

Introduction to Data Science

Lecture 02; January 15th, 2015

Ernst Henle

ErnstHe@UW.edu

Skype: ernst.predixion

Agenda (1)



- Social Interactions
- Reminder: Optional class on programming in R on January 17th 2015
9:00 AM to 12:00 noon
(<http://uweoconnect.extn.washington.edu/datasci250>)
- Review
 - Class Prerequisites
 - DFD
- Break
- Data Sharing
 - Video
 - Getting Data
 - Public Archives

Agenda (2)

- R
 - R as a calculator
 - Hello World
 - Data Structures (vector, data frame)
 - Viewing Datasets
- Quiz: Be prepared to use R and Octave
- Data Preparation (Time permitting)
- GNU-Octave (Time permitting)
 - Octave as a calculator

Class Prerequisites



- Comments on Social interactions: LinkedIn Group "**UW Data Science 2015**".
(<https://www.linkedin.com/groups?gid=6930125>)
- Review of Class Structure and Prerequisites
 - Ability to take Quiz during class
 - Ability to run VMWare on your Computer
 - Access to Catalyst (<https://catalyst.uw.edu/>)
 - Ability to submit homework to the Catalyst drop box called "**UW Data Science 2015 Homework Submission**"
(<https://catalyst.uw.edu/collectit/dropbox/ernsthe/34022>)
 - Ability to get class resources before each class from: "**UW Data Science 2015 Resources**"
(<https://catalyst.uw.edu/workspace/ernsthe/48432>)

Data Flow Diagram (DFD) Examples

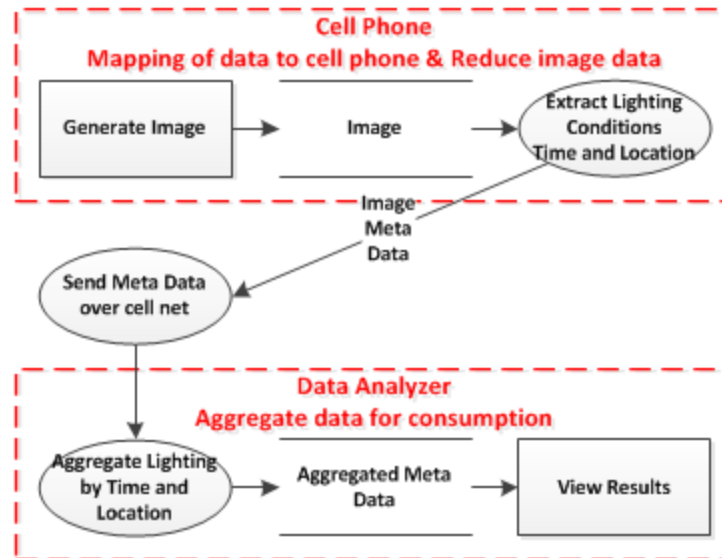
Image Aggregation DFD (1)

1. Describe, in a few sentences, a data science task that interests you. The following is one that interests me:
 1. Data are extracted and processed from images on cell phones
 2. The processed data are combined
 3. The combined data are used to derive meaning, like: Which are the popular tourist locations?
2. Construct a data flow diagram that depicts the data processing that is required to complete the task in item 1

Image Aggregation DFD (2)

- Collect and aggregate cell phone camera images
 - The image is taken (Image is mapped to cell phone)
 - Image is associated with cell location and time
 - The image data is extracted (Data Reduction)
 - The data (Image characteristics, time, and location) are sent
 - The data are collected and aggregated by location and time
 - The data are viewed

Image Aggregation DFD (3)



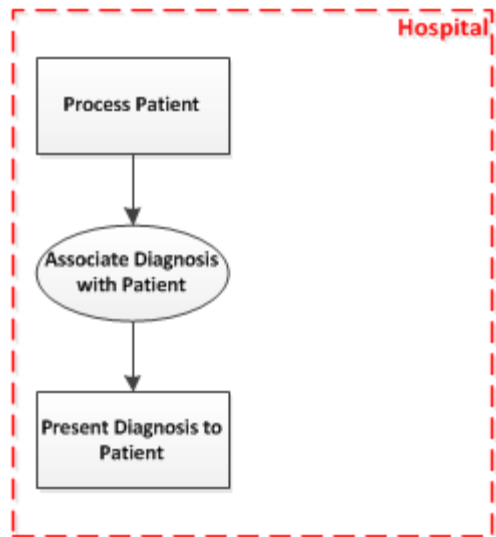
DFD: Digital Pathology (1)

Digital Pathology

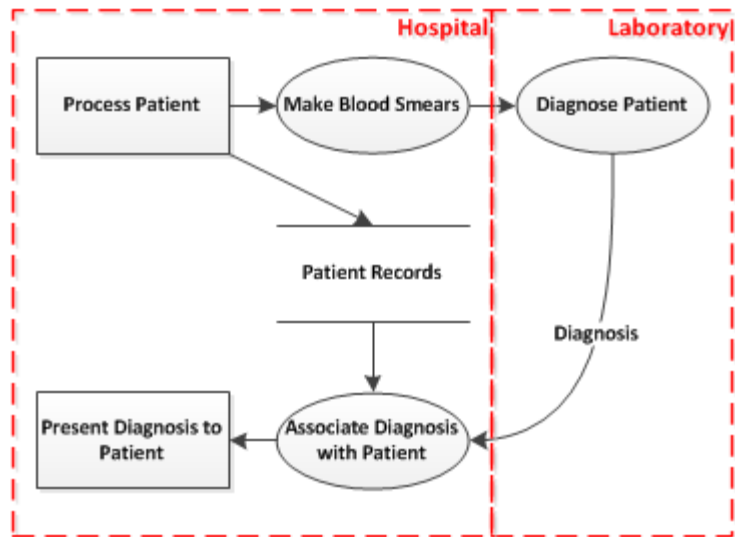
Many blood disorders manifest themselves through easily recognizable morphological changes, but the affected cells may be as few as one in a hundred thousand. Given the scarcity and cost of pathologists, it is not possible to routinely screen for these blood disorders. We would like to find an automated way of diagnosing such disorders.

We use a pathologist to score aberrant cells and correlate these findings with shape characteristics determined by image segmentation.

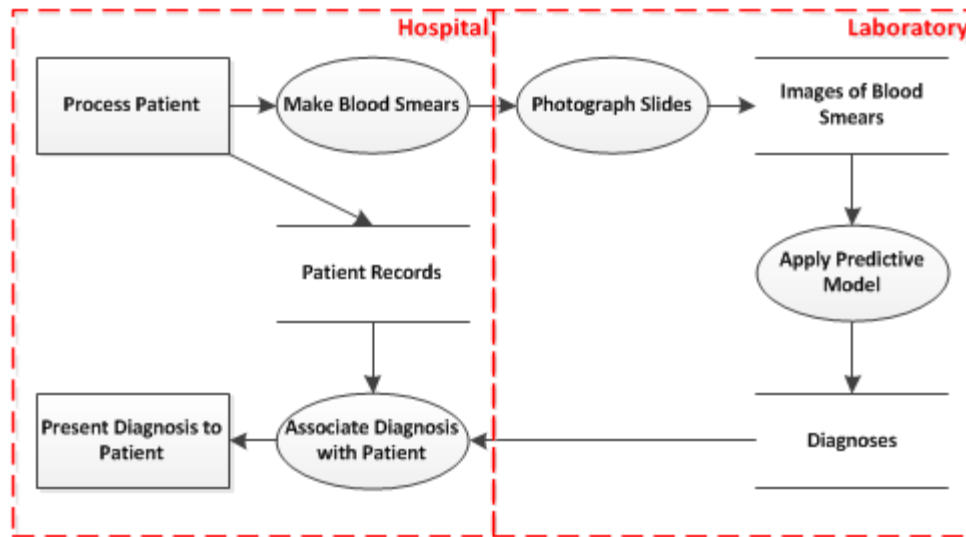
DFD: Digital Pathology (2)



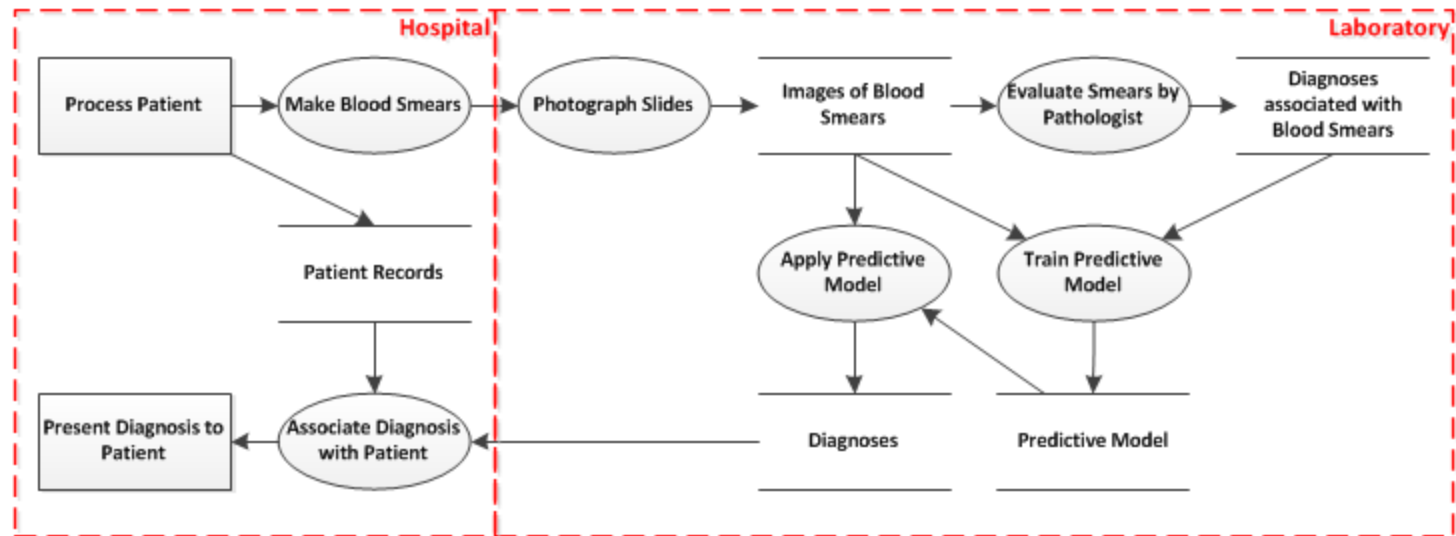
DFD: Digital Pathology (3)



DFD: Digital Pathology (4)



DFD: Digital Pathology (5)



Data Flow Diagram (DFD) Examples

Break

Data Sharing (watch these videos)

- <http://www.youtube.com/watch?v=RVZbk3GEVSw>
 - It's all there.
 - Just follow that link. You don't need the data.
 - I've already analyzed the data.
 - I can't find my data.
- <http://www.youtube.com/watch?v=RtSv0gSbCP8>
 - This is my only copy
 - The data format is unusable.
- <http://www.youtube.com/watch?v=-MIH8PkuUo4>
 - The attributes (headings) are self-evident
 - Somebody else knows how that column was calculated.
 - You can figure out for yourself what that column means.

Some Sample Archives

- <http://data.bls.gov/cgi-bin/surveymost?bls>
 - Bureau of Labor Statistics provided monthly datasets on labor since 2003. These data sets are somewhat short as they only provide a single number in a single year. You could extract a lot of data, but it would need to be compiled together as this website separates every attribute of the data.
- <http://socialcomputing.asu.edu/datasets/Twitter>
- <http://socialcomputing.asu.edu/uploads/1296759055/Twitter-dataset.zip>
 - Link provides a .zip file with GB data of twitter data
- <http://archive.ics.uci.edu/ml/datasets.html>
 - **Status:** Well-formed and existing
 - **Description:** UCI's machine learning repository, A collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.
- <http://www.ourairports.com/data/>
 - **Status:** Well-formed and existing CSV file
 - **Description:** Provide 6 CSV files for airport information, including type, name, location, and frequency of airports, and also containing website URL for different regions etc.
- <http://www.quandl.com/>
 - **Status:** Well-formed and existing CSV, Excel files
 - **Description:** Give share's price and volume for Microsoft, Oracle, IBM, HP, Dell, Cisco, Apple, and Google in every business day from September 1997 to 2013, nowadays. Updated every day.
- Archive of archives: DataSets.doc

R

- Open in R Studio: DataScience02a.R, DataScience02b.R, and DataScience02c.R

Quiz 02

- <https://catalyst.uw.edu/webq/survey/ernsthe/258306>
- You should use R and Octave during the Quiz.
- Collaborate! Check your answers with others. We did not cover everything in class!

Data Preparation

Data Preparation (0)

- Time Permitting
- DataScience02e.R

Data Preparation (1)

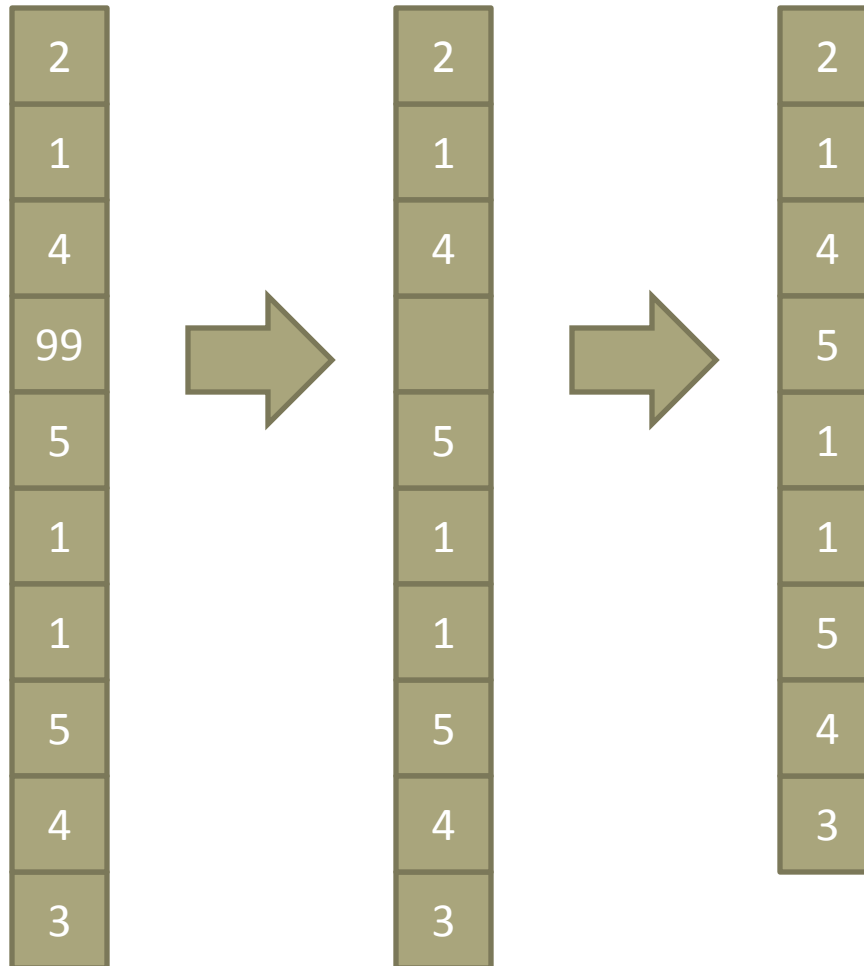
- Schematize (Shape Data)
 - Create Tables
 - Flatten Data
 - Star http://en.wikipedia.org/wiki/Star_schema
 - Snowflake http://en.wikipedia.org/wiki/Snowflake_schema
 - Specify Input vs. Target
 - Specify attributes that are neither Input nor Target
- Clean Data (Today's topic)

Data Preparation (2)

- Clean Data
 - Outlier Removal
 - Numeric
 - Remove data beyond 3 standard deviations (1, 2, 2, 3, 3, 3, 4, 4, 5, 99)
 - `x <- x[x < 10]`
 - Categorical
 - Categories: Remove cases that have less than 1% support (20 X A, 20 X B, 1 X C, 20 X D, 20 X E, 20 X F)

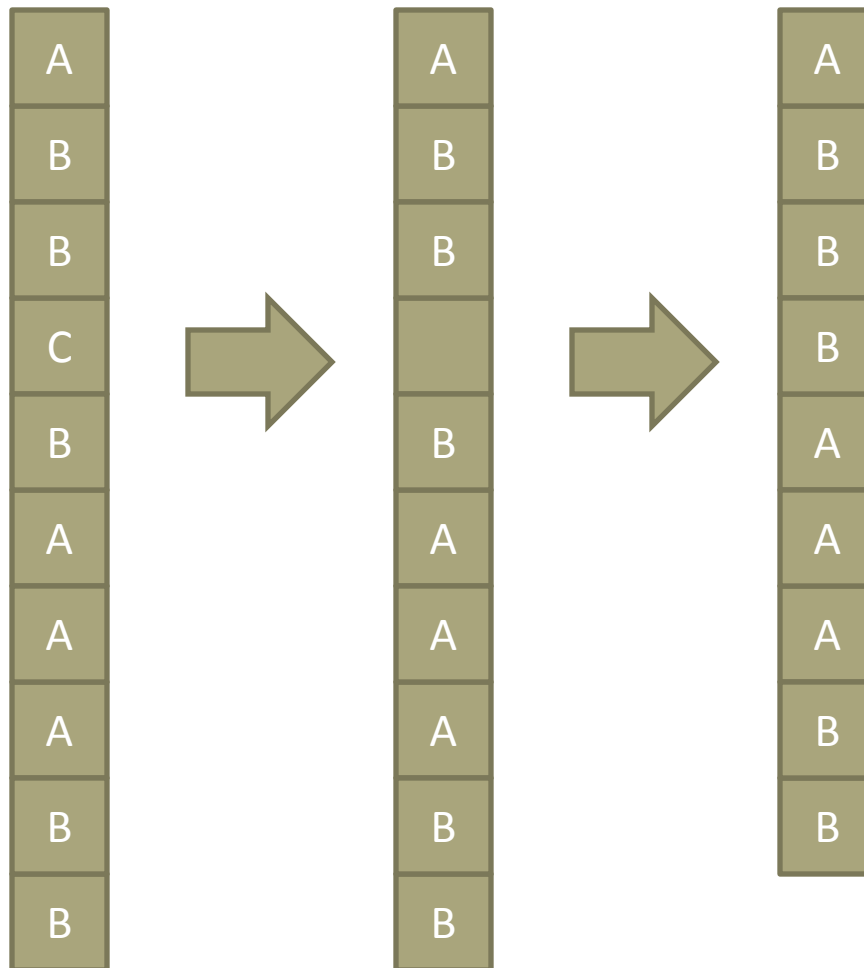
Data Preparation (3)

Outlier Removal (Numeric)



Data Preparation (4)

Outlier Removal (Category)

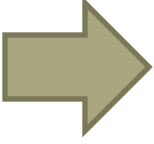


Data Preparation (5)

- Clean Data
 - Relabeling
 - Simplify (e.g. all 4 year degrees, like Bachelors, A.B. BSc, etc. as BS)
 - Example:
 - Vehicle: (Car, Automobile, Bike, Truck, Bicycle, Sedan, Coupe, Cycle, Truck, Velo, Automobile, Bike)
 - Car, Automobile, Sedan, Coupe -> Car
 - Bike, Bicycle, Cycle, Velo -> Bike
 - Truck -> Truck
 - Vehicle: (Car, Car, Bike, Truck, Bike, Car, Car, Bike, Truck, Bike, Car, Bike)
 - De-code (numbers to categories)
 - Example1: Origin: (3, 1, 2, 1, 1, 2)
 - 1 -> USA
 - 2 -> Europe
 - 3 -> Japan
 - Origin: (Japan, USA, Europe, USA, USA, Europe)
 - Example2: Origin: `x <- as.character(3, 1, 2, 1, 1, 2)`

Data Preparation (6)

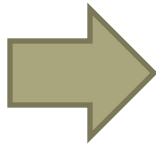
Relabeling (Simplify)

Vehicle		Vehicle
Car		Car
Bike		Bike
Velo		Bike
Truck		Truck
Bicycle		Bike
Sedan		Car
Coupe		Car
Auto		Car
Lorry		Truck
Truck		Truck

Data Preparation (7)

Relabeling (Decode)

Vehicle
1
2
2
3
2
1
1
1
3
3



Vehicle
Car
Bike
Bike
Truck
Bike
Car
Car
Car
Truck
Truck

Code	Item
1	Car
2	Bike
3	Truck

Data Preparation (8)

- Clean Data (continued)
 - Casting
 - Characters to Numbers: ("4", "-7", "X", "3") -> (4, -7, NA, 3)
 - Numbers to Characters : (4, -7, NA, 3) -> ("4", "-7", NA, "3")
 - Normalize
 - Normalize (Linear)
 - offset and multiplier $y = a + bx$ or $y = (x - c)/d$; Where: $a = -c/d$; $b = 1/d$
 - Min-Max where: $c = \min$; $d = \max - \min$
 - Z-score: where $c = \text{mean}$; $d = \text{sigma}$
 - MAD (http://en.wikipedia.org/wiki/Median_absolute_deviation) where $c = \text{median}$; $d = \text{median of differences to median}$
 - Normalize (Non-Linear)
 - Log-normalization: $y = \text{Log}(x)$ or similar
 - Equalization

Data Preparation (9)

Normalization

Orig		MM	Z
2		.3	-0.23
-1		0	-1.09
0		.1	-0.8
1		.2	-0.52
7		.1	1.2
9		1	1.78
7		.8	1.2
1		.2	-0.52
1		.2	-0.52
1		.2	-0.52

Data Preparation (10)

- Clean Data (continued)
 - Binarization Categorical to Numerical (Binary)
 - 1 column of Colors (Red, Green, Blue) -> three columns called isRed, isGreen, and isBlue
 - Color: Red, Green, Blue, Blue, Red, Red ->
 - -> isRed: 1, 0, 0, 0, 1, 1
 - -> isGreen: 0, 1, 0, 0, 0, 0
 - -> isBlue: 0, 0, 1, 1, 0, 0
 - Discretization
 - Age: (10, 23, 11, 55, 60, 32, 99, 4, 32, 33, 0) ->
 - Equal Range (0 – 33) (34 – 66) (67 – 99) -> (Low, Low, Low, Med, Med, Low, High, Low, Low, Low, Low)
 - Equal Area (0 - 11) (23 - 33) (55 - 99) -> (Low, Med, Low, High, High, Med, High, Low, Med, Med, Low)
 - Null Handling
 - (4, -7, NA, 3) ->
 - Value removal or Row Removal -> (4, -7, 3)
 - value substitution -> (4, -7, 0, 3)

Data Preparation (11)

Binarization

Vehicle		Car	Bike	Truck
Car		1	0	0
Bike		0	1	0
Velo		0	1	0
Truck		0	0	1
Bicycle		0	1	0
Sedan		1	0	0
Coupe		1	0	0
Auto		1	0	0
Lorry		0	0	1
Truck		0	0	1

Data Preparation

Break

R

- Open in R Studio: DataScience02d.R

Optional R Language Class

- Reminder: Optional class on programming in R on January 17th 2015
9:00 AM to 12:00 noon
(<http://uweoconnect.extn.washington.edu/datasci250>)

Assignment (0)

- Download the Hadoop VM (Cloudera-Training-VM-4.2.1.p-vmware_prist.zip) from this link: <http://1drv.ms/1Inp5er>. The zip file is about 2.4 GB. You may have to sign up for a Microsoft account to download the zip. If you do not want a Microsoft account, you can create a fake persona with a random email (<http://www.yopmail.com>). Indicate whether you successfully downloaded the VM image.

Assignment (1)

For all R assignment items, use the patterns described in DataScience02a.R, DataScience02b.R, and DataScience02c.R

1. Using R Data Preparation

a. Get Indian Liver Patient Dataset:

a. `url <- "http://archive.ics.uci.edu/ml/machine-learning-databases/00225/Indian%20Liver%20Patient%20Dataset%20\(ILPD\).csv"`

b. `ILPD <- read.csv(url, header=FALSE, stringsAsFactors=FALSE)`

b. Get the column headers from:

`http://archive.ics.uci.edu/ml/datasets/ILPD+\(Indian+Liver+Patient+Dataset\)#`

c. Manually construct a vector of column using

a. `headers <- c(<name1>, <name2>, ...)`

d. Associate names with the dataframe

a. `names(<dataframe>) <- headers`

Assignment (2)

2. Using R Data Exploration on ILPD
 - a. Use **head(ILPD)** to view the first 6 rows.
 - b. Determine the **mean**, **median**, and standard deviation (**sd**) of each column.
 - c. What does **na.rm = TRUE** do in `sd(x, na.rm = TRUE)`?
 - d. Create Histograms (**hist**) for each column where possible.
 - e. Use the **plot(ILPD)** function on this data frame to present a general overview of the data. You want to see a matrix of many plots. Your efforts may be thwarted because the Gender column is not numeric. You can skip the Gender column, or you can turn the gender column into a numeric column. You might need help from a fellow student, the LinkedIn group, or me. Look at the plots and answer:
 - a. What can you say about the data?
 - b. How can you tell if a vector contains continuous numbers or binary data?
 - c. How can you tell if two vectors are correlated?

Assignment (3)

3. Using Data Preparation concepts Create examples of the following data preparation processes in R:
 - a. Remove Outliers: `c(-1, 1, 5, 1, 1, 17, -3, 1, 1, 3)`
 - b. Relabel: `c('BS', 'MS', 'PhD', 'HS', 'Bachelors', 'Masters', 'High School', 'BS', 'MS', 'MS')`
 - c. Normalize: `c(-1, 1, 5, 1, 1, 17, -3, 1, 1, 3)`
 - a. Min-Max Normalization
 - b. Z-score normalization
 - d. Binarize: `c('Red', 'Green', 'Blue', 'Blue', 'Blue', 'Blue', 'Blue', 'Red', 'Green', 'Blue')`
 - e. Discretize: `c(3, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 12, 23, 23, 25, 81)`
 - a. 3 Bins of equal range
 - b. 3 Bins Equal of near equal amounts (Do this by hand. Writing equalization code is tricky)
4. Combine the assignment items 1, 2, and 3 into a single R file. In the same R-file, indicate whether you downloaded the zip with the Hadoop VM. Submit the file by Sunday 11:00 PM PST to the homework submission site on catalyst. If you cannot submit the assignment on time, please notify me before the deadline at ErnstHe@UW.edu.
5. Reading assignment:
 - http://en.wikipedia.org/wiki/Cluster_analysis
 - http://en.wikipedia.org/wiki/K-means_clustering
 - http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/

Introduction to Data Science