IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Khanh-Vu HA
24/10/2021

Outline

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

Executive Summary

- For this project, all methodologies were used are:
  - Data scraping on websites
  - Exploratory data analysis
  - Interactive visual analytics
  - Predictive analysis using classification models

- The project finds out the best launching sites and orbits, the best range of payload mass rocket can carry for successful landing as well as building a predictive models for predicting landing outcome of future missions

Introduction

- The project aimed to analyze SpaceX success to get the idea how to start a company similar to it

- The project aims to calculate the probability of reusing the first stage in rocket launching. The probability then will be applied for other company to calculate for operation cost and bid against SpaceX

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Scaping data from various sites, including Wikipedia page on mission outcomes of SpaceX launch history
- Perform data wrangling
  - Parsing HTML collected from websites and transforming it into pandas dataframe for further analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardizing data, splitting it into train and test set, and trying various classification model: logistics regression, support vector machine, KNN, decision tree to find the best performance possible

Data Collection

- Select all websites that contain information about SpaceX launch history including SpaceX website and Wikipedia page on SpaceX mission outcomes

- Using Requests and Bs4 to scrap and parse websites for information inside

- Transforming parsed data into pandas table for further analysis

Data Collection – SpaceX API

- First, using a static response object and define as a variable (static_json_url)
- Second, check the request response. It will return 200 for successful:

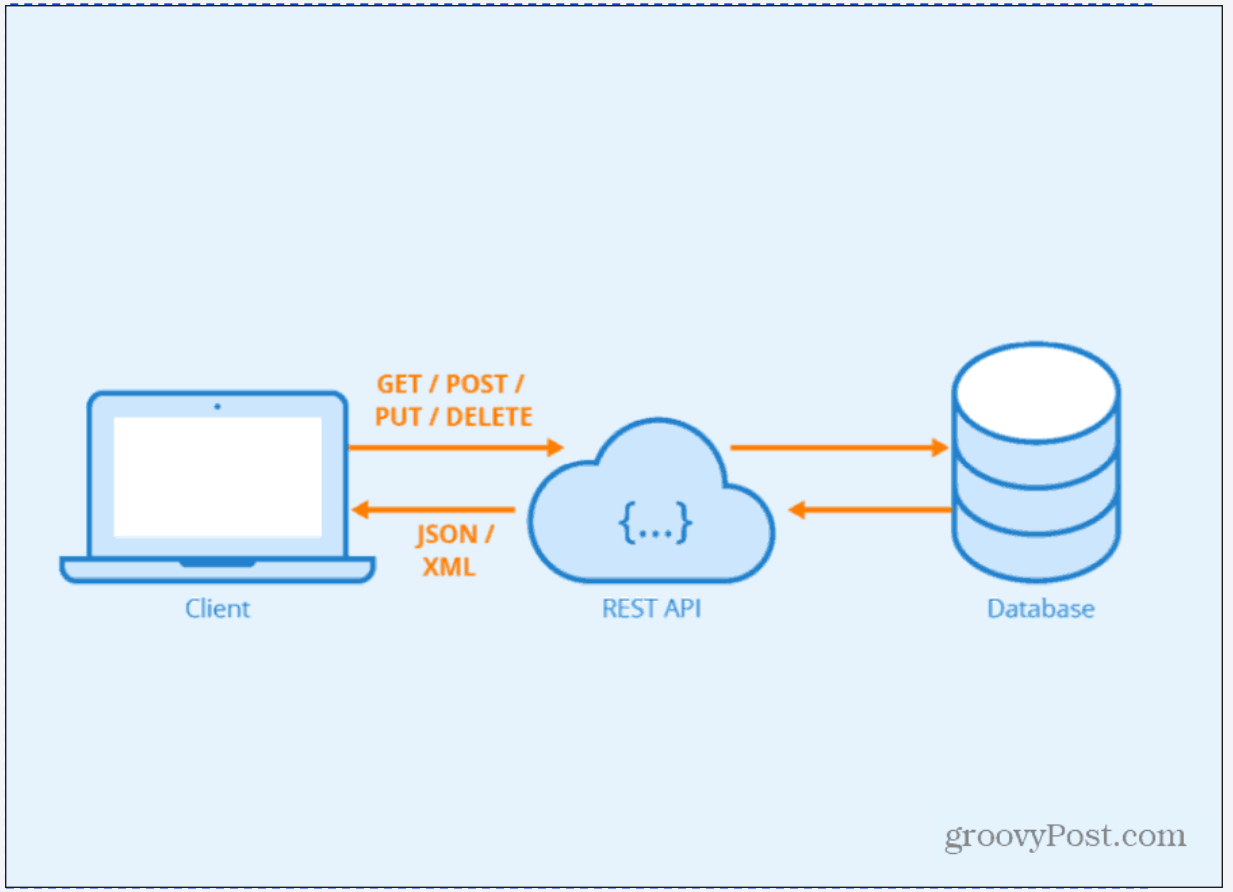
```
response.status_code
```

- Third, decoding the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize():

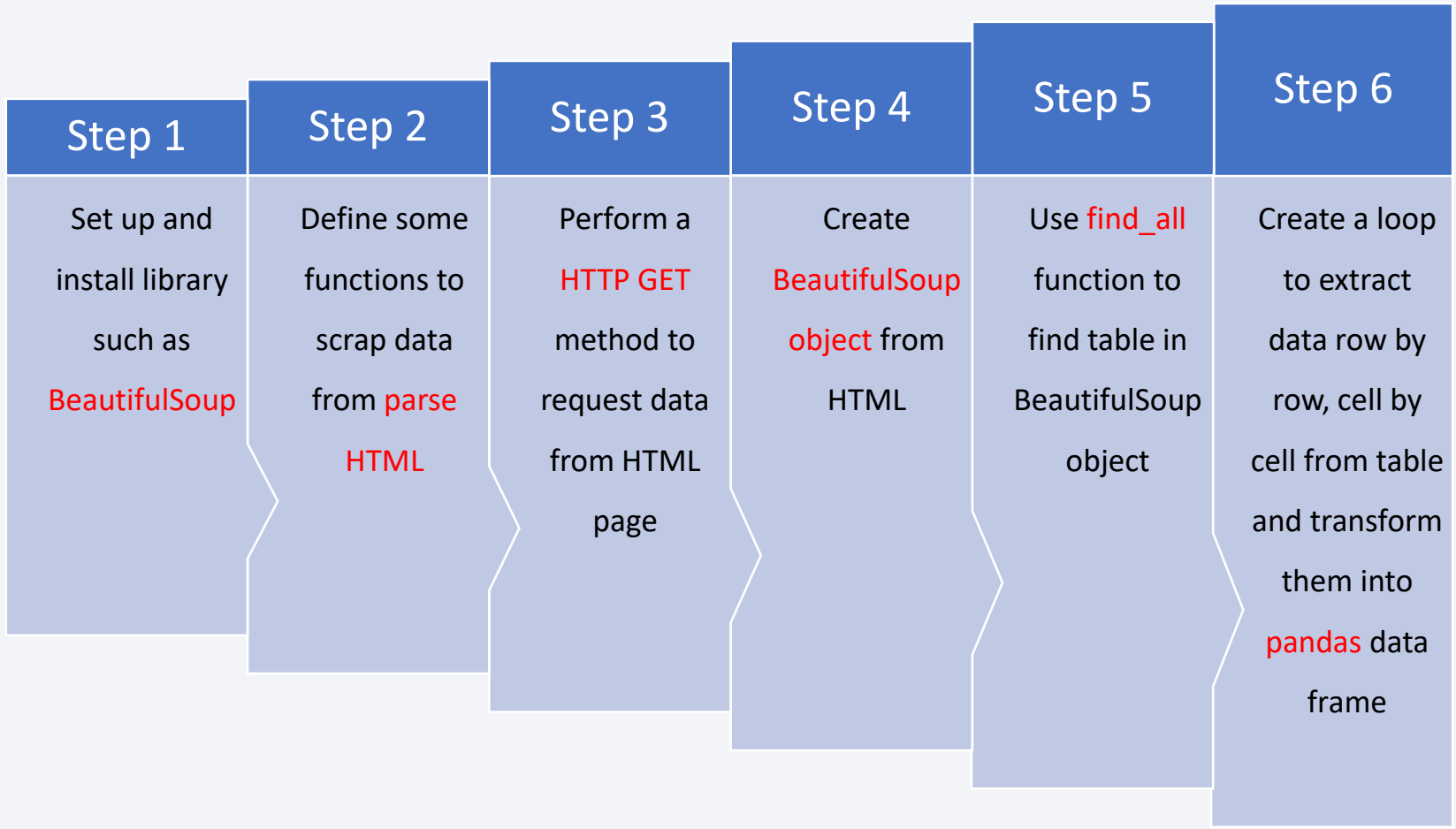
```
response2 = requests.get(static_json_url)

data = pd.json_normalize(response2.json())
```

- Github link: https://github.com/khanhvuha/ibm_watson_studio/blob/master/ADSC:%20Data%20collection%20API.ipynb



Data Collection - Scraping



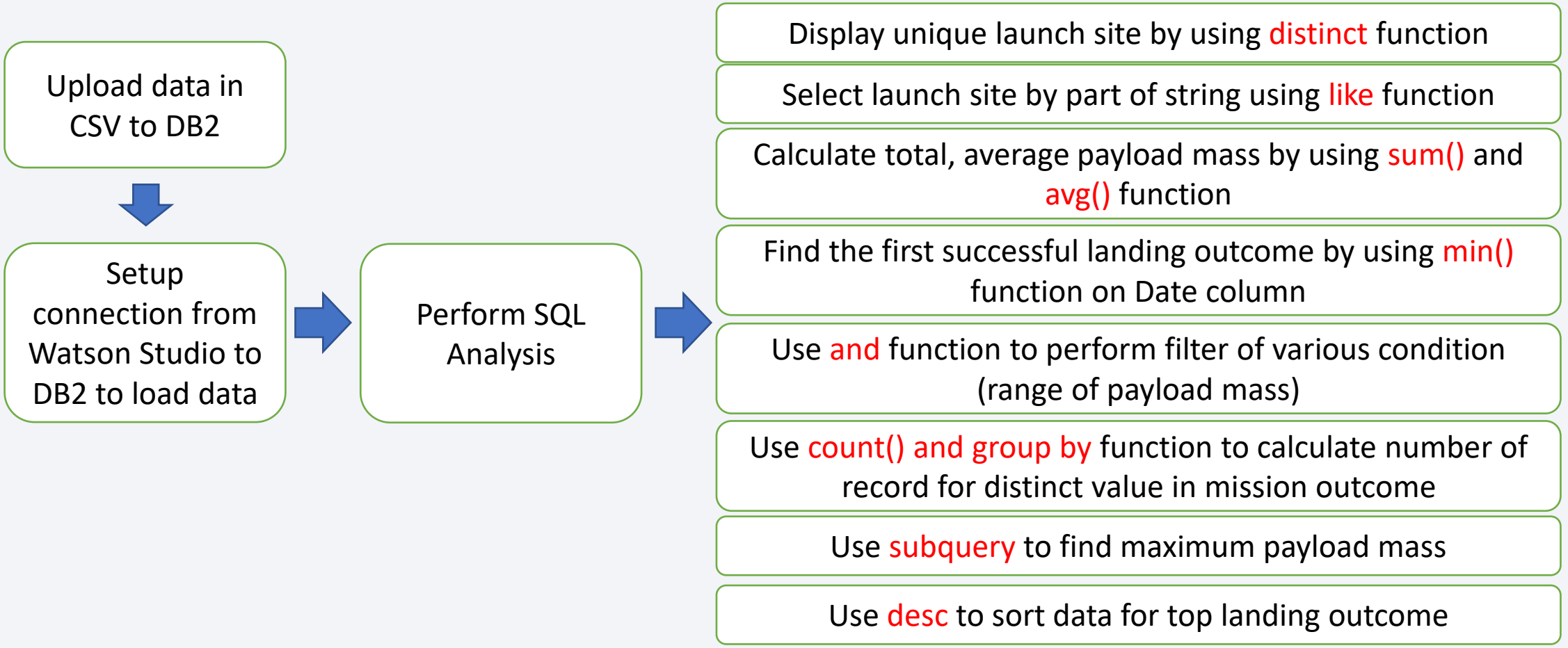
- Github link: https://github.com/khanhvuha/ibm_watson_studio/blob/master/ADSC:%20Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wranging

1. Check data type and missing values
   - df.dtypes
2. Calculate number of launch per launch site
   - df['LaunchSite'].value_counts()
3. Determine number of occurrence of each orbit
   - df['Orbit'].value_counts()
4. Calculate number and occurrence of mission outcomes
   - df['Outcome'].value_counts()
5. Create binary label from mission outcome
   - Enumerate()

- Github link: https://github.com/khanhvuha/ibm_watson_studio/blob/master/ADSC:%20EDA.ipynb

EDA with SQL



- Github link: https://github.com/khanhvuha/ibm_watson_studio/blob/master/ADSC:%20EDA%20with%20SQL.ipynb

EDA with Data Visualization

- There are 3 types of chart were plotted:

Chart 1

Scatterplot displays relation between features and mission outcome. This type of chart gives general insight about range and threshold in relation with success launch

Chart 2

Bar chart displays successful rate of each type of orbit

This type of chart provides general idea about preference orbit for future launching mission

Chart 3

Line chart displays successful rate over all of mission outcome over time

This chart indicates trend of outcome and can be used for future prediction of outcome following year

- Github link: https://github.com/khanhvuha/ibm_watson_studio/blob/master/ADSC:%20EDA%20with%20Visualization%20lab.ipynb

Build interactive map with Folium

Circle of location	Label tag of mission outcome	Distance to coastline and railway
• Add a circle and name on the location of the map of the US • To make a visible mark on the map where are the launch sites	• To provide how often each site are used and percentage of successful rate in each site	• Calculate distance from launch sites to coastline and railway • Show the idea about transporting rockets to launch sites

- Github link: https://github.com/khanhvuha/ibm_watson_studio/blob/master/ADSC:%20Data%20Visualization%20with%20Folium.ipynb

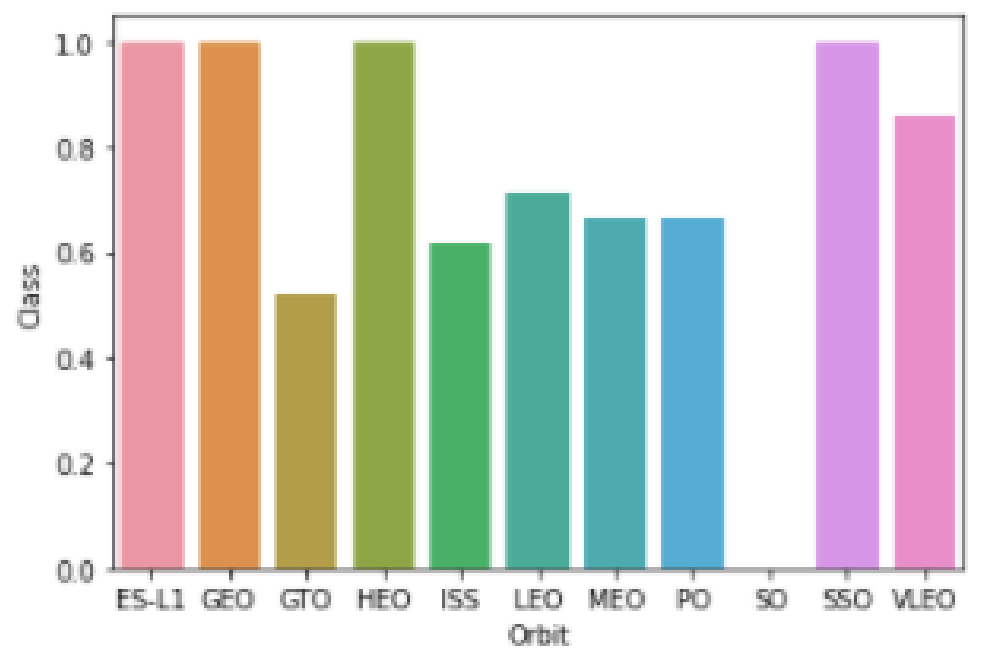
Build a Dashboard with Plotly Dash

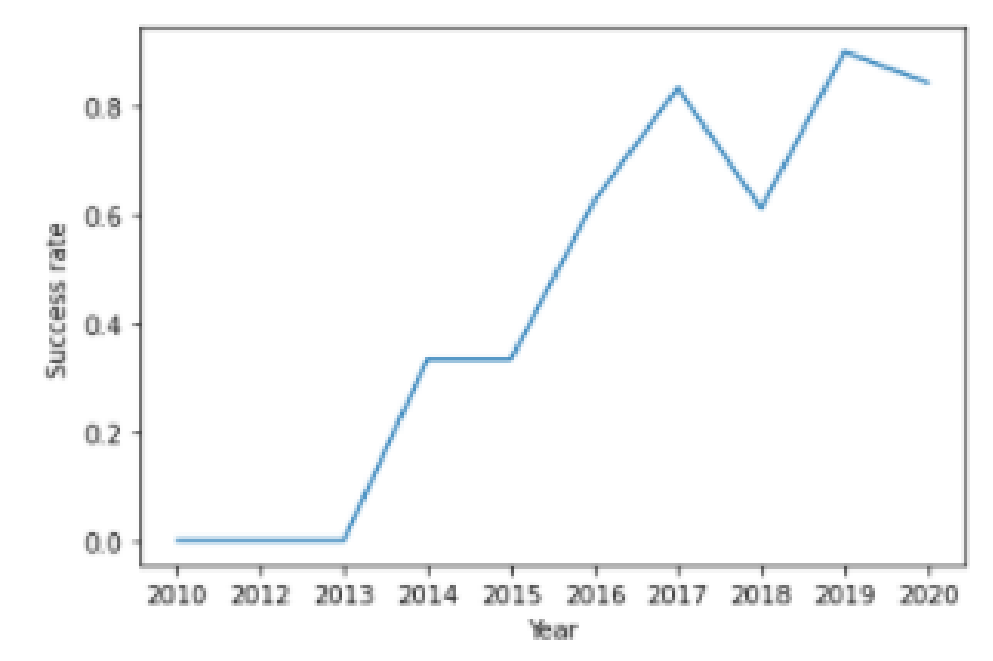
- There are 2 types of charts and interactions were added to Dashboad:
    - Pie chart: shows percentage of launch missions for each launch stie. A dropdown option was added to show percentage of successful mission outcome for each site separately
    - Scatterplot: describes relation between payload mass and mission outcomes. A slide range was added to interactively observe which range of payload mass has higher chance of successful
- Github link: https://github.com/khanhvuha/ibm_watson_studio/blob/master/ADSC%20Interactive%20Dashboard%20with%20Ploty%20Dash.ipynb

Predictive Analysis (Classification)

- The process of predictive analysis:
  - First, load data and standardize features for training using preprocessing.StandardScaler()
  - Second, split data into train and test set
  - Use train data for various model: logistics regression, support vector machine, KNN, decision tree
  - List all the parameter for each model and use grid search for the best parameter of each model with train data
  - Back test model with test data and compare result of each model
- Github link: https://github.com/khanhvuha/ibm_watson_studio/blob/master/ADSC:%20Machine%20learning%20prediction.ipynb

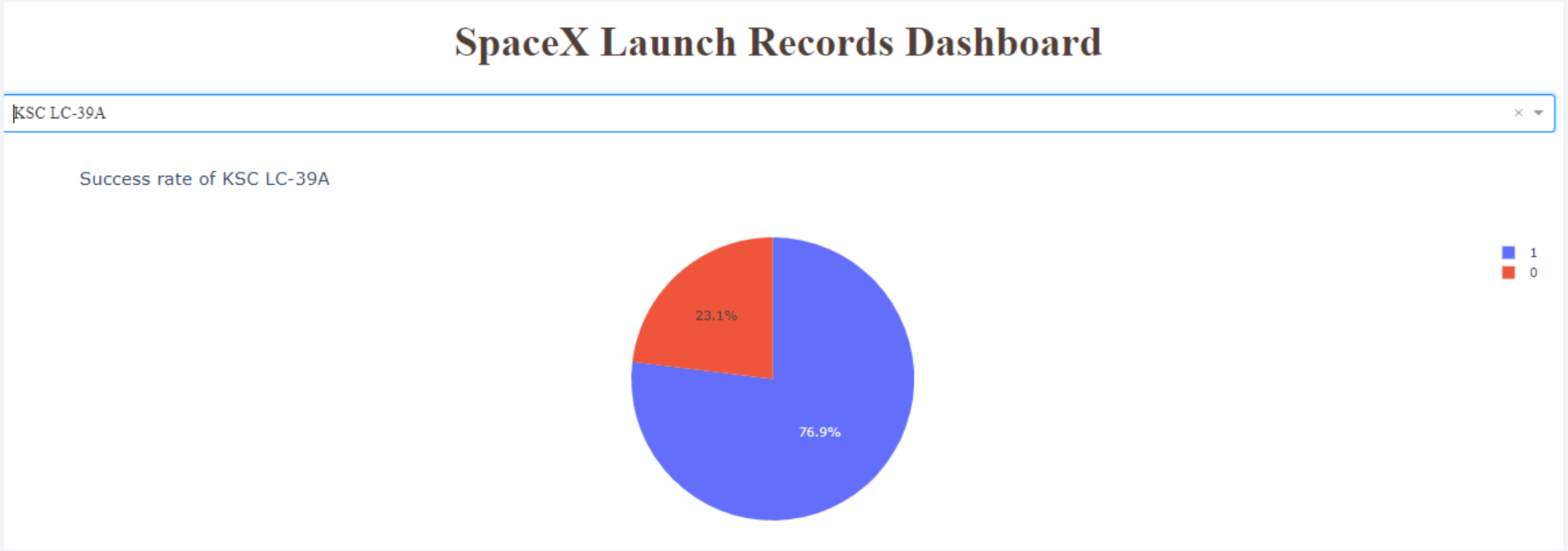
Results (1)





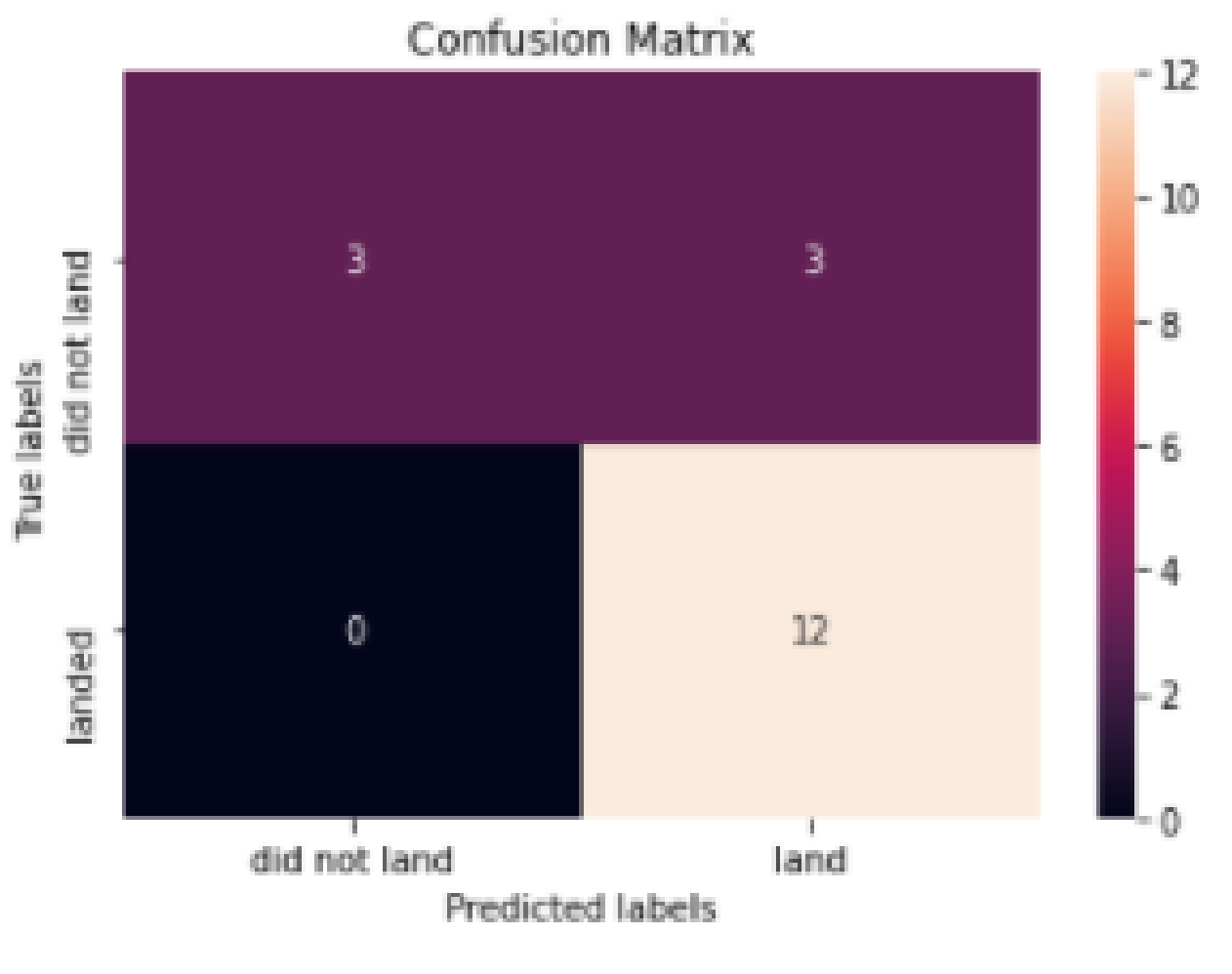
- Exploratory data analysis results:
  - ES-L1, GEO, HEO, SSO are type of orbit with most successful launch: 100%, followed by VLEO
  - The success rate overall are improving year by year, now above 80%

Results (2)



- KSC LC-39A is the site has the highest success rate in launching rocket at 76.9%

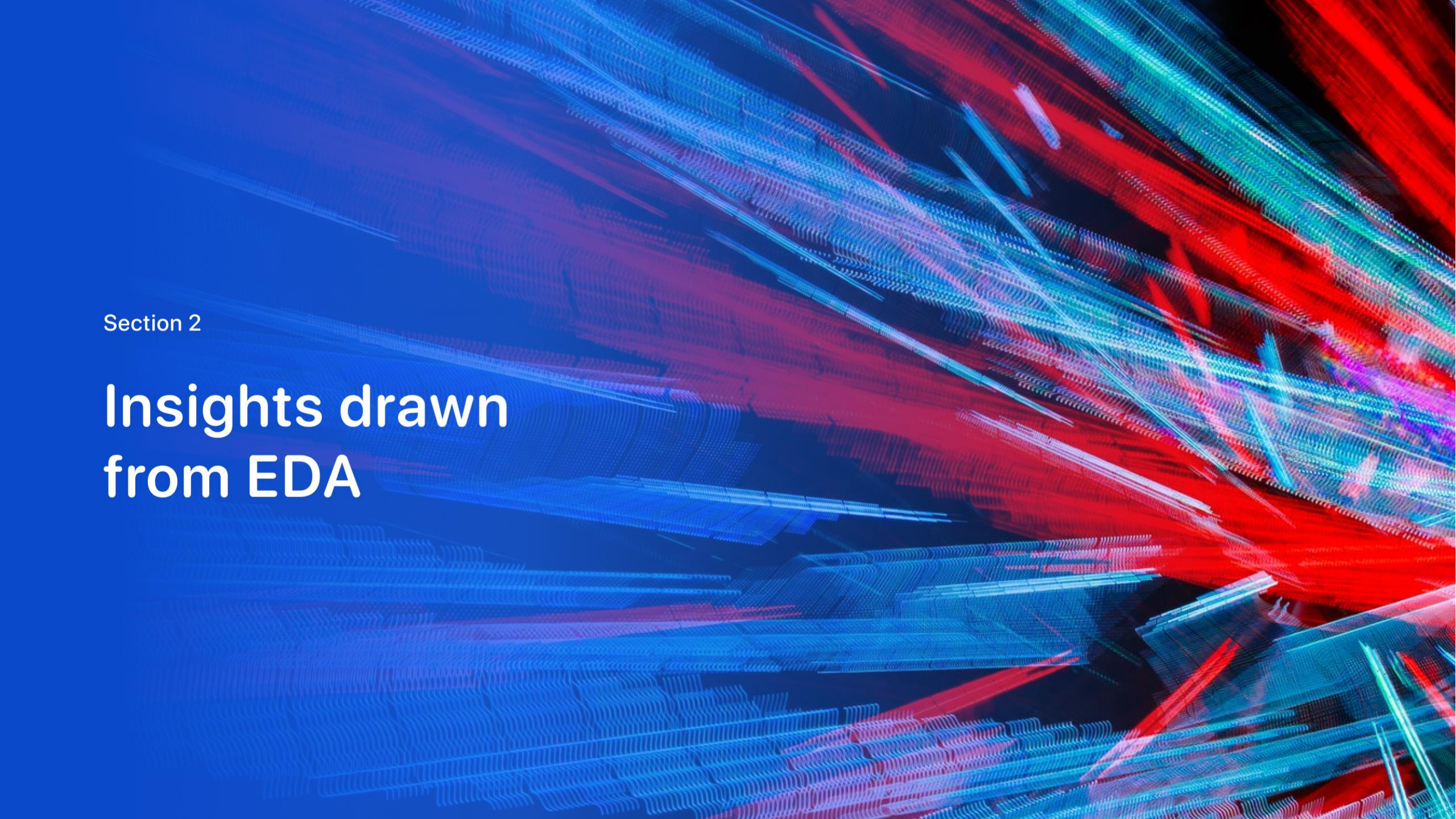
Results (2)



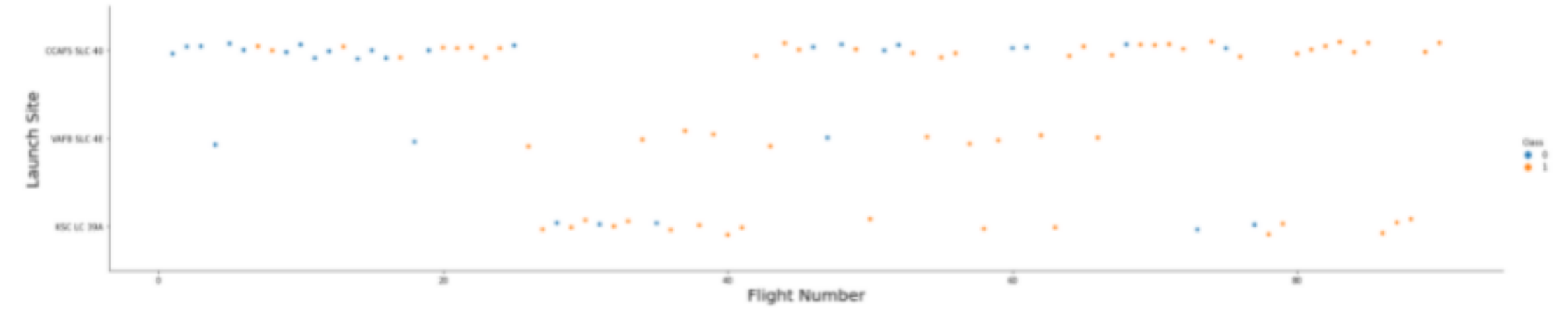
- All model have 83.3% ROC on test data
- The confusion matrix shows that models are good at predicting which rocket will land but not good at predicting which did not land

Section 2

Insights drawn from EDA

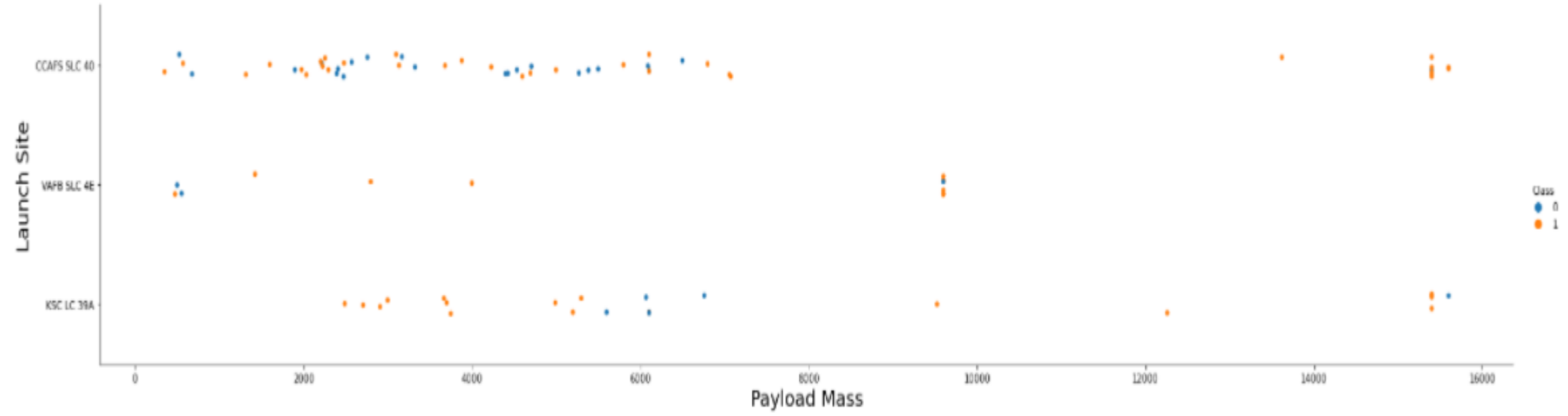


Flight Number vs. Launch Site



- Most Flight Number are in launch site CCAFS SLC 40
- From Flight Number 25 to 42 are executed at launch site KSC LC 39A and most of them are successful

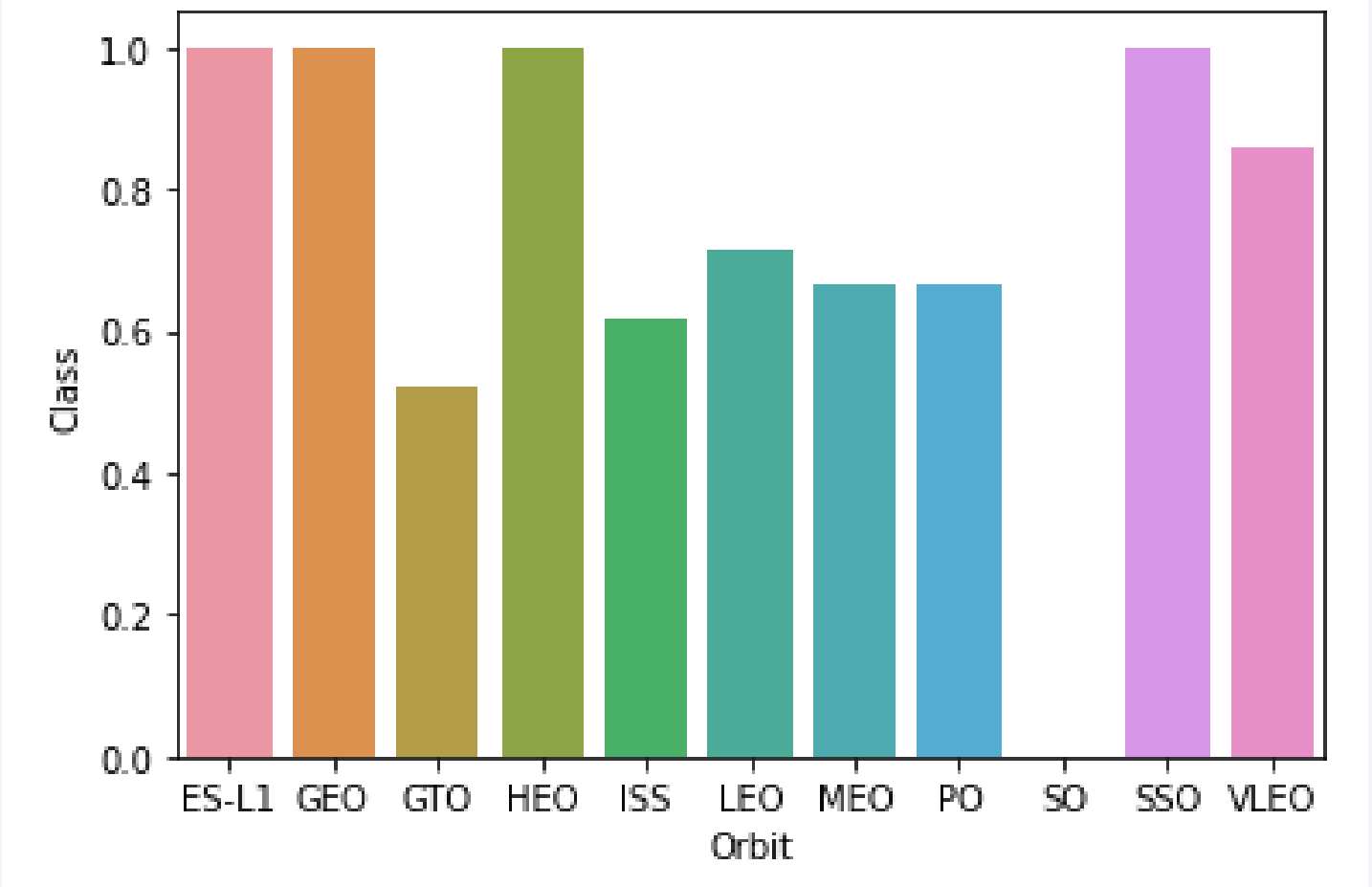
Payload vs. Launch Site



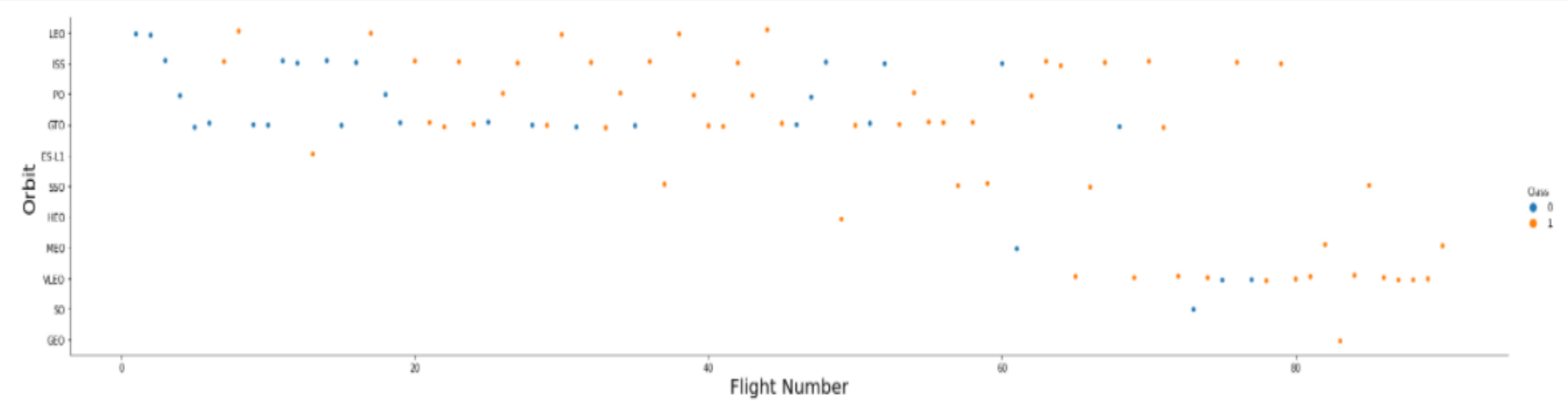
- With payload mass greater than 10000 kg, missiles were launched at CCAPS SLC 40 or KSC LC 39A
- With payload mass between 8000 and 10000 kg, missiles were launched at VAFB SLC 4E with very high success rate
- Payload mass smaller than 6000 kg launched at KSC LC 39A have very high success rate

Success Rate vs. Orbit Type

- Orbit ES-L1, GEO, HEO, SSO achieved 100% success rate
- Orbit VLEO follows with over 80% success rate
- The rest of orbit type reach around 60% success rate
- GTO only gets 50% of success rate while SO has no record on success launching mission

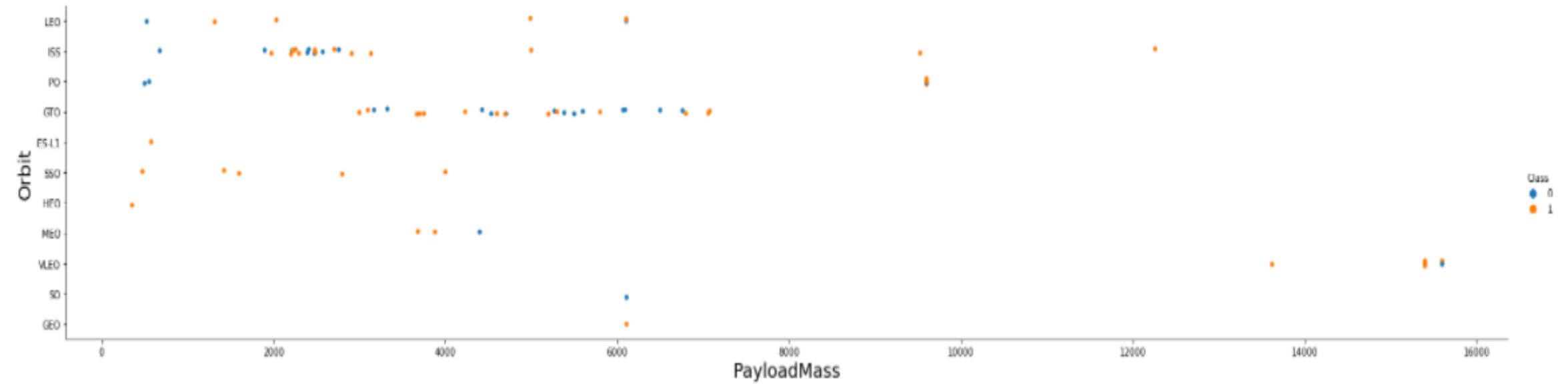


Flight Number vs. Orbit Type



- At the beginning, first 20 flights were missions to orbit LEO, ISS, PO, GTO and most of them were failed (only 5 success)
- From flight 20 to 60, still with those orbit but achieved better success mission (orange dots)
- After Flight 60, SpaceX has success mission with VLEO orbit (only 2 failed) and achieved higher success with ISS orbit (only 1 failed)
- They tried other type of orbits but mostly stay with VLEO, GTO, LEO, ISS, PO

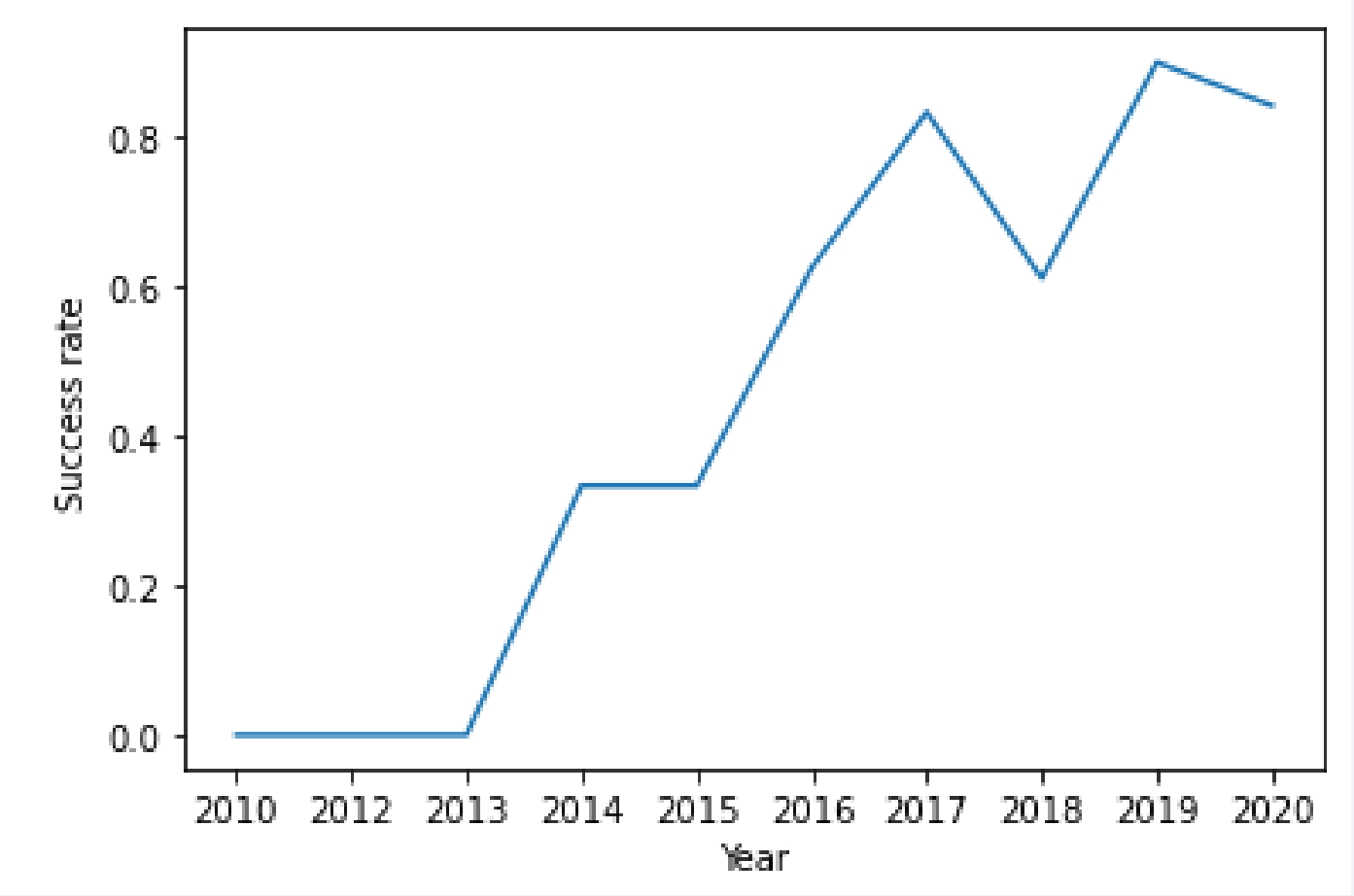
Payload vs. Orbit Type



- Payload greater than 12000 kg were tested with ISS and VLEO with only 1 failed mission
- ISS orbit was test most of the time with payload between 2000 and 3000 kg while GTO was tested with payload between 3000 and 8000 kg. The result was mixed between success and failed mission

Launch Success Yearly Trend

- Overall, there is an upward trend in success rate of SpaceX ‘s missions over the year
- Before 2014, there was no success mission. After that, the success rate rose steadily to 80% at 2017
- SpaceX success got a decline in 2018, back to 60%, but returned over 80% the following years



All Launch Site Names

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

- Query:
  - %sql select distinct launch_site from SPACEXTBL
- There are 4 sites of SpaceX:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

Launch Site Names 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Query: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
- First 5 records with launch site ‘CCA’ were in 2010, 2012 and 2013. All of them have success mission outcome but no landing attempt or failed landing attempt
- All of them aimed to reach LEO orbit with payload smaller than 1000 kg

Total Payload Mass

45,596 KG

- Query:
  - %sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)'
- Overall, NASA launched over 45.5 tons of materials and supplies using SpaceX's rockets

Average Payload Mass by F9 v1.1

- Query:
  - %sql select avg(payload_mass__kg_) from SPACEXTBL where booster_version = 'F9 v1.1'
- F9 V1.1 carried almost 3 tons of payload per mission

2928 KG/MISSION

First Successful Ground Landing Date

- Query:
  - %sql select min(date) from SPACEXTBL where landing__outcome = 'Success (ground pad)'
- December 25th 2015 was the very first date SpaceX achieved success landing on Ground Pad

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Query:
  - %sql select booster_version from SPACEXTBL where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ >4000 and payload_mass__kg_ <6000
- For payload between 4000 and 6000 kg, all successful drome ship landing were performed by 4 booster version, all of them in category ‘F9 FT’

Total Number of Successful and Failure Mission Outcomes

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Query:
  - %sql select mission_outcome, count(*) as count from SPACEXTBL group by mission_outcome
- For mission outcomes, only 1 was failed, and 100 were successful

Boosters Carried Maximum Payload

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

- Query:
  - %sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
- There were 12 boosters carried maximum payload from missions of SpaceX
- All of them are from group ‘F9 B5’

2015 Launch Records

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2015-01-10	09:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

- Query:
  - %sql select * from SPACEXTBL where landing__outcome = 'Failure (drone ship)' and year(DATE) = '2015'
- In 2015, there were 2 failed attempt landing on drone ship
- Both of them were launched by booster ‘F9 V1.1’ to LEO orbit for NASA

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

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

- Query:
  - %sql select landing__outcome, count(*) as count from SPACEXTBL where date >= '2010-06-04' and date <= '2017-03-20' group by landing__outcome order by count desc
- From 2010-06-04 to 2017-03-20, there were 31 missions of SpaceX, 10 of them had no attempt in landing missiles after use while 5 failed to land on drone ship
- 5 successful landing also on drone ship and 3 for ground pad

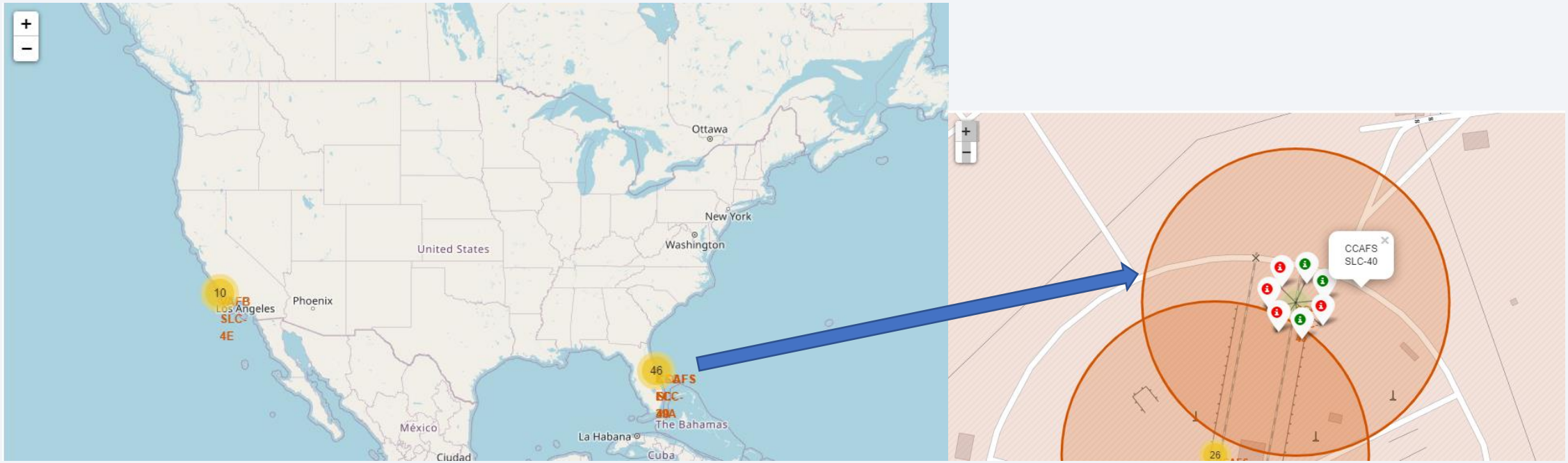
Section 4

Launch Sites Proximities Analysis

Launch site locations in the US

- All of them located near the coast, 1 in California while 3 others are in Florida

Landing outcomes in each site



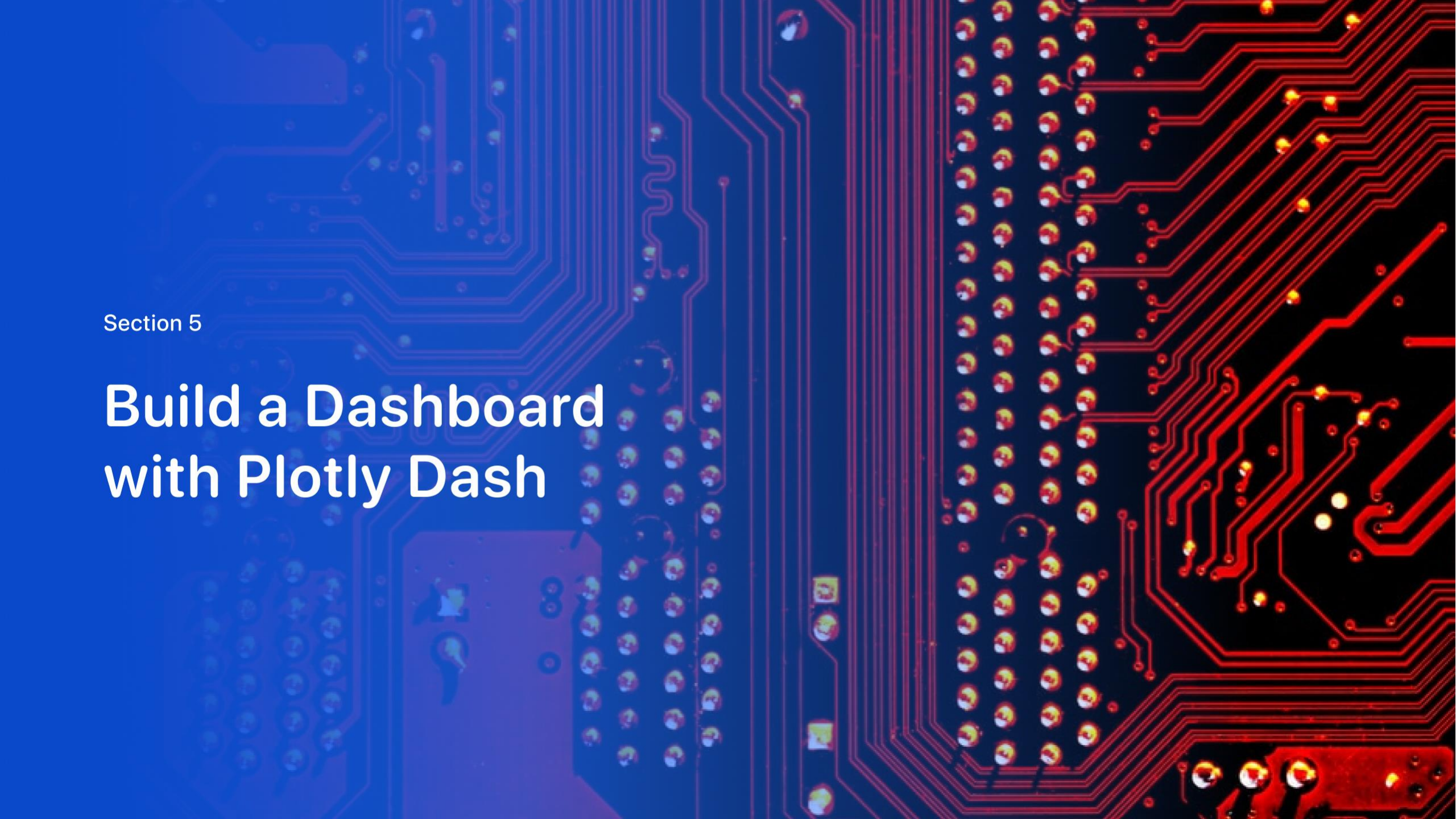
- In total, there were 56 landing attempts, 10 in the west coast and the rest in east coast
- For example, in CCAFS SLC-40 there were 3 successful landing while 4 were failed. The successful landing were marked in green while failed one in red. Each site has a circle and name tag to be marked on the map

Distance to coast and railway

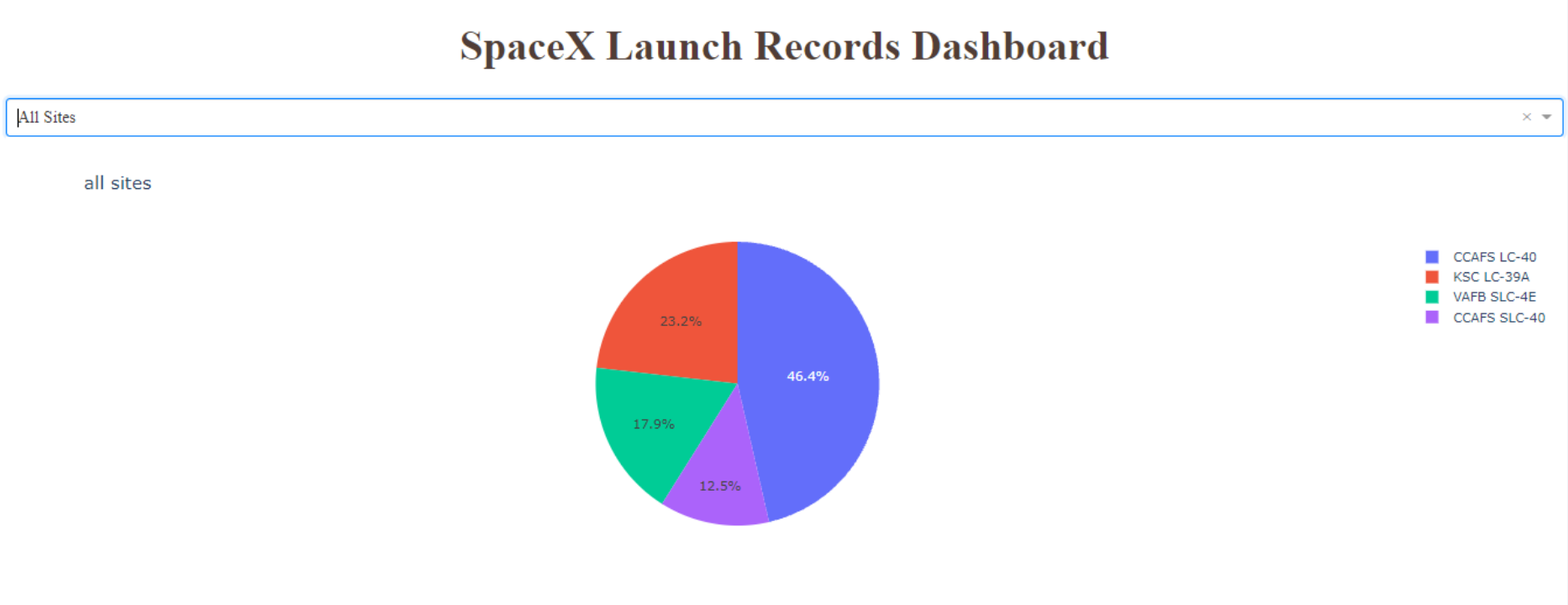
- Distance from each site to closest railways and coasts were marked in blue line with value in km
- All launch sites are very close to railways and coastline, all of them are under 2 km

Section 5

Build a Dashboard with Plotly Dash

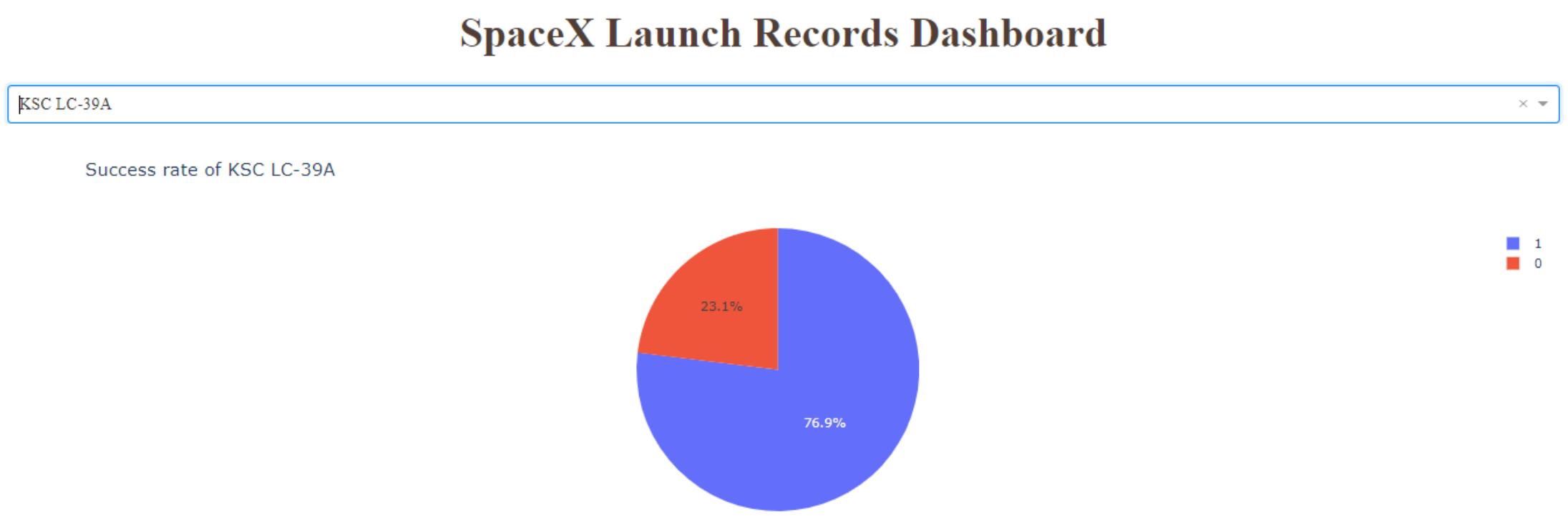


Launch records of all sites



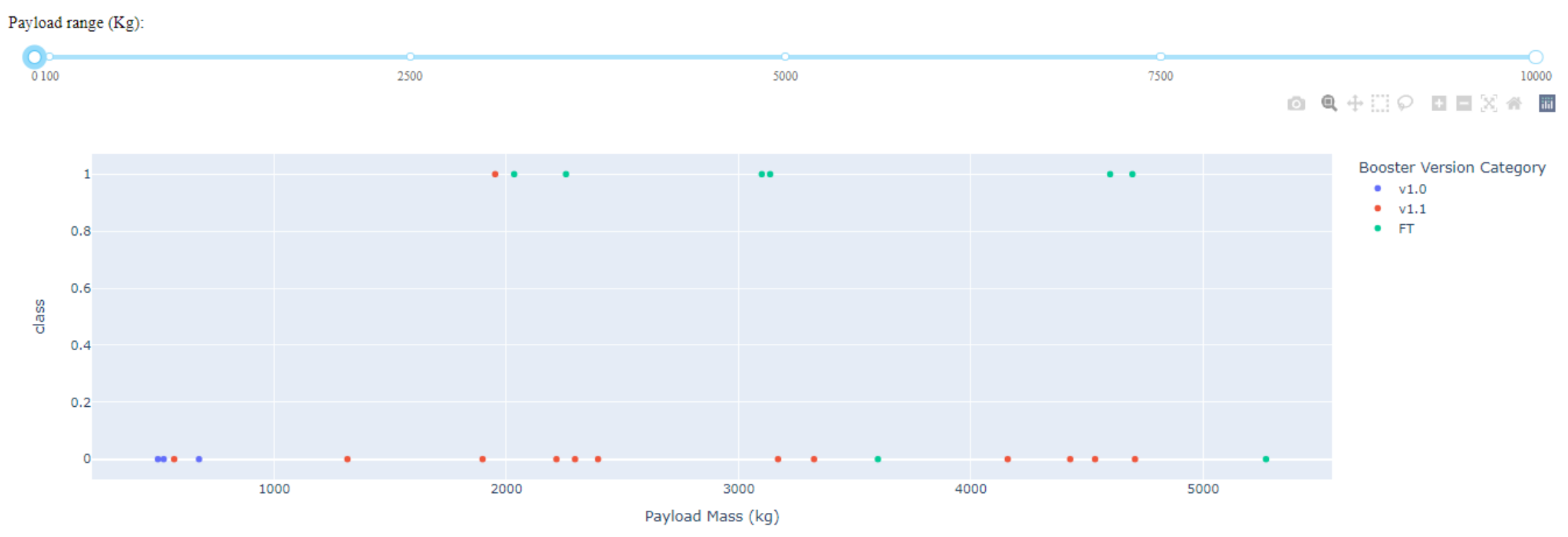
- Nearly a half (46.4%) of successful landing attempts were made at CCAFS LC-40

Launch site with highest success landing rate



- KSC LC-39A was the site with highest success landing rate at 76.9%, higher than CCAFS LC-40 at 73.1%

Payload vs booster version



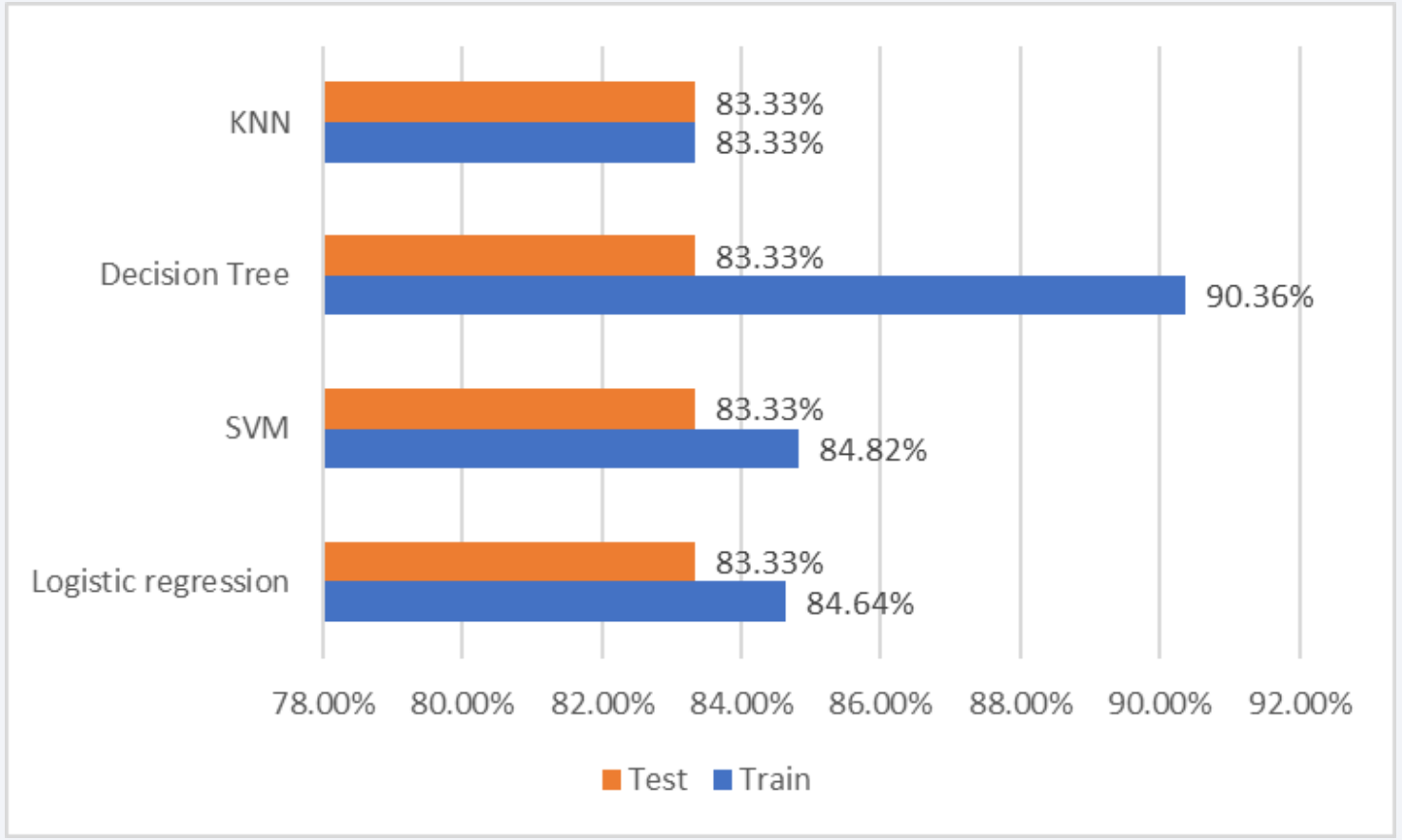
- For all sites, V1.0 has the lowest success rate (0%) when tested with small payload (<1 ton)
- V.1.1 also had only 1 success attempt when tested with wider range of payload

Section 6

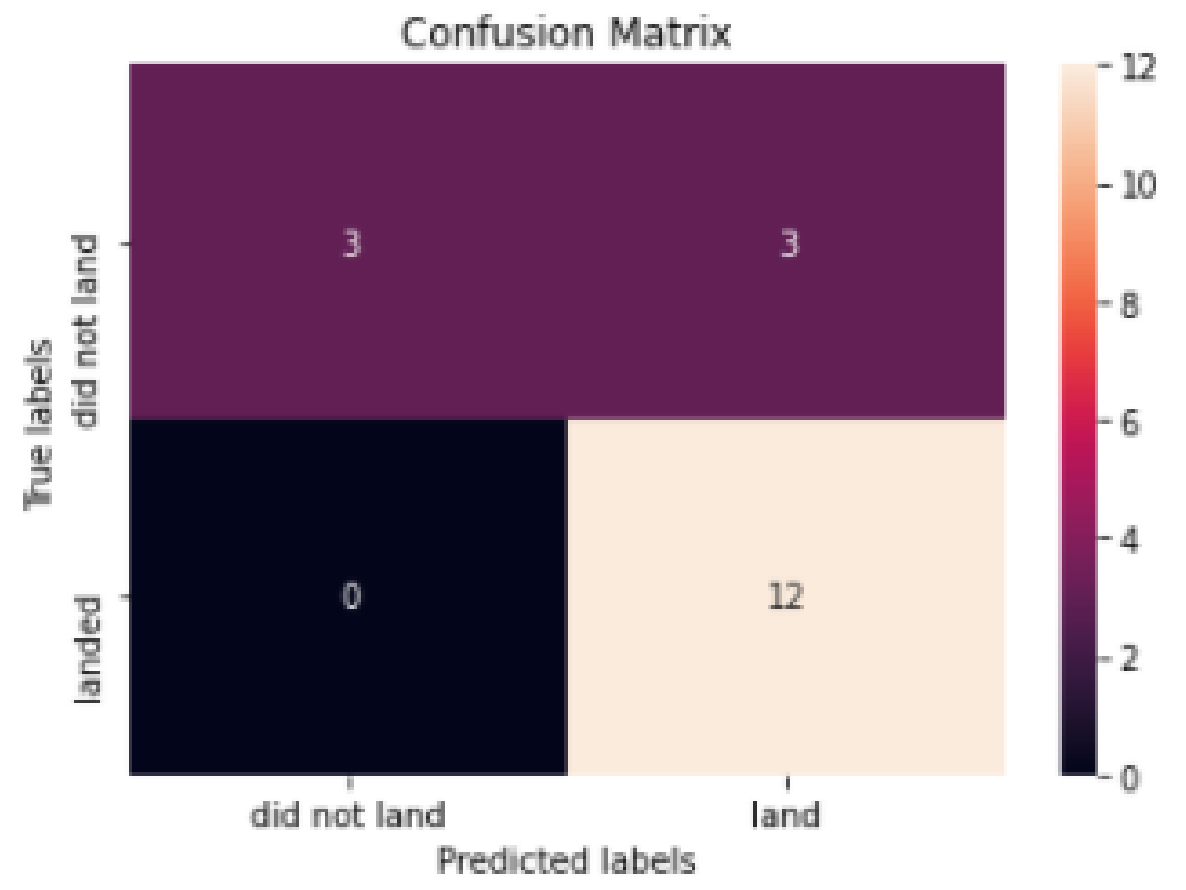
Predictive Analysis (Classification)

Classification Accuracy

- All models have same accuracy on test set at 83.3%
- Decision tree has highest accuracy on train set. However, this can be considered as an overfit model when we have wide gap between train and test set (~7%)
- Therefore, KNN can be considered as the best



Confusion Matrix of KNN



- Confusion matrix of KNN shows that:
  - Model is very good at predicting which landing attempts will success (true positive)
  - Model is not good at predicting which did not land where 3 failed attempts were classified as success (false positive)

Conclusions

Orbit ES-L1, GEO, HEO, SSO are among the best orbits for success landing outcome

KSC LC-39A has the highest landing rate

Payload greater than 12000 kg were tested with ISS and VLEO has the highest success rate

KNN is the best model for predicting outcome, good at predict success landing (100%) but not at predicting failed attempts (50%)

Appendix

- Github project:
- https://github.com/khanhvuha/ibm_watson_studio/tree/master

Thank you!