

Project 1: Data and Visualization

Assigned: 1/28/2016
Due: 2/17/2016 (via Canvas)
Points: 100

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



Source: WFAA.com

The Dallas Morning News

Violent crime up in Dallas so far in 2015, police report

For the first time in more than a year, police say violent crime is on the rise in Dallas. Police are shifting tactics to try to combat rising robberies, which are driving the violent crime spike, Chief David Brown said at a City Council public safety committee meeting Monday. Despite the current increase, the chief pledged to try and reduce crime for a 12th year in a row this year.

Dallas Morning News, March 23, 2015

You will work for this project with crime report data for Dallas published by Dallas Open Data.

<https://www.dallasopendata.com/dataset/Dallas-Police-Public-Data-RMS-Incidents-With-GeoLo/4ea4-q4ui>

The data contains Dallas Police Department (DPD) incidents for part of 2014 and 2015. We are interested in any knowledge that can be learned from this data (and any additional data that you choose to use) which can help the police department to fight crime more effectively.

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]

- Define the purpose of using data analytics/mining to achieve possible goals of the DPD, local politicians and the general public? How would you define and measure effectiveness of this data mining project? What data would be needed to judge the effectiveness? [10 point]

2. Data Understanding [80]

- Describe the meaning and type of data (scale, 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, cross-tabulation, group-wise averages. [25 points]
- The data contains locations and thus is spatial data. What could you do with this spatial data? Can you do this (exceptional work)? [5 points]

Exceptional Work [10 points]