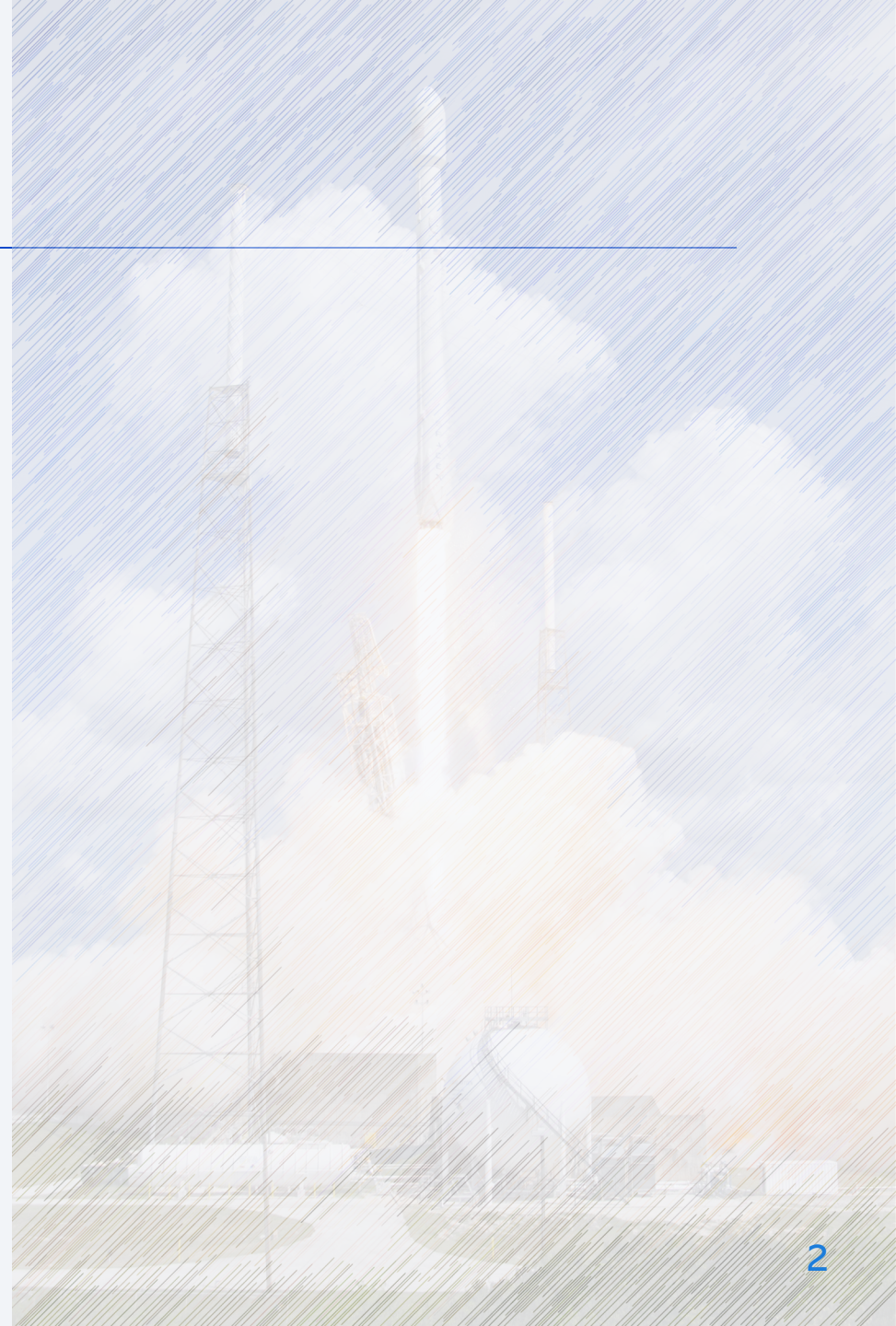
Winning Space Race with Data Science

Patrycja Naumczyk
17 February 2022

Presentation outline

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

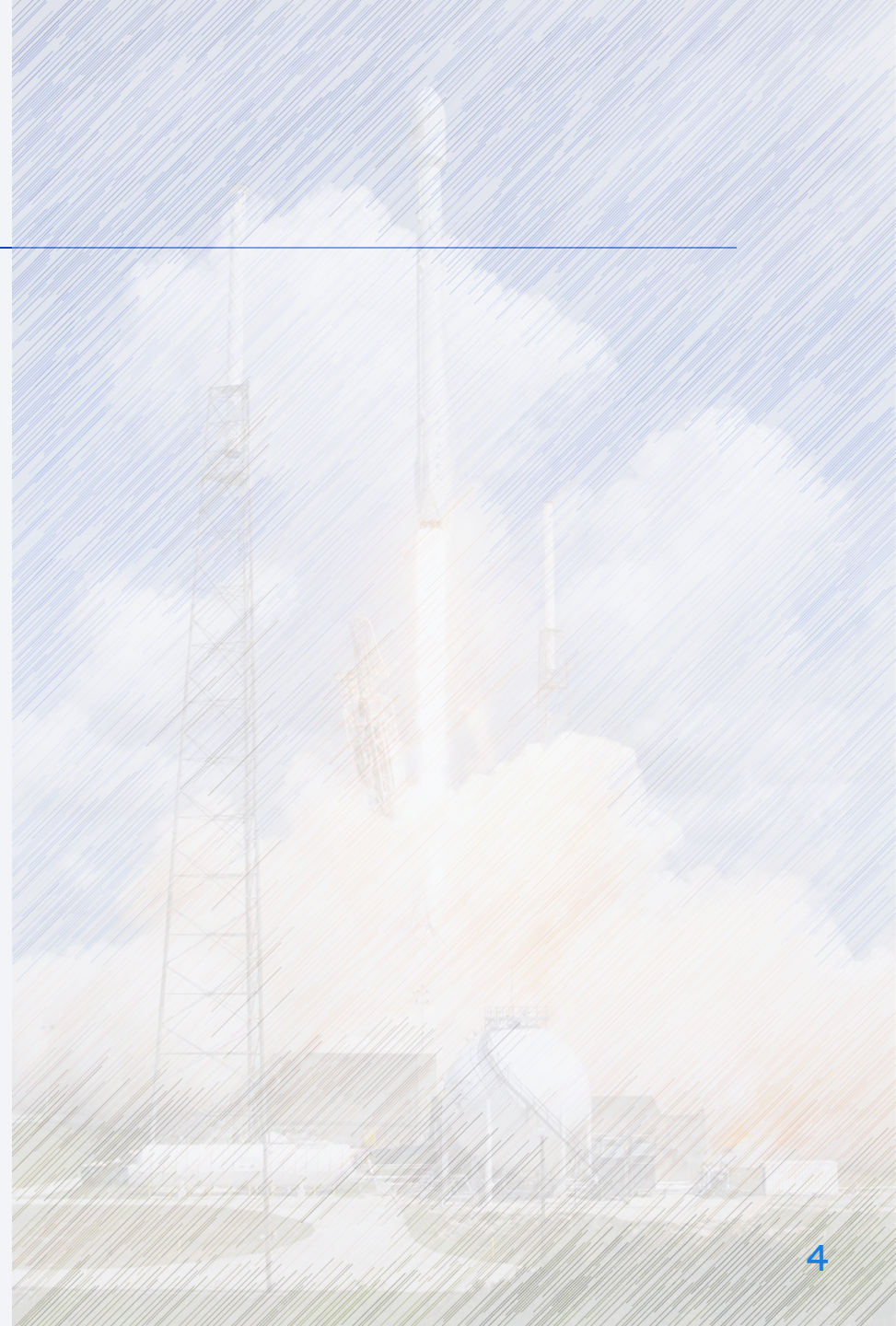


Executive Summary

- **Methodology**:
  Based on publicly available data, past SpaceX launch data was collected and parsed. Features, most likely corresponding with positive landing outcome, were selected. Upon them explanatory data analysis was performed, with visualization, SQL queries, as well as interactive dashboard and Folium maps. Additionally, four different classification models were tested for the best accuracy.
- **Results**:
  The outcome of a launch is dependent on the launch site, the orbit, the payload mass, and the booster version used. The highest probability for success comes from launches performed at the west coast, with medium payload mass (2k-6k kg), and booster version of FT. The classification models all performed comparably.

Introduction

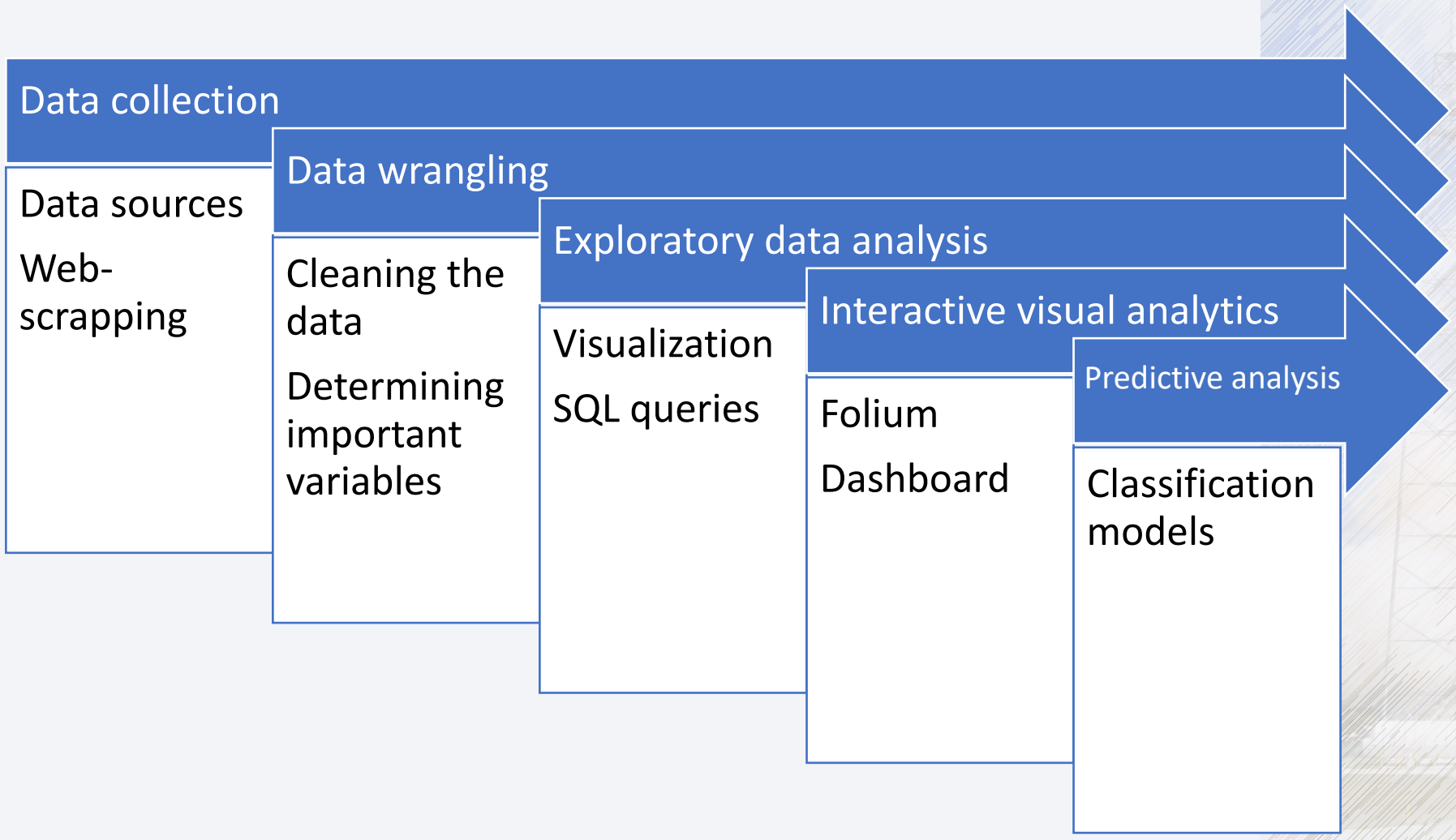
- Space Y – the new alternative to SpaceX
  - We are less experienced, but more motivated!
  - Additional founds were gathered, so let’s start moving!
- Problems to be addressed:
  - How to compete efficiently?
  - What determines the success of SpaceX launches?
  - Build a predictive model for determining success



Section 1

Methodology

Methodology - overview



When is the launch worthwhile ?

Data Collection – two-way methodology

SpaceX API

- Get response from https://api.spacexdata.com/v4/launches/past in JASON format
- Convert to pandas dataframe
- Use API for determining Booster, Launchpad, payload and core

Web-scrapping

- Get response from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922 in HTML format
- Apply BeautifulSoup to parse the data
- Extract the data to create pandas dataframe

Reference:

- SpaceX API notebook: https://github.com/patrycja-naumczyk/IBM-cert-final/blob/143c0f278c9ae9f2d288a30f76731eb11fb19f54/1_1_Data_collection.ipynb
- Web-scrapping notebook: https://github.com/patrycja-naumczyk/IBM-cert-final/blob/143c0f278c9ae9f2d288a30f76731eb11fb19f54/1_2_Web_scrapping.ipynb

Data Wrangling methodology

Basic wrangling at the data collection stage:
sampling, dealing with nulls, removing Falcon 1 launches, changing data format

Calculating missing values, identifying data types, calculating number of launches for each site, number of occurrences of each orbit

Creating launch outcome label

Reference:

- Basic wrangling notebook: https://github.com/patrycja-naumczyk/IBM-cert-final/blob/143c0f278c9ae9f2d288a30f76731eb11fb19f54/1_1_Data_collection.ipynb
- Data wrangling with basic EDA notebook: https://github.com/patrycja-naumczyk/IBM-cert-final/blob/143c0f278c9ae9f2d288a30f76731eb11fb19f54/1_3_Data_wrangling_EDA.ipynb

EDA with Data Visualization methodology

- Visualization included charts:
  - Flight Number vs. Payload Mass (with outcome)
    to verify improvement over time (and the importance of payload mass)
  - Flight Number vs Launch Site (with outcome)
    to compare different sites' outcomes
  - Payload Mass vs Launch Site (with outcome)
    to compare different sites and their payloads
  - Success Rate vs Orbit Type
    to determine which orbits have the best success rate
  - Flight Number vs Orbit Type (with outcome)
    to determine the existence of relationship
  - Payload Mass vs Orbit Type (with outcome)
    to compare success over different orbits and payloads
  - Launch success yearly trend
    to observe the trend

Reference:

- EDA with visualization notebook: https://github.com/patrycja-naumczyk/IBM-cert-final/blob/143c0f278c9ae9f2d288a30f76731eb11fb19f54/2_2_Explanatory_Data_Analysis_Vizualization.ipynb

EDA with SQL methodology

- SQL queries performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass.
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - 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

Reference:

- EDA with SQL notebook: https://github.com/patrycja-naumczyk/IBM-cert-final/blob/143c0f278c9ae9f2d288a30f76731eb11fb19f54/2_1_Explanatory_data_analysis_SQL.ipynb

Interactive Map with Folium methodology

- Following maps were created:
  - Marked launch sites by their latitude and longitude coordinates
    to visualize the origin of launches
  - Marked successful/failed launches for each site
    with MarkerCluster for better transparency
  - Calculated and marked distance from one of the sites to proximities
    verified with a PolyLine

Reference:

- Interactive map notebook: https://github.com/patrycja-naumczyk/IBM-cert-final/blob/143c0f278c9ae9f2d288a30f76731eb11fb19f54/3_1_Interactive_Dashboard.ipynb

Dashboard with Plotly Dash methodology

- Interactive dashboard was created as follows:
  - User choses whether to display stats for all sites vs particular site (dropdown menu)
  - User choses payload range for the stats (slider)
- For “All sites” presented plots were:
  - Pie-chart presenting success rate for all sites
    to determine the most successful site
  - Scatter-plot comparing Launch Outcome vs Payload Mass with marked Booster Versions for all sites
    to determine the most successful combination of the variables
- For every specific site presented plots were:
  - Pie-chart presenting success rate for the site
    to present the success rate
  - Scatter-plot comparing Launch Outcome vs Payload Mass with marked Booster Version for the site
    to determine the most successful combination for the site

Reference:

- Dashboard Python code:
  https://github.com/patrycja-naumczyk/IBM-cert-final/blob/143c0f278c9ae9f2d288a30f76731eb11fb19f54/3_2_spacex_dash_app.py

Predictive Analysis (Classification) methodology

Features selection (with dummy encoding):

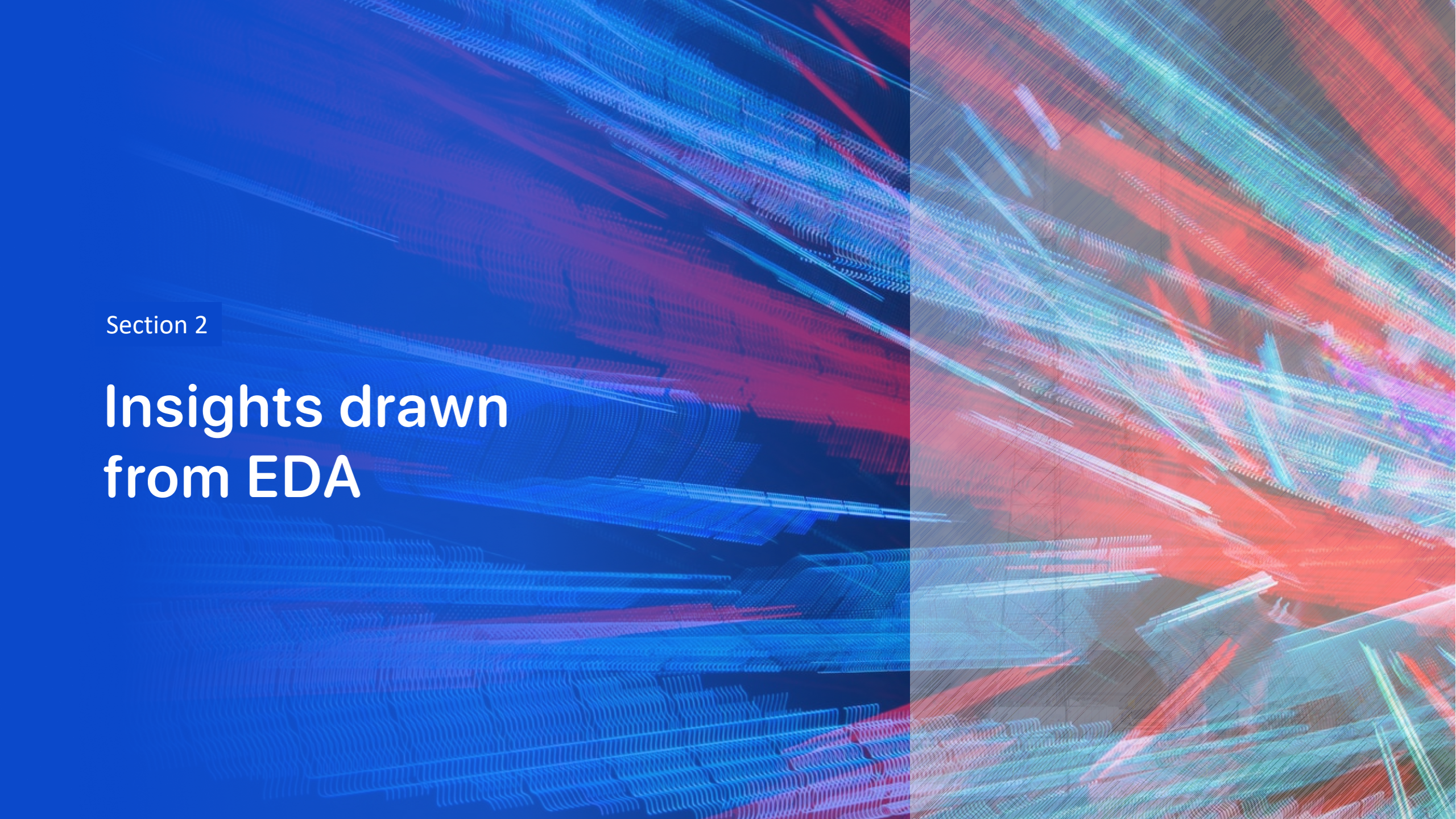
- Payload mass
- Booster version
- Orbit
- Launch site
- Flights
- Grid Fins
- Reused
- Legs
-Landing pad
-Block
- Reused count
- Serial
-Flight number

Standardization and train/test split of the data (80%/20%)

Model tested:	GridSearch cross-validation:
Logistic regression	'C':[0.01,0.1,1], 'penalty':['l2'], 'solver':['lbfgs','newton-cg']
SVM	'kernel':('linear', 'rbf','poly','rbf', 'sigmoid'), 'C': np.logspace(-3, 3, 5), 'gamma':np.logspace(-3, 3, 5)
Decision Tree	'criterion': ['gini', 'entropy'], 'splitter': ['best', 'random'], 'max_depth': [2*n for n in range(1,10)], 'max_features': ['auto', 'sqrt'], 'min_samples_leaf': [1, 2, 4], 'min_samples_split': [2, 5, 10]
K-nearest neighbours	'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'], 'p': [1,2]

Reference:

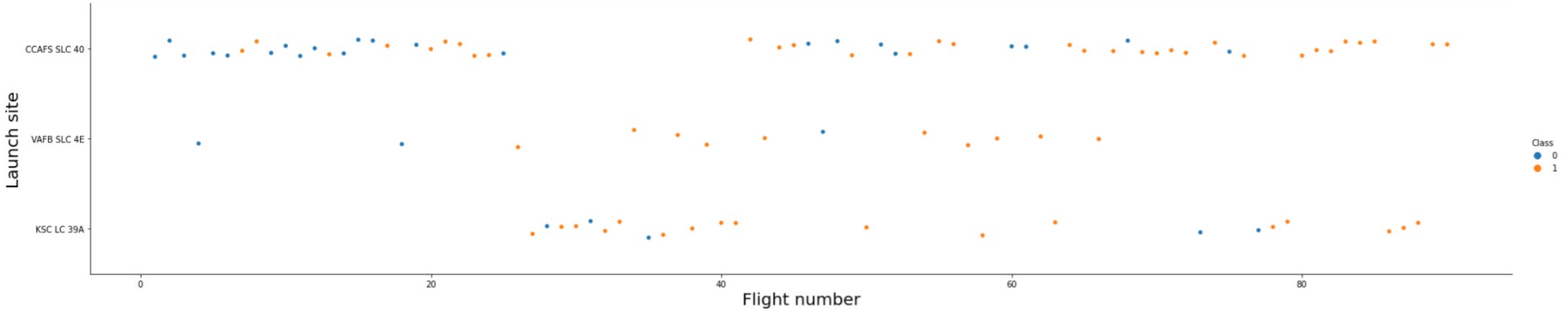
- Machine Learning notebook: https://github.com/patrycja-naumczyk/IBM-cert-final/blob/143c0f278c9ae9f2d288a30f76731eb11fb19f54/4_Machine_learning.ipynb



Section 2

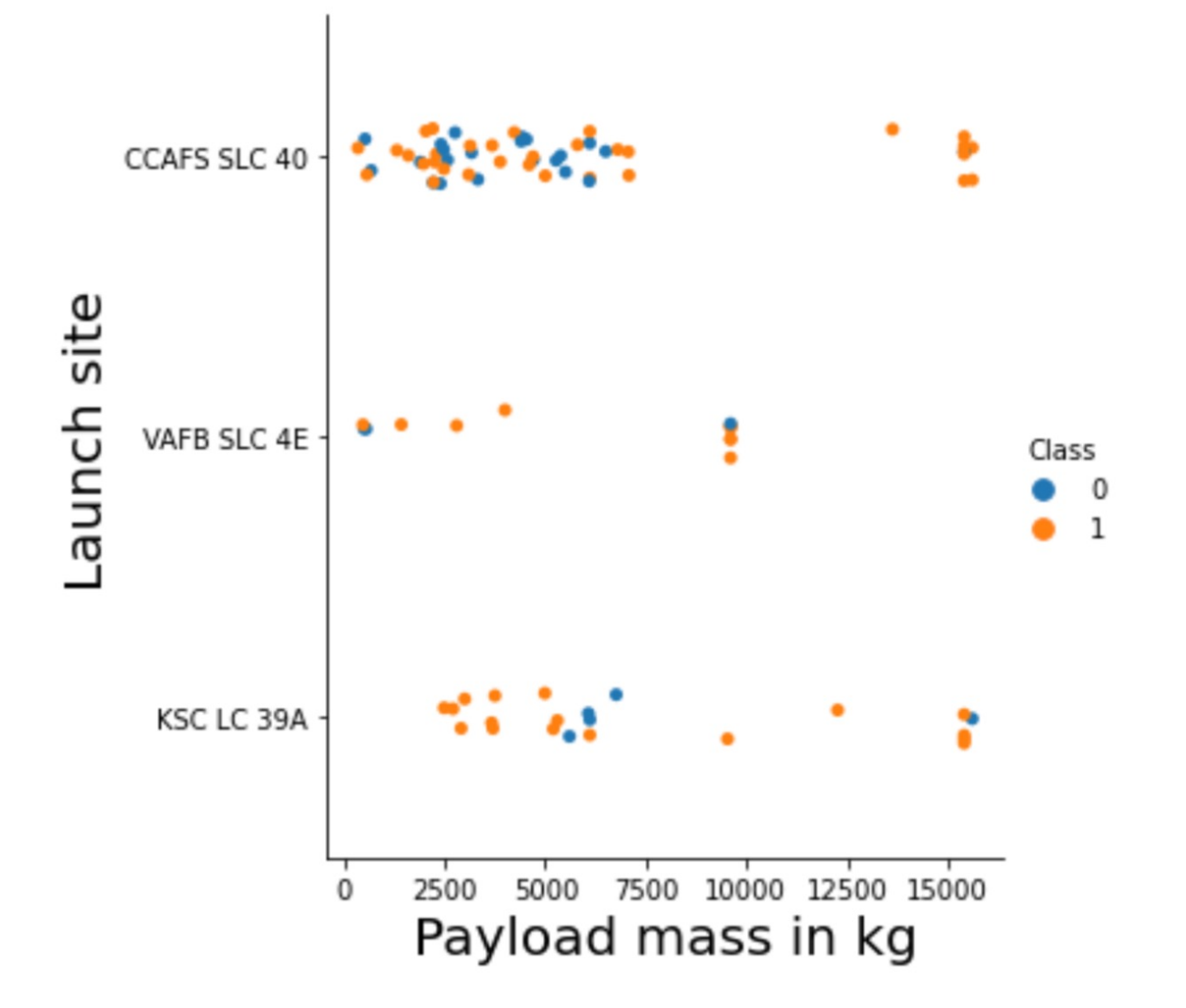
Insights drawn from EDA

Flight Number vs. Launch Site



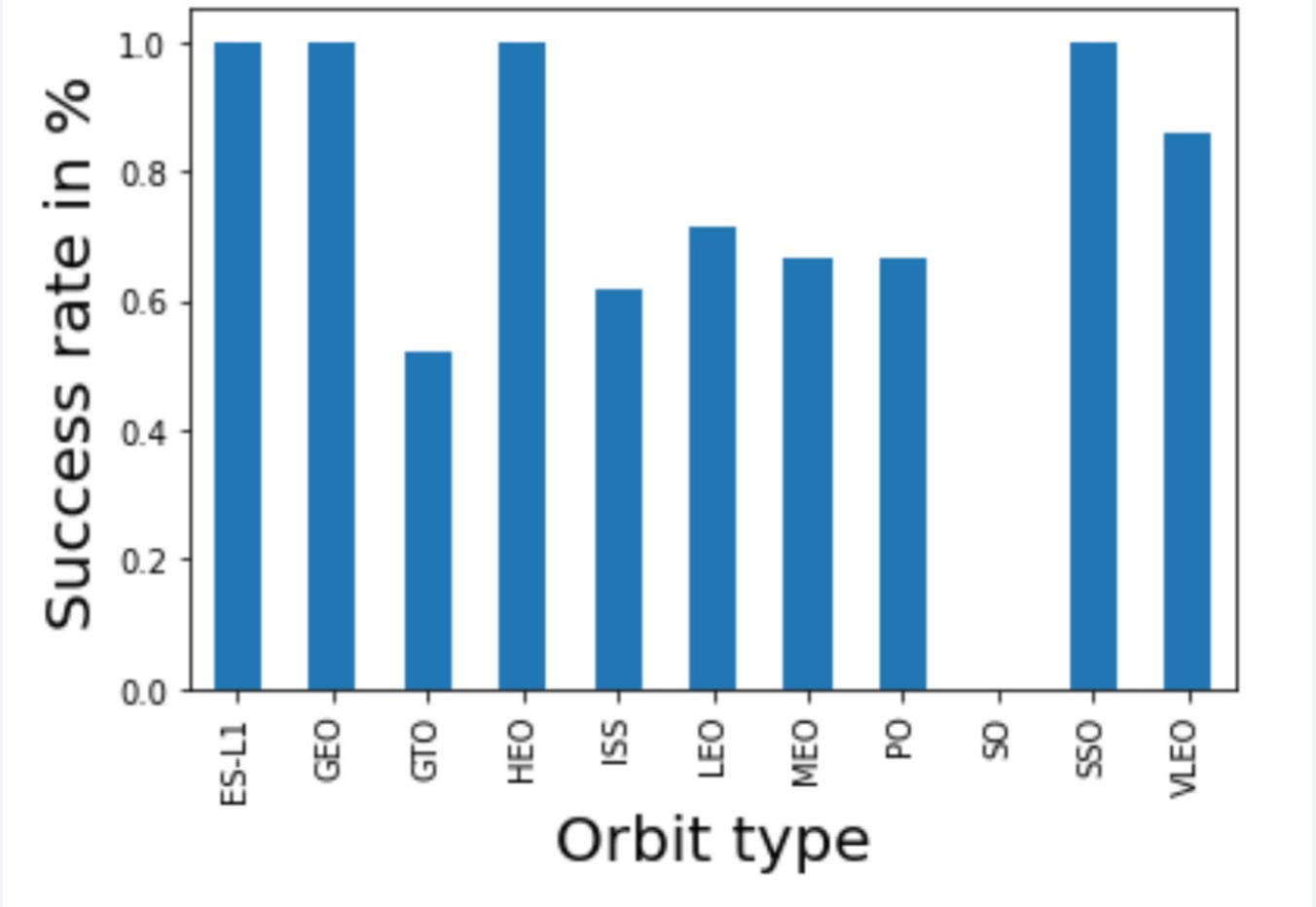
- Class depicts success label (0/blue: failure, 1/orange: success)
- VAFB-SLC launchsite appears to have the best success rate
- Most launches were performed from CCAFS SLC 40 site
- As the project progressed, the success rate increased

Payload vs. Launch Site



- Class depicts success label (0/blue: failure, 1/orange: success)
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000)

Success Rate vs. Orbit Type

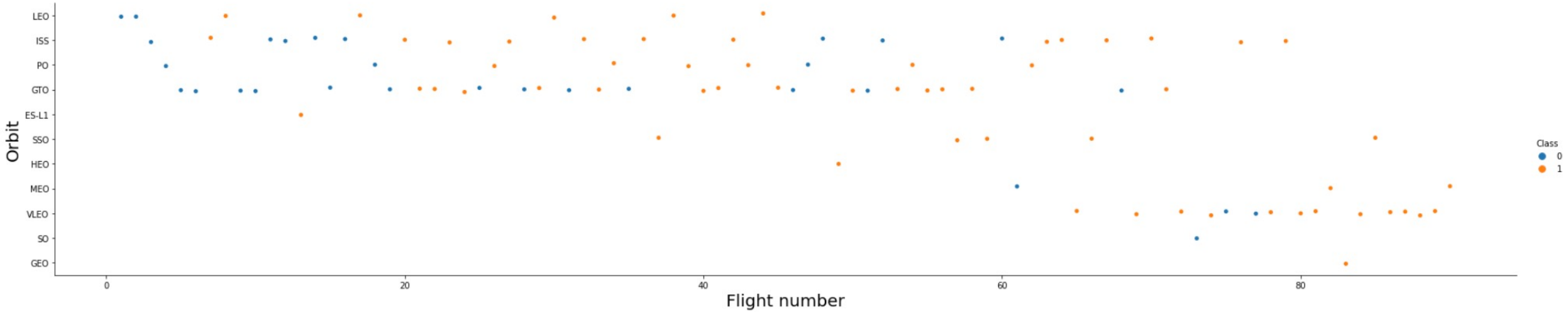


- The highest success rate was for orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO

  where all launches were a success

ES-L1: Lagrange point 1 orbit; GEO/GTO: Circular/geosynchronus orbit; HEO: Geocentric orbit above GEO; ISS: Modular space station orbit; LEO: Low Earth orbit; MEO: Intermediate circular orbit; PO: Pole orbit; SO/SSO: Sun-synchronous orbit; VLEO: Very low Earth orbit

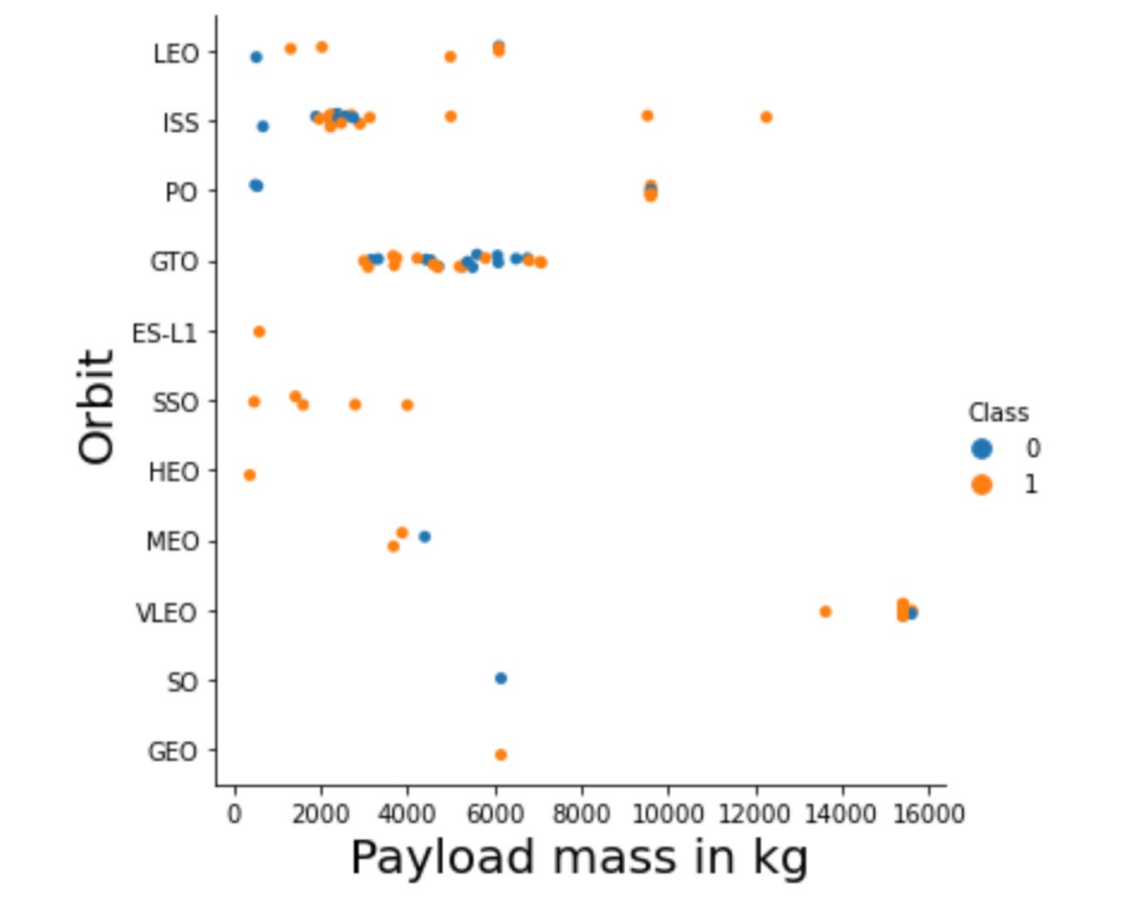
Flight Number vs. Orbit Type



ES-L1: Lagrange point 1 orbit; GEO/GTO: Circular/geosynchronus orbit; HEO: Geocentric orbit above GEO; ISS: Modular space station orbit; LEO: Low Earth orbit; MEO: Intermediate circular orbit; PO: Pole orbit; SO/SSO: Sun-synchronous orbit; VLEO: Very low Earth orbit

- Class depicts success label (0/blue: failure, 1/orange: success)
- In the LEO/VLEO orbit the success appears related to the number of flights
- No relationship between flight number when in GTO orbit

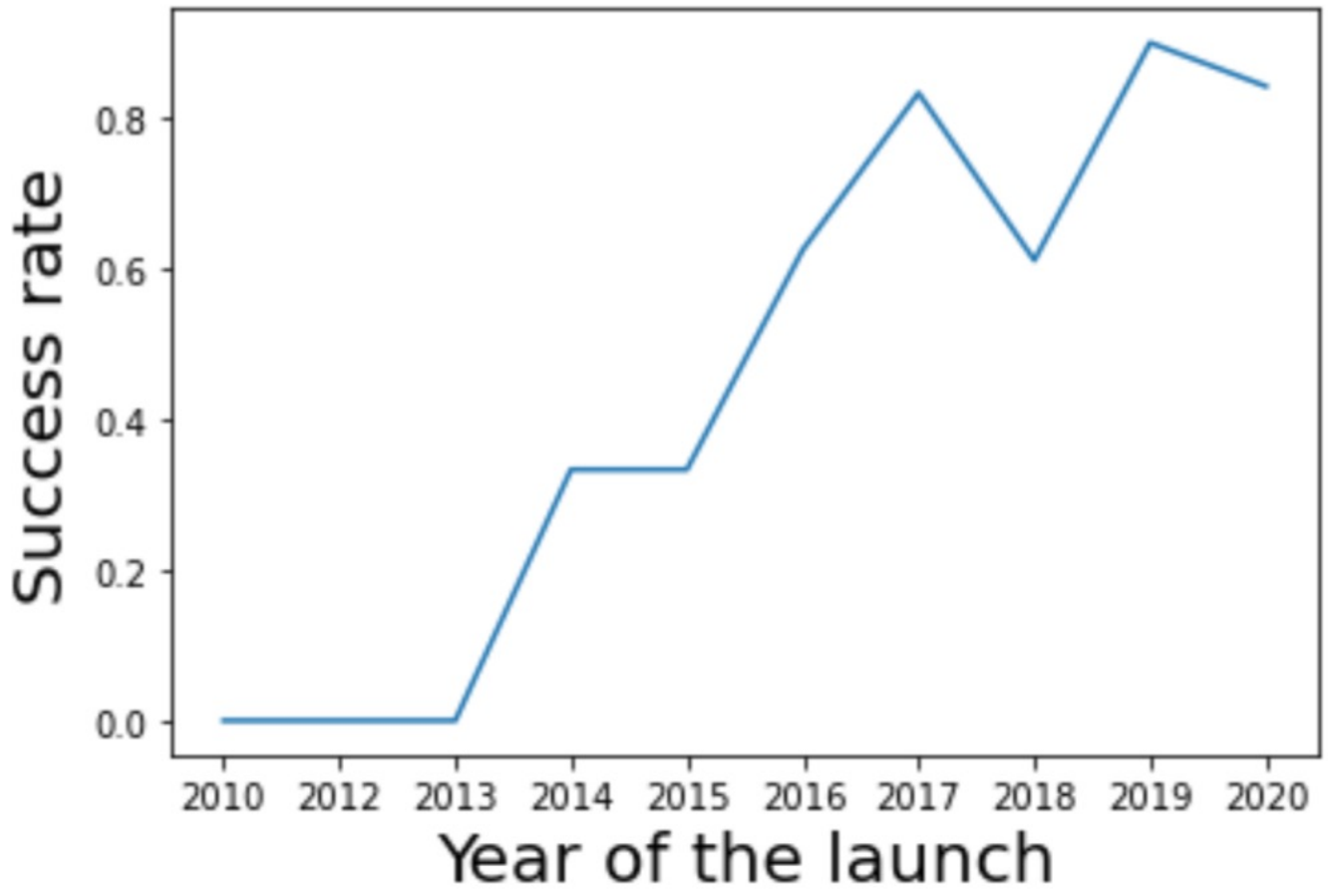
Payload vs. Orbit Type



ES-L1: Lagrange point 1 orbit; GEO/GTO: Circular/geosynchronus orbit; HEO: Geocentric orbit above GEO; ISS: Modular space station orbit; LEO: Low Earth orbit; MEO: Intermediate circular orbit; PO: Pole orbit; SO/SSO: Sun-synchronous orbit; VLEO: Very low Earth orbit

- Class depicts success label (0/blue: failure, 1/orange: success)
- With heavy payloads the successful landing or positive landing rate are more probable for Polar, LEO and ISS
- For GTO we cannot distinguish this well

Launch Success Yearly Trend



- The sucess rate kept increasing since 2013 till 2020

All Launch Site Names

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

- There were four unique launch sites
- Two in close proximity (CCAFS LC-40 and CCAFS SLC-40)

Example records from Launch Site Names Begining 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

- The results were limited to 5, as requested

Total Payload Mass

1
48213

- Total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

1
2534

- Average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

	1
	2015-12-22

- The date when the first successful landing outcome in ground pad was achieved

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

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

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

- The total number of successful and failure mission outcomes

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

- The names of the booster versions which have carried the maximum payload mass

2015 Failure Launch Records

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

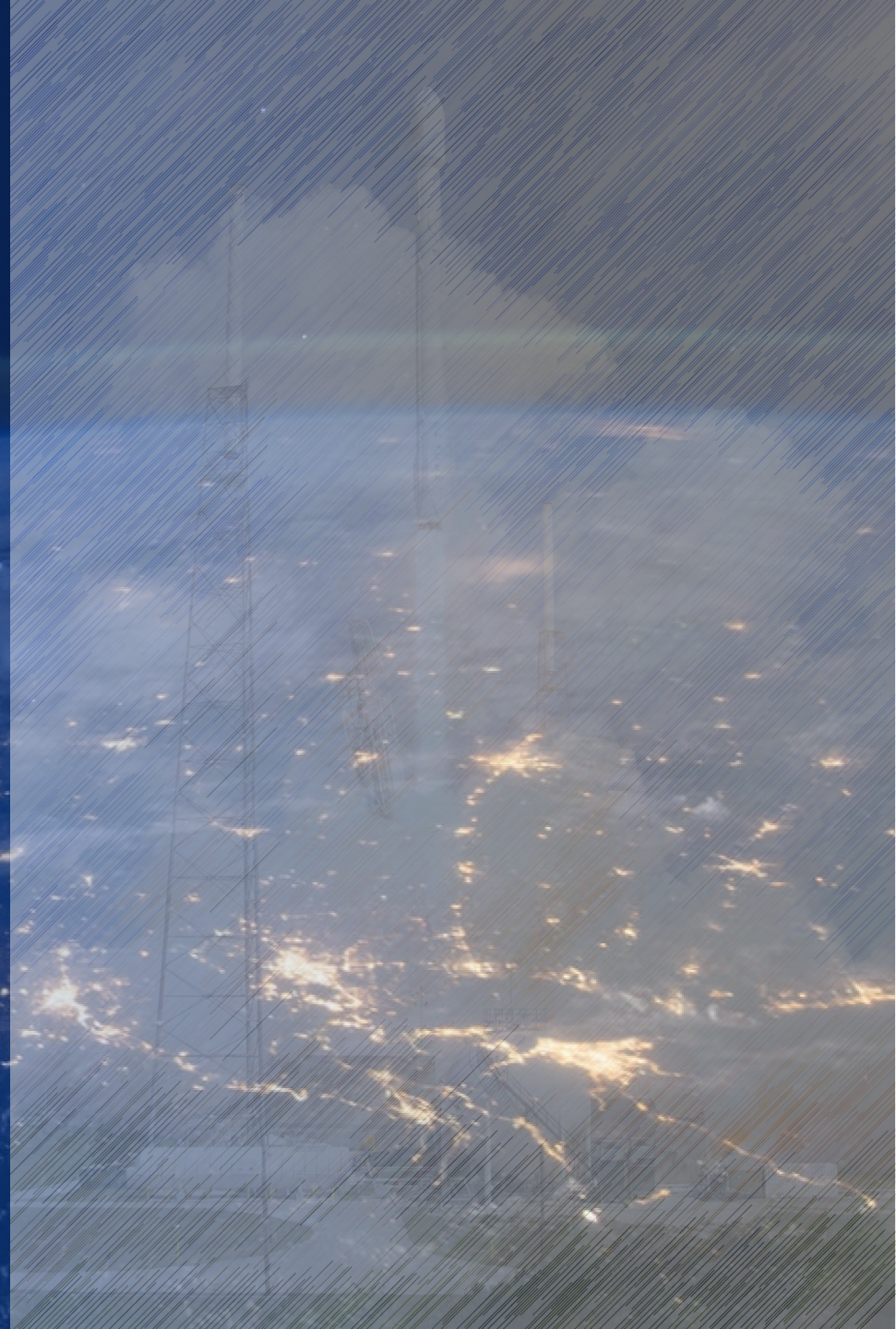
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

landing__outcome	number_of_occurences
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

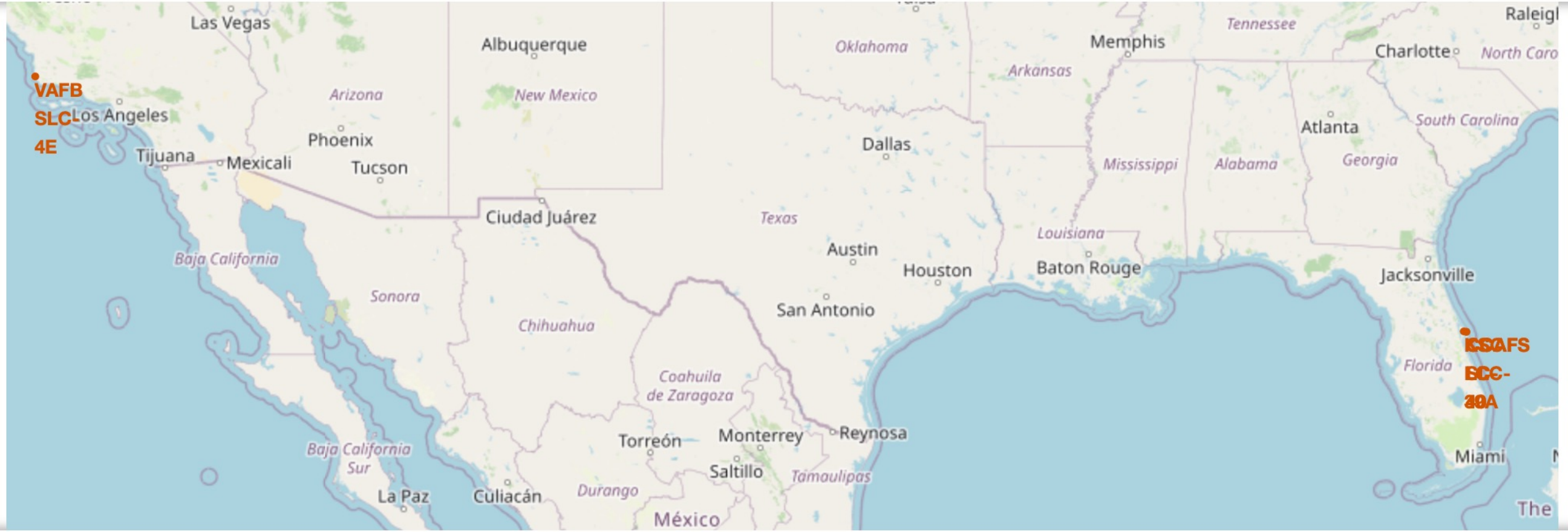
- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad) between the date 2010-06-04 and 2017-03-20
- Presented in descending order

Section 3

Launch Sites Proximities Analysis

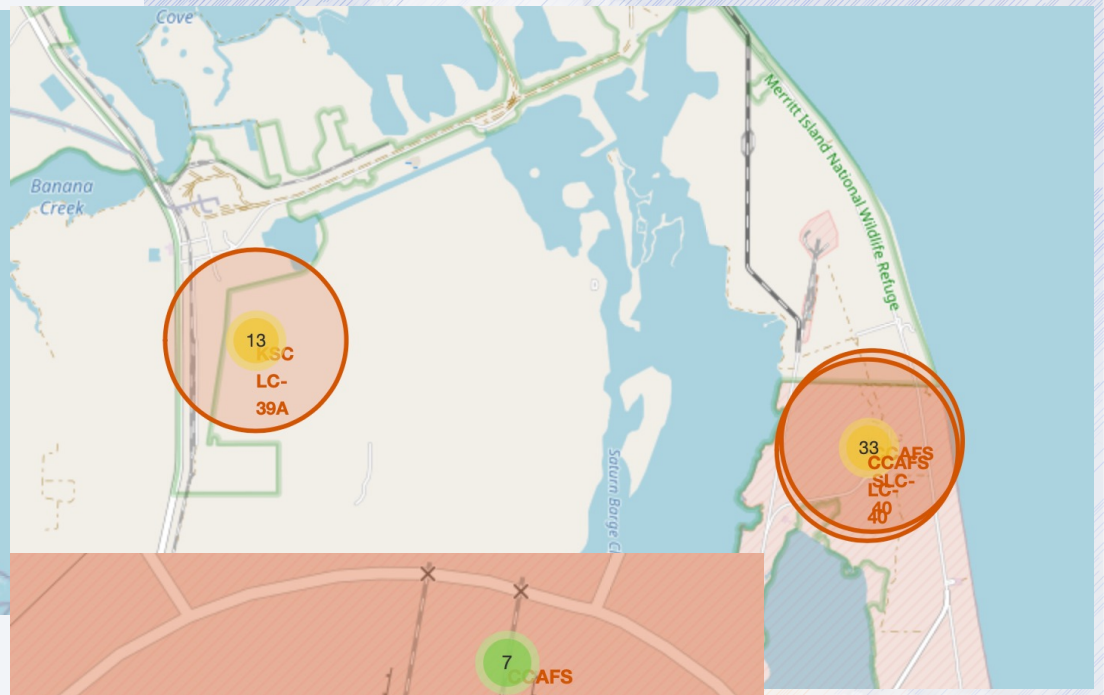


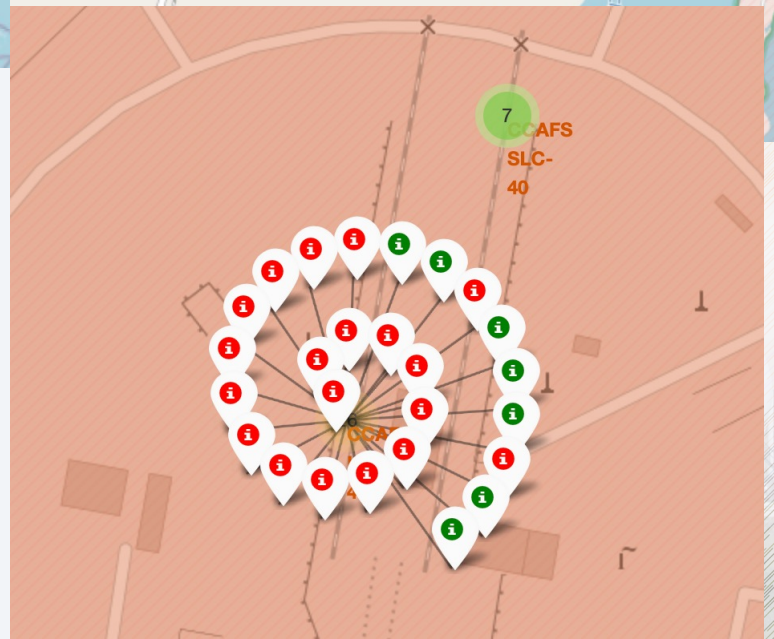
The locations of the sites



- One site (VAFB SLC 4E) was located on west coast
- Other sites were located on east coast

Success/failed launches for each site

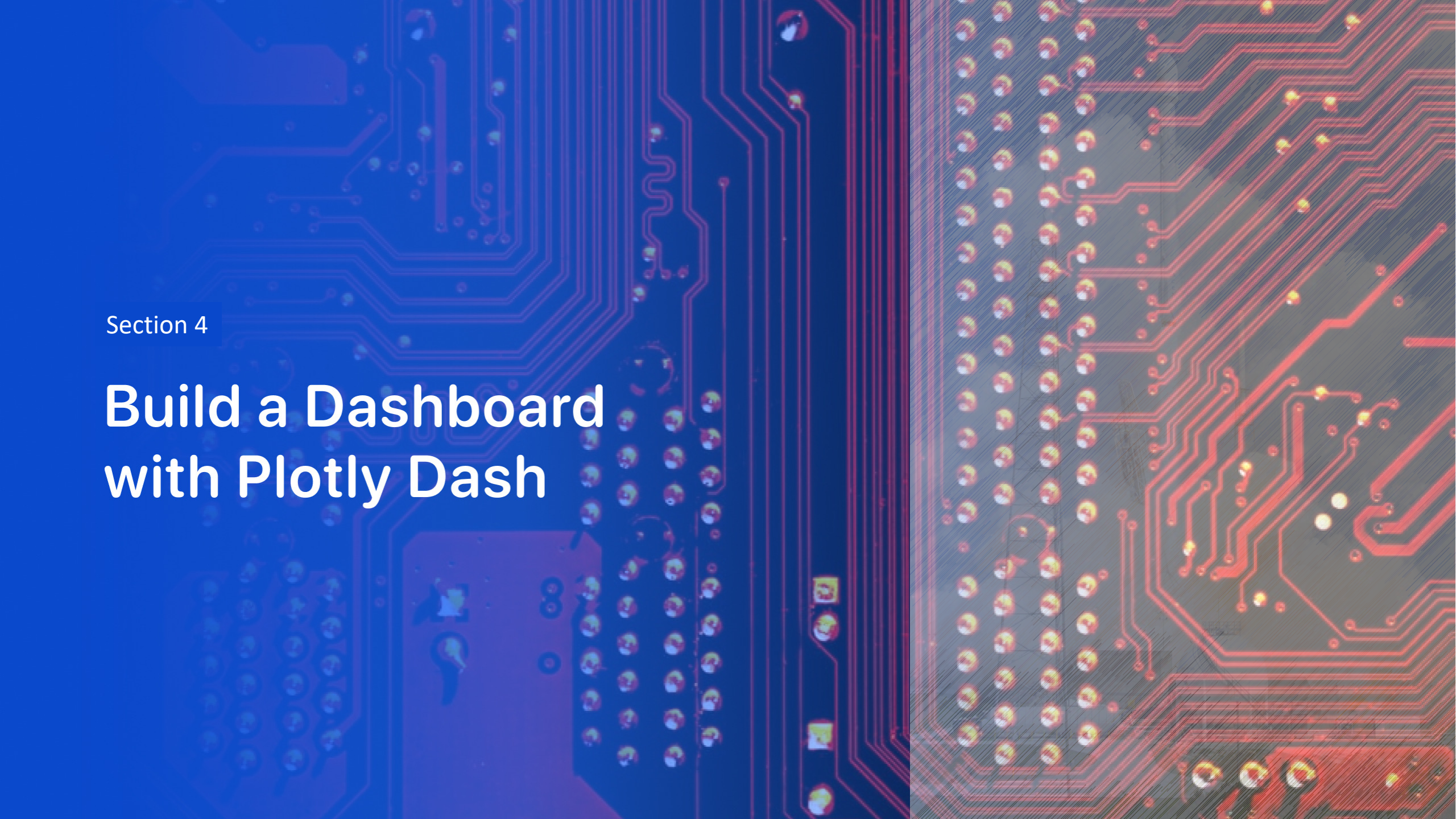




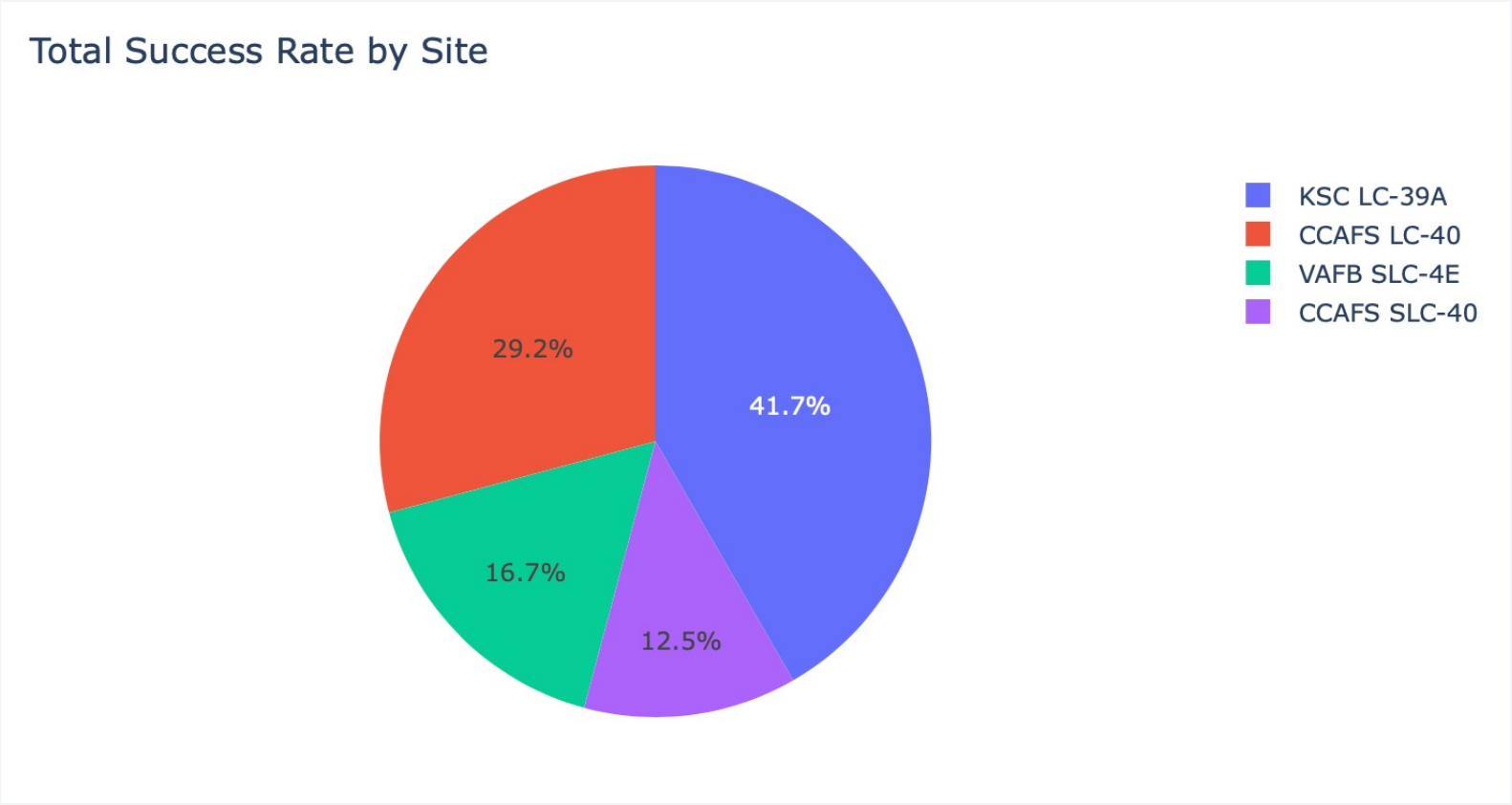
- Interactive map allowed for visualization of the success rates for each of the sites

Distance to proximities form launch sites

- Distance to proximities (i.e., coastline) was manually calculated for sites

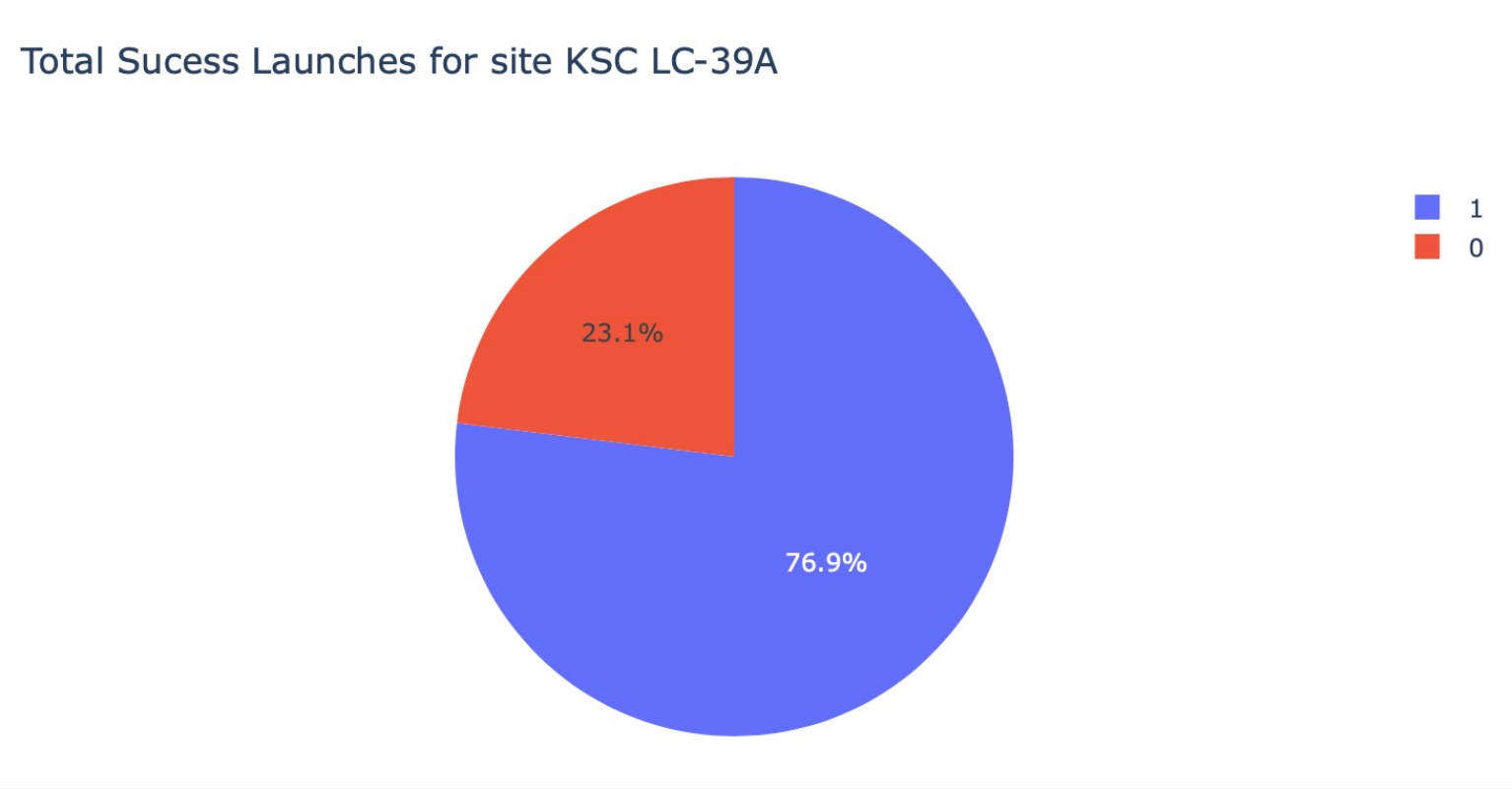

Section 4
Build a Dashboard with Plotly Dash

Launch success for all sites



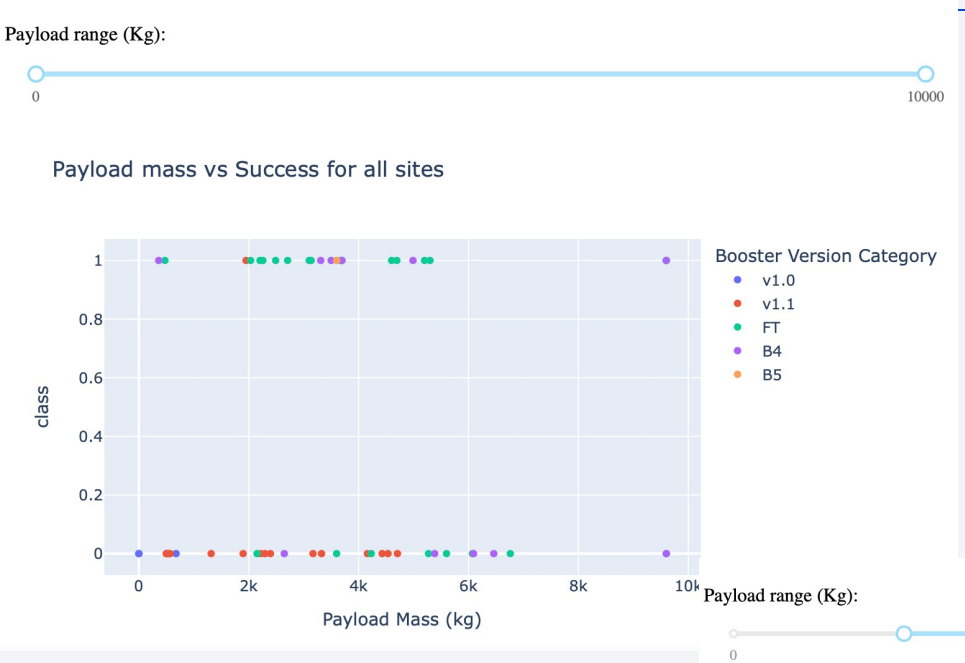
- The highest success rate was for the KSC LC-39A site
- The lowest success rate was for the CCAFS SLC-40

Success launches at the most successful site



- Number depicts success in launch
  (0/red: failure, 1/blue: success)
- For the most successful site the success rate was over 76%

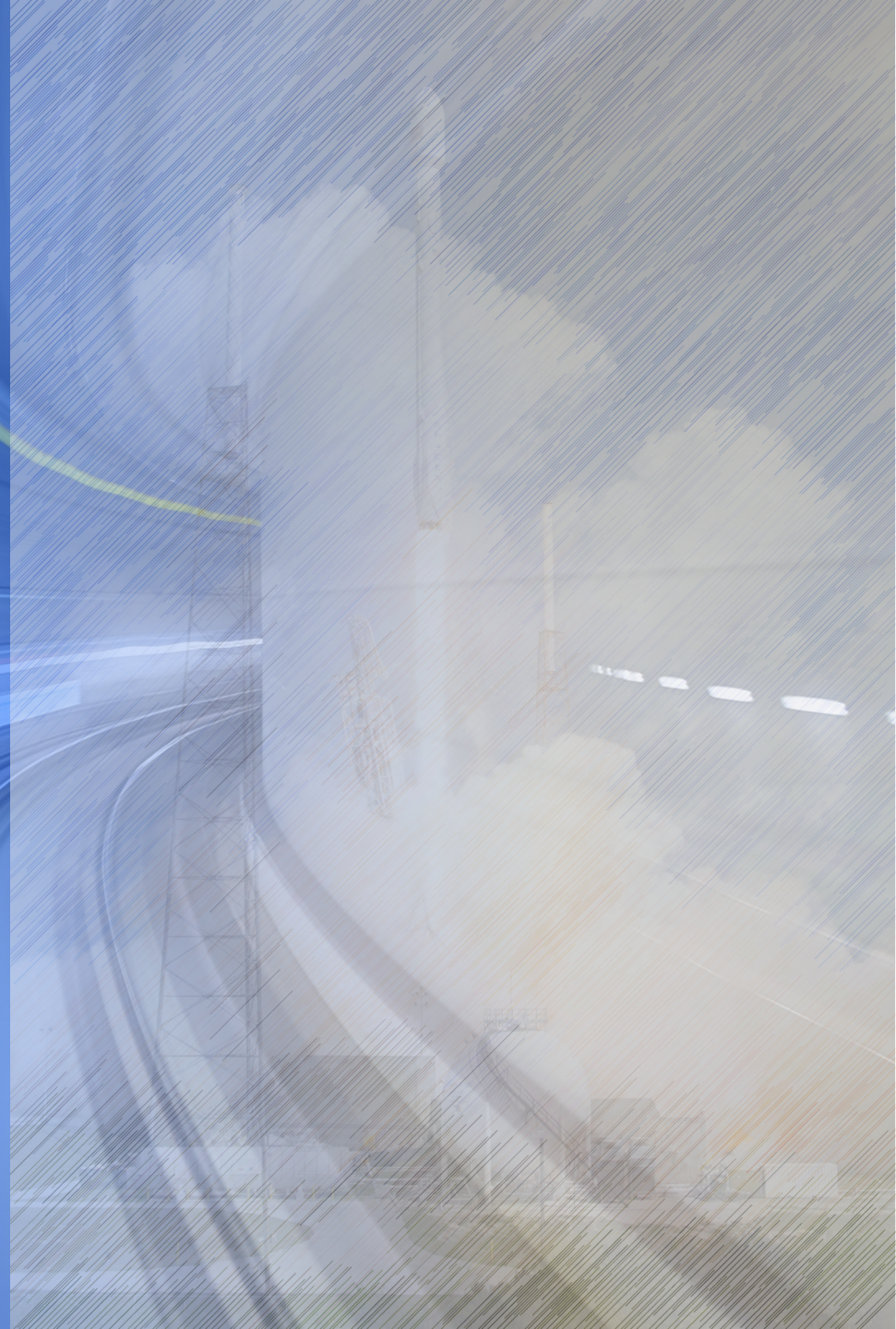
Payload mass and booster version vs success



- Class depicts success (0: failure, 1: success)
- Most of the successes came for lower payload mass (between 2000-6000kg)
- Most successful booster in that weight range was FT

Section 5

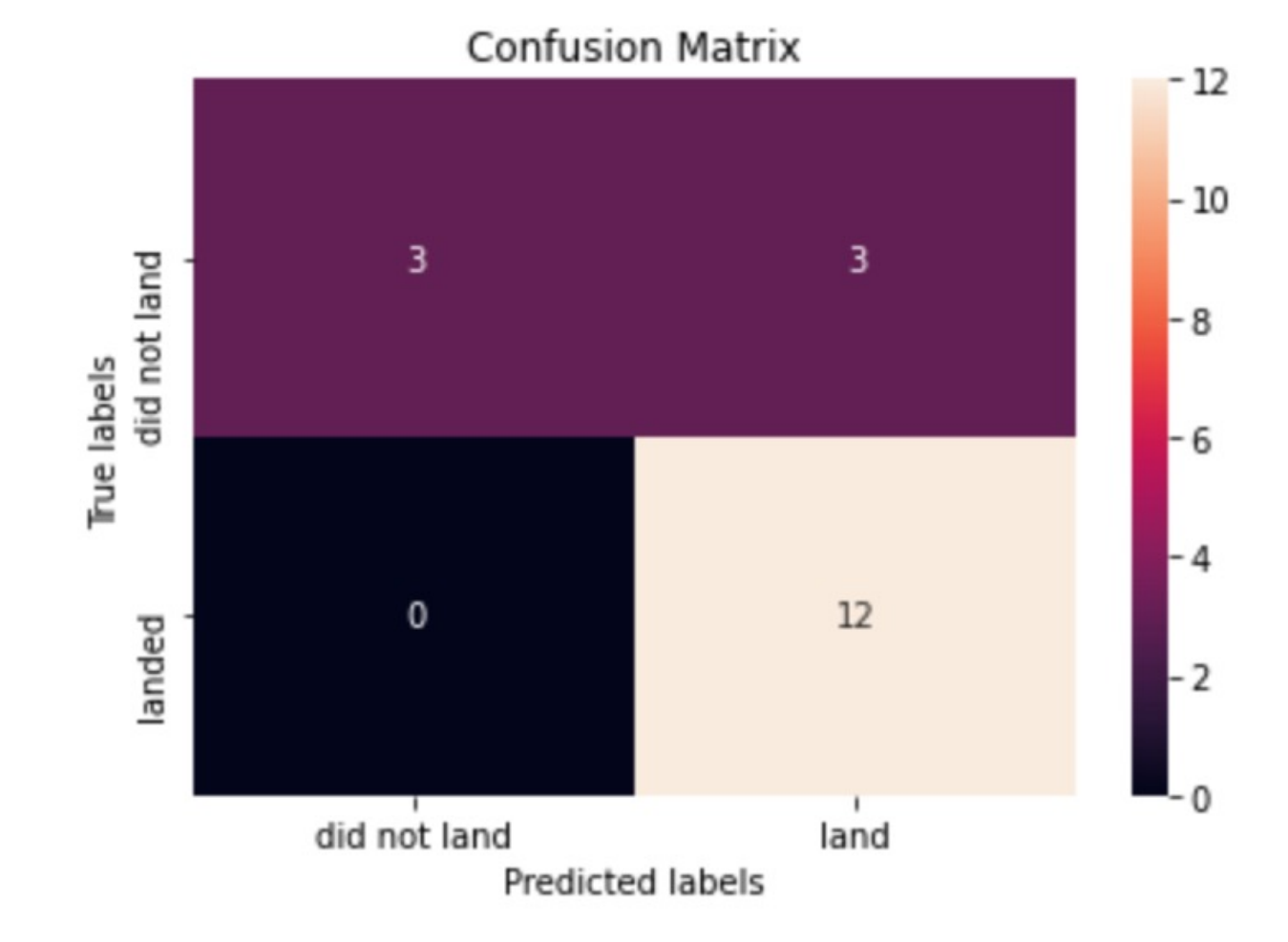
Predictive Analysis (Classification)



Classification Accuracy – models comparison

- The accuracy was tested with both train and test samples
- The performance on train data varied, yet ALL models performed identically on test data (acc=.833)

Confusion Matrix - SVM



- All models performed equally, and their confusion matrices were identical
- All models had three cases of false positives (the model predicted a success outcome, whereas the outcome was failure)

Conclusions

- VAFB-SLC launch site appears to have the best success rate, when accounting only for the site's attempts (probably due to smaller payload masses)
- Yet, the highest success rate (when accounted for all attempts) was for the KSC LC-39A site
- With heavy payloads the successful landing and positive landing rate are more probable for Polar, LEO and ISS orbits
- Most of the successes came for lower payload mass (between 2000-6000kg)
- Most successful booster in that weight range was FT
- Based on the data, an accurate model predicting the outcome was built with and accuracy of .833

Thank you!

Appendix – SQL queries

Task 1

Display the names of the unique launch sites in the space mission

```
%sql SELECT unique(launch_site) FROM SPACEXDATASET
```

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXDATASET WHERE launch_site LIKE '%CCA%' LIMIT 5
```

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(payload_mass__kg_) FROM SPACEXDATASET WHERE customer LIKE '%NASA (CRS)%'
```

Task 4 ¶

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(payload_mass__kg_) FROM SPACEXDATASET WHERE booster_version LIKE '%F9 v1.1%'
```

Appendix – SQL queries

Task 5

List the date when the first successful landing outcome in ground pad was acheived. ¶

Hint:Use min function

```
%sql SELECT MIN(DATE) FROM SPACEXDATASET WHERE landing__outcome LIKE '%Success%' AND landing__outcome LIKE '%ground pad%'
```

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT booster_version FROM SPACEXDATASET WHERE (landing__outcome LIKE '%Success%' AND landing__outcome LIKE '%drone ship%'
```

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome, COUNT(*) AS sum FROM SPACEXDATASET GROUP BY mission_outcome
```

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT booster_version FROM SPACEXDATASET WHERE payload_mass__kg_=(SELECT max(payload_mass__kg_) FROM SPACEXDATASET)
```

Appendix – SQL queries

Task 9

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

```
%sql SELECT booster_version, launch_site, landing__outcome FROM SPACEXDATASET WHERE (landing__outcome LIKE '%Failure%' AND landi
```

Task 10

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

```
: %sql SELECT landing__outcome, COUNT(*) AS number_Of_Occurences FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-2
```