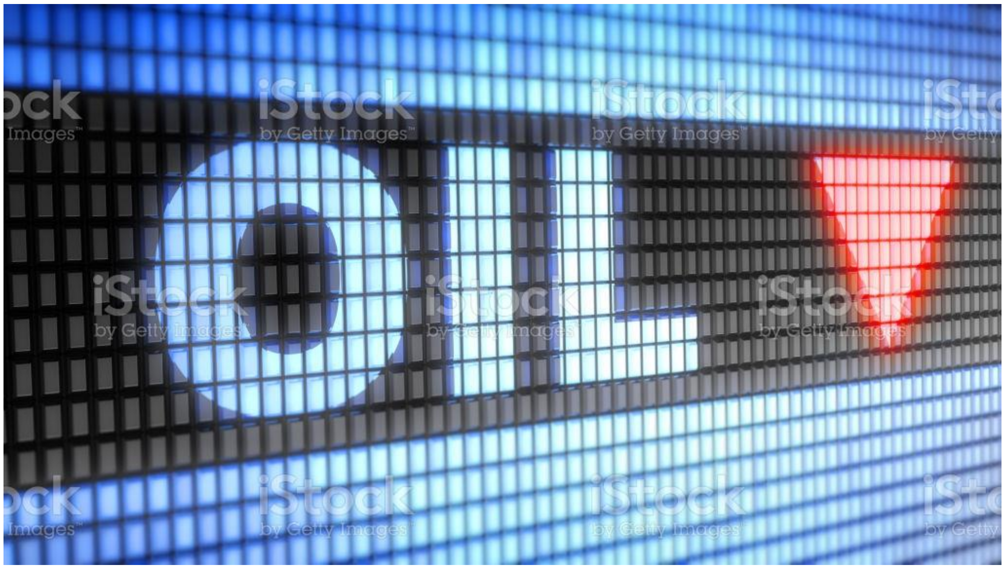
ETL Final Report

ETL Project

Nima Karimi and Kirstie McCown

Week 12: University of Western Australia Data Analytics Bootcamp



Project Title:	ETL Project
Class Instructor:	Daniel Kasatchkow
TA:	Hazar Ayaz
Project Due Date:	Saturday 29th August 2020
Date of Report:	Saturday 29th August 2020
Reporting Period:	Tuesday 25th August 2020 – Saturday 29th August 2020

Table of Contents

Annexures

Summary

Currently, the price of oil is ever changing, and sometimes for unknown cause.
We are carrying out this project to identify if there is any correlation between US Oil Pipeline Accidents and the Crude Oil Price around the same period (2010 to 2016).

Data Sources

We utilised two sets of data from Kaggle.com, one was in cvs. format and the other .xlsx:

Oil Prices
https://www.kaggle.com/rockbottom73/crude-oil-prices

US Oil Pipeline Accidents
https://www.kaggle.com/usdot/pipeline-accidents

Data Transformation

Our overall data transformation we wanted to look at the following elements, we will detail these further below for each individual data set.

- Remove any unnecessary columns
- Drop all accidents not related to crude oil
- Drop select items which are N/A or have the value of NaN
- Split accident date/time field to show only dates

Oil Prices

- Read in xlsx to Panadas DataFrame to enable visualisation and cleaning
- Rename columns so they are easier to work with
- Drop everything which is N/A
- Drop all rows that are not the same dates as what is in the accidents DataFrame

US Oil Pipeline Accidents

- Read in csv to Panadas DataFrame to enable visualisation and cleaning
- Remove any unnecessary columns and rename columns so they are easier to work with
- Look at what non null values are in the DataFrame to see if values need to be removed
- Drop all accidents not related to crude oil
- Drop everything which is N/A in the following columns: city, facility_name, country, shutd own

- Split the date/time column keeping the date in a newly created column, whist dropping the original date/time column
- Change the format of the date so that both DataFrame dates match format

Database

For our project we utilised a Postgres SQL Database, as part of our ETL process we conducted the following steps:

- Create a new Postgres Database called “oil_db”
- Create two table schema’s called “cleaned_oil” and “cleaned_accidents”
- Connect to Postgres database via our Jupyter Notebook (.ipynb file)
- Check to ensure tables are available in Postgres database and able to be connected with via our Jupyter Notebook
- Load panda's DataFrame to postgres sql tables

See: Annexure 1 and Annexure 2

Database Tables

Table Name	Number of Columns
cleaned_oil	2
cleaned_accidents	17

The above two tables were joined to create one table for further analysis.

See: Annexure 3, Annexure 4 and Annexure 5

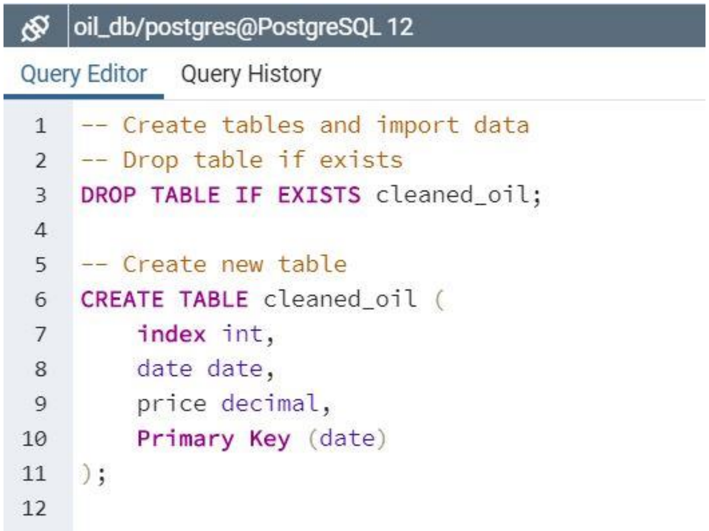
Project Conclusion

We feel that our ETL process has prepared the two datasets adequately in order to be able to further analyse and identify if there is any correlation between US Oil Pipeline Accidents and the fluctuation of Crude Oil Prices around the same period of time.

Our dataset has been prepared into two separate tables, which have then been joined to allow for further investigation and manipulation, while maintaining the integrity of each individual data set as a whole.

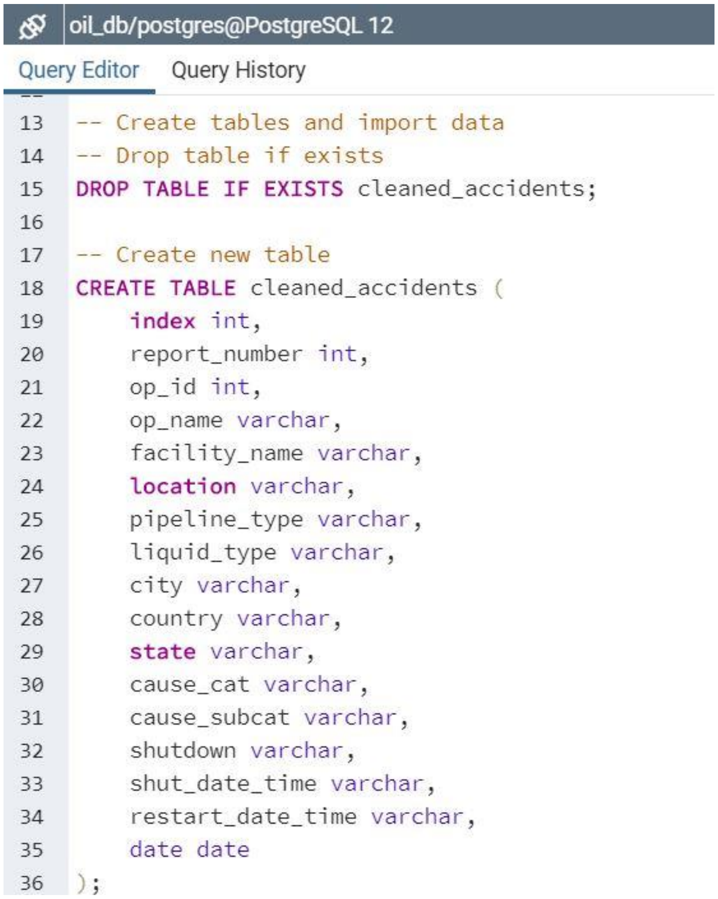
Annexures/ Figures

Annexure 1 – Cleaned Oil Table Schema



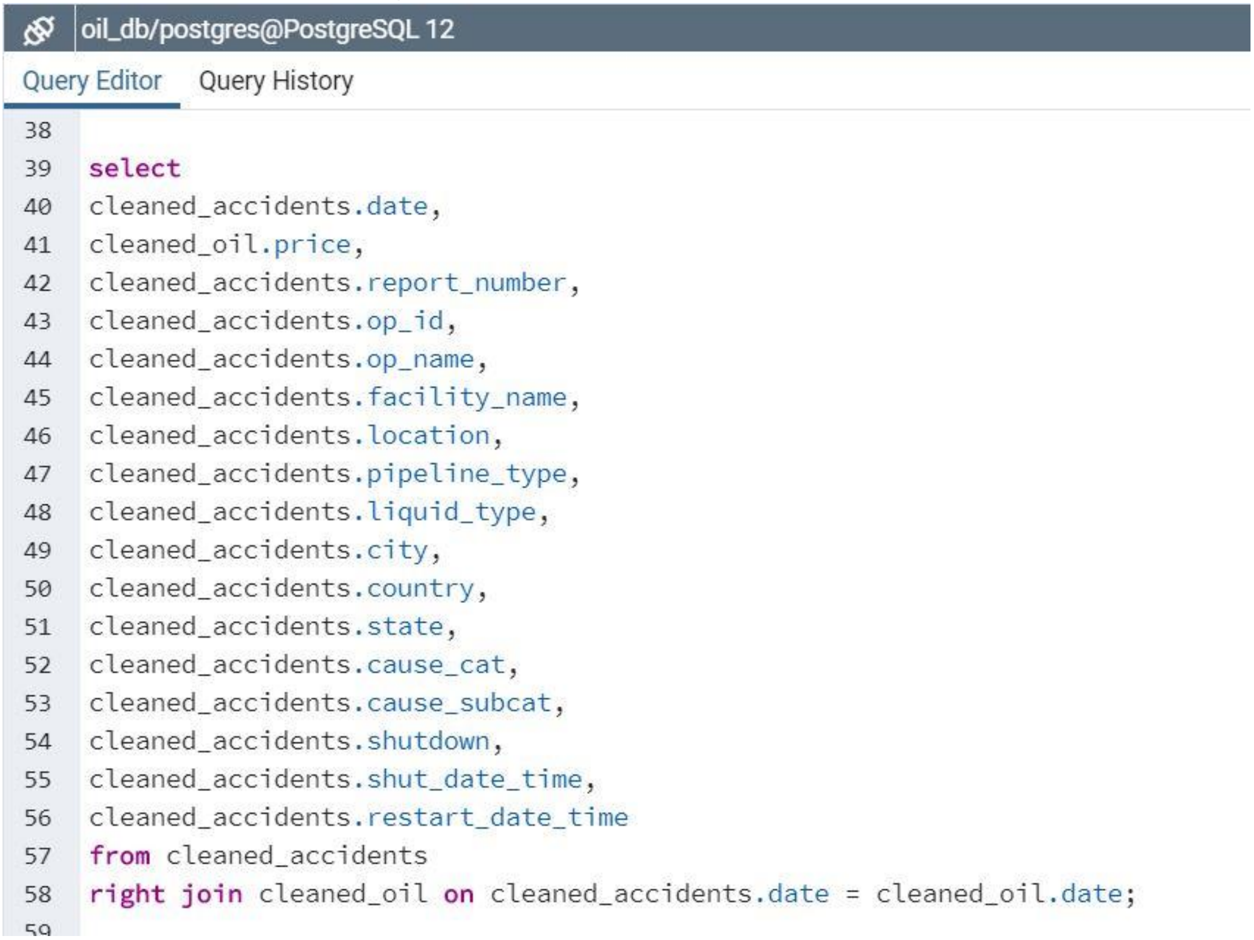
```sql
-- Create tables and import data
-- Drop table if exists
DROP TABLE IF EXISTS cleaned_oil;

-- Create new table
CREATE TABLE cleaned_oil (
    index int,
    date date,
    price decimal,
    Primary Key (date)
);

```

Annexure 2 – Cleaned Accidents Table Schema



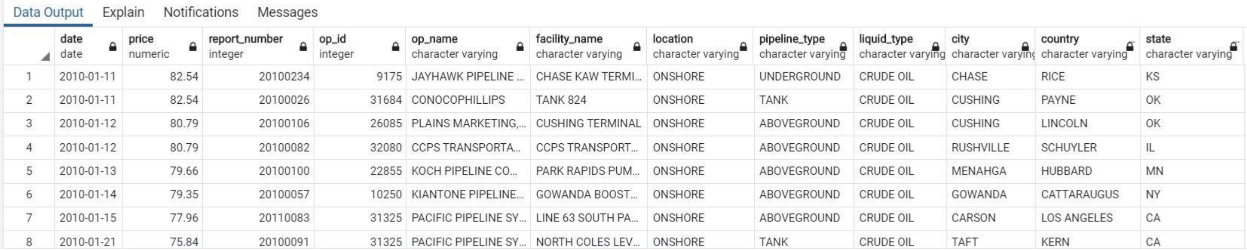
```sql
-- Create tables and import data
-- Drop table if exists
DROP TABLE IF EXISTS cleaned_accidents;

-- Create new table
CREATE TABLE cleaned_accidents (
    index int,
    report_number int,
    op_id int,
    op_name varchar,
    facility_name varchar,
    location varchar,
    pipeline_type varchar,
    liquid_type varchar,
    city varchar,
    country varchar,
    state varchar,
    cause_cat varchar,
    cause_subcat varchar,
    shutdown varchar,
    shut_date_time varchar,
    restart_date_time varchar,
    date date
);
```

Annexure 3 – Table Join Query



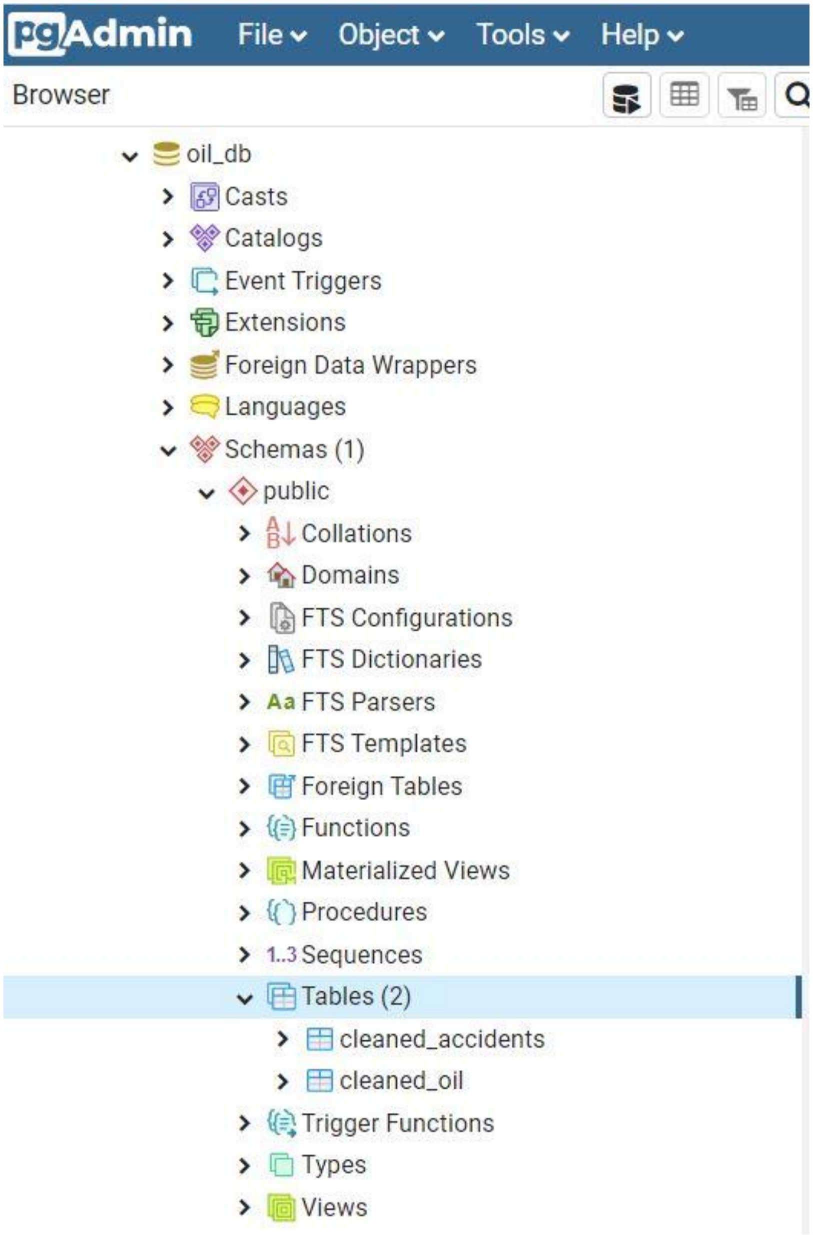
```
select
cleaned_accidents.date,
cleaned_oil.price,
cleaned_accidents.report_number,
cleaned_accidents.op_id,
cleaned_accidents.op_name,
cleaned_accidents.facility_name,
cleaned_accidents.location,
cleaned_accidents.pipeline_type,
cleaned_accidents.liquid_type,
cleaned_accidents.city,
cleaned_accidents.country,
cleaned_accidents.state,
cleaned_accidents.cause_cat,
cleaned_accidents.cause_subcat,
cleaned_accidents.shutdown,
cleaned_accidents.shut_date_time,
cleaned_accidents.restart_date_time
from cleaned_accidents
right join cleaned_oil on cleaned_accidents.date = cleaned_oil.date;
```

Annexure 4 - Joined Tables from Database



	date	price	report_number	op_id	op_name	facility_name	location	pipeline_type	liquid_type	city	country	state
	date	numeric	integer	integer	character varying	character varying	character varying	character varying	character varying	character varying	character varying	character varying
1	2010-01-11	82.54	20100234	9175	JAYHAWK PIPELINE ...	CHASE KAW TERMI...	ONSHORE	UNDERGROUND	CRUDE OIL	CHASE	RICE	KS
2	2010-01-11	82.54	20100026	31684	CONOCOPHILLIPS	TANK 824	ONSHORE	TANK	CRUDE OIL	CUSHING	PAYNE	OK
3	2010-01-12	80.79	20100106	26085	PLAINS MARKETING,...	CUSHING TERMINAL	ONSHORE	ABOVEGROUND	CRUDE OIL	CUSHING	LINCOLN	OK
4	2010-01-12	80.79	20100082	32080	CCPS TRANSPORTA...	CCPS TRANSPORT...	ONSHORE	ABOVEGROUND	CRUDE OIL	RUSHVILLE	SCHUYLER	IL
5	2010-01-13	79.66	20100100	22855	KOCH PIPELINE CO...	PARK RAPIDS PUM...	ONSHORE	ABOVEGROUND	CRUDE OIL	MENAHGA	HUBBARD	MN
6	2010-01-14	79.35	20100057	10250	KIANTONE PIPELINE...	GOWANDA BOOST...	ONSHORE	ABOVEGROUND	CRUDE OIL	GOWANDA	CATTARAUGUS	NY
7	2010-01-15	77.96	20110083	31325	PACIFIC PIPELINE SY...	LINE 63 SOUTH PA...	ONSHORE	ABOVEGROUND	CRUDE OIL	CARSON	LOS ANGELES	CA
8	2010-01-21	75.84	20100091	31325	PACIFIC PIPELINE SY...	NORTH COLES LEV...	ONSHORE	TANK	CRUDE OIL	TAFT	KERN	CA

Annexure 5 - Screenshot of Database and Schemas



- oil_db
  - Casts
  - Catalogs
  - Event Triggers
  - Extensions
  - Foreign Data Wrappers
  - Languages
  - Schemas (1)
    - public
      - Collations
      - Domains
      - FTS Configurations
      - FTS Dictionaries
      - FTS Parsers
      - FTS Templates
      - Foreign Tables
      - Functions
      - Materialized Views
      - Procedures
      - Sequences
      - Tables (2)
        - cleaned_accidents
        - cleaned_oil
      - Trigger Functions
      - Types
      - Views