Final\_Project\_DG

Dillon Geary

2022-11-16

# Introduction

## About the dataset

The data I will be using is from a data set called coffee which I got from github. This data set has rating on different beans which have different attributes associated to them. The data was last updated in 2021.A lot of the data held in the data set is not very useful, such as amount of bags made etc. Therefore, the data I will use is listed below.

## Data used for analysis

* Total Cup Points
* Species
* Region
* Country
* Altitude

Using this data I will make deductions how each variable has an affect on the other, if any.

# Table of Data

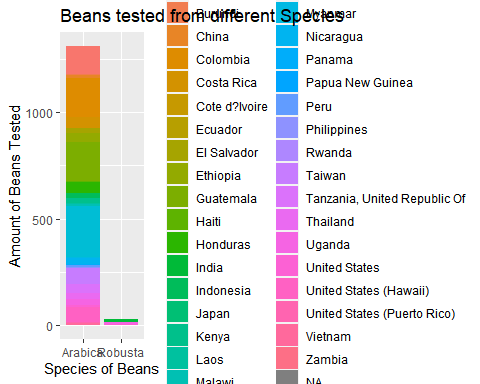
library(ggplot2)  
coffee<-read.csv("coffee.csv")  
head(coffee)

## total\_cup\_points species owner country\_of\_origin  
## 1 90.58 Arabica metad plc Ethiopia  
## 2 89.92 Arabica metad plc Ethiopia  
## 3 89.75 Arabica grounds for health admin Guatemala  
## 4 89.00 Arabica yidnekachew dabessa Ethiopia  
## 5 88.83 Arabica metad plc Ethiopia  
## 6 88.83 Arabica ji-ae ahn Brazil  
## farm\_name lot\_number mill ico\_number  
## 1 metad plc <NA> metad plc 2014/2015  
## 2 metad plc <NA> metad plc 2014/2015  
## 3 san marcos barrancas "san cristobal cuch <NA> <NA> <NA>  
## 4 yidnekachew dabessa coffee plantation <NA> wolensu <NA>  
## 5 metad plc <NA> metad plc 2014/2015  
## 6 <NA> <NA> <NA> <NA>  
## company altitude region  
## 1 metad agricultural developmet plc 1950-2200 guji-hambela  
## 2 metad agricultural developmet plc 1950-2200 guji-hambela  
## 3 <NA> 1600 - 1800 m <NA>  
## 4 yidnekachew debessa coffee plantation 1800-2200 oromia  
## 5 metad agricultural developmet plc 1950-2200 guji-hambela  
## 6 <NA> <NA> <NA>  
## producer number\_of\_bags bag\_weight  
## 1 METAD PLC 300 60 kg  
## 2 METAD PLC 300 60 kg  
## 3 <NA> 5 1  
## 4 Yidnekachew Dabessa Coffee Plantation 320 60 kg  
## 5 METAD PLC 300 60 kg  
## 6 <NA> 100 30 kg  
## in\_country\_partner harvest\_year grading\_date  
## 1 METAD Agricultural Development plc 2014 April 4th, 2015  
## 2 METAD Agricultural Development plc 2014 April 4th, 2015  
## 3 Specialty Coffee Association <NA> May 31st, 2010  
## 4 METAD Agricultural Development plc 2014 March 26th, 2015  
## 5 METAD Agricultural Development plc 2014 April 4th, 2015  
## 6 Specialty Coffee Institute of Asia 2013 September 3rd, 2013  
## owner\_1 variety processing\_method aroma flavor aftertaste  
## 1 metad plc <NA> Washed / Wet 8.67 8.83 8.67  
## 2 metad plc Other Washed / Wet 8.75 8.67 8.50  
## 3 Grounds for Health Admin Bourbon <NA> 8.42 8.50 8.42  
## 4 Yidnekachew Dabessa <NA> Natural / Dry 8.17 8.58 8.42  
## 5 metad plc Other Washed / Wet 8.25 8.50 8.25  
## 6 Ji-Ae Ahn <NA> Natural / Dry 8.58 8.42 8.42  
## acidity body balance uniformity clean\_cup sweetness cupper\_points moisture  
## 1 8.75 8.50 8.42 10 10 10 8.75 0.12  
## 2 8.58 8.42 8.42 10 10 10 8.58 0.12  
## 3 8.42 8.33 8.42 10 10 10 9.25 0.00  
## 4 8.42 8.50 8.25 10 10 10 8.67 0.11  
## 5 8.50 8.42 8.33 10 10 10 8.58 0.12  
## 6 8.50 8.25 8.33 10 10 10 8.33 0.11  
## category\_one\_defects quakers color category\_two\_defects  
## 1 0 0 Green 0  
## 2 0 0 Green 1  
## 3 0 0 <NA> 0  
## 4 0 0 Green 2  
## 5 0 0 Green 2  
## 6 0 0 Bluish-Green 1  
## expiration certification\_body  
## 1 April 3rd, 2016 METAD Agricultural Development plc  
## 2 April 3rd, 2016 METAD Agricultural Development plc  
## 3 May 31st, 2011 Specialty Coffee Association  
## 4 March 25th, 2016 METAD Agricultural Development plc  
## 5 April 3rd, 2016 METAD Agricultural Development plc  
## 6 September 3rd, 2014 Specialty Coffee Institute of Asia  
## certification\_address  
## 1 309fcf77415a3661ae83e027f7e5f05dad786e44  
## 2 309fcf77415a3661ae83e027f7e5f05dad786e44  
## 3 36d0d00a3724338ba7937c52a378d085f2172daa  
## 4 309fcf77415a3661ae83e027f7e5f05dad786e44  
## 5 309fcf77415a3661ae83e027f7e5f05dad786e44  
## 6 726e4891cf2c9a4848768bd34b668124d12c4224  
## certification\_contact unit\_of\_measurement  
## 1 19fef5a731de2db57d16da10287413f5f99bc2dd m  
## 2 19fef5a731de2db57d16da10287413f5f99bc2dd m  
## 3 0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660 m  
## 4 19fef5a731de2db57d16da10287413f5f99bc2dd m  
## 5 19fef5a731de2db57d16da10287413f5f99bc2dd m  
## 6 b70da261fcc84831e3e9620c30a8701540abc200 m  
## altitude\_low\_meters altitude\_high\_meters altitude\_mean\_meters  
## 1 1950 2200 2075  
## 2 1950 2200 2075  
## 3 1600 1800 1700  
## 4 1800 2200 2000  
## 5 1950 2200 2075  
## 6 NA NA NA

mod1 <- lm(total\_cup\_points~aroma+body+category\_two\_defects+moisture+altitude\_mean\_meters,data=coffee)

# First Plot

ggplot(data = coffee, mapping = aes(x = species,fill=country\_of\_origin))+  
 geom\_bar() + xlab("Species of Beans")+ ylab("Amount of Beans Tested")+ggtitle("Beans tested from different Species")

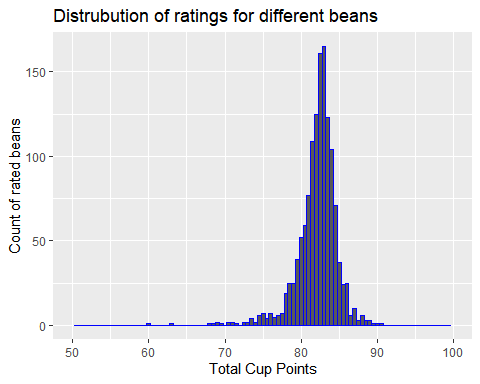
 The above bar graph shows that most of the beans that were tested come from the Arabica species. This might be because most beans in the world are the Arabica type or is the preferred one for consumers and that’s why there is more of it.

# Second Plot

ggplot(coffee) +  
geom\_histogram(mapping = aes(x = total\_cup\_points), binwidth = .5,color='Blue')+ xlab("Total Cup Points")+ ylab("Count of rated beans") + ggtitle("Distrubution of ratings for different beans") +xlim(50,100)

## Warning: Removed 1 rows containing non-finite values (stat\_bin).

## Warning: Removed 2 rows containing missing values (geom\_bar).

 Many of the cups score a total cup point above 75% this shows that most of the beans which were tested are of good quality. Thus, the average bean would still brew a good cup of coffee. It can be said that different beans therefore don’t have a huge affect on its rating.

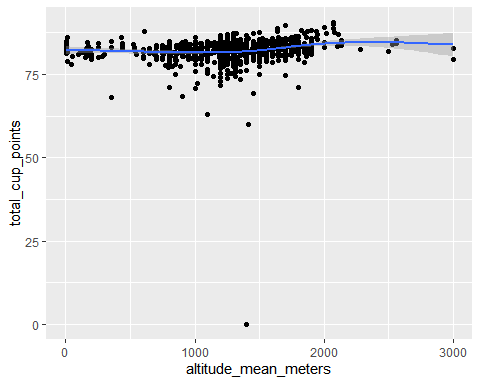
# Third plot

ggplot(mod1) +   
 geom\_point(mapping = aes(x = altitude\_mean\_meters , y=total\_cup\_points)) +   
 geom\_smooth(aes(x = altitude\_mean\_meters, y=total\_cup\_points))+xlim(0,3000)

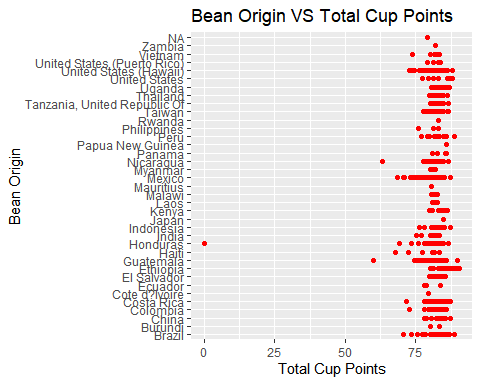
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 16 rows containing non-finite values (stat\_smooth).

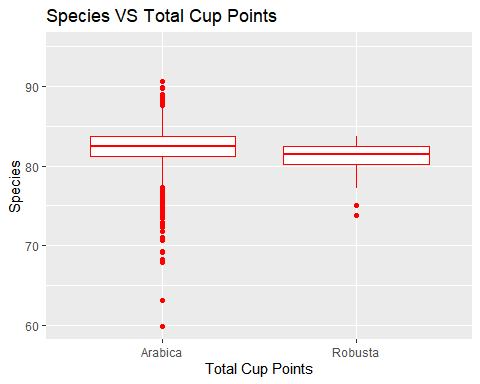
## Warning: Removed 16 rows containing missing values (geom\_point).

 In the above graph you can see that the altitude does not have an affect on the total cup rating as the distribution for all altitudes. Therefore, it can be said that there is an other variable that has an affect on the total cup points. # Fourth Plot

ggplot(data = coffee, aes(x = total\_cup\_points , y = country\_of\_origin)) +  
 geom\_point(color='red') + xlab("Total Cup Points")+ ylab("Bean Origin")+ggtitle("Bean Origin VS Total Cup Points")

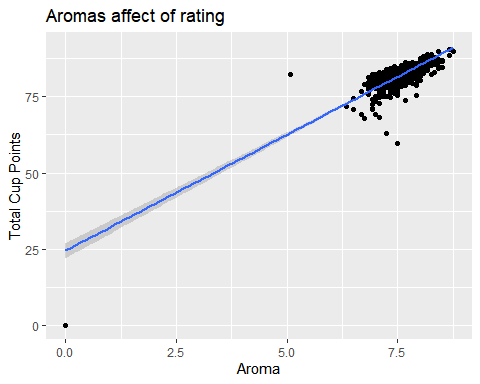
 The country that the beans are grown in also does not seem to have an effect on the total cup. The outlier from Honduras, is probably just a bag which was bad quality from the maker as it is rated far lower than the rest from the same country of origin. # Fifth Plot

ggplot(data = coffee, aes(x = species , y = total\_cup\_points)) +  
 geom\_boxplot(color='red')+ xlab("Total Cup Points")+ ylab("Species")+ggtitle("Species VS Total Cup Points")+coord\_cartesian(ylim=c(60,95))

 The Arabica beans are rated more highly than the Robusta beans. This could be due to many factor such as judges taste or the amount of beans tested from each species. It is more likely that the Arabica beans are just better because they have been more refined due to their being more of them compared to the Robusta as seen in plot 1.

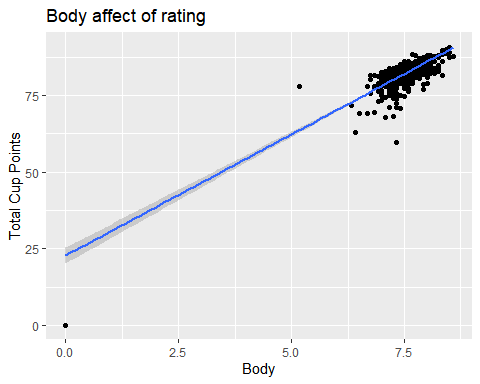
ggplot(mod1) +   
 geom\_point(mapping = aes(x = aroma , y=total\_cup\_points)) +   
 geom\_smooth(aes(x = aroma, y=total\_cup\_points), method = "lm")+xlab("Aroma")+ ylab("Total Cup Points")+ggtitle("Aromas affect of rating")

## `geom\_smooth()` using formula 'y ~ x'

 As the aroma score increases there is a trend for a increase in the total cup points. This shows that the aroma is important for a good cup of coffee.

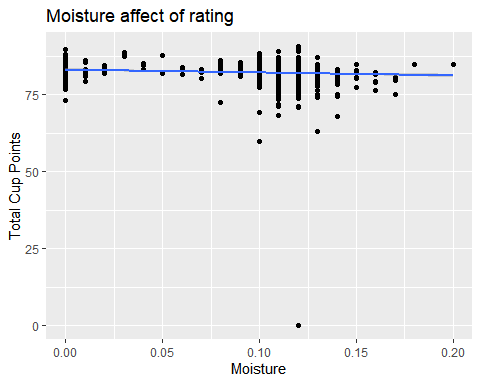
ggplot(mod1) +   
 geom\_point(mapping = aes(x = body , y=total\_cup\_points)) +   
 geom\_smooth(aes(x = body, y=total\_cup\_points), method = "lm")+xlab("Body")+ ylab("Total Cup Points")+ggtitle("Body affect of rating")

## `geom\_smooth()` using formula 'y ~ x'



ggplot(mod1) +   
 geom\_point(mapping = aes(x = moisture , y=total\_cup\_points)) +   
 geom\_smooth(aes(x = moisture, y=total\_cup\_points), method = "lm")+xlab("Moisture")+ ylab("Total Cup Points")+ggtitle("Moisture affect of rating")

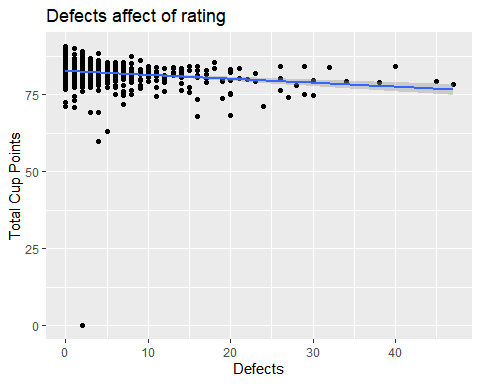
## `geom\_smooth()` using formula 'y ~ x'



The more moisture the beans has the worst the rating for a cup goes down. This could be that the moisture in the beans has a affect on the taste.

ggplot(mod1) +   
 geom\_point(mapping = aes(x = category\_two\_defects , y=total\_cup\_points)) +   
 geom\_smooth(aes(x = category\_two\_defects, y=total\_cup\_points), method = "lm")+xlab("Defects")+ ylab("Total Cup Points")+ggtitle("Defects affect of rating")

## `geom\_smooth()` using formula 'y ~ x'



The more defects there are in the beans the worse the rating.