

# Final Report - Group 4

Joanna Hsiao      Ophelia Liu

## Data Import

```
library(tidyverse)
library(lubridate)
library(patchwork)

# Read coffee futures data
futures <- read_csv("C:/Users/USER/Desktop/PBAwork/Final/Final_Version/US_CoffeeC_Futures_His.csv")

# Read USDA coffee production data
usda <- read_csv("C:/Users/USER/Desktop/PBAwork/Final/Final_Version/USDA_Coffee_Data.csv")

# Check structure
head(futures)
```

```
# A tibble: 6 x 7
  Date      Price   Open   High   Low Vol.    `Change %` 
  <chr>     <dbl>  <dbl>  <dbl>  <dbl> <chr>    <chr>    
1 12/01/2024 320.  318.  347.  294.  104.69K 0.53%  
2 11/01/2024 318.  246.  335.  241.  547.35K 29.34% 
3 10/01/2024 246.  270.  271.  243.  384.55K -9.01% 
4 09/01/2024 270.  247.  275.  240   142.26K 8.88%  
5 08/01/2024 248.  229.  259.  221.  0.02K   8.29%  
6 07/01/2024 229.  226.  250.  224.  152.99K 0.11% 
```

```
head(usda)
```

```
# A tibble: 6 x 21
  Country    Year `Beginning Stocks` `Arabica Production` `Robusta Production` ...
  <chr>      <dbl>           <dbl>            <dbl>            <dbl>            <dbl>            ...
```

	<code>&lt;chr&gt;</code>	<code>&lt;dbl&gt;</code>	<code>&lt;dbl&gt;</code>	<code>&lt;dbl&gt;</code>	<code>&lt;dbl&gt;</code>
1	Albania	1990	0	0	0
2	Algeria	1990	0	0	0
3	Angola	1990	186	20	80
4	Argentina	1990	0	0	0
5	Armenia	1990	0	0	0
6	Australia	1990	0	0	0
# i	16 more variables:	`Other Production` <code>&lt;dbl&gt;</code> , `Production` <code>&lt;dbl&gt;</code> ,			
#		`Bean Imports` <code>&lt;dbl&gt;</code> , `Roast & Ground Imports` <code>&lt;dbl&gt;</code> ,			
#		`Soluble Imports` <code>&lt;dbl&gt;</code> , `Imports` <code>&lt;dbl&gt;</code> , `Total Supply` <code>&lt;dbl&gt;</code> ,			
#		`Bean Exports` <code>&lt;dbl&gt;</code> , `Roast & Ground Exports` <code>&lt;dbl&gt;</code> ,			
#		`Soluble Exports` <code>&lt;dbl&gt;</code> , `Exports` <code>&lt;dbl&gt;</code> , `Rst,Ground Dom. Consum` <code>&lt;dbl&gt;</code> ,			
#		`Soluble Dom. Cons.` <code>&lt;dbl&gt;</code> , `Domestic Consumption` <code>&lt;dbl&gt;</code> ,			
#		`Ending Stocks` <code>&lt;dbl&gt;</code> , `Total Distribution` <code>&lt;dbl&gt;</code>			

## Data processing

```

# PROCESS FUTURES DATA
futures <- futures |>
  mutate(
    Date = mdy(Date),
    Year = year(Date)
  )

annual_price <- futures |>
  group_by(Year) |>
  summarize(
    Avg_Price = mean(Price, na.rm = TRUE),
    Min_Price = min(Price, na.rm = TRUE),
    Max_Price = max(Price, na.rm = TRUE),
    .groups = "drop"
  )

# PROCESS USDA PRODUCTION DATA

global_production <- usda |>
  group_by(Year) |>
  summarize(
    Global_Production = sum(Production, na.rm = TRUE),
    .groups = "drop"
  )

```

```

) |>
mutate(
  # Convert to million bags (production is in thousand 60kg bags)
  Global_Production_Million = Global_Production / 1000
)

# Get top 5 producers in most recent year
latest_year <- max(usda$Year, na.rm = TRUE)

top5_latest <- usda |>
  filter(Year == latest_year) |>
  arrange(desc(Production)) |>
  slice(1:5) |>
  mutate(Production_Million = Production / 1000) |>
  select(Country, Production, Production_Million)

# Merge data for analysis
merged_data <- global_production |>
  full_join(annual_price, by = "Year") |>
  arrange(Year) |>
  mutate(
    Price_Lag1 = lag(Avg_Price, 1)
  )

overlap_data <- merged_data |>
  filter(!is.na(Global_Production) & !is.na(Avg_Price))

```

## Global Trend

```

# Define Custom Theme
paper_theme <- theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 14, hjust = 0.5),
    plot.subtitle = element_text(size = 12, hjust = 0.5, color = "gray40"),
    axis.title = element_text(face = "bold"),
    panel.grid.minor = element_blank()
  )

# Create production plot

```

```

plot_production <- ggplot(overlap_data, aes(x = Year, y = Global_Production_Million)) +
  geom_line(linewidth = 1.2, color = "#2E7D32") +
  geom_point(size = 2.5, color = "#2E7D32") +
  labs(
    title = "Global Coffee Production Over Time",
    x = "Year",
    y = "Production"
  ) +
  paper_theme +
  theme(
    plot.title = element_text(face = "bold", size = 14, hjust = 0.5),
    axis.title = element_text(face = "bold"),
    panel.grid.minor = element_blank()
  ) +
  scale_x_continuous(breaks = seq(min(overlap_data$Year),
                                  max(overlap_data$Year),
                                  by = 2))

# Create price plot
plot_price <- ggplot(overlap_data, aes(x = Year, y = Avg_Price)) +
  geom_line(linewidth = 1.2, color = "#C62828") +
  geom_point(size = 2.5, color = "#C62828") +
  labs(
    title = "Coffee Futures Price (KC=F)",
    x = "Year",
    y = "Price"
  ) +
  paper_theme +
  theme(
    plot.title = element_text(face = "bold", size = 14, hjust = 0.5),
    axis.title = element_text(face = "bold"),
    panel.grid.minor = element_blank()
  ) +
  scale_x_continuous(breaks = seq(min(overlap_data$Year),
                                  max(overlap_data$Year),
                                  by = 2))

# Combine plots using patchwork
combined_plot <- plot_production / plot_price +
  plot_annotation(
    title = "Figure 1: Global Coffee Market Trends",
    subtitle = paste0("Analysis period: ", min(overlap_data$Year),

```

```

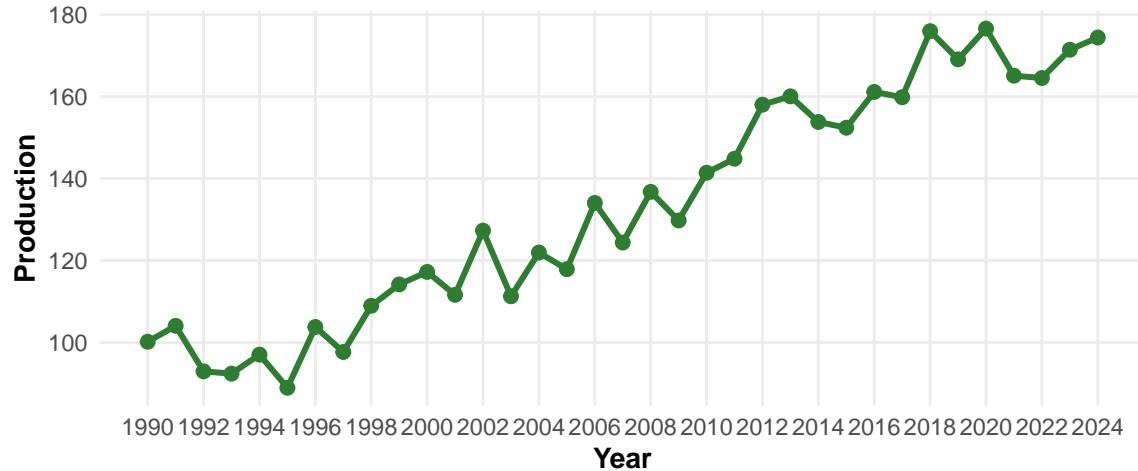
    " - ", max(overlap_data$Year)),
theme = paper_theme) +
plot_layout(heights = c(1, 1))
combined_plot

```

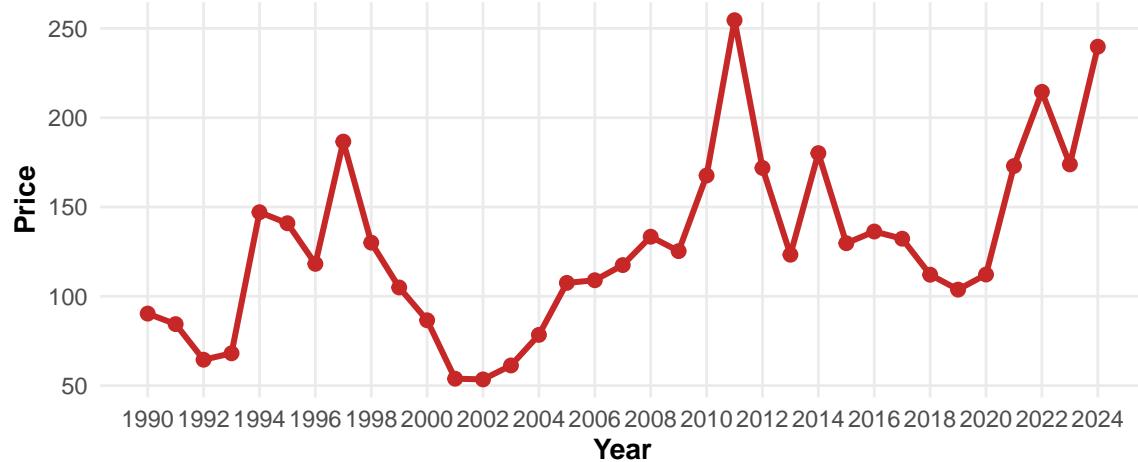
**Figure 1: Global Coffee Market Trends**

Analysis period: 1990 – 2024

### Global Coffee Production Over Time



### Coffee Futures Price (KC=F)



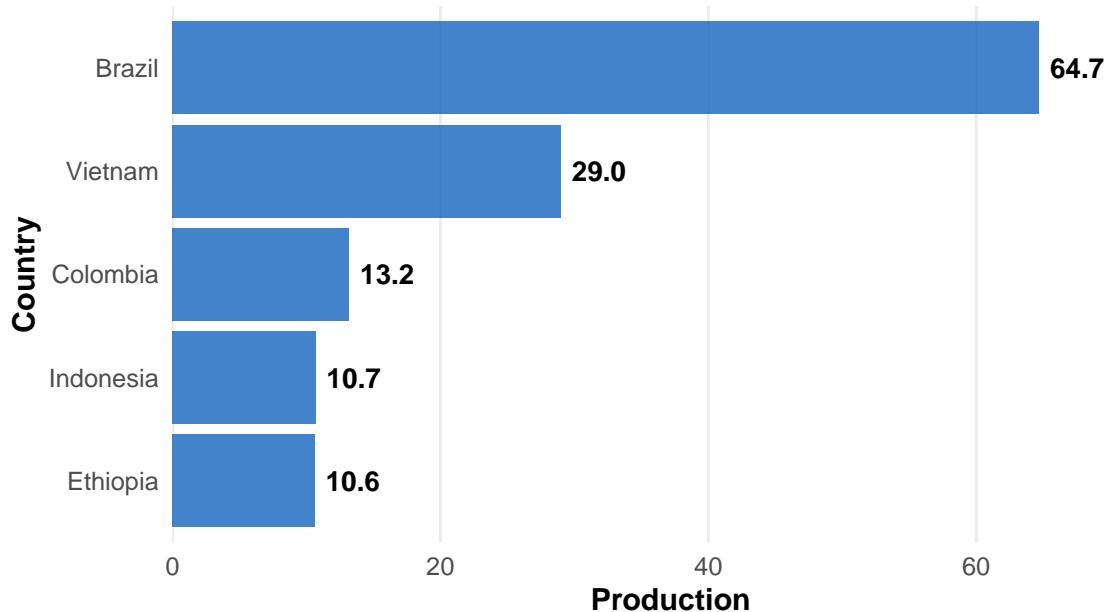
## Top 5 Producers

```
plot_top5 <- ggplot(top5_latest,
                     aes(x = Production_Million,
                         y = fct_reorder(Country, Production_Million))) +
  geom_col(fill = "#1565C0", alpha = 0.8) +
  geom_text(aes(label = sprintf("%.1f", Production_Million)),
            hjust = -0.2, size = 4, fontface = "bold") +
  labs(
    title = paste0("Figure 2: Top 5 Coffee Producing Countries 2024"),
    subtitle = "Production in Million 60kg bags",
    x = "Production",
    y = "Country"
  ) +
  paper_theme +
  theme(
    plot.title = element_text(face = "bold", size = 14, hjust = 0.5),
    plot.subtitle = element_text(size = 11, hjust = 0.5, color = "gray40"),
    axis.title = element_text(face = "bold"),
    panel.grid.minor = element_blank(),
    panel.grid.major.y = element_blank()
  ) +
  scale_x_continuous(expand = expansion(mult = c(0, 0.15)))

plot_top5
```

**Figure 2: Top 5 Coffee Producing Countries 2024**

Production in Million 60kg bags

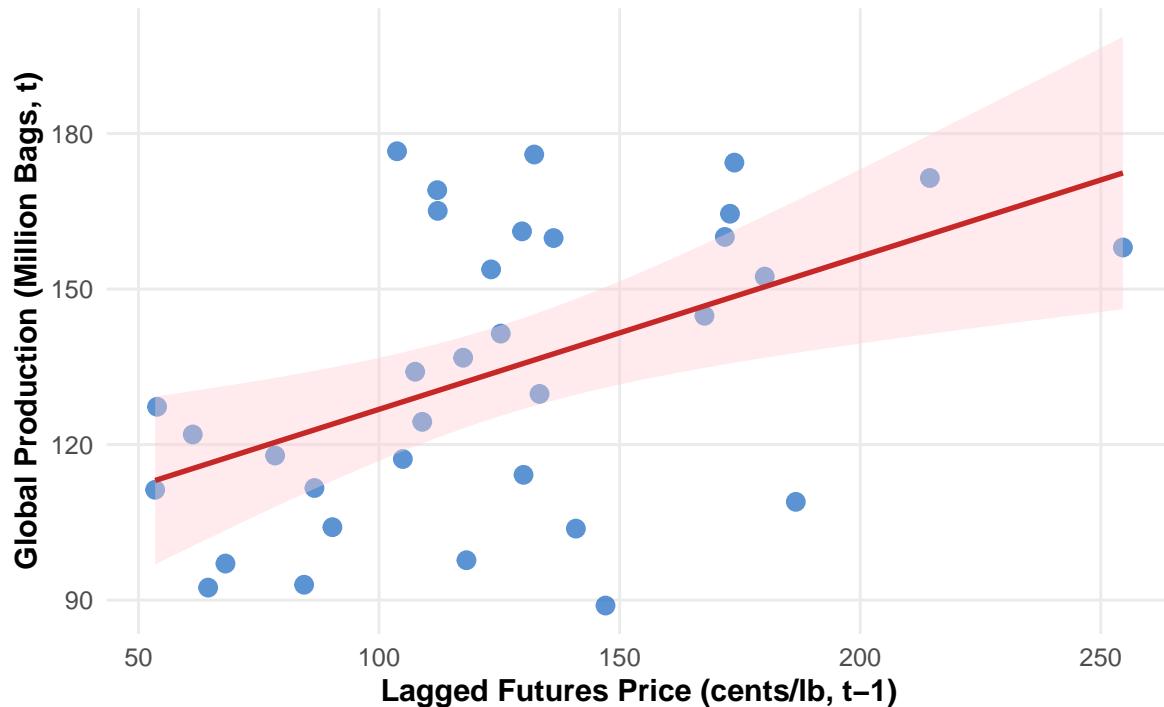


### Visualization - Scatter Plot with Regression Line

```
plot_scatter <- ggplot(merged_data, aes(x = Price_Lag1, y = Global_Production_Million)) +  
  geom_point(color = "#1565C0", size = 3, alpha = 0.7) +  
  geom_smooth(method = "lm", color = "#C62828", fill = "#FFCDD2") +  
  labs(  
    title = "Figure 3: Lagged Futures Price vs. Global Production",  
    subtitle = "Regression Analysis (1990-2024)",  
    x = "Lagged Futures Price (cents/lb, t-1)",  
    y = "Global Production (Million Bags, t)"  
) +  
  paper_theme  
  
print(plot_scatter)
```

**Figure 3: Lagged Futures Price vs. Global Production**

Regression Analysis (1990–2024)



## Regression Results

```
model_global <- lm(Global_Production_Million ~ Price_Lag1, data = merged_data)
summary(model_global)
```

Call:

```
lm(formula = Global_Production_Million ~ Price_Lag1, data = merged_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-51.734	-18.501	0.086	15.690	48.683

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	97.30050	12.47520	7.800	6.77e-09 ***

```

Price_Lag1    0.29497    0.09382   3.144  0.00358 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.97 on 32 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared:  0.236, Adjusted R-squared:  0.2121
F-statistic: 9.885 on 1 and 32 DF,  p-value: 0.003583

```

## Difference-in-Differences (DiD) Analysis

```

# Divide period group with Year 2021 (Pre/Post)
did_years <- c(2008:2010, 2012:2014)

top5_names <- top5_latest$Country

# Define Treated and Control Group
did_dataset <- usda |>
  filter(Year %in% did_years) |>
  mutate(
    Group = if_else(Country %in% top5_names, "Treated (Top 5)", "Control (Others)") ,
    Period = if_else(Year >= 2012, "Post", "Pre"),
    Production_Million = Production / 1000
  )

# Calculate DID
did_summary <- did_dataset |>
  group_by(Group, Period) |>
  summarize(Avg_Production = mean(Production_Million, na.rm = TRUE), .groups = "drop") |>
  pivot_wider(names_from = Period, values_from = Avg_Production) |>
  mutate(Change = Post - Pre)

did_value <- did_summary$Change[did_summary$Group == "Treated (Top 5)"] -
  did_summary$Change[did_summary$Group == "Control (Others)"]

print(did_summary)

# A tibble: 2 x 4
  Group          Post     Pre  Change
  <chr>        <dbl>   <dbl>   <dbl>
1 Treated (Top 5) 1000.  1000.  0.
2 Control (Others) 1000.  1000.  0.

```

```
1 Control (Others) 0.487 0.478 0.00915  
2 Treated (Top 5) 22.8   18.7   4.10
```

```
cat("\nDifference-in-Differences (DiD) Estimator:", round(did_value, 2), "Million Bags\n")
```

```
Difference-in-Differences (DiD) Estimator: 4.09 Million Bags
```