IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

José Carlos Cimorra
10/11/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Web scraping and SpaceX API to retrieve data
  - Data wrangling
  - Exploratory Data Analysis (SQL, Data Visualization and Folium)
  - Data Wrangling
- Summary of all results
  - With the data collected, EDA was used to see what was the best combination of characteristics to predict launchings, and with ML, what was the best model.

Introduction

- Project background and context

In this capstone, SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. The goal is to predict if the Falcon 9 first stage will land successfully.

- Problems you want to find answers

Factors and interactions between them that determine if the first stage will land successfully

Section 1

Methodology

Methodology

Executive Summary

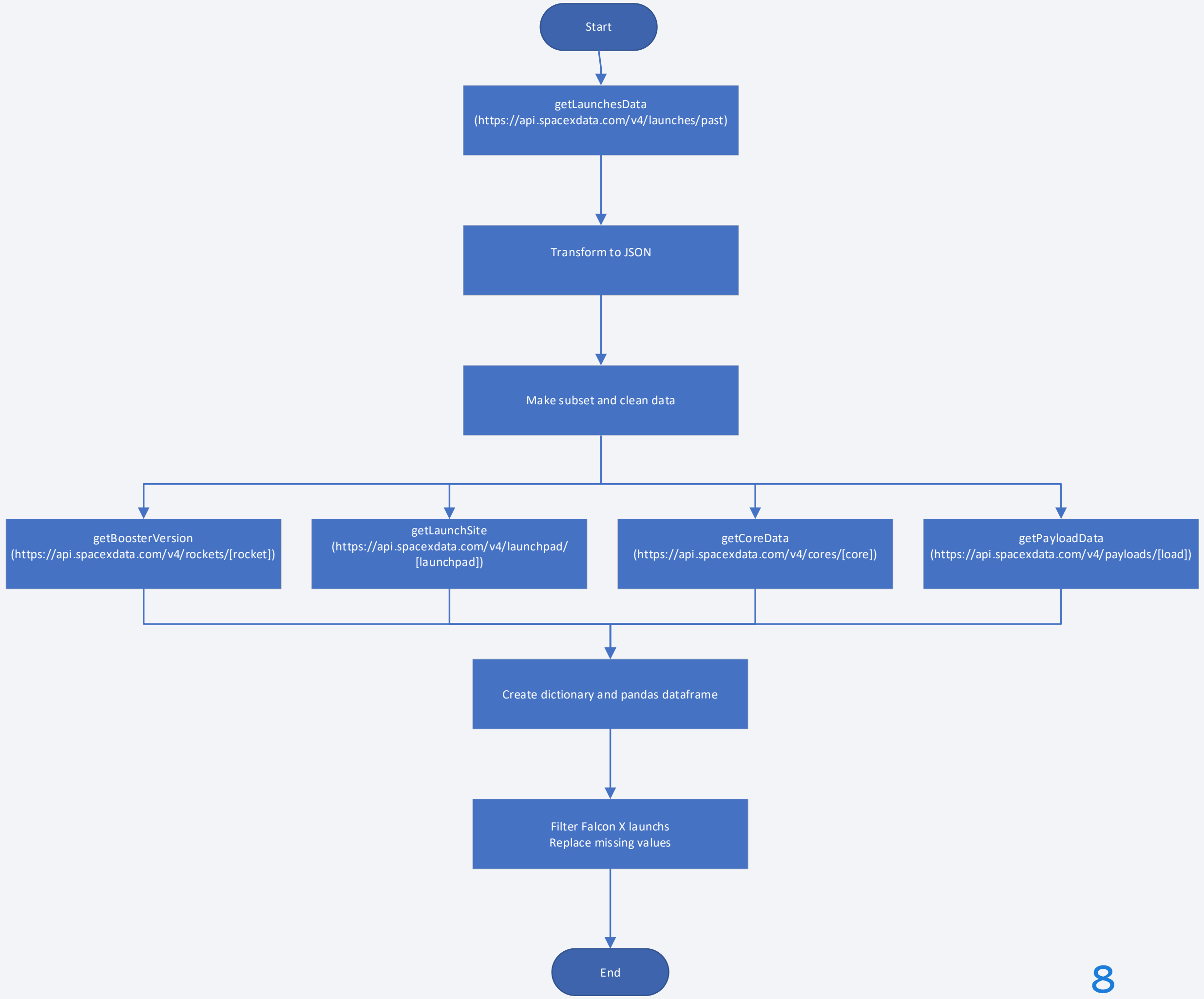
- Data collection methodology:
  - Data collected with web scraping of Wikipedia and SpaceX API
- Perform data wrangling
  - Data wrangling applying one hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

1. SpaceX API
    1. GET request to the SpaceX API
    2. Decode response as a JSON
    3. Transform response into a Pandas dataframe
    4. Clean data: search missing values and complete if necessary
2. Webscraping
    1. Websc scraping of Falcon 9 launch records from Wikipedia
    2. Parse HTML table
    3. Transform table into a Pandas dataframe

Data Collection – SpaceX API

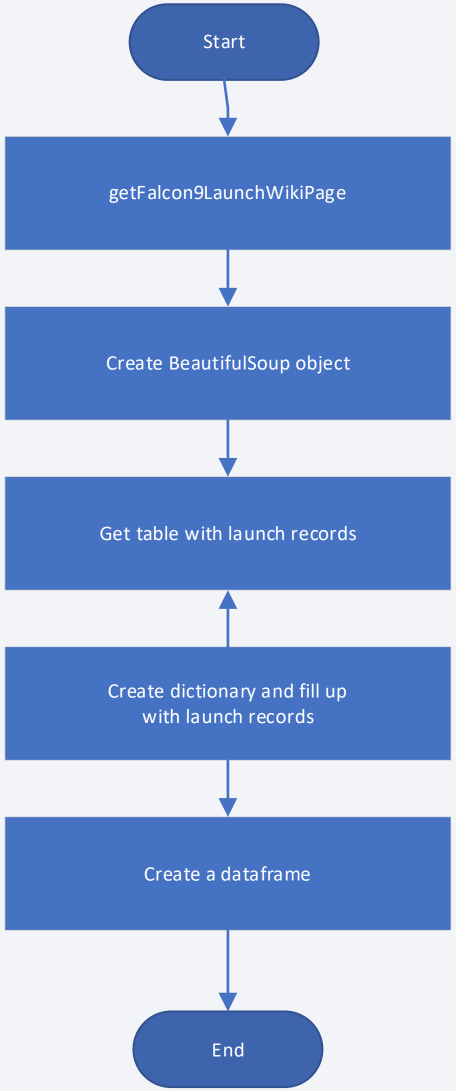
- The flowchart represents the process of data collection using SpaceX APIs and the creation of a pandas dataframe at the end
- Link to the github file here



Start
getLaunchesData
(https://api.spacexdata.com/v4/launches/past)
Transform to JSON
Make subset and clean data
getBoosterVersion
(https://api.spacexdata.com/v4/rockets/[rocket])
getLaunchSite
(https://api.spacexdata.com/v4/launchpad/[launchpad])
getCoreData
(https://api.spacexdata.com/v4/cores/[core])
getPayloadData
(https://api.spacexdata.com/v4/payloads/[load])
Create dictionary and pandas dataframe
Filter Falcon X launchs
Replace missing values
End

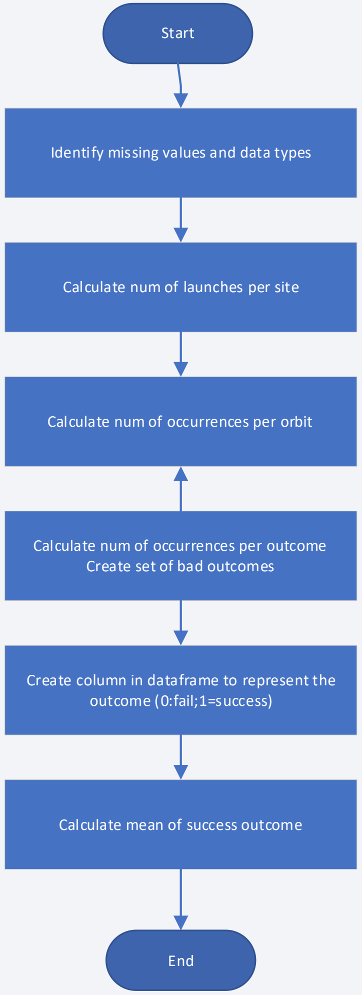
Data Collection - Scraping

- The flowchart represents the process of data collection using web scraping of the Falcon9 launch records page in Wikipedia and the creation of a pandas dataframe at the end
- Link to the github file here


Start
getFalcon9LaunchWikiPage
Create BeautifulSoup object
Get table with launch records
Create dictionary and fill up with launch records
Create a dataframe
End

Data Wrangling

- The flowchart represents the process of data wrangling using the result dataframe from last step.
- The final step determines the success rate of the rocket's first-stage landing
- Link to the github file here


Start
Identify missing values and data types
Calculate num of launches per site
Calculate num of occurrences per orbit
Calculate num of occurrences per outcome
Create set of bad outcomes
Create column in dataframe to represent the outcome (0:fail;1=success)
Calculate mean of success outcome
End

EDA with Data Visualization

- Scatter plots were used to visualize outcome depending on combinations:
  - FlightNumber vs Payload
  - LaunchSite vs FlightNumber
  - LaunchSite vs PayloadMass
  - Orbit vs FlightNumber
  - Orbit vs PayloadMass
- Bar chart was used to visualize outcome depending on the orbit
- They all were used to see the relationships of several factors in the success rate
- Also a line chart was used to see the average launch success yearly trend
- Link to the github file here

EDA with SQL

- These are the different calculations performed in the SQL EDA process with SQL queries:
  - Names of the unique launch sites
  - 5 records where launch sites begin with 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first succesful landing outcome in ground pad was achieved.
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Total number of successful and failure mission outcomes
  - Names of the booster_versions which have carried the maximum payload mass.
  - Records (month names, failure landing_outcomes in drone ship ,booster versions, launch_site) for the months in year 2015.
  - Ranking of 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.
- Link to the github file here

Build an Interactive Map with Folium

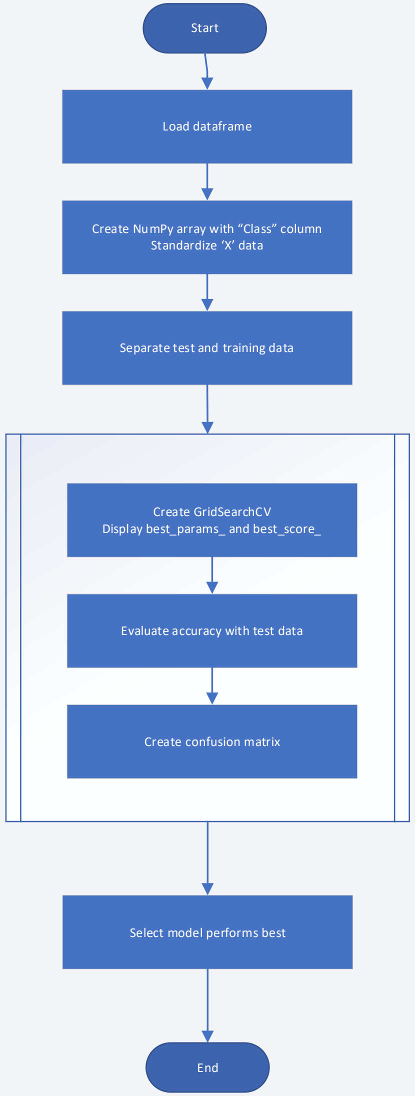
- Map objects added to the folium map
  - **Circles** to locate coordinates
  - **Popups to show coordinates description (i.e.: NASA Johnson Space Center)**
  - **Markers** to show labels with the coordinates names near the circles (and other exercises later like distance between a selected railway and a selected launch site)
  - **MarkerCluster** to add Markers with launch records (can have same coordinates)
  - **MousePosition** to show coordinates (top-right) of the mouse position
  - **PolyLine** to draw a line between a launch site and selected coastlines and railways
- Link to the github file here

Build a Dashboard with Plotly Dash

- Plots and graphs added to the dashboard
    - **Drop-down** to select the launch sites
    - **Pie chart** to show the success rate in the launch site selected
    - **Range slider** to select the payload mass
    - **Sclatter chart** to show the correlation between the payload mass and success rate
- Link to the github file here

Predictive Analysis (Classification)

- The flowchart represents the process of the predictive analysis building and evaluating classification models in the subprocess shown:
  - LogisticRegression
  - SVC
  - DecissionTreeClassifier
  - KneighborsClassifier
- Link to the github file here



Start

Load dataframe

Create NumPy array with “Class” column
Standardize ‘X’ data

Separate test and training data

Create GridSearchCV
Display best_params_ and best_score_

Evaluate accuracy with test data

Create confusion matrix

Select model performs best

End

Results

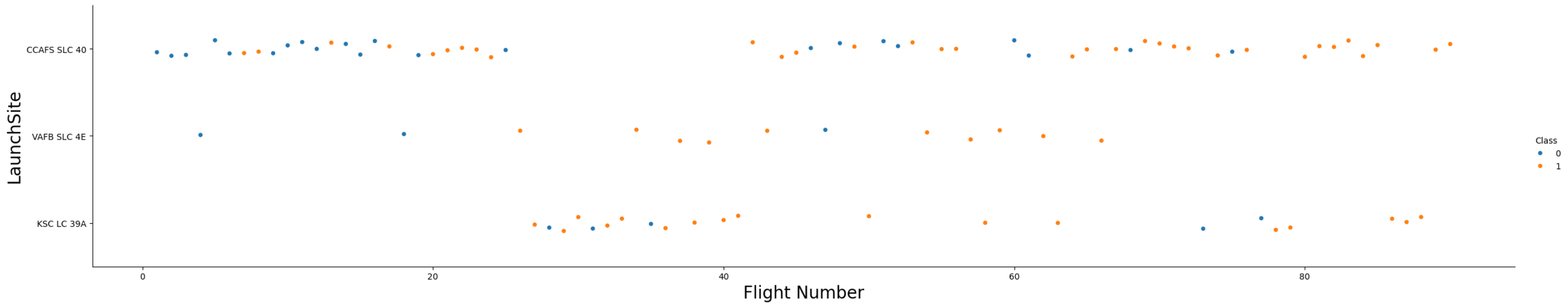
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

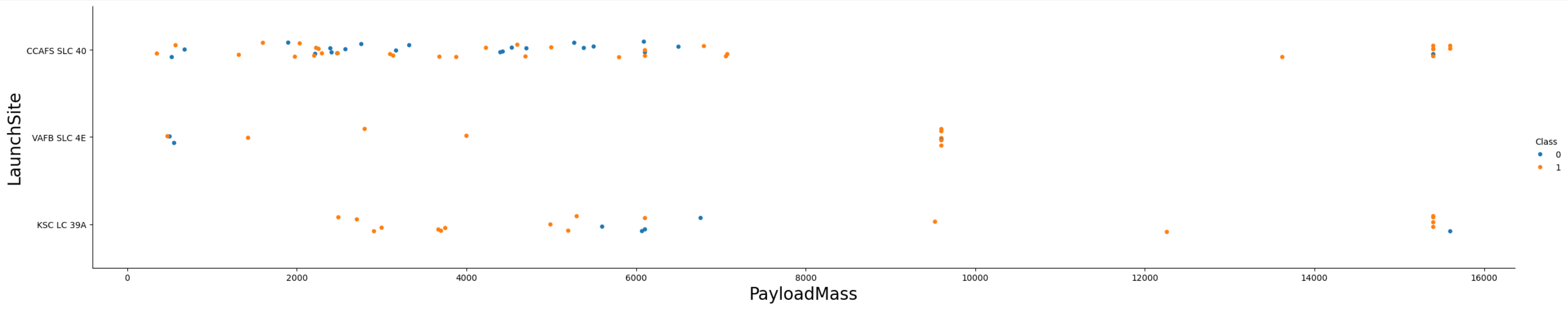
Flight Number vs. Launch Site

- CCAFS SLC 40 highest number of rockets launch tan the others, poorest results
- VAFB SLC 4E lowest number of rochets launch
- VAFB SLC 4E and KSC LC 39A launch sites have high success rates than CCAFS SLC 40



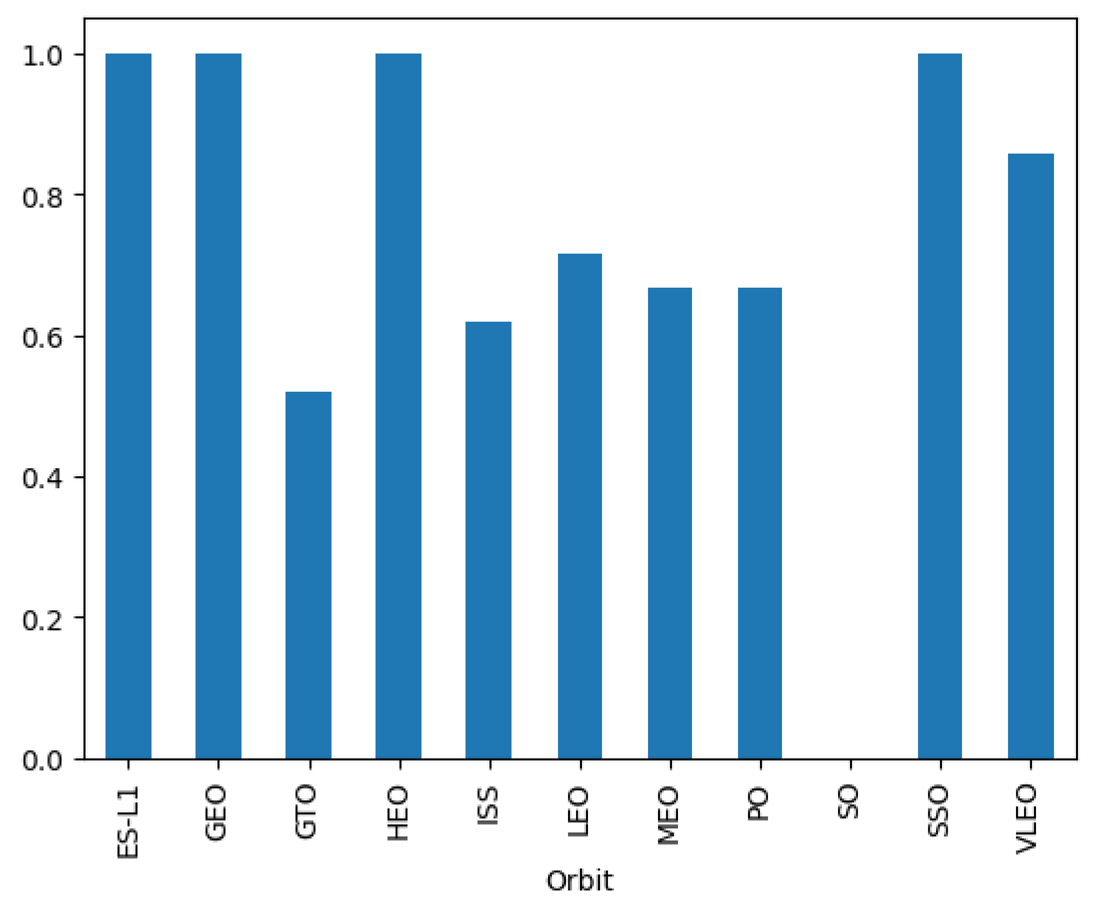
Payload vs. Launch Site

- Most of the rockets launched in all sites have less tan 9000kg payload mass
- CCAFS SLC 40 performs better with Payload mass over 7500kg
- VAFB SLC 4E performs better over 1000kg payload mass but has no launches over 10000 Kg
- KSC LC 39ª has better results with payloads under 6000Kg



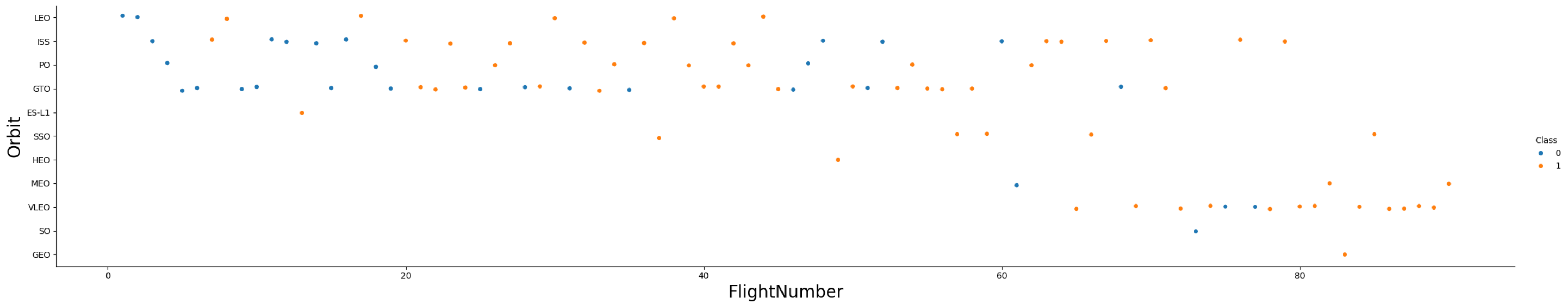
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO orbits have 100% success rate
- GTO has the lowest success rate



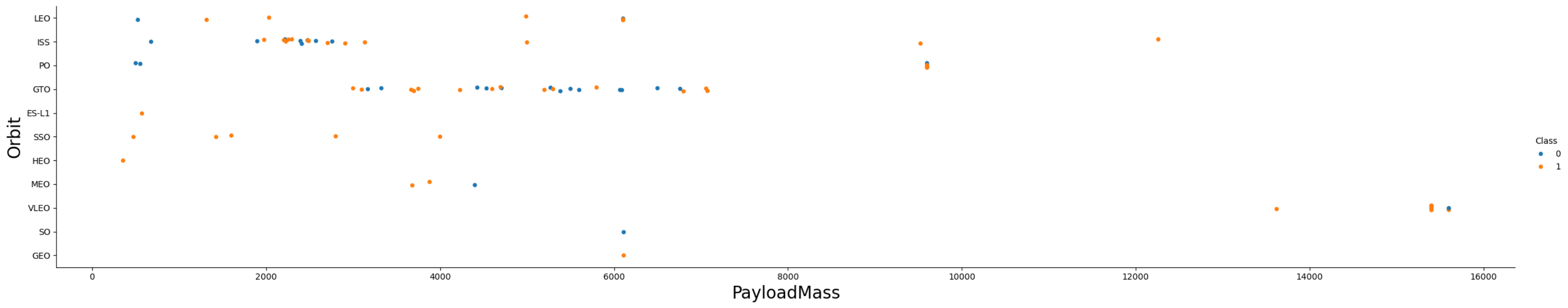
Flight Number vs. Orbit Type

- LEO, ISS, PO, GTO and VLEO were the orbits most used for rocket launches
- Rest of orbits were rarely used



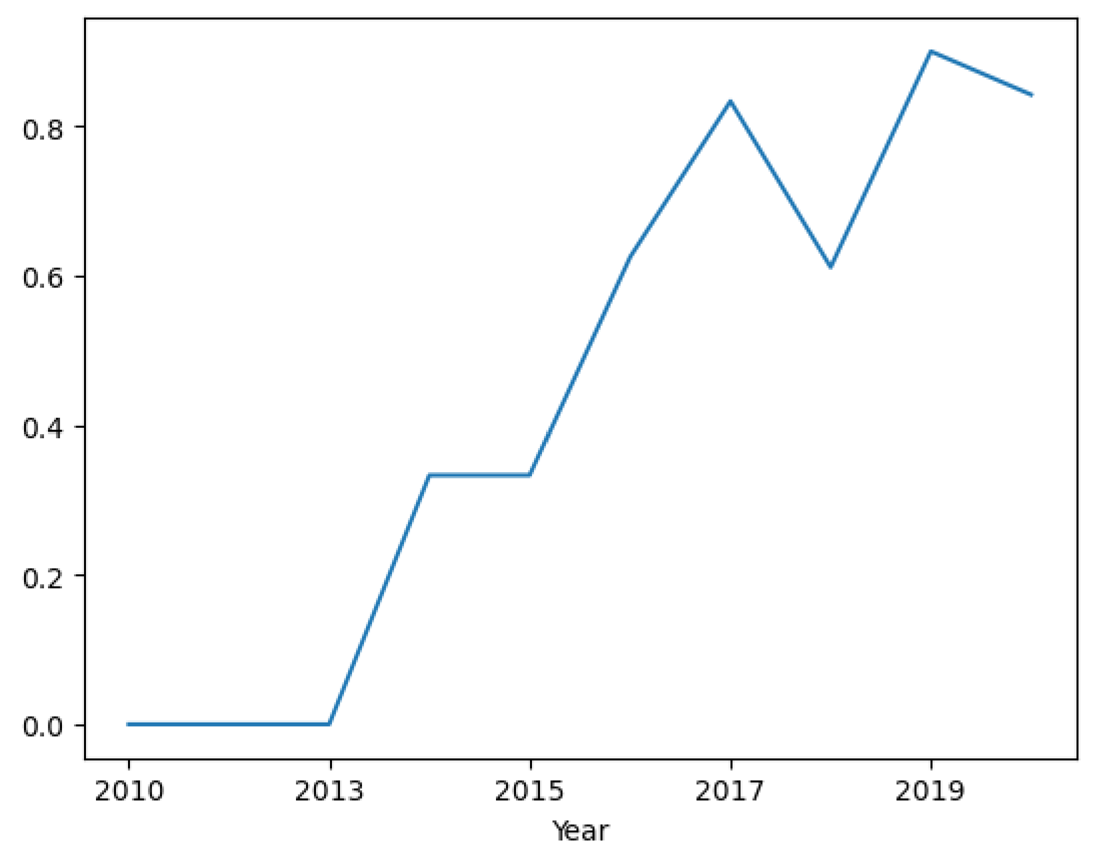
Payload vs. Orbit Type

- Higher success with higher payloads in LEO, ISS and PO
- ES-LI, SSO, HEO and MEO have 100% success with payloads under 4000Kg
- GTO appears not to have correlation of success rate and payload



Launch Success Yearly Trend

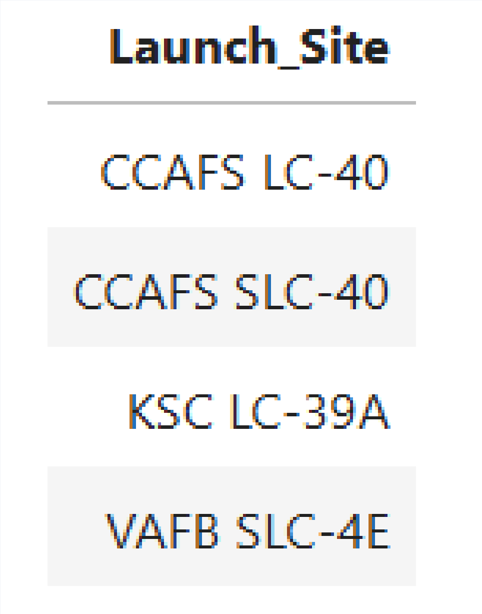
- Trend is increasing from 2013
- Decreased in 2018 but recovered in 2019
- Small decrease again in 2020



All Launch Site Names

- Unique launch sites names retrieved from database with this simple query:

```
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```



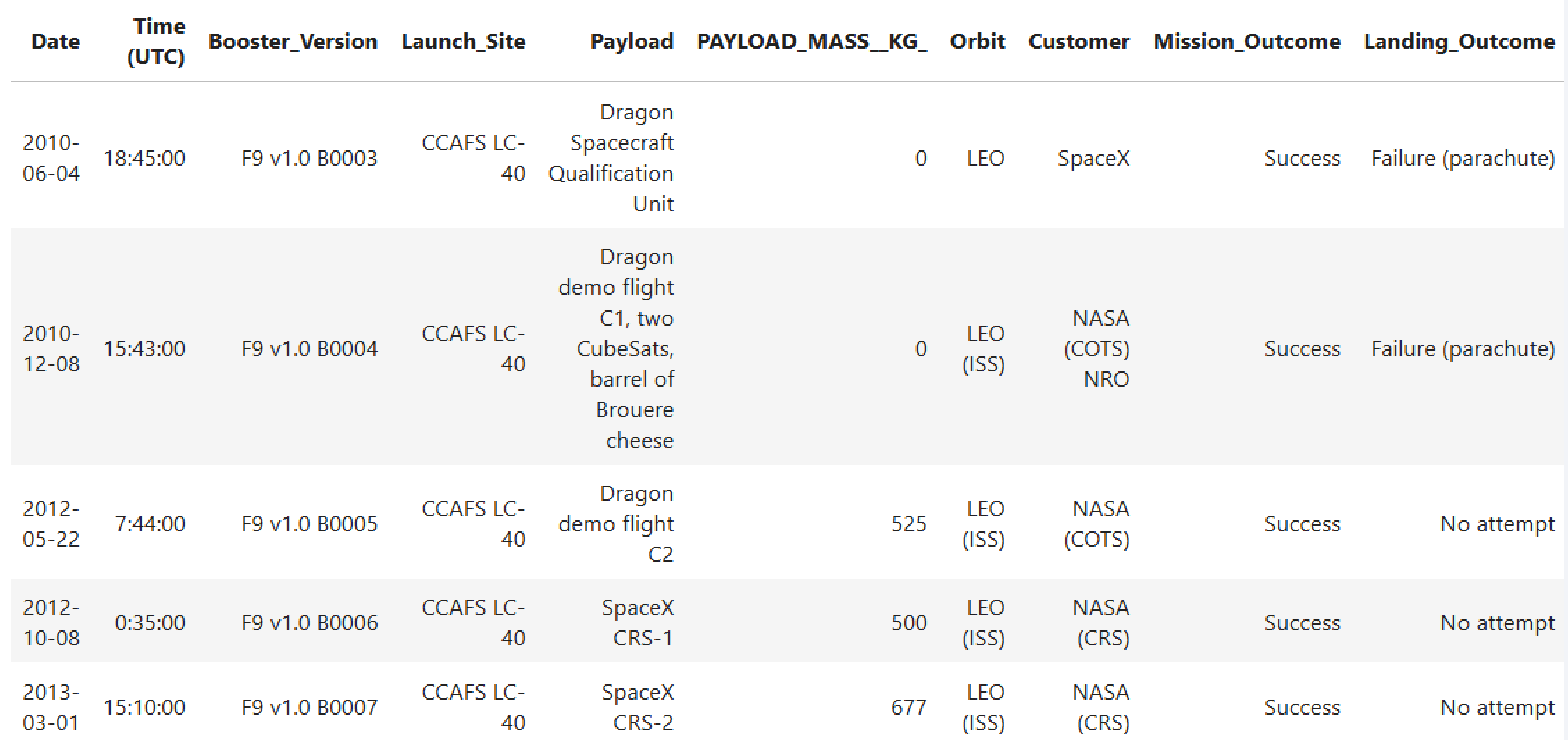
Launch_Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA’ (I)

- 5 records where launch sites begin with `CCA` retrieved from database with this simple query:

```
SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Launch Site Names Begin with 'CCA' (II)

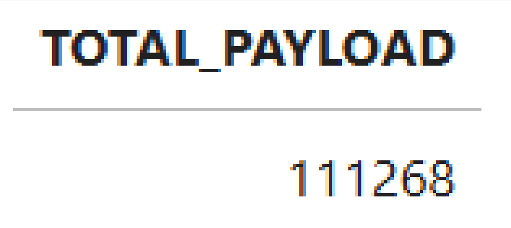


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

- Total Payload Mass carried by boosters launched by NASA (CRS) retrieved from database with this simple query:

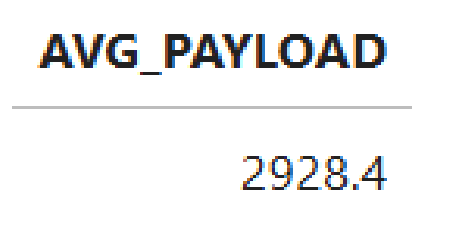
```
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD
FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';
```



TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1retrieved from database with this simple query:

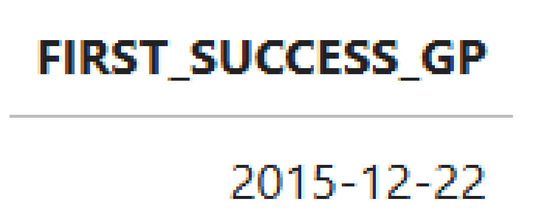


AVG_PAYLOAD
2928.4

```
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD
FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad retrieved from database with this simple query:



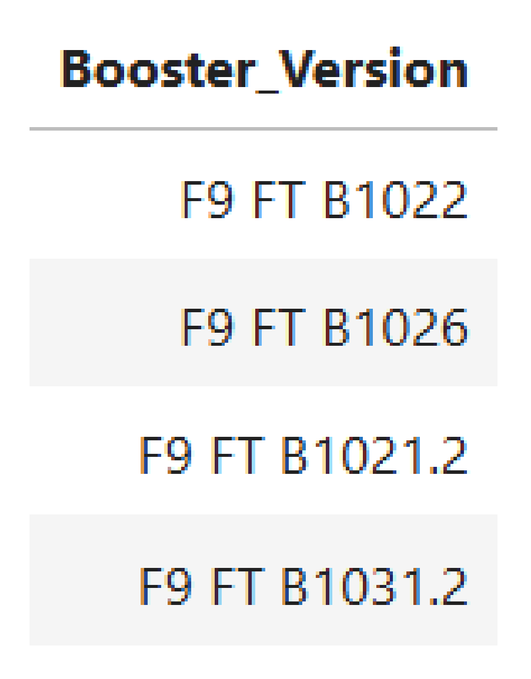
FIRST_SUCCESS_GP
2015-12-22

```
SELECT MIN(DATE) AS FIRST_SUCCESS_GP
FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 retrieved from database with this simple query:

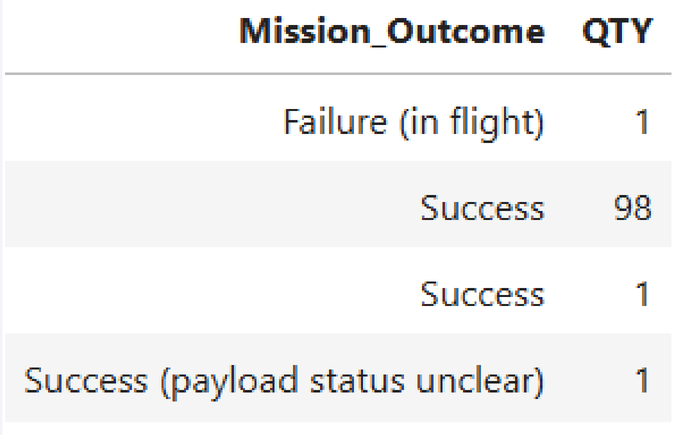
```
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
AND LANDING_OUTCOME = 'Success (drone ship)';
```



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

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes retrieved from database with this simple query:



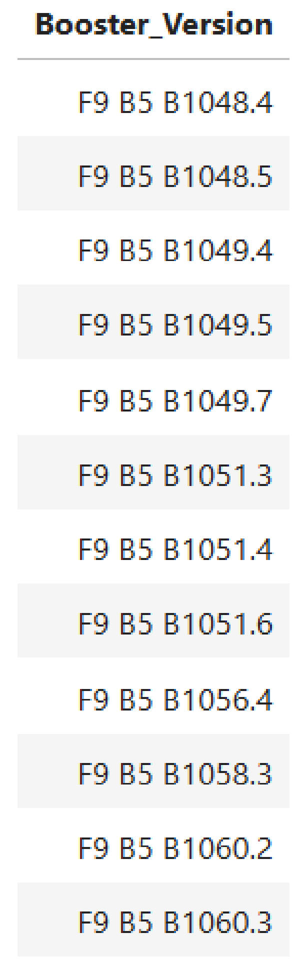
Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

```
SELECT MISSION_OUTCOME, COUNT(*) AS QTY
FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass retrieved from database with this simple query (using a subquery):

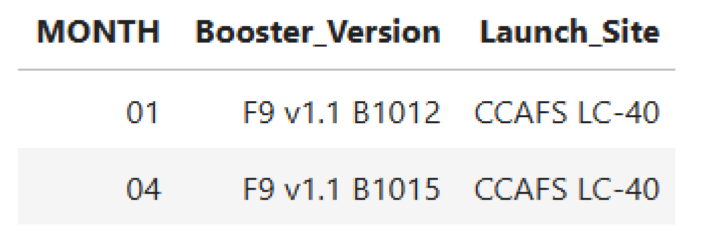
```
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)ORDER BY BOOSTER_VERSION;
```



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

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 retrieved from database with this simple query:



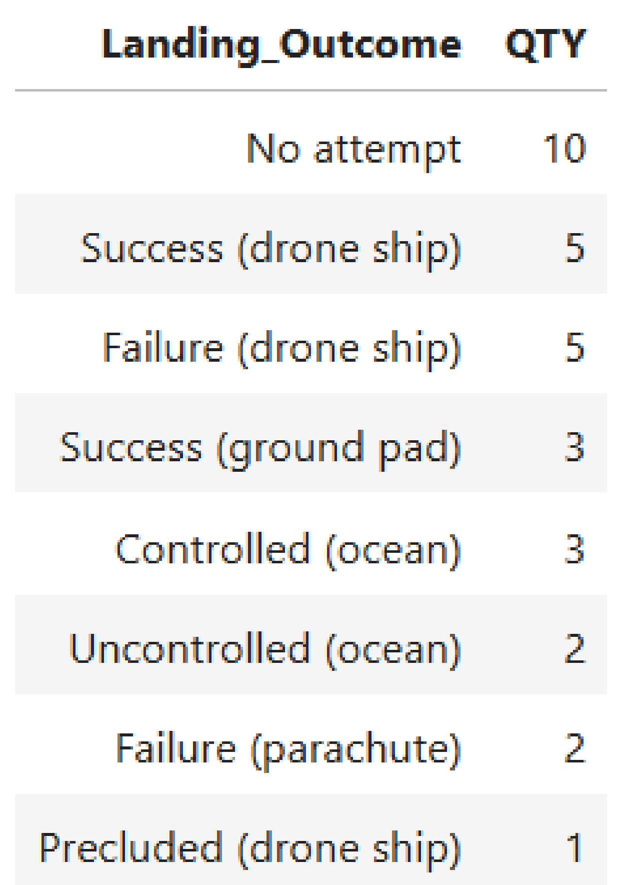
MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

```
SELECT substr(Date, 6,2) as MONTH, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)’
AND substr(Date,0,5) = '2015';
```

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

- Ranking of 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 retrieved from database with this simple query:

```
SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20’
GROUP BY LANDING_OUTCOME ORDER BY QTY DESC;
```



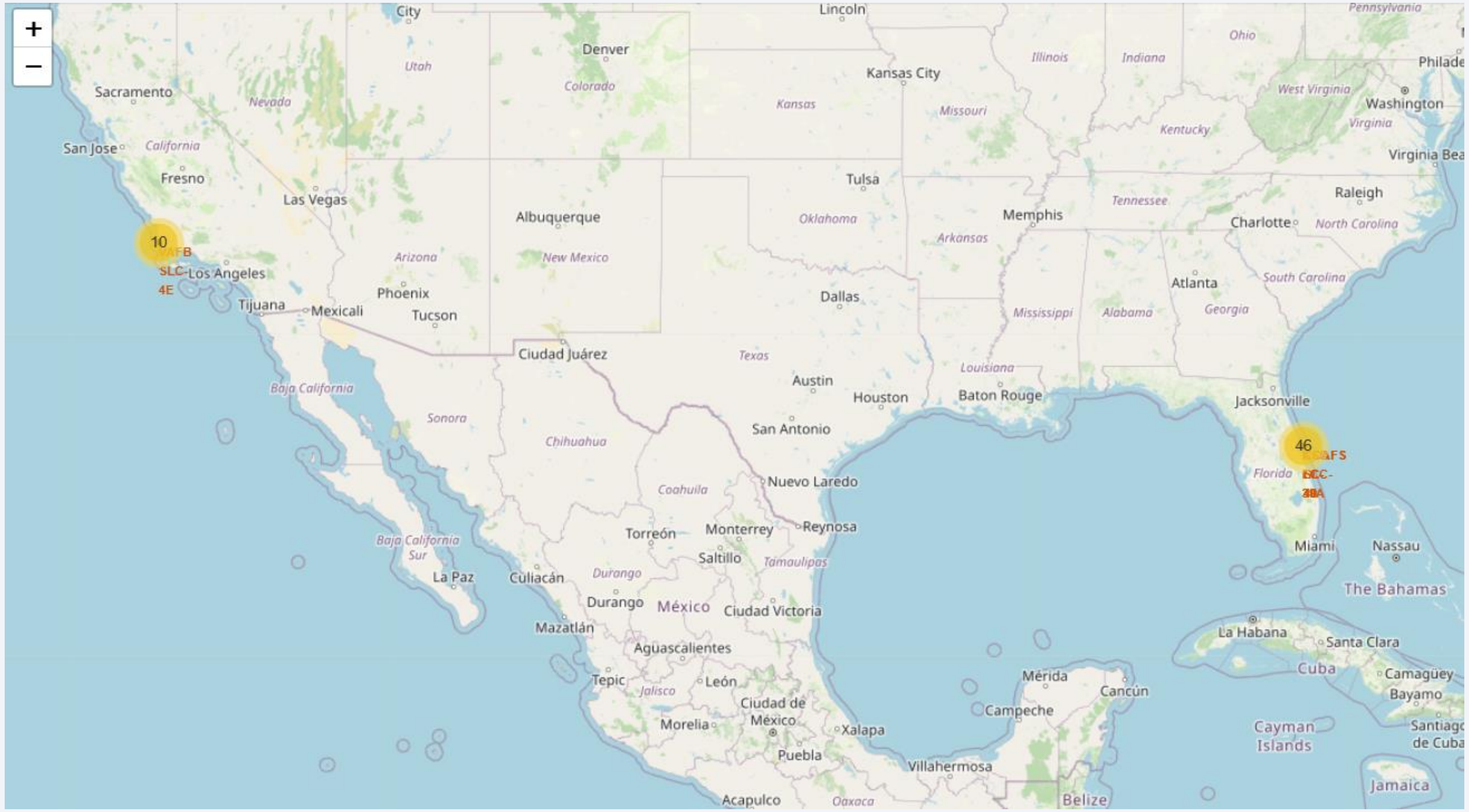
Landing_Outcome	QTY
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

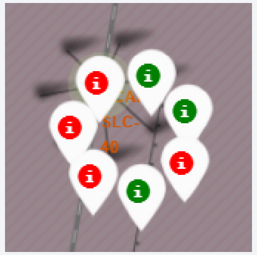
SpaceX Falcon 9 launch sites locations

- 10 of the launch sites were on the west coast of USA and the other 46 sites were on the east coast of the USA.

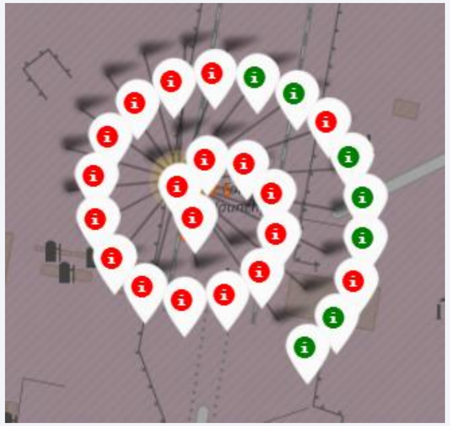


SpaceX Falcon 9 launch outcomes

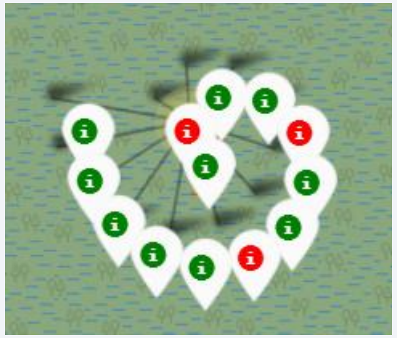
- CCAFS SLC-40



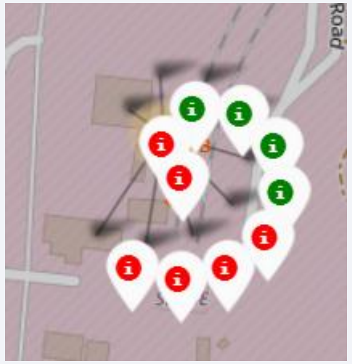
- CCAFS LC-40



- KSC LC-39A



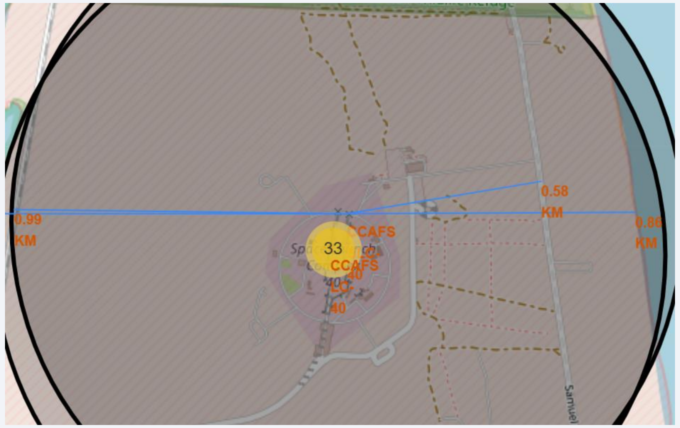
- VAFB SLC-4E



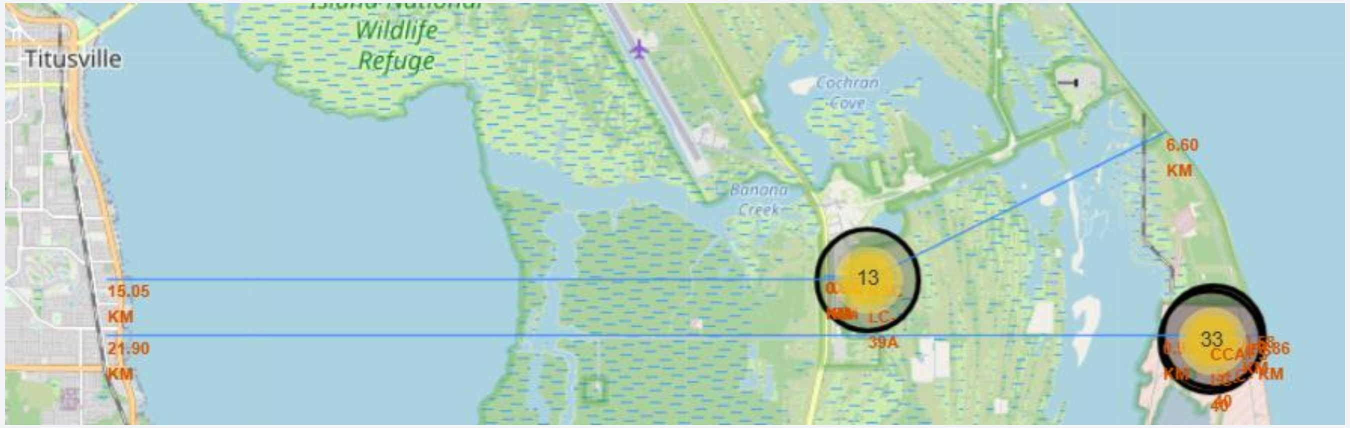
Green = successful; Red = failed

Launch site distances from different points (I)

- CCAFS SLC-40 & CCAFS LC-40



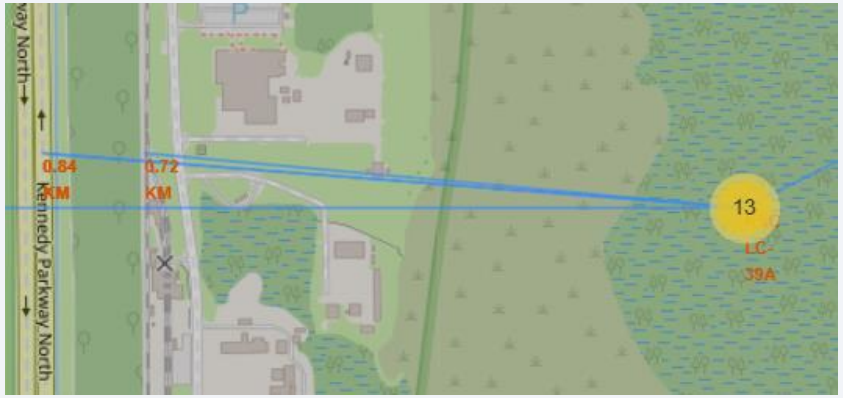
Coastline, railway and highway



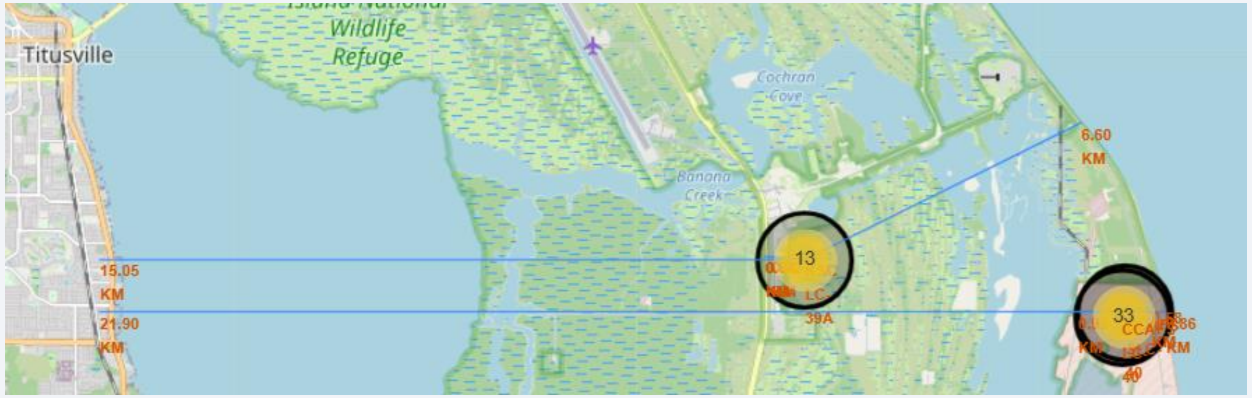
Nearest city (Titusville)

Launch site distances from different points (II)

- KSC LC-39A



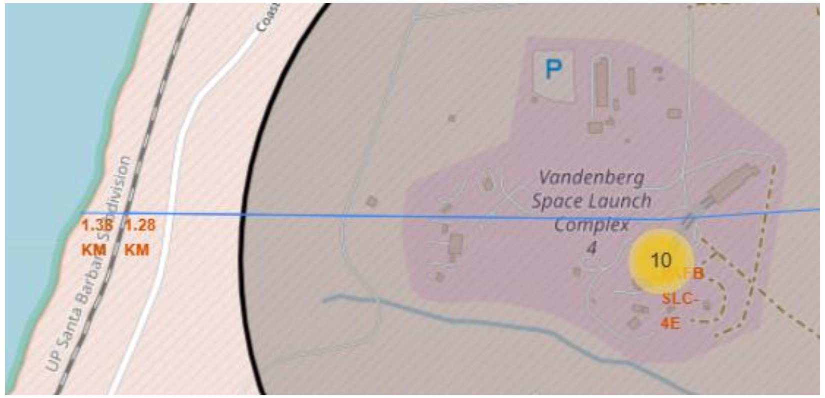
Railway and highway



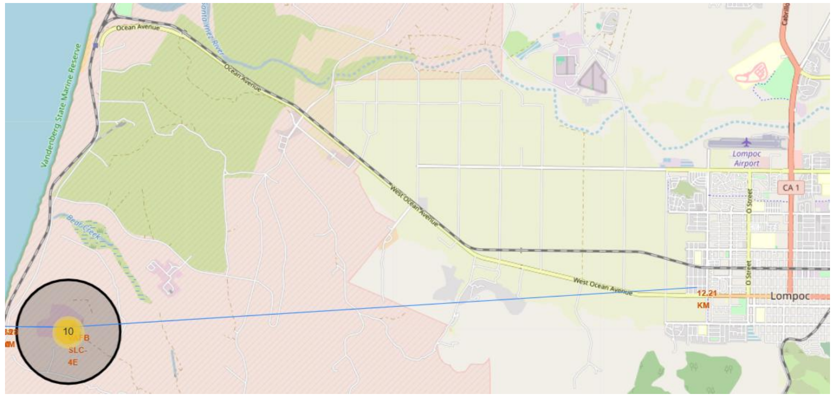
Coastline and nearest city (Titusville)

Launch site distances from different points (III)

- VAFB SLC-4E


P
Coast
Vandenberg Space Launch Complex 4
1.38 KM
1.28 KM
UP Santa Barbara Subdivision
10
VAFB SLC-4E

Railway and coastline



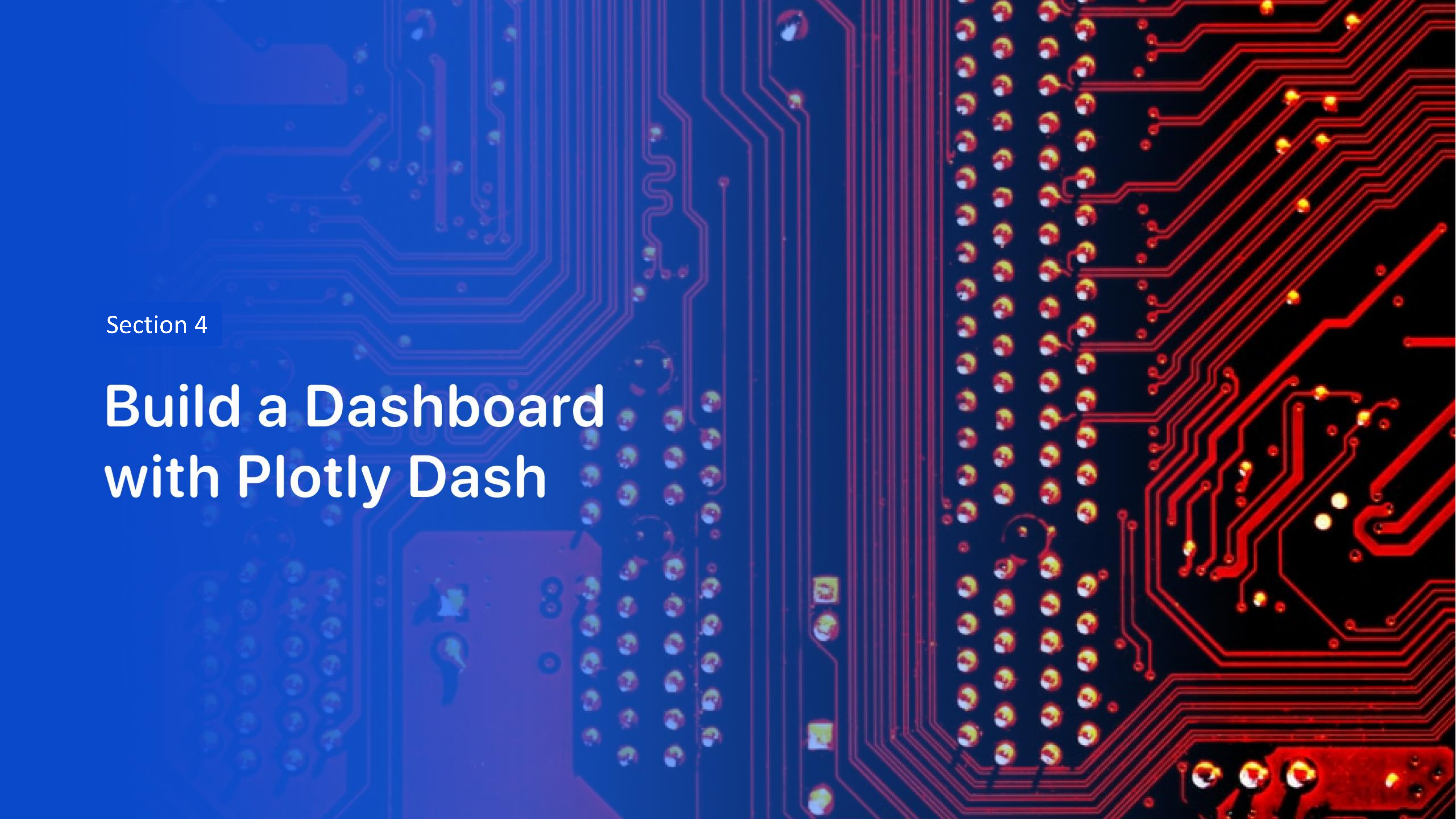
Nearest city (Lompoc)

Launch site distances from different points (IV)

- Launch sites close to the coast
  - CCAFS SLC-40 and CCAFS LC-40: near 0.9 km
  - VAFB SLC-4E: near 1.4 km
  - KSC LC-39A: near 6.6 km
- Nearest cities
  - CCAFS SLC-40 and CCAFS LC-40: near 21.9 km
  - VAFB SLC-4E: near 12.2 km
  - KSC LC-39A: near 15 km
- Nearest railways
  - CCAFS SLC-40 and CCAFS LC-40: near 1 km
  - VAFB SLC-4E: near 1.28 km
  - KSC LC-39A: near 0,72 km
- Nearest highways
  - CCAFS SLC-40 and CCAFS LC-40: near 0,58 km
  - VAFB SLC-4E: No highways near, but other roads
  - KSC LC-39A: near 0,84 km

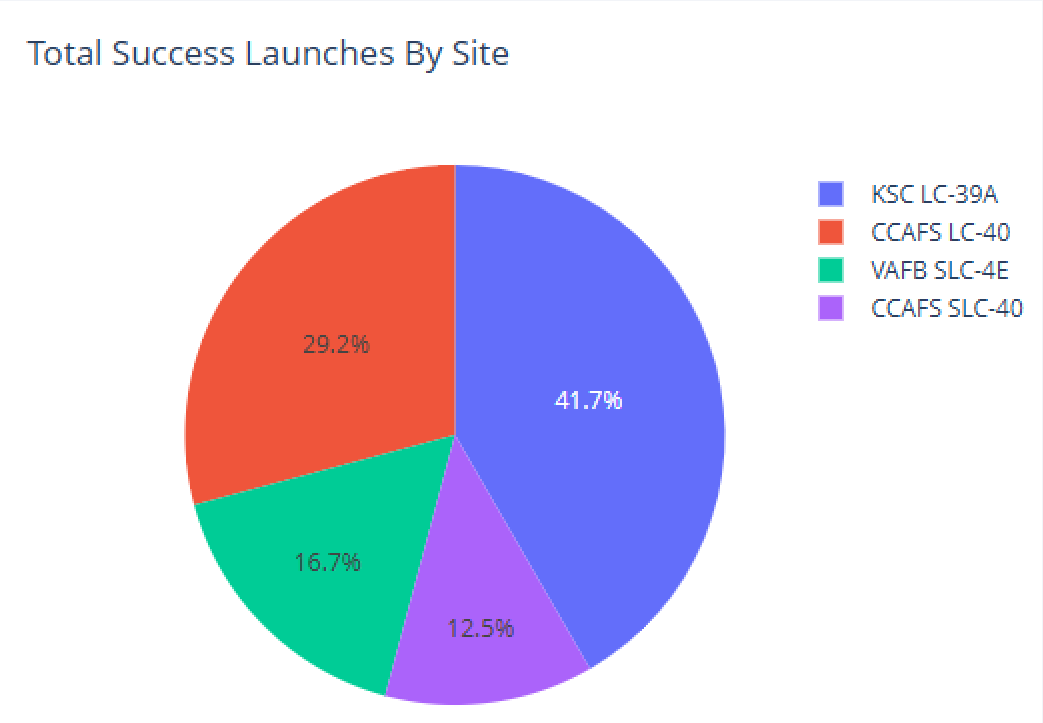
Section 4

Build a Dashboard with Plotly Dash



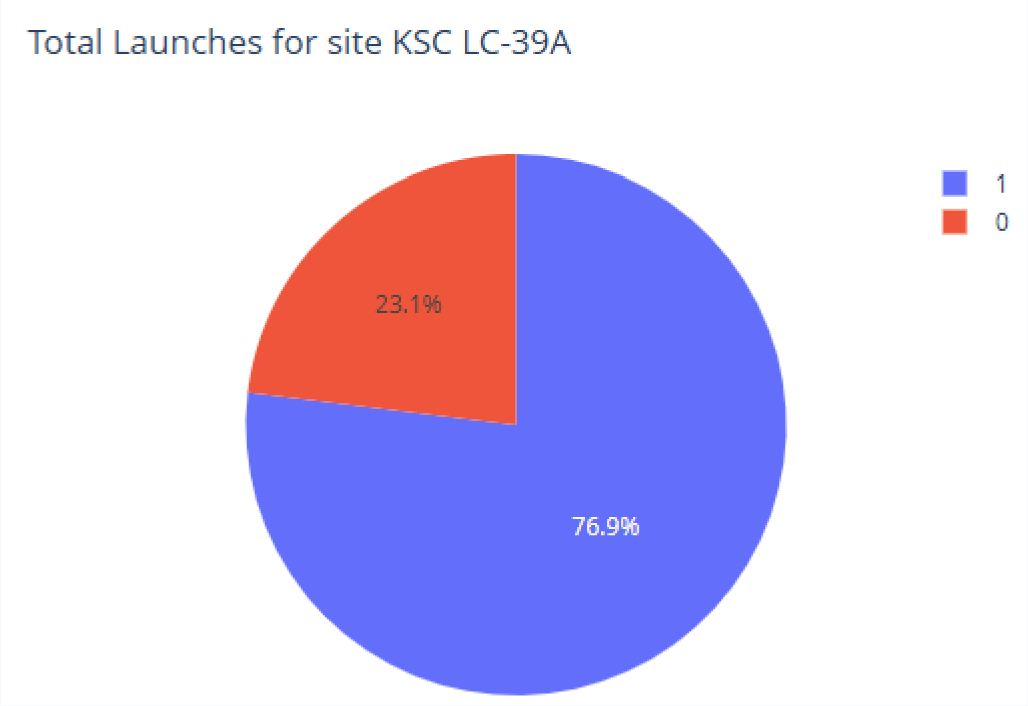
Launch success from all launch sites

- KSC LC-39A has the largest number of successful launches, with the 41,7% of the total successful launches



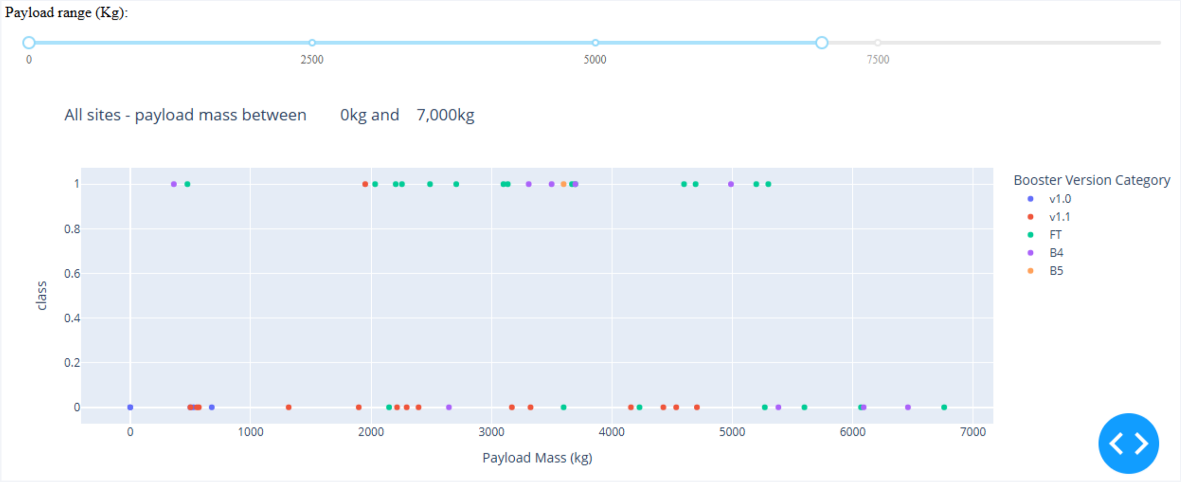
Launch site with highest launch success rate

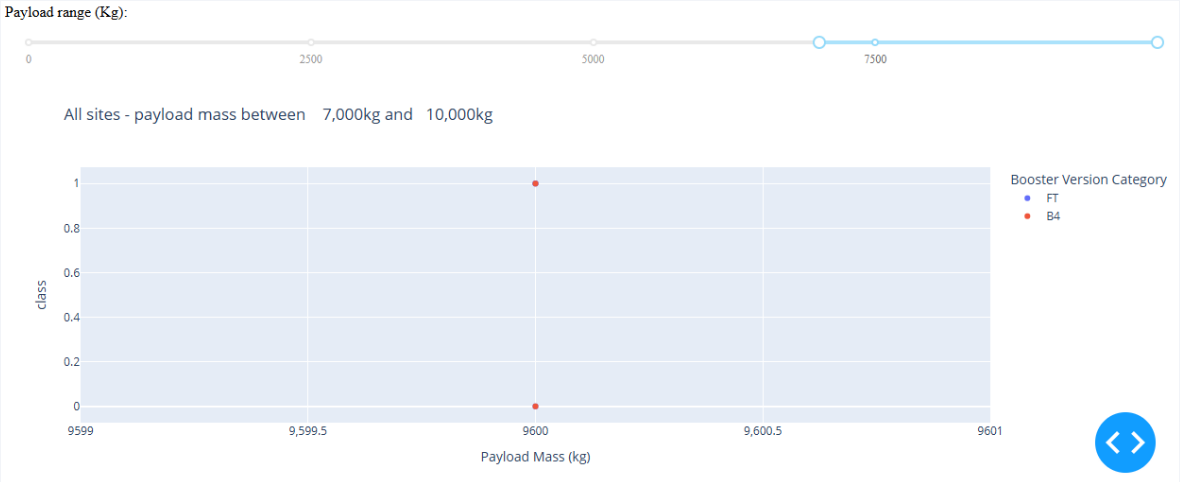
- KSC LC-39A has the highest success rate (76.9%), against CCAFS LC-40 (73,1%), VAFB SLC-4E (60%) and CCAFS LSC-40 (57.1%)



Payload vs. Launch Outcome for all sites

- Almost all of the launches have payload masses under 7.000Kg
- FT Booster has the highest success rate from 0 to 7.000Kg payload mass
- Rockets with payload masses over 5.000Kg have lower success rates



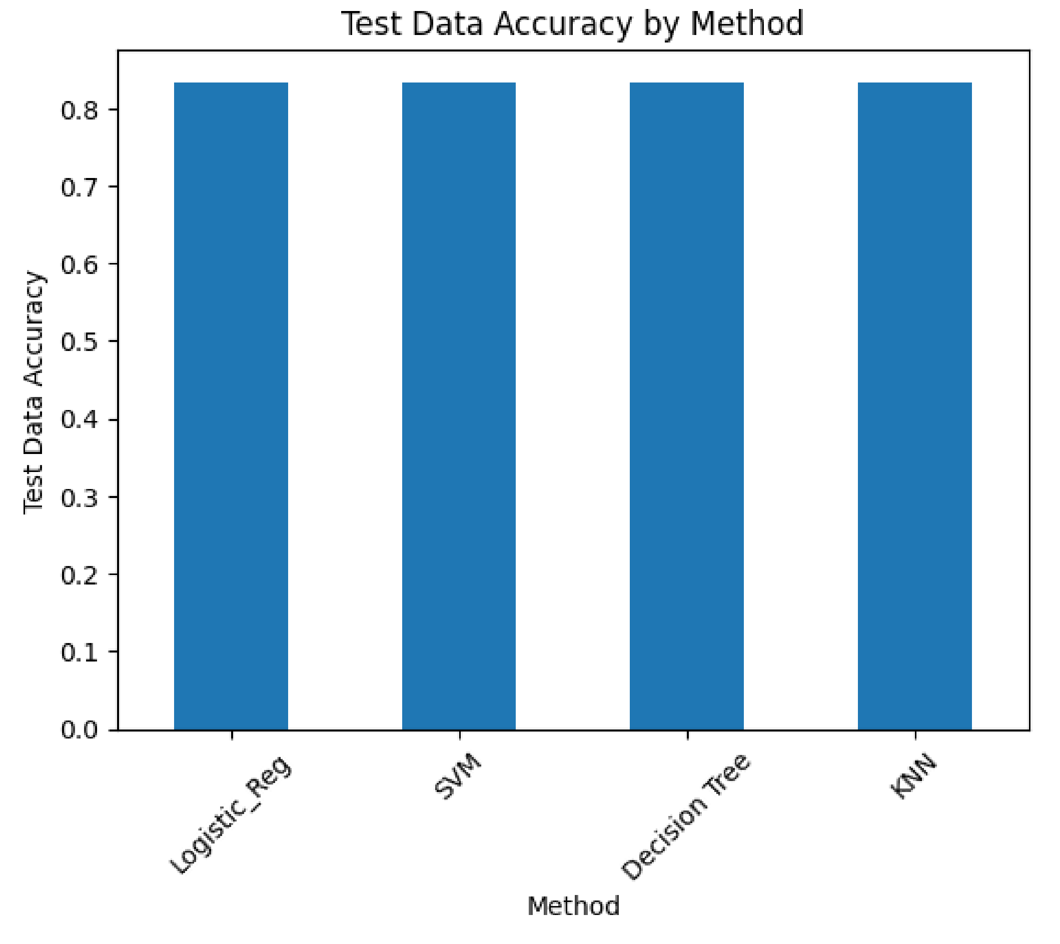


Section 5

Predictive Analysis (Classification)

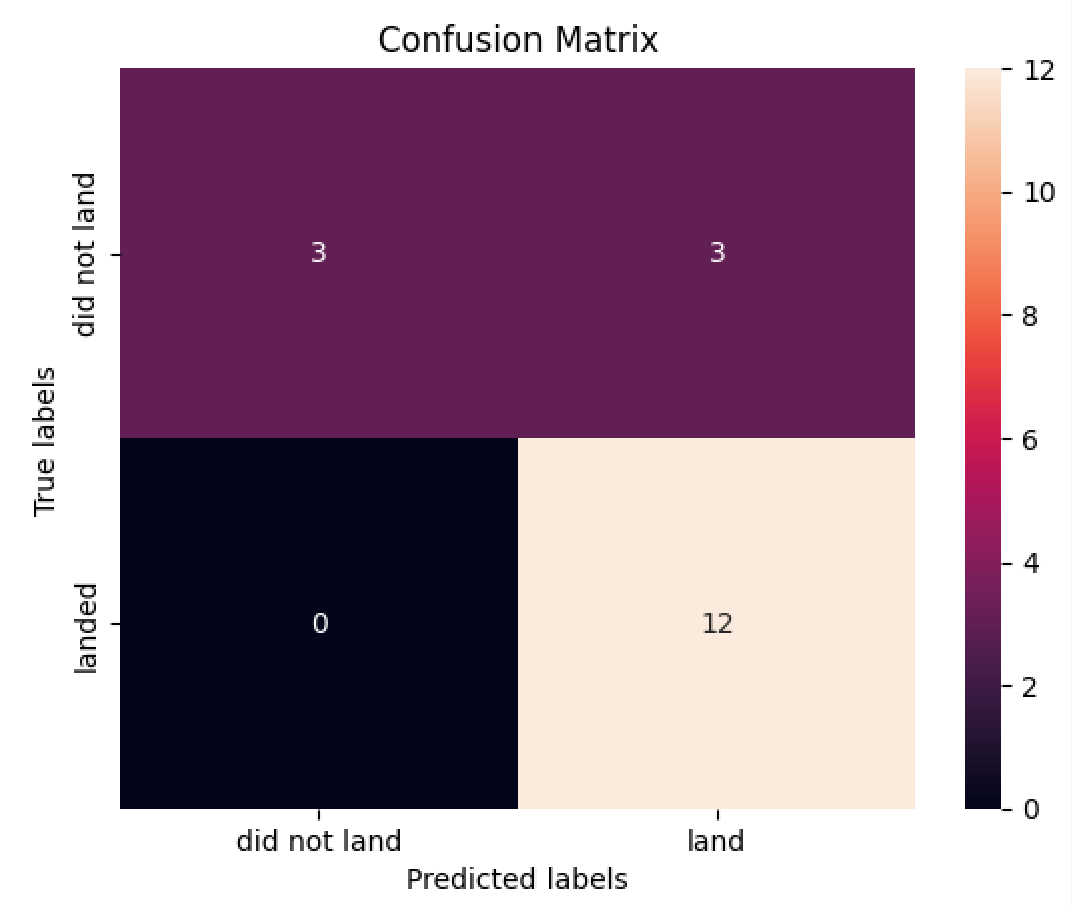
Classification Accuracy

- All the different classification models perform the same with 83,33% of accuracy so there is no better model than others.



Confusion Matrix

- The confusion matrix is also the same for all classification models
- There are 18 test samples
- Models predicted successfully 12 successful landings and 3 failed landings.
- There were no false negatives
- There were 3 false positives, as there were 3 successful predictions that weren't actually



Conclusions

- Regarding payload mass, it is recommended to use rockets with smaller payload mass to have a higher success rate
- Regarding orbits, rockets launched to VLEO,ES-L1,GEO,HEO,SSO have higher success rate so it is a parameter to be considered
- Regarding launch sites, it is preferrable to use KSC LC-39A as it is the site with better performance
- Regarding boosters, FT is the booster with better success rate
- Finally, all the different classification algorithms predict with an 83,33% accuracy, so any of them are good models to predict landing outcomes.

Thank you!