CSS 429 Data Analysis Systems

Case 2

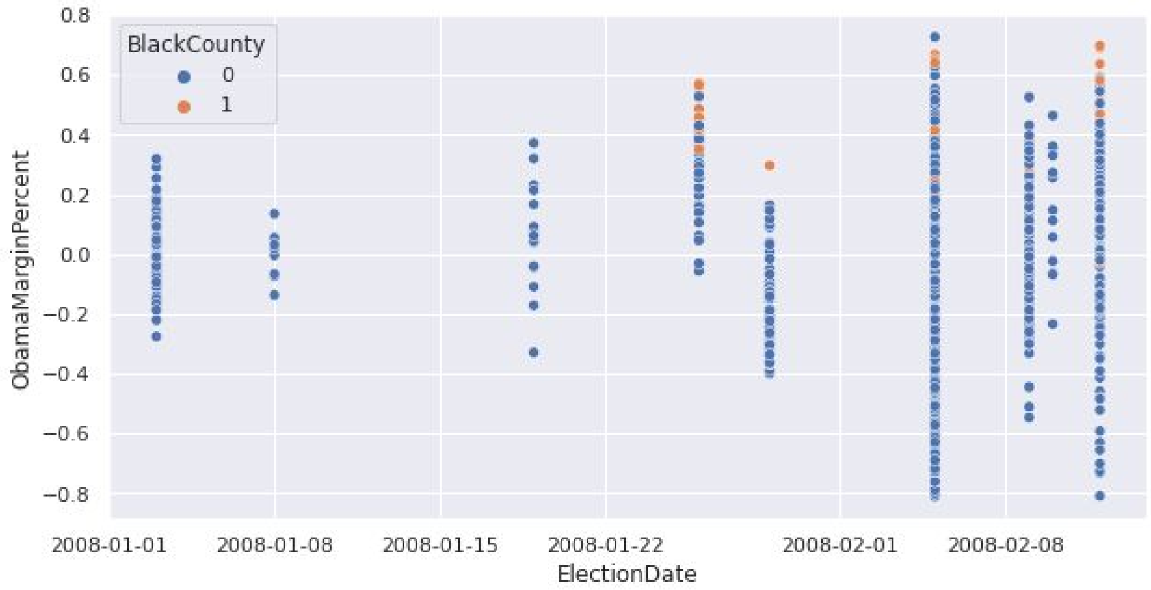
Team 7 - Abraca-data:

- Muratbekov Maksatbek
- Sarkulov Yerden
- Tolbassy Bakdaulet
- Yegemberdi Nurgeldi
- Zhakypov Aslan

Answers

1. **Pick two (or more) variables and attempt to show a relation between them via visualization. As discussed before, this requires one to formulate a question, and to communicate clearly a conclusion based on data visualization (specify the why, what, how). (Note that in this question it is not required that the relationship displayed relates to the election.)**

   Inasmuch Obama was black, we were interested in how people voted in counties where black people are more than white people and did it affect Obama's win and the election as a whole. As expected almost all of them voted for Obama. But we can't say that Obama won the election because of these black people. Because there aren't a lot of black people in the US and in their counties which didn't make a huge difference to the election generally. Although it is insignificant but we are convinced that the race of a person also affects. But this is all about black people. In those years racism still played a role after all. As you can see on this visualization, orange dots represent counties where there are more black people than whites. Blue dots, otherwise.



Additionally, we were interested in correlation of exhibits with Obama's Margin Percent and were aware about the importance of race of civils in the USA. Correlation rate was the biggest (0.44) between ***Black*** *and Obama's Margin Percent, while it showed the least (-0.41) with* ***White*** *column.*

2. Provide a model to predict the winning spread of Obama over Clinton measured as percentage of the total vote.

a. Describe clearly the core task, briefly discuss all the models you compared

We created different models changing their exhibits. Model 1 uses 31 columns, while Model 2 and Model 3 use only 5 variables. Model 4 and Model 5 are trained with data which have only 3 columns. We had a look at the correlation table and chose top rated variables. However, if there is more data in your model, in many cases it works better than others. So, we prefer to use Model 1 which is the best one among our models.

b. state which metric is being used to evaluate performance, and how did you choose a final model.

To evaluate the performance of models used R-Squared, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

We chose our model because it's performance was better.

Evaluation the models against test data using MAE, RMSE and R2:

```
Model 1:
LinearRegression
    MAE 0.1521087932625633
    RMSE 0.1980698179330634
    R2 0.5522590196973303
KNeighborsRegressor
    MAE 0.2321398180793533
    RMSE 0.29559421702980815
    R2 0.0028003723196678987
```

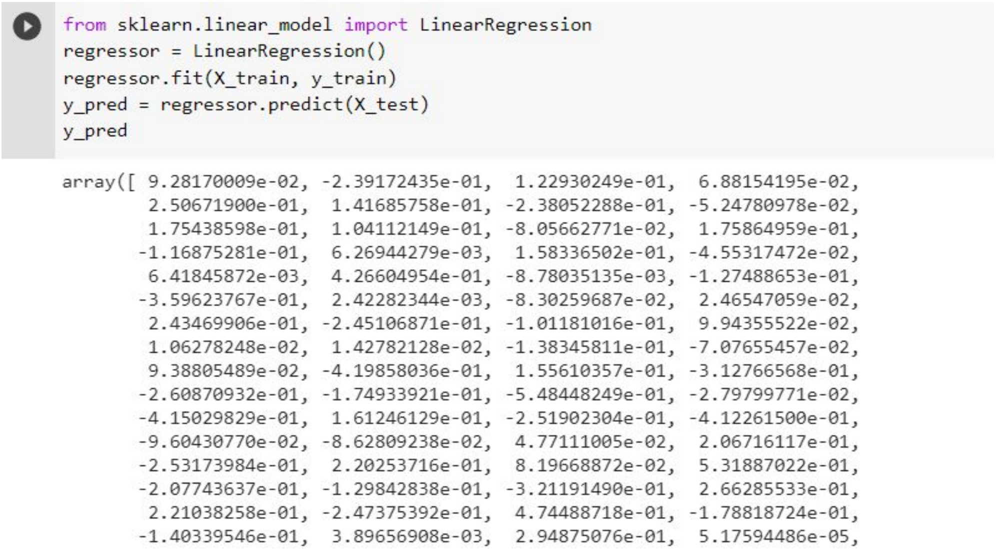
c. Apply and report a K-fold cross validation to evaluate the performance of your chosen model.

Case 2

Our Model 1 was applied by K-fold cross validation taking the ***n_split = 2*** *and evaluated, then it showed this R-Squared evaluation:*

```
LinearRegression [0.47917911 0.53035172]
KNRegressor [0.10005884 0.04327738]
```

d. **Based on your final model, predict the winning spread percentages for the test sample (provide the Python code that generates your predictions).**



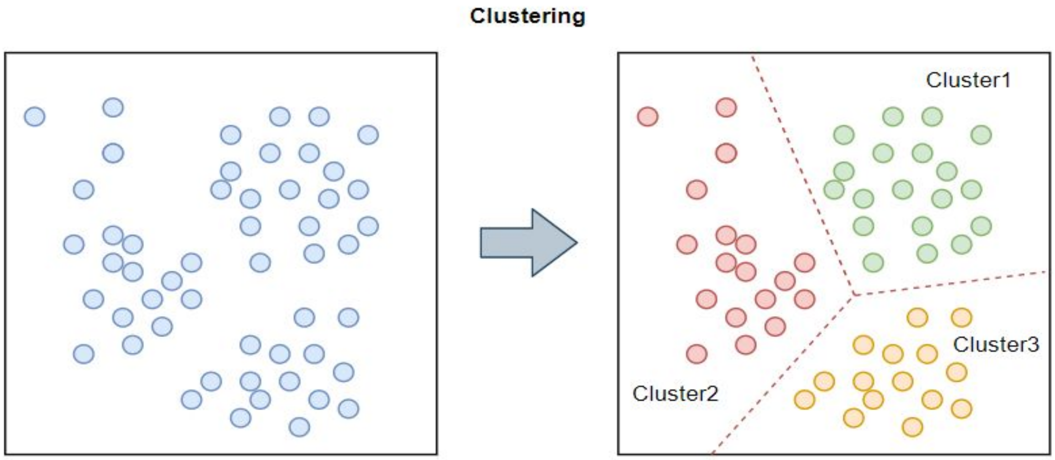
```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)
y_pred
```

```
array([ 9.28170009e-02, -2.39172435e-01,  1.22930249e-01,  6.88154195e-02,
        2.50671900e-01,  1.41685758e-01, -2.38052288e-01, -5.24780978e-02,
        1.75438598e-01,  1.04112149e-01, -8.05662771e-02,  1.75864959e-01,
       -1.16875281e-01,  6.26944279e-03,  1.58336502e-01, -4.55317472e-02,
        6.41845872e-03,  4.26604954e-01, -8.78035135e-03, -1.27488653e-01,
       -3.59623767e-01,  2.42282344e-03, -8.30259687e-02,  2.46547059e-02,
        2.43469906e-01, -2.45106871e-01, -1.01181016e-01,  9.94355522e-02,
        1.06278248e-02,  1.42782128e-02, -1.38345811e-01, -7.07655457e-02,
        9.38805489e-02, -4.19858036e-01,  1.55610357e-01, -3.12766568e-01,
       -2.60870932e-01, -1.74933921e-01, -5.48448249e-01, -2.79799771e-02,
       -4.15029829e-01,  1.61246129e-01, -2.51902304e-01, -4.12261500e-01,
       -9.60430770e-02, -8.62809238e-02,  4.77111005e-02,  2.06716117e-01,
       -2.53173984e-01,  2.20253716e-01,  8.19668872e-02,  5.31887022e-01,
       -2.07743637e-01, -1.29842838e-01, -3.21191490e-01,  2.66285533e-01,
        2.21038258e-01, -2.47375392e-01,  4.74488718e-01, -1.78818724e-01,
       -1.40339546e-01,  3.89656908e-03,  2.94875076e-01,  5.17594486e-05,
```

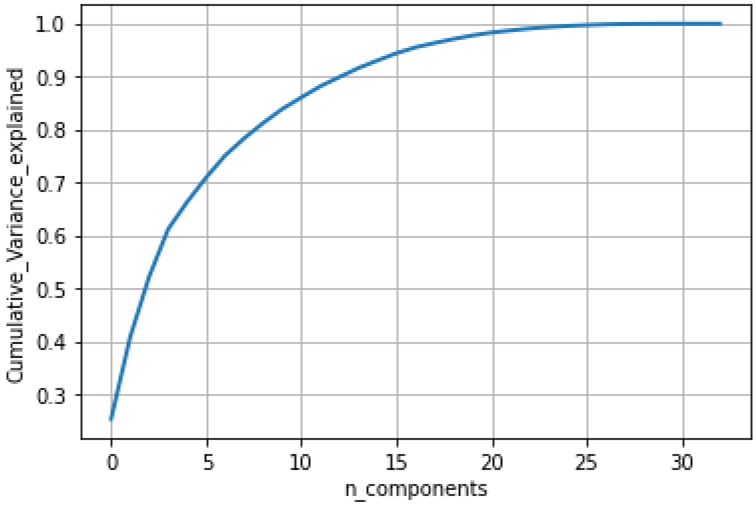
As you see y_pred is predicted Obama's margin percent over Hillary Clinton's.

3. **In order to explore the data, apply one unsupervised learning tool (e.g., k-means, principal component analysis), interpret and communicate briefly the output (e.g., clusters, latent features), and attempt to obtain insights.** *We decided to apply unsupervised learning tool to column called ‘Regions’, as we need to group them and add to the clusters, each region (Midwest, Northeast,*

West, South) will have own color and specific placement



PCA. The number of components in PCA can be decided after checking variance with respect to no_components*. Let's see some of common ways of checking the components.*

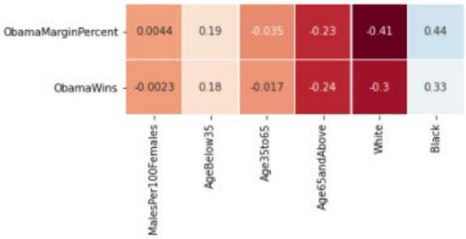


```
When n_components = 0,   variance=25.430000
When n_components = 1,   variance=40.870000
When n_components = 2,   variance=52.290000
When n_components = 3,   variance=61.260000
When n_components = 4,   variance=66.390000
When n_components = 5,   variance=71.010000
When n_components = 6,   variance=75.160000
When n_components = 7,   variance=78.400000
When n_components = 8,   variance=81.290000
When n_components = 9,   variance=83.920000
When n_components = 10,   variance=86.080000
When n_components = 11,   variance=88.200000
When n_components = 12,   variance=89.930000
When n_components = 13,   variance=91.630000
When n_components = 14,   variance=93.060000
When n_components = 15,   variance=94.420000
When n_components = 16,   variance=95.540000
When n_components = 17,   variance=96.350000
When n_components = 18,   variance=97.100000
When n_components = 19,   variance=97.780000
When n_components = 20,   variance=98.330000
When n_components = 21,   variance=98.710000
When n_components = 22,   variance=99.080000
When n_components = 23,   variance=99.370000
When n_components = 24,   variance=99.570000
When n_components = 25,   variance=99.730000
When n_components = 26,   variance=99.870000
When n_components = 27,   variance=99.940000
When n_components = 28,   variance=99.980000
When n_components = 29,   variance=99.990000
When n_components = 30,   variance=100.000000
When n_components = 31,   variance=100.000000
When n_components = 32,   variance=100.000000

Number of components to explain 90% Variance is 14
Components =  14 ;
Total explained variance =  0.91631
```

4. **Choose one candidate. What kind of advice (based on data analytics) would you provide to your candidate? For example, which voter segment to target with their campaign messages and why? Or, how to allocate resources (budget and volunteer time) across regions and why? How would you communicate such insights?**

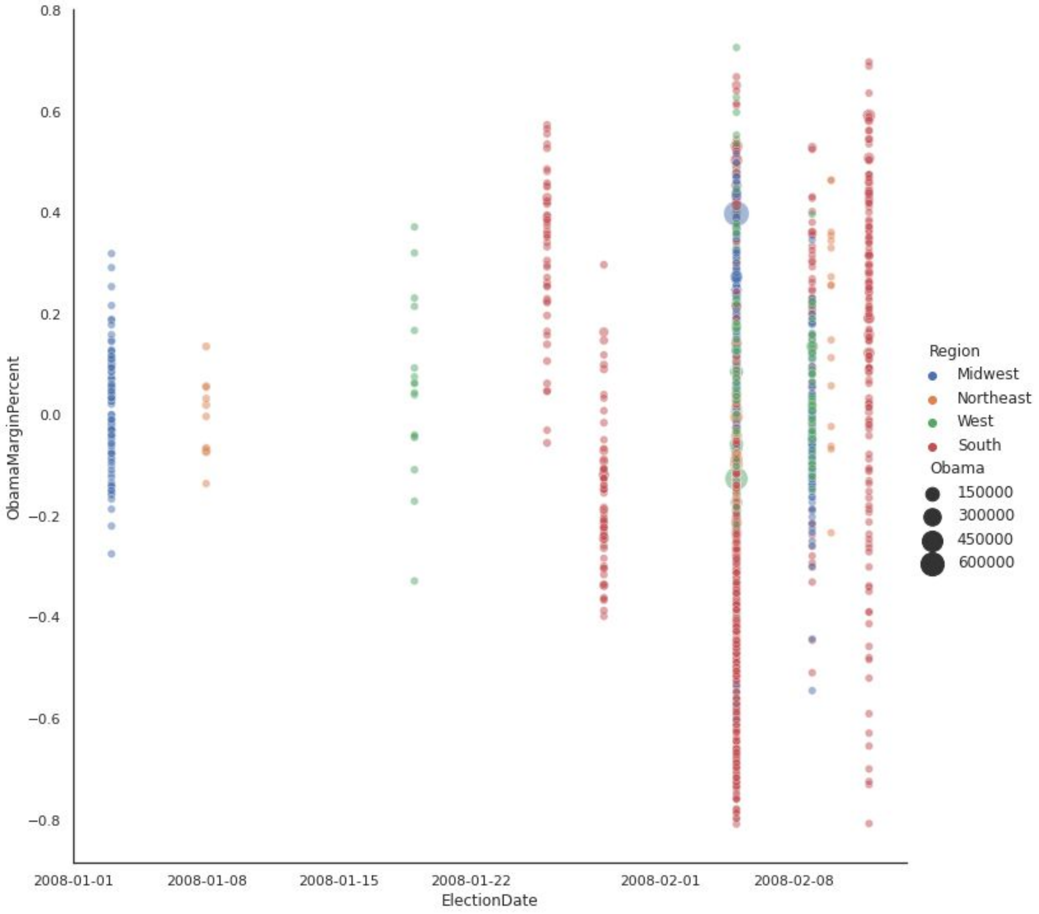
   We chose Hillary Clinton as she lost the @POTUS.
   The first thing, as we said, It is necessary to prevent racial divisions among citizens. That is, the fact that blacks did not vote for Clinton was her first defeat.



(Figure 1)

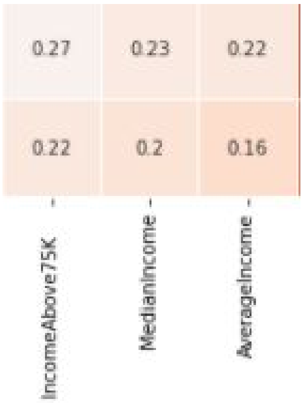
Case 2

How does Obama's Margin Percent relate with White column? It would not be an exaggeration for Clinton to talk about supporting blacks.
Secondly, Clinton needs to focus on resource allocation in the southern regions, which are overcrowded.



(Figure 2)

In addition, it would be not surprising that Clinton received a large number of votes if she worked hard in the social sphere.



IncomeAbove75K	MedianIncome	AverageIncome
0.27	0.23	0.22
0.22	0.2	0.16

(Figure 3) part of table shown in Figure 2.