Abstract

- Outline

1. Data preparation: based on the different sources of data fetched, the format is various. Sort those data sets into the same column names and formats;
2. Combine data sets wti_spot_price.csv and OXY_stock.csv, and then find whether there exists a (strong) relationship between them, via Simple Linear Regression Model, correlation check, Moving Average Model. Fetch two more stock data,tickers as 'CVX' and 'ALB', for further exploration.
3. Includes data set car_sales.csv for regression model analysis. Do the data preparation again to match the car_sales.csv timeslot and format. Apply linear and multiple regression model. Set monthly car_sales as Y, OXY price and WTI price as two predictors. Fit the model to check whether the two predictors perform well on the Y (car sales).
4. Web_scrape for one more data sets- 2017 wti_spot_price - to create a multi-index table, in order to compare the sales of the same car model in different years.

- Conclusion

1. The relationship between OXY and WTI is weak
2. Neither linear regression model nor multiple regression model perform well on predicting car sales.
3. In general, electric vehicles sales in 2018 is higher than that of in 2017.

Data Source

- API Key

1. OXY_stock.csv (from NASDAQ)

api_key='wCgHNXA4XE-naxGvQt6X'

URL:https://data.nasdaq.com/api/v3/datasets/WIKI/OXY/data.json?start_date=2015-01-01&end_date=2018-03-27&order=asc&column_index=4&api_key=wCgHNXA4XE-naxGvQt6X

Click on the URL provided above to get private api_key, which should be included in the OXY_Stock_Price.py file for web-scraping. Using the requests library and .json(), create a data set called OXY_stock.csv. The first 5 rows is shown below.

	Date	Close
0	2015-01-02	80.65
1	2015-01-05	77.66
2	2015-01-06	77.00
3	2015-01-07	77.01
4	2015-01-08	77.69

2. wti_spot_price.csv (from EIA)

api_key ='vT2MbTCiEVzdJdsxQe1VVKxVBsmwrgbJroVhrDxc'

URL:https://api.eia.gov/series/?api_key=vT2MbTCiEVzdJdsxQe1VVKxVBsmwrgbJroVhrDxc&series_id=PET.RWTC.W

Click on the URL provided above to get private api_key, which should be included in the wti_spot_price.py file for web-scraping. Using the requests library and .json(), set the startDate = '2015-01-01' and endDate = '2019-01-01' to extract a data set called wti_spot_price.csv. The first 5 rows is shown below.

	Date	Spot Price
0	2015-01-02	53.44
1	2015-01-09	48.77
2	2015-01-16	47.07
3	2015-01-23	46.46
4	2015-01-30	45.32

Tips: Combine the two date sets on the key 'Date', using pd.merge(), and then rename the columns, created a dataframe called oxy_wti. The first 5 rows is shown below.

	Date	OXY	WTI
0	2015-01-02	80.65	53.44
1	2015-01-09	77.54	48.77
2	2015-01-16	78.06	47.07
3	2015-01-23	78.85	46.46
4	2015-01-30	80.00	45.32

- Web Scraper

1. car_sales.csv

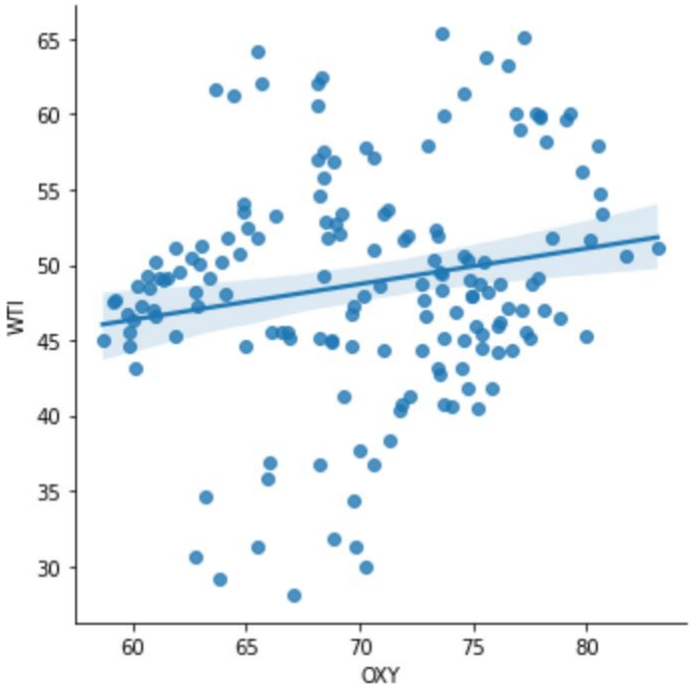
URL: https://insideevs.com/news/344007/monthly-plug-in-ev-sales-scorecard-historical-charts/

In EV_sales. Py file, using requests and beautifulsoup library to prettify the html text, and the find the content in <td>. Use the for loop get the 2018 data and create a data set called car_sales.csv. The first 5 rows is shown below.

	BRAND	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
0	Tesla Model 3*	1875	2485	3820	3750	6000	5902	14250	17800	22250	17750	18650	25250
1	Toyota Prius Prime	1496	2050	2922	2626	2924	2237	1984	2071	2213	2001	2312	2759
2	Tesla Model X*	700	975	2825	1025	1450	2550	1325	2750	3975	1225	3200	4100
3	Tesla Model S*	800	1125	3375	1250	1520	2750	1200	2625	3750	1350	2750	3250
4	Honda Clarity PHEV*	604	911	1131	1129	1639	1495	1542	1462	1997	2025	1897	2770

How Analysis works

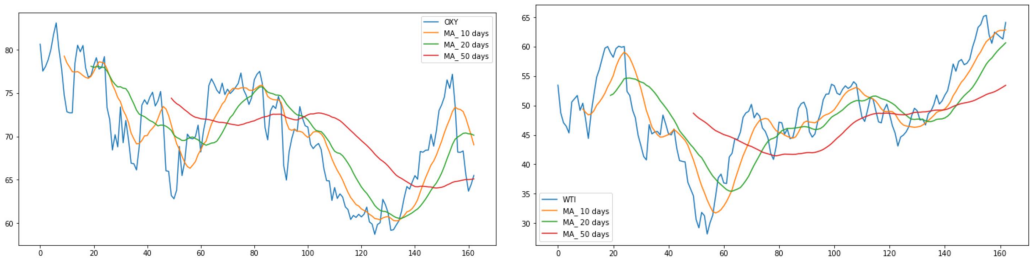
- After data preparation, we get a oxy_wti data frame, which records the price for both OXY and WTI in the same day. Let's do regression model to check their relationship. We get the plot below.



Based on the plot, it is hard to conclude there exists a strong relationship between OXY and WTI. Besides the data visualization, numerical result may be more convincing. Check the **correlation coefficient** by Pearson method, which normalized the value in the range [-1,1]. The higher absolute is , the stronger relationship exist. In this case, 0.188 is a low correlation coefficient value, which testified the assumption above.

	OXY	WTI
OXY	1.000000	0.188519
WTI	0.188519	1.000000

- Step to the **Moving Average Model**, which applied a longer timeslot for values, in order to make them less reliable on daily fluctuations.



The one on the left panel is created by OXY price, with windows= [10,20,50]
The one on the right panel is created by WTI price, with windows = [10,20,50]

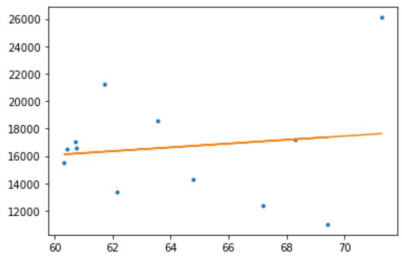
Based on the two plots above, here comes into a new assumption:
with the same moving windows, WTI price seems to be more stable than that of OXY price. Can we assume if we fit a regression model on EV_sales, WTI predictor will perform better than OXY predictor? In other words, because WTI is more stable, it can be easily to track without too much fluctuations, and hence it will be a “good” predictor.

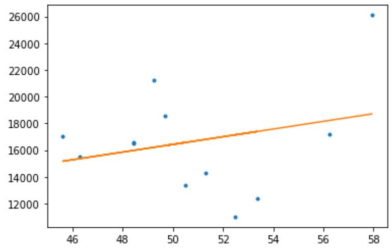
- Try to fit **linear model and multiple regression model**

Set EV_sales 2017 as Y
Set OXY and WTI as predictors X’s
Re-sort data for oxy_wti to match the format of car_2017, which has the columns sales and monthly index. Here we get a new dataframe called reg_df.

	OXY	WTI	SALES
JAN	69.42	52.49	11005
FEB	67.20	53.36	12377
MAR	63.55	49.70	18541
APR	62.13	50.49	13365
MAY	60.73	48.42	16596

Separately apply predictors for linear regression model, we get the linear model below:





The one on the left panel is created by OXY predictor.
The one on the right panel is created by WTI predictor.
Compared the two plots, it seems to confirm our assumption that WTI predictor perform better (even though neither of both fits the model well because of the LOW slope)

Try the multiple regression model, formula is 'SALES~(OXY + WTI)' and get the summary table

OLS Regression Results

Dep. Variable:	SALES	**R-squared:**	0.151
Model:	OLS	**Adj. R-squared:**	-0.038
Method:	Least Squares	**F-statistic:**	0.7991
Date:	Wed, 11 May 2022	**Prob (F-statistic):**	0.479
Time:	01:41:10	**Log-Likelihood:**	-115.26
No. Observations:	12	**AIC:**	236.5
Df Residuals:	9	**BIC:**	238.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.357e+04	2.11e+04	0.642	0.537	-3.42e+04	6.13e+04
OXY	-771.2780	828.669	-0.931	0.376	-2645.857	1103.301
WTI	1035.7962	871.378	1.189	0.265	-935.398	3006.991

In general, R-squared shows how well the regression model fits the observed data.

In this multiple regression model, R-squared = 0.151 means that only 15.1% of the data fit the regression model.
Hence, here comes a conclusion: multiple regression model is NOT well-fitted.

- Create a multi_index table for Electric Vehicle sales

The lack of data makes it difficult to find a fitted model. Instead of fitting model, I decide to compare sales of the same vehicle model 2017 and 2018, monthly.
Here are several examples:

		JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
brand	year												
Tesla Model S*	2017	900	1750	3450	1125	1620	2350	1425	2150	4860	1120	1335	4975
	2018	800	1125	3375	1250	1520	2750	1200	2625	3750	1350	2750	3250
Tesla Model X*	2017	750	800	2750	715	1730	2200	1650	1575	3120	850	1875	3300
	2018	700	975	2825	1025	1450	2550	1325	2750	3975	1225	3200	4100
Toyota Prius Prime	2017	1366	1362	1618	1819	1908	1619	1645	1820	1899	1626	1834	2420
	2018	1496	2050	2922	2626	2924	2237	1984	2071	2213	2001	2312	2759
Nissan LEAF	2017	772	1037	1478	1063	1392	1506	1283	1154	1055	213	175	102
	2018	150	895	1500	1171	1576	1367	1149	1315	1563	1234	1128	1667
Ford Fusion Energi	2017	606	837	1002	905	1000	707	703	762	763	741	731	875
	2018	640	794	782	742	740	604	522	396	480	453	1131	790

This multi_index table provides a much clearer way to compare the same vehicle model.

• Conclusion

1. Even though it seems tightly correlated with each others, there is no enough evidence to show the strong relationship between OXY and WTI.
2. WTI is more stable than OXY price, which can be considered as a better predictor.
3. Multiple regression model with formula 'SALES~(OXY + WTI)' is NOT fitted.
4. Generally, with the same vehicle model, sales of 2018 is higher than 2017 (could be considered as a new energy tendency I guess?)

• Maintainability and Extensibility?

1. Change the time slot for web scraping more data, such as setting startDate = '2010-01-01' and endDate = '2022-01-01'.
2. After including more data, compute machine learning method such as K-means clustering to increase the model accuracy.
3. Incorporate different predictors in regression model.