DATA1902
Project Stage 1 - Report

We are using three datasets for this project, Bitcoin tweets and prices, crime in Chicago and the price of gold. All three datasets are time data, meaning they all share an attribute relating the other columns to a date and time. Similarly, the two financial datasets on gold and Bitcoin are intrinsically related since they record the trajectory of the financial market over time. Whilst the Chicago crime dataset contains categories of computer-related fraud and financial theft which are connected to the domain of cryptocurrency (which Bitcoin falls under) and the movement of financial prices which the gold dataset relates to. In addition, the movement of economic markets and people's sentiments are highly related to rates of crime.

1) The Data Sources

Bitcoin Tweets & Prices

URL: https://www.kaggle.com/jaimebadiola/bitcoin-tweets-and-price/version/2

We obtained this dataset from Kaggle, where it was uploaded on the 13th of February 2019 by Jaime Badiola who used the GitHub account "GetOldTweets" to download over 17.7 million tweets related to Bitcoin (from 2017/08/01 to 2019/01/22). They used this data to form the dataset uploaded onto Kaggle. The most recent update, which is the version used for this project, was Version 2. It was updated on 2019/06/15 and made sentiment data more accurate by filtering out bot and spam tweets and eliminating their sentiment scores. As a result, the average sentiment column has not been skewed by these tweets. The data has 12 937 rows (including the header) and 16 columns.

The data is presented in hourly periods with each row containing data on the number of tweets and the sentiment data associated with those tweets within the same time period. The creator assigned a sentiment score to each tweet using the library VaderSentiment which is used to calculate how positive or negative social media texts are with a more positive tweet obtaining a higher positive number, neutral tweets obtaining a score of 0 and negative tweets obtaining smaller and smaller negative values the more negative the tweet is. The sentiment score is between 1 and -1. The dataset also contains 6 columns of bitcoin price information for each hourly time period.

These columns are:

- **Date:** the starting time of the 1 hr time span for which all data in a row corresponds to (in 24-hour time format: yyyy-mm-dd hh:mm:ss)
- **Compound_Score:** average of sentiment scores for all tweets after filtering out bot and spam tweets
- **Total Volume of Tweets**: total number of tweets, includes bot and spam tweets
- **Count_Negatives:** total number of negative tweets (doesn't include bot/spam tweets)
- **Count_Positives:** total number of positive tweets (doesn't include bot/spam tweets)
- **Count_Neutrals:** total number of neutral tweets (doesn't include bot/spam tweets)
- **Sent_Negatives:** average of all negative sentiment scores only (doesn't include bot/spam tweets)
- **Sent_Positives:** average of all positive sentiment scores only (doesn't include bot/spam tweets)
- **Count_News:** number of tweets that contains a link
- **Count_Bots:** number of tweets classified as bots
- **Open:** first bitcoin price during the 1 hr time span (units: USD)
- **High:** highest bitcoin price during the 1 hr time span (units: USD)
- **Low:** lowest bitcoin price during the 1 hr time span (units: USD)
- **Close:** last bitcoin price during the 1 hr time span (units: USD)
- **Volume (BTC):** the amount of bitcoin that underwent transactions (units: BTC)
- **Volume (Currency):** the amount of bitcoin that underwent transactions (units: USD)

The license for the data set is listed under *CC0: Public Domain* where "You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission."
Thus, we were allowed to use and modify this dataset without obtaining any permission beforehand.

Strengths: There are many useful attributes for data analysis including sentiment analysis, price data, the volume traded and the number of news and bot tweets. This data is recorded consistently at a high frequency (namely by the hour) and has very detailed values for each recording to many decimal places. It has also been scraped professionally using a legitimate library and GitHub repository.

Limitations: Some hours contain missing data, and the frequency of spelling mistakes on social media may have included or excluded tweets by accident. Since the list of 30 phrases used to accumulate data on Bitcoin tweets is not provided we cannot interpret the validity

of the sourcing to a high degree. The dataset is semi-colon separated which proves annoying to read into a data frame for most python libraries.

Chicago Crime

URL: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data

The Chicago Crime dataset is from the Chicago City Portal. It's regularly updated so that the data reflects crimes from 2001 to present (minus the last seven days). The data itself is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. The downloaded raw data set has 22 columns and 2 099 624 rows (including the header) and contains data on crimes from 2012/04/04 to 2019/09/28 and was most recently updated on 2019/10/09.

Each row is data for a single crime containing its own ID, case number, date, time (which is sometimes an estimate), crime description and location information. More specific location information is partially redacted by altering it so that the exact location is not revealed but the block where the crime took place remains the same.

The columns have the following attributes:

- **ID:** unique identifier for the reported crime
- **Case Number:** Chicago Police Department Records Division Number (also unique)
- **Date:** date and time in 12h hr time (format: mm:dd:yyyy hh:mm:ss AM/PM)
- **Block:** refers to the block (location) where the crime took place
- **IUCR:** Illinois Uniform Crime Reporting (IUCR) code. A four-digit code used to classify criminal incidents when taking individual reports.
- **Primary Type:** primary description of IUCR code
- **Description:** secondary description of IUCR code, a subcategory of the primary description
- **Location Description:** short description of the type of location where the crime occurred
- **Arrest:** True/False - whether or not an arrest was made
- **Domestic:** True/False - whether or not classified as a domestic crime according to the Illinois Domestic Violence Act
- **Beat:** the smallest police geographic area in which the crime occurred. A four-digit code
- **District:** the police district where the crime occurred. Three-digit code
- **Ward:** the ward, a city council district, where the crime occurred
- **Community Area:** community area where crime occurred, Chicago has 77 community areas numbered from 1-77

- **FBI Code:** classification code under National Incident-Based Reporting System (NIBRS)
- **X Coordinate:** the x-coordinate of the crime's location on a map of Chicago (7 digits)
- **Y Coordinate:** the y-coordinate of the crime's location on a map of Chicago (7 digits)
- **Year:** year in which crime occurred
- **Updated On:** date and time the data was last updated (mm:dd:yyyy hh:mm:ss AM/PM)
- **Latitude:** geographic latitude coordinate
- **Longitude:** geographic longitude coordinate
- **Location:** both the latitude and longitude values combined as one coordinate

The website states that any user providing any 'derivative application' of the data on the website must include the following disclaimer:

> "This site provides applications using data that has been modified for use from its original source, www.cityofchicago.org, the official website of the City of Chicago. The City of Chicago makes no claims as to the content, accuracy, timeliness, or completeness of any of the data provided at this site. The data provided at this site is subject to change at any time. It is understood that the data provided at this site is being used at one's own risk."

Thus we are able to use the data provided the above disclaimer is included wherever the project materials can be accessed/downloaded.

Strengths: The attributes of the dataset are very detailed including the "Case Number", a multiplicity of location information for later mapping as well as interesting information regarding the crime itself under the "Primary Type" and "Description" columns. It also includes whether or not the individual was arrested or if it was a domestic case. In addition, the dataset is updated regularly each week and includes up to 500 cases per day with detailed time information for each day from 2002.

Limitations: The website states that the crimes reported may not have been verified, while the "preliminary crime classifications may be changed at a later date based upon additional investigation and there is always the possibility of mechanical or human error". This poses some questions regarding the dataset's integrity and accuracy which is a great limitation against its use.

Gold Prices

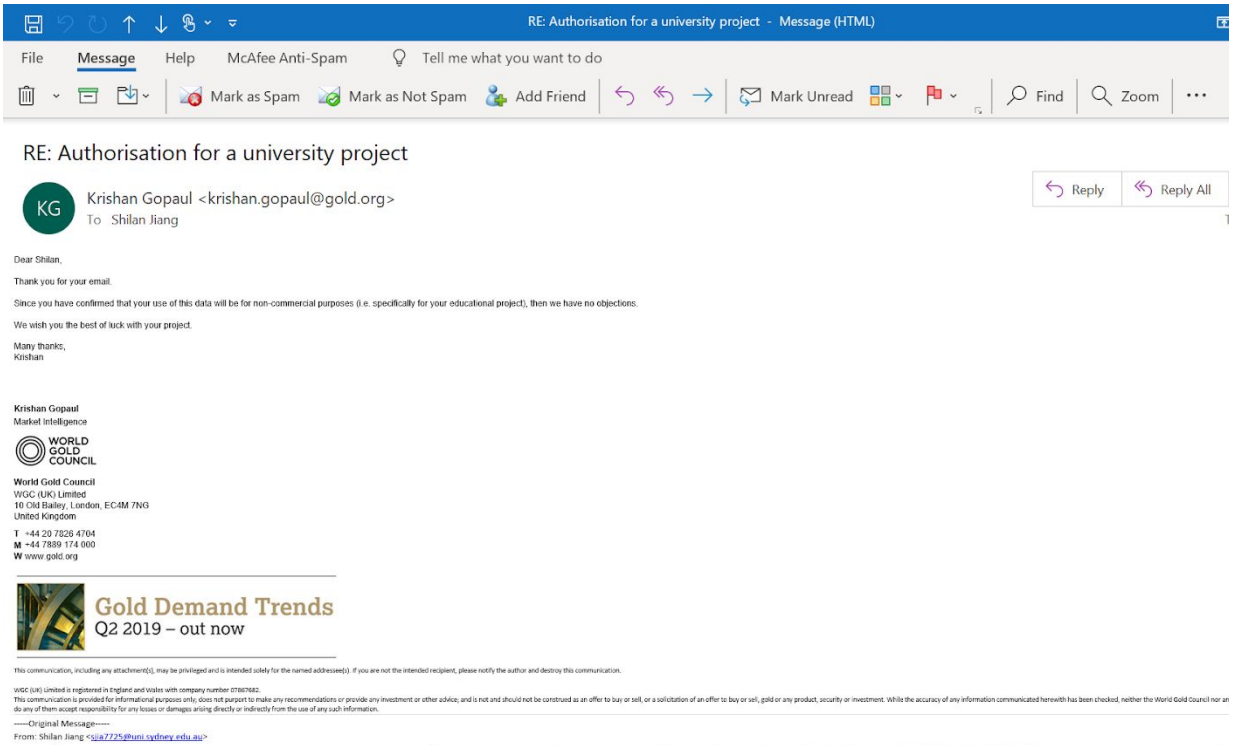
URL: https://www.gold.org/goldhub/data/gold-prices

The Gold Price dataset was obtained from Gold Hub - a website containing gold prices, market data and research from the World Gold Council. It provides data on daily gold prices from 2010-01-01 to 2019-09-27. It has 2 columns and 2542 rows, including the header row.

Under “Use of this website”, it is stated that:

> “This Website and the information and materials on this Website are provided for general informational and educational purposes only.
> You are permitted to save, display or print out information contained on this Website only for your personal, non-commercial use. Except as otherwise permitted by these terms and conditions, you are not permitted to modify, copy, scrape, distribute, transmit, display, reproduce, duplicate, publish, license, frame, link, create derivative works from, transfer or otherwise use in any manner, in whole or in part, this Website or the information and materials on this Website without the prior written authorization of the World Gold Council and with suitable accreditation to the World Gold Council. To request such authorization, please contact us at info@gold.org.”

Upon emailing them, we were granted authorisation to modify the dataset for this project. As seen below:



RE: Authorisation for a university project - Message (HTML)

File Message Help McAfee Anti-Spam Tell me what you want to do

Mark as Spam | Mark as Not Spam | Add Friend | Mark Unread | Find | Zoom

RE: Authorisation for a university project

KG Krishan Gopaul <krishan.gopaul@gold.org>
To Shilan Jiang

Reply Reply All

Dear Shilan,

Thank you for your email.

Since you have confirmed that your use of this data will be for non-commercial purposes (i.e. specifically for your educational project), then we have no objections.

We wish you the best of luck with your project.

Many thanks,
Krishan

Krishan Gopaul
Market Intelligence
WORLD GOLD COUNCIL
World Gold Council
WGC (UK) Limited
10 Old Bailey, London, EC4M 7NG
United Kingdom
T +44 20 7826 4704
M +44 7889 174 000
W www.gold.org

Gold Demand Trends
Q2 2019 – out now

-----Original Message-----
From: Shilan Jiang <sjia7725@uni.sydney.edu.au>

Column descriptions:

- **Date:** the date in the format yyyy:mm:dd
- **Value:** the price of gold for the day (units: USD per troy ounce)

Strengths: The dataset is quite simple containing only two attributes and with a single gold price for each day. This means it is easy to work with, configure and upload/download due to its low file size in comparison to the other two datasets used. There are no missing values, meaning the dataset is relatively clean and reliable already.

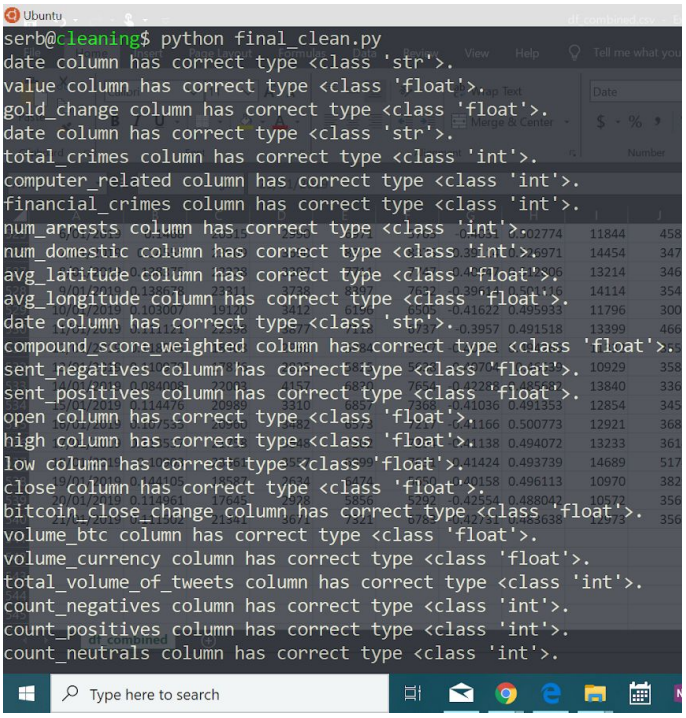
Limitations: Its simplicity means there is also not a large amount of detailed data for each entry. Also the prices for Saturday and Sunday are unavailable since gold is not traded on weekends. This means that it is necessary to extend the price for Friday over those days for completeness and comparison with the other two datasets we intend to work with.

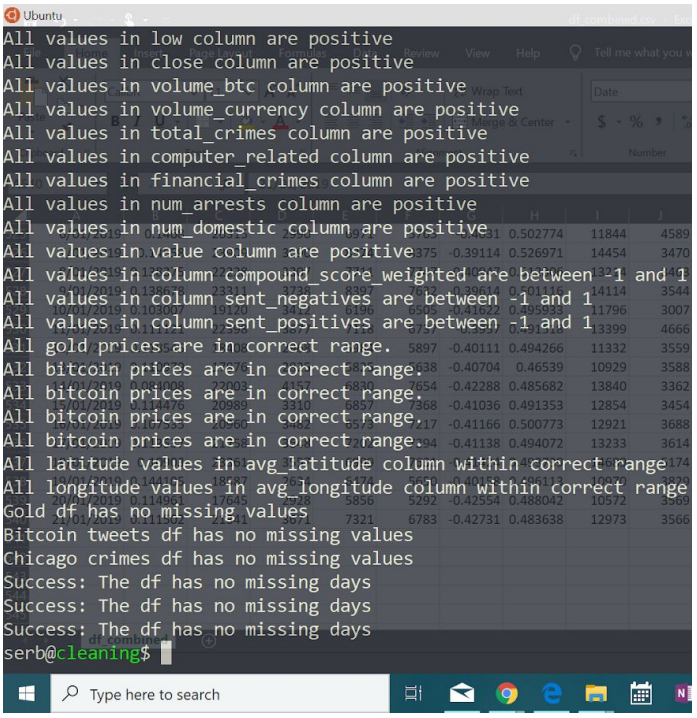
2) Data Cleaning and Transformation

Initially, it was important to establish a common date format between each dataset since they differed in terms of detail and the ordering of year, month and day. Quite annoyingly, the Chicago crimes dataset was ordered mm/dd/yyyy. After turning all dates to the format yyyy-mm-dd, the next task was to make each dataset share common date ranges. This required finding the earliest and most recent dates for each dataset and then finding the intersection of all these ranges - leading to the combined dataset ranging from 2017-08-01 to 2019-01-21.

For each dataset, the column names were changed to lowercases and snake case for ease of use, whilst the types of each column were cast according to the type of data in each respective attribute.

To save time in the cleaning process, some functions were created with one or several columns required as parameters. For example, there was a `check_type()` function which would check to see if every value in the column had the correct type. Whilst there was also a `check_greater_than_zero()` function to check whether each value in the column was greater than 0. By creating lots of functions like these with helpful error and success messages, we could abstract repetitive code to simplify our data cleaning process. The output of using such functions looks like so:





Bitcoin Tweets & Prices

For our final data set, we wanted each row to represent one day in order to make comparisons between the gold dataset and the price and tweets related to Bitcoin. However, the Bitcoin data set contained data for each hour, hence there were 24 rows for each day. Thus, we transformed the dataset from hourly data to daily data.

First of all, we need to replace all commas to decimal places since many numbers contained undesirable commas. Afterwards, the count columns such as the total volume of tweets and the count of negative tweets were simply added up using Pandas. Whilst the compound score needed to be weighted according to the volume of tweets for each hour of that day. We then took the average over the hours for positive and negative sentiments. The opening price was taken to be the first price of that day, the end price the last and the high and low as the maximum and minimum price over that day respectively. Furthermore, another column was created which showed the percentage change of the bitcoin price for that day. This was done using the nifty Pandas function `pct_change()`.

Chicago Crime

We needed each row in the Chicago crime data set to represent one day in order for it to be combined with the other datasets into one CSV file. However, this data set had roughly 500 rows for each day where each row represents a single crime. Hence, we aggregated all the

data for one day into one row. This was done by creating another column which was the total number of crimes for that day. Two more columns were also created, one for the number of computer-related crimes and one for the number of financial-related crimes for the day. This was achieved by searching through the description column and seeing whether the string 'COMPUTER' or 'FINANCIAL' was in it. We created these columns as we may want to see whether crimes related to these areas affect the financial markets.

The raw dataset also contained two columns, one for arrests and the other for domestic crimes, which contained either the string "true" or "false". Hence, we aggregated this into the number of arrests and domestic-related crimes for each day into two new columns. The latitude and longitude columns were also averaged for each day. Since the rest of the columns had values/attributes which were unique to that specific crime, these were discarded.

Gold Prices

The gold price data set had a specific date for each row which is the format we wanted it to be in. However, since gold is not traded on the weekends, there was no data for the gold price on weekends which meant that there were missing dates in the data set. To be able to combine it with the other data sets, the missing dates were added in so extra rows were created. We let the gold price for these dates to be the last price that the gold was traded at.

Furthermore, another column was created which showed the percentage change of the gold price for that day. This was accomplished using the pct_change() function from the Pandas module.

In the end, prior to our analysis, we combined the three datasets on the common "date" column using Pandas' pd.merge() function.

3) Analysis

During the transforming process, summary statistics were created for each day when the hourly data for the bitcoin tweets and Boston crimes data sets were aggregated to contain only single days.

In particular, we broke down these complex datasets into sensible ways and out emerged summary statistics. For example, the total number of tweets, computer-related crimes and the average longitude and latitude for each day was found. In the Chicago crimes data set, we aggregated over unique days to determine how many crimes there were overall in each day. We also counted how many of the crimes for a specific day had an arrest or was domestic-related.

In addition to aggregating the data set by reducing the volume of data, we also added useful information. For example, we calculated the percentage change in bitcoin and gold prices for each day and displayed these values in the “bitcoin_close_change” and “gold_change” columns respectively.

Furthermore, summary statistics were created for each of the columns in the final dataset. These summary statistics are:

- Maximum value
- Minimum value
- Standard deviation
- Mean
- Median, and
- Interquartile range

The following table was created as a result of some simple aggregating functions in the Pandas library

	A	B	C	D	E	F	G	H	I	J	K	L
1	statistics	date	compound_score_weighted	total_volume_of_tweets	count_negatives	count_positives	count_neutrals	sent_negatives	sent_positives	count_news	count_bots	open
2	max	21/01/2019	0.203200514	132408	32143	45178	48134	-0.337225421	0.554544422	69766	24900	19346.6
3	min	1/08/2017	-0.013692694	81	3	9	11	-0.529889461	0.375211111	72	58	2703.51
4	sd		0.0305515	16176.98471	3378.673303	5273.485816	6548.242714	0.022588442	0.026942707	10525.45293	1914.611867	3133.486
5	mean		0.101490422	32908.18738	5301.315399	10332.18738	12139.66976	-0.408969784	0.472599558	21373.95918	5135.014842	7309.139
6	median		0.100809401	28725	4376	8965	10623	-0.410040903	0.475148564	19002	4855	6620.63
7	iqr		0.036971925	14187	2917.5	4351.5	5834	0.027339868	0.031819739	10459	2500	3604.5

	A	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	statistics	open	high	low	close	bitcoin_close_change	volume_btc	volume_currency	gold_value	gold_change	total_crimes	computer_related	financial_crimes	num_arrests
2	max	19346.6	19870.62	18873.37	19345.49	23.68561677	7.94E+16	8.54E+17	1354.95	2.594724967	988	5	67	231
3	min	2703.51	2760	2615	2703.51	-17.23513385	208.56	845839.7	1178.4	-1.882789711	394	0	2	52
4	sd	3133.486	3302.245	2903.65	3138.789	4.73787038	7.53512E+15	2.01E+17	46.05396348	0.475434388	86.10024455	0.774375819	9.350496165	21.91012087
5	mean	7309.139	7550.785	7024.731	7311.971	0.161783808	1.40982E+15	2.68E+17	1274.062894	0.002380056	731.8831169	0.515769944	17.82931354	143.7792208
6	median	6620.63	6756.36	6501.01	6633.14	0.163491425	423726893.6	3.00E+17	1281.2	0	733	0	17	143
7	iqr	3604.5	3518.85	3404.535	3587.4	4.45069341	1500246439	3.62E+17	81	0.360274588	108	1	9	27

	A	X	Y	Z	AA
1	statistics	num_arrests	num_domestic	avg_latitude	avg_longitude
2	max	231	212	41.85873278	-87.66293781
3	min	52	70	41.83353805	-87.67646786
4	sd	21.91012087	21.39476678	0.00385782	0.00237059
5	mean	143.7792208	118.0037106	41.844896	-87.66995364
6	median	143	115	41.84497117	-87.66995042
7	iqr	27	26.5	0.005362886	0.003349373

4) References

Creative Commons. (n.d.). *CC0 1.0 Universal (CC0 1.0) Public Domain Dedication.* [online] Creative Commons. Available at: https://creativecommons.org/publicdomain/zero/1.0/ [Accessed 7. Oct 2019].

Badiola, J. (2019). *Bitcoin 17.7 million Tweets and price.* [online] Kaggle. Available at: https://www.kaggle.com/jaimebadiola/bitcoin-tweets-and-price/version/2 [Accessed 7. Oct 2019].

Pandey, P. (2018). *Simplifying Sentiment Analysis using VADER in Python (on Social Media Text)*. [online] Medium. Available at: https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f [Accessed 11. Oct 2019].

(2019) *Crimes - 2001 to present.* [online] Chicago Data Portal. Available at: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data [Accessed 10. Oct 2019].

City of Chicago (n.d.). *Data Terms of Use.* [online] City of Chicago. Available at: https://www.chicago.gov/city/en/narr/foia/data_disclaimer.html [Accessed 8. Oct 2019].

Gold Hub. (2019). *Gold Prices.* [online] World Gold Council. Available at: https://www.gold.org/goldhub/data/gold-prices [Accessed 7. Oct 2019].

World Gold Council. (n.d). *Terms and Conditions.* [online] World Gold Council. Available at: https://www.gold.org/terms-and-conditions [Accessed 7. Oct 2019].