This project will have you doing exploratory data analysis in iPython on a real-world dataset. The goal is to get fluent in working with the standard tools and techniques of exploratory data analysis, by working with a dataset where you have some basic sense of familiarity.
This project is based on Chicago Crash Data available to the public. You should explore the data and uncover interesting observations. You will need to submit all your results in three different formats (.ipynb, .pdf and .py). Make sure to have your code documented with proper comments and the exact sequence of operations you needed to produce the resulting tables and figures.

Data set

You are to download the Chicago Crash Data and perform various EDA tasks on it. You can download the data by accessing the following link and download the CSV file from the Box.
The original data set is available on Chicago Open Data but the original data includes more attributes and fields that were removed in the version available on the Box.
Here's the data set description from the original website:

Crash data shows information about each traffic crash on city streets within the City of Chicago limits and under the jurisdiction of Chicago Police Department (CPD). Data are shown as is from the electronic crash reporting system (E-Crash) at CPD, excluding any personally identifiable information. Records are added to the data portal when a crash report is finalized or when amendments are made to an existing report in E-Crash. Data from E-Crash are available for some police districts in 2015, but citywide data are not available until September 2017. About half of all crash reports, mostly minor crashes, are self-reported at the police district by the driver(s) involved and the other half are recorded at the scene by the police officer responding to the crash. Many of the crash parameters, including street condition data, weather condition, and posted speed limits, are recorded by the reporting officer based on best available information at the time, but many of these may disagree with posted information or other assessments on road conditions. If any new or updated information on a crash is received, the reporting officer may amend the crash report at a later time. A traffic crash within the city limits for which CPD is not the responding police agency, typically crashes on interstate highways, freeway ramps, and on local roads along the City boundary, are excluded from this dataset. As per Illinois statute, only crashes with a property damage value of $1,500 or more or involving bodily injury to any person(s) and that happen on a public roadway and that involve at least one moving vehicle, except bike dooring, are considered reportable crashes. However, CPD records every reported traffic crash event, regardless of the statute of limitations, and hence any formal Chicago crash dataset released by Illinois Department of Transportation may not include all the crashes listed here.

This is a large dataset with many fields. Here is a list of all attributes included:

Column Name	Description	Type
CRASH-RECORD-ID	Unique identifier for the record	Number
CRASH-DATE	Date and time of crash as entered by the officer	Date & Time
POSTED-SPEED-LIMIT	Posted speed limit	Number
TRAFFIC-CONTROL-DEVICE	Traffic control device present at crash location	Plain Text
WEATHER-CONDITION	Weather condition at time of crash	Plain Text
LIGHTING-CONDITION	Light condition at time of crash	Plain Text
FIRST-CRASH-TYPE	Type of first collision in crash	Plain Text
TRAFFIC WAY-TYPE	Traffic way type	Plain Text
ROADWAY-SURFACE-COND	Road surface condition	Plain Text
ROAD-DEFECT	Road defects	Plain Text
CRASH-TYPE	A general severity classification for the crash. Can be either Injury and/or Tow Due to Crash or No Injury / Drive Away	Plain Text
INTERSECTION-RELATED	A field observation by the police officer whether an intersection played a role in the crash. Does not represent whether or not the crash occurred within the intersection.	Plain Text
HIT-AND-RUN	Crash did/did not involve a driver who caused the crash and fled the scene without exchanging information and/or rendering aid	Number
DAMAGE	A field observation of estimated damage.	Plain Text
DATE-POLICE-NOTIFIED	Calendar date on which police were notified of the crash	Date & Time
PRIM-CONTRIBUTORY-CAUSE		Number
NUM-UNITS		Number
INJURIES-TOTAL	Total persons sustaining fatal, incapacitating, non-incapacitating, and possible injuries	Number
INJURIES-FATAL	Total persons sustaining fatal injuries in the crash	Number
INJURIES-INCAPACITATING	Total persons sustaining incapacitating/serious injuries in the crash [1]	Number
INJURIES-NON-INCAPACITATING	Total persons sustaining non-incapacitating injuries in the crash [2]	Number
INJURIES-NO-INDICATION	Total persons sustaining no injuries in the crash	Number

This is an EDA practice, so you need to delve into the data set and identify some useful insights and visualize them. But here are some of the key points every submission must include:

1. The data set need cleaning. Decide what to do with missing values and extra attributes.
2. Some attributes are more useful if you break them into several attributes. An example of this is already included in the data set where the time, day, and month of the crash are given as separate attributes. These attributes allow you to compare crashes based on the day of the week, time, or month (season). Are there other attributes that you can break down into smaller attributes to gain more information from?
3. What are some insights about the crashes and date/time? You can look into season, day of the week, day/night, lightning, weather, etc.
4. Has number of deadly crashes increased recently? Look at the data over the years. Can you identify any significant increase/decrease?
5. Investigate number and type of injuries based on the speed limit.
6. Is there a relationship between hit and run crashes and number of fatal injuries?
7. Do intersection-related crashes result in more fatal injuries?
8. Try to include visualization with your answer to these questions.

9. Come up with at least two more interesting insights and visualize them. (*Suggestions: Season/weather/road condition and fatalities, or hit and run, ...* })

You must have at least one visualization for any questions/insight you are investigating.

Rules

1. This is an **individual assignment**. It is not a group activity.
2. If you are new to Python, this project will be a lot of work, I strongly suggest you start early!
3. Please include some proper explanations for your results. Do not submit a notebook with code cells only. You need to properly describe your methods and discuss/analyze your observations.
4. We will look at the quality of your work for grading. You submission should be coherent and well documented.
5. You might find some online discussions and demos on this data set. It is okay if you look them up, but you must write your own code and analyze the data by yourself.
6. We will run your code though MOSS software to detect copying and plagiarism.

Submission

Submit everything through Gradescope and Blackboard. As mentioned above, you will need to upload:

1. The Jupyter notebook all your work is in (.ipynb file) on Blackboard
2. Python code (.py file) on Blackboard
   - You can zip .py and .ipynb files for Blackboard submission
3. PDF version of your Jupyter notebook on Gradescope

In order to make grading easier for your TA, please use the following format for naming your files:

- netid-hw2-418 {.py, .pynb, .pdf }