1. Ask Questions
   1. The problem is that honeybee populations are in decline. So, I asked the guiding question: *Why are Honeybee populations in decline?*
      1. Based on this guiding question, I knew the factors I wanted to analyze:
         1. Negative impacts
         2. Price of honey
         3. Production of honey
         4. The rate of loss depending on the state
2. Collect Raw Data
   1. I researched and found the USDA honey commodity survey, which contains data for all 50 states. I downloaded it from the USDA website and
3. Clean + Prepare Data
   1. The first thing I noticed was that the table contained 33,359 rows and 9 columns. Based on this size, I determined that SQL would be the easiest way to work with the table, rather than using excel. So, I loaded the table into Azure Data Studio.
   2. Upon opening the import wizard, I noticed that the value column contained the records, which meant that was the column I needed to extract. Several records were written as decimals, so I set the column format to Float. This way I could still conduct mathematical operations on the values column. However, I ran into an issue- several values in this column had been entered as (Z), and others had been entered as (S). I changed the import mode to VARCHAR and imported the table. I then deleted all rows where value was (S) or (Z) using:

DELETE FROM ImportTable

WHERE Value IN (S), (Z)).

* 1. After this, I set the Value column back to Float. The other issue I ran into with importing was that the State\_ANSI column contained several nulls. This was not an issue, as the State\_ANSI column was not going to be critical, so I set the State\_ANSI column to allow Nulls. This enabled importing the table.
  2. With the table now imported and ready to be organized, I began planning my design. Based on the 9 columns, I recognized that there were 3 dimensions to the data- the state the data was recorded in(location), the time the data was recorded(time), and the type of measurement being recorded(measurement).
     1. The location table was the simplest to create, so I created it first. I planned on using State\_ANSI as the primary key, which wasn’t possible at that point since the column contained NULL values. I investigated why using the query:

SELECT State\_ANSI, State

FROM ImportTable

WHERE State\_ANSI IS NULL.

I found that there was two values for ‘State’ where the State\_ANSI column was NULL- US Total, and Other States. To account for this, I found the max State\_ANSI with

SELECT MAX(State\_ANSI)

FROM ImportTable,

and found that it was 56. Knowing this, I updated the State\_ANSI column-

UPDATE ImportTable

SET State\_ANSI= CASE

WHEN STATE= ‘OTHER STATES’ THEN 57 WHEN STATE= ‘US TOTAL’ THEN 58

ELSE State\_ANSI END

WHERE State\_ANSI IS NULL.

I verified that this would not conflict with actual state codes by ensuring the Census Bureau does not have anything assigned to 57 and 58 in its ANSI lookup. Although this did impact on the import table, this replaced nulls, so it wasn’t a loss of data. After doing this, I created the dimension table with:

SELECT State\_ANSI, State, Geo\_Level

INTO StateLookup FROM

ImportTable.

I then set the State\_ANSI column as the primary key using Azure Data Studio’s table editor. This way, I had State\_ANSI as the primary key and could use the State column to filter for specific states, and the Geo\_Level (which was either STATE or NATIONAL) to filter out national totals, or to only get national totals.

* + 1. Next, I created the Timetable. There were two columns from which I intended to derive the Timetable- Year, and Period (which noted the timespan of the data). First, I deduplicated with the query:

SELECT Year, Period

FROM ImportTable

GROUP BY Year, Period

and validated that it worked by adding ORDER BY Year DESC, Period. I knew from my planning that I would want to record the change in hive amounts over time, which would be enhanced by seeing changes within the year. Based on this, I added a quarter column with the query:

SELECT

Period,

CASE

WHEN Period IN('JAN','FEB','MAR') OR PERIOD='JAN THRU MAR' OR PERIOD='FIRST OF JAN' THEN 1

WHEN Period IN ('APR','MAY','JUN') OR PERIOD = 'APR THRU JUN' OR PERIOD ='FIRST OF APR' THEN 2

WHEN Period IN ('JUL','AUG','SEP') OR Period='JUL THRU SEP' OR Period='FIRST OF JUL' THEN 3

WHEN Period IN (OCT, NOV, DEC) OR Period='OCT THRU DEC' OR Period='FIRST OF OCT' THEN 4

ELSE 0

END AS Quarter.

Next, I wanted to measure what type of timeframe was being measured, in order to help with creating the primary key. I took note of which values existed for Period with the query:

SELECT DISTINCT Period

FROM ImportTable

Next, I created the type column with

CASE

WHEN Period LIKE '%THRU%' THEN 'T'

WHEN Period LIKE '%FIRST OF%' THEN 'F'

WHEN LEN(Period)=3 THEN 'M'

WHEN Period='MARKETING YEAR' THEN 'MY'

WHEN Period='MID DEC' THEN 'YE'

ELSE 'NA'

END AS Type.

This case statement covered every existing value for Period but included an ELSE statement just to be sure. I then wanted to increase cardinality in the type column, as I was going to use it to produce a primary key. I used

CASE

WHEN PERIOD='JAN' THEN 'M1'

WHEN PERIOD='FEB' THEN 'M2'

WHEN PERIOD='MAR' THEN 'M3'

WHEN PERIOD='APR' THEN 'M4'

WHEN PERIOD='MAY' THEN 'M5'

WHEN PERIOD='JUN' THEN 'M6'

WHEN PERIOD='JUL' THEN 'M7'

WHEN PERIOD='AUG' THEN 'M8'

WHEN PERIOD='SEP' THEN 'M9'

WHEN PERIOD='OCT' THEN 'M10'

WHEN PERIOD='NOV' THEN 'M11'

WHEN PERIOD='DEC' THEN 'M12'

ELSE Period

END AS Class.

I then created the primary key with:

CASE

WHEN TYPE='M' THEN Class+RIGHT(Year,2)

WHEN TYPE IN('Q','F’) THEN Type+CAST (Quarter AS varchar)+RIGHT(Year,2)

ELSE Type+RIGHT(Year,2)

END AS Timestamp

I then selected timestamp, year, period, and quarter, using timestamp as the primary key. I then noticed that certain measures were taken on the first day of a quarter- meaning they were actually recording the prior quarter. To reflect this, I created two columns- reflective year and reflective quarter. To create this, I used the query

UPDATE Time

SET ReflectiveYear=

CASE

WHEN LEFT(Timestamp)='F' THEN Year-1

ELSE Year

END,

ReflectiveQuarter=

CASE

WHEN Quarter=1 THEN 4

ELSE Quarter-1

END

WHERE LEFT(Timestamp)='F';

UPDATE Time

SET ReflectiveQuarter=Quarter

WHERE Quarter IS NULL;

* + 1. The first problem I encountered is that the data is recorded in a way that makes it difficult to filter. For example, a record in the original table reads “HONEY, BEE COLONIES, AFFECTED BY PESTICIDES - INVENTORY, MEASURED IN PCT OF COLONIES.” Read on its own in a table it works well- it’s clear what is being measured and how it’s being measured. However, for my analysis, where I’m trying to compare several things, this would not be suitable. For example, I knew from my preparation phase that I’d want to compare different types of negative impact in PowerBI, which would involve the Legend bucket. This would be complicated with the table still normalized like this, as it would require filtering out all but those two values, rather than filtering for one value. So, in order to enable the analysis, I plotted out a STAR diagram with three dimensions: time, measures, and state.

To create the Measures table, I decided to split each data\_item entry into three components: category, subcategory, and measure. This design was optimized for PowerBI’s structure of filtering and buckets. First, I used SQL to select the distinct values of Data\_Item, adding an index with the RANK() function. In order to this, I used a subquery to select the distinct values, adding the index in the main function. This became the MeasureCode column, which functioned as the primary key. From there, I split each data\_item record into its components. To use the example from above, HONEY, BEE COLONIES, AFFECTED BY PESTICIDES - INVENTORY, MEASURED IN PCT OF COLONIES was split into three component parts-

1. Category(the general type): Colonies: Negative Impact
2. Subcategory(the specific type, meant to be used with the Legends bucket): Pesticides
3. Measure(the method of measurement): PCT OF COLONIES

This was done using CASE statements, enabling the entire column to be created in one go.

Similarly, HONEY, BEE COLONIES, AFFECTED BY VARROA MITES - INVENTORY, MEASURED IN PCT OF COLONIES became:

1. Category: Colonies: Negative Impact
2. Subcategory: Varroa Mites
3. Measure: PCT OF COLONIES
   * 1. Example:

Here’s an example of how that process plays out in the actual table. The values of this row(row 100) are:

Program:SURVEY # All rows are of the value survey- it’s not of much use

Year: 2024

Period: APR THRU JUN

Week\_Ending: *NULL* # every value of this column is null

Geo\_Level: STATE

STATE: IDAHO

State\_ANSI: 16

Watershed\_code:0 #All values of this column are 0

Commodity: HONEY # All values are HONEY

Data\_Item: HONEY, BEE COLONIES - LOSS, COLONY COLLAPSE DISORDER, MEASURED IN COLONIES

Value: 4500

State\_ANSI, STATE, and Geo\_Level are added into the StateLookup table as:

State\_ANSI: 16

State: IDAHO

Geo\_Level: STATE

Year and Period are added into the Timetable as:

Timestamp: Q224

Year: 2024

Period: APR THRU JUN

Quarter: 2

ReflectiveYear: 2024

ReflectiveQuarter: 2

Data\_Item goes into the Measures table as:

MeasureCode: 23

Category: Colonies: Lost

Subcategory: Colony Collapse Disorder

Data\_Item: MEASURED IN COLONIES

And finally, a record is created in the Reports table as:

ID: 16Q22423

Timestamp: Q224

State\_ANSI: 16

MeasureCode: 23

Value: 4500

1. Analyze Data

I intended to use the 5 Whys to address my question.

Why #1: *Why are honeybee populations in decline?*

I used the Categories column to filter out any unrelated values and sorted them by Subcategory to analyze each individual negative factor. What I found was that the impact of these negative factors is actually in decline, with the total percentage of colonies declining over time also going down. This meant the root cause was not these factors. So, I asked the next question-

Why #2: *Why are honeybee populations declining despite negative impacts also declining?*

I looked into other factors. First, I looked into the price of honey. This could be a factor, as many hives exist as part of commercial honey production. A decline in the honey industry would likely result in a decline in hives. However, the price of honey has risen at this time, likely a consequence of the decline in honeybee populations. Then I looked into the rate of colonies being added over time and found my answer- since at least 2020(2019 is an outlier due to quarter 2 being unreported), the annual amount of colonies added has gone down.

*Why #3: Why is the rate of Honeybee colony additions going down?*

I broke it into quarters, expecting to find that the majority of additions occurred in quarters 1 and 2, as that lines up with the time of the year bees exit hibernation and begin producing drones (due to a lack of drones after Spring, hive reproduction would slow down dramatically). I found this to be true. For instance, 75% of colony additions in 2022 occurred during Quarters 1 and 2. However, I discovered that the colony additions during these quarters is steadily decreasing. Which led to my next question:  
  
*Why #4: Why are colony additions down in Quarters 1 and 2?*

I analyzed honey production measured in lb per colony, which revealed that honey production is down across the board. This means the bees are doing less work per colony- indicative of a greater trend. This would also explain why colony additions are slowing down- bees are producing less, their hives don’t grow big enough, which means their hives can’t be split, which means you can’t create more colonies. Which led to my final question:

*Why #5: Why are bees producing less?*

From earlier analysis, I knew it wasn’t due to measurable negative factors. So, I conducted outside research. I found this article(link) that discusses the impact of climate change on bees. This makes sense- bees are highly sensitive to temperature, and as the heat goes up, their production will slow down. So, the reason honeybee colony inventory is declining is because new honeybee colonies are not being added fast enough to offset the continual loss of colonies, and the reason this addition is slowing down is due to climate change. This is proven true based on a national map showing where colony losses are highest- New Mexico, Arizona, and the South- areas where spring heat is highest.

1. Visualize

To demonstrate losses, I used line charts. For the first chart, I set the X axis to reflective year and reflective quarter, sorting the axis based on the reflective year and reflective quarter, ascending. I used the Value column as the Y axis, and the Category column as the legend. In order to get the correct data, I set visual-level filters to only show the Added/Replaced or Lost categories, state totals, and years between 2015 and 2023. This way, the addition and loss of colonies could be traced quarter by quarter. I added an additional line graph showing the change in percentage of colonies lost. I set the visual-level filters to category- Colonies: Lost, Geo\_Level: STATE, Year: not 2024, Data\_Item: MEASURED IN PCT OF COLONIES. Thus, I got the specific data I desired. I set the Value column as the Y axis, and set that to be average, to ensure it showed the average percentage, which was the goal. I also added a pie chart to demonstrate the type of loss colonies usually endure- the objective being to demonstrate why colony collapse disorder (accounting for 13.5% of losses) is generally overlooked. These visuals prove the causal root of the decline (bee colony additions declining) while also proving the cause is external (since the line graph shows the percentage of deadout going down).

A screenshot of a graph

Description automatically generated

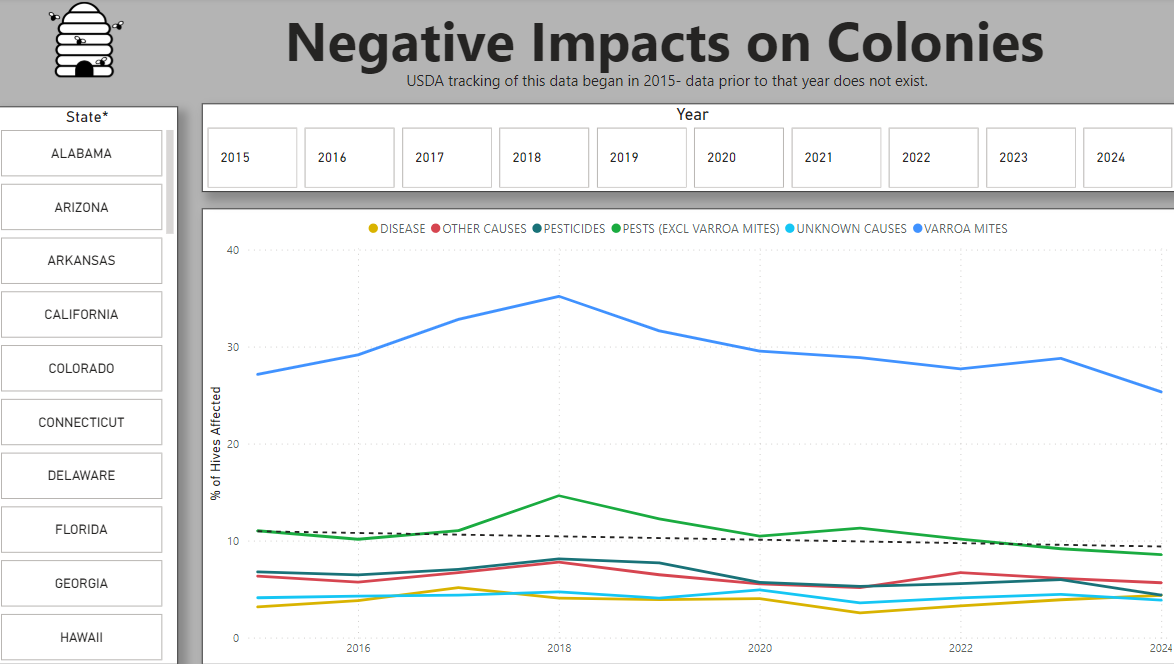
The decline map uses the filled map visual, with the State column in the Location bucket and the average percentage loss per state in the Tooltip bucket. The shading is a dynamic scale using the Value average to determine the shading.

This map shows which states/regions are experiencing the most loss, and which states have the least loss. For instance, you can see that Arizona has the highest loss, whereas Vermont is the lowest. Certain states (like Nevada) are not colored due to no data being reported.

A map of the united states

Description automatically generated

The Negative Impacts page uses a line graph to show the various negative impacts on bees and their trend over time. As it shows, Varroa Mites are the primary negative impact on bees, but is trending down. Nearly all other causes are trending down as well, with disease trending up.



This is the model that I built:

A screenshot of a computer

Description automatically generated

Recommendations:   
  
Based on the research, I recommend public outreach to strike up support for measures that aim to combat climate change, based on the fact that climate change is the root cause of the decline of honeybee populations. I also recommend funding research to help breed bees that are more resilient to heat, thus making them more capable of producing as the temperature increases.