

# SHETH L.U.J. AND SIR M.V. COLLEGE

## SUBJECT: Data Analysis with R

Aim: Identifying and handling duplicates using distinct() (R studio)

```
Attaching package: 'janitor'

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

> # Load the cacao dataset
> cacao_df <- read.csv("flavors_of_cacao.csv", stringsAsFactors = FALSE)
> # Clean column names to handle spaces and special characters
> cacao_df <- cacao_df %>%
+   clean_names()
> print("--- 1. Original dataset ---")
[1] "--- 1. Original dataset ---"
> print(head(cacao_df, 10)) # show first 10 rows
  company_maker_if_known specific_bean_origin_or_bar_name ref review_date cocoa_percent company_location rating bean_type
1 A. Morin Agua Grande 1876 2016 63% France 3.75
2 A. Morin kpime 1676 2015 70% France 2.75
3 A. Morin Atsane 1676 2015 70% France 3.00
4 A. Morin Akata 1680 2015 70% France 3.50
5 A. Morin Quilla 1704 2015 70% France 3.50
6 A. Morin Carenero 1315 2014 70% France 2.75 criollo
7 A. Morin Cuba 1315 2014 70% France 3.50
8 A. Morin Sur del Lago 1315 2014 70% France 3.75 criollo
9 A. Morin Puerto Cabello 1319 2014 70% France 3.75 criollo
10 A. Morin Pablino 1319 2014 70% France 4.00 criollo

broad_bean_origin
1 Sao Tome
2 Togo
3 Togo
4 Togo
5 Peru
6 Venezuela
7 Cuba
8 Venezuela
9 Venezuela
10 Peru

> # Identify rows that are EXACT duplicates across all columns
> duplicates_report <- cacao_df %>%
+   group_by(across(everything())) %>% # group by all columns
+   count() %>% # count occurrences
+   filter(n > 1) # keep only rows that appear more than once
```

```
Source
Console Terminal Background Jobs
R452: ~/
[1] "--- 2. Identification Report (Exact Duplicate Rows) ---"
> print(duplicates_report)
# A tibble: 0 x 10
# Groups:   company_maker_if_known, specific_bean_origin_or_bar_name, ref, review_date, cocoa_percent, company_location, rating, bean_type,
#   broad_bean_origin [0]
# i 10 variables: company_maker_if_known <chr>, specific_bean_origin_or_bar_name <chr>, ref <int>, review_date <int>, cocoa_percent <chr>,
#   company_location <chr>, rating <dbl>, bean_type <chr>, broad_bean_origin <chr>, n <int>
> # Remove rows where EVERY column is identical
> clean_exact <- cacao_df %>%
+   distinct() # removes exact duplicates
> print("--- 3. Removed Exact Duplicates (distinct) ---")
[1] "--- 3. Removed Exact Duplicates (distinct) ---"
> print(head(clean_exact, 10))
  company_maker_if_known specific_bean_origin_or_bar_name ref review_date cocoa_percent company_location rating bean_type
1 A. Morin Agua Grande 1876 2016 63% France 3.75
2 A. Morin kpime 1676 2015 70% France 2.75
3 A. Morin Atsane 1676 2015 70% France 3.00
4 A. Morin Akata 1680 2015 70% France 3.50
5 A. Morin Quilla 1704 2015 70% France 3.50
6 A. Morin Carenero 1315 2014 70% France 2.75 criollo
7 A. Morin Cuba 1315 2014 70% France 3.50
8 A. Morin Sur del Lago 1315 2014 70% France 3.75 criollo
9 A. Morin Puerto Cabello 1319 2014 70% France 3.75 criollo
10 A. Morin Pablino 1319 2014 70% France 4.00 criollo

broad_bean_origin
1 Sao Tome
2 Togo
3 Togo
4 Togo
5 Peru
6 Venezuela
7 Cuba
8 Venezuela
9 Venezuela
10 Peru

> unique_companies <- cacao_df %>%
+   distinct(company_maker_if_known, .keep_all = TRUE)
> print("--- 4. Unique Companies Only (Partial Duplicates Removed) ---")
[1] "--- 4. Unique Companies Only (Partial Duplicates Removed) ---"
> print(head(unique_companies, 10))
  company_maker_if_known specific_bean_origin_or_bar_name ref review_date cocoa_percent company_location rating
1 A. Morin Agua Grande 1876 2016 63% France 3.75
2 Acalli chulucanas, El Platanal 1462 2015 70% U.S.A. 3.75
```

# SHETH L.U.J. AND SIR M.V. COLLEGE

## SUBJECT: Data Analysis with R

The top screenshot shows the R console output for the following code:

```
> unique_companies <- cacao_df %>%  
+ distinct(company_maker_if_known, .keep_all = TRUE)  
> print("---- 4. Unique Companies Only (Partial Duplicates Removed) ----")  
[1] "---- 4. Unique Companies Only (Partial Duplicates Removed) ----"  
> print(head(unique_companies, 10))  
company_maker_if_known specific_bean_origin_or_bar_name ref review_date cocoa_percent company_location rating  
1 A. Morin Agua Grande 1876 2016 63% France 3.75  
2 Acañi Chulucanas, El Platanal 1462 2015 70% U.S.A. 3.75  
3 Adí Vanua Levu 705 2011 60% Fiji 2.75  
4 Aequare (Gándula) Los Ríos, Quevedo, Arriba 370 2009 55% Ecuador 2.75  
5 Ah Cacao Tabasco 316 2009 70% Mexico 3.00  
6 Akesson's (Pralus) Bali (west), Sukrama Family, Melaya area 636 2011 75% Switzerland 3.75  
7 Alain Ducasse Trinite 1215 2014 65% France 2.75  
8 Alexandre winak coop, Napo 1944 2017 70% Netherlands 3.50  
9 Altus aka Cao Artisan Monotombo 1728 2016 60% U.S.A. 2.75  
10 Amano Morobe 725 2011 70% U.S.A. 4.00  
bean_type broad_bean_origin  
1 Sao Tome  
2 Peru  
3 Trinitario Fiji  
4 Forastero (Arriba) Ecuador  
5 Criollo Mexico  
6 Trinitario Indonesia  
7 Trinitario Trinidad  
8 Forastero (Nacional) Ecuador  
9 Nicaragua  
10 Papua New Guinea  
> # Scenario: Keep unique chocolate bars based on company + Specific Bean Origin  
> unique_bars <- cacao_df %>%  
+ distinct(company_maker_if_known, specific_bean_origin_or_bar_name, .keep_all = TRUE)  
> print("---- 5. Unique Bars (Company + Specific Bean Origin) ----")  
[1] "---- 5. Unique Bars (Company + Specific Bean Origin) ----"  
> print(head(unique_bars, 10))  
company_maker_if_known specific_bean_origin_or_bar_name ref review_date cocoa_percent company_location rating bean_type  
1 A. Morin Agua Grande 1876 2016 63% France 3.75  
2 A. Morin Kpime 1676 2015 70% France 2.75  
3 A. Morin Atsane 1676 2015 70% France 3.00  
4 A. Morin Akata 1680 2015 70% France 3.50  
5 A. Morin Quilla 1704 2015 70% France 3.50  
6 A. Morin Carenero 1315 2014 70% France 2.75  
7 A. Morin Cuba 1315 2014 70% France 3.50  
8 A. Morin Sur del Lago 1315 2014 70% France 3.50  
9 A. Morin Puerto Cabello 1319 2014 70% France 3.75  
10 A. Morin Pablino 1319 2014 70% France 4.00  
broad_bean_origin  
1 Sao Tome  
2 Togo  
3 Togo  
4 Togo  
5 Peru  
6 Venezuela  
7 Cuba  
8 Venezuela  
9 Venezuela  
10 Peru
```

The middle screenshot shows the R console output for the following code:

```
> distinct(company_maker_if_known, specific_bean_origin_or_bar_name, .keep_all = TRUE)  
> print("---- 5. Unique Bars (Company + Specific Bean Origin) ----")  
[1] "---- 5. Unique Bars (Company + Specific Bean Origin) ----"  
> print(head(unique_bars, 10))  
company_maker_if_known specific_bean_origin_or_bar_name ref review_date cocoa_percent company_location rating bean_type  
1 A. Morin Agua Grande 1876 2016 63% France 3.75  
2 A. Morin Kpime 1676 2015 70% France 2.75  
3 A. Morin Atsane 1676 2015 70% France 3.00  
4 A. Morin Akata 1680 2015 70% France 3.50  
5 A. Morin Quilla 1704 2015 70% France 3.50  
6 A. Morin Carenero 1315 2014 70% France 2.75  
7 A. Morin Cuba 1315 2014 70% France 3.50  
8 A. Morin Sur del Lago 1315 2014 70% France 3.50  
9 A. Morin Puerto Cabello 1319 2014 70% France 3.75  
10 A. Morin Pablino 1319 2014 70% France 4.00  
broad_bean_origin  
1 Sao Tome  
2 Togo  
3 Togo  
4 Togo  
5 Peru  
6 Venezuela  
7 Cuba  
8 Venezuela  
9 Venezuela  
10 Peru
```

The bottom screenshot shows the RStudio interface with the 'unique\_companies' dataset loaded and the 'unique\_bars' dataset loaded. The Environment pane shows the following variables:

- cacao\_df: 1795 obs. of 9 variables
- clean\_exact: 1795 obs. of 9 variables
- duplicates: 0 obs. of 10 variables
- unique\_bars: 1677 obs. of 9 variables
- unique\_companies: 416 obs. of 9 variables

# SHETH L.U.J. AND SIR M.V. COLLEGE

## SUBJECT: Data Analysis with R

The screenshot displays the R Studio interface with the following components:

- Main Window:** A data frame with 21 rows and 9 columns. The columns are: company\_maker\_if\_known, specific\_bean\_origin\_or\_bar\_name, ref, review\_date, cocoa\_percent, company\_location, rating, bean\_type, and broad\_bean\_origin. The data shows chocolate products from various companies like A. Morin, Kplme, and Akata, with details on their origin, review date, cocoa percentage, and rating.
- Environment Pane (Right):** Lists data frames: cacao\_df (1795 obs. of 9 variables), clean\_exact (1795 obs. of 9 variables), duplicates... (0 obs. of 10 variables), unique\_bars (1677 obs. of 9 variables), and unique\_com... (416 obs. of 9 variables).
- Files Pane (Right):** Shows a file explorer with folders like My Pictures, My Videos, My Web Sites, and files like NetBeansProjects, Power BI Desktop, scala for DS, Sound Recordings, T050\_AnjaliT, vgsales.csv, Virtual Machines, Visual Studio 18, Visual Studio 2022, WindowsPowerShell, Student.csv, car\_price\_prediction.csv, disease\_diagnosis.csv, and flavors\_of\_cacao.csv.
- Console (Bottom):** Shows R commands and output:

```
R - R 4.5.2 - ~/...  
9 venezuela  
10 Peru  
> view(unique_companies)  
> view(cacao_df)  
> view(clean_exact)  
> |
```

# SHETH L.U.J. AND SIR M.V. COLLEGE

## SUBJECT: Data Analysis with R

The screenshot displays the R Studio environment. The main window shows a data table with 21 rows and 9 columns. The columns are: company\_maker\_if\_known, specific\_bean\_origin\_or\_bar\_name, ref, review\_date, cocoa\_percent, company\_location, rating, bean\_type, and broad\_bean\_origin. The console window at the bottom shows the following commands and output:

```
> view(unique_companies)
> view(cacao_df)
> view(clean_exact)
> view(unique_bars)
> view(unique_companies)
> |
```

The file explorer on the right shows the following files and folders:

- My Pictures
- My Videos
- My Web Sites
- NetBeansProjects
- Power BI Desktop
- scale for DS
- Sound Recordings
- T050\_AnjaliT
- vgsales.csv (1.3 MB)
- Virtual Machines
- Visual Studio 18
- Visual Studio 2022
- WindowsPowerShell
- Student.csv (163 KB)
- car\_price\_prediction.csv (156.3 KB)
- disease\_diagnosis.csv (189.1 KB)
- flavors\_of\_cacao.csv (126.5 KB)