Problem Solving With Data Challenge

There are three parts of this challenge. The first part involves UNIX, for which you can either use your own UNIX-based system (e.g., Mac) or FTP the data file to the studentweb server and run the commands there. The second part requires working with R. The third part requires working with Python. For those two parts, should download the data files to your local drive and use RStudio and Spyder as needed.  
  
Prepare a single PDF file for your submission. Pay attention to what is required by each question. Some require you to provide your code/commands and other ask for outcomes.   
  
**Part-1: Answer the following questions using UNIX. [Total 20 points]**  
The dataset (in a file named ChicagoCrimes.csv) that we will use is about crimes in Chicago. It has the following attributes: ID,Case Number,Date,Block,IUCR (Illinois Uniform Crime Reporting),Primary Type,Description,Location Description,Arrest,Domestic,Beat,District,Ward,Community Area,FBI Code,X Coordinate,Y Coordinate,Year,Updated On,Latitude,Longitude,Location.

1. How many primary types are in the data? [5 points]
2. How many are described as "ATTEMPT THEFT"? How many of those resulted in arrests? [5 points]
3. Extract all the assault cases (those with primary type "ASSAULT"), sort them by district number and store those records in a separate file. [5 points]
4. Which are the three most dangerous districts and the three safest in the data? [5 points]

Make sure to show your UNIX commands for each of these questions. Use any portion of the data used or produced as a way to support your answers (e.g., showing the first or last few records of the data). But you do not need to submit any resulting files from running these commands.  
  
**Part-2: Answer the following questions using R. [Total 30 points]**  
We will work with a dataset about population estimates. See the attached document Polulation.zip, which has two files -- Population.csv with the dataset and Population.pdf with details of this data and various attributes.

1. Plot a bar chart with top 5 states that have the highest number of international immigration for 2010, 2012, and 2014. [10 points]
2. Plot a line graph - population trends over time by (1) region, and (2) division [10 points]
3. What are the divisions that show the highest increasing rate of population (1) between 2011 and 2012, and (2) between 2013 and 2014? [10 points]

Copy your R code as well as the outcomes.  
  
**Part-3: Answer the following questions using Python. [Total 50 points]**  
We will use Online News Popularity dataset for this. Attached is a zip file OnlineNewsPopularity.zip. It contains a csv file and a text file. The former has the data and the latter has the description of the dataset.

1. [Regression] Here, the target variable is ‘shares,’ which refers to the number of shares. Do correlation analysis with it and other variables in the dataset to find the most relevant (related) factor (make sure this variable is continuous). Build a regression model using these variables to predict 'shares'. Report the model. [5 points]
2. [Regression] Make a subset of online news that was published on weekends. Do correlation analysis with 'shares' and other variables in the dataset to find the two most related factors, and build a regression model using these variables to predict 'shares'. Report the model. [5 points]
3. [Regression] Make a subset of online news that was published on weekdays. Do correlation analysis with 'shares' and other variables in the dataset to find the two most related factors, and build a regression model using these variables to predict 'shares'.  Report the model. [5 points]
4. [Model comparison] Compare the two models from the previous two questions (weekend vs. weekdays) regarding what affects the online news popularity. Report the model (provide your interpretations as you compare those two models). [5 points]
5. [Clustering] Divide the dataset into two clusters, one having less than 1400 shares, and the other having equal to or greater than 1400 shares. Extract the two features you have from the first question, and show a 2D plot with clusters marked. [10 points]
6. [Clustering] Do clustering using two different parameters (k values) for k-means. Show the plots with clusters marked. Provide your interpretations of these two clustering methods in 2-4 sentences comparing them qualitatively. [10 points]
7. [Classification] For classification, take only the two features you’ve gotten from the first question. Split the dataset into 70% for training and 30% for test using kNN. Show the resulting accuracies with three different values of k. [10 points]