Java Metrics:

Brewing the Numbers Behind U.S. Coffee Imports

Submitted by:

Samantha Doyle

STAT 7100

November 22, 2024

Abstract

COVID-19 changed many aspects of everyday life. As time goes on after COVID, researchers are looking to see if COVID had any long-lasting effects. One area that might have potential impacts is importing goods into the U.S. Coffee is consumed by millions of Americans every day, but almost all the coffee Americans drink is imported from other countries. This report aims to compare the imported coffee in U.S.D. pre-COVID and post-COVID to see if there are lasting impacts from the 2020 pandemic on the importation of coffee products. The primary statistical methods used to investigate this question will be simple linear regression and ANOVA testing. Some t-Tests will be conducted, as well.

Table of Contents

Introduction

Coffee is such an important part of American culture. 73% of Americans drink coffee every day. While researchers determine if COVID had lasting effects on importation, the report aims to decide if there are effects over a 4-year period by comparing importation values in USD for 2019 (pre-COVID) and 2023 (post COVID) to see if there is a major difference in the two years. As imports increase or decrease, costs of coffee products for Americans changes as well. This has an impact on the daily life of many Americans.

Data Cleaning

The dataset used to compare this information is from the USDA Economic Research Service (ERS). The dataset covers imports of coffee, tea, and spice products into the U.S. from other countries. The original dataset was an Excel sheet that had pivot tables used in it. To import the data set into SAS without any issues, the data was copied and pasted into a new excel sheet. The years were also changed from 1999-2023 into y1999-y2023 to make proper variable names in SAS. Once imported into SAS, variable A was reprinted as product_type with every observation getting a product type. Variable B was renamed obs_ID for observation ID and variable C was renamed exporting country.

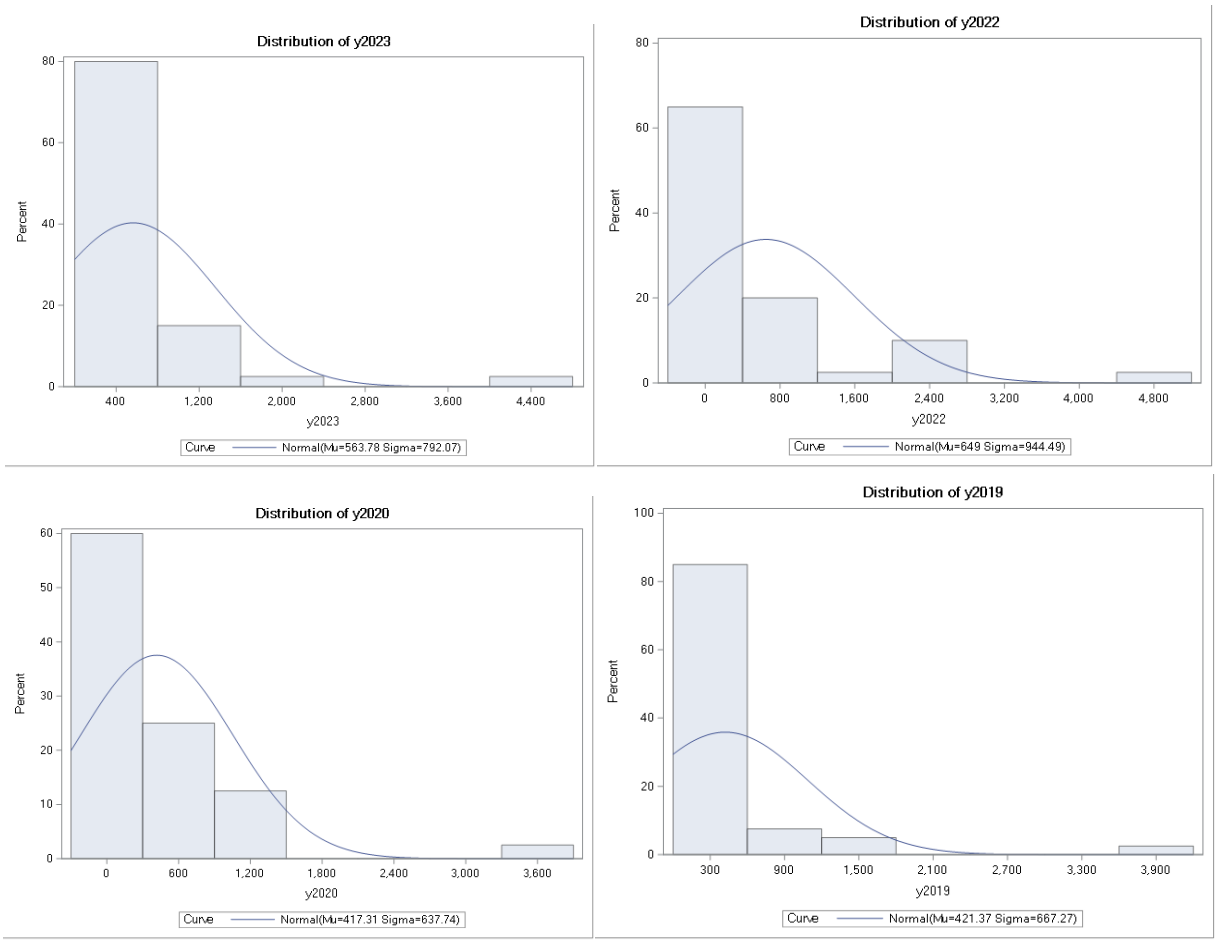
Once the variables were renamed, two types of observations were moved to two different datasets. If exporting_country was “World” it was moved to dataset *stat.worldcoffee*. This was done because these were total values which already included some of the other observation values. If exporting_country was “World (quantity)” it was moved to dataset *stat.coffee_worldquantity*. This was done because these values referred to the amount of product in weight and not in dollar amount. The resulting dataset, *stat.coffee*, after these changes had 40 observations. After these changes were made, all numeric values in the dataset were in millions of USD.

There are 32 variables in the data set. 4 categorical variables were made from the years 2019-2023. The values were separated into three categories, “Low,” “Moderate,” and “High,” using the upper and lower quartiles as boundaries. There are also two other categorical variables, exporting_country, the country the product is coming from, and product_type, the type of product being imported. An observation ID variable is also included. The remaining variables all represent a year from 1999 to 2023. These variables are numeric and represent the amount of product imported into the U.S. from other countries in millions of USD. Below is a codebook of the main dataset.

Variable Name	Label	Numeric or Categorical	Format
obs_ID	Observation ID	Numeric	# for observation from original dataset
exporting_country	Country of Export	Categorical	Brazil, Canda, China, Columbia, Guatemala, Honduras, India, Indonesia, Italy, Japan, Madagascar, Mexico, Nicaragua, Peru, Rest of World, Spain, Switzerland, Vietnam
y2023	Year 2023	Numeric	Import amounts in USD in million dollars
y2022	Year 2022	Numeric	Import amounts in USD in million dollars
y2021	Year 2021	Numeric	Import amounts in USD in million dollars
y2020	Year 2020	Numeric	Import amounts in USD in million dollars
y2019	Year 2019	Numeric	Import amounts in USD in million dollars
y2018	Year 2018	Numeric	Import amounts in USD in million dollars
y2017	Year 2017	Numeric	Import amounts in USD in million dollars
y2016	Year 2016	Numeric	Import amounts in USD in million dollars
y2015	Year 2015	Numeric	Import amounts in USD in million dollars
y2014	Year 2014	Numeric	Import amounts in USD in million dollars
y2013	Year 2013	Numeric	Import amounts in USD in million dollars
y2012	Year 2012	Numeric	Import amounts in USD in million dollars
y2011	Year 2011	Numeric	Import amounts in USD in million dollars
y2010	Year 2010	Numeric	Import amounts in USD in million dollars
y2009	Year 2009	Numeric	Import amounts in USD in million dollars
y2008	Year 2008	Numeric	Import amounts in USD in million dollars
y2007	Year 2007	Numeric	Import amounts in USD in million dollars
y2006	Year 2006	Numeric	Import amounts in USD in million dollars
y2005	Year 2005	Numeric	Import amounts in USD in million dollars
y2004	Year 2004	Numeric	Import amounts in USD in million dollars
y2003	Year 2003	Numeric	Import amounts in USD in million dollars
y2002	Year 2002	Numeric	Import amounts in USD in million dollars
y2001	Year 2001	Numeric	Import amounts in USD in million dollars
y2000	Year 2000	Numeric	Import amounts in USD in million dollars
y1999	Year 1999	Numeric	Import amounts in USD in million dollars
yr2023_levels	Year 2023 imports in 3 categorical levels	Categorical	"Low", "Moderate", "High"
yr2022_levels	Year 2022 imports in 3 categorical levels	Categorical	"Low", "Moderate", "High"
yr2020_levels	Year 2020 imports in 3 categorical levels	Categorical	"Low", "Moderate", "High"
yr2019_levels	Year 2019 imports in 3 categorical levels	Categorical	"Low", "Moderate", "High"

Data Analysis:

To begin to get a visual sense of the data, histograms were made for years 2023, 2022, 2020 and 2019.



All of the years have a right-skewed distribution. Then, descriptive statistics were found for all the numeric variables in the dataset. The values for lower quartile and upper quartile were used to as boundaries for the categorical variables: yr2019_levels, yr2002_levels, yr2022_levels and yr2023_levels.

DESCRIPTIVE STATISTICS

Variable	Label	N	N Miss	Mean	Median	Std Dev	Range	Lower Quartile	Upper Quartile
obs_ID	B	40	0	25.4500000	24.5000000	15.3755550	49.0000000	12.5000000	38.5000000
y2023	y2023	40	0	563.7800000	301.6500000	792.0681262	4443.70	133.6000000	678.0000000
y2022	y2022	40	0	648.9975000	304.6000000	944.4942599	4927.50	127.9000000	723.1500000
y2021	y2021	40	0	490.9600000	262.3500000	716.3164531	4060.90	101.2000000	510.4500000
y2020	y2020	40	0	417.3100000	238.6500000	637.7393273	3696.60	94.9500000	429.9000000
y2019	y2019	40	0	421.3650000	235.1500000	667.2714819	3867.50	94.8000000	389.8500000
y2018	y2018	40	0	422.7375000	215.4500000	679.1593694	3921.50	87.9000000	405.1000000
y2017	y2017	40	0	453.3350000	241.1000000	709.4437830	4037.20	92.4500000	458.9500000
y2016	y2016	40	0	415.5675000	222.8500000	649.6112429	3640.30	82.9000000	405.9500000
y2015	y2015	40	0	426.8100000	184.4000000	660.9026519	3525.20	80.3500000	385.1500000
y2014	y2014	40	0	414.9375000	163.6500000	631.9903134	3275.70	68.8500000	421.8000000
y2013	y2013	40	0	381.6575000	146.3000000	578.7693734	3098.50	63.8500000	444.4000000
y2012	y2012	40	0	436.2850000	145.8000000	685.9584668	3639.00	63.3000000	571.3000000
y2011	y2011	40	0	485.7900000	131.8500000	774.5214045	3666.40	64.1500000	548.4000000
y2010	y2010	40	0	313.2525000	104.3000000	494.0316838	2543.70	52.8500000	333.7500000
y2009	y209	40	0	258.0200000	91.8000000	405.7014219	2115.10	47.9500000	302.9000000
y2008	y2008	40	0	279.0400000	101.8500000	467.7016756	2504.90	41.3500000	321.2000000
y2007	y2007	40	0	239.5600000	85.0500000	396.2060878	2112.50	34.8500000	307.2000000
y2006	y2006	40	0	209.7450000	62.9500000	352.6502524	1862.50	36.8500000	222.2500000
y2005	y2005	40	0	188.5425000	52.4500000	308.2376172	1610.60	33.7500000	193.7500000
y2004	y2004	40	0	157.2175000	58.8000000	271.1049687	1546.40	27.3500000	180.1000000
y2003	y2003	40	0	143.5975000	49.8000000	250.3830660	1448.20	23.3000000	181.7500000
y2002	y2002	40	0	122.7500000	51.7500000	210.1423083	1197.10	20.5500000	160.9000000
y2001	y2001	40	0	120.0350000	53.2500000	205.0454103	1169.70	16.8500000	157.8500000
y2000	y2000	40	0	172.0900000	63.0500000	300.2249232	1526.30	18.0500000	174.5000000
y1999	y1999	40	0	180.1850000	46.4000000	304.4593127	1451.50	16.4000000	159.8000000

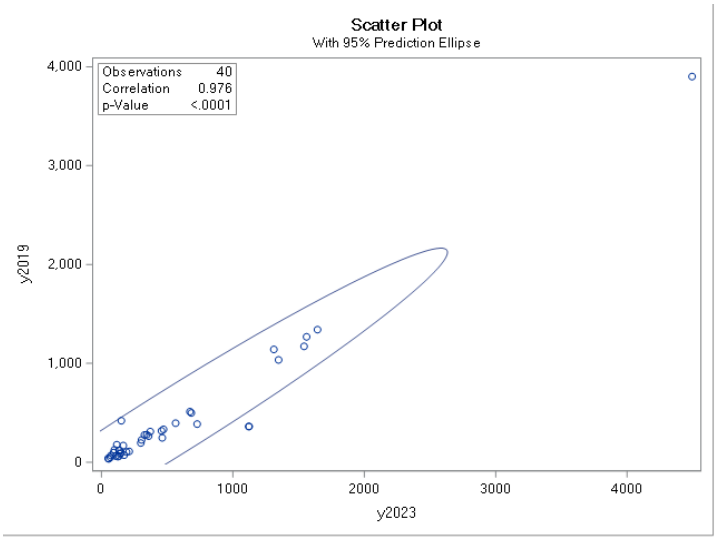
Finally, correlation coefficients were found for the 4 analysis years in the main dataset. The closest correlation variables from the main dataset are the years 2020 and 2023. However, the goal is to compare pre-COVID and post-COVID so the years 2019 and 2023 will be used for further analysis.

CORRELATION COEFFICIENTS FOR YEARS

Pearson Correlation Coefficients, N = 40 Prob > |r| under H0: Rho=0				
	y2023	y2022	y2019	y2020
y2023 y2023	1.00000	0.98168 <.0001	0.97601 <.0001	0.98450 <.0001
y2022 y2022	0.98168 <.0001	1.00000	0.97738 <.0001	0.98342 <.0001
y2019 y2019	0.97601 <.0001	0.97738 <.0001	1.00000	0.99810 <.0001
y2020 y2020	0.98450 <.0001	0.98342 <.0001	0.99810 <.0001	1.00000

To get a visual of the two comparison years, a scatterplot was created for the years 2019 and 2023. This visual confirms that a linear regression model is appropriate as all the observations fall into a linear pattern.

SCATTERPLOT OF YEARS 2019 AND 2023



To conduct comprehensive statistical analysis of this data, many hypothesis tests will be run, to get a better understanding of COVIDs impacts on the coffee industry. To start, ANOVA tests will be conducted for the years 2019 and 2023. An ANOVA test will be run for both product type and exporting country to see if there is a difference in the means across these categories. For all ANOVA tests, alpha will be 0.05.

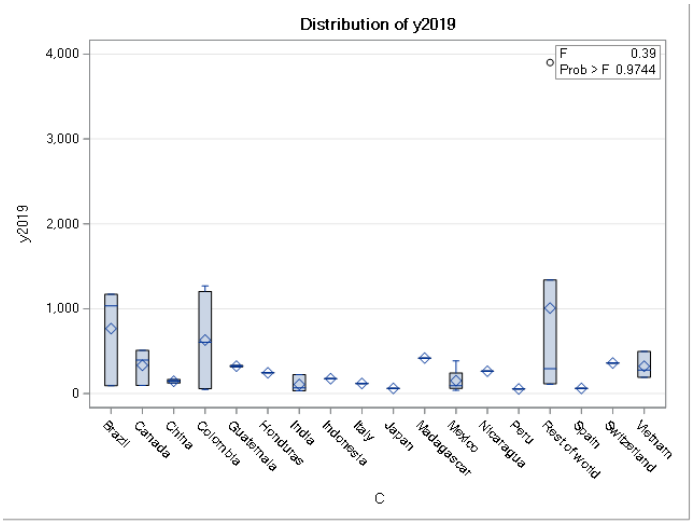
ANOVA TEST FOR Y2019 AND EXPORTING COUNTRY

H_0: The means for y2023 for all exporting countries are the same.

H_a: There is at least one means the is not equal for all the exporting countries in 2023.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	4011725.38	235983.85	0.39	0.9744
Error	22	13353072.61	606957.85		
Corrected Total	39	17364797.99			

Using the table above, the p-value is much higher than 0.05. Therefore, we fail to reject the null hypothesis. Below is a visualization of this data. While there is some variance in the means of each country, none of the variance is statistically significant.



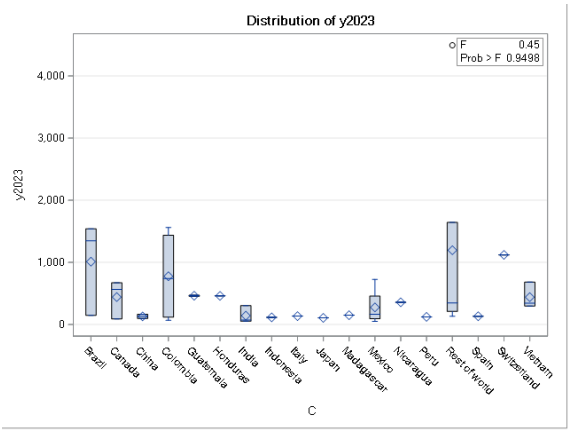
ANOVA TEST FOR Y2023 AND EXPORTING COUNTRY

H_0: The means for y2023 for all exporting countries are the same.

H_a: There is at least one means the is not equal for all the exporting countries in 2023.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	6345796.74	373282.16	0.45	0.9498
Error	22	18121708.00	823714.00		
Corrected Total	39	24467504.74			

Using the table above, the p-value is much higher than 0.05. Therefore, we fail to reject the null hypothesis. This is a similar trend to what happened in 2019. Below is a visualization of this data. While there is some variance in the means of each country, none of the variance is statistically significant.



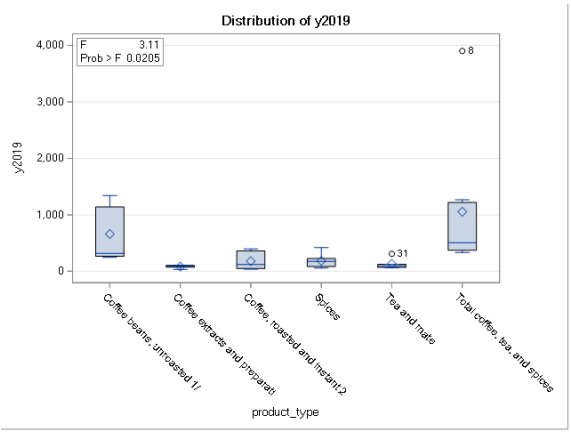
ANOVA TEST FOR Y2019 AND PRODUCT TYPE

H_0: The means for all product types will be the same over 2019.

H_a: At least one mean is difference for the product types over 2019.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5443867.92	1088773.58	3.11	0.0205
Error	34	11920930.07	350615.59		
Corrected Total	39	17364797.99			

Using the information above, the p-value=0.0205 and the table F-value is 3.11 Since $F_{table} = 3.11 > F_{5,34,0.05} = 2.494$ and $\alpha = 0.05 > p = 0.0205$, the null hypothesis will be rejected. This means there is a statistically significant difference between the means of the product types during 2019. Below is a visualization of this data.



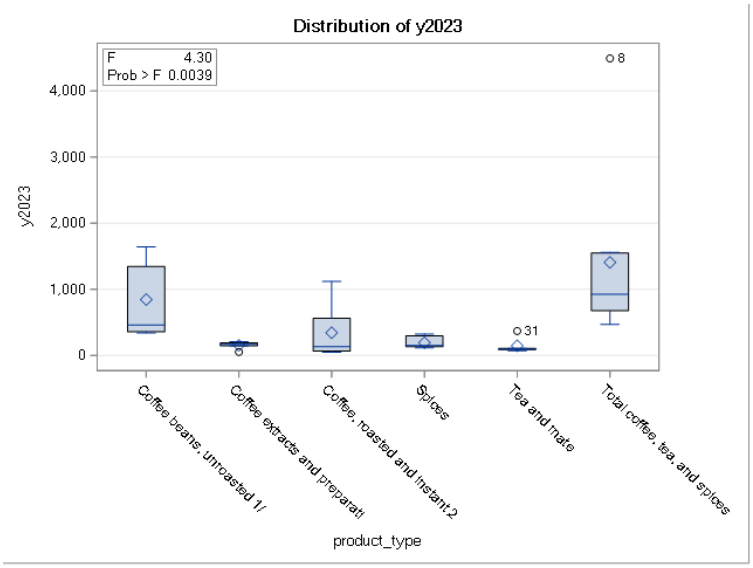
ANOVA TEST FOR Y2023 AND PRODUCT TYPE

H_0: The means for all product types will be the same over 2019.

H_a: At least one mean is difference for the product types over 2019.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9483100.24	1896620.05	4.30	0.0039
Error	34	14984404.50	440717.78		
Corrected Total	39	24467504.74			

Using the information above, the p-value=0.0039 and the table F-value is 4.30. Since$F_{table} = 4.30 > F_{5,34,0.05} = 2.494$ and $\alpha = 0.05 > p = 0.0039$, the null hypothesis will be rejected. This means there is a statistically significant difference between the means of the product types during 2023, similar to 2019. Below is a visualization of this data.



ANOVA TEST FOR Y2023 AND Y2019 LEVELS

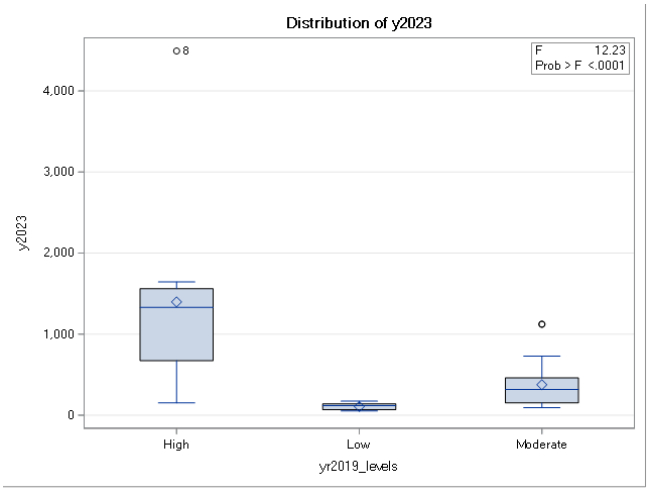
While not the best indicator of the relationship of years 2019 and 2023, the categorical variables(levels) will give some insight on the variance between the levels for 2019.

H_0: All the means across the 2019 levels are the same for 2023.

H_a: At least one mean across the 2019 levels is different for 2023.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	9736872.73	4868436.37	12.23	<.0001
Error	37	14730632.01	398125.19		
Corrected Total	39	24467504.74			

From the table, the p-value<0.0001 and the table F-value is 12.23. Since $F_{table} = 12.23 > F_{2,\,37,\,0.05} = 3.252$ and $\alpha = 0.05 > p < 0.0001$, the null hypothesis is rejected. There is at least one level that has a statistically significant difference in means. A visual of the data can be seen below.



ANOVA TEST FOR Y2019 AND Y2023 LEVELS

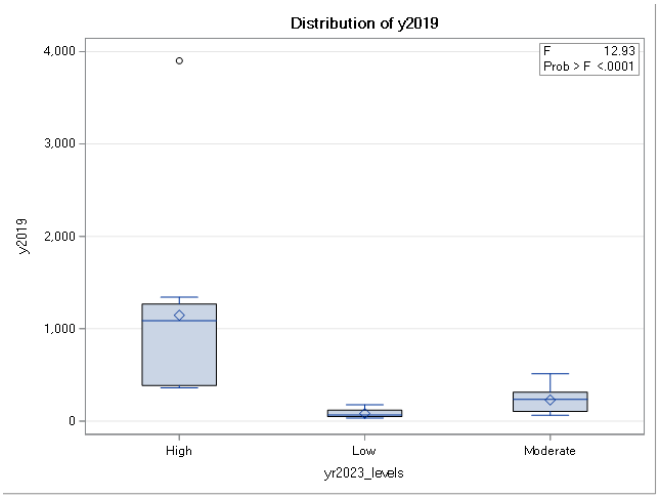
H_0: The means across the 2023 levels are the same for 2019.

H_a: At least one means across the 2023 levels is different for 2019.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7144223.41	3572111.70	12.93	<.0001
Error	37	10220574.58	276231.75		
Corrected Total	39	17364797.99			

From the table, the p-value<0.0001 and the table F-value is 12.93. Since $F_{table} = 12.93 > F_{2,37,0.05} = 3.252$ and $\alpha = 0.05 > p < 0.0001$, the null hypothesis is rejected.

There is at least one level that has a statistically significant difference in means. A visual of the data can be seen below.



PAIRED T-TEST FOR YEARS 2019 AND 2023

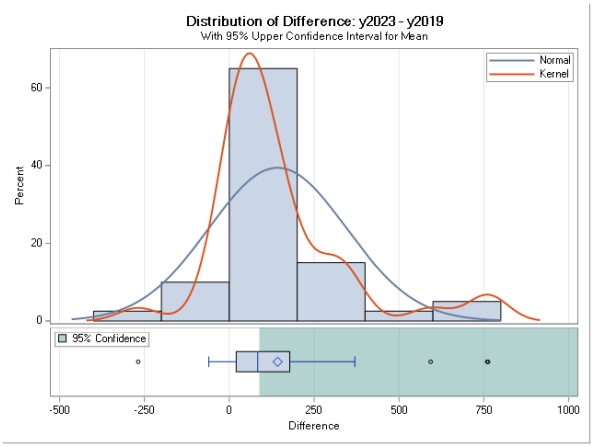
To further understand the relationship between the years 2019 and 2023, a paired t-test was run.

$$H_0\colon \mu_{2019} = \mu_{2023}$$

$$H_a\colon \mu_{2019} \neq \mu_{2023}$$

DF	t Value	Pr > t
39	4.45	<.0001

Using a t-Test distribution table, the critical value for these variables is $t_{39,0.05} = 2.023$. Since $t_{39,0.05} = 2.023 < t_{table} = 4.45$ and $\alpha = 0.05 > p < 0.0001$, the null hypothesis is rejected. Below is a visual of this data.

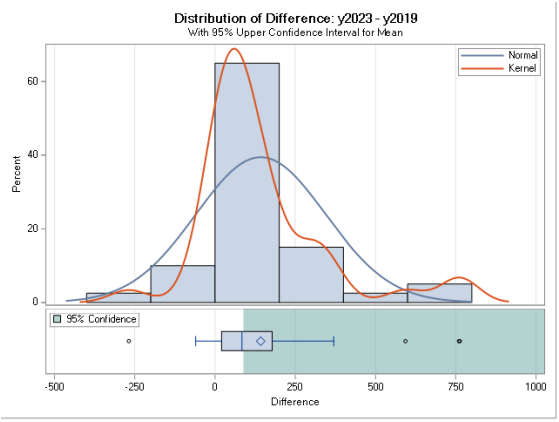


PAIRED T-TEST FOR YEARS 2020 AND 2023

While the forefront issue in this report is comparing pre- and post-COVID, there is a strong correlation between the years 2020 and 2023 as well. To get further information on this relationship, a paired t-Test was run for these years.

DF	t Value	Pr > t
39	4.45	<.0001

Using a t-Test distribution table, the critical value for these variables is $t_{39,0.05} = 2.023$. Since $t_{39,0.05} = 2.023 < t_{table} = 4.45$ and $\alpha = 0.05 > p < 0.0001$, the null hypothesis is rejected. Below is a visual of this data.



SIMPLE LINEAR REGRESSION FOR YEARS 2019 AND 2023

$H_0: \beta_1 = 0$

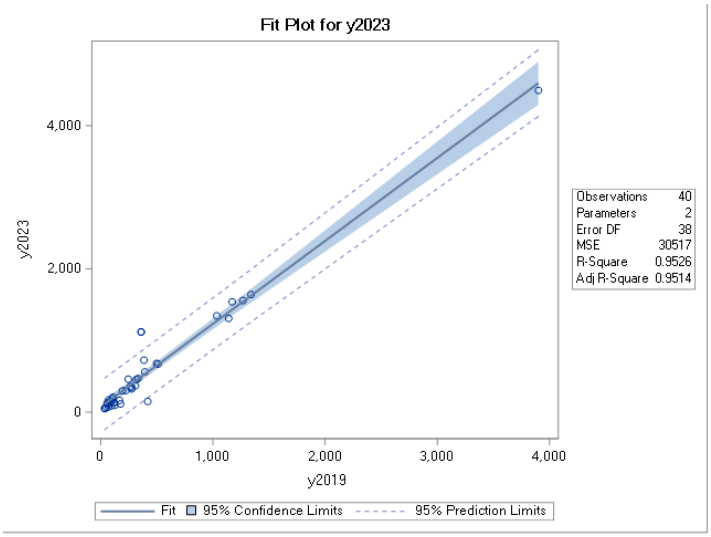
$H_a: \beta_1 \neq 0$

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	75.60602	32.78666	2.31	0.0267
y2019	y2019	1	1.15855	0.04192	27.64	<.0001

The equation built from this model would be y=75.606+1.1586x, where x is the imports into the U.S. in USD (millions) in 2019 and y is the imports into the U.S. in USD

(millions) in 2023. Since $\beta_1 = 1.15855$ and this is not zero, the null hypothesis will be rejected. Therefore, there is a linear relationship between the years 2019 and 2023.

By interpreting this equation, it can be said that for every 1 million USD in imports into the U.S. in 2019, there are 1.16 million USD in imports into the U.S. in 2023 with an initial lead by 75.6 million USD in 2023. Additionally, as seen below in the fit plot, the R-squared value is 0.9526 which means that approximately 95% of the data can be explained by the variability in the data.



Decision:

Each decision was made under each section. Below is a recap for each test.
ANOVA Y2019 AND EXPORTING COUNTRY: Fail to reject null hypothesis
ANOVA Y2023 AND EXPORTING COUNTRY: Fail to reject null hypothesis
ANOVA Y2019 AND PRODUCT TYPE: Reject null hypothesis
ANOVA Y2023 AND PRODUCT TYPE: Reject null hypothesis
ANOVA FOR Y2023 AND Y2019 LEVELS: Reject null hypothesis
ANOVA FOR Y2019 AND Y2023 LEVELS: Reject null hypothesis
PAIRED T-TEST FOR YEARS 2019 AND 2023: Reject null hypothesis
PAIRED T-TEST FOR YEARS 2020 AND 2023: Reject null hypothesis
SIMPLE LINEAR REGRESSION FOR YEARS 2019 AND 2023: Reject null hypothesis

Conclusion and Implications:

For most of the test, the null hypothesis was rejected. The main goal of this analysis was to compare the years 2019 (pre-COVID) and 2023 (post-COVID). Using the paired t-test for the years 2019 and 2023, the null hypothesis is rejected. This means there is a

statistically significant difference in the means of the two years. This suggests that COVID may have had an impact on the imports in USD of coffee into the United States. Additionally, consider the simple linear regression test where the null hypothesis was also rejected, there is additional evidence that there is a change in the imports from the years 2019 to 2023.

Limitations:

While this data is straight from the USDA ERS, there are some things that are overlooked. One thing to consider is inflation rates. Inflation rates over the 4-year period from 2019 to 2023 was approximately 4.5% every year. The dataset was also limited, with the main dataset having only 40 observations. Going forward, this analyis could be conducted on a larger data set with more observations.

Summary:

Considering the information found through the statistical analysis, there seems to be a change from 2019 to 2023 in the amount of imported coffee products. It appears there is an increase from 2019 to 2023. COVID-19 might have been one of the reasons this was the case. However, more research and statistical analysis would have to be performed to make more informed decisions. Additionally, as previously stated, there should be more consideration for the inflation rates and perhaps the correlation between COVID-19 and inflation.

References

Allen, Lark. “2024 Coffee Statistics: Consumption, Preferences, & Spending.” *Full-Service Market Research Company*, 1 Feb. 2024, www.driveresearch.com/market-research-company-blog/coffee-survey/.

Critical Values of the F-Distribution: α = 0.05, www.stat.purdue.edu/~lfindsen/stat511/F_alpha_05.pdf. Accessed 18 Nov. 2024.

Schoonjans, Frank. “T-Distribution Table (Two-Tailed).” *MedCalc*, MedCalc Software, 7 Apr. 2024, www.medcalc.org/manual/t-distribution-table.php.

“U.S. Food Imports.” *USDA ERS - U.S. Food Imports*, 3 Oct. 2024, www.ers.usda.gov/data-products/u-s-food-imports/.

“U.S. Inflation Rate 1960-2024.” *MacroTrends*, www.macrotrends.net/global-metrics/countries/USA/united-states/inflation-rate-cpi. Accessed 18 Nov. 2024.