# Data Science Exam – 26 February 2017

This document describes the exam assignment for the course “Data Science – December 2016” @ SoftUni.

## General Information

In this exam, you will be reviewing various crime data in the US. You’ll first get a high-level overview of the data, then you’ll explore the crimes in a bit more detail.

Please, **don’t** download the datasets from the Internet, **use the ones provided in the exam package instead**. The data may have been slightly modified to fit the purpose of the exam better. While testing and evaluating the exam solutions, those datasets will be used.

There are six problems. Each problem is worth 20 points. The maximum number of points for the exam is 100, the other 20 points being a bonus. The optional parts also bring bonus points and may compensate for other points.

All datasets are properties of their respective owners. They are public domain and are free to use in any public research.

## Research Topic

You’ll look at several types of data and answer questions about them using statistics, modelling, visualization, exploratory data analysis, hypothesis testing, etc.

The main question which is going to guide you through your research is: “**What makes people commit crime?**”

Of course, this is too broad a question and you’ll try to answer more concrete questions (laid out in the problems) like: “How does crime in the US evolve in the last years?”, “What crimes are most often?”, “What social and demographical factors may lead to committing a crime?” and so on.

Don’t worry too much about all these questions now, just keep the main one in mind.

**Don’t forget to document** what you do: dataset cleaning and transformations, visualization, analysis and interpretation of the results.

## Before You Start

**Create a blank folder** anywhere in your file system. You will use that to store the exam contents. Name it appropriately, e. g. **“DataScienceExam”**.

Inside this folder, **create another one** named **“datasets”**. Extract all datasets in it. You don’t need to create additional folders inside the “datasets” folder; all files can be stored there directly.

**Create a new Jupyter notebook** in the root folder and name it appropriately. In the first cell, place all **imports** you’ll be using. In the second cell, write a **title** (and subtitle if you wish), and an **author name**. If you don’t want to disclose your name, use a nickname instead.

When you are ready to upload the solution, zip everything you have: Jupyter notebook (.ipynb file), additional text files that you may have created, images, etc. Before uploading, remove the “datasets” folder.

**Note:** If you have used additional datasets to complete the exam, **provide links to them**. Please don’t upload any datasets directly (as they may be too large).

## Putting Crime in the US in Context

The dataset inside crime\_rates.zip contains information about committed crimes in the period 1975 – 2015. Read the file associated with it, and read the dataset.

The data is supposed to be tidy but feel free to perform additional tidying. Remove all missing values. The column “crimes\_percapita” reflects the number of **violent** crimes per capita, not the total number of crimes, and you might want to fix that.

* Have a look at the “months\_reported” column values. What do they mean? What percent of the rows have less than 12 months? How significant is that?
* Overall crime popularity: Create a bar chart of crime frequencies (total, not per capita). Display the type of crime and total occurrences (sum over all years). Sort largest to smallest. Are there any patterns? Which crime is most common?
* Crime popularity by year: Break down the analysis of the previous graph by year. What is the most common crime (total, not per capita) **for each year**? What is the least common one?
* Crime evolution (e. g. crime rates as a function of time): How do crime rates per capita evolve over the years? Create a plot (or a series) of plots displaying how each rate evolves. Create another plot of all crimes (total, not per capita) over the years.

## Crimes by States

* “Criminal” jurisdictions: Plot the sum of all crimes (total, not per capita) for each jurisdiction. Sort largest to smallest. Are any jurisdictions more prone to crime?
* “Criminal” jurisdictions, part 2: Create the same type of chart as above, but use the crime rates per capita this time. Are you getting the same distribution? Why? You may need data from the “population” column to answer this. Don’t perform significance tests, just inspect the plots.
* “Criminal states”: Create the same type of chart as in the first subproblem, but use the **states** instead. You can get the state name in two ways: either the first two letters of the **agency\_code** column or the symbols after the comma in the **agency\_jurisdiction** column.  
  For example, “Cincinnati, OH”, “Cleveland, OH” and “Columbus, OH” belong to the state OH (Ohio).  
  Are any states more prone to crime than others? What may be one reason?
* Plot the total number of crimes on a map. Use the results from the third subproblem in the second problem. Have a look at this answer to get help: <http://stackoverflow.com/questions/7586384/color-states-with-pythons-matplotlib-basemap>. You may want to use different colors for different totals.
* Hypothesis testing: Are crime rates per capita related to population, e. g. does a more densely populated community produce more crime (because there are more people), or less crime (because there is a better police force)? Plot the total number of crimes vs. population to find out. Is there any correlation? If so, what is it? Is the correlation significant?
* Hypothesis testing, part 2: Perform the same operation for each individual type of crime. Is any crime related to higher population? Comment your findings.

## Additional Data

One of the jurisdictions in the dataset is “Minneapolis, MN”. The datasets inside **minneapolis-incidents-crime.zip** contain information about the crimes and incidents in Minneapolis in the years 2010-2016. Unfortunately, It doesn’t come with a readme file but the column descriptions might be good enough to use.

Read, tidy up and clean the “crimes” dataset. You don’t need the “x” and “y” columns, as well as the unique identifiers "OBJECTID" and "ESRI\_OID".

Also, X and Y are actually longitude and latitude (if you’re interested in plotting the locations). Remove missing values.

Convert date strings to real dates.

If you look closely, you’ll see that the times are repeated in two columns (e.g. year and time – you can get these from the date columns). Get rid of the repetitions.

Perform any other cleaning operations as you see fit

* Total number of crimes per year: Count all crimes for years in the dataset (2010-2016). Print the total number.
* Plot how crimes evolve each year.
* Compare the previous plot to the plots in the previous exercise.  
  Note: In order to make comparison better, plot the data for all states again, but this time filter only years 2010-2016. Does the crime rate in MN have any connection to the total crime rate? What percentage of the total crime rate (in all given states) is given by MN?
* Cross-dataset matching: Get data from the previous dataset (crime rates in the US) again. This time, search only for MN and only for years 2010-2016. Do you have any results? If so, the results for total crime in MN should match in both datasets. Do they match?

Optional: Include the “incidents” dataset in your analysis in any way you like.

## Temporal Analysis

You now have data for crimes in MN for each year (problem 3, subproblem 2).

* Look at the crime categories. Which is the most popular crime category in MN overall?
* Break down the data by months. Plot the total number of crimes for each month, summed over the years. Is there a seasonal component? Which month has the highest crime rate? Which has the smallest? Are the differences significant?
* Break the results by weekday. You can get the weekday from the date (there are functions for this). Do more crimes happen on the weekends?
* Break the weekday data by crime type. Are certain types of crime more likely to happen on a given day? Comment your findings.

More advanced time series analysis is out of the scope for this course, so let’s not bother with that.

## Significant Factors in Crime

The communities.zip file contains a dataset with different features which try to explain the number of crimes per capita. The dataset contains a description.

Read and tidy up the dataset. There should be some cleaning needed. Remove all features which are listed as “not predictive” in the dataset description. Can you (and should you) get rid of the missing values? Can you replace them with something instead? Document your reasoning.

Characterize the dataset. How many observations and features?

Give three examples of numerical and three examples of categorical features. For the numerical features, show the range and for the categorical features – show the possible values.

Let’s do some exploratory data analysis on demographics.

* People by age group. There are 4 age groups, each given as a percentage of the total population (you can figure out the percentage of the fifth group).
* People by age group, part 2. Plot all percentage distributions (for each age group). What are the most common percentages for each age group? Do they correspond to the percentages you got in the previous subproblem?  
  Create a chart which displays the total percentage of people in all age groups (sum over all rows).
* People by ethnicity. There are 4 ethnicities (black, white, Asian, Hispanic). Create the same kind of chart but for ethnicity.
* Plot the distributions of the per-capita income. There are several columns. First, plot the distribution of the “overall” per capita income. Next, print all other distributions. For each distribution, print the mean, median and skewness. Incomes are usually skewed. What are your findings?
* Which two income means (by ethnicity) differ most? Is that difference significant? Perform a significance test.
* Cross-dataset matching, part 2: Use the “population density” and “violent crimes per 100K” variables to check whether there is any correlation between the two. Compare your result from this dataset to the one you got in problem 2, subproblem 5.

Optional: Perform any other exploratory data analysis you like, create visualizations and extract meaningful data.

## Predictive Modelling

Run a regression algorithm of your choice over the data. The goal is to predict the ViolentCrimesPerPop column. Don’t forget to:

1. Normalize your data
2. Prepare your data for modelling
3. Split training and testing data. Don’t forget to remove the labels from the testing data