

A screenshot of a computer

Description automatically generated

Graphical user interface, table

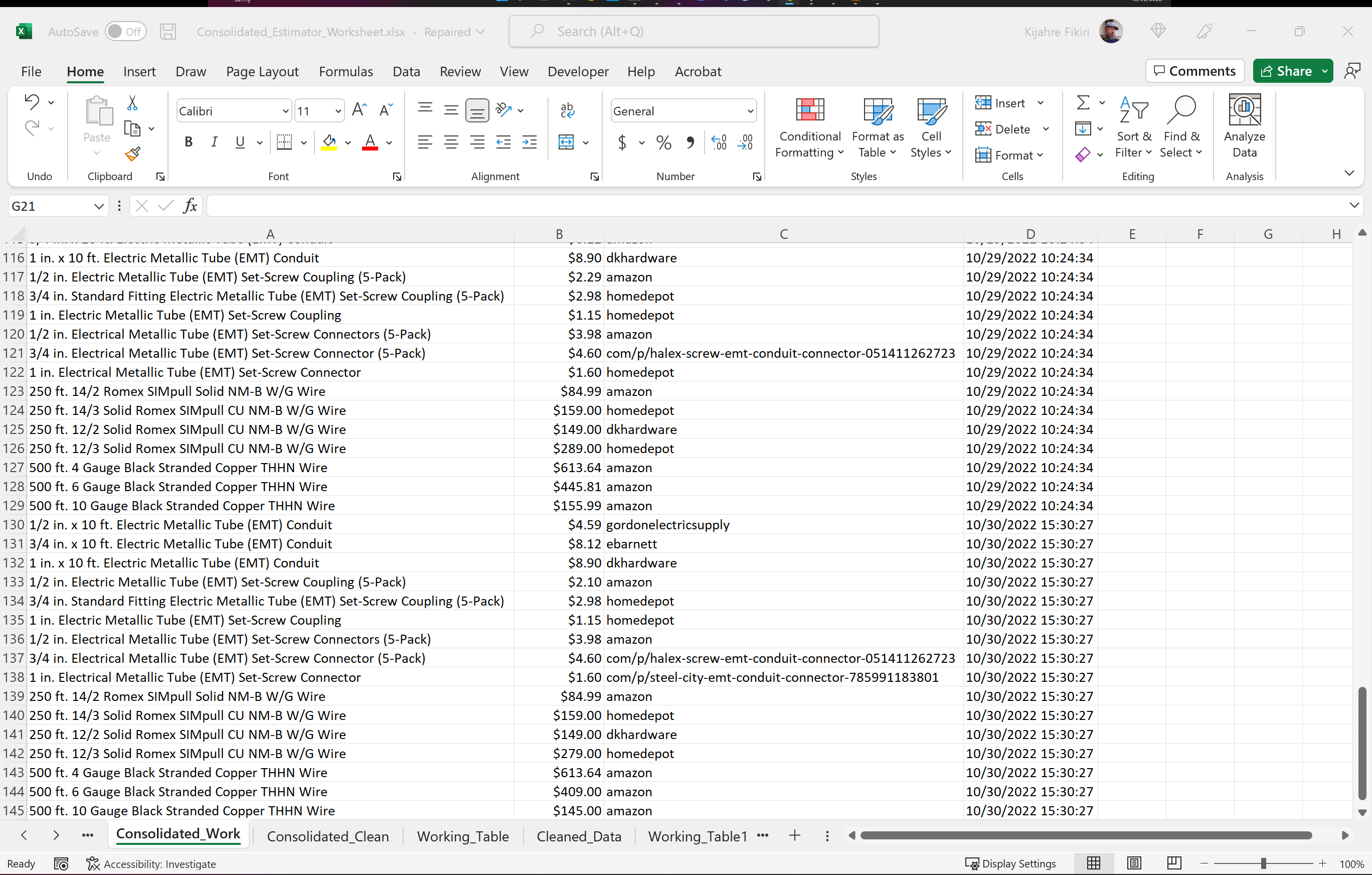
Description automatically generated with medium confidence

Graphical user interface, table

Description automatically generated

* I found a website that will run our script on a timed basis and hosts the results: deepnote link: <https://deepnote.com/workspace/estimatorproject-6166945d-6ec6-41f0-bb95-87f4a60e4841/project/WorkingAutomatedEstimator-59dc3a3e-9636-43f0-9372-adacb8d143f7/notebook/Working_Estimator_Script-9386e927ee03453397c483eb80ef1906>
* I recreated a “Normalized” set of schemas for the data:
* I sent data to pgAdmin and created databases.
* I consolidated (9) days of data and started creating a model that will establish:
  + a correlation between cable and wire
  + a correlation between conduit and fittings
  + Predict which vendor will offer the cheapest priced item
  + Predict which day of the week will offer the cheapest priced item
* I cannot export from deepnote at the moment. I know it will allow because I’ve done it once before. I suspect it won’t because I am on the free tier.
* I cannot export PGAdmin db’s so that they re-write existing db in-place
* We need historical data to create a predictive model and I don’t have it yet.

Content  
The presentation tells a cohesive story about their project, including the following:  
1. Selected topic  
Nancy  
2. Reason why they selected their topic  
Nancy3. Questions they hope to answer with the data  
Correlations: Nancy1MIN4. Description of their source of data: Kijahre  
explain obstacles we came across and why we chose to go with data set we accumulate5. Description of the data exploration phase of the project: Kijahre  
why are we choosing the features we chose  
location and vendor were not straight forward  
dissect and digest the data



Graphical user interface, application, table, Excel

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

ALTER TABLE "target\_table22"

ADD FOREIGN KEY (product\_id) REFERENCES material\_description\_table(product\_id);

SELECT \* FROM lowest\_price\_table1 limit 10;

ALTER TABLE "all\_vendor\_table1"

ADD PRIMARY KEY(vendor\_id);

SELECT \* FROM all\_vendor\_table1 limit 10;

ALTER TABLE "all\_link\_table1"

ADD PRIMARY KEY(link\_id);

SELECT \* FROM all\_link\_table1 limit 10;

**The presentation tells a cohesive story about their project, including the following:**  
Nancy: 1. Selected topic, 2. Reason why they selected their topic, 3. Questions we hope to answer with the data

Kijahre:

4. Correlations

5. Description of the source of data

* Description:
  + Since October 20th, our algorithm has been collecting data daily using a combination of Serp APIs (Home Depot product API, Google search API, a free stock exchange API, Google product search API).
  + Our ‘Estimator\_Script’ successfully returns the lowest priced items of our query and we store the remaining data for analysis (to date we have over 1250 rows of data).
* Explain obstacles we came across
  + In general, I had a lot of coding obstacles due to inexperience.
  + Comparatively, returning the lowest prices was the easiest part. But to analyze the data in a machine learning model, I had to clean the data and extract features that could be useful.
  + Initially our extracted data did not have vendor name, vendor location/region, raw materials, or date of query.
* Why we chose to go with this method of data extraction
  + We opted for using APIs because some of the material’s sites would kick me off when I would try to web scrape.
  + Also, web scraping can be tricky because different sites update often which we'll throw off things you have pre-programmed your algorithm to search and grab.
* Description of the data exploration phase of the project
  + The raw data only showed material name, price, and link.
  + To get vendor name, I split the link between the dots.
  + To get vendor locations, I fed link-created vendor names into a Google products search API.
  + Once I had addresses, I assigned regions based on zip codes.
  + Finally, I parsed the materials price column into: “about average”, “high price” and “low price”. “high price” is >=Q3, “low price” is <=Q1 and “about average” is everything else.
* Why we chose these features
  + Honestly, we chose these features because price and material name alone didn't seem sufficient to make a machine learning model.
  + Our theory was that material prices (in part at least) are somewhat governed by where the manufacturer is.
  + Looking at the queries and creating simple histograms, it seemed that prices fluctuated on different days so I made a day of the week feature.
  + I realized that half of the items we are querying are made of copper and the other half are made of steel so I made a raw materials feature. I then queried daily stock prices for copper and steel and added that as well.

6. Description of the analysis phase of the project (ML)

* data validation
  + I used a Shapiro-Wilk test to determine normality
  + I used a Pearson test to check various correlations
  + From there is was plug and play: test, train, split, scale and make different models from different features with my target being high/low prices
* correlation matrices

Sarah: 7. Result of analysis: confusion matrix / discuss specifics in data features

Notes for Sarah:

* With the results from the Pearson test we were able to answer the following questions:

1. Is there a correlation between the cost of conduit and fittings? **YES**
2. Is there a correlation between the cost of cable and the wires they require? **YES**

* We used a decision tree logistic regression model to determine:

1. Could we use a machine learning model to predict cheap prices? **ABSOLUTELY (98% accuracy and 98% precision)** **Graphical user interface, text, application

   Description automatically generated**(‘Picture31.png’ in my images folder)
2. Is there a relationship between the cost of materials and the days of the week? **SOMEWHAT BUT** **NOT A STRONG ONE (65% accuracy and 65% precision).** As it turns out, the number of returned queries increase around weekends due to increases in weekend advertising for the materials we queried. Graphical user interface, text, application

   Description automatically generated (‘Picture32.png’ in my images folder)
3. Is there a relationship between a vendor’s location or region and how much they charge? **YES**/**SOMEWHAT (73% accuracy and 69% precision)** **Graphical user interface, text, application

   Description automatically generated** (‘Picture33.png’ in my images folder)

Alexei: 8. Recommendation for future analysis/ Anything the team would have done differently: Expand on website…etc, 9. Technologies, languages, tools, and algorithms used throughout the project (short discussion)

Nancy: 10. The team demonstrates interactivity of dashboard in real time  
Notes for Alexei:

* In the future we’d add the following features:
  1. Weight of materials item being queried
  2. Size of vendor (is it a big or small company?)
* In the future we’d add try the following:
  1. A Naïve-Bayes model to account for prior probability
  2. A Regression model to predict which day of the week or which vendor is best to shop
  3. We used Python to query API’s but we might have gotten more specific results had we used JSON