IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

James Blasich
9/18/2021

Outline

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

Executive Summary

- Methodologies:
    - All algorithmic programming for data collection, data analysis, and data visualization was performed using the Python programming language with Jupyter Notebooks and IBM Cloud services such as Watson Studio and DB2.
    - Data was collected directly from SpaceX through their API, and from scraping webpages.
- Conclusions:
    - SpaceY can predict whether a rocket booster will be recovered successfully with better than 80% accuracy.
    - These predictions are based on several factors such as payload mass, booster type, launch site location, and orbital path that largely impact the recoverability of the rocket booster.
    - Accurate predictions of recovery allow for excellent cost estimates in order to competitively price launches and gain market share.

Introduction

- SpaceY is looking to become a competitor in the market of low-cost orbital launches by means of successful recovery of reusable rocket boosters.
- The primary challenge for SpaceY in entering this competitive market is to predict whether a booster will have a successful recovery.
- Successful booster recoveries significantly reduce the overall program cost by being able to reuse the primary rocket section, thus saving on materials and labor.
- The goal of this data science project is to build and train an algorithmic model that can accurately predict the chance of successful booster recoveries in order to better estimate program costs and determine if SpaceY can take a competitive advantage in the orbital launch business market.

Section 1

Methodology

Methodology

Executive Summary

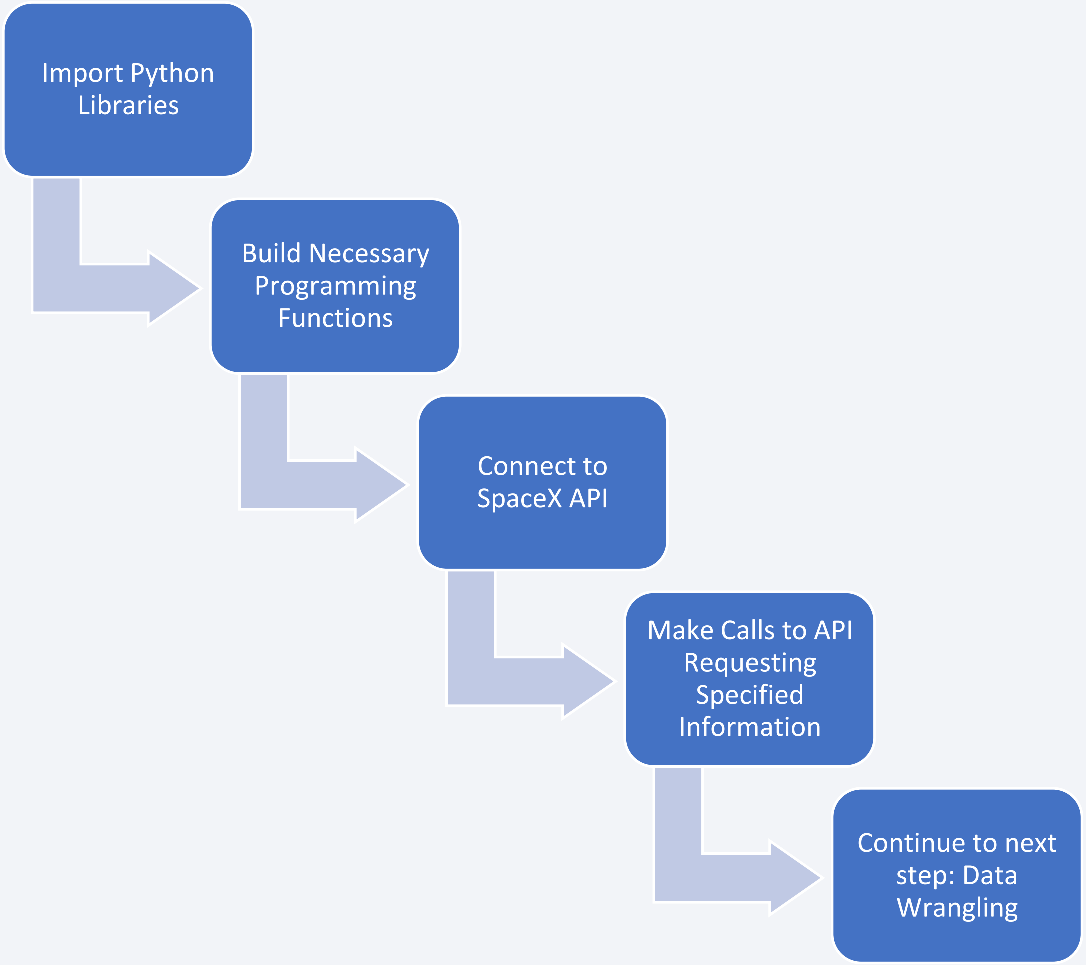
- Data collection methodology:
  - JSON files were collected from the SpaceX API and Wikipedia tables were webscraped using BeautifulSoup.
- Perform data wrangling
  - Data was cleaned and organized into dataframes using Pandas and NumPy.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Multiple machine learning models were built using the SciKitLearn library for Python.

Data Collection

- Data was collected from two sources:
  - SpaceX REST API
    - Calls were made to the API using the Python programming language.
    - This data comes directly from SpaceX's own database.
  - Wikipedia
    - SpaceX launch data tables were scraped directly from Wikipedia webpages using the Python programming language and the BeautifulSoup library for Python.

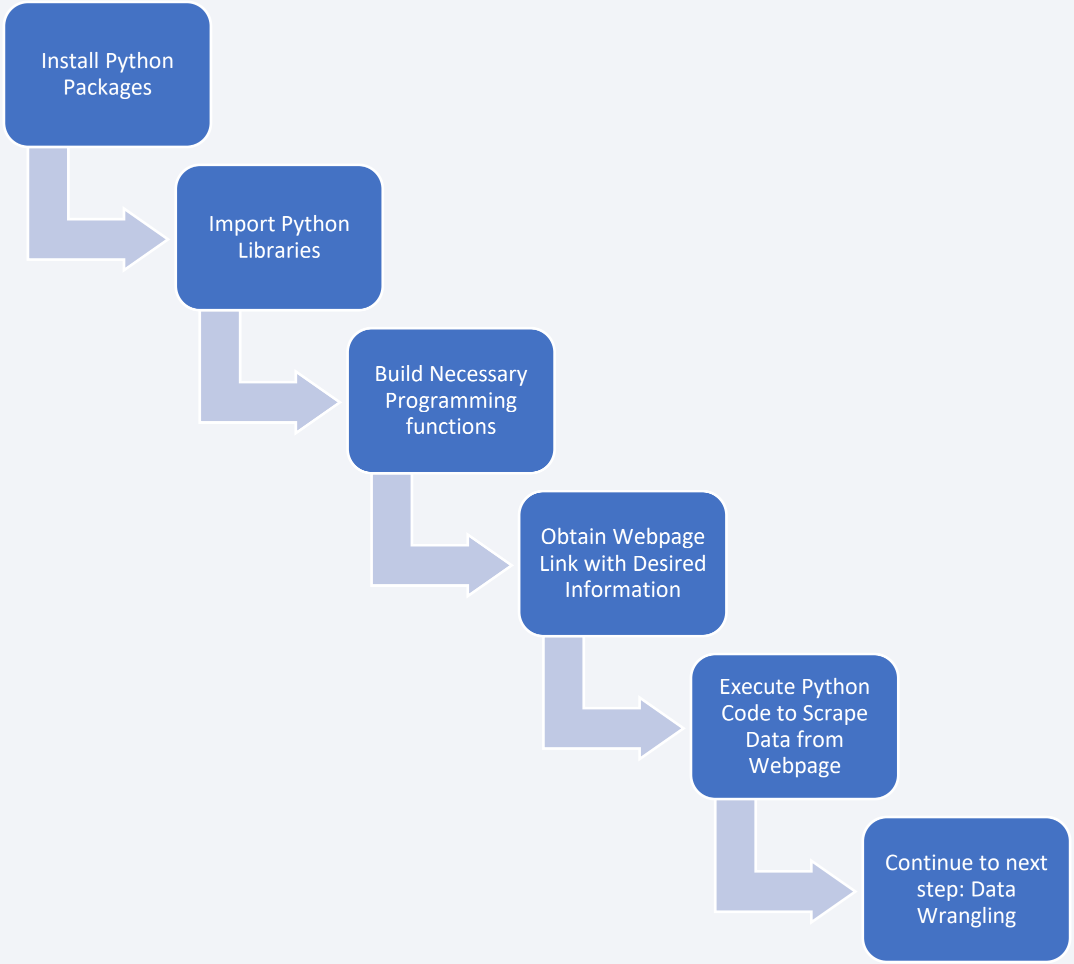
Data Collection – SpaceX API

- Process of data collection from SpaceX database using SpaceX REST API.
- REST API notebook link:
  - https://github.com/Trade4Bread/Space_Y/blob/master/API%20Data%20Collection%20Lab.ipynb



Data Collection - Scraping

- Process of collecting data by webscraping Wikipedia tables using Python and BeautifulSoup.
- Link for Webscraping with BeautifulSoup notebook:
  - https://github.com/Trade4Bread/Space_Y/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Extract relevant information from collected datasets.
- Parse, clean, and normalize data into useable Pandas dataframes with desired categorical information.
- Handle all null or missing values such as:
  - Missing payload values were treated with the mean of all known payloads.
- Links for Data Wrangling related notebooks:
  - https://github.com/Trade4Bread/Space_Y/blob/master/API%20Data%20Collection%20Lab.ipynb
  - https://github.com/Trade4Bread/Space_Y/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb
  - https://github.com/Trade4Bread/Space_Y/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

- Numerous charts were generated to gain detailed insight into the multiple launch variables that influence successful booster recoveries including:
  - Flight number vs payload mass
  - Flight number vs launch site
  - Flight number vs orbit
  - Payload mass vs launch site
  - Payload mass vs orbit
  - Success rate by orbit type
  - Success rate by year
- Link for EDA with Data Visualization notebook:
  - https://github.com/Trade4Bread/Space_Y/blob/master/EDA%20with%20Visualization.ipynb

EDA with SQL

- Many SQL queries were required to select relevant information to better understand the dataset and its correlation to successful booster recoveries. Some notable queries were:
  - First successful ground pad landing date
  - Booster versions with successful drone ship landings
  - Booster versions carrying maximum payloads
  - Failed drone ship landings and their associated booster versions and launch sites
  - Early program landing outcome totals
- Link for EDA with SQL notebook:
  - https://github.com/Trade4Bread/Space_Y/blob/master/EDA%20with%20SQL%20Lab.ipynb

Build an Interactive Map with Folium

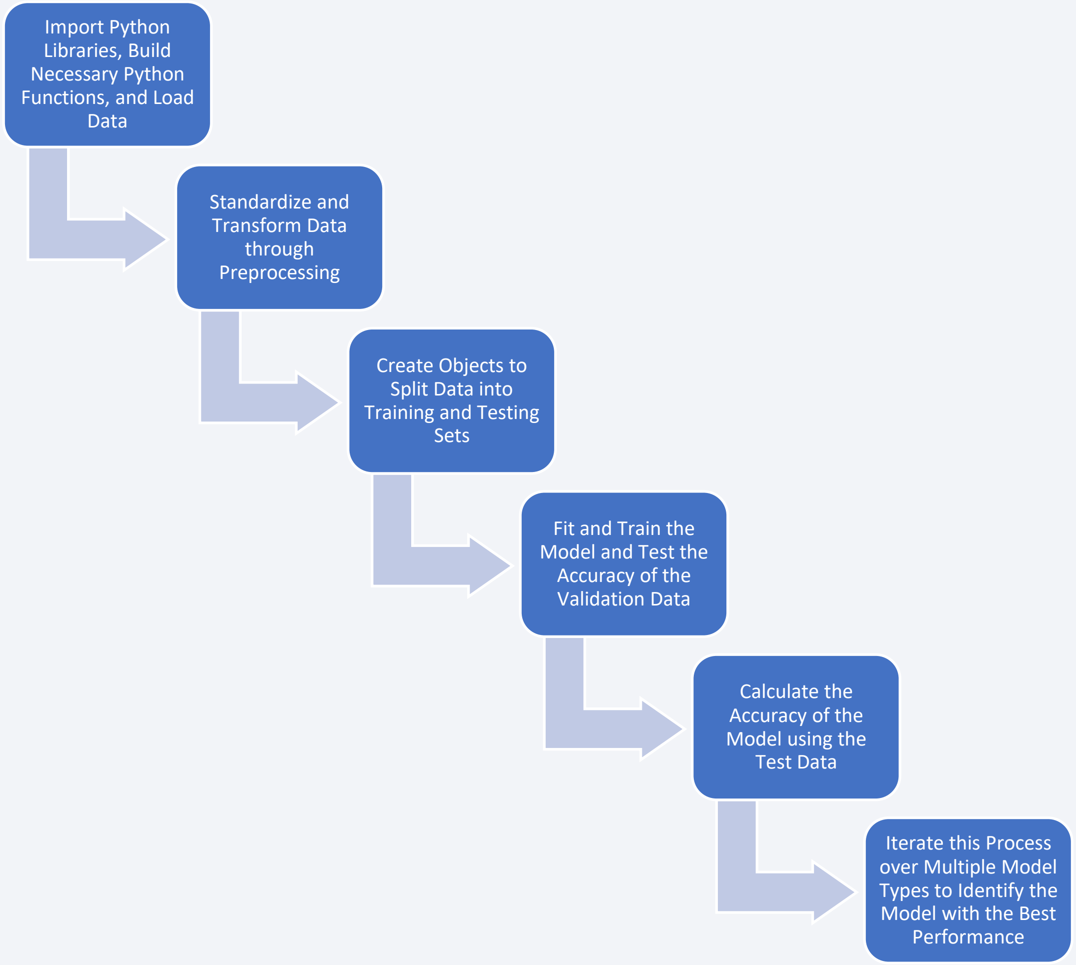
- An interactive map of launch sites and individual launches was created using the Folium package for Python. The following markers were added to the map to create greater visual understanding of the launch sites:
  - Individual circle markers for each launch site
  - Colored launch indicators for all successful/failed launches at each launch site
  - Polylines and distance markers to nearby landmarks
- These markers were added to create quick visual understanding of the geographic location of launch sites and the outcomes of each launch from the associated site.
- Link for Interactive Visual Analytics with Folium Map notebook:
  - https://github.com/Trade4Bread/Space_Y/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- The Dashboard consisted of four elements:
  - A dropdown selection for all or individual launch sites
  - A pie chart representing the success rate of launches from the selected site
  - A slider allowing for selection of different ranges of payloads
  - A scatter plot chart showing the correlation between payload and success for the selected payload and launch site
- Utilizing the dashboard with these elements allows for numerous quick visual representations of the data without having to produce high numbers of individual charts. It allows the user to gain deeper understanding of the data while ultimately requiring much less programming intensive work.
- Links for Dashboard built using Plotly Dash:
  - Code
    - https://github.com/Trade4Bread/Space_Y/blob/master/Plotly%20Dash%20Dashboard.ipynb
  - Dashboard Visual
    - https://github.com/Trade4Bread/Space_Y/blob/master/Dashboard_Full.PNG

Predictive Analysis (Classification)

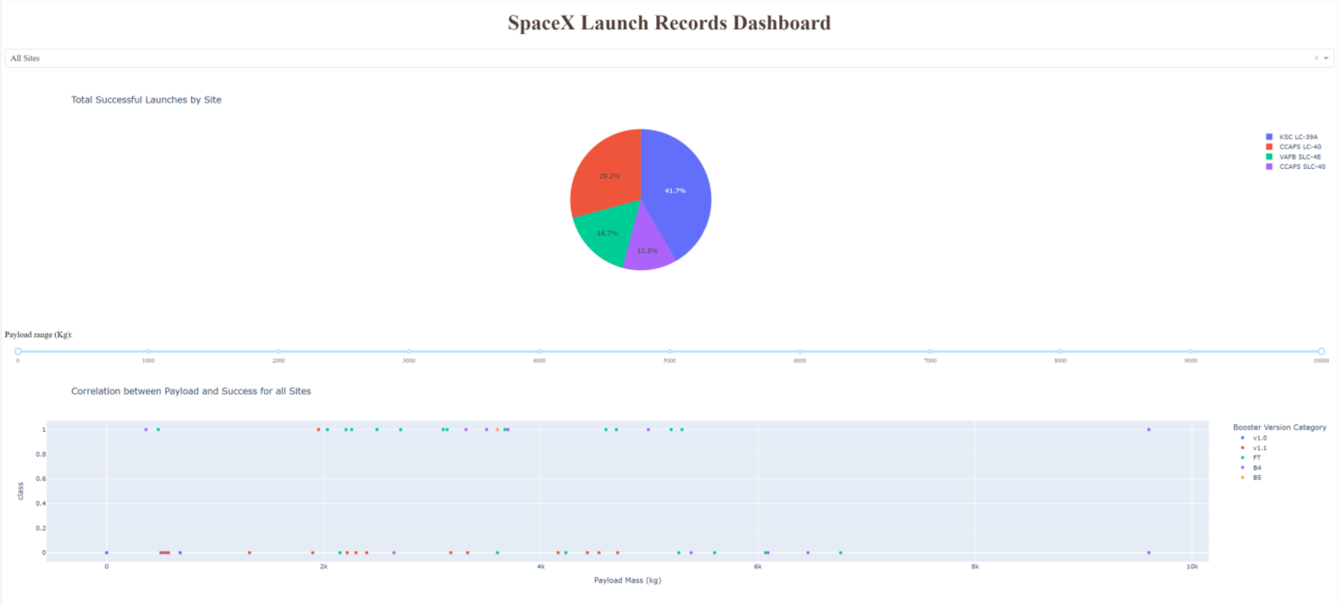
- Algorithmic prediction models were built using the Python programming language and the SciKitLearn package for Python.
- Four different machine learning algorithms were built:
  - Logistic regression
  - Support vector machine
  - Decision tree classifier
  - K nearest neighbor
- Models were evaluated using the score method from the SciKitLearn package to determine the most effective models.
- Link for machine learning predictive analysis lab:
  - https://github.com/Trade4Bread/Space_Y/blob/master/Machine%20Learning%20Prediction.ipynb



Results

- As a result of exploratory data analysis, many factors have been identified that contribute to the successful recovery of a rocket booster after launch. Some factors are:
    - Launch site
    - Recovery method
    - Booster type
    - Flight number
    - Payload mass
    - Orbit type
- Predictive analysis through machine learning algorithms shows high accuracy for all model types, with no one model performing better than others.

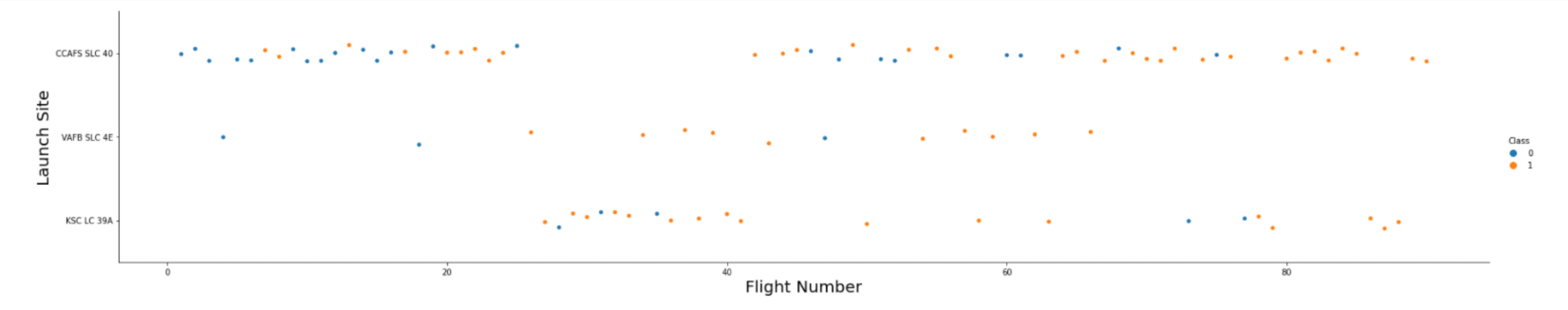
Dashboard Visual



Section 2

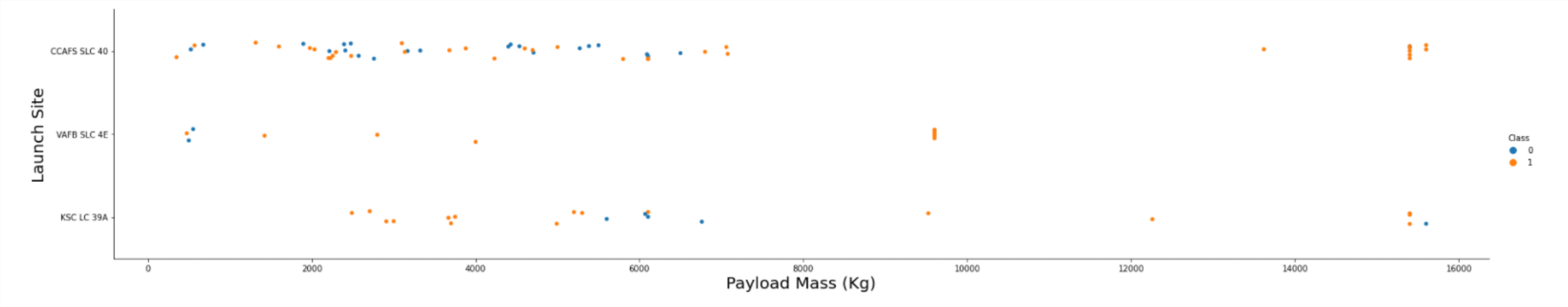
Insights drawn from EDA

Flight Number vs. Launch Site



- Early-stage flights had a low success rate.
- When KSC LC-39A was introduced as a launch site at approximately 25 flights, success rates improved across all sites.
- From flight 62 forward, there has been only four failures. We can state that after approximately 60 flights, success rates stabilize near peak performance.

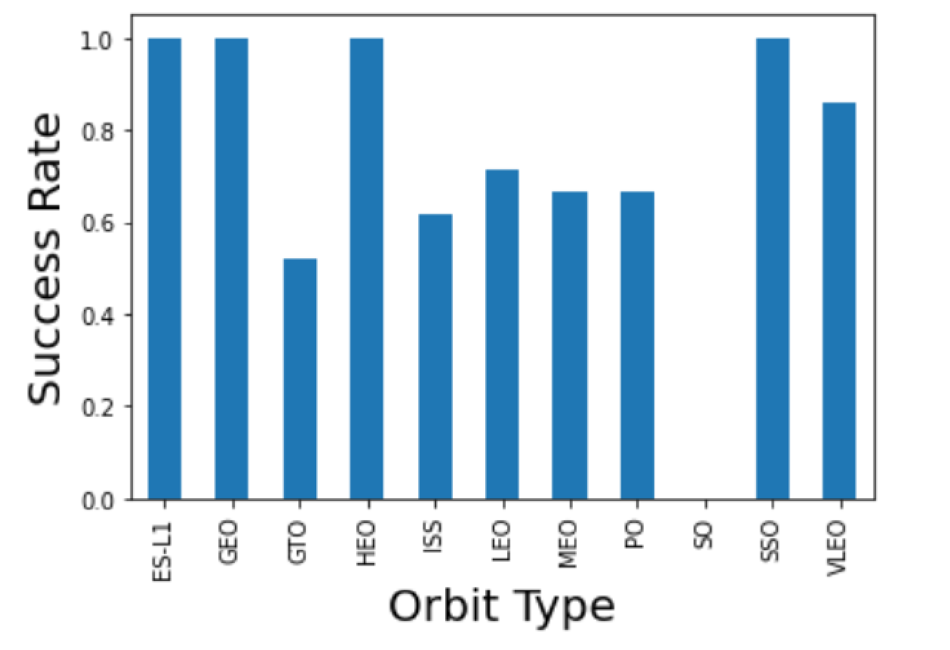
Payload vs. Launch Site



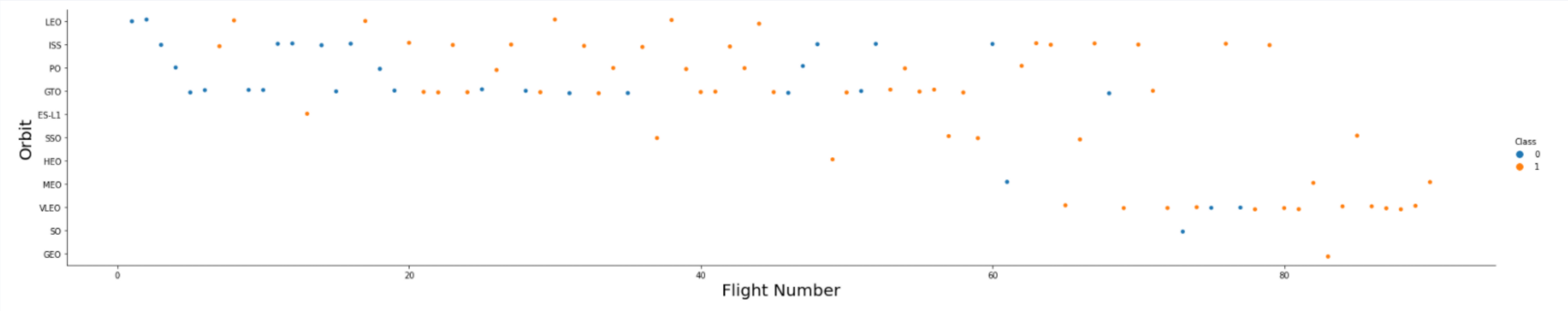
- Launch site CCAFS SLC-40 has mixed results for payloads less than 8000kg, but near perfect results for payloads exceeding 8000kg.
- Launch site KSC LC-39A has performed well except with payload localized around 6000kg
- Payloads above 8000kg have been highly successful for all sites and could signal these heavy loads are only used when confidence of success is high.

Success Rate vs. Orbit Type

- Orbit type can be correlated to success rate.
- ES-L1, GEO, HEO, and SSO have perfect mission success rates.
- SO has had zero successes.
- However, these success rates are highly tied to the quantity of launches performed for each orbit as you will see in the next slide.

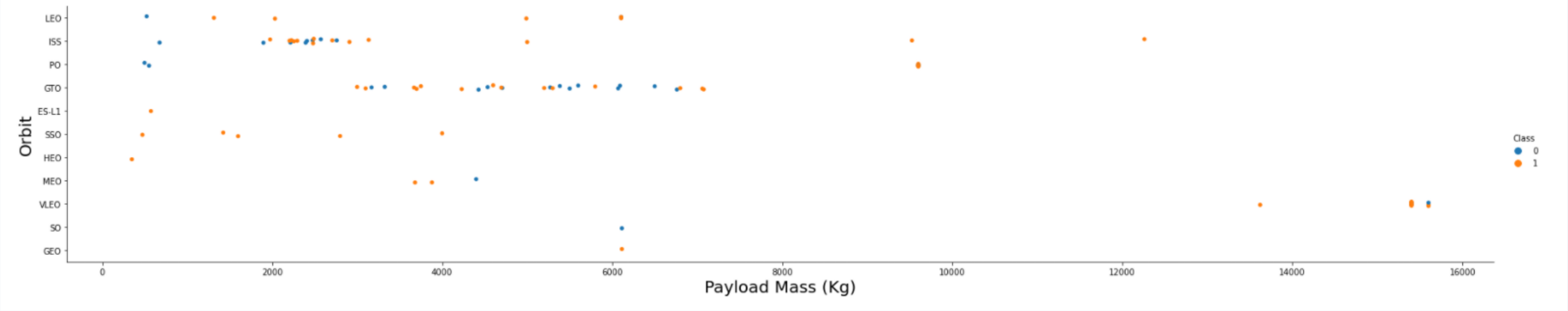


Flight Number vs. Orbit Type



- Early flights adhered to a small group of orbital paths before expanding to a full complement of orbit types around flight 60.
- Most of the high success or failure rate orbit types from the previous slide have only one flight, with the exception of SSO which has five successes and zero failures.
- The standout performer is VLEO which has a sizeable sample of flights and high success rate from those flights.

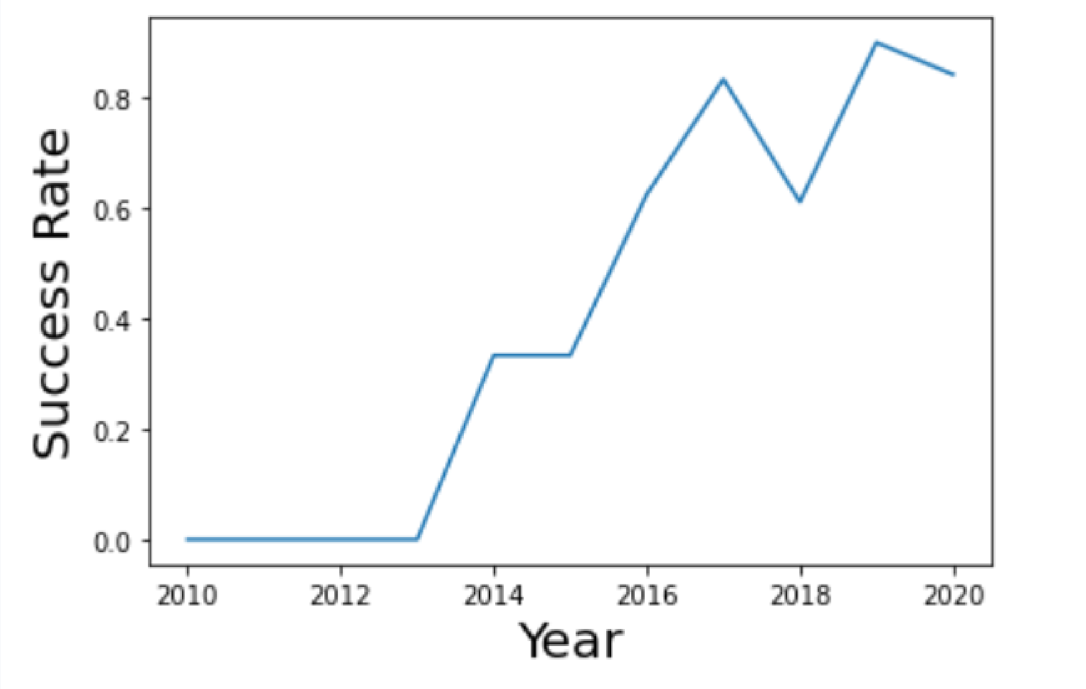
Payload vs. Orbit Type



- Heavy payloads tend to have a negative effect on GTO orbits.
- Heavy payloads have a positive effect on PO, ISS, and LEO orbits.
- VLEO orbits are exclusively heavy payloads.
- SSO orbits are exclusively lighter payloads.

Launch Success Yearly Trend

- Success rate steadily increases from program onset for about 4-5 years.
- At this point it begins to have its ups and downs.
- This could mean a peak in success rate has been reached and is shown to have stabilized around 80% success.



All Launch Site Names

Display the names of the unique launch sites in the space mission

In [6]:
```
%sql select distinct(LAUNCH_SITE) from SPACEXDATASET
```

```
 * ibm_db_sa://vvt88191:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[6]:

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

- There are four unique launch sites listed in the SpaceX program database.
- They are shown above as pulled directly from the dataset.

Launch Site Names Begin with 'CCA'

```
In [7]: %sql select * from SPACEXDATASET where LAUNCH_SITE like 'CCA%' limit 5
```

```
 * ibm_db_sa://vvt88191:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[7]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Above is an example of 5 lines from the dataset that are characterized by the launch site beginning with CCA.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [8]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXDATASET where CUSTOMER like 'NASA (CRS)'

 * ibm_db_sa://vvt88191:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
Out[8]:
```

1
45596

- Shown above is total payload mass of all launches performed for the customer NASA.
- Total mass is 45,596kg

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
In [9]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXDATASET where BOOSTER_VERSION like '%F9 v1.1'

 * ibm_db_sa://vvt88191:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
Out[9]:
```

1
2928

- Displayed above is the average payload mass for launches using the F9 V1.1 booster.
- Average mass is 2,928kg which is considered a light load.

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [12]:
```
%sql select min(DATE) from SPACEXDATASET where LANDING__OUTCOME like 'Success (ground pad)%'
```

```
 * ibm_db_sa://vvt88191:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[12]:

1
2015-12-22

- The first successful recovery of a booster on a ground pad was on 12/22/2015.
- This date is several years after the start of the program.

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [14]:
```
%sql select distinct(BOOSTER_VERSION) from SPACEXDATASET where LANDING__OUTCOME like 'Success (drone ship)%' and PAYLOAD_MASS__KG_ betwee
n 4000 and 6000
```

```
 * ibm_db_sa://vvt88191:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[14]:

booster_version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- Listed are the booster versions which have completed successful landings to drone ships with an initial payload mass of between 4000kg and 6000kg.
- Four different boosters have had successful recoveries within this payload range.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
In [19]: %sql select count(LANDING__OUTCOME) from SPACEXDATASET where LANDING__OUTCOME like '%Success%' or LANDING__OUTCOME like '%Failure%'

 * ibm_db_sa://vvt88191:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
Out[19]:
```

1
71

- Here we see the total number of missions that have resulted in either a success or failure outcome.
- 71 missions have succeeded or failed.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [26]:
```
%sql select distinct(BOOSTER_VERSION) from SPACEXDATASET where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXDATASET)
```

```
 * ibm_db_sa://vvt88191:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[26]:

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

- Above are all the different booster versions that have carried the maximum payload mass.

2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
In [23]: %sql select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE from SPACEXDATASET where LANDING__OUTCOME like 'Failure (drone ship)%' and DAT
E like '2015%'

 * ibm_db_sa://vvt88191:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[23]:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Listed above are the launches that had failed recoveries to a drone ship in the year 2015 including their booster version and launch site.
- Both failures resulted from the CCAFS LC-40 launch site.

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

In [25]:
```
%sql select LANDING__OUTCOME, count(LANDING__OUTCOME) from SPACEXDATASET where DATE between '2010-06-04' and '2017-03-20' group by LANDIN
G__OUTCOME order by count(LANDING__OUTCOME) DESC
```

```
 * ibm_db_sa://vvt88191:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[25]:

landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- Shown here are the landing outcomes from missions launched between 6/4/2010 and 3/20/2017.
- They are listed in descending order by total number of like outcomes.
- During this time, a high number of launches made no attempt of recovering the booster.

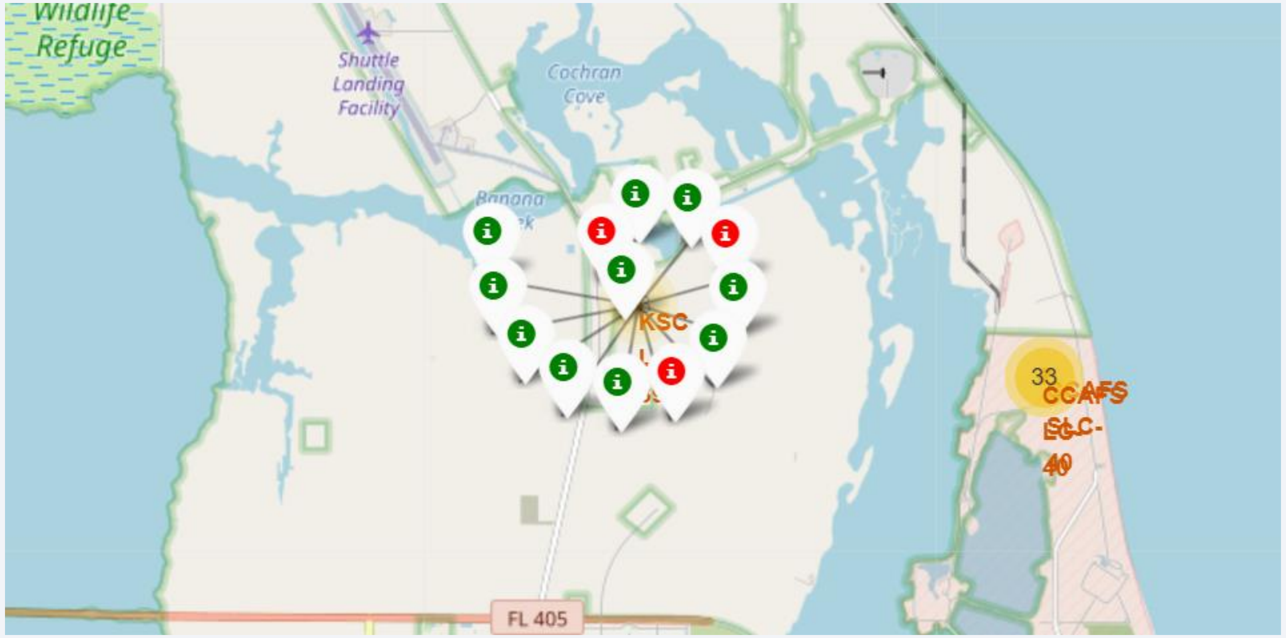
Section 4

Launch Sites Proximities Analysis

Folium Site Mapping

- Launch sites are coastal, and more southernly.
- The lower latitude of the Florida coast launch sites appear to make it the favorite for a higher volume of launch missions.

Folium Site Mapping



- Shown are launch outcomes at the most successful launch site, KSC LC-39A. Green indicates a success, and red a failure.
- Being slightly further from the coastline appears to have a positive result on the success of the launch.

Folium Site Mapping

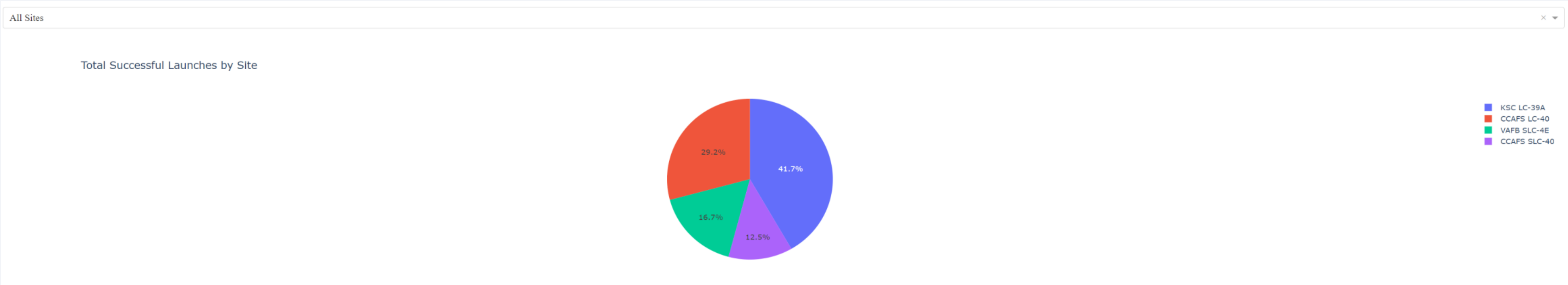

```
In [38]: # distance_railway = calculate_distance(lat1, lon1, lat2, lon2)
         lat1 = 28.57386
         lon1 = -80.58517
         lat2 = 28.563197
         lon2 = -80.576820
         distance_railway = calculate_distance(lat1, lon1, lat2, lon2)
         distance_railway

Out[38]: 1.4394602666848335
```

- Launch sites keep close proximity to railways, in this case just over 1 kilometer.
- This ensures easier delivery of large payloads and rocket parts.

Section 5

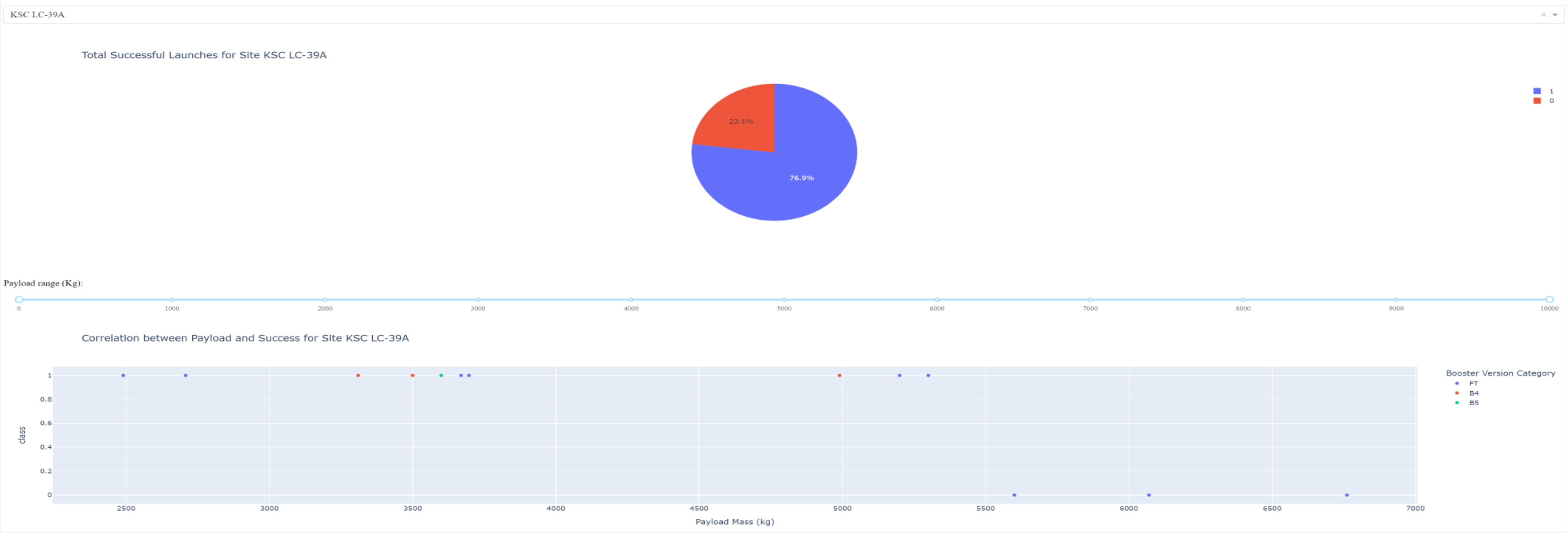
Build a Dashboard with Plotly Dash

Dashboard Visualization



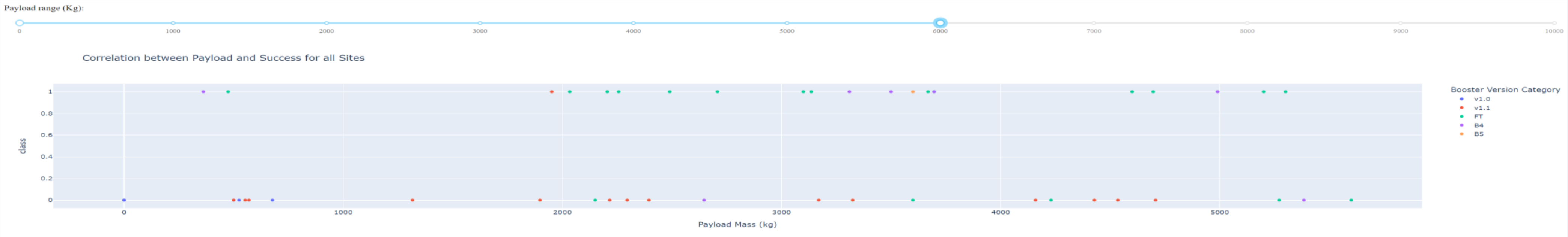
- Launch site KSC LC-39A has more successful launches than any other launch site.
- KSC LC-39A accounts for 41.7% of successful launches while accounting for only 25% of the total launches.
- From this we can determine that the selected launch site can have a high impact on the potential success of the launch and recovery mission.

Dashboard Visualization



- KSC LC-39A has a successful launch ratio of 76.9%. This is significantly higher than the next nearest site at 42.9%.
- Of note is that all successful launches came with a payload of less than 5500kg. No launches above this mass were successful, potentially limiting the capabilities of the site.

Dashboard Visualization



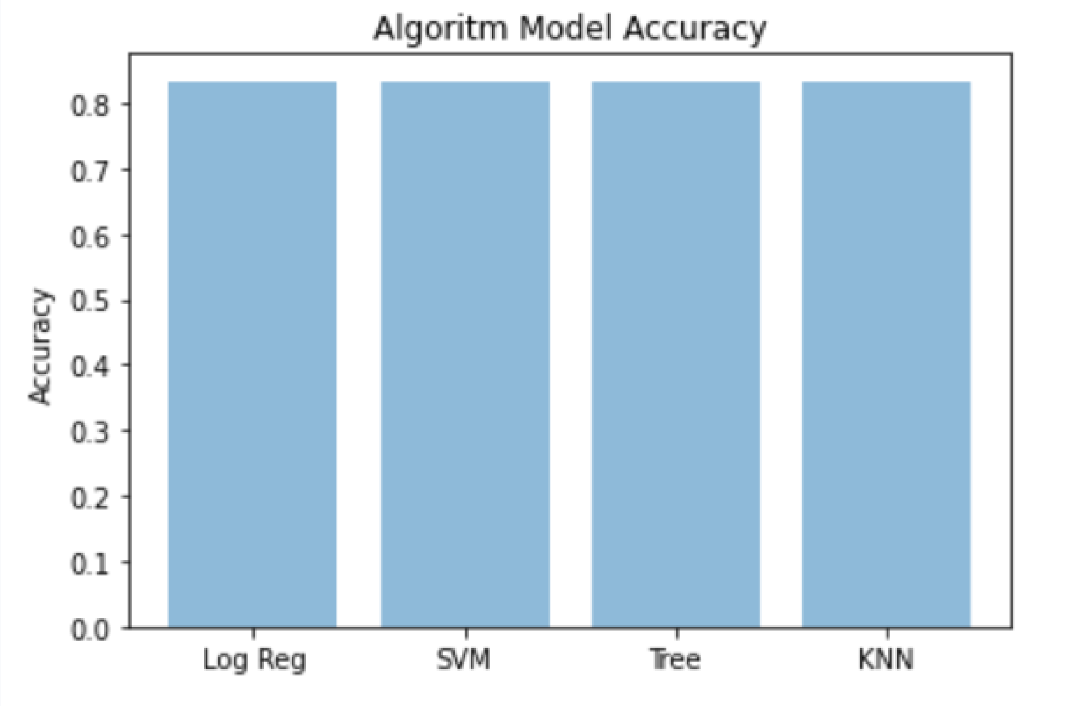
- Booster version can be seen to have a high impact on success rate. With payloads of less than 6000kg, booster version FT has by far the most successes.
- Payload mass also highly influences success rate. Payloads over 6000kg only have a singular success, and with booster version B4.

Section 6

Predictive Analysis (Classification)

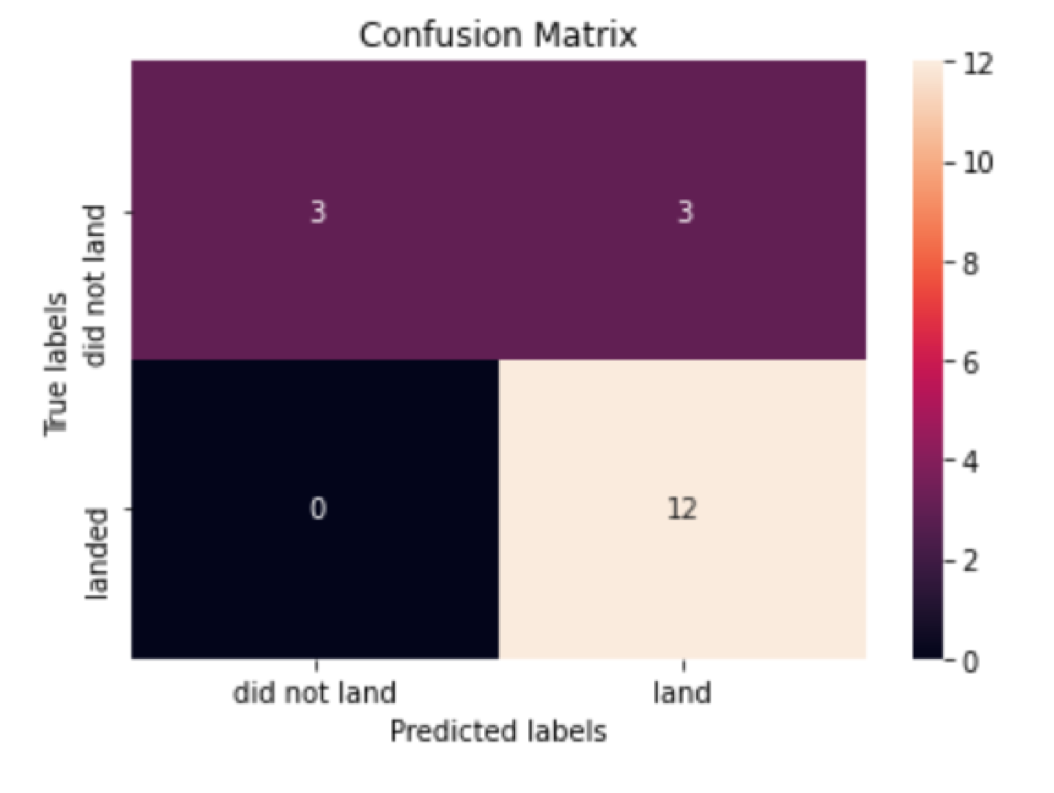
Classification Accuracy

- Accuracy for all algorithmic models is identical at 83.33%.
- This gives a high probability of accurately predicting the outcome of recovering a booster after launch.



Confusion Matrix

- The confusion matrix is identical for all algorithms.
- All launches predicted to not land were accurate.
- Three launches predicted to land were unsuccessful in their landing attempts.
- Twelve launches predicted to land successfully did so.
- Three inaccuracies out of 18 launch predictions gives model accuracies of 83.33%.



Conclusions

- Algorithms predicting successful booster recoveries have high accuracy in correctly determining launch and recovery outcome, with better than 80% prediction accuracy across all algorithmic models.
- Choosing an advantageous launch site is proven to have a high impact on successful recovery of rocket boosters.
- Number of launches highly influences mean success rate of recoveries. Early program launches tend to be less successful, with optimum success occurring beyond 60 total launches. However, the insights gained from this study can aid in greatly reducing the time to high mean success rate.
- Launch costs can be significantly reduced and accurately predicted based on these findings, allowing for deep insight into the potential competitiveness of SpaceY launch pricing.

Appendix

- Full detailed project analysis can be viewed at the following link:
  - https://github.com/Trade4Bread/Space_Y/tree/master

Thank you!