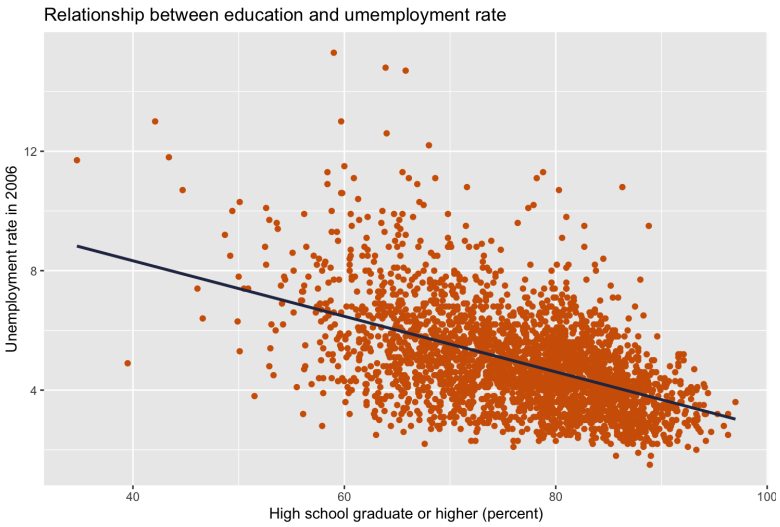
Data Science for Business Team Case 3
Team 18

Yuhe(Tiffany) Jin, Yinan Chen, Tirth Pravin Gala,
Chenjie (Angelina) Sun, Caryl Alexis Cohen

Question 1

- **Why**
  Education level plays a critical role in influencing economic development. Specifically, how does the percent of the high school graduates or higher affect the unemployment rate, and if so, what might explain this correlation?
- **What**
  From this scatterplot, we can see there's a significantly negative correlation between unemployment rate in 2018 and the percent of high school graduates or higher. The data are closely clustered around the regression line with only few outliers in the left side. That is to say, the unemployment raet tends to decrease as more people graduate from high school or higher education.



- **How**
  Individuals with higher levels of education tend to possess more extensive skills and have access to more diverse employment opportunities, thereby increasing their attractiveness and likelihood of securing jobs.

Question 2

1. **Predictive Task**

   Our core task is to build a model to predict the winning spread of Obama over Clinton measured as percentage of the total vote *(Obama_margin_percent=100*(Obama-Clinton)/TotalVote)* given the observed demographic features of the voters.

2. **Data Preparation**

   The data is split into two separate data sets, one is the training set with primaries and caucuses before February 19, 2008 and the other is the testing set after February 19, 2008. The variables *County, State, FIPS,TotalVote, Clinton, Obama, Obama_margin, Obama_wins* are excluded in the training set and *County, State, FIPS,TotalVote, Clinton, Obama* are excluded in the testing set.

3. **Potential Models**

   We run 1) linear model 2) random forest model 3) CART model to fit into the data set and compare them with each other to find the final model for prediction.

4. **Evaluation Metric**

   The metric we use to evaluate model performance is RMSE. Since a lower RMSE indicates that the model has a closer prediction to the actual values and fits the data better, we'd like to choose the model with the lowest RMSE to be our final model. The MSE for 3 models are shown below:

```
           Model      RMSE
1         Linear 16.212592
2 Random Forest  6.226764
3           CART 18.544388
```

5. **Final Model**

   Based on MSE for 3 models, we choose Random Forest Model as our final model since it has a significantly lower MSE compared with others.

6. **5-fold Cross Validation**

   We apply and report a 5-fold cross validation to evaluate the performance of Random Forest Model as our chosen model.

```
Random Forest

1737 samples
  35 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1389, 1389, 1390, 1390, 1390
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   2    17.97474  0.7093402  14.33445
  19    15.01851  0.7640204  11.57643
  37    14.78869  0.7680766  11.32888

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 37.
```

7. **Prediction**

   Based on our final model, we predict the winning spread percentages for the test sample with the R code below.

```
### Final Model
################
rf_model <- randomForest(Obama_margin_percent ~ ., election_data_train)
final_model <- predict(rf_model, election_data_test)
print(final_model)
```

Question 3

In order to explore the data, we apply kmeans as one unsupervised learning tool to *election_data*. For the clustered data, we select the following demographic variables from the original dataset and scale them to better fit into kmeans:

"AgeBelow35", "Age35to65", "Age65andAbove", "White", "Black","Asian","AmericanIndian","Hawaiian","Hispanic", "HighSchool","Bachelors", "Poverty","IncomeAbove75K", "MedianIncome", "MedicareRate", "UnemployRate", "SocialSecurityRate","DisabilitiesRate","PopDensity"

In the scatterplot, each color represents a different cluster generated by the kmeans algorithm. We can observe that the clusters are well-separated with distinct boundaries between the clusters. There are still some points overlap but not too much. Those well-separated clusters could suggest that the similar data points are effectively grouped together under kmeans.

Next, we take a look into the average of "MedianIncome", that is the median household income in 2005, and "MedicareRate", that is the medicare program enrollment in 2005 (rate per 100k persons), within each cluster as shown below:

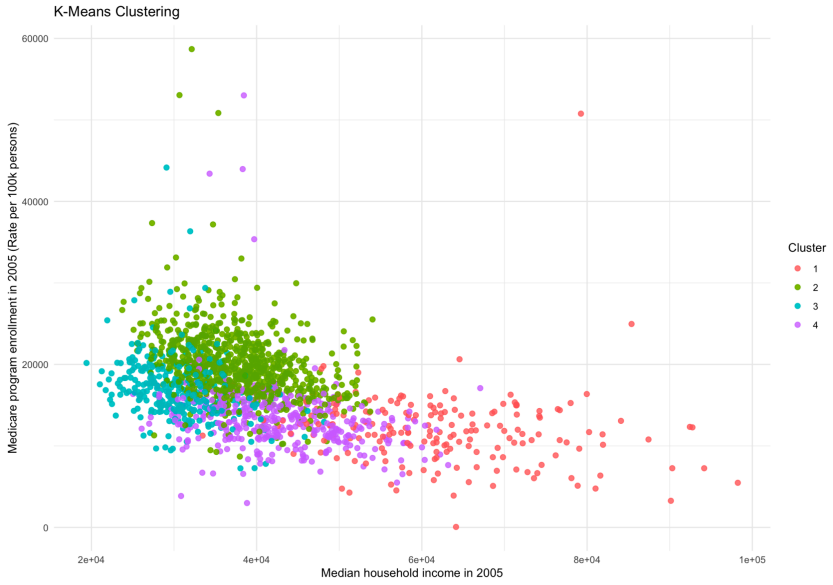
	`k_means$cluster`	Avg_MedianIncome	Avg_MedicareRate
	<int>	<dbl>	<dbl>
1	1	62303.	11874.
2	2	37577.	19974.
3	3	30870.	17018.
4	4	42406.	13365.

Specifically, Cluster 1 appears to represent higher-income households with lower Medicare enrollment, Cluster 2 represents lower-income households with higher Medicare enrollment, Cluster 3 represents lower-income households with moderate Medicare enrollment, and Cluster 4 represents moderate-income households with moderate Medicare enrollment.

That is to say, the population with higher income may have access to better private healthcare options or have higher incomes to cover medical expenses, therefore they don't need to rely on the Medicare program. However, for these households with lower income, there may be a higher

proportion of elderly or retired individuals with relatively higher reliance on the enrollment of the program. Those moderate-income households have a mix of income levels and more middle-class households, so the Medicare program enrollment remains a moderate level.

Overall, with the support of kmeans algorithm, we get a deep insight into the correlation within the *election_data*. These insights may help policymakers or related organizations to better understand the demographics and healthcare patterns within different segments of the population.



Question 4

First: Holding other conditions constant, if the Hispanic demographic has been 5% larger, Obama is 98% likely to have 0.6% fewer total voters over Clinton. If the Black demographic has been 5% larger, Obama is 99% likely to have 4.3% more total voters over Clinton.

Holding other conditions constant, if the Hispanic population increases by 5%, percentage of Obama's voters will be 0.6% points lesser. This can be said with low confidence as p-value (0.098) is more than 0.05. Similar logic of interpretation says if the Black population increases by 5%, Obama's voters will be 4.3% points more. This can be said with 99% confidence as p-value (2e^-16) is less than 0.05.

```
Call:
glm(formula = Obama_margin_percent ~ Hispanic, data = election_data_train)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.98261    0.85734  -3.479 0.000516 ***
Hispanic    -0.11511    0.06973  -1.651 0.098940 .
```

```
Call:
glm(formula = Obama_margin_percent ~ Black, data = election_data_train)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.28720    0.80164  -16.57   <2e-16 ***
Black         0.86311    0.04204   20.53   <2e-16 ***
```

Second: When we run regression considering the 5% increase in :

1. **Hispanic Population**: As the p-value (0.098) i.e., higher that 0.05, there is no further analysis which can give a clear conclusion of Hispanic population affecting Obama voters change due to change in Hispanic population. Further analysis as per counties will be insignificant.
2. **Black Population**:
    a. **Inferential Output**:

	Estimate	Std. Error	t value	Pr(>|t|)	
### xCountyCharles	-8.738e+01	1.822e+01	-4.795	2.11e-06	***
### xCountyBaltimore city	-7.619e+01	1.885e+01	-4.042	6.07e-05	***
### xCountyAroostook	-6.357e+01	1.890e+01	-3.364	0.000823	***

    b. **Inference**: Keeping other things constant, these three counties led to reduction as an effect on Obama's spread of voters compared to Clinton when 1% of Black population has increased.

    For 5% population increase, in County Charles Obama's spread was reduced by -4.37% percentage point, in County Baltimore by -3.81 and in County Aroostook by -3.18.

Question 5

We would highly recommend Candidate Obama to allocate resources as per statistical inferences. This means that if in a particular county, he is not having majority voters, he should run campaigns in those areas. Accordingly, a particular population, like Hispanic or Black population change, leads to decrease in percentage of votes of Obama. He should focus his attention on catering to those ethnic groups. To give one example, Question 4 says that 5% increase in Black population increases 4.3% point of Obama's voters compared to Clinton voters. Thus, in such a situation, Obama's team should focus their resources around other ethnic group more or bank more on this population depending on their strategy.