

# Project\_Test

Suraj Nihal

2023-08-13

## Libraries and Dataset

```
library(dplyr)
library(corrplot)
library(psych)
library(yacca)
library(REdaS)
library(ggplot2)
library(ltm)

data <- read.csv("~/Downloads/data.csv")
```

## Problem 1

I will be applying Common Factor Analysis on the Breast Cancer Dataset -

### Introduction

Determining dimensions of Breast Cancer using Common Factor Analysis (CFA)

First we will discover and visualize the data to gain insights then we will apply Common factor analysis (CFA) to determine which features effect Breast Cancer the most

The Dataset

The Breast Cancer (Wisconsin) Diagnosis dataset contains the diagnosis and a set of 30 features describing the characteristics of the cell nuclei present in the digitized image of a fine needle aspirate (FNA) of a breast mass.

Attribute Information:

- 1) ID number
- 2) Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

- a) radius (mean of distances from center to points on the perimeter)
- b) texture (standard deviation of gray-scale values)
- c) perimeter
- d) area
- e) smoothness (local variation in radius lengths)
- f) compactness ( $\text{perimeter}^2 / \text{area} - 1.0$ )
- g) concavity (severity of concave portions of the contour)
- h) concave points (number of concave portions of the contour)
- i) symmetry
- j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are re-coded with four significant digits [1]

Fine Needle Aspiration of the Breast

According to the American Cancer Society website, during a fine needle aspiration (FNA), a small amount of breast tissue or fluid is removed from a suspicious area with a thin, hollow needle and checked for cancer cells. This type of biopsy is sometimes an option if other tests show you might have breast cancer (although a core needle biopsy is often preferred) [2]

## Data Cleaning and Inspection

```
#Checking Sample Size and Number of Variables
dim(data)
```

```
## [1] 569 33
```

```
#569-Sample Size and 34 variables
```

```
#Showing head of the dataset
head(data, 3)
```

```
##           id diagnosis radius_mean texture_mean perimeter_mean area_mean
## 1    842302         M      17.99       10.38         122.8       1001
## 2    842517         M      20.57       17.77         132.9       1326
## 3  84300903         M      19.69       21.25         130.0       1203
## smoothness_mean compactness_mean concavity_mean concave.points_mean
## 1         0.11840         0.27760         0.3001         0.14710
## 2         0.08474         0.07864         0.0869         0.07017
```

```
## 3          0.10960          0.15990          0.1974          0.12790
## symmetry_mean fractal_dimension_mean radius_se texture_se perimeter_se
## 1          0.2419          0.07871          1.0950          0.9053          8.589
## 2          0.1812          0.05667          0.5435          0.7339          3.398
## 3          0.2069          0.05999          0.7456          0.7869          4.585
## area_se smoothness_se compactness_se concavity_se concave.points_se
## 1 153.40          0.006399          0.04904          0.05373          0.01587
## 2  74.08          0.005225          0.01308          0.01860          0.01340
## 3  94.03          0.006150          0.04006          0.03832          0.02058
## symmetry_se fractal_dimension_se radius_worst texture_worst perimeter_worst
## 1  0.03003          0.006193          25.38          17.33          184.6
## 2  0.01389          0.003532          24.99          23.41          158.8
## 3  0.02250          0.004571          23.57          25.53          152.5
## area_worst smoothness_worst compactness_worst concavity_worst
## 1    2019          0.1622          0.6656          0.7119
## 2    1956          0.1238          0.1866          0.2416
## 3    1709          0.1444          0.4245          0.4504
## concave.points_worst symmetry_worst fractal_dimension_worst X
## 1          0.2654          0.4601          0.11890 NA
## 2          0.1860          0.2750          0.08902 NA
## 3          0.2430          0.3613          0.08758 NA
```

```
#Showing summary of the dataset
summary(data)
```

```
##      id      diagnosis      radius_mean      texture_mean
## Min.   :    8670 Length:569      Min.   : 6.981      Min.   : 9.71
## 1st Qu.: 869218 Class :character 1st Qu.:11.700      1st Qu.:16.17
## Median : 906024 Mode  :character  Median :13.370      Median :18.84
## Mean   : 30371831      Mean   :14.127      Mean   :19.29
## 3rd Qu.: 8813129      3rd Qu.:15.780      3rd Qu.:21.80
## Max.   :911320502      Max.   :28.110      Max.   :39.28
## perimeter_mean area_mean smoothness_mean compactness_mean
## Min.   : 43.79      Min.   : 143.5      Min.   :0.05263      Min.   :0.01938
## 1st Qu.: 75.17      1st Qu.: 420.3      1st Qu.:0.08637      1st Qu.:0.06492
## Median : 86.24      Median : 551.1      Median :0.09587      Median :0.09263
## Mean   : 91.97      Mean   : 654.9      Mean   :0.09636      Mean   :0.10434
## 3rd Qu.:104.10      3rd Qu.: 782.7      3rd Qu.:0.10530      3rd Qu.:0.13040
## Max.   :188.50      Max.   :2501.0      Max.   :0.16340      Max.   :0.34540
## concavity_mean concave.points_mean symmetry_mean fractal_dimension_mean
## Min.   :0.00000      Min.   :0.00000      Min.   :0.1060      Min.   :0.04996
## 1st Qu.:0.02956      1st Qu.:0.02031      1st Qu.:0.1619      1st Qu.:0.05770
## Median :0.06154      Median :0.03350      Median :0.1792      Median :0.06154
## Mean   :0.08880      Mean   :0.04892      Mean   :0.1812      Mean   :0.06280
## 3rd Qu.:0.13070      3rd Qu.:0.07400      3rd Qu.:0.1957      3rd Qu.:0.06612
## Max.   :0.42680      Max.   :0.20120      Max.   :0.3040      Max.   :0.09744
## radius_se texture_se perimeter_se area_se
## Min.   :0.1115      Min.   :0.3602      Min.   : 0.757      Min.   : 6.802
## 1st Qu.:0.2324      1st Qu.:0.8339      1st Qu.: 1.606      1st Qu.:17.850
## Median :0.3242      Median :1.1080      Median : 2.287      Median :24.530
## Mean   :0.4052      Mean   :1.2169      Mean   : 2.866      Mean   :40.337
## 3rd Qu.:0.4789      3rd Qu.:1.4740      3rd Qu.: 3.357      3rd Qu.:45.190
## Max.   :2.8730      Max.   :4.8850      Max.   :21.980      Max.   :542.200
## smoothness_se compactness_se concavity_se concave.points_se
```

```
## Min.      :0.001713    Min.      :0.002252    Min.      :0.000000    Min.      :0.000000
## 1st Qu.:0.005169    1st Qu.:0.013080    1st Qu.:0.01509    1st Qu.:0.007638
## Median :0.006380    Median :0.020450    Median :0.02589    Median :0.010930
## Mean   :0.007041    Mean   :0.025478    Mean   :0.03189    Mean   :0.011796
## 3rd Qu.:0.008146    3rd Qu.:0.032450    3rd Qu.:0.04205    3rd Qu.:0.014710
## Max.   :0.031130    Max.   :0.135400    Max.   :0.39600    Max.   :0.052790
## symmetry_se fractal_dimension_se radius_worst texture_worst
## Min.      :0.007882    Min.      :0.0008948    Min.      : 7.93    Min.      :12.02
## 1st Qu.:0.015160    1st Qu.:0.0022480    1st Qu.:13.01    1st Qu.:21.08
## Median :0.018730    Median :0.0031870    Median :14.97    Median :25.41
## Mean   :0.020542    Mean   :0.0037949    Mean   :16.27    Mean   :25.68
## 3rd Qu.:0.023480    3rd Qu.:0.0045580    3rd Qu.:18.79    3rd Qu.:29.72
## Max.   :0.078950    Max.   :0.0298400    Max.   :36.04    Max.   :49.54
## perimeter_worst area_worst smoothness_worst compactness_worst
## Min.      : 50.41    Min.      : 185.2    Min.      :0.07117    Min.      :0.02729
## 1st Qu.: 84.11    1st Qu.: 515.3    1st Qu.:0.11660    1st Qu.:0.14720
## Median : 97.66    Median : 686.5    Median :0.13130    Median :0.21190
## Mean   :107.26    Mean   : 880.6    Mean   :0.13237    Mean   :0.25427
## 3rd Qu.:125.40    3rd Qu.:1084.0    3rd Qu.:0.14600    3rd Qu.:0.33910
## Max.   :251.20    Max.   :4254.0    Max.   :0.22260    Max.   :1.05800
## concavity_worst concave.points_worst symmetry_worst fractal_dimension_worst
## Min.      :0.0000    Min.      :0.000000    Min.      :0.1565    Min.      :0.05504
## 1st Qu.:0.1145    1st Qu.:0.06493    1st Qu.:0.2504    1st Qu.:0.07146
## Median :0.2267    Median :0.09993    Median :0.2822    Median :0.08004
## Mean   :0.2722    Mean   :0.11461    Mean   :0.2901    Mean   :0.08395
## 3rd Qu.:0.3829    3rd Qu.:0.16140    3rd Qu.:0.3179    3rd Qu.:0.09208
## Max.   :1.2520    Max.   :0.29100    Max.   :0.6638    Max.   :0.20750
## X
## Mode:logical
## NA's:569
##
##
##
##
```

```
#Showing structure of the dataset
str(data)
```

```
## 'data.frame': 569 obs. of 33 variables:
## $ id : int 842302 842517 84300903 84348301 84358402 843786 844359 84458202 844...
## $ diagnosis : chr "M" "M" "M" "M" ...
## $ radius_mean : num 18 20.6 19.7 11.4 20.3 ...
## $ texture_mean : num 10.4 17.8 21.2 20.4 14.3 ...
## $ perimeter_mean : num 122.8 132.9 130 77.6 135.1 ...
## $ area_mean : num 1001 1326 1203 386 1297 ...
## $ smoothness_mean : num 0.1184 0.0847 0.1096 0.1425 0.1003 ...
## $ compactness_mean : num 0.2776 0.0786 0.1599 0.2839 0.1328 ...
## $ concavity_mean : num 0.3001 0.0869 0.1974 0.2414 0.198 ...
## $ concave.points_mean : num 0.1471 0.0702 0.1279 0.1052 0.1043 ...
## $ symmetry_mean : num 0.242 0.181 0.207 0.26 0.181 ...
## $ fractal_dimension_mean : num 0.0787 0.0567 0.06 0.0974 0.0588 ...
## $ radius_se : num 1.095 0.543 0.746 0.496 0.757 ...
## $ texture_se : num 0.905 0.734 0.787 1.156 0.781 ...
## $ perimeter_se : num 8.59 3.4 4.58 3.44 5.44 ...
```

```
## $ area_se : num 153.4 74.1 94 27.2 94.4 ...
## $ smoothness_se : num 0.0064 0.00522 0.00615 0.00911 0.01149 ...
## $ compactness_se : num 0.049 0.0131 0.0401 0.0746 0.0246 ...
## $ concavity_se : num 0.0537 0.0186 0.0383 0.0566 0.0569 ...
## $ concave.points_se : num 0.0159 0.0134 0.0206 0.0187 0.0188 ...
## $ symmetry_se : num 0.03 0.0139 0.0225 0.0596 0.0176 ...
## $ fractal_dimension_se : num 0.00619 0.00353 0.00457 0.00921 0.00511 ...
## $ radius_worst : num 25.4 25 23.6 14.9 22.5 ...
## $ texture_worst : num 17.3 23.4 25.5 26.5 16.7 ...
## $ perimeter_worst : num 184.6 158.8 152.5 98.9 152.2 ...
## $ area_worst : num 2019 1956 1709 568 1575 ...
## $ smoothness_worst : num 0.162 0.124 0.144 0.21 0.137 ...
## $ compactness_worst : num 0.666 0.187 0.424 0.866 0.205 ...
## $ concavity_worst : num 0.712 0.242 0.45 0.687 0.4 ...
## $ concave.points_worst : num 0.265 0.186 0.243 0.258 0.163 ...
## $ symmetry_worst : num 0.46 0.275 0.361 0.664 0.236 ...
## $ fractal_dimension_worst: num 0.1189 0.089 0.0876 0.173 0.0768 ...
## $ X : logi NA NA NA NA NA NA ...
```

```
#Checking for missing values
colSums(is.na(data))
```

```
##          id          diagnosis          radius_mean
##          0              0              0
## texture_mean    perimeter_mean          area_mean
##          0              0              0
## smoothness_mean compactness_mean    concavity_mean
##          0              0              0
## concave.points_mean    symmetry_mean fractal_dimension_mean
##          0              0              0
##          radius_se          texture_se          perimeter_se
##          0              0              0
##          area_se          smoothness_se    compactness_se
##          0              0              0
##          concavity_se    concave.points_se    symmetry_se
##          0              0              0
## fractal_dimension_se    radius_worst          texture_worst
##          0              0              0
##          perimeter_worst    area_worst          smoothness_worst
##          0              0              0
##          compactness_worst    concavity_worst    concave.points_worst
##          0              0              0
##          symmetry_worst fractal_dimension_worst          X
##          0              0              569
```

```
#569 total missing values were found in X variable
```

```
#Treating Missing Values
```

```
#Sub-setting out the X variable and saving in a new dataframe
data_clean <- data[,1:32]
```

```
#Checking if new data has any missing values
colSums(is.na(data_clean))
```

```
##           id           diagnosis           radius_mean
##           0             0             0
## texture_mean    perimeter_mean    area_mean
##           0             0             0
## smoothness_mean compactness_mean    concavity_mean
##           0             0             0
## concave.points_mean    symmetry_mean    fractal_dimension_mean
##           0             0             0
## radius_se         texture_se         perimeter_se
##           0             0             0
## area_se          smoothness_se        compactness_se
##           0             0             0
## concavity_se      concave.points_se    symmetry_se
##           0             0             0
## fractal_dimension_se    radius_worst    texture_worst
##           0             0             0
## perimeter_worst        area_worst        smoothness_worst
##           0             0             0
## compactness_worst      concavity_worst    concave.points_worst
##           0             0             0
## symmetry_worst    fractal_dimension_worst
##           0             0
```

```
#no missing values found
```

```
#Removing the ID column
```

```
wbcd <- data_clean[,2:32]
```

```
#converting the diagnosis variable into a factor
```

```
wbcd$diagnosis <- factor(ifelse(wbcd$diagnosis=='B',"Benign","Malignant"))
```

```
#now converting diagnosis as a double - 1 if Malignant and 0 if Benign
```

```
wbcd_n <- wbcd %>%
  mutate_at(vars(diagnosis), as.double) %>%
  mutate(diagnosis = diagnosis - 1)
```

```
#checking the structure of the new dataframe
```

```
str(wbcd_n)
```

```
## 'data.frame':   569 obs. of  31 variables:
## $ diagnosis      : num  1 1 1 1 1 1 1 1 1 1 ...
## $ radius_mean    : num  18 20.6 19.7 11.4 20.3 ...
## $ texture_mean   : num  10.4 17.8 21.2 20.4 14.3 ...
## $ perimeter_mean : num  122.8 132.9 130 77.6 135.1 ...
## $ area_mean      : num  1001 1326 1203 386 1297 ...
## $ smoothness_mean : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
## $ compactness_mean : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
## $ concavity_mean  : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
## $ concave.points_mean : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
## $ symmetry_mean   : num  0.242 0.181 0.207 0.26 0.181 ...
## $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
## $ radius_se       : num  1.095 0.543 0.746 0.496 0.757 ...
```

```
## $ texture_se : num 0.905 0.734 0.787 1.156 0.781 ...
## $ perimeter_se : num 8.59 3.4 4.58 3.44 5.44 ...
## $ area_se : num 153.4 74.1 94 27.2 94.4 ...
## $ smoothness_se : num 0.0064 0.00522 0.00615 0.00911 0.01149 ...
## $ compactness_se : num 0.049 0.0131 0.0401 0.0746 0.0246 ...
## $ concavity_se : num 0.0537 0.0186 0.0383 0.0566 0.0569 ...
## $ concave.points_se : num 0.0159 0.0134 0.0206 0.0187 0.0188 ...
## $ symmetry_se : num 0.03 0.0139 0.0225 0.0596 0.0176 ...
## $ fractal_dimension_se : num 0.00619 0.00353 0.00457 0.00921 0.00511 ...
## $ radius_worst : num 25.4 25 23.6 14.9 22.5 ...
## $ texture_worst : num 17.3 23.4 25.5 26.5 16.7 ...
## $ perimeter_worst : num 184.6 158.8 152.5 98.9 152.2 ...
## $ area_worst : num 2019 1956 1709 568 1575 ...
## $ smoothness_worst : num 0.162 0.124 0.144 0.21 0.137 ...
## $ compactness_worst : num 0.666 0.187 0.424 0.866 0.205 ...
## $ concavity_worst : num 0.712 0.242 0.45 0.687 0.4 ...
## $ concave.points_worst : num 0.265 0.186 0.243 0.258 0.163 ...
## $ symmetry_worst : num 0.46 0.275 0.361 0.664 0.236 ...
## $ fractal_dimension_worst: num 0.1189 0.089 0.0876 0.173 0.0768 ...
```

```
#all numeric variables
```

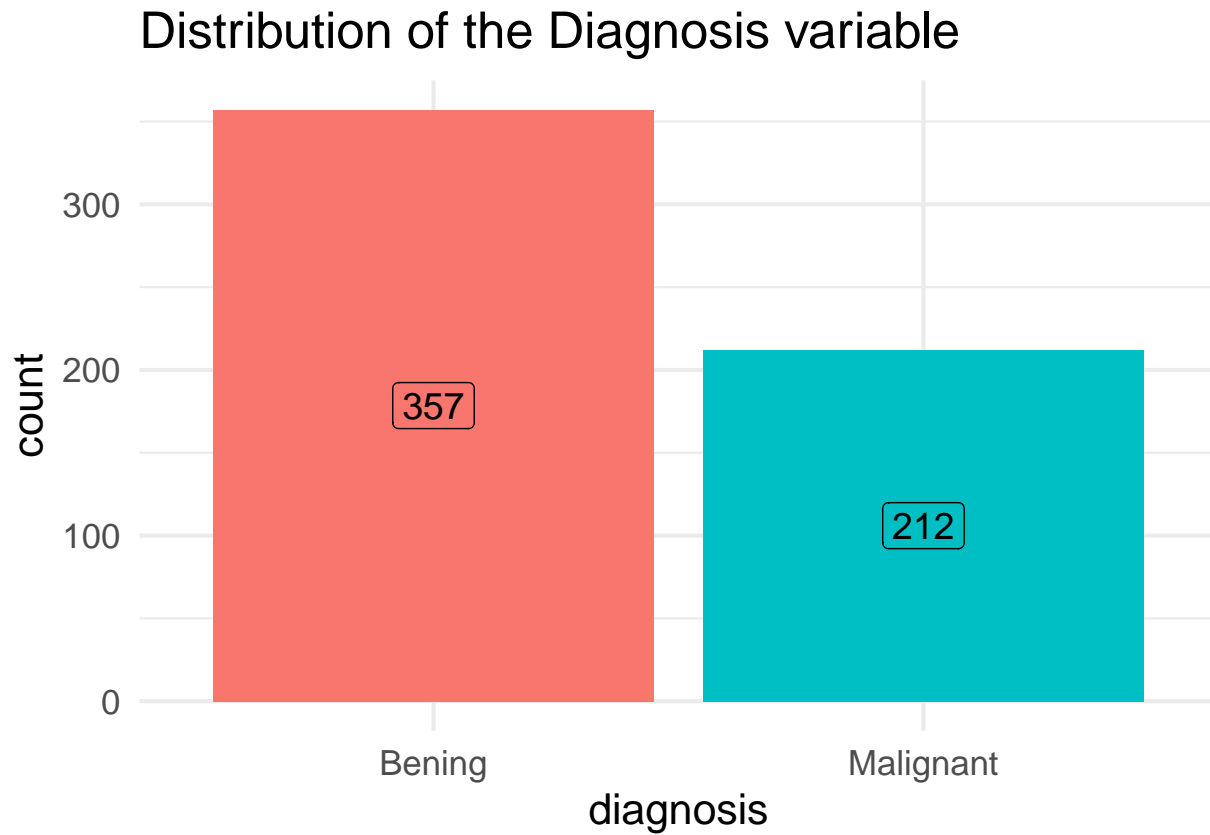
Now we know the data is clean and we can run some Visualization and Analysis

## Data Visualization

Distribution of the diagnosis variable -

```
#visualizing the diagnosis variable
p1 <- ggplot(wbcd, aes(x = diagnosis, fill = diagnosis)) +
  geom_bar(stat = "count", position = "stack", show.legend = FALSE) +
  theme_minimal(base_size = 16) +
  geom_label(stat = "count", aes(label = after_stat(count)), position = position_stack(vjust = 0.5),
    size = 5, show.legend = FALSE)

p1 +
  ggtitle("Distribution of the Diagnosis variable")
```



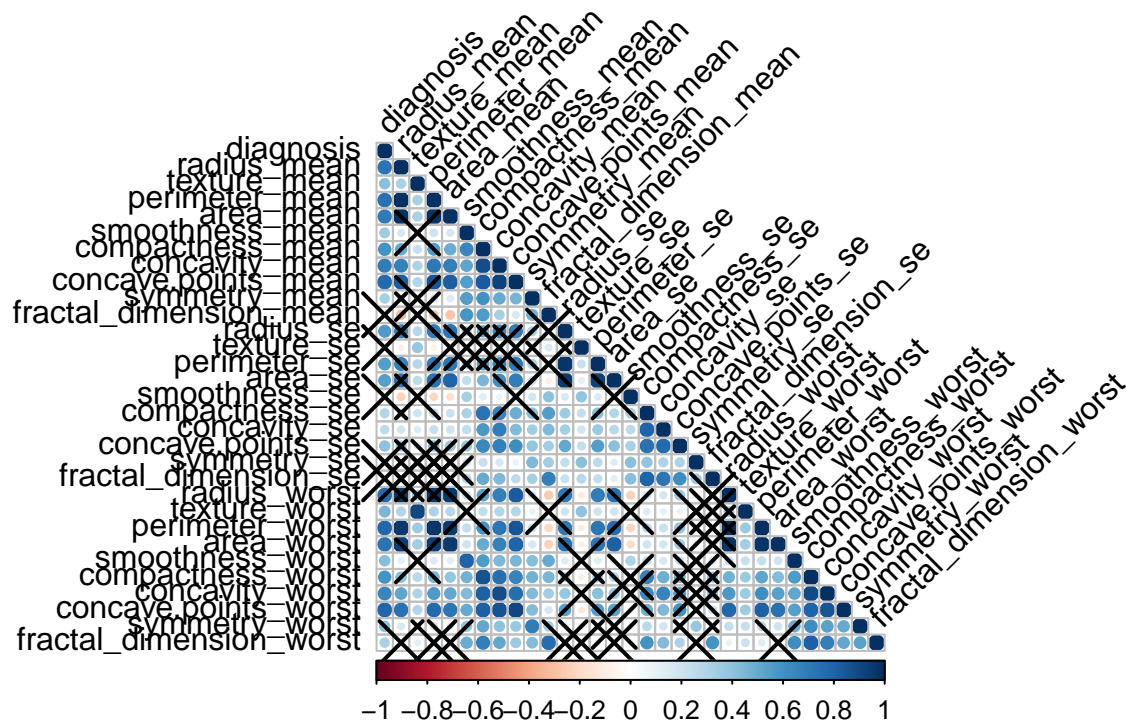
After looking at the distribution we notice that the diagnosis variable is biased.

### Correlation Analysis

In this correlation matrix, correlation coefficients which has a p-value less than 0.05 are marked with a cross (which means they are significant).

```
#checking the correlation matrix
cor(wbcd_n) %>%
  corplot(method = "circle", type = "lower", tl.col = "black", tl.srt = 45, p.mat = cor.mtest(wbcd_n)$
```





We notice that many of the independent variables are very strongly correlated, which suggests that there is multicollinearity.

## Descriptive Analysis

```
#showing descriptive analysis for the numeric dataframe
describe(wbcd_n)
```

##	vars	n	mean	sd	median	trimmed	mad	min
## diagnosis	1	569	0.37	0.48	0.00	0.34	0.00	0.00
## radius_mean	2	569	14.13	3.52	13.37	13.82	2.82	6.98
## texture_mean	3	569	19.29	4.30	18.84	19.04	4.17	9.71
## perimeter_mean	4	569	91.97	24.30	86.24	89.74	18.84	43.79
## area_mean	5	569	654.89	351.91	551.10	606.13	227.28	143.50
## smoothness_mean	6	569	0.10	0.01	0.10	0.10	0.01	0.05
## compactness_mean	7	569	0.10	0.05	0.09	0.10	0.05	0.02
## concavity_mean	8	569	0.09	0.08	0.06	0.08	0.06	0.00
## concave.points_mean	9	569	0.05	0.04	0.03	0.04	0.03	0.00
## symmetry_mean	10	569	0.18	0.03	0.18	0.18	0.03	0.11
## fractal_dimension_mean	11	569	0.06	0.01	0.06	0.06	0.01	0.05
## radius_se	12	569	0.41	0.28	0.32	0.36	0.16	0.11
## texture_se	13	569	1.22	0.55	1.11	1.16	0.47	0.36
## perimeter_se	14	569	2.87	2.02	2.29	2.51	1.14	0.76
## area_se	15	569	40.34	45.49	24.53	31.69	13.63	6.80
## smoothness_se	16	569	0.01	0.00	0.01	0.01	0.00	0.00
## compactness_se	17	569	0.03	0.02	0.02	0.02	0.01	0.00
## concavity_se	18	569	0.03	0.03	0.03	0.03	0.02	0.00
## concave.points_se	19	569	0.01	0.01	0.01	0.01	0.01	0.00
## symmetry_se	20	569	0.02	0.01	0.02	0.02	0.01	0.01

## fractal_dimension_se	21	569	0.00	0.00	0.00	0.00	0.00	0.00
## radius_worst	22	569	16.27	4.83	14.97	15.73	3.65	7.93
## texture_worst	23	569	25.68	6.15	25.41	25.39	6.42	12.02
## perimeter_worst	24	569	107.26	33.60	97.66	103.42	25.01	50.41
## area_worst	25	569	880.58	569.36	686.50	788.02	319.65	185.20
## smoothness_worst	26	569	0.13	0.02	0.13	0.13	0.02	0.07
## compactness_worst	27	569	0.25	0.16	0.21	0.23	0.13	0.03
## concavity_worst	28	569	0.27	0.21	0.23	0.25	0.20	0.00
## concave.points_worst	29	569	0.11	0.07	0.10	0.11	0.07	0.00
## symmetry_worst	30	569	0.29	0.06	0.28	0.28	0.05	0.16
## fractal_dimension_worst	31	569	0.08	0.02	0.08	0.08	0.01	0.06
##			max	range	skew	kurtosis	se	
## diagnosis		1.00	1.00	0.53	-1.73	0.02		
## radius_mean		28.11	21.13	0.94	0.81	0.15		
## texture_mean		39.28	29.57	0.65	0.73	0.18		
## perimeter_mean		188.50	144.71	0.99	0.94	1.02		
## area_mean		2501.00	2357.50	1.64	3.59	14.75		
## smoothness_mean		0.16	0.11	0.45	0.82	0.00		
## compactness_mean		0.35	0.33	1.18	1.61	0.00		
## concavity_mean		0.43	0.43	1.39	1.95	0.00		
## concave.points_mean		0.20	0.20	1.17	1.03	0.00		
## symmetry_mean		0.30	0.20	0.72	1.25	0.00		
## fractal_dimension_mean		0.10	0.05	1.30	2.95	0.00		
## radius_se		2.87	2.76	3.07	17.45	0.01		
## texture_se		4.88	4.52	1.64	5.26	0.02		
## perimeter_se		21.98	21.22	3.43	21.12	0.08		
## area_se		542.20	535.40	5.42	48.59	1.91		
## smoothness_se		0.03	0.03	2.30	10.32	0.00		
## compactness_se		0.14	0.13	1.89	5.02	0.00		
## concavity_se		0.40	0.40	5.08	48.24	0.00		
## concave.points_se		0.05	0.05	1.44	5.04	0.00		
## symmetry_se		0.08	0.07	2.18	7.78	0.00		
## fractal_dimension_se		0.03	0.03	3.90	25.94	0.00		
## radius_worst		36.04	28.11	1.10	0.91	0.20		
## texture_worst		49.54	37.52	0.50	0.20	0.26		
## perimeter_worst		251.20	200.79	1.12	1.04	1.41		
## area_worst		4254.00	4068.80	1.85	4.32	23.87		
## smoothness_worst		0.22	0.15	0.41	0.49	0.00		
## compactness_worst		1.06	1.03	1.47	2.98	0.01		
## concavity_worst		1.25	1.25	1.14	1.57	0.01		
## concave.points_worst		0.29	0.29	0.49	-0.55	0.00		
## symmetry_worst		0.66	0.51	1.43	4.37	0.00		
## fractal_dimension_worst		0.21	0.15	1.65	5.16	0.00		

We see that radius\_se (3.07), perimeter\_se (3.43), area\_se(5.42), concavity\_se (5.08) and fractal\_dimension\_se (3.90) are some of the highly skewed variables.

We see observe that texture\_worse(0.50) and diagnosis(0.53) are approximately symmetric.

The highest mean values were found for area\_mean (654.89), area\_worst (880.58) and perimeter\_worst (107.26).

The lowest mean values was for fractal\_dimension\_se (0.00).

## Statstical Analysis

Before running Factor Analysis, we need to check for the factoribility of the dataset by running the below tests

```
#Testing KMO Sampling Adequacy  
KMO(wbcd_n)
```

```
## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = wbcd_n)  
## Overall MSA = 0.84  
## MSA for each item =  
##           diagnosis      radius_mean      texture_mean  
##           0.99           0.84           0.66  
##           perimeter_mean      area_mean      smoothness_mean  
##           0.86           0.87           0.82  
##           compactness_mean      concavity_mean      concave.points_mean  
##           0.88           0.90           0.91  
##           symmetry_mean      fractal_dimension_mean      radius_se  
##           0.83           0.83           0.84  
##           texture_se      perimeter_se      area_se  
##           0.49           0.85           0.86  
##           smoothness_se      compactness_se      concavity_se  
##           0.65           0.87           0.83  
##           concave.points_se      symmetry_se      fractal_dimension_se  
##           0.84           0.58           0.81  
##           radius_worst      texture_worst      perimeter_worst  
##           0.83           0.62           0.89  
##           area_worst      smoothness_worst      compactness_worst  
##           0.83           0.76           0.86  
##           concavity_worst      concave.points_worst      symmetry_worst  
##           0.91           0.90           0.70  
## fractal_dimension_worst  
##           0.82
```

```
#Overall MSA = 0.84
```

```
#Testing Bartlett's Test of Sphericity  
bart_spher(wbcd_n)
```

```
## Bartlett's Test of Sphericity  
##  
## Call: bart_spher(x = wbcd_n)  
##  
##      X2 = 40167.506  
##      df = 465  
## p-value < 2.22e-16
```

```
#p-value < 2.22e-16
```

```
#Checking the Cronbach's Alpha  
cronbach.alpha(wbcd_n)
```

```
##  
## Cronbach's alpha for the 'wbcd_n' data-set  
##  
## Items: 31  
## Sample units: 569  
## alpha: 0.585
```

```
#raw_alpha = 0.58
```

The KMO Sampling Test gave us a MSA value of 0.84, which confirms that the sample used is sufficient. We see the the Bartlett's Test of Sphericity has a p-value of less than 0.05 demonstrating that the correlation matrix is not an identity matrix, therefore providing justification to use Factor Analysis. Usual we accept a sample only when the Cronbach's alpha is greater than 0.6 but I am making an exception with my dataset.

## Parallel Analysis

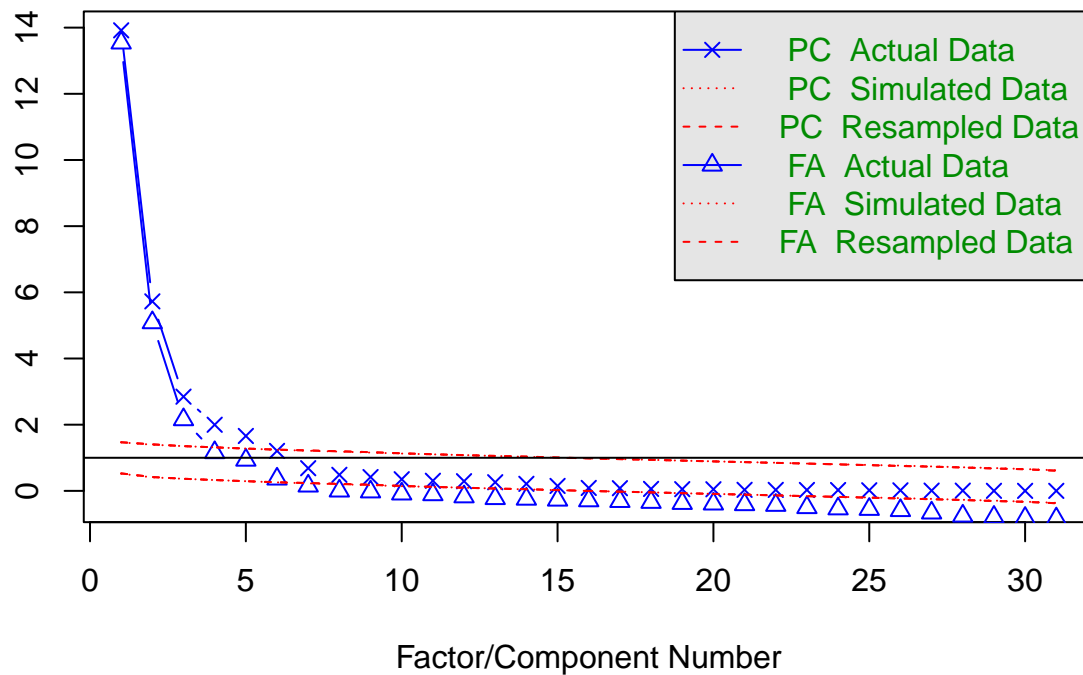
Parallel Analysis can helps us determine how many factors we need to use in factor analysis. Parallel Analysis can be used as a guess and not a final answer, it gives us something to get started.

```
#running parallel analysis  
comp <- fa.parallel(wbcd_n)
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :  
## The estimated weights for the factor scores are probably incorrect. Try a  
## different factor score estimation method.
```

eigenvalues of principal components and factor analysis

## Parallel Analysis Scree Plots



## Parallel analysis suggests that the number of factors = 6 and the number of components = 5

comp

```
## Call: fa.parallel(x = wbcd_n)
## Parallel analysis suggests that the number of factors = 6 and the number of components = 5
##
## Eigen Values of
## Original factors Resampled data Simulated data Original components
## 1 13.54 0.53 0.52 13.92
## 2 5.08 0.42 0.41 5.73
## 3 2.15 0.37 0.36 2.85
## 4 1.16 0.33 0.33 2.00
## 5 0.93 0.29 0.29 1.66
## 6 0.36 0.26 0.26 1.21
## Resampled components Simulated components
## 1 1.47 1.47
## 2 1.41 1.39
## 3 1.35 1.35
## 4 1.32 1.31
## 5 1.28 1.28
## 6 1.25 1.24
```

Parallel Analysis suggests that we should be using 6 factors but let's take a look at eigenvalues which are greater than 1

```
#checking for eigenvalues which are greater than 1
sum(comp$fa.values>1)
```

```
## [1] 4
```

Now since we know that there are four factors which have eigenvalues that are greater than 1, we will start by using 4 factors

## Common Factor Analysis

```
#Conducting factor analysis
fit = factanal(wbcd_n[,-1], 4, rotation = "varimax", lower = 0.1)
print(fit$loadings, cutoff=0.5, sort=T)
```

```
##
## Loadings:
##
## Factor1 Factor2 Factor3 Factor4
## radius_mean 0.943
## perimeter_mean 0.941
## area_mean 0.962
## concave.points_mean 0.782 0.548
## radius_se 0.828
## perimeter_se 0.814 0.503
## area_se 0.874
## radius_worst 0.939
## perimeter_worst 0.932
## area_worst 0.945
## smoothness_mean 0.596
## compactness_mean 0.828
## concavity_mean 0.631 0.690
## symmetry_mean 0.545
## fractal_dimension_mean 0.705
## compactness_se 0.686
## concavity_se 0.582
## fractal_dimension_se 0.546 0.545
## smoothness_worst 0.641
## compactness_worst 0.870
## concavity_worst 0.812
## concave.points_worst 0.660 0.671
## symmetry_worst 0.596
## fractal_dimension_worst 0.870
## texture_se 0.528
## smoothness_se 0.598
## symmetry_se 0.569
## texture_mean 0.914
## texture_worst 0.918
## concave.points_se
##
## Factor1 Factor2 Factor3 Factor4
## SS loadings 9.854 7.634 3.018 2.077
## Proportion Var 0.328 0.254 0.101 0.069
## Cumulative Var 0.328 0.583 0.684 0.753
```

```
#Displaying the summary  
summary(fit)
```

```
##           Length Class      Mode  
## converged      1  -none-   logical  
## loadings     120  loadings numeric  
## uniquenesses  30  -none-   numeric  
## correlation   900  -none-   numeric  
## criteria       3  -none-   numeric  
## factors        1  -none-   numeric  
## dof            1  -none-   numeric  
## method         1  -none-   character  
## rotmat         16  -none-   numeric  
## STATISTIC      1  -none-   numeric  
## PVAL           1  -none-   numeric  
## n.obs          1  -none-   numeric  
## call           5  -none-   call
```

Interpretation -

The variables with high factor loadings in Factor 1 are radius, parameter and area which are related to the size of the nucleus. The larger these variables are, the larger these values become.

The variables with high factor loadings in Factor 2 are those related to the distortion of the contour of the cell nucleus, such as fractal\_dimension, smoothness, compactness and concavity.

The variable with highest factor loading in Factor 3 is smoothness\_se which drives the Factor 3.

The variables with high factor loadings in Factor 4 are mainly related to texture and larger these values are, the larger these values become.

We also observe some cross-loadings between Factor 1 and Factor 2 and Factor 3 and Factor 4.

The four factors explain 75% of the variance.

Names of the components -

Factor 1 - size, as it speaks about how large a nuclei is

Factor 2 - distortion, as it describes the distorted cells outline

Factor 3 - variety, as it tells us the variety of cell nuclei

Factor 4 - texture, as it talks about the texture of the nuclei

## Conclusion

We arrive at a conclusion that there are four main characteristics which are needed to detect breast cancer and among these four characteristics, size is one of the most important characteristics to consider.

In actual medical practice, the degree of “nuclear atypia” of cells is used to classify the malignancy of breast cancer. The larger the cell nucleus, the more chromatin is increased and unevenly distributed, and the more distorted the nuclear outline, the more abnormal the cell is considered to be. [3]

The above mentioned study confirms are findings.

## References

1.Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. <https://doi.org/10.24432/C5DW2B>.

2.American Cancer Society (n.d.). Fine Needle Aspiration (FNA) of the Breast. Cancer.org. <https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html#:~:text=During%20a%20fine%20needle%20aspiration,needle%20biopsy%20is%20often>

3.Okudela K. (2014). An association between nuclear morphology and immunohistochemical expression of p53 and p16INK4A in lung cancer cells. Medical molecular morphology, 47(3), 130–136. <https://doi.org/10.1007/s00795-013-0052-x>