IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vandanapu Pranay
21/08/2022

Outline

- Executive Summary
- Introduction
- Methodology
- Insights drawn from EDA
- Launch Sites Proximities Analysis
- Build a Dashboard with Plotly Dash
- Predictive Analysis (Classification)
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection using SPACE X API and Web Scraping
  - Data Wrangling, Exploratory Data Analysis (EDA) with SQL, Visualization
  - Interactive Visual Analytics using Folium & Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - EDA allowed us to identify the best features to predict the success of launches.
  - Machine Learning Prediction has given the best model to predict the characteristics needed for the launch, using the historical data from SPACE X.

Introduction

- ## Project background and context
- The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is SpaceX.
- SpaceX's accomplishments include: Sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space.
- One reason SpaceX can do this is the rocket launches are relatively inexpensive, much of the savings is because SpaceX can reuse the first stage.
- Objective is to evaluate the viability of a company Space Y that can compete with Space X.
- ## Problems you want to find answers
- Estimate the total cost of the launches, by predicting successful landings of first stage of Rocket. So, what features like payload, launch site, orbit etc., does it depend upon?
- Which algorithm is best for this classification?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using Space X Rest API and through Web Scraping Wikipedia.
- Perform data wrangling
  - By Filtering the data, dealing with missing values, and using One hot encodings to prepare the data for Classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data processed till this step is normalized, then divided into train and test set. Four different classification models are evaluated using various evaluation metrics to find the best model

Data Collection

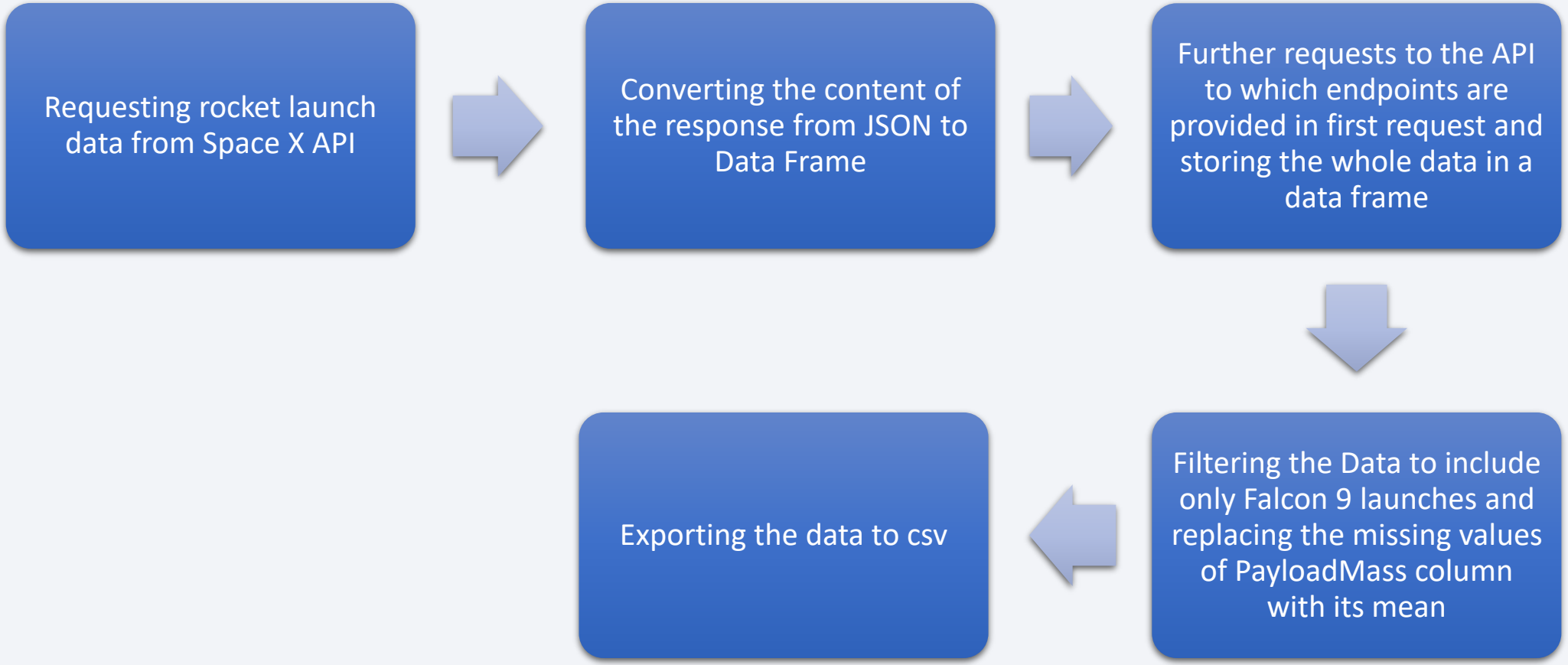
Data Collection Process involves a multiple API requests to the Space X API and Web Scraping a table from Space X's Wikipedia entry. Both the sources of data are needed to get complete information and so to make a more informed prediction.

Space X API: https://api.spacexdata.com/v4/rockets

Web Scraping of: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

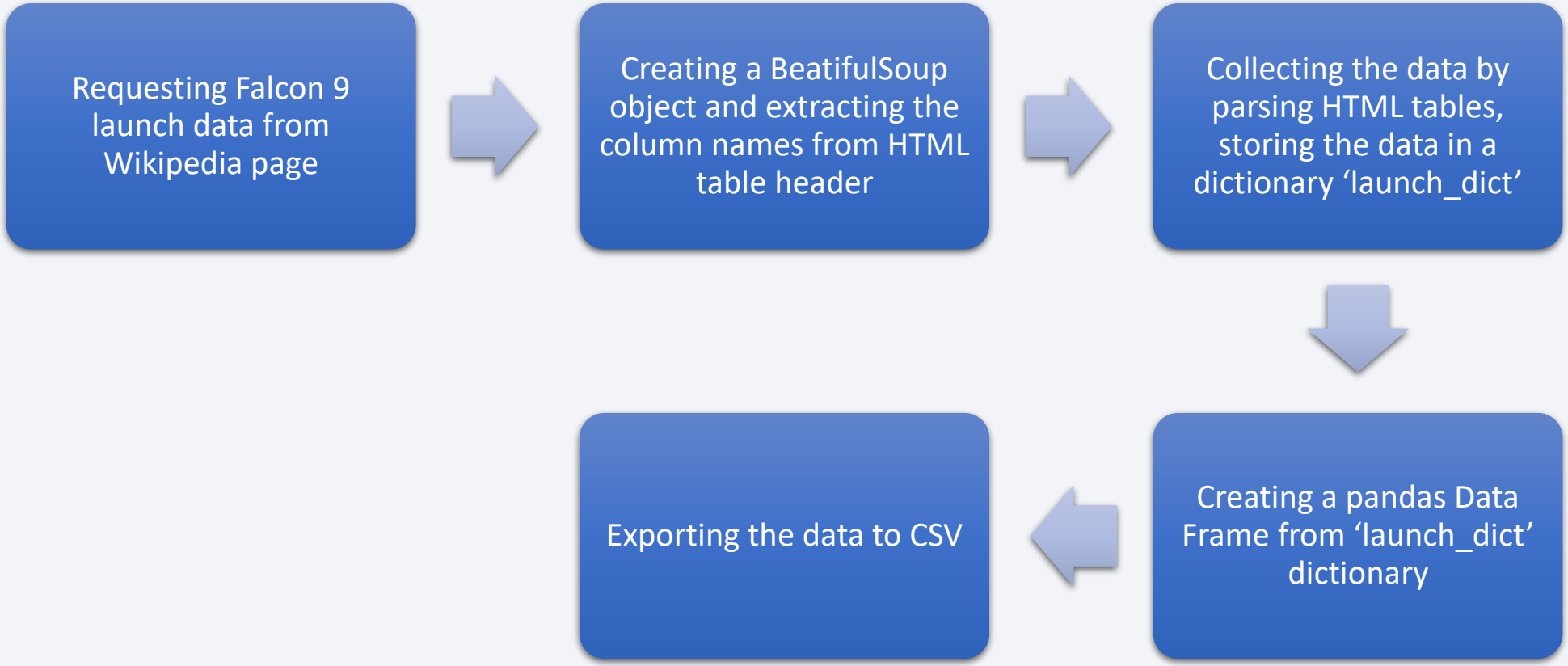
Data Collection – SpaceX API

Data Collection using API (Notebook URL)



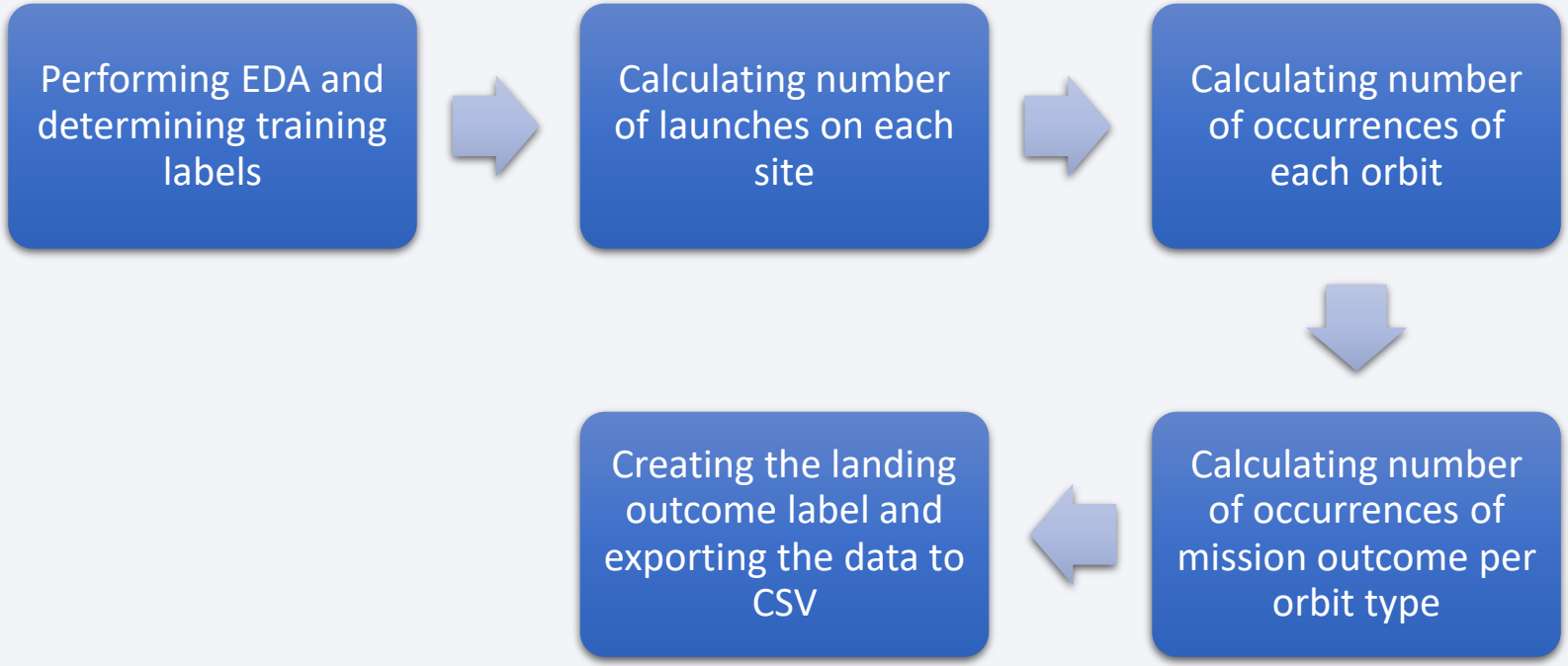
Data Collection - Scraping

Data Collection through Web Scraping (Notebook URL)



Data Wrangling

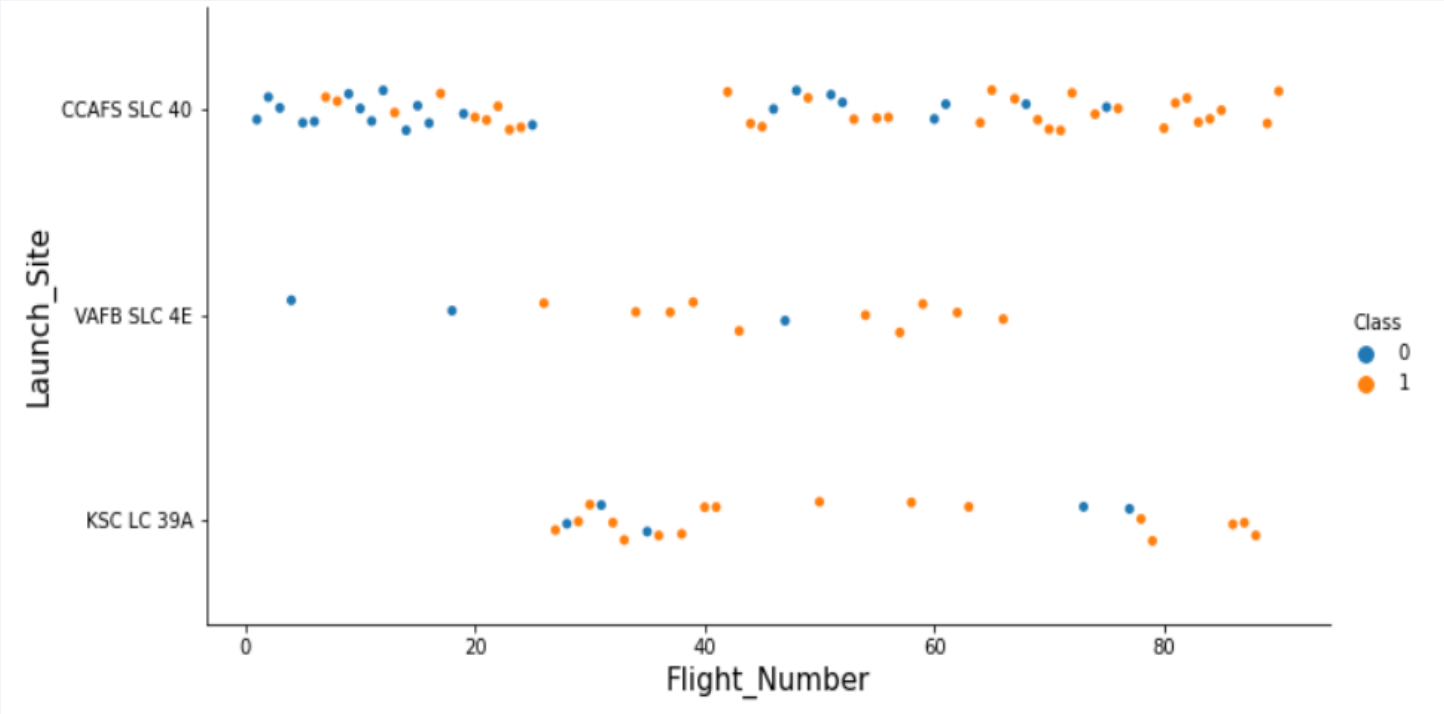
- Basic EDA is performed on the dataset and then summaries of launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type are calculated.
- The landing outcome label is created from the Outcome column. ‘1’ means successful landing, ‘0’ means unsuccessful landing.
- Data Wrangling (Notebook URL)



EDA with Data Visualization

- To explore data, scatter plots and bar plots are used to visualize the relationship between following 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.

  - The image here is a scatter plot between the features Launch Site and Flight Number
  - EDA with Data Visualization (Notebook URL)



EDA with SQL

The following SQL queries are performed:

- Names of the unique launch sites in the space mission
- Top 5 launch sites whose name begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome on ground pad was achieved
- Names of the boosters which have success in drone ships and have payload mass between 4000 and 6000kg
- Total number of successful and failed mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ships, their booster versions, and launch site names in the year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 in descending order.

EDA with SQL(Notebook URL)

Build an Interactive Map with Folium

- Task 1: Mark all launch sites on a map
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts
- Task 2: Mark the success/failed launches for each site on the map
  - Added colored Markers of success (green) and failed (red) launches using Marker Cluster to identify which launch sites have relatively high success rates
- Task 3: Calculate the distances between a launch site to its proximities
  - Added colored Lines to show distances between the Launch Site VAFB SLC-4E and its proximities like Railway, Highway, Coastline and Closest City

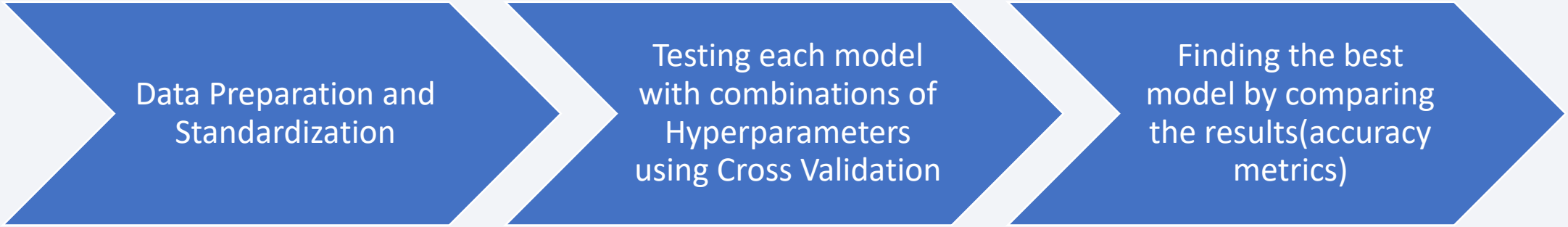
Visual Analytics with Folium(Notebook URL)

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List
  - Added a dropdown list to enable Launch Site selection
- Pie Chart showing Success Launches (All Sites/Certain Site)
  - Added a pie chart to show the total successful launches count for all sites and the Success vs Failed counts for the site, if a specific Launch Site was selected
- Slider of Payload Mass Range
  - Added a slider to select Payload range
- Scatter Chart of Payload Mass vs Success Rate for the different Booster Versions
  - Added a scatter chart to show the correlation between Payload and Launch Success
- Dashboard with Plotly Dash (Notebook URL)

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.
- Predictive Analysis (Notebook URL)

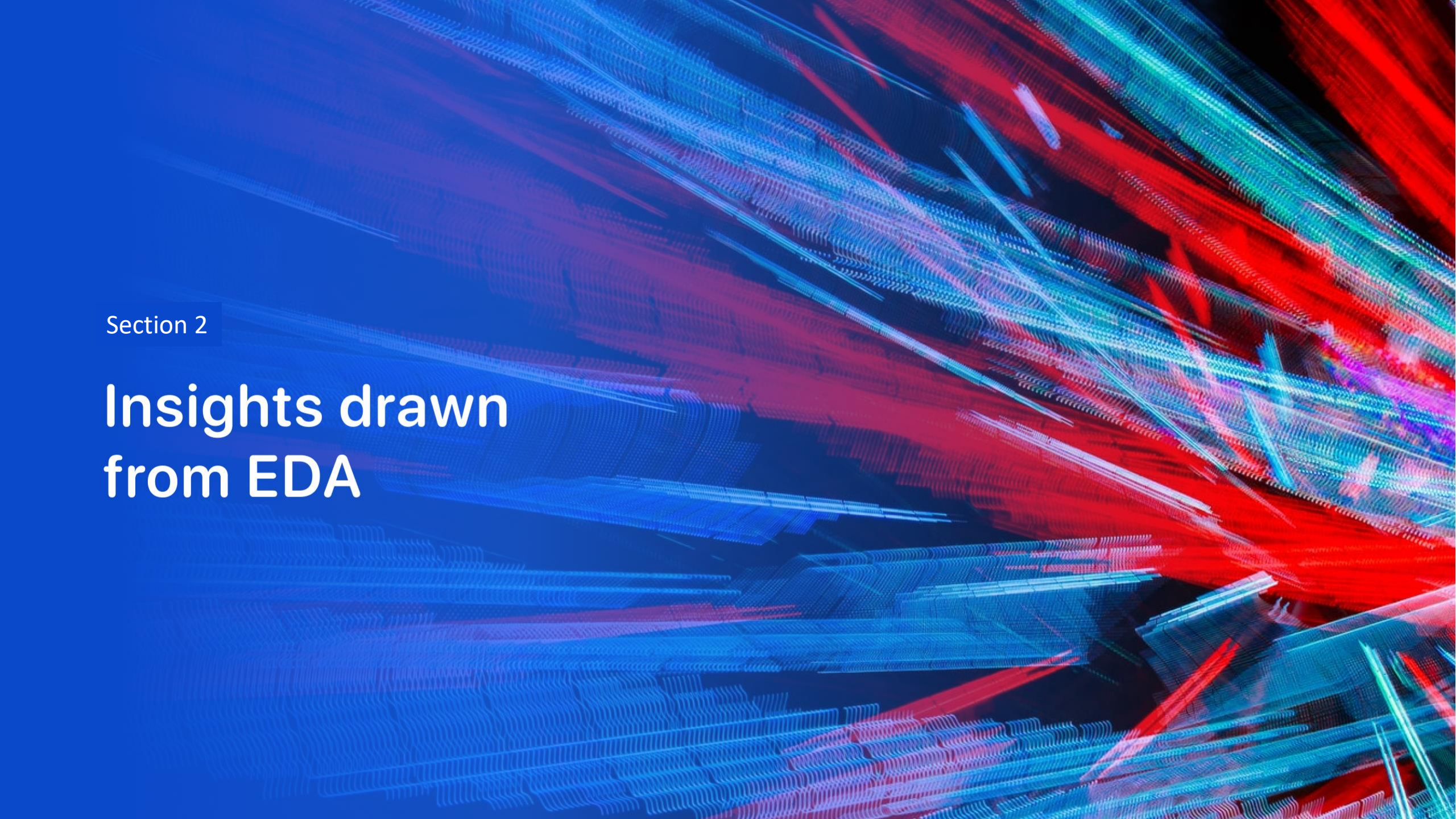


Results

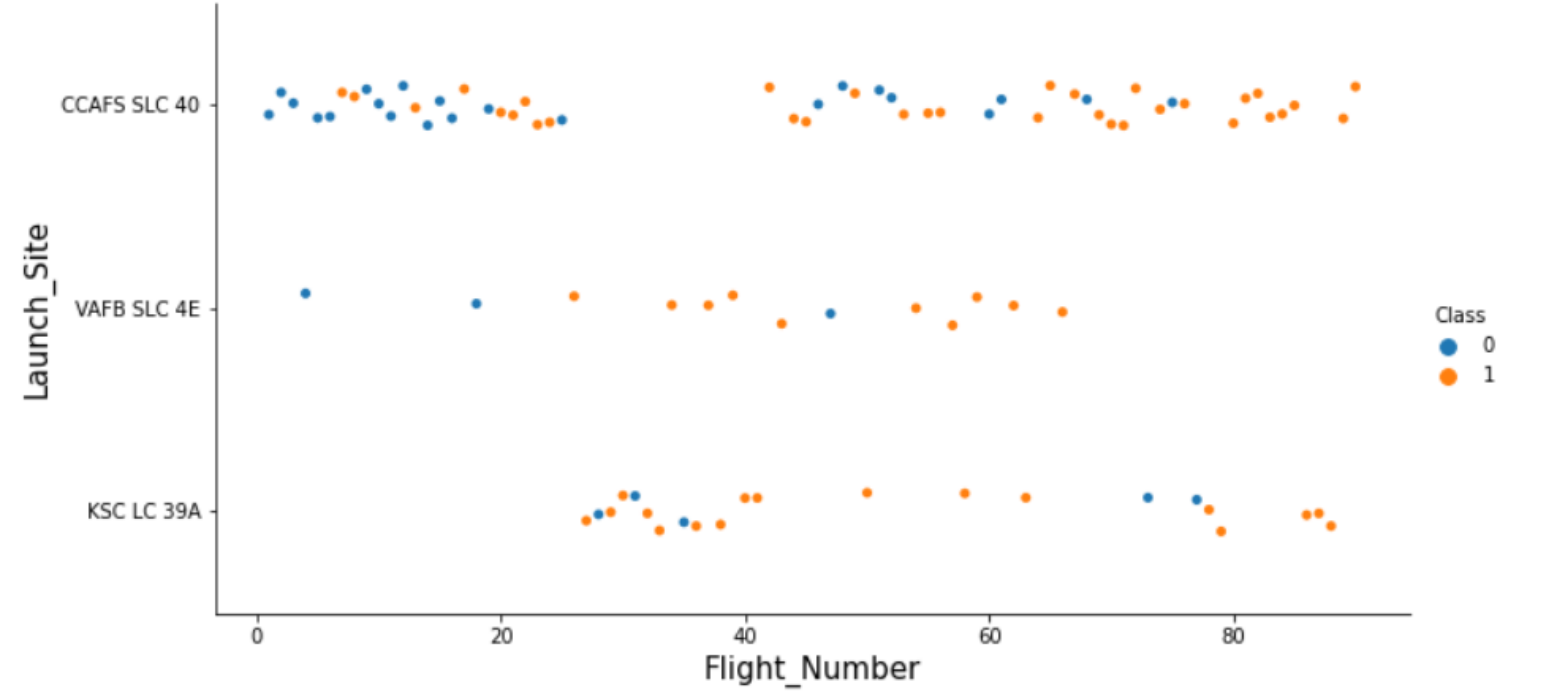
- Exploratory data analysis results
- Space X uses 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload the of F9 v1.1 booster is 2,928 kg;
- The first successful landing outcome happened in 2015 five years after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payloads above the average;
- Almost 100% of mission outcomes were successful. Two booster versions failed at landing on drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed;
- Using interactive analytics was possible to identify that launch sites use to be in safe places, near the sea , for example, and have a good logistic infrastructure around.
- Most launches happen at east cost launch sites.

Results

- Predictive analysis results
  - Predictive Analysis showed that the Decision Tree Classifier is the best model to predict successful landings, having accuracy of over 89% and accuracy for test data over 83%

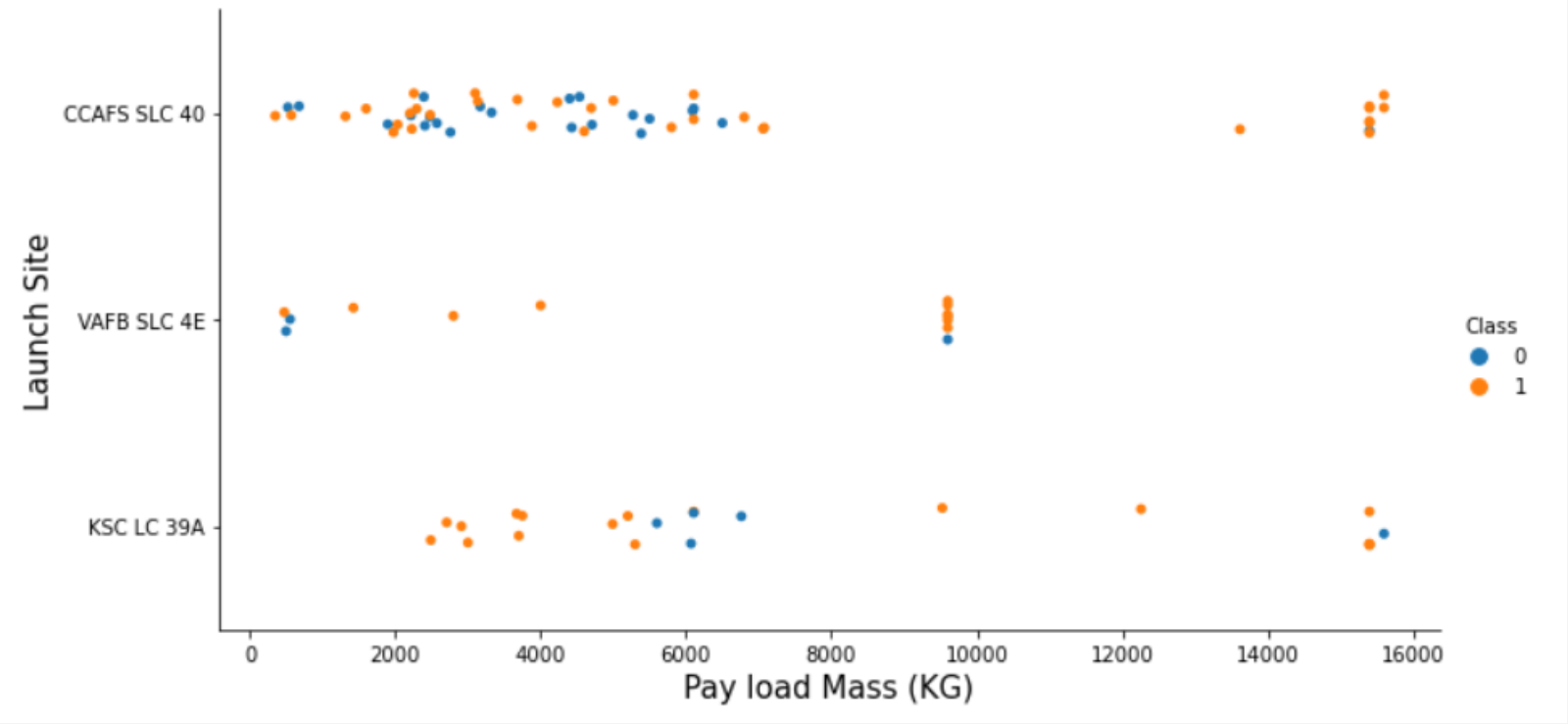

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The earliest flights all failed while the latest flights all succeeded
- The CCAFS SLC 40 launch site has about a half of all Launches Site
- VAFB SLC 4 E and KSC LC 39 A have higher success rates than CCAFS SLC 40
- We can see that each new launch has a higher rate of success

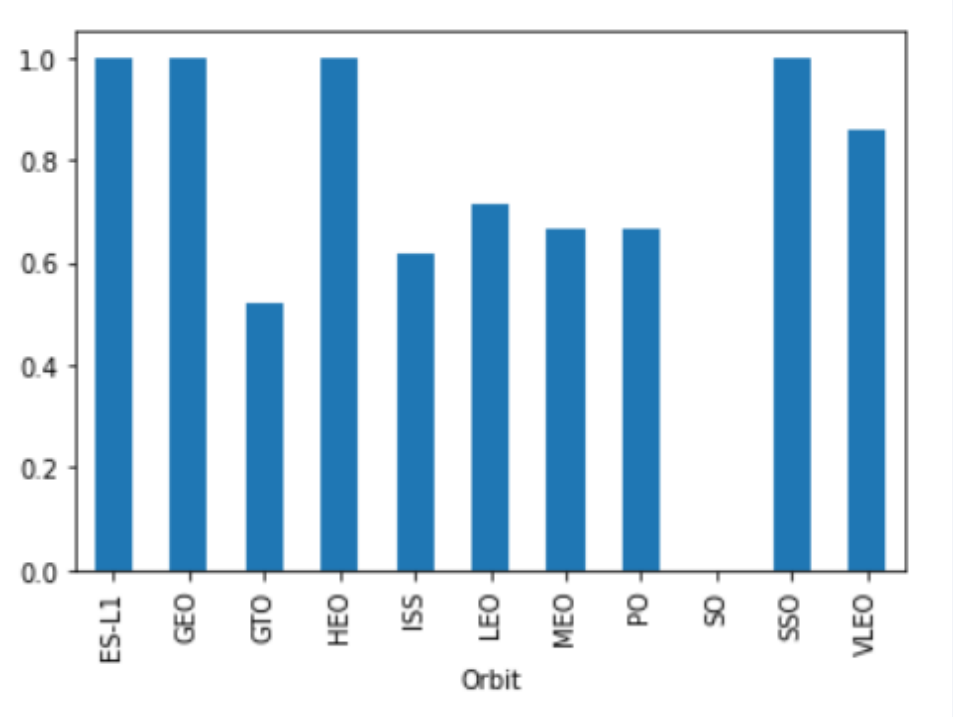
Payload vs. Launch Site



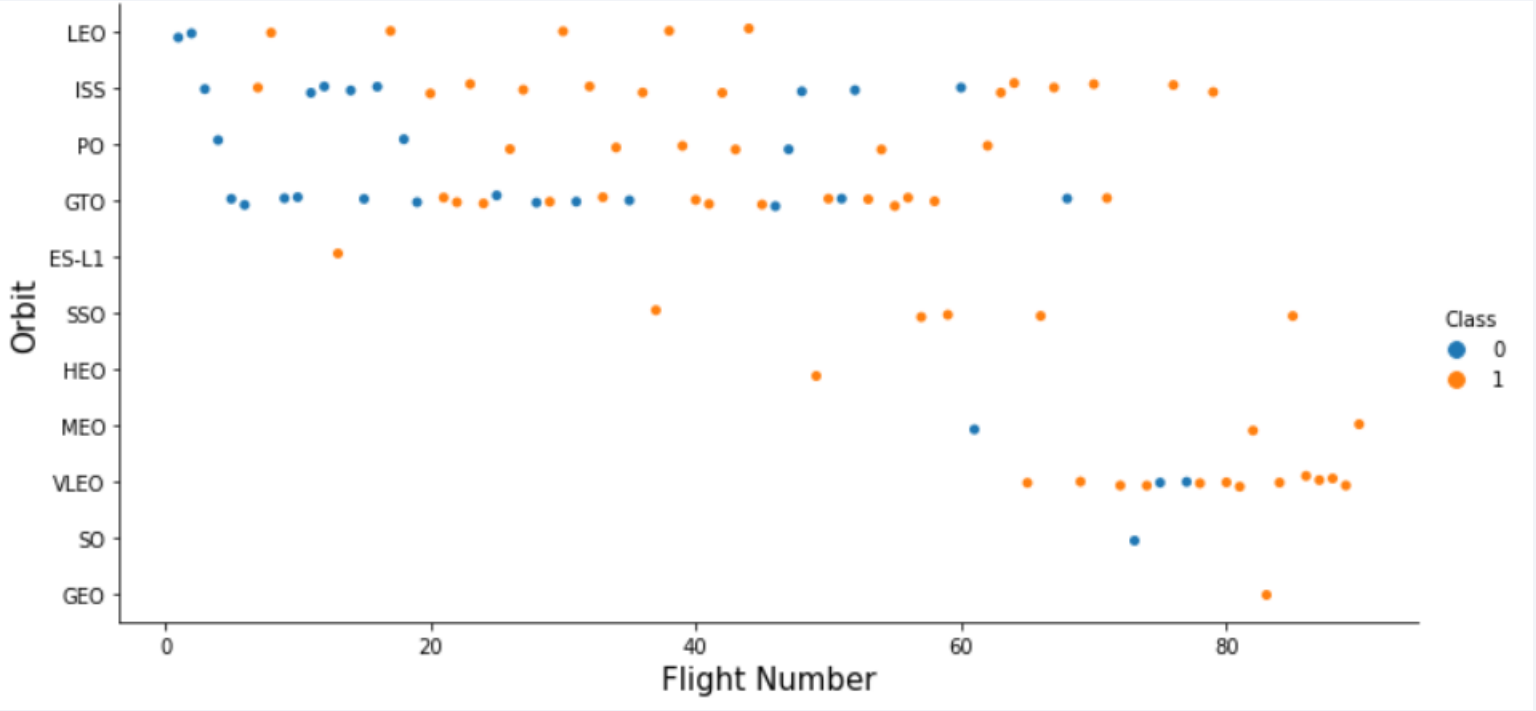
- For VAFB-SLC 4E launch site there are no rockets launched for heavy payload mass (>10,000 kg).
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100 success rate for payload mass under 5500 kg.

Success Rate vs. Orbit Type

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

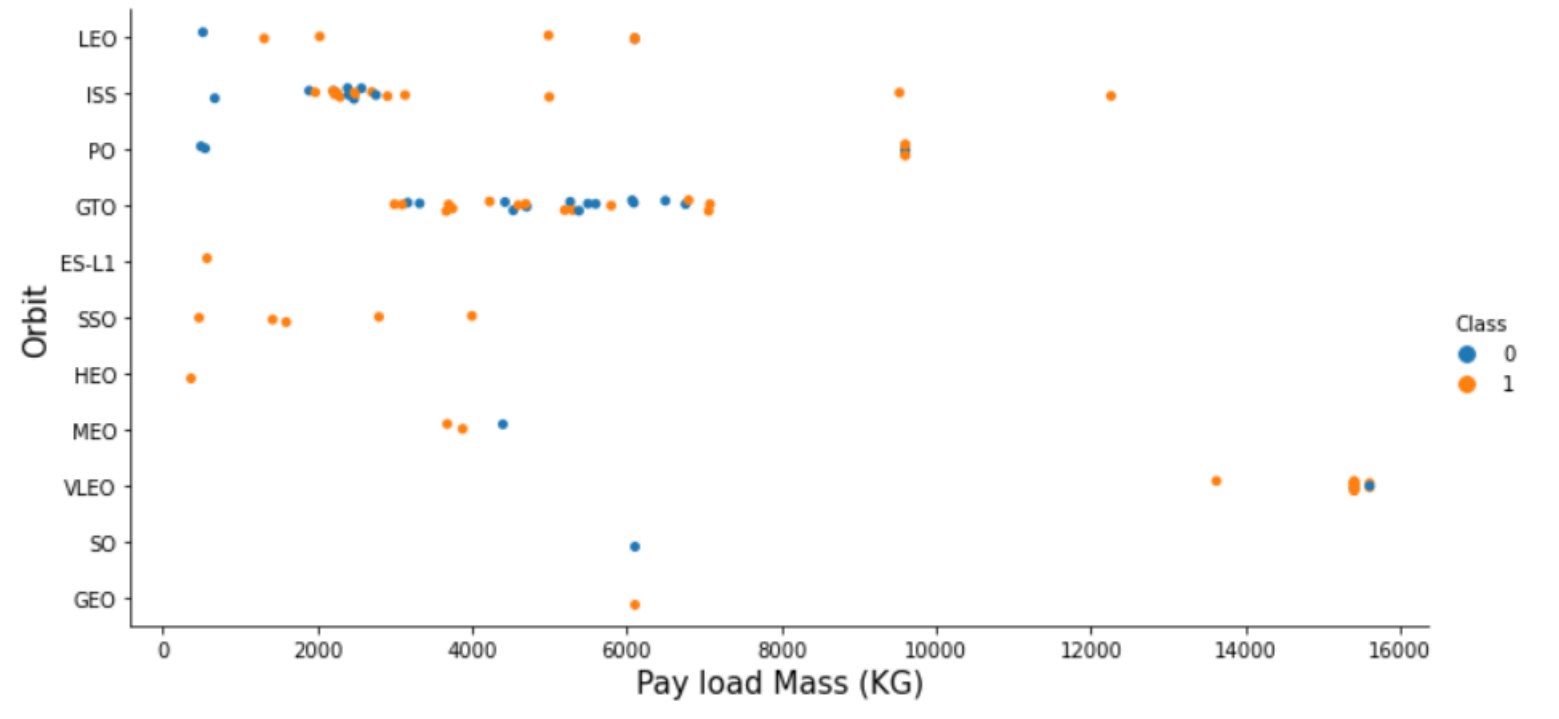


Flight Number vs. Orbit Type



- Apparently, the success rate improved over time to all orbits.
- VLEO orbit seems a new business opportunity, due to the recent increase in its frequency

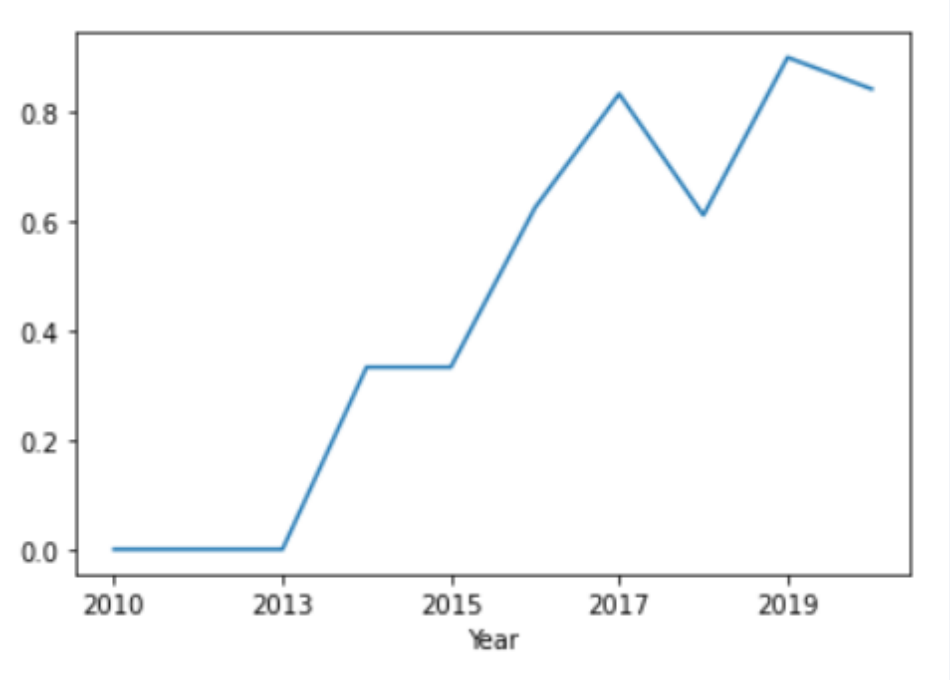
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing unsuccessful mission) are both there here

Launch Success Yearly Trend

- The success rate since 2013 kept increasing until 2020 however between 2017 and 2018 the success rate dropped about 20%
- It seems that the first three years were a period of adjusts and improvement of technology



All Launch Site Names

```
%sql select distinct(launch_site) from SPACEXDATASET
```

```
 * ibm_db_sa://yfg27871:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Display the names of the unique launch sites in the space mission. The result were 4 different launch sites.

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://yfg27871:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
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	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

5 records where launch sites begin with `CCA

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXDATASET where CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://yfg27871:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

1
45596

- Display the total payload mass carried by boosters launched by NASA (CRS). The result was 45,596 kg.

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXDATASET where BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://yfg27871:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

1
2928

- The average payload mass carried by booster version F9 v1.1 was 2928 kg.

First Successful Ground Landing Date

```
%sql select min(DATE) from SPACEXDATASET where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://yfg27871:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

1
2015-12-22

- The date when the first successful landing outcome in ground pad was achieved was on December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

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

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(mission_outcome) from SPACEXDATASET where mission_outcome = 'Success' or mission_outcome like 'Failure%'
```

```
 * ibm_db_sa://yfg27871:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

1
100

- There were a total of 100 successful and failure mission outcomes.

Boosters Carried Maximum Payload

- There are 12 different booster versions which have carried the maximum payload mass.

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

MONTH	landing_outcome	booster_version	launch_site
1	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
4	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015 The results presents one failed landing outcome in January and another in April, on the same launch site (CCAFS LC-40)

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

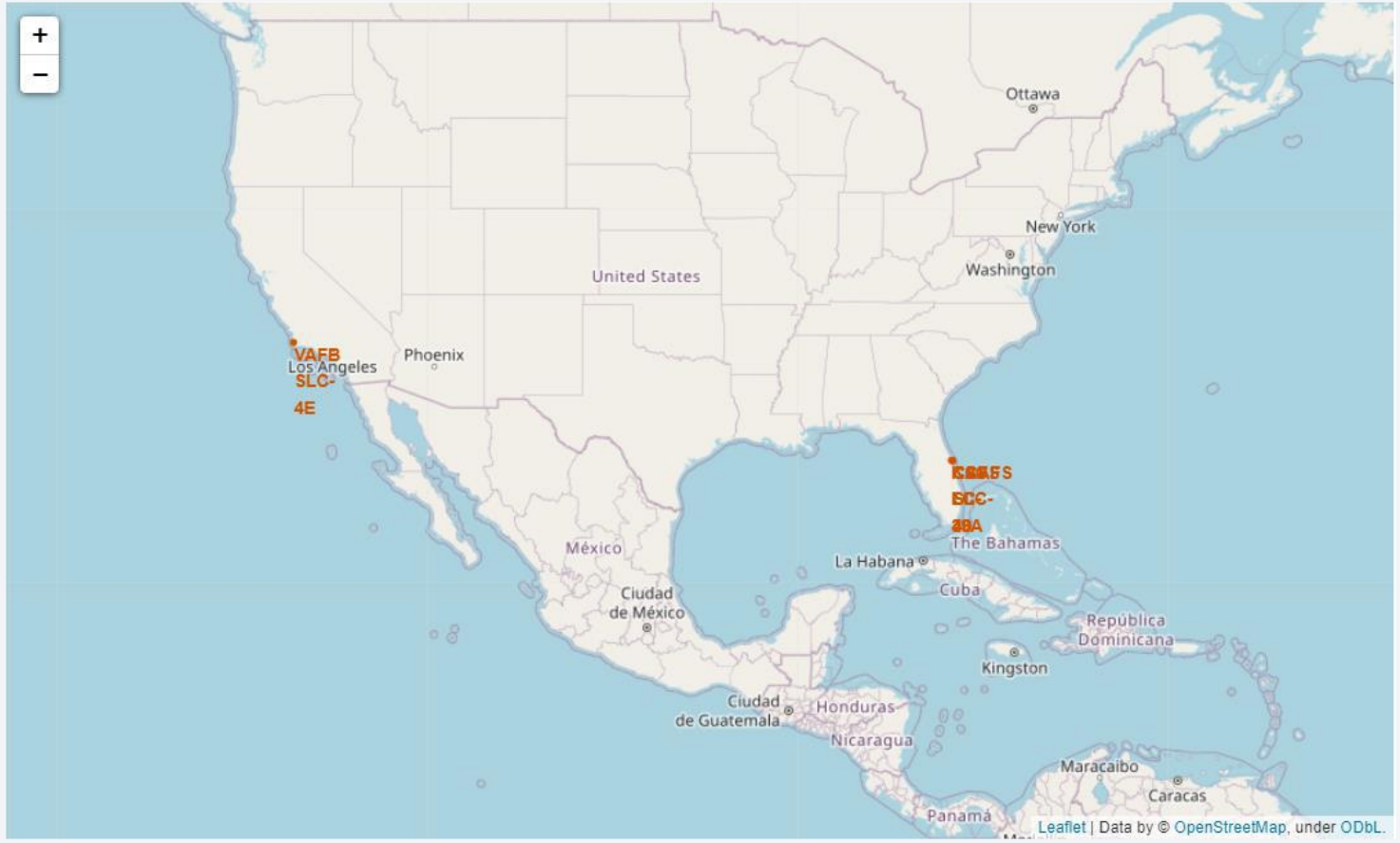
Between 2010-06-04 and 2017-03-20, there are 31 launches.

landing__outcome	Total 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

Section 3

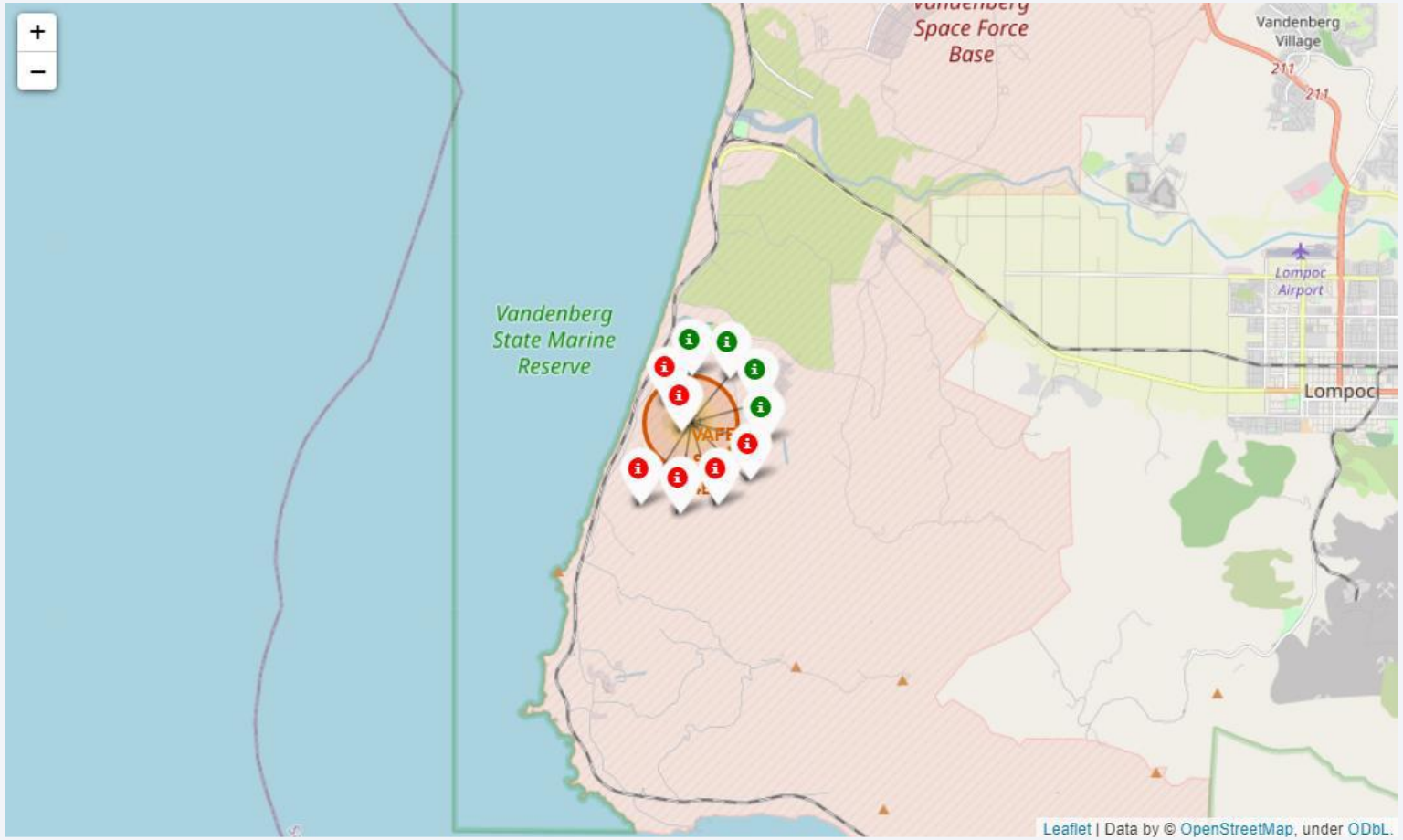
Launch Sites Proximities Analysis

Launch sites locations markers on a global map



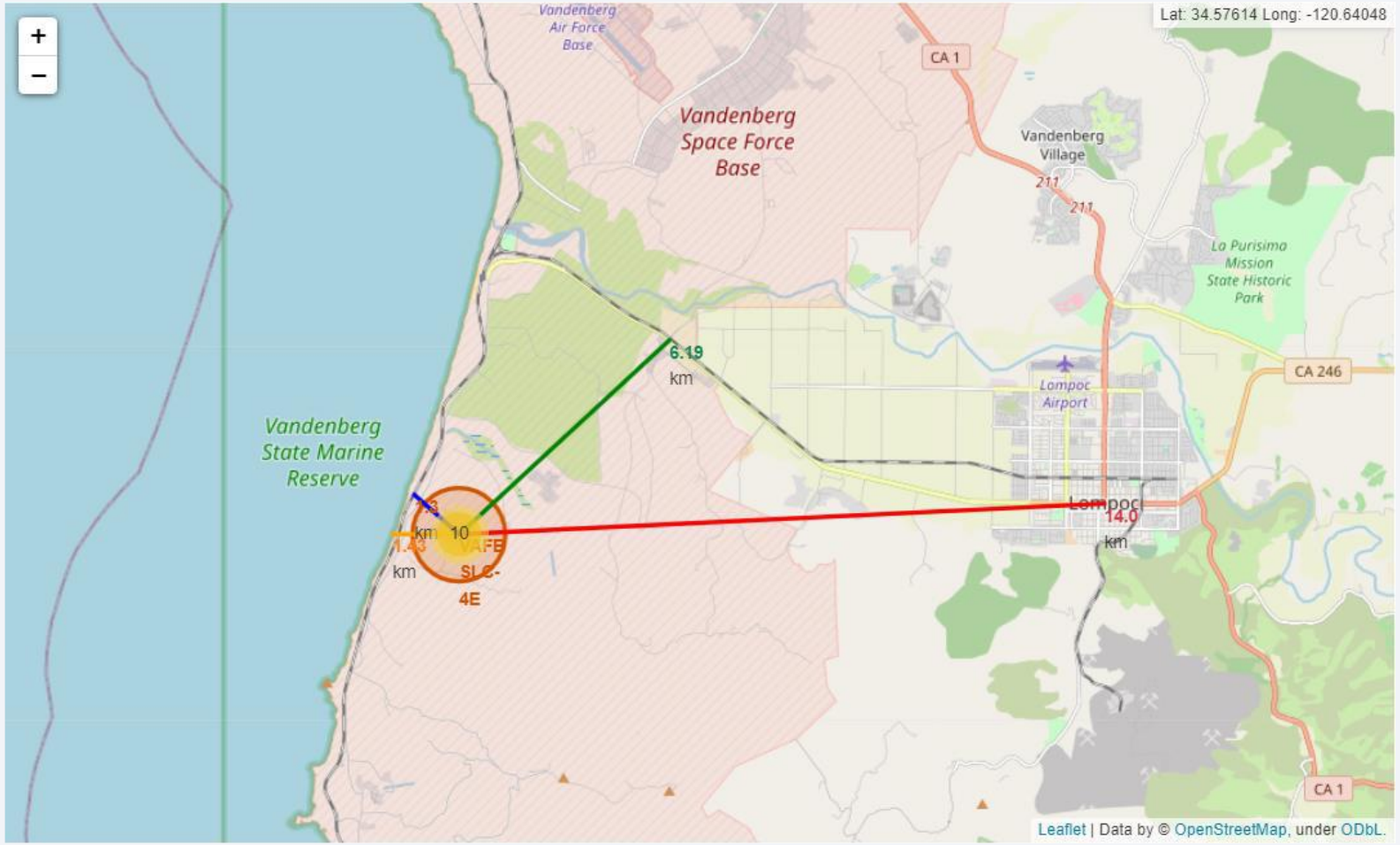
- Observations:-
- All launch sites lie in the proximity of equator.
- All launch sites lie in close proximity to the coast.

Success and failed launches for each site on the map

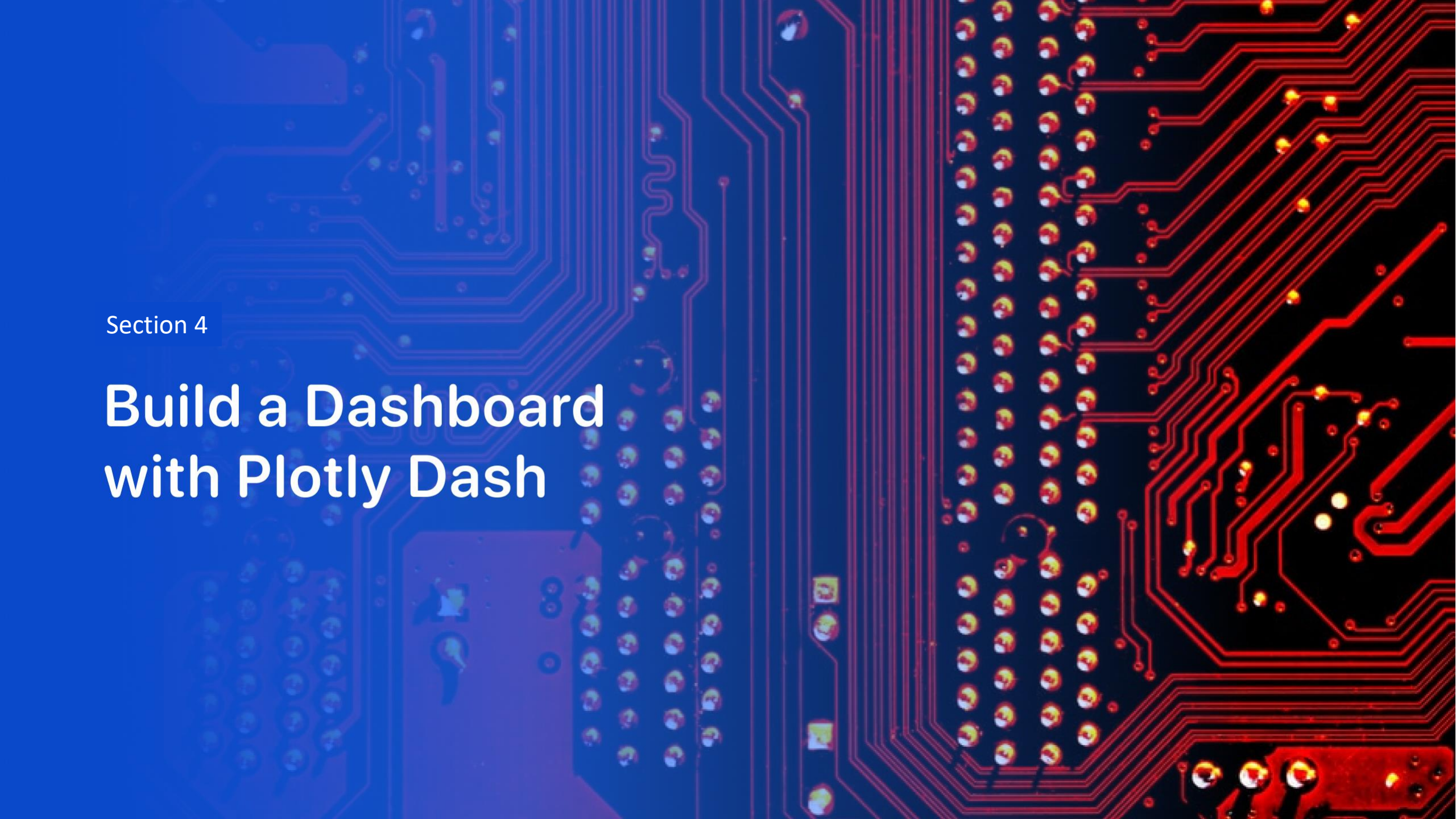


- Green marker indicates a successful launch while red marker indicates unsuccessful launch.
- From the figure, We can say that in VAFB SLC-4E launch site has 4 successful launches out of 10.

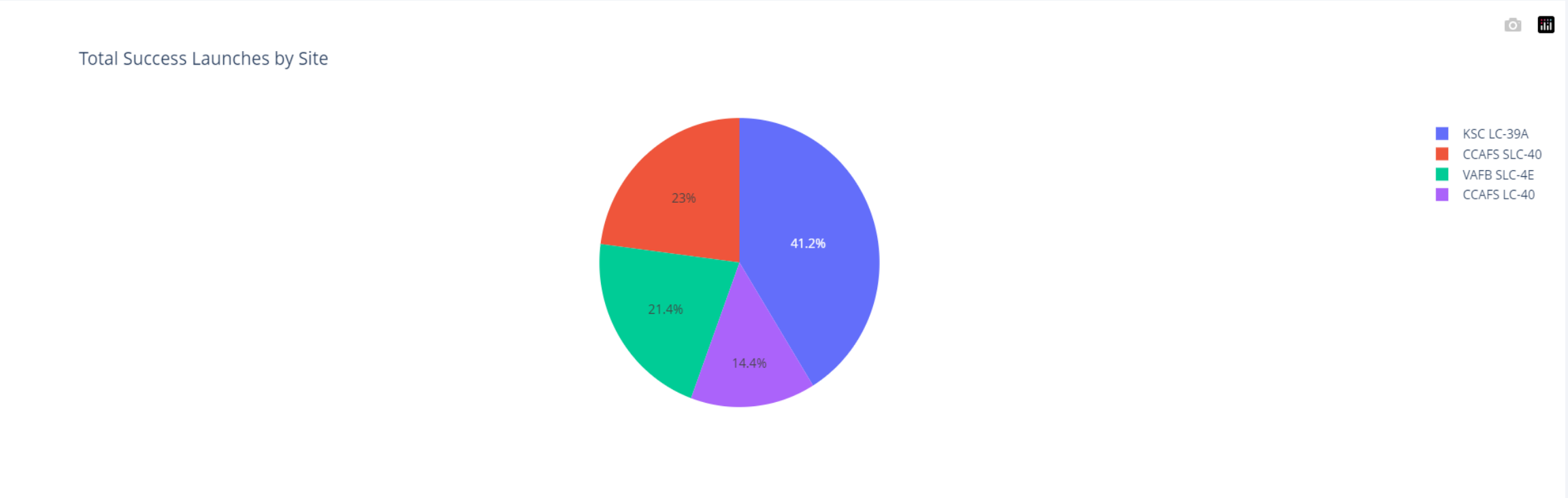
Distance from VAFB SLC-4E launch site to its proximities



- From the visual analysis of the launch site VAFB SLC-4E we can clearly see that it is; very close to railway 1.3 km), close to highway 6.19 km); very close to coastline 1.43 km)
- Also the launch site VAFB SLC-4E is relative close to its closest city Lompoc (14 km)

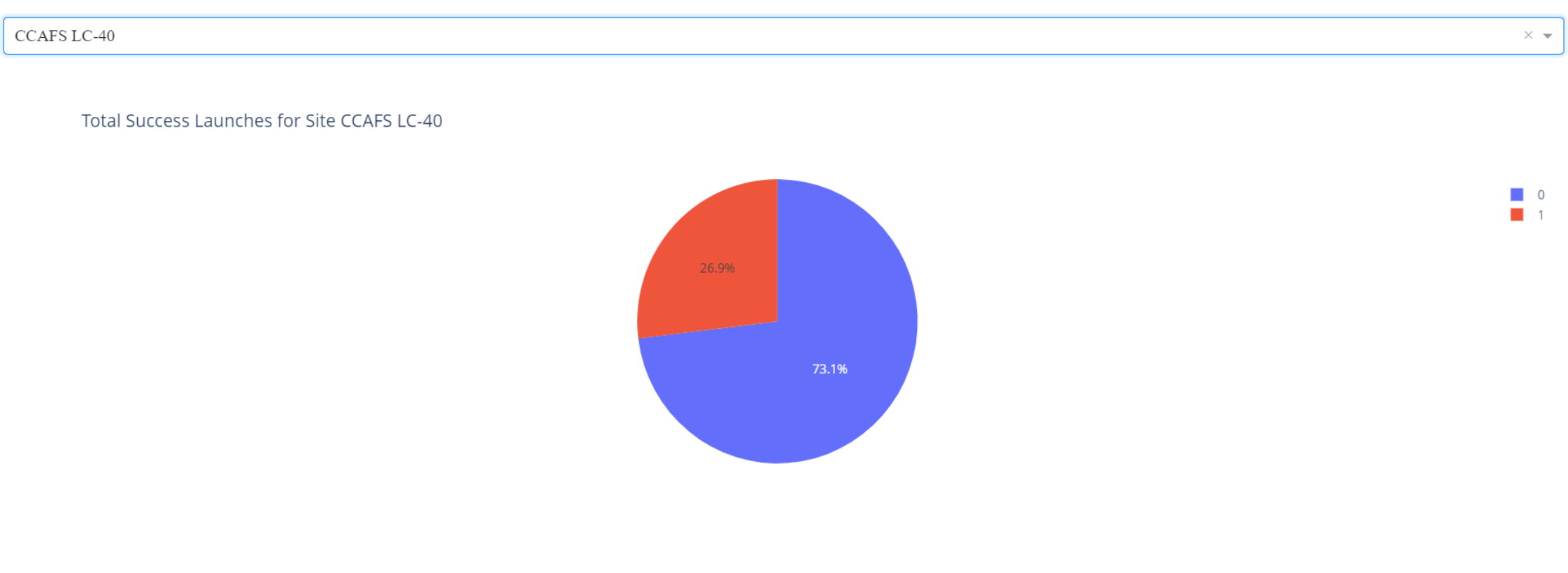

Section 4
Build a Dashboard with Plotly Dash

Successful launches by Site



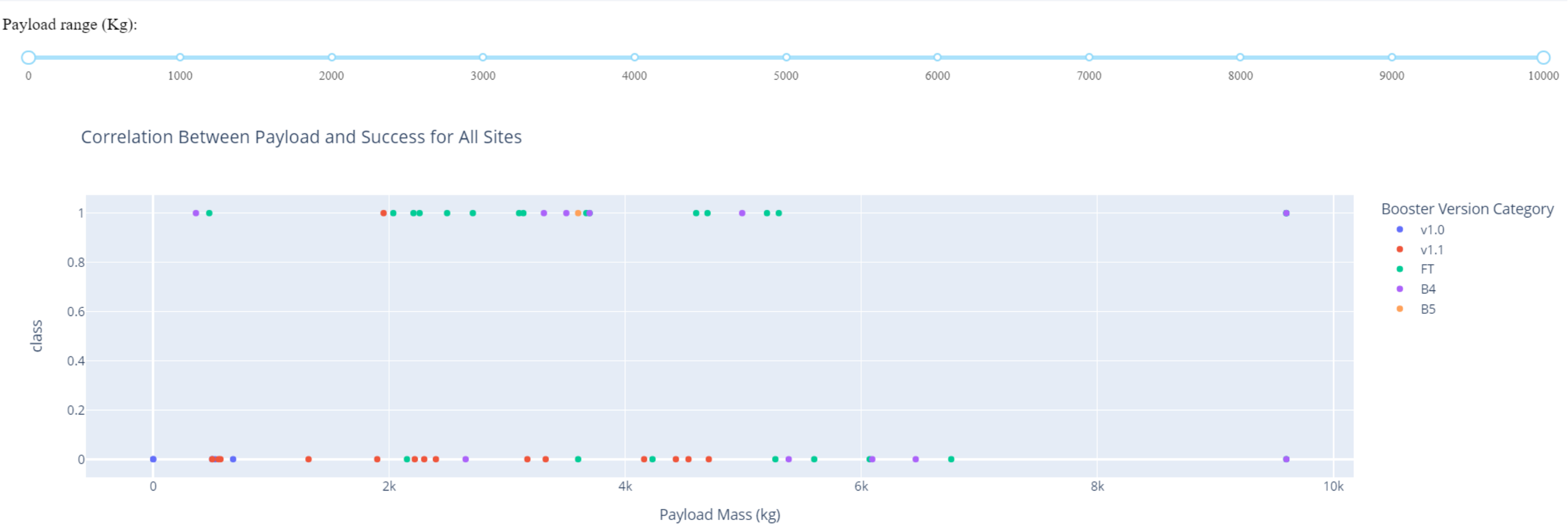
- KSC LC-39A launch site has the highest success rate of launches.
- CCAFS LC-40 launch site has the lowest success rate of launches.

Launch Success Ratio for Site CCAFS LC-40



- CCAFS LC-40 launch site has a success rate of 73.1%

Payload vs. Launch Outcome



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All the models have the same accuracy of over 83% on the test data. Since it is a small dataset (18 test samples)
- But Decision Tree Classifier has better accuracy of over 89% on the training data.
- So, we could say Decision Tree Classifier is the best model for this problem.

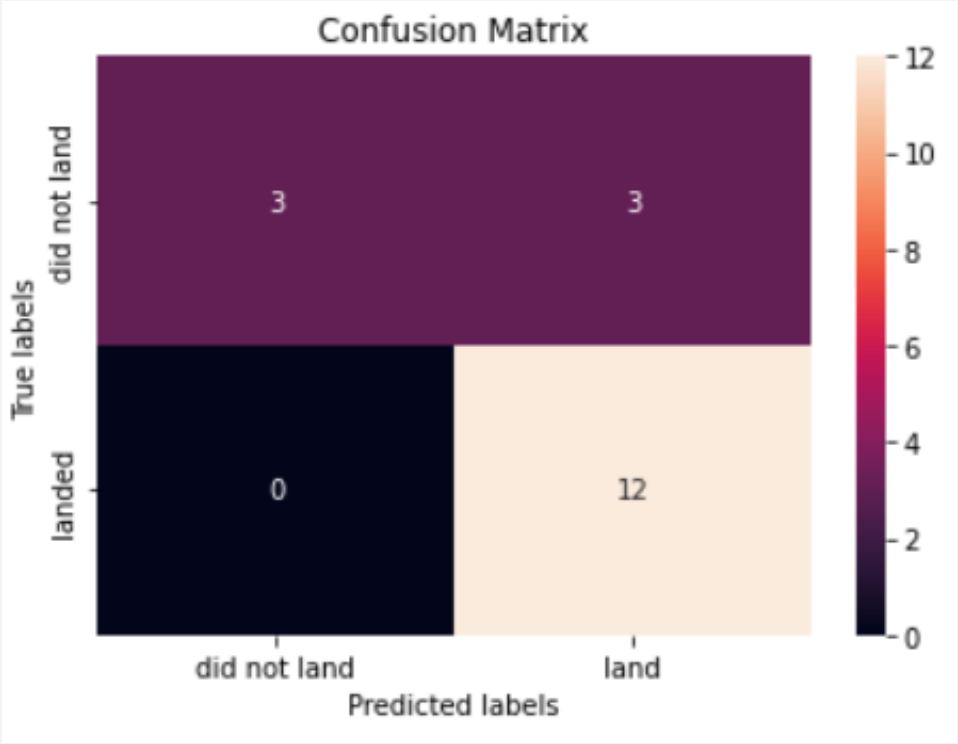
```
print('Accuracy for Logistics Regression:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearst neighbors:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression: 0.8333333333333334
Accuracy for Support Vector Machine: 0.8333333333333334
Accuracy for Decision tree: 0.8333333333333334
Accuracy for K nearst neighbors: 0.8333333333333334
```

```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'entropy', 'max_de
5, 'splitter': 'best'}
accuracy : 0.8910714285714286
```

Confusion Matrix



- Since Results came similar for all the models, Confusion Matrix is same for all.

Conclusions

- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- The success rate of launches increases over the years.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- The accuracy results are practically the same, but Decision Tree Classifier is slightly better because of better training accuracy.
- So, Decision Tree Classifier can be used to predict successful landings and increase profits.

Thank you!