Winning Space Race with Data Science

Kateryna MELESHENKO

August 2022

Outline

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

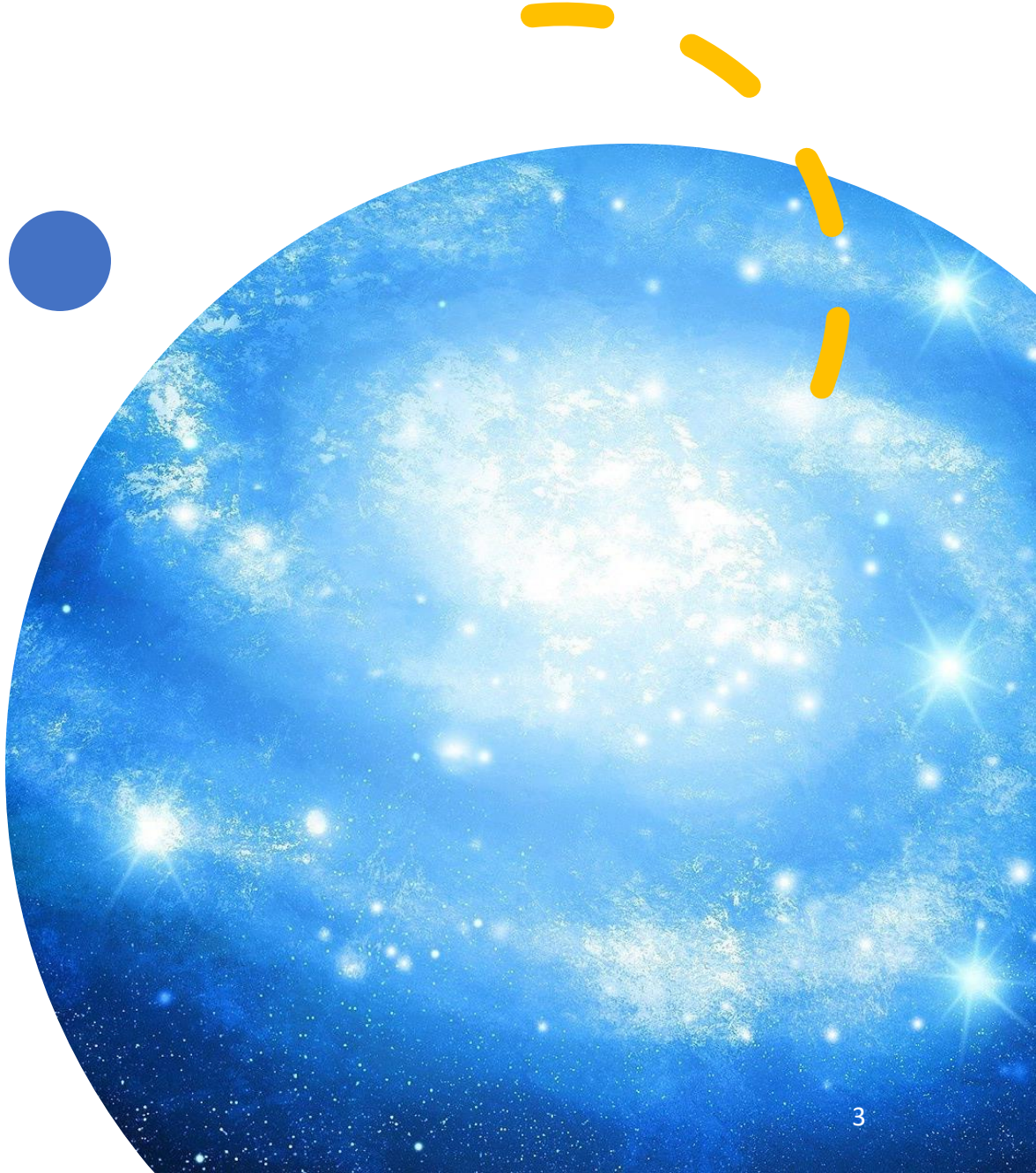
Executive Summary

Summary of methodologies

- Data collecting & data wrangling;
- Exploratary data analysis;
- Interactive visual analytics;
- Machine learning prediction.

Summary of all results

- A machine learning pipeline was built to predict if the first stage of the Falcon 9 lands successfully



Introduction

Project background

SpaceX, founded by Billionaire industrialist Allon Musk, 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. This stage does most of the work and is much larger than the second stage. However, sometimes it does not land. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

Problems we want to find answers

To determine if the first stage will land, in order to be able to determine the cost of a launch.


Section 1
Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - Data was collected from an API and Wiki page through the web scraping method.
- Perform data wrangling:
  - Outcomes were converted into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

- Perform predictive analysis using classification models:
  - Preprocessing (allowed us to standardize our data);
  - Train_test_split (allowed us to split our data into training and testing data);
  - Model training;
  - Grid Search performing (allowed us to find the hyperparameters that allow a given algorithm to perform best);
  - Determination of the model with the best accuracy (Using the best hyperparameter values and training data
  - Testing Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors.
  - Output the confusion matrix.

Data Collection

SPACEX REST API

WIKI PAGES (SCRAPING)

Data Collection – SpaceX API

Now let's start requesting rocket launch data from SpaceX API with the following URL:

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

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

SpaceX launch data were gathered from the SpaceX REST API and gave us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

GitHub URL of the completed SpaceX API calls

Data Collection – Scraping

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9
data = requests.get(static_url).text
soup = BeautifulSoup(data)
```

```
html_tables = soup.find_all('table')
first_launch_table = html_tables[2]
print(first_launch_table)
```

Falcon 9 Launch data were obtained from Wiki pages via web scraping. We used the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records. We parsed the data from those tables and converted them into a Pandas data frame for further visualization and analysis.

GitHub URL of the completed SpaceX web scraping

Data Wrangling

We created a landing outcome label from outcome column that represents the classification variable. If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully.

```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
bad_outcomes
```

```
{'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'}
```

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class = df['Outcome'].replace({'False Ocean': 0, 'False ASDS': 0, '
df['Outcome'] = df['Outcome'].astype(int)
df.info()
```

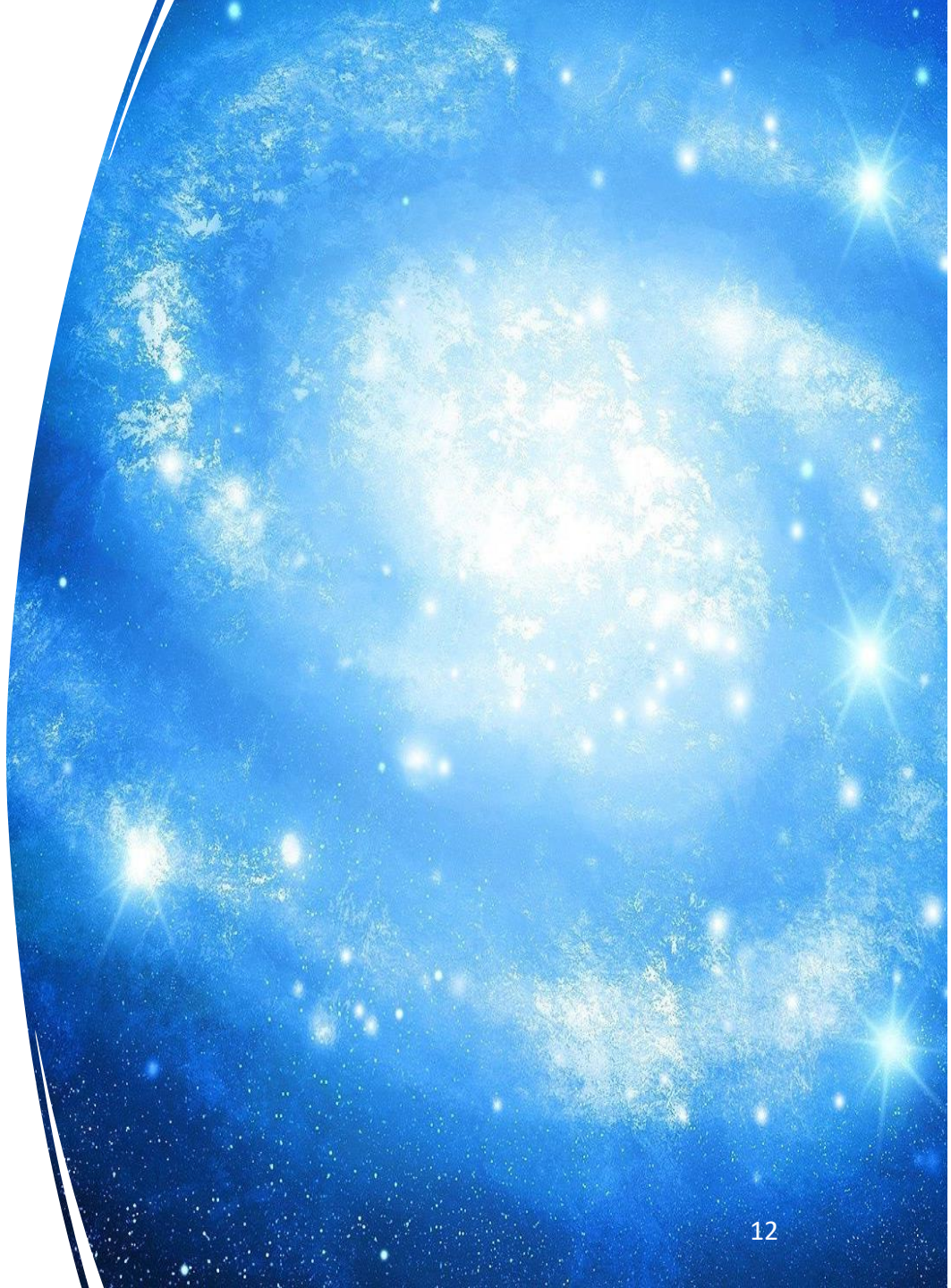
[GitHub URL of the completed SpaceX data wrangling]

EDA with Data Visualization

GitHub URL of the completed SpaceX exploratory data analysis with Pandas and Matplotlib

There were built:

- scattered plots to visualize the relationship between:
  - Flight Number and Launch Site;
  - Payload and Launch Site;
  - Flight Number and Orbit type;
  - Payload and Orbit type;
- a bar chart to visualize the relationship between success rate of each orbit type;
- a line plot to Visualize the launch success yearly trend



EDA with SQL

The next SQL queries were executed:

- Displayed the names of the unique launch sites in the space mission;
- Displayed 5 records where launch sites begin with the string 'CCA';
- Displayed the total payload mass carried by boosters launched by NASA (CRS);
- Display average payload mass carried by booster version F9 v1.1;
- Listed the date when the first successful landing outcome in ground pad was achieved;
- Listed the names of the boosters which have success in drone ship and have payload between 4000 and 6000;
- Listed the total number of successful and failure mission outcomes;
- Listed the names of the boosters' versions which have carried the maximum payload mass;
- Listed the records which will display the month names, failure landing outcomes in drone ship , booster versions, launch site for the months in year 2015;
- Ranked the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub URL of the completed SpaceX exploratory data analysis with SQL

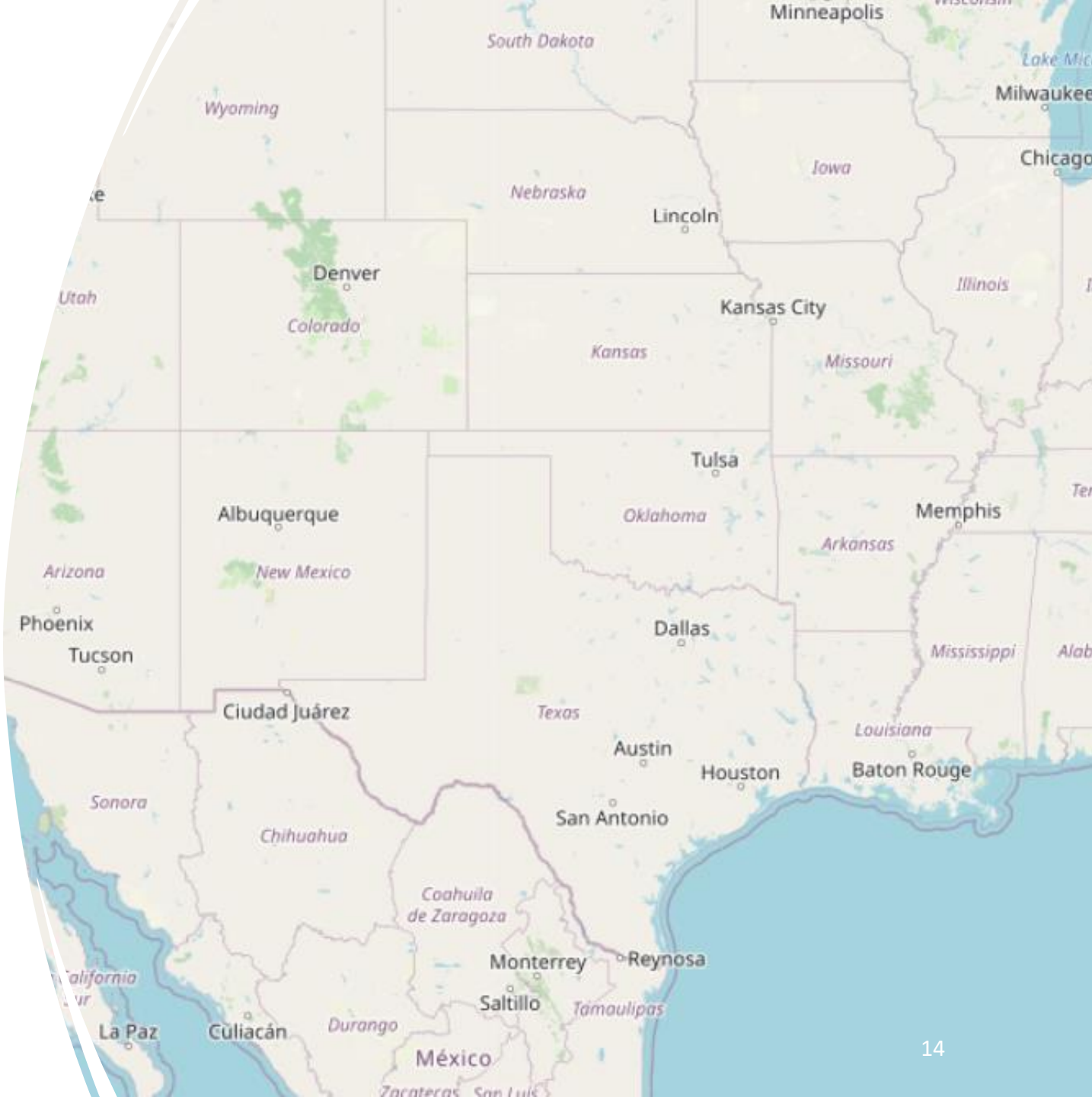
Build an Interactive Map with Folium

While building a map:

- all launch sites on a map were marked;
- the success/failed launches for each site on the map were marked;
- the distances between a launch site to its proximities was calculated

in order to perform more interactive visualization of some preliminary correlations between the launch site and success rate.

[GitHub URL of the completed SpaceX launch sites locations analysis with Folium]



Build a Dashboard with Plotly Dash

There were added to a dashboard:

- **a pie chart,** which let us define which site has the largest successful launches and which site has the highest launch success rate;
- **a scatter chart,** which let us define which payload range(s) has the highest launch success rate, which payload range(s) has the lowest launch success rate, which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate.

GitHub URL of the completed SpaceX Plotly Dash Visualization

Predictive Analysis (Classification)

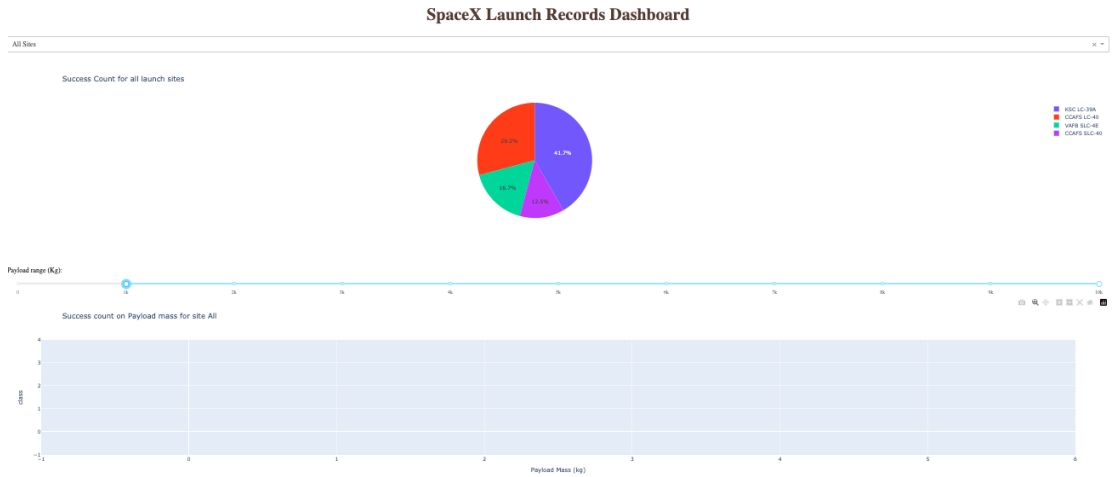
GitHub URL of the completed SpaceX Predictive Analysis

Results

During the exploratory data analysis:

- we have visualized the SpaceX launch dataset using matplotlib and seaborn and as a **result** we discovered some preliminary correlations between the launch site and success rates.
- we have visualized the SpaceX launch dataset using Plotly Dash and as a **result** could define which site has the largest successful launches and has the highest launch success rate; which payload range(s) has the highest launch success rate and the lowest launch success rate; which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate.

During the predictive analysis we used the best hyperparameter values and as a **result** we could determine the model with the best accuracy using the training data.

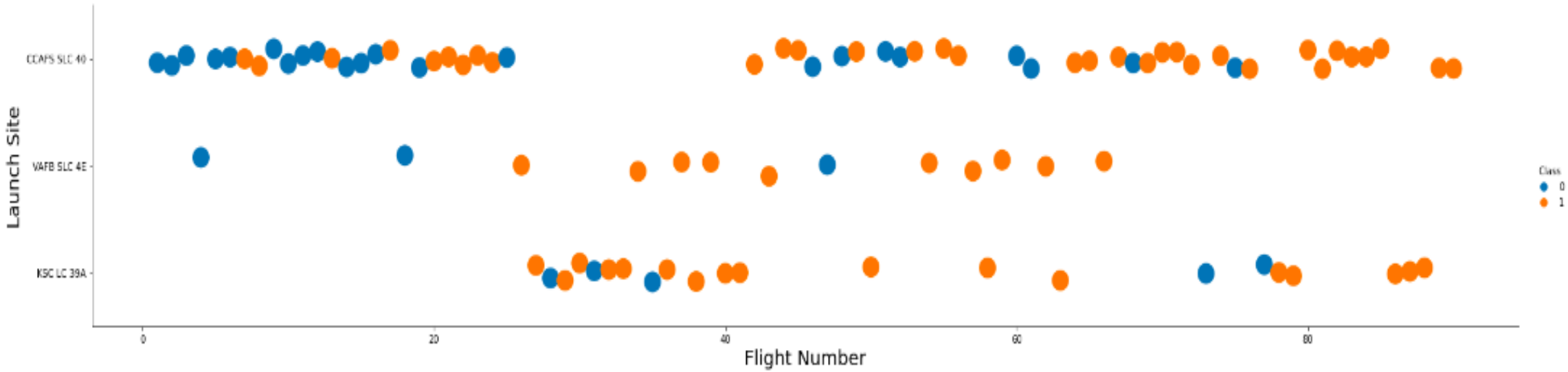



Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Different launch sites have different rates of success flights (VAFB SLC 4E has the highest rate of success flights).

Latest flight at CCAFS SLC 40 are more successful than the earliest ones.

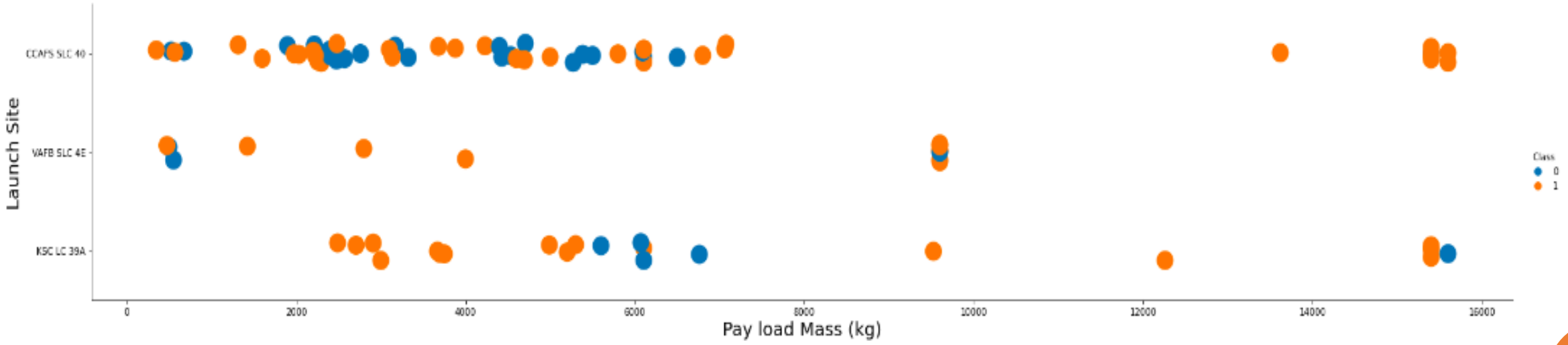


Payload vs. Launch Site

There are no rockets launched for heavy payload mass (greater than 10000 kg) at the VAFB-SLC launch site.

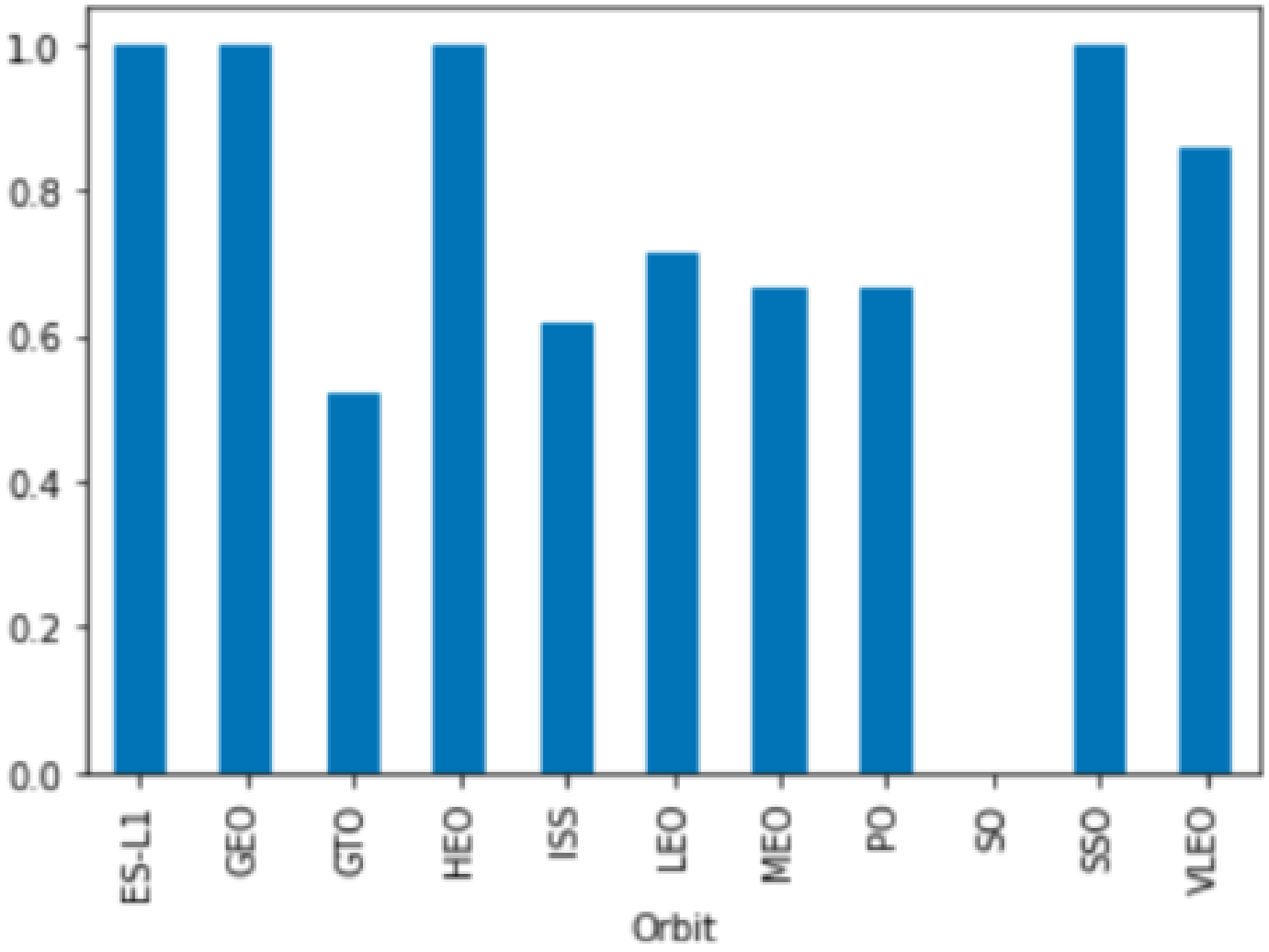
Almost all rockets with payload mass around 6000 kg have unsuccessful flights at the KSC LC 39A launch site.

All The rockets with payload more than 7000 kg have all successful flights at CCAFS SLC 40 launch site and a very high rate of successful flights at other launch sites.

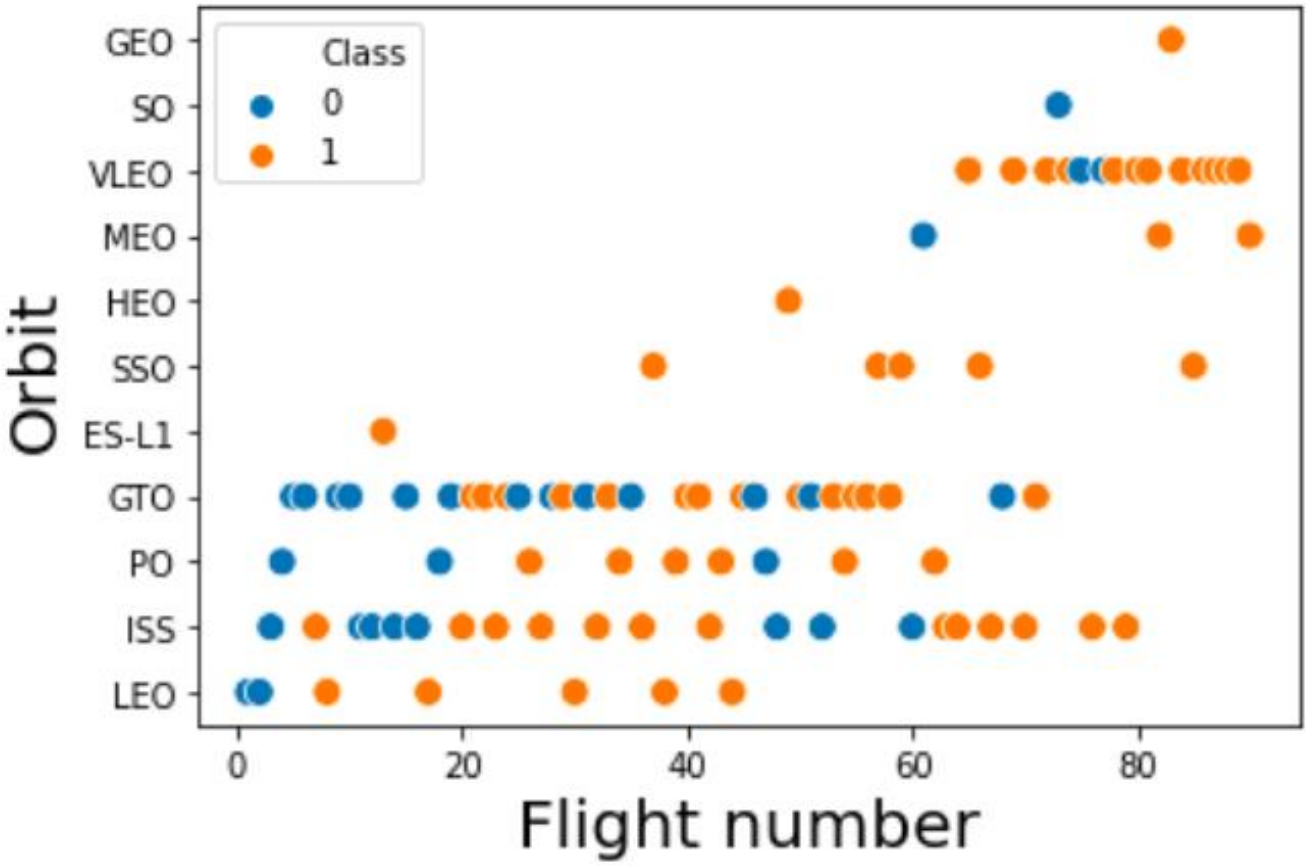


Success Rate vs. Orbit Type

- Only 4 out of 11 orbits (ES-L1, GEO, HEO, SSO) have 100% successful flights.
- SO Orbit has no successful flights at all.
- The average rate of successful flights to the other orbits is more than 60%.



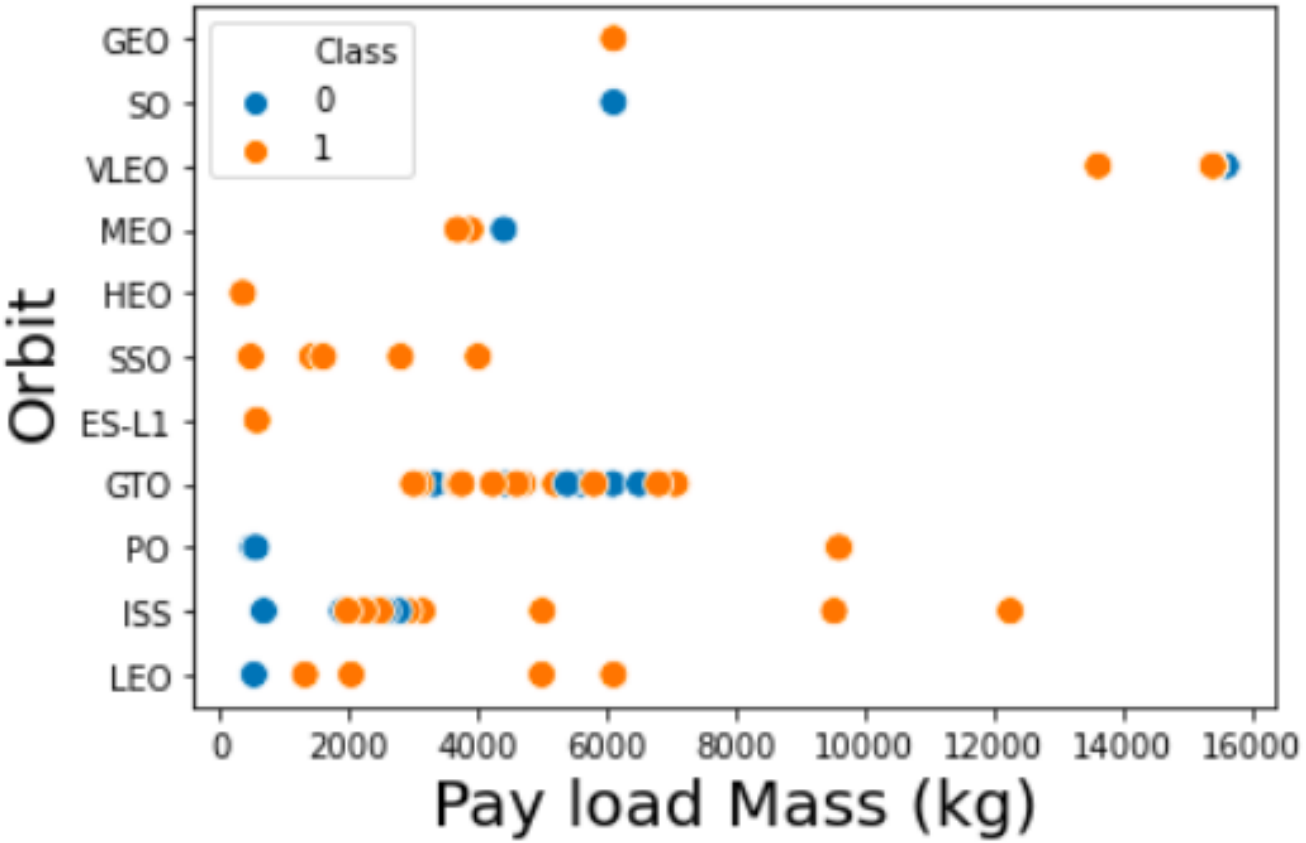
Flight Number vs. Orbit Type



- The latest flights were performed to the orbit VLEO and almost all of them are successful.

- There seems to be no relationship between flight number when in GTO orbit

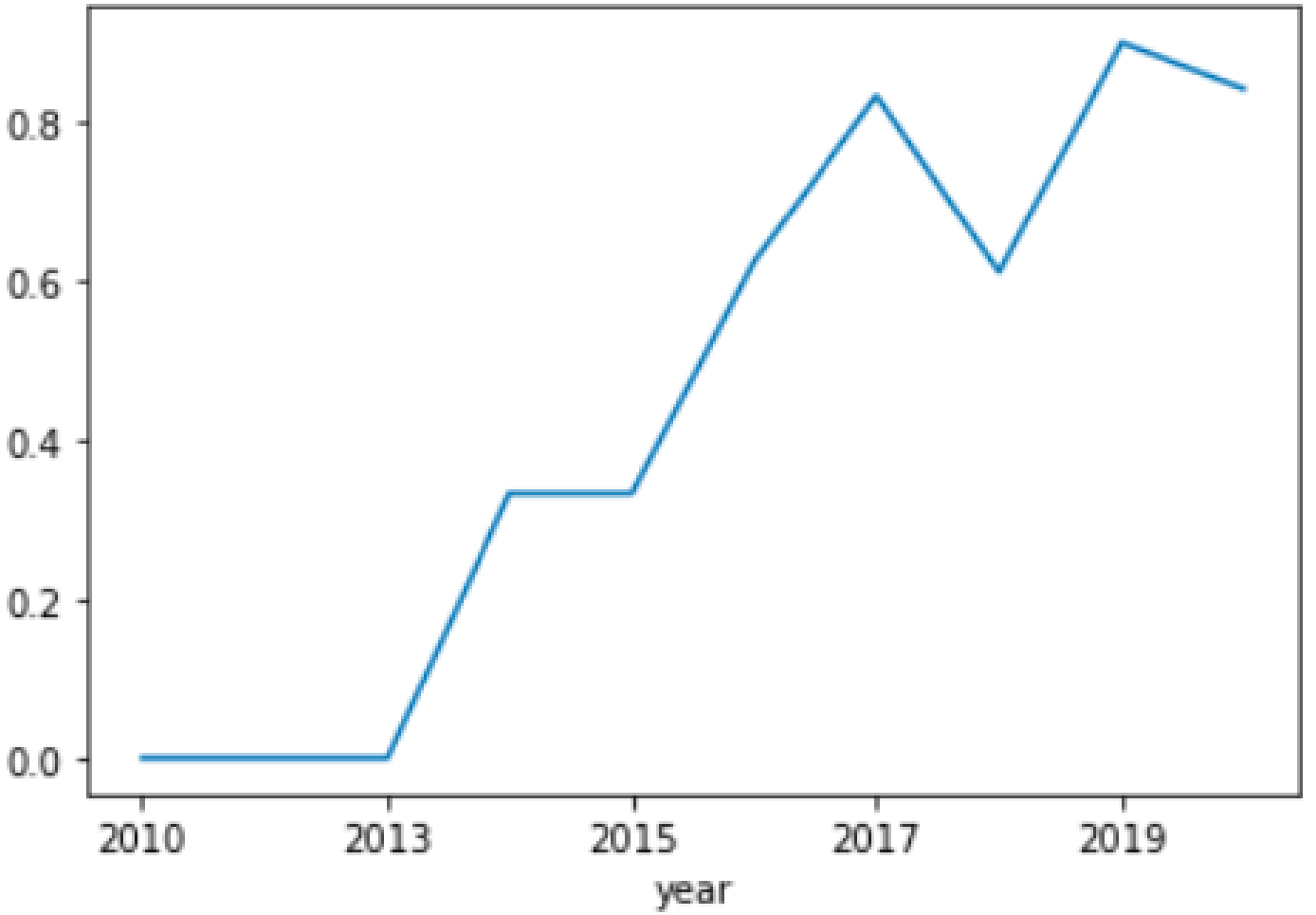
Payload vs. Orbit Type



- 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 Early Trend

- The sucess rate since 2013 kept increasing till 2020



All Launch Sites Names

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

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

There are four Launch Sites according to the SQL query result. Actually, CCAFS LC-40 and CCAFS SLC-40 are located close to each other and previously they were grouped into one site.

Launch Site Names Begin with '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
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

- The total payload mass carried by boosters from NASA is 48213 kg.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER LIKE "%NASA (CRS)%"
```

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

SUM(PAYLOAD_MASS__KG_)
48213

Average Payload Mass by F9 v1.1

- An average payload mass carried by booster version F9 v1.1 is 2928.4 kg.

```
%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

•The date when the first successful landing outcome in ground pad was achieved on the 1st March 2013.

```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE Mission_Outcome = "Success"
```

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

MIN(Date)
01-03-2013

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are only four boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000, their names are F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2.

```
%sql SELECT booster_version FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful mission outcomes is 100;
- The total number of failed mission outcomes is 1.

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

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

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

Boosters Carried Maximum Payload

```
%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

- There are 12 boosters which have carried the maximum payload mass

2015 Launch Records

```
%sql SELECT substr(Date, 4, 2) as Month, booster_version, launch_site FROM (SELECT * FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Failur
```

```
 * 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

- **In year 2015 there were only two failed in drone ship landing outcomes, in January and in April, both from CCAFS LC-40 launch site.**

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

```
%sql SELECT "Landing _Outcome", COUNT(*) AS COUNT_LAUNCHES FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%' AND
```

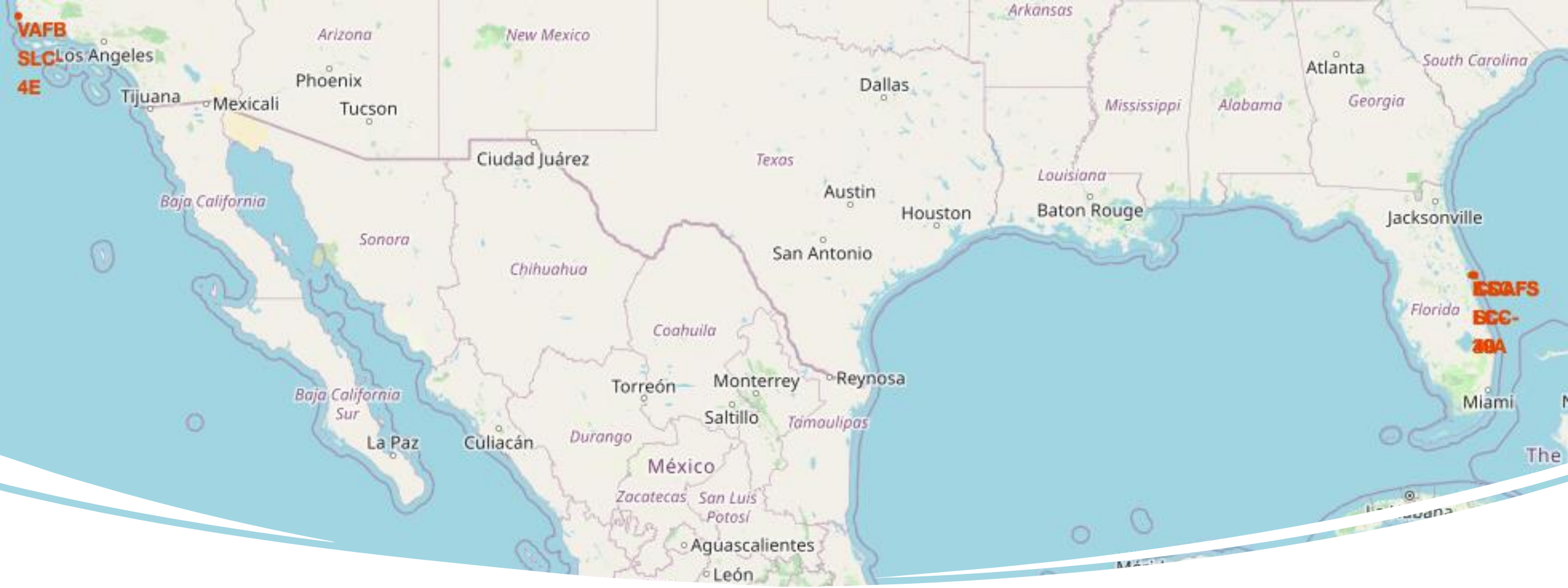
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	COUNT_LAUNCHES
Success	20
Success (drone ship)	8
Success (ground pad)	6

- **There were 34 successful landings between 04.06.2010 and 20.03.2017, 8 of wich are to a drone ship, 6 are to a ground pad, 20 without any additional information.**

Section 3

Launch Sites Proximities Analysis

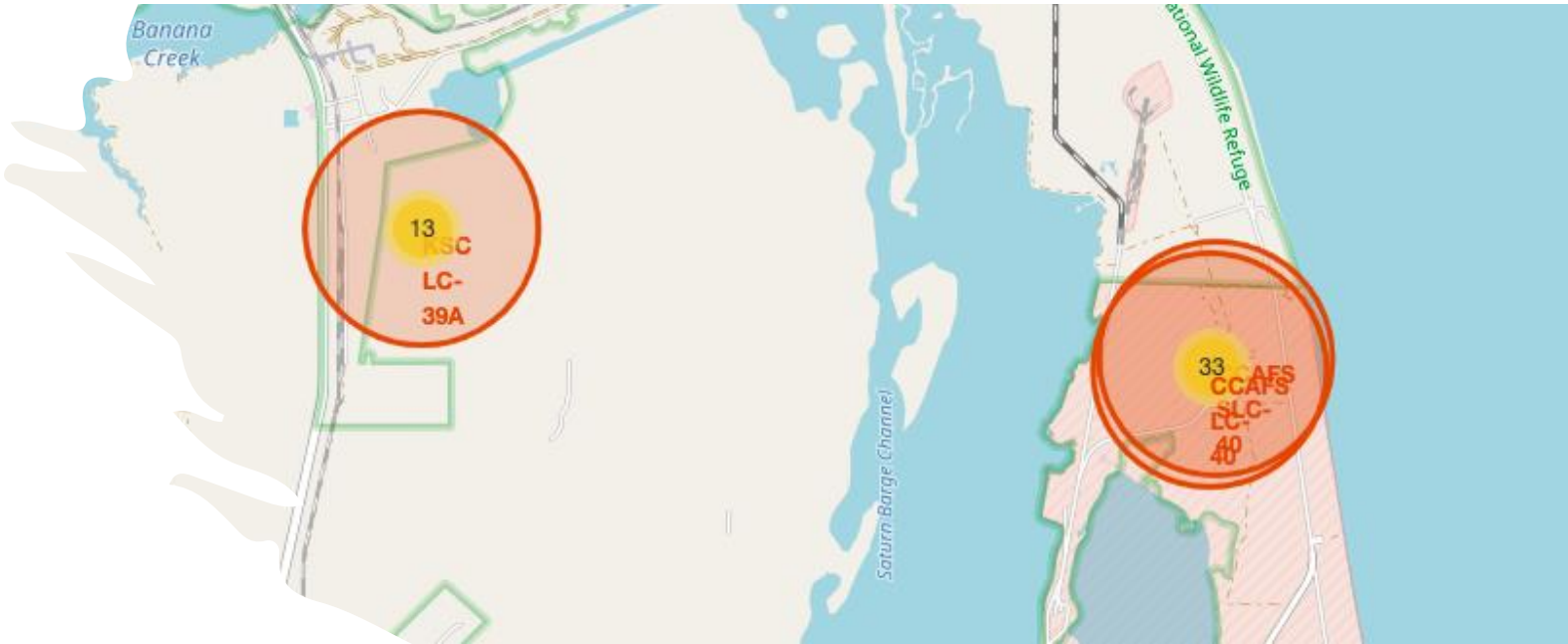


Map with Marked Launch Sites

- VAFB SLC 4E launch site is located on the West coast of the North America continent while two other launch sites (KSC LC 39A and CCAFS SLC 40) are located on the East coast.
- Meanwhile all three sites are close to the oceans.

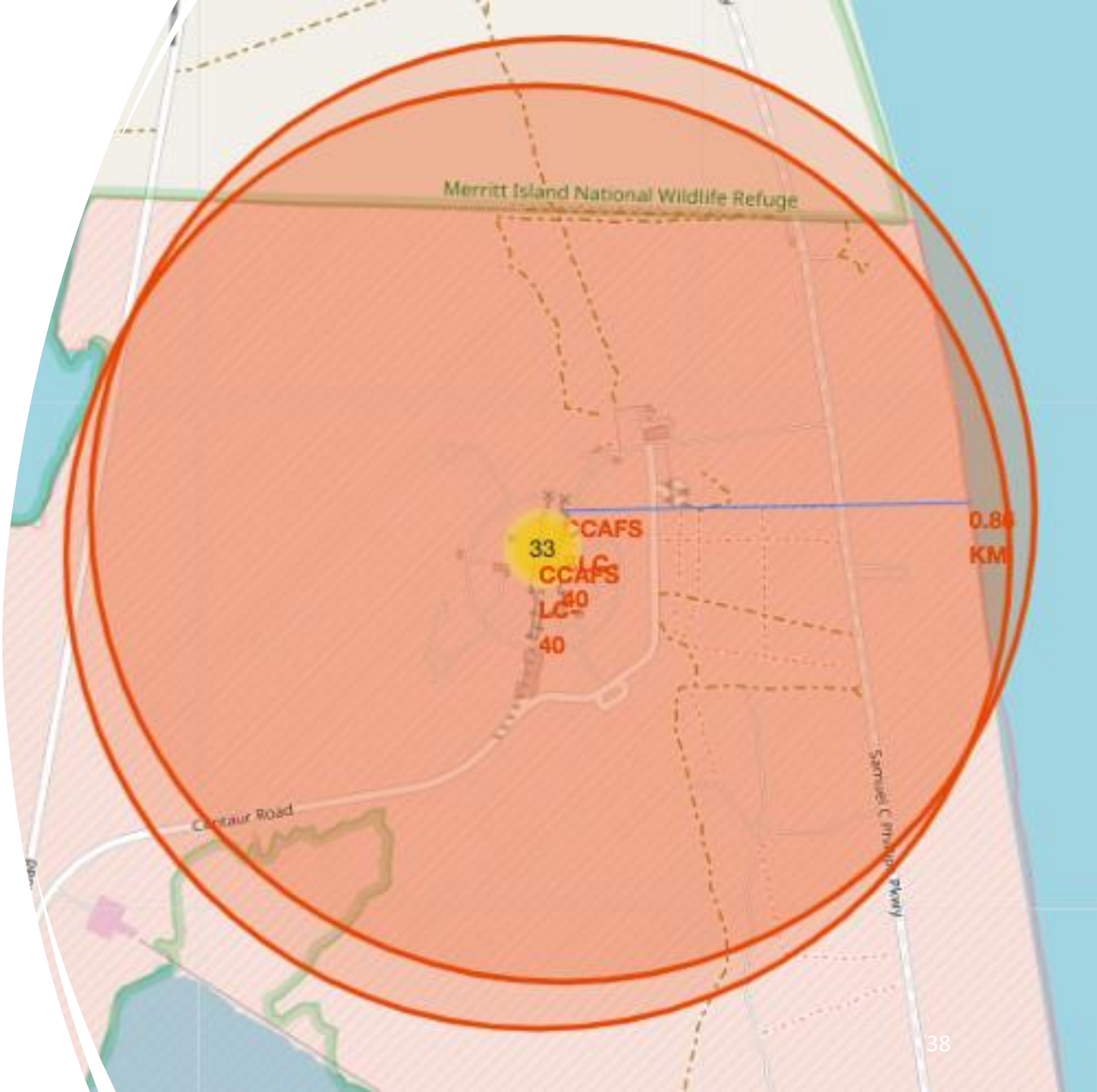
Map with Marked Successful and Failed launches

- From the color-labeled markers in marker clusters, we are able to easily identify that KSC LC-39A launch site has relatively high success rate.



Map with the Distances between a Launch Site to its proximities

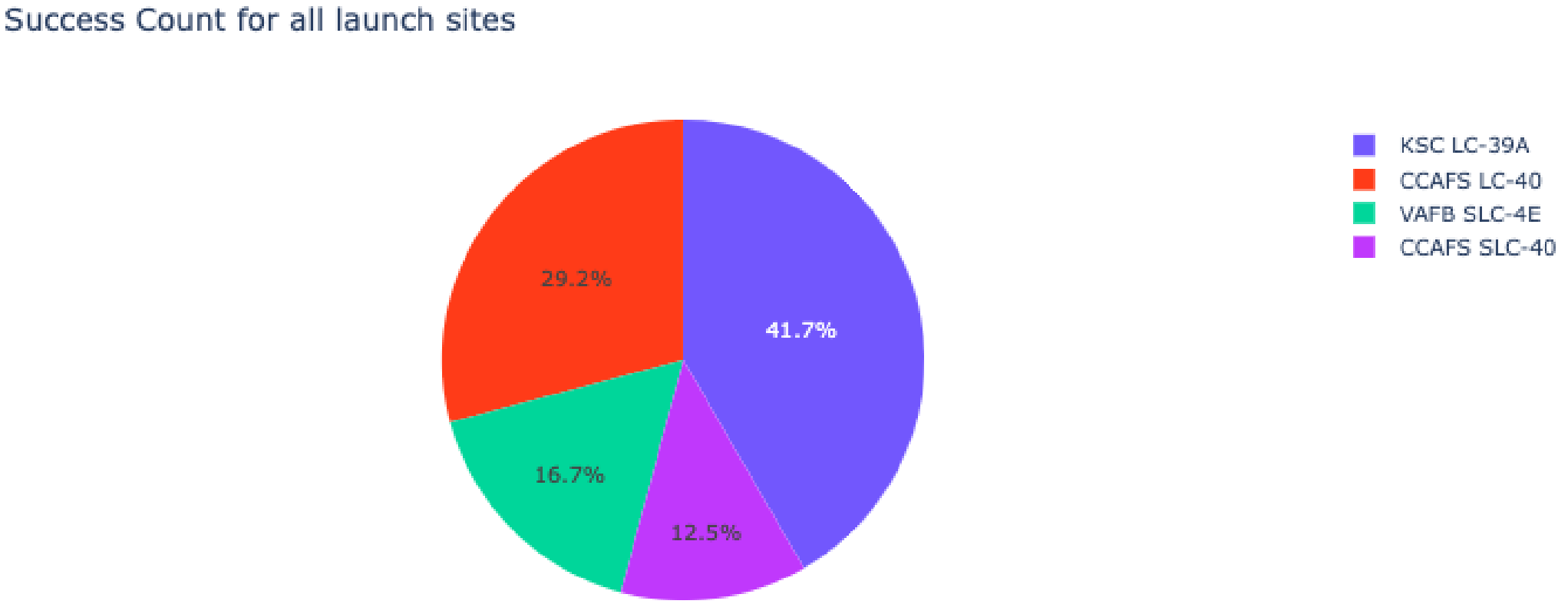
- CCAFS SLC 40 launch site is located less than 1 km from the coast, it could have impact on the launch outcomes.




Section 4
Build a Dashboard with Plotly Dash

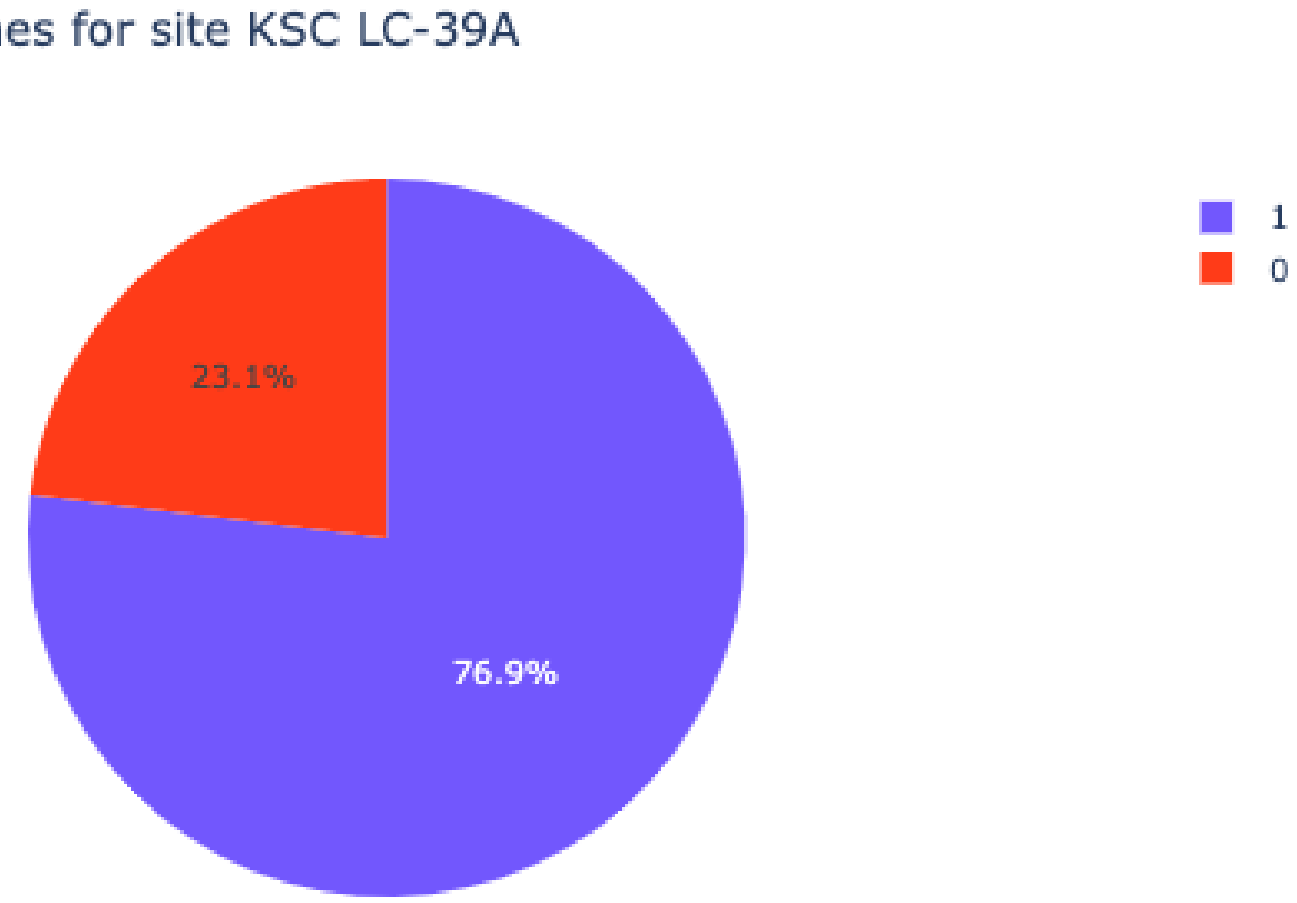
Launch Success Rates

- Pie chart illustrates that KSC LC-39A launch site has the higher number of successful flight among all launch sites. However, it doesn't mean that it's the best site for launch, it could be just an approval that the number of attempts from this site is higher



Launch Site with Highest Launch Success Ratio

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



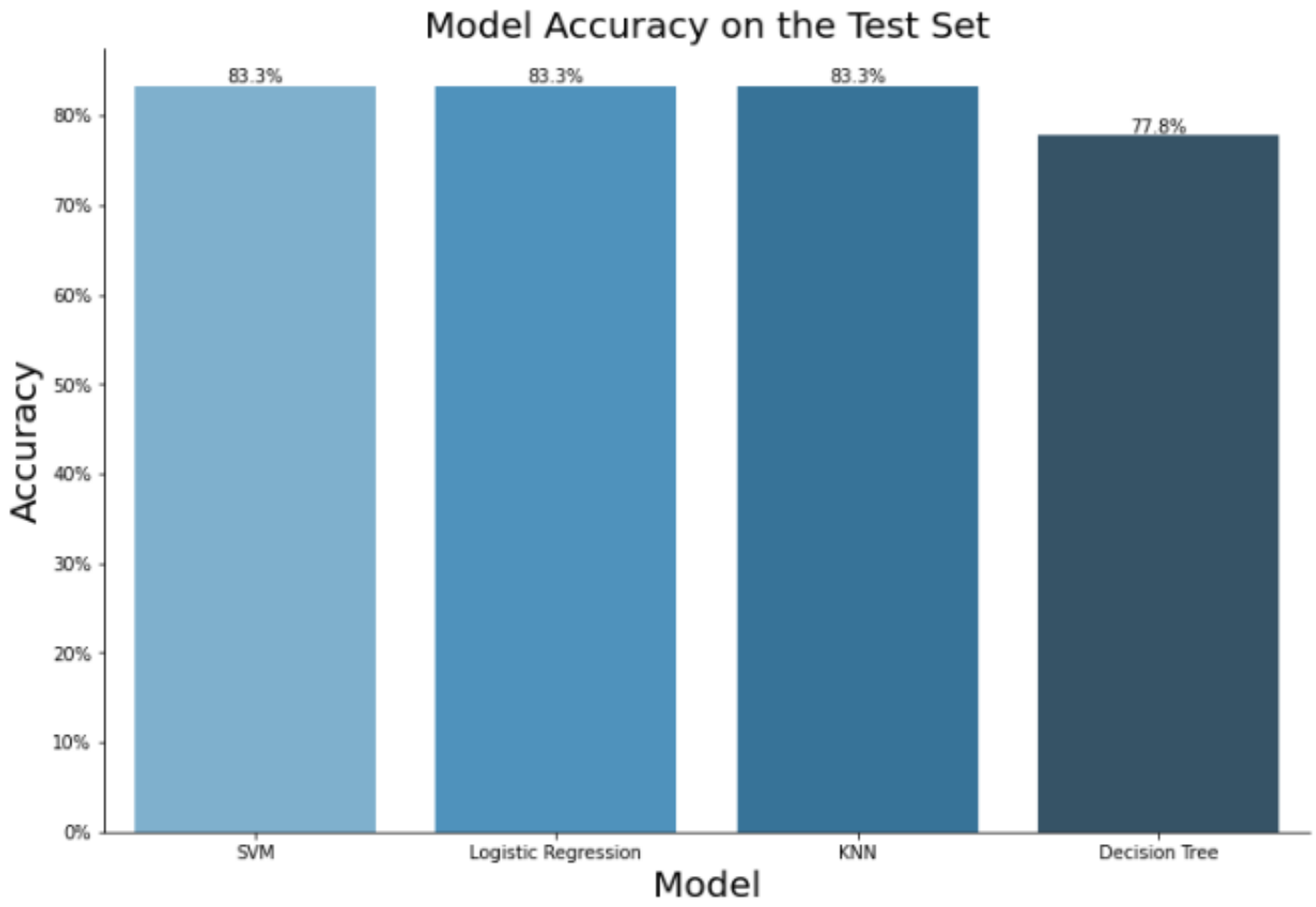
Payload vs. Launch Outcome

- The scatter plot for all sites demonstrates that the heaviest boosters (more than 5500 kg) don't have chances for their successful flights. The same we can conclude to the light boosters (500-1500 kg)


Section 5
Predictive Analysis (Classification)

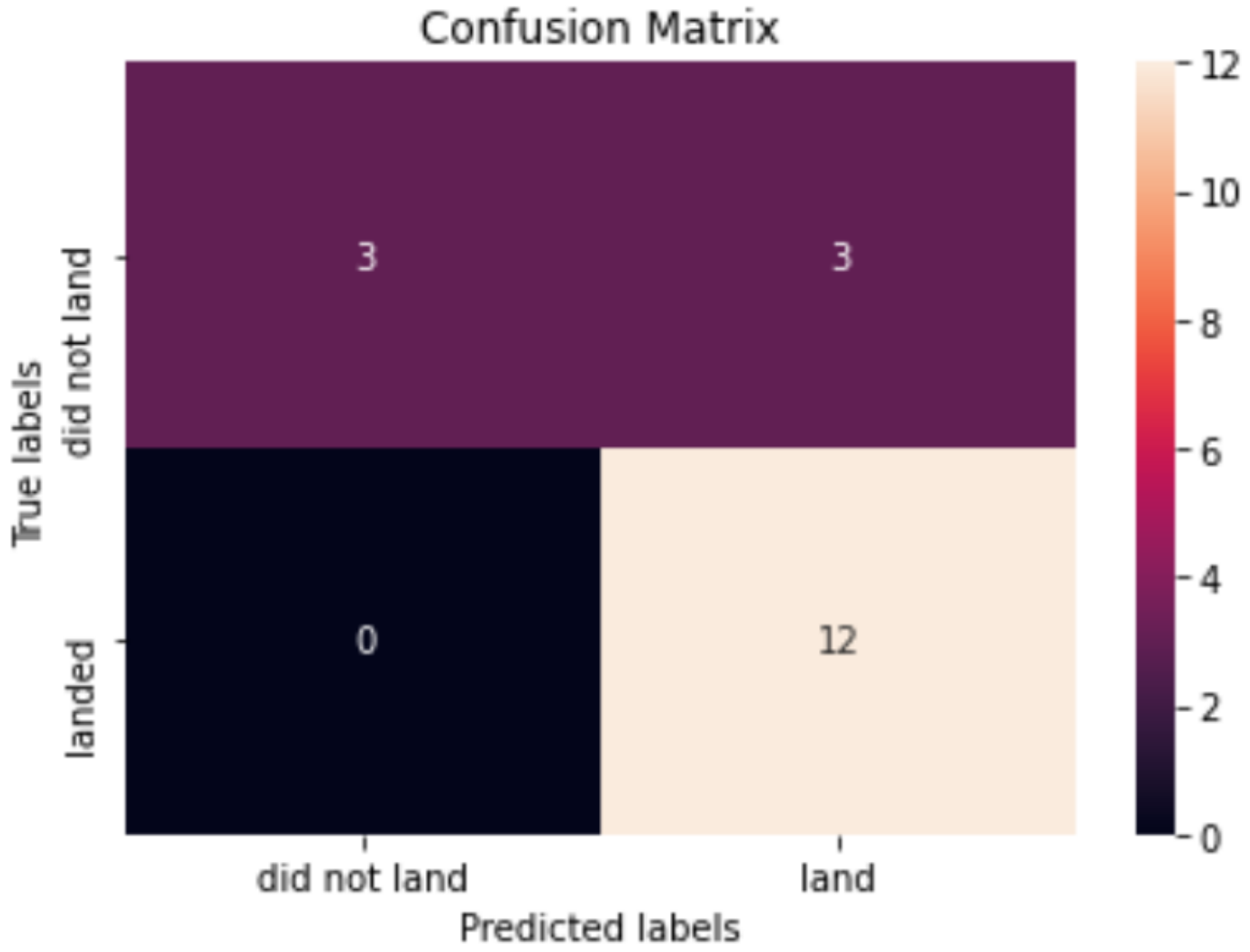
Classification Accuracy

- Three out of four models have the highest classification accuracy 83.3% which means that all of them could be used as classification models for supervised machine learning.



Confusion Matrix

- The confusion matrixes for all three appropriate models are the same and they illustrate that 83.3% of all predictions will be correct. At the same time 16.7% will be mistakenly predicted as successful.



Conclusions

In order to increase the probability of a successful launch, it's better:

- to choose orbits ES-L1, GEO, HEO, SSO as they have 100% successful flights;
- to perform a launch from KSC LC-39A launch site as it has the highest ratio of successful outcomes;
- to avoid launches of too light (till 2000 kg) and too heavy (from 5500 kg) as they have the lowest chances to the successful flights;
- to try more, because:

  - more number of flights, higher possibility of their success;

  - launch success rate is permanently increasing since 2013

Thank you!