CS 696 Intro to Big Data
Fall Semester, 2016
Assignment 3

San Diego State University -- This page last updated 10/5/16

Due Oct 20 23:59

The assignment is to be turned in as a Jupyter notebook, using either Julia 0.4.6 or Julia 0.5. To answer the questions below you will write some Julia code. The Julia code is to be placed in code cells, which will be run when your assignment is graded. In answering the questions you may at times need to write some text. The text should appear in markdown cells, not in code cells. Your notebook will consist of markdown cells interspersed with code cells forming a narrative. Some questions like “How many voters are in each country?” may have a numerical answer, but to get that answer some calculations are required. You are to show those calculations in code cells so your results can be repeated.

The assignment uses several data files which your code needs to read. The location of those files will be different on your machine than on mine. You need to define variables that contain the location of the data files like:

uk_election_file = “c:\\foo\bar\UK2010.csv”

Put all variables referencing files you read/write near the beginning of the notebook. If you do this for every file you read or create then it will make it easy to modify your notebook to run on other machines.

You can either put both parts of the assignment into one notebook or put each part into separate notebooks.

Part one. Election Data.

The goal is to detect election fraud. Download the **Election** data on the course **Data Sets** page. There two countries involved - United Kingdom and Russia. After unziping the Election Data file you will find three csv files: UK2010.csv, Russia2011_1of2.csv and Russia2011_2of2.csv.

About the data

The data contains national election results in elections held in in two countries.

UK2010.csv

The file contains the data for the 2010 election in the UK. The election was for members of parliament. The party that received the majority of seat in parliament form the government. If no party receives a majority in parliament then a coalition government is formed. The first six columns are:

Press Association Reference: A unique number for each voting district.
Constituency Name: The common name of the voting district.
Region: The geographic region where the district is located.
Election Year: The year when the election was held
Electorate: The total number of people eligible to vote in the district
Votes: The total number of votes cast.

The remaining columns show how many votes each political party received in each voting district.

Russia2011_1of2.csv
Russia2011_2of2.csv

These files contain data for the 2011 election in Russia which were for members of the Duma. The second file is a continuation of the first file. Only the first file contains column headers explaining the contents of each column. The first 21 column headers are listed below and are fairly self-explanatory. The remaining seven columns are the votes each political party received in each district.

Code for district
Number of the polling district (unique to state, not overall)
Name of district
Number of voters included in voters list
The number of ballots received by the precinct election commission
The number of ballots issued to voters who voted early
The number of ballots issued to voters at the polling
The number of ballots issued to voters outside the polling station
The number of canceled ballots
The number of ballots in mobile ballot boxes
The number of ballots in the stationary ballot boxes
Number of invalid ballots
Number of valid ballots
The number of absentee ballots received by the precinct election commission
The number of absentee ballots issued to voters at a polling station
The number of voters who voted with absentee ballots at a polling station
The number of the unused absentee ballots
The number of absentee ballots issued to voters of the territorial election commission
Number of lost absentee ballots
The number of lost ballots
The number of ballots not recorded after being obtained

Questions & Tasks

1. Data Cleaning (5 points). The first task is to read the data into dataframes and perform any needed data cleaning. Did you need to perform any data cleaning to be able to answer the simple queries given in #2?

2. Simple queries (12 points). To get a feel for the data answer the following questions.

    a. How many voters are in each country?

    b. How many votes were cast in each country?

    c. Which party received the most votes in each country?

    d. What is the mean and standard deviation of the number of voters in each district in each country?

    e. Using Gadfly produce a histogram of number of voters in each district for each country. What sort of differences or similarities are there between the two countries in this regard.

    f. What is the mean and standard deviation of the number of votes cast in each district in each country?

3. Sanity Checks (5 points). One would expect that published election data would be cleaned up but still one should check. Do the number of votes received by all parties in a district equal the number of votes cast? If not how many districts in each country do the number of votes cast do not equal the votes received.

4. Investigating the election results (28 points).

    a. For each district compute the turnout rate. That is the number of votes divided by the number of voters in each district. Using Gadfly produce a histogram of the turnout rate in each country per district. One would think that the results should approximate a normal distribution. Does it in each country? A high turnout rate could mean high interest in particular districts or it could represent ballot stuffing.

    b. In the UK election no party won a majority in Parliament. As a result the Conservatives (Con) and the Liberal Democrats (LD) formed a coalition government. Compute the total number of votes in each district won by the coalition.

    c. The party with the most vote in Russia won the election. In the UK the coalition won the election. Here we are only concerned about the votes in each district that the winners of the election received. If there was election fraud or pressure on voters to vote for a given party one might see a high percentage of votes for the winners in some districts. In particular some claim that large number of districts with near 100% voter turn out near 100% votes for the winner indicates election fraud. For each country produce a scatterplot of votes obtained by the winners in each district and the turnout rate in each district. Do you see a concentration of points near the 100% turnout rate and 100% votes for the winner?

    d. Given the number of districts in Russia the plot in c is likely to contain a large mass of points which can hide information. It is not clear if the the density of points varies in the plot. One way to handle this is to set a low alpha for the color of each point so you will see a dark areas were there are more points clustered. However this is not easy in Gadfly. One way to handle this is to select a random sample of the data. You can use **rand** or **sample** function from StatsBase.jl package. Any cluster of points should still

occur in a random sample of the data. Do the scatter plot done in c with a random sample of districts in Russia. Is it possible to select a sample size that shows several clusters of points in the plot rather than one massive blob of points?

Part Two. GRE

The goal of part two is to determine the relationship between a students GRE scores and their GPA. Download the **GRE** data on the course **Data Sets** page.

About the data

gpa-gre.csv

The file contains four columns. Each row represents a single student. The columns are:

Year - The year the student graduated. Values from 1 to 16.
GPA - The GPA of the student when they graduated.
Verbal - The verbal GRE score obtained by the student. As this is historical data the scores are in the old scale from 200-700.
Quant - The quantitative GRE score obtained by the student. Again in the old scale.

Questions & Tasks

1. Exploring the data (10 points). To visualize the data produce boxplots of the GPA earned by year. Is there any obvious trend? Are the means of the grades increasing or decreasing over time?

2. Relationship between GRE & GPA (35 points).

   a. To determine if there is a relationship between GRE and GPA compute Pearson's correlation value (r) between the GPA and Quantitative GRE scores, between the GPA and verbal scores, and between total GPA (quantitative + verbal) and GPA. Show each value. Which has a stronger relationship with the GPA?

   b. Using the quantitive GRE score as the independent variable and the GPA as the dependent variable compute the regression model. Show the estimates of the coefficients with the standard Error, t value and Pr(>|t|).

   c. Plot quantitive GRE score verses GPA data points with the regression line.

   d. Plot the residuals of the GRE score verses GPA data. Is there anything unusual about the plot to indicate the residuals are not random?

   e. What is the coefficient of determination?

   f. Using the quantitive GRE and verbal GRE scores as the independent variables and the GPA as the dependent variables compute the regression model. Show the estimates of the coefficients?

   g. Using the adjusted R^2 how much of the variation in GPA is explained by the two GRE scores?

What to turn in

Turn in one zip file containing your notebook(s) file.

Late Penalty

An assignment turned in 1-7 days late, will lose 5% of the total value of the assignment per day late. The eight day late the penalty will be 40% of the assignment, the ninth day late the penalty will be 60%, after the ninth day late the penalty will be 90%. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.