Project3B

Tina Trinh, Ian Kramer, Mary Testa, and Ben Strougal

2024-05-02

Disclaimer

Our data set keeps loading in differently which is causing inaccurate outputs in our model than what it was earlier, so a couple things like the subset regression are different than the ones used in the presentation.

Introduction

> This data was retrieved from the Coffee Quality Database courtesy of Buzzfeed Data Scientist James LeDoux. The information was collected from the Coffee Quality Institute's review pages in January 2018. This data holds very detailed information over Arabica and Robusta beans, across many countries and they are professionally rated on a 0-100 scale. There are many rates/scores for things like acidity, sweetness, fragrance, balance, etc. Here is a brief explanation of all the variables included in the data frame with 1311 observations on the following 44 variables:

Variable	Description
...1	id
Species	Species of coffee bean (arabica or robusta)
Owner	Owner of the farm
Country.of.Origin	Where the bean came from
Farm.Name	Name of the farm
Lot.Number	Lot number of the beans tested
Mill	Mill where the beans were processed
ICO.Number	International Coffee Organization number
Company	Company name
Altitude	Altitude
Region	Region where bean came from
Producer	Producer of the roasted bean
Number.of.Bags	Number of bags tested
Bag.Weight	Bag weight tested
In.Country.Partner	Partner for the country
Harvest.Year	When the beans were harvested (year)
Grading.Date	When the beans were graded
Owner.1	Who owns the beans
Variety	Variety of the beans
Processing.Method	Method for processing
Aroma	Aroma grade
Flavor	Flavor grade
Aftertaste	Aftertaste grade
Acidity	Acidity grade

Variable	Description
Body	Body grade
Balance	Balance grade
Uniformity	Uniformity grade
Clean.Cup	Clean cup grade
Sweetness	Sweetness grade
Cupper.Points	Cupper Points
Total.Cup.Points	Total rating/points (0 - 100 scale)
Moisture	Moisture Grade
Category.One.Defects	Category one defects (count)
Quakers	quakers
Color	Color of bean
Category.Two.Defects	Category two defects (count)
Expiration	Expiration date of the beans
Certification.Body	Who certified it
Certification.Address	Certification body address
Certification.Contact	Certification contact
unit_of_measurement	Unit of measurement
altitude_low_meters	Altitude low meters
altitude_high_meters	Altitude high meters
altitude_mean_meters	Altitude mean meters

The analysis begins with a linear regression, because it is by far the simplest one to run and will likely expose issues that can be resolved by choosing a different model. It is very unlikely that the linear regression will produce substantive results, but it serves as a solid base to start the analysis from.

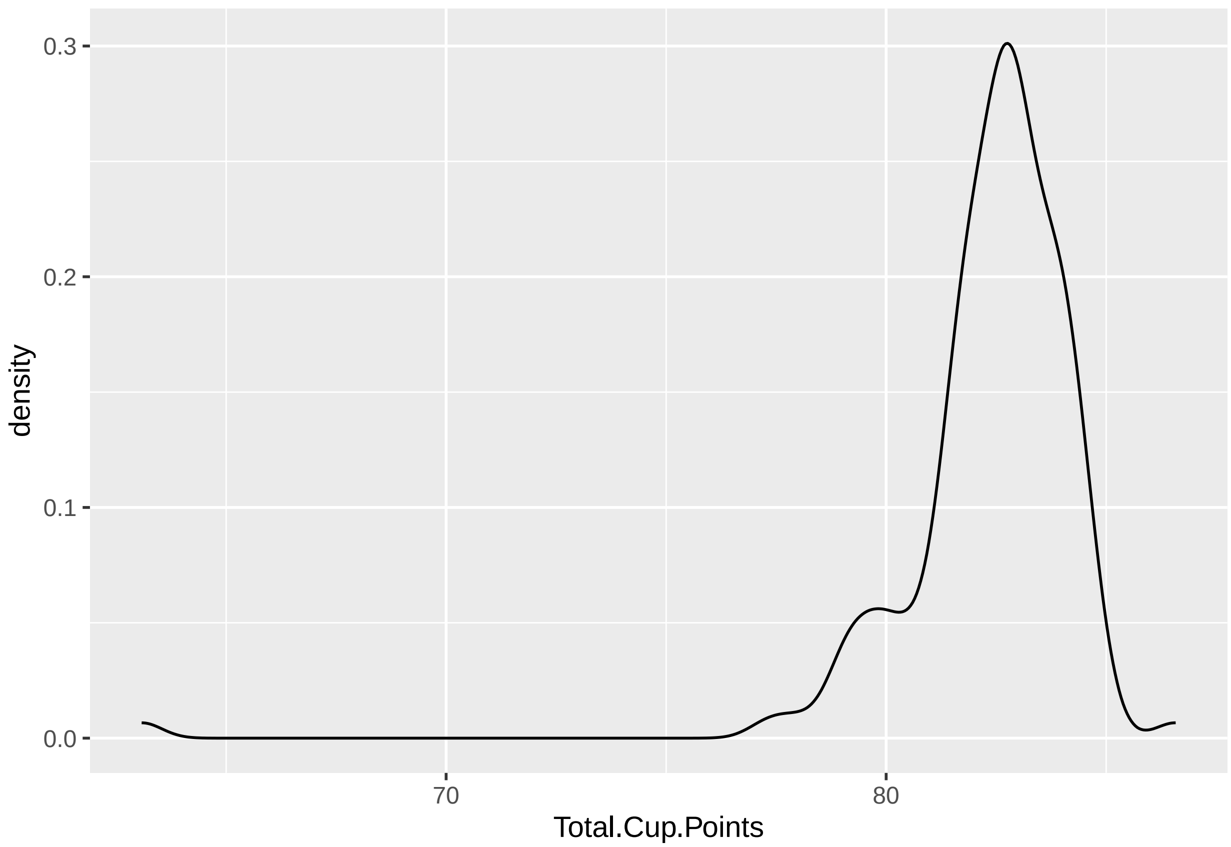
The explanatory variable chosen was total_cup_points. This was a fairly obvious choice, given that the ideal outcome was to determine what factors go into making the best cup of coffee. To this end, it was established fairly early on that the total_cup_points were simply equal to the sum of ten other variables already present in the dataset: aroma, flavor, aftertaste, acidity, body, balance, uniformity, clean_cup, sweetness, and cupper_points. Therefore, using these variables in any model would be both redundant and uninformative, as it is already known what the relation between them and the total score is.

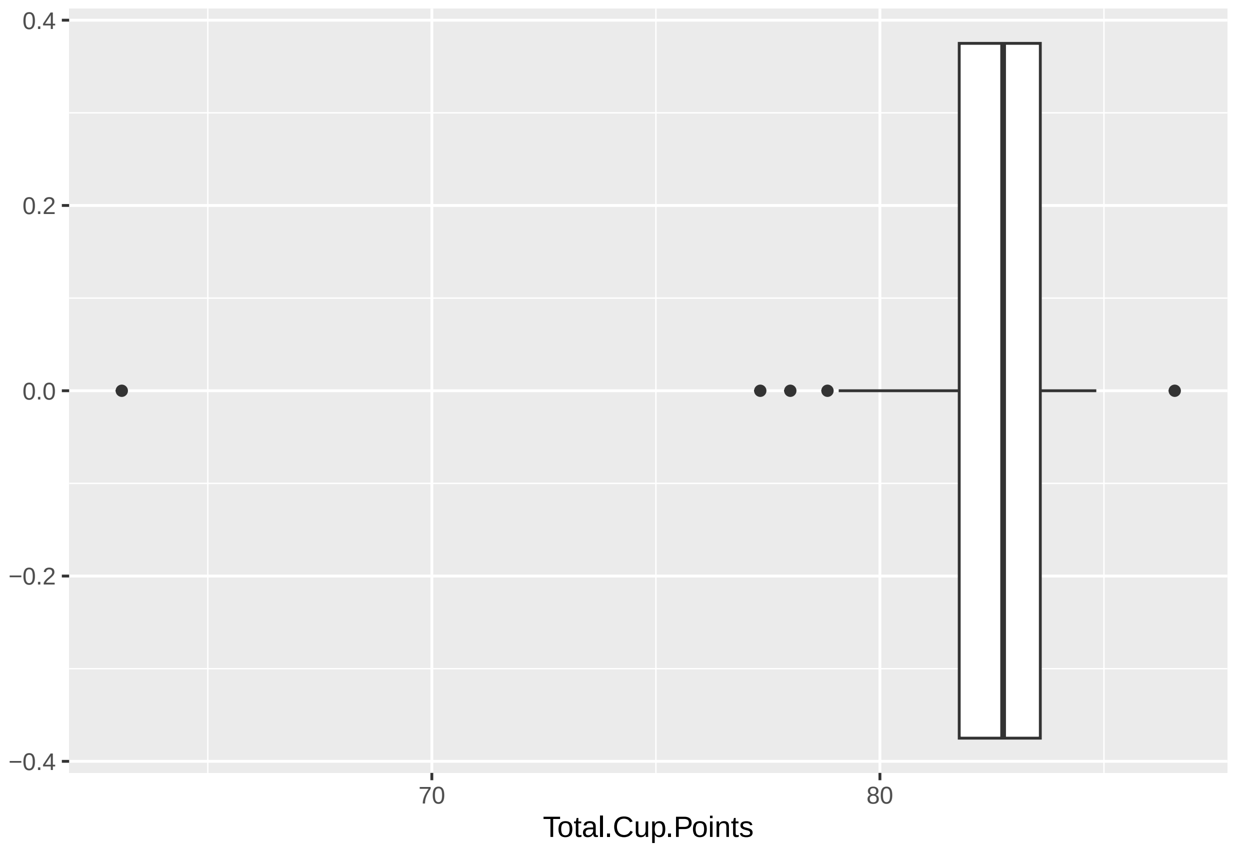
Some of the EDA was made before this realization occured, but keeping them present doesn't cause any long-term issues.

EDA

We first check if the dataset contains any NA values and found that there's 3877 NA values. We then create a new clean dataset that omit the rows that has NA value and call that new dataset `coffee.clean`

When we plot the distribution of the variable `Total.Cup.Points`, we see that distribution of the variable `Total.Cup.Points` is skewed to the left and unimodal. We can interpret this distribution as, in general, we have pretty high total rating points with the mean of around 82 points. The skewness of the distribution may suggest that there might be underlying interaction terms which are affecting the distribution of the variable. Note that there are 2 extreme outliers on this ditribution (based on the boxplot), so we'd use this distribution with caution.


0.3
0.2
0.1
0.0
density
70
80
Total.Cup.Points



```
## [1] 82.33969
```

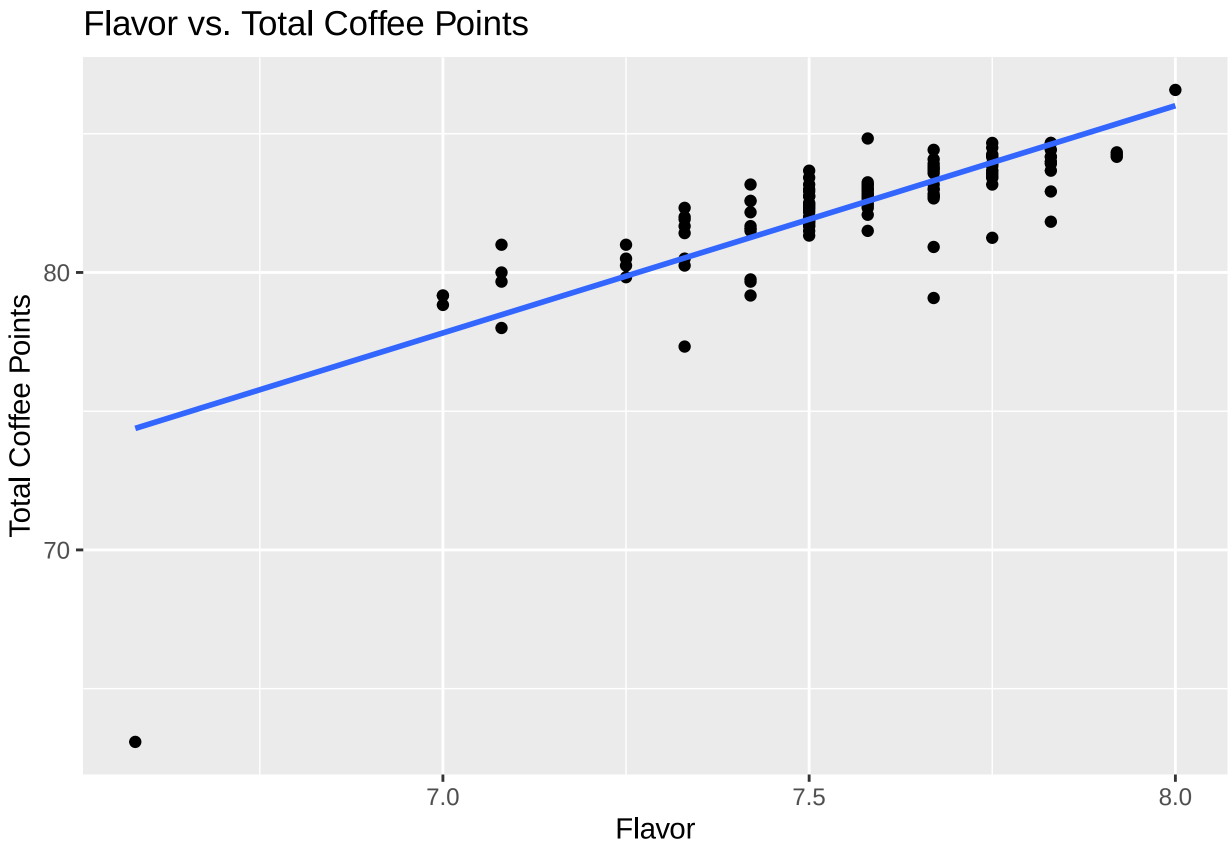
We simplified the categorical variable of country of origin of the coffee bean by assigning the countries based on their continents (Asia, South America, Africa) rather than having the specific country of where the beans came from. We assign the new dataset as `coffee.new` that contains the grouping of the countries based on their continent.

Since we want to simplify the process of model selection, we'd exclude the variables that have string values and only focus on those with numerical values then assign this new dataset as `coffee.new`.

```
## Rows: 130
## Columns: 19
## $ Number.of.Bags       <dbl> 20, 10, 150, 15, 120, 200, 275, 200, 275, 10, 230~
## $ Processing.Method    <chr> "Pulped natural / honey", "Natural / Dry", "Washe~
## $ Aroma                <dbl> 8.00, 7.92, 7.83, 8.08, 7.75, 7.92, 7.58, 7.67, 7~
## $ Flavor               <dbl> 8.00, 7.58, 7.83, 7.75, 7.83, 7.75, 7.83, 7.67, 7~
## $ Aftertaste           <dbl> 8.00, 7.83, 7.50, 7.67, 7.58, 7.67, 7.83, 7.83, 7~
## $ Acidity              <dbl> 8.25, 7.83, 8.00, 7.83, 8.00, 7.75, 8.00, 7.58, 7~
## $ Body                 <dbl> 8.00, 7.83, 7.83, 7.50, 7.92, 7.83, 7.67, 7.83, 7~
## $ Balance              <dbl> 8.17, 7.83, 7.67, 7.92, 7.75, 7.75, 7.58, 7.83, 7~
## $ Uniformity           <dbl> 10.00, 10.00, 10.00, 10.00, 10.00, 10.00, 10.00, ~
## $ Clean.Cup            <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 1~
## $ Sweetness            <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 1~
## $ Cupper.Points        <dbl> 8.17, 8.00, 8.00, 7.92, 7.83, 7.83, 7.92, 8.00, 7~
## $ Total.Cup.Points     <dbl> 86.58, 84.83, 84.67, 84.67, 84.67, 84.50, 84.42, ~
## $ Moisture             <dbl> 0.00, 0.00, 0.00, 0.10, 0.10, 0.11, 0.10, 0.00, 0~
## $ Category.One.Defects <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Quakers              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 4, 0, 0, 0, 0, 1, 0~
## $ Color                <chr> "Green", "Green", "Blue-Green", "Blue-Green", "Gr~
## $ Category.Two.Defects <dbl> 0, 0, 2, 2, 1, 1, 3, 3, 2, 0, 4, 1, 1, 2, 6, 10, ~
## $ country_group        <chr> "Asia", "Asia", "South America", "South America",~
```

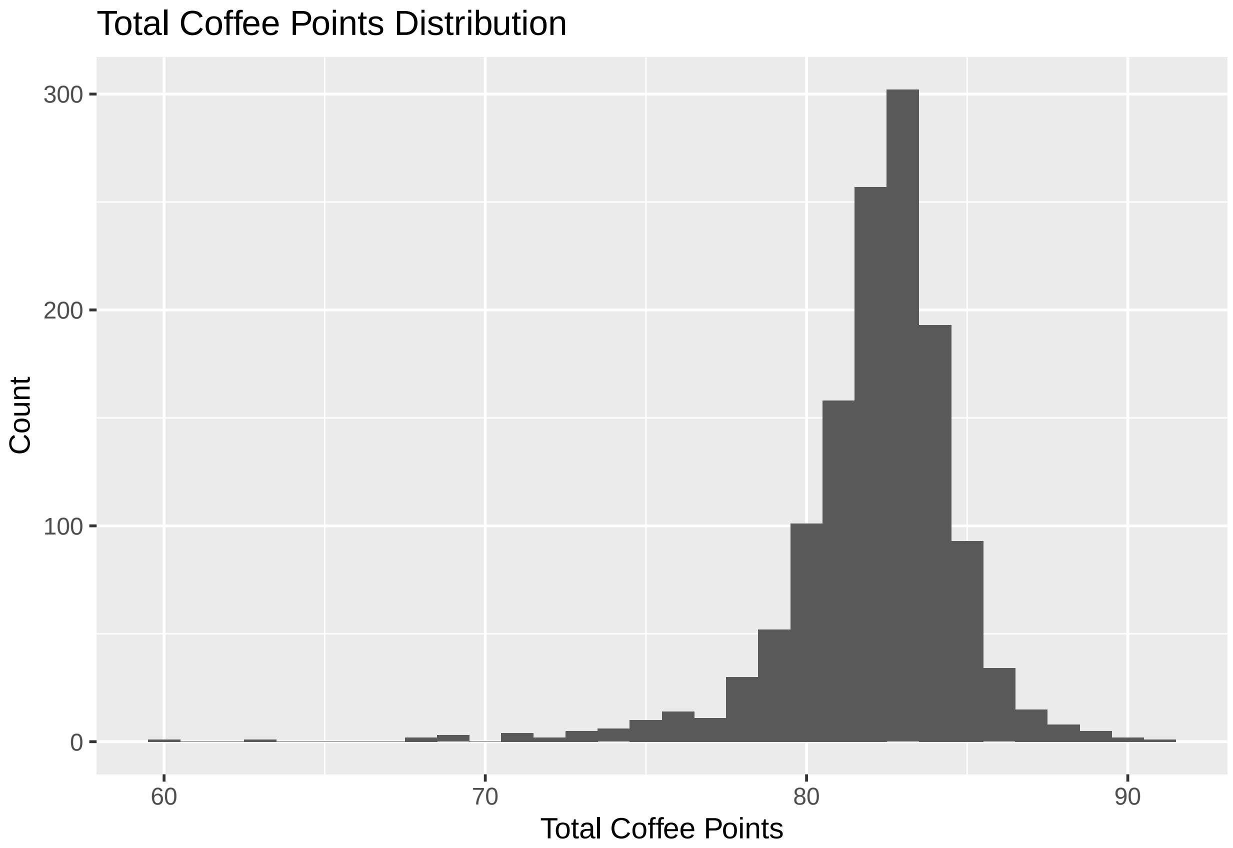
We've discussed that there might be some interaction terms that are affecting the distribution of the variable `Total.Cup.Points`, and based on the plot, it suggests that there may be an interaction between the variables 'Country.Group' and 'Flavor'. Specifically, the distribution of 'Total.Cup.Points' appears to vary across different levels of 'Flavor' within each level of 'Country.Group'. This suggests that the effect of 'Flavor' on 'Total.Cup.Points' may depend on the 'Country.Group', indicating a potential interaction between these two factors.

```
## 'geom_smooth()' using formula = 'y ~ x'
```

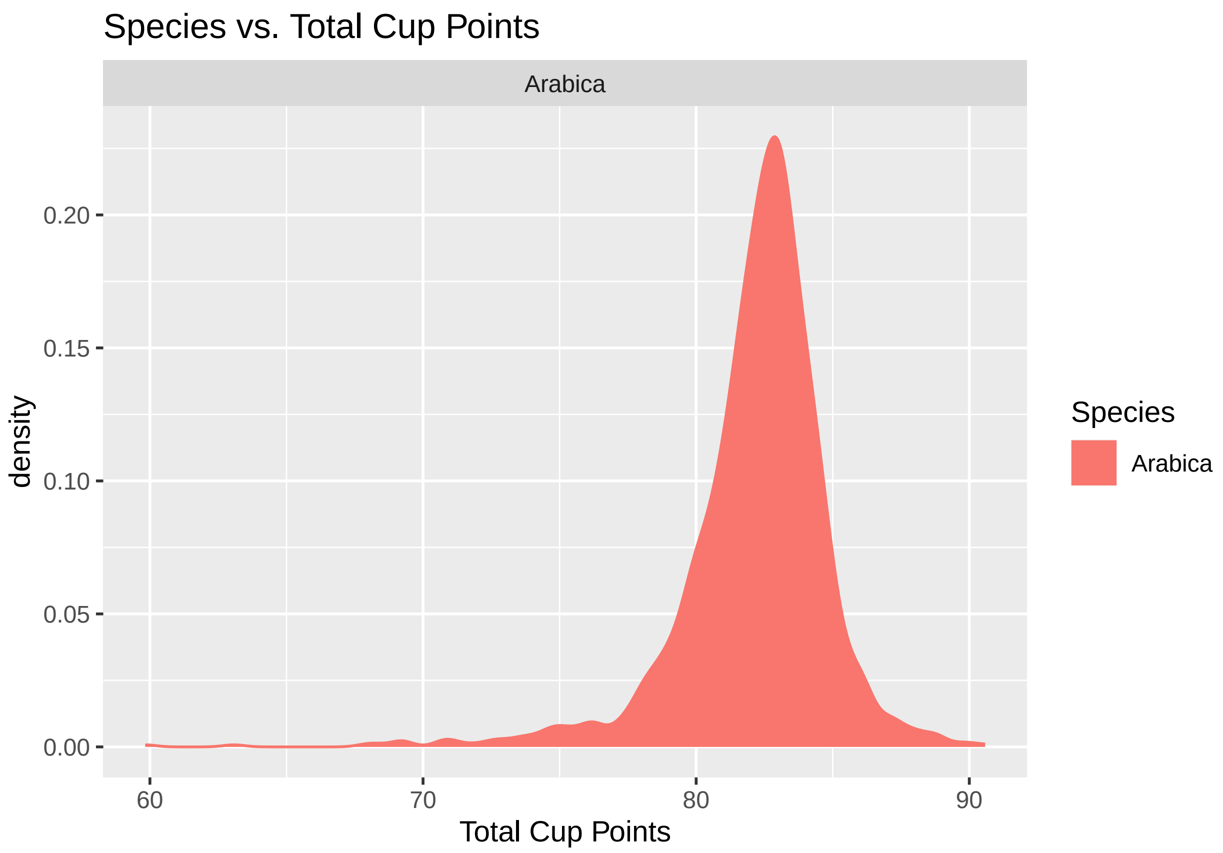


Looking at this plot, it's evident that flavor shares a clear positive relationship with total coffee points. It makes sense that as flavor increases, the perception of the cup does as well. A good tasting coffee is likely to be a good cup of joe overall. There is an outlier at the low end of the plot, and it doesn't follow along with the trend. It's rating is much lower than the plot would have predicted, so there may be other factors within that cup of coffee that make it worse besides flavor.

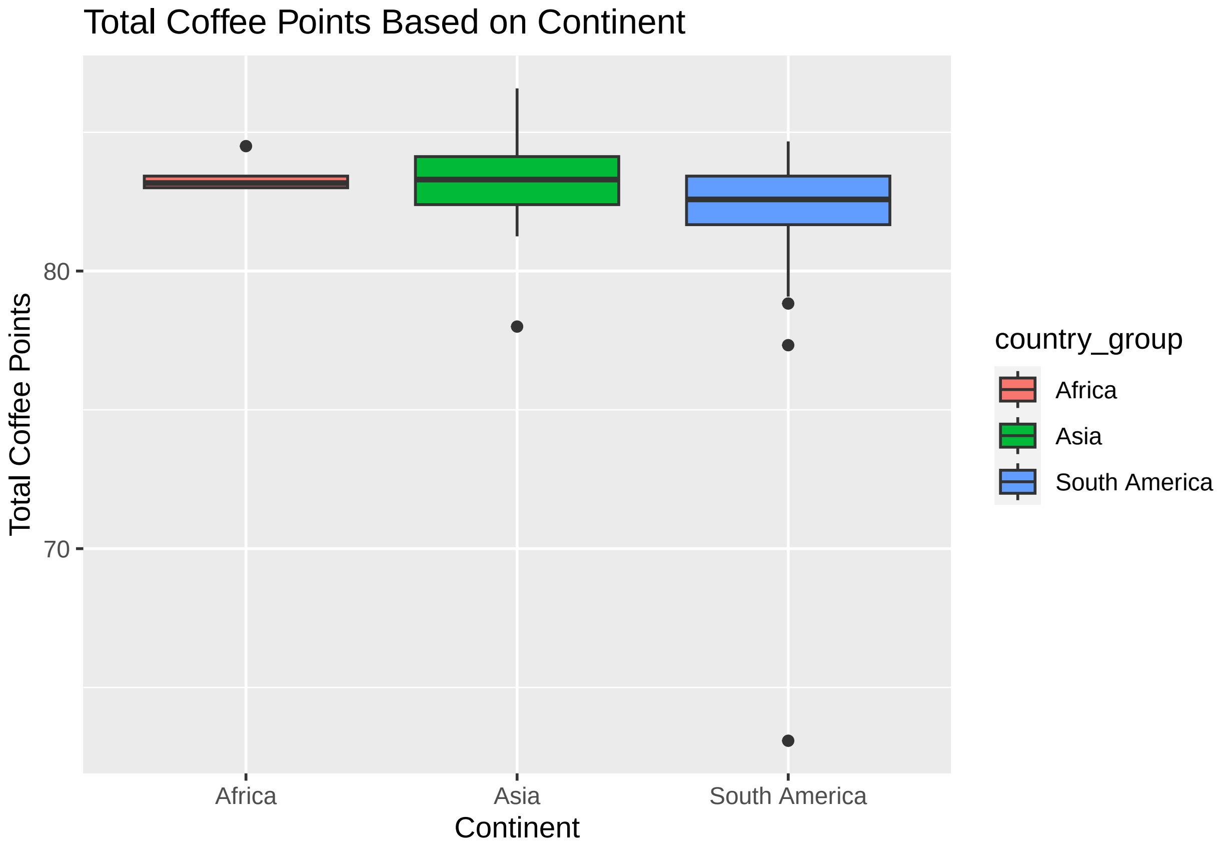
Ian’s EDA Graphs



There was one observation in the data that had a total coffee point score of 0. This is an extreme outlier, and also doesn’t really make sense logistically. It would be hard for any cup of coffee to truly score a flat 0 without there being some sort of bias in the rating. With a score that low, it could affect the model later on, so removing it from the data would be a good decision.

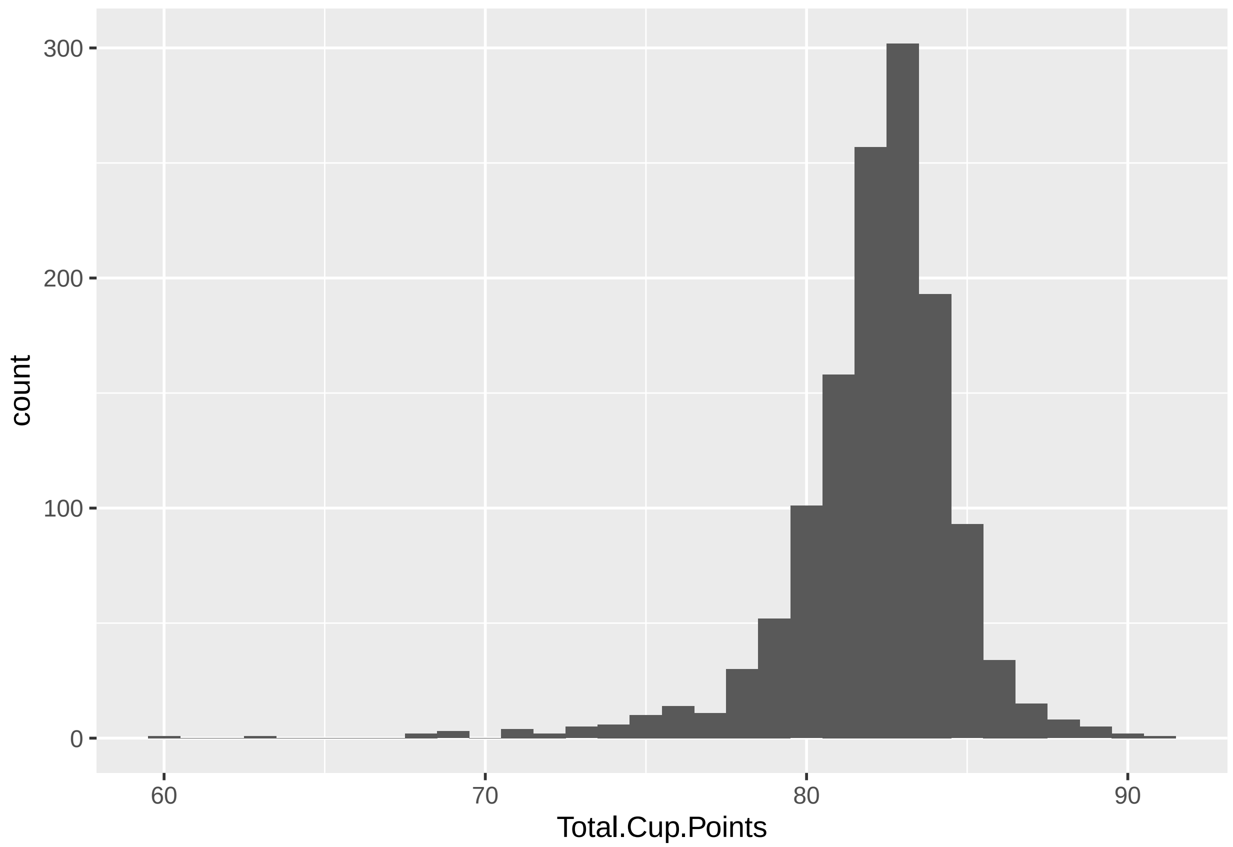


```
ggplot(data = coffee.new, aes(x = country_group, y = Total.Cup.Points, fill = country_group)) + geom_bo
```

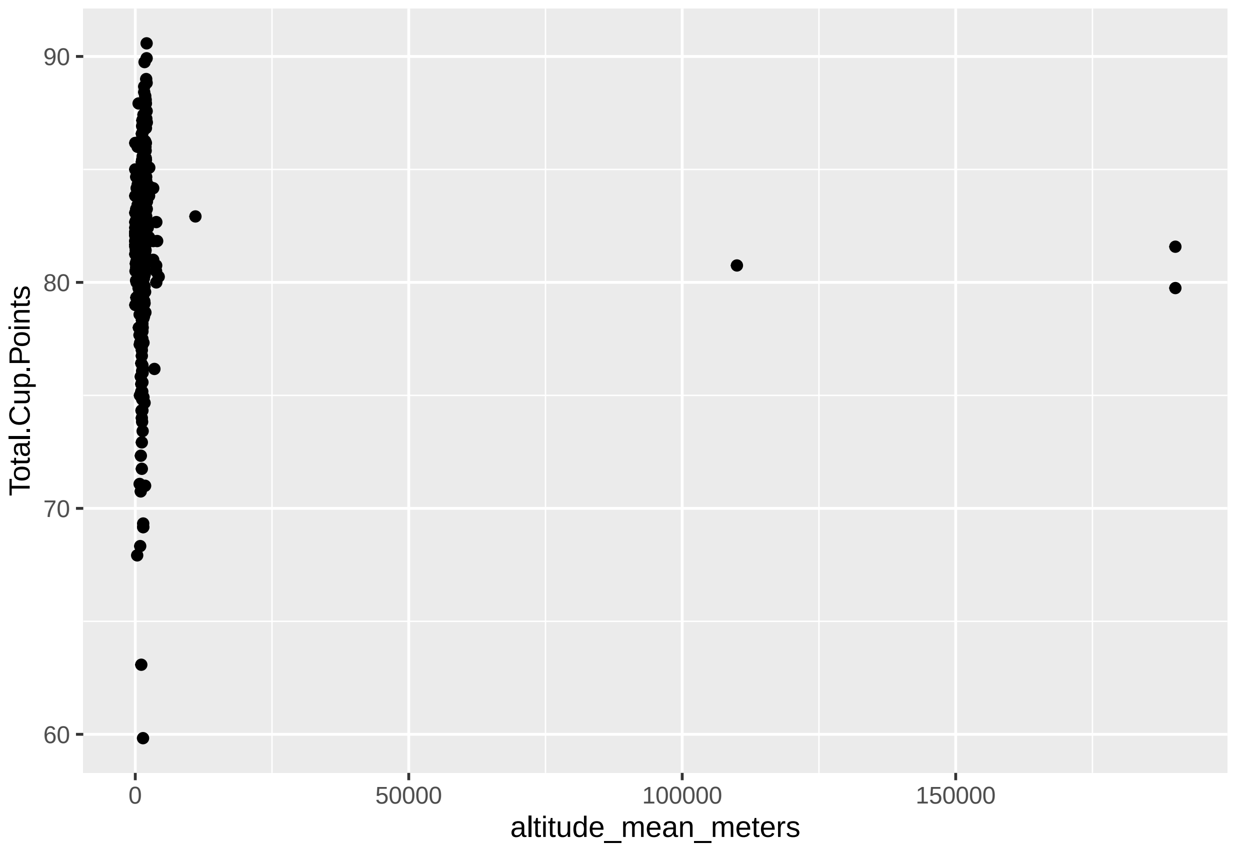


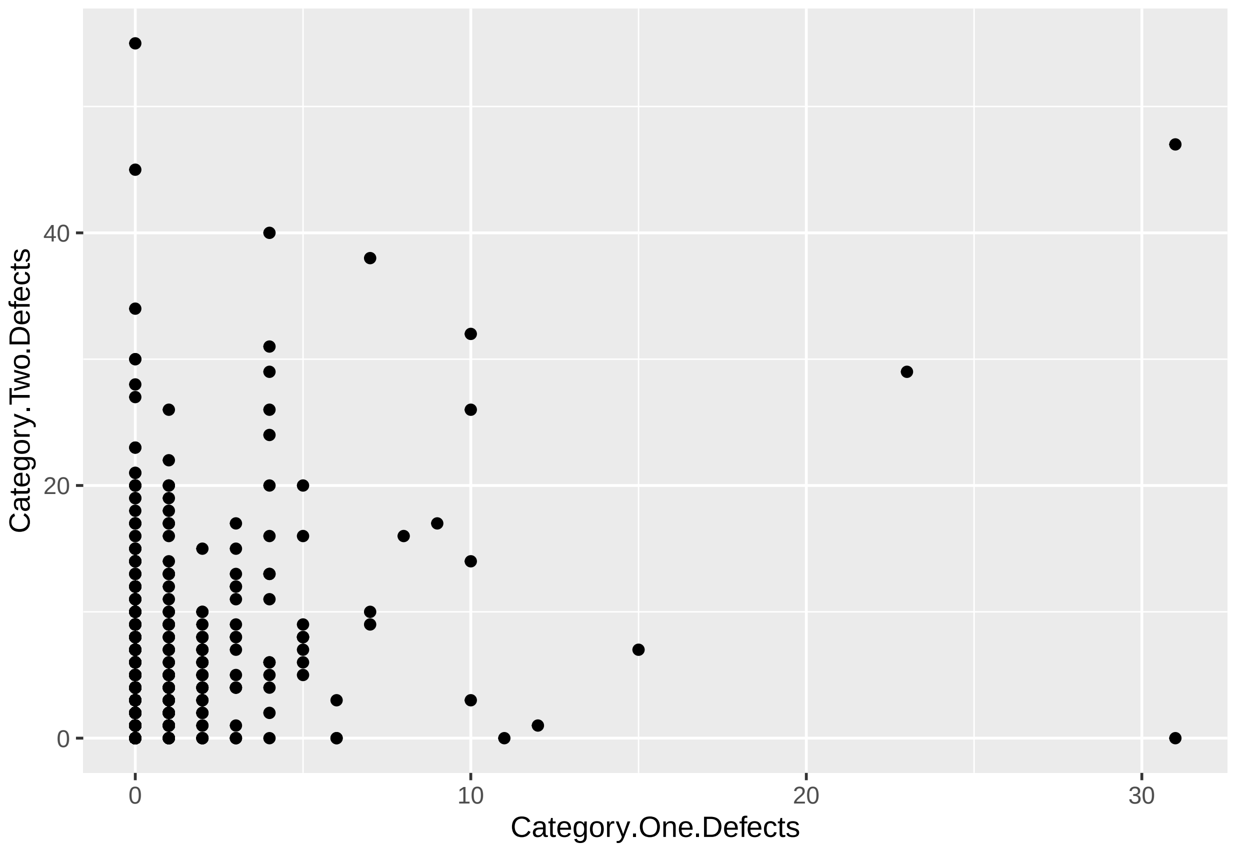
These plots look extremely similar, but the Robusta species has a lower mean than Arabica. Since the shape of the density plots are so similar in shape, it seems that Robusta as a species is very close in consistency to Arabica. Because the mean is lower though, there may be some kind of genetic issue with the bean that maybe doesn't bring out as much flavor or something like that. Overall though, the species are very comparable to one another.

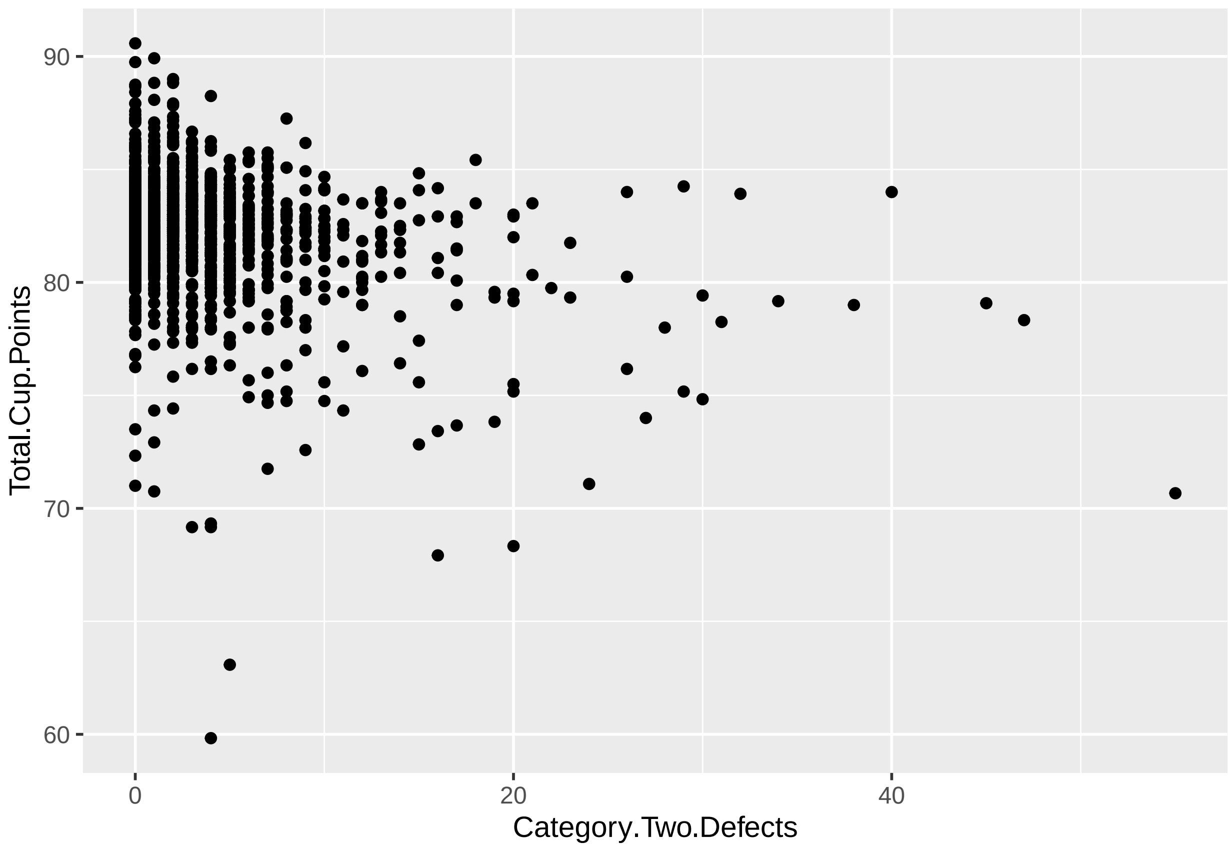
We filtered out the 0 total cup points since they are extreme low outliers.



```
## Warning: Removed 227 rows containing missing values (`geom_point()`).
```


Total.Cup.Points
90
80
70
60
0
50000
100000
150000
altitude_mean_meters


Category.Two.Defects
40
20
0
0
10
20
30
Category.One.Defects



MODEL SELECTION

```
linear.model.all <- lm(Total.Cup.Points ~ ., data=coffee.new)
summary(linear.model.all)
```

```
##
## Call:
## lm(formula = Total.Cup.Points ~ ., data = coffee.new)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0262128 -0.0042212 -0.0008627  0.0064472  0.0201168
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -9.014e-02  4.830e-02  -1.866   0.0647
## Number.of.Bags                       6.042e-06  8.560e-06   0.706   0.4818
## Processing.MethodOther               5.459e-03  5.094e-03   1.072   0.2863
## Processing.MethodPulped natural / honey -9.441e-05  4.718e-03  -0.020   0.9841
## Processing.MethodWashed / Wet       -1.313e-04  2.262e-03  -0.058   0.9538
## Aroma                                1.008e+00  7.217e-03 139.638   <2e-16
## Flavor                               9.946e-01  8.736e-03 113.848   <2e-16
## Aftertaste                           9.978e-01  7.171e-03 139.151   <2e-16
## Acidity                                 1.006e+00  6.189e-03 162.590   <2e-16
## Body                                    1.010e+00  6.377e-03 158.357   <2e-16
## Balance                                 9.919e-01  8.000e-03 123.977   <2e-16
## Uniformity                              9.963e-01  2.712e-03 367.339   <2e-16
## Clean.Cup                               1.008e+00  5.360e-03 188.001   <2e-16
## Sweetness                               9.986e-01  4.711e-03 211.947   <2e-16
## Cupper.Points                           1.000e+00  2.502e-03 399.725   <2e-16
## Moisture                               -2.283e-02  2.319e-02  -0.985   0.3270
## Category.One.Defects                    1.181e-04  3.224e-04   0.366   0.7149
## Quakers                                -8.601e-05  6.431e-04  -0.134   0.8939
## ColorBluish-Green                       5.468e-04  4.079e-03   0.134   0.8936
## ColorGreen                              1.820e-03  3.327e-03   0.547   0.5855
## Category.Two.Defects                   -2.665e-04  3.806e-04  -0.700   0.4852
## country_groupAsia                      -2.249e-03  5.544e-03  -0.406   0.6858
## country_groupSouth America              2.658e-03  4.239e-03   0.627   0.5320
##
## (Intercept)                            .
## Number.of.Bags
## Processing.MethodOther
## Processing.MethodPulped natural / honey
## Processing.MethodWashed / Wet
## Aroma                                  ***
## Flavor                                 ***
## Aftertaste                             ***
## Acidity                                ***
## Body                                   ***
## Balance                                ***
## Uniformity                             ***
## Clean.Cup                              ***
## Sweetness                              ***
## Cupper.Points                          ***
## Moisture
## Category.One.Defects
## Quakers
## ColorBluish-Green
## ColorGreen
## Category.Two.Defects
## country_groupAsia
## country_groupSouth America
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.008698 on 107 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 4.029e+05 on 22 and 107 DF,  p-value: < 2.2e-16
```

```
coffeeSub <- regsubsets(`Total.Cup.Points` ~ Category.Two.Defects + Category.One.Defects +
    Moisture + Quakers + altitude_mean_meters + Number.of.Bags, data = coffee.clean, nbest=2)
plot(coffeeSub)

model3 <- lm(Total.Cup.Points~ Category.Two.Defects + Moisture, data = coffee.new)

summary(model3)
```

To perform model selection, a subset selection of variables can be created to help choose variables to put into a linear model. After running the subset selection, the best model that can be created is a linear model with category two defects and moisture as the sole explanatory variables. Even using the best model possible for a linear model, the adjusted R squared is still extremely low. Because of this, a linear model shouldn't be used, and a different model should be found. A gamma might be better in this scenario because our data is continuous and positive.

```
naCoffee = coffee.new %>% drop_na()
```

```
stepwise <- lm(Total.Cup.Points ~ . - Flavor - Cupper.Points -
    Aroma - Aftertaste - Body - Acidity - Balance - Clean.Cup -
    Sweetness - Uniformity, data= coffee.new)
model_b <- step(stepwise, direction='backward')
```

```
## Start:  AIC=203.26
## Total.Cup.Points ~ (Number.of.Bags + Processing.Method + Aroma +
##     Flavor + Aftertaste + Acidity + Body + Balance + Uniformity +
##     Clean.Cup + Sweetness + Cupper.Points + Moisture + Category.One.Defects +
##     Quakers + Color + Category.Two.Defects + country_group) -
##     Flavor - Cupper.Points - Aroma - Aftertaste - Body - Acidity -
##     Balance - Clean.Cup - Sweetness - Uniformity
##
##                        Df Sum of Sq    RSS    AIC
## - Quakers               1     0.007 508.31 201.26
## - Number.of.Bags        1     0.131 508.43 201.29
## - Moisture              1     1.201 509.50 201.57
## - Category.One.Defects  1     1.772 510.07 201.71
## - Category.Two.Defects  1     2.049 510.35 201.78
## <none>                              508.30 203.26
## - country_group         2    17.336 525.64 203.62
## - Color                 2    18.831 527.13 203.99
## - Processing.Method     3   112.701 621.00 223.29
##
## Step:  AIC=201.26
## Total.Cup.Points ~ Number.of.Bags + Processing.Method + Moisture +
##     Category.One.Defects + Color + Category.Two.Defects + country_group
##
##                        Df Sum of Sq    RSS    AIC
## - Number.of.Bags        1     0.139 508.45 199.30
## - Moisture              1     1.272 509.58 199.59
## - Category.One.Defects  1     1.767 510.08 199.71
## - Category.Two.Defects  1     2.195 510.51 199.82
## <none>                              508.31 201.26
## - country_group         2    17.416 525.73 201.64
## - Color                 2    18.869 527.18 202.00
## - Processing.Method     3   112.780 621.09 221.31
##
## Step:  AIC=199.3
## Total.Cup.Points ~ Processing.Method + Moisture + Category.One.Defects +
##     Color + Category.Two.Defects + country_group
##
##                        Df Sum of Sq    RSS    AIC
## - Moisture              1     1.298 509.75 197.63
## - Category.One.Defects  1     1.809 510.26 197.76
## - Category.Two.Defects  1     2.068 510.52 197.82
## <none>                              508.45 199.30
## - country_group         2    18.516 526.96 199.95
## - Color                 2    20.432 528.88 200.42
## - Processing.Method     3   116.229 624.68 220.06
##
## Step:  AIC=197.63
## Total.Cup.Points ~ Processing.Method + Category.One.Defects +
##     Color + Category.Two.Defects + country_group
##
##                        Df Sum of Sq    RSS    AIC
## - Category.Two.Defects  1     1.905 511.65 196.11
## - Category.One.Defects  1     1.984 511.73 196.13
## <none>                              509.75 197.63
## - country_group         2    19.318 529.06 198.47
## - Color                 2    21.650 531.40 199.04
## - Processing.Method     3   118.554 628.30 218.81
##
## Step:  AIC=196.11
## Total.Cup.Points ~ Processing.Method + Category.One.Defects +
##     Color + country_group
##
##                        Df Sum of Sq    RSS    AIC
## - Category.One.Defects  1     1.815 513.47 194.57
## <none>                              511.65 196.11
## - country_group         2    21.955 533.61 197.58
## - Color                 2    22.767 534.42 197.77
## - Processing.Method     3   121.884 633.53 217.89
##
## Step:  AIC=194.57
## Total.Cup.Points ~ Processing.Method + Color + country_group
##
##                     Df Sum of Sq    RSS    AIC
## <none>                           513.47 194.57
## - country_group      2    20.347 533.81 195.63
## - Color              2    23.698 537.16 196.44
## - Processing.Method  3   120.838 634.30 216.05
```

We chose the "best" linear model based on their AIC, which is 194.57

```
best.linear.model <- lm(Total.Cup.Points ~ Processing.Method + Color + country_group, data = coffee.new
summary(best.linear.model)
```

```
##
## Call:
## lm(formula = Total.Cup.Points ~ Processing.Method + Color + country_group,
##     data = coffee.new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0973  -0.5239   0.2059   1.0491   4.5221
##
```

```
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             85.2267     1.2460  68.401  < 2e-16
## Processing.MethodOther                  -5.1200     1.0011  -5.114 1.18e-06
## Processing.MethodPulped natural / honey   0.7255     1.0336   0.702   0.4840
## Processing.MethodWashed / Wet           -0.1764     0.4332  -0.407   0.6846
## ColorBluish-Green                       -1.4676     0.8864  -1.656   0.1003
## ColorGreen                              -1.6676     0.7042  -2.368   0.0194
## country_groupAsia                       -0.2088     1.0837  -0.193   0.8475
## country_groupSouth America              -1.2618     0.9431  -1.338   0.1834
##
## (Intercept)                             ***
## Processing.MethodOther                  ***
## Processing.MethodPulped natural / honey
## Processing.MethodWashed / Wet
## ColorBluish-Green
## ColorGreen                              *
## country_groupAsia
## country_groupSouth America
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.052 on 122 degrees of freedom
## Multiple R-squared:  0.2343, Adjusted R-squared:  0.1903
## F-statistic: 5.332 on 7 and 122 DF,  p-value: 2.418e-05
```

Gamma

Gamma inverse link

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following objects are masked from 'package:openintro':
##
##     housing, mammals

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

```
gamma.inverse <- glm(Total.Cup.Points ~ . - Flavor - Cupper.Points -
    Aroma - Aftertaste - Body - Acidity - Balance - Clean.Cup -
    Sweetness - Uniformity, family = Gamma(link = "inverse"),  data = coffee.new)
summary(gamma.inverse)
```

```
##
## Call:
```

```
## glm(formula = Total.Cup.Points ~ . - Flavor - Cupper.Points - 
##     Aroma - Aftertaste - Body - Acidity - Balance - Clean.Cup - 
##     Sweetness - Uniformity, family = Gamma(link = "inverse"), 
##     data = coffee.new)
## 
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              1.168e-02  2.354e-04  49.597  < 2e-16
## Number.of.Bags                          -5.019e-08  2.991e-07  -0.168   0.8670
## Processing.MethodOther                   7.932e-04  1.653e-04   4.799 4.76e-06
## Processing.MethodPulped natural / honey -8.204e-05  1.651e-04  -0.497   0.6201
## Processing.MethodWashed / Wet            2.354e-05  7.804e-05   0.302   0.7634
## Moisture                                 3.980e-04  7.578e-04   0.525   0.6004
## Category.One.Defects                     6.462e-06  1.052e-05   0.614   0.5402
## Quakers                                  7.490e-07  2.263e-05   0.033   0.9737
## ColorBluish-Green                        1.866e-04  1.391e-04   1.341   0.1824
## ColorGreen                               2.237e-04  1.113e-04   2.009   0.0468
## Category.Two.Defects                     8.942e-06  1.355e-05   0.660   0.5106
## country_groupAsia                        4.383e-05  1.900e-04   0.231   0.8180
## country_groupSouth America               1.973e-04  1.465e-04   1.347   0.1804
##                                            
## (Intercept)                             ***
## Number.of.Bags                             
## Processing.MethodOther                  ***
## Processing.MethodPulped natural / honey    
## Processing.MethodWashed / Wet              
## Moisture                                   
## Category.One.Defects                       
## Quakers                                    
## ColorBluish-Green                          
## ColorGreen                              *  
## Category.Two.Defects                       
## country_groupAsia                          
## country_groupSouth America                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Gamma family taken to be 0.0006862421)
## 
##     Null deviance: 0.109580  on 129  degrees of freedom
## Residual deviance: 0.085124  on 117  degrees of freedom
## AIC: 590.6
## 
## Number of Fisher Scoring iterations: 4
```

```
gamma.inverse.back <- step(gamma.inverse, direction='backward')
```

```
## Start:  AIC=590.6
## Total.Cup.Points ~ (Number.of.Bags + Processing.Method + Aroma + 
##     Flavor + Aftertaste + Acidity + Body + Balance + Uniformity + 
##     Clean.Cup + Sweetness + Cupper.Points + Moisture + Category.One.Defects + 
##     Quakers + Color + Category.Two.Defects + country_group) - 
##     Flavor - Cupper.Points - Aroma - Aftertaste - Body - Acidity - 
##     Balance - Clean.Cup - Sweetness - Uniformity
```

```
##
##                         Df Deviance    AIC
## - Quakers               1 0.085125 588.60
## - Number.of.Bags        1 0.085144 588.63
## - Moisture              1 0.085313 588.87
## - Category.One.Defects  1 0.085386 588.98
## - Category.Two.Defects  1 0.085423 589.03
## - country_group         2 0.087686 590.33
## <none>                    0.085124 590.60
## - Color                 2 0.087908 590.65
## - Processing.Method     3 0.102304 609.63
##
## Step:  AIC=588.6
## Total.Cup.Points ~ Number.of.Bags + Processing.Method + Moisture +
##     Category.One.Defects + Color + Category.Two.Defects + country_group
##
##                         Df Deviance    AIC
## - Number.of.Bags        1 0.085145 586.63
## - Moisture              1 0.085325 586.89
## - Category.One.Defects  1 0.085386 586.98
## - Category.Two.Defects  1 0.085445 587.07
## - country_group         2 0.087698 588.38
## <none>                    0.085125 588.60
## - Color                 2 0.087915 588.70
## - Processing.Method     3 0.102317 607.86
##
## Step:  AIC=586.63
## Total.Cup.Points ~ Processing.Method + Moisture + Category.One.Defects +
##     Color + Category.Two.Defects + country_group
##
##                         Df Deviance    AIC
## - Moisture              1 0.085349 584.93
## - Category.One.Defects  1 0.085412 585.02
## - Category.Two.Defects  1 0.085446 585.08
## <none>                    0.085145 586.63
## - country_group         2 0.087890 586.69
## - Color                 2 0.088174 587.11
## - Processing.Method     3 0.102849 606.84
##
## Step:  AIC=584.94
## Total.Cup.Points ~ Processing.Method + Category.One.Defects +
##     Color + Category.Two.Defects + country_group
##
##                         Df Deviance    AIC
## - Category.Two.Defects  1 0.085627 583.35
## - Category.One.Defects  1 0.085643 583.38
## <none>                    0.085349 584.94
## - country_group         2 0.088211 585.21
## - Color                 2 0.088548 585.71
## - Processing.Method     3 0.103384 605.81
##
## Step:  AIC=583.36
## Total.Cup.Points ~ Processing.Method + Category.One.Defects +
##     Color + country_group
```

```
##
##                         Df Deviance    AIC
## - Category.One.Defects  1 0.085895 581.76
## <none>                    0.085627 583.36
## - country_group         2 0.088879 584.24
## - Color                 2 0.088995 584.41
## - Processing.Method     3 0.104160 605.13
##
## Step:  AIC=581.77
## Total.Cup.Points ~ Processing.Method + Color + country_group
##
##                      Df Deviance    AIC
## <none>                 0.085895 581.77
## - country_group      2 0.088909 582.31
## - Color              2 0.089399 583.05
## - Processing.Method  3 0.104271 603.45
```

“Best” gamma model with inverse link (the one with lowest AIC)

```
gamma.best.inverse <- glm(Total.Cup.Points ~ country_group + Color + Processing.Method, data = coffee.ne
```

Gamma log link

```
gamma.log <- glm(Total.Cup.Points ~ . - Flavor - Cupper.Points -
    Aroma - Aftertaste - Body - Acidity - Balance - Clean.Cup -
    Sweetness - Uniformity, family = Gamma(link = "log"),  data = coffee.new)
summary(gamma.log)
```

```
##
## Call:
## glm(formula = Total.Cup.Points ~ . - Flavor - Cupper.Points -
##     Aroma - Aftertaste - Body - Acidity - Balance - Clean.Cup -
##     Sweetness - Uniformity, family = Gamma(link = "log"), data = coffee.new)
##
## Coefficients:
##                                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               4.450e+00  1.945e-02 228.785  < 2e-16
## Number.of.Bags                            2.904e-06  2.435e-05   0.119   0.9053
## Processing.MethodOther                   -6.442e-02  1.299e-02  -4.961 2.41e-06
## Processing.MethodPulped natural / honey   6.501e-03  1.374e-02   0.473   0.6370
## Processing.MethodWashed / Wet            -2.067e-03  6.410e-03  -0.322   0.7477
## Moisture                                 -3.398e-02  6.258e-02  -0.543   0.5881
## Category.One.Defects                     -5.572e-04  8.635e-04  -0.645   0.5200
## Quakers                                   4.727e-06  1.858e-03   0.003   0.9980
## ColorBluish-Green                        -1.573e-02  1.156e-02  -1.362   0.1759
## ColorGreen                               -1.885e-02  9.269e-03  -2.033   0.0443
## Category.Two.Defects                     -7.515e-04  1.112e-03  -0.676   0.5003
## country_groupAsia                        -3.549e-03  1.575e-02  -0.225   0.8222
## country_groupSouth America               -1.643e-02  1.218e-02  -1.348   0.1802
##
## (Intercept)                             ***
## Number.of.Bags
## Processing.MethodOther                  ***
## Processing.MethodPulped natural / honey
## Processing.MethodWashed / Wet
## Moisture
## Category.One.Defects
## Quakers
## ColorBluish-Green
## ColorGreen                              *
## Category.Two.Defects
## country_groupAsia
## country_groupSouth America
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for Gamma family taken to be 0.0006842825)
##
##     Null deviance: 0.109580  on 129  degrees of freedom
## Residual deviance: 0.084778  on 117  degrees of freedom
## AIC: 590.07
##
## Number of Fisher Scoring iterations: 4
```

```r
gamma.log.backward <- step(gamma.log, direction= 'backward')
```

```
## Start:  AIC=590.07
## Total.Cup.Points ~ (Number.of.Bags + Processing.Method + Aroma +
##     Flavor + Aftertaste + Acidity + Body + Balance + Uniformity +
##     Clean.Cup + Sweetness + Cupper.Points + Moisture + Category.One.Defects +
##     Quakers + Color + Category.Two.Defects + country_group) -
##     Flavor - Cupper.Points - Aroma - Aftertaste - Body - Acidity -
##     Balance - Clean.Cup - Sweetness - Uniformity
##
##                        Df Deviance    AIC
## - Quakers               1 0.084778 588.07
## - Number.of.Bags        1 0.084788 588.08
## - Moisture              1 0.084979 588.36
## - Category.One.Defects  1 0.085064 588.49
## - Category.Two.Defects  1 0.085090 588.52
## - country_group         2 0.087397 589.89
## <none>                    0.084778 590.07
## - Color                 2 0.087656 590.27
## - Processing.Method     3 0.102289 609.66
##
## Step:  AIC=588.07
## Total.Cup.Points ~ Number.of.Bags + Processing.Method + Moisture +
##     Category.One.Defects + Color + Category.Two.Defects + country_group
##
##                        Df Deviance    AIC
## - Number.of.Bags        1 0.084788 586.08
## - Moisture              1 0.084985 586.37
## - Category.One.Defects  1 0.085064 586.49
## - Category.Two.Defects  1 0.085105 586.55
## - country_group         2 0.087403 587.94
## <none>                    0.084778 588.07
## - Color                 2 0.087658 588.31
## - Processing.Method     3 0.102303 607.90
##
## Step:  AIC=586.08
## Total.Cup.Points ~ Processing.Method + Moisture + Category.One.Defects +
##     Color + Category.Two.Defects + country_group
##
##                         Df Deviance    AIC
## - Moisture              1 0.084998 584.39
## - Category.One.Defects  1 0.085079 584.51
## - Category.Two.Defects  1 0.085112 584.56
## <none>                    0.084788 586.08
## - country_group         2 0.087640 586.32
## - Color                 2 0.087870 586.66
## - Processing.Method     3 0.102854 606.92
##
## Step:  AIC=584.4
## Total.Cup.Points ~ Processing.Method + Category.One.Defects +
##     Color + Category.Two.Defects + country_group
##
##                         Df Deviance    AIC
## - Category.Two.Defects  1 0.085298 582.85
## - Category.One.Defects  1 0.085318 582.88
## <none>                    0.084998 584.40
## - country_group         2 0.087982 584.87
## - Color                 2 0.088280 585.31
## - Processing.Method     3 0.103382 605.88
##
## Step:  AIC=582.86
## Total.Cup.Points ~ Processing.Method + Category.One.Defects +
##     Color + country_group
##
##                         Df Deviance    AIC
## - Category.One.Defects  1 0.085591 581.30
## <none>                    0.085298 582.86
## - country_group         2 0.088708 583.99
## - Color                 2 0.088748 584.05
## - Processing.Method     3 0.104153 605.20
##
## Step:  AIC=581.31
## Total.Cup.Points ~ Processing.Method + Color + country_group
##
##                      Df Deviance    AIC
## <none>                 0.085591 581.31
## - country_group      2 0.088742 582.07
## - Color              2 0.089184 582.74
## - Processing.Method  3 0.104266 603.52
```

“Best” gamma log

```
gamma.best.log <- glm(Total.Cup.Points ~ country_group + Color + Processing.Method, data = coffee.new,
```

Gamma identity

```
gamma.identity <- glm(Total.Cup.Points ~ . - Flavor - Cupper.Points - Aroma - Aftertaste - Body - Acidi
summary(gamma.identity)
```

```
##
## Call:
## glm(formula = Total.Cup.Points ~ . - Flavor - Cupper.Points -
##     Aroma - Aftertaste - Body - Acidity - Balance - Clean.Cup -
##     Sweetness - Uniformity, family = Gamma(link = "identity"),
##     data = coffee.new)
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          85.690852   1.606864  53.328  < 2e-16
## Number.of.Bags                        0.000128   0.001978   0.065    0.949
## Processing.MethodOther               -5.235370   1.019400  -5.136 1.13e-06
## Processing.MethodPulped natural / honey  0.512817   1.143289   0.449    0.655
## Processing.MethodWashed / Wet        -0.181949   0.526279  -0.346    0.730
## Moisture                             -2.906229   5.167422  -0.562    0.575
## Category.One.Defects                 -0.048118   0.070833  -0.679    0.498
## Quakers                               0.006219   0.152486   0.041    0.968
## ColorBluish-Green                    -1.325588   0.959922  -1.381    0.170
## ColorGreen                           -1.586753   0.771630  -2.056    0.042
## Category.Two.Defects                 -0.063328   0.091138  -0.695    0.489
## country_groupAsia                    -0.287921   1.305546  -0.221    0.826
## country_groupSouth America           -1.367713   1.013634  -1.349    0.180
##
## (Intercept)                          ***
## Number.of.Bags
## Processing.MethodOther               ***
## Processing.MethodPulped natural / honey
## Processing.MethodWashed / Wet
## Moisture
## Category.One.Defects
## Quakers
## ColorBluish-Green
## ColorGreen                           *
## Category.Two.Defects
## country_groupAsia
## country_groupSouth America
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.0006822081)
##
##     Null deviance: 0.109580  on 129  degrees of freedom
## Residual deviance: 0.084402  on 117  degrees of freedom
## AIC: 589.49
##
## Number of Fisher Scoring iterations: 5
```

```r
gamma.identity.backward <- step(gamma.identity, direction = 'backward')
```

```
## Start:  AIC=589.49
## Total.Cup.Points ~ (Number.of.Bags + Processing.Method + Aroma + 
##     Flavor + Aftertaste + Acidity + Body + Balance + Uniformity + 
##     Clean.Cup + Sweetness + Cupper.Points + Moisture + Category.One.Defects + 
##     Quakers + Color + Category.Two.Defects + country_group) - 
##     Flavor - Cupper.Points - Aroma - Aftertaste - Body - Acidity - 
##     Balance - Clean.Cup - Sweetness - Uniformity
## 
##                        Df Deviance    AIC
## - Quakers               1 0.084403 587.49
## - Number.of.Bags        1 0.084405 587.49
## - Moisture              1 0.084617 587.80
## - Category.One.Defects  1 0.084716 587.95
## - Category.Two.Defects  1 0.084730 587.97
## - country_group         2 0.087080 589.42
## <none>                    0.084402 589.49
## - Color                 2 0.087373 589.84
## - Processing.Method     3 0.102273 609.68
## 
## Step:  AIC=587.49
## Total.Cup.Points ~ Number.of.Bags + Processing.Method + Moisture + 
##     Category.One.Defects + Color + Category.Two.Defects + country_group
## 
##                        Df Deviance    AIC
## - Number.of.Bags        1 0.084406 585.50
## - Moisture              1 0.084619 585.81
## - Category.One.Defects  1 0.084719 585.96
## - Category.Two.Defects  1 0.084739 585.99
## - country_group         2 0.087083 587.45
## <none>                    0.084403 587.49
## - Color                 2 0.087373 587.88
## - Processing.Method     3 0.102288 607.93
## 
## Step:  AIC=585.5
## Total.Cup.Points ~ Processing.Method + Moisture + Category.One.Defects + 
##     Color + Category.Two.Defects + country_group
## 
##                        Df Deviance    AIC
## - Moisture              1 0.084623 583.82
## - Category.One.Defects  1 0.084724 583.97
## - Category.Two.Defects  1 0.084758 584.02
## <none>                    0.084406 585.50
## - country_group         2 0.087372 585.92
## - Color                 2 0.087536 586.16
## - Processing.Method     3 0.102860 607.00
## 
## Step:  AIC=583.83
## Total.Cup.Points ~ Processing.Method + Category.One.Defects + 
##     Color + Category.Two.Defects + country_group
## 
##                        Df Deviance    AIC
## - Category.Two.Defects  1 0.084950 582.32
## - Category.One.Defects  1 0.084972 582.35
## <none>                    0.084623 583.83
## - country_group         2 0.087737 584.50
## - Color                 2 0.087987 584.87
## - Processing.Method     3 0.103381 605.96
##
## Step:  AIC=582.33
## Total.Cup.Points ~ Processing.Method + Category.One.Defects +
##     Color + country_group
##
##                          Df Deviance    AIC
## - Category.One.Defects  1 0.085269 580.81
## <none>                    0.084950 582.33
## - Color                 2 0.088479 583.65
## - country_group         2 0.088528 583.73
## - Processing.Method     3 0.104146 605.27
##
## Step:  AIC=580.82
## Total.Cup.Points ~ Processing.Method + Color + country_group
##
##                     Df Deviance    AIC
## <none>                 0.085269 580.82
## - country_group      2 0.088567 581.81
## - Color              2 0.088949 582.39
## - Processing.Method  3 0.104260 603.59
```

"Best" identity link

```
gamma.best.identity <- glm(Total.Cup.Points ~ Processing.Method + Color + country_group, data = coffee.n
```

Compare MSE and MAE for gamma and linear

```
coffee.new.data <- coffee.new %>% mutate(predict.inverse = gamma.best.inverse$fitted.values,
                                    predict.identity = gamma.best.identity$fitted.values,
                                    predict.log = gamma.best.log$fitted.values,
                                    predict.linear = best.linear.model$fitted.values)


coffee.new.data %>% summarize(MSE.inverse = mean((Total.Cup.Points - predict.inverse)^2),
                         MSE.log = mean((Total.Cup.Points - predict.log)^2),
                         MSE.identity = mean((Total.Cup.Points - predict.identity)^2),
                         MSE.linear = mean((Total.Cup.Points -predict.linear)^2))
```

```
## # A tibble: 1 x 4
##   MSE.inverse MSE.log MSE.identity MSE.linear
##         <dbl>   <dbl>        <dbl>      <dbl>
## 1        3.98    3.97         3.95       3.95
```

```
coffee.new.data %>% summarize(MAE.inverse = mean(abs(Total.Cup.Points - predict.inverse)),
                         MAE.log = mean(abs(Total.Cup.Points - predict.log)),
                         MAE.identity = mean(abs(Total.Cup.Points - predict.identity)),
                         MAE.linear = mean(abs(Total.Cup.Points-predict.linear)))
```

```
## # A tibble: 1 x 4
##   MAE.inverse MAE.log MAE.identity MAE.linear
##         <dbl>   <dbl>        <dbl>      <dbl>
## 1        1.26    1.26         1.26       1.26
```

```
AIC(gamma.best.inverse)
```

```
## [1] 581.7692
```

```
AIC(gamma.best.log)
```

```
## [1] 581.3076
```

```
AIC(gamma.best.identity)
```

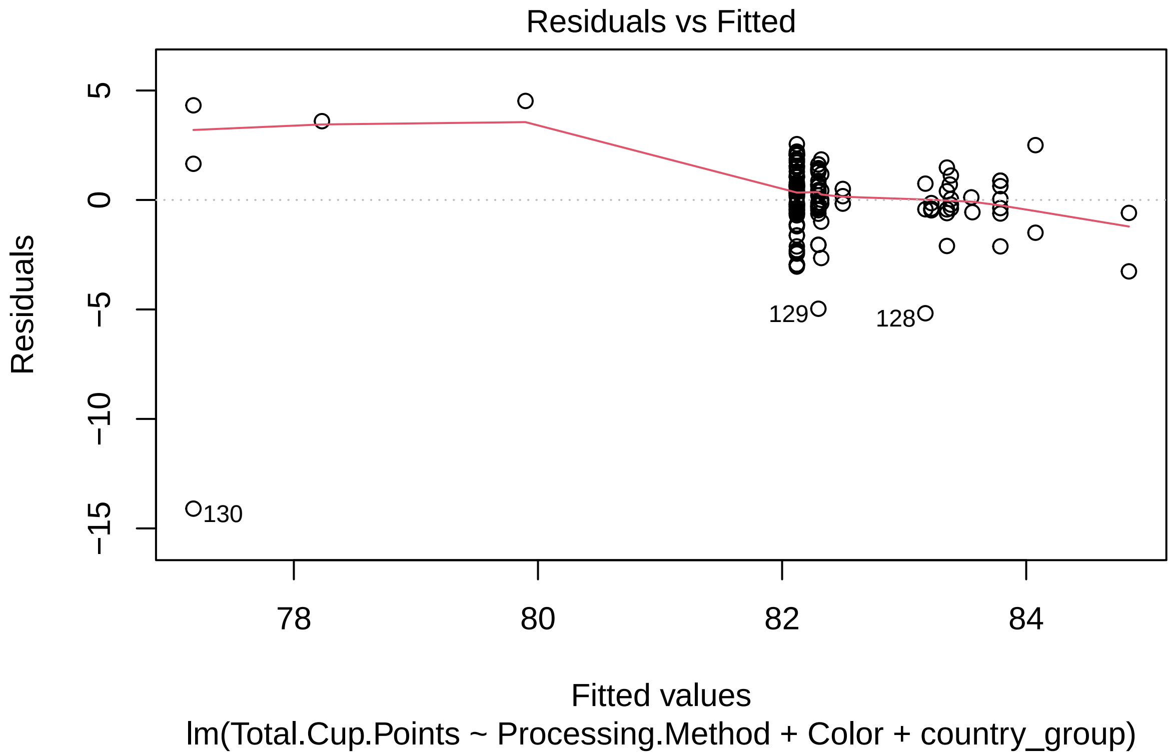
```
## [1] 580.8185
```

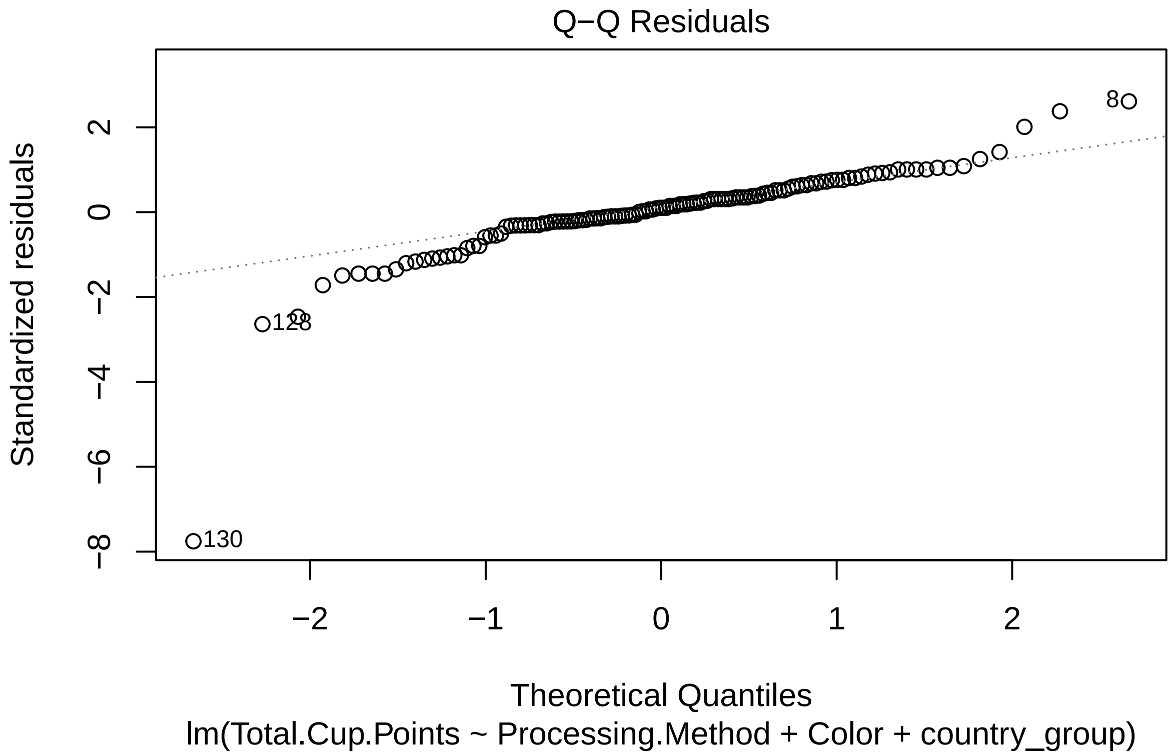
```
AIC(best.linear.model)
```

```
## [1] 565.4986
```

Linear model assumption check

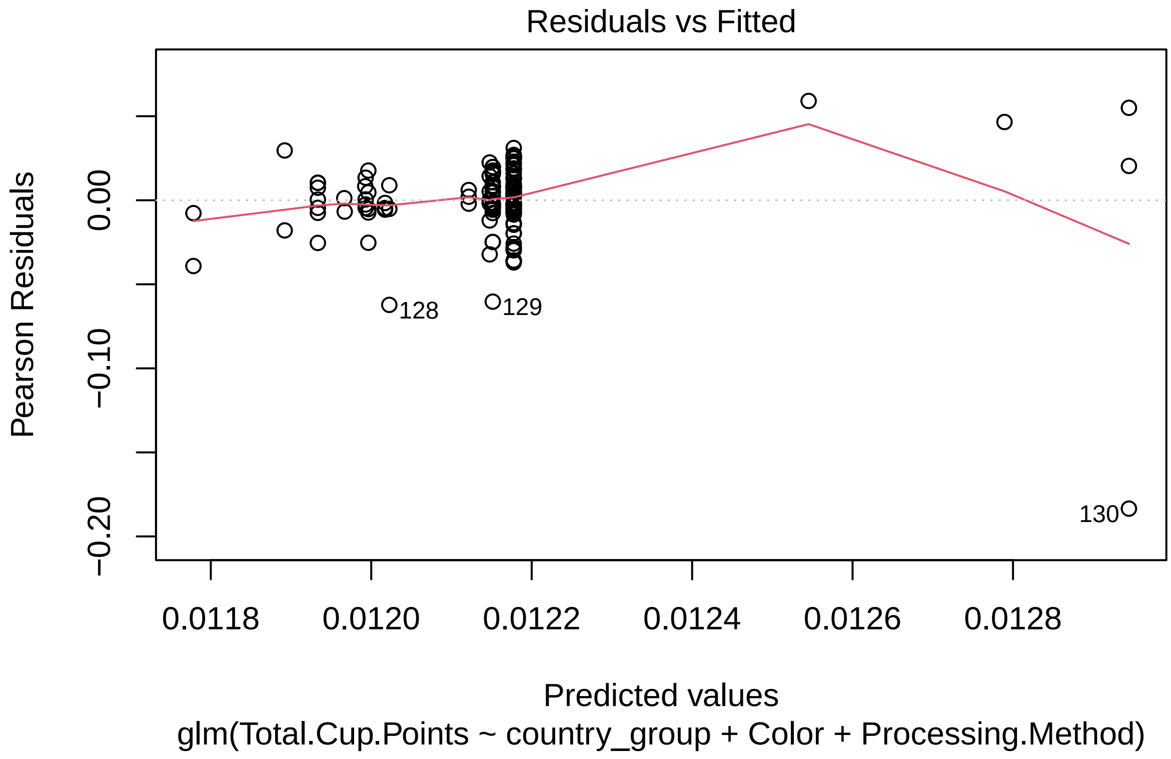
```
plot(best.linear.model, which =c(1,2))
```

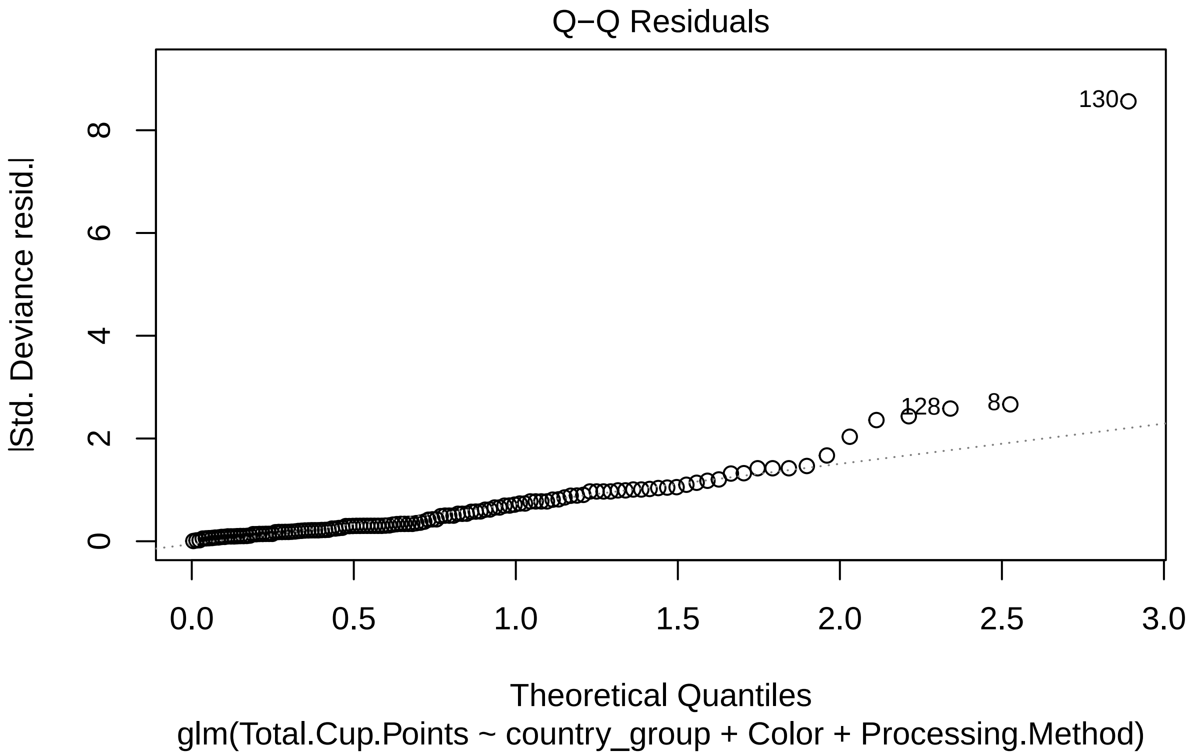

Residuals vs Fitted
Residuals
5
0
−5
−10
−15
78
80
82
84
Fitted values
lm(Total.Cup.Points ~ Processing.Method + Color + country_group)
130
129
128



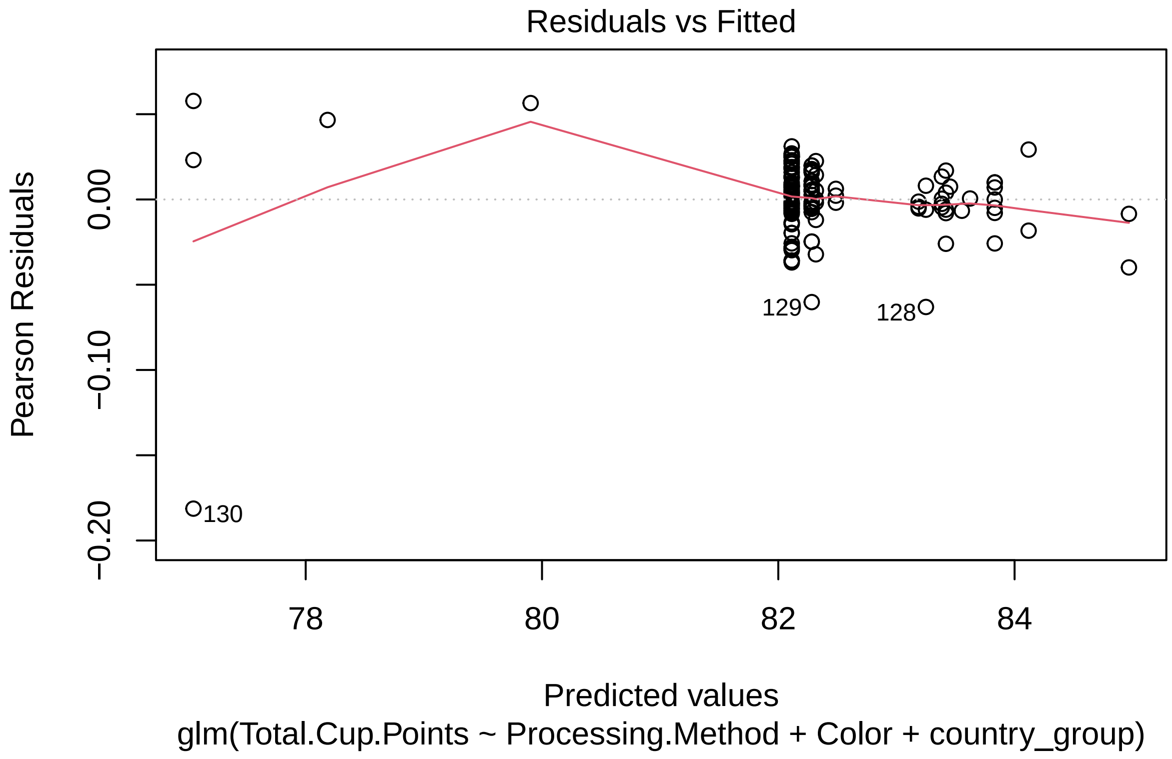
Gamma model assumption check

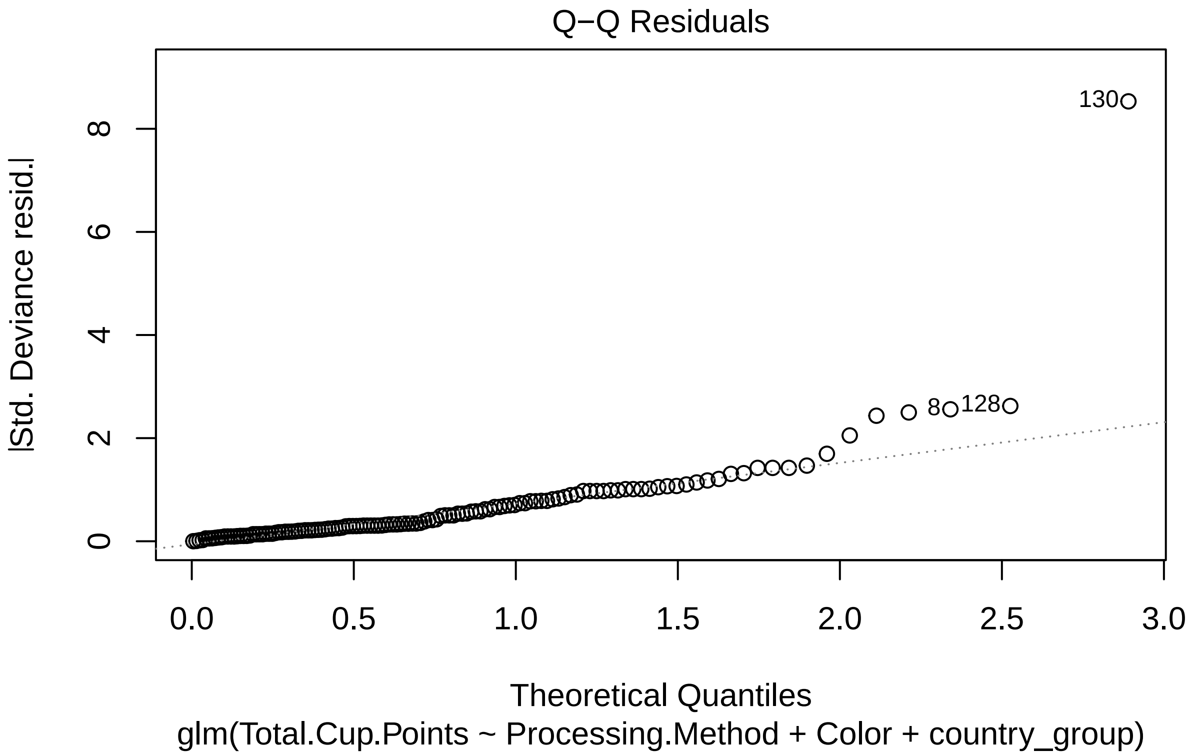
```
plot(gamma.best.inverse, which = c(1,2))
```


Residuals vs Fitted
Pearson Residuals
0.00
−0.10
−0.20
0.0118
0.0120
0.0122
0.0124
0.0126
0.0128
128
129
130
Predicted values
glm(Total.Cup.Points ~ country_group + Color + Processing.Method)

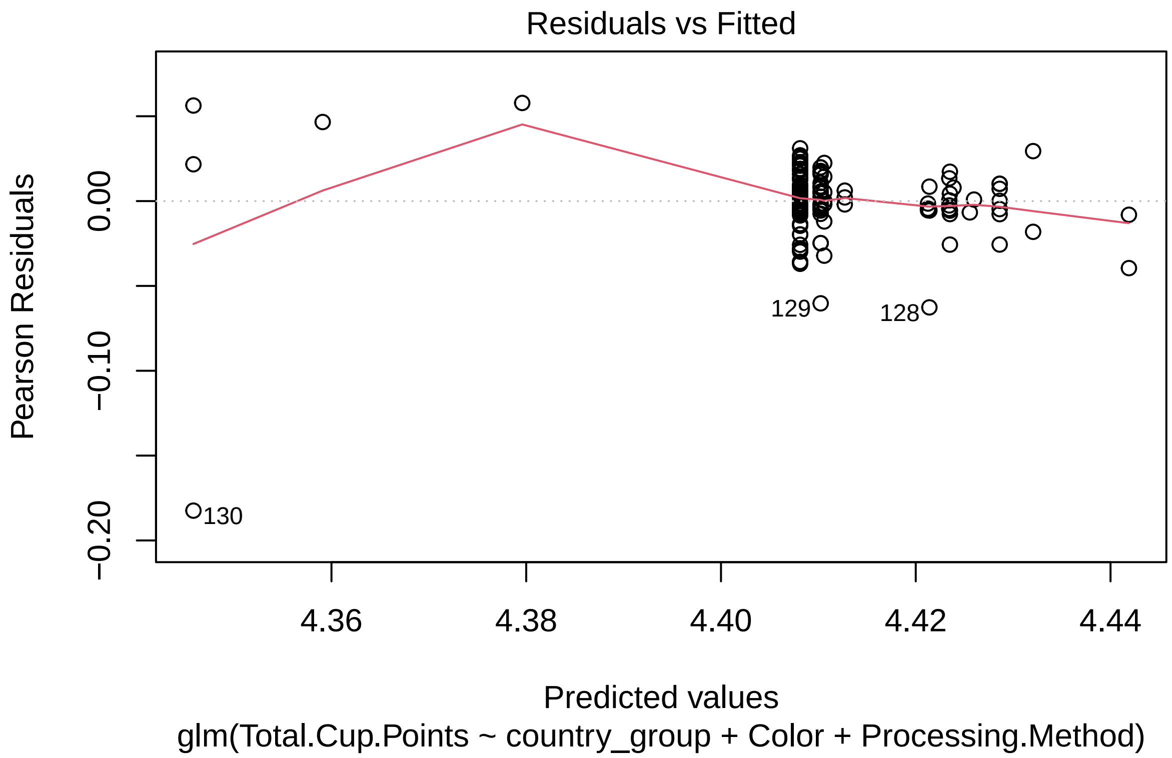


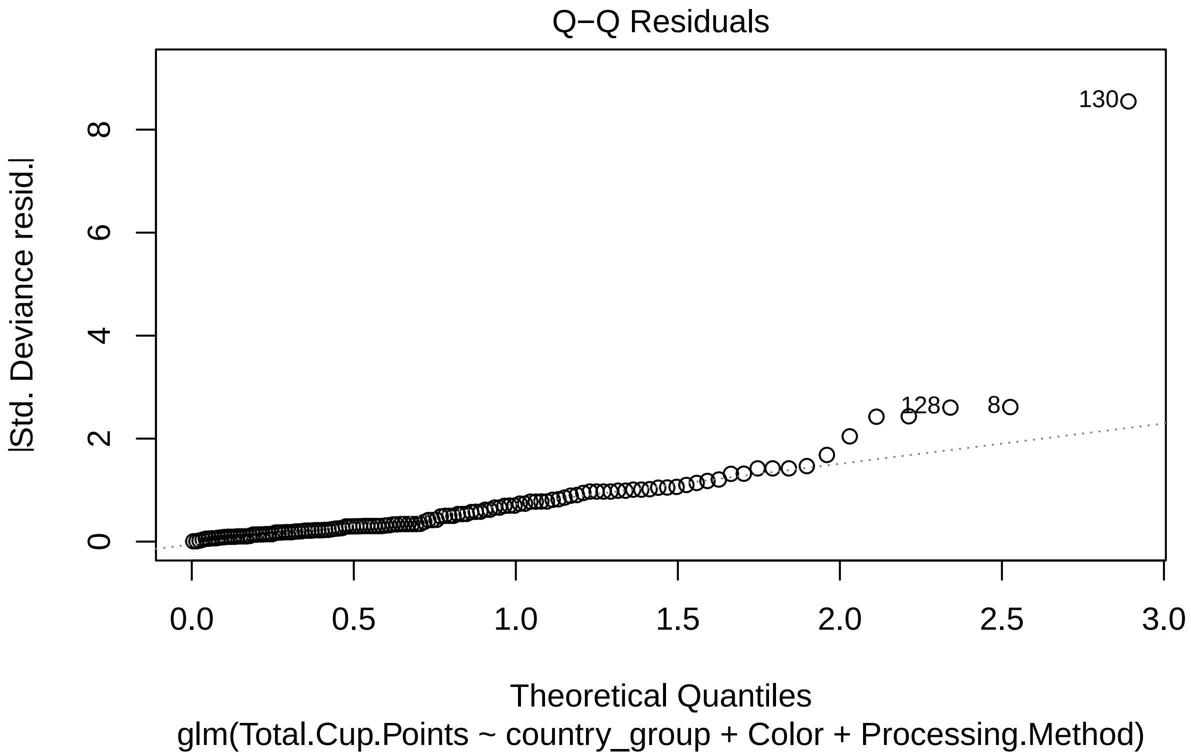
```
plot(gamma.best.identity, which =c(1,2))
```


Residuals vs Fitted
Pearson Residuals
0.00
−0.10
−0.20
78
80
82
84
130
129
128
Predicted values
glm(Total.Cup.Points ~ Processing.Method + Color + country_group)



```r
plot(gamma.best.log, which =c(1,2))
```


Residuals vs Fitted
Pearson Residuals
0.00
−0.10
−0.20
4.36
4.38
4.40
4.42
4.44
129
128
130
Predicted values
glm(Total.Cup.Points ~ country_group + Color + Processing.Method)



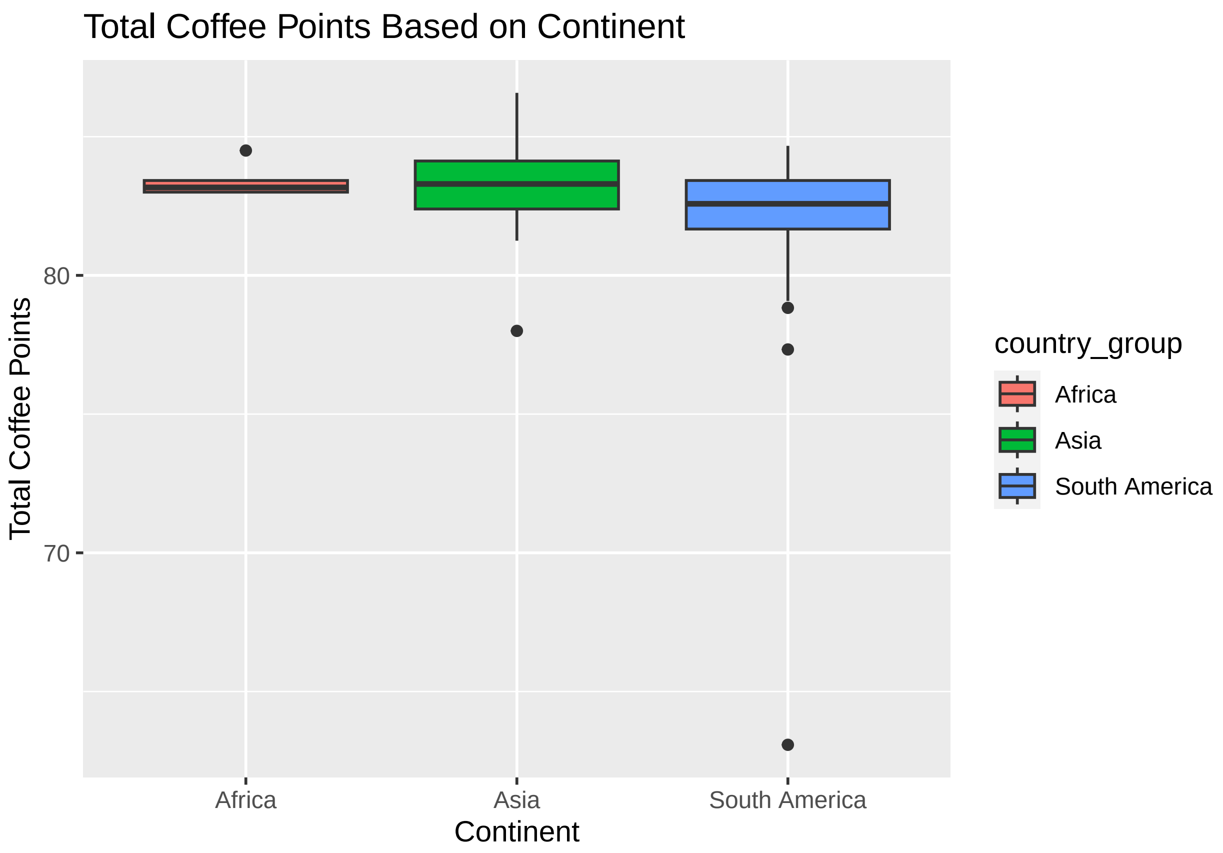
For all of the gamma Q-Q and residuals vs. fitted plots, they are nearly identical to one another. For the Q-Q plots, they follow a very straight line. Because of this, they all have evidence for normality. However, the residuals vs. fitted plots are not randomly distributed across the horizontal axis at all. There is not enough evidence to claim linearity for the gamma models.

Among the gamma models, the gamma model using the identity log function appears to be the best by checking its AIC value. The AIC is 580.82, which is marginally lower than the two other gamma models, making it the best option.

Gamma Analysis

Country Regions Plot

```
ggplot(data = coffee.new, aes(x = country_group, y = Total.Cup.Points, fill = country_group)) + geom_bo
```


Total Coffee Points Based on Continent
Total Coffee Points
80
70
Africa
Asia
South America
Continent
country_group
Africa
Asia
South America