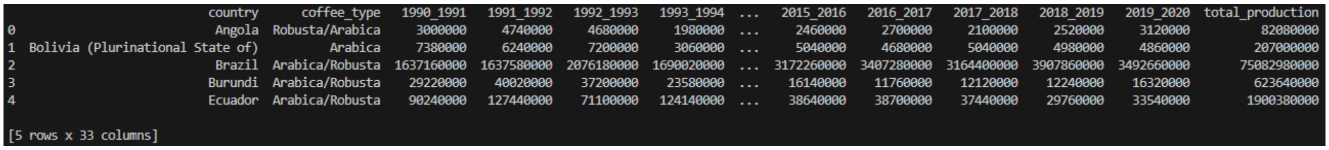
Keiffer Tan

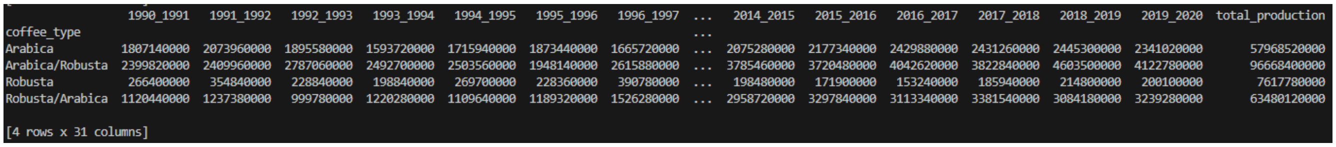
Coffee Production Analysis

Data Aggregation



```
                        country      coffee_type   1990_1991   1991_1992   1992_1993   1993_1994  ...   2015_2016   2016_2017   2017_2018   2018_2019   2019_2020  total_production
0                        Angola  Robusta/Arabica     3000000     4740000     4680000     1980000  ...     2460000     2700000     2100000     2520000     3120000          82080000
1  Bolivia (Plurinational State of)        Arabica     7380000     6240000     7200000     3060000  ...     5040000     4680000     5040000     4980000     4860000         207000000
2                        Brazil  Arabica/Robusta  1637160000  1637580000  2076180000  1690020000  ...  3172260000  3407280000  3164400000  3907860000  3492660000       75082980000
3                       Burundi  Arabica/Robusta    29220000    40020000    37200000    23580000  ...    16140000    11760000    12120000    12240000    16320000         623640000
4                       Ecuador  Arabica/Robusta    90240000   127440000    71100000   124140000  ...    38640000    38700000    37440000    29760000    33540000        1900380000

[5 rows x 33 columns]
```

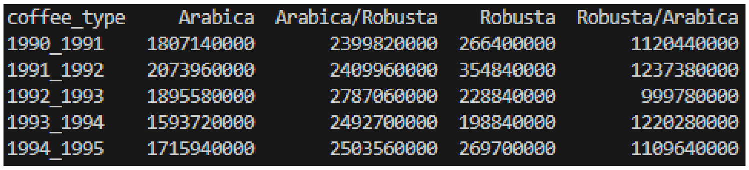


```
                  1990_1991   1991_1992   1992_1993   1993_1994   1994_1995   1995_1996   1996_1997  ...   2014_2015   2015_2016   2016_2017   2017_2018   2018_2019   2019_2020  total_production
coffee_type                                                                                          ...
Arabica          1807140000  2073960000  1895580000  1593720000  1715940000  1873440000  1665720000  ...  2075280000  2177340000  2429880000  2431260000  2445300000  2341020000       57968520000
Arabica/Robusta  2399820000  2409960000  2787060000  2492700000  2503560000  1948140000  2615880000  ...  3785460000  3720480000  4042620000  3822840000  4603500000  4122780000       96668400000
Robusta           266400000   354840000   228840000   198840000   269700000   228360000   390780000  ...   198480000   171900000   153240000   185940000   214800000   200100000        7617780000
Robusta/Arabica  1120440000  1237380000   999780000  1220280000  1109640000  1189320000  1526280000  ...  2958720000  3297840000  3113340000  3381540000  3084180000  3239280000       63480120000

[4 rows x 31 columns]
```

- By using groupby[(‘coffee_type’]) combined with .sum() I am able to aggregate the amount of coffee produced for each coffee type by year.
- I had also dropped the Country column since it is not relevant currently

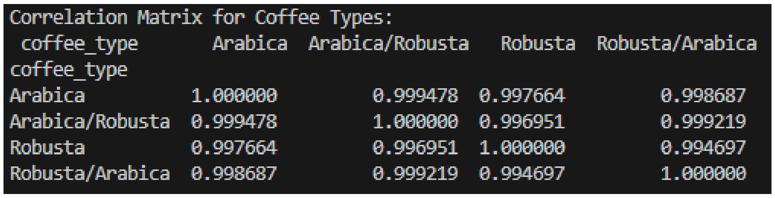
Data Transformation



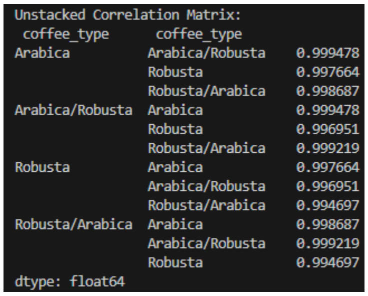
```
coffee_type     Arabica  Arabica/Robusta    Robusta  Robusta/Arabica
1990_1991    1807140000       2399820000  266400000       1120440000
1991_1992    2073960000       2409960000  354840000       1237380000
1992_1993    1895580000       2787060000  228840000        999780000
1993_1994    1593720000       2492700000  198840000       1220280000
1994_1995    1715940000       2503560000  269700000       1109640000
```

- To transpose the data, I just used the .transpose() function which easily moved year as the index and type of coffee to columns

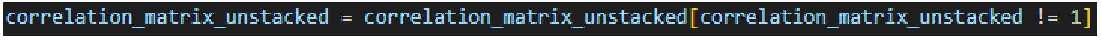
Correlation Analysis



```
Correlation Matrix for Coffee Types:
 coffee_type       Arabica  Arabica/Robusta   Robusta  Robusta/Arabica
coffee_type
Arabica          1.000000         0.999478  0.997664         0.998687
Arabica/Robusta  0.999478         1.000000  0.996951         0.999219
Robusta          0.997664         0.996951  1.000000         0.994697
Robusta/Arabica  0.998687         0.999219  0.994697         1.000000
```



```
Unstacked Correlation Matrix:
 coffee_type      coffee_type
Arabica          Arabica/Robusta    0.999478
                 Robusta            0.997664
                 Robusta/Arabica    0.998687
Arabica/Robusta  Arabica            0.999478
                 Robusta            0.996951
                 Robusta/Arabica    0.999219
Robusta          Arabica            0.997664
                 Arabica/Robusta    0.996951
                 Robusta/Arabica    0.994697
Robusta/Arabica  Arabica            0.998687
                 Arabica/Robusta    0.999219
                 Robusta            0.994697
dtype: float64
```

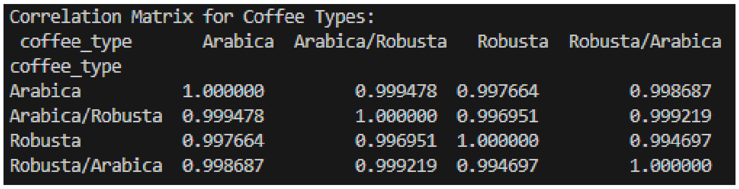
- For correlation analysis, I used the .corr() function on the cleaned and transposed dataframe which returned the first image above.
- By using .abs().unstack() I am able to retrieve the dataframe as a series with multiple indexes.
- I further cleaned the correlation analysis by removing the self-correlation values
  - 

```
correlation_matrix_unstacked = correlation_matrix_unstacked[correlation_matrix_unstacked != 1]
```

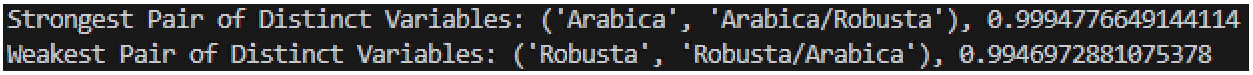
Questions

Keiffer Tan

1. Examine the correlation matrix. Which two coffee types have the **strongest** correlation in production volumes over the years? What might this imply about their production dynamics?
2. Identify the two coffee types with the **weakest** correlation. Discuss possible reasons for this weak relationship and any external factors that might influence these production types differently.



```
Correlation Matrix for Coffee Types:
 coffee_type       Arabica  Arabica/Robusta   Robusta  Robusta/Arabica
coffee_type
Arabica          1.000000         0.999478  0.997664         0.998687
Arabica/Robusta  0.999478         1.000000  0.996951         0.999219
Robusta          0.997664         0.996951  1.000000         0.994697
Robusta/Arabica  0.998687         0.999219  0.994697         1.000000
```



```
Strongest Pair of Distinct Variables: ('Arabica', 'Arabica/Robusta'), 0.9994776649144114
Weakest Pair of Distinct Variables: ('Robusta', 'Robusta/Arabica'), 0.9946972881075378
```

- This is interesting because even though there is a strongest and weakest pair of correlations, all of them are very high at 0.99 on a scale from 0 – 1. The correlations in production years are possibly from the fact that Arabica and Robusta coffee beans are produced in the same regions which could closely align their demands.
- Also, that Arabica and Arabica/Robusta blends can have similar demands due to Arabica still being in the blend. This can also be true for Robusta and Robusta/Arabica blends as well
- Going back to the initial point, there is the “weakest pair” in relation to the strongest pair, but in general it still holds a very strong correlation.