

PROJECT 2 – ETL

ASX - 200

1st of August 2022

Group 9;

David Salim
Christopher Mai
Johan Ehrhardt
Krystal Enweya



CONTENTS;

1. Project Aim

2. Extraction

2.1 Table 1

2.2 Table 2

3. Transformation

3.1 Removing columns

3.2 Dropping duplicates

3.3 Using functions

3.4 Formatting data

3.5 Formatting data type

4. Load

4.1 Create SQL database

4.2 Creating connections

4.3 Loading data frames into SQL

4.4 Joining tables

4.5 Creating ‘View’ function

5. Schema

6. ERD Diagram

1. Project aim;

The aim of our project was to collect the data regarding the ASX 200. From this collected data the aim was to create data bases and join tables to provide a relevant and effective summary of the ASX 200

2. Extraction;

Extraction of the data used within this project can be divided based on content of tables;

Table 1, 'ASX 200 Companies'

- Retrieved from the following website; <https://www.marketindex.com.au/asx200>
- Data regarding the ASX 200
- Base data from web therefore using web scrapping
- Splinter and google chrome driver manager used to interact with the web
- Beautiful soup and for loop used to extract data from the table
- Pandas used to create data frames from the extracted data

Code ↑ Company	↓ Price	↓ Chg	↓ % Chg	High	Low	Volume	↓ Mkt Cap	↓ 1 Year
↑ A200 Betashares Australia 200 ETF	\$116.21	+0.71	+0.62%	\$116.30	\$115.59	46,534	\$2.2B	-6.97%
- A2M The a2 Milk Company Ltd	\$4.54	0.00	0.00	\$4.60	\$4.49	2,382,120	\$3.4B	-22.53%
↓ AAA Betashares Australian High Interest Cash ETF	\$50.08	-0.05	-0.10%	\$50.08	\$50.07	43,341	\$1.8B	-0.02%
↓ ABP Abacus Property Group	\$2.92	-0.01	-0.34%	\$2.96	\$2.89	1,158,318	\$2.6B	-7.89%
↑ AFI Australian Foundation Investment Company Ltd	\$8.02	+0.02	+0.25%	\$8.14	\$7.99	434,943	\$9.9B	-4.75%
↑ AGL AGL Energy Ltd	\$8.45	+0.07	+0.84%	\$8.45	\$8.325	1,110,766	\$5.7B	+16.87%
↑ AIA Auckland International Airport Ltd	\$6.71	+0.04	+0.60%	\$6.74	\$6.65	471,451	\$9.9B	-1.32%
↑ AIZ Air New Zealand Ltd	\$0.565	+0.02	+3.67%	\$0.565	\$0.545	719,103	\$1.9B	-60.21%
↑ AKE Allkem Ltd	\$11.79	+0.51	+4.52%	\$11.92	\$11.30	4,585,892	\$7.5B	+42.91%
↑ ALD Ampol Ltd	\$34.01	+0.51	+1.52%	\$34.14	\$33.50	680,050	\$8.1B	+20.39%
↓ ALL Aristocrat Leisure Ltd	\$35.21	-0.09	-0.26%	\$35.55	\$35.10	1,910,979	\$23B	-15.10%
↑ ALQ Als Ltd	\$11.66	+0.02	+0.17%	\$11.75	\$11.53	910,430	\$5.6B	-8.48%
↓ ALU Altium Ltd	\$30.64	-0.53	-1.70%	\$31.63	\$30.56	284,540	\$4B	-9.62%
↑ ALX Atlas Arteria	\$7.76	+0.01	+0.13%	\$7.96	\$7.71	5,021,120	\$7.4B	+23.76%
↑ AMC Amcor Plc	\$18.58	+0.09	+0.49%	\$18.58	\$18.43	1,896,189	\$13B	+17.97%
↑ AMP AMP Ltd	\$1.105	+0.02	+1.84%	\$1.11	\$1.085	9,524,894	\$3.6B	+6.25%

Figure 1. Raw data, from web

Table 2, 'ASX Companies'

- Retrieved from the following website; <https://www.listcorp.com/asx/>
- Data regarding ASX Companies
- CSV file download
- Pandas to extract data
- Pandas used to create data frames from the extracted data

A	B	C	D	E	F	G	H	I	J
1	Code	Company	Link	Market Cap	Last trade	Change	%Change	Sector	
2	ASX:BHP	BHP Group	https://www.listcorp.com/asx/BHP	193988000	38.68	0	0	Materials	
3	ASX:CBA	Commonwealth Bank	https://www.listcorp.com/asx/CBA	169967000	100.77	0	0	Financials	
4	ASX:CSL	CSL Limited	https://www.listcorp.com/asx/CSL	140369000	289.84	0	0	Health Care	
5	ASX:NAB	National Australia Bank	https://www.listcorp.com/asx/NAB	963842000	30.6	0	0	Financials	
6	ASX:WBC	Westpac Banking Corporation	https://www.listcorp.com/asx/WBC	750292000	21.51	0	0	Financials	
7	ASX:MQG	Macquarie Group	https://www.listcorp.com/asx/MQG	691368000	181.13	0	0	Financials	
8	ASX:ANZ	ANZ Bank	https://www.listcorp.com/asx/ANZ	636240000	22.9	0	0	Financials	
9	ASX:WDS	Woodside Petroleum	https://www.listcorp.com/asx/WDS	607030000	31.98	0	0	Energy	
10	ASX:FMG	Fortescue Metals Group	https://www.listcorp.com/asx/FGM	575766000	18.34	0	0	Materials	
11	ASX:WES	Wesfarmers	https://www.listcorp.com/asx/WES	524968000	46.63	0	0	Consumer Discretionary	
12	ASX:WOW	Woolworths Group	https://www.listcorp.com/asx/WOW	454485000	37.52	0	0	Consumer Staples	
13	ASX:TLS	Telstra	https://www.listcorp.com/asx/TLS	450623000	3.89	0	0	Communication Services	
14	ASX:TCL	Transurban	https://www.listcorp.com/asx/TCL	440955000	14.51	0	0	Industrials	
15	ASX:GMG	Goodman	https://www.listcorp.com/asx/GMG	369721000	20.7	0	0	Real Estate	
16	ASX:RIO	Rio Tinto	https://www.listcorp.com/asx/RIO	362678000	97.83	0	0	Materials	
17	ASX:COL	Coles Group	https://www.listcorp.com/asx/COL	249084000	18.75	0	0	Consumer Staples	
18	ASX:STO	Santos	https://www.listcorp.com/asx/STO	242935000	7.3	0	0	Energy	
19	ASX:ALL	Aristocrat Leisure	https://www.listcorp.com/asx/ALL	240094000	35.3	0	0	Consumer Discretionary	
20	ASX:QBE	QBE Insurance	https://www.listcorp.com/asx/QBE	176063000	11.53	0	0	Financials	
21	ASX:S32	South32	https://www.listcorp.com/asx/S32	174955000	3.81	0	0	Materials	
22	ASX:ASX	ASX Limited	https://www.listcorp.com/asx/ASX	168447000	88.26	0	0	Financials	
23	ASX:NCM	Newcrest Mining	https://www.listcorp.com/asx/NCM	166210000	19.3	0	0	Materials	
24	ASX:REA	REA Group	https://www.listcorp.com/asx/REA	163204000	125.06	0	0	Communication Services	
25	ASX:SHL	Sonic Health	https://www.listcorp.com/asx/SHL	160879000	34.27	0	0	Health Care	
26	ASX:WTC	WiseTech	https://www.listcorp.com/asx/WTC	159865000	50.1	0	0	Information Technology	
27	ASX:RHC	Ramsay HealthCare	https://www.listcorp.com/asx/RHC	159851000	70.2	0	0	Health Care	
28	ASX:BXB	Brambles	https://www.listcorp.com/asx/BXB	157955000	11.45	0	0	Industrials	
29	ASX:JHX	James Hardie	https://www.listcorp.com/asx/JHX	154119000	35.2	0	0	Materials	

Figure 2. Raw data, as CSV

3. Transformation;

The following data cleaning was performed in order to transform the data into concise and relevant pieces information;

3.1 Removing columns not relevant to our aim including;

From Table 1;

- High price
- Low price
- Volume
- Market cap

From Table 2;

- Link
- Last trade
- change
- % change

3.2 Dropping duplicates to ensure high data validity

3.3 Using functions such as;

- 'Fill na' for table 1 data to ensure our sample size remains (200)
- 'Drop na' for table 2 data

3.4 Modifying data to present aesthetically;

- Removing '\$' from each row
- Removing 'ASX:' prefix from row company titles by renaming content of columns
- Sort data to allow for more efficient tracking

3.5 Ensure datatypes are appropriate for each variable in each column

- E.g. transforming 'Object' -> 'Float'

company_code	company_name	price_29jul22	change	percent_change	one_year_percent_change
0 A200	Betashares Australia 200 ETF	116.21	0.71	0.62	-6.97
1 A2M	The a2 Milk Company Ltd	4.54	0.00	0.00	-22.53
2 AAA	Betashares Australian High Interest Cash ETF	50.08	-0.05	-0.10	-0.02
3 ABP	Abacus Property Group	2.92	-0.01	-0.34	-7.89
4 AFI	Australian Foundation Investment Company Ltd	8.02	0.02	0.25	-4.75

Figure 3. Transformed 'ASX 200' data from Table 1

```

# show data
print(len(transformed_asx_company_df))
transformed_asx_company_df.head()
✓ 0.2s
2217

```

	company_code	company_name	market_cap	sector
0	14D	1414 Degrees Limited (ASX:14D)	21208500	Industrials
1	1AD	AdAlta Limited (ASX:1AD)	15709200	Health Care
2	1AE	Aurora Energy Metals Limited (ASX:1AE)	26270400	Materials
3	1AG	Alterra Limited (ASX:1AG)	11128800	Consumer Staples
4	1MC	Morella Corporation Limited (ASX:1MC)	93171900	Materials

Figure 4. Transformed 'ASX Company' data from Table 2

4. Loading;

4.1 Create SQL database 'asx_db' in PostgreSQL

- Create two tables to insert values into
 - o Asx 200 table
 - o Asx_companies table

4.2 Create connection in jupyter notebook to PostgreSQL to allow loading of data

4.3 Load data frames created into the sql database

```

1 select * from asx_companies

```

Data output Messages Notifications

	company_code [PK] character varying (10)	company_name character varying (100)	market_cap numeric	sector character varying (100)
1	14D	1414 Degrees Limited (ASX:14D)	21208500	Industrials
2	1AD	AdAlta Limited (ASX:1AD)	15709200	Health Care
3	1AE	Aurora Energy Metals Limited (ASX:1AE)	26270400	Materials
4	1AG	Alterra Limited (ASX:1AG)	11128800	Consumer Staples
5	1MC	Morella Corporation Limited (ASX:1MC)	93171900	Materials
6	1ST	1st Group Limited (ASX:1ST)	9017220	Health Care
7	1VG	Victory Goldfields Limited (ASX:1VG)	7114760	Materials
8	29M	29Metals Limited (ASX:29M)	714759000	Materials

Total rows: 1000 of 2217 Query complete 00:00:00.205 Ln 1, Col 28

Figure 5. Asx_companies table

```

1 select * from asx_200;

```

Data output Messages Notifications

	company_code [PK] character varying (10)	company_name character varying (100)	price_29Jul22 numeric	change numeric	percent_change numeric	one_year_percent_change numeric
1	A200	Betashares Australia 200 ETF	116.21	0.71	0.62	-6.97
2	A2M	The a2 Milk Company Ltd	4.54	0.0	0.0	-22.53
3	AAA	Betashares Australian High Interest Cash ETF	50.08	-0.05	-0.1	-0.02
4	ABP	Abacus Property Group	2.92	-0.01	-0.34	-7.89
5	AFI	Australian Foundation Investment Company Ltd	8.02	0.02	0.25	-4.75
6	AGL	AGL Energy Ltd	8.45	0.07	0.84	16.87
7	AIA	Auckland International Airport Ltd	6.71	0.04	0.6	-1.32
8	AIZ	Air New Zealand Ltd	0.565	0.02	3.67	-60.21

Total rows: 200 of 200 Query complete 00:00:00.124 Ln 1, Col 23

Figure 6. Asx_200 table

4.4 Joining tables on ‘company_code’

4.5 Create a ‘view’ for the joined table ‘asx_200_data’ to prevent manual joining at each moment that joint data is to be analysed

The screenshot shows a database interface with a query window containing the following SQL command:

```
1 select * from asx_200_data
```

The results pane displays a table with 8 rows of data, representing the joined table 'asx_200_data'. The columns are:

company_code	company_name	price_29jul22	percent_change	one_year_percent_change	sector	market_cap
A200	Betashares Australia 200 ETF	116.21	0.62	-6.97	Financials	2170880000
A2M	Tha2 Milk Company Ltd	4.54	0.0	-22.53	Consumer Staples	3272090000
ABP	Abacus Property Group	2.92	-0.34	-7.89	Real Estate	2570200000
AFI	Australian Foundation Investment Company Ltd	8.02	0.25	-4.75	Financials	9839250000
AGL	AGL Energy Ltd	8.45	0.84	16.87	Utilities	5570350000
AIA	Auckland International Airport Ltd	6.71	0.6	-1.32	Industrials	9674660000
AIZ	Air New Zealand Ltd	0.565	3.67	-60.21	Industrials	1835810000
AKE	Allkem Ltd	11.79	4.52	42.91	Materials	7090760000

Total rows: 200 of 200 Query complete 00:00:00.135 Ln 1, Col 27

Figure 7. Asx_200_data , joint table

5. Schema;

```
project_no_2.sql
1 -- create asx_200 table
2 CREATE TABLE asx_200 (
3   company_code VARCHAR(10) NOT NULL,
4   company_name VARCHAR(100) NOT NULL,
5   price_29jul22 DEC,
6   change DEC,
7   percent_change DEC,
8   one_year_percent_change DEC,
9   PRIMARY KEY (company_Code)
10 );
11
12 -- create asx_companies table
13 CREATE TABLE asx_companies (
14   company_code VARCHAR(10) NOT NULL,
15   company_name VARCHAR(100) NOT NULL,
16   market_cap DEC,
17   sector VARCHAR(100) NOT NULL,
18   PRIMARY KEY (company_code)
19 );
20
21 -- Join tables on company_code
22 SELECT a200.company_code, a200.company_name, a200.price_29jul22, a200.percent_change, a200.one_year_percent_change, a.sector , a.market_cap
23 FROM asx_200 a200
24 LEFT JOIN asx_companies a
25 ON a200.company_code = a.company_code;
26
27 -- create view
28 CREATE VIEW asx_200_data AS
29 SELECT a200.company_code, a200.company_name, a200.price_29jul22, a200.percent_change, a200.one_year_percent_change, a.sector , a.market_cap
30 FROM asx_200 a200
31 LEFT JOIN asx_companies a
32 ON a200.company_code = a.company_code;
```

Figure 8. Schema

6. ERD Diagram:

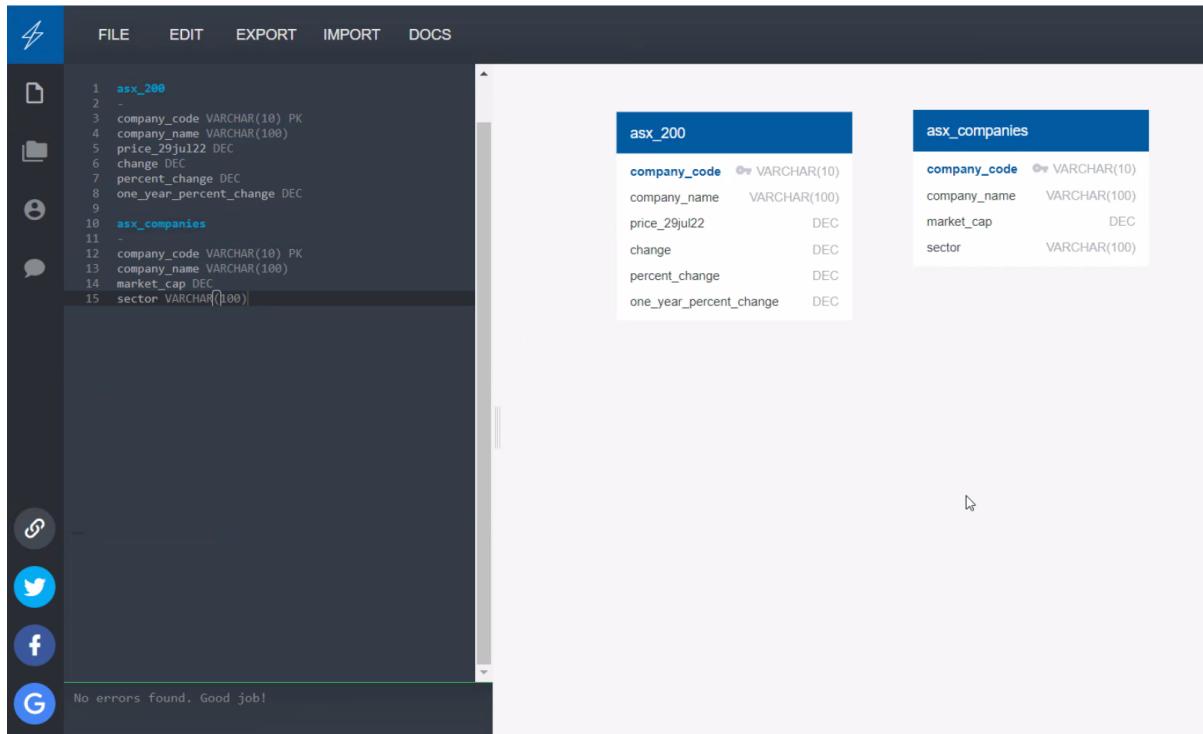


Figure 9. ERD diagram