IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jean C. Sobreira
2023-12-08

Outline

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

Executive Summary

- The data was collected from public SpaceX API and SpaceX Wikipedia. The most important variables were selected and used to predict the success of landing.

- Four machine learning models were generated: Logistic Regression, Suport Vector Machine, Decision Tree Classifier and K Nearest Neighbour. Both Logistic Regression and KNN produced the best accuracy rate of 0.94.

Introduction

- Project background and context
  - Commercial space travels are a billionaire industry
  - Space X has the best pricing offered and largely due to the capability of reuse part of a rocket (Stage 1)
  - Space Y wants to compete with Space X

- Problems you want to find answers
  - Space Y wants to predict successful Stage 1 recovery, so they can predict more precisely their costs.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from public SpaceX API and SpaceX Wikipedia.
- Perform data wrangling
  - Both datasets were combined and then the most important variables were selected.
  - One variable was created due to classify successful and unsuccessful landings
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The models were tuned using GridSearchCV

Data Collection

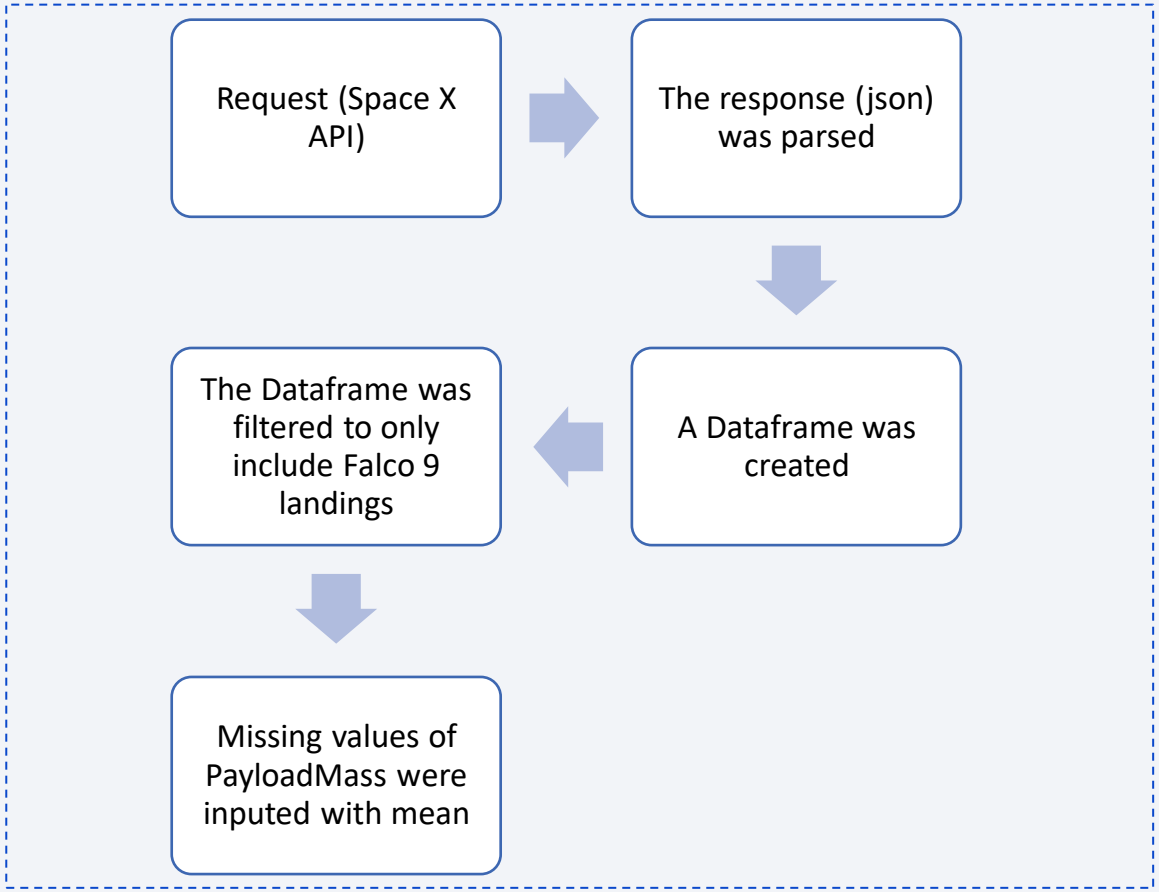
- The data was collected from public SpaceX API and web scraping from SpaceX Wikipedia.
- The next slides show flowcharts from both processes.

Data Collection – SpaceX API

- The flowchart aside shows the process used to collect data from SpaceX API.

- GitHub URL:

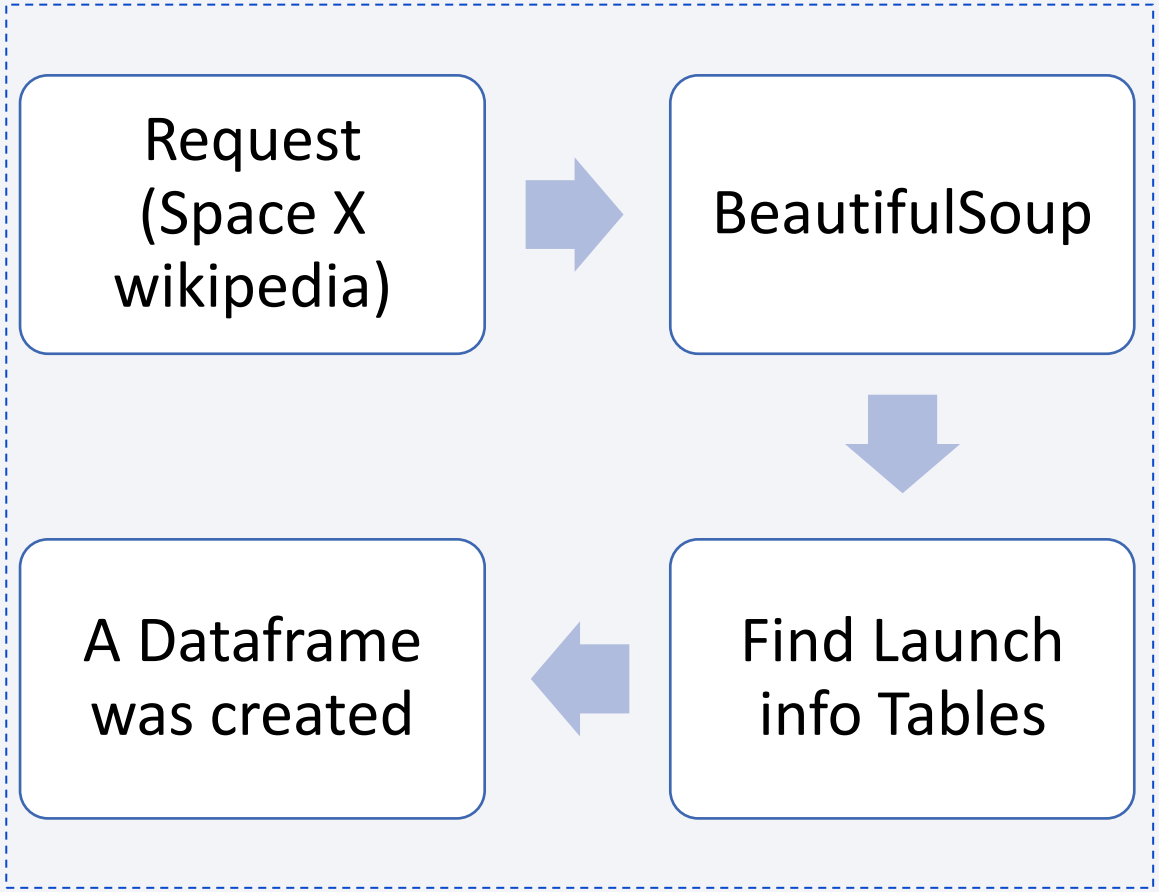
- [ibm-data-science/jupyter-labs-spacex-data-collection-api.ipynb at main · jackniet987/ibm-data-science (github.com)]



Data Collection - Scraping

- The flowchart aside shows the process used to collect data from SpaceX Wikipedia using web scrapping.

- GitHub URL:
- ibm-data-science/jupyter-labs-webscraping.ipynb at main · jackniet987/ibm-data-science (github.com)



Data Wrangling

- Feature engineering was necessary to create a column named “Class”, which shows if the Stage 1 landed (1) or not (0).
- GitHub URL:
- [ibm-data-science/labs-jupyter-spacex-Data wrangling.ipynb at main · jackniet987/ibm-data-science (github.com)]

EDA with Data Visualization

- Some variables (Flight Number, Payload Mass, Launch Site, Orbit, Class and Year) were analysed to generate insights about which features are more relevant to predict the success of "Stage 1" landing.
- Plots Used:
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend
- Scatter plots, bar plots, line charts were used to compare the relationship between variables.
- GitHub URL:
- [ibm-data-science/jupyter-labs-eda-dataviz.ipynb at main · jackniet987/ibm-data-science (github.com)]

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass.
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20
- GitHub URL: [ibm-data-science/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · jackniet987/ibm-data-science (github.com)]

Build an Interactive Map with Folium

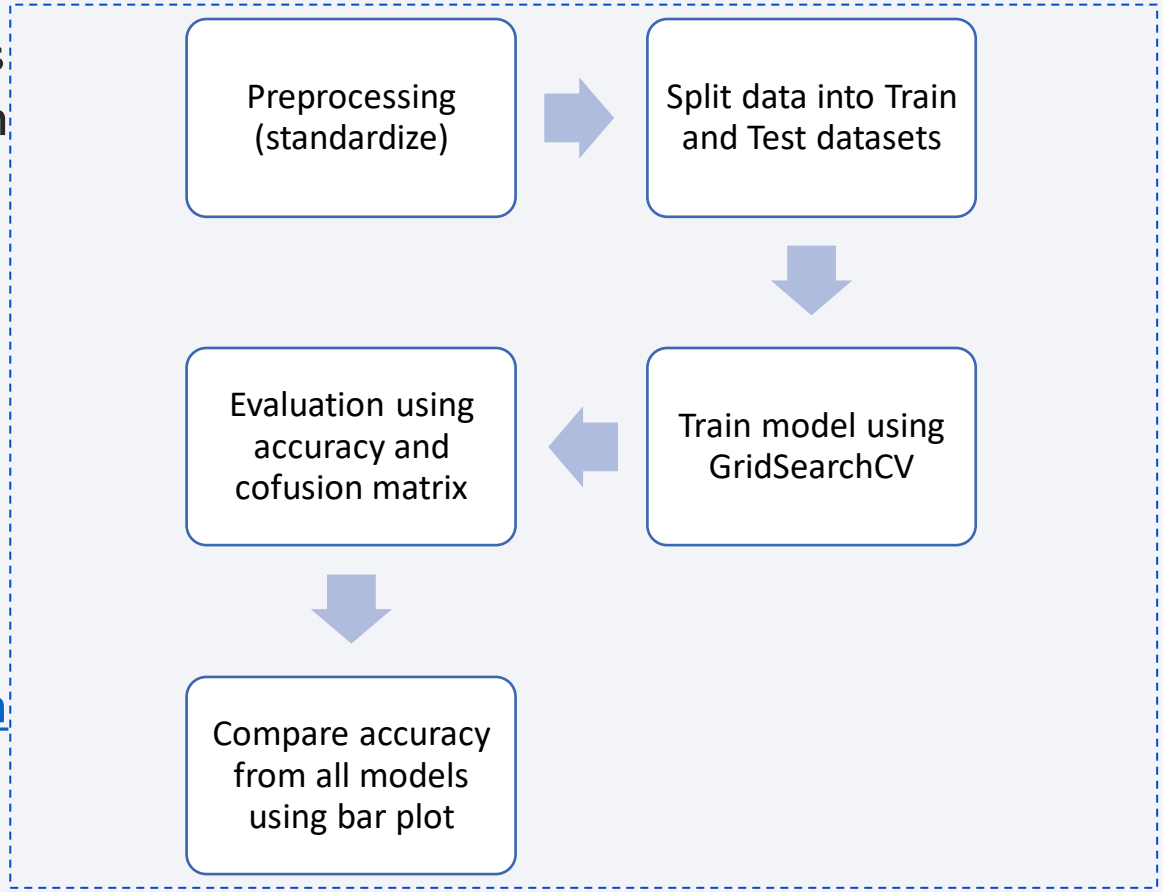
- The folium map was used to show the successful and failure landings in each launch site.
- It's possible to see also the proximities of Florida's launch site to nearest railway, highway, coastline and city.
- GitHub URL:
- ibm-data-science/lab_jupyter_launch_site_location.ipynb at main · jackniet987/ibm-data-science (github.com)

Build a Dashboard with Plotly Dash

- The Dashboard includes a pie chart and a scatter plot.
- The pie chart shows the success rate by launch site.
- The scatter plot show the success rate by payload mass.
- GitHub URL:
- ibm-data-science/spacex_dash_app.py at main · jackniet987/ibm-data-science (github.com)

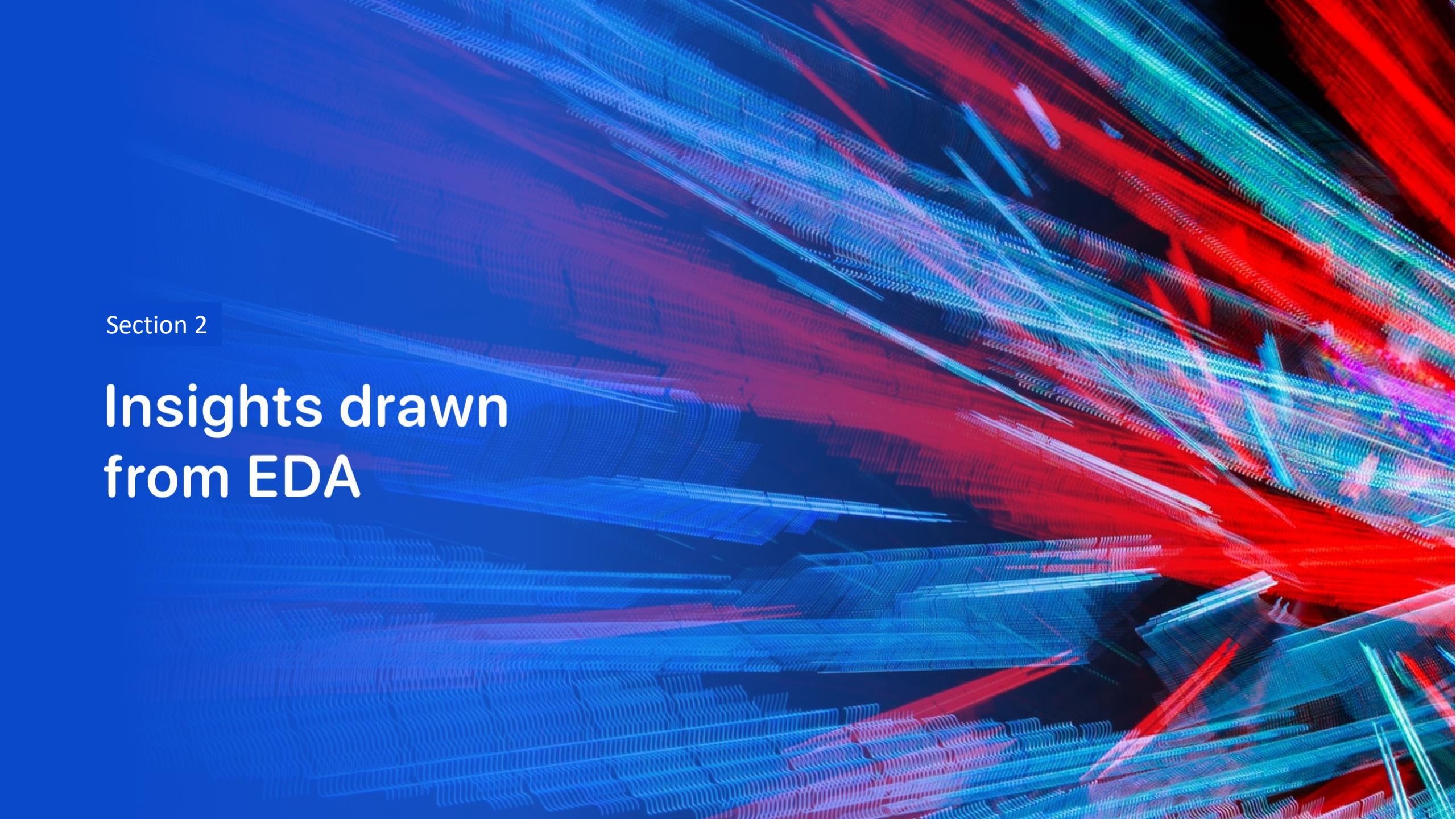
Predictive Analysis (Classification)

- The data was preprocessed, the dataset was splited into train and test datasets, and then different models were trained, those models were evaluated using accuracy metric, improved using GridSearchCV and then the accuracy of the models were compared to find the best preforming model.
- GitHub URL:
- [ibm-data-science/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb at main · jackniet987/ibm-data-science (github.com)]

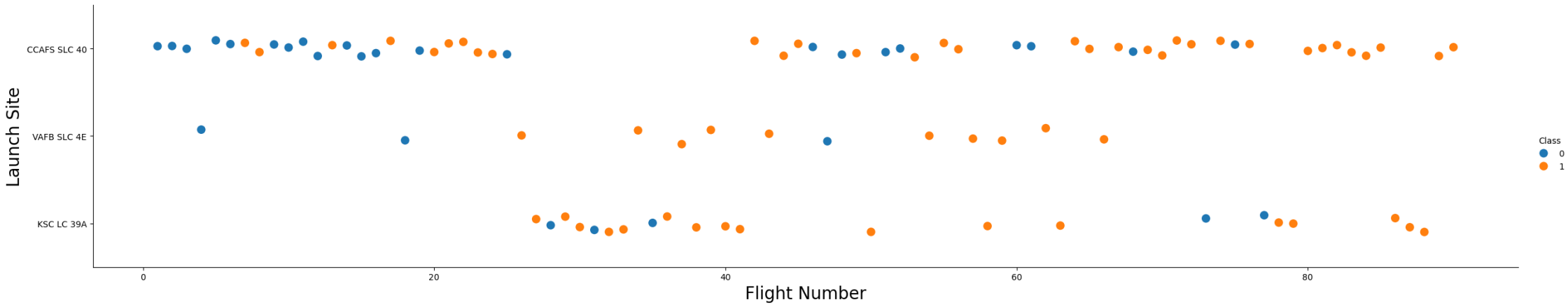


Results

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

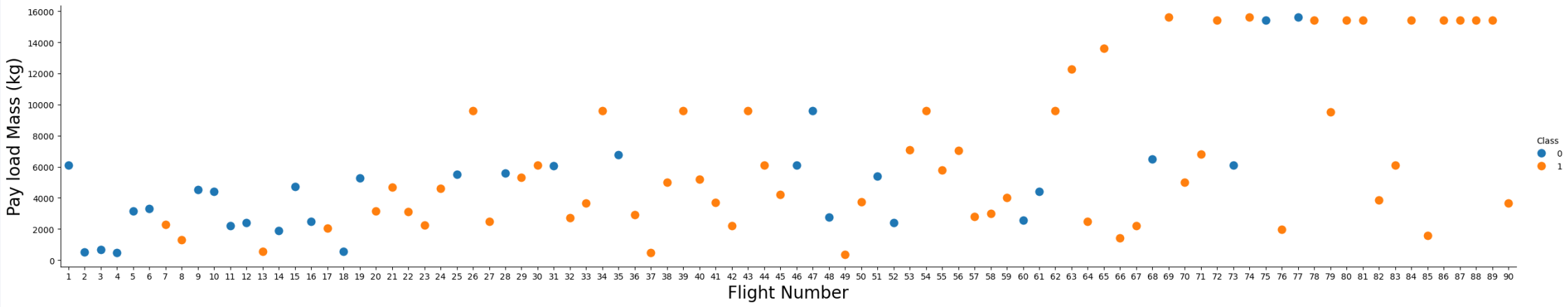

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Blue points indicate failure, orange points indicate success launch.
- Success rate increases over time, strongly after 20th flight.
- CCAFS appears to be the main launch site as it has the most volume.

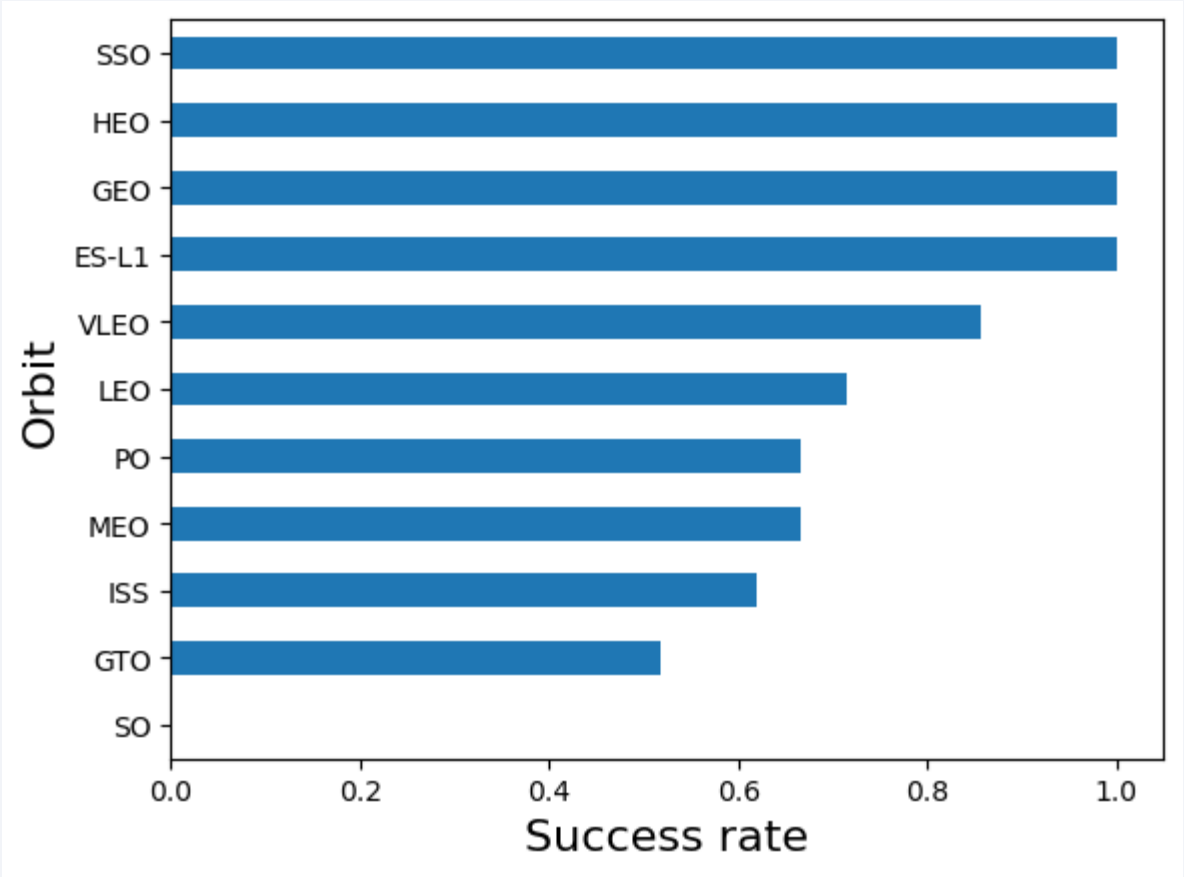
Payload vs. Launch Site



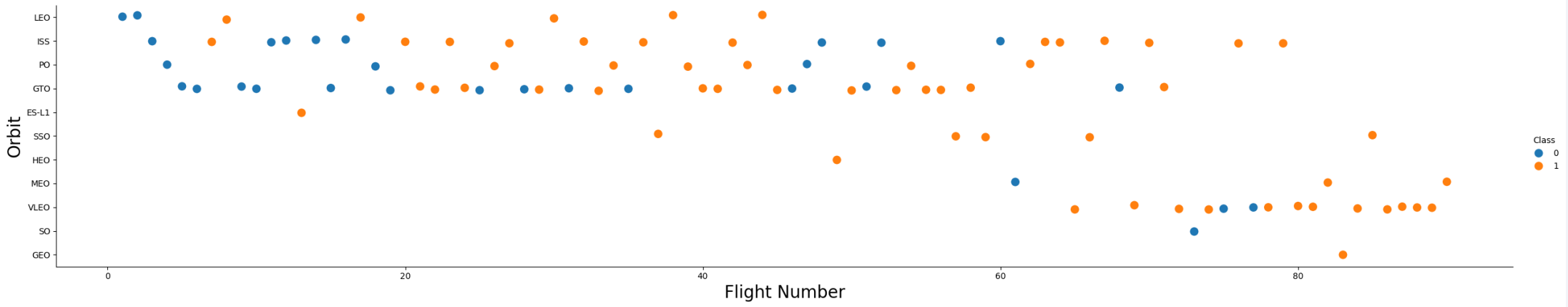
- Blue points indicate failure, orange points indicate success launch.
- Success rate increases over time.
- Different launch sites also seem to use different payload mass

Success Rate vs. Orbit Type

- Success rate scale 0.0 as 0% and 1.0 as 100%
- SSO, HEO, GEO and ES-L1 have 100% success rate.
- SO has 0% of success rate.

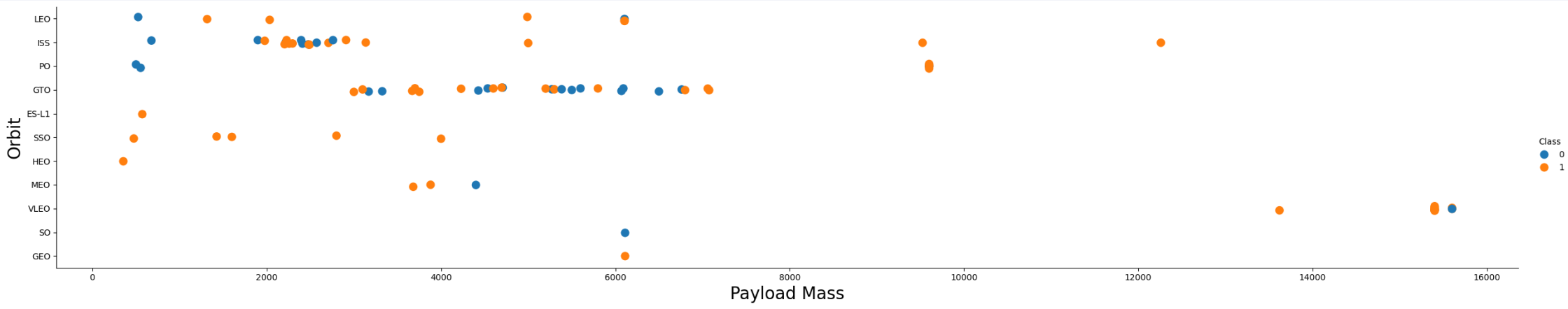


Flight Number vs. Orbit Type



- Blue points indicate failure, orange points indicate success launch.
- LEO orbit the Success appears related to the number of flights.
- Launch Outcome seems to correlate with this preference.
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits.

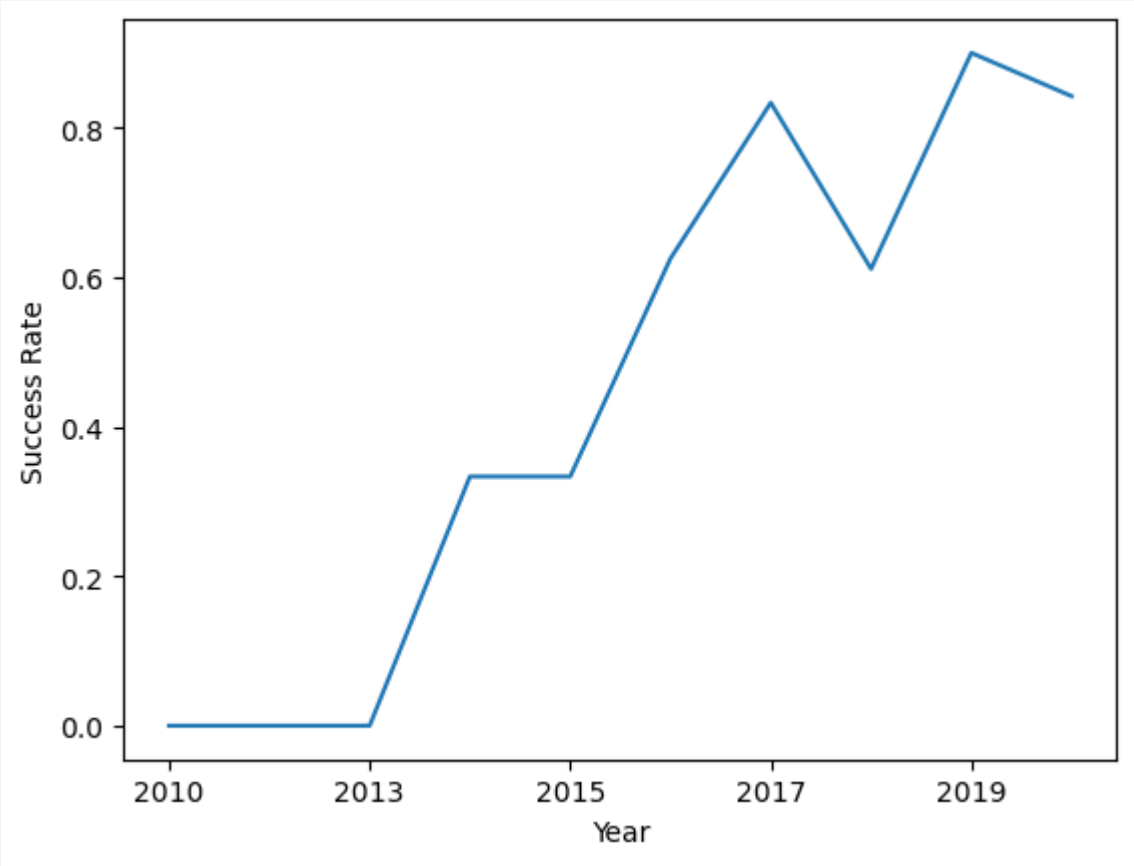
Payload vs. Orbit Type



- Blue points indicate failure, orange points indicate success launch.
- Payload mass seems to correlate with orbit
- Successful landing rate are more for LEO and ISS
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend

- Success rate increases after 2013.
- In recent years, the success rate is around 80%.



All Launch Site Names

- The query aside shows the Launch sites names.
- CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same launch site with data entry errors.
- CCAFS LC-40 was the previous name. Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

```
[15]: %%sql

      select distinct Launch_Site from SPACEXTABLE
```

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

[15]:

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

Launch Site Names Begin with 'CCA'

[14]:
```
%%sql
select * from SPACEXTABLE
where Launch_Site like '%CCA%'
limit 5
```

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

[14]:

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

- 5 records where launch sites begin with ‘CCA’ from SPACEXTABLE

Total Payload Mass

- This query shows the total payload mass in KG by NASA.
- CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

```
[8]: %%sql

     select
     sum(PAYLOAD_MASS__KG_) as total_payload_mass
     from SPACEXTABLE
     where Customer = 'NASA (CRS)'

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

[8]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- This query shows the average payload mass carried by booster version F9 v1.1.
- Average payload mass of F9 1.1 is on the low end of our payload mass range.

```
[23]: %%sql

select
Booster_Version,
avg(PAYLOAD_MASS__KG_) as avg_payload_mass
from SPACEXTABLE
where Booster_Version = 'F9 v1.1'
group by Booster_Version
order by Booster_Version
```

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

[23]:

Booster_Version	avg_payload_mass
F9 v1.1	2928.4

First Successful Ground Landing Date

- This query shows date of the first successful landing outcome on ground pad.
- First ground pad landing wasn't until the end of 2015.
- Successful landings in general appear starting 2014.

```
[9]: %%sql

     select
     Landing_Outcome,
     min(Date) as first_success_landing
     from SPACEXTABLE
     where Landing_Outcome = 'Success (ground pad)'
     group by Landing_Outcome
```

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

[9]:

Landing_Outcome	first_success_landing
Success (ground pad)	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query shows the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

[33]:
```
%%sql

select
Booster_Version
from SPACEXTABLE
where Landing_Outcome = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

[33]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- This query shows the total number of successful and failure mission outcomes.
- Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.
- SpaceX appears to achieve its mission outcome nearly 99% of the time.

[39]:
```
%%sql

select
trim(Mission_Outcome),
count(*) as total
from SPACEXTABLE
group by trim(Mission_Outcome)
order by trim(Mission_Outcome)
```

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

[39]:

trim(Mission_Outcome)	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- This query shows the name of the booster which have carried the maximum payload mass.
- The maximum payload was 15,600 kg.
- This likely indicates payload mass correlates with the booster version that is used.

```
[42]: %%sql

      select
      Booster_Version,
      max(max) maximum_payload_mass
      from (
      select
      Booster_Version,
      max(PAYLOAD_MASS__KG_) as max
      from SPACEXTABLE
      group by Booster_Version ) T1
```

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

[42]:

Booster_Version	maximum_payload_mass
F9 B5 B1048.4	15600

2015 Launch Records

- This query shows the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- There were two such occurrences

[45]:
```
%%sql

select
substr(Date,0,5) as year,
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)'
group by substr(Date, 6,2)
```

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

[45]:

year	month	Landing_Outcome	Booster_Version	Launch_Site
2015	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- This query shows the rank of counting of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
- There are one type of successful landing outcomes Success (ground pad) and one of failure (drone ship).

```
[10]: %%sql

select
Landing_Outcome,
count(*) as Total
from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
and Landing_Outcome in ('Failure (drone ship)', 'Success (ground pad)')
group by Landing_Outcome
order by Total desc
```

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

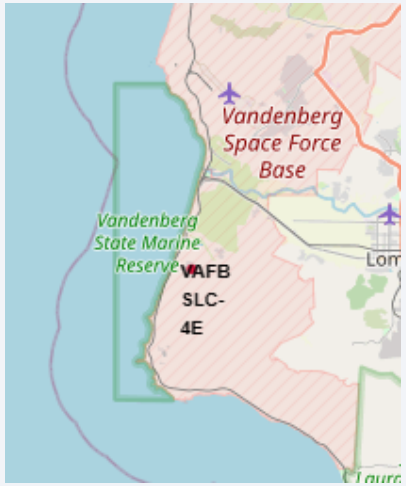
[10]:

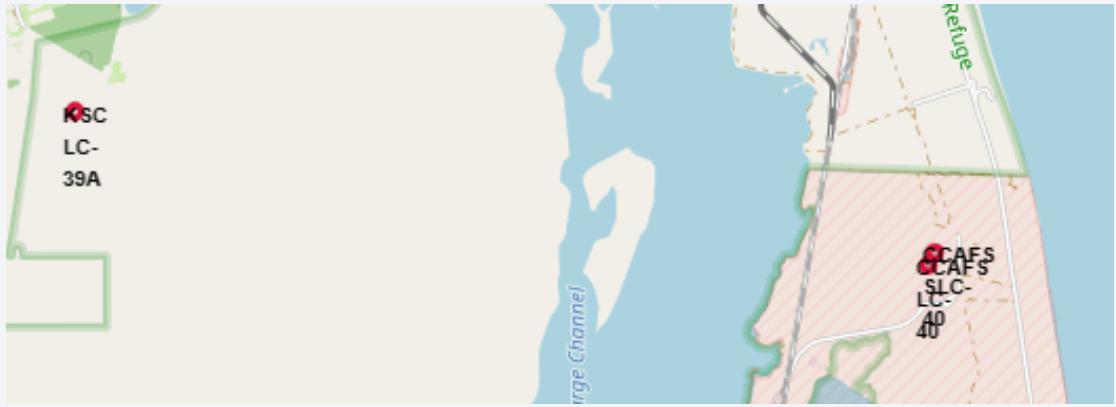
Landing_Outcome	Total
Failure (drone ship)	5
Success (ground pad)	3


Section 3
Launch Sites
Proximities Analysis

Launch Site Locations

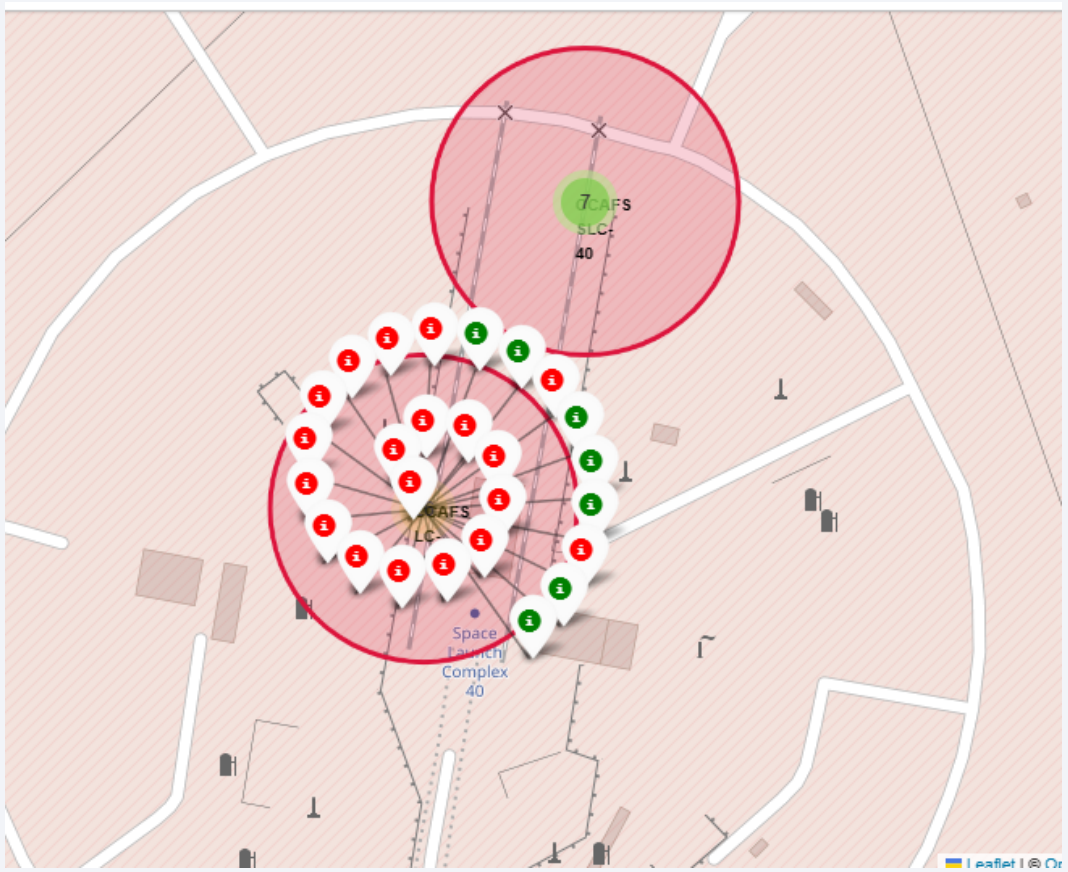
- The map to the right shows the launch sites all over the USA.
- The map to the bottom left shows the launch site in California and the map to the bottom right show de launch sites in Florida.





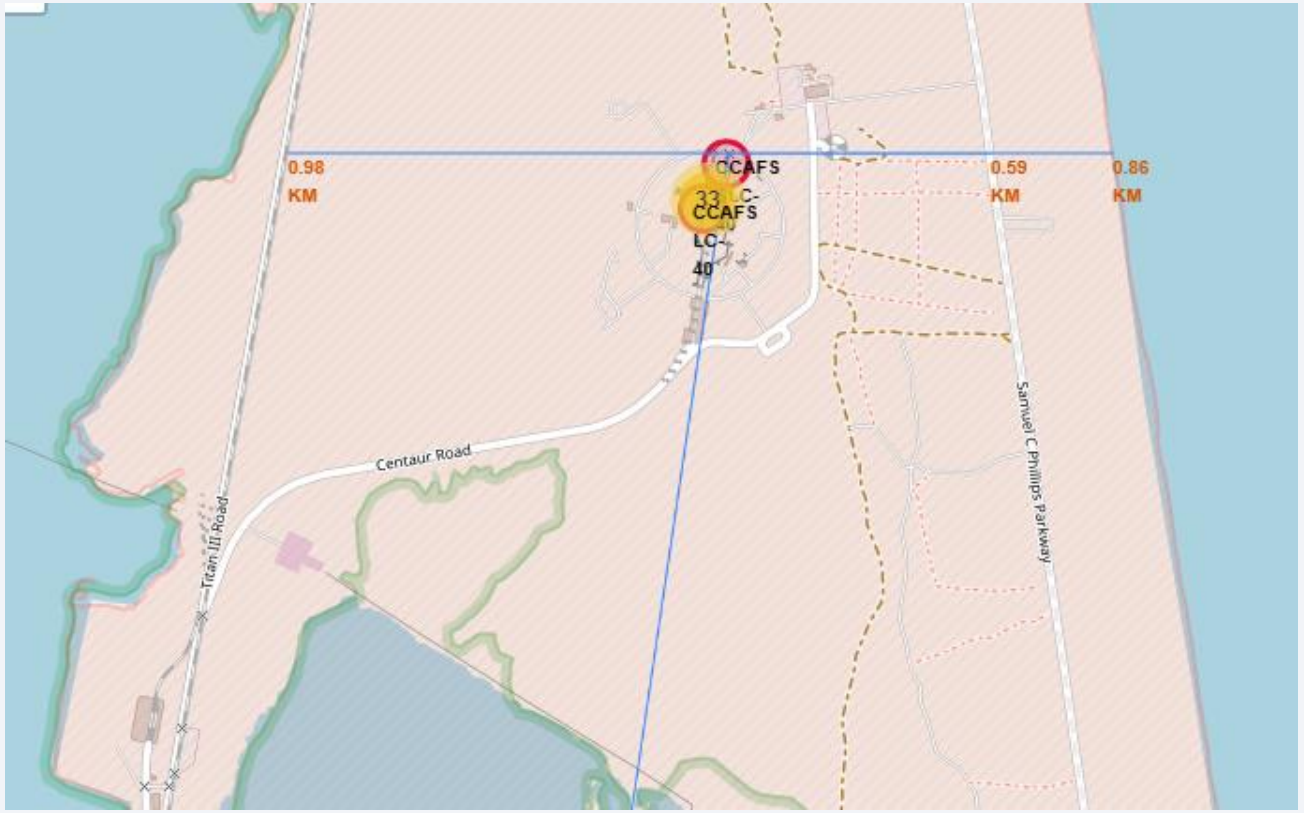
Color-coded markers

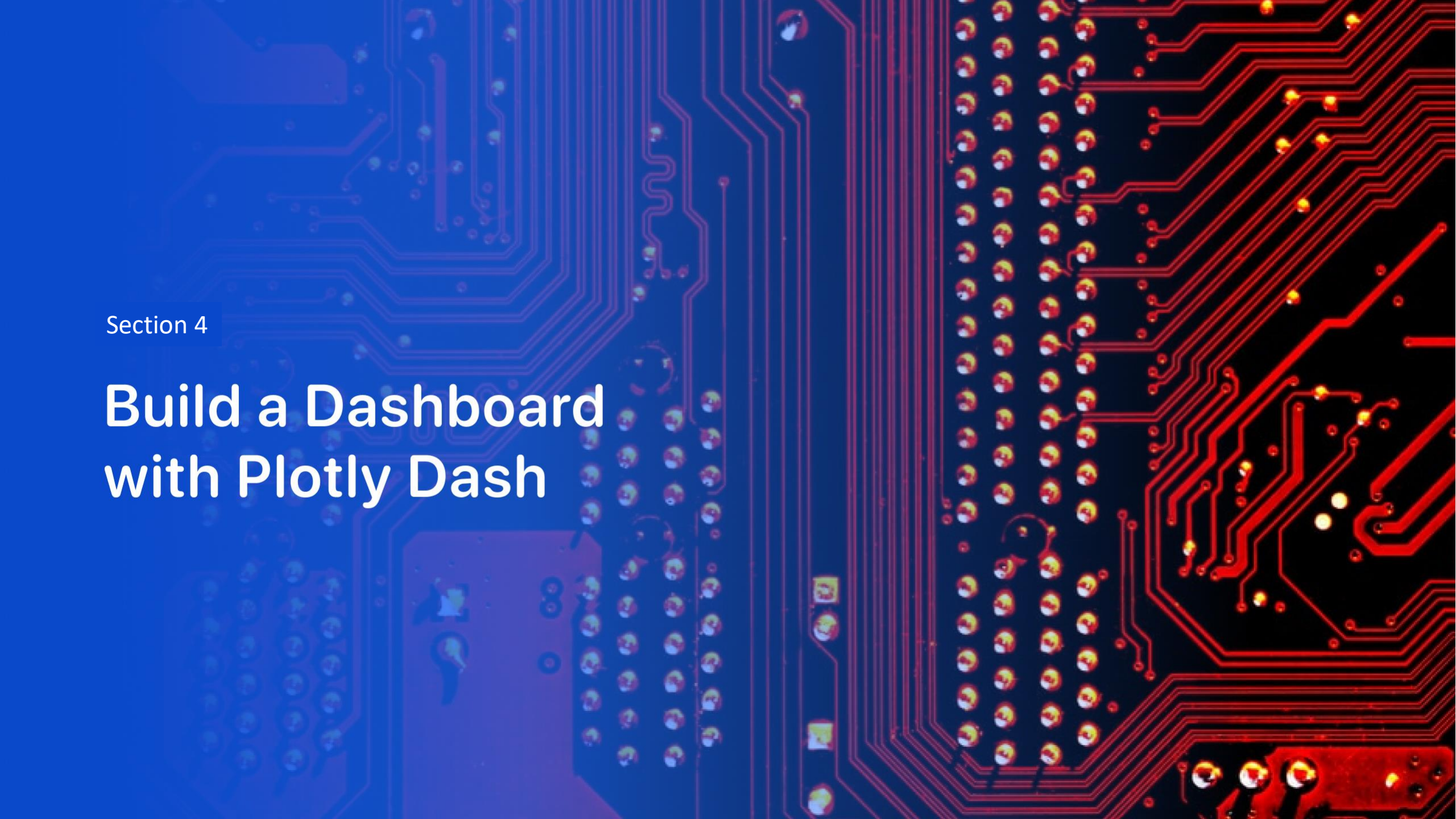
- The map shows the success landing (green icon) and the failure landing (red icon) for each launch site.
- In this exemple we have 7 success landing.



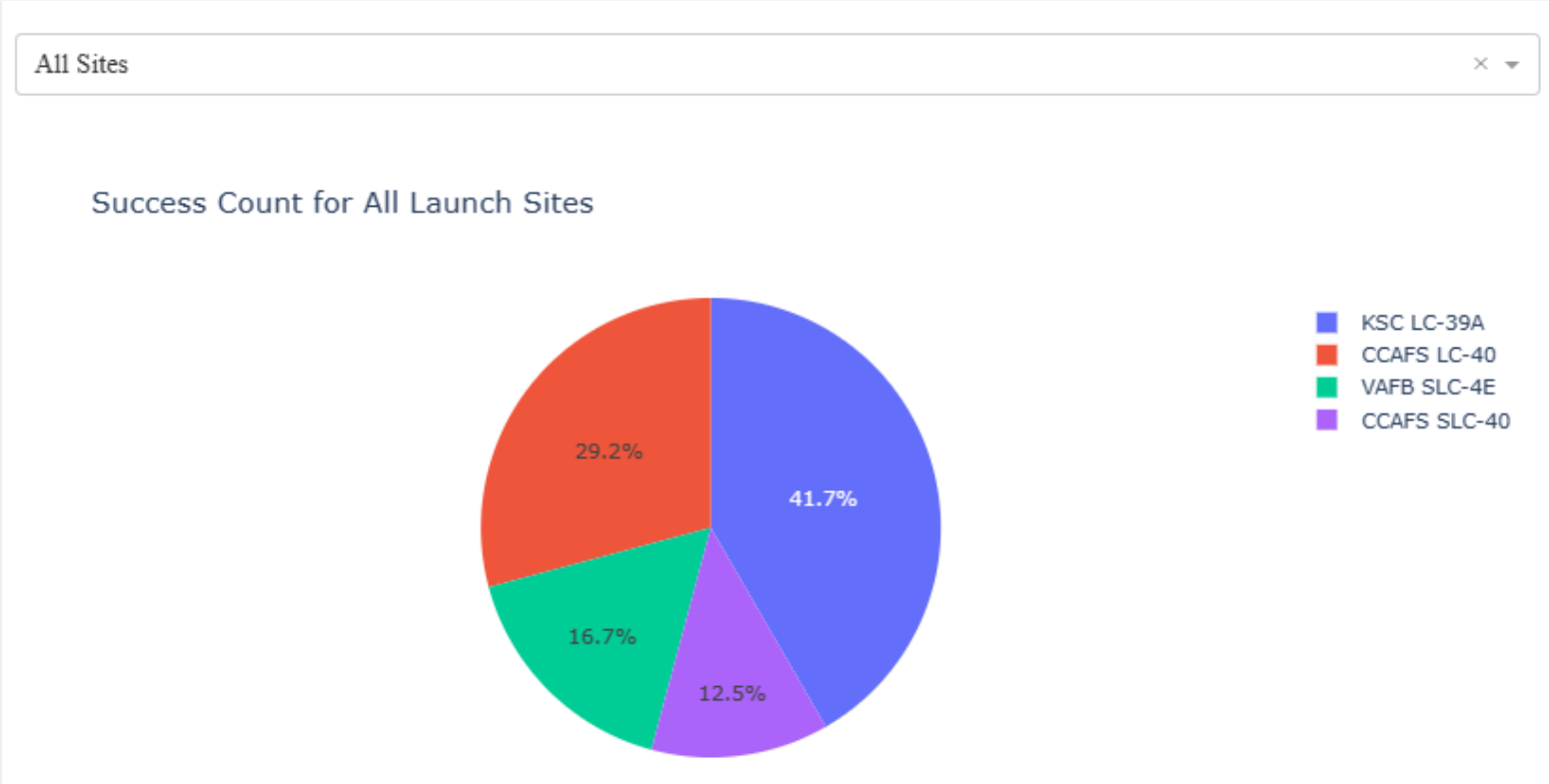
Infrastructure proximities

- This map shows the launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed



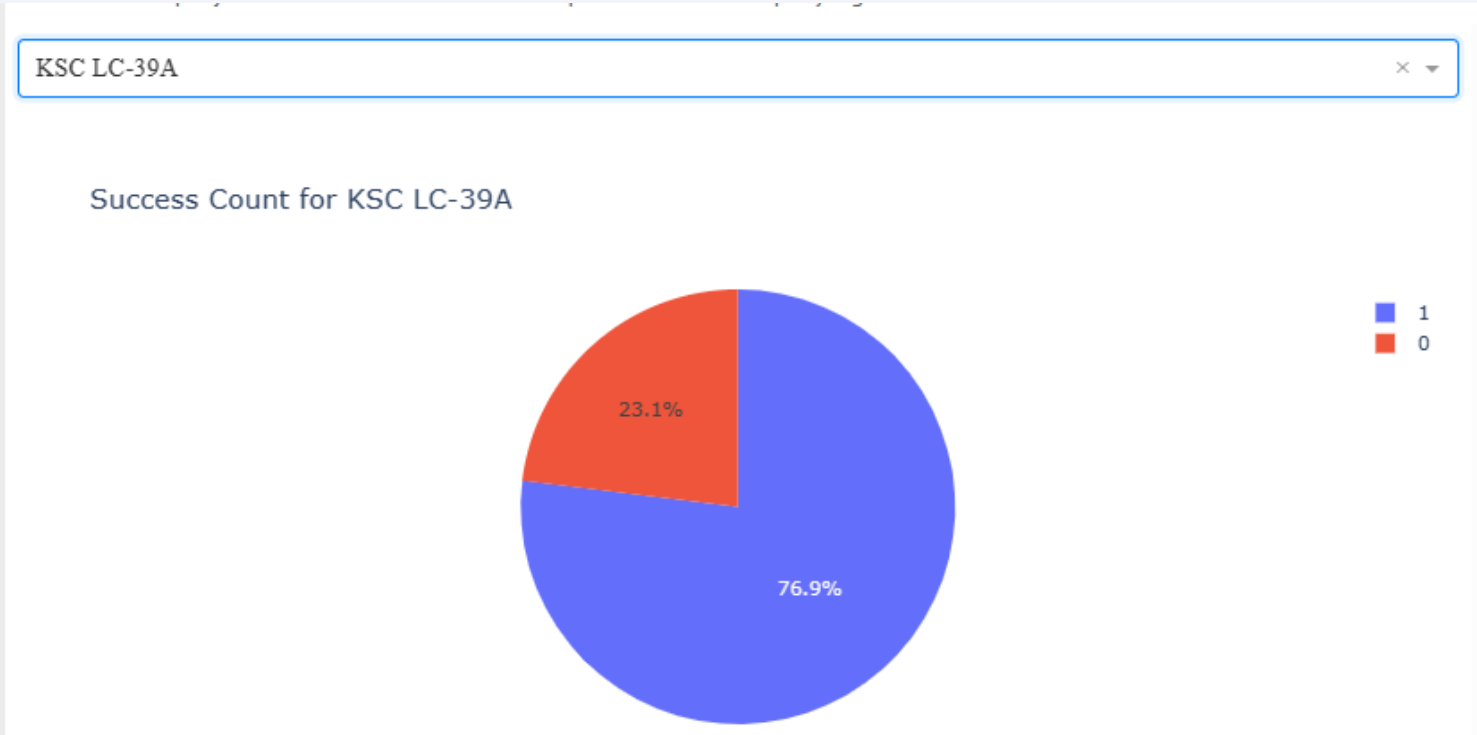

Section 4
Build a Dashboard with Plotly Dash

Success launch rate by site



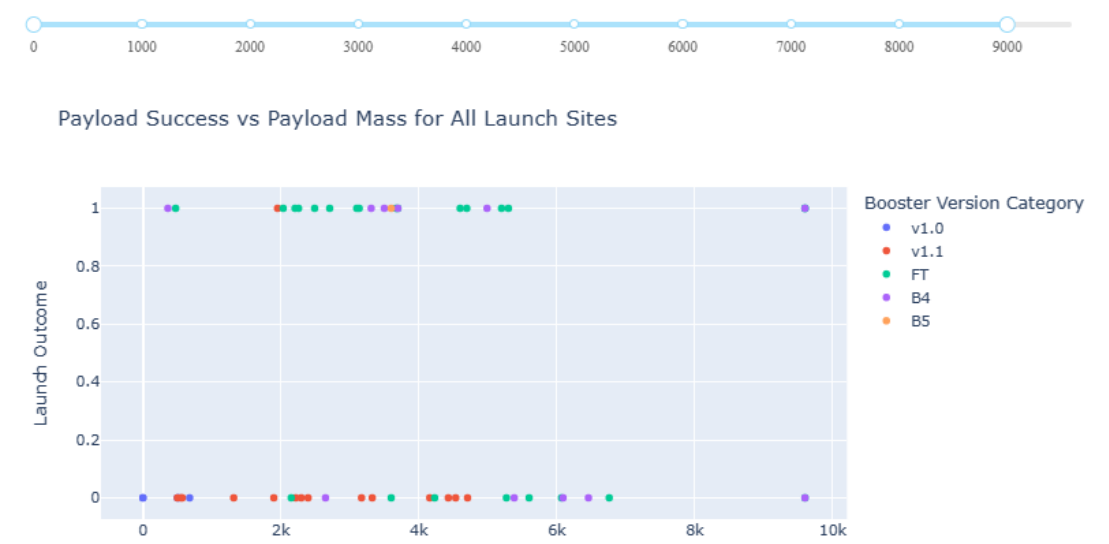
- KSC LC-39A has the highest success rate.
- CCAPS SLC-40 has the smallest success rate. The small sample or difficulty in launch from west cost may explain this.

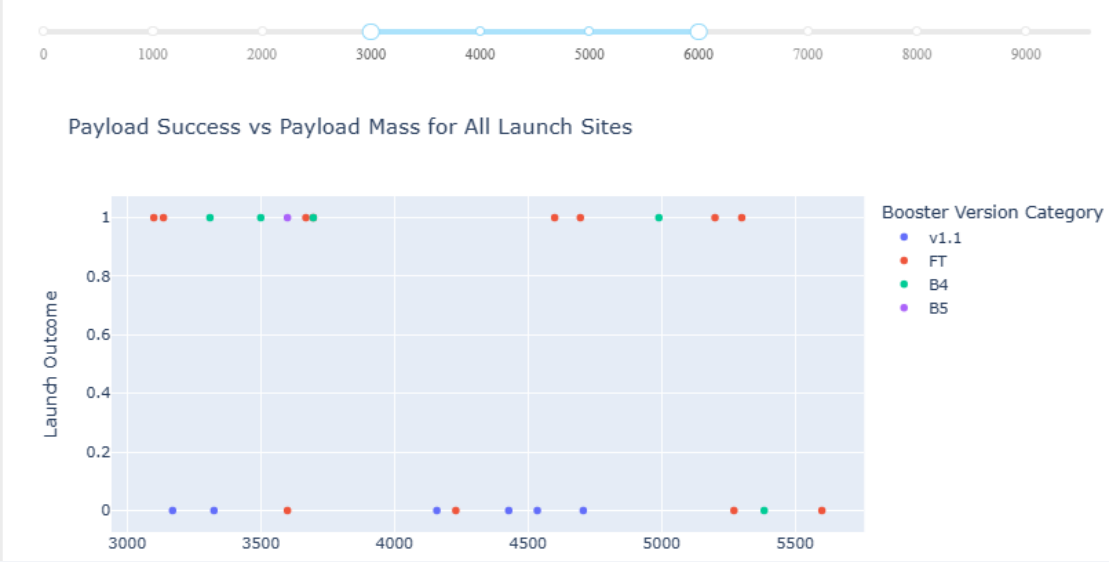
Highest success rate site



- KSC LC-39A has the highest success rate around 77%

Payload success vs Payload mass for all sites



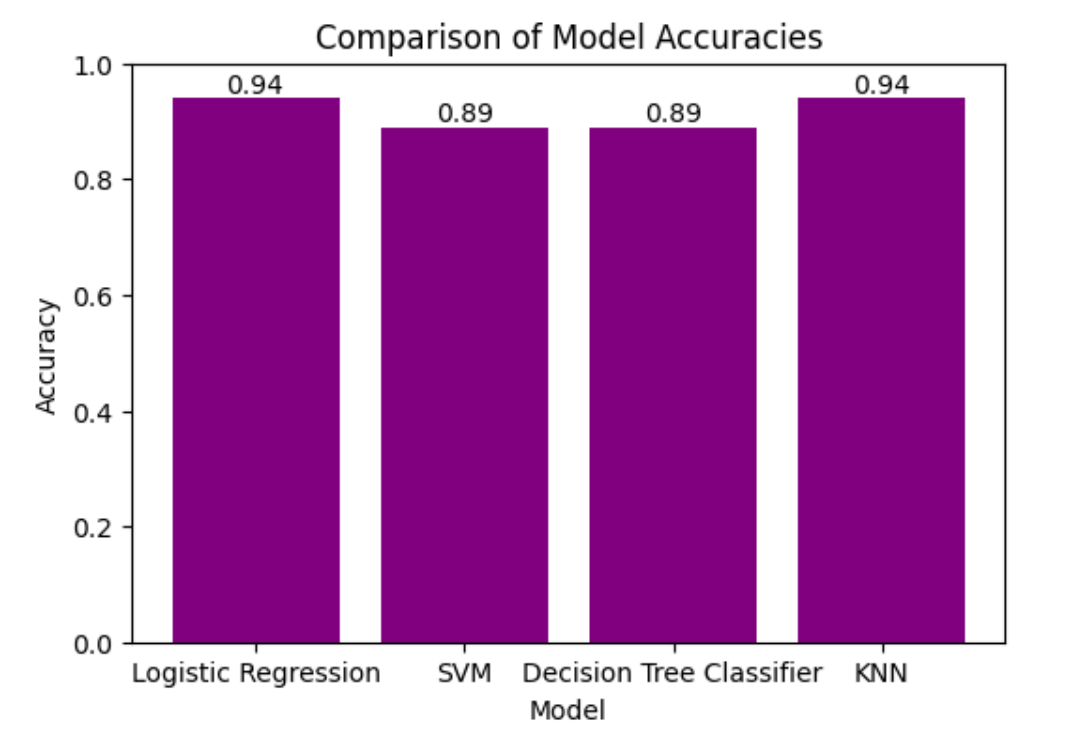


- Launch outcome equals 0 means failure and launch outcome equals 1 means success.
- The booster version is shown by the colored points.

Section 5

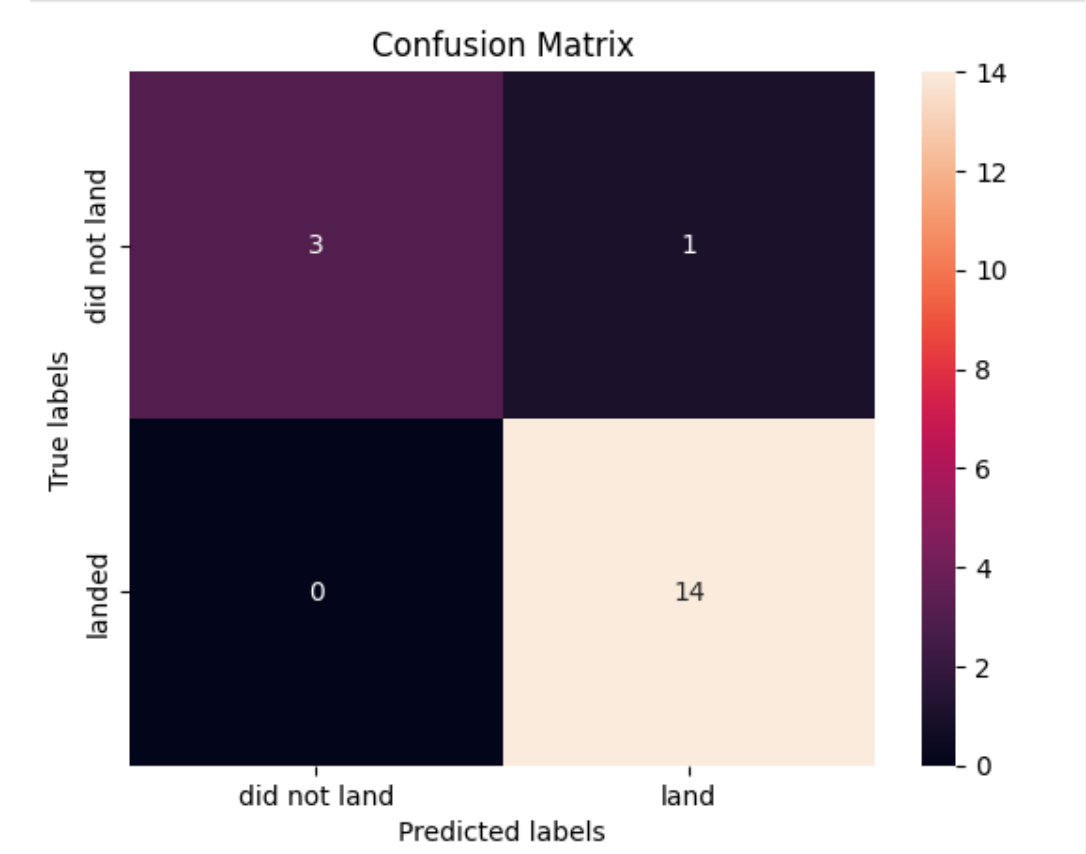
Predictive Analysis (Classification)

Classification Accuracy



- Logistic Regression and KNN have the same accuracy.
- It should be noted that test size is small at only sample size of 18.
- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.

Confusion Matrix



- The correct predictions are in diagonal from top left to bottom right.
- Just 1 prediction was wrong, when predicted that it would land and actually did not land.
- Our models over predict successful landings.

Conclusions

- Our primary objective involved the intricate development of a sophisticated machine learning model, meticulously designed to forecast the success of the Stage 1 landing for Space Y. In pursuit of comprehensive communication of our discoveries, we ingeniously crafted an insightful dashboard.
- Remarkably, the predictive capabilities of our model manifested with an astonishing accuracy rate of 94%. However, in the pursuit of continuous enhancement, we advocate for the collection of additional data. This augmentation in data acquisition aims to meticulously refine our model selection process and further elevate the overall predictive accuracy.
- The implications of our endeavors extend beyond the mere realm of technical achievements. The prowess of our model empowers SpaceX to foresee, with unparalleled precision, whether a launch will culminate in a successful Stage 1 landing prior to its initiation. This strategic foresight not only informs decision-making processes but also confers a substantial commercial advantage upon Space Y.

Appendix

- GitHub repository url:
- jackniet987/ibm-data-science: Repositório referente ao curso de Data Science da IBM (github.com)
- All prints come from jupyter notebooks, python files available in github repository.

Thank you!