IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Juan Manuel Ortiz Cahun
31/03/2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Methodology Summary

- Data was collected using SpaceX API
- Data was filtered selecting the attributes of interest. Standard dataset cleaning was carried out and some data were changed
- Exploratory data analysis (EDA) using visualization, and interactive visual analytics using Folium and Plotly Dash were carried out
- Four classification models were evaluated, each one was optimized

Results Summary

- There is growing costumer interest in the VLEO as target orbit.
- Kennedy Space Center had the highest landing success rate.
- Decision tree accuracy percentage was **88.57 and 83.33 %** when using training and testing data, respectively.

Introduction

- In recent years, the commercial space flights have become a booming industry.
- A company capable of putting cargo and people into space at **affordable prices** will **win this new space race**
- The key factor for achieving this is reuse as much as possible of the rocket
- There are some major competitors in this sector, SpaceX being the biggest.
- Our company **SpaceY** wants to lead the market and offer **better prices** to our costumers

SpaceX's Falcon 9

Introduction

- For doing so we need to **predict the price SpaceX will offer and bet against it**
- Some of the factors we think will influence the landing's success*, and therefore the cost, are:
  - Payload mass
  - Launch site
  - Target orbit
  - Type of technology used
- Throughout this report we will explore these factors and develop a model to predict the launch's success

* Successfully retrieving the first stage of the rocket

Section 1

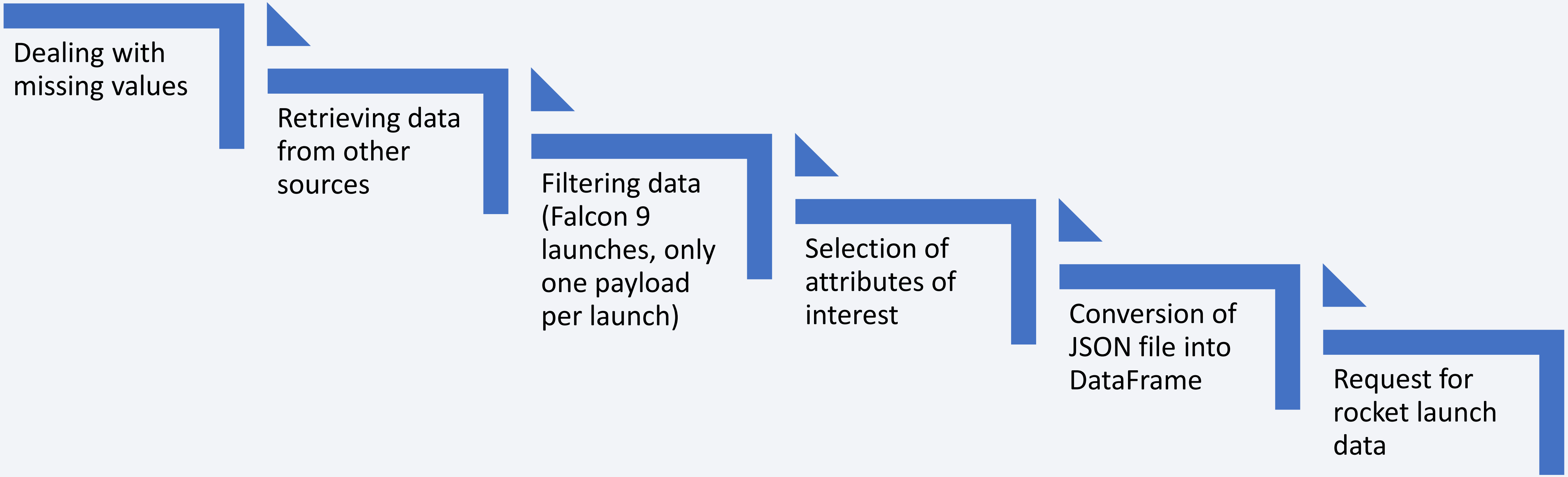
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API
- Perform data wrangling
  - Data was filtered selecting the attributes of interest and only Falcon 9 launches. Standard dataset cleaning was carried out and some data types of the attributes of interest were changed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four classification models were evaluated, each was optimized iterating through their hyperparametrs and using a cross validation value of 10

Data Collection



Dealing with missing values

Retrieving data from other sources

Filtering data (Falcon 9 launches, only one payload per launch)

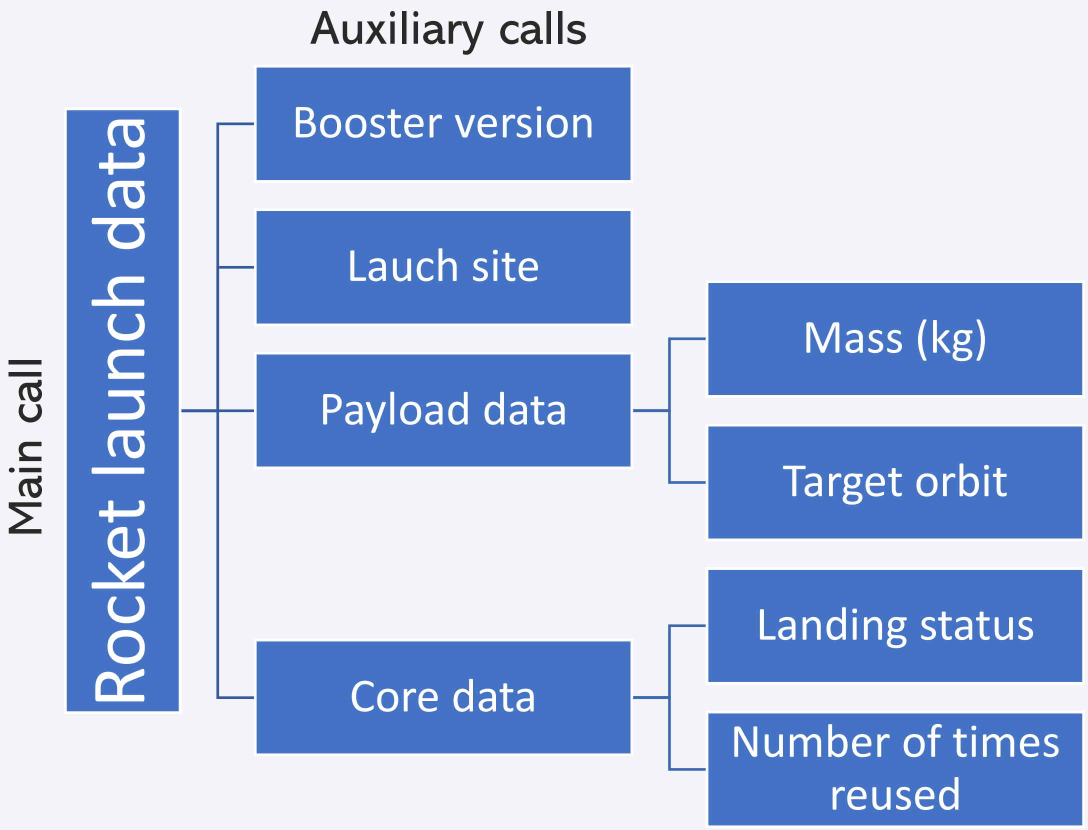
Selection of attributes of interest

Conversion of JSON file into DataFrame

Request for rocket launch data

Data Collection – SpaceX API

SpaceX REST API calls



The main call retrieves the dataset with all past SpaceX launches.
Knowing the general information of each launch, the subsequent calls obtain the rest of the details. For example, the main call obtains the launch site, but the "Launch site" call obtains the launch site coordinates.

https://github.com/OmarRodriguezG/Applied-Data-Science-Capstone/blob/0624ef417cc740d25eaa53fc35bba195d6f740da/data-collection-api.ipynb

Data Wrangling

- Missing values were addressed. They were substituted by the mean
- Target orbits were identified
- Landing status was converted from string to numerical
  - 1 for success
  - 0 for failure

https://github.com/OmarRodriguezG/Applied-Data-Science-Capstone/blob/0624ef417cc740d25eaa53fc35bba195d6f740da/Data%20wrangling.ipynb

EDA with Data Visualization

- The following charts were plotted:
  - Flight number vs Payload mass
  - Flight number vs Launch site
  - Payload mass vs Launch site
  - Payload mass vs Target orbit
  - Bar chart of Target orbit vs Success rate
  - Line plot of Year vs Success rate

Scatter plot. Landing status for marker color

- Scatter plots were used to quickly explore all the independent variables and identify which of them are correlated with each other and to a successful landing.
- When using the success rate as a dependent variable, simpler plots can be used.

https://github.com/OmarRodriguezG/Applied-Data-Science-Capstone/blob/0624ef417cc740d25eaa53fc35bba195d6f740da/Exploratory_analysis.ipynb

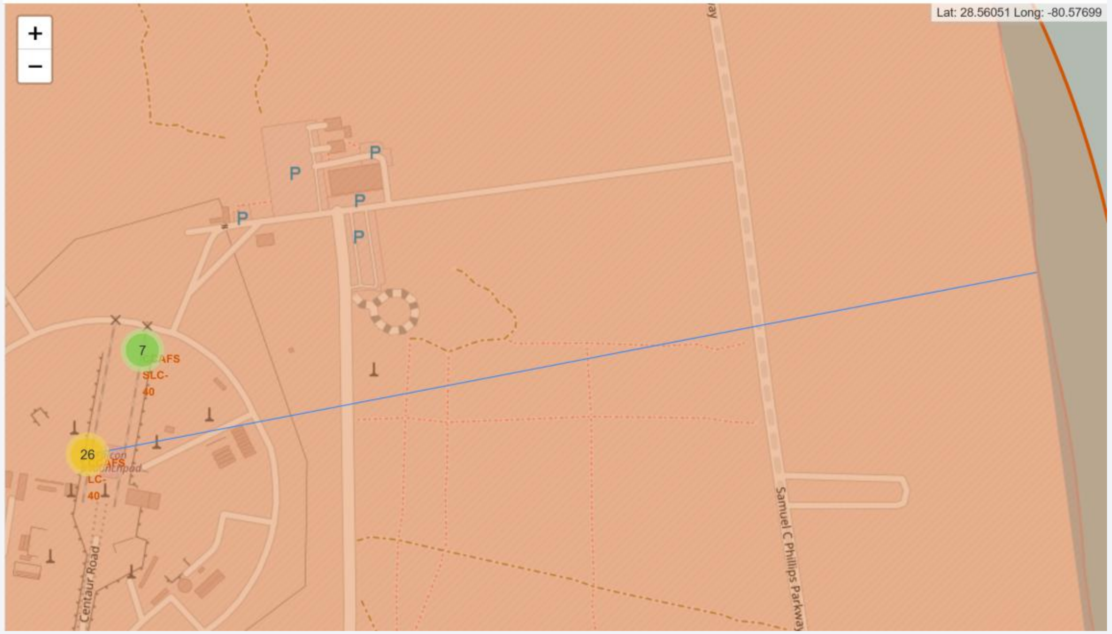
EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Select the unique launch sites
  - Find the payload mass put in orbit by customer
  - Find the average mass carried by the Falcon 9 v1.1
  - Find the name of the boosters in which its drone ship landing was successful having carried a payload mass greater than 4000 and 6000
  - Find the number of successful and failed missions
  - Filter the successful and failed missions using the date

https://github.com/OmarRodriguezG/Applied-Data-Science-Capstone/blob/0624ef417cc740d25eaa53fc35bba195d6f740da/sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- A map of the United States was plotted.
  - Circles with a 1-kilometer radius were drawn around the launch sites. Their names were added as a label
  - A maker was created for every launch
  - A marker cluster was created to group nearby launches to prevent to overcrowd the launch site
  - Lines were drawn to various points of interest such as coastline, railroads and highways



https://github.com/OmarRodriguezG/Applied-Data-Science-Capstone/blob/0624ef417cc740d25eaa53fc35bba195d6f740da/Folium.ipynb

Build a Dashboard with Plotly Dash

Two types of graphs were used

Pie chart

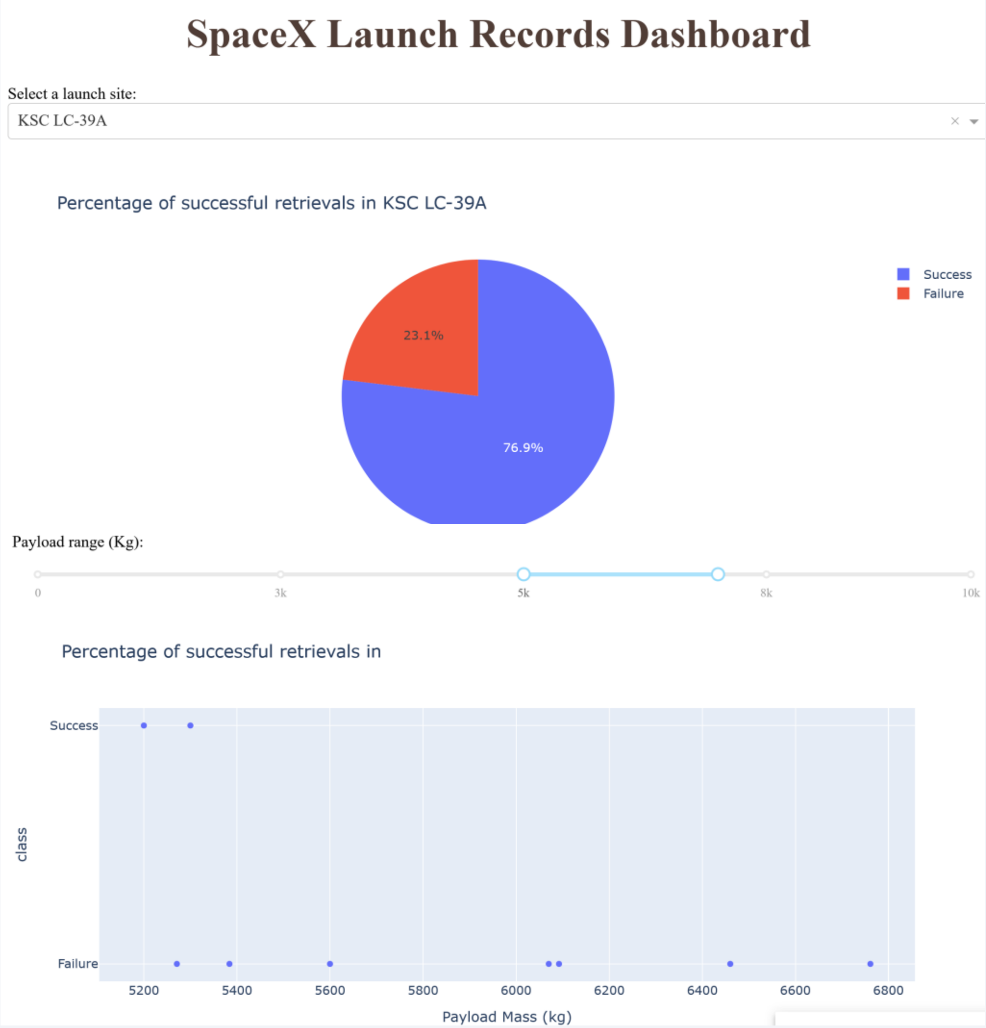
- Showing the number of successful landings by launch site
- Showing the success rate by launch site

- A dropdown menu was added to choose between all launch sites or a specific one

Scatter plot

- Compares the payload mass with the landing status to find a correlation

- A slider was added to select the payload mass range for analysis



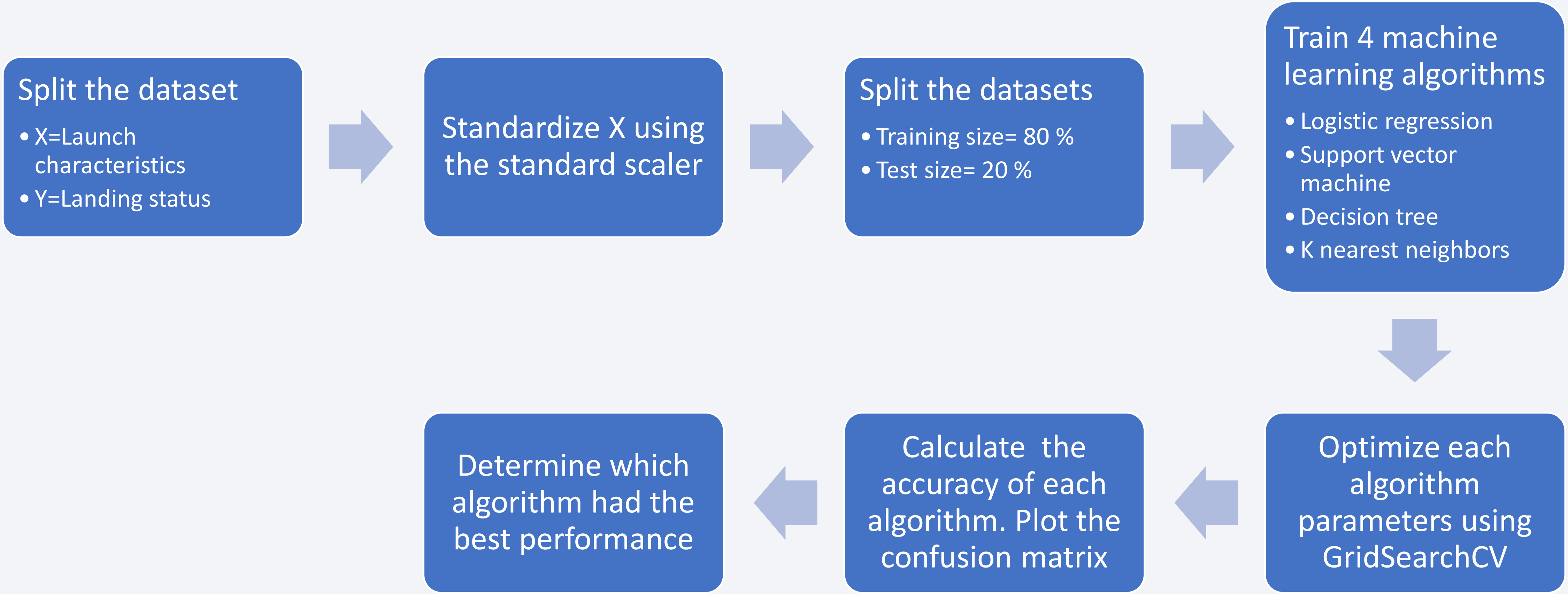
SpaceX Launch Records Dashboard

Select a launch site: KSC LC-39A

Payload range (Kg):

https://github.com/OmarRodriguezG/Applied-Data-Science-Capstone/blob/0624ef417cc740d25eaa53fc35bba195d6f740da/Dashboard_final.ipynb

Predictive Analysis (Classification)



Split the dataset

- X=Launch characteristics
- Y=Landing status

Standardize X using the standard scaler

Split the datasets

- Training size= 80 %
- Test size= 20 %

Train 4 machine learning algorithms

- Logistic regression
- Support vector machine
- Decision tree
- K nearest neighbors

Optimize each algorithm parameters using GridSearchCV

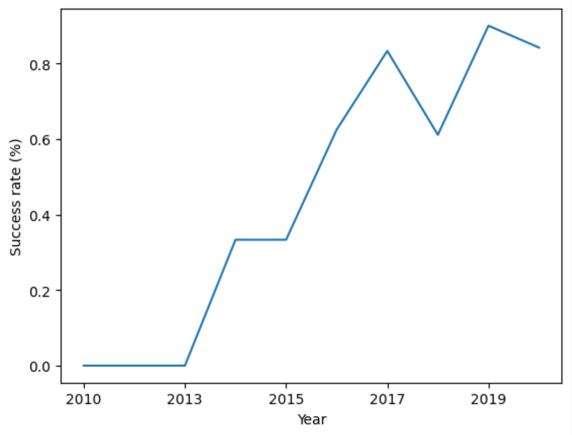
Calculate the accuracy of each algorithm. Plot the confusion matrix

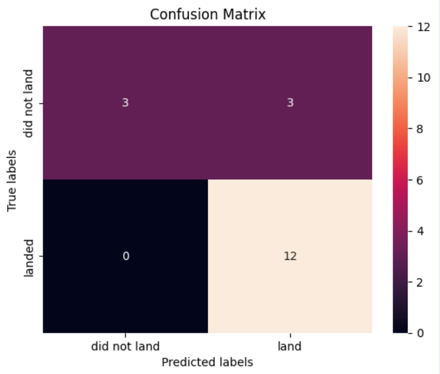
Determine which algorithm had the best performance

https://github.com/OmarRodriguezG/Applied-Data-Science-Capstone/blob/0624ef417cc740d25eaa53fc35bba195d6f740da/Machine%20Learning%20Prediction.ipynb

Results

- Here is an overview of the results. In the following sections they will be presented and discussed in detail
- Exploratory data analysis results
  - With each new flight and the passage of time , the success rate increased
  - KSC LC 39A launch site had a higher success rate
- Interactive analytics demo in screenshots
  - Payload mass seems to have a positive impact in the landing status
- Predictive analysis results
  - Decision tree algorithm had the best performance using training data. When test data was used, all four algorithms had the same performance

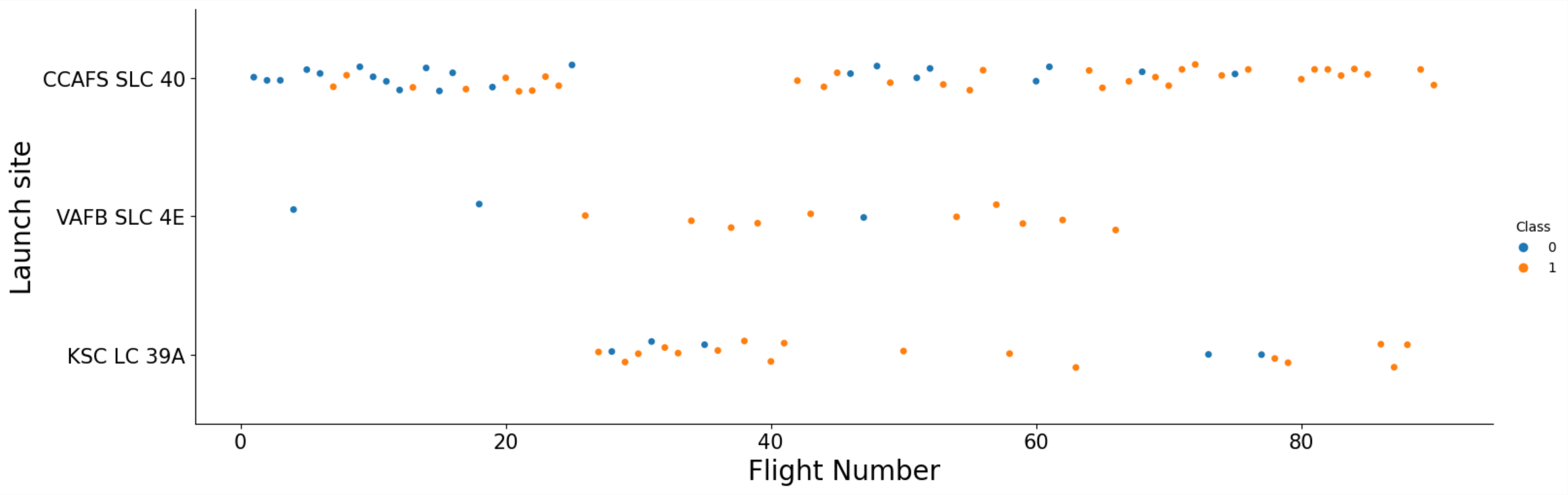




Section 2

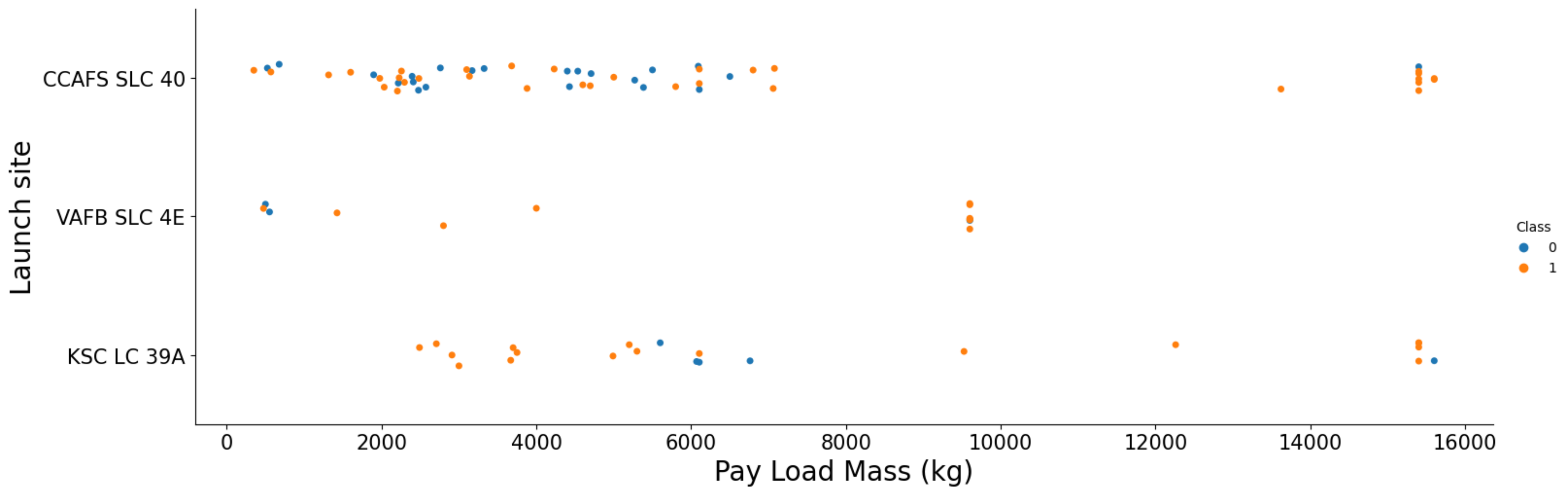
Insights drawn from EDA

Flight Number vs. Launch Site



- Most launches have occurred in Florida, in the Cape Canaveral Air Force Station (CCAFS) and the Kennedy Space Center (KSC). These launch sites are closer to the equator, where the rockets can take advantage of earth's rotational speed.
- As the flight number increases, the number of successful landings increases.

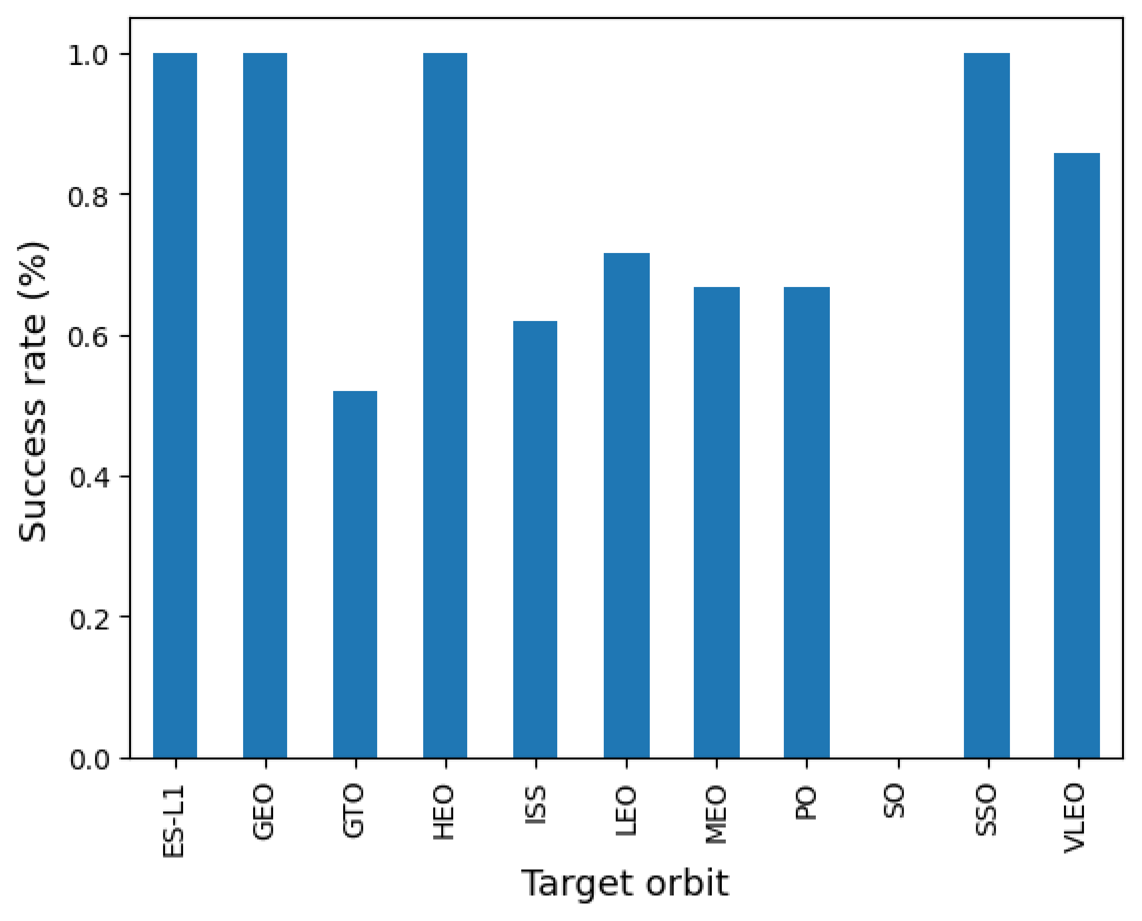
Payload vs. Launch Site



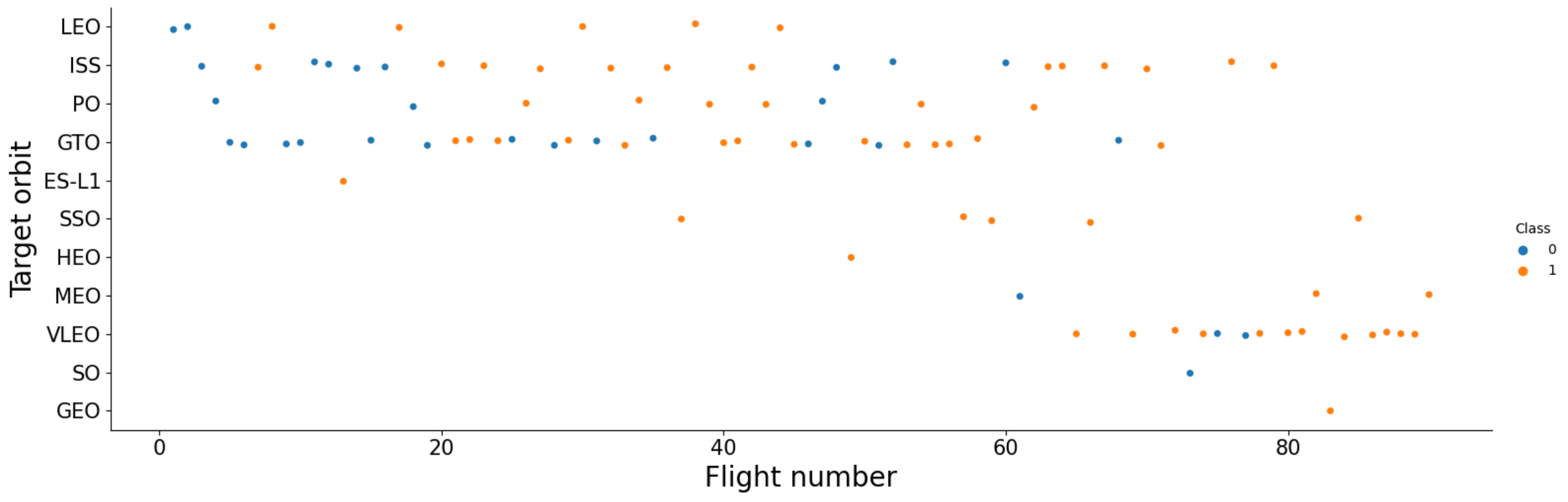
- Most launches carry a payload mass lower to 8,000 kg.
- Over a payload of 8000 kg most of the landings are successful, independently of the launch site.
- However, this correlation doesn´t imply causation. With the passage of time flights carried a greater payload, but also the SpaceX's expertise increased, what may have caused a landing methodology refinement and, in turn, more successful landings.

Success Rate vs. Orbit Type

- Most target orbits have a success rate superior to 60 %. Only the Geostationary transfer orbit (GTO) and one Sun-synchronous orbit (SO) have an inferior success rate, 51 % and 0 %, respectively.
- The First Lagrange Point (ES-L1), Geostationary orbit (GEO), Highly Elliptical Orbit (HEO) and Sun-synchronous orbit (SSO) have a 100 % success rate

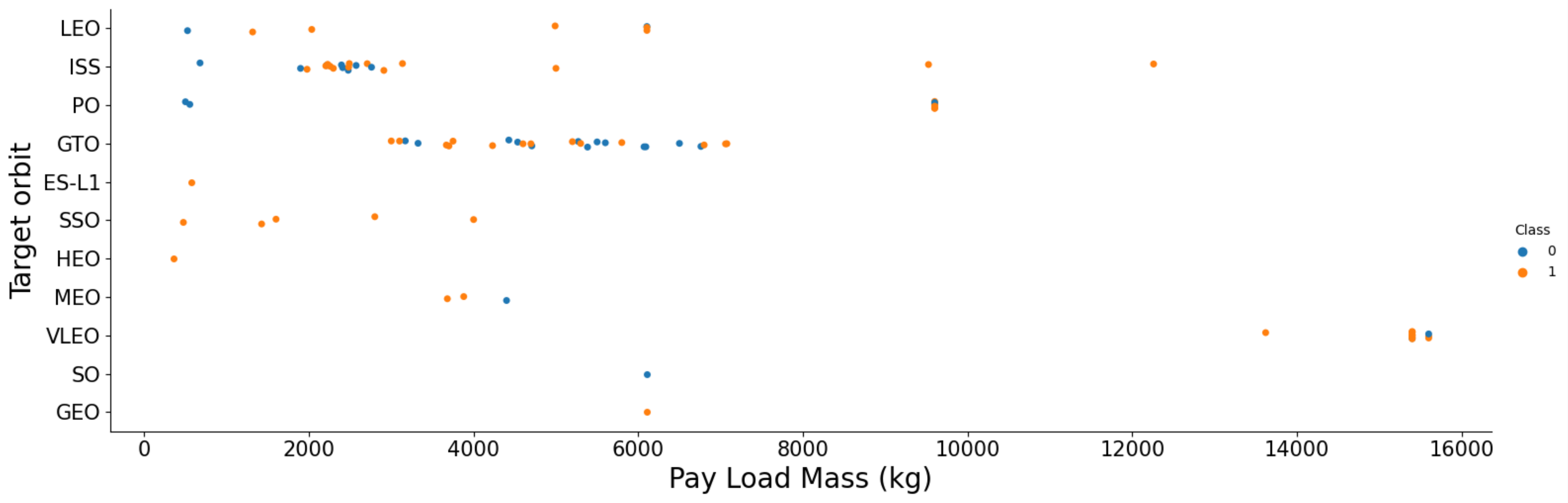


Flight Number vs. Orbit Type



- Many launches had a target orbit relatively low, i.e. LEO and VLEO. When these were the target orbits, most landings were successful.
- With passage of time more flights have VLEO as target orbit.
- One important target orbit is GTO. It is farthest orbit and, coincidentally, have a mixed record of landings.

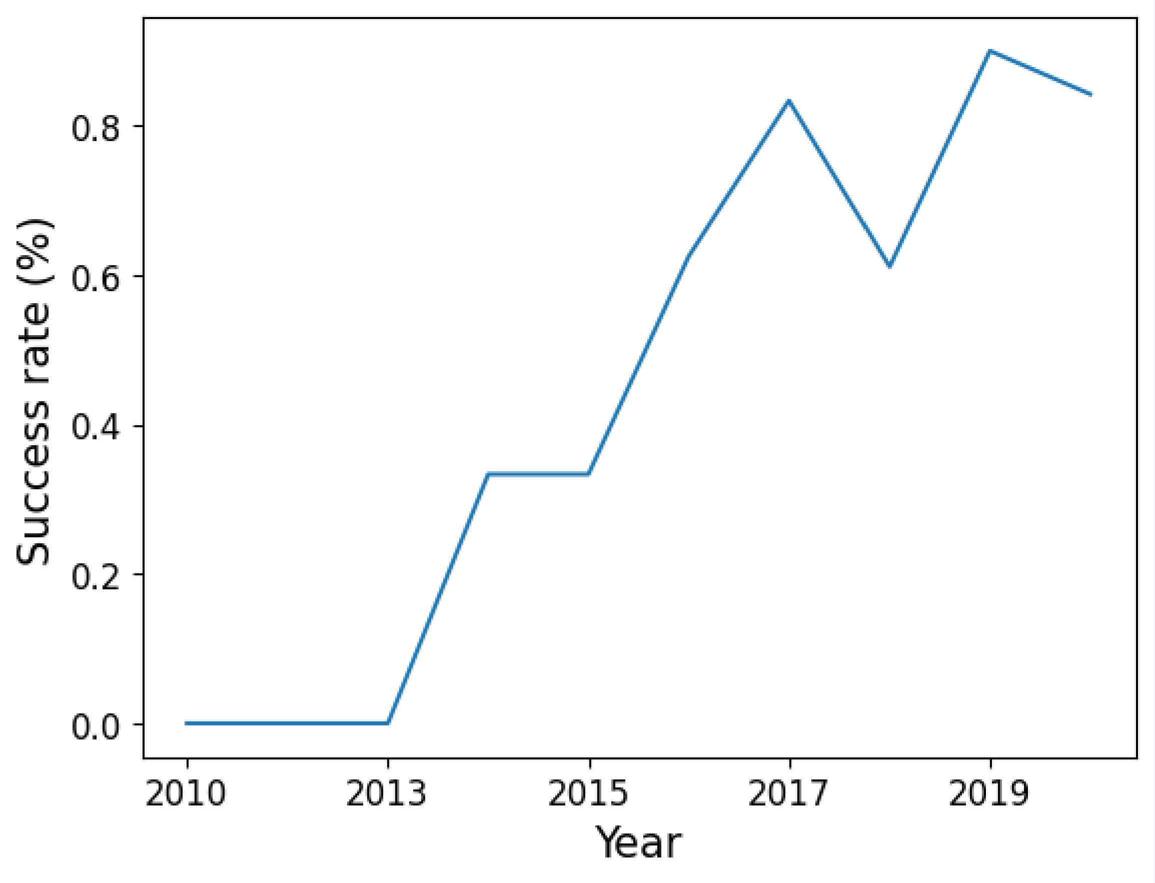
Payload vs. Orbit Type



- All the flights with VLEO as target orbit carried a payload mass superior to 13000 kg.
- Most of the flights to the ISS orbit had a payload mass of 2000-3000 kg
- All flights to the GTO carried a payload mass between 3000 and 7000 kg

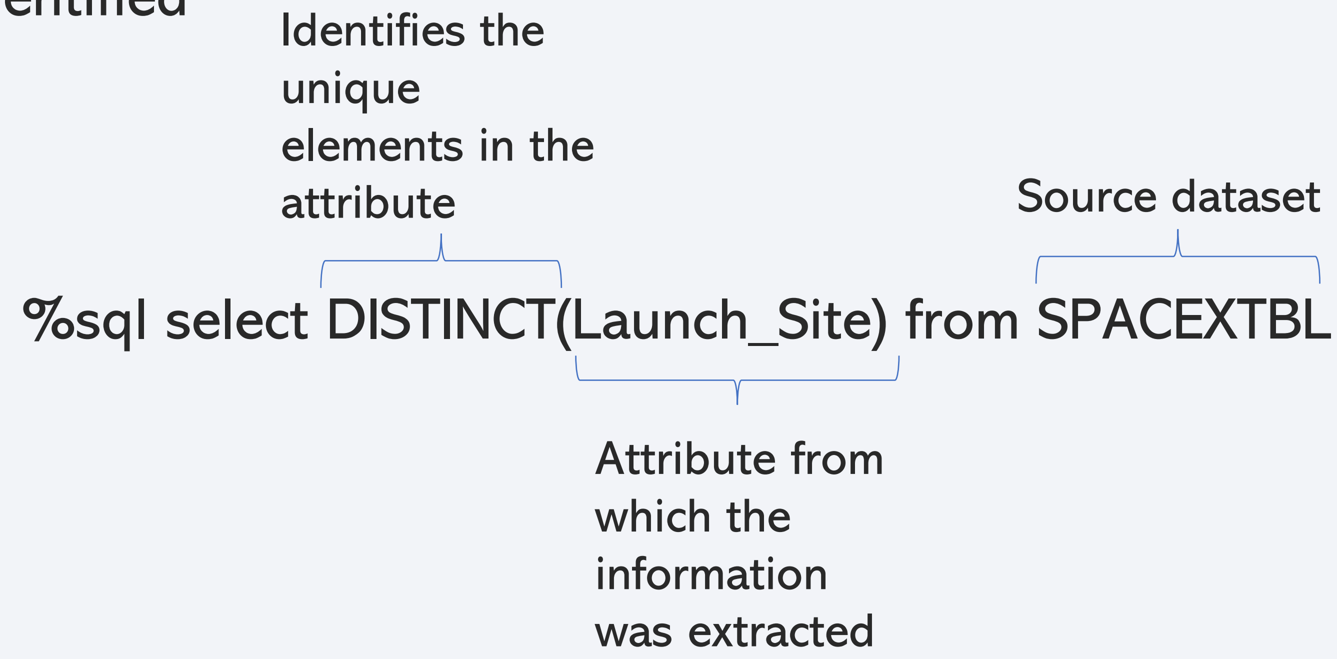
Launch Success Yearly Trend

- Like was previously stated, with the passage of time the landing success rate increased, independently of payload mass, target orbit, etc. With time, the SpaceX team expertise controlling the landing parameters increased, which in turn, increased the success rate.



All Launch Site Names

- Using the following SQL query all launch sites names were identified



Identifies the unique elements in the attribute

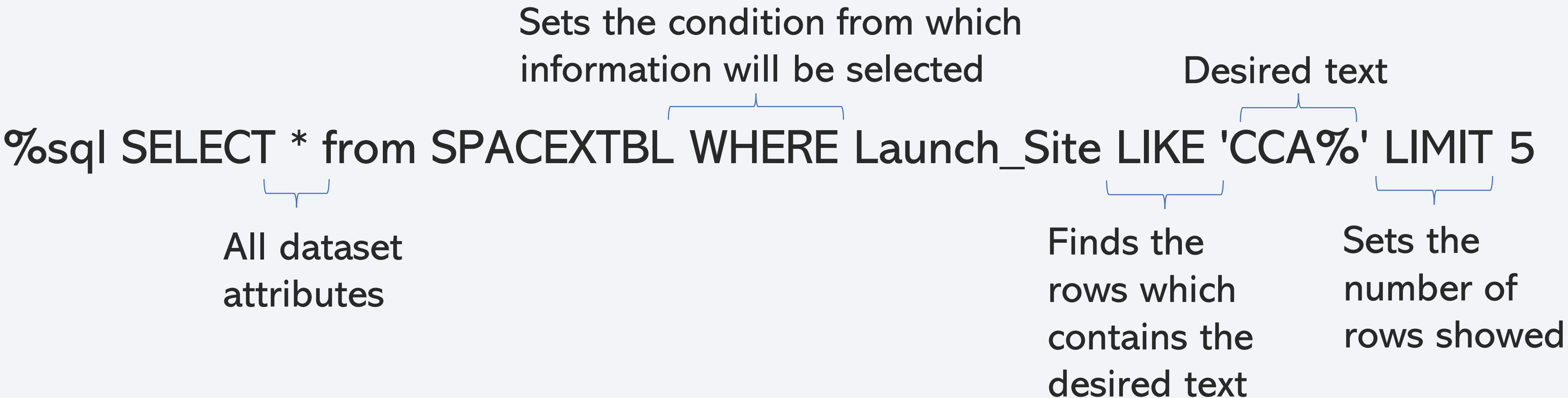
Source dataset

```
%sql select DISTINCT(Launch_Site) from SPACEXTBL
```

Attribute from which the information was extracted

Launch sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'



Sets the condition from which information will be selected

Desired text

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

All dataset attributes

Finds the rows which contains the desired text

Sets the number of rows showed

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

Total Payload Mass carried by boosters from NASA

Mathematical function applied to the attribute

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE Customer LIKE '%CRS%’
```

Total Payload Mass= 48213 KG

Average Payload Mass by F9 v1.1

- In this case there are not new commands needing explanation, so, the query is pasted as it is

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

Average Payload Mass= 6210.36 KG

First Successful Ground Landing Date

Selects the attribute’s smaller element. For dates, the oldest

```
%sql SELECT MIN(Date) from SPACEXTBL WHERE Landing_Outcome='Success’
```

In this case the condition is equal because it's known the required outcome

Earliest successful ground landing date= 2018-07-22

Although SpaceX started operations in 2010, the first successful ground landing occurred in 2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- It is required to find the Booster version where a condition with two clauses

First clause

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

Second clause. Actually, there are 3 clauses because the range was limited using another AND, but this could be achieved using the function BETWEEN

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

Total Number of Successful and Failure Mission Outcomes

- First, it was necessary to identify all the mission outcomes. To do this, the following query was used

```
%sql SELECT DISTINCT(Mission_Outcome) from
SPACEXTBL
```

Mission_Outcome
Success
Failure (in flight)
Success (payload status unclear)
Success

- Knowing the possible outcomes, a query was formulated searching for all of them having the word “Success” in them

```
%sql SELECT COUNT(Mission_Outcome) from
SPACEXTBL WHERE Mission_Outcome LIKE
'Success%’
```

Number of successful missions= 100

Number of failed missions= 1

Boosters Carried Maximum Payload

- For this a subquery is needed. First, is necessary to find the maximum payload mass in the dataset. Then, knowing the value, all the Boosters that carried that payload were selected.

Main query

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

Subquery

Booster_Version	
F9 B5 B1048.4	F9 B5 B1049.5
F9 B5 B1049.4	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1058.3
F9 B5 B1056.4	F9 B5 B1051.6
F9 B5 B1048.5	F9 B5 B1060.3
F9 B5 B1051.4	F9 B5 B1049.7

2015 Launch Records

This function extracts the month (expressed as a number) in which the launch took place

```
%sql SELECT substr(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site from
SPACEXTBL WHERE Date LIKE '2015%'
```

substr(Date, 6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

```
%sql SELECT Landing_Outcome FROM SPACEXTBL WHERE
Date BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY
Date DESC
```

Modifies the data order

Rearranges the data depending on the selected attribute

The first ten landing statuses are presented

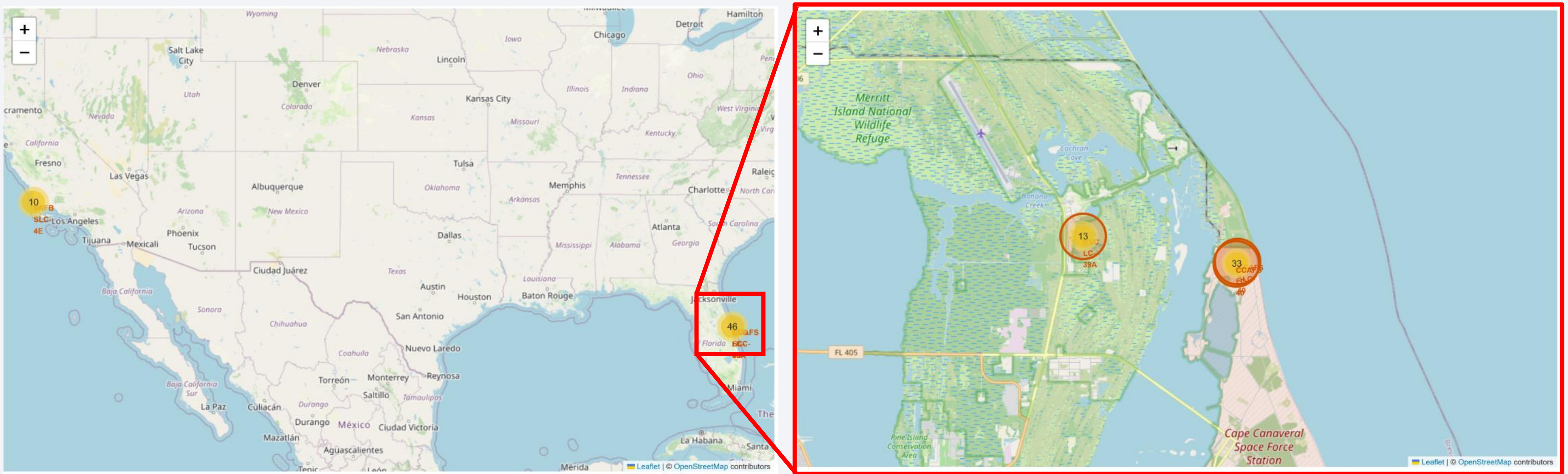
```
Landing_Outcome
No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Success (drone ship)
```

Section 3

Launch Sites Proximities Analysis

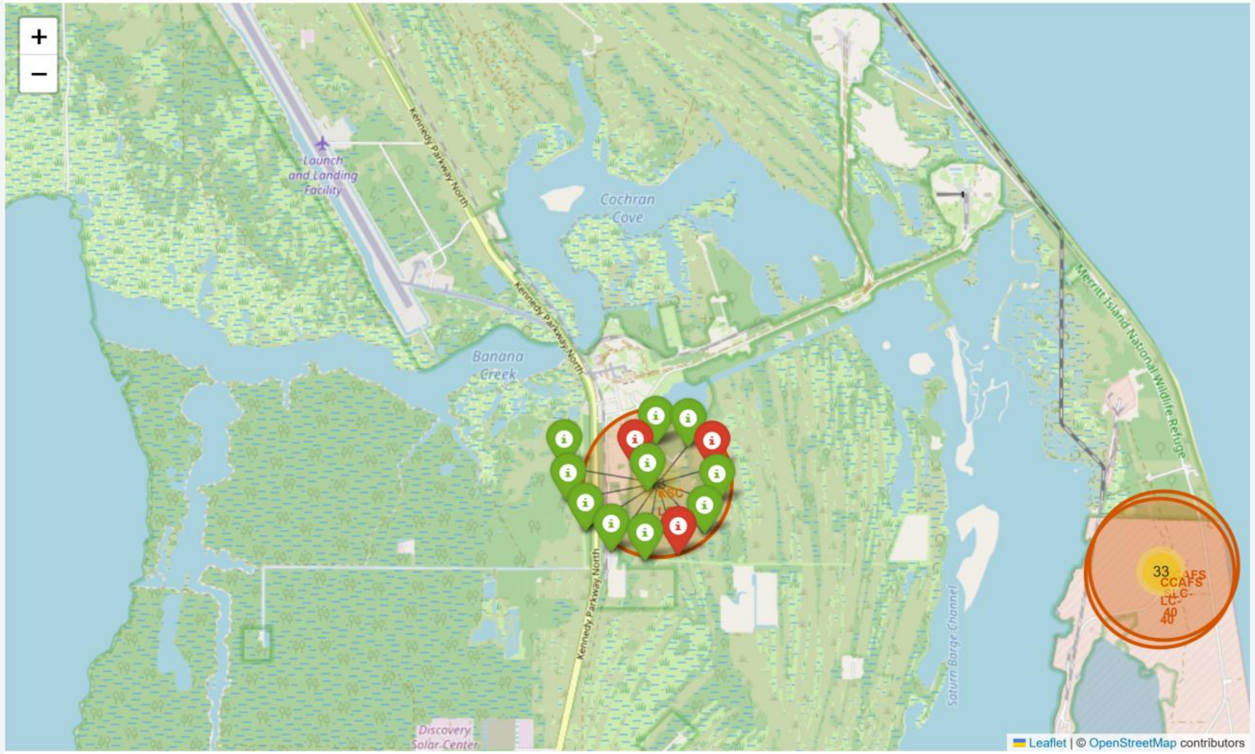
Launch sites' location

- There are four launch sites in America:
  - One in California, VAFB SLC-4E
  - Three in Florida, KSC LC-39A, CCAFS LC-40 and CCAFS SLC-40. The last two are very close to each other.
- All of them are as close as possible to the Equator. At this point, the rockets take advantage of the Earth's rotational speed to escape the atmosphere



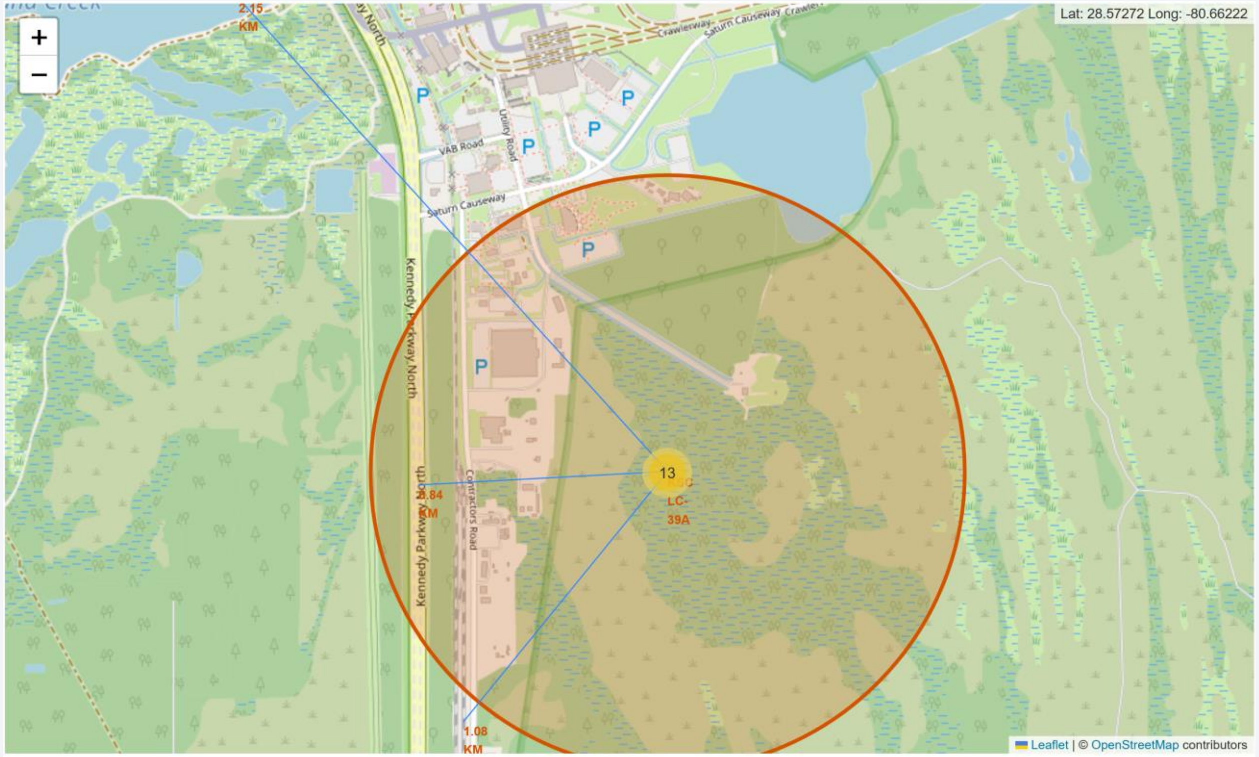
Launches according to their landing status

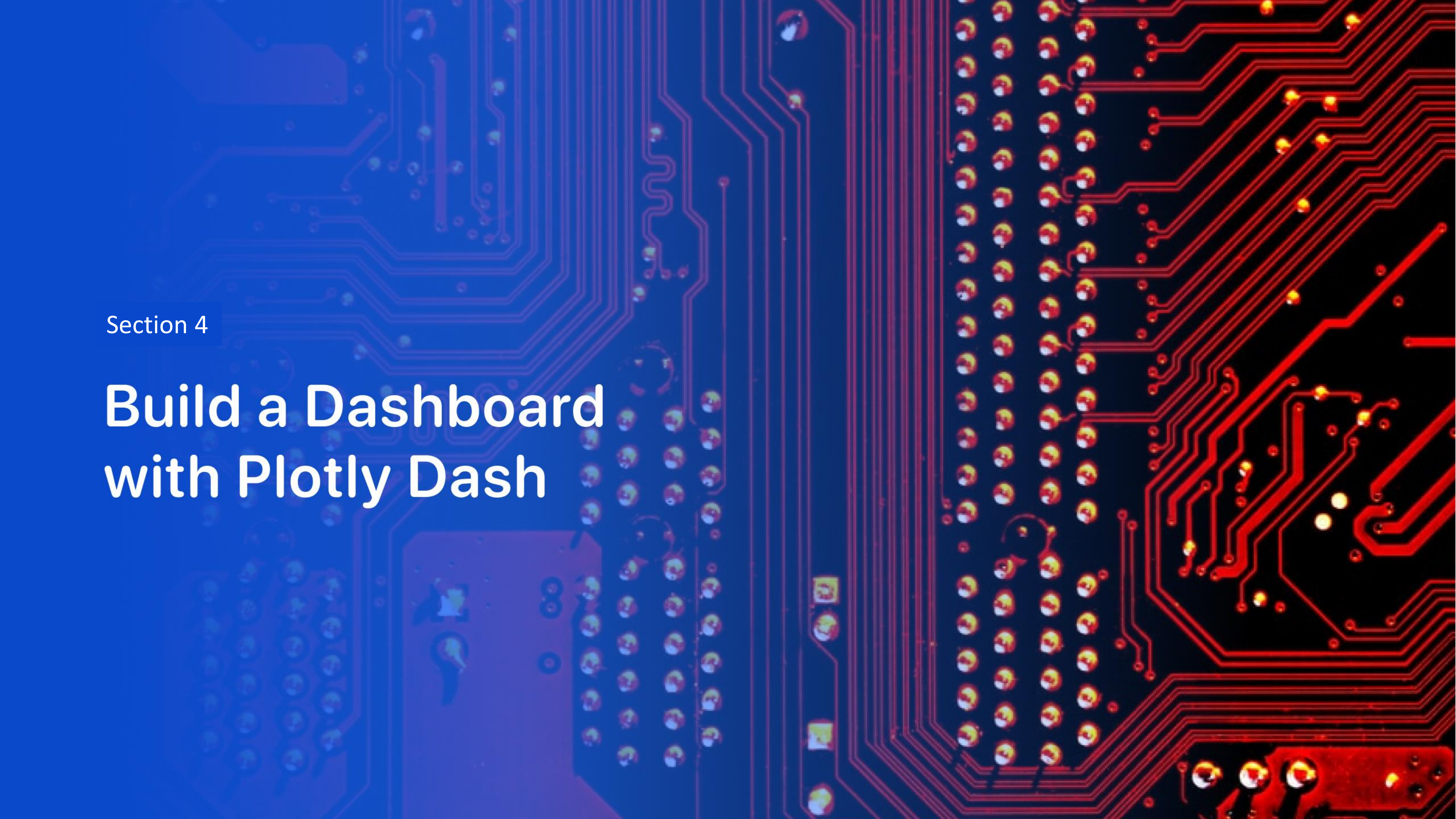
Most launches were done in Florida. Of these, the KSC LC-39A (Kennedy Space Center) had the highest success rate. In the screenshot it can be see it that three of them failed, the rest landed successfully.



Launch site surroundings and their importance

The KSC was chosen because its success rate. Exploring its surroundings it can be seen that the launch site is close to the coastline, but not as close as other launch sites. This can be helpful because being more inland shields the launch site of the wind streams. Likewise, around the launch site there are fewer buildings, which alters less the wind streams



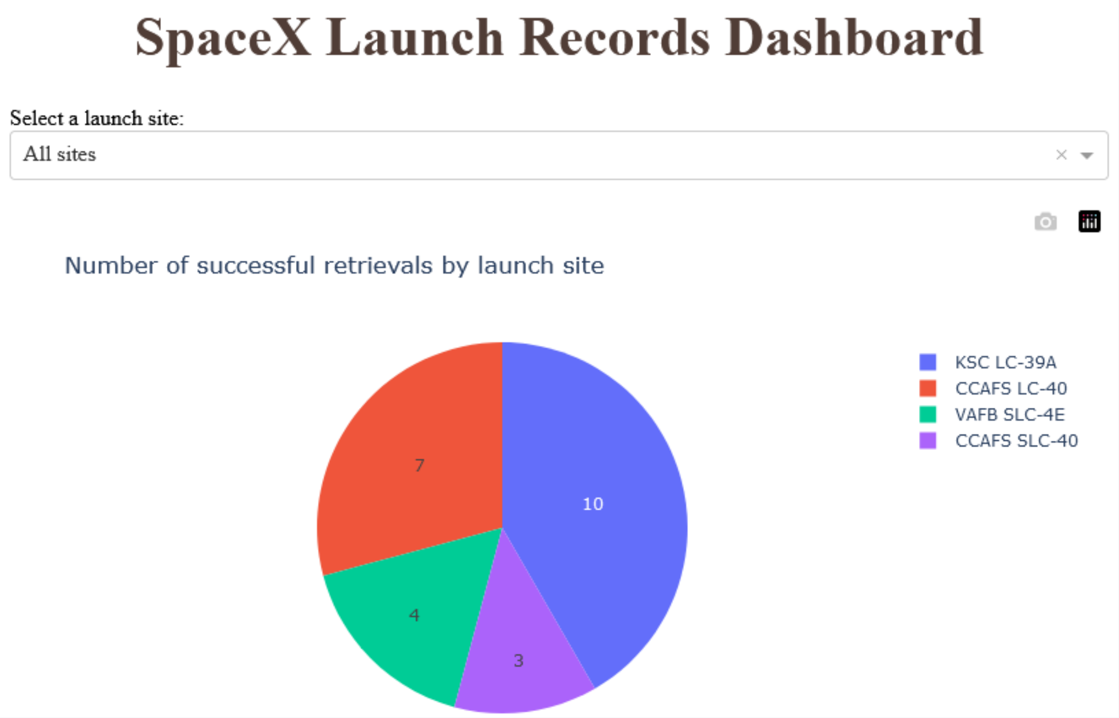


Section 4

Build a Dashboard with Plotly Dash

Successful landings per launch site

- Comparing all launch sites, it can be seen that, as stated before, the Kennedy Space Center has the highest number of successful landings, having 10 of them.
- Closely followed by the Cape Canaveral Air Force Station with 7 successful landings.

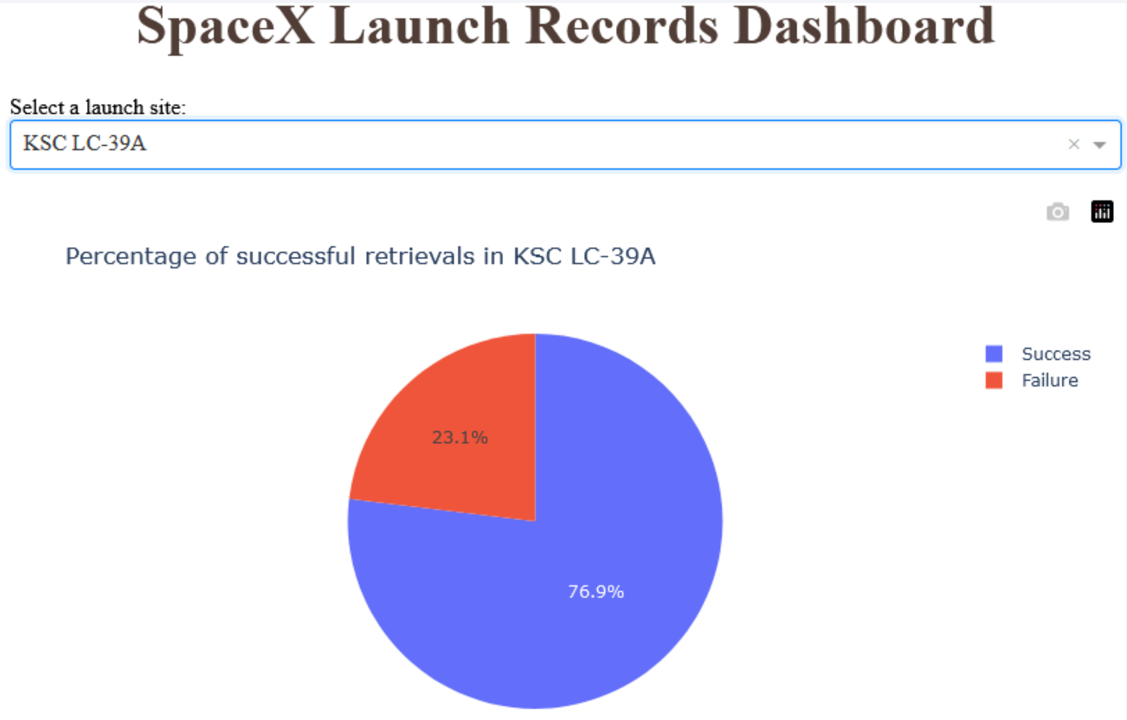


SpaceX Launch Records Dashboard

Select a launch site: All sites

Deeper analysis of KSC LC-39A

- Selecting KSC LC-39A it can be seen that 77 % of launches made here had a successful landing.



SpaceX Launch Records Dashboard

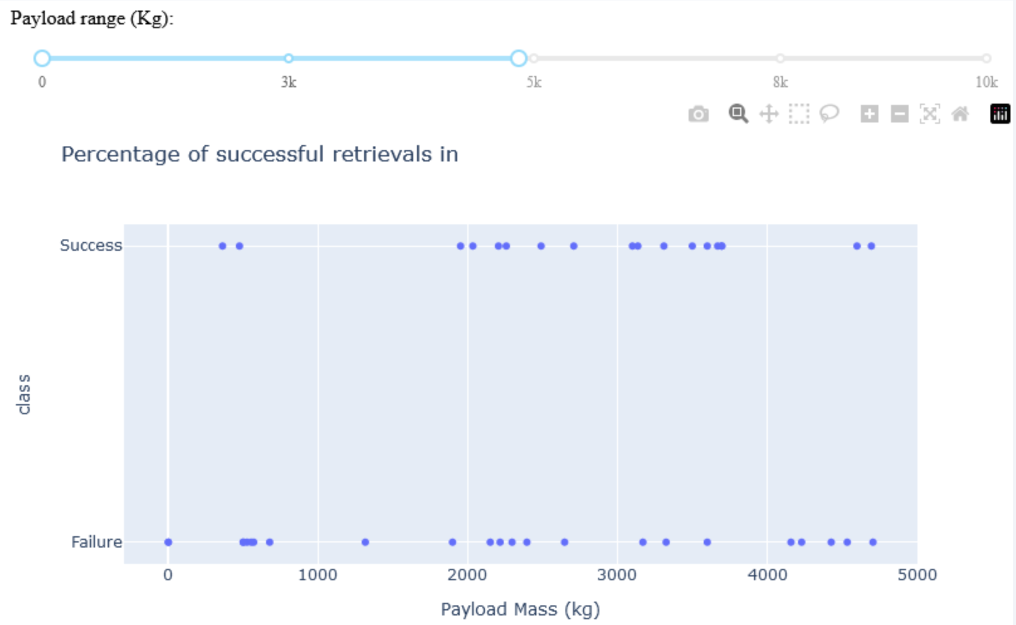
Select a launch site:

KSC LC-39A

Percentage of successful retrievals in KSC LC-39A

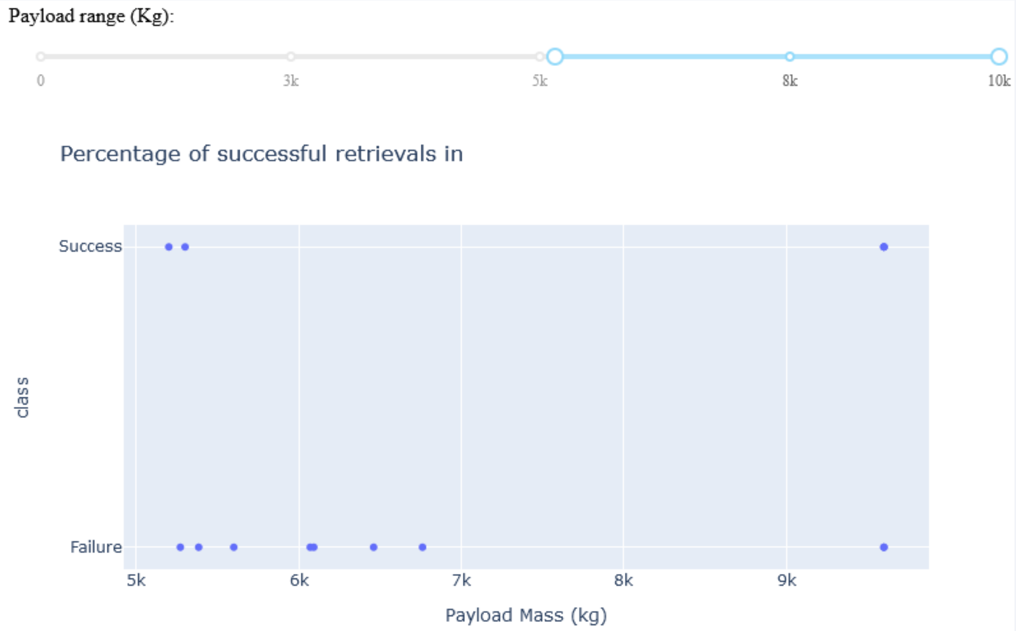
Successful landings according to their payload mass

- The dataset used for this dashboard has 56 records, the SpaceX's first 56 flights. As stated before, the success rate wasn't great in the beginning, and accordingly many landings ended in failure. This is especially critical in the payload range of 5k to 10k kg, where only three of them were successful.



Payload range (Kg):

Percentage of successful retrievals in



Payload range (Kg):

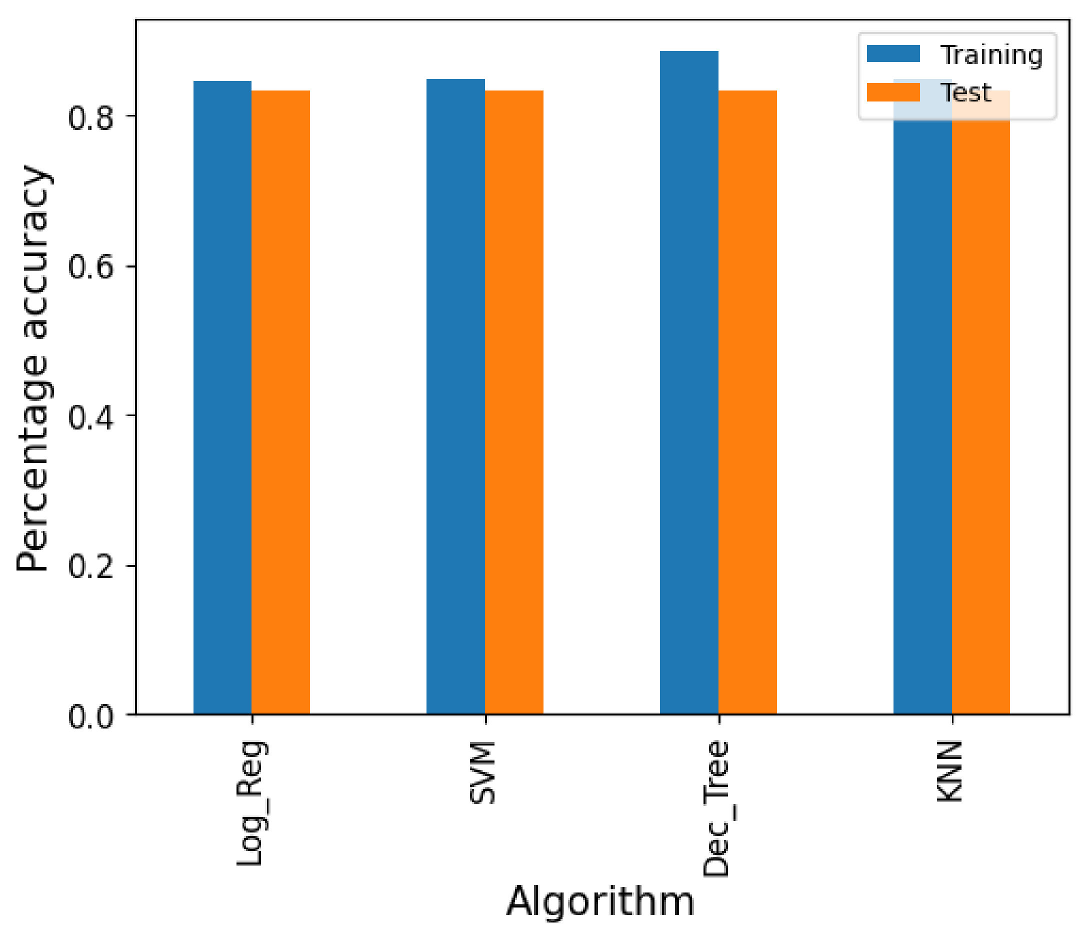
Percentage of successful retrievals in

Section 5

Predictive Analysis (Classification)

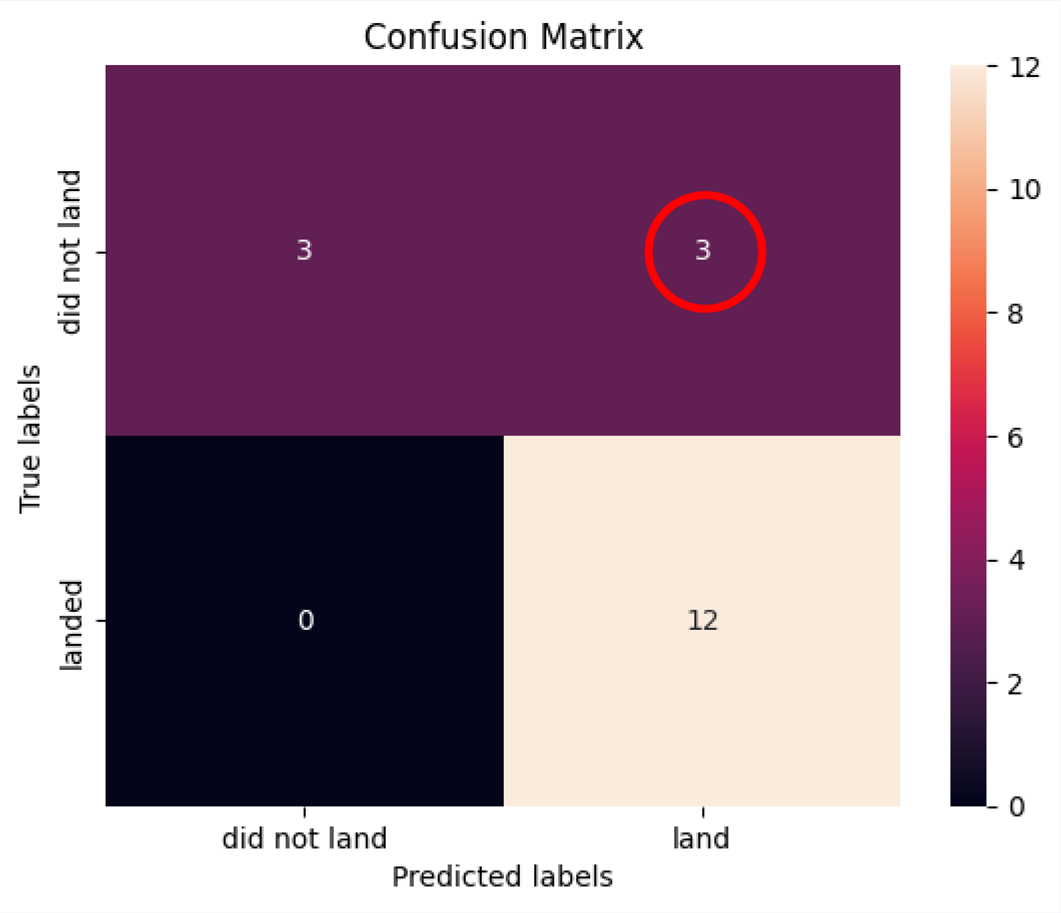
Classification Accuracy

- Decision tree had the best performance when using training data, 88.57 % accuracy. However, when using testing data **all** the models had the same performance, 83.33 % accuracy



Confusion Matrix

- All four models had 3 false positive predictions. The models predicted a successful landing, but in reality, it did not land. These cases are circled in red.



Conclusions

- SpaceX success rate increased with the passage of time independently of the flight parameters, which points to the importance of the expansion of the team's expertise to had better odds
- There is growing costumer interest in the VLEO as target orbit. Efforts must be made to optimize the launching and landing parameters for this orbit.
- Kennedy Space Center had the highest landing success rate, which makes it the preferred launch site.
- Decision tree had the **best performance** using training data. Since all models had the **same performance** when using testing data, decision tree is preferred. Its accuracy percentage was **88.57 and 83.33 %** when using training and testing data, respectively.

Thank you!