IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Samuel Castillo
May 2025

Outline

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

Executive Summary

- Summary of methodologies:
  - Data collection through SpaceX API
  - Data collection with web scraping
  - Data wrangling
  - Exploratory data analysis using SQL
  - Exploratory data analysis for data visualization
  - Interactive visual analytics with Folium
  - Interactive dashboard with Ploty Dash
  - Machine Learning prediction
- Summary of all results
  - Exploratory data analysis
  - Interactive visual analytics
  - Predictive analysis

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - What are the key factors in determining whether a launch will be successful?
  - How these factors are correlated and what weight they carry for a successful launch?
  - Which method is the most appropriate for predicting whether a launch will be successful?

Section 1

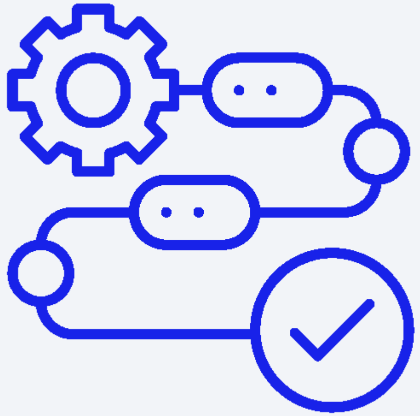
Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data were collected through SpaceX API and web scraping from wikipedia page List of Falcon 9 and Falcon Heavy launches
- Perform data wrangling
    - One hot encoding applied to categorical features for Machine Learning and irrelevant features dropped
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Logistic regression, SVM, decision tree and KNN where the models explored for ML prediction

Data Collection



- Data were collected using SpaceX API using the get request.
- We obtained the response content as a Json and turn it into a Pandas data frame
- Filtered the data frame to only include Falcon 9 launches
- We dealt with missing values
- Then we continued our data collection using web scraping from wikipedia page List of Falcon 9 and Falcon Heavy Launches with BeautifulSoup. We extracted a Falcon 9 HTML table and parsed this table and converted into a Pandas data frame.

Data Collection – SpaceX API

- Get request in SpaceX API to collect data
- Decode the response content as a Json and turn it into a Pandas data frame
- Clean data to obtain relevant data
- GitHub URL of the completed SpaceX API notebook: https://github.com/SamuelCastillo14/SpaceX-Falcon-9-Landing-Prediction/blob/6348ea11fb13018edbef7229490e63ad1f3c793d/jupyter-labs-spacex-data-collection-api_hecho.ipynb

Task 1: Request and parse the SpaceX launch data using the GET request

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

We should see that the request was successfull with the 200 status response code

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

```
response.status_code
```

```
200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
# Use json_normalize meethod to convert the json result into a dataframe
data= pd.json_normalize(response.json())
```

Using the dataframe `data` print the first 5 rows

```
# Get the head of the dataframe
data.head()
```

Data Collection - Scraping

- Complete our data collection with web scraping using Beautiful Soup.
- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse this table and convert it into a Pandas dataframe
- GitHub URL of the completed web scraping notebook: https://github.com/SamuelCastillo14/SpaceX-Falcon-9-Landing-Prediction/blob/6348ea11fb13018edbef7229490e63ad1f3c793d/jupyter-labs-webscraping_hecho.ipynb

TASK 1: Request the Falcon9 Launch Wiki page from its URL

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

```
# use requests.get() method with the provided static_url
# assign the response to a object
response= requests.get(static_url)
```

Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup= BeautifulSoup(response.content, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
# Use soup.title attribute
soup.title
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

Data Wrangling

- Perform some exploratory data analysis of our interest: number of launch on each site, number and occurrence of each orbit, number and occurrence of mission outcome of the orbits, etc
- Apply one hot encoding to categorical features
- Determine training labels
- GitHub URL of completed data wrangling notebook: https://github.com/SamuelCastillo14/SpaceX-Falcon-9-Landing-Prediction/blob/6348ea11fb13018edbef7229490e63ad1f3c793d/labs-jupyter-spacex-Data%20wrangling_hecho.ipynb

```
df.head(5)
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

EDA with Data Visualization

- We have used 3 types of charts in this section: scatter plots, bar chart and line chart
- 3 scatter plots to explore the relationship between Flight number & Launch site, Payload Mass & Launch site, Flight number & Orbit type.
- A bar chart to visualize the relationship between Success rate & Orbit type
- A line chart to check the launch success yearly trend
- GitHub URL of completed EDA with data visualization notebook: https://github.com/SamuelCastillo14/SpaceX-Falcon-9-Landing-Prediction/blob/6348ea11fb13018edbef7229490e63ad1f3c793d/edadataviz_hecho.ipynb

EDA with SQL

- We have performed several SQL queries of interest:
- Names of the unique launch sites in the space mission
- Total payload mass carried by boosters launched by NASA (CRS)
- Total number of successful and failure mission outcomes
- All the booster versions that have carried the maximum payload mass
- Check the link below to discover all the SQL queries
- GitHub URL of completed EDA with SQL notebook: https://github.com/SamuelCastillo14/SpaceX-Falcon-9-Landing-Prediction/blob/6348ea11fb13018edbef7229490e63ad1f3c793d/jupyter-labs-eda-sql-coursera_sqllite_hecho.ipynb

Build an Interactive Map with Folium

- The map objects we have created and added to a folium map are such as:
    - Markers
    - Circles
    - Lines
    - Marker clusters
    - Labels

- We have added these objects to identify launch sites, mark success/failed launches for each site on the map and calculate distances between launch sites and locations in the proximities
- GitHub URL of completed interactive map with Folium map: https://github.com/SamuelCastillo14/SpaceX-Falcon-9-Landing-Prediction/blob/6348ea11fb13018edbef7229490e63ad1f3c793d/lab_jupyter_launch_site_location_hecho2.ipynb

Build a Dashboard with Plotly Dash

- We have added next plots/graphs and interactions to a dashboard:
  - Launch Site Drop-down Input Component
  - A callback function to render success-pie-chart based on selected site dropdown
  - A Range Slider to Select Payload
  - A callback function to render the success-payload-scatter-chart scatter plot
- GitHub URL of completed Plotly Dash lab: https://github.com/SamuelCastillo14/SpaceX-Falcon-9-Landing-Prediction/blob/6348ea11fb13018edbef7229490e63ad1f3c793d/spacex-dash-app.py

Predictive Analysis (Classification)

- To built, evaluate, improve, and found the best performing classification model, we have studied 4 models: Logistic regression, SVM, decission tree and KNN.
- In each model, we have split data into training and test data, create a GridSearchCV object with a value of 10 for cross validation and using parameters to find the best parameters
- Then we have calculated the accuracy of each model using the method score and plotted its corresponding confusion matrix
- Finally we have determinated that all the models practically give the same results
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose: https://github.com/SamuelCastillo14/SpaceX-Falcon-9-Landing-Prediction/blob/6348ea11fb13018edbef7229490e63ad1f3c793d/SpaceX_Machine%20Learning%20Prediction_Part_5_hecho.ipynb

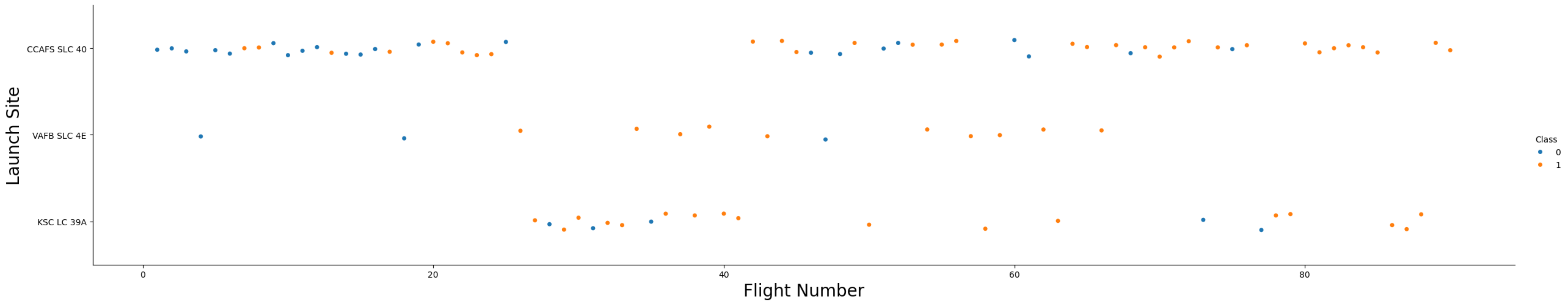
Results

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

Section 2

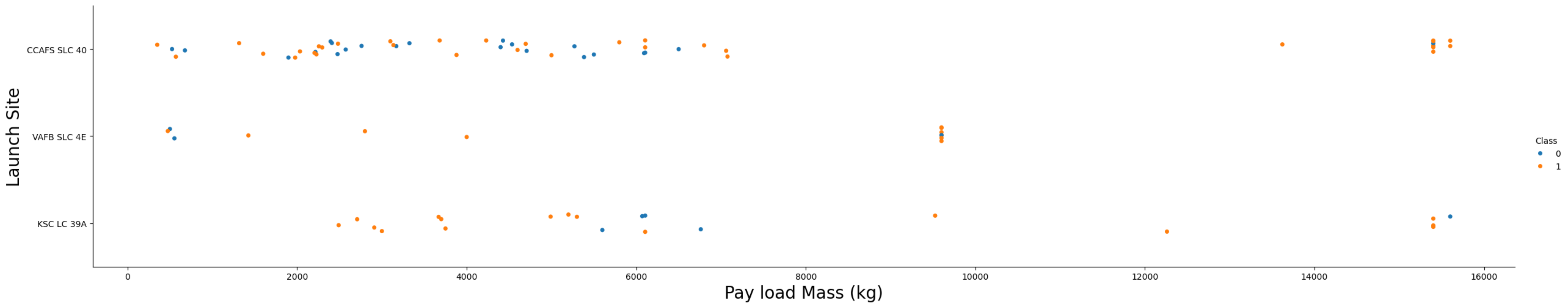
Insights drawn from EDA

Flight Number vs. Launch Site



- The higher the number of flights, the higher the rate of successful launches from the launch site.

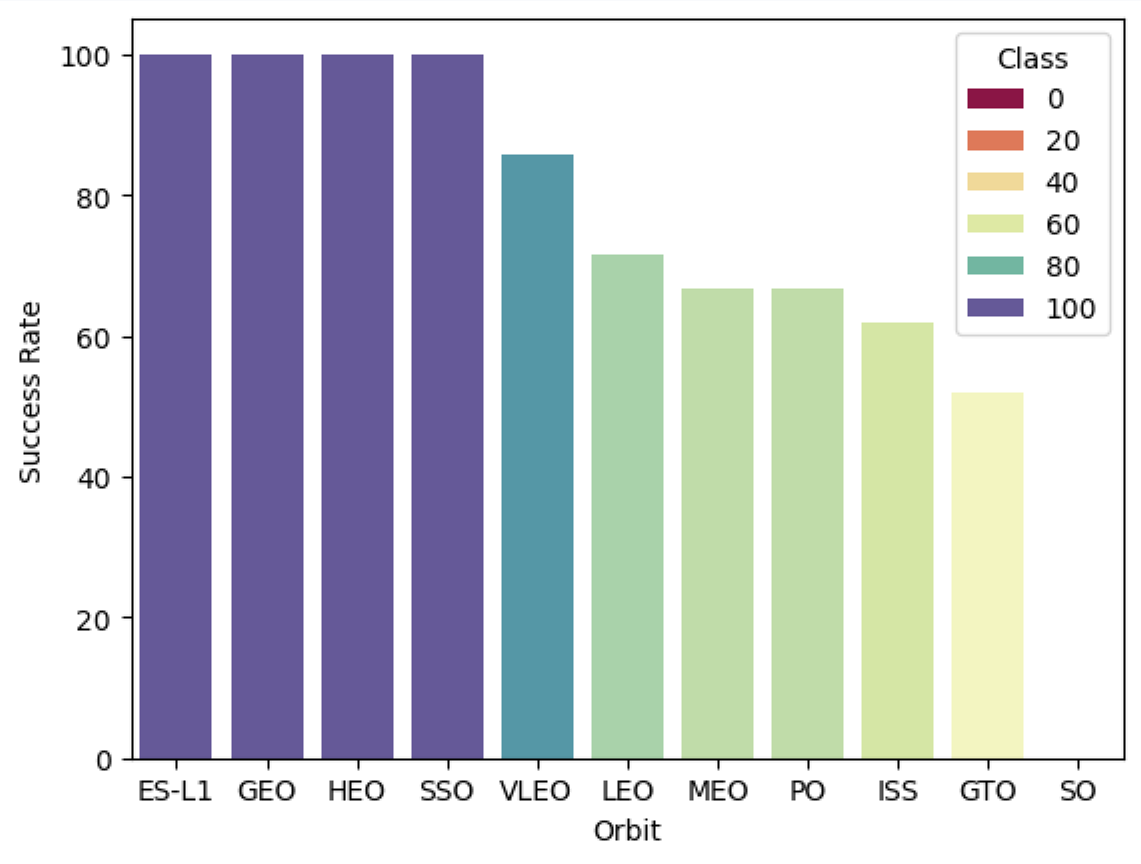
Payload vs. Launch Site



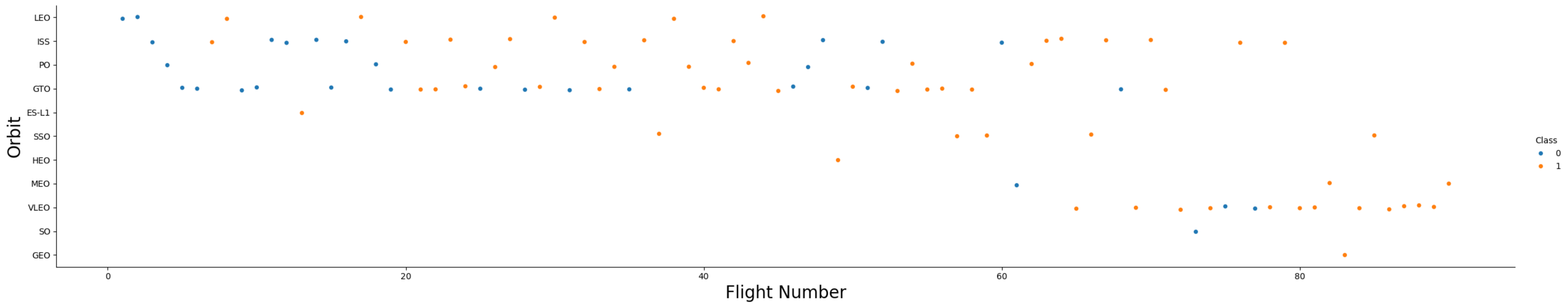
- For CCAFS SLC 40, the higher the payload, the higher the rate of successful launches.
- For the rest of the launch sites this pattern is not as noticeable, so it is not clear that there is as much influence when considering the load as a decisive factor in determining whether a launch will be successful or not.
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000)

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have the maximum success rate with 100%
- GTO has approximately 50% of success rate
- The lowest success rate belongs to SO orbit with 0%

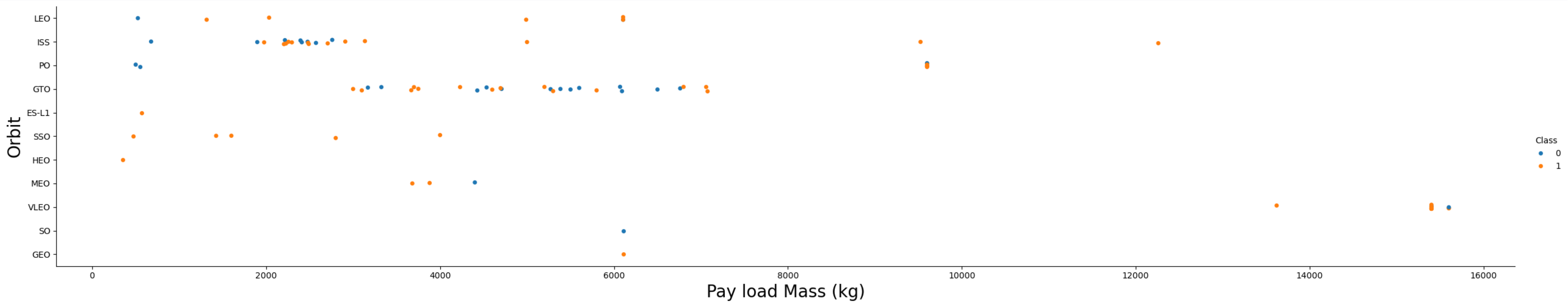


Flight Number vs. Orbit Type



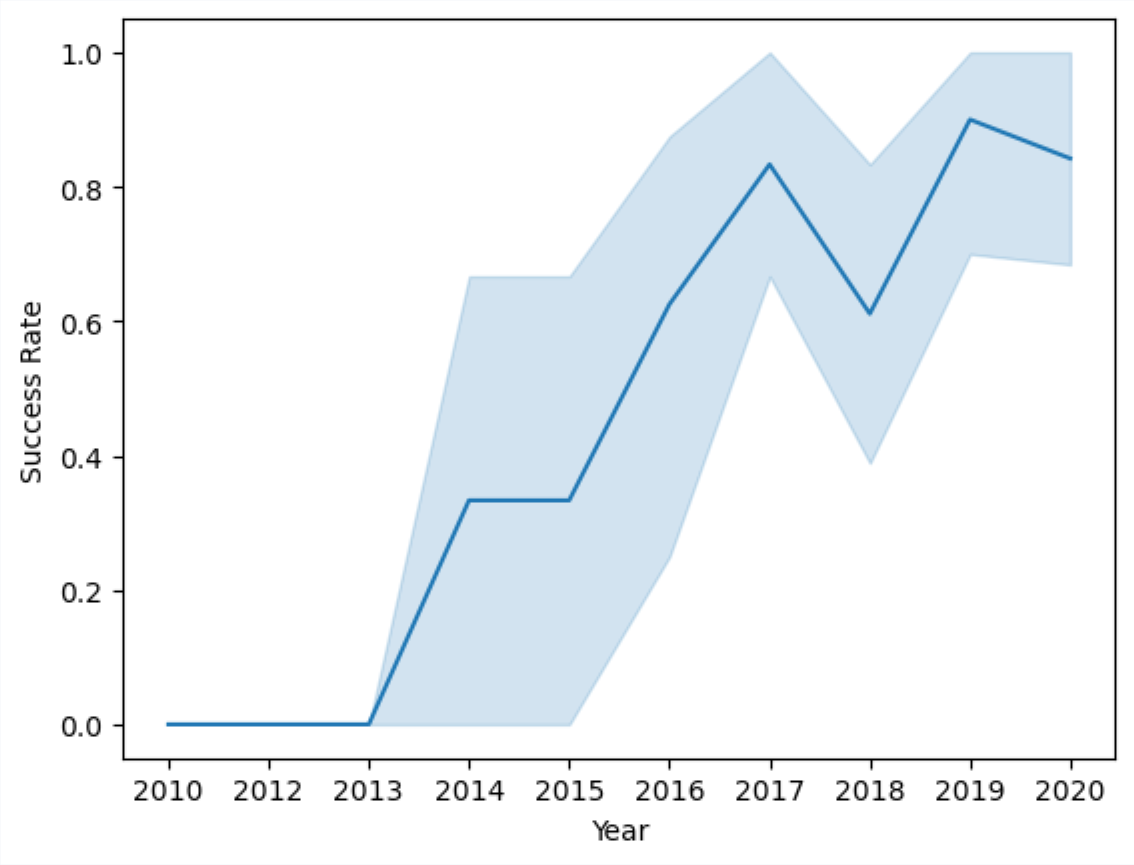
- In the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



- There is a clear upward trend from 2013 to 2020.

All Launch Site Names

- Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT("Launch_Site") FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

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

- We use SELECT DISTINCT because we want to be returned only different values.

Launch Site Names Begin with 'CCA'

- Display 5 records where launch sites begin with the string 'CCA'

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

```
 * sqlite:///my_data1.db
,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	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

- We use LIKE operator for the string 'CCA%' to filter the results, also we use % to match any character after CCA and we finish with LIMIT 5 to show only 5 records

Total Payload Mass

- Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer= 'NASA (CRS)'

 * sqlite:///my_data1.db
,Done.
,,,,,,,,,,,
```

SUM(PAYLOAD_MASS__KG_)
45596

- We use SUM() function to return the total sum and we filter with WHERE in customer column to return required results

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
,Done.
,,,,,,,,,,,
```

AVG(PAYLOAD_MASS__KG_)
2928.4

- We use AVG() function to return the average value of the PAYLOAD_MASS__KG_ column. Then we filter required results using WHERE and LIKE

First Successful Ground Landing Date

- List the date when the first successful landing outcome in ground pad was achieved

```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'

 * sqlite:///my_data1.db
,Done.
```

MIN(Date)
2015-12-22

- We use MIN() function to retrieve the smallest (first) value of the date column, then we filter the required result using WHERE in Landing_Outcome column

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT DISTINCT(Booster_Version) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ >4000 AND PAYLOAD_MASS__KG_ <6000
```

```
 * sqlite:///my_data1.db
,Done.
,,,,,,,,,,,,,,,,,,,,
```

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

- We use SELECT DISTINCT to retrieve only different values, then we filter using WHERE in Landind_Outcome column and we also use AND operators to filter out all other conditions

Total Number of Successful and Failure Mission Outcomes

- List the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) AS Total FROM SPACEXTBL GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
,Done.
,,,,,,,,,,,,,,,,,,,,,,,,
```

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

- We use COUNT() function and GROUP BY to count the number of rows that belong to each group. We also use the keyword AS to create the alias Total to improve the readability.
- As a curiosity, in the results we can observe a possible failure derived from data cleaning

Boosters Carried Maximum Payload

- List all the booster_versions that have carried the maximum payload mass. Use a subquery.

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * sqlite:///my_data1.db
,Done.
```

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

- In this case we use a subquery that is acting like a filter to retrieve the required results.
- For the subquery, we use MAX() function to return the largest value

2015 Launch Records

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

```
%sql SELECT SUBSTR(Date,6,2) AS Month, Booster_Version, Launch_site FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date,0,5) = '2015'
```

```
 * sqlite:///my_data1.db
,Done.
```

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- SQLite does not support monthnames. So we need to use SUBSTR(Date, 6,2) as month to get the months and SUBSTR(Date,0,5)='2015' for year
- We also use WHERE to filter data by failed landing_outcomes in drone ship

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending

```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Total
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
AND Landing_Outcome = 'Failure (drone ship)' OR Landing_Outcome = 'Success (ground pad)'
GROUP BY Landing_Outcome ORDER BY Total DESC
```

```
 * sqlite:///my_data1.db
,Done.
,,,,,,,,,,,,,,,,
```

Landing_Outcome	Total
Success (ground pad)	9
Failure (drone ship)	5

- For greater convenience, we use %%sql to write our query in several lines

- We use COUNT() function and GROUP BY to count the number of rows that belong to each group. We also use the alias Total, WHERE to filter date in the indicate period using AND operator. We also filter by Landing_Outcome and use OR operator for the conditions. Finally ORDER BY alias Total and DESC to sort the data in descending order

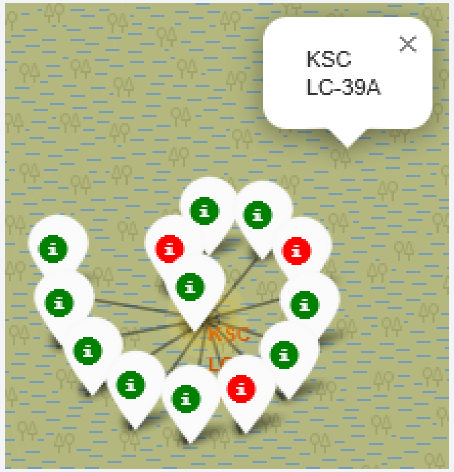
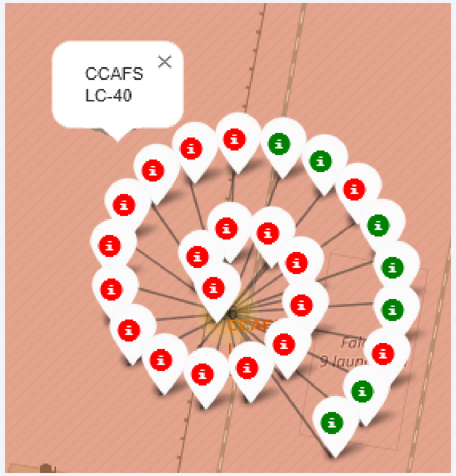
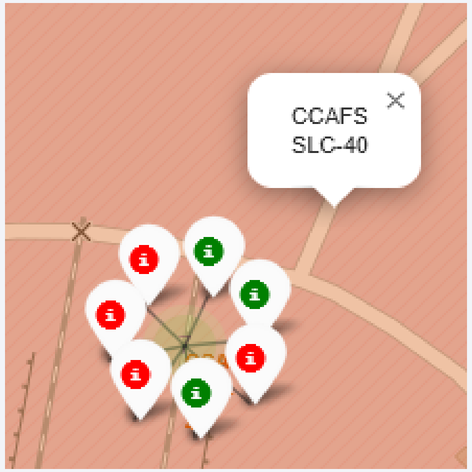
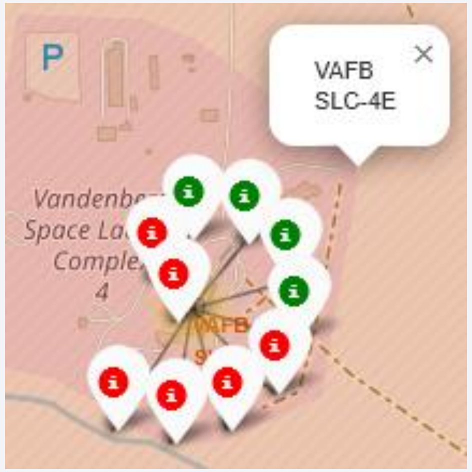
Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites Global Map

- All launch sites are in proximity to Equator line
- All launch sites are in very close proximity to the coast
- 2 main locations: California (Pacific coast) and Florida (Atlantic coast)

Success/Failed Lauches

Florida	Florida	Florida	California
			

- We can observe that there is a higher number of launches from Florida.
- In terms of the rate of successful launches (marked in green), KSC LC 39A (Florida) has the highest with almost 77% success rate.
- CCAFS LC-40 (Florida) has the worst success rate with almost 27%.

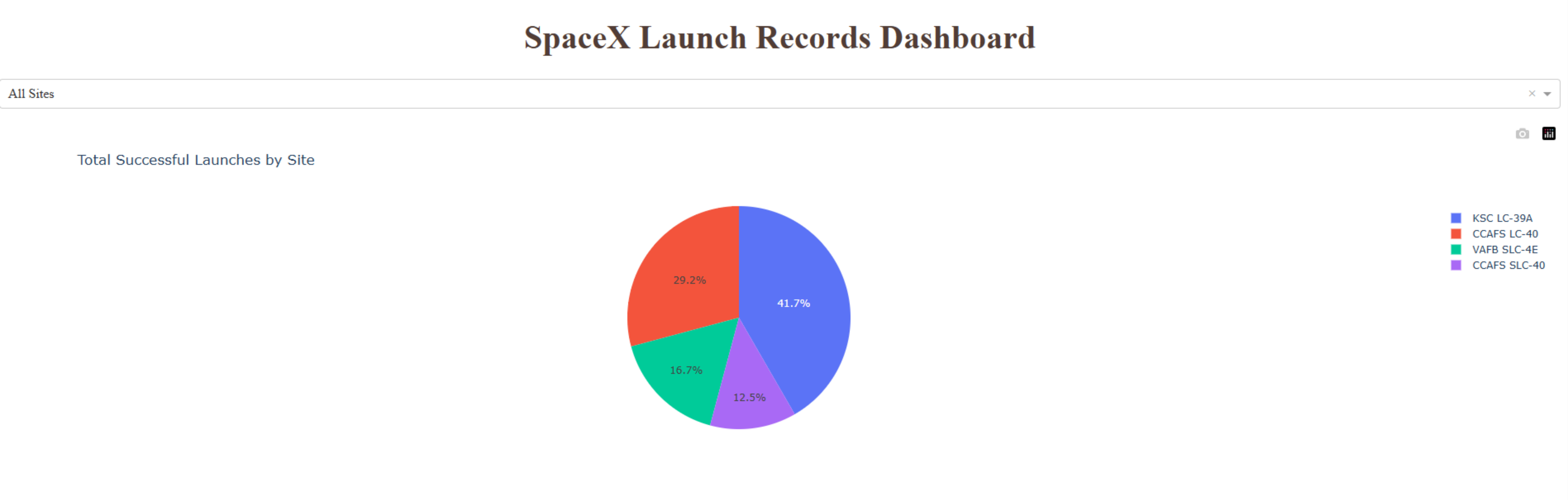
Distances between Launch Sites to its Proximities

- For obvious safety reasons, there is a safety distance between launch sites and certain locations in their proximities, such as cities, highways, railways or the coastline.
- To give a couple of examples, the city of Titusville is almost 16 kilometres from the nearest launch site. On the other hand, the distance between the nearest coastline and the CCAFS SLC-40 launch site is almost 1 kilometre.

Section 4

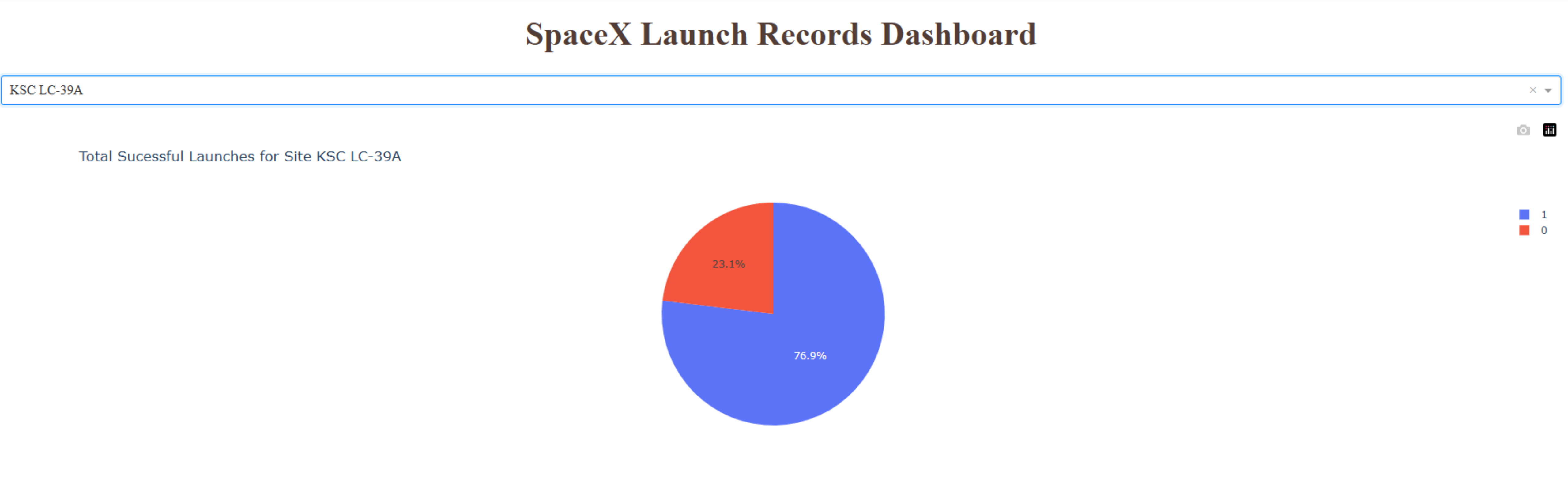
Build a Dashboard with Plotly Dash

Launch Succes Count for All Sites



- The launch site with the highest success count is KSC LC 39A with just over 41%, followed at some distance by CCAFS LC-40 with just over 29%, trailing VAFB SLC-4E with almost 17% and CCAFS SLC-40 with 12.5%.

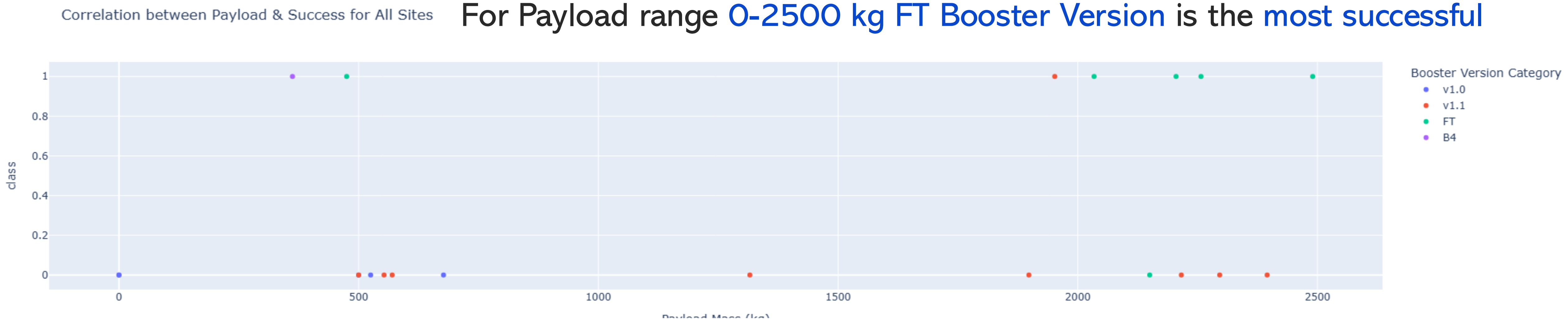
Launch Site with Highest Success Ratio



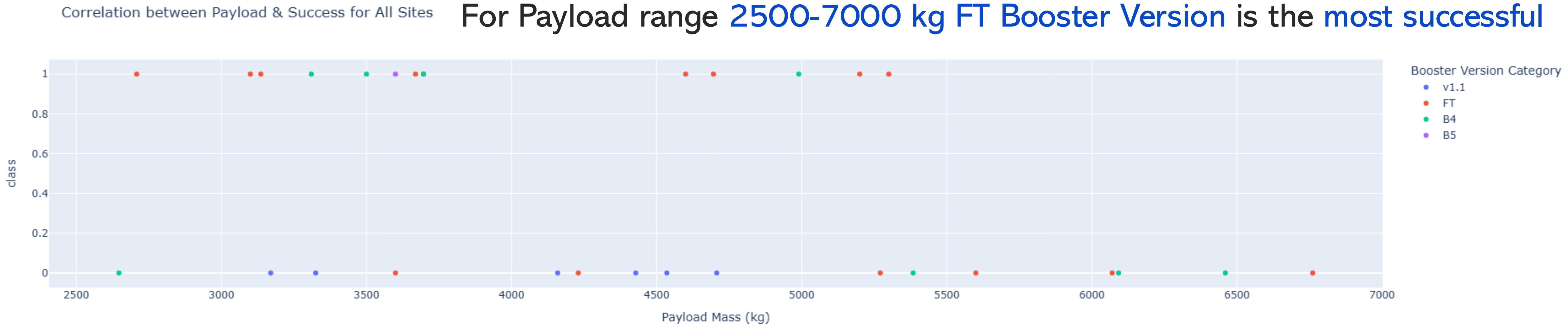
- The launch site with the highest success ratio is KSC LC 39A with almost 77%.

Payload vs Launch Outcome

For Payload range 0-2500 kg FT Booster Version is the most successful



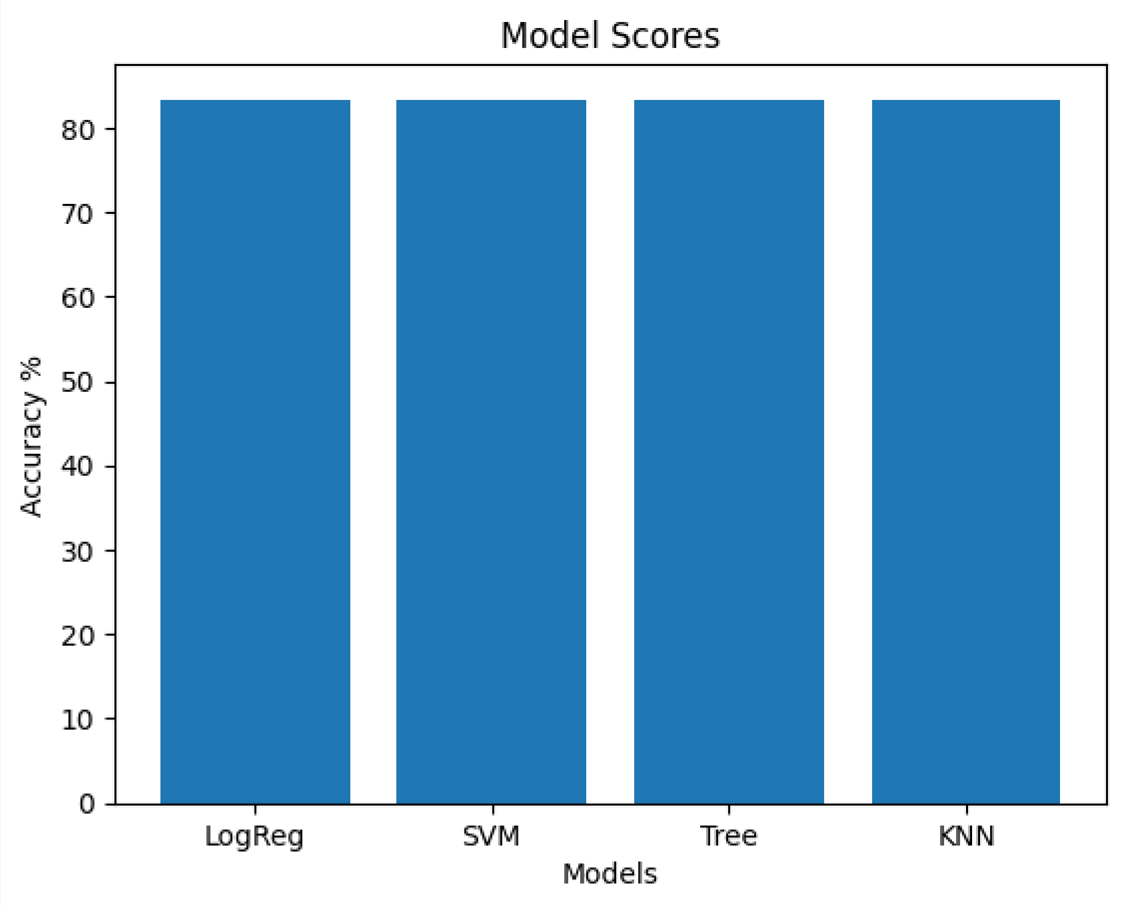
For Payload range 2500-7000 kg FT Booster Version is the most successful

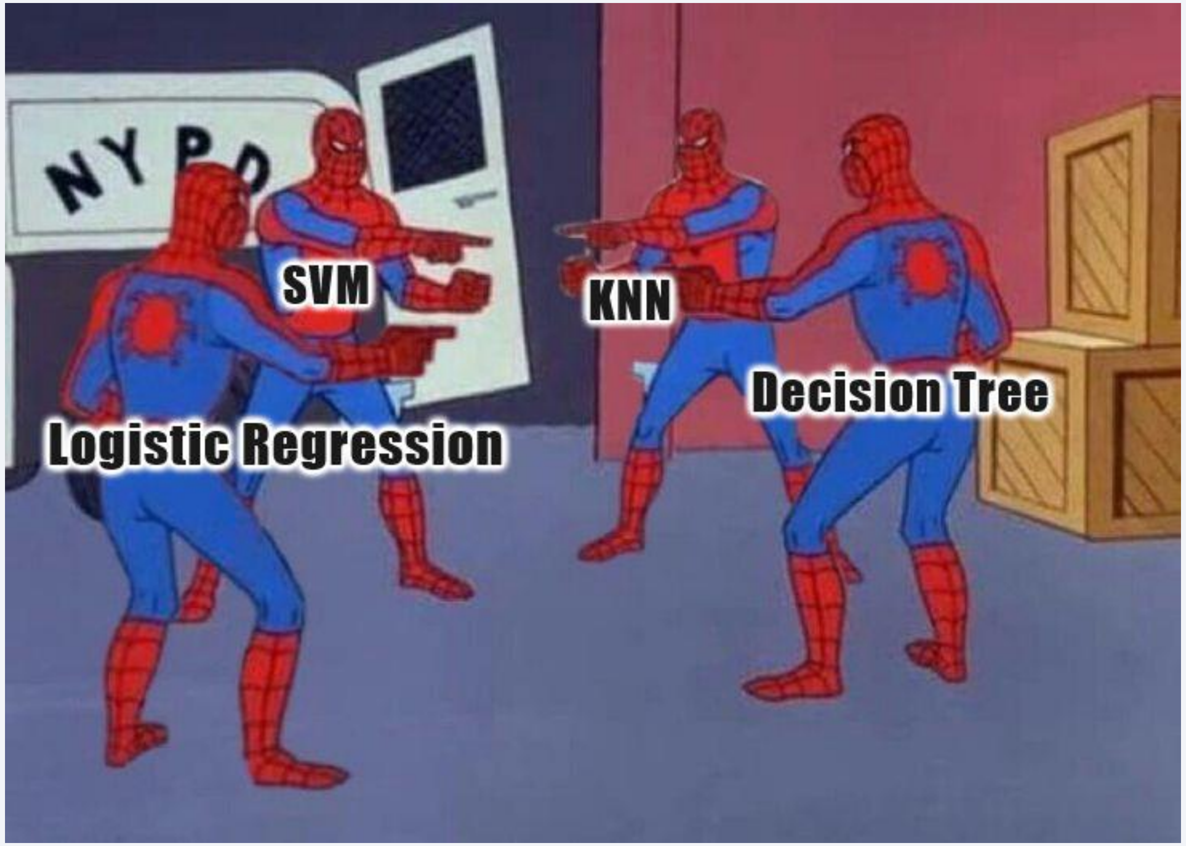


Section 5

Predictive Analysis (Classification)

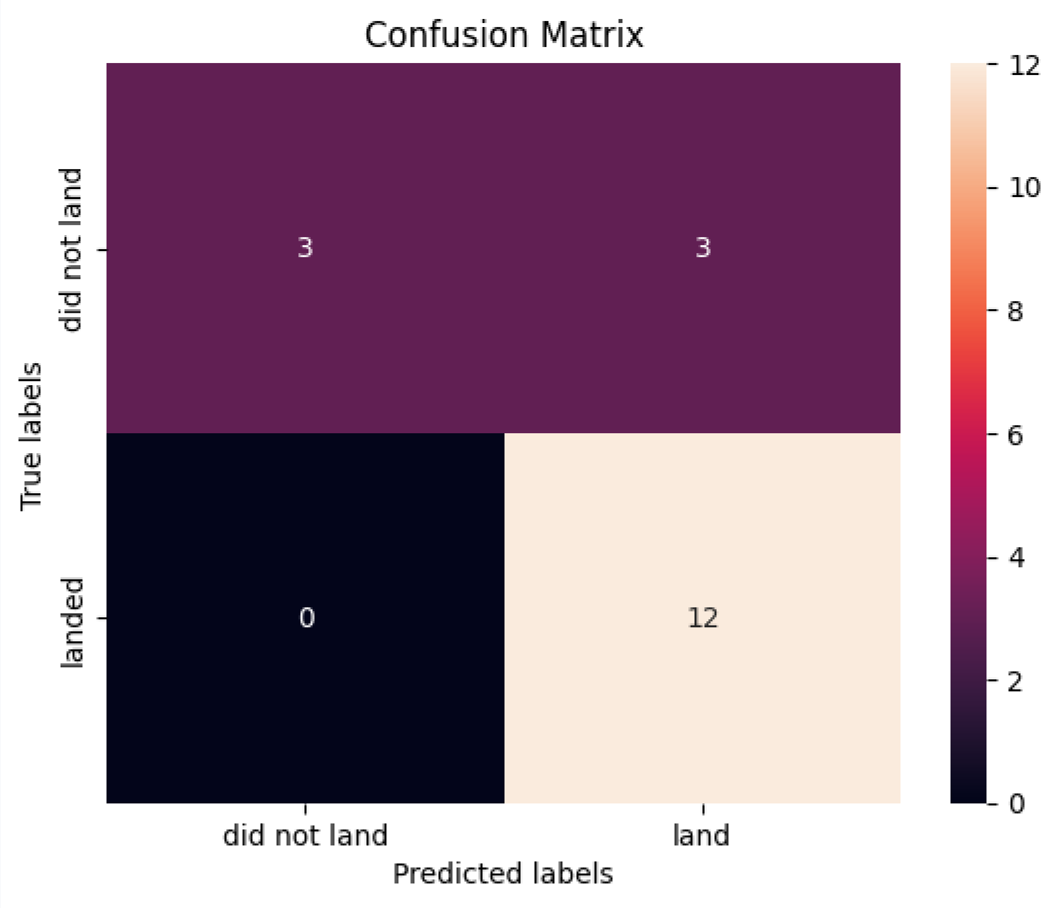
Classification Accuracy





Practically all these algorithms give the same result

Confusion Matrix



- We obtained the same confusion matrix in all models.
- Consequently, we can draw the same conclusions for all of them.
- The models can distinguish between the different classes. We see that the problem is false positives.
- True positives: 12. False positives 3

Conclusions (1/2)

- Succes rate has a clear upward trend from 2013 to 2020
- ES-L1, GEO, HEO and SSO orbits have the maximum success rate with 100%
- The higher the number of flights, the higher the rate of successful launches from the launch site
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000)
- All launch sites are in proximity to Equator line. This helps to make the weather conditions more favourable, which helps to increase the success of the launches
- All launch sites are in very close proximity to the coast. It should be noted that some rockets land on platforms over the sea. Since reusing the first stage of the rocket is a key factor, it is logical that the launch sites are not far from the sea

Conclusions (2/2)

- For obvious safety reasons, there is a safety distance between launch sites and certain locations in their proximities, such as cities, highways, railways or the coastline
- The launch site with the highest success ratio is KSC LC 39A with almost 77%
- The launch site with the lowest success ratio is CCAFS SLC-40 with just 57%
- FT Booster Version is the most successful in different payload ranges
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS orbits.
- Logistic regression, SVM, decision tree and KNN practically give the same result as predictors. With the same problem: false positives

Appendix

- GitHub URL of the repository: https://github.com/SamuelCastillo14/SpaceX-Falcon-9-Landing-Prediction/tree/main
- Falcon 9 web: https://www.spacex.com/vehicles/falcon-9/
- Falcon 9 wikipedia web: https://es.wikipedia.org/wiki/Falcon_9

Thank you!