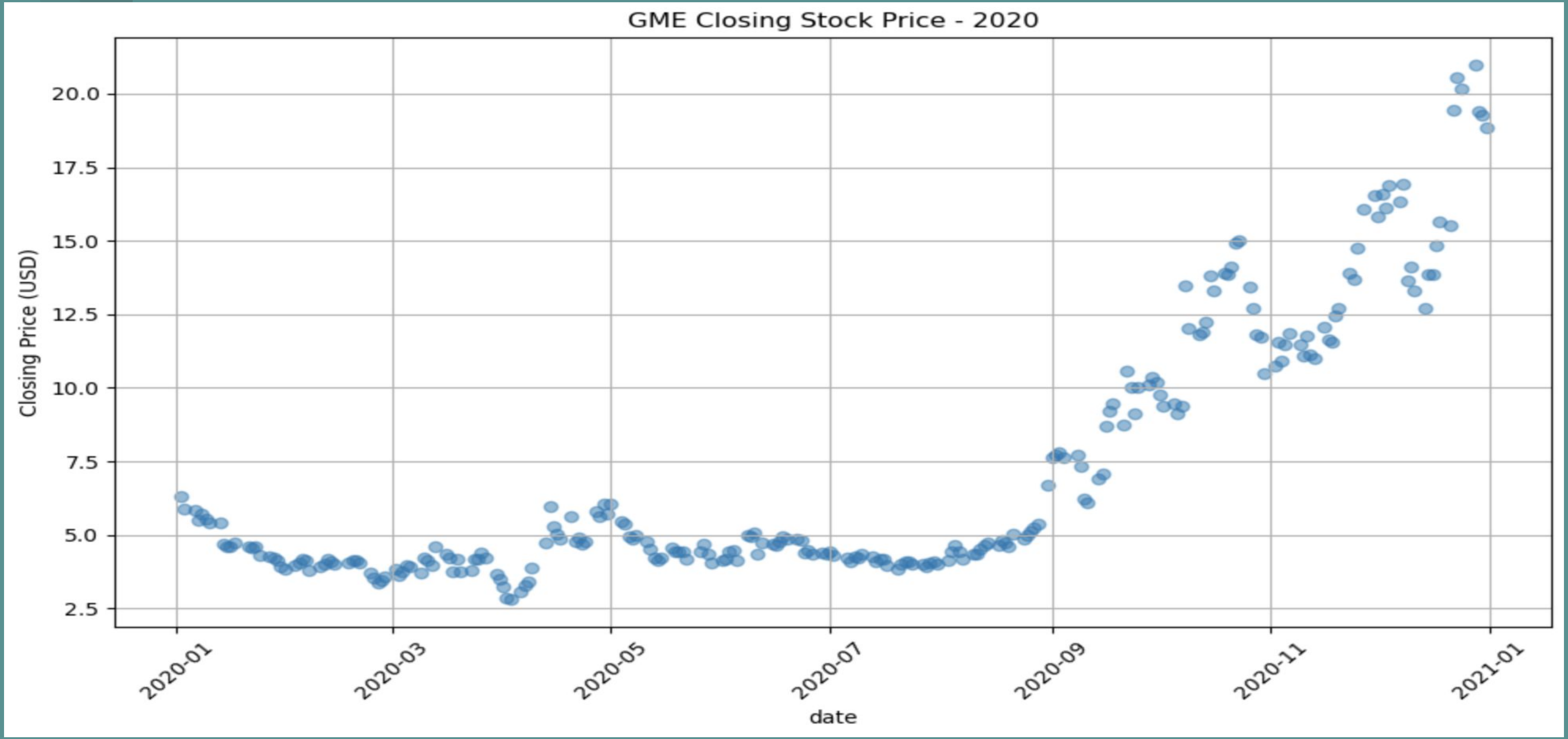
Data Analytics Project 1

Stock Market Analysis

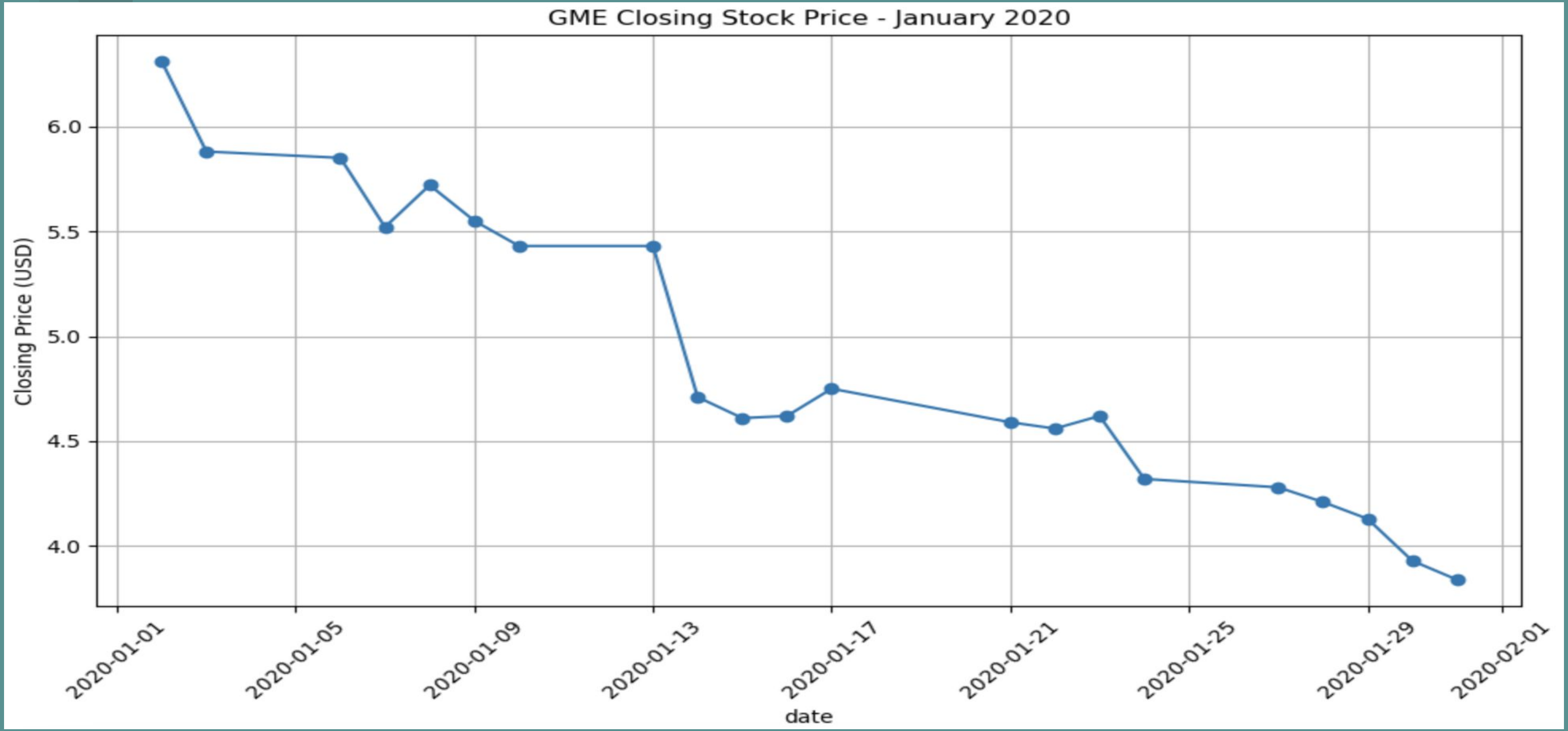
Problem Statement

- The Stock Market is extremely complicated Machine with many different factors influencing it
- Can the internet influence the stock market?
- Does Covid affect the stock market?
- All Datasets are from Kaggle
- https://www.kaggle.com/datasets/leukipp/reddit-finance-data/data
- https://www.kaggle.com/datasets/hananxx/gamestop-historical-stock-prices?rvi=1
- https://www.kaggle.com/datasets/stefanoleone992/mutual-funds-and-etfs?select=MutualFund+prices+-+L-P.csv

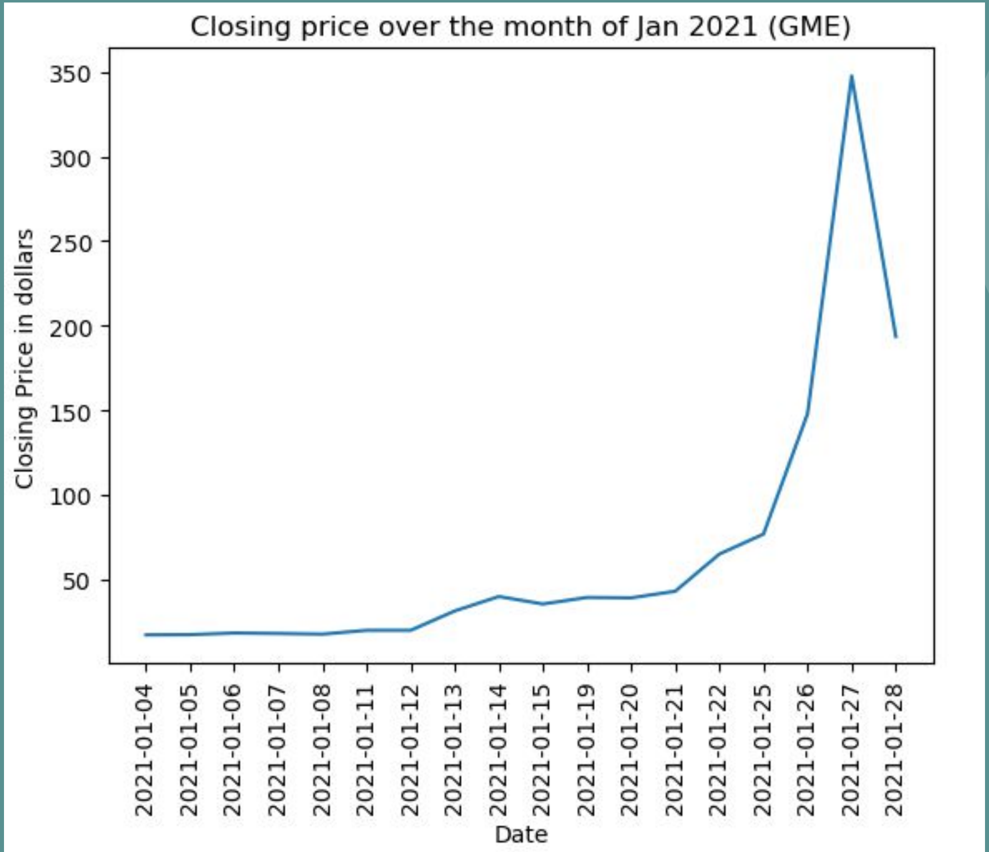
GME Stock Prices through 2020



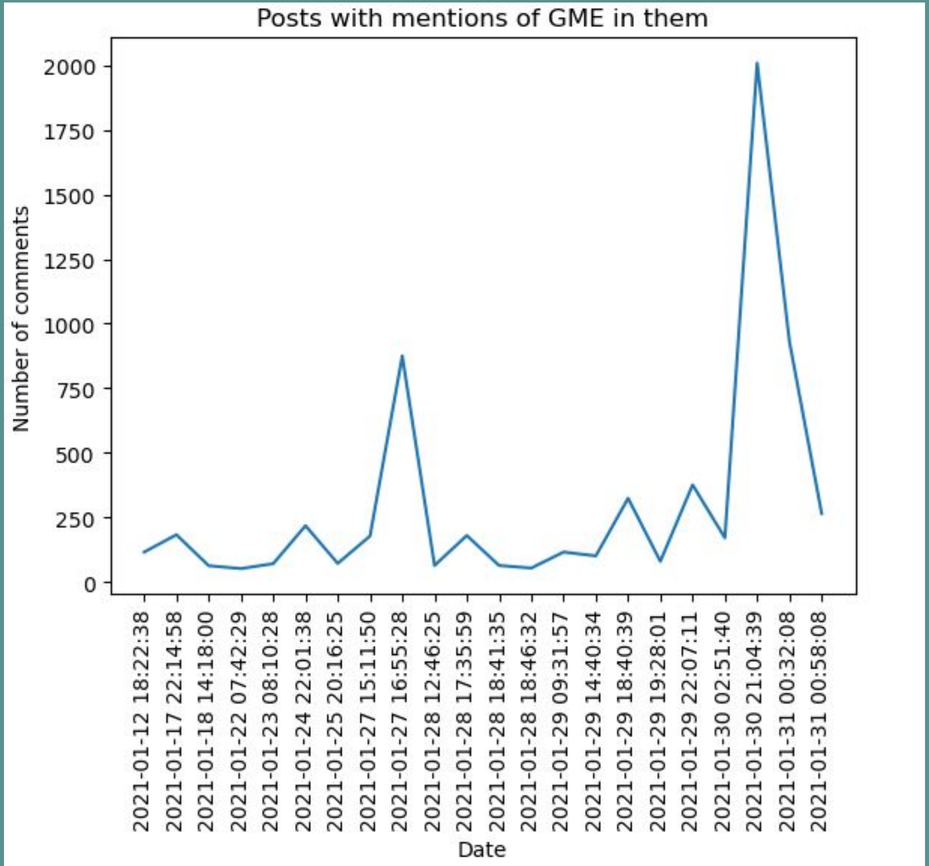
GME Stock Price for January 2020

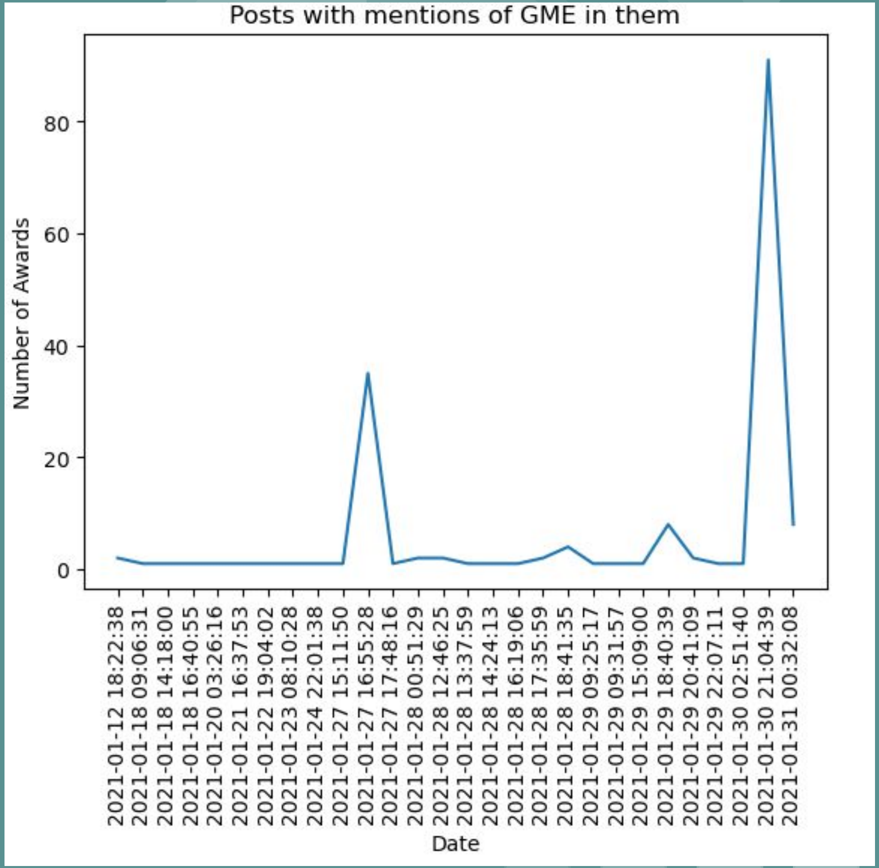


GME Stock Prices through Jan 2021

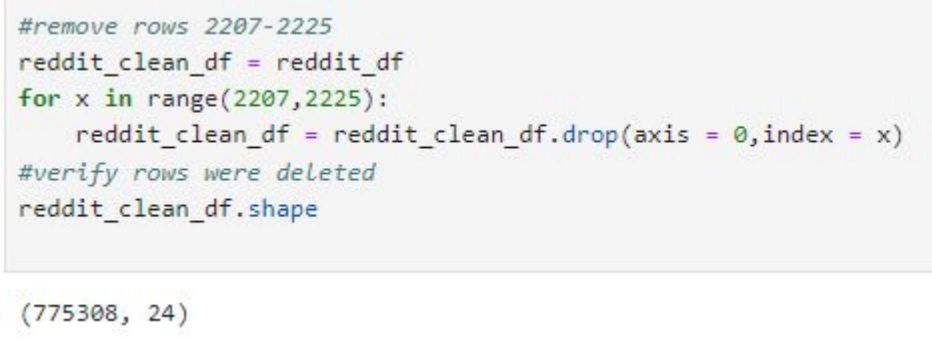


Reddit Posts with GME in it



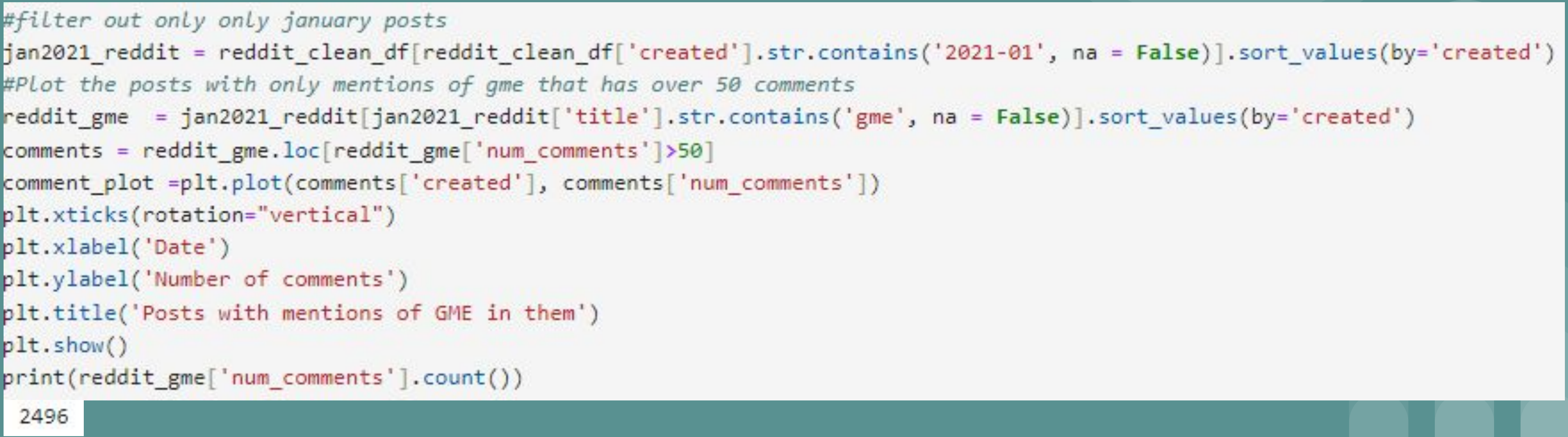


Code for filtered data



```
#remove rows 2207-2225
reddit_clean_df = reddit_df
for x in range(2207,2225):
    reddit_clean_df = reddit_clean_df.drop(axis = 0,index = x)
#verify rows were deleted
reddit_clean_df.shape
```

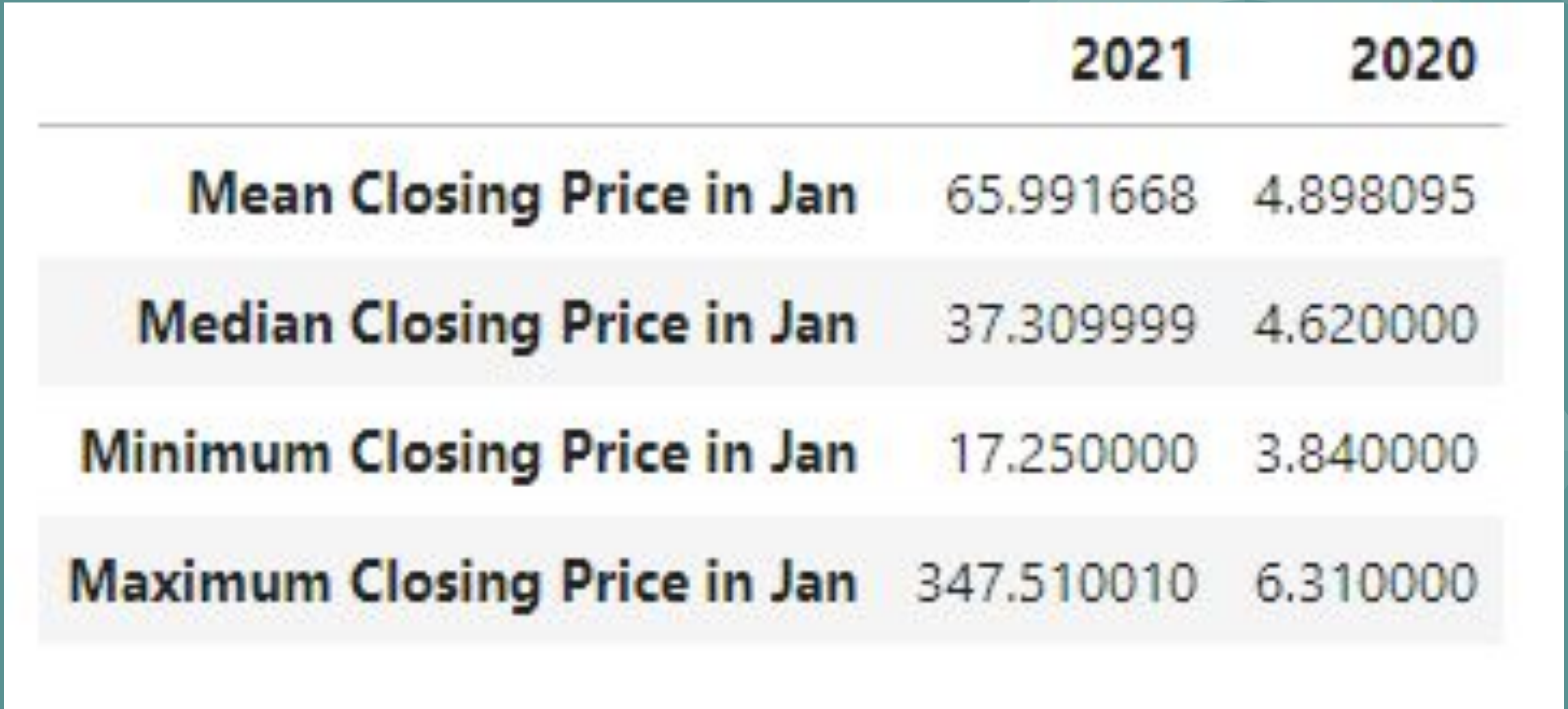
```
(775308, 24)
```



```
#filter out only only january posts
jan2021_reddit = reddit_clean_df[reddit_clean_df['created'].str.contains('2021-01', na = False)].sort_values(by='created')
#Plot the posts with only mentions of gme that has over 50 comments
reddit_gme  = jan2021_reddit[jan2021_reddit['title'].str.contains('gme', na = False)].sort_values(by='created')
comments = reddit_gme.loc[reddit_gme['num_comments']>50]
comment_plot =plt.plot(comments['created'], comments['num_comments'])
plt.xticks(rotation="vertical")
plt.xlabel('Date')
plt.ylabel('Number of comments')
plt.title('Posts with mentions of GME in them')
plt.show()
print(reddit_gme['num_comments'].count())
```

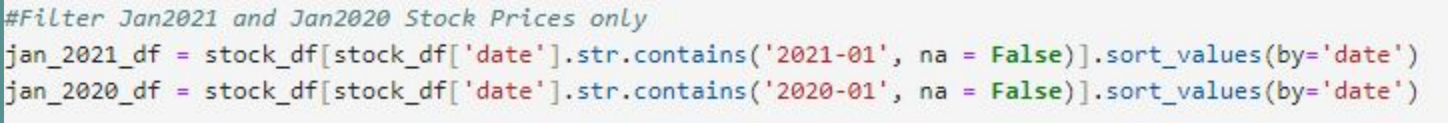
```
2496
```

GME Stock Prices through Jan 2021

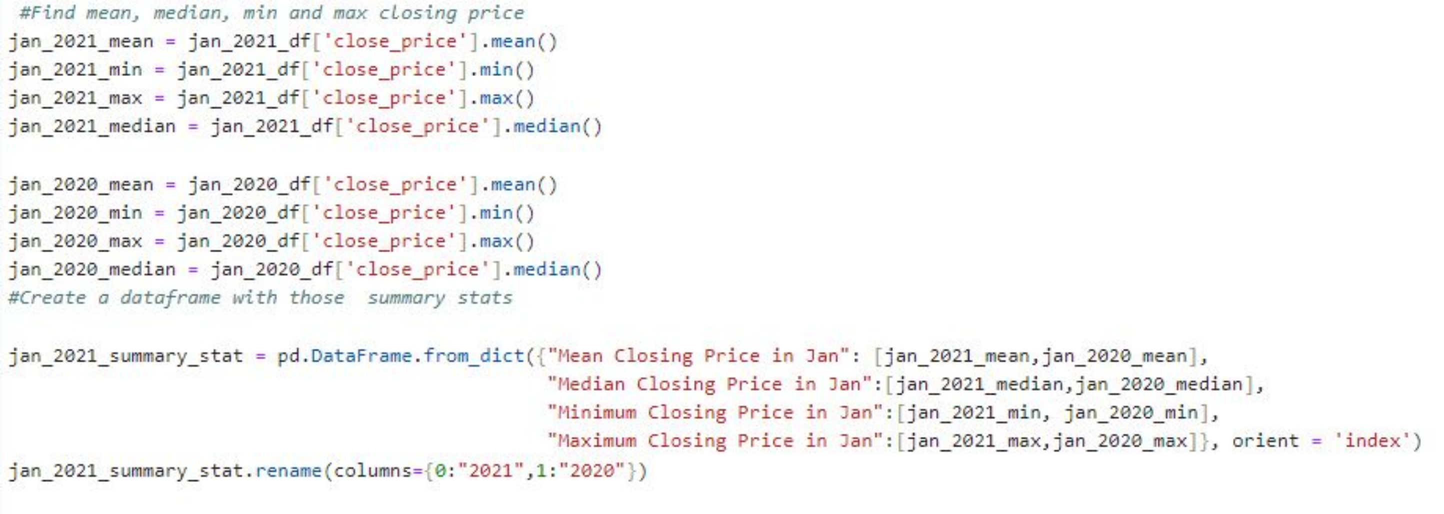


	2021	2020
Mean Closing Price in Jan	65.991668	4.898095
Median Closing Price in Jan	37.309999	4.620000
Minimum Closing Price in Jan	17.250000	3.840000
Maximum Closing Price in Jan	347.510010	6.310000

Code for filtered data

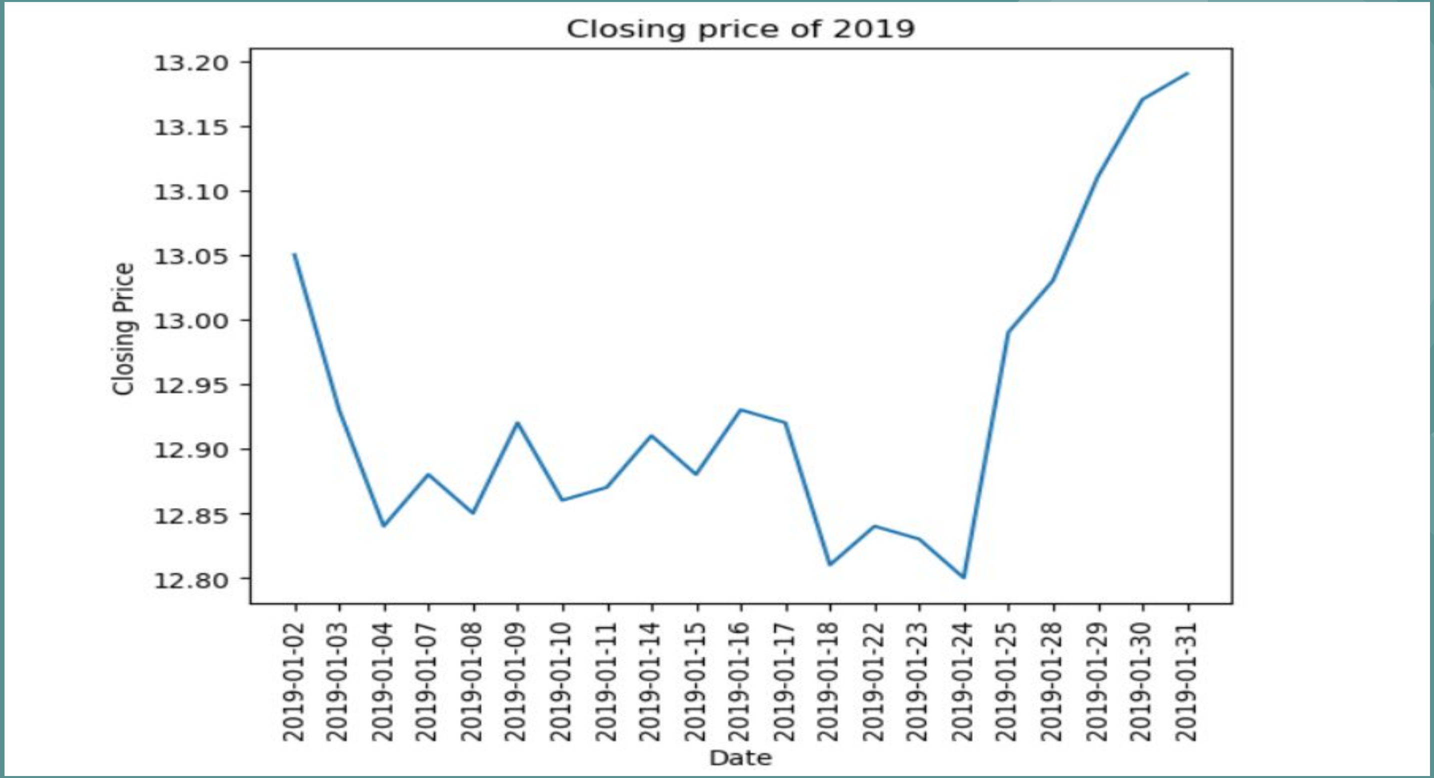


```
#Filter Jan2021 and Jan2020 Stock Prices only
jan_2021_df = stock_df[stock_df['date'].str.contains('2021-01', na = False)].sort_values(by='date')
jan_2020_df = stock_df[stock_df['date'].str.contains('2020-01', na = False)].sort_values(by='date')
```

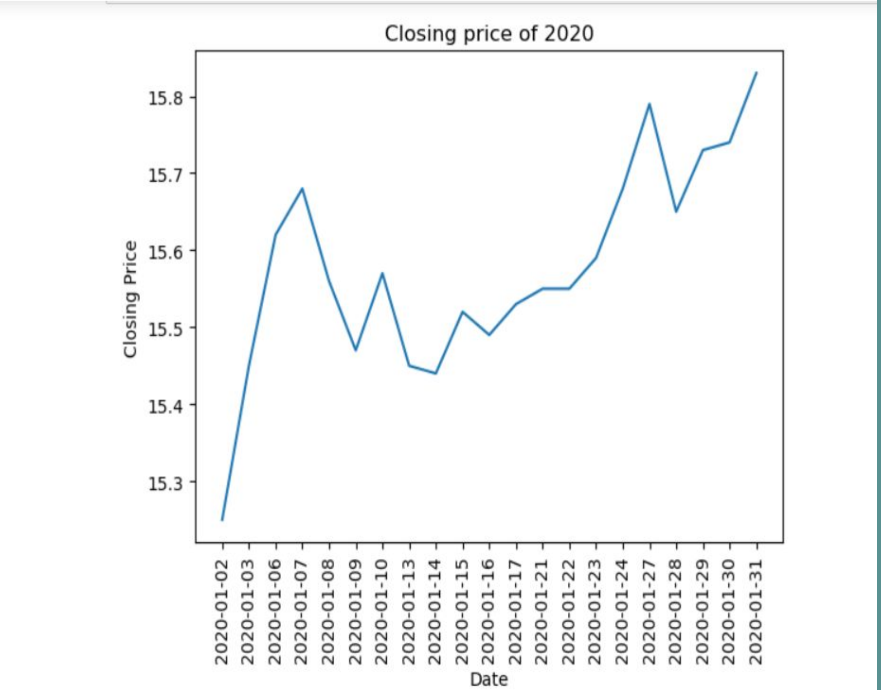


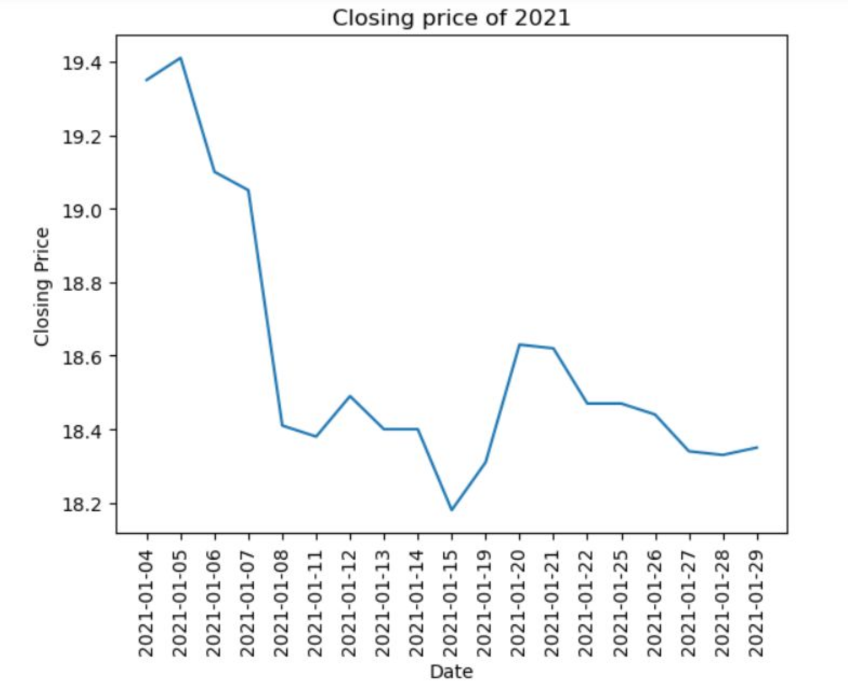
```
 #Find mean, median, min and max closing price
jan_2021_mean = jan_2021_df['close_price'].mean()
jan_2021_min = jan_2021_df['close_price'].min()
jan_2021_max = jan_2021_df['close_price'].max()
jan_2021_median = jan_2021_df['close_price'].median()

jan_2020_mean = jan_2020_df['close_price'].mean()
jan_2020_min = jan_2020_df['close_price'].min()
jan_2020_max = jan_2020_df['close_price'].max()
jan_2020_median = jan_2020_df['close_price'].median()
#Create a dataframe with those  summary stats

jan_2021_summary_stat = pd.DataFrame.from_dict({"Mean Closing Price in Jan": [jan_2021_mean,jan_2020_mean],
                                                "Median Closing Price in Jan":[jan_2021_median,jan_2020_median],
                                                "Minimum Closing Price in Jan":[jan_2021_min, jan_2020_min],
                                                "Maximum Closing Price in Jan":[jan_2021_max,jan_2020_max]}, orient = 'index')
jan_2021_summary_stat.rename(columns={0:"2021",1:"2020"})
```

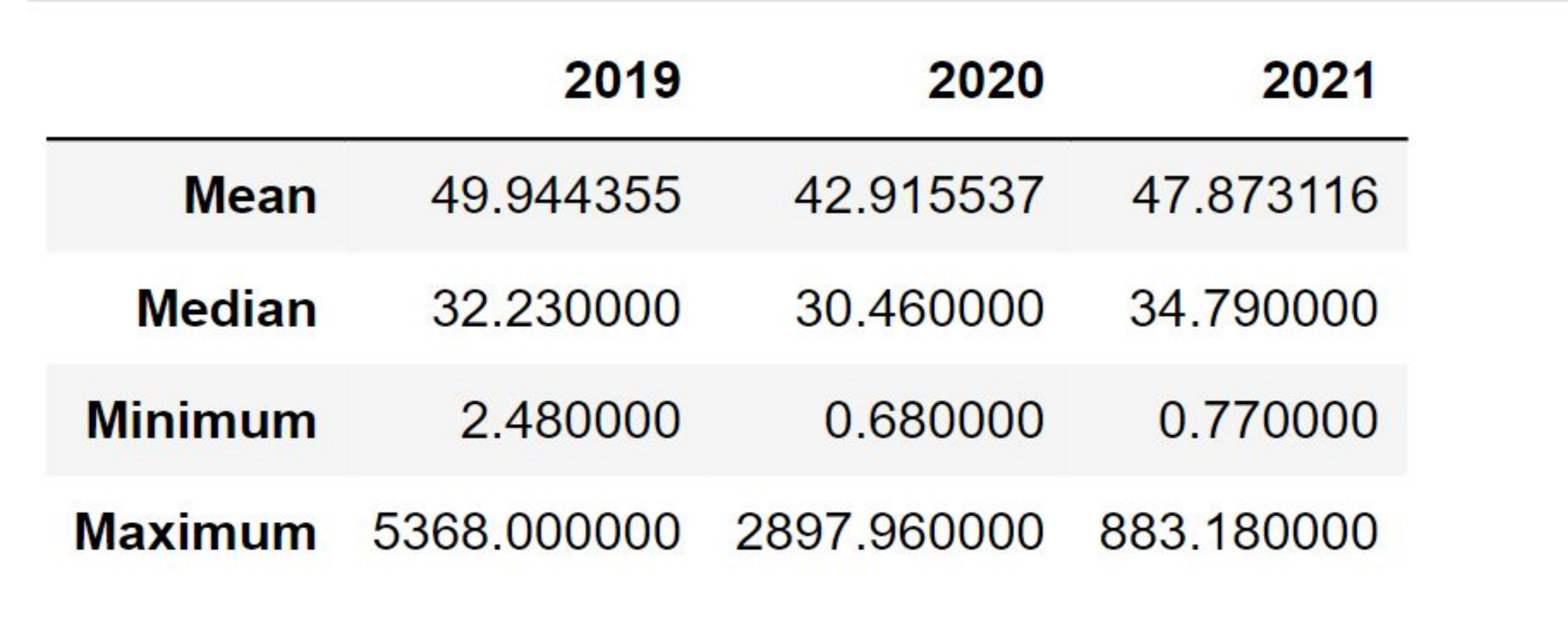
ETF Stock prices of 2019



ETF Stock prices of 2020&2021







	2019	2020	2021
Mean	49.944355	42.915537	47.873116
Median	32.230000	30.460000	34.790000
Minimum	2.480000	0.680000	0.770000
Maximum	5368.000000	2897.960000	883.180000

Summary

- GME experienced a massive increase from Jan 2020 to Jan 2021
- Reddit Posts correlated to the price increases
- ETF Average prices dipped in 2020 from 2019 then increased in 2021
- Maximum Price decreased over the years for ETF