

Impact of Canadian Crop Prices on Crop Production

Foundations of Data Science - Group Project

Muskaan Brar¹, Jeush Stanly², Monika Tomar³, Jason Wilcox⁴,
Elmira Sepehri⁵, and David Fishman⁶

Email ids: ¹mbrar44@uwo.ca, ²jeush.stanly@mail.utoronto.ca,
³25monikatomar25@gmail.com, ⁴jason.wqost@slmails,
⁵elmira.sphr@gmail, ⁶davjfish@gmail.com

December 2023

Contents

1	Summary	3
1.1	Data source	3
1.1.1	What is the nature of the data you chose?	3
1.1.2	Why is it interesting?	3
1.2	Analysis	3
1.2.1	What were the challenges?	3
1.2.2	How did you overcome them?	4
1.3	Conclusions	5
1.3.1	What did you learn about your dataset?	5
2	Objectives	6
2.1	What are the goals of the analysis and why did you choose them?	6
2.2	What question(s) do you want to answer?	6
2.3	What hypothesis(es) do you have and what is your approach to tackle the problem?	6
3	Data Preparation	7
3.1	What was your data source (e.g., web scraping, corporate data, a standard machine learning data set, open data, etc.)?	7
3.2	How good was the data quality?	7
3.3	What did you need to do to procure it?	7
3.4	What tools or code did you need to use to prepare it for analysis?	7
3.5	What challenges did you face?	7
4	Analysis	10
4.1	What trends, correlations, and/or patterns do you see in the data?	10
4.1.1	Poultry	10
4.1.2	Hog	10
4.1.3	Wheat	10
5	Conclusions	13
5.1	What did you learn about your data set?	13

1 Summary

1.1 Data source

1.1.1 What is the nature of the data you chose?

The dataset we first selected was sourced from the Government of Canada's Open Government Portal. It records the monthly price of various crop, commodity and livestock products produced in Canada. The raw data contains the price of each farm product, the month the price was captured, a product description, and the province of production. Our desire to perform analysis regarding the relationship between price and production led us to seek additional datasets containing production data for selected commodities. This production data was also sourced from Government of Canada's Open Government Portal. It recorded the monthly production of farm products by province of production over a period of years.

1.1.2 Why is it interesting?

The datasets we eventually selected were interesting as between them they contain a time series of prices for different products and from different provinces, and the production for those products over of a period of over time, from January 1980 to the present. The duration of the dataset provided an opportunity to find patterns and the possibility of a relationship between price and the amount of production within the regions of Canada. Analysis of these trends could provide us with insights with current day problems such as inflation and the cost of living crisis.

1.2 Analysis

1.2.1 What were the challenges?

Within our chosen datasets of Canadian farm product prices and production spanning from 1980 to present, our analysis required us to overcome a series of challenges that demanded a number of solutions.

The format of the certain features within the data, such as the REF_DATE were recorded as text and as such were not immediately usable, leading us to perform some feature engineering.

The structure of the pricing dataset allowed for differing units of measured to utilized within the data for a singled commodity.

The Wheat production dataset also could contain multiple entries for any given year which made comparison to other commodities problematic.

The differing ranges of data of the farm products represented in the dataset posed a challenge in terms of selecting the best products for analysis

The publication of distinct the price and production datasets further added layers of complexity. Our analysis required these be merged.

The presence of null values in the merged pricing and production datasets added also required additional action prior to our final analysis.

Finally, the limited overlap in time series, resulted in a comparatively small sample size, and led to the selection of suitable techniques suitable for smaller datasets.

1.2.2 How did you overcome them?

To overcome the challenges presented in previous section, we took a number of steps to prepare the data and create a model for analysis

The date feature `REF_DATE` as imported as a text and feature engineering was leveraged to change its datatype to a date.

During the validation of the model, we noted that the Wheat production data reported production in March, July and December, as opposed once a year. As the March and July values were the same the previously reported December value, we were able to make the assumption that the December values were the correct ones. Accordingly, we selected the the December values and re-indexed the data to match the price data prior to merging.

The source of the data noted that the unit of measure may not be consistent. As part of our data cleaning we validated that the unit of measure for our the selected products was indeed consistent.

When reviewing a visual representing of the time series data for all provinces. It became evident that all the province had a similar pattern, with the exception of Newfoundland and Labrador. Therefore we elected to exclude Newfoundland and Labrador from our subsequent analysis.

To analyze the impact of product production on product price it was needed to add production data to our model. Using the Government of Canada's Open Government Portal we search and found production data for our selected product products. After similar feature engineering to correct the `REF_DATE` feature, we merged the production data using the product description as a key.

The availability of price and production data led us to select Meat Chickens, Wheat and Hogs as products for final analysis on relationship between price and production. The availability of price and production data also required us to drop null values resulting in few data points for comparison.

To evaluate the relationship between price and production we applied the technique of linear regression on the resulting data model.

1.3 Conclusions

1.3.1 What did you learn about your dataset?

After reviewing the liner regression results for Meat Chickens, Wheat and Hogs we found the following:

For Meat Chickens, the model intercept was found to be 26.4810 and the coefficient price variable is 50.9459.

$$\text{production} = 26.4810 + 50.9459 \times \text{price}$$

The F-staticic for this model was 12.47 and the changes of observing this statistic under a normal distribution is less than 0.05%. Therefore, with a p-value set to 0.05, we would reject the null hypothesis.

An R2 value of 0.675 means approximately 67% of the variance of the data can be accounted for by this model. We lead us to the conclusion that price does affect production in Meat Chickens and supports the regression of the null hypothesis.

The model for Wheat had an intercept of 1.285e4 and the coefficient for the price variable is 41.3883

$$\text{production} = 1.285e4 + 41.3883 \times \text{price}$$

The F-staticic for this Wheat was 8.065 and the changes of observing this statistic under a normal distribution is less than 0.05%. Therefore, with a p-value set to 0.05, we would reject the null hypothesis.

The R2 is 0.28 which means approximately 30% of the variance of the data can be accounted for by this model. There is more unexplained variability in these data compared to the Meat Chickens

Finally the model for Hogs has an F-staticic of 0.3441 and the chances of observing this statistic under a normal distribution is approximately 56%. Therefore, with a p-value set to 0.05, we are not in a position to reject the null hypothesis.

Based on these results, we can say that only price affects production only in some farm products produced in Canada. This combined with some understanding of of Canadian economic policy, show that affect to be stronger in products without any form of Supply Management involvement from Government.

2 Objectives

2.1 What are the goals of the analysis and why did you choose them?

The goal of this analysis is to explore and better understand changes in the production of agricultural crops in Canada. Understanding these fluctuations can yield valuable insights into the Canadian economy and the agricultural sector.

Statistics Canada collects copious amounts of data via the Census of Agriculture (REF) thus making these datasets excellent candidates for analysis in our group project.

2.2 What question(s) do you want to answer?

While the collections of elements affecting the total production of agricultural products are complex and multifaceted, this report we focus solely on a single variables: price.

2.3 What hypothesis(es) do you have and what is your approach to tackle the problem?

The null hypothesis for this anal

NOTES: supply management, commodities, large investments

3 Data Preparation

3.1 What was your data source (e.g., web scraping, corporate data, a standard machine learning data set, open data, etc.)?

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Architecto at aut beatae commodi consequatur eaque fugit, harum hic impedit in incidunt inventore odio officia ratione tenetur totam veritatis. Debitis, magnam.

3.2 How good was the data quality?

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Architecto at aut beatae commodi consequatur eaque fugit, harum hic impedit in incidunt inventore odio officia ratione tenetur totam veritatis. Debitis, magnam.

3.3 What did you need to do to procure it?

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Architecto at aut beatae commodi consequatur eaque fugit, harum hic impedit in incidunt inventore odio officia ratione tenetur totam veritatis. Debitis, magnam.

3.4 What tools or code did you need to use to prepare it for analysis?

For chickens, the raw had to be transformed... etc. Talk about NL data being excluded and the provincial data being aggregated by annual mean; see Figure 1

Please refer to Figure 2

3.5 What challenges did you face?

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Architecto at aut beatae commodi consequatur eaque fugit, harum hic impedit in incidunt inventore odio officia ratione tenetur totam veritatis. Debitis, magnam.

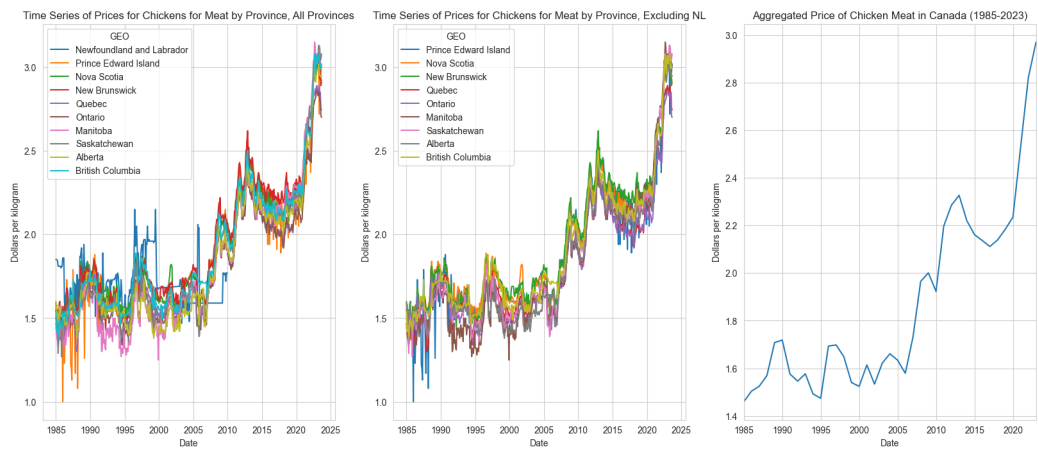


Figure 1: This figure shows the raw data, and aggregated forms of chicken prices. MORE TEXT NEEDED?

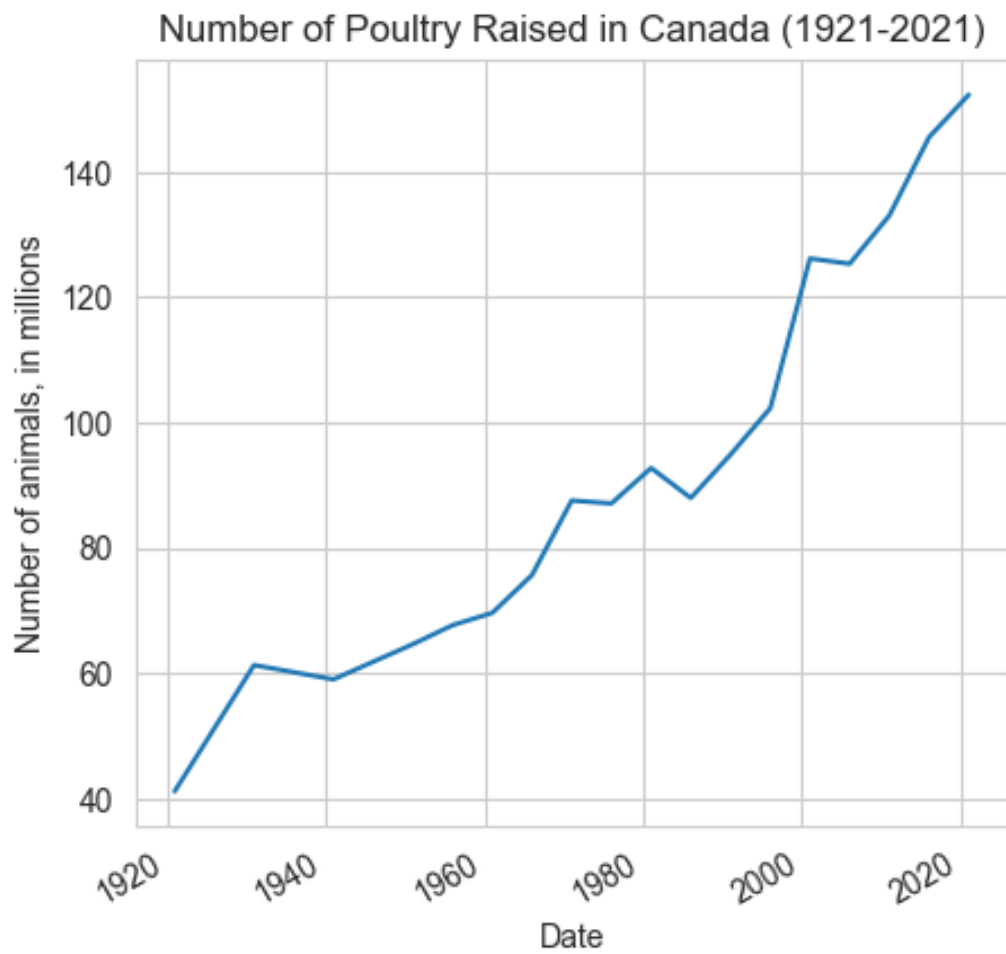


Figure 2: This figure shows the number of chickens counted on chicken farms in Canada between 1921-2021.

4 Analysis

4.1 What trends, correlations, and/or patterns do you see in the data?

4.1.1 Poultry

A scatter plot of chicken production can be observed in Figure 3.

The results from the OLS Figure 4

4.1.2 Hog

4.1.3 Wheat

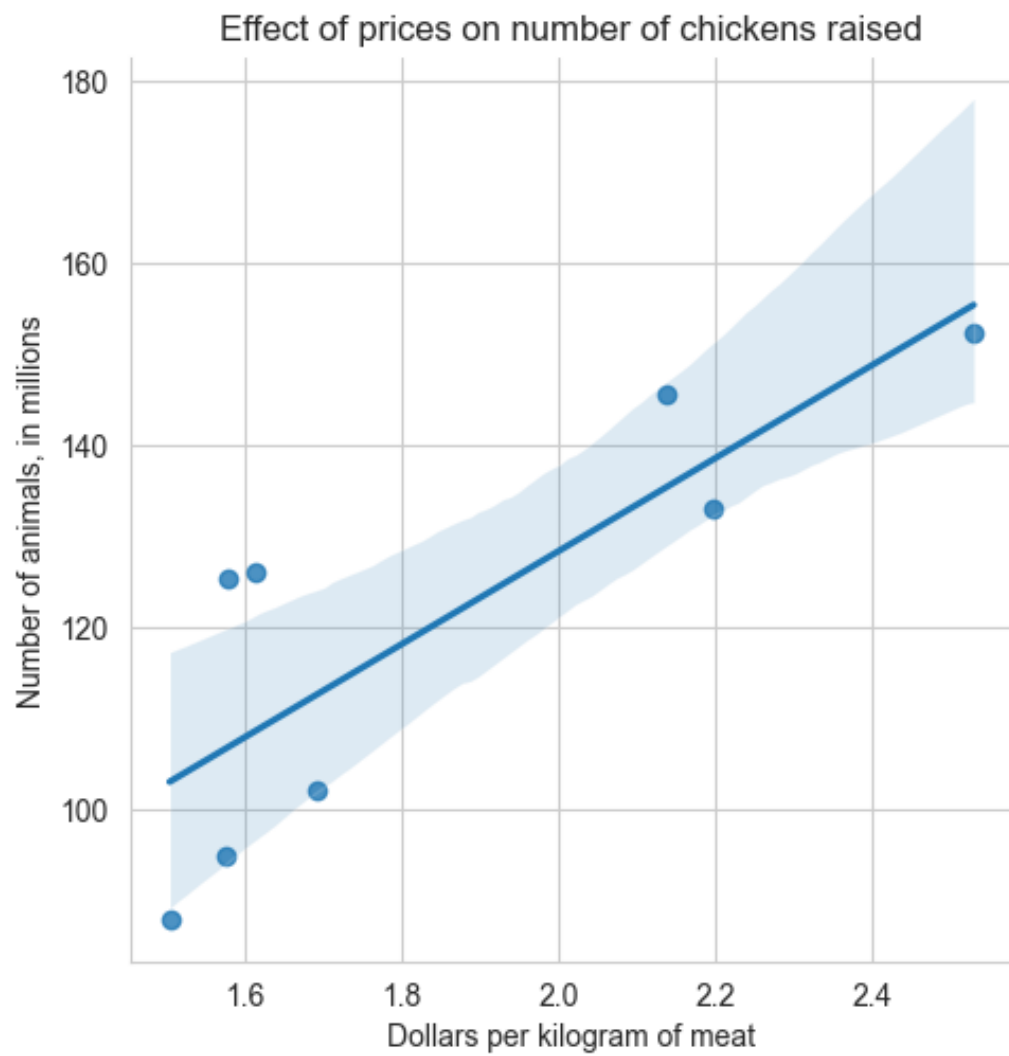


Figure 3: A scatter plot of chicken prices vs. chicken production.

OLS Regression Results						
Dep. Variable:	production			R-squared:	0.675	
Model:	OLS			Adj. R-squared:	0.621	
Method:	Least Squares			F-statistic:	12.47	
Date:	Sun, 10 Dec 2023			Prob (F-statistic):	0.0124	
Time:	16:19:44			Log-Likelihood:	-31.609	
No. Observations:	8			AIC:	67.22	
Df Residuals:	6			BIC:	67.38	
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t 	[0.025	0.975]
prices	50.9459	14.428	3.531	0.012	15.642	86.250
Eins	26.4810	27.235	0.972	0.368	-40.161	93.123
Omnibus:	2.302	Durbin-Watson:		1.404		
Prob(Omnibus):	0.316	Jarque-Bera (JB):		0.929		
Skew:	0.382	Prob(JB):		0.629		
Kurtosis:	1.516	Cond. No.		12.7		

Figure 4: Results from an Ordinary Least Squares Regression.

5 Conclusions

5.1 What did you learn about your data set?

Scratch:

This analysis is complex because it is not clear which variable should be the response and which are the covariates. For example, we can imagine a scenario where the increase of prices results in the increase of production, but it is also conceivable that the various levels of production will drive prices changes.