IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aditi Tambat
August 21, 2022

Outline

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

Executive Summary

- Data acquired via API, web scraping of Wikipedia
- Exploratory Data Analysis using SQL, Plotly visualizations
- Launch site locations explored using Folium
- Interactive dashboard created using Dash
- Machine learning algorithms trained & tested for prediction of landing success classification
- Algorithms optimized with grid search, accuracy score of 0.83 achieved

Introduction

- SpaceX has attempted to safely land rocket boosters for reuse since 2010
- First successful ocean landing was in 2014, ground landing in 2015
- Landing success rate has increased over time
- Can the success or failure of a landing attempt be predicted using factors such as flight number, orbit type, payload mass, etc?

Section 1

Methodology

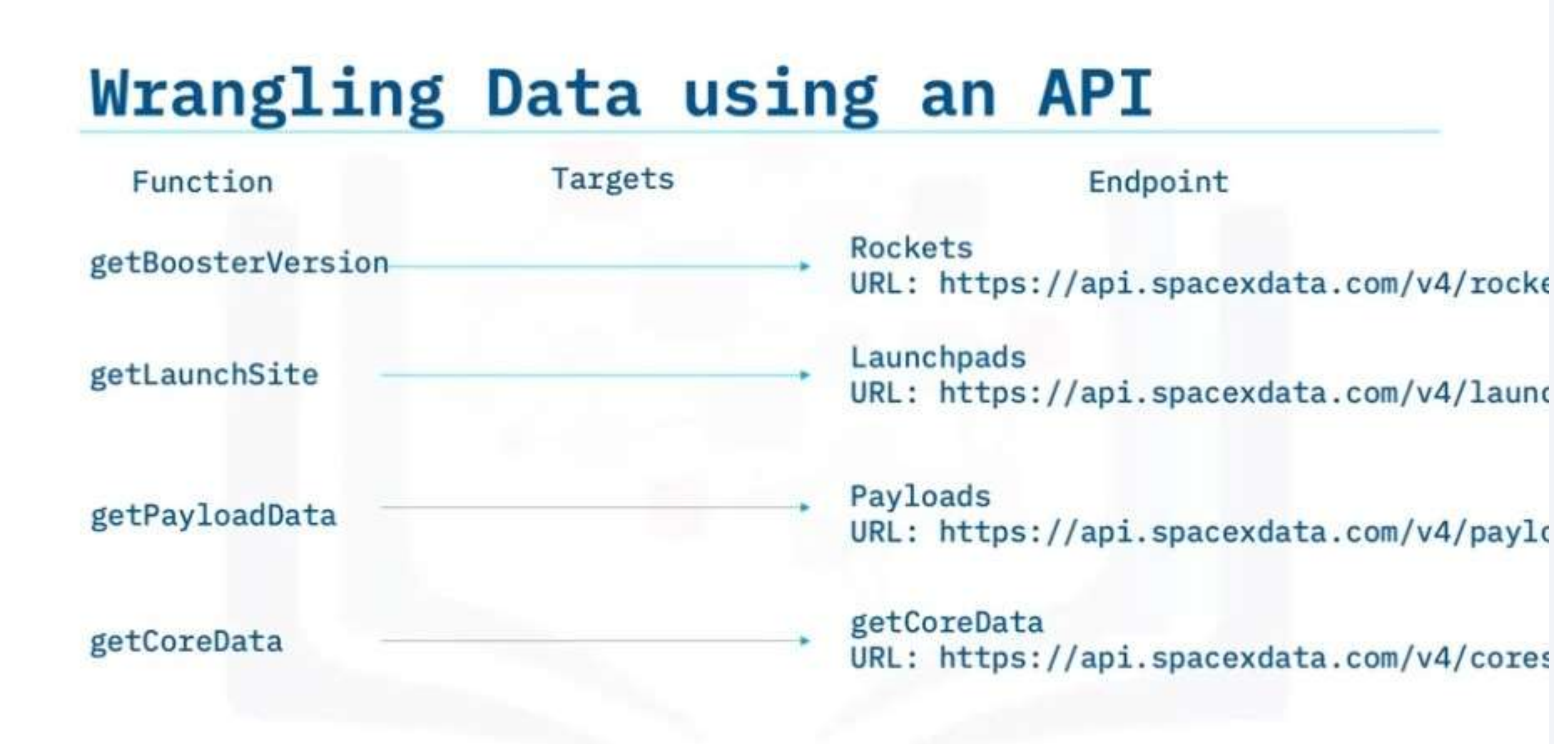
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using API calls and web scraping of Wikipedia
- Perform data wrangling
  - Data was processed to isolate booster type and address missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four different classification models were optimized for the data and evaluated

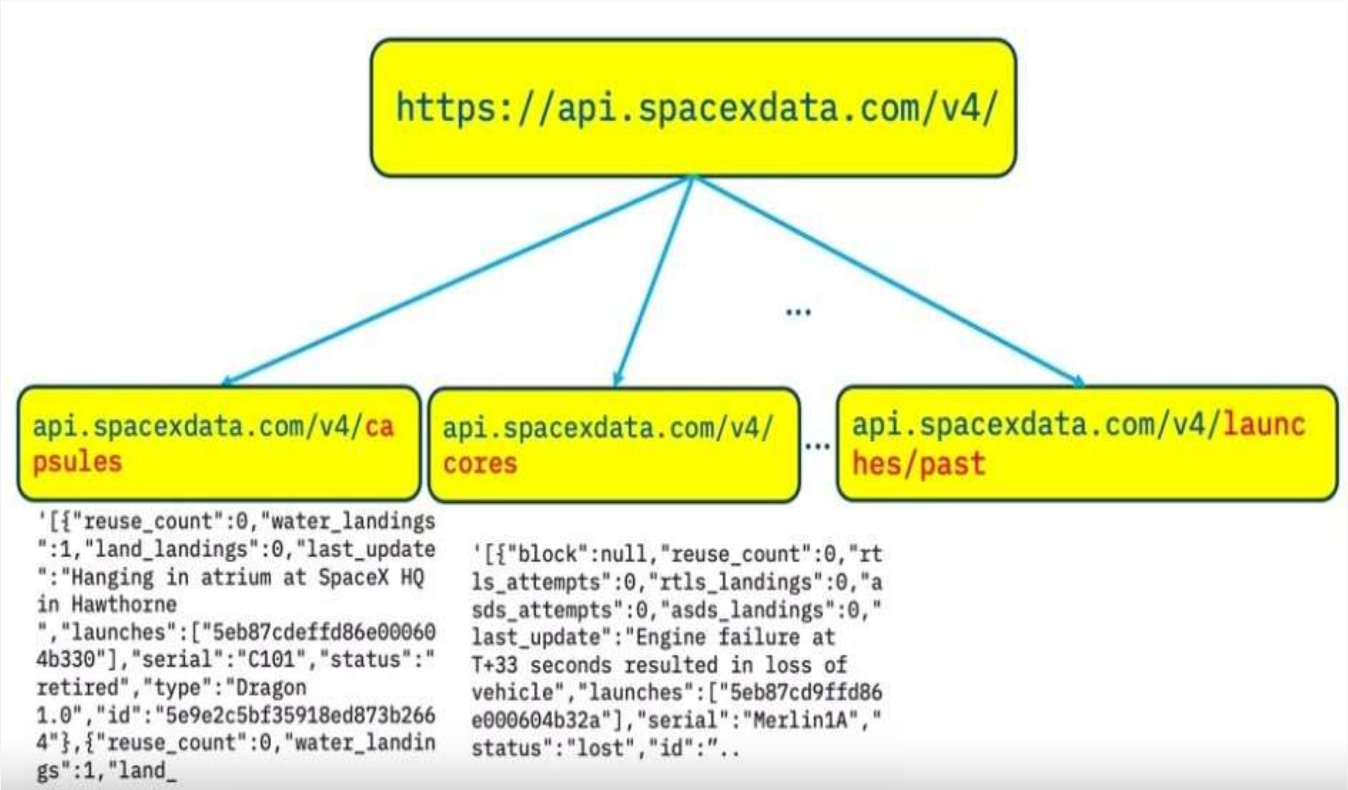
Data Collection

- Data was collected using API calls from spacexdata.com, as well as web scraping the Wikipedia page for Falcon 9 Launches



Data Collection – SpaceX API

- Data gathered from SpaceX API - Past launches
  - Data included launch date, launch site, booster version, payload mass, launch outcome, etc.
- Notebook located at https://github.com/aerrico1/capstone/blob/master/Data%20Collection%20API.ipynb



Data Collection - Scraping

- Web scraping used on Wikipedia page for Falcon 9 Launches
    - Data included launch date, launch site, booster version, launch outcome, customer, payload, etc.
- Notebook located at https://github.com/aerrico1/capstone/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCAFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Data Wrangling

- Data were filtered for booster version- only Falcon 9 launches used
- Missing values for payload mass were replaced with mean payload mass

```
data_falcon9.isnull().sum()

FlightNumber       0
Date               0
BoosterVersion     0
PayloadMass        5
Orbit              0
LaunchSite         0
Outcome            0
Flights            0
GridFins           0
Reused             0
Legs               0
LandingPad        26
Block              0
ReusedCount        0
Serial             0
Longitude          0
Latitude           0
dtype: int64
```

```
data_falcon9.isnull().sum()

FlightNumber       0
Date               0
BoosterVersion     0
PayloadMass        0
Orbit              0
LaunchSite         0
Outcome            0
Flights            0
GridFins           0
Reused             0
Legs               0
LandingPad        26
Block              0
ReusedCount        0
Serial             0
Longitude          0
Latitude           0
dtype: int64
```

- Data wrangling notebooks:
  - https://github.com/aerrico1/capstone/blob/master/Data%20Collection%20API.ipynb
  - https://github.com/aerrico1/capstone/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb
  - https://github.com/aerrico1/capstone/blob/master/EDA.ipynb

EDA with Data Visualization

- Charts plotted to find potential relationships between variables
  - Payload mass vs. flight number
  - Launch site vs. flight number
  - Launch site vs. payload mass
  - Success rate vs. orbit type
  - Orbit type vs. flight number
  - Orbit type vs. payload mass
  - Success rate vs. year
- Notebook located at https://github.com/aerrico1/capstone/blob/master/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- SQL queries performed:
  - Launch site names, launch sites beginning with “CCA”
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster F9 v1.1
  - Date of first successful landing using ground pad
  - Boosters with successful drone ship landings and payload mass between 4000 and 6000 kg
  - Total number of successful and failed outcomes, booster & date of failures
  - Booster versions which have carried maximum payload mass
  - List of landing outcomes ranked by count
- Notebook located at https://github.com/aerrico1/capstone/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

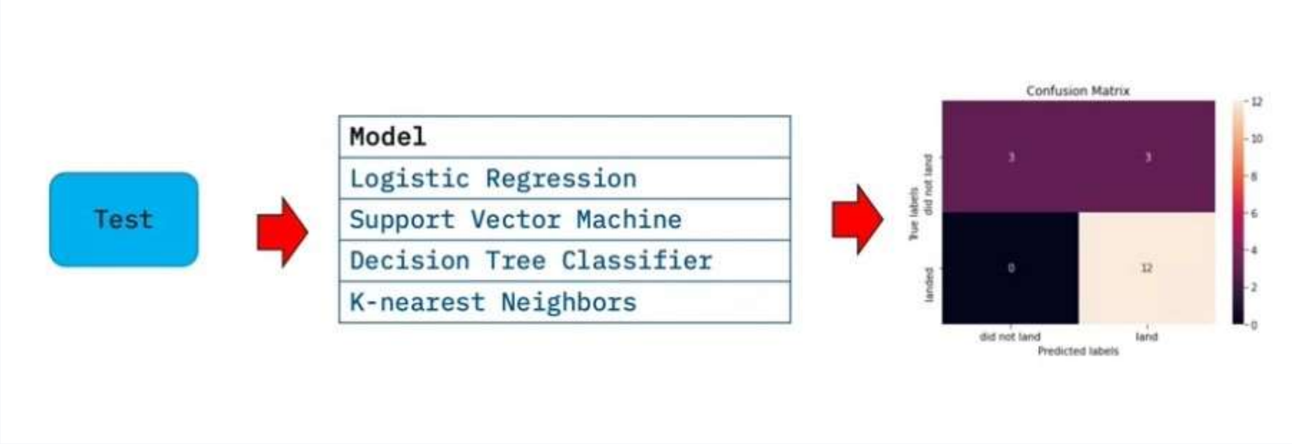
- Launch sites mapped using Folium
  - Markers added indicating successful and failed landing attempts launched from each site
  - Lines and distances to nearest coastline, highway, rail, and city added to map
- Maps used to determine how launch site locations are chosen
- Notebook located at https://github.com/aerrico1/capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Dashboard created with Plotly Dash
  - Pie chart displaying success rate for each individual launch site, and for all launch sites combined
  - Interactive scatter plot using site dropdown menu and payload mass slider to plot landing outcome against payload mass, with booster version included as data point color
- Dashboard can be used to visualize the factors that create a successful landing
- Python file located at https://github.com/aerrico1/capstone/blob/master/SpaceX%20Dash%20App.py

Predictive Analysis (Classification)

- Four different machine learning models were trained and tested using standardized split data and best parameters were found using Grid Search
  - Logistic regression, support vector machine, decision tree, K-nearest neighbors
- Confusion matrix plotted for test runs of all models



- Notebook located at https://github.com/aerrico1/capstone/blob/master/Machine%20Learning%20Prediction.ipynb

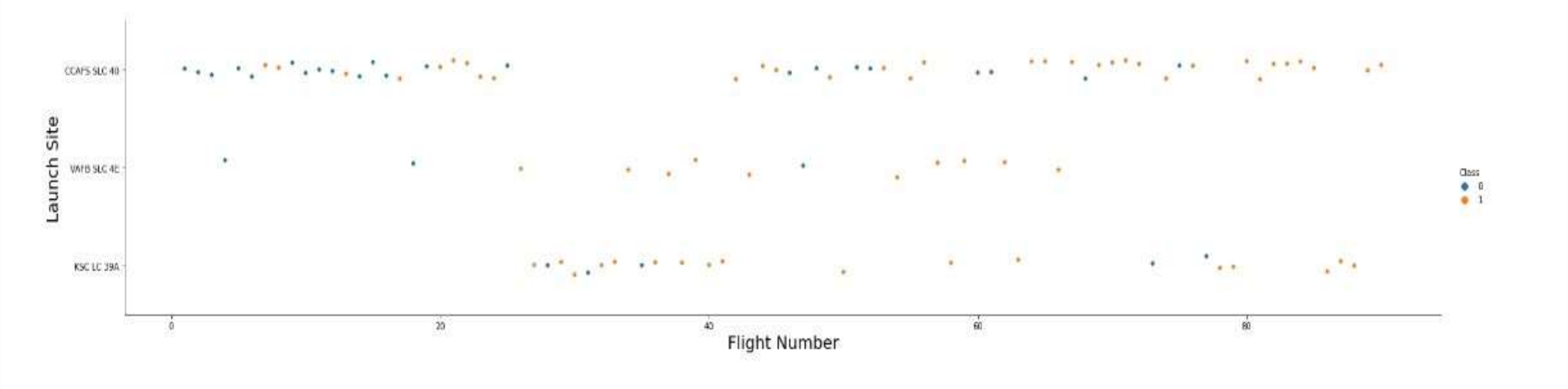
Results

- Exploratory data analysis revealed that landing success depends on factors such as launch site, flight number, payload mass, and orbit type
- Interactive analytics demo in screenshots can be found in Section 5
- Predictive analysis was performed on four machine learning algorithms, which were optimized using grid search.
- All algorithms had equivalent performance, with an accuracy score of 0.83.

Section 2

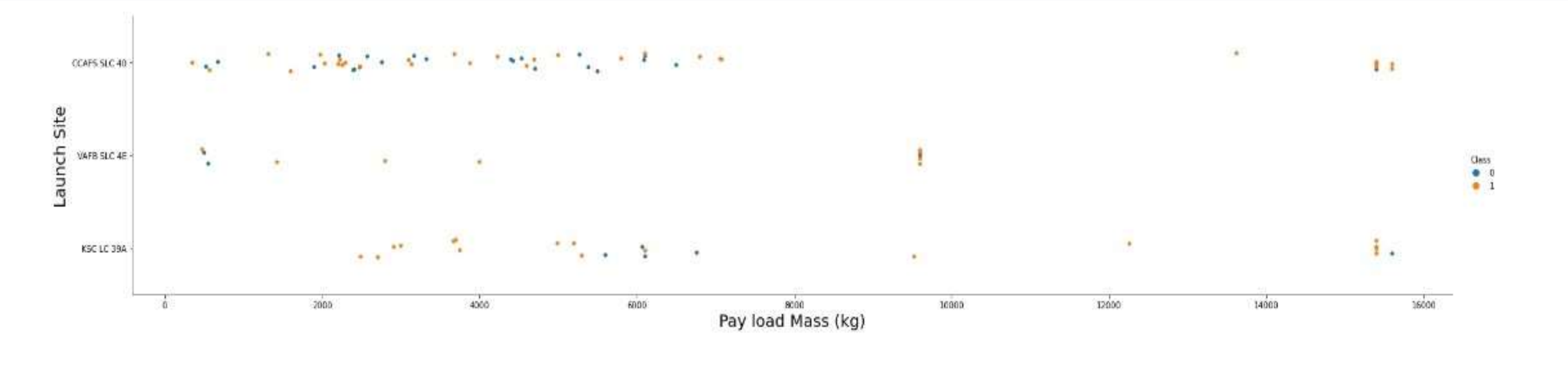
Insights drawn from EDA

Flight Number vs. Launch Site



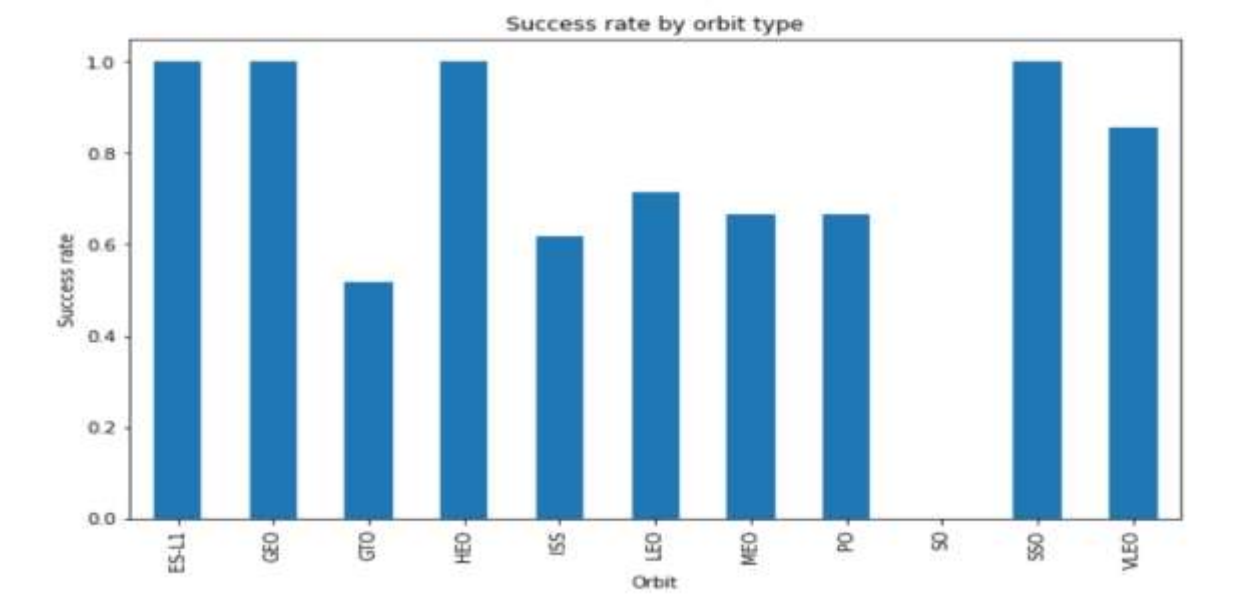
- This plot shows how the preferred launch site has varied over time, and how successful landings have become more common.

Payload vs. Launch Site



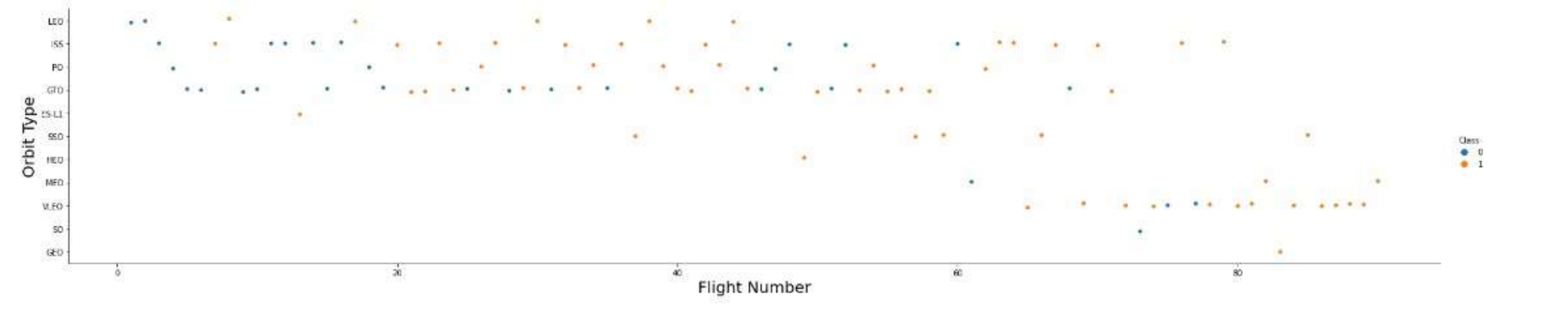
- This plot shows the distributions of payload mass, as well as the payload capabilities of each launch site. For example, VAFB SLC 4E is never used with payloads over 10,000 kg.

Success Rate vs. Orbit Type



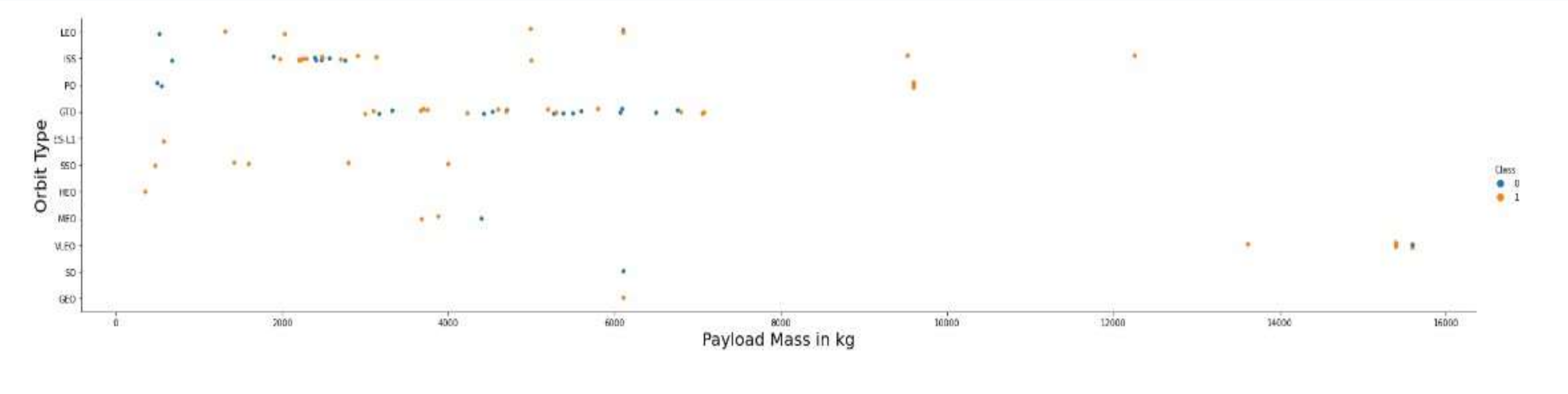
- This plot shows the landing success rate for various types of orbits. ES-L1, GEO, HEO, and SSO have high success rates, while other orbit types have more mixed results.

Flight Number vs. Orbit Type



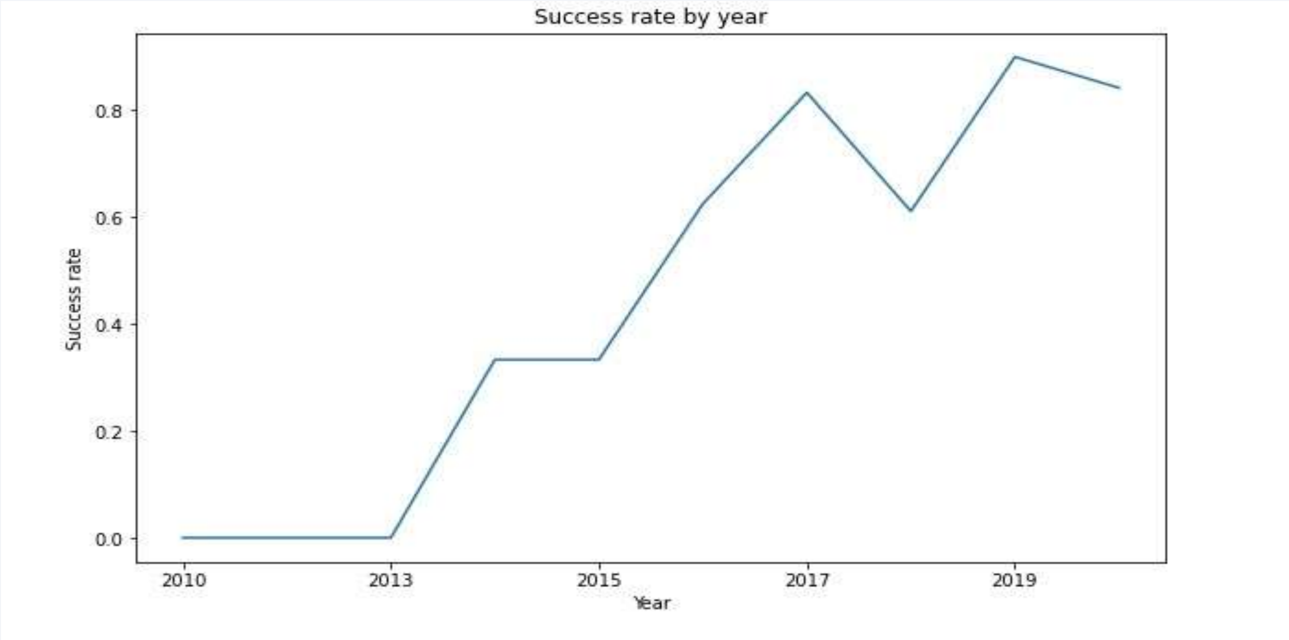
- This plot shows the different types of orbits used over time, and their outcomes. SpaceX appears to have transitioned from Low Earth Orbit to Very Low Earth Orbit. LEO landings become more successful with increasing flight number, other orbits such as GTO did not.

Payload vs. Orbit Type



- This plot shows the orbit types used with different payload masses, as well as the landing outcome. ISS takes lower payloads, GTO is used for mid-range, and VLEO is used for the heaviest payloads. Heavier payloads are more likely to have successful landing outcomes for certain orbit types, such as PO, LEO, and ISS.

Launch Success Yearly Trend



- This chart shows how the landing success rate has changed by year. Since the first successful landing, the success rate has trended upward.

All Launch Site Names

%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXDATASET;

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

- Returns the names of unique launch sites used

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXDATASET \
WHERE LAUNCH_SITE LIKE 'CCA%' \
FETCH FIRST 5 ROWS ONLY;
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Returns the first 5 records where the launch site name begins with the string “CCA”

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXDATASET \
WHERE CUSTOMER='NASA (CRS)';
```

- Result: 45,596
- Returns total payload mass (in kg) carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXDATASET \

WHERE BOOSTER_VERSION LIKE '%F9 v1.1%';

- Result: 2534
- Returns the average payload mass (in kg) carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) FROM SPACEXDATASET \
WHERE LANDING__OUTCOME='Success (ground pad)';
```

- Result: 2015-12-22
- Returns the date of the first successful landing outcome where the ground pad was used

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXDATASET \

WHERE (LANDING__OUTCOME='Success (drone ship)' AND
(PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000));
```

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

- Returns the unique names of boosters which have successfully landed on the drone ship after carrying a payload between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as count FROM SPACEXDATASET \
GROUP BY MISSION_OUTCOME;
```

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Returns the number of successful and failure mission outcomes

Boosters Carried Maximum Payload

%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXDATASET \

WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- Returns the unique names of booster versions which have carried the maximum payload mass

2015 Launch Records

```
%sql SELECT DATE, LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXDATASET \

WHERE (LANDING__OUTCOME='Failure (drone ship)' AND (DATE BETWEEN '2015-01-01' AND '2015-12-31'));
```

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Returns the launch records from failed drone ship landing attempts in 2015

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

```
%sql SELECT LANDING__OUTCOME, COUNT(*) as count FROM SPACEXDATASET \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY LANDING__OUTCOME \
ORDER BY COUNT DESC;
```

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

- Returns a list of landing outcomes between the selected dates, ranked by count in descending order

Section 4

Launch Sites Proximities Analysis

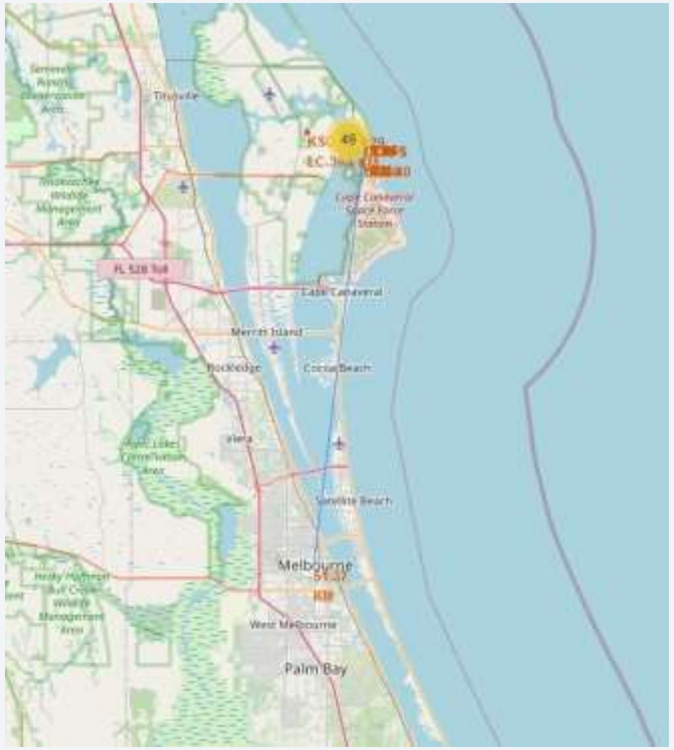
Folium map of launch sites

- Some launch sites are too close together to be viewed individually at this scale

Map with outcome indicators

- Green markers indicate successful landings; red markers indicate failed landings

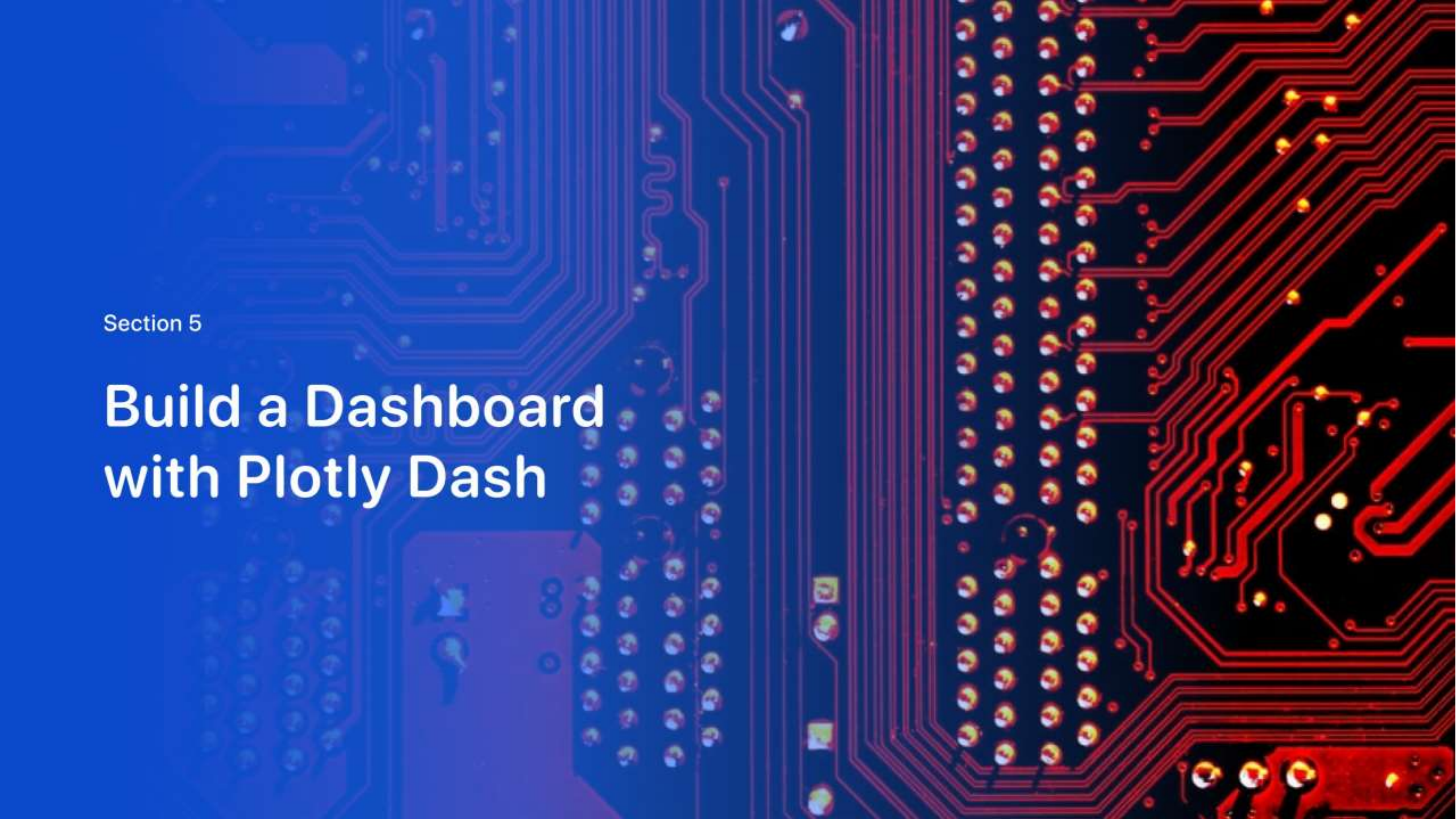
Distances to nearest features



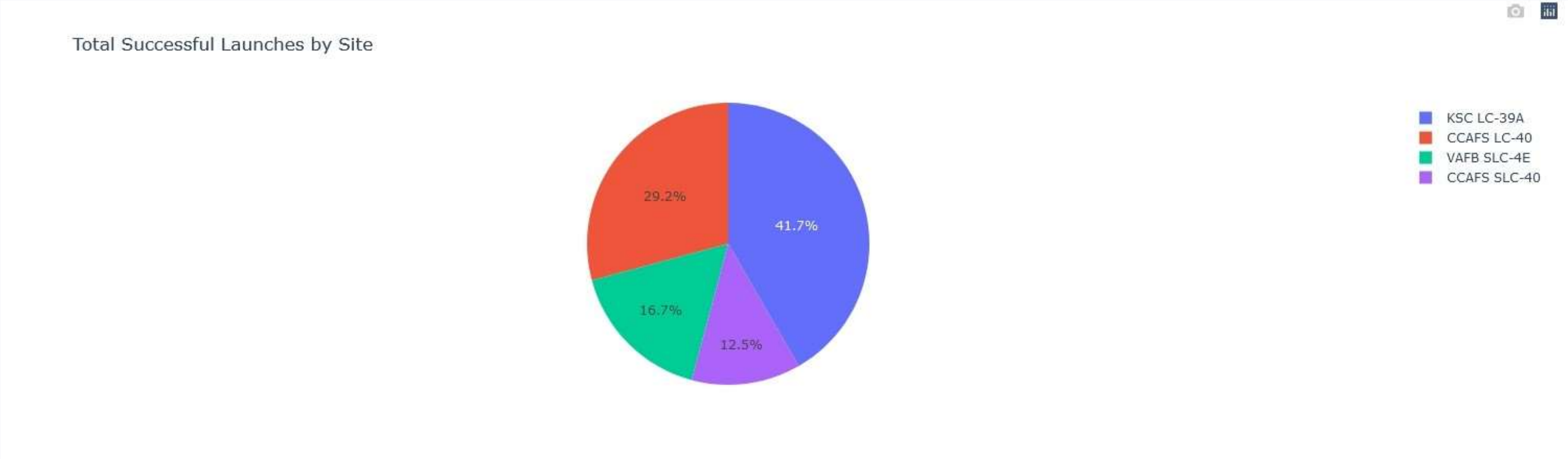
- Map shows short distances (<2 km) to nearest coastline, highway, and railway, but long distance (>50 km) to nearest city

Section 5

Build a Dashboard with Plotly Dash

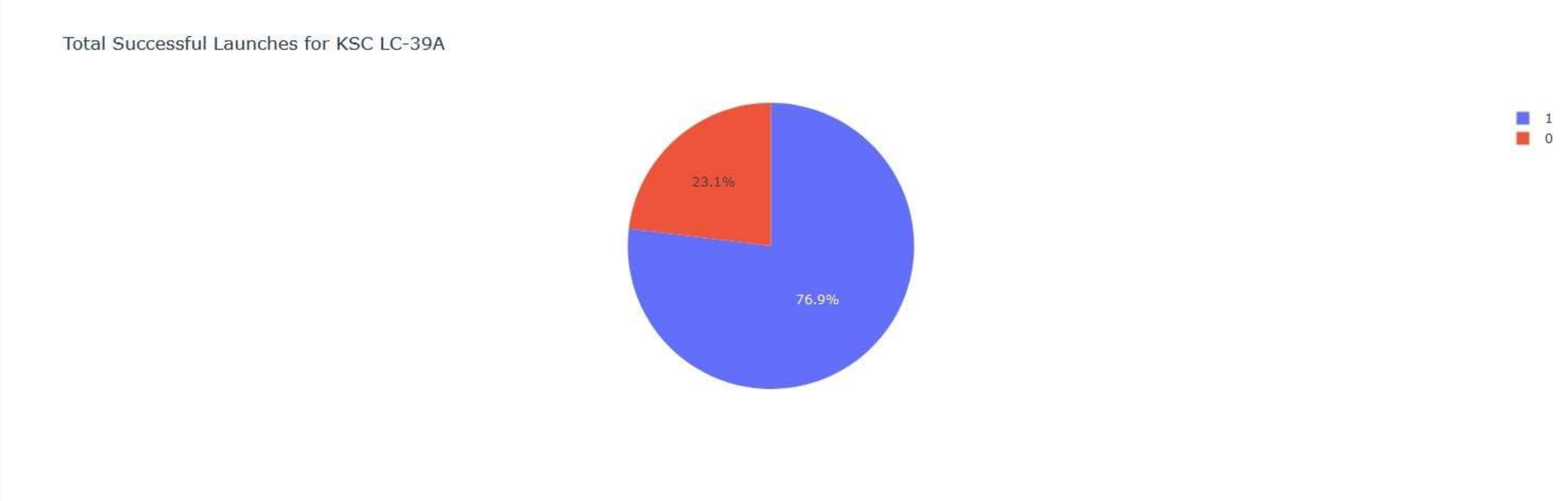


Dashboard – success count for all sites



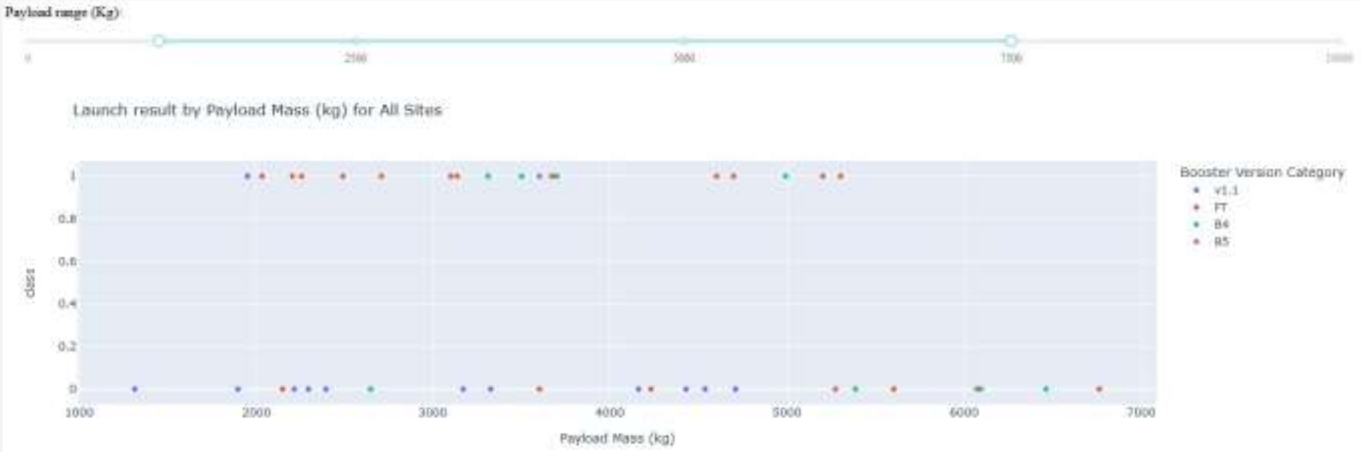
- Pie chart shows the percentage of successful landings launched from each site

Dashboard – single site pie chart



- Pie chart shows landing outcomes for KSC LC-39A, which had the highest success rate of all sites

Dashboard – launch result vs. payload mass

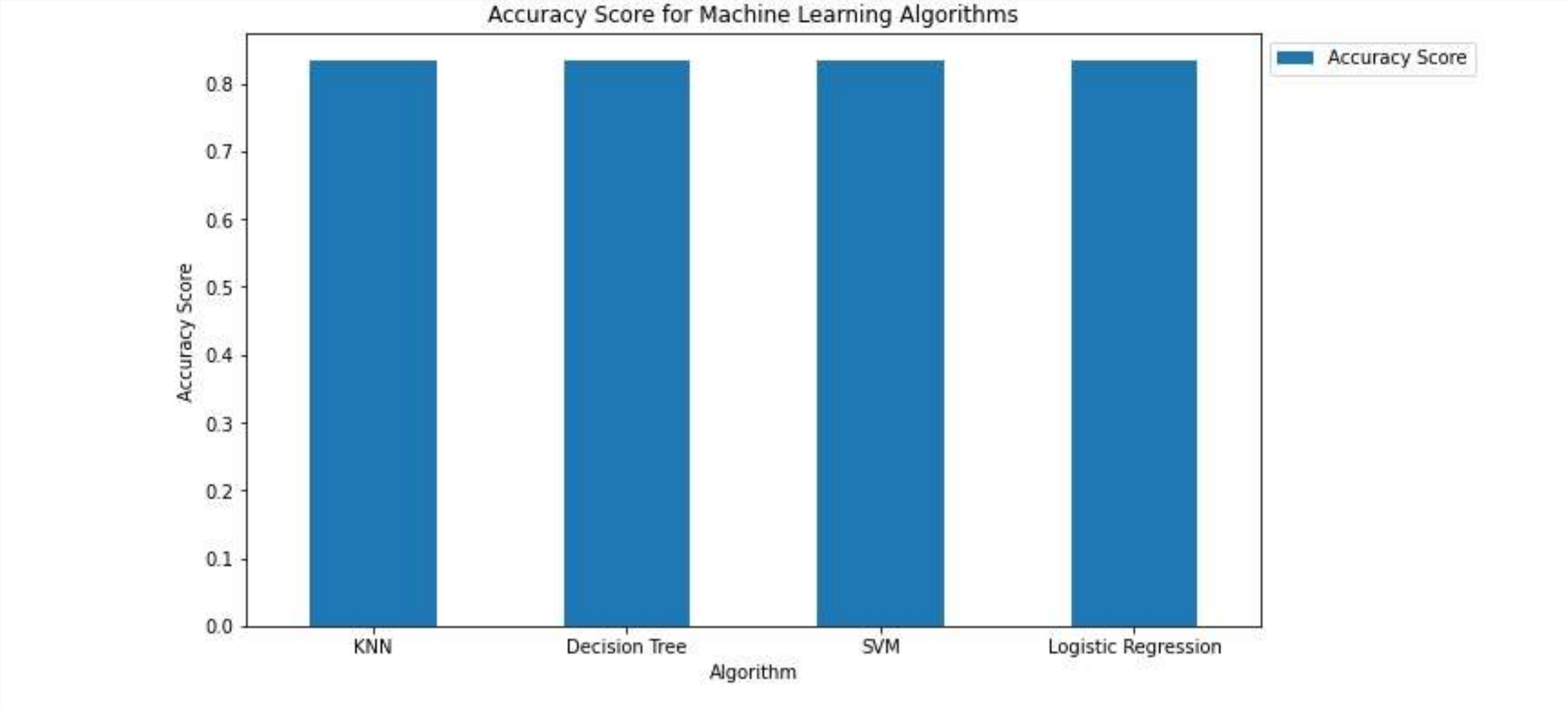


- Payload range selectable using slider
- FT booster had highest success rate
- Success rate highest when payload between 2000 and 5500 kg

Section 6

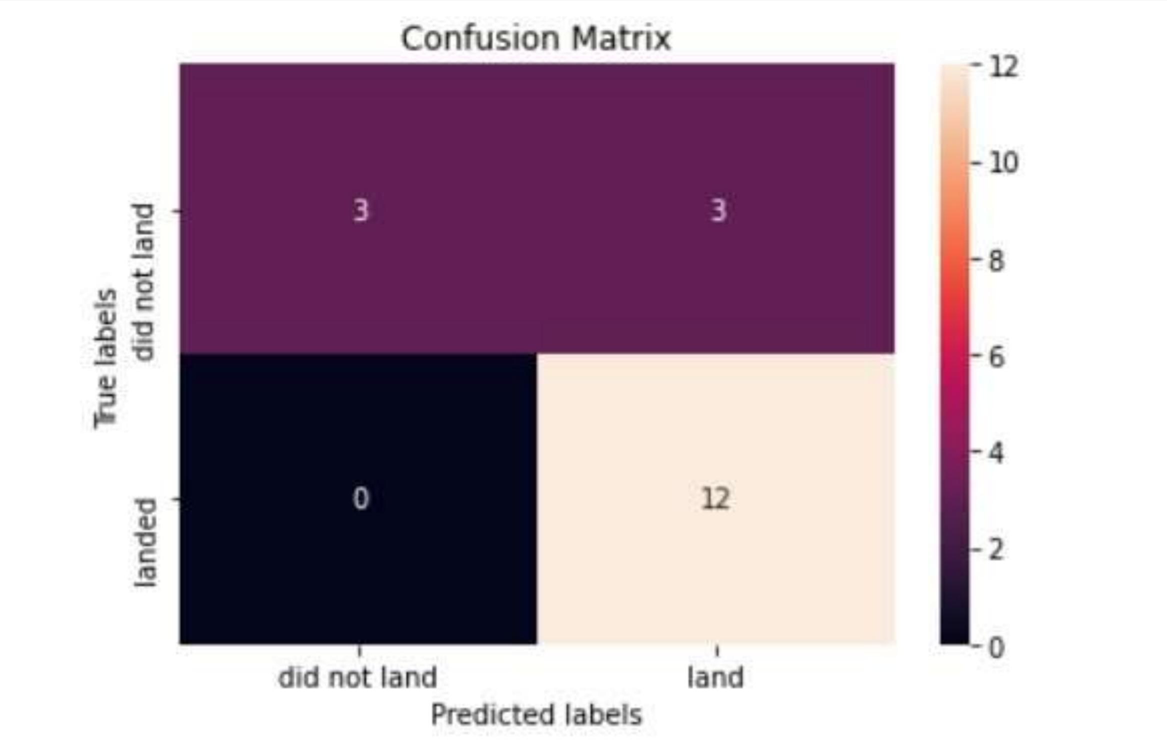
Predictive Analysis (Classification)

Classification Accuracy



- All models evaluated had equivalent accuracy, at 83%

Confusion Matrix



- The confusion matrix shows true negatives (3), false positives (3), false negatives (0), and true positives (12). The confusion matrix was the same for all models.

Conclusions

- The outcome of attempted SpaceX booster landings can be predicted with fairly good accuracy
- More testing data is likely needed to do a more comprehensive evaluation of the different machine learning models
- Launch sites are chosen to minimize distance to coastlines, highways, and rail lines, and to maximize distance to inhabited areas
- The most successful landing outcomes occurred when the payload mass was between 2000 and 5500 kg
- The landing success rate has increased over time as SpaceX has amassed more experience with the project

Appendix

- All code can be found at https://github.com/aerrico1/capstone

Thank you!