**Sampling Plan**

Objective – use a representative sample to determine variables that predict wine points.

Target population – Top 10 countries in terms of total wines represented in the dataset.

Population Frame - 280,902 wines and their information and attributes

Operational Procedures – Data is scraped from Wine Enthusiast Magazine’s website.

Statistical Tools – Excel is used to manipulate and summarize the data. Top 10 countries are identified and compared via pivot tables. The separate countries’ data is imported into R and cleaned so that each country’s data frame is the same in preparation for further analysis.

# Sampling and cross-validation process

The population dataset is too large to run advanced analytics methods on, so sampling is required. The client requires information on success of various wines by region, country, price, and variety. They want to target future inventory with rating, price, and popularity in mind. Staying true to the project scope, the variables most important to the client are selected for the top 10 countries. A random sampling does not produce sample data representative of the population data. There is too much sampling error and the final analysis would be misguided. By selecting the top 10 countries, analyzing them separately, and identifying similarities and differences, a much better conclusion can be made in the end. Also, separating the data allows for advanced analytics methods to be utilized. The top 10 countries, United States, France, Italy, Spain, Portugal, Chile, Argentina, Australia, Austria, and New Zealand represent 95% of the population dataset and contain the wines that are more easily accessible to the client.

Cross-validation methods used include a train/test data split (70%/30%) and K-fold for each country’s sample. Using the training data to develop predictive models and the test data to identify overfitting and overall error is essential to verifying the fit and validity of a statistical model. K-fold cross-validation is necessary for developing a random forest predictive model. This was done in R. The code is provided below.

# Statistical Modeling Prep - Data Cleaning and Cross-Validation

Load in necessary packages.

options(tidyverse.quiet = TRUE)  
library(titanic)  
library(tidyverse)

## Warning: package 'stringr' was built under R version 3.5.3

library(mice)

## Loading required package: lattice

##   
## Attaching package: 'mice'

## The following object is masked from 'package:tidyr':  
##   
## complete

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(VIM)

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

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

## VIM is ready to use.   
## Since version 4.0.0 the GUI is in its own package VIMGUI.  
##   
## Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##   
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':  
##   
## sleep

library(readr)  
library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(GGally)

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

library(leaps)  
library(readxl)  
library(caret)

##   
## Attaching package: 'caret'

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

library(ranger)

# US

Read in raw dataset.

US <- read\_excel("US.xlsx")  
glimpse(US)

## Observations: 116,901  
## Variables: 8  
## $ country <chr> "US", "US", "US", "US", "US", "US", "US", "US", "U...  
## $ designation <chr> "Martha's Vineyard", "Special Selected Late Harves...  
## $ points <dbl> 96, 96, 96, 95, 95, 95, 95, 95, 95, 95, 95, 95, 95...  
## $ price <dbl> 235, 90, 65, 65, 60, 48, 48, 185, 90, 325, 75, 24,...  
## $ province <chr> "California", "California", "Oregon", "Oregon", "C...  
## $ region\_1 <chr> "Napa Valley", "Knights Valley", "Willamette Valle...  
## $ variety <chr> "Cabernet Sauvignon", "Sauvignon Blanc", "Pinot No...  
## $ winery <chr> "Heitz", "Macauley", "Ponzi", "BergstrÃ¶m", "Blue ...

Select pertinent variables that will be used in further analysis.

USwine = US %>% dplyr::select(c(points,price,province,region\_1,variety))

Convert character data types to factors.

USwine = USwine %>% mutate(province = as.factor(province)) %>%  
 mutate(region\_1 = as.factor(region\_1)) %>%  
 mutate(variety = as.factor(variety))

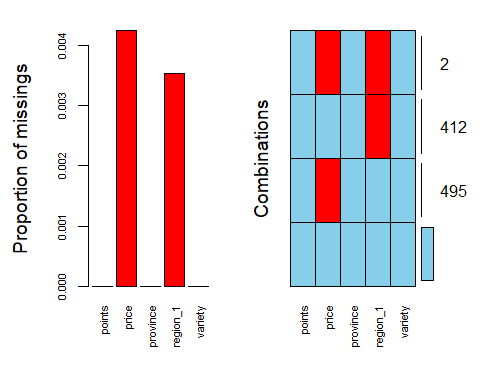
Identify missing data as “NA”

USwine$points[USwine$points==""] = NA  
USwine$province[USwine$province==""] = NA  
USwine$region\_1[USwine$region\_1==""] = NA  
USwine$variety[USwine$variety==""] = NA  
  
str(USwine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 116901 obs. of 5 variables:  
## $ points : num 96 96 96 95 95 95 95 95 95 95 ...  
## $ price : num 235 90 65 65 60 48 48 185 90 325 ...  
## $ province: Factor w/ 27 levels "America","Arizona",..: 3 3 20 20 3 20 20 20 20 3 ...  
## $ region\_1: Factor w/ 284 levels "Adelaida District",..: 152 88 276 36 234 188 57 57 276 55 ...  
## $ variety : Factor w/ 274 levels "Abouriou","Aglianico",..: 33 200 170 170 170 170 170 170 54 33 ...

Visualize missing data.

vim\_plot = aggr(USwine, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Attempt to impute missing data rather than fully remove.

imp\_price = mice(USwine, m=1, method='pmm', printFlag=FALSE)

## Warning: Number of logged events: 12

Imputation fails because of memory allocation so all NAs are removed.

USwine = USwine %>% drop\_na()  
str(USwine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 115992 obs. of 5 variables:  
## $ points : num 96 96 96 95 95 95 95 95 95 95 ...  
## $ price : num 235 90 65 65 60 48 48 185 90 325 ...  
## $ province: Factor w/ 27 levels "America","Arizona",..: 3 3 20 20 3 20 20 20 20 3 ...  
## $ region\_1: Factor w/ 284 levels "Adelaida District",..: 152 88 276 36 234 188 57 57 276 55 ...  
## $ variety : Factor w/ 274 levels "Abouriou","Aglianico",..: 33 200 170 170 170 170 170 170 54 33 ...

#write.table(USwine, file = "USwine.csv", row.names=F, sep = ",")

#### Cross-Validation (Train-Test Split and Kfold)

set.seed(1234)   
UStrain.rows = createDataPartition(y = USwine$points, p=0.7, list = FALSE)   
UStrain = USwine[UStrain.rows,]   
UStest = USwine[-UStrain.rows,]

USfit\_control = trainControl(method = "cv", number = 10)

# France

Read in raw dataset.

France <- read\_excel("France.xlsx")  
glimpse(France)

## Observations: 43,191  
## Variables: 8  
## $ country <chr> "France", "France", "France", "France", "France", ...  
## $ designation <chr> "La BrÃ»lade", "ChÃ¢teau Montus Prestige", "Le Pig...  
## $ points <dbl> 95, 95, 95, 90, 90, 90, 90, 90, 90, 90, 91, 91, 91...  
## $ price <dbl> 66, 90, 290, 22, 60, 69, 68, 42, 28, NA, 45, 38, 8...  
## $ province <chr> "Provence", "Southwest France", "Southwest France"...  
## $ region\_1 <chr> "Bandol", "Madiran", "Cahors", "Vin Mousseux", "Ch...  
## $ variety <chr> "Provence red blend", "Tannat", "Malbec", "Sparkli...  
## $ winery <chr> "Domaine de la BÃ©gude", "Vignobles Brumont", "ChÃ...

Select pertinent variables that will be used in further analysis.

Francewine = France %>% dplyr::select(c(points,price,province,region\_1,variety))

Convert character data types to factors.

Francewine = Francewine %>% mutate(province = as.factor(province)) %>%  
 mutate(region\_1 = as.factor(region\_1)) %>%  
 mutate(variety = as.factor(variety))

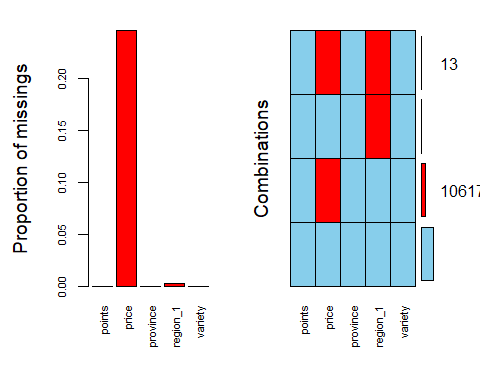
Identify missing data as “NA”

Francewine$points[Francewine$points==""] = NA  
Francewine$province[Francewine$province==""] = NA  
Francewine$region\_1[Francewine$region\_1==""] = NA  
Francewine$variety[Francewine$variety==""] = NA  
  
str(Francewine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 43191 obs. of 5 variables:  
## $ points : num 95 95 95 90 90 90 90 90 90 90 ...  
## $ price : num 66 90 290 22 60 69 68 42 28 NA ...  
## $ province: Factor w/ 11 levels "Alsace","Beaujolais",..: 9 11 11 6 10 10 4 8 8 8 ...  
## $ region\_1: Factor w/ 426 levels "Aloxe-Corton",..: 16 257 101 420 106 139 116 350 159 350 ...  
## $ variety : Factor w/ 163 levels "Abouriou","AligotÃ©",..: 112 150 62 141 117 143 26 129 109 129 ...

Visualize missing data.

vim\_plot = aggr(Francewine, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Francewine = Francewine %>% drop\_na()  
str(Francewine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 32483 obs. of 5 variables:  
## $ points : num 95 95 95 90 90 90 90 90 90 91 ...  
## $ price : num 66 90 290 22 60 69 68 42 28 45 ...  
## $ province: Factor w/ 11 levels "Alsace","Beaujolais",..: 9 11 11 6 10 10 4 8 8 4 ...  
## $ region\_1: Factor w/ 426 levels "Aloxe-Corton",..: 16 257 101 420 106 139 116 350 159 108 ...  
## $ variety : Factor w/ 163 levels "Abouriou","AligotÃ©",..: 112 150 62 141 117 143 26 129 109 26 ...

#write.table(Francewine, file = "Francewine.csv", row.names=F, sep = ",")

#### Cross-Validation (Train-Test Split and Kfold)

set.seed(1234)   
Francetrain.rows = createDataPartition(y = Francewine$points, p=0.7, list = FALSE)   
Francetrain = Francewine[Francetrain.rows,]   
Francetest = Francewine[-Francetrain.rows,]

Francefit\_control = trainControl(method = "cv", number = 10)

# Italy

Read in raw dataset.

Italy <- read\_excel("Italy.xlsx")  
glimpse(Italy)

## Observations: 43,018  
## Variables: 8  
## $ country <chr> "Italy", "Italy", "Italy", "Italy", "Italy", "Ital...  
## $ designation <chr> "Ronco della Chiesa", "Vigna Piaggia", "Riserva", ...  
## $ points <dbl> 95, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 91...  
## $ price <dbl> 80, NA, 135, 29, 23, 29, 39, 30, 90, 50, 100, 60, ...  
## $ province <chr> "Northeastern Italy", "Tuscany", "Tuscany", "Tusca...  
## $ region\_1 <chr> "Collio", "Brunello di Montalcino", "Brunello di M...  
## $ variety <chr> "Friulano", "Sangiovese", "Sangiovese", "Sangioves...  
## $ winery <chr> "Borgo del Tiglio", "Abbadia Ardenga", "Carillon",...

Select pertinent variables that will be used in further analysis.

Italywine = Italy %>% dplyr::select(c(points,price,province,region\_1,variety))

Convert character data types to factors.

Italywine = Italywine %>% mutate(province = as.factor(province)) %>%  
 mutate(region\_1 = as.factor(region\_1)) %>%  
 mutate(variety = as.factor(variety))

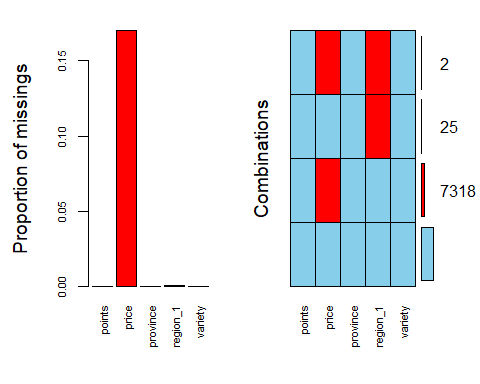
Identify missing data as “NA”

Italywine$points[Italywine$points==""] = NA  
Italywine$province[Italywine$province==""] = NA  
Italywine$region\_1[Italywine$region\_1==""] = NA  
Italywine$variety[Italywine$variety==""] = NA  
  
str(Italywine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 43018 obs. of 5 variables:  
## $ points : num 95 90 90 90 90 90 90 90 90 90 ...  
## $ price : num 80 NA 135 29 23 29 39 30 90 50 ...  
## $ province: Factor w/ 10 levels "Central Italy",..: 4 9 9 9 9 9 9 9 9 9 ...  
## $ region\_1: Factor w/ 403 levels "Abruzzo","Aglianico d'Irpinia",..: 113 54 54 400 74 74 74 74 54 54 ...  
## $ variety : Factor w/ 201 levels "Aglianico","Albana",..: 60 157 157 157 157 146 146 146 157 157 ...

Visualize missing data.

vim\_plot = aggr(Italywine, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Italywine = Italywine %>% drop\_na()  
str(Italywine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 35673 obs. of 5 variables:  
## $ points : num 95 90 90 90 90 90 90 90 90 90 ...  
## $ price : num 80 135 29 23 29 39 30 90 50 100 ...  
## $ province: Factor w/ 10 levels "Central Italy",..: 4 9 9 9 9 9 9 9 9 9 ...  
## $ region\_1: Factor w/ 403 levels "Abruzzo","Aglianico d'Irpinia",..: 113 54 400 74 74 74 74 54 54 54 ...  
## $ variety : Factor w/ 201 levels "Aglianico","Albana",..: 60 157 157 157 146 146 146 157 157 157 ...

#write.table(Italywine, file = "Italywine.csv", row.names=F, sep = ",")

#### Cross-Validation (Train-Test Split and Kfold)

set.seed(1234)   
Italytrain.rows = createDataPartition(y = Italywine$points, p=0.7, list = FALSE)   
Italytrain = Italywine[Italytrain.rows,]   
Italytest = Italywine[-Italytrain.rows,]

Italyfit\_control = trainControl(method = "cv", number = 10)

# Spain

Read in raw dataset.

Spain <- read\_excel("Spain.xlsx")  
glimpse(Spain)

## Observations: 14,913  
## Variables: 8  
## $ country <chr> "Spain", "Spain", "Spain", "Spain", "Spain", "Spai...  
## $ designation <chr> "Carodorum SelecciÃ³n Especial Reserva", "Numanthi...  
## $ points <dbl> 96, 95, 95, 95, 95, 95, 95, 90, 90, 86, 90, 90, 90...  
## $ price <dbl> 110, 73, 65, 110, 80, 79, 220, 17, 26, NA, 22, 17,...  
## $ province <chr> "Northern Spain", "Northern Spain", "Northern Spai...  
## $ region\_1 <chr> "Toro", "Toro", "Toro", "Toro", "Ribera del Duero"...  
## $ variety <chr> "Tinta de Toro", "Tinta de Toro", "Tinta de Toro",...  
## $ winery <chr> "Bodega Carmen RodrÃ­guez", "Numanthia", "Maurodos...

Select pertinent variables that will be used in further analysis.

Spainwine = Spain %>% dplyr::select(c(points,price,province,region\_1,variety))

Convert character data types to factors.

Spainwine = Spainwine %>% mutate(province = as.factor(province)) %>%  
 mutate(region\_1 = as.factor(region\_1)) %>%  
 mutate(variety = as.factor(variety))

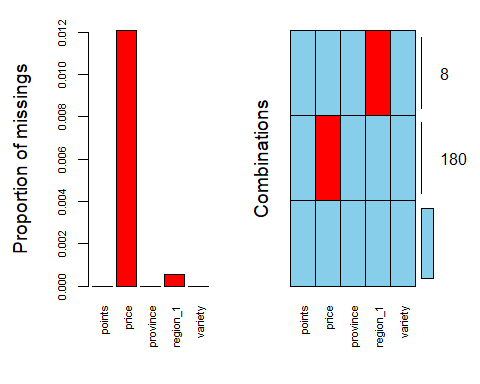
Identify missing data as “NA”

Spainwine$points[Spainwine$points==""] = NA  
Spainwine$province[Spainwine$province==""] = NA  
Spainwine$region\_1[Spainwine$region\_1==""] = NA  
Spainwine$variety[Spainwine$variety==""] = NA  
  
str(Spainwine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 14913 obs. of 5 variables:  
## $ points : num 96 95 95 95 95 95 95 90 90 86 ...  
## $ price : num 110 73 65 110 80 79 220 17 26 NA ...  
## $ province: Factor w/ 8 levels "Andalucia","Catalonia",..: 6 6 6 6 6 6 6 4 1 5 ...  
## $ region\_1: Factor w/ 87 levels "Alella","Alicante",..: 63 63 63 63 46 50 63 44 26 27 ...  
## $ variety : Factor w/ 135 levels "Airen","AlbarÃ­n",..: 113 113 113 113 105 112 113 59 73 64 ...

Visualize missing data.

vim\_plot = aggr(Spainwine, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Spainwine = Spainwine %>% drop\_na()  
str(Spainwine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 14725 obs. of 5 variables:  
## $ points : num 96 95 95 95 95 95 95 90 90 90 ...  
## $ price : num 110 73 65 110 80 79 220 17 26 22 ...  
## $ province: Factor w/ 8 levels "Andalucia","Catalonia",..: 6 6 6 6 6 6 6 4 1 6 ...  
## $ region\_1: Factor w/ 87 levels "Alella","Alicante",..: 63 63 63 63 46 50 63 44 26 50 ...  
## $ variety : Factor w/ 135 levels "Airen","AlbarÃ­n",..: 113 113 113 113 105 112 113 59 73 105 ...

#write.table(Spainwine, file = "Spainwine.csv", row.names=F, sep = ",")

#### Cross-Validation (Train-Test Split and Kfold)

set.seed(1234)   
Spaintrain.rows = createDataPartition(y = Spainwine$points, p=0.7, list = FALSE)   
Spaintrain = Spainwine[Spaintrain.rows,]   
Spaintest = Spainwine[-Spaintrain.rows,]

Spainfit\_control = trainControl(method = "cv", number = 10)

# Portugal

Read in raw dataset.

Portugal <- read\_excel("Portugal.xlsx")  
glimpse(Portugal)

## Observations: 11,013  
## Variables: 8  
## $ country <chr> "Portugal", "Portugal", "Portugal", "Portugal", "P...  
## $ designation <chr> NA, "Premium", "MarquÃªs de Marialva RosÃ© Bruto",...  
## $ points <dbl> 91, 91, 86, 86, 86, 86, 89, 93, 93, 92, 92, 92, 91...  
## $ price <dbl> 23, 15, 12, 10, 12, 7, 20, 55, 41, 40, 35, 95, 18,...  
## $ province <chr> "Alentejano", "Alentejo", "Beira Atlantico", "Dour...  
## $ region\_1 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ variety <chr> "Portuguese Red", "Portuguese Red", "Baga", "Portu...  
## $ winery <chr> "Herdade do Rocim", "Adega Cooperativa de Borba", ...

Select pertinent variables that will be used in further analysis. (Region\_1 has no data)

Portugalwine = Portugal %>% dplyr::select(c(points,price,province,variety))

Convert character data types to factors.

Portugalwine = Portugalwine %>% mutate(province = as.factor(province)) %>%  
 mutate(variety = as.factor(variety))

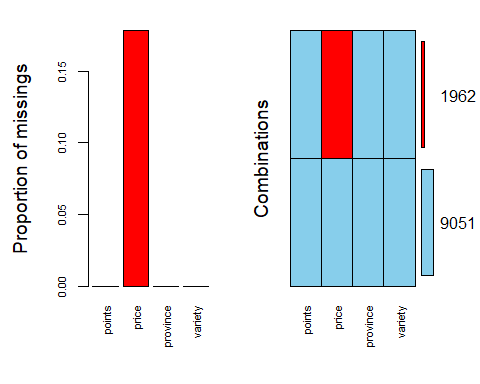
Identify missing data as “NA”

Portugalwine$points[Portugalwine$points==""] = NA  
Portugalwine$province[Portugalwine$province==""] = NA  
Portugalwine$variety[Portugalwine$variety==""] = NA  
  
str(Portugalwine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 11013 obs. of 4 variables:  
## $ points : num 91 91 86 86 86 86 89 93 93 92 ...  
## $ price : num 23 15 12 10 12 7 20 55 41 40 ...  
## $ province: Factor w/ 45 levels "Alenquer","Alentejano",..: 2 3 7 13 33 2 13 13 13 2 ...  
## $ variety : Factor w/ 96 levels "Albana","Alfrocheiro",..: 57 57 13 57 57 57 84 57 57 57 ...

Visualize missing data.

vim\_plot = aggr(Portugalwine, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Portugalwine = Portugalwine %>% drop\_na()  
str(Portugalwine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 9051 obs. of 4 variables:  
## $ points : num 91 91 86 86 86 86 89 93 93 92 ...  
## $ price : num 23 15 12 10 12 7 20 55 41 40 ...  
## $ province: Factor w/ 45 levels "Alenquer","Alentejano",..: 2 3 7 13 33 2 13 13 13 2 ...  
## $ variety : Factor w/ 96 levels "Albana","Alfrocheiro",..: 57 57 13 57 57 57 84 57 57 57 ...

#write.table(Portugalwine, file = "Portugalwine.csv", row.names=F, sep = ",")

#### Cross-Validation (Train-Test Split and Kfold)

set.seed(1234)   
Portugaltrain.rows = createDataPartition(y = Portugalwine$points, p=0.7, list = FALSE)   
Portugaltrain = Portugalwine[Portugaltrain.rows,]   
Portugaltest = Portugalwine[-Portugaltrain.rows,]

Portugalfit\_control = trainControl(method = "cv", number = 10)

# Chile

Read in raw dataset.

Chile <- read\_excel("Chile.xlsx")  
glimpse(Chile)

## Observations: 10,288  
## Variables: 8  
## $ country <chr> "Chile", "Chile", "Chile", "Chile", "Chile", "Chil...  
## $ designation <chr> "Ecos de Rulo Single Vineyard El ChequÃ©n Estate",...  
## $ points <dbl> 89, 89, 89, 89, 89, 92, 89, 89, 89, 89, 89, 90, 92...  
## $ price <dbl> 20, 19, 30, 61, 19, 30, 20, 20, 20, 25, 17, 24, 30...  
## $ province <chr> "Marchigue", "Colchagua Valley", "CuricÃ³ Valley",...  
## $ region\_1 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ variety <chr> "CarmenÃ¨re", "Red Blend", "Cabernet Sauvignon-Cab...  
## $ winery <chr> "ViÃ±a Bisquertt", "Estampa", "Santa Alba", "Casa ...

Select pertinent variables that will be used in further analysis. (Region\_1 has no data)

Chilewine = Chile %>% dplyr::select(c(points,price,province,variety))

Convert character data types to factors.

Chilewine = Chilewine %>% mutate(province = as.factor(province)) %>%  
 mutate(variety = as.factor(variety))

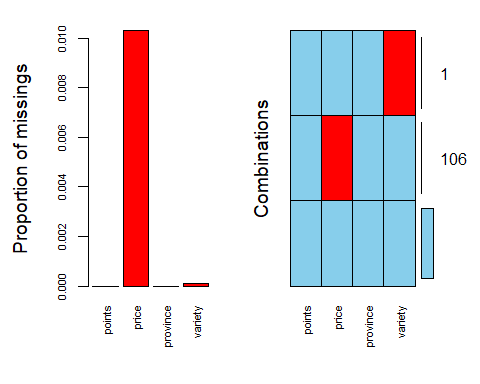
Identify missing data as “NA”

Chilewine$points[Chilewine$points==""] = NA  
Chilewine$province[Chilewine$province==""] = NA  
Chilewine$variety[Chilewine$variety==""] = NA  
  
str(Chilewine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 10288 obs. of 4 variables:  
## $ points : num 89 89 89 89 89 92 89 89 89 89 ...  
## $ price : num 20 19 30 61 19 30 20 20 20 25 ...  
## $ province: Factor w/ 44 levels "Aconcagua Costa",..: 31 16 19 32 16 16 28 9 6 16 ...  
## $ variety : Factor w/ 70 levels "Bordeaux-style Red Blend",..: 14 45 7 45 45 1 6 43 58 6 ...

Visualize missing data.

vim\_plot = aggr(Chilewine, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Chilewine = Chilewine %>% drop\_na()  
str(Chilewine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 10181 obs. of 4 variables:  
## $ points : num 89 89 89 89 89 92 89 89 89 89 ...  
## $ price : num 20 19 30 61 19 30 20 20 20 25 ...  
## $ province: Factor w/ 44 levels "Aconcagua Costa",..: 31 16 19 32 16 16 28 9 6 16 ...  
## $ variety : Factor w/ 70 levels "Bordeaux-style Red Blend",..: 14 45 7 45 45 1 6 43 58 6 ...

#write.table(Chilewine, file = "Chilewine.csv", row.names=F, sep = ",")

#### Cross-Validation (Train-Test Split and Kfold)

set.seed(1234)   
Chiletrain.rows = createDataPartition(y = Chilewine$points, p=0.7, list = FALSE)   
Chiletrain = Chilewine[Chiletrain.rows,]   
Chiletest = Chilewine[-Chiletrain.rows,]

Chilefit\_control = trainControl(method = "cv", number = 10)

# Argentina

Read in raw dataset.

Argentina <- read\_excel("Argentina.xlsx")  
glimpse(Argentina)

## Observations: 9,431  
## Variables: 8  
## $ country <chr> "Argentina", "Argentina", "Argentina", "Argentina"...  
## $ designation <chr> "The Apple Doesn't Fall Far From The Tree", "Alego...  
## $ points <dbl> 91, 91, 86, 86, 88, 87, 88, 94, 86, 86, 86, 86, 88...  
## $ price <dbl> 30, 25, 12, 15, 26, 15, 15, 50, 10, 23, 15, 14, 25...  
## $ province <chr> "Mendoza Province", "Mendoza Province", "Mendoza P...  
## $ region\_1 <chr> "Mendoza", "Mendoza", "Tupungato", "Valle de Uco",...  
## $ variety <chr> "Malbec", "Malbec", "RosÃ©", "Malbec", "Bordeaux-s...  
## $ winery <chr> "Matias Riccitelli", "Navarro Correas", "Zorzal", ...

Select pertinent variables that will be used in further analysis. (Region\_1 has no data)

Argentinawine = Argentina %>% dplyr::select(c(points,price,province,region\_1,variety))

Convert character data types to factors.

Argentinawine = Argentinawine %>% mutate(province = as.factor(province)) %>%  
 mutate(region\_1 = as.factor(region\_1)) %>%  
 mutate(variety = as.factor(variety))

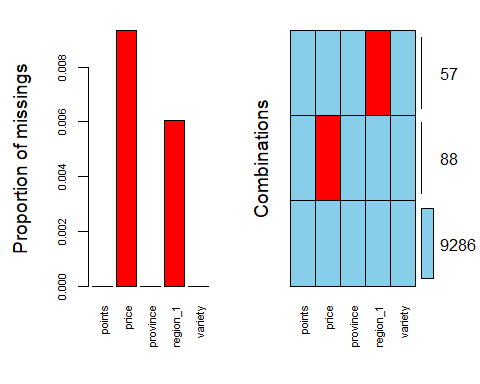
Identify missing data as “NA”

Argentinawine$points[Argentinawine$points==""] = NA  
Argentinawine$province[Argentinawine$province==""] = NA  
Argentinawine$region\_1[Argentinawine$region\_1==""] = NA  
Argentinawine$variety[Argentinawine$variety==""] = NA  
  
str(Argentinawine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 9431 obs. of 5 variables:  
## $ points : num 91 91 86 86 88 87 88 94 86 86 ...  
## $ price : num 30 25 12 15 26 15 15 50 10 23 ...  
## $ province: Factor w/ 2 levels "Mendoza Province",..: 1 1 1 1 1 1 1 1 2 1 ...  
## $ region\_1: Factor w/ 34 levels "Agrelo","Alto Valle de Uco",..: 19 19 30 32 16 30 1 19 27 19 ...  
## $ variety : Factor w/ 76 levels "Barbera","Bonarda",..: 26 26 54 26 3 26 11 11 58 26 ...

Visualize missing data.

vim\_plot = aggr(Argentinawine, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Argentinawine = Argentinawine %>% drop\_na()  
str(Argentinawine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 9286 obs. of 5 variables:  
## $ points : num 91 91 86 86 88 87 88 94 86 86 ...  
## $ price : num 30 25 12 15 26 15 15 50 10 23 ...  
## $ province: Factor w/ 2 levels "Mendoza Province",..: 1 1 1 1 1 1 1 1 2 1 ...  
## $ region\_1: Factor w/ 34 levels "Agrelo","Alto Valle de Uco",..: 19 19 30 32 16 30 1 19 27 19 ...  
## $ variety : Factor w/ 76 levels "Barbera","Bonarda",..: 26 26 54 26 3 26 11 11 58 26 ...

#write.table(Argentinawine, file = "Argentinawine.csv", row.names=F, sep = ",")

#### Cross-Validation (Train-Test Split and Kfold)

set.seed(1234)   
Argentinatrain.rows = createDataPartition(y = Argentinawine$points, p=0.7, list = FALSE)   
Argentinatrain = Argentinawine[Argentinatrain.rows,]   
Argentinatest = Argentinawine[-Argentinatrain.rows,]

Argentinafit\_control = trainControl(method = "cv", number = 10)

# Australia

Read in raw dataset.

Australia <- read\_excel("Australia.xlsx")  
glimpse(Australia)

## Observations: 7,286  
## Variables: 8  
## $ country <chr> "Australia", "Australia", "Australia", "Australia"...  
## $ designation <chr> NA, "The Dagger", "Applejack Vineyard", "The Dead ...  
## $ points <dbl> 91, 90, 88, 90, 90, 88, 90, 90, 98, 90, 90, 89, 96...  
## $ price <dbl> 36, 20, 42, 65, 65, 25, 25, 33, 125, 17, 25, 34, 1...  
## $ province <chr> "Victoria", "South Australia", "Victoria", "South ...  
## $ region\_1 <chr> "Mornington Peninsula", "Adelaide Hills", "Yarra V...  
## $ variety <chr> "Pinot Noir", "Pinot Noir", "Pinot Noir", "Shiraz"...  
## $ winery <chr> "Moorooduc", "Riposte", "Giant Steps", "D'Arenberg...

Select pertinent variables that will be used in further analysis.

Australiawine = Australia %>% dplyr::select(c(points,price,province,region\_1,variety))

Convert character data types to factors.

Australiawine = Australiawine %>% mutate(province = as.factor(province)) %>%  
 mutate(region\_1 = as.factor(region\_1)) %>%  
 mutate(variety = as.factor(variety))

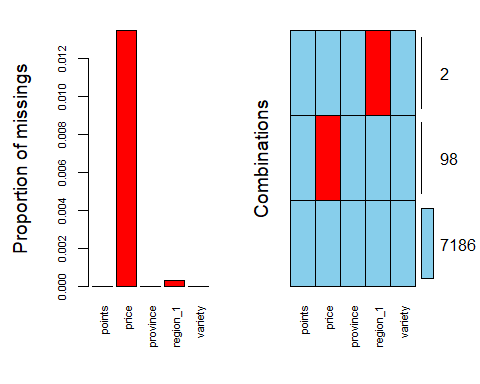
Identify missing data as “NA”

Australiawine$points[Australiawine$points==""] = NA  
Australiawine$province[Australiawine$province==""] = NA  
Australiawine$region\_1[Australiawine$region\_1==""] = NA  
Australiawine$variety[Australiawine$variety==""] = NA  
  
str(Australiawine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 7286 obs. of 5 variables:  
## $ points : num 91 90 88 90 90 88 90 90 98 90 ...  
## $ price : num 36 20 42 65 65 25 25 33 125 17 ...  
## $ province: Factor w/ 7 levels "Australia Other",..: 6 4 6 4 5 4 4 6 4 4 ...  
## $ region\_1: Factor w/ 86 levels "Adelaide","Adelaide Hills",..: 54 2 85 49 79 26 10 60 10 2 ...  
## $ variety : Factor w/ 87 levels "Alicante Bouschet",..: 50 50 50 68 50 17 68 28 68 50 ...

Visualize missing data.

vim\_plot = aggr(Australiawine, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Australiawine = Australiawine %>% drop\_na()  
str(Australiawine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 7186 obs. of 5 variables:  
## $ points : num 91 90 88 90 90 88 90 90 98 90 ...  
## $ price : num 36 20 42 65 65 25 25 33 125 17 ...  
## $ province: Factor w/ 7 levels "Australia Other",..: 6 4 6 4 5 4 4 6 4 4 ...  
## $ region\_1: Factor w/ 86 levels "Adelaide","Adelaide Hills",..: 54 2 85 49 79 26 10 60 10 2 ...  
## $ variety : Factor w/ 87 levels "Alicante Bouschet",..: 50 50 50 68 50 17 68 28 68 50 ...

#write.table(Australiawine, file = "Australiawine.csv", row.names=F, sep = ",")

#### Cross-Validation (Train-Test Split and Kfold)

set.seed(1234)   
Australiatrain.rows = createDataPartition(y = Australiawine$points, p=0.7, list = FALSE)   
Australiatrain = Australiawine[Australiatrain.rows,]   
Australiatest = Australiawine[-Australiatrain.rows,]

Australiafit\_control = trainControl(method = "cv", number = 10)

# Austria

Read in raw dataset.

Austria <- read\_excel("Austria.xlsx")  
glimpse(Austria)

## Observations: 6,402  
## Variables: 8  
## $ country <chr> "Austria", "Austria", "Austria", "Austria", "Austr...  
## $ designation <chr> NA, NA, NA, "The Butcher", "The Butcher", "Classic...  
## $ points <dbl> 93, 93, 93, 93, 93, 93, 93, 93, 92, 92, 92, 92, 92...  
## $ price <dbl> 21, 19, 25, NA, NA, 27, NA, 19, 17, 18, 17, 18, 22...  
## $ province <chr> "Thermenregion", "Burgenland", "Carnuntum", "Burge...  
## $ region\_1 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ variety <chr> "St. Laurent", "Blauburgunder", "BlaufrÃ¤nkisch", ...  
## $ winery <chr> "Johanneshof Reinisch", "Meinklang", "Muhr-Van der...

Select pertinent variables that will be used in further analysis.

Austriawine = Austria %>% dplyr::select(c(points,price,province,variety))

Convert character data types to factors.

Austriawine = Austriawine %>% mutate(province = as.factor(province)) %>%  
 mutate(variety = as.factor(variety))

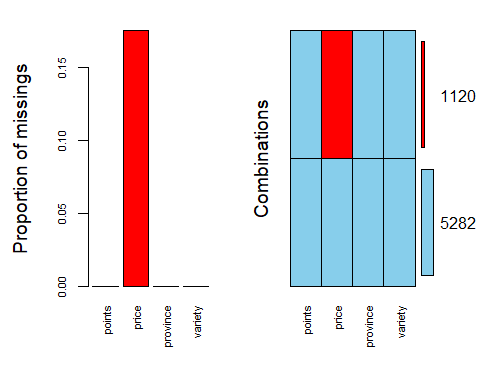
Identify missing data as “NA”

Austriawine$points[Austriawine$points==""] = NA  
Austriawine$province[Austriawine$province==""] = NA  
Austriawine$variety[Austriawine$variety==""] = NA  
  
str(Austriawine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 6402 obs. of 4 variables:  
## $ points : num 93 93 93 93 93 93 93 93 92 92 ...  
## $ price : num 21 19 25 NA NA 27 NA 19 17 18 ...  
## $ province: Factor w/ 30 levels "Ã–sterreichischer Perlwein",..: 22 4 5 4 4 15 15 22 26 8 ...  
## $ variety : Factor w/ 62 levels "Austrian Red Blend",..: 53 3 5 38 5 38 21 38 21 21 ...

Visualize missing data.

vim\_plot = aggr(Austriawine, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Austriawine = Austriawine %>% drop\_na()  
str(Austriawine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 5282 obs. of 4 variables:  
## $ points : num 93 93 93 93 93 92 92 92 92 92 ...  
## $ price : num 21 19 25 27 19 17 18 17 18 22 ...  
## $ province: Factor w/ 30 levels "Ã–sterreichischer Perlwein",..: 22 4 5 15 22 26 8 26 8 15 ...  
## $ variety : Factor w/ 62 levels "Austrian Red Blend",..: 53 3 5 38 38 21 21 21 21 21 ...

#write.table(Austriawine, file = "Austriawine.csv", row.names=F, sep = ",")

#### Cross-Validation (Train-Test Split and Kfold)

set.seed(1234)   
Austriatrain.rows = createDataPartition(y = Austriawine$points, p=0.7, list = FALSE)   
Austriatrain = Austriawine[Austriatrain.rows,]   
Austriatest = Austriawine[-Austriatrain.rows,]

Austriafit\_control = trainControl(method = "cv", number = 10)

# NZ

Read in raw dataset.

NZ <- read\_excel("New Zealand.xlsx")  
glimpse(NZ)

## Observations: 4,660  
## Variables: 8  
## $ country <chr> "New Zealand", "New Zealand", "New Zealand", "New ...  
## $ designation <chr> "MatÃ©'s Vineyard", "MatÃ©'s Vineyard", "Coddingto...  
## $ points <dbl> 94, 94, 93, 93, 93, 92, 92, 92, 92, 92, 92, 92, 92...  
## $ price <dbl> 57, 57, 46, 33, 40, 63, 49, 25, 34, 65, 61, 29, 28...  
## $ province <chr> "Kumeu", "Kumeu", "Kumeu", "Marlborough", "Central...  
## $ region\_1 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ variety <chr> "Chardonnay", "Chardonnay", "Chardonnay", "Sauvign...  
## $ winery <chr> "Kumeu River", "Kumeu River", "Kumeu River", "Spy ...

Select pertinent variables that will be used in further analysis.

NZwine = NZ %>% dplyr::select(c(points,price,province,variety))

Convert character data types to factors.

NZwine = NZwine %>% mutate(province = as.factor(province)) %>%  
 mutate(variety = as.factor(variety))

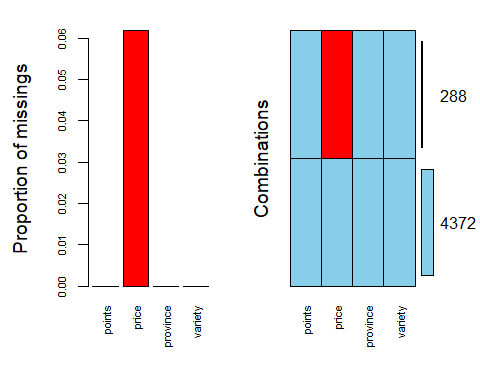
Identify missing data as “NA”

NZwine$points[NZwine$points==""] = NA  
NZwine$province[NZwine$province==""] = NA  
NZwine$variety[NZwine$variety==""] = NA  
  
str(NZwine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 4660 obs. of 4 variables:  
## $ points : num 94 94 93 93 93 92 92 92 92 92 ...  
## $ price : num 57 57 46 33 40 63 49 25 34 65 ...  
## $ province: Factor w/ 28 levels "Awatere Valley",..: 10 10 10 11 3 13 10 16 10 3 ...  
## $ variety : Factor w/ 40 levels "Bordeaux-style Red Blend",..: 9 9 9 30 26 23 9 9 9 23 ...

Visualize missing data.

vim\_plot = aggr(NZwine, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



NZwine = NZwine %>% drop\_na()  
str(NZwine)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 4372 obs. of 4 variables:  
## $ points : num 94 94 93 93 93 92 92 92 92 92 ...  
## $ price : num 57 57 46 33 40 63 49 25 34 65 ...  
## $ province: Factor w/ 28 levels "Awatere Valley",..: 10 10 10 11 3 13 10 16 10 3 ...  
## $ variety : Factor w/ 40 levels "Bordeaux-style Red Blend",..: 9 9 9 30 26 23 9 9 9 23 ...

#write.table(NZwine, file = "NZwine.csv", row.names=F, sep = ",")

#### Cross-Validation (Train-Test Split and Kfold)

set.seed(1234)   
NZtrain.rows = createDataPartition(y = NZwine$points, p=0.7, list = FALSE)   
NZtrain = NZwine[NZtrain.rows,]   
NZtest = NZwine[-NZtrain.rows,]

NZfit\_control = trainControl(method = "cv", number = 10)