



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Galen Shen
September 6, 2024



Outline

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

Executive Summary

Methodologies

- Data Collection (web scraping, SpaceX Rest API)
- Data Wrangling (filtering, missing values handling, one hot encoding)
- Exploratory Data Analysis (data visualizations and SQL)
- Interactive folium map
- Plotly dashboard
- Predictive analysis (classification)

Summary of all results

- EDA allowed for identification of variables to test in model for launch outcomes
- The decision tree model among the four trialed performed the best for the task of accurately predicting launch outcome

Introduction

- The objective of this work is to identify what features of a rocket launch contribute most diagnostically to the success of that launch.
- This work is undertaken in the context that company SpaceY seeks to better understand the decision-making and resource usage of SpaceX.
- Specifically SpaceY aims to better understand when SpaceX will reuse its first stage rockets and the overall price of the launch.
- Thus, the model will be developed to predict based on key variables whether SpaceX will reuse its first stage rocket.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
 - SpaceX Rest API, Wikipedia web scraping
- Perform data wrangling
 - Filtering, missing values handling, one hot encoding for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - Tried four ML models to determine best fit to the data

Data Collection

- API requests from SpaceX REST API and Web Scraping from SpaceX's Wikipedia

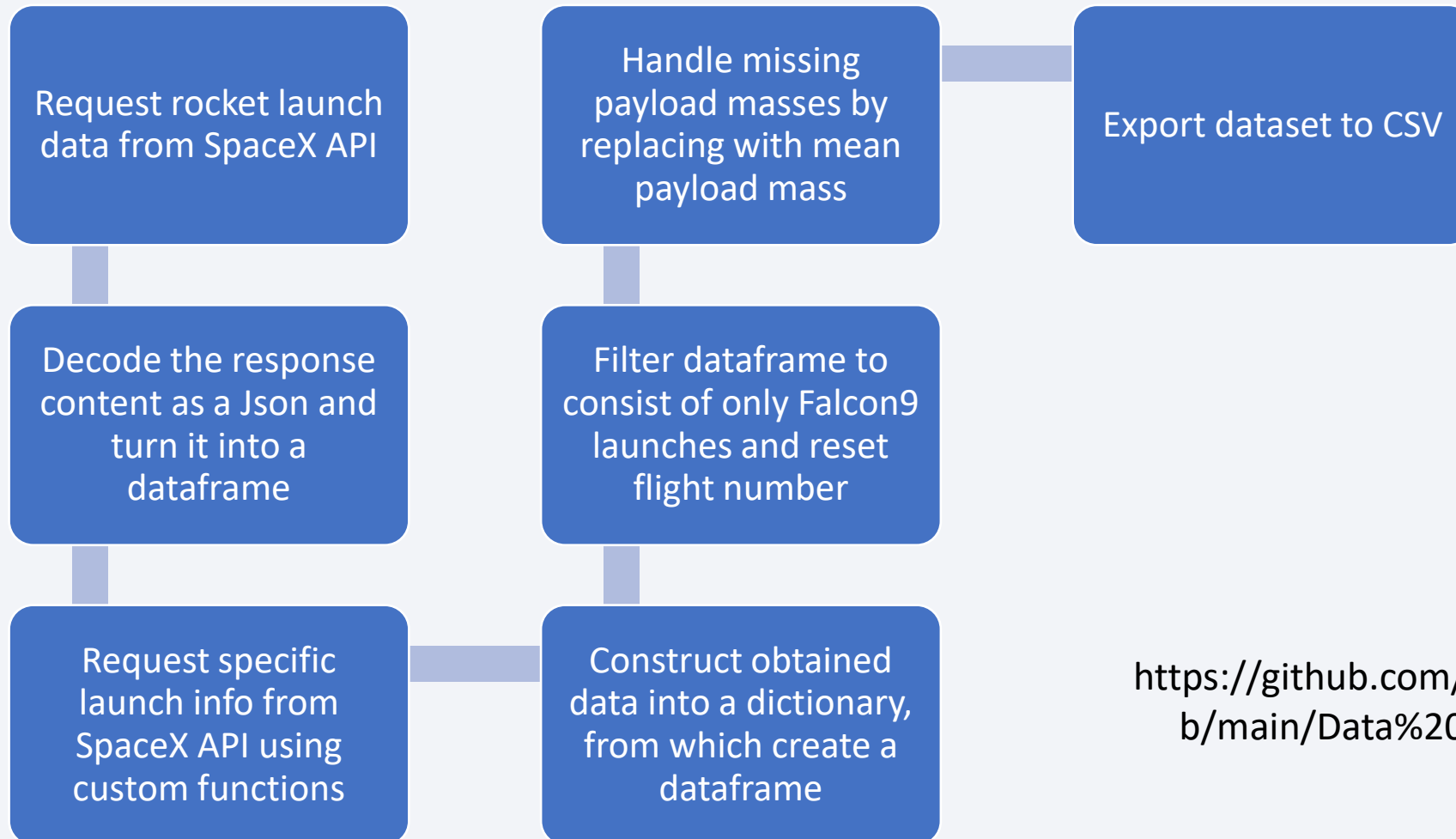
SpaceX REST API:

Flight No., Date, Booster version, Payload mass, Orbit, Launch site, Outcome, Flights, Grid fins, Reused, Legs, Landing pad, Block, Reused count, Serial, Longitude, Latitude

Wikipedia Web Scraping:

Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Booster version, Booster landing, Date, Time

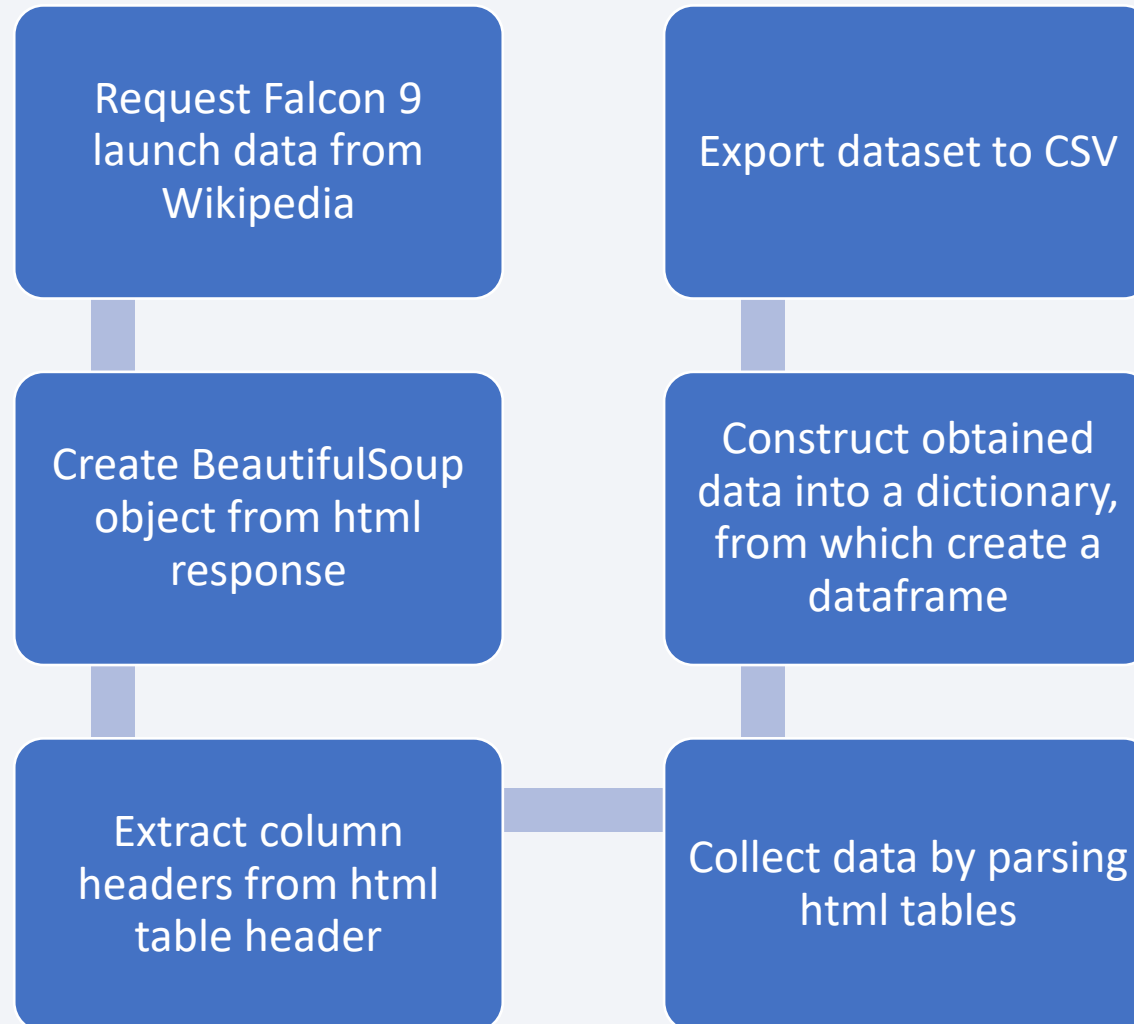
Data Collection – SpaceX API



Link:

<https://github.com/galenisabella/IBMfinal/blob/main/Data%20Collection%20API.ipynb>

Data Collection - Scraping



Link:

<https://github.com/galenisabella/IBMfinal/blob/main/Data%20Collection%20Web%20Scraping.ipynb>

Data Wrangling

Perform EDA and determine training labels

Identify percentage of missing values for each attribute of the dataset

Export dataset to CSV

Calculate number of launches on each site

Create landing outcome label from Outcome column

Calculate number and occurrence of each orbit

Calculate number and occurrence of mission outcome of the orbits

In our dataset, we identify several cases for which the booster did not land successfully. Specifically, we seek to identify where a launch was attempted but failed to successfully complete its landing. There are two primary ways in which a failed landing case is captured in the data, with a False or a None label. For instance, False ASDS means the mission was unsuccessfully landed to a drone ship while None ASDS represents a failure to land. Correspondingly, True ASDS represents a successful landing to a drone ship. This pattern is mirrored across landing types.

The success or failure of a launch is converted to training labels as either a '0' (failure) or '1' (success).

Link:


<https://github.com/galenisabella/IBMfinal/blob/main/Data%20Wrangling.ipynb>

EDA with Data Visualization

Charts plotted:

Scatterplots

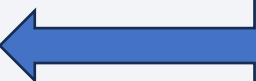
- Flight number vs payload mass, success of landing
- Flight number vs launch site, success of landing
- Payload mass vs launch site, success of landing
- Flight number vs orbit type, success of landing
- Payload mass vs orbit type, success of landing



Visualize relationship between variables and whether strength of correlation warrants inclusion in ML model

Bar plot

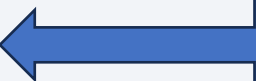
- Orbit type vs success rate



Compare discrete categories

Line plot

- Success rate yearly trend



Demonstrate trend over time

Link:

<https://github.com/galenisabella/IBMfinal/blob/main/EDA%20Data%20Viz.ipynb>

EDA with SQL

Performed the following queries:

- Displayed names of unique launch sites in the space mission
- Displayed 5 records where launch sites beg with string 'CCA'
- Displayed total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed date when first successful landing outcome in ground pad was achieved
- Listed names of boosters which have success in drone ship and have payload mass >4000 and <6000
- Listed total number of successful and failed mission outcomes
- Listed the names of booster versions which have carried maximum payload mass
- Listed record displaying month names, failed drone ship landings, booster version, launch site by month in 2015
- Ranked the count of landing outcomes between June 4, 2010, and March 20, 2017, in descending order

Link:

<https://github.com/galenisabella/IBMfinal/blob/main/EDA%20SQL.ipynb>

Build an Interactive Map with Folium

The folium map included the following map objects:

- Markers for all launch sites, with circle to highlight area, popup label, and associated text information for latitude, longitude, and site name
 - Markers were colored based on the success or failure of the launch
- Markers were clustered to identify overall success rate for launches at a particular site
- Colored lines were added to display distances between launch site KSC LC-39A and nearby infrastructures such as railways, highways, and cities as well as the nearest coastline.

Link:

<https://github.com/galenisabella/IBMfinal/blob/main/Interactive%20Viz%20Analytics%20with%20Folium.ipynb>

Build a Dashboard with Plotly Dash

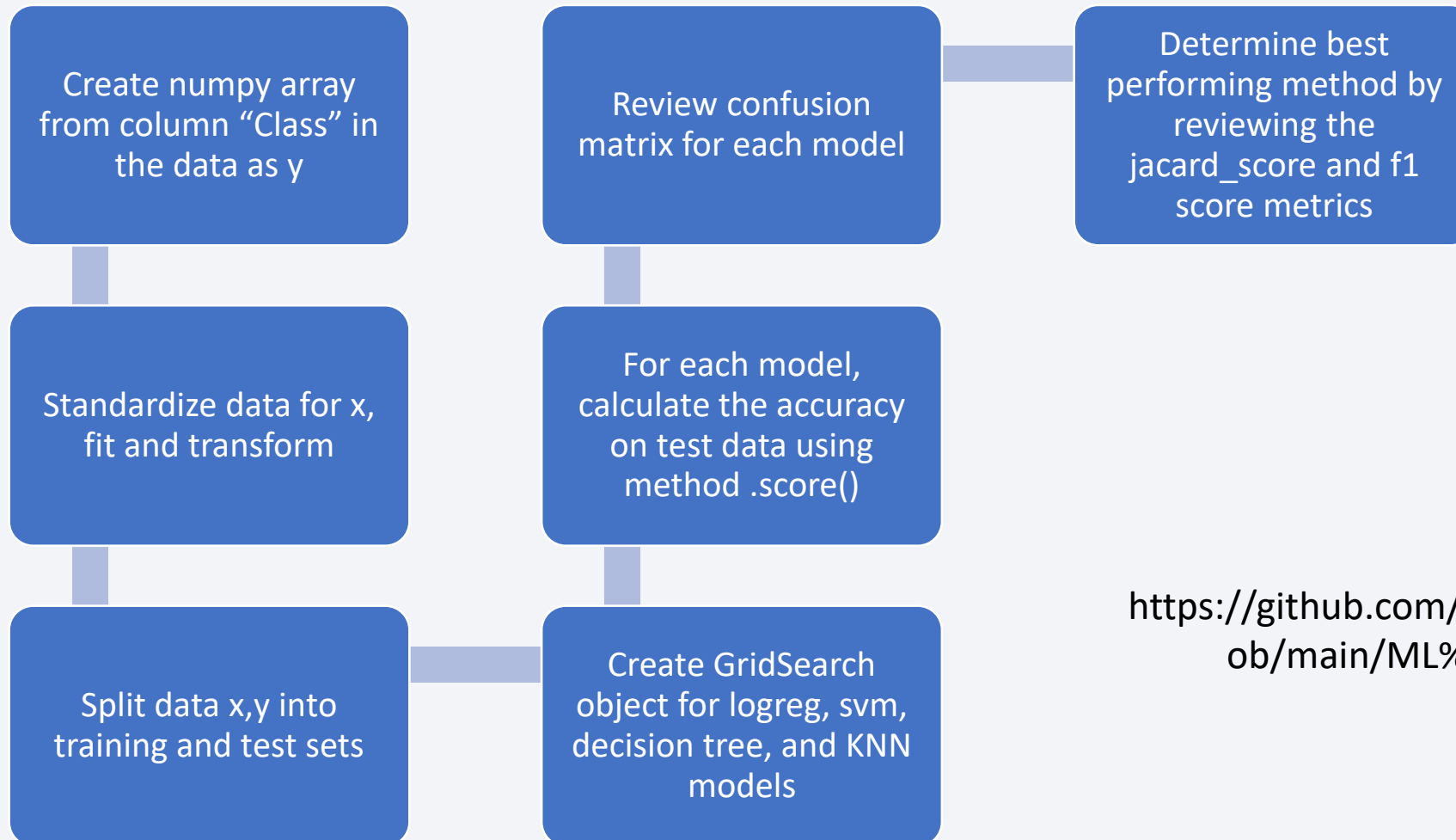
The dashboard includes the following:

- Launch sites list – dropdown menu for selection
- Success of launches across all sites, success rate at specific sites – pie chart
- Payload mass range – slider to select range for payload mass
- Payload mass vs. success rate for different boosters – scatter plot demonstrating correlation

Link:

<https://github.com/galenisabella/IBMfinal/blob/main/Dash%20App.py>

Predictive Analysis (Classification)



Link:
<https://github.com/galenisabella/IBMfinal/blob/main/ML%20Prediction.ipynb>

Results

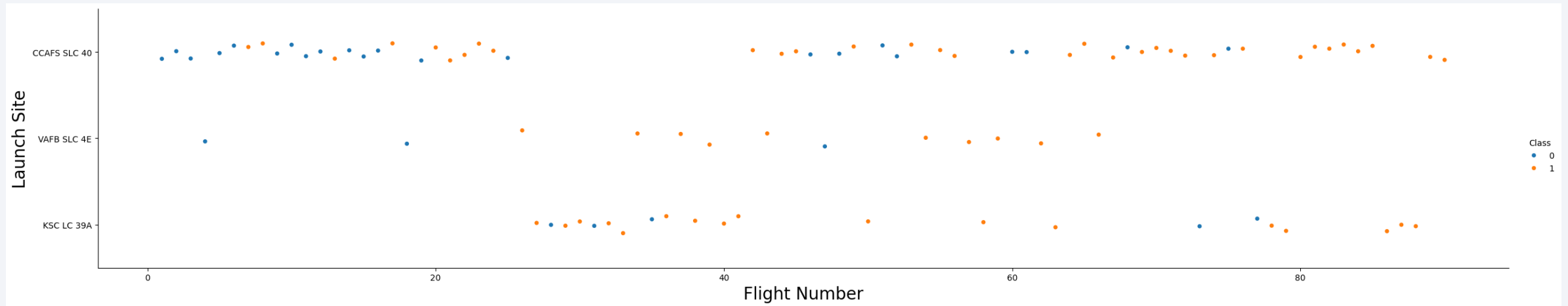
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

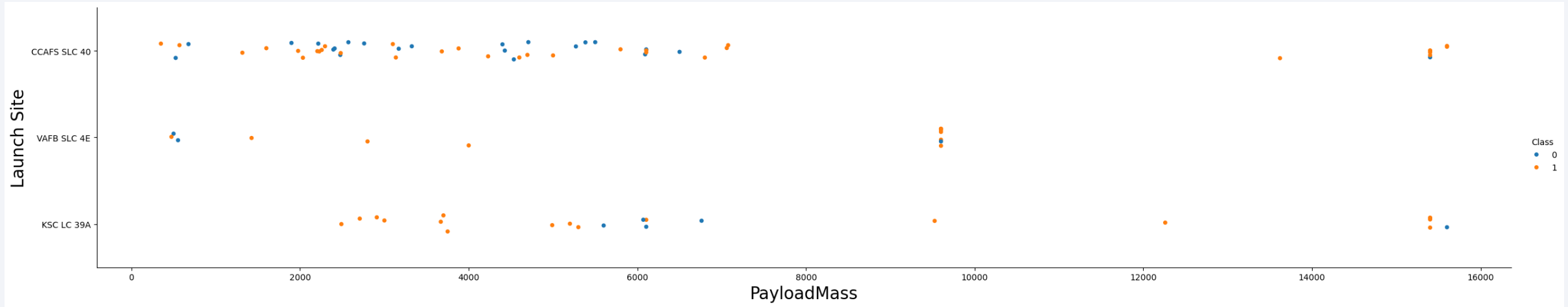
Insights drawn from EDA

Flight Number vs. Launch Site



- Earlier flights more often failed than succeeded, a trend which reversed in later flights.
- The CCAFS SLC 40 launch site accounts for the greatest number of launches across the different sites.
- Both VAFB SLC 4E and KSC LC 39A have higher success rates than CCAFS SLC 40. Both of these sites also have a higher percentage of their overall flights launching later than CCAFS SLC 40.
- It can be assumed from viewing these trends that launches became more likely to succeed with more prior launch failures and successes to reference across landing sites.

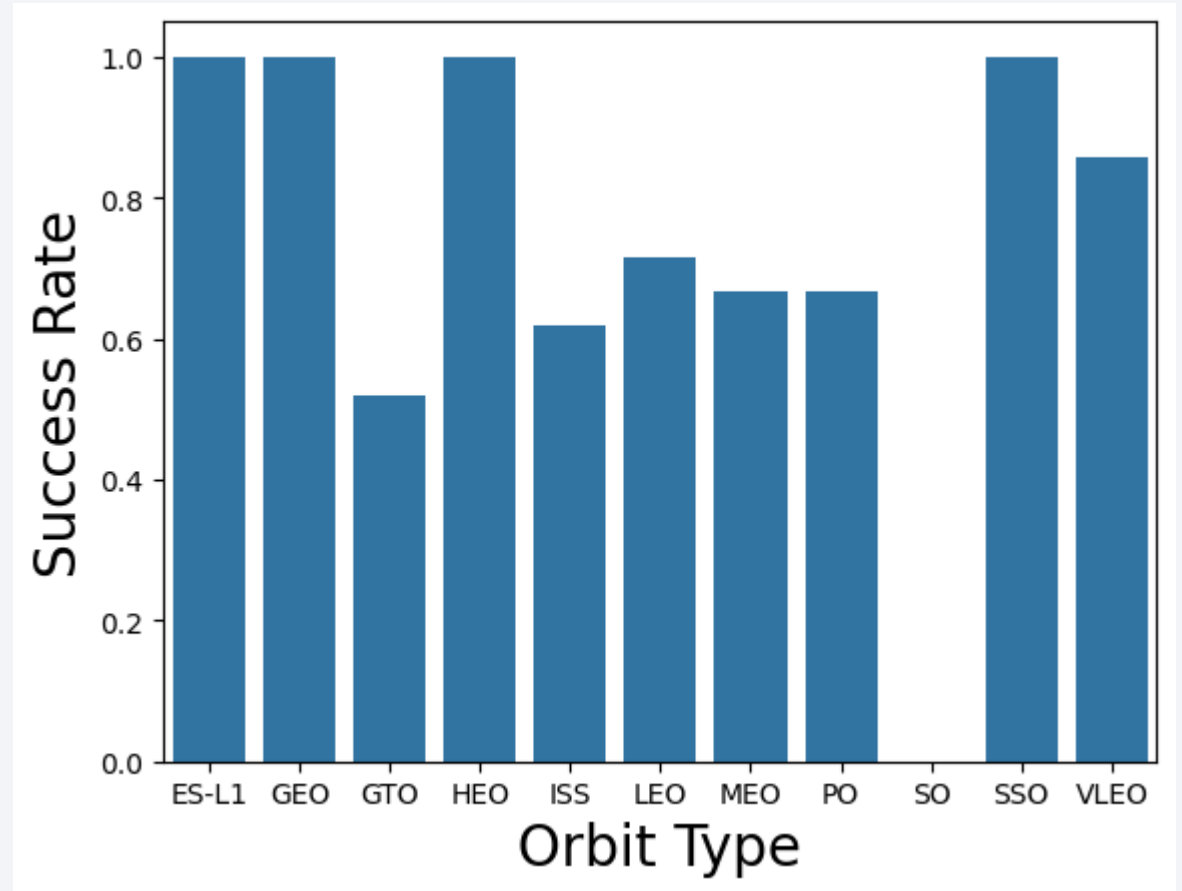
Payload vs. Launch Site

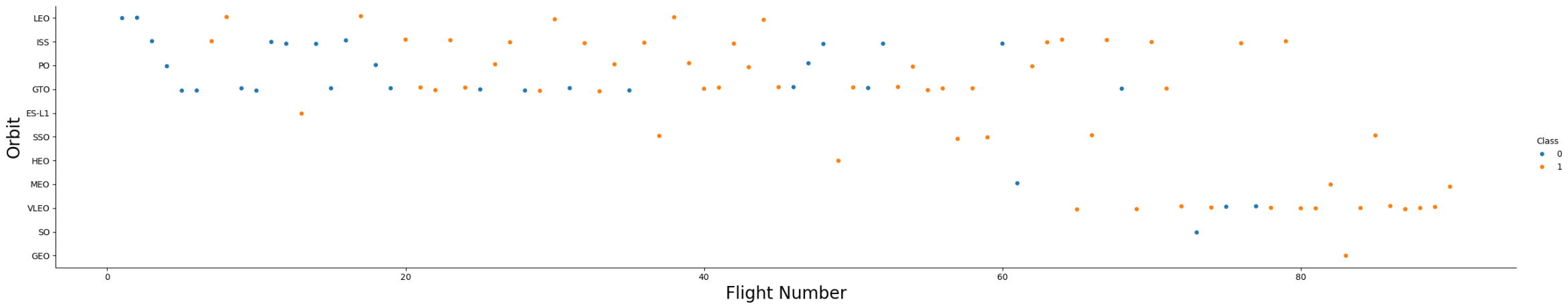


- Far more payloads of <6000kg are launched than over that value
- Having a payload mass of >6000kg is correlated with a greater landing success rate agnostic of launch site
- For the CCAFS SLC 40 and VAFB-SLC sites, payloads of masses under 6000kg brought a mixed rate of landing success, while for KSC LC 39A payloads in this range nearly always landed successfully

Success Rate vs. Orbit Type

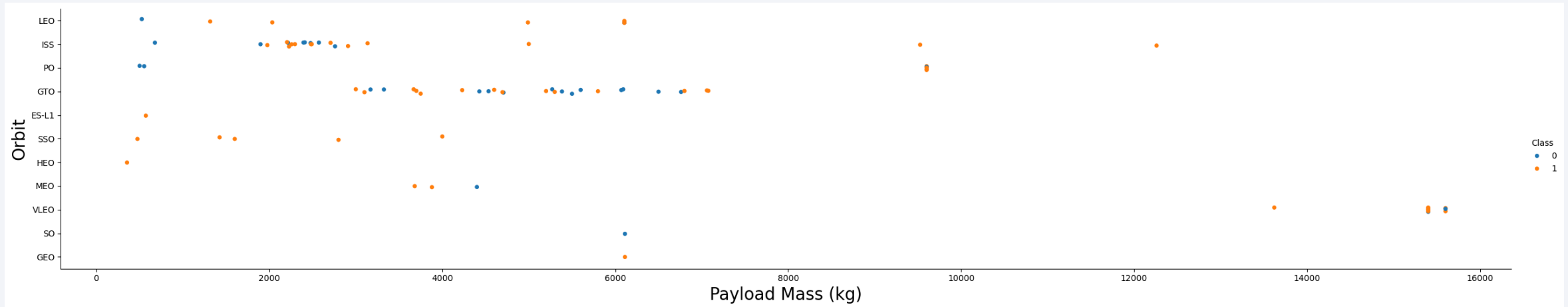
- Success rates of each orbit:
 - 100%: ES-L1, GEO, HEO, SSO
 - 0%: SO
 - 50-85%: GTO, ISS, LEO, MEO, PO, VLEO





- For LEO, landing success appears correlated with the number of flights. Conversely, for the other orbits this correlation does not appear to hold.

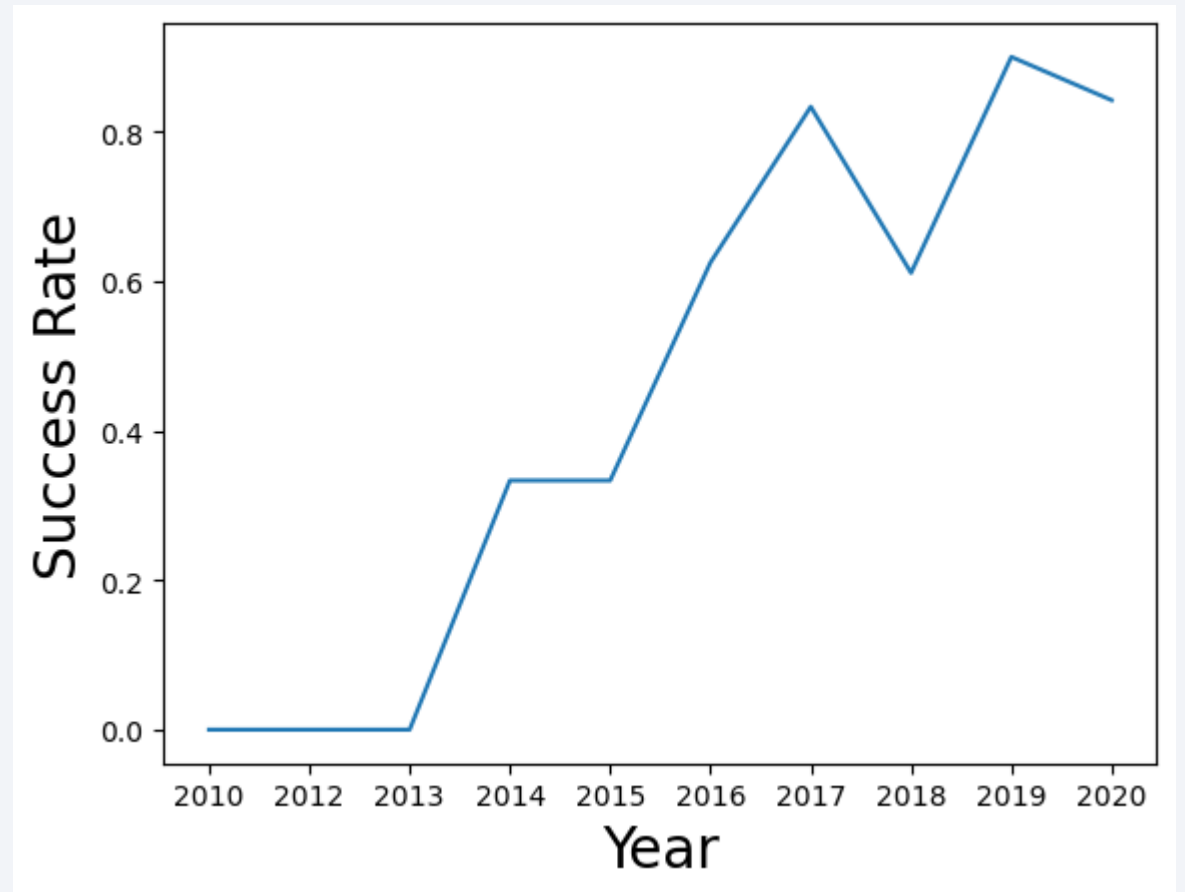
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are greater for Polar, LEO and ISS. However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- An increasing success rate is observed 2013-2020.



All Launch Site Names

- There are four unique launch sites in the dataset.

```
%sql select distinct launch_site from SPACEXTBL;
```

```
* sqlite:///my\_data1.db
```

```
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

Python

```
* sqlite:///my\_data1.db
```

Done.

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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 are a selection of launches from Cape Canaveral

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS_KG_) as total_payload_mass from SPACEXTBL where Customer = "NASA (CRS)";
```

```
* sqlite:///my\_data1.db
```

```
Done.
```

total_payload_mass

45596

- Payload masses from customer NASA (CRS) were summed to return the total payload mass

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS_KG_) as average_payload_mass from SPACEXTBL where booster_version like '%F9 v1.1%';

* sqlite:///my_data1.db
Done.

average_payload_mass
2534.6666666666665
```

- The data was filtered by booster version to identify launches using the F9 v1.1 booster, from which then the average payload mass was calculated.

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from SPACEXTBL where landing_outcome = 'Success (ground pad)';

* sqlite:///my\_data1.db
Done.

first_successful_landing
2015-12-22
```

- The data was filtered by landing outcome on Success (ground pad), then observing the first successful landing date under this outcome condition

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql select booster_version from SPACEXTBL where landing_outcome = 'Success (drone ship)' and PAYLOAD_MASS_KG_
between 4000 and 6000;
```

```
* sqlite:///my\_data1.db
```

```
Done.
```

Booster_Version

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

- The data was filtered by landing outcome on Success (drone ship) and payload mass $4000 < \text{payload mass} < 6000$.

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXTBL group by mission_outcome;
```

```
* sqlite:///my\_data1.db
```

```
Done.
```

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- From the table, we observe 1 failure in flight, 99 successes, and 1 success with an unclear payload status.

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXTBL where PAYLOAD_MASS_KG_ = (select(max(PAYLOAD_MASS_KG_)) from SPACEXTBL);
```

```
* sqlite:///my\_data1.db
```

```
Done.
```

Booster_Version

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

The booster versions were filtered for those which had carried the maximum payload mass, resulting in 12 F9 B5s.

2015 Launch Records

```
%%sql select substr(Date,6,2) as month, Date, booster_version, launch_site, landing_outcome from SPACEXTBL  
| where landing_outcome = 'Failure (drone ship)' and substr(Date,1,4)='2015';
```

* [sqlite:///my_data1.db](#)

Done.

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- There were two failed landing outcomes for drone ship landings in 2015. Both were launched from CCAFS LC-40 and their booster versions are observed in the table above.

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

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTBL
where Date between '2010-06-04' and '2017-03-20'
group by landing_outcome
order by count_outcomes desc;
```

* [sqlite:///my_data1.db](#)

Done.

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

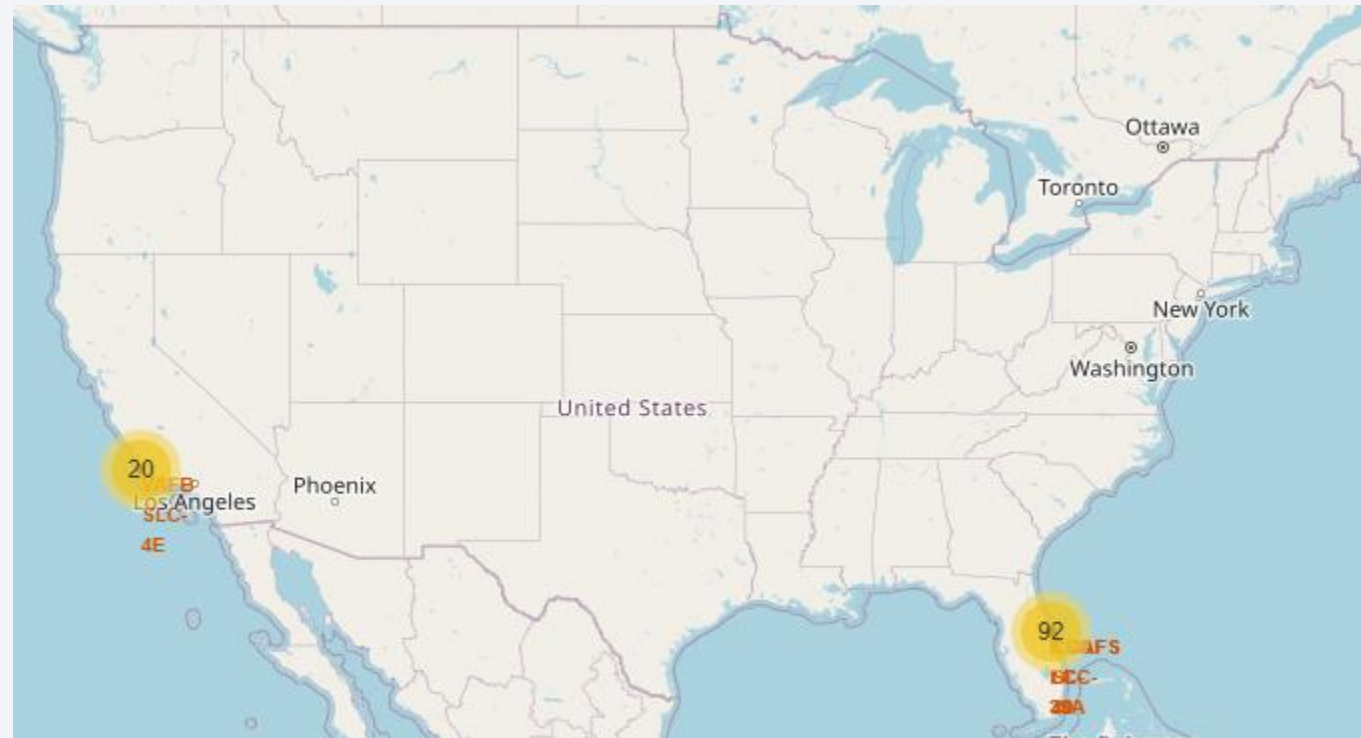
- The landing outcomes that took place between 2010-06-04 and 2017-03-20 are listed to the left by frequency

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

Launch Sites Proximities Analysis

Launch sites



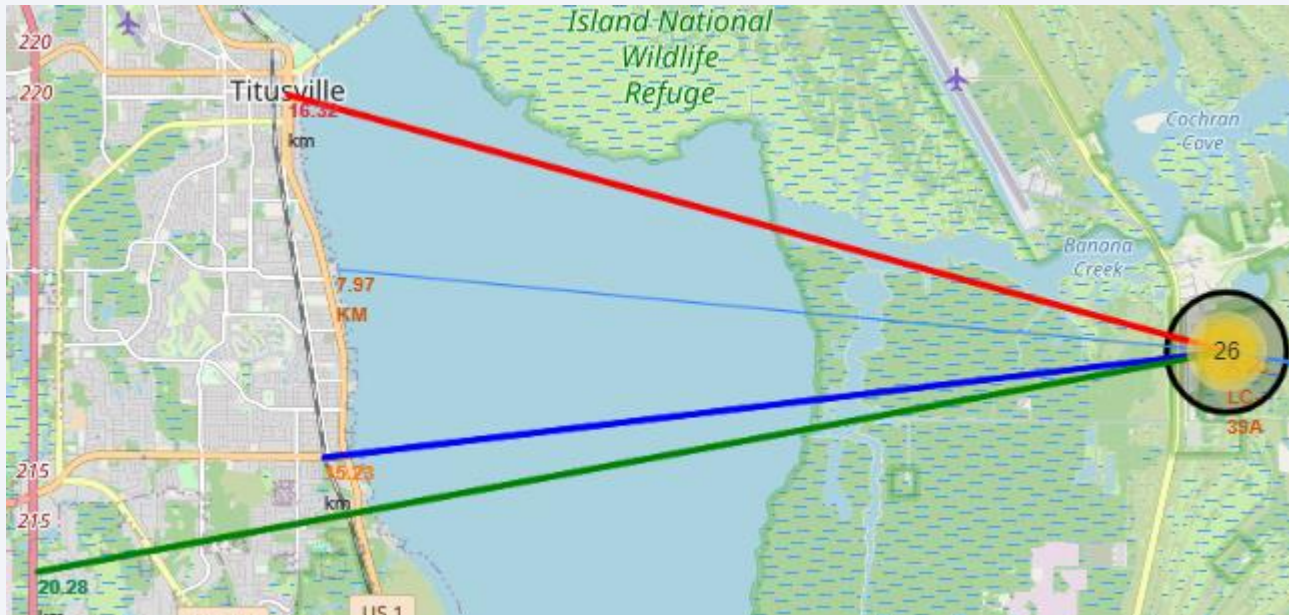
- The above map visualizes the location of launch sites catalogued in the data. These are uniformly on the coast, with a majority being in Florida.

Launch outcomes by site



- The success or failure of a given launch is indicated by an associated green or red labeled marker. As can be seen in this visual, launches are grouped by launch site. Here, we see the launch outcomes for site CCAFS SLC-40

Distances from launch site KSC LC-39A



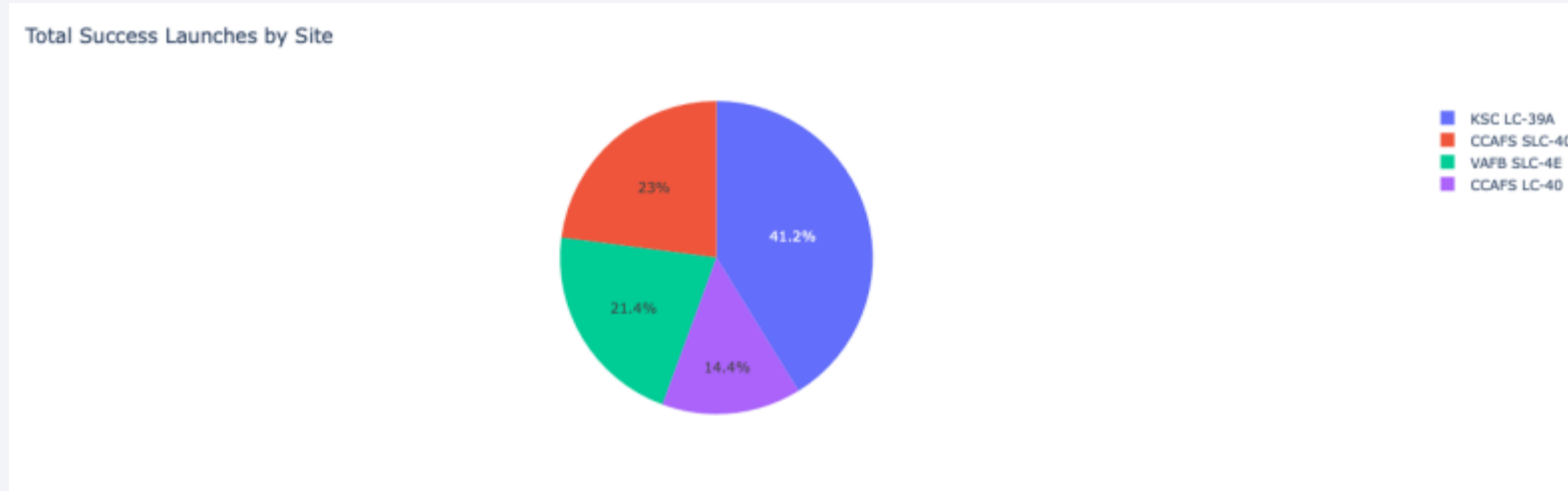
- Visualized are the distances to the nearest...
 - City: 16.32km
 - Coastline: 7.97km
 - Railway: 15.23km
 - Highway: 20.28km



Section 4

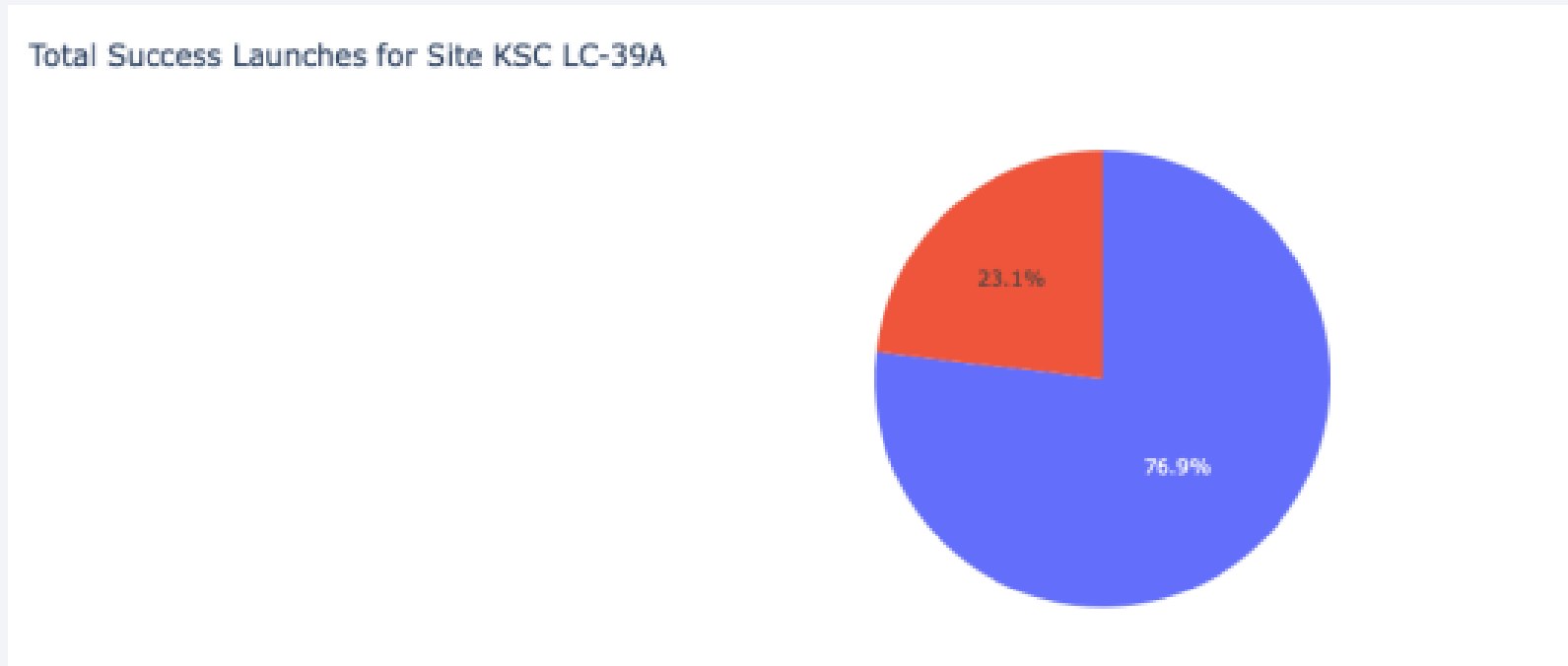
Build a Dashboard with Plotly Dash

Launch success count across all sites



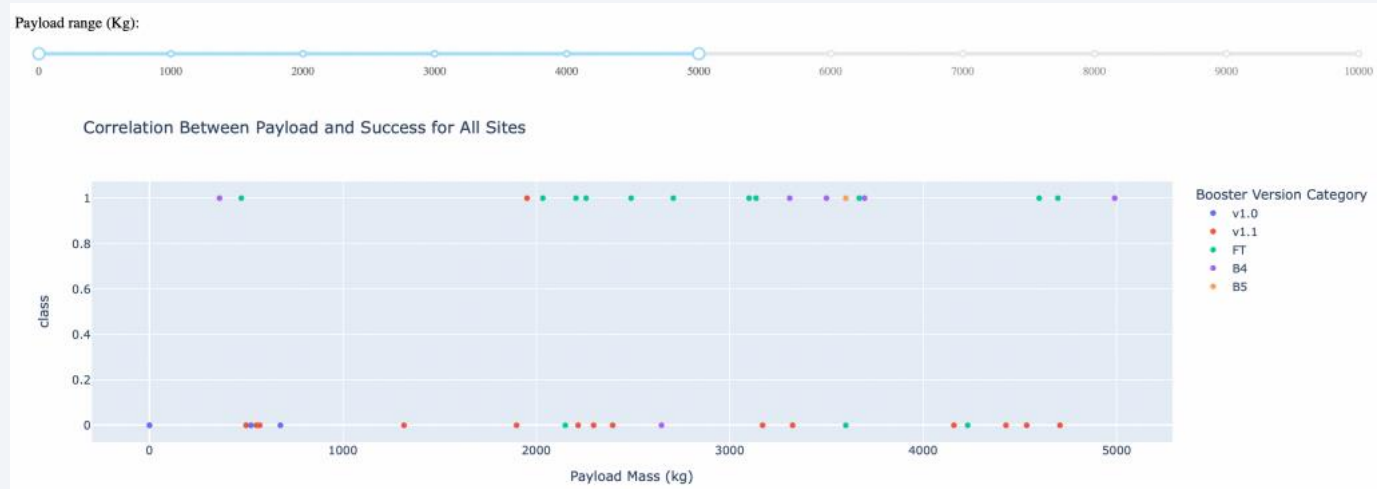
- KSC LC-39A demonstrates the greatest number of launch successes out of the four launch sites.

Launch success rate for KSC LC-39A



- KSC LC-39A launches had a 76.9% success rate, with a total of 10 successful landings and 3 failures.

Payload mass vs. launch outcome by booster version across all sites



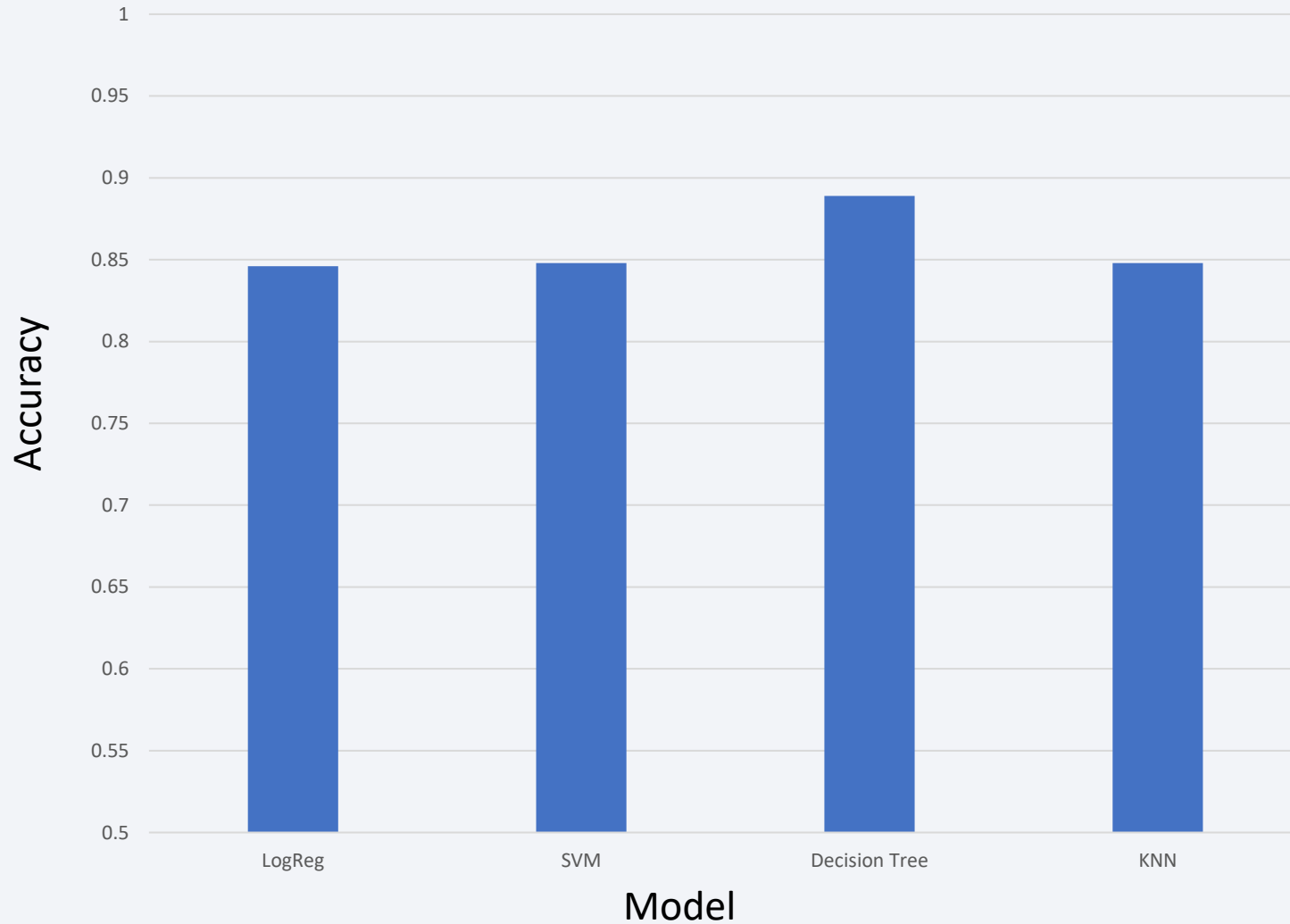
- Payloads between 2000 and 5500kg demonstrated the highest rate of success.



Section 5

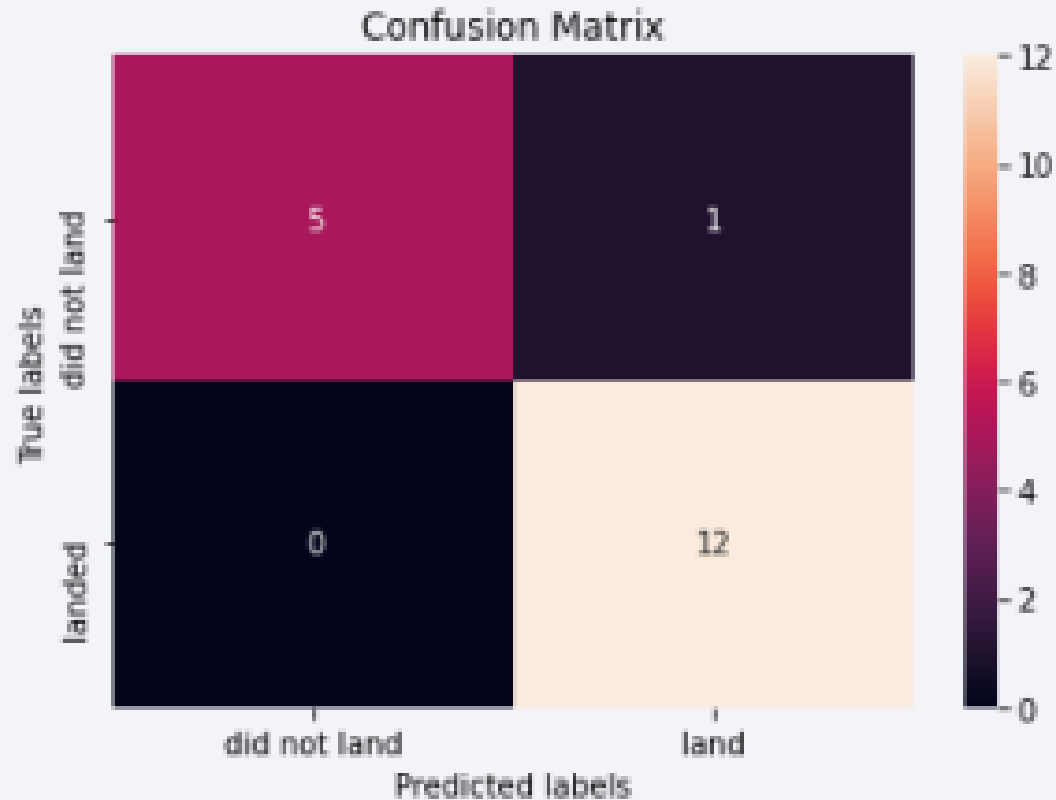
Predictive Analysis (Classification)

Classification Accuracy



- By a narrow margin, the decision tree model performed the best among the models tested

Confusion Matrix



- The confusion matrix for the decision tree, as the best performing model, demonstrates effective classification of true outcomes, but returns one false positive.

Conclusions

- Launch sites tend to be in proximity to the equator line and the coast
- The decision tree model performed the best in accurately predicting launch outcome, though this finding is caveated in several ways – most notably the small test data set which contained only 18 observations.
- From the dataset, we can additionally observe certain aspects of successful launches:
 - KSC LC-39A has the greatest success with launches among launch sites
 - Orbits ES-L1, GEO, HEO and SSO have 100% success rate for launches
 - Success of launches trends positively over time
 - Payload mass positively correlates with launch success for certain orbits, but not all orbits

Thank you!

