IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thierry Baudez
8 March 2025

Outline

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

Executive Summary

- Summary of methodologies and --> results:
  - Data collection via SpaceX APIs and Wikipedia scraping
  - Data wrangling to clean and prepare data
  - EDA with data visualisation (plots) and SQL --> Booster version, payload masses, orbit types, launch sites identified as key features
  - Visualise launch sites on interactive maps with Folium --> Coastline-, and equator-based with train lines; civil infrastructure and settlements kept at a distance
  - Dashboard with Plotly Dash --> F1, 2000-6000 kg, CCAFS LSC-40 is the best combination
  - Predictive analysis via classification --> Decision Tree is the best classifier.

Introduction

- At SpaceX we save $100 million per re-usable rocket in the right circumstances.
- Our differentiator? The recovery of the first stage.
- Problem:
  - Which circumstances are conducive to recovery of the first stage?
  - How can we predict the successful recovery of the first stage in future launches?

Section 1

Methodology

Methodology

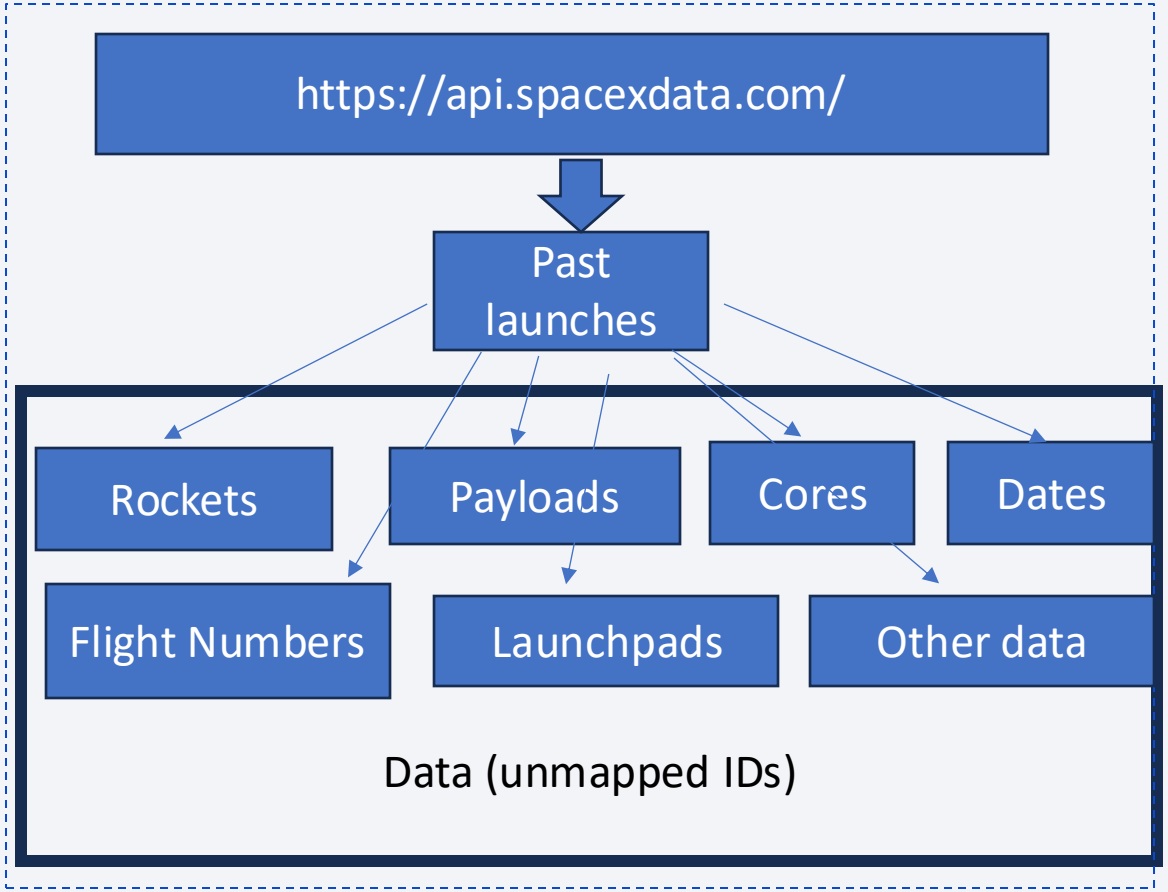
Executive Summary

- Data collection methodology
  - SpaceXdata API provides input for rockets, launchpads, payloads, cores, past launches
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API (1)

- SpaceXdata API provides input for past launches with detailed info on rockets, launchpads, payloads, cores, flight numbers and dates (see flowchart)

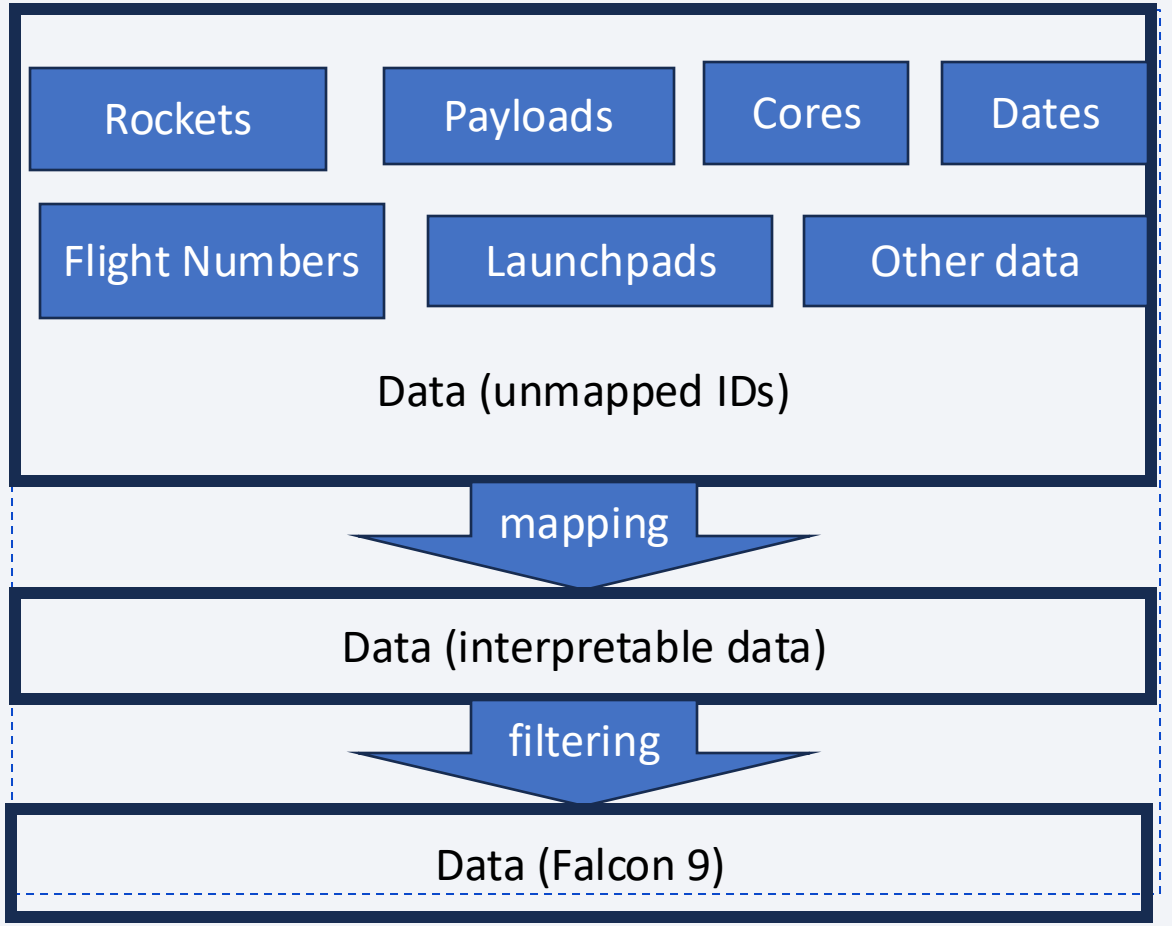
- Notebook available on GitHub: https://github.com/MrGeegor/CapStone_DataScience/blob/main/jupyter-labs-spacex-data-collection-api.ipynb
  (see upper right corner of each slide)



On GitHub

Data Collection – SpaceX API (2)

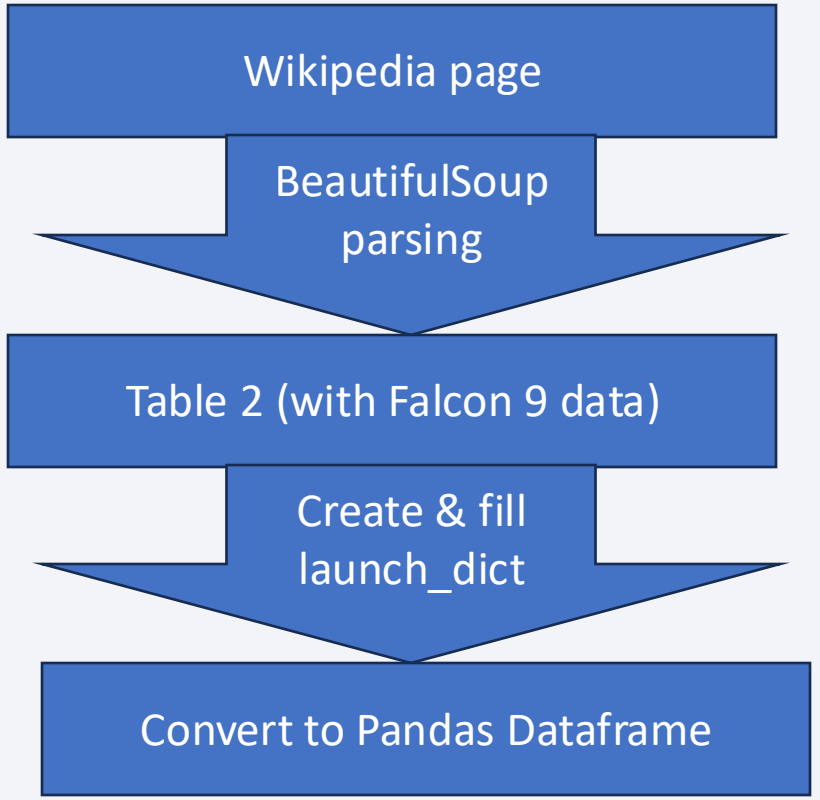
- Mapped ID-values to usable valuesl:
  - Rockets: booster names
  - Payload: mass of payload, orbit
  - Launchpad: name of launch site, longitude, latitude
  - Cores: landing outcome, landing type, # flights, gridfins, core reuse, legs, landing pad, core block, serial no
- Filtered on Falcon 9 data



On GitHub

Data Collection - Scraping

- Wikipedia page (9 June 2021 static) gets parsed with BeautifulSoup focusing on Falcon 9 launches
- Wikipedia page Table 2 html provided column names for empty dictionary launch_dict
- Filled dictionary with row value from the same Table 2
- Converted launch_dit to Pandas Dataframe



On GitHub

Data Wrangling (1)

- Do we have complete data?
- No. There are missing values in:
  - Payload Mass (5)
  - LandingPad (26)
- These missing values were replaced with mean values.

```
In [44]: data_falcon9.isnull().sum()

Out[44]: FlightNumber      0
         Date              0
         BoosterVersion    0
         PayloadMass       5
         Orbit             0
         LaunchSite        0
         Outcome           0
         Flights           0
         GridFins          0
         Reused            0
         Legs              0
         LandingPad       26
         Block             0
         ReusedCount       0
         Serial            0
         Longitude         0
         Latitude          0
         dtype: int64
```

On GitHub

Data Wrangling (2)

- How much data do we have on the various Launch Sites?
- Counting values with value_counts from the df.LaunchSite column

```
In [7]: # Apply value_counts() on column LaunchSite
        df['LaunchSite'].value_counts()

Out[7]: LaunchSite
        CCAFS SLC 40    55
        KSC LC 39A      22
        VAFB SLC 4E     13
        Name: count, dtype: int64
```

On GitHub

Data Wrangling (3)

- How much data do we have on the various Orbit Types?
- Counting values with value_counts from the df.Orbit column

```
In [8]:  # Apply value_counts on Orbit column
         df['Orbit'].value_counts()

Out[8]:  Orbit
         GTO      27
         ISS      21
         VLEO     14
         PO        9
         LEO       7
         SSO       5
         MEO       3
         HEO       1
         ES-L1     1
         SO        1
         GEO       1
         Name: count, dtype: int64
```

On GitHub

Data Wrangling (4)

- What are the landing outcomes?
- These results are saved in the newly created Outcome column (see visual).
- The outcomes with 'True' in their titles point to successful outcomes with a specific landing:
  - ASDS on a drone ship
  - RTLS on a ground pad
  - Ocean in the ocean
- 'False' and 'None' in the titles point to failures

```
In [9]:
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes

Out[9]:
Outcome
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: count, dtype: int64
```

Data Wrangling (5)

- Simplifying the results of outcomes to 2 classes: success (1) or failure (0)
- In newly created DataFrame column Class (see visual):
  - All 'False' and 'None' outcomes are saved as 0
  - All 'True' outcomes are saved as 1
- Additionally: we get to calculate the overall success rate: 66.6667%

```
In [19]: df.head(5)
```

Out[19]:

LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

On GitHub

EDA with Data Visualization

Finding the features affecting success rate (Class)

- Flight number vs Launch Site vs Class -> scatter plot (relationship identification)
- Payload Mass vs Launch Site vs Class -> scatter plot (idem)
- Success ratio (Class means) per Orbit Type -> bar chart (visualisation of sums for 1 feature)
- Flight number vs Orbit Type -> scatter plot (relationship identification)
- Payload Mass vs Orbit Type -> scatter plot (idem)
- Success rate over time -> line graph (visualisation of time-series with variance)

On GitHub

EDA with SQL

- Displayed the names of the unique launch sites
- Displayed 5 records where the launch sites begin with the string 'CCA'
- Displayed the toal payload mass carried by boosters launched by NASA (CRS)
- Displayed the average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome on a ground pad was achieved.
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster_versions which have carried the maximum payload mass.
- Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Ranked the count of landing outcomes between the dates 2010-06-04 and 2017-03-20, in descending order.

On GitHub

Build an Interactive Map with Folium

- Map objects added:
  - Circles to indicate the launch sites
  - Icons to create text labels that give the names of the launch sites
  - Green and red markers to indicate successful and failed launches respectively.
  - MarkerClusters to visually make it possible to have various green and red markers per location.
  - 'Mouse position' for immediate feedback on longitudinal and lattitudinal coordinates wherever the mouse pointer shows on the map.
  - Distance markers and lines to indicate with a line and a label the distance between any two points of interest.

On GitHub

Build a Dashboard with Plotly Dash

- Elements created:
  - A dropdown list with the 4 launch sites, allowing for a filtering of the graphs below according to the chosen location. By default 'All Sites' is selected, i.e. all launch sites.
  - A pie chart for an easy and quick visualisation of the success vs failure rate of launches for the selected site. By default it shows the success vs failure rate for all launch sites.
  - A slider to intuitively select the minimum and maximum range of the payload mass one wishes to explore on the following graph (see next point).
  - A scatter plot shows the correlation between payload mass and launch success, with colored differentiation by booster version.

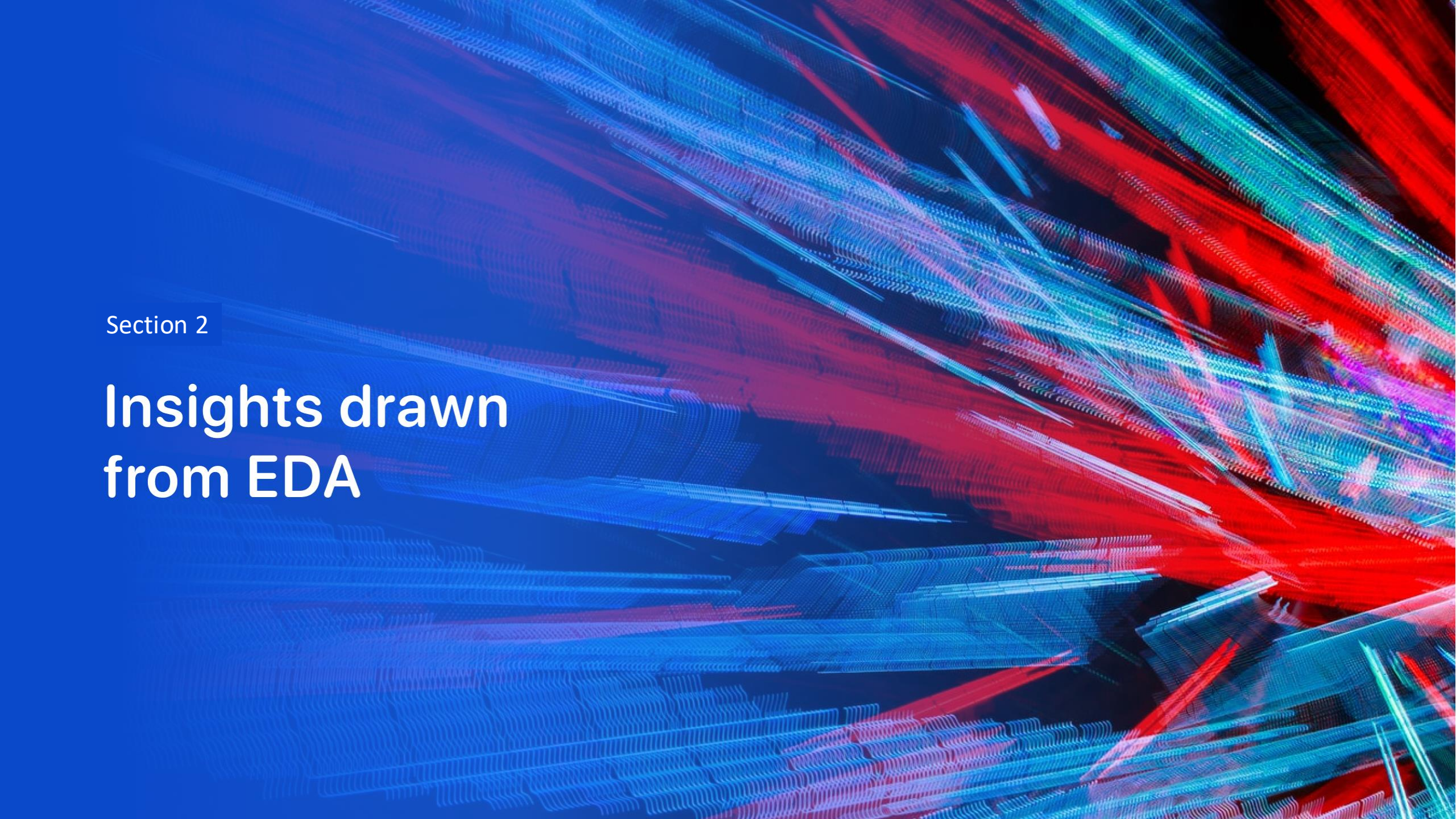
On GitHub

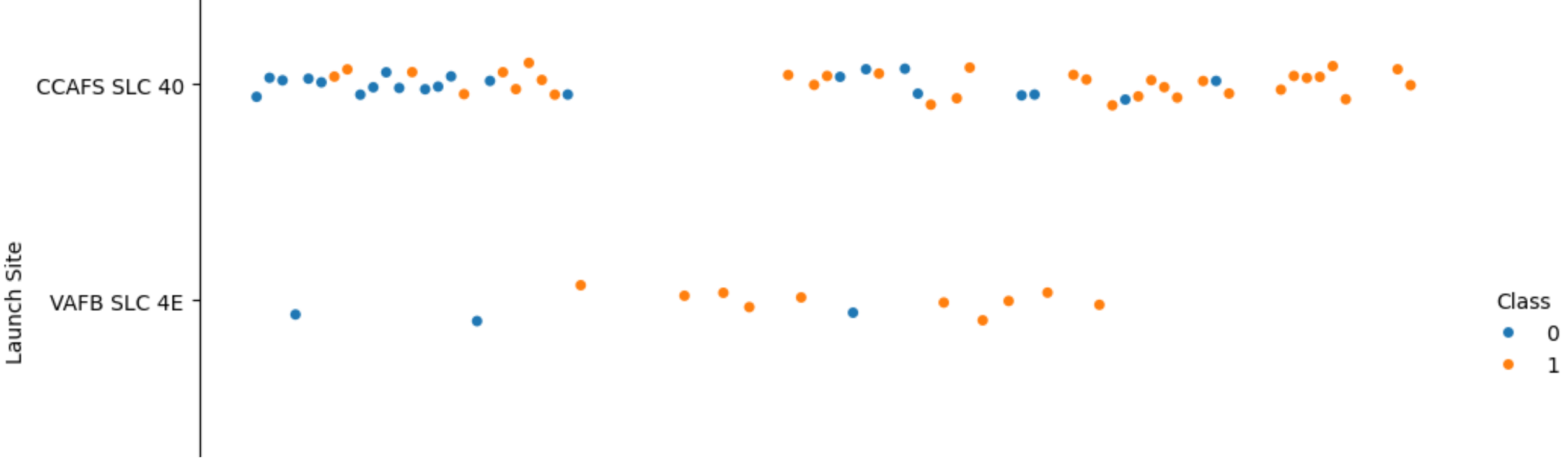
Predictive Analysis (Classification)

1. Defining X (=all columns - 'Class') and Y(='Class')
2. Perform a train, test, split with a 20% test size (i.e. 18 test samples)
3. Standardise both X_train and X_test data sets to supress features with large numbers
4. Next is pitting various models against each other with accuracy, R2 scores and confusion matrices. They all get trained through GridSearch Cross Validation (10 folds):
    - Logistic regression
    - Support vector machine
    - Decision tree
    - K nearest neighbours

Results

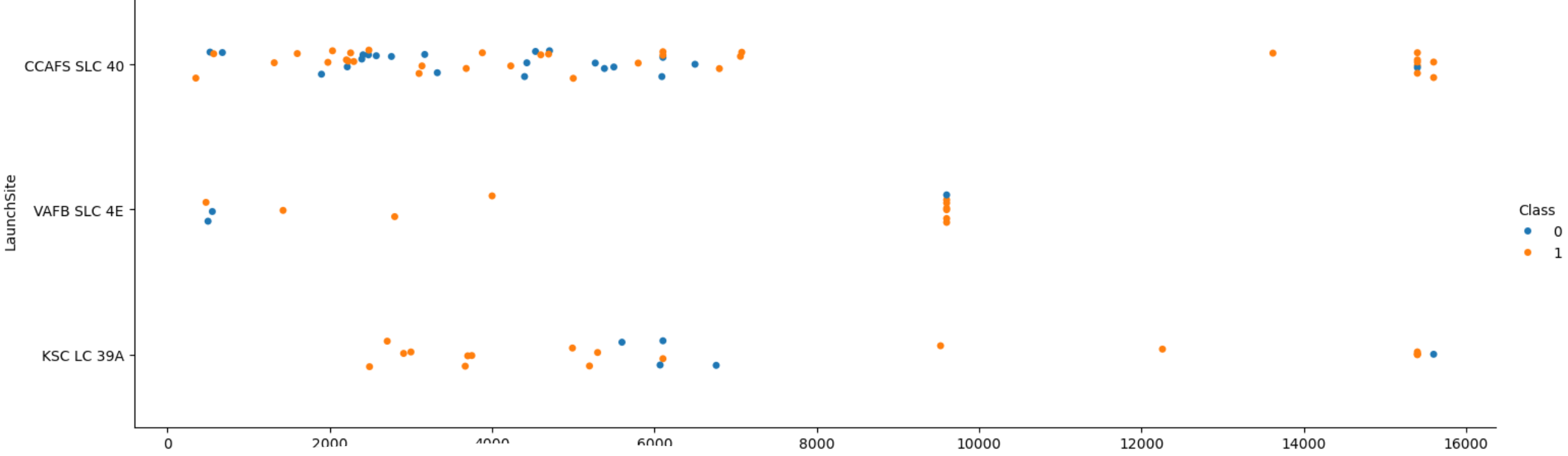
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA



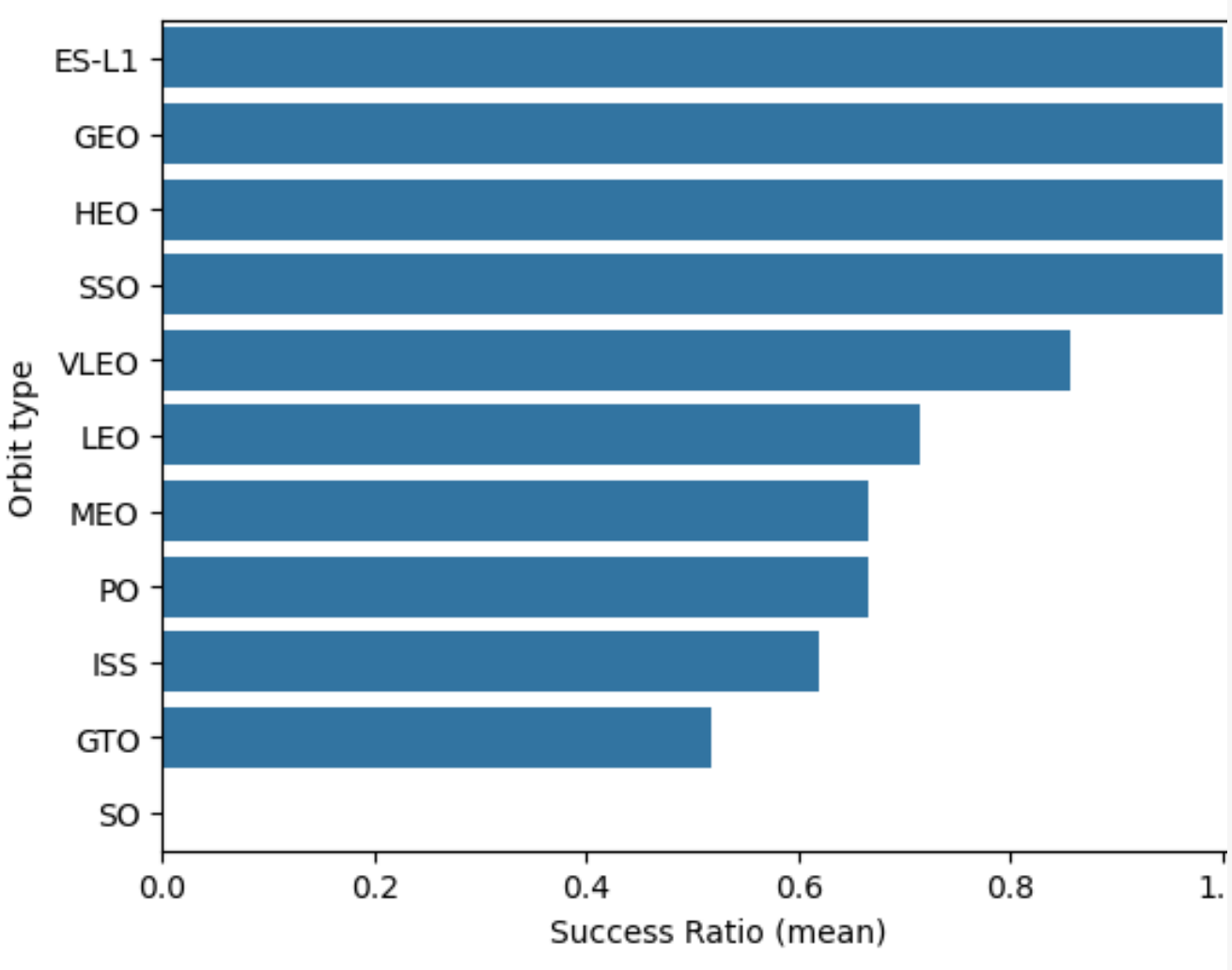
Flight Number vs. Launch Site

- Does the number of flights per launch site affect success ratio? (See scatter plot)
  - CCAFS SLC 40 had most flights. Earlier flights have a poor success ratio, with most attempts failing to land the first stage. Note the drastic improvement in success with the flight numbers in the 40-90 range.
  - VAFB SLC 4E was used the least: only 13 records. 3 failed to return the first stage (2 in the early flight numbers of 0-20, and 1 around 50). The rest were successful.
  - KSC LC 39A has 22 records: 5 failed (3 in the 20-40 flight number zone, 2 in the 70-80 zone). The rest were successful.
  - Summary: The later flight numbers perform much better than the initial ones, as each launch provides data feeding into the next one.



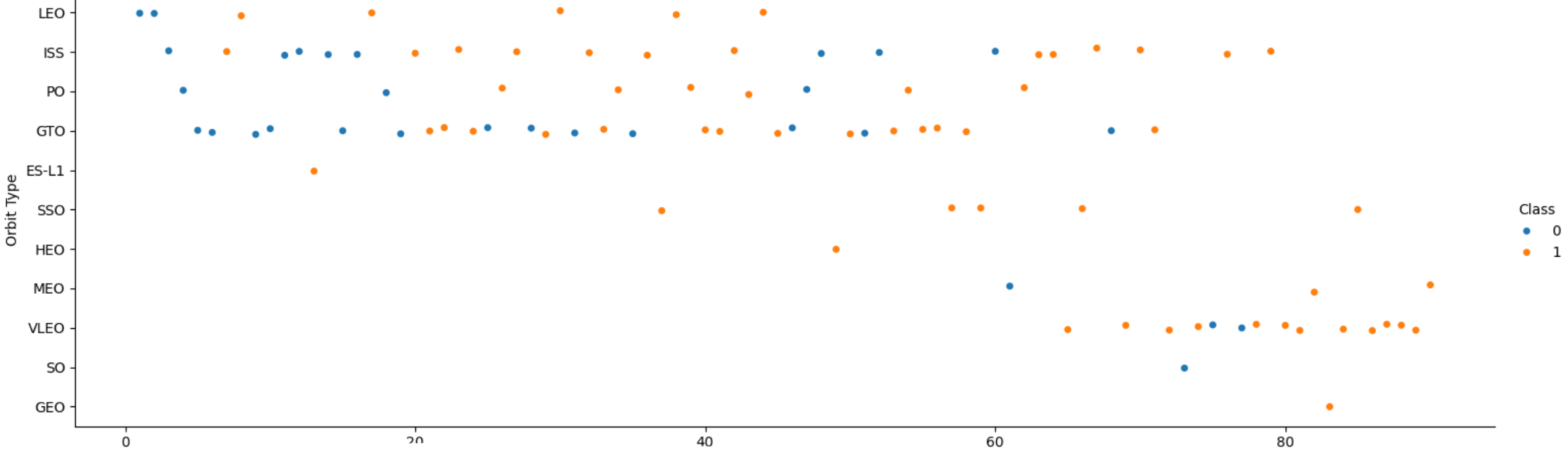
Payload vs. Launch Site

- Is there a relationship between payload mass and launch sites? (See scatter plot)
- Yes: certain launch sites lend themselves more to higher payload masses than others. This might be related to the Orbit Types that each launch site lends itself to.
- Note how there are no launches on VAFB-SLC 4L for payloads above 10000 kg



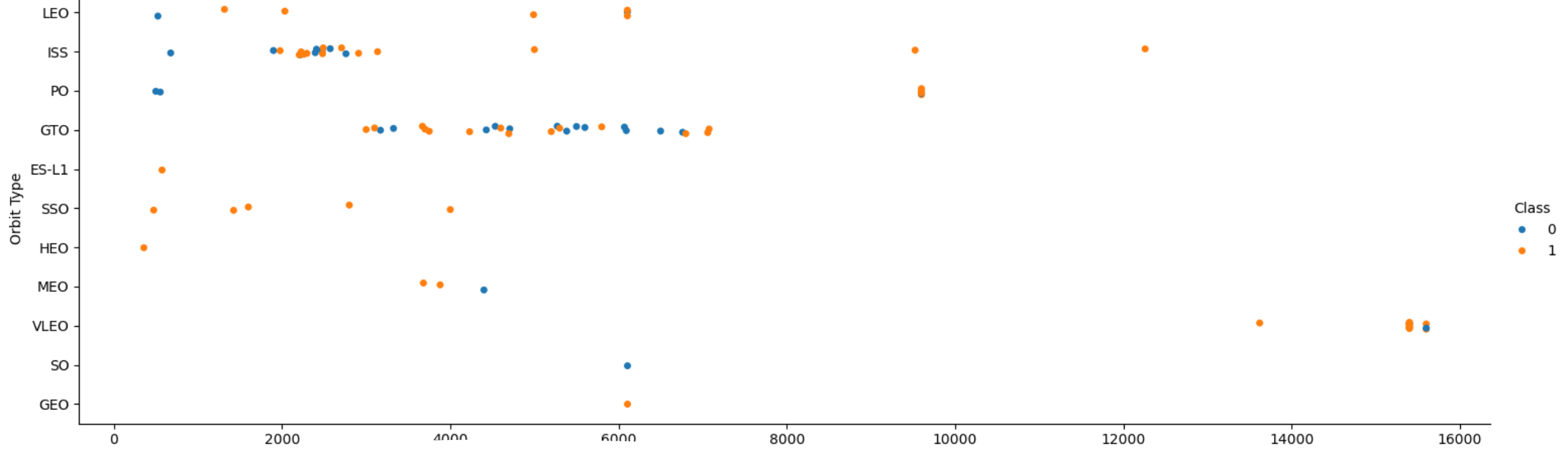
Success Rate vs. Orbit Type

- Does the orbit matter for the success rate? (See bar chart)
- Yes: certain orbits have a perfect score, while others show mixed results:
- Class
- Orbit
- ES-L1 1.000000
- GEO 1.000000
- HEO 1.000000
- SSO 1.000000
- VLEO 0.857143
- LEO 0.714286
- MEO 0.666667
- PO 0.666667
- ISS 0.619048
- GTO 0.518519
- SO 0.000000



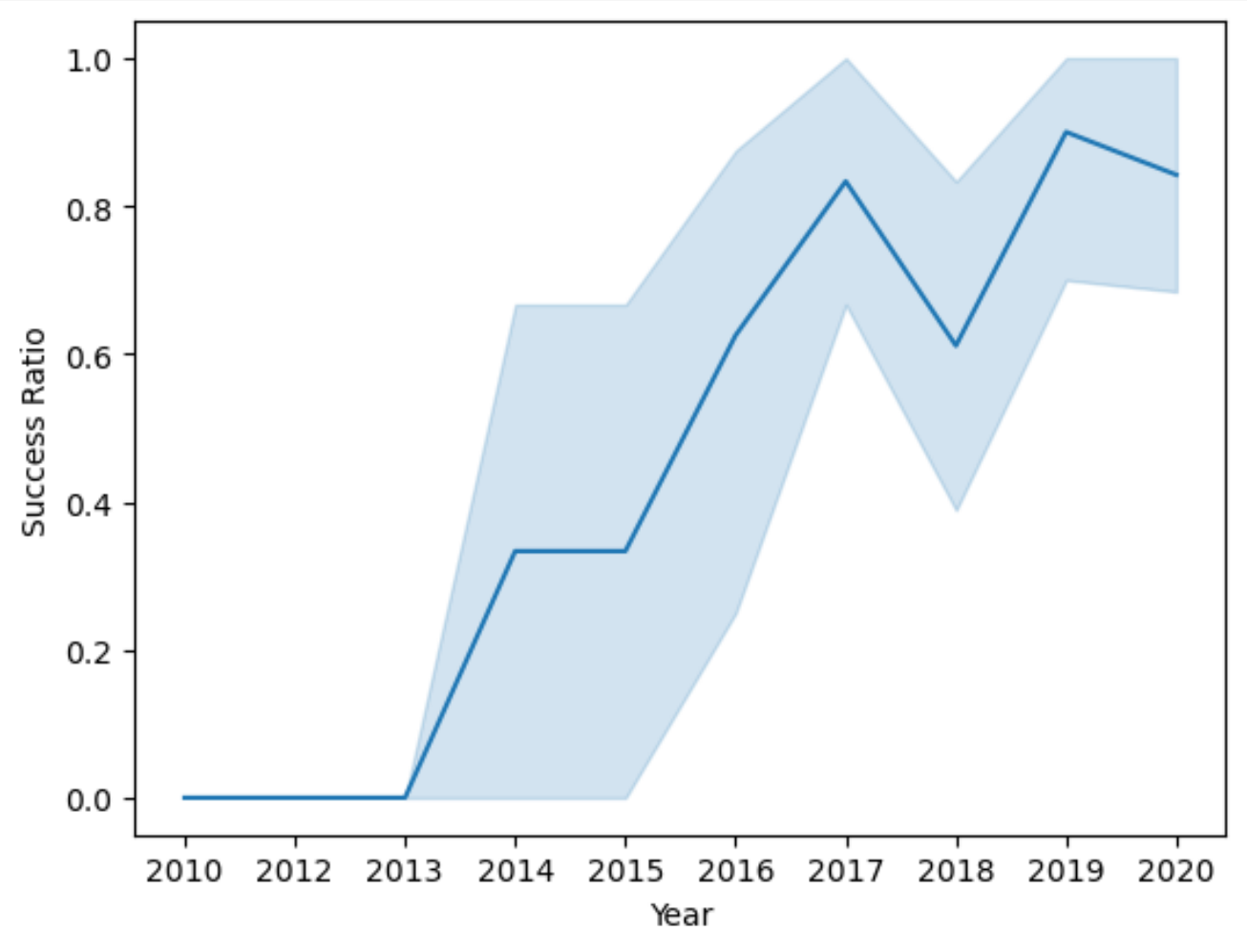
Flight Number vs. Orbit Type

- Do the scores for the Orbit Types hold true in light of repeated flights? I.e. which Orbit Types show consistency?
  (See scatterplot)
- It becomes clear that the results for GEO, SO, HEO, ES-L1 issue from single flights. These results don't show any reliable info on success or failure.
- The most successful Orbit Types are:
  - SSO -> 1.00000
  - VLEO -> 0.857143
  - LEO -> O.714286



Payload vs. Orbit Type

- Do the scores for the Orbit Types relate to the Payload Masses? (See scatterplot)
- There tends to be a higher success rate with heavier Payload Masses.
- One exception is GTO, where the results are mixed across Payload Masses.



Launch Success Yearly Trend

- Is the assumption that the rate improves over time actually correct?
- Yes. The trend is clearly upwards from 2013 onwards.

All Launch Site Names

Out[10]:

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

Running an SQL query for distinct Launch_Site values gives 4 unique Launch Site names.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Launch Site Names Begin with 'CCA'

- The CCA string can be found by adding % as a placeholder symbol in the SQL query.
- The query displays the first 5 entries (see visual)

Total Payload Mass

- Total Payload Mass launched by NASA (CRS): 45596 kg
- Given that the majority of the payloads are between 2000 and 8000 kg, as observed in the EDA section, it is probable that NASA has supported SpaceX at least 5 times, and maybe as much as 20 times.

In [26]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE CUSTOMER LIKE 'NASA (CRS)';
```

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

Out[26]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534.67 kg
- That indicates that other boosters might be preferred for larger payloads.

In [15]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

Out[15]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

```
[16]: %%sql
      SELECT MIN(Date) FROM SPACEXTABLE
      WHERE Landing_Outcome LIKE 'Success (ground pad)';

       * sqlite:///my_data1.db
      Done.
[16]:  MIN(Date)
      ------------
      2015-12-22
```

First Successful Ground Landing Date

- The first successful landing on a ground pad happened on 22 December 2015.

[26]:
```
%%sql
SELECT Booster_Version FROM SPACEXTABLE
WHERE (Landing_Outcome LIKE 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ >4000
    AND PAYLOAD_MASS__KG_ <6000);
```

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

[26]:

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2
- Note that the payloads carried by these F9 boosters are significantly higher already than the average payloads carried by the F9 v1.1 boosters viewed earlier.

```
[19]: %sql SELECT DISTINCT Mission_Outcome FROM SPACEXTABLE;
```

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

[19]:

Mission_Outcome
Success
Failure (in flight)
Success (payload status unclear)
Success

```
[20]: %%sql
SELECT COUNT(Mission_Outcome) FROM SPACEXTABLE
WHERE Mission_Outcome LIKE '%Failure%';
```

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

[20]:

COUNT(Mission_Outcome)
1

```
[21]: %%sql

SELECT COUNT (Mission_Outcome) FROM SPACEXTABLE
WHERE Mission_Outcome LIKE '%Success%';
```

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

[21]:

COUNT (Mission_Outcome)
100

Total Number of Successful and Failure Mission Outcomes

- Failure = 1
  Success = 100
- Mission outcome and landing outcome measure different things. While many landing outcomes were failures, they were so by design. I.e. they failed on purpose, which then gets noted as a mission success.
- The track record of 1 mission failure vs 100 mission successes is very impressive.

[27]:
```
%%sql

SELECT Booster_Version FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ =
    (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

[27]:

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

Boosters Carried Maximum Payload

- The boosters which have carried the maximum payload mass:
  - 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
- We see many iterations of the F9 B5 model used to carry maximum payloads.

```
%%sql
SELECT
    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)';
```

```
 * 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

2015 Launch Records

- These are the details found:

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

- The totaled ranking of landing outcomes confirms that SpaceX approach to re-using phase 1 of rockets only become successful iteratively. They had controlled successes and failures on various types of platforms, controlled and uncontrolled crashes into the ocean and an exercise with parachutes. But all of these were accompanied by 10 landings where they did not even attempt to recover phase 1. Presumably, at the beginning of their learning curve; testing, validating and improving their approach.

In [25]:
```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Landing_Outcome_Count
    FROM SPACEXTABLE
WHERE
    Date<'2017-03-20' AND Date >'2010-06-04'
    GROUP BY Landing_Outcome
    ORDER BY Landing_Outcome_Count DESC;
```

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

Out[25]:

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

Section 3

Launch Sites Proximities Analysis

4 Launch Sites: 1 West, 3 East

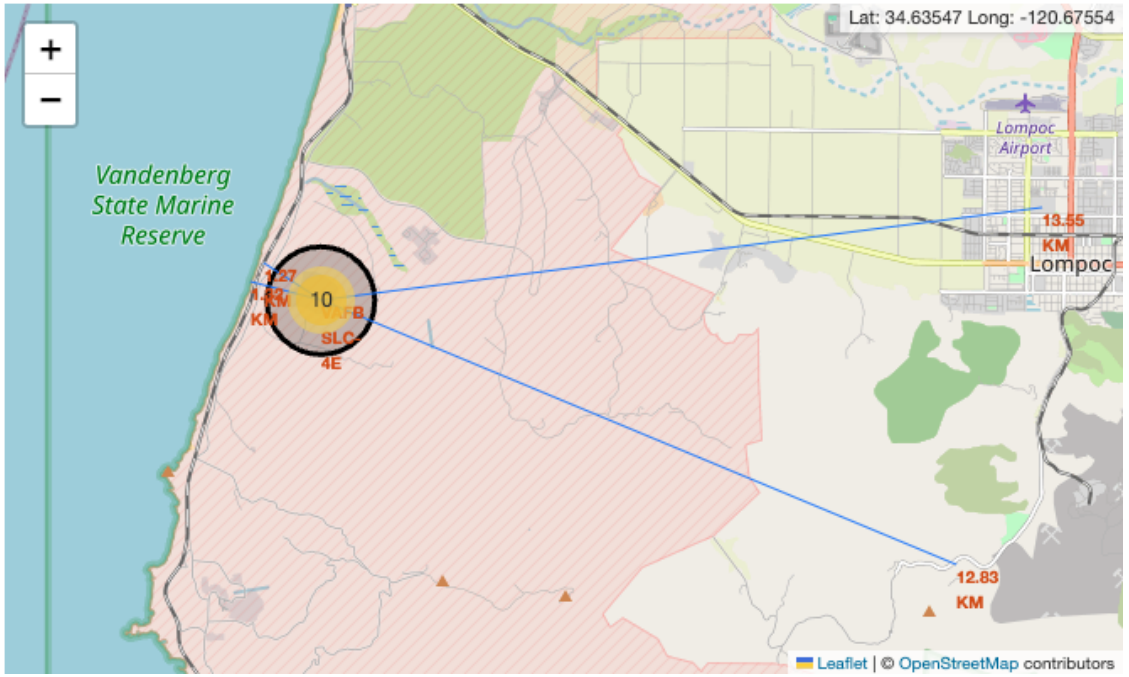
- All launch sites are situated at the coast.
  - VAFB SLC-4E is on the West-coast
  - KSC LC-39A, CCAFS LC-40 and CCAFS SLC-40 are on the East coast. In fact they are so close to each other that on this map they all feature as just one dot for all three.

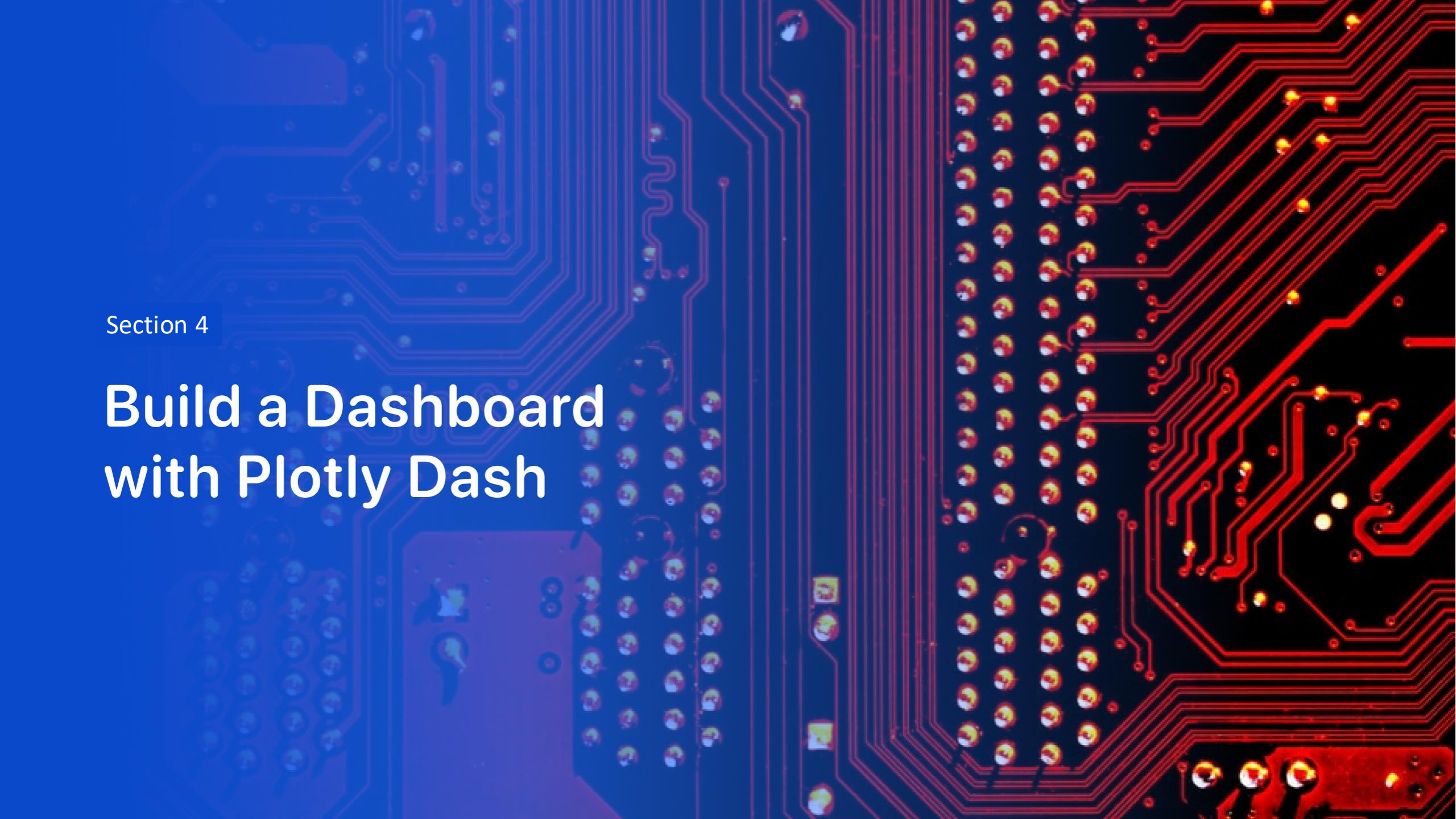
Launch Successes Visualised per Launch Site

- MarkerClustering allows for the visualisation of many labels for the same coordinates. In this case: successful (green) vs failed (red) launches per launch site.
- Visualised here is the 40% success / 60% failure launch rate for the VAFB SLC-4E launch site.

Launch Site Proximity to Infrastructure

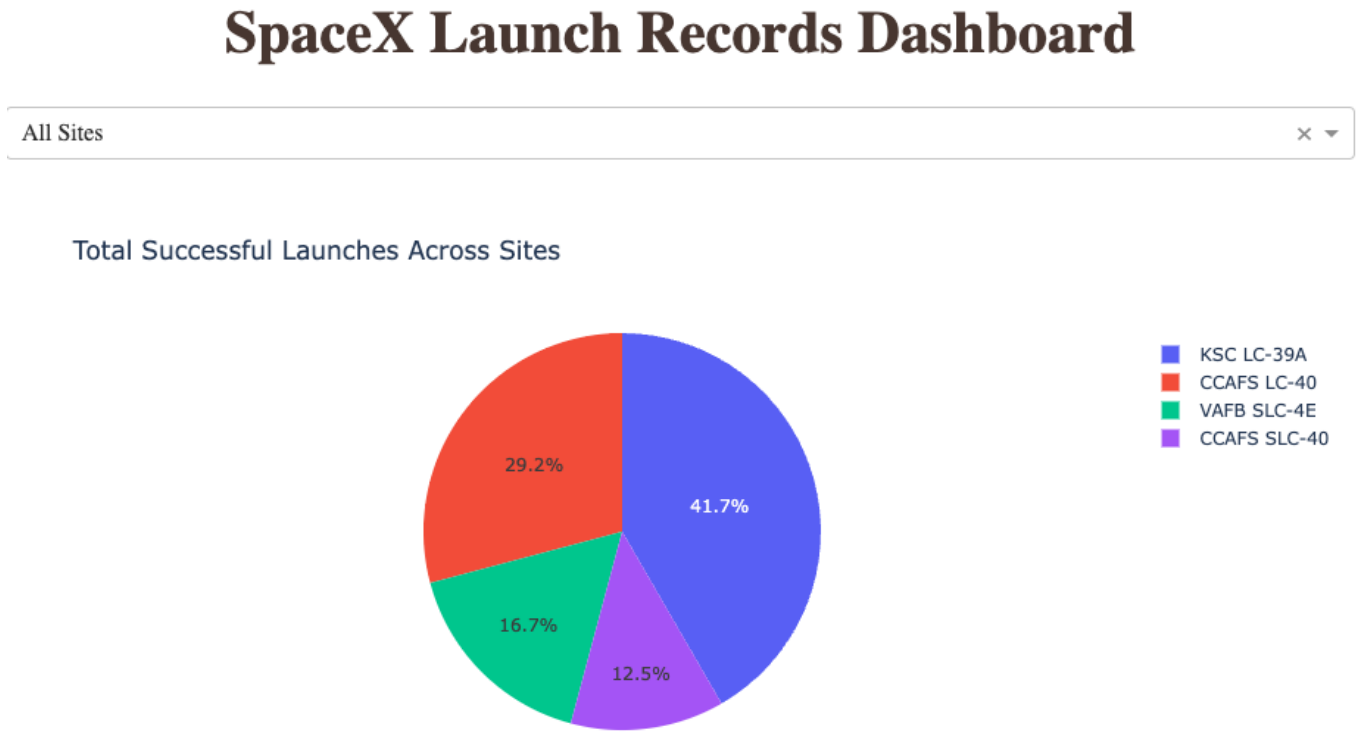
- Proximity of launch site VAFB SLC-4E to
  - Train tracks: 1.27 km
  - Coast: 1.32 km
  - Motorway: 12.83 km
  - Nearest town: 13.55 km
- The launch site is a solid 10+ km away from civilians (motorway, town)
- A train track and the coast are in close proximity, which could allow for large transports and pose less of a risk in case of incidents at the launch site.

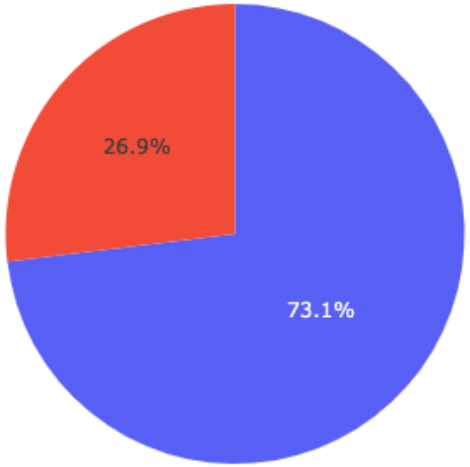



Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches Across Sites

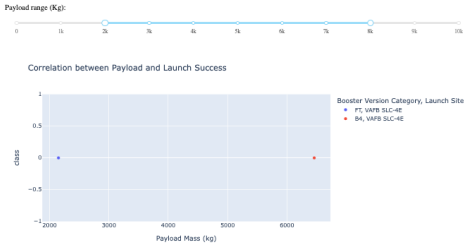
- 41.7% of all successful launches happened from KSC LC-39A. That's more than 4 out of 10; indicating that the characteristics of this launch site (orbit types, payload mass, …) lend itself particularly well to successful launches.
- With 29.2% nearby CCAFS LC-40 is a solid number 2. The fact that CCAFS SLC-40, which is almost at the same location, garnered only 12.5% suggests a different infrastructure or capabilities for both locations.
- VAFB SLC-4E is the sole West-coast location with 16.7% of the successful launches.

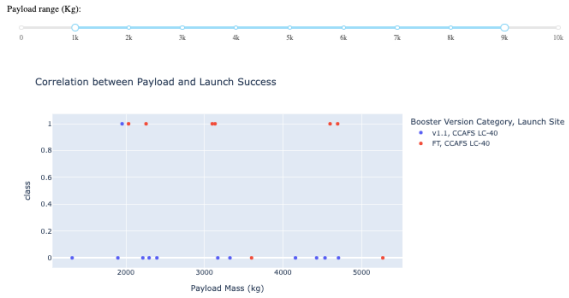




Success vs Failure at CCAFS LC-40

- This pie chart puts into perspective the overall 41.7% success ratio for launches from CCAFS LC-40.
- Those 41.7% constitute 73.1% of the successful launches, with 26.9% allowed to fail.





Payload Mass vs Success

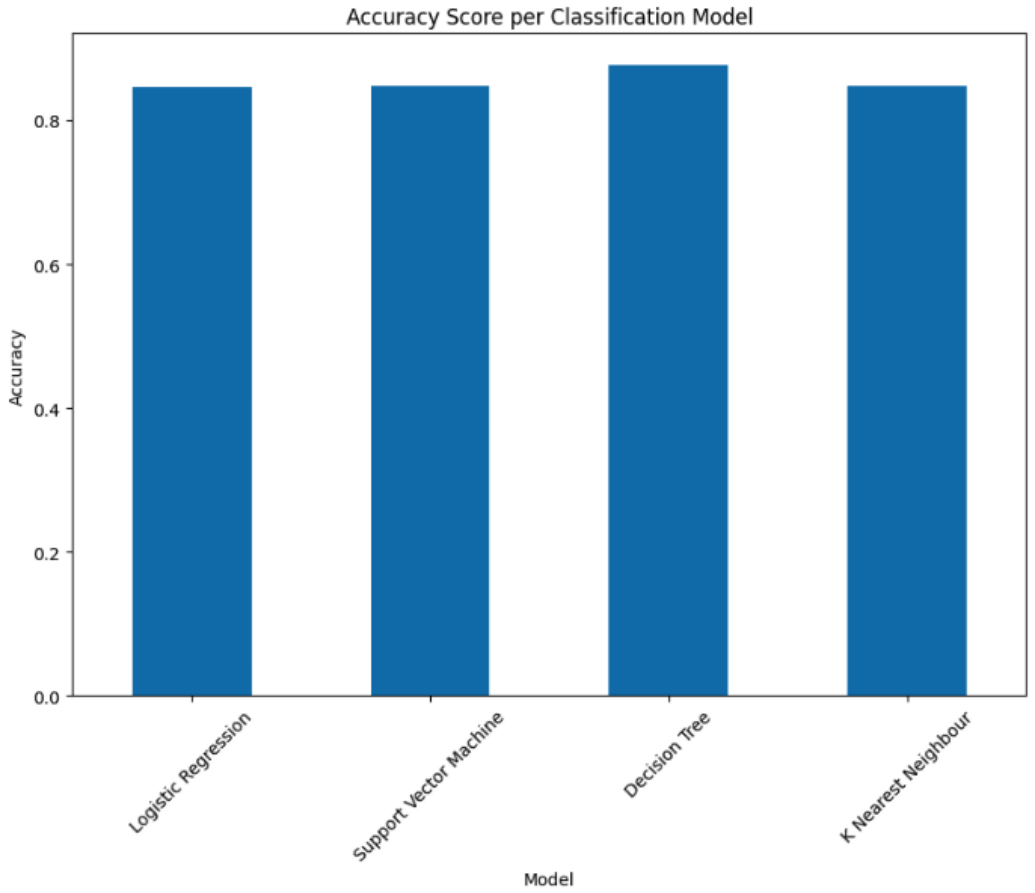
- The top 3 combinations of booster versions, payload mass and launch sites for successful returns of phase 1 are:
    1. FT boosters carrying between 2000 and 6000 kg payload departing from CCAFS LC-40
    2. FT boosters carrying between 2000 and 6000 kg payload departing from KSC LC-39A (closer to 6000 kg and above these boosters were mostly used on this launch site for failure testing)
    3. B4 boosters carrying between 2000 and 6000 kg payload departing from KSC LC-39A
- What is the 'best' combination for failure?
    1. Head and shoulders above the rest V1.1 boosters carrying between 0 and 6000 kg departing from CCAFS LC-40

Section 5

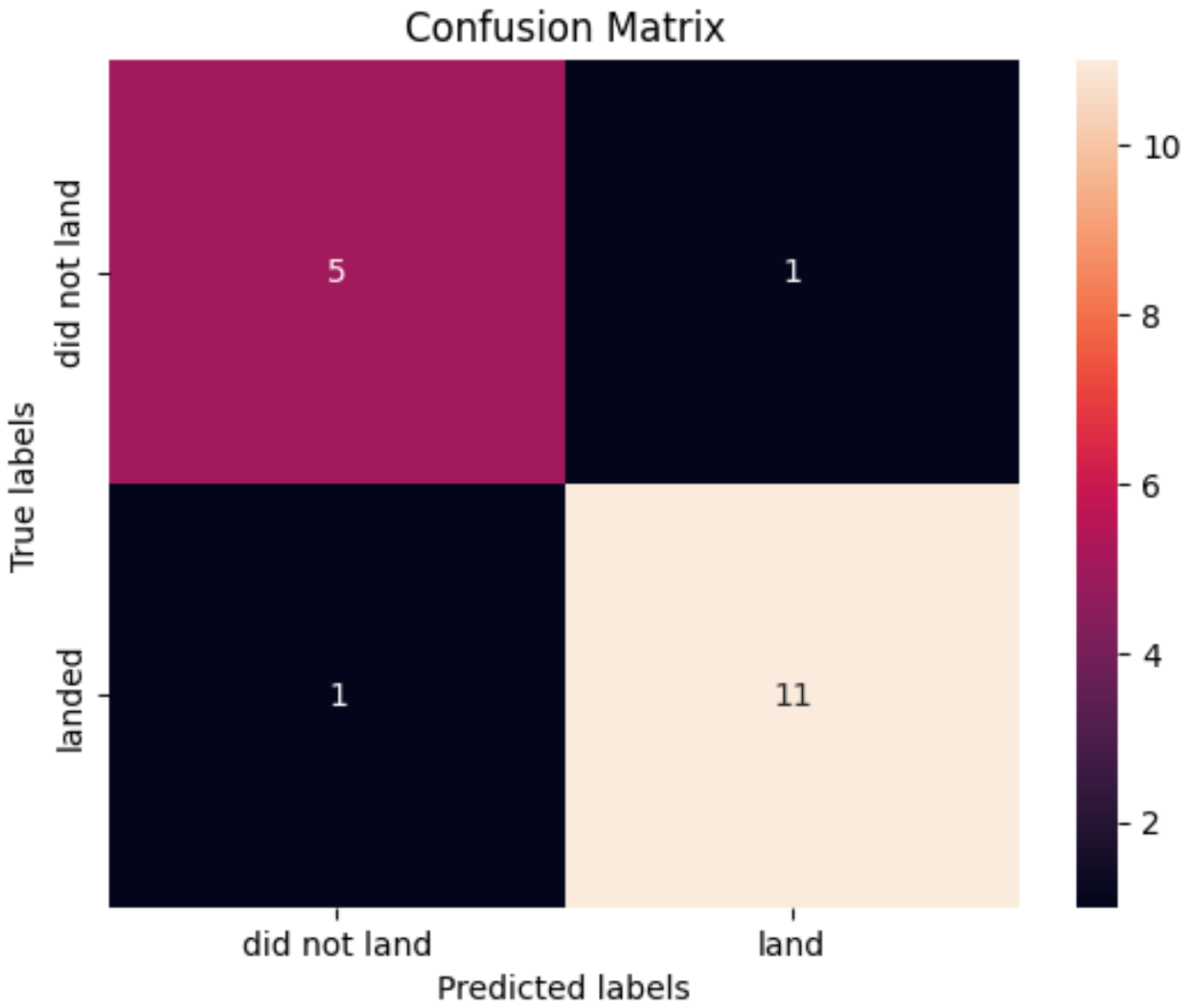
Predictive Analysis (Classification)

Classification Accuracy

- The decision tree offers the highest accuracy at 87.68%.
- The other models come close with scores around 84-85%



	Model	Accuracy
0	Logistic Regression	0.846429
1	Support Vector Machine	0.848214
2	Decision Tree	0.876786
3	K Nearest Neighbour	0.848214



Confusion Matrix

- The decision tree model still makes mistakes, though the mistakes are much cheaper than with the other models:
  - Other models predict 3 false positives and 0 false negatives; this model reduces the (very costly) false positives.
- The true negative score is higher here (from 3 for other models to 5 here), overall strengthening the reliability of the model for financial decisions.

Conclusions

- Experience with booster versions, payload masses and orbit types allowed by launch sites generally increase the successful landing of phase 1 rockets. These are all significant features to include into prediction models.
- Overall the SpaceX team is adept at predicting whether a launch will have a successful landing or not with 1 failed prediction vs 100 successful predictions in the studied timespan.
- The high success came at a cost, since next to the various landing attempts in different circumstances, they also launched 10 rockets without any expectation of recovery at all.
- The 4 SpaceX launch sites are located at the East (3) and West (1) coastlines of the US, near the equator. All locations are away from civilian towns or motorways, while still connected to rail infrastructure and water ways.
- The best combination of booster version, payload masses and launch site in the studied period is F1, 2000-6000 kg, CCAFS LSC-40.
- Our best predictive model is the decision tree classifier with an accuracy of 87.68%, with a clear improvement over other models for correctly reclassifying (very costly!) false positives as true negatives.

Appendix

Thank you!