Database of Quarterly Crude Oil Prices and Estimated Percent Changes Compared with Quarterly Australian Consumer Price Index data.

ETL Project

Group 5

By:

Masi Mapiye, Angadjeet Sanghera, Musah Abdulai

March 19, 2022

Introduction

We were tasked with finding two data sources and performing the ETL process. We extracted our CPI data from the Australian Bureau of Statistics (ABS) website, and the Crude Oil data from the Trading Economics website before transforming it by doing the necessary cleaning with the help of Pandas. We then loaded it into an SQL database in order to join the two data sets in preparation for analysis.

We aimed at creating a database of quarterly crude oil prices, calculating the percent changes, and comparing this with ABS-provided quarterly Australian Consumer Price Index data. The only changes we made to the ABS data was to select the appropriate CPI grouping, and to limit it to the years 1983 to 2021 in order to have parity with the Crude Oil data range.

Below we summarise the steps one would follow to reproduce our findings.

Data Extraction

- Australian Bureau of Statistics (ABS) – xlsx data file.
    - Consumer Price Index Australia – we selected the "All group CPI Australia".
    - Utilise data only from 1983 to 2021.
- tradingeconomics.com/commodity/crude-oil – csv data file
    - Convert from monthly to quarterly,
    - Calculate the percentage change for each quarter
- Data was cleaned in Jupyter Notebook files, with the use of Pandas

Websites

1. https://www.abs.gov.au/statistics/economy/price-indexes-and-inflation/consumer-price-index-australia/dec-2021
2. https://tradingeconomics.com/commodity/crude-oil

Data Transformation

Step 1: CPI & Crude Oil Price Data – Data Cleaning

- We removed all of the State columns from data CPI and only maintained the "All group CPI Australia" data,
- Using Panda functions, we cleaned our datasets and removed all unwanted rows so that our start date is 1983-03-01,
- Unwanted columns were removed and some remaining columns were renamed for simplification,
- Formatted date column and set date as the index.

	Unnamed: 0	Index Numbers ; All groups CPI ; Sydney ;	Index Numbers ; All groups CPI ; Melbourne ;	Index Numbers ; All groups CPI ; Brisbane ;	Index Numbers ; All groups CPI ; Adelaide ;	Index Numbers ; All groups CPI ; Perth ;	Index Numbers ; All groups CPI ; Hobart ;	Index Numbers ; All groups CPI ; Darwin ;	Index Numbers ; All groups CPI ; Canberra ;	Index Numbers ; All groups CPI ; Australia ;	...	Percentage Change from Corresponding Quarter of Previous Year ; All groups CPI ; Australia ;	Percentage Change from Previous Period ; All groups CPI ; Sydney ;	Percentage Change from Previous Period ; All groups CPI ; Melbourne ;	F
147	1983-03-01 00:00:00	34.4	34.4	33.9	33.9	34.2	35	37.1	34.9	34.3	...	11.4	2.1	2.1	
148	1983-06-01 00:00:00	35.1	35.4	34.3	34.8	34.8	35.6	37.8	35.5	35	...	11.1	2	2.9	
149	1983-09-01 00:00:00	35.5	35.9	35.1	35.3	35.8	36.1	38.3	36	35.6	...	9.2	1.1	1.4	
150	1983-12-01 00:00:00	36.2	37	35.8	36	36.5	36.9	38.8	36.9	36.5	...	8.6	2	3.1	
151	1984-03-01 00:00:00	36.1	36.8	35.9	36	36.2	37	39	36.9	36.3	...	5.8	-0.3	-0.5	

Figure 1: Unnecessary Columns, CPI

	date	price
0	1983-03-30T00:00:00	29.27
1	1983-04-04T00:00:00	30.63
2	1983-05-02T00:00:00	30.25
3	1983-06-01T00:00:00	31.38
4	1983-07-01T00:00:00	32.00

```
# Check column data types to see if datetime functions can be performed
crudeOil_df.dtypes
```

```
date      object
price    float64
dtype: object
```

```
# Change date column's datatype
crudeOil_df['date'] = pd.to_datetime(crudeOil_df['date'])
crudeOil_df.head()
```

	date	price
0	1983-03-30	29.27
1	1983-04-04	30.63
2	1983-05-02	30.25
3	1983-06-01	31.38
4	1983-07-01	32.00

Figure 2: Date Cleanup, Crude Oil

Step 2: CPI & Crude Oil Price Data – Restructure

- Grouped the monthly Crude Oil price data into quarterly data,
- With the Crude Oil quarterly data, we calculated the percentage changes, utilising the Pandas *.pct_change* function.
- Generated a new column in both datasets, called *report_month*, for parity,
- Reordered columns prior to csv export.

```
# Select months of Sep, Dec, Mar & Jun only (Quarterly Data)
month_list = [9, 12, 3, 6]
qrtCrudeOil_df = crudeOil_df[crudeOil_df['date'].dt.month.isin(month_list)]
qrtCrudeOil_df
```

	date	price
0	1983-03-30	29.27
3	1983-06-01	31.38
6	1983-09-01	30.36
9	1983-12-01	29.60
12	1984-03-01	30.85

Figure 3: Monthly to Quarterly, Crude Oil

```
clean_cpi_data_df['report_month'] = pd.to_datetime(clean_cpi_data_df['Date']).dt.to_period('M')
clean_cpi_data_df.head()
```

	Date	CPI	Percent_Change_CPI	report_month
147	1983-03-01	34.3	2.1	1983-03
148	1983-06-01	35	2	1983-06
149	1983-09-01	35.6	1.7	1983-09
150	1983-12-01	36.5	2.5	1983-12
151	1984-03-01	36.3	-0.5	1984-03

Figure 4: Generation of New report_month Column, CPI

```
# Change column order
qrtCrudeOil_df = qrtCrudeOil_df[['date', 'report_month', 'price', 'percent_change']]
qrtCrudeOil_df.head()
```

	date	report_month	price	percent_change
0	1983-03-30	1983-03	29.27	NaN
3	1983-06-01	1983-06	31.38	7.208746
6	1983-09-01	1983-09	30.36	-3.250478
9	1983-12-01	1983-12	29.60	-2.503294
12	1984-03-01	1984-03	30.85	4.222973

Figure 5: Reordering of Columns, Crude Oil

Data Loading – PostgreSQL

- We created a database with two tables, *cpi_data* & *crude_oil,* in PostgreSQL and set *report_month* in the *cpi_data* table as the primary key.
- A left join was done on the tables on the report_month columns

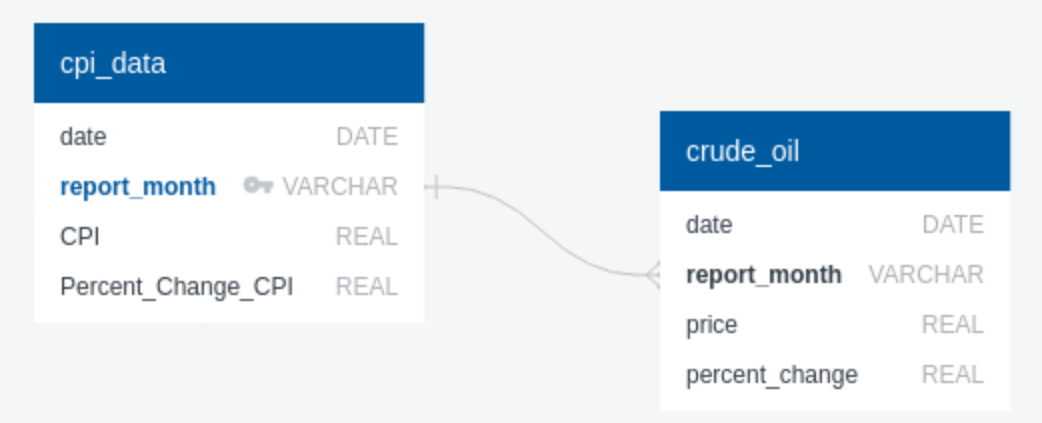

Figure 6: ERD plot

```
5  -- Join tables in order to make comparison
6  SELECT cd.report_month, cd.cpi, cd.percent_change_cpi, co.price, co.percent_change
7      FROM cpi_data as cd
8      LEFT JOIN crude_oil as co
9          ON cd.report_month = co.report_month
```

Data Output | Explain | Messages | Notifications

	report_month character varying	cpi real	percent_change_cpi real	price real	percent_change real
1	1983-03	34.3	2.1	29.27	[null]
2	1983-06	35	2	31.38	7.20875
3	1983-09	35.6	1.7	30.36	-3.25048
4	1983-12	36.5	2.5	29.6	-2.50329

Figure 7: SQL Query for Analysis