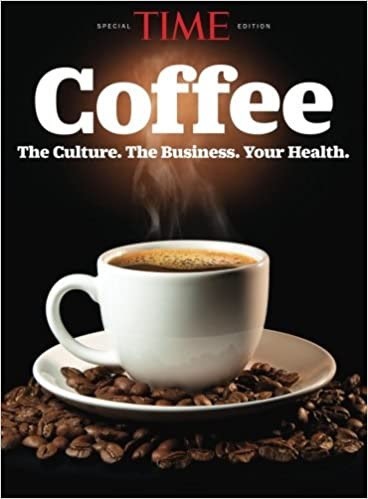
# Introduction



Around four years ago I was given a copy of Time Magazine’s specialty issue on Coffee together with a French press as a gift. At the time, I was satisfied with a regular instant cup of joe and did not know much about the vastness and culture of the industry. However, it was thanks to these gifts that I was able to learn a lot about coffee, such as the two major species of beans (Arabica and Robusta),the tasting process done by connoisseurs to rank various coffees(called “cupping”), about the altitude, climate and countries various coffees grow around the world. If you read this specialty issue by Time, you probably not only got a more expensive interest piqued (if you haven’t already), but also probably learned enough to hold your own with the the best of the coffee snobs out there.

(PSA- this blog is not sponsored by Time Magazine, but I won’t say no if I got an offer!)

In this blog post we’re going to examine the coffee\_ratings dataset released back in the beginning of July 2020 in the Tidy Tuesday Project by R4DS. I initially started analyzing this dataset seeking to answer a lot of questions. But, because there is so much to discover and analyze from this relatively small dataset, I thought it is best to try to focus my question on a very simple one:

Where in the world can I find the best coffee beans?

While this question seems simple enough. There is a lot to uncover to answer this question.

# Our Data (Some Exploratory Data Analysis)

## Loading our data

I am loading the data with the tidytuesdayR package, if you want you can load the raw data with the readr package’s read\_csv() function as well.

# A Quick Glimpse

library(tidyverse)

coffee\_ratings<-tuesdata$coffee\_ratings glimpse(coffee\_ratings)

## Rows: 1,339

## Columns: 43

## $ total\_cup\_points 90.58, 89.92, 89.75, 89.00, 88.83, 88.83,

88.75, 88.67, 88.42, 88.25, 88.08, 87.92, 87.92, 87.92, 87.8...

## $ species "Arabica", "Arabica", "Arabica", "Arabica", "Arabica", "Arabica", "Arabica", "Arabica", "Arabica", "Ar...

## $ owner "metad plc", "metad plc", "grounds for health admin", "yidnekachew dabessa", "metad plc", "ji-ae ahn",... ## $ country\_of\_origin "Ethiopia", "Ethiopia", "Guatemala", "Ethiopia", "Ethiopia", "Brazil", "Peru", "Ethiopia", "Ethiopia",... ## $ farm\_name "metad plc", "metad plc", "san marcos barrancas \"san cristobal cuch", "yidnekachew dabessa coffee pla...

## $ lot\_number NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "YNC-06114", NA, NA, NA, NA, N...

## $ mill "metad plc", "metad plc", NA, "wolensu", "metad plc", NA, "hvc", "c.p.w.e", "c.p.w.e", "tulla coffee f...

## $ ico\_number "2014/2015", "2014/2015", NA, NA, "2014/2015", NA, NA, "010/0338", "010/0338", "2014/15", NA, "unknown...

## $ company "metad agricultural developmet plc", "metad agricultural developmet plc", NA, "yidnekachew debessa cof...

## $ altitude "1950-2200", "1950-2200", "1600 - 1800 m",

"1800-2200", "1950-2200", NA, NA, "1570-1700", "1570-1700",...

## $ region "guji-hambela", "guji-hambela", NA, "oromia", "guji-hambela", NA, NA, "oromia", "oromiya", "snnp/kaffa... ## $ producer "METAD PLC", "METAD PLC", NA, "Yidnekachew Dabessa Coffee Plantation", "METAD PLC", NA, "HVC", "Bazen ...

## $ number\_of\_bags 300, 300, 5, 320, 300, 100, 100, 300, 300,

50, 300, 10, 10, 1, 300, 10, 1, 150, 3, 250, 10, 250, 14, 1...

## $ bag\_weight "60 kg", "60 kg", "1", "60 kg", "60 kg",

"30 kg", "69 kg", "60 kg", "60 kg", "60 kg", "60 kg", "1 kg",...

## $ in\_country\_partner "METAD Agricultural Development plc", "METAD Agricultural Development plc", "Specialty Coffee Associat...

## $ harvest\_year "2014", "2014", NA, "2014", "2014", "2013",

"2012", "March 2010", "March 2010", "2014", "2014", "2014"...

## $ grading\_date "April 4th, 2015", "April 4th, 2015", "May 31st, 2010", "March 26th, 2015", "April 4th, 2015", "Septem...

## $ owner\_1 "metad plc", "metad plc", "Grounds for

Health Admin", "Yidnekachew Dabessa", "metad plc", "Ji-Ae Ahn",...

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ## $ variety  "Other", NA, NA, "Other" | | | | NA, "Other", "Bourbon", NA, "Other", NA,  , NA, "Other", "Other", NA, NA, "Othe... | | | |
| ## $ processing\_method | | | | "Washed / Wet", "Washed / Wet", NA, | | | |
| "Natural / Dry", "Washed | | | | / Wet", "Natural / Dry", "Washed / Wet", ... | | | |
| ## $ aroma | | | | 8.67, 8.75, 8.42, 8.17, 8.25, 8.58, 8.42, | | | |
| 8.25, 8.67, 8.08, 8.17, | | | | 8.25, 8.08, 8.33, 8.25, 8.00, 8.33, ... | | | |
| ## $ flavor | | | | 8.83, 8.67, 8.50, 8.58, 8.50, 8.42, 8.50, | | | |
| 8.33, 8.67, 8.58, 8.67, | | | | 8.42, 8.67, 8.42, 8.33, 8.50, 8.25, ... | | | |
| ## $ aftertaste | | | | 8.67, 8.50, 8.42, 8.42, 8.25, 8.42, 8.33, | | | |
| 8.50, 8.58, 8.50, 8.25, | | | | 8.17, 8.33, 8.08, 8.50, 8.58, 7.83, ... | | | |
| ## $ acidity | | | | 8.75, 8.58, 8.42, 8.42, 8.50, 8.50, 8.50, | | | |
| 8.42, 8.42, 8.50, 8.50, | | | | 8.33, 8.42, 8.25, 8.25, 8.17, 7.75, ... | | | |
| ## $ body | | | | 8.50, 8.42, 8.33, 8.50, 8.42, 8.25, 8.25, | | | |
| 8.33, 8.33, 7.67, 7.75, | | | | 8.08, 8.00, 8.25, 8.58, 8.17, 8.50, ... | | | |
| ## $ balance | | | | 8.42, 8.42, 8.42, 8.25, 8.33, 8.33, 8.25, | | | |
| 8.50, 8.42, 8.42, 8.17, | | | | 8.17, 8.08, 8.00, 8.75, 8.00, 8.42, ... | | | |
| ## $ uniformity | | | | 10.00, 10.00, 10.00, 10.00, 10.00, 10.00, | | | |
| 10.00, 10.00, 9.33, | | | 10.00, 10.00, 10.00, 10.00, | | 10.00, 9.33,... | | |
| ## $ clean\_cup |  | | 10, 10, 10, 10, 10, | | 10, 10, | 10, 10, 10, 10, | |
| 10, 10, 10, 10, | 10, | | 10, 10, 10, 10, 10, 10, 10, | | 10, 10, | 10... | |
| ## $ sweetness | 10.00, 10.00, 10.00, 10.00, | | | | | 10.00, 10.00, | |
| 10.00, 9.33, 9.33, | | | 10.00, 10.00, 10.00, 10.00, 10.00, 9.33, ... | | | |  |
| ## $ cupper\_points | | | 8.75, 8.58, 9.25, 8.67, 8.58, 8.33, | | | | 8.50, |
| 9.00, 8.67, 8.50, | | 8.58, | | 8.50, 8.33, 8.58, 8.50, 8.17, 8.33, ... | | |  |
| ## $ moisture | |  | | 0.12, 0.12, 0.00, 0.11, 0.12, 0.11, | | | 0.11, |
| 0.03, 0.03, 0.10, | | 0.10, | | 0.00, 0.00, 0.00, 0.05, 0.00, 0.03, ... | | |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ## | $ category\_one\_defects | | | | | 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, | 0, 0, |
| 0, | 0, 0, 0, 0, | 0, | 0, | 0, | 0, | 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ... |  |
| ## | $ quakers |  |  |  |  | 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, | 0, 0, |
| 0, | 0, 0, 0, 0, | 0, | 0, | 0, | 0, | 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ... |  |

## $ color "Green", "Green", NA, "Green", "Green", "Bluish-Green", "Bluish-Green", NA, NA, "Green", NA, NA, NA, N...

## $ category\_two\_defects 0, 1, 0, 2, 2, 1, 0, 0, 0, 4, 1, 0, 0, 2,

2, 0, 0, 2, 0, 8, 0, 2, 0, 0, 1, 2, 2, 1, 3, 0, 2, 1, 2, 0, ...

## $ expiration "April 3rd, 2016", "April 3rd, 2016", "May 31st, 2011", "March 25th, 2016", "April 3rd, 2016", "Septem...

## $ certification\_body "METAD Agricultural Development plc", "METAD Agricultural Development plc", "Specialty Coffee Associat...

## $ certification\_address "309fcf77415a3661ae83e027f7e5f05dad786e44", "309fcf77415a3661ae83e027f7e5f05dad786e44", "36d0d00a37243...

## $ certification\_contact "19fef5a731de2db57d16da10287413f5f99bc2dd", "19fef5a731de2db57d16da10287413f5f99bc2dd", "0878a7d4b9d35...

## $ unit\_of\_measurement "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "m", "ft", "m", "m", "m", "m", "m", "m", "...

## $ altitude\_low\_meters 1950.0, 1950.0, 1600.0, 1800.0, 1950.0, NA,

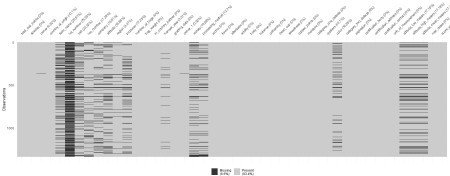
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NA, 1570.0, 1570.0, 1795.0, | 1855.0, | 1872.0, | 1943.0, | 609.6,... |  |
| ## $ altitude\_high\_meters | 2200.0, | 2200.0, | 1800.0, | 2200.0, 2200.0, | NA, |
| NA, 1700.0, 1700.0, 1850.0, | 1955.0, | 1872.0, | 1943.0, | 609.6,... |  |
| ## $ altitude\_mean\_meters | 2075.0, | 2075.0, | 1700.0, | 2000.0, 2075.0, | NA, |
| NA, 1635.0, 1635.0, 1822.5, | 1905.0, | 1872.0, | 1943.0, | 609.6,... |  |

A quick glimpse of our data (no pun intended) is enough to indicate that our dataset is far from clean.

It also looks like there is missing data everywhere. Lets see how much.

## Missing Data

library(naniar) vis\_miss(coffee\_ratings)



Thankfully, it’s not as bad as I thought it was going to be. For the nature of my question I am only going to using the total\_cupper\_points, country\_of\_origin, grading\_date and species variables which all have little to no missing data (I thought this would be more of an issue, but looking back at it I’m thankful it isn’t for this case.)

# Quantites of Coffee per Country

As stated in the description of our dataset

“These data were collected from the Coffee Quality Institute’s review pages in January 2018.”

(I am not sure how grammatical that phrase is but ok.)

To better understand our data, lets look at the frequencies of our data in terms of countries listed in our data set. Because there is only one instance of missing data, we will remove it from our plots for aesthetic reasons.

library(ggthemes)

# Need to make a new transformed dataset for this visualization

(

country\_table<-coffee\_ratings %>%

count(country\_of\_origin = factor(country\_of\_origin)) %>% mutate(pct = prop.table(n)) %>%

arrange(-pct) %>% tibble()

)

## # A tibble: 37 x 3

## country\_of\_origin n pct ##

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## | 1 | Mexico |  | 236 | | 0.176 |
| ## | 2 | Colombia |  | 183 | | 0.137 |
| ## | 3 | Guatemala |  | 181 | | 0.135 |
| ## | 4 | Brazil |  | 132 | | 0.0986 |
| ## | 5 | Taiwan |  | 75 | | 0.0560 |
| ## | 6 | United States | (Hawaii) | 73 | | 0.0545 |
| ## | 7 | Honduras | |  | 53 | 0.0396 |
| ## | 8 | Costa Rica | |  | 51 | 0.0381 |
| ## | 9 | Ethiopia | |  | 44 | 0.0329 |
| ## | 10 | Tanzania, United Republic | | Of | 40 | 0.0299 |

## # ... with 27 more rows

# Together with my knowledge of ggplot and google, these visualizations became possible

ggplot(

country\_table %>% filter(country\_of\_origin != "NA"), mapping = aes(

x = reorder(country\_of\_origin, n), y = pct,

group = 1,

label = scales::percent(pct)

)

) +

theme\_fivethirtyeight() + geom\_bar(stat = "identity",

fill = "#634832") +

geom\_text(position = position\_dodge(width = 0.9), # move to center of bars

hjust = -0.05,

#Have Text just above bars size = 2.5) +

labs(x = "Country of Origin",

y = "Proportion of Dataset") + theme(axis.text.x = element\_text(

angle = 90,

vjust = 0.5,

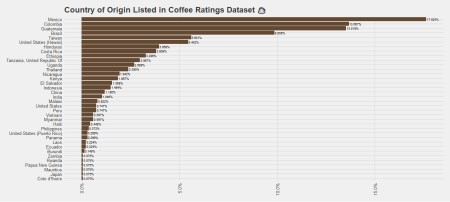
hjust = 1

)) +

ggtitle("Country of Origin Listed in Coffee Ratings Dataset " ) + #

This Emoji messes up this line in R markdown but hey, it scale\_y\_continuous(labels = scales::percent) +

# looks good. coord\_flip()



From a brief look at our table and bar chart we see that **over 54% of our dataset consists of coffees from Mexico, Columbia, Guatemala and Brazil**. But this only tells us part of the story, what species of coffees do we have in our dataset from each country?

Before looking at that lets look at the overall Arabica/Robusta proportion in our dataset:

# Need to make a new transformed dataset for this visualization

species\_table<-coffee\_ratings %>% count(species = factor(species)) %>% mutate(pct = prop.table(n)) %>% tibble()

ggplot(species\_table,mapping=aes(x=species,y=pct,group=1, label=scales::percent(pct)))+

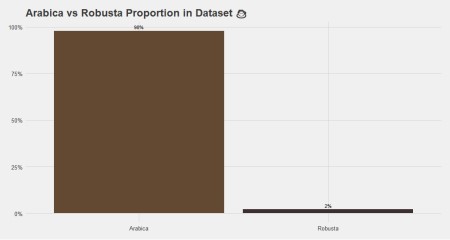
theme\_fivethirtyeight()+ geom\_bar(stat="identity",

fill=c("#634832","#3b2f2f"))+

geom\_text(position = position\_dodge(width=0.9), # move to center of bars

vjust=-0.5, #Have Text just above bars size = 3)+

scale\_y\_continuous(labels = scales::percent)+ ggtitle("Arabica vs Robusta Proportion in Dataset ")



Wow! only 2% of Coffee in our dataset is from Robusta beans! But if you think about this in context, this shouldn’t be too much of a suprise. Robusta coffee is primarily used in instant coffee,espresso and filler for coffee blends. The reason why Robusta coffee beans are not

graded proportionately as Arabica beans are is due to the fact that the quality of these bitter, earthy beans are usually not as desirable to coffee drinkers as their smoother, richer Arabica counterparts.

With that in mind, lets see how the breakdown proportionally per country:

# Need to make a new transformed datasets for this visualization

(

arabica\_countries<-coffee\_ratings %>% filter(species =="Arabica") %>%

count(species=factor(species), country=country\_of\_origin) %>%

mutate(pct = prop.table(n)) %>% arrange(-n) %>%

tibble()

)

## # A tibble: 37 x 4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ##  ## | species | country |  | n | | pct |
| ## 1 | Arabica | Mexico |  | 236 | | 0.180 |
| ## 2 | Arabica | Colombia |  | 183 | | 0.140 |
| ## 3 | Arabica | Guatemala |  | 181 | | 0.138 |
| ## 4 | Arabica | Brazil |  | 132 | | 0.101 |
| ## 5 | Arabica | Taiwan |  | 75 | | 0.0572 |
| ## 6 | Arabica | United States | (Hawaii) | 73 | | 0.0557 |
| ## 7 | Arabica | Honduras | |  | 53 | 0.0404 |
| ## 8 | Arabica | Costa Rica | |  | 51 | 0.0389 |
| ## 9 | Arabica | Ethiopia | |  | 44 | 0.0336 |
| ## 10 | Arabica | Tanzania, United Republic | | Of | 40 | 0.0305 |

## # ... with 27 more rows

ggplot(arabica\_countries %>% filter(country!="NA"), mapping=aes(x=reorder(country,n),y=pct,group=1,label=scales:

:percent(pct))) + theme\_fivethirtyeight()+ geom\_bar(stat="identity",

fill="#634832")+

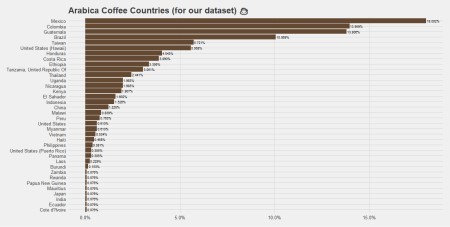
geom\_text(position = position\_dodge(width = 0.9), # move to center of bars

hjust = -0.05,

#Have Text just above bars size = 2.5) +

ggtitle("Arabica Coffee Countries (for our dataset) ") + scale\_y\_continuous(labels = scales::percent) +

coord\_flip()



(

robusta\_countries<-coffee\_ratings %>% filter(species =="Robusta") %>% count(species = factor(species),

country=country\_of\_origin) %>% mutate(pct = prop.table(n)) %>% arrange(-n) %>%

tibble()

)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## ##  ## | # | A tibble: 5 x 4 species country | n | pct |
| ## | 1 | Robusta India | 13 | 0.464 |
| ## | 2 | Robusta Uganda | 10 | 0.357 |
| ## | 3 | Robusta Ecuador | 2 | 0.0714 |
| ## | 4 | Robusta United States | 2 | 0.0714 |
| ## | 5 | Robusta Vietnam | 1 | 0.0357 |

ggplot(robusta\_countries %>% filter(country!="NA"), mapping=aes(x=reorder(country,n),y=pct,group=1,label=scales:

:percent(pct))) + theme\_fivethirtyeight()+ geom\_bar(stat="identity",

fill="#3b2f2f")+

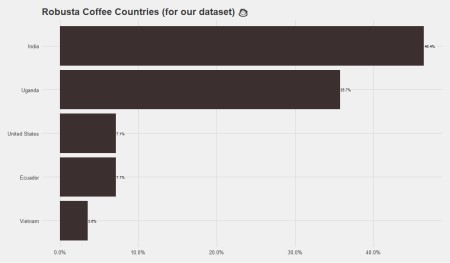
geom\_text(position = position\_dodge(width = 0.9), # move to center of bars

hjust = -0.05,

#Have Text just above bars size = 2.5) +

ggtitle("Robusta Coffee Countries (for our dataset) ") + scale\_y\_continuous(labels = scales::percent) +

coord\_flip()



The Robusta coffees that we have in this dataset are mostly from India and Uganda, with a few coffees from the Ecuador, the United States and Vietnam. With that being known, Lets look at the Arabica/Robusta ratio for countries that we have Robusta Data on.

coffee\_ratings %>%

filter(country\_of\_origin %in% c("India","Uganda","Ecuador","United States","Vietnam")) %>%

count(country\_of\_origin,species) %>% group\_by(country\_of\_origin)

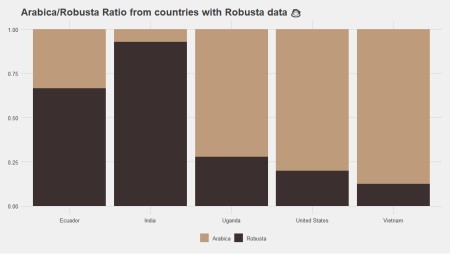
|  |  |  |  |
| --- | --- | --- | --- |
| ## ## ##  ## | # # | A tibble: 10 x 3  Groups: country\_of\_origin country\_of\_origin species | [5]  n |
| ## | 1 | Ecuador Arabica | 1 |
| ## | 2 | Ecuador Robusta | 2 |
| ## | 3 | India Arabica | 1 |
| ## | 4 | India Robusta | 13 |
| ## | 5 | Uganda Arabica | 26 |
| ## | 6 | Uganda Robusta | 10 |
| ## | 7 | United States Arabica | 8 |
| ## | 8 | United States Robusta | 2 |
| ## | 9 | Vietnam Arabica | 7 |
| ## | 10 | Vietnam Robusta | 1 |

ggplot(coffee\_ratings %>% filter(country\_of\_origin %in% c("India","Uganda","Ecuador","United States","Vietnam")),

mapping=aes(x=country\_of\_origin,fill=species))+ theme\_fivethirtyeight()+

geom\_bar(position="fill")+ scale\_fill\_manual(values=c("#BE9B7B", "#3b2f2f"))+ theme(legend.title = element\_blank())+

ggtitle("Arabica/Robusta Ratio from countries with Robusta data ")



Now that we have better understanding of where our coffees come from, we can get into trying to answer the question of **where** the best coffee beans are in the world.

Well, it depends.

# What type? What year?

It would be nice to just pick out the highest rated coffee and be done with it, but that wouldn’t tell us anything (or really motivate a blog post). We need to consider is when was a given coffee graded. That can tell us the performance of a given country’s over time. Additionally, we need to consider the species of bean- where is the best ranked Arabica coffee from? Where is the best Robusta coffee from?

Before we can answer this question, we need to clean the grading\_date and convert them into the date data from. Thankfully, the lubridate package will help us with doing this relatively easy. After that we will formulate our data set with the dplyr package to get the data in the form we need for our visualization.

|

library(lubridate)

# Getting the year data

coffee\_ratings$new\_dates<-coffee\_ratings$grading\_date %>% mdy() coffee\_ratings$score\_year<- coffee\_ratings$new\_dates %>% year()

# Dataset for visualizations

(

top\_annual\_score<- coffee\_ratings %>% group\_by(species,

score\_year, country\_of\_origin) %>%

summarise(max\_points = max(total\_cup\_points)) %>% filter(max\_points == max(max\_points)) %>% arrange(-max\_points)

)

## # A tibble: 15 x 4

## # Groups: species, score\_year [15]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ##  ## | species | score\_year | country\_of\_origin | max\_points |
| ## 1 | Arabica | 2015 | Ethiopia | 90.6 |
| ## 2 | Arabica | 2010 | Guatemala | 89.8 |
| ## 3 | Arabica | 2013 | Brazil | 88.8 |
| ## 4 | Arabica | 2012 | Peru | 88.8 |
| ## 5 | Arabica | 2016 | China | 87.2 |
| ## 6 | Arabica | 2014 | Costa Rica | 87.2 |
| ## 7 | Arabica | 2011 | Brazil | 86.9 |
| ## 8 | Arabica | 2017 | Honduras | 86.7 |
| ## 9 | Arabica | 2018 | Kenya | 84.6 |
| ## 10 | Robusta | 2014 | Uganda | 83.8 |
| ## 11 | Robusta | 2017 | India | 83.5 |
| ## 12 | Robusta | 2015 | India | 83.2 |
| ## 13 | Robusta | 2012 | India | 82.8 |
| ## 14 | Robusta | 2016 | India | 82.5 |
| ## 15 | Robusta | 2013 | India | 81.2 |

ggplot(top\_annual\_score, mapping=aes(x=score\_year,

y=max\_points, label=paste0(score\_year,"\n",country\_of\_origin,"\n",

max\_points),

color=country\_of\_origin))+

theme\_fivethirtyeight()+

geom\_text(position = position\_dodge(width = 0.9), # move to center of bars

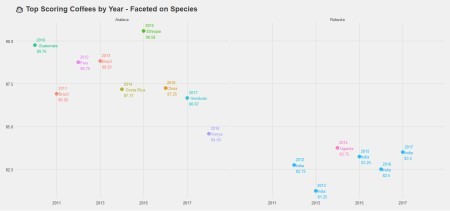
hjust =-0.2,

#Have Text just above bars size =3.5) +

geom\_point(size=4,

alpha=0.8)+ theme(legend.position = "none")+ facet\_wrap(~species)+

ggtitle(" Top Scoring Coffees by Year - Faceted on Species ")



From our visualization and table we see for Arabica beans, the top coffee varied from country to country for a given year. However for Robusta, India seemed to have dominated with consistent wins from 2012 – 2017 with an exception of Uganda beating them in 2014.

Overall, for our given timespan in our dataset, for Arabica beans (as well as our entire dataset) Ethiopia scored the highest with a score of 90.58 and for Robusta Beans Uganda had the highest score of 83.75.

The overall summary for of scores for Arabica and Robusta beans accross the years is plotted in the below visualization with boxplots.

(arabica\_robusta\_average\_score<- coffee\_ratings %>%

group\_by(species) %>%

summarise(average\_score = mean(total\_cup\_points), lower\_ci = mean(total\_cup\_points) -

1.96\*sqrt(var(total\_cup\_points)/length(total\_cup\_points)), upper\_ci = mean(total\_cup\_points) +

1.96\*sqrt(var(total\_cup\_points)/length(total\_cup\_points)))

)

## # A tibble: 2 x 4

## species average\_score lower\_ci upper\_ci ##

## 1 Arabica 82.1 81.9 82.3

## 2 Robusta 80.9 80.0 81.8

ggplot(coffee\_ratings,mapping=aes(x=score\_year,y=total\_cup\_

points,group=score\_year))+ theme\_fivethirtyeight()+ geom\_boxplot(color="#3b2f2f")+ coord\_flip()+ facet\_wrap(~species)+

geom\_hline(data=arabica\_robusta\_average\_score, mapping=aes(yintercept=average\_score), size= 0.5)+

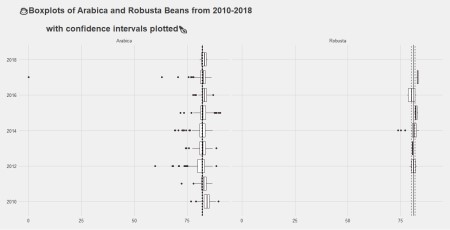
geom\_hline(data=arabica\_robusta\_average\_score, mapping=aes(yintercept=lower\_ci), linetype="dashed",

size= 0.5)+ geom\_hline(data=arabica\_robusta\_average\_score,

mapping=aes(yintercept=upper\_ci), linetype="dashed",

size= 0.5)+

ggtitle("Boxplots of Arabica and Robusta Beans from 2010-2018 \n with confidence intervals plotted")



Besides for some outliers on the lower end of the scoring range, most of these coffees in this dataset are on average score around 80 or above. What can be implied from here is that the coffees that come in to be graded by the Coffee Quality Institute are usually those which have are assumed to be high in quality.

# Conclusion

Its not surprising for our data set that Robusta beans scored poorer than their Arabica counterparts. That is something that anyone with some background in coffee will tell you- Arabica is generally more desirable by coffee drinkers and Robusta is usually used for instant coffee, Espresso and filler for coffee blends.