Winning Space Race with Data Science

Gregory Fountain
16 Oct 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Visualization
  - Interactive visual analytics
  - Predictive Analysis
- Summary of all results
  - EDA results
  - Dashboard Analytics Results
  - Predictive Analysis results

Introduction

- Project background and context
  - SpaceX has been considered one of the most successful companies in the commercial space age, making space travel more affordable and accessible. Much of the savings that Space X sees is because of their ability to reuse the first stage rocket motors. What can other companies, like Space Y, do to compete with Space X and reduce their operational costs?
- Questions to Answer
  - How do variables such as payload mass, launch site, types of orbit and number of flights affect the success of a first stage landing?
  - Does the rate of successful landings increase over the years?
  - What algorithms provide the greatest accuracy for predicting the success or failure of future rockets?

Section 1

Methodology

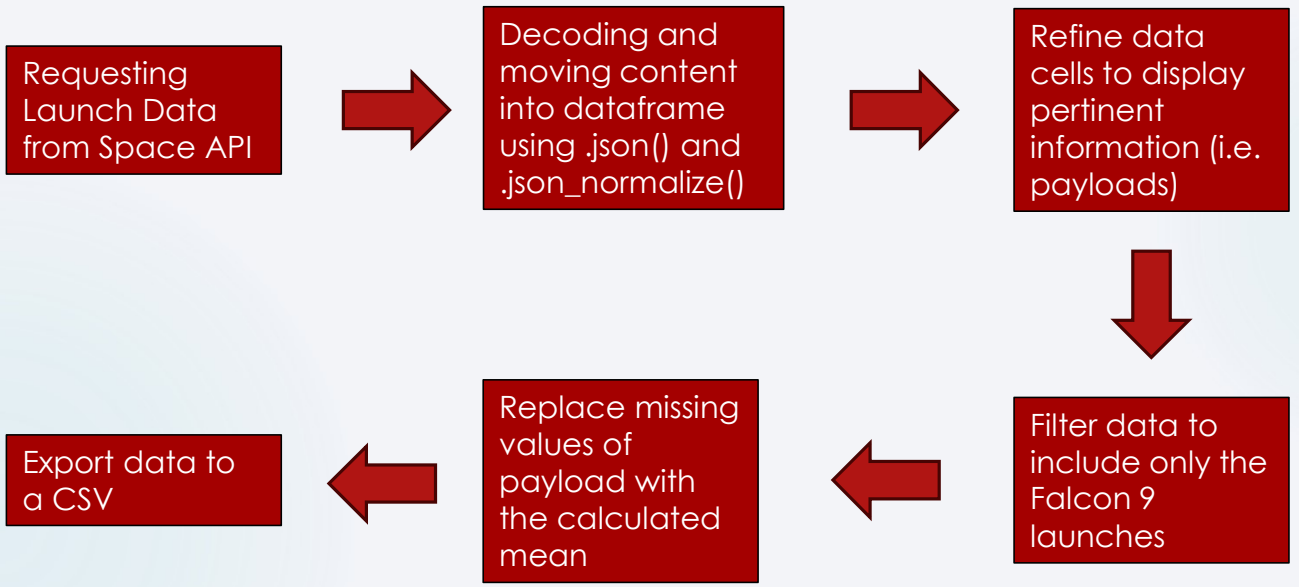
Methodology

- Data collection methodology:
  - Using SpaceX API
  - Using Webscraping from Wikipedia
- Perform data wrangling
  - Filtering Data
  - Dealing with missing values
  - Preparing data for binary classification
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models

Data Collection

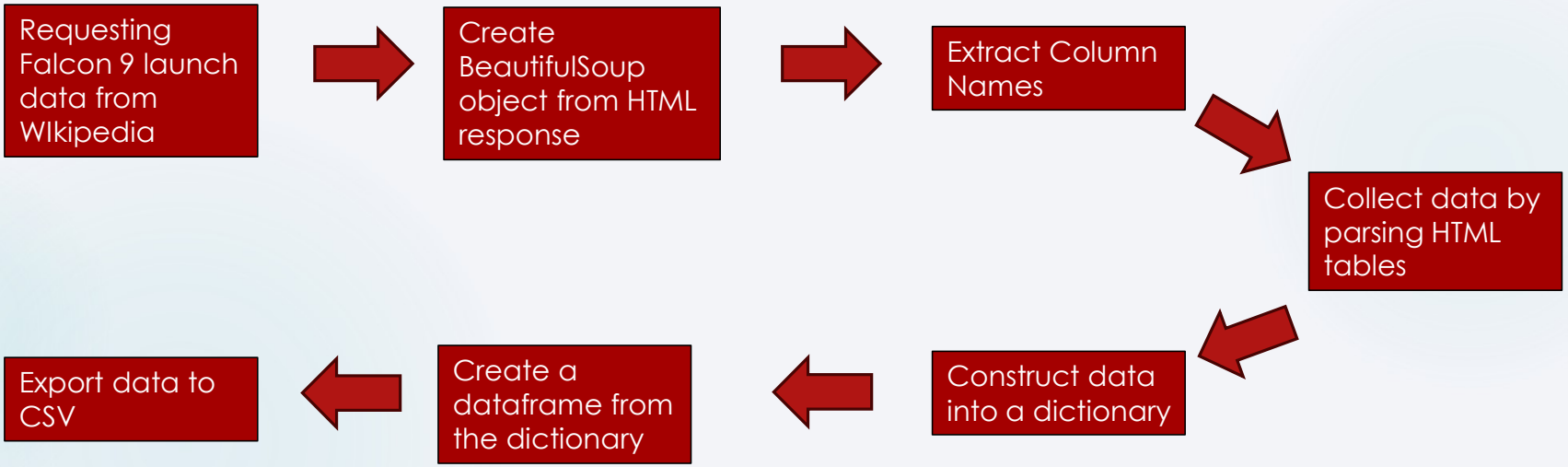
- Data was collected through a combination of:
  - SpaceX API
  - Web Scraping from Wikipedia page on SpaceX's Falcon 9 launches (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Both methods were used to get the most complete information about launches and recoveries for more detailed analysis and included data in the following categories:
  - Flight #, Booster Versions, Payload, Orbit Types, Launch Sites, Recovery Outcomes, and Landing Pad Types.

Data Collection – SpaceX API



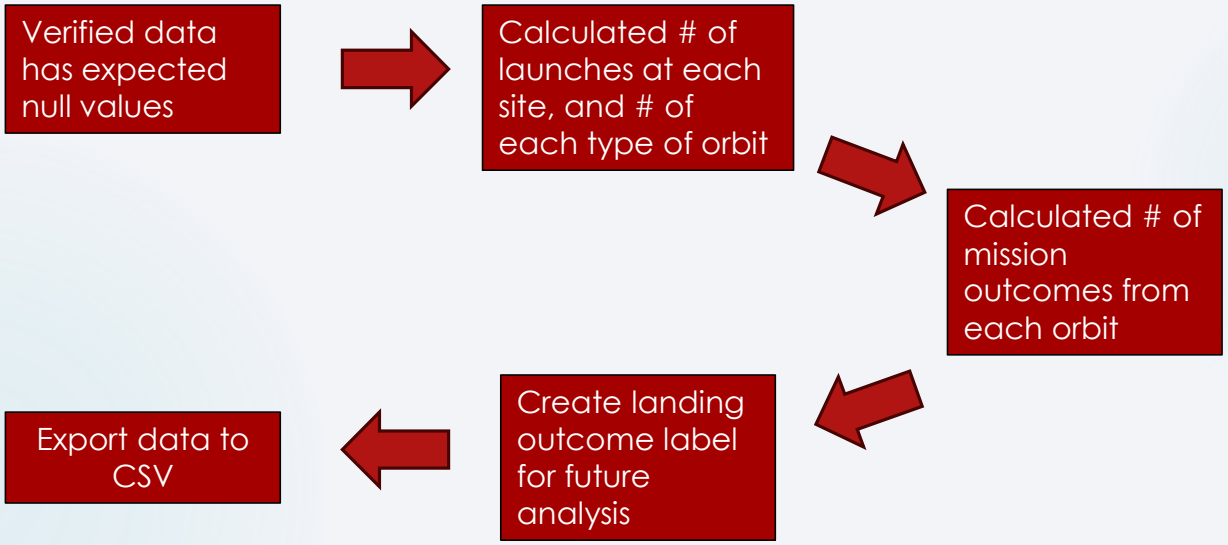
- GitHub Data Collection API

Data Collection – Webscraping



- GitHub Data Collection-Webscraping

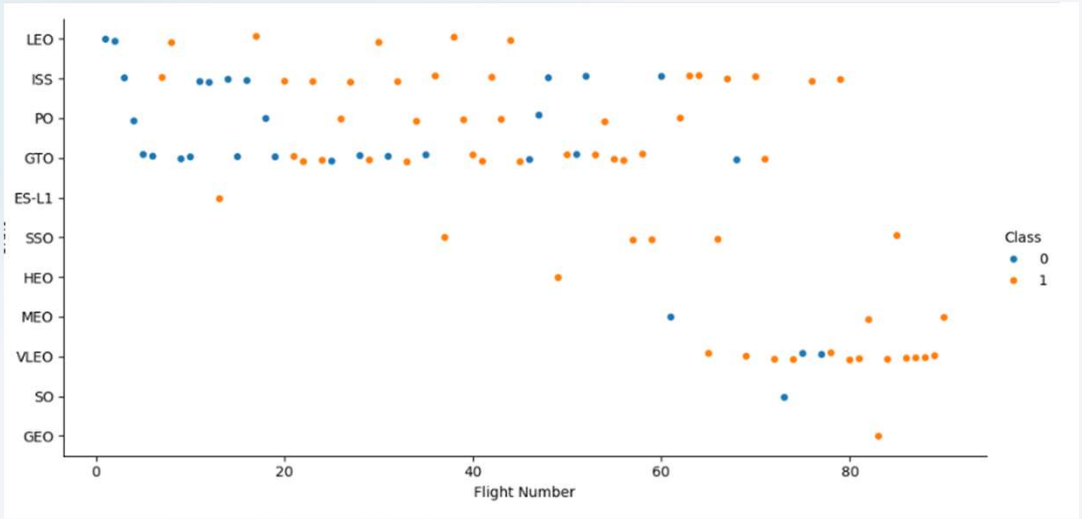
Data Wrangling

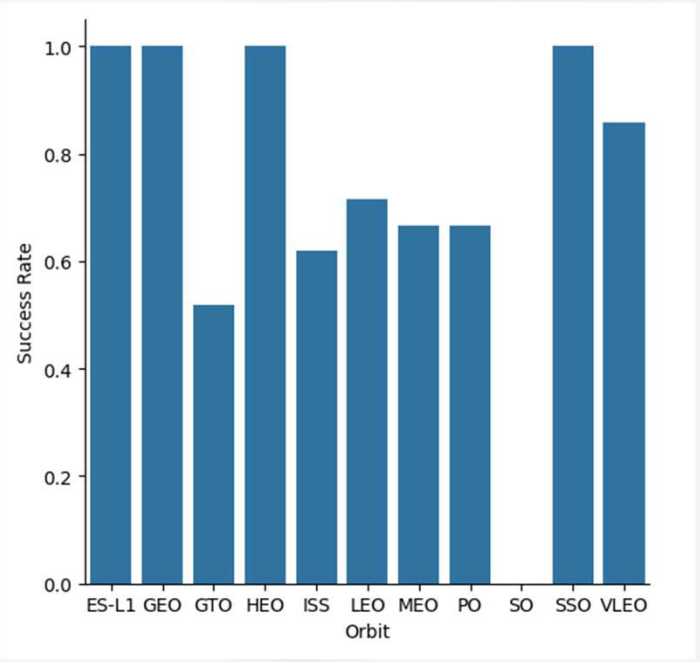


- GitHub Data Wrangling

EDA with Data Visualization

- Explored data via scatterplots (Payload Mass vs Flight Number, Launch site vs flight number, launch site vs payload mass, and bar charts (success rate and orbit)
- Github – EDA with Data Visualization





EDA with SQL

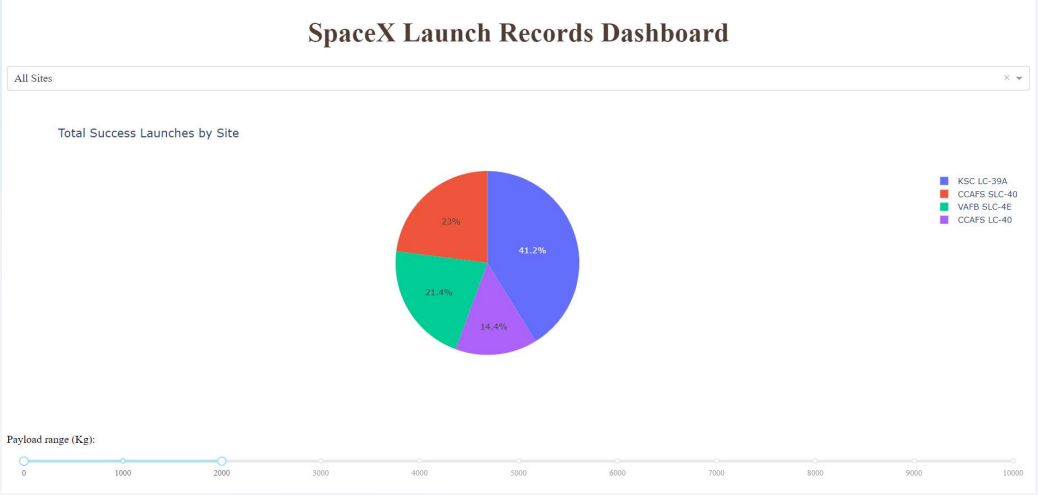
- The following SQL queries were performed:
  - Names of unique launch sites
  - Top 5 Launch sites whose name begin with 'CCA'
  - Total payload mass carried by boosters on NASA (CRS) missions
  - Date the first successful landing to a ground pad was accomplished
  - Total number of successful and failure mission outcomes
  - Names of booster versions which carried the max payload
  - Failed landing outcomes to drone ship, what the booster version was, and the launch site names for year 2015
  - Rank of landing outcomes based off overall count between 2010-06-04 and 2017-03-20
- Github – EDA with SQL

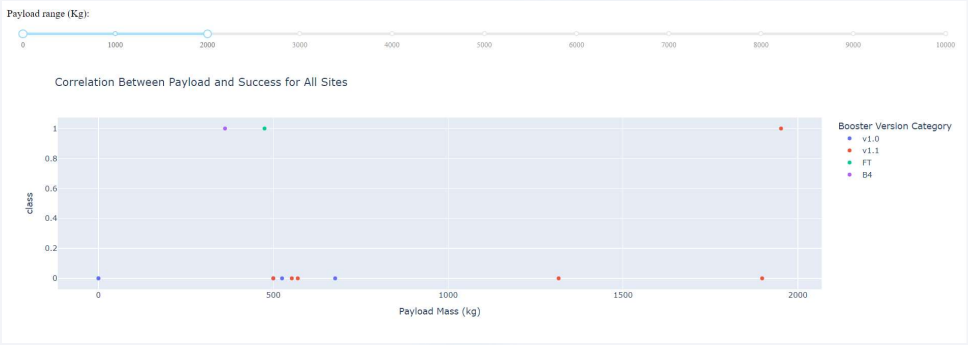
Build an Interactive Map with Folium

- Markers added to indicate launch sites
- Circles added to highlight area such as the NASA Johnson Space Center, and launch sites
- Marker clusters indicates groups of launches from within a site
- Lines were used to indicate distances from key features (railways, coastlines) from launch sites
- Github – Interactive Map with Folium

Build a Dashboard with Plotly Dash

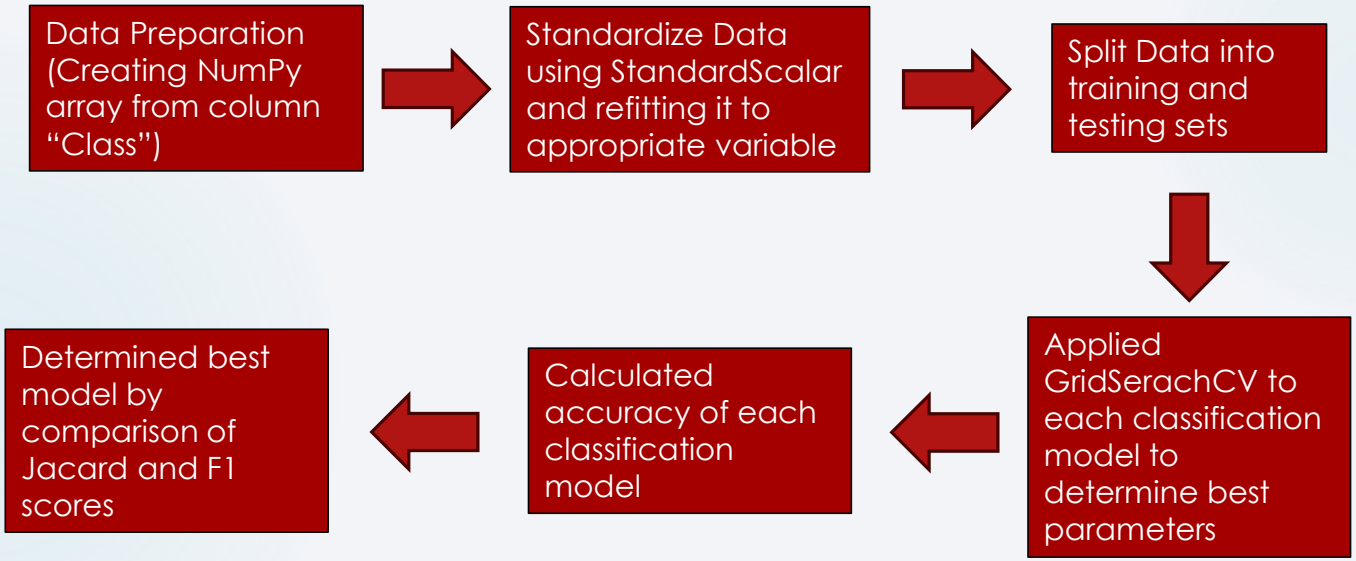
- Created dropdown to select desired sites for display
- Pie chart depicted on the dashboard for total successful launches by site
- Scatter plot depicts successes and failures by Booster Version and Payload Mass vis selection of payload on the slider bar
- [Github – Dashboard with Plotly]



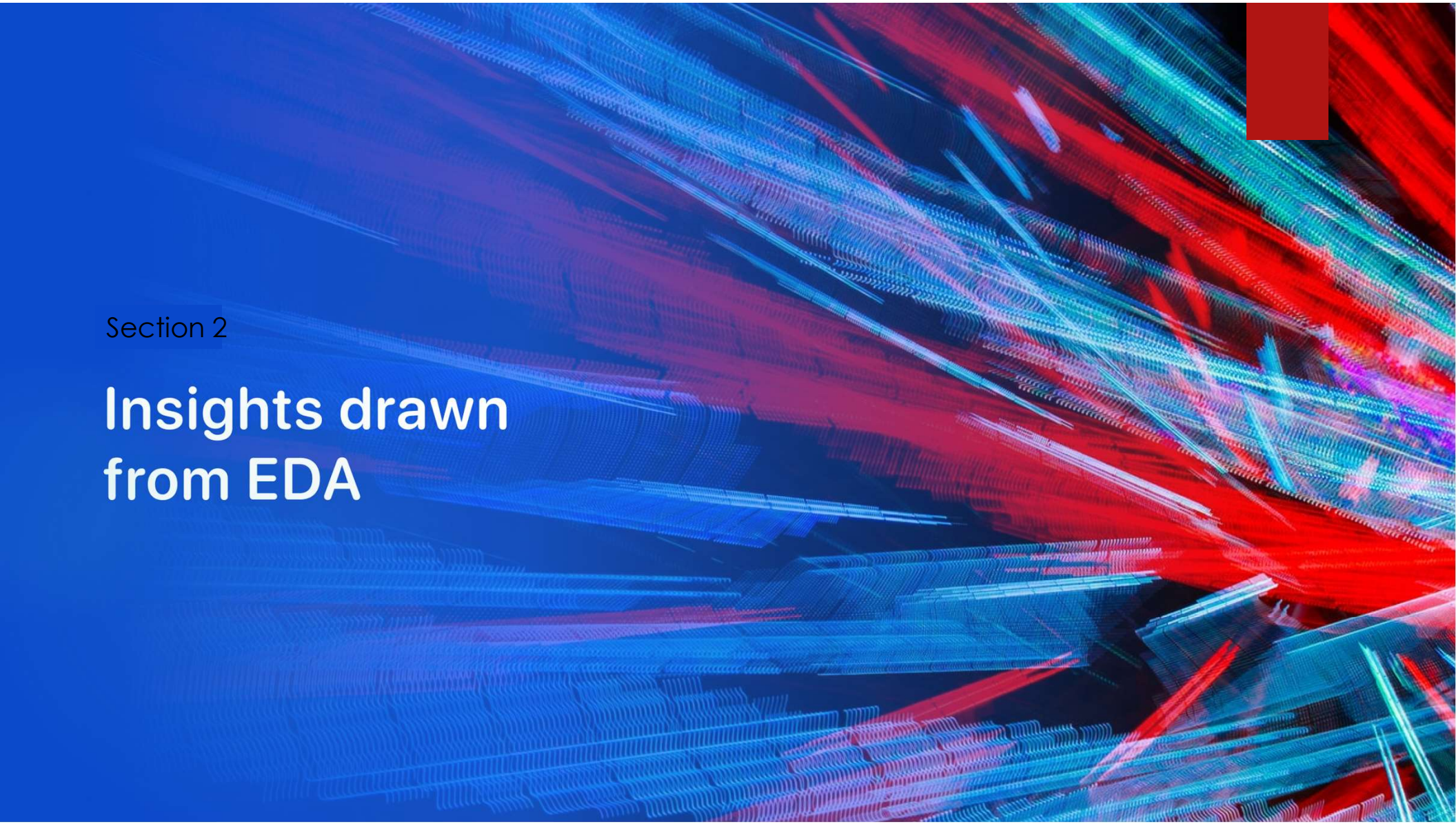


Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree, and K nearest neighbors.



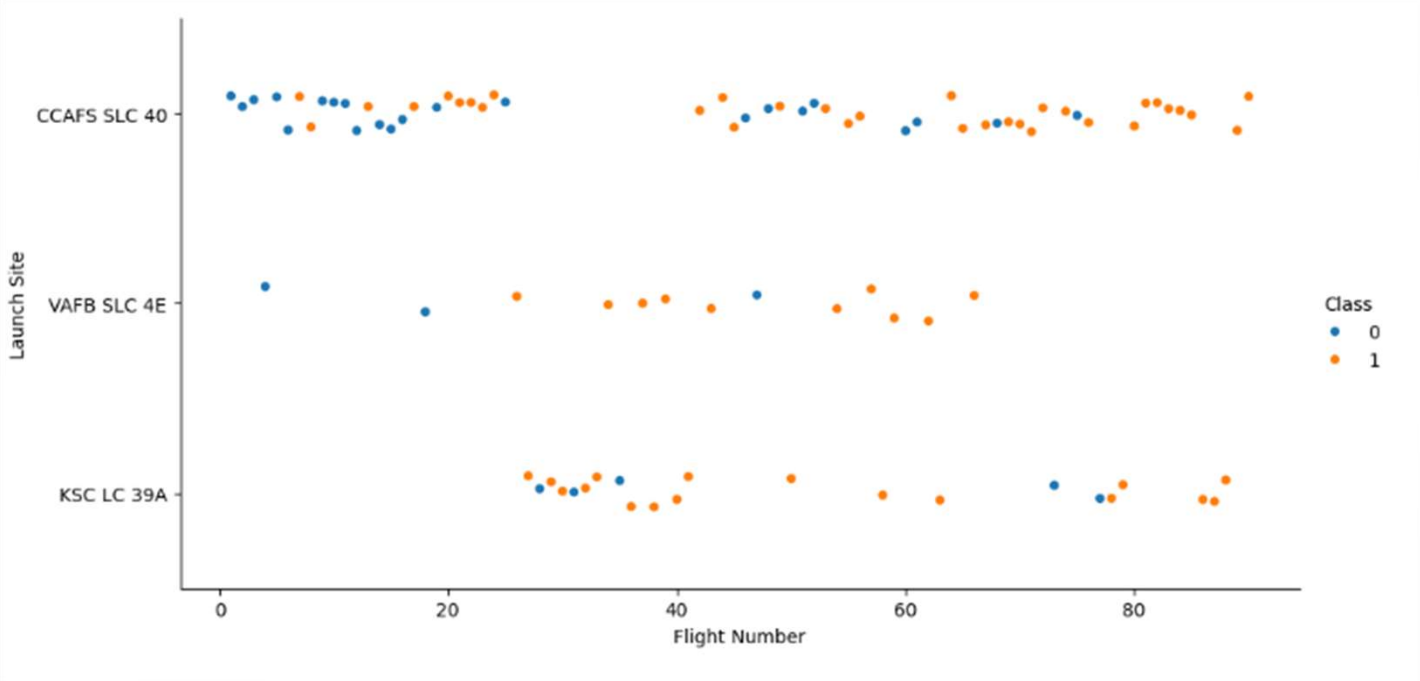
- Github – Machine Learning Prediction



Section 2

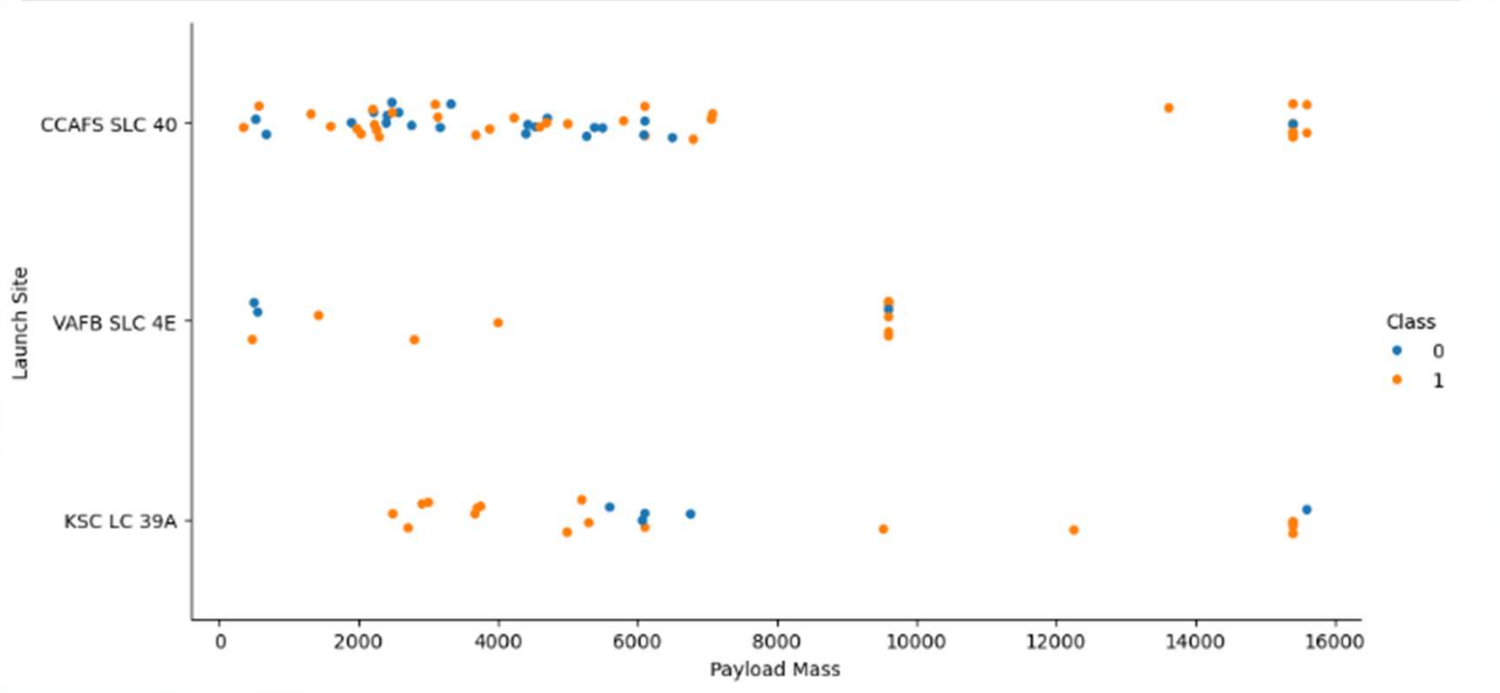
Insights drawn from EDA

Flight Number vs. Launch Site



- All flight recoveries after Flight #80 have been successes
- CCAFS SLC 40 has most total launches and has the highest # of recent successes
- CCAFS SLC 40 & KSC LC 39A are more popular for launches since Flt # 60

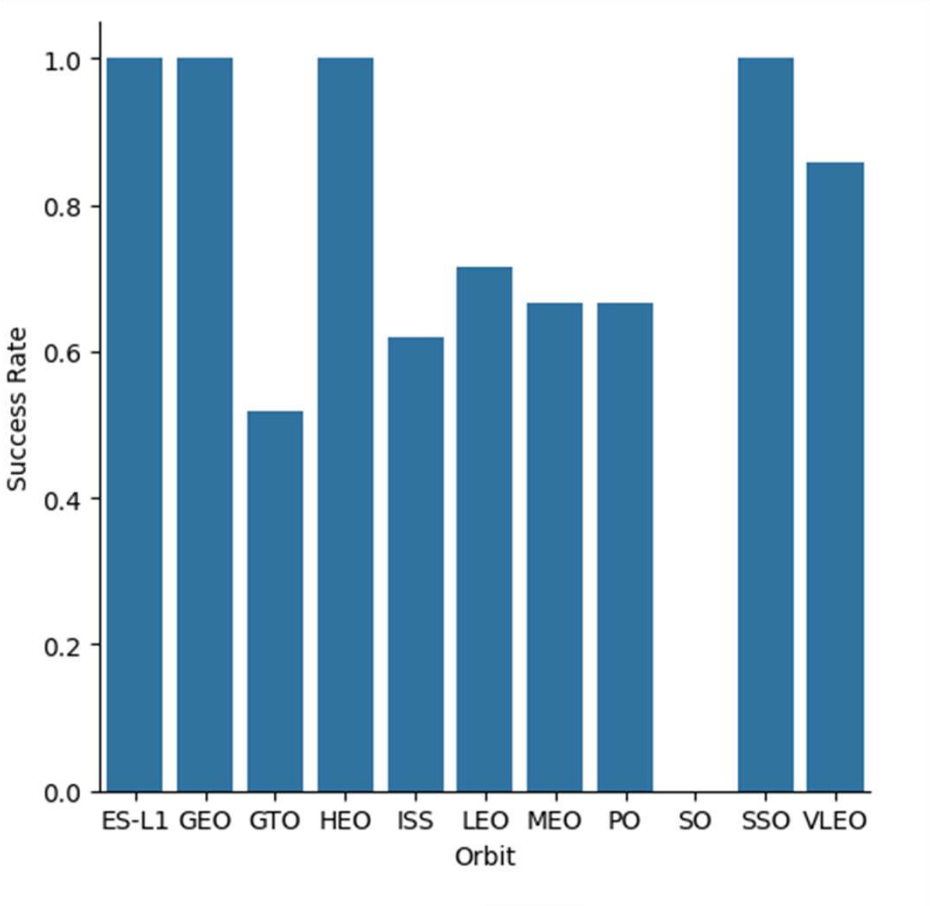
Payload vs. Launch Site



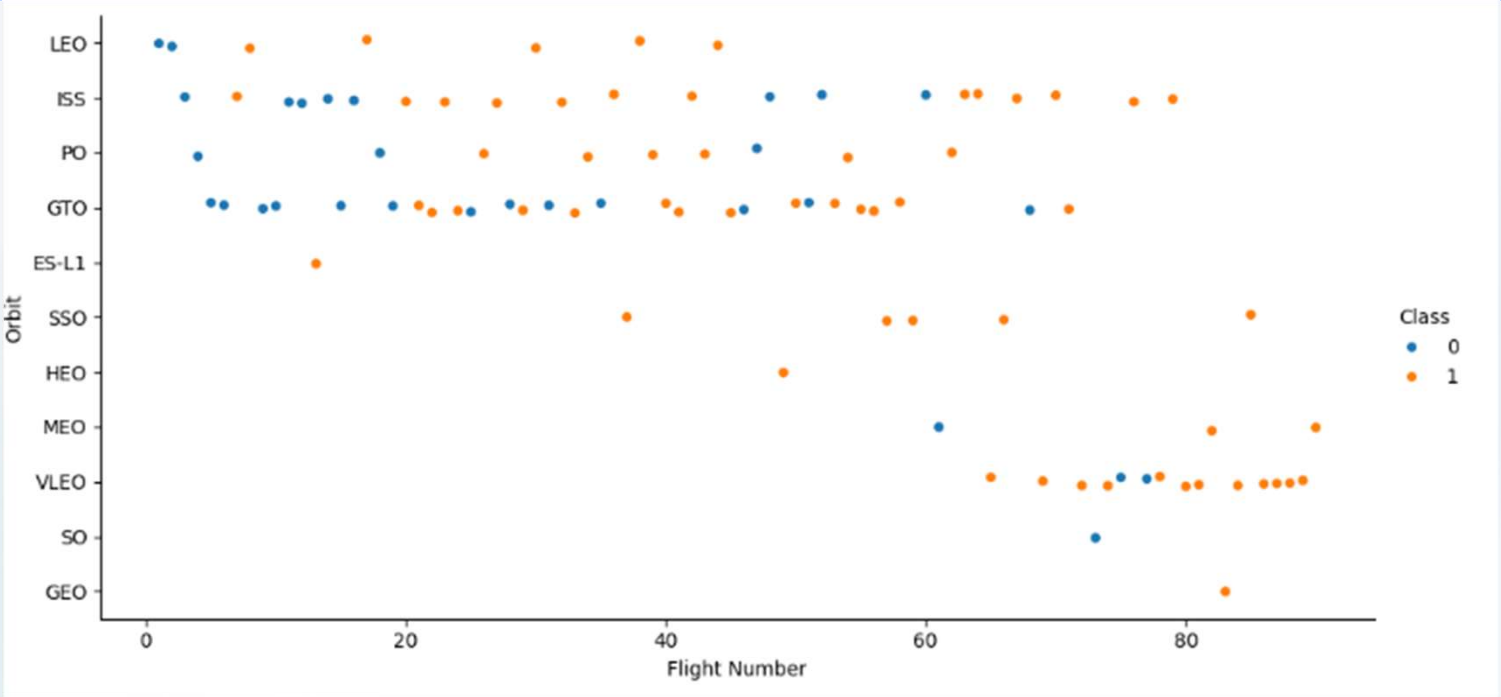
- As payload mass at each site increases relative success rate increases
- VAFB SLC 4E has good success rate of high payloads but hasn't launched a mission with payload > 10000
- KSC LC 39A HAS 100% success rate for payloads < 5500 kg

Success Rate vs. Orbit Type

- Orbits with 100% success:
  - ES-L1, GEO, HEO and SSO
- Orbit SO has 0% success
- All remaining orbits have between 50-90% success rate

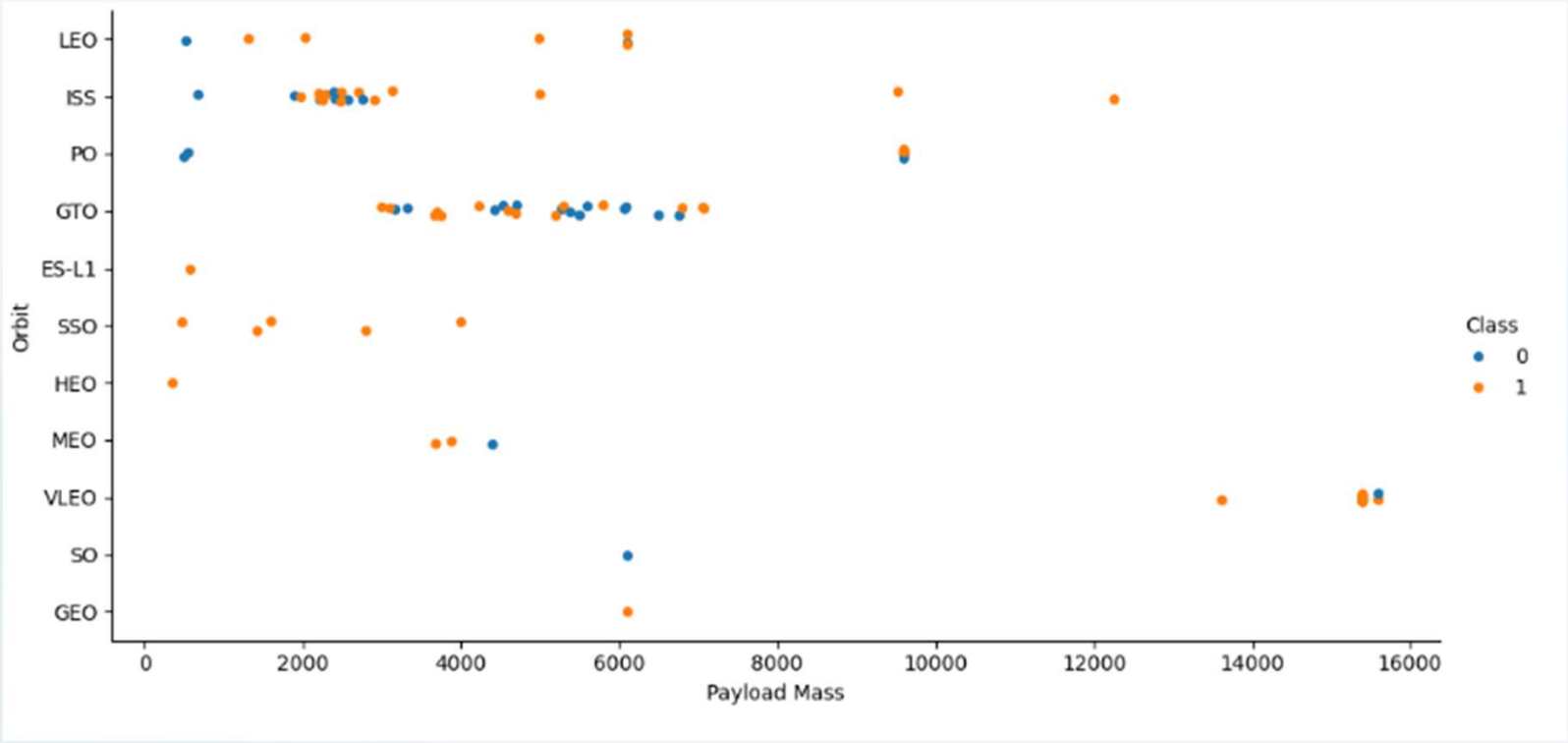


Flight Number vs. Orbit Type



- Early orbit types: LEO, ISS, PO, and GTO don't have a correlation to success in relation to Flight #
- Orbit VLEO success rate has improved with more flights

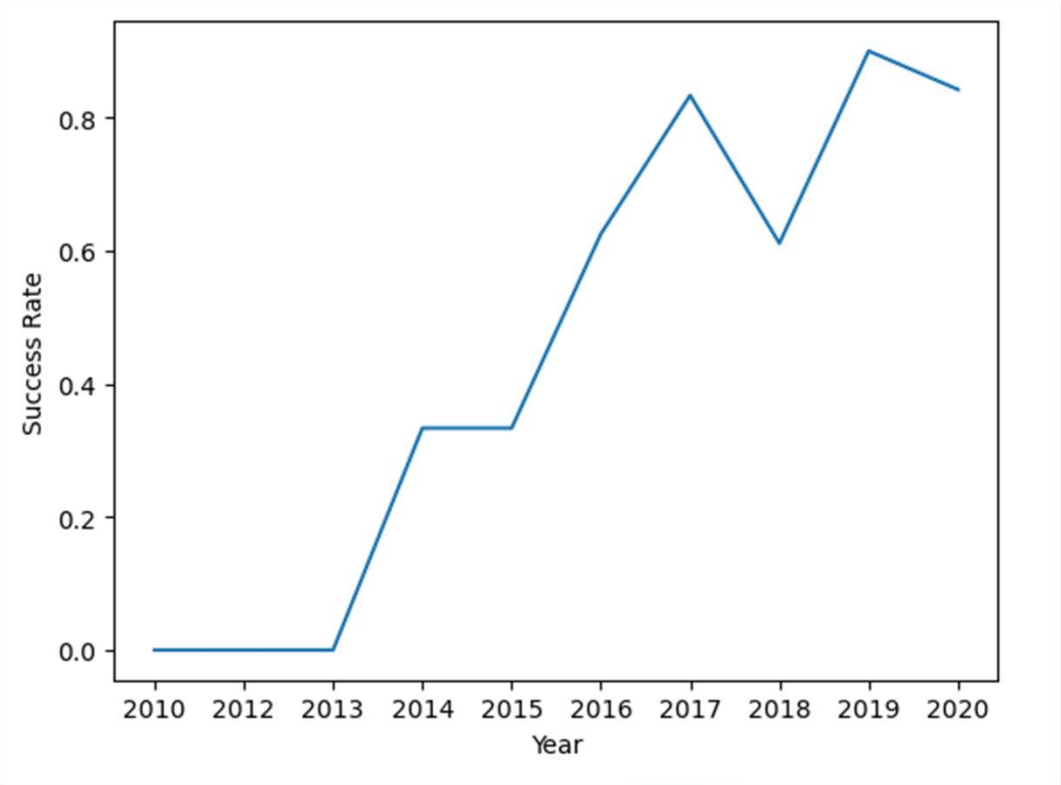
Payload vs. Orbit Type



- GTO orbits have a weak correlation between payload and landing success
- ISS orbits have an improved landing success rate with higher payloads

Landing Success Yearly Trend

- Generally, landing success rate has increased over time with a brief decline in 2018



All Launch Site Names


```
%sql select distinct "launch_site" from  SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
 * sqlite:///my_data1.db
Done.
```

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

- Display of unique launch sites within the data set

Launch Site Names Begin with 'CCA'


```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

- Displaying first 5 records of launch sites whose name includes ‘CCA’

Total Payload Mass


```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTABLE where Customer like 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

total_payload_mass
45596

- Displaying total payload mass from NASA (CRS) missions

Average Payload Mass by F9 v1.1


```
%sql select avg(PAYLOAD_MASS__KG_) as avg_payload_mass from SPACEXTABLE where "booster_version" like "F9 v1
```

```
 * sqlite:///my_data1.db
Done.
```

avg_payload_mass
2534.6666666666665

- Displaying the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date


```
%sql select min(Date) from SPACEXTABLE where "landing_outcome" = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
2015-12-22

- Date of first successful ground pad landing

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- List of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes


```
%sql select "mission_outcome", count (*) as total from SPACEXTABLE group by "mission_outcome"
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Display of number of mission successes and failures

Boosters Carried Maximum Payload

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

- List of booster names which have carried the maximum payload mass

2015 Launch Records


```
%%sql select substr(Date,6,2) as month,Date, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE
    where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5) = '2015'
```

```
 * sqlite:///my_data1.db
Done.
```

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

- List Launch site, booster version and date of failed landings to a drone ship in 2015

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


```
%%sql select Landing_Outcome, count (*) as count_outcome from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome order by count_outcome desc
```

```
 * sqlite:///my_data1.db
Done.
```

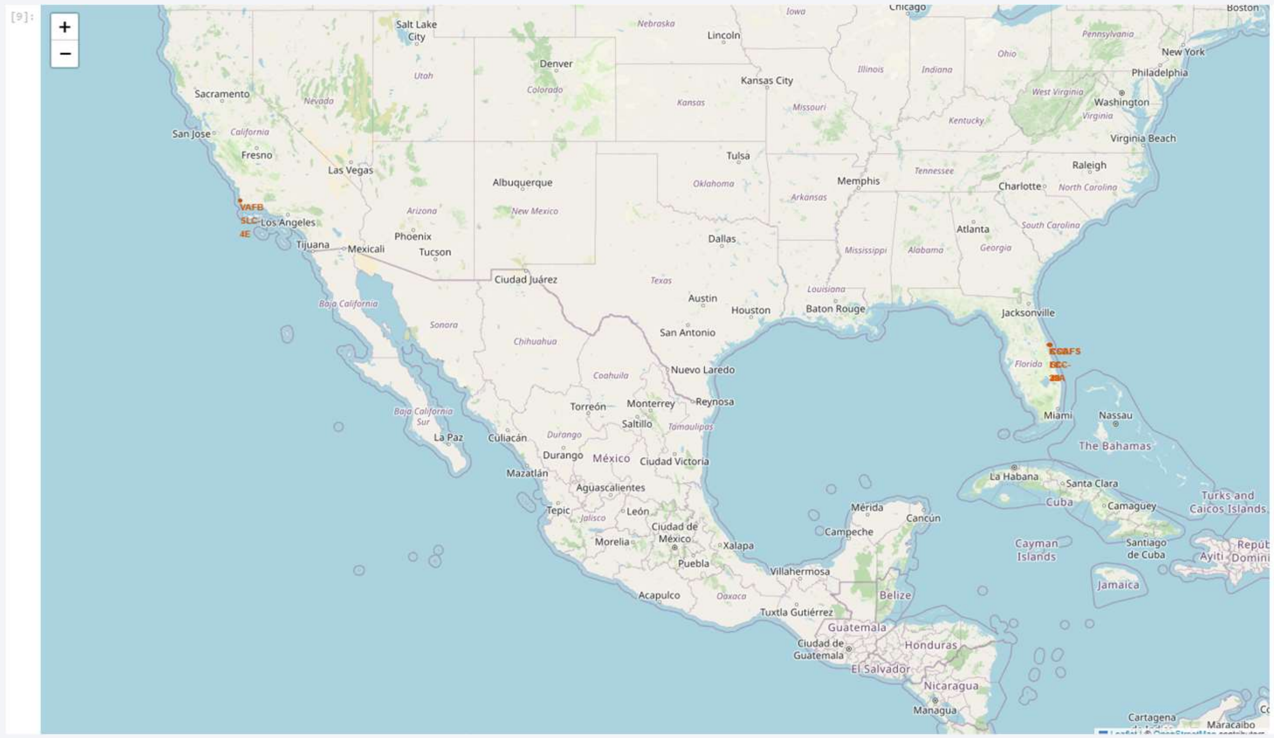
Landing_Outcome	count_outcome
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Ranking of number of landing outcomes between 2010-06-04 and 2017-03-20

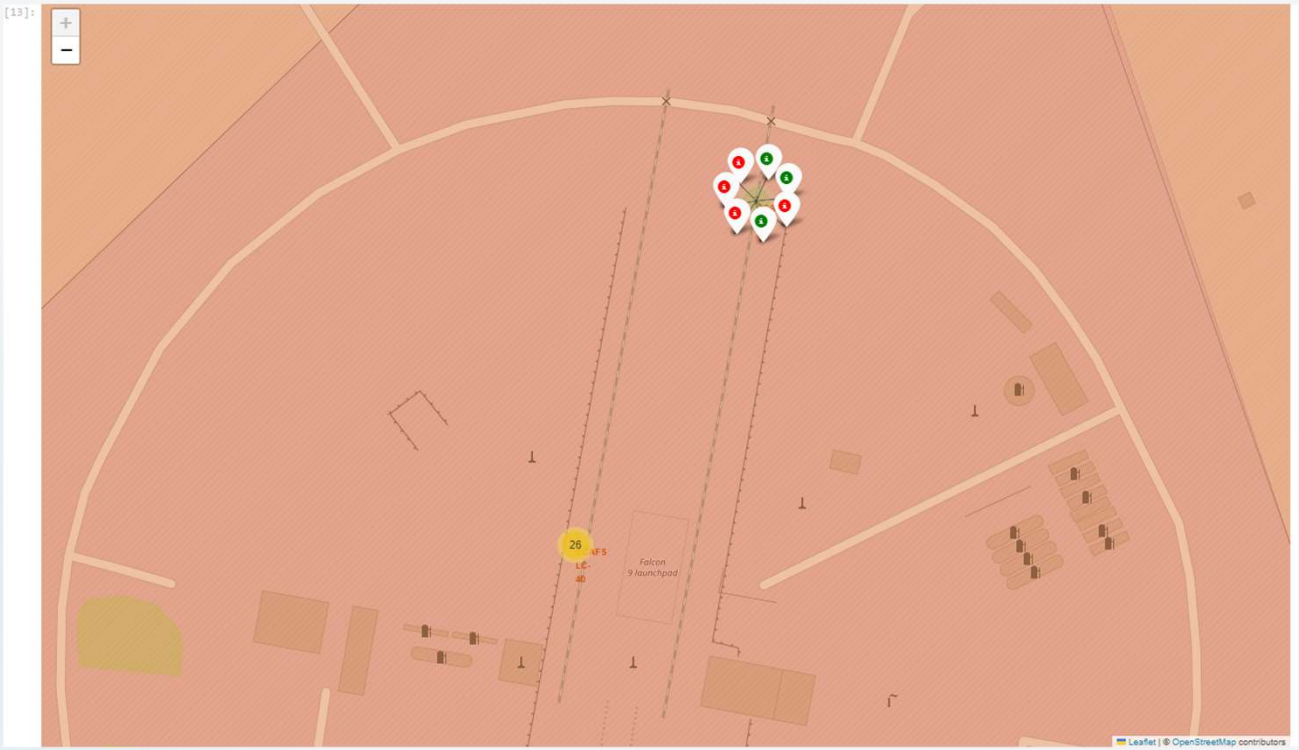

Section 3
Launch Sites
Proximities Analysis

All Launch Site Locations

- The map to the left shows all launch site locations pulled from the Falcon 9 dataset



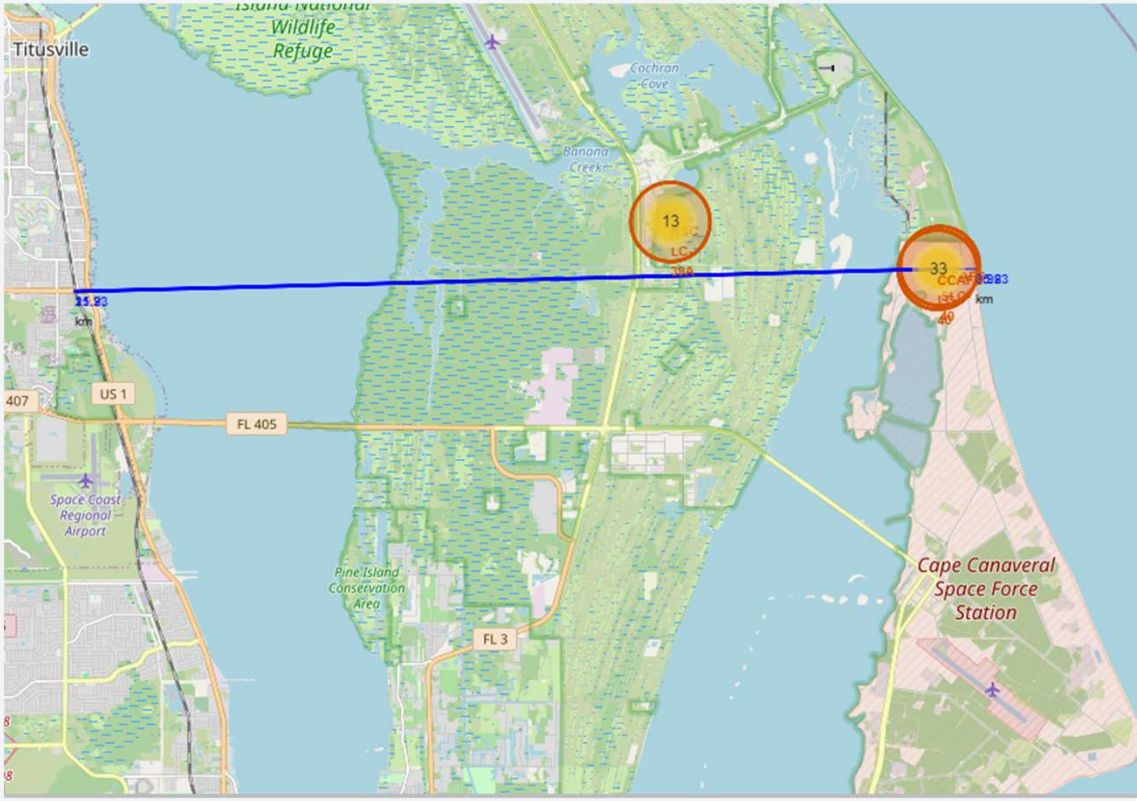
Successful vs Unsuccessful Landing Outcomes

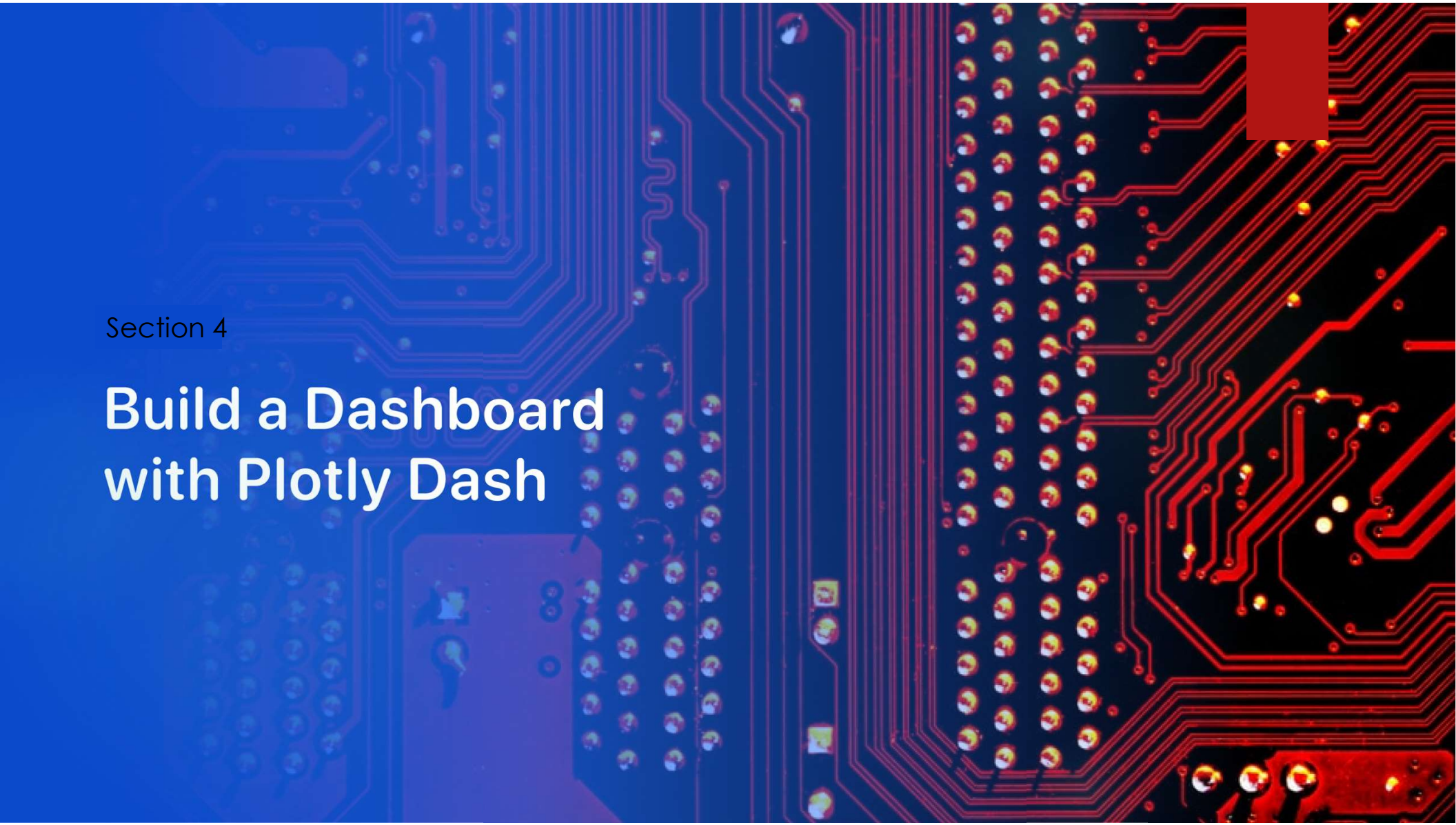


- We've added color coded launch positions to show which launches resulted in successful or unsuccessful landings at completion of the mission

Proximity of Launch Sites to Key Features

- To the right we see the distance and direction from our launch site to the nearest railway and costline

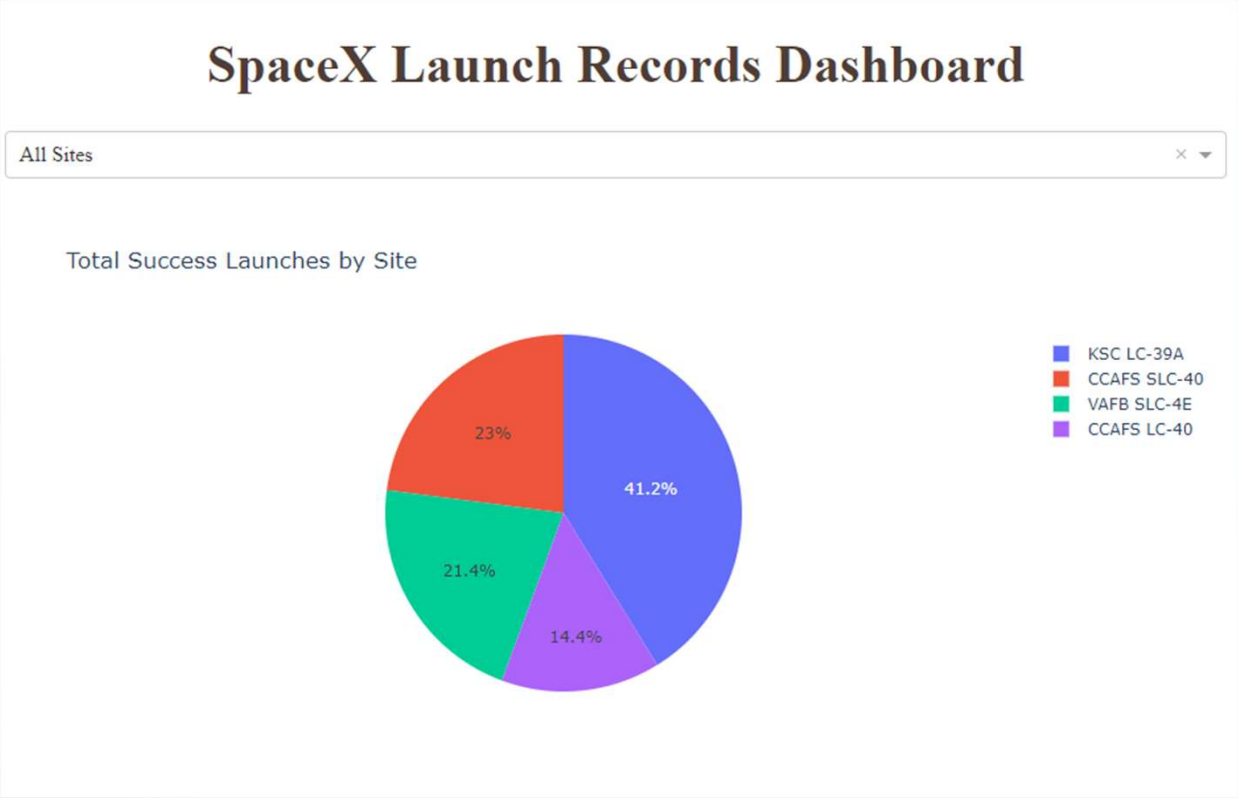




Section 4

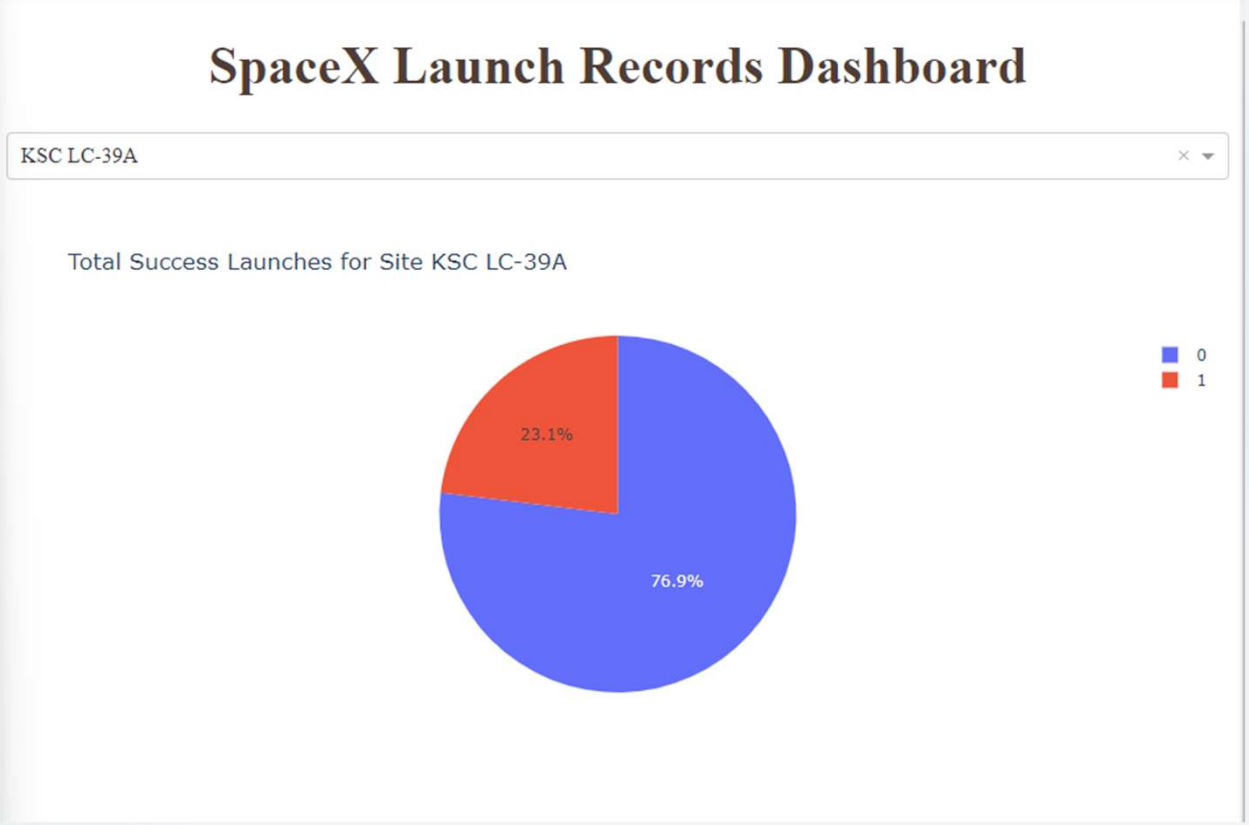
Build a Dashboard with Plotly Dash

Distribution of Successes by Launch Site



- This shows that KSC LC-39A has the best success rate of the launch sites

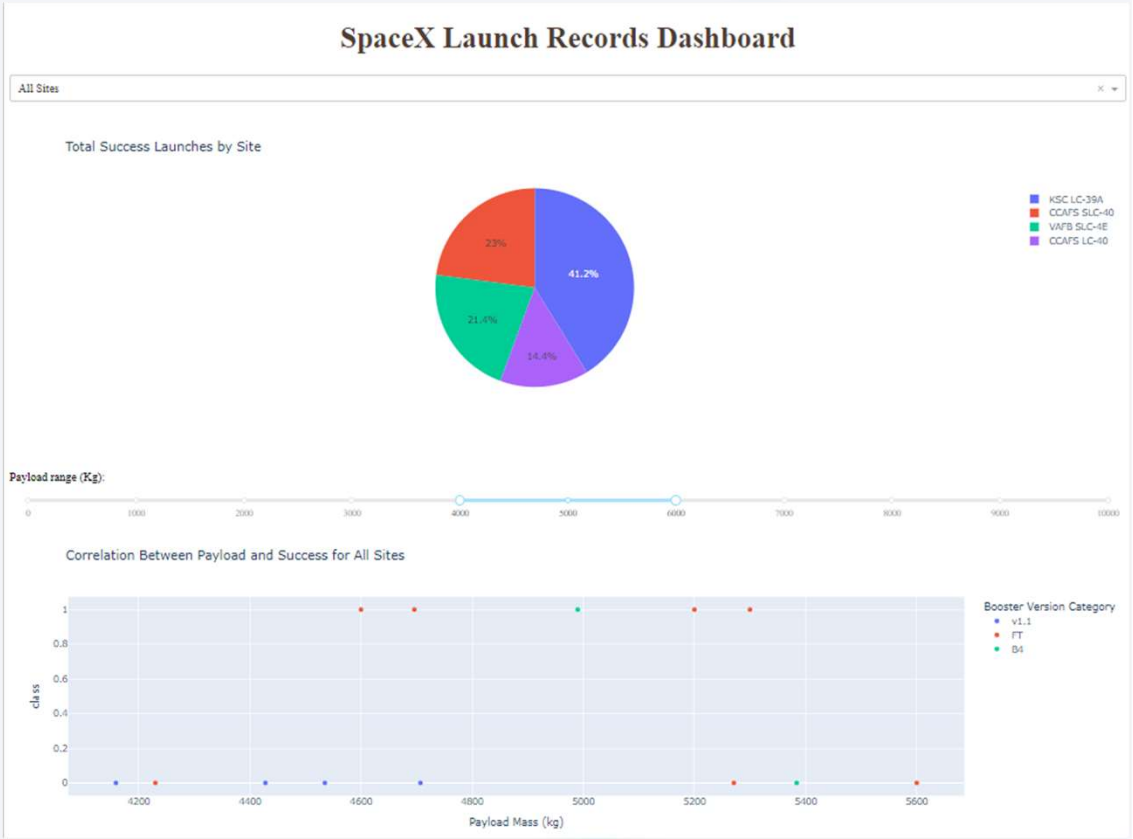
KSC LC-39A Success Rate



- KSC LC-39A has the highest success rate of all the launch sites at 76.9%

Success at all Sites with Higher Payloads

- For mid range payloads, the FT booster gives the best success from all of the launch sites.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- While using just the test set alone did not yield insight into the best model, using the entire dataset showed that SVM will give the most accurate predictions

Scores and Accuracy for Test Set

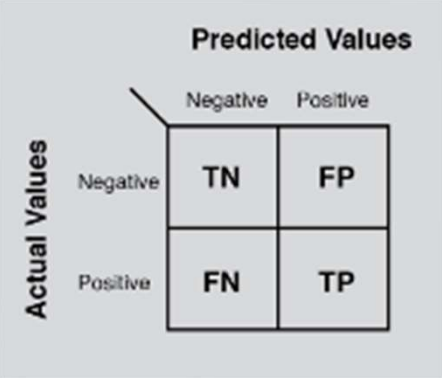
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.750000	0.800000
F1_Score	0.888889	0.888889	0.857143	0.888889
Accuracy	0.833333	0.833333	0.777778	0.833333

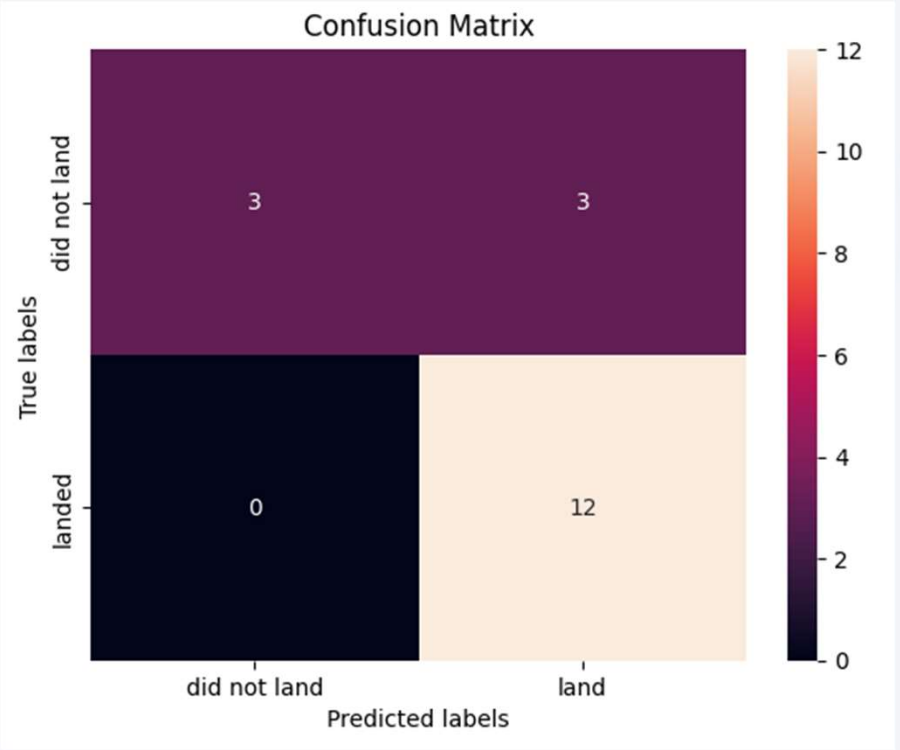
Scores and Accuracy for Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.797297	0.819444
F1_Score	0.909091	0.916031	0.887218	0.900763
Accuracy	0.866667	0.877778	0.833333	0.855556

Confusion Matrix

- Examining the Confusion Matrix see that largest problem was False Positives, but overall, the SVM yielded the most accurate predictions.





Conclusions

- SVM gave us the most accuracy out of the different models
- KSC LC-39A has the highest success rate of all the launch sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
- Success rate has increased over time as more flights have occurred

Appendix

- Github Main Page for IBM Data Science Capstone
- Thank you to Coursera and the Instructors for the Course

Thank you!