HarvardX - Data Science Professional Certificate Capstone Project - Chocolate Bar Rating System -

Anass Latif

April 20, 2019

Contents

1. Introduction / Overview / Executive Summary

Background and Motivation

Chocolate is one of the most popular candies in the world. Canadian eat an average of **6.4** kilograms of chocolate a year, which base on an average bar size, is at least 160 chocolate bars per year, per person. However, citizens around the world have different tastes for different kinds of chocolate and not all chocolate bars are created equal!

This **Harvard Data Science Capstone** is the final assignment for HarvardX - Data Science Professional Certificate from Harvard University.

The purpose of this report is to highlight customer satisfaction of chocolate bars based on different variables such as cocoa percentage, cocoa bean type, cocoa bean origin, manufacturer contry, etc.

Dataset

For this assignment, we will go throught all the steps to create a predictive model using the Chocolate Bar Rating dataset, available in Kaggle.

This dataset contains expert ratings of over 1,750 individual chocolate bars, along with information on their regional origin, percentage of cocoa, the variety of chocolate bean used and where the beans were grown.

These ratings were compiled by Brady Brelinski, Founding Member of the Manhattan Chocolate Society. For up-to-date information, as well as additional content (including interviews with craft chocolate makers), please see his website: Flavors of Cacao.

Goal

The objective of this report is to predict, in the most accurate and comprehensive way, the ***Chocolate Bar Rating Class*** by implementing, testing and validating different machine learning algorithms.

The chocolate rating class is based on the **Flavors of Cacao Rating System**

- 5= Elite (Transcending beyond the ordinary limits) - Rating = 5.00
- 4= Premium (Superior flavor development, character and style) - Rating = 4.00 to 4.75
- 3= Satisfactory(3.0) to praiseworthy(3.75) (well made with special qualities) - Rating = 3.00 to 3.75
- 2= Disappointing (Passable but contains at least one significant flaw) - Rating = 2.00 to 2.75
- 1= Unpleasant (mostly unpalatable) - Rating = 1.00 to 1.75

Key Steps

To achieve the project objectives, we will follow a comprehensive machine learning workflow:

1. Download the Chocolate Bar Ratings dataset
2. Explore the dataset to discover the data and the available features. We will use some exploratory techniques such as data description, preparation, exploration and visualization.
3. Split the dataset into training (`training_set`) and test (`validation_set`) datasets
4. Develop and train different predictive models and algorithms in order to find a recommendation model with the best possible outcome (Accuracy).

5. Explain the results and conclude.

All the project will be made through RStudio (version 3.5.3) using some useful packages (eg: dplyr, tidyverse, lubridate, caret, etc.).

This report doesn't display the R code used to generate the information. All scripts are available in My GitHub Repository.

2. Methods / Analysis

Data Preparation

Dataset Generation

The generated dataset is downloaded from a CSV file available in my GitHub repository. The file contains some non-printable characters that have been removed during the generation.

Dataset Description

The dataset contains ***1795*** observations (rows) of ***9*** variables (columns).

There is some missing values (***962*** na in total):

Column_Name	Missing_Values
CompanyName	0
ChocolateBarName	0
Reference	0
ReviewYear	0
CocoaPercentage	0
CompanyCountry	0
Rating	0
BeanType	888
BeanOrigin	74

The features / variables identified in the dataset are:

- `CompanyName`: Name of the company manufacturing the Chocolate bar.
- `ChocolateBarName`: The specific species, the geo-region of origin for the bar, or the bar name.
- `Reference`: A value linked to when the review was entered in the database. Higher = more recent.
- `ReviewYear`: Year of publication of the review.
- `CocoaPercentage`: Cocoa percentage (darkness) of the chocolate bar being reviewed.
- `CompanyCountry`: Manufacturer base country.
- `Rating`: Expert rating for the Chocolate bar from 1 to 5 with 0.25 increment.
- `BeanType`: The variety (breed) of bean used, if provided.
- `BeanOrigin`: The broad country or geo-region of origin for the bean, if provided.

Let's display a sample of the dataset.

CompanyName	ChocolateBarName	Reference	ReviewYear	CocoaPercentage	CompanyCountry	Rating	BeanType	BeanOrigin
A. Morin	Agua Grande	1876	2016	63%	France	3.75	NA	Sao Tome
A. Morin	Kpime	1676	2015	70%	France	2.75	NA	Togo
A. Morin	Atsane	1676	2015	70%	France	3.00	NA	Togo
A. Morin	Akata	1680	2015	70%	France	3.50	NA	Togo
A. Morin	Quilla	1704	2015	70%	France	3.50	NA	Peru
A. Morin	Carenero	1315	2014	70%	France	2.75	Criollo	Venezuela

Dataset Preprocessing (Feature Selection and Engineering)

Based on our data description, we notice that the data is not normalized and needs some cleaning and transformation to be usable in our Exploratory Data Analysis.

We apply the following cleaning and transformation rules to our dataset:

- Standardize the `BeanOrigin` column data based on the Country Name or Sub-Region of the Country based on ISO-3166 standards
- Correct the Misspellings in `CompanyCountry` column to be compliant with ISO-3166 standards
- Group `BeanType` by specy. Five main bean type groups are identified: Criollo, Forastero, Nacional, Trinitario and Blend
- Replace missing values for `BeanType` to `Blend` based on the assumption
  - Multiple countries for `BeanOrigin` and `BeanType` is missing (`na`)
  - `ChocolateBarName` contains `Blend` or `blend` or `,` and `BeanType` is missing (`na`)
- Convert `CocoaPercentage` column to `Numeric` by removing `%` sign and rounding to the nearest integer

In addition, we engineer new features (variables) that we might use to build our preditive model:

- Create a new variable `RatingScale` based on the ***Flavor Of Cocoa Rating System***
- Create a new variable `BeanOriginGeoRegion` providing the Geo-region based on the `BeanOrigin`
- Create a new variable `CompanyGeoRegion` providing the Geo-region based on the `CompanyCountry`

We notice that the `BeanOrigin` are pipe-separated values after the transformation. We might need this feature `BeanOrigin` to predict the Chocolate bar rating. However, we should extract individual value for more consistent and robust estimate.

Finally, we convert all variables to their corresponding data type and remove `Reference` feature as it is time related to `ReviewYear`.

After preprocessing the data, the `chocolate_data_clean` dataset looks like:

ChocolateBarName	CompanyName	CompanyCountry	CompanyGeoRegion	BeanType	BeanOrigin	BeanOriginGeoRegion	CocoaPercentage	ReviewYear	Rating	RatingClass
Agua Grande	A. Morin	France	Western Europe	NA	Sao Tome and Principe	Central Africa	63	2016	3.75	30-Satisfactory
Kpime	A. Morin	France	Western Europe	NA	Togo	Western Africa	70	2015	2.75	20-Disappointing
Atsane	A. Morin	France	Western Europe	NA	Togo	Western Africa	70	2015	3.00	30-Satisfactory
Akata	A. Morin	France	Western Europe	NA	Togo	Western Africa	70	2015	3.50	30-Satisfactory
Quilla	A. Morin	France	Western Europe	NA	Peru	South America	70	2015	3.50	30-Satisfactory
Carenero	A. Morin	France	Western Europe	Criollo	Venezuela	South America	70	2014	2.75	20-Disappointing

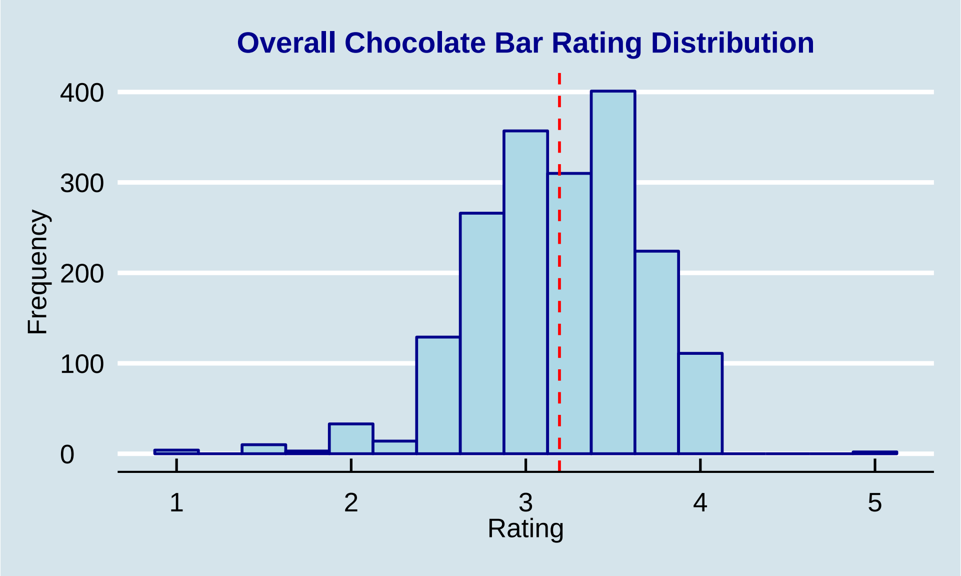
To summarize, the features that could be selected in the machine learning models to predict the `Chocolate Bar Rating` are:

Feature_Name	Data_Type	Distinct_Values	Missing_Values
ChocolateBarName	character	1039	0
CompanyName	factor	416	0
CompanyCountry	factor	54	0
CompanyGeoRegion	factor	17	0
BeanType	factor	6	607
BeanOrigin	factor	54	74
BeanOriginGeoRegion	factor	13	74
CocoaPercentage	numeric	42	0
ReviewYear	factor	12	0
Rating	numeric	13	0
RatingClass	factor	5	0

Exploratory Data Analysis (EDA)

Let's explore our `chocolate_data_clean` dataset using some visualization techniques to build more comprehensive understanding of the data. So, here are some questions that we raise:

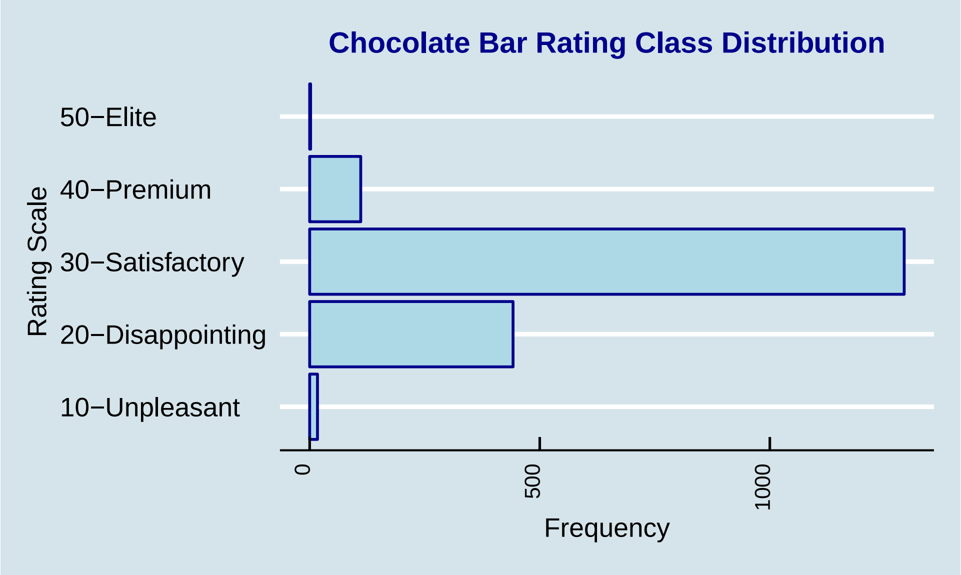
1. What is the Overall Chocolate Bar Ratings distribution?



The figure *"Overall Chocolate Bar Rating Distribution"* shows the rating distribution in our dataset. The vertical dashed line represents the overall rating average μ (***3.193***) across all Chocolate bars. We notice also that the rating range from 1 to 4 with an exception for a few Chocolate bars that are rated as Elite or Unpleasant chocolate.

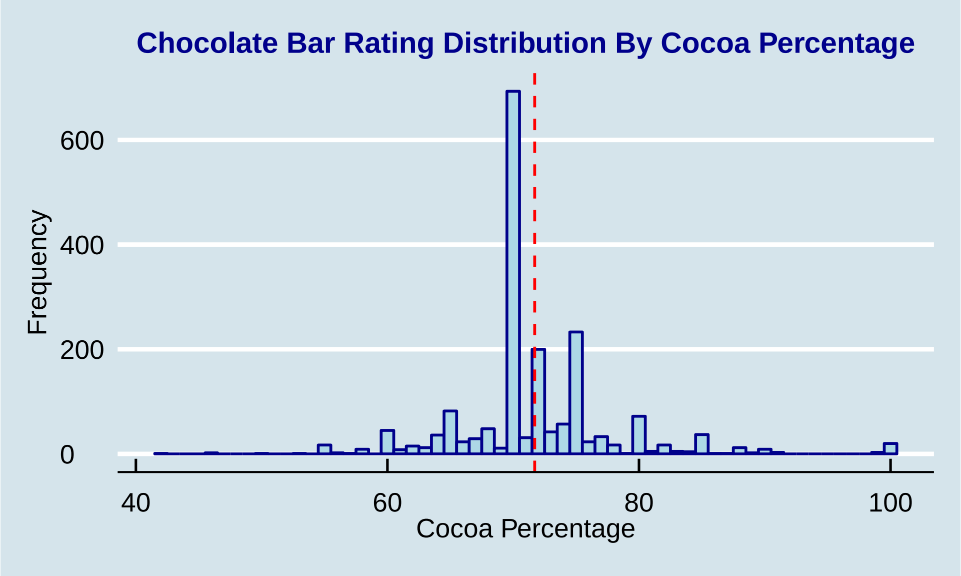
We also notice that the most common rating is 3.5 by around 400 ratings followed closely by 3.0.

2. What is the Chocolate Bar Rating Class Distribution?



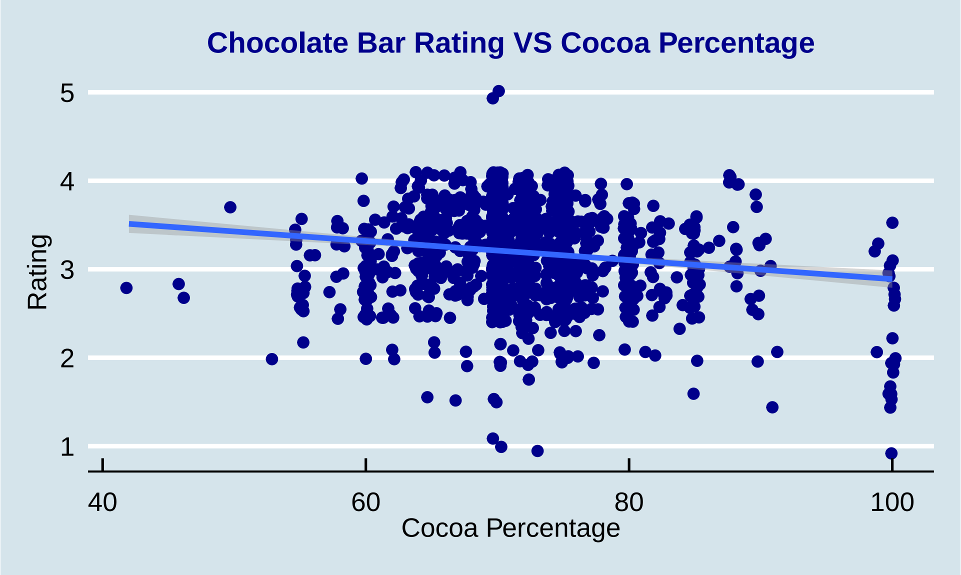
The figure *"Chocolate Bar Rating Class Distribution"* confirm the rating districution. Most of the chocolate bar ratings are *"Satisfactory"* (Class 3) followed by the *"Premium"* Chocolate bars.

3. What is the Chocolate Bar Ratings distribution by Cocoa Percentage?



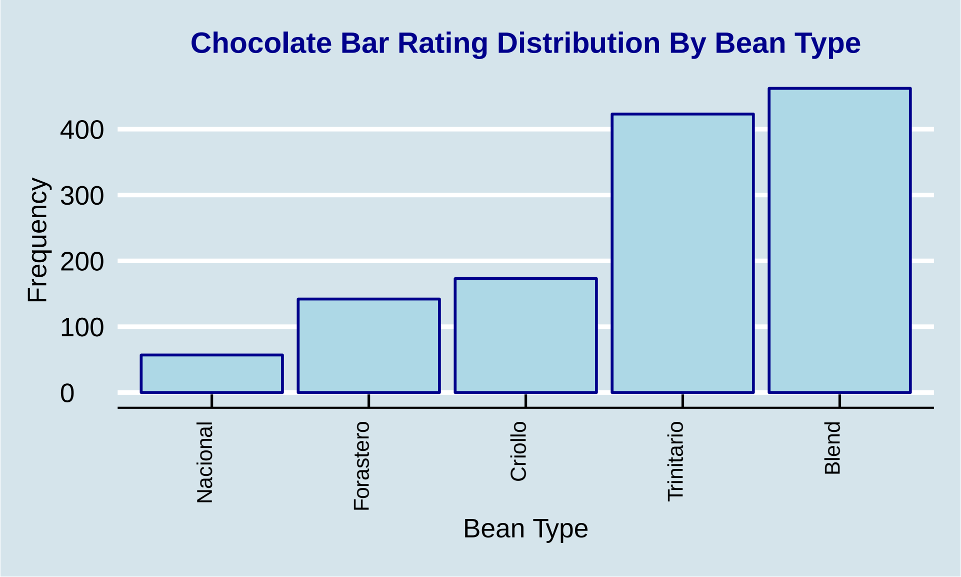
The figure *"Chocolate Bar Rating Distribution By Cocoa Percentage"* shows the rating distribution by Cocoa Percentage. The vertical dashed line represents the Cocoa Percentage average μ (***71.707***) across all Chocolate bars. The figure shows that most of the chocolate bars are made with around 65% and 75% of cacao in the bar, with around 700 chocolate bars being made with 70% cacao.

4. Is there any correlation between Chocolate Bar Ratings and Cocoa Percentage?

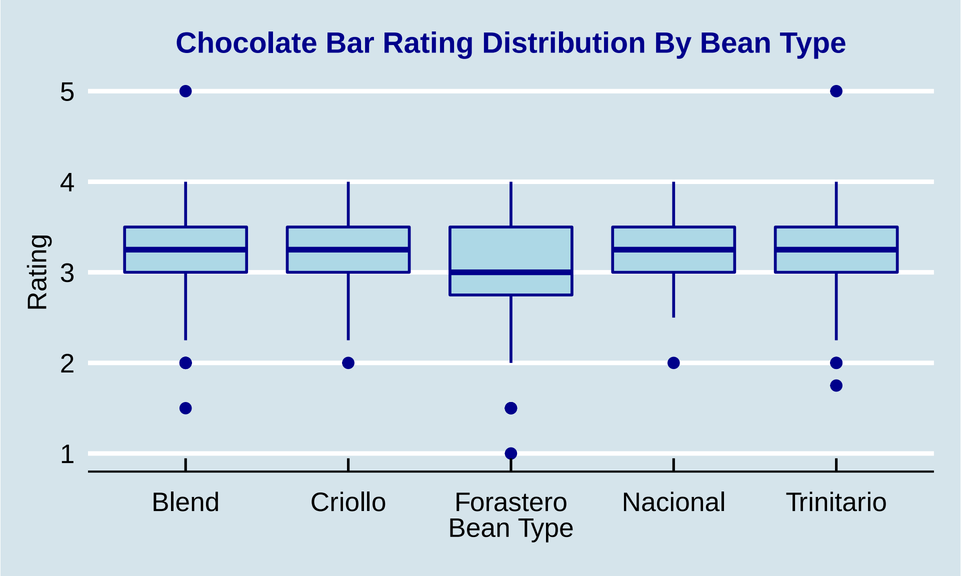


This figure *"Chocolate Bar Rating VS Cocoa Percentage"* shows that as the cocoa percent increases, the perceived rating of chocolate decreases slightly.

5. What is the Chocolate Bar Ratings distribution by Bean Type?

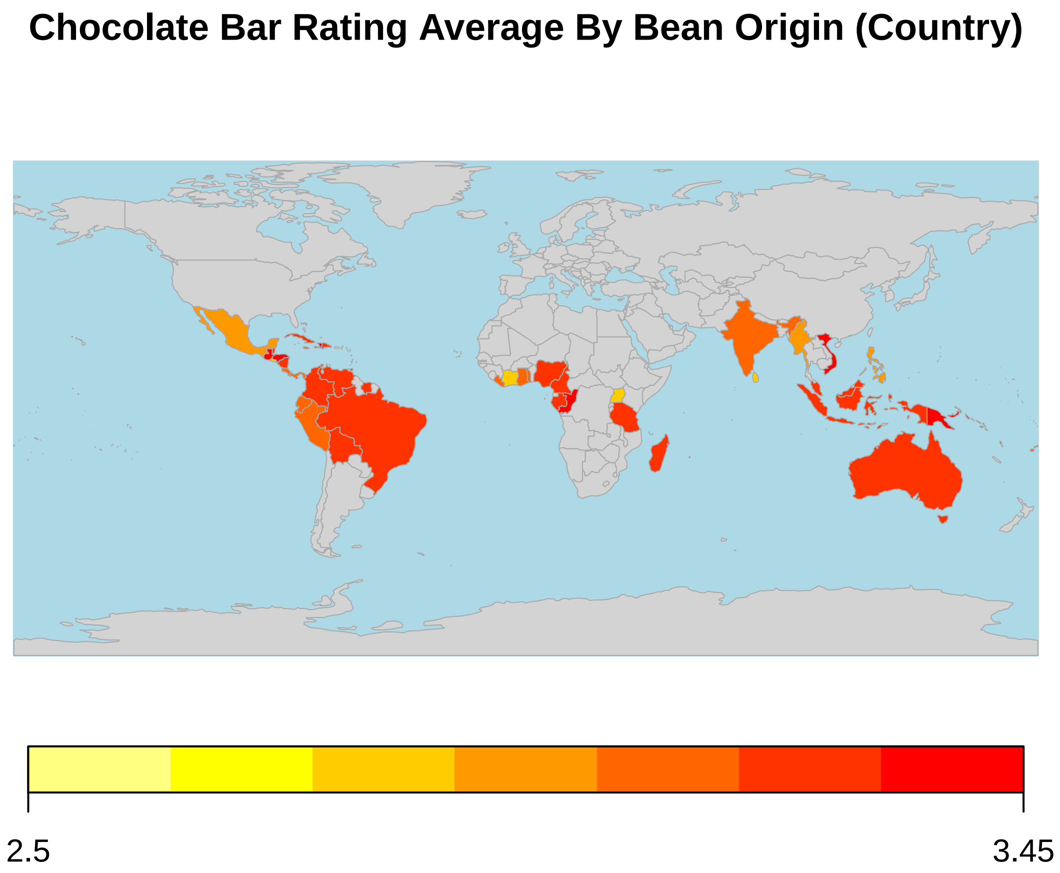


The figure *"Chocolate Bar Rating Distribution By Bean Type"* shows the rating distribution by Bean Type. The most bean types used are Blend and Trinitario. However, most of the manufacturers don't provide Bean Type (607 observations). We assume that they are preserving the secret of their recepies.



The above Box-Plot figure shows no such difference in the data distribution by Bean Type. This confirms that there is no strong correlation between Ratings and Bean Type.

6. What Is The Chocolate Bar Rating Average By Bean Origin (Country)?



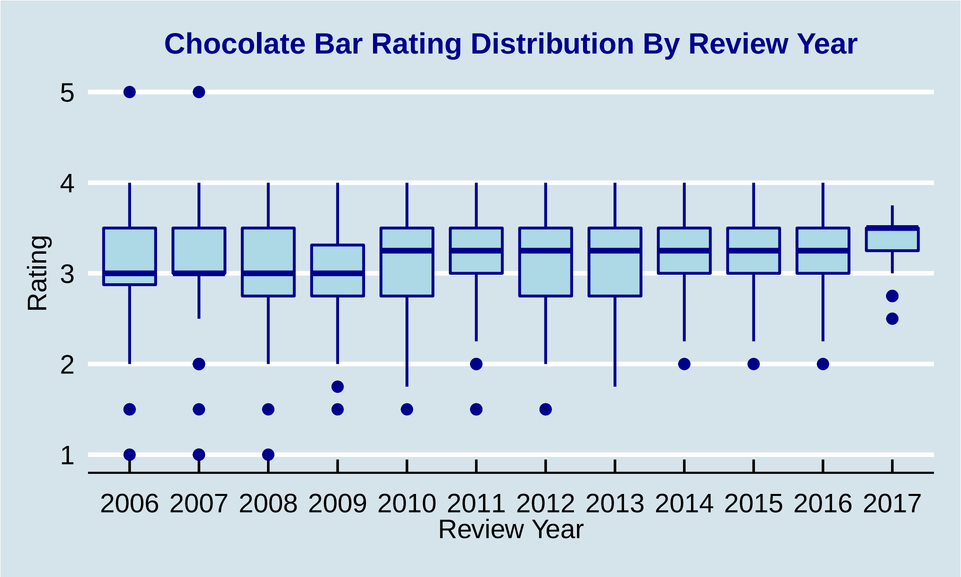
The map *"Chocolate Bar Rating Distribution By Bean Origin (Country)"* shows the rating world distribution by Bean Origin. The most rated chocolate bars have bean origin from Venezuela, Ecuador, Dominican Republic, Peru and Madagascar.

7. What is the top 10 ranking of the Chocolate Bar Rating Average by Bean Origin (Country)?

BeanOrigin	Rating_Count	Rating_Average
Haiti	10	3.45
Honduras	15	3.35
Guatemala	29	3.35
Papua New Guinea	47	3.33
Republic of the Congo	10	3.33
Vietnam	38	3.32
Indonesia	21	3.30
Brazil	59	3.29
Madagascar	156	3.27
Cuba	11	3.25

The table above shows that the best rated chocolate bars are made from cocoa beans that grown in Caribbean and South Pacific.

8. What is the Chocolate Bar Ratings distribution by Review Year?



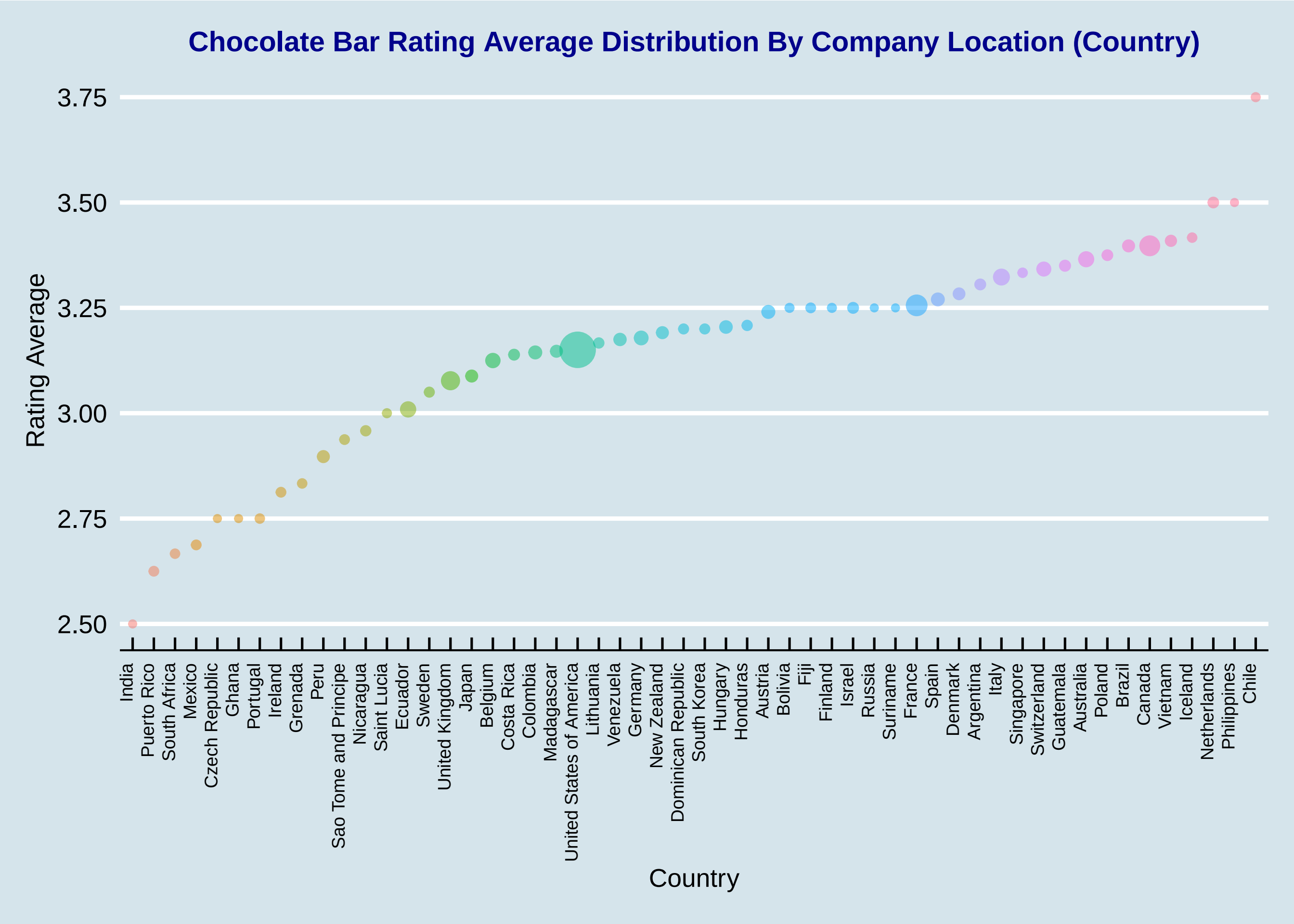
The figure *"Chocolate Bar Rating Distribution By Review Year"* shows that the number of chocolate bars ratings varies from year to year. But, the most of the chocolate ratings are distributed around the average. Also, we notice that there is less variation in the ratings in the recent years.

9. What are the top 10 Chocolate Bar Rating Average by Company Name (at least 10 ratings)?

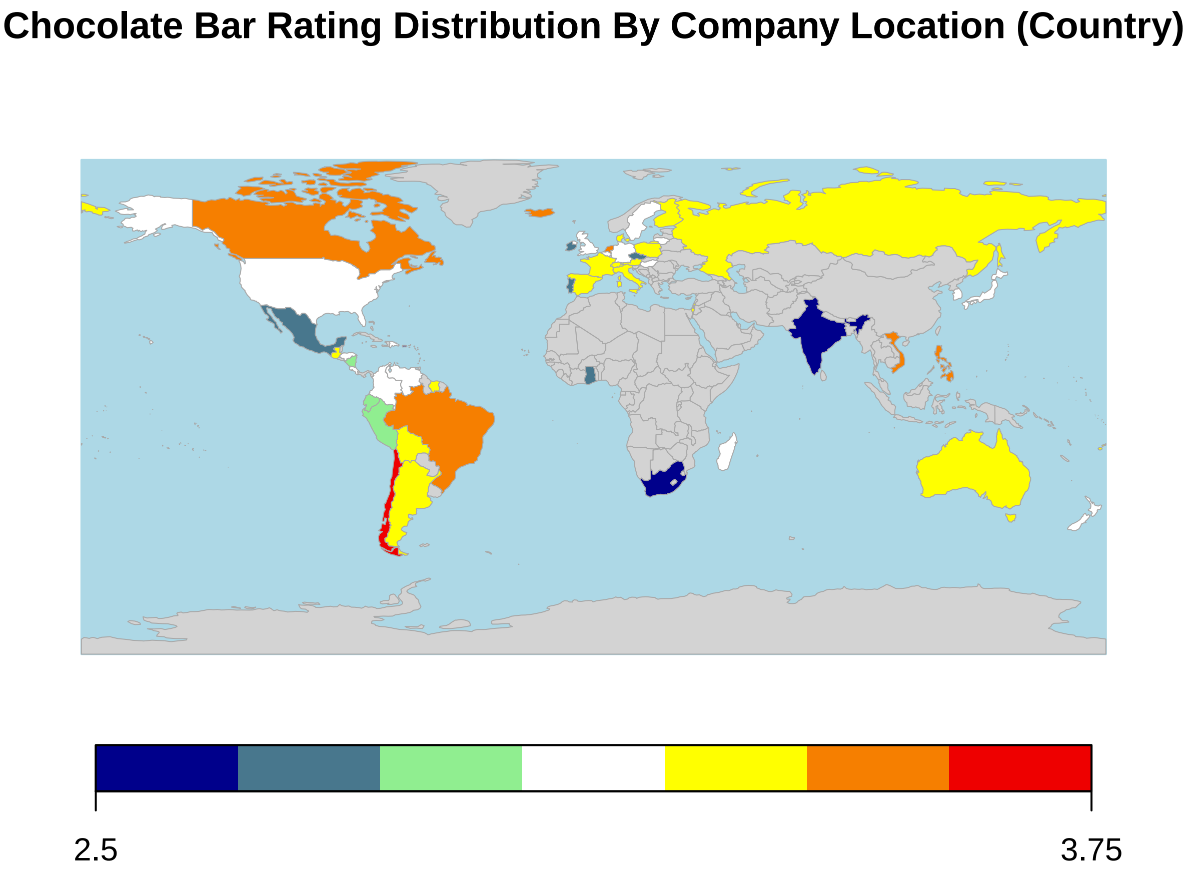
CompanyName	CompanyCountry	Rating_Count	Rating_Average
Amedei	Italy	13	3.85
Idilio (Felchlin)	Switzerland	10	3.77
Soma	Canada	67	3.66
Arete	United States of America	22	3.53
Smooth Chocolator, The	Australia	16	3.52
Duffy's	United Kingdom	13	3.50
Pierre Marcolini	Belgium	16	3.50
Domori	Italy	22	3.48
Bonnat	France	31	3.48
Marou	Vietnam	10	3.45

The table above shows that only Amedei from Italy and Idilio (Felchlin) from Switzerland are the most rated companies with a Premium quality chocolate bars (Based on at least 10 ratings).

10. What is the Chocolate Bar Rating Average Distribution By Company Location (Country)?



The figure *"Chocolate Bar Average Rating Distribution By Company Location (Country)"* shows that chocolate bar ratings vary depending on the Company location (Country).


Chocolate Bar Rating Distribution By Company Location (Country)
2.5
3.75

The map *"Chocolate Bar Rating Distribution By Company Location (Country)"* shows the rating world distribution of chocolate companies.

11. What are the top 10 Chocolate Bar Rating Average by Company Country (at least 10 ratings)?

CompanyCountry	Rating_Count	Rating_Average
Vietnam	11	3.41
Canada	146	3.40
Brazil	17	3.40
Australia	52	3.37
Guatemala	10	3.35
Switzerland	38	3.34
Italy	65	3.32
Denmark	15	3.28
Spain	25	3.27
France	168	3.26

The table above shows that the companies from Vietnam, Canada and Brazil are the most rated companies (Based on at least 10 ratings).

Model Building, Training and Validation

Key steps

After an in-depth Exploratory Data Analysis, we are ready to build, train and test different algorithms and models to reach our goal that provide the best accuracy on the validation dataset.

Based on our Exploratory Data Analysis, most of the available features are categorical (except the `CocoaPercentage`). Our models will be using the following features to predict the `RatingClass`:

- `CocoaPercentage`
- `BeanType`
- `BeanOrigin` (Country)
- `CompanyCountry`
- `ReviewYear`

To build and train the different models, we will proceed, for each model, in five sequential steps:

1. Split our `chocolate_data_clean` into two datasets with a breakdown 70% and 30%:

- Training dataset `training_set`
- Validation dataset `validation_set`

2. Define and Build the model
3. Train and tune the algorithm in the training dataset `training_set`
4. Validate the algorithm by runing the predictions in the validation dataset `validation_set`
5. Iterate over the models until goal satisfaction

Our goal is to predict with the highest acuracy the `RatingClass` of the Chocolate Bars. So, we decided to test the following classification and regression models:

- Model 1: Support Vector Machine (SVM)
- Model 2: K-Nearest Neighbors (KNN)
- Model 3: Random Forest (RF)
- Model 4: Learning Vector Quantization (LVQ)
- Model 5: Stochastic Gradient Boosting Machine (GBM)
- Model 6: Classification And Regression Tree (CART)

For all these models, we will be using the `caret` package for its simplicity. This package provide a simplified way to build machine learning algorithm with almost the same syntax.

Model 1 - Support Vector Machine (SVM)

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N-the number of features) that distinctly classifies the data points.

The equation of the support vector classifier is:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x_i, y_i)$$

- S are the support vectors
- α is a weight value which is non-zero for all support vectors and otherwise 0
- $K(x_i, y_i)$ is the Kernel Function that will use the "Radial kernel"

$$K(x, y) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} y_{ij})^2)$$

Let's apply this model to our `Chocolate Bars Rating` dataset and verify the overall accuracy on the `validation_set` dataset.

sigma	C	Accuracy	Kappa	AccuracySD	KappaSD
0.086	0.25	0.713	0.000	0.005	0.00
0.086	0.50	0.713	0.000	0.005	0.00
0.086	1.00	0.708	0.001	0.011	0.03

	Measure_Value
Accuracy	0.715
Kappa	0.000
AccuracyLower	0.666
AccuracyUpper	0.761
AccuracyNull	0.715
AccuracyPValue	0.526
McnemarPValue	NaN

Model_Id	Model_Method	Accuracy_On_Training	**Accuracy_On_Validation**
SVM	Support Vector Machine	0.713	**0.715**

The predicted ***Accuracy*** on the `validation_set` dataset for the ***Support Vector Machine*** is about ***0.715***.

Model 2 - K-Nearest Neighbors (KNN)

K-nearest neighbor algorithm is very simple. It works based on minimum distance from the query instance to the training samples to determine the K-nearest neighbors. After we gather K nearest neighbors, we take simple majority of these K-nearest neighbors to be the prediction of the query instance.

Let's apply this model to our `Chocolate Bars Rating` dataset and verify the overall accuracy on the `validation_set` dataset.

k	Accuracy	Kappa	AccuracySD	KappaSD
5	0.679	0.035	0.032	0.072
7	0.685	0.008	0.030	0.085
9	0.689	0.004	0.024	0.063

	Measure_Value
Accuracy	0.699
Kappa	0.014
AccuracyLower	0.649
AccuracyUpper	0.745
AccuracyNull	0.715
AccuracyPValue	0.776
McnemarPValue	NaN

Model_Id	Model_Method	Accuracy_On_Training	**Accuracy_On_Validation**
SVM	Support Vector Machine	0.713	**0.715**
KNN	K-Nearest Neighbors	0.689	**0.699**

The predicted ***Accuracy*** on the `validation_set` dataset for the ***K-Nearest Neighbors*** is about ***0.699***.

Model 3 - Random Forest (RF)

Random Forest is a supervised learning algorithm. It creates a forest and makes it somehow random. The "forest" it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Let's apply this model to our `Chocolate Bars Rating` dataset and verify the overall accuracy on the `validation_set` dataset.

mtry	Accuracy	Kappa	AccuracySD	KappaSD
2	0.714	0.000	0.005	0.000
61	0.676	0.095	0.033	0.111
121	0.667	0.100	0.031	0.100

	Measure_Value
Accuracy	0.715
Kappa	0.000
AccuracyLower	0.666
AccuracyUpper	0.761
AccuracyNull	0.715
AccuracyPValue	0.526
McnemarPValue	NaN

Model_Id	Model_Method	Accuracy_On_Training	**Accuracy_On_Validation**
SVM	Support Vector Machine	0.713	**0.715**
KNN	K-Nearest Neighbors	0.689	**0.699**
RF	Random Forest	0.714	**0.715**

The predicted ***Accuracy*** on the `validation_set` dataset for the ***Random Forest*** is about ***0.715***.

Model 4 - Learning Vector Quantization (LVQ)

The Learning Vector Quantization algorithm (or LVQ for short) is an artificial neural network algorithm that lets you choose how many training instances to hang onto and learns exactly what those instances should look like.

Let's apply this model to our `Chocolate Bars Rating` dataset and verify the overall accuracy on the `validation_set` dataset.

size	k	Accuracy	Kappa	AccuracySD	KappaSD
129	1	0.664	0.071	0.032	0.083
129	6	0.685	0.018	0.024	0.063
129	11	0.695	0.006	0.017	0.052
193	1	0.674	0.095	0.035	0.089
193	6	0.685	0.010	0.021	0.055
193	11	0.698	0.007	0.016	0.058
258	1	0.668	0.073	0.040	0.094
258	6	0.686	-0.004	0.019	0.050
258	11	0.697	0.002	0.018	0.053

	Measure_Value
Accuracy	0.710
Kappa	0.041
AccuracyLower	0.660
AccuracyUpper	0.756
AccuracyNull	0.715
AccuracyPValue	0.617
McnemarPValue	NaN

Model_Id	Model_Method	Accuracy_On_Training	**Accuracy_On_Validation**
SVM	Support Vector Machine	0.713	**0.715**
KNN	K-Nearest Neighbors	0.689	**0.699**
RF	Random Forest	0.714	**0.715**
LVQ	Learning Vector Quantization	0.698	**0.710**

The predicted ***Accuracy*** on the `validation_set` dataset for the ***Learning Vector Quantization*** is about ***0.71***.

Model 5 - Stochastic Gradient Boosting Machine (GBM)

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

Let's apply this model to our `Chocolate Bars Rating` dataset and verify the overall accuracy on the `validation_set` dataset.

	shrinkage	interaction.depth	n.minobsinnode	n.trees	Accuracy	Kappa	AccuracySD	KappaSD
1	0.1	1	10	50	0.711	0.043	0.016	0.062
4	0.1	2	10	50	0.701	0.038	0.019	0.071
7	0.1	3	10	50	0.695	0.048	0.025	0.074
2	0.1	1	10	100	0.705	0.061	0.021	0.070
5	0.1	2	10	100	0.700	0.065	0.019	0.058
8	0.1	3	10	100	0.685	0.044	0.024	0.072
3	0.1	1	10	150	0.704	0.070	0.021	0.065
6	0.1	2	10	150	0.692	0.062	0.022	0.068
9	0.1	3	10	150	0.681	0.054	0.025	0.072

	Measure_Value
Accuracy	0.704
Kappa	0.041
AccuracyLower	0.654
AccuracyUpper	0.750
AccuracyNull	0.715
AccuracyPValue	0.701
McnemarPValue	NaN

Model_Id	Model_Method	Accuracy_On_Training	**Accuracy_On_Validation**
SVM	Support Vector Machine	0.713	**0.715**
KNN	K-Nearest Neighbors	0.689	**0.699**
RF	Random Forest	0.714	**0.715**
LVQ	Learning Vector Quantization	0.698	**0.710**
GBM	Stochastic Gradient Boosting Machine	0.711	**0.704**

The predicted ***Accuracy*** on the `validation_set` dataset for the ***Stochastic Gradient Boosting Machine*** is about ***0.704***.

Model 6 - Classification And Regression Tree (CART)

Classification And Regression Trees or Decision Trees are commonly used in data mining with the objective of creating a model that predicts the value of a target (or dependent variable) based on the values of several input (or independent variables) either continuous or categorical.

The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any should be. The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions.

Let's apply this model to our `Chocolate Bars Rating` dataset and verify the overall accuracy on the `validation_set` dataset.

cp	Accuracy	Kappa	AccuracySD	KappaSD
0.006	0.704	0.031	0.022	0.072
0.008	0.705	0.021	0.021	0.069
0.014	0.711	0.000	0.006	0.013

	Measure_Value
Accuracy	0.715
Kappa	0.000
AccuracyLower	0.666
AccuracyUpper	0.761
AccuracyNull	0.715
AccuracyPValue	0.526
McnemarPValue	NaN

Model_Id	Model_Method	Accuracy_On_Training	**Accuracy_On_Validation**
SVM	Support Vector Machine	0.713	**0.715**
KNN	K-Nearest Neighbors	0.689	**0.699**
RF	Random Forest	0.714	**0.715**
LVQ	Learning Vector Quantization	0.698	**0.710**
GBM	Stochastic Gradient Boosting Machine	0.711	**0.704**
CART	Classification And Regression Tree	0.711	**0.715**

The predicted ***Accuracy*** on the `validation_set` dataset for the ***Classification And Regression Tree*** is about ***0.715***.

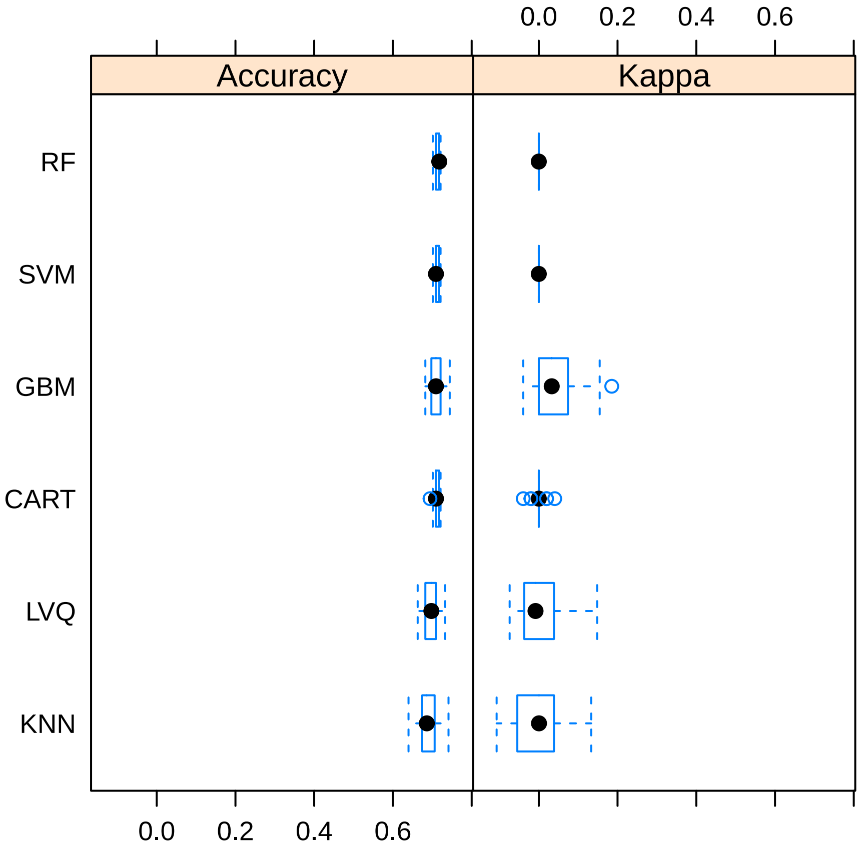
3. Results

Here is the summary of the Accuracy measures after building, training and validating different models on the `validation_set` dataset:

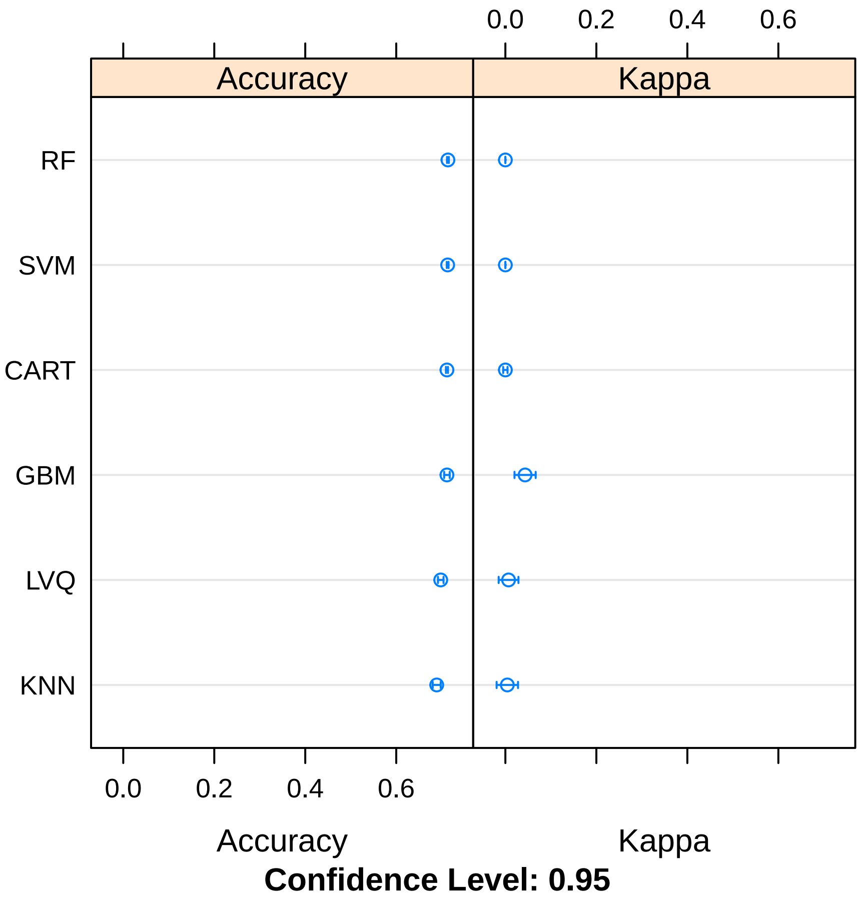
Model_Id	Model_Method	Accuracy_On_Training	**Accuracy_On_Validation**
RF	Random Forest	0.714	**0.715**
SVM	Support Vector Machine	0.713	**0.715**
CART	Classification And Regression Tree	0.711	**0.715**
LVQ	Learning Vector Quantization	0.698	**0.710**
GBM	Stochastic Gradient Boosting Machine	0.711	**0.704**
KNN	K-Nearest Neighbors	0.689	**0.699**

Based on the `Accuracy` measure only, we can observe that the model that predict our `Chocolate Bar Rating Class` with the best Accuracy of ***71.507*** % is the ***Random Forest***. But, all the classification models used in this project give us slightly similar results.

However, we need to compare all the models based on the Accuracy, Kappa value and the confidence intervall (95%).



Based on this Box-plot, the best model to predict the `Chocolate Bar Rating Class` is ***Random Forest*** followed by ***Support Vector Machine*** and ***Stochastic Gradient Boosting Machine***.



Based on this Dot-plot and the Confidence interval at 95%, the best model to predict the `Chocolate Bar Rating Class` is ***Random Forest*** followed by ***Support Vector Machine*** and ***Classification And Regression Tree***.

This confirms that the classification models used in this project give us slightly similar results. So, we will be validation the ***Random Forest*** model as the one that predict most accuratly the `Chocolate Bar Rating Class`.

4. Conclusion

In this project, we have used an iterative applied machine learning approach to implement a Chocolate Bar Recommendation System to predict a Chocolate Bar Rating Class.

We have explored how data preprocessing and data visualization can impact the complex machine learning model building phase. We learned about different data pre-processing techniques and tried out a few on the chocolate bar dataset.

We also explored a few data visualization tools and discussed how visualization can impact modeling itself. Each visualization tool has its own significance in story telling, and it's important to understand which ones can be used with particular types of data.

Moreover, we have build, train and validate multiple Classification and Regression algorithms to predict the Quality of the Chocolate Bars. We could also build and train other models and algorithms that could improve the Accuracy.

The most challenging impediment that we might encounter when implementing those additional models and going further in the overall optimizations are both the dataset size (not large enought) and the data quality within the dataset.

Finally, based on our analysis, we can conclude that for all chocolate bars made with one of the five bean types we have included in our data, the ratings for these chocolate bars decrease as the chocolate becomes more bitter, or as the cocoa percent in the chocolate bar increases. From this, we can take away that people prefer sweeter over bitter chocolate for bars made from these five bean types.