**Project Proposal – The Queen Bees**

**Executive Summary**

At the end of ENTITY Academy’s Data Science program, The Queen Bees expect to complete a final project regarding USDA production of commodities such as wheat, corn, sugar, and honey along with the factors that contribute to the level of production of said commodities. We intend to make known the correlation (if any) of honey production with the production of other USDA commodities, as well as analyze the number of colonies and yield per colony changes, and levels of production value.

**Business Objectives**

To showcase the skills that The Queen Bees have acquired throughout the Data Science program. We will be using R, Python, and Tableau to wrangle, analyze, and visualize the “USDA Production Supply And Distribution” and “honey production” datasets made available on Kaggle.

Questions to Evaluate Throughout this Project:

**(Pre-Analyses)** - Honey Production Data (Neha & Sindy)  
*Visual analyses per state, heat map, & statistical analyses*

1. What states have the highest production value (price per lb)? – *heat map - Max/ Average of priceperlb per state.*
2. How has yield, numcol, and honey production changed over time, per state and as a country? -  
   *High producer state analysis vs low producing state analysis*
3. Other relevant / interesting findings

**Group Analyses** - Commodities vs. Honey  
*(Each team member answers following questions based on their category)*

1. Does honey production correlate with the production of any other USDA commodities in your category/group? Which are more highly correlated? - *Correlation*
2. Have the production levels of each commodity in your category changed over time? - *Linear Modeling / MANCOVA*
3. Other relevant / interesting findings about your category

**(Post-Analyses)** - USDA Global Data (Sara, Nic, Sonya)  
*Analyzing only top 3-10 honey-correlated commodities*

1. How are commodity production levels trending globally in more recent years? – *Time analyses, linear modeling, MANCOVA*
2. Can we make any predictive analyses about honey production based on correlation patterns? *Predictive analysis / conclusions*
3. Looking at another aspect of Attribute Description (import/export/domestic consumption), what conclusions can we make globally based off of our production findings? *Visual modeling, heat map*

At the very end of the project, The Queen Bees will explain our work in layman’s terms, and present our findings to the students, faculty, staff, and potential employers, along with other interested parties via Zoom, or other telecommunication software.

**Background**

In the United States, the honeybee generates approximately $700 million due to the production of honey and the use of beeswax in an array of consumer products from candles to soaps, to reusable wraps replacing single-use plastic products. Since 1947 the number of managed honeybee colonies has dropped from 6 million to less than 2.5 million today. Commercial beekeepers have been suffering from high levels of hive loss, crippling pollination demand for a wide variety of crops, many of which are market-traded commodities.[[1]](#footnote-1)

The estimated value of the added revenue to crop production from pollinators is $18 billion with honeybees making up a significant portion of this value. Honeybee decline was first noticed as an issue in 2006, at which time it was determined that the collapse of many colonies was due to a combination of pesticide use and pollutants, agricultural production practices, and factors related to loss of habitat.[[2]](#footnote-2)

The health of the honeybee population is critical to 70% of plants grown and has been linked to improved yields in many other products that do not require pollinators for survival.[[3]](#footnote-3) With honeybee health critical to many of the commodities currently on the market, is there a link between the production of honey, an indicator of colony health, and the production of other market traded commodities?

**Scope**

With honeybee health critical to many of the commodities currently on the market, we are trying to figure out is there a link between the production of honey, and the production of other market traded commodities?

We’ll be using what we learned in the program.

For now, we are going to focus on just commodity production affected by honey production and not the monetary value of how it affects inflation.

We are planning to keep the effects of climate change and/ or other factors for causes of colony losses as secondary data in case needed.

**Functional requirements**

* **Programs / languages to be used –**

R, Python, Tableau, Trello Kanban, GitHub.

* **Information about your dataset and source –**

We’ll be using two datasets -

* 1. **USDA\_Production\_Supply\_And\_Distribution –** It comes from Kaggle. It provides information on USDA ‘s estimations of Production, Supply and Distribution of soft commodities worldwide over the years.

Some variables of the **USDA\_Production\_Supply\_And\_Distribution** dataset is provided below:

* *Commodity\_Description:* Gives the product name.
* *Country\_Code:* Codes for Different Countries.
* *Country\_Name:* Country’s Full Name.
* *Year:* Year for the Attribute event occurrence
* *Description:* Type of data (Production, Consumption, Distribution and their derivatives)
* *Unit\_Description:* Unit of Measure. This is helpful because not all commodities use the same UOM as standard.
* *Value:* Quantitative value

**2. HoneyProduction** – This dataset also comes from Kaggle.

This dataset provides insight into honey production supply and demand in America by state from 1998 to 2012.

Some variables of the honeyproduction dataset are provided below:

* *numcol*: Number of honey producing colonies. Honey producing colonies are the maximum number of colonies from which honey was taken during the year. It is possible to take honey from colonies which did not survive the entire year
* *yieldpercol*: Honey yield per colony. Unit is pounds
* *totalprod*: Total production (numcol x yieldpercol). Unit is pounds
* *stocks*: Refers to stocks held by producers. Unit is pounds
* *priceperlb*: Refers to average price per pound. Unit is dollars.
* *prodvalue*: Value of production (totalprod x priceperlb). Unit is dollars.
* **Variables and analyses you'll utilize to answer the questions in the business objectives section**

Some of the analyses that will be used are:

*Visual analyses per state, heat map, & statistical analyses*

*heat map - Max/ Average of priceperlb per state.*

*High producer state analysis vs low producing state analysis*

*Correlation*

*Linear Modeling / MANCOVA*

*Time analyses, linear modeling, MANCOVA*

*Predictive analysis / conclusions*

*Visual modeling, heat map*

**Personnel requirements**

We are a team of 5. Our instructor is Dr. Saleh.

The team members will share different responsibilities every week and the Scrum Master will keep rotating weekly. We’ll use Slack to discuss the Daily Stand Ups.

By 12 pm EST, all teammates will Slack their Daily Stand-Up points.

Team meetings are scheduled every Saturday at 12 EST.

Team meeting with Dr. Saleh is scheduled for Mondays at 8pm EST.

Dependability (committing to deadlines) and Active Involvement (everyone contributes) will help in the Project’s success.

**Delivery schedule**

**Week 1:** *Project Planning* – Create group, come up with name, topic, establish group dynamics. Create Kanban & Set up GitHub. Set meeting times. Import dataset into R for quick overview and light data wrangling. Study the dataset and ask questions. Come up with evaluation questions & analysis. Educate ourselves on bees, honey production, agriculture, and commodities. Create project proposal. Finish Lesson 1.

**Week 2:** *Data Wrangling* – Finalize analyses with Instructor. Determine analysis specifications. Assign evaluation question. Clean up, recode, transform, and organize datasets. Combine datasets. Test for assumptions. Look at possible correlations. Is the data normally distributed? What are some predictive models we can make from it? Visualize the data to see if there is any interesting findings. Choose next week’s scrum master. Present work to team. Create week 2 summary document. Perform code review and submit all code used for data wrangling. Finish Lesson 2.

**Week 3:** *Exploratory Analyses* – Look at correlation matrices. Create visual graphics of data. Integrate data into Tableau to begin exploring visuals. Perform Modeling/Optimization in R and Machine Learning in Python. Ensure all data is ready for final analyses. Choose next week’s scrum master. Present work to team. Create week 3 summary document. Perform code review and submit all code used for analyses. Finish Lesson 3.

**Week 4:** *Final Analysis & Conclusions* – Perform final analyses. Go through the iterative process. Present findings to team. Review and validate findings from the previous week and draw insights/conclusions. Choose next week’s Scrum Master. Create Week 4 summary document. Perform code review and submit all code used for conclusions. Finish Lesson 4.

**Week 5:** *Data Visualization & Reporting –* Establish presentation structure. Compile findings into a Power Point slideshow. Integrate Tableau graphics. Refine presentation Aesthetics. Set the tone of the presentation. Go over ppt with Dr. Saleh. Choose next week’s Scrum master. Create Week 5 summary document. Perform code review and submit all code used for data visualization & reporting.

**Week 6:** *Presentation! –* Make final touches to the Power Point presentation. Practice presentation delivery several times. Work on public speaking tips and feedback shared within group. Utilize presentation checklist. Record presentation. Present!!! Create Week 6 summary document. Finish Final lesson. Celebrate!!

**Other requirements**

**Assumptions**

All statistical programming and BI software should be up to date for proposal use. All datasets seem viable and provide relevant information to the proposal questions provided. Each team member is prepared for their portion of the work on the proposal schedule and other roadblocks aside. Changes in the variables may impact the project deadline and quality.

**Limitations**

As long as all the statistical software, analysis, and other project tasks run according to schedule, there should be no delays in the final project. Limitations with the data may come into play, and there may be other variables that we have not foreseen as necessary. They may also be other variables that we realize are important to the final proposal questions. These unforeseen variables may result in delays.

**Risks**

There are no risks to motivation from the team, or assistance to the group from leadership and mentors. There is the possibility of setbacks in the timeline or unforeseen issues with the study, however; we should be right on track or ahead of schedule. The risks to the project are minimal and failure to complete is highly unlikely.

1. “Impacts on the Food Supply.” *Center for Food Safety*, Center for Food Safety, 2022, https://www.centerforfoodsafety.org/issues/304/pollinators-and-pesticides/impacts-on-the-food-supply. [↑](#footnote-ref-1)
2. Nowierski, Posted by Robert M., et al. “Pollinators at a Crossroads.” *USDA*, USDA, 29 July 2021, https://www.usda.gov/media/blog/2020/06/24/pollinators-crossroads. [↑](#footnote-ref-2)
3. “Impacts on the Food Supply” [↑](#footnote-ref-3)