Final Project Introduction

Group number 7

Members:

Minhaz Khan, Bobak Ahmar, Vincent La, Navin Chandradat, Truman Zheng

Tasks:

data cleaning, summary of data, introduction - Truman

Organizing presentation/presenting - Truman, Minhaz, Vincent

Anaylizing data/performing various test - Everyone (idea: each of us analize different variables)

Putting everything together/conclusions - Navin, Bobak

Introduction

Breast cancer is a malignant cell growth in the breast. if it is left untreated the cancer can spread to other parts of the human body and it can be very deadly. there are generally two type of tumors non-cancerous and cancerous and the difference between the two is important, Benign tumor is non-cancerous and not dangerous on its own, but a malignant tumor, means the mass is cancerous.

summary of the data

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
# preview of the data
mydata = read_csv("Project3-Data.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   id = col_integer(),
##   diagnosis = col_character()
## )

## See spec(...) for full column specifications.
```

```r
mydata
```

```
## # A tibble: 569 x 32
##         id diagnosis radius_mean texture_mean perimeter_mean area_mean
##      <int> <chr>           <dbl>        <dbl>          <dbl>     <dbl>
##  1 8.42e5 M                18.0         10.4           123.      1001
##  2 8.43e5 M                20.6         17.8           133.      1326
##  3 8.43e7 M                19.7         21.2           130       1203
##  4 8.43e7 M                11.4         20.4            77.6      386.
##  5 8.44e7 M                20.3         14.3           135.      1297
##  6 8.44e5 M                12.4         15.7            82.6      477.
##  7 8.44e5 M                18.2         20.0           120.      1040
##  8 8.45e7 M                13.7         20.8            90.2      578.
##  9 8.45e5 M                13           21.8            87.5      520.
## 10 8.45e7 M                12.5         24.0            84.0      476.
## # ... with 559 more rows, and 26 more variables: smoothness_mean <dbl>,
## #   compactness_mean <dbl>, concavity_mean <dbl>, `concave
## #   points_mean` <dbl>, symmetry_mean <dbl>, fractal_dimension_mean <dbl>,
## #   radius_se <dbl>, texture_se <dbl>, perimeter_se <dbl>, area_se <dbl>,
## #   smoothness_se <dbl>, compactness_se <dbl>, concavity_se <dbl>,
## #   `concave points_se` <dbl>, symmetry_se <dbl>,
## #   fractal_dimension_se <dbl>, radius_worst <dbl>, texture_worst <dbl>,
## #   perimeter_worst <dbl>, area_worst <dbl>, smoothness_worst <dbl>,
## #   compactness_worst <dbl>, concavity_worst <dbl>, `concave
## #   points_worst` <dbl>, symmetry_worst <dbl>,
## #   fractal_dimension_worst <dbl>
```

```r
# number of variables we have
num_var = ncol(mydata) - 1
num_var
```

```
## [1] 31
```

```r
# number of observation we have
num_obs = nrow(mydata)
num_obs
```

```
## [1] 569
```

```r
# the number of each type of tumor
table(mydata$diagnosis)
```

```
## 
##   B   M 
## 357 212
```

```r
# small summary of a few variables
summary(mydata$radius_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   6.981  11.700  13.370  14.127  15.780  28.110
```

```r
summary(mydata$texture_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.71   16.17   18.84   19.29   21.80   39.28
```

```
summary(mydata$perimeter_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.79   75.17   86.24   91.97  104.10  188.50
```

```
summary(mydata$perimeter_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.79   75.17   86.24   91.97  104.10  188.50
```