GROUP WORK PROJECT # 3
GROUP NUMBER: 846

MScFE 600: FINANCIAL DATA

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Shailza Virmani	India	virmanishailza@gmail.com	
Harshil Sumra	India	harshilsumra1997@gmail.com	
Wisdom Yakanu	Ghana	ellinamwisdom@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).	
Team member 1	**Shailza Virmani**
Team member 2	**Harshil Sumra**
Team member 3	**Wisdom Yakano**

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed. **Note:** You may be required to provide proof of your outreach to non-contributing members upon request.
N/A

Introduction and Structure of the report

In this report, we'll be focussing on the below two themes:

- Primary theme - Statistical Related Risk: Volatility & Statistical Related Risk: Correlation
- Secondary theme - Fallout: Model Failure & Crises

The report is divided in four different sections:

1. Theoretical Set-up and definitions
2. Illustrations with real world data
3. Literature reviews
4. Conclusion and future research direction

Theoretical set up and definitions

In today's world, where all industries are quite data driven, especially the fintech industry, the utilization of data for the purpose of analysis and forecast has become a very common occurrence. In the financial market, sound data analysis and forecasting are critical as literally billions ride on the decision made based on these analysis.
Most of the aforementioned statistical analysis and modeling is rooted in the core mathematics and the assumptions of each application scenario. For example we can do fundamental analysis at different levels, that is, Company, industry or country levels.Or every model like Capital asset pricing model comes with its own set of defining assumptions which make the analysis easier but also limits the practicality of the analysis and the model has definitely strayed from the real world scenarios.
Here, we'll be focusing on Statistical Risk. It is the quantitative value associated with a particular kind of risk derived from statistical methods. For example, the required rate of return of any investment should be the sum of risk free rate, maturity premium, inflationary premium, liquidity premium and Default risk premium.

$$r = r_{rf} + r_{inflation} + r_{maturity} + r_{liquidity} + r_{default\ risk}$$

Above returns are the returns corresponding to the compensation for given risk inherent to the financial product or investment in consideration. These are calculated using statistical methods.

In case of parameter estimation, the statistical error comes from accuracy of the estimated measure. This accuracy is hugely dependent on the model definition, the corresponding assumptions and the quality of sample data used. From the context of Modern portfolio theory, the statistical risk of investment is measured based on several factors:

1. Variance- is a measure of dispersion of data with respect to the mean of the sample data. It's square root is called volatility/standard deviation.
$$Var(r) = \sigma^2 = \Sigma(r_i - E(r))/(n - 1)$$
$$Std\ dev(r) = \sigma = (\Sigma(r_i - E(r))/(n - 1))^{0.5}$$
2. Alpha - is the measure of the given product's risk/return relative to the market benchmark return/risk.
3. Beta - is the measure of Correlation between the return of an investment product w.r.t. market benchmark return.
4. Sharpe ratio - is the measure of excess return over the risk free rate and its relative scale in comparison the standard deviation of the return of the investment.
$$Sharpe\ Ratio = (E(r_p) - E(r_f))/\sigma_p$$
5. R-square value - tells us about the accuracy of our model definition based on the percentage of total variation explained by our model.

Above risk factors are easily visible in the case of Capital Asset Pricing Model(CAPM), Arbitrage Pricing model or even Multi- factor pricing model like Fama French three factor model.

Based on **FAMA FRENCH 3 Factor model[7]**, we state below:
$$R_i - R_f = \alpha + \beta_1(R_m - R_f) + \beta_2 * SMB + \beta_3 * HML + \epsilon$$
Here,
$\alpha = R_p - (R_f + \beta_p(R_m - R_f))$
$R_f = Risk\ free\ rate$
$R_i = return\ of\ security\ i$
$R_m = return\ of\ market\ benchmark$
SMB = Size premium (small minus big)
HML = value premium (High minus low)
$\epsilon = stochastic\ error\ term$

Another example of correlation and its practical impact is the belief that increase in global economic integration has led to higher correlation between economies[10]. Or we can also do this study in a comparatively micro scale, if we take a single economy in consideration, then the study of linkages between different industrial sectors and the study of supposed causality between these industries can give us crucial insight into how exactly the given economy is functioning.

Correlation can also help us study the causality between a firm's financial statements and stock prices or even the derivatives having the company stock as the underlying.

Correlation as a concept helps a lot in the ease of devising efficient diversification strategies[6]. The basic idea is that our current portfolio's correlation should at least be quite low, whether positive or negative(preferably), so as to fulfill the criterion of diversification, that is, the new portfolio has lower variance than the individual constituents of the portfolio.

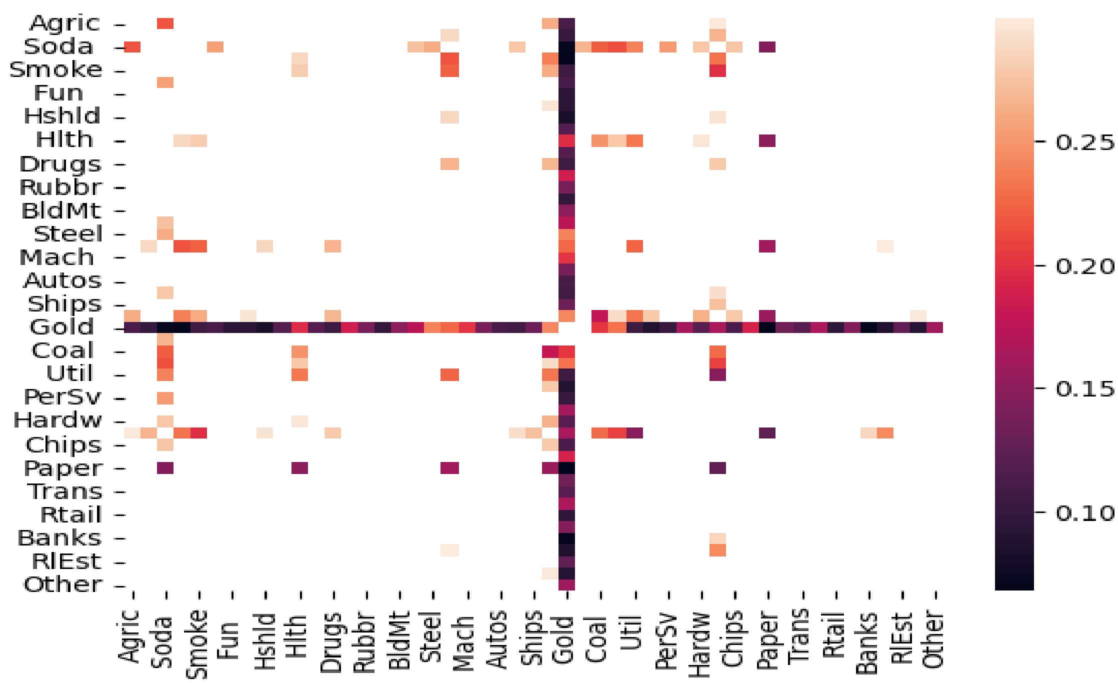
The idea of discussing the second theme here is that it is crucial to understand that the market is inherently dynamic in nature. Using these models comes with the belief that no matter what history repeats itself in the market context. That is why market/equity research is a field which needs constant monitoring. And there is no such thing as 'One model for all situations' as of right now. It may become a reality in future with improvement in the field of Data Science and Artificial Intelligence. In today's era of High frequency trading, these automated algorithms have been found to have had a drastic impact on the market on several occasions. If these model failures are not caught very fast then it can cause exponentially large losses. For example, Knight Capital's errant software sent it on a 1 hour shopping spree worth roughly $7Bn which it could not afford[8]. Another example would be 2010's Flash crash[3]. Any sort of Black swan event would basically invalidate the currently accurate models because such an event will basically have quite an impact on the market defining parameters.

Illustration with Real World Data

1. Fama-French 3 Factor model

a. Industry wise Correlation

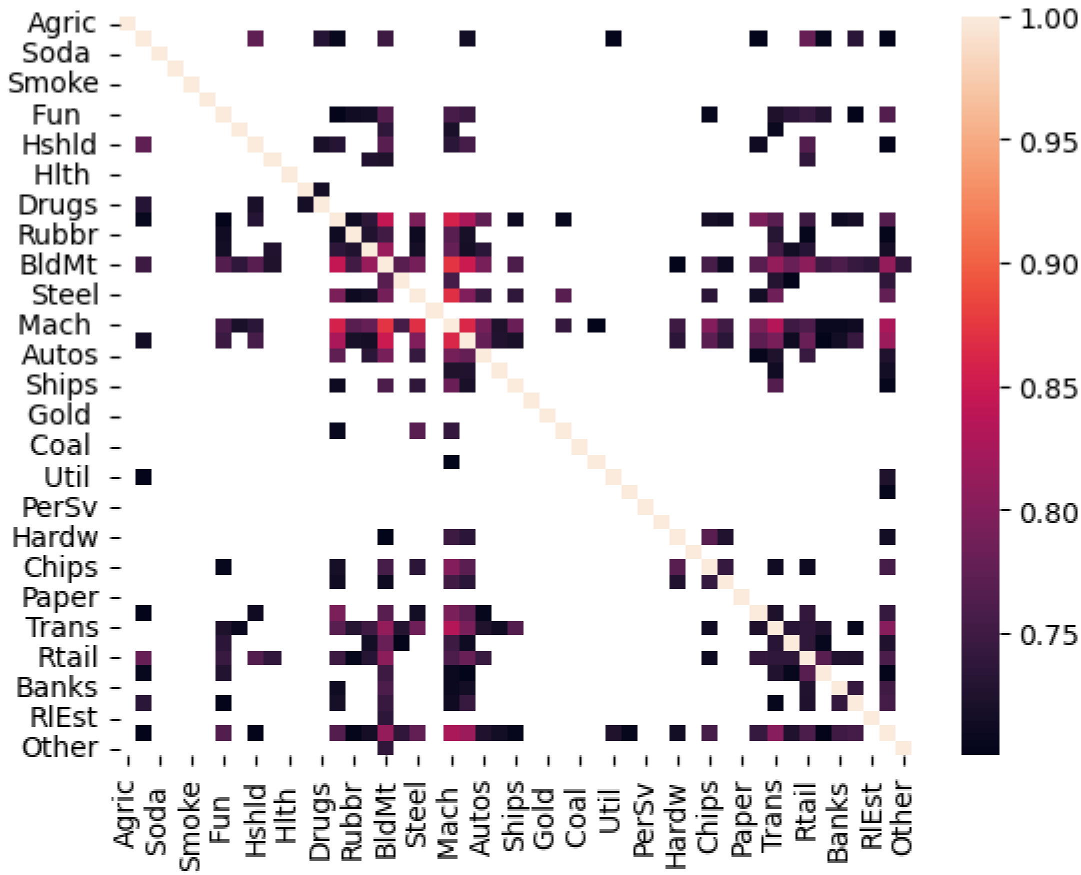
1) Low correlation cases



Most Identifiable fact here is that gold has quite a low correlation with almost all of the other industries displaying the fact how this market still functions on the most basic demand and supply basis.

We can also observe low correlation between agriculture and Soda industry.

2) High Correlation cases



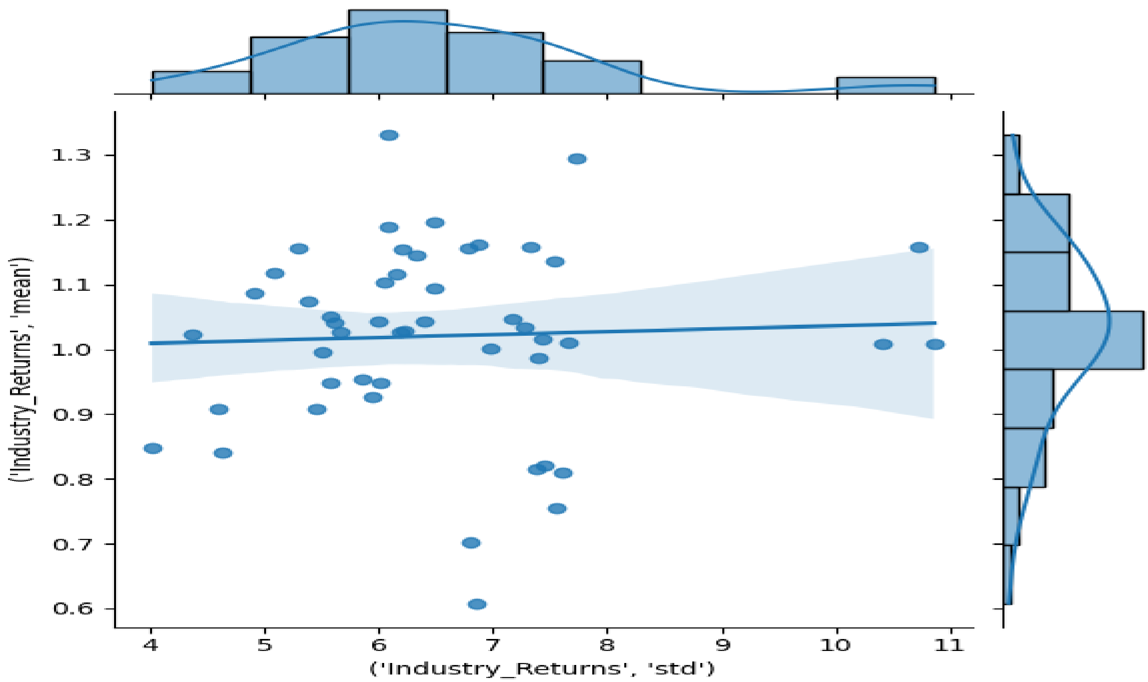

```
In [198]: #Correlation analysis
          value_wted_returns.iloc[:,1:].corr()['Agric'].sort_values().tail(10)

Out[198]: Steel     0.566235
          Insur     0.568157
          Food      0.574797
          Fin       0.592978
          Chems     0.601359
          Trans     0.603829
          BldMt     0.613849
          ElcEq     0.614002
          Mach      0.622234
          Agric     1.000000
          Name: Agric, dtype: float64
```

It is clearly identifiable that the Agriculture sector has high correlation with almost all sectors, that is how basic and necessary it is. Same is observed in case of coal and corresponding industries impacted by it.

2. Regression results

a. Industry returns and standard deviations have a positive slope. Proof of the fact that risk and reward go hand in hand. Deviations are due to excessively large and diverse dataset both in case of time line and variety of industry.



b. We found α to be significant for below industries. It goes to show the slight independence of their risk/return characteristics from the overall economic system.

In [200]:
```
df_t_test['Reject_Ho']=np.where((df_t_test['t_score']>t_stat_critical_value)|(df_t_test['t_score']<-t_stat_critical_value),1,0)
df_t_test[df_t_test['Reject_Ho']==1]
```

Out[200]:

Industry	alpha	t_score	Reject_Ho
Smoke	0.532657	2.271469	1
Drugs	0.409976	1.763794	1
Steel	-0.483452	-1.933346	1
RlEst	-0.623108	-2.511268	1

c. Overall Industry analysis

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 re_bar   R-squared:                       0.504
Model:                            OLS   Adj. R-squared:                  0.504
Method:                 Least Squares   F-statistic:                 1.172e+04
Date:                Tue, 07 Feb 2023   Prob (F-statistic):               0.00
Time:                        21:45:54   Log-Likelihood:            -1.0285e+05
No. Observations:               34643   AIC:                         2.057e+05
Df Residuals:                   34639   BIC:                         2.057e+05
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0163      0.026     -0.631      0.528      -0.067       0.034
Mkt-RF         1.0278      0.006    169.123      0.000       1.016       1.040
HML            0.1396      0.009     15.861      0.000       0.122       0.157
SMB            0.2279      0.009     26.130      0.000       0.211       0.245
==============================================================================
Omnibus:                     7824.968   Durbin-Watson:                   1.576
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           142739.118
Skew:                           0.612   Prob(JB):                         0.00
Kurtosis:                      12.869   Cond. No.                         4.78
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

All three β's are significant with market risk premium having the most largest proportional impact. We have Adj- R-Squared score of 0.5 which is pretty good considering the diversity of our data.

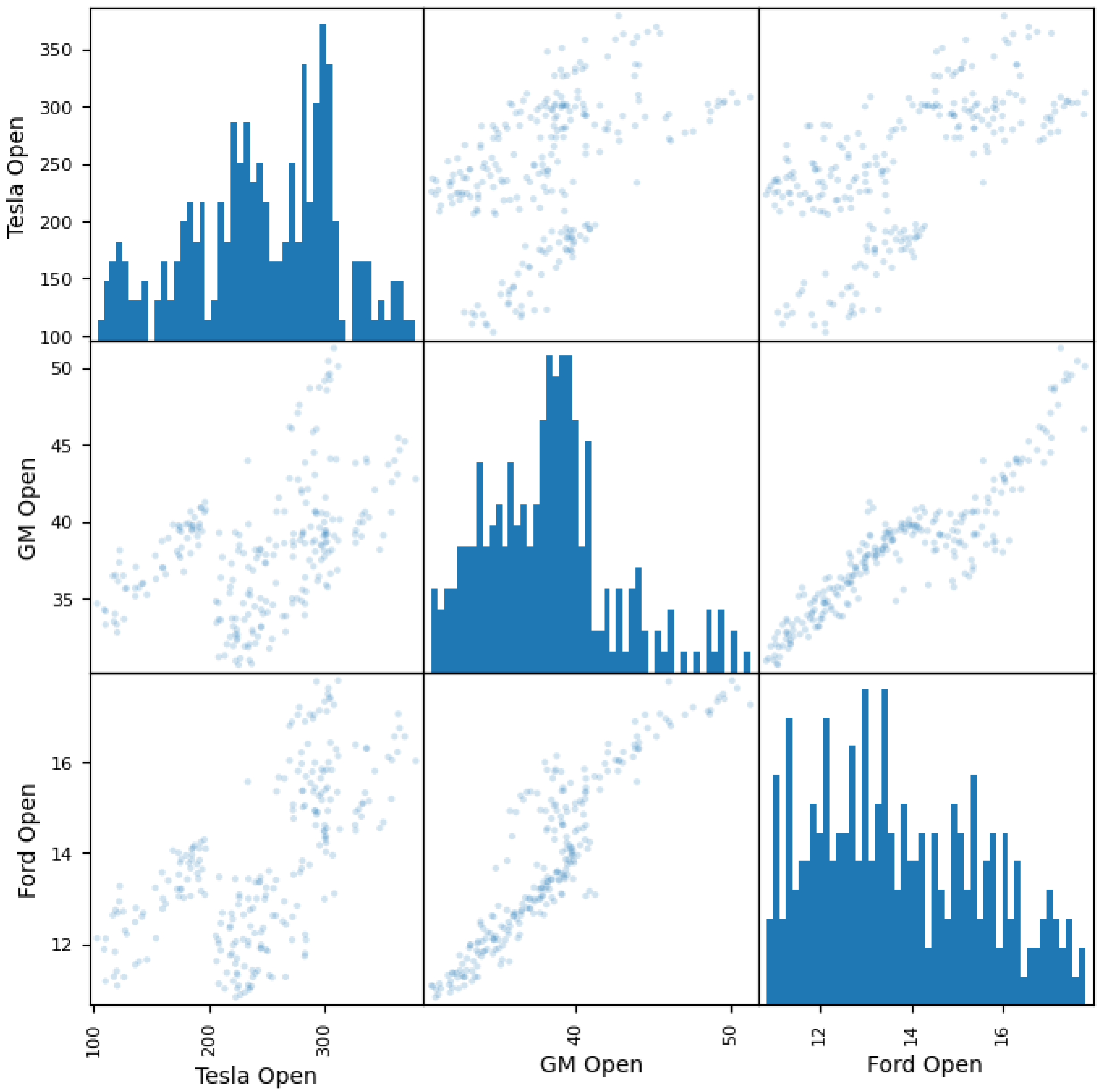
Correlation between Tesla, Ford and GM stock prices.

We conclude that there is a high degree of correlation between Ford and GM. Tesla stock prices are comparatively less correlated with both Ford and Gm.

```
In [237]: tesla['Open'].corr(ford['Open'])
Out[237]: 0.6315345206188344

In [238]: tesla['Open'].corr(gm['Open'])
Out[238]: 0.4539597606104765

In [239]: gm['Open'].corr(ford['Open'])
Out[239]: 0.9065758928982761
```

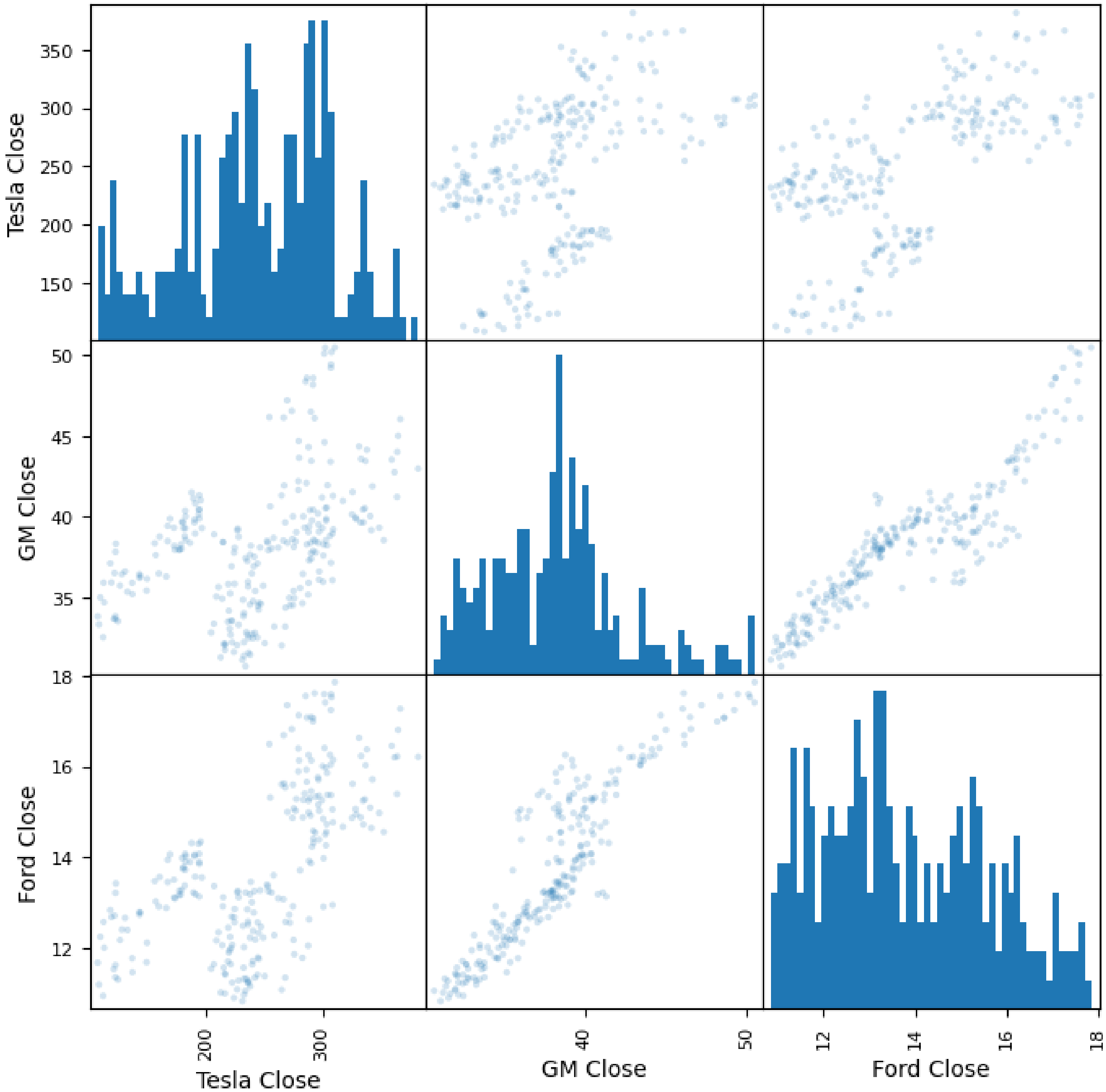


Further analysis also showed that tesla is most volatile followed by Ford and then GM given their respective price levels.

```
In [251]: comparison.describe().loc['mean',:]/comparison.describe().loc['std',:]

Out[251]: Tesla Open    3.900512
          GM Open       9.155948
          Ford Open     7.624141
          dtype: float64
```

Similar pattern of correlation is also observed for 1 day returns of the three stocks, as visible from below scatter matrix.



Literature Review

1. **2010 Flash Crash -** (Andrei Kirilenko, Albert S. Kyle, Mehrdad Samadi, Tugkan Tuzun, The Flash Crash: The Impact of High Frequency Trading on an Electronic Market) [3]

This is a prime example of the case where High Frequency Trading can accelerate both the crash and recovery in the market. On **6th May 2010,** almost all the leading US stock indices across all the product classes tumbled and then later rebounded within the hour. This event caused a loss of roughly 1 trillion USD. The markets were already in quite a turbulent downtrend state which got accelerated due to Emini S&P contracts and the subsequent selling orders executed by the HFT trading algorithm. This paper suggested redesigning of the market itself rather than imposing taxes and order limits, which add to market inefficiency. HFT is a game where the firms aim to minimize latency which can sometimes lead to such unforeseen events. It is suggested that inclusion of brief trading.pauses would allow for the market to stabilize from time to time.

2. **Correlation in economies in times of Crisis [11]**

In this paper, the author investigated most of the largest market downturn events ranging from 1987 (Black Monday), 1998 (Russian crisis), 2001 (bursting of dot-com bubble), and 2008 (Subprime Mortgage Crisis). They aimed to capture the correlation matrices of some of the major economies using the eigenvalues and eigenvectors in order to try and link the consecutive high volatility across two economies based on their correlation with each other. The study utilizes the 'Random matrix theory' This theory helps them simulate, adjust and finally validate the correlation matrix. It also helped in setting up hierarchical linkages between economies. Based on the research, the author gave the belo conclusions:

a. Similar behavior of markets in case of high volatility scenarios.
b. Certain indices distributions were not independent but exhibited the properties of joint distributions.
c. Relation of the average correlation and average volatility increased with increase in Spearman rank correlation test, instead of Pearson Correlation, indicating non-linear traits. Covariance between the same can be used as a good indicator of when the crisis occurs.

3. **Correlation between global stock markets [10]**

In this article, they investigated whether stock markets are becoming more correlated as the degree of international integration and business has increased over the past few years. They considered the stock market data from Japan, Europe, Asia Ex Japan and

US from the library of Kenneth R. French. They observed a consistent rise in correlation between US and other stock markets from 1990 to 2008, followed by no significant change thereafter. It was also concluded that there are more economic ties between the US and Europe as they consistently had a higher correlation. They also did correlation analysis on factor,m value, size and momentum, basis too. It concluded the same thing about the US-Europe economic ties. In conclusion, they stated any discrepancy in correlation between different regions is due to time difference otherwise globalization and ease of financial flow has led to higher correlation across Global stock markets.

References

1. https://insight.factset.com/the-fundamental-differences-of-fundamental-and-statistical-risk-models
2. https://www.investopedia.com/terms/r/riskmeasures.asp
3. https://www.epfl.ch/schools/cdm/wp-content/uploads/2019/03/Kyle-paper.pdf
4. Jablonowski, M. (2006). Statistical Risk and its Treatment. In: Precautionary Risk Management. Palgrave Macmillan, London. https://doi.org/10.1057/9780230627659_1
5. https://corporatefinanceinstitute.com/resources/equities/2010-flash-crash/
6. https://www.investopedia.com/articles/financial-theory/09/uncorrelated-assets-diversification.asp
7. https://www.investopedia.com/terms/f/famaandfrenchthreefactormodel.asp
8. https://www.henricodolfing.com/2019/06/project-failure-case-study-knight-capital.html
9. https://www.bis.org/publ/qtrpdf/r_qt1409v.htm
10. https://caia.org/blog/2021/09/19/are-stock-markets-becoming-more-correlated
11. https://www.researchgate.net/publication/48202792_Correlation_of_financial_markets_in_times_of_crisis