Forecasting Sunflower Seeds and Oil Production in Kazakhstan Using Regression and Time Series Models

Azhar Idrisheva

May 2, 2025

Capstone Thesis, Spring 2025
Department of Mathematics, School of Sciences and Humanities
Nazarbayev University
Supervisor: Dr. Piotr Skrzypacz
Second reader: Dr. Dongming Wei

Abstract

Kazakhstan is considered a major global producer of sunflower seeds, ranking among top 10 countries worldwide. Sunflower cultivation and sunflower oil production experience a rapid growth in the country, holding a potential to greatly benefit the country's economic and agricultural sectors. This capstone project aims to develop statistical models to predict sunflower yields and sunflower oil production until the year 2035. Using historical data, the project uses regression and time series analysis to identify patterns and impact of variables. By leveraging these findings, the project aims to support effective decision-making in agricultural sector to improve sustainable practices and enabling the effective scaling of sunflower production in Kazakhstan.

1 Introduction

The objective of this capstone project is to predict the production of sunflower seeds and oil in Kazakhstan for the next decade. The analysis is based on historical data, including the volume of sunflower seeds harvested, export and import volumes, prices of sunflower products, and the monetary value of trade in sunflower oil and seeds.

Three different predictive modeling techniques will be implemented: a Multiple Polynomial Regression Model, an ARIMA Model, and a LOESS Regression Model. This project aims to identify the strengths and limitations of each approach when applied to Kazakhstan's agricultural context. By evaluating these models' performance, this project shows how different approaches may give unique insights into complex data.

This work contributes to ongoing efforts to increase the efficiency of Kazakhstan's agricultural sector by applying data-driven methods to a real-world case. Forecasting will help optimize resource allocation, improve supply chain management, and help the country acquire the position of a major and reliable exporter. The results may serve as a foundation for further research and support better decisions in the agricultural industry.

Year	Rank
2000-2010	from 12th–13th to top 10
2015-2020	7th - 8th
2021-2022	7th
2023	5th-6th

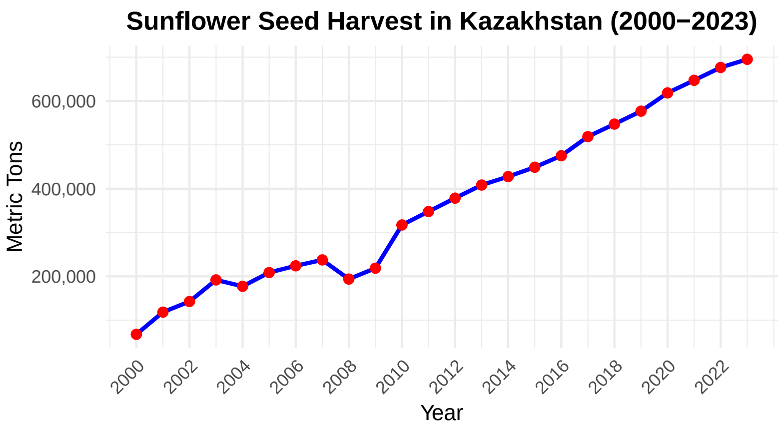
Table 1:
Kazakhstan Rankings according to USDA and FAOSTAT

2 Dataset Description

The data covers 2000 to 2023 and was sourced from the National Bureau of Statistics of the Republic of Kazakhstan and international databases such as FAOSTAT and World Bank.

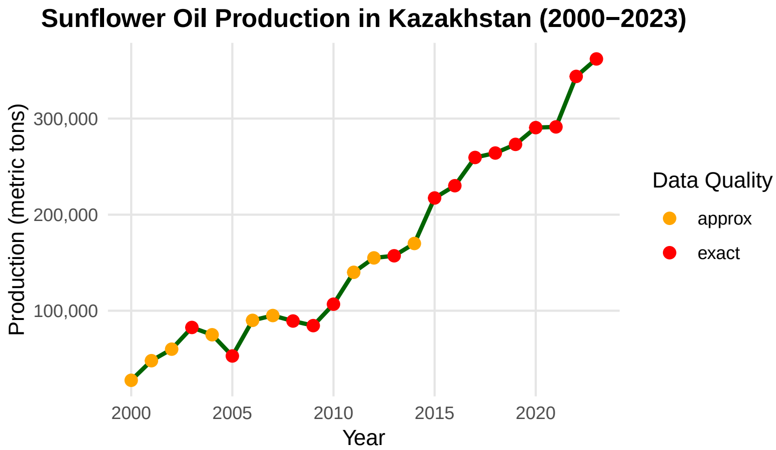
Year	Amount	Year	Amount
2000	67 800	2012	378 400
2001	118 400	2013	408 300
2002	142 600	2014	427 600
2003	191 900	2015	448 900
2004	177 300	2016	475 200
2005	208 700	2017	518 700
2006	224 100	2018	547 300
2007	237 500	2019	576 800
2008	193 800	2020	618 400
2009	218 600	2021	647 200
2010	317 200	2022	676 500
2011	347 900	2023	695 100

Table 2: Annual data on the quantity of sunflower seeds harvested in Kazakhstan (in metric tons) from 2000 to 2023



Year	Amount	Year	Amount
2000	∼ 27 600	2012	∼ 155 000
2001	∼ 48 000	2013	157 200
2002	∼ 60 000	2014	∼ 170 000
2003	82 600	2015	217 373
2004	∼ 75 000	2016	230 141
2005	52 900	2017	259 521
2006	∼ 90 000	2018	264 150
2007	∼ 95 000	2019	273 159
2008	89 300	2020	290 600
2009	84 500	2021	291 360
2010	106 800	2022	343 900
2011	∼ 140 000	2023	362 100

Table 3: Annual data on the quantity of sunflower oil produced in Kazakhstan (in metric tons) from 2000 to 2023



Note: Not all data on the amounts of sunflower oil produced are available for specific years. For this reason, data on sunflower harvest was multiplied by the average oil yield coefficient (0.42-0.45 for Kazakhstan) and adjusted by the percent of processing capacity. This estimate also takes underutilization of capacities in the early years, losses during storage, and export of seeds without processing into account.

Year	Volume (t)	Avg Price ($/t)	Top 3 Destinations (% share)
2000	4 150	480	Uzbekistan (62%), Russia (28%), Kyrgyzstan (7%)
2001	6 820	510	Uzbekistan (58%), Russia (30%), Tajikistan (9%)
2002	9 350	535	Russia (51%), Uzbekistan (33%), Afghanistan (11%)
2003	12 100	580	Russia (49%), Uzbekistan (27%), Tajikistan (18%)
2004	15 300	610	Russia (47%), Tajikistan (29%), Uzbekistan (21%)
2005	18 200	620	Russia (51%), Tajikistan (33%), Mongolia (8%)
2006	22 400	655	Russia (44%), China (23%), Tajikistan (22%)
2007	27 800	725	China (38%), Russia (35%), Afghanistan (15%)
2008	19 600	1020	China (41%), Afghanistan (29%), Russia (17%)
2009	24 300	880	China (45%), Afghanistan (26%), Uzbekistan (14%)
2010	42 700	890	Afghanistan (39%), China (27%), Uzbekistan (19%)
2011	67 500	950	China (43%), Afghanistan (22%), EU (15%)
2012	83 200	1100	China (46%), EU (21%), Afghanistan (18%)
2013	97 800	1075	China (49%), EU (23%), Iran (12%)
2014	104 600	1150	China (51%), EU (22%), Uzbekistan (11%)
2015	112 500	1025	China (54%), Uzbekistan (23%), EU (10%)
2016	134 200	1080	China (57%), Uzbekistan (20%), Iran (9%)
2017	158 700	1125	China (59%), Uzbekistan (17%), EU (12%)
2018	187 200	1150	EU (38%), China (31%), Uzbekistan (14%)
2019	203 800	1210	EU (42%), China (33%), Iran (10%)
2020	213 400	1320	China (47%), Iran (19%), EU (17%)
2021	254 600	1425	China (53%), EU (18%), Uzbekistan (13%)
2022	286 100	1540	China (52%), EU (21%), Uzbekistan (15%)
2023	317 500	1490	China (61%), Uzbekistan (17%), EU (11%)

Table 4: Kazakhstan Sunflower Oil Exports from 2000 to 2023

Year	Volume (t)	Value (in 1000 USD)	Main Source Countries (% Share)
2000	4 850	3 400	Russia (60%), Ukraine (35%)
2001	5 200	3 600	Russia (62%), Ukraine (33%)
2002	6 100	4 300	Russia (65%), Ukraine (30%)
2003	7 500	5 200	Russia (60%), Ukraine (30%), Argentina (5%)
2004	9 800	6 900	Russia (65%), Ukraine (30%)
2005	14 500	10 200	Russia (70%), Ukraine (25%)
2006	18 300	13 100	Russia (72%), Ukraine (23%)
2007	22 000	16 500	Russia (75%), Ukraine (20%)
2008	25 400	19 800	Russia (78%), Ukraine (17%)
2009	20 100	14 200	Russia (80%), Ukraine (15%)
2010	34 700	24 900	Russia (75%), Ukraine (20%), Belarus (3%)
2011	38 200	28 600	Russia (78%), Ukraine (18%)
2012	42 500	32 000	Russia (80%), Ukraine (15%)
2013	45 100	34 200	Russia (82%), Ukraine (13%)
2014	48 300	36 800	Russia (85%), Ukraine (10%)
2015	50 200	39 500	Russia (88%), Ukraine (7%)
2016	52 000	41 000	Russia (90%), Ukraine (5%)
2017	58 400	46 500	Russia (92%), Ukraine (3%)
2018	64 800	54 700	Russia (93%), Ukraine (2%)
2019	60 100	51 300	Russia (94%), Ukraine (1%)
2020	45 600	49 800	Russia (95%), Ukraine (1%)
2021	39 500	59 200	Russia (96%), Ukraine (1%)
2022	30 200	44 700	Russia (98%), Turkey (1%)
2023	24 500	39 800	Russia (99%)

Table 5: Kazakhstan Sunflower Oil Imports from 2000 to 2023

Indicator	Value
Average price per liter of sunflower oil	~$1,66
Average price per 100 grams of sunflower seeds	~$0,78

Table 6: Average sunflower oil and seed prices in Kazakhstan

3 Statistical Models

3.1 LOESS regression model

LOESS (locally estimated scatterplot smoothing) - a non-parametric regression technique that fits multiple simple models to localized subsets of the data.

1. **Local Weighted Least Squares Regression**

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} w_i(x_0) \cdot (y_i - \beta_0 - \beta_1(x_i - x_0))^2$$

- y_i - actual value of point i.
- β_0 and β_1 - intercept and slope of the local regression.
- $w_i(x_0)$ shows how much weight is given to each point i when fitting for x_0 (closer points have bigger w_i).

2. Tricube Weight Function

$$w_i(x_0) = \begin{cases} \left(1 - \left|\frac{x_i - x_0}{d(x_0)}\right|^3\right)^3, & \text{if } |x_i - x_0| < d(x_0) \\ 0, & \text{otherwise} \end{cases}$$

- This is how LOESS decides how much importance each point x_i has when estimating for x_0.
- $d(x_0)$ - distance from x_0 to the farthest data point inside the span window.
- If x_i is far from x_0 (beyond $d(x_0)$), its weight is 0, so it does not affect the local regression.
- The curve $\left(1 - \left|\frac{x_i - x_0}{d(x_0)}\right|^3\right)^3$ smoothly decreases weight as x_i gets further from x_0.

3. LOESS Estimate

$$\hat{y}(x_0) = \hat{\beta}_0$$

Since $(x_0 - x_0) = 0$, the prediction at x_0 is just the intercept $\hat{\beta}_0$ from the local regression. So for each x_0, we take $\hat{\beta}_0$ as the smoothed value.

3.2 Time series model

ARIMA(p,d,q) model (AutoRegressive Integrated Moving Average) equation:

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right) (1 - L)^d y_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

- p - order of autoregressive part
- d - degree of differencing
- q - order of moving average part
- L - lag operator
- ε_t - white noise

1. **Mean Absolute Percentage Error (MAPE)**: Measures average prediction accuracy as a percentage.

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

2. **Root Mean Squared Error (RMSE)**: Quantifies absolute deviation in the original units.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (A_t - F_t)^2}$$

where A_t = actual value, F_t = forecasted value, and n = number of observations.

3.3 Multiple Linear Regression Model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

- $\hat{y}$ - predicted value of the dependent variable.
- β_0 - y-intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ - coefficients for each independent variable $(x_1, x_2, \ldots, x_n)$, representing the change in $\hat{y}$ for a one-unit increase in the respective x, holding other x's constant.
- ϵ - error term.

$\text{Seed\_Production} = \beta_0 + \beta_1 \cdot \text{Year} + \beta_2 \cdot \text{Oil\_Production} + \beta_3 \cdot \text{Lagged\_Seed\_Production} + \beta_4 \cdot \text{Lagged\_Export\_Volume} + \beta_5 \cdot \text{Lagged\_Import\_Volume} + \beta_6 \cdot \text{Export\_Avg\_Price} + \beta_7 \cdot \text{Import\_Value} + \varepsilon$

$\text{Oil\_Amount} = \beta_0 + \beta_1 \cdot \text{Year} + \beta_2 \cdot \text{Year}^2 + \beta_3 \cdot \text{Seed\_Amount} + \beta_4 \cdot \text{Export\_Volume} + \beta_5 \cdot \text{Import\_Volume} + \beta_6 \cdot \text{Import\_Price\_Per\_Ton} + \beta_7 \cdot \text{Export\_Price\_Per\_Ton} + \beta_8 \cdot \text{Oil\_Price\_USD} + \beta_9 \cdot \text{Seed\_Price\_Per\_Ton\_USD} + \varepsilon$

4 Results and Discussion

4.1 ARIMA model for sunflower harvest

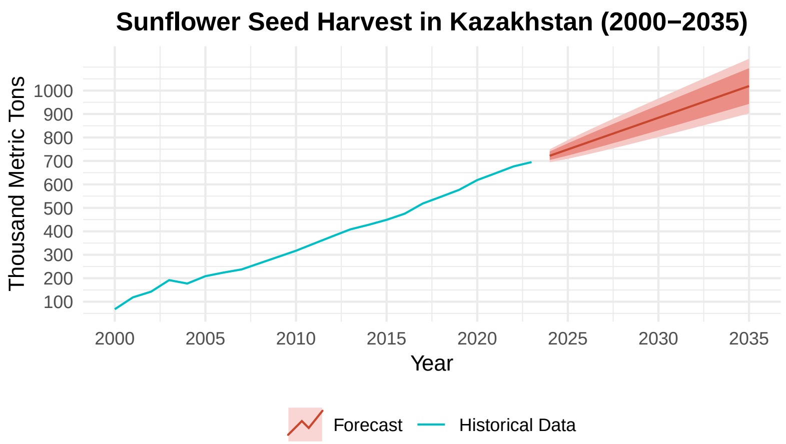
Model	**AIC**	**BIC**	**MAPE (%)**	**RMSE (Metric Tons)**
ARIMA(0,1,0)	478.3	481.1	3.5	12 450
ARIMA(1,1,1)	478.5	483.2	3.2	12 450
ARIMA(0,1,1)	480.3	484.1	4.1	14 800
ARIMA(1,1,0)	482.7	486.5	5.6	16 300

Table 7: Comparison of ARIMA Models

	ARIMA(0,1,0)	**ARIMA(1,1,1)**	**Difference**
Forecast (2035)	868,000 tons	872,000 tons	$<$ 0.5% variation
Drift Term	+15,049 tons/yr	+14,900 tons/yr	Nearly identical

Table 8:

ARIMA(0,1,0) is chosen since it has the lowest AIC and BIC and explains variance most efficiently. ADF test showed non-stationarity (p-value > 0.05), so we need first-order differencing to achieve stationarity. Residuals satisfied the Ljung-Box test with $p = 0.39 > 0.05$, failing to reject the null hypothesis of independence. This confirms the residuals are white noise (no autocorrelation), validating the model's adequacy for forecasting. The model's low MAPE and RMSE demonstrate its forecasting precision. While ARIMA(1,1,1) shows marginally lower MAPE and residual variance, its additional parameters were statistically insignificant ($p > 0.05$) and failed a likelihood ratio test against the simpler ARIMA(0,1,0) model.



Year	Forecast	95% Confidence Interval
2024	722 373.9	696 512.4 – 748 235.4
2025	749 647.8	713 074.2 – 786 221.5
2026	776 921.7	732 128.4 – 821 715.1
2027	804 195.7	752 472.7 – 855 918.6
2028	831 469.6	773 641.6 – 889 297.6
2029	858 743.5	795 396.1 – 922 090.9
2030	886 017.4	817 594.4 – 954 440.4
2031	913 291.3	840 144.0 – 986 438.6
2032	940 565.2	862 980.8 – 1 018 149.6
2033	967 839.1	886 058.0 – 1 049 620.3
2034	995 113.0	909 340.3 – 1 080 885.8
2035	1 022 387.0	932 800.2 – 1 111 973.7

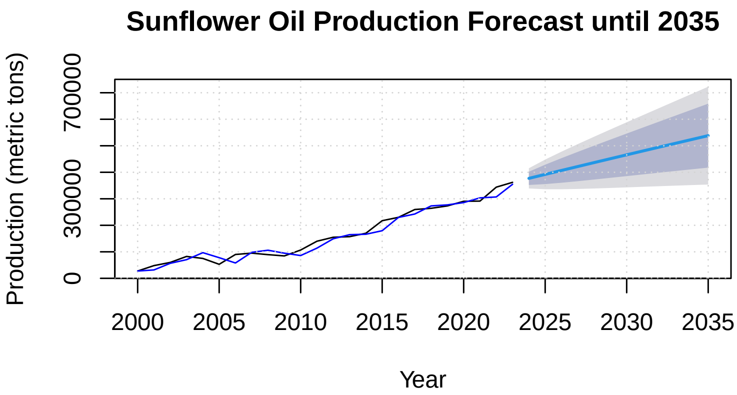
Table 9: ARIMA (0,1,0) tonne amounts forecasts and 95% Confidence Intervals (2024–2035)

4.2 ARIMA model for sunflower oil production

Model	AIC	BIC	MAPE (%)	RMSE (Metric Tons)
ARIMA(0,1,0)	527.67	528.80	13.01	21 749.71
ARIMA(1,1,0)	527.14	529.41	11.92	20 537.10
ARIMA(0,1,1)	527.84	530.11	12.11	20 873.81
ARIMA(1,1,1)	522.61	526.01	10.77	17 170.46

Table 10: Comparison of ARIMA Models

ADF test with $p = 0.6693 > 0.05$ showed non-stationarity, so first-order differencing is needed to achieve stationarity. Residuals satisfied the Ljung-Box test with $p = 0.39 > 0.05$, failing to reject the null hypothesis of independence. This confirms the residuals are white noise (no autocorrelation), validating the model's adequacy for forecasting.



Year	Forecast	95% Confidence Interval
2024	376 643.5	342 984.1 – 410302.8
2025	391 187.0	343 585.5 – 438 788.5
2026	405 730.4	347 430.7 – 464 030.1
2027	420 273.9	352 955.2 – 487 592.6
2028	434 817.4	359 552.8 – 510 082.0
2029	449 360.9	366 912.7 – 531 809.1
2030	463 904.3	374 850.1 – 552 958.6
2031	478 447.8	383 244.8 – 573 650.8
2032	492 991.3	392 013.3 – 593 969.3
2033	507 534.8	401 094.6 – 613 975.0
2034	522 078.3	410 442.9 – 633 713.7
2035	536 621.7	420 022.4 – 653 221.1

Table 11: AUTOARIMA Forecast of sunflower oil production in Kazakhstan from 2024 to 2035 (in metric tons)

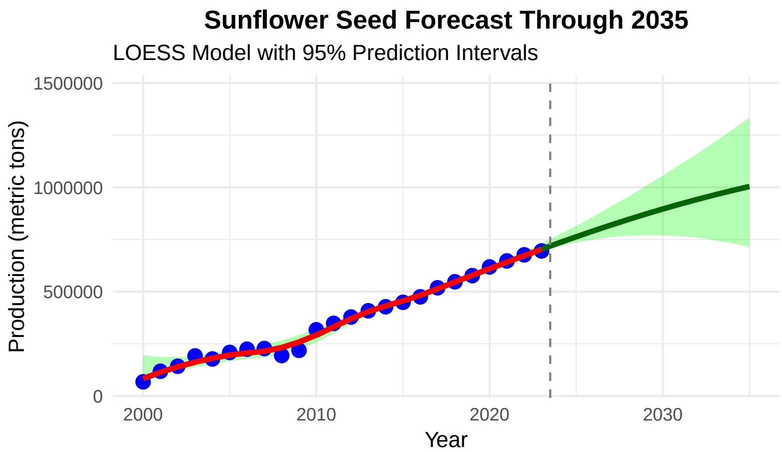
Year	Forecast	95% Confidence Interval
2024	372 907.7	338 234.9 – 407 580.4
2025	386 130.9	341 284.1 – 430 977.6
2026	400 095.9	347 989.4 – 452 202.4
2027	414 288.8	356 081.3 – 472 496.3
2028	428 551.6	364 897.8 – 492 205.5
2029	442 836.0	374 187.2 – 511 484.7
2030	457 126.9	383 828.8 – 530 425.1
2031	471 419.9	393 751.9 – 549 087.9
2032	485 713.5	403 909.3 – 567 517.8
2033	500 007.3	414 266.3 – 585 748.4
2034	514 301.2	424 796.4 – 603 806.0
2035	528 595.0	435 478.5 – 621 711.6

Table 12: ARIMA (1,1,1) Forecast of sunflower oil production in Kazakhstan from 2024 to 2035 (in metric tons)

	AUTOARIMA	ARIMA(1,1,1)	Practical Difference
Forecast (2035)	536 621.7 tons	528 595.0 tons	∼ 1.5% variation
Drift Term	+14 543.48 tons/yr	+14 293.876 tons/yr	Nearly identical

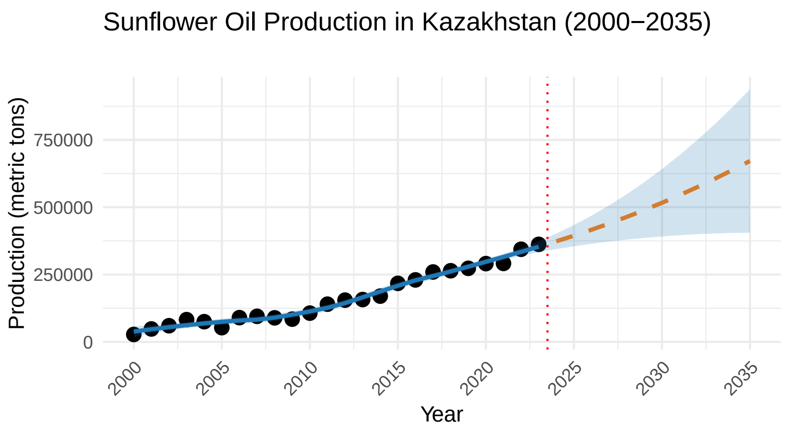
Table 13:

4.3 LOESS regression model



Year	Predicted	95% Confidence Interval
2024	733 489.5	681 215.0 – 775 764.0
2025	764 308.6	709 182.8 – 819 434.4
2026	795 111.2	724 874.4 – 865 348.1
2027	825 869.0	738 294.6 – 913 443.4
2028	856 557.6	749 422.3 – 963 693.0
2029	887 156.6	758 226.3 – 1 016 086.9
2030	917 648.3	764 672.1 – 1 070 624.4
2031	948 017.6	768 726.0 – 1 127 309.1
2032	978 251.7	770 356.4 – 1 186 147.0
2033	1 008 339.5	769 534.4 – 1 247 144.6
2034	1 038 271.4	766 233.7 – 1 310 309.0
2035	1 068 039.0	760 431.0 – 1 375 647.0

Table 14: Forecasted values with 95% confidence intervals

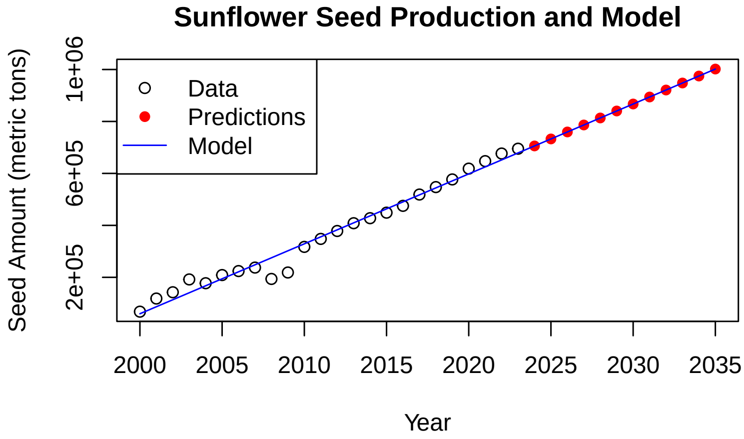


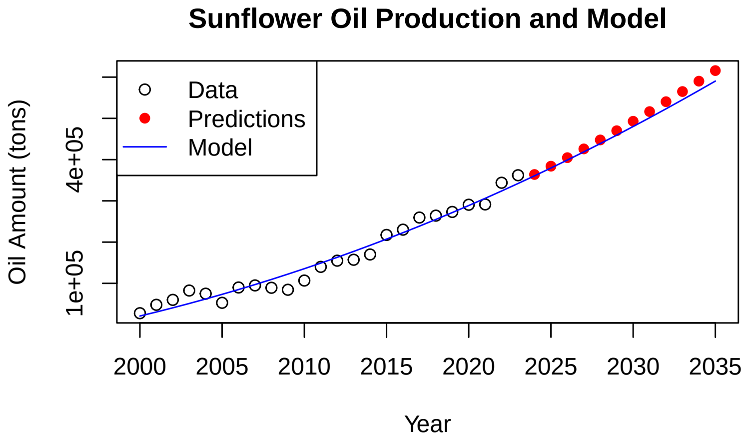
Year	**Forecast**	**95% Confidence Interval**
2024	373 452.0	344 625.1 – 402 278.9
2025	394 380.9	354 894.8 – 433 867.0
2026	416 379.9	364 070.1 – 468 689.7
2027	439 532.4	372 257.3 – 506 807.6
2028	463 902.6	379 513.6 – 548 291.7
2029	489 540.4	385 871.0 – 593 209.7
2030	516 485.1	391 347.7 – 641 622.5
2031	544 768.2	395 953.9 – 693 582.4
2032	574 415.1	399 695.4 – 749 134.8
2033	605 446.8	402 575.4 – 808 318.2
2034	637 880.2	404 595.0 – 871 165.5
2035	671 729.8	405 754.7 – 937 704.9

Table 15: LOESS prediction of sunflower oil production with 95% confidence intervals (2024–2035)

It was found that LOESS is not ideal for prediction, especially long-term, because it is fundamentally a non-parametric smoothing technique designed to capture local trends within the observed data range rather than model underlying structural patterns. It can only give short-term estimates (1–2 years) if trends are stable. The prediction line after 2023 is just LOESS averaging the last few data points which results in unreliability.

4.4 Multiple linear regression model





Year	Sunflower Seed Amounts	Sunflower Oil Amounts
2024	751 670.8	363 924.1
2025	789 608.5	384 034.9
2026	828 394.3	404 703.5
2027	868 028.4	425 929.9
2028	908 510.6	447 714.0
2029	949 841.2	470 056.0
2030	992 019.9	492 955.7
2031	1 035 046.9	516 413.2
2032	1 078 922.0	540 428.6
2033	1 123 645.4	565 001.7
2034	1 169 217.1	590 132.5
2035	1 215 636.9	615 821.2

Table 16: Predicted values by polynomial degrees (2024–2035)

Model	RMSE (t)	MAE (t)	R^2	MAPE (percent)
ARIMA(0,1,0)	24 117	13 980	0.9837	6.004
LOESS	20 690	15 270	0.988	7.50
Multiple Linear Reg.	17 034.52	13 954.16	0.9871	4.74

Table 17: Model Performance Comparison for sunflower seed production prediction.

Model	RMSE (t)	MAE (t)	R^2	MAPE (percent)
ARIMA(1,1,1)	16 442.33	12 317.15	0.9710	10.77
LOESS	12 060.0	10 250	0.9853	9.9
Multiple Linear Reg.	11 411.18	9 584	0.9811	9.35

Table 18: Model Performance Comparison for sunflower oil prediction.

The results above show that the multiple linear regression model outperformed ARIMA and LOESS models in predicting both sunflower seed and sunflower oil production. For sunflower seed production, the multiple linear regression model obtained the lowest RMSE and MAE, together with the highest R^2. Similarly, for sunflower oil prediction, the same model showed better performance with the lowest RMSE, MAE, and highest R^2 value. These results indicate that this model effectively captures the underlying relationships in the data and provides better forecasts. The relatively low MAPE values further confirm the model's higher suitability for accurate percentage-based error assessment.

5 Conclusions and Outlooks

By considering historical data and trade dynamics, this capstone project provided a comprehensive framework for forecasting sunflower seed and oil production in Kazakhstan. It explored the forecasting of sunflower seed and sunflower oil production in Kazakhstan using three different modeling approaches. While the multiple linear regression model offers simplicity in modeling relationships between a dependent variable and several independent variables, LOESS provides flexibility in capturing local variations, though it comes with lower predictive reliability. The ARIMA model, on the other hand, is particularly useful for understanding trends. Together, these models provide complementary insights into the country's sunflower products production dynamics. These results may benefit Kazakhstan's agricultural sector by improving the resource allocation strategy and sustainability. The findings emphasize the importance of choosing modeling techniques based on certain goals and nature of the data, and may support efforts in agricultural planning and policy development in Kazakhstan.

References

[1] National Bureau of Statistics of the Republic of Kazakhstan. (2025). *Kazakhstan's Economic Indicators*. https://stat.gov.kz

[2] Food and Agriculture Organization of the United Nations. (2025). *FAOSTAT: Crops and livestock products*. https://www.fao.org/faostat/

[3] United States Department of Agriculture Foreign Agricultural Service. (2025). *Sunflowerseed Explorer – Kazakhstan. USDA FAS.* https://ipad.fas.usda.gov/

[4] Observatory of Economic Complexity. (2025). *Sunflower Seeds in Kazakhstan.* https://oec.world/en/profile/bilateral-product/sunflower-seeds/reporter/kaz

[5] World Bank. (2025). *Sunflower Seeds in Kazakhstan.* https://oec.world/en/profile/bilateral-product/sunflower-seeds/reporter/kaz

[6] Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. D. (2022). *Regression: Models, Methods and Applications.* (2nd ed.). Springer. https://doi.org/10.1007/978-3-662-63882-8

[7] Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis.* (2nd ed.). Springer. https://doi.org/10.1007/b98888

A Appendix

Listing 1: Sunflower seed production analysis with outliers (ARIMA)

```

library(forecast)
library(tseries)
library(ggplot2)
library(dplyr)
library(scales)

sunflower_data <- data.frame(
  Year = 2000:2023,
  Metric_Tons = c(67800, 118400, 142600, 191900, 177300, 208700, 224100,
                  237500, 193800, 218600, 317200, 347900, 378400, 408300,
                  427600, 448900, 475200, 518700, 547300, 576800, 618400,
                  647200, 676500, 695100)
)


ts_data <- ts(sunflower_data$Metric_Tons, start = 2000, frequency = 1)


outliers <- tsoutliers(ts_data)
print(paste("Outliers detected at positions:", toString(outliers$index)))


ts_clean <- ts_data
ts_clean[outliers$index] <- outliers$replacements

outlier_points <- data.frame(
  Year = time(ts_data)[outliers$index],
  Value = ts_data[outliers$index],
  Label = paste("Outlier", c(2008, 2009))
)

ggplot() +

  geom_line(aes(x = time(ts_data), y = ts_data, color = "Original Data"),
            linewidth = 0.8) +
  geom_point(aes(x = time(ts_data), y = ts_data, color = "Original Data"))
     +

```

```
  geom_line(aes(x = time(ts_clean), y = ts_clean, color = "Outliers Removed"),
            linewidth = 0.8, linetype = "dashed") +

  geom_point(data = outlier_points,
             aes(x = Year, y = Value, fill = "Outliers"),
             shape = 21, size = 4, color = "red") +
  geom_text(data = outlier_points,
            aes(x = Year, y = Value, label = Label),
            vjust = -1.5, hjust = 0.5, color = "red") +

  scale_color_manual(name = "Series",
                     values = c("Original Data" = "#1f77b4",
                                "Outliers Removed" = "#ff7f0e")) +
  scale_fill_manual(name = "", values = c("Outliers" = "red")) +
  scale_x_continuous(breaks = seq(2000, 2023, by = 2)) +
  scale_y_continuous(labels = comma) +
  labs(title = "Sunflower Seed Production in Kazakhstan (2000-2023)",
       subtitle = "Outliers at 2008-2009 treated with linear interpolation",
       x = "Year", y = "Production (Metric Tons)") +
  theme_minimal(base_size = 12) +
  theme(legend.position = "bottom",
        panel.grid.minor = element_blank())

adf_original <- adf.test(ts_data)
print(paste("ADF p-value (original):", round(adf_original$p.value, 4)))

adf_clean <- adf.test(ts_clean)
print(paste("ADF p-value (cleaned):", round(adf_clean$p.value, 4)))

adf_diff <- adf.test(diff(ts_clean))
print(paste("ADF p-value (differenced):", round(adf_diff$p.value, 4)))

set.seed(123)
best_model <- auto.arima(ts_clean,
                         seasonal = FALSE,
                         stepwise = FALSE,
                         approximation = FALSE,
                         max.p = 3,
                         max.q = 3,
                         trace = TRUE)

summary(best_model)

model_order <- paste(arimaorder(best_model), collapse = ",")

checkresiduals(best_model, plot = TRUE)
```

```r
forecast_values <- forecast(best_model, h = 12)

forecast_df <- data.frame(
  Year = seq(2024, 2035),
  Point_Forecast = as.numeric(forecast_values$mean),
  Lo_95 = as.numeric(forecast_values$lower[,2]),
  Hi_95 = as.numeric(forecast_values$upper[,2])
)


ggplot() +

  geom_line(aes(x = time(ts_clean), y = ts_clean, color = "#1f77b4"),
            linewidth = 0.8) +
  geom_point(aes(x = time(ts_clean), y = ts_clean, color = "#1f77b4")) +


  geom_ribbon(data = forecast_df,
              aes(x = Year, ymin = Lo_95, ymax = Hi_95),
              fill = "grey70", alpha = 0.3) +
  geom_line(data = forecast_df,
            aes(x = Year, y = Point_Forecast),
            color = "#d62728", linewidth = 0.8) +
  geom_point(data = forecast_df,
             aes(x = Year, y = Point_Forecast),
             color = "#d62728") +


  scale_x_continuous(breaks = seq(2000, 2035, by = 5)) +
  scale_y_continuous(labels = comma, limits = c(0, NA)) +
  labs(title = "Sunflower Production Forecast (2024-2035)",
       subtitle = paste("Best Model: ARIMA(", model_order, ")", sep = ""),
       x = "Year", y = "Production (Metric Tons)") +
  theme_minimal(base_size = 12) +
  theme(panel.grid.minor = element_blank())


results <- data.frame(
  Year = c(2000:2023, 2024:2035),
  Actual = c(ts_data, rep(NA, 12)),
  Cleaned = c(ts_clean, rep(NA, 12)),
  Forecast = c(rep(NA, 24), forecast_values$mean),
  Forecast_Lo_95 = c(rep(NA, 24), forecast_values$lower[,2]),
  Forecast_Hi_95 = c(rep(NA, 24), forecast_values$upper[,2])
)

write.csv(results, "sunflower_production_forecast.csv", row.names = FALSE)


print(forecast_df)
```

Listing 2: Sunflower oil production analysis with outliers (ARIMA)

```r

library(forecast)
library(ggplot2)

sunflower_data <- data.frame(
  Year = 2000:2023,
  Amount = c(27600, 48000, 60000, 82600, 75000, 52900, 90000, 96000,
             89300, 84500, 106800, 140000, 155000, 157200, 170000,
```

```
                217373, 230141, 259521, 264150, 273159, 290600, 291360,
                343900, 362100)
)


med_neighbors <- median(c(sunflower_data$Amount[8], sunflower_data$Amount
    [11])) # 2007 and 2010 values
sunflower_data$Amount[9:10] <- med_neighbors  # Replacing 2008-2009


ts_data <- ts(sunflower_data$Amount, start = 2000)

calculate_metrics <- function(model, data) {
  res <- residuals(model)
  fitted <- fitted(model)
  return(data.frame(
    Residual_Variance = var(res, na.rm = TRUE),
    RMSE = sqrt(mean((data - fitted)^2, na.rm = TRUE)),
    MAPE = mean(abs((data - fitted)/(data + 1e-10)) * 100, na.rm = TRUE)
  ))
}


models <- list(
  ARIMA010 = Arima(ts_data, order = c(0,1,0)),
  ARIMA110 = Arima(ts_data, order = c(1,1,0)),
  ARIMA011 = Arima(ts_data, order = c(0,1,1)),
  ARIMA111 = Arima(ts_data, order = c(1,1,1)),
  AutoARIMA = auto.arima(ts_data, seasonal = FALSE, stepwise = FALSE)
)


metrics <- lapply(models, calculate_metrics, data = ts_data)
metrics_df <- do.call(rbind, metrics)
metrics_df$Model <- names(models)
metrics_df$AIC <- sapply(models, AIC)
metrics_df$BIC <- sapply(models, BIC)


metrics_df <- metrics_df[, c("Model", "AIC", "BIC", "Residual_Variance", "
    RMSE", "MAPE")]


print(metrics_df)


best_model <- models[[which.min(metrics_df$AIC)]]
forecast_values <- forecast(best_model, h = 12)


autoplot(forecast_values) +
  ggtitle(paste("Sunflower Oil Production Forecast\nBest Model:", names(
      which.min(metrics_df$AIC)))) +
  xlab("Year") +
  ylab("Production (Metric Tons)") +
  theme_minimal()
```

Listing 3: Sunflower seed production analysis with outliers (LOESS)

```
library(ggplot2)
library(dplyr)
```

```
sunflower <- data.frame(
  Year = 2000:2023,
  Amount = c(67800, 118400, 142600, 191900, 177300, 208700, 224100, 237500,
             193800, 218600, 317200, 347900, 378400, 408300, 427600, 448900,
             475200, 518700, 547300, 576800, 618400, 647200, 676500, 695100)
)


all_years <- data.frame(Year = 2000:2035)

loess_fit <- loess(Amount ~ Year, data = sunflower,
                   span = 0.75, control = loess.control(surface = "direct"))


pred <- predict(loess_fit, newdata = all_years, se = TRUE)
all_years$Amount <- pred$fit
all_years$se <- pred$se.fit


ggplot() +

  geom_point(data = sunflower, aes(x = Year, y = Amount), size = 3, color = "black") +


  geom_line(data = all_years, aes(x = Year, y = Amount), color = "blue", linewidth = 1) +


  geom_ribbon(data = all_years,
              aes(x = Year, ymin = Amount - 1.96*se, ymax = Amount + 1.96*se),
              fill = "blue", alpha = 0.2) +


  geom_line(data = all_years %>% filter(Year >= 2024),
            aes(x = Year, y = Amount), color = "red", linewidth = 1.2, linetype = "dashed") +


  annotate("text", x = 2030, y = 500000, label = "Predictions", color = "red", size = 5) +


  geom_vline(xintercept = 2023.5, linetype = "dotted", color = "black", linewidth = 1) +


  labs(title = "Sunflower Seed Harvest in Kazakhstan (2000-2035)",
       subtitle = "LOESS Model with Predictions to 2035",
       x = "Year", y = "Amount (metric tons)") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(2000, 2035, by = 5)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
future_pred <- all_years %>%
  filter(Year >= 2024) %>%
  mutate(CI_lower = Amount - 1.96*se,
         CI_upper = Amount + 1.96*se) %>%
  select(Year, Amount, CI_lower, CI_upper)

cat("\nFuture Predictions (2024-2035):\n")
print(future_pred, row.names = FALSE)
```

Listing 4: Sunflower oil production analysis with outliers (LOESS)

```r
library(ggplot2)
library(dplyr)

sunflower_oil <- data.frame(
  Year = 2000:2023,
  Amount = c(27600, 48000, 60000, 82600, 75000, 52900, 90000, 95000,
             89300, 84500, 106800, 140000, 155000, 157200, 170000, 217373,
             230141, 259521, 264150, 273159, 290600, 291360, 343900,
             362100)
)


loess_model <- loess(Amount ~ Year, data = sunflower_oil,
                     span = 0.65, control = loess.control(surface = "direct"
                     ))


all_years <- data.frame(Year = 2000:2035)
predictions <- predict(loess_model, newdata = all_years, se = TRUE)
all_years$Amount <- predictions$fit
all_years$se <- predictions$se.fit


ggplot() +

  geom_point(data = sunflower_oil, aes(x = Year, y = Amount), size = 3) +


  geom_line(data = all_years[all_years$Year <= 2023,],
            aes(x = Year, y = Amount), color = "#1f77b4", linewidth = 1) +


  geom_line(data = all_years[all_years$Year >= 2024,],
            aes(x = Year, y = Amount), color = "#ff7f0e", linewidth = 1,
            linetype = "dashed") +


  geom_ribbon(data = all_years,
              aes(x = Year, ymin = Amount - 1.96*se, ymax = Amount + 1.96*
              se),
              fill = "#1f77b4", alpha = 0.2) +


  geom_vline(xintercept = 2023.5, linetype = "dotted", color = "red") +


  labs(title = "Sunflower Oil Production in Kazakhstan (2000-2035)",
       subtitle = " ",
       x = "Year",
       y = "Production (metric tons)",
```

```r
      caption = " ") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(2000, 2035, by = 5)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


future_predictions <- all_years %>%
  filter(Year >= 2024) %>%
  mutate(CI_lower = Amount - 1.96*se,
         CI_upper = Amount + 1.96*se) %>%
  select(Year, Predicted = Amount, CI_lower, CI_upper)

cat("\nLOESS Predictions for Sunflower Oil Production (2024-2035):\n")
print(future_predictions, row.names = FALSE)
```

Listing 5: Sunflower seed production analysis with outliers (Multiple linear regression)

```r
library(dplyr)
library(zoo)
library(ggplot2)



# Sunflower Seed Data (Table 2)
seed_data <- data.frame(
  Year = 2000:2023,
  Seed_Production = c(
    67800, 118400, 142600, 191900, 177300, 208700, 224100, 237500,
    193800, 218600, 317200, 347900, 378400, 408300, 427600, 448900,
    475200, 518700, 547300, 576800, 618400, 647200, 676500, 695100
  )
)

# Sunflower Oil Production Data (Table 3)
oil_production_data <- data.frame(
  Year = 2000:2023,
  Oil_Production = c(
    27600, 48000, 60000, 82600, 75000, 52900, 90000, 95000,
    89300, 84500, 106800, 140000, 155000, 157200, 170000, 217373,
    230141, 259521, 264150, 273159, 290600, 291360, 343900, 362100
  )
)

# Sunflower Oil Exports (Table 4)
export_data <- data.frame(
  Year = 2000:2023,
  Export_Volume = c(
    4150, 6820, 9350, 12100, 15300, 18200, 22400, 27800,
    19600, 24300, 42700, 67500, 83200, 97800, 104600, 112500,
    134200, 158700, 187200, 203800, 213400, 254600, 286100, 317500
  ),
  Export_Avg_Price = c(
    480, 510, 535, 580, 610, 620, 655, 725,
    1020, 880, 890, 950, 1100, 1075, 1150, 1025,
    1080, 1125, 1150, 1210, 1320, 1425, 1540, 1490
  )
)

# Sunflower Oil Imports (Table 5)
import_data <- data.frame(
  Year = 2000:2023,
```

```
  Import_Volume = c(
    4850, 5200, 6100, 7500, 9800, 14500, 18300, 22000,
    25400, 20100, 34700, 38200, 42500, 45100, 48300, 50200,
    52000, 58400, 64800, 60100, 45600, 39500, 30200, 24500
  ),
  Import_Value = c(
    3400, 3600, 4300, 5200, 6900, 10200, 13100, 16500,
    19800, 14200, 24900, 28600, 32000, 34200, 36800, 39500,
    41000, 46500, 54700, 51300, 49800, 59200, 44700, 39800
  )
)



combined_data <- merge(seed_data, oil_production_data, by = "Year", all.x = TRUE)
combined_data <- merge(combined_data, export_data, by = "Year", all.x = TRUE)
combined_data <- merge(combined_data, import_data, by = "Year", all.x = TRUE)


combined_data <- combined_data %>%
  arrange(Year) %>%
  mutate(
    Oil_Production = na.locf(Oil_Production, na.rm = FALSE),
    Export_Volume = na.locf(Export_Volume, na.rm = FALSE),
    Import_Volume = na.locf(Import_Volume, na.rm = FALSE),
    Import_Value = na.locf(Import_Value, na.rm = FALSE),
    Export_Avg_Price = na.locf(Export_Avg_Price, na.rm = FALSE)
  )


combined_data <- combined_data %>%
  mutate(
    Lagged_Seed_Production = lag(Seed_Production, n = 1),
    Lagged_Export_Volume = lag(Export_Volume, n = 1),
    Lagged_Import_Volume = lag(Import_Volume, n = 1)
  )


combined_data <- combined_data[-1, ]


model <- lm(Seed_Production ~ Year + Oil_Production + Lagged_Seed_Production +
              Lagged_Export_Volume + Lagged_Import_Volume + Export_Avg_Price + Import_Value,
            data = combined_data)


summary(model)


library(car)
vif_values <- vif(model)
print("VIF values:")
print(vif_values)
```

```r
plot(model)



future_years <- data.frame(Year = 2024:2035)


oil_model <- lm(Oil_Production ~ Year, data = tail(combined_data, 5))
future_years$Oil_Production <- predict(oil_model, newdata = future_years)


export_model <- lm(Export_Volume ~ Year, data = combined_data)
future_years$Lagged_Export_Volume <- predict(export_model, newdata = data.frame(Year = 2023:2034))


import_model <- lm(Import_Volume ~ Year, data = combined_data)
future_years$Lagged_Import_Volume <- predict(import_model, newdata = data.frame(Year = 2023:2034))


export_price_model <- lm(Export_Avg_Price ~ Year, data = combined_data)
future_years$Export_Avg_Price <- predict(export_price_model, newdata = future_years)


import_value_model <- lm(Import_Value ~ Year, data = combined_data)
future_years$Import_Value <- predict(import_value_model, newdata = future_years)



predictions <- numeric(length = nrow(future_years))


previous_seed_production <- last(combined_data$Seed_Production) # Get last actual seed production

for (i in 1:nrow(future_years)) {

  current_year_data <- data.frame(
    Year = future_years$Year[i],
    Oil_Production = future_years$Oil_Production[i],
    Lagged_Seed_Production = previous_seed_production,
    Lagged_Export_Volume = future_years$Lagged_Export_Volume[i],
    Lagged_Import_Volume = future_years$Lagged_Import_Volume[i],
    Export_Avg_Price = future_years$Export_Avg_Price[i],
    Import_Value = future_years$Import_Value[i]
  )


  predictions[i] <- predict(model, newdata = current_year_data)


  previous_seed_production <- predictions[i]
}

future_years$Predicted_Seed_Production <- predictions

print(future_years)

```

```
plot_data <- rbind(
  data.frame(Year = combined_data$Year, Seed_Production = combined_data$
      Seed_Production, Type = "Historical"),
  data.frame(Year = future_years$Year, Seed_Production = future_years$
      Predicted_Seed_Production, Type = "Predicted")
)


ggplot(plot_data, aes(x = Year, y = Seed_Production, color = Type)) +
  geom_line() +
  geom_point() +
  labs(
    title = "Sunflower Seed Production: Historical Data and Predictions",
    x = "Year",
    y = "Seed Production (Metric Tons)",
    color = "Data Type"
  ) +
  theme_minimal() +
  scale_x_continuous(breaks = seq(2000, 2035, by = 5))


ggsave("sunflower_seed_production_plot.png", width = 8, height = 6, units =
     "in")
```

Listing 6: Sunflower oil production analysis with outliers (Multiple linear regression)

```


# Table 2: Sunflower Seed Harvest
sunflower_seed_data <- data.frame(
  Year = c(2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009,
      2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020,
      2021, 2022, 2023),
  Seed_Amount = c(67800, 118400, 142600, 191900, 177300, 208700, 224100,
      237500, 193800, 218600, 317200, 347900, 378400, 408300, 427600,
      448900, 475200, 518700, 547300, 576800, 618400, 647200, 676500,
      695100)
)

# Table 3: Sunflower Oil Production
sunflower_oil_data <- data.frame(
  Year = c(2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009,
      2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020,
      2021, 2022, 2023),
  Oil_Amount = c(27600, 48000, 60000, 82600, 75000, 52900, 90000, 95000,
      89300, 84500, 106800, 140000, 155000, 157200, 170000, 217373, 230141,
       259521, 264150, 273159, 290600, 291360, 343900, 362100)
)

# Table 4: Sunflower Oil Exports
sunflower_export_data <- data.frame(
  Year = c(2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009,
      2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020,
      2021, 2022, 2023),
  Export_Volume = c(4150, 6820, 9350, 12100, 15300, 18200, 22400, 27800,
      19600, 24300, 42700, 67500, 83200, 97800, 104600, 112500, 134200,
      158700, 187200, 203800, 213400, 254600, 286100, 317500),
  Export_Price_Per_Ton = c(480, 510, 535, 580, 610, 620, 655, 725, 1020,
      880, 890, 950, 1100, 1075, 1150, 1025, 1080, 1125, 1150, 1210, 1320,
      1425, 1540, 1490)
```

```r
)

# Table 5: Sunflower Oil Imports
sunflower_import_data <- data.frame(
  Year = c(2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023),
  Import_Volume = c(4850, 5200, 6100, 7500, 9800, 14500, 18300, 22000, 25400, 20100, 34700, 38200, 42500, 45100, 48300, 50200, 52000, 58400, 64800, 60100, 45600, 39500, 30200, 24500),
  Import_Value_1000USD = c(3400, 3600, 4300, 5200, 6900, 10200, 13100, 16500, 19800, 14200, 24900, 28600, 32000, 34200, 36800, 39500, 41000, 46500, 54700, 51300, 49800, 59200, 44700, 39800)
)

# Table 6:  Average sunflower oil and seed prices in Kazakhstan
kazakhstan_prices <- data.frame(
  Item = c("Sunflower oil (per liter)", "Sunflower seeds (per 100 grams)"),
  Price_USD = c(1.66, 0.78)
)


merged_data <- merge(sunflower_seed_data, sunflower_oil_data, by = "Year")
merged_data <- merge(merged_data, sunflower_export_data, by = "Year")
merged_data <- merge(merged_data, sunflower_import_data, by = "Year")


merged_data$Year_Squared <- merged_data$Year^2


merged_data$Import_Price_Per_Ton <- merged_data$Import_Value_1000USD * 1000 / merged_data$Import_Volume


merged_data$Oil_Price_KZT <- kazakhstan_prices$Price_USD[kazakhstan_prices$Item == "Sunflower oil (per liter)"]
merged_data$Seed_Price_KZT <- kazakhstan_prices$Price_USD[kazakhstan_prices$Item == "Sunflower seeds (per 100 grams)"]


merged_data$Seed_Price_Per_Ton_KZT <- merged_data$Seed_Price_KZT / 0.0001


model <- lm(Oil_Amount ~ Year + Year_Squared + Seed_Amount + Export_Volume + Import_Volume + Import_Price_Per_Ton + Export_Price_Per_Ton + Oil_Price_KZT + Seed_Price_Per_Ton_KZT, data = merged_data)


summary(model)

 "sunflower_seed_predictions_2024_2035.md"
future_seed_data <- data.frame(
  Year = 2024:2035,
  Seed_Amount = c(728732.2, 763786.3, 800072.7, 837591.3, 876342.2, 916325.4, 957540.8, 999988.5, 1043668.4, 1088580.6, 1134725.1, 1182101.8)
)


future_years <- data.frame(Year = 2024:2035)
future_years$Year_Squared <- future_years$Year^2
```

```
future_years$Seed_Amount <- future_seed_data$Seed_Amount

export_model <- lm(Export_Volume ~ Year, data = merged_data)
future_years$Export_Volume <- predict(export_model, newdata = future_years)

import_model <- lm(Import_Volume ~ Year, data = merged_data)
future_years$Import_Volume <- predict(import_model, newdata = future_years)

import_price_model <- lm(Import_Price_Per_Ton ~ Year, data = merged_data)
future_years$Import_Price_Per_Ton <- predict(import_price_model, newdata = future_years)

export_price_model <- lm(Export_Price_Per_Ton ~ Year, data = merged_data)
future_years$Export_Price_Per_Ton <- predict(export_price_model, newdata = future_years)


future_years$Oil_Price_KZT <- kazakhstan_prices$Price_USD[kazakhstan_prices$Item == "Sunflower oil (per liter)"]
future_years$Seed_Price_Per_Ton_KZT <- kazakhstan_prices$Price_USD[kazakhstan_prices$Item == "Sunflower seeds (per 100 grams)"] / 0.0001


predictions <- predict(model, newdata = future_years)


predictions_df <- data.frame(Year = 2024:2035, Predicted_Oil_Amount = predictions)


print(predictions_df)


plot(merged_data$Year, merged_data$Oil_Amount, main = "Sunflower Oil Production and Model", xlab = "Year", ylab = "Oil Amount (metric tons)",
     xlim = c(min(merged_data$Year), max(future_years$Year)),
     ylim = c(min(merged_data$Oil_Amount), max(c(merged_data$Oil_Amount, predictions)))
     )
points(future_years$Year, predictions, col = "red", pch = 16)

curve_years <- seq(min(merged_data$Year), max(future_years$Year), by = 1)
curve_data <- data.frame(Year = curve_years)
curve_data$Year_Squared <- curve_years^2


seed_model <- lm(Seed_Amount ~ Year, data = merged_data)
curve_data$Seed_Amount <- predict(seed_model, newdata = curve_data)
curve_data$Export_Volume <- predict(export_model, newdata = curve_data)
curve_data$Import_Volume <- predict(import_model, newdata = curve_data)
curve_data$Import_Price_Per_Ton <- predict(import_price_model, newdata = curve_data)
curve_data$Export_Price_Per_Ton <- predict(export_price_model, newdata = curve_data)
curve_data$Oil_Price_KZT <- kazakhstan_prices$Price_USD[kazakhstan_prices$Item == "Sunflower oil (per liter)"]
curve_data$Seed_Price_Per_Ton_KZT <- kazakhstan_prices$Price_USD[kazakhstan_prices$Item == "Sunflower seeds (per 100 grams)"] / 0.0001
predicted_curve_values <- predict(model, newdata = curve_data)
```

```r

lines(curve_years, predicted_curve_values, col = "blue")
legend("topleft", legend = c("Data", "Predictions", "Model"), col = c("
    black", "red", "blue"), pch = c(1, 16, NA), lty = c(NA, NA, 1))
```