IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mykola Purdes
22.11.2024

Outline

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

Executive Summary

- Summary of methodologies we used:

✓ Data Collection from site using API and Web Scraping

✓ Data Wrangling with Python

✓ Exploratory Data Analysis using SQL

✓ Exploratory Data Analysis with Visualization using Pandas and Matplotlib libs

✓ Interactive Visual Analytics with Folium and Plotly Dash

✓ Making Predictions using Machine Learning

- Summary of all results:

✓ Data Collection from Public Recourses, Preprocessing and Cleaning of the Data about Launches

✓ Defining the Most important Factors of a Launch Success with Exploratory Data Analysis, Visualizations and Dashboard

✓ Booster Landing Prediction using Machine Learning

Introduction

- **Project background and context**

In this project we predict if the Falcon 9 first stage (or Booster) will land successfully. The first stage of the rocket is the most expensive and is crucial factor for determining the cost of the rocket launch. Therefore, if we able to predict the success of the first stage landing and probability of the success, we will be able to define the right cost of the launch and make competitive price offer for customers. Also, we can define main factors that influence the launch success the most.

- **We want to find answers to the next problems**

- Define and analyze the factors that are essential for the first stage landing success.
- Define the launch sites proximity.
- Making prediction of the success of the launch and its accuracy.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

SpaceX launch data is gathered from the SpaceX REST API using Python request library.

We also gathered data about SpaceX launches by Web scraping Wiki site «List of Falcon 9 and Falcon Heavy launches» with Python BeautifulSoup package.

- Perform data wrangling

We performed some Exploratory Data Analysis (EDA) to get some general information about places of launches, orbits and outcomes and to find some patterns in the data.

We also determined landing outcomes as training labels and marked them with 0 and 1.

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

How to build, tune, evaluate classification models

Data Collection

We collected data about Falcon Launches through API and by Web Scraping.

1. Collecting data through SpaceX API.

We used Python request library to get data from SpaceX site. Our response was in the form of a list of JSON objects, which we converted to Pandas dataframe with json_normalize function. Then we cleaned the data: removed unnecessary data, extracted and transformed data to get only useful and needed features.

Then we filtered the data to get only data about Falcon 9 spaceship.

2. Collecting data by Web Scraping.

We used Python BeautifulSoup package to request and get HTML Wiki page, extract table about Falcon9 Launches and cleaned it with custom Python functions. Then we created the Pandas dataframe and assigned extracted and cleaned information to its columns.

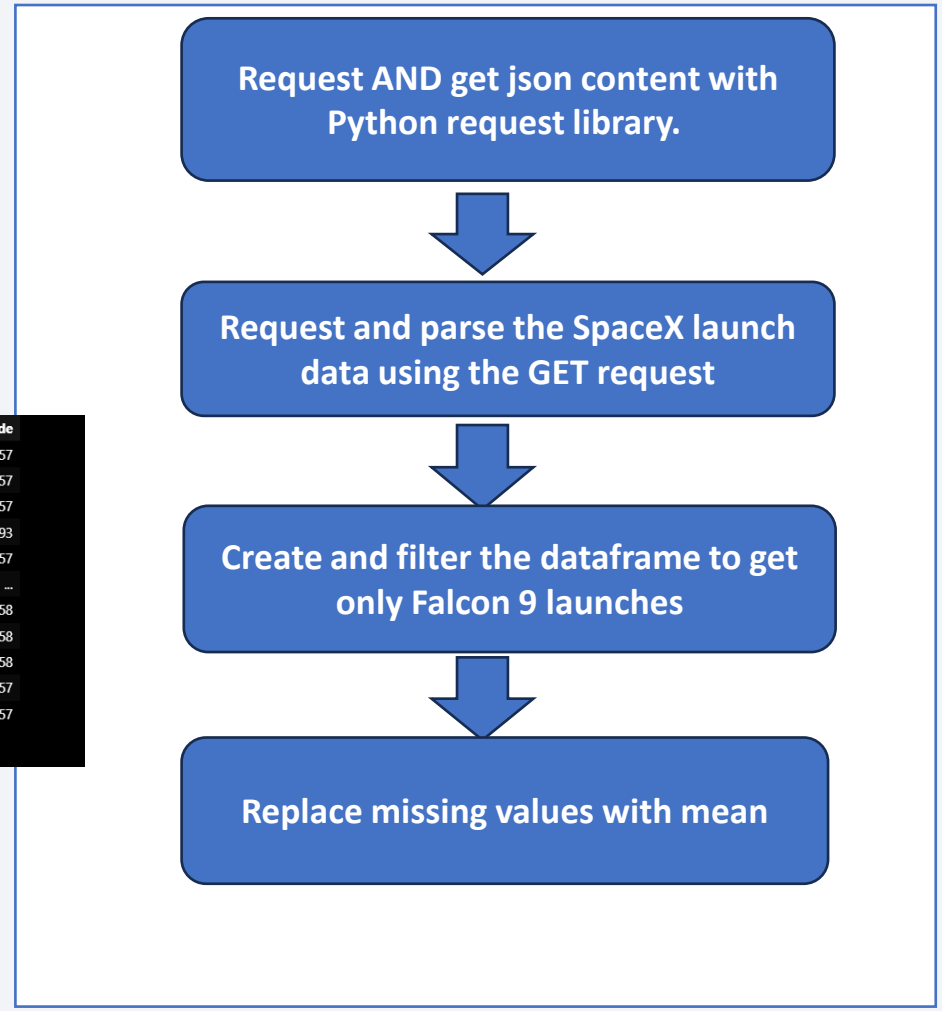
Data Collection – SpaceX API

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

Collected and cleaned data about Falcon 9 launches:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060	-80.603956	28.608058
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058	-80.603956	28.608058
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1051	-80.603956	28.608058
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	B1060	-80.577366	28.561857
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1062	-80.577366	28.561857

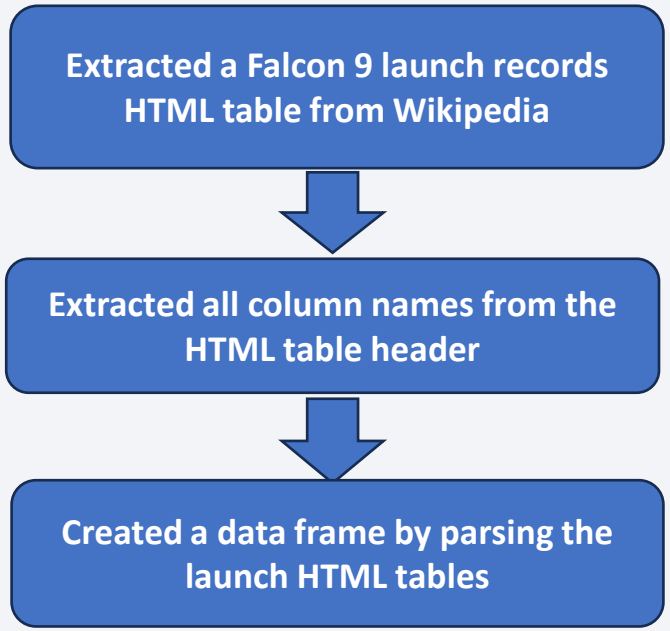
90 rows × 17 columns



- GitHub URL: https://github.com/NickPurdes/CourseraProjects/blob/main/SpaceX-%20Data%20Collection-API.ipynb

Data Collection - Scraping

- The launch records are stored in a HTML table on https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches site
- We webscraped Falcon 9 launch records with BeautifulSoup library



- GitHub URL: https://github.com/NickPurdes/CourseraProjects/blob/main/SpaceX-%20Data%20Collection-%20Webscraping.ipynb

Data Wrangling

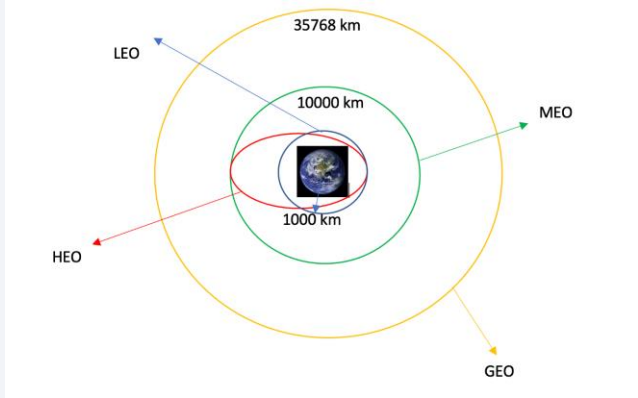
We performed Exploratory Data Analysis (EDA) and determined Training Labels

- EDA:

a)Calculated the number of launches on each site

b)Calculated the number and occurrence of each orbit

c)Calculated the number and occurrence of mission outcome of the orbits



- Created training labels from Outcome column

EDA:
- number of launches on each site
- number and occurrence of each orbit
- number and occurrence of mission outcome

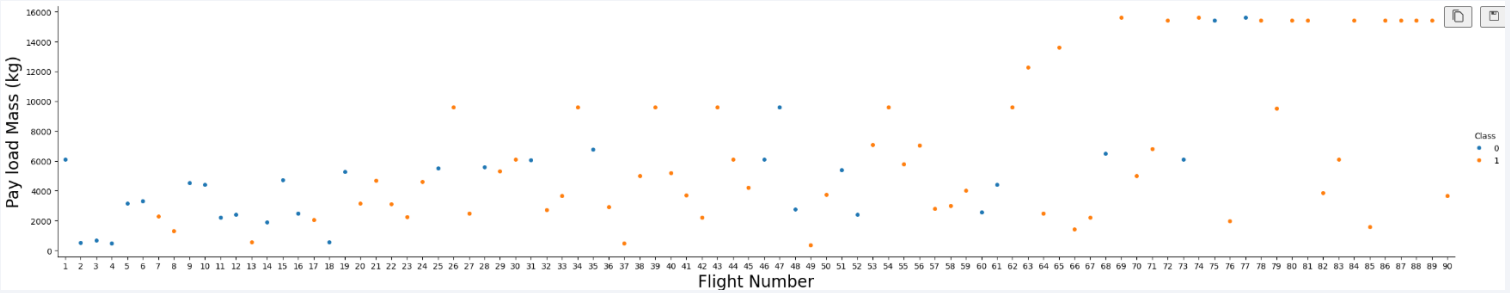
Creating training labels:
- Define class as 1 if successful landing and 0 if not
- Append classes to the Table

GitHub URL: https://github.com/NickPurdes/CourseraProjects/blob/main/SpaceX%20-%20Data%20Wrangling.ipynb

EDA with Data Visualization

We used scatterplots and barplots to visualize the relationship between pairs of features:

- Flight Number vs Payload Mass



- Flight Number vs Launch Site
- Launch Site vs Payload Mass
- Orbit Type vs Flight Number
- Payload vs Orbit Type

We used linear graph to observe Launch Success Yearly Trend

GitHub URL: https://github.com/NickPurdes/CourseraProjects/blob/main/SpaceX-%20EDA-Visualization.ipynb

EDA with SQL

I performed SQL queries on SpaceX Table to define:

- the names of the unique launch sites in the space mission
- 5 first records where launch sites begin with `CCA`
- the 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
- the total number of successful and failure mission outcomes
- the names of the booster versions which have carried the maximum payload mass. Use a subquery
- the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
- 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

GitHub URL: https://github.com/NickPurdes/CourseraProjects/blob/main/SpaceX-%20EDA-SQL%20SQLite.ipynb

Build an Interactive Map with Folium

Map objects such as markers, circles, PolyLine, mouse position, marker clusters were created and added to a folium map to execute launch sites proximities analysis:

- Markers were created to add each site's location on a map using site's latitude and longitude coordinates
- Circles are used to add a highlighted circle area with a text label on a specific coordinate
- Marker clusters are used to add more visual details about launches – how many were successful and how many were not
- Mouse position is used to define coordinates of any object on the Folium map (in the high right corner)
- PolyLine is used to draw a line between two map objects

GitHub URL https://github.com/NickPurdes/CourseraProjects/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Two interactive plots/graphs were added to a dashboard:

- Launch success ratio pie chart to analyze overall launch success and success rate for each area individually. It helps to see more details and to compare different results.
- Payload vs. Launch Outcome scatter plot with possibility to tune payload range in order to see more clear and detailed information on the scatter plot. It also may be useful if we want to define which boosters are more effective and for which payload masses.

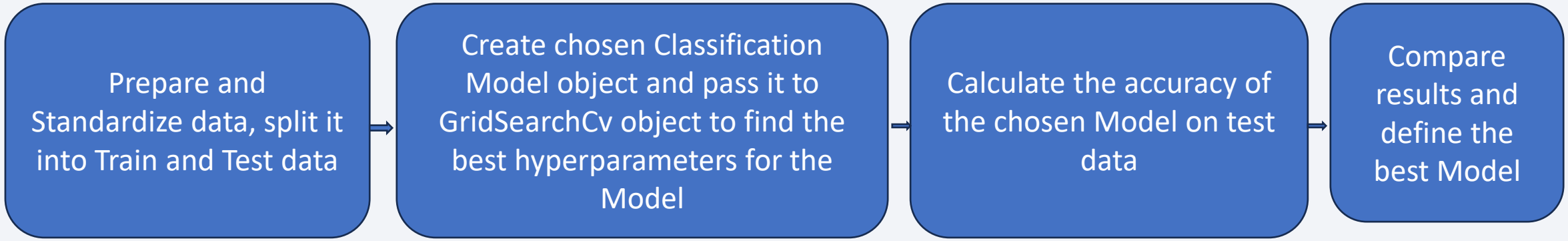
GitHub URL https://github.com/NickPurdes/CourseraProjects/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

We used Machine Learning to define the best classification model among:

- Logistic regression
- SVM
- Decision tree
- K nearest neighbors

We followed the next procedure:



GitHub URL
https://github.com/NickPurdes/CourseraProjects/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

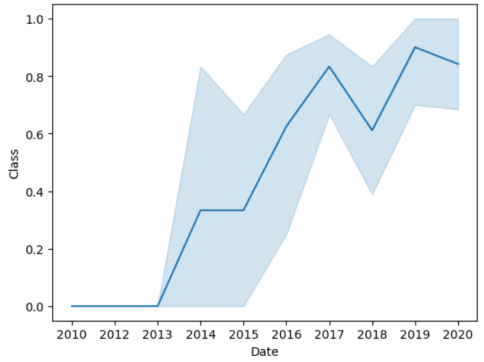
Exploratory data analysis results:

- the success rate since 2013 kept increasing till 2020, thus flight number (time and experience in other words) is crucial factor for launch success compared to others – launch site, orbit or payload mass.
- Space X used 4 different launch sites
- The total payload carried by boosters from NASA is 45596 kg – SpaceX major client
- The average payload mass carried by booster version F9 v1.1 is 2928,4 kg
- The first successful landing outcome on ground pad was on 2015-12-22 – five years later after the first launch
- Only 1outcome was a Failure, other 100 were successful
- All the boosters which have carried the maximum payload mass were of B5 category
- The boosters failed at landing in drone ships in 2015 were F9 v1.1 B1012 and F9 v1.1 B1015
- There were 10 cases with no attempts to land a booster

Results

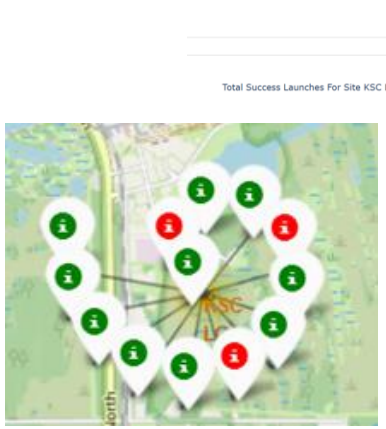
Interactive analytics demo in screenshots:

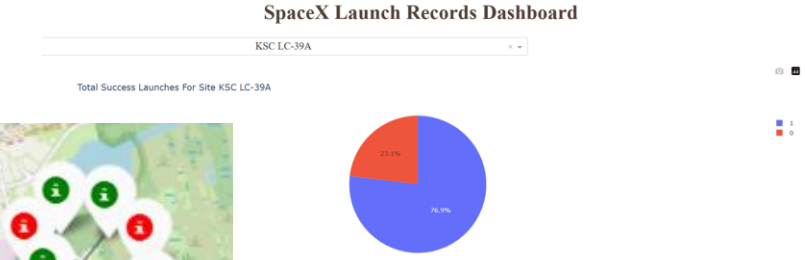
- The success rate since 2013 kept increasing till 2020. Only in 2018 success rate seriously fallen from 0,8 to 0,6.



- All launch sites located in the USA as close to the Equator line as possible to use the maximum initial Earth's speed. We also see that they all are near sea to provide safety and landing of the first stage on sea platform when necessary.
- 10 launches where from Lompoc area and 46 launches where from area near Cape Canaveral
- Launch sites is very close to railway and highway to provide the launch with all necessary materials and good logistic.

- KSC LC-39A area is a site with the highest launch success ratio. 76,9% of launches where successful from this place



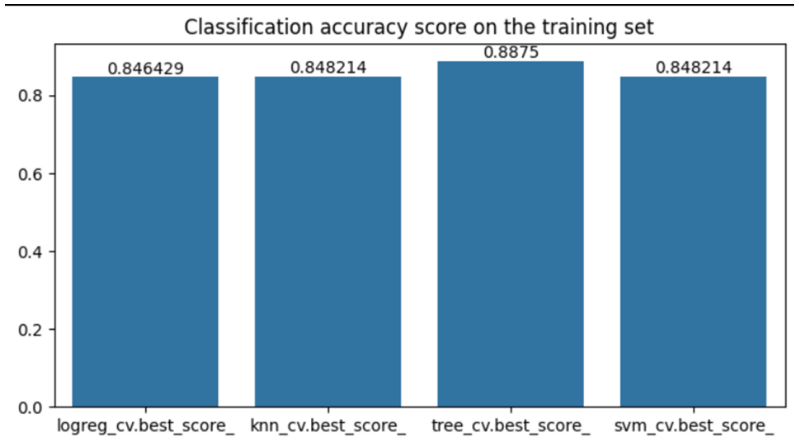


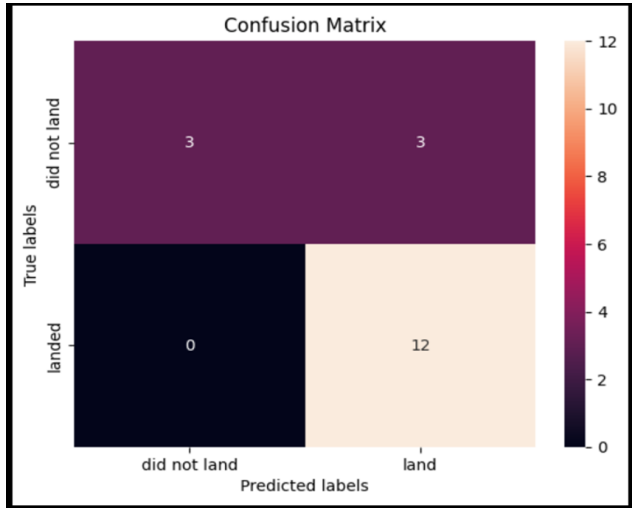
Results

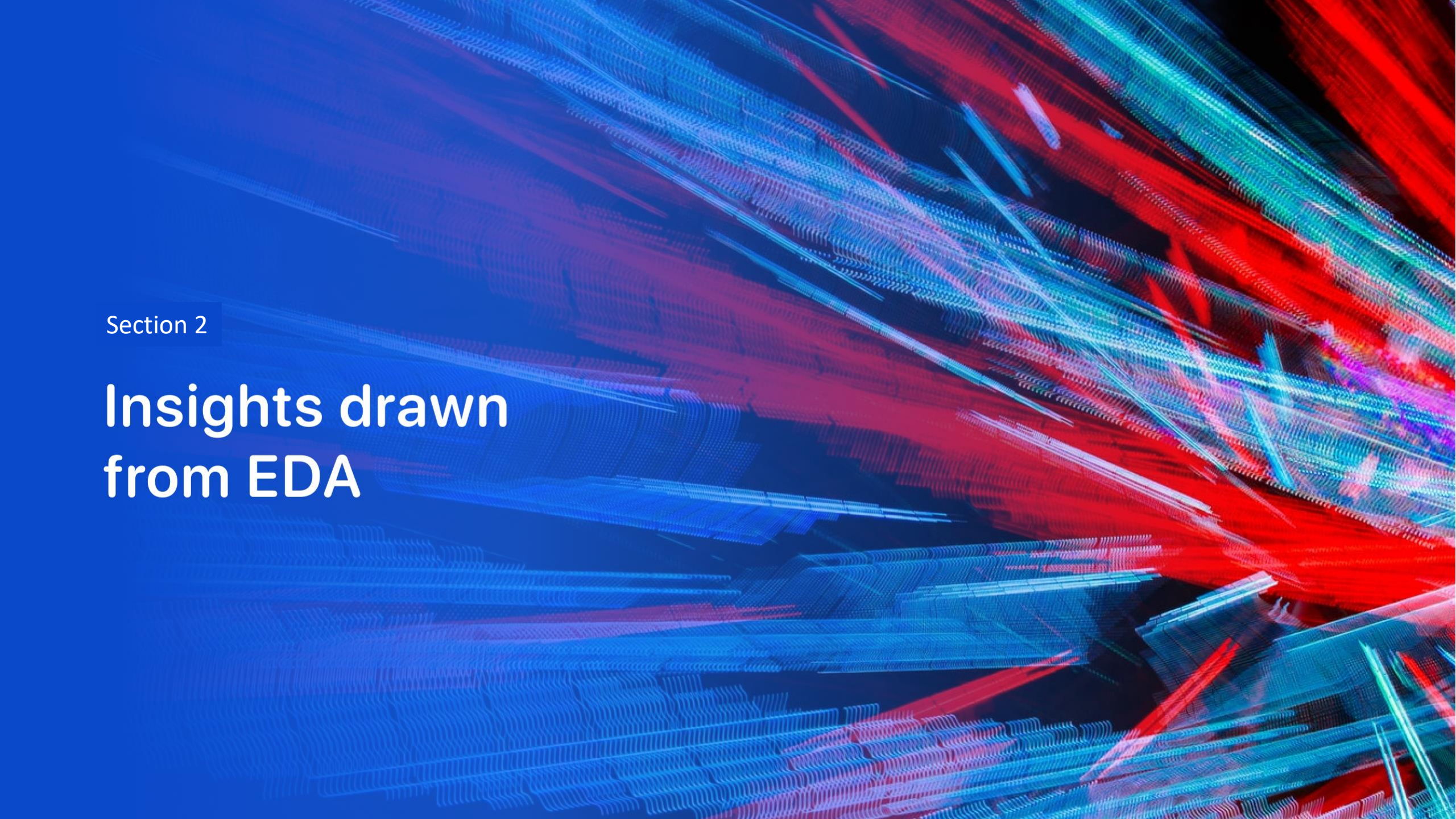
Predictive analysis results:

- Classification with Decision Tree has the highest classification accuracy score on the training set 88,75%. Because all Models have the same accuracy on the test 83,33%, we can choose this algorithm as the most promising classifier for our model

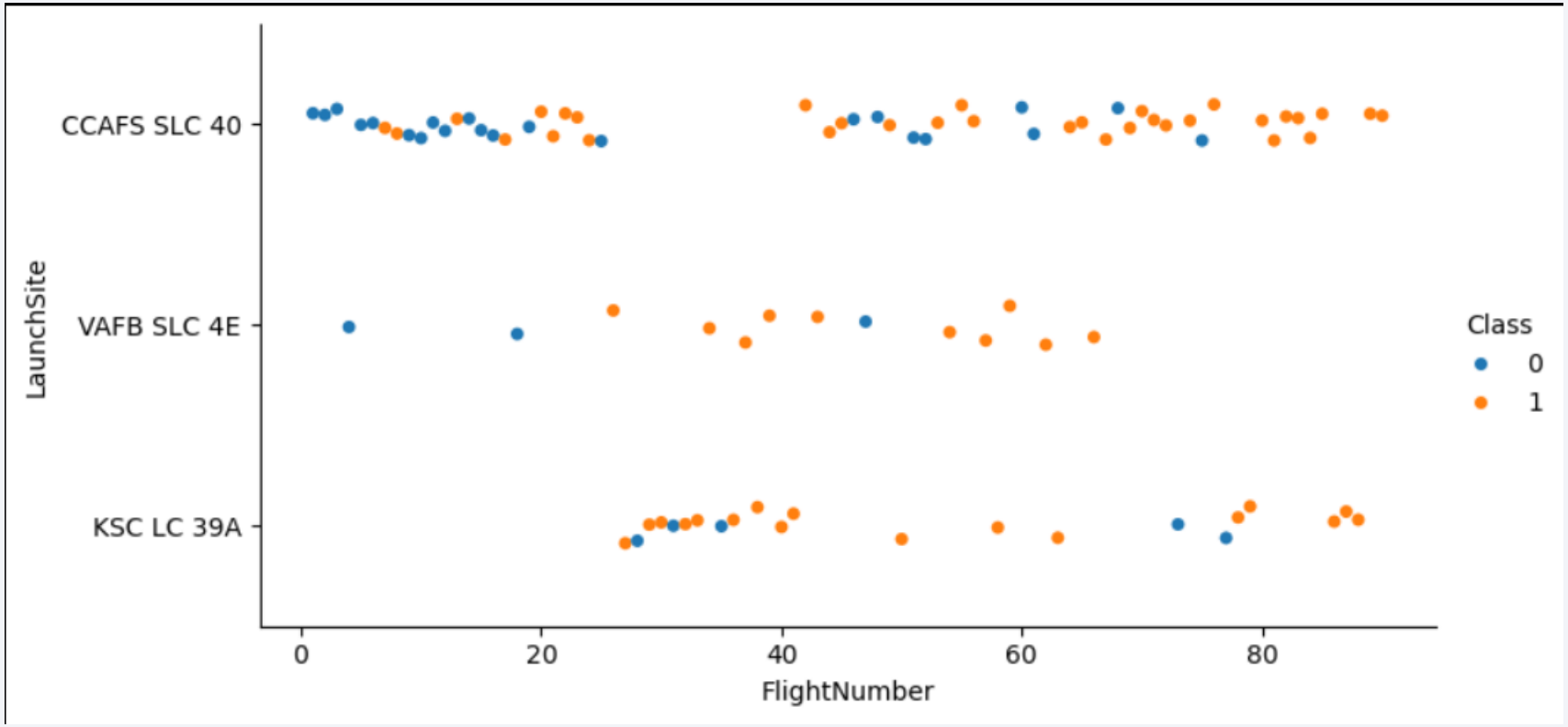
- The confusion matrix for Decision Tree classifier shows that all 12 successfully landed boosters where predicted right. But from 6 booster that where not landed 3 where marked as successfully predicted. So, accuracy score on test data for Decision Tree is 83,33% (as well as for other classifiers)





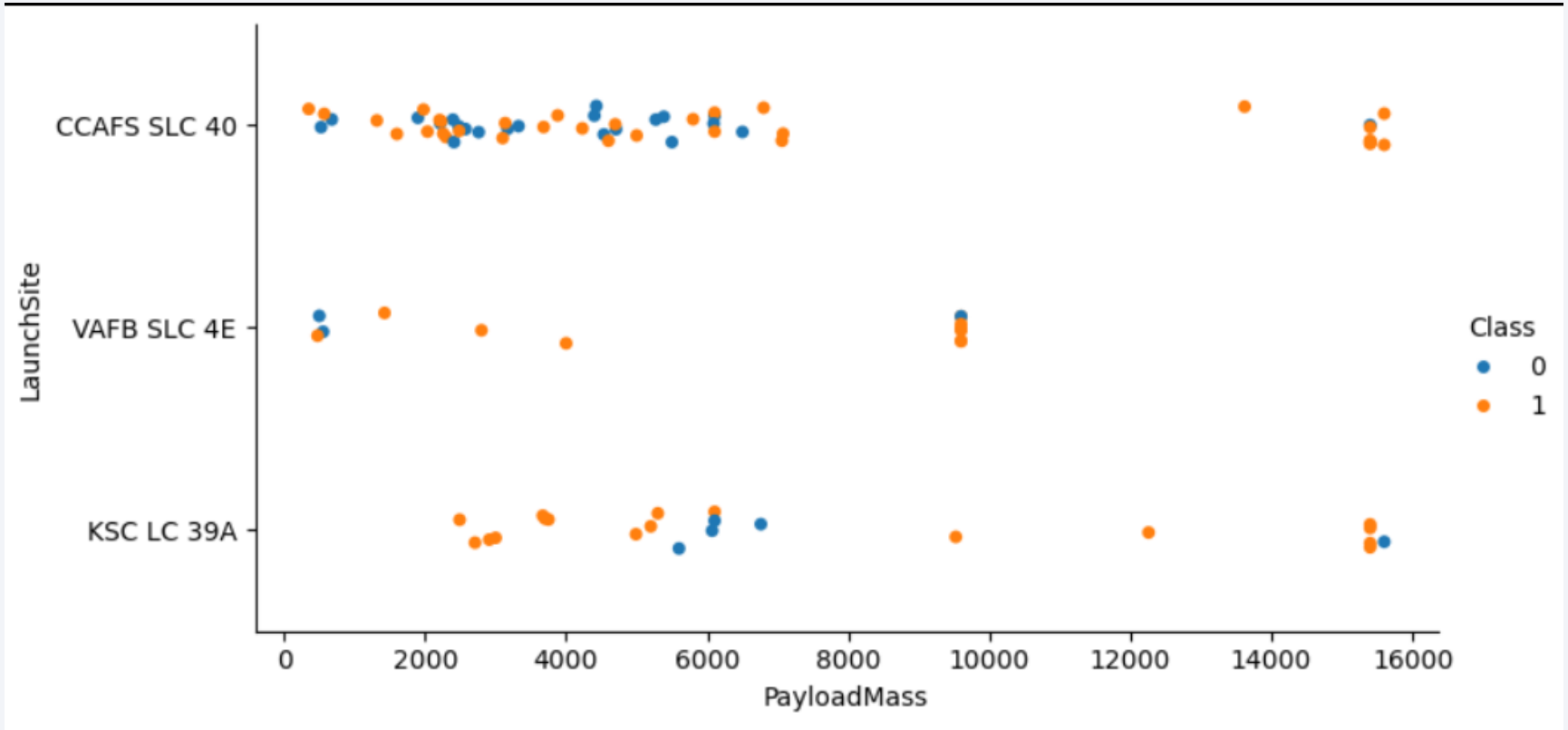

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



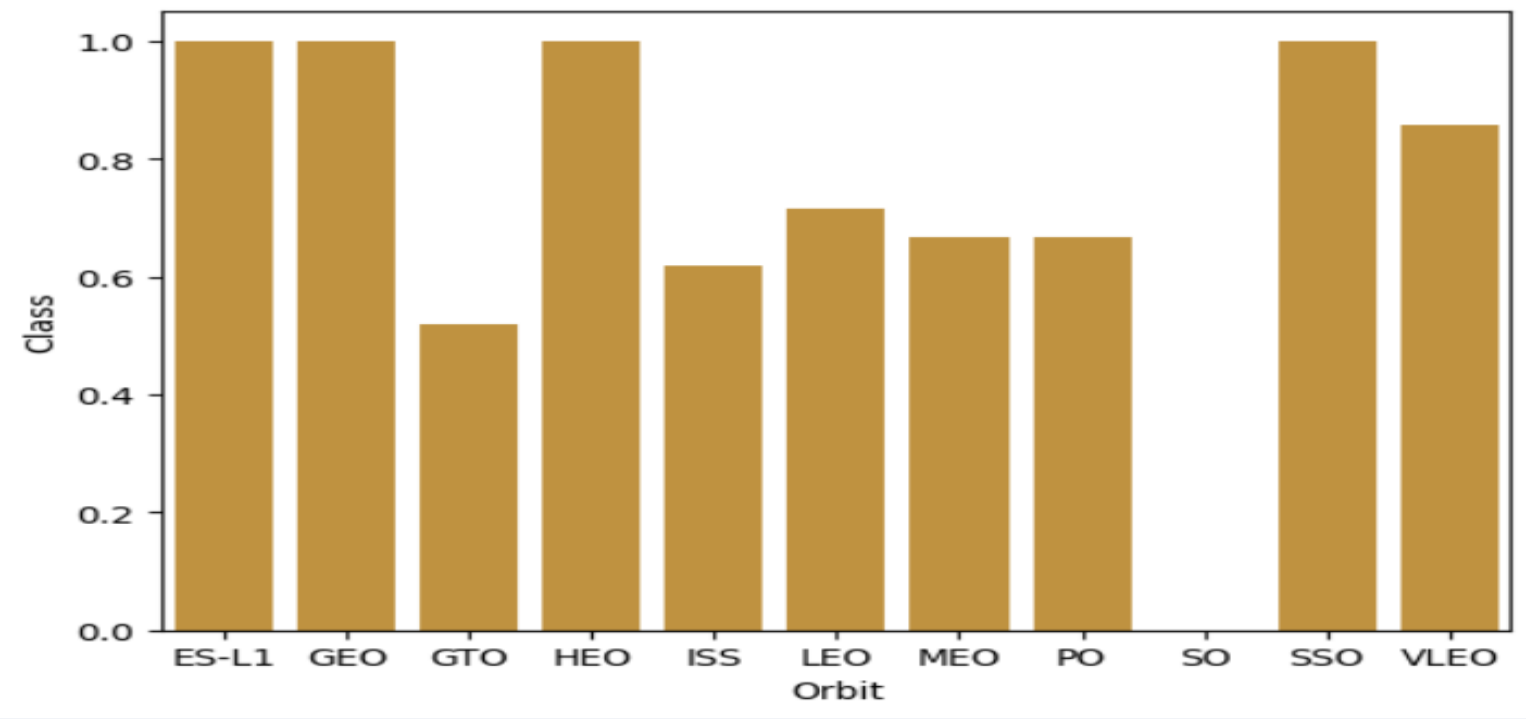
- Class 1 belongs to successful outcome of a mission. We clearly see that when Flight Number growth the first stage is more likely to land successfully for all Launch Sites. Therefore a Flight Number is important factor of success and a Launch Site is not.

Payload vs. Launch Site



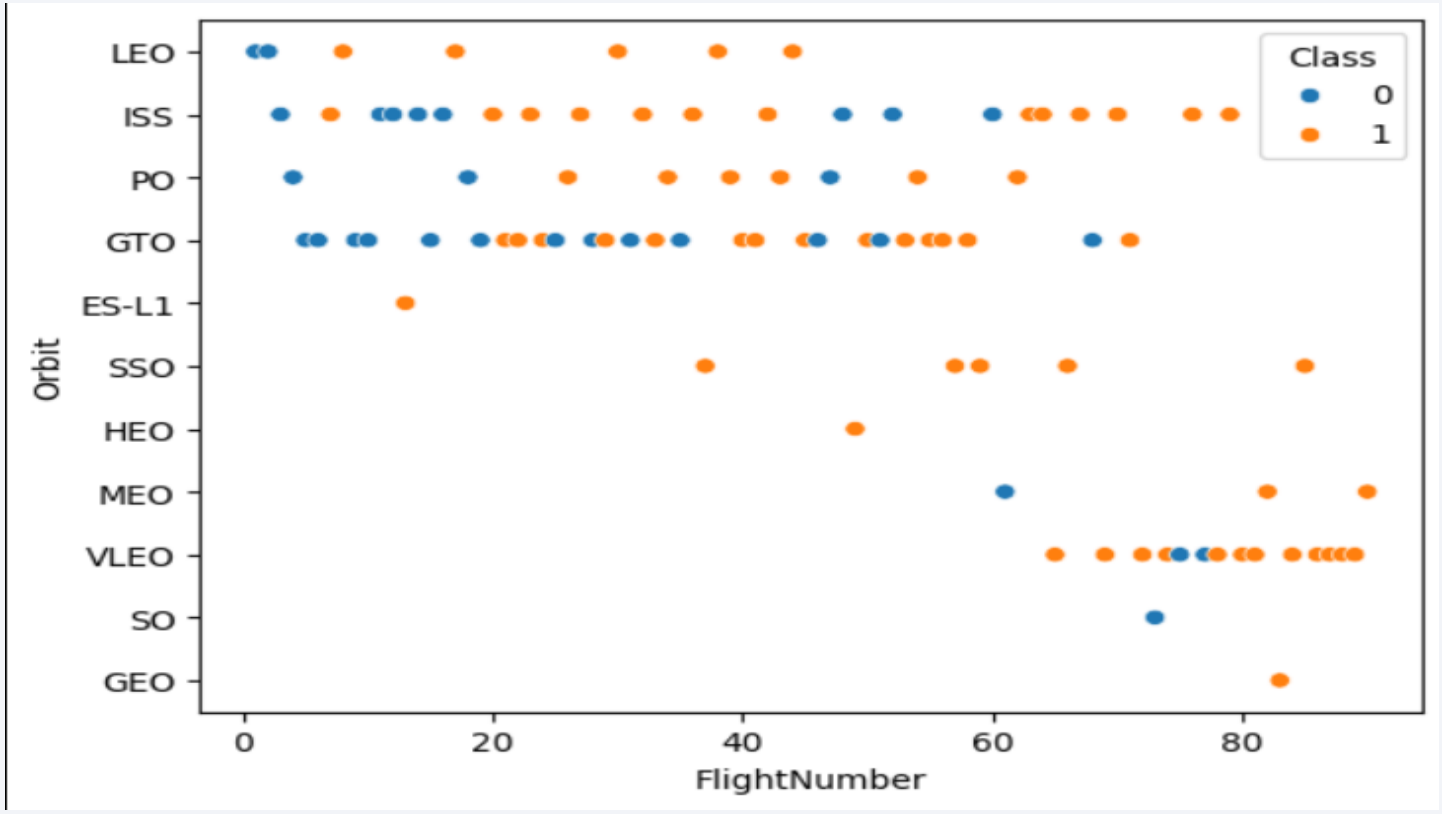
- We see that for the VAFB-SLC launch site there were no rockets launched for payload mass greater than 10000 kg
- With payload mass increase more than 8000 kg the possibility of success increases too for all launch sites

Success Rate vs. Orbit Type



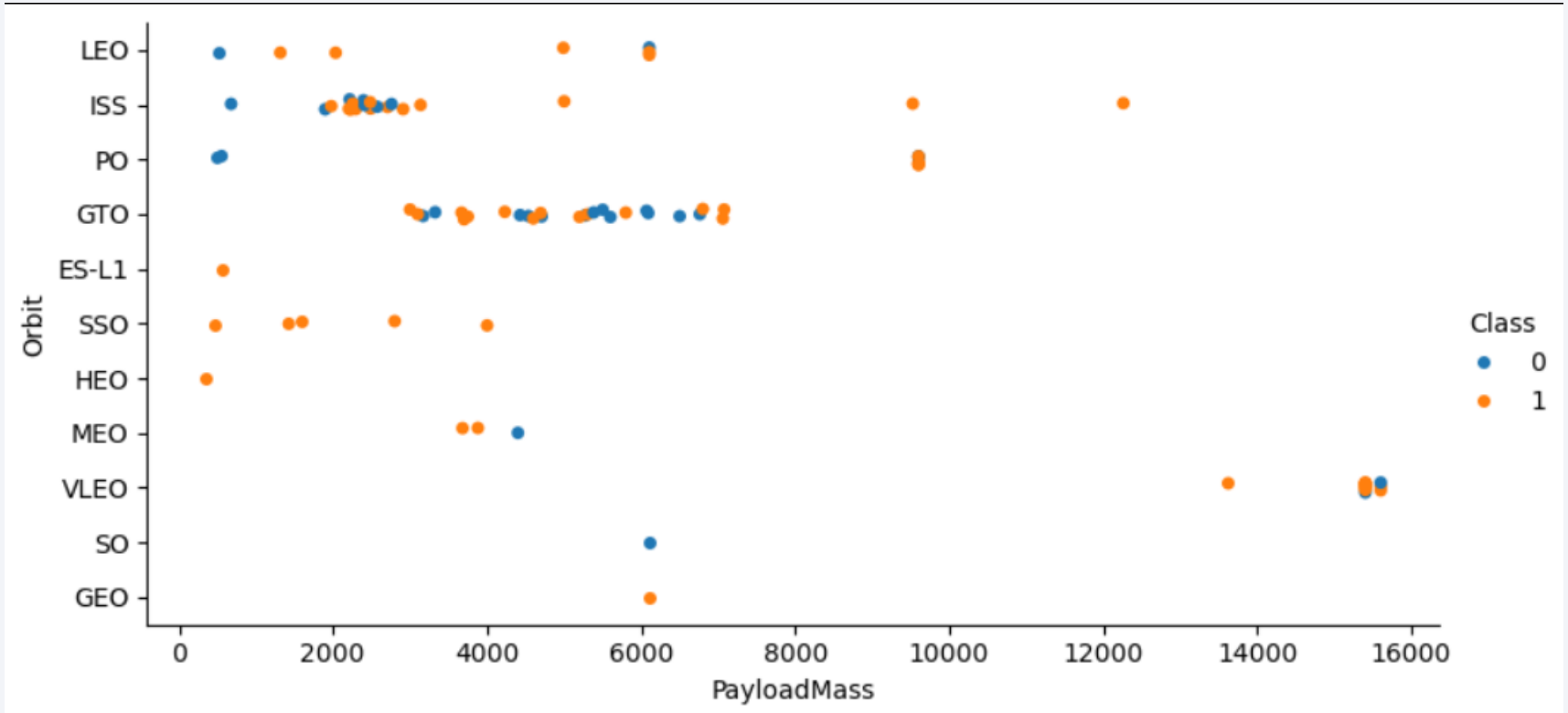
ES-L1, GEO, HEO and SSO orbits have the highest success rate 1, which means all launches to these orbits were successful, GTO is the least successful orbit with success rate near 0,5 (only 50% launches were successful).

Flight Number vs. Orbit Type



While some orbits are 100% successful (3 of 4 these had only 1 launch!), with flight number growth other orbits are becoming more successful too. We see that first launches were mostly unsuccessful while all launches after 80th were all successful for all orbit types. Hence flight number is much more important factor of success than orbit.

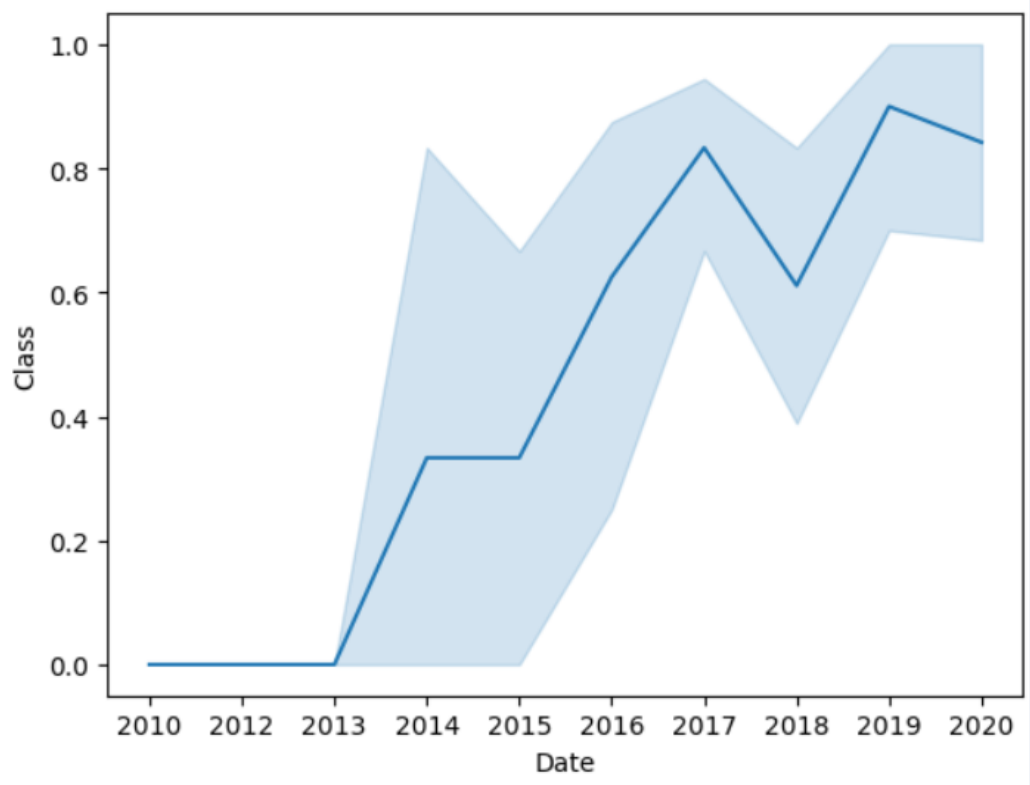
Payload vs. Orbit Type



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

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



We can observe that the success rate since 2013 kept increasing till 2020. Only in 2018 success rate seriously fallen from 0,8 to 0,6.

All Launch Site Names

The names of the unique launch sites are

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

I used next query with DISTINCT function

%sql select distinct Launch_Site from SPACEXTABLE

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

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

SQL query used

%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5

First 2 launches were demo without any payload and landing failure, next 3 were without attempt to land

Total Payload Mass

The total payload carried by boosters from NASA

total payload by NASA(CRS)
45596

I used next query with sum function and filter ‘where’ for NASA (CRS)

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

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2928,4 kg

avg payload mass for booster F9 v1.1
2928.4

I used next query with avg function

```
%sql select avg(PAYLOAD_MASS__KG_) as  'avg payload mass for booster F9 v1.1' from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

First Successful Ground Landing Date

The first successful landing outcome on ground pad was on 2015-12-22

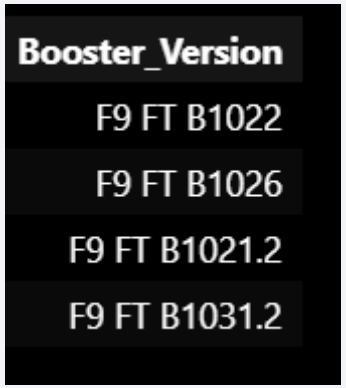
first succesful landing outcome in ground pad date
2015-12-22

- I used next query with min function

```
%sql select min(Date) as  'first succesful landing outcome in ground pad date' from
SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are



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

I used next query

```
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone
ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

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

I used grouping in my query

%sql select Mission_Outcome, count(Mission_Outcome) as Outcome_Sum from SPACEXTABLE group by Mission_Outcome

We can easily see that only 1outcome marked as Failure, other 100 (98+1+1) are successful

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass are

I used next query with subquery

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_=(select
max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

All the boosters which have carried the maximum payload mass were of B5 category

2015 Launch Records

List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

I used next multirow query

```
%%sql  select substr(Date,6,2) as month, Date ,Booster_Version, Launch_Site, Landing_Outcome
        from SPACEXTBL where Landing_Outcome='Failure (drone ship)' and substr(Date,0,5)='2015'
```

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

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

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

- I used next query with grouping and filtering

```
%%sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE  where Date BETWEEN '2010-06-04' AND '2017-03-20'  group by Landing_Outcome  order by count(Landing_Outcome) desc
```

As we see there were 10 cases with no attempts to land a booster

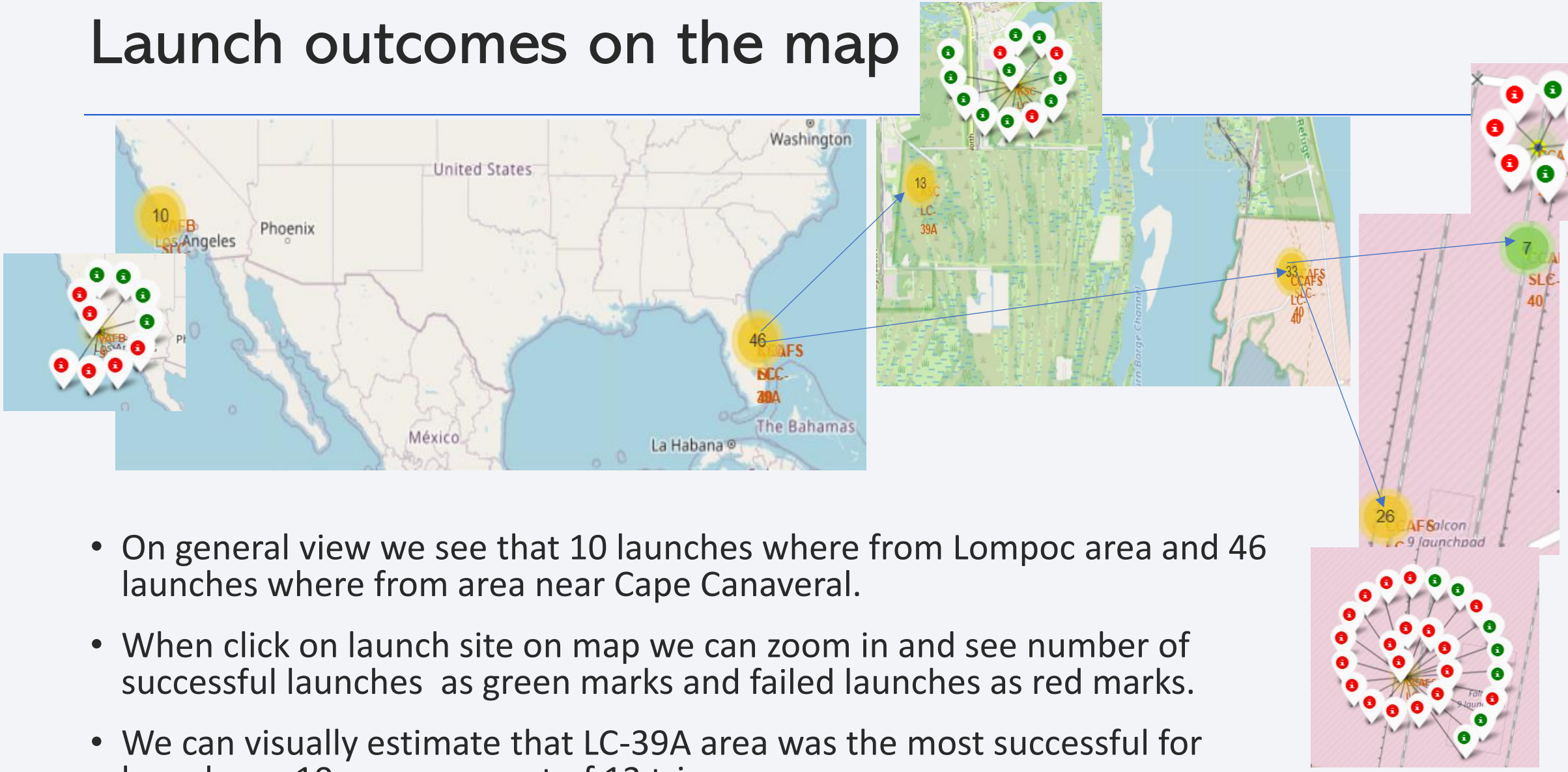
Section 3

Launch Sites Proximities Analysis

All launch sites' location

- All launch sites located in the USA as close to the Equator line as possible to have the maximum initial Earth's speed. We also see that they all are near sea to provide safety and landing of the first stage on sea platform when necessary. Three of four sites are very close to each other.

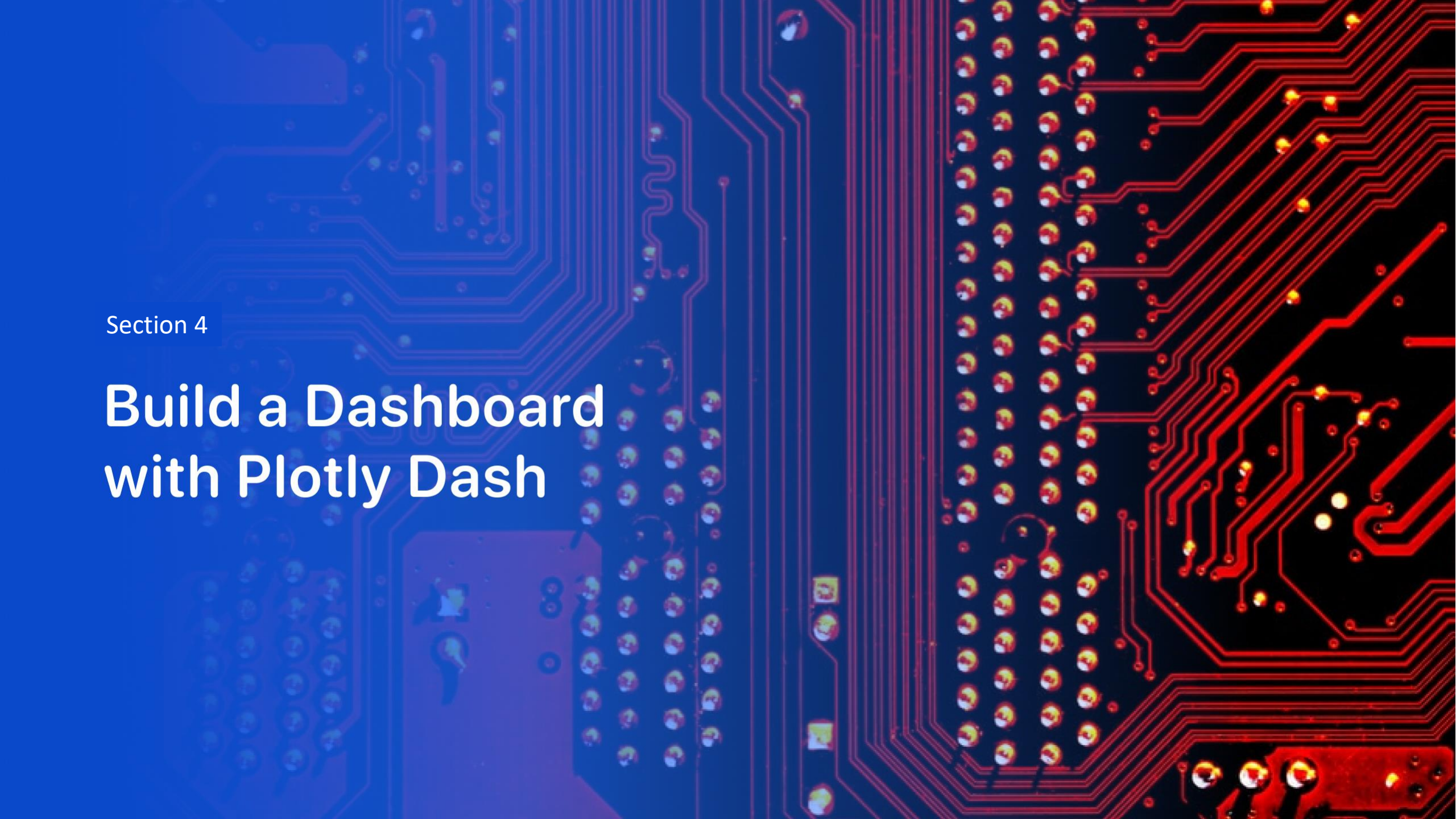
Launch outcomes on the map



- On general view we see that 10 launches where from Lompoc area and 46 launches where from area near Cape Canaveral.
- When click on launch site on map we can zoom in and see number of successful launches as green marks and failed launches as red marks.
- We can visually estimate that LC-39A area was the most successful for launches – 10 successes out of 13 tries.

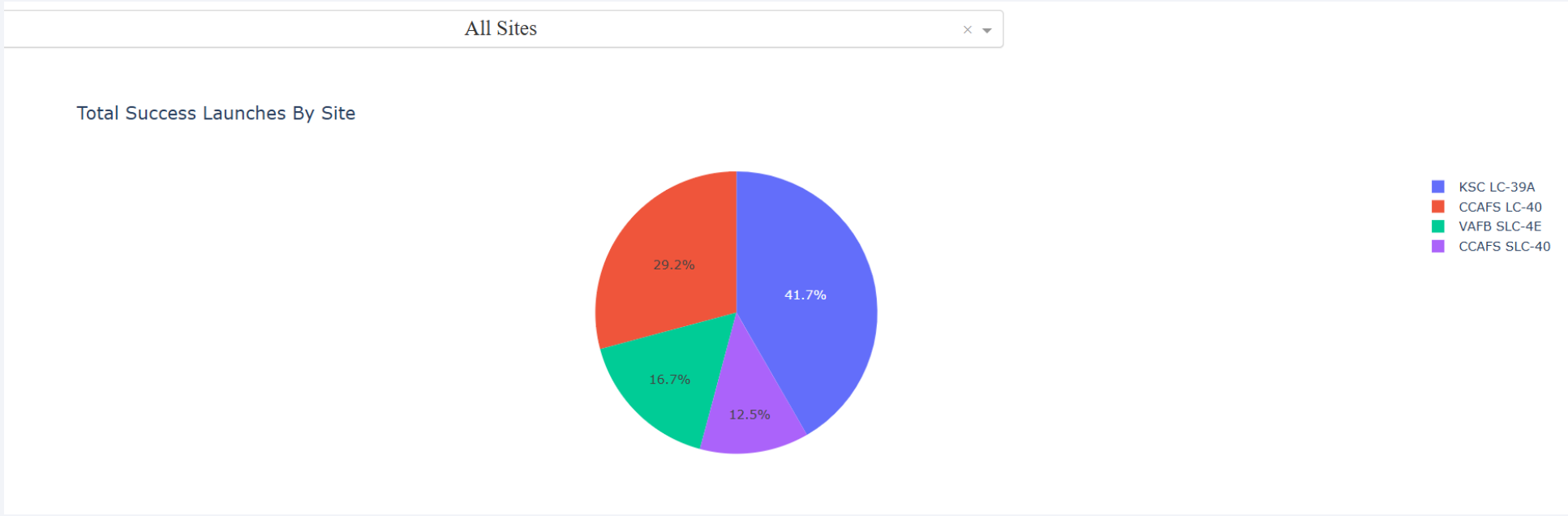
SLC-4E launch site infrastructure

- On the map we can see that SLC-4E launch site is very close to railway and highway (near 1 - 1,25 km) to provide the launch with all necessary materials and good logistic. It is also close to seashore line (1,4 km) to provide launches safety and booster landing in the sea if needed.


Section 4
Build a Dashboard with Plotly Dash

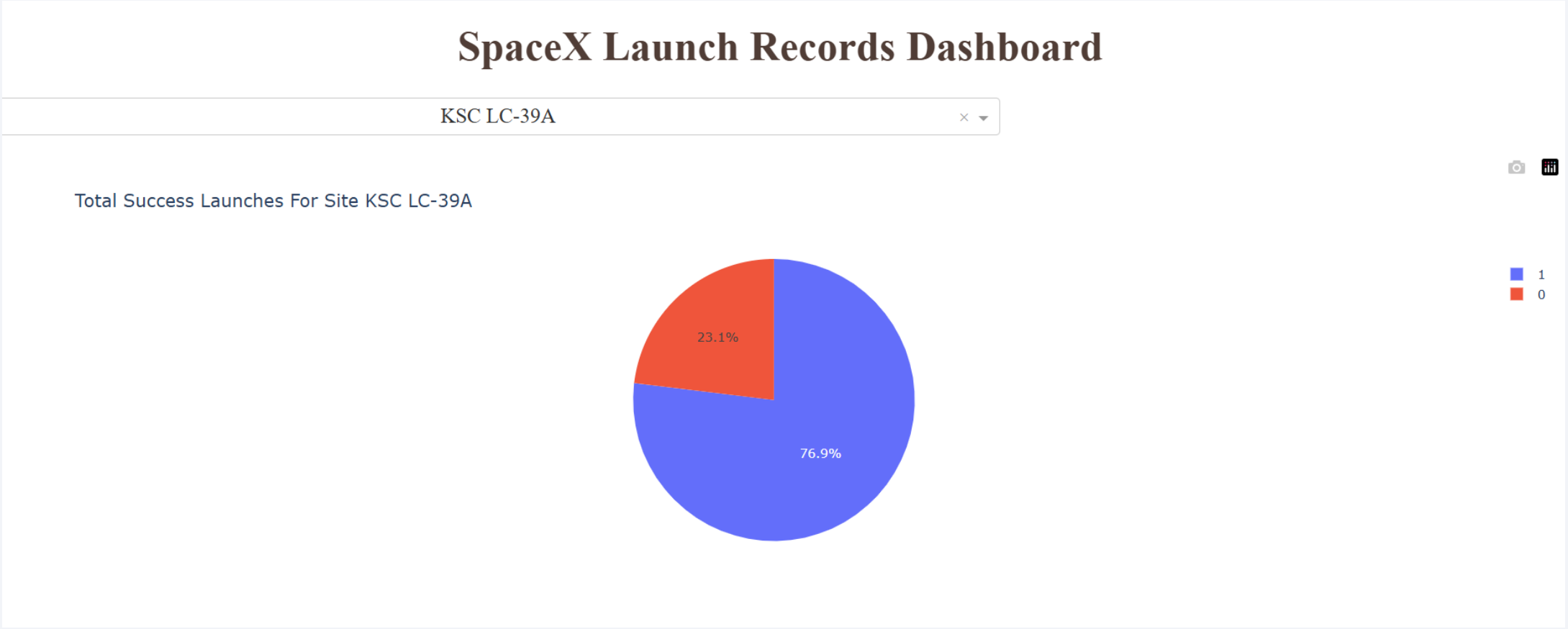
SpaceX Launch Records Dashboard

Total Success Launch By Site



The launch site seems to be an important factor for the mission success. The KSC LC-39A area is the most successful place for a launch.

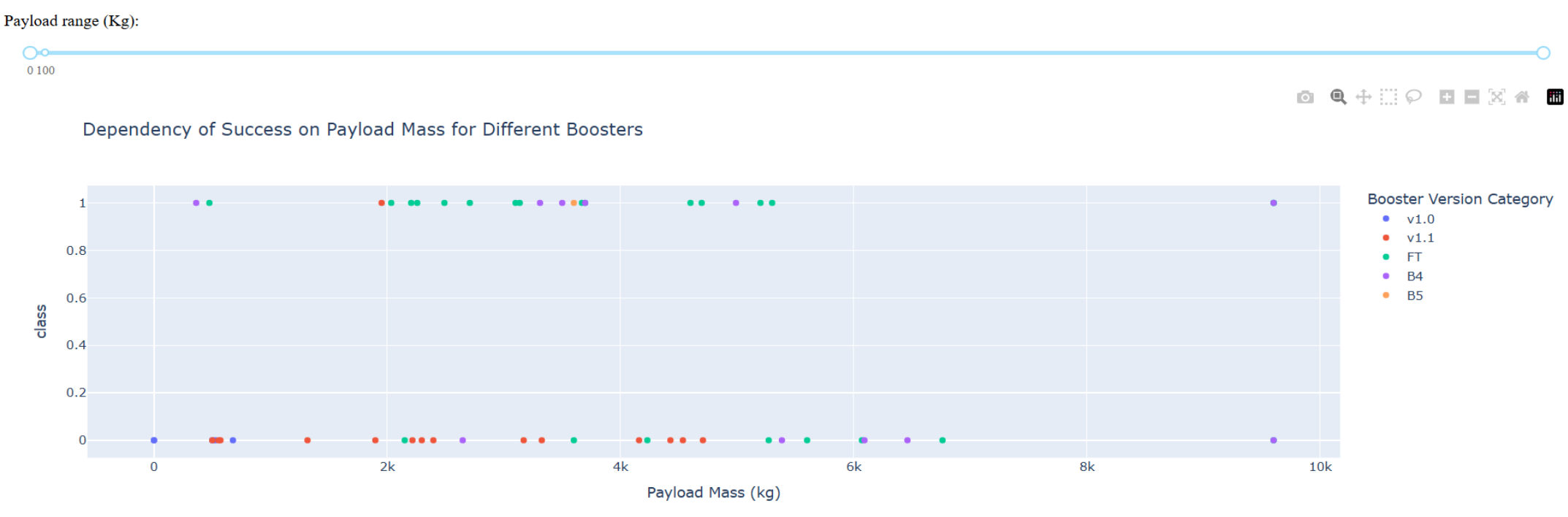
Launch success ratio pie chart for KSC LC-39A area



KSC LC-39A area is a site with the highest launch success ratio. 76,9% of launches where successful from this place.

Payload vs. Launch Outcome

The whole payload range



- Class 1 belongs to successful missions. For the whole payload range we can see that payload mass is crucial factor for the success. With payload mass more than 6000 kg only 1 of 6 missions was successful.

Payload vs. Launch Outcome

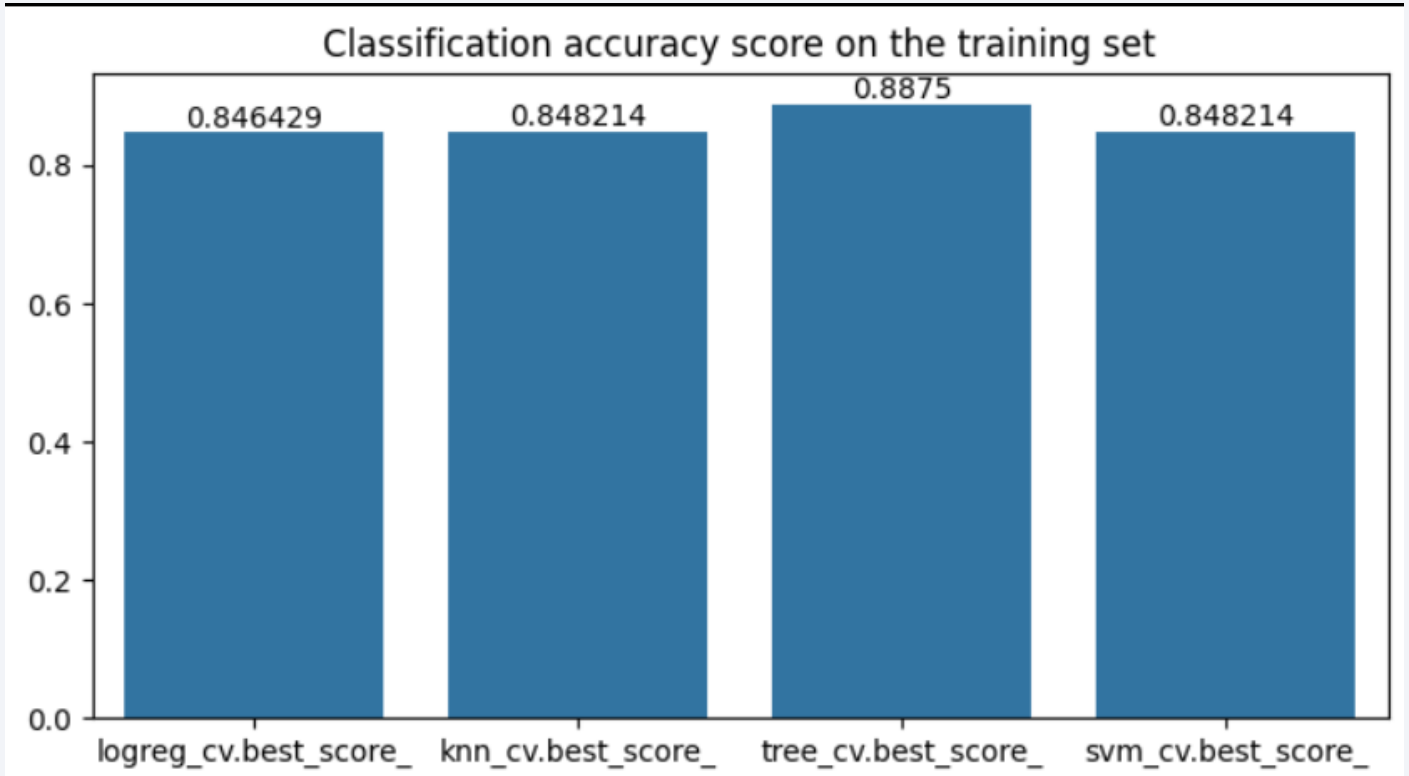
The payload range 100 - 6000 kg

- We can limit payload range from 100 to 6000 kg to see that FT, B4 and B5 boosters are the most successful, with B5 (orange point) we had only one and successful launch. Boosters v1.0 and v1.1 had a very poor performance.

Section 5

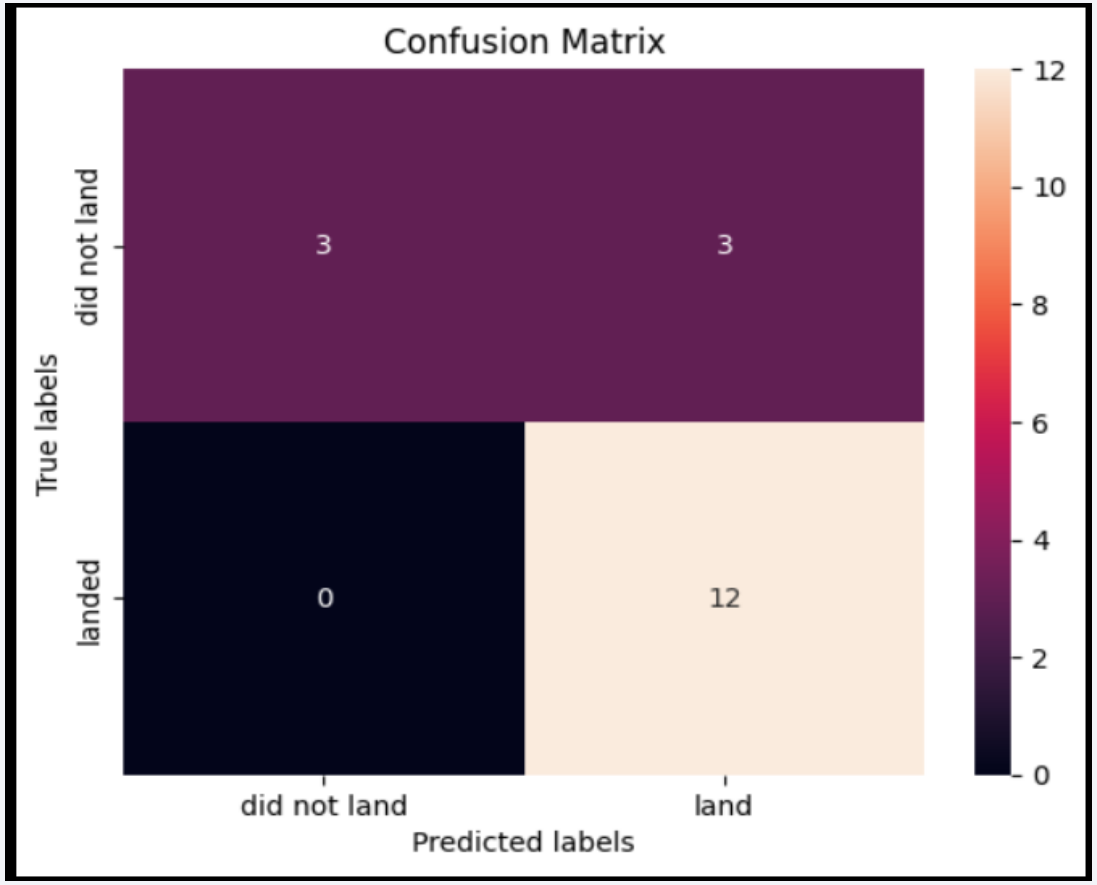
Predictive Analysis (Classification)

Classification Accuracy



As we can see from the bar chart above classification with Decision Tree has the highest classification accuracy score on the training set 88,75%, so we can choose this algorithm as the most promising classifier for our model

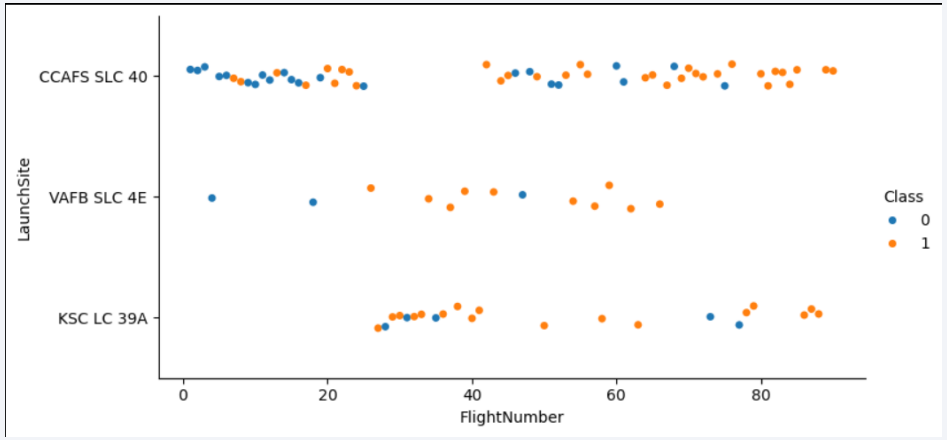
Confusion Matrix



The confusion matrix for Decision Tree classifier shows that all 12 successfully landed boosters where predicted right. But from 6 booster that where not landed 3 where marked as successfully predicted. So, accuracy score on test data for Decision Tree is 83,33% (as well as for other classifiers).

Conclusions

- We collected data about SpaceX launches from different sources, preprocess and cleaned it.
- Clearly the most important factor for the first stage landing success is time and experience as there were no successful landings during first 5 years (with many launches there were no even attempts to land a booster) and for last 2 years success rate was 80-90%
- What about launch site factor? On the dashboard we see that KSC LC-39A area is a site with the highest launch success ratio - 76,9%. But when we look at FlightNumber vs LaunchSite scatterplot (with more data by the way) we see that first launches from KSC LC-39A area were made only after first 25-th launches - experimental, with no attempt to land or simply unsuccessful due to the lack of experience. When we dissect these first 25 launches all sites seems to have pretty much the same success ratio. It seems launch site is not an important factor for success.



Conclusions

- The same point as previous is about Orbit Type. Even more – for many Orbits we don't have enough data as there were only 1 flight to them. Hence Orbit type seems doesn't look an important factor as well as Launch Site.
- Payload Mass factor. During EDA we could clearly see that Payload is an important factor and with Payload more than 7000 kg is very low risk of failure.
- From dashboard Payload vs Launch Outcome we see that we need to use FT, B4 and B5 booster to have successful outcome
- Decision Tree has the highest classification accuracy score on the training set 88,75%, so we can choose this algorithm as the most promising classifier to predict a successful booster landing.

Appendix

Data sets:

- https://github.com/NickPurdes/CourseraProjects/blob/main/spacex_launch_dash.csv
- https://github.com/NickPurdes/CourseraProjects/blob/main/dataset_part_2.csv

Thank you!