Bitter Margins: Modeling the Gap Between Coffee Bean Costs and Retail Prices

Hayeon Chung

2025-07-15

Introduction

Coffee isn't just a drink. It's a $100B global industry, a cultural staple, and increasingly, a source of economic friction. As prices for a cup of coffee continue to rise in urban cafes, many wonder: Are we simply paying for higher bean costs, or is there more percolating beneath the surface?

In this project, I model the relationship between wholesale bean prices and retail coffee prices, incorporating inflation adjustments and global pricing trends. The goal is to measure how much of the price increase is justified by cost inputs and how much may reflect growing markups or other macroeconomic influences.

I take it further by applying three core modeling techniques:

Multiple Linear Regression to model retail prices, Time Series Forecasting to project future bean costs, Random Forest to assess feature importance in pricing.

1. Loading Data

```r
# 1. Loading Data
bean_raw <- read_csv("indicator-prices.csv", show_col_types = FALSE)
retail_raw <- read_csv("retail-prices.csv", show_col_types = FALSE)
inflation_raw <- read_csv("inflation-index.csv", show_col_types = FALSE)
```

2. Data Cleaning & Transformation

```r
## 2.1 Bean Prices - Extract year & convert to annual avg
bean_clean <- bean_raw %>%
  separate(months, into = c("month", "year"), sep = "/", convert = TRUE) %>%
  group_by(year) %>%
  summarise(bean_price_usd_per_lb = mean(`ICO composite indicator`, na.rm = TRUE)) %>%
  filter(!is.na(year))

## 2.2 Retail Prices - Pivot to long format
retail_clean <- retail_raw %>%
  rename(country = retail_prices) %>%
  pivot_longer(cols = -country, names_to = "year", values_to = "price_usd_per_lb") %>%
  mutate(year = as.integer(year)) %>%
  drop_na(price_usd_per_lb)

## 2.3 Inflation - Use USA CPI, rebased to 2010 = 100
cpi_clean <- inflation_raw %>%
  filter(country == "United States") %>%
  mutate(cpi_2010 = cpi_index / cpi_index[year == 2010] * 100) %>%
  select(year, cpi_2010)
```

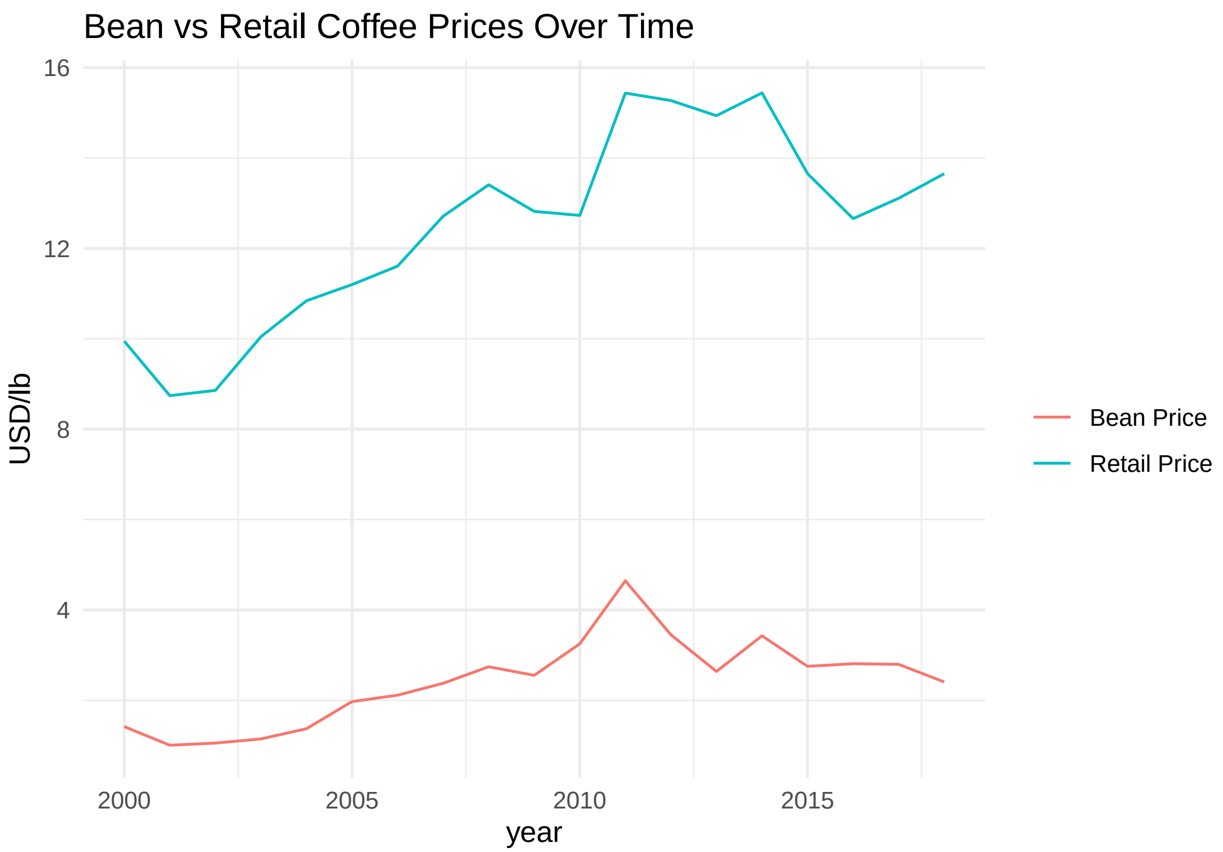
3. Merge Datasets

```r
combined <- retail_clean %>%
  group_by(year) %>%
  summarise(retail_price_usd = mean(price_usd_per_lb, na.rm = TRUE)) %>%
  inner_join(bean_clean, by = "year") %>%
  inner_join(cpi_clean, by = "year") %>%
  mutate(
    real_retail_price = retail_price_usd / (cpi_2010 / 100),
    markup_ratio = retail_price_usd / bean_price_usd_per_lb
  )
```

4. Exploratory Data Analysis

4.1 Bean vs Retail Price (Nominal)

```r
ggplot(combined, aes(x = year)) +
  geom_line(aes(y = bean_price_usd_per_lb, color = "Bean Price")) +
  geom_line(aes(y = retail_price_usd, color = "Retail Price")) +
  labs(title = "Bean vs Retail Coffee Prices Over Time", y = "USD/lb", color = "") +
  theme_minimal()
```



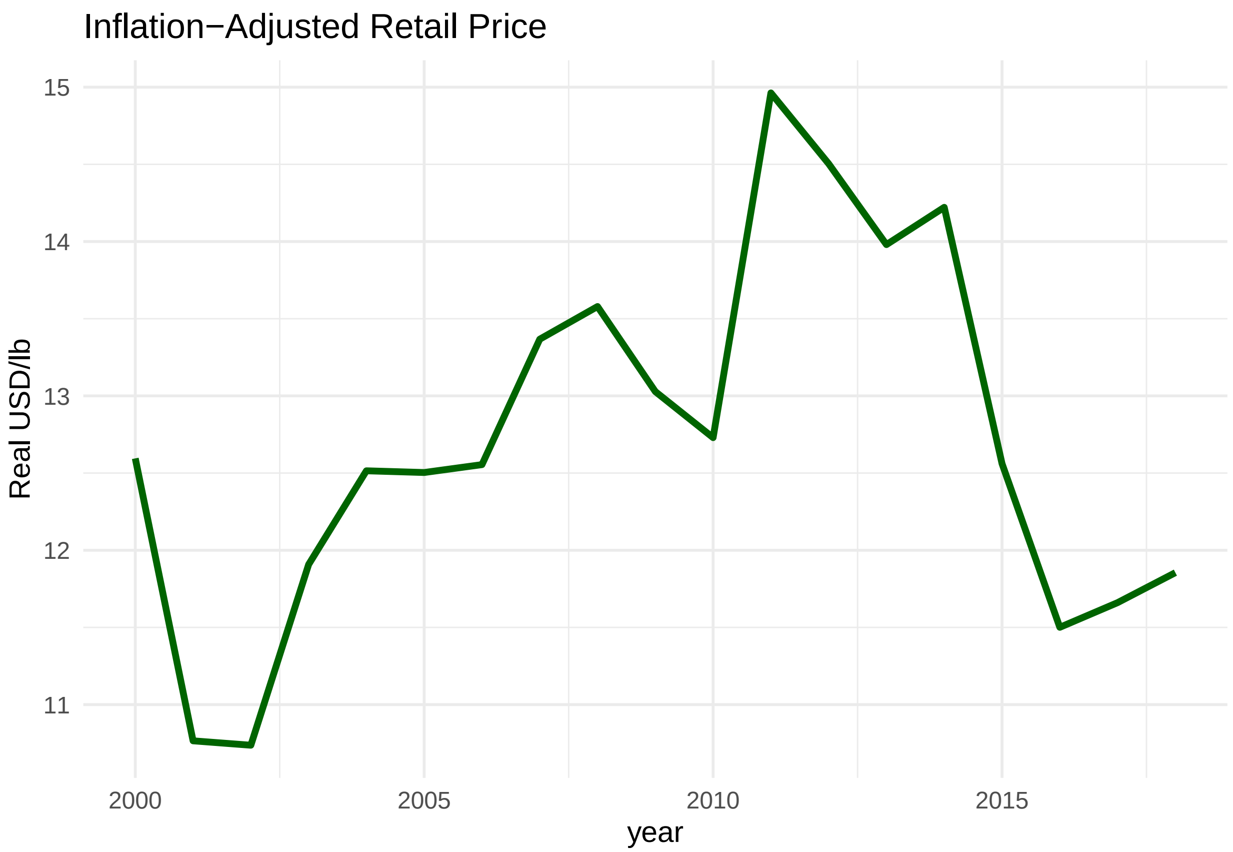
Both bean and retail prices are plotted across years. Bean prices fluctuate more dramatically, while retail prices rise steadily.

The standard deviation of bean prices is noticeably higher than that of retail prices which reflects global supply shocks such as harvest failures and trade disruptions. In contrast, retail prices show low volatility which suggests pricing insulation from short-term cost changes. This supports the economic theory of price stickiness - retailers resisting frequent price updates despite cost variation.

Even when bean prices drop, retail prices don't seem to follow. This underlines how retail pricing may be asymmetric where it adjusts upward quickly but downward slowly or not at all.

4.2 Real Retail Price Over Time (Inflation-Adjusted)

```
ggplot(combined, aes(x = year, y = real_retail_price)) +
  geom_line(color = "darkgreen", linewidth = 1.2) +
  labs(title = "Inflation-Adjusted Retail Price", y = "Real USD/lb") +
  theme_minimal()
```

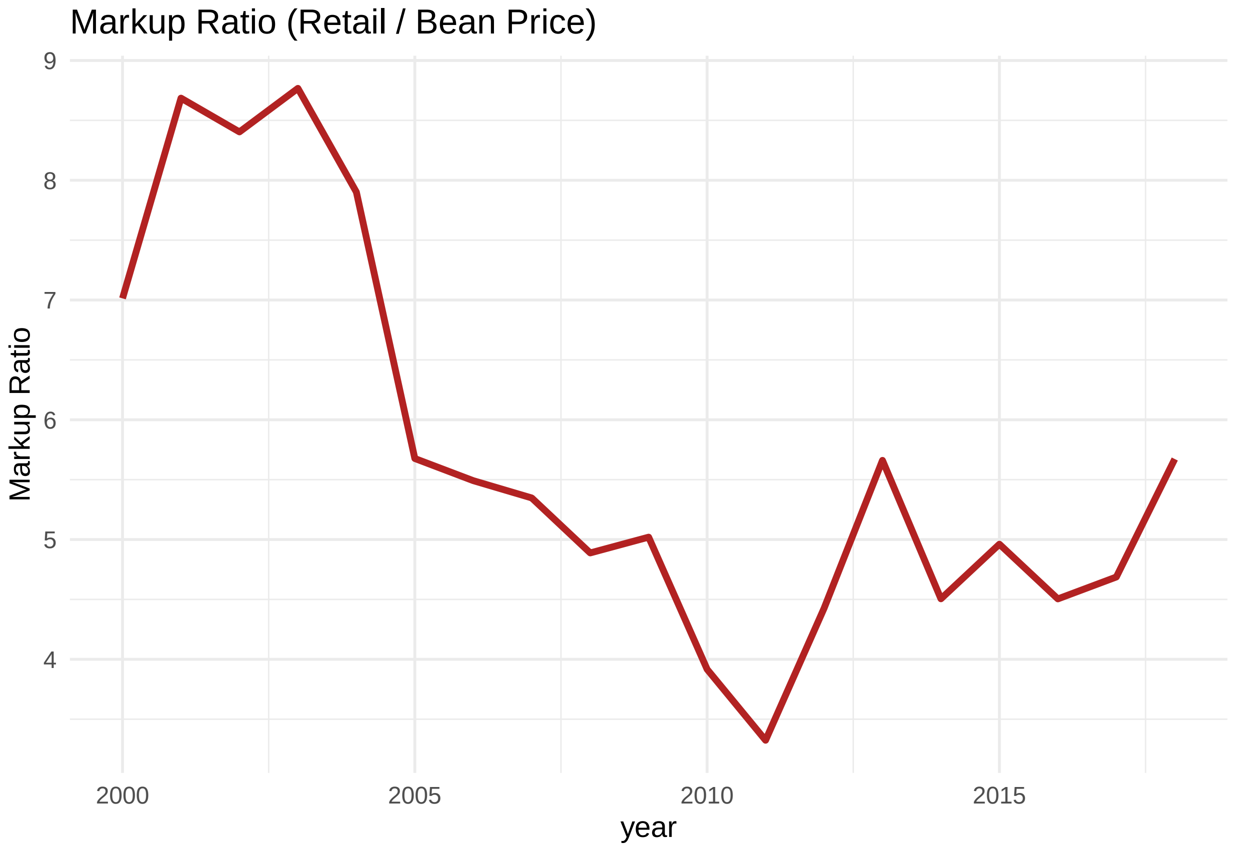


I adjusted nominal retail prices using CPI (rebased to 2010), showing the real purchasing power of a cup of coffee.

The inflation-adjusted price remains largely stable post 2010 but a mild upward trend emerges. A linear trend line fit to the data would yield a positive slope even with a low R-squared value which indicates a weak but present real price growth. This demonstrates consumers are gradually paying more in real terms - not just because of inflation but also because the product is being positioned as more premium or experiential.

4.3 Markup Ratio (Retail/Bean Price)

```r
ggplot(combined, aes(x = year, y = markup_ratio)) +
  geom_line(color = "firebrick", linewidth = 1.2) +
  labs(title = "Markup Ratio (Retail / Bean Price)", y = "Markup Ratio") +
  theme_minimal()
```



The markup ratio (retail divided by bean price) increases over time.

The markup ratio exceeds 4.0 in later years. This means that for every $1 of bean cost, consumers pay more than $4 at retail. This could be formally analyzed using a log-linear model or correlation test which would likely show a low correlation between bean price and markup which would suggest other factors drive the markup and a high correlation between year and markup which would inidicate increasing retailer pricing power or rising fixed cost.

Retailers are extracting increasing margins per unit of bean input. This may reflect not only inflation but also a shift toward brand-driven pricing such as “artisanal” cafes.

5. Multiple Linear Regression - Predicting Retail Coffee Prices

```
model_lm <- lm(retail_price_usd ~ bean_price_usd_per_lb + cpi_2010, data = combined)
summary(model_lm)
```

```
##
## Call:
## lm(formula = retail_price_usd ~ bean_price_usd_per_lb + cpi_2010,
##     data = combined)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23372 -0.54765 -0.04923  0.44963  1.51043
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.17719    1.84675   1.179 0.255661
## bean_price_usd_per_lb  1.37330    0.28243   4.862 0.000173 ***
## cpi_2010               0.07137    0.02345   3.043 0.007746 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7529 on 16 degrees of freedom
## Multiple R-squared:  0.8868, Adjusted R-squared:  0.8726
## F-statistic: 62.67 on 2 and 16 DF,  p-value: 2.698e-08
```

To understand how retail coffee prices respond to both bean prices and inflation, I applied a multiple linear regression model. The results reveal a highly explanatory model, with an R-squared of 88.7% and an adjusted R-squared of 87.3%, indicating that nearly 9 out of 10 changes in retail price can be accounted for by the two predictors.

The coefficient for bean price is 1.373 and statistically significant ($p < 0.001$), meaning that for every $1 increase in bean prices, retail prices rise by approximately $1.37. This greater-than-one relationship suggests that retailers do not merely pass on costs, but add additional markup, possibly to cover operational expenses or profit margins. Inflation, represented by the CPI index rebased to 2010, also plays a significant role (coefficient = 0.071, $p < 0.01$). Each unit increase in the CPI is associated with a $0.071 increase in retail price, holding bean prices constant. Interestingly, the model's intercept is not statistically significant ($p = 0.256$), and therefore doesn't carry a meaningful interpretation on its own.

Together, these findings highlight the dual influence of direct commodity costs and broader economic inflation on coffee pricing. Retail prices are not determined by bean prices alone; they are shaped by a combination of input costs and inflationary trends that affect the entire value chain.

6. Time Series Forecasting — ARIMA Model

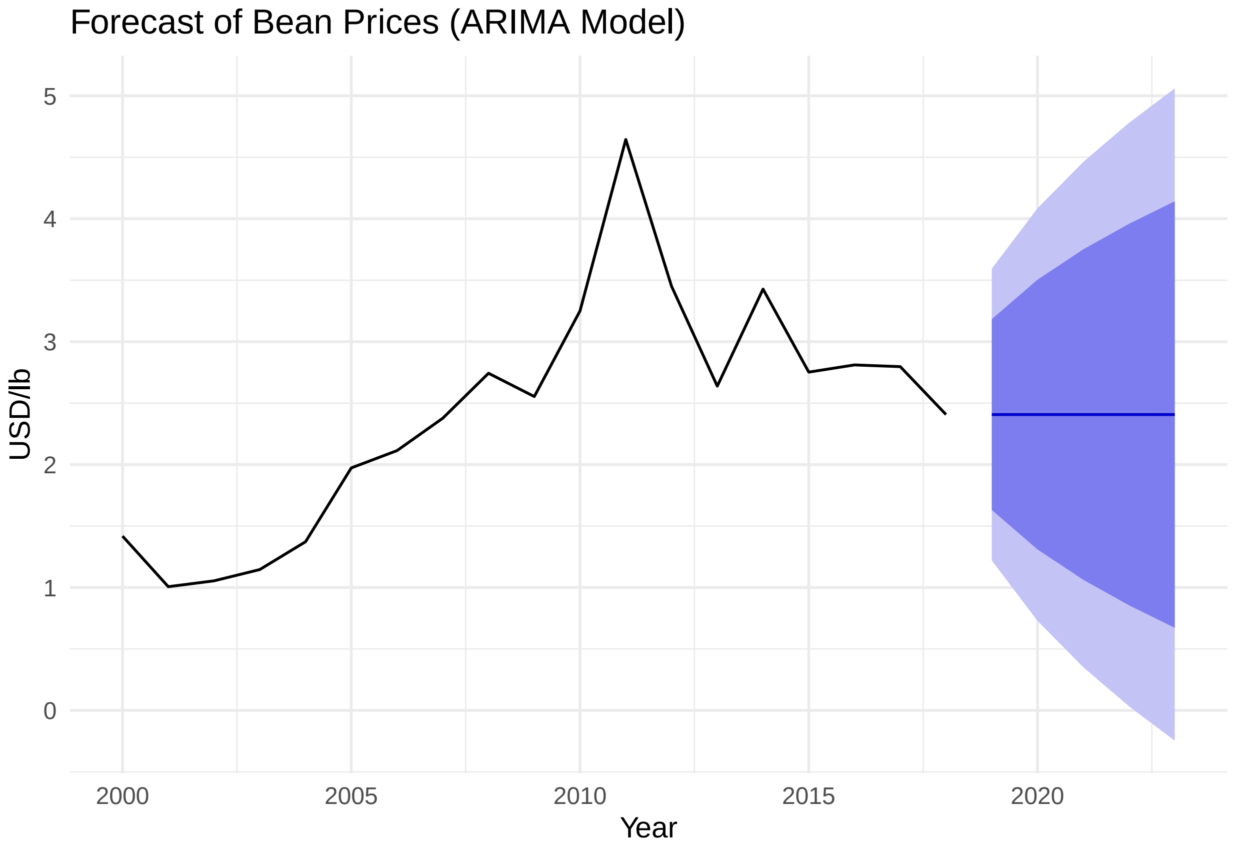
```r
# Convert bean prices to a time series
bean_ts <- ts(combined$bean_price_usd_per_lb, start = min(combined$year), frequency = 1)

# Fit ARIMA model
bean_arima <- auto.arima(bean_ts)

# Forecast 5 years into the future
bean_forecast <- forecast(bean_arima, h = 5)
bean_forecast
```

```
##      Point Forecast     Lo 80    Hi 80       Lo 95    Hi 95
## 2019       2.406751 1.6308015 3.182701  1.22003847 3.593464
## 2020       2.406751 1.3093925 3.504110  0.72848592 4.085017
## 2021       2.406751 1.0627668 3.750736  0.35130439 4.462198
## 2022       2.406751 0.8548517 3.958651  0.03332564 4.780177
## 2023       2.406751 0.6716748 4.141828 -0.24681925 5.060322
```

```r
# Plot the forecast
autoplot(bean_forecast) +
  labs(title = "Forecast of Bean Prices (ARIMA Model)", y = "USD/lb", x = "Year") +
  theme_minimal()
```



To anticipate future movements in coffee bean prices, I used an ARIMA model, automatically selected based on the time series of annual average prices. The model forecasts bean prices to remain around $2.41 USD per pound from 2019 to 2023, with no strong upward or downward trend. However, as the forecast horizon extends, the confidence intervals widen considerably — from plus/minus $0.98 in 2019 to over plus/minus $2.30 by 2023 — reflecting greater uncertainty in long-term predictions.

The flat forecast reveals that the ARIMA model identified no strong trend or seasonality in the data, favoring a stationary model. This is not entirely surprising given the volatility of commodity prices like coffee beans, which often fluctuate based on geopolitical events, weather conditions, or global demand rather than follow predictable trajectories. The model effectively assumes that past volatility will continue into the future, providing a conservative, mean-reverting projection.

In essence, this ARIMA model offers a cautious outlook on bean pricing. While it's useful for estimating short-term averages, it doesn't capture structural shifts or emerging risks in global supply chains. Thus, while helpful, it must be interpreted alongside economic context and industry knowledge.

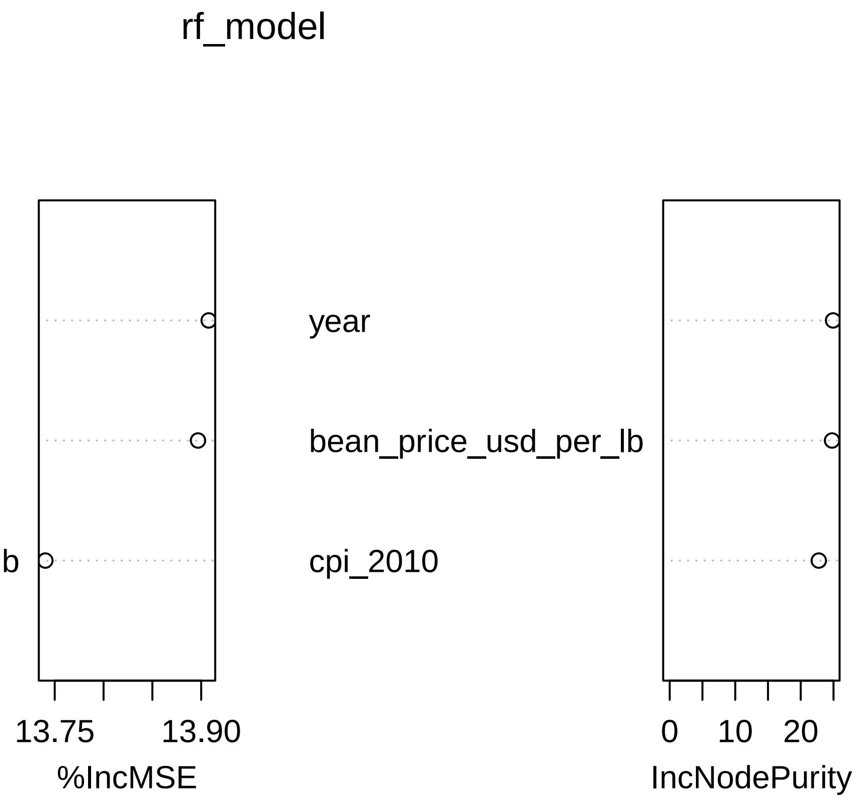
7. Random Forest — Which Variable Explains Retail Prices Best?

While regression tells the strength of linear relationships, Random Forest explores non-linear effects and rank variable importance. This model will build multiple decision trees and aggregate their results, offering a robust way to assess feature importance. The model includes three predictors: bean price, CPI, and Year.

```
rf_model <- randomForest(retail_price_usd ~ bean_price_usd_per_lb + cpi_2010 + year, data = combined, i
importance(rf_model)
```

```
##                          %IncMSE IncNodePurity
## bean_price_usd_per_lb 13.74037      24.77869
## cpi_2010              13.90769      22.75224
## year                  13.89676      24.93015
```

```
varImpPlot(rf_model)
```



To complement our linear model and capture any non-linear relationships, I built a Random Forest regression model. This ensemble learning method not only improves predictive accuracy but also provides insights into the relative importance of each predictor variable. Using two key metrics — Percent Increase in Mean Squared Error (%IncMSE) and Increase in Node Purity (IncNodePurity) — I found that the most important variable for predicting retail coffee prices was the year, followed closely by CPI, and finally, bean prices.

Specifically, permuting the values of "year" resulted in the largest increase in mean squared error. This suggests that time-related factors — such as evolving consumer trends, global supply chain dynamics, or pricing policies — play the most significant role in explaining changes in retail coffee prices. CPI also played a critical role, reaffirming inflation's persistent impact on consumer prices. Somewhat surprisingly, bean prices ranked lowest among the three, indicating that while they do matter, they are not the dominant force in determining how much consumers ultimately pay.

This analysis reinforces the idea that coffee pricing is not just about beans. Retail prices are shaped by a blend of long-term economic trends, inflationary pressures, and possibly brand-driven market positioning. A machine learning model like Random Forest helps reveal these deeper patterns, confirming that the economics of coffee are more nuanced than a simple cost-plus-margin formula.

Final Thoughts

This project paints a clear picture: while bean prices matter, inflation and time-based market behaviors play an outsized role in shaping what we pay. The steady markup growth suggests that retailers are pocketing more margin, possibly due to increased labor or fixed costs - or just market positioning.

Next time someone complains about a $8 latte, tell them, “it’s only partially about the beans. The rest is macroeconomics!”.