IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Piotr Borowski
16.04.2023

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

OUTLINE

- The methods used for this project were data collection, web scraping, data cleaning, exploratory data analysis, data visualization and machine learning simulation
- The study found variables that are likely to affect whether the landing of the falcon 9 will be successful
- The best machine learning model ended up using decision tree method which predicted a 66% chance of success for Falcon 9 mission

EXECUTIVE SUMMARY

INTRODUCTION

- The goal of this project is to predict the chance of successful landing for Falcon 9 rocket and what variables are the most likely to affect the success of the mission

Section 1

Methodology

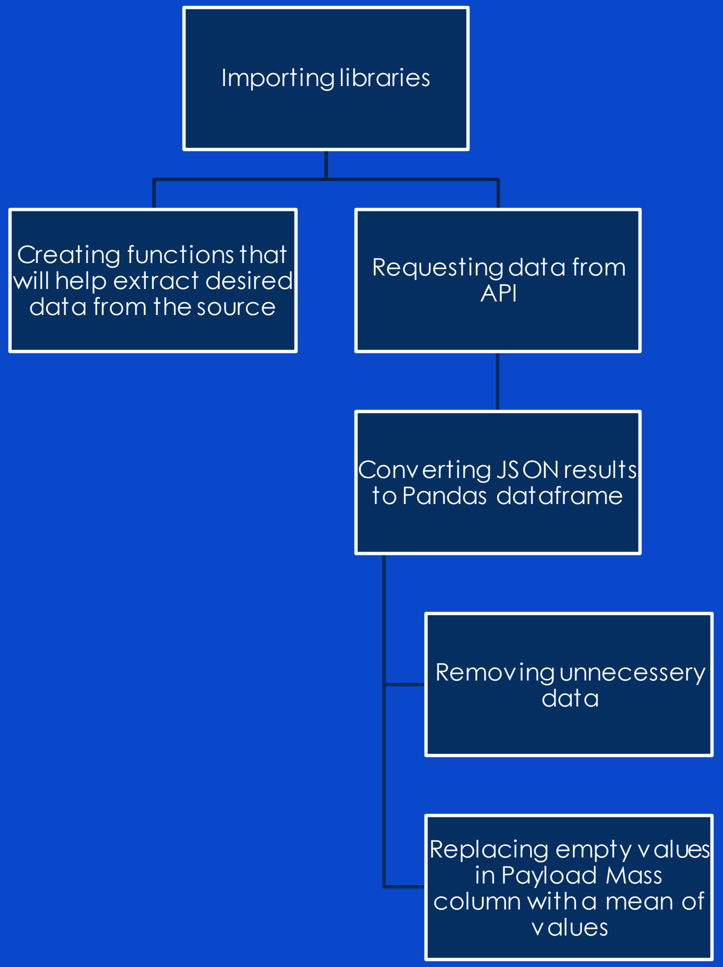
Methodology

- Data collection:
  - Part of the data was collected from the SpaceX API
  - The rest was web scraped from Falcon 9 and Falcon Heavy Launches records on Wikipedia
- Data wrangling:
  - Data was categorized based on the outcome of the landing and number of occurances on each orbit
- Exploratory data analysis (EDA) using visualization and SQL
  - Data was analyzed with usage of Python visualization tools and necessery data was extracted using SQL
- Creating Interactive visual analytics using Folium and Plotly Dash
  - Creating maps in Folium and Dashboards in Plotly
- Predictive analysis using classification models

Data Collection

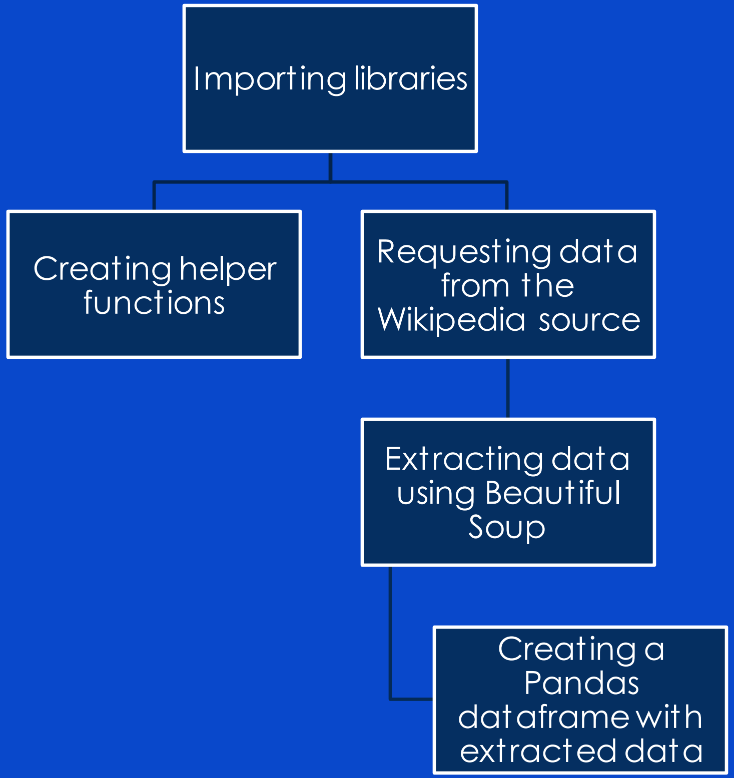
- Data was collected using SpaceX API and web scraping the Wikipedia page. The entire process was completed in Python.
- Data needed to be cleaned since much of the provided information was not useful for the purposes of this analysis.

Data Collection – SpaceX API



- Entire process done in Python can be found here

Data Collection - Scraping



- Entire process done in Python can be found [here]

Data Wrangling

- Data was inspected to avoid data type problems
- Another column was created to specify mission success. As it turns out, landings of 2 out of every 3 missions ends up successfully.

Entire process done in Python can be found here

EDA with Data Visualization

- To visualize the data cat plot, scatter plots, bar chart and a line chart were used since they were the best use in each case to help draw conclusion from them and help analyze the data.

- Entire process done in Python can be found here

EDA with SQL

- Querying for launch sites used in missions
- Querying for total and average payload mass carried by specific boosters
- Querying for the date of the first successful landing in ground pad
- Querying for boosters that were used to carry specific mass, with success in a drone ship
- Querying for successful and failing missions
- Querying for boosters that were able to carry maximum amount of payload mass with the use of a subquery
- Grouping the data depending on the month in 2015 that ended up failing the landing
- Ranking the count of successful landings between the years of 2010 and 2017.

- All SQL queries can be found here

Building an Interactive Map with Folium

- The launch sites were marked with pointers to the sum of launches on each launching site. When zooming in the pointers reveal the outcome of each launch.
- The distance to proximities was calculated and connecting lines were drawn to visualize the distance of launch site to relevant proximities.
- Entire process done in Python can be found here(github) or here(nbviewer to load the maps)

Building a Dashboard with Plotly Dash

- In the dashboard you can select:
- data for specific launch sites
- pie chart of total success launches by launching site
- slider for selecting a specific payload mass range carried by a booster, with a scatter plot connected to instantly visualize the results.

- Code can be found here

Predictive Analysis (Classification)

- Models were built using 80%/20% train/test split, which meant the test data contained only 18 samples. Models included logistic regression, k nearest neighbours, support vector machine and decision tree classifier.
- Test data turned out to be too small for the needs of models, which means additional research is needed.
- The code can be found here

Results

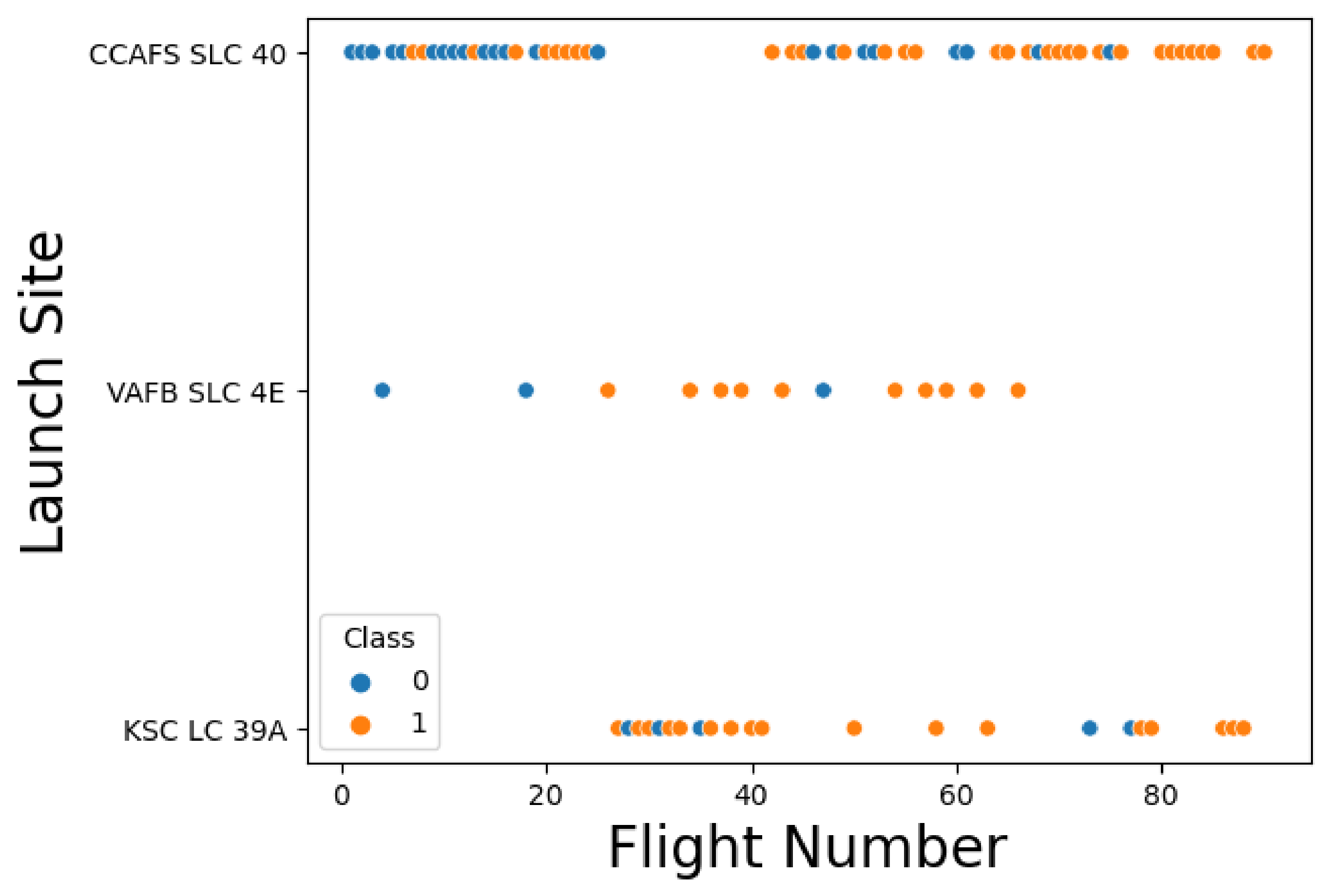
- The launching site with the highest chance of success was KSLC-39A. Statistically the Falcon 9 would want to have an FT booster and one of the 4 of best performing orbits. It's important to remember that even though payload mass doesn't seem to affect success rate, it may affect the choice of boosters, orbits and perhaps launching site
- Best model had a high chance of a false negative and was created on a small sample size therefore it is hard to conclude if it is accurate. Further analysis is needed

Section 2

Insights drawn from EDA

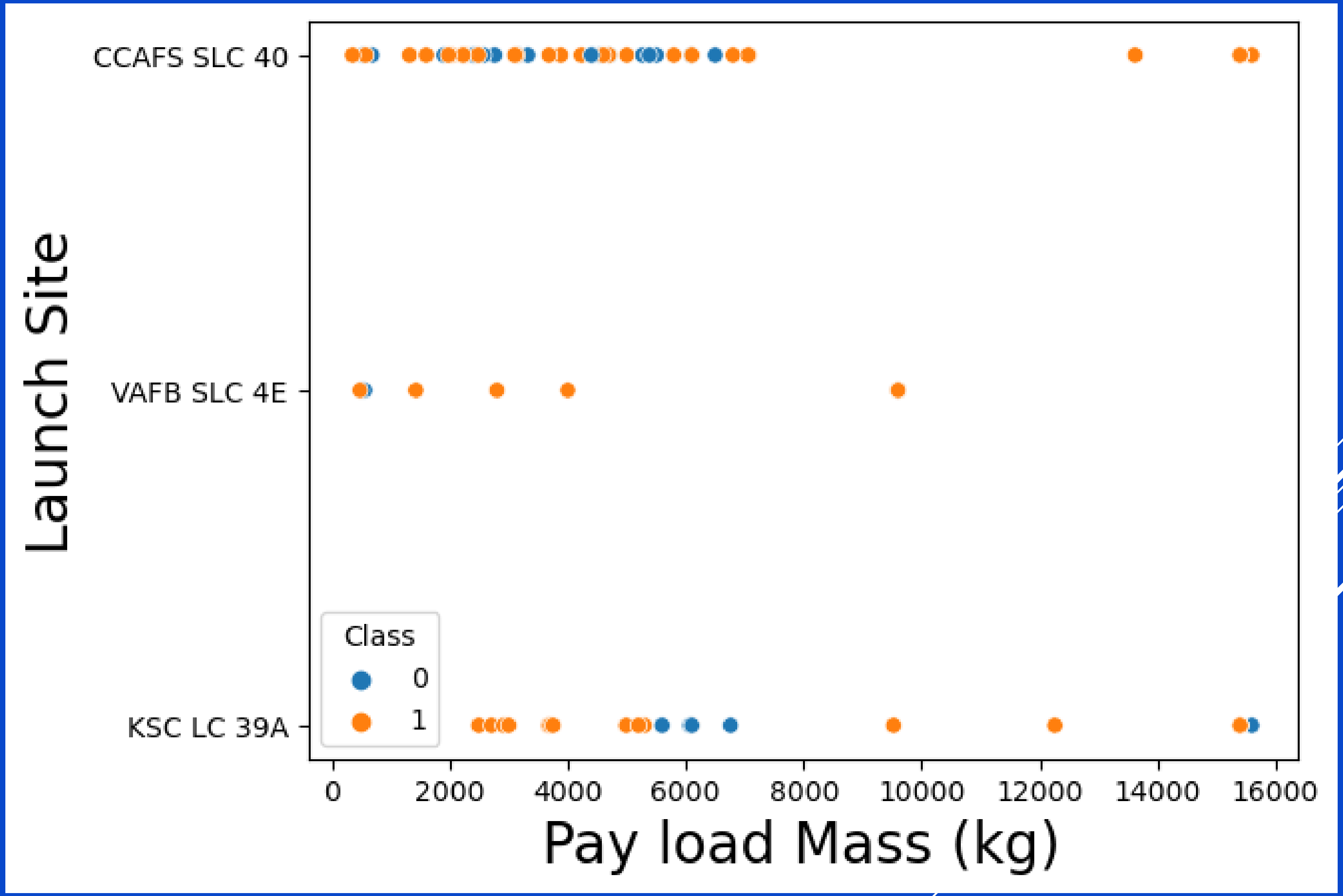
Flight Number vs. Launch Site

- For each Launch Site, later flights have the chance to be successful increasing significantly.
- Early missions have a higher chance of failure than 50% for VAFB and CCAFS
- Launch site VAFB SLC 4E had the lowest amount of flights in the group.



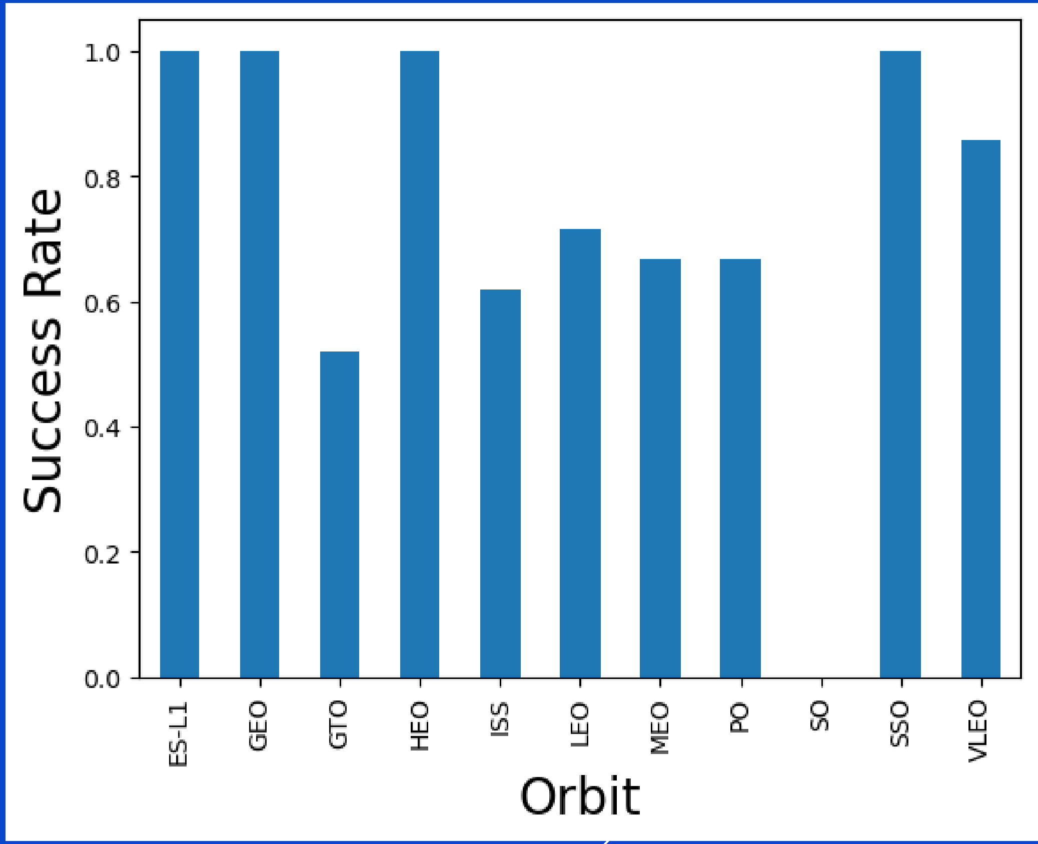
Payload vs. Launch Site

- Plot doesn't suggest a strong relationship between payload mass and success of the mission
- Rockets from launch site VAFB didn't transfer more than 10000 kg of payload mass while other launch sites had several of them



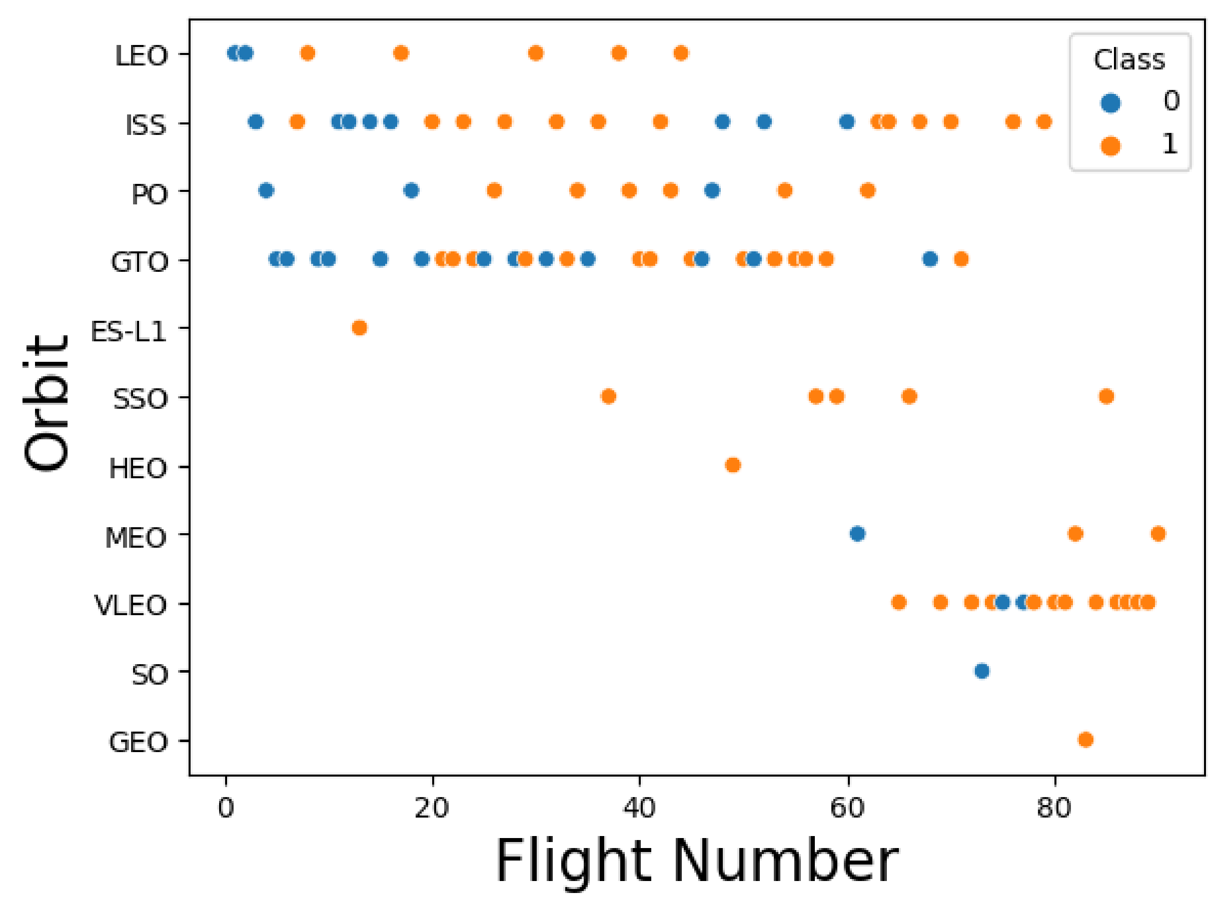
Success Rate vs. Orbit Type

- 4 orbits have success rates of 100% or very close to it
- Most challenging was orbit GTO where only around 50% of missions were successes



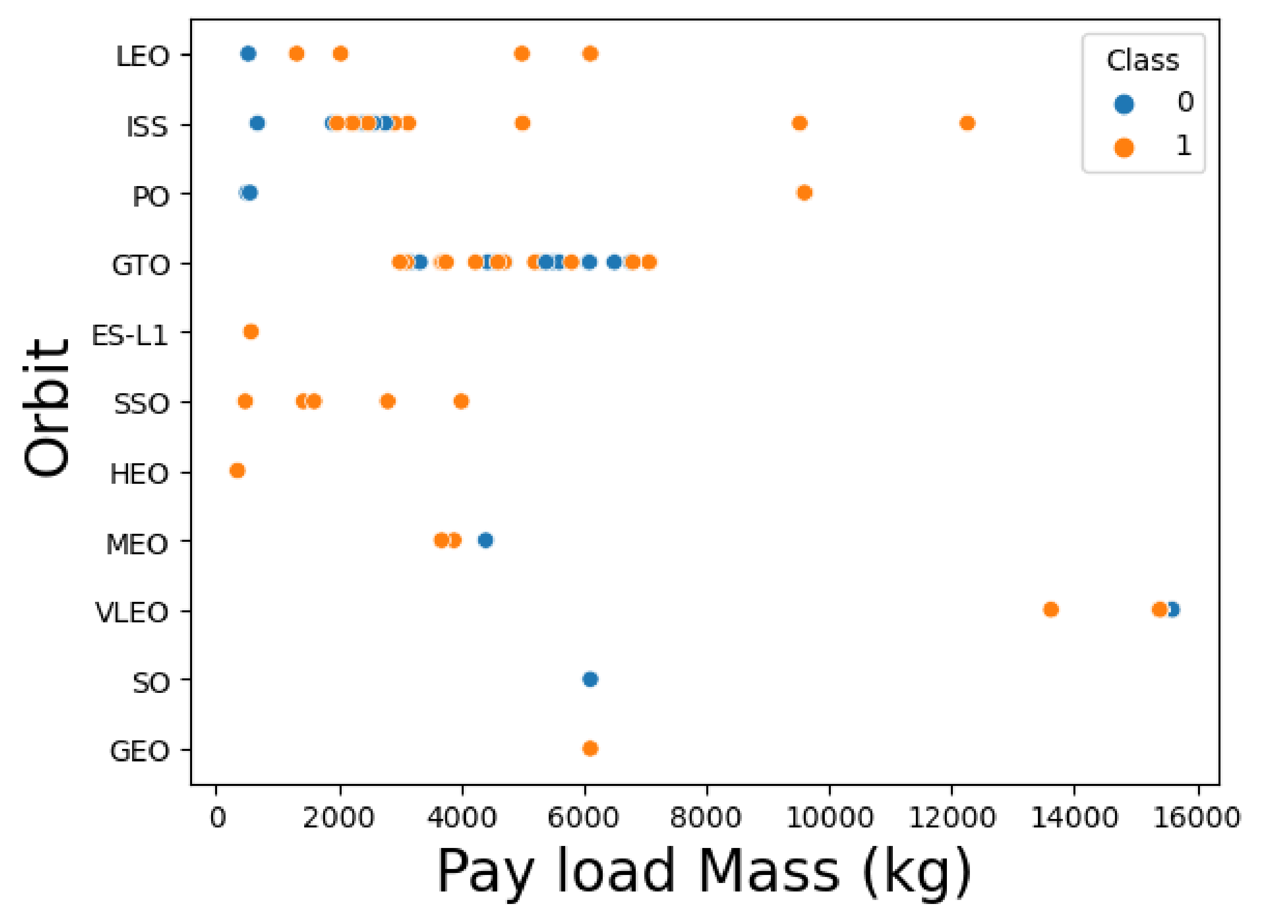
Flight Number vs. Orbit Type

- Later flights have a higher chance of success
- Missions on ISS and GTO orbits had relatively high chance of failure even after 40 flights



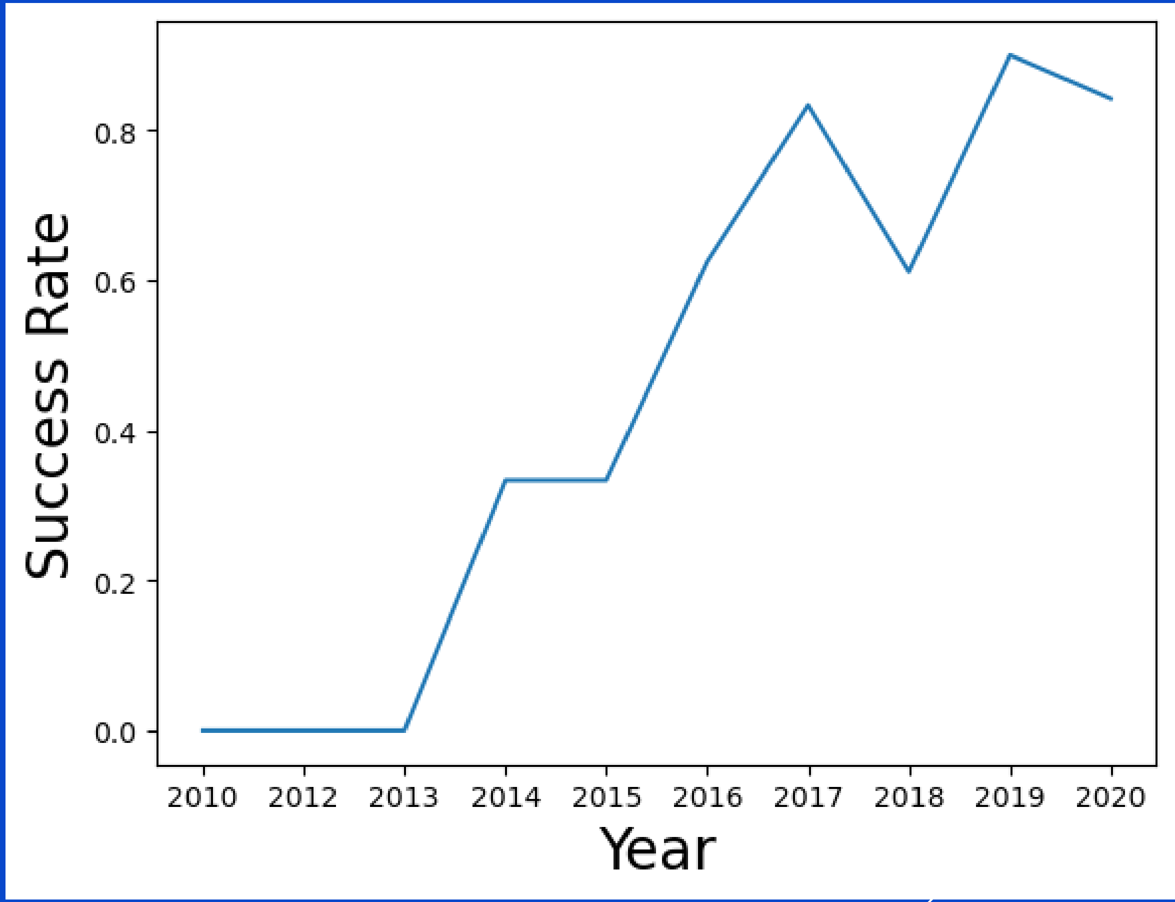
Payload vs. Orbit Type

- Again Payload mass seems not to have a high influence on a success of a mission
- Lowest payload mass flights on LEO, ISS and PO all ended up failing



Launch Success Yearly Trend

- Launch success was relatively low in the beginning with 30% in 2014, but rose from 2015-2017 to about 85%, dropped in 2018 but reached apex in 2019 of above 90%
- Success rate was much higher in late 2010s and 2020 than middle 2010s



All Launch Site Names

- There were 4 unique launch sites
- Information was extracted using a simple select distinct query

```
%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

Launch Site Names Begin with 'CCA'

- Selecting 5 launch sites records that begin with 'CCA' can be achieved with adding % to CCA which means return cells beginning with CCA and have anything afterwards

```
%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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total payload mass for NASA(CRS) missions was 45596 kg
- The total payload mass can be calculated using a sum query

```
%sql select sum(PAYLOAD_MASS__KG_) as 'total payload mass by NASA(CRS)' from SPACEXTBL where Customer = 'NASA (CRS)'
```

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

total payload mass by NASA(CRS)
45596

Average Payload Mass by F9 v1.1

- Average payload mass for boosters F9 v1.1 was 2534.66 kg
- To query average payload mass we need to use the sum function with a where clause

```
%sql select avg(PAYLOAD_MASS__KG_) as 'average payload mass of F9 v1.1' from SPACEXTBL where Booster_Version Like 'F9 v1.1%'
```

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

average payload mass of F9 v1.1
2534.6666666666665

First Successful Ground Landing Date

- First successful ground landing happened 22nd december 2015
- The date of first successful ground landing can be queried a number of ways, e.g by ordering by date and limiting a query to 1 result

```
%%sql
select * from SPACEXTBL
where "Landing _Outcome" = "Success (ground pad)" order by date desc limit 1;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
22-12-2015	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are listed below
- Result can be queried with a where and clauses

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_ , "Landing _Outcome" from SPACEXTBL where "Landing _Outcome" ='Success (drone ship)' and PAYLOAD_MASS__KG_ >
```

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

Booster_Version	PAYLOAD_MASS__KG_	Landing _Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Almost all missions were classified as successes(despite not all of them landing successfully)
- To query for this a group by clause is needed

```
%sql select Mission_Outcome, count(*) as 'Succesful missions' from SPACEXTBL where Mission_Outcome Like 'Success%' group by Mission_Outcome order by
```

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

Mission_Outcome	Succesful missions
Success	98
Success	1
Success (payload status unclear)	1

```
%sql SELECT Mission_Outcome, count(*) as 'Failure_missions' from SPACEXTBL where Mission_Outcome Like 'Failure%' group by 'Mission_Outcome' order by M
```

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

Mission_Outcome	Failure_missions
Failure (in flight)	1

Boosters Carried Maximum Payload

- Several different versions of F9 B5 boosters carried maximum payload
- One of the ways to query for this is 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

2015 Launch Records

- There were 2 failing landings in 2015 , in january and april, both at launch site CCAFS LC-40
- To query for this we need to extract month and year from date, this can be done differently depending on which SQL database is being used

```
%%sql select substr(Date, 4, 2) as Month, `Landing _Outcome`, Booster_Version, Launch_Site, count(*) as 'Occurances' from SPACEXTBL
where substr(Date, 7, 4) = '2015' and `Landing _Outcome` = 'Failure (drone ship)'
group by substr(Date, 4, 2) order by substr(Date, 4, 2)
```

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

Month	Landing _Outcome	Booster_Version	Launch_Site	Occurances
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	1
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	1

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

- Between the date 2010-06-04 and 2017-03-20, more successful landings happened at drone ships than ground pads
- To query for this we need to extract dates, group by landing outcomes and order by counts of landing outcomes

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql
SELECT "Landing _Outcome", count("Landing _Outcome") as LANDING_OUTCOME_COUNT, DATE as Last_Landing
from SPACEXTBL where substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) between '20100604'
and '20170320'and "Landing _Outcome" like "Success%"
group by "Landing _Outcome" order by count("Landing _Outcome") desc;
```

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

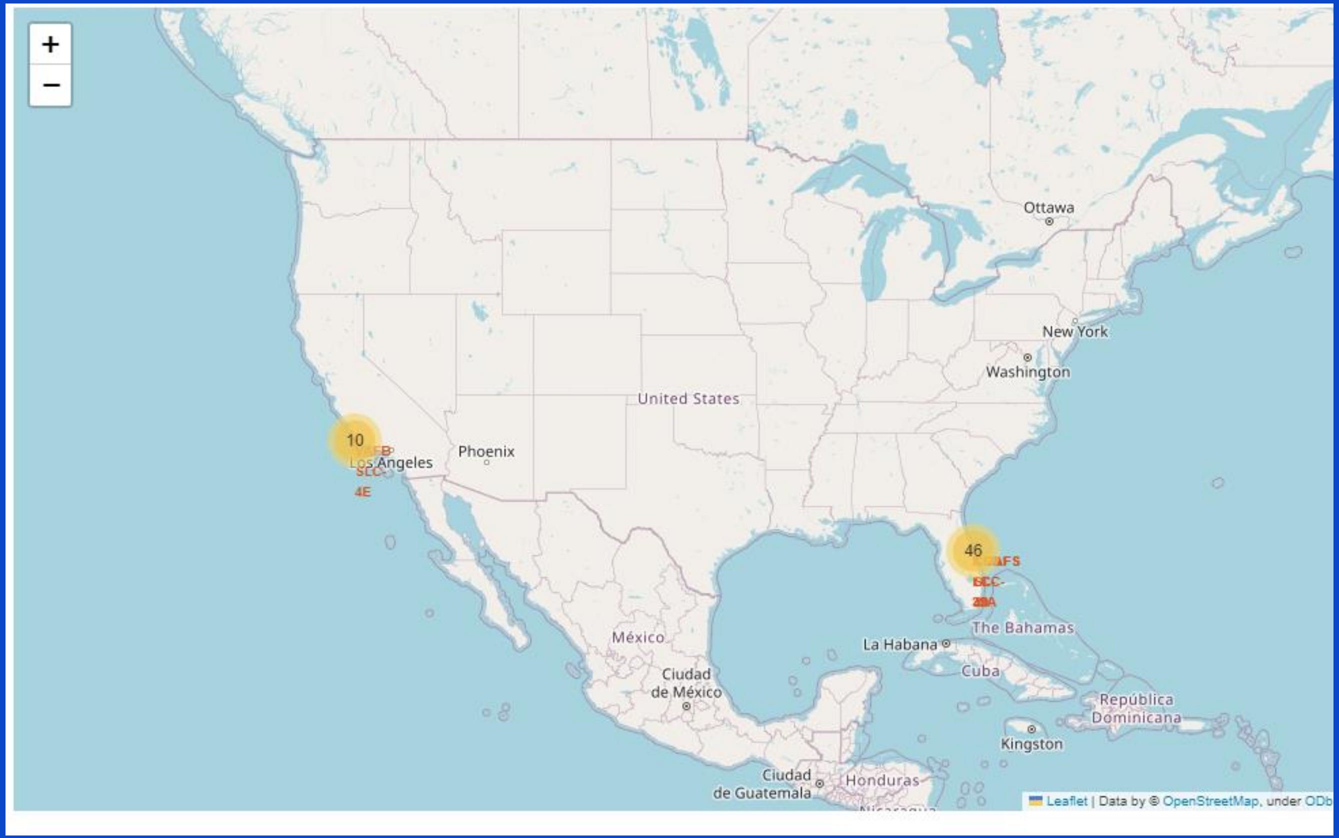
Landing _Outcome	LANDING_OUTCOME_COUNT	Last_Landing
Success (drone ship)	5	08-04-2016
Success (ground pad)	3	22-12-2015

Section 3

Launch Sites Proximities Analysis

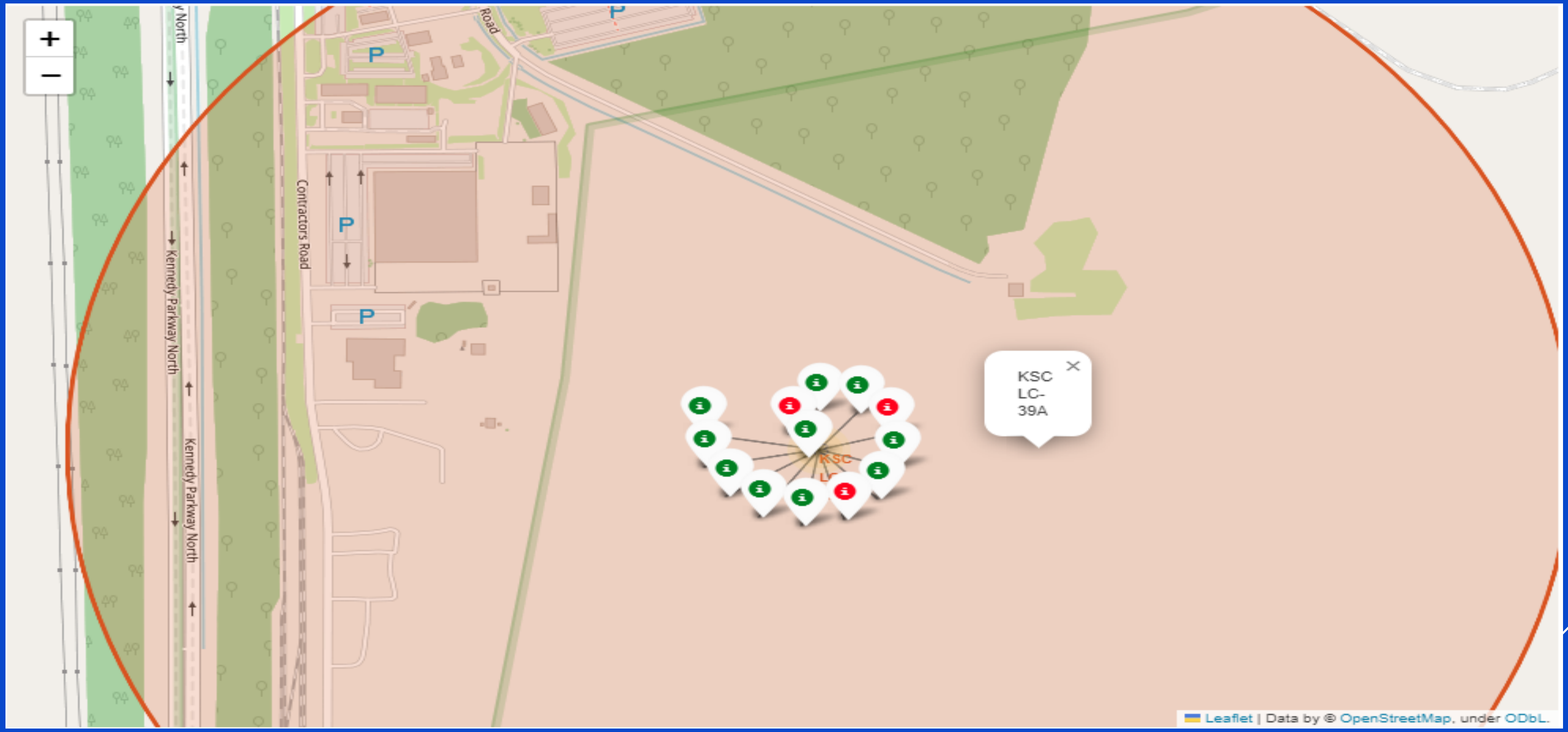
Folium launching sites

- 3 of the launching sites are very close to each other, the 4th is on the west side of the coast



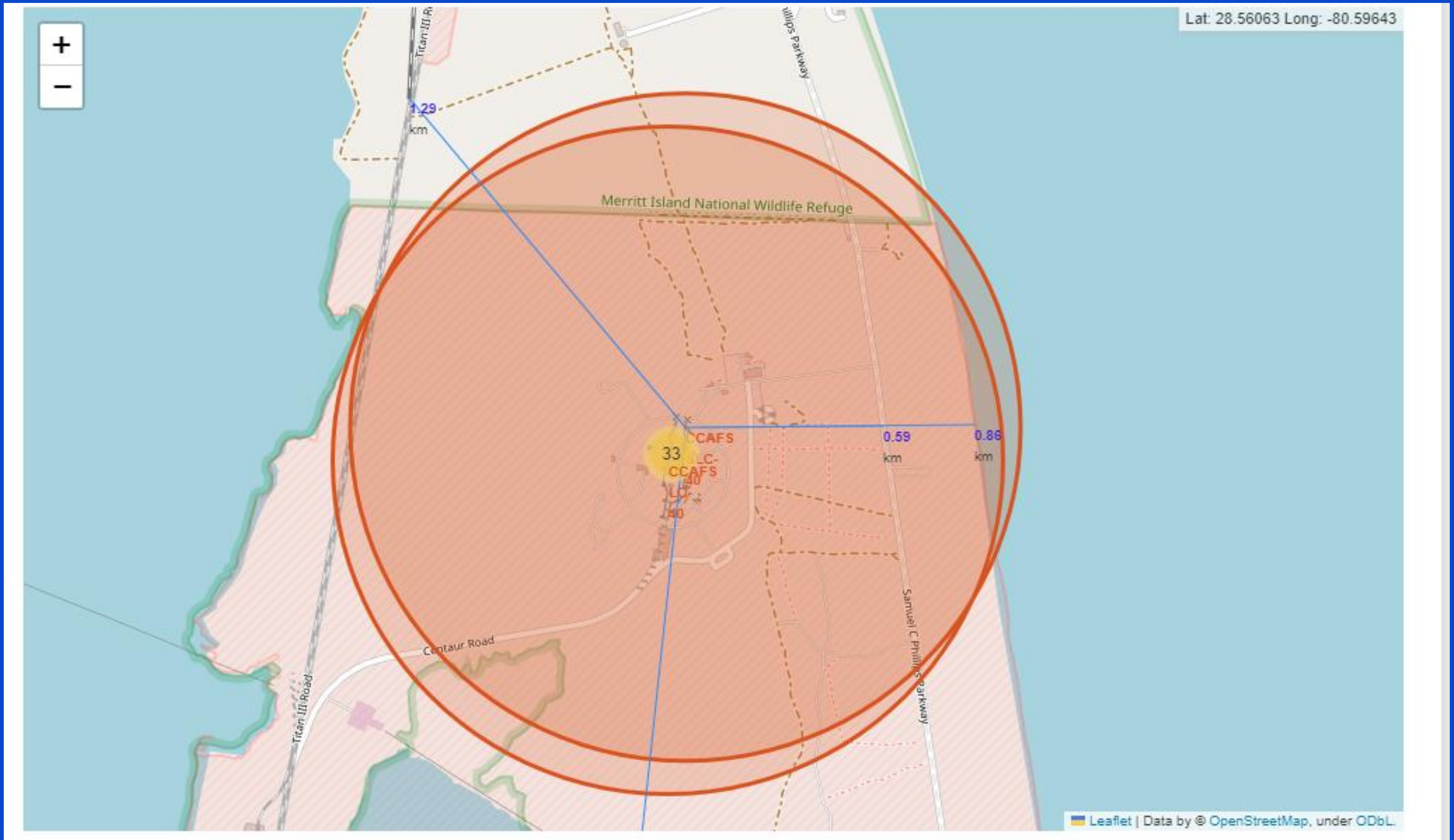
Examining launching sites

- On each launching site circle we can observe successful and unsuccessful landings



Specifics of launching sites

Launching sites are very close to means of transport, on land transportation can potentially generate high costs. On the other hand launching sites are not close to cities probably to avoid affecting them in any way

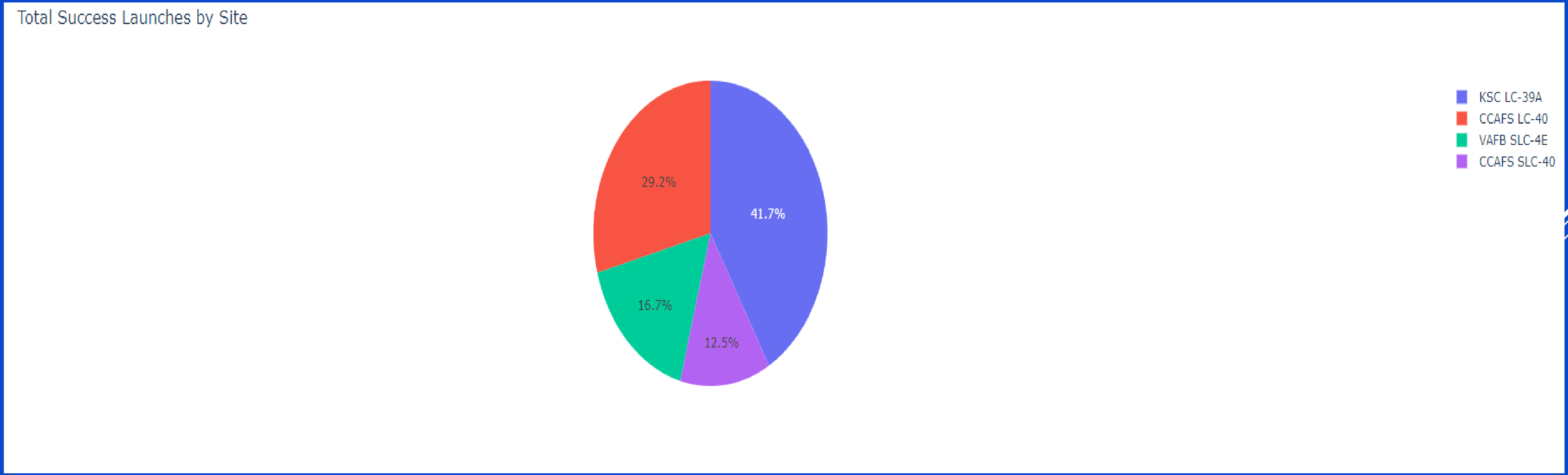


Section 4

Build a Dashboard with Plotly Dash

Total success launches by launching site

- Most successful launches were achieved at KSC LC-39A launching site , followed by CCAFS LC-40, VAFB SLC-4E and CCAFS SLC-40

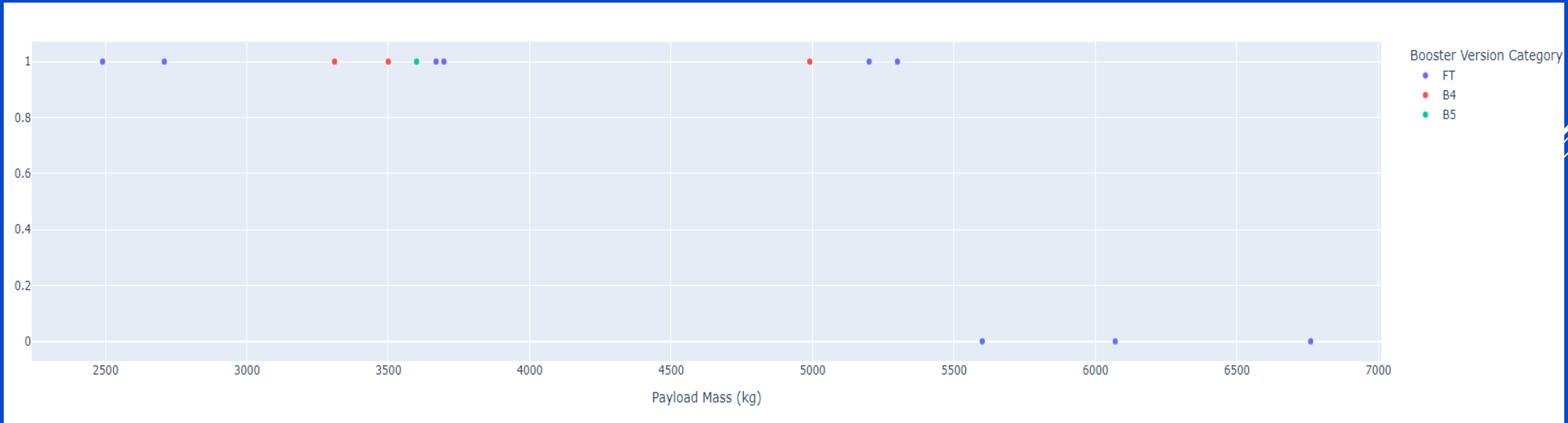


Total success launches at KSLC-39A

KSLC-39A had the success rate of 76.9%

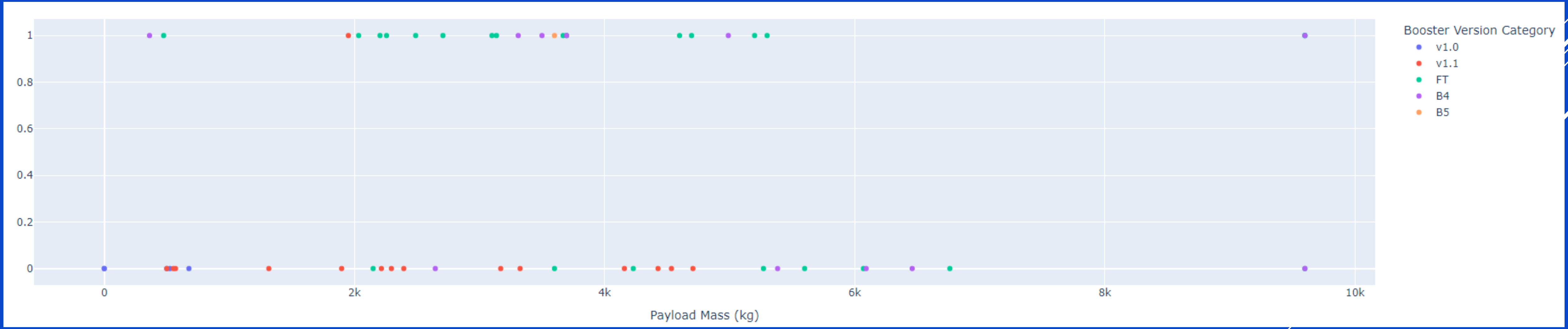
KSC LC-39A site success rate vs payload mass

Launches from this launching site were relatively low on payload mass, launches above 5.5 tones ended up as failures



Success rate based on payload mass and booster version

- FT booster had the highest success rate
- Rockets with payload mass above 6 tones generally ended up failing

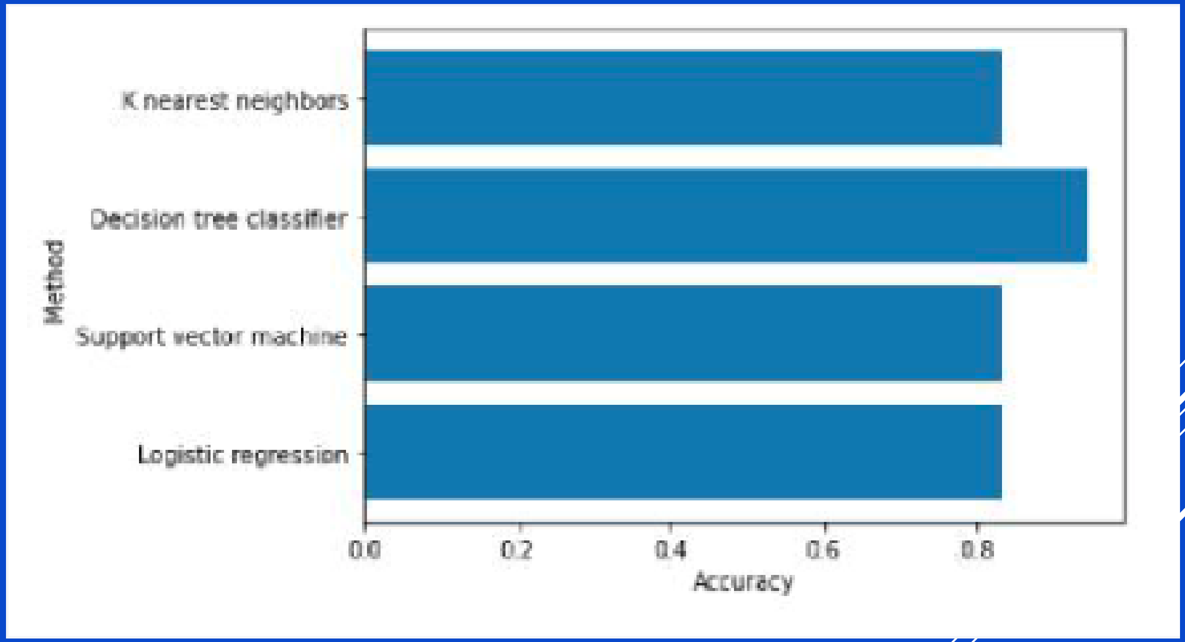


Section 5

Predictive Analysis (Classification)

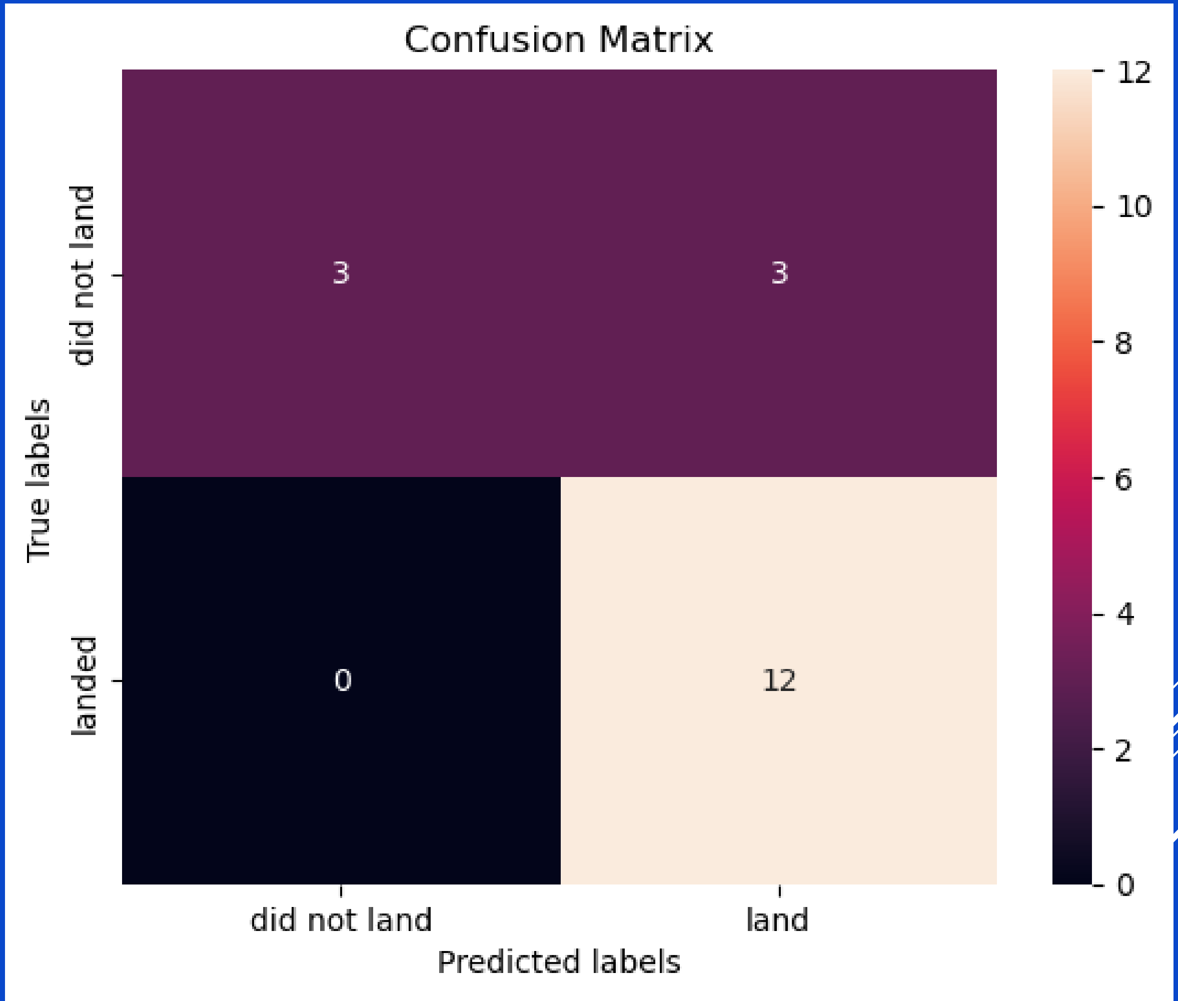
Classification Accuracy

- The highest accuracy had the Decision tree classifier



Confusion Matrix

- Confusion matrix had 3 false positives and 0 false negatives, which can mean that it may not be accurate enough for use due to small sample size



Conclusions

Launching site with the best chance of success is KSLC-39A

There are several orbits with high performance that can be chosen to the task

The best performing booster is FT booster

According to decision tree model the falcon 9 have a 66% chance of success

The model might not be highly accurate due to low sample size and high amount of false negatives therefore further analysis is recommended

Appendix

- https://github.com/borp879/ibm-data-science-course

Thank you!