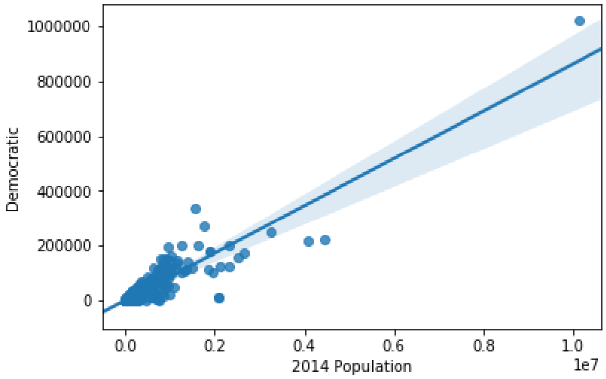
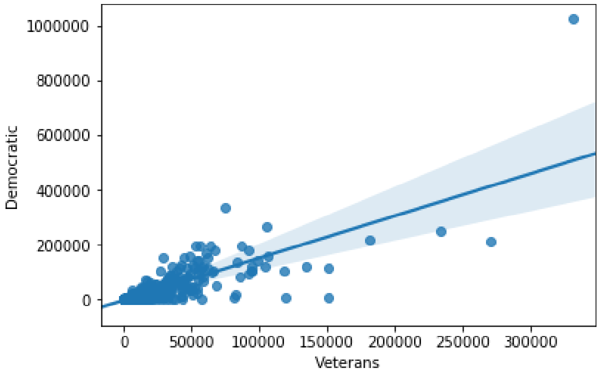
Project-2 Report

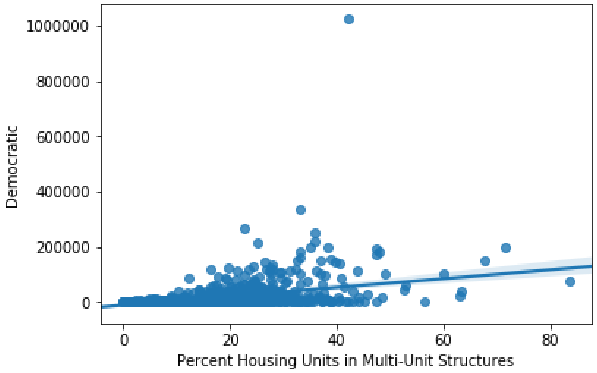
Ques 1:

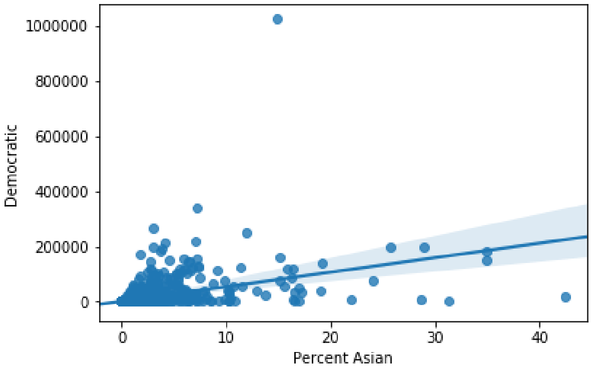
Since the data set has **less observations** we used **Cross Validation Method** to partition the dataset into 5 folds. 4-folds used for Training and 1-fold for testing.

Ques 3: Democratic Votes (Linear Regression)









Democratic Votes Vs 2014 Population: R squared Value **0.7676021019051809**	Democratic Votes Vs Veterans: R squared Value 0.6182486015453023
Democratic Votes Vs Percent Housing Units: R squared Value 0.22689551351538348	Democratic Votes Vs Percent Asians: R squared Value 0.20545255049235156

Out of the above Variables the Linear model with **“2014 Population”** as the independent variable best represents the proportion of variance of the “Democratic Votes” Variable.

Selecting Variables: Used Lasso Regression to find the attributes having high coefficients and tried each of those as independent variable.

Ques 4: Democratic Votes (Multiple Regression)

Variable Type	Variables	Adj R Squared Value
ALL Variables in demographics	'2014 Population', 'Percent Under 5 Years', 'Percent Under 18 Years', 'Percent 65 and Older', 'Percent Female', 'Percent White', 'Percent Black or African American', 'Percent American Indian and Alaska Native', 'Percent Asian', 'Percent Native Hawaiian and Other Pacific Islander', 'Percent Two or More Races', 'Percent Hispanic or Latino', 'Percent White, not Hispanic or Latino', 'Veterans', 'Percent Foreign Born', 'Percent High School or Higher', "Percent Bachelor's Degree or Higher", 'Median Household Income'	0.7603270121122873
Population and Races	'2014 Population', 'Percent White', 'Percent Black or African American', 'Percent American Indian and Alaska Native', 'Percent Asian', 'Percent Native Hawaiian and Other Pacific Islander', 'Percent Two or More Races', 'Percent Hispanic or Latino	0.7667590644612492
Combination	'2014 Population', 'Percent White', 'Veterans'	0.7705446070199521
Combination	'2014 Population', 'Percent Under 18 Years', 'Percent Female', 'Percent White', 'Veterans'	**0.7757726373543314**
Population and Gender	'2014 Population', 'Percent Female'	0.7686740692176488
Population and Degree	'2014 Population', 'Percent High School or Higher', "Percent Bachelor's Degree or Higher"	0.7673850267089972

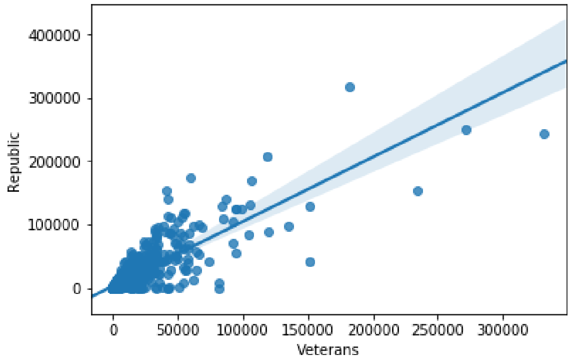
From the above table the best attributes that gave the highest adjusted R square value for the test dataset using **cross validation with 5 folds** were **'2014 Population', 'Percent Under 18 Years', 'Percent Female', 'Percent White', 'Veterans'** with **77.6%** adjusted R square value.

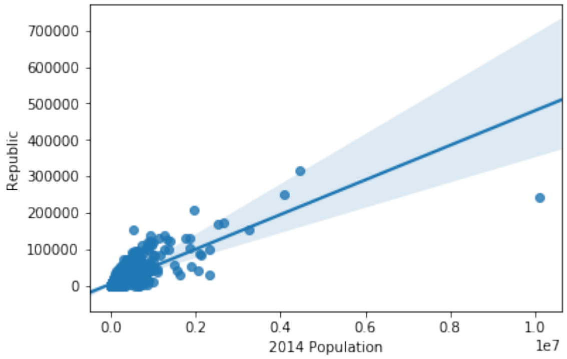
Selection of Variable:

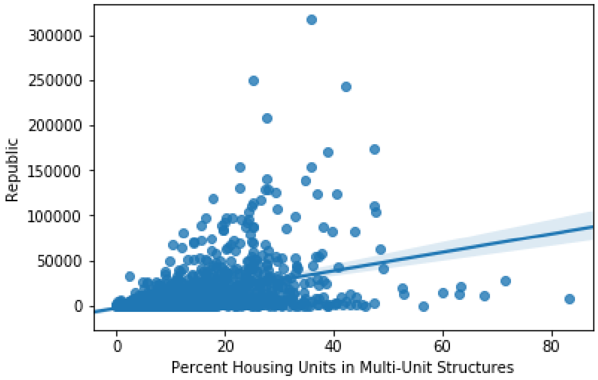
Tried multiple combinations of demographic attributes based on categories such as Race, Gender, Degree, Population etc

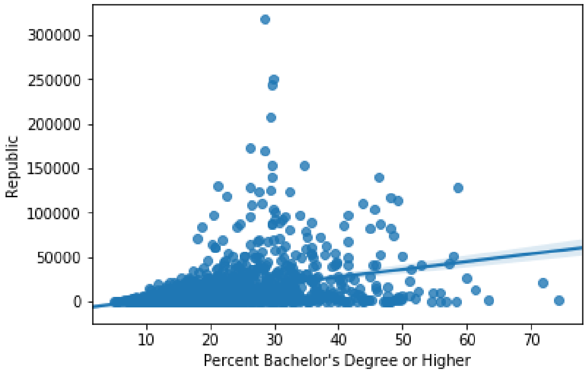
Used Lasso Regression to find the attributes having high coefficients and tried those as combination of variables.

Ques 5: Republican Votes (Linear Regression)









Republican Votes Vs Veterans: R squared Value 0.688409212246721	Republican Votes Vs 2104 Population: R squared Value 0.4419662405327318
Republican Votes Vs Percent Housing Units: R squared Value 0.1879172662191234	Republican Votes Vs Percent Bachelor’s Degree: R squared Value 0.13124021344937936

Out of the above Variables the Linear model with **“Veterans”** as the independent variable best represents the proportion of variance of the “Republican Votes” Variable.

Selecting Variables: Used Lasso Regression to find the attributes having high coefficients and tried each of those as independent variable.

Ques 6: Republican Votes (Multiple Regression)

Variable Type	Variables	Adj R Squared Value
ALL Variables in demographics	'2014 Population', 'Percent Under 5 Years', 'Percent Under 18 Years', 'Percent 65 and Older', 'Percent Female', 'Percent White', 'Percent Black or African American', 'Percent American Indian and Alaska Native', 'Percent Asian', 'Percent Native Hawaiian and Other Pacific Islander', 'Percent Two or More Races', 'Percent Hispanic or Latino', 'Percent White, not Hispanic or Latino', 'Veterans', 'Percent Foreign Born', 'Percent High School or Higher', "Percent Bachelor's Degree or Higher", 'Median Household Income'	0.5973249554223555
Population and Races	'2014 Population', 'Percent White', 'Percent Black or African American', 'Percent American Indian and Alaska Native', 'Percent Asian', 'Percent Native Hawaiian and Other Pacific Islander', 'Percent Two or More Races', 'Percent Hispanic or Latino	0.39305314200845354
Combination	'Percent Female', 'Veterans', 'Percent Two or More Races', "Percent Bachelor's Degree or Higher", 'Percent Under 18 Years'	**0.7032299624779152**
Combination	'2014 Population', 'Percent Under 18 Years', 'Percent Female', 'Percent White', 'Veterans'	0.6883803543923029
Population and Gender	'2014 Population', 'Percent Female'	0.4479955836316444
Population and Degree	'2014 Population', 'Percent High School or Higher', "Percent Bachelor's Degree or Higher"	0.4603491035109062

From the above table the best attributes that gave the highest adjusted R square value for the test dataset using **cross validation with 5 folds** were **''Veterans', 'Percent Female', 'Percent White'** with **70.33%** adjusted R square value.

Selection of Variable:

Tried multiple combinations of demographic attributes based on categories such as Race, Gender, Degree, Population etc.

Used Lasso Regression to find the attributes having high coefficients and tried those as combination of variables.

Ques 7: Classification of Party

Classifier	Attributes	Accuracy and F1-Score
KNN	'2014 Population', 'Population Percent Change', 'Percent Female', 'Percent White', 'Percent American Indian and Alaska Native', 'Percent Asian', 'Percent Native Hawaiian and Other Pacific Islander', 'Percent Hispanic or Latino', 'Percent White, not Hispanic or Latino', 'Percent Foreign Born', 'Percent High School or Higher',"Percent Bachelor's Degree or Higher",'Percent Housing Units in Multi-Unit Structures'	**Accuracy: 0.8204724409448819** **F1-Score: 0.62**
Decision Tree	'2014 Population', '2010 Population', 'Population Percent Change', 'Percent Under 5 Years', 'Percent 65 and Older', 'Percent Black or African American', 'Percent American Indian and Alaska Native', 'Percent Asian','Percent Two or More Races', 'Percent White, not Hispanic or Latino', 'Veterans', 'Percent Foreign Born', 'Housing Units', 'Percent Housing Units in Multi-Unit Structures', 'Homeownership Rate', 'Households', 'Persons per Household','Percent Living in Same House +1 Years', 'Median Household Income', 'Per Capita Income', 'Percent Below Poverty Level', 'Accommodation and Food Services Sales', 'Land Area'	Accuracy: 0.7732283464566929 F1_Score: 0.6108108108108108]
Random Forrest	'2014 Population', 'Percent Under 18 Years', 'Percent 65 and Older', 'Percent Female', 'Percent White', 'Percent Black or African American', 'Percent American Indian and Alaska Native', 'Percent Asian', 'Percent Native Hawaiian and Other Pacific Islander', 'Percent Two or More Races', 'Percent Hispanic or Latino', 'Percent White, not Hispanic or Latino', 'Veterans', 'Percent Foreign Born', 'Percent High School or Higher', "Percent Bachelor's Degree or Higher"	Accuracy: 0.8204724409448819 F1-Score: 0.5899280575539568
Gaussian Naïve Bayes	'Percent Under 18 Years', 'Percent 65 and Older', 'Percent Female', 'Percent White', 'Percent Black or African American','Veterans', 'Percent Two or More Races', 'Percent Hispanic or Latino', 'Percent White, not Hispanic or Latino', 'Veterans', 'Percent Foreign Born', "Percent Bachelor's Degree or Higher"	Accuracy: 0.8 F1-Score: 0.5916398713826366

Best Model: K-Nearest Neighbor

Performance: **Accuracy:0.8204724409448819; F1-Score:0.62**

Selection of parameters: Used Cross validation on the training data to get select the best hyperparameter i.e. the number of closest neighbors to take.

Selection and Attributes: used the best attributes selected from the seaborn plots in project1 and Used trial and error on the validation set to get the best accuracy and f1-score.

Ques 8: Clustering

Clustering Technique	**Attributes**	Supervised Metrics	Unsupervised Metrics
K-Means	'2014 Population', 'Percent Under 5 Years', 'Percent Female', 'Percent White', 'Percent Black or African American', 'Percent American Indian and Alaska Native', 'Percent Asian', 'Percent Native Hawaiian and Other Pacific Islander', 'Percent Two or More Races', 'Percent Hispanic or Latino'	**Accuracy: 0.72185430463** **F1_score: 0.4378585086**	**Adj Rand Index: 0.14159850** **Silhouette Coeff: 0.263016385**
Ward's Linkage	'2014 Population', 'Percent Under 18 Years', 'Percent 65 and Older', 'Percent Female', 'Percent White', 'Percent Black or African American', 'Percent American Indian and Alaska Native', 'Percent Asian', 'Percent Native Hawaiian and Other Pacific Islander', 'Percent Two or More Races', 'Percent Hispanic or Latino', 'Percent White, not Hispanic or Latino', 'Veterans', 'Percent Foreign Born', 'Percent High School or Higher', "Percent Bachelor's Degree or Higher", 'Median Household Income', 'Percent Housing Units in Multi-Unit Structures'	**Accuracy:** 0.70813623 **F1_score:** 0.3836163	**Adj Rand Index:** 0.141598 **Silhouette Coeff:** 0.263016

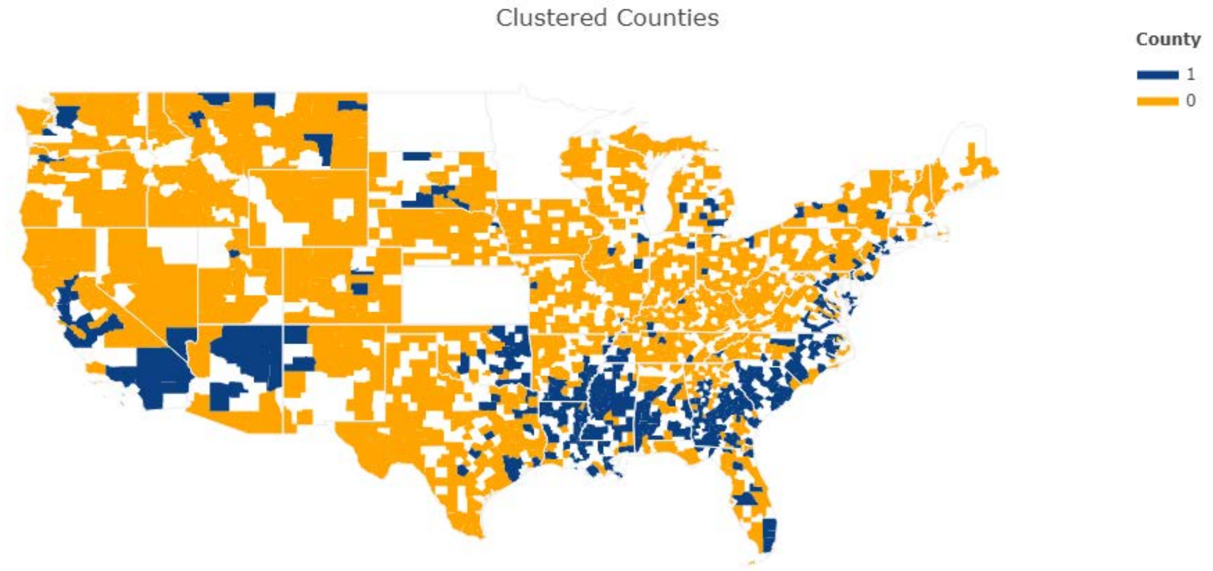
Best Model: K-Means

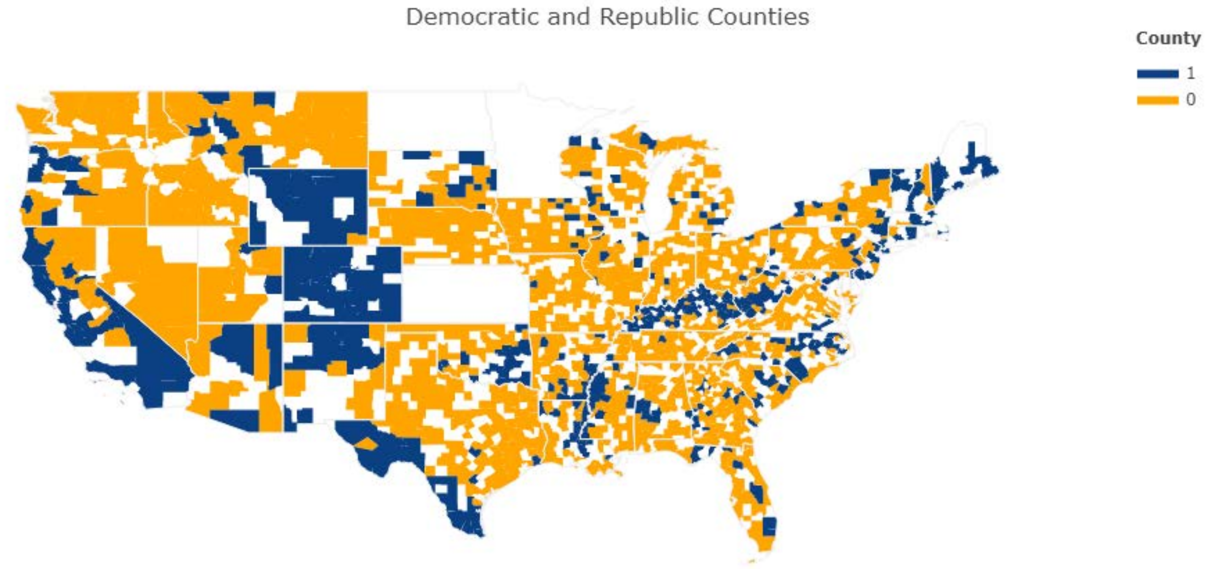
Performance: Accuracy: 0.721854304 **F1_score:** 0.437858508
Adj Rand Index: 0.14159850 **Silhouette Coeff**: 0.2630163

Selection of parameters: Used Cross validation on the training data to get select the best hyperparameter i.e. select best method of initialization

Selection and Attributes: used the best attributes selected from the seaborn plots in project1 and Used trial and error on the validation set to get the best accuracy and f1-score.

Ques 9: Plotting Clustered Counties on US Map





The counties in the clustered map are more clustered together in a region based on the party of the counties closest to it.