**Instructions**

You may use any tools and language, however, it is highly suggested that you use Jupyter notebook as it is a primary tool that we use on the team. Our team also is more Python-focused versus R or Julia. Please show/explain your work.

**Background**

You work at a magic potato trading company. The company earns fees by storing potatoes for clients as well as finding buyers and sellers of potatoes on behalf of clients. The storage fees earned by the company are based off the total market value of the potatoes stored by the clients. These potatoes are magic- they do not spoil while in storage, they can be shrunk so they take up very little space, and the delivery of the potatoes does not incur any cost. Potatoes must be kept separate from each other and therefore the cost to store different potato types may differ. Storage rates will be determined by traders and based off of numerous factors.

You work on the data science team and your role is to work with data to create reports, models, and make the business smarter and more efficient

**Two parts**

Analyse the given sets of data and answer the questions below. Help the management team understand what is happening in the potato market and what client activity looks like.

Create a web tool that helps capture and save potato prices seen in the market

**Part 1**

Background

You are given the *Data\_Files* excel sheet which contains information about potatoes, the company's clients, and a snapshot of client potato positions for a period of time. A quantity of Null represents no position. Management has some questions regarding these data and would like to know your interpretations. You will present your findings at the quarterly management meeting

Questions

* **Which clients have the largest potato stockpile based on market value?**

Based on market value: Client ID=27, 38, 82

* **Which clients are the most active?**

Based on quantity: Client ID=27, 38, 82

Based on count of distinct Product ID: Client ID=38, 50

Based on count of days: Client 17, 27, 30, 38, 44, 49, 50, 52, 64, 74, 82, 89, 94

* **Which potatoes are most activity traded?**

Based on sum of distinct Client ID: Product ID=16

Based on quantity: Product 73, 93

* **What client activity trends do you see?**
  + Indicators1(per client): count of distinct potato type, stockpile, quantity, count of distinct clients   
    Indicators2(overall trend): average of indicators1
  + **The count of distinct potato type per client per day**   
    Indicators1: There are four types of client activity trends   
    -Type1: basically steady and highly active (Client 38, 74)   
    -Type2: unsteady and lowly active (Client 44,27)   
    -Type3: Extremely high volatility(Client 50)   
    -Type4: basically steady and lowly active (Most Clients)   
    Indicators2: May-Jun(Down); Jun-Jul(Up); Jul-Aug(Unsteady); Aug-Sep(Down)   
    Summary: apart from extreme cases, most clients are less active after August and quantity value of many fluctuations are not recorded
  + **The total stockpile**   
    Indicators1:There are four types of client activity trends   
    -Type1: extreme large stockpile(Client 38, 27, 82)   
    -Type2: steady stockpile with similar trend (Client 74, 30, 89, 64)   
    -Type3: unsteady stockpile (Client 77)   
    -Type4: steady stockpile without similar trend (rest of clients)   
    Indicators2:There is a drop of average stockpile at the start/end of month.   
    May-Jun(Down); Jun-Aug(Up);Aug-Sep(Down-Up)
  + **The sum of quantity**: quite similar with sum of stockpile
  + **The count of distinct client**: more clients join in mid-Aug

|  |  |  |
| --- | --- | --- |
|  | Quantity | Sum of distinct Product ID |
| Focus Each Client | Some clients increased large storage in a short time (ID:38, 82, 27) | Client 38 storage a diverse product;  Client 50 quickly increases the type of potato in a short time;  May-Jul(Up); Jul-Aug(Keep);Aug-Sep(Valley) |

* **What potato price trends do you see?**  
  -Basically the trend of price are steady for most potatoes, but some always have high price(Product ID 11,236,227), rest of them are of relatively low price.

-Majority prices show a "valley" around Aug 24th-25th

-The average price of potatoes per day May-Jun(Up); Jun-Aug(Keep); Aug-Sep(Valley)

* **Are there any factors that can help predict potato prices?**

Based on previous analysis of potato prices, I think there are two main factors that can help predict potato prices:

* **Potato fundamental information** (potato\_info): those factors can be used to describe the quality of product, which decide the inner value of potato. I think we can use some regression models to find significant factors.
  + **Market trend**: Although most potato show low volatility cross the time, most potato still show a common “valley” trend in August, which shows the price can be influenced by factors out of those fundamental potato information to influence the overall market. I think it is possible to build Time Series model to predict market average potato price.
* **Are there any factors that can help predict client activity?**

I evaluate the client activity from metrics like stockpile, quantity, type of potato and count of distinct potato.

* + **Price:** the price of potato can influence the storage quantity of certain potato
  + **Product ID:** the type of potato can influence the quantity
* **Summarize what has happened during this period**
  + May-Jun: Clients’ activity was steady with low volatility, and stockpile was decreasing as coming to the end of May; potato price increased a little bit
  + Jun-Jul: Clients became more active in trading different types of potato, and stockpile was increasing and some clients left; More potato types joined and the price continue to grow
  + Jul-Aug: Clients became less active and stockpile continue to increase; the potato price is steady and active potato was decreasing
  + Aug-Sep: Clients became less active and decreased the number of potato types, stockpile and the active potato types decreased significantly firstly and increased later may due to the change of price, at the same time more clients join to store potato; Most all potato price got a sudden decrease at the mid-Aug.
  + Interesting take away:
  + There is usually a decrease at the end/start of a month

|  |  |  |
| --- | --- | --- |
| Client ID | Quantity | Sum of potato type |
| 38 | June: increase storage of Product 242  July: increase storage of Product 146 | May-July: steady up trend  August: down-up trend |
| 27 | June: increase storage of Product 251  August: increase storage of Product 281 when other clients tried to sell stored potato | Steady |
| 50 | Steadily cycling around 15k | June & July: sudden increase |

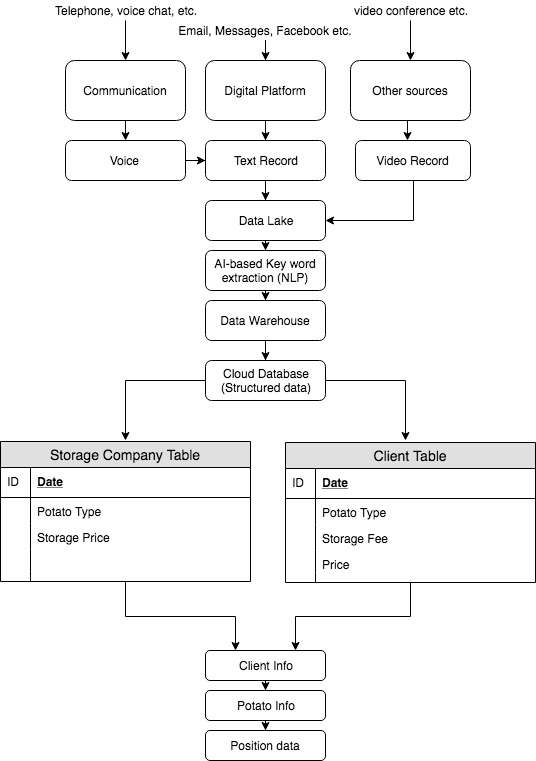
**Part 2**

Background

Sales people and traders often receive storage prices offered by other storage companies. This information is valuable to know as it helps traders price storage within the potato market as there are no central mechanisms for price discovery within this market. However capturing this information is tricky as there is not a standard manner in which this information is disseminated. Word of mouth, conversations over the telephone, over Bloomberg chat, or over email are the ways in which price information can be passed. Additionally, when information is passed, the structure of the messages will not always be in a standard format and pieces of information can be missing. For example, one client can give you the number of potatoes being stored while another client may not. However, the fee and the type of potato will always be given.

Questions

* Design and draw a high level diagram of how you would go about building something that would help capture data from these various mediums.



* How would you structure these data to make it useful for traders?

-Merge data sources

Based on the diagram, traders can get the information from two sources: other storage companies and clients. The storage company can provide storage price (fee) for each type of potato and clients can provide the storage information of each type of potato. Based on these information, trader can structure the data with one record per potato type per client per day, and add them to position data. At the same time, traders can add new features about storage fee of each potato type. Here is a simple structured data frame:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Date | Product ID | Client ID | Quantity | Market Price | Storage Company1 | Storage Company2 | … |
| 2020-01-01 | 1 | 1 | 20 | 13 | 5% | 4.3% |  |
| … | … | … | … | … | … | … |  |
| 2020-01-01 | 1 | 5 | 13 | 13 | 5% | 4.3% |  |
| 2020-01-01 | 2 | 1 | 12 | 14 | 6% | 5% |  |

-Derive New Features

Apart from features directly transformed from data sources, traders can derive new features to help better predict the price of each type of potato, activity of clients and also the storage price in potato market. Here are some examples of new derived features:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| New features used to predict potato price | The number of client storing this type of potato | New features used to predict client activity | The total Market value | Storage Price | Average storage price |
| The total stored quantity | The total types of potato stored | Median storage price |
| The maximum proportion of quantity owned by one clients | The maximum proportion of storing one type of potato | Volatility per storage company |
| Volatility | volatility | Standard deviation of all storage company |
| … | … | … |

* What are potential uses of these data?

Those data are used to help trader predict potato price, client activity and storage rate, specifically:

1. Potato price

Trader can use the data to run a model and find out important factors that influence the price of potato. Those factors can come from fundamental aspect shown in the potato\_info, and also market aspect like the overall price trend of market and storage distribution among different clients. Traders can combine the effect of those factors by creating a signal index to show the trend of potato price. If the price of one type of potato price is likely to increase, trader can quickly buy the potato and sell it in the future.

1. Client activity

The client activity indicates the trend of client to buy or sell a certain type of potato, specifically it is to predict the quantity of one type of potato of one client. Traders can also find out factors that reflect the storage quantity of potato, it may come from the client information aspect as well as potato aspect like the price. By predicting the quantity, traders or sales people can sell the potato to clients who show the trend of increasing the storage of certain type of potato.

1. Storage fee

As the storage company, the storage fee is charged based on market value of stored potato. So to benefit the company from this aspect. If the client has the tendency to increase the storage when the price is low, I would charge more on the storage fee, however I cannot set it too high since other storage companies may provide lower fee than mine

1. Recommendation System

When a new potato comes in, since there is no historical data, we can use the potato information to cluster the new potato and make predictions of its price

When a new client comes in, we can use the historical data to build a recommendation system. For example, new client was clustered to a group of clients who have similar trading pattern, and the pattern shows they prefer holding a long position of a certain type of potato, and we can recommend the potato to the new client

1. Trading Strategy

As the seller/trader, if I find the client would like to buy a certain type of potato and other client would like the sell the potato at another price, it is a point for trader to make a decision to match the seller and buyer and optimize the price. (not sure if it is legal to make the margin)

Create a simple web app where sales people or traders can input the information that is passed to them

<https://www.w3schools.com/code/tryit.asp?filename=G22MPNGEJ31G>