IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mathias Dalmás
09-26-2023

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data collection with API
  - Data collection with web scrapping
  - Data Wrangling
  - Exploratory analysis with SQL
  - Exploratory analysis with Data Visualization tools
  - Interactive Visual Analytics with Folium and Dash
  - Machine Learning prediction
- **Summary of all results**
  - Exploratory Data analysis results
  - Interactive analytics in screenshots
  - Predictivey analytics results

Introduction

- Project background and context
  - Space X 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 space X 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 space X for a rocket launch. In this lab, you will create a machine learning pipeline to predict if the first stage will land given the data from the preceding labs.
- Problems you want to find answers
  - What factors determine if the rocket land successfully
  - The interaction amongst various features that determine the success rate of a successful landing
  - What operating conditions needs to be in place to ensure a successful landing program

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API → https://api.spacexdata.com/v4/rockets/
  - Webscraping → https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Perform data wrangling
  - Collect data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  - Data was collected, on this step was normalized and divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

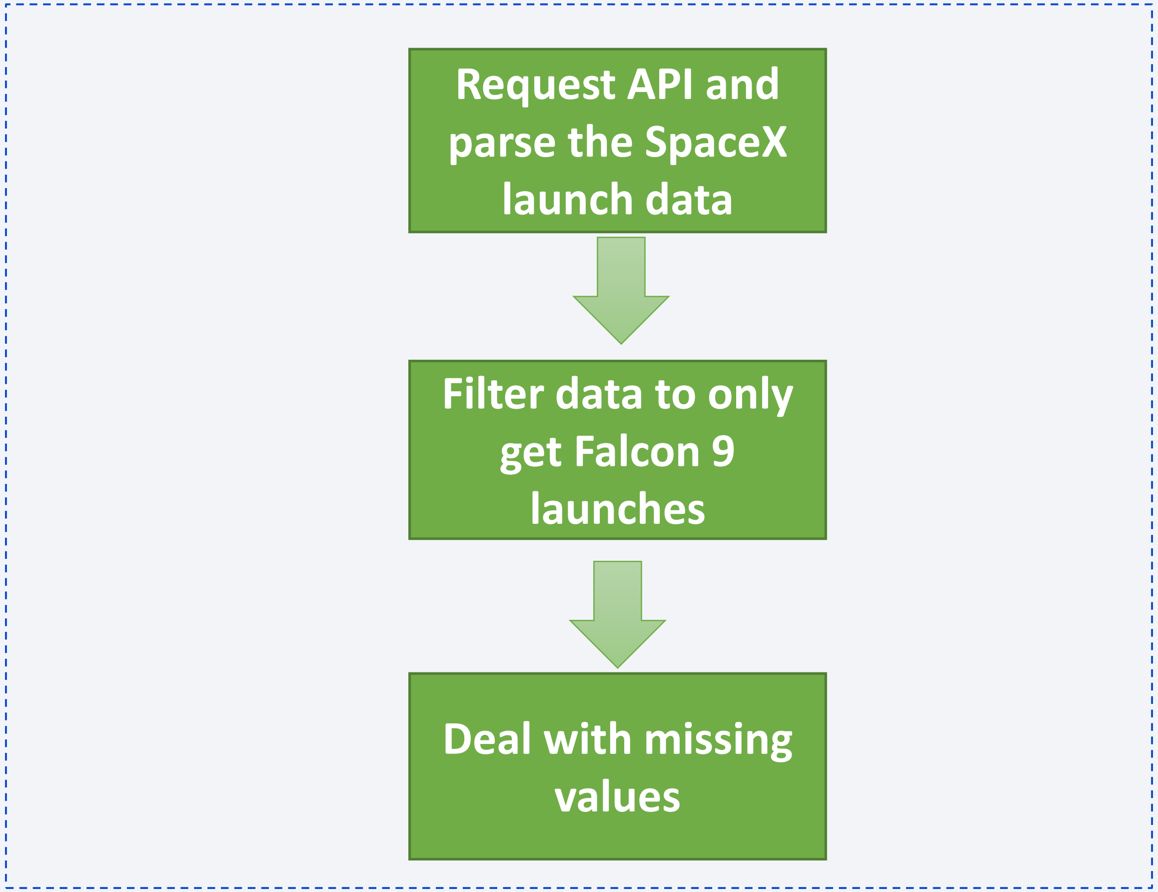
Data Collection

- Describe how data sets were collected.
- Data sets were collected using SpaceX API (https://api.spacexdata.com/v4/rockets/) and Webscrapping on web (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

Data Collection – SpaceX API

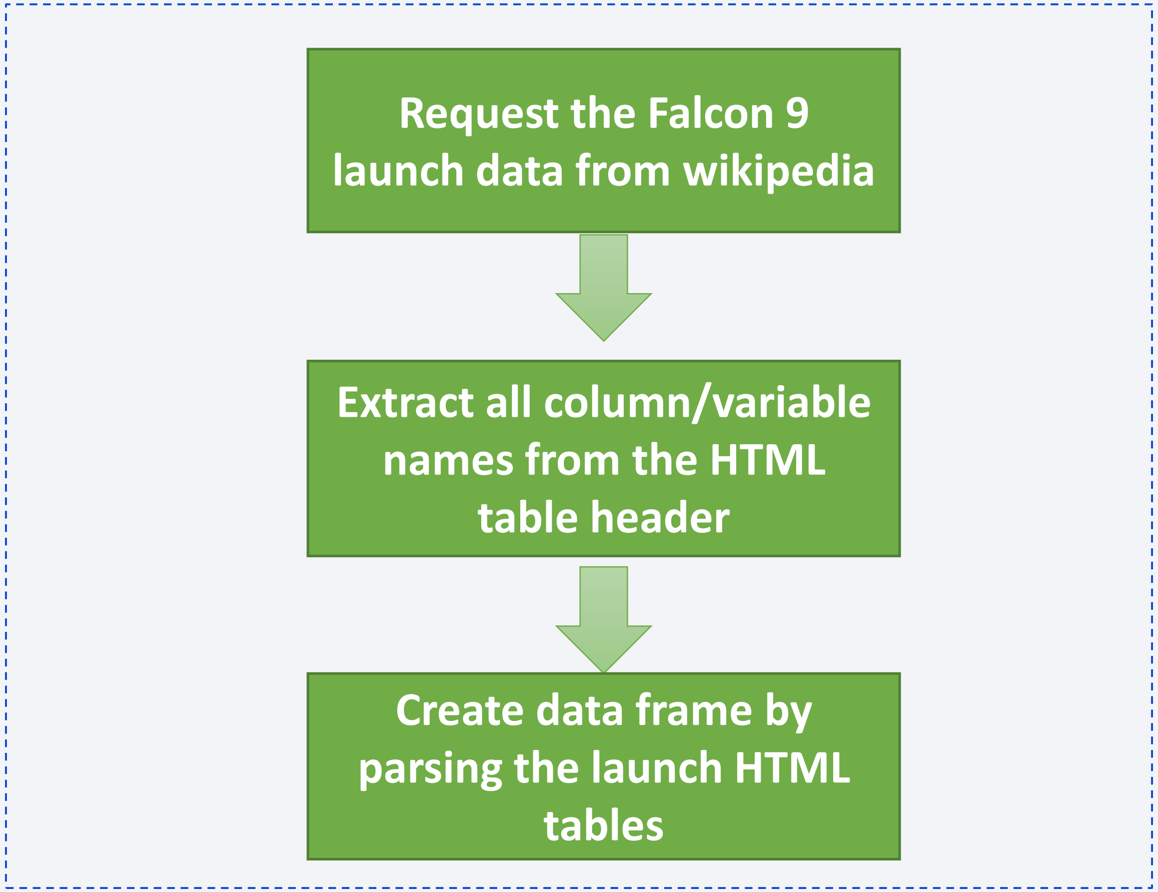
- SpaceX offers a public API from where data can be obtained and used.
- This API was used according to the flowchart beside.
- Once information is obtain it is move to a .csv file.

- GitHub URL: https://github.com/dalmasm/applied_ds_capstone/blob/Data-Science/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Data from SpacesX launches can be get from Wikipedia
- Data was downloaded regading flowchart beside.
- Github URL: https://github.com/dalmasm/applied_ds_capstone/blob/Data-Science/jupyter-labs-webscraping.ipynb



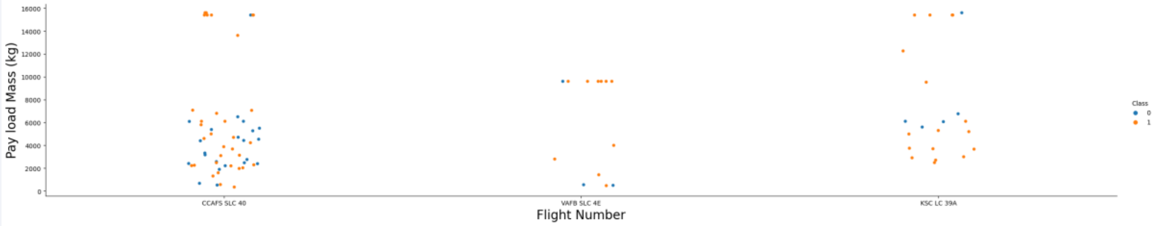
Data Wrangling

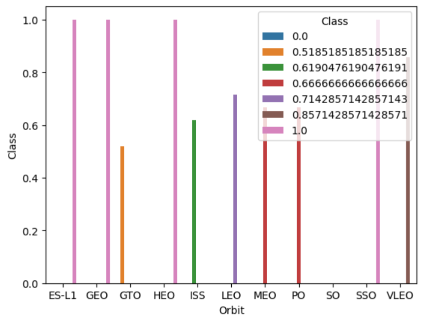
- Initialize some Exploratory data analysis (EDA) was performed on dataset
- Then, the summary launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated
- Finally, the landing outcome label was created for Outcome column

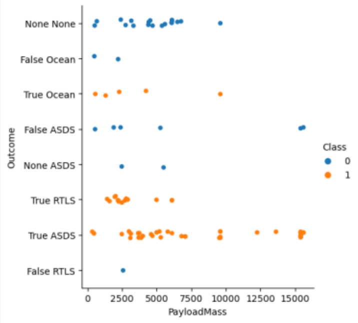
- Github URL: https://github.com/dalmasm/applied_ds_capstone/blob/Data-Science/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Scatterplots and barplots were used to visualize relationship pair of features:
  - Payload mass vs flight number
  - Launch site vs flight number
  - Launch site vs payload mass
  - Orbit and flight number
  - Payload and orbit







Github URL: https://github.com/dalmasm/applied_ds_capstone/blob/Data-Science/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- The following querys were done using SQL:
  - Names of unique launch sites in the space mission
  - 5 records where launch site starts with “CCA”
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - The date when the first successful landing outcome in ground pad was achieved.
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass. Use a subquery
  - List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

Github URL: https://github.com/dalmasm/applied_ds_capstone/blob/Data-Science/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Map
- Markers indicate points like launch site
- Circles indicate highlighted areas around specific coordinates
- Marker clusters indicate groups of event in each coordinate, like launches in launch site.
- Lines are use to indicate distances between two coordinates.

Github URL:
https://github.com/dalmasm/applied_ds_capstone/blob/Data-Science/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Interactive chart using Plotly Dash
- Pie charts showing total launches on certain sites
- Scatter plot showing the relationship with Outcome and Payload mass (Kg) for the different booster version

Github URL: https://github.com/dalmasm/applied_ds_capstone/blob/Data-Science/SpaceX%20Plotly.py

Predictive Analysis (Classification)

- Data was loaded using pandas and numpy, then transformed and splitter into training and testing.
- Different machine learning models were built and tune different hyperparameters using GridSearchCV.
- Accuracy was used as metric for model evaluated. Models were improved using feature engineering and algorithm tuning.
- Best performing classification model was found.

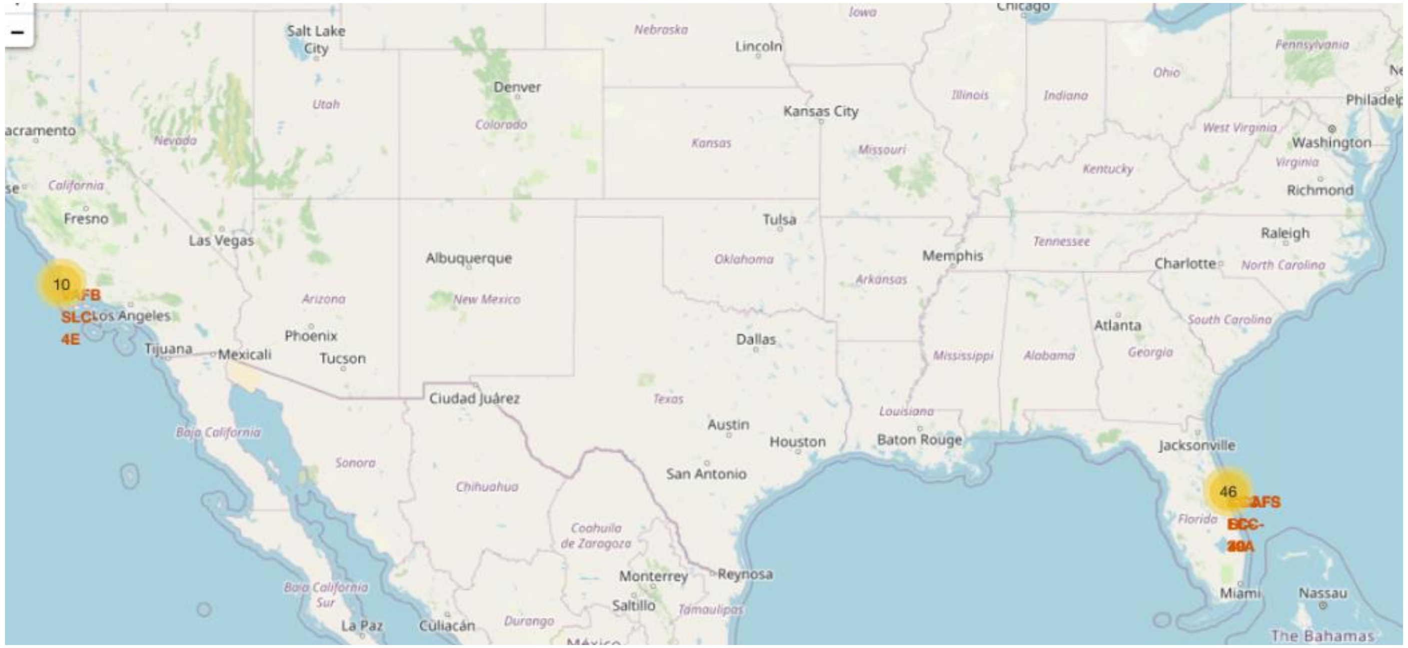
GitHub URL: https://github.com/dalmasm/applied_ds_capstone/blob/Data-Science/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Space X uses 4 differents launch sites.
- The first launches were done to Space X itself and NASA
- The average payload of F9 v1.1 booster is 2,928Kg
- The first success landing outcome happened in 2015
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above avg.
- Almost 100% of mission outcomes were successful.
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
- The number of landing outcomes become better as year passed.

Results

- Using interactive analytics was possible to identify that launch sites used to be in safety places, near sea. And with good logistic infrastructure nearby.
- On east cost occurred more launches than in west coast (46 vs 10).

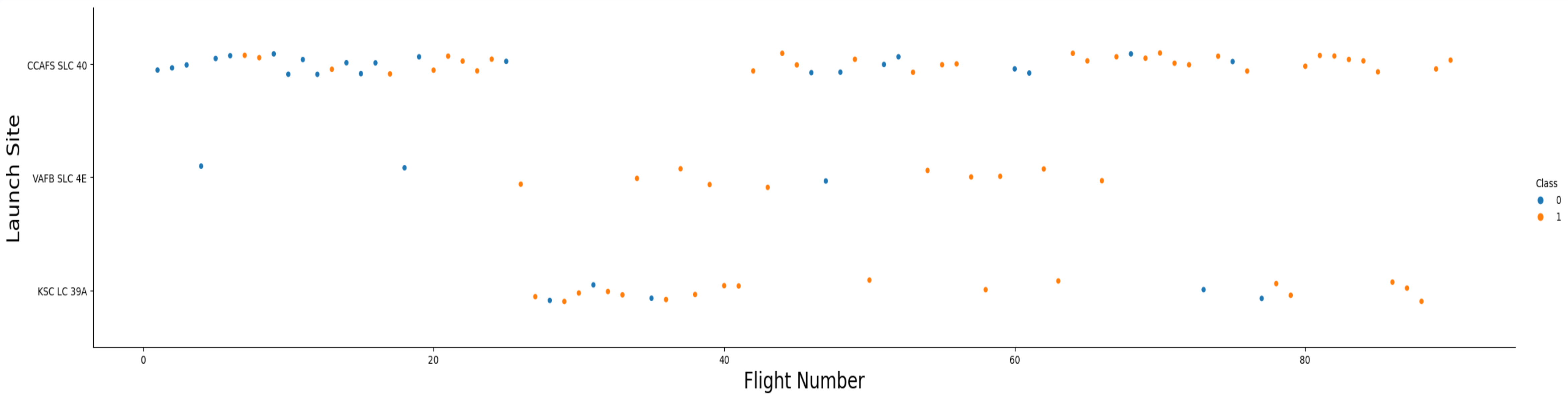


Section 2

Insights drawn from EDA

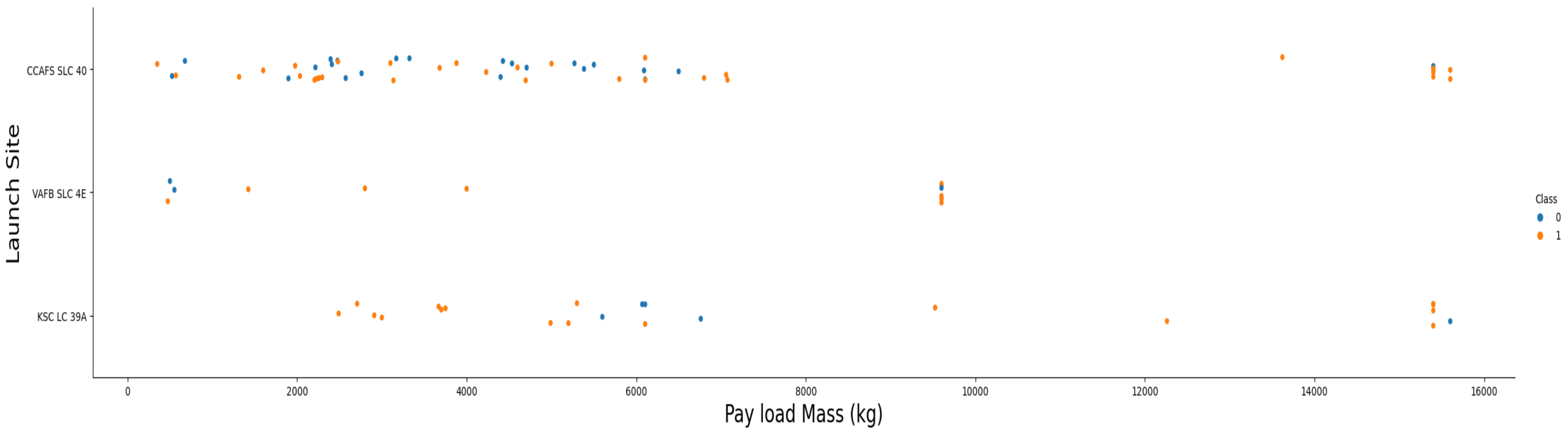
Flight Number vs. Launch Site

- Success rate increase as flights increase.
- Biggest number on launches occurred in CCADS SLC 40 site.



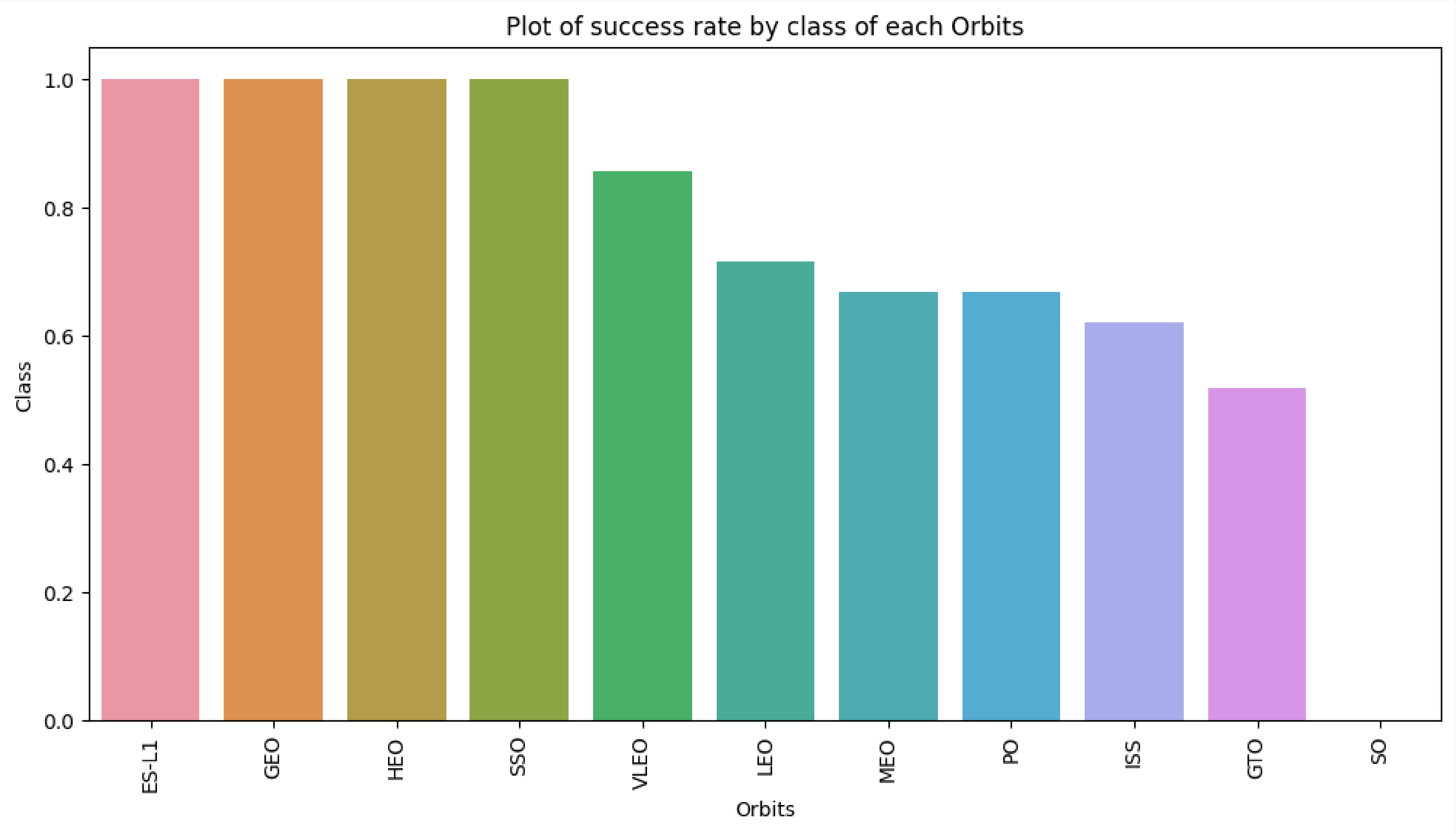
Payload vs. Launch Site

- The greater the Payload Mass the higher the successful rate on SLC40 site
- On VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).



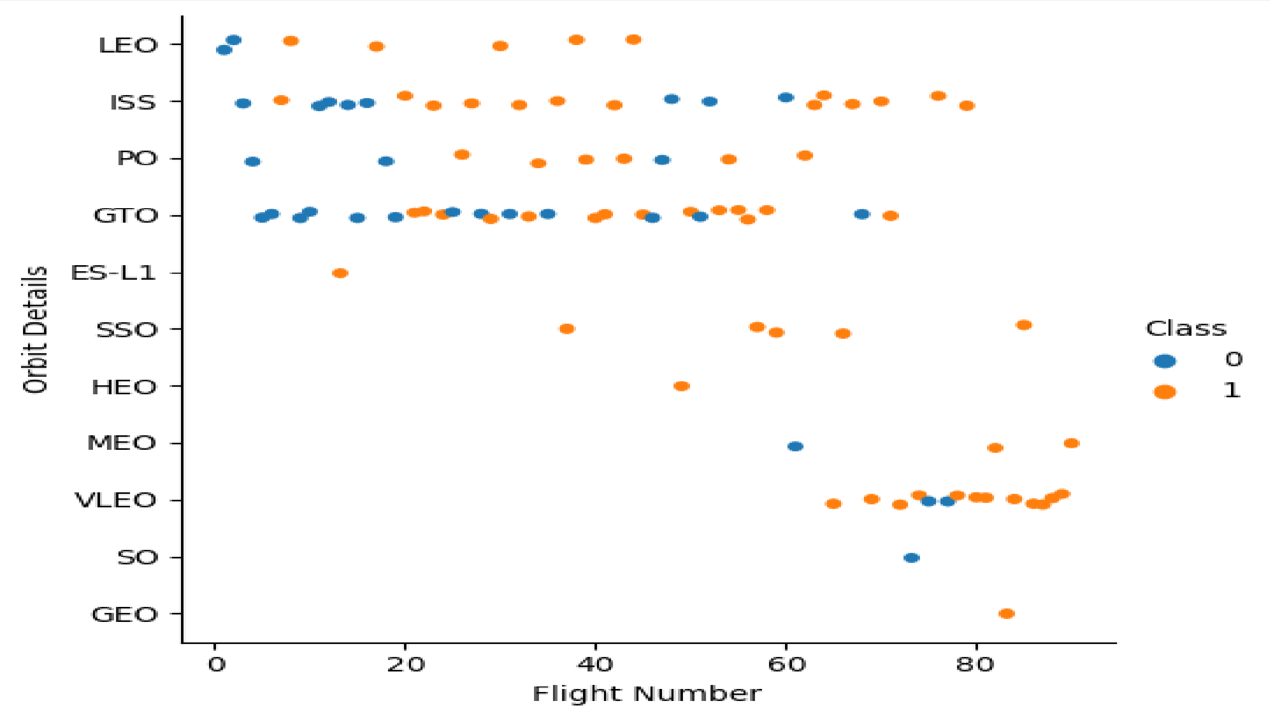
Success Rate vs. Orbit Type

- On the plot you can see sites ordered from higher success rato to lower.

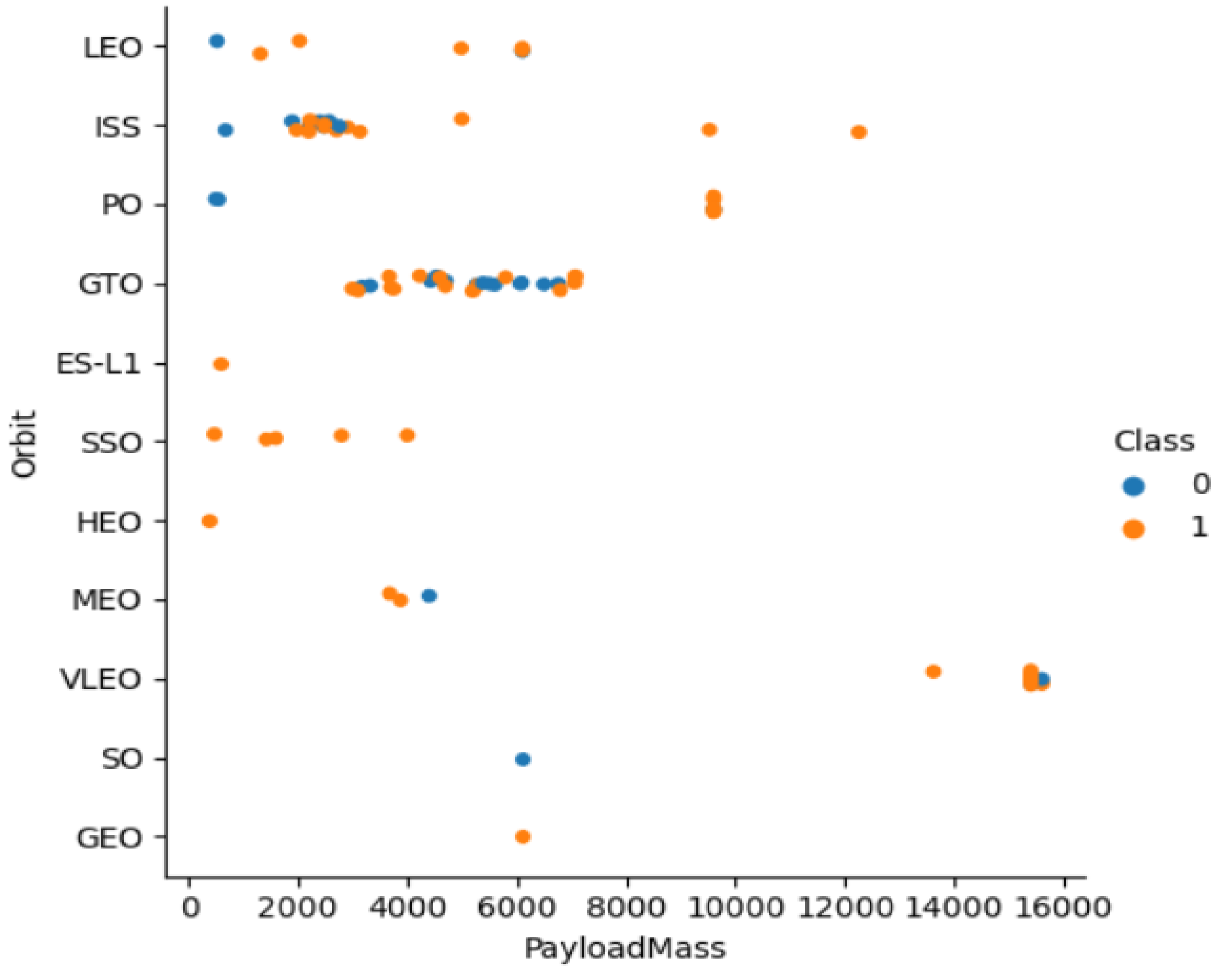


Flight Number vs. Orbit Type

In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



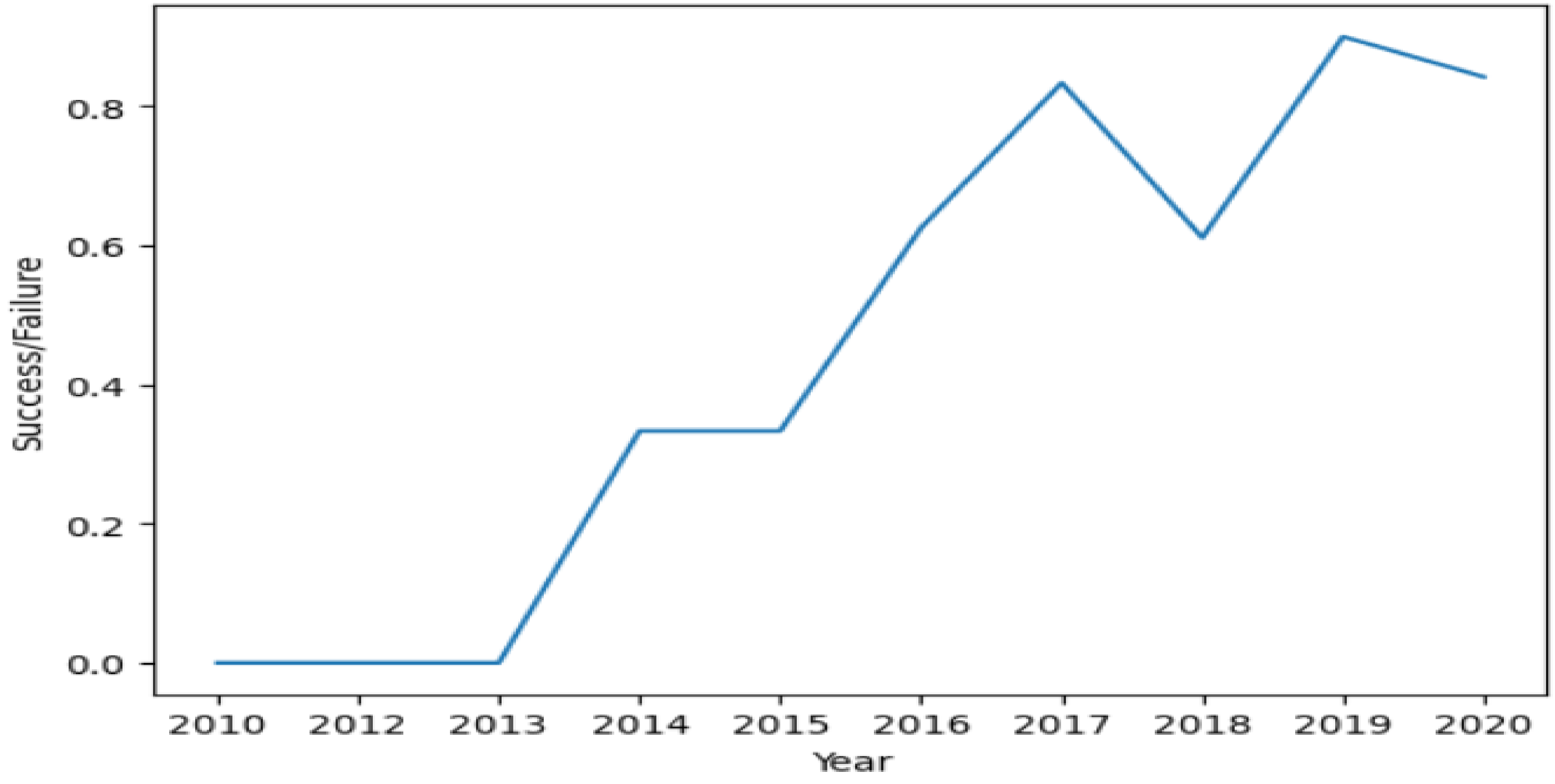
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Using “Distinct” on SQL you can find that there are 4 unique launch sites:

```
%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'

- Next you can see 5 records that begin with ‘CCA’. As a special mension all outcomes are “Success”. It is found using the command “LIMIT” on SQL

```
%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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
2010-08-12	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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- Next you can find that the Total payload mass carried 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

Average Payload Mass by F9 v1.1

- If we filter booster version by ‘F9 v1.1’ we get the average payload of 2928.4Kg

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

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- First successful ground landing was on December 22 of 2015

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

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We have 4 successful drone ship landing with payload btwn 4000 and 6000 Kg. Two of them in 2016 and the other two in 2017.

```
%sql SELECT * FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lan
2016-06-05	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	
2017-11-10	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	

Total Number of Successful and Failure Mission Outcomes

- We get a total of 100 success and only 1 failure.

List the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) from SPACEXTBL GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- List of the booster versions which carried the maximum payload mass:

```
%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

- In 2015 we had 2 failures on B1012 and B1015 ob F9 v1.1 booster versions:

```
%sql SELECT substr(Date,4,2) as month,Date, BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome FROM SPACEXTBL WHERE substr(Date,1,4)='2015' and L
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
5-	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
5-	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- 32 landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

```
%sql SELECT Landing_Outcome, COUNT(*) as count_outcomes from SPACEXTBL WHERE DATE Between '2010-06-04' and '2017-03-20' order by count_outco
```

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

Landing_Outcome	count_outcomes
Failure (parachute)	32

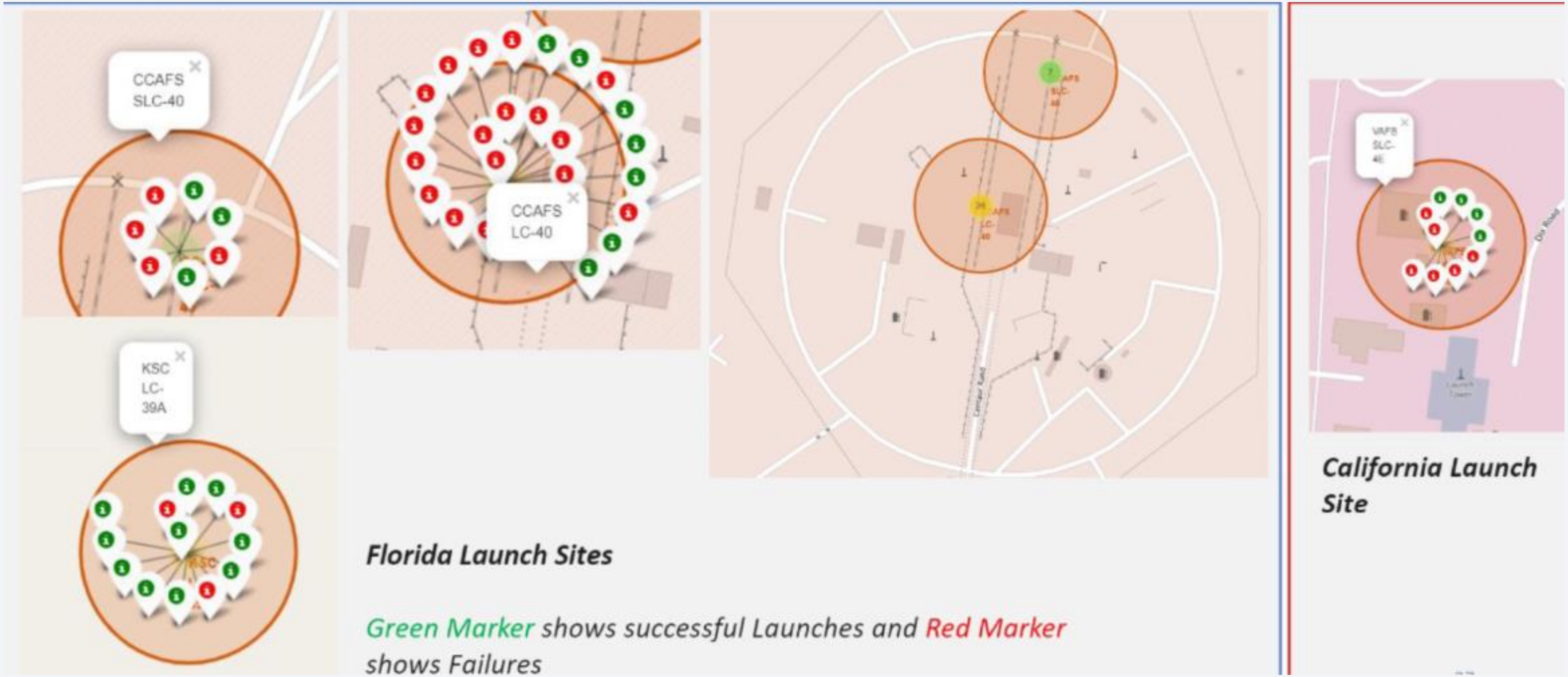
Section 3

Launch Sites Proximities Analysis

Global map with launch sites

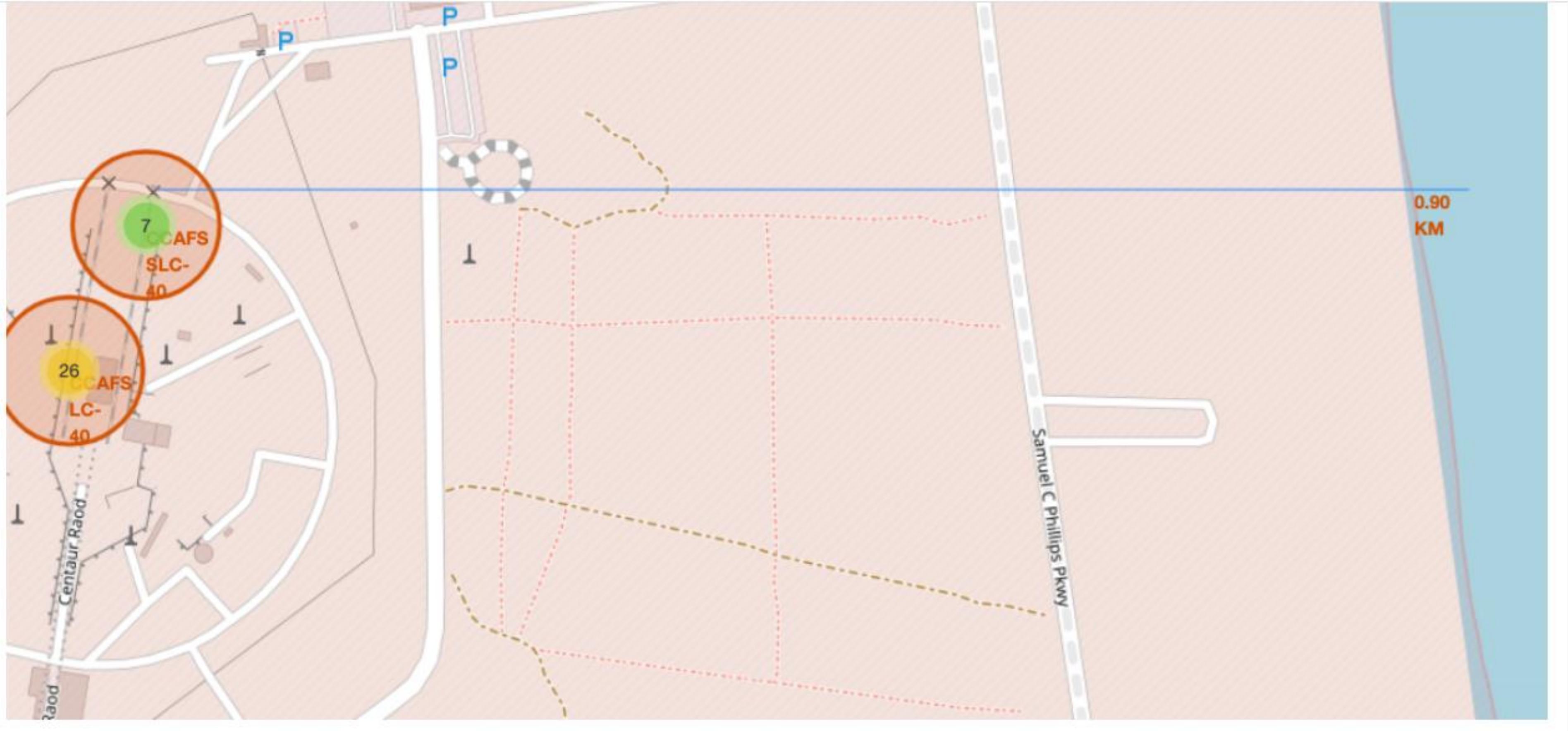
- Looking at the global map we can see that launch sites are on USA coast (west & east):

Markers showing launch sites



Distance of launch site to sea

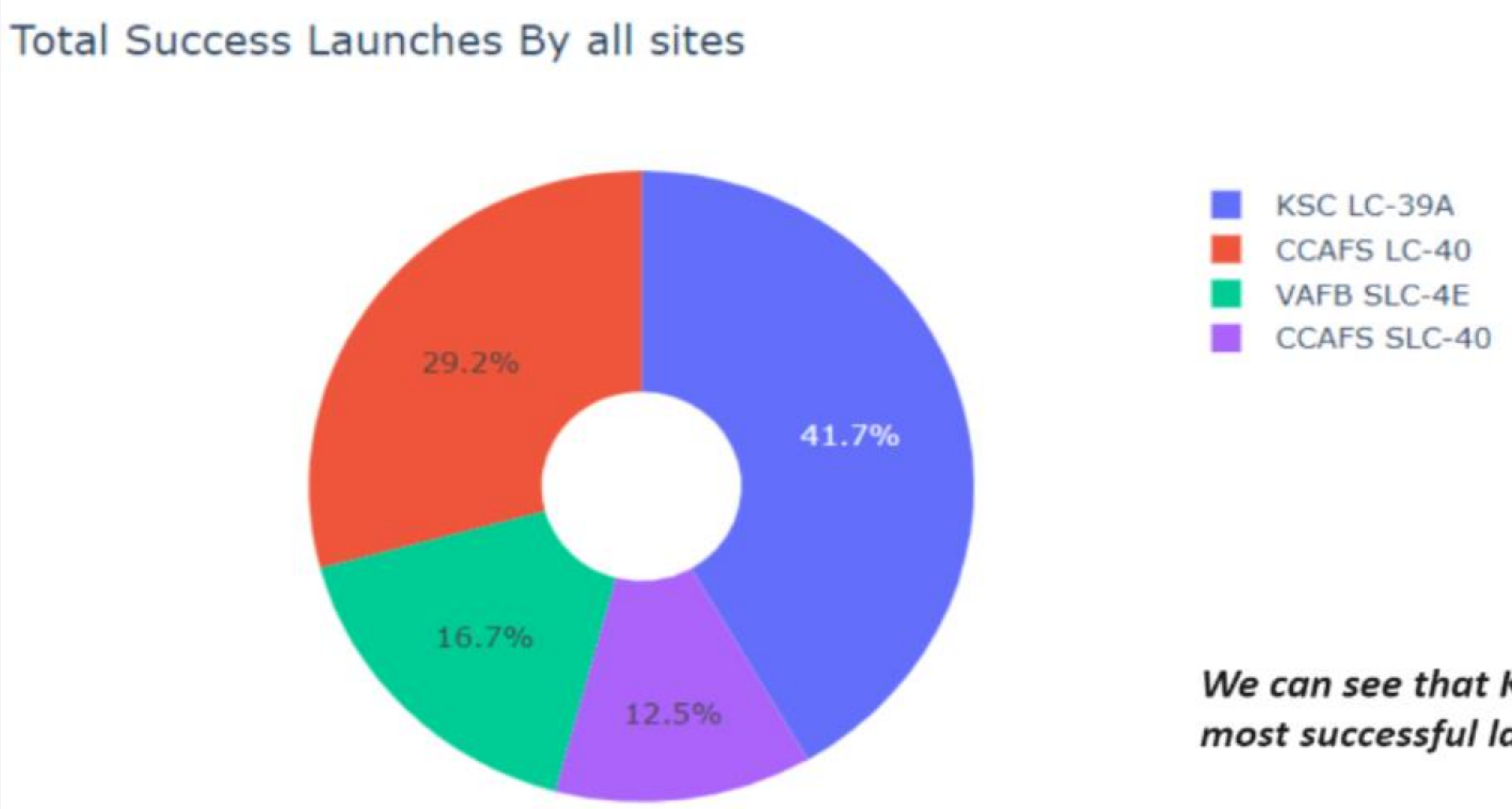
- Here you can see a distance from launch point to sea



Section 4

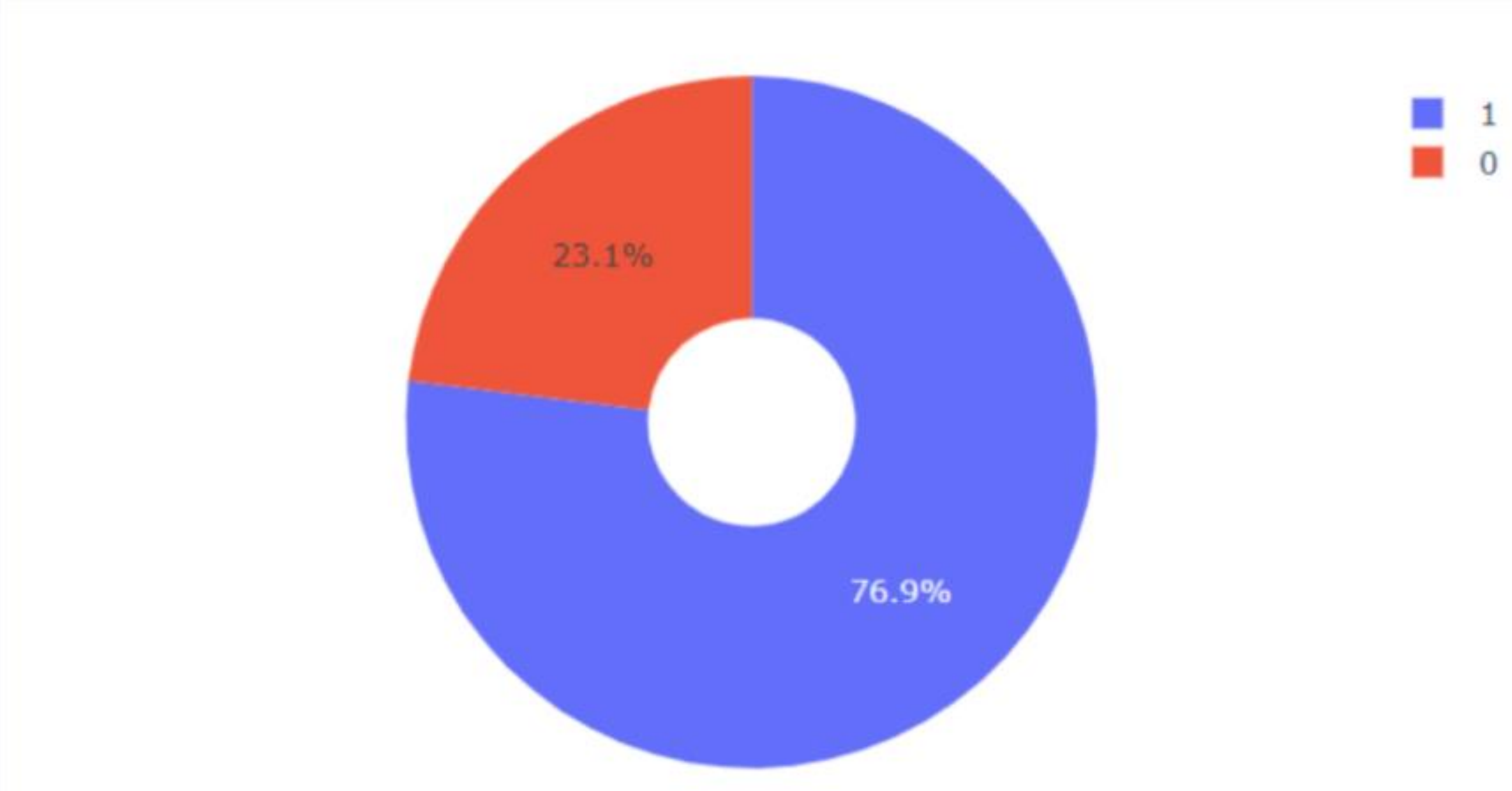
Build a Dashboard with Plotly Dash

Pie chart showing success rate per site



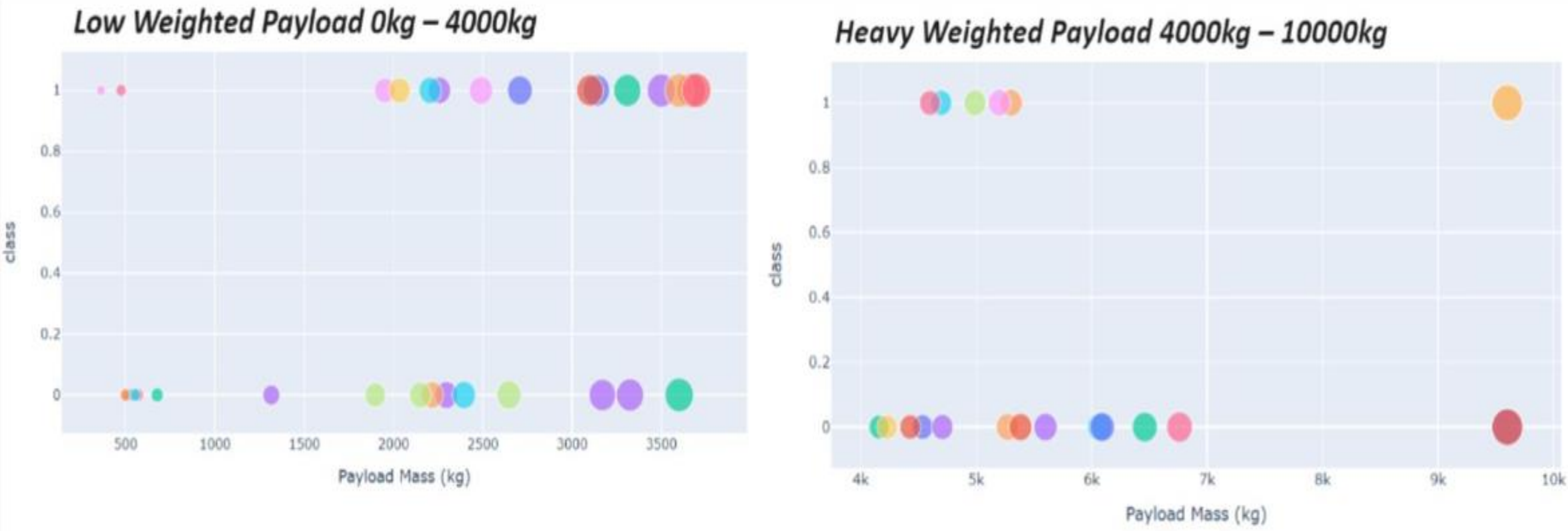
We can see that KSC LC-39A had the most successful launches from all the sites

KSC LC-39A success rate



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Success rate on Low & High payload



We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

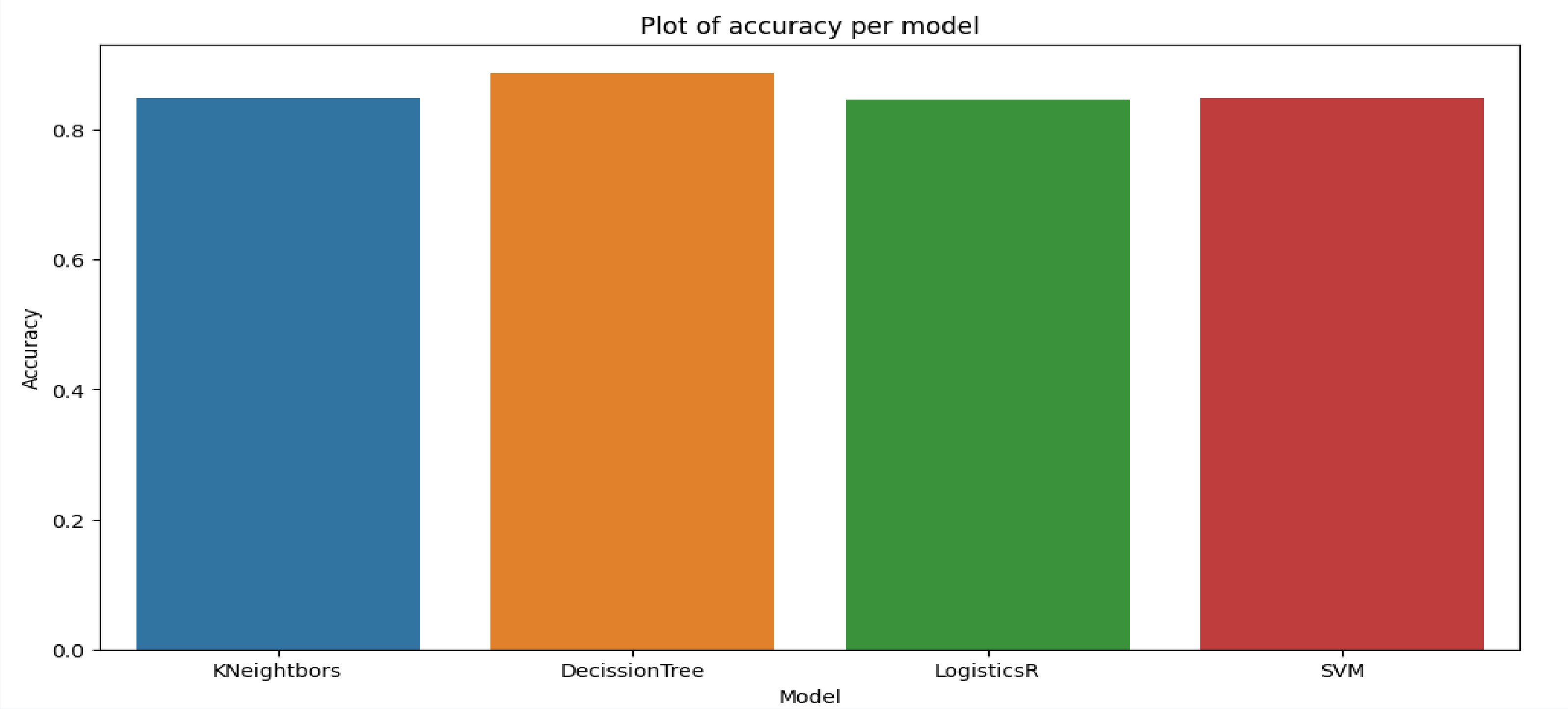
Section 5

Predictive Analysis (Classification)

Classification Accuracy

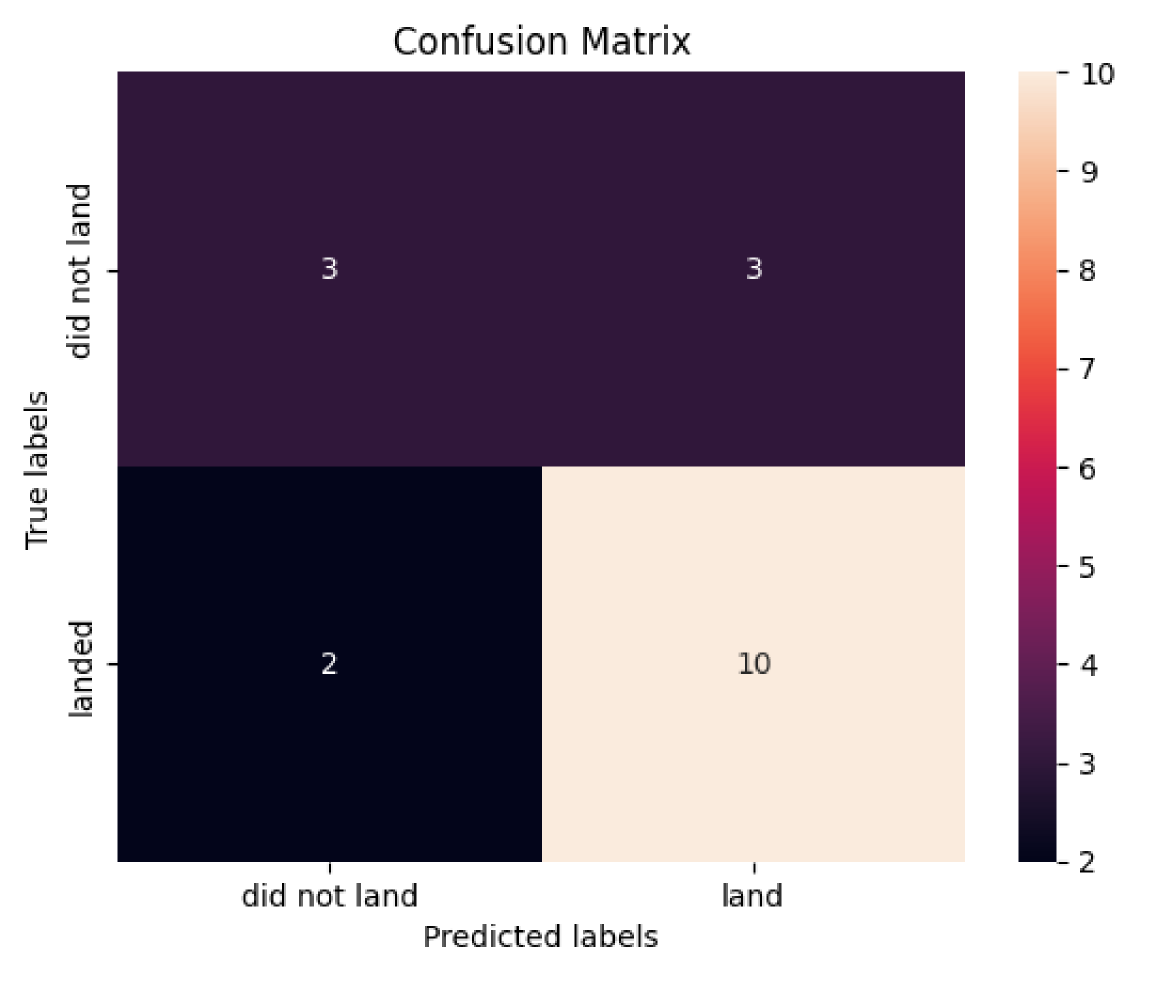
- As you can se in the chart below, the model with best accuracy is Decission Tree with 0.8875.
- Best parameter found are:

Best Parameters {'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 2, 'splitter': 'best'}



Confusion Matrix

- On confusion Matrix of Dessicion Tree methond we can find that we have 10 successful predicted “land” and 3 “not land”, On the other hand we have 5 unsuccessful predictions.



Conclusions

We can conclude that:

- Success rate increase as amount of flights amount increase
- Launch success rate increase as years passed from 2013 to 2020
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate
- KSC LC-39A had the most successful launches rate
- Decision Tree classifier is the best ML algorithm for prediction in this case but all methods have high accuracy and similar.

Thank you!