BAN 502 - Course Project

Phase 1

The dataset for this project comes from the City of Chicago's Data Portal and contains information about crime in Chicago. To make the dataset a bit more manageable (particularly if you are using an older computer) I have reduced the size of the dataset by sampling 15,000 rows from the original dataset (original size was over 267,000 rows). Note that sampling a large dataset is often a valid approach to working with "big" data.

For Phase 1 you will conduct a thorough **exploratory/descriptive analysis of the dataset**. Please **DO NOT** build any predictive models (e.g., logistic regression, trees, etc.) in this phase. In Phase 2 you will build predictive models to predict the variable "Arrest" (described in the data dictionary below).

Assume that your "audience" for this work are non-technical decision-makers and law enforcement officials.

Phase 1 Deliverables: * There are **three deliverables** for Phase 1
* Deliverable 1: A PowerPoint presentation summarizing your findings from Phase 1. The presentation should be no more than seven slides (including a title slide). Your findings should indicate which variables may be strong predictors of "Arrest" as well as any other interesting descriptive findings. You should include a charts/visuals in the presentation. There should NO VISIBLE R CODE in this deliverable. As noted above, you should assume that the target audience for the deliverable is relatively "non-technical."
* Deliverable 2: A knitted Word document of your Phase 1 R work.
* Deliverable 3: A "flexdashboard" dashboard containing four visualizations that you find to be important. You should upload dashboard to shinyapps.io and include a link to your dashboard in your Canvas submission comments. There is a separate document in the Phase 1 assignment description on Canvas that describes how to create a "flexdashboard".
* Submit the deliverables via Canvas.

Hints/Suggestions/Warnings for Phase 1:
* Carefully review the data dictionary below before proceeding with your work!
* I provide some insight below on working with date data.
* If you wish to do some mapping with Latitude or Longitude that would fine, but I would be extremely careful if you have any plans to try to use these variables in your predictive models.
* You may consider identifying some crimes that are relatively infrequent. You may choose to exclude those crimes from your analysis.

For this project you may need some or all of the packages that we have used in this course. Be sure to install and library packages as needed.

Data dictionary: * ID: Unique identifier for each crime
* Case Number: Chicago PD unique identifier
* Date: Best estimate for time and date of crime (Date given in M/D/Y format)
* Block: Partially redacated address of the crime
* IUCR: Illinois Uniform Crime Reporting code (codes described here: https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e/data)
* Primary Type: Primary type of crime (by IUCR code)
* Description: Secondary description of crime (by IUCR code)
* Location Description: Description of incident location
* Arrest: Indicates whether or not an arrest has been in the case
* Domestic: Indicates whether or not incident involved domestic violence
* Beat: Smallest police division of geographic area for where the crime occurred (Map of Chicago Beats here: https://data.cityofchicago.org/Public-Safety/Boundaries-Police-Beats-current-/aerh-rz74)
* Ward: City Council District where crime occurred
* Community Area: Community Area where crime occurred (Map of Community Areas here: https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6)
* FBI Code: Code of crime by FBI Code (Codes described here: http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html)
* X and Y Coordinates: Partially redacted location of the crime on an official Illinois map

* Year: Year the incident occurred
* Updated On: Date and time of last update of the crime's record
* Latitude and Longitude: Partially redacted latitude and longitude of the location of the crime * Location: Combination of Latitude and Longitude

Data cleaning/preparation tasks:
* Delete the following columns: ID, Case Number, and Updated On, X Coordinate, Y Coordinate, and Location.
* Convert the Date object to an appropriate R Date/Time object with the following code (uses the lubridate package, for more info on working with dates in R with the lubridate package visit the link below https://lubridate.tidyverse.org/):

```
chicago = chicago %>% mutate(Date = mdy_hms(Date))
```

- If you wish to extract the month, day, hour, etc. from the converted date, you can then use code such as:

```
chicago = chicago %>% mutate(Hour = hour(Date)) #creates new variable in dataset with the
#hour extracted from each date/time.
#See the lubridate package reference above for the other time objects that can be extracted
```

- Carefully consider which variables should be factors (categorical) and then convert them. For example, you would probably want to convert your new Hour variable to a factor.

- Create appropriate visualizations (charts, tables, etc.) to examine the relationship between variables and the "Arrest" variable.
- You may decide to create new variables.