FinTech Project Part A
Data Design and Analysis Report

Option #2

Nont Fakungkun

z5317972

Station #1: Data Lake

Product Description

The product is a finance-based product that provides business insights to add value to your clients. It uses a broad spectrum of market and sentiment indicators, news, weather, shipping and other heterogeneous data streams to provide insights that can help your clients make better decisions.

The product is designed for small and medium sized businesses, and can be used to track a variety of financial metrics, including:

- Market prices
- Cash flow
- Weather forecasts
- Correlation between market prices and weather

The product provides tools like reports, charts and graphs to help you analyse the data and identify trends.

Product Basic Use and Interface

The product will utilise various data sources, including a corn and wheat price database, weather database, and client cashflow records, to generate meaningful insights for clients. It will analyse and correlate these data sets to uncover trends, patterns, and relationships that can impact businesses in the agriculture and finance sectors.

The product can be used either online or offline. The online version is accessible through a web browser, and the offline version is available as a desktop software.

The interface will be user-friendly and intuitive where clients can access and interact with the generated insights. The main dashboard provides a high-level overview of the data, and the side panels provide more detailed information. Users will be able to explore visualisations, reports, and dashboards that provide a comprehensive view of market conditions, sentiment analysis, weather impacts, and financial performance.

Input Data

The input data consists of 5 database i.e., 4 excel sheets (1-4), and 1 json file (5):

1. Wheat trading data from 1 August 2018 to 31 July 2020
2. Corn trading data from 1 August 2018 to 31 July 2020
3. Bank account flows of 5 clients from 25 June 2019 to 26 June 2020
4. Relevant weather data from 1 July 2019 to 31 July 2020
5. Relevant news data in json format

What needs to be calculated and stored and cleaned

The weather database has some blanks space which needs estimation to fill the space. The blank space needs to respect the original data type of each section. The wind speed column has some rows labelled as Calm. I assume that this means the wind speed is 0. This needs to be converted to 0 in order to maintain the data type as integer (see Figure 1 in appendix). I then changed all column names by removing all whitespaces, and converted dates into a correct format.

The corn and wheat price history datasets are treated in the same way. To make them suitable for analysis with python code, I have removed unnecessary data like last price, bid and ask price, and open Interest. I also sort the data rows in ascending order based on dates using python, reverse_row.py. The processed dataset is saved as new files.

Output Data

The blank spaces in the weather dataset are filled by the mean of the existing value for each section. Other problems mentioned earlier are also addressed (see Figure 2 in appendix).

Processed corn and wheat price history data sheet. They are now in simpler form suitable for analysis (see Figure 3 and 4 in appendix).

From these processed datasets and other datasets, the data is imported into python code in the form of data frames. The data is cleaned to make it ready for analysis in further state.

1. Wheat trading data from 1 August 2018 to 31 July 2020
2. Corn trading data from 1 August 2018 to 31 July 2020
3. Bank account flows of 5 clients from 25 June 2019 to 26 June 2020
4. Relevant weather data from 1 July 2019 to 31 July 2020
5. Relevant news from JSON file.

How would you define Station #1

The data lake serves as the central repository for all of the raw data that an organisation collects. However, before the data can be analysed, it needs to be cleaned and processed. This includes removing errors, filling blank spaces and simplifying the databases. This mainly helps convert datasets into cleaned datastreams suitable for analysis in further stations.

Station #2: Prompt and Features Engineering

What are the inputs to achieve the results relevant to this station and delivering on financial product

Considering the business insight that this product should deliver, we need to think about what our client would want to know. The main problem for the small and medium sized agricultural entrepreneurs is what crops and when they should start planting in order to meet their financial goal. This goal is subjective and doesn't need to be maximised, for example, they might want to sell all of their agricultural products by Q3 even though the revenue is not at maximum.

We would need a thorough analysis of the total revenue of each crop. This includes how much crops can grow on the same amount of area as well as the price unit at that moment.

Hence, the inputs needed would be the price history of both crops for at least 1 year, so that we can conduct a full seasonal price analysis. If we have more than 1 complete year, we can also compare the price on the same date at different years.

In another perspective, if the users are instead contract traders, they would focus solely on the settlement price which the contracts are being traded at. They would want to know if the current price for each price is considered cheap or expensive. They will have a few simple questions, what and when to buy and then when to sell. So, they need a clear analysis of relative price, and price estimation.They would want to know which contracts give them more chance of making profit and if this is the right time to buy or sell.

What are the requirements in terms of data collection and data formats

As mentioned, the dataset containing required information needs to cover at least 1 full complete year. Clearly record the price and volume of each crops as well as weather data in details e.g., temperature, rainfall, wind gust, and humidity. The recorded data has to be labelled with date of recording and the date for all dataset has to be aligned for at least 1 full complete year so that we can have a thorough analysis of relationship between price, volume and weather for every season within a year.

Describe your core features

The core features of the financial product are then given corn and wheat settlement price and trading volume. These features are essential for both farmers and traders.

- Settlement price is the price at which a futures contract is settled at the end of the trading day. The settlement price is determined by the market and reflects the current supply and demand for the particular asset.

- Trading volume is the number of contracts that are traded on a given day. Trading volume is an indicator of market liquidity. A high trading volume indicates high activity in the market, which makes it easier to make transactions.

These features can be used to generate insights that can be used by farmers and traders to make informed decisions.

Price distribution - shows the distribution of prices for a particular crop over time. This information can be used to identify trends in the market and to make predictions. It is also helpful when determining relative price which means whether the price is currently underpriced or overpriced (see Figure 5 and 6 in appendix).

Seasonal price - shows the average price for a particular crop during different seasons of the year. As the data is from the Chicago Board of Trade, this seasonal analysis should be based on seasons period in the USA. This information can be used to help farmers plan their planting, harvesting, and stocking schedules.

Price correlation - shows how the price of a particular crop is correlated with the price of other crops or the price at different years. This information can be used to determine relationship strength between different contracts and years.

Trading volume - shows the volume of contracts that are traded for a particular crop on a given day. This information can be used to see how active the market is for a particular crop and to identify opportunities to buy or sell contracts at a good price.

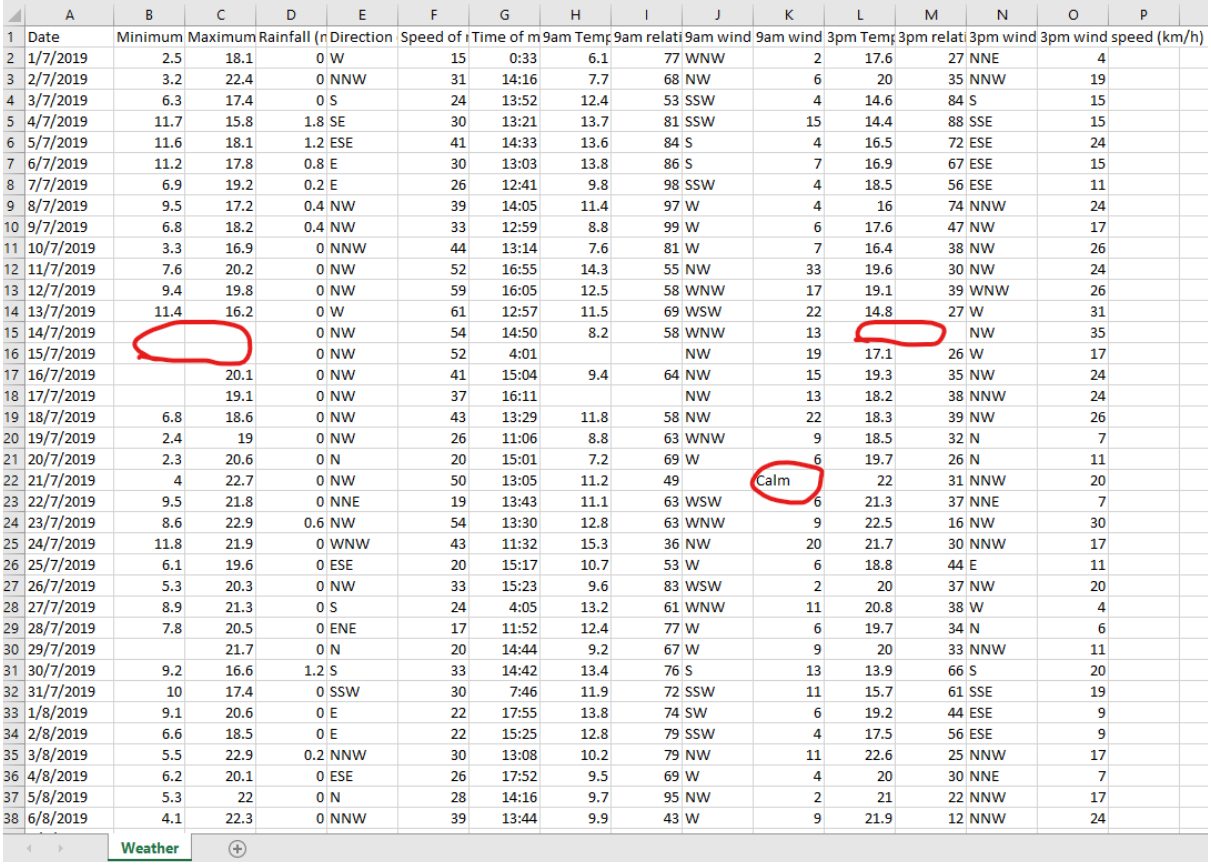
Farmers can decide when and what crops to plant using the settlement price and trading volume data. Farmers could choose to plant corn rather than wheat, for instance, if the settlement price for corn is higher than the settlement price for wheat. Farmers can also use the trading volume data to see how active the market is for a particular crop. Selling the crop at a good price might be harder if there is little trading activity.

Traders can use the settlement price and trading volume data to make decisions about what contracts they should trade, and when to buy and sell the contracts. Traders may choose to buy a contract if the settlement price is increasing, or sell a contract if the settlement price is falling. Traders can also use the trading volume data to see how active the market is for a particular contract. If there is high trading volume, it may be easier to buy or sell the contract at a good price.

How would you define Station #2

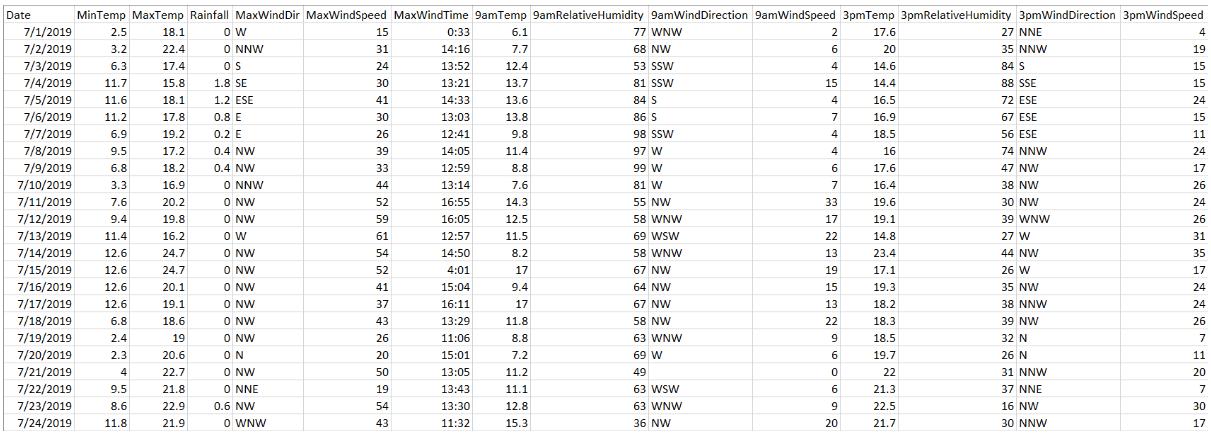
Features engineering is the process of selecting and transforming portions of essential raw data into features that are relevant to the problem being solved and can be used in the analysis. This requires us to have a clear understanding of the product delivery and what features are required in order to achieve that. The selected core feature will form the foundation of the analysis in further stations.

Appendix



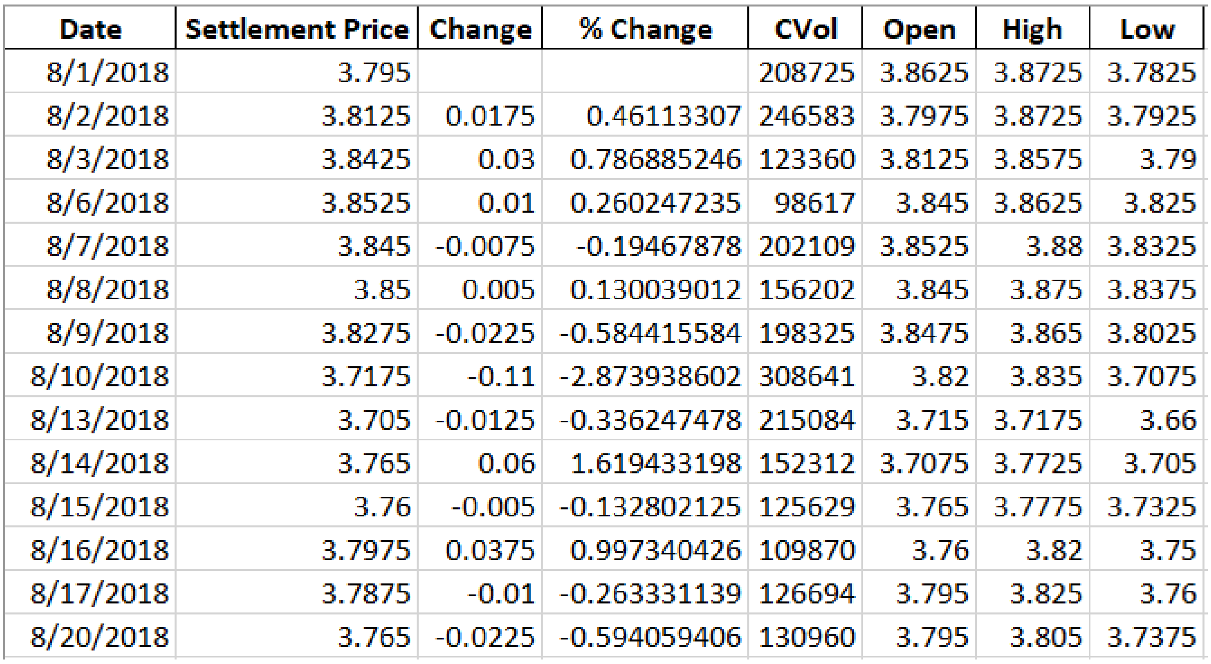
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Date	Minimum	Maximum	Rainfall (n	Direction	Speed of r	Time of m	9am Temp	9am relati	9am wind	9am wind	3pm Temp	3pm relati	3pm wind	3pm wind speed (km/h)	
2	1/7/2019	2.5	18.1	0	W	15	0:33	6.1	77	WNW	2	17.6	27	NNE	4	
3	2/7/2019	3.2	22.4	0	NNW	31	14:16	7.7	68	NW	6	20	35	NNW	19	
4	3/7/2019	6.3	17.4	0	S	24	13:52	12.4	53	SSW	4	14.6	84	S	15	
5	4/7/2019	11.7	15.8	1.8	SE	30	13:21	13.7	81	SSW	15	14.4	88	SSE	15	
6	5/7/2019	11.6	18.1	1.2	ESE	41	14:33	13.6	84	S	4	16.5	72	ESE	24	
7	6/7/2019	11.2	17.8	0.8	E	30	13:03	13.8	86	S	7	16.9	67	ESE	15	
8	7/7/2019	6.9	19.2	0.2	E	26	12:41	9.8	98	SSW	4	18.5	56	ESE	11	
9	8/7/2019	9.5	17.2	0.4	NW	39	14:05	11.4	97	W	4	16	74	NNW	24	
10	9/7/2019	6.8	18.2	0.4	NW	33	12:59	8.8	99	W	6	17.6	47	NW	17	
11	10/7/2019	3.3	16.9	0	NNW	44	13:14	7.6	81	W	7	16.4	38	NW	26	
12	11/7/2019	7.6	20.2	0	NW	52	16:55	14.3	55	NW	33	19.6	30	NW	24	
13	12/7/2019	9.4	19.8	0	NW	59	16:05	12.5	58	WNW	17	19.1	39	WNW	26	
14	13/7/2019	11.4	16.2	0	W	61	12:57	11.5	69	WSW	22	14.8	27	W	31	
15	14/7/2019			0	NW	54	14:50	8.2	58	WNW	13			NW	35	
16	15/7/2019			0	NW	52	4:01			NW	19	17.1	26	W	17	
17	16/7/2019		20.1	0	NW	41	15:04	9.4	64	NW	15	19.3	35	NW	24	
18	17/7/2019		19.1	0	NW	37	16:11			NW	13	18.2	38	NNW	24	
19	18/7/2019	6.8	18.6	0	NW	43	13:29	11.8	58	NW	22	18.3	39	NW	26	
20	19/7/2019	2.4	19	0	NW	26	11:06	8.8	63	WNW	9	18.5	32	N	7	
21	20/7/2019	2.3	20.6	0	N	20	15:01	7.2	69	W	6	19.7	26	N	11	
22	21/7/2019	4	22.7	0	NW	50	13:05	11.2	49		Calm	22	31	NNW	20	
23	22/7/2019	9.5	21.8	0	NNE	19	13:43	11.1	63	WSW	6	21.3	37	NNE	7	
24	23/7/2019	8.6	22.9	0.6	NW	54	13:30	12.8	63	WNW	9	22.5	16	NW	30	
25	24/7/2019	11.8	21.9	0	WNW	43	11:32	15.3	36	NW	20	21.7	30	NNW	17	
26	25/7/2019	6.1	19.6	0	ESE	20	15:17	10.7	53	W	6	18.8	44	E	11	
27	26/7/2019	5.3	20.3	0	NW	33	15:23	9.6	83	WSW	2	20	37	NW	20	
28	27/7/2019	8.9	21.3	0	S	24	4:05	13.2	61	WNW	11	20.8	38	W	4	
29	28/7/2019	7.8	20.5	0	ENE	17	11:52	12.4	77	W	6	19.7	34	N	6	
30	29/7/2019		21.7	0	N	20	14:44	9.2	67	W	9	20	33	NNW	11	
31	30/7/2019	9.2	16.6	1.2	S	33	14:42	13.4	76	S	13	13.9	66	S	20	
32	31/7/2019	10	17.4	0	SSW	30	7:46	11.9	72	SSW	11	15.7	61	SSE	19	
33	1/8/2019	9.1	20.6	0	E	22	17:55	13.8	74	SW	6	19.2	44	ESE	9	
34	2/8/2019	6.6	18.5	0	E	22	15:25	12.8	79	SSW	4	17.5	56	ESE	9	
35	3/8/2019	5.5	22.9	0.2	NNW	30	13:08	10.2	79	NW	11	22.6	25	NNW	17	
36	4/8/2019	6.2	20.1	0	ESE	26	17:52	9.5	69	W	4	20	30	NNE	7	
37	5/8/2019	5.3	22	0	N	28	14:16	9.7	95	NW	2	21	22	NNW	17	
38	6/8/2019	4.1	22.3	0	NNW	39	13:44	9.9	43	W	9	21.9	12	NNW	24	

Figure 1: A snapshot of a portion of weather database showing blank data space and data with unaligned format



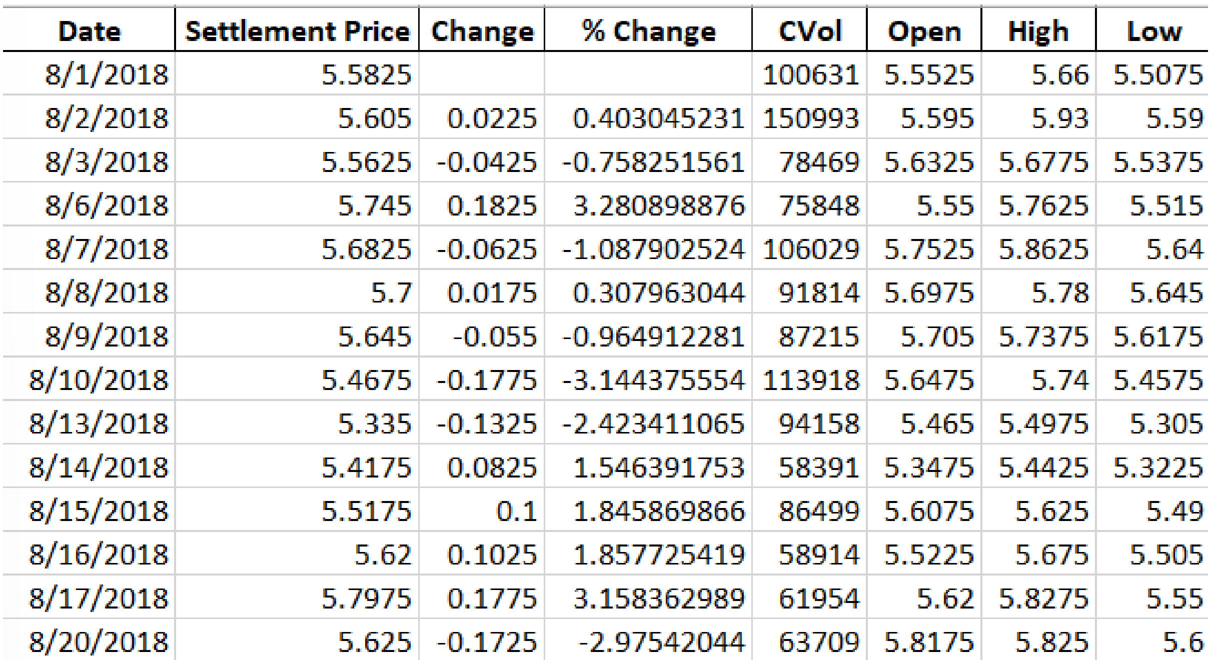
Date	MinTemp	MaxTemp	Rainfall	MaxWindDir	MaxWindSpeed	MaxWindTime	9amTemp	9amRelativeHumidity	9amWindDirection	9amWindSpeed	3pmTemp	3pmRelativeHumidity	3pmWindDirection	3pmWindSpeed
7/1/2019	2.5	18.1	0	W	15	0:33	6.1	77	WNW	2	17.6	27	NNE	4
7/2/2019	3.2	22.4	0	NNW	31	14:16	7.7	68	NW	6	20	35	NNW	19
7/3/2019	6.3	17.4	0	S	24	13:52	12.4	53	SSW	4	14.6	84	S	15
7/4/2019	11.7	15.8	1.8	SE	30	13:21	13.7	81	SSW	15	14.4	88	SSE	15
7/5/2019	11.6	18.1	1.2	ESE	41	14:33	13.6	84	S	4	16.5	72	ESE	24
7/6/2019	11.2	17.8	0.8	E	30	13:03	13.8	86	S	7	16.9	67	ESE	15
7/7/2019	6.9	19.2	0.2	E	26	12:41	9.8	98	SSW	4	18.5	56	ESE	11
7/8/2019	9.5	17.2	0.4	NW	39	14:05	11.4	97	W	4	16	74	NNW	24
7/9/2019	6.8	18.2	0.4	NW	33	12:59	8.8	99	W	6	17.6	47	NW	17
7/10/2019	3.3	16.9	0	NNW	44	13:14	7.6	81	W	7	16.4	38	NW	26
7/11/2019	7.6	20.2	0	NW	52	16:55	14.3	55	NW	33	19.6	30	NW	24
7/12/2019	9.4	19.8	0	NW	59	16:05	12.5	58	WNW	17	19.1	39	WNW	26
7/13/2019	11.4	16.2	0	W	61	12:57	11.5	69	WSW	22	14.8	27	W	31
7/14/2019	12.6	24.7	0	NW	54	14:50	8.2	58	WNW	13	23.4	44	NW	35
7/15/2019	12.6	24.7	0	NW	52	4:01	17	67	NW	19	17.1	26	W	17
7/16/2019	12.6	20.1	0	NW	41	15:04	9.4	64	NW	15	19.3	35	NW	24
7/17/2019	12.6	19.1	0	NW	37	16:11	17	67	NW	13	18.2	38	NNW	24
7/18/2019	6.8	18.6	0	NW	43	13:29	11.8	58	NW	22	18.3	39	NW	26
7/19/2019	2.4	19	0	NW	26	11:06	8.8	63	WNW	9	18.5	32	N	7
7/20/2019	2.3	20.6	0	N	20	15:01	7.2	69	W	6	19.7	26	N	11
7/21/2019	4	22.7	0	NW	50	13:05	11.2	49		0	22	31	NNW	20
7/22/2019	9.5	21.8	0	NNE	19	13:43	11.1	63	WSW	6	21.3	37	NNE	7
7/23/2019	8.6	22.9	0.6	NW	54	13:30	12.8	63	WNW	9	22.5	16	NW	30
7/24/2019	11.8	21.9	0	WNW	43	11:32	15.3	36	NW	20	21.7	30	NNW	17

Figure 2: A snapshot of a portion of processed weather database



Date	Settlement Price	Change	% Change	CVol	Open	High	Low
8/1/2018	3.795			208725	3.8625	3.8725	3.7825
8/2/2018	3.8125	0.0175	0.46113307	246583	3.7975	3.8725	3.7925
8/3/2018	3.8425	0.03	0.786885246	123360	3.8125	3.8575	3.79
8/6/2018	3.8525	0.01	0.260247235	98617	3.845	3.8625	3.825
8/7/2018	3.845	-0.0075	-0.19467878	202109	3.8525	3.88	3.8325
8/8/2018	3.85	0.005	0.130039012	156202	3.845	3.875	3.8375
8/9/2018	3.8275	-0.0225	-0.584415584	198325	3.8475	3.865	3.8025
8/10/2018	3.7175	-0.11	-2.873938602	308641	3.82	3.835	3.7075
8/13/2018	3.705	-0.0125	-0.336247478	215084	3.715	3.7175	3.66
8/14/2018	3.765	0.06	1.619433198	152312	3.7075	3.7725	3.705
8/15/2018	3.76	-0.005	-0.132802125	125629	3.765	3.7775	3.7325
8/16/2018	3.7975	0.0375	0.997340426	109870	3.76	3.82	3.75
8/17/2018	3.7875	-0.01	-0.263331139	126694	3.795	3.825	3.76
8/20/2018	3.765	-0.0225	-0.594059406	130960	3.795	3.805	3.7375

Figure 3: A snapshot of a portion of the processed corn price database



Date	Settlement Price	Change	% Change	CVol	Open	High	Low
8/1/2018	5.5825			100631	5.5525	5.66	5.5075
8/2/2018	5.605	0.0225	0.403045231	150993	5.595	5.93	5.59
8/3/2018	5.5625	-0.0425	-0.758251561	78469	5.6325	5.6775	5.5375
8/6/2018	5.745	0.1825	3.280898876	75848	5.55	5.7625	5.515
8/7/2018	5.6825	-0.0625	-1.087902524	106029	5.7525	5.8625	5.64
8/8/2018	5.7	0.0175	0.307963044	91814	5.6975	5.78	5.645
8/9/2018	5.645	-0.055	-0.964912281	87215	5.705	5.7375	5.6175
8/10/2018	5.4675	-0.1775	-3.144375554	113918	5.6475	5.74	5.4575
8/13/2018	5.335	-0.1325	-2.423411065	94158	5.465	5.4975	5.305
8/14/2018	5.4175	0.0825	1.546391753	58391	5.3475	5.4425	5.3225
8/15/2018	5.5175	0.1	1.845869866	86499	5.6075	5.625	5.49
8/16/2018	5.62	0.1025	1.857725419	58914	5.5225	5.675	5.505
8/17/2018	5.7975	0.1775	3.158362989	61954	5.62	5.8275	5.55
8/20/2018	5.625	-0.1725	-2.97542044	63709	5.8175	5.825	5.6

Figure 4: A snapshot of a portion of the processed wheat price database

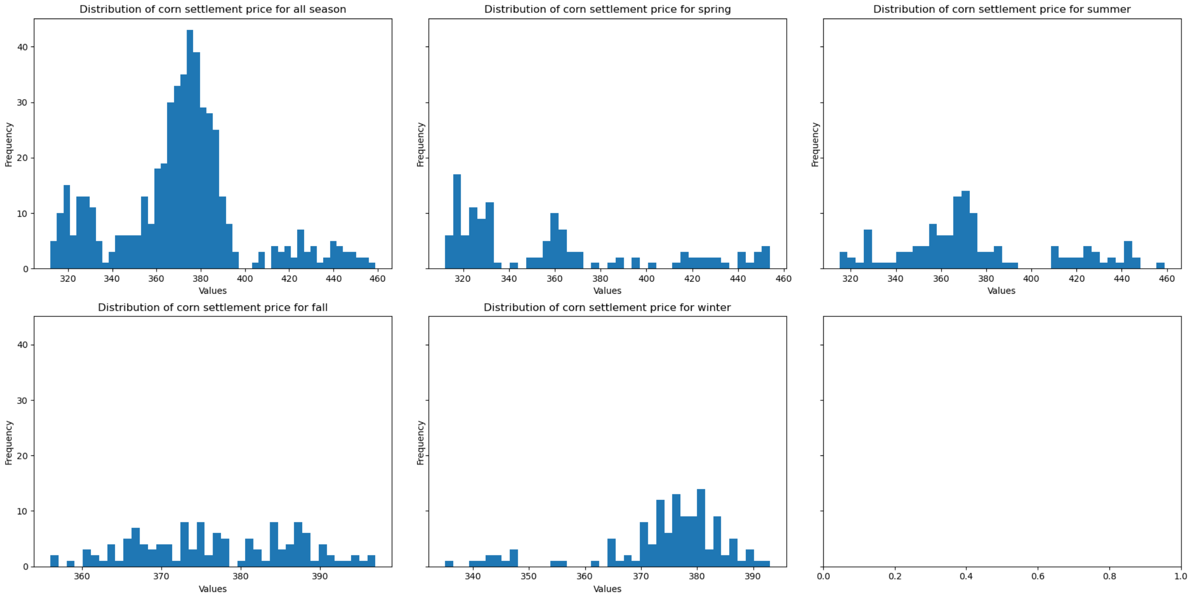


Figure 5: A snapshot of overall and seasonal histograms of corn settlement price

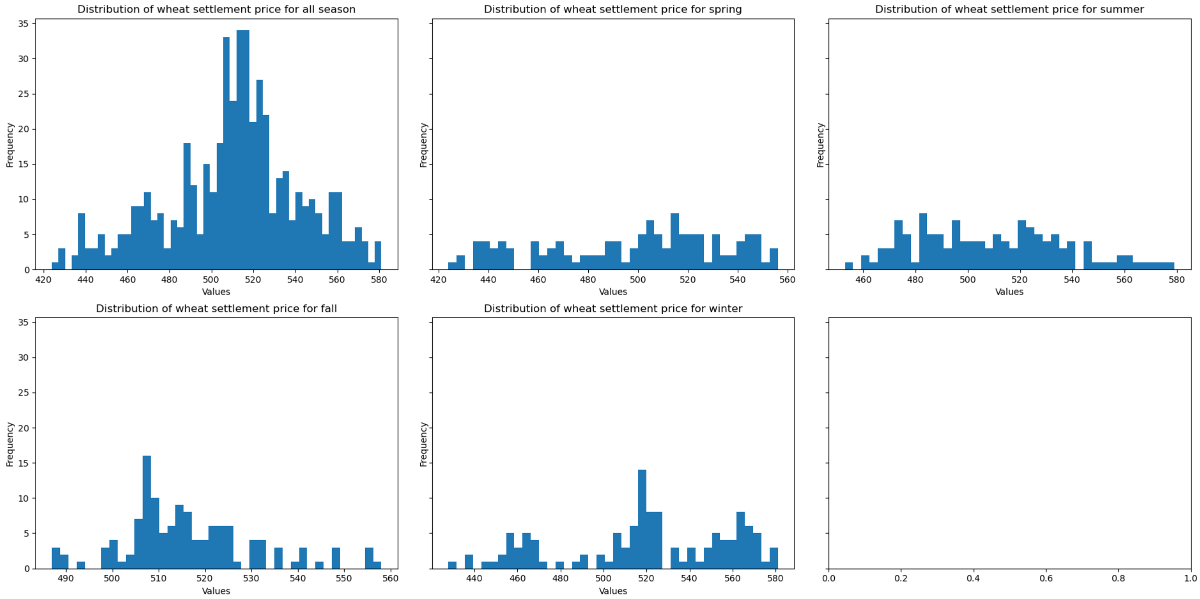


Figure 6: A snapshot of overall and seasonal histograms of wheat settlement price