IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Linh Mai

11 July 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Data is collected from online sources and wrangled to determine training labels. Visualization and SQL is applied to perform Exploratory Data Analysis (EDA). More insights are highlighted in interactive visual analytics using Folium and Plotly Dash. Classification models are employed in predictive analysis.
- EDA analysis shows that factors including time, launch site, booster versions, payload mass, orbit type, landing type having impacts on a success launch. Interactive tools highlight the success ratio for each launch site and their geographical patterns. All predictive model perform equally in term of accuracy.

Introduction

- Context:
  - SpaceX advertises Falcon 9 rocket launches with a much lower cost than competitors, thanks to its ability to reuse the first stage. If we can determine if the first stage will land, we can determine the cost of a launch and the probability that SpaceX will reuse the first stage.
- Objectives:
  - predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

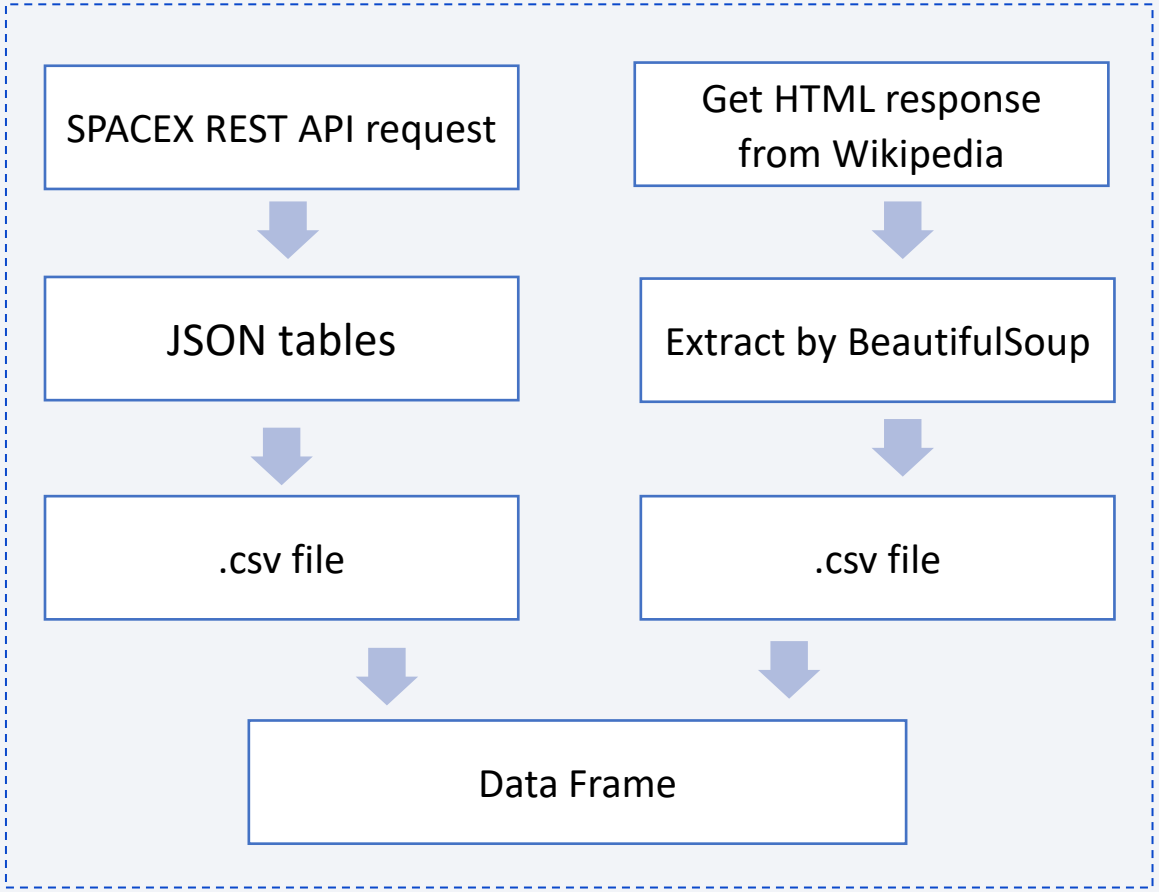
Executive Summary

- Data collection methodology:
  - Requesting data from API and web-scrapping from Wikipedia.
- Perform data wrangling
  - Cleaning and standardizing data before determining training labels.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four classification models have been built using training/testing sample and evaluated to determine the best performer.

Data Collection

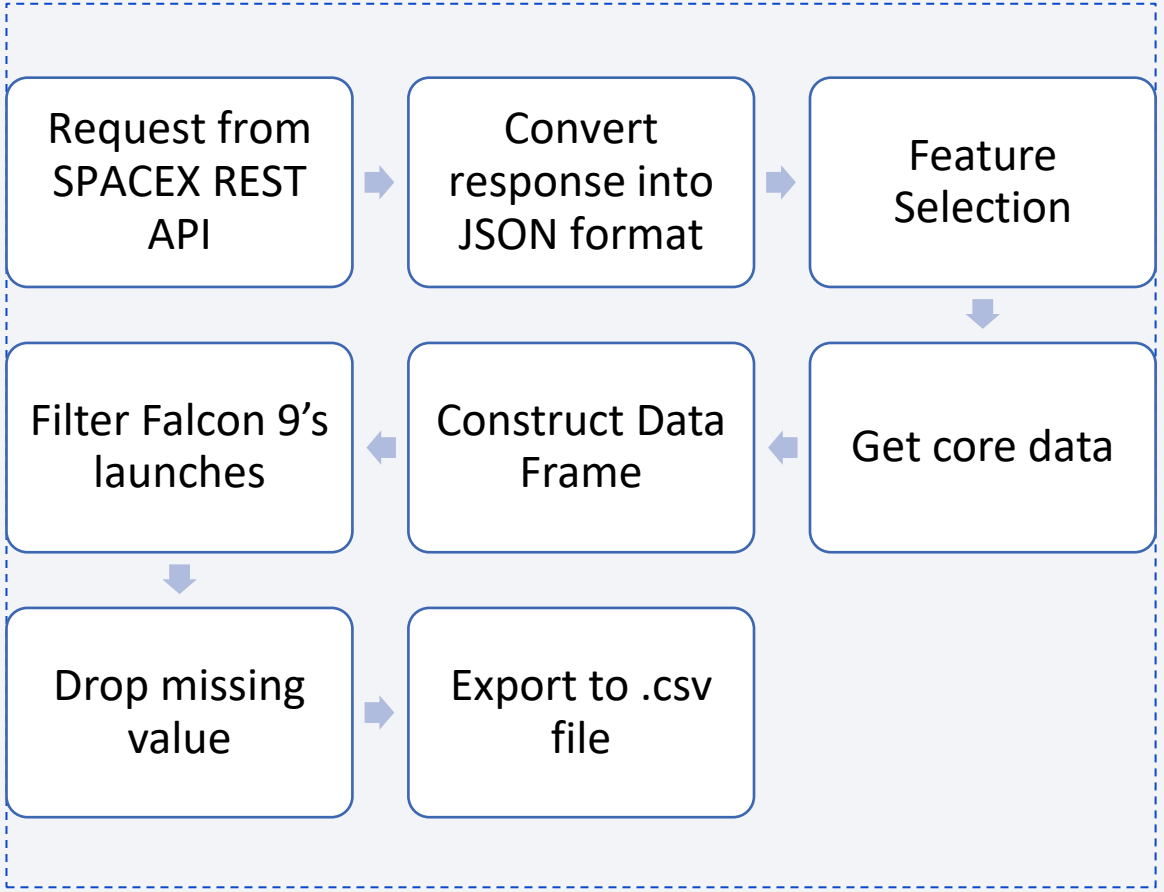
Launch data is gathered from the SpaceX REST API and web scrapped from Wikipedia

Collected data includes: time, sites, booster, launch and landing specifications, mission and outcome.



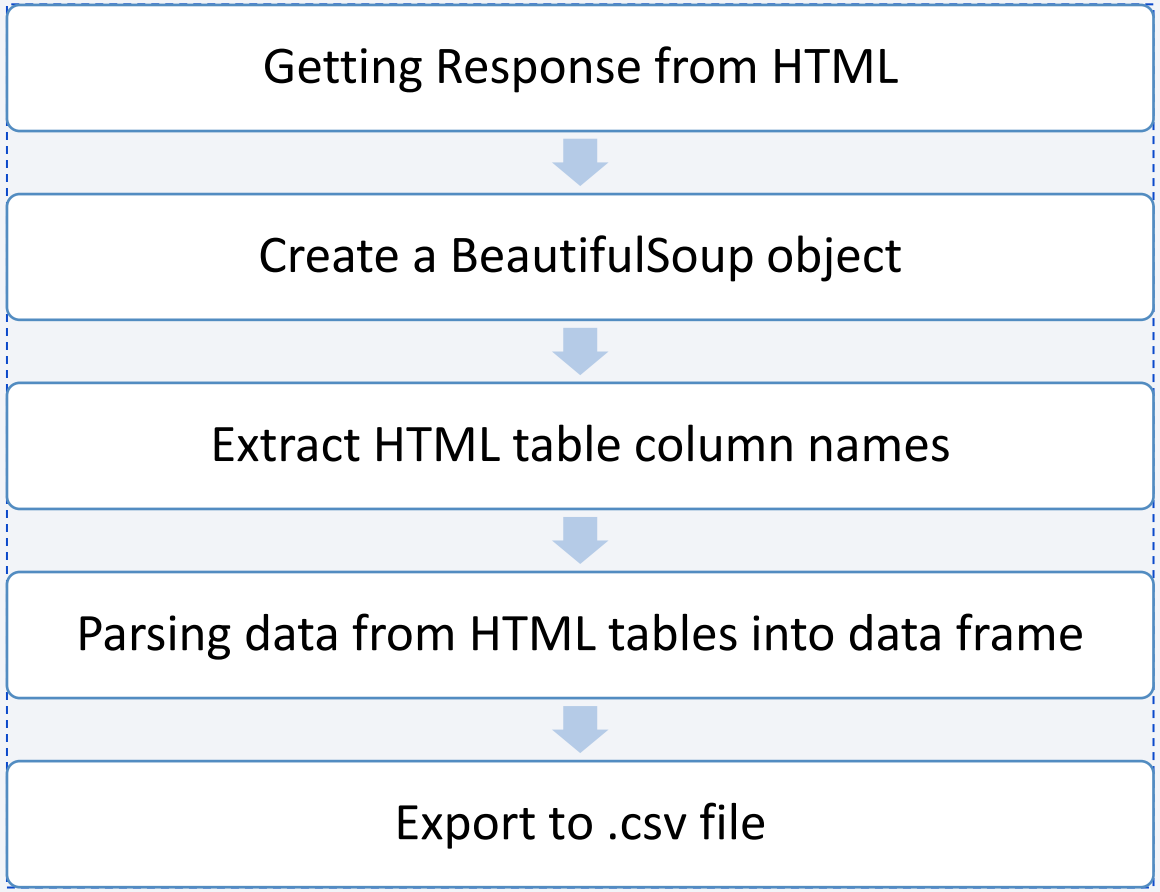
Data Collection – SpaceX API

- Data is request from the SpaceX API and then normalized
- Notebook



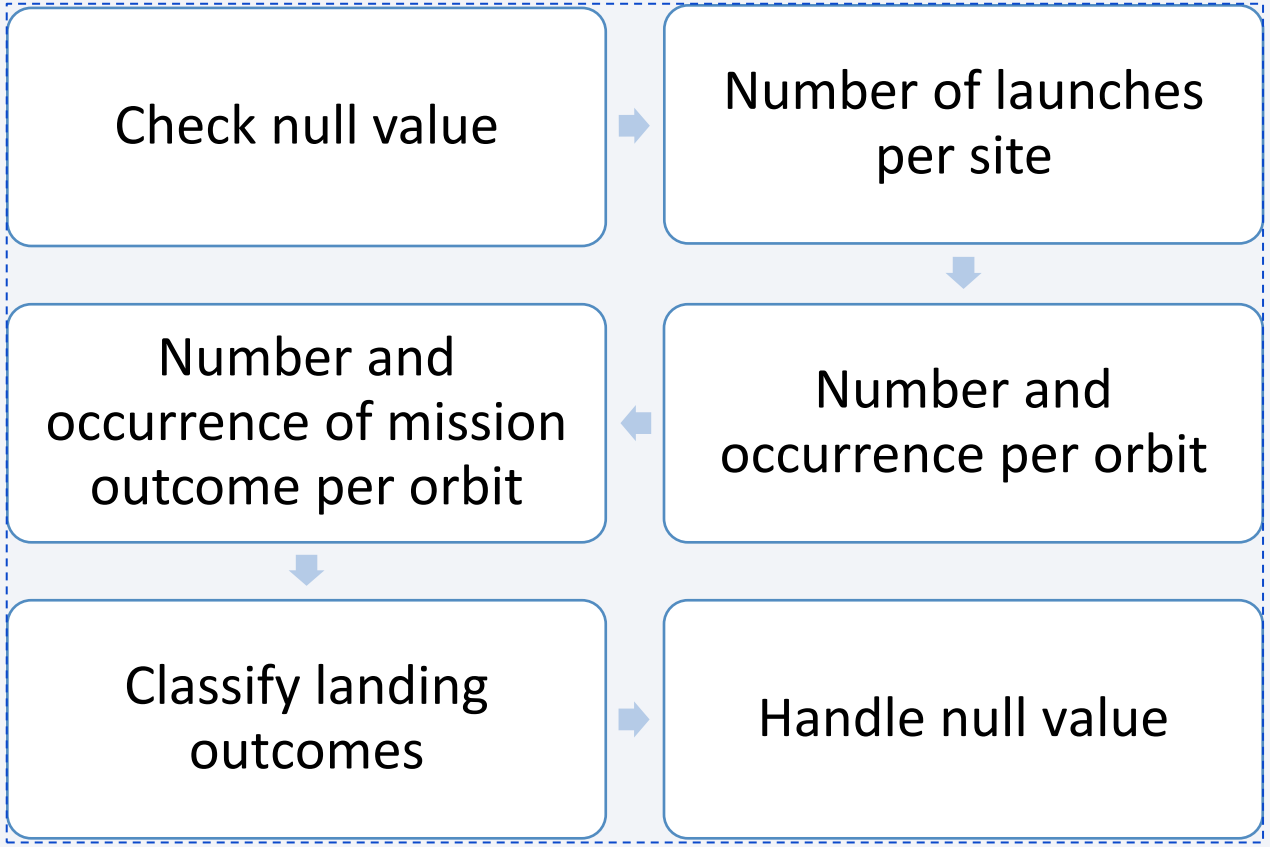
Data Collection - Scraping

- Falcon 9 launch records in HTML tables are extracted from Wikipedia 'Falcon9 Launches' page, then convert to data frame.
- Notebook



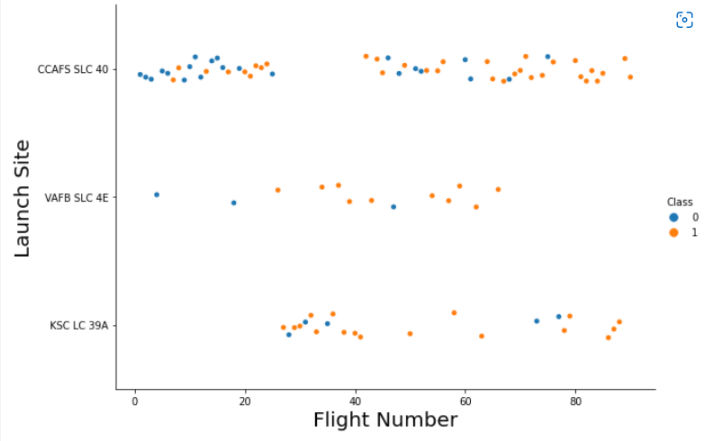
Data Wrangling

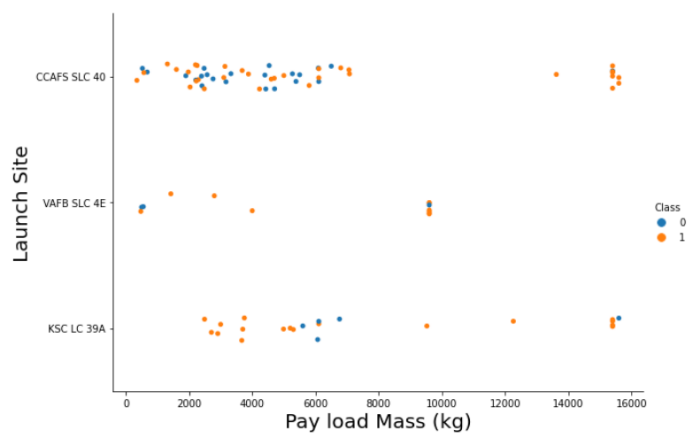
- Relevant data is transformed and analyzed.
- Training labels are determined to be landing outcome class variable.
- Notebook

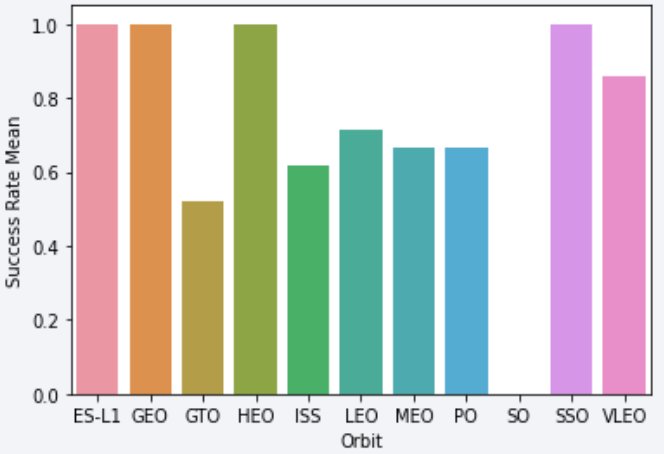


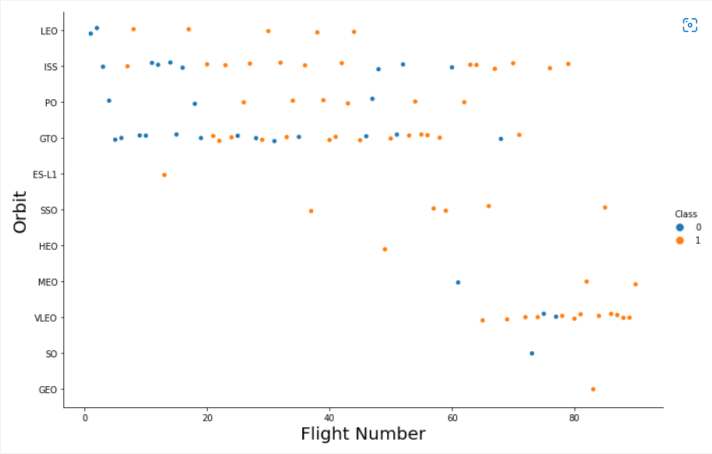
EDA with Data Visualization

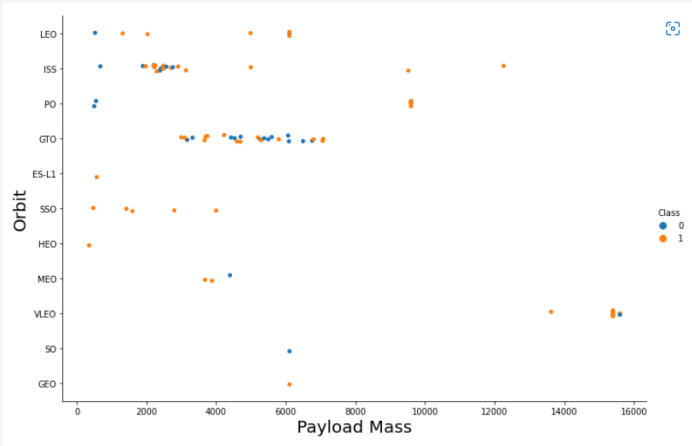
Notebook

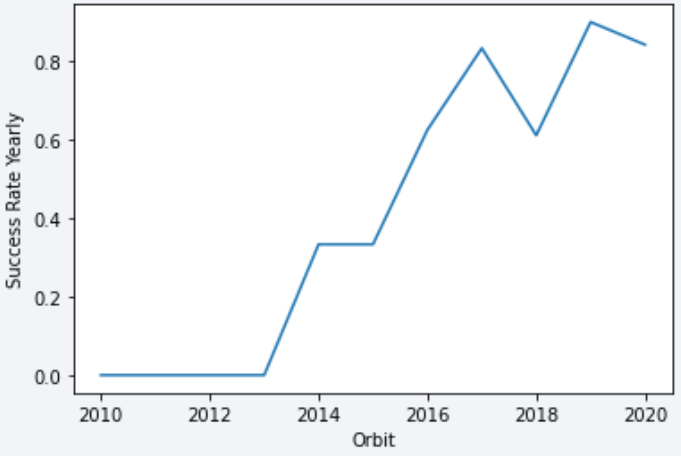












EDA with SQL

Notebook

SQL queries are to get insights from the dataset, including:

- Launch sites lists
- Records where launch sites begin with 'CCA'
- Total payload mass carried by NASA's boosters
- Average payload mass carried by booster version F9 v1.1
- Date of the first successful landing outcome in ground pad
- Boosters which have success in drone ship, 4000 < payload mass < 6000
- Total number of successful and failure mission outcomes
- Booster versions which have carried the maximum payload mass.
- Failure landings in drone ship in year 2015.
- Successful landing outcomes from 04-06-2010 to 20-03-2017.

Build an Interactive Map with Folium

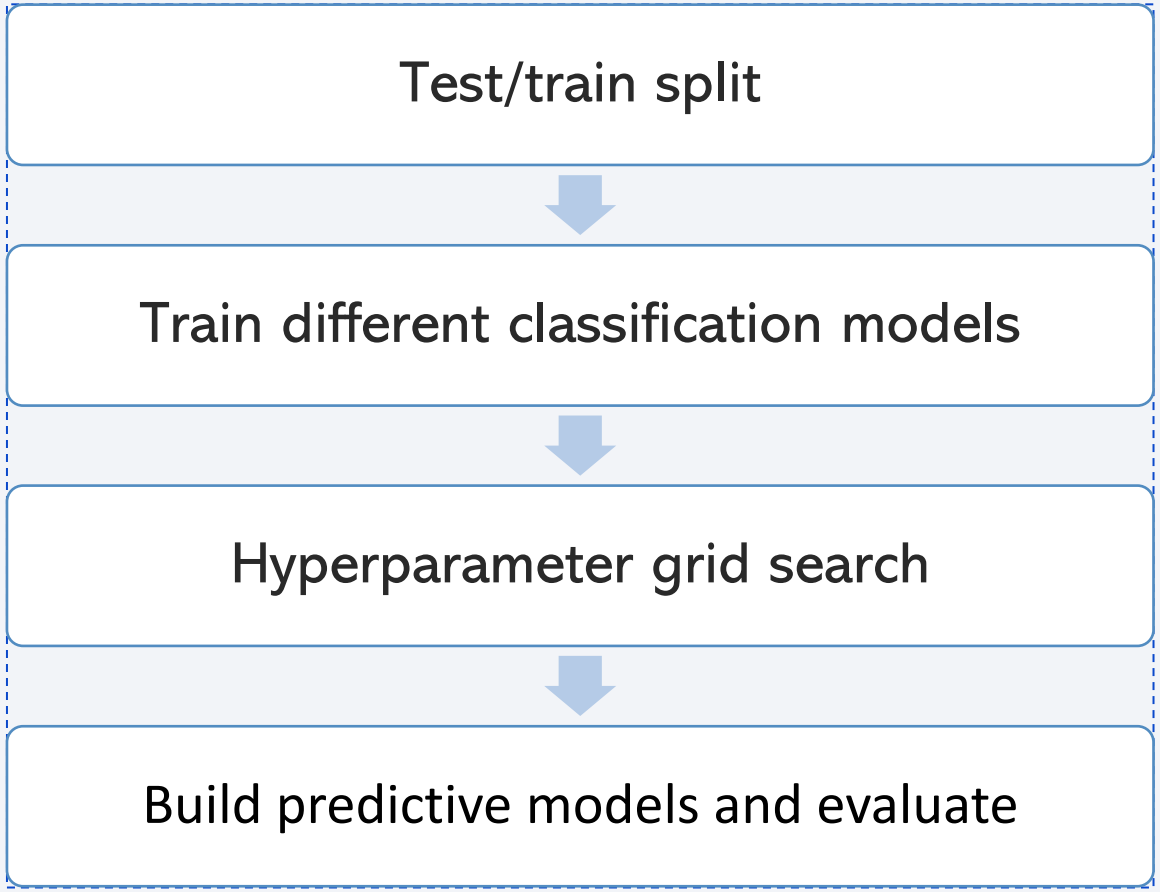
- All launch sites are marked on a map, with circles and labels popup sites' names.
- Green/red markers are added indicating the success/failed launches for each site.
- Distances and polylines between a launch site to its proximities are included.
- These tools are aim to find some geographical patterns about launch sites and their impacts on launch success rate, such as their location, proximity to Equator line and the coastline, success rates for each site, the proximity to transport facilities and urban areas.
- Notebook

Build a Dashboard with Plotly Dash

- Pie charts of total success launches for all sites and for each site, with a dropdown list to choose displaying options.
- Scatter point charts showing the correlation between payload and success for all sites with a range slider of payload mass from 0 to 10000 kg.
- These plots and interactions might help to get insights from the dataset more easily than with static graphs.
- Plotly Dash lab

Predictive Analysis (Classification)

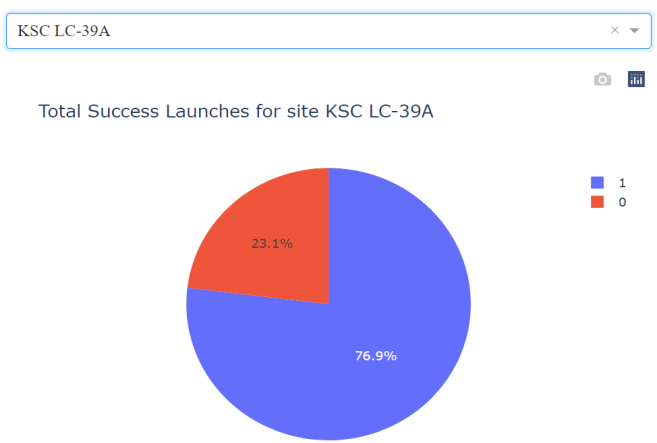
- We consider four methods: Logistic Regression, Support Vector Machine, K Nearest Neighbor, Tree Classifications.
- Predictive models are analyzed through confusion matrix and evaluated by accuracy score.
- Notebook



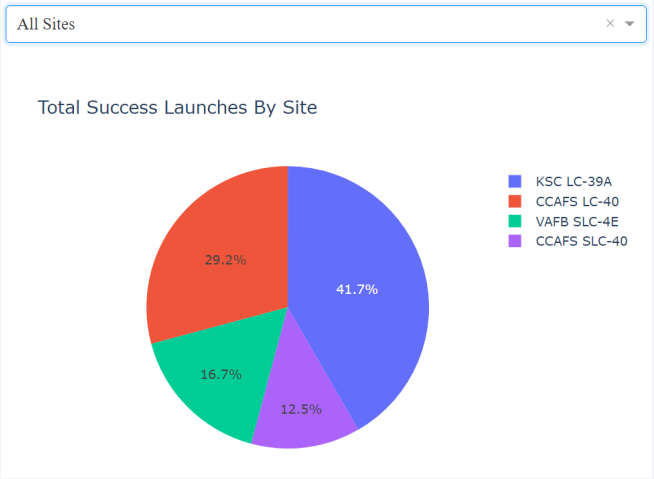
Results

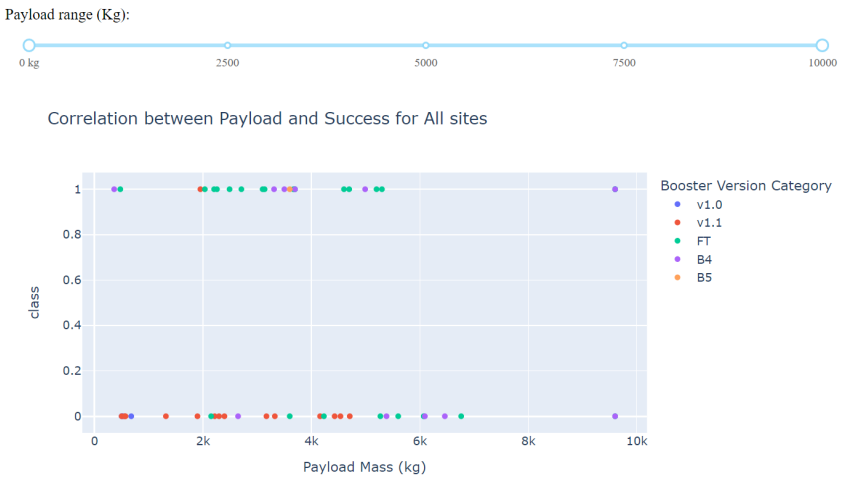
- Factors having impacts on success launching include time, launch site, booster versions, payload mass, orbit type, landing type.
- The optimal launch site is in proximity to Equator line, coastline, transport facilities, and within a certain distance from urban areas.

Results


KSC LC-39A
Total Success Launches for site KSC LC-39A
1
0
23.1%
76.9%

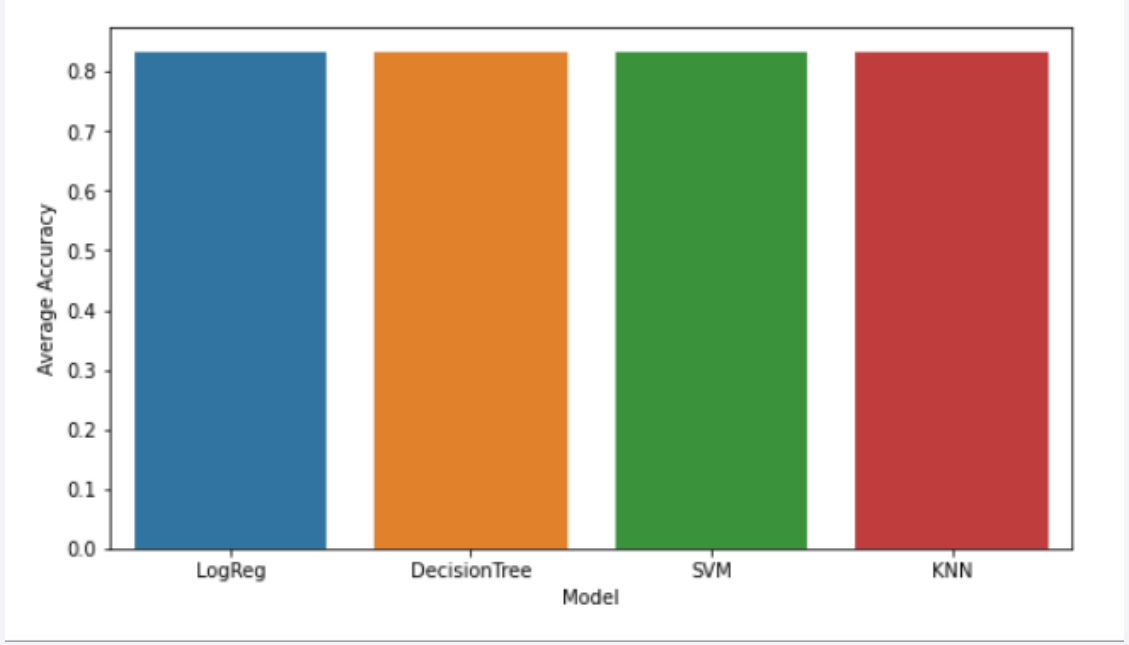

Payload range (Kg):
0 kg
2500
5000
7500
10000
Correlation between Payload and Success for All sites
Booster Version Category
v1.1
FT
B4
B5
class
Payload Mass (kg)

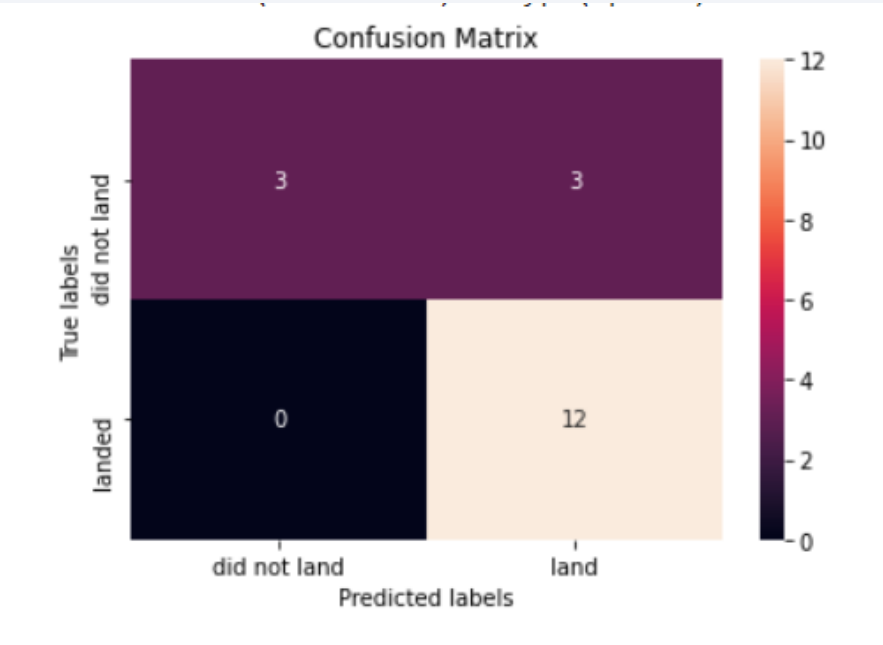

All Sites
Total Success Launches By Site
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
29.2%
41.7%
16.7%
12.5%


Payload range (Kg):
0 kg
2500
5000
7500
10000
Correlation between Payload and Success for All sites
Booster Version Category
v1.0
v1.1
FT
B4
B5
class
Payload Mass (kg)

Results

- Logistic Regression, Tree Classifications, Support Vector Machine and K Nearest Neighbor all achieve the same prediction accuracy for this dataset.



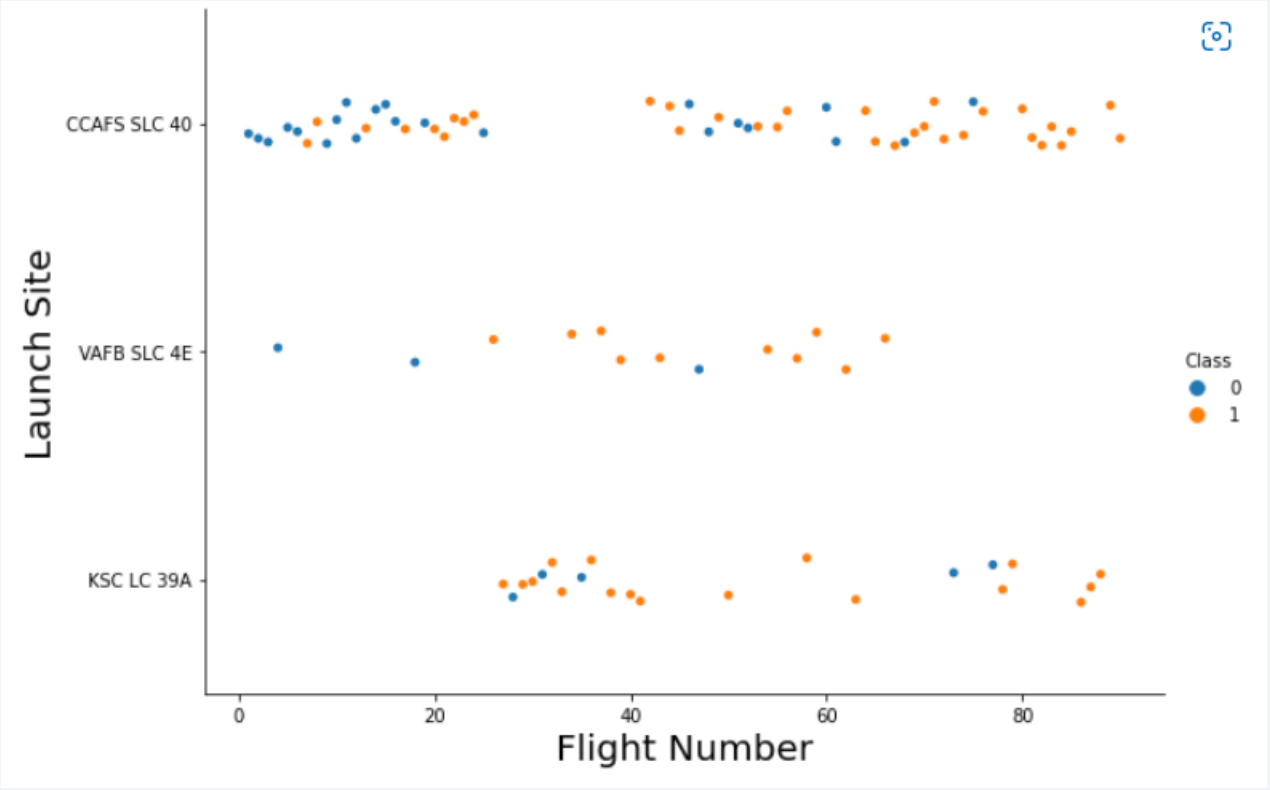


Section 2

Insights drawn from EDA

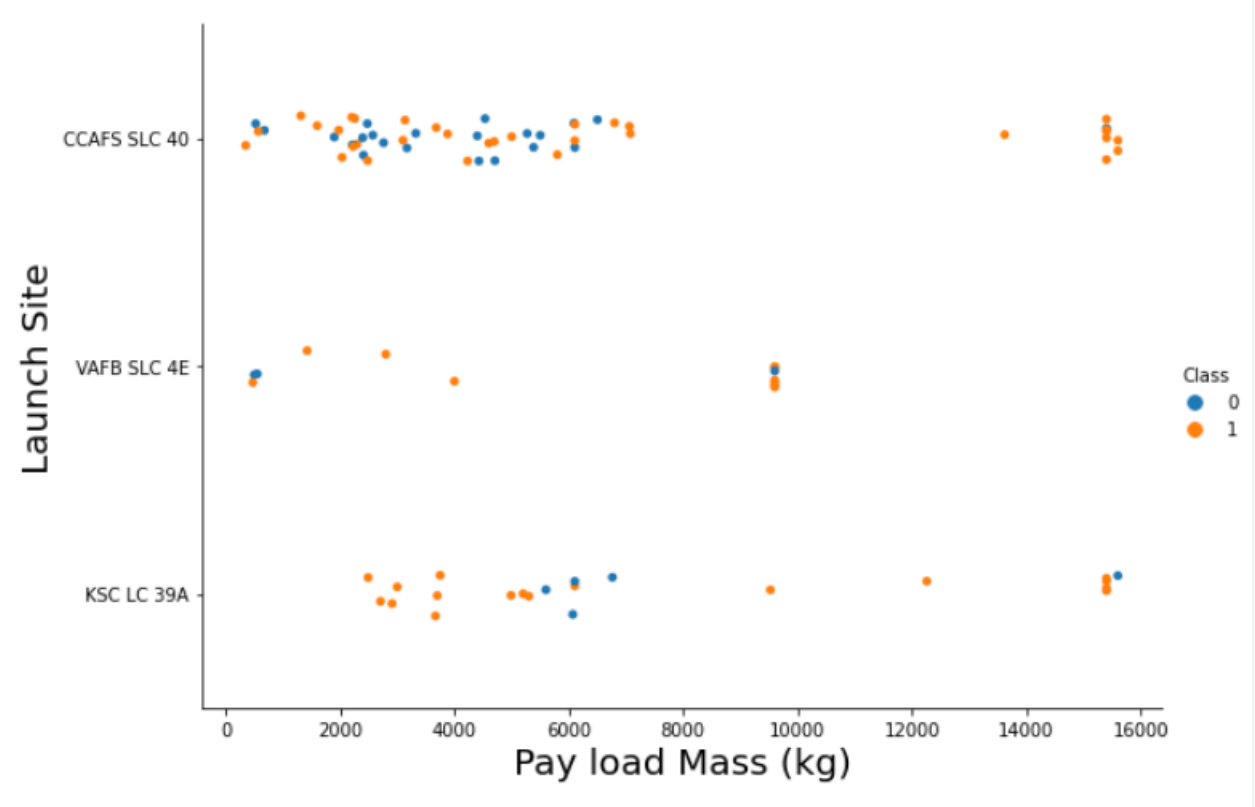
Flight Number vs. Launch Site

- Most launches in the earlier period were conducted in CCAFS LC-40.
- CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%



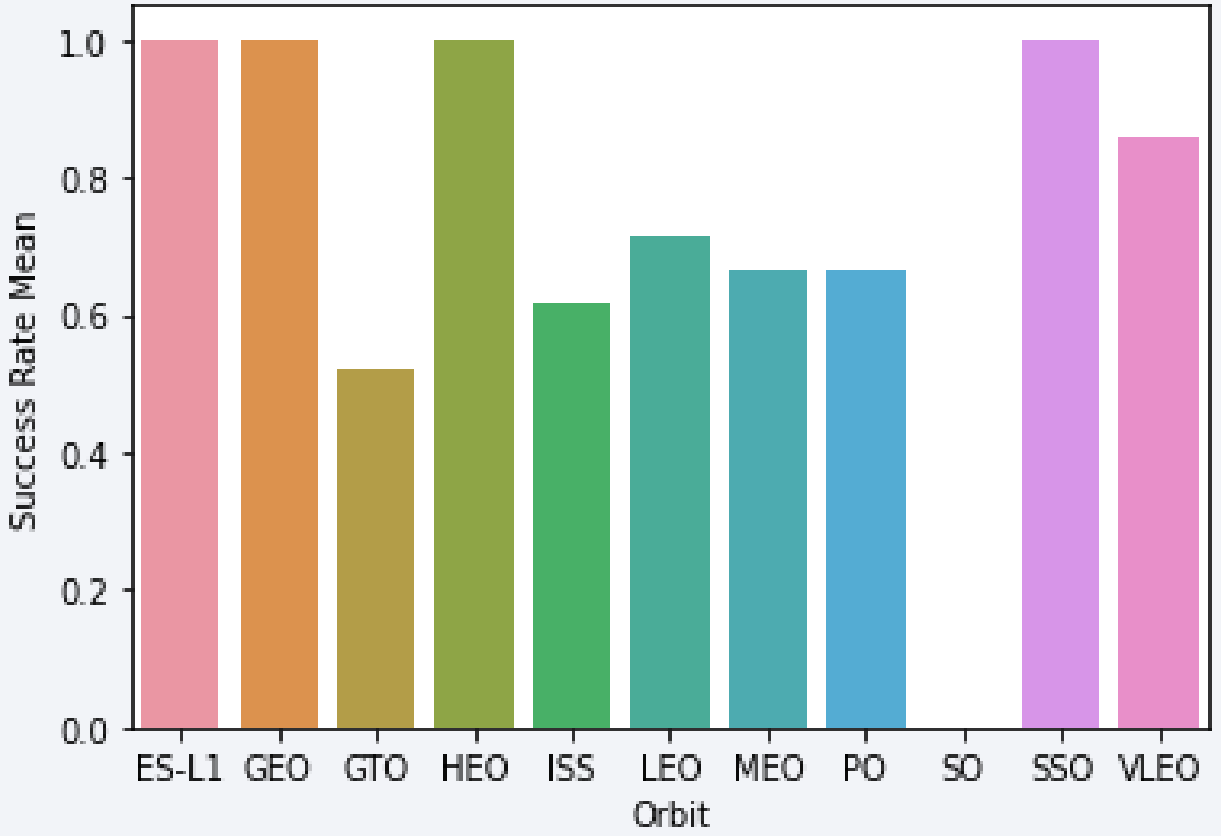
Payload vs. Launch Site

- In VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).



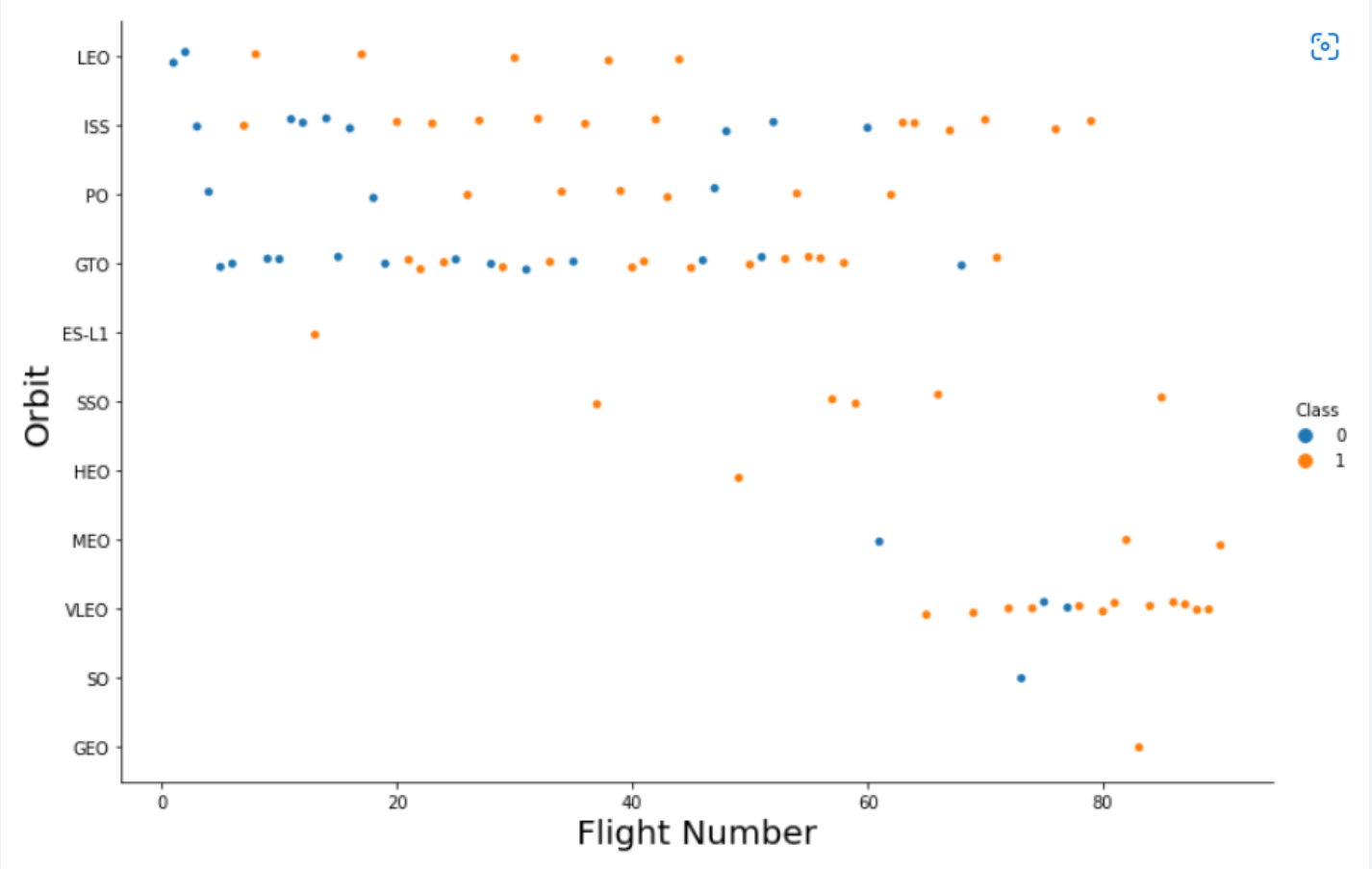
Success Rate vs. Orbit Type

- There is no successful landing in the SO orbit.
- ES-L1, GEO and SSO have 100% successful landings.
- Except SO, all orbit types have successful rate over 50%.



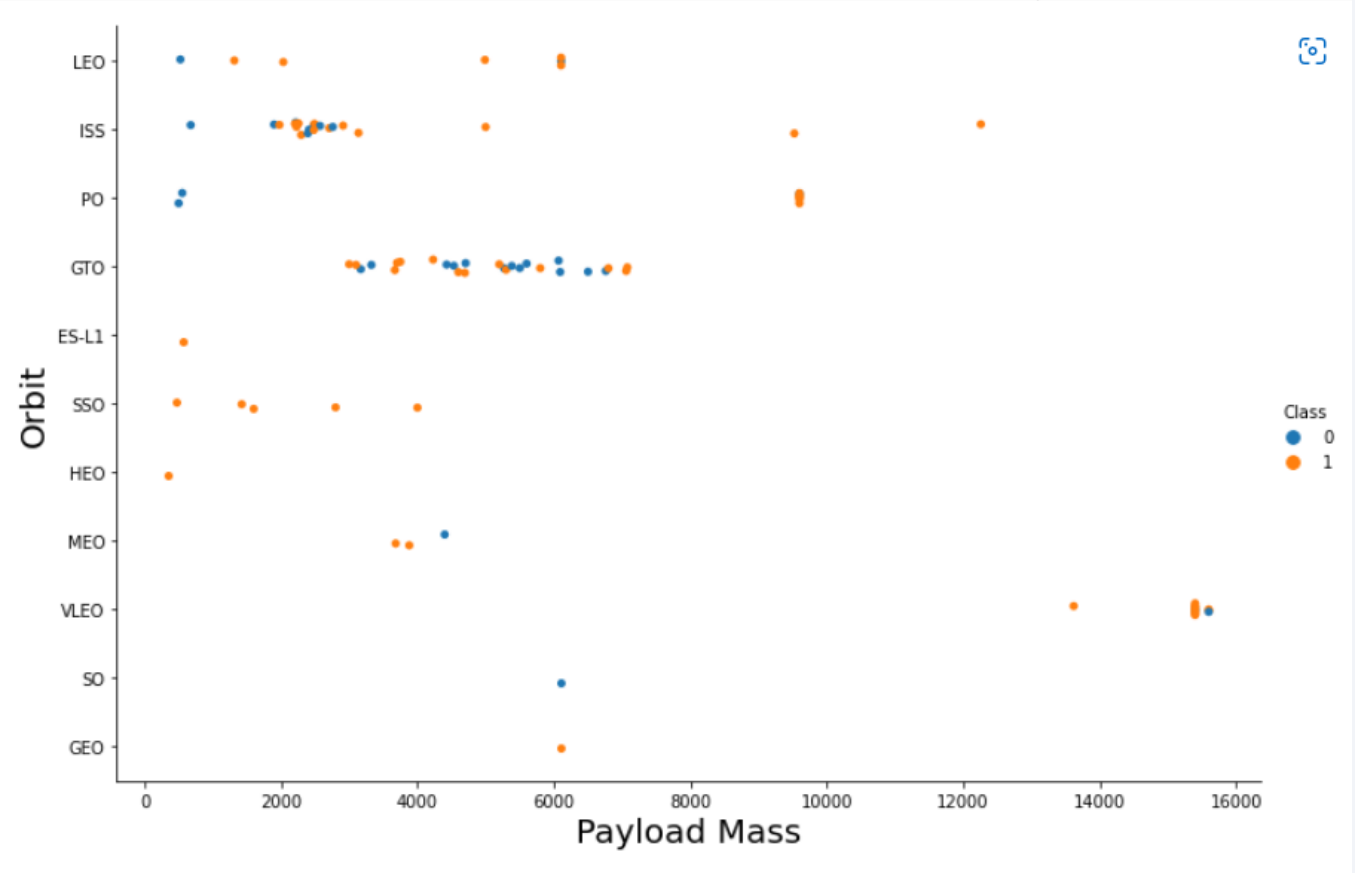
Flight Number vs. Orbit Type

- In the LEO orbit the Success appears related to the number of flights
- There seems to be no relationship between flight number when in GTO orbit



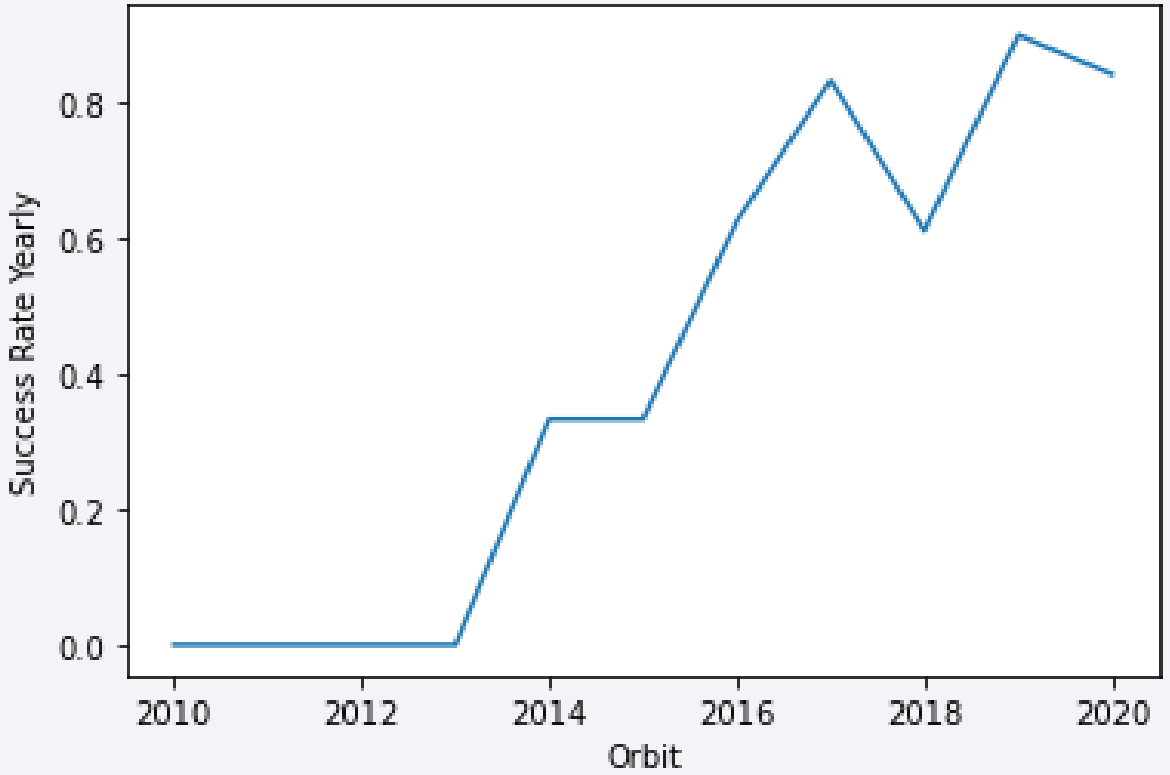
Payload vs. Orbit Type

- With heavy payloads, the successful landing or positive landing rate are higher for Polar, LEO and ISS.
- For GTO, the trend is unclear as both positive landing rate and negative landing (unsuccessful mission) are scattered.



Launch Success Yearly Trend

- There was no success launch until 2013;
- It started increasing from 2013 and kept the upward trend till 2020.



All Launch Site Names

- Launch site "CCAFS SLC-40" is previously known as "CCAFS LC-40“, therefore there are four site names listed while in fact only three launch sites exist.

```
[ ]  %sql select distinct launch_site from SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- Launch site ‘CCAFS SLC-40’ is previously known as ‘CCAFS LC-40’.

```
[11] %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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) NRO	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
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload mass carried by boosters ordered by NASA Commercial Resupply Services is 48,213 kilogram.

```
[13] %%sql
     select sum(payload_mass_kg) from SPACEXTBL
       where customer like '%NASA (CRS)%';
```

```
 * sqlite:///my_data1.db
Done.
sum(payload_mass_kg)
48213
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is approximately 2,534.67 kilogram. This counts all booster versions of F9 v1.1 model (for example, F9 v1.1 B1003, F9 v1.1 B1011 etc).

```
[14] %%sql
     select avg(payload_mass_kg) from SPACEXTBL
     where booster_version like '%F9 v1.1%';
```

```
 * sqlite:///my_data1.db
Done.
avg(payload_mass_kg)
2534.6666666666665
```

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was carried out by Falcon 9 FT model version B1019 in 22 December 2015, five and a half year from the first mission by Falcon 9 v1.0 model in 04 June 2010.

```
%sql select min(date), booster_version from SPACEXTBL where "landing_outcome" = "Success (ground pad)";
```

```
 * sqlite:///my_data1.db
Done.
```

min(date)	booster_version
2015-12-22	F9 FT B1019

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are 14 boosters which have successfully landed on drone ship and had payload mass between 4000 and 6000.

```
%%sql
select Booster_Version from SPACEXTBL
where "landing_outcome" = "Success (drone ship)"
and 4000< payload_mass_kg < 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

booster_version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- The rate of successful outcomes is 99 percent. Among 101 mission, there is only one failure outcome and one with payload status unclear. The difference between mission outcomes and landing outcome implies that most mission outcomes were planned.

```
%%sql
select mission_outcome, count(*)
from SPACEXTBL
group by mission_outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

- The maximum payload mass is 15,600 kg, carrying by F9 B5 B10 version, carrying 15,600 kg.

```
%%sql
select booster_version, payload_mass_kg from SPACEXTBL
where payload_mass_kg = (select max(payload_mass_kg) from SPACEXTBL);
```

```
 * sqlite:///my_data1.db
Done.
```

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

2015 Launch Records

- Both missions attempted to land in drone ship in 2015 were failed. Both were carried out by F9 v1.1 versions and took off from launch site CCAFS LC-40.

```
%%sql
select substr(Date, 6, 2) as month, "landing_outcome", booster_version, launch_site
from SPACEXTBL
where "landing_outcome" = "Failure (drone ship)"
and substr(date,1,4) = '2015';
```

```
 * sqlite:///my_data1.db
Done.
```

month	landing_outcome	booster_version	launch_site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Between 2010-06-04 and 2017-03-20, there were 31 missions conducted in total. A half of landing attempts (52.38%) were on drone ship with the success rate of 45.45%.

```
[67] %%sql
     select "landing_outcome", count(*) as launch_count from SPACEXTBL
     where date between '2010-06-04' and '2017-03-20'
     group by "landing_outcome"
     order by launch_count desc;
```

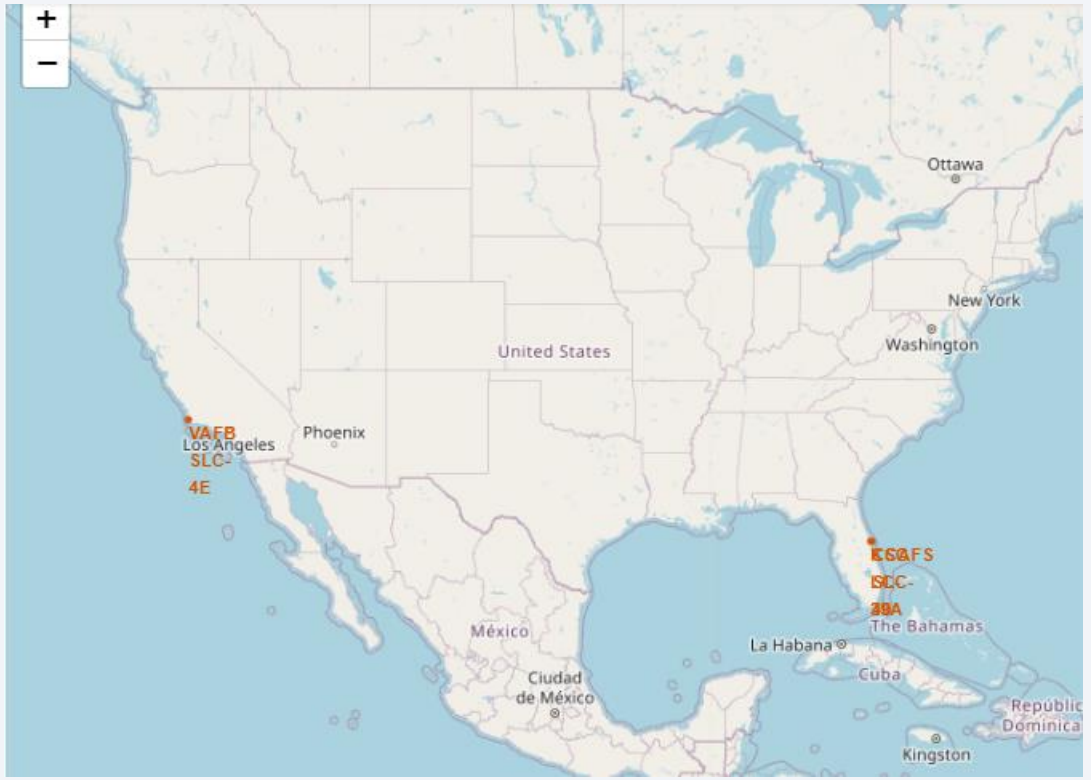
```
 * sqlite:///my_data1.db
Done.
```

landing_outcome	launch_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

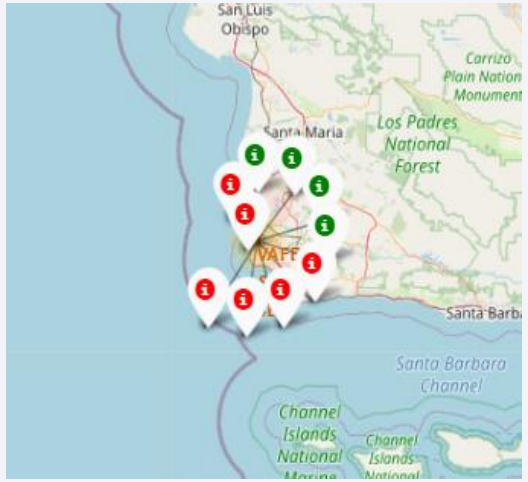

Section 3
Launch Sites
Proximities Analysis

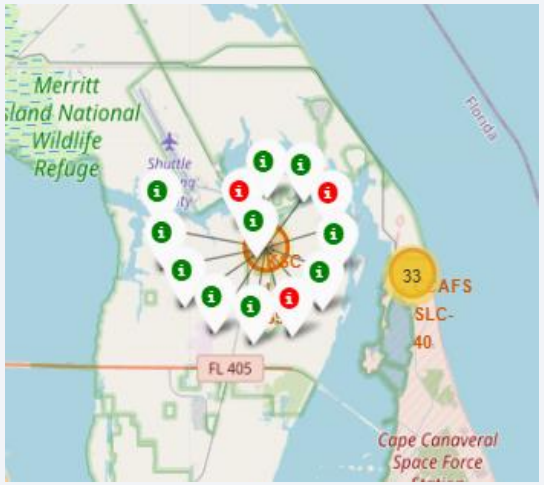
Launch Sites marked on map

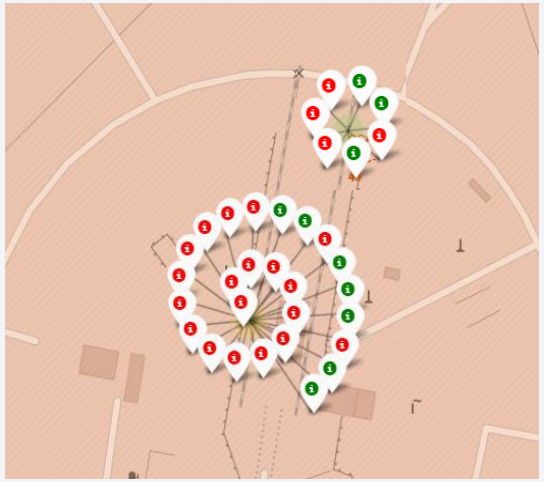
- All three launch sites are in proximity to the Equator line and the coast.
- Both sites CCAFS SLC-40 and KSC LC-39A are in the same area in the east Florida coast, while site VAFB SLC-4E site is located to the South California coast.



Sites marked for success/failed launches



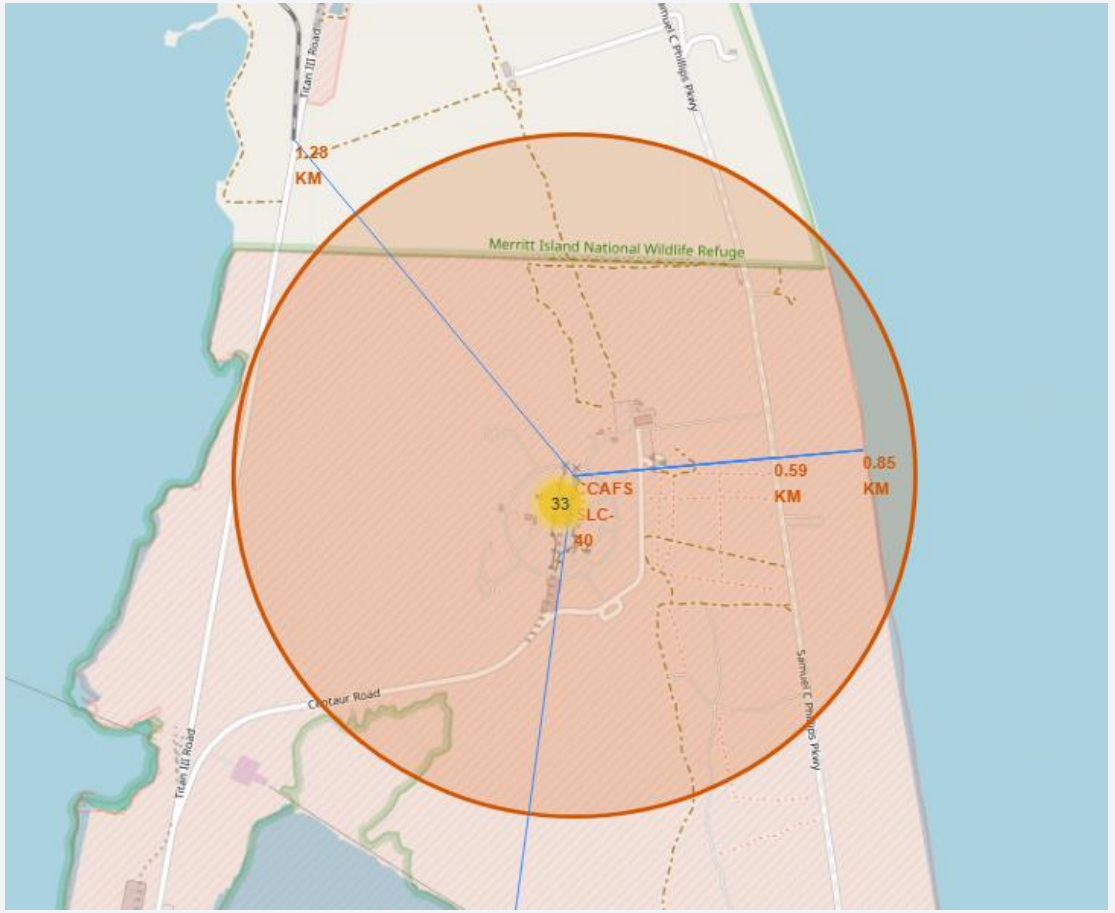


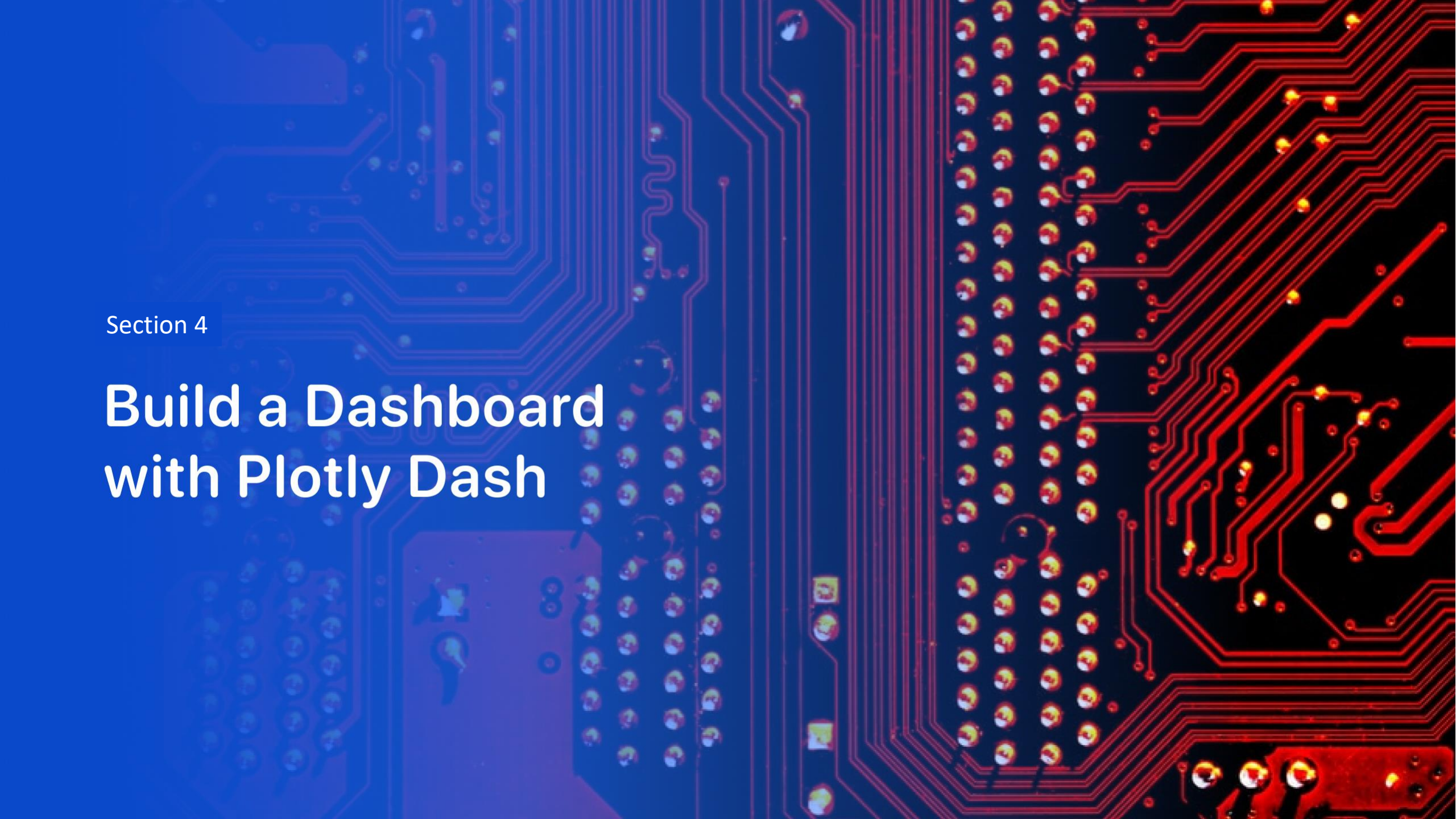


- Site KSC LC-39A appears to have relatively high success rates (76.92 %), in comparison to CCAFS SLC-40 (previously CCAFS LC-40) and VAFB SLC-4E (30.30 % and 40.00 %, respectively).

Distances from CCAFS SLC-40 to its proximities

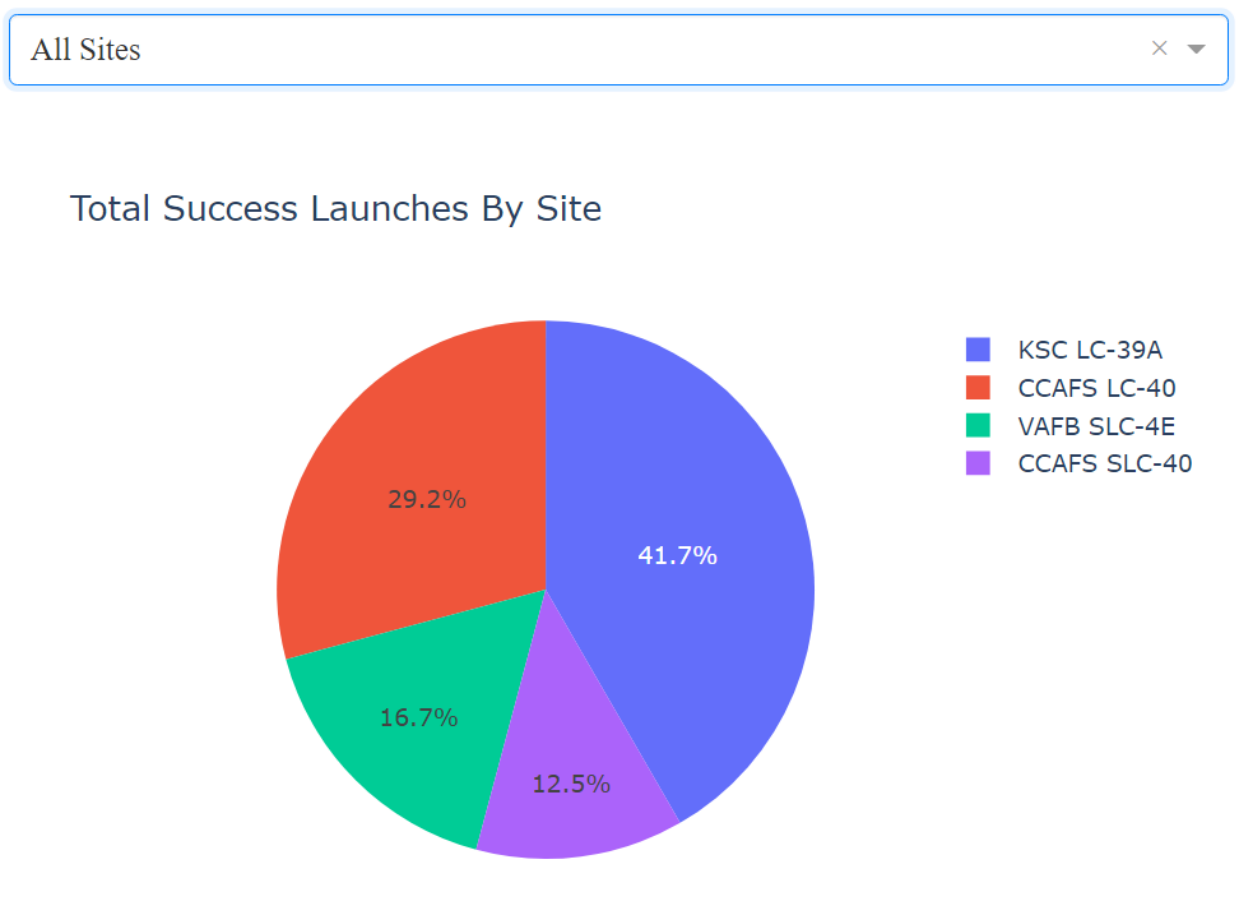
- Launch site CCAFS SLC-40 is well connected to transport facilities, including highway (0.59 km), railway (1.28 km), coastline (0.85 km).
- The nearest large city with an international airport is Melbourne, which is outside of neighborhood but still within a convenient distance (52.07 km).
- An optimal launch site should be in proximity to the Equator line and the coastline, as well as well-connected to transport facilities.




Section 4
Build a Dashboard with Plotly Dash

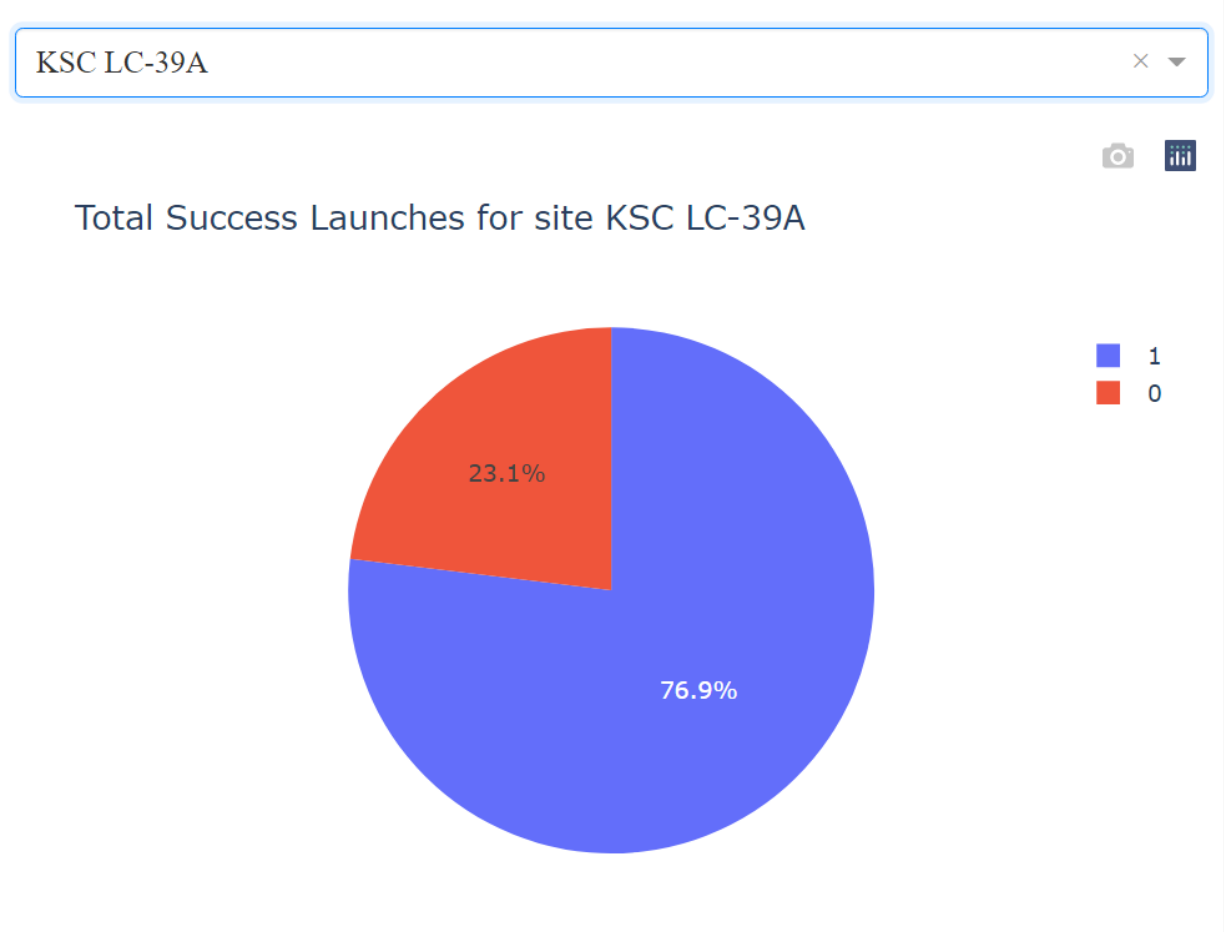
Total success launch by site

- KSC LC-39A accounts for 41.7 % of total success launches, equal proportion with site CCAFS LC-40/ CCAFS SLC-40.



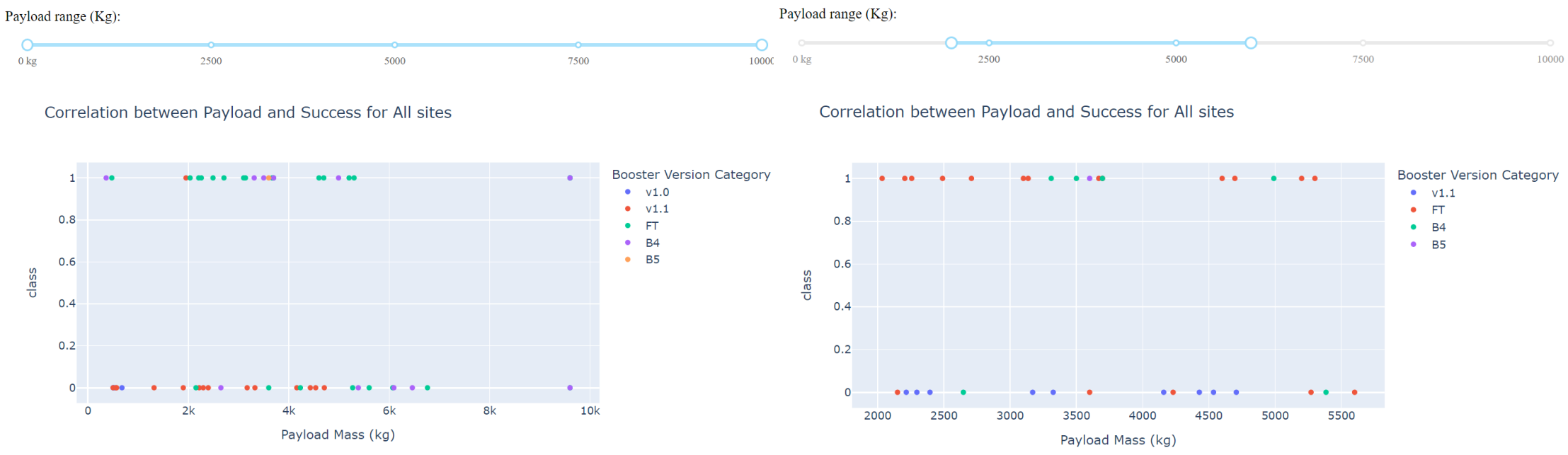
Total Success Launches for site KSC LC-39A

- KSC LC-39A has highest success launch rate among all sites, averagely 3 in 4 launches are success.



Correlation between Payload and Success for All sites

- Small-lift boosters (payload <2000 kg) are more likely to fail than medium-lift boosters.
- Most success launches have payload mass in range of 2000-6000 kg.

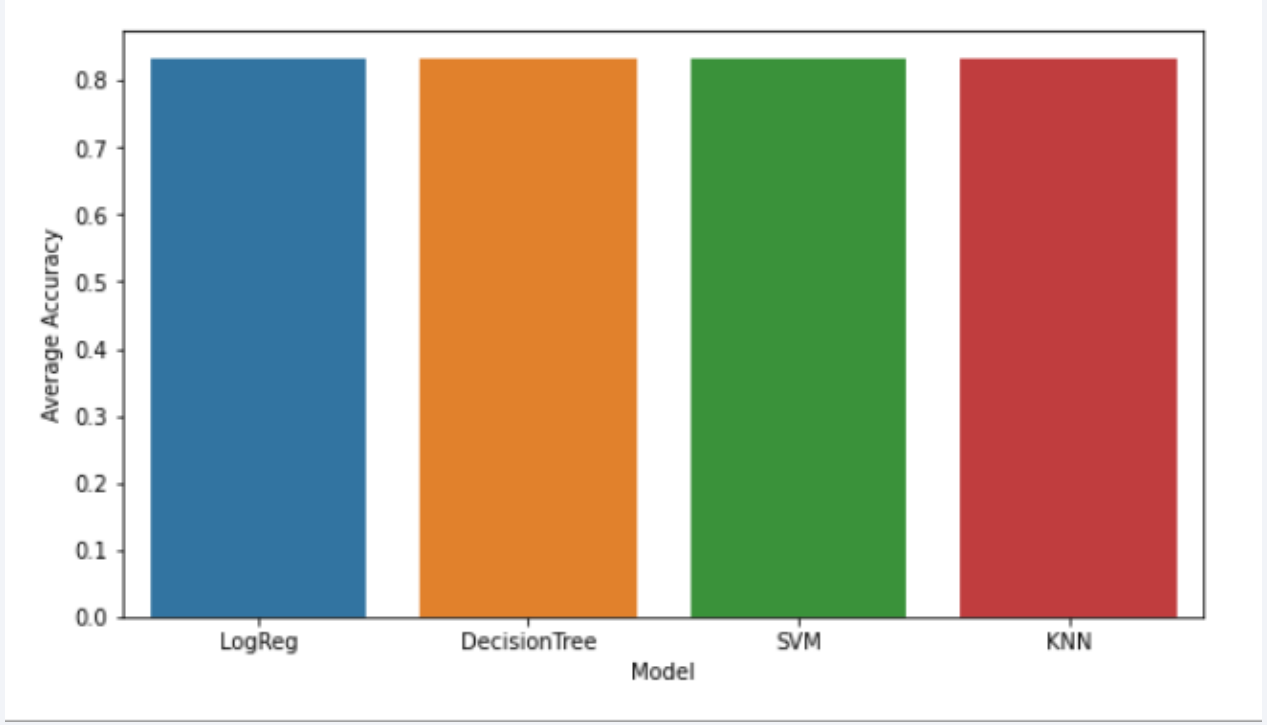


Section 5

Predictive Analysis (Classification)

Classification Accuracy

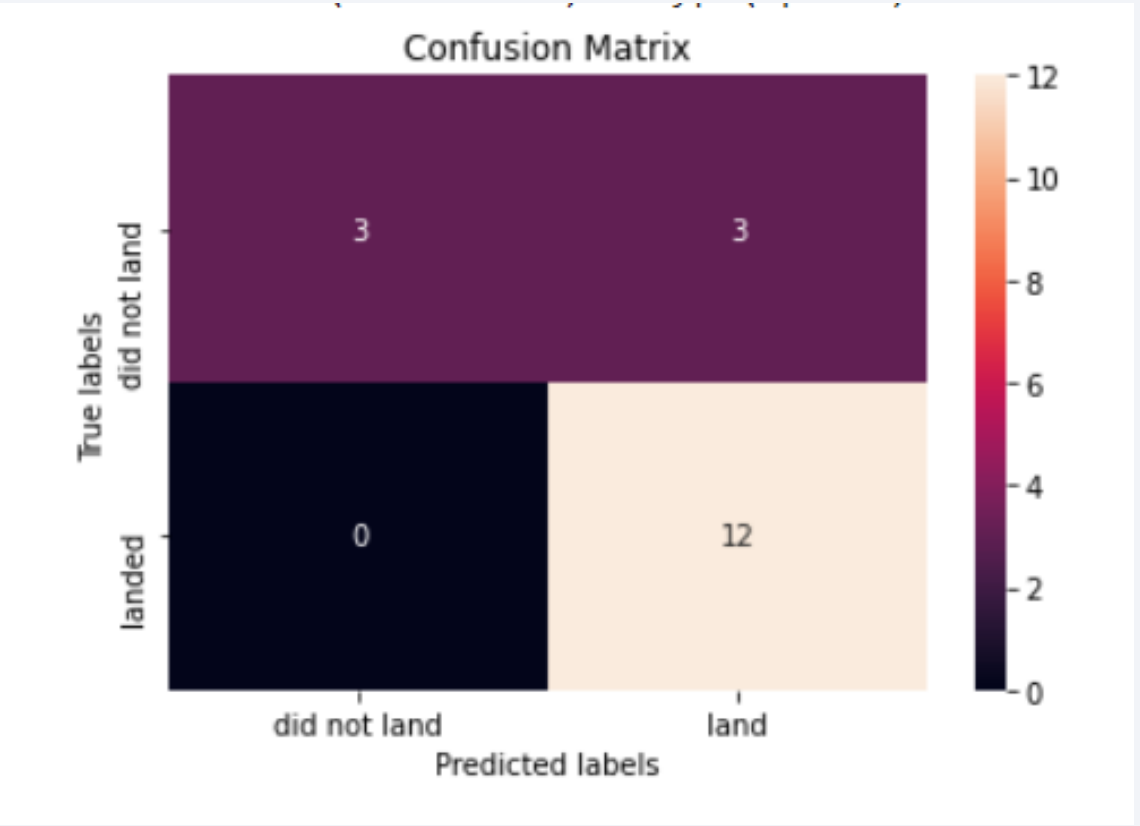
- All models achieve the similar accuracy (83%), which can be partly explained by their non-complexity and small dataset.
- In term of best accuracy, Tree Classifications has the highest score (87%)



Model	Accuracy	Best Accuracy
Logistic Regression	0.83	0.85
Tree Classifications	0.83	0.88
Support Vector Machine	0.83	0.85
K Nearest Neighbors	0.83	0.85

Confusion Matrix

- Confusion Matrices are similar for all models
- Precision $= \frac{12}{12+3} = 0.80$
- Recall $= \frac{12}{12+0} = 1.00$
- F1 score $= \frac{12}{12+\frac{1}{2}(3+0)} = 0.89$
- Predictive models have high sensitivity. The major problem is false positives.



Conclusions

- Decision Tree, SVM, KNN, and Logistic Regression models achieve the similar prediction accuracy. Decision Tree performs better in term of best score.
- Medium-lift boosters perform better than small-lift boosters.
- The success rates increase over time as SpaceX gradually improves their technologies.
- KSC LC 39A had the highest success launches ratio among all the sites.
- Orbit GEO, HEO, SSO, ES L1 has 100 percent success rate.

Thank you!