Winning Space Race with Data Science

Shehab Emad

7th January, 2022

Outline

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

Executive Summary

- The methodologies utilized to analyze the data were as follows:
  - ✓ Data Collection using Web Scraping and SpaceX API.
  - ✓ Exploratory Data Analysis with SQL, Data Visualizations using Plotly libraries and Folium maps and other interactive visual analytics.
  - ✓ Machine Learning Predictions using Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.
- Summary of all methods:
  - ✓ It was easy and possible to collect all necessary data.
  - ✓ Exploratory Data Analysis helped to differentiate and identify best features for Machine Learning Predictions.
  - ✓ GridSearchCV was utilized to find the best parameters for machine learning models. All four produced similar results with accuracy rate of about 83.33%.

Introduction

SpaceX is the leading company in the space travel industry. The fact that SpaceX has made space travel economical and accessible for everyone is one of the key factors that has made it a leader.

Problem

- The objective is to find the viability of the new company SpaceY to compete with SpaceX.

Desirable Outcomes

- The best way to predict the total cost of launches, by predicting successful landings of the first stage of rockets;
- To find the best location for launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data from SpaceX was gathered from two sources.

  - SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - WebScraping (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

- Perform data wrangling

  - After analysis and summarization, collected data was enhanced by the creation of a "landing outcome" label.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL

  SQL Magic was used to execute SQL codes in Jupyter Noteboook

- Perform interactive visual analytics using Folium and Plotly Dash

  Up until this point, data had been gathered, normalised, and divided into training and testing sets. Four machine learning classification models—Logistic Regression, K Nearest Neighbor, Support Vector Machine, and Decision Trees—were used for predictions. These models were tested on the test set using various evaluation metrics, including F1-scores, Jaccard scores, Classification reports, and Confusion matrices, after having been trained using the training set.

Data Collection

Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

Space X API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude and Latitude.

Wikipedia Webscrape Data Columns:

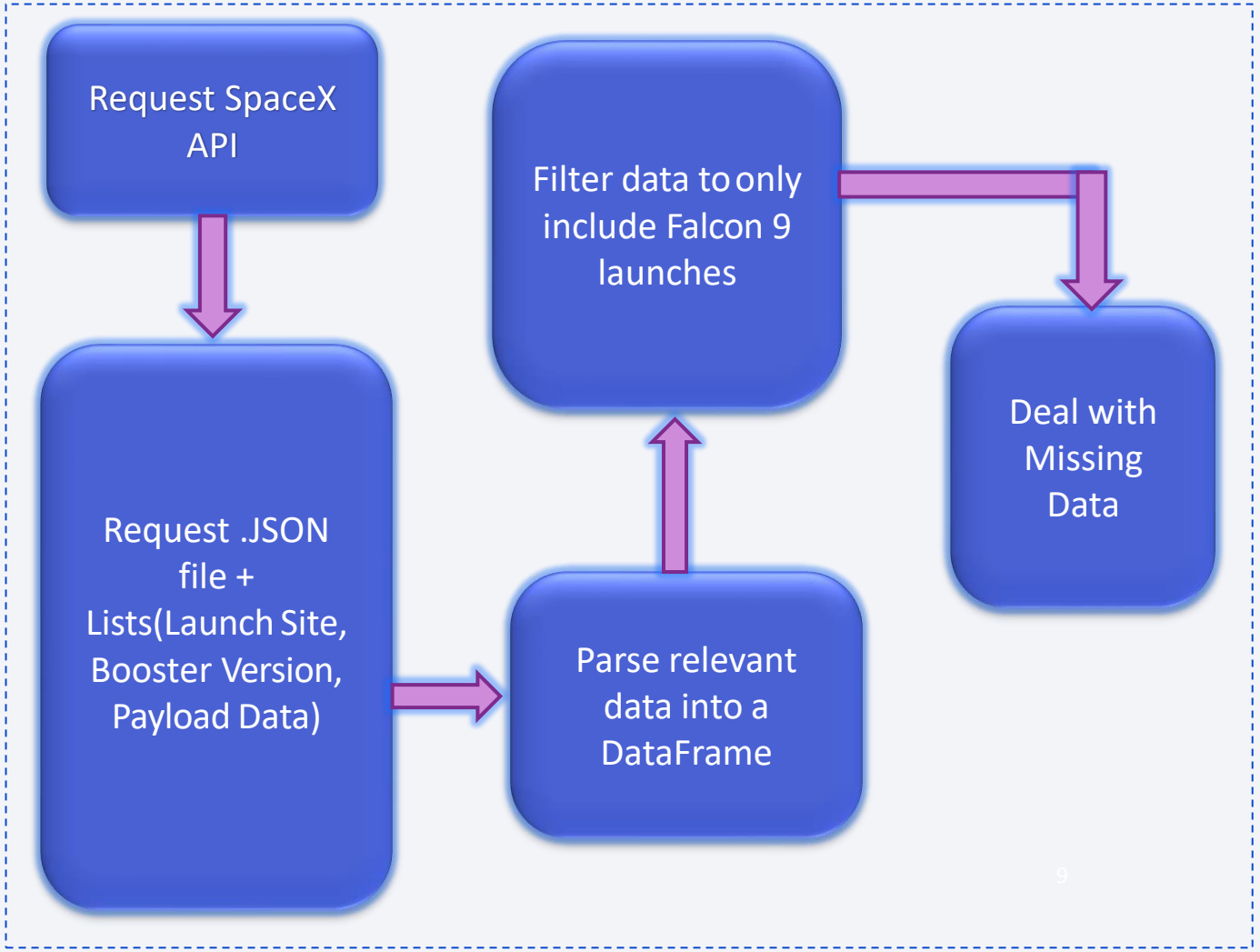
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date and Time.

Data Collection – SpaceX API

SpaceX offers a public API from where data can be extracted and then used;

The data was obtained and used according to the flowchart beside.

Source Code:

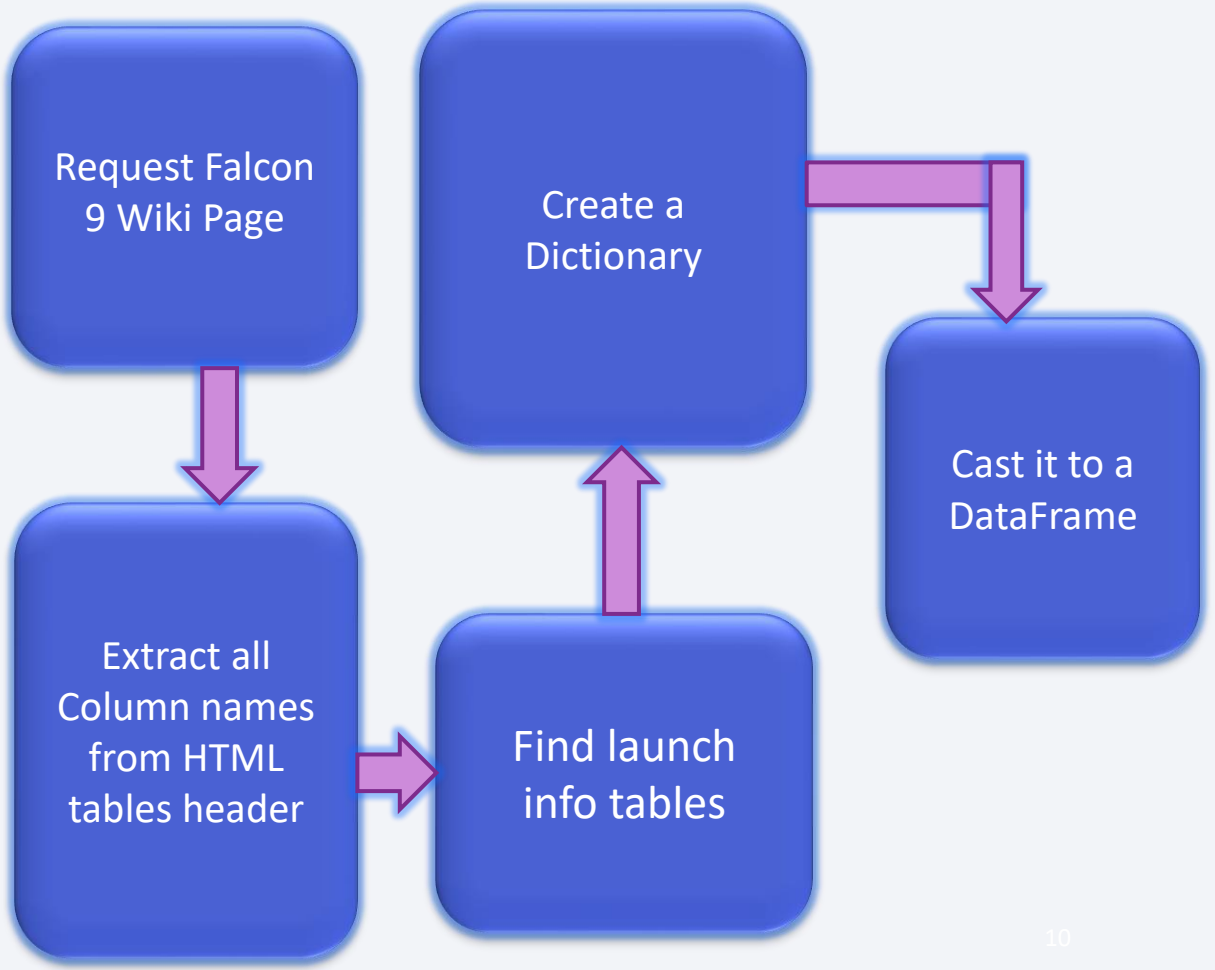


Data Collection - Scraping

Data from Wikipedia's page was also collected through WebScraping.

Data are extracted from Wikipedia according to the flowchart.

Source Code:

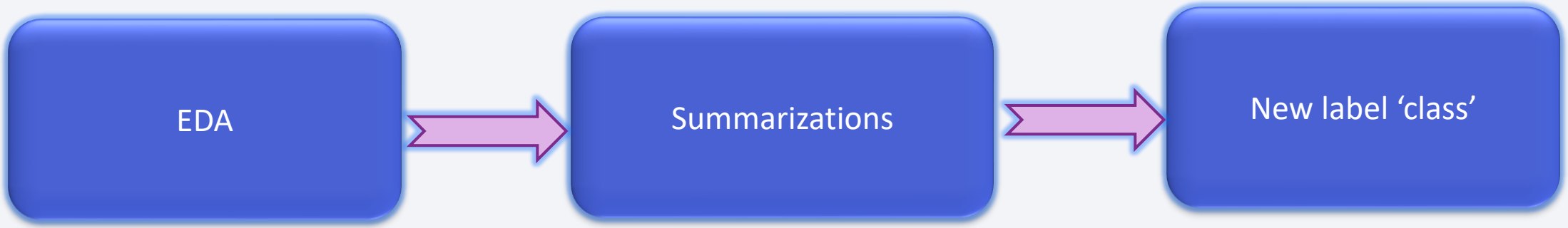


Data Wrangling

- Summaries such as ‘Launches per site’, ‘Occurrences of each orbit’ and ‘Occurrences of mission outcome per orbit’ were calculated.
- Outcome column has two components: ‘Mission Outcome’ and ‘Landing Location’.
- Finally created a new training column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise from Outcome column after performing data wrangling.

True ASDS, True RTLS, & True Ocean – set to -> 1

None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0



- *Source Code:*

EDA with Data Visualization

To explore data Scatter plots, Line charts, and Bar plots were used to compare relationships between variables to decide if any direct or indirect relationship exists.

The variables Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year were visualized graphically.

Plots Used:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend.

Source Code:

EDA with SQL

The dataset was loaded into IBM DB2 Database and it was queried using SQL Python integration. The following queries were performed:

- Names of the unique launch sites in the space mission.
- 5 records where launch sites begin with the string 'CCA'
- The total payload mass carried by boosters launched by NASA (CRS).
- Average payload mass carried by booster version F9 v1.1
- The date when the first successful landing outcome in ground pad was achieved.
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes.
- The names of the booster versions which have carried the maximum payload mass.
- the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- 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.

Source Code:

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 co-ordinates like NASA Johnson Space Centre.
- Lines speak of the distance between two co-ordinates.
- Marker Clusters indicate group of events in each co-ordinate for e.g. launches in a launch site.

Source code:

Build a Dashboard with Plotly Dash

- We have used a pie chart and a scatter plot.

- Pie chart is used to show the distribution of successful landings across all launch sites and can also be used to show individual launch site's success rates.

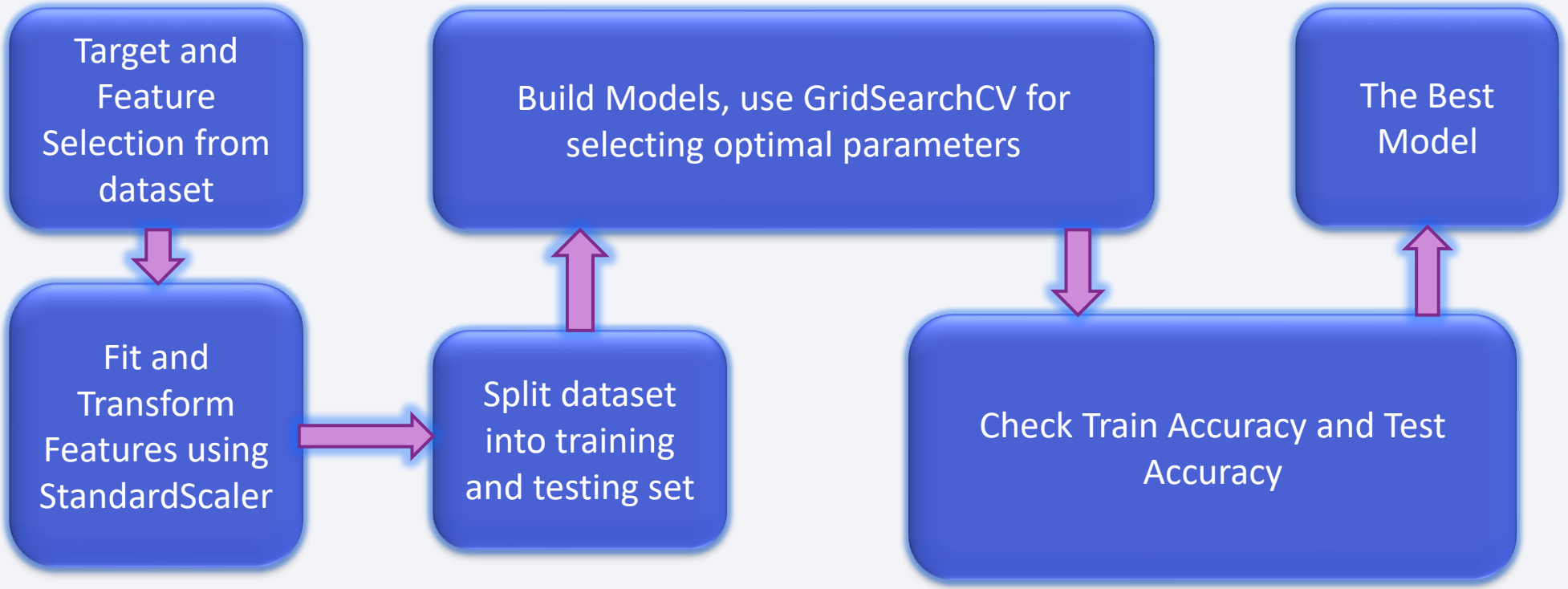
- Scatter plot takes two inputs: All sites or any individual launch site and Payload mass on a slider between 0 and 10000 kg.

- This combination allowed us to quickly analyze the relation between Payload and launch sites so that we can select best launch sites.

Source Code:

Predictive Analysis (Classification)

Four Machine Learning Models were compared- Logistic regression, Support Vector Machine, K Nearest Neighbor and Decision Trees



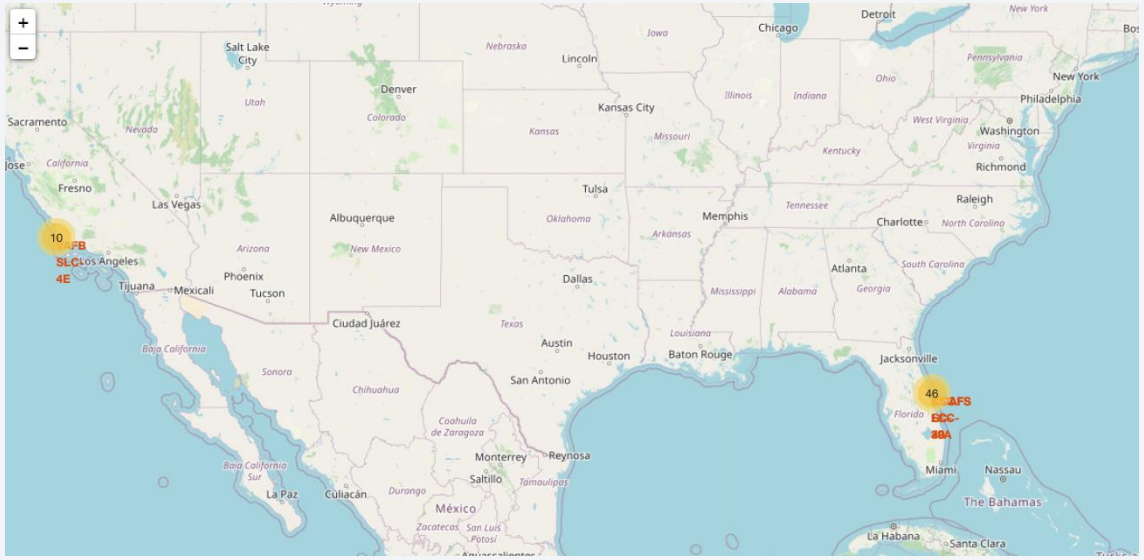
Source Code:

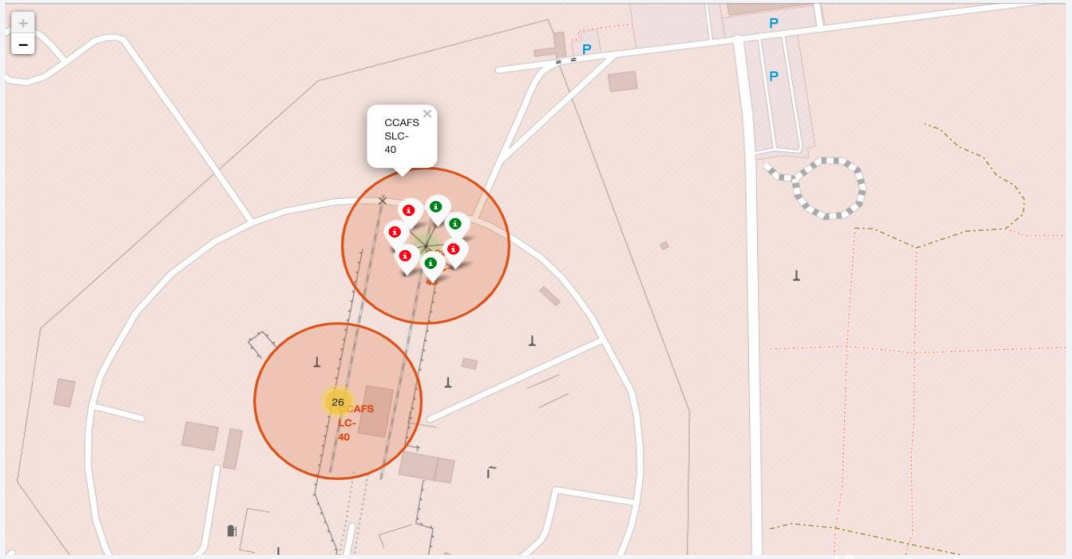
Results

- Exploratory data analysis results
  - SpaceX uses 4 different launch sites
  - The first launches were done to Space X itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 five year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

Results

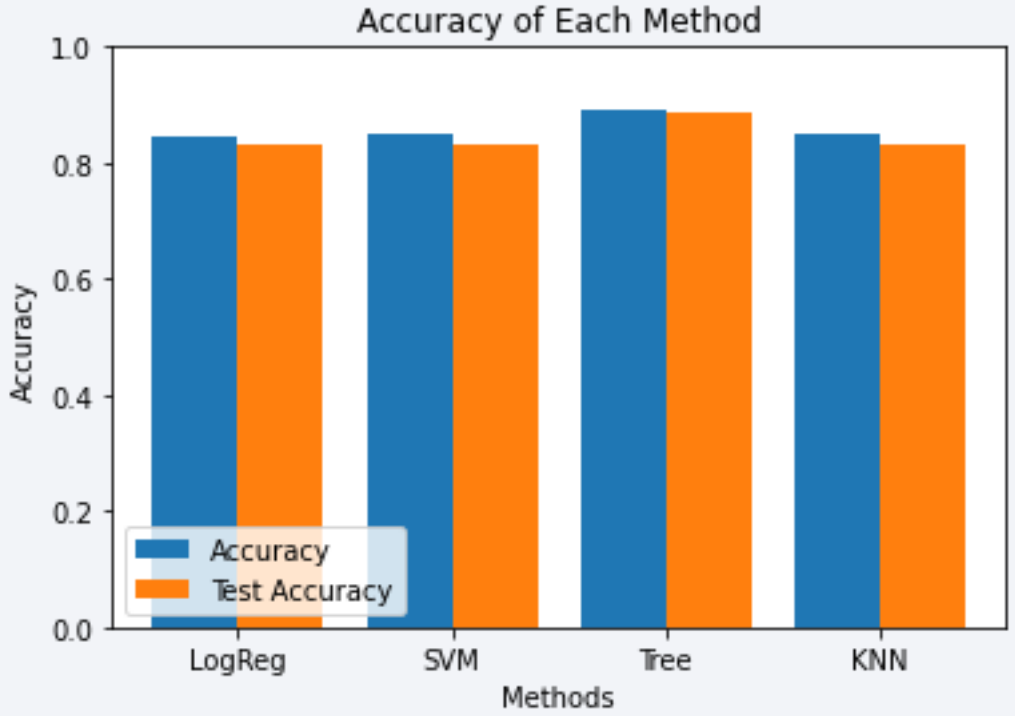
- Interactive analytics demo in screenshots
  - Interactive analytics allowed for the identification of the fact that launch sites historically had solid logistical infrastructure and were located in secure areas, such as close to the sea.
  - Most launches take place at East Coast launch locations,





Results

- Predictive analysis results
  - Decision Tree Classifier is the best model to predict successful landings, with accuracy over 88% for training data and accuracy for test data over 89%, according to predictive analysis.

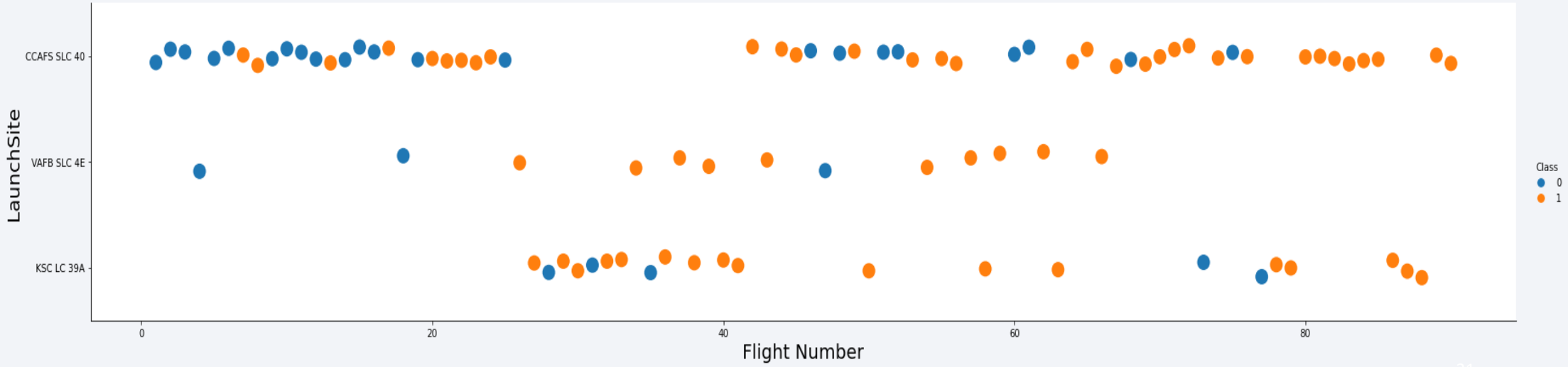


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

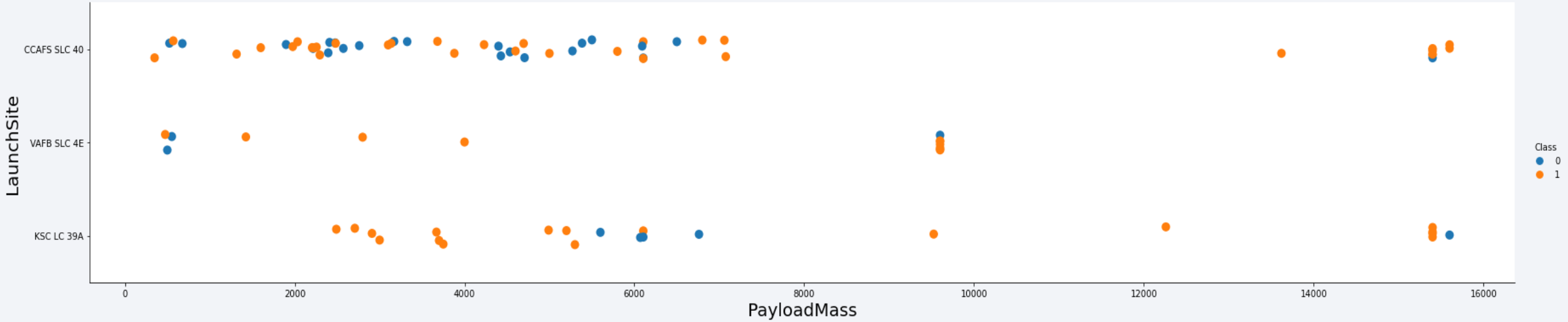
- The best launch site (nowadays) is CCAFS SLC 40 where most of the successful launches have taken place recently.
- Graphic suggests an increase in success rate over time (indicated in Flight Number).
- Most likely, there was a huge development around flight 20 that greatly improved the success rate.



Blue indicates unsuccessful launch; Orange indicates successful launch.

Payload vs. Launch Site

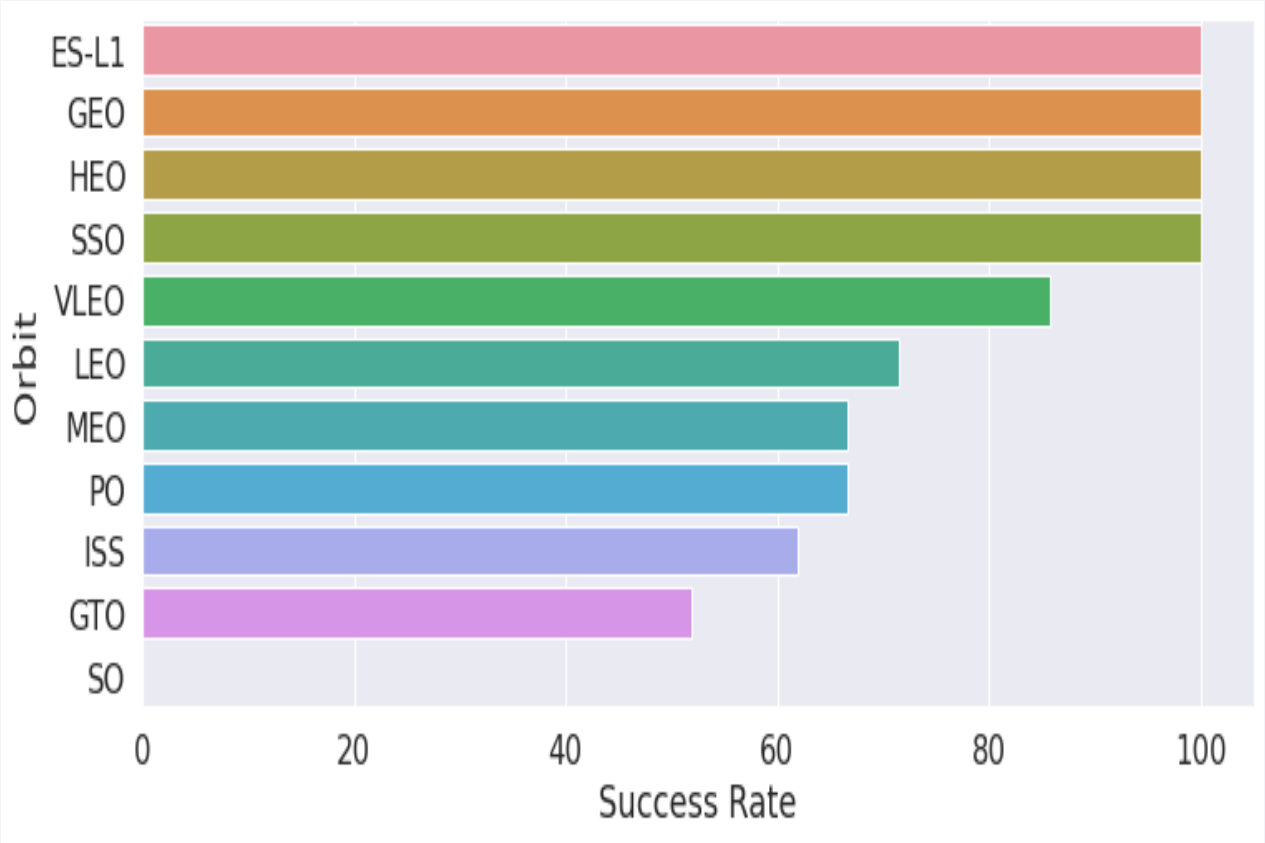
- Payload mass appears to fall mostly between 0-7000 kg.
- Payloads over 12,000kgs seems to be possible for only CCFAS SLC 40 and KSC LC 39A.



Blue indicates unsuccessful launch; Orange indicates successful launch.

Success Rate vs. Orbit Type

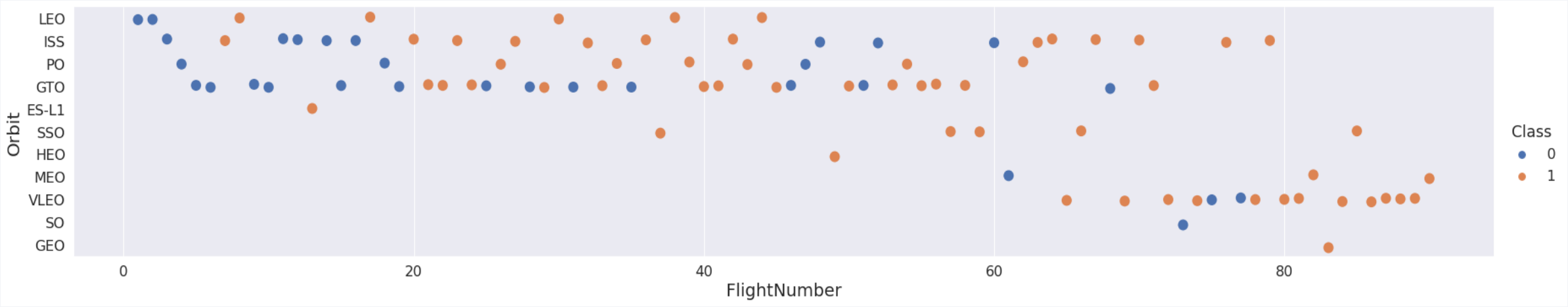
- 100% Success Rates (sample in parenthesis)
  - ✓ ES-L1(1)
  - ✓ GEO(1)
  - ✓ HEO(1)
  - ✓ SSO(5)
- Followed by VLEO(14) which has a success rate above 80%
- LEO(7), MEO(3) and PO (9) with success rates between 65% to 70%
- ISS (21) and GTO(27) have success rates between 50% to 65%



Success Rate Scale: 0 as 0% and 1 as 100%

Flight Number vs. Orbit Type

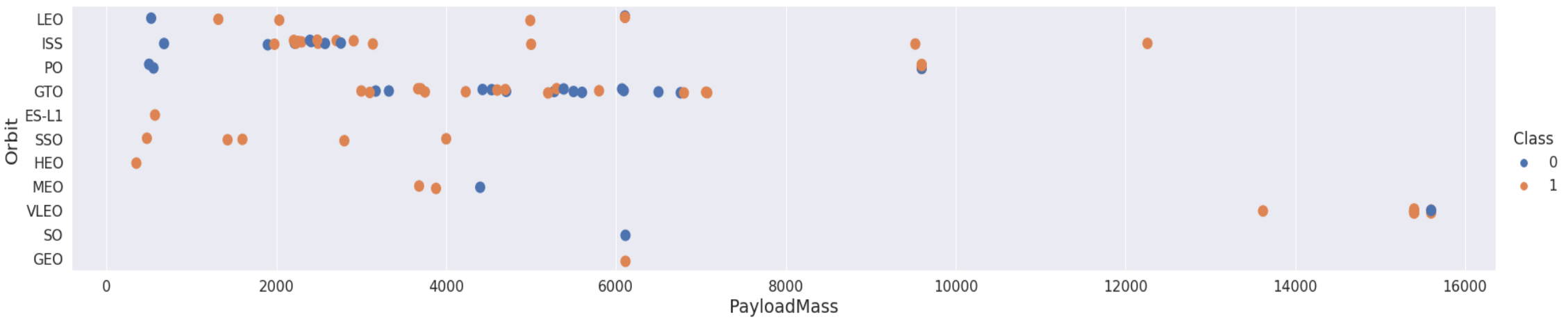
- Launch Orbit preferences changed over Flight Number.
- Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits



Blue indicates unsuccessful launch; Orange indicates successful launch

Payload vs. Orbit Type

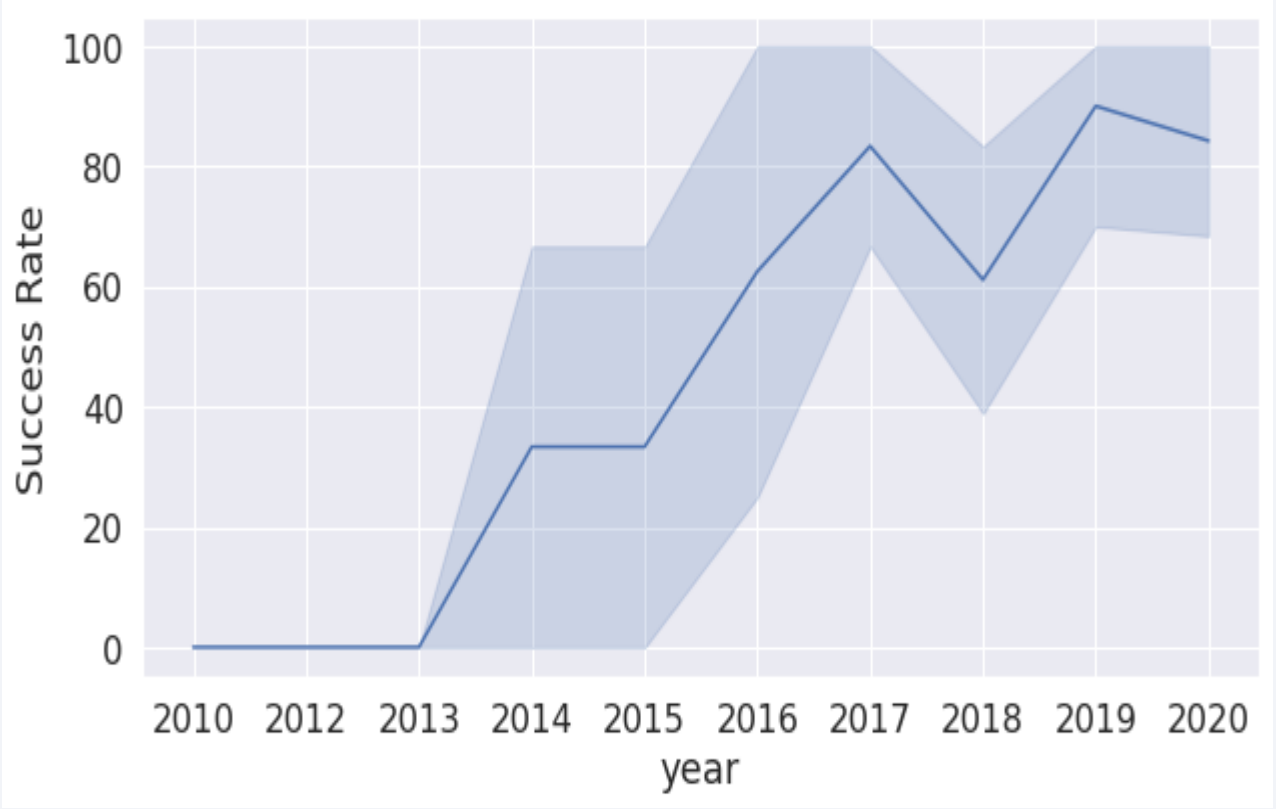
- VLEO (one of the most successful orbits) has relatively high Payload(in Kgs).
- ISS has the most diverse range of Payloads.
- GTO has a range between 3000 to 7000 Kgs
- SSO which has 100% success rate is in the range 0-4000kgs which is on the lower end.



Blue indicates unsuccessful launch; Orange indicates successful launch

Launch Success Yearly Trend

- Success rates improved over time
- The trend shows there has been a steep jump in success rates after 2015
- In recent years success rate has been around 80%.
- However there was a dip in 2018 where success rate decreased to 60%
- Success rates have only been increasing after 2013. It seems like first three years were the years of improvements and build ups of technology.



All Launch Site Names

- There are four unique launch sites as per data:
  - CCAFS LC-40(Old name for CCAFS SLC-40)
  - CCAFS SLC-40 (Cape Canaveral Space Launch Complex 40) (Brevard County, Florida)
  - KSC LC-39A (Kennedy Space Center Launch Complex 39) (Merritt Island, Florida)
  - VAFB SLC-4E(Vandenberg Space Launch Complex 4) (Vandenberg Space Force Base, California)

```
In [6]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX ORDER BY 1;
         * ibm_db_sa://yxq36934:***@125f9f61-9715-46f9-9399-c
        Done.
Out[6]:
```

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

SQL QUERY

Launch Site Names Begin with 'CCA'

Five records where launch sites begin with `CCA`

```
%sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://yxq36934:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

7]:

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

Here we can see first five entries of Cape Canaveral launches.

Total Payload Mass

This query sums the total payload mass in kg where NASA was the customer.

```
In [7]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEX WHERE CUSTOMER  LIKE '%CRS%'
 * ibm_db_sa://yxq36934:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
Out[7]:
```

total_payload
48213

Commercial Resupply Services (CRS) are a series of flights awarded by NASA for the delivery of cargo and supplies to the International Space Station (ISS)

Average Payload Mass by F9 v1.1

This query calculates average payload mass or launches which used booster version F9 v1.1

```
In [9]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1'
 * ibm_db_sa://yxq36934:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
Out[9]:
avg_payload
2928
```

Avg Payload = 2928

Average payload mass of ‘F9 1.1’ is on the low end on the payload mass range (0 to 16000)

First Successful Ground Landing Date

- First ground pad landing wasn't until the end of 2015.

```
: %sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEX WHERE LANDING__OUTCOME = 'Success (ground pad)';
 * ibm_db_sa://yxq36934:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
10]:  first_success_gp
      2015-12-22
```

First Successful Ground Landing =>
22nd December 2015

- Successful landings in general appears to start from 2014.

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

```
%%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEX
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
AND LANDING__OUTCOME = 'Success (drone ship)';
```

```
 * ibm_db_sa://yxq36934:***@125f9f61-9715-46f9-91
Done.
```

8]:

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

Total Number of Successful and Failure Mission Outcomes

- SpaceX appears to achieve successful mission outcome nearly 99% of the time.
- Interestingly, only one launch has an unclear payload status
- Unfortunately one failed in flight which can also suggest that this failure was intended.

```
: %%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEX
  GROUP BY MISSION_OUTCOME
  ORDER BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://yxq36934:***@125f9f61-9715-46f9-9399-c8177b21
Done.
```

12]:

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

Boosters Carried Maximum Payload

- This query returns the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions are very similar and all are of the F9 B5 B10xx.x versions.

```
%%sql SELECT DISTINCT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEX
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX)
ORDER BY BOOSTER_VERSION;
```

```
 * ibm_db_sa://yxq36934:***@125f9f61-9715-46f9-9399-c8177b21803b.c1og
Done.
```

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

- This likely indicates payload mass correlates with the booster version that is used.

2015 Launch Records

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 (stage 1 failed to land on a drone ship)

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE)
= 2015
```

```
 * ibm_db_sa://yxq36934:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- There were two such occurrences both for the launch_site CCAFS LC-40.

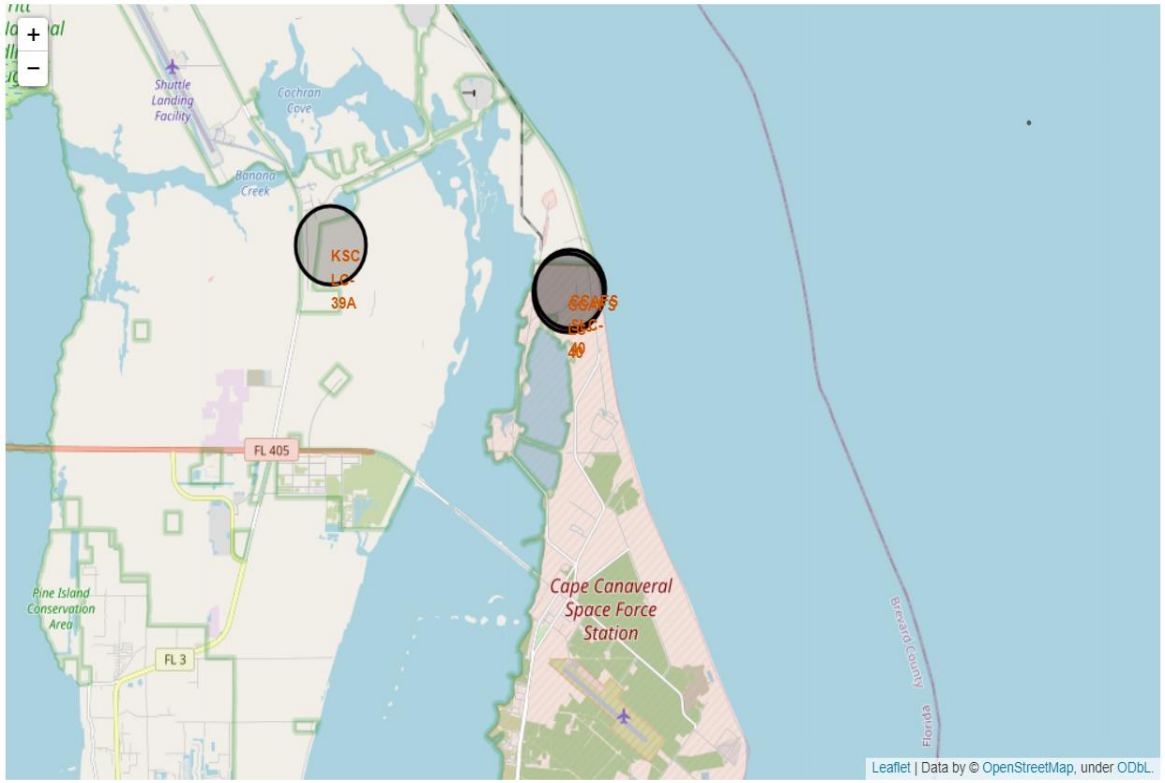
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

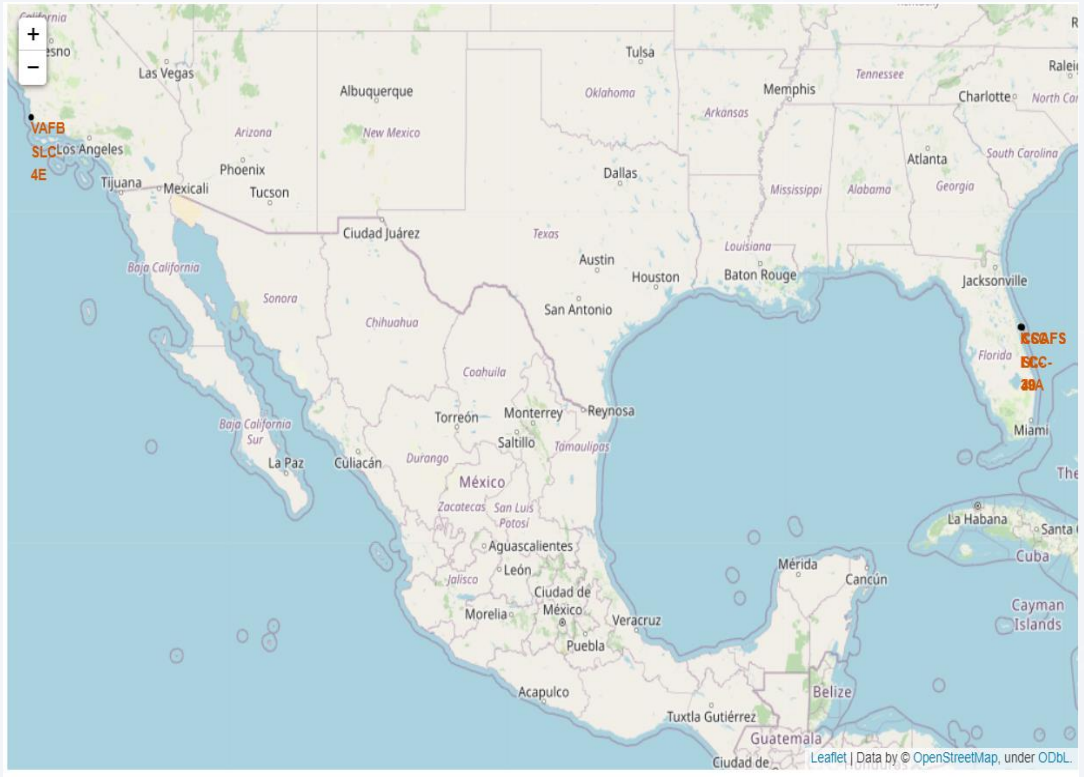
- This query returns a list of all landings between 2010-06-04 and 2017-03-20 inclusively.
- There were 8 successful landings in total during this time period.
- There were 7 unsuccessful landings in total during this time period.
- There were high numbers of 'No attempt' landing outcomes which must be taken into consideration.

landing__outcome	qty
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 Site Locations





- The left map shows the closer look of Florida launch sites
- The right map shows all launch sites relative US map.
- Launch sites are near sea, probably for safety, but are not too far away from roads and railway lines.

Launch Markers By Site

Fig (1)

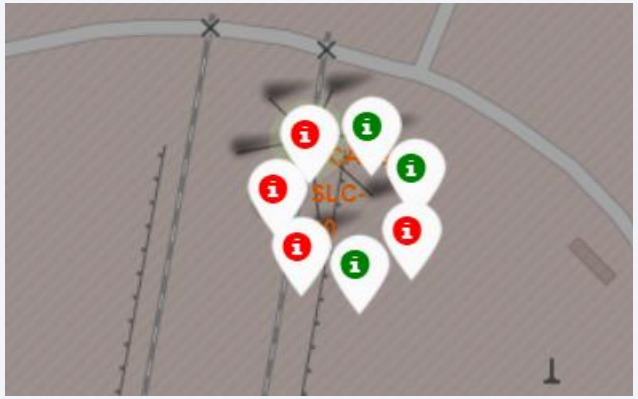


Fig (2)

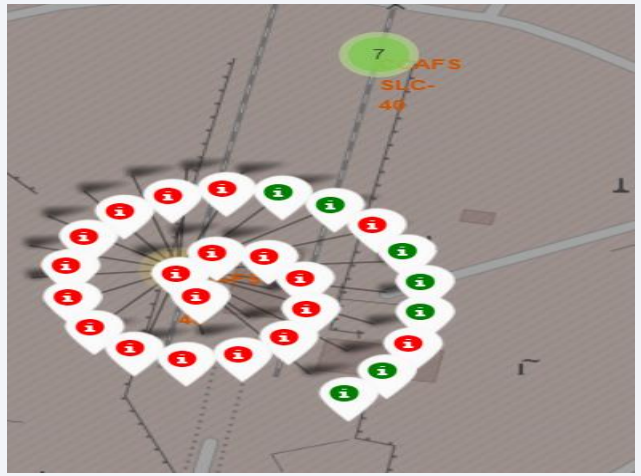


Fig (3)

- Successful landing (green icon) and failed landing (red icon).
- In this example VAFB SLC-4E (Fig 1) shows 4 successful landings and 6 failed landings.
- Fig(4) shows KSC LC-39 launch site
- Fig(2) and Fig(3) show CCAFS SLC-40 and CCAFS LC-40

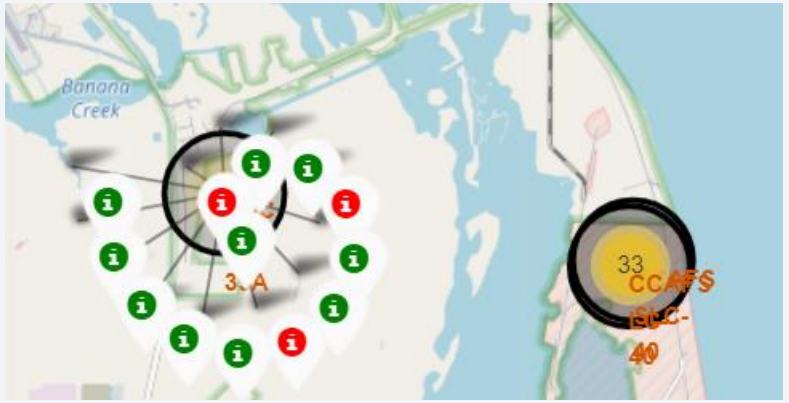
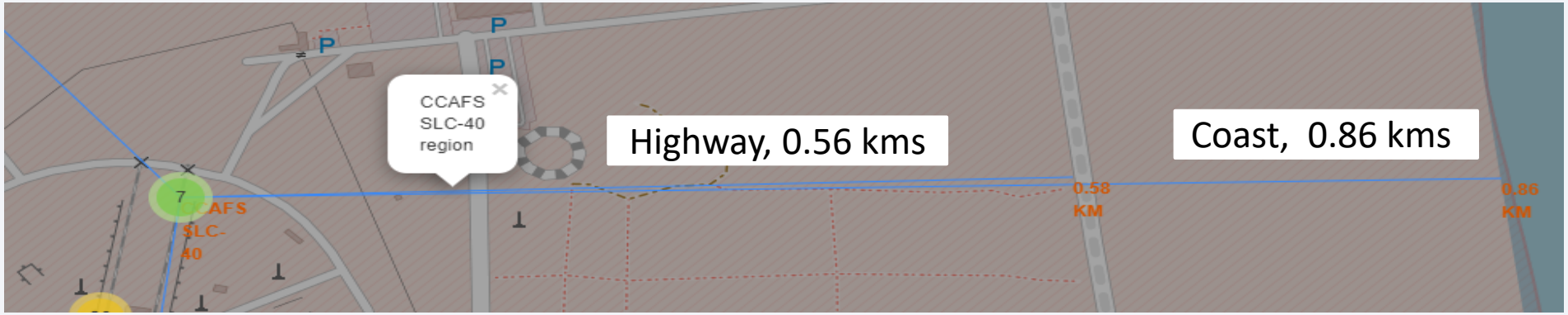
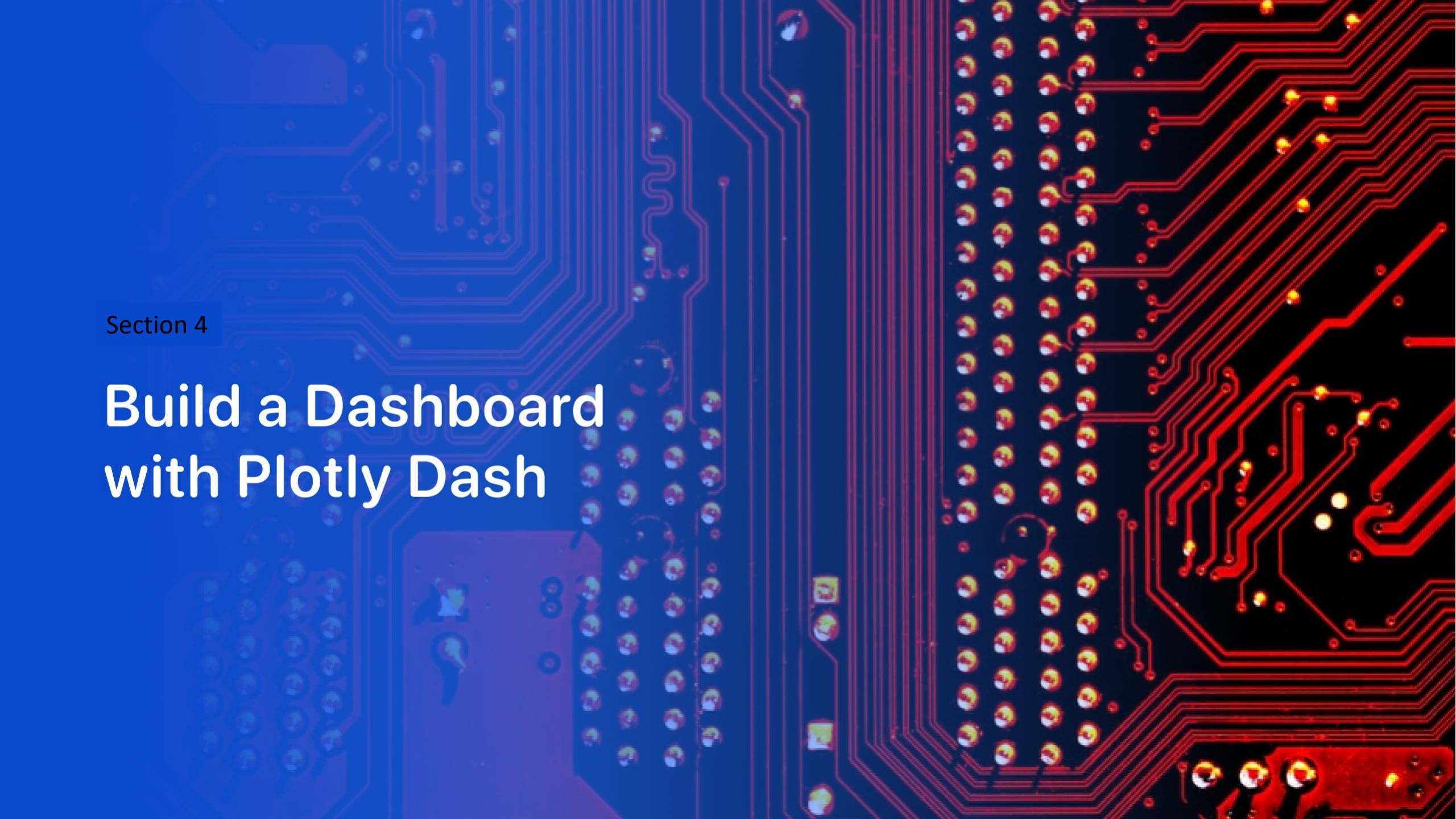


Fig (4)

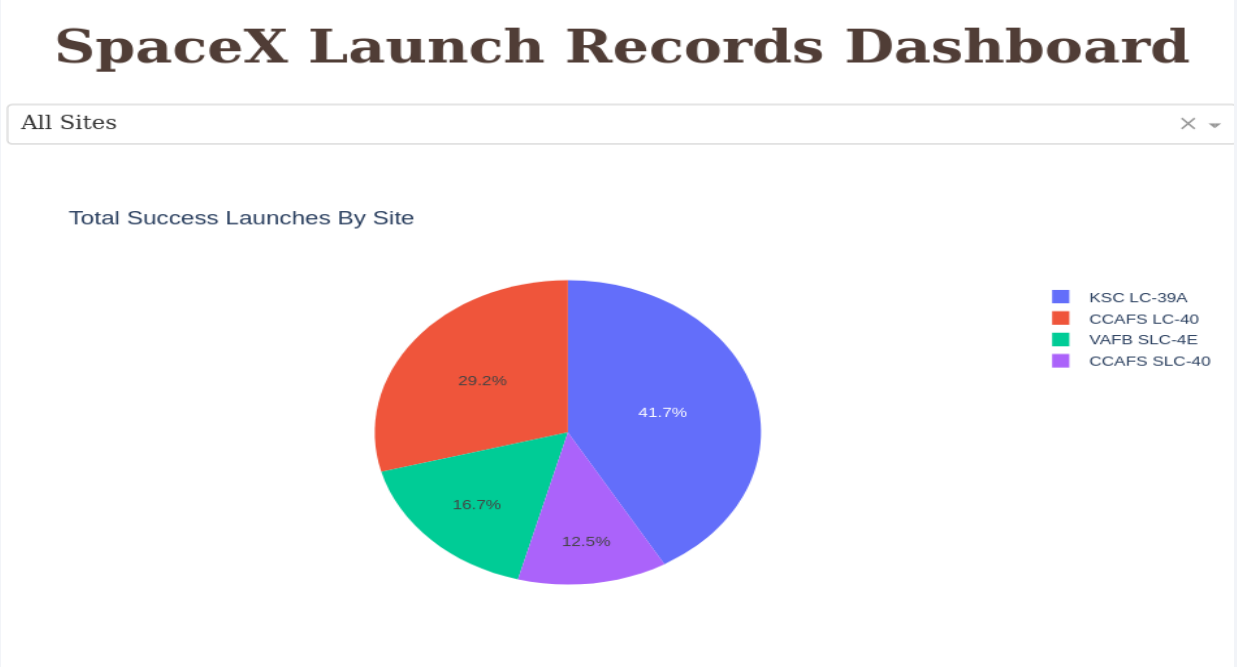
Key Location Proximities



- Here taking CCAFS SLC-40 as an example, it can be said that, launch sites are very close to railways which makes supply transportation easy and inexpensive.
- Launch sites are close to highways for human and supply transport.
- Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.


Section 4
Build a Dashboard with Plotly Dash

Successful Launches By Site

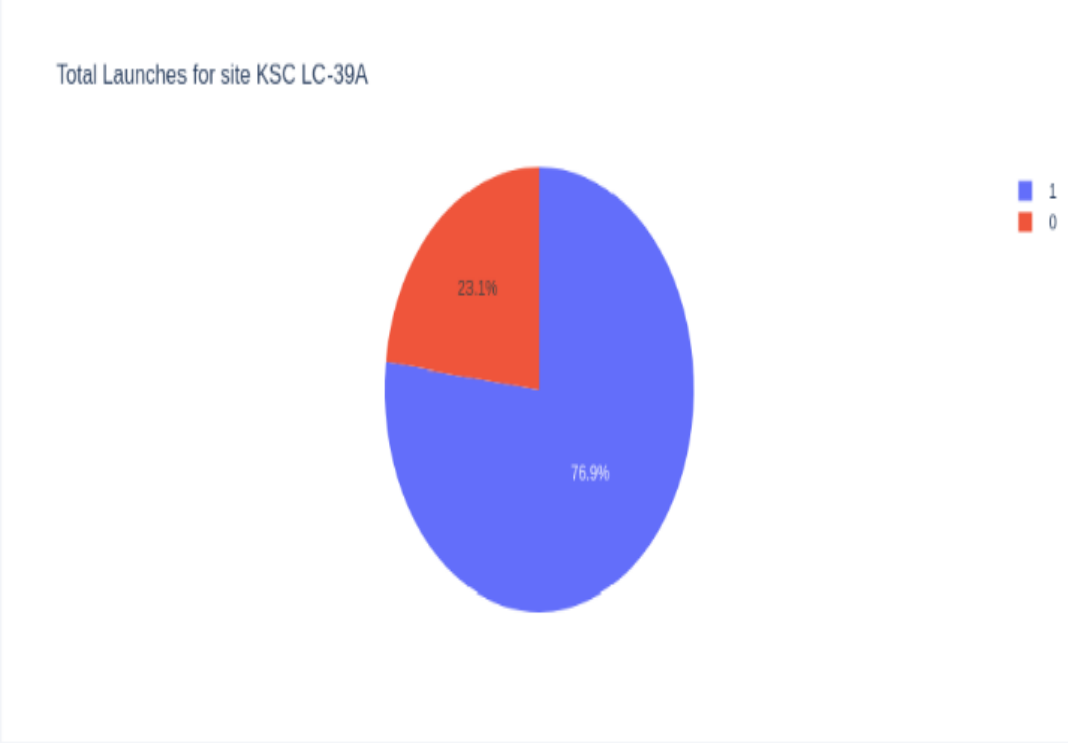


- CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC LC 39-A have the same amount of successful landings, but a majority of the successful landings were achieved before the name change.
- VAFB SLC-4E has the smallest share of successful landings.

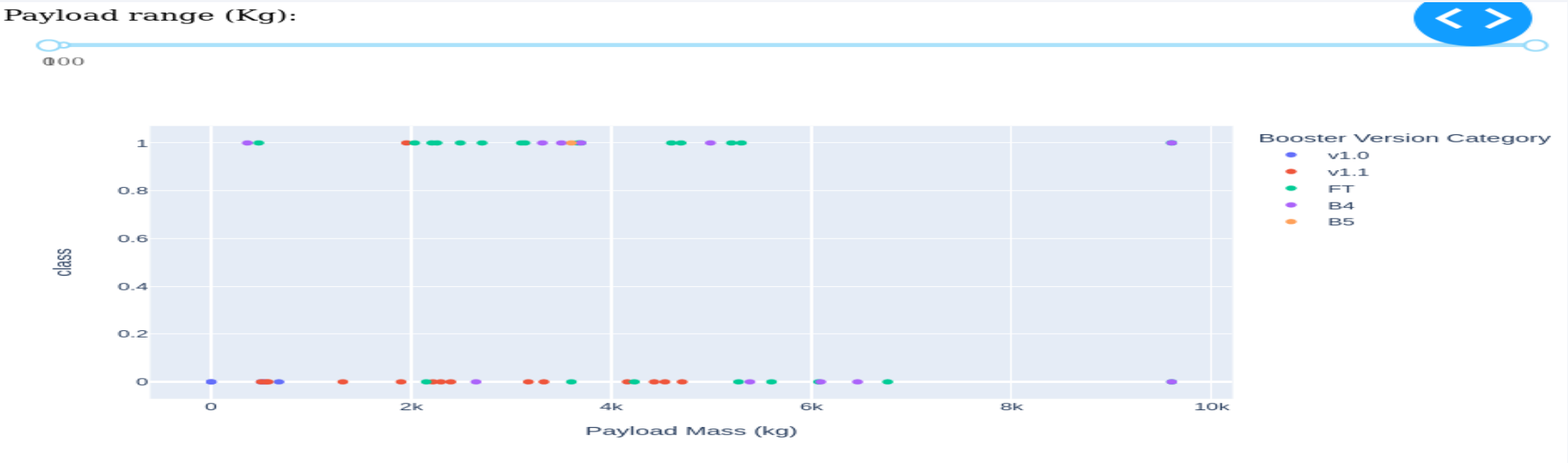
Launch Site with Highest Success Ratio

- KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.
- 76.9% of launches are successful from this site.

KSC LC-39A Success Rate (blue=success)



Payload Vs Launch Outcome



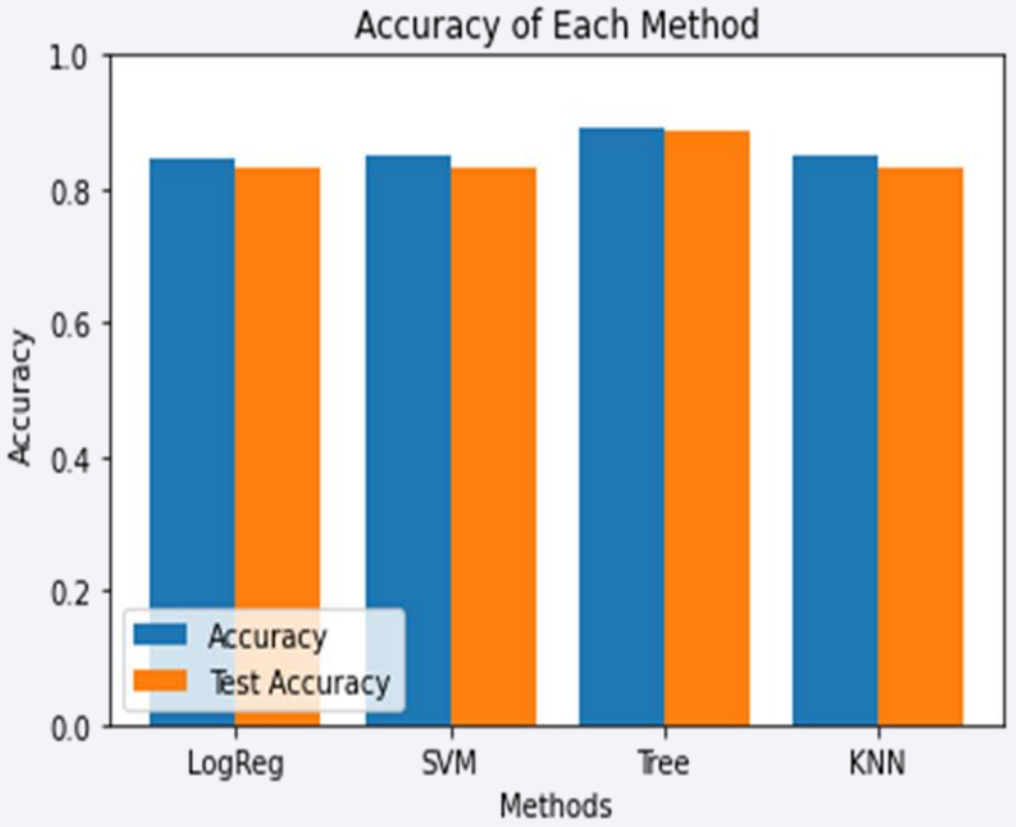
- Class indicates 1 for successful landing and 0 for unsuccessful landing.
- Payloads under 6000 kgs and FT booster version combination has high success rates.
- There is less data to estimate risk of launches over Payload Mass 7000 kgs.

Section 5

Predictive Analysis (Classification)

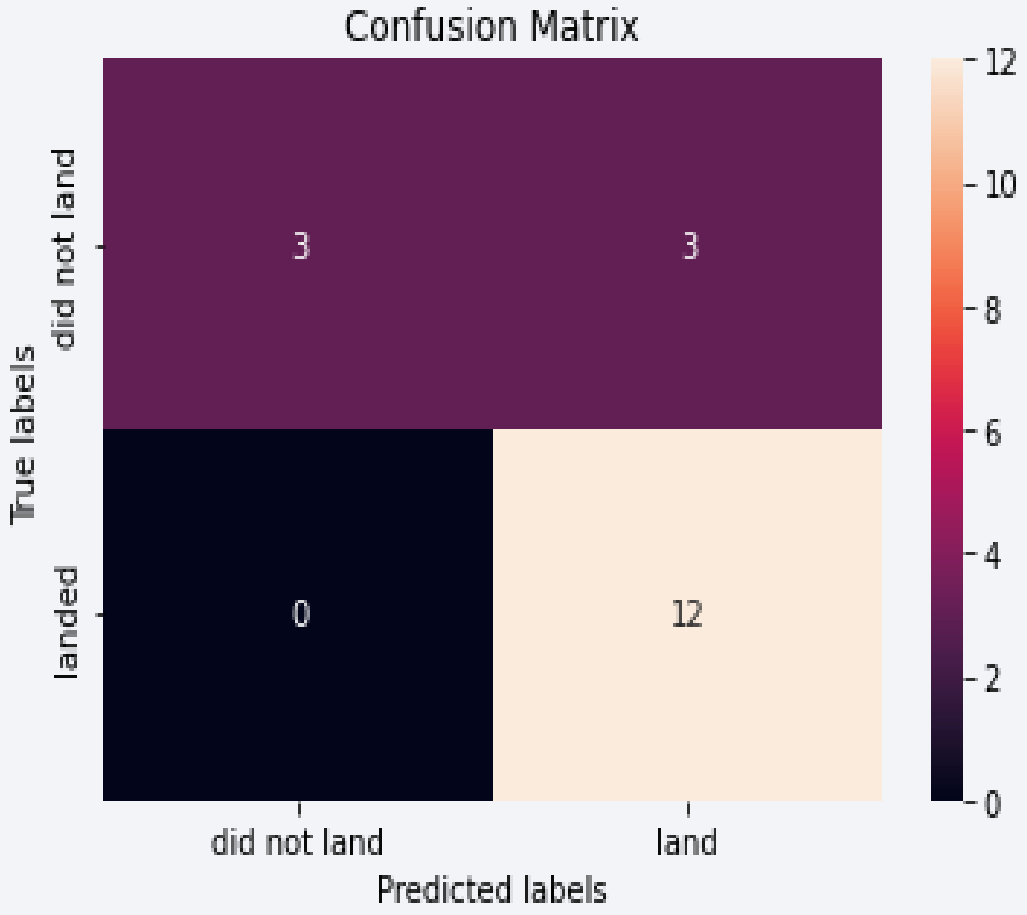
Classification Accuracy

- Four Models were tested, namely, Logistic Regression, SVM, Decision Trees and KNN and their accuracy can be shown beside.
- In our case, the best model which has highest classification accuracy is Decision Tree model with an accuracy of 88%.
- Rest all three models performed equally and have accuracy around 83%
- It should be noted that test size is small and is only contain 18 samples. This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.



Confusion Matrix

- The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 0 unsuccessful landings when the true label was landed.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landings
- The model predicted 3 successful landings when true label was unsuccessful landing.
- Our models over predict successful landings.



Confusion Matrix of Decision Tree Classifier

Conclusions

Problem: To develop a machine learning model for Space Y who wants to compete against SpaceX.

- The goal of model is to predict whether Stage 1 will successfully land or fail to land.
- Different data sources (API and Wiki Page) were analyzed.
- The best launch site is KSC LC 39-A
- Launches above payloads 7000kg are more successful.
- Decision Tree Classifier can be used to predict successful landings and increase profits.
- However, If possible more data should be collected to determine the best machine learning model and improve accuracy.

Appendix

GitHub Repository:

Thank you!