

Exploring the accuracy of ChatGPT's data analysis and visualizations



Ginny Boehme (she/her), Science Librarian
Miami University



**Science has a
stats problem**

Statistics in medical research: Common

miCommon Mistakes in Statistical and Methodological Practices
Corof Sport Management Research

AssS How to get wrong results from good experimental data: a
in survey of incorrect applications of regression

Towards better research: A discussion of
some common mistakes in statistical
analyses

PhD Students and the Most Frequent Mistakes During Data
Interpretation by Statistical Analysis Software

AI to the rescue?

Generative AI is democratizing skills...

- GenAI is leveling the playing field and lowering the bar for entry in so many areas:
 - Writing
 - Coding
 - Graphic design
 - Data analysis
 - Data visualization
 - etc.



Image courtesy: [pexels.com/photo/persons-hand-with-silver-ring-5716037/](https://www.pexels.com/photo/persons-hand-with-silver-ring-5716037/)

...but is it right?

- GenAI training datasets are heavily based on existing literature and user feedback
 - Garbage in = Garbage out
- Hallucinations (i.e., incorrect information presented as factual) are still quite common
- Ethical questions should also be considered

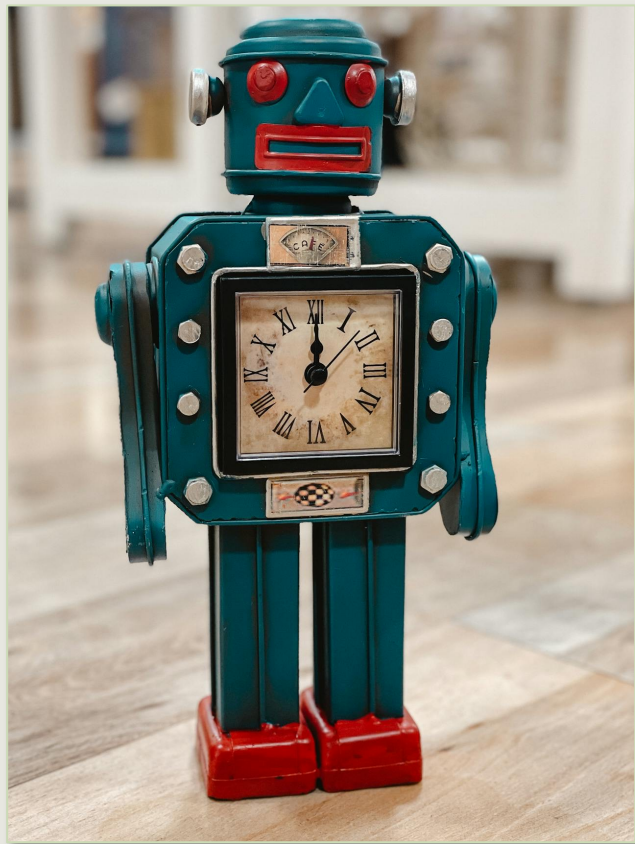


Image courtesy: pexels.com/photo/blue-and-red-robot-clock-9026734/

Students and stats

- Anxiety around statistics is common, as is uncertainty around quantitative methods
- Gender and level of achievement in mathematics courses have been shown to influence attitudes and self-confidence



Image courtesy: pexels.com/photo/close-up-on-hands-holding-pen-over-textbook-5527519/

Data Analyst (formerly Code Interpreter) is available to paid users of OpenAI's ChatGPT (as of July 2024)

My GPTs

+ Create

BO

ChatGPT

Data Analyst

Explore GPTs

Top Picks

Writing

Productivity

Research & Analysis

Education

Lifestyle

→

By ChatGPT

GPTs created by the ChatGPT team

1

DALL·E

Let me turn your imagination into imagery.

By ChatGPT

2

Data Analyst

Drop in any files and I can help analyze and visualize your data.

By ChatGPT

Exploring Data Analyst: Methods

- Uploaded longitudinal bird banding dataset
- Interrogated GPT about dataset
 - Attempted to get into student mindset/perspective
 - General questions about data
 - Specific questions about analyses
 - Additional requests to create visualizations
 - Wanted to see basic analyses and figures
- Initial "conversation" occurred May 2024; repeated Feb 2025

The Dataset

- Excel file
- One year of data per sheet: 2004 through spring of 2022
- Did not clean or manipulate
- Contains information about individual birds humanely captured, banded, and released at local avian research station
 - General info (species, age, sex, etc.)
 - Morphometrics (wing length, tail length, skull size, etc.)
 - Capture info (net location, date, time, etc.)



Image courtesy: Kyra Harvey
<https://www.usgs.gov/media/images/banding-american-woodcock>

Relevant Data

Wing Chord



Age

Indeterminable	0
After Hatching Year	1
Hatching Year	2
Not Used	3
Local	4
Second Year	5
After Second Year	6
Third Year	7
After Third Year	8
Not Attempted	9

Sex



Data Analyst Overview

The Good

Can quickly and easily perform analyses and transformations

The Bad

Takes some underlying assumptions at face value, although others have been improved since 2024

The Ugly

Visualizations have been improved since 2024, but may take several rounds of edits to be usable



Data Analyst: The Good

Query: extract data
into new file

Let's start by creating a new file with just the NOCA entries for 2016.
Are you able to create a CSV file of this data?



I've created a CSV file containing just the NOCA entries from 2016. You can download it using the link below:

[Download NOCA 2016 data \[-\]](#)



Let's add in the NOCA entries from 2015 and 2017 as well



I've added the NOCA entries from 2015 and 2017 to the previous data from 2016, and created a new CSV file containing all these entries. You can download this combined dataset here:

[Download NOCA 2015-2017 data \[-\]](#)

Data Analyst: The Good

Query: extract data
into new file

The model is more
polite than it used to
be.

Let's start by creating a new file with just the NOCA entries for 2016. Are you able to create a CSV file of this data?

I have created a CSV file containing the **Northern Cardinal (NOCA) data for 2016**. You can download it using the link below:

[Download NOCA_2016_data.csv](#)

Let me know if you need any modifications or further analysis! [-]



Let's add in the NOCA entries from 2015 and 2017 as well

I have created a CSV file containing **Northern Cardinal (NOCA) data for 2015, 2016, and 2017**. You can download it using the link below:

[Download NOCA_2015_2017_data.csv](#)

Let me know if you need any further modifications or analyses! [-]

Data Analyst: The Good

Query: run t-test comparing weights between sexes

May 2024

Let's continue with a T-test

>_

Before proceeding with a T-test to compare the weights between male and female Northern Cardinals, it's necessary to check if the variances of these two groups are equal using Levene's test. This will determine whether to use equal or unequal variances in the T-test. Let's perform Levene's test for equality of variances first:

Levene's test for equality of variances is approximately 0.000, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the variances between the two groups are not equal.

Now, let's proceed with the T-test comparing the weights between the two groups.

It seems there is a significant difference in weights between the two groups, possibly due to the difference in sex.

Let's convert the

Let's convert these columns to a proper numeric format and then rerun the T-test. I'll handle the conversion and perform the test: [\[↗\]](#)

After ensuring the data were correctly formatted as numeric types, the T-test for independent samples comparing weights between male and female Northern Cardinals was successfully conducted:

- **T-Statistic:** 4.901
- **P-Value:** approximately 2.41×10^{-6}

This very small p-value indicates a statistically significant difference in weights between males and females. The result suggests that there are meaningful differences in weight distributions between the sexes for Northern Cardinals within the data from 2015 to 2017.

Would you like to explore anything further related to this or perhaps another aspect of the dataset? [\[↗\]](#)

Data Analyst: The Good

Query: run t-test comparing weights between sexes

Provides the same basic info as in 2024, but takes a more "can-do" approach (less text-heavy, with less explanatory commentary).

Let's run a t-test on these to see if there is any significance

t-Test Results (Comparing Male and Female NOCA Morphometrics)

1. Weight (WT) Comparison

- **t-statistic:** 4.91
- **p-value:** 2.29×10^{-6} (very significant)
- **Interpretation:** There is a **statistically significant difference** in weight between males and females ($p < 0.05$). Males and females have significantly different average weights.

2. Wing Chord (WC) Comparison

- **t-statistic:** 3.79
- **p-value:** 0.00022 (highly significant)
- **Interpretation:** There is also a **statistically significant difference** in wing chord length between males and females ($p < 0.05$).

Data Analyst: The Good ~~Bad~~

After ensuring the data were correctly formatted as numeric types, the T-test for independent samples comparing weights between male and female Northern Cardinals was successfully conducted:

- **T-Statistic:** 4.901
- **P-Value:** approximately 2.41×10^{-6}

May 2024

t-Test Results (Comparing Male and Female NOCA Morphometrics)

1. Weight (WT) Comparison

- **t-statistic:** 4.91
- **p-value:** 2.29×10^{-6} (very significant)

Feb 2025

Data Analyst: The Bad

Query: least frequently captured species

for the year 2016, which species was captured the least frequently?



In 2016, the species captured the least frequently was labeled as "WAVI," with only 1 capture recorded. [\[↗\]](#)

Are there any other birds recorded in that year that were captured only once?



In 2016, there were several species that were captured only once, not just "WAVI". Here's the list of these species, each with only one capture:

- WAVI
- WIFL
- DOWO

Data Analyst: The ~~Bad~~ Good

Query: least
frequently
captured species

It's learning...

For the year 2016, which species was captured the least frequently?

In 2016, the following species were captured the least frequently, each only once:

- **WAVI** (Warbling Vireo)
- **WIFL** (Willow Flycatcher)
- **DOWO** (Downy Woodpecker)
- **LOWA** (Louisiana Waterthrush)
- **PIWO** (Pileated Woodpecker)
- **BTNW** (Black-throated Green Warbler)
- **TRFL** (Tropical Flycatcher)
- **SLTA** (Summer Tanager)

Data Analyst: The Bad

Query: linear
regression
comparing age to
wing chord

Data Analyst: The Bad

Qu
rec
co
wi

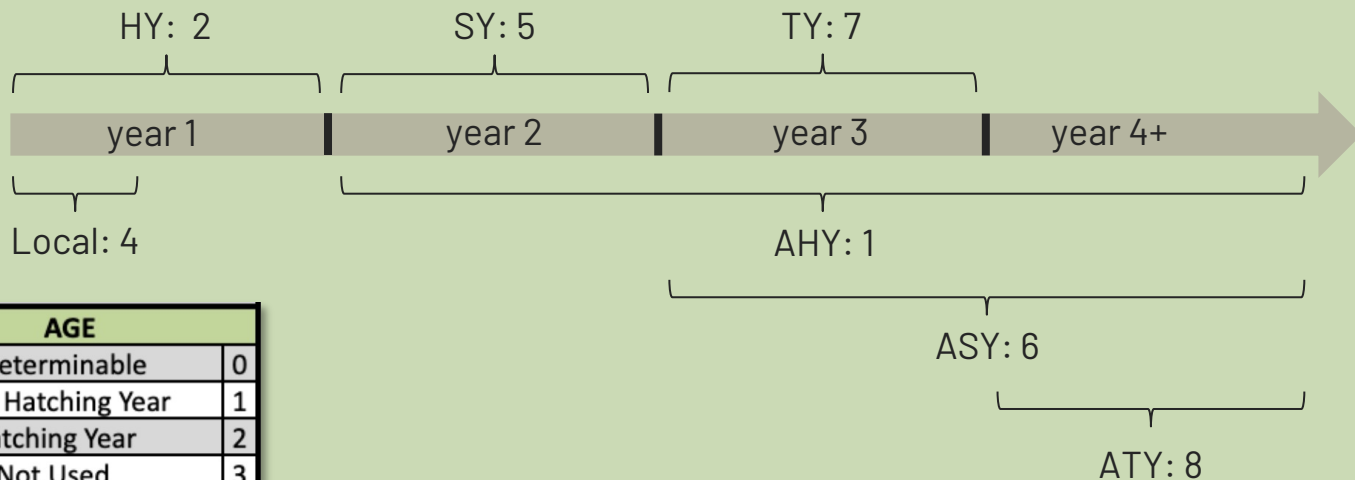
Let's refresh: Linear Regression

- Used to compare two **numerical** variables
- Wing chord → measured in millimeters, **numerical**
- Age → presented on a number-based scale, **not numerical**

Data Analyst: The Bad

Qu
rec
co
wi

Why is age not numerical here?



AGE	
Indeterminable	0
After Hatching Year	1
Hatching Year	2
Not Used	3
Local	4
Second Year	5
After Second Year	6
Third Year	7
After Third Year	8
Not Attempted	9

Data Analyst: The Bad

Query: linear
regression
comparing age to
wing chord

Your underlying assumption about the age data is wrong. This variable is purely categorical

Thank you for clarifying that the "Age" variable is purely categorical without any inherent order. With this understanding, we should consider alternative approaches to analyze the relationship between "Age" and "WC" (wing cord length) that better respect the categorical nature of the age data.

Here are two more appropriate methods to analyze the relationship between a purely categorical variable and a numeric outcome:

1. Analysis of Variance (ANOVA):

- **Purpose:** ANOVA tests whether there are statistically significant differences between the means of three or more unrelated groups. In this case, it would test if the mean wing cord length differs across the different age categories.
- **Assumptions:** The data should meet assumptions like homogeneity of variances (equal variances across groups) and the dependent variable should be approximately normally distributed within groups.

Data Analyst: T

Query: linear
regression
comparing age to
wing chord

It still takes multiple
rounds before the
model offers the
possibility that age
is categorical here.

Lin

Mod

-
-
-
-
-
-
-

Feb 2025

So age is numerical rather than categorical?

Great question! Whether **Age** should be treated as a **numerical** or **categorical** variable depends on how you plan to analyze it.

Is Age Numerical?

Yes, **Age** is **inherently numerical** because it represents a measurable quantity (e.g., 1, 2, 5, etc.). This allows for:

- **Mathematical operations** (e.g., calculating means, regression analysis).
- **Trends over time** (e.g., does wing chord decrease with age?).
- **Correlation analysis** (e.g., age vs. weight).

Can Age Be Treated as Categorical?

Yes, **in some contexts**. If age represents **distinct life stages** rather than a continuous variable, it might be better treated as **categorical**:

- **Example:**
 - 1 (Hatch-year)
 - 2 (After hatch-year)
 - 5 (Older adults)
 - U (Unknown)

Data Analyst: The Bad

Query: linear regression comparing sex to wing chord

Also not an appropriate analysis, but the model runs the test anyway.

Feb 2025

I want to do a linear regression comparing sex and wing chord

Linear Regression: Sex vs. Wing Chord (WC)

Since **Sex** is a **categorical variable** (Male = "M", Female = "F"), we need to **convert it into numerical form** for regression analysis. The best way to do this is by **dummy coding**:

- Assign **Male = 0** (reference group)
- Assign **Female = 1** (dummy variable)

I'll now perform the **linear regression** to see if **Sex significantly predicts Wing Chord (WC)**.

It looks like there are **missing or invalid values** in the dataset, which is causing the regression to fail. I will clean the dataset by removing rows where the **Sex or Wing Chord (WC) values are missing or invalid** and re-run the analysis. [\[↩\]](#)

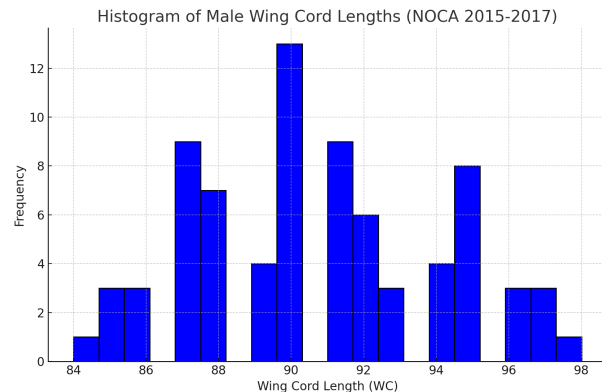
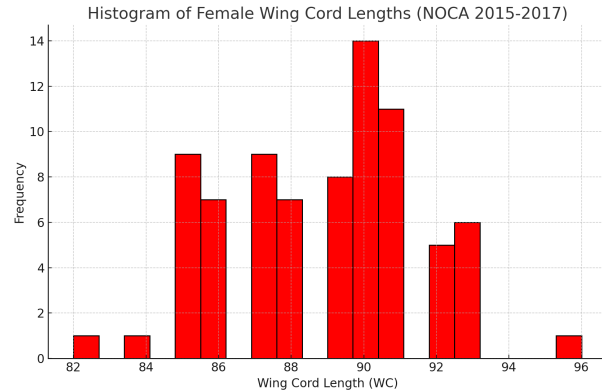
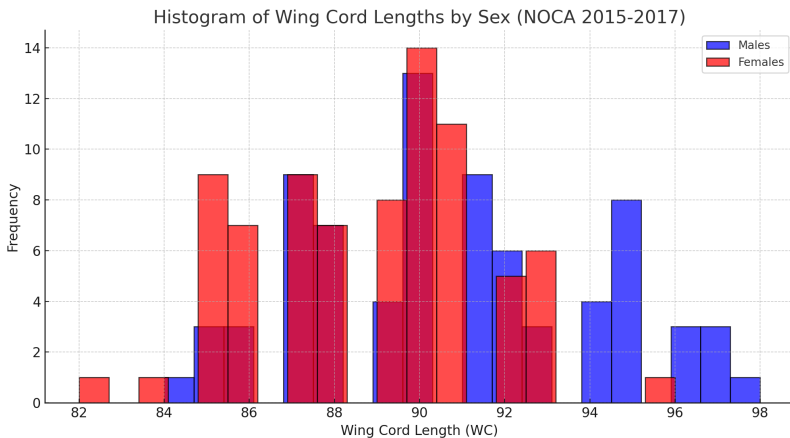
Linear Regression Results: Sex vs. Wing Chord (WC)

Model Summary

Data Analyst: The Ugly

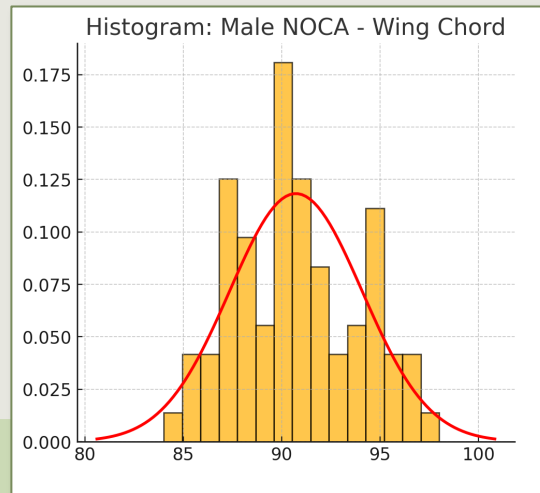
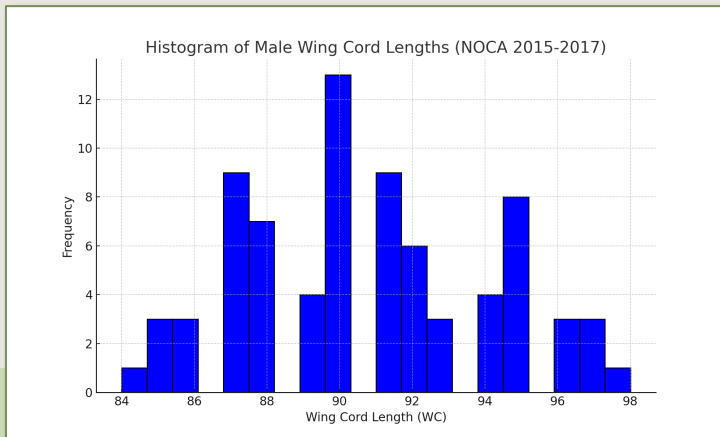
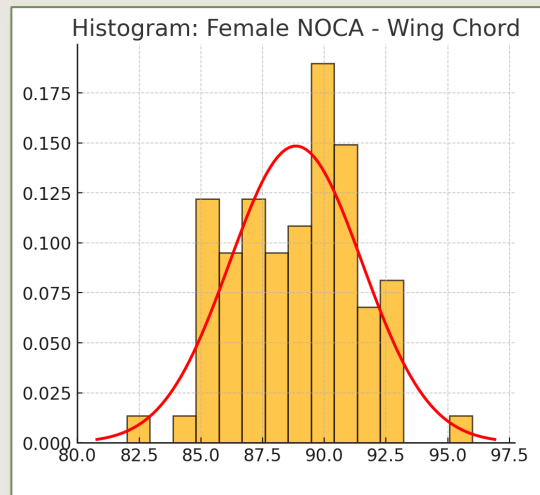
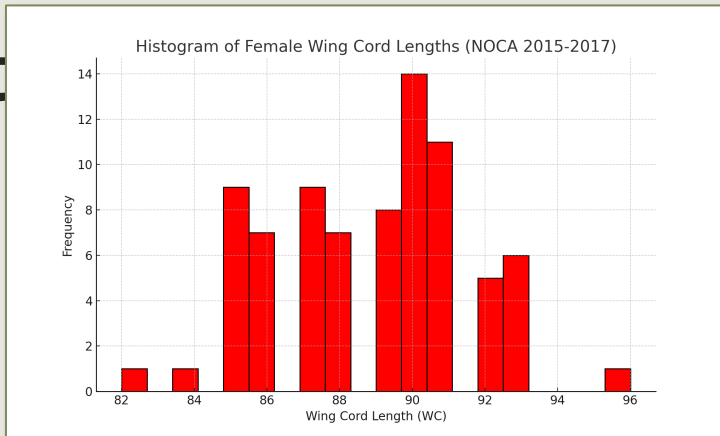
Query: histograms of wing chords by sex

Please show me the histograms for both sexes before we do anything else



Data

Query:

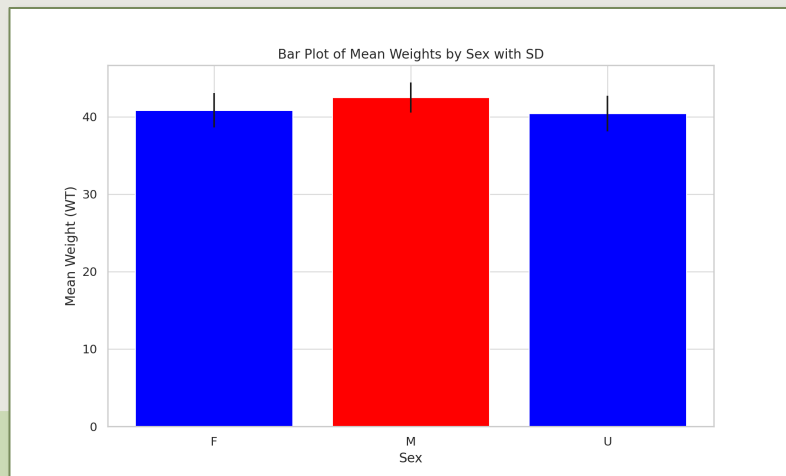
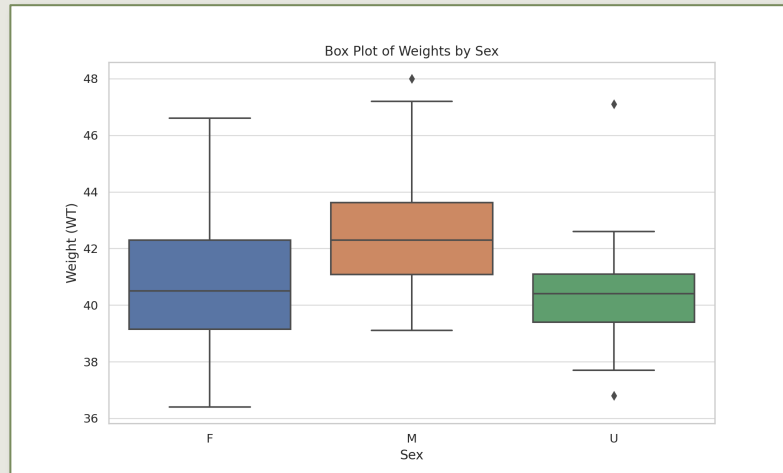
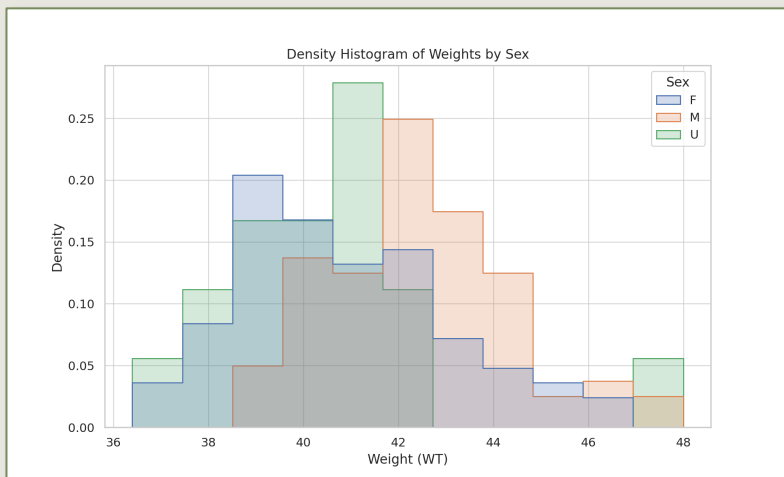


May 2024

Feb 2025

Data Analyst: The Ugly

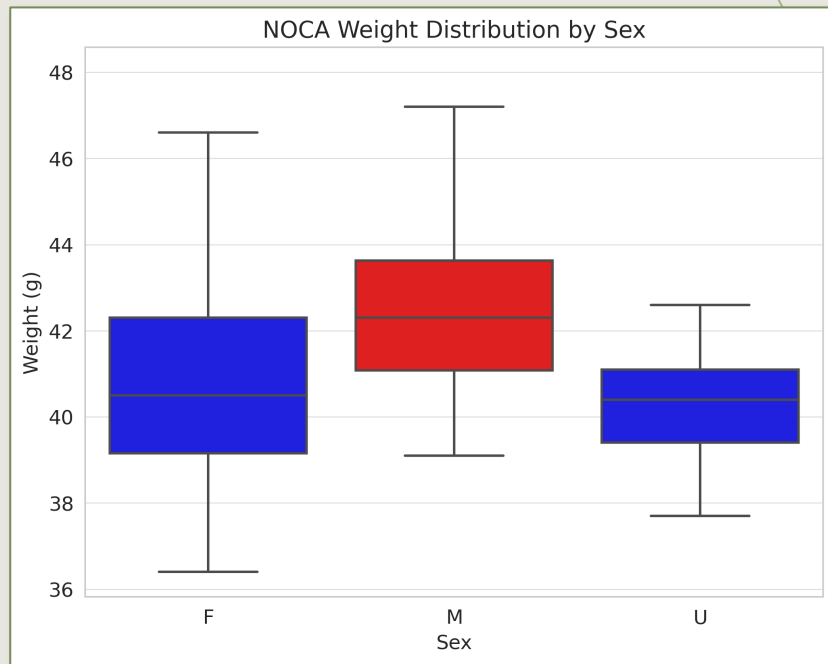
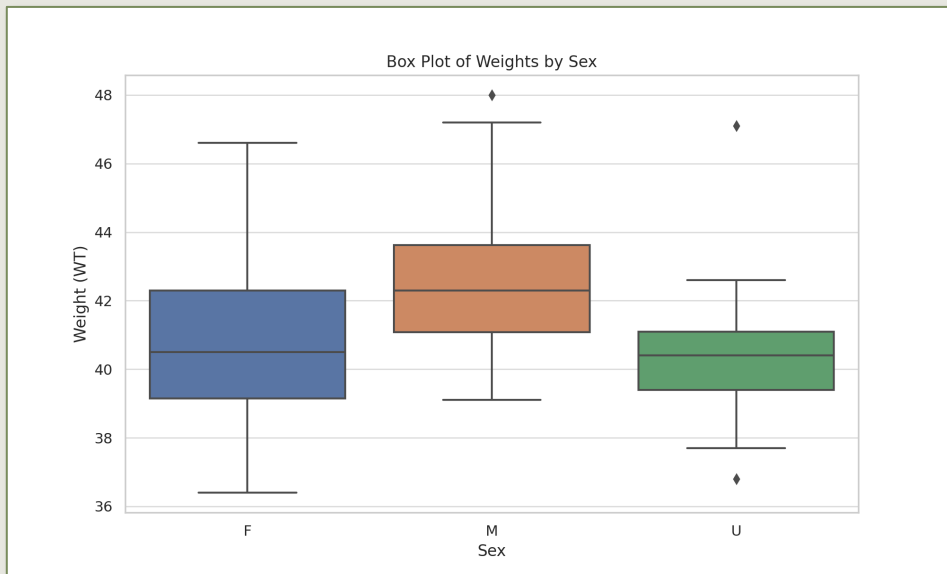
Query: provide other figures showing comparison of sex to wing chord



May 2024

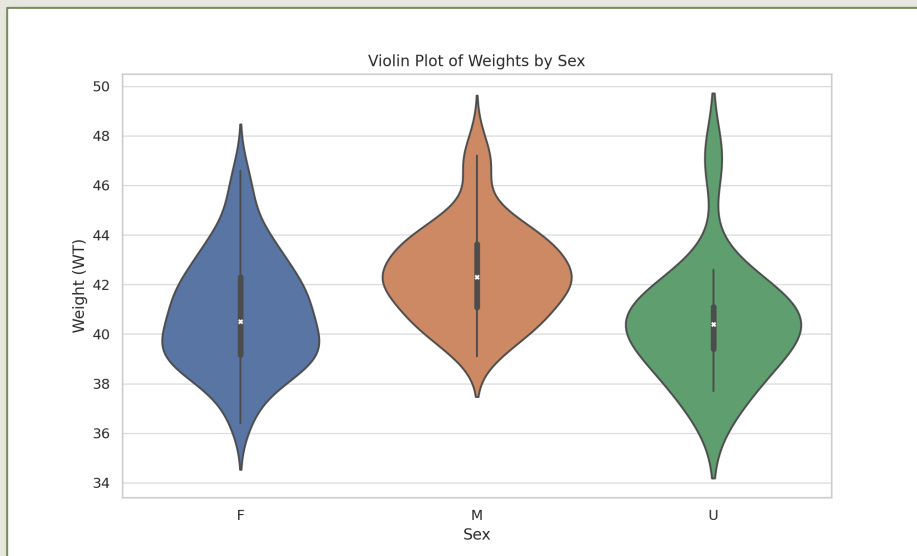
Data Analyst: The Ugly

Query: provide other figures showing comparison of sex to wing chord

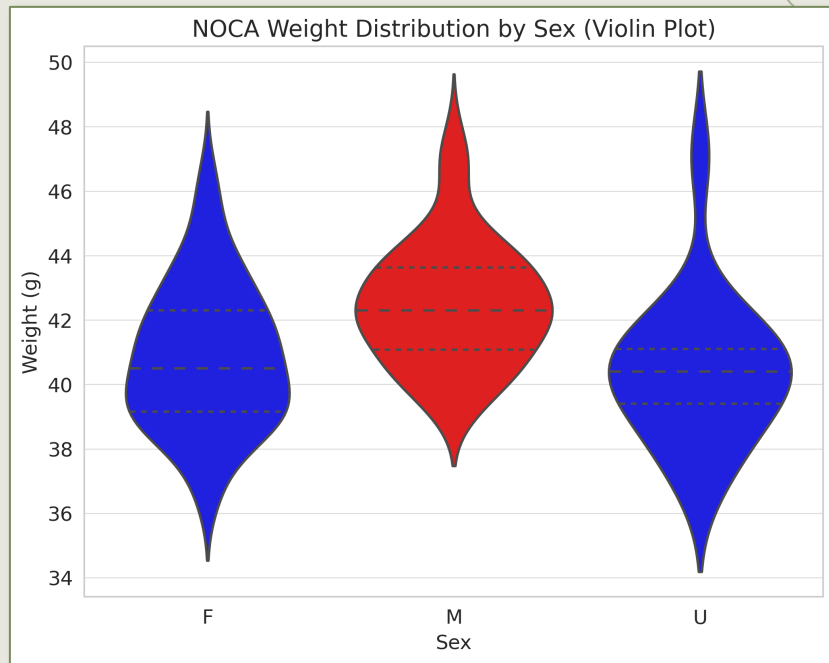


Data Analyst: The Ugly

Query: provide other figures showing comparison of sex to wing chord



May 2024



Feb 2025

Data Analyst Decision-Making

Then

Tended to require more prompting and direction from the user in order to make decisions about which analyses to run.

Now

Has the capability to "reason" and conduct analyses with very minimal input from user

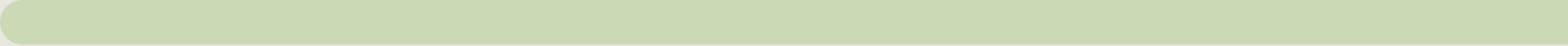


increased potential to introduce error and bias, particularly with large datasets and large questions with many possible variables

See example:

<https://chatgpt.com/share/67cfc3a5-e350-800f-8469-c99738b8ebea>





**As with all generative AI
tools, use with caution.**

Mistakes will be made; untrained students are unlikely to notice.



Thank you!

Questions?

boehmemv@miamioh.edu

Special thanks to:

Avian Research & Education Institute
Dr. David Russell

Link to cited references:

<https://zbib.org/7c2305c6d891465b93987763216752c9>

CREDITS: This presentation template was created by **Slidesgo**,
including icons by **Flaticon**, and infographics & images by **Freepik**.