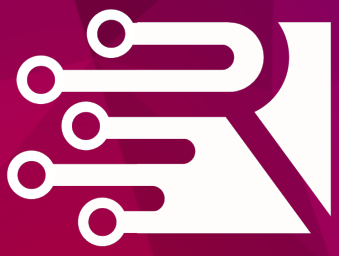
University of Texas at San Antonio

2024 Rowdy Datathon

Data Challenge



Juan B. Gutiérrez

October 5, 2024

Contents

Chapter 1

2024 Data Challenge - Monarch Butterflies

1.1 Problem Description

NOTE: The following describes a hypothetical scenario. While ecological pressures on pollinators are real, the crisis outlined below has not occurred... yet.

As an intern at a national laboratory, you are tasked with responding to a global crisis. Over the past year, there has been a precipitous decline in pollinator populations worldwide. Immediate action is required to reverse the trend; however, a thorough understanding of the problem is necessary before proposing a solution.

Your responsibility is to select and analyze relevant data, interpret your findings, and recommend a strategy. Multiple teams are working on different aspects of this issue, and your group has been assigned to focus on monarch butterflies.

Your report must address the following:

- A detailed description of monarch butterfly population fluctuations.
- An investigation of factors that may be contributing to these population changes, and the potential consequences for human well-being.
- A recommendation for resource allocation, prioritizing the most significant factors identified in your analysis.
- A focused analysis of the monarch butterfly migration through Texas, with comparisons to other states.

Your team will deliver an executive report supported by an appendix containing technical details.

1.2 Motivation

The collapse of pollinator populations would have severe consequences for human food production worldwide. In Texas, the disappearance of pollinators would severely affect the state's agriculture, economy, and ecosystems. Key pollinator-dependent crops, such as melons, peaches, and pumpkins, would see significant declines in yields, reducing local food diversity and raising prices. Livestock producers would face higher feed costs due to reduced forage crops like alfalfa, affecting meat and dairy production. The loss of pollinators would also disrupt ecosystems, harming biodiversity and wildlife habitats. Economically, regions that depend on agriculture, such as the Rio Grande Valley, would experience downturns, leading to job losses and economic strain across the state.

- **Decline in Crop Yields**: Pollinator-dependent crops, such as fruits, nuts, and vegetables, would experience significant yield reductions. Crops like almonds, apples, and melons would face near-total failure.
- **Loss of Crop Diversity**: Nutrient-rich foods that rely on pollination, such as berries, nuts, and many vegetables, would become scarce, leading to reduced dietary diversity and over-reliance on staple crops like wheat and corn.

- **Increased Food Prices**: Reduced supply of pollinator-dependent crops would increase food prices, making fruits, vegetables, and nuts less accessible, especially in low-income areas.
- **Manual Pollination Costs**: Manual pollination for high-value crops like almonds and vanilla would raise production costs, further driving up food prices. For large-scale crops, this is often unfeasible.
- **Loss of Ecosystem Services**: The decline of pollinated wild plants would lead to habitat loss and biodiversity declines, affecting the wider ecosystem and wildlife that depend on these plants.
- **Impact on Animal Products**: Pollinator-dependent forage crops like alfalfa and clover would decline, increasing livestock feed costs and reducing dairy and meat production.
- **Nutritional Impacts**: The reduction of fruits and vegetables would result in poorer human nutrition, increasing the risk of vitamin deficiencies and diet-related diseases.
- **Economic Impacts**: Agriculture sectors reliant on pollinator-dependent crops would suffer economic losses, particularly in regions that export high-value crops. This would affect the broader food supply chain.

1.3 Monarch Butterflies

Figure 1.3-1: Female monarch butterfly. Photography by Kenneth Dwain Harrelson, May 29, 2007. Image source: MediaWiki, https://w.wiki/BJJh. Accessed on September 3, 2024. Used under CC BY-SA 3.0 license.

Monarch butterflies (*Danaus plexippus*) are well-known for their striking orange and black wings and their extraordinary migratory behavior. Native to North America, monarchs are unique in their long-distance, multigenerational migration. They travel from breeding grounds in the United States and Canada to overwintering sites in central Mexico, primarily in the oyamel fir forests of Michoacán and Mexico State. A smaller western population migrates to coastal California for overwintering.

The monarch's life cycle consists of four stages: egg, larva (caterpillar), pupa (chrysalis), and adult. Monarchs lay their eggs exclusively on milkweed plants, as the caterpillars feed solely on milkweed. This plant provides the caterpillars with cardenolides, chemical compounds that make them toxic to many predators. As adults, monarchs feed on nectar from various flowers, which contributes to pollination.

The monarch migration spans several generations. Monarchs born in late summer and early fall enter a reproductive diapause, allowing them to live up to eight months, long enough to complete the migration to Mexico. These monarchs overwinter in Mexico, then begin the journey north in spring, reproducing along the way. The migratory relay continues with the next generations, which live only a few weeks during the spring and summer.

1.3.1 Importance of Monarch Butterflies

The decline of monarch butterfly populations is a significant ecological issue with broader implications for biodiversity, ecosystems, and human interests. Monarch populations have experienced dramatic decreases over the past few decades, primarily due to habitat loss, climate change, pesticide use, and changes in agricultural practices. The importance of addressing this decline goes beyond merely preserving a single species; it reflects the interconnectedness of ecological systems and the role of pollinators like monarchs in maintaining the health of ecosystems.

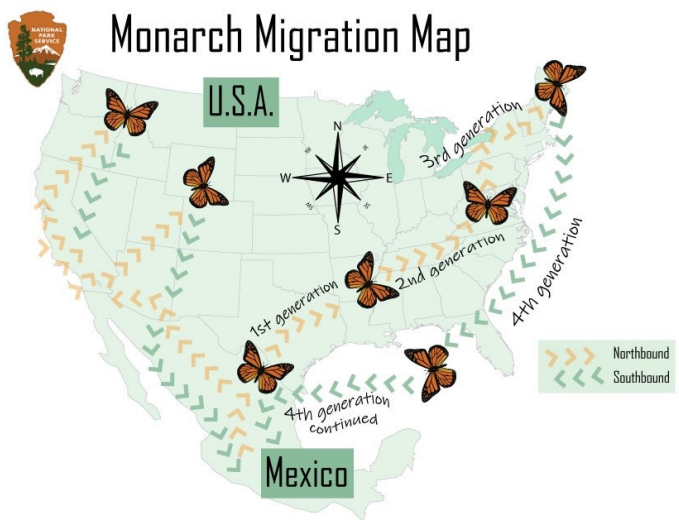


Figure 1.3-2: Monarch butterfly migration pattern. Image source: MediaWiki, https://w.wiki/BJJ$. Accessed on September 3, 2024. This image or media file contains material based on a work of a National Park Service employee, created as part of that person's official duties. As a work of the U.S. federal government, such work is in the public domain in the United States.

One major driver of population decline is the loss of milkweed, the only plant on which monarch caterpillars feed. Habitat destruction, especially due to agricultural expansion and urban development, has led to a sharp reduction in milkweed availability. The widespread use of herbicides, particularly those linked to genetically modified crops, has further exacerbated the issue. Pesticides used in farming not only destroy milkweed but also directly harm monarchs, both in their larval and adult stages.

Climate change also plays a critical role in the decline of monarch populations. Extreme weather events such as droughts, storms, and unseasonal temperature fluctuations can destroy monarch habitats and disrupt migration patterns. Warmer temperatures may alter the timing of migration and breeding, leading to mismatches between monarchs and the availability of necessary resources like milkweed and nectar plants. The monarchs' overwintering sites in Mexico and California are particularly vulnerable to deforestation and climate-related habitat shifts.

The decline of monarchs is important not only for the species itself but also for the ecosystems they inhabit. Monarchs are key pollinators, and their migration

contributes to the pollination of various plants, some of which are important to agriculture. A loss of monarchs would disrupt these pollination services, potentially affecting biodiversity and food production.

Furthermore, monarchs are a flagship species, meaning their conservation helps raise awareness about broader environmental issues. Protecting monarchs can serve as a catalyst for efforts to conserve other pollinators and habitats, benefiting biodiversity as a whole. The monarch migration is a natural phenomenon with cultural and educational significance, particularly in regions where the butterflies overwinter or pass through during migration. The loss of this species would represent a cultural and environmental degradation.

Addressing the decline of monarch populations is therefore essential for maintaining ecosystem balance, supporting biodiversity, and protecting the broader environmental and cultural importance of this iconic species.

1.4 Data Description

Your team has specific data sources for this analysis. You can only use approved data available at the online folder for the Rowdy Datathon data challenge.

US Environmental Protection Agency Air Quality System (EPA AQS) API
Data: API.
(https://aqs.epa.gov/aqsweb/documents/data_api.html). This API is the primary place to obtain row-level data from the EPA's Air Quality System (AQS) database.

US Environmental Protection Agency Air Quality System Pre-generated Files
Data: Select data appropriate for the inquiry from the web page.
(https://aqs.epa.gov/aqsweb/airdata/download_files.html). This page contians meterological and air quality data since 1980.

A Temporal Map of Monarch Migration
Data: Web reference.
(https://www.texasento.net/fall_peak.htm). Map compiled by Mike Quinn, a Texas entomologist.

Journey North
Data: Web scraping.
https://journeynorth.org/ Journey North is a citizen science platform that tracks the migration of species like monarch butterflies and birds across North America. It encourages public participation by allowing individuals to report sightings, contributing to real-time maps and data on migration patterns.

eButterfly
Data: Web scraping.
(https://www.e-butterfly.org/ebapp/en/species/profile/307). eButterfly provides rich data sources for basic information on butterfly abundance, distribution, and phenology at a variety of spatial and temporal scales around the globe.

US Department of Agriculture Pesticide Data Program
Data: USDA_PDP_AnalyticalResults.csv,
USDA_PDP_AnalyticalResults_DataDictionary.pdf
(https://www.ams.usda.gov/datasets/pdp/pdpdata). The Pesticide Data Program (PDP) is a national pesticide residue monitoring program and produces the most comprehensive pesticide residue database in the U.S.

Inter-university Consortium for Political and Social Research (ICPSR): County-level crop area in the USA 1840-2017
Data: 115795-V3.zip
(https://www.openicpsr.org/openicpsr/project/115795/version/V3/view) openICPSR is a repository for the deposit of replication data sets for researchers who need to publish their raw data associated with a journal article so that other researchers can replicate their findings.

1.5 Criteria

Results will be evaluated according to the following requirements:

- **Skill level**: The level of inquiry must be proportional to skills. A participant who claims a lower level of expertise could be disqualified.
- **Compelling presentation**: You must enable your colleagues to share your numerical and graphical results directly with legislators and citizens through executive summaries. This lay audience should find your summaries and implications to be understandable and convincing. Express your results in ways that can be acted on.
- **Analysis comprehension**: Before a single line of code is written, before a single byte of raw data is processed, you must be able to tell the story of what is the progression of steps that will be undertaken in analysis.
- **Sound technical methods**: You may cite the analyses of others, but your supervisors want to see the methods that you have invented or adopted for calculations (which should be accompanied by error bars, if possible). Your colleagues must have confidence in your results in order to present them to others.
- **Awareness of the data context**: All data have bias. Before, during and after analysis, it is essential to identify biases in the data and articulate clearly how these biases influence all steps of analysis and interpretation.
- **Reproducible results**: You must have your results confirmed by an independent team. That is, enable the independent team to replicate your results by describing your data and methods in detail.

Table 1.1: Evaluation Criteria

CRITERION	% WEIGHT
1. Compelling presentation - Informative	10%
2. Compelling presentation - Understandable	10%
3. Analysis comprehension	20%
4. Sound technical methods	20%
5. Awareness of the data context	20%
6. Reproducible results	20%

Chapter 2

Analysis Activities

2.1 Data Source

You should have received a flash drive containing all necessary files. As a backup, the data is also available in the online folder for the Rowdy Datathon data challenge.

2.2 Tasks for All Preparation Levels

2.2.1 Background Information

Review the Preface and Chapters 1, 2, and 3 in the *Rowdy Datathon Supplementary Material* to familiarize yourself with the context of the challenge.

2.2.2 Generative AI

If you attended the seminar on Large Language Models (LLMs), use the LLM API as discussed. Otherwise, complete the following steps:

- Register for a free account on ChatGPT.
- In the bottom left corner of the interface, next to your name, click the three dots and select "*Custom Instructions.*"
- In the field labeled "*What would you like ChatGPT to know about you to provide better responses?*," write something similar to, "*I am a student working on a data competition. My areas of expertise are...*"
- In the field labeled "*How would you like ChatGPT to respond?*," include instructions such as "*I need source code with ample documentation and verification steps*," along with any other specific preferences.

Customizing ChatGPT allows you to define parameters that improve the quality of responses. You may need to iterate on these instructions to achieve optimal results.

2.3 Tasks for the Beginner Level

1. Access the file:
   USDA_PDP_AnalyticalResults.csv
   This file contains data on pesticide use across all U.S. states.
2. Follow the instructions in Section 2.2 of *"2024 Rowdy Datathon Supplementary Material"* to create a color map that shows each state's pesticide concentration, with a focus on the pesticide with the highest concentration in each state.

3. Visit the Journey North website and map the number of monarch butterfly sightings by state. You can collect data by hand or using the Web scraping approach described in Chapter 4 of *"2024 Rowdy Datathon Supplementary Material"* .

2.4 Tasks for the Intermediate Level

1. Visit Journey North and create a map displaying the number of monarch butterfly sightings by county. Since the dataset only includes city and state information, use an LLM API to programmatically infer the county from the city. Follow the instructions in Section 2.2 of *"2024 Rowdy Datathon Supplementary Material"* to color each county according to the number of monarch sightings.

2. Utilize the U.S. Environmental Protection Agency (EPA) Air Quality System Pre-generated Files or the EPA AQS API to gather air quality and temperature data for the dates relevant to your study of Journey North data.

3. Perform a statistical analysis to investigate whether there is a significant correlation between temperature, air quality, and monarch butterfly decline.

2.5 Tasks for the Advanced Level

1. Can you infer pesticide use by county based on the Pesticide Data Program (PDP)? If so, create a map illustrating this data. What additional data would be required to make this inference more precise? Study the file 115795-V3.zip.

2. Conduct a statistical analysis to examine whether there is a significant correlation between pesticide use and the decline in monarch butterfly populations.

3. One of your team members suggested that the National Institutes of Health (NIH) and the Centers for Disease Control and Prevention (CDC) maintain maps of disease incidence, such as Parkinson's. There is speculation of a correlation between pesticide use and infant mortality. Investigate whether there is a connection between pesticide use, the decline in monarch butterflies, and the incidence of human disease.

Acknowledgements

Acknowledgements for the Rowdy Datathon must go beyond mere formalities, for the event is a monumental effort that reflects a commitment to making data accessible and comprehensible to all.

Special thanks are due to the Student Chapter of the Association for Computing Machinery at the University of Texas at San Antonio. The event would have not been possible without the support of the School of Data Science at UTSA and the National Security Agency. To all contributors, your collective efforts have culminated in a Datathon that serves as both a platform for applied data science and as an educational crucible for emerging data analysts.

The phrase "*Data is everywhere, therefore data should be for everyone*" is not merely a tagline; it encapsulates the philosophy that guided us through countless meetings, debugging sessions, manual tests, and problem-solving endeavors. This project is not isolated but forms a part of our larger mission: to guide aspiring data analysts through the multi-faceted landscape of data science. The Datathon aims to bridge the gaps in a fragmented educational landscape, offering a holistic view of data analytics that is sorely needed in the field. Thank you for your time, your expertise, and most importantly, your unwavering commitment to the democratization of data.