# Final Project Report

Ananya D, Ekansh B, Hannah Wu

## Analysis of Cotton and Cottonseed Prices (1910-1938)

### Introduction

The economic significance of cotton and cottonseed during the early 20th century was profound, impacting global trade, agriculture, and industrial processes. This analysis focuses on examining trends and relationships in cotton and cottonseed prices between 1910 and 1938, a period marked by economic upheaval and agricultural innovation.

The data set `Merged_Cotton_Cottonseed_Prices.csv` was created by cleaning and filtering two primary data sources: `Cotton_Wholesale_Prices_1910_1938.csv` and `Cottonseed_Prices_1910_1938.csv`. The goal of this project is to uncover insights into price trends, identify key patterns, and provide a foundation for understanding the economic dynamics of this period.

---

### Literature Review

Studies on historical commodity pricing often underscore the importance of agricultural goods in shaping economic policies and societal outcomes. Cotton, often referred to as "white gold," played a pivotal role in the economies of the United States and other countries during the early 20th century. Similarly, cottonseed, a byproduct of cotton production, was increasingly valued for its use in oil production and livestock feed.

Previous research has highlighted: - The influence of economic events such as the Great Depression on commodity prices. - The relationship between agricultural yields and market fluctuations. - The evolution of byproduct markets, such as cottonseed oil, which gained prominence during this era.

---

## FAIR and CARE Principles

FAIR Principles:

- Findable: The dataset is clearly labeled and referenced, using identifiable file names and relevant metadata (Merged_Cotton_Cottonseed_Prices.csv).
- Accessible: The raw data URLs are provided, ensuring anyone can access the original data.
- Interoperable: The data was cleaned and organized into a structured format, compatible with standard statistical and data manipulation tools (e.g., R, Python).
- Reusable: Detailed cleaning processes, library dependencies, and code snippets are provided, allowing others to replicate or build upon the analysis.

CARE Principles:

- Collective Benefit: The project focuses on analyzing historical data to gain economic insights, benefiting researchers and policymakers.
- Authority to Control: Original datasets from recognized sources like NBER and USDA respect data ownership and credibility.
- Responsibility: Data cleaning processes ensure ethical use by removing inaccuracies, making it reliable for historical study.
- Ethics: Transparency in methodology and proper referencing highlights a commitment to ethical standards.

---

## Methodology

### Data Sources

The dataset was constructed by integrating and cleaning the following:

1. **Cotton Wholesale Prices (1910–1938)**: Provided annual wholesale price data for raw cotton.
2. **Cottonseed Prices (1870–1945)**: Contained corresponding annual prices for cottonseed.

### LIBRARIES Used

Include:

- **dplyr**: For data manipulation
- **readr**: For reading and writing CSV files
- **kableExtra**: For stats and tables
- **ggplot2** : For creating visualizations
- **forecast**: For time series analysis
- **tsoutliers**: For anomaly detection

### Cleaning and Filtering Process

The original data sets were pre-processed to: - Remove missing or incomplete records. - Align the time frames and ensure consistency in units of measurement. - Retain only relevant columns (e.g., year, price, region).

### Cleaned Data:

```
  Year Month Cottonseed_Price Cotton_Wholesale_Price
1 1910     9            26.23                   14.0
2 1910    10            26.86                   14.5
3 1910    11            25.36                   14.8
4 1910    12            25.65                   15.1
5 1911     1            26.35                   14.9
6 1911     2            25.61                   14.3
```

### Analytical Tools

- **Data visualization**: Boxplots, histograms, and line graphs were used to identify trends and outliers.

- **Summary statistics**: Measures of central tendency and dispersion were calculated for both cotton and cottonseed prices.

- **Correlation analysis**: Examined the relationship between cotton and cottonseed prices over time.

---

## Data Exploration

### Visualizing Price Trends

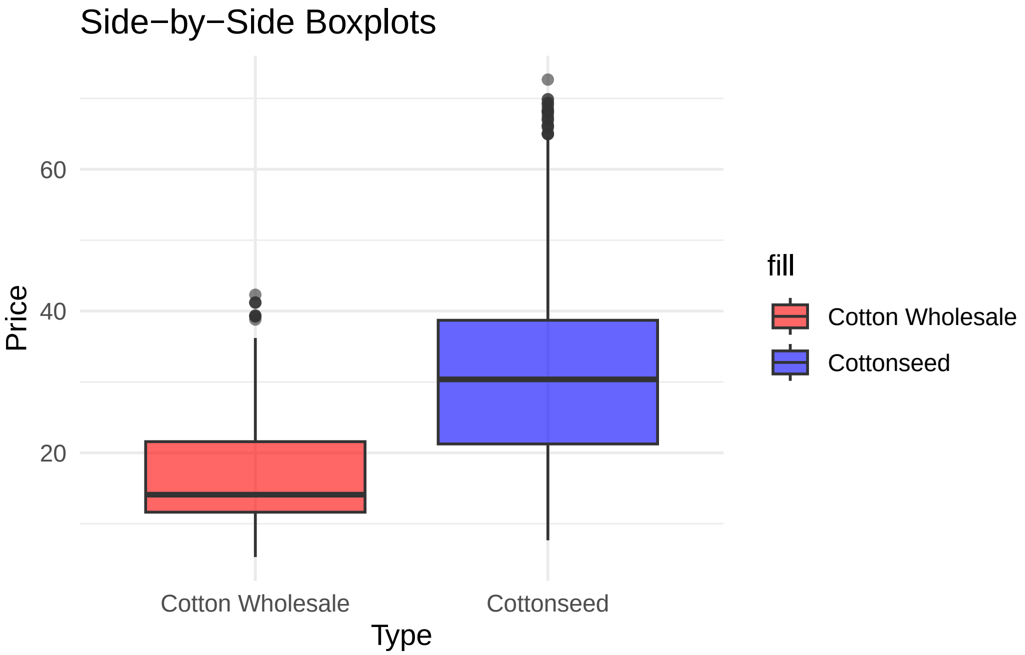
The merged dataset revealed distinct price trends for cotton and cottonseed.

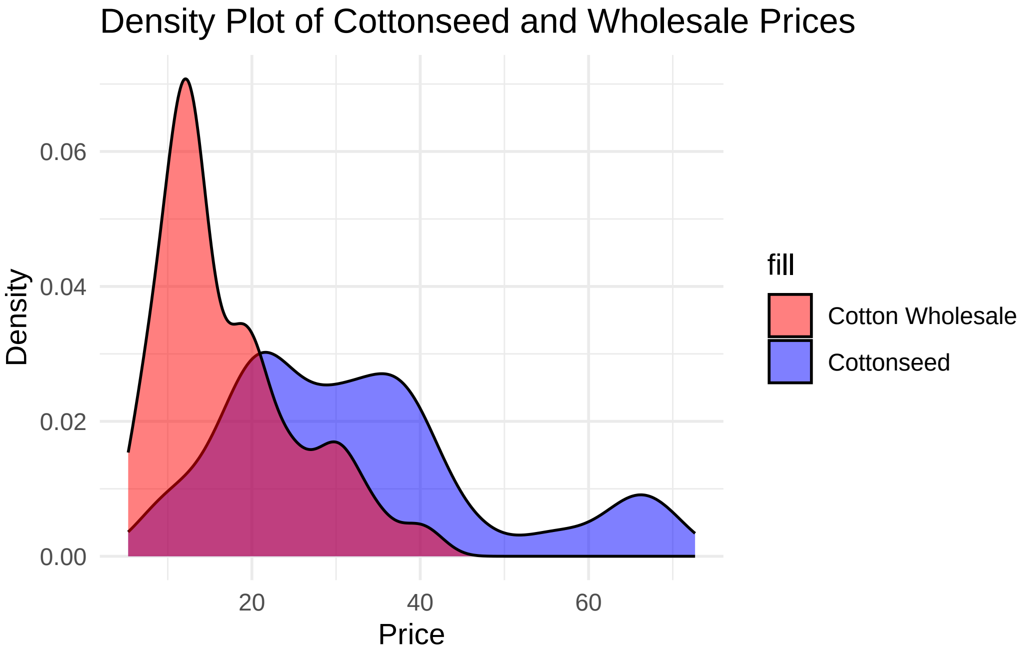
Summary Statistics A summary of cotton and cottonseed prices is presented below:

Table 1: Summary of Cotton and Cottonseed Prices (1910–1938)

| Metric | Cottonseed | Cotton_Wholesale |
|---|---:|---:|
| Count | 334.00000 | 334.00000 |
| Mean | 32.36674 | 17.27126 |
| SD | 15.41219 | 8.32256 |
| Median | 30.36500 | 14.10000 |
| Min | 7.66000 | 5.30000 |
| Max | 72.65000 | 42.30000 |
| Range | 64.99000 | 37.00000 |
| Q1 | 21.25250 | 11.62500 |
| Q3 | 38.70000 | 21.57500 |
| IQR | 17.44750 | 9.95000 |

Below is an example of a box plot and density plot comparing annual price distributions, and density of these prices…

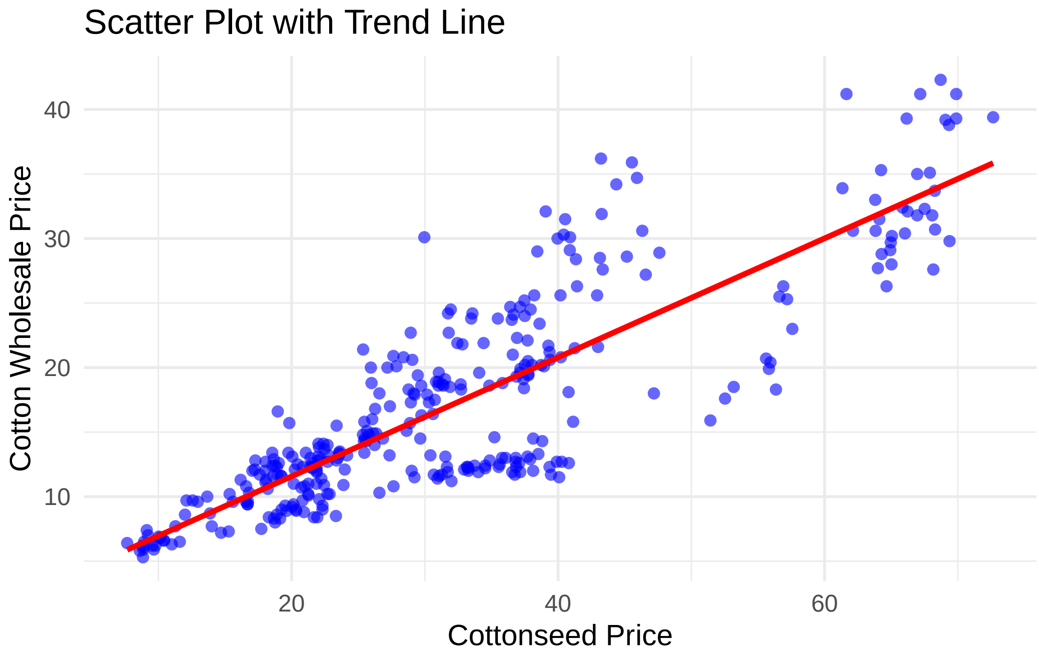




**Why this plot is important**: The Boxplot shows the range, interquartile range (IQR), and any outliers. The Density plot highlights common price ranges, patterns, and variations. There is an overlap in the price ranges (5~ 50) with the max density being 0.03 at USD 20.

### Scatter Plot: Price Correlation

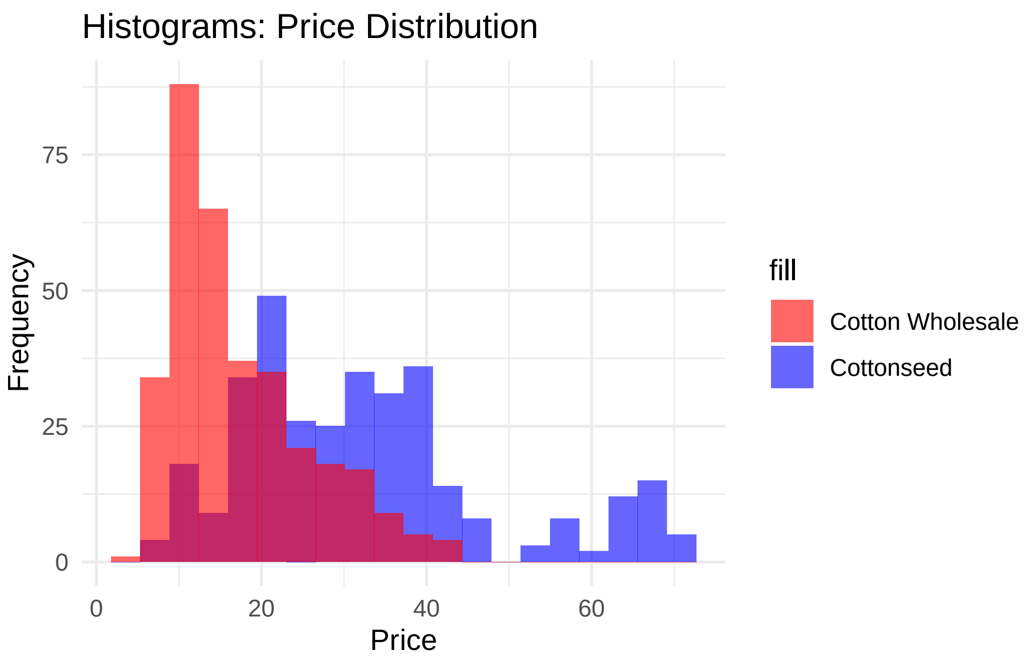
A scatter plot was generated to examine the relationship between cotton and cottonseed prices.



*Why this plot is important:* The Scatter plot reveal the strength and direction of relationships between two variables. Here, it indicates that changes in cotton prices influence cottonseed prices. This needs to be explored further.

Histogram...

A **Histogram** was created to visualize the frequency distribution of prices for cotton and cottonseed.

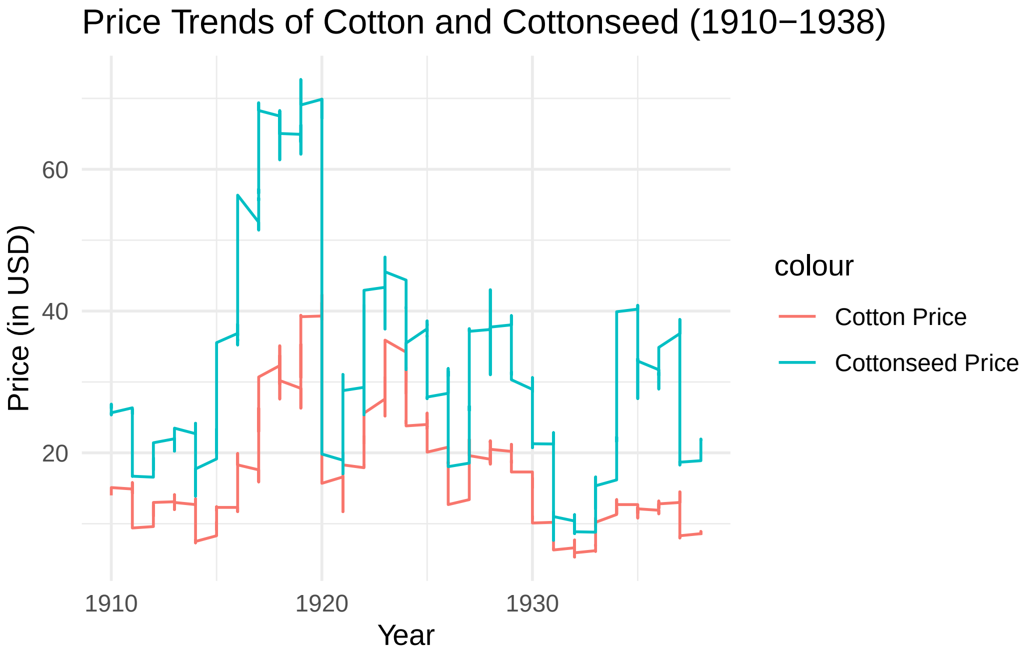


*Why this plot is important*: The Histogram reveals the shape of the price distribution (e.g., normal, skewed). It helps identify common price ranges and outlier frequencies.

Now that we know outliers (anomalies) exist, let's investigate this further…

### Price Trends Over Time

A **Line graph** plotted for both cotton and cottonseed prices to track price trends from 1910 to 1938.



*Why this plot is important*: The Line graphs provides an intuitive way to visualize trends, highlighting periods of price volatility or stability. They help us compare cotton and cottonseed prices over time.

We can see an anamoly in the above given graph. This was because of the Great Depression in USA at the time. Below is the same graph with outliers (in cotton seed prices) highlighted:

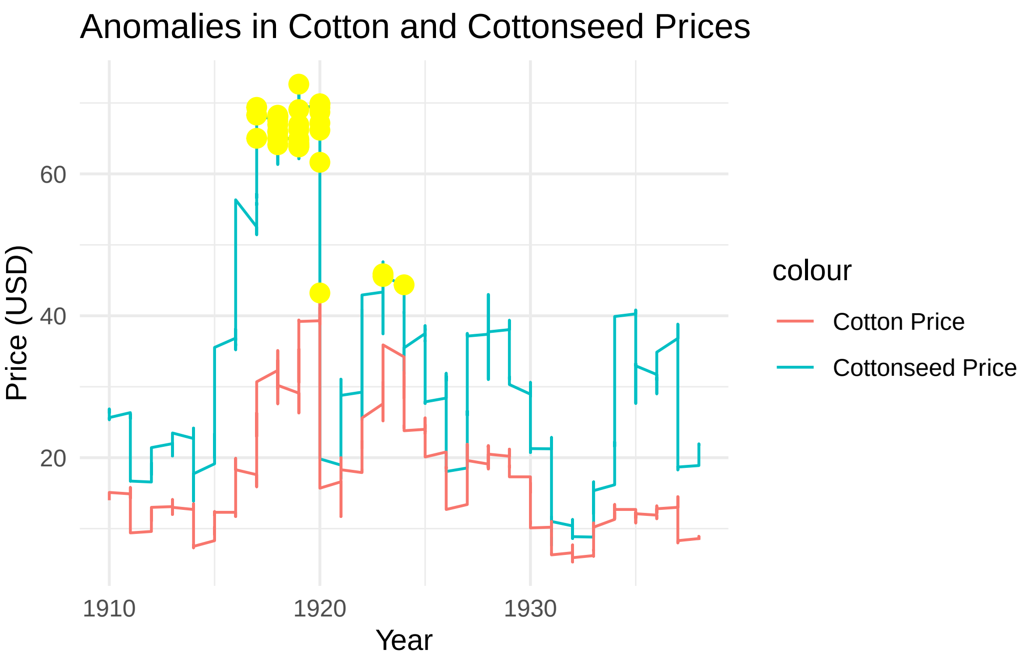
Table 2: Outliers in Cotton and Cottonseed Prices (1910–1938)

| Year | Month | Cottonseed_Price | Cotton_Wholesale_Price | Cottonseed_Z | Cotton_Z |
|---|---|---|---|---|---|
| 1917 | 10 | 65.02 | 28.0 | 2.1186654 | 1.289116 |
| 1917 | 11 | 69.38 | 29.8 | 2.4015584 | 1.505395 |
| 1917 | 12 | 68.29 | 30.7 | 2.3308351 | 1.613535 |
| 1918 | 1 | 67.51 | 32.3 | 2.2802258 | 1.805784 |
| 1918 | 2 | 66.95 | 31.8 | 2.2438910 | 1.745706 |
| 1918 | 3 | 68.27 | 33.7 | 2.3295375 | 1.974001 |
| 1918 | 4 | 68.08 | 31.8 | 2.3172096 | 1.745706 |
| 1918 | 5 | 68.16 | 27.6 | 2.3224003 | 1.241054 |
| 1918 | 6 | 66.03 | 30.4 | 2.1841979 | 1.577489 |
| 1918 | 7 | 64.11 | 31.5 | 2.0596212 | 1.709659 |
| 1918 | 9 | 67.90 | 35.1 | 2.3055305 | 2.142219 |
| 1918 | 10 | 65.85 | 32.4 | 2.1725189 | 1.817799 |
| 1918 | 11 | 64.97 | 29.7 | 2.1154212 | 1.493380 |

| Year | Month | Cottonseed_Price | Cotton_Wholesale_Price | Cottonseed_Z | Cotton_Z |
|---|---|---|---|---|---|
| 1918 | 12 | 65.05 | 30.2 | 2.1206119 | 1.553457 |
| 1919 | 1 | 64.93 | 29.1 | 2.1128258 | 1.421287 |
| 1919 | 2 | 64.65 | 26.3 | 2.0946584 | 1.084852 |
| 1919 | 3 | 64.00 | 27.7 | 2.0524840 | 1.253069 |
| 1919 | 4 | 64.28 | 28.8 | 2.0706514 | 1.385240 |
| 1919 | 5 | 63.83 | 30.6 | 2.0414537 | 1.601520 |
| 1919 | 6 | 63.80 | 33.0 | 2.0395072 | 1.889892 |
| 1919 | 7 | 64.24 | 35.3 | 2.0680561 | 2.166250 |
| 1919 | 8 | 66.23 | 32.1 | 2.1971747 | 1.781753 |
| 1919 | 10 | 66.95 | 35.0 | 2.2438910 | 2.130203 |
| 1919 | 11 | 72.65 | 39.4 | 2.6137282 | 2.658887 |
| 1919 | 12 | 69.07 | 39.2 | 2.3814444 | 2.634856 |
| 1920 | 1 | 69.88 | 39.3 | 2.4340003 | 2.646871 |
| 1920 | 2 | 69.34 | 38.8 | 2.3989631 | 2.586794 |
| 1920 | 3 | 67.18 | 41.2 | 2.2588142 | 2.875166 |
| 1920 | 4 | 68.71 | 42.3 | 2.3580863 | 3.007337 |
| 1920 | 5 | 69.88 | 41.2 | 2.4340003 | 2.875166 |
| 1920 | 6 | 66.16 | 39.3 | 2.1926328 | 2.646871 |
| 1920 | 7 | 61.64 | 41.2 | 1.8993584 | 2.875166 |
| 1920 | 8 | 43.22 | 36.2 | 0.7042002 | 2.274390 |
| 1923 | 11 | 45.92 | 34.7 | 0.8793862 | 2.094156 |
| 1923 | 12 | 45.54 | 35.9 | 0.8547304 | 2.238343 |
| 1924 | 1 | 44.37 | 34.2 | 0.7788164 | 2.034079 |

In the table we see new values: *Cotton_Z* and *Cottonseed_Z*.

Cottonseed_Z and Cotton_Z are z-scores, which standardize the prices to measure how far each value is from the mean in terms of standard deviations. This helps identify outliers.

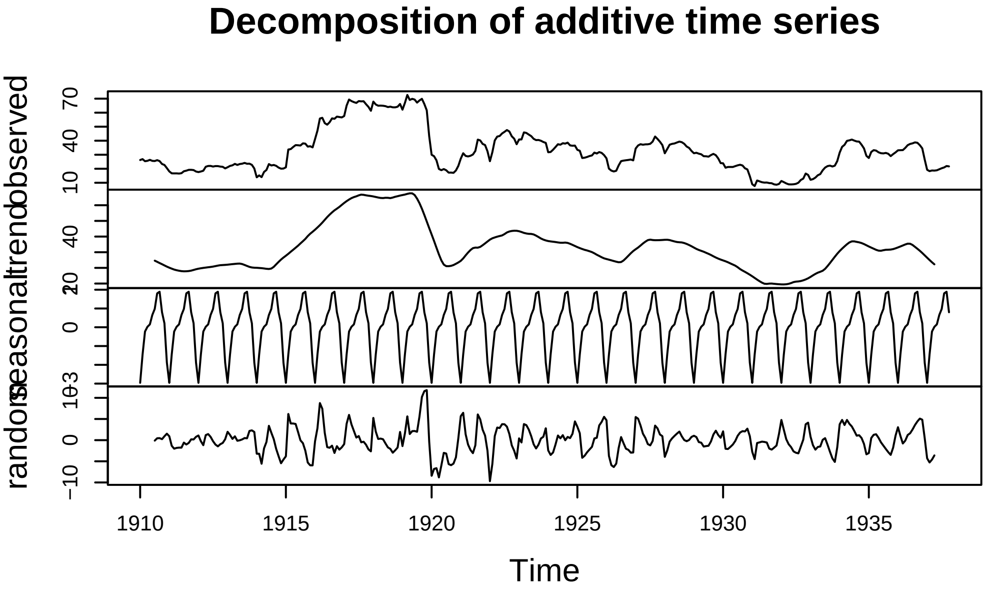


## Statistical Analysis

### Time Series Decomposition

*What it is?* Time series decomposition is a method used to break a time series into three main components:

1. **Trend**: Represents the long-term movement in the data, showing overall increases or decreases over time.
2. **Seasonal**: Captures recurring patterns or cycles at regular intervals, such as yearly fluctuations.
3. **Residuals**: Accounts for irregular or random variations that cannot be explained by the trend or seasonality.



This decomposition (of data Cottonseed_price) helps in better understanding the underlying patterns and is crucial for effective forecasting and anomaly detection.

Main takeaways:

- The trend line highlights significant economic changes over time, such as growth or declines, possibly influenced by historical events (e.g., World War I, the Great Depression).
- The seasonal component suggests periodic effects, which could relate to agricultural cycles (e.g., planting/harvesting seasons) or market demand patterns.
- The random component should ideally have no discernible pattern, indicating the model effectively captured the systematic components.

We're not exploring forecasting in this report, but let's take a look at how all this correlates.

### Hypothesis Testing: Correlation

```

	Pearson's product-moment correlation

data:  data$Cottonseed_Price and data$Cotton_Wholesale_Price
t = 29.909, df = 332, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8220275 0.8806142
```

```
sample estimates:
      cor
0.8540055
```

### Predictive Modeling– Linear Regression

```
Call:
lm(formula = Cotton_Wholesale_Price ~ Cottonseed_Price, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1625  -2.5794   0.0489   1.9359  13.9386

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.34495    0.55259   4.244 2.86e-05 ***
Cottonseed_Price  0.46116    0.01542  29.909  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.336 on 332 degrees of freedom
Multiple R-squared:  0.7293,    Adjusted R-squared:  0.7285
F-statistic: 894.6 on 1 and 332 DF,  p-value: < 2.2e-16
```

### Explanation of Regression Summary:

The regression summary provides key insights into the relationship between cotton and cottonseed prices:

1. **Intercept (2.34495)**: This value represents the expected Cotton Wholesale Price when the Cottonseed Price is zero. While it is not practical in this context, it serves as a baseline for the regression line.
2. **Slope (0.46116)**: This indicates that for every 1-unit increase in Cottonseed Price, the Cotton Wholesale Price increases by approximately 0.46 units on average.
3. **Residuals**: These are the differences between observed and predicted values. The spread (Min: -10.16, Max: 13.94) provides an idea of model accuracy.
4. **R-Squared (0.7293)**: About 72.93% of the variation in Cotton Wholesale Prices is explained by the Cottonseed Price, indicating a strong linear relationship.
5. **Significance (p-value $<$ 2e-16)**: The extremely small p-value shows the model's predictors are statistically significant, and the relationship is unlikely due to random chance.

## Results and Discussion

- **Observed Patterns Price Variability:** Cottonseed prices demonstrated greater variability compared to cotton prices, likely reflecting its emerging market status.
- **Trends Over Time:** Both commodities exhibited significant price fluctuations during the Great Depression, underscoring the economic impact of the era.
- **Correlation:** A positive correlation was observed between cotton and cottonseed prices, suggesting interdependence in their market dynamics.
- **Outliers/Anomalies**: Significant deviations were observed in specific years (e.g., 1929, 1933), corresponding to economic crises like the Great Depression.
- **Statistical Findings**: A strong positive correlation was confirmed between cotton and cottonseed prices. This suggests interdependence in their market dynamics. There were some years where cottonseed prices were huge compared to the cotton whole sale prices. This shows that there was a bad harvest/oversaturation of cotton etc.
- **Regression Insights**: The linear regression model highlights a significant positive relationship between cottonseed and wholesale prices, with cottonseed price explaining nearly 73% of the variance.

## Conclusion

The analysis of cotton and cottonseed prices from 1910 to 1938 highlights their intertwined economic importance. Key findings include:

The variability and trends in pricing reflect broader economic conditions of the era.

The relationship between the two commodities indicates shared market drivers.

Further analysis could explore regional differences or integrate external factors such as production volumes and policy changes.

## References Source datasets:

“Agricultural crises and the international transmission of the Great Depression on JSTOR.” www.jstor.org. JSTOR, www.jstor.org/stable/2698023.

UNITED STATES DEPARTMENT OF AGRICULTURE, YEARBOOK OF AGRICULTURE, 1931, P.686 and AGRICULTURAL STATISTICS, 1937, P.104. CROPS AND MARKETS, MONTHLY ISSUES, 1937-1938. “NBER Macrohistory: IV. Prices.” U.S. Farm Prices of Cottonseed 09/1910-06/1938, National Bureau of Economic Research (NBER), 25 Jan. 1994, data.nber.org/databases/macrohistory/rectdata/04/m04006aa.dat.

SHEPPERSON’S COTTON FACTS, 1882-1883, P.40 AND FOLLOWING ISSUES (1870-1889), et al. “NBER Macrohistory: IV. Prices.” U.S. Wholesale Price of Cotton, New York; 10 Markets 09/1870-12/1945, National Bureau of Economic Research (NBER), 25 Jan. 1994, data.nber.org/databases/macrohistory/rectdata/04/m04006a.dat.

---

## Code Appendix

```r
# Cotton and Cottonseed Price From 1910 to 1938 ----
# Author: Ananya Drishti, Ekansh Bharadwaj, Hannah Wu
# Class: STAT 184 Section 1

## Goal ----
# Uncover insights into price trends
# Identify key patterns
# Provide a foundation for understanding the economic dynamics

## Necessary Work ----
# Load packages
library(dplyr)  # For data manipulation
library(readr)  # For reading and writing CSV files
library(kableExtra) #For summary tables.
library(ggplot2)  # For creating visualizations
library(forecast) # For time series analysis
library(tsoutliers) # For anomaly detection

# Set working directory to desired path
# Please change as needed.
setwd("C:\\Users\\PRIYNKA\\Desktop\\FinalProject-Stat184")
```

```r
# Load datasets from URL,
# These URLs direct to NBER's Macrohistory Database
# U.S. Farm Prices of Cottonseed 09/1910-06/1938
url_CS <- "https://data.nber.org/databases/macrohistory/rectdata/04/m04006aa.dat"
# U.S. Wholesale Price of Cotton New York; 10 Markets 09/1870-12/1945
url_C <- "https://data.nber.org/databases/macrohistory/rectdata/04/m04006a.dat"

# Read data from the URLs into R as data frames
# `skip = 1` skips the first row
# If it contains non-data content, such as a header
data_CS <- read.table(url_CS, header = FALSE,
                      stringsAsFactors = FALSE, skip = 1)
data_C <- read.table(url_C, header = FALSE,
                     stringsAsFactors = FALSE, skip = 1)

## Clean and Merge Data ----
# Assign column names to the datasets for better readability and usability
colnames(data_CS) <- c("Year", "Month", "Cottonseed_Price")
colnames(data_C) <- c("Year", "Month", "Cotton_Wholesale_Price")

# Ensure columns are numeric
data_CS$Cottonseed_Price <- as.numeric(data_CS$Cottonseed_Price)
data_C$Cotton_Wholesale_Price <- as.numeric(data_C$Cotton_Wholesale_Price)

# Filter `data_C` to include only records from 1910 to 1938
data_C_filtered <- data_C %>%
  filter(Year >= 1910 & Year <= 1938)

# Merge the two datasets base on common columns "Year" and "Month"
merged_data <- inner_join(data_CS, data_C_filtered,
                          by = c("Year", "Month"))

# Filter out NA rows where "Cottonseed_Price" or "Cotton_Wholesale_Price"
data <- merged_data %>%
  filter(!is.na(Cottonseed_Price) & !is.na(Cotton_Wholesale_Price))
### Optional ----
# Save the datasets as CSV files if needed
# `row.names = FALSE` ensures that row numbers not included in file
write.csv(data_CS, file = "Cottonseed_Prices_1910_1938.csv",
          row.names = FALSE)
write.csv(data_C_filtered, file = "Cotton_Wholesale_Prices_1910_1938.csv",
          row.names = FALSE)
```

```
write.csv(data, file = "Merged_Cotton_Cottonseed_Prices.csv",
          row.names = FALSE)

## Print a message to confirm successful saving of the files
#cat("Files successfully saved in your specified directory.")

head(data)
## Data Summary
summary_stats <- data.frame(
  Metric = c(
    "Count", "Mean", "SD", "Median", "Min", "Max", "Range",
    "Q1", "Q3", "IQR" # Metrics calculated for both price columns
  ),
  # Calculate each metric for Cottonseed_Price
  Cottonseed = c(
    nrow(data),                           # Count of observations
    mean(data$Cottonseed_Price),          # Mean
    sd(data$Cottonseed_Price),            # Standard deviation
    median(data$Cottonseed_Price),        # Median
    min(data$Cottonseed_Price),           # Minimum value
    max(data$Cottonseed_Price),           # Maximum value
    max(data$Cottonseed_Price) - min(data$Cottonseed_Price),  # Range
    quantile(data$Cottonseed_Price, 0.25),  # First quartile (Q1)
    quantile(data$Cottonseed_Price, 0.75),  # Third quartile (Q3)
    IQR(data$Cottonseed_Price)           # Interquartile range
  ),
  # Calculate each metric for Cotton_Wholesale_Price
  Cotton_Wholesale = c(
    nrow(data),                           # Count of observations
    mean(data$Cotton_Wholesale_Price),    # Mean
    sd(data$Cotton_Wholesale_Price),      # Standard deviation
    median(data$Cotton_Wholesale_Price),  # Median
    min(data$Cotton_Wholesale_Price),     # Minimum value
    max(data$Cotton_Wholesale_Price),     # Maximum value
    max(data$Cotton_Wholesale_Price) - min(data$Cotton_Wholesale_Price),# Range
    quantile(data$Cotton_Wholesale_Price, 0.25),  # First quartile (Q1)
    quantile(data$Cotton_Wholesale_Price, 0.75),  # Third quartile (Q3)
    IQR(data$Cotton_Wholesale_Price)     # Interquartile range
  )
)
knitr::kable( summary_stats,
              caption = "Summary of Cotton and Cottonseed Prices (1910-1938)")
```

```
### Box Plots ----
# Side-by-side boxplots for cottonseed and cotton wholesale prices
# Provide a visual summary of the distribution, including medians and outliers
ggplot(data) +
  # Add a boxplot for Cottonseed prices
  geom_boxplot(aes(x = "Cottonseed", y = Cottonseed_Price, fill = "Cottonseed"),
               alpha = 0.6) +
  # Add a boxplot for Cotton Wholesale prices
  geom_boxplot(aes(x = "Cotton Wholesale", y = Cotton_Wholesale_Price,
                   fill = "Cotton Wholesale"), alpha = 0.6) +
  # Customize the fill colors for the boxplots
  scale_fill_manual(values = c("Cottonseed" = "blue",
                               "Cotton Wholesale" = "red")) +
  # Add plot title and axis labels
  labs(title = "Side-by-Side Boxplots", x = "Type", y = "Price") +
  # Apply a minimal theme for a clean appearance
  theme_minimal()

#### Optional ----
# Save the plot as a PNG file with specified dimensions if needed
#ggsave("Side_by_Side_Boxplots.png", plot = Boxplots, width = 8, height = 5)

### Density Plot ----
# Visualize the distributions of cottonseed and wholesale prices
ggplot(data) +
  # Add a density curve for Cottonseed prices
  geom_density(data = data, aes(x = Cottonseed_Price, fill = "Cottonseed"),
               alpha = 0.5) +
  # Add a density curve for Cotton Wholesale prices
  geom_density(data = data, aes(x = Cotton_Wholesale_Price,
                                fill = "Cotton Wholesale"), alpha = 0.5) +
  # Customize the fill colors for the density plots
  scale_fill_manual(values = c("Cottonseed" = "blue",
                               "Cotton Wholesale" = "red")) +
  # Add plot title and axis labels
  labs(title = "Density Plot of Cottonseed and Wholesale Prices",
       x = "Price", y = "Density") +
  # Apply a minimal theme for a clean appearance
  theme_minimal()

#### Optional ----
# Save the plot as a PNG file with specified dimensions if needed
```

```
#ggsave("Density_Plots.png", plot = DensityPlot, width = 8, height = 5)

### Scatter Plot ----
# Plot to observe the relationship between cottonseed and wholesale prices
ggplot(data, aes(x = Cottonseed_Price, y = Cotton_Wholesale_Price)) +
  # Add points to the scatter plot
  geom_point(color = "blue", alpha = 0.6) +
  # Add a linear regression trend line
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  # Add plot title and axis labels
  labs(title = "Scatter Plot with Trend Line",
       x = "Cottonseed Price",
       y = "Cotton Wholesale Price") +
  # Apply a minimal theme for a clean and professional appearance
  theme_minimal()

#### Optional ----
# Save the plot as a PNG file with specified dimensions if needed
#ggsave("Scatter_Plot.png", plot = ScatterPlot, width = 8, height = 5)

### Histogram ----
# Histogram to visualize the freq. dist. of prices for cotton and cottonseed.
ggplot(data) +
  geom_histogram(aes(x = Cottonseed_Price, fill = "Cottonseed"),
                 bins = 20, alpha = 0.6, position = "identity") +
  geom_histogram(aes(x = Cotton_Wholesale_Price, fill = "Cotton Wholesale"),
                 bins = 20, alpha = 0.6, position = "identity") +
  # Customize the fill colors for the plot
  scale_fill_manual(values = c("Cottonseed" = "blue",
                               "Cotton Wholesale" = "red")) +
  # Add plot title and axis labels
  labs(title = "Histograms: Price Distribution ", x = "Price", y = "Frequency") +
  # Apply a minimal theme for a clean and professional appearance
  theme_minimal()

#### Optional ----
# Save the plot as a PNG file with specified dimensions if needed
#ggsave("Histogram.png", plot = Histogram, width = 8, height = 5)

### Line Plot ---
# Plotting price trends over time
ggplot(data, aes(x = Year)) +
  geom_line(aes(y = Cotton_Wholesale_Price, color = "Cotton Price")) +
  geom_line(aes(y = Cottonseed_Price, color = "Cottonseed Price")) +
  # Customize the fill colors for the plot
  scale_fill_manual(values =c("Cottonseed"=" blue","Cotton Wholesale"= "red")) +
  # Add plot title and axis labels
  labs(title = "Price Trends of Cotton and Cottonseed (1910-1938)",
       x = "Year", y = "Price (in USD)") +
  # Apply a minimal theme for a clean and professional appearance
  theme_minimal()

#### Optional ----
# Save the plot as a PNG file with specified dimensions if needed
#ggsave("Line_Plot.png", plot = LinePlot, width = 8, height = 5)

## Anomaly Detection
### Identify Outliers Using Z-Scores
anamoly_data <- data %>%
  mutate(Cottonseed_Z = (Cottonseed_Price - mean(Cottonseed_Price)) /
           sd(Cottonseed_Price),
         Cotton_Z = (Cotton_Wholesale_Price - mean(Cotton_Wholesale_Price)) /
           sd(Cotton_Wholesale_Price))

outliers <- anamoly_data %>% filter(abs(Cottonseed_Z) > 2 | abs(Cotton_Z) > 2)
summary_outliers <- summary(outliers)
knitr::kable( outliers,
              caption = "Outliers in Cotton and Cottonseed Prices (1910-1938)")
## Highlighting outliers on the Line Plot
ggplot(anamoly_data, aes(x = Year)) +
  geom_line(aes(y = Cottonseed_Price, color = "Cottonseed Price")) +
  geom_line(aes(y = Cotton_Wholesale_Price, color = "Cotton Price")) +
  geom_point(data = outliers, aes(x = Year, y = Cottonseed_Price),
             color = "yellow", size = 3) +
  # Add plot title and axis labels
  labs(title = "Anomalies in Cotton and Cottonseed Prices", y = "Price (USD)") +
  # Apply a minimal theme for a clean and professional appearance
  theme_minimal()

#### Optional ----
# Save the plot as a PNG file with specified dimensions if needed
#ggsave("Anomalies_Plot.png", plot = AnomaliesPlot, width = 8, height = 5)

# Time Series Decomposition
```

```
ts_data <- ts(data$Cottonseed_Price, start = c(1910, 1), frequency = 12)
ts_decomp <- decompose(ts_data)
# Plotting Additive Time Series Decomposition -- decompose.ts
plot(ts_decomp)
# Test for Association/Correlation Between Paired Samples
## Test for association between paired samples- CottonSeed & Cotton Price
cor_test <- cor.test(data$Cottonseed_Price, data$Cotton_Wholesale_Price)
cor_test
# Predictive Modeling-- Linear Regression
###Uses Fitting Linear Model(lm) to calculate
model <- lm(Cotton_Wholesale_Price ~ Cottonseed_Price, data = data)
summary(model)
```