IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Francisco
8/14/24

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA – Exploratory data analysis with SQL and visualization
  - Interactive visual analytics with Folium and Plotly Dash
  - Predictive analysis using classification models
- Summary of all results
  - Exploratory data analysis results
  - Exploratory data analysis results
  - Predictive analysis results

Introduction

- **Project background and context**

  SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- **Problems to find answers**

  - Determine if Stage 1 will land successfully or not depending on variables such as payload mass, launch site, booster, etc.
  - Find out which feature has the most influence on the result.
  - Best algorithm for prediction.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Request and parse the SpaceX launch data using the GET request on API and web.
- Perform data wrangling
  - Missing values were replaced, and landing outcomes were converted to classes (0 or 1)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data Collection – SpaceX API

  First, we get the information through the Spacex API with the GET request method. Then we decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize().

- Data Collection - Scraping

  Here, we perform web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches. First, we use the GET request method and then the library BeautifulSoap was used to parse the html response. With the parse we obtained the desired table.

Data Collection – SpaceX API

GitHub URL of the completed SpaceX API calls notebook

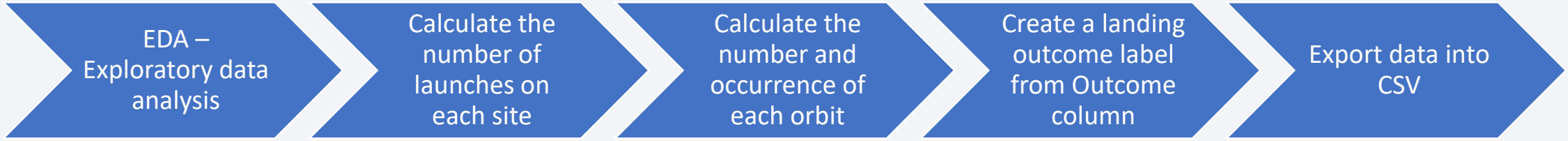
Data Collection - Scraping

GitHub URL of the completed web scraping notebook

Data Wrangling

We performed an exploratory data analysis (EDA) and calculated some important data.

Then, we created a CSV file with the selected data.



GitHub URL of the completed data wrangling notebook

EDA with Data Visualization

We plotted the followings types of charts:

- **Catplot:** to see how the FlightNumber (indicating the continuous launch attempts.) and Payload/Launch site variables would affect the launch outcome. Also, to visualize the relationship between “Payload Mass and Launch Site”, “FlightNumber and Orbit type” and “Payload Mass and Orbit type”.
- **Barplot:** to visualize the relationship between success rate of each orbit type.
- **Line chart:** to visualize the launch success yearly trend.

GitHub URL of the completed EDA with data visualization notebook

EDA with SQL

The SQL queries were:

- Display the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin 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 succesful 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. Use a subquery.
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- 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.

GitHub URL of the completed EDA with SQL notebook

Build an Interactive Map with Folium

Objects added to the folium map:

- Circle and marker at Launch sites coordinates with a popup label showing their names.
- Markers clusters.
- Lines to indicate distance between different places.
- Markers to indicate the outcome of the landing.

GitHub URL of the completed interactive map with Folium map notebook

Build a Dashboard with Plotly Dash

Plots and interactions used in the dashboard:

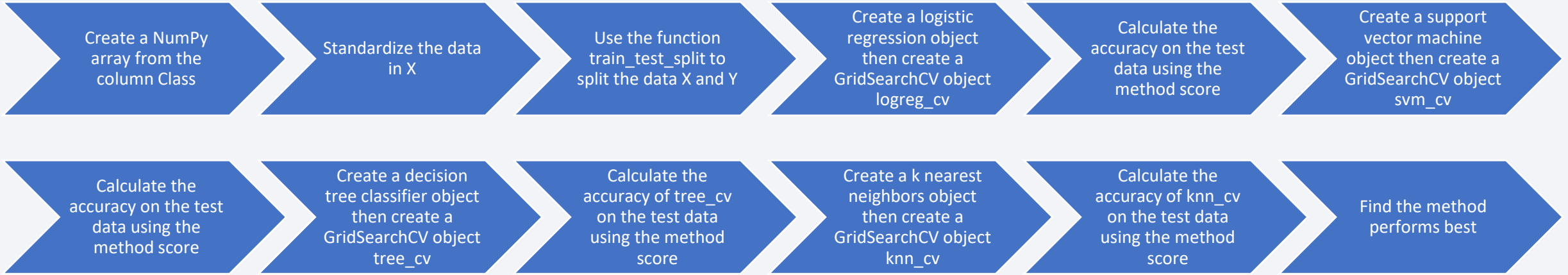
- Dropdown list to select a launch site.
- Pie plot to compare the success of the mission.
- Scatter plot: Outcome of the mission vs Payload mass.
- Range slider for the scatter plot to change the Payload mass range.

GitHub URL of the completed interactive Plotly Dash lab notebook

Predictive Analysis (Classification)

First, we loaded the dataframe and then we standardized the data. Then, we apply each method of classification using GridSearchCV.

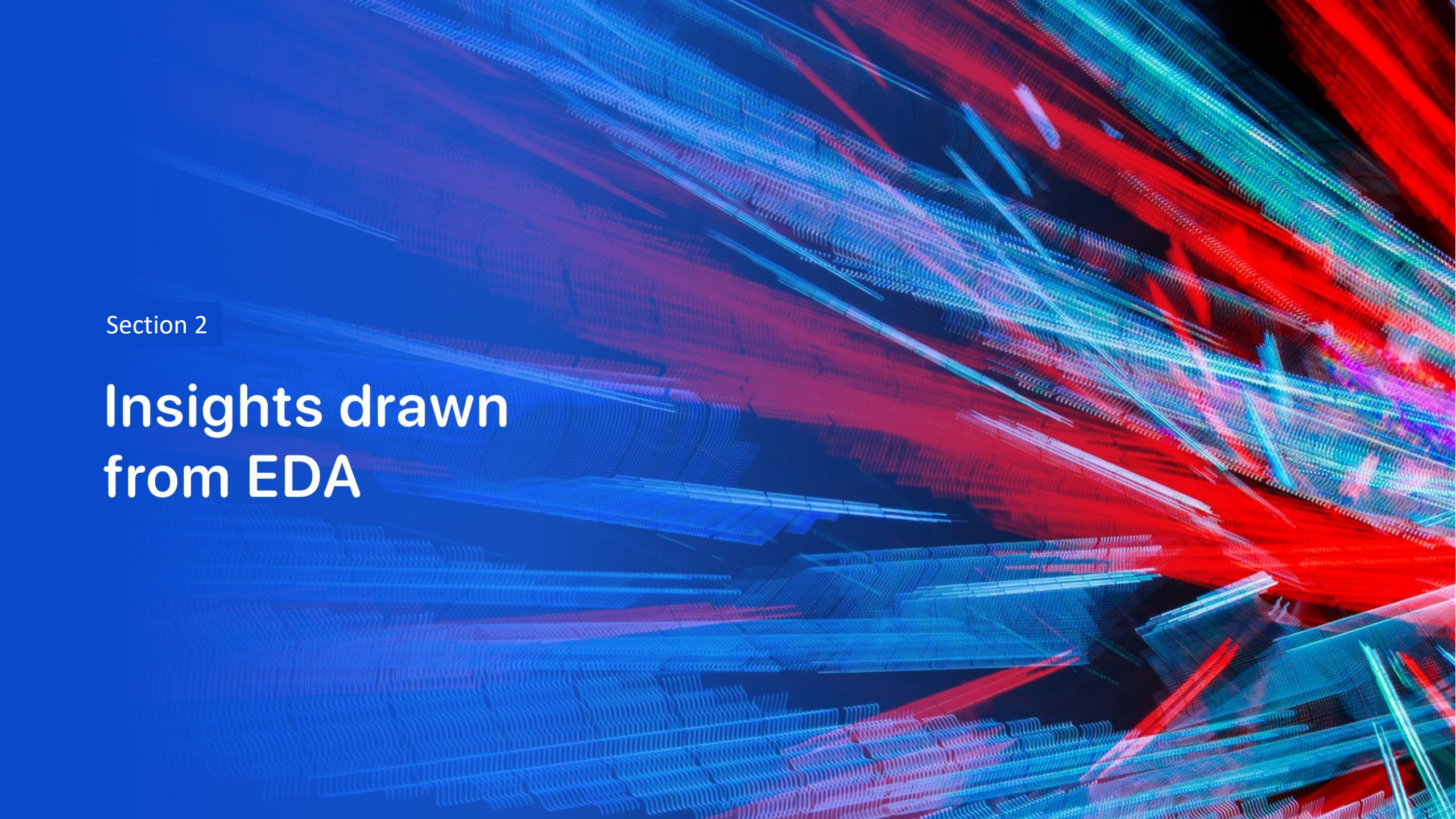
Finally, we compare each method score to find the best performing classification model.



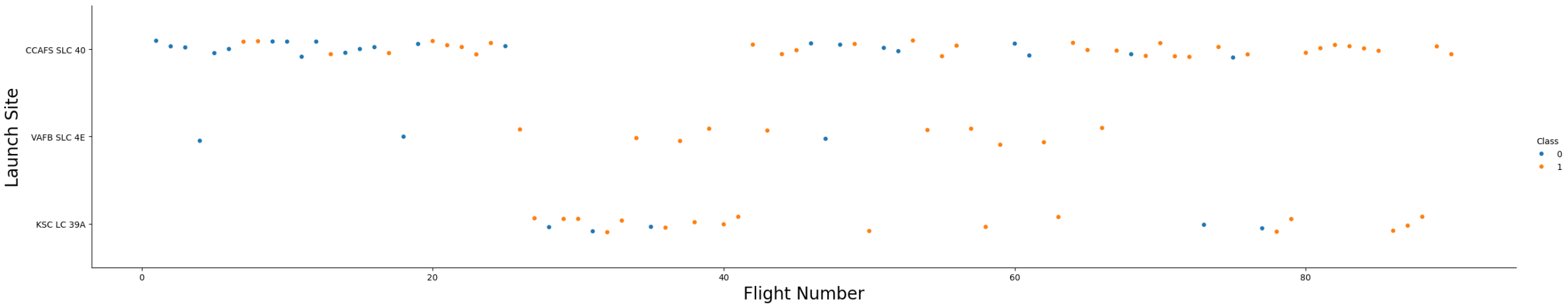
GitHub URL of the completed predictive analysis lab notebook

Results

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

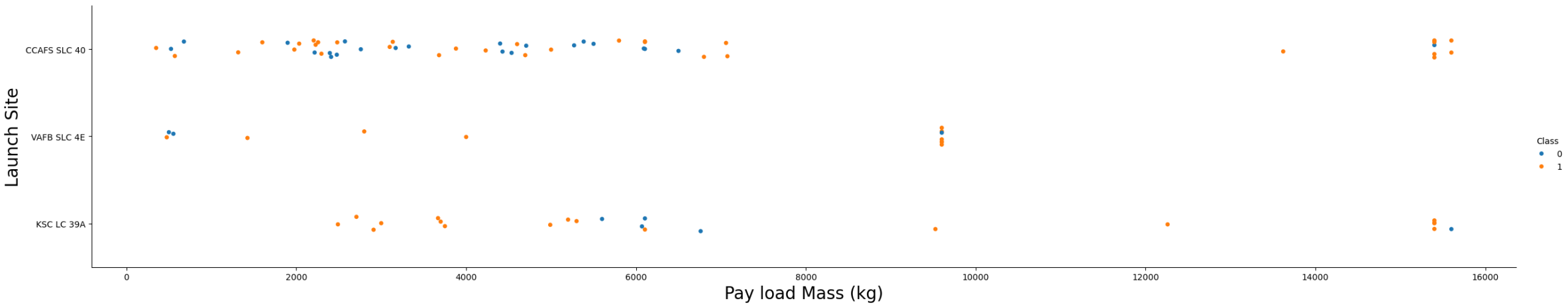

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



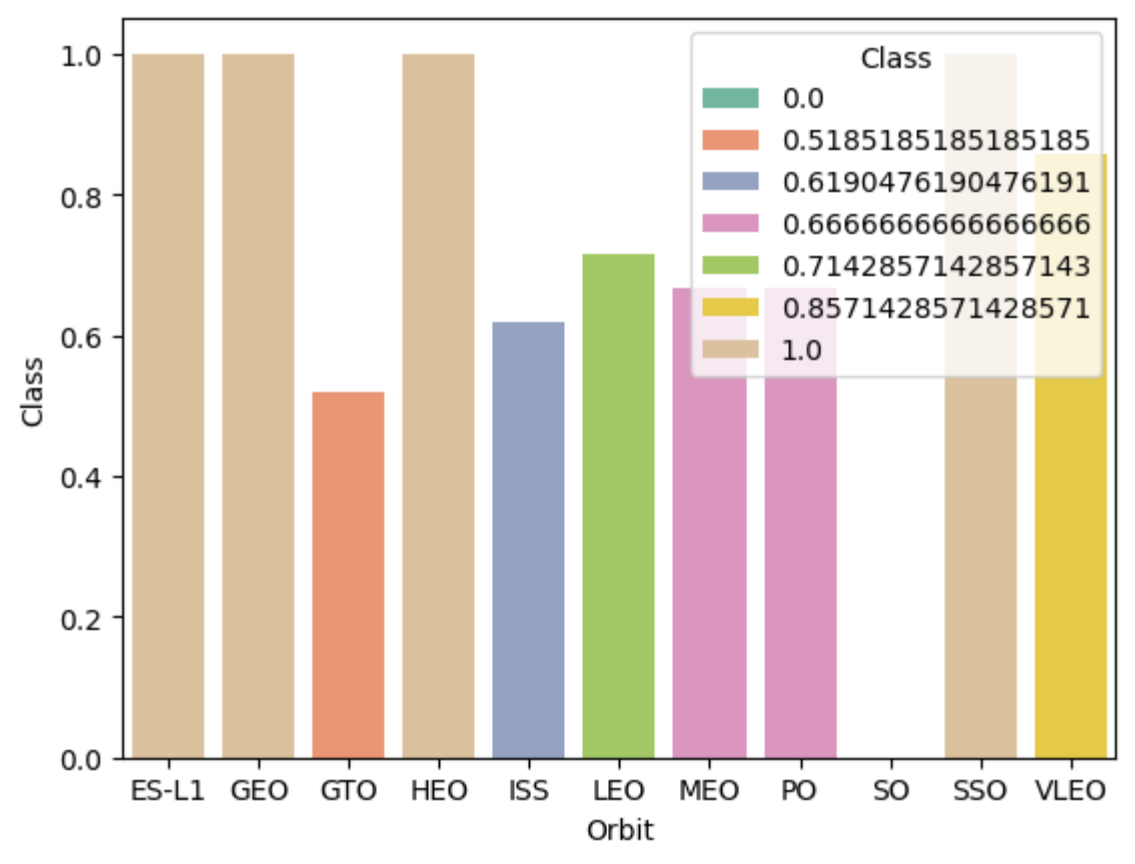
We see that as the flight number increases, the first stage is more likely to land successfully at all launch sites.

Payload vs. Launch Site



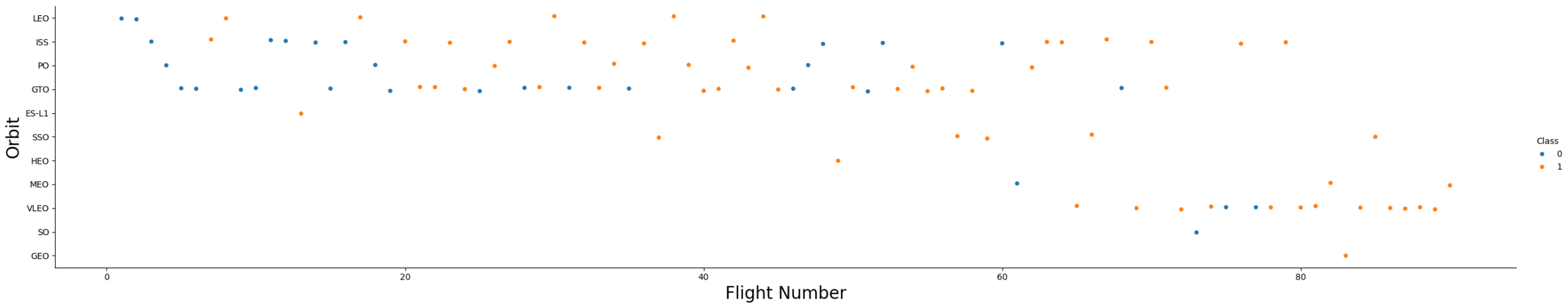
Now if we observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type



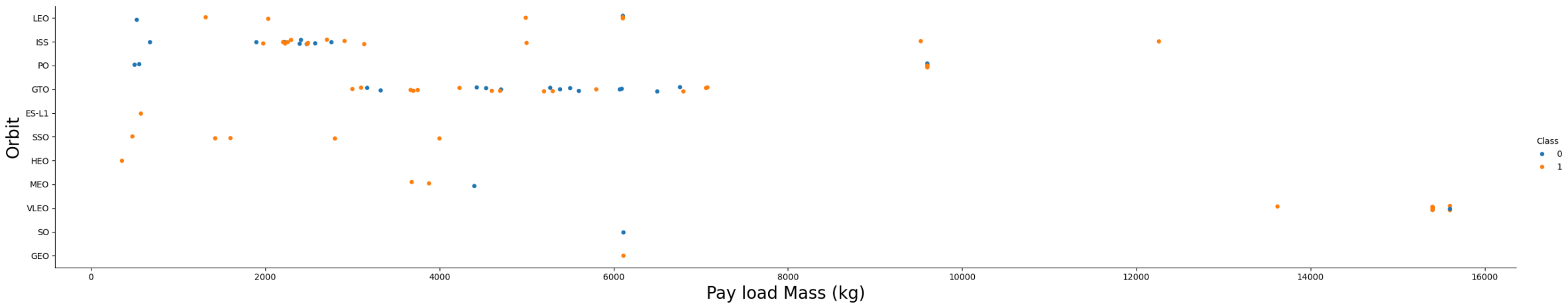
ES-L1, GEO, HEO and SSO have the most successful rate

Flight Number vs. Orbit Type



We can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success

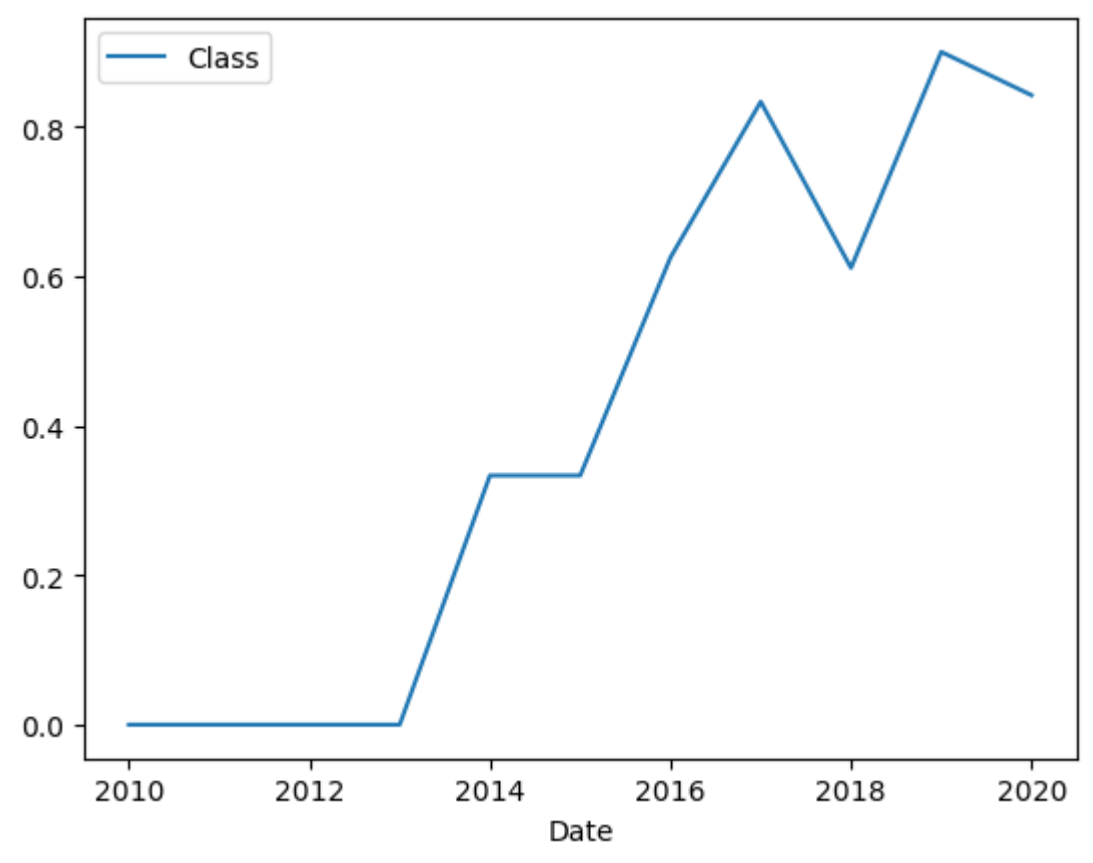
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more 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



You can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

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

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

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	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

Display 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

```
%sql SELECT sum(PAYLOAD_MASS__KG_) AS Total_Payloadmass_nasa FROM SPACEXTABLE WHERE Customer LIKE "NASA (CRS)";
```

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

Total_Payloadmass_nasa
45596

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

Average Payload Mass by F9 v1.1

```
%sql SELECT avg(PAYLOAD_MASS__KG_) AS AVG_Payloadmass_F9V1_1 FROM SPACEXTABLE WHERE Booster_Version LIKE "F9 v1.1";
```

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

AVG_Payloadmass_F9V1_1
2928.4

Display 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

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000) AND Landing_Outcome='Success (
```

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

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome,COUNT(*) FROM SPACEXTABLE group by Mission_Outcome;
```

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

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

List the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

```
 * 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

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

2015 Launch Records

```
%sql SELECT substr(Date, 6,2) AS Month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome
```

```
 * 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)

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

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

```
%sql SELECT Landing_Outcome, COUNT(*) AS Total FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' group by La
```

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

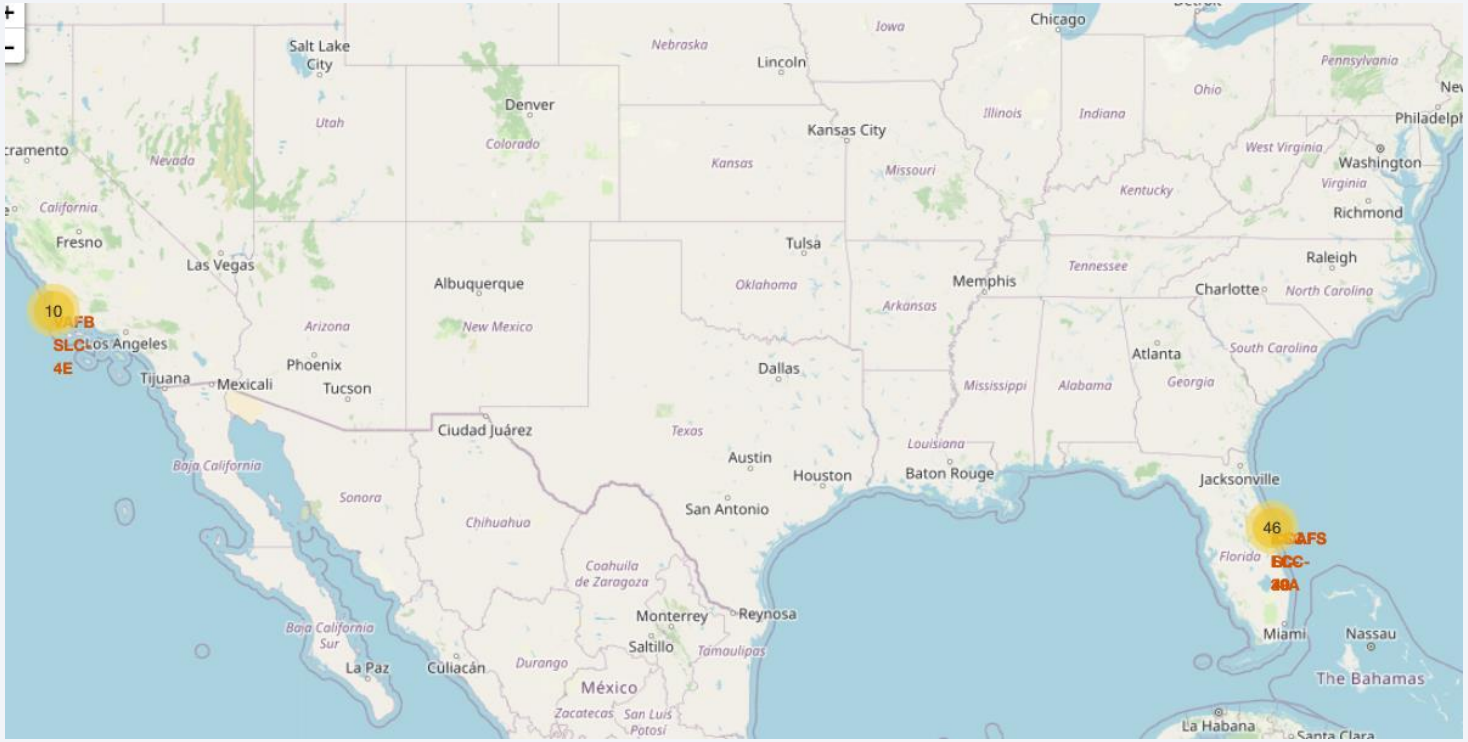
Landing_Outcome	Total
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

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.

Section 3

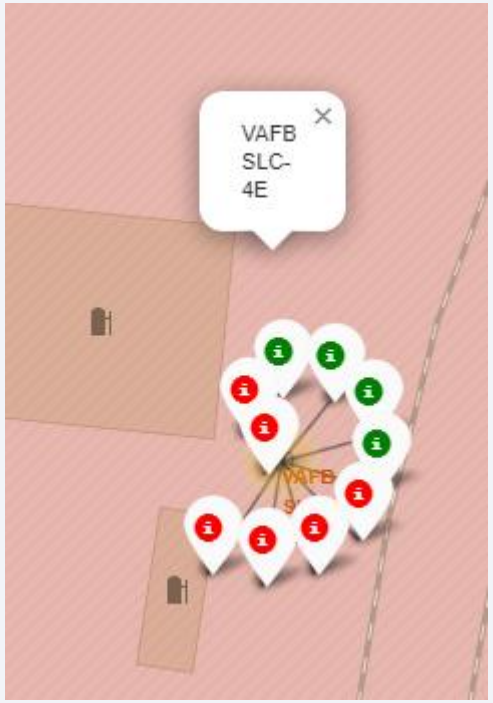
Launch Sites Proximities Analysis

Global map and launch sites



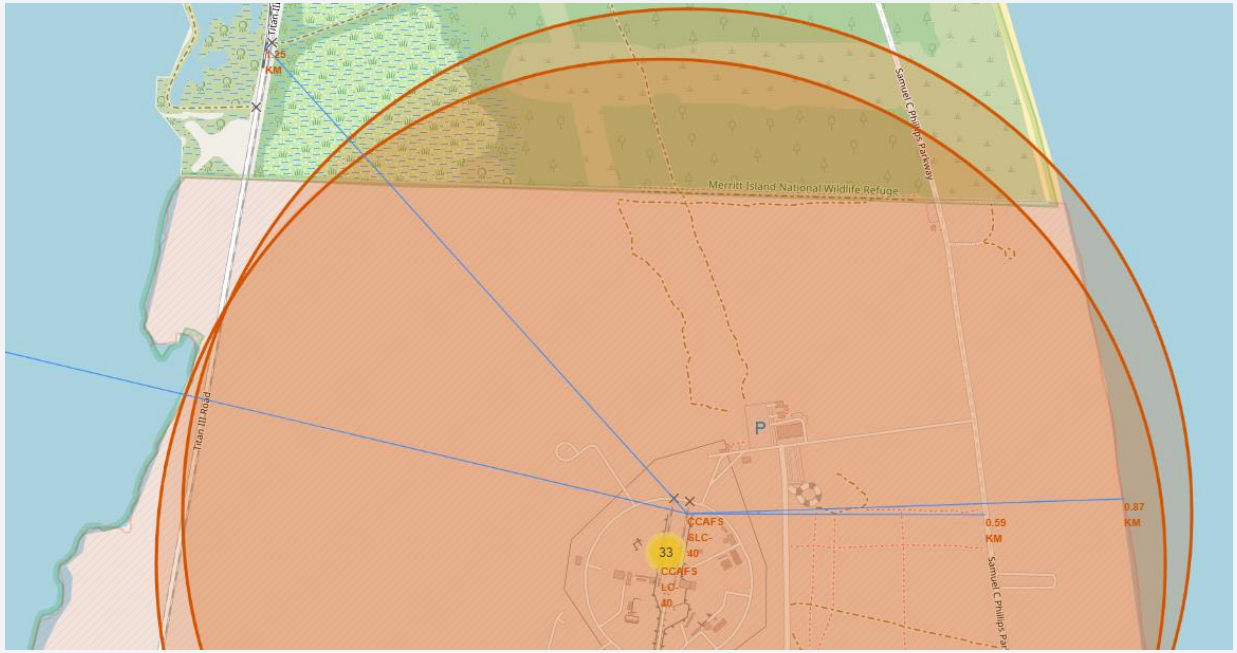
The most launches were made at the east coast.

Outcomes at VAFB SLC-4E



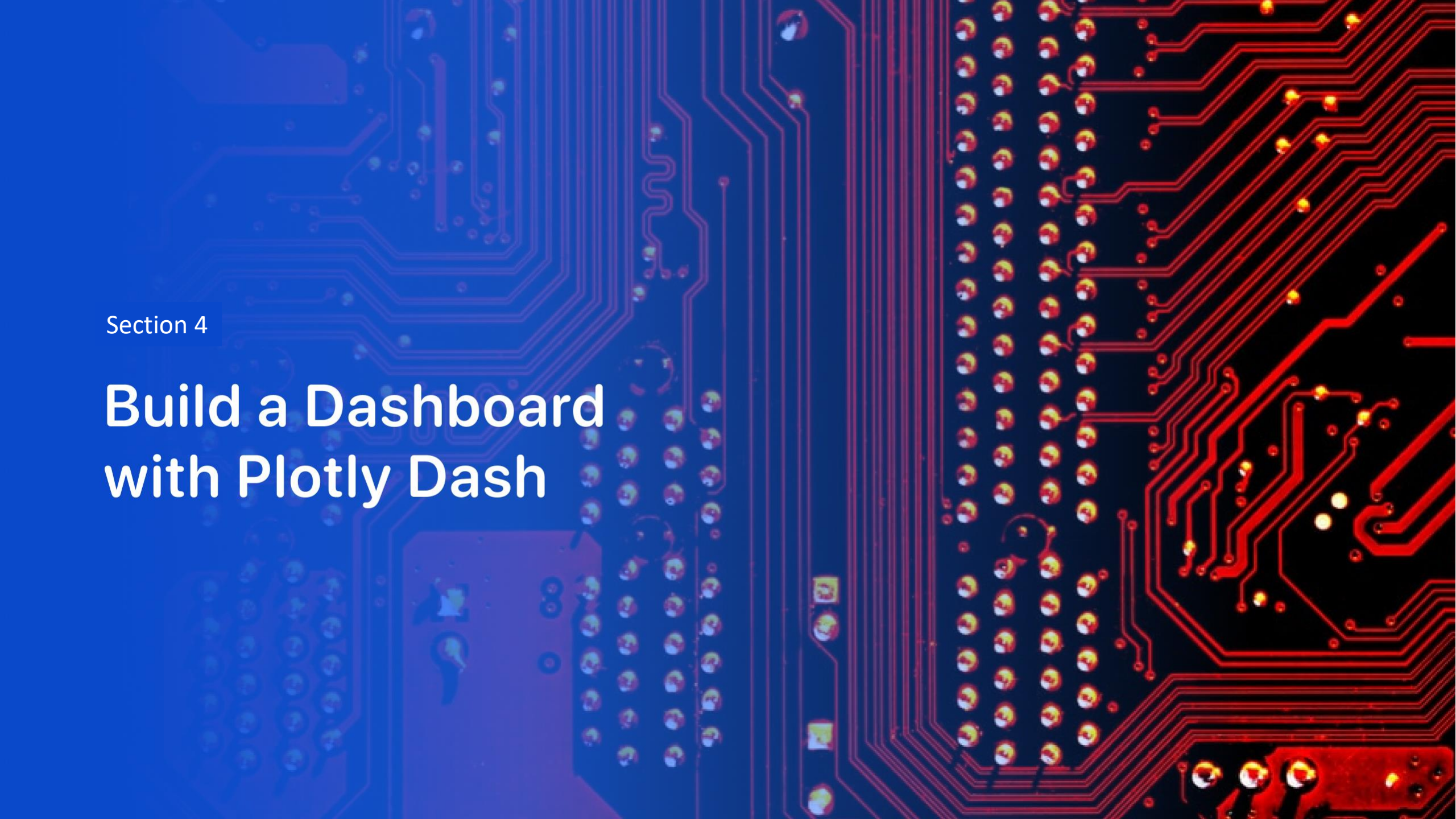
At this launch site there are more unsuccessful lands than successful.

Distance between a launch site an places

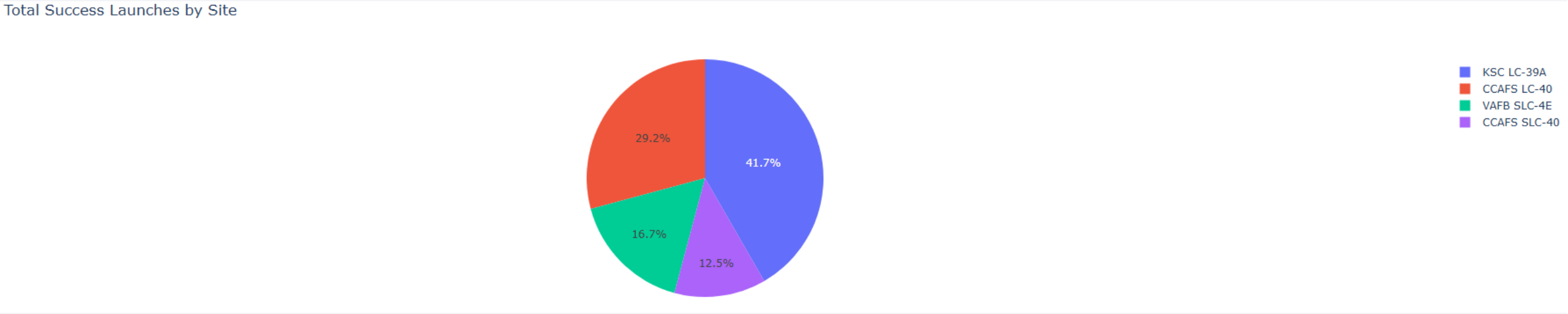


Distance from CCAFS SLC-40 to:

- Coast: 0.87 km
- Highway: 0.59 km
- Railway: 1.25 km
- Titusville: 23.43 km

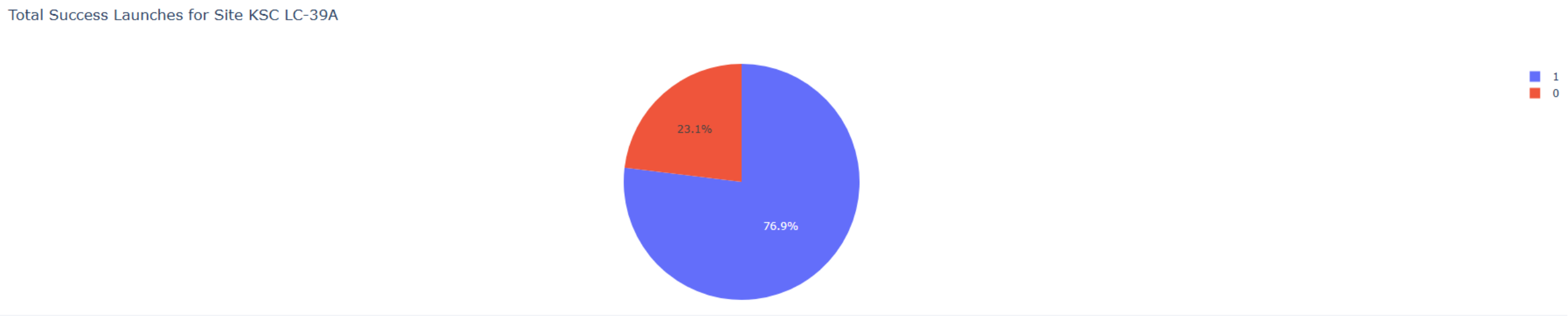

Section 4
Build a Dashboard with Plotly Dash

Total success launches by site



The most successful launch site is KSC LC-39A

Total success launches for site KSC LC-39A



The success rate is 76.9% for KSC LC-39A

<Dashboard Screenshot 3>

The payload between 2500 kg and 5000 kg are more successful than the ones between 5000 kg and 7500 kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

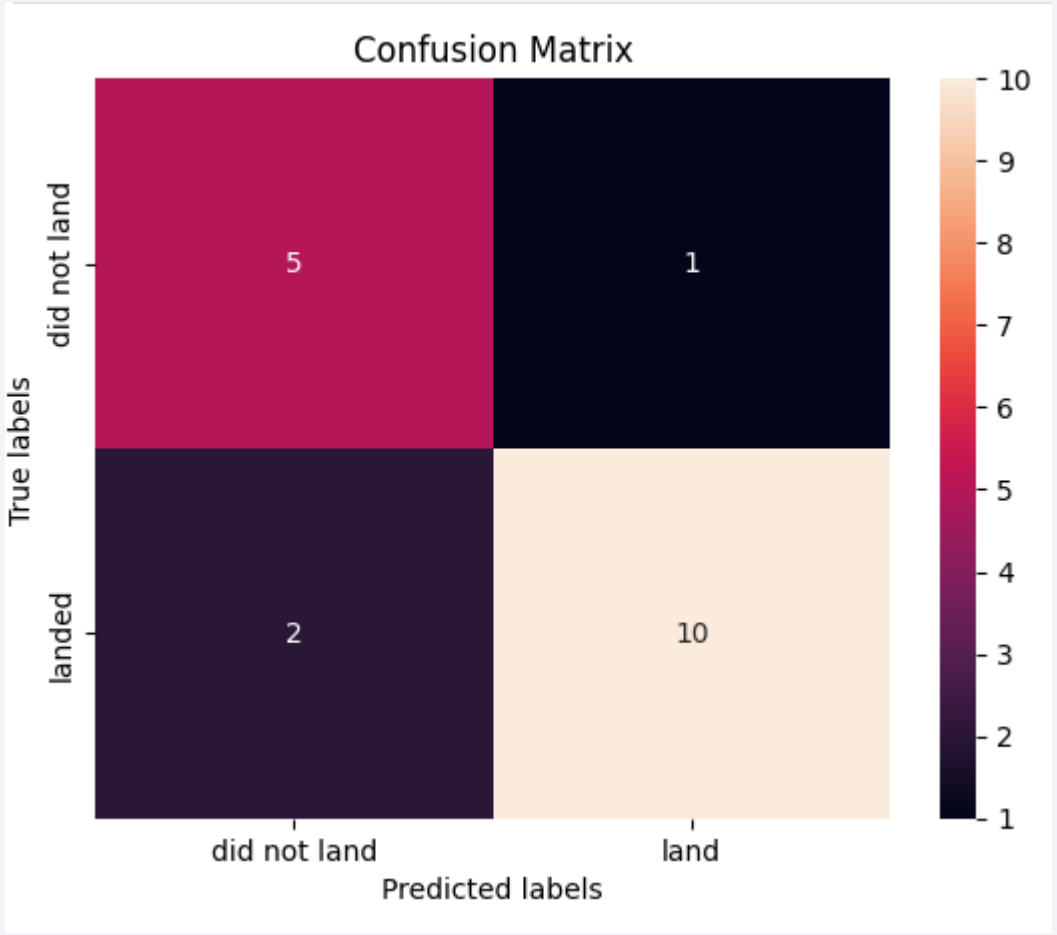
- The accuracy of all models are the same.

- We can infer that the decision tree classifier will have the best accuracy in the real world because it has the best model score (0.9).

Method	Score
Logistic Regression	0.833333
SVM	0.833333
Decision Tree Classifier	0.833333
KNN	0.833333

Confusion Matrix

- We can se that the TN and TP are higher than FN and FP.
- That indicates that the model classifies in a correct way the test data.



Conclusions

- As the flight number increases, the first stage is more likely to land successfully at all launch sites.
- Orbits ES-L1, GEO, HEO and SSO have the most successful rate.
- The success rate since 2013 kept increasing till 2020
- The most launches were made at the east coast.
- The most successful launch site is KSC LC-39A
- Decision tree classifier is the best method to fit the data

Thank you!