Exploratory Data Analysis

The dependent variable of interest percent\_republican and the 7 independent explanatory variables have the following summary statistics:

average\_age percent\_republican\_2012 percent\_white

Min. :18.00 Min. : 5.978 Min. : 2.50

1st Qu.:37.10 1st Qu.:50.538 1st Qu.:67.65

Median :39.90 Median :60.791 Median :86.30

Mean :39.86 Mean :59.653 Mean :78.75

3rd Qu.:42.80 3rd Qu.:70.277 3rd Qu.:94.30

Max. :62.50 Max. :95.862 Max. :99.20

percent\_degree percent\_uninsured

Min. : 3.70 Min. : 3.10

1st Qu.:13.10 1st Qu.:14.00

Median :16.80 Median :17.70

Mean :19.01 Mean :17.99

3rd Qu.:22.60 3rd Qu.:21.50

Max. :71.00 Max. :46.00

percent\_unemployed average\_income percent\_republican

Min. : 0.800 Min. : 0 Min. : 4.122

1st Qu.: 5.800 1st Qu.:22332 1st Qu.:54.960

Median : 7.500 Median :24825 Median :66.715

Mean : 7.706 Mean :25461 Mean :63.600

3rd Qu.: 9.300 3rd Qu.:27549 3rd Qu.:75.033

Max. :28.300 Max. :56674 Max. :95.273

Also of importance for an initial explanatory analysis is a list of the correlation coefficient between each independent explanatory variable and the dependent variable percent\_republican, shown below.

#average\_age 0.3325677

#percent\_republican\_2012 0.9347904

#percent\_white 0.5365550

#percent\_uninsured 0.1950089

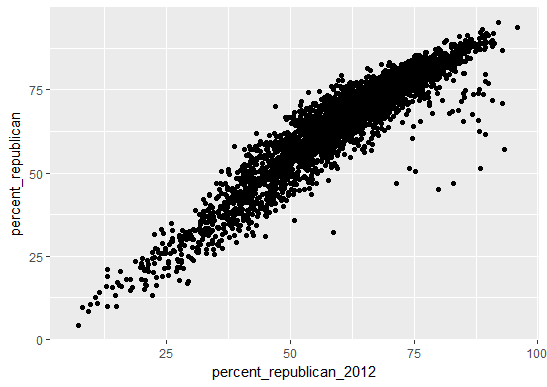
#percent\_degree -0.4936432

#percent\_unemployed -0.2728920

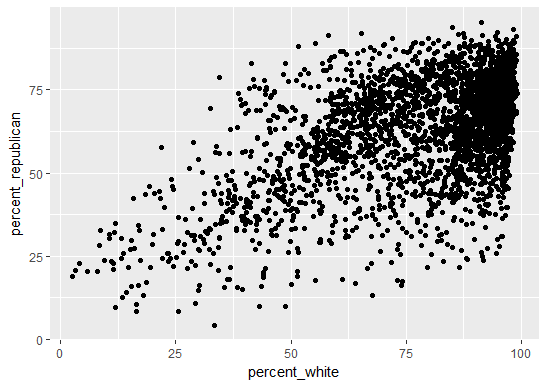
#average\_income -0.1996266

Let’s now take a closer look at our 8 varibales.

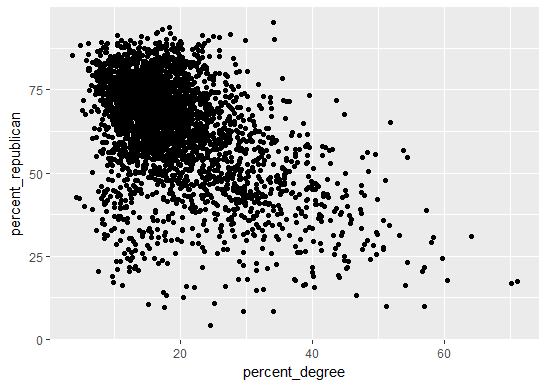
1. Percent\_republican: This variable is measuring the Republican percentage of the vote in the 2016 presidential election. It appears to be roughly normally distributed with the median and mean relatively close in value(66.7% and 63.6%) respectively. The first quintile is 55.0% and the 3rd quintile is 75.0% with a min of 4.1% and a max of 95.3%.
2. Percent\_republican 2012: This variable is the most strongly correlated with percent\_republican with a correlation coefficient of 0.93, given that party affiliation usually doesn’t change over 4 years. Like percent\_republican it is normally distributed, with the median and mean being very close in value(60.8% and 59.7%) respectively. Below is a plot showing the strong linear correlation between percent\_republican 2012 and percent\_republican.



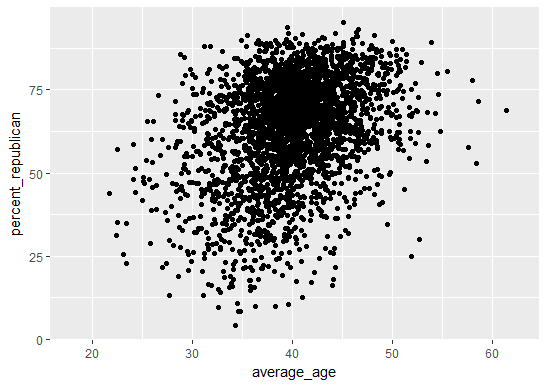
1. Percent\_white: After percent\_republican\_2012, this is the 2nd most highly correlated variable with percent\_republican with a correlation coefficient of 0.54, as seen in the scatter plot below. This variable appears to be skewed, with the median being greater than the mean (86.3% vs 78.8% respectively). In the scatter plot, it can be seen that there is a very high concentration of points on the right side with percent\_white between% 90 to 100% and percent\_republican between 50 and 75%.



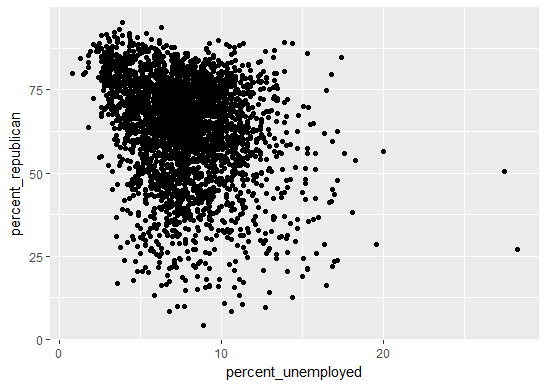
1. Percent\_degree: This variable has a moderately strong negative correlation with percent\_republican on -0.49. The scatter plot below shows this correlation. Most of the points lie in the upper left corner with percent\_degree between 10 to 20% and percent\_republican between 55% and 75%. Most of the counties where percent\_degree is 40% or more have a Republican share of the vote that is less than 50%. It appears to be approximately normally distributed, perhaps with a slight skew, with a median of 16.8% and mean of 19.0%.



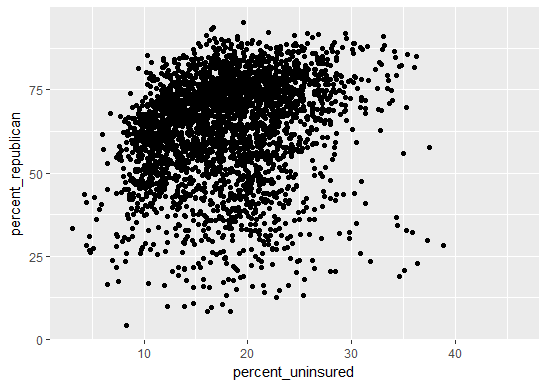
1. Average\_age: This variable is normally distributed, with median and mean being virtually the same (39.90 years and 39.86 years respectively). This variable has a small to moderate-sized positive correlation with percent\_republican; the correation coefficient is 0.33. A plot is shown below.



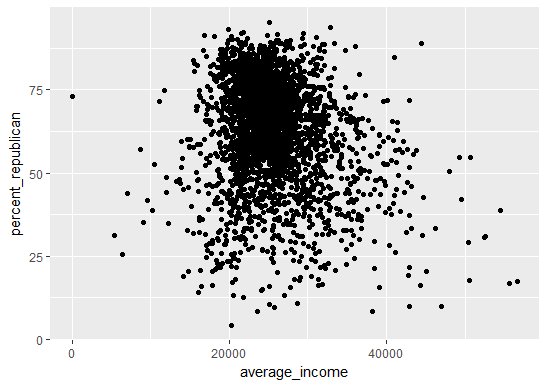
1. Percent\_unemployed: This variable is normally distributed, with median and mean being virtually the same (7.5% and 7.7% respectively). This variable has a small to moderate-sized negatve correlation with percent\_republican; the correation coefficient is -0.27. A plot is shown below.



1. Percent\_uninsured: This variable is normally distributed, with median and mean being virtually the same (17.70% and 17.99 respectively). This variable has a small positive correlation with percent\_republican; the correlation coefficient is 0.195. This would mean that the R-squared value is less than 0.04; percent\_unemployed explains less than 4% of the variability in percent\_republican, which does not indicate a good fit for a model from an initial standpoint. A plot is shown below. An inspection of the plot shows that it is hard to discern a clear positive linear correlation.



1. Average\_income: This variable is normally distributed, with median and mean both being around $25,000. This variable has a small negative correlation with percent\_republican; the correlation coefficient is approximately -0.20. This would mean that the R-squared value is labout 0.04; the individual average annual income explains less than 4% of the variability in the 2016 share of the Republican vote, which does not indicate a good fit for a model from an initial standpoint. A plot is shown below. An inspection of the plot shows that it is hard to discern a clear negative linear correlation.



Based on this initial exploratory analysis, it appears that percent\_republican\_2012, percent\_white, and percent\_degree will be good explanatory variables for our model. Machine testing will need to be done to determine if the other 4 independent variables will also be good explanatory variables. Based on this initial assessment, there is reason to suspect that percent\_unemployed and average\_income may not be good variables for our linear regression model.