IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Maurício Rocha
November 2024

Outline

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

Executive Summary

The goal of this research was to predict the success of first stage landing of SpaceX's Falcon 9 to determine the cost of a launch and therefore provide other space agencies the ability to decide if they bid against SpaceX for a rocket launch.

- Summary of methodologies:
  - Collect data through API and Web scraping;
  - Transform data through data wrangling;
  - Conduct exploratory data analysis with SQL and data visuals;
  - Build an interactive map with folium to analyze launch site proximity;
  - Build a dashboard to analyze launch records interactively with Plotly Dash;
  - Finally, build a predictive model to predict if the first stage of Falcon 9 will land successfully.
- Summary of all results
  - It was possible to collected valuable data from public sources;
  - EDA (Exploratory Data Analysis) 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

- Project background and context:
  - With the recent successes in private space travel, space industry is becoming more and more mainstream and accessible to general population. Cost of launch continues to remain a key barrier for new competitors to enter the space race;
  - SpaceX with its first stage reuse capabilities offers a key advantage against its competitors. Each SpaceX launch costs around 62 million dollar and SpaceX can reuse stage 1 for future launches. This provides SpaceX a unique advantage where other competitors are spending around 165 mission plus for each launch
- Problems you want to find answers
  - Determine if the first stage of SpaceX Falcon 9 will land successfully;
  - Impact of different parameters/variables on the landing outcomes (e.g., launch site, payload mass, booster version, etc.);
  - Correlations between launch sites and success rates.

Section 1

Methodology

Methodology

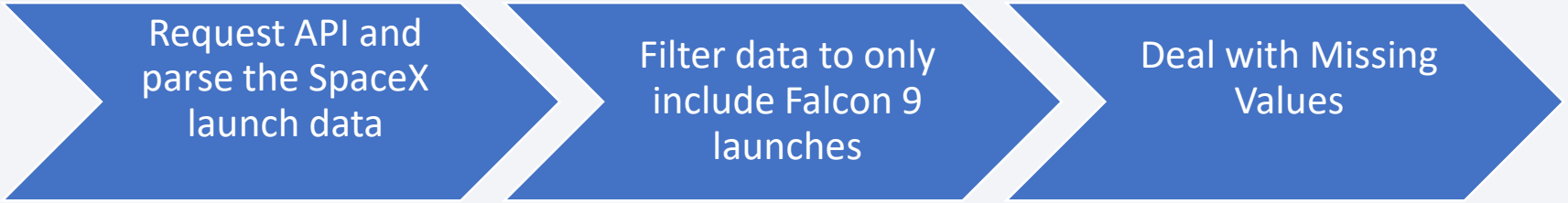
Executive Summary

- Data collection methodology:
  - SpaceX API;
  - Web scrap from List of Falcon/ 9/ and Falcon Heavy launches - Wikipedia
- Performed data wrangling
  - Determined labels for training the supervised models by converting mission outcomes into training labels (0-unsuccessful, 1-successful)
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

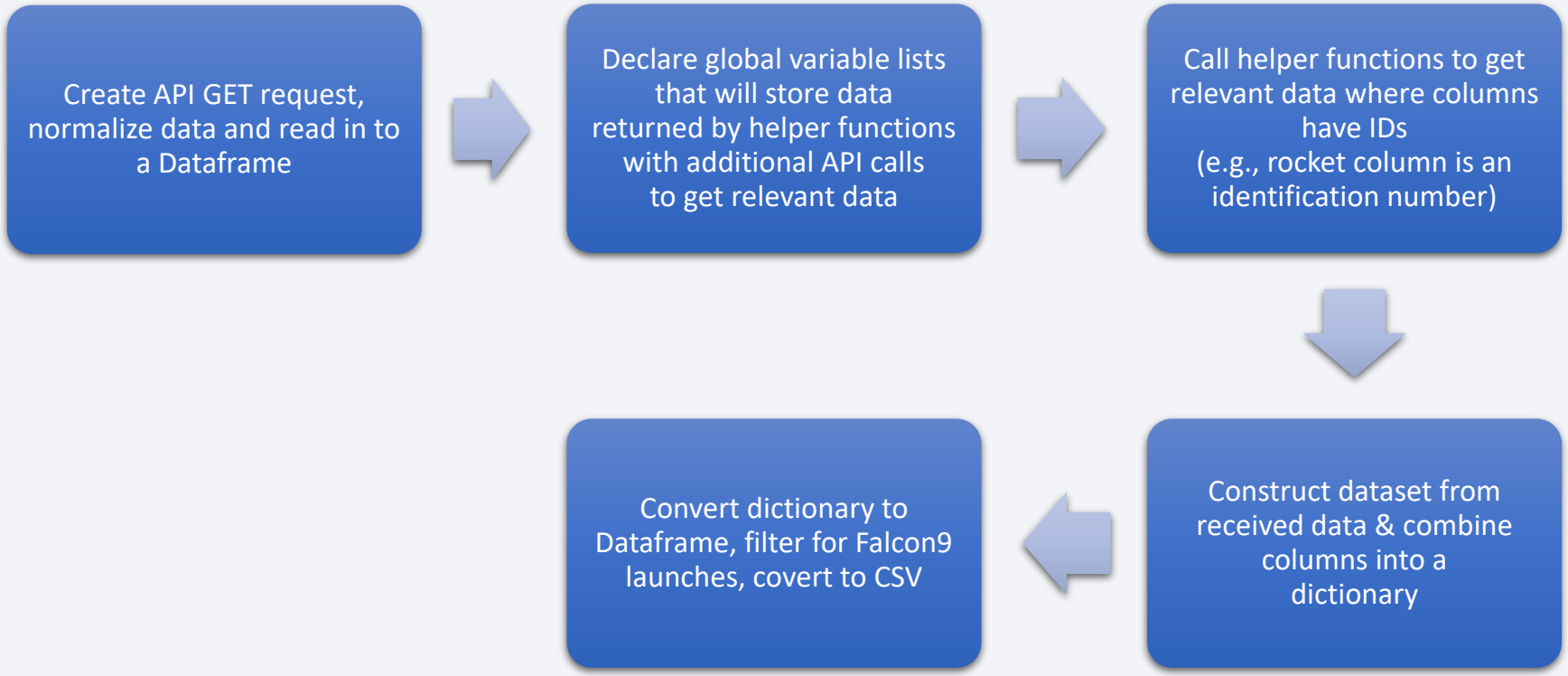
- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

SpaceX API:



Web scraping data from Wiki:

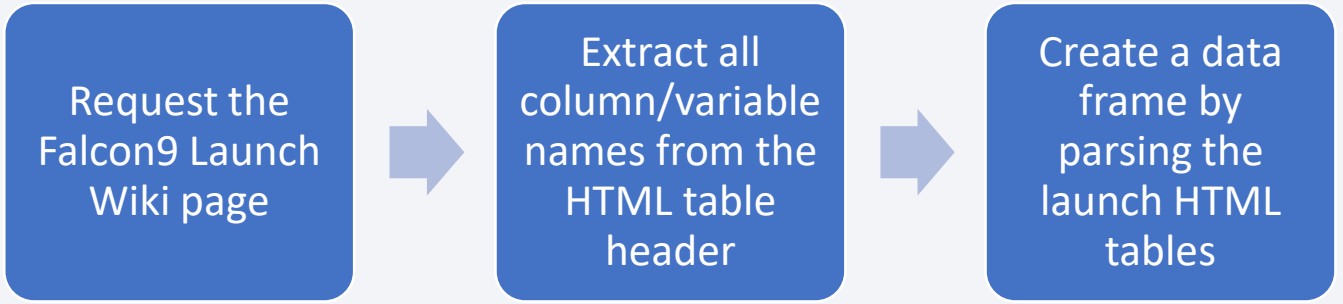
Data Collection – SpaceX API



- GitHub

Data Collection - Scraping

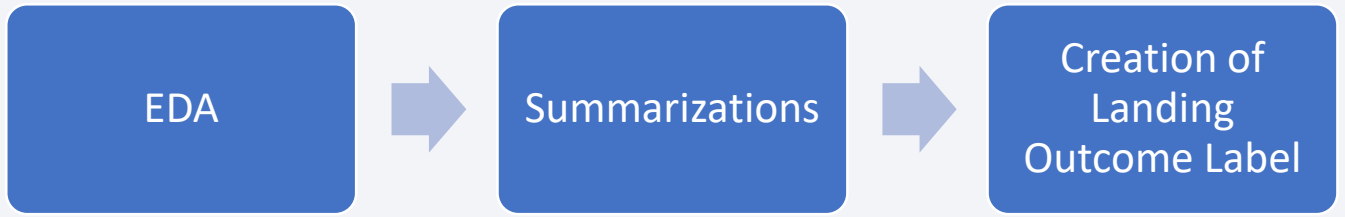
- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.



- GitHub

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 calculated.
- Finally, the landing outcome label was created from Outcome column.



- GitHub

EDA with Data Visualization

- To explore data, scatterplots and barplots 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,

EDA with SQL

- The following SQL queries were 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 in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (groundpad)) between the date 2010-06-04 and 2017-03-20.

- GitHub

Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities. We answered some question for instance:
    - Are launch sites near railways, highways and coastlines.
    - Do launch sites keep certain distance away from cities

- GitHub

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash.
- We plotted pie charts showing the total launches by a certain sites.
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

- GitHub

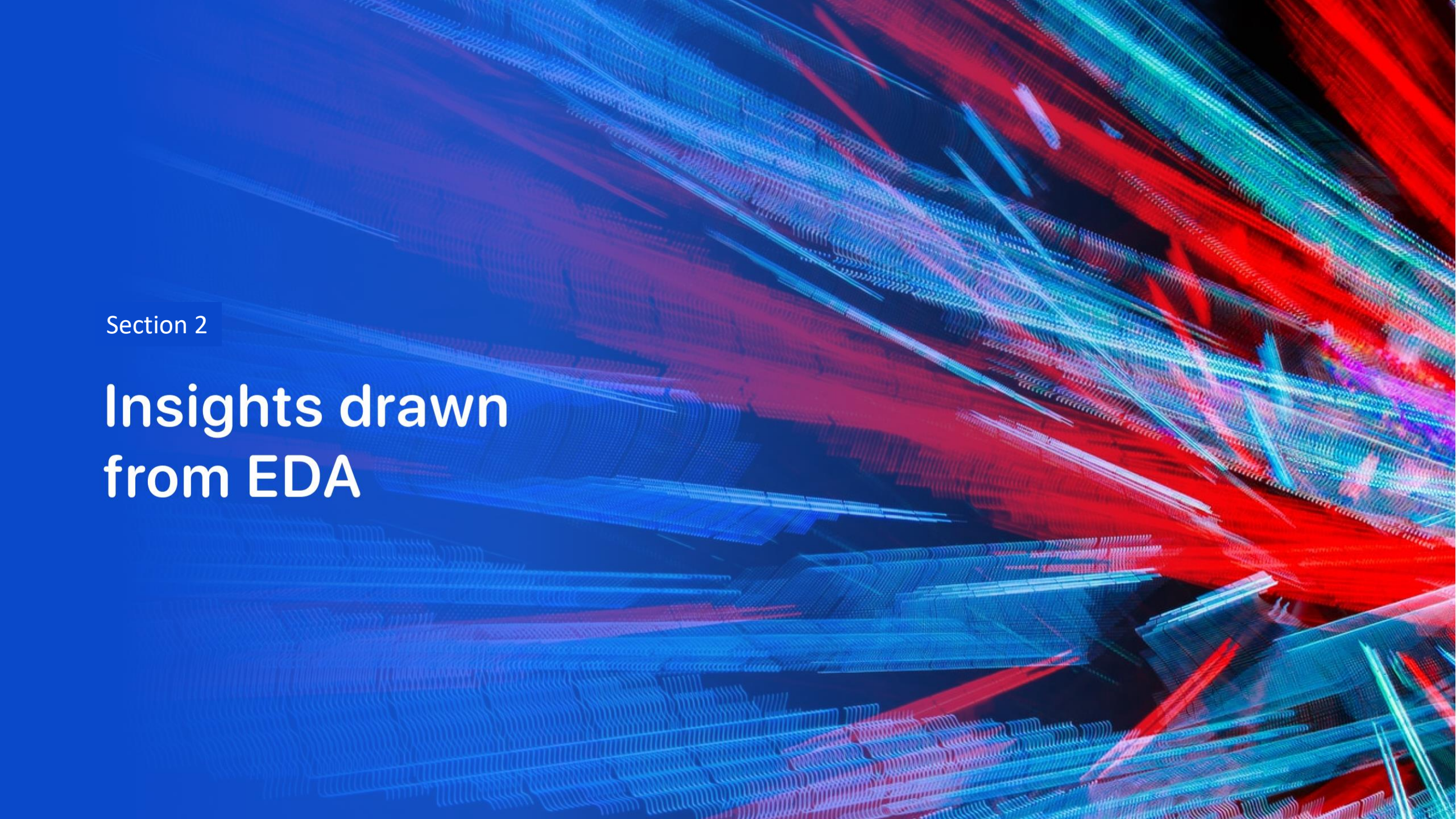
Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model.

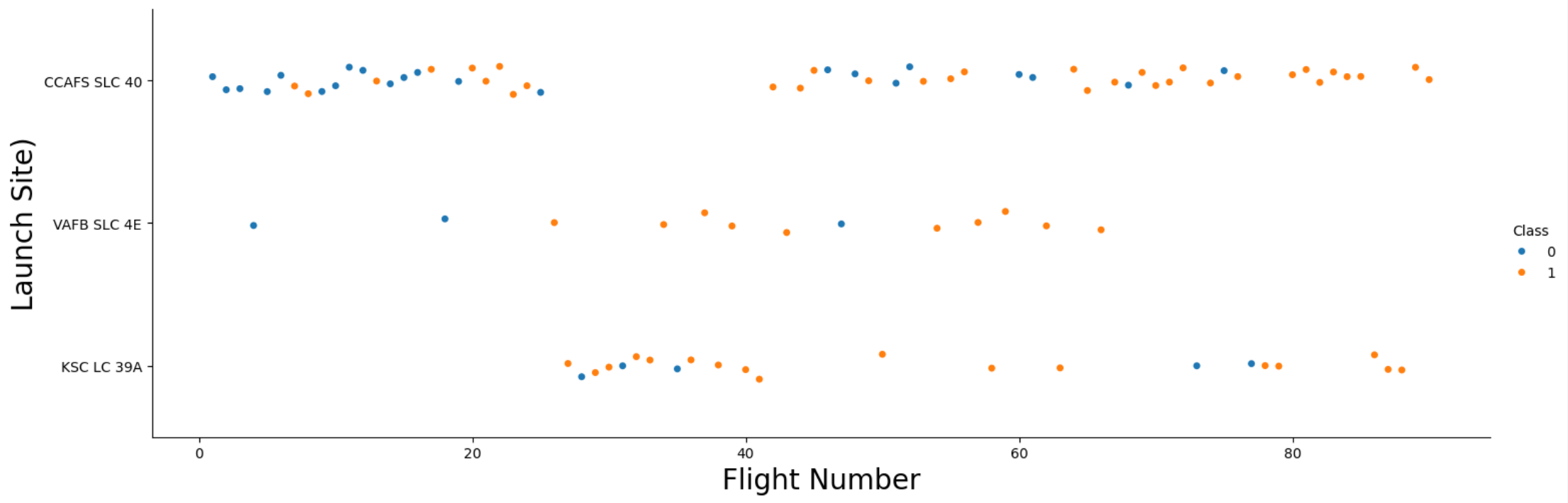
- GitHub

Results

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

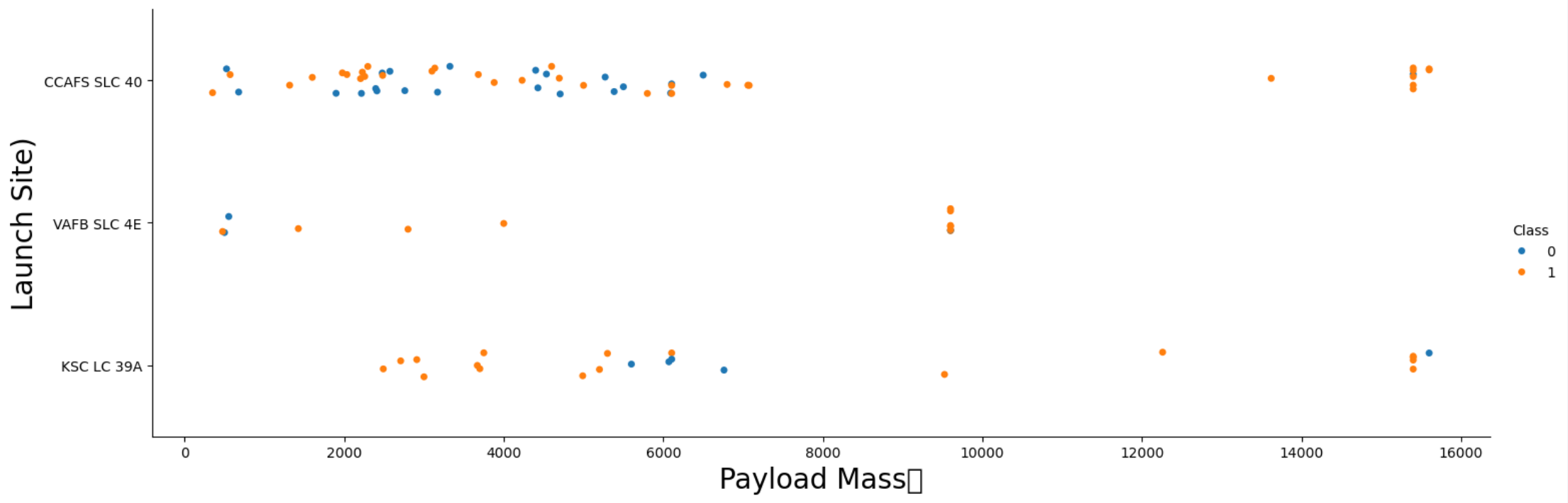

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A;
- It's also possible to see that the general success rate improved over time.

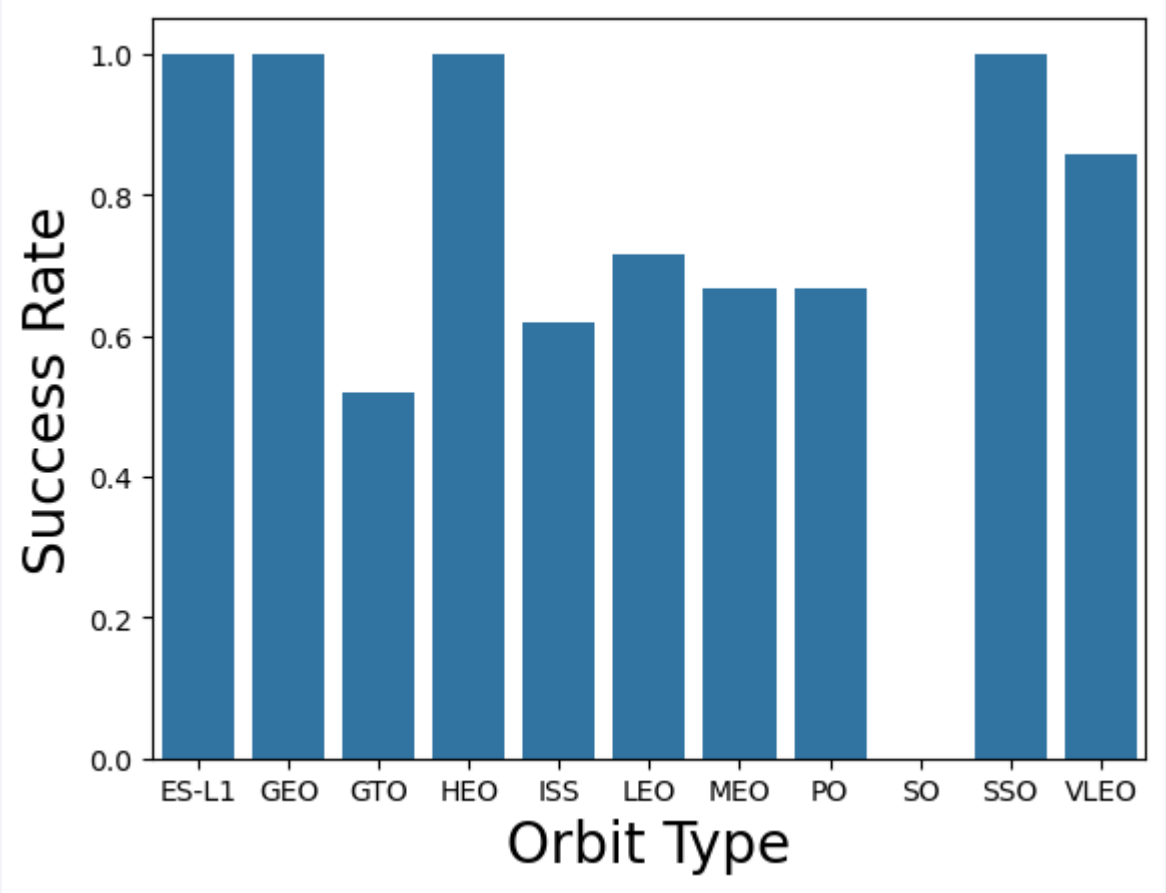
Payload vs. Launch Site



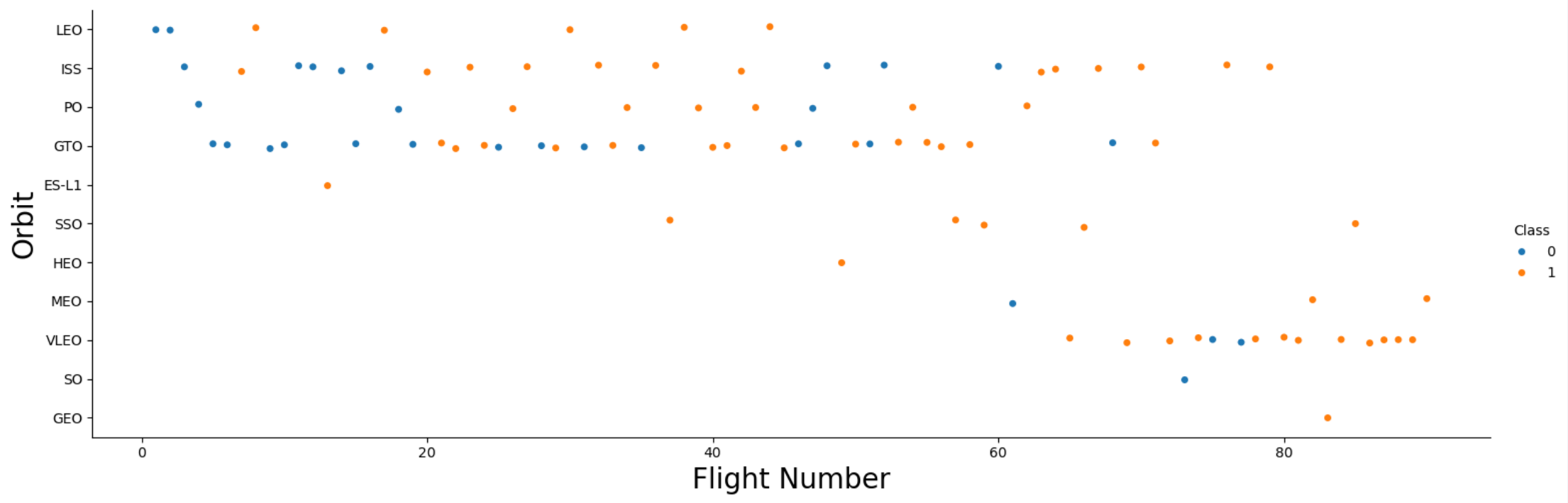
- Payloads over 9,000kg have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

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

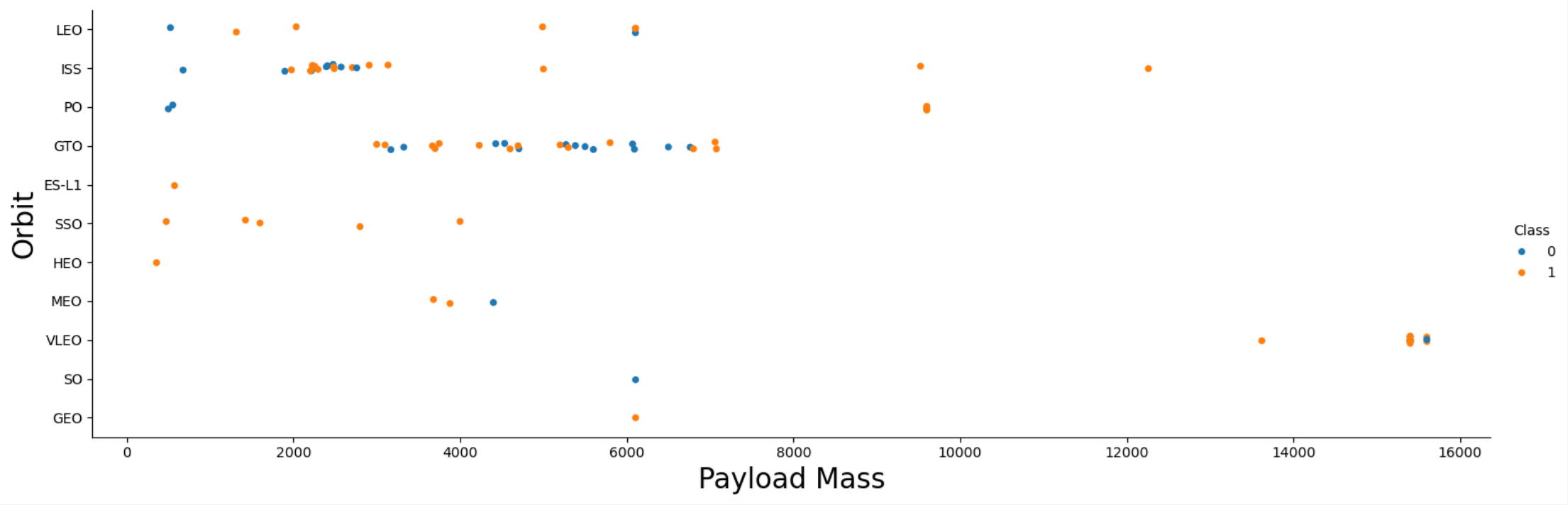


Flight Number vs. Orbit Type



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

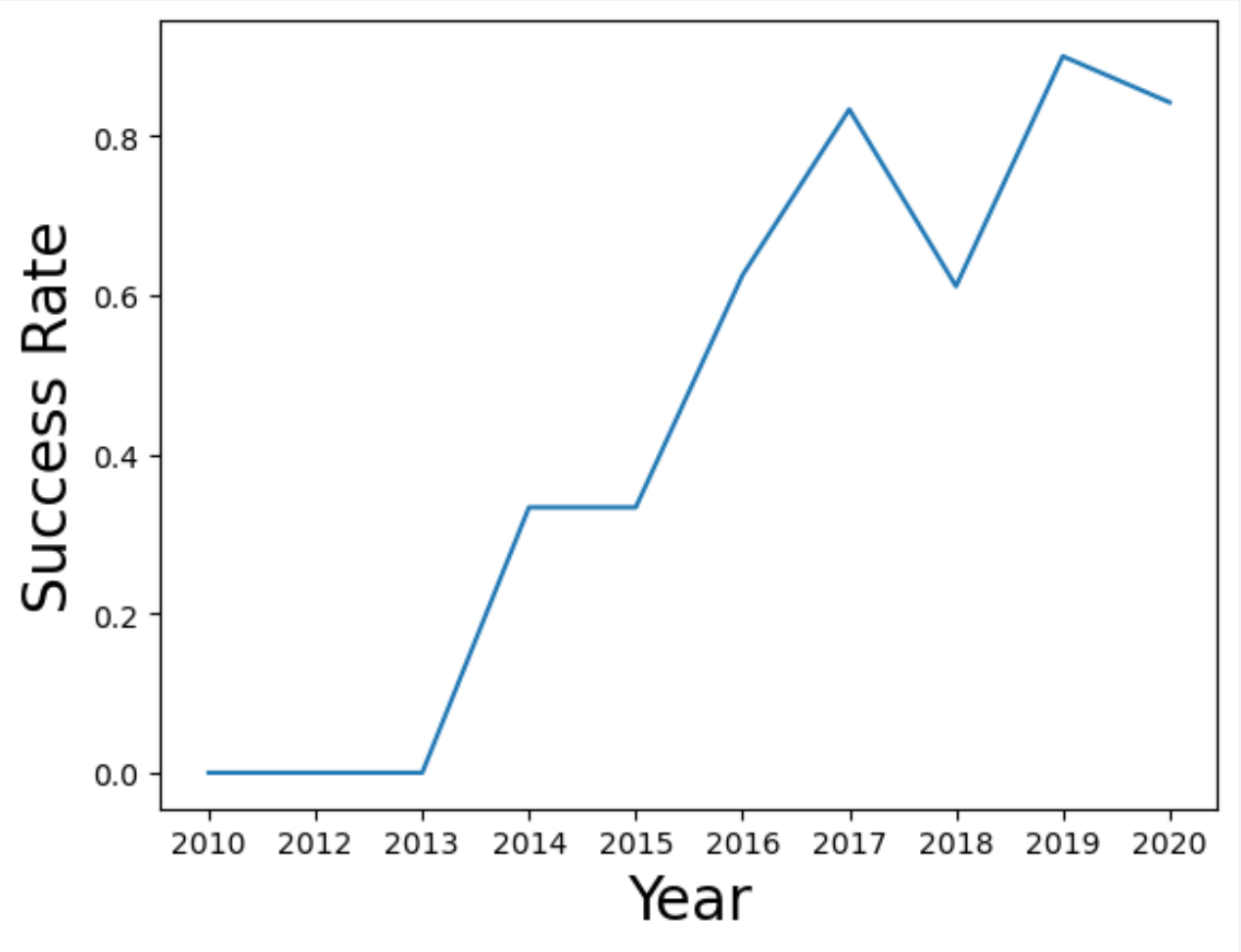
Payload vs. Orbit Type



- Apparently, 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;
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2017 (stable in 2014);
- After 2015 started increasing;
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- Query:

  pd.read_sql("select distinct Launch_Site from SPACEXTBL", con)

- Explanation:

  'distinct' returns only unique values from the queries column (Launch_Site)

- Result:

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

Launch Site Names Begin with 'KSC'

- Query:

  pd.read_sql("SELECT * FROM SPACEXTBL where launch_site like 'KSC%' Limit 5", con)

- Explanation:
  - Using keyword ‘Like’ and format ‘KSC%’, returns records where ‘Launch_Site’ column starts with “KSC”.
  - Limit 5, limits the number of returned records to 5

- Result:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
19/02/2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16/03/2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30/03/2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01/05/2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
15/05/2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- Query:

```
pd.read_sql("SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'",
con)
```

- Explanation:

  'sum' adds column 'PAYLOAD_MASS_KG' and returns total payload mass for customers named 'NASA (CRS)

- Result:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Query:

  pd.read_sql("SELECT avg(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'", con)

- Explanation:

  'avg' keyword returns the average of payload mass in 'PAYLOAD_MASS_KG' column where booster version is 'F9 v1.1')

- Result:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Query:

```
pd.read_sql("SELECT min(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)'",con)
```

- Explanation:
  - ‘min(Date)’ selects the first or the oldest date from the ‘Date’ column where first successful landing on group pad was achieved
  - Where clause defines the criteria to return date for scenarios where ‘Landing_Outcome’ value is equal to ‘Success (drone ship)
- Result:

min(Date)
08/04/2016

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query:

```
pd.read_sql("SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE
Landing_Outcome='Success (ground pad)' and PAYLOAD_MASS__KG_ between 4000 and 6000",con)
```

- Explanation:
  - The query finds the booster version where payload mass is greater than 4000 but less than 6000 and the landing outcome is success in ground pad
  - The ‘and’ operator in the where clause returns booster versions where both conditions in the where clause are true
- Result:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1032.1	5300
F9 B4 B1040.1	4990
F9 B4 B1043.1	5000

Total Number of Successful and Failure Mission Outcomes

- Query:

  pd.read_sql("SELECT substr(Mission_Outcome,1,7) as Mission_Outcome, count(*) FROM SPACEXTBL group by 1",con)

- Explanation:
  - The ‘group by’ keyword arranges identical data in a column in to group
  - In this case, number of mission outcomes by types of outcomes are grouped in column

- Result:

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

Boosters Carried Maximum Payload

- Query:

```
pd.read_sql("SELECT Booster_Version,PAYLOAD_MASS__KG_ FROM SPACEXTBL where  PAYLOAD_MASS__KG_
= (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)",con)
```

- Explanation:
  - The sub query returns the maximum payload mass by using keywork ‘max’ on the payload_mass column
  - The main query returns booster versions and respective payload mass where payload mass is maximum with value of 15600
- Result:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2017 Launch Records

- Query:

```
pd.read_sql("""SELECT    CASE substr(Date, 6, 2)         WHEN '01' THEN 'January'        WHEN '02' THEN 'February'        WHEN '03' THEN 'March'
WHEN '04' THEN 'April'        WHEN '05' THEN 'May'        WHEN '06' THEN 'June'        WHEN '07' THEN 'July'        WHEN '08' THEN 'August'
WHEN '09' THEN 'September'        WHEN '10' THEN 'October'        WHEN '11' THEN 'November'        WHEN '12' THEN 'December'    END as
Month,    Booster_Version,    Launch_Site,    Landing_OutcomeFROM    SPACEXTBLWHERE    substr(Date, 0, 5) = '2017'    AND Landing_Outcome
= 'Success (ground pad)';""",con)
```

- Explanation:
  - The query lists landing outcome, booster version, and the launch site where landing outcome is failed in drone ship and the year is 2017
  - The ‘and’ operator in the where clause returns booster versions where both conditions in the where clause are true
  - The ‘year’ keywork extracts the year from column ‘Date
  - The results identify launch site as ‘CCAFS LC-40’ and booster version as F9 v1.1 B1012 and B1015 that had failed landing outcomes in drop ship in the year 2017
- Result:

Month	Booster_Version	Launch_Site	Landing_Outcome
February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

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

- Query:

```
pd.read_sql("SELECT Landing_Outcome, count(*) FROM SPACEXTBL where Date between '2011-06-04' and '2017-03-20' group by Landing_Outcome order by 2 desc",con)
```

- Explanation:
  - The ‘group by’ key word arranges data in column ‘Landing_Outcome’ into groups
  - The ‘between’ and ‘and’ keywords return data that is between 2010-06-04 and 2017-03-20
  - The ‘order by’ keyword arranges the counts column in descending order
  - The result of the query is a ranked list of landing outcome counts per the specified date
- Result:

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

Section 3

Launch Sites Proximities Analysis

SpaceX Falcon9 - Launch Sites Map

- Figure 1 displays the Global map with Falcon 9 launch sites that are located in the United States (in California and Florida). Each launch site contains a circle, label, and a popup to highlight the location and the name of the launch site. It is also evident that all launch sites are near the coast.
- Figure 2 and Figure 3 zoom in to the launch sites to display 4 launch sites:
- • VAFB SLC-4E (CA)
- • CCAFS LC-40 (FL)
- • KSC LC-39A (FL)
- • CCAFS SLC-40 (FL)

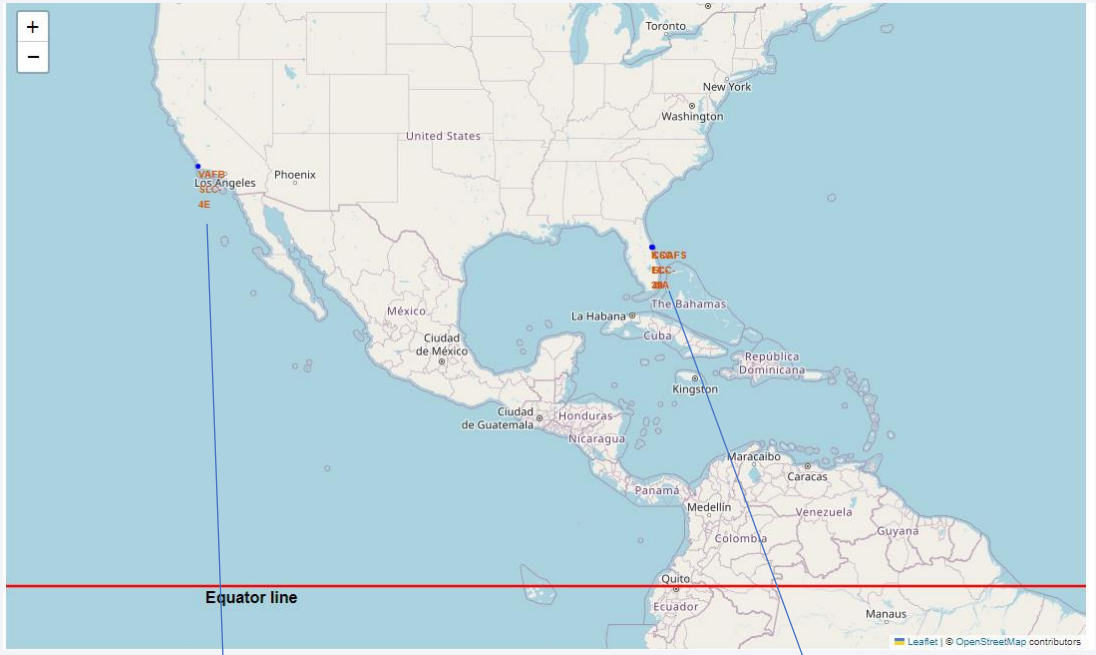


Figure 1

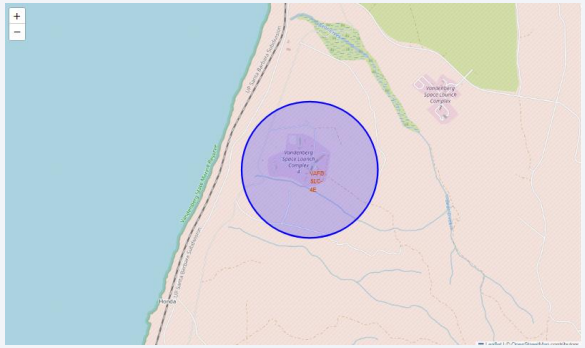


Figure 2

Figure 3

SpaceX Falon9 – Success/Failed Launch Map for all Launch Sites

- Figure 4 is the US map with all the Launch Sites. The numbers on each site depict the total number of successful and failed launches
- Figure 5 and 6 zoom in to each site and displays the success/fail markers with green as success and red as failed

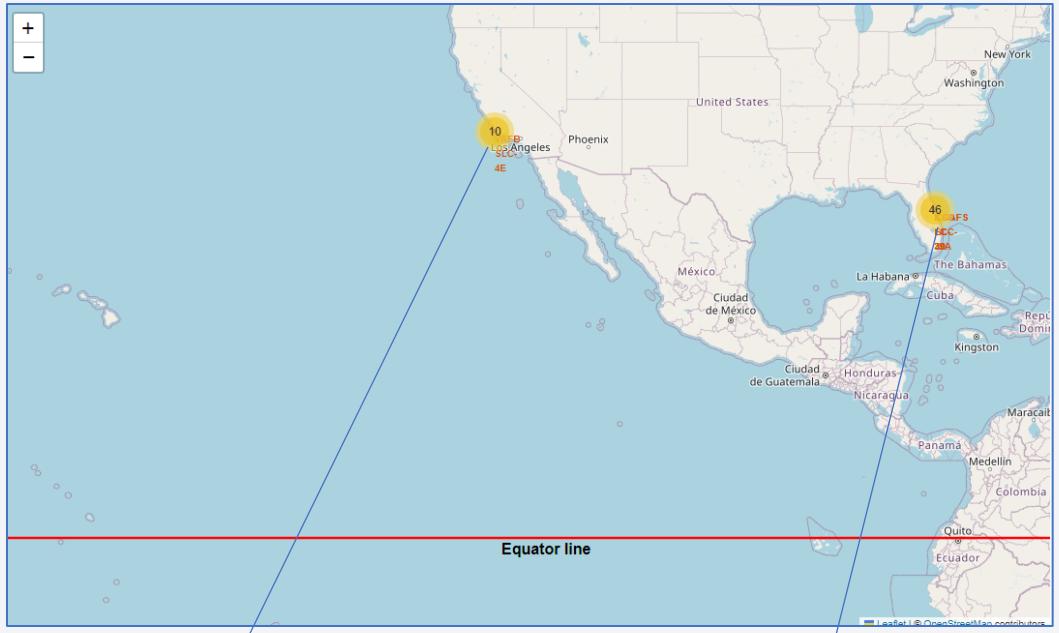


Figure 4

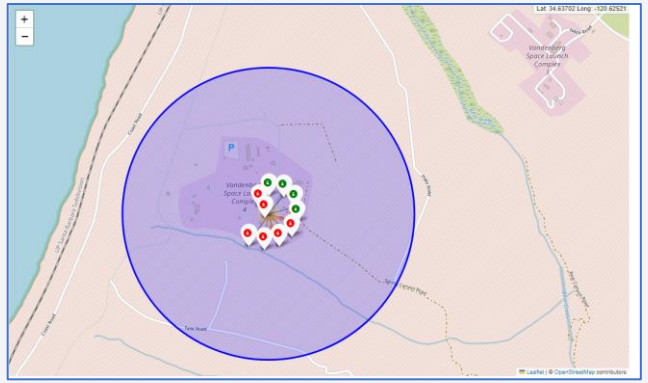


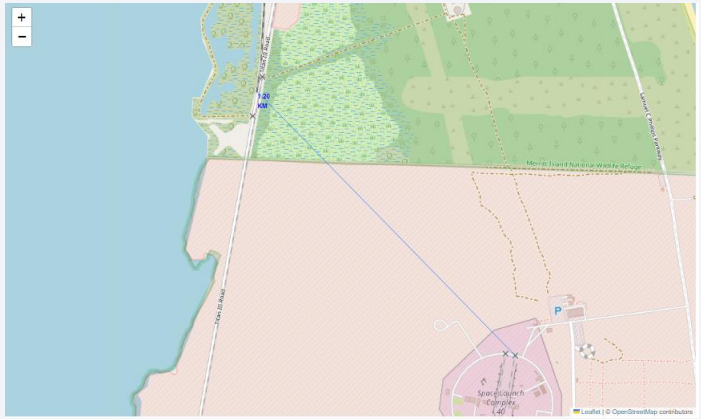
Figure 5

Figure 6

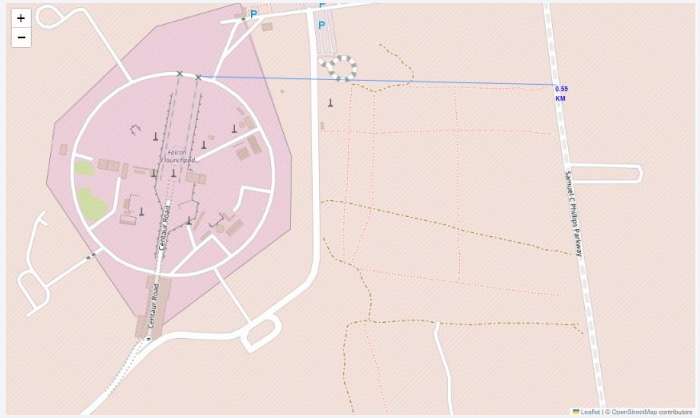
SpaceX Falcon9 – Launch Site to proximity Distance Map

- In general, cities are located away from the Launch Sites to minimize impacts of any accidental impacts to the general public and infrastructure. Launch Sites are strategically located near the coastline, railroad, and highways to provide easy access to resources.

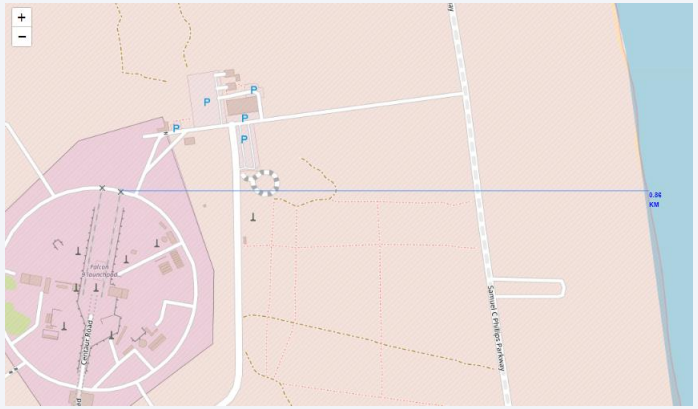
#Closest city



#Closest railway

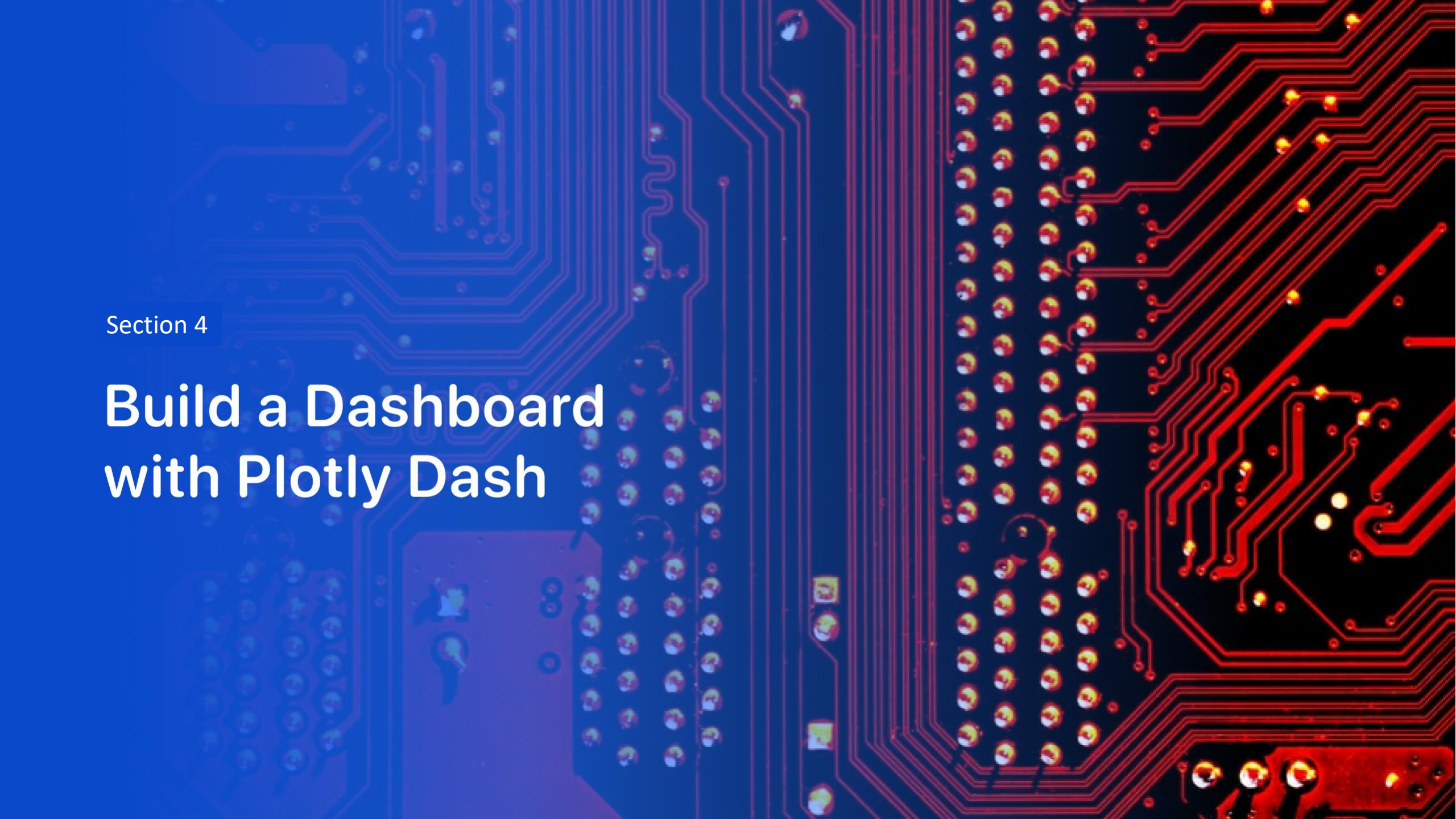


#Closest highway

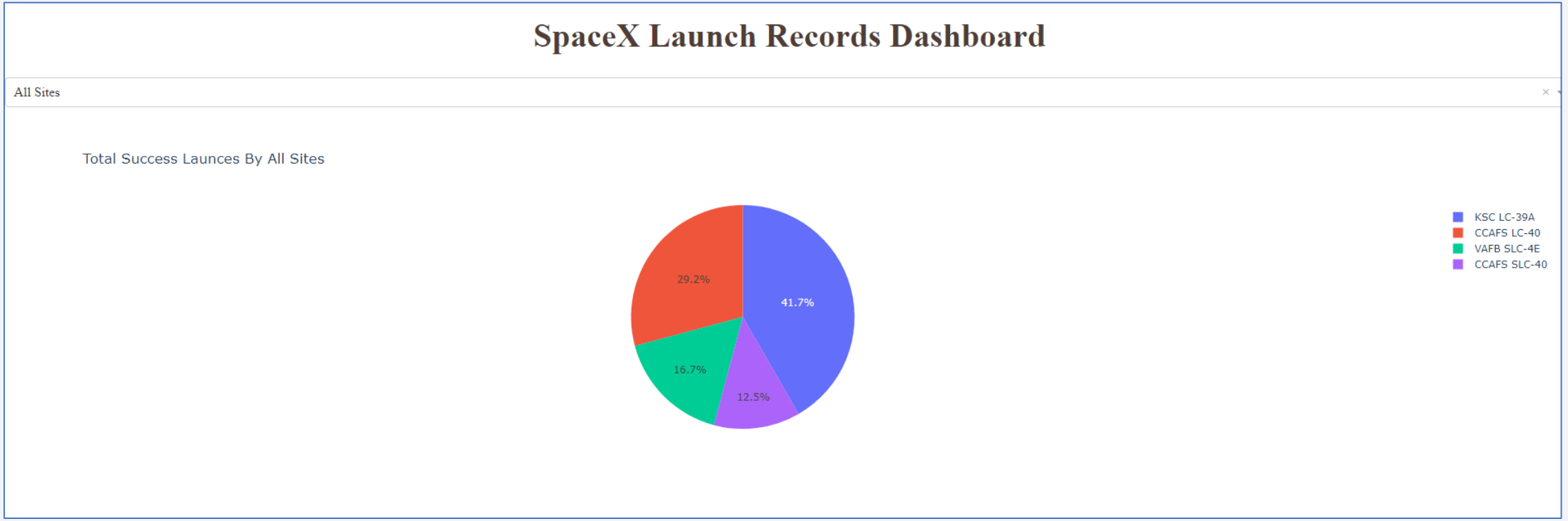


#Closest coastline

Figure 6

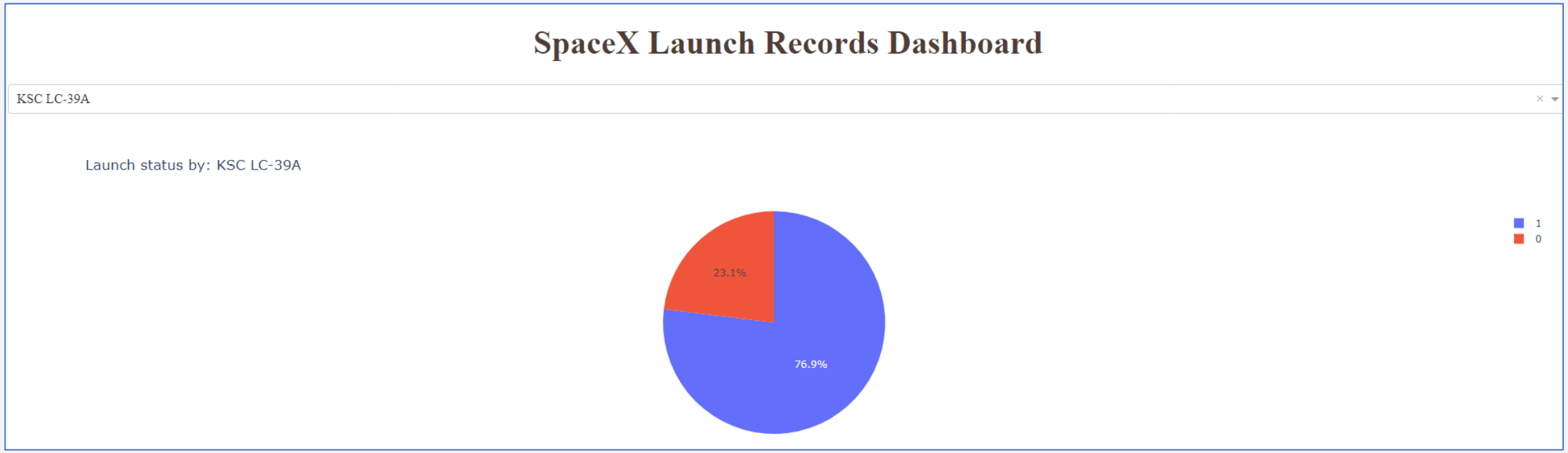

Section 4
Build a Dashboard with Plotly Dash

Launch Success Counts For All Sites



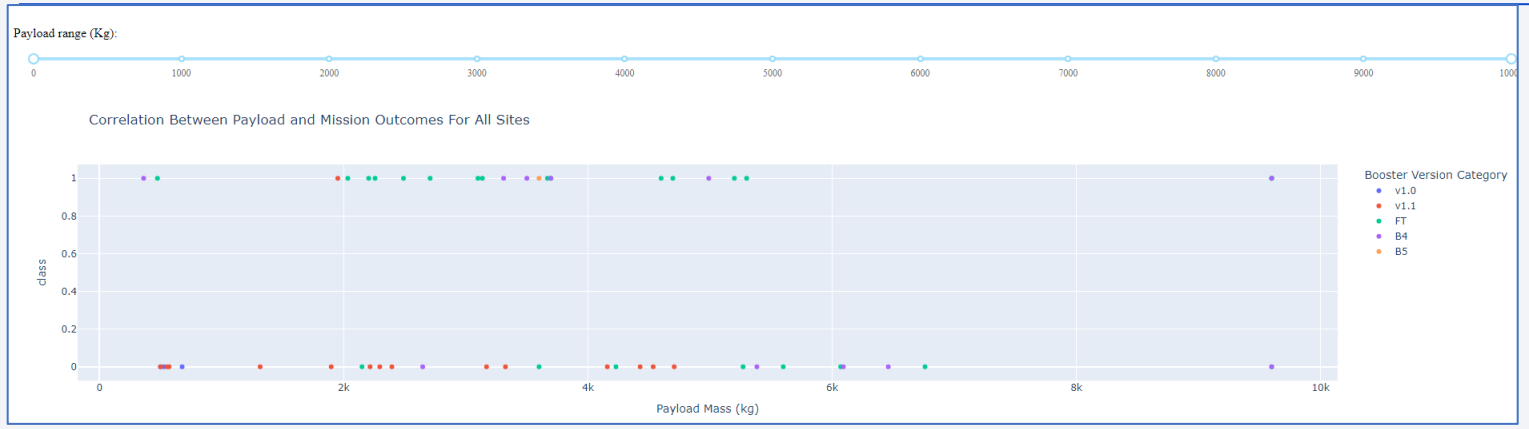
- Launch Site ‘KSC LC-39A’ has the highest launch success rate
- Launch Site ‘CCAFS SLC- 40’ has the lowest launch success rate

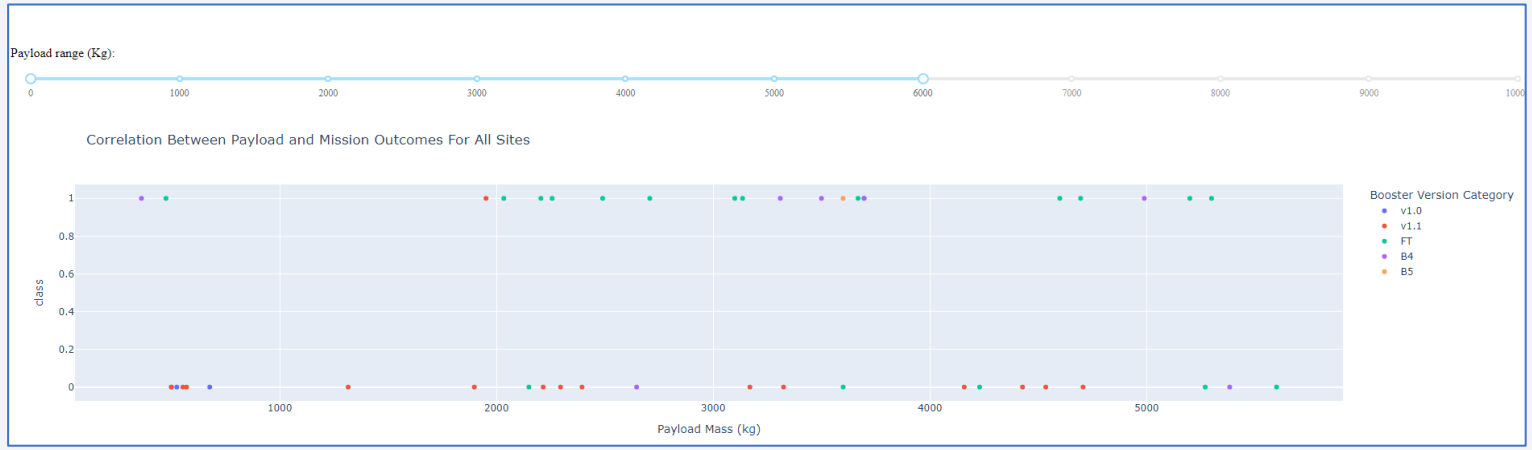
Launch Site with Highest Launch Success Ratio



- KSC LC-39A Launch Site has the highest launch success rate and count
- Launch success rate is 76.9%
- Launch success failure rate is 23.1%

Launch Site with Highest Launch Success Ratio





- Most successful launches are in the payload range from 1952 to about 5300
- Booster version category ‘FT’ has the most successful launches
- Only booster with a success launch when payload is greater than 6k is ‘B4’

Section 5

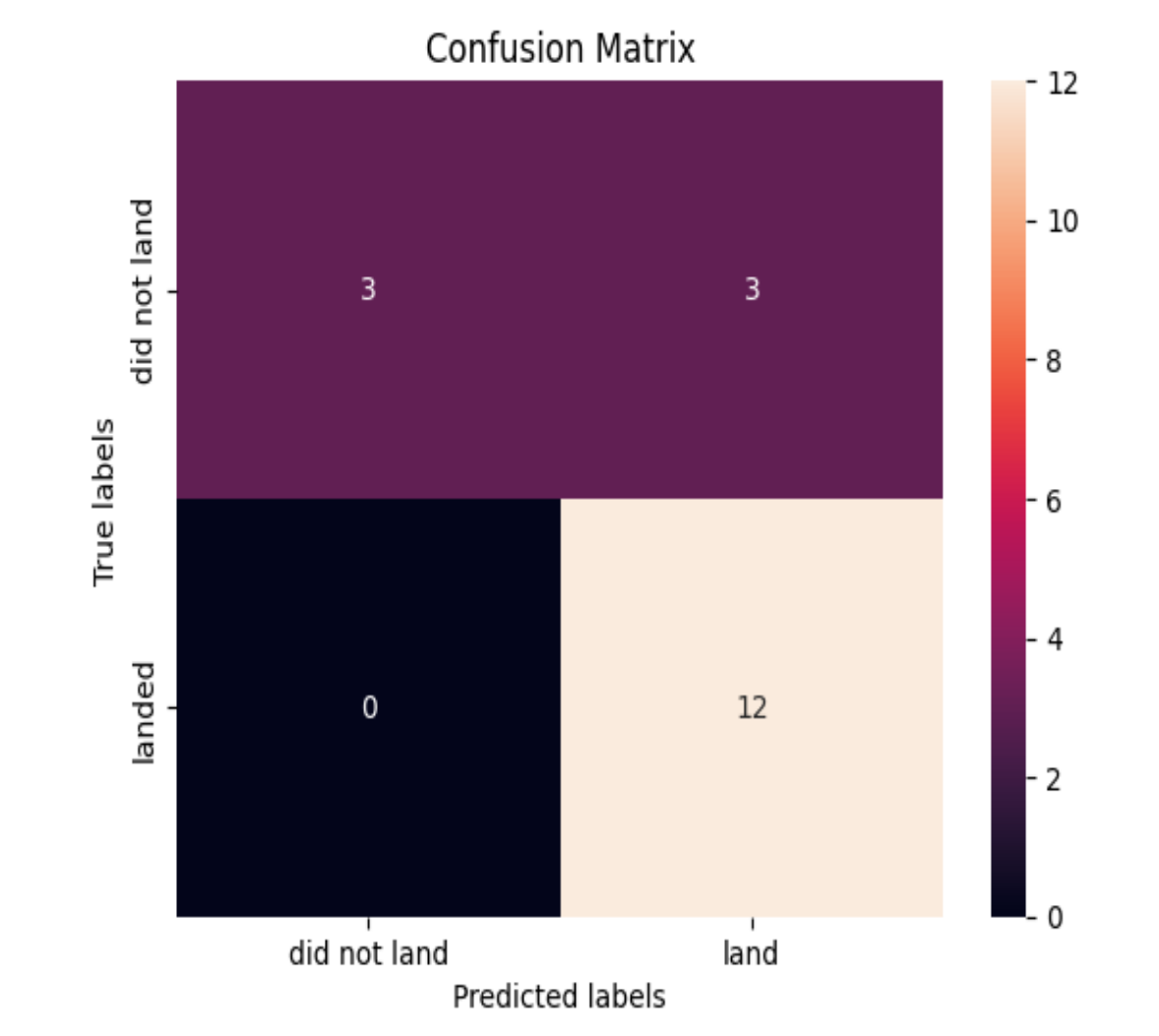
Predictive Analysis (Classification)

Classification Accuracy

Algo Type	Accuracy Score	Test Data Accuracy Score
Decision Tree	0.875000	0.833333
KNN	0.848214	0.833333
SVM	0.848214	0.833333
Logistic Regression	0.846429	0.833333

- Based on the Accuracy scores and, Decision Tree algorithm has the highest classification score with a value of .8750
- Accuracy Score on the test data is the same for all the classification algorithms based on the data set with a value of .8333
- Given that the Accuracy scores for Classfication algorithms are very close and the test scores are the same, we may need a broader data set to further tune the models

Confusion Matrix



- The confusion matrix is same for all the models (LR, SVM, Decision Tree, KNN)
- Per the confusion matrix, the classifier made 18 predictions
- 12 scenarios were predicted Yes for landing, and they did land successfully (True positive)
- 3 scenarios (top left) were predicted No for landing, and they did not land (True negative)
- 3 scenarios (top right) were predicted Yes for landing, but they did not land successfully (False positive)
- Overall, the classifier is correct about 83% of the time ((TP + TN) / Total) with a misclassification or error rate ((FP + FN) / Total) of about 16.5%

Conclusions

- As the numbers of flights increase, the first stage is more likely to land successfully
- Success rates appear go up as Payload increases but there is no clear correlation between Payload mass and success rates
- Launch success rate increased by about 80% from 2013 to 2020
- Launch Site ‘KSC LC-39A’ has the highest launch success rate and Launch Site ‘CCAFS SLC-40’ has the lowest launch success rate
- Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rates and orbit GTO the lowest
- Lunch sites are located strategically away from the cities and closer to coastline, railroads, and highways
- The best performing Machine Learning Classfication Model is the Decision Tree with an accuracy of about 87.5%. When the models were scored on the test data, the accuracy score was about 83% for all models. More data may be needed to further tune the models and find a potential better fit.

Thank you!