IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ibai Mugica
3/4/2024

Executive Summary

- Methods
  - Data about flights is collected from different sources
  - Exploration of data is done through different numerical and visual methods
  - A number of predictive methods are put to the test.
- Results
  - Success rates are steadily growing
  - Larger payload are riskier
  - Predictive modeling is possible and accurate

Introduction

- Economic success of SpaceX as an orbit transport business depends on the reusability of the boosters, and their successful landings
- Identifying what parameters correlate or successful landings can help understand what work and what does not, to have some certainty and be able to operate a bussiness

Section 1

Methodology

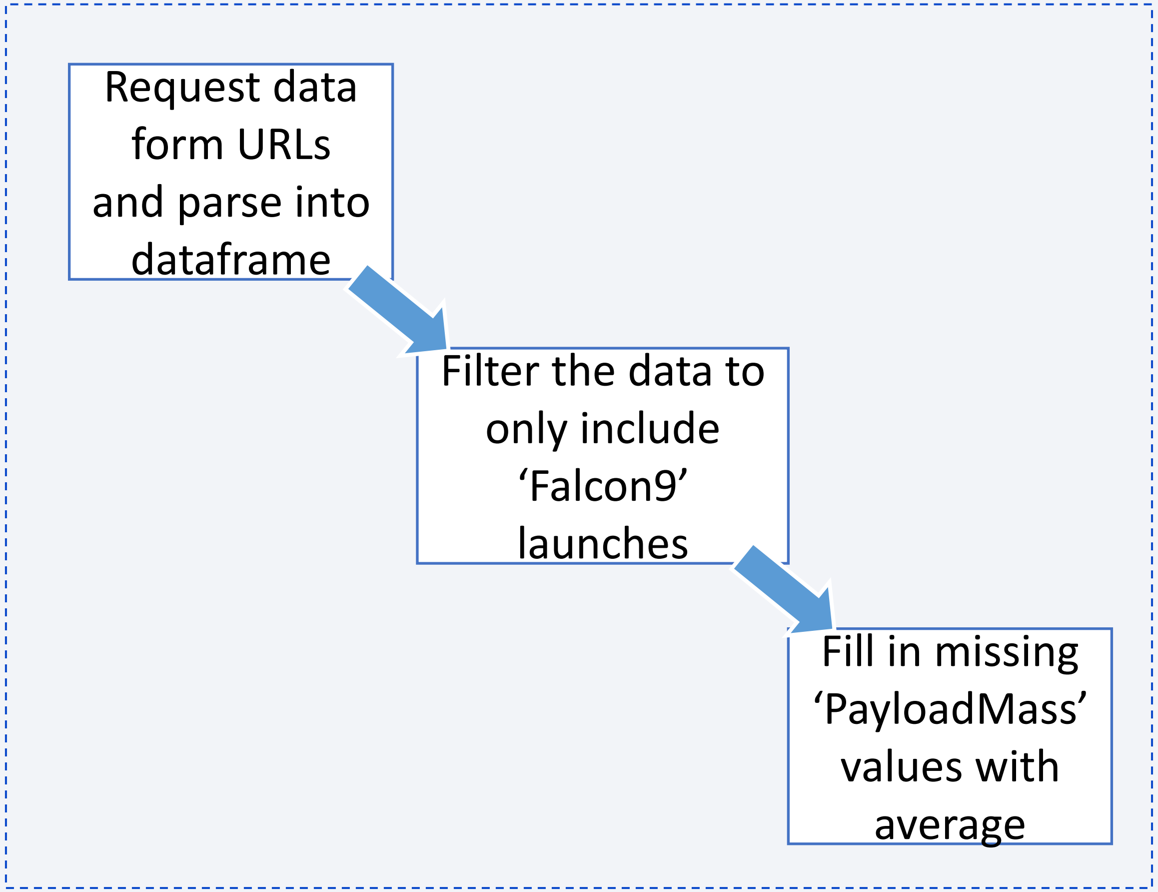
Methodology

Executive Summary

- Data collection methodology:
  - Webscraping
- Perform data wrangling
  - SQL, pandas dataframes
- 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

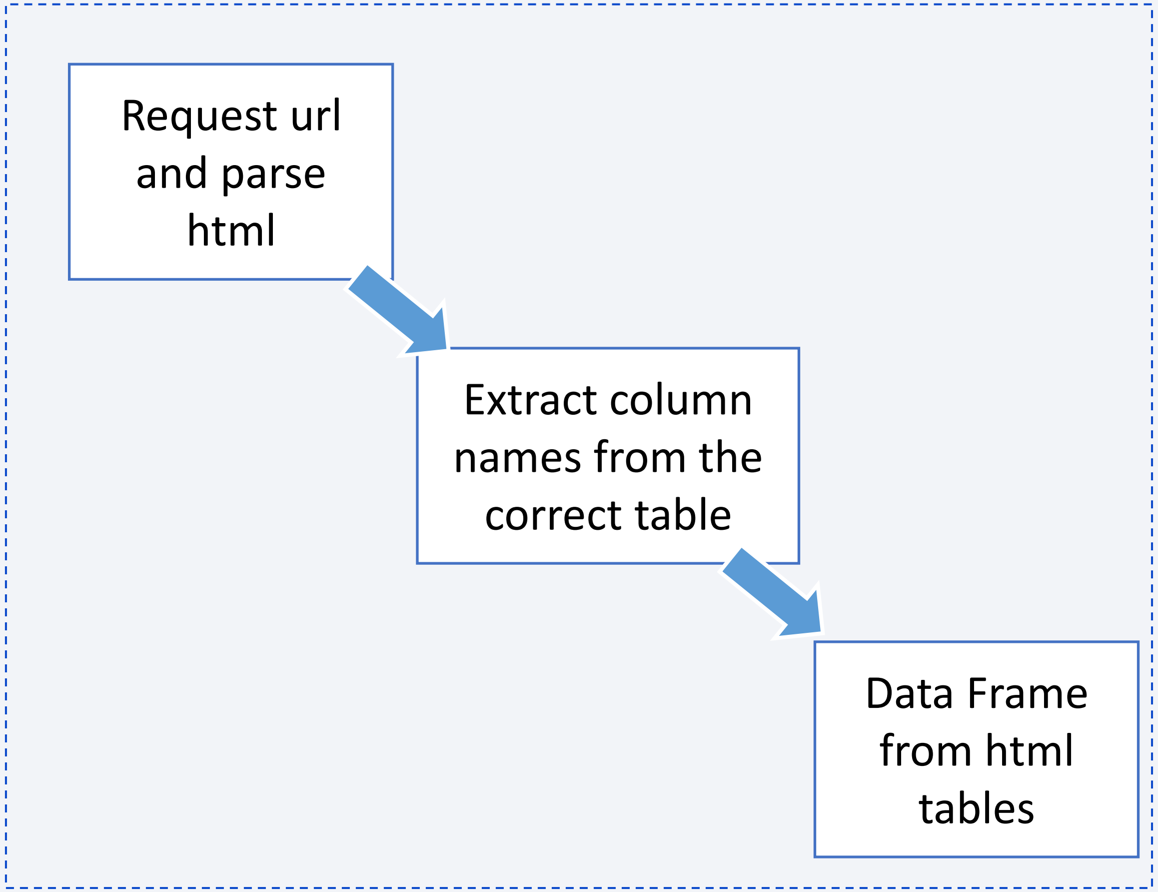
Data Collection – SpaceX API

- The request downloads data from different parts (urls) of the API. Part of it comes in as a JSON dictionary. The dictionary is parsed into a pandas data frame.
- Keep only the ‘Falcon9’ launches from dataframes
- Fill in the missing ‘PayloadMass’ values with the average recorded values.
- https://github.com/imugica/DataScience_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Get more data from launches by scraping the Wikipedia article. Html is parsed with Beautifulsoup
- Extract column names form the correct table to use as variables
- Iterate through the html table rows and copy the data into a pandas data frame
- https://github.com/imugica/DataScience_Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Calculate the number of launches on each site

```
LaunchSite
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
```

- Calculate number of trips to each orbits and types of landing outcome

```
Orbit
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
```

- Create ‘Class’ column with a binary indicator for success

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

- https://github.com/imugica/DataScience_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- The exploratory plots allow to identify the relationship between some of the Flight variables.
- The interesting variables are:
  - Mission Success/Failure
  - Flight number
  - Launch Site
  - Payload Mass
  - Orbit

https://github.com/imugica/DataScience_Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Unique Launch sites
  - Explore CCAFS launch sites
  - Total payload for NASA customer
  - Average payload of F9 v1.1 booster
  - First success in ground pad
  - Mid-payload Boosters that successfully landed on drone ships
  - Total of success and failure landings
  - Booters that carry the max payload
  - Most popular landing outcomes 2010-2017

https://github.com/imugica/DataScience_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers and circles are added to show the launch locations on a world map. Launch locations are labeled
- Markers are green for success and red for failure.
- Cursor with coordinates allows us to calculate the distance between the launch site and interesting landmarks

- https://github.com/imugica/DataScience_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Build a Dashboard with Plotly Dash

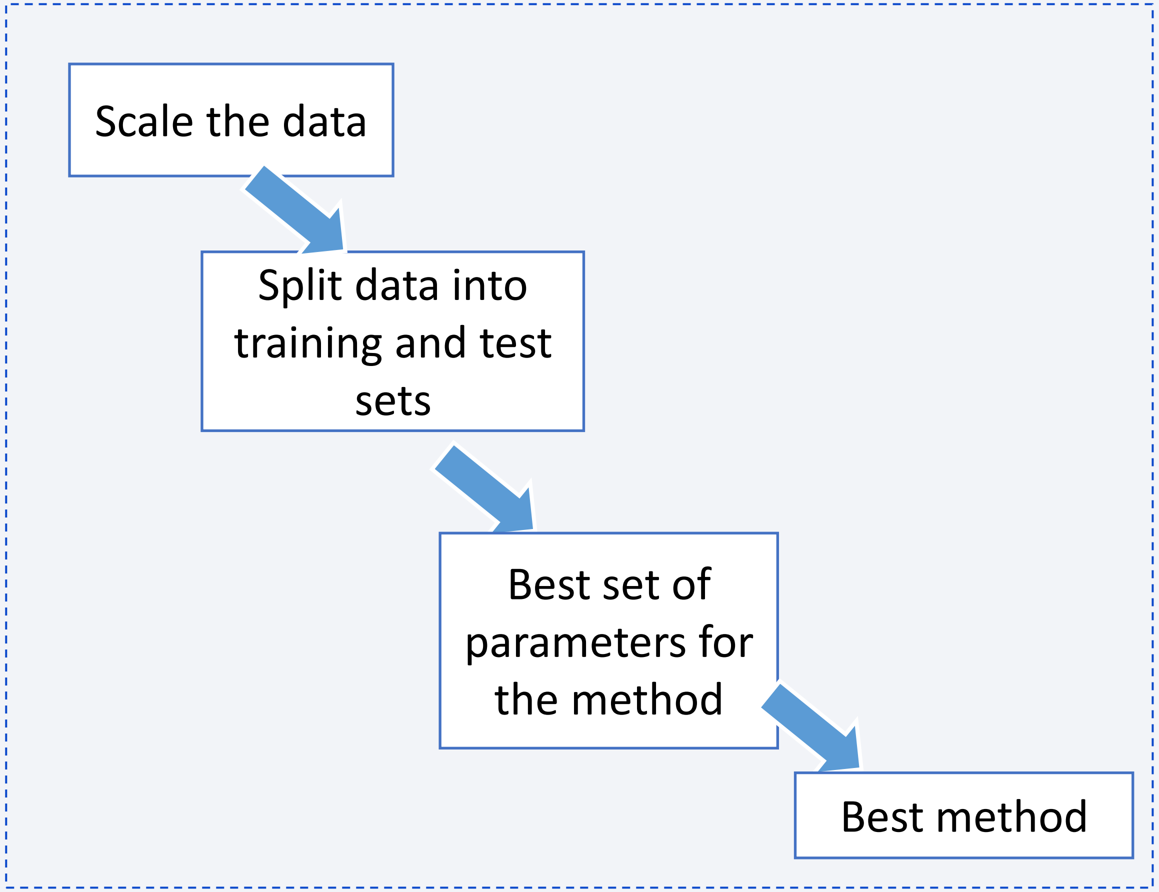
- The interactive pie charts show the relative success of launch sites.
- A dropdown menu changes to show a pie of each site success ratio
- A slider selects the successes and failures of a given payload range.

- https://github.com/imugica/DataScience_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The data is scaled and split into training and test sets.
- The best set of parameters is found
- The accuracy of the methods is evaluated

- https://github.com/imugica/DataScience_Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

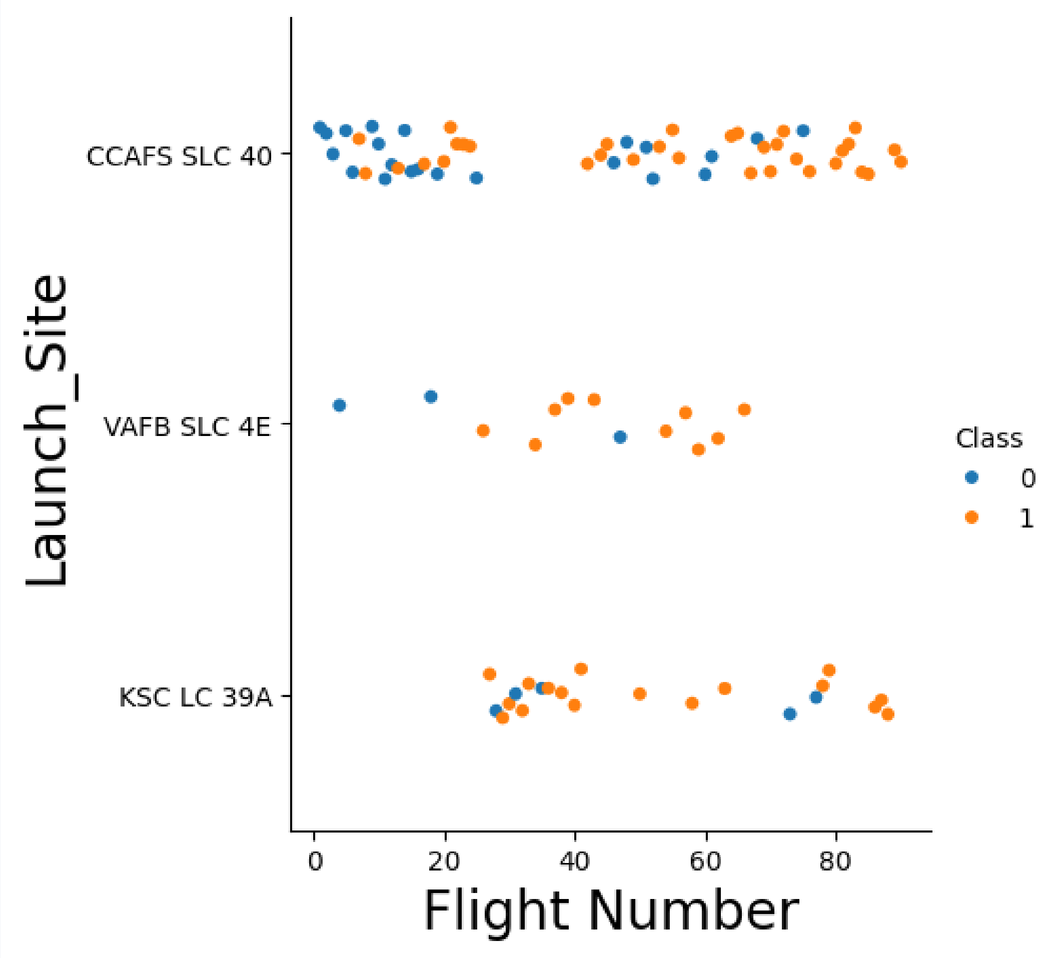
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

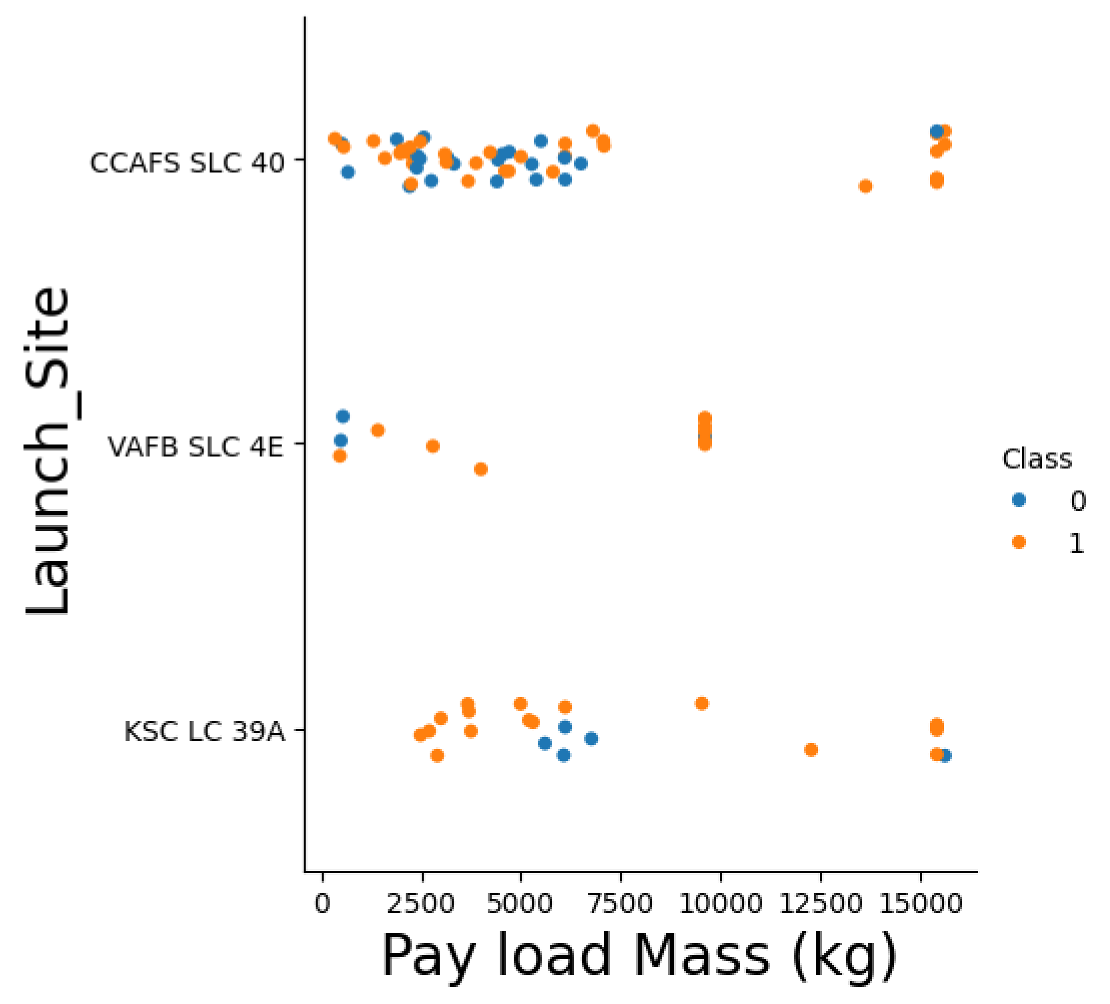
Flight Number vs. Launch Site

- Th plot shows larger failure rate in the *CCAFS SLC 40* site for the first bunch of launches. They briefly stopped using *CCAFS SLC 40* in favor of *VAFB SLC 4E* and *KSC LC 39A*. Success rate improved after that.



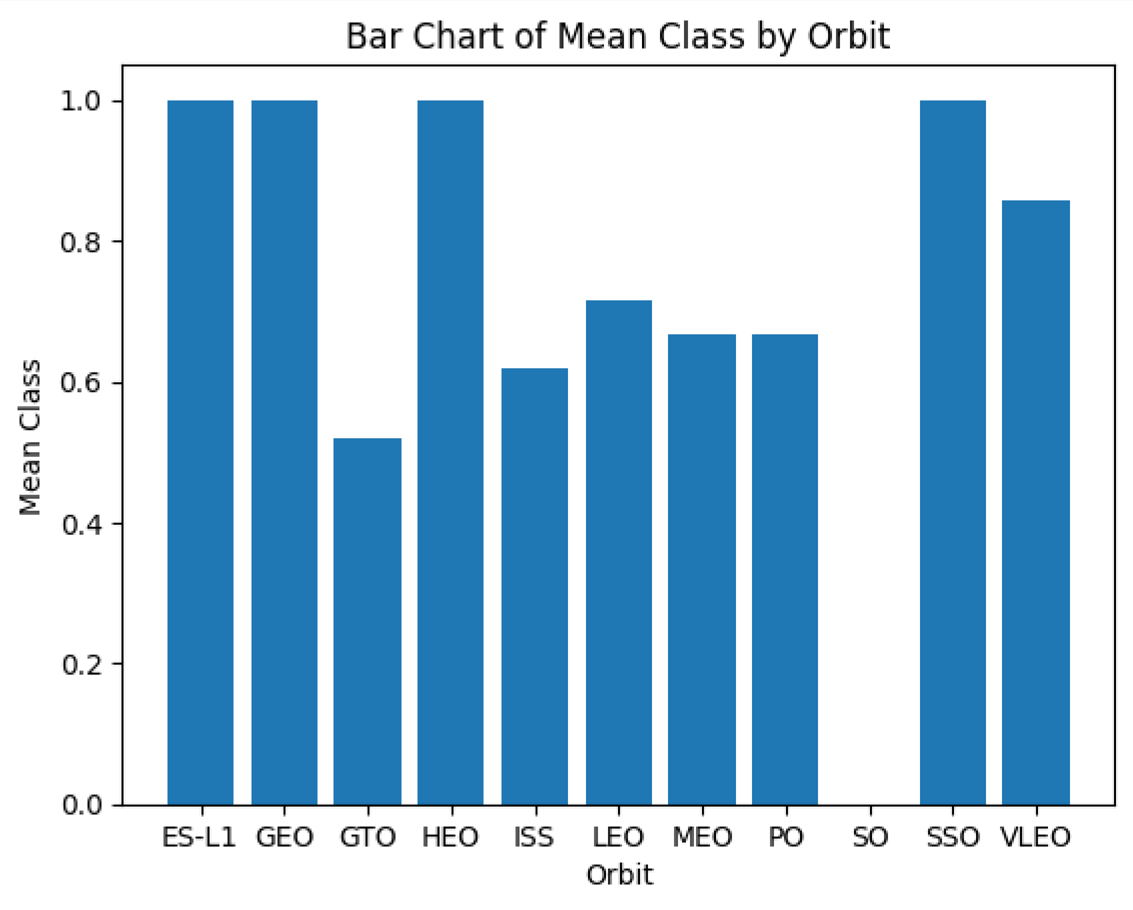
Payload vs. Launch Site

- The plot shows there is no larger launches than 10000Kg in the *VAFB SLC 4E*. Launches less than 7500kg vary much more in Payload mass.
- Most of the failures happened in the *CCAFS SLC 40* site



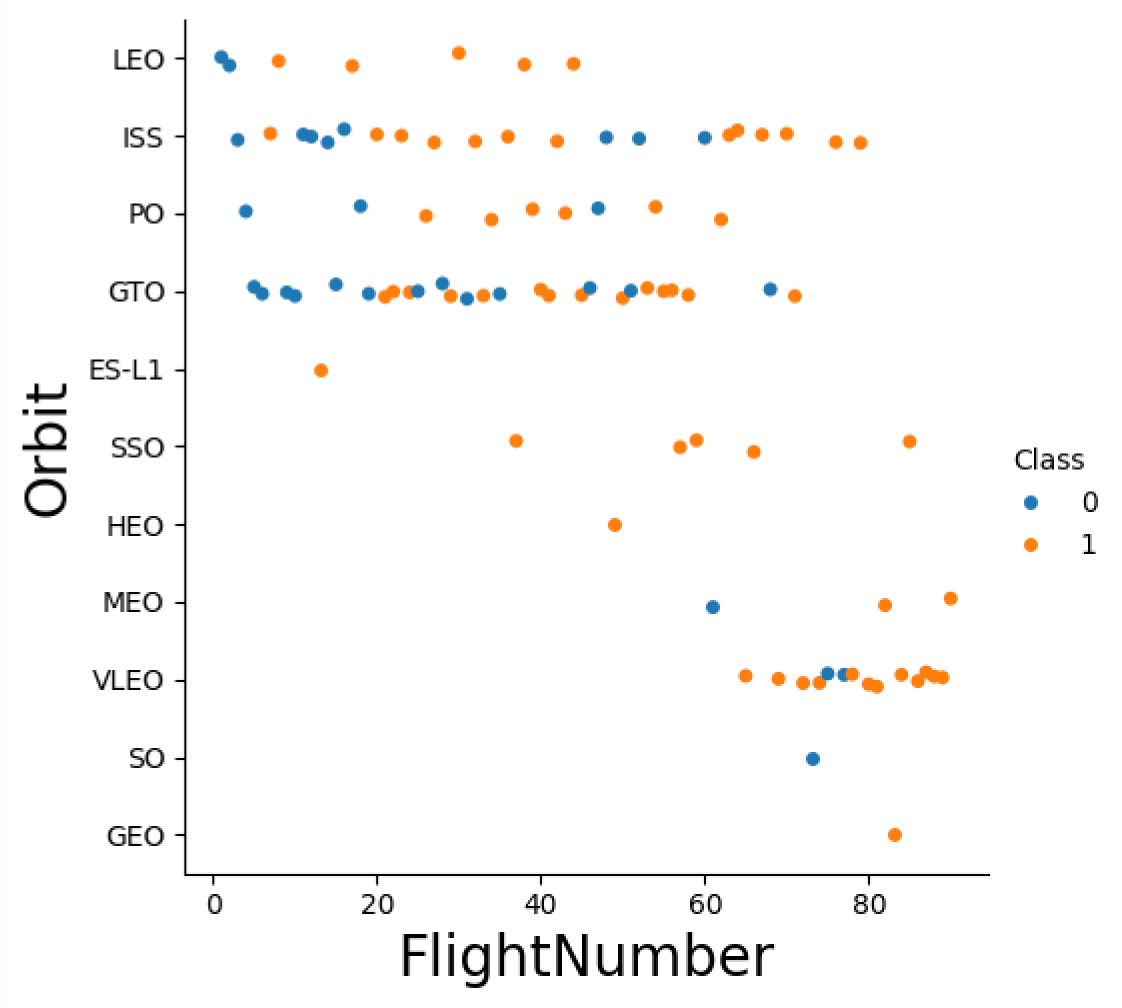
Success Rate vs. Orbit Type

- *ES-L1*, *GEO*, and *SSO* have the highest success rate, whereas *GTO* has the lowest.



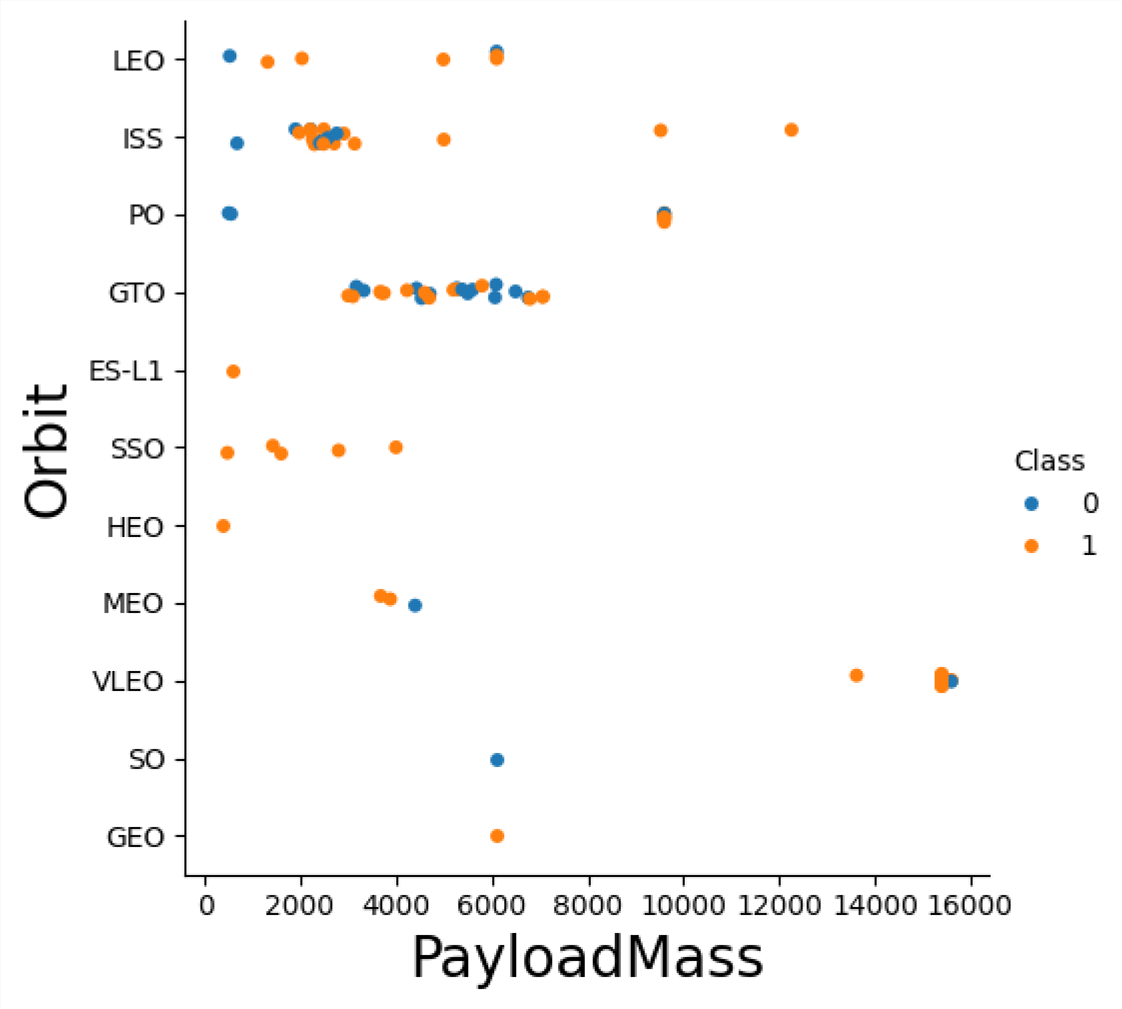
Flight Number vs. Orbit Type

- LEO orbit went from initial failures to having consistent success after a flew flights.
- There is no apparent relation between flight number and GTO orbit



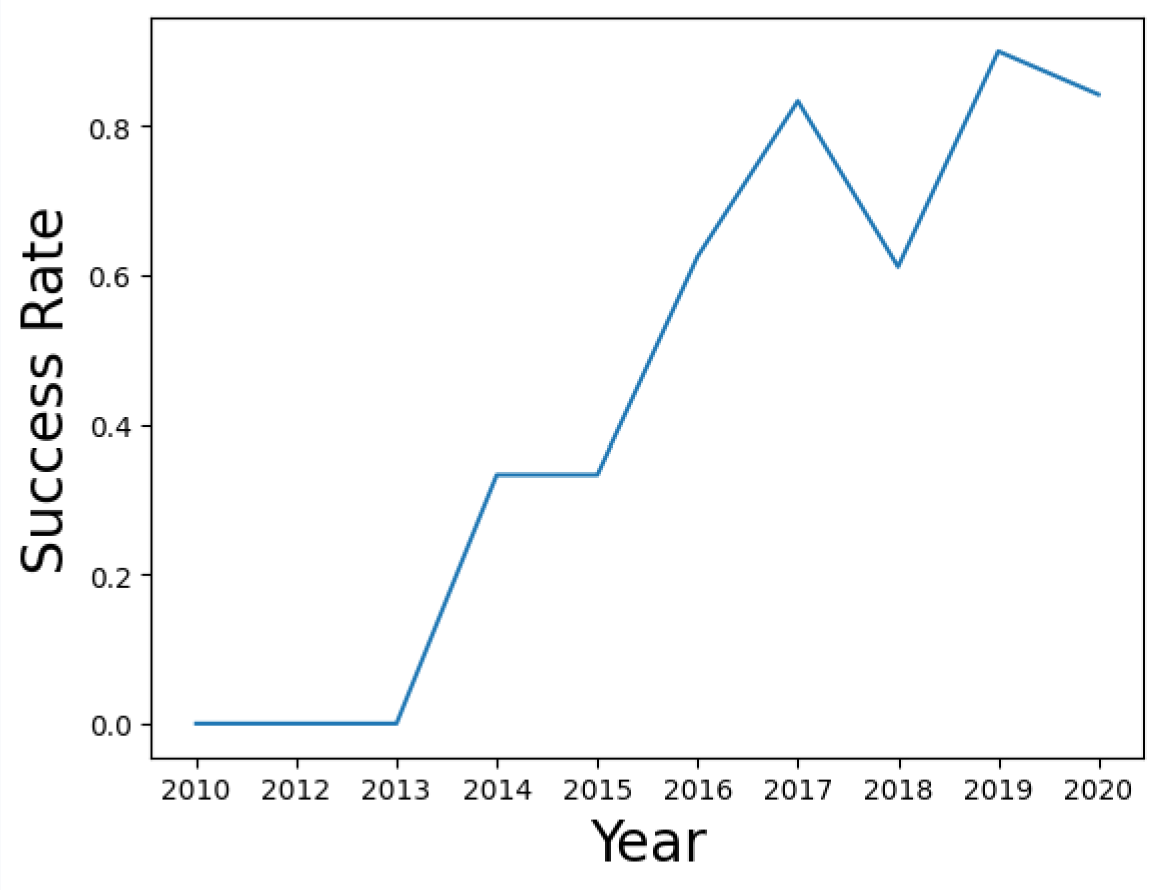
Payload vs. Orbit Type

- LEO, ISS, Polar orbits have better success rate at larger payloads
- SSO, has a good success rate with smaller payloads
- Only heavy launches to VLEO



Launch Success Yearly Trend

- Success rate started increasing in 2013, it stayed the same in 2014. Since 2015 it has been steadily increasing, except for 2018.



All Launch Site Names

- %sql select distinct "Launch_Site" from SPACEXTABLE

- The keyword DISTINCT ensures that only unique values are returned.

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

Launch Site Names Begin with 'CCA'

- %sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5;
- The *Like* keyword can be used to query rows with a given string value (in this case “CCA…” in Launch site)

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like 'NASA (CRS)’

sum(PAYLOAD_MASS__KG_)
45596

- With the *Where* keyword, we can filter the customer (NASA), and the summation function will calculate the total payload mass

Average Payload Mass by F9 v1.1

- %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1’

avg(PAYLOAD_MASS__KG_)
2928.4

- With the *Where* keyword, we can filter the booster version (F9), and the average payload is calculated with a function

First Successful Ground Landing Date

- %sql select min(Date) from SPACEXTABLE where Landing_Outcome like 'Success (ground pad)’

min(Date)
2015-12-22

- The min() function is able to sort the date format from the data and the keyword *Like* is used to identify the successful outcomes

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql select distinct(Booster_Version) from SPACEXTABLE where PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
- The Distinct keyword looks for the unique booster versions, and the where keyword limits the search to the range payload

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- %sql select count(*) from SPACEXTABLE where Landing_Outcome like 'Fail%' or Landing_Outcome like 'Succ%’;

count(*)
71

- *Count* keyword returns the number of instances and the *like* keyword narrows the search to the successful and failure missions

Boosters Carried Maximum Payload

- %sql select distinct (Booster_Version) from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);

- We use a subquery to filter the rows with max payload. The outer query keeps the unique booster version names.

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

2015 Launch Records

- %sql select substr(Date,6,2) as Month, Booster_Version , Launch_Site from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome like 'Fail%';
- The *Where* keyword filters the 2015 year rows by looking at the Date column “and” the launches that start with the ‘Fail’ string.

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

- %sql select Landing_Outcome, count(*) as Outcome_Count from SPACEXTABLE where Date > '2010-06-04' and Date < '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;

- The *Where* keyword uses the 'Date' column to filter the time period. The *as* keyword allows *order by* to sort in descending order.

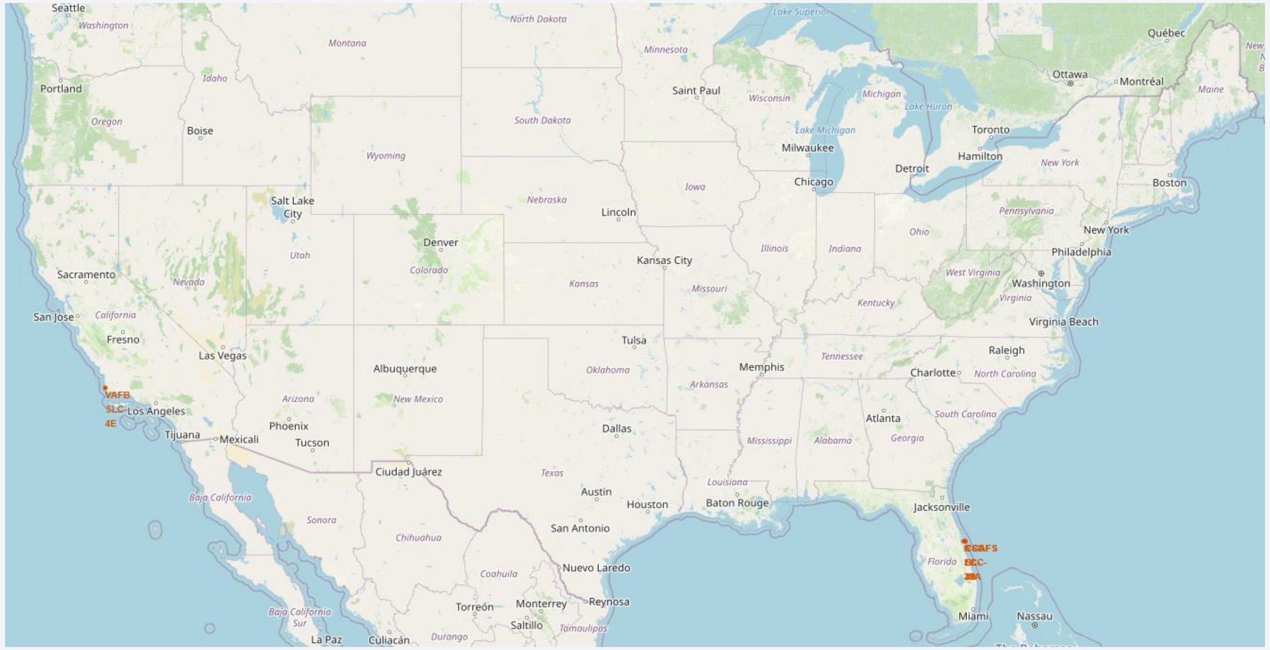
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

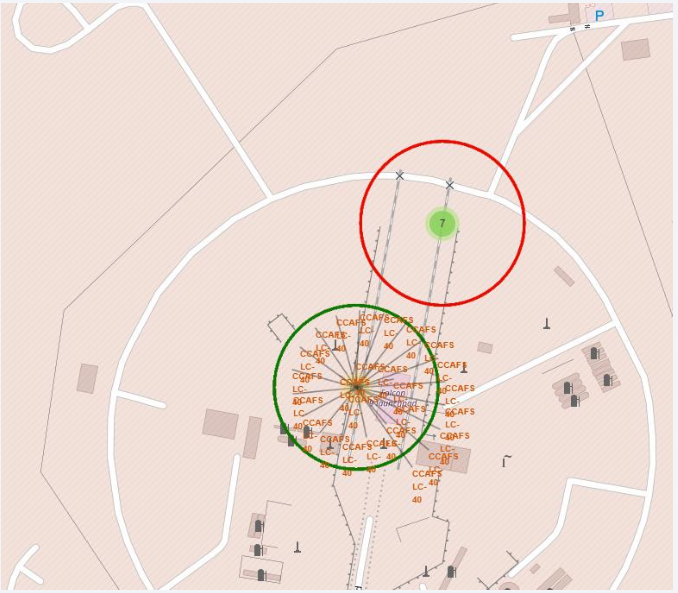
Launch Sites

- Filtered the database to keep unique Launch sites with coordinate columns
- We iterate through the list to add markers with the launch site name on the map
- There are three launch sites near Orlando FL (*CCAFS LC-40, CCAFS SLC-40, KSC LC-39A*), and another one (*VAFB SLC-4E*) near L.A. CA.



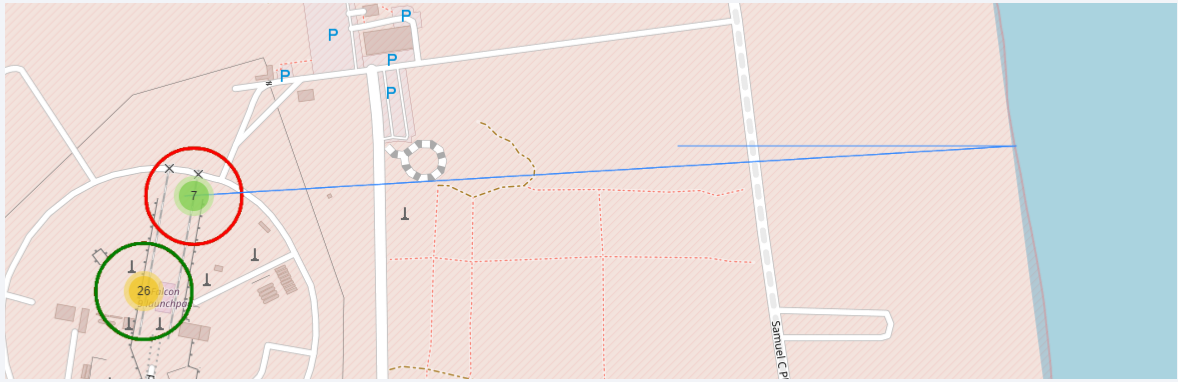
Local Success/Failure

- Add markers with launch site labels. They are colored green for success and red for failure
- Navigating we can see how some launch sites are more successful than others.



Proximity to coast

- The easternmost launch sites are 0.9Km away from the coast

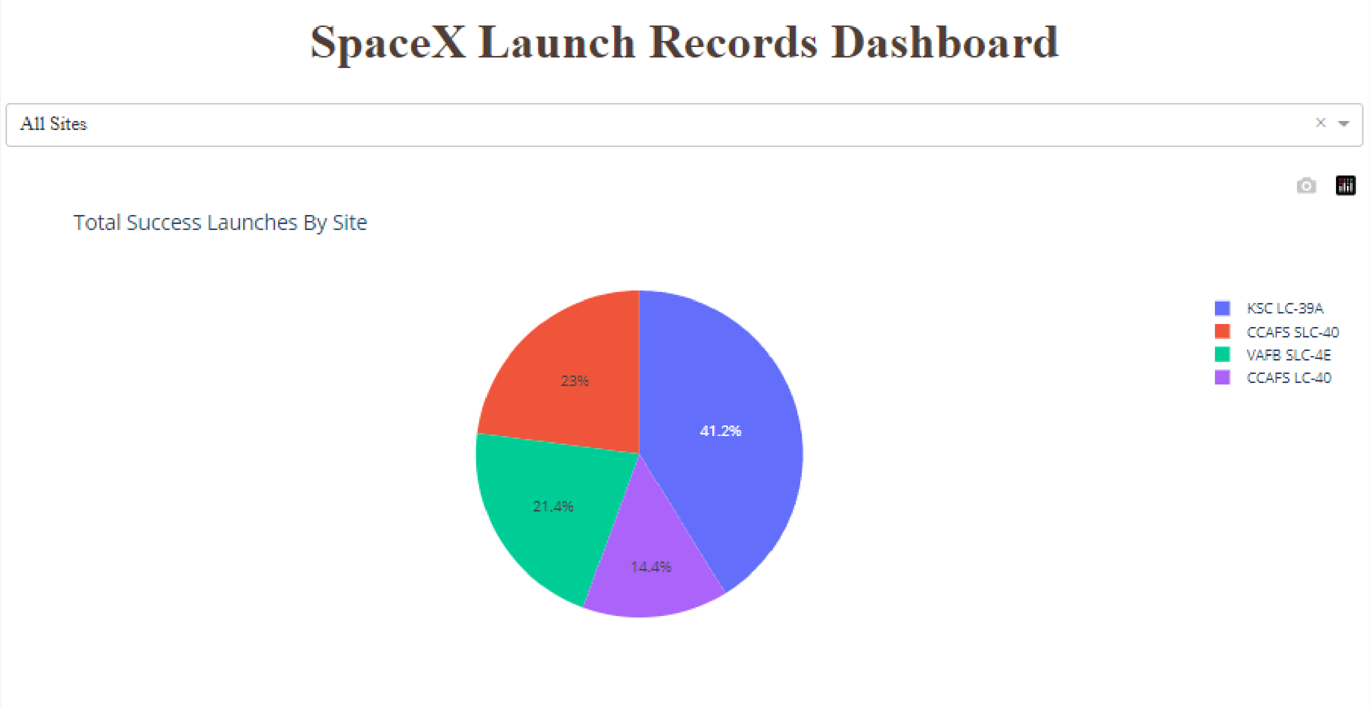


Section 4

Build a Dashboard with Plotly Dash

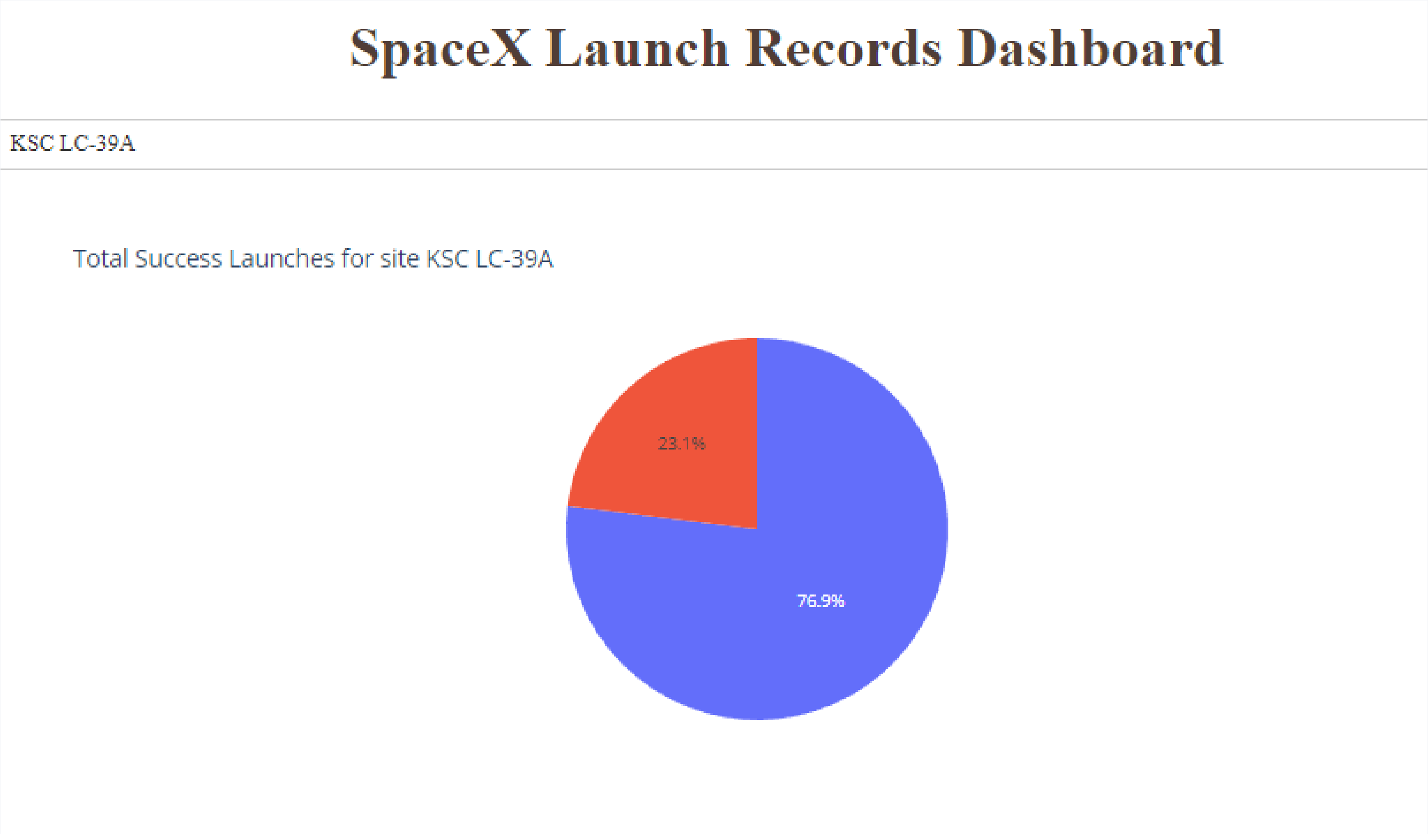
Success Launches by site

- The pie chart shows how the *KSC LC 39A* is the most successful launch site



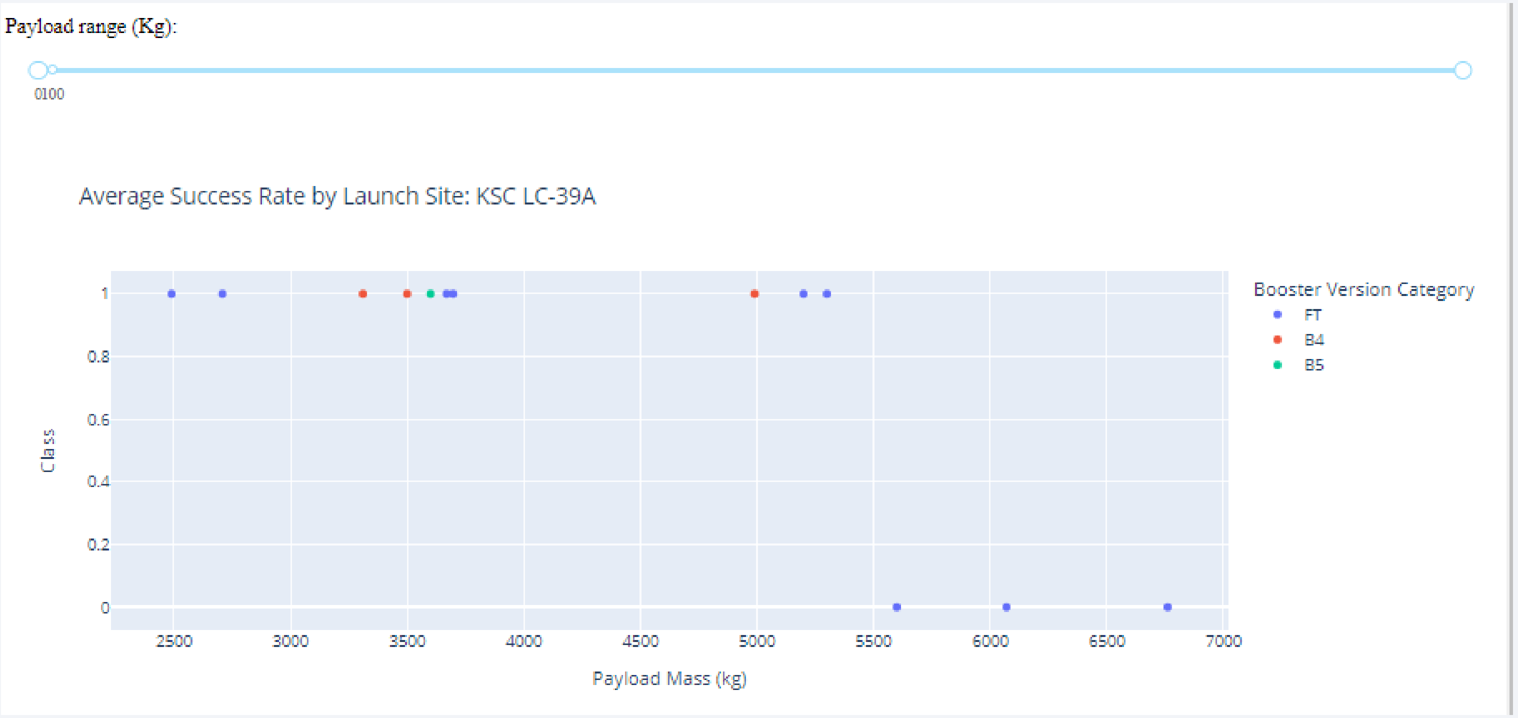
KSC LC 39A Success rate

- The Success ratio of KSC LC 39A is 76.9%



Success of different Payloads

- All failures happen for higher payloads. Logically, sending a heavier load to orbit is less safe.

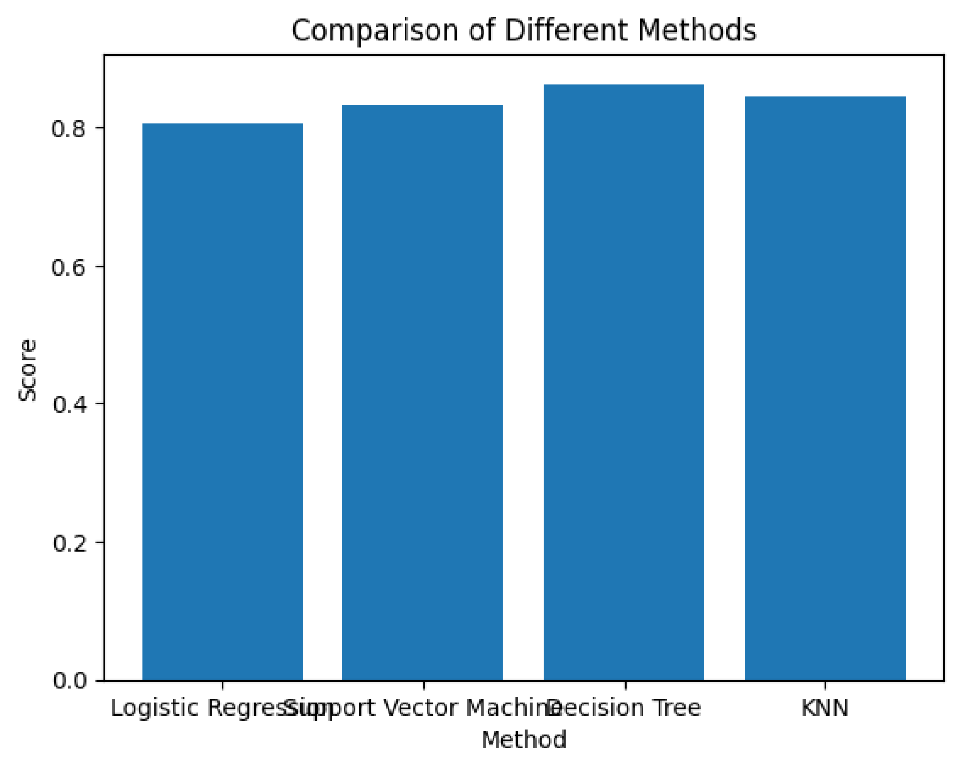


Section 5

Predictive Analysis (Classification)

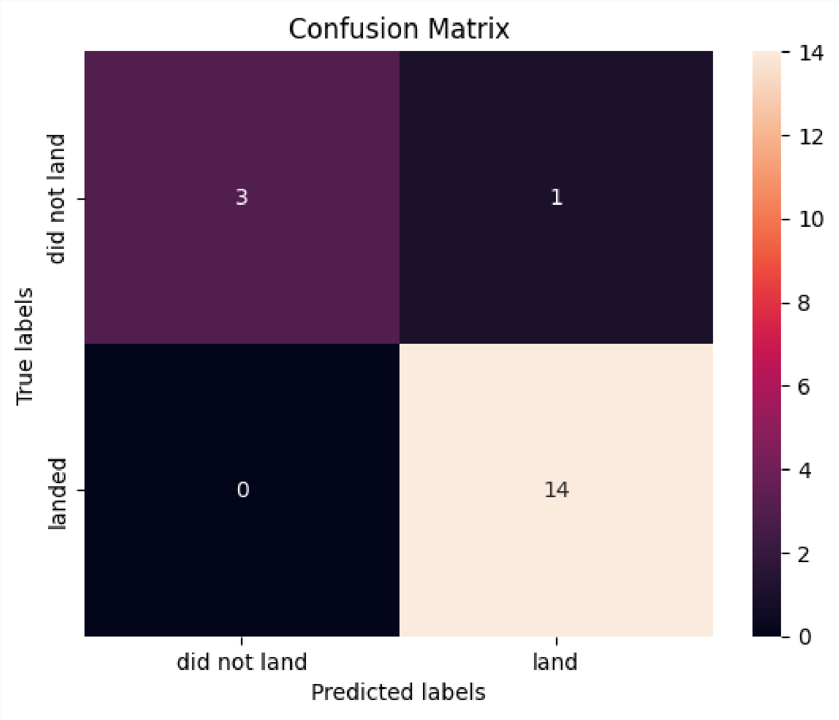
Classification Accuracy

- The highest classification accuracy is for the Decision Tree.
- It could be because it is very god at handling no numerical data.



Confusion Matrix

- The Only mistake of the decision tree, is a False Positive:
  - The rocket did not land, but it was predicted to land.



Conclusions

- Success rate is steadily increasing on a 2-3 year basis.
- Success over payload showed that higher payloads are more risky launches.
- We can build a predictive model, and it is most accurate with a decision tree.

Thank you!