Analyzing Relationships Between COVID-19 Trends, Mobility, and the S&P 500 Index

1. Introduction

The COVID-19 pandemic caused unprecedented disruptions to global financial markets and societal behavior. As governments implemented lockdowns and restrictions, mobility patterns drastically changed, businesses adapted their operations, and stock markets experienced significant volatility. This research explores the intricate relationships between COVID-19 case trends, population mobility metrics, and the performance of the S&P 500 index during this unique period in history.

The pandemic provides a rare opportunity to examine how public health emergencies simultaneously affect financial markets and human behavior. By analyzing mobility data alongside market performance, we can better understand how changes in human movement patterns correlate with market fluctuations during crisis periods. This knowledge could help investors, policymakers, and businesses better prepare for and respond to future pandemic events or similar societal disruptions.

This project utilizes data science techniques to integrate multiple datasets, including Google Mobility reports, COVID-19 case statistics, and S&P 500 index values. We aim to uncover meaningful patterns and potentially predictive relationships between these variables through statistical analysis and machine learning approaches.

2. Data

This project integrates data from multiple sources to analyze the relationships between COVID-19 metrics, mobility patterns, and stock market performance.

2.1 Data Sources

The primary dataset was created by merging data from the following sources:

1. Google Mobility Reports: Data showing movement trends across different categories of places, downloaded from Google's COVID-19 Community Mobility Reports (https://www.google.com/covid19/mobility/). This data tracks percentage changes in visits to various location categories compared to a pre-pandemic baseline.

2. COVID-19 Statistics: Daily new case counts and deaths for the United States, obtained from Worldometers (https://www.worldometers.info/coronavirus/country/us/).

3. S&P 500 Index Values: Daily S&P 500 index closing values and returns, sourced from the Federal Reserve Economic Data (FRED) database

(https://fred.stlouisfed.org/series/SP500).

The resulting dataset covers the period from early 2020 through late 2020, capturing the initial waves of the pandemic and market response.

2.2 Data Dictionary

Column Name	Data Type	Description	Source
Date	Date	Date of observation (YYYY-MM-DD)	All
SP500_Close	Float	S&P 500 index closing value	FRED
Is_trading_day	Boolean	TRUE if the date is a trading day for the S&P 500, FALSE otherwise	FRED
New_Cases	Integer	Number of new confirmed COVID-19 cases	Worldometers

New_Deaths	Integer	Number of new COVID-19 deaths	Worldometers
New_cases_7d_avg	Float	7-day moving average of new COVID-19 cases	Worldometers
New_cases_12d_avg	Float	12-day moving average of new COVID-19 cases	Worldometers
Retail_and_recreation_percent_change_from_baseline	Float	% change in visits to retail/recreation locations from baseline	Google Mobility
Grocery_and_pharmacy_percent_change_from_baseline	Float	% change in visits to grocery/pharmacy locations from baseline	Google Mobility
Parks_percent_change_from_baseline	Float	% change in visits to parks from baseline	Google Mobility

Transit_stations_percent_change_from_baseline	Float	% change in visits to transit stations from baseline	Google Mobility
Workplaces_percent_change_from_baseline	Float	% change in visits to workplaces from baseline	Google Mobility
Residential_percent_change_from_baseline	Float	% change in time spent at residential locations from baseline	Google Mobility
New_Cases_lag3	Integer	New COVID-19 cases, lagged by 3 days	Worldometers (Derived)
New_Cases_lag7	Integer	New COVID-19 cases, lagged by 7 days	Worldometers (Derived)
New_Deaths_lag3	Integer	New COVID-19 deaths, lagged by 3 days	Worldometers (Derived)

New_Deaths_lag7	Integer	New COVID-19 deaths, lagged by 7 days	Worldometers (Derived)

3. Analysis

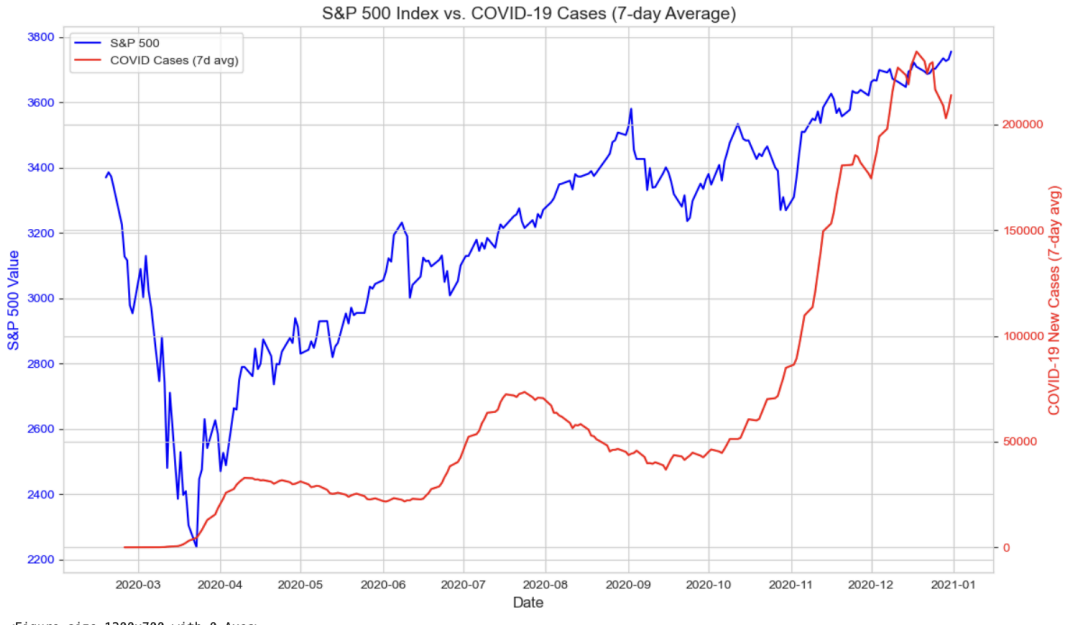
3.1 Exploratory Data Analysis

Initial exploratory analysis revealed several key patterns in the data. We observed that mobility metrics experienced dramatic changes at the onset of the pandemic, with retail, workplace, and transit mobility showing sharp declines while residential mobility increased significantly. The S&P 500 saw an initial steep drop followed by a recovery despite continued COVID-19 case growth.

Basic statistical analysis showed that the S&P 500 daily returns had a mean of 0.000486 with a standard deviation of 0.014033. COVID-19 cases showed high variability with a mean of 76,257 daily new cases and a standard deviation of 109,918 during the period analyzed.

Missing value patterns were consistent with the nature of lagged variables, and these missing values were addressed during data preparation. Some significant relationships were identified between mobility metrics and S&P 500 data, with varying correlation strengths.

VISUALIZATION 1: Time series chart showing S&P 500 and COVID-19 cases over time



3.2 Hypothesis Testing

We tested several hypotheses to explore the relationships between mobility metrics, COVID-19 cases, and market performance:

Hypothesis 1: Workplace mobility is correlated with S&P 500 returns

- Correlation: -0.185, p-value: 0.0046
- Interpretation: There is a statistically significant negative correlation between workplace mobility and market returns, suggesting that as workplace mobility decreased (more people working from home), the market tended to show slightly positive performance.

Hypothesis 2: S&P 500 returns differ on days with high vs. low COVID cases

- t-statistic: 0.904, p-value: 0.3672
- Mean return on high COVID days: 0.0027
- Mean return on low COVID days: -0.0026
- Result: Failed to reject null hypothesis
- Interpretation: No significant difference in returns between high and low COVID days was detected in the data.

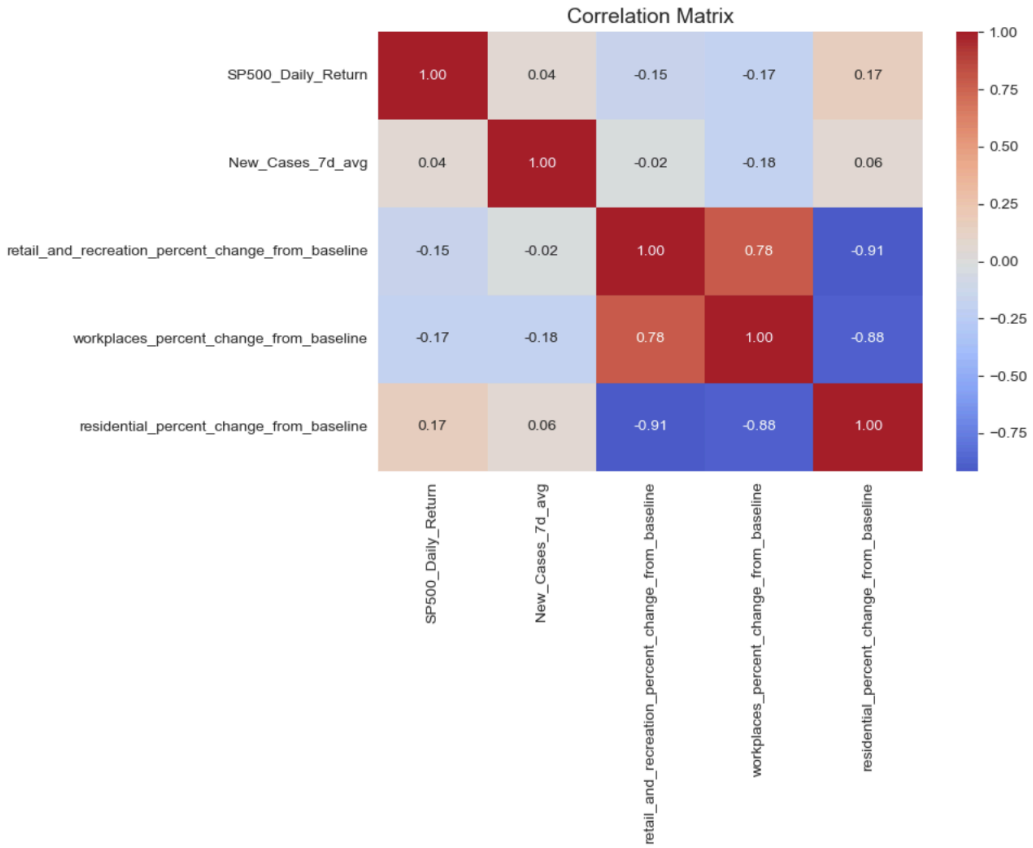
Residential Mobility vs. S&P Returns:

- Correlation: 0.178, p-value: 0.0064

- Interpretation: A significant positive relationship exists between increased residential mobility (more time spent at home) and market returns.

Analysis was performed on 233 trading days with complete data.

VISUALIZATION 2: Correlation matrix heatmap showing relationships between key variables]



<Figure size 1200x700 with 0 Axes>

3.3 Machine Learning Models

3.3.1 Regression Models

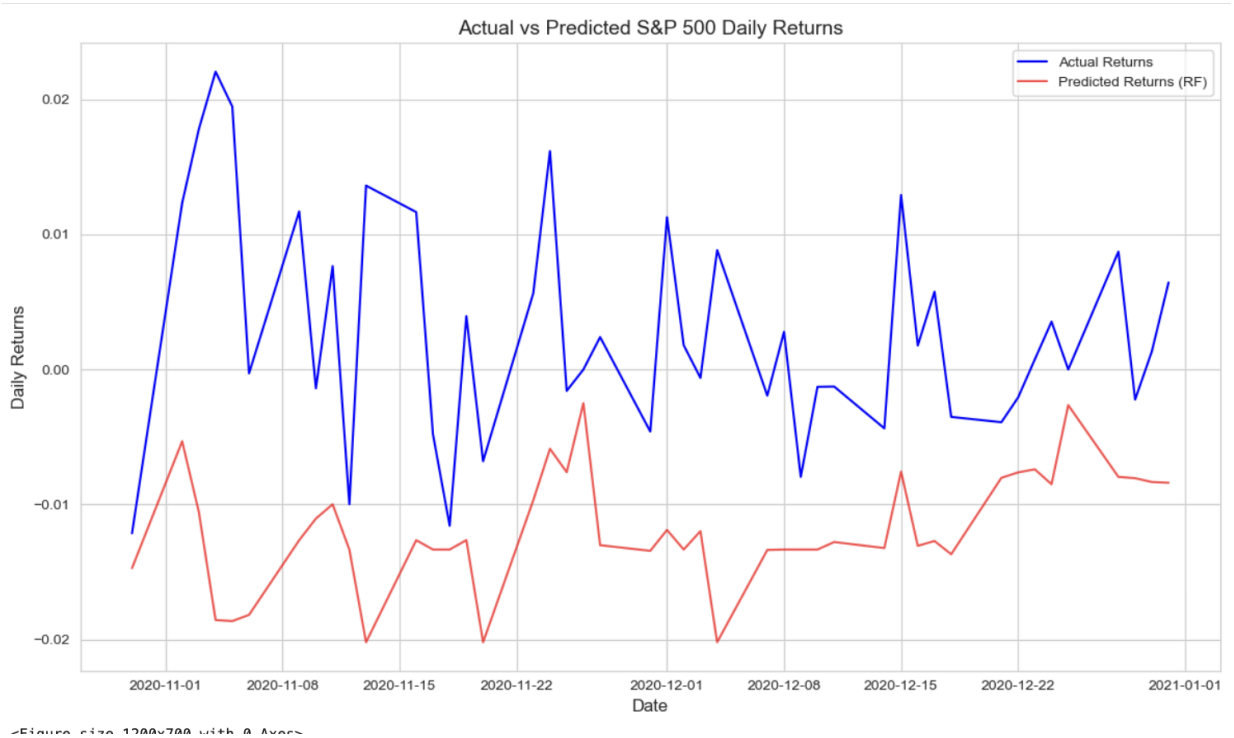
We implemented several regression models to predict S&P 500 values based on mobility metrics and COVID-19 data:

1. Linear Regression
    - R^2 Score: -1.316
    - Poor performance suggests non-linear relationships
2. Ridge Regression
    - R^2 Score: -1.316
    - Similar performance to linear regression

3. Random Forest Regressor
    - R^2 Score: -1.5285
    - Despite handling non-linearities better, it still underperformed

The negative R^2 scores indicate that these models performed worse than simply using the mean as a predictor, suggesting complex relationships that weren't adequately captured by the features used.

VISUALIZATION 3: Actual vs Predicted S&P 500 Daily Returns



3.3.2 Classification Models

We also developed classification models to predict whether the market would go up or down based on our features:

1. Logistic Regression
    - Accuracy: 0.5333
    - F1 Score: 0.6057
2. Random Forest Classifier
    - Accuracy: 0.4444
    - F1 Score: 0.4741
3. Support Vector Machine
    - Accuracy: 0.5111

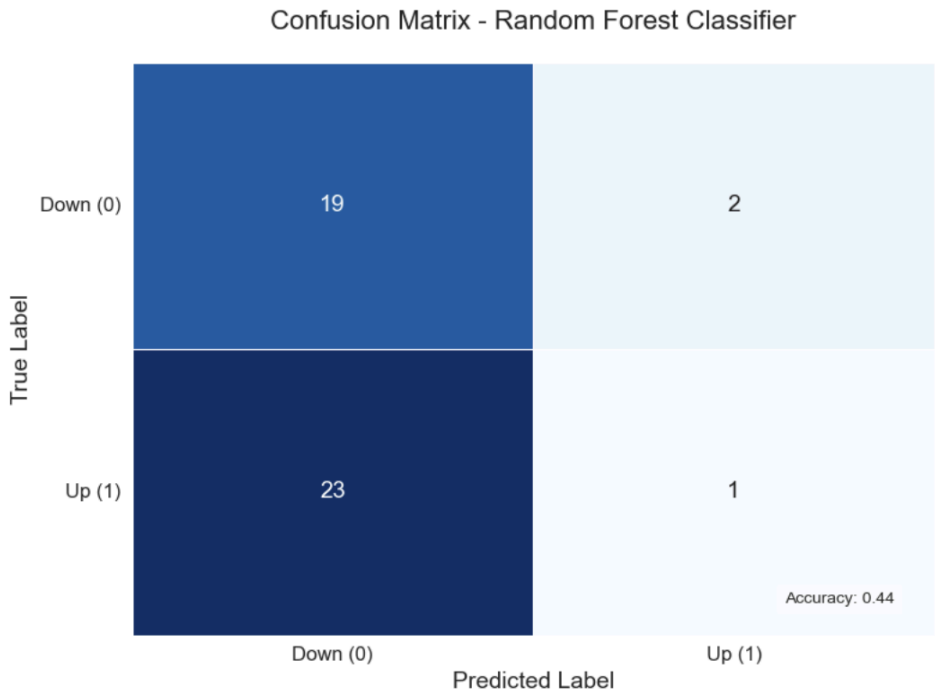
- F1 Score: 0.3529

Logistic Regression performed best among classification models, though its performance was only moderately better than random chance.

Feature importance analysis from the Random Forest model revealed that the most important features were:

1. New_Cases_7d_avg (0.290758)
2. New_Cases_7d_avg (0.278969)
3. workplaces_percent_change_from_baseline (0.170835)
4. retail_and_recreation_percent_change_from_baseline (0.150851)
5. residential_percent_change_from_baseline (0.095794)

VISUALIZATION 4: Confusion matrix for the best classification model



3.4 Time Series Analysis

We conducted time series analysis to forecast S&P 500 values using ARIMA models. The best ARIMA model (1,0,1) achieved an RMSE of 305.41 on the test set. This model provides useful baseline forecasting capabilities but has limitations during high-volatility periods.

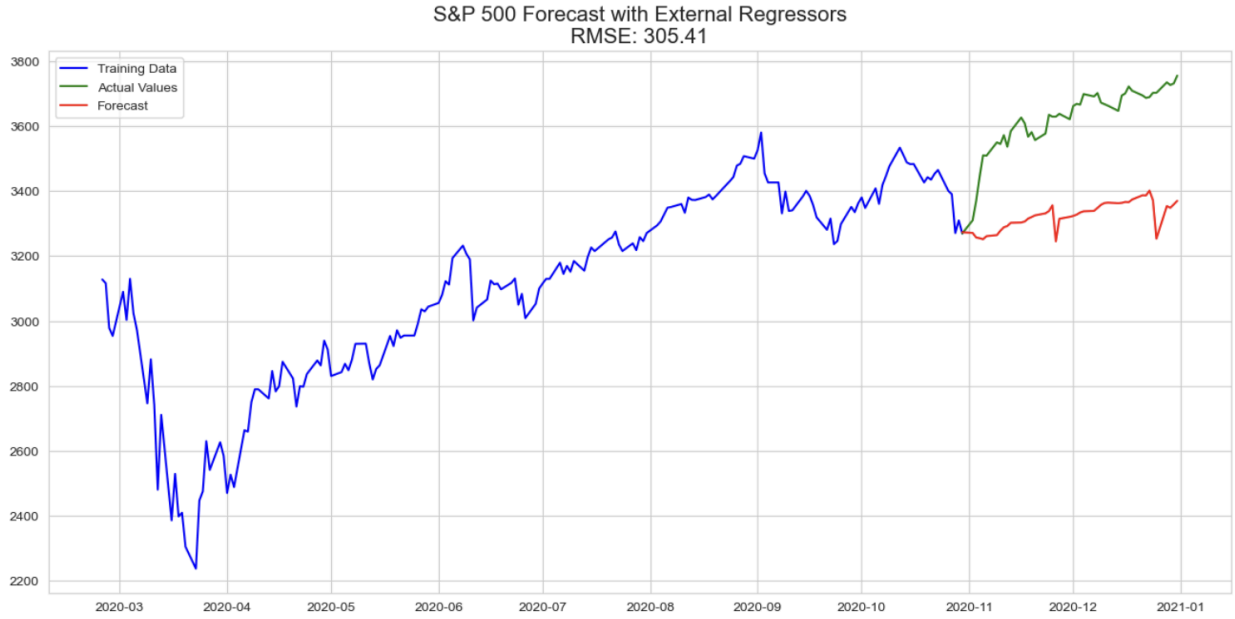
The SARIMA modeling revealed:

- Best ARIMA Parameters: (0, 1, 2)

- Improved ARIMA RMSE: 359.81

The model diagnostics showed that while the residuals were generally well-behaved, some patterns suggested room for improvement with more complex models or additional features.

VISUALIZATION 5: ARIMA model forecast vs actual S&P 500 values



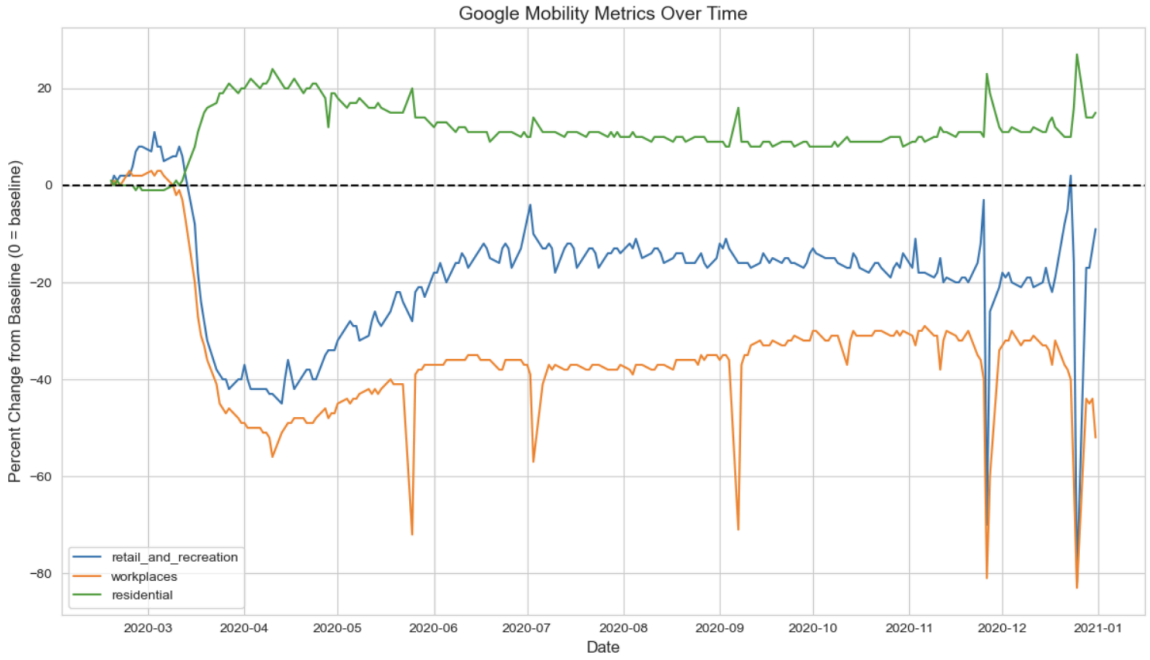
3.5 Correlation Analysis

Key correlations between S&P 500 Returns and mobility metrics:

- retail_and_recreation_percent_change_from_baseline: -0.155
- workplaces_percent_change_from_baseline: -0.175
- residential_percent_change_from_baseline: 0.168
- New_Cases_7d_avg: 0.043

These correlations reveal that increased time at home (residential mobility) was positively associated with market performance, while decreased retail activity and workplace attendance showed negative correlations with market returns.

VISUALIZATION 6: Google Mobility Metrics Over Time chart showing the different mobility categories



4. Conclusion

Our analysis of COVID-19 data, mobility metrics, and S&P 500 performance reveals several vital insights about the relationship between pandemic behaviors and market dynamics.

The findings suggest that there is indeed a measurable relationship between mobility metrics and market performance, with residential mobility showing a positive correlation with market returns and workplace mobility showing a negative correlation. This indicates that as people spent more time at home and less time at workplaces during the pandemic, the market tended to perform better, contrary to what might be intuitively expected.

Surprisingly, direct correlations between COVID-19 case numbers and market performance were weaker than expected, suggesting that markets were more responsive to changes in human behavior (as measured by mobility) than the raw case counts themselves. This might indicate that the market was pricing in the economic impact of behavioral changes rather than responding directly to public health metrics.

The predictive models developed in this study demonstrate the challenges of forecasting market movements even with novel data sources. While our models achieved limited predictive success, they highlight mobility data's potential value as a factor in market analysis during crisis periods.

This study's limitations include potential inconsistencies between trading and non-trading days and the extraordinary nature of the pandemic period, which may limit generalizability to normal market conditions.

Future work could expand on this research by incorporating sentiment analysis from news and social media, including additional economic indicators, and extending the time frame to examine how these relationships evolved in later phases of the pandemic. Additionally, more sophisticated machine learning approaches might yield improved predictive performance, particularly deep learning models that can better capture temporal dependencies.

This research contributes to understanding how unprecedented societal disruptions affect financial markets. It highlights the potential value of alternative data sources, such as mobility metrics, in economic analysis during crisis periods.