IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

DBel

September 5, 2023

Outline

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

Executive Summary

To be a competitive player in the commercial space flight industry, it is required to substantially lower flight costs by reusing booster rockets. We analyzed data from SpaceX flights regarding which flight parameters were associated with a successful booster landing and with the ability to reuse boosters. We applied API data access and webscraping to obtain relevant data on SpaceX missions. Using data visualization techniques, SQL queries and geolocation mapping we specified the conditions under which a booster landing is favored, including the optimal launch site, the target orbit, payload ranges and years of experience it takes. We developed an interactive dashboard enabling visual evaluation of success rates stratified by launch site, booster versions and payload mass. Lastly, we developed a machine learning model with 83% accuracy, i.e. much better than a 50% chance guess, that can be deployed to assess if a potential mission is going to lead to a successful landing or not and, thus, provide a basis for an executive decision if a prospect mission can be accommodated by lower flight costs or if it is not worth taking a budgetary risk.

Introduction

- SpaceX is a pioneer in commercial space rocket launches. It has been providing services to a growing number of clients, including governmental agencies such as NASA. One of the reasons for its stand-alone status within the industry and its prevalence over competitors is its technical ability to reuse launch booster rockets. This allows them to substantially lower the costs per flight using Falcon 9 rockets down to 62 Mio., compared to offers of 165 Mio. by its competitors.
- The prerequisite for the reusage of boosters is the ability to land them safely back to Earth after the mission had launched and the booster detached from its cargo.
- As an aspiring aeronautics company, we aim to analyze what the optimal conditions are to recover a booster, thus potentially enabling us to lower costs per flight. If we restrict our launch offers to missions that have a high chance of a successful booster landing, we can predictably minimize the costs and obtain part of the market share that is currently occupied by SpaceX.

Section 1

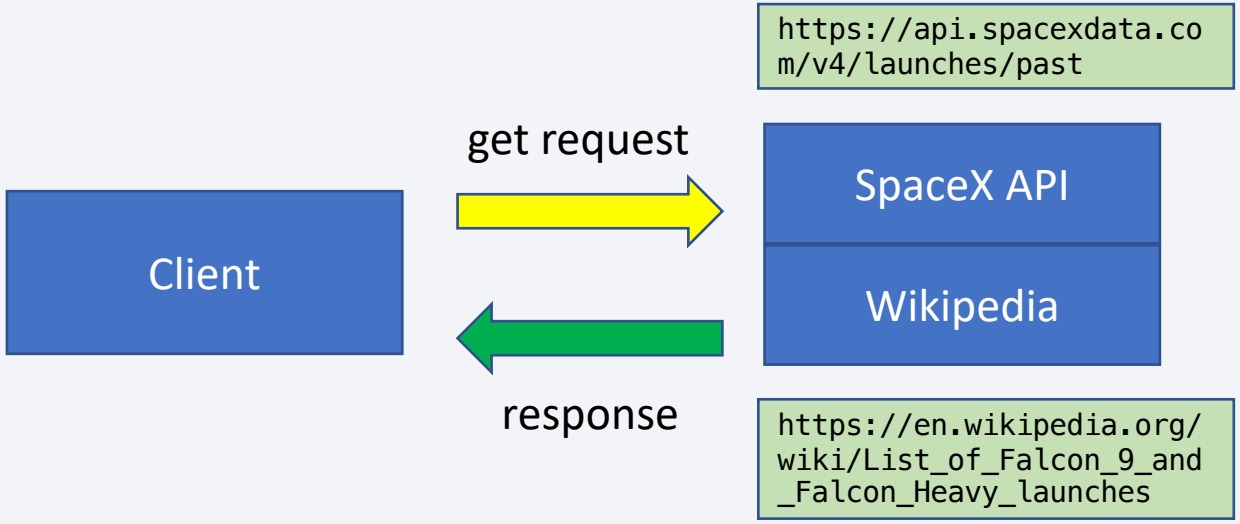
Methodology

Methodology

- Data collection methodology:
  - API calls; Webscraping and html parsing
- Perform data wrangling
  - Outcome variable engineering
- 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 on past launches of Falcon 9 rockets were obtained from:
  - a static version of the official SpaceX API
  - a static version of a Wikipedia page on Falcon 9 launches
- "requests" API call and "BeautifulSoup" html parser python libraries were used to extract the json and html data, respectively



Data Collection – SpaceX API

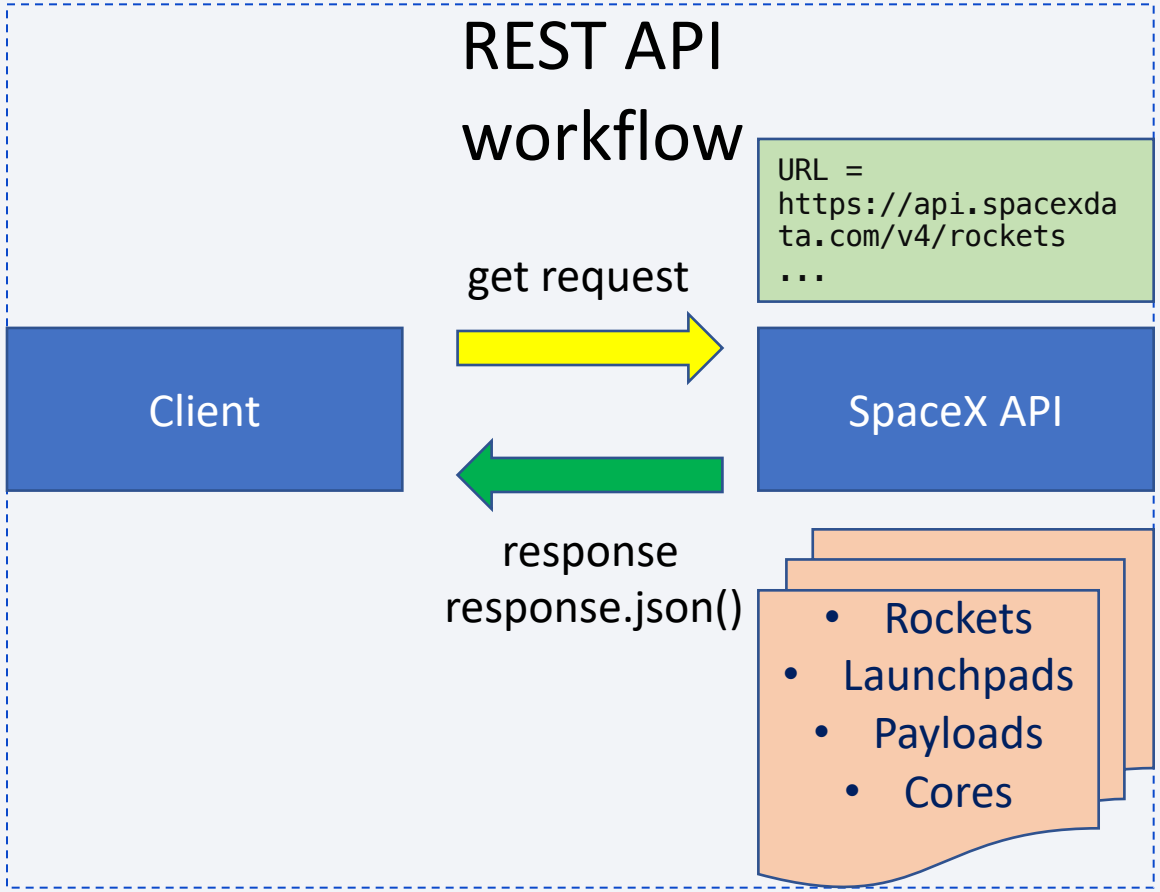
- Extracted textual data assigned to rocket associated codes (such as "5e9d0d95eda69955f709d1eb") as obtained from the SpaceX API;

  using REST API call libraries:

  - requests
  - requests_toolbelt

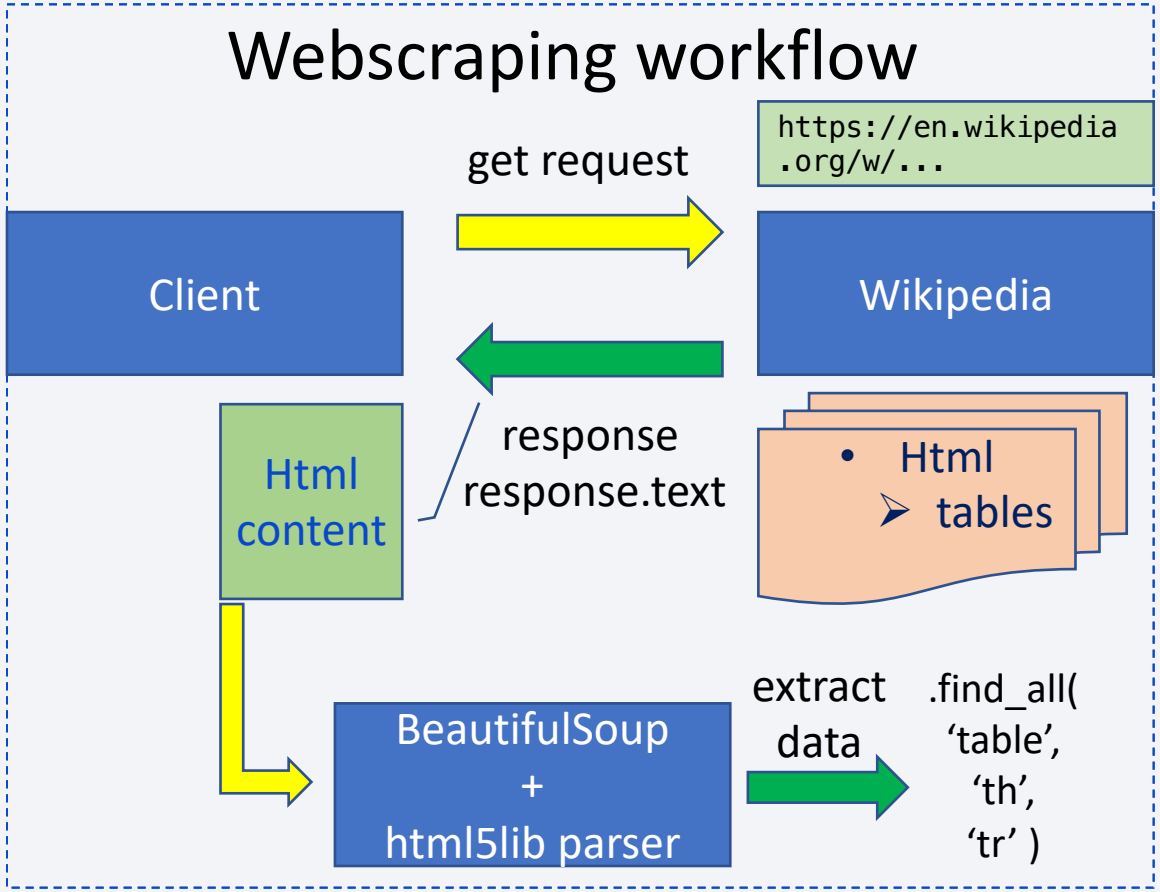
  and respective functions:

  - request.get(URL/queryvalue)
  - sessions.BaseUrlSession(URL).get(queryvalue)
- Filtered for Falcon 9 launches
- Substituted missing Payload mass values by the arithmetic mean
- Code: https://github.com/dblnk/IBM_DataScience_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



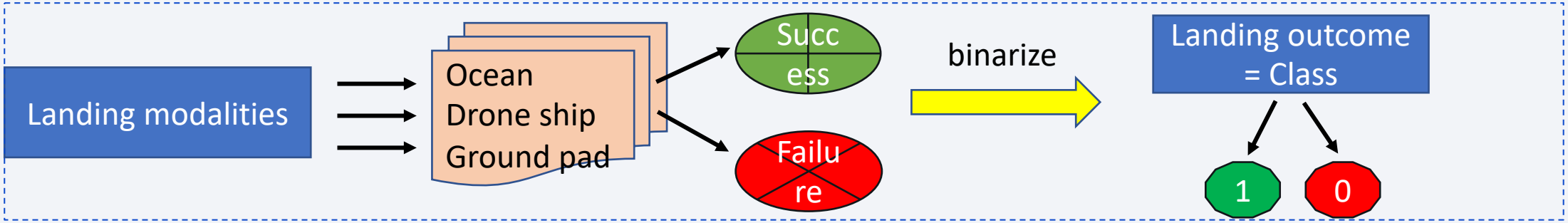
Data Collection - Scraping

- Obtained historical data on SpaceX launches in html format from Wikipedia using HTTP request
- Parsed the html code with BeautifulSoup
- Extracted relevant data from the soup object using find_all() and relevant tag names
- Code: https://github.com/dblnk/IBM_DataScience_Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Since we are primarily interested in whether a booster can be successfully landed or not, and do not need to differentiate between different landing modalities, we create a binary Outcome variable, called “Class”, with “1” representing “Success” and “0” representing “Failure” to land.



- Code: https://github.com/dblnk/IBM_DataScience_Capstone/blob/main/jupyter-labs-spacex-data-wrangling.ipynb

EDA with Data Visualization

- Different parameters were visually evaluated for their potential contribution to the success of a booster landing:
  - Does geographic location of the launch site matter?
  - Are some launch sites better suited for different payload mass ranges?
  - Are some orbit launches harder to land from?
  - Which orbit targets require more experience and which are more suitable for which payload mass ranges?
  - How many years of experience are needed to achieve consistently high landing success rates?
- Code: https://github.com/dblnk/IBM_DataScience_Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries in sqlite3 using sql magic were performed to retrieve potentially important insights from the Booster Launch data. Some of them are listed here (see Appendix):
  - **Identify unique launch sites:** *select distinct("Launch_Site") from SPACEXTABLE*
  - **Obtain total payload mass launched by NASA (CRS):**

    select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Customer" like 'NASA (CRS)%'
  - **Obtain average payload mass carried by F9 v1.1 rockets:**

    select round(avg(PAYLOAD_MASS__KG_),2) from SPACEXTABLE where "Booster_Version" like 'F9 v1.1%'
  - **Identify mission outcomes:**

    select sum(case when "Mission_Outcome" like '%Success%' then 1 else 0 end) as "Number of successful missions",

    sum(case when "Mission_Outcome" like '%Fail%' then 1 else 0 end) as "Number of failed missions"

    from SPACEXTABLE
  - **List landing outcomes in the first 7 years of operation:**

    select "Landing_Outcome", count(*) as "Count" from SPACEXTABLE where Date between "2010-06-04" and "2017-03-20" group by "Landing_Outcome" order by "Count" desc
- Find complete code here:
  https://github.com/dblnk/IBM_DataScience_Capstone/blob/main/jupyter-labs-eda_sqllite.ipynb

Build an Interactive Map with Folium

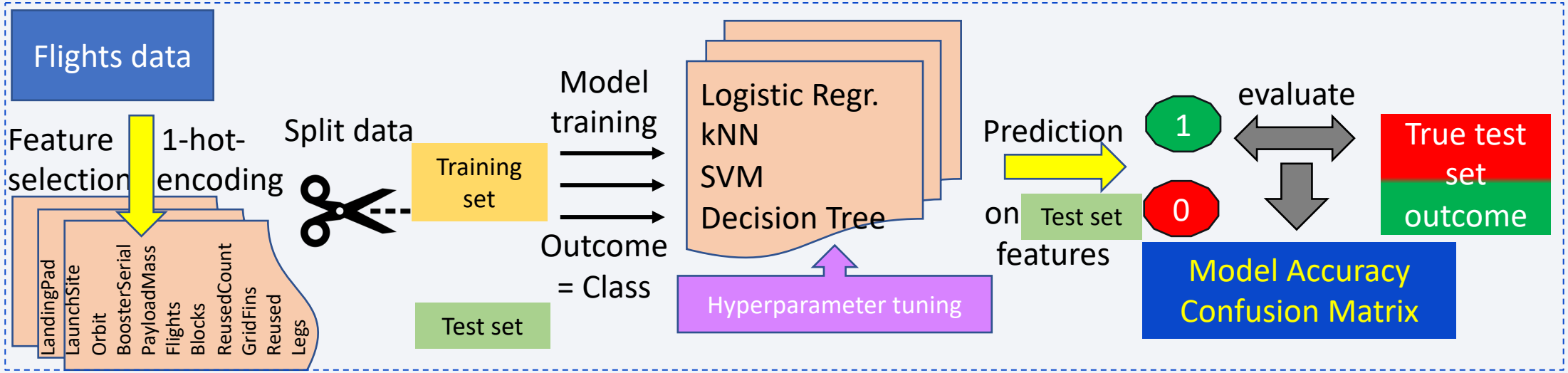
- folium library was used to map launch site locations onto a global folium.Map object
- folium.Circle and folium.map.Marker objects were added to locate SpaceX launch sites on the map
- MarkerCluster was populated with folium.map.Marker objects with successful and failed landings at each of the respective launch sites, enabling a quick overview over total attempted landings and success rate per site
- MouseMarker was used to obtain geolocations of geo- and topographic features such as coastline, towns, railroads and highway.
  A geometric function enabled calculation of distances to the launch sites. Distances were added using folium.Marker and connecting lines added using folium.PolyLine
- Full code available at:
  https://github.com/dblnk/IBM_DataScience_Capstone/blob/main/jupyter-labs-launch-site-location.ipynb

Build a Dashboard with Plotly Dash

- Built a dashboard using dash and plotly python libraries which displays
  - ➢ the success rate of the different launch sites
  - ➢ payload mass ranges among successful and failed landings
  - ➢ the booster versions associated with failed and successful landing attempts
- The user can select the data for all sites or choose any single site.
  The user can also specify the desired payload mass range to be displayed
- These plots will enable exploring which launch sites, booster versions and payload masses were associated with successful landings. That will help select the most promising specifications for successful missions, such as the right booster version, payload to carry or launch site to operate from.
- Code: https://github.com/dblnk/IBM_DataScience_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

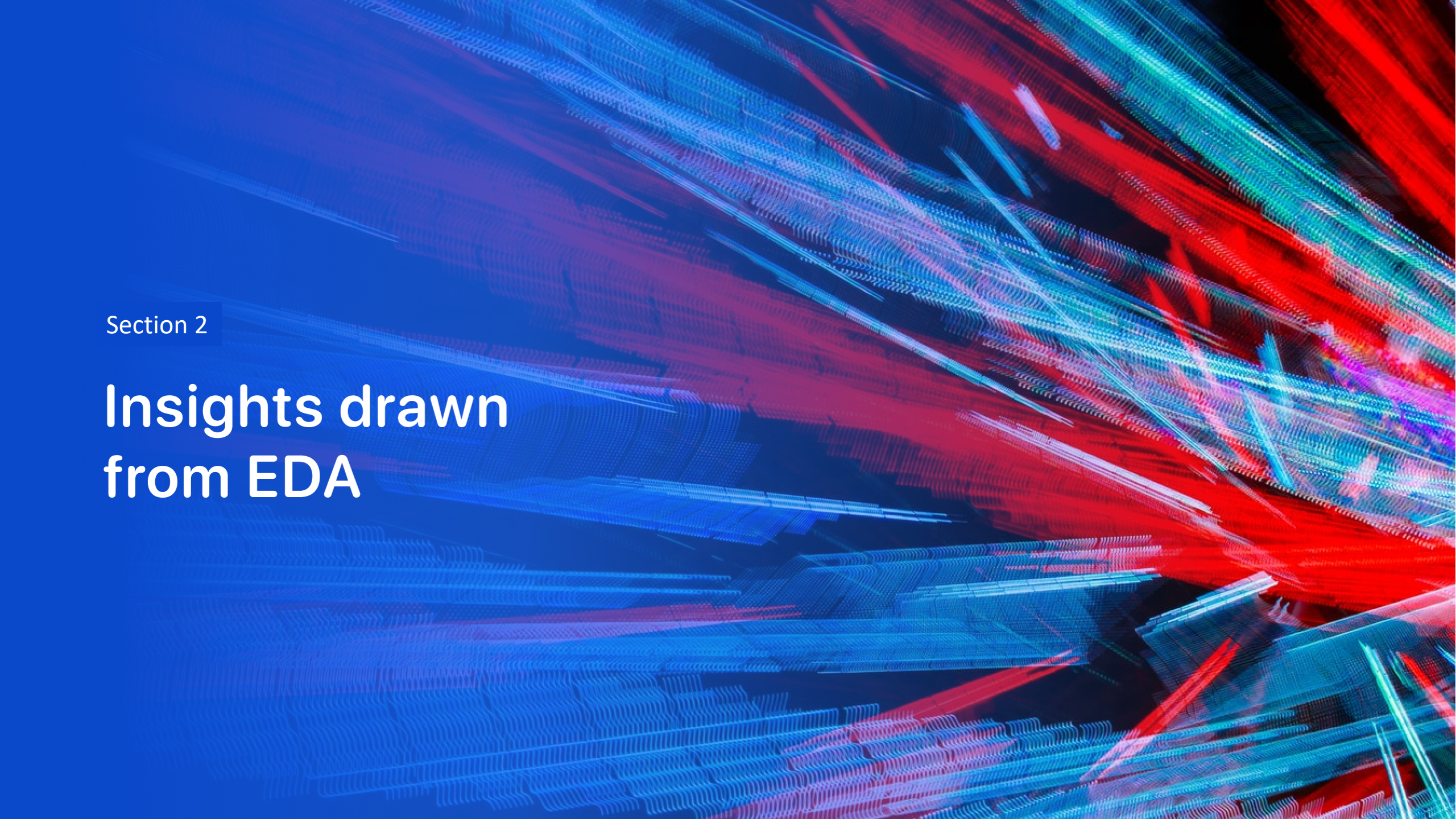
- Features likely to associate with landing outcome as extrapolated from EDA were selected and if binary, then one-hot-encoded.
- Dataset was split into training and test set, allocating 20% of data to testing.
- Using hyper-parameter tuning, logistic regression, kNN, SVM and Decision Tree models were trained on the training set and evaluated on the test set.



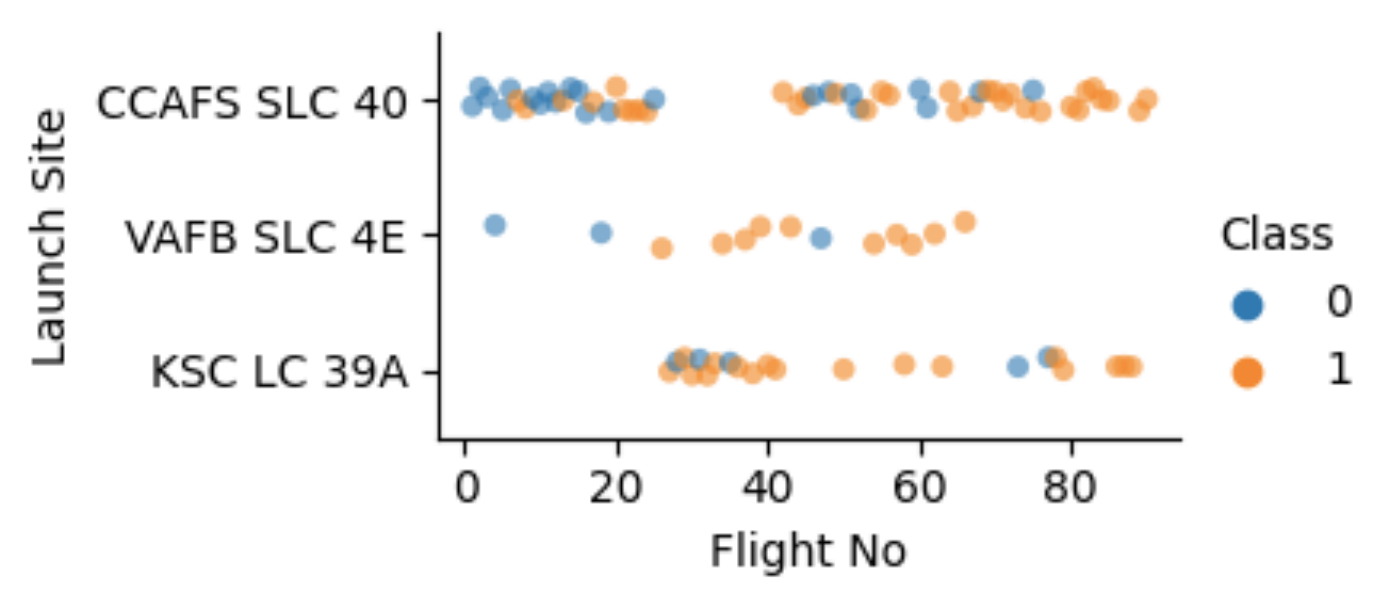
- Code: https://github.com/dblnk/IBM_DataScience_Capstone/blob/main/jupyter-labs-spacex-machine-learning.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

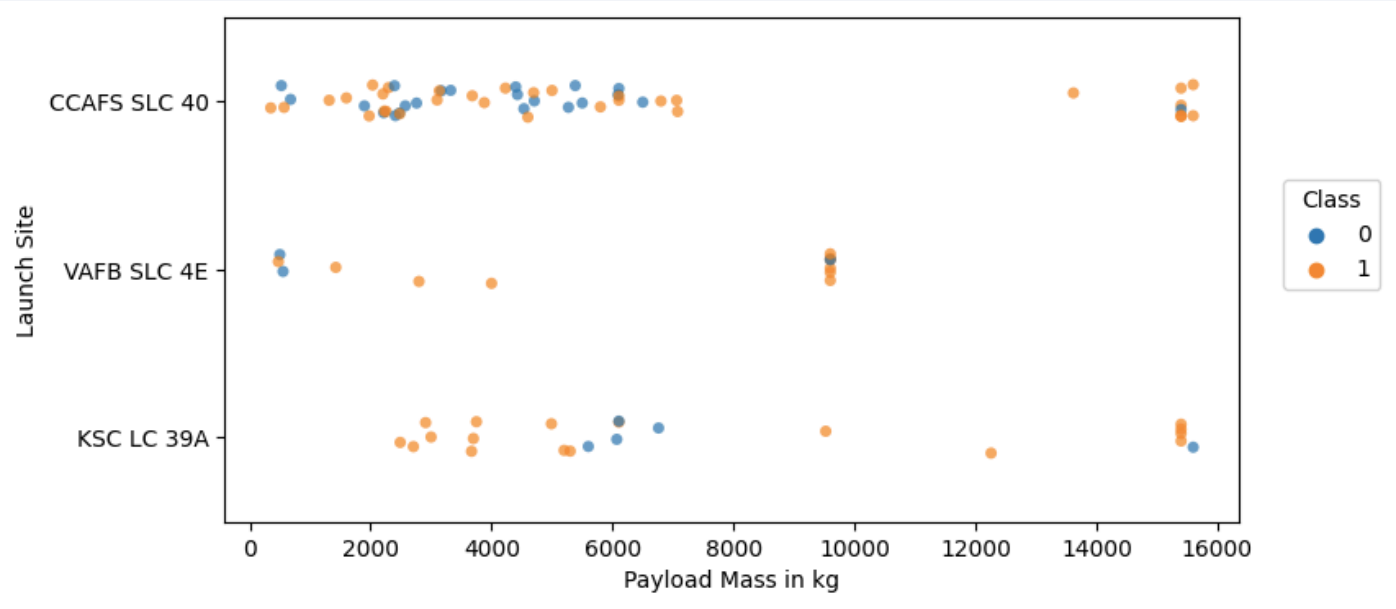

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The Cape Canaveral (CCAFS) site in Florida was the pioneering site and therefore accumulated many failures during the first flights. The success rate improved by a margin.
- The adjacent Kennedy Space Center (KSC) site seems to have served as a substitute for CCAFS during the period of flights No. 20-40 and was operated intermittently afterwards. It seems to have profited from the learning curve of the CCAFS site as relatively few failed landings were recorded.
- The Vandenberg Airforce Base in California (VAFB) has rarely been used. However, except for the first two flights having failed, it associated with a high success rate. It hasn't been used for the past 25 flights.

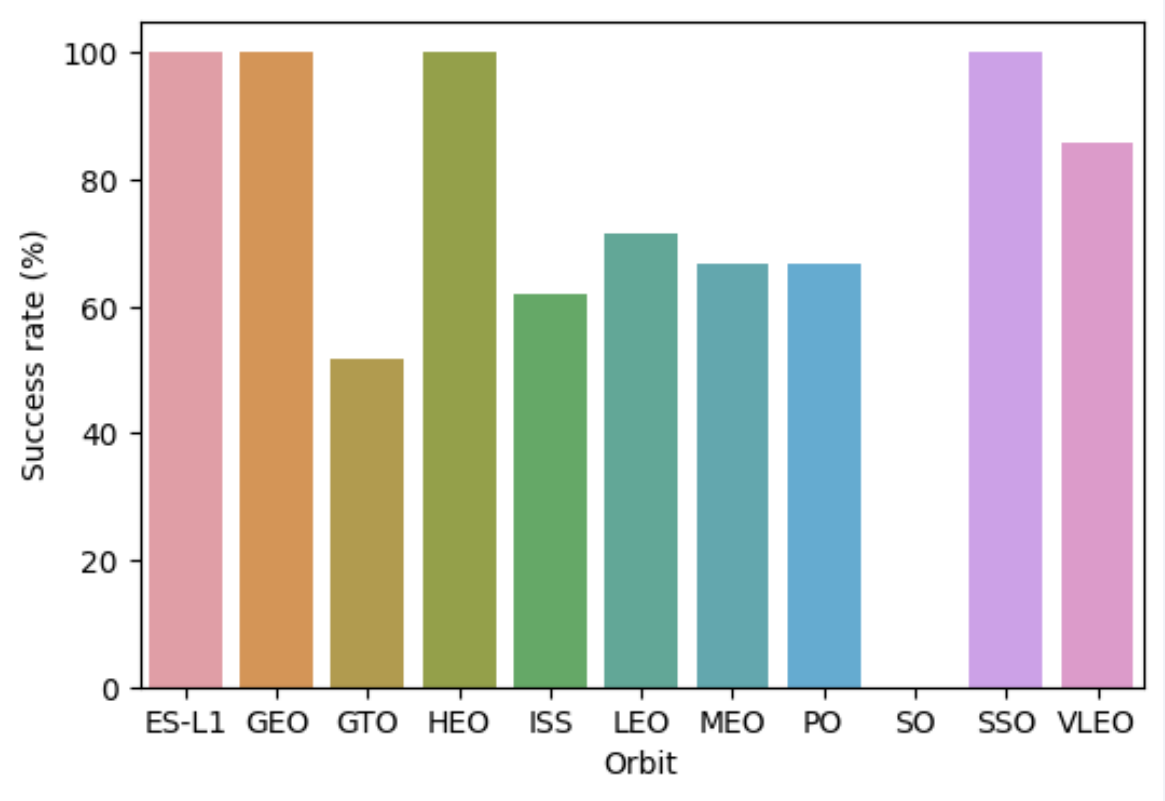
Payload vs. Launch Site



- Heavy payloads seems to land more successfully than light loads
- The VAFB site has not been used for loads heavier than 10,000 kg
- Otherwise, there is no clear relationship between payload and success at the CCAFS site
- KSC has recorded almost all failures at payloads of around 6,000 kg

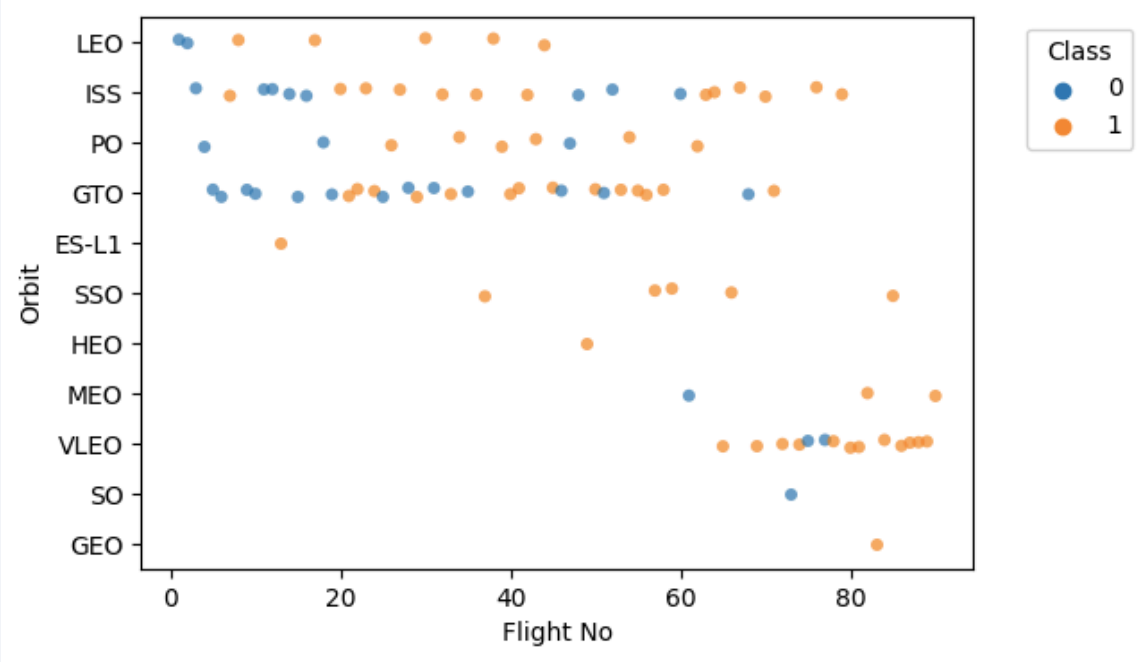
Success Rate vs. Orbit Type

- We must be careful in interpreting the results as only following orbits were targeted in >= 5 flights:
  - GTO (27)
  - ISS (21)
  - LEO (7)
  - PO (9)
  - SSO (5)
  - VLEO (14)
- Among these, GTO orbit has been the riskiest, while VLEO and SSO seem to be the safest for a successful landing.



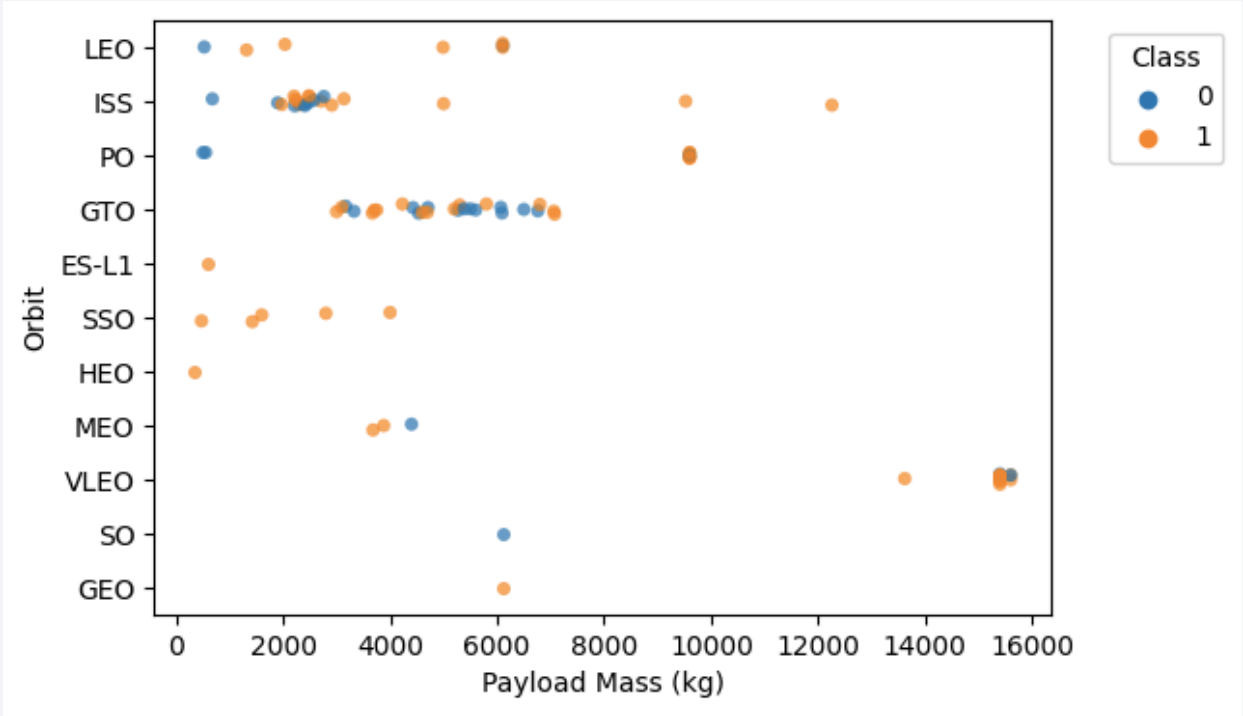
Flight Number vs. Orbit Type

- The low overall success rate of the GTO orbit is most likely due to it being most frequently targeted during the first years of the program
- In contrast, VLEO and SSO have been targeted relatively late in the program and profited from the gained experience
- Therefore, it is hard to tell if particular orbit types are more promising for a successful landing



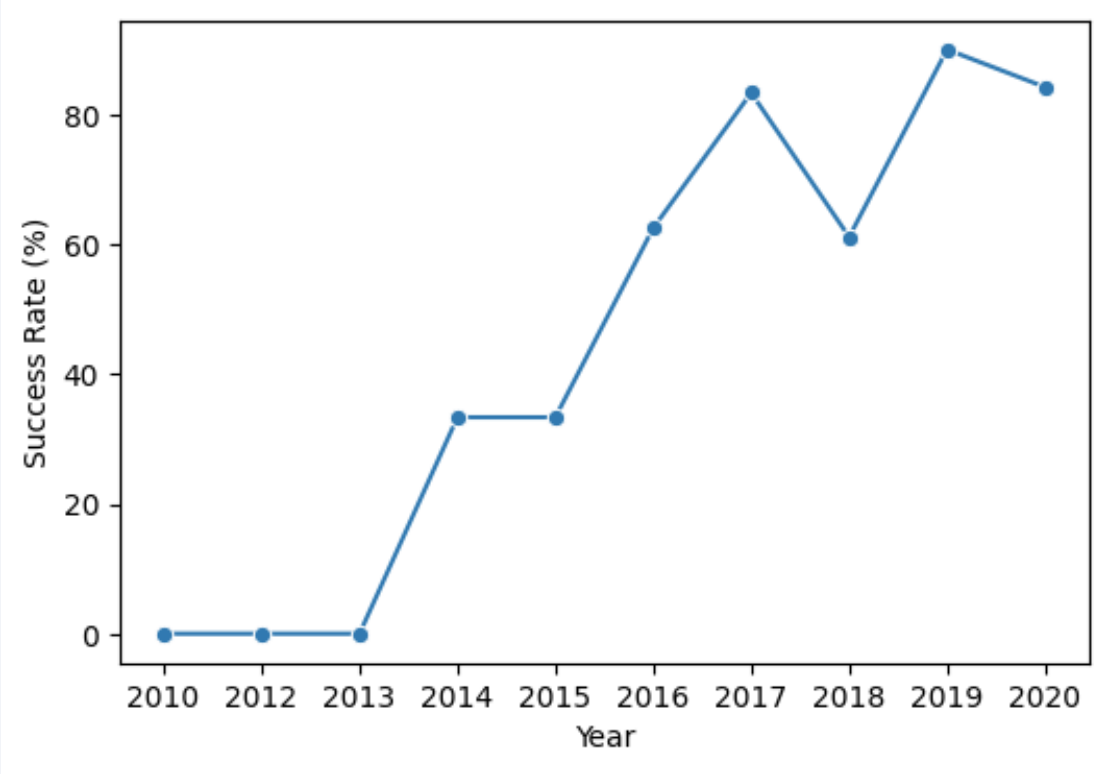
Payload vs. Orbit Type

- Some orbits were used for very heavy payloads (VLEO) while some for light loads (SSO)
- GTO was most often used for mid range payloads (4,000 – 7,000 kg)
- ISS has the widest range of payloads with most flights carrying around 2,000 – 3,000 kg



Launch Success Yearly Trend

- Examining the development of the SpaceX program, we can identify a clear learning curve with first successful landings recorded in the 5th year of the program.
- After 2014 - a continuous improvement reaching around 80-90% success rate in 2019-2020.



All Launch Site Names

- Found the names of the unique launch sites:

UNIQUE LAUNCH SITES
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- SpaceX operates out of Florida with two Cape Canaveral and one Kennedy Space Center sites, as well as out of California using the Vandenberg Air Force base.

Launch Site Names Begin with 'CCA'

- Listed 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
06-04-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
12-08-2010	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)	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
10-08-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
03-01-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- Cape Canaveral was the among the first sites to be operated out of as early as in 2010. No or only light payload, including a humoristic barrel of cheese was used on the first two missions.
- NASA seems to be the main customer during the early days in Cape Canaveral.

Total Payload Mass

- Calculated the total payload carried by boosters from NASA (CRS):

TOTAL PAYLOAD MASS CARRIED BY NASA(CRS) in KG
48213

- NASA transported 48 tons of material during the time investigated

Average Payload Mass by F9 v1.1

- Calculated the average payload mass carried by booster version F9 v1.1

AVERAGE PAYLOAD MASS CARRIED BY F9 v1.1 BOOSTERS in KG
2534.67

- F9 v1.1 boosters seem to be suited for moderate loads of 2.5 tons on average.

First Successful Ground Landing Date

- Found the date of the first successful landing outcome on a ground pad:

Date of First Ground
Pad Landing Success:
2015-12-22

- It took SpaceX five years of development to hit the first successful booster landing on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listed the names of boosters which have successfully landed on a drone ship and had a payload mass greater than 4000 but less than 6000:

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- Successful landings on a drone ship were achieved using F9 FT boosters carrying relatively heavy loads of 4 to 6 tons

Total Number of Successful and Failure Mission Outcomes

- Calculated the total number of successful and failed missions:

Number of successful missions	Number of failed missions
100	1

- SpaceX is very successful in their missions in general with only 1% failure rate.

Boosters Carried Maximum Payload

- Listed the names of the boosters which have carried the maximum payload mass:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

- SpaceX uses F9 B5 booster versions for transportation of maximum loads of 15.6 tons.

2015 Launch Records

- Listed the failed landing outcomes on drone ships, their booster versions, and launch site names for the year 2015:

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Not sure why the execs wanted to have this data.
- Avoiding months of April and October to land back on drone ships due to the weather conditions would not be a robust conclusion. That's because, in 2015, SpaceX was still gathering technical experience.

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

- Ranked 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:

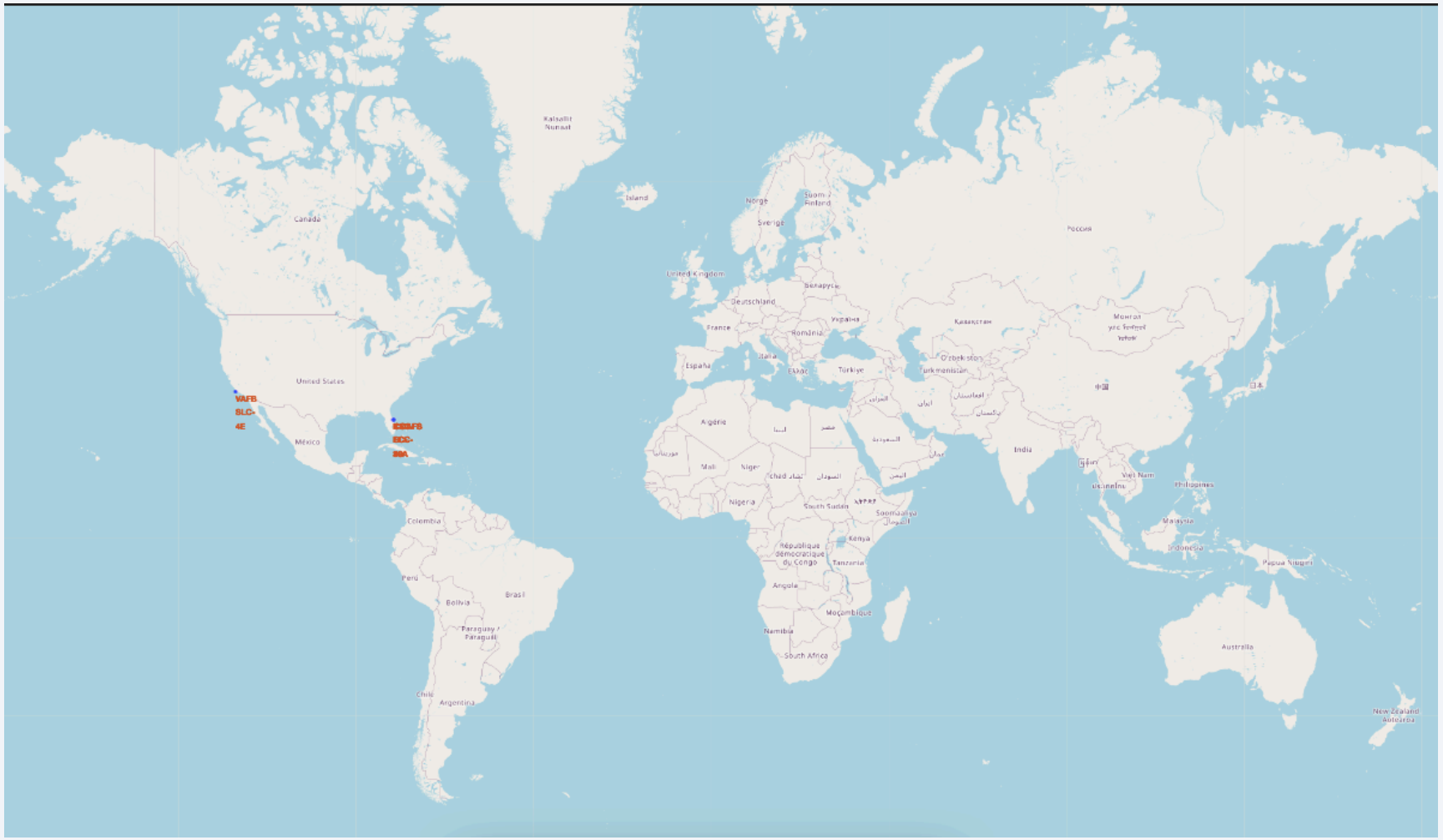
Landing_Outcome	Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship	1
Failure (parachute)	1

- 1 out of 3 times SpaceX was not attempting to land. All ground pad landings were successful. Drone ship and ocean landings were prone to fail half of the time.


Section 3
Launch Sites
Proximities Analysis

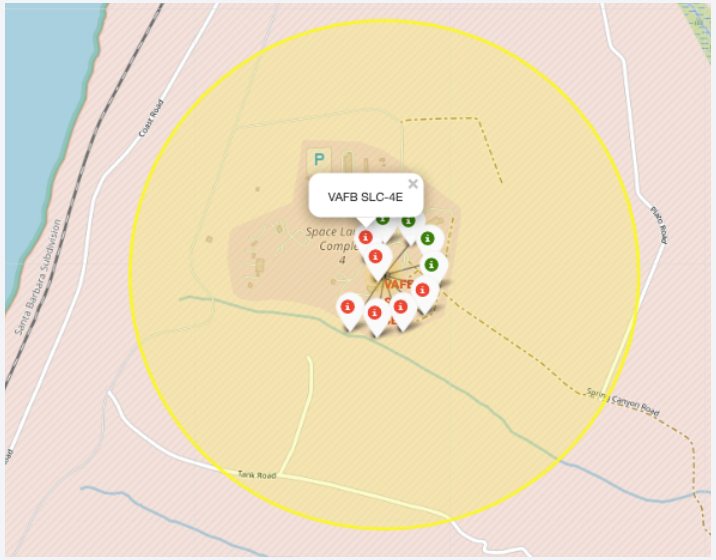
Launch site locations

- Global map shows that SpaceX operates out of the USA.
- All four launch sites are on the ocean coast, northern of the Tropic of Cancer.
- Risk for mainland damages is reduced
- Advantageous latitude for targeting orbits

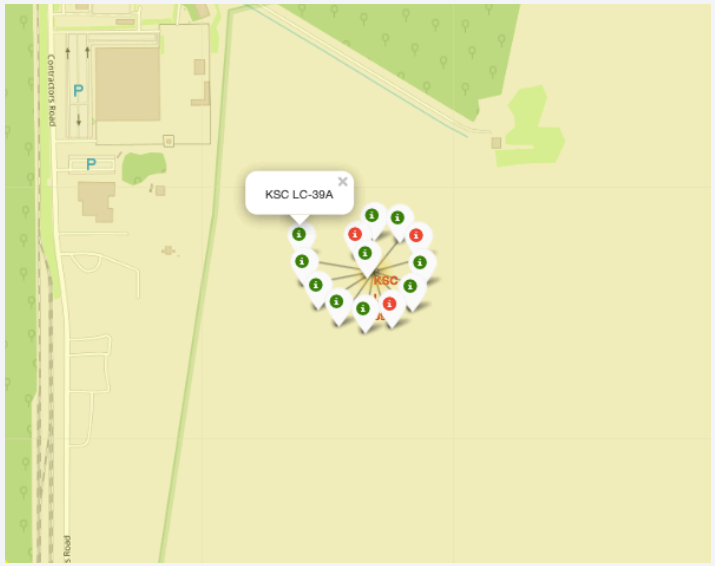


Landing outcomes per launch site

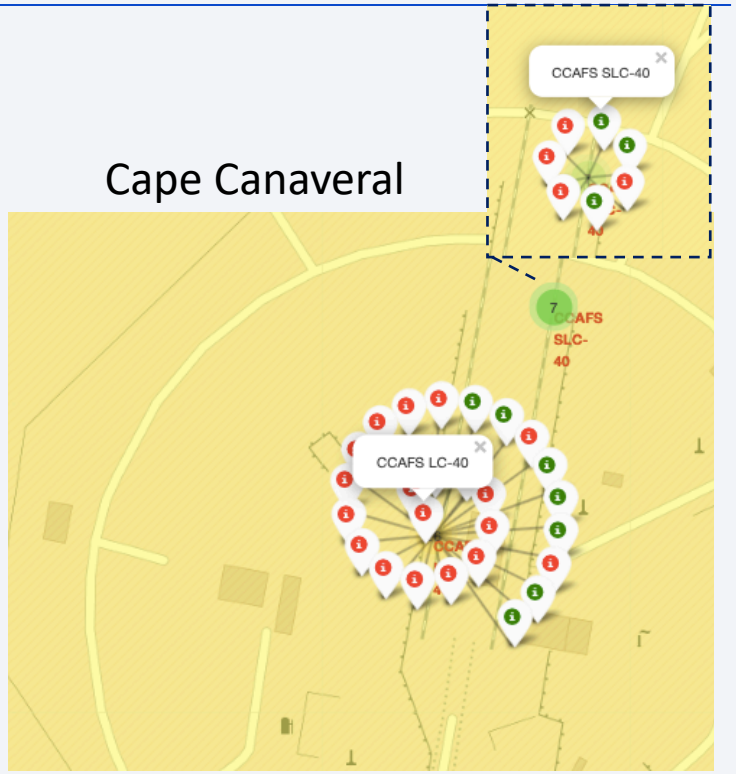
Vandenberg AirForce Base



Kennedy Space Center



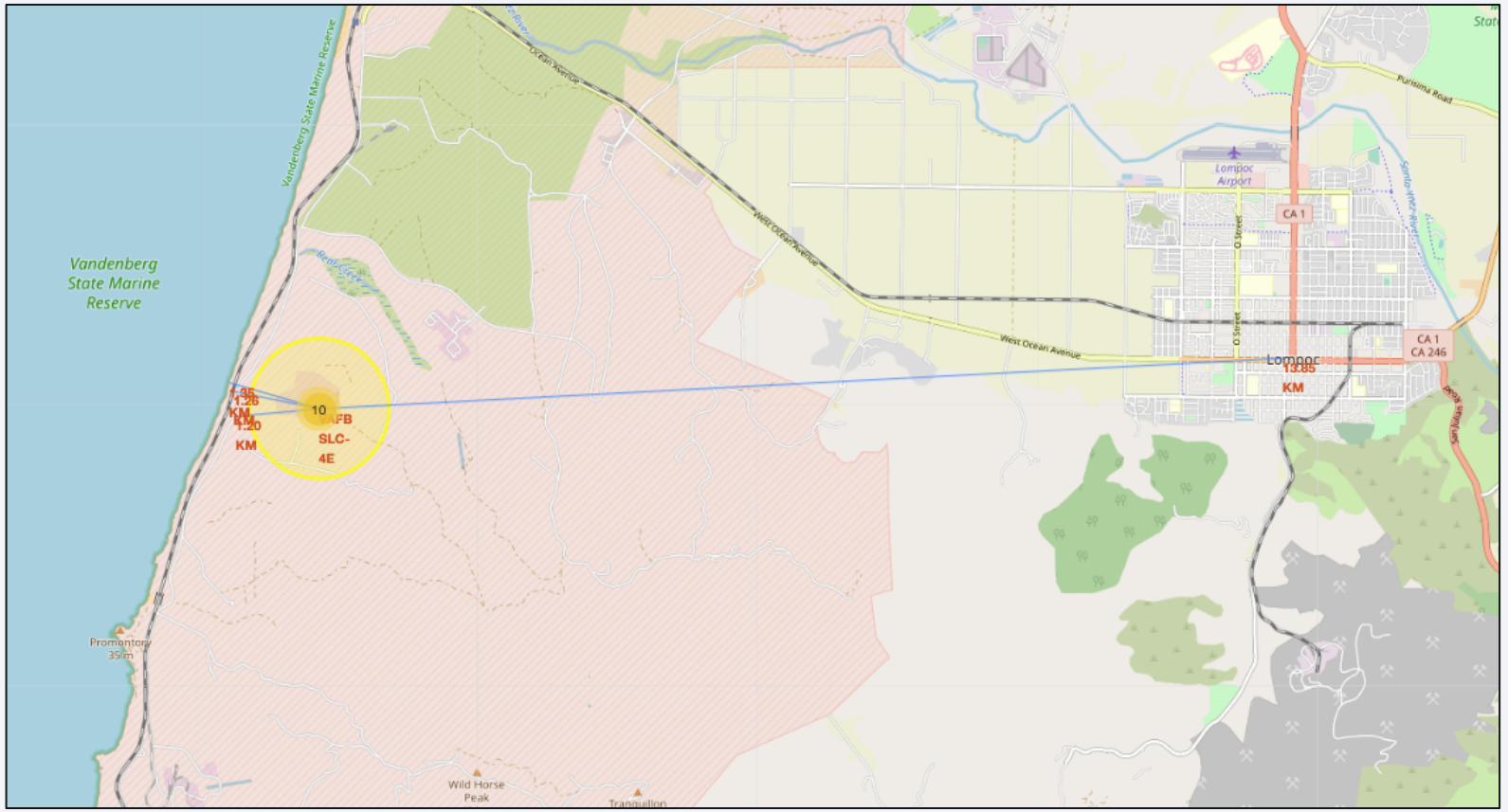
Cape Canaveral



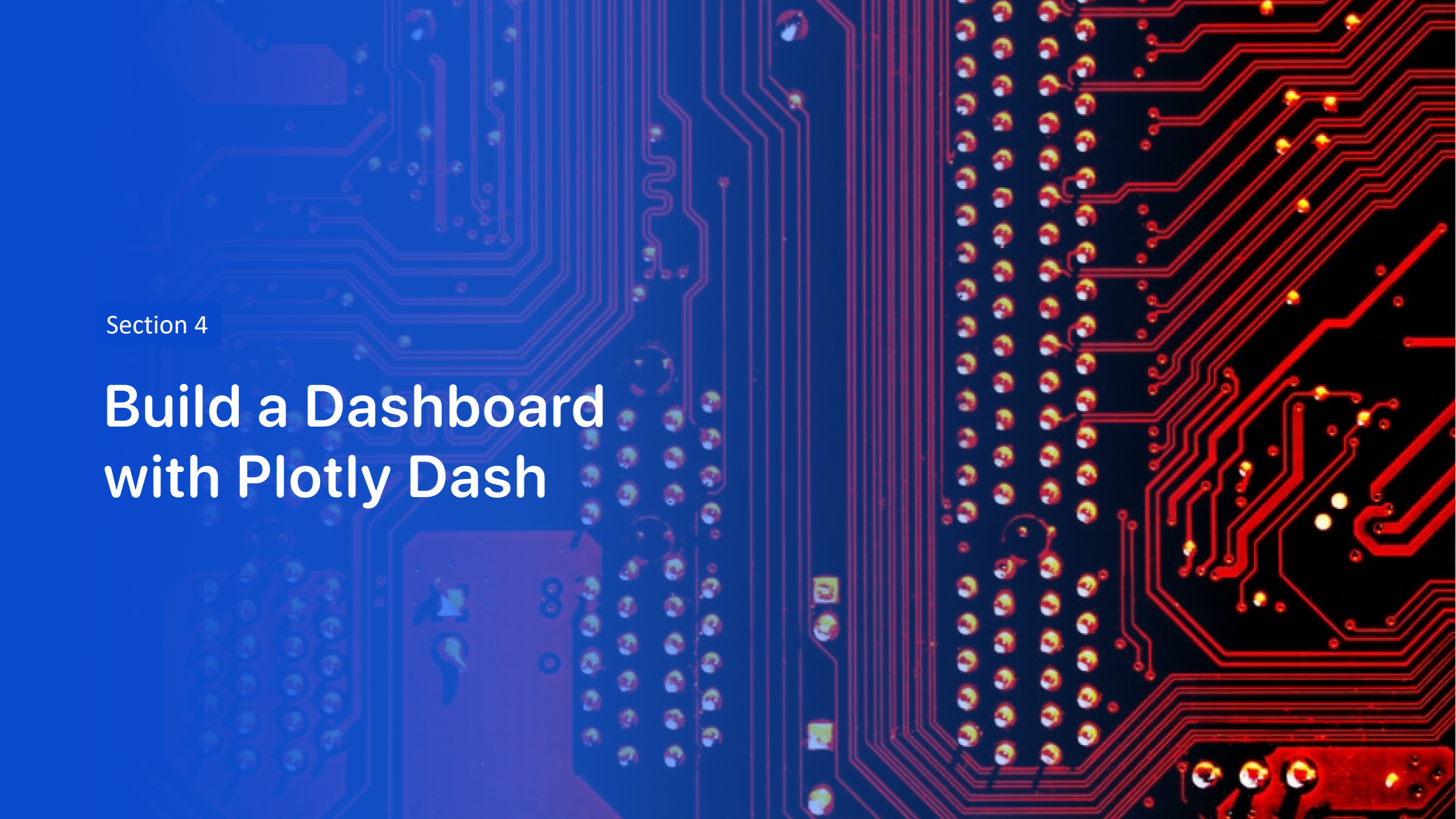
- Cape Canaveral has the most experience and has accrued a lot of failed landings, similarly to the less frequented Vandenberg site with 10 landings. Landings in later stages of the program were much more often successful than in early stages
- Kennedy Space Center seems to have been more successful overall, likely because it was operated during mid stages of the program as shown on slide 18 and profited from the expertise gained during Cape Canaveral landings.

Infrastructure and geography of launch sites

- All SpaceX launch sites are within 1-2 km distance away from the coast, as well as from a railroad and a highway. At the same time, nearest towns are 15-20 km away. Thus, logistic ease of delivery and safety of US citizens are ensured.

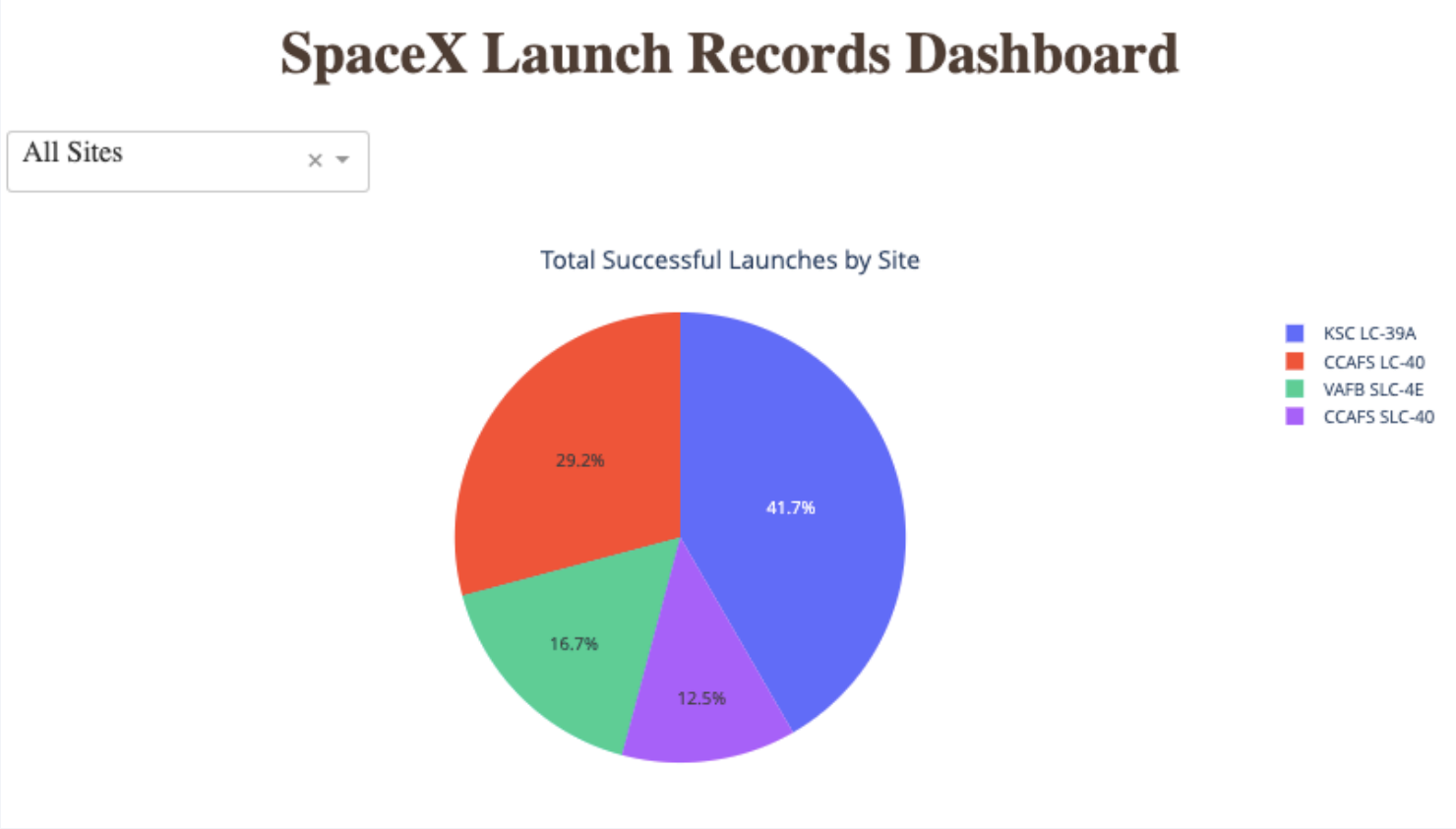


Vandenberg Air Force Base on Pacific Ocean coast


Section 4
Build a Dashboard with Plotly Dash

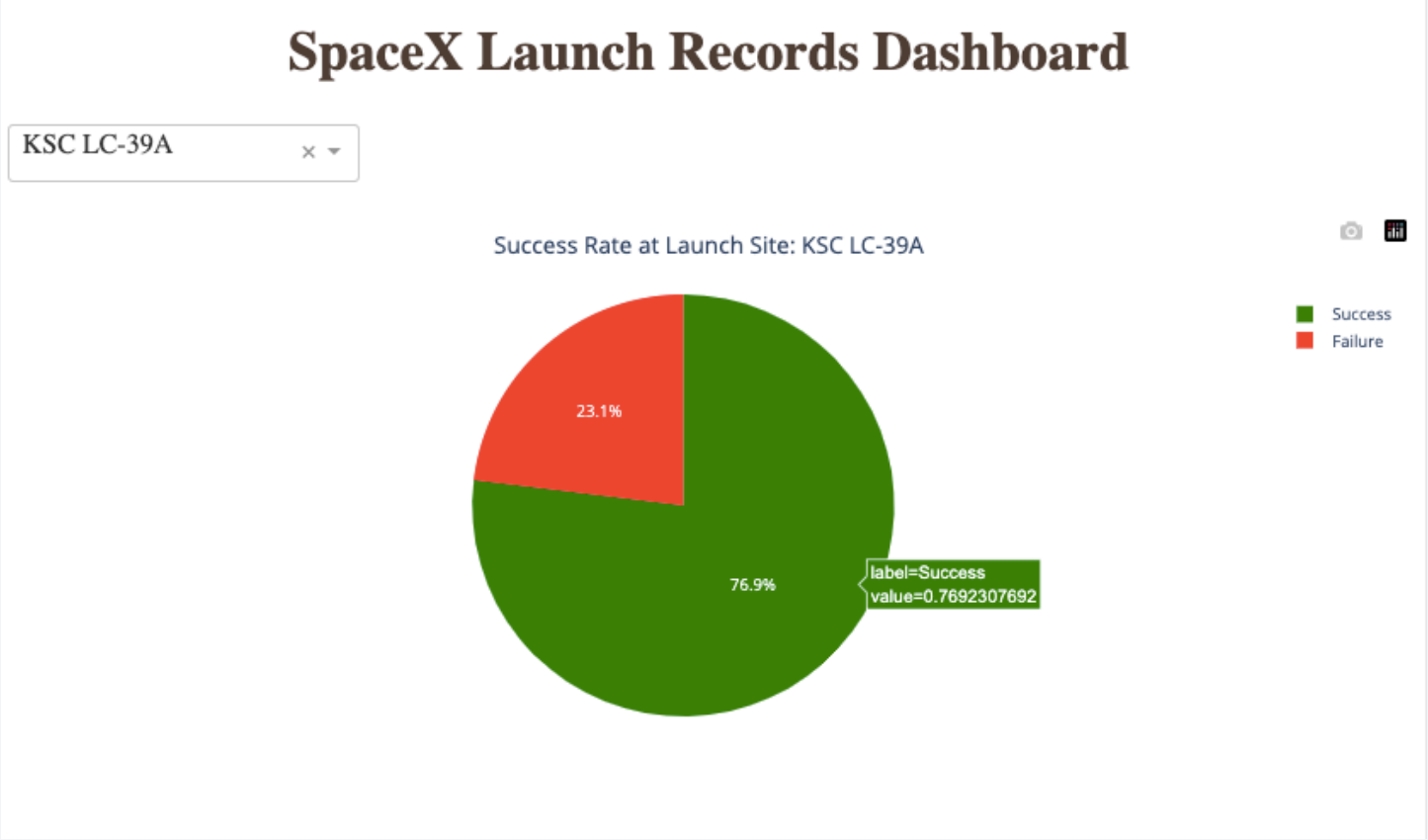
Launch Site shares in successful landings

- The most successful site is Kennedy Space Center showing as many successful landings as both Cape Canaveral sites together
- Vandenberg Air Force Base has only launched 1/6 of successful landings. However, we saw before that it is not as frequently used as other launch sites.

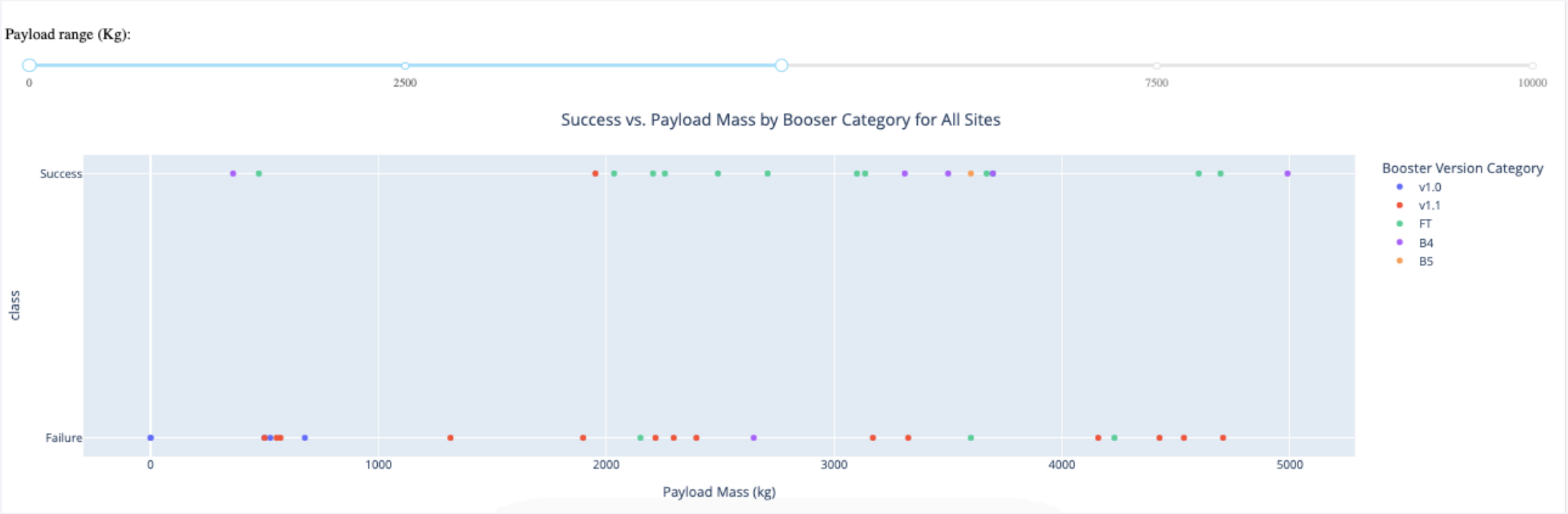


Launch Site with highest landing success rate

- The launch site with the highest landing success rate is the Kennedy Space Center with ca. 3 out of 4 missions achieving landing completion.



Influence of payload mass on Outcome by booster



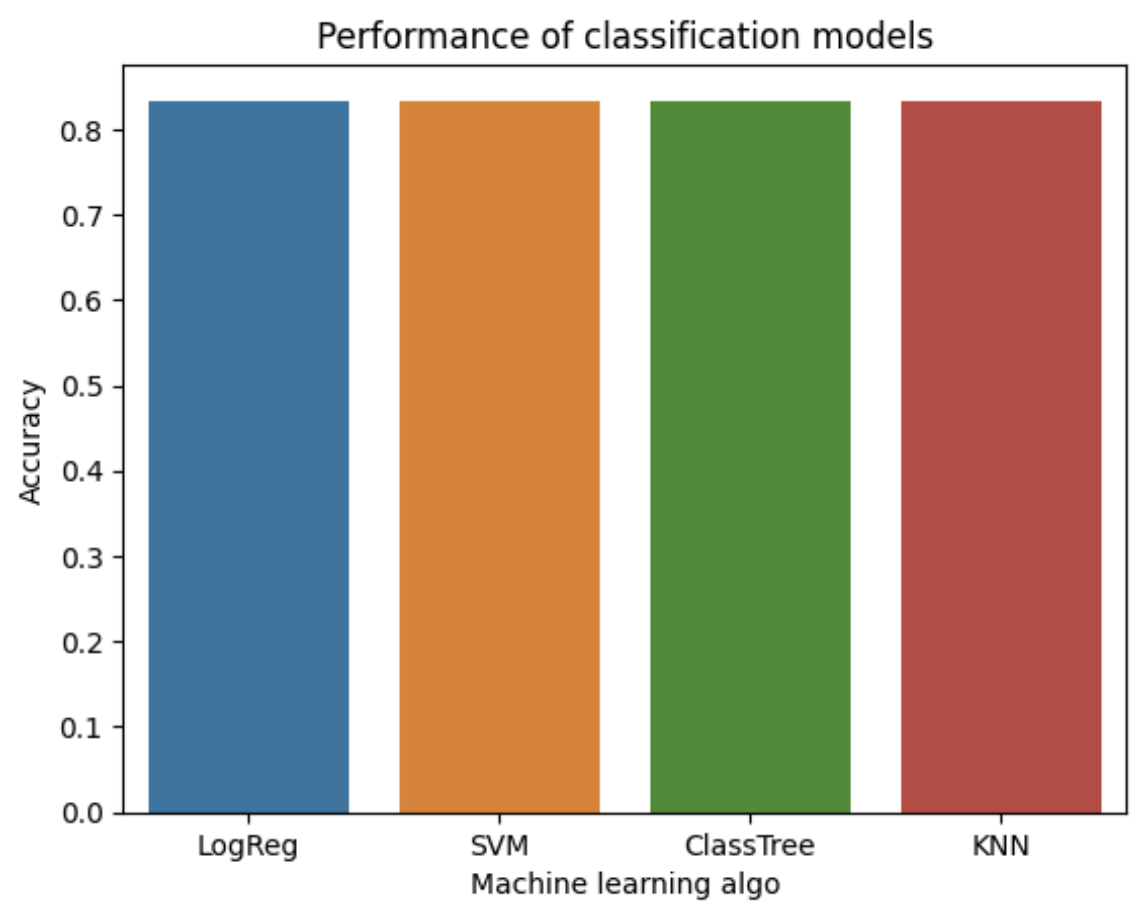
- Early booster versions v1.0 and v1.1 failed most of the time
- Moderate payloads of 2 - 4 tons were more likely to land
- Booster Versions FT followed by B4 partook in most successful landings
- Not enough data for B5 booster (only 1 landing attempt, but successful)

Section 5

Predictive Analysis (Classification)

Classification Accuracy

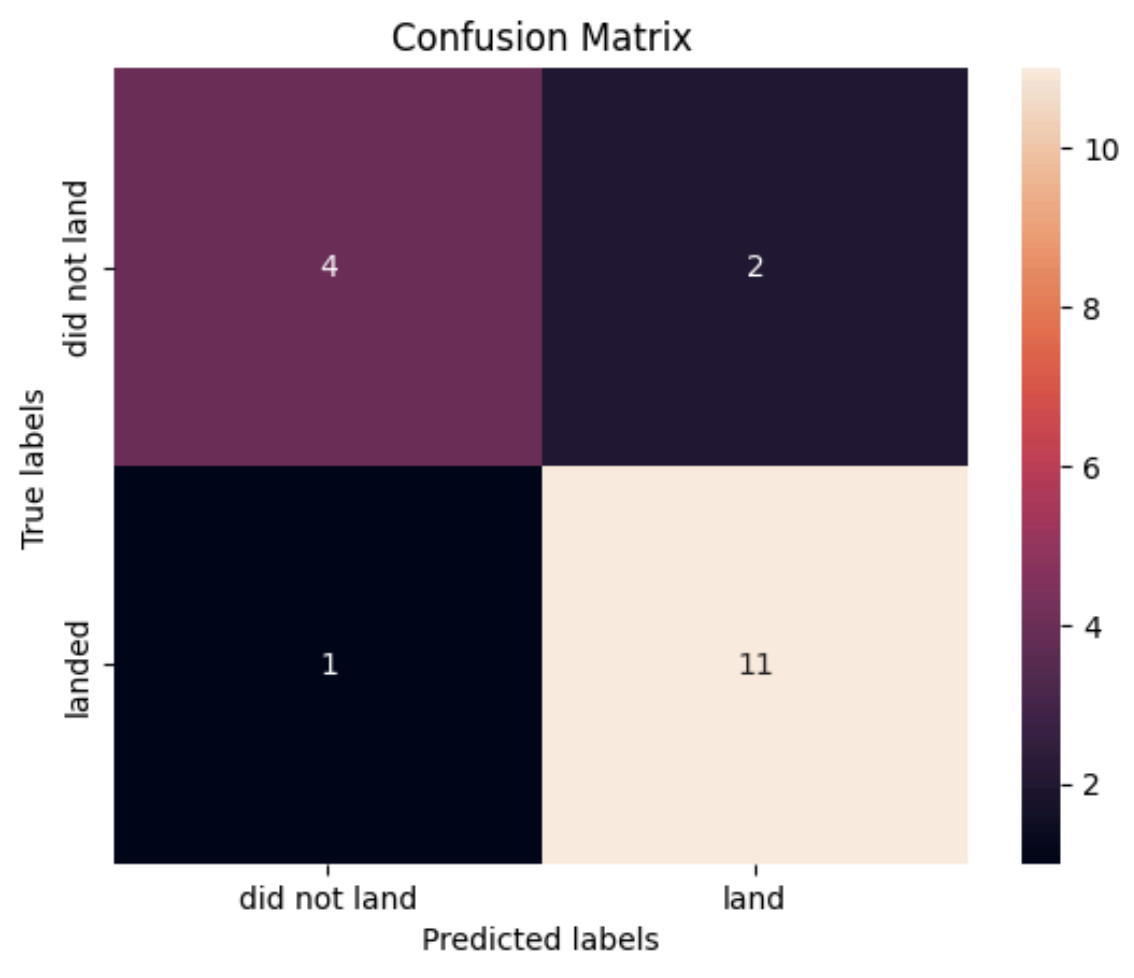
- All models yielded the same accuracy when evaluated on the test set.
- Our training set had 72 observations and the test set only 18.
- We had 82 features.
- Likely there were only a few features that were truly predictive, and all models latched onto them.
- Ideally, need more data to solidify models.



Confusion Matrix of Decision Tree model

- All models have same accuracy, yet decision trees are most interpretable and had fewest false positives (i.e. losing a booster = money).
- The decision tree model had a recall of 11/12 = 91.7% and a precision of 11/13 = 84.6%, i.e. there were 2 false positives and 1 false negative.
- Weighted average F1-score was 83%, same as the overall accuracy.

	precision	recall	f1-score	support
0	0.80	0.67	0.73	6
1	0.85	0.92	0.88	12
accuracy			0.83	18
macro avg	0.82	0.79	0.80	18
weighted avg	0.83	0.83	0.83	18



Conclusions

- Advisable to cooperate with Kennedy Space Center as launch site. Else, find a site with very close connection to a railroad, highway and coast (~1-2 km), and decent distance to residential areas (~15 km).
- Target VLEO orbit for heavy and SSO for light payloads. Avoid GTO orbit missions.
- Use FT and B4 booster builds. Use B5 builds for ultra heavy payloads.
- Landing on ground pads preferred.
- Implemented an interactive dashboard displaying success rates for launch site, booster versions and payload mass.
- Using our Decision Tree prediction model, predict with 83% accuracy if a potential mission will be accompanied by a landing success.
- Our model suggests that having a booster with ”Legs” correlates with landing success (see Appendix).

Appendix

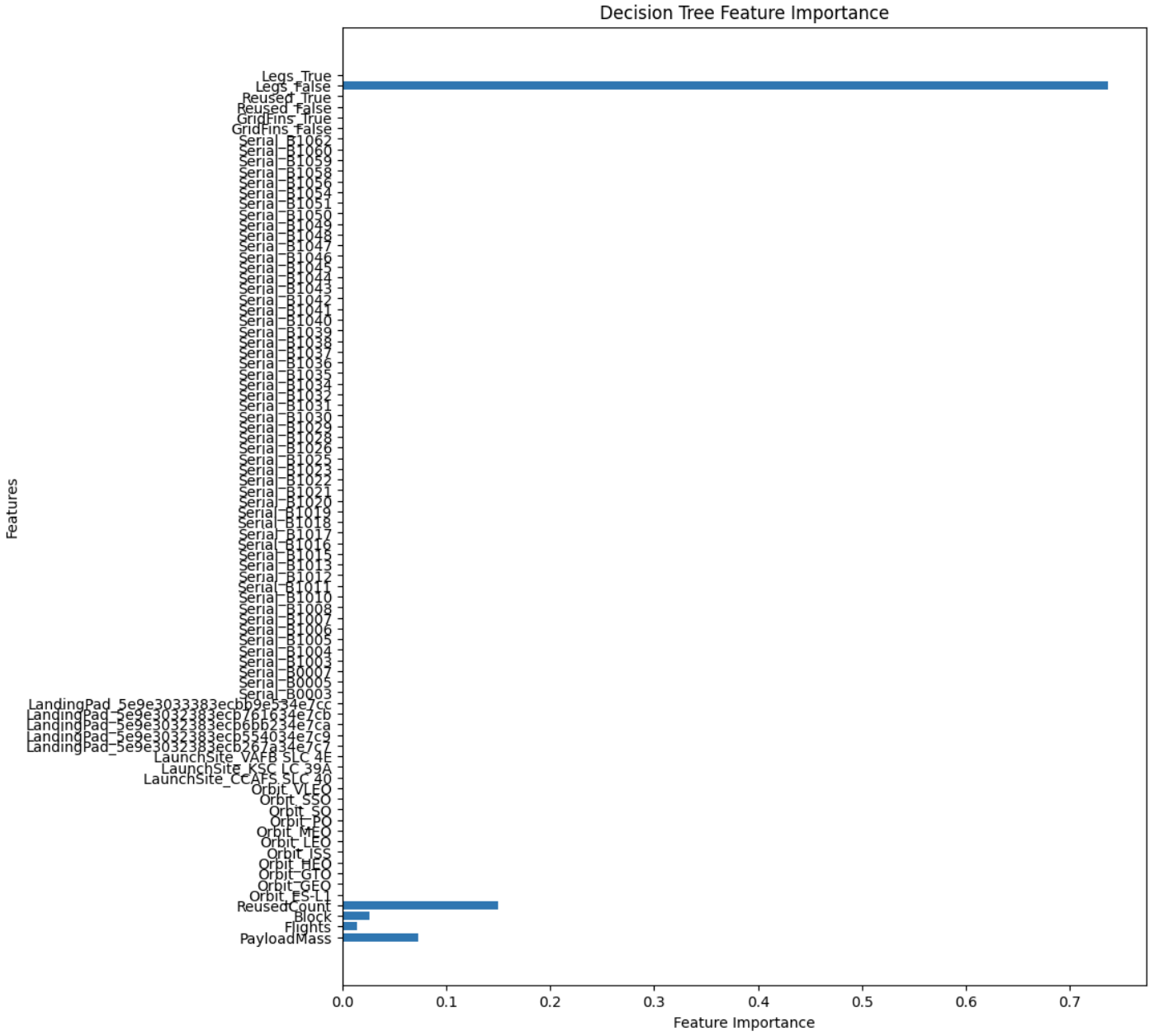
- All SQL queries used for EDA:

```
%load_ext sql
%sql sqlite:///my_data1.db
1. %sql select distinct("Launch_Site") as 'UNIQUE_LAUNCH_SITES' from SPACEXTABLE
2. %sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5
3. %sql select sum(PAYLOAD_MASS__KG_) as "TOTAL PAYLOAD MASS CARRIED BY NASA(CRS) in KG" from SPACEXTABLE where "Customer" like
'NASA (CRS)%'
4. %sql select round(avg(PAYLOAD_MASS__KG_),2) as "AVERAGE PAYLOAD MASS CARRIED BY F9 v1.1 BOOSTERS in KG" from SPACEXTABLE
where "Booster_Version" like 'F9 v1.1%'
5. %sql select min(Date) as "First Ground Pad Landing Success" from SPACEXTABLE where "Landing_Outcome" == 'Success (ground
pad)'
6. %sql select "Booster_Version", PAYLOAD_MASS__KG_, "Landing_Outcome" from SPACEXTABLE where "Landing_Outcome" == 'Success
(drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
7. %sql select \
sum(case when "Mission_Outcome" like '%Success%' then 1 else 0 end) as "Number of successful missions", \
sum(case when "Mission_Outcome" like '%Fail%' then 1 else 0 end) as "Number of failed missions" \
from SPACEXTABLE
8. %sql select distinct "Booster_Version", PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select
MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE) ORDER BY "Booster_Version"
9. %sql select substr("Date", 6, 2) as "Month", "Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where
substr(Date, 0, 5) = "2015" and "Landing_Outcome" == "Failure (drone ship)"
10. %sql select "Landing_Outcome", count(*) as Count from SPACEXTABLE where Date between "2010-06-04 "and "2017-03-20" group by
"Landing_Outcome" order by "Count" desc
```

Appendix

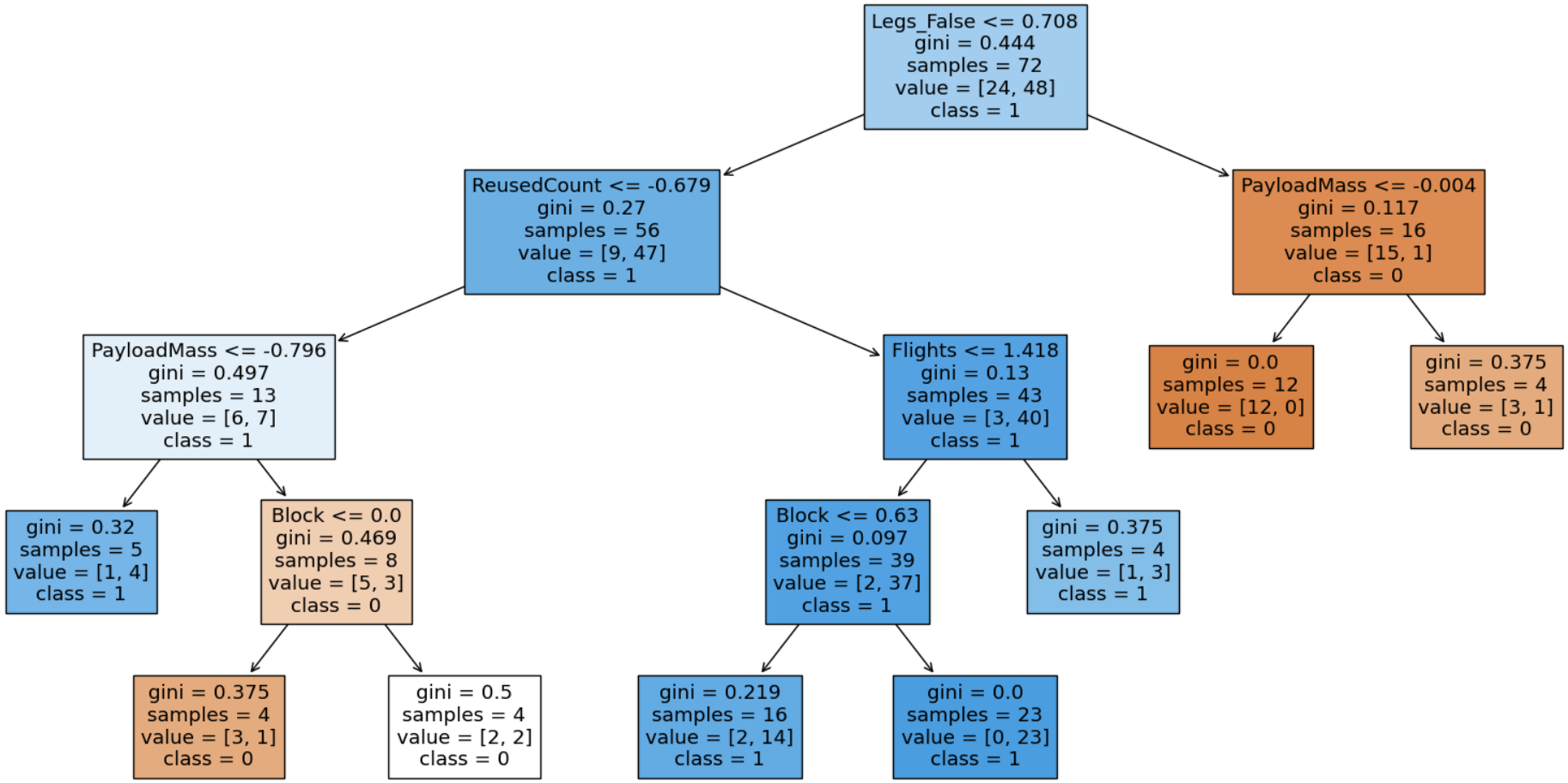
- Feature importance of the Decision Tree model:

‘Legs_False' is the single most important feature contributing to the prediction accuracy of the Decision tree, followed with a large margin by 'ReusedCount', 'PayloadMass', 'Block' and 'Flights’.



Appendix

- The Decision Tree of the Decision Tree model:



Thank you!