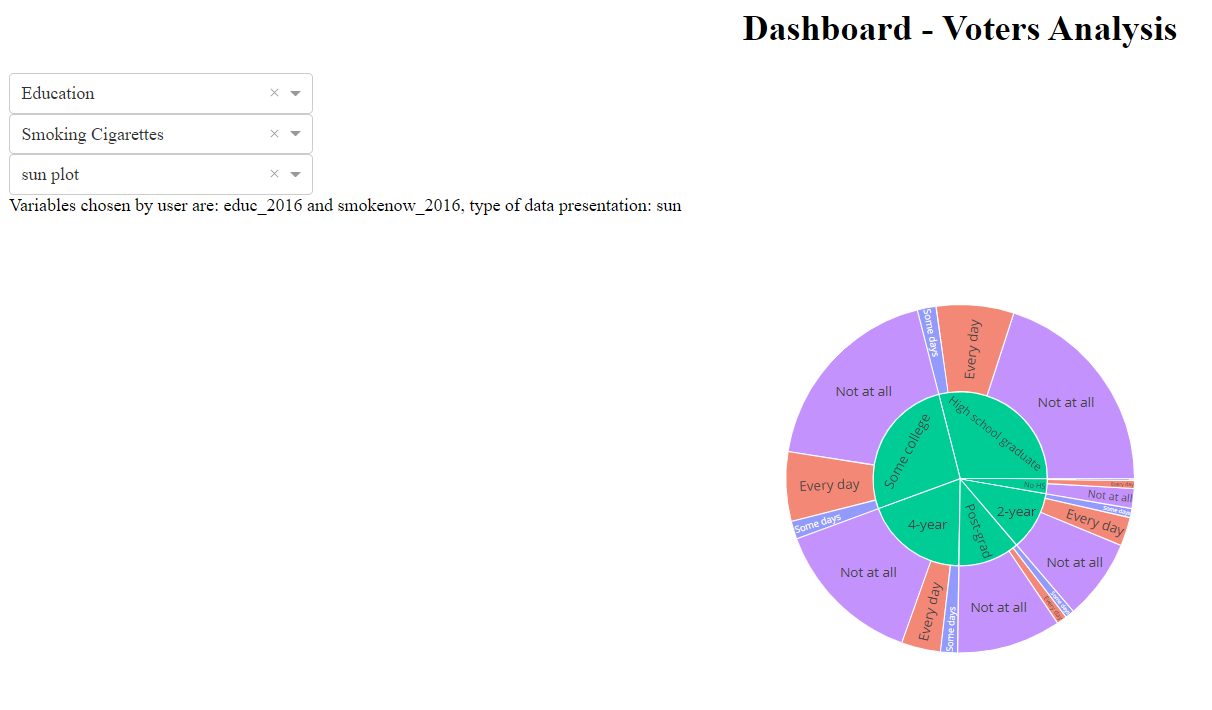
Plot below represents on X axis education level of surveyed people (4 year college, post-graduate study, high school graduate, some college, 2 year college, no high school).

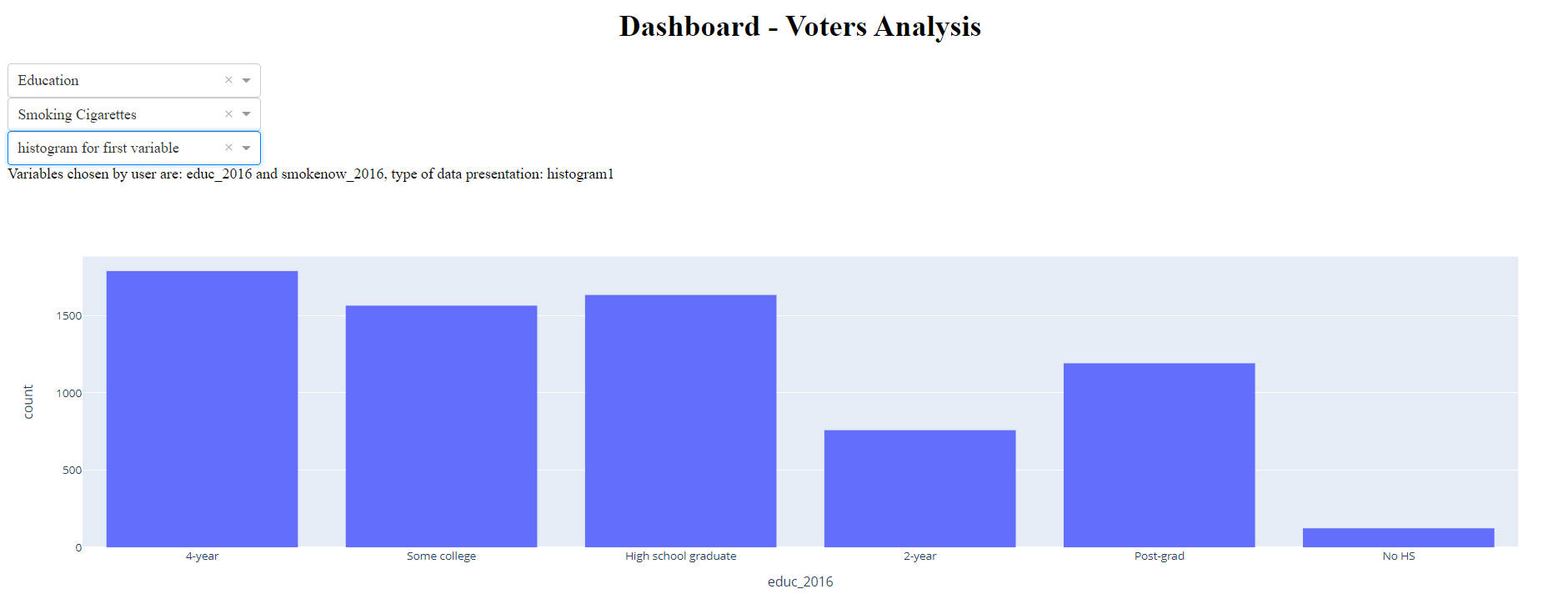
Second variable smoke100\_baseline represents answer to question: “Did you smoke at least 100 cigarettes in your whole life?”

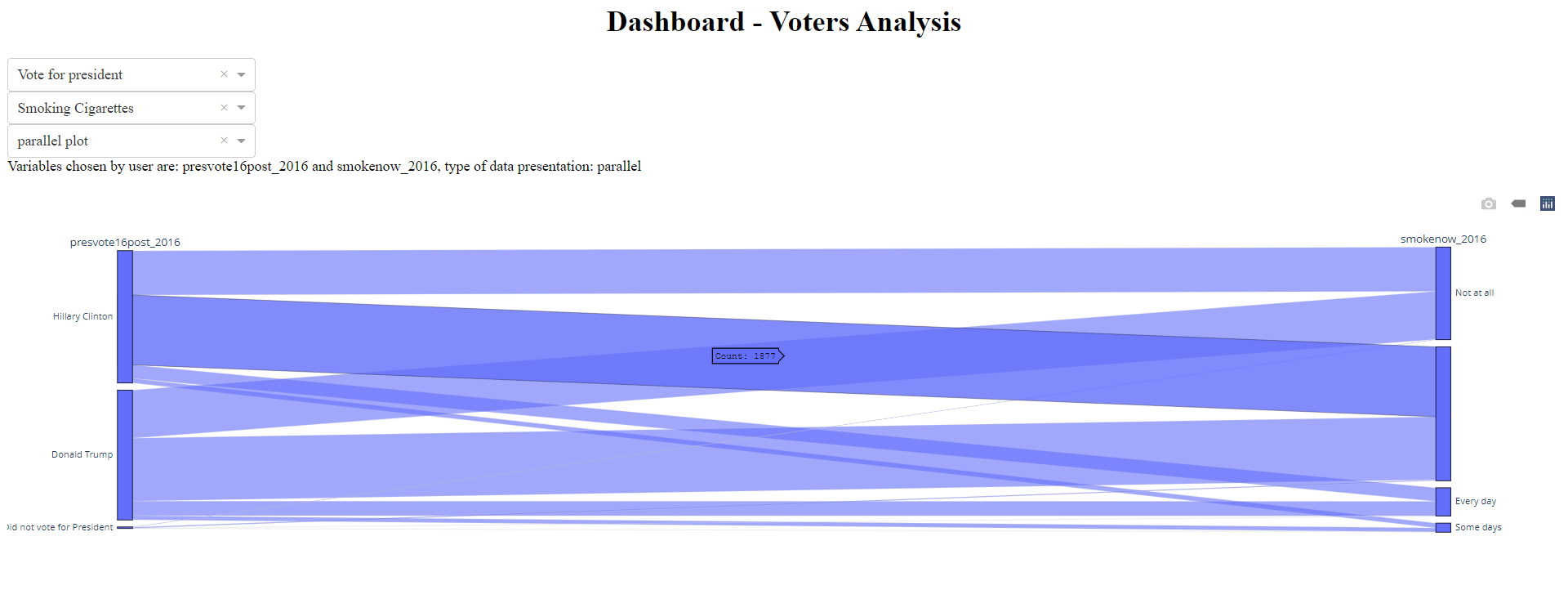
I decided to analyze this correlation because we can clearly see difference for people with higher education (4 year, Post-grad) and other levels of education.



Visual analysis tool output:



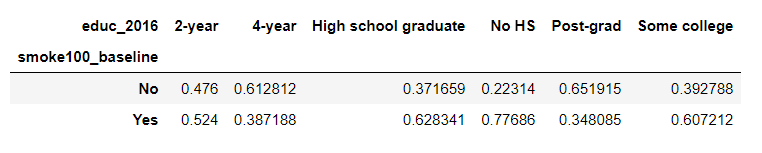




Statistical analysis:

From normalized crosstab we can clearly see percentage of those who didn’t smoke 100 cigarettes in their life is significantly higher for people with higher education:

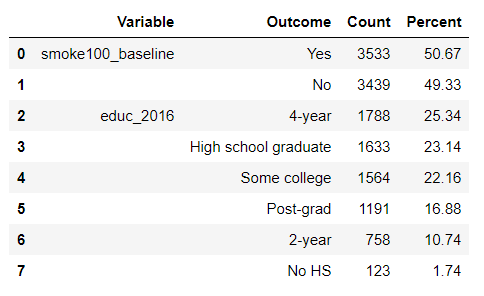
(65% and 61% for Post-grad and 4-year vs 22% for No High School and 37% for High school graduates)



Pearson Chi-square coefficient which is very high proves that there is indeed relation between these variables and differences are not caused by random distribution.

|  |  |
| --- | --- |
| Pearson Chi-square ( 5.0) = | 413.3128 |

Amount of ‘Yes’ and ‘No answers is almost equal and amount of people with different education levels is quite evenly distributed (excluding those without High School graduation).



1. Comparison exploration vs. Confirmation

For 210 found insights in data most wasn’t correlated in statistically significant way. A lot of variables didn’t show big differences when comparing their values or their “perceived correlations” were caused only by misrepresentation of values in data.