Final Project – Milk Cows & Production

Ander Eguiluz

Bellevue University

Milk Cows and Production

Business Problem

The U.S. dairy industry is a critical part of the agricultural economy, contributing significantly to the national output. As milk production continues to evolve, understanding long-term trends is essential for forecasting future production, optimizing resource allocation, and informing policy decisions. This project focuses on analyzing trends in both milk cow populations and milk production across different U.S. regions from 1970 to the present. By identifying changes in cow populations and production efficiency, the project provides insights into regional disparities and overall production growth. These insights will help dairy farmers, policymakers, and supply chain businesses prepare for future market conditions and resource needs.

Background/History

Milk production has been central to U.S. agriculture since the 19th century. Technological advances in milking equipment, improved feed, and more efficient herd management have led to steady increases in milk production per cow over the past few decades. Meanwhile, changing climate conditions and shifts in economic priorities have influenced regional disparities in milk cow populations. This project examines how these factors have impacted both milk cow populations and total milk production across various U.S. regions. By understanding historical trends, stakeholders in the dairy industry can better anticipate future shifts and address challenges related to climate change, economic pressure, and evolving consumer demand.

Data Explanation

The dataset used in this project is sourced from data.gov and the USDA Economic Research Service. It consists of two main components: milk cow population data and milk production data. The milk cow data includes information on the number of cows across different U.S. regions, recorded annually. The milk production data details the average milk produced per cow, as well as the total milk production in million pounds. In preparation for analysis, irrelevant columns, such as "Percent of U.S. milk," were removed. Rows with missing or non-numeric values were dropped to maintain data integrity. The data was then reshaped to a long format, allowing for time series analysis of milk cow populations and production trends. This restructuring enabled more detailed year-by-year comparisons across regions and states, ensuring the data was ready for forecasting and analysis.

Methods

To analyze both milk cow populations and milk production trends, I employed several statistical techniques. First, time series models were used to forecast future trends. For both milk cow data and production data, the ARIMA (AutoRegressive Integrated Moving Average) model was applied, given its effectiveness in handling time series data. Additionally, correlation heatmaps were generated to explore relationships between the number of milk cows and total milk production. Data visualizations, such as line charts, were created to highlight trends in the number of cows over time, and these visualizations were paired with production trends to understand how the two variables interact. In the preprocessing stage, steps such as data cleaning, reshaping, and handling missing values ensured that the dataset was well-prepared for analysis.

Analysis

The analysis revealed significant findings in both milk production and milk cow trends across the U.S. Over the past several decades, there has been a steady increase in milk production, driven in part by improvements in cow productivity. However, regional differences in milk cow populations have also played a critical role. The data shows that while some regions, such as the Northeast and Lake States, have experienced consistent growth in both cow populations and production, others, such as the Corn Belt, have seen more variable trends. The time series forecasts indicate that milk production is expected to continue growing, with the number of cows stabilizing or decreasing in some regions due to efficiency gains per cow. The residual analysis from the ARIMA model also highlighted potential outliers in production data, which may be explained by sudden shifts in dairy farming practices or economic pressures in certain regions.

Code

Milk Production:

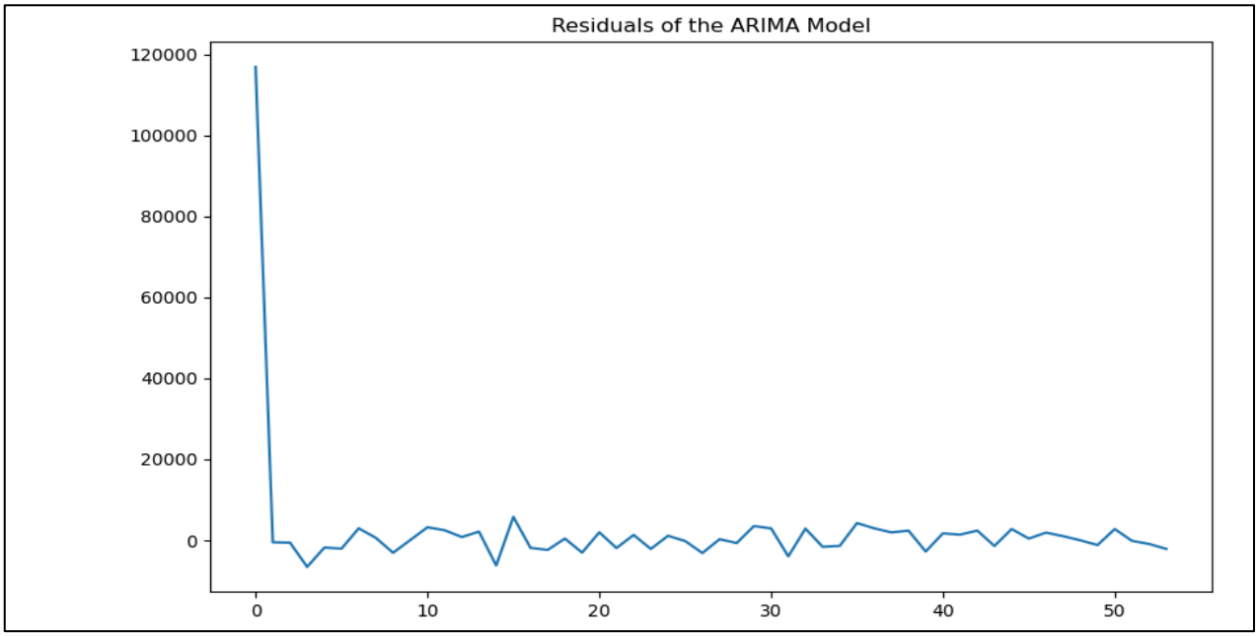
```
In [24]: import pandas as pd

         # Loading data from the "Milk production" sheet
         milk_production = pd.read_excel('Milk Cows and Production.xlsx', sheet_name='Milk production', skiprows=1)

         # Identifying and dropping columns that contain "Percent of U.S. milk" in any row
         columns_to_drop = milk_production.columns[milk_production.apply(lambda x: x.astype(str).str.contains('Percent of U.S. milk production').any())]

         # Dropping identified columns
         milk_production_cleaned = milk_production.drop(columns=columns_to_drop)

         # Dropping the first column
         milk_production_cleaned = milk_production_cleaned.drop(milk_production_cleaned.columns[0], axis=1)

         # Dropping rows that contain NaN values
         milk_production_cleaned = milk_production_cleaned.dropna()

         # Dropping rows that have "Million pounds" in any cell
         milk_production_cleaned = milk_production_cleaned[~milk_production_cleaned.apply(lambda row: row.astype(str).str.contains('Million pounds').any(), axis=1)]

         # Removeing regional rows
         regions = ["Northeast", "Lake States", "Corn Belt", "Northern Plains", "Appalachia", "Southeast", "Southern Plains", "Mountain", "Delta States", "West Coast", "Other St
         milk_production_cleaned = milk_production_cleaned[~milk_production_cleaned.iloc[:, 0].isin(regions)]

         # Removing the "United States" row
         milk_production_cleaned = milk_production_cleaned[milk_production_cleaned.iloc[:, 0] != "United States"]

         # Resetting the index
         milk_production_cleaned.reset_index(drop=True, inplace=True)

         # Displaying the cleaned data
         print(milk_production_cleaned.head())

         # Reshaping the data to long format
         df_long = pd.melt(milk_production_cleaned, id_vars=['Back to content page.'], var_name='Year', value_name='Milk Production')

         # Renaming the 'Back to content page.' column to 'State'
         df_long = df_long.rename(columns={'Back to content page.': 'State'})

         # Converting Year to integer
         df_long['Year'] = df_long['Year'].astype(int)

         # Ensuring 'Milk Production' is numeric
         df_long['Milk Production'] = pd.to_numeric(df_long['Milk Production'], errors='coerce')

         # Dropping any remaining NaN values that may have resulted from the conversion
         df_long = df_long.dropna()

         # Displaying the reshaped DataFrame
         print(df_long.head())
```

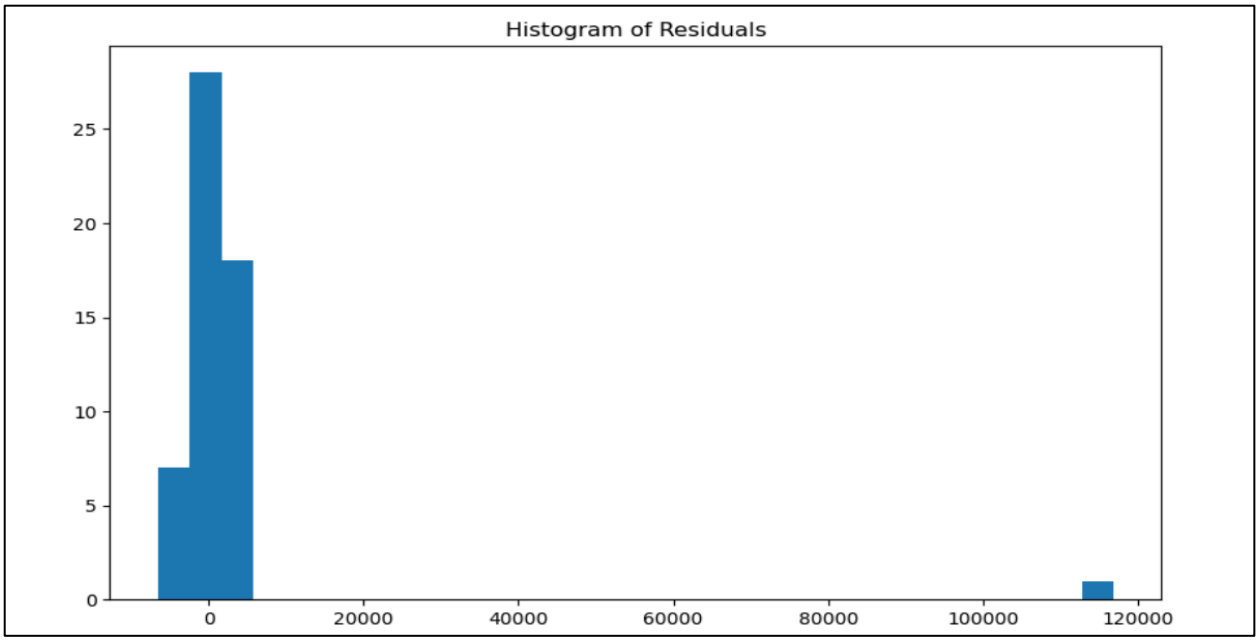
```
  Back to content page.  1970  1971  1972  1973  1974  1975  1976  1977  1978  \
0                 Maine   619   629   638   614   611   629   628   638   641
1         New Hampshire   356   359   353   337   329   336   335   339   341
2               Vermont  1970  2025  2039  1948  1945  2009  2093  2100  2136
3         Massachusetts   658   658   629   595   593   601   598   597   572
4          Rhode Island    75    70    69    64    63    63    60    57    55

    ...  2014  2015  2016  2017  2018  2019  2020  2021  2022  2023
0   ...   599   594   630   630   618   621   591   571   554   543
1   ...   282   282   287   273   249   239   236   227   219   208
2   ...  2666  2666  2727  2729  2683  2697  2603  2567  2554  2536
3   ...   233   217   221   211   202   193   200   195   188   179
4   ...  17.1  15.9  14.1    13  11.5  10.6  10.9  10.1    10   9.7

[5 rows x 55 columns]
          State  Year  Milk Production
0         Maine  1970            619.0
1  New Hampshire  1970            356.0
2       Vermont  1970           1970.0
3  Massachusetts  1970            658.0
4   Rhode Island  1970             75.0
```

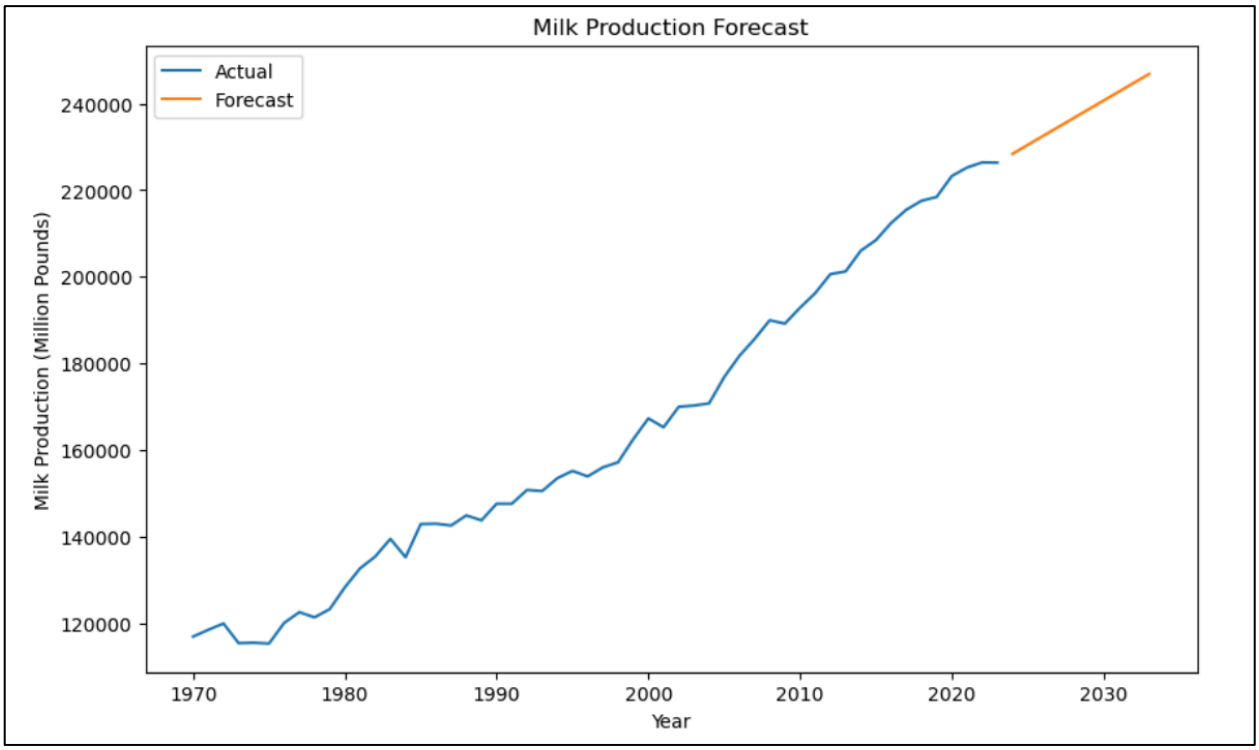
In [25]:
```
import statsmodels.api as sm
from statsmodels.tsa.arima.model import ARIMA
import matplotlib.pyplot as plt

# Aggregating data by year
annual_production = df_long.groupby('Year')['Milk Production'].sum().reset_index()

# Ensuring the 'Milk Production' column is numeric
annual_production['Milk Production'] = pd.to_numeric(annual_production['Milk Production'])

# Defining the ARIMA model
model = ARIMA(annual_production['Milk Production'], order=(1, 1, 1))

# Fitting the model
model_fit = model.fit()

# Summary of the model
print(model_fit.summary())
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:        Milk Production   No. Observations:                   54
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -490.576
Date:                Sun, 28 Jul 2024   AIC                            987.153
Time:                        19:00:45   BIC                            993.064
Sample:                             0   HQIC                           989.426
                                  - 54
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.0000      0.007    146.995      0.000       0.987       1.013
ma.L1         -0.9991      0.157     -6.368      0.000      -1.307      -0.692
sigma2      6.345e+06   1.67e-08    3.8e+14      0.000    6.35e+06    6.35e+06
===================================================================================
Ljung-Box (L1) (Q):                   1.92   Jarque-Bera (JB):                 1.21
Prob(Q):                              0.17   Prob(JB):                         0.55
Heteroskedasticity (H):               0.38   Skew:                            -0.37
Prob(H) (two-sided):                  0.05   Kurtosis:                         2.98
===================================================================================
```

In [26]:
```
# Plotting residual errors
residuals = model_fit.resid
plt.figure(figsize=(10, 6))
plt.plot(residuals)
plt.title('Residuals of the ARIMA Model')
plt.show()

# Plotting density of residuals
plt.figure(figsize=(10, 6))
plt.hist(residuals, bins=30)
plt.title('Histogram of Residuals')
plt.show()

# Forecasting future values
forecast = model_fit.forecast(steps=10)
print(forecast)

# Plotting actual vs predicted
plt.figure(figsize=(10, 6))
plt.plot(annual_production['Year'], annual_production['Milk Production'], label='Actual')
plt.plot(range(annual_production['Year'].iloc[-1] + 1, annual_production['Year'].iloc[-1] + 11), forecast, label='Forecast')
plt.xlabel('Year')
plt.ylabel('Milk Production (Million Pounds)')
plt.title('Milk Production Forecast')
plt.legend()
plt.show()
```




Histogram of Residuals
25
20
15
10
5
0
0
20000
40000
60000
80000
100000
120000


Milk Production Forecast
Actual
Forecast
240000
220000
200000
180000
160000
140000
120000
Milk Production (Million Pounds)
1970
1980
1990
2000
2010
2020
2030
Year

Milk Cows:

```
import pandas as pd

# Loading data from the "Milk cows" sheet
milk_cows = pd.read_excel('Milk Cows and Production.xlsx', sheet_name='Milk cows', skiprows=1)

# Dropping the first column
milk_cows_cleaned = milk_cows.drop(milk_cows.columns[0], axis=1)

# Dropping rows that contain NaN values
milk_cows_cleaned = milk_cows_cleaned.dropna()

# Removing regional rows
regions = ["Northeast", "Lake States", "Corn Belt", "Northern Plains", "Appalachia", "Southeast", "Southern Plains", "Mountain
milk_cows_cleaned = milk_cows_cleaned[~milk_cows_cleaned.iloc[:, 0].isin(regions)]

# Removing the "United States" row
milk_cows_cleaned = milk_cows_cleaned[milk_cows_cleaned.iloc[:, 0] != "United States"]

# Resetting the index
milk_cows_cleaned.reset_index(drop=True, inplace=True)

# Displaying the cleaned data
print(milk_cows_cleaned.head())

# Reshaping the data to long format
df_long = pd.melt(milk_cows_cleaned, id_vars=['Back to content page.'], var_name='Year', value_name='Milk Cows')

# Renaming the 'Back to content page.' column to 'State'
df_long = df_long.rename(columns={'Back to content page.': 'State'})

# Converting Year to integer
df_long['Year'] = df_long['Year'].astype(int)

# Ensuring 'Milk Production' is numeric
df_long['Milk Cows'] = pd.to_numeric(df_long['Milk Cows'], errors='coerce')

# Dropping any remaining NaN values that may have resulted from the conversion
df_long = df_long.dropna()
```

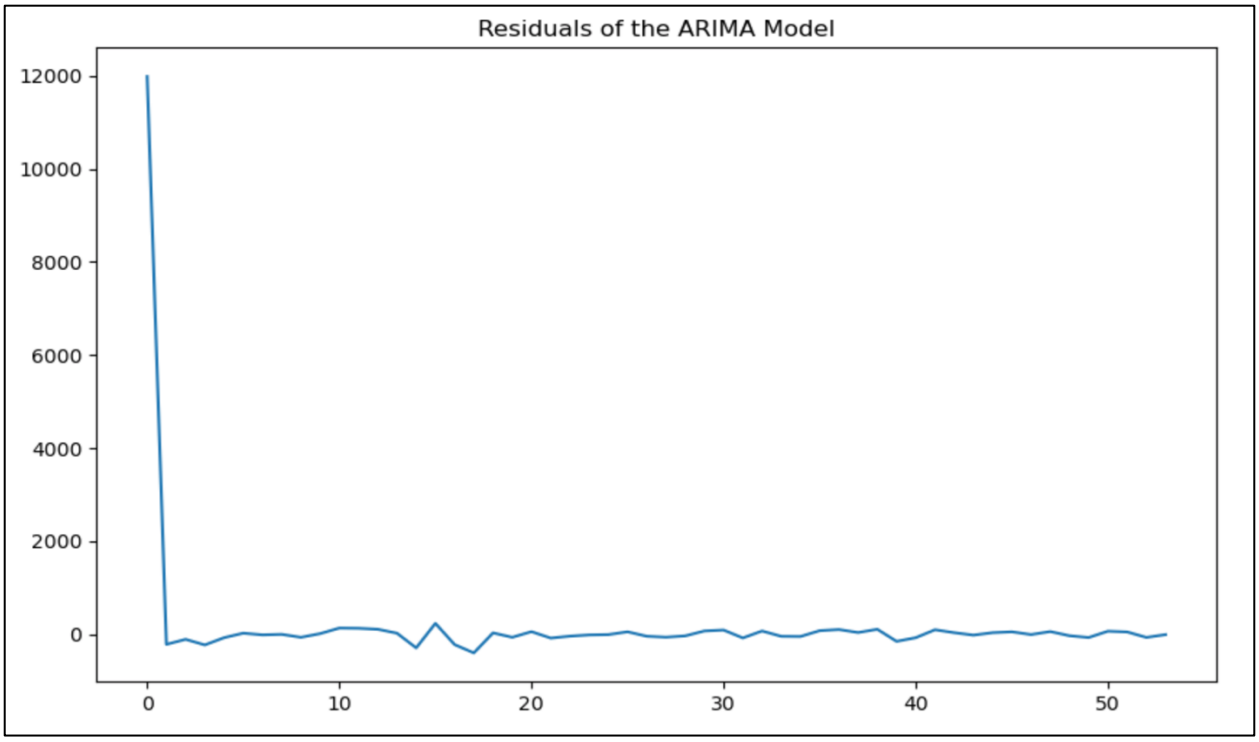
```
# Displaying the reshaped DataFrame
print(df_long.head())
```

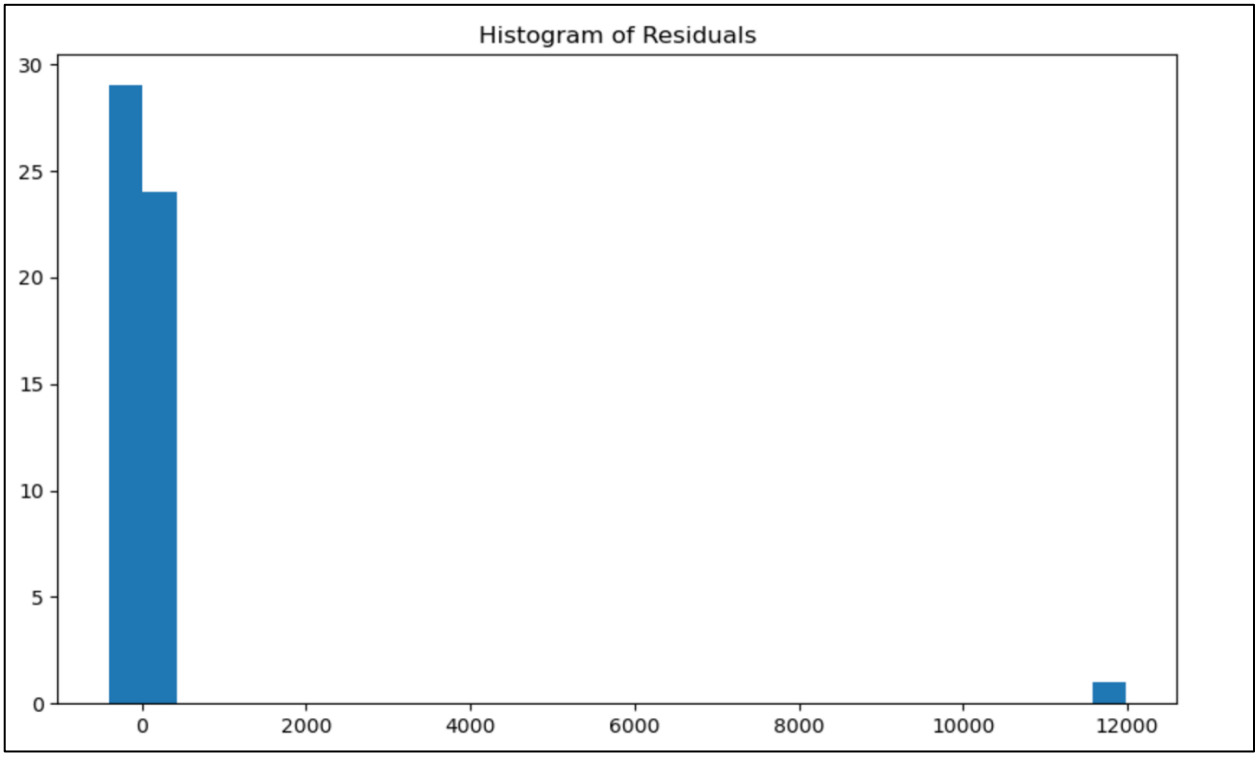
```
  Back to content page.   1970   1971   1972   1973   1974   1975   1976  \
0                 Maine   62.0   61.0   61.0   60.0   60.0   61.0   59.0
1         New Hampshire   36.0   35.0   33.0   33.0   32.0   33.0   31.0
2               Vermont  194.0  195.0  195.0  191.0  193.0  193.0  194.0
3         Massachusetts   60.0   59.0   57.0   55.0   54.0   55.0   54.0
4          Rhode Island    6.9    6.3    6.2    5.9    5.9    6.0    5.5

    1977   1978  ...   2014   2015   2016   2017   2018   2019   2020   2021  \
0   58.0   58.0  ...   30.0   30.0   30.0   30.0   30.0   29.0   27.0   27.0
1   31.0   31.0  ...   14.0   14.0   14.0   13.0   12.0   11.0   11.0   11.0
2  191.0  186.0  ...  132.0  132.0  130.0  129.0  127.0  126.0  122.0  120.0
3   51.0   49.0  ...   13.0   12.0   12.0   12.0   11.0   10.0   10.0   10.0
4    5.0    4.7  ...    0.9    0.9    0.8    0.8    0.7    0.6    0.5    0.5

    2022   2023
0   26.0   25.0
1   10.0   10.0
2  118.0  117.0
3    9.0    9.0
4    0.5    0.5

[5 rows x 55 columns]
               State  Year  Milk Cows
0              Maine  1970       62.0
1      New Hampshire  1970       36.0
2            Vermont  1970      194.0
3      Massachusetts  1970       60.0
4       Rhode Island  1970        6.9
```

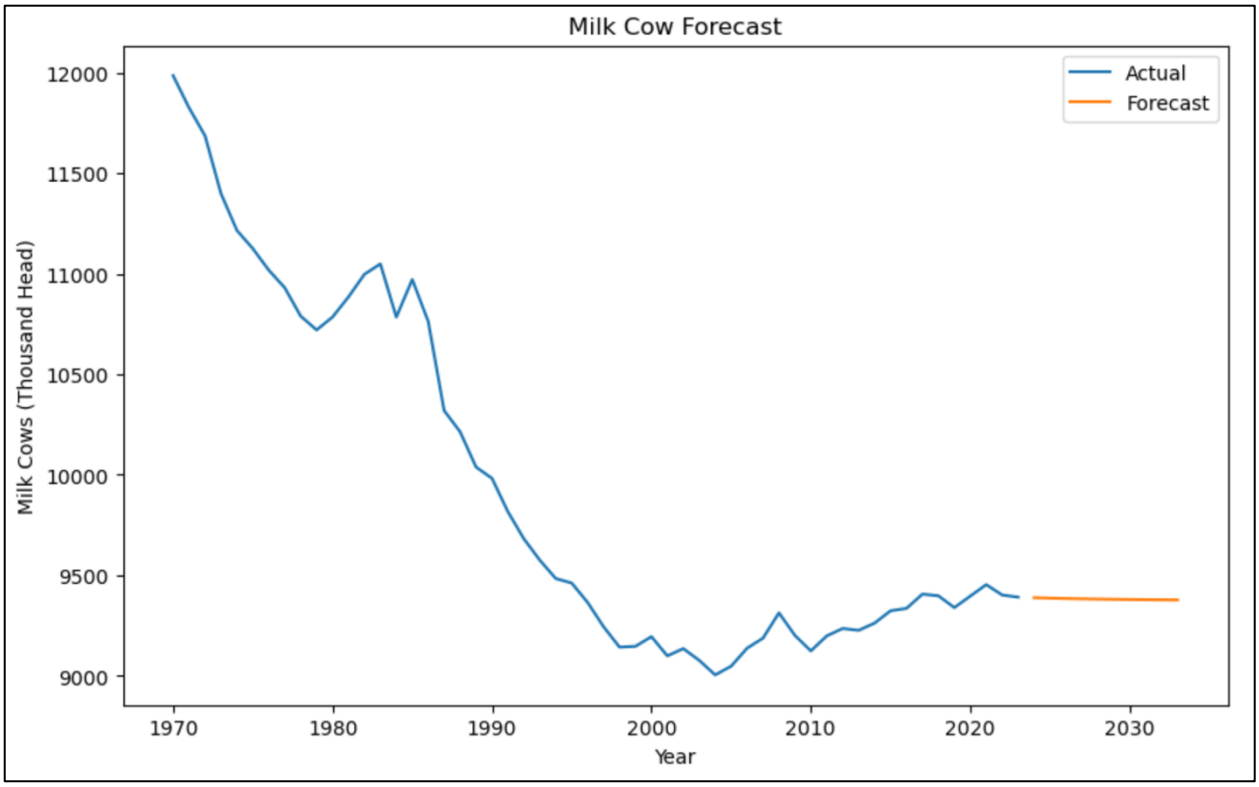
```python
import statsmodels.api as sm
from statsmodels.tsa.arima.model import ARIMA
import matplotlib.pyplot as plt

# Aggregating data by year
annual_cows = df_long.groupby('Year')['Milk Cows'].sum().reset_index()

# Ensuring the 'Milk Production' column is numeric
annual_cows['Milk Cows'] = pd.to_numeric(annual_cows['Milk Cows'])

# Defining the ARIMA model
model = ARIMA(annual_cows['Milk Cows'], order=(1, 1, 1))

# Fitting the model
model_fit = model.fit()

# Summary of the model
print(model_fit.summary())
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Milk Cows   No. Observations:                   54
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -325.476
Date:                Wed, 18 Sep 2024   AIC                            656.952
Time:                        18:07:30   BIC                            662.863
Sample:                             0   HQIC                           659.225
                                  - 54
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.8606      0.175      4.918      0.000       0.518       1.204
ma.L1         -0.6202      0.217     -2.855      0.004      -1.046      -0.194
sigma2      1.183e+04   1477.233      8.009      0.000    8936.536    1.47e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                18.70
Prob(Q):                              0.82   Prob(JB):                         0.00
Heteroskedasticity (H):               0.17   Skew:                            -1.04
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.04
===================================================================================
```

```python
# Plotting residual errors
residuals = model_fit.resid
plt.figure(figsize=(10, 6))
plt.plot(residuals)
plt.title('Residuals of the ARIMA Model')
plt.show()

# Plotting density of residuals
plt.figure(figsize=(10, 6))
plt.hist(residuals, bins=30)
plt.title('Histogram of Residuals')
plt.show()

# Forecasting future values
forecast = model_fit.forecast(steps=10)
print(forecast)

# Plotting actual vs predicted
plt.figure(figsize=(10, 6))
plt.plot(annual_cows['Year'], annual_cows['Milk Cows'], label='Actual')
plt.plot(range(annual_cows['Year'].iloc[-1] + 1, annual_cows['Year'].iloc[-1] + 11), forecast, label='Forecast')
plt.xlabel('Year')
plt.ylabel('Milk Cows (Thousand Head)')
plt.title('Milk Cow Forecast')
plt.legend()
plt.show()
```


Residuals of the ARIMA Model
12000
10000
8000
6000
4000
2000
0
0
10
20
30
40
50


Histogram of Residuals
30
25
20
15
10
5
0
0
2000
4000
6000
8000
10000
12000



Conclusion

The findings from this project confirm the long-standing growth in U.S. milk production, with notable regional disparities in both milk cow populations and production rates. The ARIMA model's forecast suggests that milk production will continue to increase, driven by both regional trends and improvements in production efficiency. However, the forecast for milk cow populations shows varying trends across regions, with some areas experiencing declines due to rising production per cow. This analysis provides valuable insights for stakeholders in the dairy industry, allowing for more informed decision-making regarding resource allocation and strategic planning. Further refinement of the models, particularly to address outliers, may be necessary to improve accuracy and provide even more actionable insights.

Assumptions

Several assumptions were made throughout this analysis. First, the accuracy and reliability of the data sourced from data.gov and the USDA Economic Research Service are assumed to be high. It is also assumed that the analysis captures the key factors driving milk production trends, despite not including exogenous variables such as feed costs, technological advancements, or policy changes, which could have significant impacts on future trends. Additionally, it was assumed that historical patterns in milk cow populations and production can be used to forecast future trends, though sudden shifts in technology, policy, or climate could alter these patterns.

Limitations

This project is limited by the exclusion of regional aggregate data, which could have provided a broader understanding of U.S. dairy trends. Another limitation is the ARIMA model's sensitivity to outliers, which were present in both the milk cow and production data. These outliers likely represent sudden changes in dairy farming practices or economic conditions that the model did not fully capture. Additionally, the analysis is descriptive, meaning that it focuses on identifying trends rather than exploring causal relationships between variables. To fully understand the factors driving changes in milk cow populations and production, further analysis would be needed, including consideration of additional variables.

Challenges

Several challenges were encountered during this project, particularly with respect to the data itself. Inconsistent and missing values in the historical dataset required careful cleaning to ensure the accuracy of the analysis. Additionally, the presence of outliers in both milk cow and production data posed challenges for accurately modeling trends. Investigating these outliers

required additional effort to determine whether they were genuine reflections of sudden shifts in the dairy industry or anomalies within the data. Finally, balancing the different trends in production efficiency and cow population changes across regions required careful analysis to ensure that regional differences were appropriately captured.

Future Uses/Additional Applications

This analysis has the potential to be extended beyond its current scope. Future research could incorporate exogenous variables, such as feed costs, technological advancements, or economic policies, to better understand the drivers of milk production trends. Applying more advanced forecasting techniques, such as machine learning models, could also help capture the finer variations missed by the ARIMA model. Additionally, the dataset could be used to explore causal relationships between dairy farming practices, cow populations, and production growth, providing deeper insights into the factors influencing the U.S. dairy industry. These applications could prove beneficial for policymakers and businesses looking to make data-driven decisions.

Recommendations

Based on the findings from this project, several recommendations can be made. First, dairy producers should focus on scaling up resources and optimizing their supply chains to accommodate the expected increase in milk production. In regions where cow populations are declining, farmers should prioritize efficiency gains per cow to maintain production levels. Policymakers should also consider implementing policies that support sustainable dairy farming practices, ensuring that production growth is achieved without compromising environmental and economic stability. Finally, refining the forecasting models by incorporating additional variables, such as feed costs or technological changes, could help improve the accuracy of predictions and further support decision-making.

Implementation Plan

To implement the recommendations outlined in this report, several steps should be taken. In the short term, additional data on factors such as feed costs and technological advancements should be integrated into the analysis to refine the model's accuracy. In the long term, more advanced forecasting techniques, such as machine learning models, should be explored. Collaboration with key stakeholders—such as farmers, economists, and policymakers—will be crucial to ensure that the insights from this analysis are applied effectively in real-world decision-making. By aligning the findings with industry trends, the dairy sector can better prepare for future challenges and opportunities.

Ethical Assessment

Several ethical considerations must be addressed in this project. First, ensuring the accuracy and transparency of the data used in the analysis is critical to avoid drawing misleading conclusions. Handling the data responsibly is essential to prevent biases in interpretation, ensuring that the analysis remains objective. Additionally, the broader impact of the findings on the dairy industry, particularly on small-scale farmers, must be considered. Regular reviews of the methodology and results should be conducted to minimize biases and ensure that the project provides fair and actionable insights to all stakeholders, promoting sustainability and equity in the dairy sector.

References

U.S. Department of Agriculture. *Dairy Data.* Economic Research Service. Retrieved from https://www.ers.usda.gov/data-products/dairy-data/