IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

ZIYUAN SUI
07/01/2022

Outline

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

Executive Summary

- The following methodologies were used to analyze data:
  - Data Collection using web scraping and SpaceX API;
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
  - Machine Learning Prediction.
- Summary of all results
  - It was possible to collected valuable data from public sources;
  - EDA allowed to identify which features are the best to predict success of launchings;
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

- The objective is to evaluate the viability of the new company Space Y to compete with SpaceX.
- The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;
- Where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

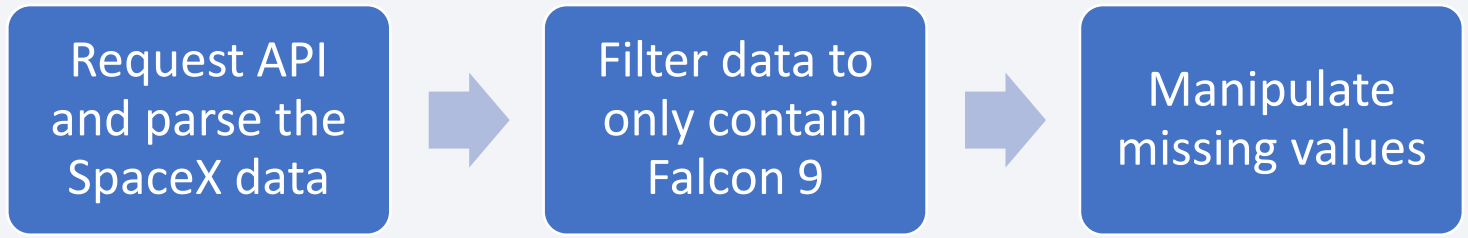
- Data collection methodology:
  - Data from SpaceX was obtained from 2 sources:
  - SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL

Data Collection

- Data were collected from SpaceX API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches) by using web scraping.

Data Collection – SpaceX API

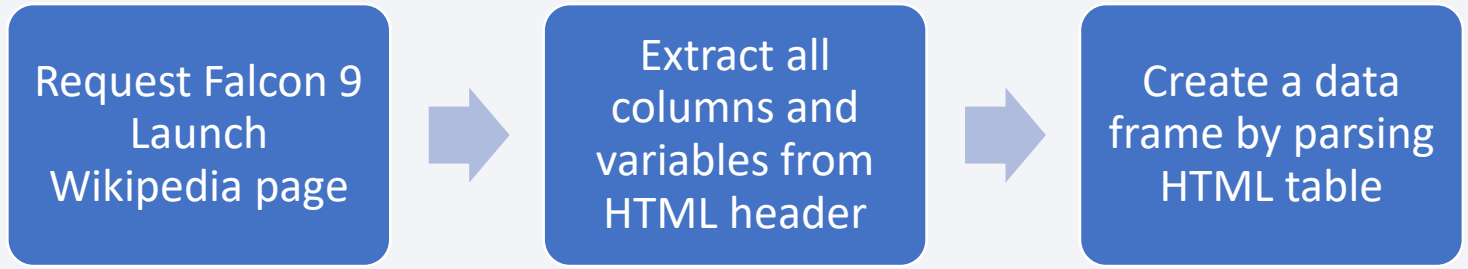
- SpaceX public API can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.



https://github.com/zaynsui625/Coursera/blob/main/IBM%20Data%20Science/Capstone/Data%20Collection%20API.ipynb

Data Collection - Scraping

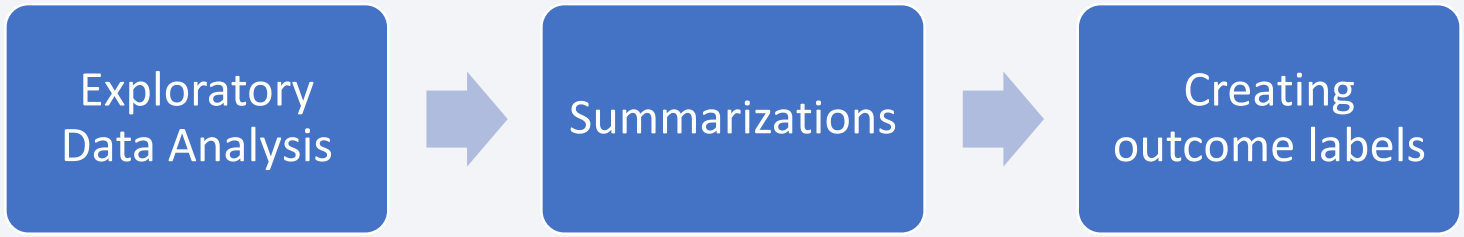
- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.



https://github.com/zaynsui625/Coursera/blob/main/IBM%20Data%20Science/Capstone/Data%20Collection%20Web%20Scraping.ipynb

Data Wrangling

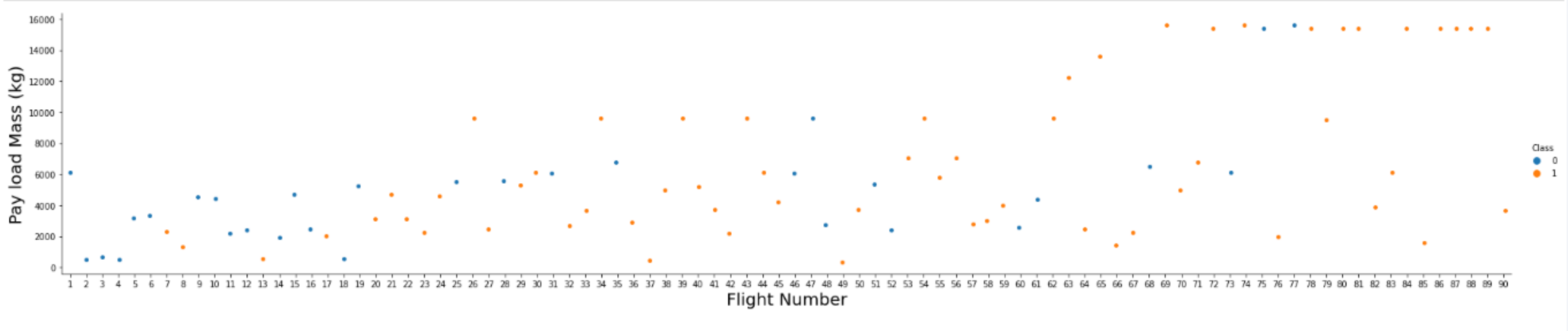
- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were formed.
- Finally, the landing outcome label was created from Outcome column.



https://github.com/zaynsui625/Coursera/blob/main/IBM%20Data%20Science/Capstone/Data%20Wrangling%20EDA.ipynb

EDA with Data Visualization

- To explore SpaceX data, scatter plots, cat plots and bar plots were used to visualize the relationship between pair of features:
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



https://github.com/zaynsui625/Coursera/blob/main/IBM%20Data%20Science/Capstone/EDA%20Data%20Visualization.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
- Markers indicate points like launch sites;
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
- Marker clusters indicates groups of events in each coordinate;
- Lines are used to indicate distances between two coordinates.

https://github.com/zaynsui625/Coursera/blob/main/IBM%20Data%20Science/Capstone/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

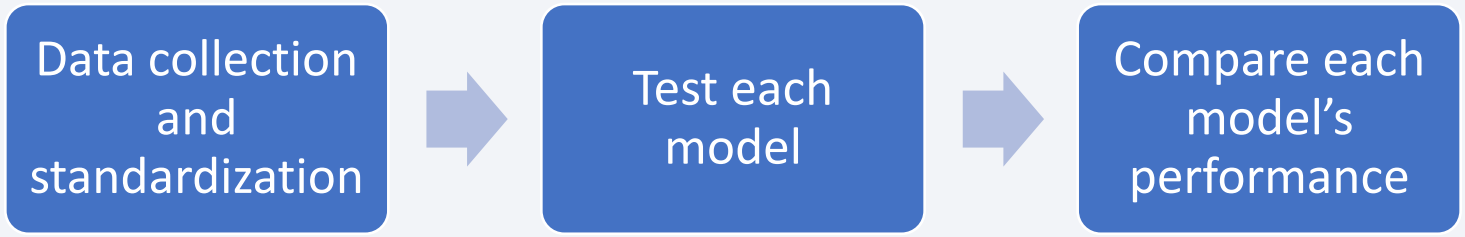
Build a Dashboard with Plotly Dash

- I built the following dashboards for analysis:
  - Percentage of launches by site
  - Payload range

https://github.com/zaynsui625/Coursera/blob/main/IBM%20Data%20Science/Capstone/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Predictive Analysis (Classification)

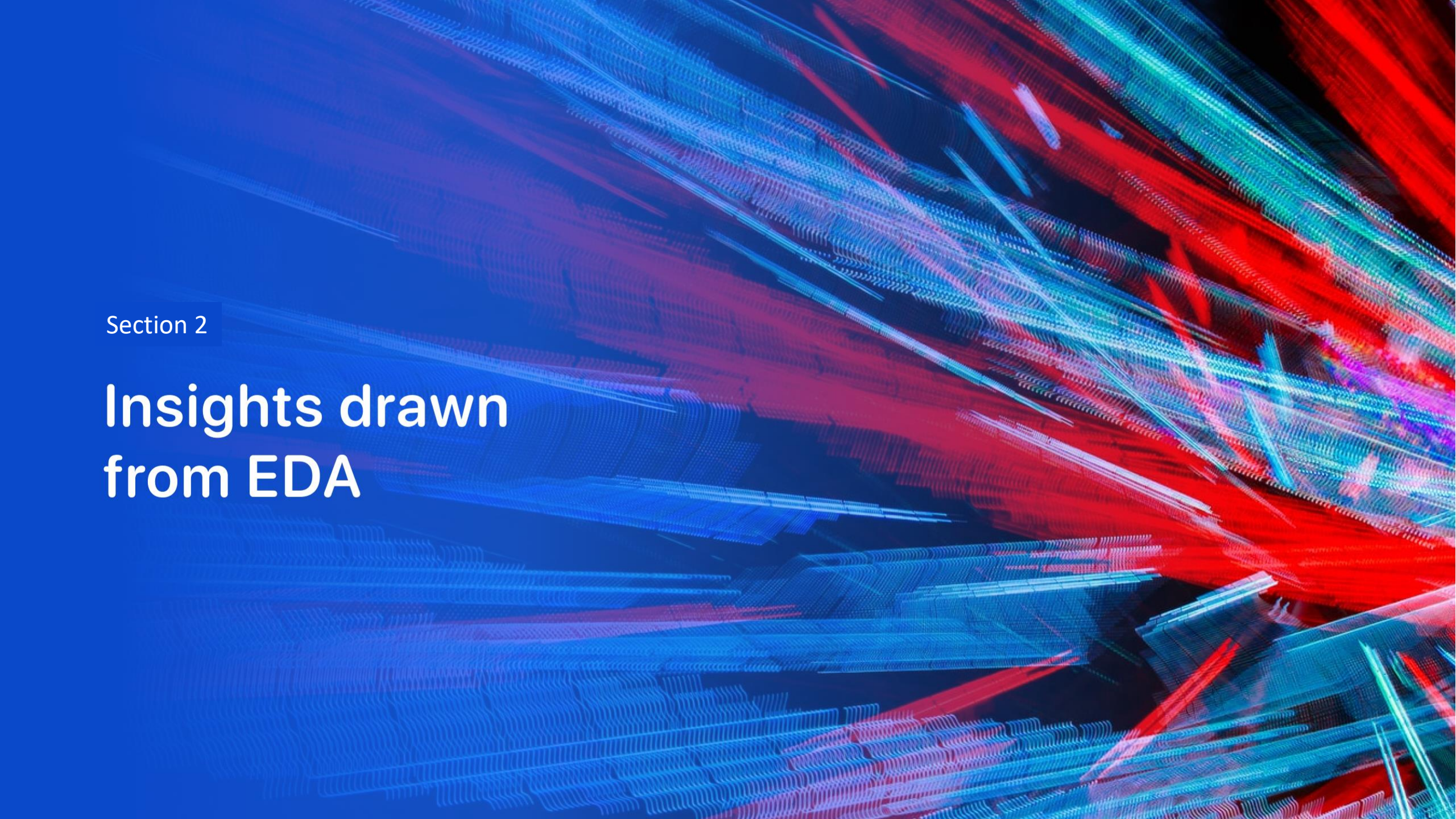
- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.



https://github.com/zaynsui625/Coursera/blob/main/IBM%20Data%20Science/Capstone/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

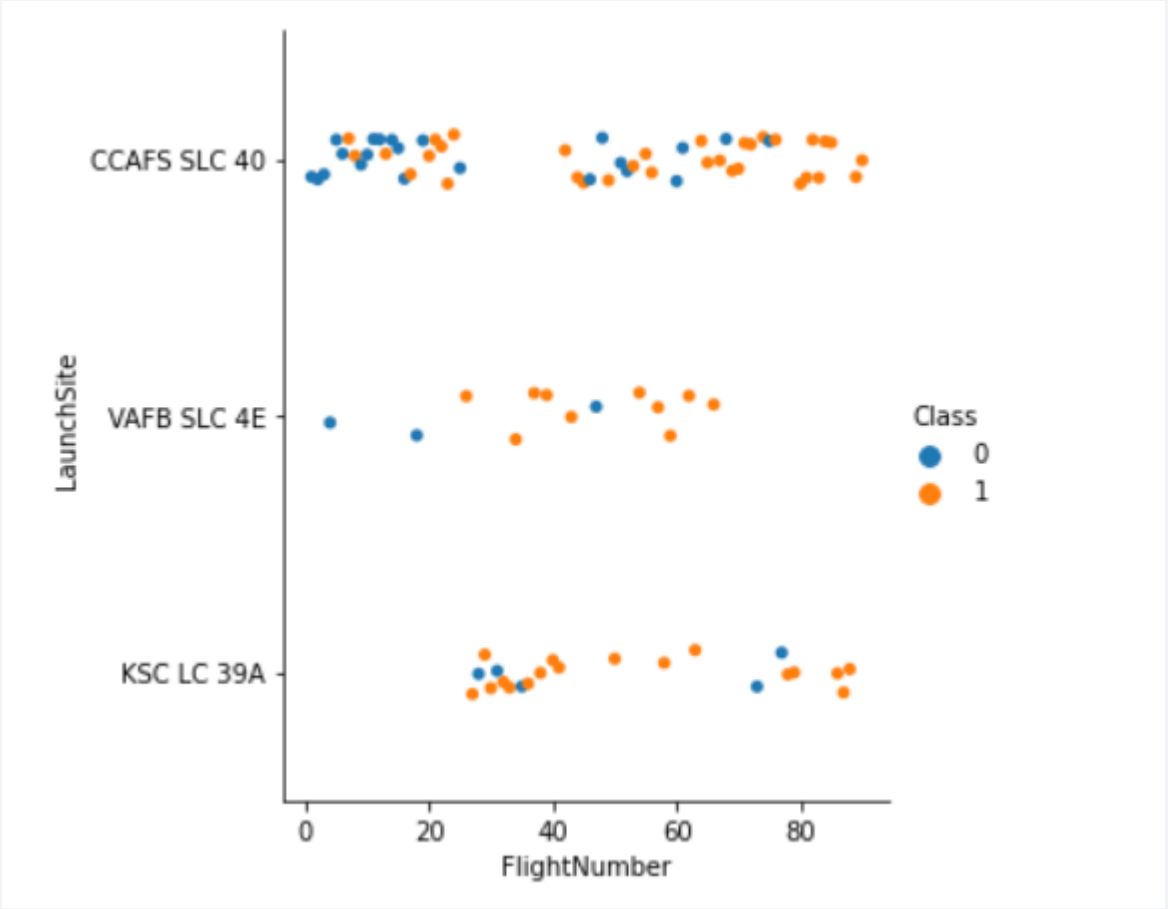
Results

- The best launch site is KSC LC 39A;
- Launches above 5,500kg are less risky;
- It turns out that most of mission outcomes are successful. The success rate was low in the first few years, but successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- SVM, Logistic Regression, and KNN can be used to predict successful landings and increase profits.


Section 2
Insights drawn from EDA

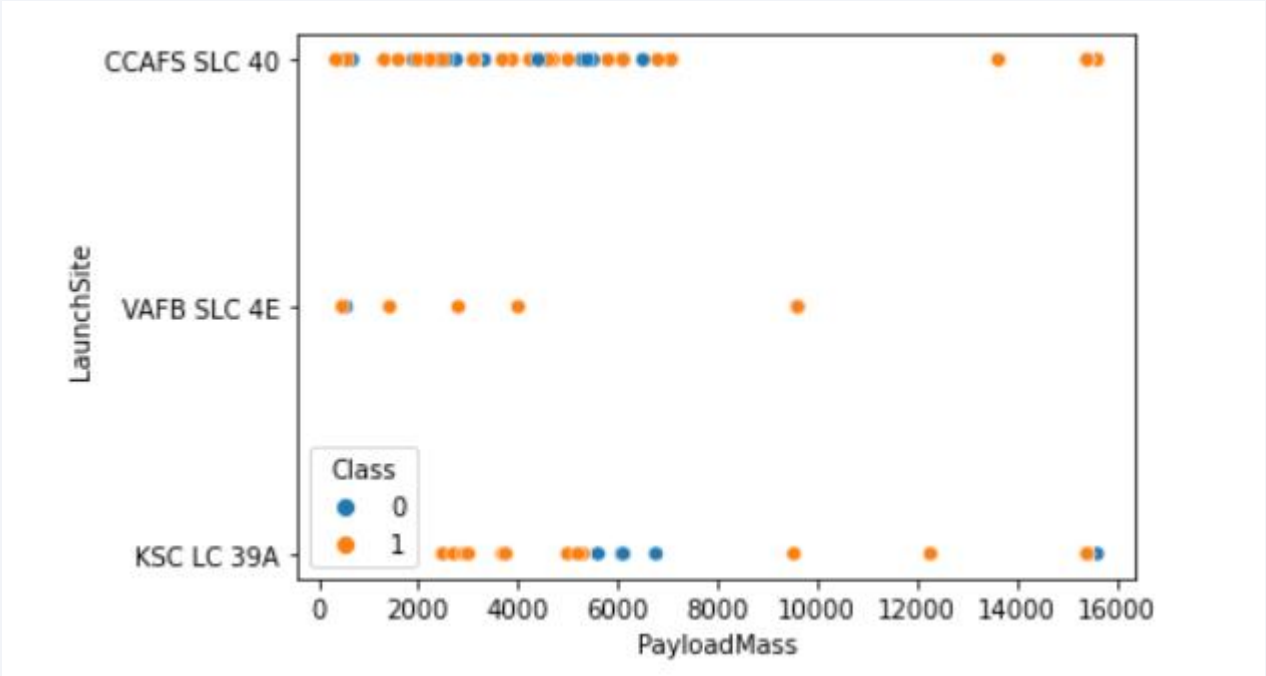
Flight Number vs. Launch Site

- According to the plot, CCAF5 SLC 40 launched the most rockets followed by KSC LC 39A and VAFB SLC 4E;
- The general success rate improves over time with the increase of the flight number.
- VAFB SLC 4E has the least number of failed cases, KSC LC 39A has the highest successful launching rate.



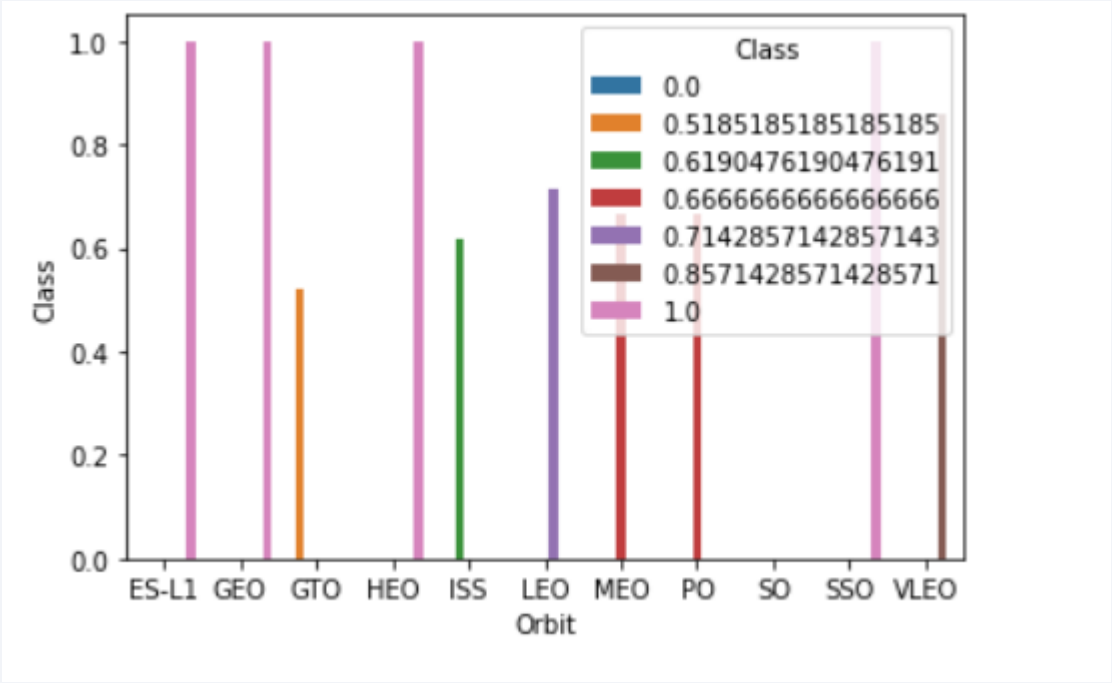
Payload vs. Launch Site

- Payloads over 8,000kg have a relatively high success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.
- Payloads under 5,500kg have higher success rates on KSC LC 39A and VAFB SLC 4E.



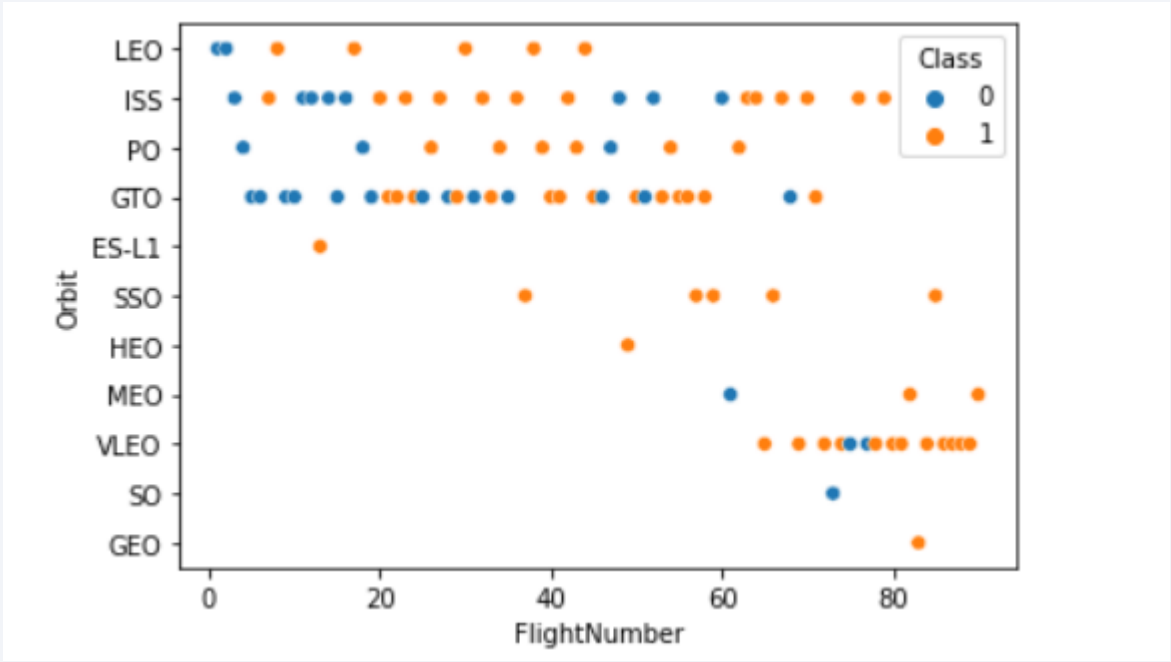
Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
- ES-L1& GEO & HEO & SSO.
- Followed by:
- VLEO (above 80%) and LFO (above 70%).



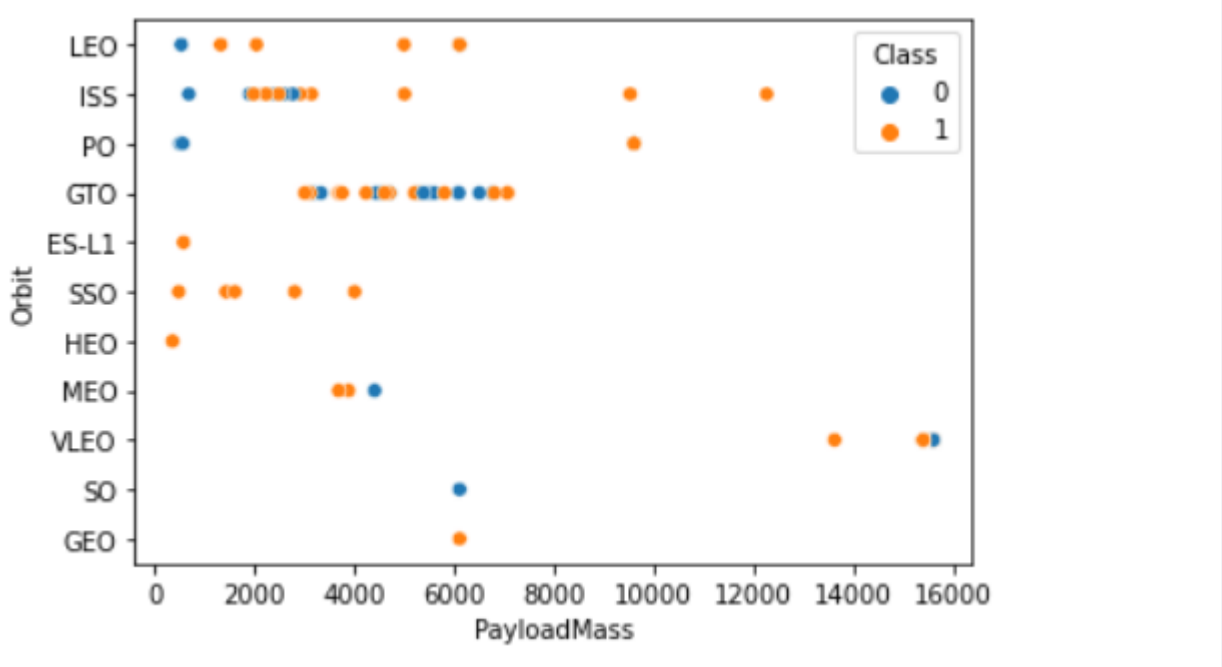
Flight Number vs. Orbit Type

- Success rate for each orbit increases over time with the increase of fight number.
- Orbit VLEO launches frequently and has a high success rate.
- Orbit SO seems to be given up after the first launch failure.



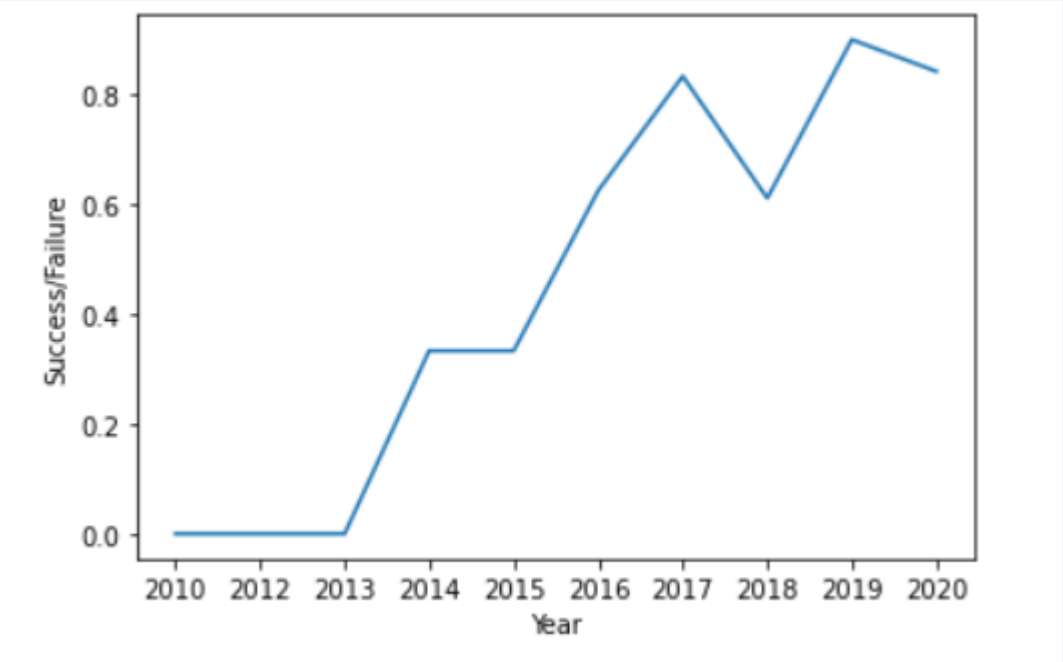
Payload vs. Orbit Type

- There is no relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success.
- VLEO, SO, GEO, HEO, ES-L1 have fixed payloads.



Launch Success Yearly Trend

- Success rate started increasing in 2013 and increased significantly until 2020;
- SpaceX was developing its models in the first three years leads to a zero-success rate.



All Launch Site Names

- According to data, there are four launch sites:

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

- They are derived by selecting distinct “launch_site” values from the dataset.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Total Payload (kg)
111.268

- Total payload was calculated by summing all payloads whose codes contain ‘CRS’.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

Avg Payload (kg)
2.928

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

Min Date
2015-12-22

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first occurrence, that happened on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes.

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

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

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- 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

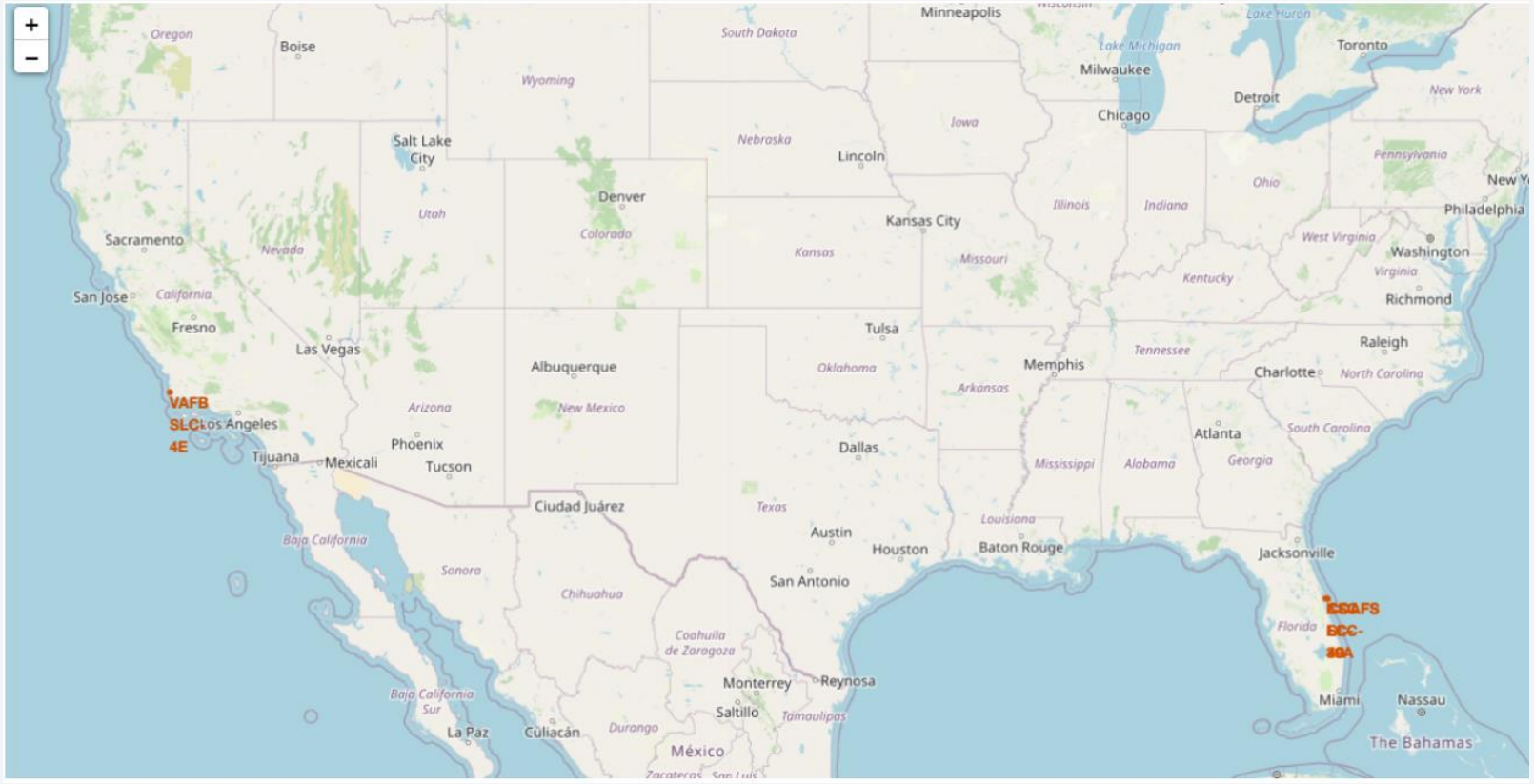
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

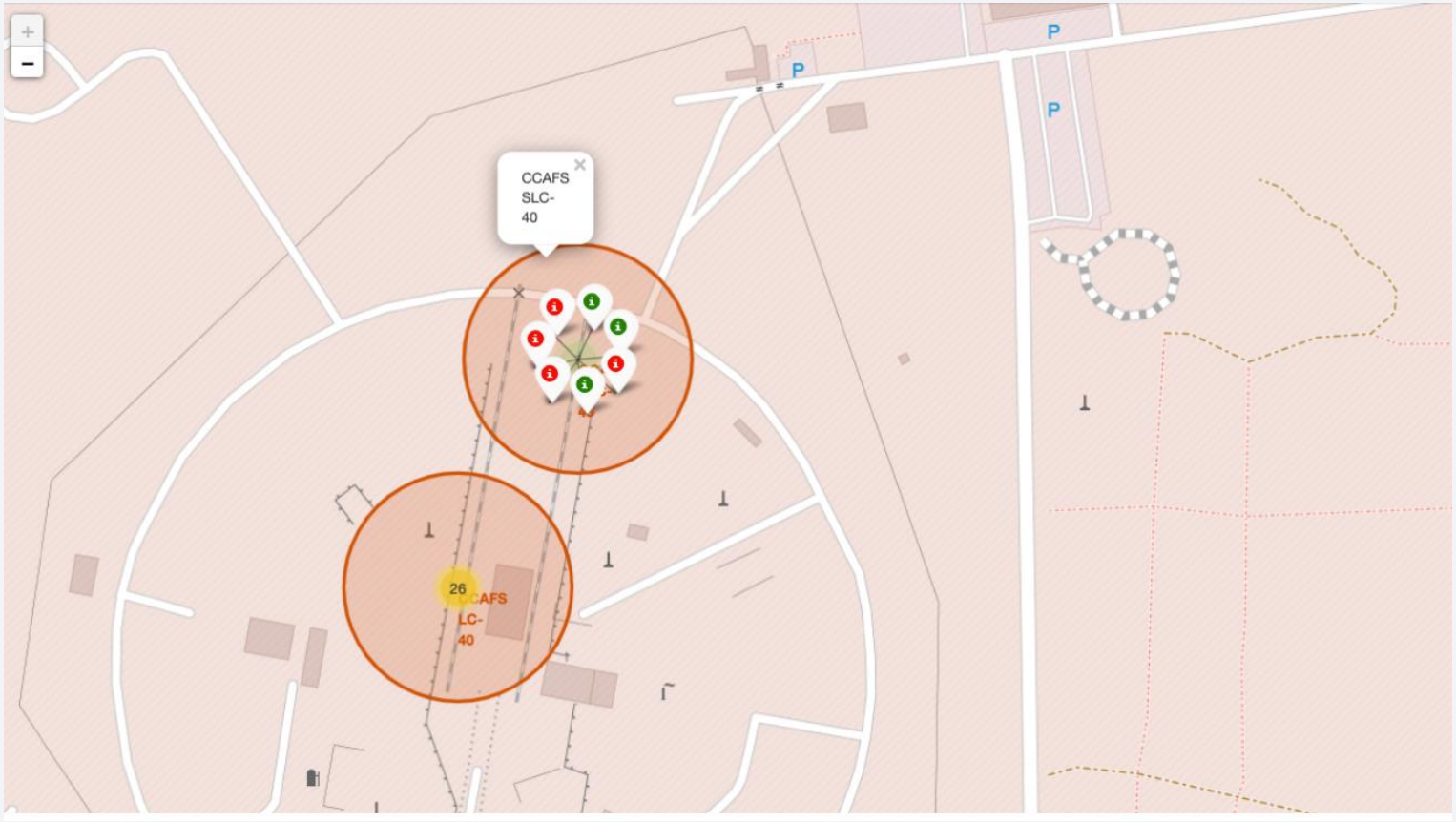
All launch sites

- Launch sites are near sea but not too far from roads and railroads.
- This is safe for launching and it is easy for the transportation.



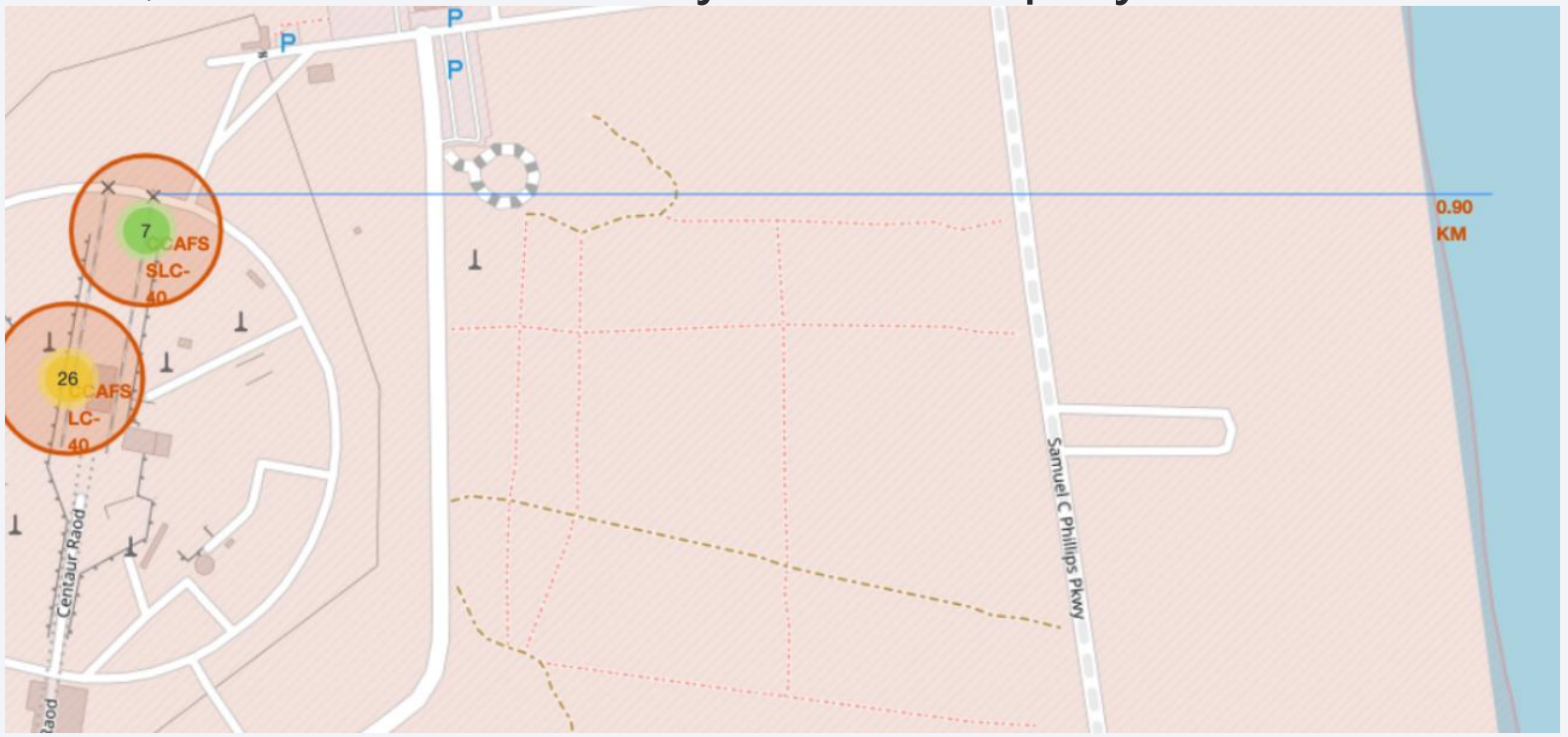
Launch Outcomes by Site

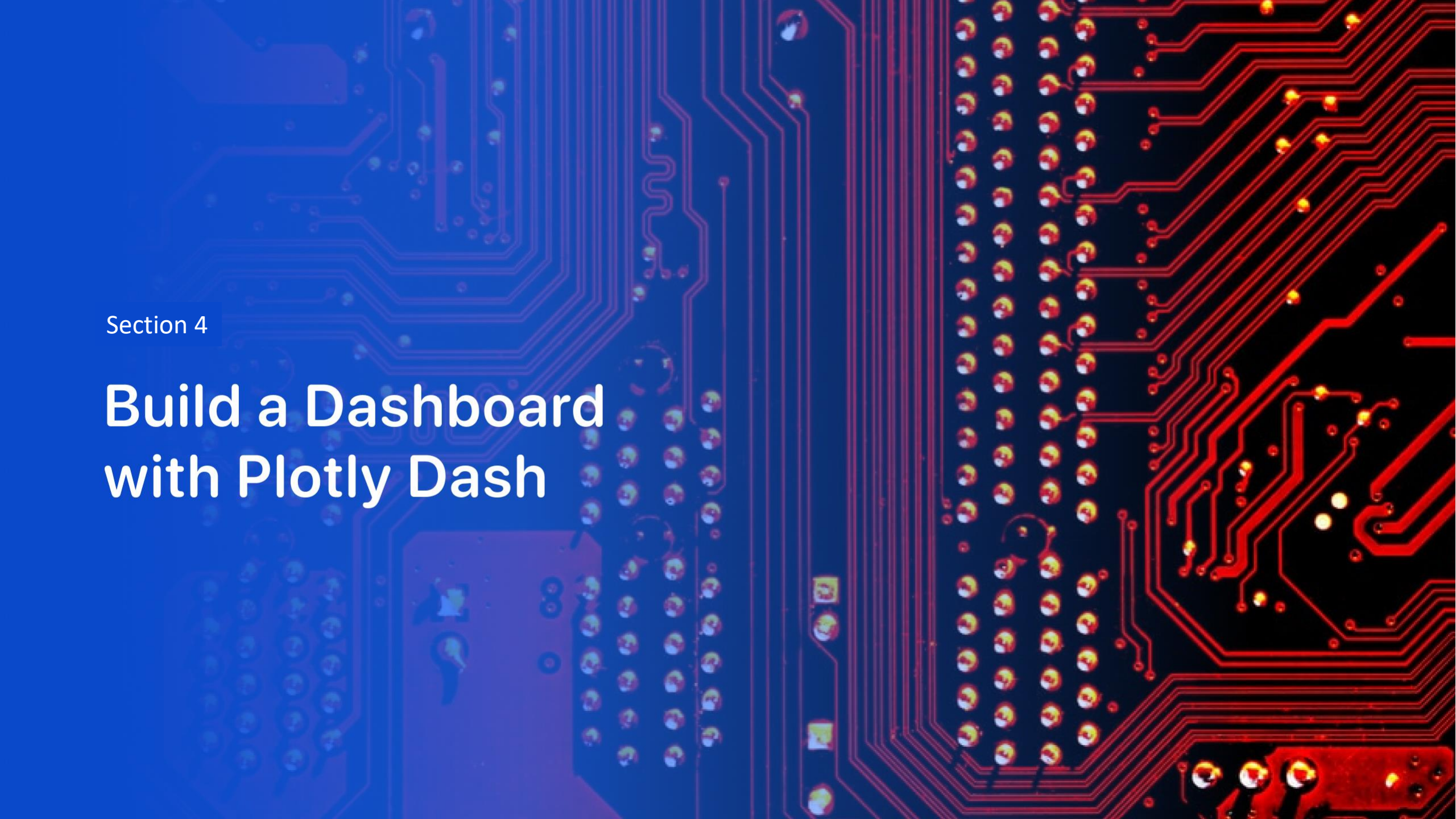
- Example of CCAFS SLC-40 launch site launch outcomes



Coast Line

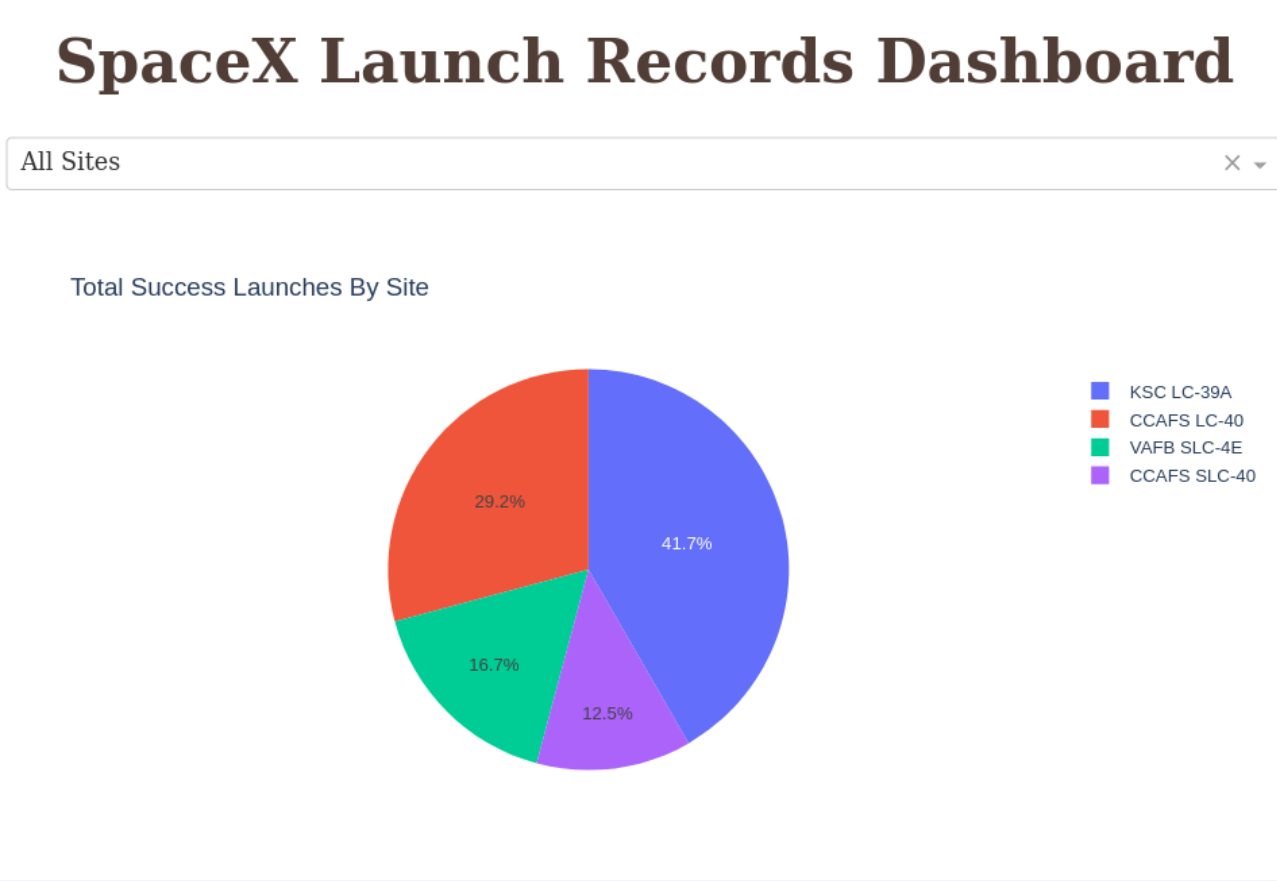
- Launch Site CCAFS SLC-40 is only 900 meters away from the coast line, which is pretty safe for the launching.
- It is also beside the road, which means it is easy for the company to transport rocket.




Section 4
Build a Dashboard with Plotly Dash

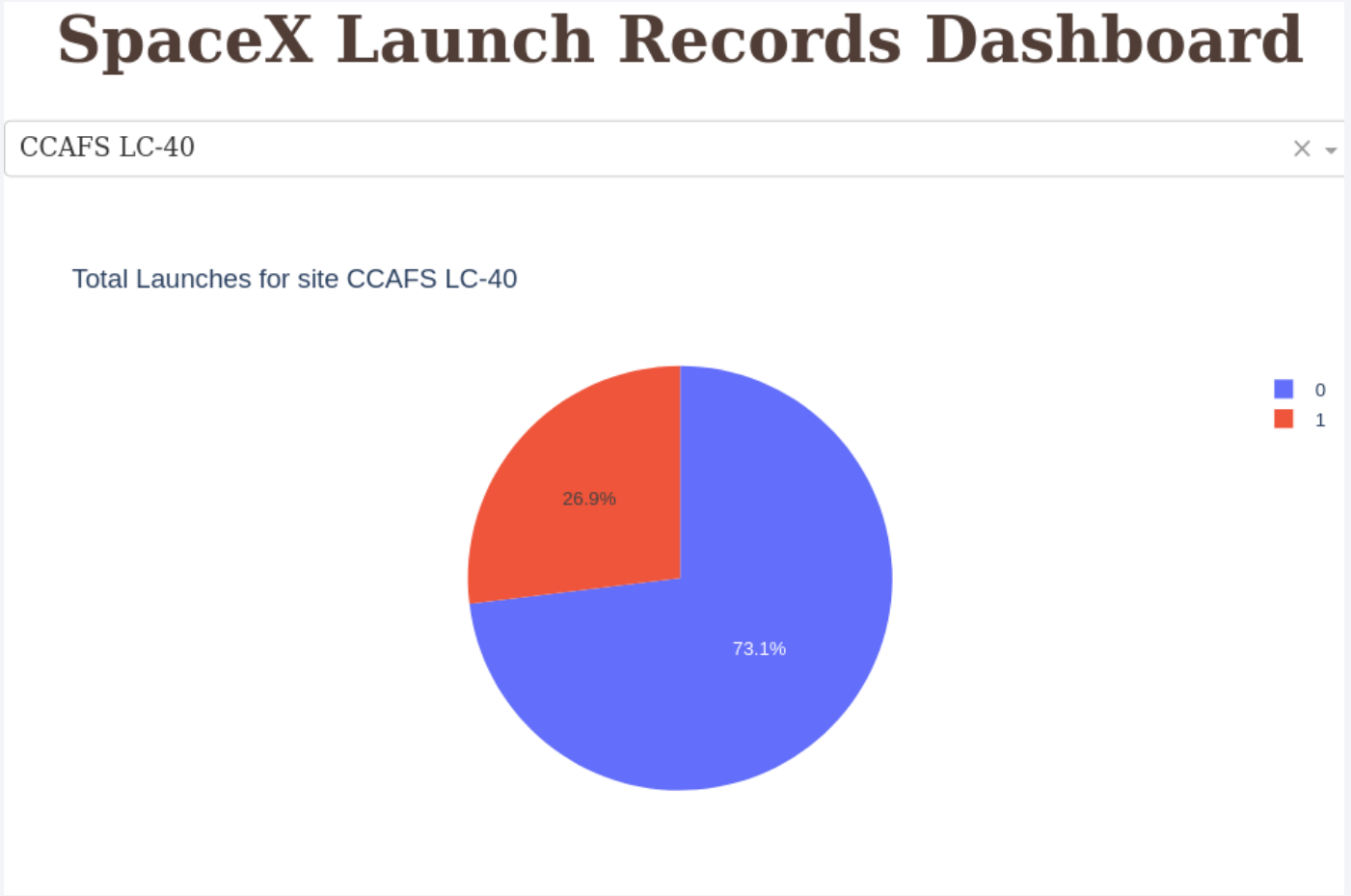
Successful launches by sites

- KSC LC-39A has the largest number of successful launches.
- CCAFS SLC-40 has the least number of successful launches.



Success Rate for CCAFS LC-40

- CCAFS LC-40 has the lowest success rate of 26.9%



Payload & Booster Version

- If payload is under 6k, FT booster version is the most successful one. On the other hand, v1.1 is the most unsuccessful one.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Logistic Regression, SVM, and KNN share the same highest accuracy:83.34%

Find the method performs best:

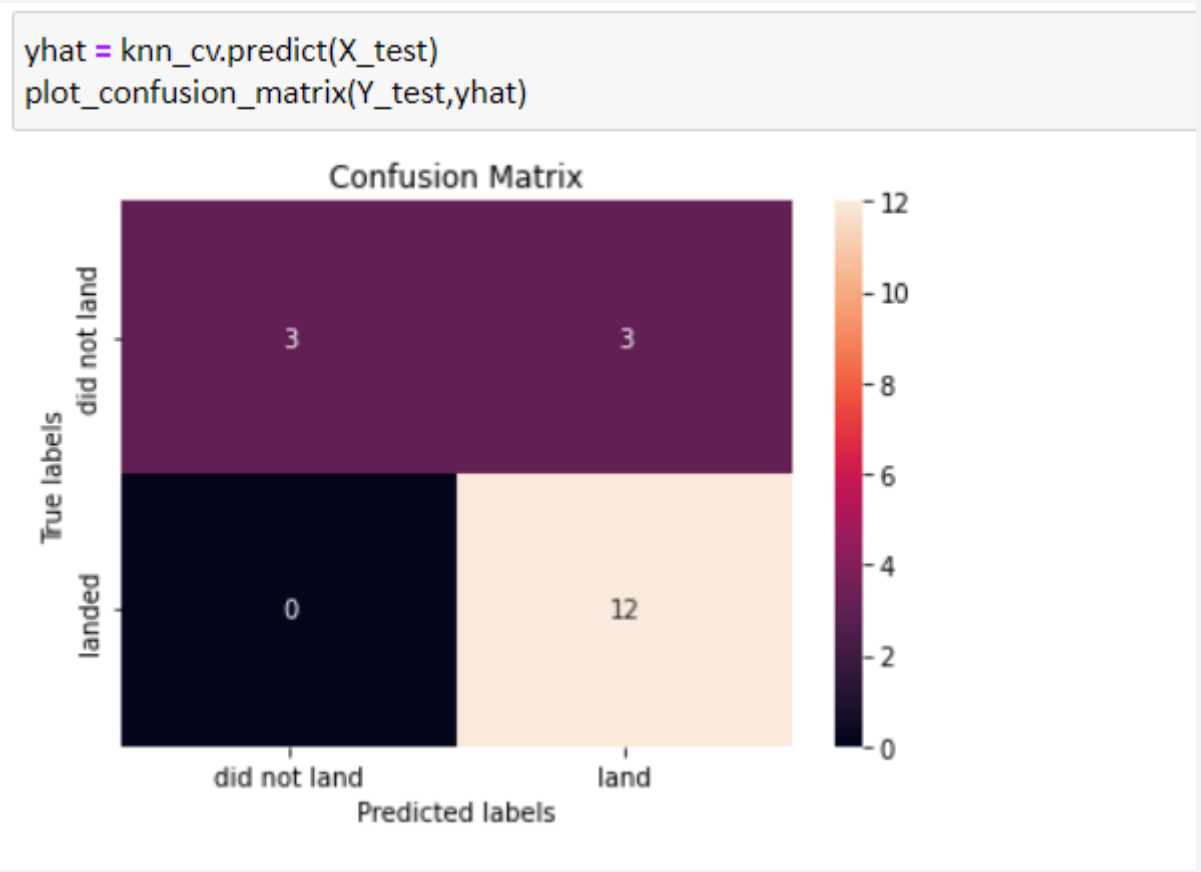
```
print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.7777777777777778
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

Confusion Matrix

- Confusion matrix of KNN, SVM, Logistic Regression proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- The best launch site is KSC LC 39A;
- Launches above 5,500kg are less risky;
- It turns out that most of mission outcomes are successful. The success rate was low in the first few years, but successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- SVM, Logistic Regression, and KNN can be used to predict successful landings and increase profits.

Thank you!