IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Oskar Graeb
19.07.2022

Outline

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

Executive Summary

- Summary of methodologies:

- Data collected using SpaceX API
- Which later was cleaned and edited using mostly pandas library.
- EDA process done using tools like numpy, pandas and matplotlib, seaborn, folium, dash for visualization to extract important information.
- Using different machine learning algorithms, it was possible to build a model with good accuracy that allowed for prediction of mission outcome.

- Summary of results

Using data from SpaceX API it was possible to find out and predict which rocket flights will be successful with landing first part of the rocket back on earth with approximately 90% accuracy, thus allowing to predict which launch will allow to reuse first stage.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- In this presentation we dive deep into the data to find out which parameters are most important to predict possibility of reusing first stage of a rocket.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX REST API and also web scraped from Wikipedia using python module called BeautifulSoup.
- Perform data wrangling
  - Describe how data was processed
- 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

- Data collection methodology:

Data was collected using SpaceX REST API and also web scraped from Wikipedia using module called BeautifulSoup

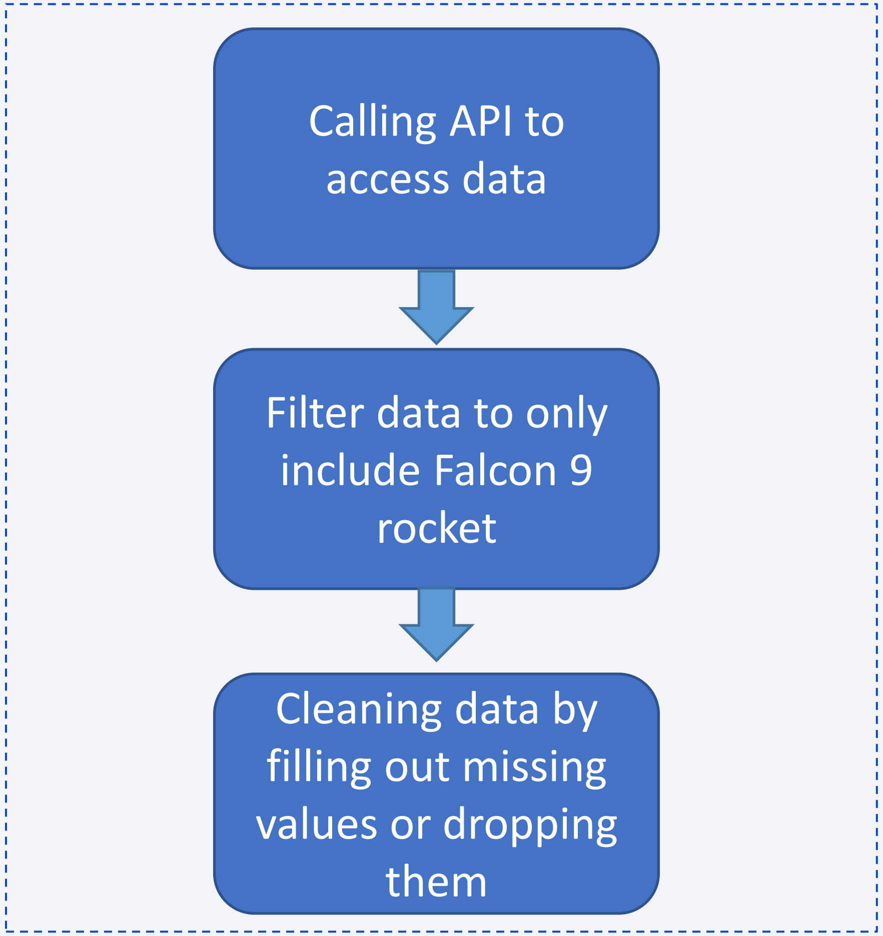
- Process:

1. Get request from SpaceX different API’s, which later was merged using pandas
2. Webscraping from Wikipedia page to extract additional information
3. Data wrangling to process it into more friendly form
4. Exploring the data and visualizing connections (EDA process)

Data Collection – SpaceX API

- SpaceX offers free API that allows to get their own rocket flight data. To access it we used requests module.

- Link to relevant notebook:
- Data-collection



Calling API to access data

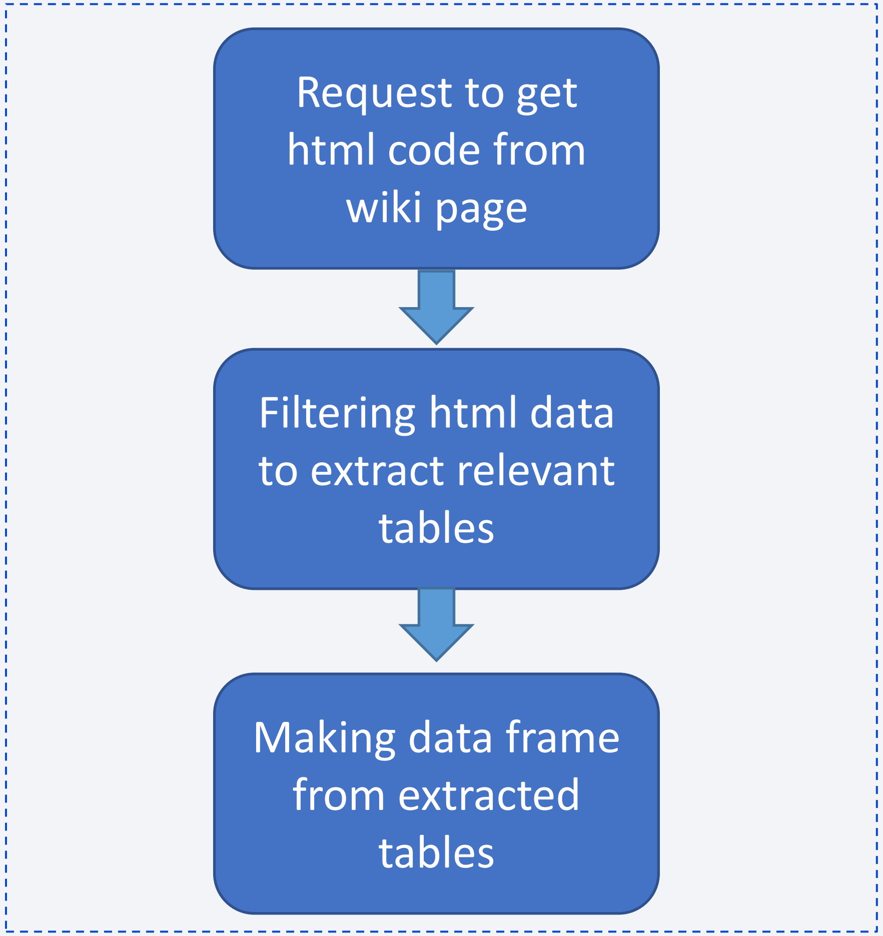
Filter data to only include Falcon 9 rocket

Cleaning data by filling out missing values or dropping them

Data Collection - Scraping

- Data was parsed from Wikipedia page, using get request command. Link: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- Link to relevant notebook:
- Webscraping



Request to get html code from wiki page

Filtering html data to extract relevant tables

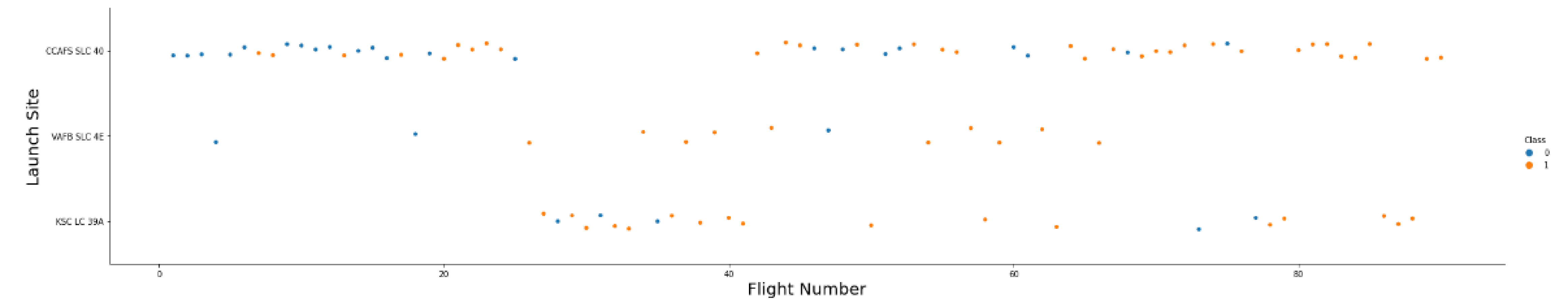
Making data frame from extracted tables

Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. At this stage we added new column named "Class" to our data that defined if landing was success or failure which was represented by 1 for success and 0 for failure.
- Link to relevant notebook:
- Data Wrangling

EDA with Data Visualization

- To explore data and find useful features, bar plots and scatter plots were used. For example, chart below shows correlation between number of flights and launch site



- Link to relevant notebook:
- [Data Visualization]()

EDA with SQL

- **Summary of performed SQL queries.**

- Showed names of the unique launch sites in the space missions
- Displayed 5 records where launch sites begin with the string 'KSC'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the succesful landing outcome in drone ship was achieved
- List of names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Total number of successful and failed missions
- Using subquery listed names of the booster versions which have carried the maximum payload mass
- Shown records which displayed the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Ranked the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

- **Link to relevant notebook:**
- EDA SQL

Build an Interactive Map with Folium

- **Folium is python package that allows to work with geospatial data and maps much more easily. To visualize data some folium objects were used.**
- Markers to indicate points e.g. geographical position of launch sites
- Circles to highlight areas around specific points, like NASA Johnson Space Center
- Lines were used to show distance between two coordinates like launch site and highways or coast line.
- Marker clusters allowed to group events in coordinate for example separate launches on launch site
- **Link to relevant notebook:**
- Geolocation with Folium

Build a Dashboard with Plotly Dash

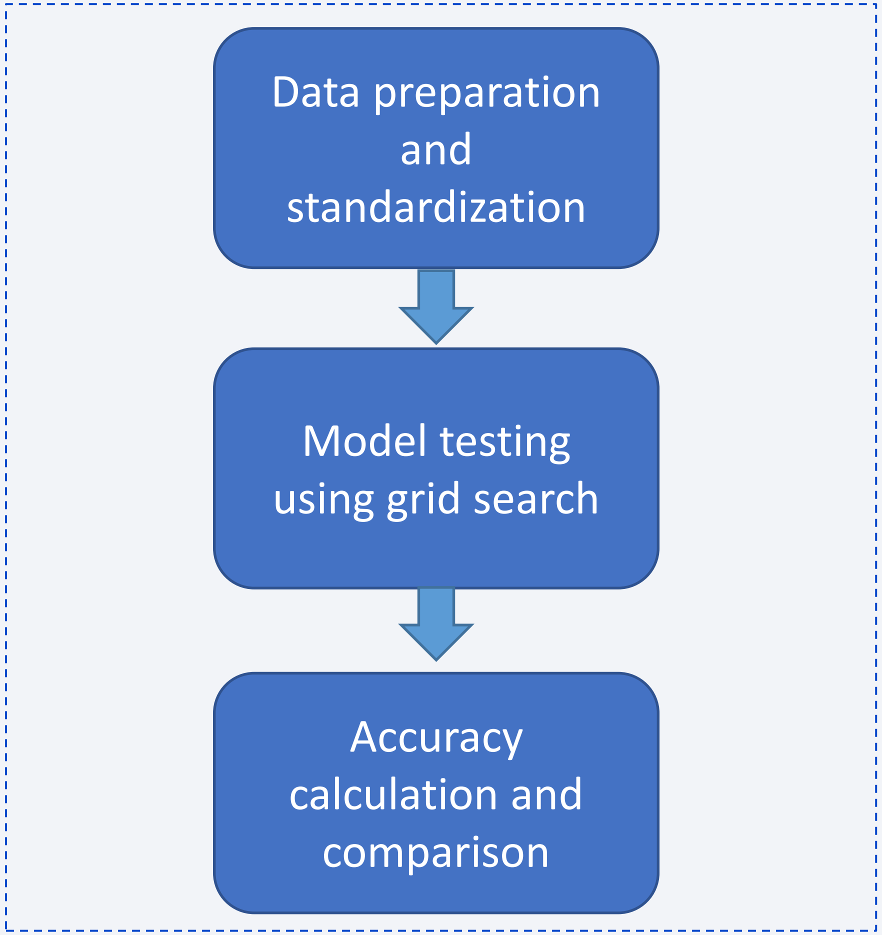
- **To better understand data interactive plots/graphs were added, for example:**
- Dropdown menu with option to select specific launch sites, or to show all.
- Payload range slider to better inspect specific data.
- Graph showing comparison between different booster versions and relation between payload mass and outcome of mission.

- **Link to relevant notebook:**
- Dashboard using plotly dash

Predictive Analysis (Classification)

- To predict outcome of mission with highest accuracy 4 different machine learning models were build and tested on different parameters. Models that were used in comparison were: logistic regression, support vector machine, decision tree and k nearest neighbors.

- **Link to relevant notebook:**
- ML models


Data preparation and standardization

Model testing using grid search

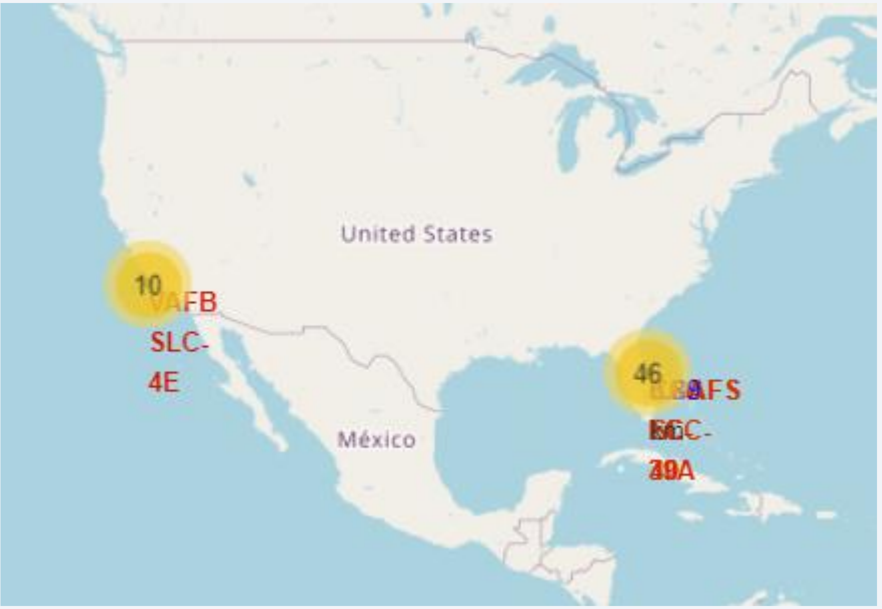
Accuracy calculation and comparison

Results

- Exploratory data analysis results:
- SpaceX uses 4 different launch sites
- First landings were on SpaceX and NASA ground
- Average payload of F9 v1.1 booster was 2928,4 kg
- First successful landing on drone ship happened in 06/05/2016
- Boosters that succeeded to land with payload between 4000 and 6000 kg were F9 FT B1032.1,F9 B4 B1040.1, F9 B4 B1043.1
- Only 12 out of 79 landing attempts ended with failure
- More time passes the higher are chances of successful landing except for year 2020 where that percentage dropped slightly

Results

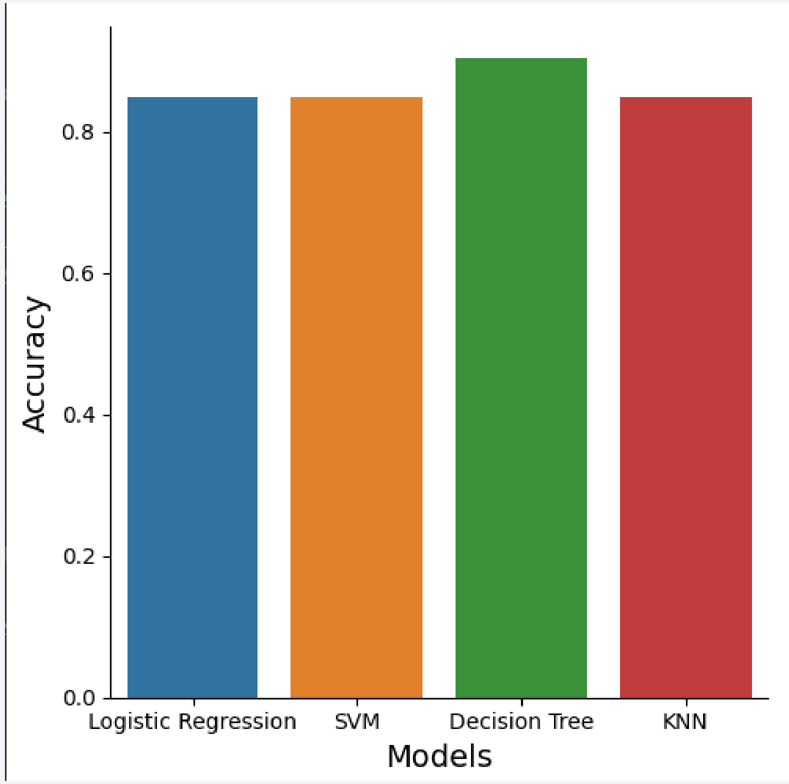
- Using interactive analytics like folium it was possible to identify launch sites and their proximity to different infrastructures around them. What was found is that most launches happened at east coast.

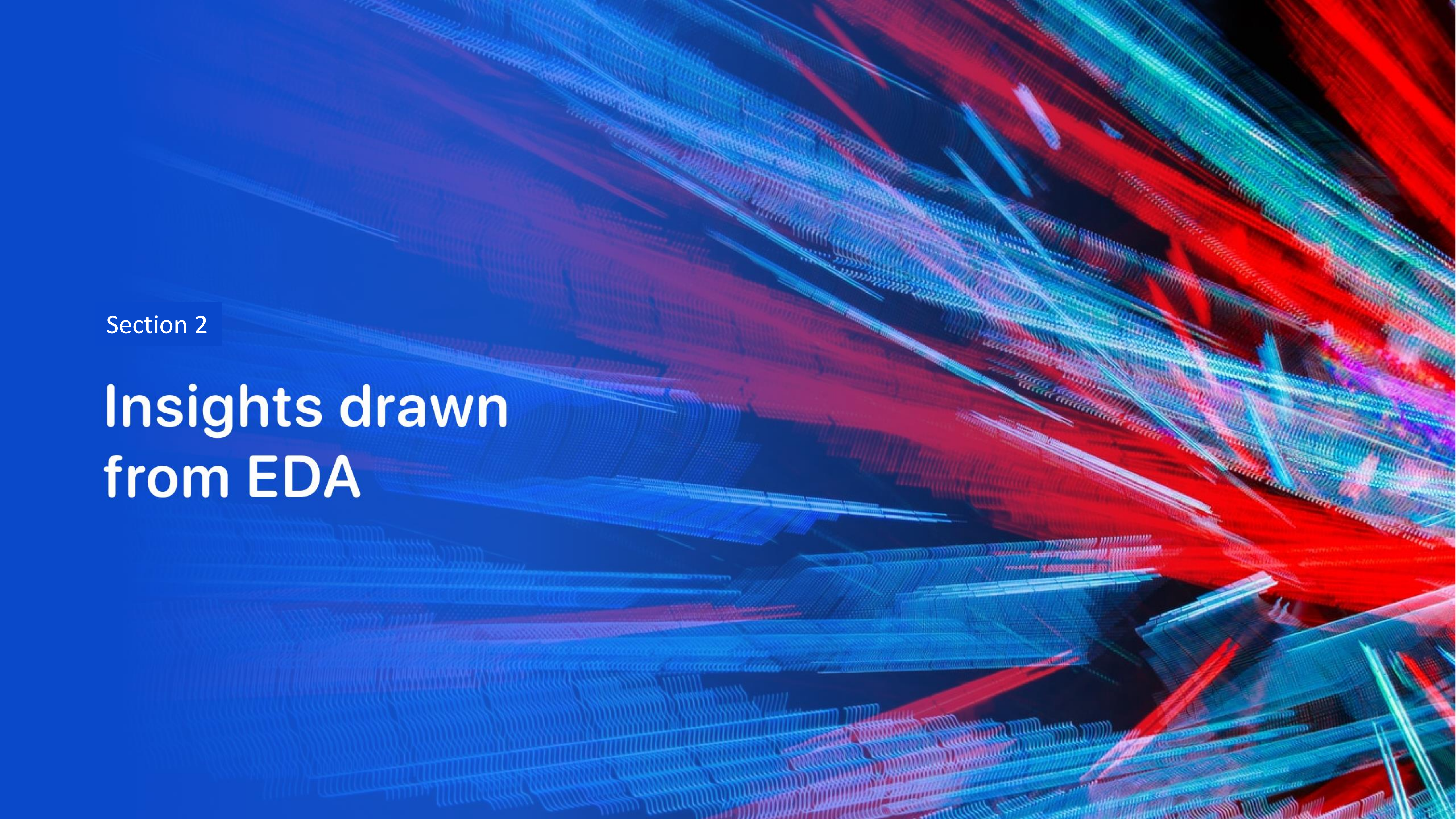


13
KSC
LC-
39A
33
CCAFS
SLC-
40
0.89
km
Wildlife Refuge
Barge Channel

Results

- Predictive analysis showed that decision tree classifier was most promising and scored of little bit over 90% accuracy on test data set!

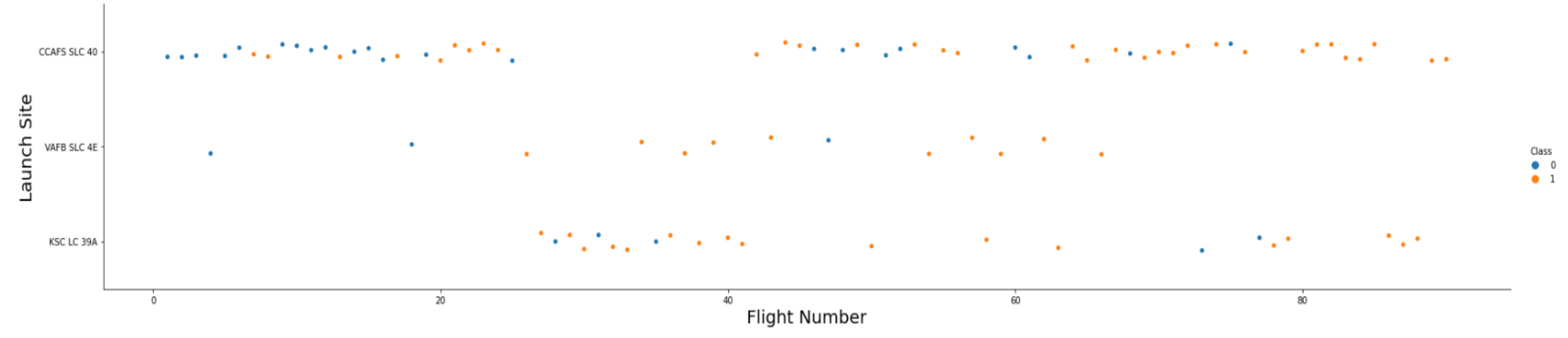




Section 2

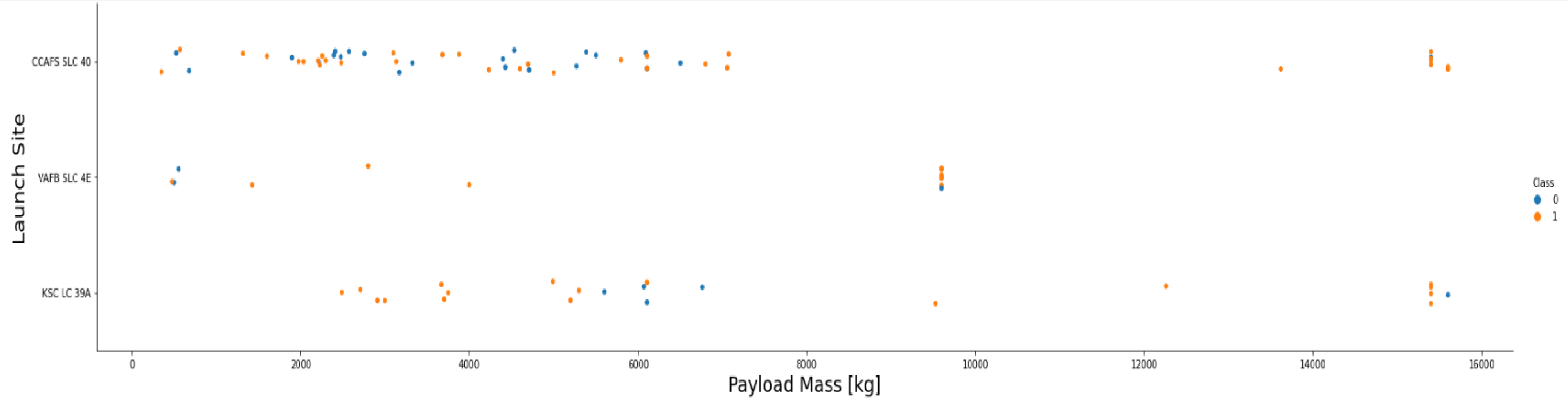
Insights drawn from EDA

Flight Number vs. Launch Site



- Analyzing the plot above we can deduct that the overall best launch site is CCAF5 SLC 40, most of it recent launches ended up being successful.
- Overall success rate improved for every launch site.

Payload vs. Launch Site

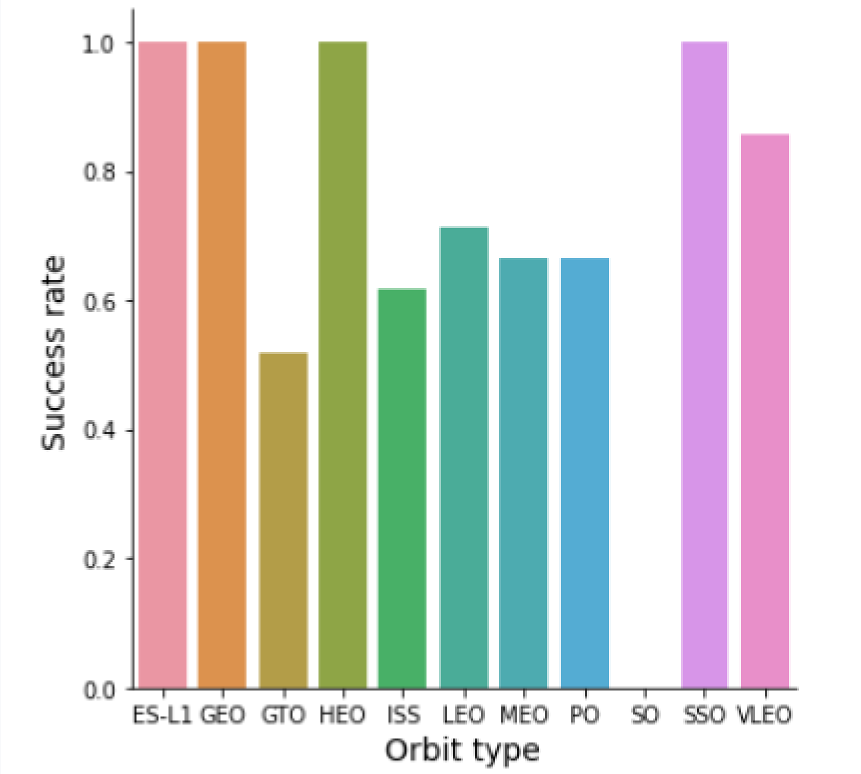


- Heavier payloads that means above 9,000 kg have great success rate, this can be the case because first launches were with lower payload, which only increased when success rate was satisfactory.
- Payloads heavier than 12,000 kg seems to be only possible on 2 launch sites, KSC LC 39A and CCAFS SLC 40.

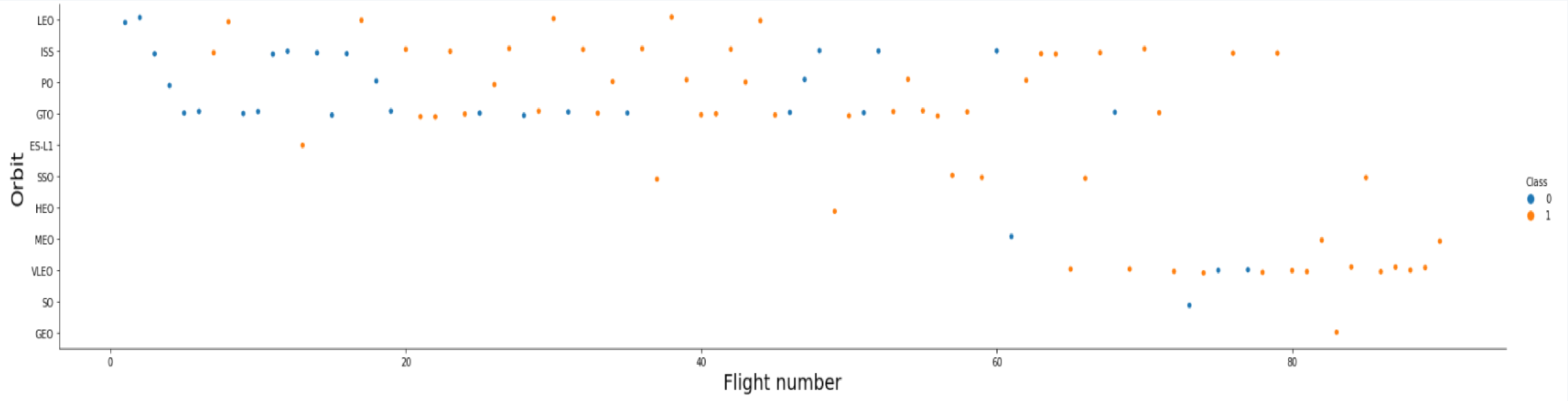
Success Rate vs. Orbit Type

- Orbits with highest success rate are:
- ES-L1;
- GEO;
- HEO
- SSO

- While the lowest success rate goes to SO with 0% success rate, and GTO with only 50%

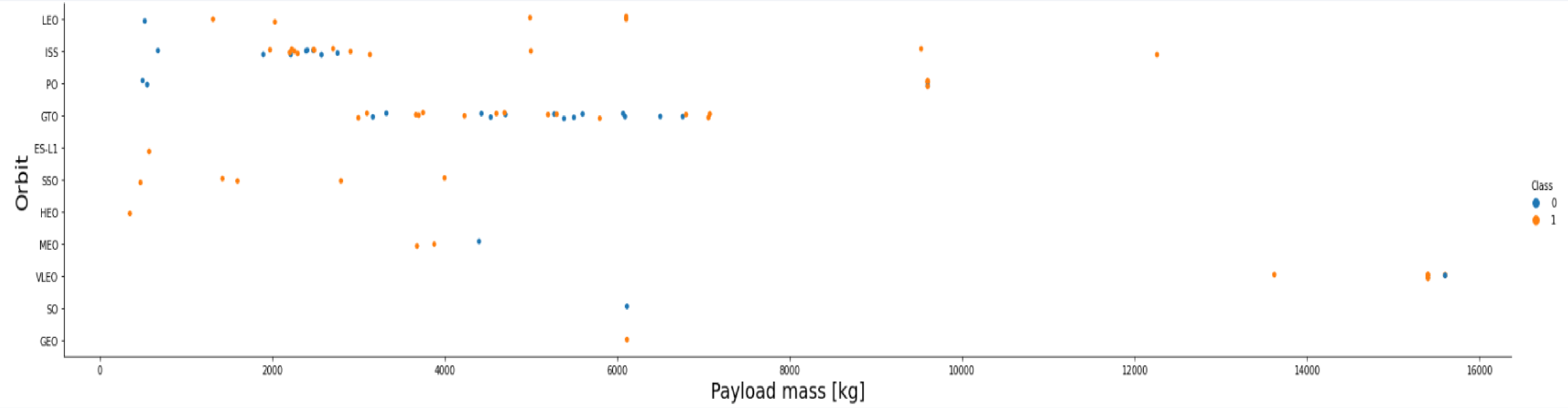


Flight Number vs. Orbit Type



- Over time success rate improved for all orbits
- Recently VLEO flights increased in popularity

Payload vs. Orbit Type



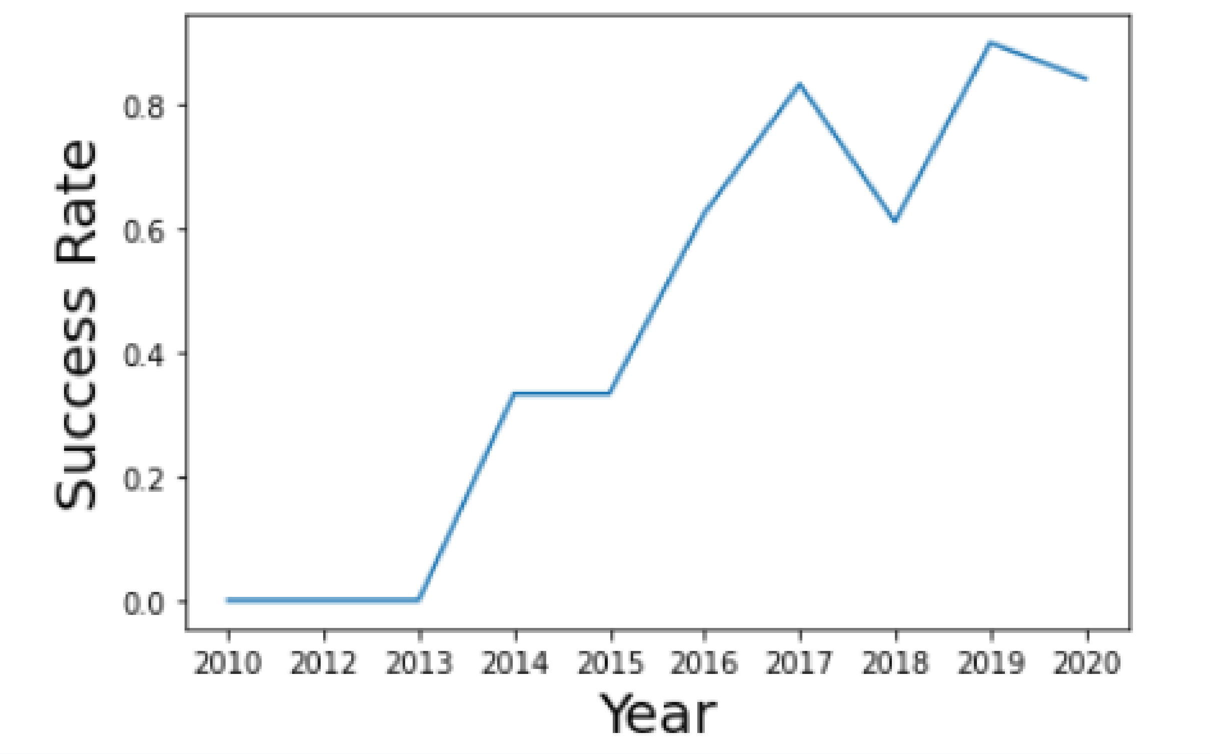
- There is no strong relation between payload and success rate for flights to orbit GTO

- ISS orbit flights are most diversified in terms of payload and have good success rate

- SO and GEO orbits are one of the least popular

Launch Success Yearly Trend

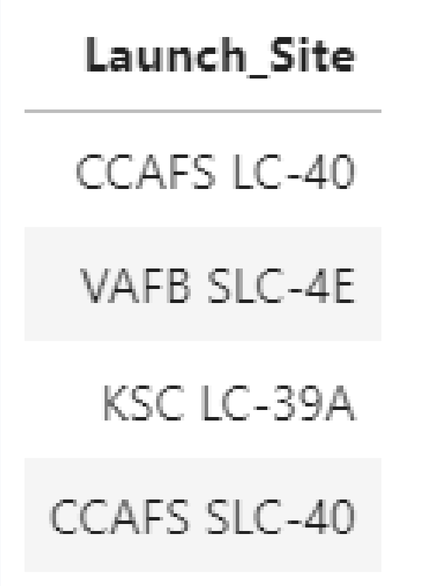
- First three years were mostly focused on improving technology and there was no successful landing yet.
- Succes rate was increasing very fast until 2017 when it plunged down and came back up in 2019 to drop a little bit again in 2020.



All Launch Site Names

- According to data there are only 4 launch sites:

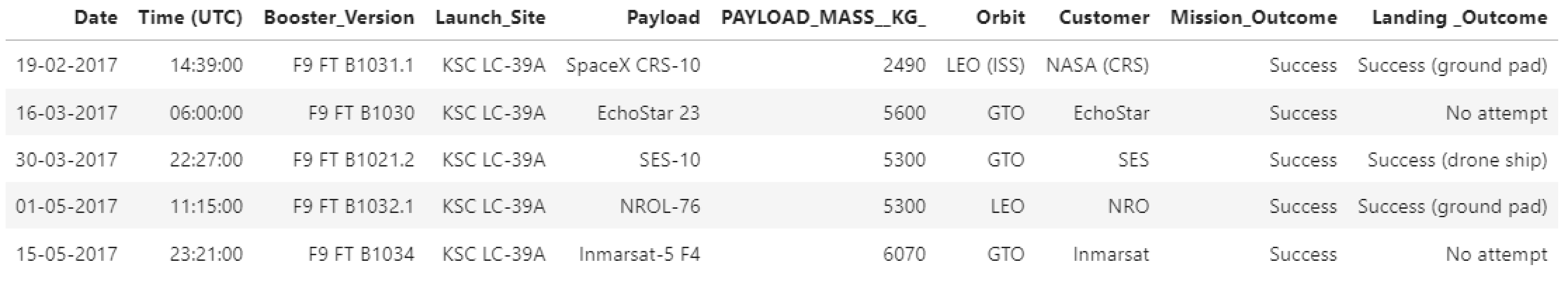
- They were obtained by parsing distinct values of Launch_Site column from dataset.



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

Launch Site Names Begin with 'KSC'

- 5 records where launch sites’ names start with `KSC`

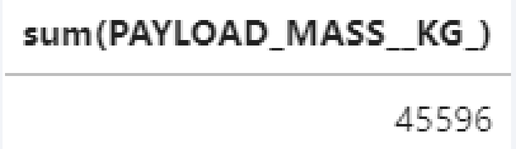


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16-03-2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

- Query shows 5 samples of KSC launches

Total Payload Mass

- Total overall payload that was carried by all boosters from 2013 to 2020 is:

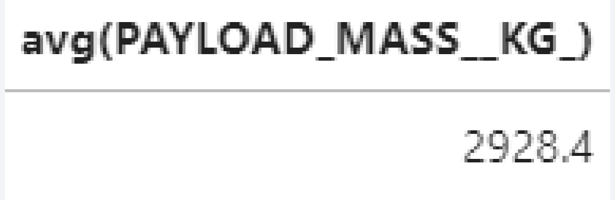


sum(PAYLOAD_MASS__KG_)
45596

- It was calculated by adding up every payload mass from dataset.

Average Payload Mass by F9 v1.1

- Average payload mass carried by F9 v1.1 booster was:

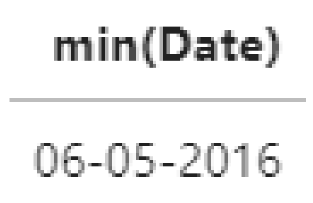


avg(PAYLOAD_MASS__KG_)
2928.4

- This number was calculated by filtering out any occurrence where booster version wasn’t equal to F9 v1.1 and then summing up all payload mass values.

First Successful drone ship landing Date

- First successful landing on drone ship happened at:

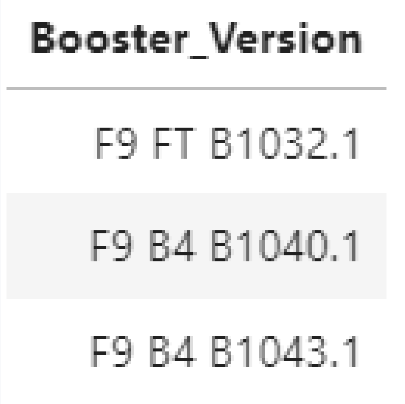


min(Date)
06-05-2016

- Date was calculated by querying dataset for minimum date where location was equal drone ship and class was equal to successful (in our case 1).

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

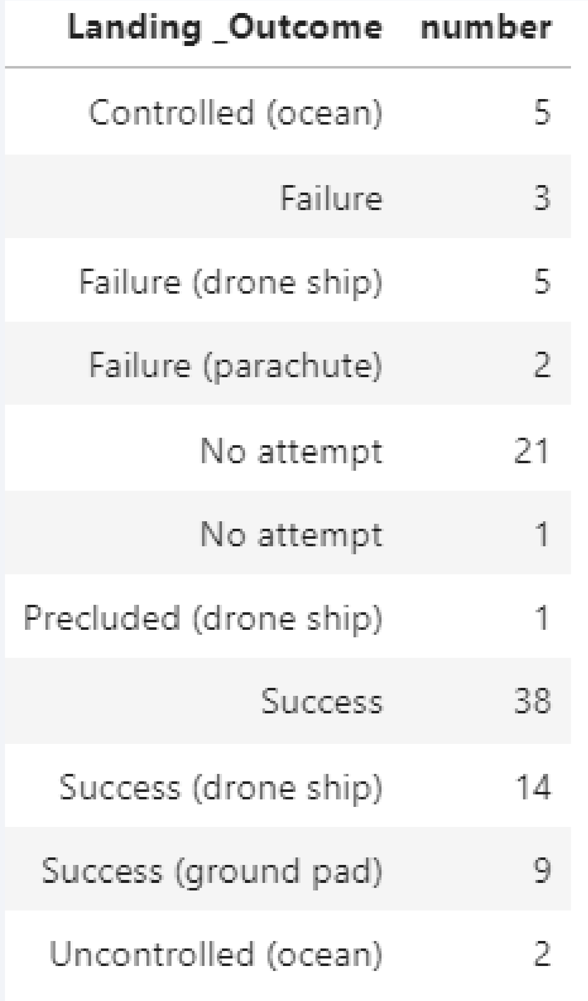


Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

- Selecting distinct booster versions where payload had to been between 4000 and 6000 kg we've got results shown above.

Total Number of Successful and Failure Mission Outcomes

- Querying data for distinct landing outcomes, we received few values. Ratio of failures to success was 12 to 79, and there was 22 launches with no landing attempts.

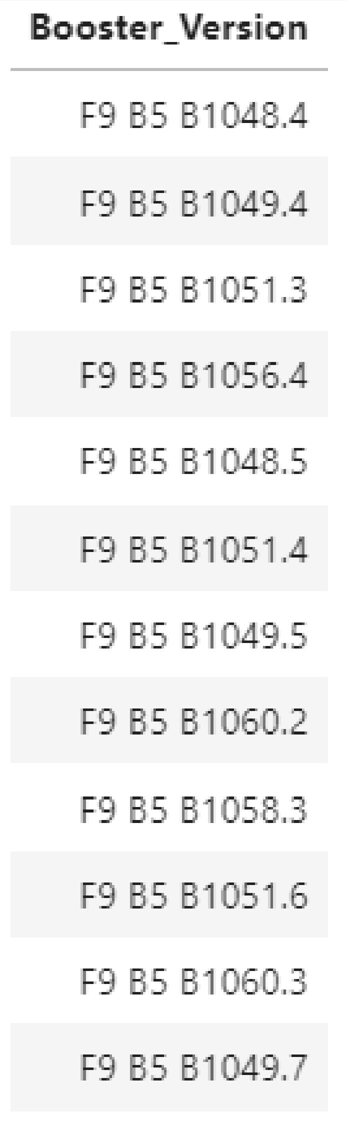


Landing _Outcome	number
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

- Boosters that carried maximum payload:

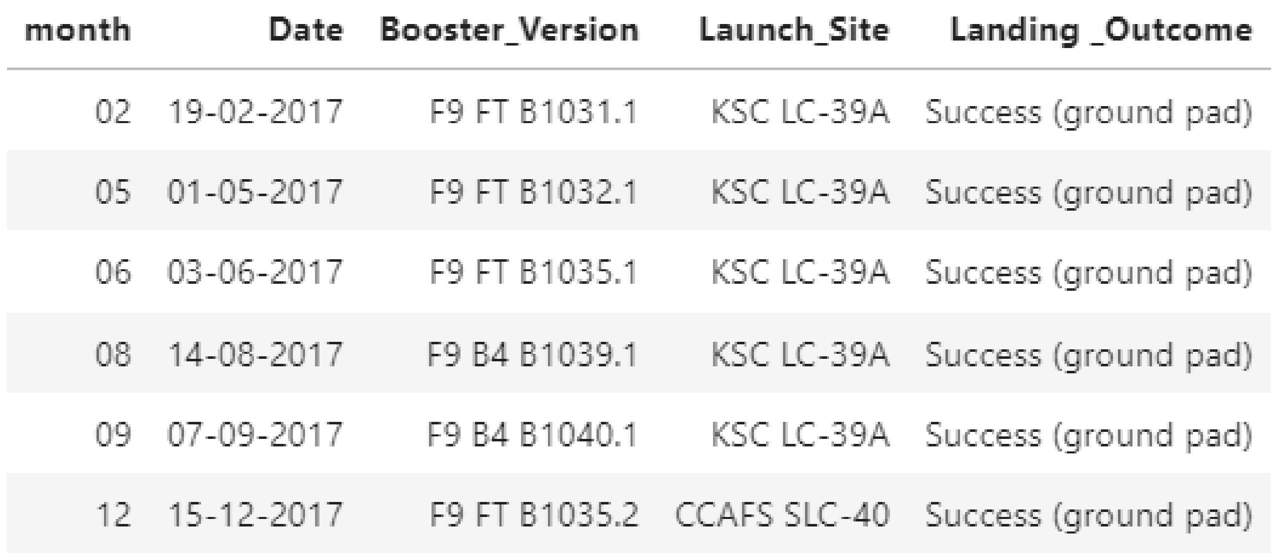
- This list was acquired by finding maximum payload value and then using it to find distinct booster versions.



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

2017 Launch Records

- List of launch records with month value, exact date, booster version, launch site and landing outcome in year 2017

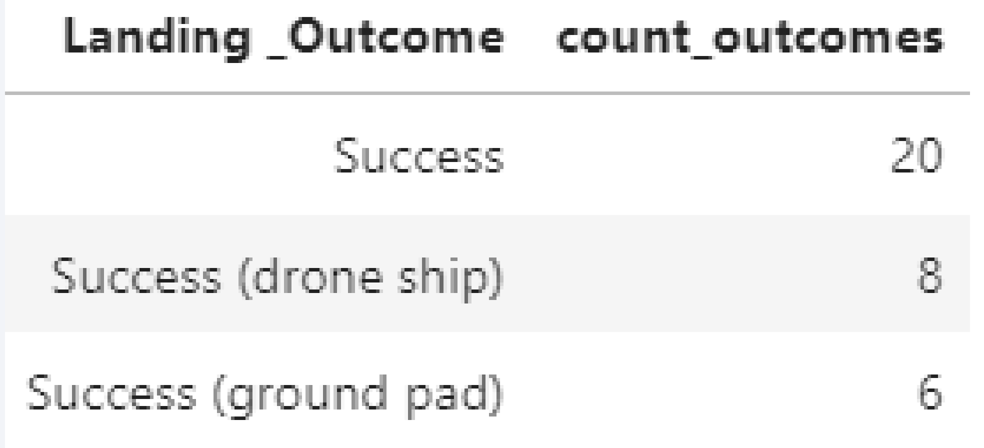


month	Date	Booster_Version	Launch_Site	Landing _Outcome
02	19-02-2017	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
05	01-05-2017	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
06	03-06-2017	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
08	14-08-2017	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
09	07-09-2017	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
12	15-12-2017	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

- We can see that there were 6 different launches, closest between 2 launches was 1 month and 2 days.

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

- Rank of summarized successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order



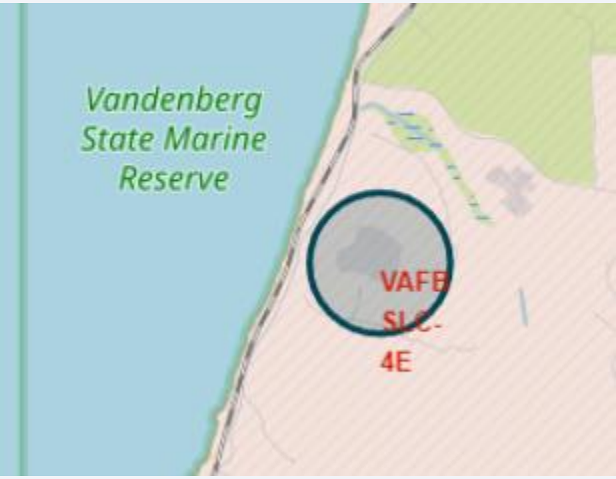
Landing _Outcome	count_outcomes
Success	20
Success (drone ship)	8
Success (ground pad)	6

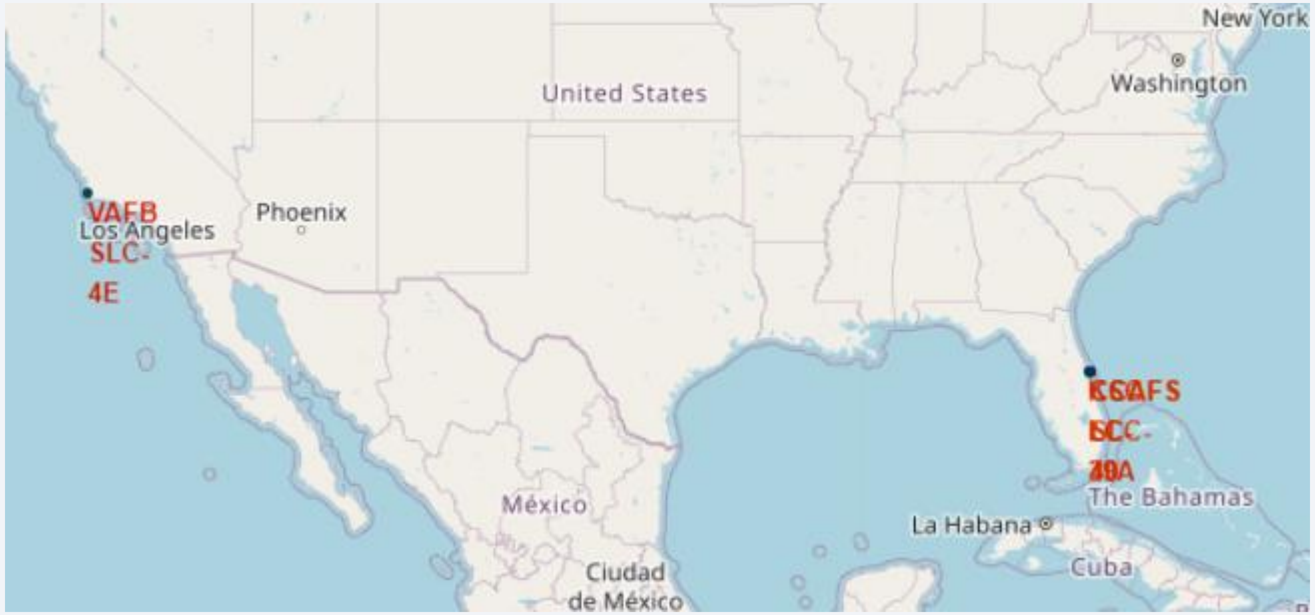
- In these 7 years 34 successful landings were achieved.

Section 3

Launch Sites Proximities Analysis

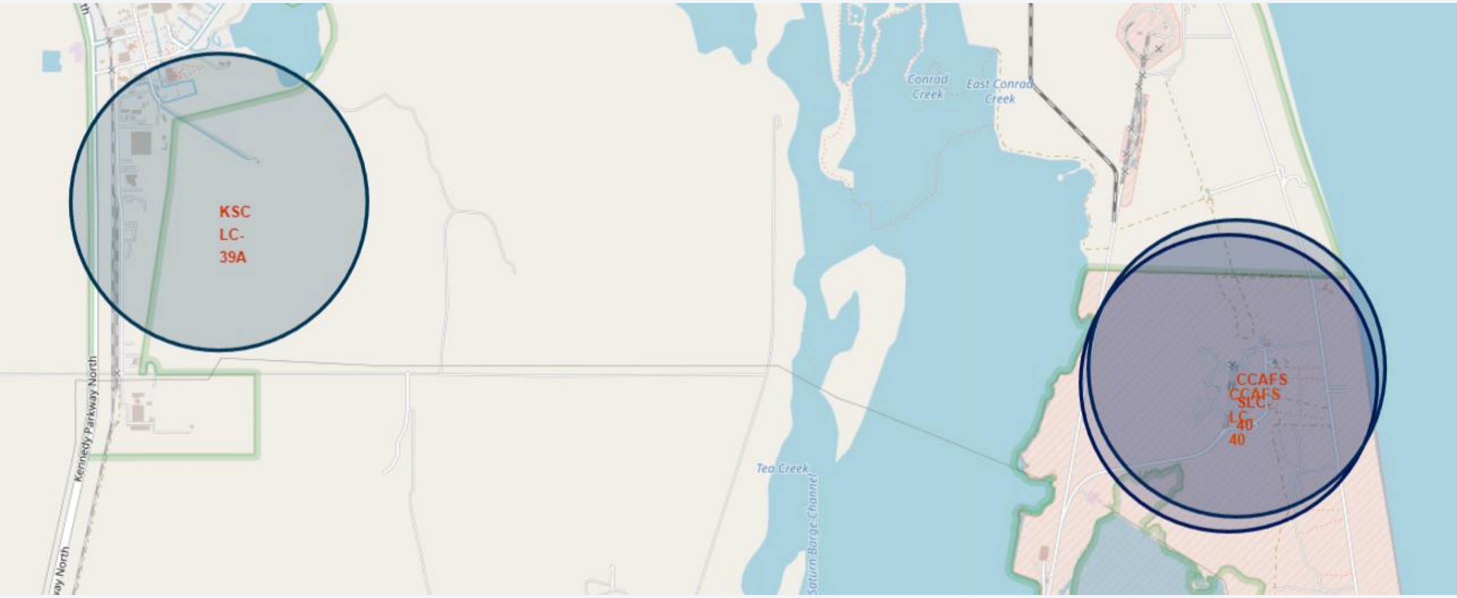
All launch sites

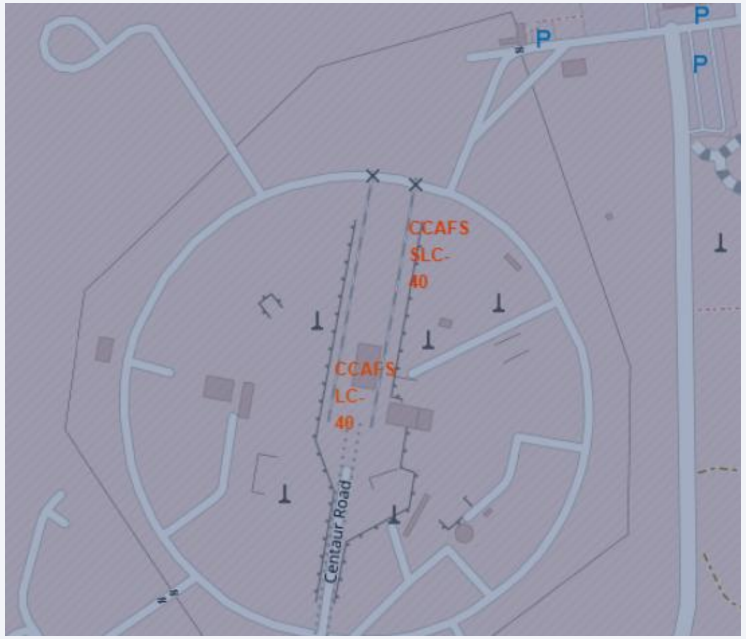




- We notice that all launch sites, are near oceans, but are relatively close to some cities.

Launch sites up close

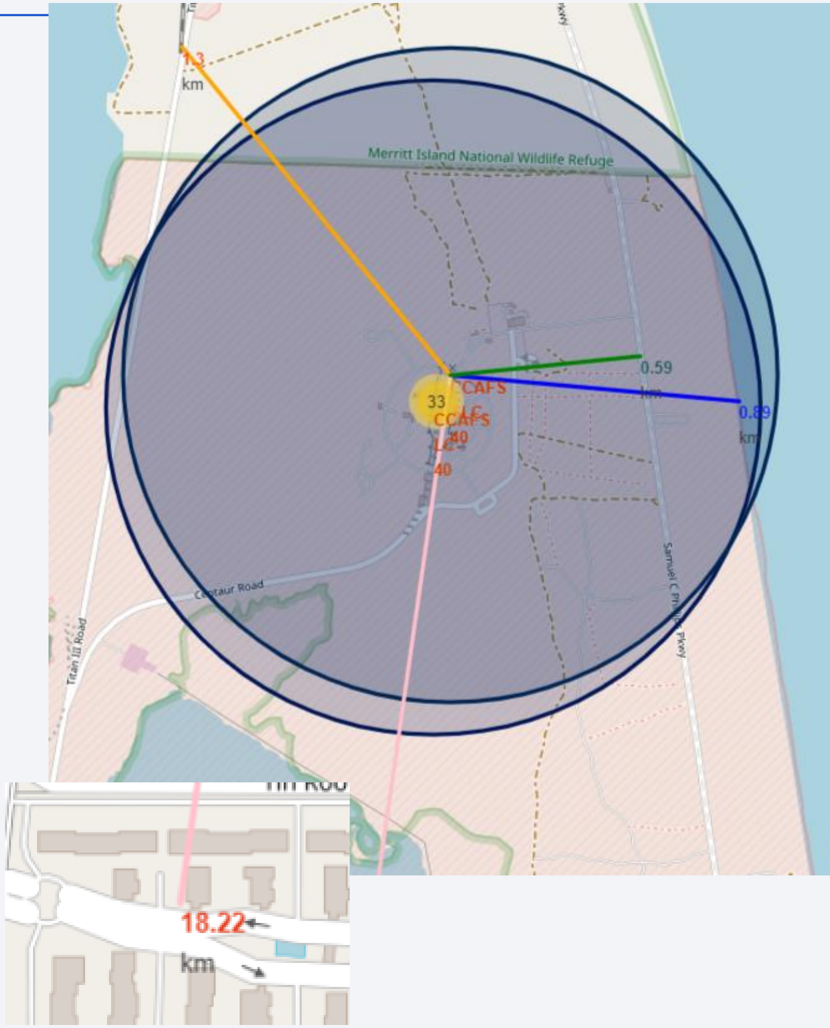


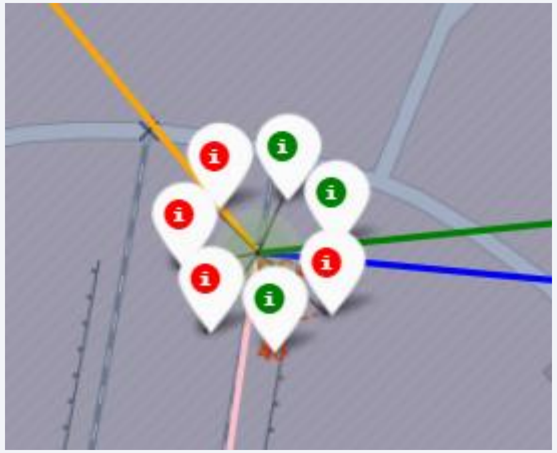


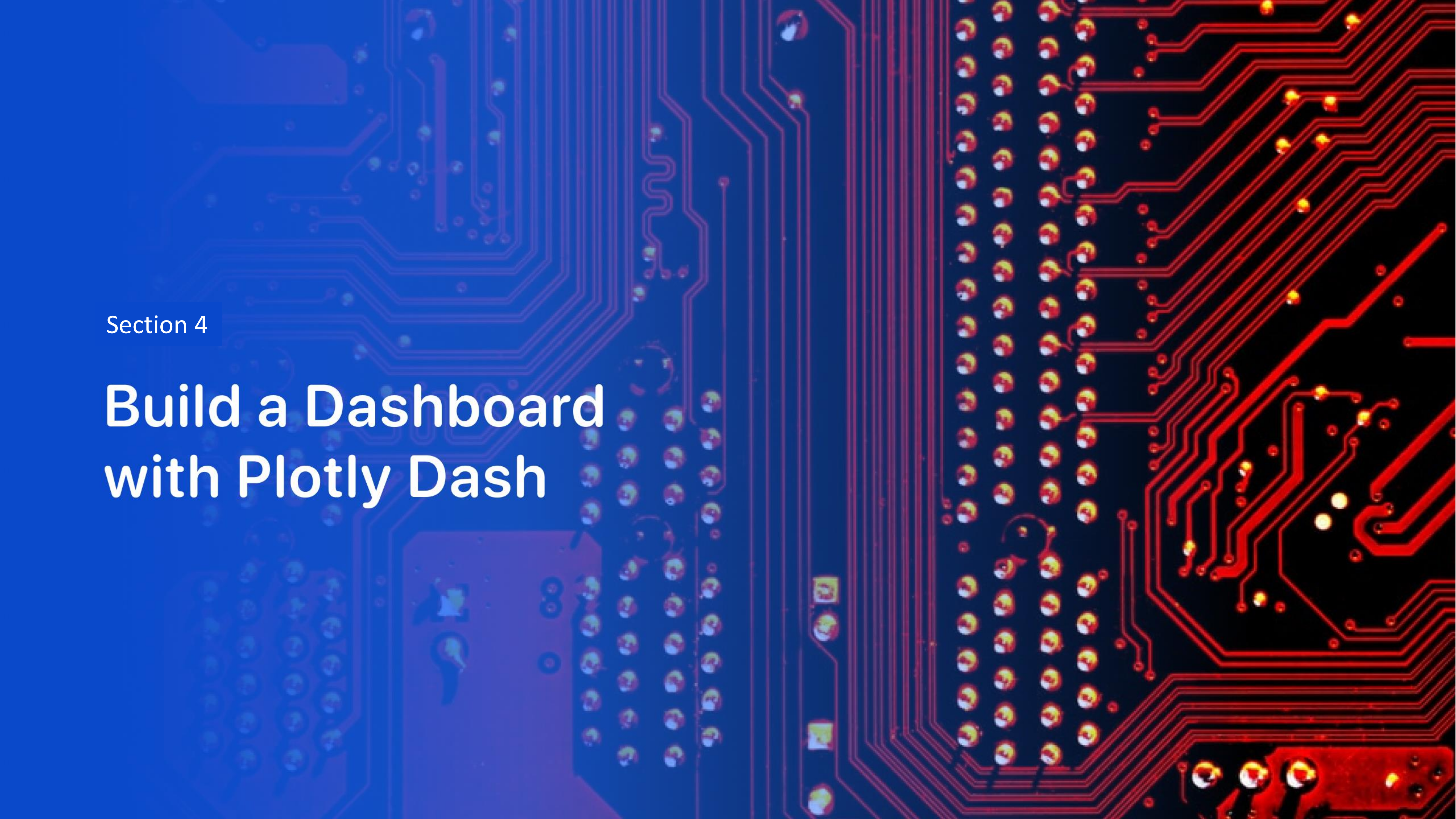
- We can see that there are 2 launch sites almost on top each other and third one in close proximity
- Highways and train tracks are also in vicinity

Distances between different infrastructures

- As seen on screenshots we could see different outcomes of some launches below, and on the right side there are highlighted paths with their length in real life to nearest key infrastructure points. For example closest highway, train tracks, coast line and city.
- Closest city to any of the launch sites is always at least 20km away to ensure safety in case of rocket malfunctioning.



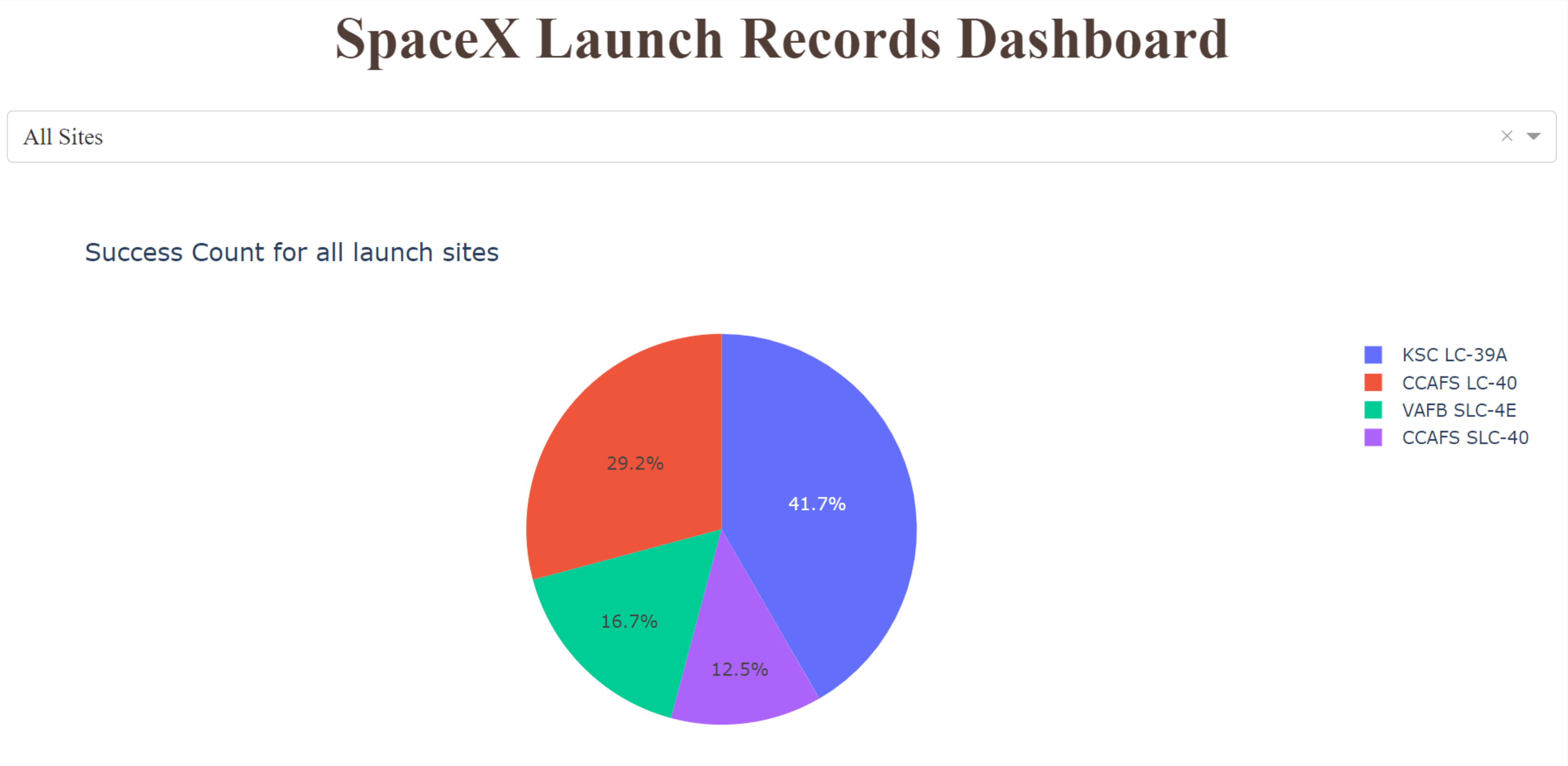




Section 4

Build a Dashboard with Plotly Dash

Distribution of successful launches by Site



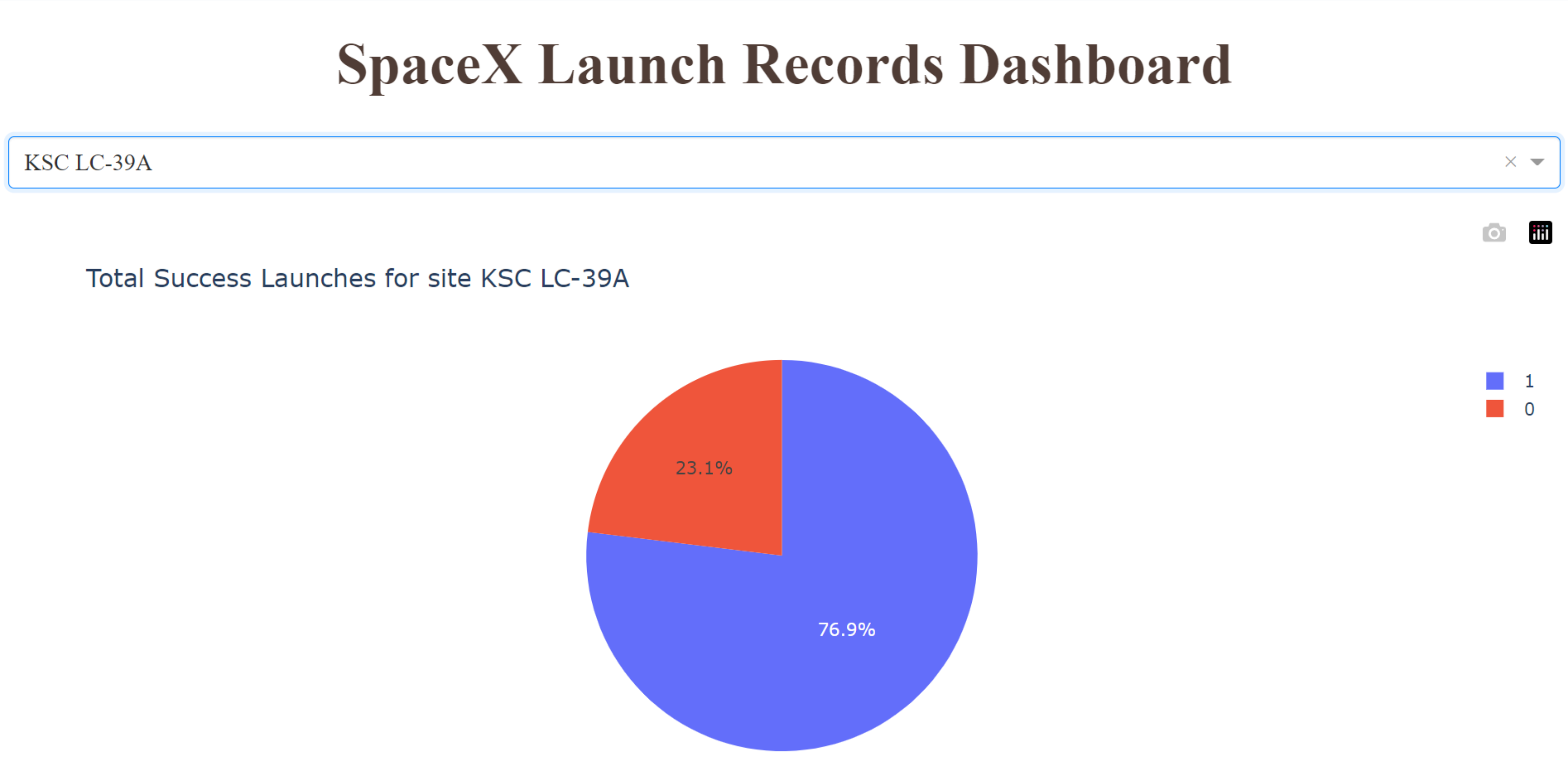
SpaceX Launch Records Dashboard

All Sites

Success Count for all launch sites

- According to pie chart, launch sites have huge impact on mission outcome.

Success ratio for best launch site

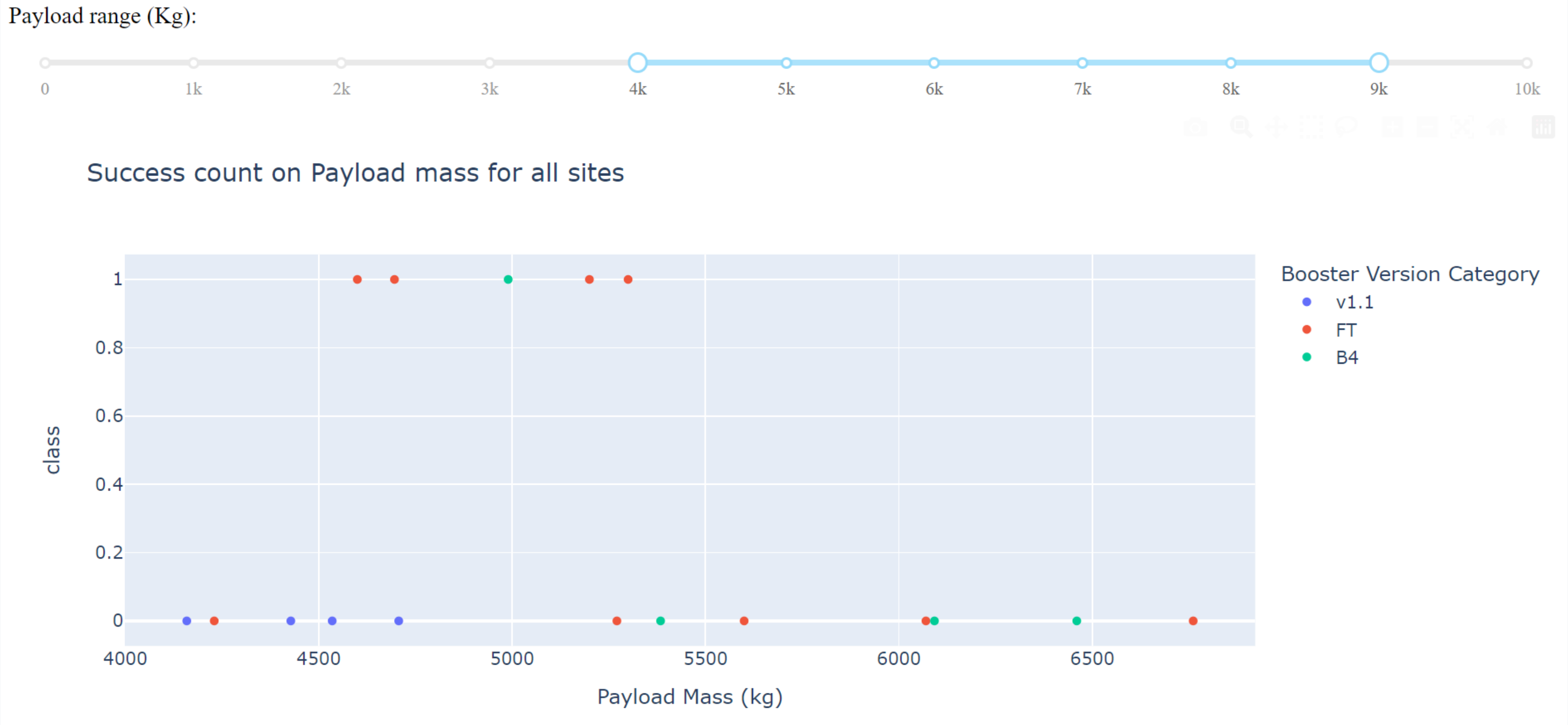


SpaceX Launch Records Dashboard

KSC LC-39A

- According to data, 76.9 launches from KSC LC-39A were successful meaning that this launch site has highest success rate out of all.

Success rate for payloads between 4-9 tons



- All successful landing were between 4,5 and 5,5 tons. Only boosters that managed to land successfully in this payload range are FT and B4.

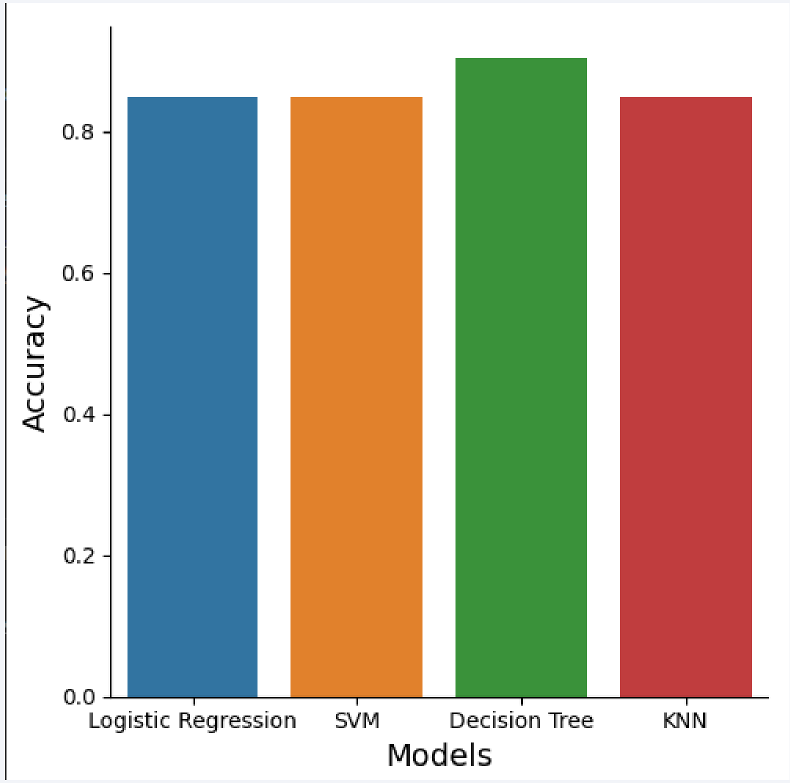
Section 5

Predictive Analysis (Classification)

Classification Accuracy

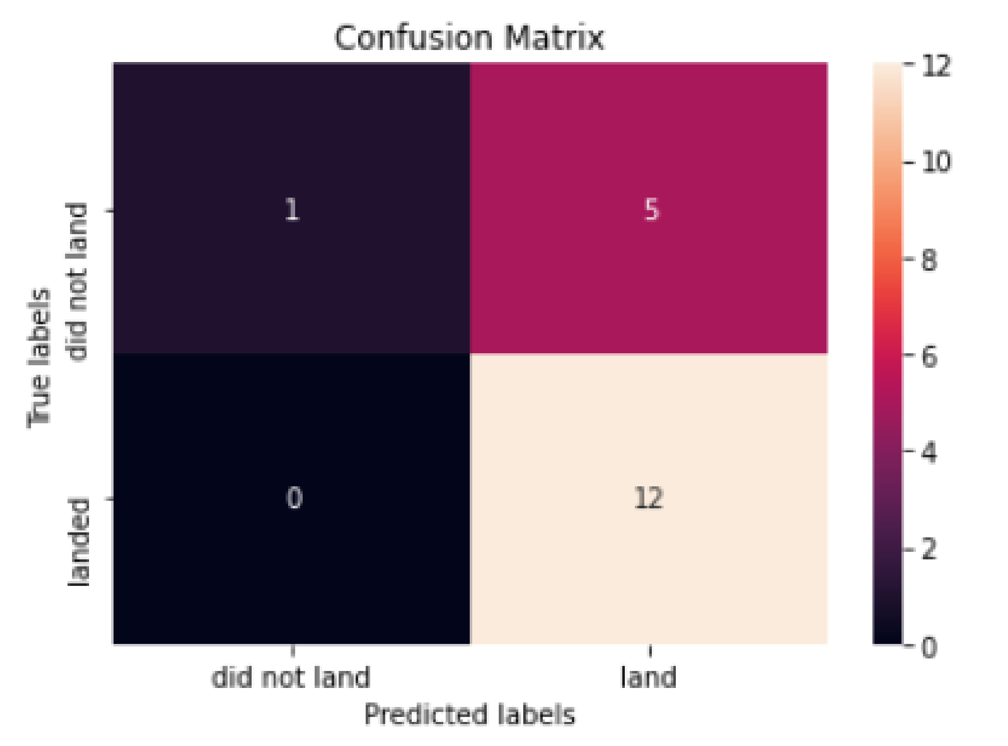
- Four classification machine learning models were tested with different parameters using grid search technique. Training was done on separate dataset to counter effect of overfitting on training data.

- Model that had highest accuracy for training data set was decision tree classifier



Confusion Matrix

- As seen on confusion matrix, decision tree might have had best accuracy but had overall worst recall out of all models because of higher number of false positives



Conclusions

- Key points from presentation:
- The overall best launch site was KSC LC-39A;
- Higher payload launches that means above 7 tons are less risky then those below this mark
- Most of missions nowadays end up being successful but high growth of success rate has slowed down and stopped
- Decision tree classifier even thought it’s pretty simple model it can be used to predict correct values of landing out come, but be aware of lower recall score.

Appendix

- Machin learning notebook seems to have some weird problem between np.int and sklearn module where it is very verbose about it and clutters viewing space. That happens when fitting model from scratch so please try hiding outcome cell.

Thank you!