**Tania Nguyen - Database**

**As Peyton mention, we will start with our database.**

**I will be explaining our process of how we created our database which includes-**

**Brainstorming using ERD**

**Database using AWS**

**Creating the tables - using PostgresSQL**

**Connection btw tables using an Inner Join**

**Before creating our database and tables we had to plan and brainstorm using ERD**

ERD: Entity–relationship model

We decided to break the data into 3 tables

We decided to connect PLU table to Avocado table by avg price and Avocado table to Size table by type.

I will be defining each column that is presented in each table

Under the PLU table: PLU - Price look-up codes

* Average Price - the average price of a single avocado
* The 3 different PLU codes- total number of avocados with the respective PLU code sold
* Geography of where the product is sold

Under Avocado table:

* The type (conventional vs organic)
* **Average price**
* Date - The date of the observation
* The year it was sold
* Total Volume - Total number of avocados sold

Under the Size table:

* **The type** (conventional vs organic)
* The 3 different sized bags of Avocado
* Total bags sold

We created this schema to answer our main question Where are the best locations for testing a new product for the best avocado prices?

AS mentioned We decided to connect PLU table to Avocado table by avg price and Avocado table to Size table by type. This would give us the best idea of where avocados would be purchased at a higher rate in the US. Taking this result would indicate where we should test our new product in.

**The process and how I got the data into the table:**

1. Took the clean data which was done by Pete, to see what columns were in the CSV.
2. Using AWS - (Amazon Web Services)created a database and S3 buckets to contain the different CSVs using Python to feed into the 3 different tables that will be created.
3. Using Postgres - used SQL code to create the tables based from our ERD.
4. To double-check whether the tables were created correctly, Select \* from (table) to see the table that was created

**Database Join:**

**Next slide:** This image shows the inner join on average price created between PLU table and Avocado table

This shows the total volume of the avocados sold within that location.

Now to see the connection btw database to MCL. I’m going to turn it over to Pete, where he will be explaining the machine learning portion.

Some relevant columns in the dataset:

13 columns, 30021 rows

Date - The date of the observation

Average Price - the average price of a single avocado

type - conventional or organic

year - the year

geography - the city or region of the observation

Total Volume - Total number of avocados sold per month

4046 - Total number of avocados with PLU 4046 sold

4225 - Total number of avocados with PLU 4225 sold

4770 - Total number of avocados with PLU 4770 sold