FINA4359: Big Data Analytics Applied

Towards Quantitative Finance

Final Project - Option 2

Forecasting Securities’ Return of

US Leisure & Travel Industry using Alternative Data

**Group Members**

|  |  |
| --- | --- |
| **Amruthraghav Gopalakrishnan**  **Wu Zijing**  **Chan Ming Ho** | **3035436674**  **3035556644**  **3035483794** |

**Table of Content**

[**1. Motivation**](#_ibhgr7b6ye1v) **3**

[**2. Alternative Data of Choice**](#_5bj2d6xzfm9b) **3**

[**3. JETS ETF**](#_c98k6ys7jqyb) **5**

[**4. Predicting JETS ETF using number of air-travel passengers**](#_bdac2c7fynnr) **5**

[4.1 TSA checkpoint travel numbers at US Airports](#_cyuna02nki7i) 5

[4.2 Data Scraping & Collection](#_69hupll56fvh) 6

[4.3 Regressing JETS ETF return against weekly TSA changes](#_6skjlgo7vi4m) 7

[4.4 Evaluation/Interpretation of the regressions](#_bwdxfjp8d3i7) 8

[**5. Predicting JETS ETF return using Trips by Distance Dataset**](#_ff1p9z7jtvfc) **9**

[5.1 Description](#_7hkgccmw1ih6) 9

[5.2 Collection](#_vglko37itpj3) 10

[5.3 Results of regression](#_wnc85nsje5qp) 11

[5.4 Interpretation of regression results](#_783tc9nofbu1) 13

[**6. Predicting JETS ETF and Travel Website Stock’s return using Online Web Traffic - Google Trends and Alexa Rank**](#_c4nqkl21rd9o) **15**

[6.1 Google Trends](#_3ssfagtzwtbs) 15

[6.1.1 Data Scraping /Collection](#_g0cbt0b6j6tf) 16

[6.1.2 Results of the regressions on JETS and BKNG(Quantitative)](#_nk5j0v1plmxf) 17

[6.1.3 Interpretation of the regressions(Qualitative)](#_6803a82gdon) 19

[6.2 Alexa Rank - Amazon Web Traffic](#_43o3xzujtyu6) 19

[6.2.1 Data Scraping /Collection](#_qjbk359icfpy) 19

[6.2.2 Regressing Stock Returns from JETs ETF and Individual Travel Website Stocks](#_2ks2894bmpy0) 20

[6.2.3 Results of the regressions](#_voq804rt3stc) 20

[6.2.4 Evaluation/Interpretation of the regressions](#_tvm8bd2cvls2) 23

[6.3 Why does amazon web traffic have a significant role in stock return compared to google trends?](#_ft1l87hhjbu9) 23

[**7. Predicting Stock returns of Companies with Theatres as Primary Business using Box Office Revenue**](#_6hqtjmnsm6pp) **24**

[7.1 Data Scraping](#_5ofeox1ka4ly) 24

[7.2 Regressing Stock Returns of Companies with Theatres as the primary business](#_teo5powswnw) 26

[7.3 Results of the regressions](#_9zura5xo1013) 26

[Future Work](#_dujqga39dodr) 28

# 1. Motivation

The COVID-19 pandemic (‘C19’) has wreaked havoc on both Wall Street and main street. This is particularly true for any businesses within the leisure and travel industry where physical encounterment is the fundamental core of its business model. As of today, both their financial results and trading performance have yet to recover from its pre-C19 level, albeit seeing gradual improvement since the beginning of 2021 underpinned by a fast vaccination rollout. Investors eyeing opportunities to “buy-the-dip” in the industry might want to know whether now is the inflection point when stocks from the sector are ready to take off. Our project intends to investigate this subject.

Our scope of research focuses on establishing connections between securities within the industry traded on US exchanges and alternative datasets. We will develop codes that scrap or obtain readily available alternative data either free-of-charge or at costs typical retail investors can afford. Such alternative datasets are high-frequency, real-time, non-mainstream indicators that do not often get reported or receive minimal media attention (i.e. certainly not weekly jobless claims, nonfarm payroll, etc). We then see whether such alternative datasets tracking the real economy can be applied as a (lead) indicator to predict stock prices with linear regression analysis.

# 2. Alternative Data of Choice

As suggested by its name, alternative data are somewhat under-appreciated datasets that potentially carry precious information about a particular firm, sector or market. The ability to provide a different perspective compared to conventional data enables alternative data to provide timely and unique insights regarding investment opportunities. These datasets are often used by hedge funds and portfolio managers, giving them a unique perspective and helps in generating alpha - excess return independent of the market.

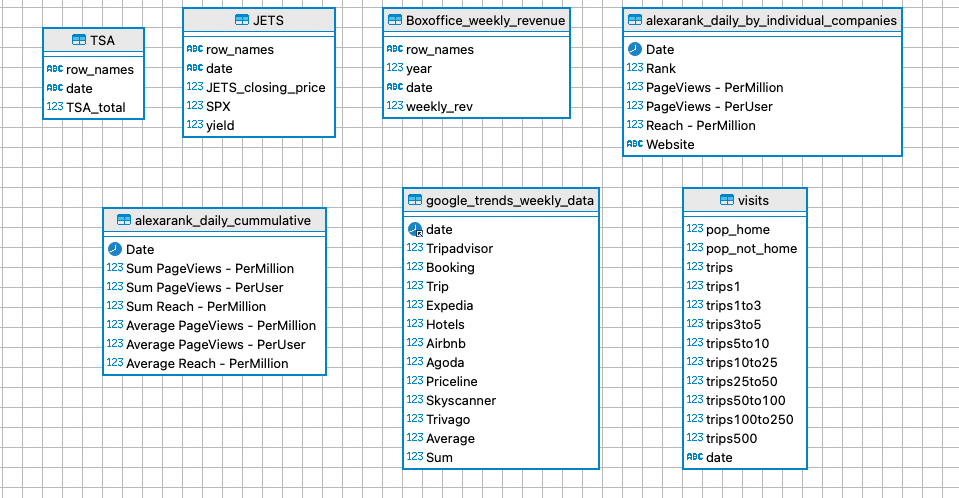
Most of the alternative datasets are collected via automation techniques such as scraping or API calls from corresponding data vendors in our projects. It ensures that data can be updated regularly and empowers anyone who has access to the code to extend our work with as little hassle as possible.

The datasets of our choice are a broad representation of how the entire travel and leisure industry might mount its comeback. Such datasets do not clash/overlap with each other, where each dataset represents a unique feature/dimension of the industry. Together, they should give a comprehensive and compelling picture concerning the travel and leisure industry.

A summary of the alternative datasets related to the travel and leisure industry included in this report is shown as follows. Detailed descriptions on the collection methodology can be found under Sections 4 - 6.

|  |  |  |
| --- | --- | --- |
| **Data** | **Timeframe & Frequency** | **Transformations?** |
| Google Trends of various Online Travel Agencies (OTAs) | May 2016 - May 2021, Weekly | Individual and Cumulative (sum, mean) traffic calculations for weekly data |
| Alexa Rank from AmazonWeb-Traffic Data for Travel Websites | Dec 2019 - May 202, Weekly | Individual Daily and weekly data with three metrics- Pageviews Per Million(PPM), Reach Per View and Rank |
| Movie Box Office from ‘BoxofficeMojo’ | Dec 2019 - May 2021, Weekly | Weekly Revenue data already calculated |
| TSA checkpoint travel numbers at US Airports | Jan 2019 - May 2021, Daily | Compute trailing 7 days average number of passengers, then sum up the rolling average of the week and compute weekly percentage changes. |
| Trips by Distance | Dec, 2019 -- March 13th, 2021, CSV, Last updated April 17th 2021 | Weekly trip occurrences grouped by a distance of the trip |
| U.S. Global Jets ETF (NYSE:JETS) | May 2016 (Inception Date) - May 2021, Daily | Compute daily return from daily stock prices and transform into weekly return |

We have collected the afore-mentioned data into a SQL database with the following database schema:



# 3. JETS ETF

Our project focuses on predicting securities return within the leisure and travel industry, specifically the U.S. Global Jets ETF (NYSE: JETS). Stocks of individual travel and leisure companies are also examined where appropriate.

The U.S. Global Jets ETF tracks a basket of companies within the air travel industry, most airlines, airports and terminal-service companies. The list of companies JETS ETF tracked is shown as follows:

|  |  |
| --- | --- |
| **Company** | **Weighting** |
| Southwest Airlines (NYSE: LUV) | 10.92% |
| American Airlines (NASDAQ: AAL) | 10.54% |
| Delta Airlines (NYSE: DAL) | 10.06% |
| United Airlines (NASDAQ: UAL) | 9.57% |
| Allegiant Travel (NASDAQ: ALGT) | 4.99% |
| Spirit Airlines (NYSE: SAVE) | 4.86% |
| Alaska Air (NYSE: ALK) | 4.53% |
| JetBlue Airways (NASDAQ: JBLU) | 4.25% |
| Skywest (NASDAQ: SKYW) | 3.43% |
| Air Canada (TSE: AC) | 3.31% |

The top 10 are all airlines mainly serving the North American passenger market, accounting for 66.5% of the entire ETF. Note that non-North American companies and airlines such as Japan Airlines and China Southern Airlines are also included in the JETS ETF, albeit insignificant portions.

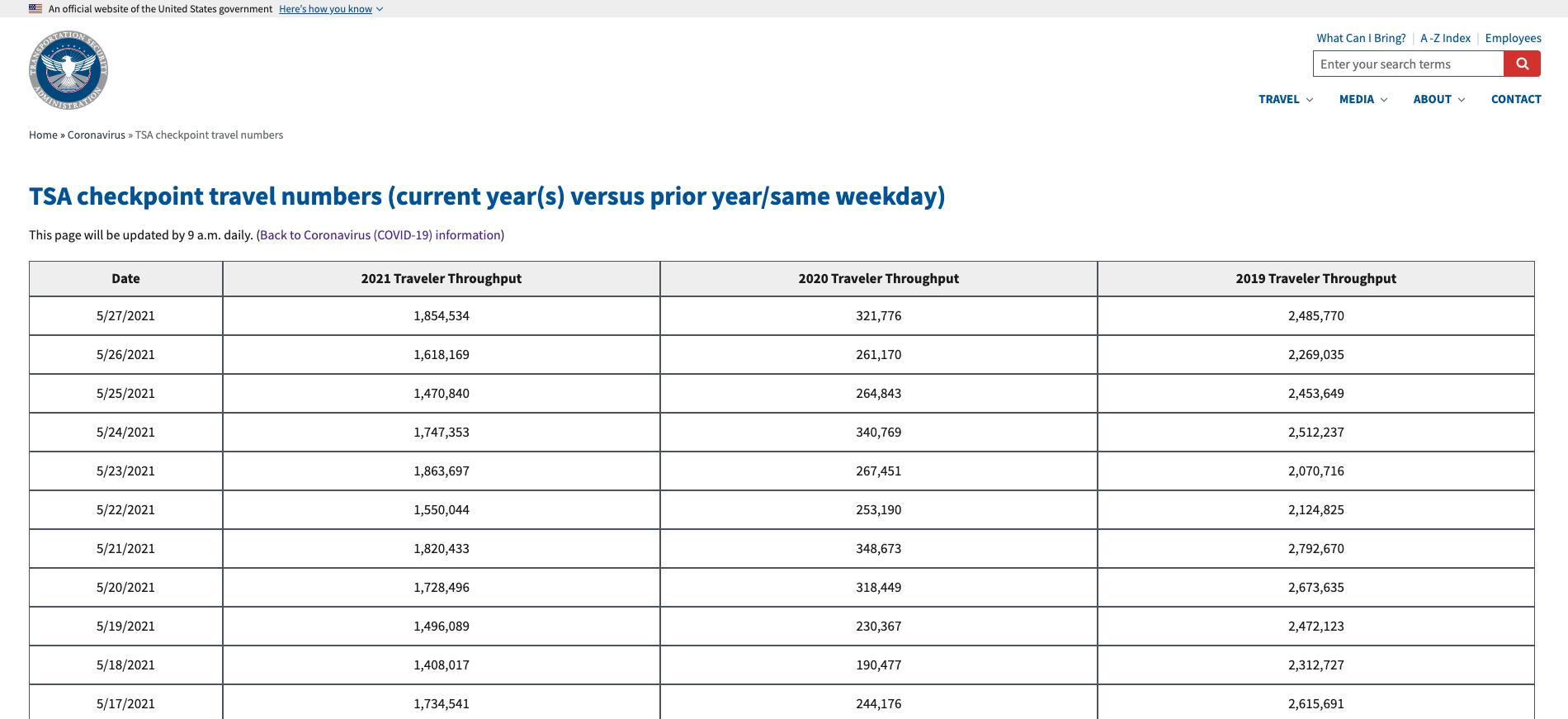
# 

# 4. **Predicting JETS ETF** using **the number of air-travel passengers**

## 4.1 TSA checkpoint travel numbers at US Airports

TSA is the security system at every airport in the US. Passengers either taking domestic and international flights are required to pass through the TSA checkpoint. At the beginning of the C19 pandemic, TSA started publishing the number of passengers who pass through the security checkpoint at a daily frequency and the corresponding number in 2019.

## 4.2 Data Scraping & Collection



TSA Checkpoint travel numbers, available at

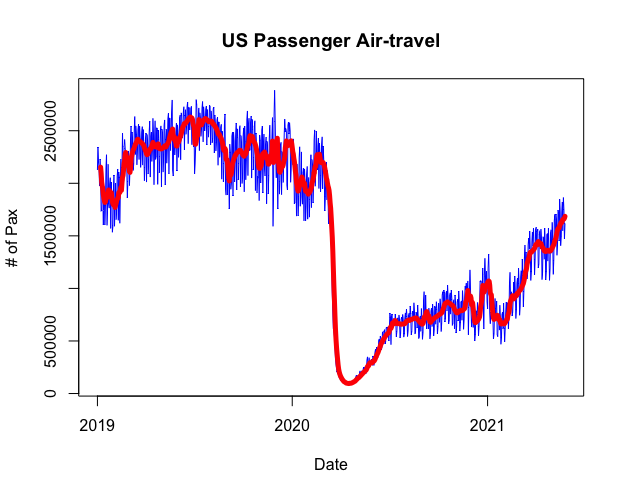
<https://www.tsa.gov/coronavirus/passenger-throughput>

As shown in the screen capture above, the TSA checkpoint travel number is stored as a HTML table. Hence, we develop a scrapping code to extract the data from the website. Here is the skeleton of the scrapping code:

1. Open the URL in a Google Chrome browser.
2. Install and make use of a Google Chrome extension “SelectorGadget” to identify the CSS Selector Path of the table elements.
3. (For R language) Install and load a web scraping package “rvest”. Paste the URL alongside the corresponding CSS Selector Path to extract elements of the table (i.e. date and TSA checkpoint number)
4. (For R language) Further processing and exporting into data.frame object for further analysis.

Before applying TSA checkpoint data in the regression analysis, it requires further transformation and processing. This is because 1) the raw data is of a daily frequency, and our intended regression is of weekly frequency (i.e. regressing weekly return in JETS against the weekly changes of TSA); 2) Daily data contains too many fluctuations and hence is of little value if noises are not smoothened out. Hence, we apply the following transformation:

1. Compute the trailing seven days rolling average (i.e. the rolling average on the date 1 Jan 2021 is computed by taking the simple average of the TSA daily travel number from 26 Dec 2020 to 1 Jan 2021, inclusive of both starting and ending dates)
2. Sum up the daily rolling average of that particular week to produce the total number of air-travel passengers. Then, compute weekly change.

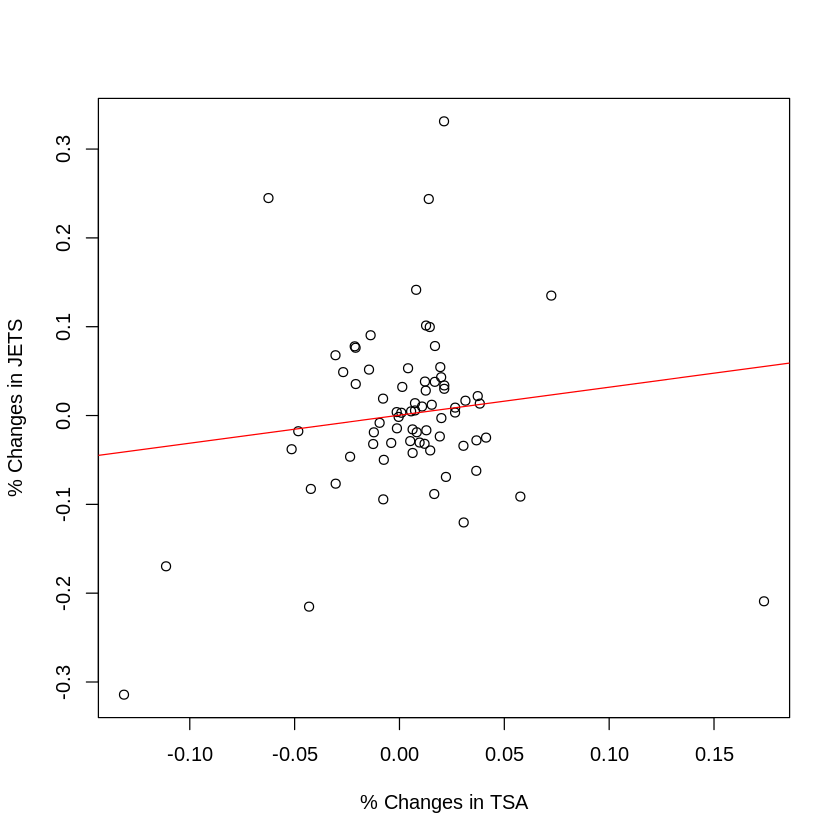
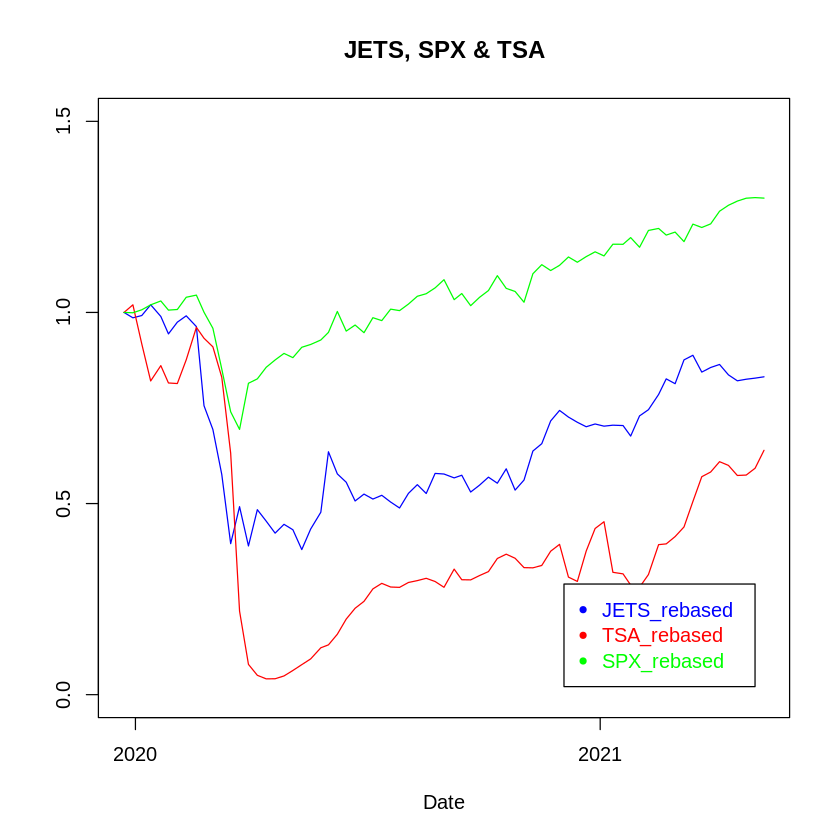


## 4.3 Regressing JETS ETF return against weekly TSA changes

At this point, we have the alternative data (i.e. independent variable) ready. Next, we proceed to obtain data on the dependent variables. To be specific, we will need stocks/securities prices, S&P 500 Index (SPX) prices, alongside risk-free rate proxied by ten year US Government Treasuries Yield. We obtain the data mentioned above using the R package ‘quantmod’. However, ‘quantmod’ package provides data of a daily frequency. Therefore, they are subsequently modified into weekly frequency.

After all data is ready, we will run 3 models for the dependent variable (i.e. JETS ETF). First Model is the CAPM model that regresses JETS ETF excess returns against S&P 500 excess returns by subtracting risk-free interest rates to examine how predictive the traditional CAPM model is. The second model is the ‘Data Model’ that regresses JETS ETF returns against the weekly change in alternative data to examine how predictive the alternative data on its own explain the variations of the dependent variable. The final model is the ‘modified CAPM model’ that regresses JETS ETF excess returns on both S&P 500 excess returns and the weekly change in alternative data to examine whether alternative data improves the predictive power of the CAPM model. We will report the R-squared, t-statistics and p-value for each of these regressions.

The following diagram shows how the JETS ETF, overall market (as proxied by S&P 500 Index) and TSA number changes since 1st Jan 2020. The three series are all ‘rebased,’ allowing easier comparison and visualization of data series of varying degree of magnitude.



Regression results of the three models are shown as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Regression Model** | **Independent Variable** | **Multiple R** | **t-statistics and p-value** |
| CAPM | Weekly Changes in Market Excess Returns | 0.016 | 1.069 and 0.289 |
| TSA | Weekly Changes in TSA | 0.0007 | 0.230 and 0.819 |
| CAPM + TSA | Weekly Changes in Excess Market Returns and Weekly Changes in TSA | 0.01721 | 1.075 and 0.286; 0.284 and 0.777 |

## 4.4 Evaluation/Interpretation of the regressions

Intuitively, such a dataset should give an accurate proxy to the number of air-travel passengers and a barometer gauging the demand of air-travel. Given how far the industry has fallen during the pandemic, hence it makes perfect sense to estimate returns of JETS ETF with the use of TSA number. From a theoretical point of view, trading performance of JETS ETF should show a statistically significant positive correlation with TSA number, and from there, we can develop an explicit mathematical formula to predict JETS ETF returns - a lucrative trading strategy that hedge funds would love to exploit.

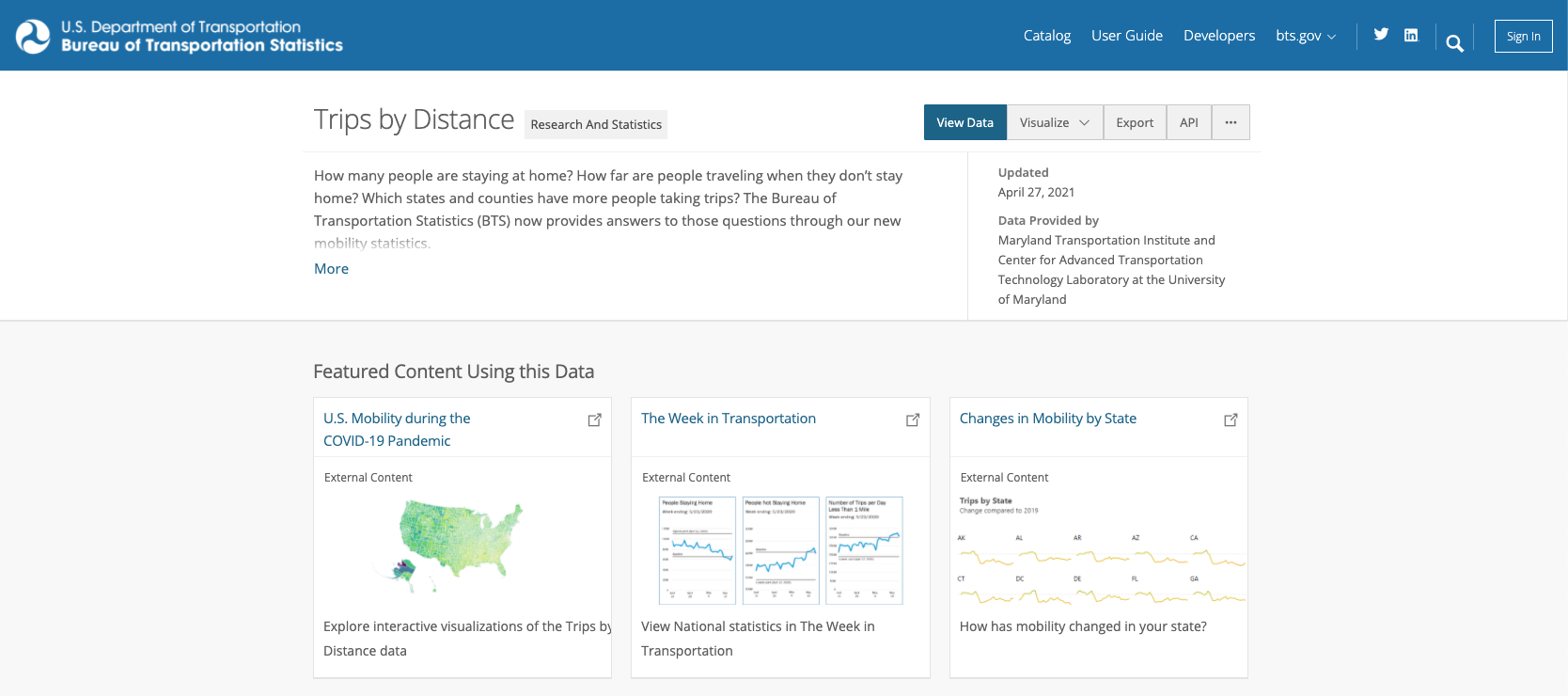
However, we discovered that the number of air-travel passengers, the most intuitive and fundamental driver of airlines’ performance, surprisingly has an insignificant correlation with airlines’ stock price. Even after we control for overall market return (Improved CAPM Model), it still produces rather disappointing results. Below are a list of possible explanations:

1. **‘The market is inefficient’**: It fails to price in the most fundamental drivers of airline stocks in the midst of a pandemic - number of passengers. If one were to believe in this reason behind low R-squared and statistical insignificance, he can profit from developing a programme that exploits the seemingly market mispricing of the real economy.
2. **‘It really doesn’t matter’**: Number of passengers perhaps may not be an important driver to airlines’ stock prices at all. This is manifested by the fact that regression between TSA and JETS ETF prices has a low R-squared, suggesting some other factors might be at play.
3. **‘Blame the outliers’**: It is possible that the sudden and catastrophic plunge in both stock prices and number of passengers in Mar 2020 explains the insignificance. However, we did attempt to remove the ‘outlier data’ and the result is still disappointing.
4. **‘Market takes time to price in the info’**: When some changes are happening in the real economy, the market might need some time to price in this new piece of information, suggesting that we might need to consider running a distributed lag model where we would predict current value of the dependent variable using past/lagged value of the independent variable.
5. **‘Market already priced in the info’**: This is the exact opposite of Reason 4, where the market has already anticipated and priced in the fact that certain changes in the real economy will take place. In this case, one might even consider using the JETS ETF to predict the TSA number.

# 5. Predicting JETS ETF return using Trips by Distance Dataset

## **5.1 Description**

Trips by Distance is a dataset collected and updated by Maryland Transportation Institute and Center for Advanced Transportation Technology Laboratory at the University of Maryland to record the number of people staying at home and not staying at home on a daily basis on a county and state level by sourcing millions of anonymized mobile device data. A metric for a trip is any detected movement that results in a stay longer than 10 minutes from an individual’s home. This dataset accounts for a wide encompassing range of transportation including personal car, public transit, rail and air travel.



## 

## **5.2 Collection**

The original data was discovered on the U.S Department of Transportation Bureau of Transportation Statistics website. The data from 2019, 2020 and 2021 was downloaded and merged together. The state-level data was isolated from the county level information and uploaded to a mySQL database table under “visits”. Subsequently, the state-level data was aggregated to present a nationwide statistic daily. In correspondence to the weekly JETs ETF and SPX data, an initial of 2019-12-22 was set to propagate a 7-day aggregate. The resulting table holds the weekly cumulative frequency of the following fields:

|  |  |
| --- | --- |
| population at home  population not at home  total trips  trips less than a mile  trips between 1~2 miles  trips between 3~5 miles | trips between 5~10 miles  trips between 10~25 miles  trips between 25~50 miles  trips between 50~100 miles  trips between 100~250 miles  trips greater than 500 miles |

Regressing returns

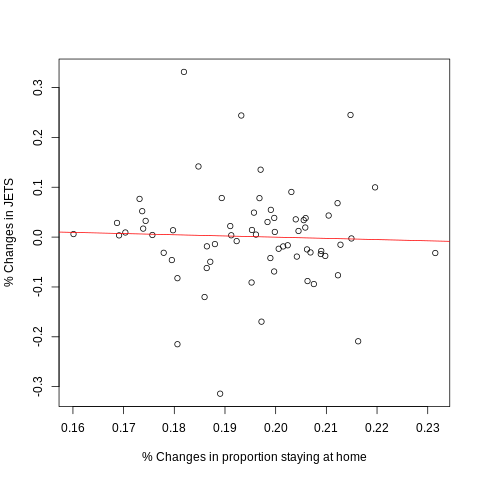
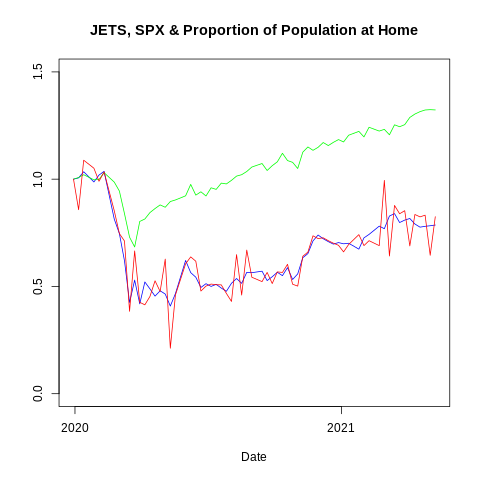
The data queried from mySQL was bundled into six new categories:

|  |  |
| --- | --- |
| Proportion of population at home  Total trips  Short trips  Medium trips  Long trips  “Super” trips | population at home/total reported population  Remains unchanged  Trips between 1~10 miles  Trips between 10~50 miles  Trips between 50~250 miles  Trips greater than 500 miles |

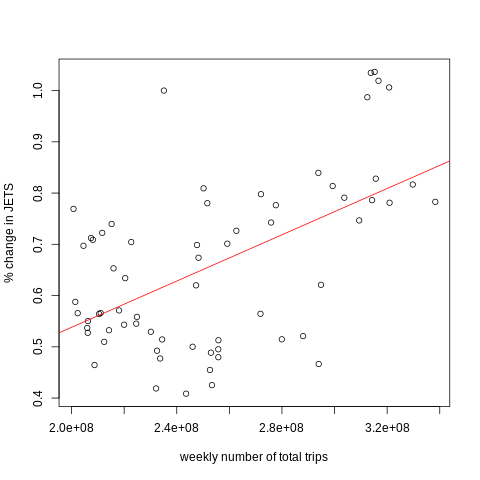
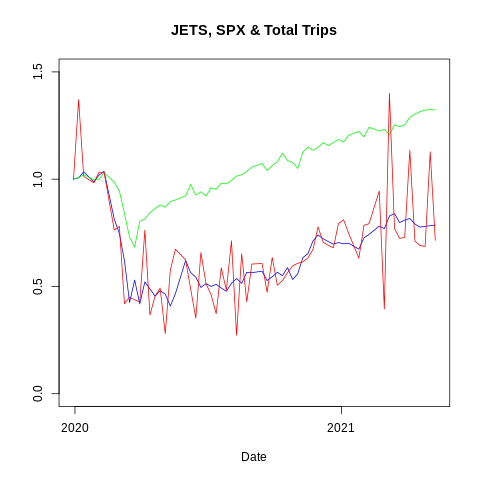
A series of models were tested based on the above data. The first model uses the proportion of population at home and tracks the weekly percent change. The remaining five models track the percentage weekly change in frequency of trips for total, short, medium, long and “super” trips.

## 5.3 Results of regression

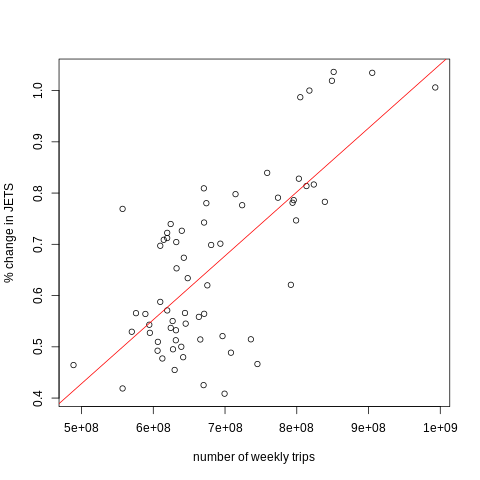
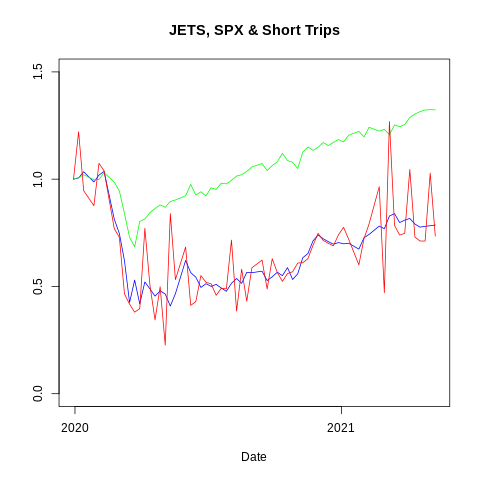
Proportion at Home



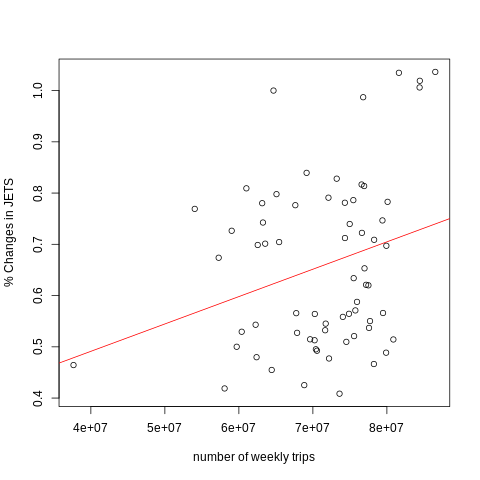
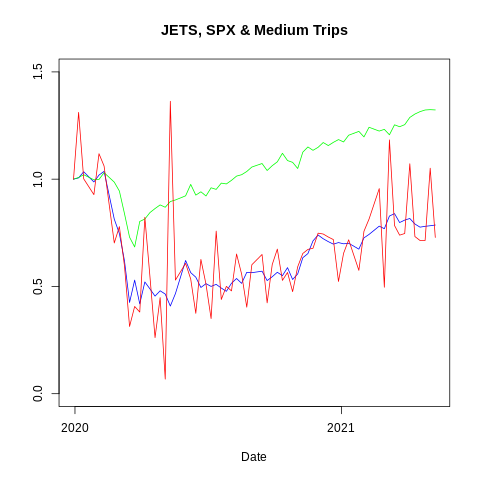
Total Trips



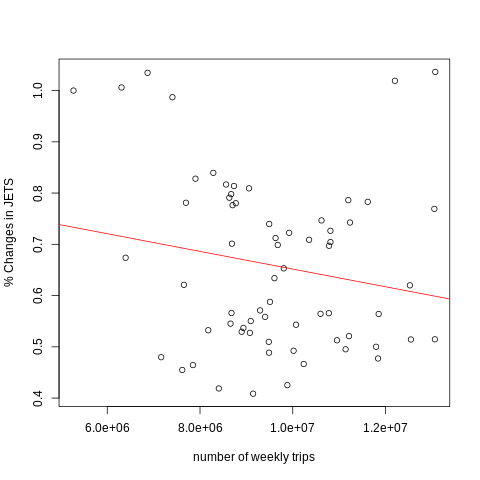
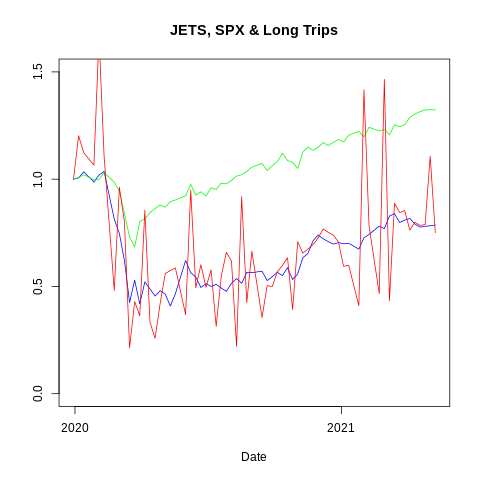
Short Trips



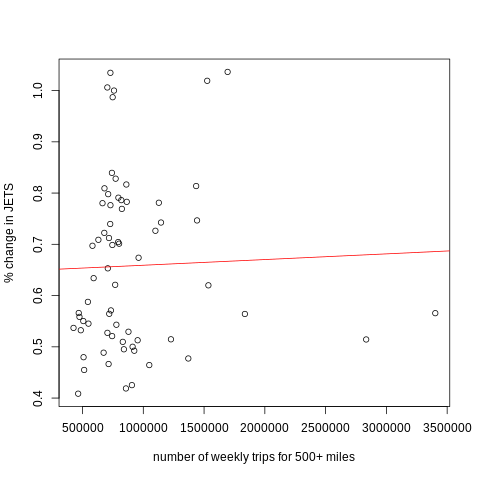
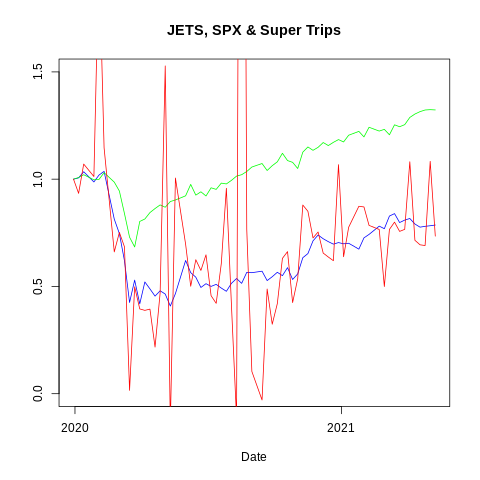
Medium Trips



Long Trips



Super Trips



## 5.4 Interpretation of regression results

|  |  |  |  |
| --- | --- | --- | --- |
| Regression Model | Independent Variable | Multiple R-Squared | T-Statistic and P-Value |
| CAPM | Weekly Changes in Market Returns | 0.01607 | 1.077 and 0.285 |
| Proportion at Home | % change in the weekly proportion of population at home | 0.001276 | -0.286 and 0.776 |
| Total Trips | % weekly change in frequency of total trips | 0.2982 | 5.215 and 0.781 |
| Short Trips | % weekly change in frequency of trips between 1~10 miles | 0.5083 | 8.133 and 1.88e-11 |
| Medium Trips | % weekly change in frequency of trips between 10~50 miles | 0.07422 | 2.265 and 0.0269 |
| Long Trips | % weekly change in frequency of trips between 50~250 miles | 0.03168 | -1.447 and 0.153 |
| Super Trips | % weekly change in frequency of trips greater than 500 miles | 0.001107 | 0.266 and 0.791 |

Interestingly, the result of the regression showed that shorter trips reflect the changes in JETs ETF within our different trip lengths. Evidently, as the distance of the trip increases, the resulting applied regression yields lower R-squared values. In terms of trend, the change in JETs ETF is less sensitive to fluctuations in shorter trips than opposed to longer trips.

In regards to the total number of trips, the R-squared value takes on a moderate value of 0.2982. This value can be explained because it appears to account for a bulk package of all trip lengths. Therefore, the data points not only include higher-performing metrics such as short trips but also take into account low performing metrics such as long and “super” trips. The model runs on the per cent change in the weekly proportion of the population at home yields disappointing results with an R-Squared value of 0.001276.

The results of these regressions allow us to develop a comprehensive and clear picture of how fluctuations in travel and leisure industry securities reflect real-life phenomena. Our models suggest that a large increase in shorter trips generally leads to an improvement in the travel and leisure industry securities. In contrast, longer trips are less reflective of the travel and leisure environment.

Shorter trips can be viewed as the strongest metric relative to the rest because an increase in short trips reflects more activity, which arguably promotes a more positive investor outlook in an industry that is hurt by a sudden decline in activity over the recent period. Rather than a direct link to a particular industry, this source is a proxy to overall changes in economic activity. It can be even suggested that the movement of JETs ETF is driven largely by overall market conditions rather than airline-specific factors.

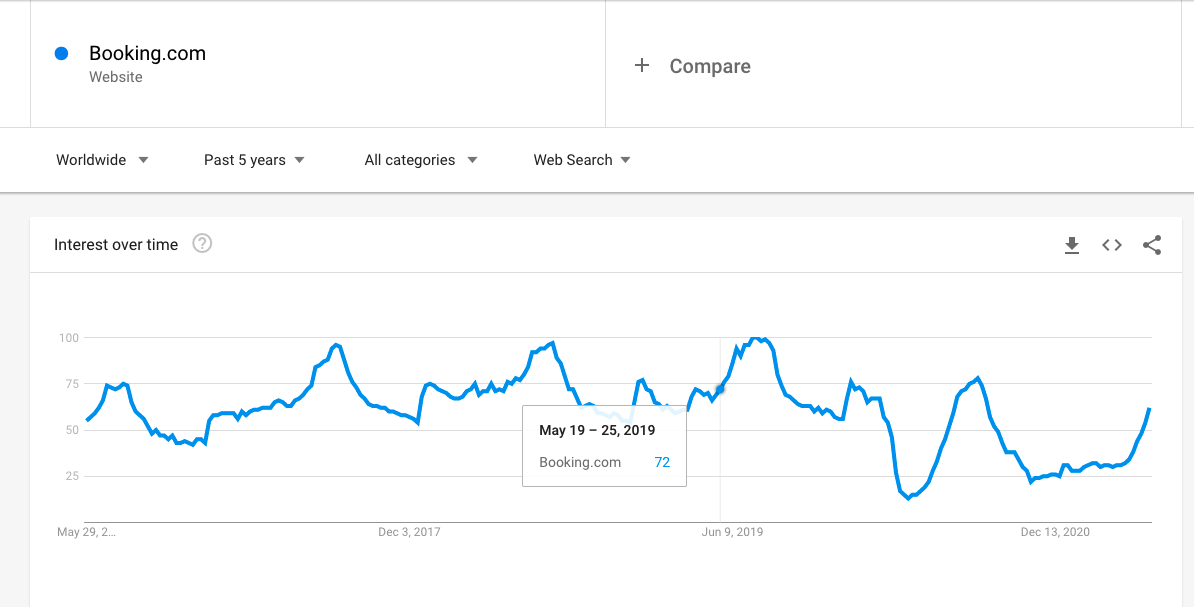
Furthermore, the sudden downturn from a positive to negative trend between the medium trip model and longer trip models can also be speculated upon. Long/super trips are seen as mutually exclusive with flights. This means that reported individuals on a trip of 100 to 500+ miles are likely to be traveling. In other words, the reported data are substitutes in nature. It makes intuitive sense that JETs ETF faces a downwards movement when less people are reportedly traveling over this vast distance.

The results also explain why the earlier model for TSA may have been more ineffective. The TSA number is a highly industry-specific metric that does not necessarily capture all the driving aspects of JETs ETF. On the other hand, the trips by distance dataset cover a more comprehensive overview of developments in the wider economy.

In terms of evaluating the data source, it can be mentioned that the reliability and completeness of the dataset contribute to the model effectiveness. These use a collection procedure that weighs across millions of mobile devices to ensure that the reported data represents the population by addressing the geographic and temporal sample variation issues often observed in a single data source. The quality standard of this dataset covers the day to day movements of a population and therefore, plays a contributory role in the comparatively stronger outcomes in some of the models.

# 6. Predicting JETS ETF and Travel Website Stock’s return using Online Web Traffic - Google Trends and Alexa Rank

## 6.1 Google Trends

**Google Trends** is a [search trend](https://www.wordstream.com/blog/ws/2015/03/02/search-trends-by-state) feature that shows how frequently a given search term is entered into Google’s search engine relative to the site’s total search volume over a given period of time. Google Trends can be used for comparative keyword research and to discover event-triggered spikes in [keyword search volume](https://www.wordstream.com/blog/ws/2017/01/23/keyword-search-volume). 

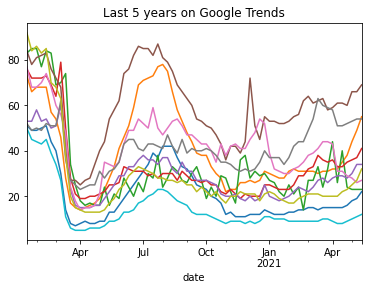
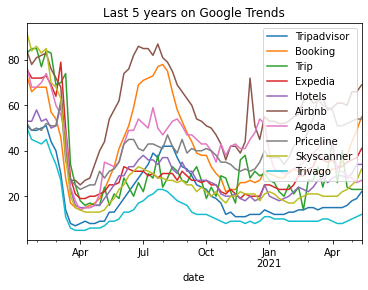
Google Trends Website and Data format

### 6.1.1 Data Scraping /Collection

We decided to obtain the google trends data for the top 10 travel websites(according to another data providerSimilarWeb). These websites account for the highest percentage of flight tickets and hotels booked between 2018-2020. Data was collected using a python script with the help of the pytrends library which provides a simple interface for downloading the data from google trends. This package does the scraping and makes the entire process easier. However, we did encounter a problem when we passed “travel website names'' as parameters to find its traffic.

While searching for distinct keywords, google uses a knowledge graph to keep track of different entities that relate to the same word. Hence when we pass the word “trip”, it can refer to the “trip(verb): that people take” and “trip(proper noun): Travel Company”. In order to distinguish the trends between the two entities, we have a unique identifier called “freebase-ID”. This identifier helps refer to a particular entity in a knowledge graph in API calls. Therefore it was important to scrape the “freebase\_ID” from the wikidata website for each travel website. The “freebase-ID” is then used as a parameter and used to retrieve the dataset.

Once the data for the 10 websites were collected, the row-wise sum and mean for weekly data was computed and appended. This gives an approximate of the total web traffic generated from google on these top 10 travel websites. The data was then written to the database. In order to obtain the latest data, one would just be required to run the code again and the table in the database gets updated alongside as well.

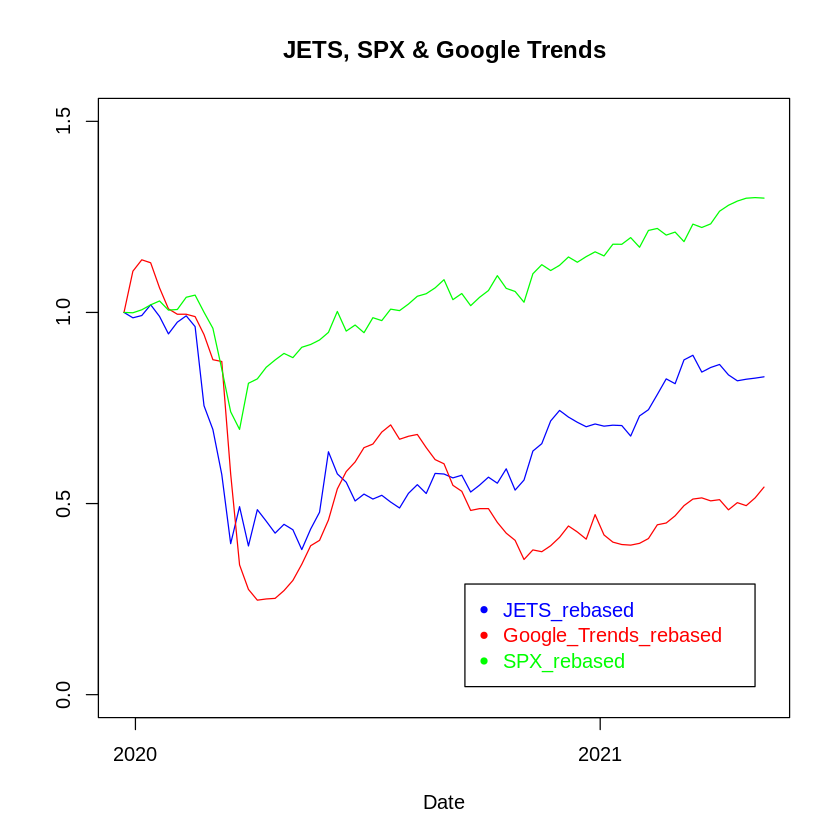
Google trends of the top 10 websites

We can witness the huge drop in google trends from March to April and its comeback by july.

### 6.1.2 Results of the regressions on JETS and BKNG(Quantitative)

As mentioned previously, we have 3 different regression models that were performed on 2 different securities - JETS and Travel Website (Trip, Bookings, Expedia, Airbnb).

**JETS ETF:**

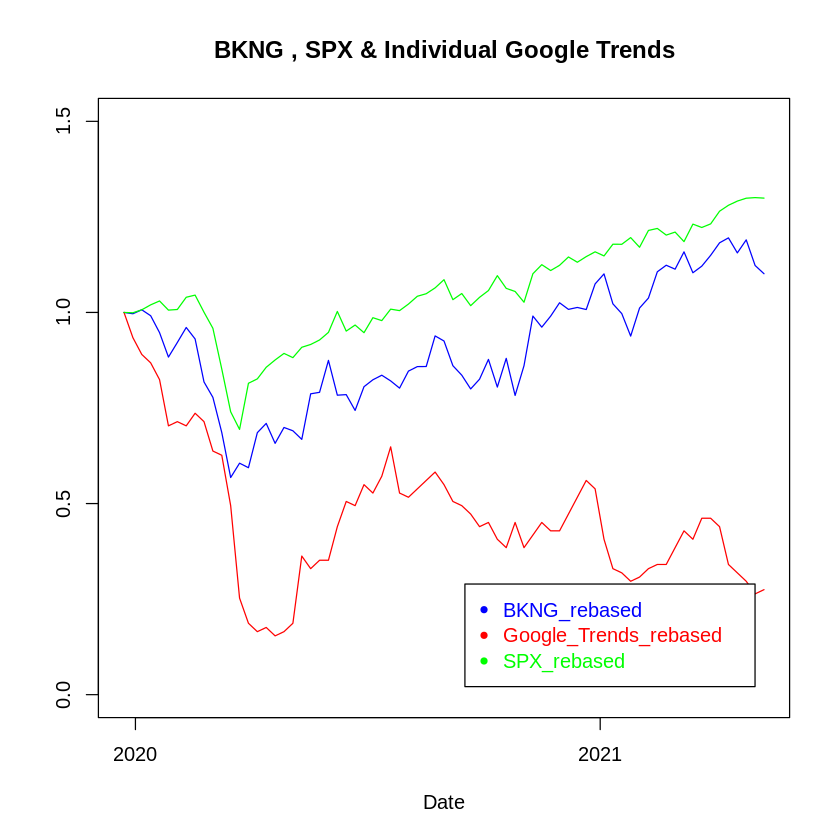
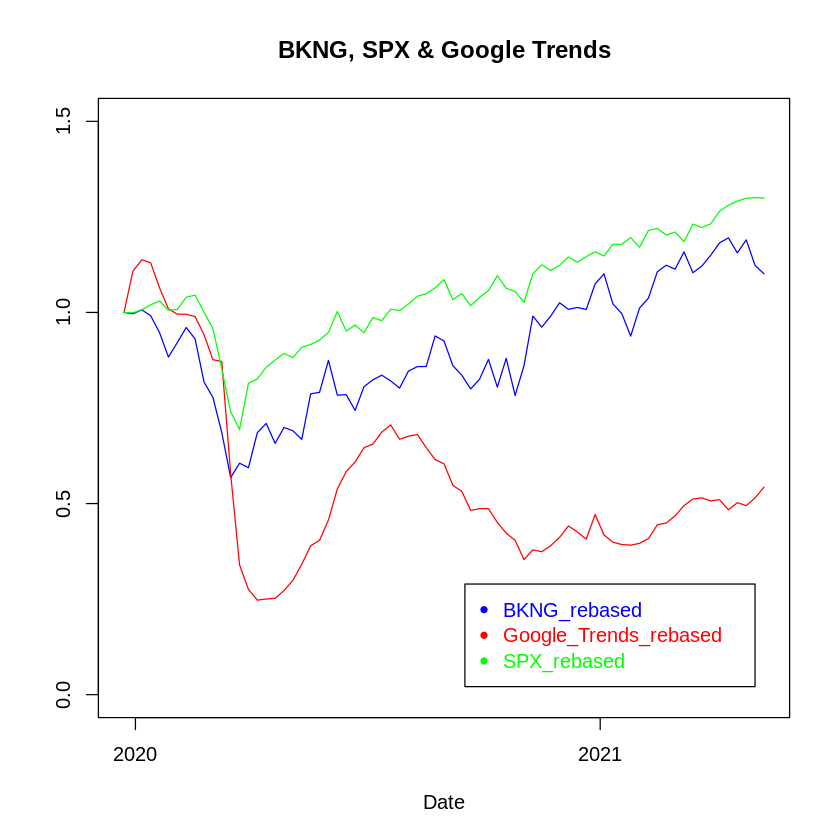


Dependent Variable is the same for all the 3 models: Weekly Change In JETS Returns

|  |  |  |  |
| --- | --- | --- | --- |
| Regression Model | Independent Variable | Multiple R-Squared | T-Statistic and P-Value |
| CAPM | Weekly Changes in Market Returns | 0.05124 | 1.958 and 0.0541 |
| Google Trends | Weekly Changes in Google Trends | 0.04165 | 1.757 and 0.0833 |
| CAPM + Google Trends | Weekly Changes in Market Returns and Weekly Changes in Google Trends | 0.07126 | 1.494 and 0.140, 1.229 and 0.223 |

**Top 10 Travel Website Individual Stocks:**

The code provides flexibility in choosing any of the top 10 stocks and calculates the 3 regressions. We will just show how Bookings.com(BKNG) returns reflect the cumulative google trends data for all websites and also its individual google trends data.



Dependent Variable: Weekly Change In BKNG Stock Returns

|  |  |  |  |
| --- | --- | --- | --- |
| Regression Model | Independent Variable | Multiple R-Squared | T-Statistic and P-Value |
| CAPM | Weekly Changes in Market Returns | 0.01607 | 1.077 and 0.285 |
| Cumulative Google Trends1 | Weekly Changes in Google Trends | 0.02685 | 1.400 and 0.166 |
| CAPM + Cumulative Google Trends | Weekly Changes in Market Returns and Weekly Changes in Google Trends | 0.03331 | 0.684 and 0.496, 1.118 and 0.268 |
| Individual Google Trends2 | Weekly Changes in Google Trends | 0.0004668 | -0.182 and 0.856 |
| CAPM + Individual Google Trends | Weekly Changes in Market Returns and Weekly Changes in Google Trends | 0.05412 | 1.993 and 0.0502, -0.462 and 0.6458 |

1 - Cumulative Google Trends: Google Trends for all the top 10 websites

2 - Individual Google Trends: Google Trends for a particular website

### 6.1.3 Interpretation of the regressions(Qualitative)

With the above data summarisation, we can clearly see that Google Trends is not a significant variable in determining the changes in returns in JETS or BKNG. For over a decade now, many investors have been trading based on word hits on social media platforms like twitter. Google trends provide similar data for searches generated from google, which could be used to predict stock movements and return predictions. Hence the popularity of an investment may be related to the investor's attention, which google searches and trends can act as a proxy.

One might intuitively expect an increase in a trend, which might lead to an increase in revenue of a sector or a business thereby raising the returns. For example, In this digital age, anyone who wants to book a flight has a higher chance of booking online by searching on google. This leads to an increase in google search volumes, which might translate to improved revenue.

However, after the regressions that were performed, it was clearly evident that our hypothesis wasn’t the case and that Google Trends data does not carry a significant role in determining the returns in JETS ETF nor in the individual stocks return. This implies that the dataset has already been priced in or does not have a relationship whatsoever (like mentioned in the previous section, “**The market is inefficient**”, “**It really doesn’t matter**” and **“Blame the outlier”**  )

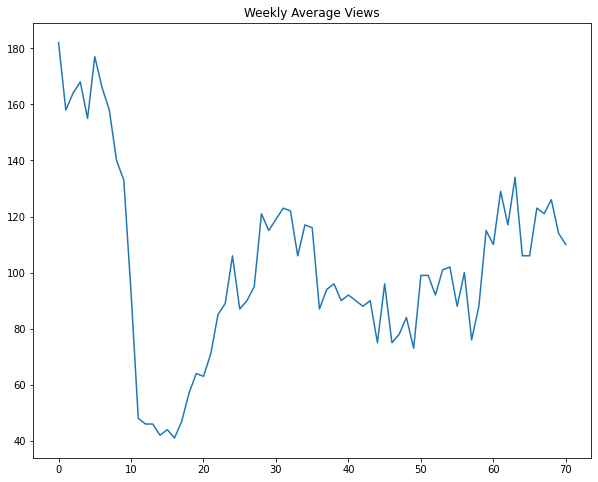
## 

## 6.2 Alexa Rank - Amazon Web Traffic

Alexa rank is a global ranking system that ranks millions of websites in order of popularity. It’s calculated by looking at the estimated average daily unique visitors and number of page views for a given site over the past 3 months. This amazon web service provides the Rank, Pageviews per Million and Pageviews per user for any valid website. As it is a paid service, we were very careful on the number of API calls and the amount of data that was collected.

### 6.2.1 Data Scraping /Collection

We decided to download the amazon web traffic for the top 10 travel websites. Data was collected using a python script with API call. AWS had provided the documentation for the API call and the list of parameters that had to be sent to receive a valid json response. Each API call could fetch upto 31 days of data from the start date. It provides a simple interface for downloading the data from their database. The json response was highly nested and we had to extract only the relevant features and convert them into a suitable dataframe. We also had to ensure that the website passed as a parameter was a valid website. The data frame was downloaded to csv files and loaded back again into one single csv for all dates and websites. This was then written to the database. The row-wise sum and mean, were calculated for all the websites on a daily basis and written to the database. Following this we also calculated the amazon web traffic in weekly intervals along with its mean and sum for all the columns. This gives an approximate of the total web traffic for the top 10 travel websites. Upon performing these transformations, the data was then written again to the database. To obtain the latest data, one would just be required to run the code again and it rewrites the table on the database as well.



Weekly Pageviews Per Million of top 10 travel websites

from Jan 2020(week 1) to May 2021(week 70)

### 6.2.2 Regressing Stock Returns from JETs ETF and Individual Travel Website Stocks

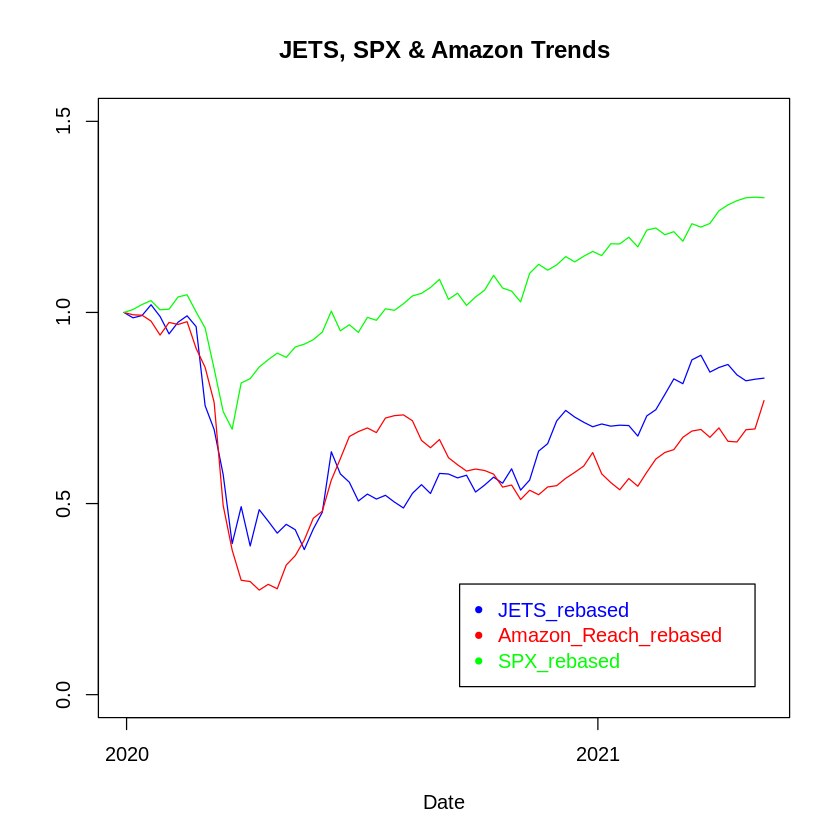
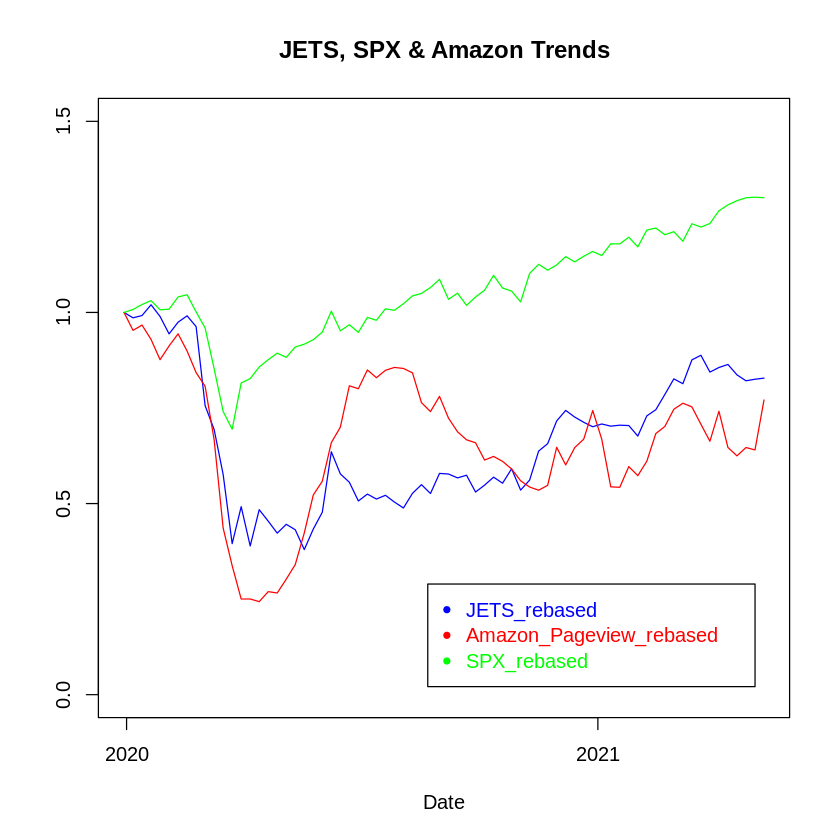
In order to have all the data for regression, we needed to include the stock/security price, market price of S&P 500 (SPX) and risk-free rate (10 year US Government Treasury Yield) just like the previous time. Upon obtaining the required information using various packages, it was then merged with the amazon web traffic table. After this, we changed the merged data to a weekly basis. Following this, we calculated the weekly returns for the JETS/Individual Stock, along with the weekly change in google trends and SPX weekly return. We also calculated the excess return for the weekly returns by reducing the yield rate(per week). We also “rebased” the weekly returns/changes to ensure the arbitrary starting point is 1. The future weekly changes are then multiplied to 1 that was rebased.

Once this was done, we had constructed 3 regression models. First Model was the CAPM model, which regressed security returns on market returns. The second model was the Data Model, which regressed security returns on the Amazon Web Data. The Final Model was an improvised model that had regressed security returns on both the market returns and the google trends data. All the 3 models show us the R-square value and significance of the different independent variables for calculating the dependent variables. This helps to evaluate the underlying dataset

### 6.2.3 Results of the regressions

As mentioned previously, we have 3 different regression models that were performed on 2 different securities - JETS and Travel Website (Trip, Bookings, Expedia, Airbnb).

**JETS ETF:**



JETS ETF return vs Change in Amazon Pageview Rebased vs Market Return

JETS ETF return vs Change in Amazon Reach Rebased vs Market Return

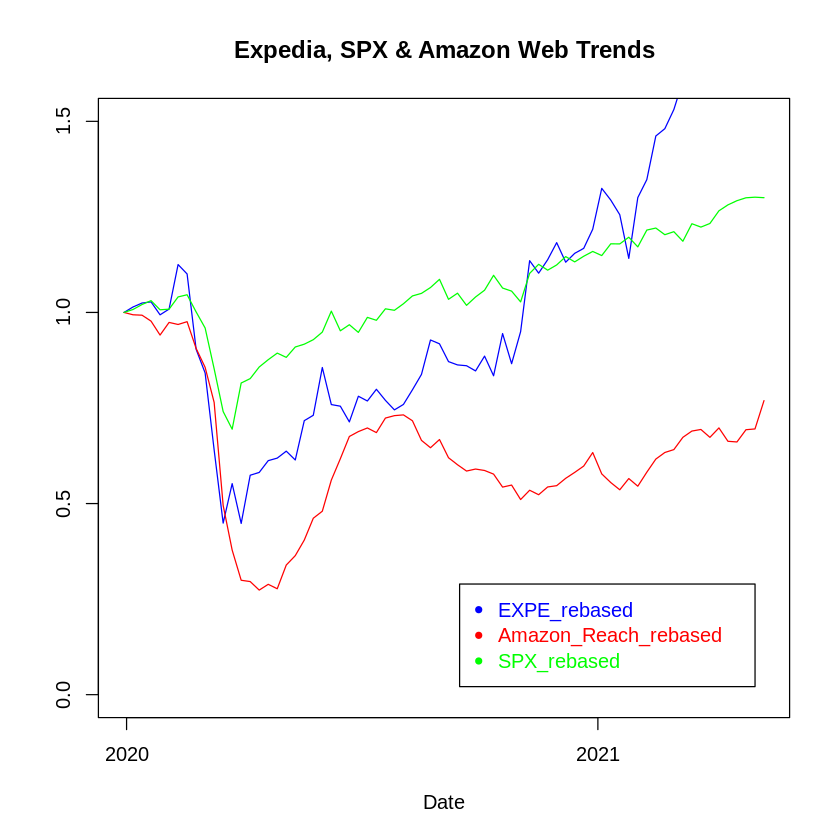
Dependent Variable: Weekly Change In JETS Returns

|  |  |  |  |
| --- | --- | --- | --- |
| Regression Model | Independent Variable | Multiple R-Squared | T-Statistic and P-Value |
| CAPM | Weekly Changes in Market Returns | 0.372 | 6.439 and 1.29e-08 |
| Amazon Web Traffic | Weekly Changes in Amazon Web Traffic | 0.05822 | 2.080 and 0.0412 |
| CAPM + Amazon Web Traffic | Weekly Changes in Market Returns and Weekly Changes in Amazon Web Traffic | 0.3806 | 5.993 and 8.4e-08, 0.980 and 0.33 |

**Top 10 Travel Website Individual Stocks:**

The code provides flexibility in choosing any of the top 10 websites and calculates the 3 regressions. For the sake of the report, we will just show how Bookings.com(BKNG) returns reflect the cumulative amazon web traffic data for all websites and also its individual amazon web traffic data.

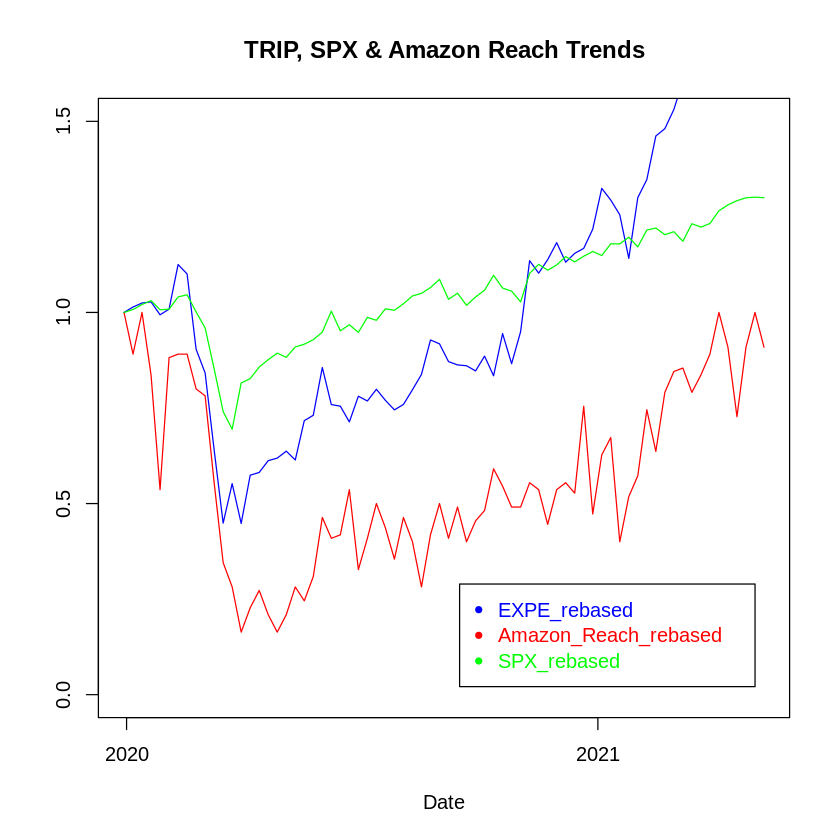
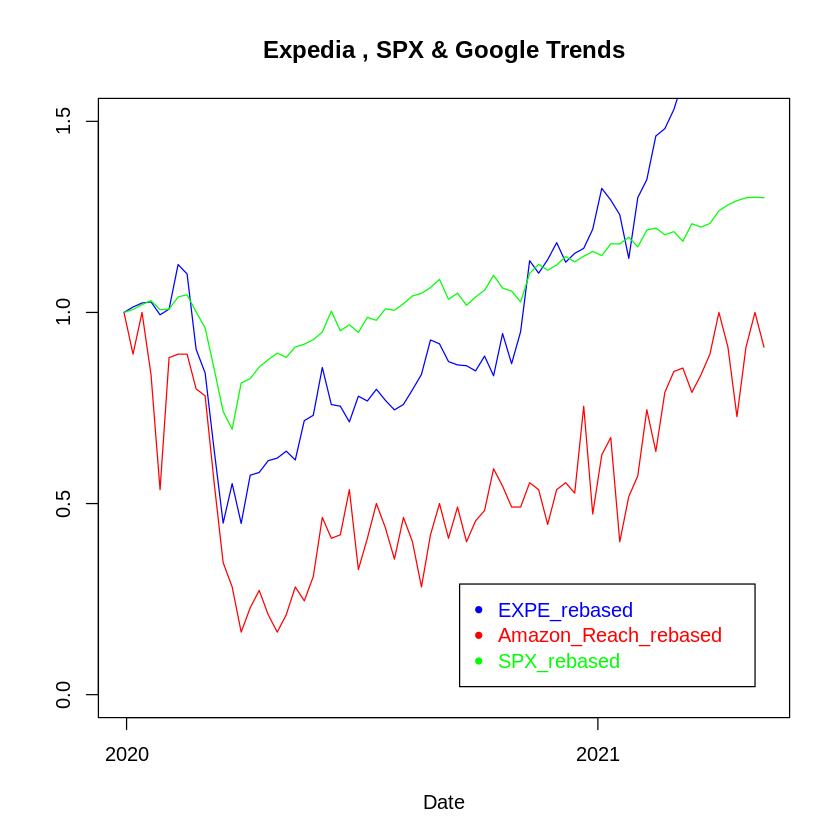
Cumulative Amazon Web Trends data



EXPE return vs Change in Amazon Pageview Rebased vs Market Return

EXPE return vs Change in Amazon Reach Rebased vs Market Return

Individual Amazon Web Trends data



EXPE return vs Change in Amazon Pageview Rebased vs Market Return

EXPE return vs Change in Amazon Reach Rebased vs Market Return

Dependent Variable: Weekly Change In Individual Stock Returns

|  |  |  |  |
| --- | --- | --- | --- |
| Regression Model | Independent Variable | Multiple R-Squared | T-Statistic and P-Value |
| CAPM | Weekly Changes in Market Returns | 0.02083 | 1.220 and 0.226 |
| Cumulative Amazon Web Traffic | Weekly Changes in Amazon Web Traffic | 0.1191 | 3.076 and 0.00299 |
| CAPM + Cumulative Amazon Web Traffic | Weekly Changes in Market Returns and Weekly Changes in Amazon Web Traffic | 0.1376 | 1.216 and 0.2281, 2.633 and 0.0104 |
| Individual Amazon Web Traffic | Weekly Changes in Amazon Web Traffic | 0.06687 | 2.240 and 0.0283 |
| CAPM + Amazon Web Traffic | Weekly Changes in Market Returns and Weekly Changes in Amazon Web Traffic | 0.06864 | 0.362 and 0.718, 1.882 and 0.064 |

### 6.2.4 Evaluation/Interpretation of the regressions

In the first table, the CAPM had a decent R squared score of .3 and market returns proved to be significant. We can also clearly see that change in Amazon Web Traffic is a significant variable in determining the changes in JETS ETF. Whereas in the second table, we can see market returns are not a significant factor in determining Expedia stock returns. However, in the other subsequent regressions, we can see that change is cumulative and individual amazon web traffic are significant factors in determining the Expedia Stock returns.

We have seen that Amazon web trends provide a dataset that could be used to understand stock movements and return predictions. After the regressions, we witness the significance of Amazon web trends on individual stocks' return. This implies that the dataset might not have been priced in.

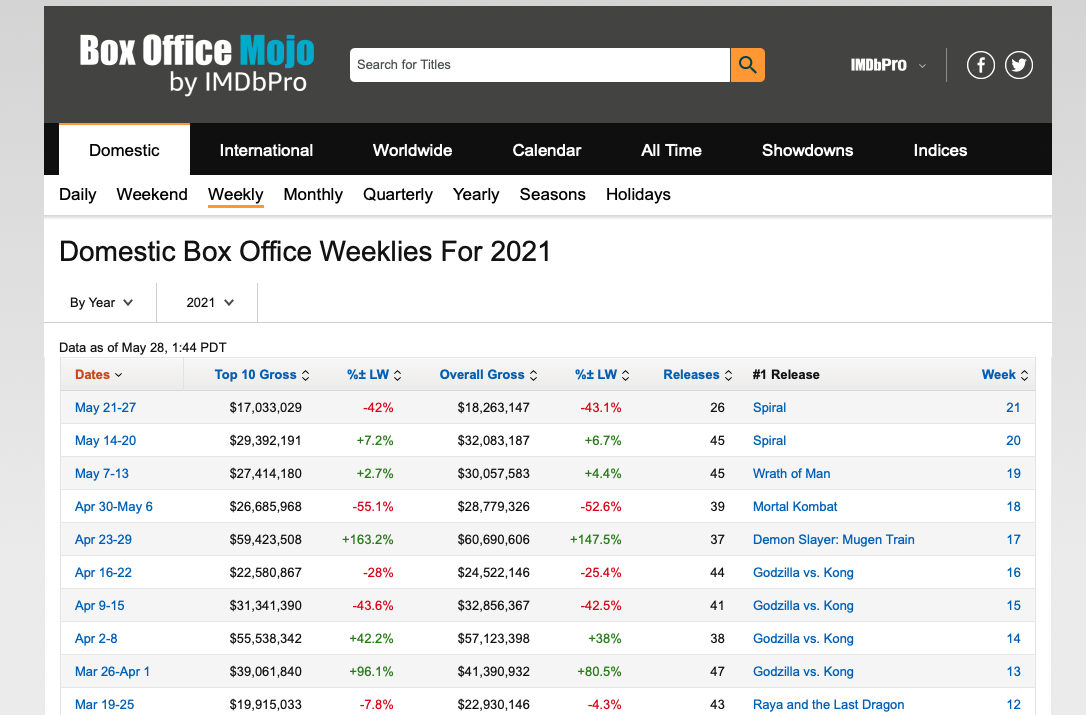
## 6.3 Why does amazon web traffic have a significant role in stock return compared to google trends?

We believe that amazon web data being a paid service has not been priced into the market yet. Whereas on the other hand, google trends is openly available to the public for free. This factor has led to investors using this dataset extensively and hence reducing the arbitrage.

Another reason why we expect that google trends might not necessarily be significant in stock returns is that google only shows the trends from their servers and does not show how many people are actually visiting their websites. Amazon on the other hand provides us with information regarding the number of people and number of pages a user is visiting for the respective websites. This in one way helps to understand and predict their revenue as many firms have a conversion ratio. The more people who visit the website, the higher the sales. Higher sales might lead to higher profits which translates into improved returns. This might be the reason why amazon web traffic is able to be a better predictor for returns on JETS and Individual returns compared to google trends.

# 7. Predicting Stock returns of Companies with Theatres as Primary Business using Box Office Revenue

Box Office Revenue is a data source that is operated by “Box Office Mojo” which is owned by IMDB, a subsidiary of Amazon. The website tracks domestic (i.e., US) box office at a daily frequency, and box office of countries worldwide at a weekly frequency. The dataset will be tested against top 3 stocks with theatres and cinemas as the primary business - AMC, IMAX and CNK.

*Box Office Mojo 2021 Weekly Domestic Box Office* 

## 7.1 Data Scraping

We will use the US box office data at the weekly frequency. Box Office Mojo has a made-ready page that aggregates the daily number into weekly numbers, which provides the following information:

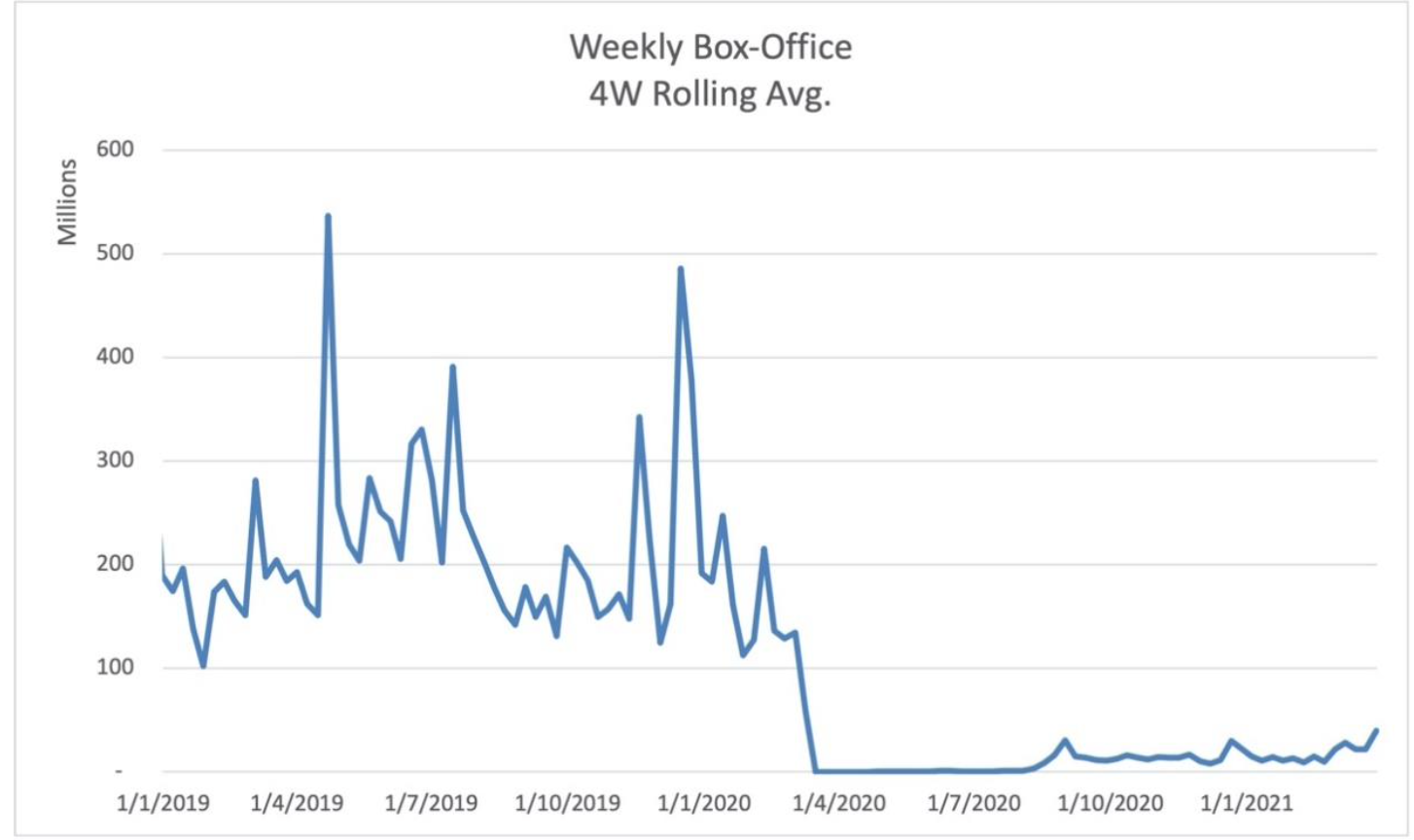
1. Week number; start & end dates of the week
2. The gross box office of top 10 (in terms of box office) movies aired during the week
3. The gross box office of all movies aired during the week
4. Number of releases aired during the week
5. #1 release of the week

We will only be using **items 1 and 3** in order to see if the underlying data can predict the stocks that have theatres as their primary business.

The data scraper was built in R using the “Rvest” Package. We first define a data.frame object “boxoffice”, such that once the data is loaded in R and manipulated into the desired format, it will be stored in the data.frame which makes analysis and writing into a database easy. We Specify the URL that the “rvest” package should look for, in this case, the Box Office Mojo weekend box office page.Our objective is to retrieve multiple years of data and store in a format that can be used later for analysis, Box Office Mojo stored different years of box office figures in different paths under the same host.

Therefore, we use a for-loop to control the program such that we can arrive at the corresponding weekly box office page of various years. To specify the portion of web content that we would like to look up, we will make use of the functions “html\_node()” and “html\_text()”. The function “html\_node()” takes the URL and a css selector as input. Passing the output of the “html\_node()” function to the function “html\_text()” yields the desired information as a string. One could obtain the css selector of the desired web content using the free Google Chrome extension “SelectorGadget”.

The second inner for-loop is designed to retrieve the weekly revenue data. The structure is similar to that of Step 4. The result is stored in the variable “temp”. Then, we remove non-number characters in the string before converting the data from a string format to a numeric format, where the result is stored inside the variable “week\_[week\_number]\_rev”. The third inner for-loop is designed to retrieve the weekly number of releases data.

Once this is done, we write the data.table into the database. 

*Weekly Box Office 4 Week Rolling Average 2019 - 2021*

## 7.2 Regressing Stock Returns of Companies with Theatres as the primary business

In order to have all the relevant data for regression, we needed to include the stock/security price, market price of S&P500 (SPX) and risk-free rate(10 year US Government Treasury Yield) just like the previous time. The rest of the regression follows the same approach as performed in other datasets. Following which the three different regression models were constructed again.

## 

## 7.3 Results of the regressions

**Top 3 stocks with Theatre and Cinemas as the primary business**

The code provides flexibility in choosing between AMC, IMAX and CNK and calculates the 3 regression. For the sake of the report, we will just show how AMC Entertainment Holdings(AMC) returns reflect the box office revenue on a weekly basis.

Graph shows the relationship between AMC Returns Rebased, 

Market Returns Rebased and the Box Office Revenue Rebased

Dependant Variable: Weekly change in stock returns of AMC

|  |  |  |  |
| --- | --- | --- | --- |
| Regression Model | Independent Variable | Multiple R-Squared | T-Statistic and P-Value |
| CAPM | Weekly Changes in Market Returns | 0.003358 | 0.763 and 0.446 |
| Box Office Revenue | Weekly Changes in Box Office Revenue | 0.01984 | 1.871 and 0.063 |
| CAPM + Box Office Revenue | Weekly Changes in Market Returns and Weekly Changes in Box Office Revenue | 0.02152 | 0.545 and 0.5868, 1.787 and 0.0757 |

7.4 Evaluation/Interpretation of the regressions(Qualitative)

From the above table, we can see that CAPM model and the Box Office Model are not of any use. The above regression shows very low R square scores and are not significant at any values. This was disappointing given our main motivation behind choosing the Box-Office revenue was due to its ability to act as a proxy for the revenue of the business and hence act as a predictor for stock returns. However as seen in the above table, that wasn’t the case.

We also have a slight belief that the stock price might lead its revenue, especially with the COVID restrictions. That is the stock price is acting as a predictor for revenue(which is vice versa of our initial assumption). For example, when NYC decided to open all the theatres to the public, most of the individual stocks were up by 8-10% on the day of the news. But it needs to be understood that the revenue wouldn’t increase until the stipulated day/week in which movie theatres are allowed to be opened. This implies that the stock price is leading the revenue and their stock price acts as a predictor for box office revenue.

# 8. Conclusion

We realize that as a whole, the use of indicators tracking the real economy produces rather disappointing results, showing low statistical significance and low R-squared. Across all regressions, the use of alternative data is not as effective as the use of market return to predict JETS ETF returns. This might indirectly prove a hypothesis that the main street and wall street are detached and disconnected.



The Economists, 9 May 2020 Edition

Cover Page

We also observe that the more specific the data depicts about the leisure and travel industry, the less effective it is to predict the JETS ETF’s return. While web traffic of OTAs, box office and TSA all are data that specifically describe how the travel and leisure industry might mount its comeback, they are overwhelmingly disappointing and are of little use in predicting returns of JETS ETF. On the contrary, data that reflect how the broader real economy and less specific about the industry is the best alternative dataset to predict JETS ETF.

Paid data more powerful? -> Put that as a point under Google v.s. AWS

This conclusion is a more general conclusion across all datasets

# 9. Future Work

Our main focus in this project was to assess the leisure and travel industry in the backdrop of COVID-19 in an attempt to model its recovery. It is important to note that the current sources of data and models are primarily focusing on the COVID-19 recovery window. Therefore, it's relevance is limited. In the future, as other factors take over, the intuitive use of data such as trips by distance and booking information is likely to diminish. Our future work consists of introducing new models that are sustainable over a longer term. As we build a more comprehensive range of data, both industry-specific and macroeconomic, it will help us develop a more comprehensive end-to-end picture of the factors at play.

Similarly, we wish to explore other travel and leisure securities or even other assets classes. This project is limited to certain stocks because we believe it is a good representation of the travel and leisure industry. However, in the long run, we wish to compile a wider array of securities. We also intend to group stocks into portfolios and see how their returns vary. These portfolios should address all the core aspects of the industry.

Evidently, the travel and leisure market is always changing and we can expect it to resume to more traditional movements after recovering from this flash period. To maintain a strong assessment, it is crucial to continually source a wide range of securities, leverage new data and fine-tune models accordingly.