**Presentation Outline of Project**

* **Selected topic**

To save estimators time, we are using APIs to search material prices of multiple electrical materials suppliers. With the data collected, we produce two tables. One that is updated daily with the lowest prices and another that is also updated daily and consolidated. The consolidated data set will be used to establish price correlations between materials and make predictions.

* **Reason topic was selected**

I have been an electrical estimator for five years. In the beginning, when we were estimating jobs, we would use a Seasonally generated list of materials with prices. soon after the pandemic began, my boss instructed me to check material prices daily before submitting bids. Estimates that took 3 hours, now took six hours.

* **Description of the source of data**

We've chosen 16 commonly used electrical materials items to query. After some research, we determined that the Home Depot product API does the best job of grabbing electrical material items we tell it to look for by product id. We then Plug the 16 items into a Google search algorithm and extract the material name, the price, and the website link. Unfortunately, the Google search API does not have a vendor category so to extract the name of the vendor, we use Python to split the link. Our main source of data is generated daily and can contain up to twenty-seven vendors. We also have historical data but the best we could find only covers two vendors (Home Depot and Lowes) over a 120 day period.

* **Questions the team hopes to answer with the data**

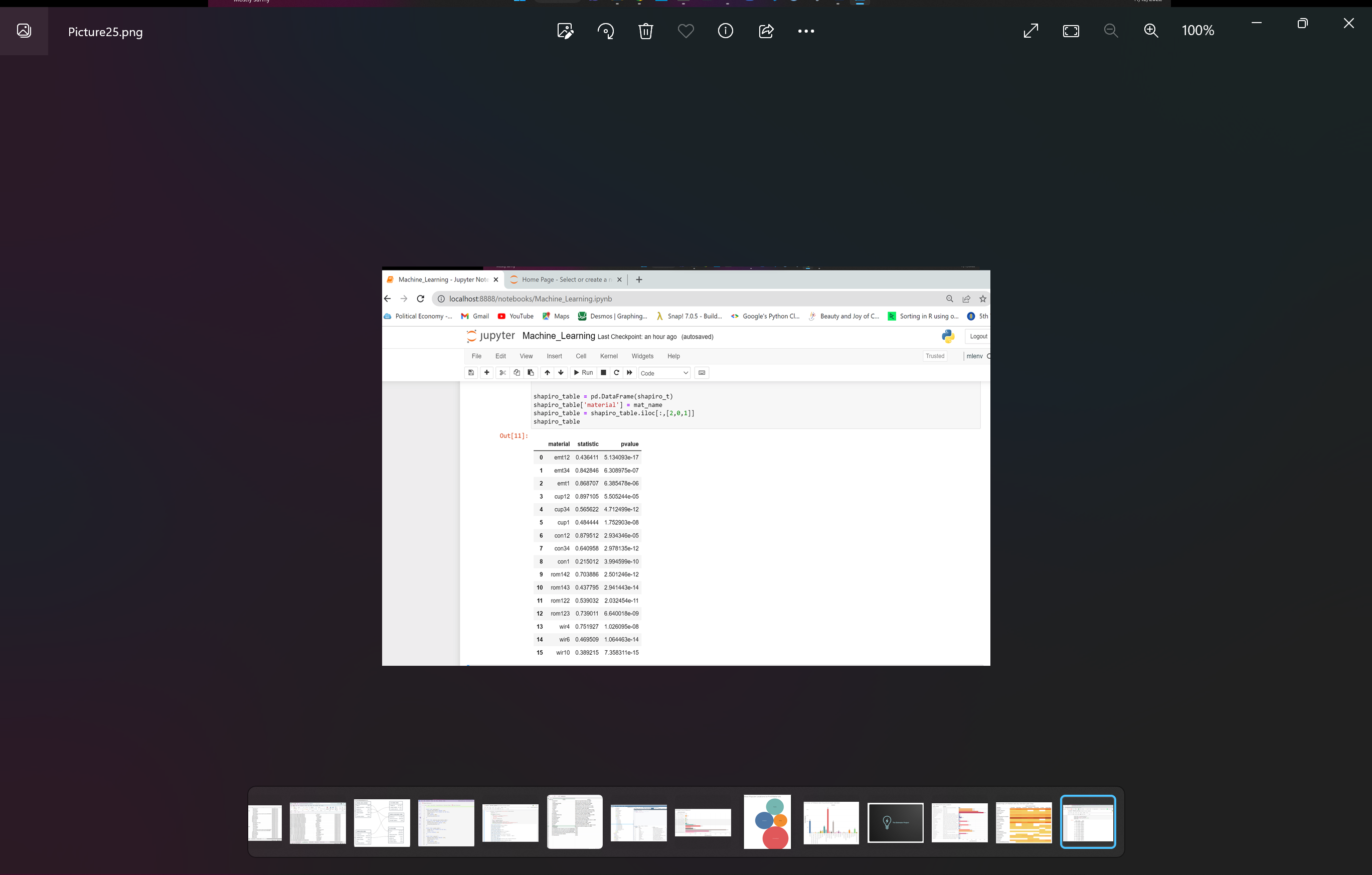
Small items, typically purchased with large materials, are not directly priced. Rather, a percentage is tagged to the cost of the primary material to account for the cost of the small items. It’s impractical to search out each individual small item and yet, the margin on electrical construction is so slim that even the price of small items can make a difference on the profit. The materials we are looking at can be placed in four categories: conduit, conduit fittings, cable and the wires that go inside the cables.

**Questions:**

* Is there a correlation between the cost of conduit and fittings?
* Is there a correlation between the cost of cable and the wires they need?
* Is there a correlation between the cost of materials and the days of the week—in other words, is there a day of the week where materials are more expensive or less expensive And if so can we predict it?
* Is there a relationship between a vendor’s location or region and how much they charge?
* Is there a relationship between the size of the company and the discount it charges?
* **Description of the data exploration phase of the project**

The API approach was our alternative to web scraping, but the data collected still requires a lot of cleaning. There are scaling issues, and naming issues (different manufacturers have slightly different names for the same items). Once we have created vendor columns, we then plug those vendor names into a slightly different Google shopping API and that returns addresses. From the address column, we create a region column. There is also a formatted datetime column and a synthetic column based on the datetime that gives us the day of the week. The system is not perfect, but it works well. Since October 20th, we have been collecting and consolidating data. We currently have over 1000 rows of cleaned data with several features (prices, location, region, day of the week) and targets (high prices, low prices, and about average prices).

* **Description of the analysis phase of the project**

1. I ran the data through a Shapiro-Wilk test to determine if our material prices are normally distributed. As you can see from the table below, all p-values are less than .05 which means that our material prices data is not normally distributed (see “Picture25.png”). 
2. I checked to see if there was a correlation between:
   1. Conduit and coupling (fitting)
   2. Conduit and connector (fitting)
   3. #4 wire and #6 wire
   4. #6 wire and #10 wire
   5. #10 wire and # 4 wire

Results: some of the tests came close (#10 wire and # 4 wire and 1” conduit to 1” fittings) but none of the items I tested were greater than the 0.05 p-value needed to establish correlation (see “Picture26.png”). Graphical user interface, text

Description automatically generated In hindsight, there is likely a correlation between the price of finished materials and the raw materials needed to make them (copper to wire and steel to conduit). I’m planning to add the following features:

* Price of copper per pound (10/20-11/19)
* Price of steel per pound (10/20-11/19)
* Weight of materials item being queried
* Size of vendor (is it a big or small company?)

1. Next I kept the price column and label encoded: material description, vendor, date, day of week, region, and target price to create a new data frame. I ran a correlation matrix and saw some connections (see “Picture27.png”). Chart

   Description automatically generated
2. With the data encoded, I dropped “about average” to define a target of either “low price” or “high price”. With this done, I was able to separate my data into features (X) and target (y).
3. Finally, I split the data: test train split (Training data = 75% and Test data = 25%), scaled the data, created a **Decision Tree** model, fit the model and made predictions: Based on material name, price, vendor, date, day of week and region could I predict high or low prices (see “Picture28.png”)

Graphical user interface, text, application

Description automatically generated

1. Last, I created a confusion matrix and printed my results (“Picture29.png”).

Text

Description automatically generated