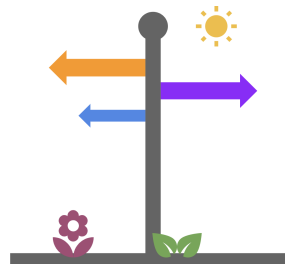


## Follow Your Own Case Study Path



### Scenario

You are a junior data analyst working for a business intelligence consultant. You have been at your job for six months, and your boss feels you are ready for more responsibility. He has asked you to lead a project for a brand new client — this will involve everything from defining the business task all the way through presenting your data-driven recommendations. You will choose the topic, ask the right questions, identify a fresh dataset, and ensure its integrity, conduct analysis, create compelling data visualizations, and prepare a presentation.

### Ask

Five questions will guide your case study:

1. What type of company does the client represent, and what are they asking?

My client is an economic consulting firm providing advisory services to small-to-medium sized egg-farm businesses (SMEs). Due to the current inflationary market, they operate on thin profit margins as they try to keep and expand intermediary distribution channels. Therefore, they require high quality analysis from prospective data analysts. They are asking me to model the relationship between egg prices and egg production, to see if in the aggregate a trend emerges that they can capitalize on. Many would believe that as prices increase so does production, in theory stabilizing prices.

2. What are the key factors involved in the business task investigated?

Key factors involved are understanding market dynamics of the egg industry, whether price shocks come from endogenous (that is, market-related such as supply-demand fluctuations) or exogenous (determined by factors outside of market conditions, such as the

Covid-19 global pandemic or the 2024 bird-flu affecting poultry farms). Other key factors include whether demand is generally stable given a certain price range. That is, how predictable its price elasticity of demand is.

### 3. What type of data is appropriate for the analysis?

For this analysis, up-to-date aggregate industry data from public data collection agencies would be ideal. The historical data should be based on industry-wide averages so as to not let regional fluctuations distort the overall price and production average of eggs across time. While hour-by-hour or day-to-day datasets would be ideal, they are computationally expensive to run a regression analysis on. And given that monthly data serves as a decent proxy as market shocks in production or pricing often have a significant lag, this project looks at the business task using month-by-month datasets.

### 4. Where was that data obtained?

In order to properly conduct this analysis, proper sourcing of data must be done. The best dataset found to study variations in historical egg production was obtained from USDA's National Agricultural Statistics Services. They are a highly credible government agency that hires statisticians, industry-experts, and statistical technicians and who produce reports used in industry and government policymaking due to their high-reliability. Specifically, their customizable census which was adapted to display on Animals & Products looking at the US national organic eggs sales measured in dozens. While there were more precise and accurate parameters online to look at production, the datasets were generally either too small in size, had inconsistent time intervals for data reporting, or were too old/no longer relevant. Therefore, the chosen dataset for production was established to be the most fitting giving the compromises made. the data is ROCC.

The best dataset found to study variation in national egg price averages was obtained from the St Louis Federal Reserve Economic (FRED) Database, but was compiled by the Bureau of Labor Statistics. FRED is maintained by the US Federal Reserve Bank of St. Louis and it aggregates data from trusted sources widely used by economists and policymakers. The data is therefore ROCC.

The sources are attached herein:

Data Sources with link:


- Production: [USDA/NASS QuickStats Ad-hoc Query Tool](#)
  - Eggs - Production, measured in dozens
- Price: [Average Price: Eggs, Grade A, Large \(Cost per Dozen\) in U.S. City Average \(APU0000708111\) | FRED | St. Louis Fed](#)

### 5. Who is the audience, and what materials will help present to them effectively?

The intended audience will be the economic consulting advisory firm which will present this analysis to business owners and executives of small and medium sized poultry farms in order to make data-informed decisions.

- Documentation of any cleaning or manipulation of data

After finding the two dataset sources the data was cleaned. The data formats needed to be matched and the year and month needed to be formatted so as to match rows and total data size. The price data had more data points so those had to be removed to get a better and equally varied datasets. A link to the google sheets cleaned dataset is attached hereunder:

 GDAC Egg PED project

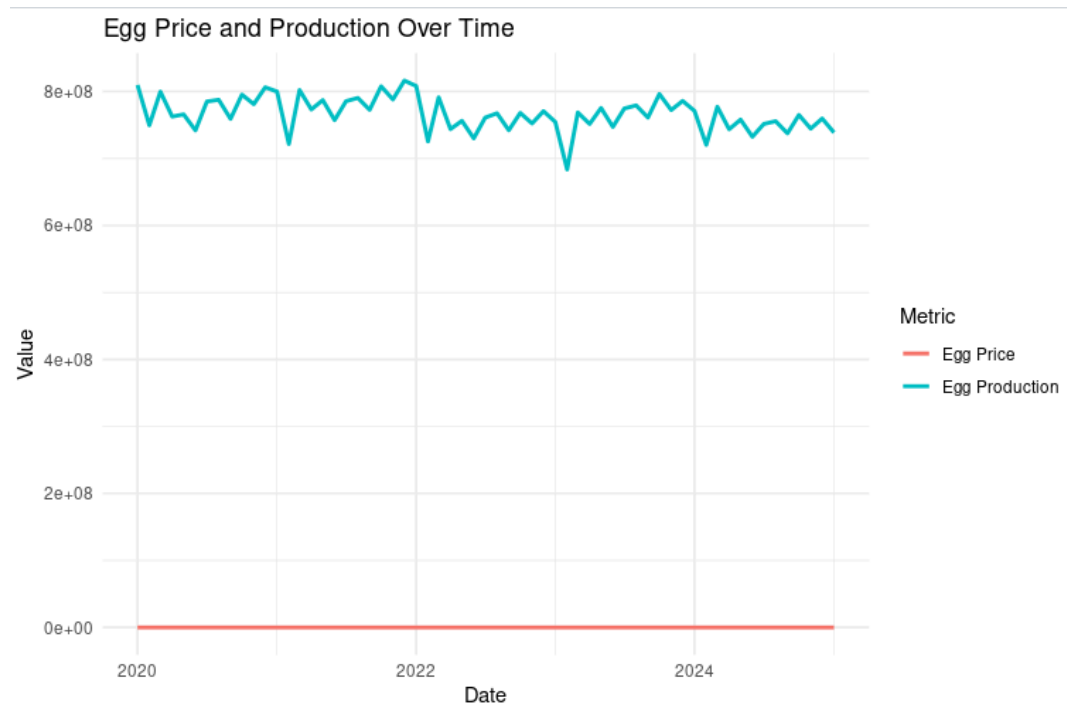
That notwithstanding, further issues with data and code knowledge arose. The data was initially cleaned wrong multiple times. The date, price, and production needed to be right next to each other as columns within the same table in simple form for R to properly read them without requiring overly complicated data manipulation. The research and experimentation also suggests that csv files work better than excel files. With less issues such as too many columns, wrong header rows, or date formatting incompatibility.

After loading the necessary ggplot2 and making R read the dataset, the unnecessary index column still had to be removed. Further resources from stackoverflow.com were needed to gain the skills to convert the date column to an appropriate date format. Converting numeric columns and manually reshaping the data for ggplot also required stackoverflow research, such as using the function rbind() to combine both datasets.

Here is my first batch of code:

```
ggplot(PlotData, aes(x = Date, y = Value, color = Metric)) +  
  geom_line(size = 1) +  
  labs(title = "Egg Price and Production Over Time",  
        x = "Date",  
        y = "Value",  
        color = "Metric") +  
  theme_minimal()
```

This code had several issues, as both variables' values use different scales: The egg production-approximate series is in millions of sales, while the egg price series looks at the cost of a dozen of eggs, making the egg price series appear flat. The graph is attached hereunder:



In order to fix this issue, it was necessary to normalize the data so they are more visually comparable. Using stackoverflow research, the `sec.axis` function was used in the code to display two y-axes (one for price and one for production).

The modified ggplot2 code batch is attached hereunder:

```
library(ggplot2)
```

```
PriceProd <- read.csv("Final Egg_Price_and_Production_Data.csv")
```

```
PriceProd <- PriceProd[, c("Date", "Egg_Production", "Egg_Price")]
```

```

PriceProd$Date <- as.Date(PriceProd$Date, format="%Y-%m-%d")

View(PriceProd$Date)

PriceProd$Egg_Production <- as.numeric(PriceProd$Egg_Production)

PriceProd$Egg_Price <- as.numeric(PriceProd$Egg_Price)

PriceProd$Egg_Price_Scaled <- PriceProd$Egg_Price * (max(PriceProd$Egg_Production) / max(PriceProd$Egg_Price))

ggplot(PriceProd, aes(x = Date)) +

  geom_line(aes(y = Egg_Production, color = "Egg Production"), linewidth = 1) +

  geom_line(aes(y = Egg_Price_Scaled, color = "Egg Price"), linewidth = 1) +

  scale_y_continuous(

    name = "Egg Production (millions)",

    sec.axis = sec_axis(~ . / (max(PriceProd$Egg_Production) / max(PriceProd$Egg_Price)), name = "Egg Price (USD)")

  ) +

  labs(title = "Egg Price and Production Over Time",

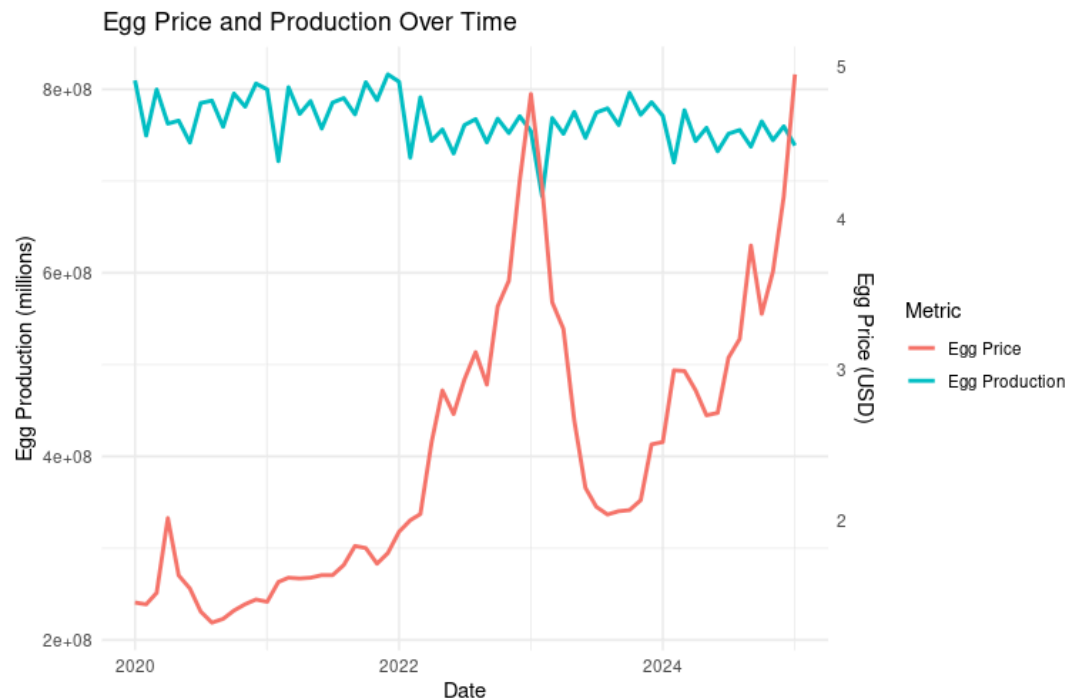
    x = "Date",

    color = "Metric") +

  theme_minimal()

```

The new Graph looks like this:



This graph better represents a visual relationship between egg prices and egg production. But there does not appear to be a strong correlation between egg prices and production. If egg production had a strong influence on egg prices, farmers would expect an inverse relationship (higher production equals lower prices and vice versa). However, this pattern is not clearly visible in the graph, leading to the conclusion that, given the datasets and the monthly time interval, there is no significant direct relationship. Egg production appears relatively stable, while prices are volatile.

But this could be due to the normalization not being done correctly or lacking statistical correlation manipulation. Since there has been a very well-established supply-demand model asserting relationship between price and production of consumable goods<sup>1</sup>.

<sup>1</sup> Fox, K. A. (1951). Relations Between Prices, Consumption, and Production. *Journal of the American Statistical Association*, 46(255), 323–333. <https://doi.org/10.2307/2280510>

This suggests other factors (such as inflation, feed costs, supply chain disruptions, or consumer demand) may play a greater role in price fluctuations than production itself.

### **Further Exploration**

The dataset limitations and scope of the research have a strong need for further exploration. The visualization needs more analysis such as correlation coefficient between egg production and prices to find a statistical relationship. And testing for lags would also be useful to find which variable affects the other.

- Top high-level insights based on the analysis

The top insights are that:

1. Egg production alone does not dictate egg prices, so farmers should not expect increases in production to lead to high profits, but should look at input costs, exogenous factors such as a bird flue, and supply chain forces to increase profitability.
2. Egg prices have occasional spikes that could be leveraged to maximize profits by predicting when they might occur and adjust production accordingly.
  - a. Farmers should track market trends and adjust their supply accordingly. Strategies to follow, given the data results would be that if egg prices are expected to rise, delaying sales could result in higher profits; and if prices are expected to fall, selling eggs quickly may be the best strategy.