Data Mining Fall 2017

Project 1: Data and Visualization

Assigned: 8/31/2017
Due: 9/24/2017 (via Canvas)
Points: 100

Please submit your report in **PDF format.**

Mark Wilson / Getty Images

BuzzFeedNEWS

We're Sharing A Vast Trove Of Federal Payroll Records

The data, obtained by BuzzFeed News via the Freedom of Information Act, covers more than 40 years of U.S. government employment. You can download and start exploring it today.
The dataset contains hundreds of millions of rows and stretches all the way back to 1973. It provides salary, title, and demographic details about millions of U.S. government employees, as well as their migrations into, out of, and through the federal bureaucracy. In many cases, the data also contains employees' names...

Posted on May 24, 2017, at 1:25 p.m. by Jeremy Singer-Vine

URL: https://www.buzzfeed.com/jsvine/sharing-hundreds-of-millions-of-federal-payroll-records?utm_term=.jbk33NbAx#.uwVQQ8Vad

We are interested in how federal, non-military employment was impacted by the transition from president George Bush to president Barack Obama to learn about the impact of the difference in policy between the two administrations on the federal government. What agencies increased/decreased staffing size? What agencies had more or less turnover? Do those changes align with stated policy goals? Is agency staffing impacted by events?

Write a report covering in detail all steps of the project. The results have to be reproducible using your report. Carefully describe every assumption and every step in your report. Also, mention any program/code/additional data that you are using for your analysis.

Follow the CRISP-DM framework

1. Business Understanding [10]

- Investigate what the major policy differences between the two administrations were. How would these changes be reflected in federal employment? Were there major incidents during the two administrations that could be the result of or trigger changes in employment? [10 point]

2. Data Understanding [80]

- Describe the meaning and type of data (scale of measurement, values, etc.) for the variables in the data file(s). [10 point]
- Verify the data quality. Are there missing values? Duplicate data? Outliers? Are those mistakes? How do you deal with these problems? [10 Points]
- Give simple appropriate statistics (e.g., range, mode, mean, median, variance, counts) for each variable and describe what they mean, especially, if you found something interesting. **Note:** You can also use data from other sources for comparison. [10 points]
- Visualize the most important variables appropriately (at least 5 attributes). **Important:** Provide an interpretation for each chart and explain for each variable why you chose the used visualization. Charts without explanation are useless! [20 points]
- Explore relationships between variables with appropriate methods (a minimum of 5 relationships). Use, for example, scatter plots, correlation, boxplots, cross-tabulation, group-wise averages. [25 points]
- The data contains data for over 40 years and thus has a temporal component. What could you do by analyzing changes over time? Can you do some if this (exceptional work)? [5 points]

Exceptional Work [10 points]