Lab 11: Data Cleaning Skills

Yasmine Pascal

04/22/2022

## Instructions

This document is the template you will use to type up your responses to the lab exercises that you can find on Canvas.

To complete the lab type your BRIEF answers and the R code in the spaces provided below, according to these guidelines:

* In general, if an exercise is numbered you will submit the R code that you used to answer the question in the gray area between the three ` marks.
* If you need to submit an answer in addition to the code type this between the line that starts #### and the beginning of the gray area for the code.
* Do not ever type the View() command in a lab report. This works only in RStudio on your computer, so don’t include that command in your lab report or you will not be able to compile the document, a process called *knitting*.
* To produce a document that you can submit on Canvas just click on the Knit button in the upper left quadrant of RStudio.
* We recommend pressing the Knit button after entering each answer: it’s very common to get errors, and they’re easier to fix if you can easily isolate the code that caused the error.
* The Knit button will create two files: one is a .Rmd file (R Markdown) and the other is an .docx file. You will need to submit both these documents for the lab assignment on Canvas.

## Preliminaries

#### Exercise 1

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.6 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.4 ✓ stringr 1.4.0  
## ✓ readr 2.1.1 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggformula)

## Loading required package: ggstance

##   
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':  
##   
## geom\_errorbarh, GeomErrorbarh

## Loading required package: scales

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

## Loading required package: ggridges

##   
## New to ggformula? Try the tutorials:   
## learnr::run\_tutorial("introduction", package = "ggformula")  
## learnr::run\_tutorial("refining", package = "ggformula")

library(mosaic)

## Registered S3 method overwritten by 'mosaic':  
## method from   
## fortify.SpatialPolygonsDataFrame ggplot2

##   
## The 'mosaic' package masks several functions from core packages in order to add   
## additional features. The original behavior of these functions should not be affected by this.

##   
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':  
##   
## mean

## The following object is masked from 'package:scales':  
##   
## rescale

## The following objects are masked from 'package:dplyr':  
##   
## count, do, tally

## The following object is masked from 'package:purrr':  
##   
## cross

## The following object is masked from 'package:ggplot2':  
##   
## stat

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,  
## quantile, sd, t.test, var

## The following objects are masked from 'package:base':  
##   
## max, mean, min, prod, range, sample, sum

library(dplyr)  
library(janitor)

##   
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(stringr)

## Read in the data

#### Exercise 2

18979 players are included in the data set and there are 77 variables

library(readr)  
FIFA21 <- read\_csv("fifa21\_raw.csv")

## Rows: 18979 Columns: 77

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (22): photoUrl, LongName, playerUrl, Nationality, Positions, Name, Team&...  
## dbl (55): Age, ↓OVA, POT, ID, BOV, Growth, Attacking, Crossing, Finishing, H...

##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(FIFA21)  
dim(FIFA21)

## [1] 18979 77

## Cleaning activities

#### Exercise 3

The name of the 19th column before cleaning is “LoanDateEnd” and after cleaning is “loan\_date\_end”.

summary(FIFA21)

## photoUrl LongName playerUrl Nationality   
## Length:18979 Length:18979 Length:18979 Length:18979   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Positions Name Age ↓OVA   
## Length:18979 Length:18979 Min. :16.00 Min. :47.00   
## Class :character Class :character 1st Qu.:21.00 1st Qu.:61.00   
## Mode :character Mode :character Median :25.00 Median :66.00   
## Mean :25.19 Mean :65.72   
## 3rd Qu.:29.00 3rd Qu.:70.00   
## Max. :53.00 Max. :93.00   
## POT Team&Contract ID Height   
## Min. :47.00 Length:18979 Min. : 41 Length:18979   
## 1st Qu.:67.00 Class :character 1st Qu.:210135 Class :character   
## Median :71.00 Mode :character Median :232424 Mode :character   
## Mean :71.14 Mean :226405   
## 3rd Qu.:75.00 3rd Qu.:246926   
## Max. :95.00 Max. :259216   
## Weight foot BOV BP   
## Length:18979 Length:18979 Min. :48.00 Length:18979   
## Class :character Class :character 1st Qu.:62.00 Class :character   
## Mode :character Mode :character Median :67.00 Mode :character   
## Mean :66.75   
## 3rd Qu.:71.00   
## Max. :93.00   
## Growth Joined LoanDateEnd Value   
## Min. : 0.000 Length:18979 Length:18979 Length:18979   
## 1st Qu.: 0.000 Class :character Class :character Class :character   
## Median : 4.000 Mode :character Mode :character Mode :character   
## Mean : 5.417   
## 3rd Qu.: 9.000   
## Max. :26.000   
## Wage ReleaseClause Attacking Crossing   
## Length:18979 Length:18979 Min. : 42.0 Min. : 6.00   
## Class :character Class :character 1st Qu.:222.0 1st Qu.:38.00   
## Mode :character Mode :character Median :263.0 Median :54.00   
## Mean :248.9 Mean :49.69   
## 3rd Qu.:297.0 3rd Qu.:63.00   
## Max. :437.0 Max. :94.00   
## Finishing HeadingAccuracy ShortPassing Volleys Skill   
## Min. : 3.00 Min. : 5.00 Min. : 7.00 Min. : 3.0 Min. : 40.0   
## 1st Qu.:30.00 1st Qu.:44.00 1st Qu.:54.00 1st Qu.:30.0 1st Qu.:222.0   
## Median :49.00 Median :55.00 Median :62.00 Median :44.0 Median :269.0   
## Mean :45.84 Mean :51.94 Mean :58.77 Mean :42.7 Mean :256.5   
## 3rd Qu.:62.00 3rd Qu.:64.00 3rd Qu.:68.00 3rd Qu.:56.0 3rd Qu.:310.0   
## Max. :95.00 Max. :93.00 Max. :94.00 Max. :90.0 Max. :470.0   
## Dribbling Curve FKAccuracy LongPassing   
## Min. : 5.00 Min. : 4.00 Min. : 5.00 Min. : 5.00   
## 1st Qu.:49.00 1st Qu.:35.00 1st Qu.:31.00 1st Qu.:43.00   
## Median :61.00 Median :49.00 Median :41.00 Median :56.00   
## Mean :55.59 Mean :47.25 Mean :42.39 Mean :52.73   
## 3rd Qu.:68.00 3rd Qu.:61.00 3rd Qu.:55.00 3rd Qu.:64.00   
## Max. :96.00 Max. :94.00 Max. :94.00 Max. :93.00   
## BallControl Movement Acceleration SprintSpeed   
## Min. : 5.00 Min. :122.0 Min. :13.00 Min. :12.00   
## 1st Qu.:54.00 1st Qu.:289.0 1st Qu.:57.00 1st Qu.:57.00   
## Median :63.00 Median :327.0 Median :67.00 Median :67.00   
## Mean :58.53 Mean :317.7 Mean :64.36 Mean :64.41   
## 3rd Qu.:69.00 3rd Qu.:356.0 3rd Qu.:74.00 3rd Qu.:74.00   
## Max. :96.00 Max. :464.0 Max. :97.00 Max. :96.00   
## Agility Reactions Balance Power ShotPower   
## Min. :14.00 Min. :24.00 Min. :12.00 Min. :122.0 Min. :18.0   
## 1st Qu.:55.00 1st Qu.:56.00 1st Qu.:56.00 1st Qu.:264.0 1st Qu.:48.0   
## Median :66.00 Median :62.00 Median :66.00 Median :302.0 Median :59.0   
## Mean :63.37 Mean :61.66 Mean :63.93 Mean :296.6 Mean :57.8   
## 3rd Qu.:74.00 3rd Qu.:68.00 3rd Qu.:74.00 3rd Qu.:334.0 3rd Qu.:68.0   
## Max. :96.00 Max. :95.00 Max. :97.00 Max. :444.0 Max. :95.0   
## Jumping Stamina Strength LongShots   
## Min. :15.00 Min. :12.00 Min. :16.00 Min. : 4.00   
## 1st Qu.:58.00 1st Qu.:55.00 1st Qu.:57.00 1st Qu.:32.00   
## Median :65.00 Median :66.00 Median :66.00 Median :51.00   
## Mean :64.58 Mean :62.65 Mean :64.79 Mean :46.79   
## 3rd Qu.:73.00 3rd Qu.:73.00 3rd Qu.:74.00 3rd Qu.:62.00   
## Max. :95.00 Max. :97.00 Max. :97.00 Max. :94.00   
## Mentality Aggression Interceptions Positioning Vision   
## Min. : 50.0 Min. : 9.00 Min. : 3.00 Min. : 2.0 Min. : 9.00   
## 1st Qu.:227.5 1st Qu.:44.00 1st Qu.:25.00 1st Qu.:40.0 1st Qu.:45.00   
## Median :263.0 Median :58.00 Median :53.00 Median :55.0 Median :55.00   
## Mean :254.2 Mean :55.59 Mean :46.39 Mean :50.3 Mean :53.86   
## 3rd Qu.:297.0 3rd Qu.:69.00 3rd Qu.:64.00 3rd Qu.:64.0 3rd Qu.:64.00   
## Max. :421.0 Max. :96.00 Max. :91.00 Max. :95.0 Max. :95.00   
## Penalties Composure Defending Marking   
## Min. : 6.00 Min. :12.00 Min. : 20.0 Min. : 3.00   
## 1st Qu.:39.00 1st Qu.:50.00 1st Qu.: 83.0 1st Qu.:29.00   
## Median :49.00 Median :59.00 Median :159.0 Median :52.00   
## Mean :48.06 Mean :58.07 Mean :139.9 Mean :46.56   
## 3rd Qu.:60.00 3rd Qu.:67.00 3rd Qu.:191.0 3rd Qu.:63.00   
## Max. :92.00 Max. :96.00 Max. :272.0 Max. :94.00   
## StandingTackle SlidingTackle Goalkeeping GKDiving   
## Min. : 5.00 Min. : 4.00 Min. : 10.00 Min. : 2.00   
## 1st Qu.:27.00 1st Qu.:25.00 1st Qu.: 48.00 1st Qu.: 8.00   
## Median :55.00 Median :52.00 Median : 53.00 Median :11.00   
## Mean :47.71 Mean :45.64 Mean : 81.44 Mean :16.42   
## 3rd Qu.:65.00 3rd Qu.:63.00 3rd Qu.: 59.00 3rd Qu.:14.00   
## Max. :93.00 Max. :90.00 Max. :440.00 Max. :90.00   
## GKHandling GKKicking GKPositioning GKReflexes TotalStats   
## Min. : 2.00 Min. : 2.00 Min. : 2.00 Min. : 2.00 Min. : 747   
## 1st Qu.: 8.00 1st Qu.: 8.00 1st Qu.: 8.00 1st Qu.: 8.00 1st Qu.:1452   
## Median :11.00 Median :11.00 Median :11.00 Median :11.00 Median :1627   
## Mean :16.21 Mean :16.08 Mean :16.22 Mean :16.52 Mean :1595   
## 3rd Qu.:14.00 3rd Qu.:14.00 3rd Qu.:14.00 3rd Qu.:14.00 3rd Qu.:1781   
## Max. :92.00 Max. :93.00 Max. :91.00 Max. :90.00 Max. :2316   
## BaseStats W/F SM A/W   
## Min. :232.0 Length:18979 Length:18979 Length:18979   
## 1st Qu.:327.0 Class :character Class :character Class :character   
## Median :356.0 Mode :character Mode :character Mode :character   
## Mean :355.7   
## 3rd Qu.:384.0   
## Max. :498.0   
## D/W IR PAC SHO   
## Length:18979 Length:18979 Min. :25.00 Min. :16.00   
## Class :character Class :character 1st Qu.:61.00 1st Qu.:44.00   
## Mode :character Mode :character Median :68.00 Median :56.00   
## Mean :67.45 Mean :53.46   
## 3rd Qu.:75.00 3rd Qu.:64.00   
## Max. :96.00 Max. :93.00   
## PAS DRI DEF PHY   
## Min. :25.00 Min. :25.00 Min. :12.00 Min. :28.00   
## 1st Qu.:51.00 1st Qu.:57.00 1st Qu.:35.00 1st Qu.:58.00   
## Median :58.00 Median :64.00 Median :53.00 Median :65.00   
## Mean :57.68 Mean :62.88 Mean :49.87 Mean :64.37   
## 3rd Qu.:64.00 3rd Qu.:69.00 3rd Qu.:63.00 3rd Qu.:71.00   
## Max. :93.00 Max. :95.00 Max. :91.00 Max. :91.00   
## Hits   
## Length:18979   
## Class :character   
## Mode :character   
##   
##   
##

FIFA21 <- clean\_names(FIFA21)

#### Exercise 4

The median weight of players in the FIFA21 data is 165 lbs.

tally(~weight, data= FIFA21)

## weight  
## 110lbs 115lbs 117lbs 119lbs 121lbs 123lbs 126lbs 128lbs 130lbs 132lbs 134lbs   
## 1 2 1 5 17 12 19 44 41 153 96   
## 137lbs 139lbs 141lbs 143lbs 146lbs 148lbs 150lbs 152lbs 154lbs 157lbs 159lbs   
## 181 239 341 535 401 508 795 566 1496 648 1024   
## 161lbs 163lbs 165lbs 168lbs 170lbs 172lbs 174lbs 176lbs 179lbs 181lbs 183lbs   
## 956 957 1458 880 893 995 684 1109 523 622 515   
## 185lbs 187lbs 190lbs 192lbs 194lbs 196lbs 198lbs 201lbs 203lbs 205lbs 207lbs   
## 444 423 285 224 266 128 156 87 68 52 34   
## 209lbs 212lbs 214lbs 216lbs 218lbs 220lbs 223lbs 225lbs 227lbs 229lbs 236lbs   
## 30 15 9 13 3 8 4 4 4 3 1   
## 243lbs   
## 1

FIFA21<- mutate(FIFA21, weight = str\_remove\_all(weight, "lbs"))  
FIFA21 <- mutate( FIFA21,as.numeric(weight))  
favstats(~weight, data= FIFA21)

## Warning in FUN(eval(formula[[2]], data, .envir), ...): Auto-converting character  
## to numeric.

## min Q1 median Q3 max mean sd n missing  
## 110 154 165 176 243 165.3878 15.58793 18979 0

#### Exercise 5

The median height of players in the FIFA21 data set is 6’1”

tally(~height, data = FIFA21)

## height  
## 5'1" 5'10" 5'11" 5'2" 5'3" 5'4" 5'5" 5'6" 5'7" 5'8" 5'9" 6'0" 6'1"   
## 2 2654 2188 4 15 33 188 332 953 1069 2304 2859 2001   
## 6'2" 6'3" 6'4" 6'5" 6'6" 6'7" 6'8" 6'9"   
## 2108 1049 797 280 116 20 6 1

FIFA21 <- mutate(FIFA21, height = str\_remove\_all(height, "[[\*\*'\*\*]]"))  
FIFA21 <- mutate(FIFA21, height = str\_remove\_all(height, '"'))  
FIFA21 <- mutate(FIFA21, height = as.numeric(height))  
favstats(~height, data=FIFA21)

## min Q1 median Q3 max mean sd n missing  
## 51 59 61 510 511 175.1805 196.229 18979 0

#### Exercise 6

FIFA21<- separate(FIFA21, height, sep = 1, into=c("feet", "inches"))  
FIFA21 <- mutate(FIFA21, feet = as.numeric(feet))  
FIFA21 <- mutate(FIFA21, inches = as.numeric(inches))  
FIFA21 <- mutate(FIFA21, hgt\_inches = feet\*12 +inches)  
favstats(~hgt\_inches, data = FIFA21)

## min Q1 median Q3 max mean sd n missing  
## 61 69 71 73 81 71.34807 2.684666 18979 0

#### Exercise 7

10005 players are categorized as average in the FIFA21 data set.

FIFA21<- mutate(FIFA21, hgt\_group = case\_when(hgt\_inches >= 73 ~ "tall", TRUE ~case\_when(hgt\_inches >= 69 ~ "average", TRUE ~ "short")))  
tally(~hgt\_group, data = FIFA21)

## hgt\_group  
## average short tall   
## 10005 2596 6378

#### Exercise 8

FIFA21 <- mutate(FIFA21, hgt\_group = fct\_relevel(hgt\_group,c("short", "average", "tall")))  
gf\_bar(~hgt\_group, data = FIFA21)

