IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Wim van der Plas
wvdplas@ppid.nl

10 November 2023

Executive Summary

In this project it is proven that with use of historic data on SpaceX launches and data science techniques, it is possible to get insight on which combinations of launch site, payload mass and booster type can maximize the chance on a successful landing of the first stage of a rocket.

This insight can be used in setting the commercial price of a customer launch request. When there is a big chance on success, the price can be sharp. When the successful landing is questionable, the price should be realistic.

In the sales process, the sales team should consult the data science team to help in determining the success rates before sending commercial proposals to customers.

Outline

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

Introduction

SpaceX, the company of well-known founder Elon Musk, advertises Falcon 9 rocket launches on its website with a cost of 62 million dollar. Although this is a considerable amount of money, it is relatively inexpensive compared to competitors. Their launches cost up to 165 millions.

Much of the savings is accomplished because SpaceX can reuse the first stage. This will only be with a controlled landing of the first stage. And this is not with all missions. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.

In this project we will use data science techniques to determine the key factors that influence the chance of a successful landing of the first stage of a rocket.

Main question to answer is which factors help in predicting a successful landing of the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data has been retrieved from Spacex sources describing the specifications of launches in the past
  - WIKI pages have been used to retrieve data from tables on web pages
- Perform data wrangling
  - Handle missing values, prepare high-level insights and prepare data for further analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - With the retrieved data, prediction models of different types have been trained, optimized and ranked on their accuracy

Data Collection

During the project we have used different techniques to gather data about finished launch missions:

- Using API requests: returned JSON format results will be converted to a data frame to be analyzed
- Using web scraping: data on Wiki pages related to Falcon 9 launches will be parsed to a data frame for analysis and visualization

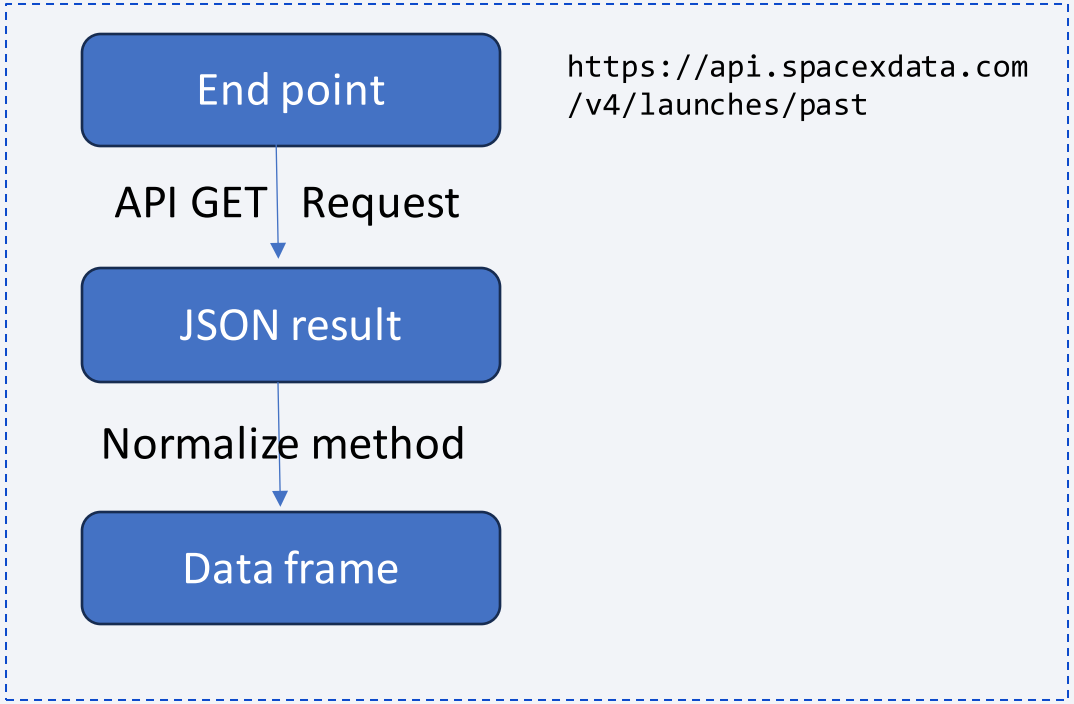
Flowcharts of the two types of data gathering follow in the next slides.

Data Collection – SpaceX API

In this step I have worked on retrieving data from the supplied end point. The result has been formatted from JSON to a Pandas Data frame.

Supplied functions have been used to enrich the data with BoosterVersion, Launch site and Payload data.

And as final step I have solved the issue with missing values for Payload mass (substituted with mean value)



GitHub URL of the completed SpaceX API calls notebook:

https://github.com/WimvanderPlas/DataScienceCourse/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Using WIKI pages with information on Falcon 9 launches, I have created a BeautifulSoup object. Filtering the object on specific tables, I have filled a Pandas data frame containing Falcon 9 launch records.

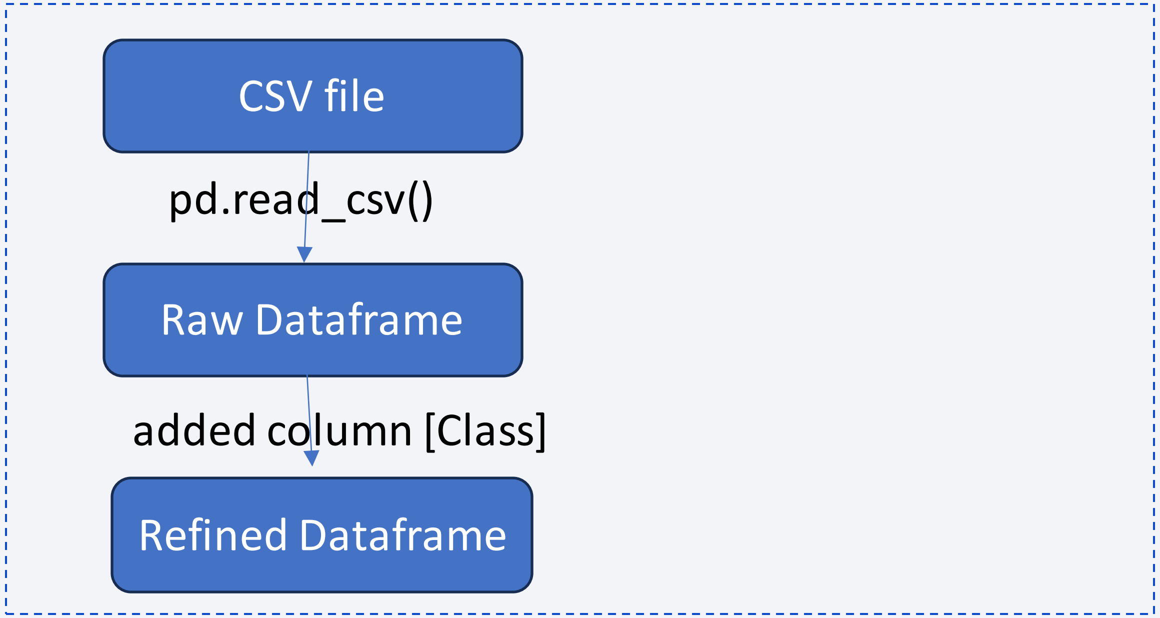
GitHub URL of the completed web scraping notebook:

https://github.com/WimvanderPlas/DataScienceCourse/blob/main/jupyter-labs-webscraping(1).ipynb

Data Wrangling

CSV data has been loaded in a data frame for analyzing:

- percentage missing values in columns
- number of launches per site
- the number and occurrence of each orbit type
- the number and occurrence of mission outcome per orbit type



Finally I added a column Class to the data frame which represents the landing outcome

GitHub URL of the completed data wrangling notebook:

https://github.com/WimvanderPlas/DataScienceCourse/blob/main/labs-jupyter-spacex-Data_wrangling.ipynb

EDA with Data Visualization (1)

In this chapter, several charts have been plotted to explore relationships between characteristics of a launch and the landing outcome. These relationships will help in predicting whether a launch will be successful or not.

Charts plotted:

- FlightNumber vs. PayloadMass
- FlightNumber vs. LaunchSite
- PayloadMass vs. LaunchSite
- Success rate of each Orbit type
- FlightNumber vs. Orbit type

EDA with Data Visualization (2)

Charts plotted (continued):

- PayloadMass vs. Orbit type
- Launch success yearly trend

And as last steps, in preparation of setting up a predictive analysis model, features have been selected that affect the success rate and OneHotEncoding has been applied to them.

GitHub URL of the completed EDA with data visualization notebook:

https://github.com/WimvanderPlas/DataScienceCourse/blob/main/jupyter-labs-eda-dataviz.ipynb

Build an Interactive Map with Folium

In this part three tasks have been executed:

1. Create a Folium map with circles and markers for each launch site. This to visualize where the launch sites are located on the continent
2. Mark the success/failed launches for each site. The added markers and green/red colour give insight on the success rate per location
3. Calculate the distances between a launch site and its proximities. For site VAFB SLC-4E lines have been added to coast line and nearest highway, railway and city, including a marker showing the distance in km. The added info gives information on considerations where to build a launch site.

GitHub URL of the completed interactive map with Folium notebook:

https://github.com/WimvanderPlas/DataScienceCourse/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

EDA with SQL (1)

Queries performed in this chapter:

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

```
SELECT DISTINCT Launch_site FROM SPACEXTABLE
```

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

```
SELECT * FROM SPACEXTABLE WHERE Launch_site like 'KSC%' LIMIT 5
```

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

```
SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

```
SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```

- List the date where the succesful landing outcome in drone ship was acheived.

```
SELECT MIN(DATE) from SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)'
```

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

```
SELECT Booster_Version from SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)' AND
PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_< 6000
```

EDA with SQL (2)

- List the total number of successful and failure mission outcomes

```
SELECT Mission_Outcome , Count(Mission_Outcome) from SPACEXTABLE GROUP BY Mission_Outcome
```

- List the names of the booster_versions which have carried the maximum payload mass.

```
SELECT Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT
MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

- List the records which will display the month names, succesful landing_outcomes in ground pad, booster versions, launch_site for the months in year 2017

  SELECT Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)

- 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

  SELECT Landing_Outcome, COUNT(Landing_Outcome) as Number FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Number DESC

GitHub URL of the completed EDA with SQL notebook:

https://github.com/WimvanderPlas/DataScienceCourse/blob/main/jupyter-labs-eda-sql-edx_sqllite.ipynb

Build a Dashboard with Plotly Dash (1)

The dashboard built in this topic consists of two parts:

A site selection box with a pie chart. If "All sites" is selected, the pie chart shows the Total success launches per Site. When a specific site is selected in the selection box, the pie chart shows the successful and unsuccessful launches of that site.

A callback function was added to redraw the pie chart on each change in the site selection.

Build a Dashboard with Plotly Dash (2)

The other part is a scatter plot showing the relation between Payload Mass and Succes. The scatter plot takes the site selection in account. And the dots in the plot are given a unique color representing the Booster Version Category. In the scatter plot there is also a slider added to narrow the range of payloads shown in the plot.

A callback function was added to redraw the scatter plot on each change in the site selection or payload mass.

The pie and scatter plot help in the decision from which site, with which booster version a certain payload should be launched to have the best chance on a successful landing of the first stage.

GitHub URL of the completed Plotly Dash lab:

https://github.com/WimvanderPlas/DataScienceCourse/blob/main/spacex_dash_app%20final.py

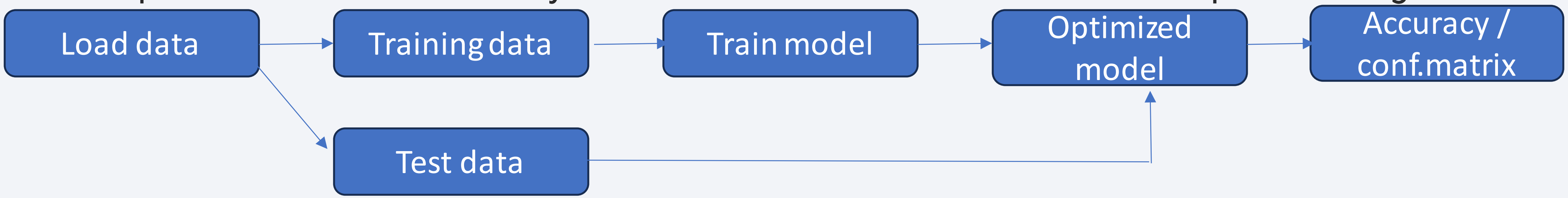
Predictive Analysis (Classification)

The CSV launch data loaded from the SN website was loaded into a data frame. The data frame was loaded into a NumPy array.

The data in the NumPy array was standardized and split into training data and test data.

After that the train data has been used to train different classification models (Logistic Regression, Support Vector Machine, Decision Tree, K Nearest Neighbour) . For each model GridSearch was used to find the optimal hyperparameters set.

For the optimized model the accuracy was calculated and a confusion matrix was plotted using the test



GitHub URL of the completed predictive analysis lab notebook:

https://github.com/WimvanderPlas/DataScienceCourse/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

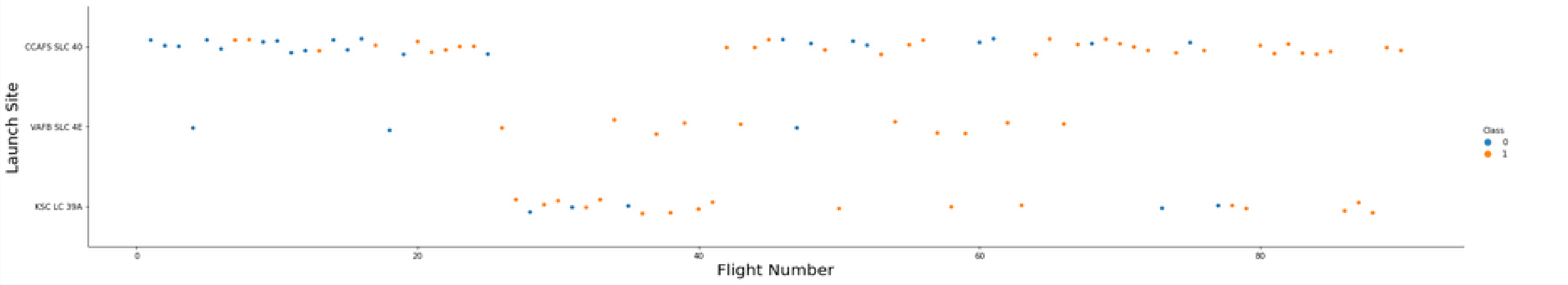
Results

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

Section 2

Insights drawn from EDA

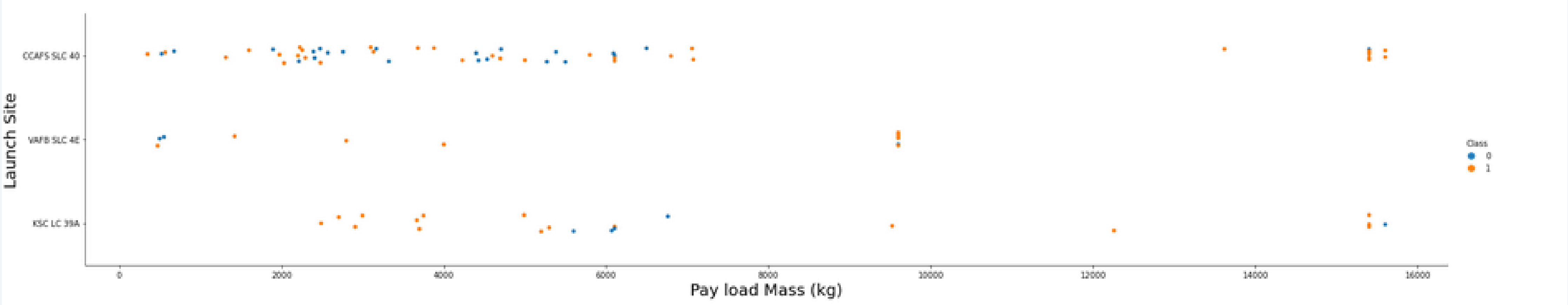
Flight Number vs. Launch Site



On the picture above, you can see a scatter plot of Flight Number vs. Launch Site

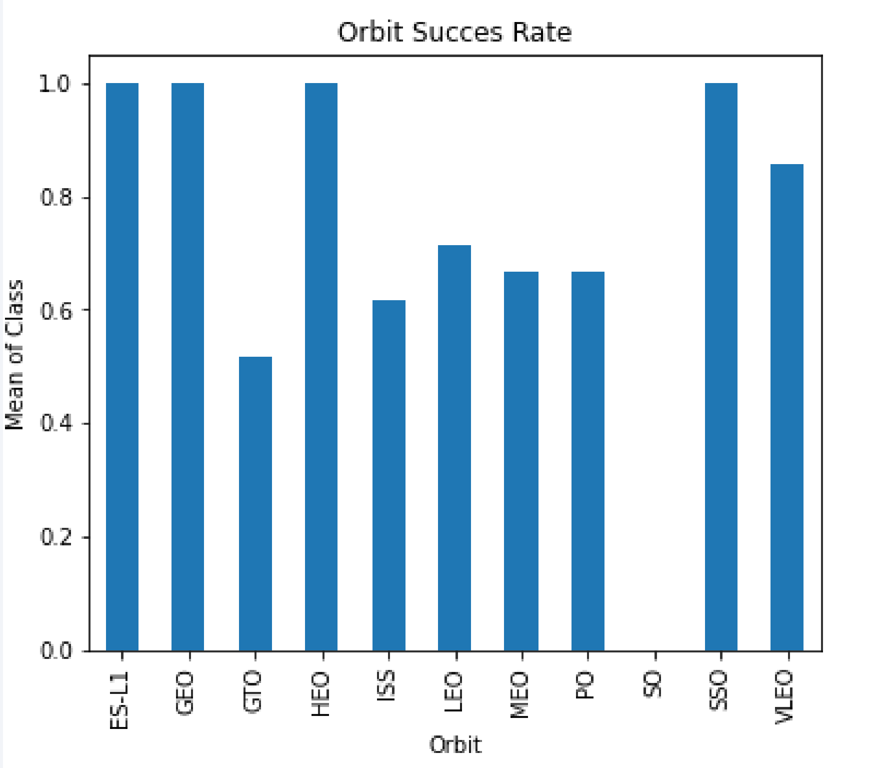
Looking at the plot it is clear that each launch site develops greater success (orange dots) numbers as the flight numbers increase. VAFB has developed a greater chance on success earlier in time. KSC is very successful in flight numbers from about 40 on.

Payload vs. Launch Site



Looking at a scatter plot of Payload vs. Launch Site you can see that two of three launch-site have a payload range where the chance of successful first stage landings is more likely. For example the KSC LC-39A site is highly successful in the payload range until 5500 kg and from 9500 kg on. The VAFB SLC-4E has a flawless record from 1950 kg on. And CCAFS SLC-40 has its best range in 7000 kg payloads and higher.

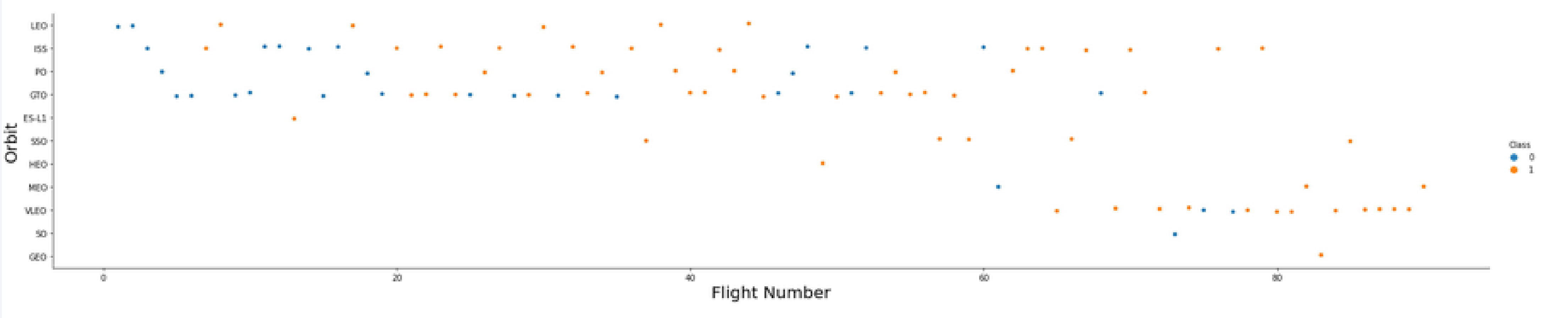
Success Rate vs. Orbit Type



The bar chart on the left shows the mean success rate of each orbit type.

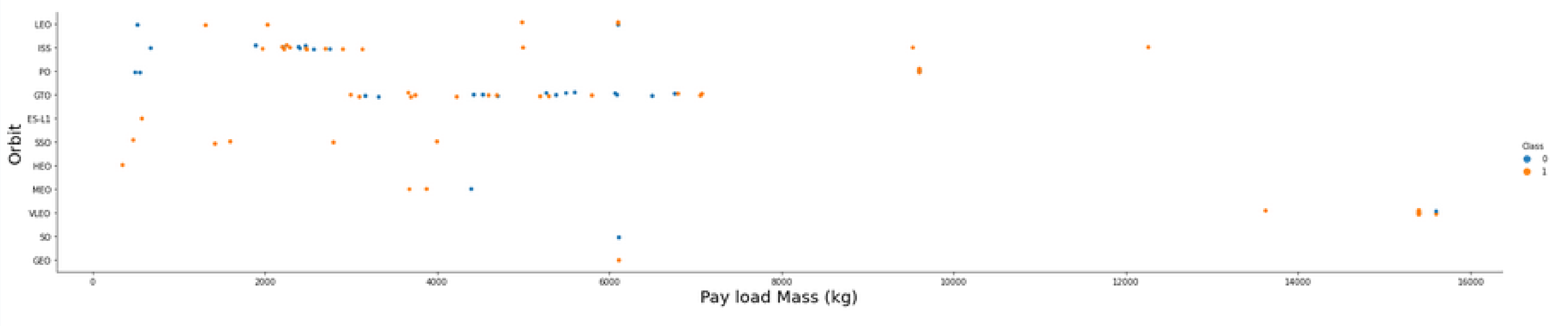
ES-L1, GEO, HEO and SSO have the highest success rates. Rates being a mean of 1 gives full confidence for reusability of stage 1.

Flight Number vs. Orbit Type



In the scatter plot on this page, showing the relationship between Flight Number and Orbit type, there is no clear trend. You can see that as Flight numbers increase, there is a bigger chance on success. But that was already a conclusion on the plot of Flight Number vs. Launch Site slide.

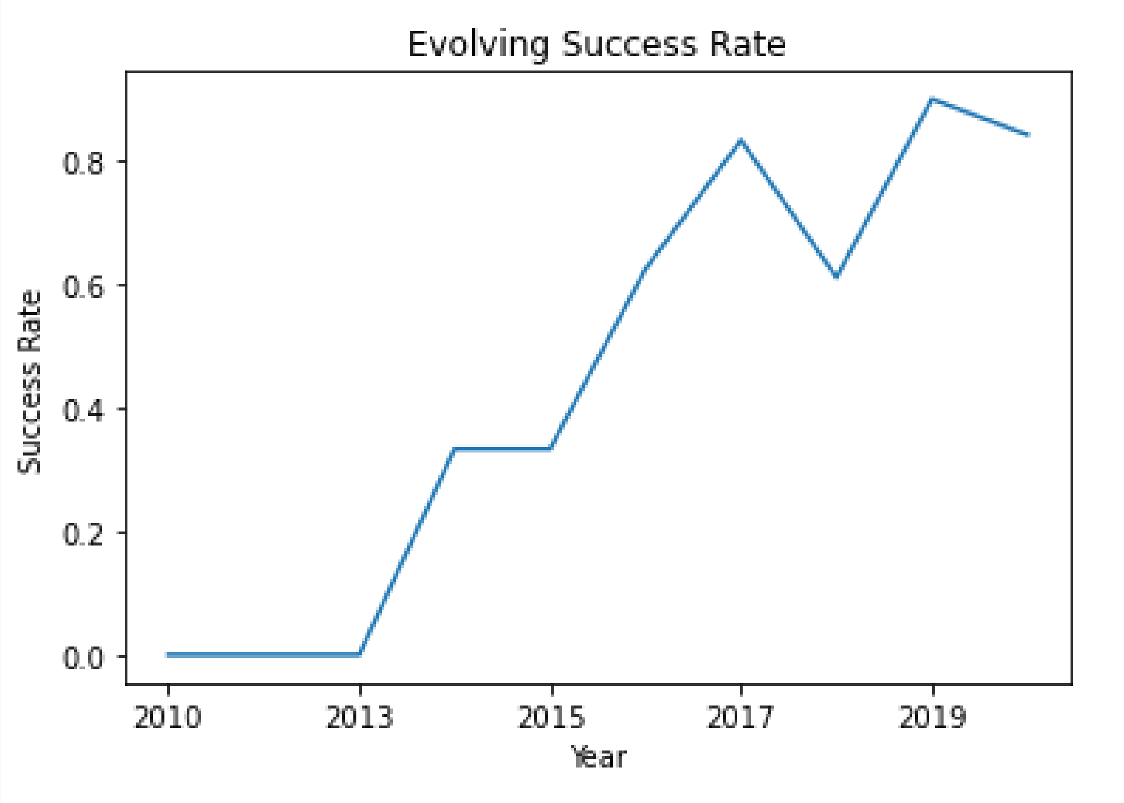
Payload vs. Orbit Type



With heavy payloads the successful landing rate is higher for Polar, LEO and ISS.

For GTO it is hard to see a relationship; both positive landing rate and negative landing (unsuccessful missions) are both there.

Launch Success Yearly Trend



In the line chart of yearly average success rate, you can see that between the years 2013 and 2017 there has been an increasing trend in the success rate with 2014 as steady rate year.

In 2018 there has been a temporary drop in the trend, increasing again in 2019. But the trend is broken again towards 2020.

Conclusion is that the steady increase until 2017 is discontinued in the years after. However, the success rate nowadays is high.

All Launch Site Names

I have used the following query to display the names of the unique launch sites in the space mission:

```
SELECT DISTINCT Launch_site FROM SPACEXTABLE
```

The query result is:

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

Launch Site Names Begin with 'KSC'

For finding 5 records where the launch sites' names starts with `KSC`, I have used the query:

```
SELECT * FROM SPACEXTABLE WHERE Launch_site like 'KSC%' LIMIT 5
```

The LIMIT parameter tailors the output of the query to not more than 5 lines. The query result is as follows:

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

Total Payload Mass NASA

Summarizing the total payload that was carried for customer NASA, was done using the query:

```
SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

The query result (meaning that over 45 tons of load was carried for this specific customer) is:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

To calculate the average payload mass carried by booster version F9 v1.1, I used the query:

```
SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```

The query result shows that for this specific booster type, the load of all launches summarized and divided by the number of flights, gives an average of almost 3 tons.

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

For finding the date of the first successful landing outcome on a drone ship, I have selected all the lines with "Success (drone ship)" as landing outcome, and in this set determined the earliest date with the MIN(DATE) parameter.

```
SELECT MIN(DATE) from SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)'
```

Result of the query shows that on the 27th of May in 2016 the first successful drone ship landing has been made.

MIN(DATE)
2016-05-27

Successful Drone Ship Landing with Payload between 4000 and 6000

The assignment to list the names of boosters which have successfully landed on a drone ship and had payload mass greater than 4000 but less than 6000, was solved using query:

```
SELECT Booster_Version from SPACEXTABLE WHERE Landing_Outcome = 'Success (ground
pad)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_< 6000
```

Using the load boundaries and the landing outcome status, gives the results for the following booster types:

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

In order to calculate the total number of successful and failure mission outcomes, I have ran the following query:

```
SELECT Mission_Outcome , Count(Mission_Outcome) from SPACEXTABLE GROUP BY Mission_Outcome
```

All rows have been grouped based on the value of the column Mission_Outcome and the and he rows in a group have been counted. The result is shown below.

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

Boosters Carried Maximum Payload

Assignment was to list the names of the boosters which have carried the maximum payload mass. The following query was run:

```
SELECT Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE WHERE PAYLOAD_MASS__KG_ =

(SELECT MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

With a subquery the maximum payload in the table was retrieved and in the query the booster versions that have carried this payload have been selected.

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

In order to list the records which will display the month names, successful landing outcomes in ground pad, booster versions and launch site for the months in year 2017, I have setup the query:

```
SELECT substr(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE WHERE
    substr(Date,0,5) ='2017' AND Landing_Outcome = 'Success (ground pad)' ORDER BY Month
```

With the substr statement, specific parts of the date string have been set for selecting records and displaying the results. Query result is:

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
03	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
07	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

For the assignment to 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, the following query has been setup:

```
SELECT Landing_Outcome, COUNT(Landing_Outcome) as Number FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04'
AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Number DESC
```

With the date boundaries, grouping by landing outcome and ordering by the number for that group in descending order, the following subset on the right was selected:

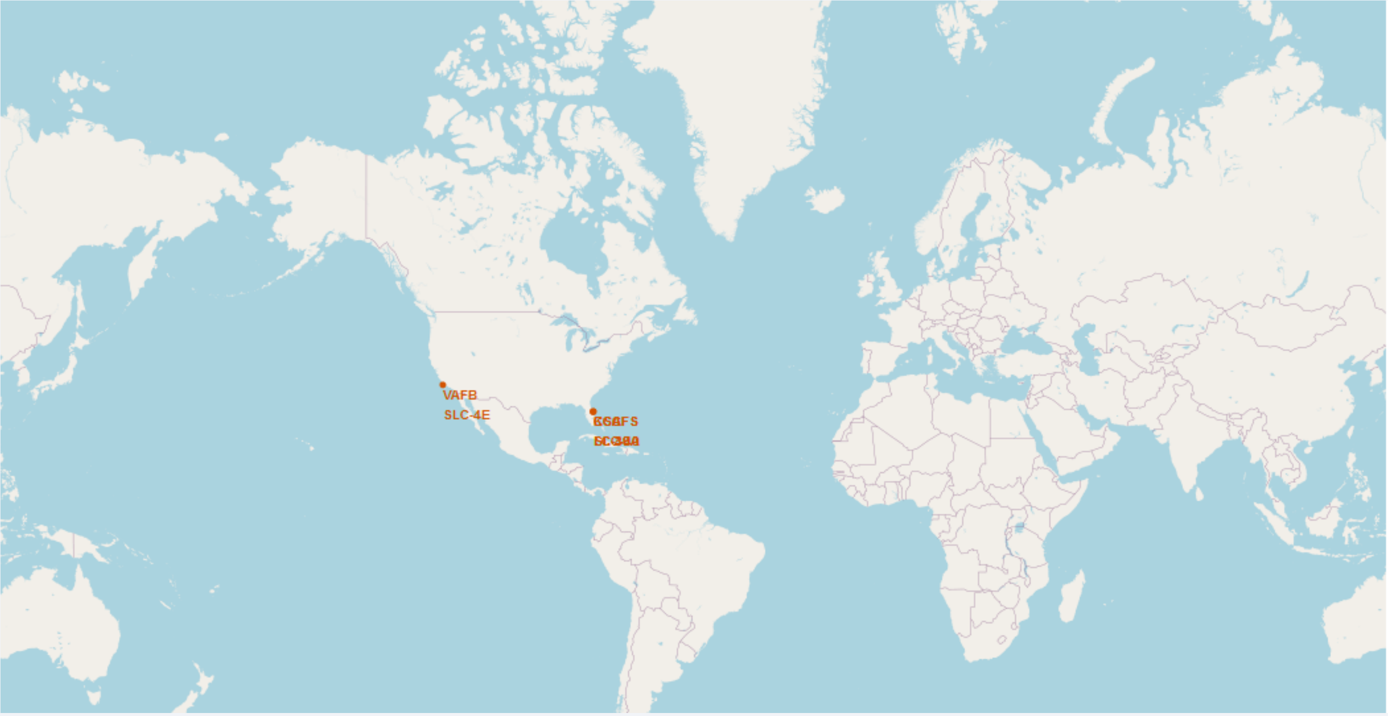
Landing_Outcome	Number
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

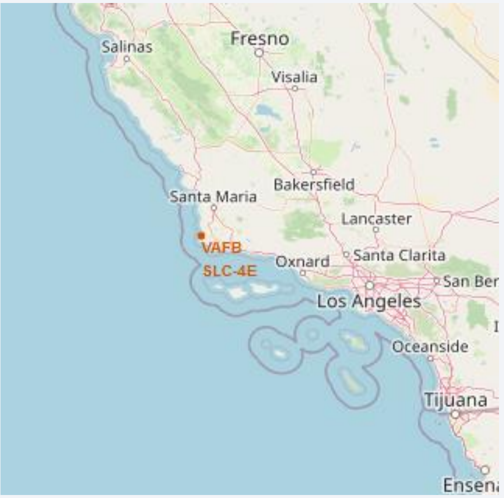
Map with all launch sites (1)

Using Folium a map was made with markers for all SpaceX launch sites. The results is as follows:



Map with all launch sites (2)

Because the launch sites on a global scale map are hard to distinguish, a zoomed in map of west coast and east coast launch sites is added:



Findings on the screenshot:

The Florida launch sites are the closest to the equator line. All launch sites are in very close proximity to the coast. VandenBerg is on the West coast, the other three are on the East coast

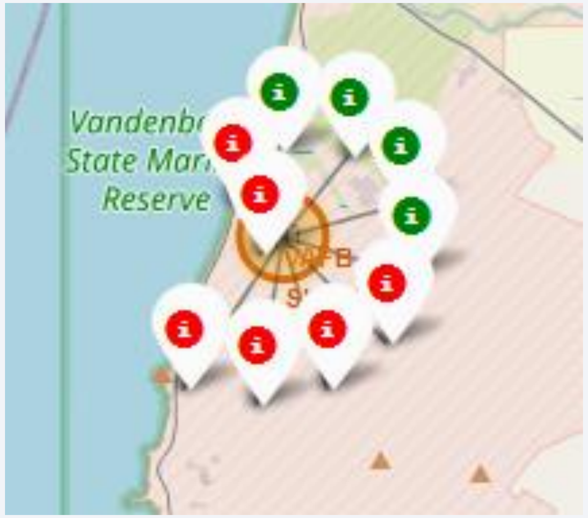
Launching from a location near the equator, the object gets the advantage of the earth's spinning speed.

The earth is turning in East direction. Launching from the East coast gives the space aircraft an additional boost due to the rotational speed of Earth.

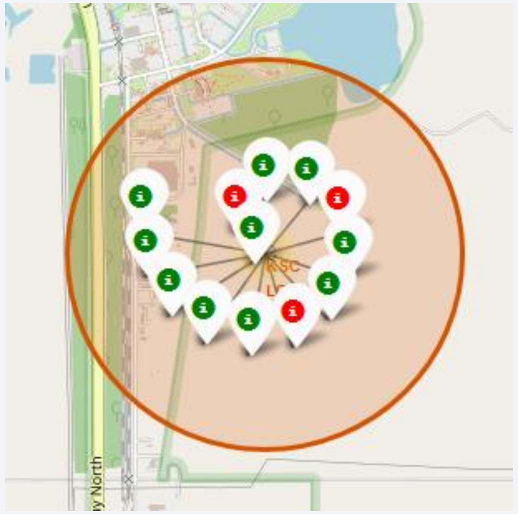
Launching from the Eastern location in Eastward direction the object will be flying over the ocean. If anything might go wrong, it is likely that the object or parts of the object will land in the sea.

Launch outcome per site

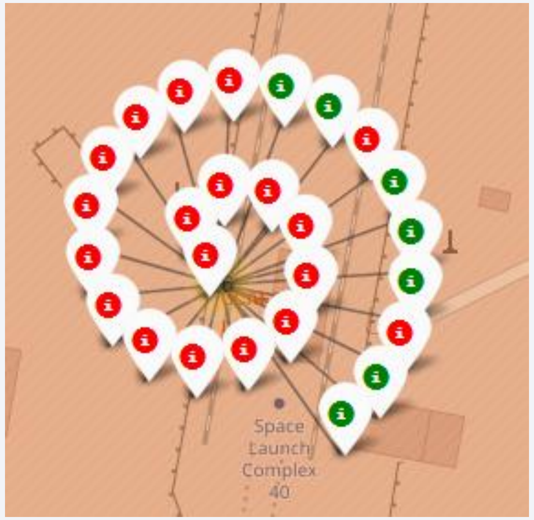
Using markers with a red or green color, representing a successful or unsuccessful landing, insight has been created on how successful a site is.



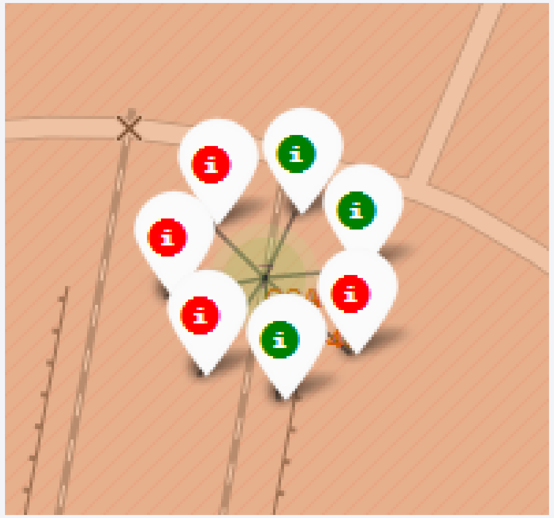
VAFB SLC-4E



KSC LC-39A



CCAFS LC-40

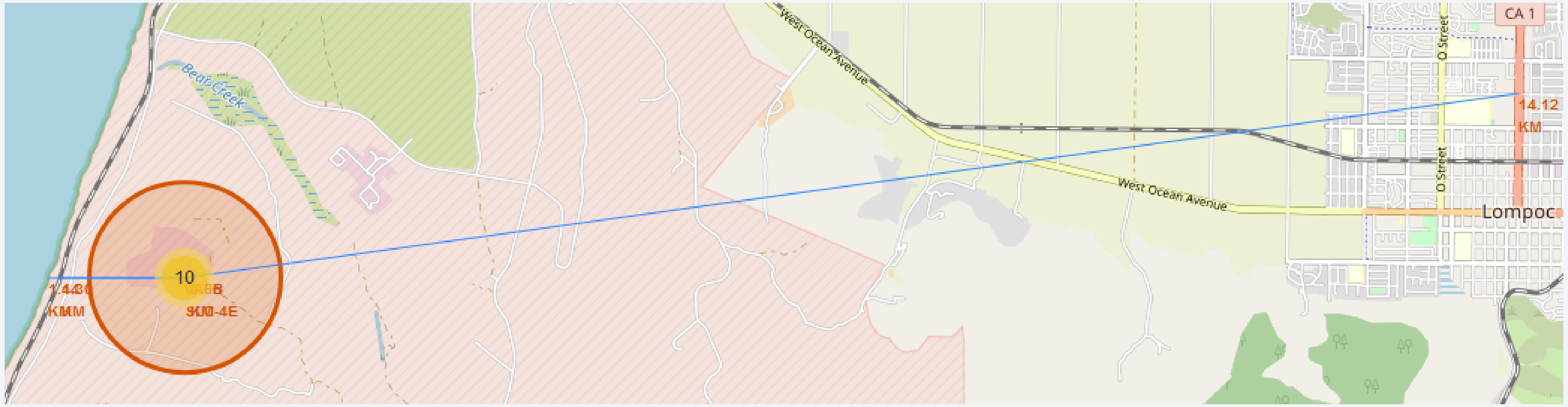


CCAFS SLC-40

Looking at the markers, it is clear that the KSC LC-39A site has the highest success rate.

Explore launch site surroundings

For one launch site, VAFB SLC-4E, the Folium map has been extended with markers and distance indicators for coast line, closest highway, closest railway and nearest city:



Launch sites are in close proximity to railways, highways and coastline. Being close to railways and highways has advantages on logistics. Rockets, payload, parts and supporting goods need to be transported to the launch site. The proximity to coastline has safety advantages. The sooner a missile is away from inhabited area the better it is in case of issues. Launch sites keep a certain distance away from cities for the same safety reasons mentioned for coastline proximity.

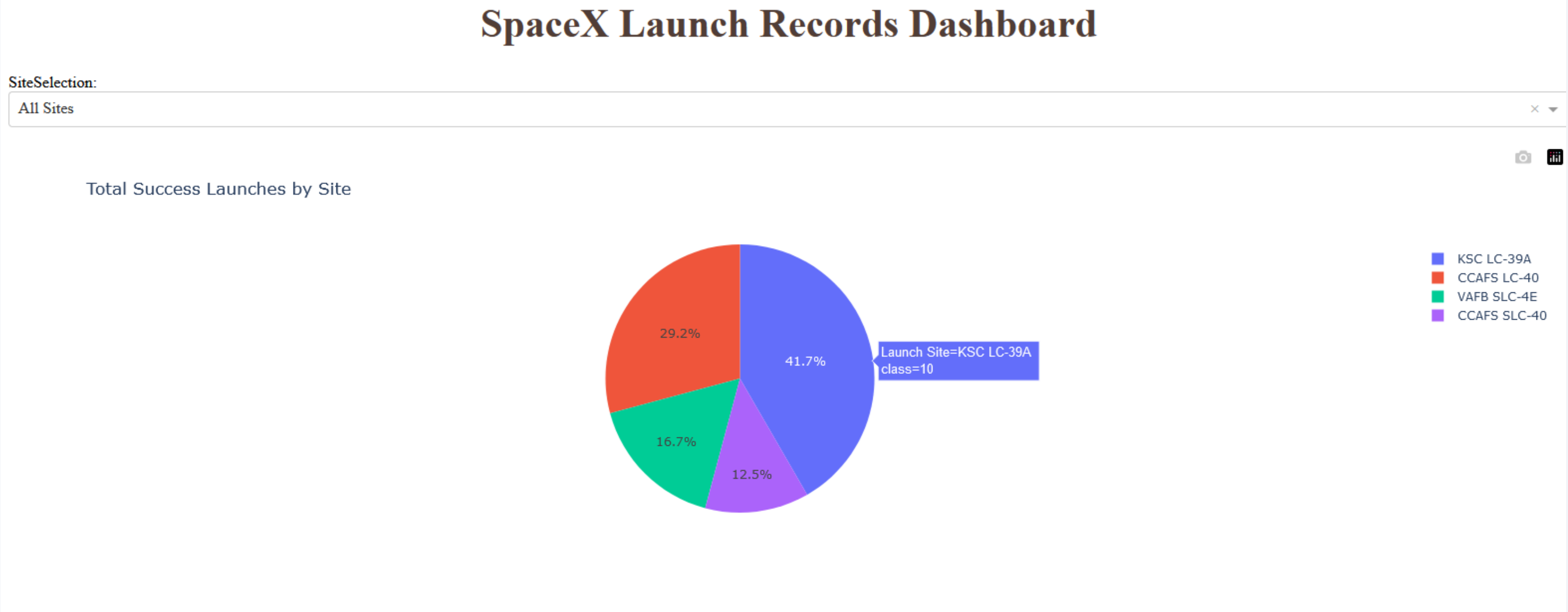
Section 4

Build a Dashboard with Plotly Dash

Launch Success counts for all sites

The pie chart below shows the launch success count for all sites, giving the share in the total for each site as a percentage.

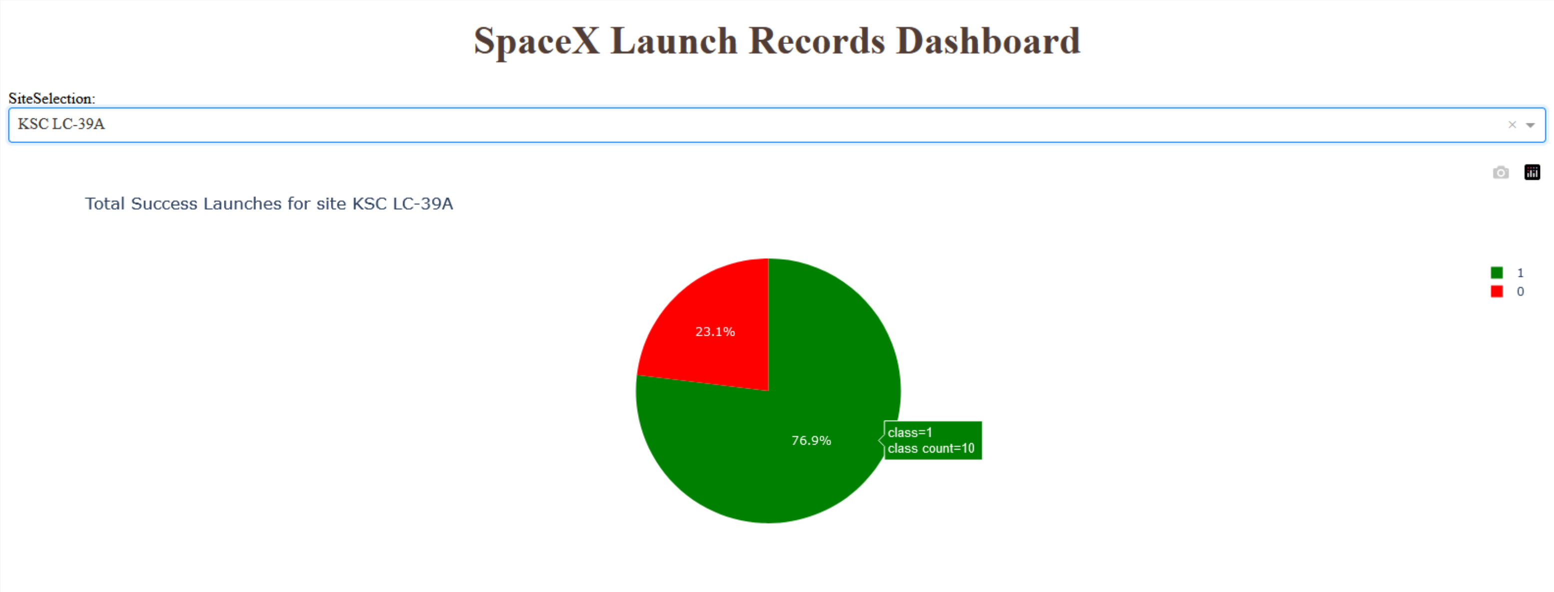
From the figure it is clear that the largest contribution in successful launches is made by the site KSC LC-39A (10 times)



Most successful launch site

With the selection box connected to the displayed pie chart, it is possible to select the launch site with the highest success rate.

In this situation the most successful site is KSC LC-39A. It not only has the nominal advantage compared to other sites, but also the ratio between success and failure is best for this site. A percentage of 76.9 to be precise.



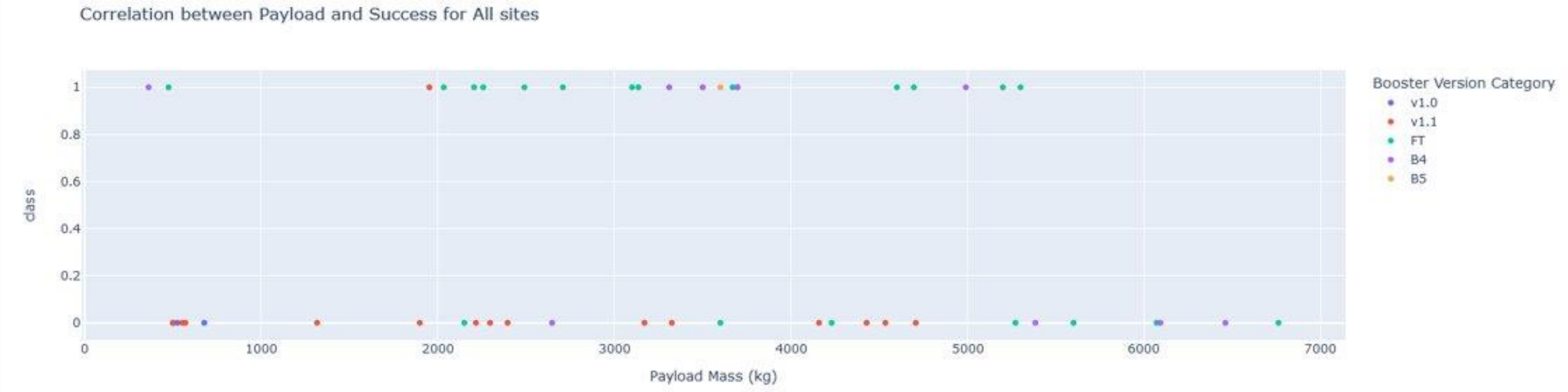
Payload vs. Landing Outcome plot (1)

Using the payload range slider, you can get insights on the relation between Payload and Landing Outcome. Each time the range slider is adjusted, the scatter plot will be redrawn (because of the callback function in the code).

It is hard to show all combinations of Payload, Booster type and Success, so the most interesting I have seen have been selected for this presentation:

- Unfiltered: all payloads, all boosters
- Payloads < 5000 kg, booster type v1.1
- Payloads between 2000 kg and 5500 kg, booster type FT

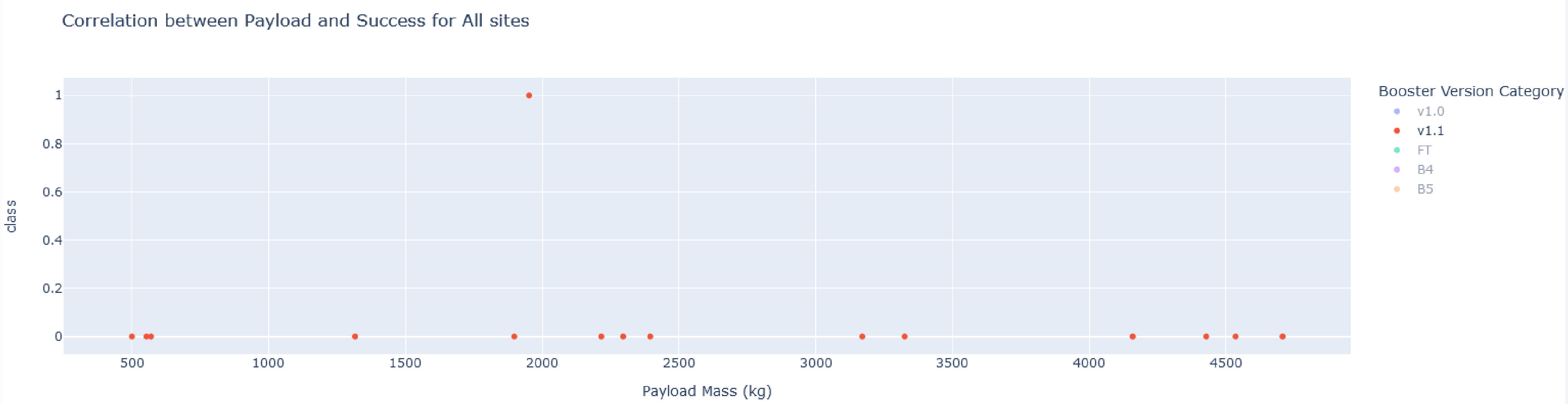
Payload vs. Landing Outcome plot (2)



Unfiltered: having all launches plotted, it is clear that there are successful payload ranges and unsuccessful ranges. For example:

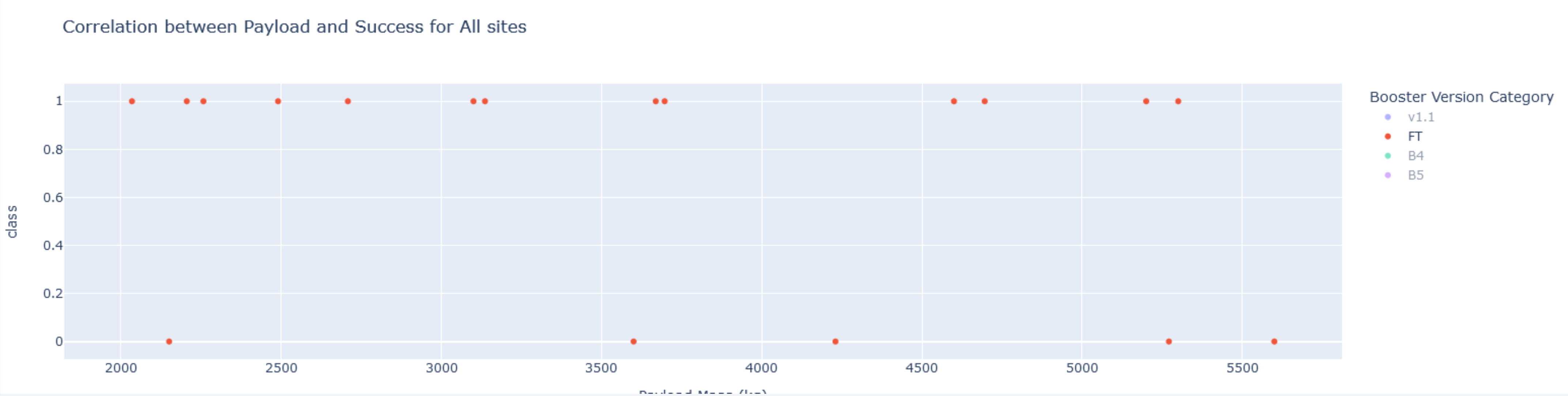
- Payloads < 500 kg (however small in number) have always been successful
- In the range between 1900 and 4000 kg the bigger part of the launches were successful
- In the rage between 4000 and 5300 kg, the bigger part is unsuccessful
- In the range over 5300 kg, there is not a single successful landing outcome

Payload vs. Landing Outcome plot (3)



Payloads < 5000 kg, booster type v1.1 : this seems an unsuccessful combination because in the whole series of landings there was only one mission that was successful.

Payload vs. Landing Outcome plot (4)



Payloads between 2000 kg and 5500 kg, booster type FT:

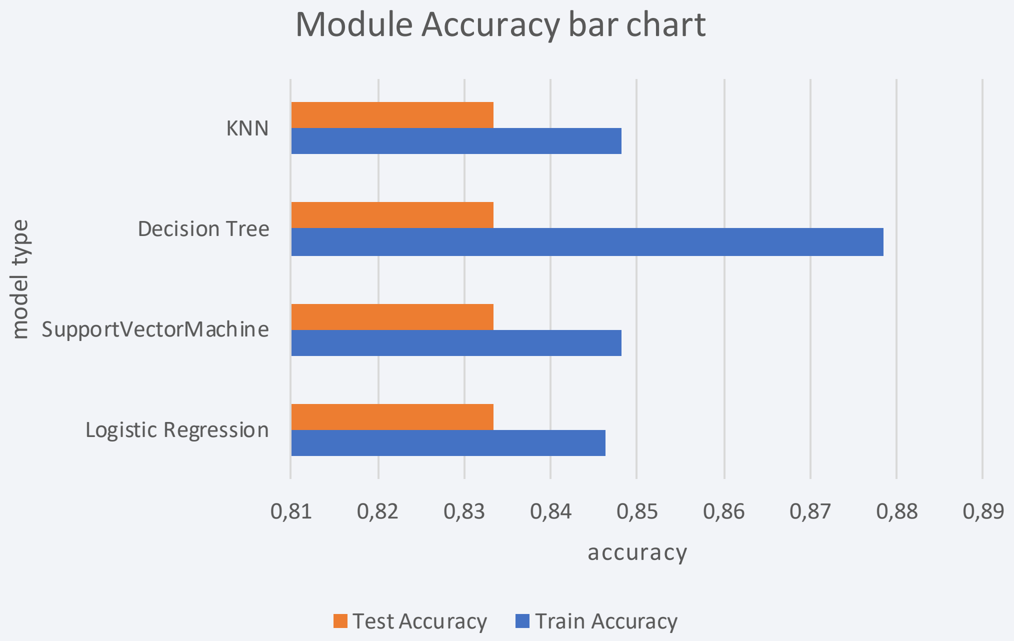
This proves to be a successful combination. Majority of the launches had a successful landing outcome

Section 5

Predictive Analysis (Classification)

Classification Accuracy

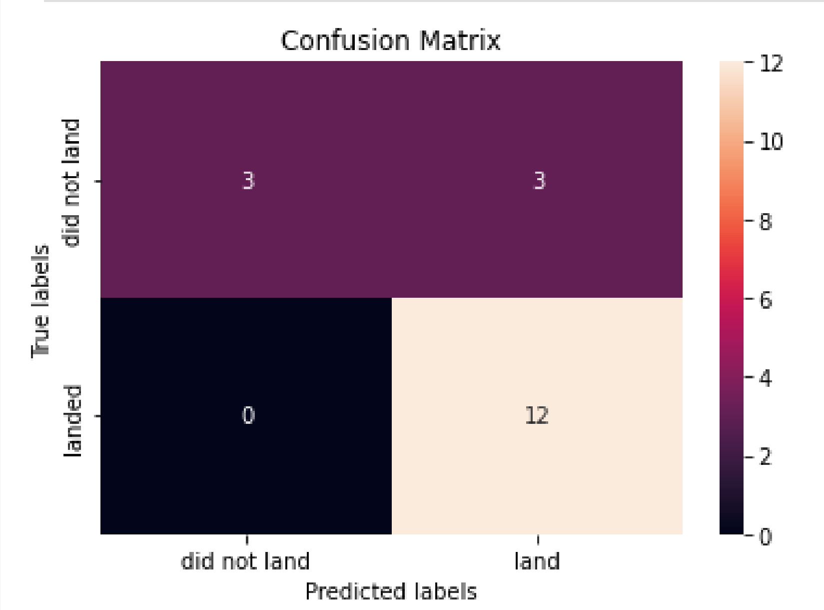
For this topic, predictive models of different kinds have been made and optimized. The accuracy of the models with training and test data has been visualized in a bar chart here:



In the bar chart you can see that the test accuracy of all models is identical. There is a difference in training accuracy; the Decision Tree model performs best.

Confusion Matrix

Shown below is the Confusion matrix of the best performing model (Decision Tree):



Category of concern in the matrix are the three outcomes predicted as "land" but in true life "did not land".

Conclusions (1)

- As the experience in launching increases over years, the success rate is also increasing. There is a learning curve.
- Launch sites are in close proximity to railways, highways and coastline. Being close to railways and highways has advantages on logistics. Rockets, payload, parts and supporting goods need to be transported to the launch site. The proximity to coastline has safety advantages. The sooner a missile is away from inhabited area the better it is in case of issues. Launch sites keep a certain distance away from cities for the same safety reasons mentioned for coastline proximity.
- However there are launch locations with a high success rate, my conclusion is that the success is only partly related to the location. Launching from a specific location is not a guarantee for success

Conclusions (2)

- Choosing the right combination of Payload Mass and Booster Type is highlyimportant. For example as we have seen for masses between 2000 kg and 5500 kg. Here a combination with booster type FT is almost a guarantee for success.

Appendix (1)

Data sources/sets used during this project:

Website	https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
	https://api.spacexdata.com/v4/rockets

API endpoints	https://api.spacexdata.com/v4/launches/past
	https://api.spacexdata.com/v4/rockets
	https://api.spacexdata.com/v4/launchpads
	https://api.spacexdata.com/v4/payloads/
	https://api.spacexdata.com/v4/cores/

CSV files	https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_2.csv

Appendix (2)

Datasets created during this project:

CSV files	https://github.com/WimvanderPlas/DataScienceCourse/blob/main/dataset_part_1.csv
	https://github.com/WimvanderPlas/DataScienceCourse/blob/main/dataset_part_2.csv
	https://github.com/WimvanderPlas/DataScienceCourse/blob/main/dataset_part_3.csv
	https://github.com/WimvanderPlas/DataScienceCourse/blob/main/spacex_web_scraped.csv


Thank you!