**Dependencies Used:**

from serpapi import GoogleSearch

import pandas as pd

import numpy as np

from datetime import datetime

from openpyxl import load\_workbook

from sqlalchemy import create\_engine

import math

import matplotlib.pyplot as plt

import seaborn as sn

from scipy.stats import shapiro

from scipy.stats import lognorm

from scipy.stats.stats import pearsonr

from sklearn import tree

from sklearn.preprocessing import LabelEncoder

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import confusion\_matrix, accuracy\_score, classification\_report

from sklearn.datasets import load\_iris

from sklearn.naive\_bayes import GaussianNB

**Slide Presentation Outline**

5. Description source of data

* Data:
  + Collecting data daily since October 20th using a combination of APIs.
  + The process:
    - Plug in a list of items
    - Standardize the names of the items
    - Filter out scaling issues
    - Create (2) data sets:
      * lowest prices
      * all prices
  + ”All prices” produces about 50 entries a day and to date we have over 1250 rows of data (returning the lowest prices was the easiest part).
* Obstacles:
  + Feature definition
  + Coding Inexperience
  + Data cleaning (is a beast!)
  + Tutoring support
    - Every tutor is not proficient in everything.
    - Tutors are human too
      * *Within a week, I lost 2 tutors I spent time building a repour with and was told by another that what I wanted to do “couldn’t be done…so I did it!”*
  + Not every manufacturer has an API (let alone a free tier).
* So why API over Web Scaping
  + Some sites kick you off when you try to web scrape!
  + An HTML update can throw off your pre-programmed search and grab.
* Feature Selection
  + vendor: split the link between the dots.
  + location: fed an f-string into a Google products search API:
    - f”where is <’vendor\_name’> address?”
  + zip: split from location.
  + region: assigned based on zip codes.
  + raw\_material: assigned based on our materials
  + raw\_prices: retrieved from daily stock prices of copper and steel
  + target\_price:
    - “high price” is >=Q3
    - “low price” is <=Q1
    - “about average” is everything else.
* Why These Features
  + Theory:
    - Material prices are somewhat governed by location of manufacturer.
    - Day of week could be a pricing factor.
    - Finished product prices depend on raw material prices

6. Analysis phase of the project (ML)

* data validation
  + Shapiro-Wilk test to determine normality
  + Pearson test to check various correlations
  + correlation matrices
  + Feature selection
  + Split, test, train, scale
  + Make models:
    - Decision tree to predict lowest price by all features
    - Decision tree to predict lowest price by dow
    - Decision tree to predict lowest price by region
    - Naïve-Bayes to predict lowest price by all features

1. **ML: explanation of changes between segment 2**

In addition to the models we already had, we ran a Naïve-Bayes to account for prior probability.

1. **Description of how we trained the model and additional training**

We imported ‘train\_test\_split’ from sklearn.model\_selection and split our data 75% training data and 25% test data.

1. **Description of current accuracy and address our 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**
   3. Can we use a machine learning model to predict cheap prices? **ABSOLUTELY (99% accuracy and 98% precision)** but may be overfit (Picture35.png)
   4. Is there a relationship between the cost of materials and the days of the week? **YES, BUT** **NOT AS STRONG (67% accuracy and 71% precision).** We noticed that accuracy and precision has been increasing as we increase our data.Also, the number of returned queries increases around weekends due to increased weekend advertising.

queried. (Picture35.png)

* 1. Is there a relationship between a vendor’s location or region and how much they charge? **YES, BUT** **NOT AS STRONG (69% accuracy and 75% precision).** We noticed that precision has been increasing as we increase our data. (Picture37.png)

Slide Notes

Slide 5: “**data will still need a lot of cleaning”**

Slide 6: “**To date we have over 1200 rows of data”**

Slide 7: “**A consolidated table is used for further analysis”**

Slide 8: “**Our target is lowest prices”**

Slide 9: “**As you can see by the black squares, vendor and region have a strong correlation”**

Slide 12: “**The same job takes an experienced estimator 3 hours to do”**