IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bianca H. Armbrister
28th September, 2021

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- In order to create our prediction model, we used the following methodologies on the collected data:
  - Publicly available data about SpaceX Falcon 9 rocket launches was collected from the SpaceX API and Wikipedia.
  - The data was cleaned of any data unrelated to the rockets themselves. Missing numerical data was replaced with a mean value.
  - The cleaned data was analyed to create a model that can be used to predict the success of stage one launches based on past performance.
- Based on the analysis of the data SpaceX would see more successful launches using the Falcon 9 rockets equipped with FT boosters with a payload of no less than 2,000 kg and no more than 4,000 kg for maximum launch success.

  Using a decision tree based model, we can predict the successful outcome of a launch based on the payload,

Introduction

SpaceX is a leader in the commercial space travel industry with a formidable reputation in sending manned missions into space.

In order for SpaceY to compete with this level of success, the company should utilize the best configuration of launch conditions, as well as minimize the cost of launches by the reuse of the boosters.

Through analysis of the SpaceX launch data, we are able to predict whether or not a first stage booster is will be reused.

Section 1

Methodology

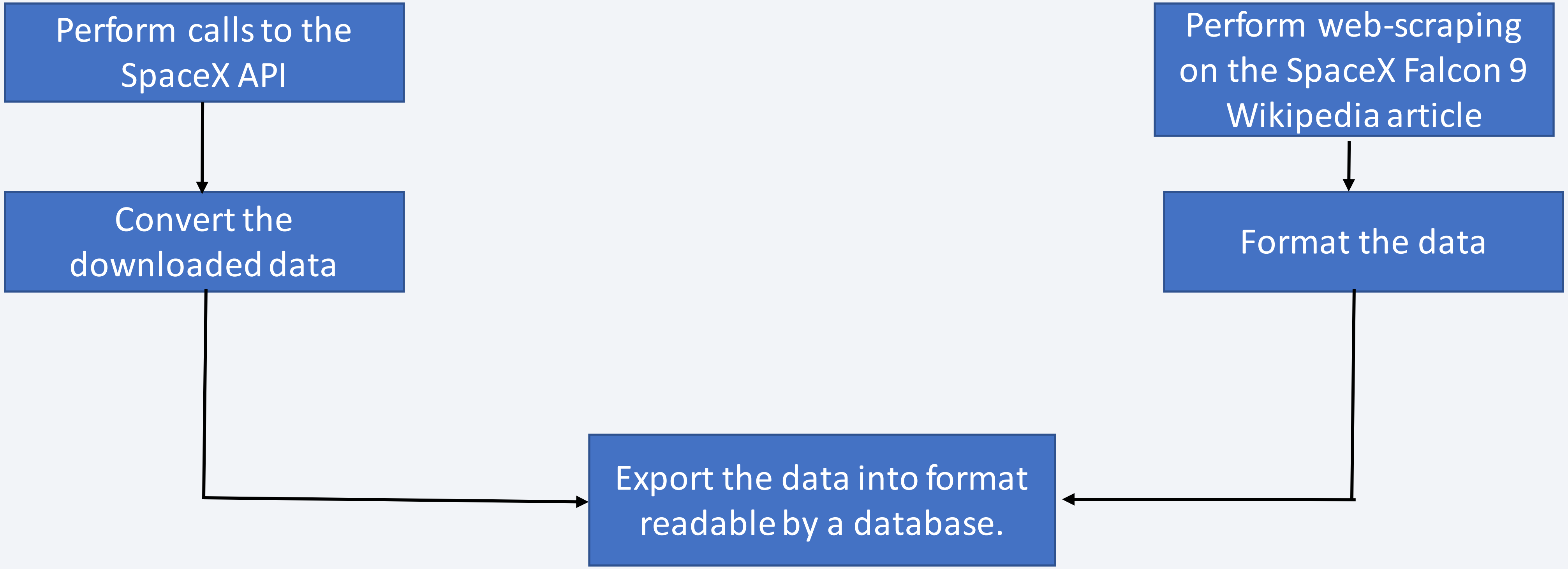
Methodology

Executive Summary

- Data collection methodology:
  - Data of past rocket launch records was obtained using datasets obtained from the SpaceX API, and historical launch data web-scraped from Wikipedia.
- Perform data wrangling
  - The API dataset and web-scraped data were processed into dataframes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using the above analysis methods, a model was built and trained to perform predictions on launch outcomes.

Data Collection

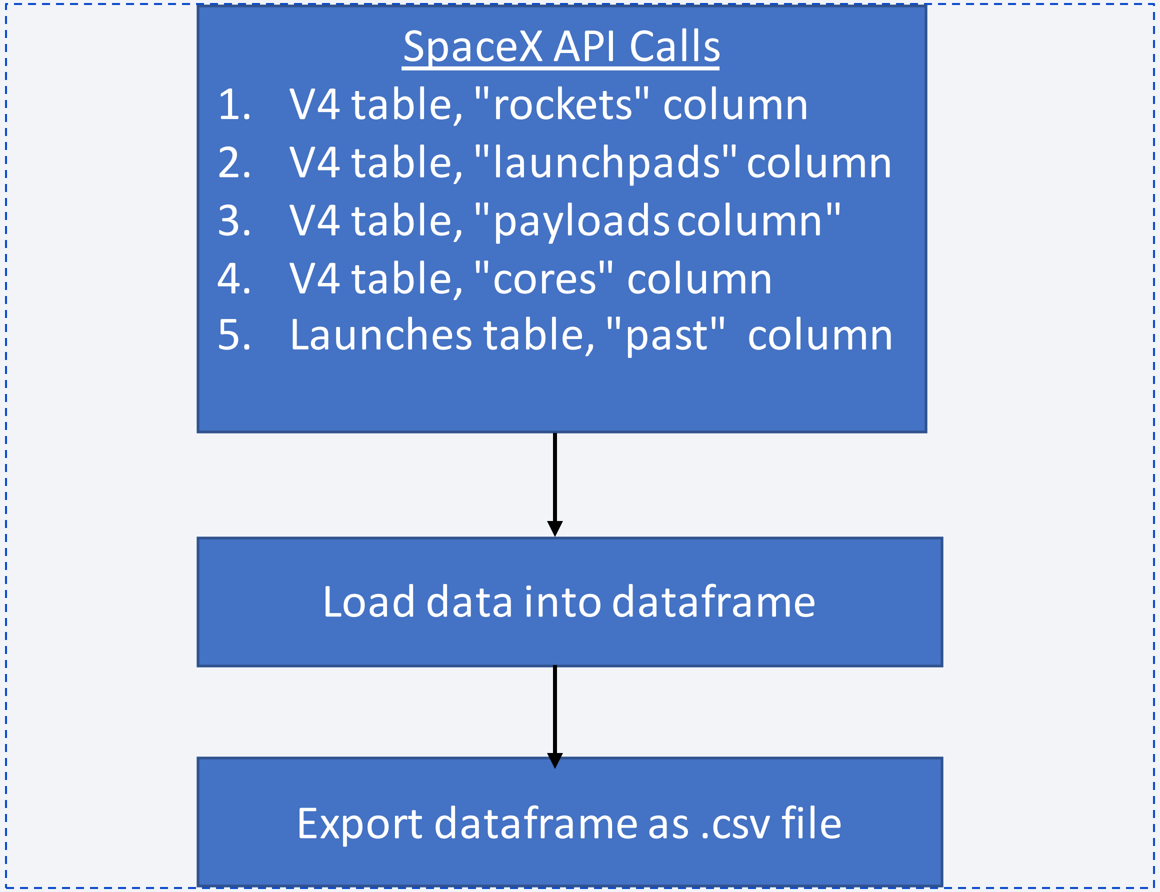
- Data sets were collected by parsing the data from the SpaceX API and web-scraping the Wikipedia article on Falcon 9.



Data Collection – SpaceX API

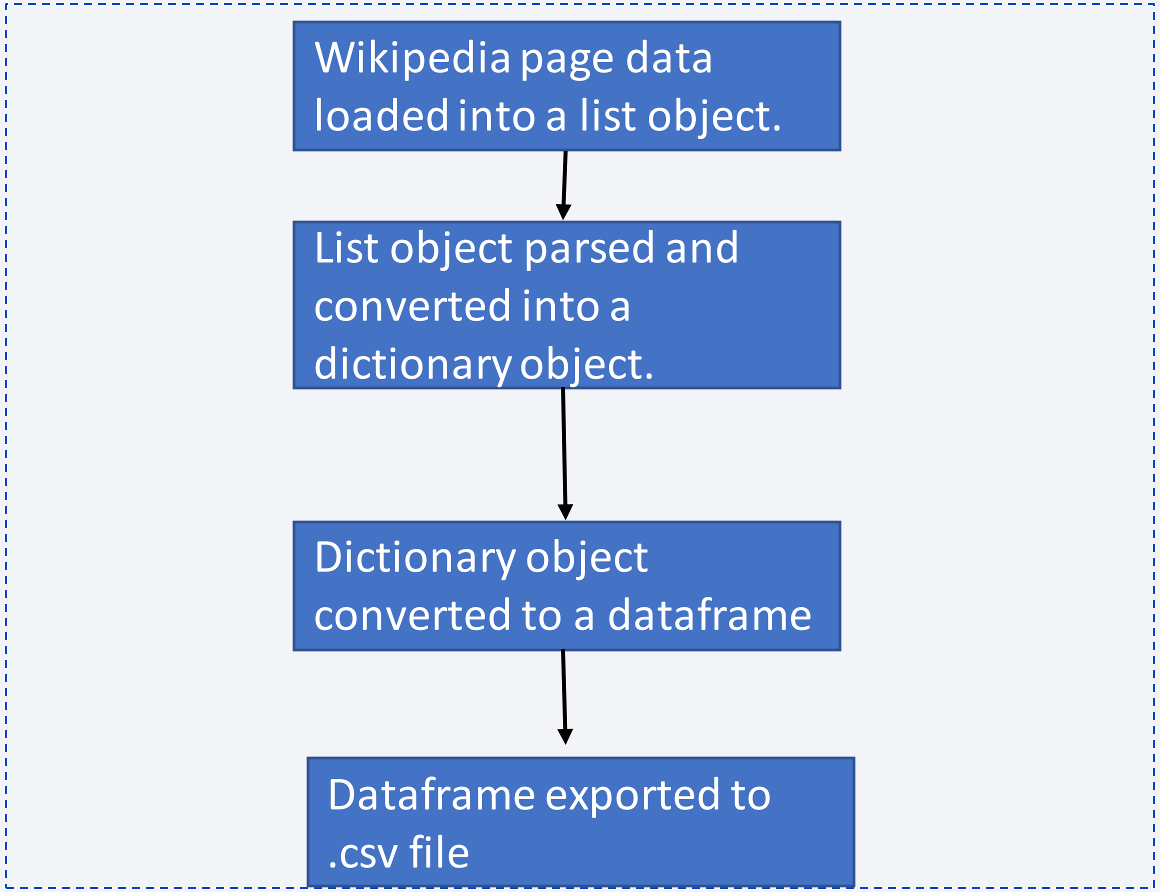
- The columns extracted from the SpaceX REST API includes:
    1. name of the booster used
    2. coordinates and name of the launchpad site
    3. payload mass and orbit of the rocket
    4. outcome of each launch and the details of the core used
    5. past launch data

- The SpaceX API calls notebook for review purposes is located at: https://github.com/GigaCat-86/MLCapstone/blob/2a66a1b453f9bc551412b3d7384c1d975f4a734b/Data%20Collection%20API.ipynb



Data Collection - Scraping

- Wikipedia's Falcon 9 and Falcon Heavy Launches Records page was scraped for data which was then converted into a .csv file.
- The web scraping notebook can be found at : https://github.com/GigaCat-86/MLCapstone/blob/2a66a1b453f9bc551412b3d7384c1d975f4a734b/Data%20Collection%20and%20Web%20Scraping.ipynb



Data Wrangling

Read .csv file into a dataframe

Determine the percentage of missing values for each column along with their data types.

Calculate the number of launches per site.

Calculate the number of occurrences of each orbit.

Calculate the number and occurrence of each mission type per orbit

Classify each outcome as either 0 (unsuccessful) or 1 (successful).

Calculate the average success rate of launches using Falcon 9 type boosters.

The Notebook pertaining to this process is accessible from the following link: https://github.com/GigaCat-86/MLCapstone/blob/2a66a1b453f9bc551412b3d7384c1d975f4a734b/EDA%20Lab.ipynb

EDA with Data Visualization

Type of Chart Plotted	Data Used	Reason
Scatter Point	Payload Mass, Flight Number, Class	To get an idea of the success and failure rates of different paylods for each launch.
Scatter Point	Launch Site, Flight Number, Class	To see the relationship between the launch sites and the different launches.
Scatter Point	Payload, Launch Site, Class	To see if there is any relationship between the payload of the rockets and their launch sites.
Bar Chart	Orbit, Class	To see which orbits have the highest success rate.
Scatter Point	Flight Number, Orbit, Class	To see if there is any relationship between the orbit and the flight number.
Scatter Point	Payload Mass, Orbit, Class	To see if there is any relationship between the payload of the rocket and the orbit.
Line Chart	Year, Class	To see the launch success trend over time.

The external reference for this process is located at: https://github.com/GigaCat-86/MLCapstone/blob/2a66a1b453f9bc551412b3d7384c1d975f4a734b/EDA%20with%20Visualization.ipynb

EDA with SQL

The SQL queries performed on the data:

- the unique launch site names
- records for launch sites that begin with "KSC"
- the total payload mass of boosters launched by NASA (CRS)
- the average payload mass carried by the FP v1.1 booster
- the date of the first successful landing in a drone ship
- the names of the boosters with a payload between 4,000 and 6,000 kg with successful landings in ground pad
- total number of success and failure outcomes
- the names of the boosters that have carried the maximum booster payload
- records by month for successful in ground pad landing results in the year 2017.
- the number of successful landing outcomes between 2010-06-04 and 2017-03-20

The queries are viewable in the following notebook: https://github.com/GigaCat-86/MLCapstone/blob/59d8c86466d13efd39a5d6a6bb42724277daa6f4/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

On the interactive map, the following markers were added:

Marker Type	Reason
Circle	To indicate the position of each launch site
Marker Cluster	To illustrate the number of successful and unsuccessful launches at each site.
Lines	To show the distance between site KSC LC-39A and the nearest city, highway, railway, and coastline.

The notebook containing the map visualizations can be found in the following link. You will need to download and run it in Jupyter Notebooks (or equivalent) as the maps are not visible directly from github: https://github.com/GigaCat-86/MLCapstone/blob/4a2c9bdd7264d7a14c38112570e651400d8737de/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

A dashboard was created incorporating a dropdown box with a list of launch sites, and a slider with a numerical range of 0 to 10,000.

The interactive elements are used to manipulate:

1. A pie chart that shows the percentage of successful launches when 'All Sites' is selected, and the success vs failure ratio when a particular site is selected.
2. A scatter plot that illustrates the outcome of the launch depending on the payload selected using the slider, or the successful vs unsuccessful launches for each booster site for a particular site and selected payload.

These interactions were chosen to assist the with determining the success rate of launches across sites for different payload values.

The notebook has been included here for reference: https://github.com/GigaCat-86/MLCapstone/blob/4a2c9bdd7264d7a14c38112570e651400d8737de/Module%203%20Dashboard.ipynb

Predictive Analysis (Classification)

Create testing and training sets from the numerical data of the F9 rocket

Create predictive models based on the training data

Evaluate the accuracy of the models against the testing data

Score the accuracy of each model.

The notebook containing the full details of this process can be found at: https://github.com/GigaCat-86/MLCapstone/blob/03d89ebc5d27c805c985fdad540a32849fea74bd/Machine%20Learning%20Prediction.ipynb

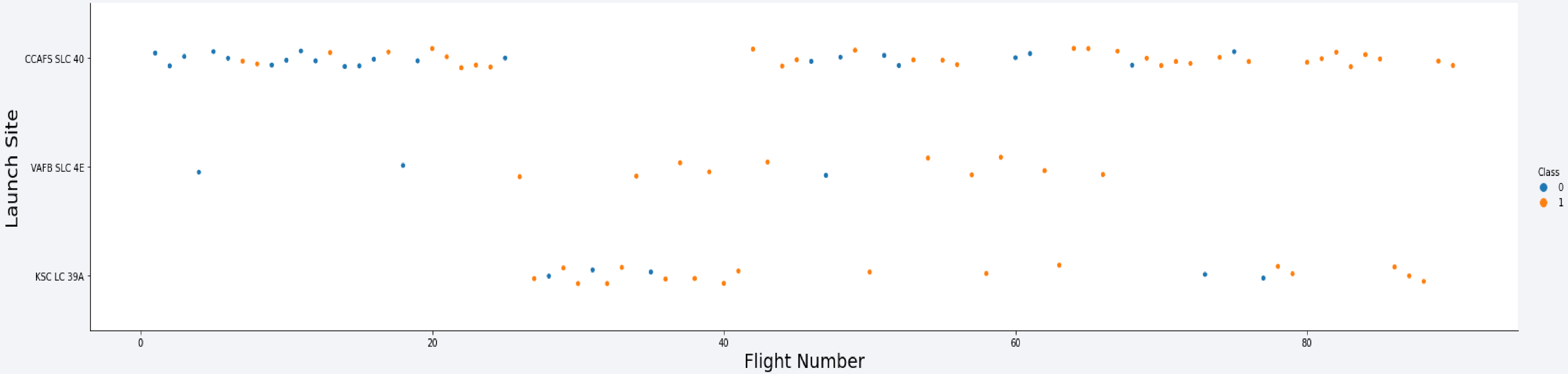
Results

- Exploratory data analysis results
  - Most successful F9 launches occur for the GTO orbit with a payload between 2,000 and 4,000 kg.
- Interactive analytics demo in screenshots
  - Within the payload range of 2,000 and 4,000 kg., most successful launches utilized the FT series of boosters.
- Predictive analysis results
  - The model built using the tree diagram method is the best at predicting whether a launch will end in success or failure.

Section 2

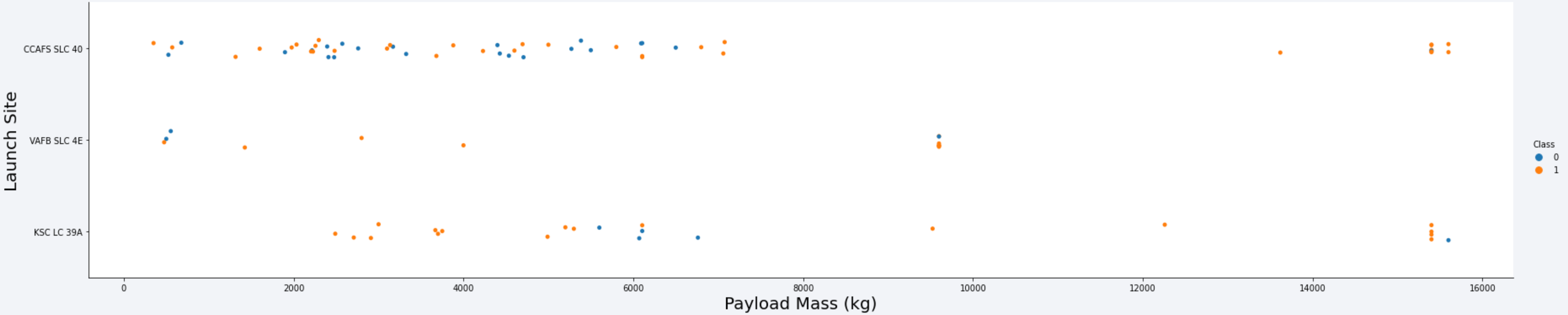
Insights drawn from EDA

Flight Number vs. Launch Site



From the scatter plot, we note that while the vast majority of launches were performed at launch site CCAFS SLC-40, launch site KSC LC-3A had a higher rate of successful launches overall.

Payload vs. Launch Site



The scatter plot above shows the success and failure rate of launches at each site in relation to the payload mass of the rocket launched.

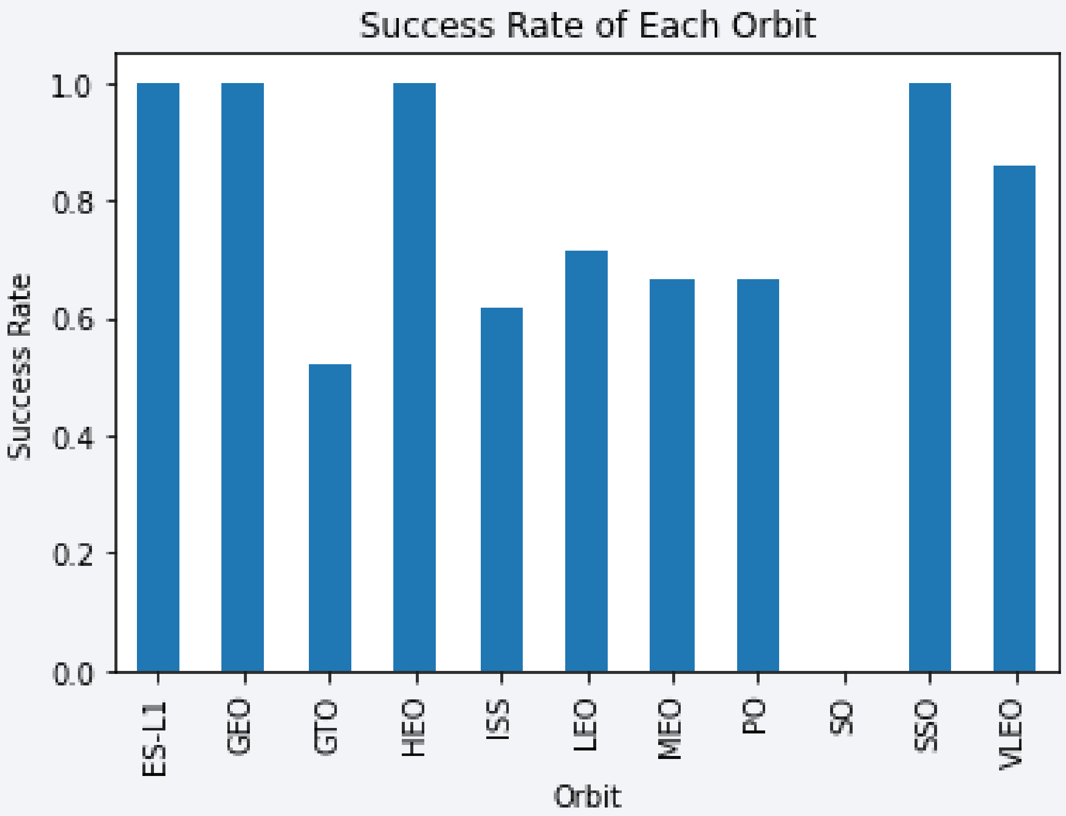
Here we note that the rate of failure seems to increase for payload masses above 4,000 kg. We also note that we see fewer launches performed below the 2,000 kg payload.

Success Rate vs. Orbit Type

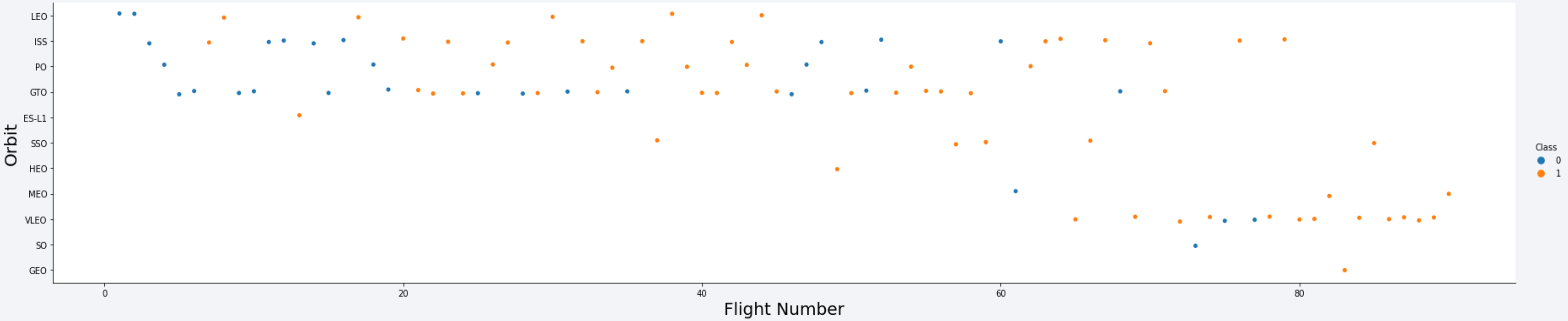
We wanted to see if there was any relationship between the success of the launch and the orbit being used.

The bar chart illustrates the rate of success for each orbit utilized by the Falcon 9 rocket.

The highest success rates are achieved using the ES-L1, GEO, HEO, SSO, and FLFO orbits.



Flight Number vs. Orbit Type

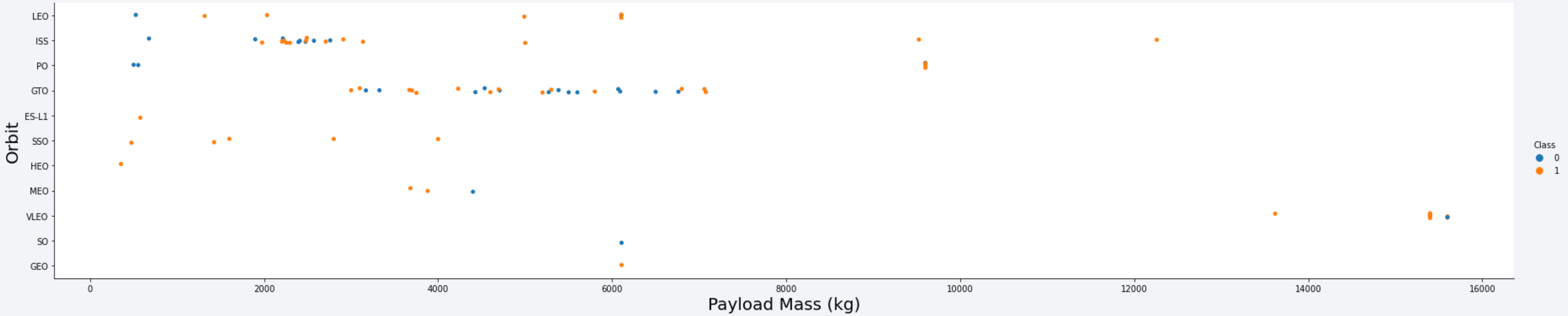


In order to see if there is any relationship between the flight number and the orbit type, we created a scatter plot to map each flight number to the orbit used along with whether it was a succcessful launch.

The higher success rates of the ES-L1, GEO, HEO, SSO, and VLEO orbits can be attributed to the low number of flights.

The GTO orbit, which showed a low success rate in the previous slide, does not seem to be influenced by the number of launches to this orbit.

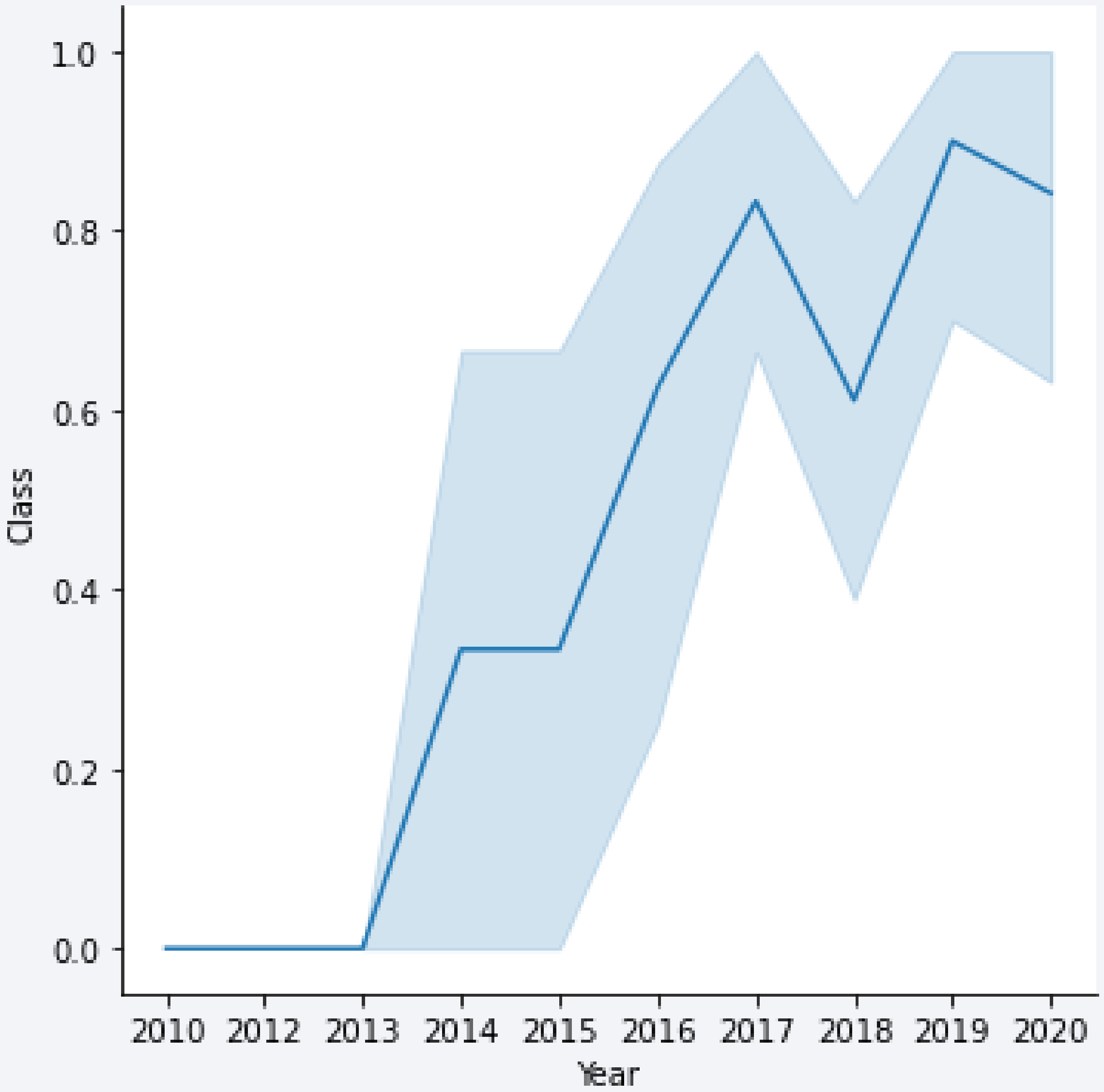
Payload vs. Orbit Type



Next we examined the relationship between the orbit type and the payload of the rocket.

We noted a significant increase in the failure rate in the GTO orbit when utilizing payloads above 4,000 kg. However, the success of launches to the PO and ISS orbits seem to benefit from the increase.

Launch Success Yearly Trend



Launch success showed marked increase beginning in 2014, declined in 2017, and then rebounded in 2018.

Success rates started to decline again sometime after the beginning of 2019 and continued into 2020.

All Launch Site Names

In order to streamline our analysis, we extracted the names of the individual launch sites:

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'KSC'

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Here we can see five records from the same database where the launch site name begins with "KSC".

Note that the first entry is for a launch request from NASA.

Total Payload Mass

We ran a query in the database in order to see the total payload mass of launches requested by NASA. The results of which are illustrated below.

```
%%sql select SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass
from SPACEXTBL
where CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://kzs78114:***@dashdb-txn-sbox-yp-dal09-08.services.dal.bluemix.net:50000/BLUDB
Done.
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

Next we calculated the avereage payload mass of the Falcon 9 booster:

```
%%sql select AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass
from SPACEXTBL
where BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://kzs78114:***@dashdb-txn-sbox-yp-dal09-08.services.dal.bluemix.net:50000/BLUDB
Done.
```

average_payload_mass
2928.400000

First Successful Drone Ship Landing Date

```
%%sql select MIN(DATE) AS First_Successfule_Drone_Ship_Landing
from SPACEXTBL
where LANDING__OUTCOME = 'Success (drone ship)'
```

```
 * ibm_db_sa://kzs78114:***@dashdb-txn-sbox-yp-dal09-08.services.dal.bluemix.net:50000/BLUDB
Done.
```

first_successfule_drone_ship_landing
2016-04-08

We performed a query on our data and discovered that the first succcessful landing onto a drone ship was achieved on 8th April 2016.

Successful Ground Pad Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on a ground pad and had payload mass greater than 4000 but less than 6000 are:

booster_version	payload_mass__kg_
F9 FT B1032.1	5300
F9 B4 B1040.1	4990
F9 B4 B1043.1	5000

Total Number of Successful and Failure Mission Outcomes

Here we see the total number of successful and failure mission outcomes.

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass are:

boosters_with_maximum_payload_mass
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2017 Launch Records

Below are the successful ground pad landing results by month for the year 2017.

launch_month	landing__outcome	booster_version	launch_site
February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
December	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Here we can see a count of successful landing outcomes between the date 2010-06-04 and 2017-03-20, noted in descending order

successful_landing_type	number_of_successful_landings
Success (drone ship)	5
Success (ground pad)	3

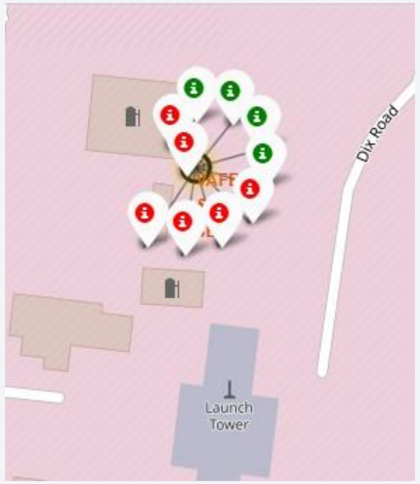
Section 4

Launch Sites Proximities Analysis

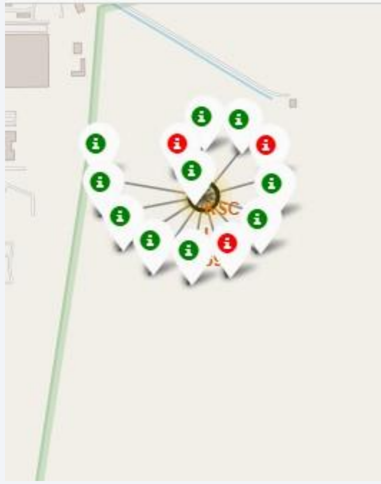
Map of All Launch Site Locations

All of the launch sites appear to be located near to coastlines in the continental United States, mainly in the state of Florida.

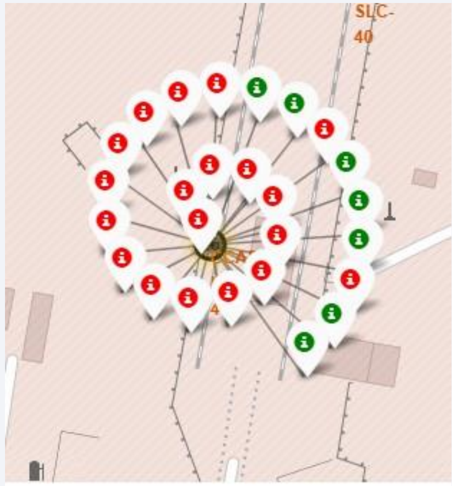
Launch Outcomes



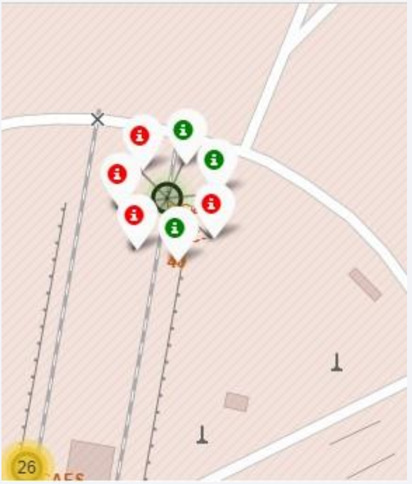
VAFB SLC-4E



KSC LC-39A



CCAFS LC-40

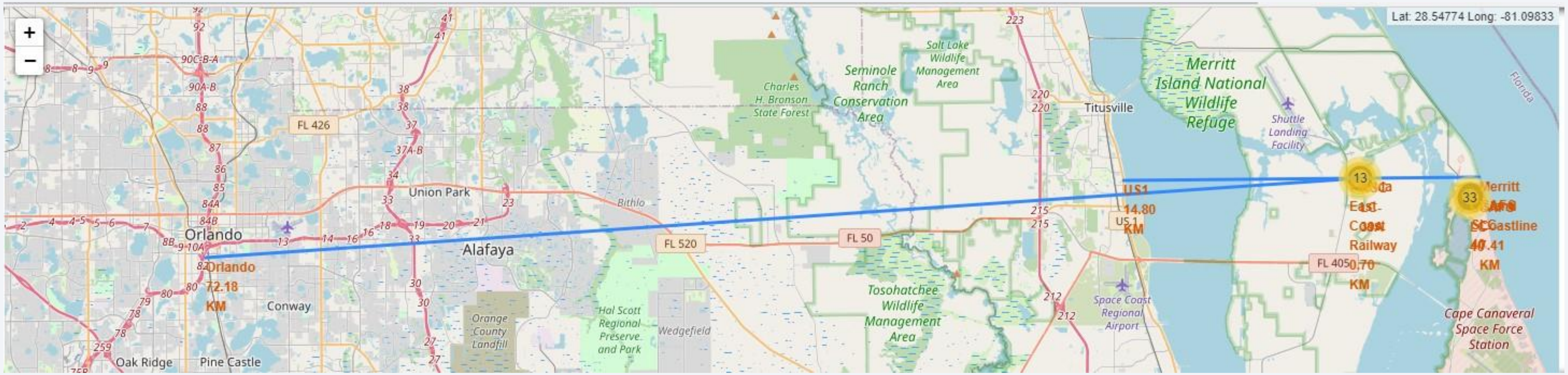


CCAFS SLC-40

Success rates for each site were added to our map with the successful launches shown in green, and unsuccessful launches shown in red.

As we can see, site KSC LC-39A has the most success per launch.

Proximities to Site KSC LC-39A



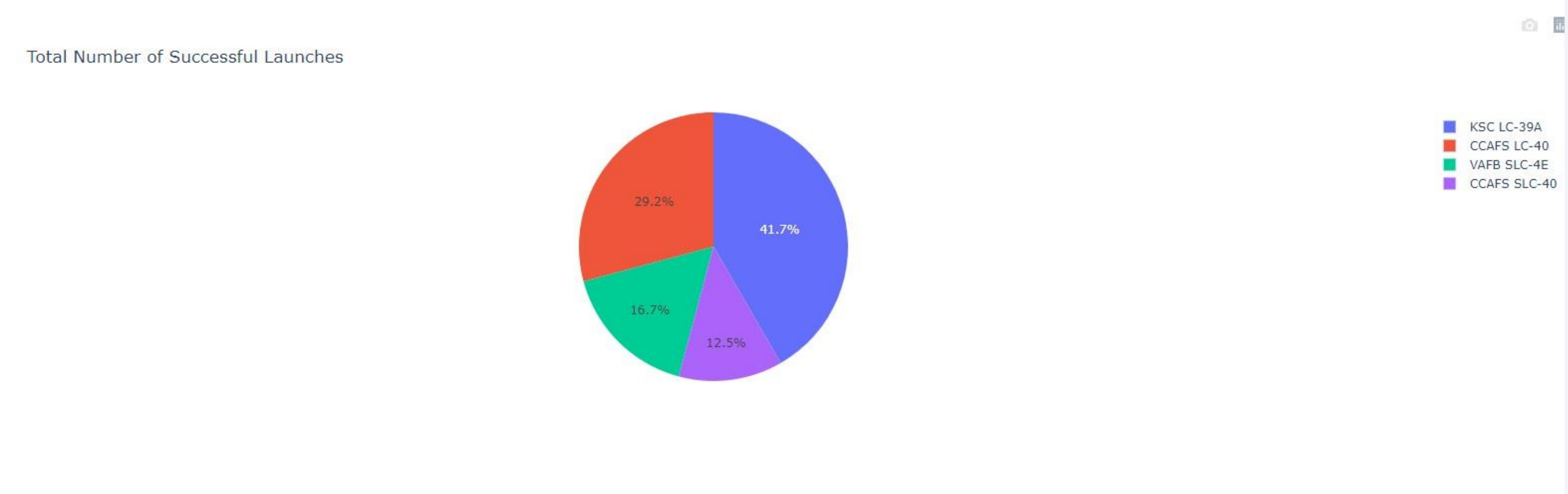
Based on its success rate, we decided to examine its location. The site is located:

- 72.18 km from its nearest city, Orlando.
- 14.80 km from the nearest highway, US1
- 0.70 km from the nearest railway, Florida East Coast Railway
- 7.41 km from the nearest coastline, Merritt island Coastline

Section 5

Build a Dashboard with Plotly Dash

Rate of Success Across All Launch Sites



As illustrated in the chart above, the highest success rate was achieved by the KSC LC-39A launch site.

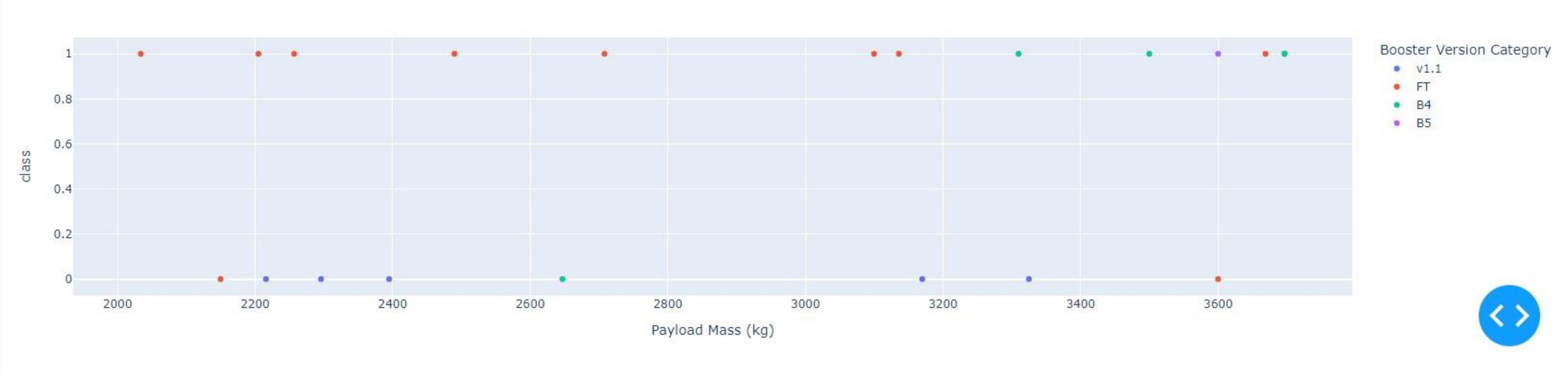
Success Rate at the KSC LC-39A Launch Site

The pie chart shows the success vs failure rate of launches at the KSC LC-39A launch site.

The blue shaded area is the success percentage and the red shaded area is the failure percentage.

As we can see, this site achieved a very high rate of launch success (76.9%)

Payload Success and Failure for Booster Range of 2,000 – 4,000 kg



As we noted in our previous analysis, we tended to see a higher rate of failure for launches above 4,000 kg., and that very few launches are performed for payloads less than 2,000 kg.

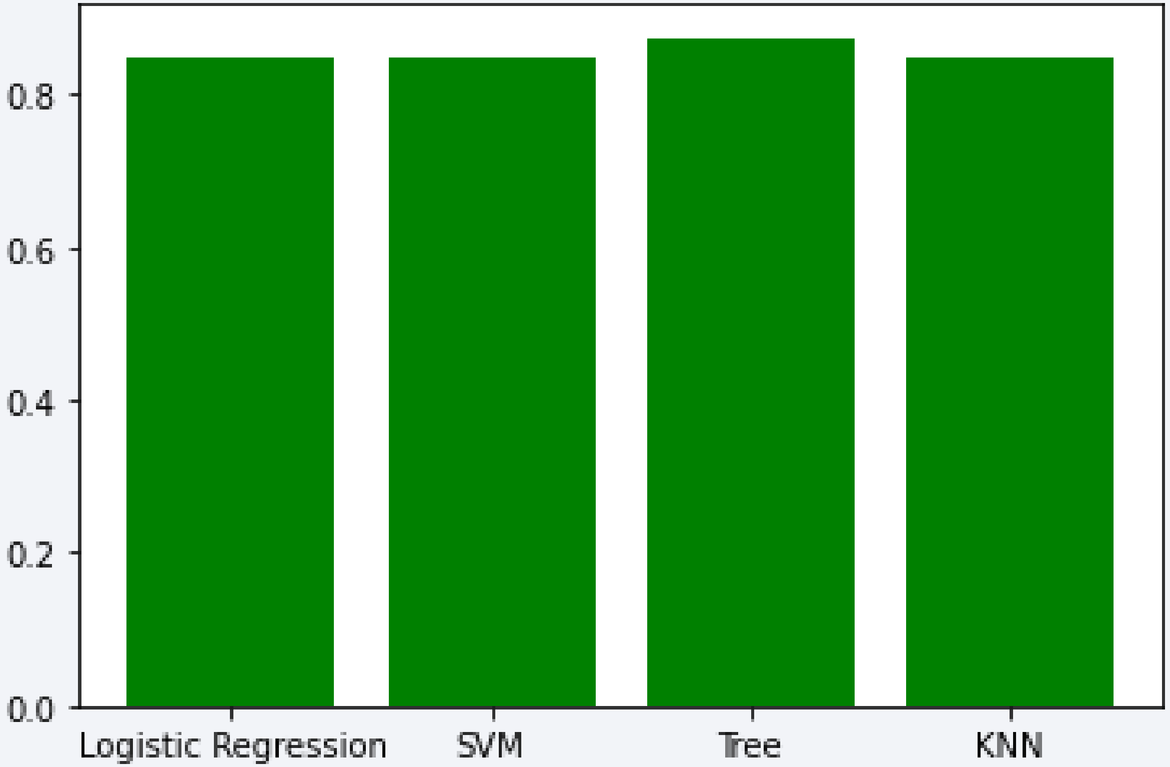
Our scatter plot shows the boosters used for the successful launches (class 1) and unsuccessful (class 0) in the aforementioned payload range. Note that the FT booster has the most successful results.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

From our analysis of each prediction method, the Tree method gives us slightly better results over other models.



Confusion Matrix – Tree Model

Our tree model produces the following results when scored against the training data, our model produced:

- Three accurate predictions for landing failure
- Three incorrect predictions for landing failure
- No incorrect predictions for successful landing
- Twelve accurrate predictions for successful landings

Conclusions

- Successful launches of the F9 rocket have a higher success rate when utilizing the FT series of boosters.
- Launches have a better chance of success when launched with payloads between 2,000 and 4,000 kg for the GTO orbit.
- Landing success is higher on drone ships than ground pads.
- Launch sites should be built close to coastlines in order to maximize the use of drone ships.
- Using a decision tree based model we can predict the success of a launch with a high degree of accuracy.

Appendix

- Data Collection with API Dataset: https://github.com/GigaCat-86/MLCapstone/blob/59d8c86466d13efd39a5d6a6bb42724277daa6f4/dataset_part_1.csv
- Data Collection with Web-Scraping Dataset: https://github.com/GigaCat-86/MLCapstone/blob/59d8c86466d13efd39a5d6a6bb42724277daa6f4/spacex_web_scraped.csv
- Exploratory Data Analysis Dataset: https://github.com/GigaCat-86/MLCapstone/blob/2a66a1b453f9bc551412b3d7384c1d975f4a734b/dataset_part_2.csv
- Exploratory Data Analysis with Visualization Dataset: https://github.com/GigaCat-86/MLCapstone/blob/59d8c86466d13efd39a5d6a6bb42724277daa6f4/dataset_part_3.csv

Thank you!