Winning Space Race with Data Science

<Pelin Okutan>
<07 April 2022>

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection with API and Web Scraping
  - Data wrangling
  - EDA with SQL
  - EDA with data visualization
  - Interactive visual analytics with Folium
  - Interactive visual analytics with Dashboard
  - Predictive Analysis (Classification)
- Summary of all results
  - Data can be acquired in many ways using public sources
  - The launch success was dependent on various factors
  - The most important factor is the training effect, more successful launches over time
  - Interactive data analysis is helpful for analysis and understanding
  - Different classification techniques achieved good results

Introduction

- Project background and context
  - Privately owned space companies are becoming more important. i.e. among them SpaceX has cost advantage
  - I am acting as a data scientist for a start up company to complete the project with SpaceX
- Problems you want to find answers
  - Predict if the first stage of SpaceX Falcon 9 will land successful
  - Get insights for SpaceX cost structure
  - Advantage in bidding against SpaceX

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data is collected using SpaceX API and Wikipedia web scraping
- Perform data wrangling
  - Filter out non-falcon9 data
  - Replace nulls with mean
  - One hot encoding technique
  - Convert results into classes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression, SVM, decision tree, KNN

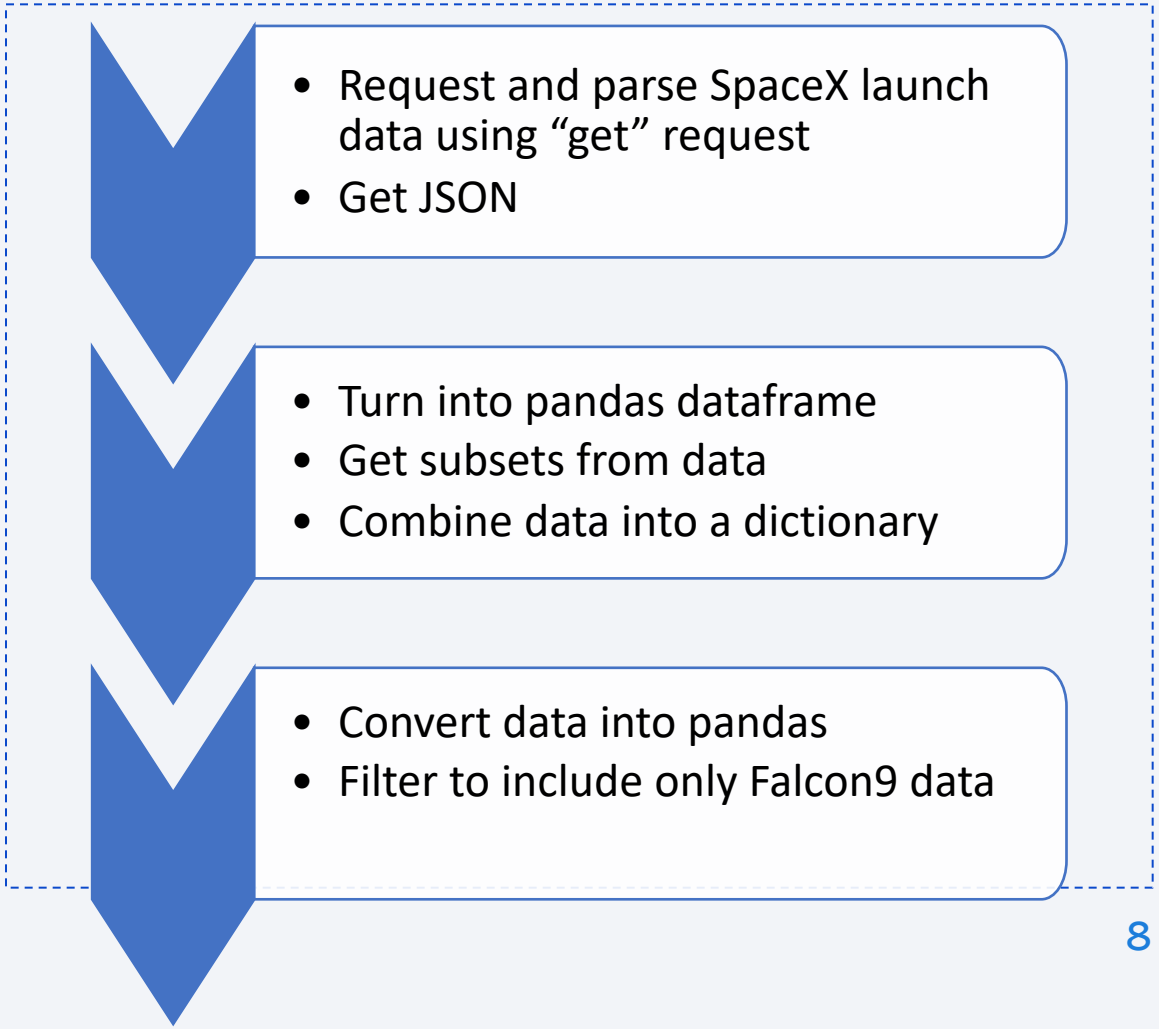
Data Collection

- Data is collected using “get” request to SpaceX API.
  - The response was decoded as JSON
  - JSON turned into a pandas dataframe
  - Data was cleaned and checked for missing values

- Web scraping from Wikipedia for Falcon9 launch using beautifulSoup
- Data was converted to and stored as .xls

Data Collection – SpaceX API

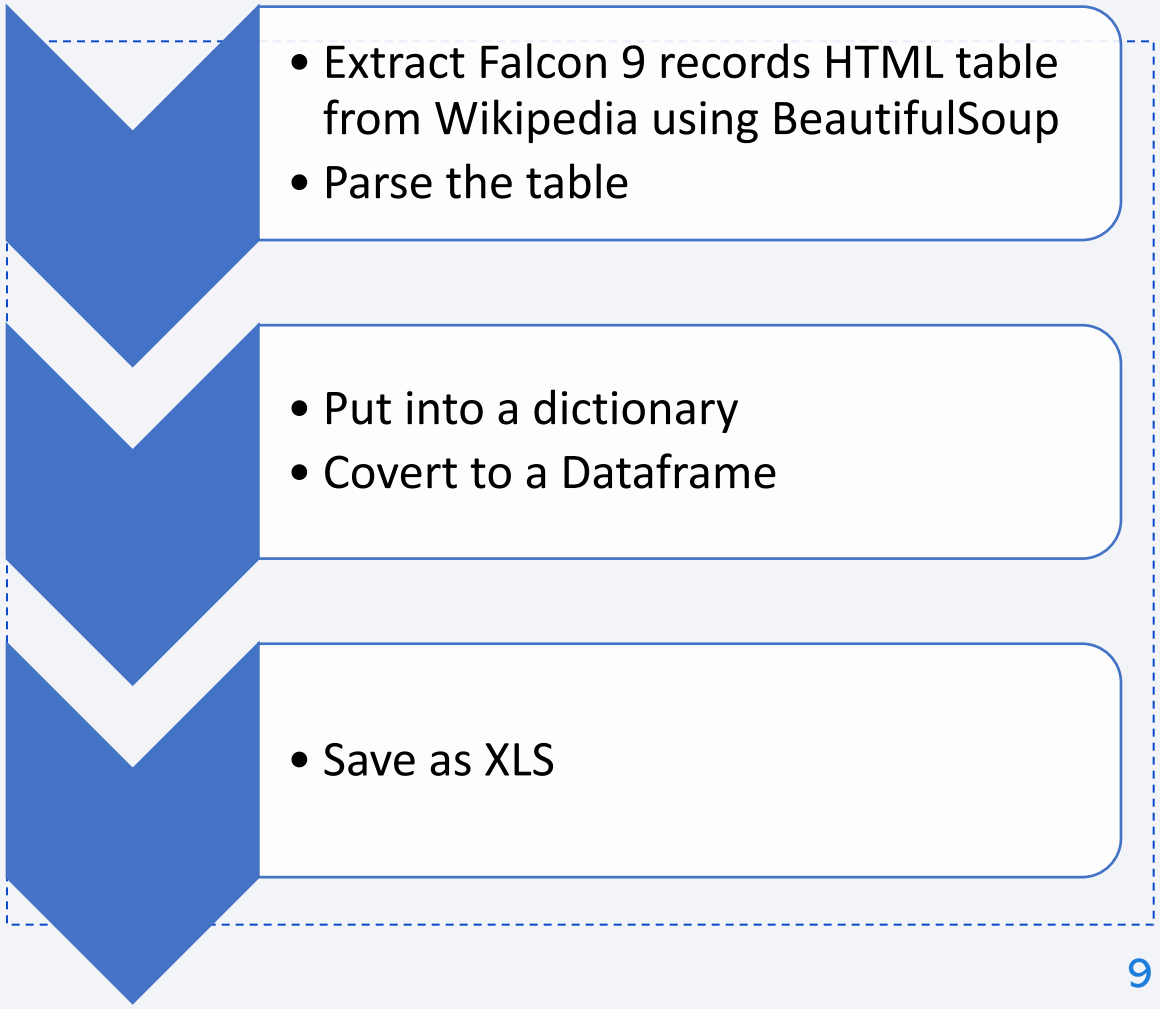
- Data collection with SpaceX REST calls
- https://github.com/pelinle/Python-for-Data-Science-IBM/blob/main/notebook:final_notebook_obQo8Hvyl.ipynb



Data Collection - Scraping

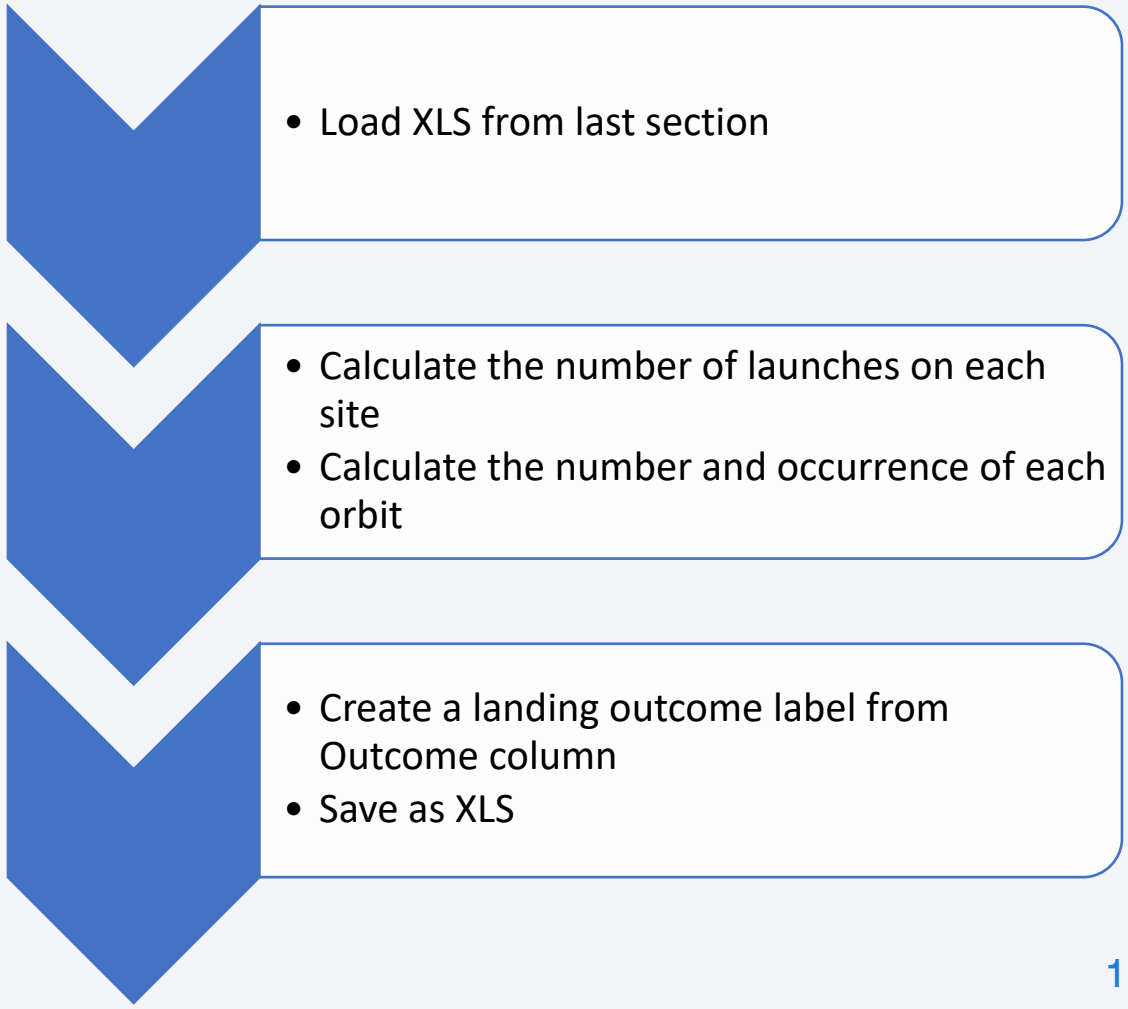
- Web scraping process using BeutifulSoup

- https://github.com/pelinle/Python-for-Data-Science-IBM/blob/main/Webscraping.ipynb



Data Wrangling

- Data was processed according to the flowchart
- https://github.com/pelinle/Python-for-Data-Science-IBM/blob/main/notebook:data_wrangling_xDICO6LfZ.ipynb



EDA with Data Visualization

- Plotted Charts:
  - Launch site vs Flight Number (How are the sites distributed and are there more successful flights?)
  - Launch site vs PayLoadMass (How is the payloadmass distributed over the sites of the flights?)
  - Yearly trend (over time, team becomes more trained and landings become more successful)
  - For each orbit:
    - Success rate (Is the success of the launches dependant on the orbits?)
    - Flight number (How are the orbits distributed over the flight number and are there any links to success rate?)
    - Payloadmass (Are the payloadmasses connected to the orbits and is the success rate linked to this?)
- https://github.com/pelinle/Python-for-Data-Science-IBM/blob/main/notebook:jupyter-labs-eda-dataviz_(1)_KVv0aOuG5.ipynb

EDA with SQL

- The following questions were addressed:
  - names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA’
  - total payload mass carried by boosters launched by NASA (CRS)
  - average payload mass carried by booster version F9 v1.1
  - date when the first successful landing outcome in ground pad was achieved
  - names of the boosters which have success in drone ship and have payload 6000> mass >4000
  - total number of successful and failure mission outcomes
  - names of the booster_versions which have carried the maximum payload mass
  - failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

- https://github.com/pelinle/Python-for-Data-Science-IBM/blob/main/notebook:EDA_with_SQL_tB72pHAMj.ipynb

Build an Interactive Map with Folium

- Map Objects:
  - Marker() to create marks
  - Markercluster() for clustered markers
  - Circle() to create circles around the location
  - Popups
  - MousePosition() to get coordinates
  - PolyLine() to draw lines between positions
- https://github.com/pelinle/Python-for-Data-Science-IBM/blob/main/notebook:lab_jupyter_launch_site_location_Uzz0NZ7hT.ipynb

Build a Dashboard with Plotly Dash

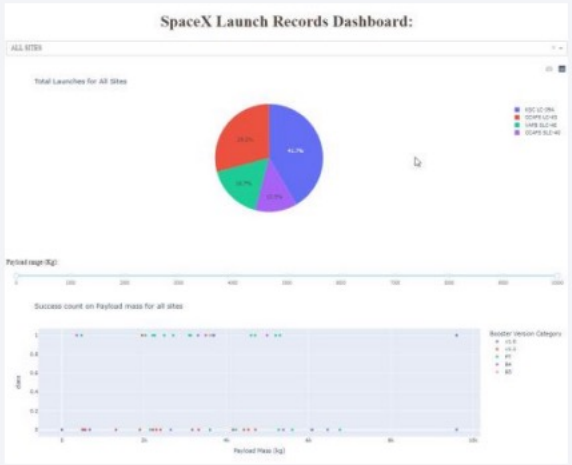
- Interactive Web Application using Plotly Dash
  - Pie Chart to show the success rate with a dropdown menu to choose between launch sites or display the successrate for all launch sites
  - Scatter plot of the payloadmass vs outcome with a slider to specify the range of payload massess
- https://github.com/pelinle/Python-for-Data-Science-IBM/blob/main/spacex_dash_app.py

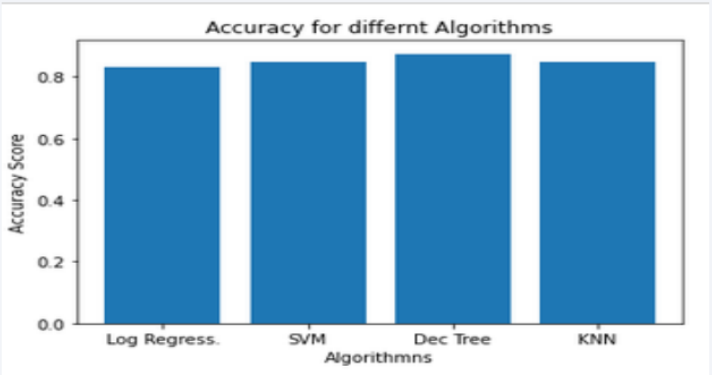
Predictive Analysis (Classification)

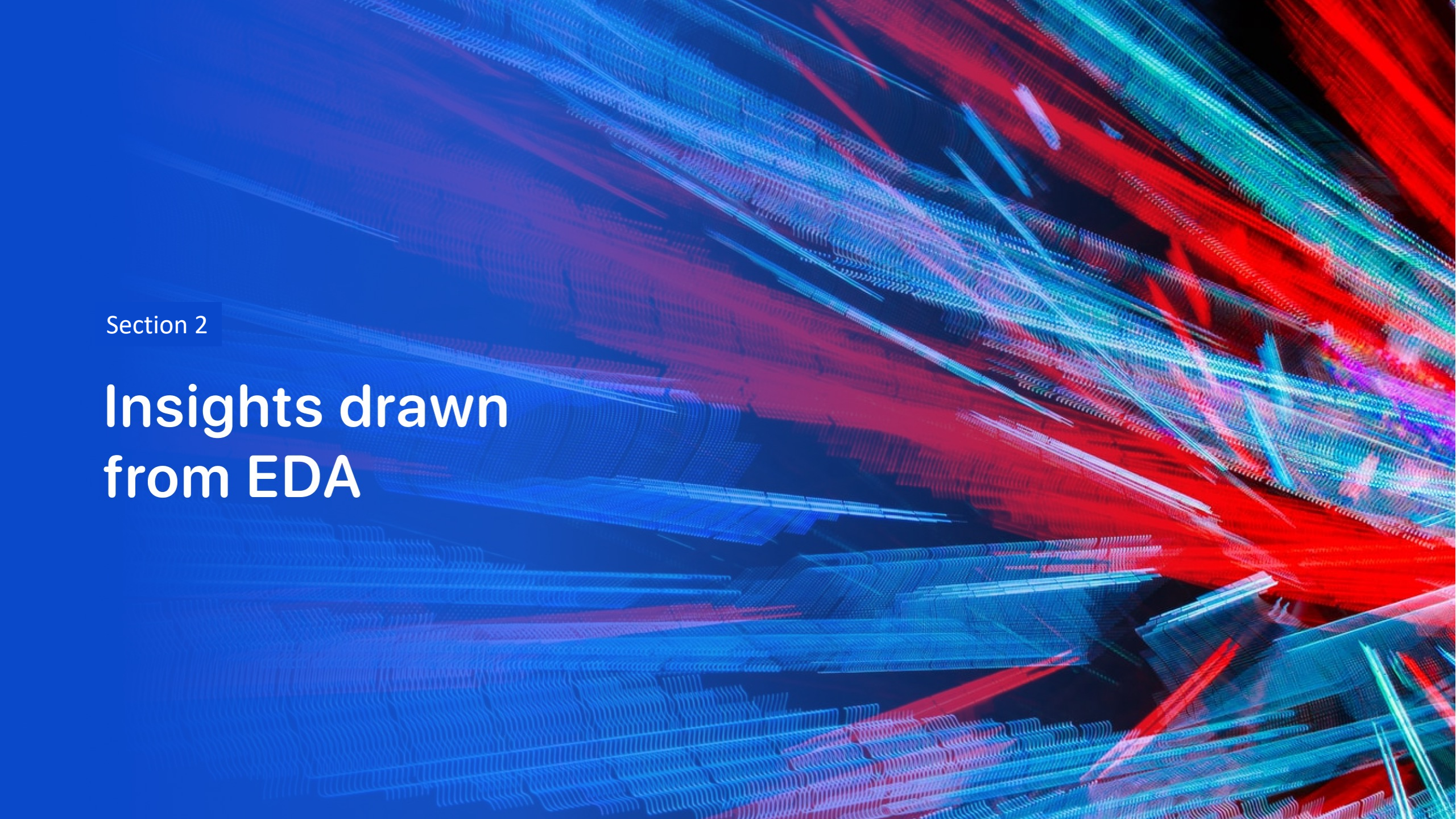
- Machine Learning models were applied including:
  - Logistic regression
  - SVM
  - Decision tree
  - KNN
- GridsearchCV for tuning the hyperparameters
- Prediction accuracy as metrics for determining the best model
- https://github.com/pelinle/Python-for-Data-Science-IBM/blob/main/notebook:SpaceX_Machine_Learning_Prediction_Part_5_(1)_zbCwzquHJ.ipynb

Results

- Exploratory data analysis results show that the success factor increased over time
- Interactive analytics demo was helpful for visualization
- Predictive analysis results was accurate

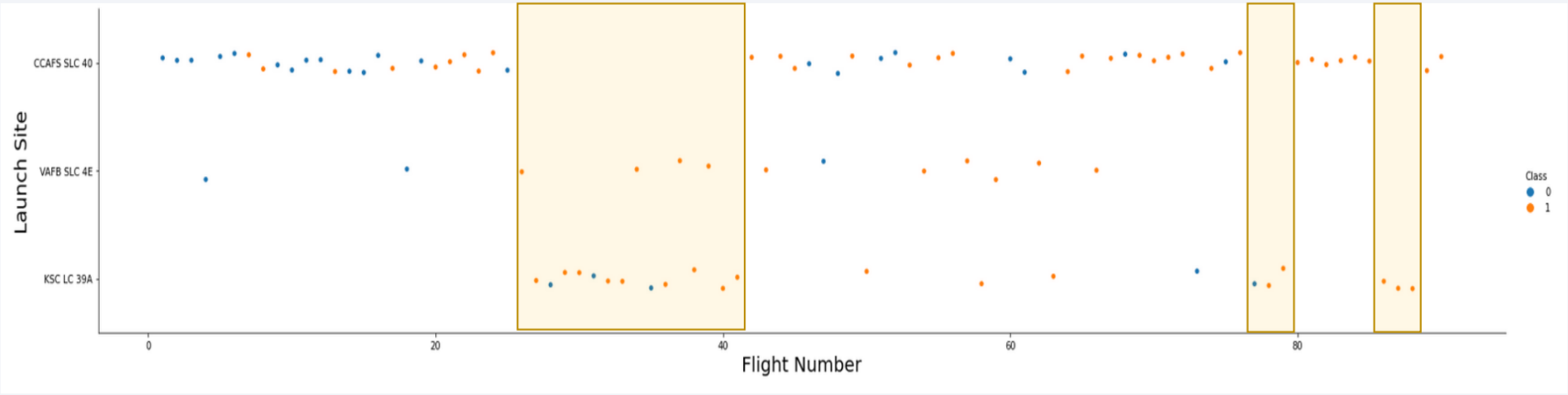





Section 2
Insights drawn from EDA

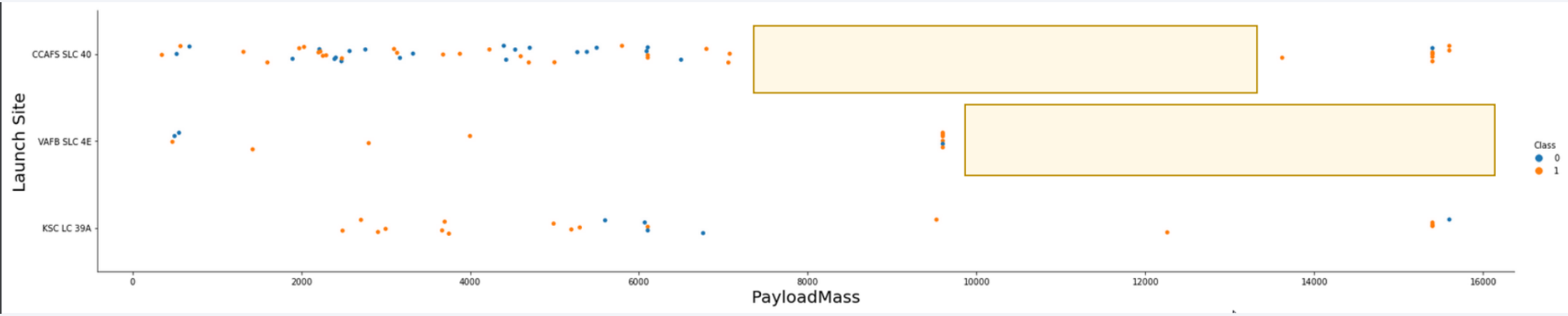
Flight Number vs. Launch Site

- As the flight number increases successful flights also increases
  - This seems to be true for all the sites, but only for CCAFS SLC 40, data is succifient to support this claim
- Most flights are from CCAFS SLC 40
  - Yellow marked flight number slots has no flights in CCAFS SLC 40
  - VAFB SLC 40 is not much used



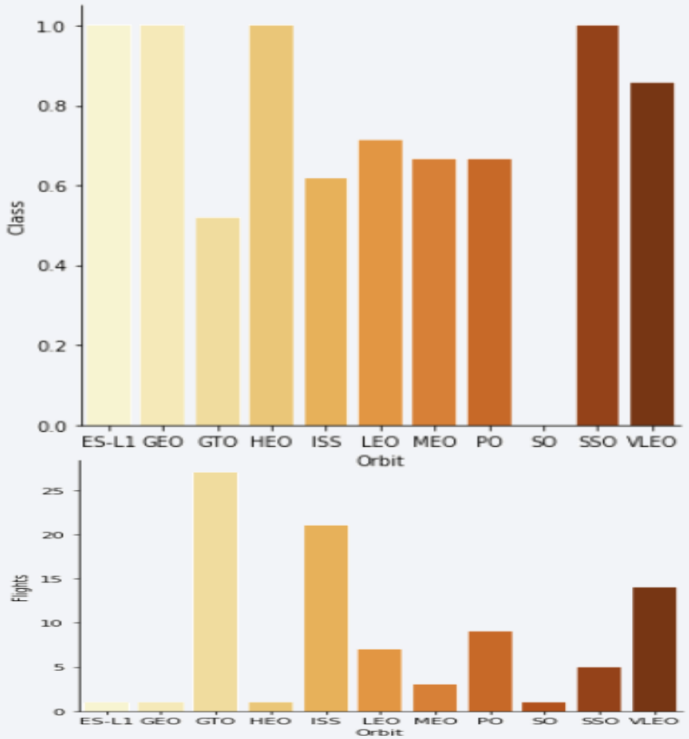
Payload vs. Launch Site

- Interesting are the gaps (marked in yellow) :
  - For VAFB SLC 40 no flights with masses > 10000
  - For CCAFS SLC 40 no flights between 7500 and 13000
  - Might be because of masses were concentrated on For VAFB SLC 40
  - KSC LC 39 less frequently used, but with all masses



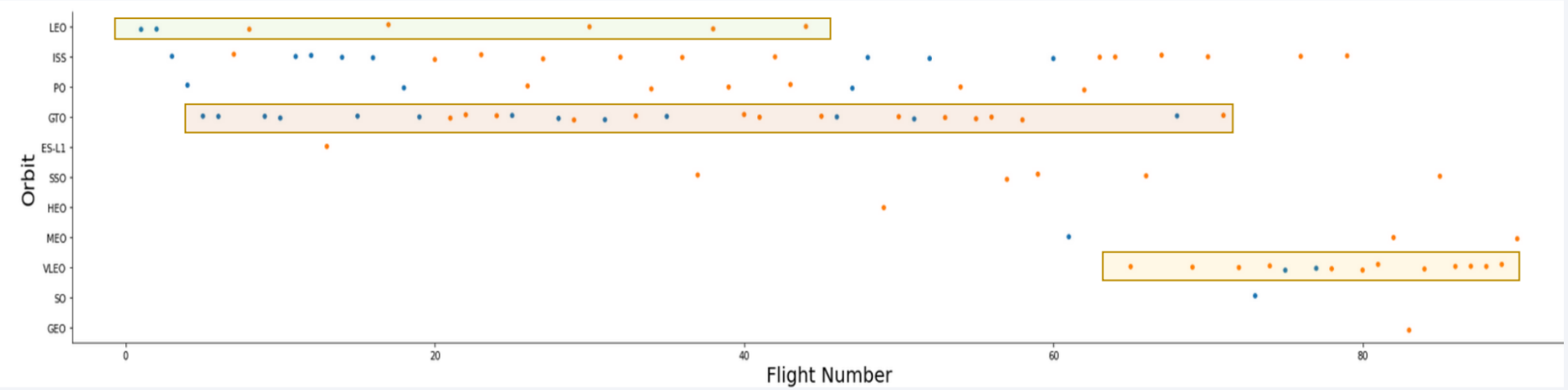
Success Rate vs. Orbit Type

- SO is not very successful, while ESL1, GTO, HE0 are very successful
  - If we plot the number of flights per orbit (lower graph) we see that these data are based on only one occurrence
- Taking only into account the orbits with a number of flights of at least 5
  - SSO and VLEO are more successful
  - GTO is less successful
- But only the launches for GTO, ISS and VLEO have a number >10, only here the statistics is OK.



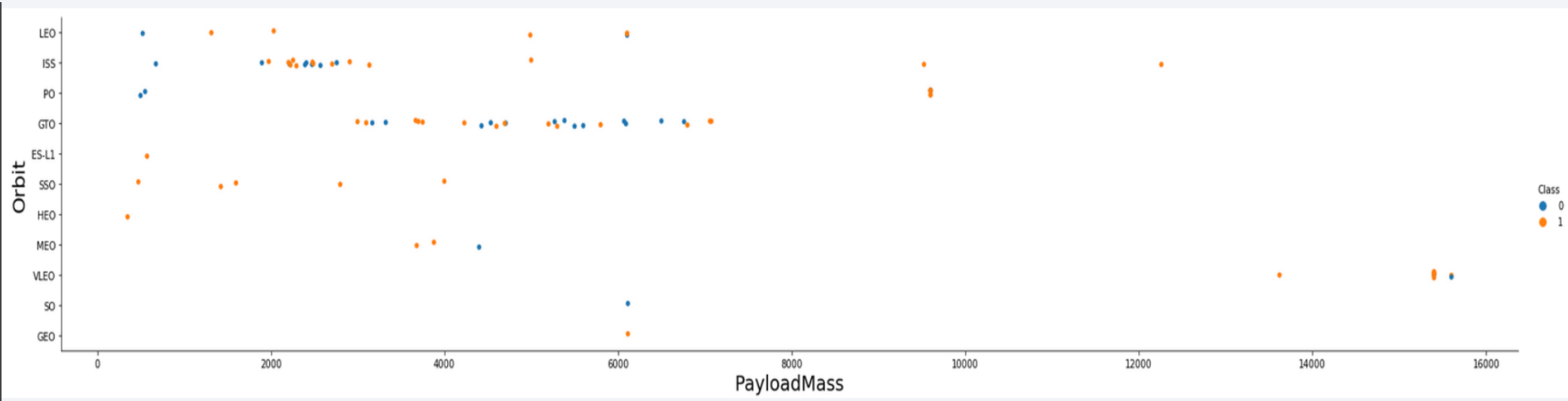
Flight Number vs. Orbit Type

- The conclusions from the last slide are also supported by this graph:
- Generally, the higher the flight number, the better the success rate
  - For orbits with a typically higher flight number, the success rate is better (e.g, the VLEO, marked in yellow)
  - For LEO (marked in green) the “training” effect is clearly visible
  - For GTO (marked in red) there are also some failures also for high flight numbers. Here I would look for other potential cause



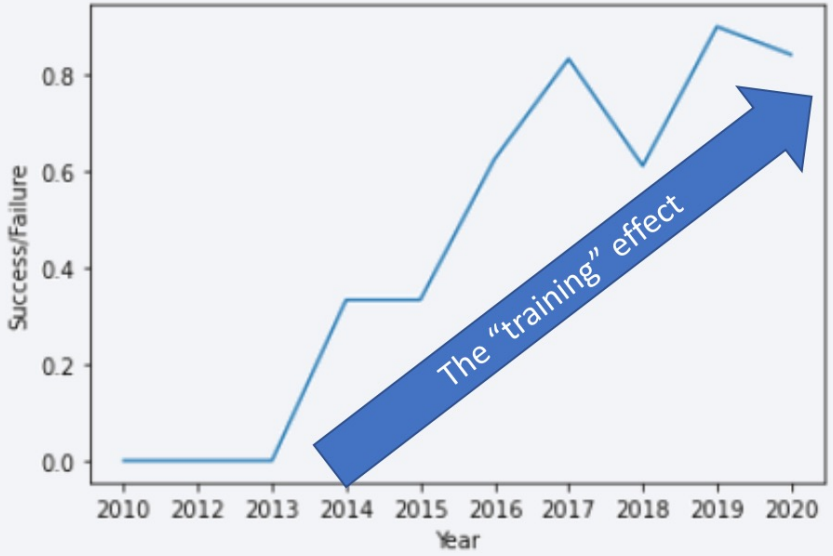
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS (marked green)
- for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here (marked yellow)
    - But also smaller spread in payload massnumbers. Here I woud look for other potential causes



Launch Success Yearly Trend

- Stagnant before 2013
- After 2013, Launch success increased with time
- After around 2019 saturation effect, no marked increase possible
- A clear visualization of the “training” effect that we also saw in the prior slides.



All Launch Site Names

- I used the distinct statement
- There are 4 different Launch Sites, but CFAFS LC40 appears only until 2016, CFAFS SLC40 afterwards.
- Maybe one of them was closed and other one opened.

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- All of them is from CCAFS LC-40
- For these first 5 the landing was either a failure or not attempted

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

- the total payload carried by boosters from NASA: **619967**
- the average payload carried by boosters from NASA: **6138 kg**

PAYLOADMASS
619967

Average Payload Mass by F9 v1.1

- the average payload mass carried by booster version F9 v1.1 : **2534 kg**
- Much less than the average payload mass **(6138 kg)**

Avg Payload Mass F9 v1.1
2534

First Successful Ground Landing Date

- The first successful landing on ground pad was in 2015
- This was followed by several further successful landings on Ground path

First sucesfull landing on Ground path
2015-12-22

sucesfull landing on Ground path
2015-12-22
2016-07-18
2017-02-19
2017-05-01
2017-06-03
2017-08-04
2017-09-07
2017-12-15
2018-01-08

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

- Acquired with BETWEEN statement

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

Total Number of Successful and Failure Mission Outcomes

- Most of the missions were successful
- Only one failure in flight
- One success, but unclear payload status

MISSION_OUTCOME	
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The maximum Payload mass was 15600 kg
- In the table this name of the boosters that carried the maximum payload is listed
- As a reference, I also selected the date: obviously, all the missions with the maximum payload mass are quite recent

BOOSTER_VERSION	MAXPAYLOADMASS	DATE
F9 B5 B1048.4	15600	2019-11-11
F9 B5 B1049.4	15600	2020-01-07
F9 B5 B1051.3	15600	2020-01-07
F9 B5 B1056.4	15600	2020-02-17
F9 B5 B1048.5	15600	2020-03-18
F9 B5 B1051.4	15600	2020-04-22
F9 B5 B1049.5	15600	2020-06-04
F9 B5 B1060.2	15600	2020-09-03
F9 B5 B1058.3	15600	2020-10-06
F9 B5 B1051.6	15600	2020-10-18
F9 B5 B1060.3	15600	2020-10-04
F9 B5 B1049.7	15600	2020-11-25

2015 Launch Records

- The list below is the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- There were only 2 occurrences
- Both from the same launch site

DATE	BOOSTER_VERSION	LAUNCH_SITE	LANDING__OUTCOME
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Below Is the 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

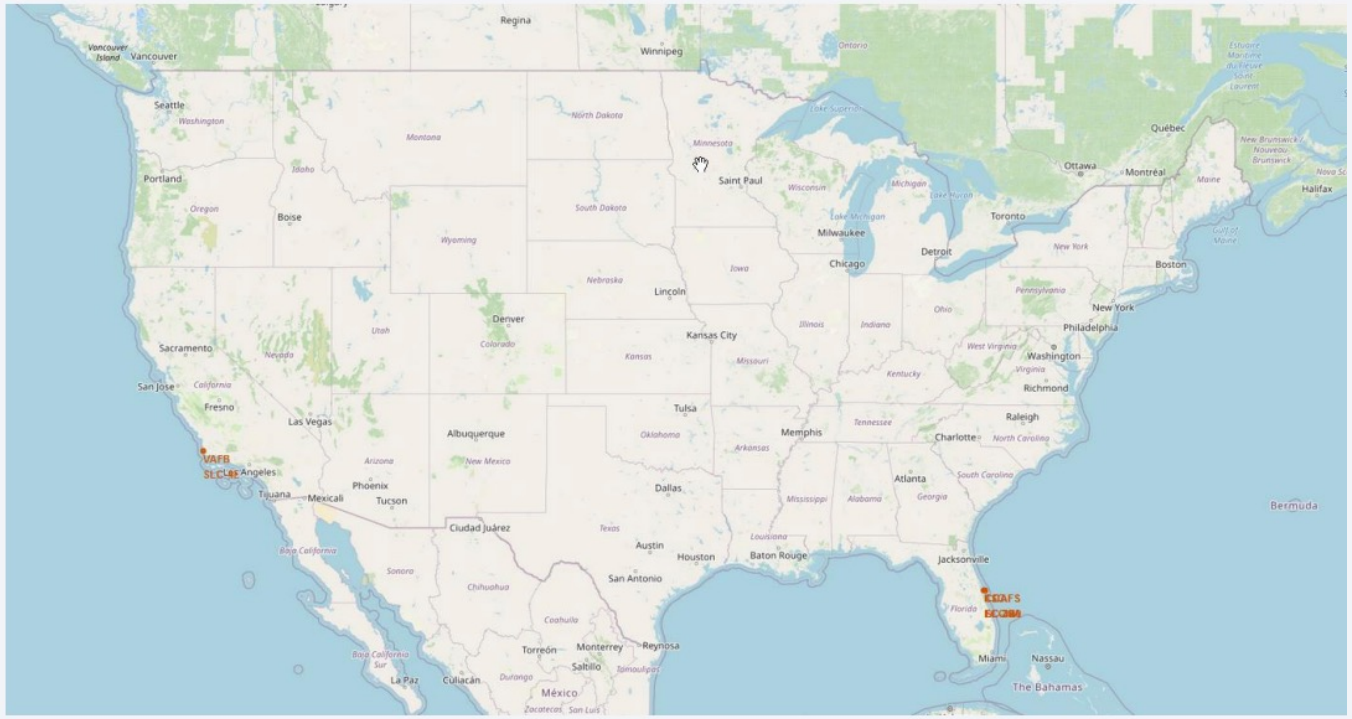
LANDING__OUTCOME	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- In the majority of cases there was no attempt
- The successful and unsuccessful outcomes were nearly equally distributed


Section 3
Launch Sites
Proximities Analysis

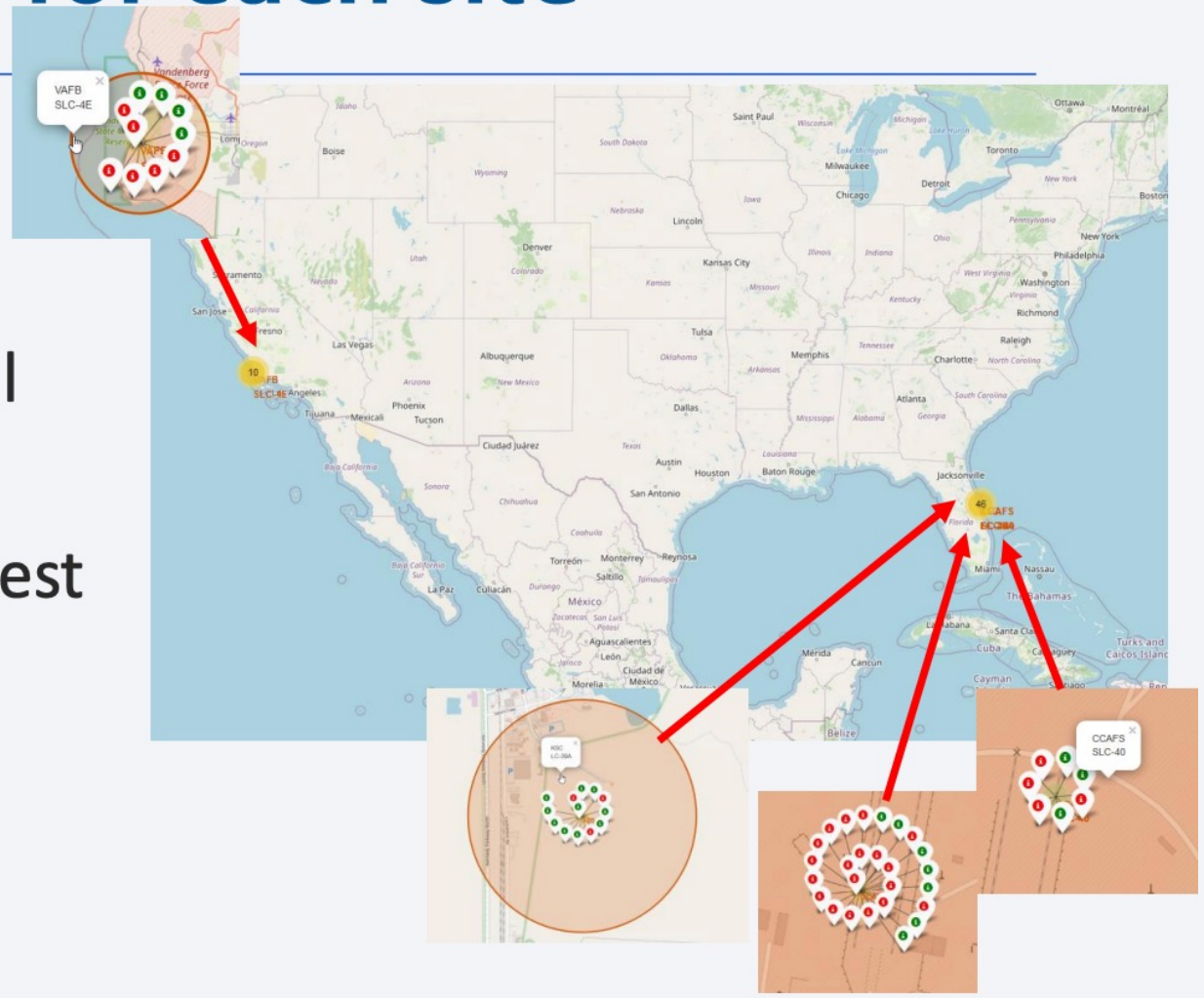
Position of the launch sites

- All launch sites are in the US, in Florida and in California
- The launch sites as close to the equator as possible within the USA
  - This reduces the fuel necessary to launch the rocket, https://en.wikipedia.org/wiki/Near_equatorial_orbit
- All launch sites are in very close proximity to the coast
  - Probably reduces the risk in case a launch fails



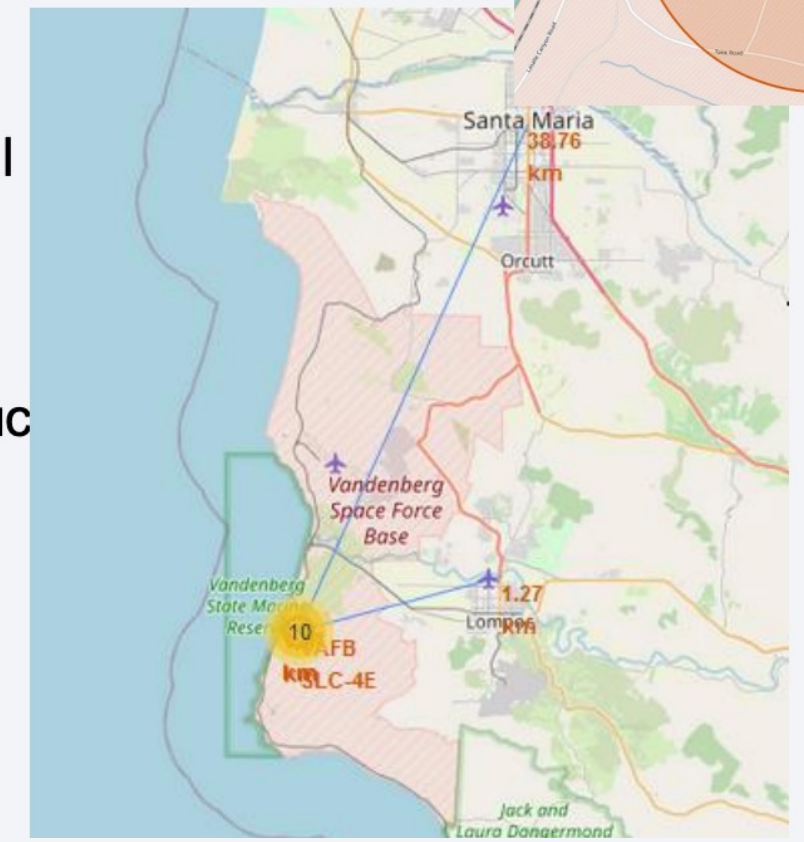
Success/failed launches for each site

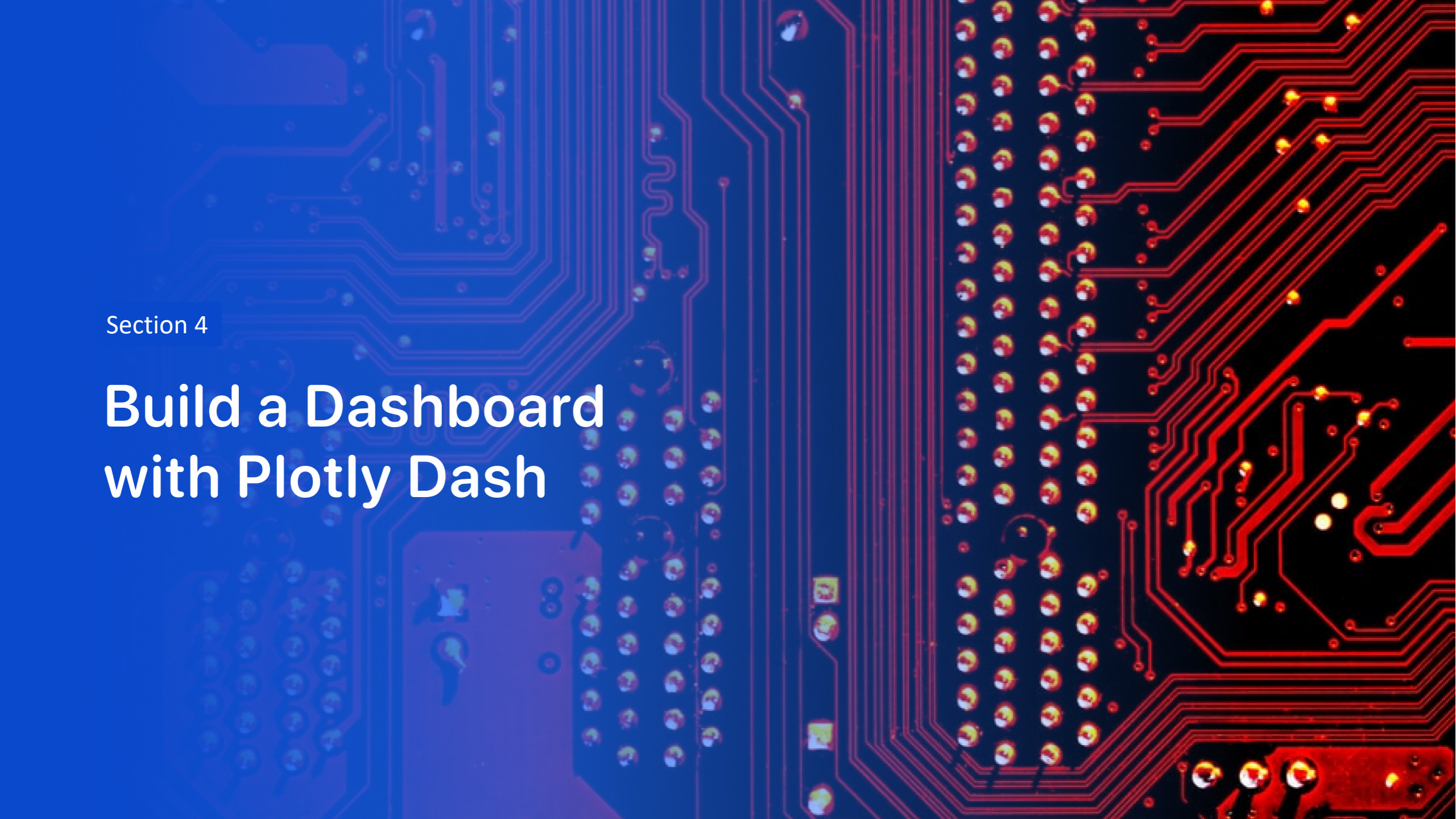
- The successful launches are marked green
- The number of launches differ for all sites
- KSC-LC 39 A is the site with the highest success rate



Distance between one launch site and its proximities

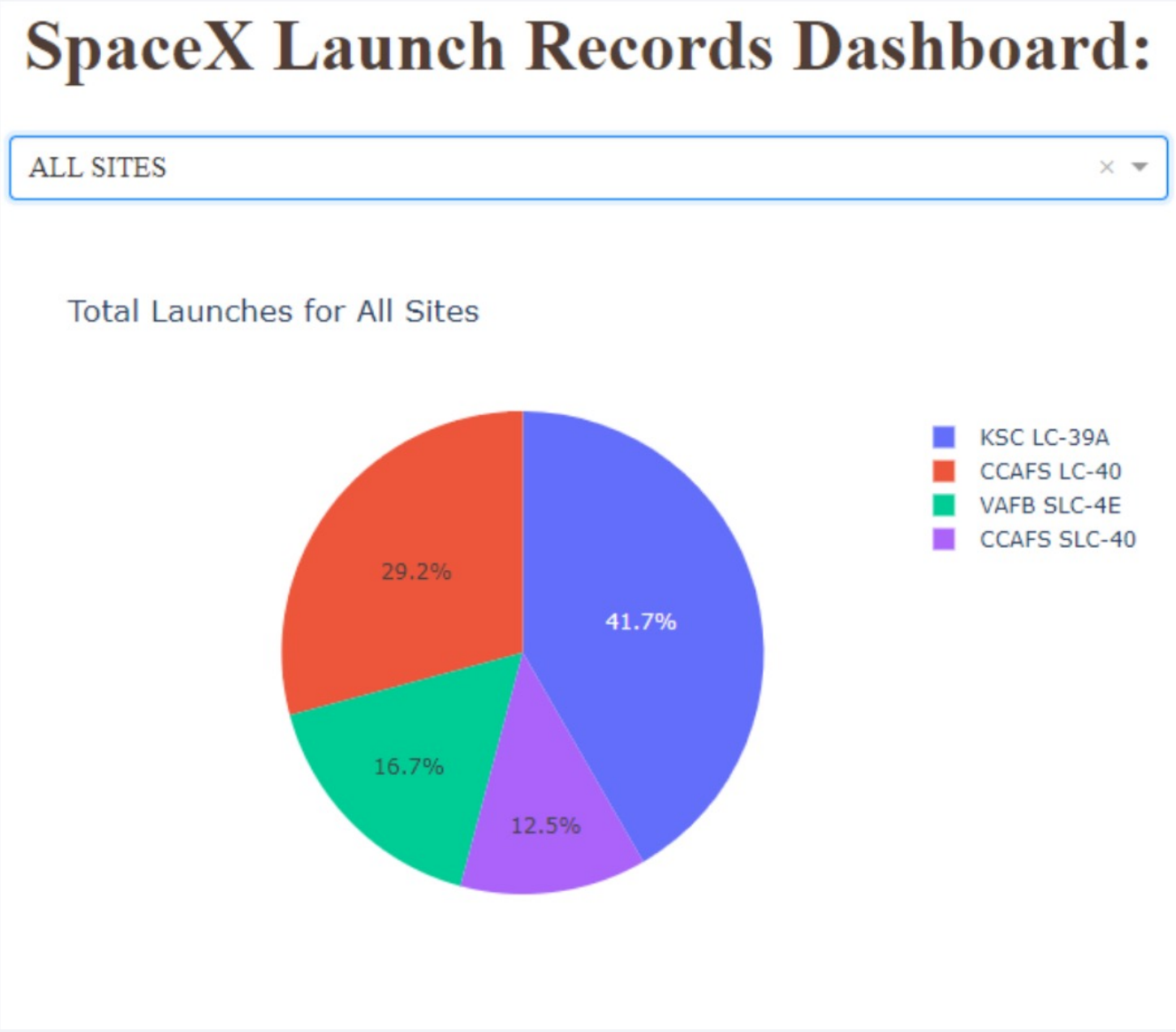
- Distance to Coast: 1.34 km
  - Possible reason: if a launch fails, the rocket can be destroyed over the sea
- Distance to Railroad: 1.27 km
  - Possible reason: supply to the launch site by rail facilitated
- Distance to Highway: 1.27 km
  - Possible reason: supply to the launch site by truc is facilitated
- Distance to city: 38.76 km
  - Possible reason: minimize the risk for the population if a launch fails




Section 4
Build a Dashboard with Plotly Dash

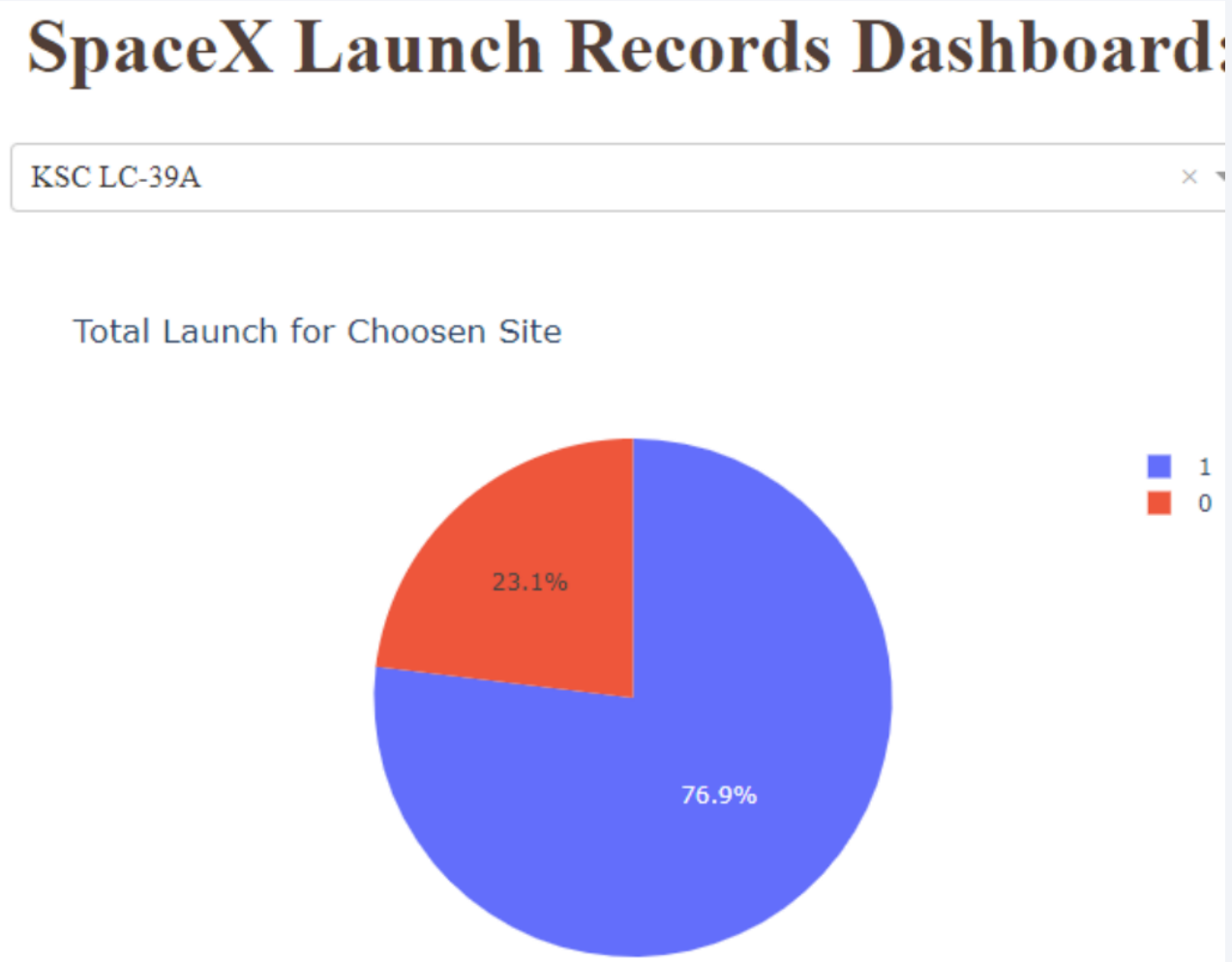
Pie chart, launch success count for all sites

- KSC LC-39A had the most success launches
- CCAFS LC-40 was also quite successful



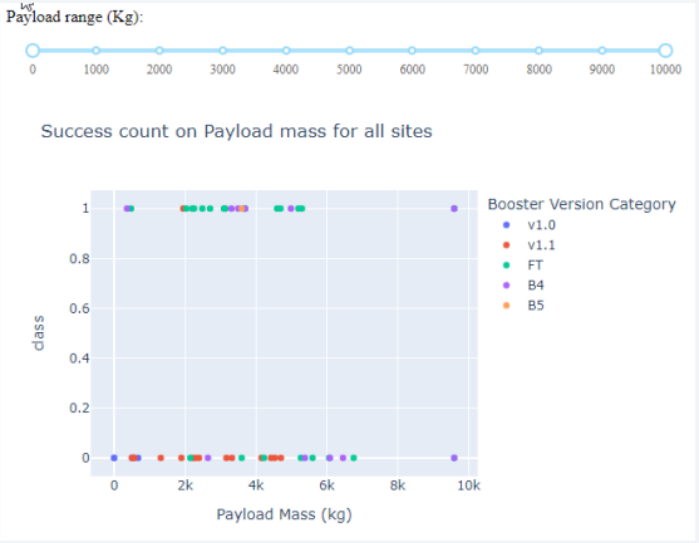
Launch site with the highest launch success ratio

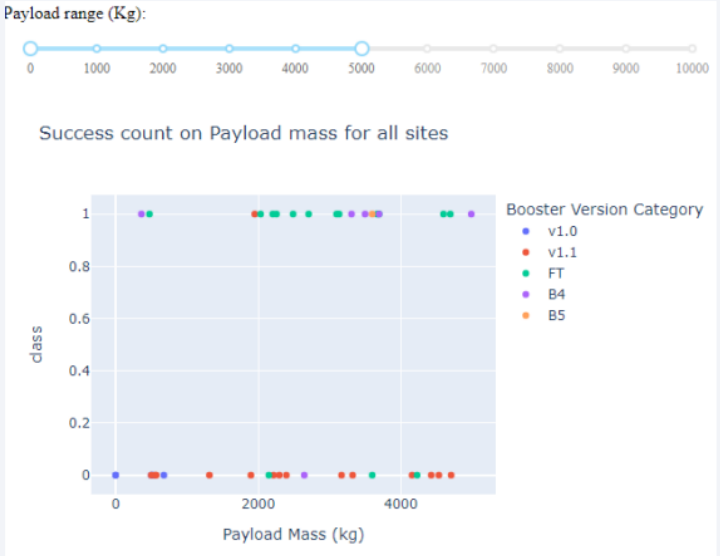
- KSC LC 39A
  - 76.9% success rate
  - 23.1% failure rate

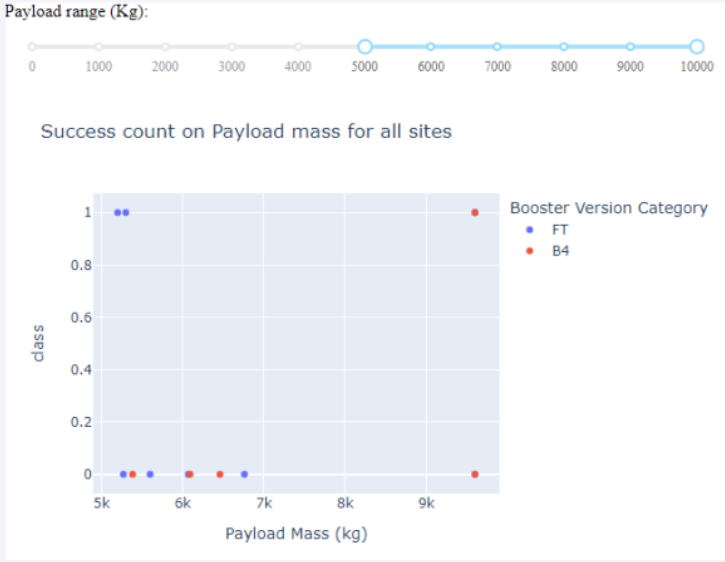


Scatter plot Payload vs Lauch Outcome

- Results for all weights:
  - Booster category V 1.1. is less successful
  - Booster category FT is more successful
- Low weight (up to 5000 kg) vs high weight (5000-10000 kg)
  - There are more successes for the low weight launches





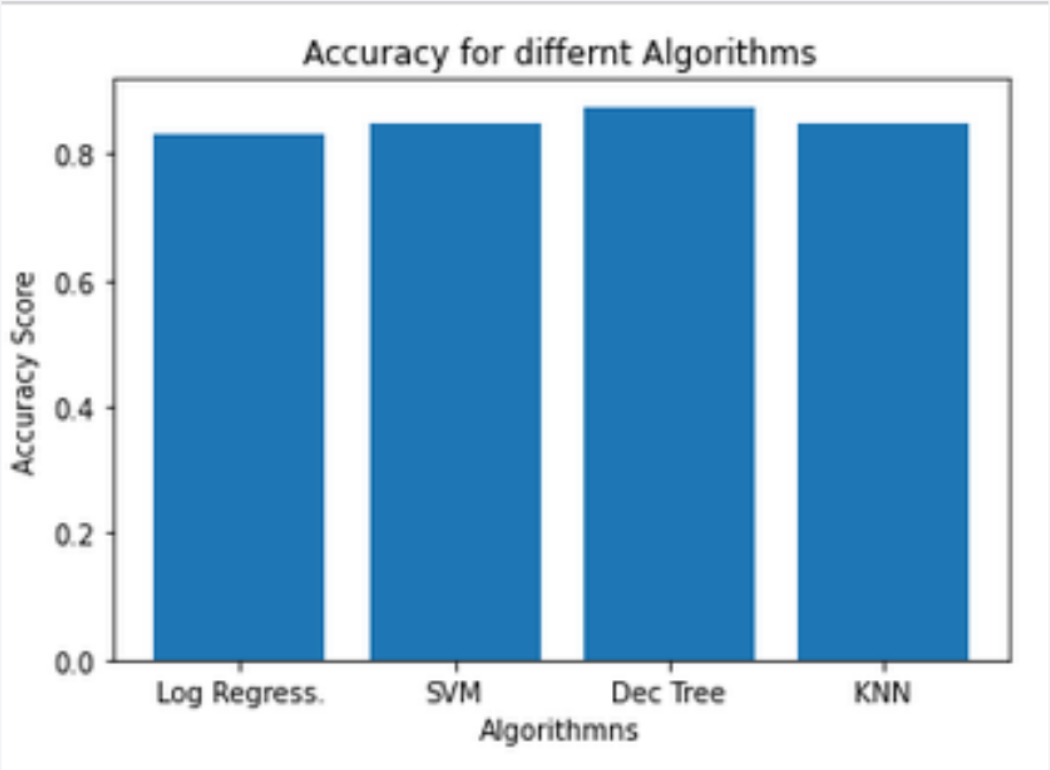


Section 5

Predictive Analysis (Classification)

Classification Accuracy

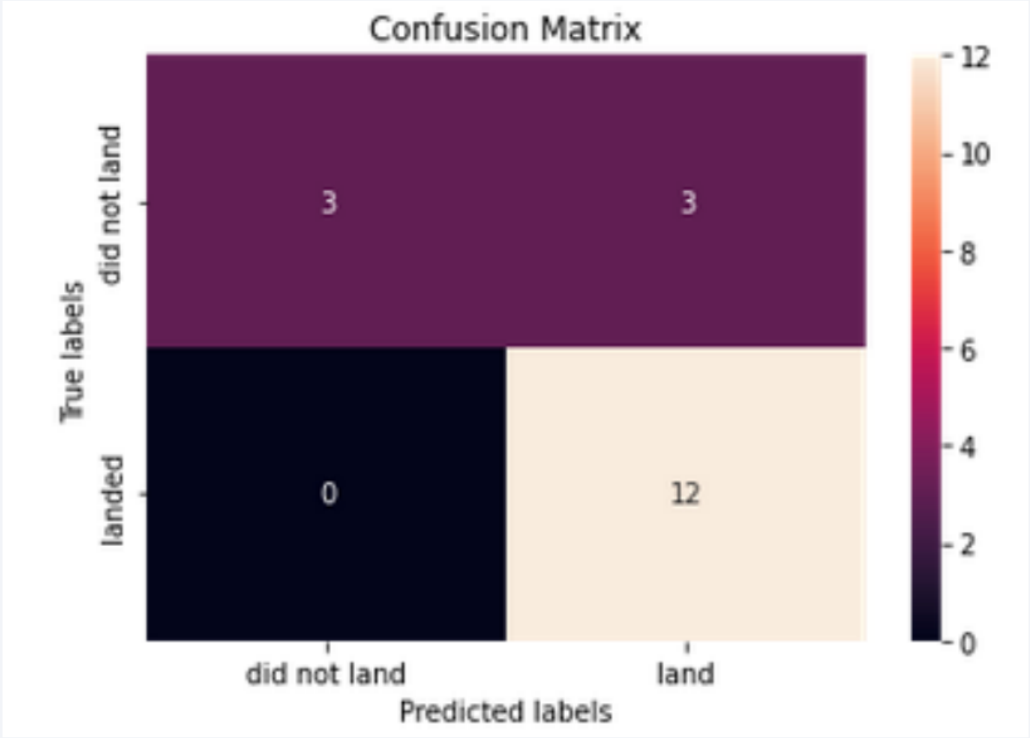
- The desision Tree gives the best accuracy
- However, the differences are not too marked




```
Out[43]: {'Log Regress.': 0.8333333333333334,
          'SVM': 0.8482142857142856,
          'Dec Tree': 0.875,
          'KNN': 0.8482142857142858}
```

Confusion Matrix

- Shown is the confusion matrix of the decision tree model
- 3 wrong predictions
  - In each of these cases it was wrongly predicted that the rocket stage did land



Conclusions

- There is a good amount of progress in launch outcomes with respect to increasing number of launches each year.
- Some sites such as KSC LC -39A had the highest success rate in comparison to other launch sites.
- We can clearly see that space x is leading the space race, and there are some major improvements in the rate of success of landing of falcon 9 stage one.
- The success rate was also dependent on the orbit and payload mass, we saw that ISS and VLEO orbits had a good success rate.
- Support Vector Machine was a suitable model to predict if the stage one would land or not, it had an accuracy of 83%

Thank you!