IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Péter Tóth
4th June 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected using SpaceX REST API and web scraping Wikipedia
  - Python DataFrames for cleaning and transforming data
  - Exploratory data analysis (EDA) using visualization and SQL
  - Interactive visual analytics using Folium and Plotly Dash
  - Perform predictive analysis using classification models
- Summary of all results
  - The success rates are increasing, showing a good learning curve
  - Orbit type, payload and launch site matters a lot
  - Launch site should be carefully selected: KSC LC-39A is the most successful site, both in absolute and relative terms
  - The machine learning model can predict the success with 83.3% accuracy

Introduction

Project background and context

- The commercial space age is here, companies are making space travel affordable. Perhaps the most successful company is SpaceX. One reason SpaceX is so successful is that can do this is that the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each.
- Much of the savings is because SpaceX can reuse the first stage of the rocket. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch and this information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Goals
  - Determine the main factors in a successful landing of the first stage of rocket
  - Determine the best machine-learning model to forecast the outcome of the landing
- Terminology
  - In this presentation success and successful means successful landing of the first stage of rocket, irrespective of the mission's success as puting the payload on orbit.

Section 1

Methodology

Methodology

Executive Summary

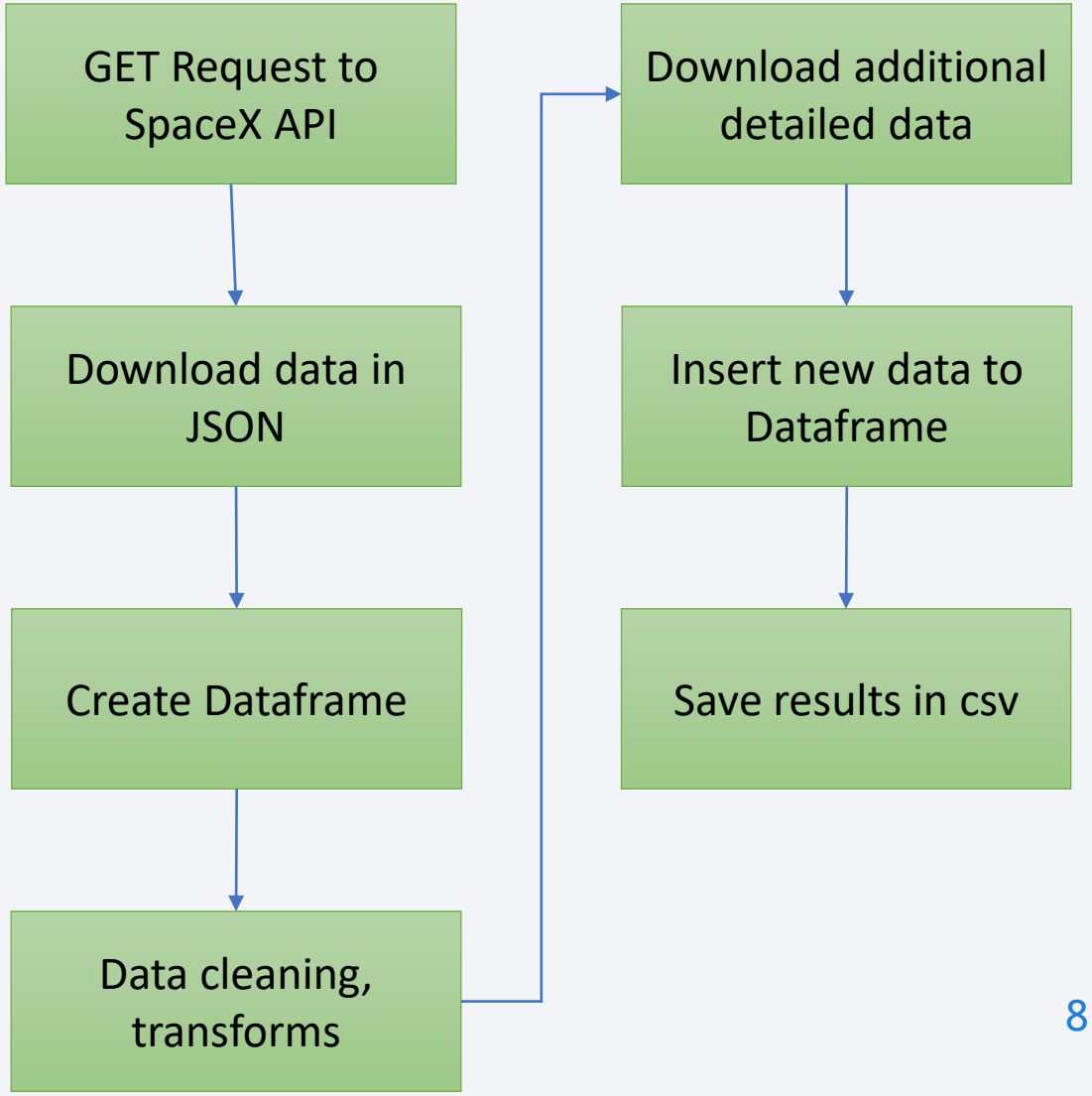
- Data collection methodology:
  - Data was collected using SpaceX REST API and web scraping Wikipedia
- Perform data wrangling
  - Python DataFrames, cleaning and transforming data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression, SVM, decision tree and KNN models with train-test separation
  - Winning model chosen on basis of accuracy

Data Collection

- Data coming from 2 main sources:
  - SpaceX's own API
    - We use GET requests for the API
    - Results coming in JSON
    - Processing with pandas DataFrames
  - Web scraping from Wikipedia
    - Use web request to the website
    - Use BeautifulSoup to process the raw data
    - Load and further process data in pandas DataFrames

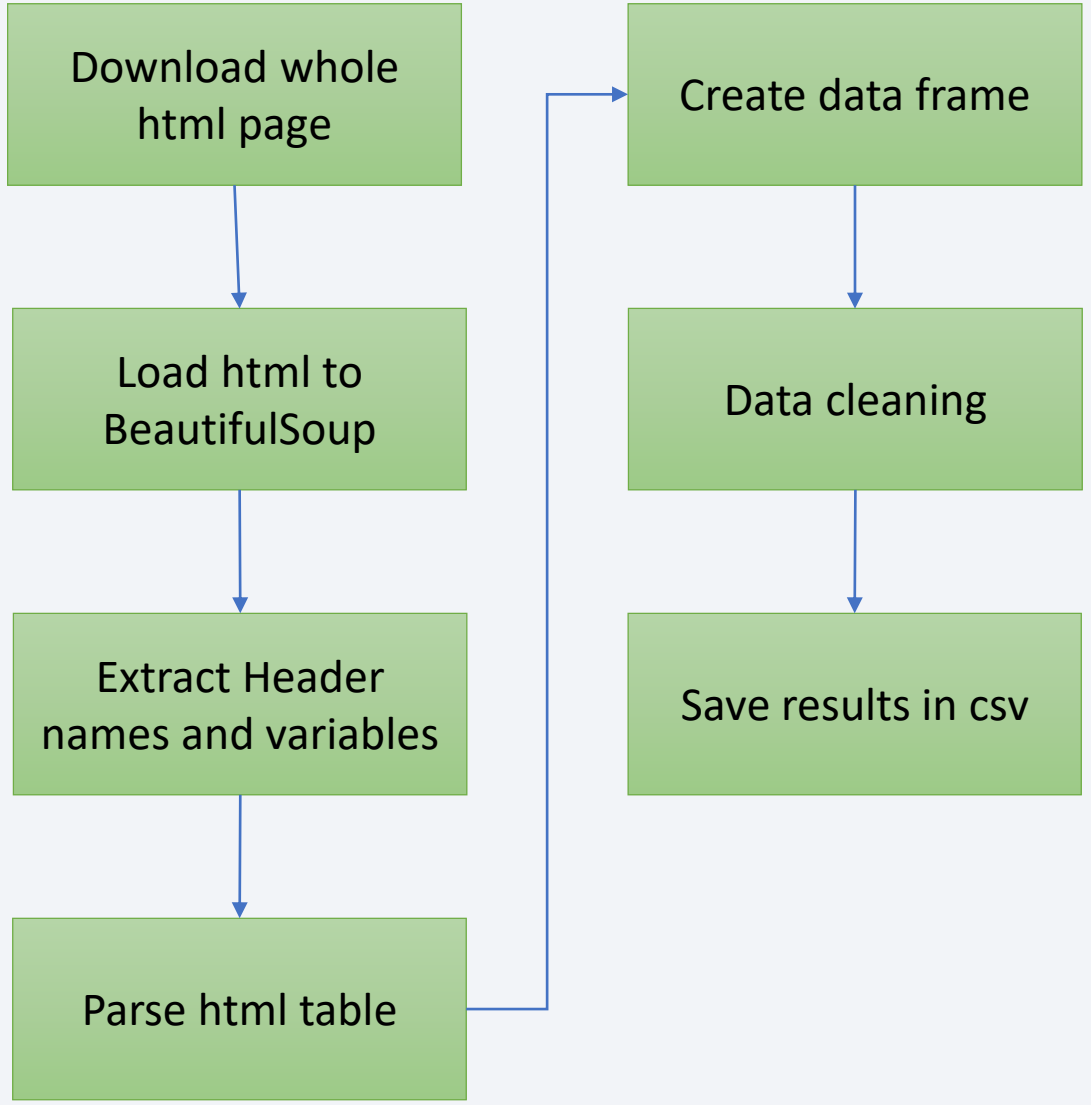
Data Collection - SpaceX API

- GET request to SpaceX API
- Parse response in JSON
- Normalize to DataFrame
- Transform data
- Remove unwanted columns
- Filter unwanted rows
- Replace some null values with mean
- Download additional data
- Export result to csv
- GitHub link:
  - https://github.com/devtpc/IBM-Data-Science/blob/main/10%20-%20Applied%20Data%20Science%20Capstone/01-data-collection-api.ipynb



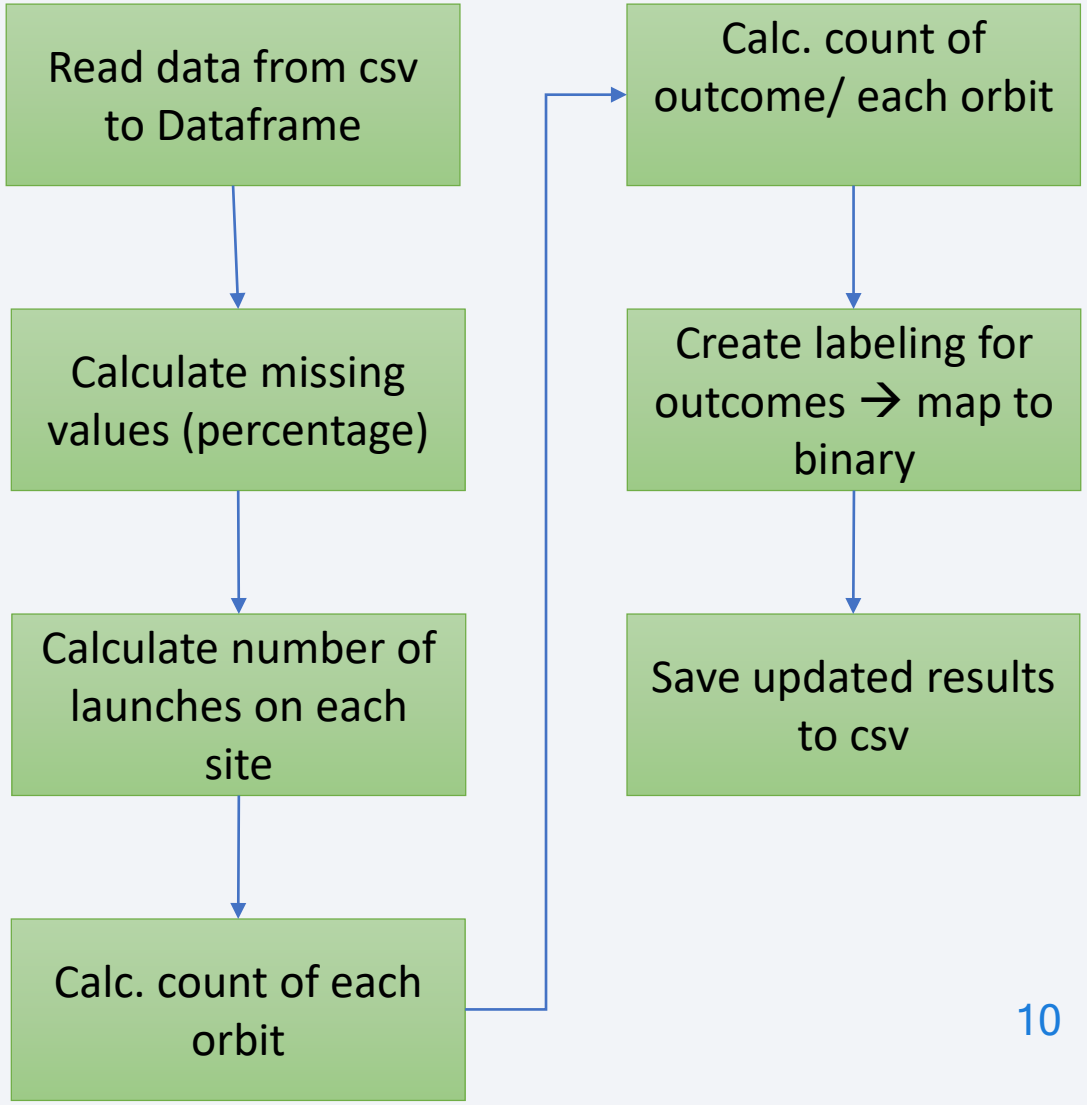
Data Collection - Scraping

- Download html page from Wikipedia with GET request
- Create BeautifulSoup object from html text
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the HTML table
- Save result as csv
- GitHub link:
  - https://github.com/devtpc/IBM-Data-Science/blob/main/10%20-%20Applied%20Data%20Science%20Capstone/02-webscraping.ipynb



Data Wrangling

- Data analysis to find some patterns
- Read data to Dataframe, and calculate missing values, and nr. of items in different categories by launch sites/orbit
- Create binary 0/1 labeling for outcomes
- Save data back to csv
- GitHub link:
  - https://github.com/devtpc/IBM-Data-Science/blob/main/10%20-%20Applied%20Data%20Science%20Capstone/03-data-wrangling.ipynb



EDA with Data Visualization

- The following charts were created in order to visualize, and explore the possible relationships:
  - Flight Number – Payload mass scatter plot
  - Flight Number – Launch site scatter plot
  - Payload mass – Launch site scatter plot
  - Orbit – Success Rate barchart
  - Flight number – Orbit scatter plot
  - Payload – Orbit scatter plot
  - Year – Success Rate line chart
- Github url: https://github.com/devtpc/IBM-Data-Science/blob/main/10%20-%20Applied%20Data%20Science%20Capstone/05-eda-dataviz.ipynb

EDA with SQL

- The following SQL Queries were performed:
  - List of 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 successful 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
  - Booster versions which have carried the maximum payload mass
  - Records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
  - Count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- GitHub url: https://github.com/devtpc/IBM-Data-Science/blob/main/10%20-%20Applied%20Data%20Science%20Capstone/04-eda-sql.ipynb

Build an Interactive Map with Folium

- Markers with the name of the launch sites and coordinates were added to an interactive map to visualize the launch sites on a map.
- Colored markers with the status of each launch were added to the map, indicating the success in red or green (not successful / successful)
  - As there are many launches on the same site, Clustered Maps were used
  - This enabled to easily visualize the number of launches and their success rates on a map
- In a randomly chosen location lines and measurements were added from the site to the nearest coast, highway, railway and city in order to determine, how important their proximity to the site is

- GitHub url: https://github.com/devtpc/IBM-Data-Science/blob/main/10%20-%20Applied%20Data%20Science%20Capstone/06-launch_site_location.ipynb

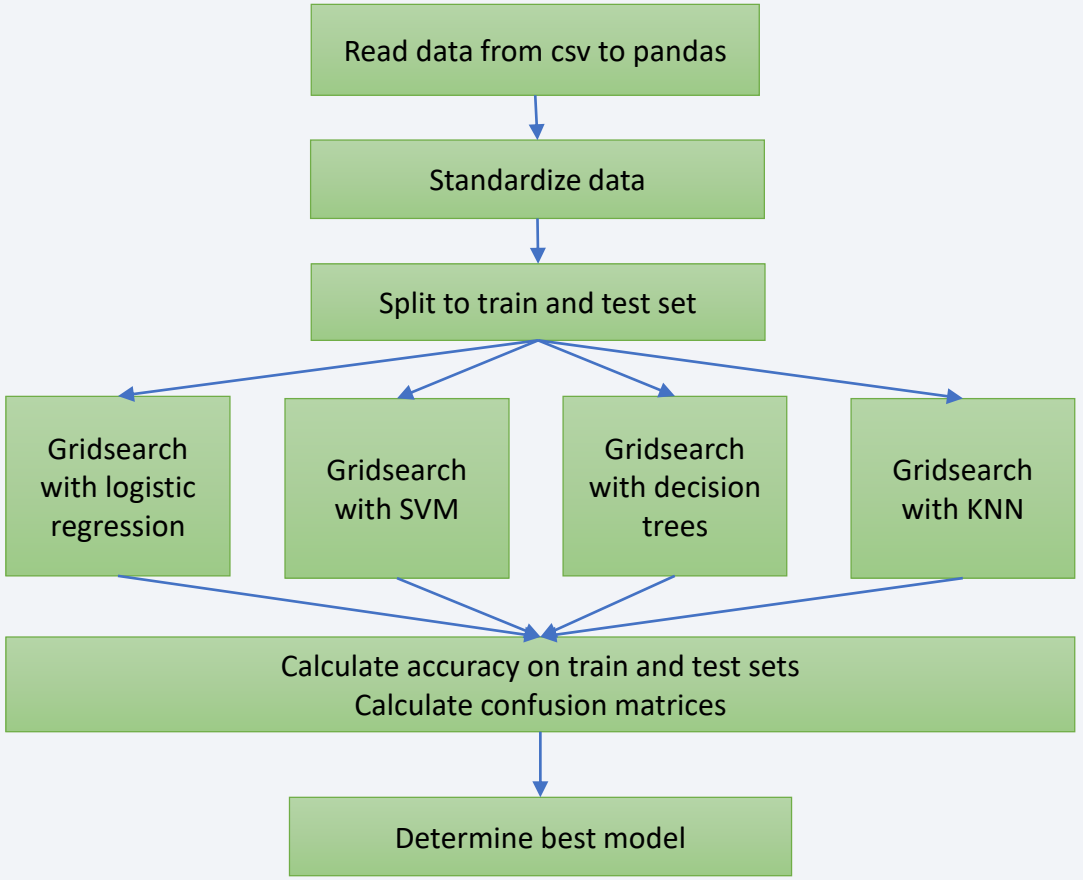
Build a Dashboard with Plotly Dash

- Added interactive charts to enable interactive visualizations, and deeper data understanding
  - Pie chart about successful launches by all sites
  - Pie chart to show success/non-success rate by individual launch sites
  - Scatter plot between relationship between Payload mass and outcome/success
    - Filterable by sites/all site
    - Color-coded booster version
    - Filterable range for payload mass

GitHub url: https://github.com/devtpc/IBM-Data-Science/blob/main/10%20-%20Applied%20Data%20Science%20Capstone/07-spacex_dash_app.py

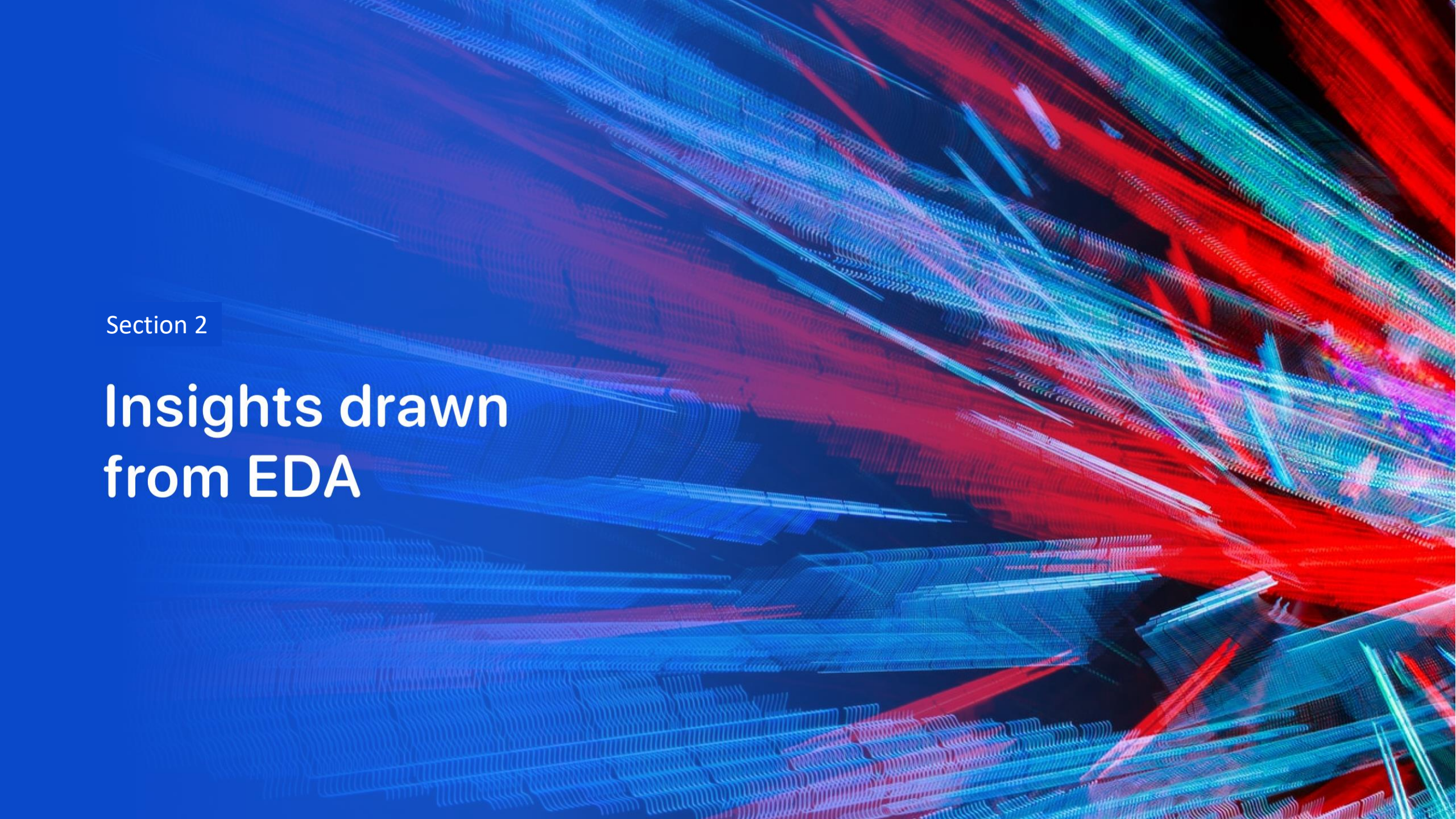
Predictive Analysis (Classification)

- For prediction, the loaded and standardized dataset was slit to train and test sets
- On train sets we used four methods: logistic regression, SVM, KNN and decision trees. Within each model the best hyperparameters were chosen with gridsearch
- We calculated the accuracy and the confusion matrices for each model and choose the best-fitting one.
- GitHub link:
  - https://github.com/devtpc/IBM-Data-Science/blob/main/10%20-%20Applied%20Data%20Science%20Capstone/08-machine_learning_prediction.ipynb



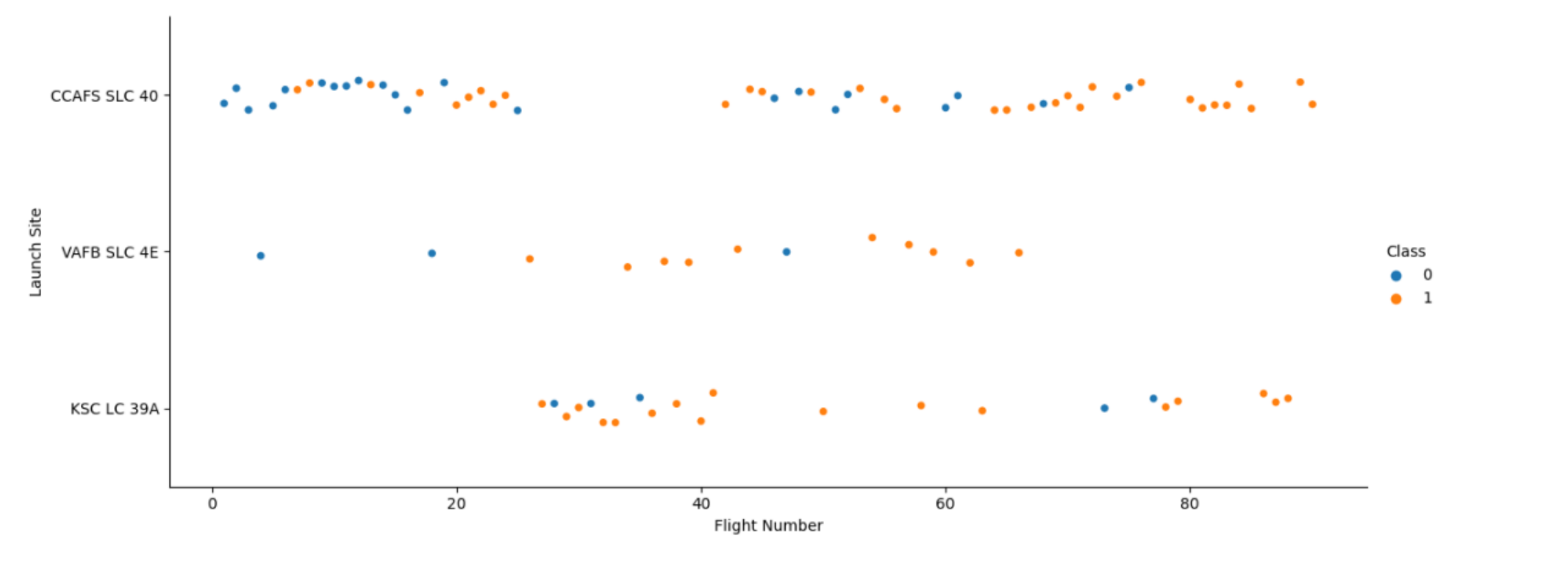
Results

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


Section 2
Insights drawn from EDA

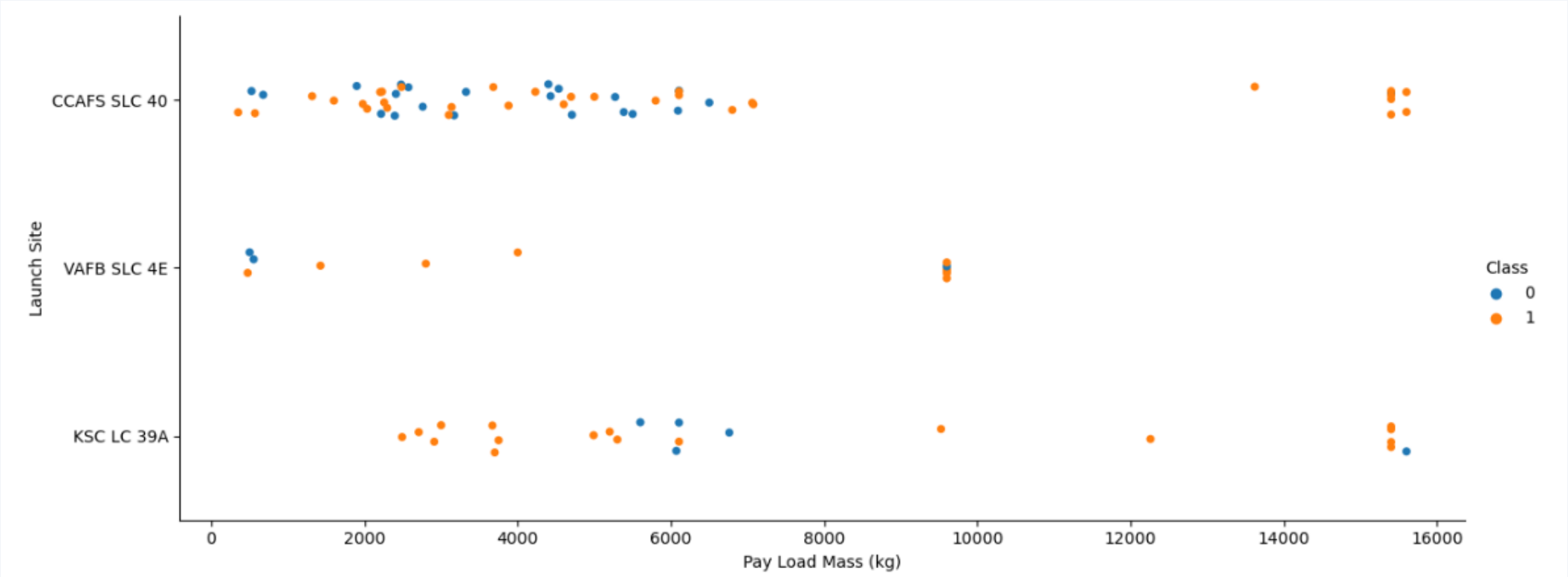
Flight Number vs. Launch Site

- As there are more flights (Flight Number increases) , there are more successful (orange) flights in every launch site



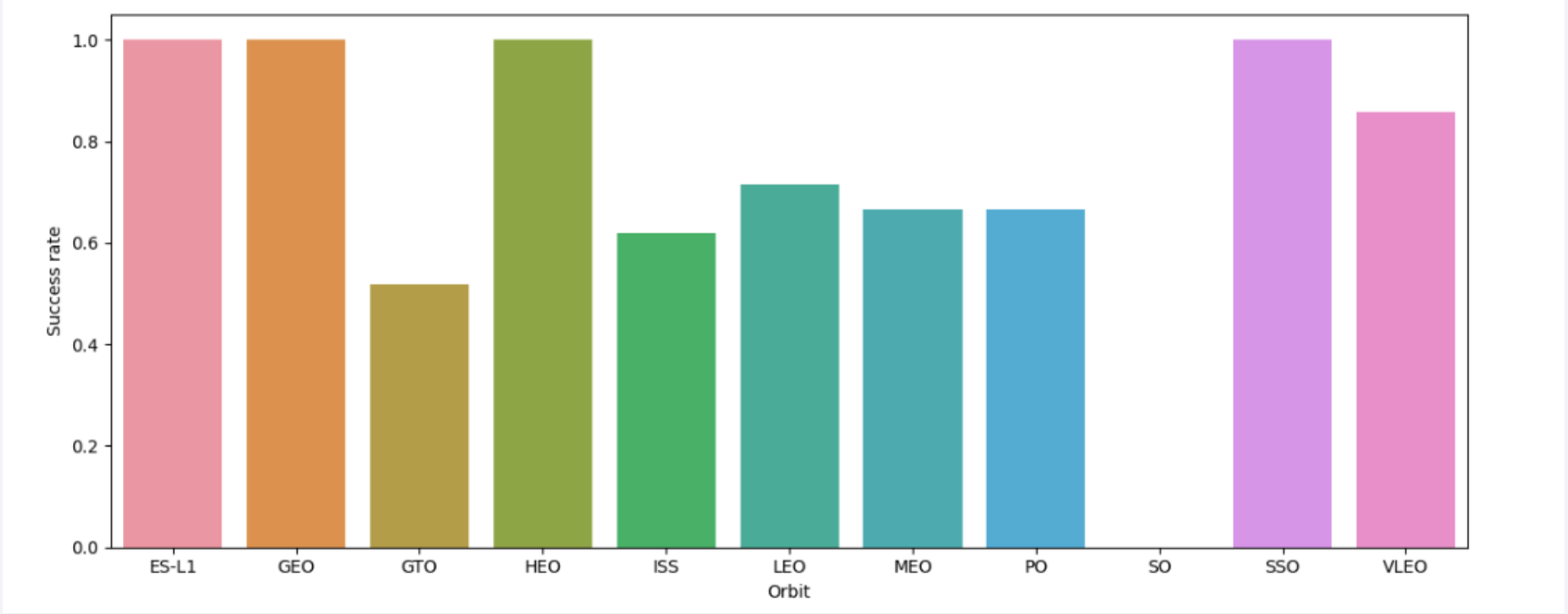
Payload vs. Launch Site

- The very heavy loads (above 8000 kg) are generally more successful
- In KSC LC 39A payloads around 6000 kg are not very successful, lighter and heavier are more successful
- IN VAFB SLC 4E there are no heavy payload (>10000) launches



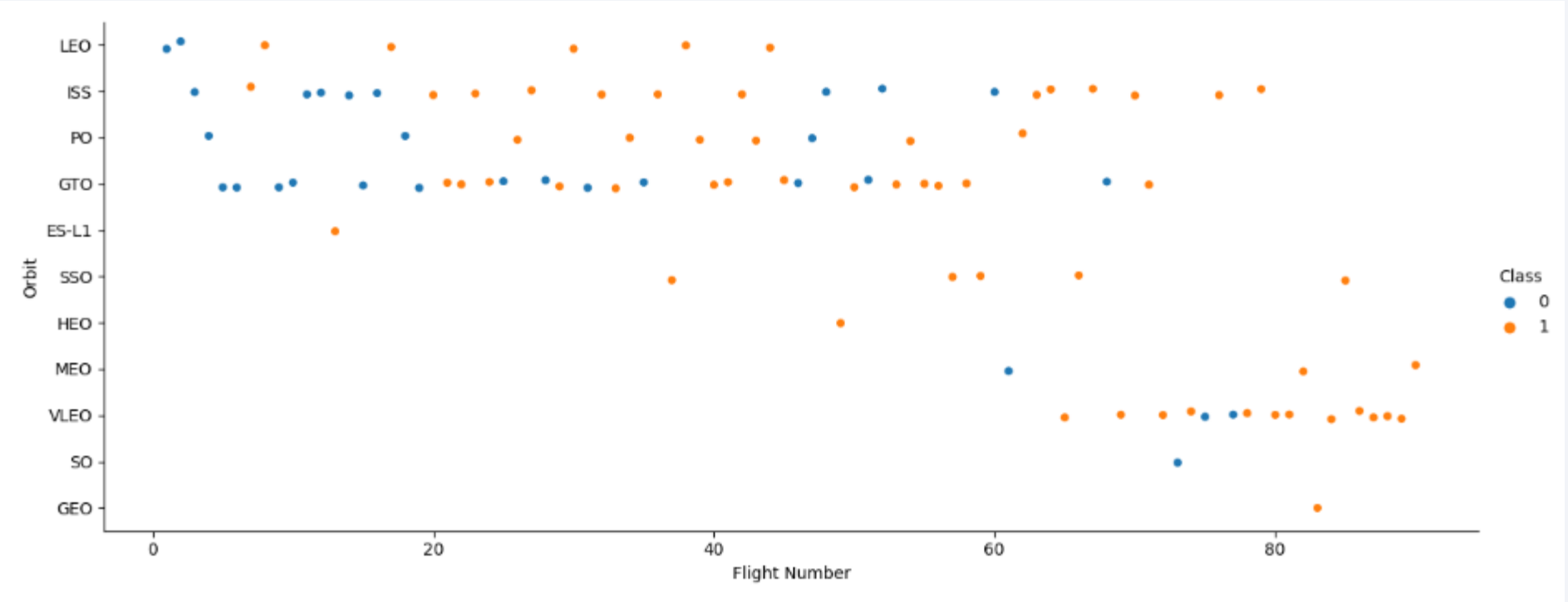
Success Rate vs. Orbit Type

- Orbit type is very important in success.
  - E-L1, GEO, HEO, SSO has 100% success rate, VLEO has close to 90%
  - SO has 0% success rate
  - Other types are in between, in these cases other factors could be more important



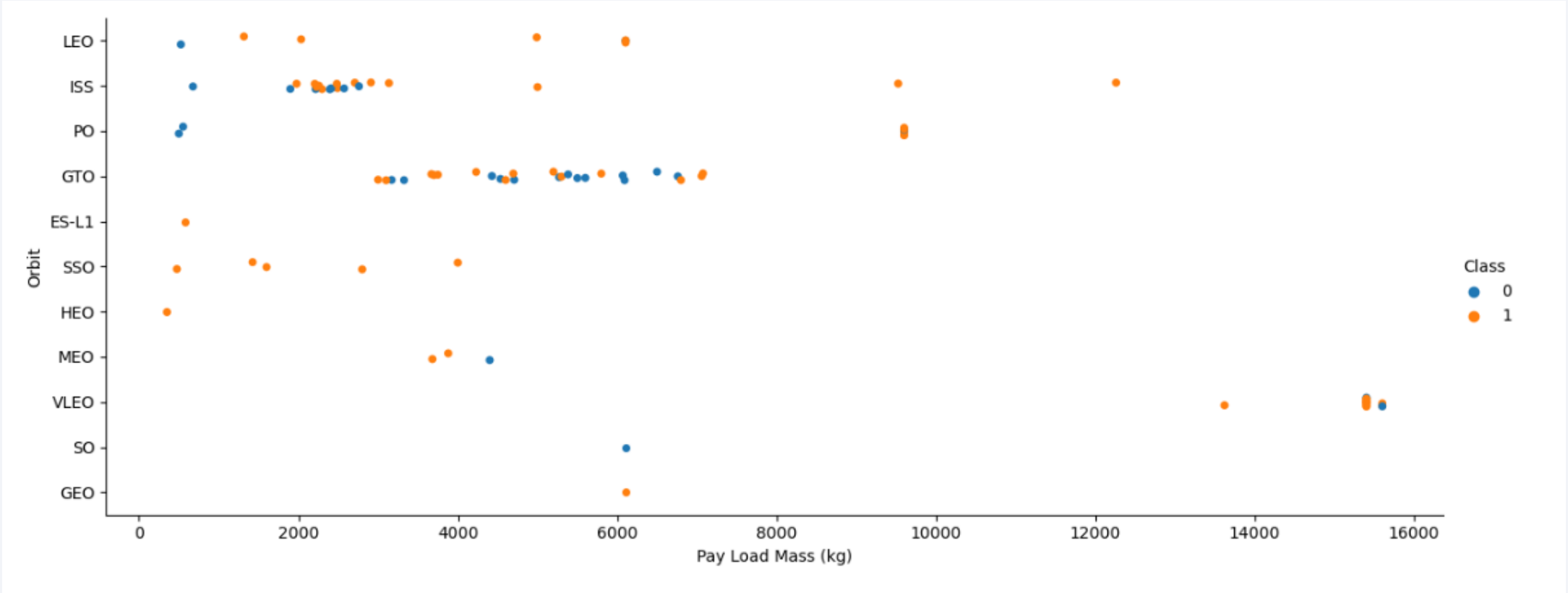
Flight Number vs. Orbit Type

- As there are more flights (Flight Number increases) , there are more successful (orange) flights in every orbit (especially on LEO), with the exception of GTO



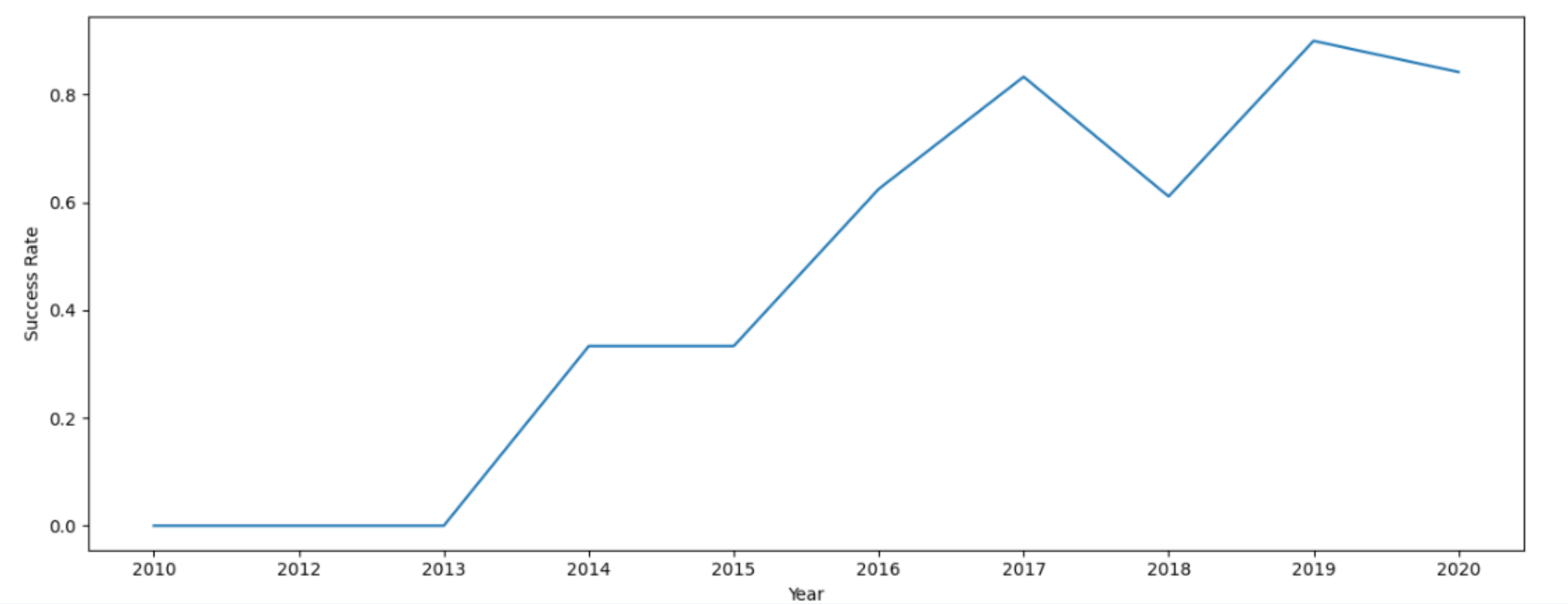
Payload vs. Orbit Type

- In LEO, ISS, PO orbits heavier payloads seem to be more successful.



Launch Success Yearly Trend

- The success rate has an increasing yearly trend



All Launch Site Names

- The distinct Launch sites were selected:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

```
[8]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

[8]:

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

None

Launch Site Names Begin with 'CCA'

- Used LIKE ’CCA%’ to select sites begin with CCA
- LIMIT 5 enables to retrieve only 5 records

Display 5 records where launch sites begin with the string 'CCA'

[9]:
```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

[9]:

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	0.0	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	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total payload mass carried by boosters from NASA is 455596 kg
- The SUM() function adds the individual masses

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
[10]: %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Customer" = "NASA (CRS)"
```

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

[10]:

SUM("PAYLOAD_MASS__KG_")
45596.0

Average Payload Mass by F9 v1.1

- Average Payload is 2534,66 kg
- AVG() function calculates us the aggregate
- LIKE ”F9 v1.1%” was used to include the sub-types of F9 v1.1, too

Task 4

Display average payload mass carried by booster version F9 v1.1

```
[20]: %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booster_Version" LIKE "F9 v1.1%"
      # "Booster_Version" = "F9 v1.1%" is not good, there are data like "F9 v1.1  B1003"
```

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

[20]:

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- MIN() function gives us the lowest value →2015-12-22
- Unlike other SQL-s (like MySQL) SQL Lite does not automatically convert string to Date, so had to convert to YYYY-MM-dd format, which is good for Date ordering

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

[18]:
```
# %sql SELECT MIN(Date) FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)"
# straightforward solution gives 01/08/2018, which is bad result, orders dates like string, 01 comes first, not 2018

%sql SELECT MIN((substr(Date, 7, 4) || '-' || substr(Date, 4, 2) || '-' || substr(Date, 1, 2))) FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)"
```

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

[18]:

MIN((substr(Date, 7, 4) || '-' || substr(Date, 4, 2) || '-' || substr(Date, 1, 2)))
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- AND is used to take into account both conditions
- BETWEEN is used
- Result are 4 values: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

[28]:
```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (drone ship)" AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000
```

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

[28]:

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

Total Number of Successful and Failure Mission Outcomes

- Had to use GROUP BY
- Results are: 99 success, 1 Success (payload status unclear), 1 Failure (in flight)

Task 7

List the total number of successful and failure mission outcomes

```
[29]: %sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") FROM SPACEXTBL GROUP BY ("Mission_Outcome")
```

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

[29]:

Mission_Outcome	COUNT("Mission_Outcome")
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- There are 12 results
- Distinct was used
- In the subquery we had to use the MAX() function

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

[30]:
```
%%sql
SELECT DISTINCT "Booster_Version" FROM SPACEXTBL
WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

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

[30]:

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

- There were 2 failed landings on drone ships in 2015, both on Launch site CCAFS LC-40, with booster versions F9 v1.1 B1012 and F9 v1.1 B1015
- For the query, substr() had to be used

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

[31]:
```
%%sql
SELECT substr(Date, 4, 2) as month, "Landing_Outcome", "Booster_Version", "Launch_Site"
FROM SPACEXTBL
WHERE "Landing_Outcome" = "Failure (drone ship)" AND substr(Date,7,4)='2015'
```

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

[31]:

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

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

- GROUP BY and ORDER BY was used.
- First is „No attempt” with 10 occasion
- Then comes 5-5 Success (ground pad) and Success (drone ship)
- The followings are
  - Controlled (ocean) – 3
  - Uncontrolled (ocean) – 2
  - Precluded (drone ship) – 1
  - Failure (parachute) – 1

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

[37]:
```
%%sql
SELECT "Landing_Outcome", COUNT("Landing_Outcome") cnt
FROM SPACEXTBL
WHERE (substr(Date, 7, 4) || '-' || substr(Date, 4, 2) || '-' || substr(Date, 1, 2)) >= '2010-06-04'
 AND (substr(Date, 7, 4) || '-' || substr(Date, 4, 2) || '-' || substr(Date, 1, 2)) <= '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY cnt DESC
```

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

[37]:

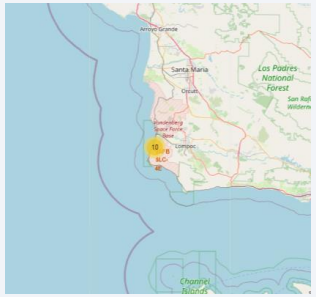
Landing_Outcome	cnt
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

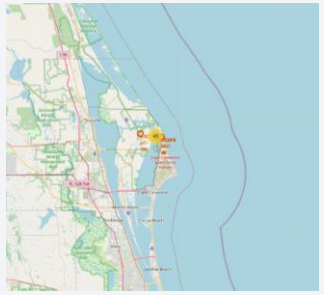
Section 3

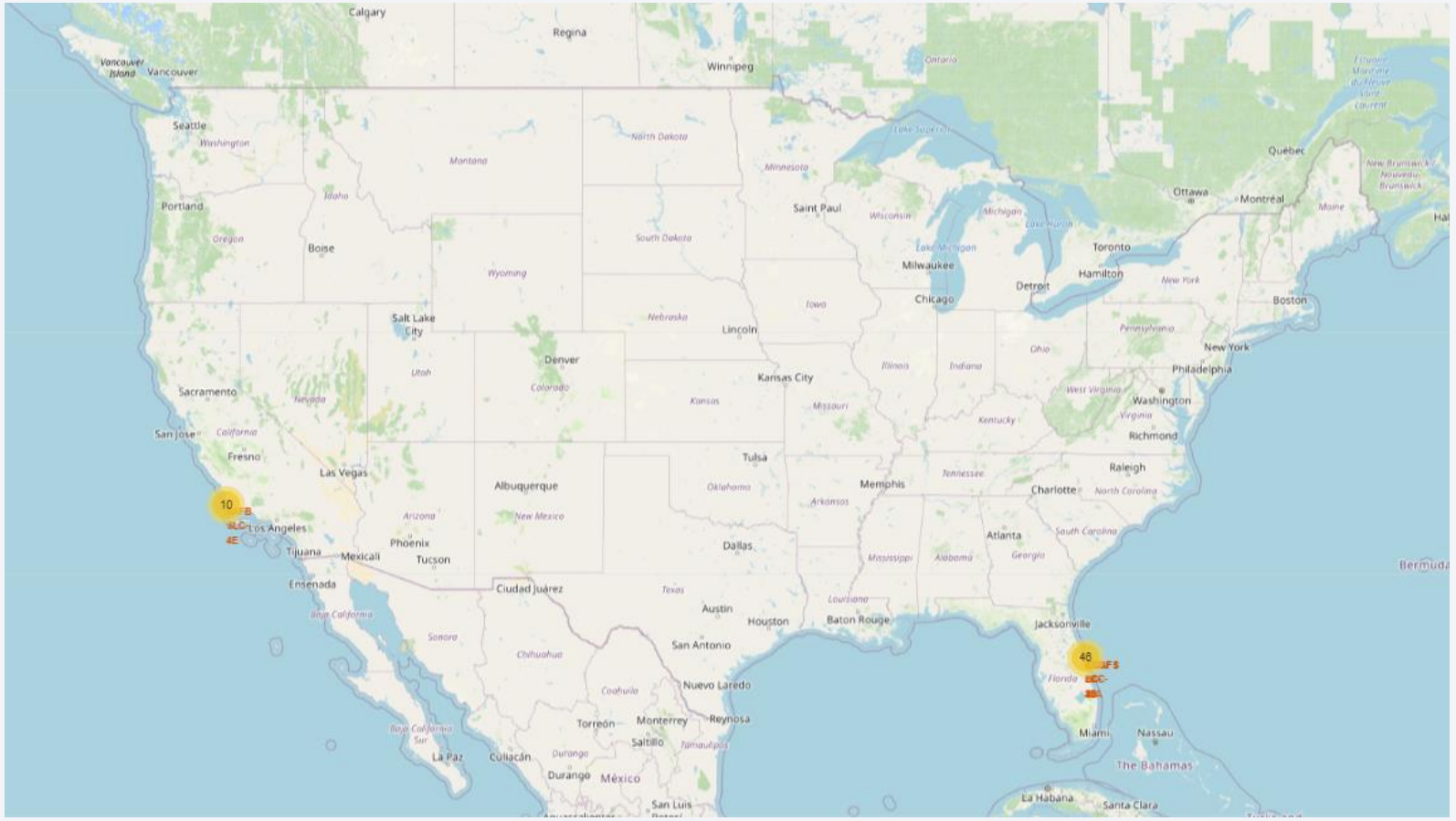
Launch Sites Proximities Analysis

Location of Launch Sites

- All launch sites are located in the continental USA
  - Close to the coasts (California, Florida)
  - As south / close to the Equator as possible



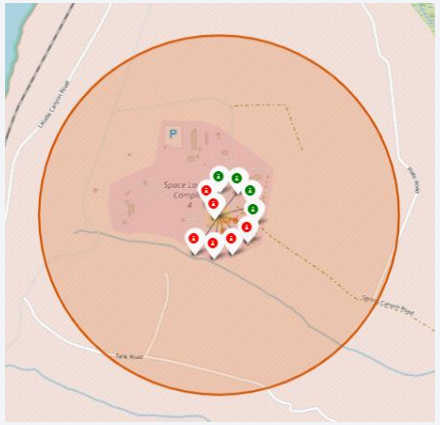




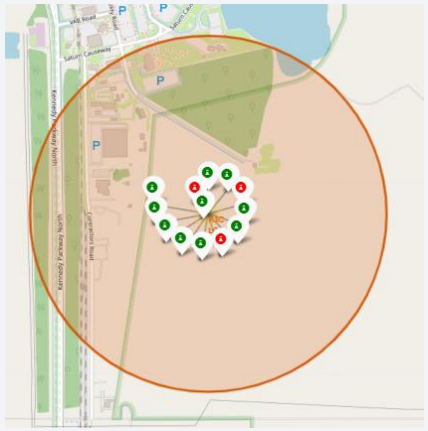
Launch Site data

- Number of markers shows total launches, greens are successful, reds are not
- It's easy to see, that
  - CCAFS LC-40 has the most launches (most markers)
  - KSC LC-39A is the most successful (most green markers, absolutely and relatively both)

VAFB SLC-4E

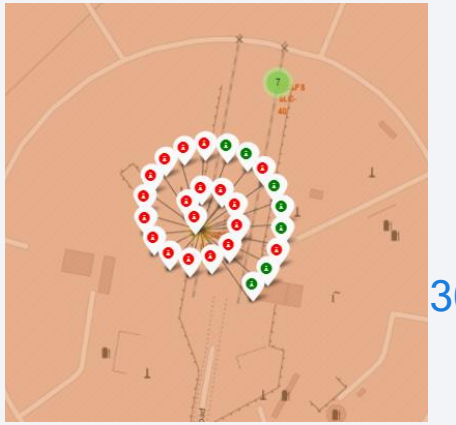


KSC LC-39A



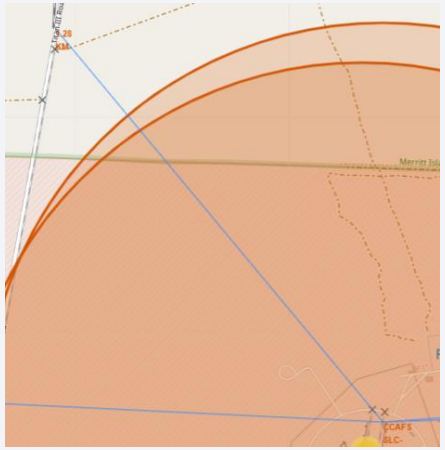
CCAFS SCL-40

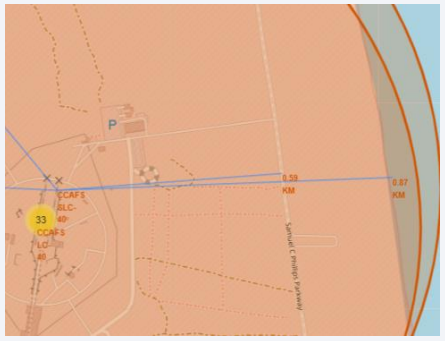
CCAFS LC-40

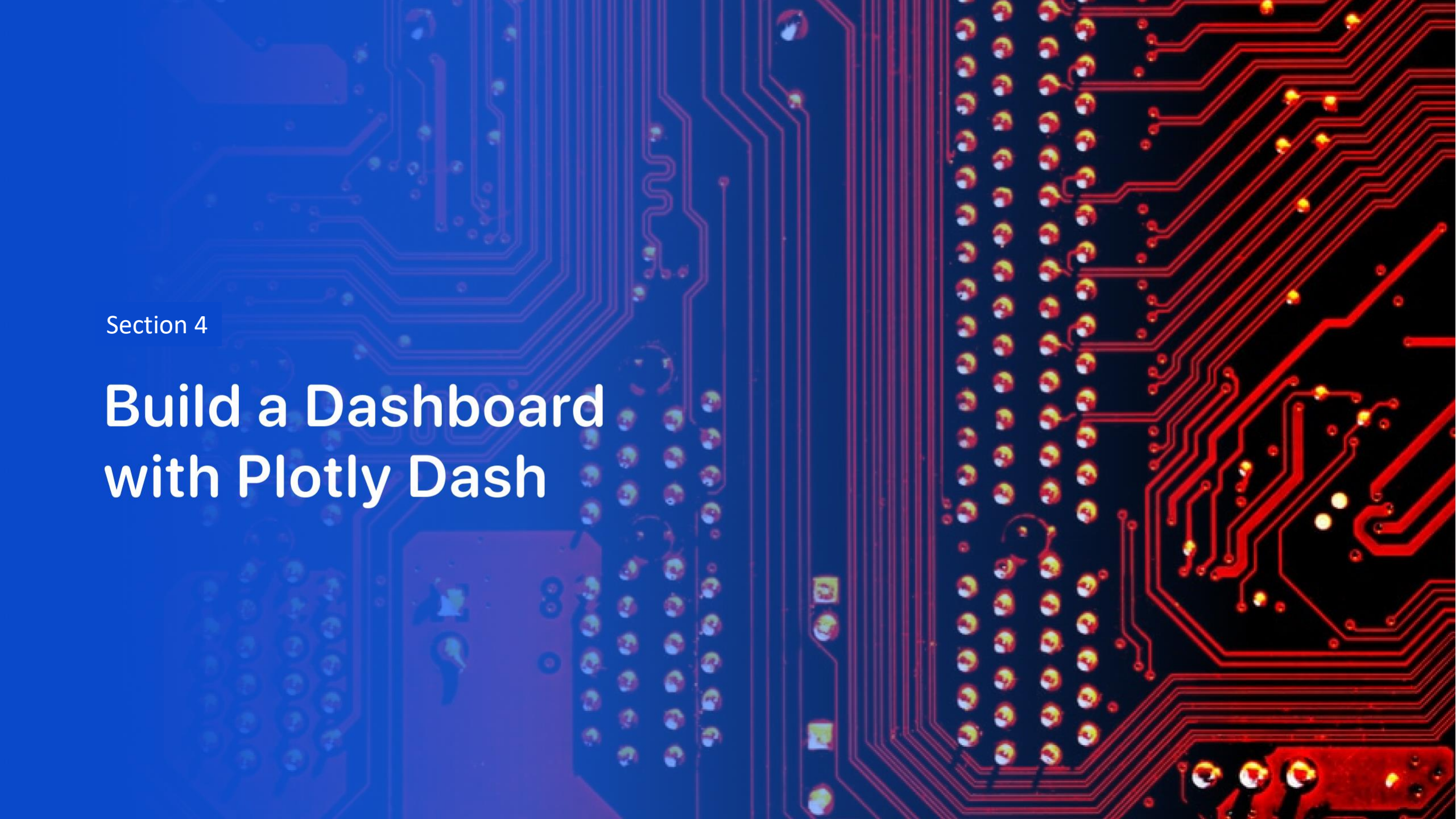


Distances from important points

- The randomly selected CCAFS SLC-40 launch site is:
  - 21,59 km. from the nearest city
  - 0,59 km from the nearest road/highway
  - 1,28 km from the nearest railway line
  - 0,87 km from the nearest coast
- It’s safe to say that it is important for a launch site to be
  - Close to the transportation facilities (road, railway, water)
  - Relatively far from the cities

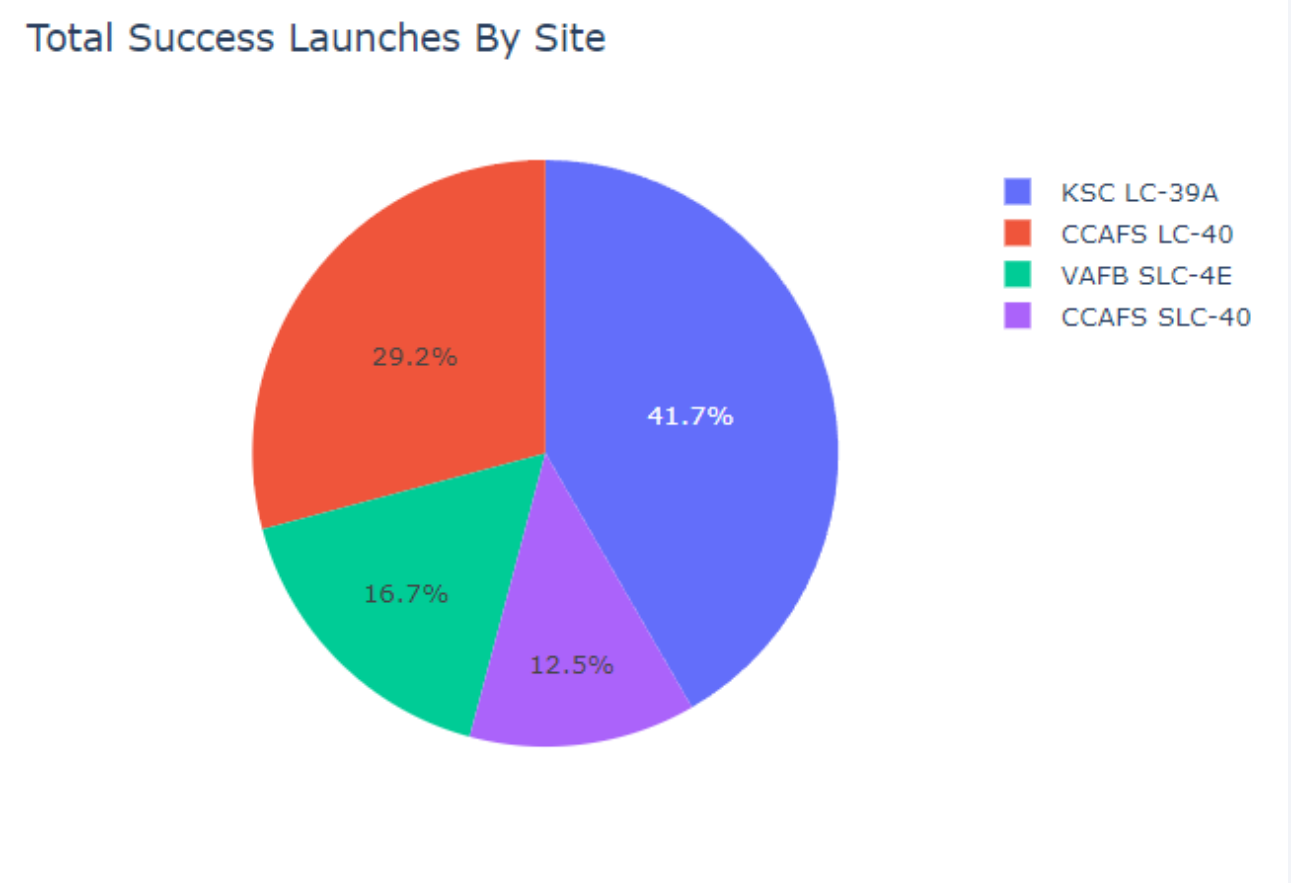





Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site

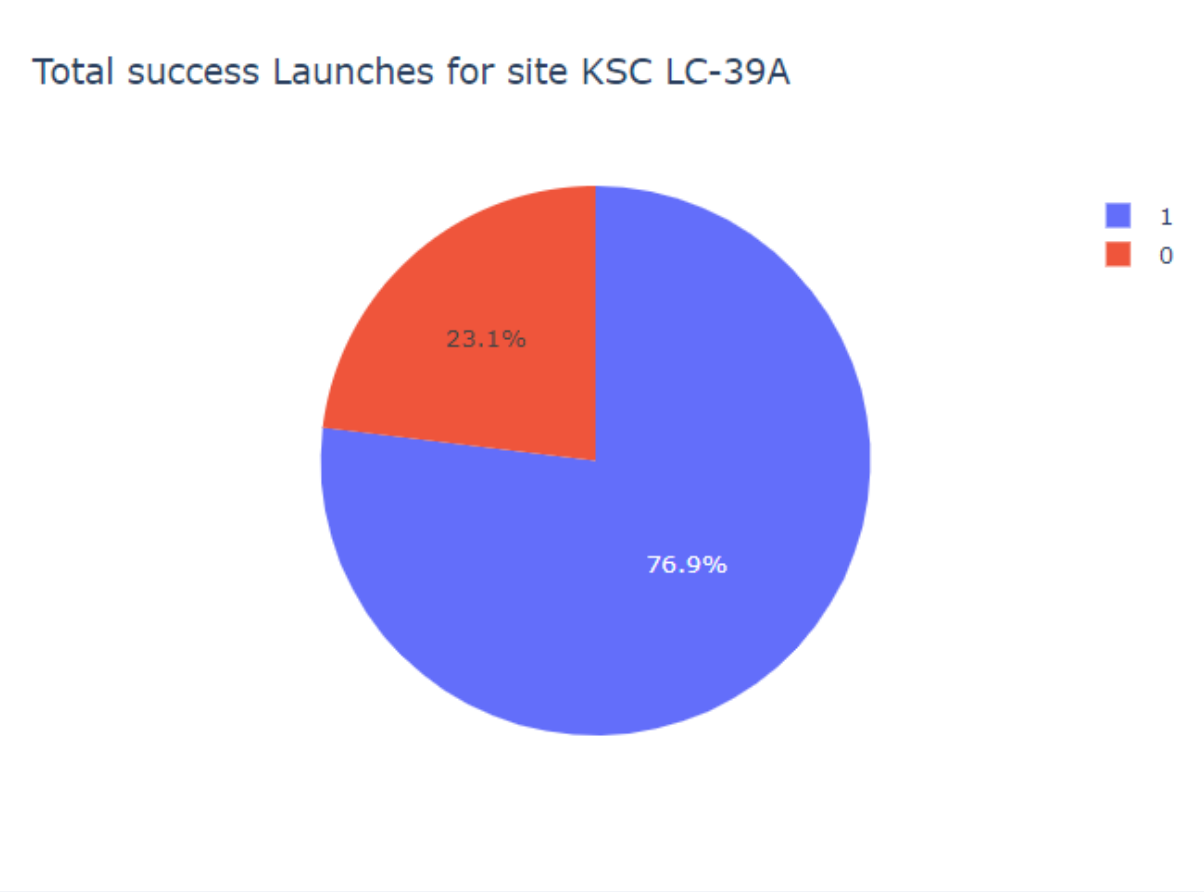
- **KSC LC-39A** site has the most successful launches, **41,7 %** of the total
- Other sites in descending order of success are:
  - CCAFS LC-40 (29,2%)
  - VAFB SLC-4E (16,7%)
  - CCAFS SCL-40 (12,5%)
- This means, that 41,7% of the succesful launches are from **KSC LC-39A**



Most Successful launch site

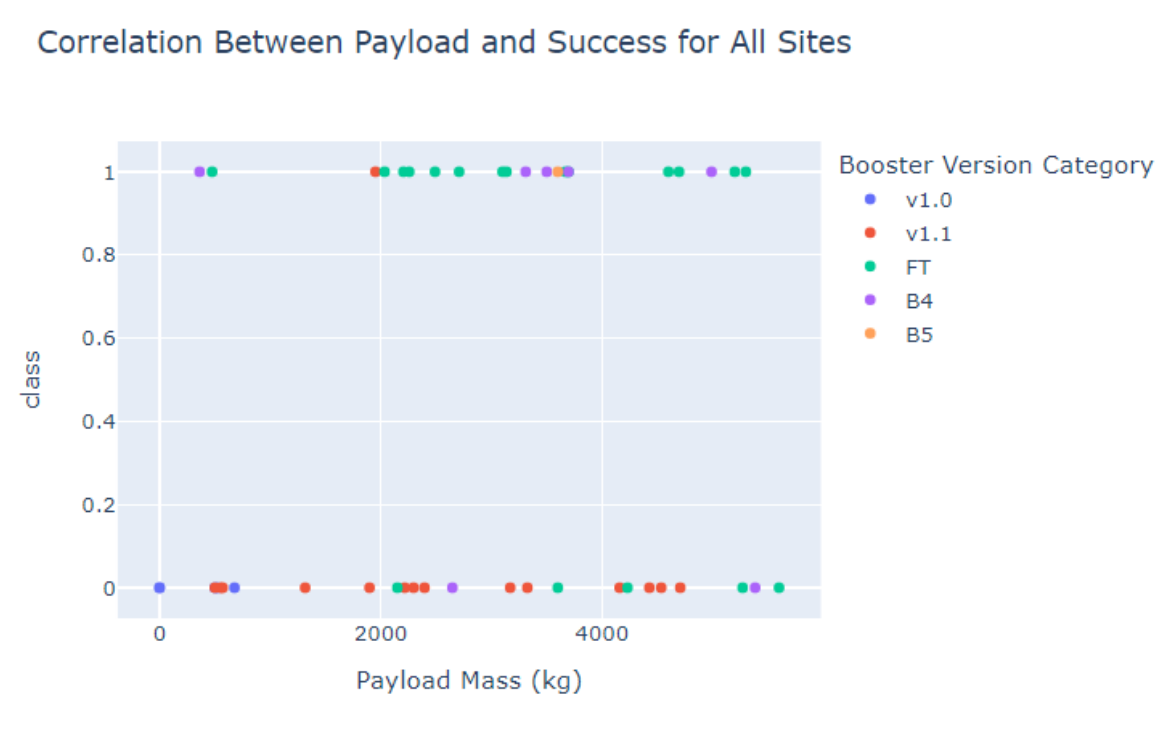
- KSC LC-39A site has the highest success ratio with 76,9%

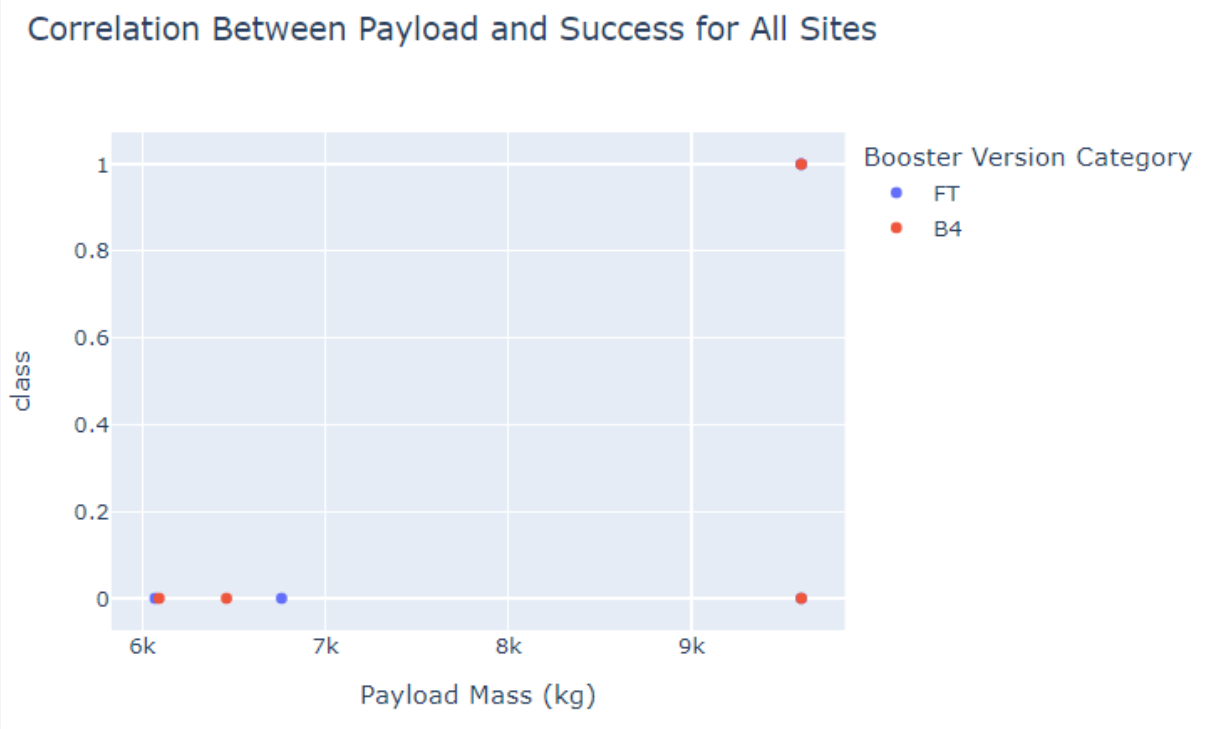
- This means, that 76,9% of the first stage of rockets launched from KSC LC-39A site are successfully landed



Payload size matters

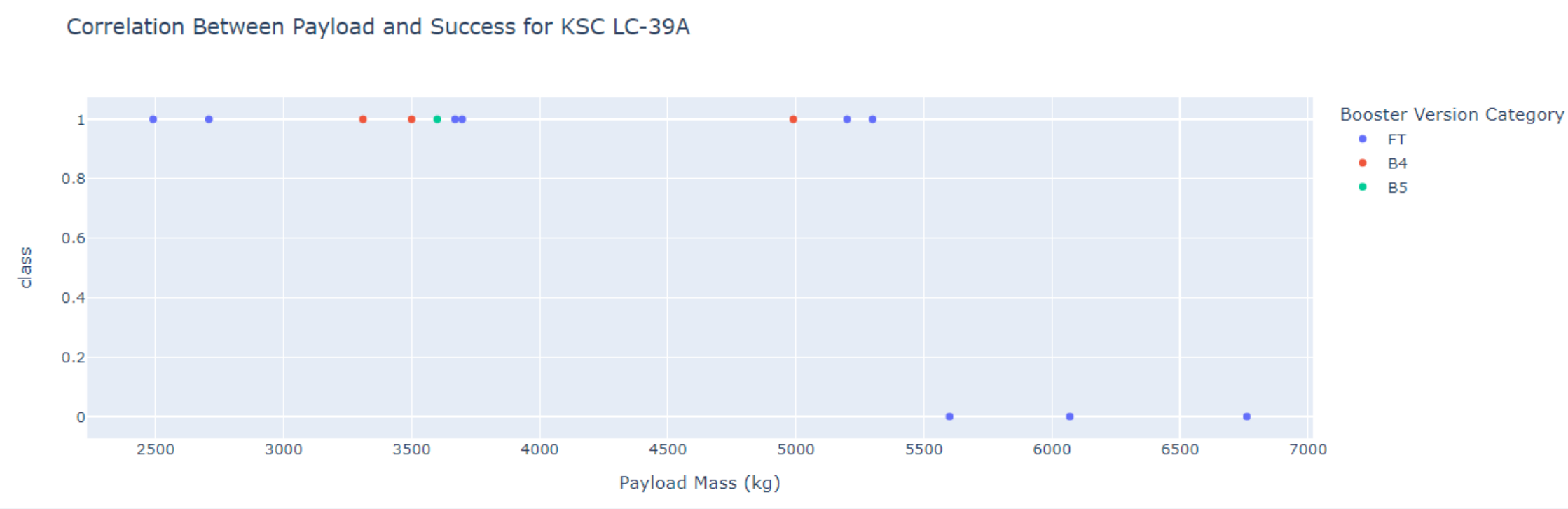
- There are much more succesful launches, when the payload are smaller, namely below 6000 kg, than payload between 6000 and 10000 kg.





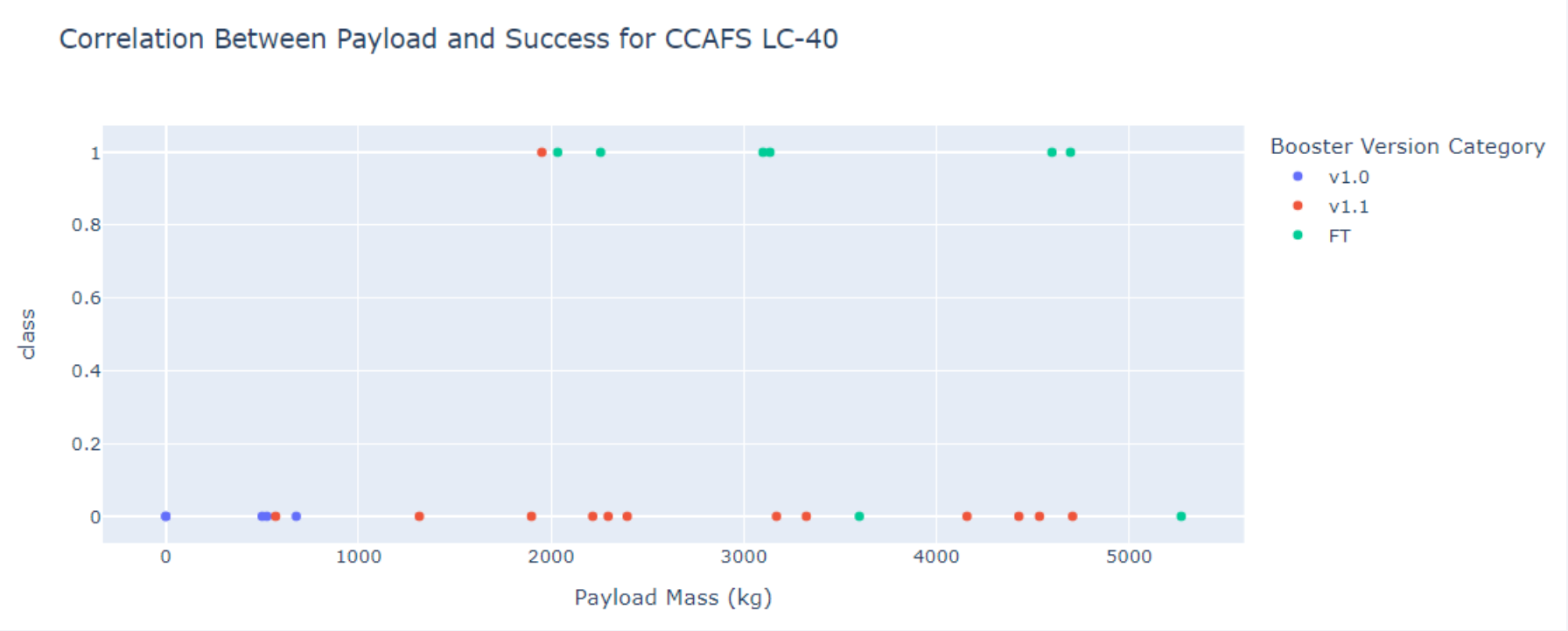
Site KSC LC-39A – success in light weights

- On Site KSC LC-39A
  - all launches below 5500 kg are successful,
  - all launches above 5500 kg are **NOT** successful,



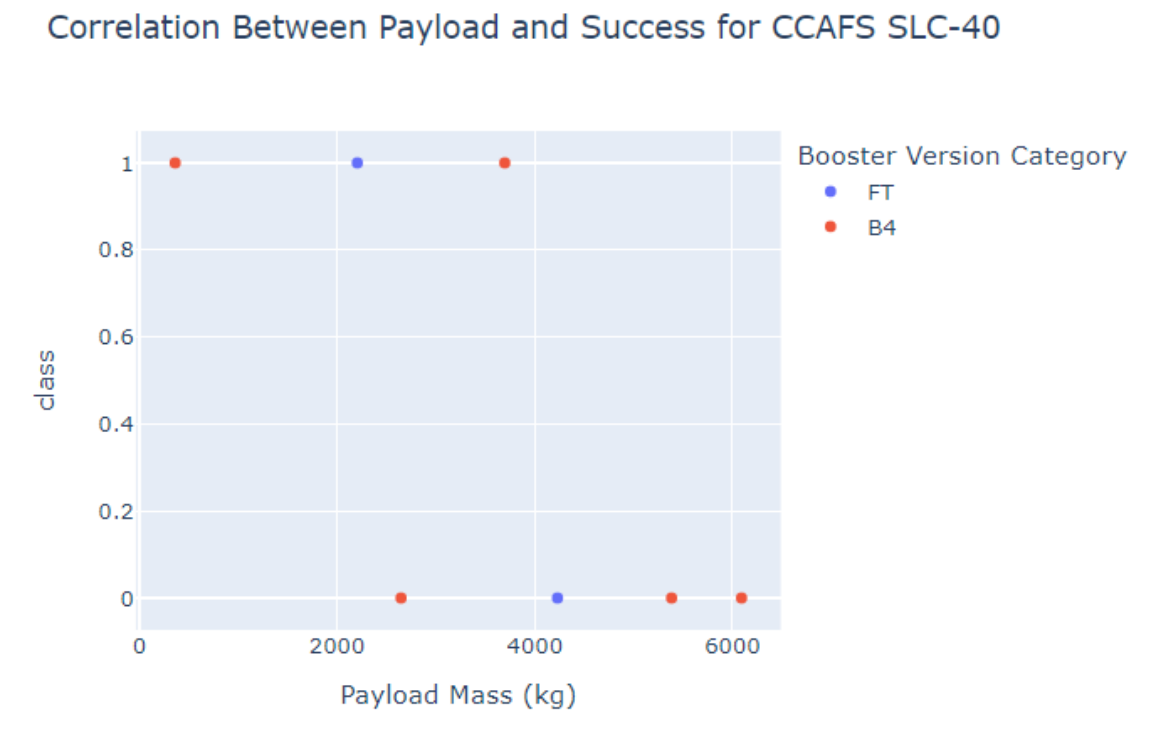
Site CCAFS LC-40 - FT Booster Success

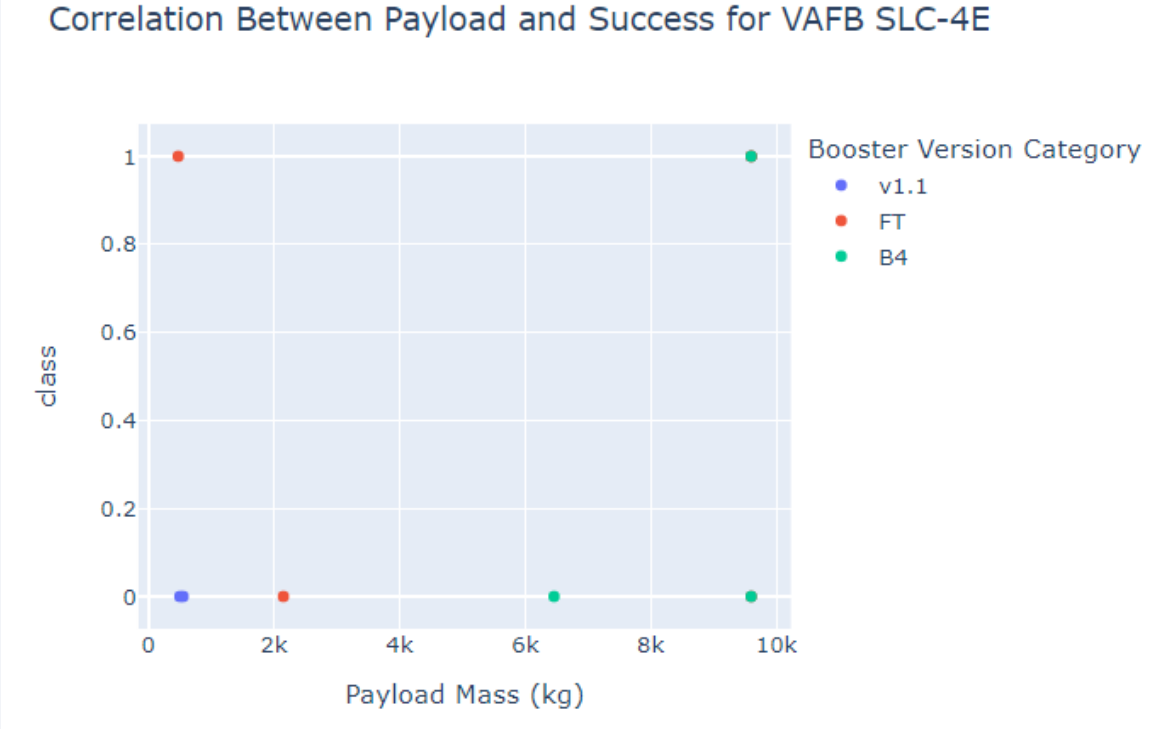
- On Site CCAFS LC-40 almost all successful launches are with Booster version FT.
- All v1.0, and almost all v1.1 versions are not successful



Other sites

- CAFS SLC-40 and VAFB SCL-4E have very few data points.
  - On CAFS SLC-40 all launches above 4000 kg. are not successful
  - VAFB SCL-4E does not show clear tendency





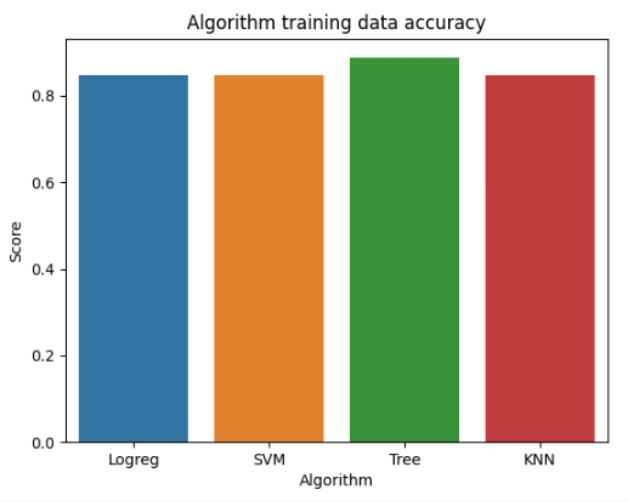
Section 5

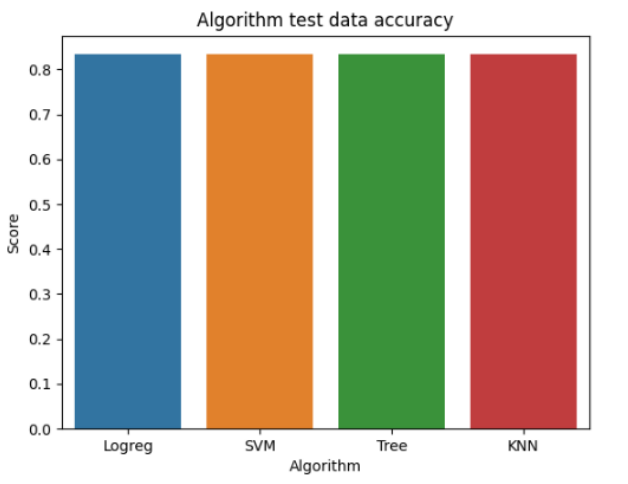
Predictive Analysis (Classification)

Classification Accuracy

- On the training data, the decision tree method has a slightly better classification accuracy, than the other methods

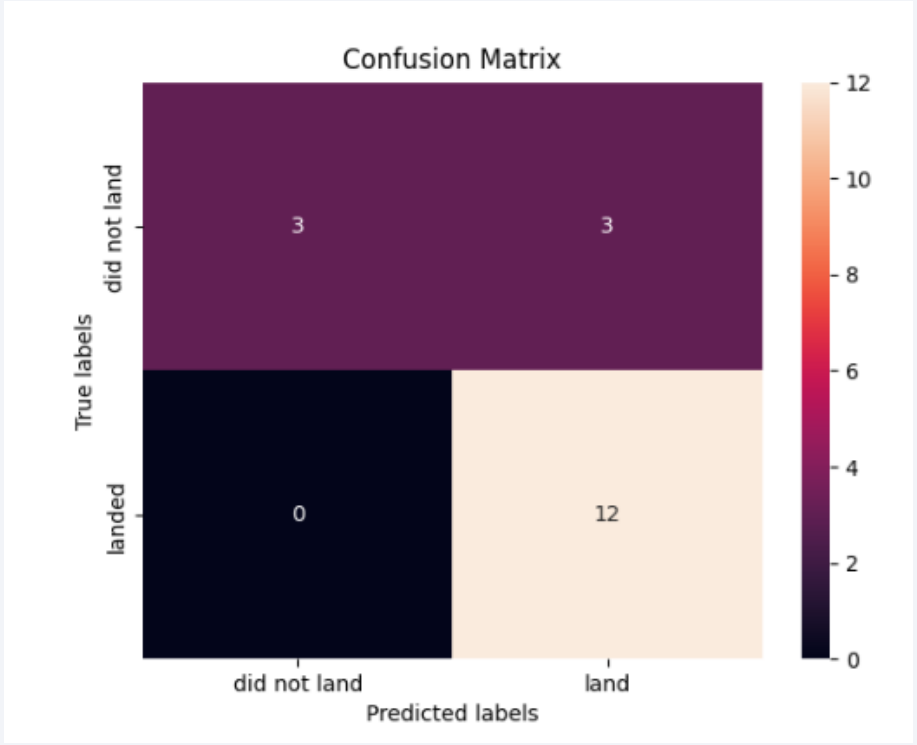
- However, on test data **all four models has the same accuracy** (83,33%)





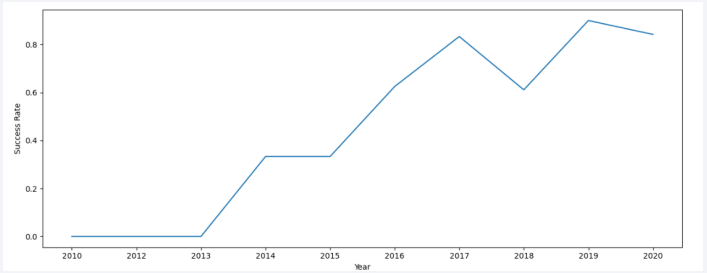
Confusion Matrix

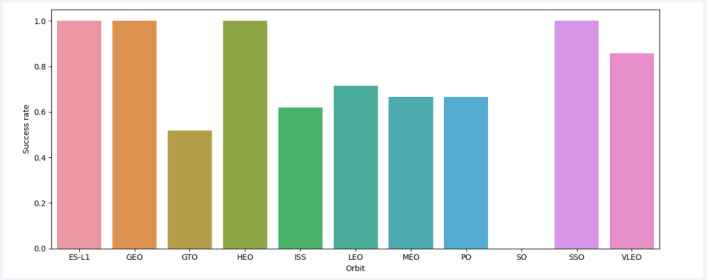
- All models have the same Confusion Matrix
- He accuracy is 83,3%, which is good
- There are 0 false negatives, which is very good.
- The main problem is the false positives, which is 3 (upper-left corner). This means, that we predict, that these will land, but in practice they will not, causing additional costs to the company.

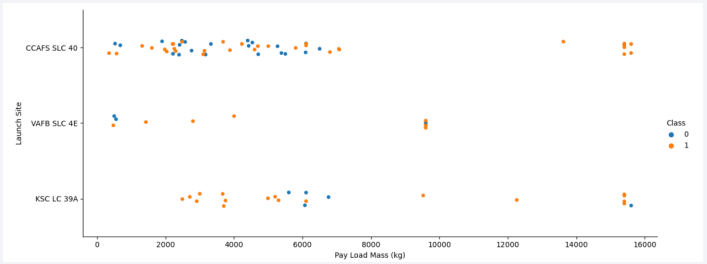


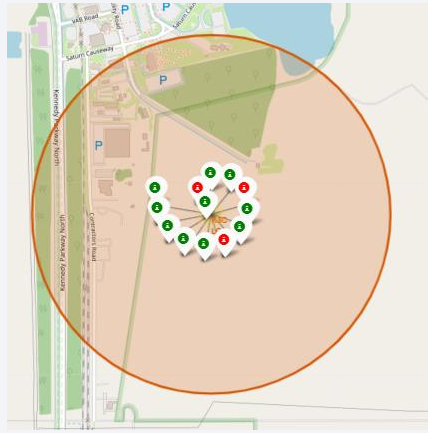
Conclusions

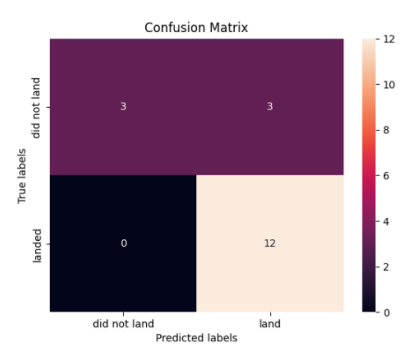
- The success rates are increasing, showing a good learning curve
- Orbit type, payload and launch site matters a lot
- Launch site should be carefully selected: KSC LC-39A is the most successful site, both in absolute and relative terms
- The machine learning model can predict the success with 83.3% accuracy









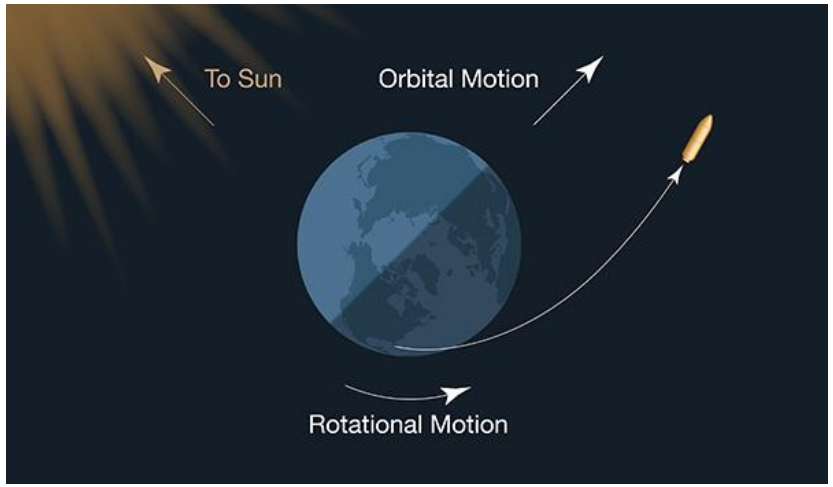


Other Insights

- Why a launch site close to the Equator is better:

  „If a spacecraft is launched from a site near Earth's equator, it can take optimum advantage of the Earth's substantial rotational speed. Sitting on the launch pad near the equator, it is already moving at a speed of over 1650 km per hour relative to Earth's center. This can be applied to the speed required to orbit the Earth (approximately 28,000 km per hour). Compared to a launch far from the equator, the equator-launched vehicle would need less propellant, or a given vehicle can launch a more massive spacecraft."

  (Source of text and picture: https://solarsystem.nasa.gov/basics/chapter14-1/)

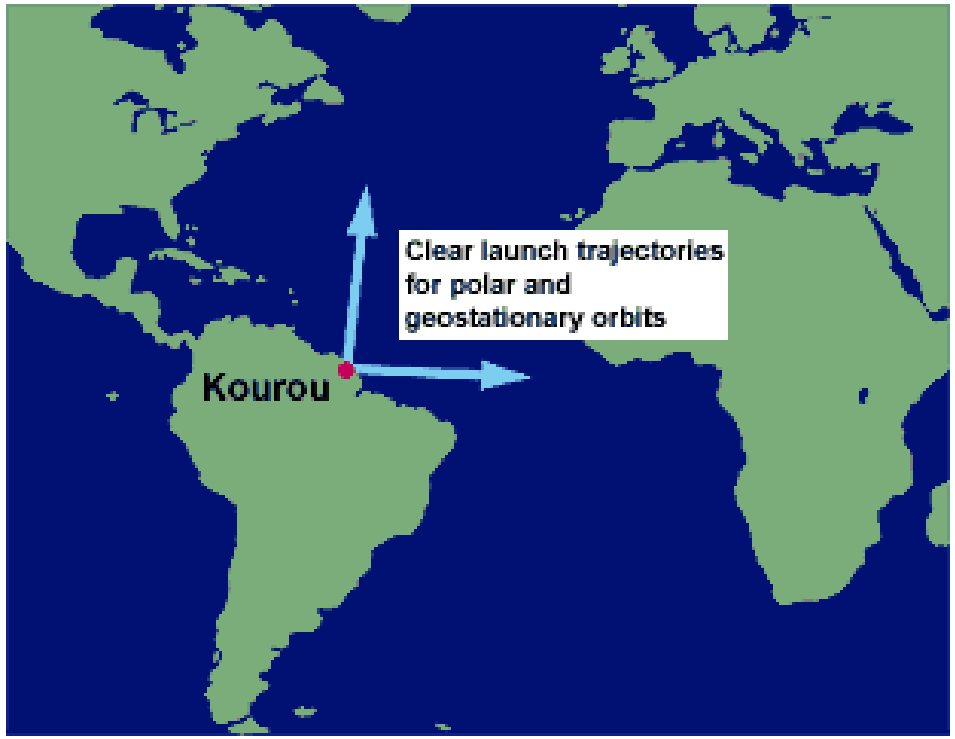


Other Insights

- The Launch sites in ex Soviet Union / (Russia/Kazakhstan) are also on the south, closer to the Equator *(source of picture: https://www.rferl.org/a/russia-cosmodrome-vostochny-rockets-space-troubled-project/27651519.html)*

Other Insights

- The main French launch site (Kourou) is in French Guinea, only 5° North from the Equator, being much more ideal for launces than US and Russia sites. *(Source of picture: https://www.eumetsat.int/launches-orbits)*



Appendix

GitHub link for the codes/notebooks:

https://github.com/devtpc/IBM-Data-Science/tree/main/10%20-%20Applied%20Data%20Science%20Capstone

Thank you!