CST 383: Intro to Data Science

Dr. Glenn Bruns

# Project 1: Montgomery County Crash Data

In this mini-project you'll preprocess, explore, and visualize vehicle crash data from Montgomery County, Maryland. You will work in teams (most teams of 2 members) in class, and outside of class.

Teams: See the [project teams sheet](https://docs.google.com/spreadsheets/d/1KAPwivKZPGoxU5uIlopCOAzk-ns6jNfATQ6yb7OCnHk/edit?usp=sharing).

**Purpose**. To get experience with data cleaning, data exploration, and data visualization on a real-world data set.

**Due date**: See iLearn

**Output**. A report on your work in the form of a Jupyter notebook showing your code and output. Your report should contain the following sections:

* Intro: explain what you will do in a sentence or two
* Data: Explain in a few sentence where the data comes from, what it is about, etc.
* Initial data exploration
* Data cleaning and preprocessing
* Data exploration and visualization
* Conclusions. One paragraph about what you learned, issues you faced, etc.

I want the report to be professional. Focus on the data and what it says. Don’t tell me the story of happened to you while you were doing the work. Also, don’t put lengthy text output from code in your report.

**Process**. I expect you to follow the process we have discussed in class:

1. Acquire the “Crash Reporting - Drivers Data” from <https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Drivers-Data/mmzv-x632>

You should download data in csv format from this page (look for ‘export’ button). Record information about where and when you got the data. Never modify the raw data file you downloaded!

1. Do you know exactly what the data means? Make sure you carefully read information about how the data was obtained, what it means, etc.
2. Create a Jupyter notebook for your work. Write code to read the data as a data frame.
3. Convert the time to a proper Pandas time, and use the time to index the data. See the slides on dates/times in Pandas in the lecture on Pandas aggregation.
4. Do some initial exploration of the data. How many rows and columns? What are the types of the columns (ints, floats, strings, …?). Don’t forget the ‘info' and 'describe' commands.
5. Do data preprocessing. Is there any missing, NA, or corrupted data? Identify and treat them appropriately.
6. Do data exploration and visualization. Look at each of the columns. Which are interesting? What kinds of single and multiple-variable visualizations can you do? Think histograms, bar plots, scatter plots, etc. Most of the variables in this data set are categorical, so you'll want to think about ways to plot multiple categorical variables at once (e.g. stacked or grouped bar plots).
7. While you’re exploring the data, make a note of questions that come to mind that you’d like to answer.

Review the lectures on data preprocessing, exploration, and visualization to make sure you're taking advantage of the methods we learned in class.

Please don’t use any machine learning algorithms in this project. Also, do not use packages not covered in lecture, unless you get permission from Dr. Bruns. It’s good to get a strong foundation with the basic tools in Python. I don't recommend plotting your data on a map. However, you can plot the lat/long position data.

The project is short so keep your goals modest. You do not need to use a large number of features.

I want teams to consider and pursue their own approaches, so please don't discuss your team's work with other teams.

**Ideas to explore**. There are lots of things to look at in this data set. Let your interest and curiosity drive you. Here are examples of things you could look into:

* what are the most common kinds of crashes?
* are certain kinds of crashes more common at certain times and places?
* are some kinds of crashes more common on certain kinds of roads?

These are just a few examples. There are lots of variables in this data and so lots of things to look into.

**Submission**. Each team will submit **two files**: a single .ipynb file plus the HTML file generated from the notebook. The team member responsible for submitting is shown on the project teams sheet -- only the submitter of each team will submit.

Late submissions absolutely not accepted, so don’t submit at the last minute.

I will also ask you to complete a questionnaire in which you will estimate how much of the project work was completed by you and your teammates.

**Grading.** This project is worth 8% of your course score; the final project is worth 12% of your course score. The project will be graded on a scale of 0-100. Grading will be based on these elements:

* 40% technical strength
  + did you use techniques learned in class?
  + did you use them correctly?
  + did you show overall good technical judgement?
* 30% content
  + how much work did you accomplish in the time allotted?
* 20% professional quality
  + is your final report of professional quality?
  + does your report look professional (i.e. good writing, no typos, well-organized)
  + did you show good overall judgement?
* 10% creativity
  + did you do something interesting or unexpected?
  + did you do research to go beyond our coverage of material in class?
* teamwork and participation
  + did you participate?
  + did you attend class every project day?
  + were you a good team player?
  + I will adjust individual grades based on participation. Normally, the adjustment will be between -10 and 10 points, but reserve the right to deduct up to 50 points if I believe your participation was very low.

I will grade subjectively within each of these parts. You will get about 70% of max if you do adequate work, 80% if you do good work, 90-100% if you do very good work and go beyond the basic requirements.