IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ashrita Moola
12/28/2024

Outline

Executive Summary

Methodologies

This project uses machine learning predict the price of each launch and find out if SpaceX will reuse the first stage after landing. The following steps were used:

- Collecting the data with SpaceX REST API web scraping
- Data wrangling cleans the data to make it usable
- Analyze the data with SQL to calculate statistics of failed and successful launches, sorting it by booster type, landing, and payload mass
- Visualize the data with matplotlib and seaborn to find relationships between each of the features by class
- Find launch sites successes and distances with Folium
- Build models to predict landing outcomes with Logistic Regression, Support Vector Machine (SVM), K-nearest neighbors (KNN), and decision tree

Results

- Since 2013, SpaceX has consistently improved their success rate with a slight drop in 2018
- Logistic Regression (threshold=.67) was slightly outperformed other models with an accuracy of .89
- Most of the launch sites were close to the coast
- There is 100% success rate for ES_L1, GEO, HEO, SSO orbits
- KSC LC 39A has the highest success rate though CCAFS SLC 40 has been used the most

Introduction

Background

SpaceX is a rising leader in the space industry as it brings 165 million dollar missions to just 65 million. This is mostly due its reuse of the first stage of the rocket. Using public data, we will create models to predict whether the first stage will land and what that means for the price of the launch.

Goals of this Project

- Rate of successful launches
- Best launch sites and booster versions
- Best predictive classification model for successful landing
- How payload mass, orbit, number of flights, and launch site affect first-stage landing success

Section 1

Methodology

Methodology

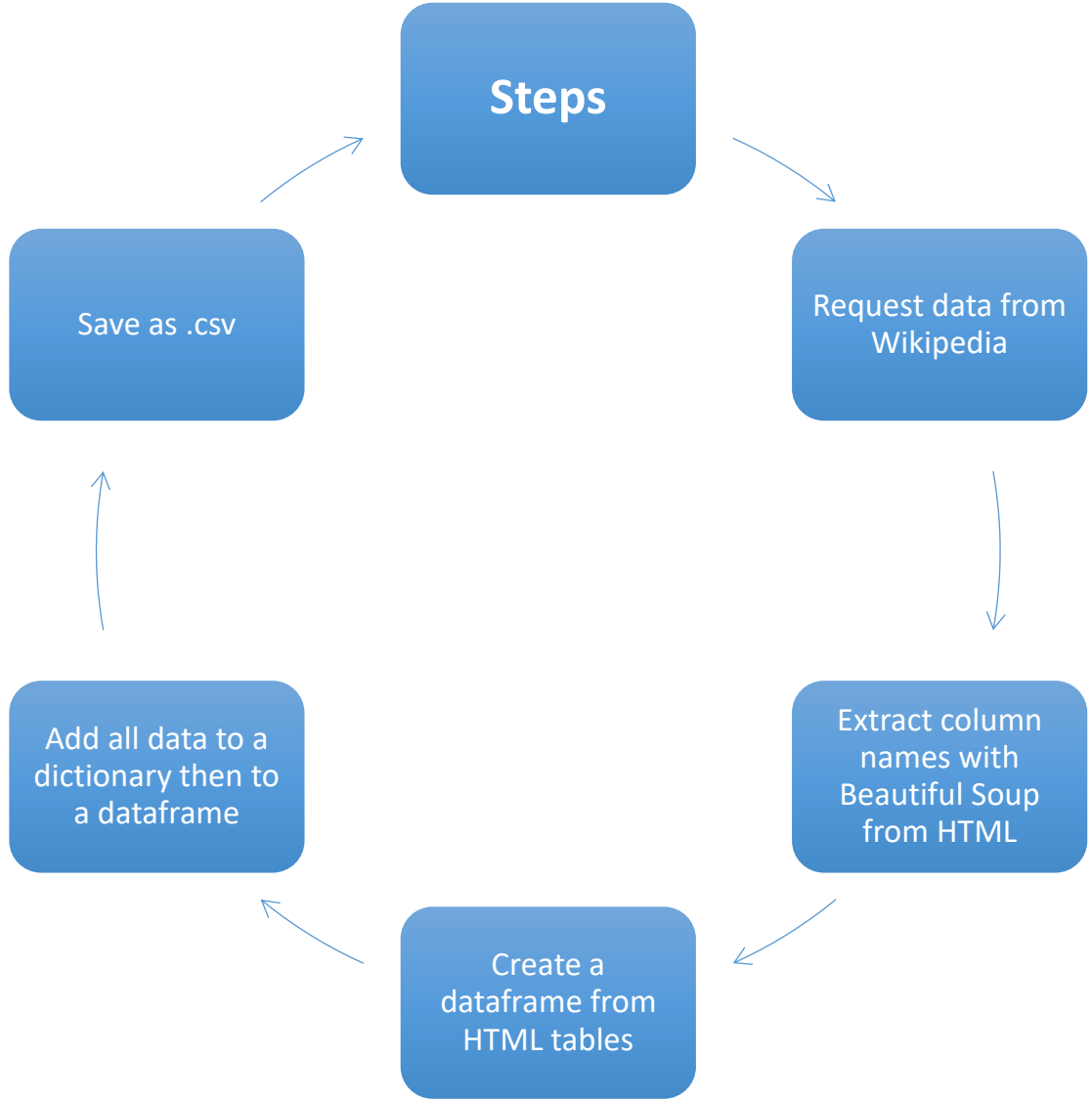
Steps

- ✓Collecting the data with SpaceX REST API and webscraping
- ✓Clean the data with data wrangling
- ✓Analyze the data with
- ✓Visualize the data on a dashboard with Folium, Plotly
- ✓Visualize on graphs with matplotlib and seaborn
- ✓Build models to predict landing outcomes with Logistic Regression, Support Vector Machine (SVM), K-nearest neighbors (KNN), and decision tree

Data Collection – SpaceX API

https://github.com/orangeostick/spacex-project/blob/dc41e600567bae711a36c1b072954f47a3e82997/edadataviz.ipynb

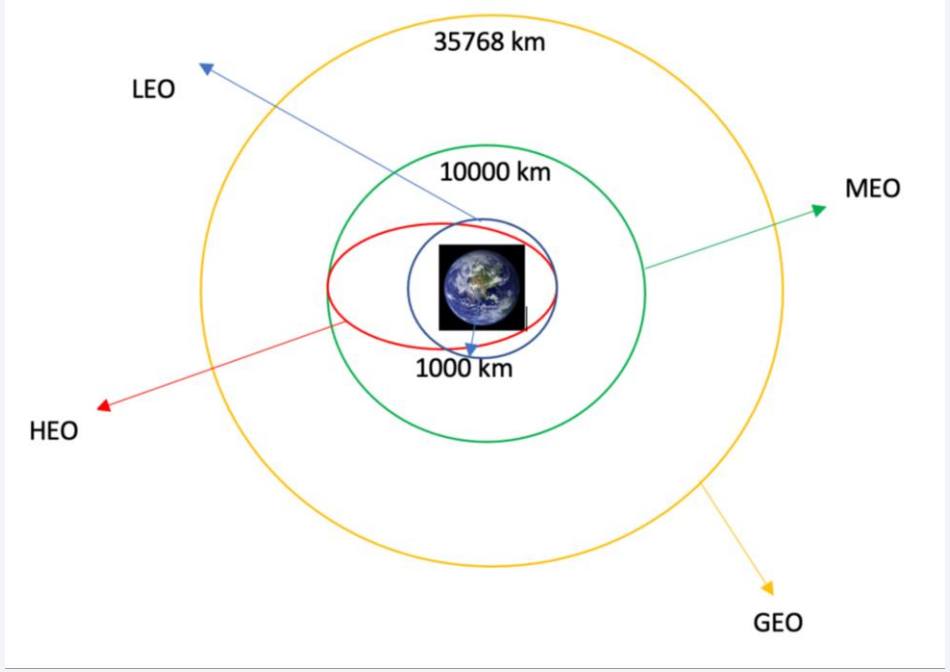
Data Collection – Web Scraping



https://github.com/AlvaroOrtiz2001/spacex-project/blob/662e99b8cc79018c269c93fdbdf6823e48a0cda7/jupyter-labs-webscraping%20(FINAL).ipynb

Data Wrangling

- **Import** Libraries and Define Auxiliary Functions
- **Calculate** the number of **launches** for each launch site
- **Calculate** the number and occurrence of each **orbit**
- Assign 0 or 1 to **failed** and **successful** launches
- **Save** as .csv



https://github.com/orangeostick/spacex-project/blob/dc41e600567bae711a36c1b072954f47a3e82997/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Relationships

- Flight # vs Payload
- Flight # Launch Site
- Payload Mass (kg) vs Launch Site
- Payload Mass (kg) vs Orbit type

Graphs

- Scatter plot (catplot in seaborn) shows if a relationship could exist
- Bar graph shows comparisons between discrete categories

https://github.com/orangeostick/spacex-project/blob/dc41e600567bae711a36c1b072954f47a3e82997/edadataviz.ipynb

EDA with SQL

```
%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

Queries

- Names of **unique** launch sites
- Names of launch sites starting with '**CCA**'
- **Total payload mass** carried by **boosters** launched by **NASA (CRS)**
- **Average payload mass** carried by booster version **F9 v1.1**
- **Date** of first **successful landing** outcome in ground pad
- **Total** # of successful and failure mission **outcomes**
- **Names** of booster versions carrying maximum payload mass

```
%sql select mission_outcome, COUNT(*) AS Total FROM SPACEXTBL 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

- **Names** of successfully landed boosters in drone ship and payload mass 4000-6000
- **Month** names, **failure** landings in drone ship, **booster** versions, **launch site** for the months in **2015**
- **Rank** the # 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

https://github.com/orangeostick/spacex-project/blob/dc41e600567bae711a36c1b072954f47a3e82997/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers indicating launch sites

- Mark blue circle for **NASA Johnson Space Center** with label of coordinates
- Mark red circle for all launch sites with label of coordinates

Colored markers based on launch outcome to show **success rate** of launch sites

- Green marker if **successful**
- Red marker if **failed**

Distances from launch sites to **proximities**

- Lines from **CCAFS SLC 40** launch site to closest railroad, coastline, highway, and city

https://github.com/orangeostick/spacex-project/blob/dc41e600567bae711a36c1b072954f47a3e82997/lab_jupyter_launch_site_location.ipynb

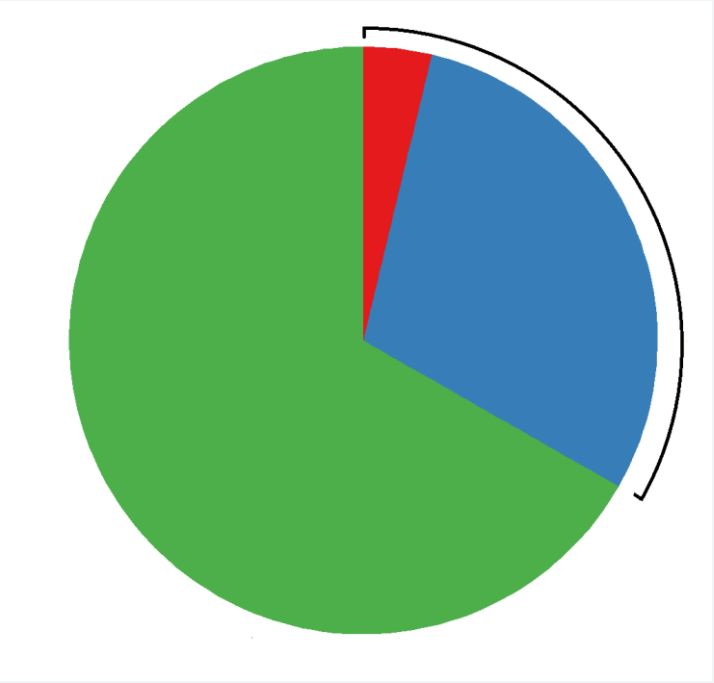
Build a Dashboard with Plotly Dash

Payload Mass Slider – Can change the mass to see change in relation to booster version

Launch Site Dropdown – can change launch site for pie chart to check distribution

Pie chart – shows percentage of successes compared to other launch sites

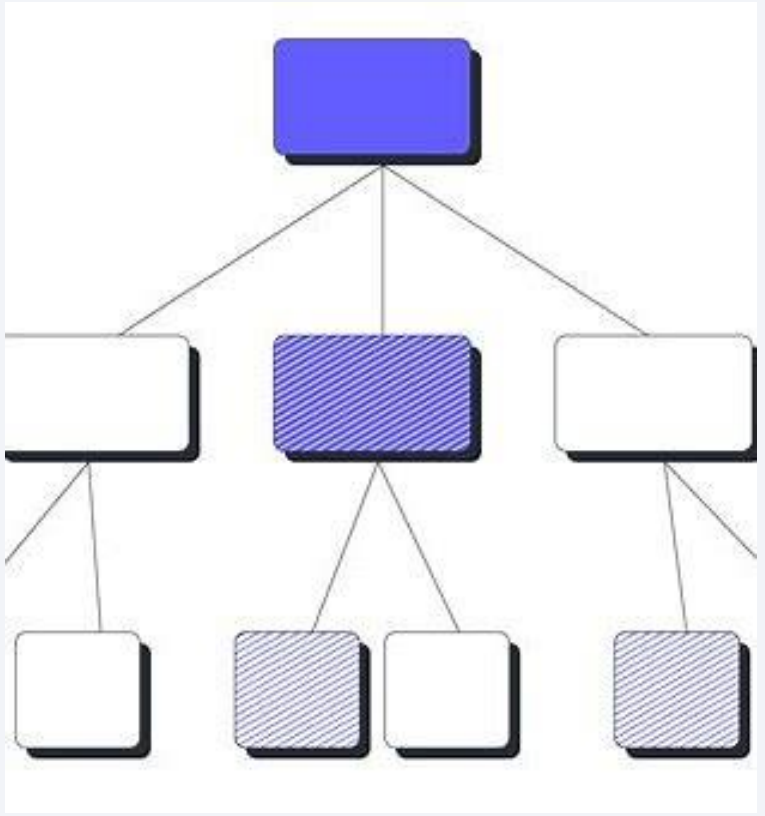
Scatter Plot – shows the correlation between payload mass and success rate for each launch sites and sorted by booster version



https://github.com/orangeostick/spacex-project/blob/db00fd81a725f556f1db04034677d6a9d606a119/SpaceX_Interactive_Visual_Analytics_Plotly.py

Predictive Analysis (Classification)

- Create a NumPy array from ‘Class’ data
- Standardize with StandardScalar, fit and transform data
- Split the data with train_test_split
- Create GridSearchCV with cv=10 for optimization
- Apply this on Logistic Regression, SVM, KNN, and decision tree
- Calculate accuracy with .score() and confusion matrix
- Pick best model by graphing scores on bar graph



https://github.com/orangeostick/spacex-project/blob/db00fd81a725f556f1db04034677d6a9d606a119/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

Exploratory data analysis results

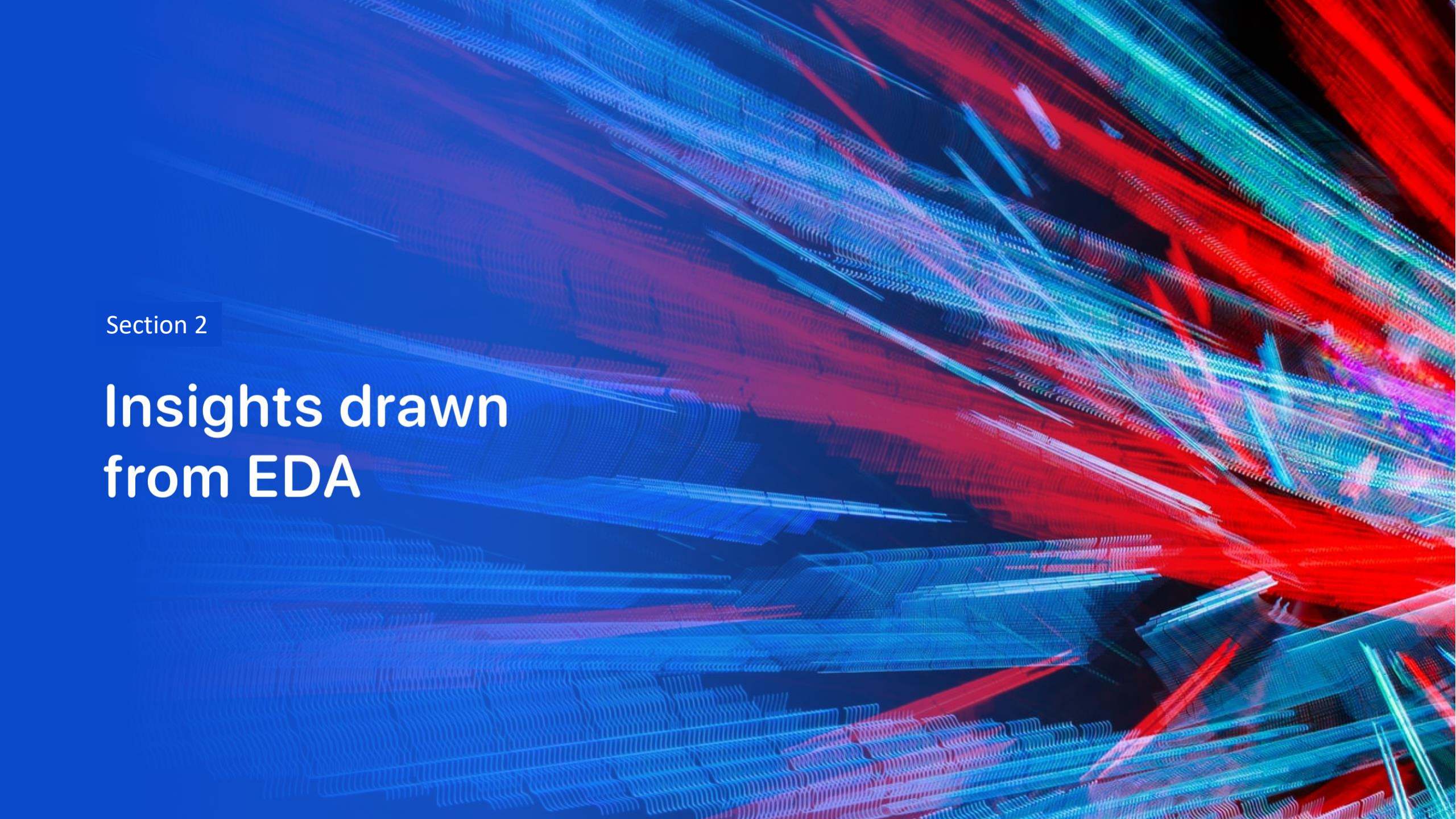
- There is 100% success rate for ES_L1, GEO, HEO, SSO orbits
- KSC LC 39A has the highest success rate though CCAFS SLC 40 has been used the most
- Since 2013, SpaceX has consistently improved their success rate with a slight drop in 2018

Interactive analytics

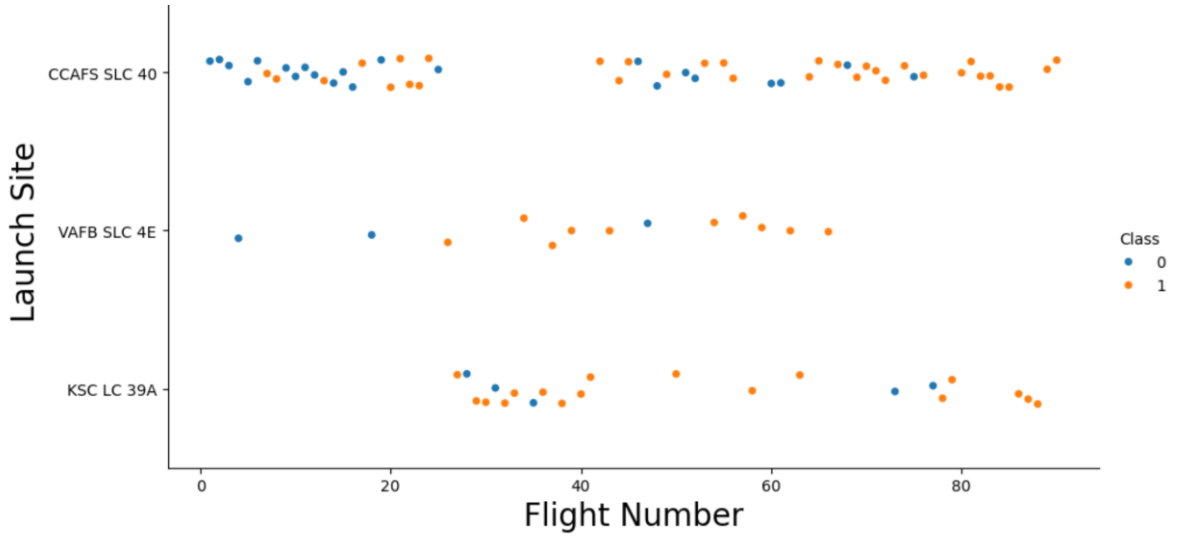
- Most of the launch sites were close to the coast

Predictive analysis results

- Logistic Regression (threshold=.67) was slightly outperformed other models with an accuracy of .89


Section 2
Insights drawn from EDA

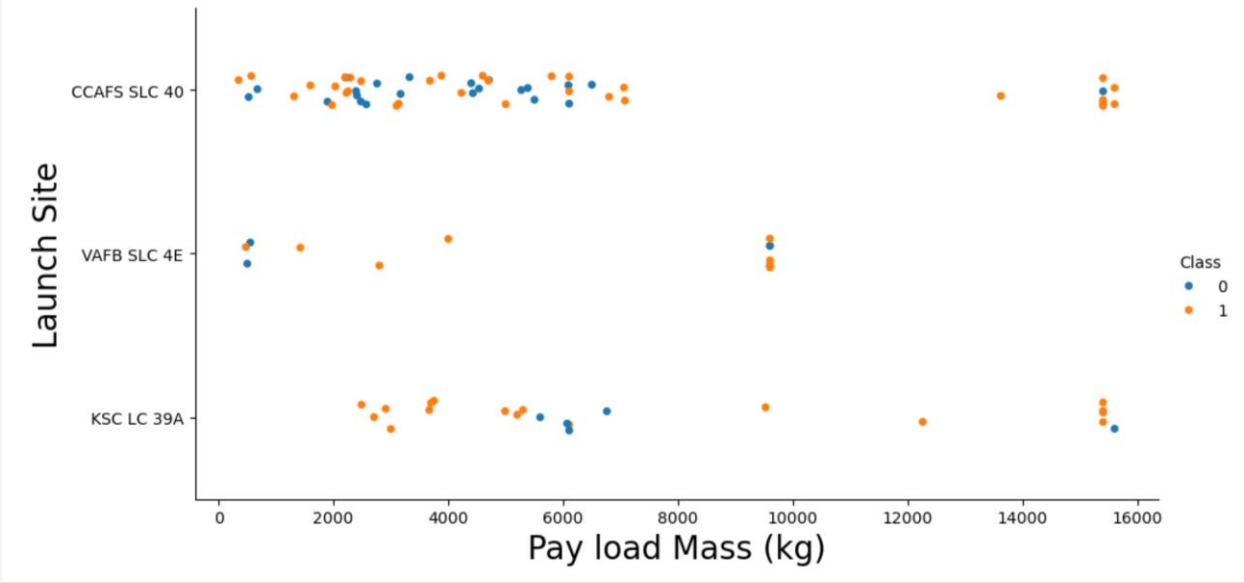
Flight Number vs. Launch Site

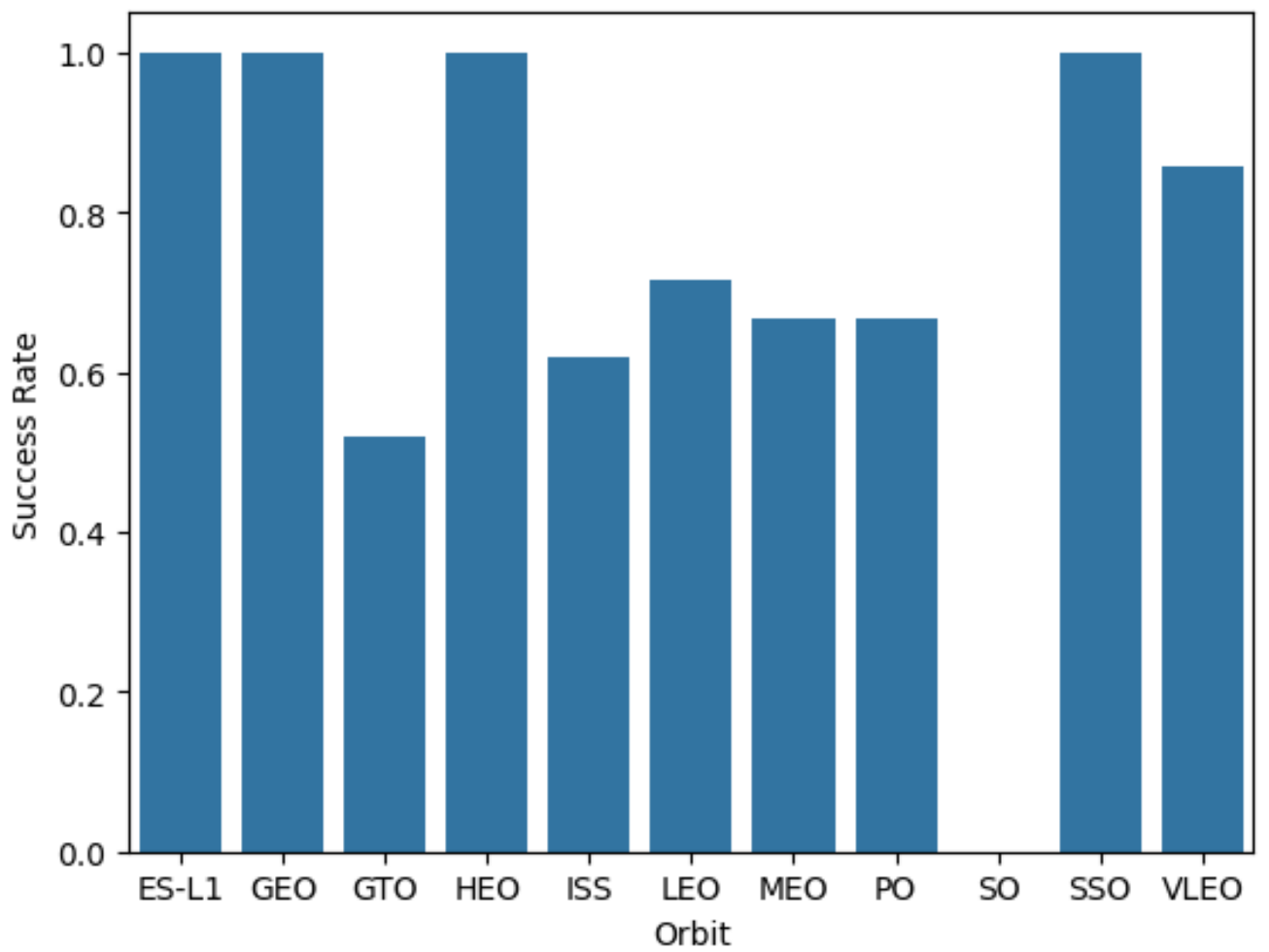


- The success rate for each launch site increased with more flights (orange=success, blue=fail)
- KSC LC 39A had the highest rate of success
- CCAFS SLC had the most launches
- Success rate increased with a higher number of launches

Payload vs. Launch Site

- Most of the launches had a payload mass around 0-8000 kg
- KSC LC 39A has the most success
- VAFB SLC 4E has the fewest launches but high success
- Above 7500 kg, almost every launch was successful



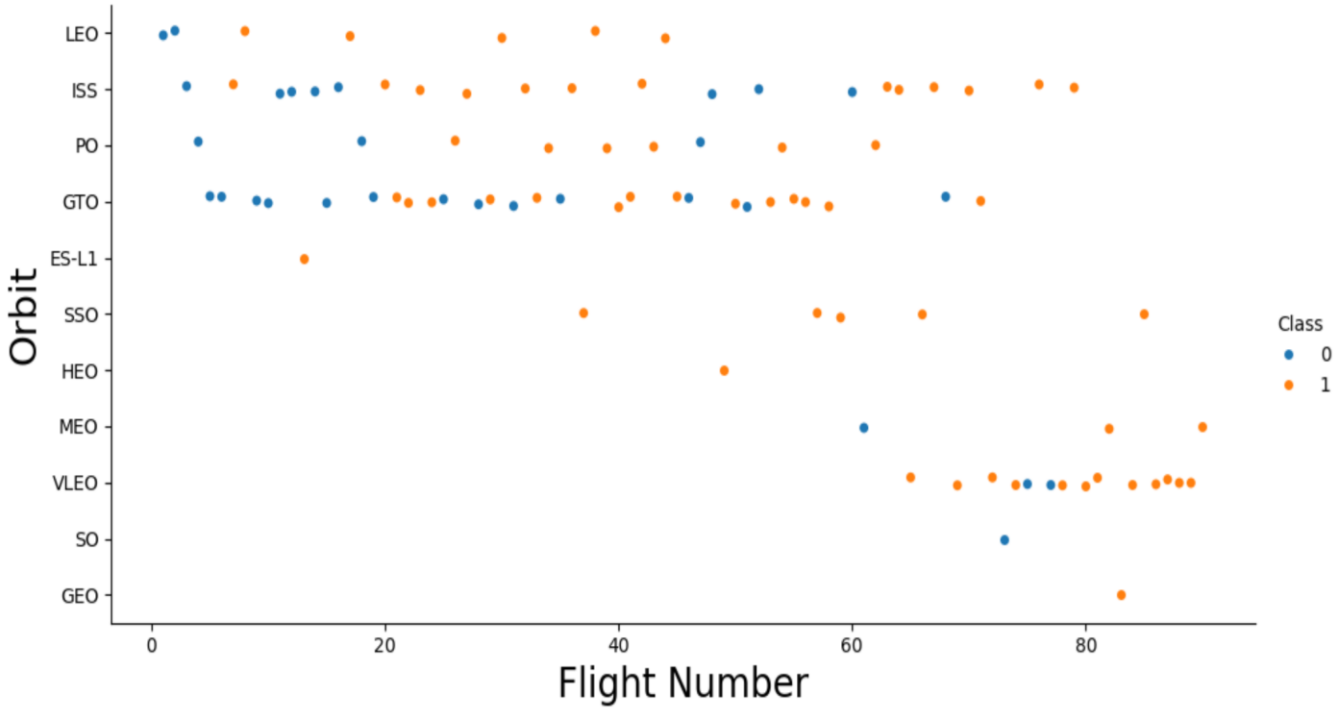


Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have **100%** success rates
- GTO, ISS, LEO, MEO, PO, and VLEO are in **median** success
- SO and GTO have the **lowest** success rates

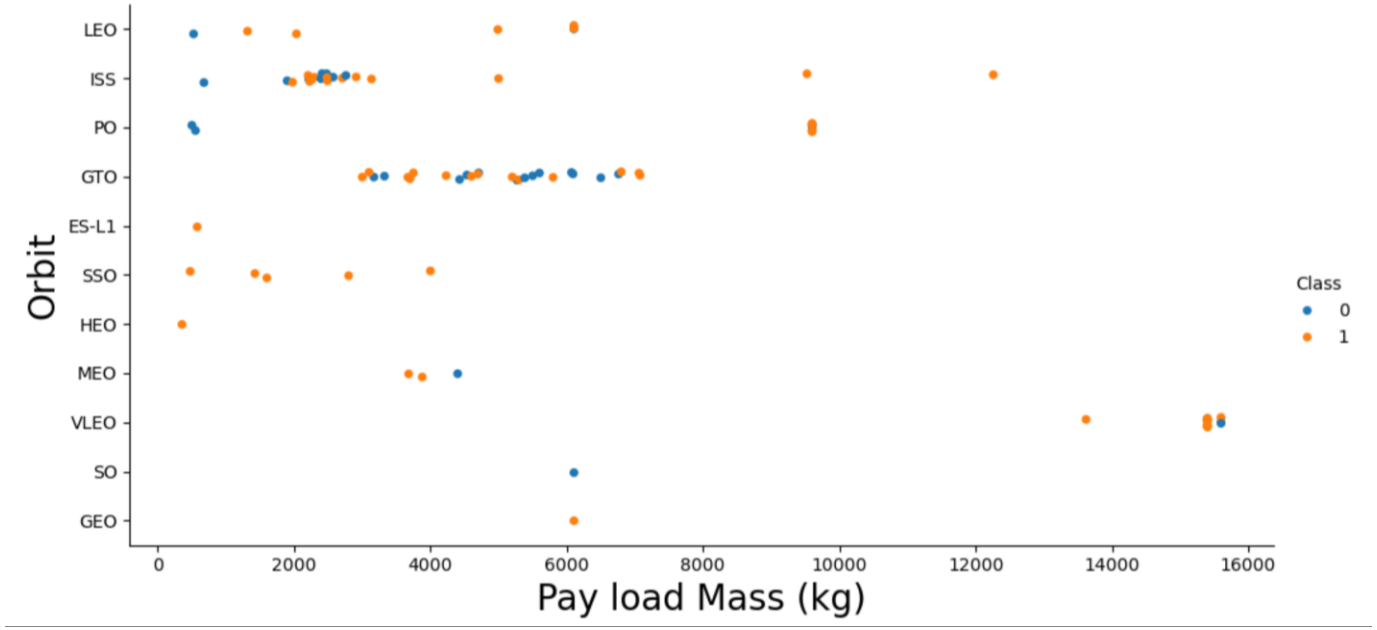
Flight Number vs. Orbit Type

- ISS and GTO have been launched the most
- Success rate typically increases with flight number



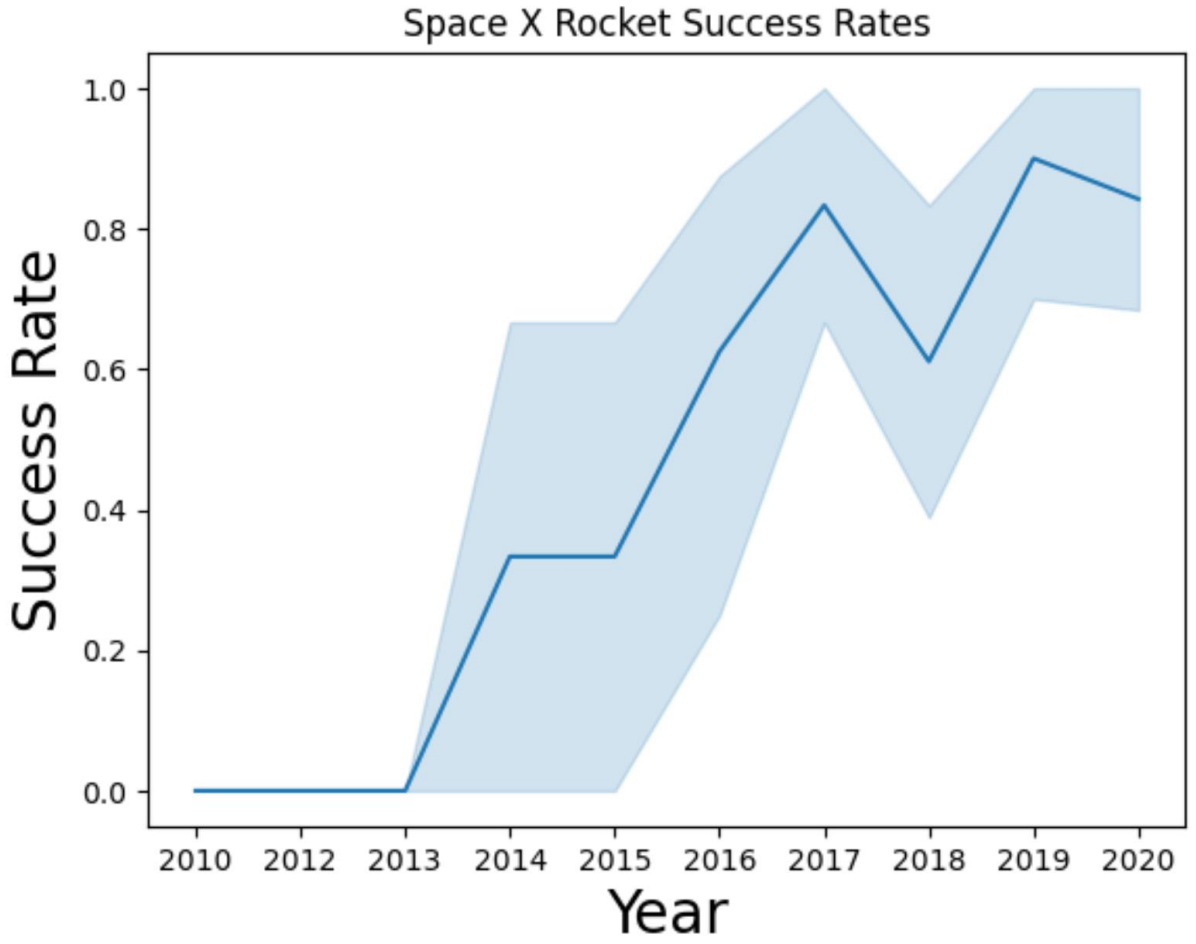
Payload vs. Orbit Type

- High payloads are best for VLEO, PO, and ISS
- GTO and ISS have most launches
- LEO, SSO, ES-L1, HEO, MEO, and GEO have high success rates below 6500 kg payload



Launch Success Yearly Trend

- Overall, the success rates increased from 2010-2020
- There was a lull 2014-2015
- A steep drop occurred 2017-2018 before rising once again



All Launch Site Names

```
%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'

First 5 records where launch site names start with ‘CCA’

```
%sql SELECT * FROM SPACEXTBL 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	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

Total Payload Mass

Sum of all payload mass where customer is NASA (CRS)

```
%sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)'
```

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

sum(payload_mass__kg_)
45596

Average Payload Mass by F9 v1.1

Average the payload mass of F9 v1.1 boosters

```
%sql select avg(payload_mass__kg_) from SPACEXTBL where booster_version = 'F9 v1.1'
```

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

avg(payload_mass__kg_)
2928.4

First Successful Ground Landing Date

```
%sql select min(date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Name of booster versions with a successful drone ship landing and a payload between 4000-6000

```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and payload_mass__kg_ BETI
```

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

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

Total Number of Successful and Failure Mission Outcomes

Count # of Successful and Failed missions

```
%sql select mission_outcome, COUNT(*) AS Total FROM SPACEXTBL 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

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

2015 Launch Records

Based on month

```
%sql select substr(Date, 6, 2) as month, booster_version, launch_site, landing_outcome from SPACEXTBL where landing
```

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

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

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

Count sorted by landing outcome

```
%sql SELECT landing_outcome, COUNT(landing_outcome) AS Total FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '20
```

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

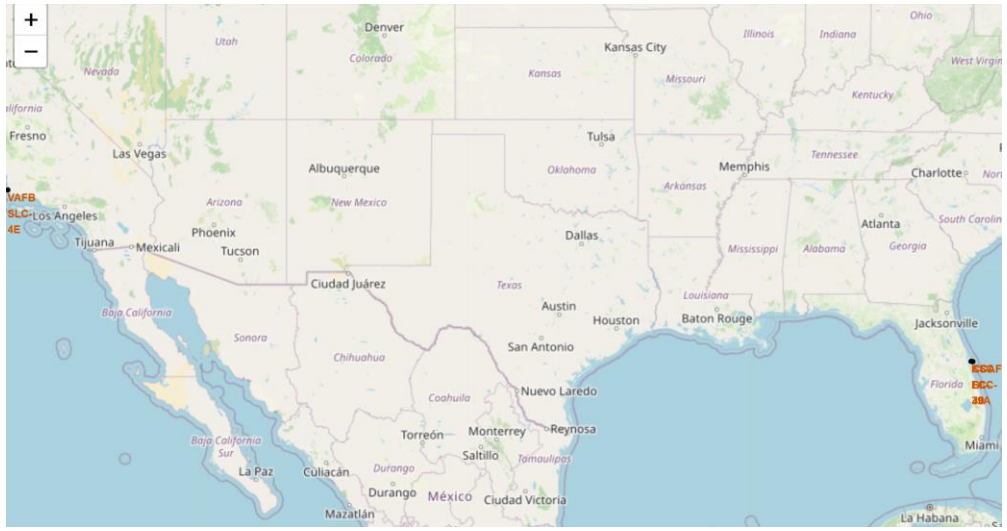
Landing_Outcome	Total
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

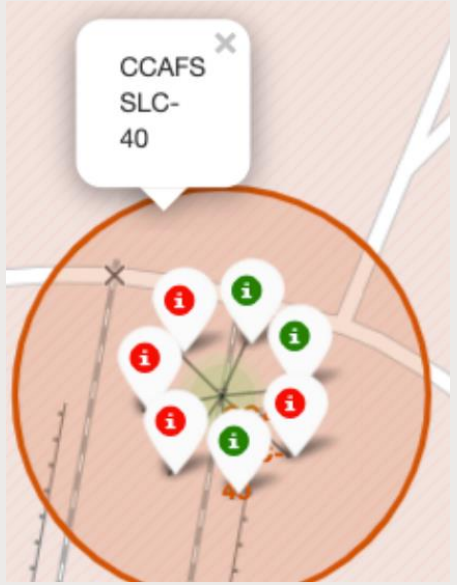
Launch Sites

- The launch sites are very **close** to the **coastline** so the **rocket** will fall into the **ocean** instead of a populated area if something goes **wrong**
- They are also **close** to the **equator** because the natural **rotation** of the Earth helps save on **fuel**



Launch Outcomes

- **Markers**
- Green markers for successful landings
- Red for failed landing

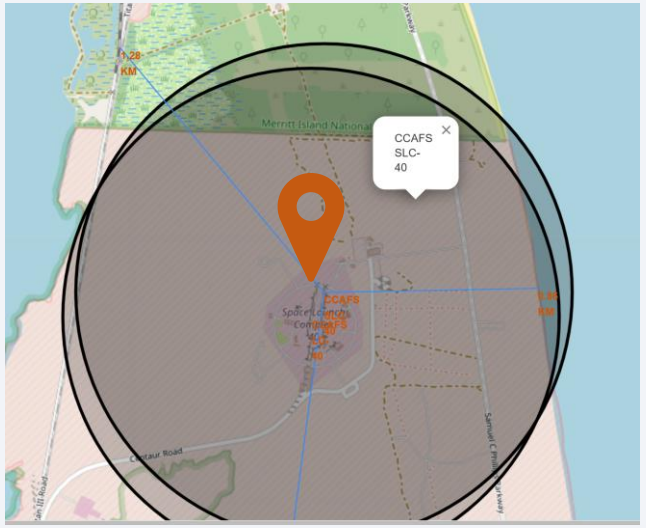


Distance to Proximities

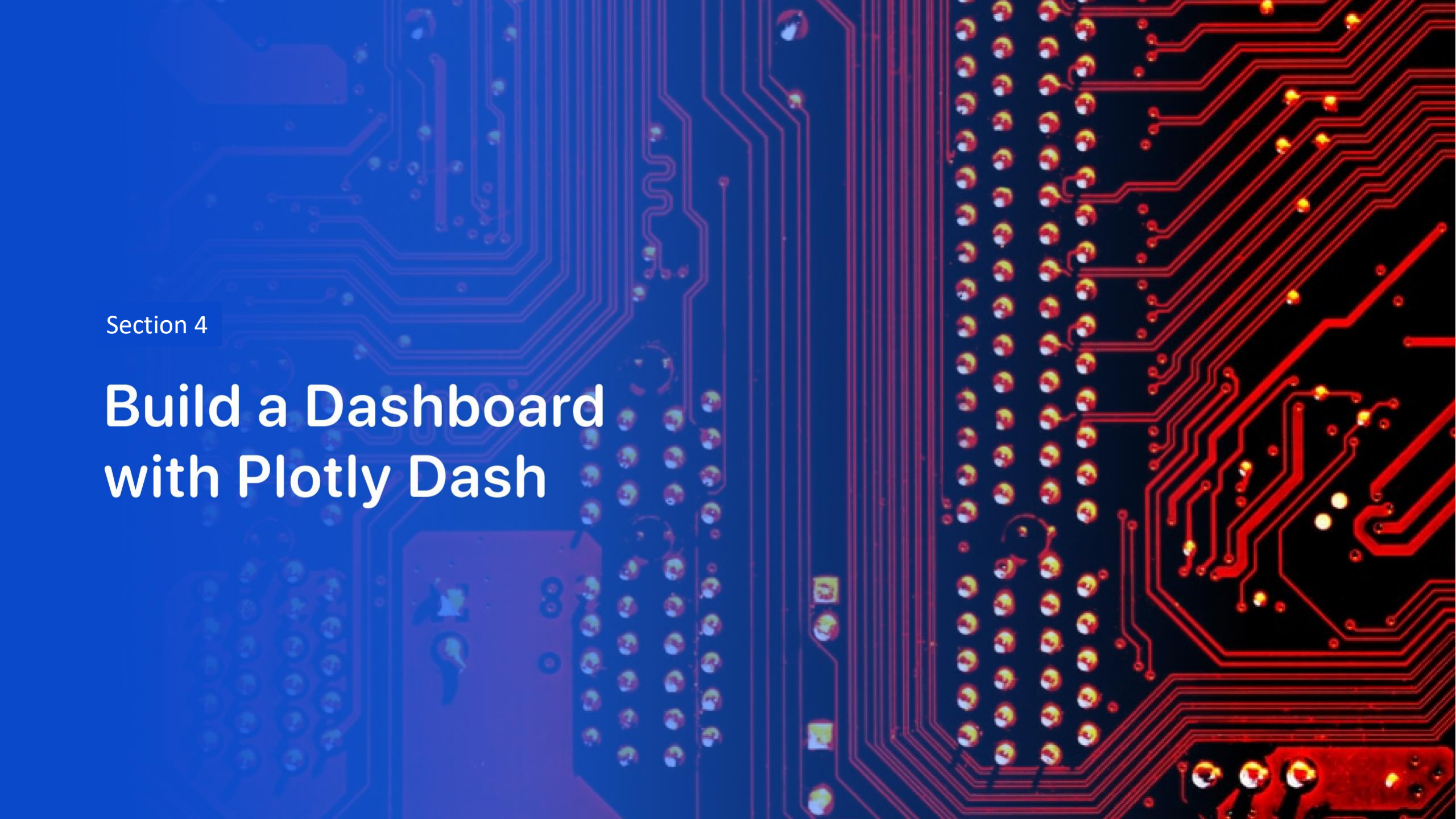
0.86 KM TO **COASTLINE** WHICH IS CRUCIAL IN CASE OF A FAILED LAUNCH TO DEPOSIT THE FAILED ROCKET

21.96 KM TO NEAREST **RAILWAY** AND **26.88 KM** TO NEAREST **HIGHWAY** BOTH OF WHICH ARE NECESSARY TO QUICKLY TRANSPORT PARTS

23.23 KM TO NEAREST **CITY**

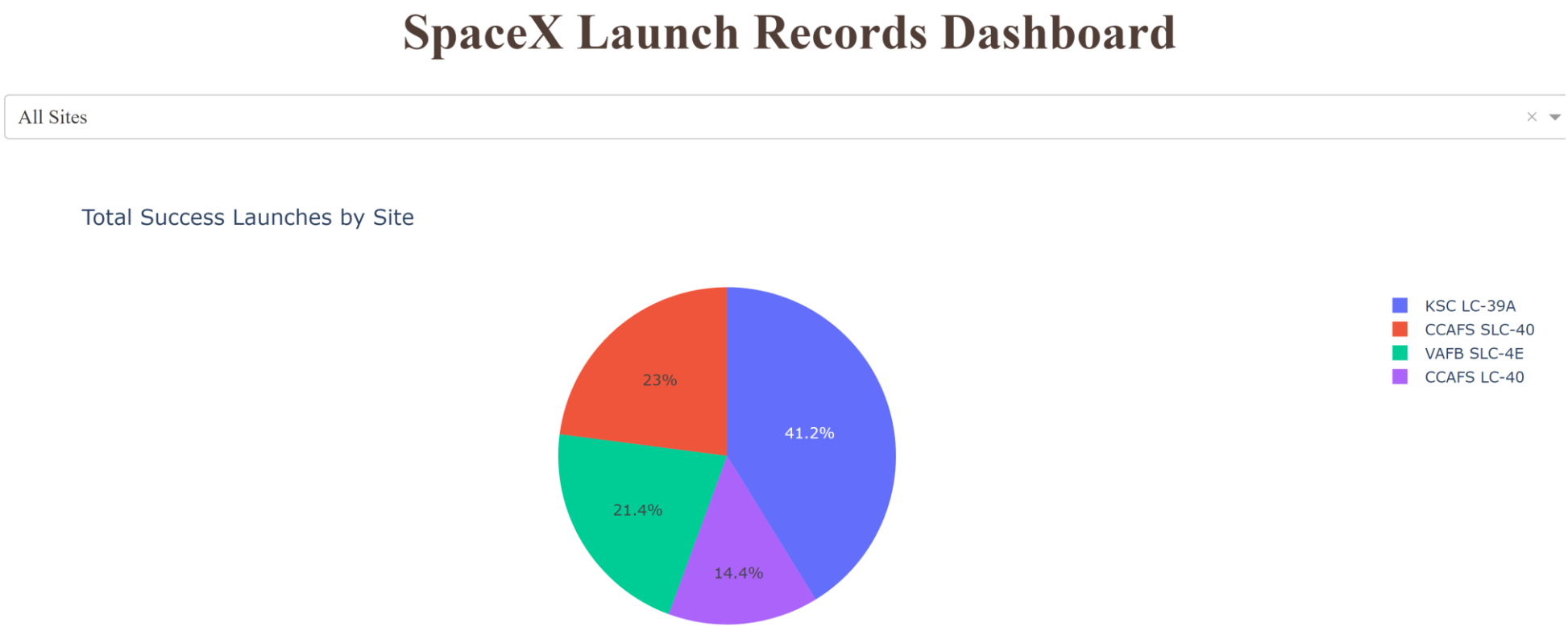


Distance to coastline and railroad


Section 4
Build a Dashboard with Plotly Dash

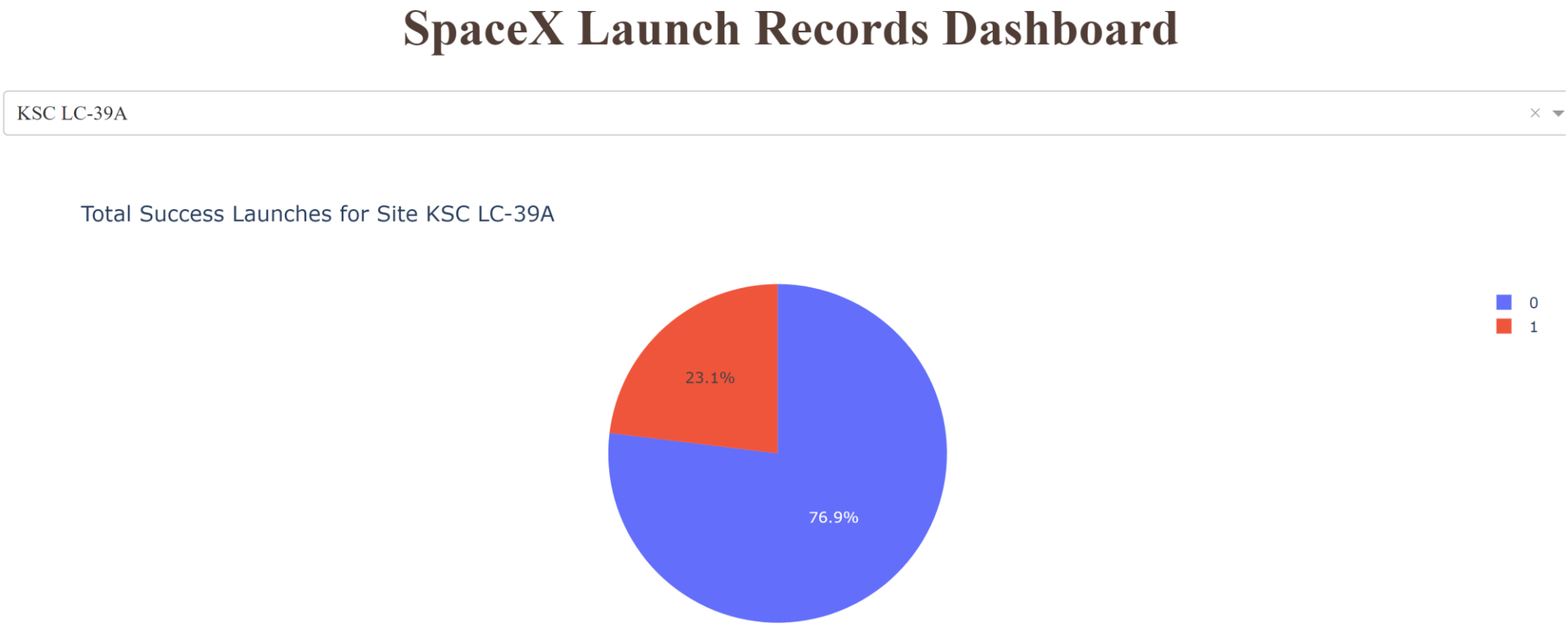
Launch Success Pie Chart

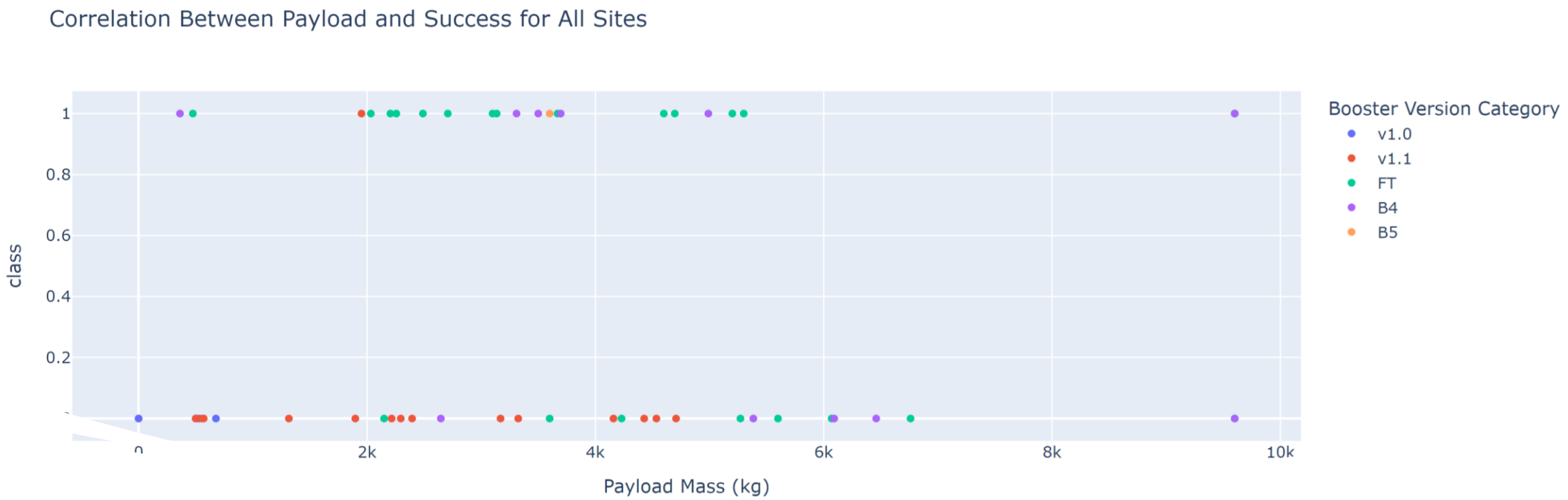
KSC LC-39A has the most **success** out of all launch sites (41.2%)



Launch Site with Highest Success Rate

KSC LC-39A has the highest total success rate out of all the launch sites (76.9%)





Payload vs. Launch Outcome (for all sites)

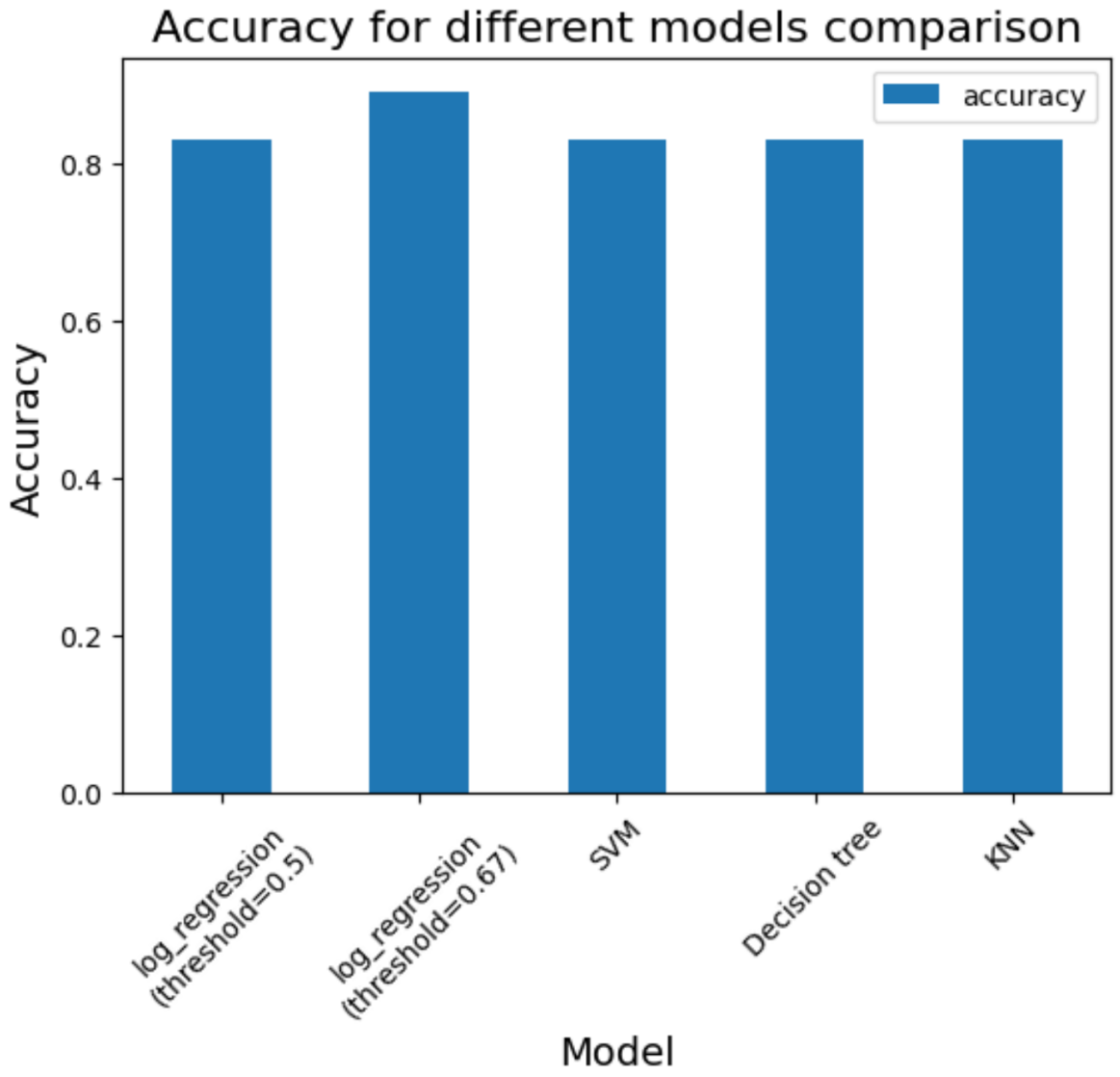
- **B4** boosters carried the **heaviest** payloads
- Payloads between **2k and 6k** were very **successful** (1=success, 0=failure)

Section 5

Predictive Analysis (Classification)

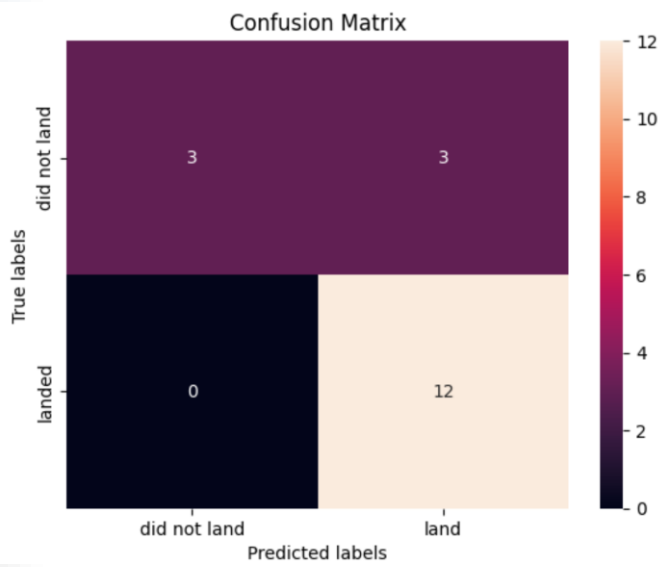
Classification Accuracy

Logistic Regression (threshold .67) has the highest accuracy while the rest are all about .80



Confusion Matrix

- 12 True Positives
- 3 False Positives
- 0 False Negatives
- 3 True Negatives



Conclusions

- Logistic Regression (threshold .67) has the highest accuracy while the rest are all about .80
- KSC LC-39A has the most success out of all launch sites (41.2%)
- The launch sites are very close to the coastline so the rocket will fall into the ocean if something goes wrong
- They are also close to the equator because the Earth's rotation improves fuel efficiency
- Success rate typically increases over time and more launches
- There is 100% success rate for ES_L1, GEO, HEO, SSO orbits

Thank you!