Bart.R

ciastkow

2019-11-10

##### Designation START  
  
library('tidyverse')  
library('here')  
library('ggthemes')  
library('stringr')  
  
wine\_data <-  
 read.csv(here("data", "input", "winemag-data-130k-v2.csv"))  
  
  
  
unwanted\_array = list( 'Š'='S', 'š'='s', 'Ž'='Z', 'ž'='z', 'À'='A', 'Á'='A', 'Â'='A', 'Ã'='A', 'Ä'='A', 'Å'='A', 'Æ'='A', 'Ç'='C', 'È'='E', 'É'='E',  
 'Ê'='E', 'Ë'='E', 'Ì'='I', 'Í'='I', 'Î'='I', 'Ï'='I', 'Ñ'='N', 'Ò'='O', 'Ó'='O', 'Ô'='O', 'Õ'='O', 'Ö'='O', 'Ø'='O', 'Ù'='U',  
 'Ú'='U', 'Û'='U', 'Ü'='U', 'Ý'='Y', 'Þ'='B', 'ß'='Ss', 'à'='a', 'á'='a', 'â'='a', 'ã'='a', 'ä'='a', 'å'='a', 'æ'='a', 'ç'='c',  
 'è'='e', 'é'='e', 'ê'='e', 'ë'='e', 'ì'='i', 'í'='i', 'î'='i', 'ï'='i', 'ð'='o', 'ñ'='n', 'ò'='o', 'ó'='o', 'ô'='o', 'õ'='o',  
 'ö'='o', 'ø'='o', 'ù'='u', 'ú'='u', 'û'='u', 'ý'='y', 'ý'='y', 'þ'='b', 'ÿ'='y' )  
  
  
wine\_data\_Reduced <- wine\_data %>%  
 #filter(grepl("serv", designation, ignore.case=TRUE)) %>%  
 mutate(designation=tolower(designation)) %>%  
 mutate(designation=str\_replace(designation,"[\\d¡\\;ã\\-£±ºÃª©]+","")) %>%  
 #mutate(designation=str\_replace(designation,"\\W+"," ")) %>%  
 #mutate(designation=gsub("[\\d¡\\;ã\\-£±ºÃª©]+","",designation)) %>%  
 mutate(designation=gsub("\\W+"," ",designation)) %>%  
 mutate(designation=str\_replace(designation,"[ãÃ©\\w]+ser[ãÃ©\\w]+","Reserve")) %>%   
 mutate(designation=chartr(paste(names(unwanted\_array), collapse=''),paste(unwanted\_array, collapse=''),designation)) %>%   
 mutate(designation=ifelse(grepl("Reserve",designation),"Reserve",designation)) %>%   
 mutate(designation=ifelse(grepl("extra dry",designation),"Extra Dry",ifelse(grepl("(dry|trocken)",designation),"Dry",designation))) %>%   
 mutate(designation=ifelse(grepl("brut",designation),"Brut",designation)) %>%   
 mutate(designation=ifelse(grepl("(estate|grand|casa)",designation),"Estate",designation)) %>%   
 mutate(designation=ifelse(grepl("single",designation),"Single Vineyard",designation)) %>%   
 mutate(designation=ifelse(grepl("(klassik|classic|tradition|vintage)",designation),"Classic Vintage",designation)) %>%   
 mutate(designation=ifelse(grepl("ros",designation),"Rose",designation)) %>%   
 mutate(designation=ifelse(grepl("barrel s",designation),"Barrel Sample",designation)) %>%   
 mutate(designation=ifelse(grepl("(old v|vieilles)",designation),"Old Vine",designation)) %>%   
 mutate(designation=ifelse(grepl("(vineyard|ranch|alpha|branco|broquel)",designation),"Some Vineyard",designation)) %>%   
 mutate(designation=ifelse(grepl("(barrel|crianza|cuve)",designation),"Barrel",designation)) %>%   
 mutate(designation=ifelse(grepl("unoaked",designation),"UnOaked",designation)) %>%   
 mutate(designation=ifelse(grepl("cuve prestige",designation),"Finest Champagne",designation)) %>%   
 mutate(designation=ifelse(grepl("(blanc|white|bianco)",designation),"White",designation)) %>%   
 mutate(designation=ifelse(grepl("(red|tinto|bussia)",designation),"Red",designation)) %>%   
 mutate(designation=ifelse(grepl("(nouveau|proprietary|signature|selec|premier)",designation),"Signature",designation)) %>%   
 mutate(designation=ifelse(grepl("lot",designation),"Lot",designation)) %>%   
 mutate(designation=ifelse(grepl("late",designation),"Late Harvest",designation)) %>%   
 mutate(designation=ifelse(grepl("(oak|roble)",designation),"Oak",designation)) %>%   
 mutate(designation=ifelse(grepl("(organic|cannubi)",designation),"Organic",designation)) %>%   
 mutate(designation=ifelse(grepl("(port|colheita)",designation),"Port",designation)) %>%   
 mutate(designation=ifelse(grepl("(collection|premium|prestige|limited)",designation),"Premium",designation)) %>%   
 mutate(designation=ifelse(grepl("clone",designation),"Clone",designation)) %>%   
 mutate(designation=ifelse(grepl("(block|bin)",designation),"Block",designation))   
 #select(designation) %>%  
 #unique() %>%   
 #View()  
  
 summary(wine\_data\_Reduced)

## X country   
## Min. : 0 US :54504   
## 1st Qu.: 32493 France :22093   
## Median : 64985 Italy :19540   
## Mean : 64985 Spain : 6645   
## 3rd Qu.: 97478 Portugal: 5691   
## Max. :129970 Chile : 4472   
## (Other) :17026   
## description   
## Cigar box, cafÃ© au lait, and dried tobacco aromas are followed by coffee and cherry flavors, with barrel spices lingering on the finish. The wood gets a bit out front but it still delivers enjoyment. : 3   
## Gravenstein apple, honeysuckle and jasmine aromas show on the relatively boisterous nose of this bottling from a large vineyard on Highway 46 east of Paso Robles. There is compellingly grippy texture to the sip, with ripe flavors of pear and honeydew melon. A salty acidity takes it to the next level.: 3   
## Ripe plum, game, truffle, leather and menthol are some of the aromas you'll find on this earthy wine. The tightly wound palate offers dried black cherry, chopped sage, mint and roasted coffee bean alongside raspy tannins that leave a mouth-drying finish. : 3   
## Seductively tart in lemon pith, cranberry and pomegranate, this refreshing, light-bodied quaff is infinitely enjoyable, both on its own or at the table. It continues to expand on the palate into an increasing array of fresh flavors, finishing in cherry and orange. : 3   
## Stalky aromas suggest hay and green herbs, with raspberry in the backdrop. It's hot and short in terms of mouthfeel, with herbal flavors leading the way and berry fruit running behind. Dry red fruit and herbal notes dominate the finish. : 3   
## This zesty red has pretty aromas that suggest small red berry, blue flower and a whiff of moist soil. The vibrant palate offers sour cherry, pomegranate and a hint of anise alongside zesty acidity and refined tannins. : 3   
## (Other) :129953   
## designation points price province   
## Length:129971 Min. : 80.00 Min. : 4.00 California:36247   
## Class :character 1st Qu.: 86.00 1st Qu.: 17.00 Washington: 8639   
## Mode :character Median : 88.00 Median : 25.00 Bordeaux : 5941   
## Mean : 88.45 Mean : 35.36 Tuscany : 5897   
## 3rd Qu.: 91.00 3rd Qu.: 42.00 Oregon : 5373   
## Max. :100.00 Max. :3300.00 Burgundy : 3980   
## NA's :8996 (Other) :63894   
## region\_1 region\_2   
## :21247 :79460   
## Napa Valley : 4480 Central Coast :11065   
## Columbia Valley (WA): 4124 Sonoma : 9028   
## Russian River Valley: 3091 Columbia Valley : 8103   
## California : 2629 Napa : 6814   
## Paso Robles : 2350 Willamette Valley: 3423   
## (Other) :92050 (Other) :12078   
## taster\_name taster\_twitter\_handle  
## :26244 :31213   
## Roger Voss :25514 @vossroger :25514   
## Michael Schachner:15134 @wineschach :15134   
## Kerin Oâ\200\231Keefe :10776 @kerinokeefe:10776   
## Virginie Boone : 9537 @vboone : 9537   
## Paul Gregutt : 9532 @paulgwineÂ : 9532   
## (Other) :33234 (Other) :28265   
## title   
## Gloria Ferrer NV Sonoma Brut Sparkling (Sonoma County) : 11   
## Korbel NV Brut Sparkling (California) : 9   
## Segura Viudas NV Extra Dry Sparkling (Cava) : 8   
## Gloria Ferrer NV Blanc de Noirs Sparkling (Carneros) : 7   
## Ruinart NV Brut RosÃ© (Champagne) : 7   
## Segura Viudas NV Aria Estate Extra Dry Sparkling (Cava): 7   
## (Other) :129922   
## variety winery   
## Pinot Noir :13272 Wines & Winemakers: 222   
## Chardonnay :11753 Testarossa : 218   
## Cabernet Sauvignon : 9472 DFJ Vinhos : 215   
## Red Blend : 8946 Williams Selyem : 211   
## Bordeaux-style Red Blend: 6915 Louis Latour : 199   
## Riesling : 5189 Georges Duboeuf : 196   
## (Other) :74424 (Other) :128710

dim(wine\_data\_Reduced)

## [1] 129971 14

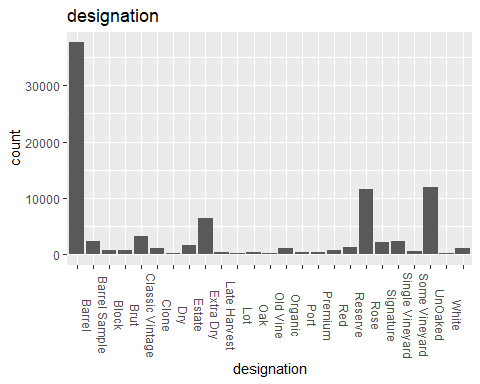
wine\_data\_Reduced <- wine\_data\_Reduced %>%   
 group\_by(designation) %>%   
 summarize(count\_obs=n()) %>%   
 filter(count\_obs > 1) %>%   
 arrange (desc(count\_obs)) %>%   
 slice(1:25)  
  
  
  
summary(wine\_data\_Reduced)

## designation count\_obs   
## Length:25 Min. : 180   
## Class :character 1st Qu.: 413   
## Mode :character Median : 1158   
## Mean : 3588   
## 3rd Qu.: 2386   
## Max. :37632

dim(wine\_data\_Reduced)

## [1] 25 2

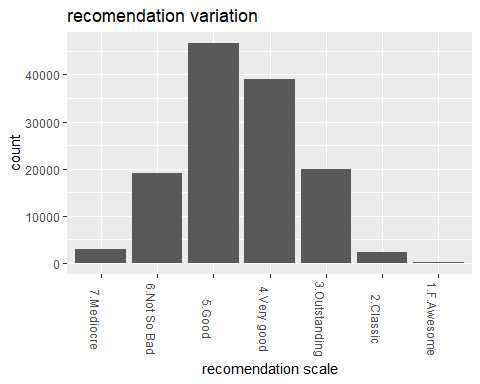
ggplot(wine\_data\_Reduced,aes(x=designation,y=count\_obs)) + geom\_col() +  
 labs(x = "designation",   
 y = "count") +  
 ggtitle("designation") +  
 theme(axis.text.x = element\_text(angle=270))



#### Designation END  
  
  
  
  
  
  
  
  
################### Feature Generation ##################  
  
# TODO Decide on factors for price groups budget, ... premium, ultra premium? then use cut to devide them into buckets.  
  
# Turn score into categories per https://www.winespectator.com/articles/scoring-scale  
wine\_data\_stats <-  
 wine\_data %>% mutate (point\_cat = cut(  
 wine\_data$points,  
 breaks = c(0, 73, 76, 79, 82, 85, 88, 91, 94, 97, 100),  
 labels = c(  
 "10.Not Recomended",  
 "9.Should Avoid",  
 "8.Not so Good",  
 "7.Mediocre",  
 "6.Not So Bad",  
 "5.Good",  
 "4.Very good",  
 "3.Outstanding",  
 "2.Classic",  
 "1.F.Awesome"  
 )  
 )) %>%   
 group\_by(point\_cat) %>%   
 summarize(count\_obs=n()) %>%   
 filter(count\_obs > 1)  
  
  
head(wine\_data\_stats)

## # A tibble: 6 x 2  
## point\_cat count\_obs  
## <fct> <int>  
## 1 7.Mediocre 2925  
## 2 6.Not So Bad 19035  
## 3 5.Good 46740  
## 4 4.Very good 38995  
## 5 3.Outstanding 19860  
## 6 2.Classic 2287

ggplot(wine\_data\_stats,aes(x=point\_cat,y=count\_obs)) + geom\_col() +  
 labs(x = "recomendation scale",   
 y = "count") +  
 ggtitle("recomendation variation") +  
 theme(axis.text.x = element\_text(angle=270))



# TODO maybe change levels to ordered and also drop unused ones  
  
# Length of title  
wine\_data <-  
 wine\_data %>% mutate (title\_length = nchar(as.character(wine\_data$title)))  
  
# Add a column to indicate wheather the wine includes a vintage # might also use str\_extract to get that vintage  
wine\_data <-  
 wine\_data %>% mutate (includes\_vintage = grepl("(19\\d{2}|20\\d{2})", title))  
  
head(wine\_data)

## X country  
## 1 0 Italy  
## 2 1 Portugal  
## 3 2 US  
## 4 3 US  
## 5 4 US  
## 6 5 Spain  
## description  
## 1 Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.  
## 2 This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016.  
## 3 Tart and snappy, the flavors of lime flesh and rind dominate. Some green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented.  
## 4 Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semidry finish.  
## 5 Much like the regular bottling from 2012, this comes across as rather rough and tannic, with rustic, earthy, herbal characteristics. Nonetheless, if you think of it as a pleasantly unfussy country wine, it's a good companion to a hearty winter stew.  
## 6 Blackberry and raspberry aromas show a typical Navarran whiff of green herbs and, in this case, horseradish. In the mouth, this is fairly full bodied, with tomatoey acidity. Spicy, herbal flavors complement dark plum fruit, while the finish is fresh but grabby.  
## designation points price province  
## 1 VulkÃ  Bianco 87 NA Sicily & Sardinia  
## 2 Avidagos 87 15 Douro  
## 3 87 14 Oregon  
## 4 Reserve Late Harvest 87 13 Michigan  
## 5 Vintner's Reserve Wild Child Block 87 65 Oregon  
## 6 Ars In Vitro 87 15 Northern Spain  
## region\_1 region\_2 taster\_name  
## 1 Etna Kerin Oâ\200\231Keefe  
## 2 Roger Voss  
## 3 Willamette Valley Willamette Valley Paul Gregutt  
## 4 Lake Michigan Shore Alexander Peartree  
## 5 Willamette Valley Willamette Valley Paul Gregutt  
## 6 Navarra Michael Schachner  
## taster\_twitter\_handle  
## 1 @kerinokeefe  
## 2 @vossroger  
## 3 @paulgwineÂ   
## 4   
## 5 @paulgwineÂ   
## 6 @wineschach  
## title  
## 1 Nicosia 2013 VulkÃ  Bianco (Etna)  
## 2 Quinta dos Avidagos 2011 Avidagos Red (Douro)  
## 3 Rainstorm 2013 Pinot Gris (Willamette Valley)  
## 4 St. Julian 2013 Reserve Late Harvest Riesling (Lake Michigan Shore)  
## 5 Sweet Cheeks 2012 Vintner's Reserve Wild Child Block Pinot Noir (Willamette Valley)  
## 6 Tandem 2011 Ars In Vitro Tempranillo-Merlot (Navarra)  
## variety winery title\_length includes\_vintage  
## 1 White Blend Nicosia 34 TRUE  
## 2 Portuguese Red Quinta dos Avidagos 45 TRUE  
## 3 Pinot Gris Rainstorm 45 TRUE  
## 4 Riesling St. Julian 67 TRUE  
## 5 Pinot Noir Sweet Cheeks 83 TRUE  
## 6 Tempranillo-Merlot Tandem 53 TRUE

# Add a column to indicate wheather the wine includes some variation of reserve  
wine\_data <-  
 wine\_data %>% mutate (is\_reserve = grepl("[Rr][ei]serv[ea]", designation))  
  
  
# TODO This is not quite right because there is still a few where they have included a year that is not the vintage  
#vintage = str\_extract(title, "(19\\d{2}|20\\d{2})") )  
  
# TODO Maybe a regex for if the title has accents in it (ie seems forign and fancy)  
  
# Get Names Vector  
dput(colnames(wine\_data))

## c("X", "country", "description", "designation", "points", "price",   
## "province", "region\_1", "region\_2", "taster\_name", "taster\_twitter\_handle",   
## "title", "variety", "winery", "title\_length", "includes\_vintage",   
## "is\_reserve")

levels(wine\_data$taster\_name)

## [1] "" "Alexander Peartree" "Anna Lee C. Iijima"  
## [4] "Anne KrebiehlÂ MW" "Carrie Dykes" "Christina Pickard"   
## [7] "Fiona Adams" "Jeff Jenssen" "Jim Gordon"   
## [10] "Joe Czerwinski" "Kerin Oâ\200\231Keefe" "Lauren Buzzeo"   
## [13] "Matt Kettmann" "Michael Schachner" "Mike DeSimone"   
## [16] "Paul Gregutt" "Roger Voss" "Sean P. Sullivan"   
## [19] "Susan Kostrzewa" "Virginie Boone"

c(  
 "X",  
 "country",  
 "description",  
 "designation",  
 "points",  
 "price",  
 "province",  
 "region\_1",  
 "region\_2",  
 "taster\_name",  
 "taster\_twitter\_handle",  
 "title",  
 "variety",  
 "winery",  
 "vintage",  
 "includes\_vintage",  
 "point\_cat",  
 "title\_length",  
 "is\_reserve"  
)

## [1] "X" "country"   
## [3] "description" "designation"   
## [5] "points" "price"   
## [7] "province" "region\_1"   
## [9] "region\_2" "taster\_name"   
## [11] "taster\_twitter\_handle" "title"   
## [13] "variety" "winery"   
## [15] "vintage" "includes\_vintage"   
## [17] "point\_cat" "title\_length"   
## [19] "is\_reserve"

cf <- fct\_lump(wine\_data$taster\_name, n = 5)  
levels(cf)

## [1] "" "Kerin Oâ\200\231Keefe" "Michael Schachner"  
## [4] "Roger Voss" "Virginie Boone" "Other"

? fct\_lump