IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sasanka Madawalagama

2024. 04. 04

Outline

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

Executive Summary

- **The commercial Space Age is here.**
Commercialisation of space flights only became possible with the reduced cost of space flights.

- **SpaceX** is renowned for its work in making space flights affordable by reusing the boosters.

- This study analysed past data of SpaceX Falcon9 launches to **get insights** and build a **prediction model** for successful recovery.

- Proper **data collection and wrangling methods**, as well as exploratory data analysis (**EDA**) with **interactive visual outputs**, were used.

- With the collected data several **Machine Learning Algorithms** were trained to find the best performing model to predict the success of a recovery

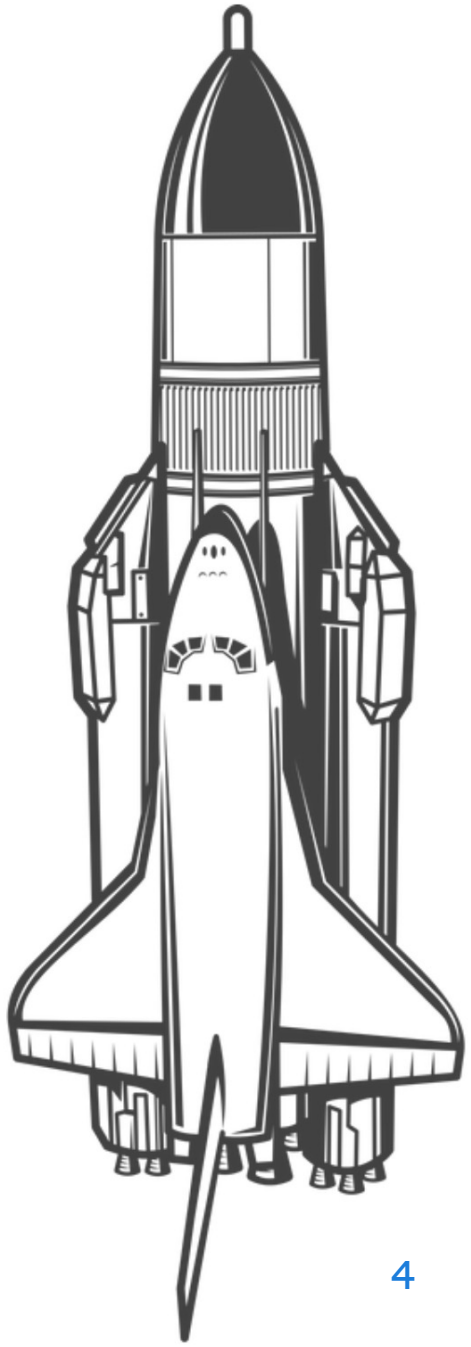
Introduction

- The commercial space age is here, companies are making space travel affordable for everyone.
- SpaceX is the most successful in commercial space operations.
  - SpaceX makes space flights affordable by Reusing the First Stage of the flight.
  - SpaceX flight only cost 62 million USD while the competitors cost upward of 165 million USD.

Objective:

Collect and Analyze the historical data from SpaceX launches to

- Get valuable insights into SpaceX Flights
- Predict successful recovery of the first stage



Section 1

Methodology

Methodology

- Data collection:
  - Using SpaceX REST API and applying web scraping techniques
- Perform data wrangling
  - Data is pre-processed by filtering the data, handling the 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
  - Different ML techniques (SVM, Logistic Regression, KNN, Decision Tree) were applied to find the best predictive model.
  - Optimum model parameters were found by applying Grid Search

Data Collection

1. Request Data

- Data is obtained from SpaceX API
- Parsed the SpaceX launch data (using the GET request)
- API helper functions were used to deal with different API endpoints

2. Cleaning the Data

- Obtained Data were filtered to include Falcon9 Launches
- Missing Data Values were fixed

Data Collection – SpaceX API

- Data is gathered from SpaceX REST API. The API gives data about launches, the rockets used, payload, specifications, outcomes, etc.
- There are different endpoints for each type of data, so multiple helper functions were used to get data from each endpoint.
  - Base Data → .json object about base information about launched
  - Specific data → When the base data only contains coded information (rocet ID), it was required to do different API calls to get specific data, such as booster and launch data.
- Obtained Data for the Analysis
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

API Calls to SpaceX API

1. Base Data (https://api.spacexdata.com/v4/launches/past)
2. Booster Data (https://api.spacexdata.com/v4/rockets/)
3. Launch Data (https://api.spacexdata.com/v4/launchpads/)
4. Payload (https://api.spacexdata.com/v4/payloads/)
5. Core Data (https://api.spacexdata.com/v4/cores/)

GitHub URL to the Notebook: https://github.com/sasankamadawalagama/ibm_ds_coursera/blob/main/W11_data_collection_api.ipynb

Data Collection - Scraping

Scraping the data from Falcon9 Wikipedia Page

- https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- BeautifulSoup Object was created

Extract all column (variable) names from the HTML header

- Extracted variable names:
  Flight No., Date and time ( ), Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome]

Creating a pandas DataFrame from the data for further analysis

- A dictionary was created frist to simplify the phrasing process
- 121 records were obtained

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

GitHub URL to the Notebook: https://github.com/sasankamadawalagama/ibm_ds_coursera/blob/main/W12_webscraping.ipynb

Data Wrangling

DATA WRANGLING STEPS

1. In data wrangling, web scraped data was converted into a usable format. i.e. pandas DataFrame with desired observations
2. Null values were observed
3. Creating the Binary Landing Outcome column (dependent variable)
   A. Since this work is carried out to determine the successful recovery of the second stage of the Falcon 9 booster, it' s required to obtain data for the landing outcome. (success or failure)
   B. The dataset doesn't provide binary landing outcomes, directly.
   C. It was calculated by
      A. First observing the outcome description
      B. Identifying each outcome category

DATA WRANGLING RESULTS

There are a total 90 recodes in the dataset for Falcon 9 landings and 60 of the landings were observed to be successful.
Success Rate = 66.6%

Table: Coding to convert Outcome text to binary outcome column (dependent variable)

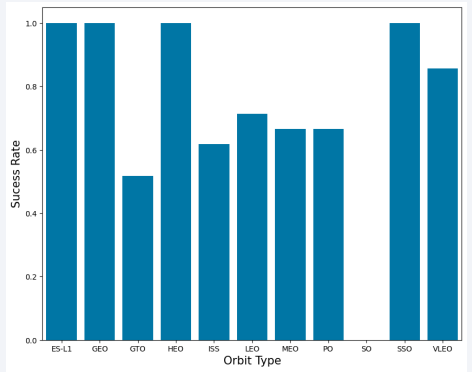
Outcome	Description	Binary Outcome (Success/Failure)
True ASDS	Successfully landed to a drone ship	1
None None	Failure to land	0
True RTLS	Successful landing on a ground pad	1
False ASDS	Successfully landed to a drone ship	0
True Ocean	Successful landing in ocean	1
False Ocean	Unsuccessful landing in ocean	0
None ASDS	Failure to land	0
False RTLS	Unsuccessful landing on a ground pad	0

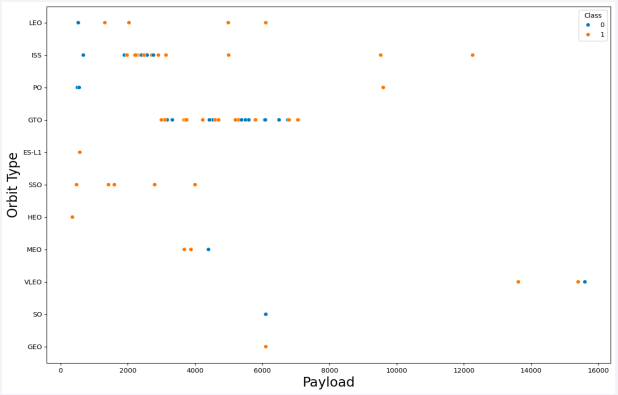
GitHub URL to the Notebook: https://github.com/sasankamadawalagama/ibm_ds_coursera/blob/main/W13_data_wrangling.ipynb

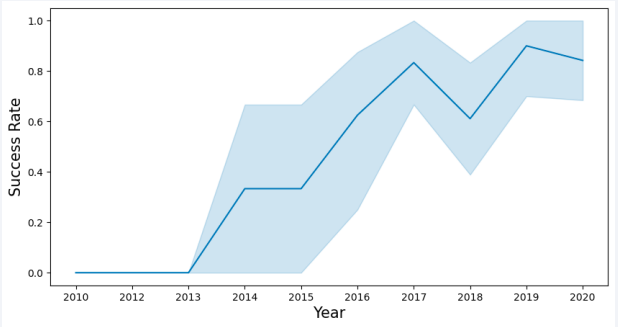
EDA with Data Visualization

The following charts were produced to identify the patterns in successful landings

- Flight Number vs Payload
- Flight Number vs Launch Site
- Payload vs Launch Site
- Success Rate of Each Orbit
- Flight Number vs Orbit
- Payload vs Orbit







GitHub URL to the Notebook: https://github.com/sasankamadawalagama/ibm_ds_coursera/blob/main/W22_eda-dataviz.ipynb

EDA with SQL

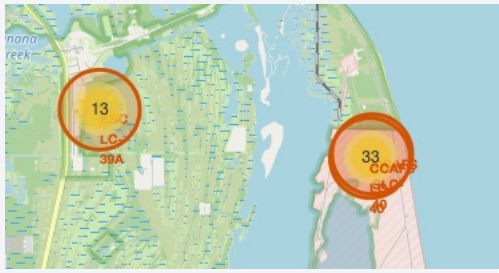
Exploratory Data Analysis (EDA) was performed with SQL. The following is the summary of queries used

- Names of unique launch sites
- 5 records where launch sites begin with ‘CCA’
- Total payload mass carried by boosters launch NASA (CRS)
- Average payload carried by booster F9 v1.1
- Date of the first successful landing outcome in ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions which have carried the max payload
- 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 2010-06-04 and 2017-03-20, in descending order.

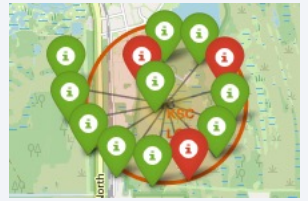
GitHub URL to the Notebook: https://github.com/sasankamadawalagama/ibm_ds_coursera/blob/main/W21_sql.ipynb

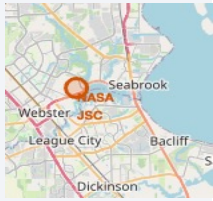
Build an Interactive Map with Folium

- **All Launch Sites were Mapped**
  - All four launch sites were mapped with latitude and longitude provided
  - Folium Circle object was used to mark the lauch site
  - Folium Map Marker object was used to mark the corresponding name



- **Launch Outcomes Corresponding to Each Launch Site was Mapped Indicating Success/Failure**
  - Folium Marker Cluster object was used to map the outcomes because the same coordinate is used in many missions
  - Sucessful outcomes were marked in green while uncussefull outcomes were marked in red



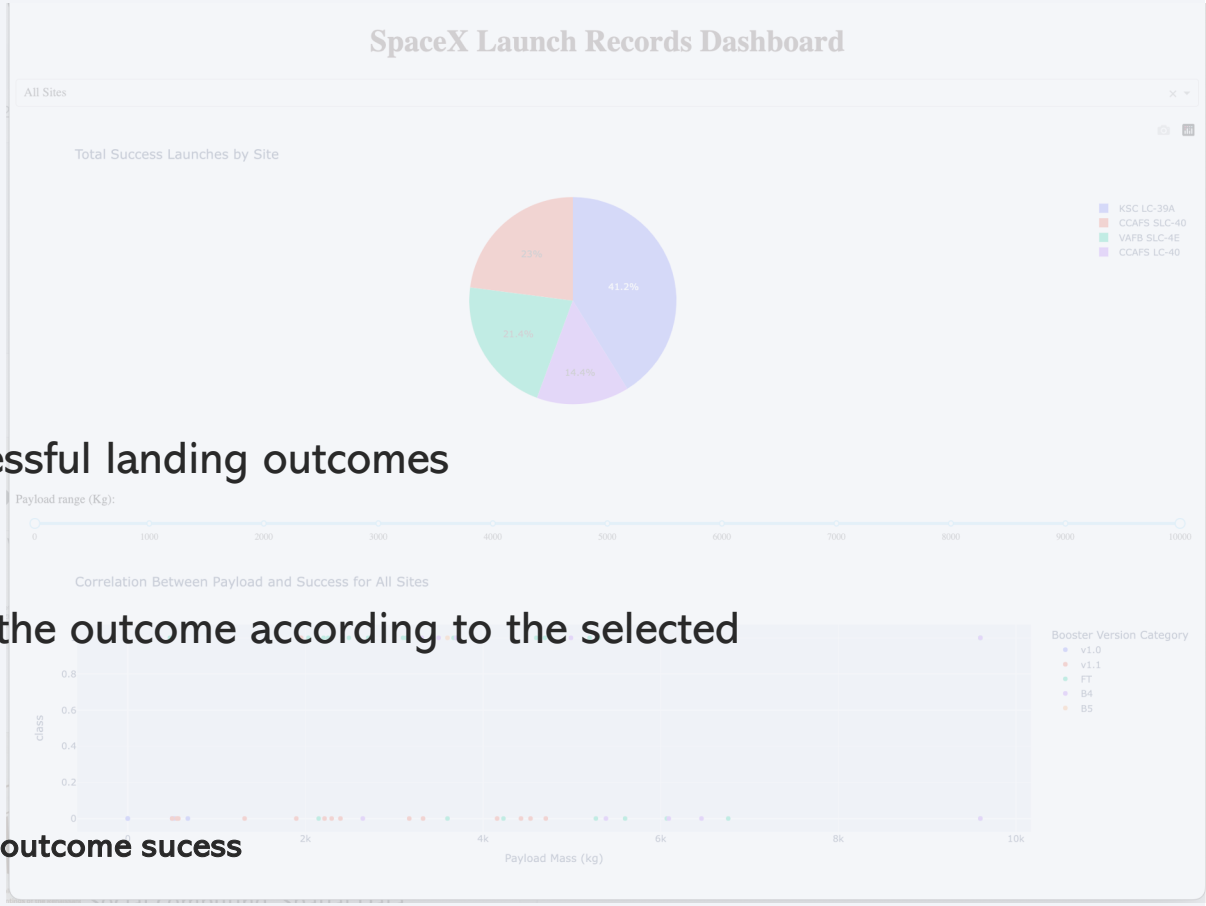


- **Mapping Distances Between Launch Sites to Proximities**
  - Distance to the nearest coastline, railway, city, and highway is calculated from the VAFB SLC-4E Launch Site.
  - Distances were marked using Folium Polyline object

GitHub URL to the Notebook: https://github.com/sasankamadawalagama/ibm_ds_coursera/blob/main/W31_folium_map_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Plotly Dash application was developed to perform interactive visual analytics on SpaceX launch data



Components

- **Launch Site Drop-down Input Component**
  - Allows to select a particular launch site or all sites
- **Pie Chart Showing Successful Outcomes**
  - To visualize the proportion of successful and unsuccessful landing outcomes
- **Slider to Select Payload Mass Range**
  - Users can select the payload mass range to visualise the outcome according to the selected range.
- **Scatter Chart Showing Payload Mass vs Success Rate**
  - **To visualize the correlation between the payload mass and mission outcome sucess**

GitHub URL to the Notebook: https://github.com/sasankamadawalagama/ibm_ds_coursera/blob/main/W32_spacex_dash_app.py

Predictive Analysis (Classification)

- With the collected data, Machine Learning Pipeline was developed to predict the successful mission outcome
- Four different ML learning methods are used to find the best
  - Support Vector Machine (SVM)
  - Logistic Regression
  - Decision Tree
  - K Nearest Neighbour
- Grid Search was used to identify the best model parameters
- Models were evaluated using
  - Accuracy, F1 Score, Jaccard Coefficient

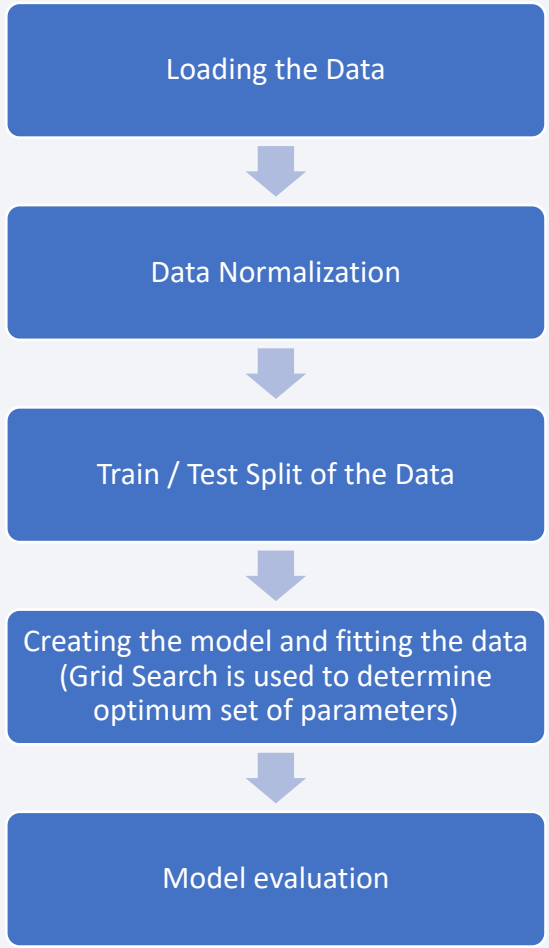


Figure: Model development pipeline

GitHub URL to the Notebook: https://github.com/sasankamadawalagama/ibm_ds_coursera/blob/main/W41_SpaceX_Machine_Learning_Prediction_Part_5.ipynb

Results: EDA

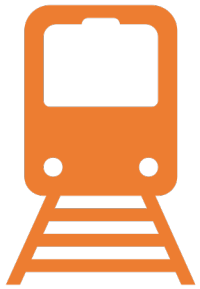
Early flights were done in the CCAFS SLC 40 launch site, which has higher number of failures at the early stage of SpaceX operations.

Most of the launches have payloads below 8000kg.

The success of recovery depends on the orbit.

Early flights got into LEO, ISS, PO, and GTO orbits.

Results: Visual Analytics



Launch Sites are located near a means of transportation like a highway or railway

Launch sites are near the equator

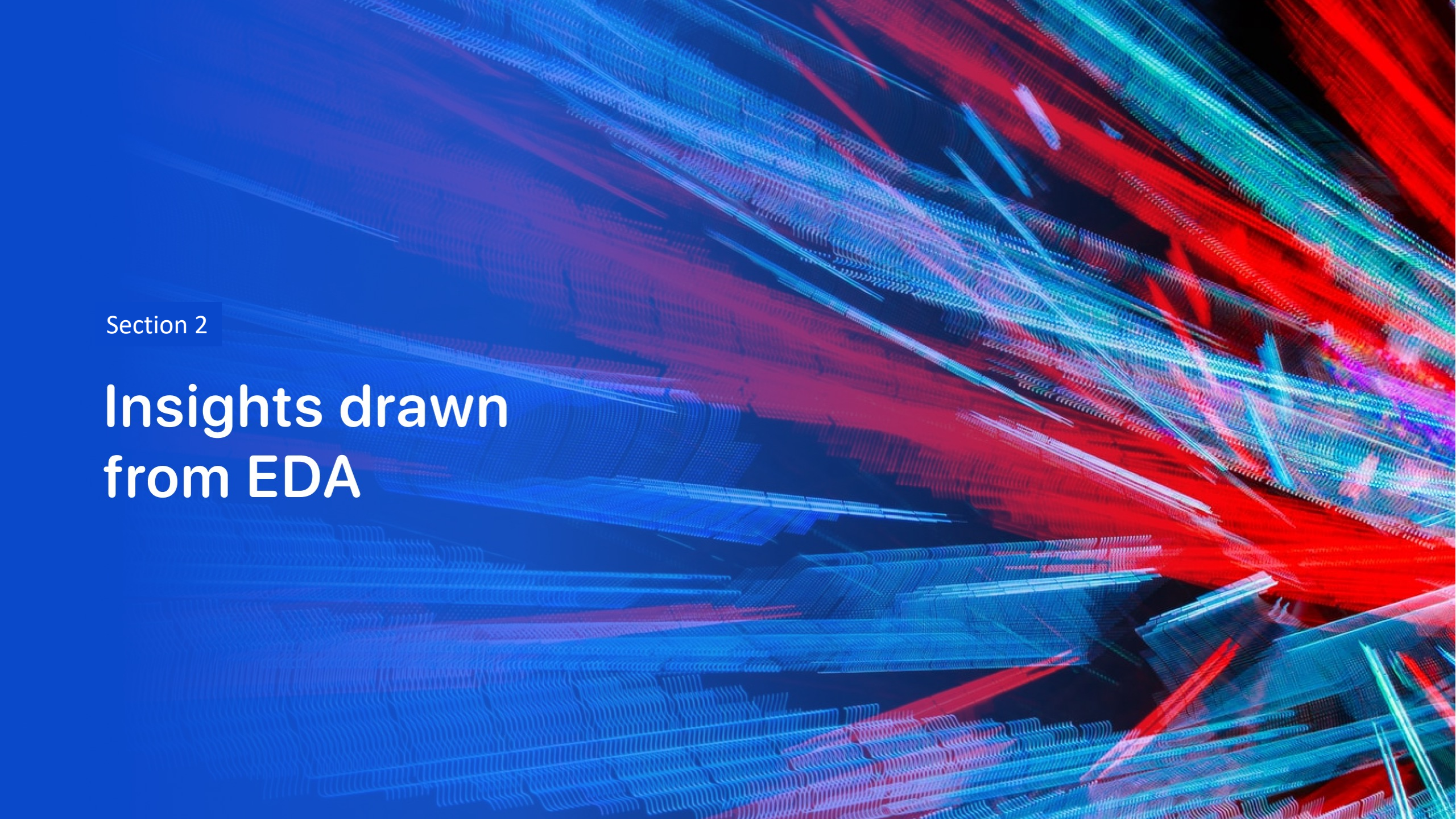
Launch sites are located in remote areas closer to coastline to minimise the damage to the general public if the launch went wrong

Results: Predictive Analysis

Several ML models were trained and evaluated.

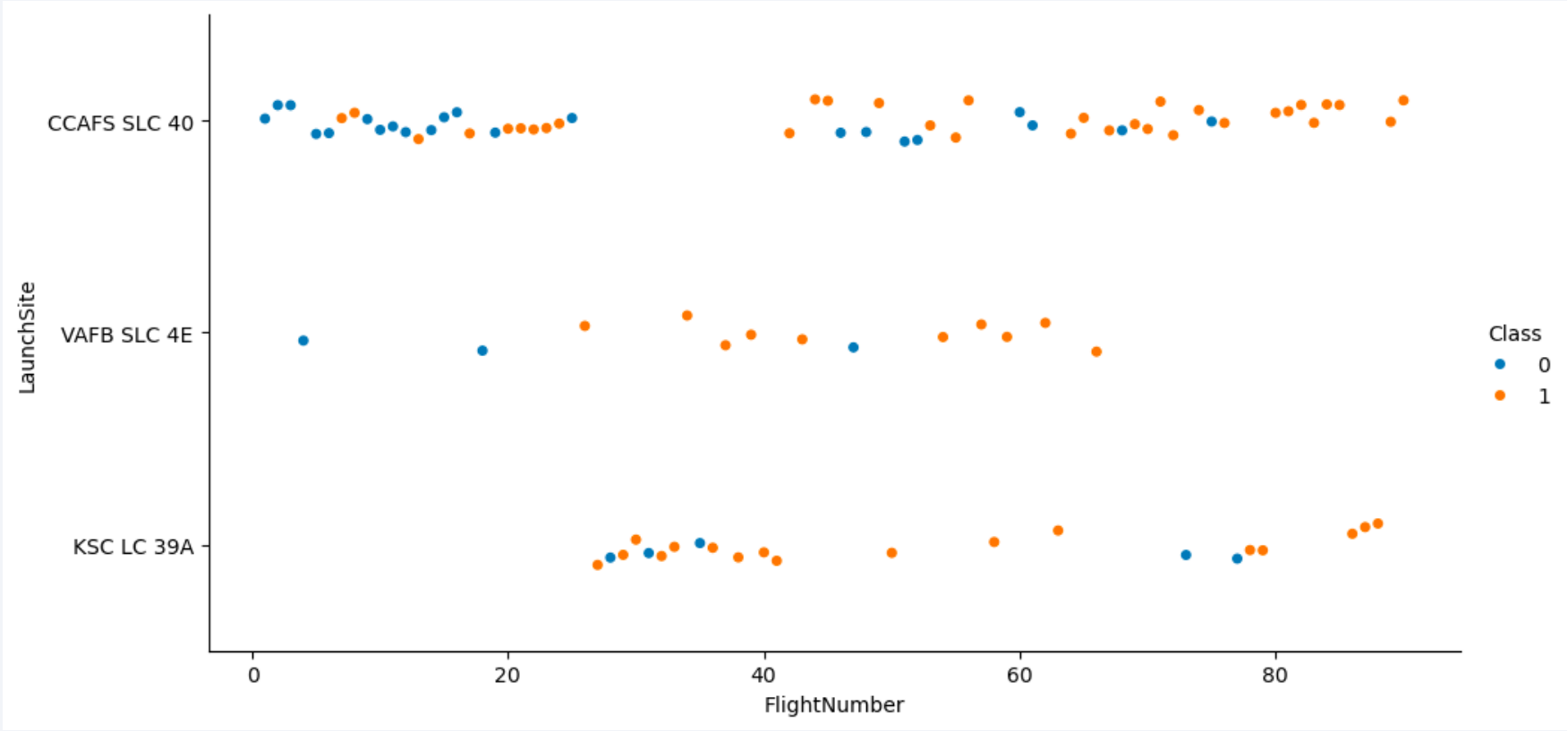
All models give about the same accuracy

To increase the accuracy of the models, the amount of data need to be increased


Section 2
Insights drawn from EDA

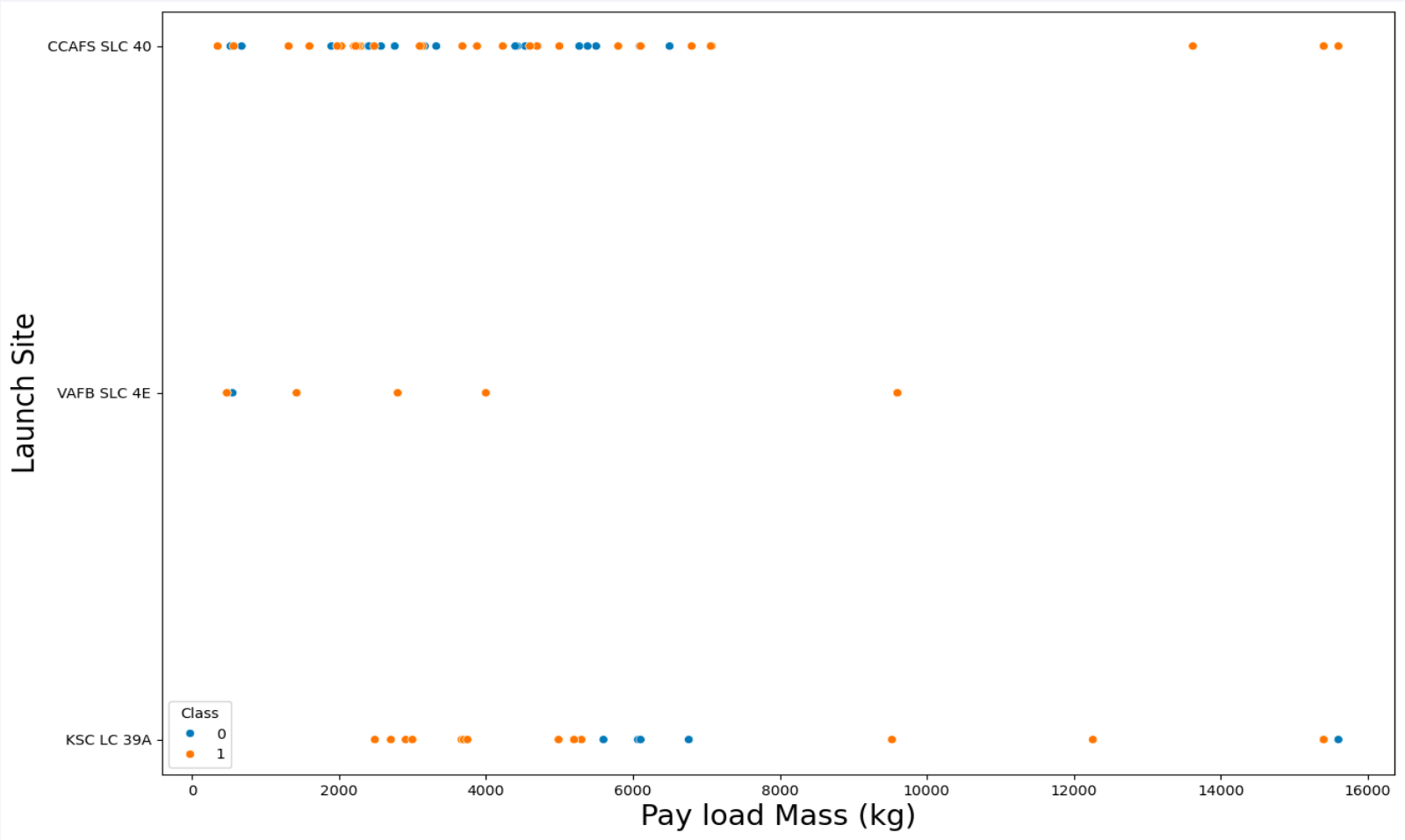
Flight Number vs. Launch Site

- Early SpaceX flights took place in the CCAFS SLC 40 launch site
- CCAFS SLC 40 has the most number of SpaceX launches
- Earlier flights have more failure rate



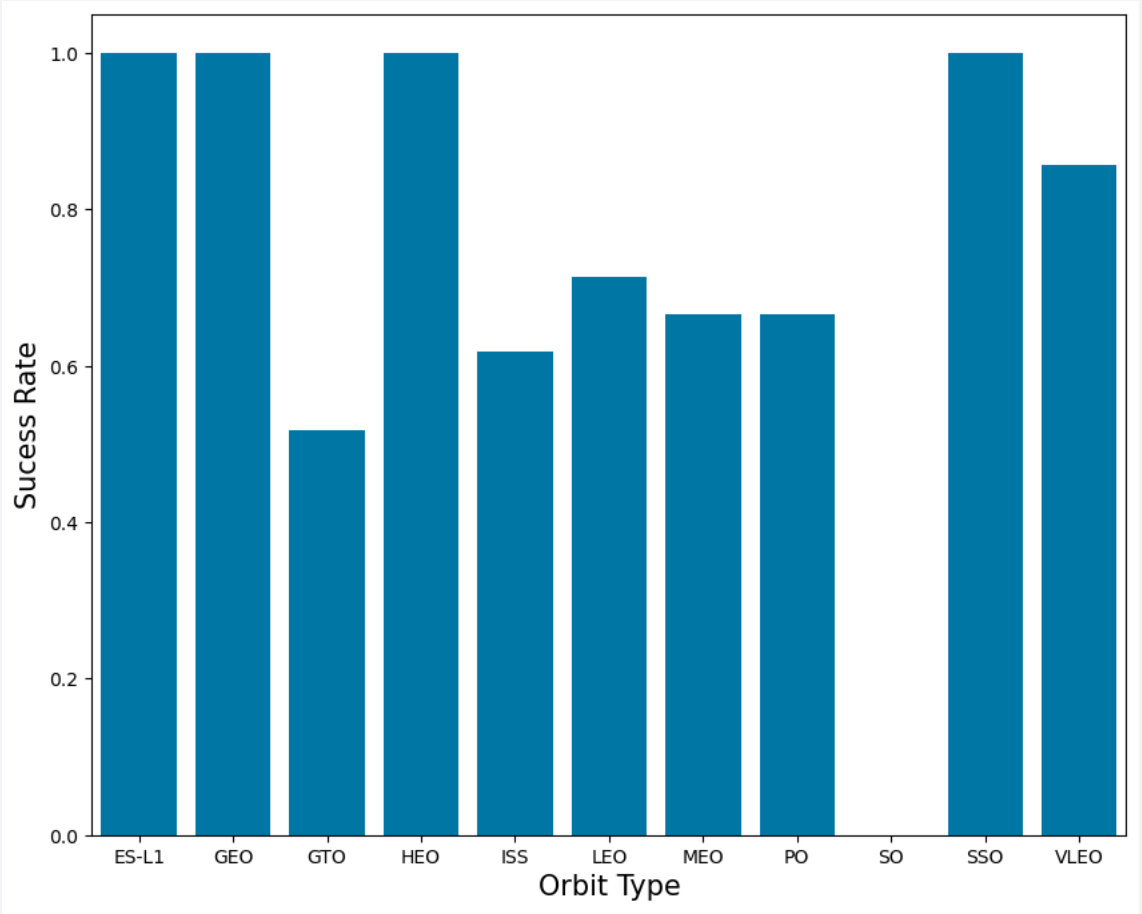
Payload vs. Launch Site

- Most of the flights had less than 8000 kg payload.
- The VAFB-SLC 4E launch site has the minimum number of launches and also the maximum payload for this site is only 10000 kg.



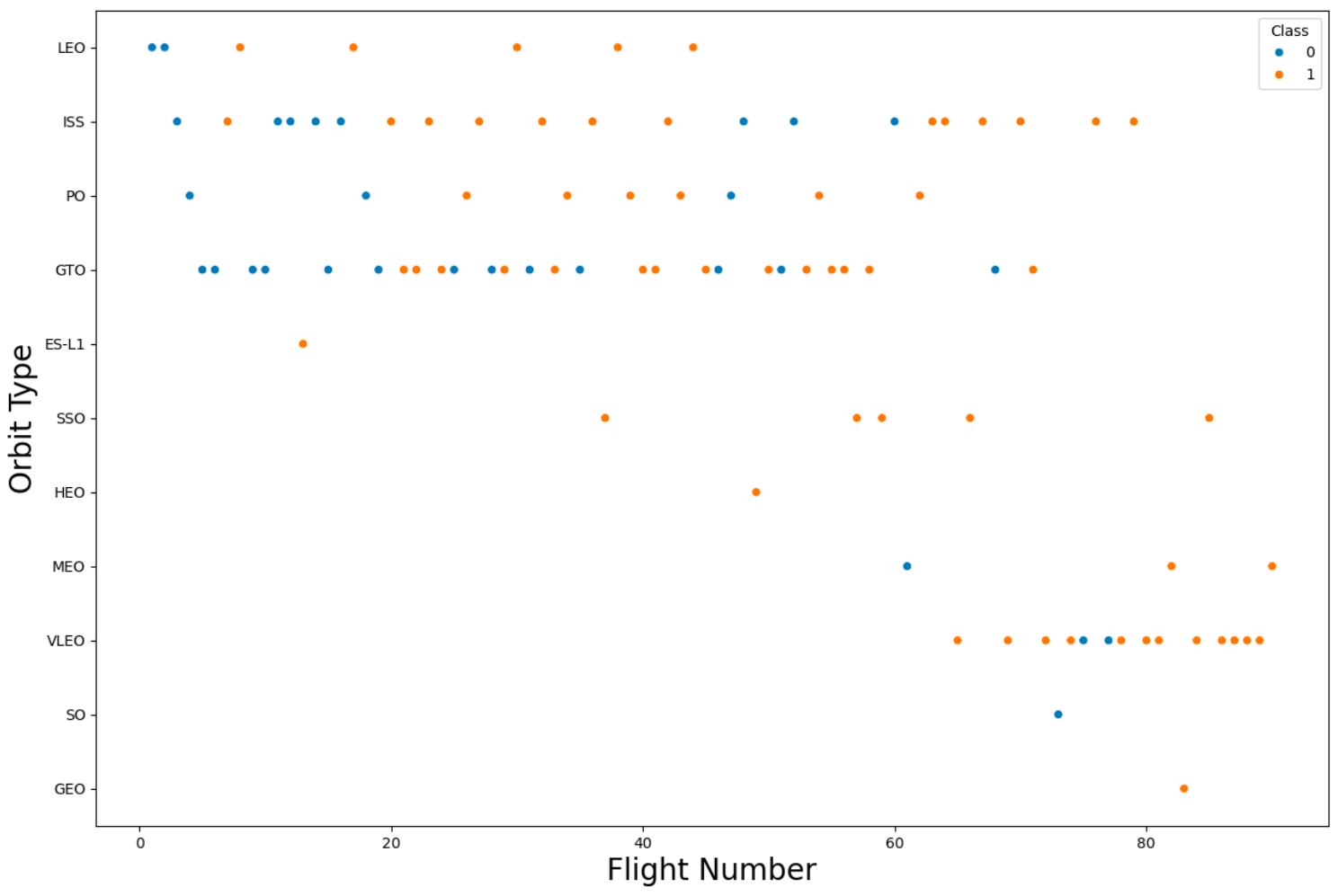
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO orbits have 100% success rate
- Flights to SO orbit failed to recover the first stage for any of the flights.
  (0 success rate)
- GTO, ISS, Leo, MEO, PO, VLEO orbits have a success rate between 0.45 and 0.85



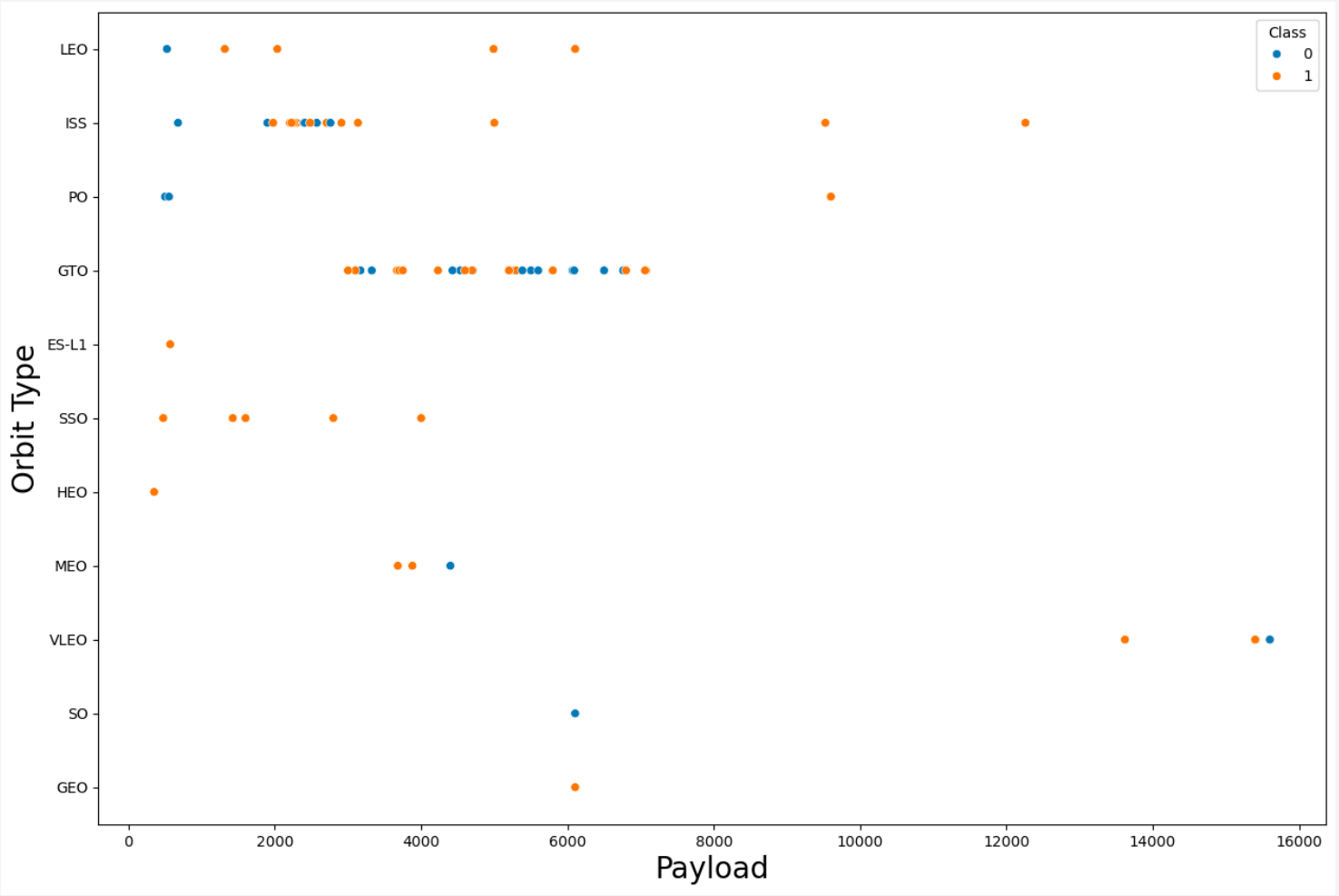
Flight Number vs. Orbit Type

- The success rate of the flights increased with the number of flights. i.e. At the early stage of SpaceX missions, the chance of failure to recover the first stage was high
- Early flights were performed to LEO, ISS, PO, and GTO orbits
- There was only one flight to the ES-L1 orbit



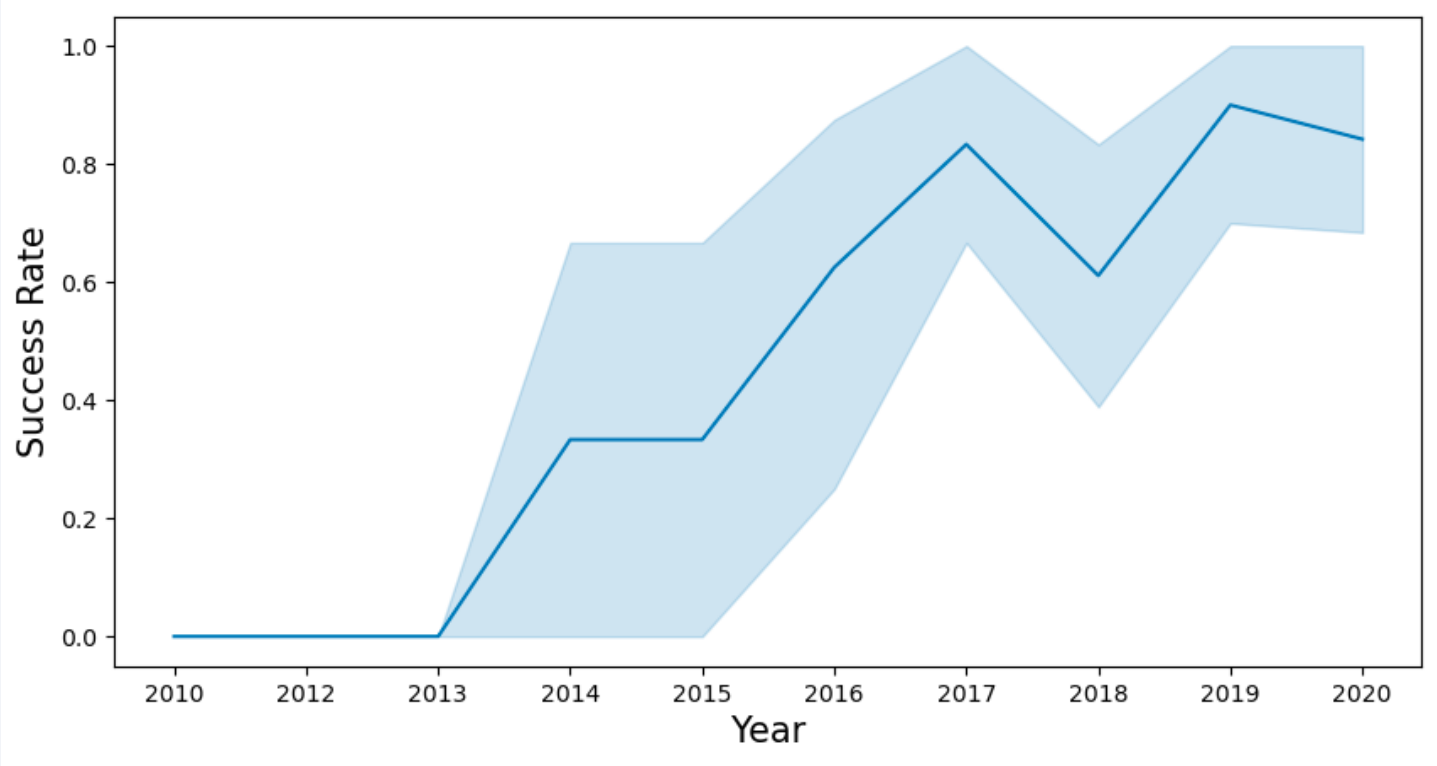
Payload vs. Orbit Type

- With heavy payloads, the successful landing or positive landing rate is higher for Polar, LEO and ISS.
- However, for GTO, we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are there.
- GTO orbit has mixed success but with a relatively low range of payloads
- The highest payload is recorded in VLEO orbit where 2 flights were successfully recovered, 1 unsucessfull.



Launch Success Yearly Trend

- There was no success from 2010 to 2013
- The success rate since 2013 kept increasing till 2017 (stable in 2014) , and after 2015, it started increasing.
- The success rate decreased from 2017-2018, but the increasing trend continues.



All Launch Site Names

All Site Names

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

```
%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

- Select Distinct is used to avoid repetition

Launch Site Names Begin with 'CCA'

```
 * 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

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5
```

- Launch_Site LIKE “CCA%” is used in the query to find all the records with the launch site beginning with CCA
- LIMIT 5 is used to only get 5 records

Total Payload Mass

The total payload carried by boosters from NASA

48213 kg

```
%%sql
SELECT SUM([PAYLOAD_MASS__KG_]) AS [Total_Payload_Mass_NASA]
FROM SPACEXTABLE
WHERE Customer LIKE "%NASA (CRS)%"
```

- SUM is used to calculate the total
- The calculation is based only on the records WHERE Customer is NASA

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

Total_Payload_Mass_NASA
48213

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1

2534.66 kg

```
%%sql
SELECT AVG([PAYLOAD_MASS__KG_]) AS [Average_Payload_Mass]
FROM SPACEXTABLE
WHERE Booster_Version LIKE "%F9 v1.1%"
```

- AVG is used to calculate the average
- Query results are limited to only consider booster version F9 v1.1

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

Average_Payload_Mass
2534.6666666666665

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad

2015-12-22

```
%%sql
SELECT MIN("Date")
FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (ground pad)";
```

- MIN is used to get the first date
- Query is limited to output successful ground pad landings

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

MIN("Date")
2015-12-22
```

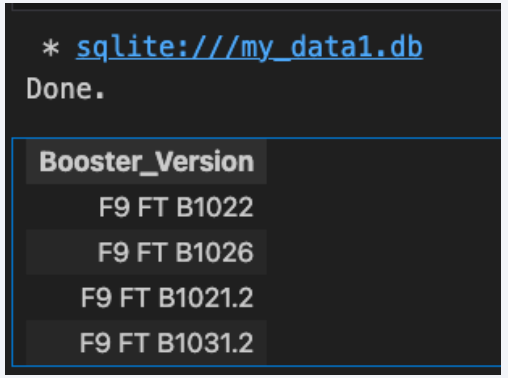
Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

```
%%sql
SELECT DISTINCT [Booster_Version]
FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (drone ship)" AND
PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

- DISTINCT is used to avoid repetition
- The query is filtered to give successful drone ship landing outcomes
- BETWEEN is used to limit the output to give results between 4000 and 6000 kgs of payload masses.



Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

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

```
%%sql
SELECT [Mission_Outcome], COUNT(*) AS [Num_Outcome]
FROM SPACEXTABLE
GROUP BY [Mission_Outcome];
```

- Aggregate function COUNT is used to get the total number of outcomes
- The outcome is categorised by MISSION_OUTCOME column

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass

- 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);
```

Sub Query is used to get the maximum payload carried by any of SpaceX flights

The query is limited to give the booster names that have carried the payload obtained from the sub-query.

```
 * 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

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
%%sql
SELECT substr(Date,6,2) as month, DATE, BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome]
FROM SPACEXTBL
where [Landing_Outcome] = "Failure (drone ship)" and substr(Date,0,5)="2015";
```

- SUBSTR function is used to get the month from data
- Again, SUBSTR function is used to get the year from date
- Conditions were provided to limit the failed results that occurred in 2015

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

Rank of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

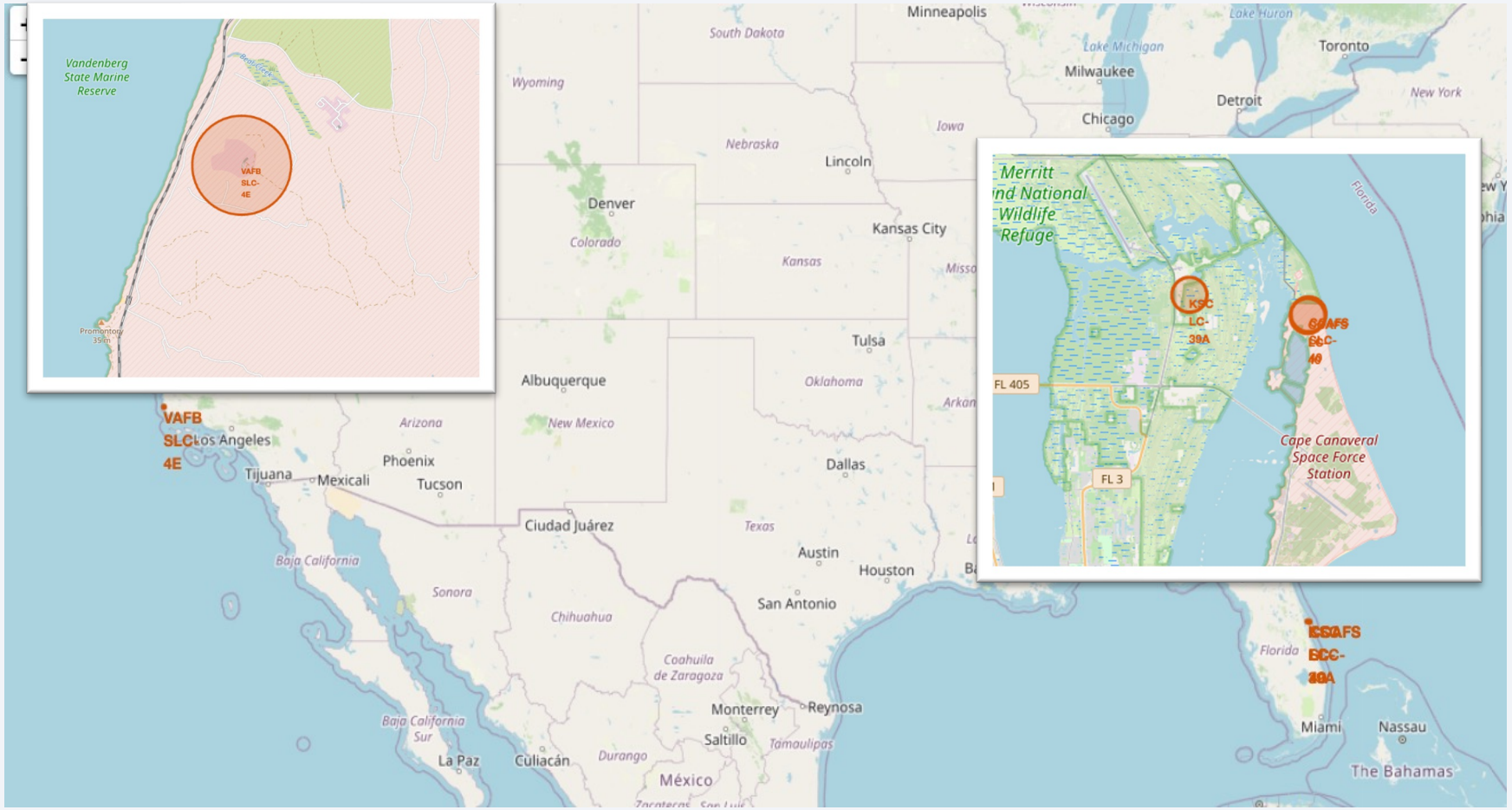
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

```
%%sql
SELECT [Landing_Outcome], count(*) as count_outcomes
FROM SPACEXTBL
WHERE DATE BETWEEN "2010-06-04" AND "2017-03-20 "
GROUP BY [Landing_Outcome]
ORDER BY count_outcomes DESC;
```

- BETWEEN is used to get the results limited to required period
- The results are grouped by landing outcome
- ORDER BY is used to arrange the query results according to the count
- DESC is used to get the results in descending order

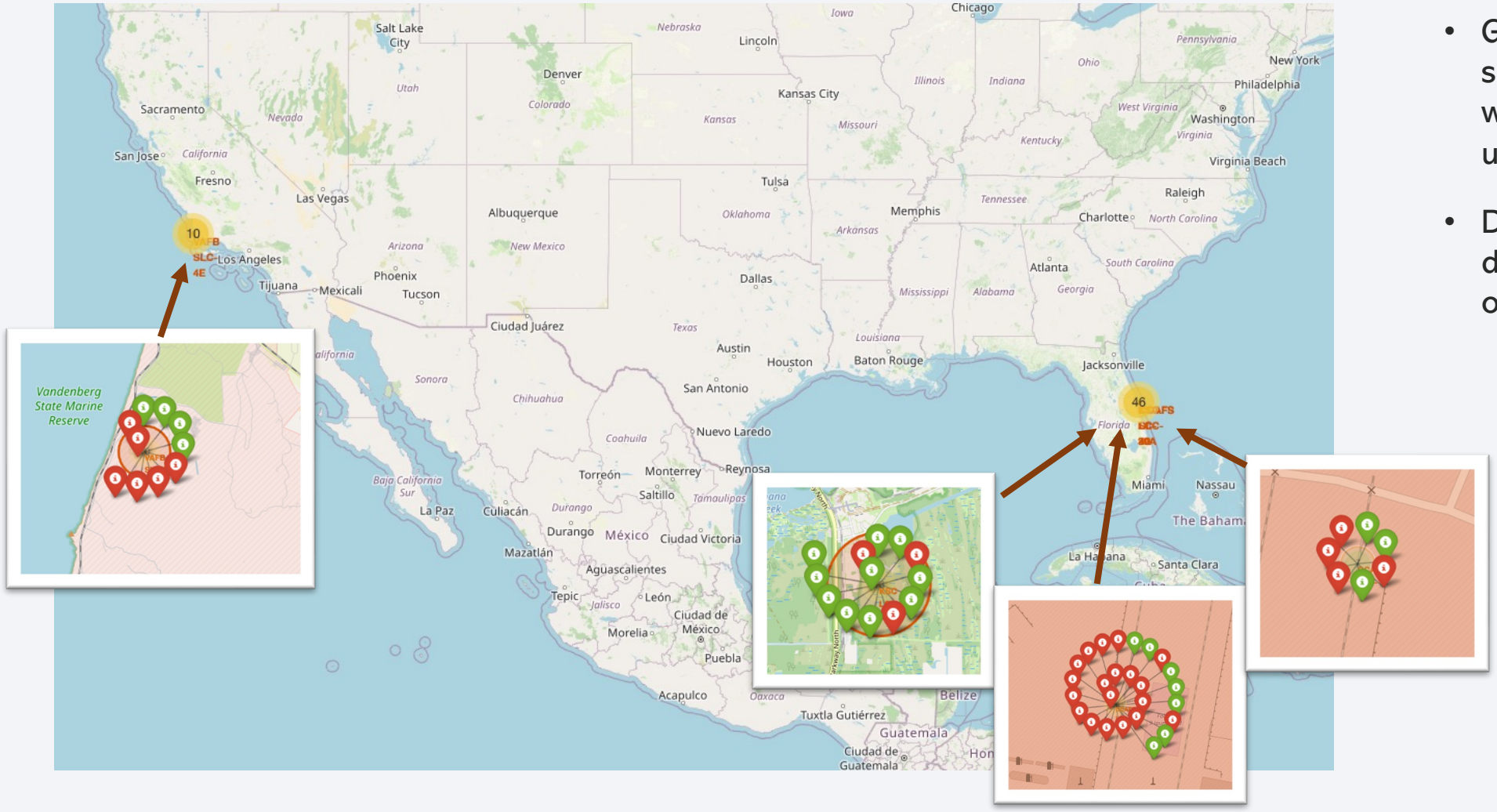

Section 3
Launch Sites
Proximities Analysis

Launch Site Map



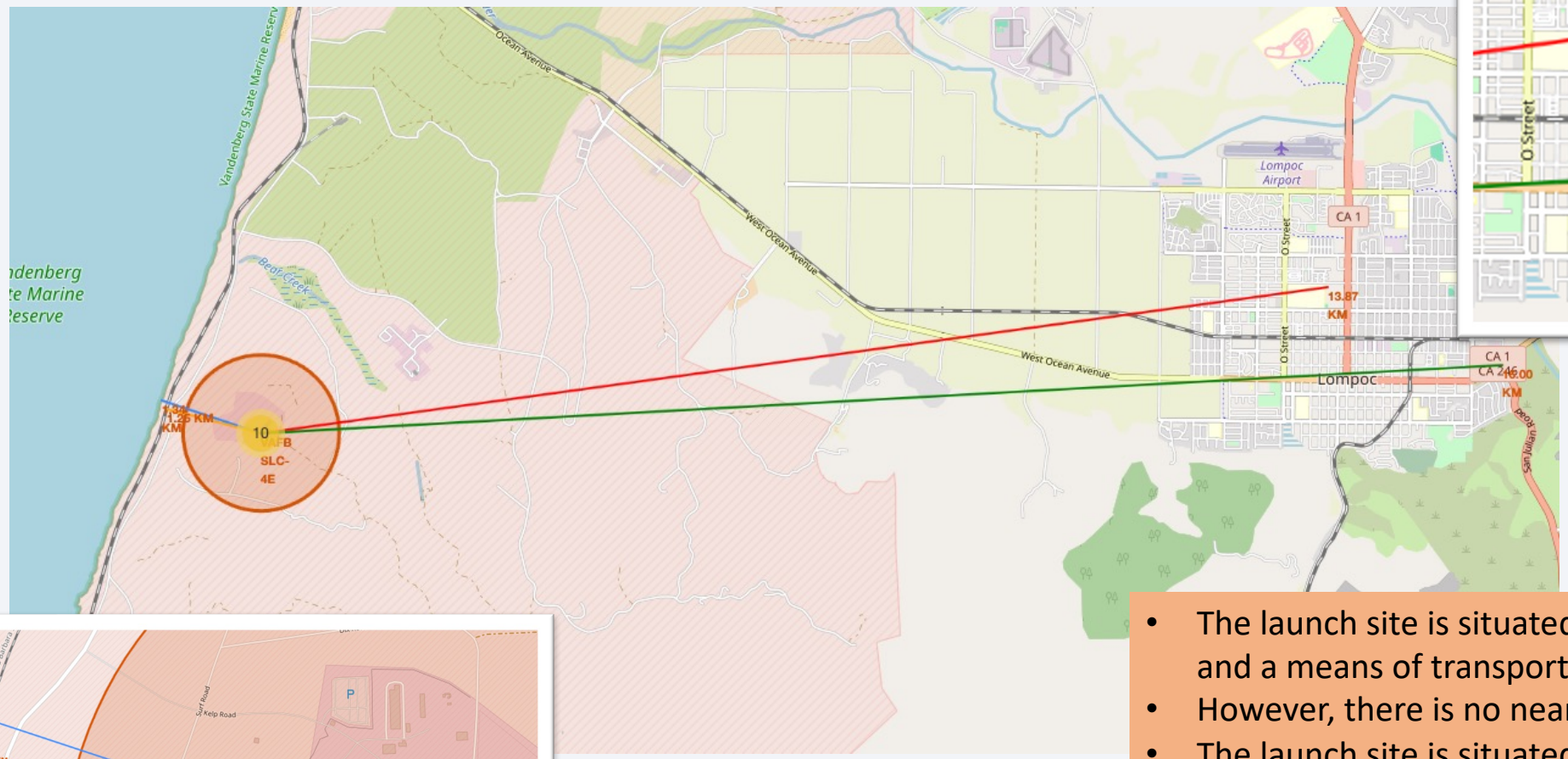
- All Sites are located near coast line and remote areas to minimize civil damage
- All launch sites are closer to the equator to get more help from the earths rotation energy

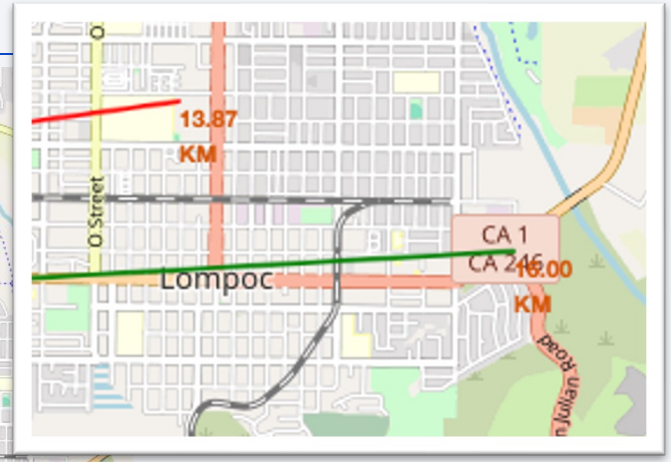
Map of Launch Outcomes



- Green markers represent successful outcomes while red is unsuccessful.
- Different sites have different successful outcome rates

VAFB SLC-4E Launch Site Proximity Map

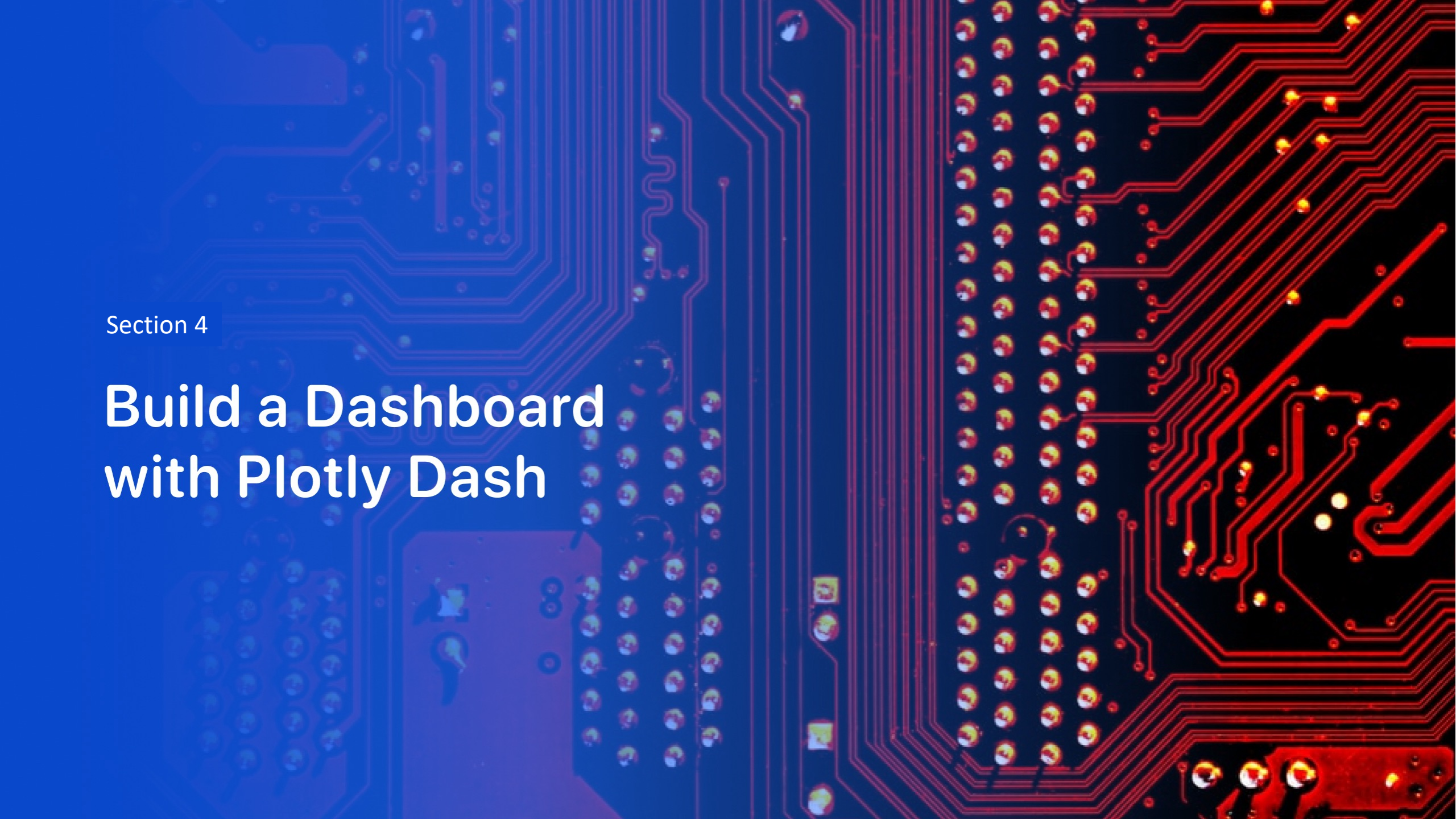




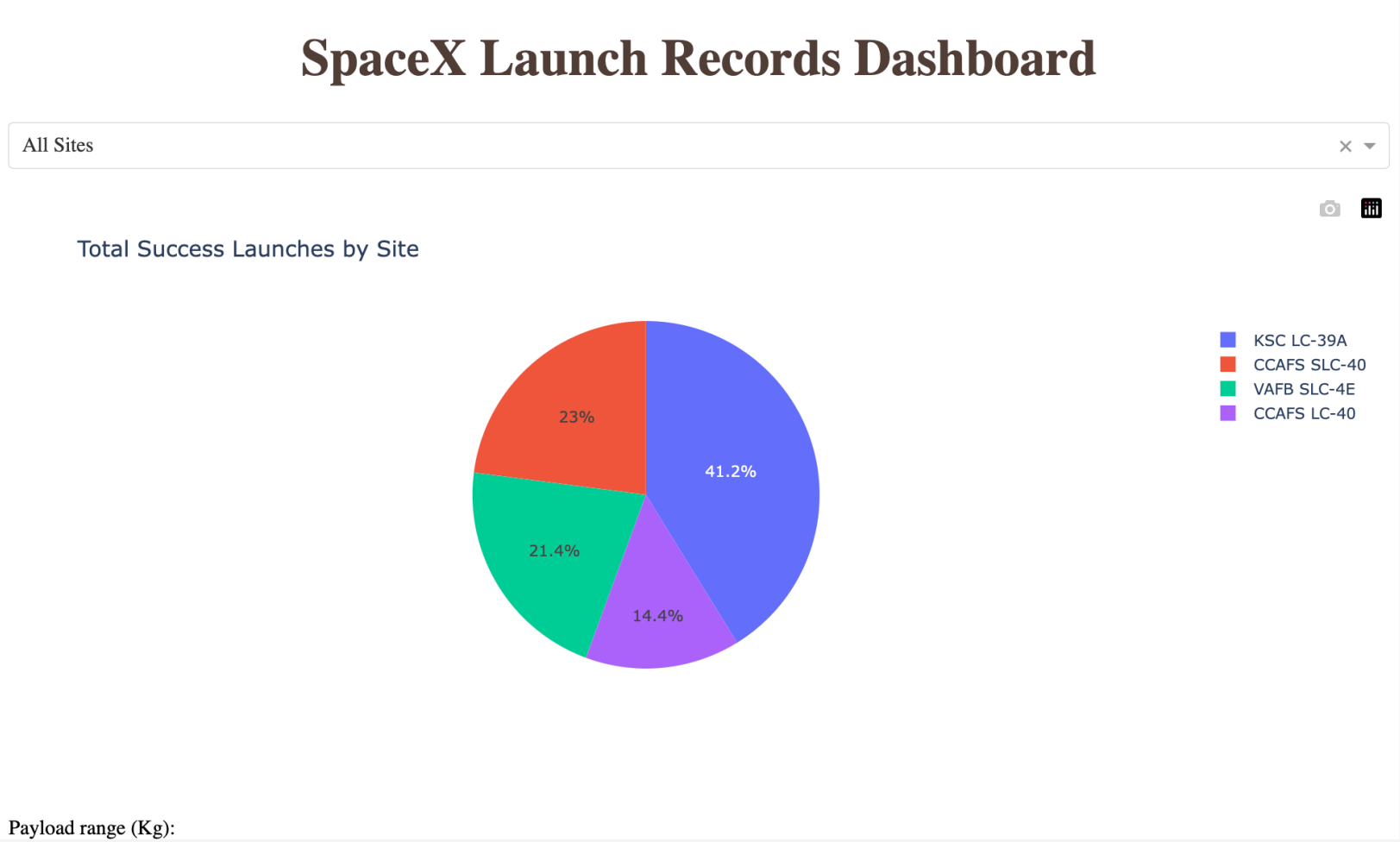
Distance to City:
13.37 km
Distance to Highway:
16.00 km

Distance to Coastline: **1.34 km**
Distance to Railway: **1.25 km**

- The launch site is situated very near to the coastline and a means of transportation (railway)
- However, there is no nearby highway
- The launch site is situated 16 km away from city center considering safty reasons

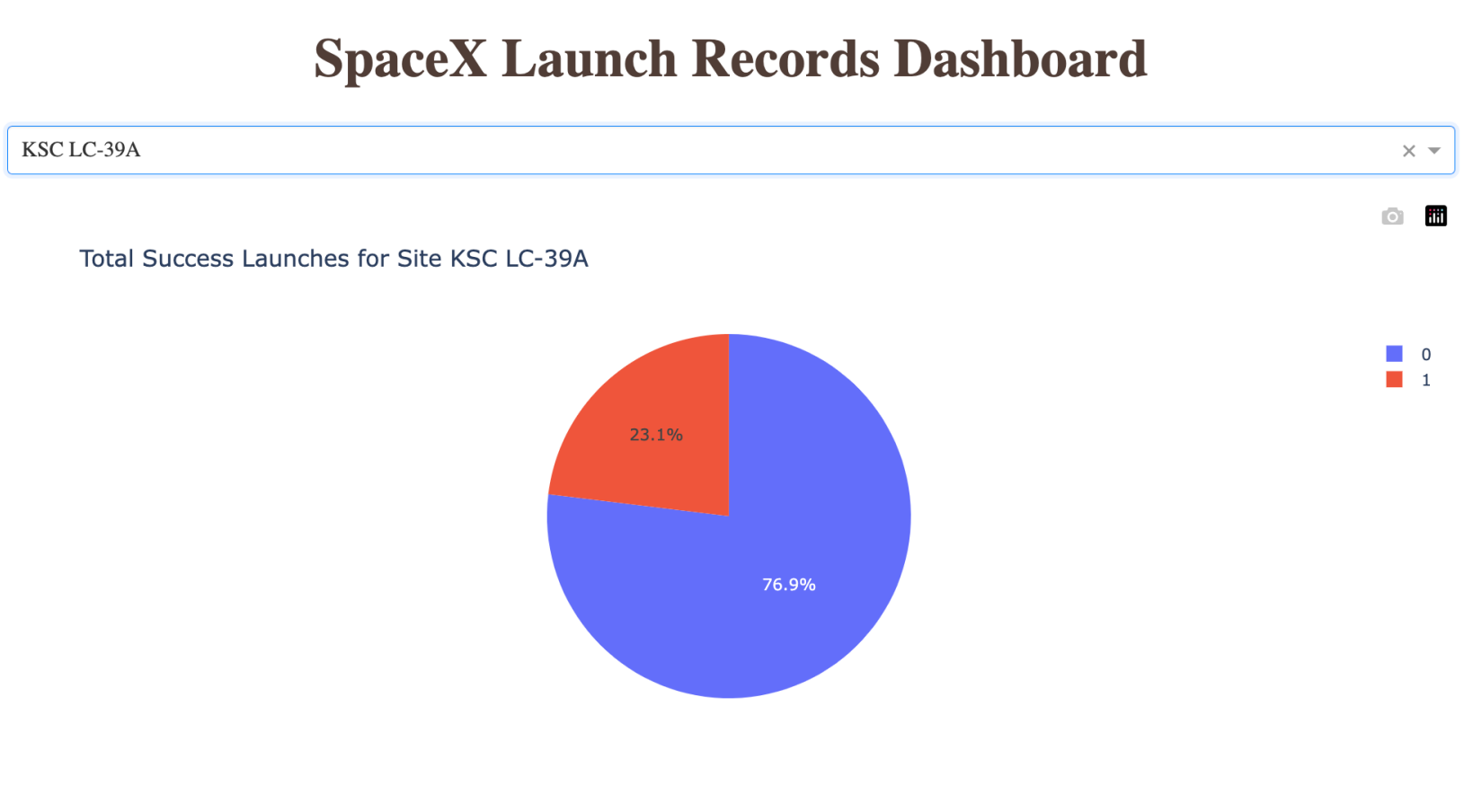

Section 4
Build a Dashboard with Plotly Dash

Launch Site Success Rate



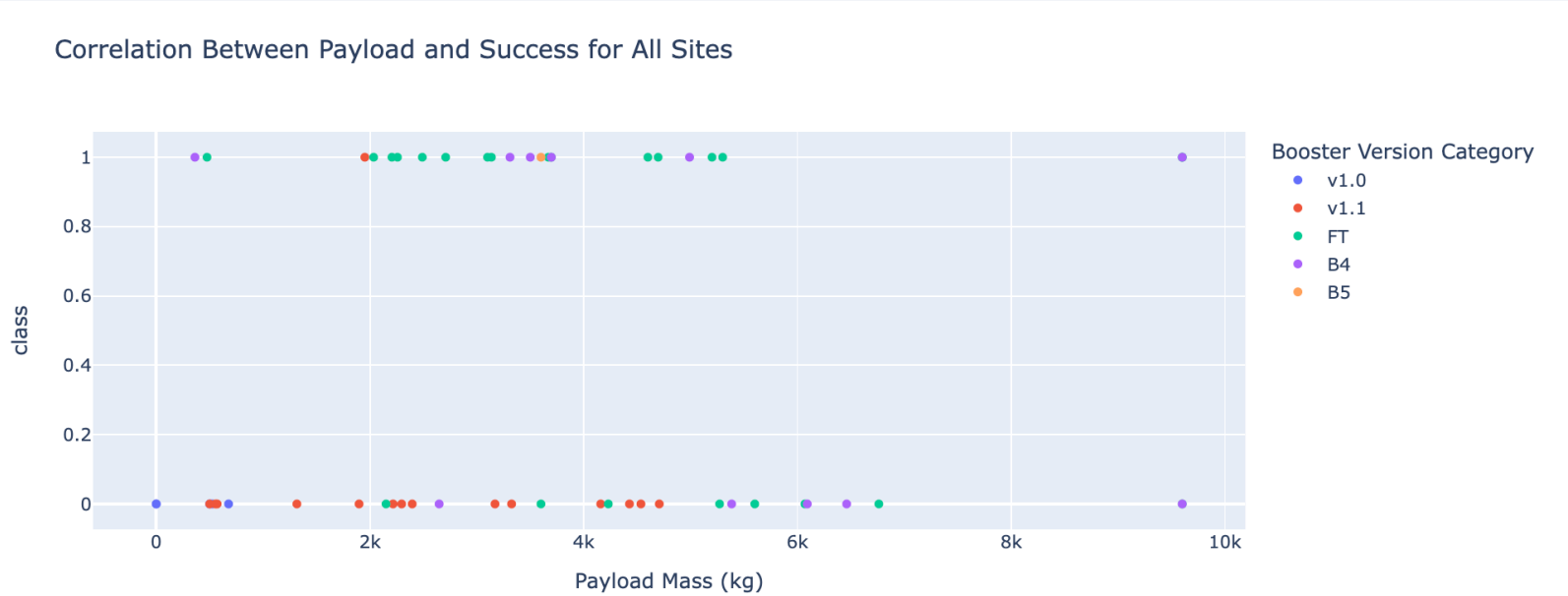
- KSC LC 39A launch site has the highest success rate of 41.2%

Total Success Launches : KSL LC 29A



- KSL LC 29A is the launch site with the highest launch success ratio.
- Among all flig ht s KSL LC 29A had, 76.9% of the flights were sucessfull

Payload vs. Launch Outcome

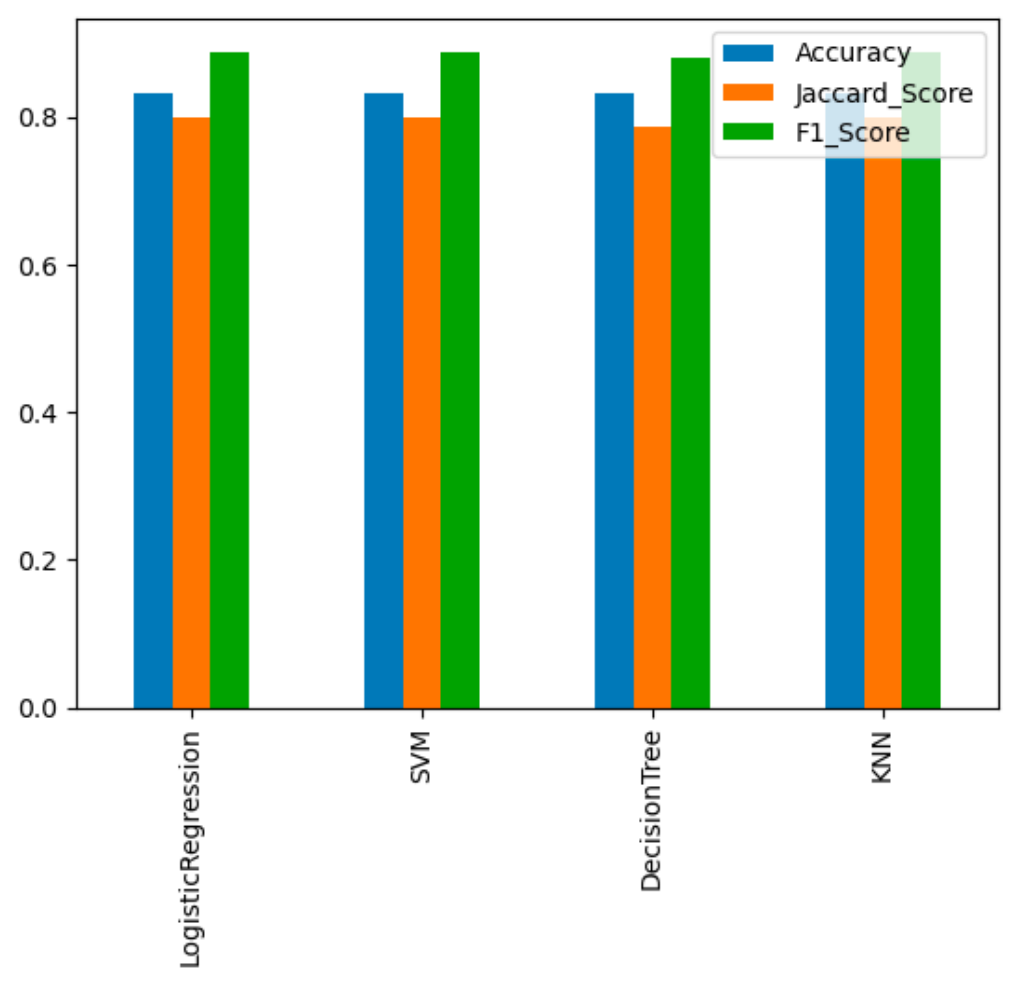


- 1 indicate a successful outcome and 0 is unsuccessful
- Payloads between 2000 kg and 5000kg have the highest success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy



	LogisticRegression	SVM	DecisionTree	KNN
Accuracy	0.833	0.833	0.833	0.833
Jaccard_Score	0.800	0.800	0.786	0.800
F1_Score	0.889	0.889	0.880	0.889

- 4 ML models were compared
- SVM, Logistic Regression, KNN have same classification accuracy in terms of raw accuracy, F1 score, and Jaccard coefficient
- However Decision Tree shows slightly less accuracy

Confusion Matrix: KNN

Confusion Matrix Summarize the performance of classification algorithm

For KNN

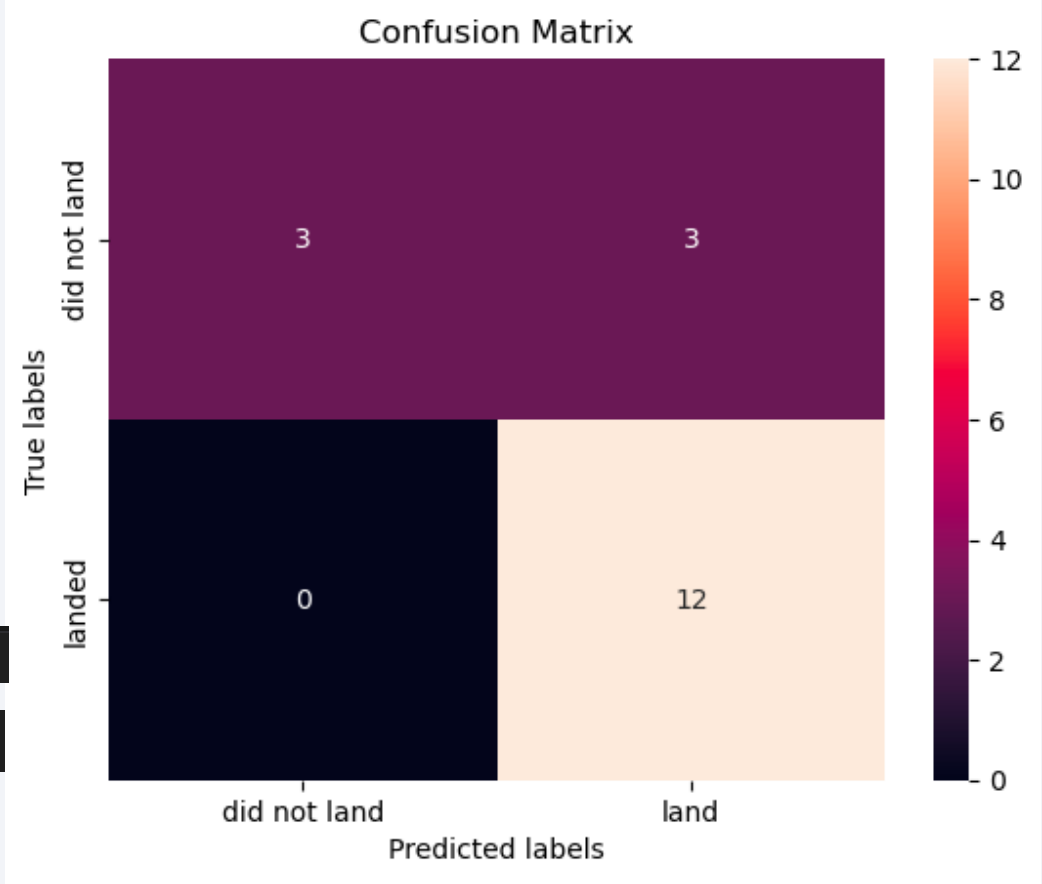
- 12 True Positives
- 3 True Negative
- 3 False Positives
- 0 False Negatives

Precision = 12/15 = 0.8,

Recall = 12/12 = 1

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
```

```
accuracy (KNN) on test data: 0.8333333333333334
```



Conclusions

- **Mission Success:**
  - Mission success increases over time. Earlier missions had a high probability of failure to recover the first stage of the Falcon 9 booster, but the success rate drastically increased in the later missions.
  - After 2019, the success rate was over 80%.
  - The success rate depends on the orbit. ES L1, GEO, HEO, and SSO have a 100% success rate.
- **Launch Sites:**
  - All launch sites were located near the coastline away from the city centre, but in close proximity to a means of transportation.
  - All launch sites were near the equator.
- **Prediction Model:**
  - With the given data, it was possible to train a machine learning model and predict the successful outcome with 0.833 accuracy
  - However, only 90 records were used to train the model, and data from up to 2020 was used in this analysis. The model's accuracy may increase if we incorporate the data from recent years.

Appendix

- Link to GitHub Repo:

  https://github.com/sasankamadawalagama/ibm_ds_coursera

- Reference:

Getting Started with Data Science: Making Sense of Data with Analytics: Making Sense of Data with Analytics (IBM Press)

Thank you!