# Week 1

By Rachel Huang and Jackson Maris
University of North Carolina Wilmington Statistical Data Mining and Machine Learning
May 25, 2018

# Outline

Project 1:

- Numerical and graphical summaries of age and gender using the MORPH-II partial dataset

Project 2:

- Analyzing the relation between age, gender and race using the entire MORPH-II dataset

Project 3:

- Creating regression models and classifying data from the MORPH-II partial dataset

# Project 1

Process:

- Two vectors were created for age and gender.
- The 7th character from the end of the filename, either “M” or “F”, indicates gender.
- The 6th and 5th characters from the end of the filename indicate the age.
- The remaining 2,568 columns are the corresponding Bio-Inspired Features (BIF) for each person.

| | Filename | BIF1 | BIF2 | BIF3 | BIF4 | BIF5 |
|---|---|---|---|---|---|---|
| 1 | 022066_01M64.JPG | 204 | 204 | 199 | 203 | 213 |

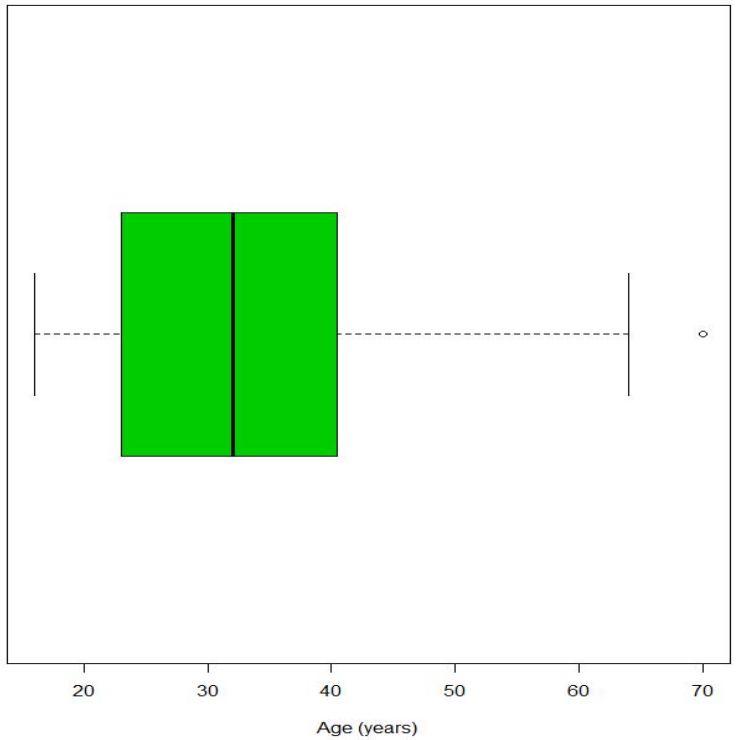
Part of a row in the dataset.

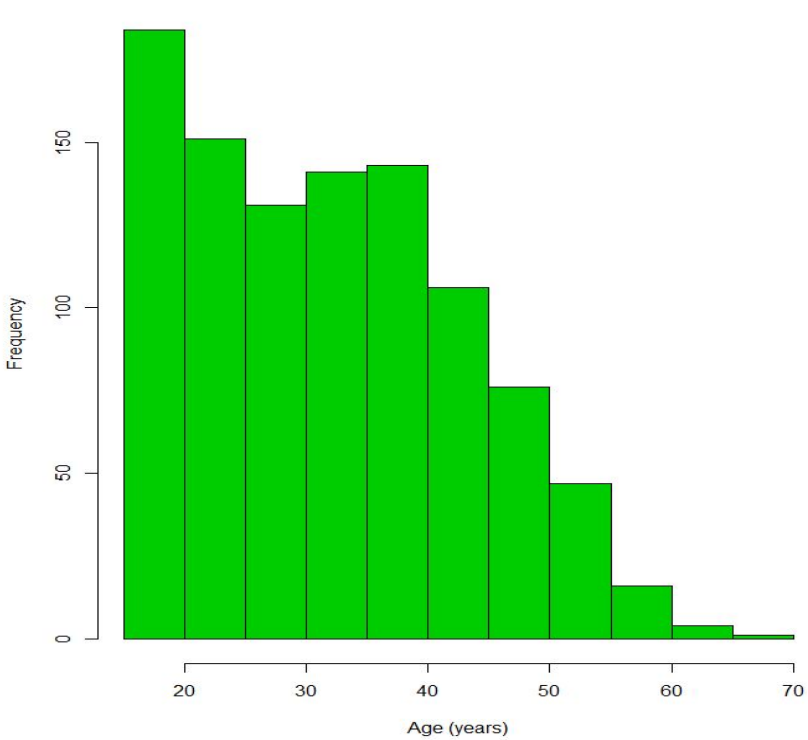
# Project 1 Age Analysis

NUMERICAL SUMMARY OF AGE

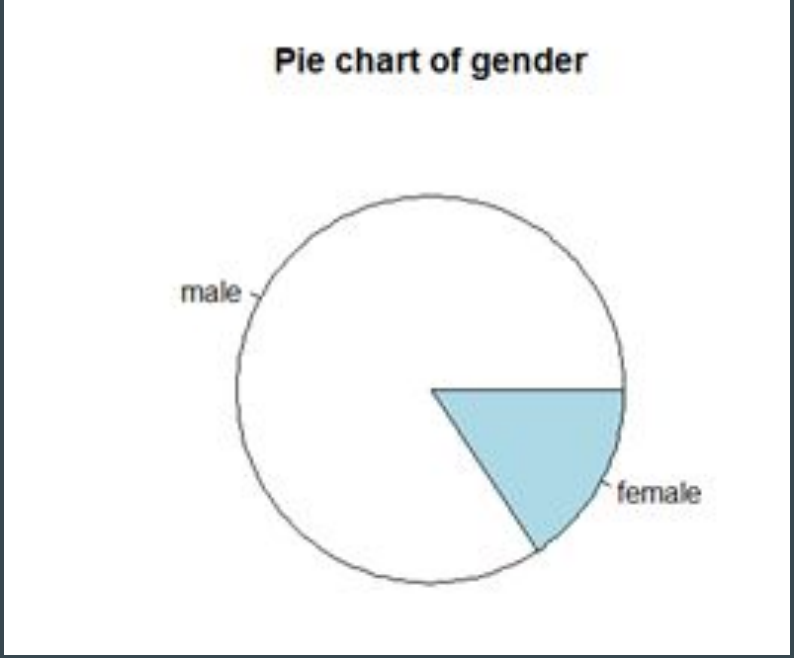
| Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|
| 16 | 23 | 32 | 32.41 | 40.25 | 70 |

# Project 1 Age Graphs



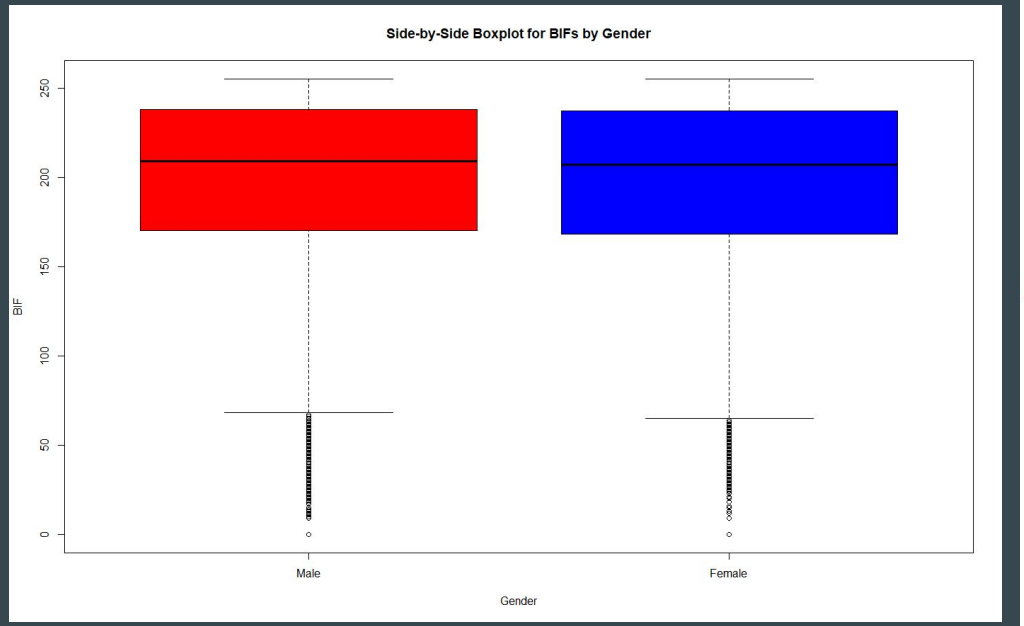



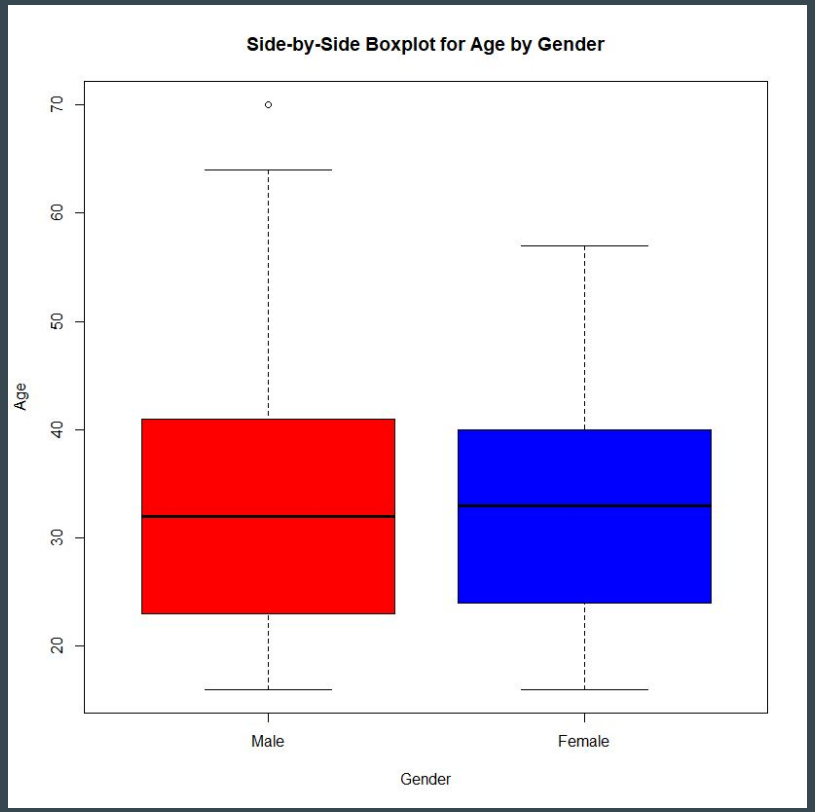
# Project 1 Gender Analysis




NUMERICAL SUMMARY OF GENDER

| | Male | Female |
|---|---|---|
| Frequency | 843 | 157 |


Side-by-Side Boxplot for BIFs by Gender
BIF
0
50
100
150
200
250
Male
Female
Gender


Side-by-Side Boxplot for Age by Gender
Age
20
30
40
50
60
70
Male
Female
Gender

# Project 2

Process:

- Analyzed the difference between a clean and dirty data set.
- Looked at relationship between:
  - Gender and race
  - Gender and additional arrests
- Combining the partial and full Morph-II data sets

Morph_2008_nonCommercial.csv

# of males: 11,459

# of females: 2,159

# distinct number of subjects: 13, 618

| | Black | White | Hispanic | Asian | Other | Total |
|---|---|---|---|---|---|---|
| Male | 8,838 | 2,070 | 517 | 49 | 15 | 11,489 |
| Female | 1,494 | 634 | 30 | 6 | 5 | 2,169 |
| Total | 10,332 | 2,704 | 547 | 55 | 20 | 13,658 |

morphII_cleaned_v2.csv

# of males; 11,458

# of females: 2,159

# of subjects: 13,617

| | Black | White | Hispanic | Asian | Other | Total |
|---|---|---|---|---|---|---|
| Male | 8,829 | 2,056 | 507 | 47 | 19 | 11,458 |
| Female | 1,491 | 628 | 28 | 4 | 8 | 2,159 |
| Total | 10,320 | 2,684 | 535 | 51 | 27 | 13,617 |

# Additional images

| | 1 | 2 | 3 | 4 | 5+ | Total |
|---|---|---|---|---|---|---|
| Male | 2350 | 3606 | 1975 | 1135 | 2020 | 11086 |
| Female | 478 | 0 | 352 | 172 | 360 | 1362 |
| Total | 2828 | 3606 | 2372 | 1307 | 2380 | 12448 |

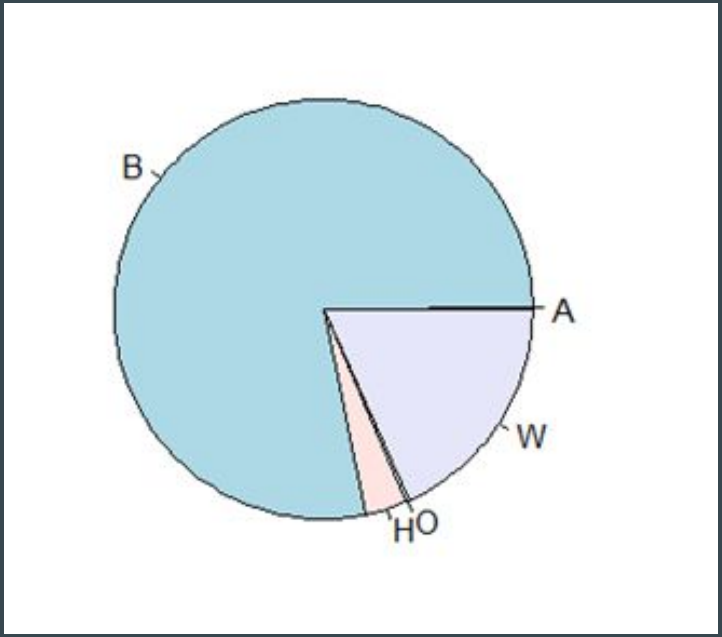
| | 1 | 2 | 3 | 4 | 5+ | Total |
|---|---|---|---|---|---|---|
| Male | 2350 | 3606 | 1975 | 1135 | 2020 | 11086 |
| Female | 478 | 712 | 352 | 172 | 360 | 2074 |
| Total | 2828 | 4318 | 2327 | 1307 | 2380 | 13440 |

# Decade-of-Life

| | <20 | 20-29 | 30-39 | 40-48 | 50+ | Total |
|---|---|---|---|---|---|---|
| Male | 1966 | 3068 | 2556 | 2071 | 755 | 10416 |
| Female | 294 | 541 | 590 | 433 | 99 | 1957 |
| Total | 2260 | 3609 | 3146 | 2504 | 854 | 12373 |

| | <20 | 20-29 | 30-39 | 40-49 | 50+ | Total |
|---|---|---|---|---|---|---|
| Male | 1968 | 3943 | 3370 | 2794 | 990 | 13065 |
| Female | 294 | 672 | 754 | 570 | 140 | 2430 |
| Total | 2262 | 4615 | 4124 | 3364 | 1130 | 15495 |

# Combining MORPH-II datasets




| Race | Asian | Black | Hispanic | White | Other |
|---|---|---|---|---|---|
| | 2 | 781 | 32 | 183 | 2 |

# Project 3

Process:

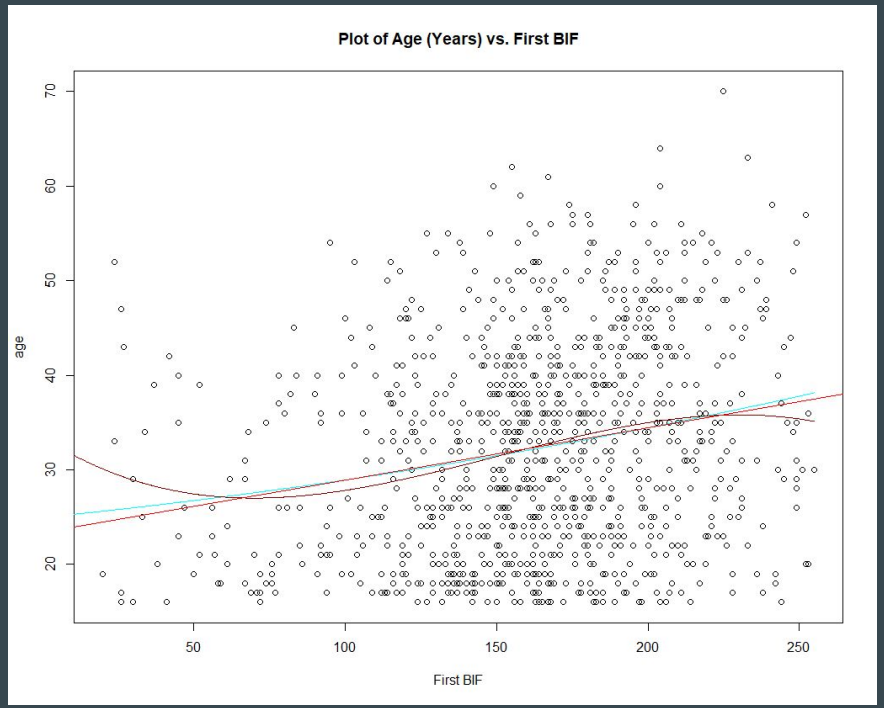
- The first 20 BIF points were used to calculate the linear, quadratic and polynomial regressions.
- Logistic regression used the entire BIF data set to build models.

Unfinished:

- LDA, QDA, and KNN
- Splitting the data into training and testing data

| | Filename | BIF1 | BIF2 | BIF3 | BIF4 | BIF5 |
|---|---|---|---|---|---|---|
| 1 | 022066_01M64.JPG | 204 | 204 | 199 | 203 | 213 |

# Linear, Quadratic and Polynomial Regression




Adjust R^2:

- Linear: .0446
- Quadratic: .04403
- Polynomial: .04738

# Logistic Regression

Accuracy: .931

Error: .069

Sensitivity: .77

Specificity: .96

| | Male | Female |
|---|---|---|
| Male | 810 | 36 |
| Female | 33 | 121 |

# Questions?