IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniela Reynoso
March-2025

Outline

Executive Summary

- Summary of methodologies
  - Data Collection through API and Web Scrapping
  - Data Wrangling
  - Exploratory Data Analysis (EDA)
    - SQL
    - Data Visualization
    - Interactive Visual Analytics with Folium
  - Machine Learning
- Summary of all results
  - EDA allowed to get insights and decide which features are more relevant to predict success of launchings.
  - ML models were over 80% of confidence in predicting success of launchings.

Introduction

- Project background and context

  Space X advertises Falcon 9 with a cost of 62 million dollars, much shipper compared to other providers. Much of the savings are because the first stage is reusable. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

- Problems you want to find answers

  Which factors influence the most the landing outcome?

  Is there any correlation between variables and how it affects the outcome?

  Which operating conditions are needed to increase the probability of a successful landing?

Section 1

Methodology

Methodology

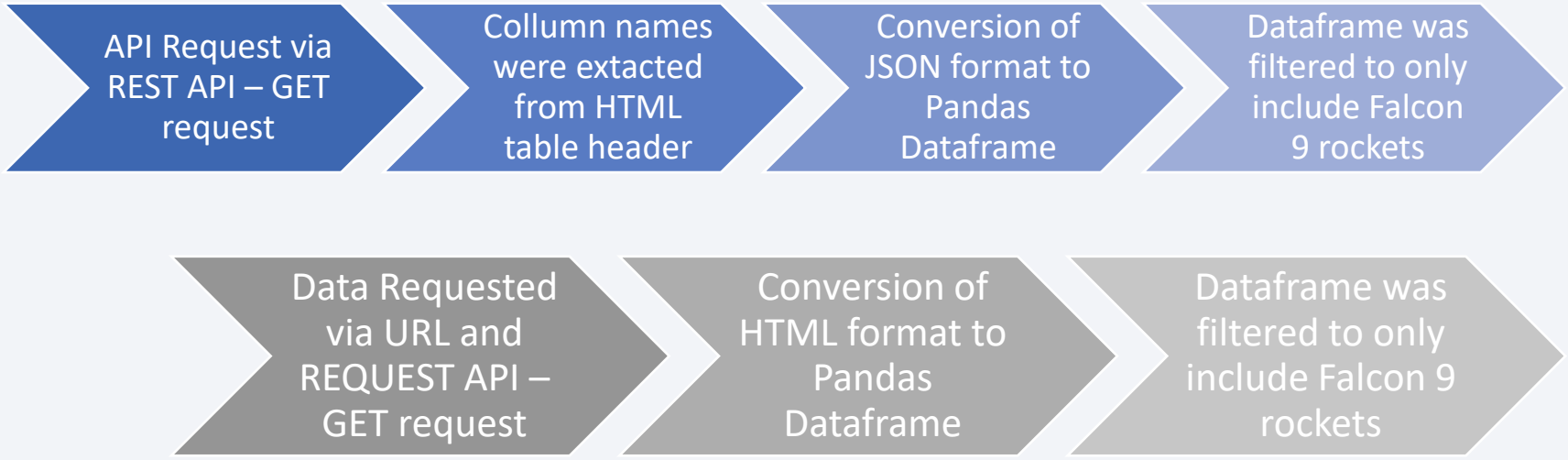
Executive Summary

- Data collection methodology:
  - Data was collected via SpaceX Rest API and through Web Scrapping from Wikipedia.
- Perform data wrangling
  - Data was prepared, using one-hot encoding technique for categorical values and an additional value was added as label: the landing outcome.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Grid search, cross-validation were techniques used to train different models.
  - Every model was evaluated with their metrics: confusion matrix, precision, f1-score, etc.

Data Collection

Data was collected from these links:

- SpaceX API: https://api.spacexdata.com/v4/rockets/
- Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches



Links to notebooks

Data Collection – SpaceX API

GET Request to obtain the Space X API adata

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

Normalizing data

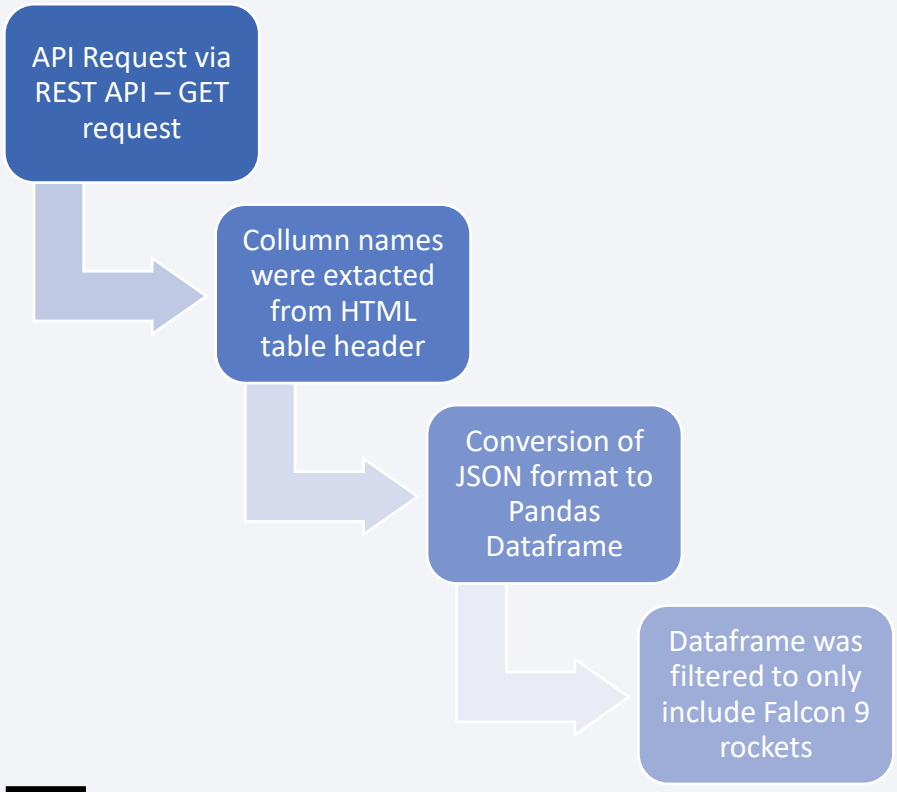
```
# Use json_normalize meethod to convert the json result into a dataframe
response = requests.get(static_json_url).json()
df = pd.json_normalize(response)
```

Assigning columns

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

Filtering dataframe

```
data_falcon9 = df.loc[df['BoosterVersion']!="Falcon 1"]
```



Links to notebooks

Data Collection - Scraping

GET Request to obtain the SpaceX Wikipedia data

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
response = requests.get(static_url)
```

Parsing HTML to text content

```
soup = BeautifulSoup(response.content, 'html.parser')
```

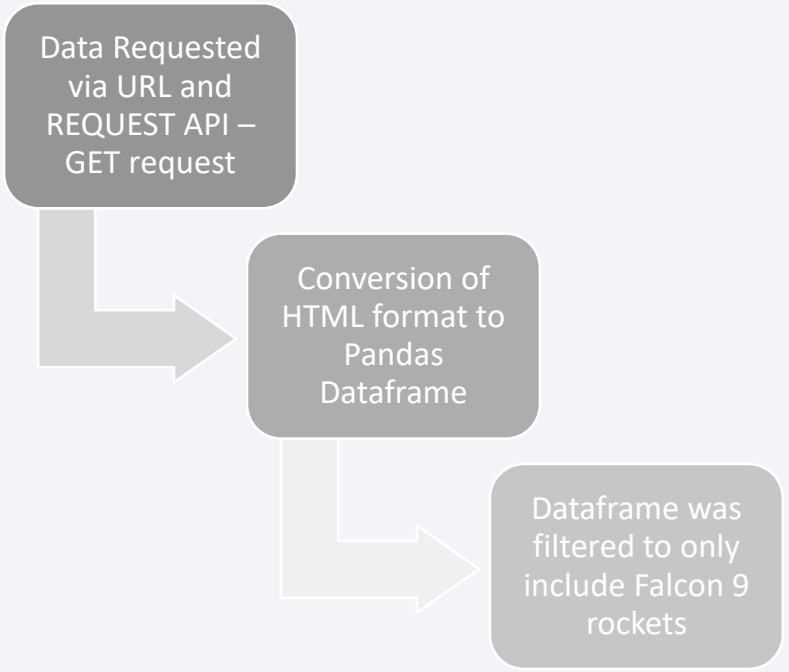
Getting tables and their titles

```
html_tables = soup.find_all('table')
```

```
table = first_launch_table.find_all('th')
for row in table:
    name = extract_column_from_header(row)
    if name is not None and len(name) > 0:
        column_names.append(name)
```

Creating dataframe

```
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```



Links to notebooks

Data Wrangling

- Data was **filtered** using the BoosterVestion column, keeping only Falcon 9 launches.
- **Data types** were checked, so they were appropriate with the column description.
- **Missing data** was deal by replacing with mean value, for example, in column PayloadMass.
- **Feature Engineering** was used, a new column of type boolean was created, this feature was called *class*, it contained the outcome, 0 if was not successful, 1 if it was.


Link to notebook

EDA with Data Visualization

Exploratory Data Analysis (EDA) are used to understand the data main features, find patterns and understand the relation between features. Also, helps to spot outliers, insights and understand data distribution.

The charts plotted were:

Scatter plot

- To visualize the **distribution** and **relation** between *Flight Number – Launch Site, Payload – Launch Site, Flight Number – Orbit Type, Payload – Orbit Type*.

Bar chart

- To visualize the **frequency** of *success rate* on each *orbit type*.

Line plot

- To visualize the **trend** on the *launch success* through the *years*.

Link to notebook

EDA with SQL

SQL queries performed:

Link to notebook

- SELECT DISTINCT Launch_Site FROM SPACEXTBL
- SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
- SELECT SUM(payload_mass__kg_) FROM spacextbl WHERE payload LIKE '%CRS%'
- SELECT AVG(payload_mass__kg_) FROM spacextbl WHERE booster_version LIKE '%F9 v1.1%'
- SELECT MIN(date) FROM spacextbl WHERE mission_outcome LIKE 'Success'
- SELECT booster_version FROM spacextbl WHERE landing_outcome LIKE '%Success (drone ship)%' AND payload_mass__kg_ BETWEEN 4000 AND 6000
- SELECT mission_outcome, COUNT(*) FROM spacextbl GROUP BY mission_outcome
- SELECT DISTINCT booster_version FROM spacextbl WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM spacextbl)
- SELECT date, landing_outcome, booster_version, launch_site FROM spacextbl WHERE landing_outcome LIKE '%Failure%' AND DATE LIKE '%2015%'
- SELECT landing_outcome, COUNT(*) AS counts FROM spacextbl WHERE date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing_outcome ORDER BY counts DESC

Build an Interactive Map with Folium

To create a better visualization with Folium these objects were used:

Link to notebook

Circle

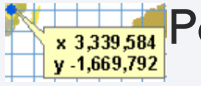
- folium.Circle()
  - A highlighted circle is added, to identify the area of NASA JSC as an initial centre location.

Marker

- folium.map.Marker()
  - To create a marker at a specific launch location.
- MarkerCluster()
  - To create cluster markers of successful and failed launches.

Line

- folium.PolyLine()
  - To mark the distance pf the launch sites to the coast, railways, highways, and major cities.



Position

- MousePosition()
  - To display the latitude and longitude of the mouse cursor's position on a map.

Build a Dashboard with Plotly Dash

- The next plots and graphs were added to the plotly dashboard:
  - Drop-down list: displayed the launch sites.
  - Pie chart: showed the success rates filtered by launch site selected in the drop-down list.
  - Range slider: to select the payload mass.
  - Scatter plot: to show the correlation between the payload mass and the success launches.

Link to notebook

Predictive Analysis (Classification)

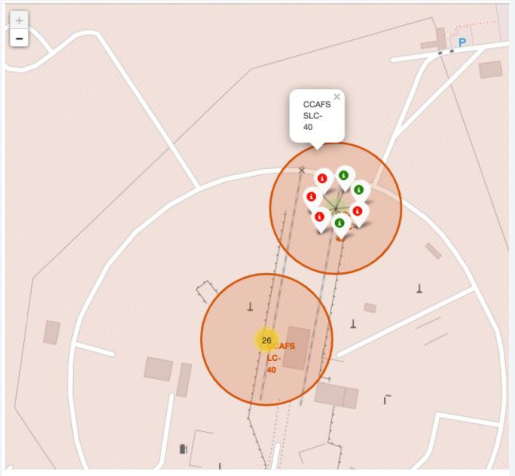
The predictive analysis process consisted in the next steps:

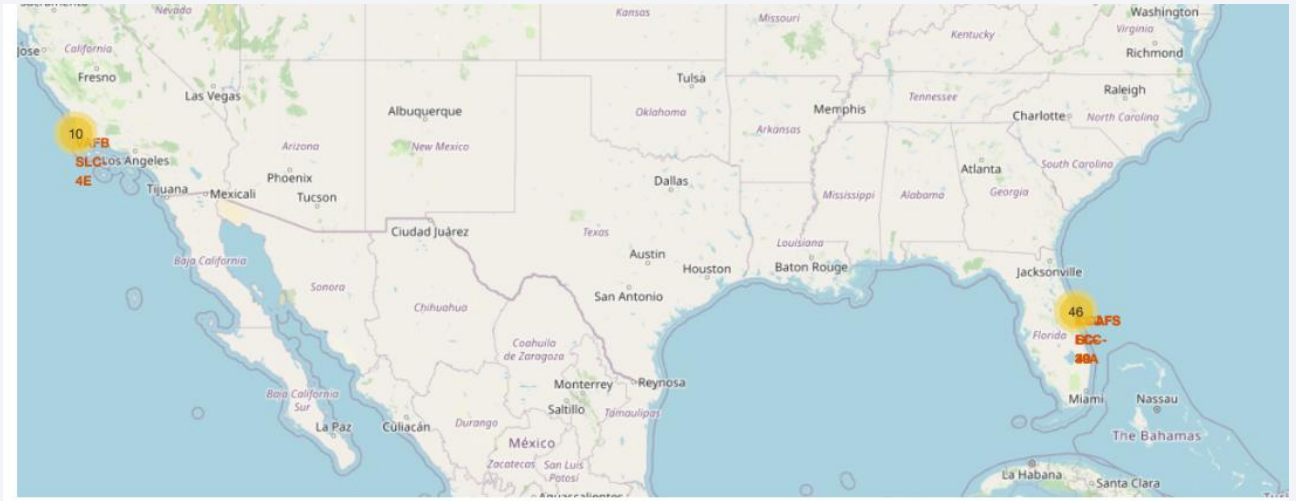
- Import data from Falcon9 launches, from csv to dataframe.
- Create a column “class” to label if the launch was a failure (0) or a success (1).
- Data cleaning (treatment of missing values, change data types) and normalization.
- Split data into training (80%) and testing (20%) sets.
- GridsearchCV was used to find the best hyper-parameters and get the best model configuration.
- Finally, each model (Logistic regression, SVM, DT, K-NN) was evaluated by their metrics and confusion matrix.

Link to notebook

Results

- Exploratory data analysis results
  - The first successful landing outcome was achieved in 2015
  - SpaceX's Falcon 9 achieved a total of 99 success missions.
  - As the flight number increases also the success rate for launch pad location CCAFS SLC 40
- Interactive analytics demo in screenshots





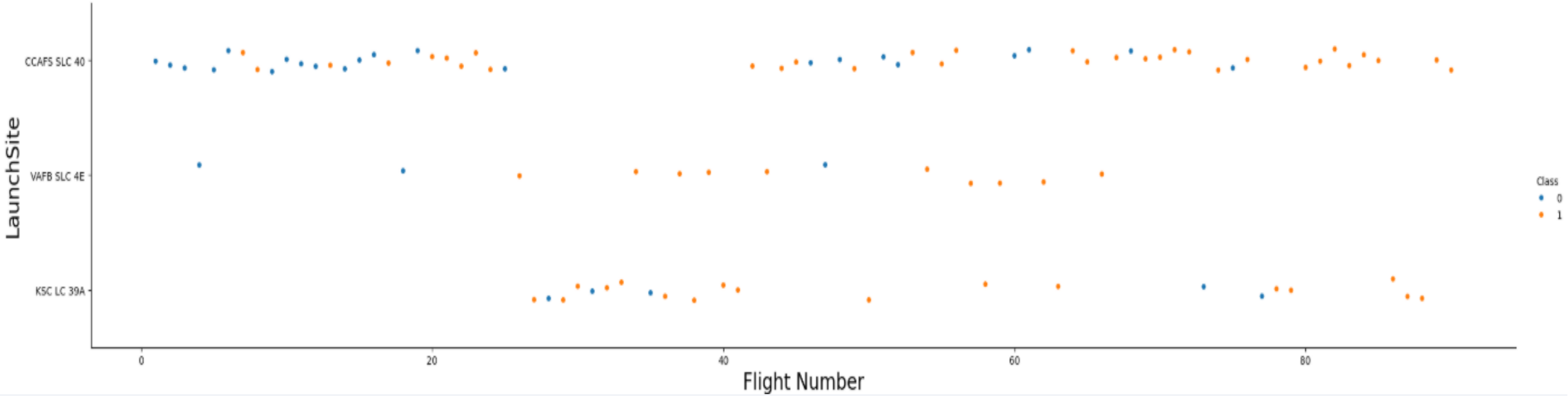
Results

- Predictive analysis results
  - Four different ML algorithms for classification were used.
    - Logistic regression
    - Decision Tree
    - Support Vector Machine
    - K - Nearest Neighbor
  - All the model's accuracy were around 83%.

Section 2

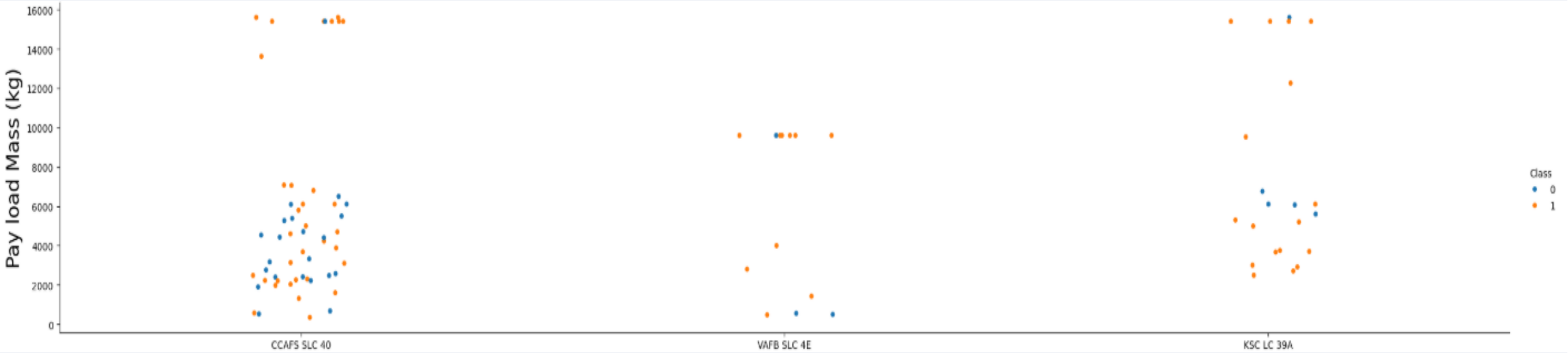
Insights drawn from EDA

Flight Number vs. Launch Site



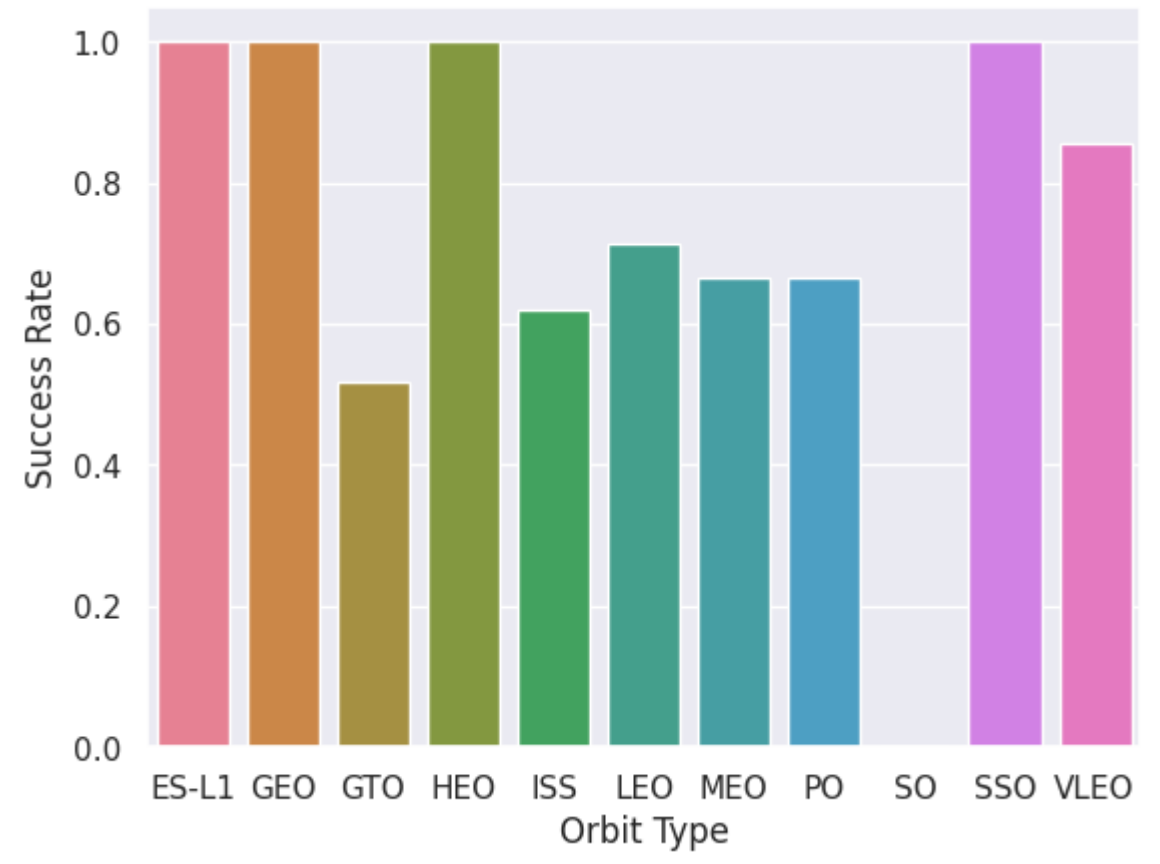
- As the flight number increases, the launchsite CCAFS SLC 40 has a better success rate.
- Launch site VAFB SLC 4E has the highest success ratio.

Payload vs. Launch Site



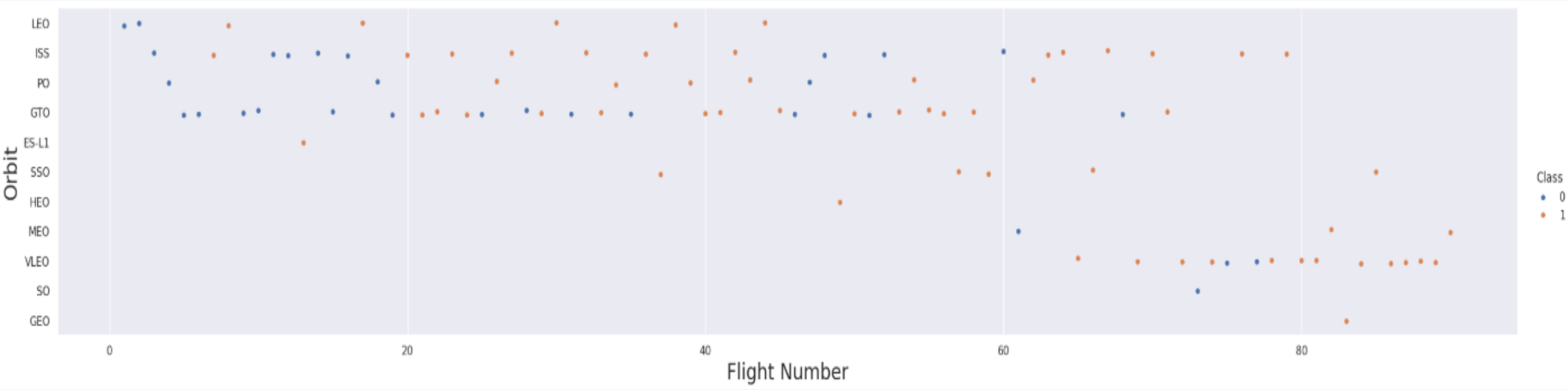
- It seems that the higher the payload mass in CCAFS SLC 40 the higher the success rate.
- VAFB SLC 4E launch site has the lower payload mass registered, below 11,000 kg.

Success Rate vs. Orbit Type



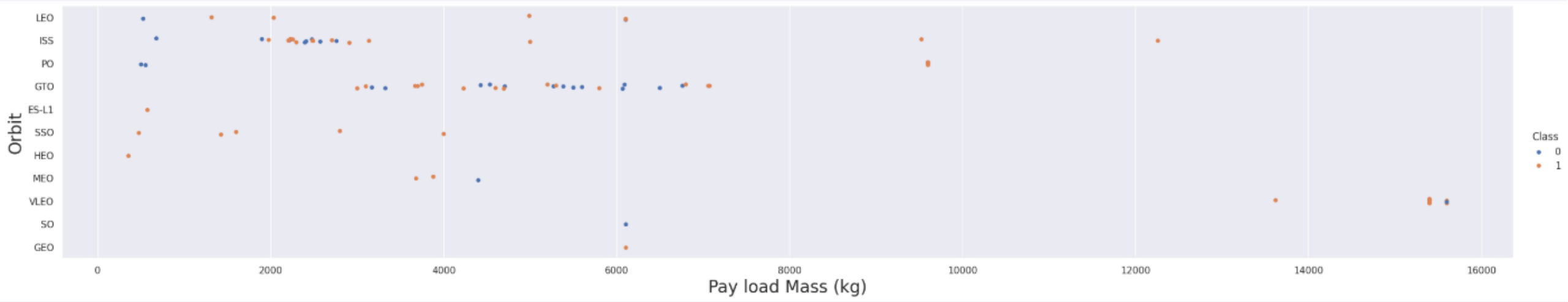
- Orbits ES-L1, GEO, HEO and SSO have a success rate of 1
- Orbit GTO is the lowest with a success rate of 0.5

Flight Number vs. Orbit Type



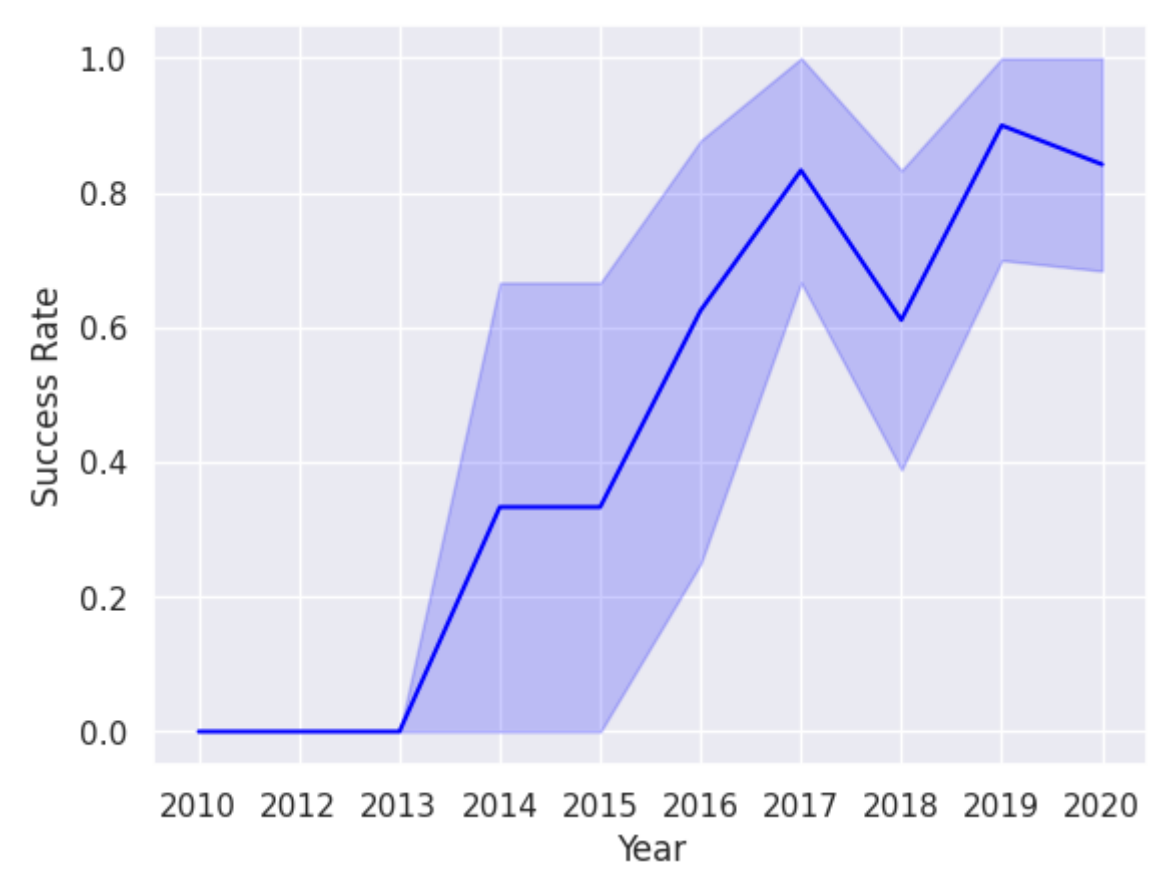
- As the flight number increase also the success rate, specially in VLEO orbit.
- In orbit GTO the flight numbers and success rate seem to be more random.

Payload vs. Orbit Type



- A relationship between the increase of payloads and successful landings can be observed in orbit PO, LEO and ISS.

Launch Success Yearly Trend



- Success rate has been increasing since 2013.
- In 2017 was a little drop down but it recovered in the next year.

All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

- With the word DISTINCT it can be identified the unique Launch site names

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- With the word LIKE it can be identified the Launch sites that contain the string CCA at the beginning.
- With the word LIMIT the number of registers to show cand be controlled, in this case, only 5.

Total Payload Mass

```
%sql SELECT SUM(payload_mass__kg_) FROM spacextbl WHERE payload LIKE '%CRS%'
```

SUM(payload_mass__kg_)
111268

- The total payload carried by boosters from NASA (CRS) is 111,268 kg.
- This query sums all the payload mass of registers that contain the word CRS in the column payload.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(payload_mass__kg_) FROM spacextbl WHERE booster_version LIKE '%F9 v1.1%'
```

AVG(payload_mass__kg_)
2534.6666666666665

- The average payload mass carried by booster version F9 v1 is 2,534.66 kg
- This query average all the payload mass of registers that contain the words F9 v1.1 in the column booster_version.

First Successful Ground Landing Date

```
%sql SELECT MIN(date) FROM spacextbl WHERE mission_outcome LIKE 'Success'
```

MIN(date)
2010-06-04

- The first successful mission outcome was on 2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT booster_version FROM spacextbl WHERE landing_outcome LIKE '%Success (drone ship)%' AND payload_mass__kg_ BETWEEN 4000 and 6000
```

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

- With the word LIKE it can be identified the landing outcomes that contain the string ‘Success (drone ship)’.
- Whit the word BETWEEN we can filtered the registers with payload mass above 4,000 kg and below 6,000 kg

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, COUNT(*) FROM spacextbl GROUP BY mission_outcome
```

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

- Grouping by mission outcome, the result observed is 99 for success and 1 for failure.

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT booster_version FROM spacextbl WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM spacextbl
```

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

- With the word DISTINCT the unique booster versions are displayed, and with the word WHERE it is filtered to the registers with the maximum payload mass.

2015 Launch Records

```
%sql SELECT date, landing_outcome, booster_version, launch_site FROM spacextbl WHERE landing_outcome LIKE '%Failure%' AND DATE LIKE '%2015%'
```

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

- With the word WHERE all the failing landing outcomes are filtered and also with the word LIKE the date is filtered to the year 2015.
- Two registers are the result of the query.

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

```
%sql SELECT landing_outcome, COUNT(*) AS counts FROM spacextbl WHERE date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing_outcome ORDER BY counts DESC
```

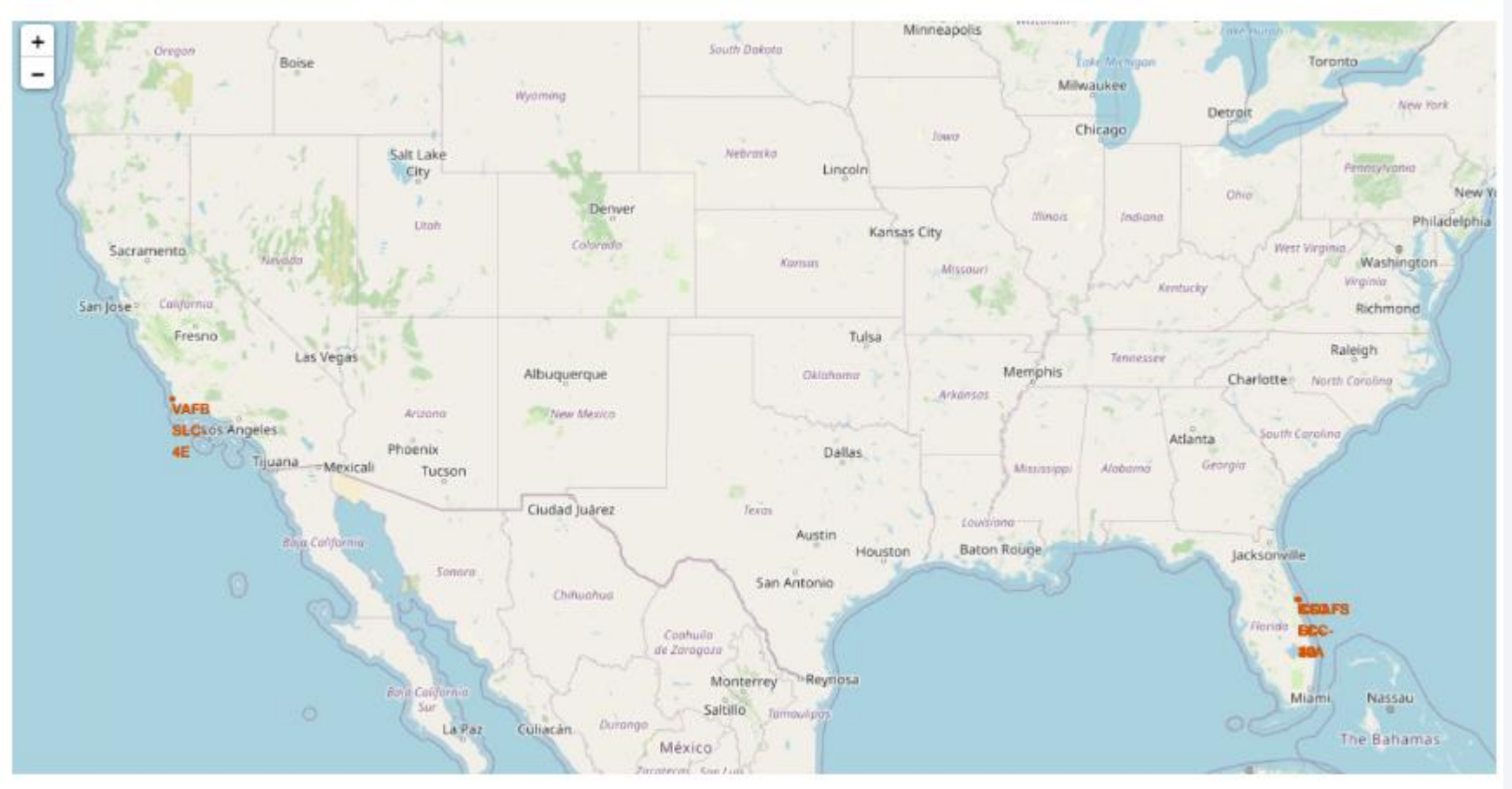
Landing_Outcome	counts
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- This query ranks 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.

Section 3

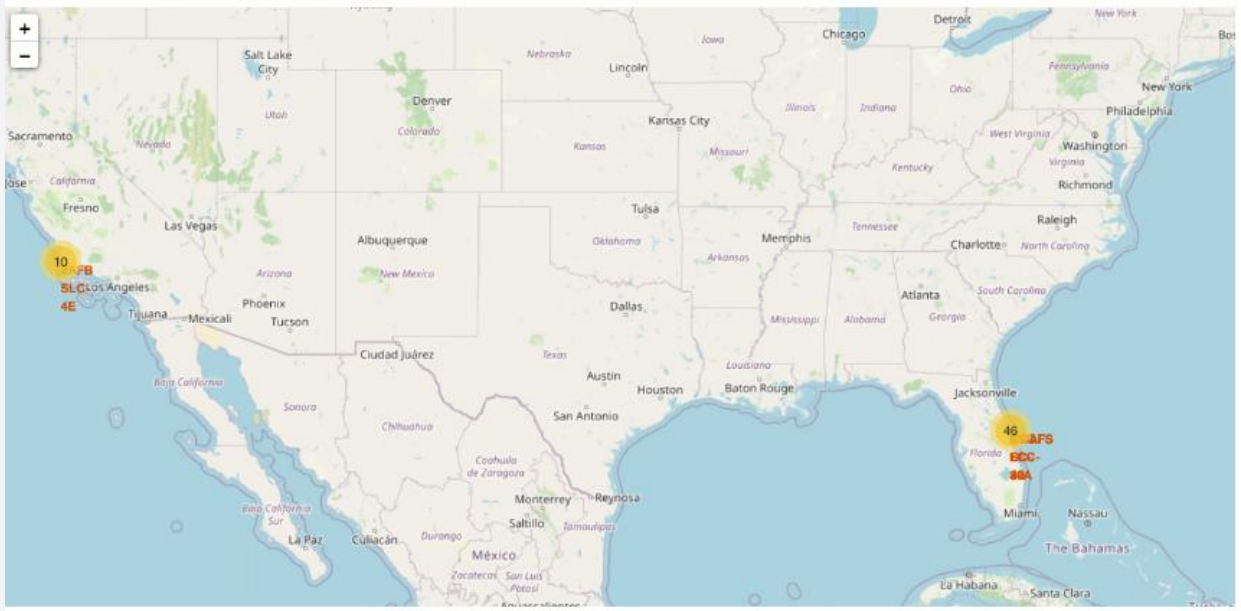
Launch Sites Proximities Analysis

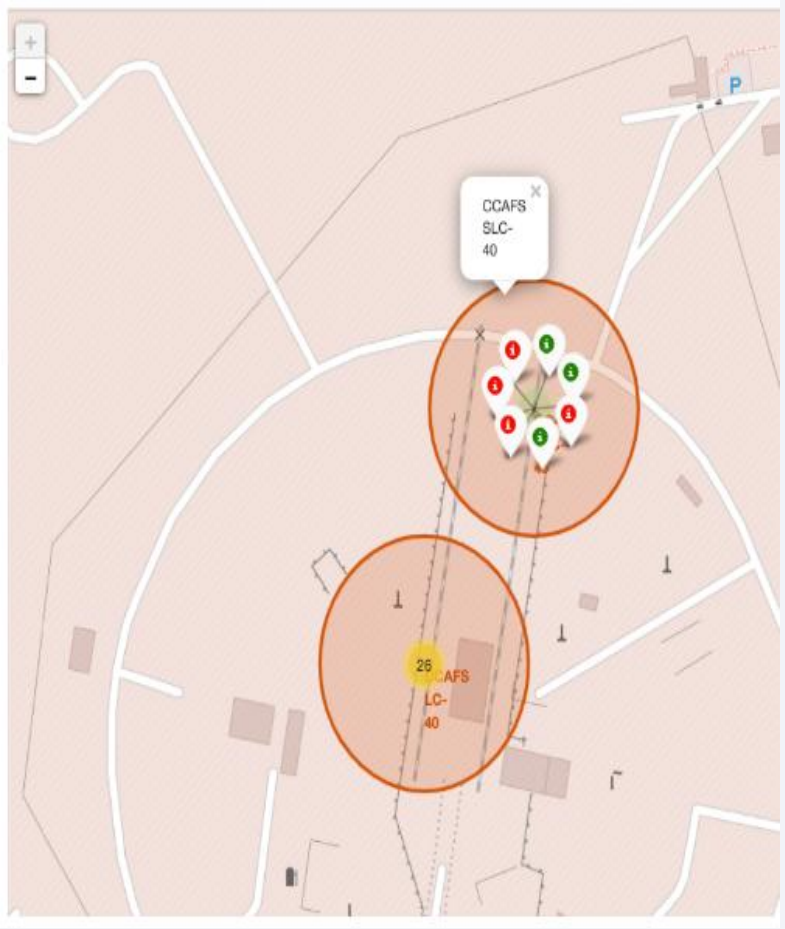
All launch sites on map



- All the launch sites are close to the coast, and in the United States of America, far to the equator line.

Marks showing the success and failed lunches





- Looking Florida launches with more detail, since it has the majority of launches.
- Green markers show successful launches, red shows failures.

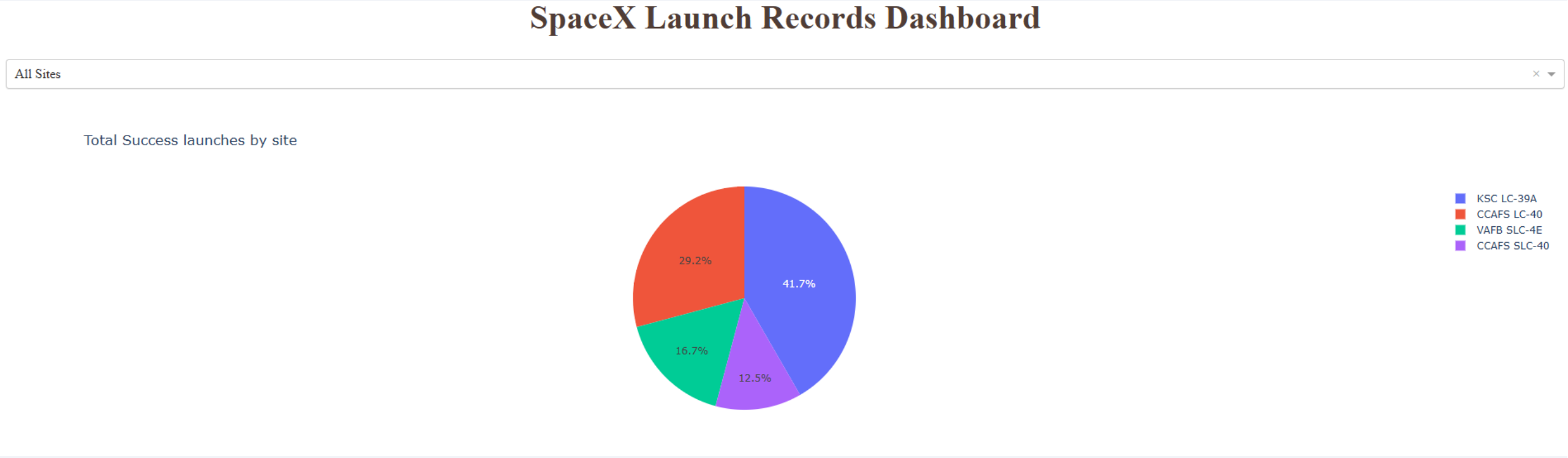
Calculation of distances between a launch site to its proximities

- The launch site proximitie to the coastline is below 1km
- There are no railways or highways close. This means the launch sites keep certain distance away from cities.

Section 4

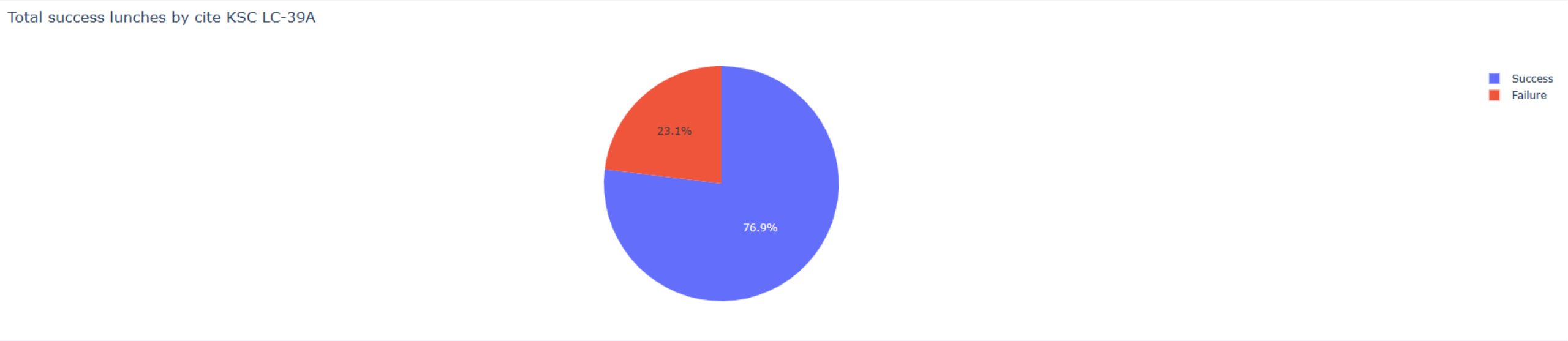
Build a Dashboard with Plotly Dash

Piechart of launch success count for all sites



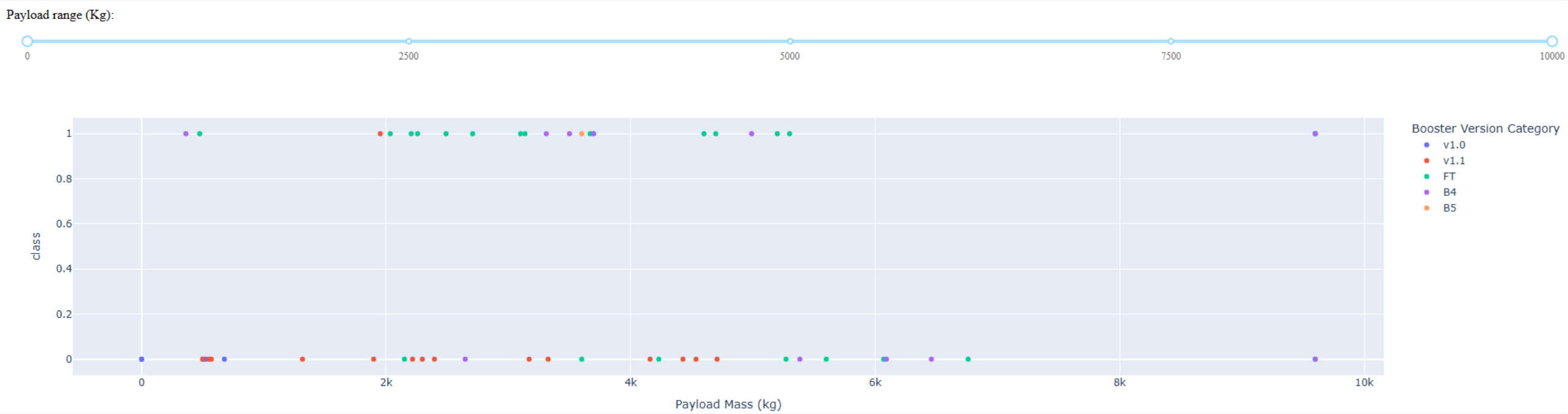
- KSC LC-39A and CCAFS LC-40 are the sites with the highest success ratio.
- CCAFS SLC-40 has the lowest success ratio with 12.5%

Launch site with highest launch success ratio



- KSC LC-39A launch site has the highest success ratio with 76.9% of success launches, and only 23.1% of failures.

Payload vs Launch Outcome

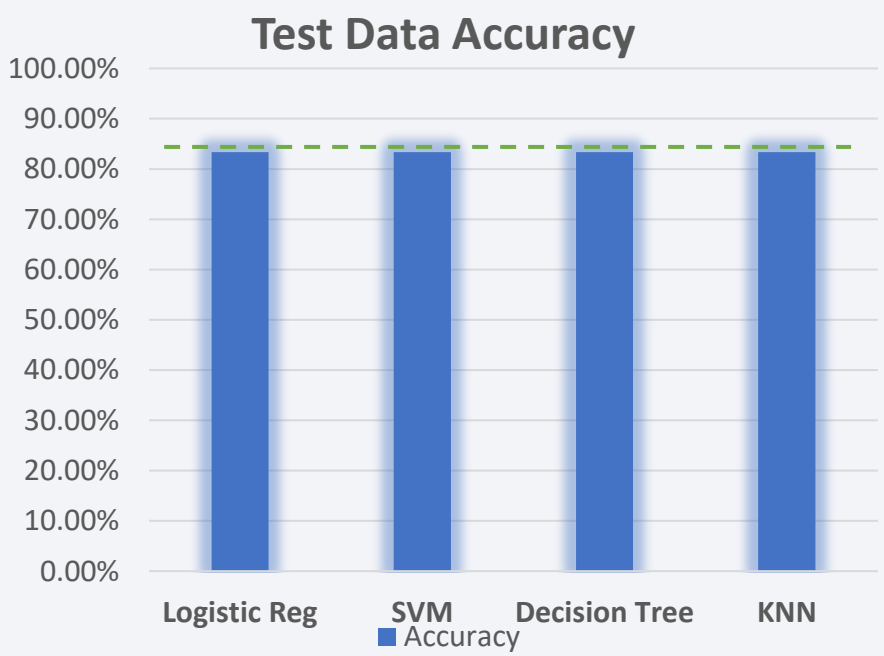


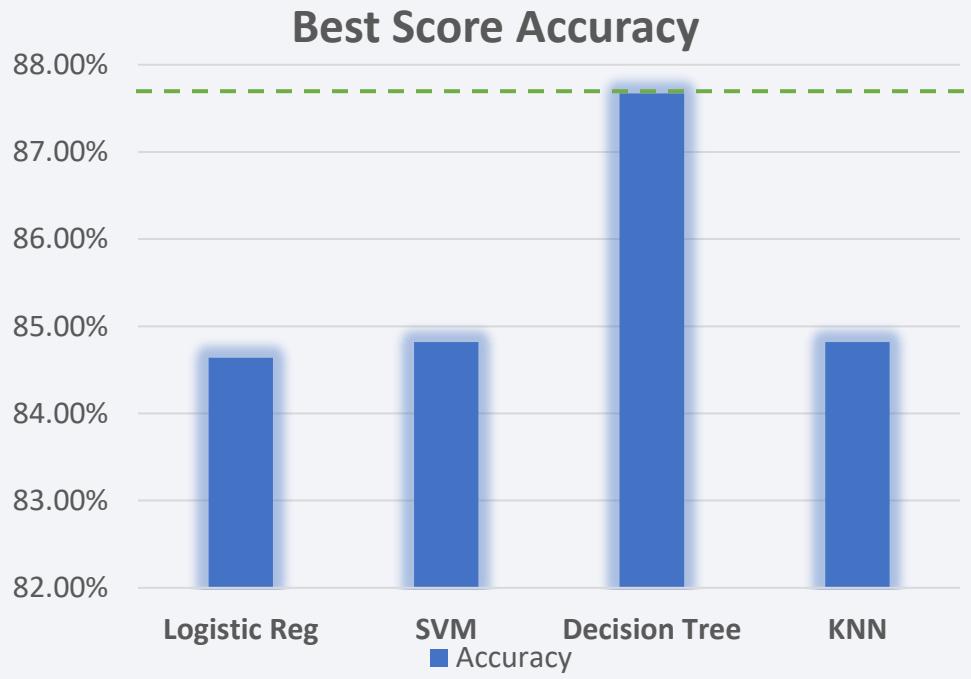
- FT Booster version has the highest success rate with payload above 2k.
- There is only one launch, with booster version B4, that has a success with payload mass above 6k.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



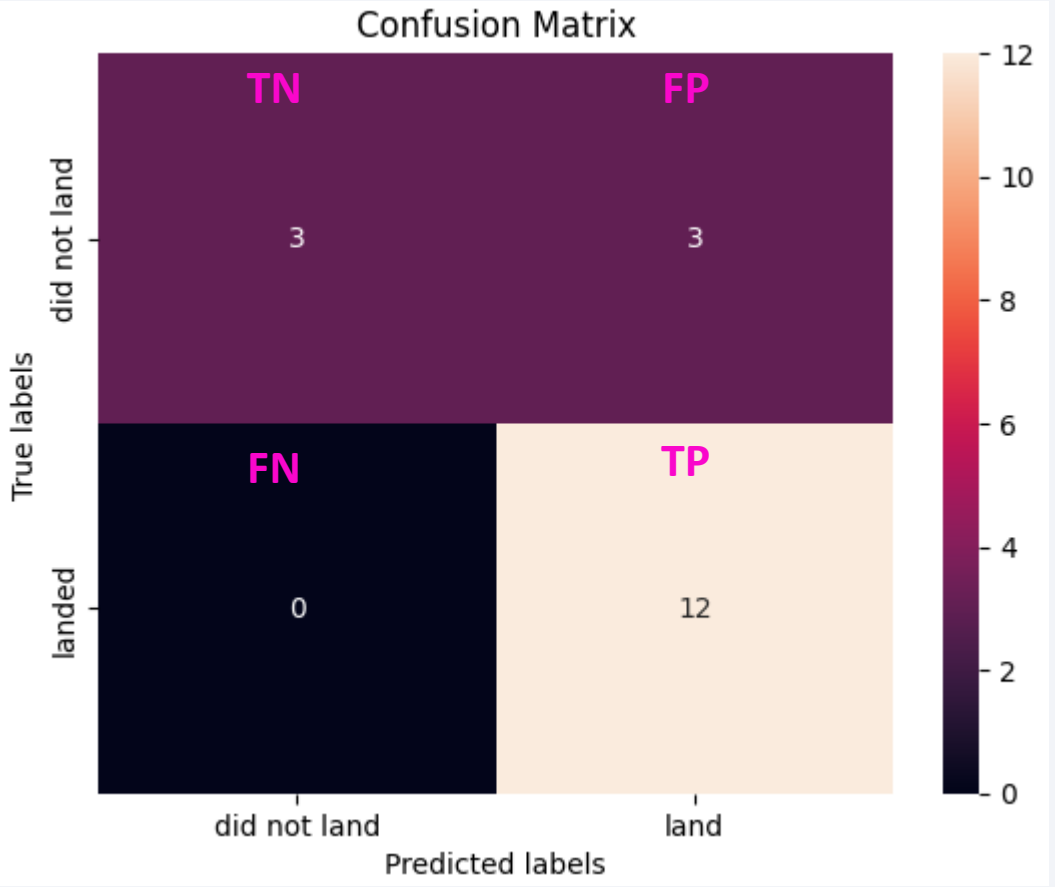


- The four models have a high accuracy, with 83.3%, on test data.
- Decision Tree has the highest best score, with 87.6%

Confusion Matrix

- In the confusion matrix can be observed 3 False Positive (Type I Error) but 0 False Negative (Type II Error).

- Specificity = 0.5
- Precision= 0.8
- Sensitivity= 1
- Accuracy= 0.83



Conclusions

- The outcome of the landing can be predicted with an accuracy above 80%
- A successful outcome can be predicted much better than a failure outcome (higher sensitivity than specificity).
- Orbits, Launch Site and Payload mass are the most important features to predict the outcome of the landing.

Appendix

Relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets can be found at this GitHub.

Thank you!