Winning Space Race with Data Science

Geoffrey Bowles
6/12/2023

Contents

Executive Summary

Summary of methodologies

- Collect data using SpaceX REST API and web scraping techniques
- Wrangle data to create success/failure outcome variable
- Explore data with data visualization techniques
- Analyze data with SQL
- Explore launch site success rates and proximities to geographical markers
- Visualize launch sites with most success and payload ranges
- Build models to predict landing outcomes

Exploratory Data Analysis

- Launch success has improved over time
- KSC LC-39A landing site has the highest success rate
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

Visualization/Analytics

- Most launch sites are near the equator and all are close to the coast

Predictive Analytics

- All models performed similarly
- The decision tree model slightly outperformed the rest using the .best_score_

Introduction

Background

SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

Explore

- How payload mass, launch shite, number of flights, and orbits affect landing success
- Rate of successful landings over time
- Best predictive model for successful landing

Section 1

Methodology

Methodology

Steps

- Data collection methodology
  - Rocket Launch Data from SpaceX API with Requests via Python
  - Falcon9 Launch Data from Wikipedia with BeautifulSoup via Python
- Perform data wrangling
  - Data was processed with one-hot encoding via Python
- Explore data using SQL and data visualization
- Visualize data with Folium and Plotly Dash
- Build predictive models using classification models to predict landing outcomes
  - How to build, tune, evaluate classification model

Data Collection

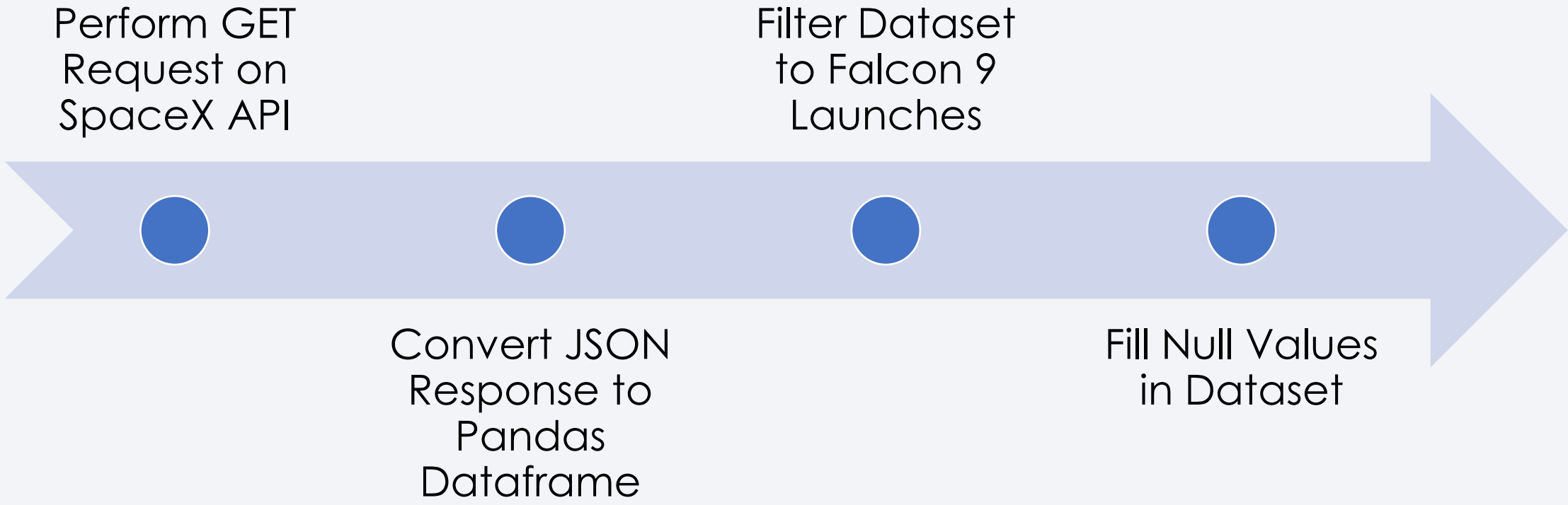
SpaceX API Data

- Pulled data from the API using the Requests module in Python
- Converted the JSON response into a Pandas dataframe using pd.json_normalize()
- Created subsets of data using lists
- Constructed a working dataframe from the subsets

Falcon9 Launch Data

- Web scraped data from Wikipedia using the BeautifulSoup module in Python
- Looped through all tables using soup.find_all()
- Parsed each table and stored all information into a custom dictionary
- Created a working dataframe using pd.from_dict()

Data Collection \ SpaceX API



GitHub Reference: https://github.com/omf-gbowles/ibm-ds-capstone/blob/master/01-spacex-data-collection-api.ipynb

Data Collection \ Web Scraping

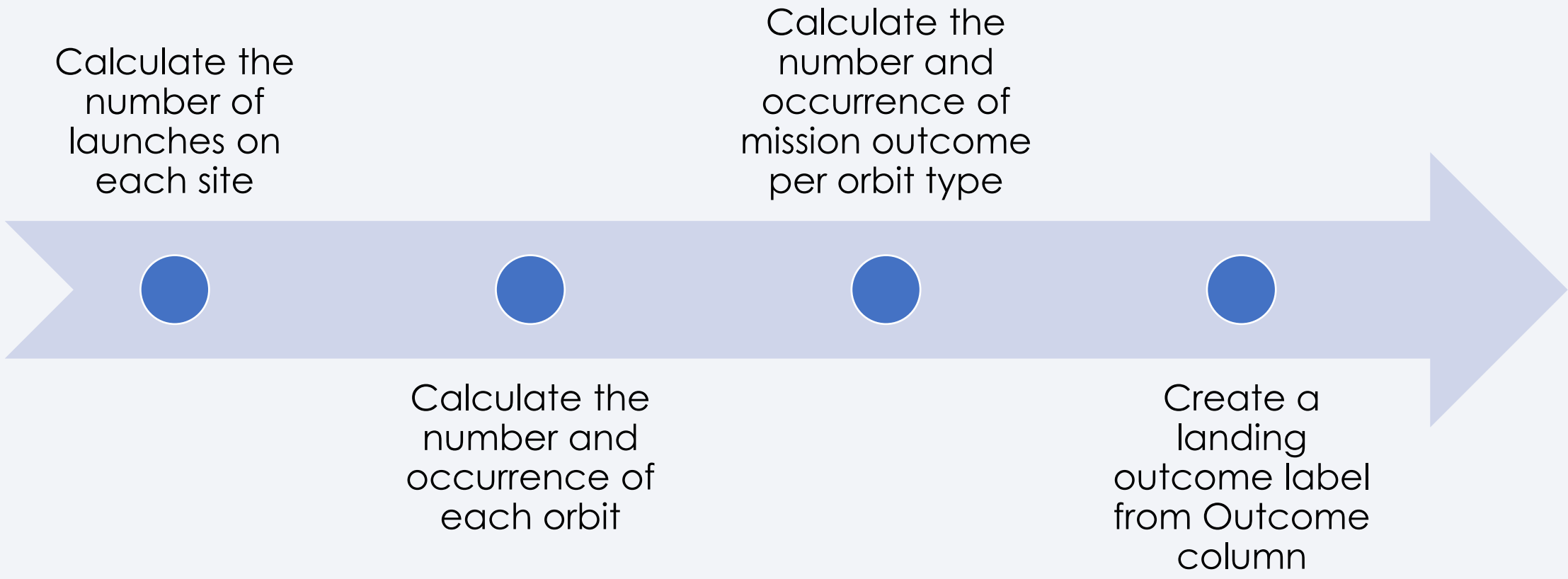
Perform GET request of Falcon9 Launch Wiki page

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

GitHub Reference: https://github.com/omf-gbowles/ibm-ds-capstone/blob/master/02-spacex-data-collection-webscraping.ipynb

Data Wrangling



GitHub Reference: https://github.com/omf-gbowles/ibm-ds-capstone/blob/master/03-spacex-data-wrangling.ipynb

EDA with SQL

Perform the following SQL queries to explore the data:

- Select the distinct launch sites in the space mission
- Select 5 records where the launch site begins with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing outcomes in drone ship, booster versions and launch sites for the months in 2015
- Rank the count of successful landing outcomes between 6/4/2010 and 3/20/2017

GitHub Reference: https://github.com/omf-gbowles/ibm-ds-capstone/blob/master/04-spacex-eda-sql.ipynb

EDA with Data Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit Type

Analysis

- View relationships by using scatter plots
- Show comparisons among discrete categories with bar charts

GitHub Reference: https://github.com/omf-gbowles/ibm-ds-capstone/blob/master/05-spacex-eda-data-viz.ipynb

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- Marked **all launch sites** on a folium map with **red circles** based on coordinates

Colored Markers of Launch Outcomes

- Assigned the feature launch outcomes (failure or success) to **class 0** (**failure**) **and 1** (**success**)
- Added **green** (**success**) and **red** (**failure**) markers to each site to show success rates

Distances Between a Launch Site to Proximities

- Added **colored lines** for calculated distances between launch sites and proximities such as **nearest coastline, railway, highway, and city**

GitHub Reference: https://github.com/omf-gbowles/ibm-ds-capstone/blob/master/06-spacex-folium.ipynb

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range

- Allow user to select payload mass range

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

- Allow user to see the correlation between Payload and Launch Success

GitHub Reference: https://github.com/omf-gbowles/ibm-ds-capstone/blob/master/07-spacex-dash.py

Predictive Analysis (Classification)

Charts

- **Create** NumPy array
- **Standardize** the data with StandardScaler
- **Split** the data using train_test_split
- **Create** a GridSearchCV object with cv=10
- **Apply** GridSearchCV on different algorithms:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K-Nearest Neighbor
- **Calculate** accuracy on the test data using .score()
- **Access** the confusion matrix for all models
- **Identify** the best model using Jaccard_Score, F1_Score and Accuracy

GitHub Reference: https://github.com/omf-gbowles/ibm-ds-capstone/blob/master/08-spacex-predictive-analysis.ipynb

Results Summary

Exploratory Data Analysis

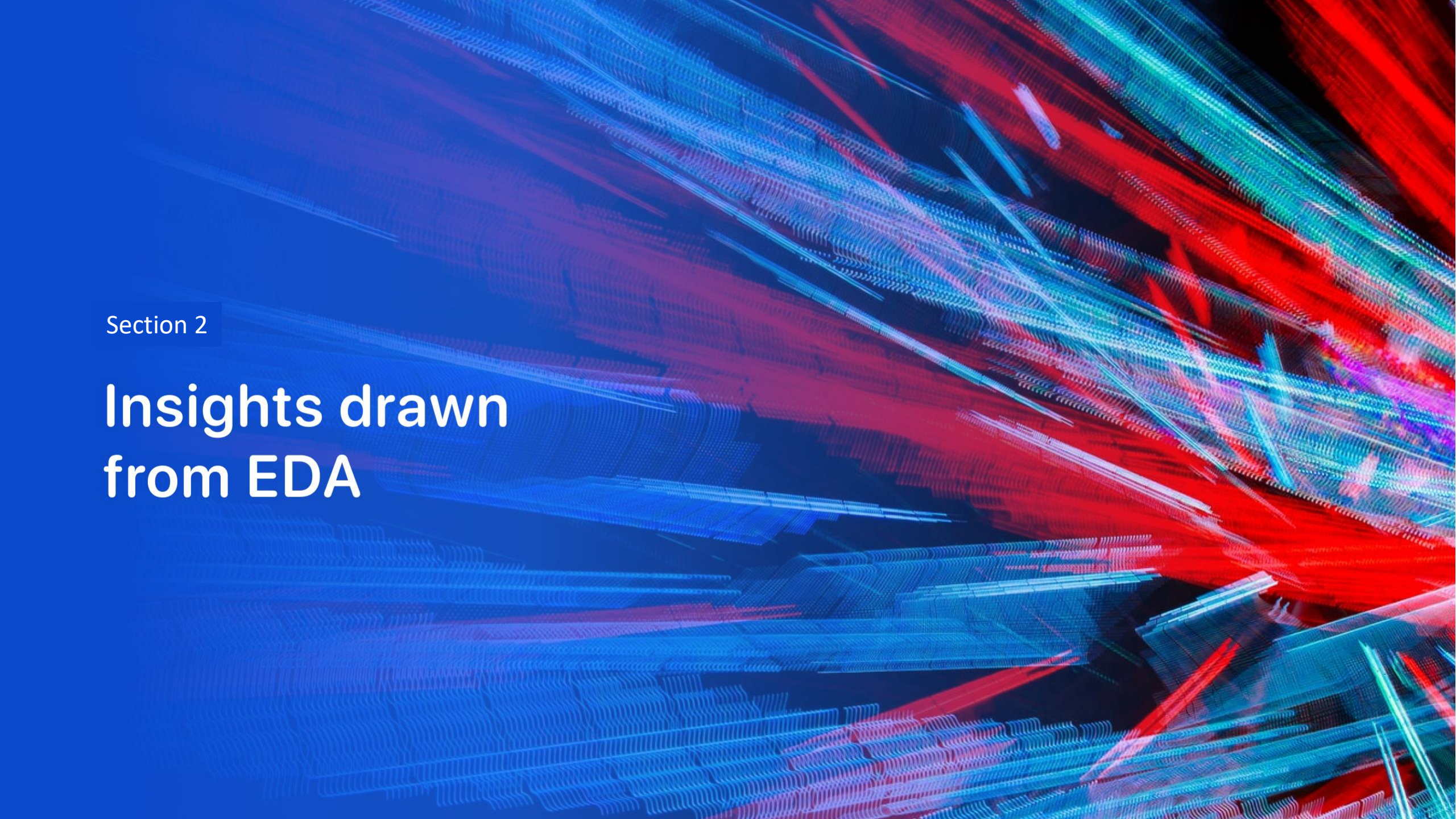
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

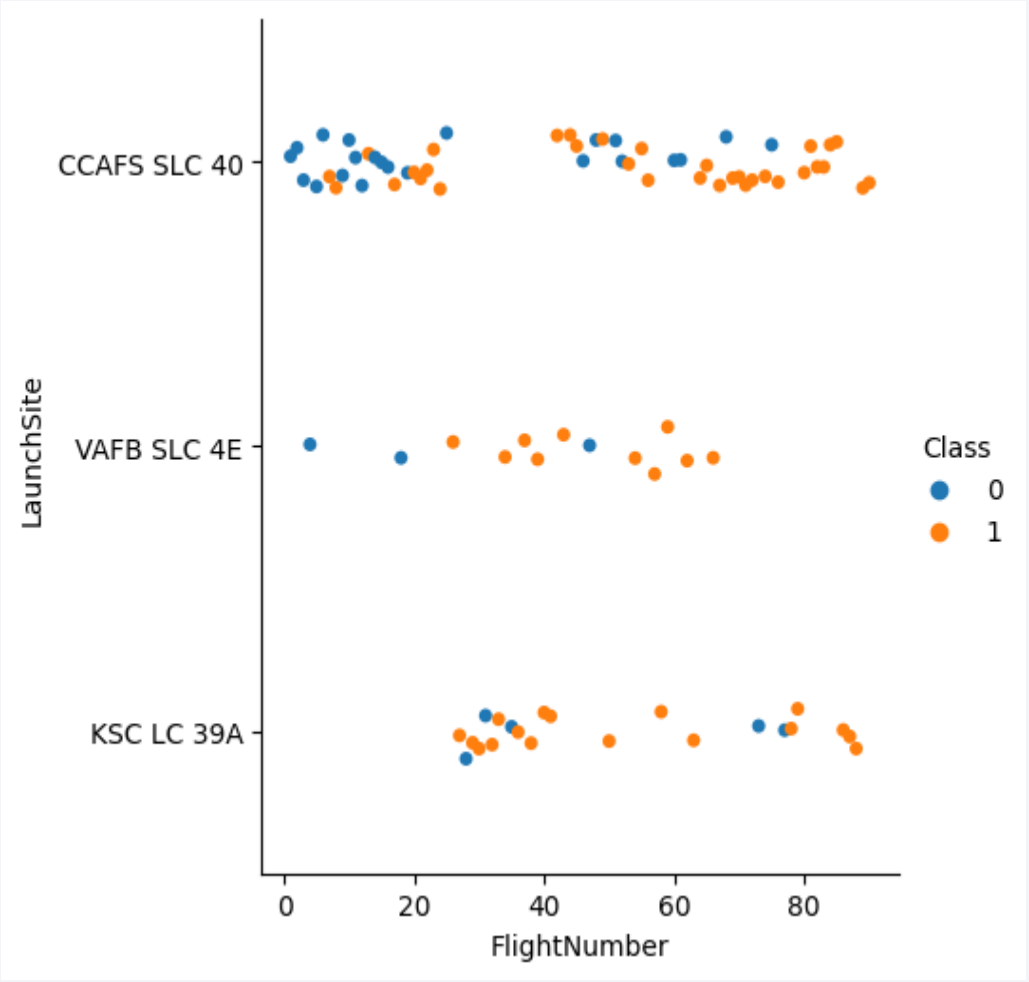
- Decision Tree model is the best predictive model for the dataset


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis

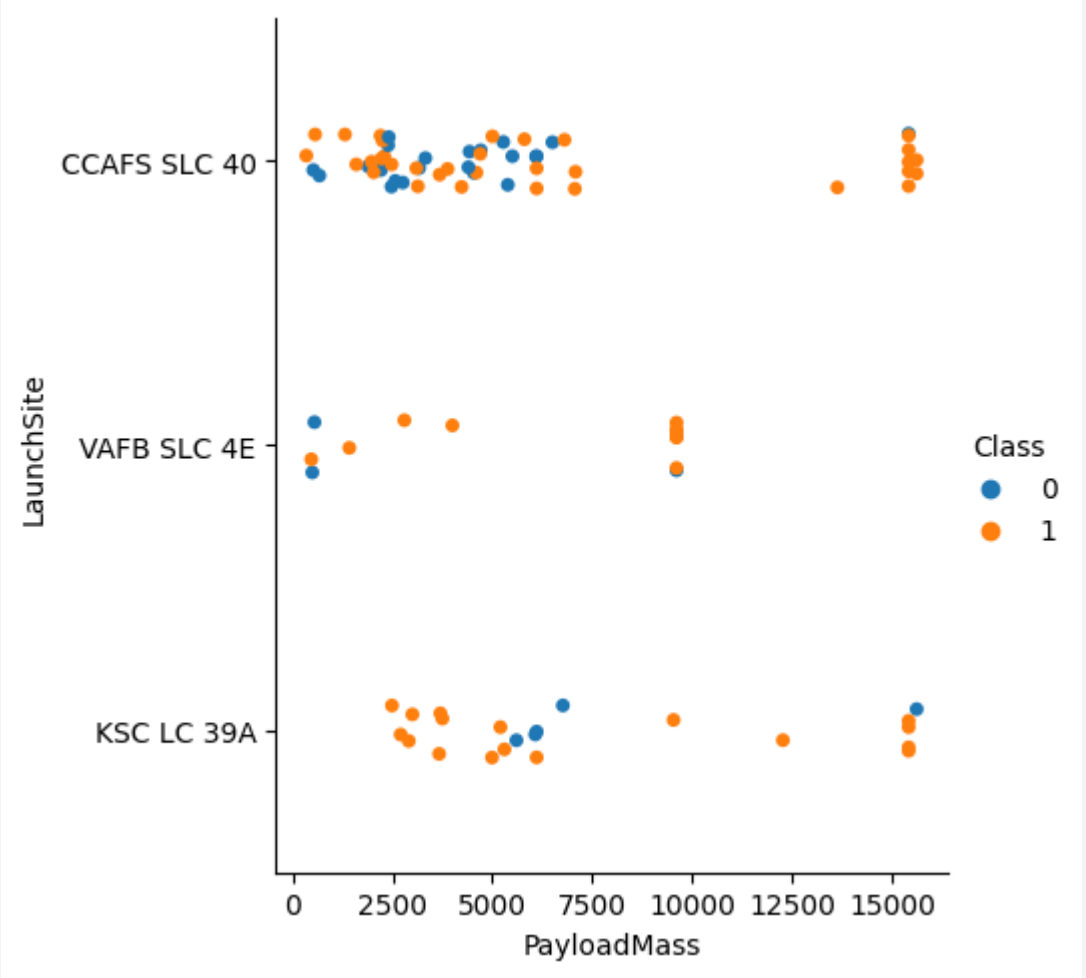
- Earlier flights had a lower success rate (**blue = fail**)
- Later flights had a higher success rate (**orange = success**)
- While CCAFS SLC 40 had more launches, the other two launch sites had higher success rates



Payload vs. Launch Site

Exploratory Data Analysis

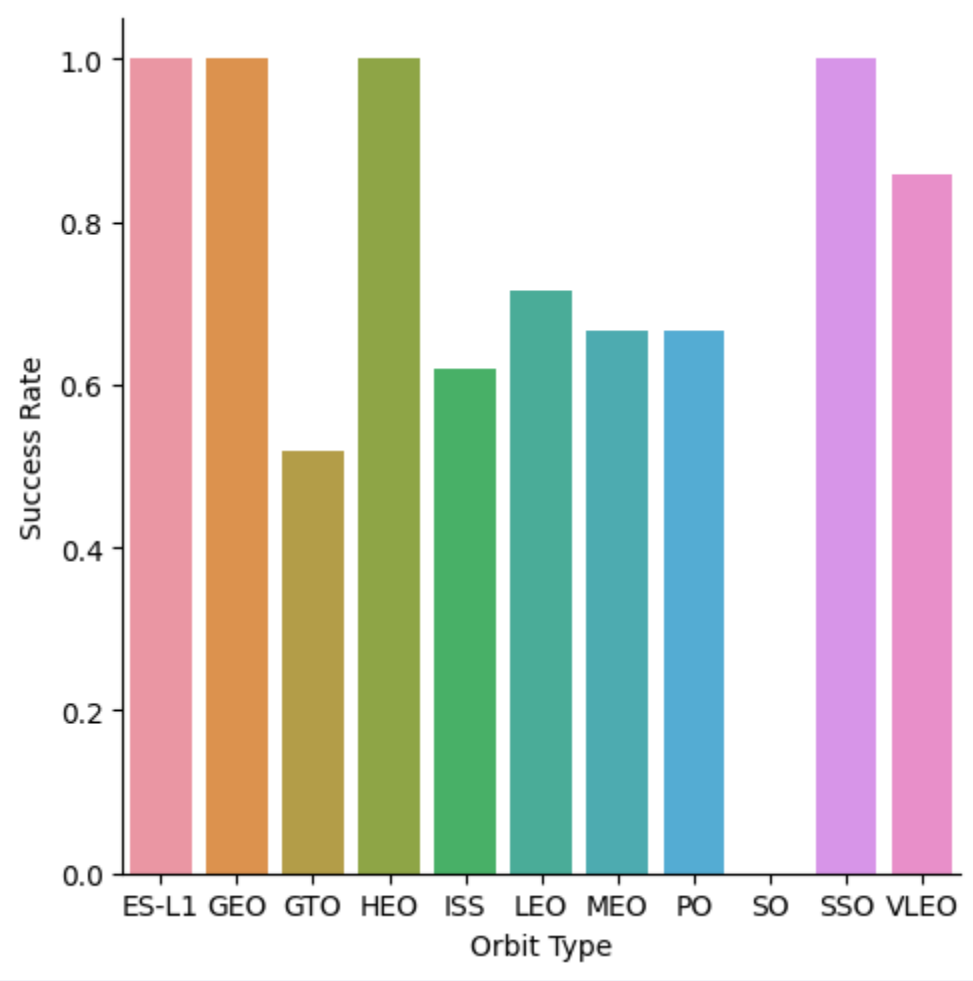
- **Greater payload mass** typically leads to **more successful launches**
- Launches with payload mass greater than ~7,500 kg were very successful
- KSC LC 39A has a 100% success rate for launches less than ~5,500 kg



Success Rate vs. Orbit Type

Exploratory Data Analysis

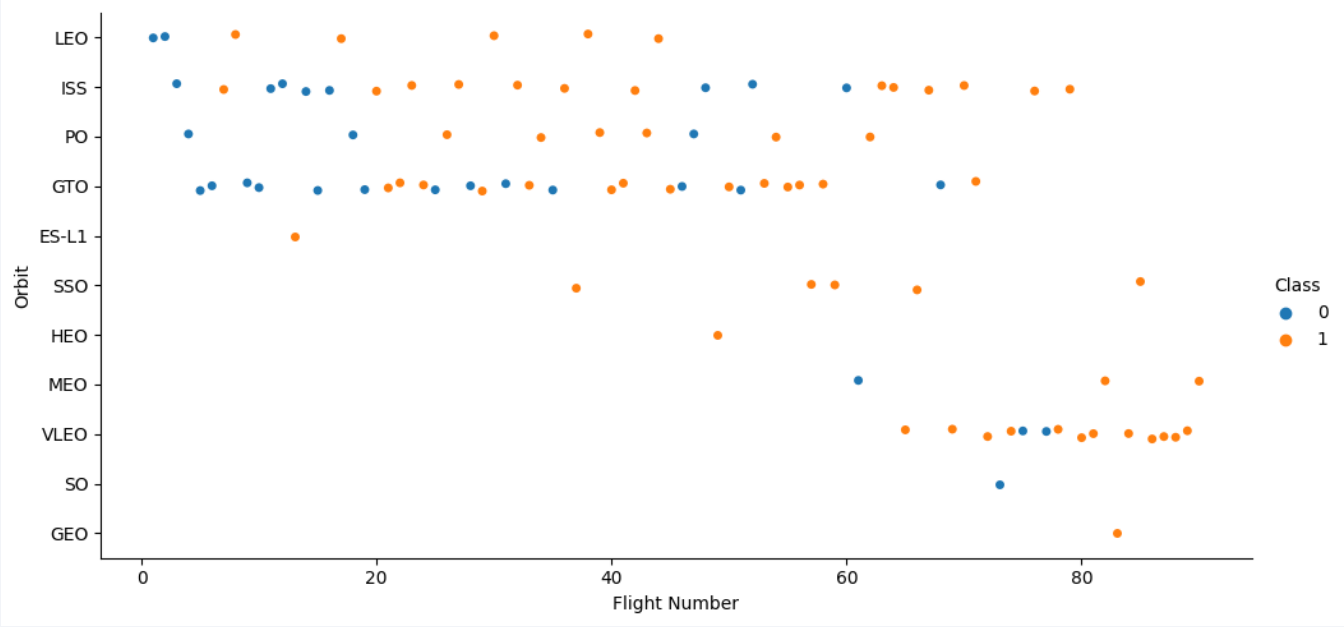
- **100% Success Rate**: ES-L1, GEO, HEO, SSO
- **50-80% Success Rate**: GTO, ISS, LEO, MEO, PO
- **0% Success Rate**: SO



Flight Number vs. Orbit Type

Exploratory Data Analysis

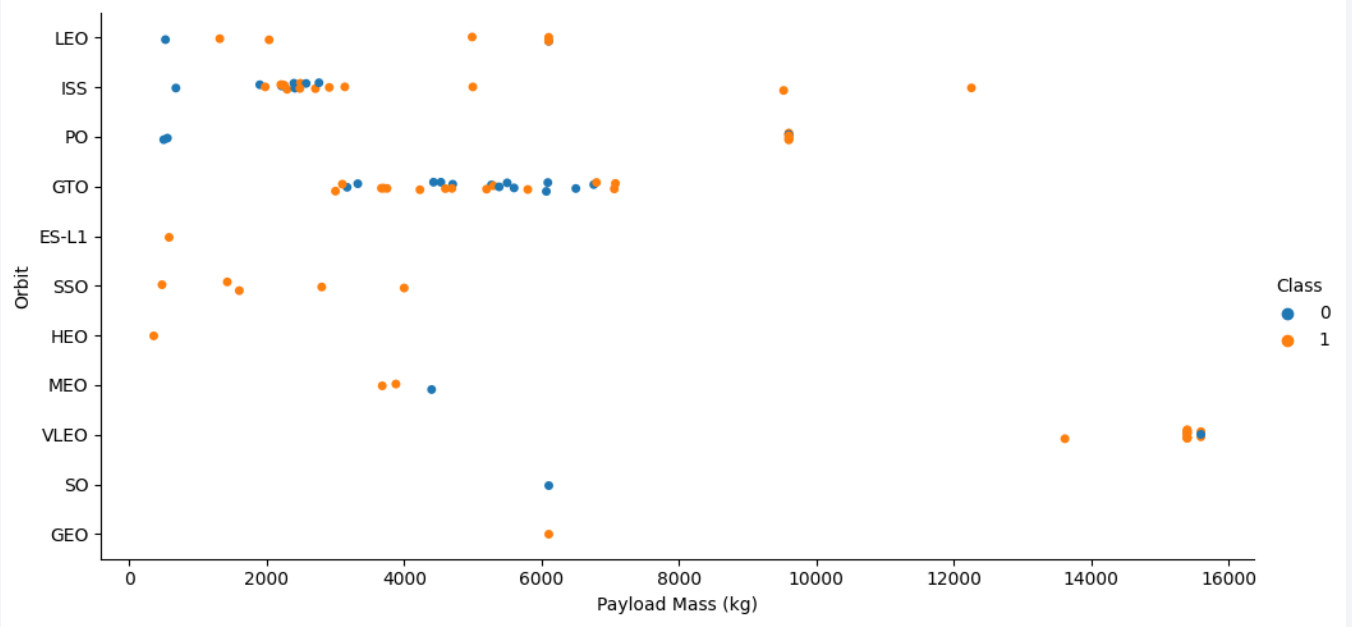
- The success rate typically increases with the number of launches for each orbit
- GTO is less likely to follow this pattern compared to other orbits



Payload vs. Orbit Type

Exploratory Data Analysis

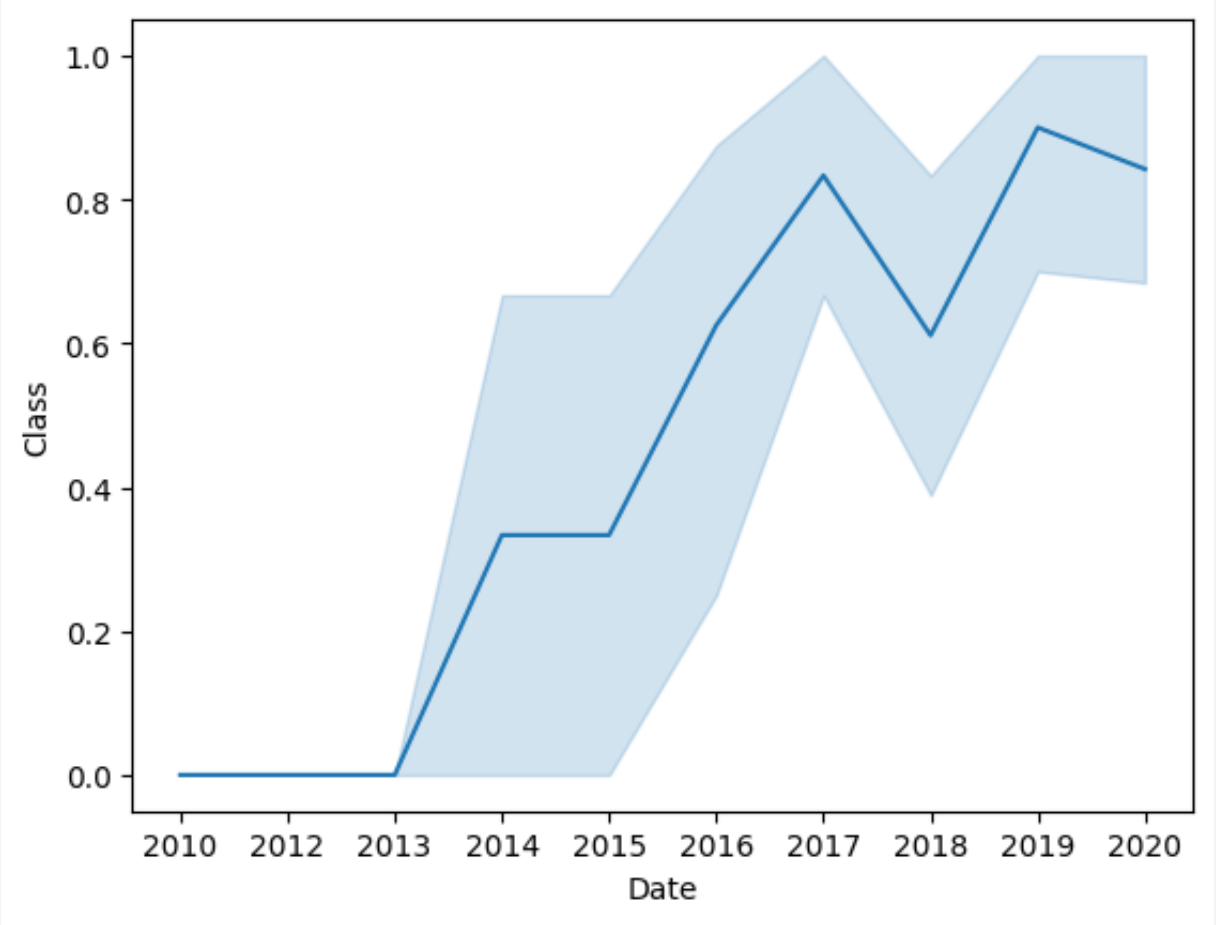
- LEO, ISS, and PO orbits perform better with heavier payloads
- GTO has mixed results amongst all launches with varying payloads



Launch Success Yearly Trend

Exploratory Data Analysis

- The success rate improved from 2013-2017 and again from 2018-2019
- The success rate decreased from 2017-2018 and again from 2019-2020
- Overall, the success rate has improved from 2013



Launch Site Information

Launch Site Names

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

```
%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

Records with Launch Site starting 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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Payload Mass

Total Payload Mass

- **45,596 kg** (total) carried by boosters launched by **NASA (CRS)**

```
%sql select sum(payload_mass__kg_) as total_payload_mass \
    from spacextbl \
    where customer = 'NASA (CRS)'
```

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

total_payload_mass
45596.0

Average Payload Mass

- **2,928.4 kg** (average) carried by booster version **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
2928.4

First Successful Ground Landing Date

First Successful Landing

- January 8, 2018

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

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

first_successful_landing
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

Successful Boosters

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

```
%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

Total Number of Successful and Failure Mission Outcomes

Mission Outcomes

- 1 Failure
- 100 Success

```
%sql select case \
        when mission_outcome like 'Failure%' then 'Failure' \
        when mission_outcome like 'Success%' then 'Success' end as outcomes, \
        count(*) as total_number \
    from spacextbl \
    group by outcomes
```

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

outcomes	total_number
None	898
Failure	1
Success	100

Boosters

Carrying Max Payload

- 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

```
%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 Failed Launch Records – Drone Ship

Showing month, date, booster version, launch site and landing outcome

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

```
%sql select substr(date, 4, 2) as month, date, booster_version, launch_site, landing_outcome \
    from spacextbl \
    where landing_outcome = 'Failure (drone ship)' \
        and substr(date, 7, 4) = '2015'
```

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

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

Landing Outcomes Between 6/4/2010 and 3/20/2017

Showing landing_outcome and count_outcomes in descending order

Landing_Outcome	Count_Outcomes
Success	20
Success (drone ship)	8
Success (ground pad)	7

```
%sql select landing_outcome, count(*) as count_outcomes \
    from spacextbl \
    where landing_outcome like 'Success%' \
        and date between '04-06-2010' and '20-03-2017' \
    group by landing_outcome \
    order by count_outcomes desc
```

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

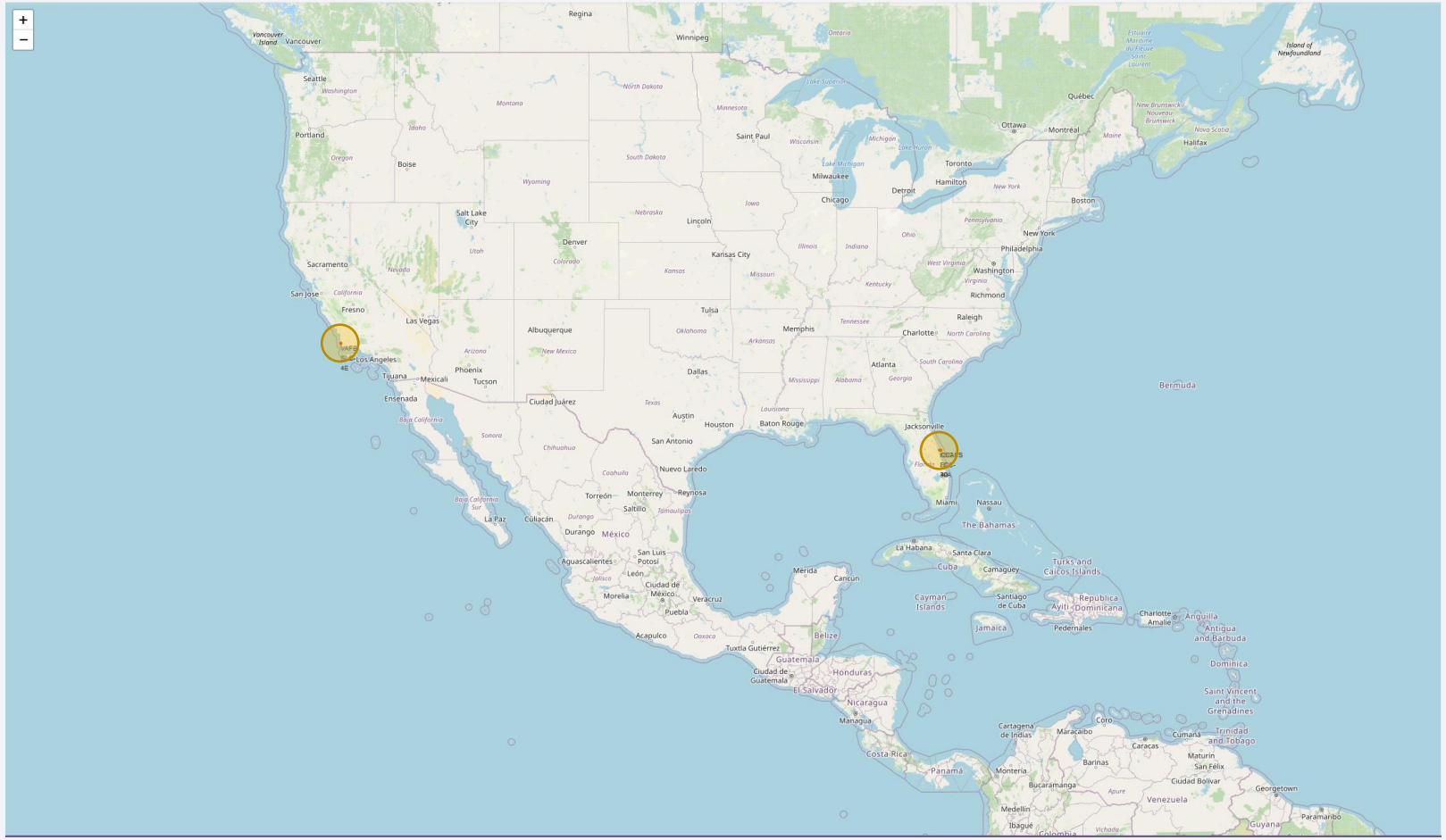
Landing_Outcome	count_outcomes
Success	20
Success (drone ship)	8
Success (ground pad)	7

Section 3

Launch Sites Proximities Analysis

Launch Sites

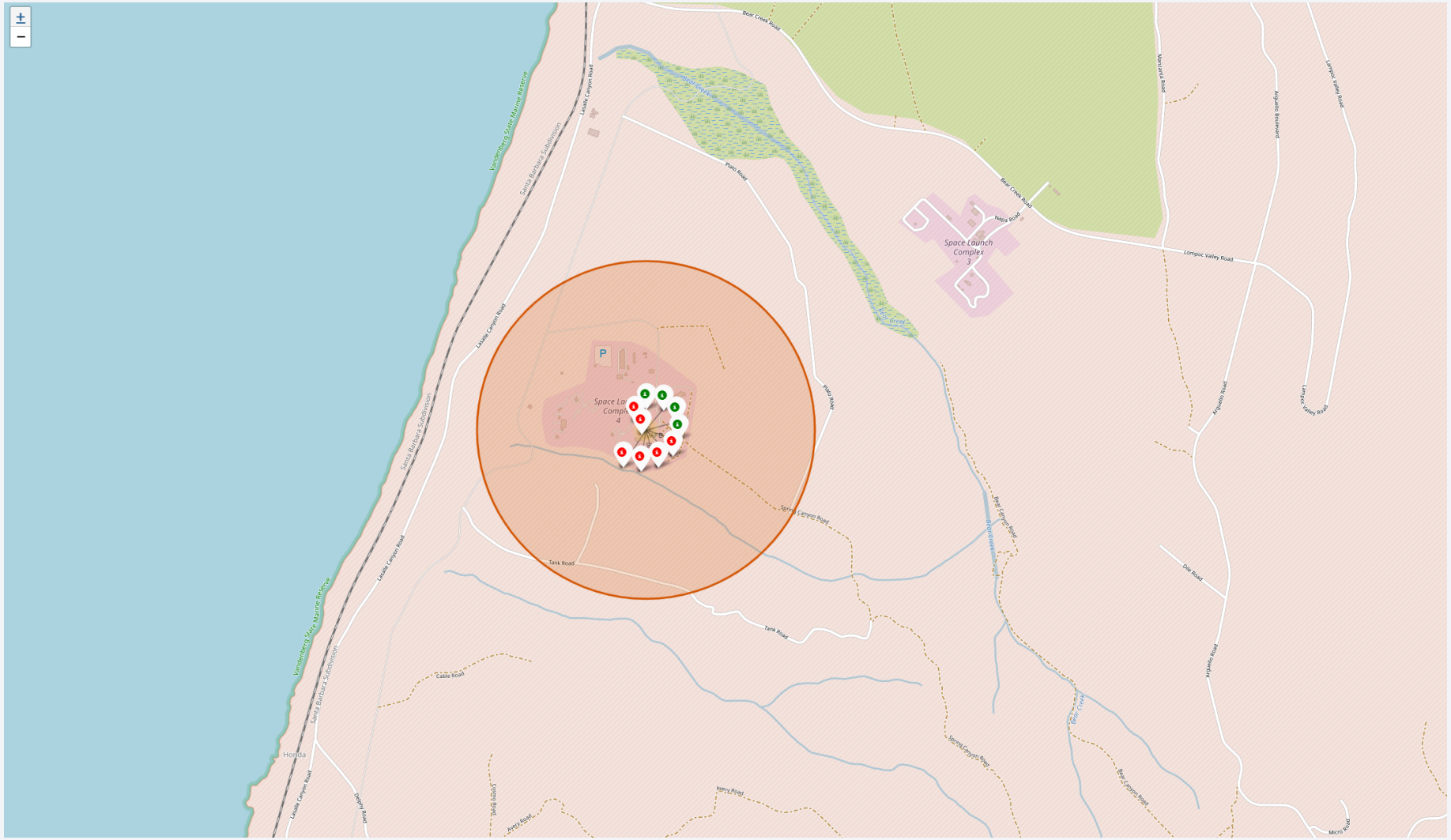
Sites **closer to the equator** gain an advantage when launching due to the **additional natural boost** of Earth's rotation – helping **save costs** of extra fuel and boosters



Launch Outcomes

Outcomes for each Launch Site:

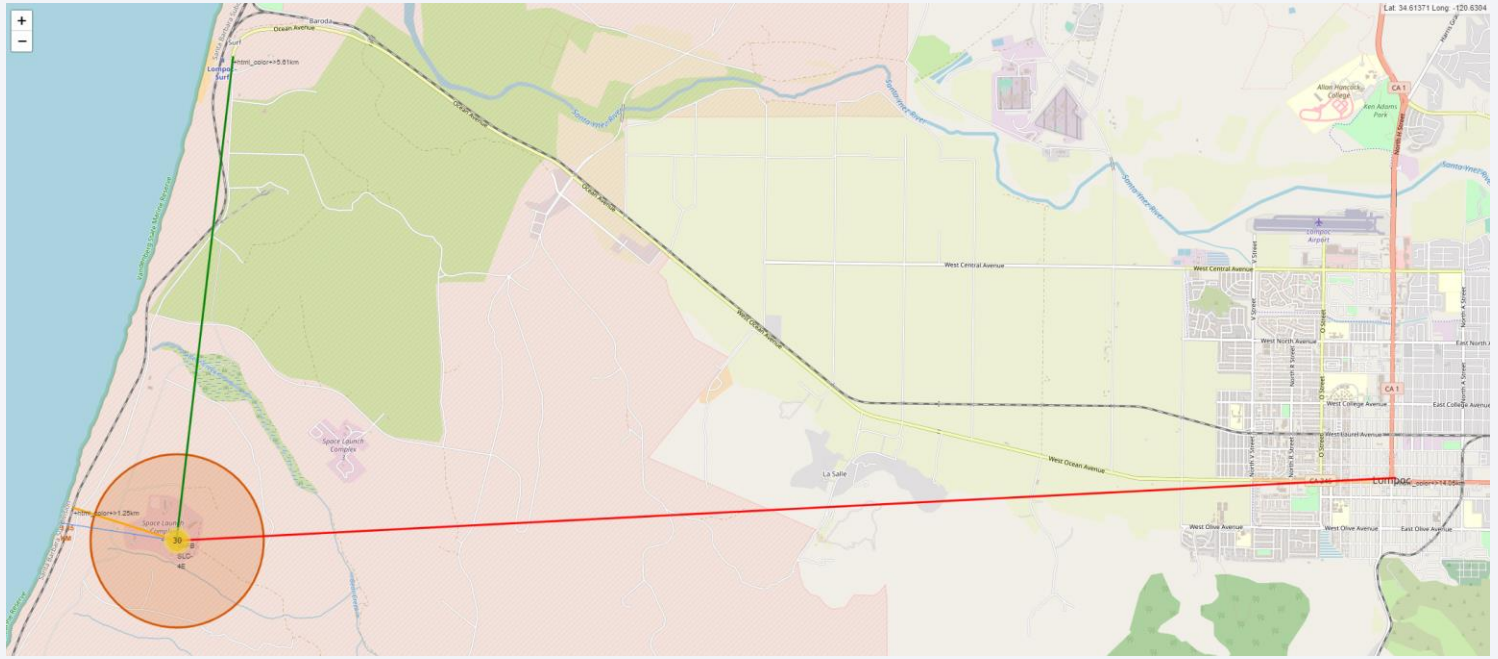
- **Green** markers for successful launches
- **Red** markers for failed launches
- Launch site **VAFB SLC-4E** has a **4/10 success rate (40%)**

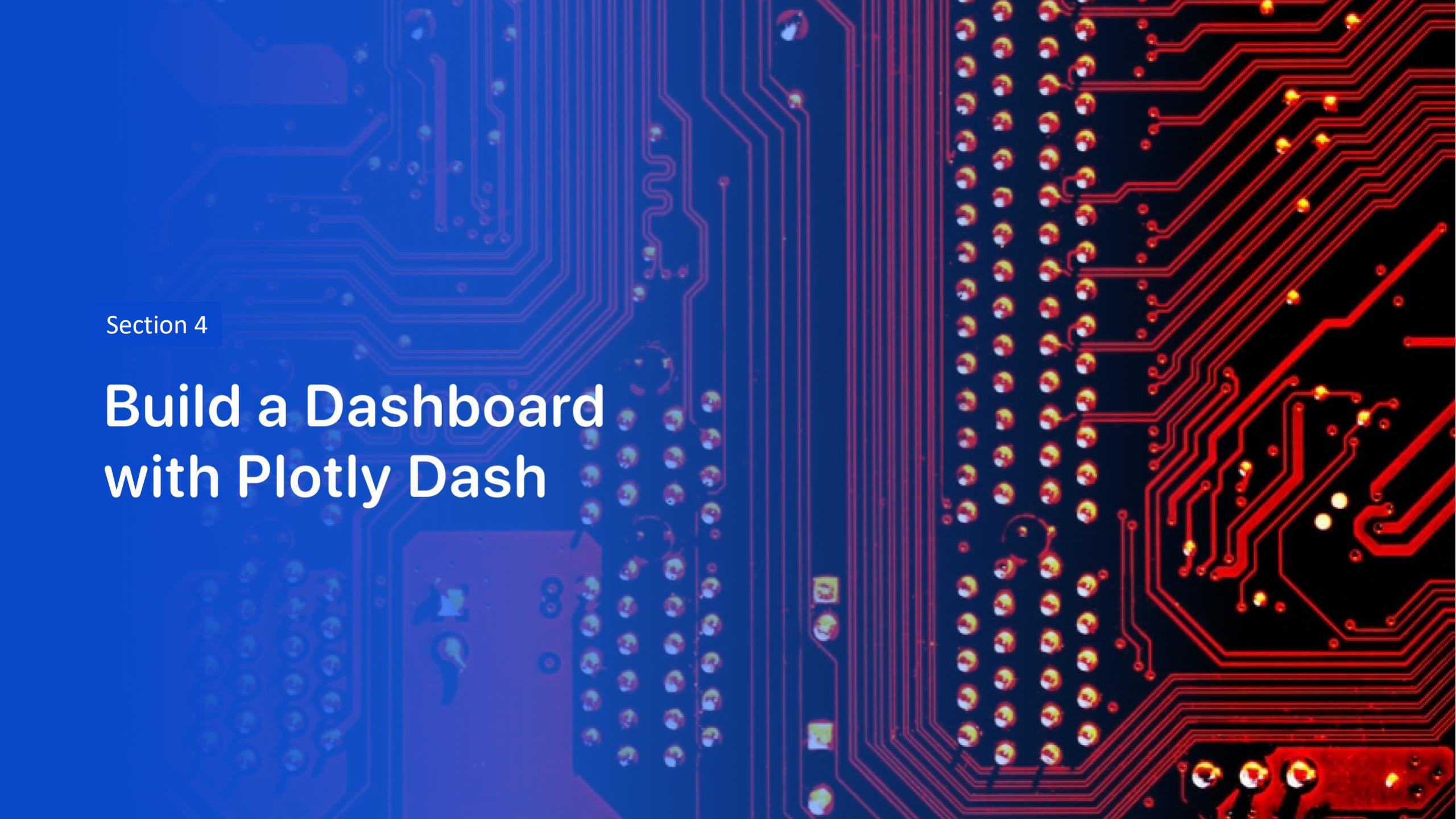


Distance to Proximities

VAFB SLC-4E

- **1.35 km** from the nearest coastline
- **1.25 km** from the nearest railway
- **14.05 km** from the nearest city
- **5.61 km** from the nearest highway

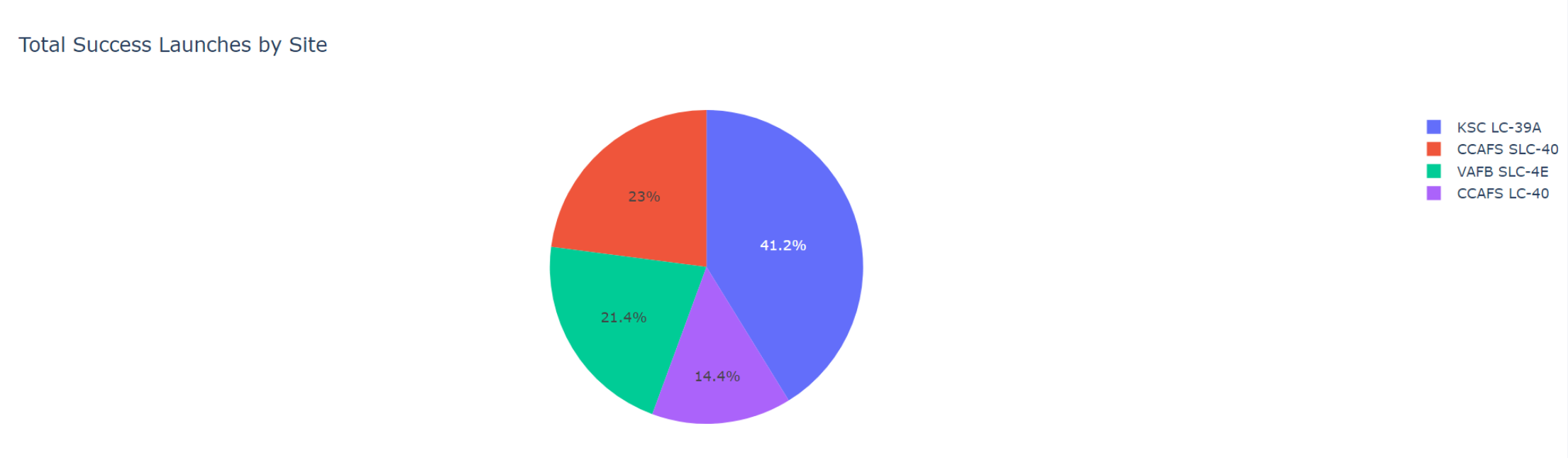



Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard

Success as Percent of Total

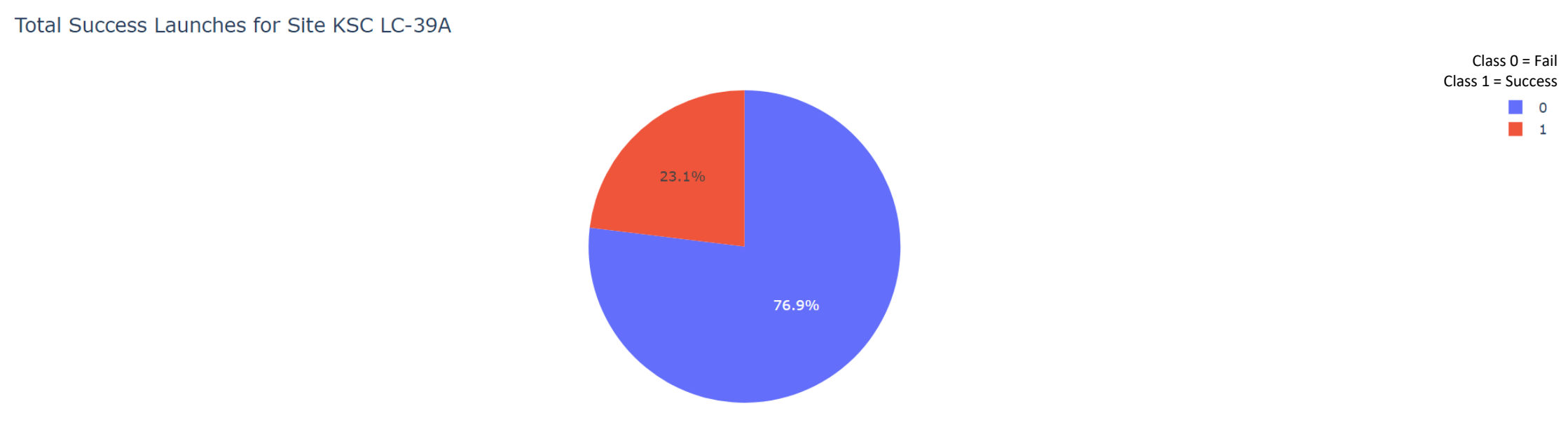
- **KSC LC-39A** has the **most** successful launches amongst launch sites **(41.2%)**
- **CCAFS LC-40** has the **least** successful launches amongst launch sites **(14.4%)**



SpaceX Most Successful Launch Site

Success as Percent of Total

- **KSC LC-39A** has the **highest success rate** amongst launch sites **(76.9%)**
- 10 successful launches and 3 failed launches



Payload Mass and Success

By Booster Version

- **Payloads between 2,000 kg** and **4,000 kg** have the **highest success rate**
- 1 indicates successful launches and 0 indicates failed launches

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Accuracy

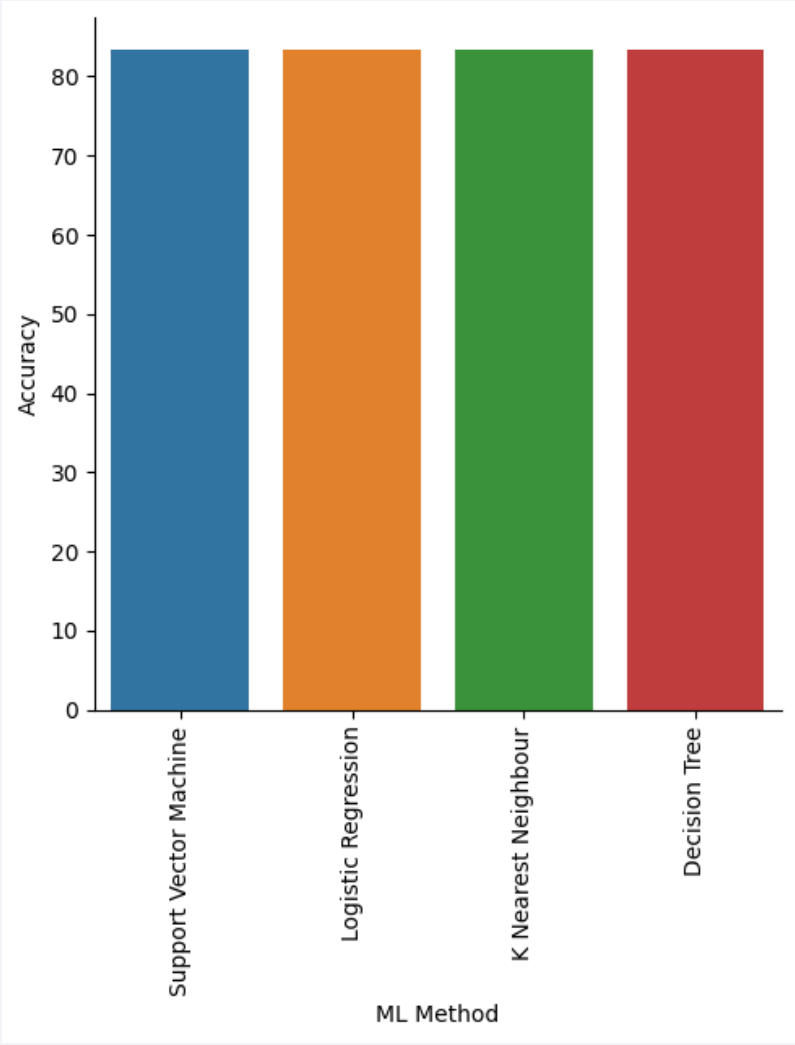
- In terms of score and accuracy, all models performed around the same level
- The decision tree classifier had the highest classification accuracy when looking at .best_score_

```
models = {'KNeighbors':knn_cv.best_score_,
          'DecisionTree':tree_cv.best_score_,
          'LogisticRegression':logreg_cv.best_score_,
          'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

✓ 0.0s

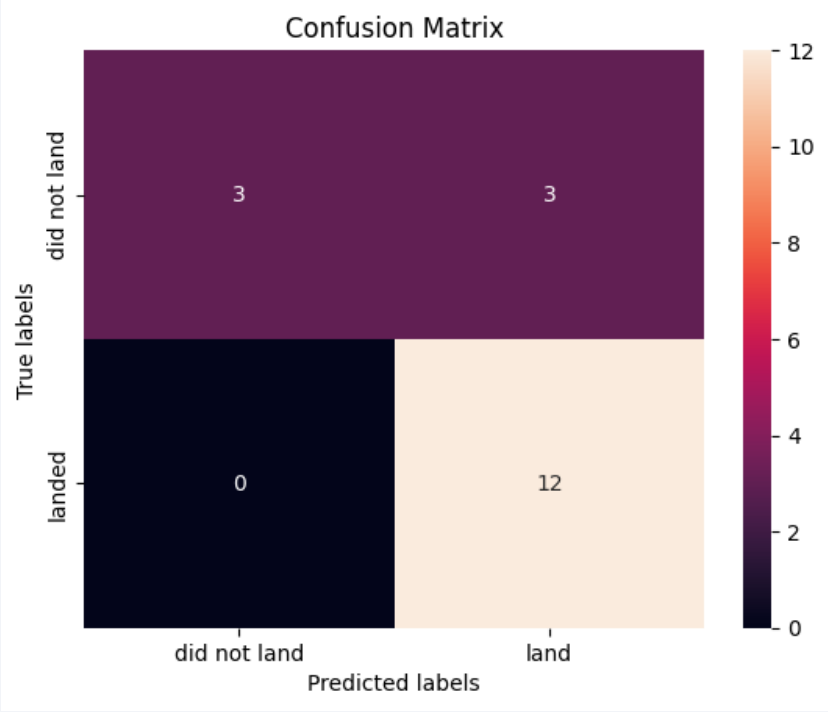
```
Best model is DecisionTree with a score of 0.8767857142857143
Best params is : {'criterion': 'gini', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'best'}
```



Confusion Matrix

Performance Summary

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical across the models
- Type 1 errors (False Positives) are not good results
- Confusion Matrix Outputs:
  - 12 True positive
  - 3 True negative
  - **3 False positive**
  - 0 False negative



Conclusions

Research

- Most launch sites are near the equator and close to a coastline
- Launch success rate improved from 2013 to 2020
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
- The higher the payload mass (kg), the higher the success rate
- The decision tree model slightly outperformed all other models

Thank you!