Sell in May and Go Away ?

Final Data Science Capstone by Fred Etter - July, 2019

Introduction

- The first question I set out to answer was if the weather in NYC has an affect on stock prices. I had read this some years ago in an article or book and have since discovered many links and studies on the subject.
- Upon researching and exploring the data in this field, other questions sur
- faced and have been included those in the presentation that follows.
- The method to answer these questions was a combination of visual representations, statistics, supervised and unsupervised learning.

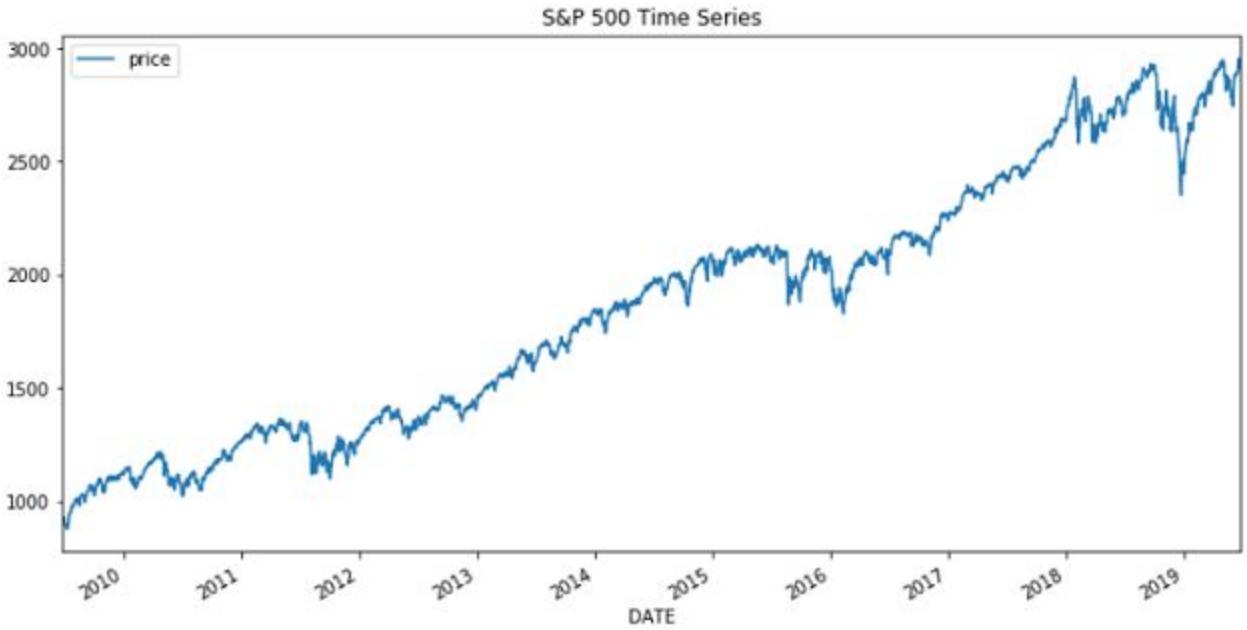
Questions

1. Does the day of the week have any affect on stocks, bonds, gold, oil or btc ?
2. Do “winter” months perform better than “summer” months for the S&P 500 ?
3. Does the 3rd Presidential Year perform better in winter than summer for the S&P 500 ?
4. Does any month typically perform better for the S&P 500 ?
5. Is it useful to build a SL model to predict gold price change based on day of the week ?
6. Does the daily temperature or sunshine in NYC affect stock prices ?
7. Is it useful to build a SL model using all examined data to predict price change of the S&P 500 ?
8. Can unsupervised learning capture relationships among these securities ?

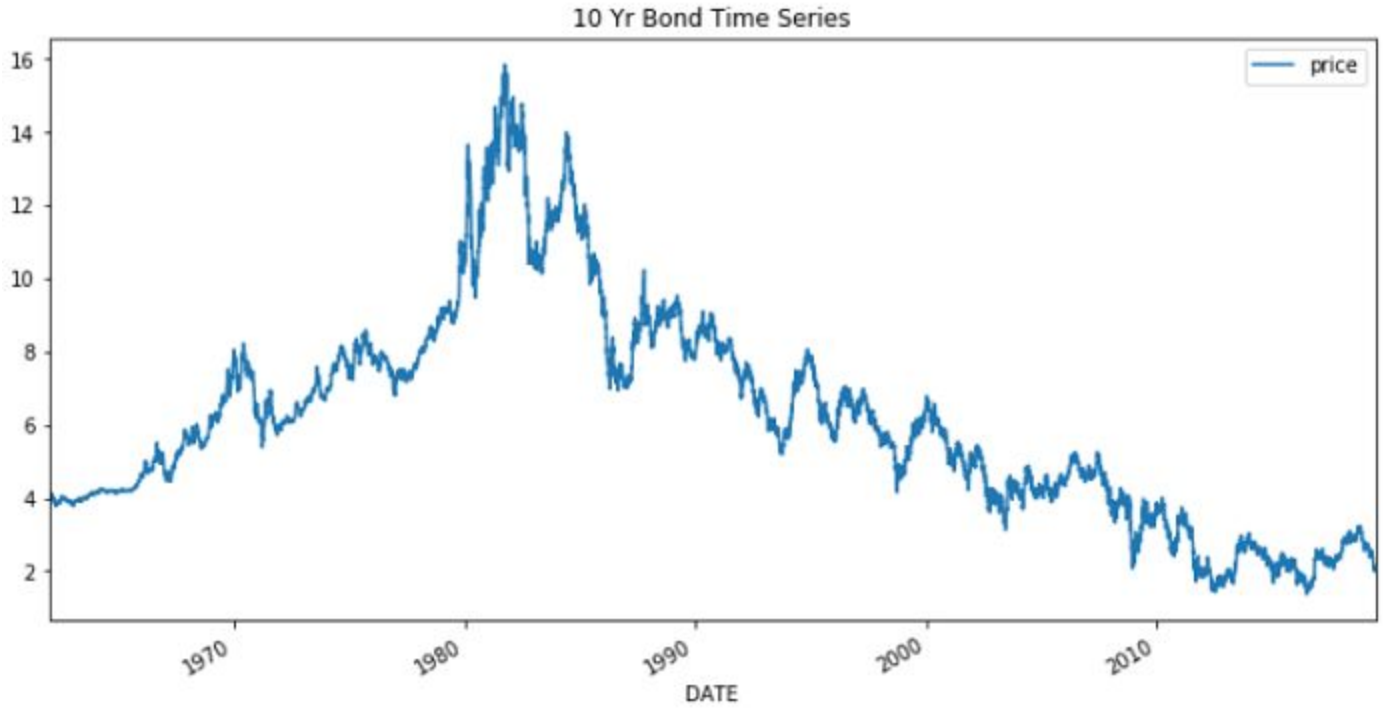
The Data

- FRED - The Federal Reserve Bank of St. Louis - financial data
  - S&P 500 (stocks)
  - 10 Yr Bond
  - Gold
  - Oil
  - Bitcoin
- Meteoblue.com - weather data
  - Daily sunshine in minutes in NYC for past 10 years
  - Daily temperature in NYC for past 10 years

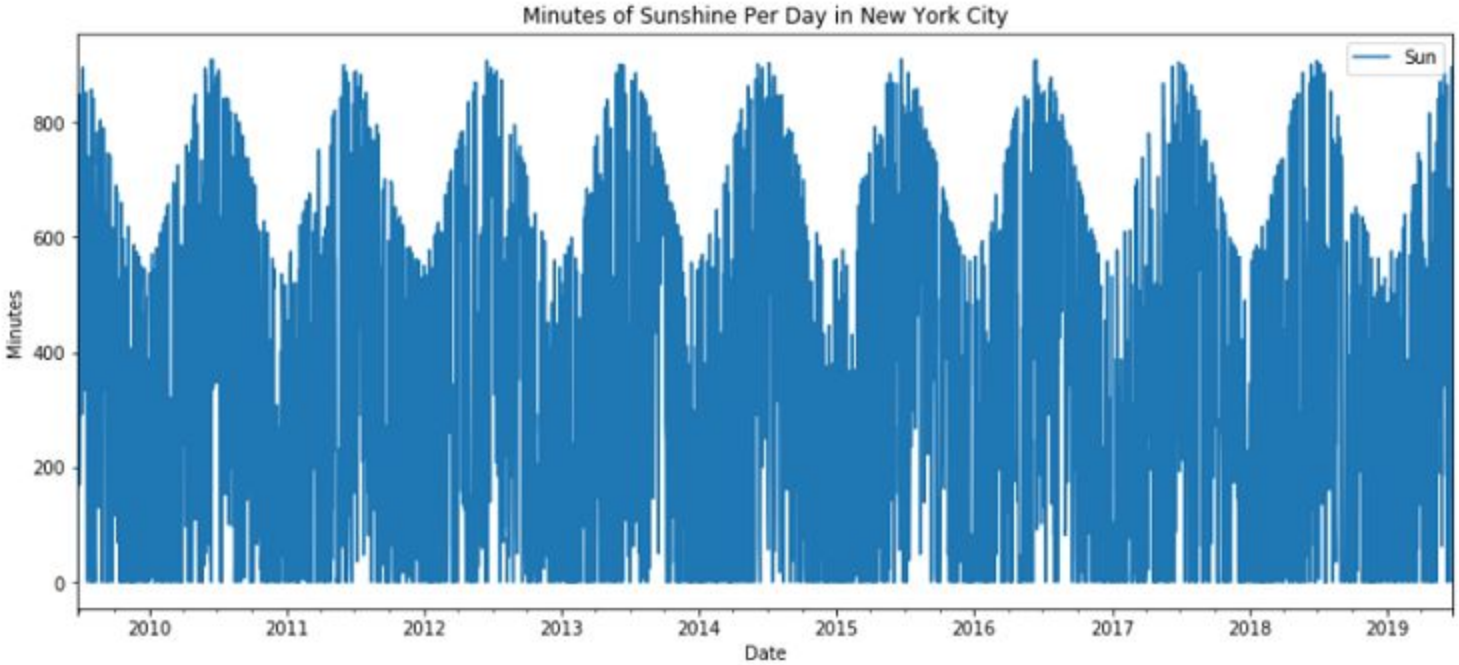
Financial data - S&P 500



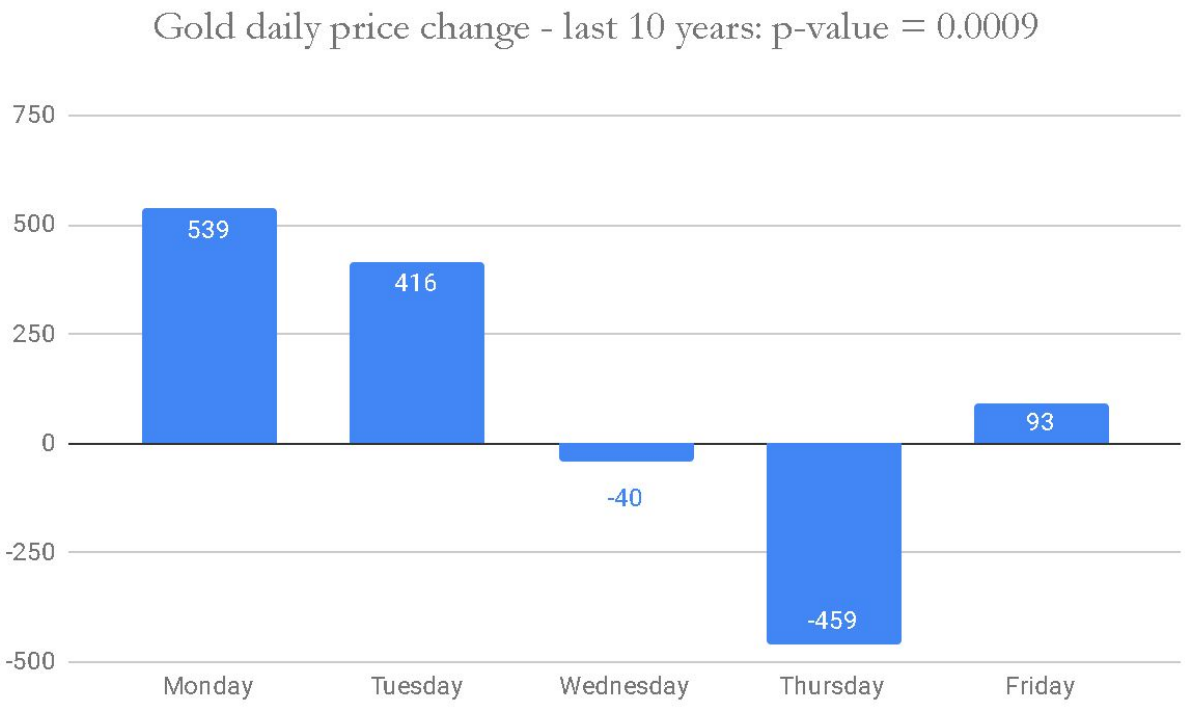
10 Yr Bond: 1962 - 2019



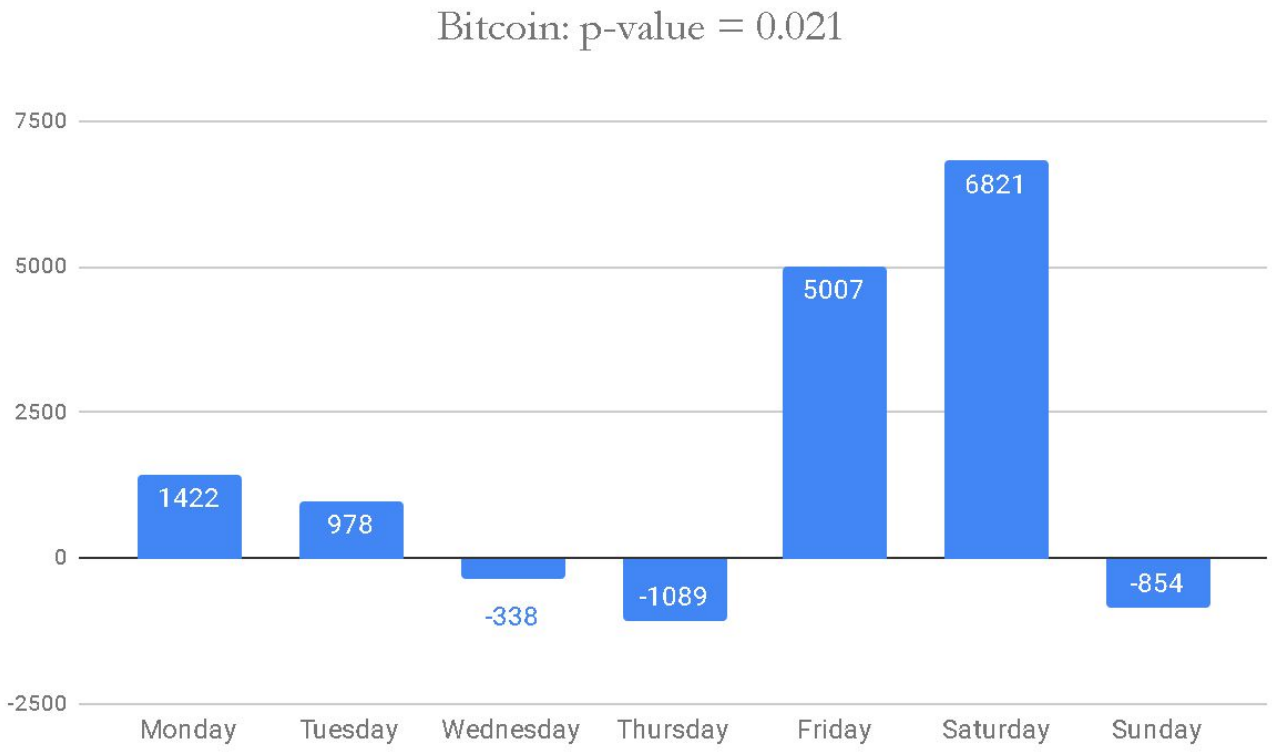
Weather data - sunshine minutes in NYC



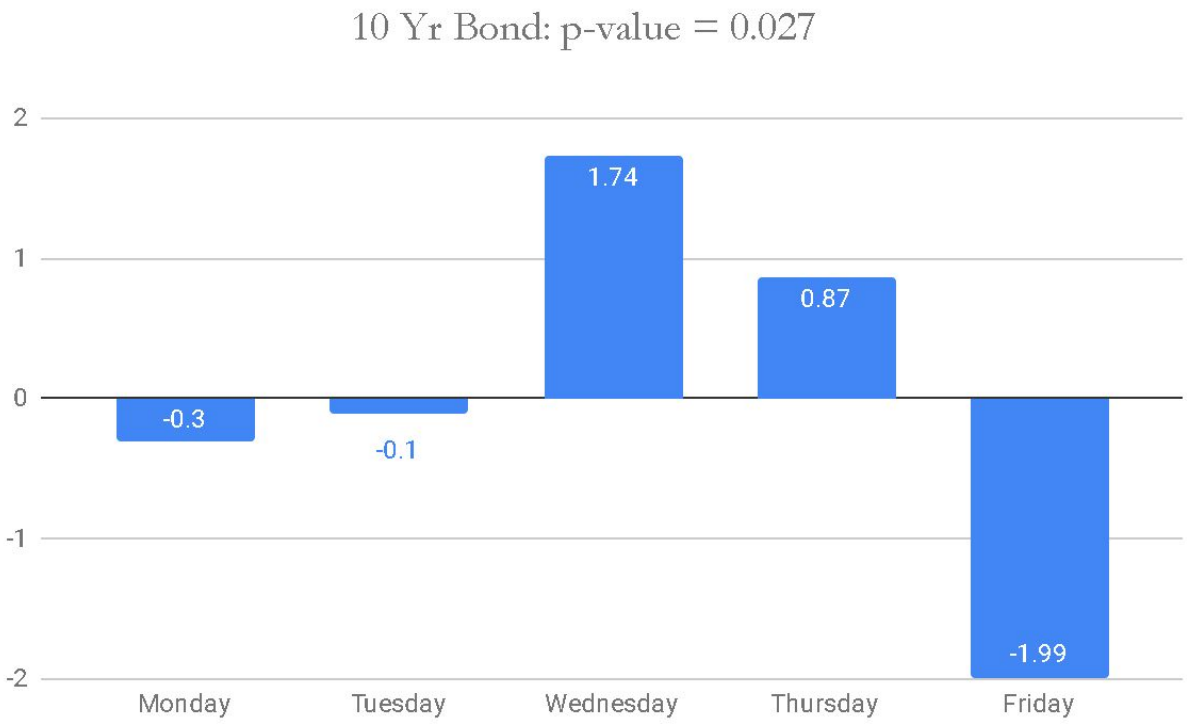
Question 1: Does the day of the week matter ? (anwer: it depends)



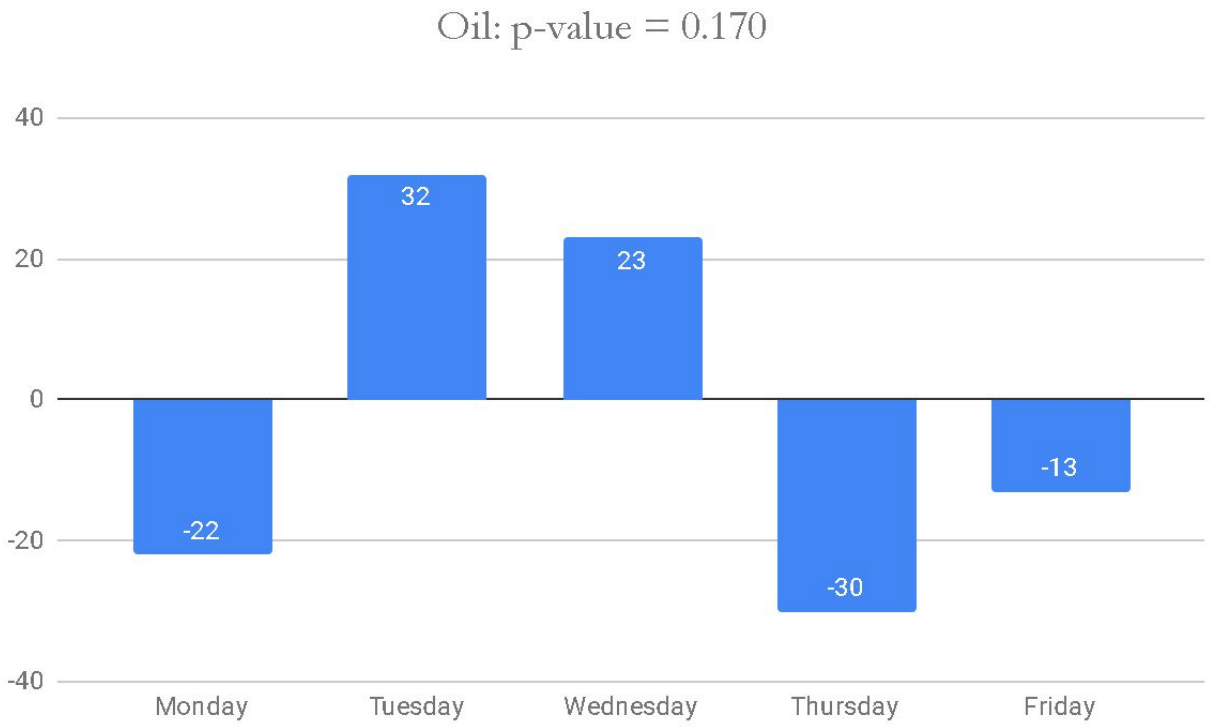
Bitcoin


Bitcoin: p-value = 0.021
7500
5000
2500
0
-2500
1422
978
-338
-1089
5007
6821
-854
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

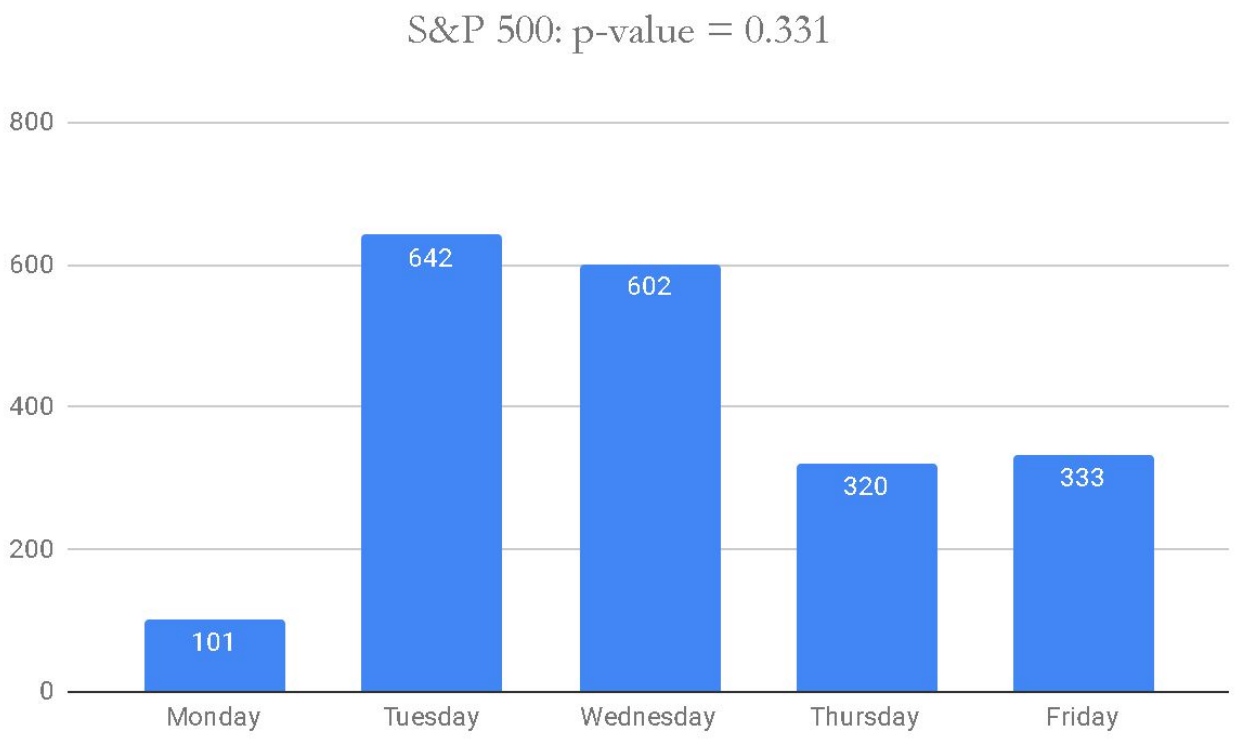
10 Yr Bond


10 Yr Bond: p-value = 0.027
2
1
0
-1
-2
-0.3
-0.1
1.74
0.87
-1.99
Monday
Tuesday
Wednesday
Thursday
Friday

Oil


Oil: p-value = 0.170
40
20
0
-20
-40
32
23
-22
-30
-13
Monday
Tuesday
Wednesday
Thursday
Friday

S&P 500


S&P 500: p-value = 0.331
800
600
400
200
0
101
642
602
320
333
Monday
Tuesday
Wednesday
Thursday
Friday

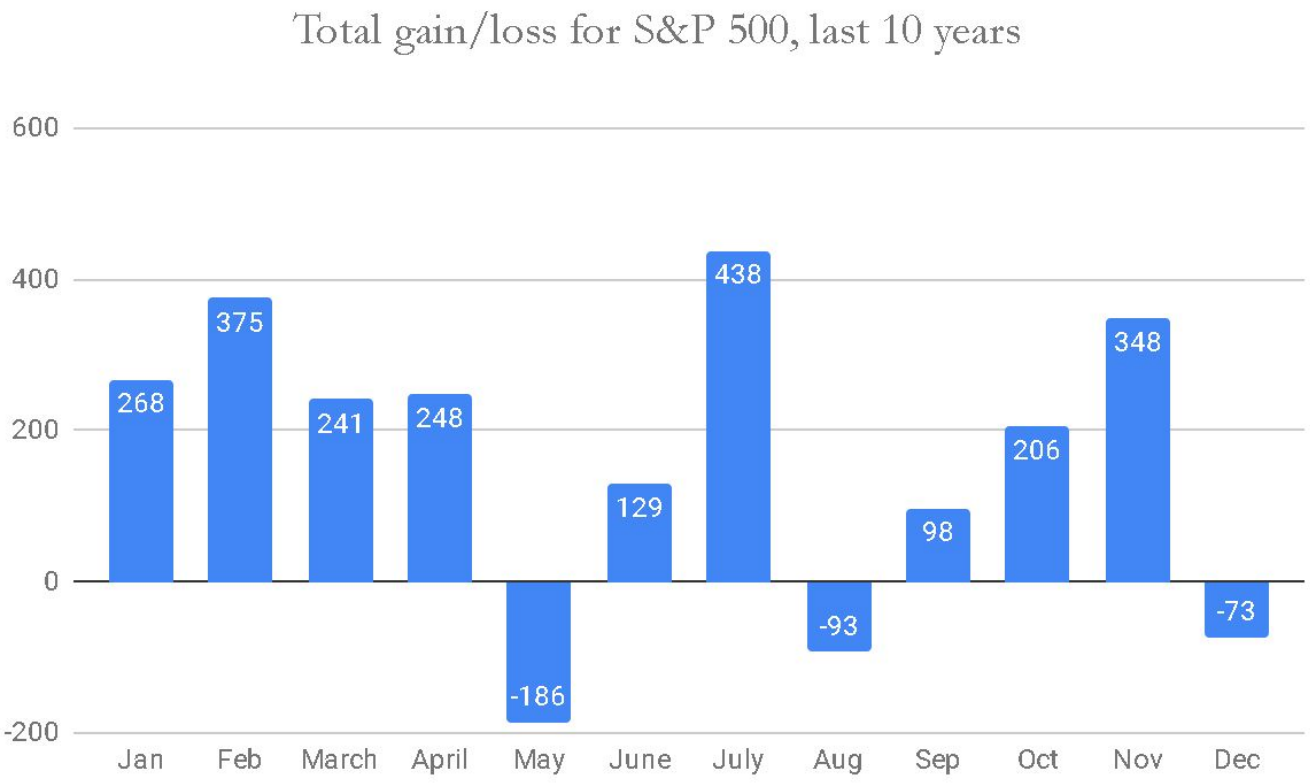
Question 2: Do winter months perform better than summer ? (anwer: no)

- This is for the S&P 500 only over the last 10 years
- ‘Winter’ is Nov 1 - April 30
- Average daily return for summer: 0.24
- Average daily return for winter: 0.56
- Is this significant ?
- p-value = 0.332
- Not significant
- This is the ‘sell in May and go away theory’

Does 3rd Presidential Year yield better returns in ‘winter’ ? (answer: no)

- For S&P 500 over last 10 years
- Average daily return for summer: -0.06
- Average daily return for winter: 0.45
- Is this significant ?
- p-value = 0.134
- Not significant

Question 4: Does any month perform better than another ? (answer: yes; p-value = 0.03)



Question 5: Is it useful to build a SL model to predict gold price change using day of week ? (answer: possibly)

- 1 for up, 0 for down
- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier
- 80% train, 20% test
- Balanced data
- Accuracy = 0.49 for all
- Too many false negatives

```
[49]: df_gold.head()
```

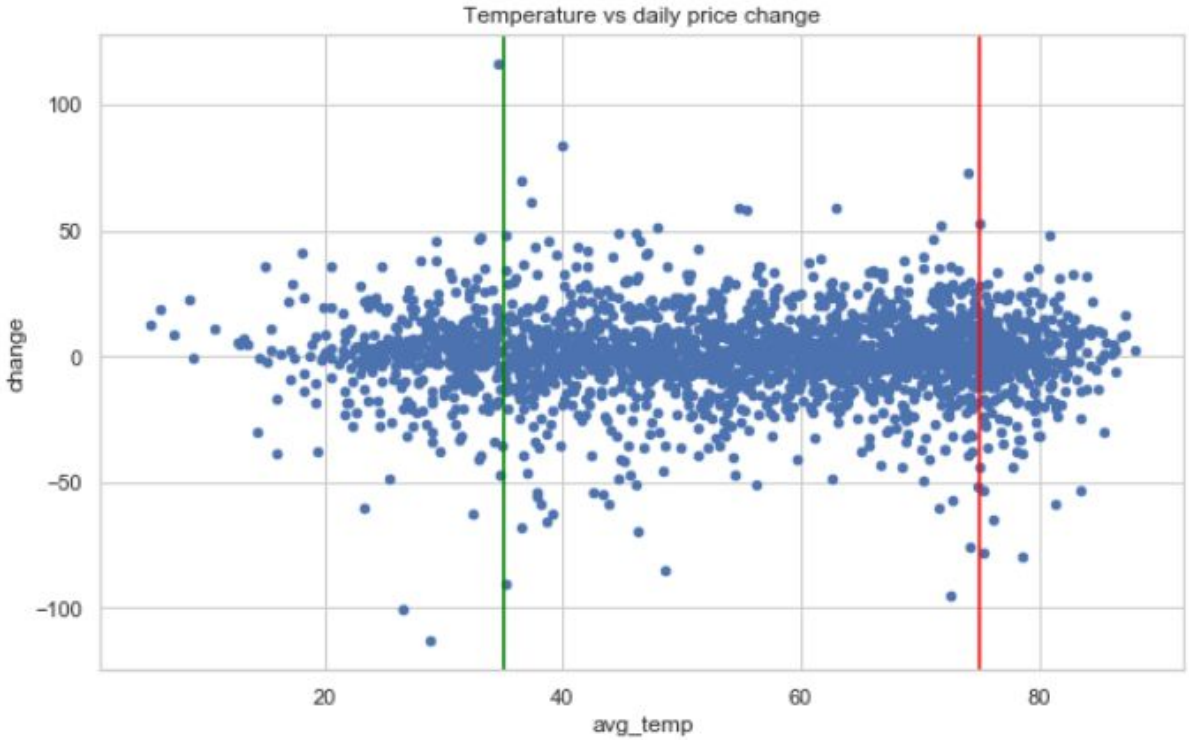
[49]:

	binary	zero	one	two	three	four
0	1	0	0	0	0	1
1	0	1	0	0	0	0
2	0	0	1	0	0	0
3	1	0	0	1	0	0
4	1	0	0	0	1	0

Question 6: Does the daily temperature or sunshine in NYC affect stock prices ? (answer: no)

- Testing the overall price gain/loss for each segment
- For temperature:
  - Units are degrees Fahrenheit
  - Used 55 (avg. temp in NYC) for cutoff
  - Also tested above 75 vs below 35
- For sunshine:
  - Units are minutes of sunshine per day
  - Used 300, 500, and 700 minutes per day as cutoff values
- No p-values were above 0.05, so not significant.

Temperature vs. S&P 500 price change



Daily sunshine vs. S&P 500 price change

Question 7: Is it useful to build a SL model using all data to predict price change of the S&P 500 ? (answer: possibly)

- Used regression this time
- Linear Regression
- Random Forest Regressor
- Gradient Boosting Regressor
- 80% train, 20% test
- R- squared (test) around 0 for all
- RF and GB yields 30%, 60% for training

	DATE	weekday	month	day	change	avg_temp	Sun
0	2009-06-26	4	6	26	0.00	76.91	418.70
1	2009-06-29	0	6	29	8.33	75.43	848.28
2	2009-06-30	1	6	30	-7.91	77.41	608.00
3	2009-07-01	2	7	1	4.01	73.40	168.86
4	2009-07-02	3	7	2	-26.91	71.23	225.66

Question 8: Can unsupervised learning capture relationships among these 4 securities ? (answer: possibly)

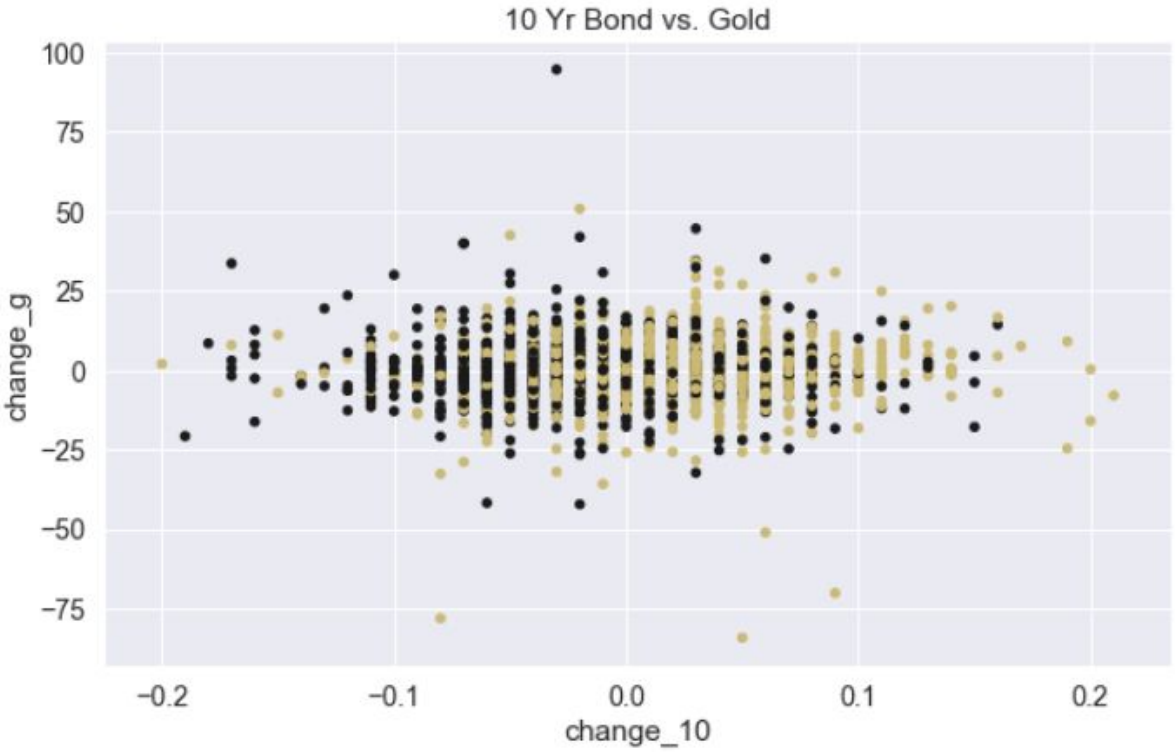
- No bitcoin because lack of data
- Binarized ‘change’ column
- Used K-Means and Mean Shift

```
Comparing k-means clusters against the data:
col_0     0    1   2   3   4    5   6    7   8   9
row_0
0       288  277  44  38  42  130  64  131  74  42
1       353  343  69  49  52  163  71  145  79  43
```

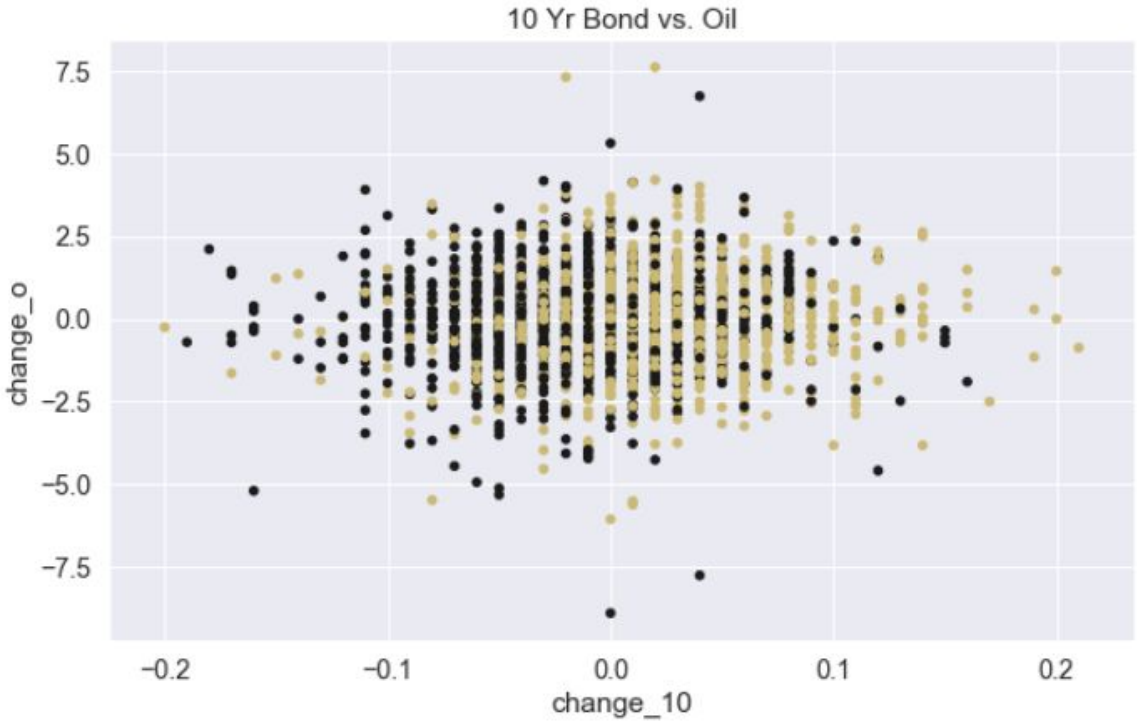
```
Comparing the assigned categories to the ones in the data:
col_0     0    1   2   3  4   5  6  7
row_0
0       480  481  54  83  1  28  1  2
1       576  583  88  93  2  25  0  0
```

	dt	change	change_10	change_g	change_o
2492	2019-06-19	8.71	-0.03	-4.03	-0.98
2493	2019-06-20	27.72	-0.02	-0.53	-1.13
2494	2019-06-21	-3.72	0.06	2.46	1.49
2495	2019-06-24	-5.11	-0.05	-11.58	0.28
2496	2019-06-25	-27.97	-0.02	1.74	-0.25

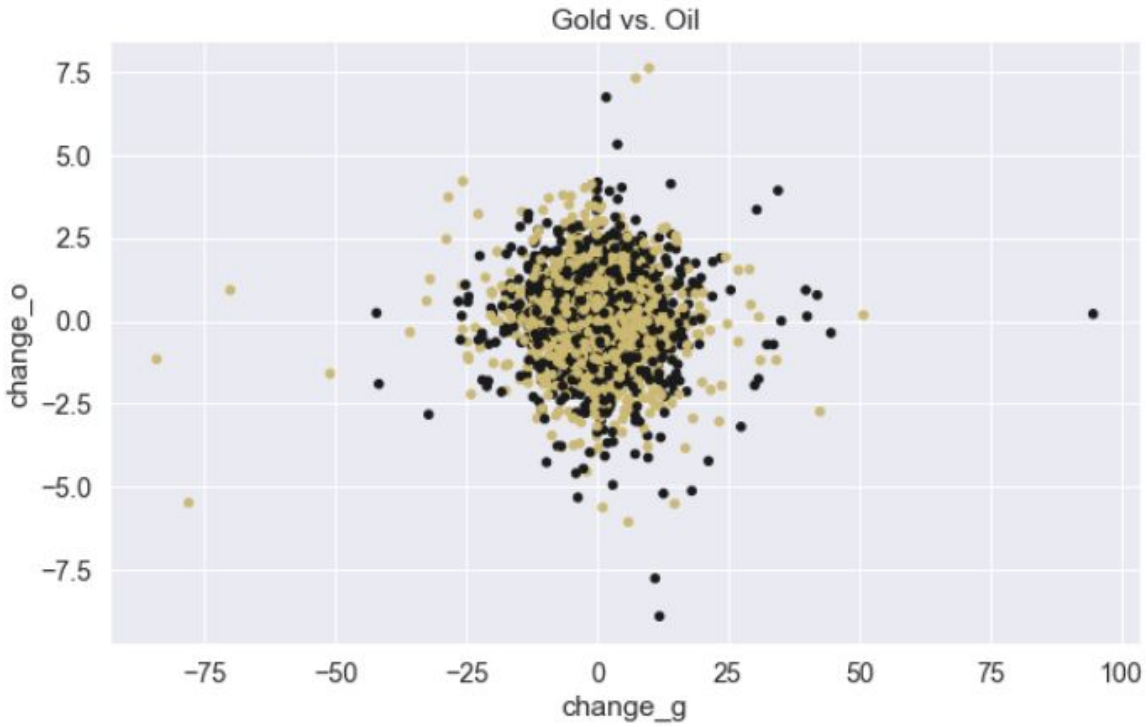
Scatterplot: 10 Yr Bond, Gold, Oil with binary S&P 500



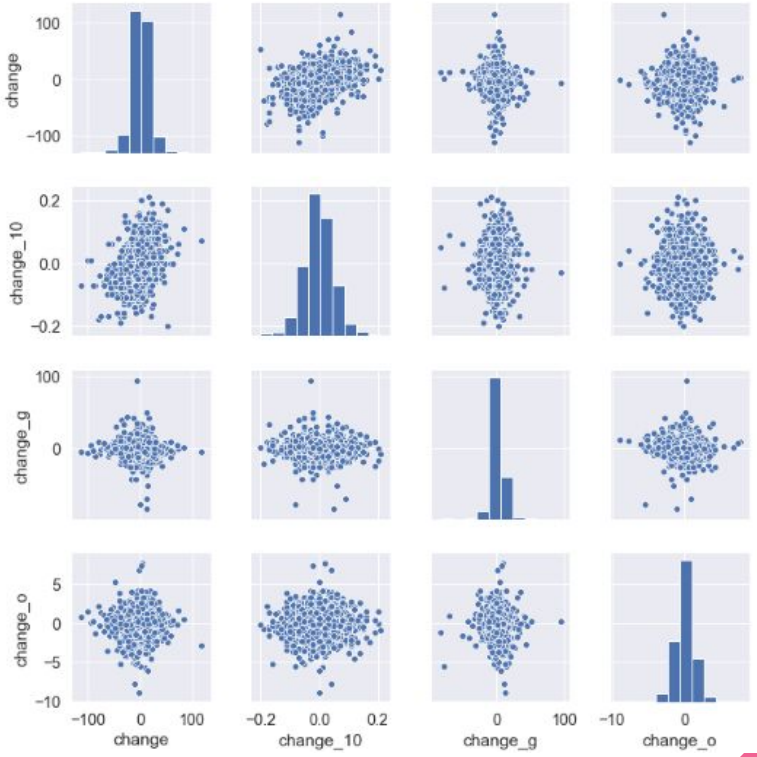
10 Yr Bond vs. Oil



Gold vs. Oil


Gold vs. Oil
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
change_o
−75
−50
−25
0
25
50
75
100
change_g

Pairplot - S&P 500, 10 Yr Bond, Gold, Oil



Conclusion

This study found no affect or no correlation for the following relationships:

- Day of the week vs price change of oil and S&P 500
- Winter does not perform better or worse than summer for S&P 500
- 3rd Presidential Year is not better or worse
- Daily temperature or minutes of sunshine vs price change of S&P 500

Affects or correlations were found in the following:

- Day of the week vs price change for 10 Yr Bond, gold, bitcoin
- July performed statistically better than May for the S&P 500

Inconclusive affects or correlations:

- Supervised Learning to predict price change based on day of the week for gold
- Supervised Learning to predict price change of S&P 500 using weather, day, month, day of week
- Unsupervised Learning to discover relationships between S&P 500, 10 Yr, gold, oil

Other considerations / further study

- Ground truth = 0.5 for classification models
- Could have investigated impact of daily snowfall or rain in NYC
- Could have tried different algorithms for supervised learning and unsupervised learning
- Could have checked correlation between month and price change for other securities (only investigated S&P 500)
- Links to studies that show bullish ‘3r Presidential Year’ and ‘sell in may’ theory
  - https://bullmarkets.co/3rd-year-presidential-cycle-bullish-stocks/
  - https://www.marketwatch.com/story/heres-the-real-story-behind-sell-in-may-and-go-away-2017-04-25
- Links that correlate weather to higher stock performance:
  - https://www.jstor.org/stable/3094570?seq=1#page_scan_tab_contents
  - https://sites.uci.edu/dhirshle/files/2011/02/Good-Day-Sunshine-Stock-Returns-and-the-Weather.pdf