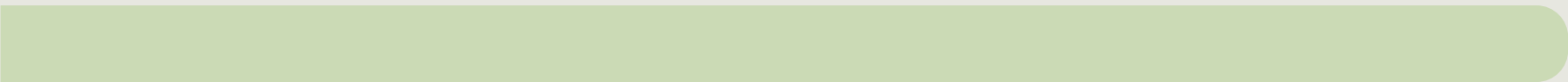


Exploring the accuracy of ChatGPT's Data Analyst tool

Implications for teaching data literacy



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



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
 - Excel file
 - One year of data per sheet: 2004 through spring of 2022
 - Did not clean
- Interrogated GPT about dataset
 - Attempted to get into student mindset/perspective
 - General questions about data
 - Specific questions about analyses
 - Additional requests to create visualizations
- "Conversation" occurred May 2024

Data Analyst: Overview

The Good

Can quickly and easily perform analyses and transformations

The Bad

Takes underlying assumptions at face value

The Ugly

Most visualizations look good at first, but do not stand up to scrutiny



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: run t-test comparing weights between sexes

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, indicating 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 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

Query: linear regression comparing age to wing chord

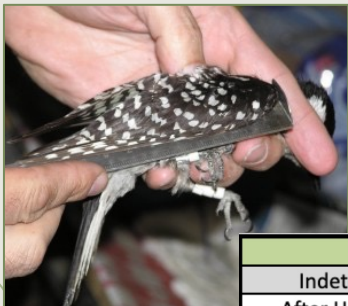


Photo by Bart Paxton
<https://cobbirds.org/2009/09/12/success-of-red-cockaded-woodpecker-translocations-in-virginia/>

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

I want to perform a linear regression comparing age to wing cord for all NOCA in this sheer



To perform a linear regression analysis comparing age to wing cord lengths for all Northern Cardinal (NOCA) data, we'll first need to ensure that both variables are in appropriate numeric formats and handle any missing values.

We will:

1. Convert 'Age' and 'WC' (Wing Cord) to numeric types if they aren't already.
2. Remove any rows with missing values in these columns.
3. Perform a linear regression to investigate the relationship between these two variables.

Let's start by preparing the data and then perform the regression analysis.

Data Analyst: The Bad

Query: linear regression comparing age to wing chord

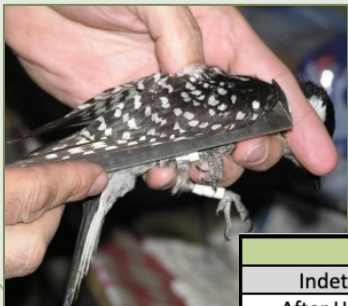


Photo by Bart Paxton
<https://cobbirds.org/2009/09/12/success-of-red-cockaded-woodpecker-translocations-in-virginia/>

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

Is a regression an appropriate analysis to perform when you have one categorical variable and one numeric variable?

Using regression to analyze relationships between a categorical independent variable and a numeric dependent variable is technically appropriate, but it requires careful handling. Here's a closer look at when and how regression can be applied in such scenarios:

- Categorical Independent Variable:** When the independent variable is categorical, it can be incorporated into a regression model through the use of dummy variables, as we did. This transforms each category level into a binary indicator variable, allowing the regression model to estimate separate intercepts (mean responses) for each category relative to a reference category. This approach effectively compares the mean of the dependent variable across the different levels of the categorical variable.
- Nature of Categorical Data:** The suitability also depends on the nature of the categorical data:
 - Ordinal Data:** If the categories have a natural order (e.g., stages of development), treating them as continuous might make sense if you expect a linear trend across

Data Analyst: The Bad

Query: linear
regression
comparing age to
wing chord

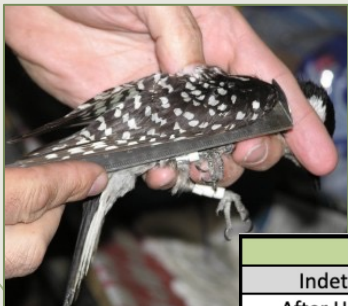


Photo by Bart Paxton
<https://ccb.birds.org/2009/09/12/success-of-red-cockaded-woodpecker-translocations-in-virginia/>

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

In the original spreadsheet, what type of variable is Age?

The "Age" variable, as observed in the data you provided, seems to represent categories rather than continuous numerical data. Often in biological studies, especially involving animals like birds, "Age" is recorded in discrete classes (such as hatch-year, juvenile, adult) or based on developmental stages that can be visually confirmed.

To understand exactly how "Age" was recorded and what each value represents (whether they are integers signifying years, categorical stages, or another coding system), we would need to refer back to any documentation, metadata, or variable definitions accompanying your dataset. If you don't have this information at hand, examining unique values and their frequency within the dataset could provide further clues on how "Age" was used.

Would you like me to analyze the unique values and frequencies of the "Age" variable from the data to give a better picture of how it's structured?

- **Ordinal Data:** If the categories have a natural order (e.g., stages of development), treating them as continuous might make sense if you expect a linear trend across

Data Analyst: The Bad

Query: linear
regression
comparing age to
wing chord

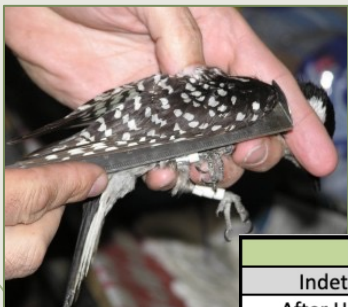


Photo by Bart Paxton
<https://ccbbirds.org/2009/09/12/success-of-red-cockaded-woodpecker-translocations-in-virginia/>

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

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:

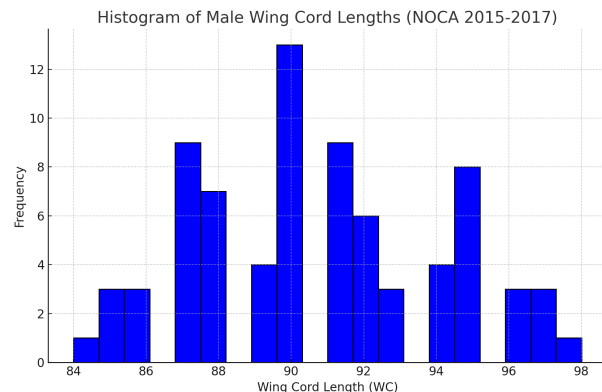
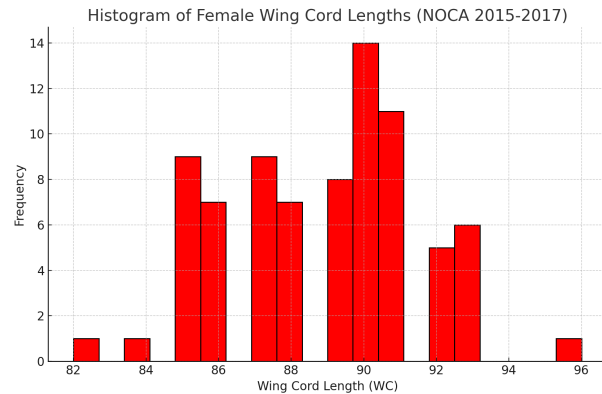
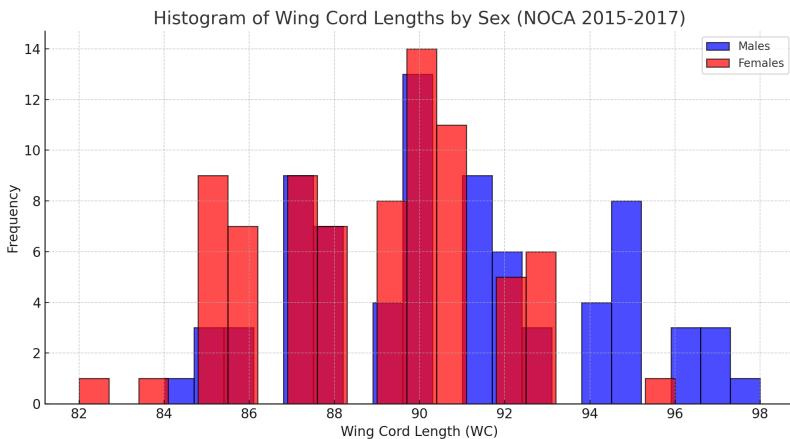
1. Analysis of Variance (ANOVA):

- **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: The Ugly

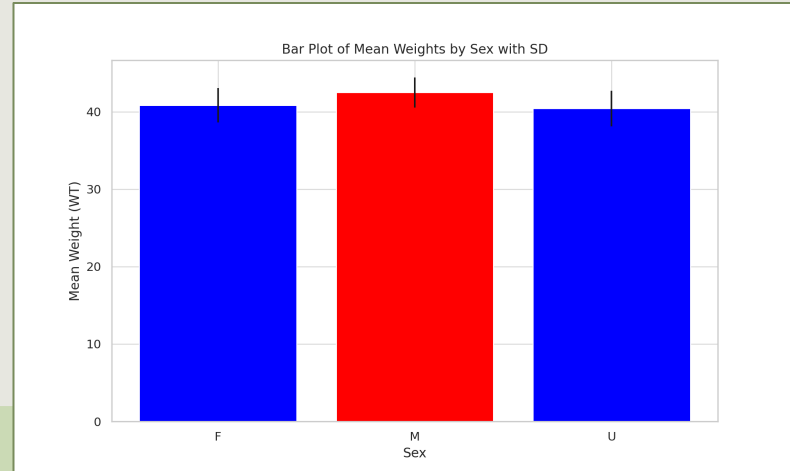
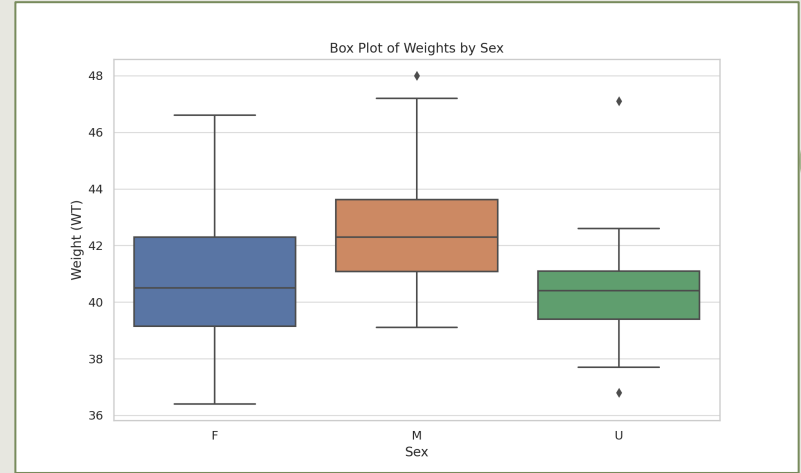
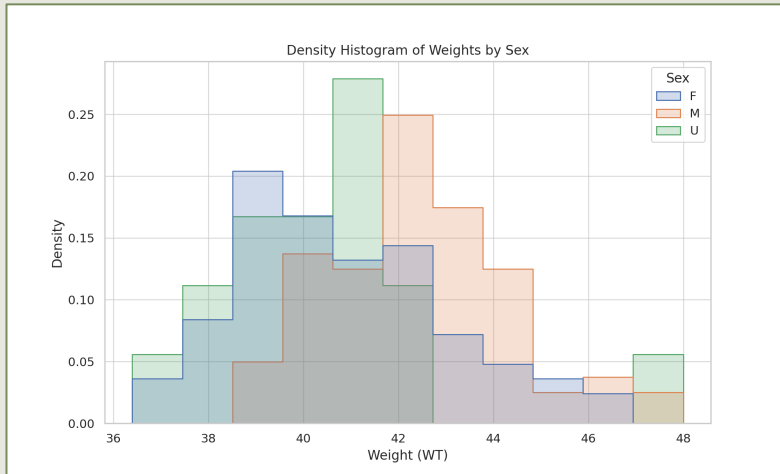
Query: histograms of wing chords by sex

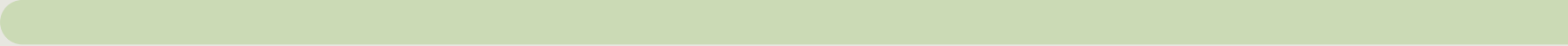
Please show me the histograms for both sexes before we do anything else



Data Analyst: The Ugly

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





**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 Russel

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