plot

Kamila Bukirova

6/4/2020

## R Markdown

I want to create different types of plots for my data based on the nature of variables. For this I am using the book by K. Healy(2019). Firstly, I will use the data from SPSS, stratified in python.

## re-encoding from UTF-8

## follow\_up\_2 hnc\_exit\_stat2  
## Min. : 4 Censor :1494   
## 1st Qu.:3904 confirmed cancer : 506   
## Median :4609 Last Participant Contact" : 18   
## Mean :4168 Death : 10   
## 3rd Qu.:5050 no time at risk : 0   
## Max. :5885 confirmed in situ/borderline/LMP/Non-targetCancer: 0   
## (Other) : 0   
## hnc\_exitage hnc\_exitdays   
## Min. :55.00 Min. : 15   
## 1st Qu.:69.00 1st Qu.:3909   
## Median :74.00 Median :4619   
## Mean :73.85 Mean :4178   
## 3rd Qu.:78.00 3rd Qu.:5064   
## Max. :89.00 Max. :5898   
##   
## educat   
## some college :445   
## 12 years or completed high school :431   
## College Graduate" :359   
## postgraduate :325   
## Post High School Training Other Than College":294   
## (Other) :173   
## NA's : 1   
## marital occupat pipe   
## Married Or Living As Married:1502 retired :920 never :1602   
## widowed : 166 working :821 current pipe smoker: 23   
## divorced : 269 homemaker:147 Former Pipe Smoker : 401   
## separated : 21 disabled : 62 NA's : 2   
## never married : 70 other : 45   
## (Other) : 26   
## NA's : 7   
## cigar sisters brothers asp   
## Never :1646 ONE :641 one :622 no :1023   
## Current Cigar Smoker: 43 none :500 none :481 yes : 999   
## Former Cigar Smoker : 335 TWO :387 two :411 NA's: 6   
## NA's : 4 THREE :213 three :250   
## FOUR :129 four :126   
## (Other):154 (Other):135   
## NA's : 4 NA's : 3   
## ibup asppd ibuppd hyperten\_f hearta\_f   
## no :1469 none :973 none :1429 no :1373 no :1818   
## yes : 557 1/Day :438 2+/day : 131 yes : 652 yes : 207   
## NA's: 2 3-4/week:180 3-4/week : 127 NA's: 3 NA's: 3   
## <2/month:135 <2/month : 109   
## 2+/day :122 2-3/month: 94   
## (Other) :177 (Other) : 136   
## NA's : 3 NA's : 2   
## stroke\_f emphys\_f bronchit\_f diabetes\_f polyps\_f arthrit\_f   
## no :1969 no :1947 no :1908 no :1881 no :1840 no :1267   
## yes : 56 yes : 78 yes : 117 yes : 143 yes : 184 yes : 759   
## NA's: 3 NA's: 3 NA's: 3 NA's: 4 NA's: 4 NA's: 2   
##   
##   
##   
##   
## osteopor\_f divertic\_f gallblad\_f bq\_age race7   
## no :1944 no :1887 no :1802 Min. :54.00 white, non-hispanic:1812   
## yes : 75 yes : 135 yes : 222 1st Qu.:58.00 black, non-hispanic: 93   
## NA's: 9 NA's: 6 NA's: 4 Median :62.00 hispanic : 37   
## Mean :62.41 asian : 67   
## 3rd Qu.:66.00 pacific Islander : 10   
## Max. :75.00 American Indian : 8   
## missing : 1   
## smoked\_f smokea\_f rsmoker\_f ssmokea\_f cigpd\_f   
## no : 101 Min. : 8.00 no :1522 Min. : 0.00 11-20 :708   
## yes:1927 1st Qu.:16.00 yes : 405 1st Qu.:27.00 1-10 :427   
## Median :17.00 NA's: 101 Median :40.00 21-30 :404   
## Mean :18.09 Mean :36.56 31-40 :243   
## 3rd Qu.:19.00 3rd Qu.:51.00 41-60 :117   
## Max. :71.00 Max. :72.00 (Other):128   
## NA's :107 NA's :279 NA's : 1   
## filtered\_f cig\_stat cig\_stop   
## filter :1415 never smoked cigarettes : 101 Min. : 0.00   
## non-filter : 348 current cigarette smoker: 405 1st Qu.: 1.00   
## about equal: 161 former cigarette smoker :1522 Median :13.00   
## NA's : 104 Mean :15.26   
## 3rd Qu.:26.00   
## Max. :54.00   
## NA's :108   
## cig\_years pack\_years bmi\_curc state   
## Min. : 0.50 Min. : 0.25 0-18.49 : 14 minnesota :264   
## 1st Qu.:18.00 1st Qu.: 16.50 18.5-24.999:644 pennsylvania:230   
## Median :30.00 Median : 33.00 25-29.999 :895 michigan :209   
## Mean :28.73 Mean : 38.80 30+ :444 wisconsin :181   
## 3rd Qu.:40.00 3rd Qu.: 52.50 NA's : 31 missouri :140   
## Max. :60.00 Max. :255.00 (Other) :994   
## NA's :137 NA's :140 NA's : 10   
## bmi\_20c bmi\_50c colon\_comorbidity liver\_comorbidity  
## 0-18.49 : 156 0-18.49 : 19 no :1998 no :1925   
## 18.5-24.999:1523 18.5-24.999:950 yes : 23 yes : 96   
## 25-29.999 : 289 25-29.999 :781 NA's: 7 NA's: 7   
## 30+ : 28 30+ :245   
## NA's : 32 NA's : 33   
##   
##   
## fh\_cancer hnc\_fh\_cnt hnc\_fh   
## no : 868 Min. :0.0000 no :1930   
## yes :1159 1st Qu.:0.0000 yes, immediate family member : 32   
## NA's: 1 Median :0.0000 possibly-relative or cancer type not clear: 63   
## Mean :0.0158 NA's : 3   
## 3rd Qu.:0.0000   
## Max. :2.0000   
## NA's :3   
## is\_dead\_with\_cod is\_dead mortality\_exitage  
## not confirmed dead:1363 not confirmed dead:1363 Min. :60.00   
## dead : 664 dead : 663 1st Qu.:75.00   
## NA's : 1 NA's : 2 Median :79.00   
## Mean :79.08   
## 3rd Qu.:84.00   
## Max. :95.00   
## NA's :1   
## mortality\_exitstat mortality\_exitdays ph\_any\_bq   
## death : 664 Min. : 436 no :1947   
## last NDI/Cutoff:1106 1st Qu.:5518 yes : 81   
## refusal : 244 Median :6439 unknown: 0   
## other : 13 Mean :6087   
## NA's : 1 3rd Qu.:7150   
## Max. :8077   
## NA's :1   
## center arm sex   
## university of minnesota :383 intervention:1034 male :1279   
## henry ford health system :302 control : 994 female: 749   
## university of pittsburgh :241   
## washington university in st louis :236   
## university of colorado :222   
## marschfield clinic research foundation:206   
## (Other) :438   
## agelevel rndyear\_cat   
## <=59 :656 1993-1995: 599   
## 60-64:681 1996-1998:1148   
## 65-69:470 1999-2001: 281   
## >=70 :221   
##   
##   
##   
## educ\_new\_cat  
## less than 11 years :173   
## 12 years or completed high school :431   
## post high school training other than college:294   
## some college :445   
## college graduate :359   
## postgraduate :325   
## NA's : 1   
## marital\_Recoded race\_recoded   
## married or living as married:1502 white, non-hispanic:1812   
## widowed : 166 black, non-hispanic: 93   
## divorced or separated : 290 hispanic : 37   
## never married : 70 asian : 67   
## other : 19   
##   
##   
## state\_recoded bmi\_difference\_recoded difference\_age   
## foreign country : 97 negative difference: 49 Min. : 0.000   
## northeast : 351 no difference :1942 1st Qu.: 3.000   
## midwest :1084 1-2 difference : 0 Median : 6.000   
## south : 183 3 difference : 2 Mean : 5.228   
## west : 282 NA's : 35 3rd Qu.: 6.000   
## u.s. territories: 0 Max. :19.000   
## NA's : 31 NA's :1   
## difference\_age\_cat bmi\_difference\_cat\_50\_cur\_recoded  
## difference of 0 : 67 negative difference: 111   
## <5 years difference: 614 no difference :1298   
## >=5 years :1346 1-2 difference : 584   
## NA's : 1 3 difference : 6   
## NA's : 29   
##   
##   
## hnc\_exit\_stat\_new Rfollow\_ follow\_up\_norm   
## negative hnc:1522 Min. :0.0004931 Min. :-3.294436   
## positive hnc: 506 1st Qu.:0.2503698 1st Qu.:-0.673714   
## Median :0.4992604 Median :-0.001854   
## Mean :0.5002465 Mean :-0.000002   
## 3rd Qu.:0.7507396 3rd Qu.: 0.674104   
## Max. :1.0000000 Max. : 3.294436   
## NA's :1   
## RFR001 hnc\_exitdays\_norm Rsmokea\_ smokea\_f\_norm   
## Min. :0.0004931 Min. :-3.294436 Min. :0.00078 Min. :-3.16297   
## 1st Qu.:0.2502465 1st Qu.:-0.673714 1st Qu.:0.30036 1st Qu.:-0.52335   
## Median :0.5007396 Median : 0.001854 Median :0.43883 Median :-0.15393   
## Mean :0.5002465 Mean :-0.000006 Mean :0.50026 Mean : 0.00121   
## 3rd Qu.:0.7507396 3rd Qu.: 0.674879 3rd Qu.:0.71629 3rd Qu.: 0.57187   
## Max. :1.0000000 Max. : 3.294436 Max. :1.00000 Max. : 3.27917   
## NA's :1 NA's :107 NA's :108   
## Rssmokea ssmokea\_f\_norm Rcig\_sto cig\_stop\_norm   
## Min. :0.07433 Min. :-1.44429 Min. :0.1122 Min. :-1.21470   
## 1st Qu.:0.25071 1st Qu.:-0.67224 1st Qu.:0.2466 1st Qu.:-0.68518   
## Median :0.49171 Median :-0.02078 Median :0.5013 Median : 0.00326   
## Mean :0.50029 Mean : 0.01625 Mean :0.5003 Mean : 0.02619   
## 3rd Qu.:0.75672 3rd Qu.: 0.69578 3rd Qu.:0.7422 3rd Qu.: 0.65010   
## Max. :1.00000 Max. : 3.13556 Max. :1.0000 Max. : 3.27902   
## NA's :279 NA's :280 NA's :108 NA's :109   
## Rcig\_yea norm\_cig\_years Rpack\_ye pack\_years\_norm   
## Min. :0.00106 Min. :-3.07355 Min. :0.00053 Min. :-3.27428   
## 1st Qu.:0.25568 1st Qu.:-0.65671 1st Qu.:0.25238 1st Qu.:-0.66701   
## Median :0.48810 Median :-0.02983 Median :0.50477 Median : 0.01195   
## Mean :0.50026 Mean : 0.00018 Mean :0.50026 Mean : 0.00039   
## 3rd Qu.:0.74960 3rd Qu.: 0.67324 3rd Qu.:0.75132 3rd Qu.: 0.67866   
## Max. :1.00000 Max. : 3.27473 Max. :1.00000 Max. : 3.27428   
## NA's :137 NA's :138 NA's :140 NA's :141   
## mortality\_age\_norm RFR002 Rmortali   
## Min. :-3.178573 Min. :0.00074 Min. :0.0004933   
## 1st Qu.:-0.628993 1st Qu.:0.26468 1st Qu.:0.2503700   
## Median : 0.005565 Median :0.50222 Median :0.4997533   
## Mean : 0.001640 Mean :0.50025 Mean :0.5002467   
## 3rd Qu.: 0.743257 3rd Qu.:0.77134 3rd Qu.:0.7501233   
## Max. : 3.178573 Max. :0.99926 Max. :1.0000000   
## NA's :1 NA's :1 NA's :1   
## mortality\_exitdays\_norm bmi\_20 bmi\_50 bmi\_curr   
## Min. :-3.294298 Min. :15.21 Min. :16.82 Min. :15.62   
## 1st Qu.:-0.673714 1st Qu.:20.08 1st Qu.:23.01 1st Qu.:24.02   
## Median :-0.000618 Median :21.70 Median :25.06 Median :26.58   
## Mean :-0.000002 Mean :22.12 Mean :25.63 Mean :27.18   
## 3rd Qu.: 0.673714 3rd Qu.:23.73 3rd Qu.:27.43 3rd Qu.:29.41   
## Max. : 3.294298 Max. :47.23 Max. :47.84 Max. :58.58   
## NA's :2 NA's :32 NA's :33 NA's :31

##Just to try ## Including Plots

I will take two numeric variables from my data and will construct the geom\_point for looking at the relation between the cigarette years and hnc exitage

## Warning: package 'tidyverse' was built under R version 3.6.3

## -- Attaching packages ------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 1.0.0  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 3.6.3

## Warning: package 'tibble' was built under R version 3.6.3

## Warning: package 'tidyr' was built under R version 3.6.3

## Warning: package 'readr' was built under R version 3.6.3

## Warning: package 'purrr' was built under R version 3.6.3

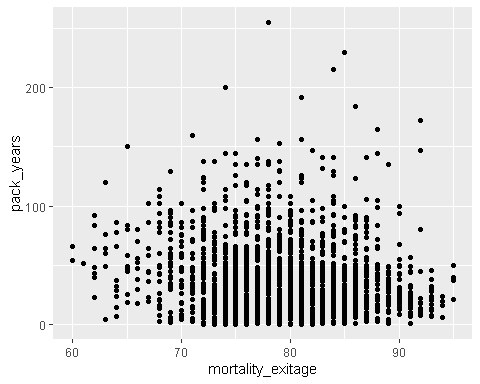
## Warning: package 'dplyr' was built under R version 3.6.3

## Warning: package 'stringr' was built under R version 3.6.3

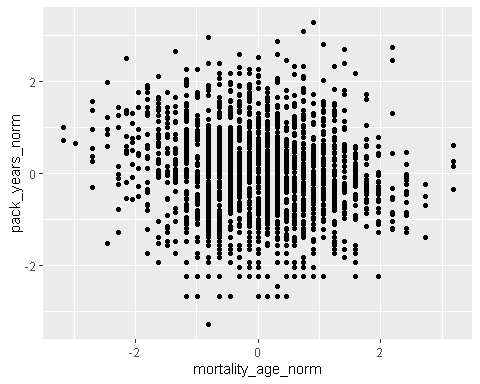
## Warning: package 'forcats' was built under R version 3.6.3

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

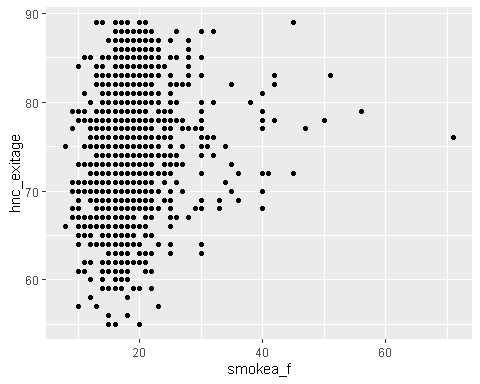
## Warning: Removed 141 rows containing missing values (geom\_point).

 and normalized graph alongside.

## Warning: Removed 142 rows containing missing values (geom\_point).

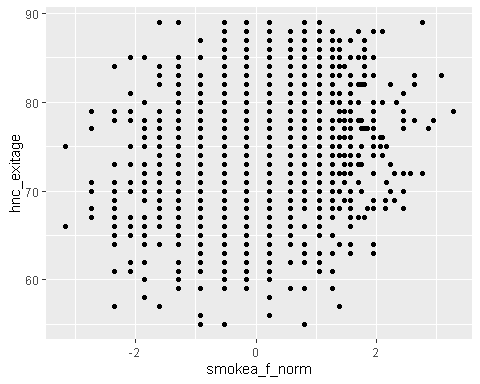
 We did the same for the age of start smoking and age of the hnc was firstly diagnosed

## Warning: Removed 107 rows containing missing values (geom\_point).

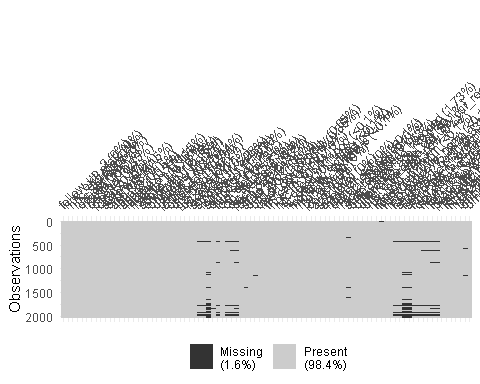


and normalized one

## Warning: Removed 108 rows containing missing values (geom\_point).

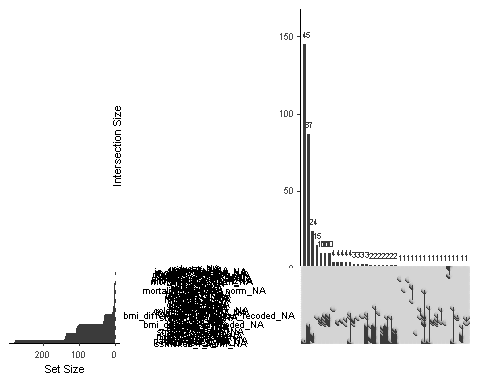
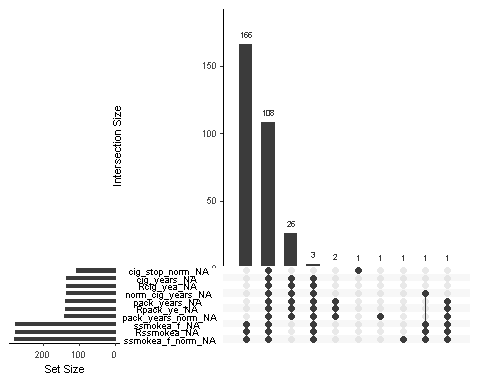
 I want to take a look at the missingness pattern

## Warning: package 'naniar' was built under R version 3.6.3



and at the missingness patterns for the data prepared for imputation

## [1] 68

 and for just 10 variables in a row 

I want to check the missingness in variables 