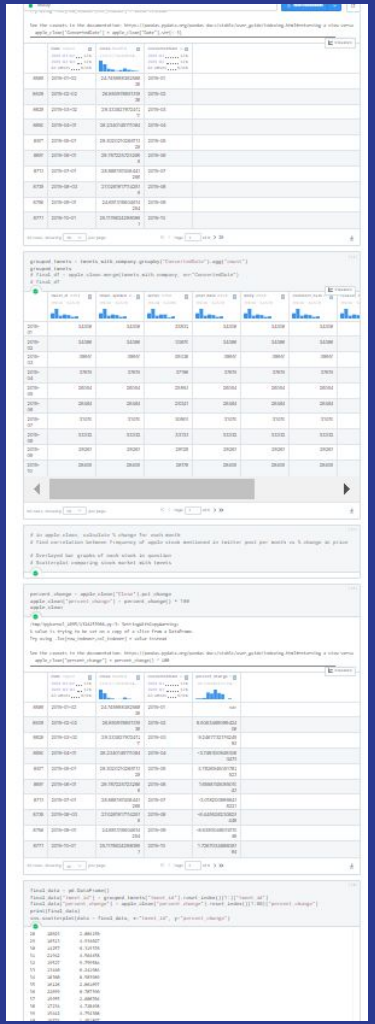
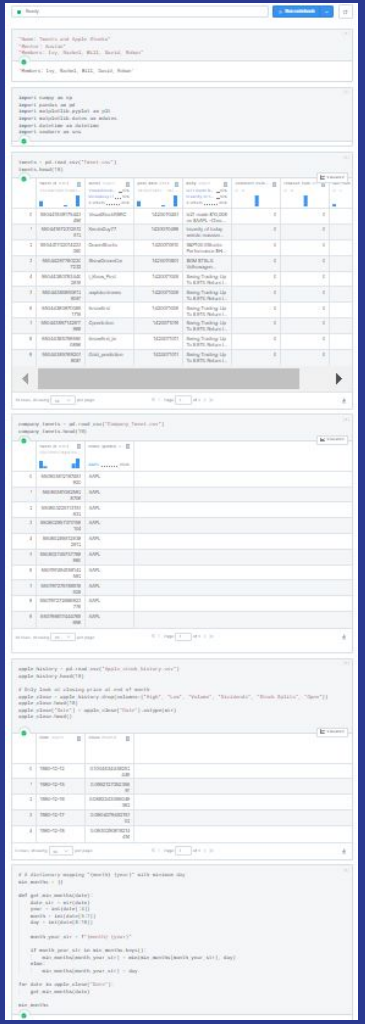
Twitter and Stocks

Rohan Ganguli, David Li, Rachel Park, William Rothman, Ivy Sim

Mentored by Austin Nicola Ardisaputra






Null Distribution of Slope Estimates
Frequency
Slope

Background I

- Why Apple?

Background I

- Why Apple?
  - Apple products are integral to our way of life
  - FAANG! All of us here love technology
  - Also, Apple is a good case study

Background II

- Why Twitter?

Background II

- Why Twitter?
  - Primary source of public communication of large companies and public officials
  - Like Apple, everyone uses it
  - Most publicly available dataset with stock tickers

Our Hypothesis

Null: A company's tweet frequency does not affect their stock price.

Our Hypothesis

Null: A company's tweet mention frequency does not affect their stock price.

Alternative: A company's tweet mention frequency affects their stock price.

Methodology

1. **Data Collection:** Twitter, Apple stock…
2. **Data Cleaning & Preprocessing:** Standardization
3. **Data Analysis**
    a. **Data visualization:** Matplotlib, Seaborn, Plotly
    b. **Hypothesis testing**

Our Datasets

- Apple Stock Data
  - Shows the stock history of Apple, showing the opening price (according to each date), in addition to the high, low, and closing prices. Also has some additional information, such as volume traded.
- Tweets
  - Shows data about the tweets that were posted, including post date, number of comments, retweets, likes, etc..

Date	Open	High	Low	Close	Volume
Date	Opening Price	High Value	Low Value	Closing Price	Volume Traded
11Dec80 – 2Jan22	0.04 – 181	0.04 – 183	0.04 – 179	0.04 – 182	0 – 7.4
1980-12-12	0.1004534438252449	0.10089017369088088	0.1004534438252449	0.1004534438252449	469033600
1980-12-15	0.09564946514251238	0.09564946514251238	0.09521273523569107	0.09521273523569107	175884800
1980-12-16	0.08866106065234197	0.08866106065234197	0.08822433650493622	0.08822433650493622	105728000
1980-12-17	0.09040794521570206	0.0908446750799829	0.09040794521570206	0.09040794521570206	86441600
1980-12-18	0.09302908182144165	0.09346580579661565	0.09302908182144165	0.09302908182144165	73449600
1980-12-19	0.09870654344558716	0.09914327334405434	0.09870654344558716	0.09870654344558716	48630400
1980-12-22	0.10351057350635529	0.10394808490351318	0.10351057350635529	0.10351057350635529	37363200
1980-12-23	0.10787859559059143	0.10831532549389696	0.10787859559059143	0.10787859559059143	46950400
1980-12-24	0.11355603486299515	0.11399276471060095	0.11355603486299515	0.11355603486299515	48003200
1980-12-26	0.12403828650712966	0.12447501643882311	0.12403828650712966	0.12403828650712967	55574400

tweet_id	writer	post_date	body	comment_num	retweet_num	like_num
550441509175443456	VisualStockRSRC	1420070457	lx21 made $10,008on$ AAPL -Check it out! htt...	0	0	1
550441672312512512	KeralaGuy77	1420070496	Insanity of today weirdo massive selling. $aap...	0	0	0
550441732014223360	DozenStocks	1420070510	S&P100 #Stocks Performance HDLOW $SBUX$TGT...	0	0	0
550442977802207232	ShowDreamCar	1420070807	GMTSLA: Volkswagen Pushes 2014 Record Recal...	0	0	1
550443807834402816	i_Know_First	1420071005	Swing Trading: Up To 8.91% Return In 14 Days h...	0	0	1
...	...	...	...	...	...	...
1212159765914079234	TEEELAZER	1577836383	That SPY SPX puuump in the last hour was the...	1	0	6
			In 2020 I may start			

Our Datasets

- Company
  - Shows data associating the company ticker symbol with the company name. Total of 6 different ticker symbols.
- Company_Tweets
  - Shows data with a tweet id along with the corresponding ticker symbol.

	ticker_symbol	company_name
0	AAPL	apple
1	GOOG	Google Inc
2	GOOGL	Google Inc
3	AMZN	Amazon.com
4	TSLA	Tesla Inc
5	MSFT	Microsoft

	tweet_id	ticker_symbol
0	550803612197457920	AAPL
1	550803610825928706	AAPL
2	550803225113157632	AAPL
3	550802957370159104	AAPL
4	550802855129382912	AAPL
...	...	...
4336440	1212158772015034369	TSLA
4336441	1212159099632267268	TSLA
4336442	1212159184931717120	TSLA
4336443	1212159838882533376	TSLA
4336444	1212160015332728833	TSLA

Data Science Cycle

- Dropped a lot of unnecessary columns within each dataset
- Cleaned the data to only consist of years after 2015.

```
apple_history = pd.read_csv("Apple_stock_history.csv")
apple_history.head(10)

# Only look at closing price at end of month
apple_close = apple_history.drop(columns=["High", "Low", "Volu
apple_close.head(10)
apple_close["Date"] = apple_close["Date"].astype(str)
apple_close.head()
```

	Date object	Close float64
0	1980-12-12	0.1004534438252

```
start_dates = apple_close["Date"].apply(first_day_of_month_and_after_2015)
apple_clean = apple_close[start_dates]
apple_clean
```

	Date object	Close float64
	2015-01-02 ..... 1.2% 2015-02-02 ..... 1.2% 83 others 97.6%	21.6727752685546... 
8799	2015-11-02	27.77659034729004
8819	2015-12-01	27.0115127563476

Data Science Cycle

- Merged datasets together
- Calculated the data we need: percent change in stock price

```
tweets_with_company = company_tweets.merge(tweets, on="tweet_id")
tweets_with_company = tweets_with_company[tweets_with_company["ticker_symbol"] == "AAPL"]
tweets_with_company['post_date'] = pd.to_datetime(tweets_with_company['post_date'], unit='s')
tweets_with_company['date'] = pd.to_datetime(tweets_with_company['post_date'].apply(lambda date: date.
# only include >= 2015-1
#date_before = datetime.date(2015, 1, 1)
#tweets_with_company = tweets_with_company[tweets_with_company['post_date'].dt.date > date_before]
tweets_with_company['ConvertedDate'] = tweets_with_company['date'].dt.strftime('%Y-%m')
tweets_with_company
```

Visual

	tweet_id int64	ticker_symbol o...	writer object	post_date dateti...	body object	comment_num i...
0	550803612197457920	AAPL	SentiQuant	2015-01-01 23:59:49	#TOPTICKERTWEETS $AAPL $IMRS...	

```
percent_change = apple_clean["Close"].pct_change
apple_clean["percent_change"] = percent_change() * 100
apple_clean
```

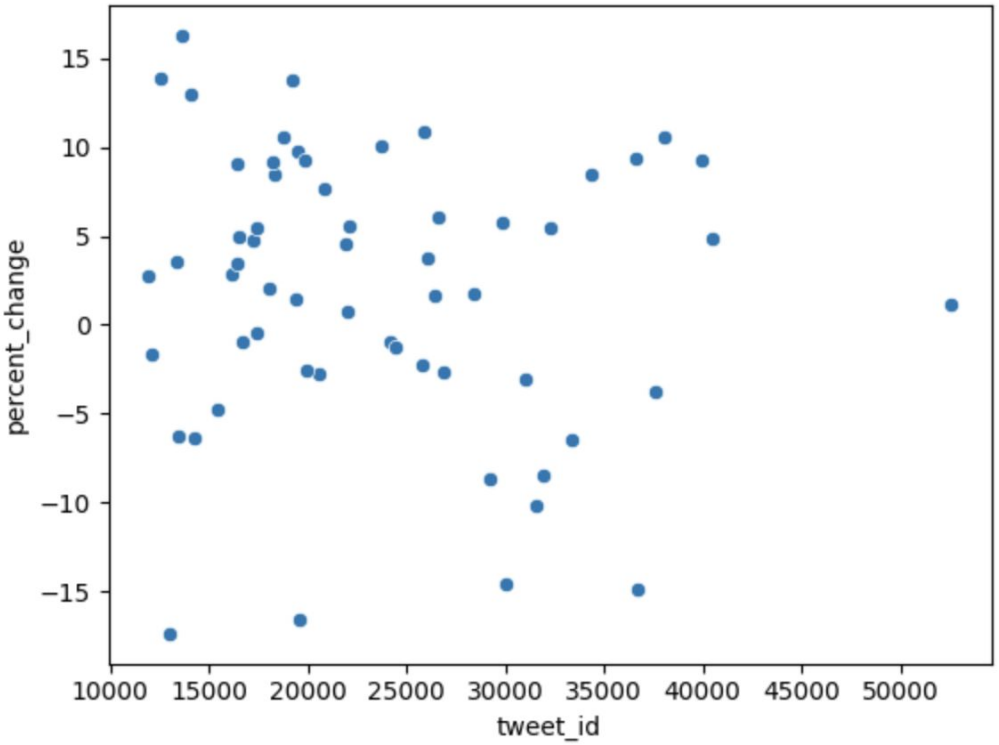
```
/tmp/ipykernel_107/1324153966.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user
  apple_clean["percent_change"] = percent_change() * 100
```

	Date object	Close float64	ConvertedDate o...	percent_change f...
	2015-01-02 ..... 1.2% 2015-02-02 ..... 1.2% 83 others 97.6%	21.6727752685546... 	2015-01 ..... 1.2% 2015-02 ..... 1.2% 83 others 97.6%	-19.3768611717247... 
8589	2015-01-02	24.74599838256836	2015-01	nan
8609	2015-02-02	26.85097885131836	2015-02	8.506346909942408
8628	2015-03-02	29.33382797241211	2015-03	9.246773217624593

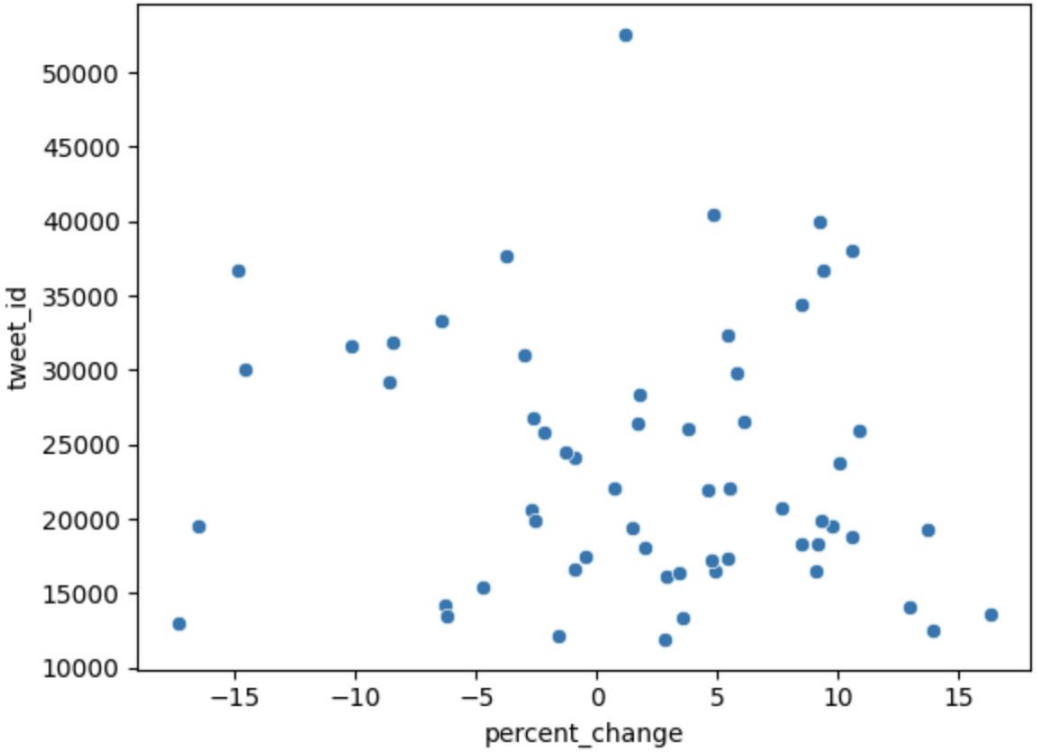
Visualization

- We used scatter plots to visualize the numerical relationship between the number of tweets and the corresponding monthly stock price change.
- Using line graphs or bar graphs could not accurately portray the high of variation data points.

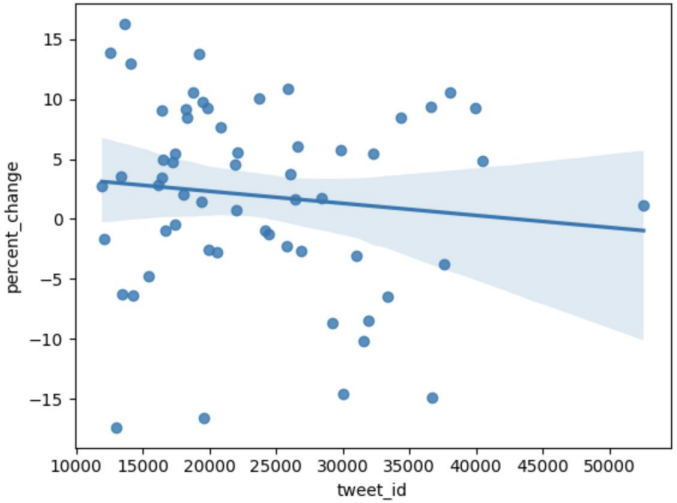
of Tweets vs. Percent Change

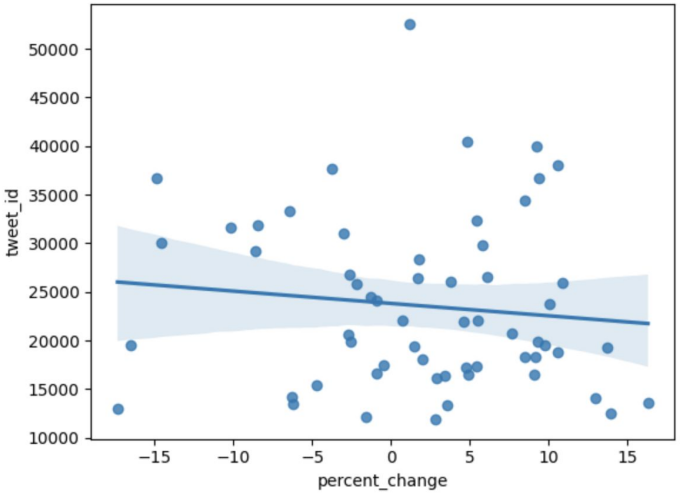


Percent Change vs. # of Tweets



Regression Plots

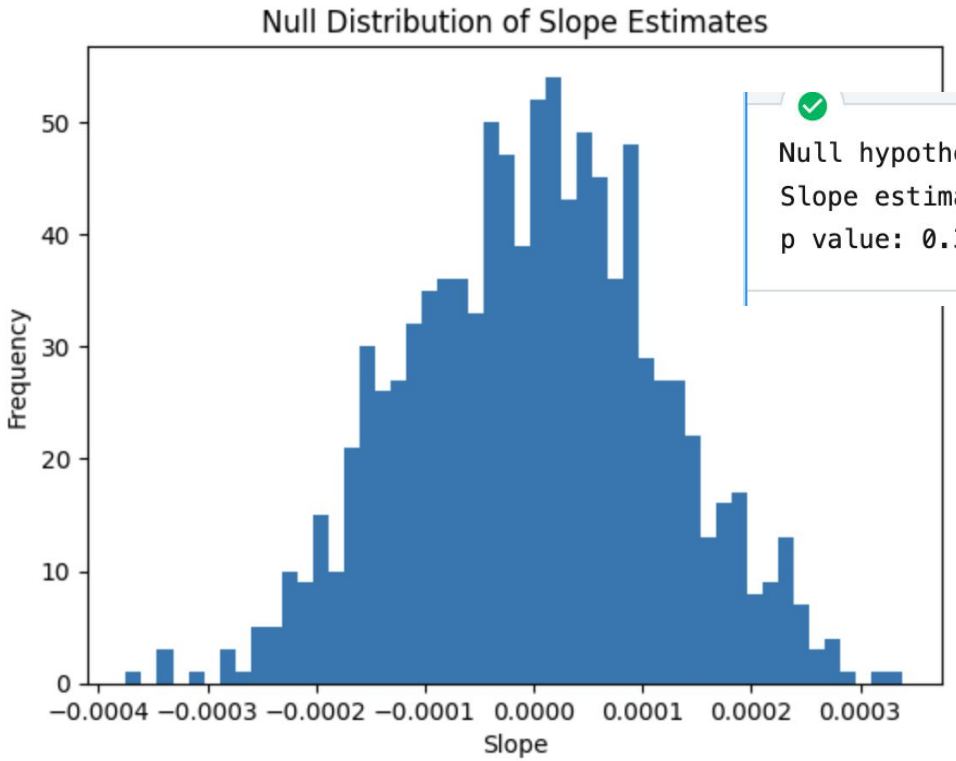




Bootstrap Test

- Null Hypothesis: The slope of the regression line is equal to 0.
- Perform regressions on the bootstrapped data
- 1000 simulation
- Calculate the p-value

Bootstrap Visualization




```
Null hypothesis: The slope of the regression line is equal to 0.
Slope estimate: -0.00010062913646617312
p value: 0.395
```

Further Steps

Further Steps

1. Using a prediction model - although the results turned out to be relatively inconclusive, it would be interesting to generate a Machine Learning based prediction model to estimate the amount of percent change that correlates to the amount of tweets and vice-versa.

Further Steps

1. Using a prediction model - although the results turned out to be relatively inconclusive, it would be interesting to generate a Machine Learning based prediction model to estimate the amount of percent change that correlates to the amount of tweets and vice-versa.
2. Expanded dataset - the tweets and stock ticker/valuation only dated back to 2015, and only used monthly data - conclusions drawn from a wider ranging dataset that maybe used daily values and insights could be more comprehensive.

Thank you.