**Slide 3 Script:**

Hello everyone, we are the Queen Bees, and this is our final project for Entity’s Data Science Course, in partnership with Woz-U and Learning Source

**Slide 5 Script:**

I am Sara Slocum

I have college credits from both Santa Monica and Miami Dade Colleges, focused on Finance, Music, and Performing Arts.

I never completed a degree program, however my knowledge also comes from a lot of real work experience and professional certifications.

Most recently, these certifications include my Series 6 and 65 for Investment Advising as well as my life and health insurance certifications, but also I have held certs in real estate, cycling and group fitness instruction, lifeguarding cpr and first aid.

I’m currently self-employed, before becoming an investment advisor I was a retail store manager.

I consider myself a creative, empathetic leader.

Healing Manuka Honey & Signature Honey Latte

**Slide 9 Script:**

We used two datasets for this project:

The majority of our data came from a dataframe we found and Kaggle, but ultimately was pulled directly from the USDA govt website, pertaining to supply, production, trade, and consumption data for commodities worldwide. This data is listed yearly per country, dating from the 1960s all the way to 2022.

However, this dataset did not include Honey, so we found a second dataset on Kaggle that surveyed total honey production in the US by state per year, from 1998 -2012

We ended up adding all the state production values together to get total US production yearly, and appended this information to our USDA dataset.

The bulk of our analysis was focused on US Production values of commodities from 1998 – 2012 as it relates to honey.

**Slide 13 Script:**

Now that we have some background on Honey, lets look at how it fits into US Agriculture as a whole.

In this table on the left you will see the average yearly production, from 1998-2012 of our top producing crops, with the yearly average in Metric Tons, and its percentage of the US total underneath. Notice the top 3 crops significantly outproduce the rest: Corn at a whopping 53% of total US production, Milk at about 15% and Wheat at 11%. The next most produced commodity is Chicken, and it falls to a measly 2.5%. All of the other commodities not listed in this table, combined, make up the remaining 10% of total US production.

Since honey was not in the top 8 commodities produced by volume, anyone want to guess what percent of Honey makes up total US crop production?

Although Honey production can be indirectly influential to US agriculture as a whole, it only accounts for .01% of total US production. That number might seem small, but to put into perspective how much Honey we really produce, I’ve converted the yearly average production of 78,920 Metric Tons to number of Elephants, weighing on average 6 Metric Tons. Honey production might be small in comparison to other commodities, but its influence on US agriculture is as mighty as over 13,000 elephants.

Now, in the pie chart is the total sum of production between 1998-2012. It is more or less in line with what we see on the left table. Our data also showed that domestic consumption values paralleled production. With Corn, Dairy, and Wheat being the top consumed commodities in the US. Looking at a modern nutrition table, we can see that Americans have a lot of work to do in terms of being healthy. Although Corn is a vegetable, it is one of the only vegetables with no nutritional value, so maybe we ought to start increasing production and consumption of other, more nutritious fruits and vegetables.

One noticeable inconsistency we saw in production to consumption, is that although Orange was the top produced fruit in the US, Apples were actually the most consumed fruit, not just in the US but worldwide. We believe this is because oranges are being grown primarily for Orange juice.

**Slide 17 Script:**

With this background information in mind we’ll move into the main bulk of our analyses, evaluating how each commodity in our data set correlates with honey.

**Slide 18 Script:**

Our entire analysis was a 6-step process, beginning with a correlation matrix. Once we had that, we singled out the commodities whose production values correlated most highly with honey production values. We defined “High Correlation” to be at 65% or higher.

We then ran linear regression analyses on our high correlators to look further at how their production values have changed from 1998-2012, and to study the linear relationships between each commodity and honey during this time period. We took note of the equations produced by our regressions, that define each commodity’s relationship with honey to use later for predictive forecasting.

The last two steps in our process make up our post-analysis, answer our last evaluation question, about whether we can make future predictions on honey production. Since the honey data only went to 2012, but the other commodity data went to 2022, we first ran an ARIMA analysis on each high-correlator to obtain 2023 forecasts.

Our last step in the process was plug in the 2023 forecasted values to each commodity’s linear regression equation and solve for honey. This gave us estimated values for honey production in 2023, one for each highly correlated commodity. We then created a data frame with these predicted honey values, plotted the points into a scatterplot, and ran a K-Means cluster to observe the spread of the prediction points and make conclusions about our 2023 forecasts.

**Slide 23 Script:**

Category 4 of our commodity deep-dive analysis is the nuts and oils group.

This includes:

Almonds, Filberts, Pistachios, Walnuts, Peanut Oil, Olive Oil, Rapeseed Oil, Cottonseed Oil, & Soybean Oil

* Quick question, can anyone tell me what filberts are?
* This is actually something I learned from this project, Filberts is another term for Hazelnuts, one of my favorite nuts

Anyway, upon running the correlation matrix here, we found four high correlators with Honey:

Almonds at 81%, Walnuts at 74%, Rapeseed Oil at 73%, and Soybean Oil at 67% correlation

Here is the linear trend plot with Honey Production Values on our x axis and commodity production value on our y. We had to create an individual axis for soybean oil to make it plottable, as the production values far exceed the other commodities by tens of millions.

It is clear to see that all the high correlators in the nuts & oils category are negatively correlated with honey production. As the production of Honey increases, the production values of these four commodities decrease. An interesting find here is that the fluctuation in production values between the commodities all occur similarly at the same levels of honey production. Notice the decline across the board in these commodities as honey production increases from 75k to around 78k, and then the sharp spike up, peaking as honey hits 80k, followed by another drop. This is something I’d like to look further into given more time.

At the bottom are the equations received from each linear regression analysis.

To wrap up the analyses on Nuts & Oils, we found that Honey accounts for:

* + 63% of Almond production
  + 52% of Walnut production
  + 50% of Rapeseed Oil production
  + And 41% of Soybean Oil production on any given year

Sonya will take it away for our miscellaneous commodity analysis

**Slide 30 Script:**

Last but not least, we want to give some disclaimers about the analyses we have done, and touch on where we would go next to further this study.

Limitations

The three biggest limitations we faced during our study were Time, Data, and Knowledge

- Our dataset is so large, and there are so many facets to look at this topic, we just didn’t have enough time to accomplish everything we wanted to do

- As we got deep into wrangling this data, we found ourselves questioning some of the accuracy of the entries. Certain things didn’t add up or make sense. Although the majority of the data is pulled directly from the USDA.gov website, we might want try to find other agricultural data elsewhere that is more comprehensive and thorough, to compare to or update our existing data.

- We also realized some of the questions we wanted to answer required analyses beyond what we learned in the course.

- Because of these limitations combined, there are some things we wanted to do that we just couldn’t get to. For example, upon running our linear regressions, we ran into some minor issues with our test assumptions, that given more time and knowledge we would adjust for.

Next Steps

With that said, here are the next steps we want to take to continue improving and expanding on our project:

- Finding Better Data to improve data set – perhaps from some other official sources on global agricultural production and trade

- Fixing the Homogeneity issues common across many of the linear regressions we did on each commodity

- Take bigger effort on removing outliers, which is a massive wrangling task considering the size of our dataset

- Reading the many 400 page texts Professor Raetano gave us on ARIMAs

- to develop deep understanding of them and be able to run them all manually, fine-tuning each parameter, with the goal of improving the R2 scores we got

- Looking into other factors that effect production values and incorporating them to be able to make more precise forecasts and explain causation

- Exploring the massive amounts of global data on Trade and Consumption in the USDA dataset

Although we still have plenty to do with this existing project, we already have an idea in mind for our next project. Please stay tuned for the Queen Bee’s next project (drum roll please)…

**Slide 31 Script:**

How could the information we have based on our data and analyses impact futures pricing?

That wraps up our presentation of US Honey Production, as it relates to other commodities. We would like to dedicate the last few minutes to Q & A to answer any questions you may have.

***Slide 19 Script:***

*Upon attempting to wrangle and analyze our data, we quickly realized how extensive the dataset was,*

*Wrangling alone was difficult, but trying to run an analysis with every commodity all at once just wouldn’t work because of multicollinearity. Altogether, the commodities were overfitting the data. So, we divided our data up into 5 groups that each of us could manage.*

*Category 1 is the Meats & Dairy group*

*Category 2 is Fruits & Vegetables*

*Category 3 is Grains & Oats*

*Category 4 is Nuts & Oils*

*And Category 5 is all the miscellaneous commodities that didn’t quite fit into any of the other groups*

*All of the values of commodity production you will see are listed in Metric Tons*