1. Identify at least 3 different datasets and perform some initial exploration.

1. [Kickstarter projects](https://www.kaggle.com/kemical/kickstarter-projects)
2. [Credit Card Fraud](https://www.kaggle.com/mlg-ulb/creditcardfraud)
3. [IMB HR Analytics Attrition Dataset](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset)

**Section 1 – Week 8 –**

* Provide an introduction that explains the problem statement you are addressing. Why would someone be interested in this?
  + How can I make my KickStarter campaign a success?
    - To improve the chances of their kickstarter being successful and secure funding
  + How can we determine a fraudulent transaction?
    - Fraud costs American businesses over $650 billion dollars each year
  + How can I keep my employees?
    - Someone would be interested in this because attrition costs companies millions of dollars due to the hiring process and training of new employees to replace those who have left the company.
* Draft 5-10 Research questions that focus on the problem statement.
  + Kickstarter
    - Are there certain types/category of campaigns that are more successful?
    - How much money should you ask for?
    - Is there a time period for the campaign that works better than others?
    - What is the average contribution of a backer?
    - Is there a better time of year to launch a campaign?
  + Fraud
    - How can you deal with a skewed data set?
    - What are the factors that can be identified?
  + Attrition
    - Does business travel impact attrition?
    - How does age play a role in how long someone says with a company?
    - How important is pay?
    - What are the controllable factors for reducing attrition?
* Provide a concise explanation of how you plan to address this problem statement.
  + I plan to perform basic data analysis and correlation to help answer some of these questions. I plan to review things like the mean, median and mode of some of the factors that are of interest.
* Discuss how your proposed approach will address (fully or partially) this problem.
  + I should understand what is important vs unimportant factors in the data. What factors are controllable (for example, companies have control over things like how much someone makes) verses uncontrollable (how old an employee is).
* Do some digging on a dataset that you can use to address the issue.
  + Original source where the data was obtained is cited and, if possible, hyperlinked.
    1. <https://www.kaggle.com/kemical/kickstarter-projects>
    2. <https://www.kaggle.com/mlg-ulb/creditcardfraud>
    3. <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>
  + Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).
    1. 1
    2. This data set is european cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. It only contains numerical values but it doesn’t tell too much about what those values mean.
    3. I learned that this was a fictional dataset created by data scientists.
* Identify the packages that are needed for your project.
  + I’ll likely need ggplot2, randomForest, rpart, cluster, rattle, class, purr, and factoextra for my project
* What types of plots and tables will help you to illustrate the ﬁndings to your research questions?
  + Scatter plots, data tables, correlation tables, box plots
* What do you not know how to do right now that you need to learn to answer your research questions?
  + Should I look into neural networks for any of these?

Sources:

<https://www.kaggle.com/kemical/kickstarter-projects>

<https://www.kaggle.com/mlg-ulb/creditcardfraud>

<https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>

<https://www.allbusiness.com/the-true-cost-of-fraud-direct-costs-2-5222152-1.html>