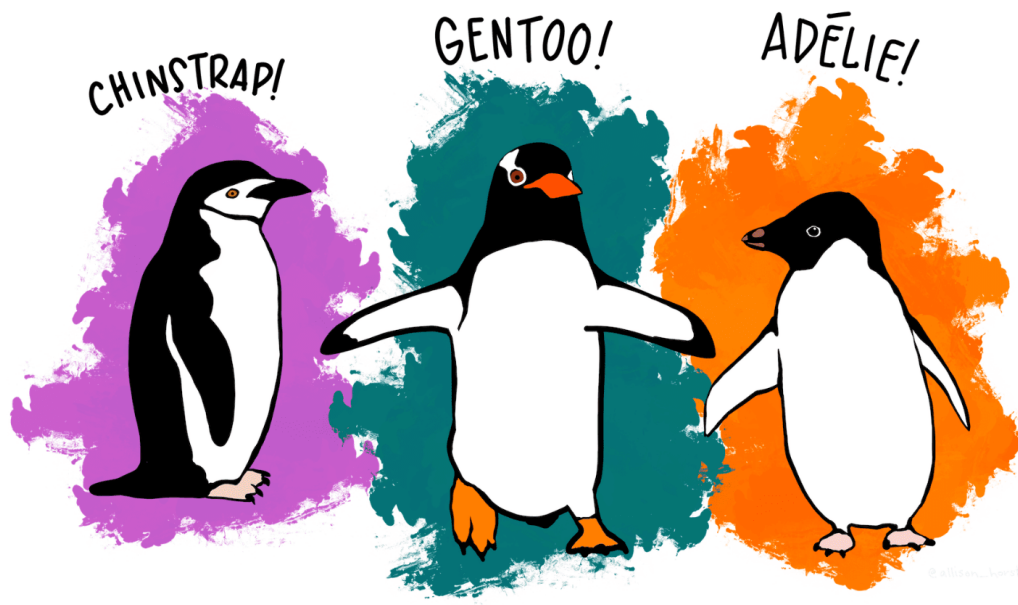


Data Preparation (Penguins)

Delete all variables

```
rm( list = ls() )
```

Meet the Palmer penguins



Data Preparation and first Insights

Import Data

- Download the data from Moodle
- Import the data using the function `read.csv()`
- Create a dataframe for it using the function `data.frame()`
- Have a look at the first rows using the function `head()`
- Have a look at the last rows using the function `tail()`

```
data = read.csv("C:/Users/Dari-Laptop/Desktop/FH Karnten - Master - AppDs/StatisticsAppDSLaptop/penguins.csv")
```

What are the column names of the data?

Use the function `names()`

```
names(data)
```

```
## [1] "studyName"      "Sample.Number"   "Species"
## [4] "Region"         "Island"          "Stage"
## [7] "Individual.ID"   "Clutch.Completion" "Date.Egg"
## [10] "Culmen.Length..mm." "Culmen.Depth..mm." "Flipper.Length..mm."
## [13] "Body.Mass..g."   "Gender"          "Delta.15.N..o.oo."
## [16] "Delta.13.C..o.oo." "Comments"
```

What are the data types of each column?

```
str(data)
```

```
## 'data.frame':   344 obs. of  17 variables:
## $ studyName      : chr  "PAL0708" "PAL0708" "PAL0708" "PAL0708" ...
## $ Sample.Number  : int   1  2  3  4  5  6  7  8  9 10 ...
## $ Species        : chr   "Adelie Penguin (Pygoscelis adeliae)" "Adelie Penguin (Pygoscelis adeliae)" ...
## $ Region         : chr   "Anvers" "Anvers" "Anvers" "Anvers" ...
## $ Island         : chr   "Torgersen" "Torgersen" "Torgersen" "Torgersen" ...
## $ Stage          : chr   "Adult, 1 Egg Stage" "Adult, 1 Egg Stage" "Adult, 1 Egg Stage" "Adult, 1 Egg Stage" ...
## $ Individual.ID   : chr   "N1A1" "N1A2" "N2A1" "N2A2" ...
## $ Clutch.Completion : chr   "Yes" "Yes" "Yes" "Yes" ...
## $ Date.Egg       : chr   "11/11/07" "11/11/07" "11/16/07" "11/16/07" ...
## $ Culmen.Length..mm. : num  39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
## $ Culmen.Depth..mm. : num  18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
## $ Flipper.Length..mm.: int   181 186 195 NA 193 190 181 195 193 190 ...
## $ Body.Mass..g.    : int   3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
## $ Gender          : chr   "MALE" "FEMALE" "FEMALE" "" ...
## $ Delta.15.N..o.oo. : num  NA 8.95 8.37 NA 8.77 ...
## $ Delta.13.C..o.oo. : num  NA -24.7 -25.3 NA -25.3 ...
## $ Comments        : chr   "Not enough blood for isotopes." "" "" "Adult not sampled." ...
```

Data types:

- Nominal:
 - studyName
 - Species
 - Region
 - Island
 - Gender
 - Comments
 - Clutch.Completion
- Ordinal:
 - Individual.ID
 - Date

- Discrete:
 - Body Mass
 - Sample.Number
 - Flipper.length
- Continuous:
 - Culmen.Length
 - Culmen.Depth
 - Delta.15
 - Delta13

Delete some columns

Delete the following columns since we won't need them in this class:

- studyName,
- Region,
- Stage,
- Clutch.Completion,
- Date.Egg,
- Delta.15.N..o.oo.,
- Delta.13.C..o.oo.,
- Comments

Use therefore the function `%in%`

```
drop <- c('studyName', 'Region', 'Stage', 'Clutch.Completion', 'Date.Egg', 'Delta.15.N..o.oo.',
          'Delta.13.C..o.oo.', 'Comments')

df = data[,!(names(data) %in% drop)]

head(df)
```

##	Sample.Number	Species	Island	Individual.ID
## 1	1	Adelie Penguin (Pygoscelis adeliae)	Torgersen	N1A1
## 2	2	Adelie Penguin (Pygoscelis adeliae)	Torgersen	N1A2
## 3	3	Adelie Penguin (Pygoscelis adeliae)	Torgersen	N2A1
## 4	4	Adelie Penguin (Pygoscelis adeliae)	Torgersen	N2A2
## 5	5	Adelie Penguin (Pygoscelis adeliae)	Torgersen	N3A1
## 6	6	Adelie Penguin (Pygoscelis adeliae)	Torgersen	N3A2

##	Culmen.Length..mm.	Culmen.Depth..mm.	Flipper.Length..mm.	Body.Mass..g.	Gender
## 1	39.1	18.7	181	3750	MALE
## 2	39.5	17.4	186	3800	FEMALE
## 3	40.3	18.0	195	3250	FEMALE
## 4	NA	NA	NA	NA	
## 5	36.7	19.3	193	3450	FEMALE
## 6	39.3	20.6	190	3650	MALE

Rearrange the data

Put the column “Individual.ID” to first column

```
reorder = c("Individual.ID", names(df[,names(df)!="Individual.ID"]))
reorder
```

```
## [1] "Individual.ID"      "Sample.Number"      "Species"
## [4] "Island"             "Culmen.Length..mm." "Culmen.Depth..mm."
## [7] "Flipper.Length..mm." "Body.Mass..g."      "Gender"
```

```
df <- df[, reorder]
head(df)
```

```
##   Individual.ID Sample.Number      Species      Island
## 1          N1A1             1 Adelie Penguin (Pygoscelis adeliae) Torgersen
## 2          N1A2             2 Adelie Penguin (Pygoscelis adeliae) Torgersen
## 3          N2A1             3 Adelie Penguin (Pygoscelis adeliae) Torgersen
## 4          N2A2             4 Adelie Penguin (Pygoscelis adeliae) Torgersen
## 5          N3A1             5 Adelie Penguin (Pygoscelis adeliae) Torgersen
## 6          N3A2             6 Adelie Penguin (Pygoscelis adeliae) Torgersen
##   Culmen.Length..mm. Culmen.Depth..mm. Flipper.Length..mm. Body.Mass..g. Gender
## 1                39.1                18.7                181            3750  MALE
## 2                39.5                17.4                186            3800 FEMALE
## 3                40.3                18.0                195            3250 FEMALE
## 4                 NA                 NA                 NA              NA
## 5                36.7                19.3                193            3450 FEMALE
## 6                39.3                20.6                190            3650  MALE
```

Rename columns

Rename the columns to

- IndividualID
- Species
- Island
- CulmenLength(mm)
- CulmenDepth(mm)
- FlipperLength(mm)
- BodyMass(g)
- Gender

```
## Rename
names(df)[1] = "IndividualID"
names(df)[2] = "Species"
names(df)[3] = "Island"
names(df)[4] = "CulmenLength(mm)"
names(df)[5] = "CulmenDepth(mm)"
names(df)[6] = "FlipperLength(mm)"
names(df)[7] = "BodyMass(g)"
names(df)[8] = "Gender"

head(df)
```

```
##   IndividualID Species      Island CulmenLength(mm)
```

```
## 1      N1A1      1 Adelie Penguin (Pygoscelis adeliae)      Torgersen
## 2      N1A2      2 Adelie Penguin (Pygoscelis adeliae)      Torgersen
## 3      N2A1      3 Adelie Penguin (Pygoscelis adeliae)      Torgersen
## 4      N2A2      4 Adelie Penguin (Pygoscelis adeliae)      Torgersen
## 5      N3A1      5 Adelie Penguin (Pygoscelis adeliae)      Torgersen
## 6      N3A2      6 Adelie Penguin (Pygoscelis adeliae)      Torgersen
##      CulmenDepth(mm) FlipperLength(mm) BodyMass(g) Gender Gender
## 1              39.1              18.7              181    3750    MALE
## 2              39.5              17.4              186    3800 FEMALE
## 3              40.3              18.0              195    3250 FEMALE
## 4              NA              NA              NA      NA
## 5              36.7              19.3              193    3450 FEMALE
## 6              39.3              20.6              190    3650    MALE
```

What are the three types of Species?

Use therefore the function `unique()`

```
# unique
```

```
unique(df$Species)
```

```
##      [1]      1      2      3      4      5      6      7      8      9     10     11     12     13     14     15     16     17     18
##     [19]     19     20     21     22     23     24     25     26     27     28     29     30     31     32     33     34     35     36
##     [37]     37     38     39     40     41     42     43     44     45     46     47     48     49     50     51     52     53     54
##     [55]     55     56     57     58     59     60     61     62     63     64     65     66     67     68     69     70     71     72
##     [73]     73     74     75     76     77     78     79     80     81     82     83     84     85     86     87     88     89     90
##     [91]     91     92     93     94     95     96     97     98     99    100    101    102    103    104    105    106    107    108
##    [109]    109    110    111    112    113    114    115    116    117    118    119    120    121    122    123    124    125    126
##    [127]    127    128    129    130    131    132    133    134    135    136    137    138    139    140    141    142    143    144
##    [145]    145    146    147    148    149    150    151    152
```

Rename the Species types

Rename the types to

- Adele
- Chinstrap
- Gentoo

```
# rename species
```

```
df$Species[df$Species=="Adelie Penguin (Pygoscelis adeliae)"] = "Adele"
df$Species[df$Species=="Chinstrap penguin (Pygoscelis antarctica)"] = "Chinstrap"
df$Species[df$Species=="Gentoo penguin (Pygoscelis papua)"] = "Gentoo"

unique(df$Species)
```

```
##      [1] "1"    "2"    "3"    "4"    "5"    "6"    "7"    "8"    "9"    "10"   "11"   "12"
##     [13] "13"   "14"   "15"   "16"   "17"   "18"   "19"   "20"   "21"   "22"   "23"   "24"
```

```
## [25] "25" "26" "27" "28" "29" "30" "31" "32" "33" "34" "35" "36"
## [37] "37" "38" "39" "40" "41" "42" "43" "44" "45" "46" "47" "48"
## [49] "49" "50" "51" "52" "53" "54" "55" "56" "57" "58" "59" "60"
## [61] "61" "62" "63" "64" "65" "66" "67" "68" "69" "70" "71" "72"
## [73] "73" "74" "75" "76" "77" "78" "79" "80" "81" "82" "83" "84"
## [85] "85" "86" "87" "88" "89" "90" "91" "92" "93" "94" "95" "96"
## [97] "97" "98" "99" "100" "101" "102" "103" "104" "105" "106" "107" "108"
## [109] "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" "119" "120"
## [121] "121" "122" "123" "124" "125" "126" "127" "128" "129" "130" "131" "132"
## [133] "133" "134" "135" "136" "137" "138" "139" "140" "141" "142" "143" "144"
## [145] "145" "146" "147" "148" "149" "150" "151" "152"
```

```
head(df)
```

```
##   IndividualID Species                Island CulmenLength(mm)
## 1         N1A1      1 Adelie Penguin (Pygoscelis adeliae)    Torgersen
## 2         N1A2      2 Adelie Penguin (Pygoscelis adeliae)    Torgersen
## 3         N2A1      3 Adelie Penguin (Pygoscelis adeliae)    Torgersen
## 4         N2A2      4 Adelie Penguin (Pygoscelis adeliae)    Torgersen
## 5         N3A1      5 Adelie Penguin (Pygoscelis adeliae)    Torgersen
## 6         N3A2      6 Adelie Penguin (Pygoscelis adeliae)    Torgersen
##   CulmenDepth(mm) FlipperLength(mm) BodyMass(g) Gender Gender
## 1              39.1              18.7         181   3750  MALE
## 2              39.5              17.4         186   3800 FEMALE
## 3              40.3              18.0         195   3250 FEMALE
## 4               NA               NA          NA     NA
## 5              36.7              19.3         193   3450 FEMALE
## 6              39.3              20.6         190   3650  MALE
```

Missing Values

How many rows contain missing values?

```
summary(df)
```

```
##   IndividualID      Species      Island      CulmenLength(mm)
## Length:344      Length:344      Length:344      Length:344
## Class :character Class :character Class :character Class :character
## Mode  :character Mode  :character Mode  :character Mode  :character
##
##
##
##   CulmenDepth(mm) FlipperLength(mm) BodyMass(g)      Gender
## Min.   :32.10   Min.   :13.10   Min.   :172.0   Min.   :2700
## 1st Qu.:39.23   1st Qu.:15.60   1st Qu.:190.0   1st Qu.:3550
## Median :44.45   Median :17.30   Median :197.0   Median :4050
## Mean   :43.92   Mean   :17.15   Mean   :200.9   Mean   :4202
## 3rd Qu.:48.50   3rd Qu.:18.70   3rd Qu.:213.0   3rd Qu.:4750
## Max.   :59.60   Max.   :21.50   Max.   :231.0   Max.   :6300
## NA's    :2      NA's    :2      NA's    :2      NA's    :2
```

```
##      Gender
## Length:344
## Class :character
## Mode  :character
##
##
##
##
```

How many rows have empty entries in the qualitative columns?

Data Imputation

What about data imputation? Check the rows with NA values. Can you apply data imputation on it?

Create Dataframe DataWithoutGender

- Select the rows where the gender is unknown and create a new data set for it. Name it “PenguinsWithoutGender”
- Delete the rows with NA values in the quantitative columns by using the function `na.omit()`
- replace “.” by “ ” in the column Gender
- save it to a csv file called “DataWithoutGender.csv” (we will classify the missing gender data later).

For the **csv saving**:

- Use either `write.csv()` or `write.csv2()` depending on your system’s language: if english use `write.csv()` if german use `write.csv2()`
- use `row.names = FALSE` to exclude the row indices → for a smoother workflow

Create Dataframe PenguinsWithoutMissingValues

- Delete the rows with NA values in the quantitative columns by using the function `na.omit()`
- Delete the rows with missing values in the gender column
- save it to a csv file called “PenguinsWithoutMissingValues.csv”

For the **csv saving**:

- Use either `write.csv()` or `write.csv2()` depending on your system’s language: if english use `write.csv()` if german use `write.csv2()`
- use `row.names = FALSE` to exclude the row indices → for a smoother workflow