IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Reinis R. Ruza
2022/11/09

Outline

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

Executive Summary

- Our startup, Space Y, has applied a data science approach to predict Falcon 9 rocket first-stage booster-retrieval rate (BRR) with high accuracy.
- To do so, we utilised the SpaceX API and scraped information available online to collect information on all SpaceX launches, such as booster type, payload mass, target orbit, launch site, booster retrieval success (BRS), and others. The data was cleaned up and used for exploratory data analysis, which showed that although some parameters might influence BRS, their individual contributions to the mission outcome were low.
- Because of this, we used parameter optimisation via grid search for different classification algorithms. The ones examined here were logistic regression, support vector machine, decision trees, and K-nearest neighbours. The best performing parameters for each algorithm were then used to score their performance, using an 80%/20% train/test split of our dataset. All algorithms performed equally well on the testing subset, although there were only 18 entries in it, so more data would be required to accurately estimate model accuracy

Introduction

- The success of SpaceX by the introduction of reusable first-stage boosters has shown that the space industry can provide lucrative business opportunities. Despite its achievements, SpaceX only has a 66% booster retrieval rate (BRR), thus driving up cost of its launches.
- If Falcon 9 booster retrieval success could be predicted for each mission, this would enable our startup, SpaceY, to provide more accurate costs for each mission, as well as allow our clients to make adjustments to their mission parameters based on their risk tolerance.
- Therefore, the principal question of this research was –
  - **Can we predict whether Falcon 9 booster retrieval for a specific mission will be successful, and can we do so with high accuracy?**

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API, Web Scraping
- Perform data wrangling
  - Data filtering, formatting, filling out missing datapoints
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Parameter gridsearch, evaluation of different models

Data Collection

SpaceX API

- Past launch data downloaded using SpaceX API as a .json file using *requests.get()*
- .json file converted into a Pandas dataframe using *pd.json_normalize*
- Most of the past launch data, such as rocket used and launch site identified by ID – convert to real values by making further calls to the SpaceX API

Web Scraping

- Downloaded Falcon 9 launch Wikipedia entry using *requests.get()*
- Converted downloaded page to a BeautifulSoup object
- Identify table containing launch data using *soup.find_all(name='table')*
- Identify the table columns, then parse each row, adding each value into a dictionary
- Finally, dictionary converted into a Pandas dataframe

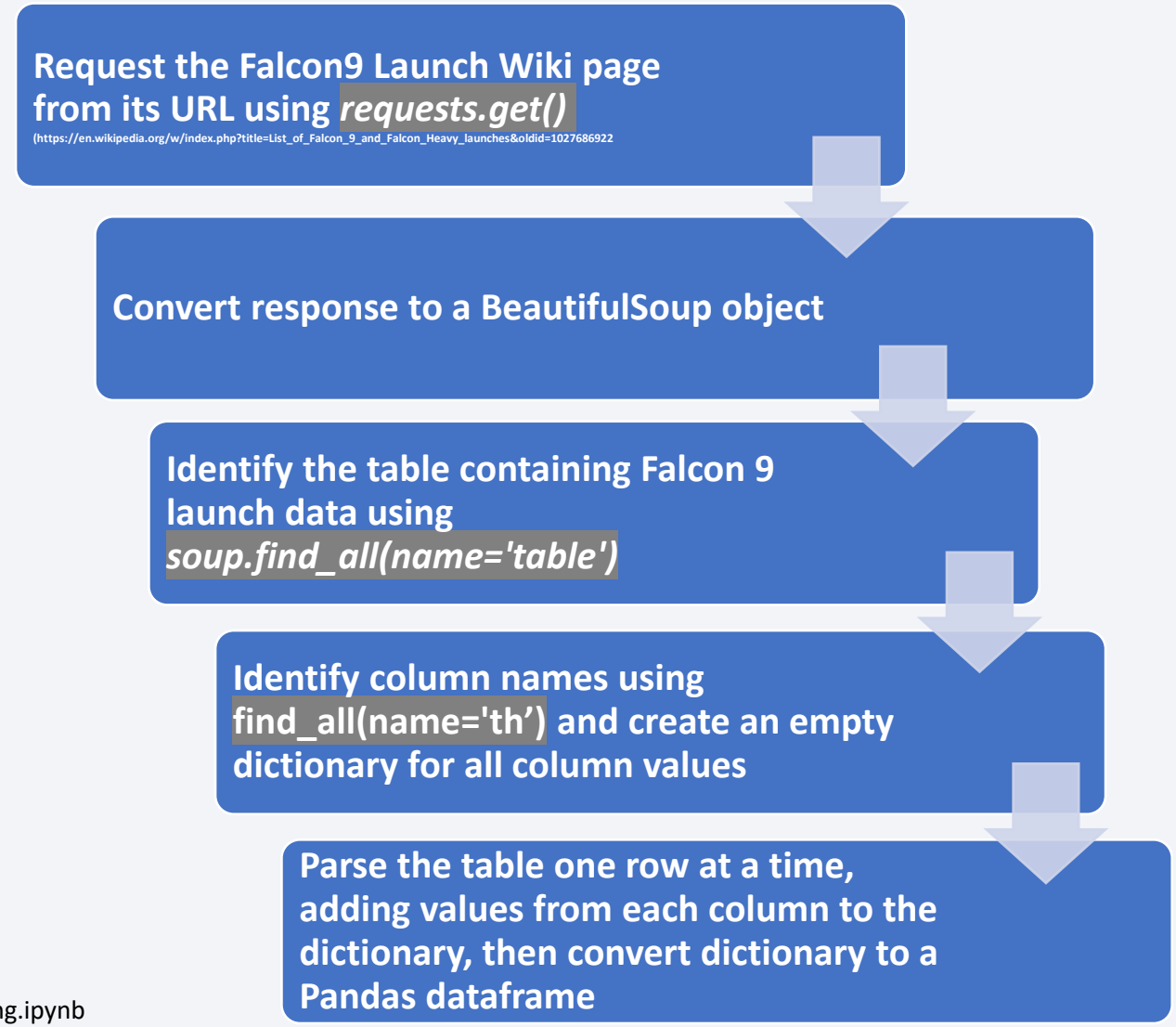
Data Collection – SpaceX API

- To collect data using the SpaceX API, made several REST calls using *requests.get()*
- First downloaded past flight data – where parameters for each launch were identified by ID values
- To obtain real values behind each ID, made 4 separate REST calls to different parts of the API
- All data was stored in a dictionary, which was then converted into a Pandas dataframe

https://github.com/Leocadion/IBM-Data-Science-Final-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- To obtain more information about each launch, data was scraped from Wikipedia
- After downloading webpage HTML and converting it to a BeautifulSoup object, the table containing relevant info was identified
- The table was then parsed and its values written into a dictionary
- Finally, the dictionary was converted into a Pandas dataframe



https://github.com/Leocadion/IBM-Data-Science-Final-Project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Dataframe rows containing information for boosters other than Falcon 9 were dropped

Removed rows with multiple cores or payloads

Datetime values were converted to date only using *pd.to_datetime()*

Checked how many null values were present in each column using *df.isnull().sum()*

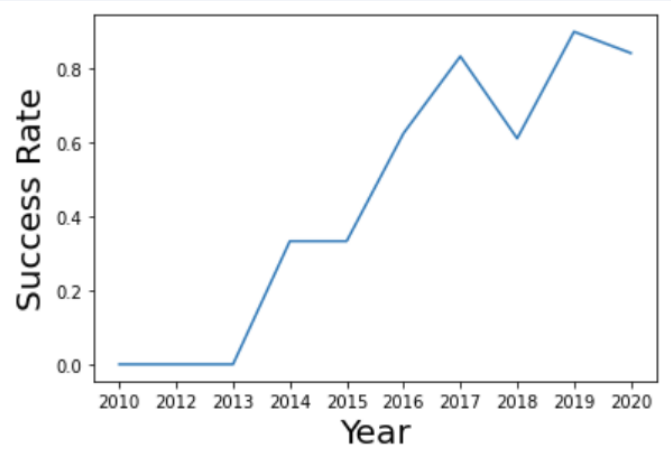
This identified 5 missing values in the PayloadMass column and 26 missing values in the LandingPad column

Calculated mean payload mass using *.mean()*, and replaced missing payload values with the mean value using *.replace(np.nan, mean)*

https://github.com/Leocadion/IBM-Data-Science-Final-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Success rate of launches seemed to increase with the flight number (see fig.), therefore created several plots to see what other parameters might have affected this:
  - Catplot: flight number vs payload mass (coloured by success)
  - Catplot: flight number and launch site (coloured by success)
  - Catplot: flight number and target orbit (coloured by success)
- Then plotted further graphs looking at relationships between these parameters:
  - Catplot: payload mass and launch site (coloured by success)
  - Catplot: payload mass and target orbit (coloured by success)
  - Bargraph: average success rate for each target orbit



https://github.com/Leocadion/IBM-Data-Science-Final-Project/blob/main/jupyter-labs-eda-dataviz.ipynb

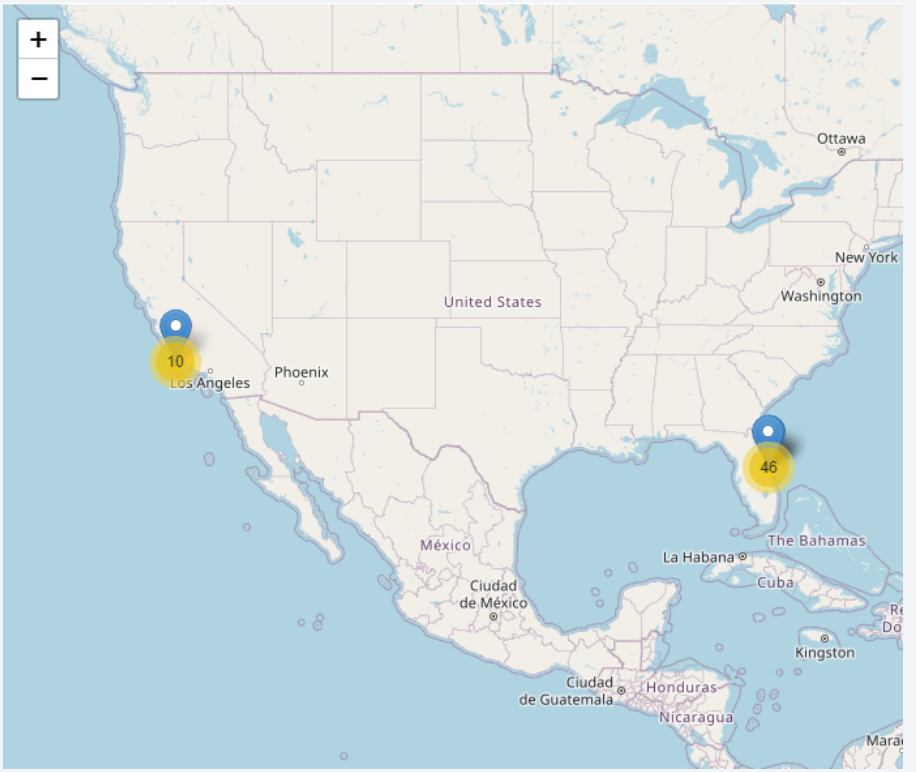
EDA with SQL

- Performed SQL queries to:
  - Identify the different launchsites used for flights (SELECT DISTINCT launch_site from spacex)
  - Select launches from CCAFS LC-40 and SLC-40
  - Identify 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 successful 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for 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, in descending order

https://github.com/Leocadion/IBM-Data-Science-Final-Project/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

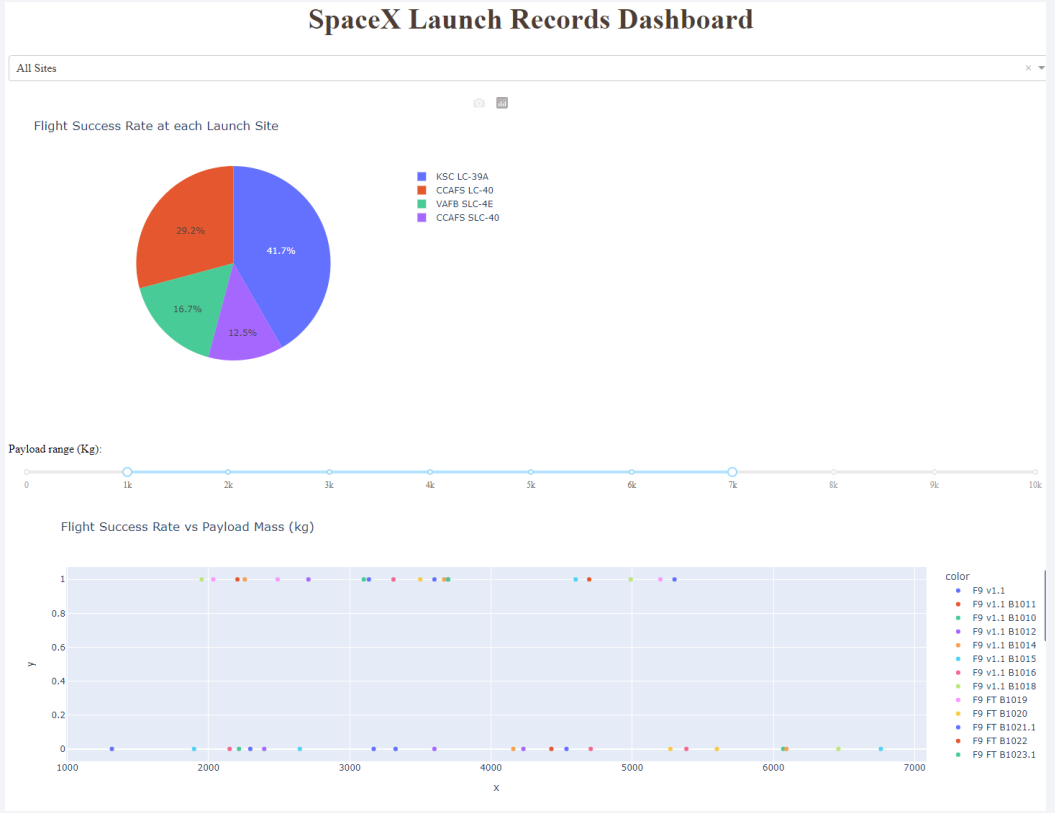
- Created an interactive map of all Falcon 9 launch sites to identify any geographical determinants of launch success:
  - Added markers of launch sites
  - Added markers indicating failed/successful launches at each site
  - Measured distances from launch sites to geographically important locations, such as highways, cities, and coastlines, added corresponding lines and distance markers.



https://github.com/Leocadion/IBM-Data-Science-Final-Project/blob/main/lab_jupyter_launch_site_location.ipynb

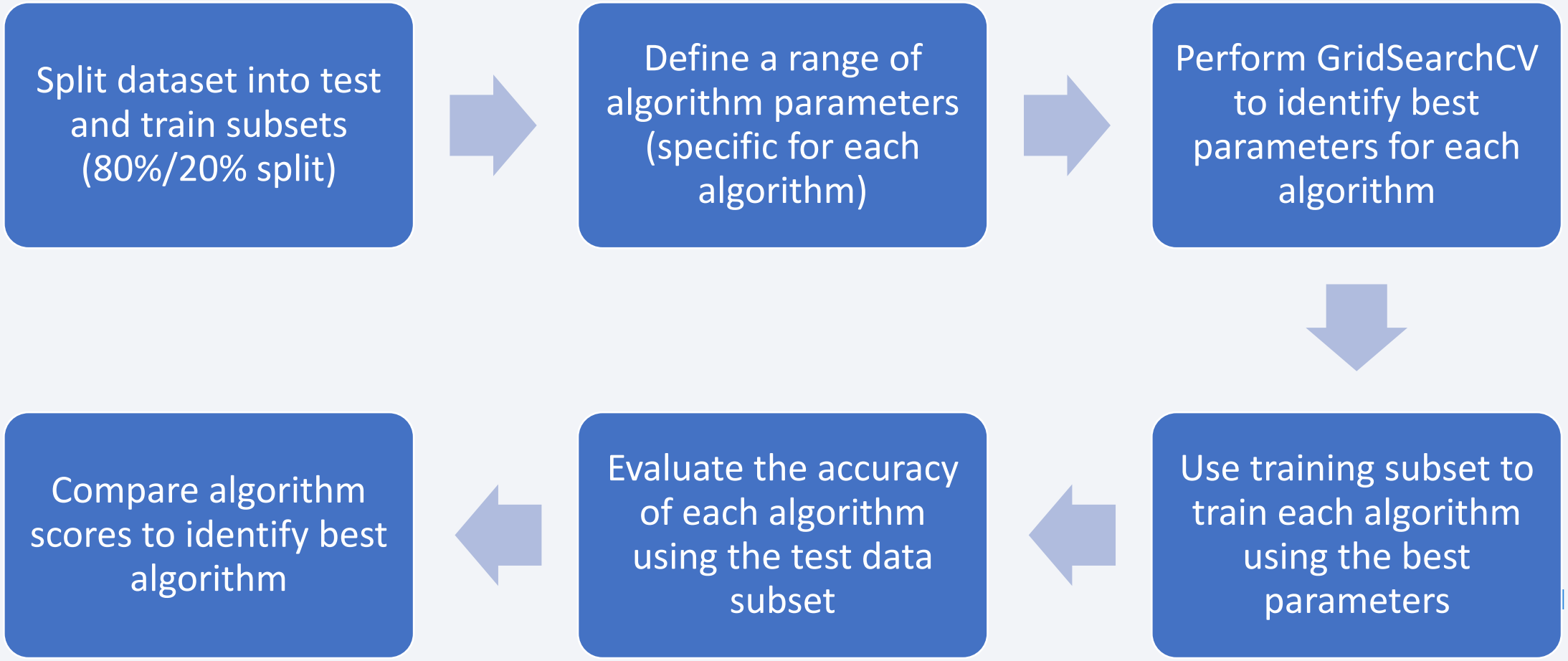
Build a Dashboard with Plotly Dash

- Created an online dashboard to facilitate exploration of the relationship between launch site, payload weight, booster type, and success rate
- Allows selection of specific launch site via dropdown bar
- Shows launch success rate at specified sites
- Also has a scatterplot showing flight success rate vs payload, coloured by booster type
- Payload mass range visualised can be adjusted via slider.



Predictive Analysis (Classification)

- Tested logistic regression, support vector machine, decision tree and K-nearest neighbour classification algorithms.



https://github.com/Leocadion/IBM-Data-Science-Final-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

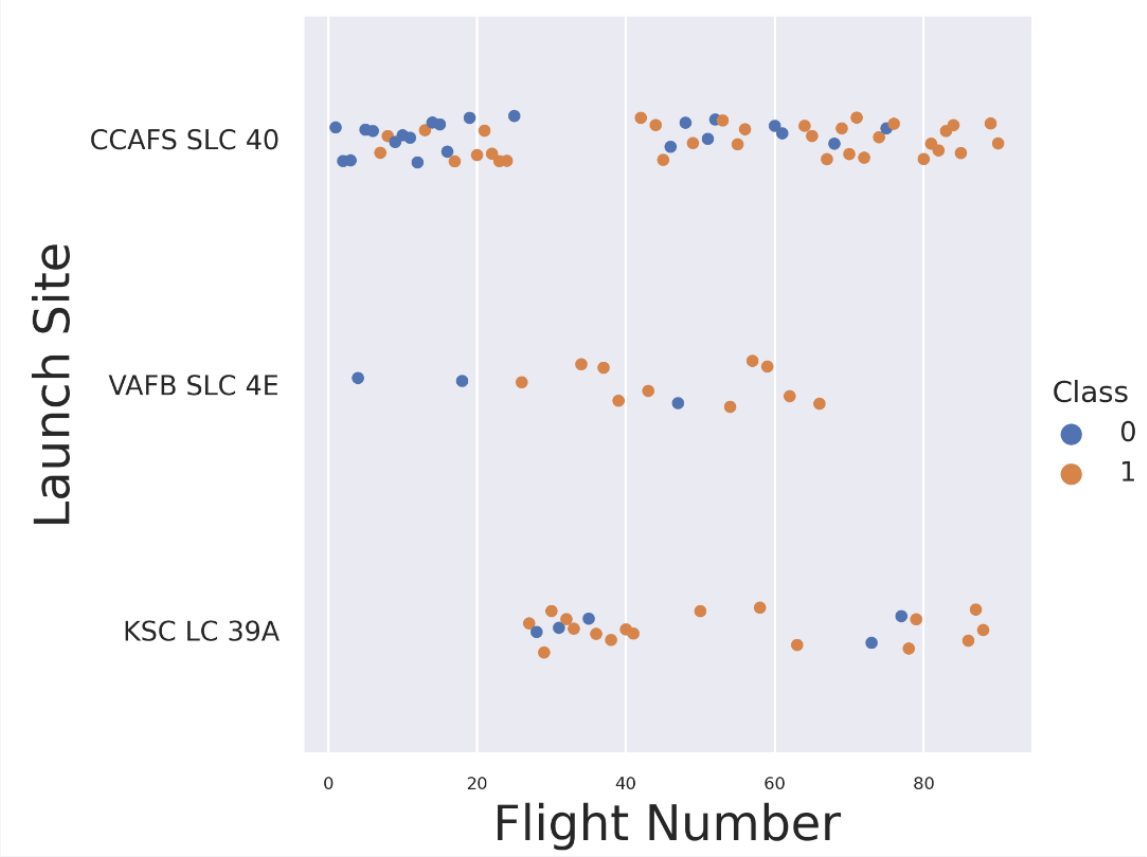
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

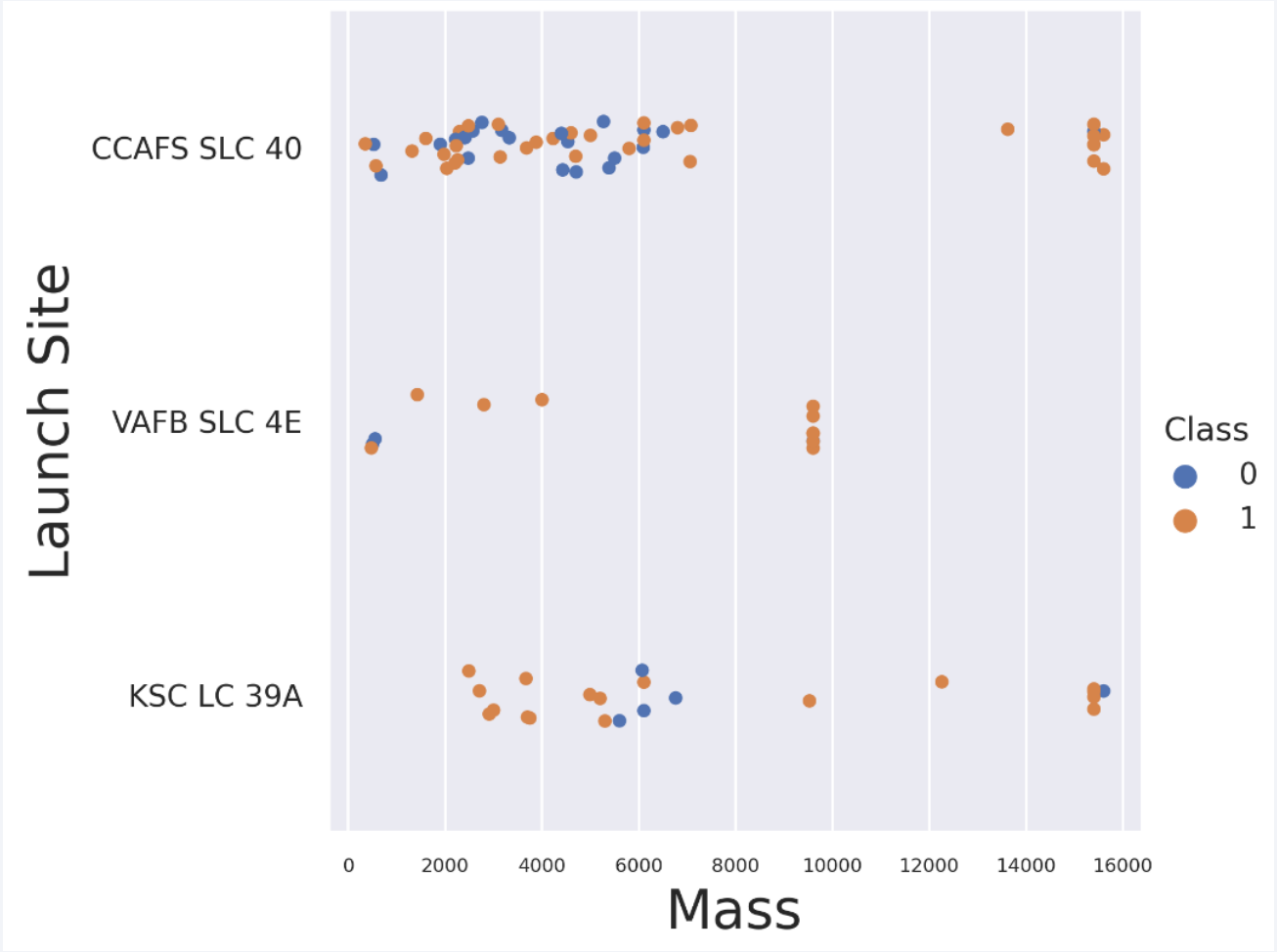
Flight Number vs. Launch Site

- The first 25 flights were conducted at CCAFS SLC40, and they were mostly unsuccessful.
- After this initial run, then next 15 flights were conducted at KSC LC 39A, and they were more successful in general.
- After that, most of the following flights were conducted at CCAFS SLC40 again, most of which were successful, although some still took place at KSC LC 39A
- VAFB SLC 4E was used only sporadically, probably when other launch sites were unavailable.



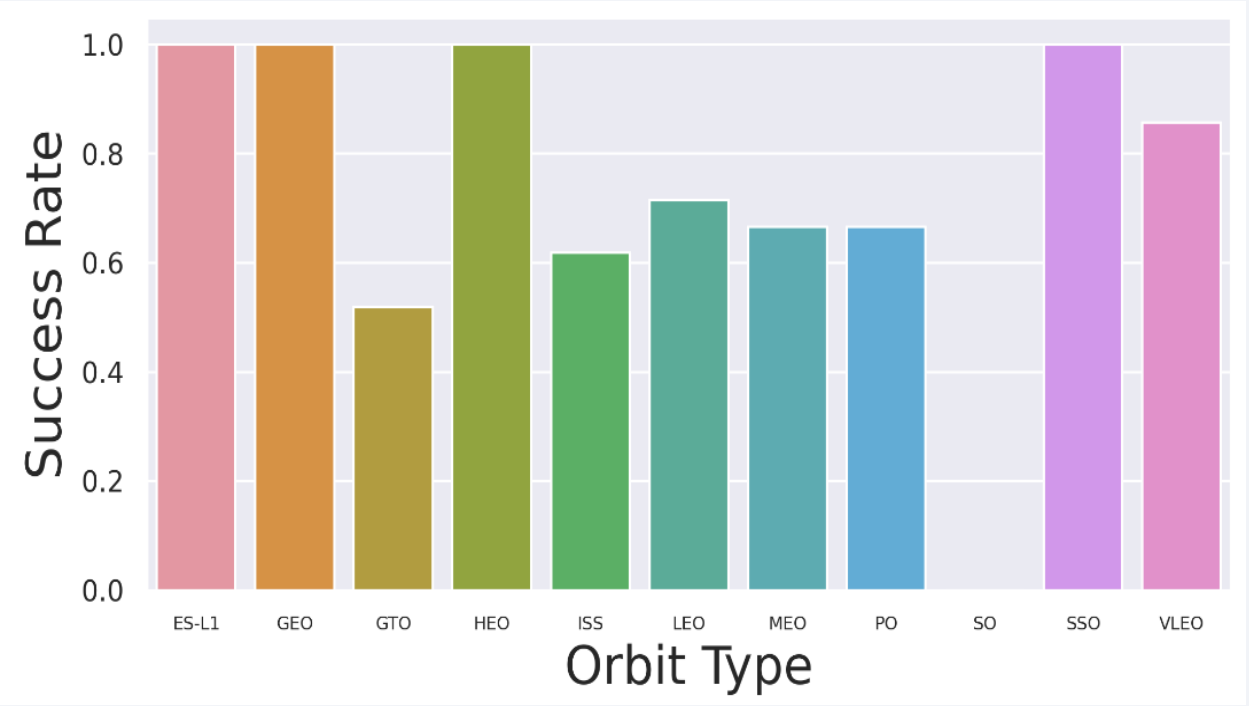
Payload vs. Launch Site

- CCAFS SLC 40 is used for both light (up to 8t) and heavy payloads (+13t)
- VAFB SLC 4E is used for light and medium, but not heavy payloads
- KSC LC 39A used for all payload types



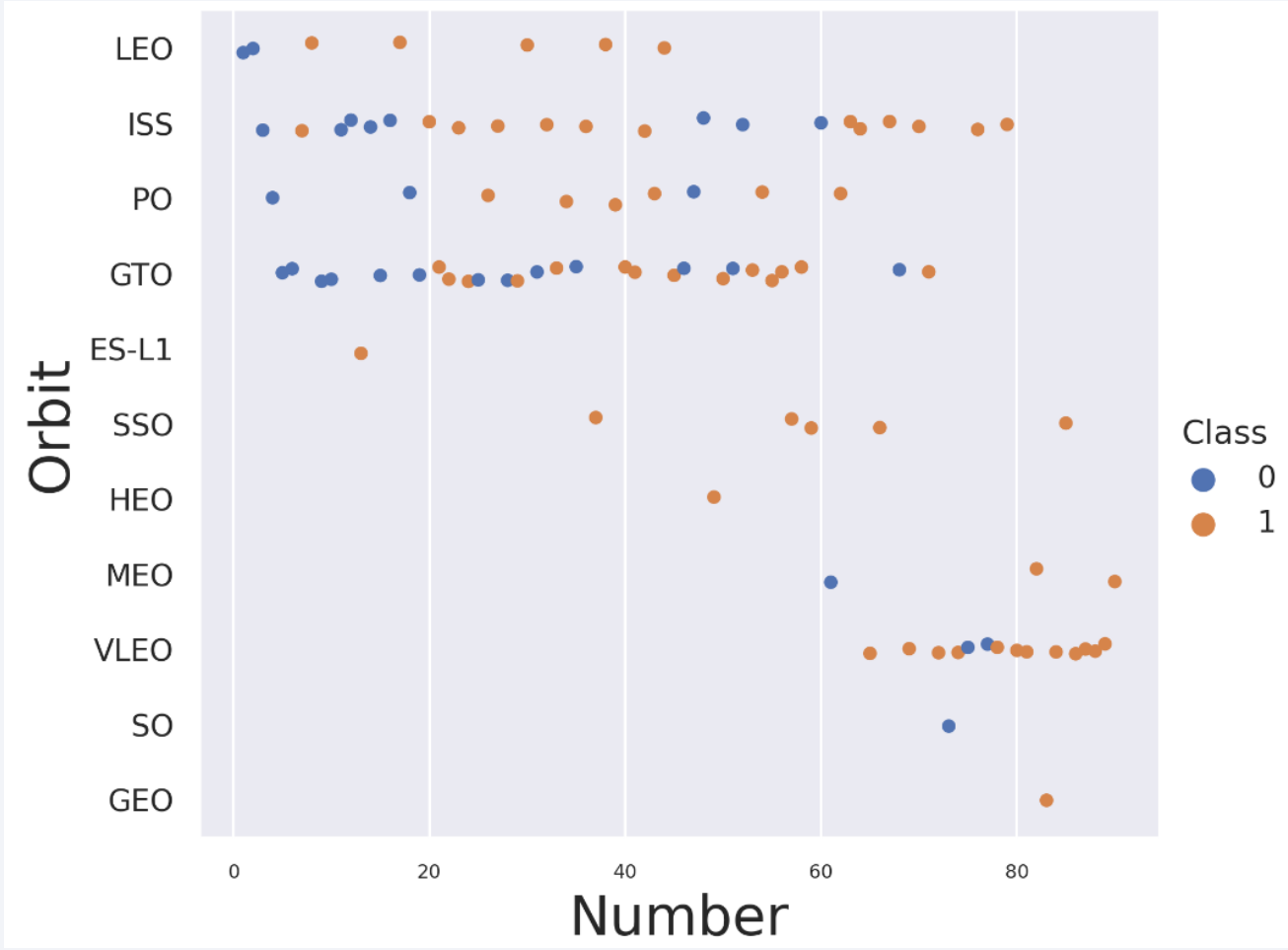
Success Rate vs. Orbit Type

- The highest success rate (100%) are for orbit types:
  - ES-L1, GEO, HEO, SSO
- High success rate (80%):
  - VLEO
- Medium success rates (50%-70%):
  - GTO, ISS, LEO, MEO, PO
- 0% Success Rate:
  - SO



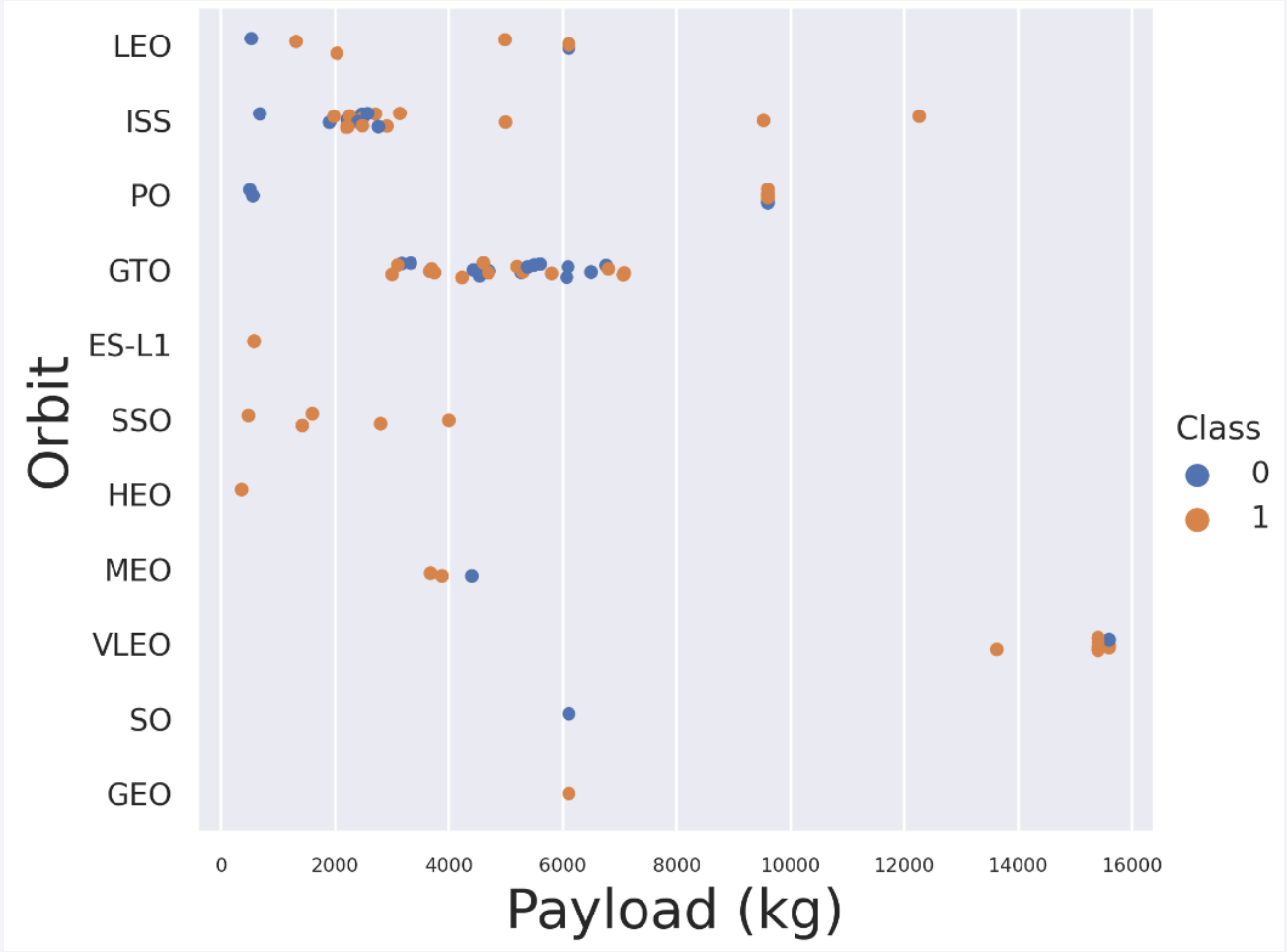
Flight Number vs. Orbit Type

- If the data in this chart is cross-referenced to the previous chart (success rate per orbit type), we can see that most orbit types with 100% success rate have only been targeted once, therefore their success rate is not statistically significant (with the exception of SSO)
- Seems like initially SpaceX was mostly focused on LEO, ISS, PO, and GTO orbits, but now the most popular orbit is VLEO, which has a very high success rate
- The ISS orbit initially had an interval of successful runs from flight number 20 – 40. The first chart (launch site vs flight number) shows that for these flights KSC LC39 was used



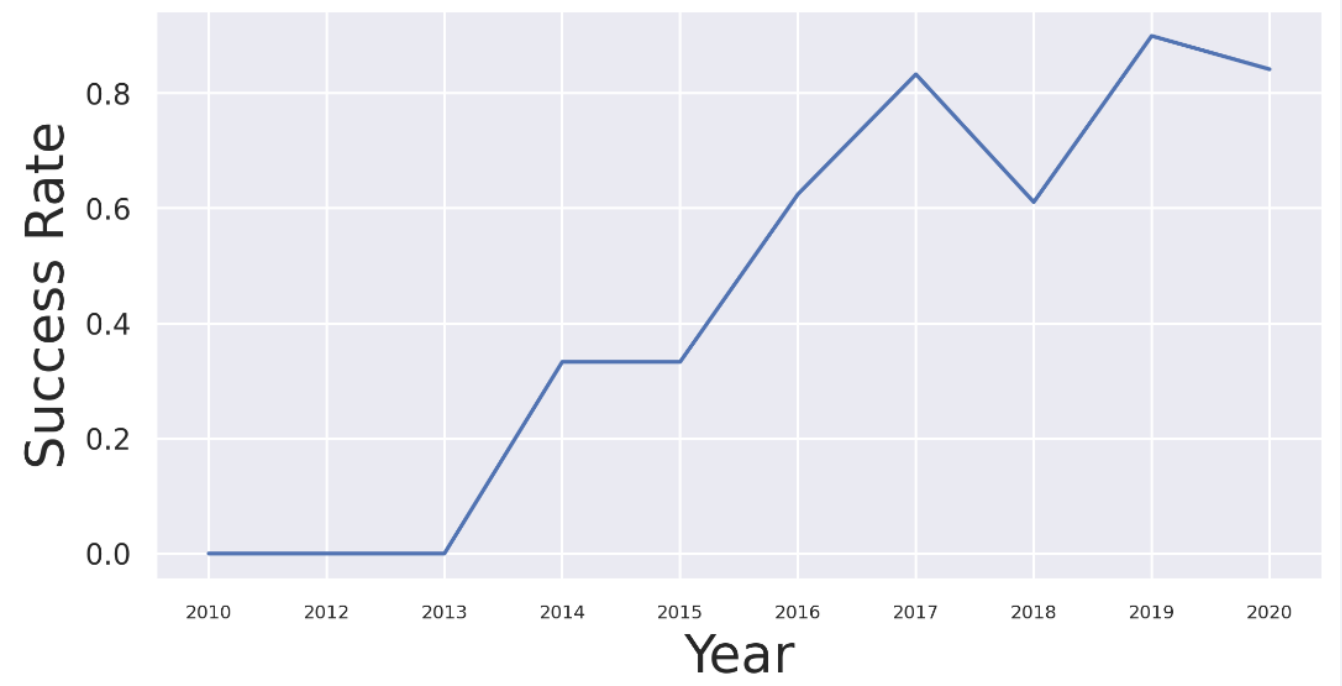
Payload vs. Orbit Type

- All heavy payloads (+13t) were targeted only to VLEO orbit
- Medium payloads (8-13t) were targeted to ISS and PO orbits only
- All other orbits were used for light payloads only (up to 8t)
- HEO and ES-L1 orbits were used only for very light payloads (less than 1t)



Launch Success Yearly Trend

- The graph shows that the average flight success rate has been steadily increasing every year since 2013
- Previous graphs did not give a clear indication of why this might be, so other methods, such as machine learning are necessary to estimate the impact of each variable on the success rate



All Launch Site Names

- The SQL query returned all distinct values in the launch_site column of our database
- This identified three unique launch sites in our data

```
%%sql
SELECT DISTINCT launch_site from spacex
```

```
 * ibm_db_sa://pgz47648:***@55fbc997-9266
Done.
```

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

Launch Site Names Begin with 'CCA'

- The query returned 5 records from our database (LIMIT 5) where launch site started with CCA by using LIKE ‘CCA%’ (where % is a wildcard)

```
%%sql
SELECT * from spacex WHERE launch_site LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://pgz47648:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Total Payload Mass

- Used function SUM() to return total payload mass where client was NASA
- SUM() was limited to NASA payloads by using WHERE customer = ‘NASA (CRS)’

```
%%sql
SELECT SUM(payload_mass__kg_) as total_mass
from spacex WHERE customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://pgz47648:***@55fbc997-9266-43
Done.
```

total_mass
22007

Average Payload Mass by F9 v1.1

- Calculated the average payload mass carried by booster version F9 v1.1
  - Used function AVG()
  - Limited to a specific booster version by using WHERE booster_version = ‘F9 v1.1’

```
%sql SELECT AVG(payload_mass__kg_) as avg_payload
from spacex WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://pgz47648:***@55fbc997-9266-4331-afd3
Done.
```

avg_payload
3676

First Successful Ground Landing Date

- Found the date of the first successful landing outcome on ground pad:
  - Used MIN() function on date column
  - Limited landing outcome to ground pad by using WHERE

```
%sql SELECT MIN(date) as first_ground_success
from spacex where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://pgz47648:***@55fbc997-9266-4331-afd3-888b05e
Done.
```

first_ground_success
2017-01-05

Successful Drone Ship Landing with Payload between 4000 and 6000

- Found the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - Used DISTINCT to skip duplicates
  - Limited query to a specific landing outcome by using WHERE
  - Specified a certain mass range by using WHERE and a set of conditions

```
%%sql
SELECT DISTINCT booster_version from spacex
WHERE landing__outcome = 'Success (drone ship)'
and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000
```

```
 * ibm_db_sa://pgz47648:***@55fbc997-9266-4331-afd3-888b0
Done.
```

booster_version
F9 FT B1031.2
F9 FT B1022

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes within the database:
  - Created a nested query for both successful and failed missions where used COUNT() and wildcard with LIKE to filter out relevant entries

```
%%sql
SELECT (SELECT COUNT(landing__outcome) as success
from spacex WHERE landing__outcome like 'Success%'),
(SELECT COUNT(landing__outcome) as fail from spacex
WHERE landing__outcome like 'Fail%') from spacex limit 1
```

```
 * ibm_db_sa://pgz47648:***@55fbc997-9266-4331-afd3-888b0
Done.
```

success	fail
27	5

Boosters Carried Maximum Payload

- Listed the names of the booster which have carried the maximum payload mass:
  - Remove duplicate entries by DISTINCT
  - Created a nested query to allow the usage of a MAX() function within a condition

```
%%sql
SELECT DISTINCT booster_version from spacex
where payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) from spacex)
```

```
 * ibm_db_sa://pgz47648:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2i
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1058.3
F9 B5 B1060.2

2015 Launch Records

- Listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:
  - Selected specific columns from all entries with specific conditions for landing outcome and date with a wildcard

```
%%sql
SELECT date, landing__outcome, booster_version, launch_site from spacex
where landing__outcome = 'Failure (drone ship)' and date like '%2015%'
```

```
 * ibm_db_sa://pgz47648:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io9
Done.
```

DATE	landing__outcome	booster_version	launch_site
2015-10-01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40

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

- Ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:
  - Used COUNT() to get total number of each landing outcome
  - Restricted query to a specific date range with WHERE
  - Grouped results by landing outcome using GROUP BY
  - Ordered by using ORDER BY DESC

```
%%sql
SELECT landing__outcome, COUNT(landing__outcome) as count from spacex
WHERE date > '2010-06-04' and date < '2017-03-20'
GROUP BY landing__outcome
ORDER BY count DESC
```

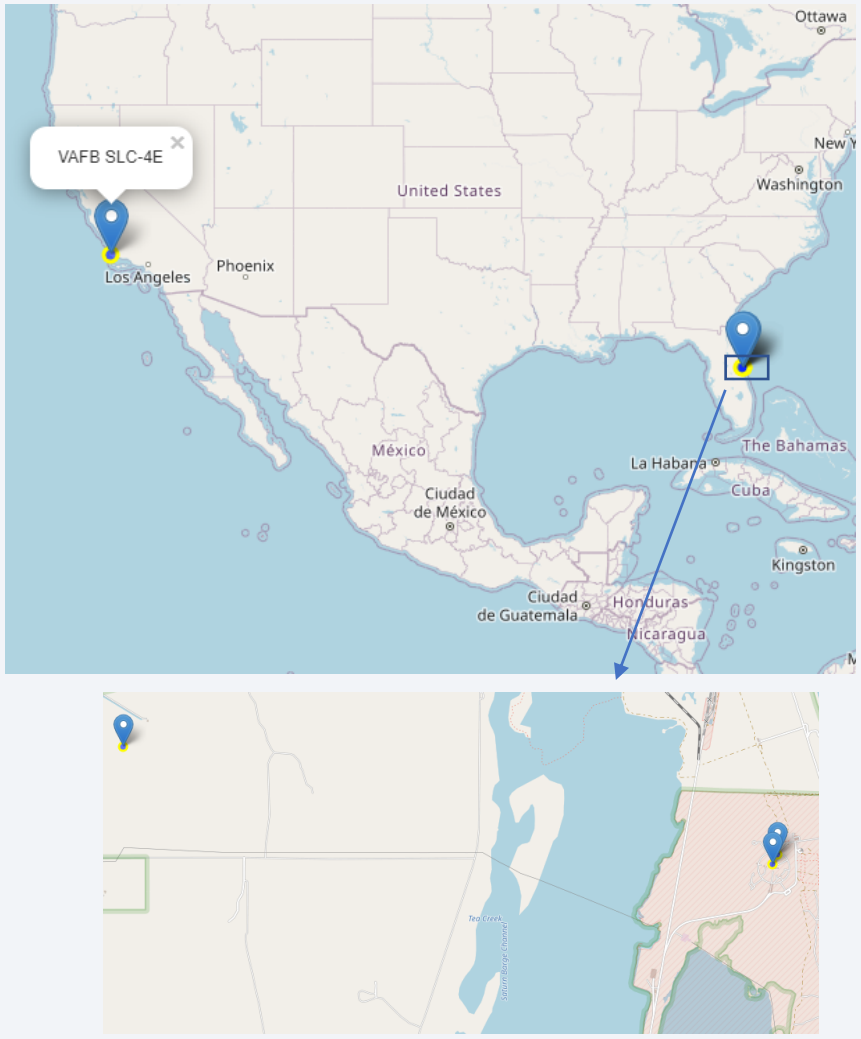
```
 * ibm_db_sa://pgz47648:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io
Done.
```

landing__outcome	COUNT
No attempt	7
Failure (drone ship)	2
Success (drone ship)	2
Success (ground pad)	2
Controlled (ocean)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

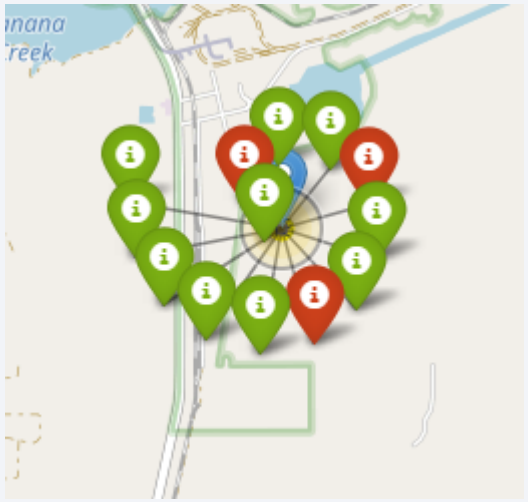
Falcon 9 Launch Sites



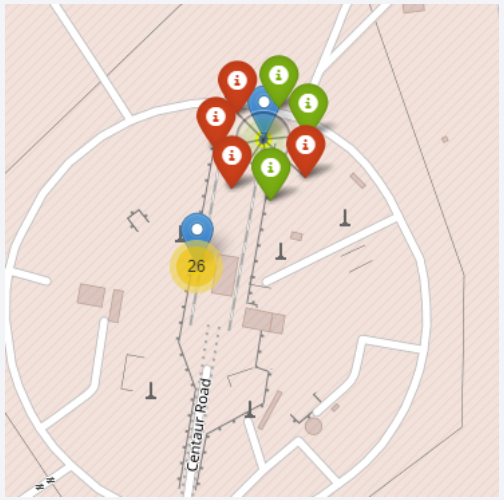
- There are 4 launch sites used for Falcon 9 flights
  - VAFB SLC4-E located on the west coast, California
  - The other three (KSC LC-39A, CCAFS LC-40 and CCAFS SLC-40) are all located very close to one another, on the east coast, Florida, at the Cape Canaveral Space Force Station
- The VAFB SLC4-E site is located further north than the other sites

Launch Outcomes at Each Site

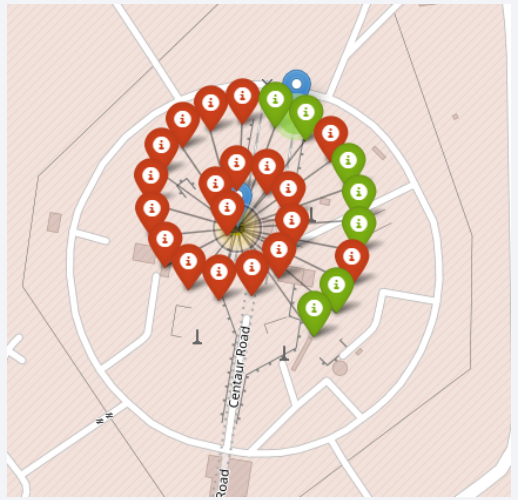
- Most launches were conducted at CCAFS LC-40, about a third of which were unsuccessful
- The CCAFS SLC-40, which is located right next to it, had the least launches , but about a half of them were successful (probably less statistically significant)
- KSC LC-39A, which is slightly to the east and further away from the coastline had a very high success rate



KSC LC-39A



CCAFS SLC-40

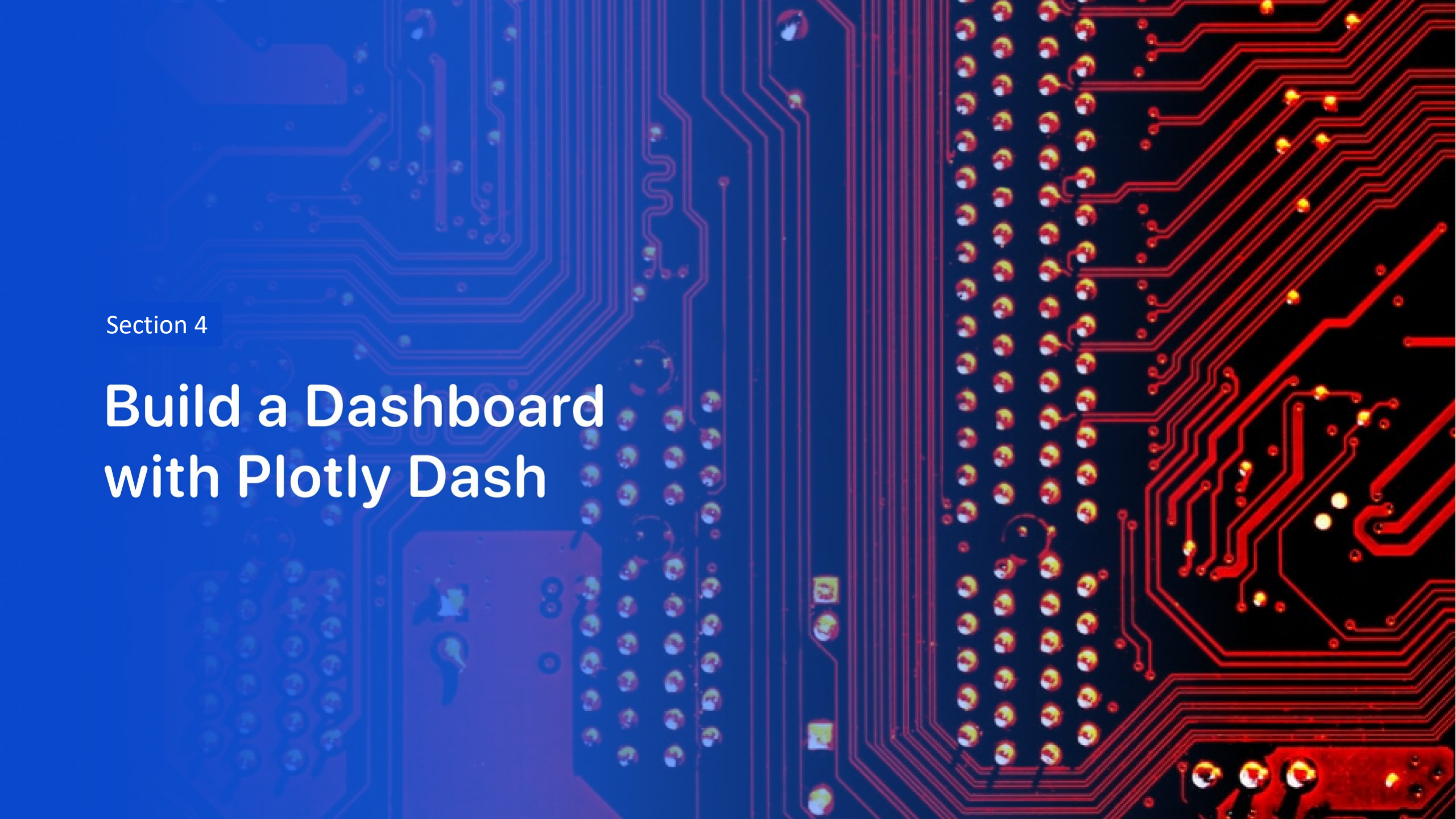


CCAFS LC-40

VAFB SLC4-E

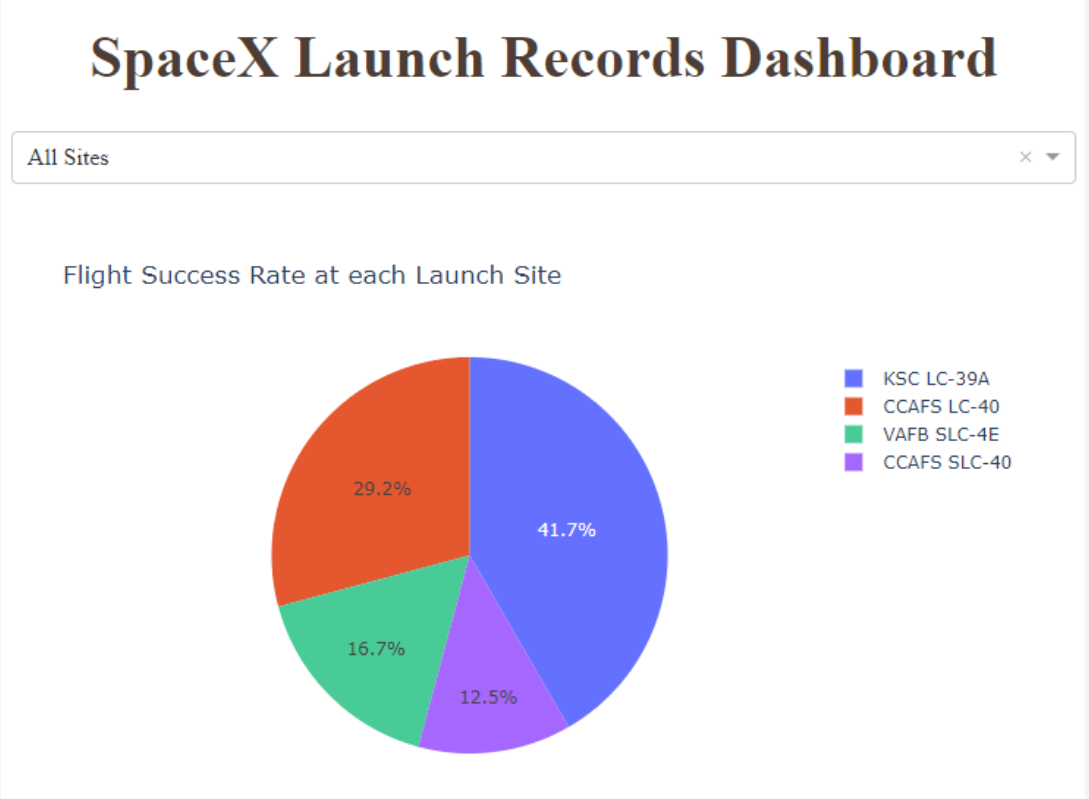
Launch Site Distance from the Coastline

- The CCAFS launch sites are less than a kilometer away from the coastline
- This might explain their relatively low success rates observed in the previous slide


Section 4
Build a Dashboard with Plotly Dash

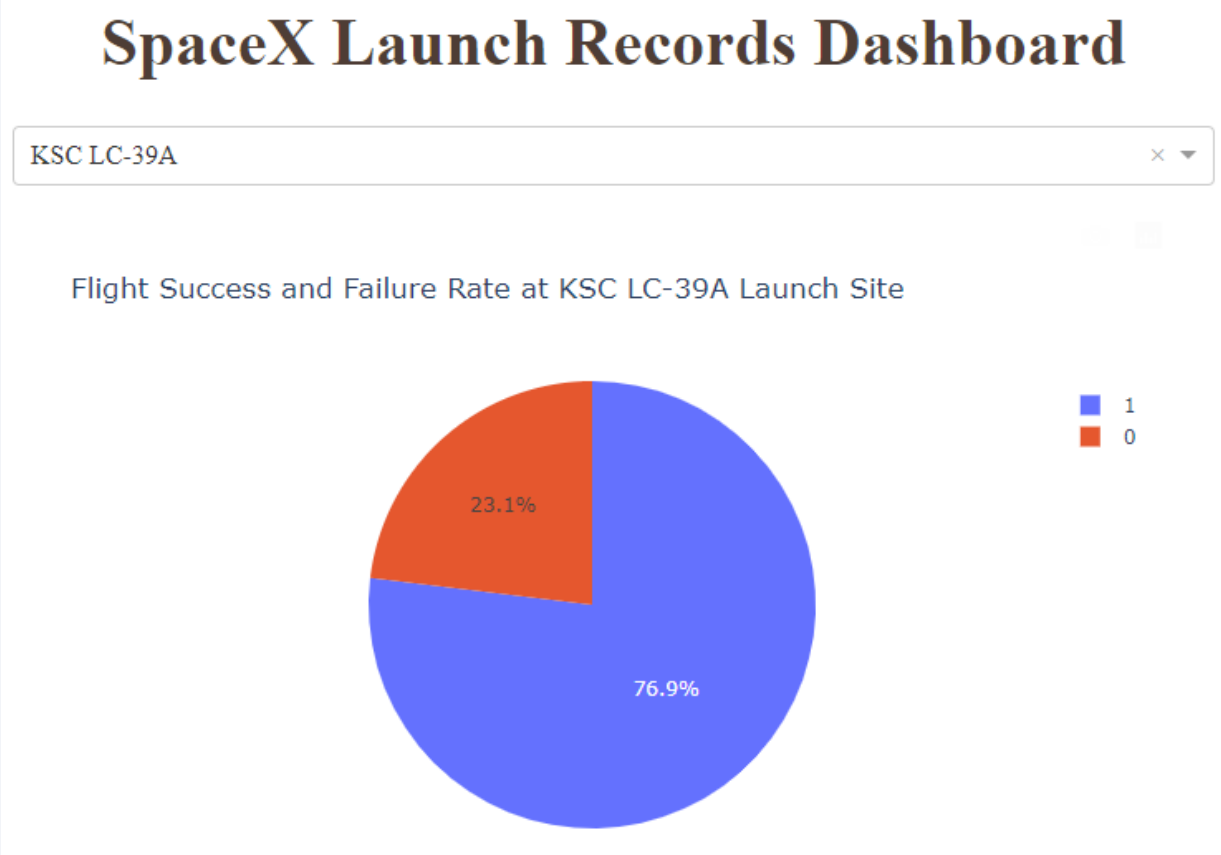
Successful flights at each Launch Site

- Similarly to the data visualised on a map, the pie chart showing successful flights at all launch sites indicates that KSC LC-39A launch site had the highest amount of successful launches (41.7%)
- Lowest amount was for CCAFS SLC-40 (12.5% of all successful flights)
- The CCAFS LC-40, which was used for most of the launches, had the 2nd highest success rate (29.2%)



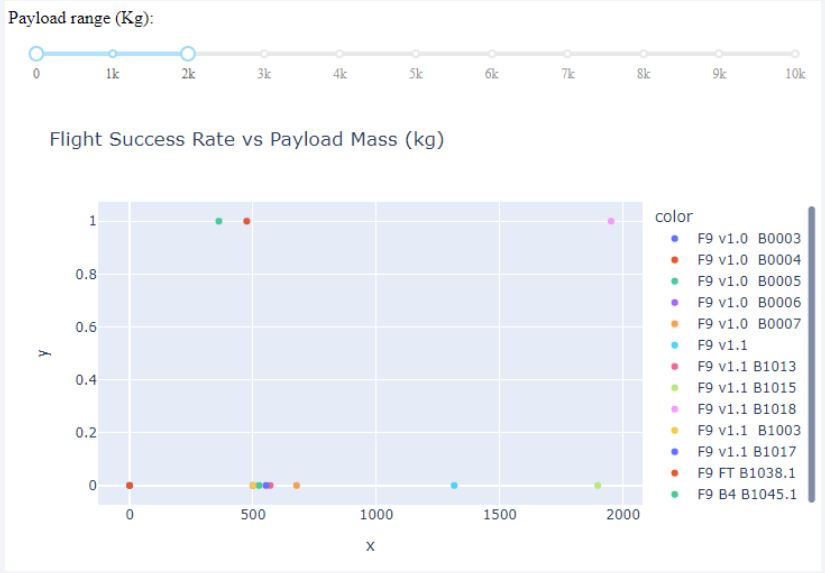
Successful flights at KSC LC-39A

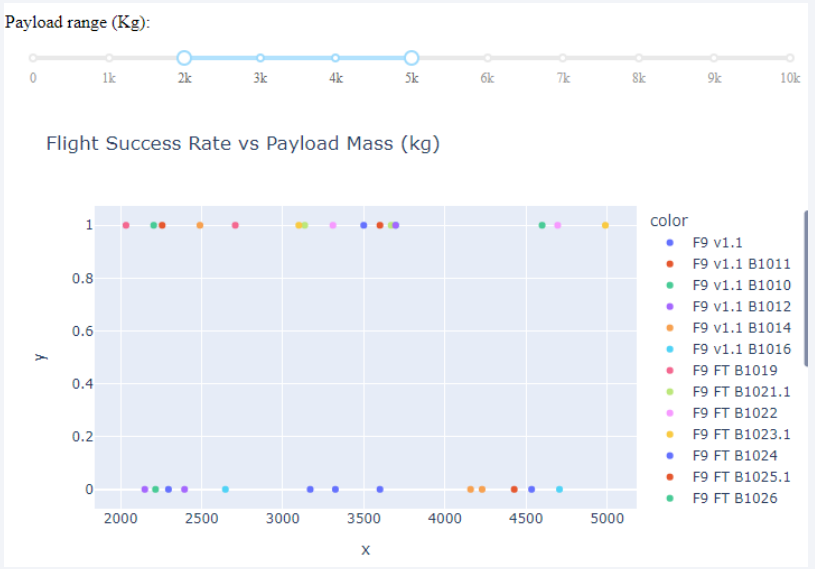
- The most successful launch site, KSC LC-39A, had a flight success rate of 76.9%

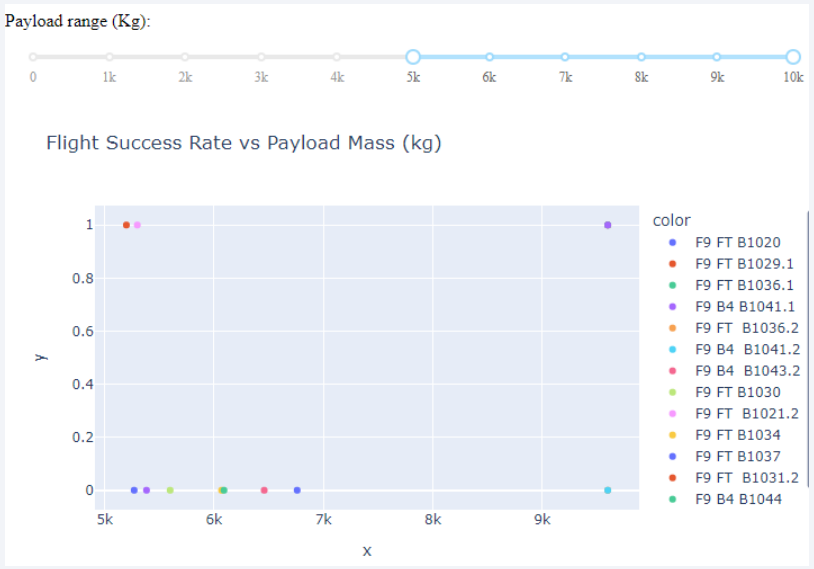


Flight success rate at different payloads

- Selecting only the highest payloads (5-10t) showed that most of the flights were unsuccessful
- Similarly, selecting the lowest payloads (up to 2t) showed that most flights were again unsuccessful
- Most of the successful flights were found in the payload mass range between 2 and 5t





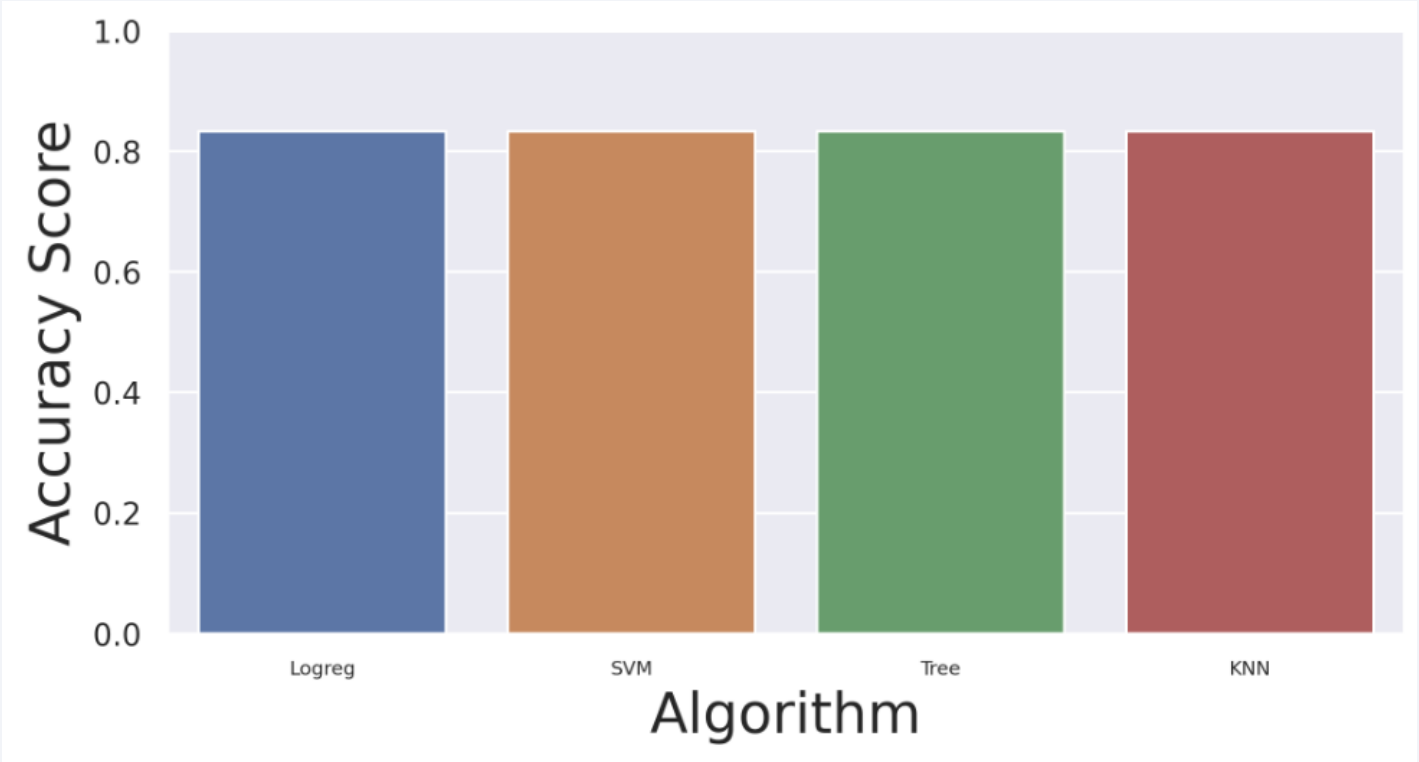


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- After finding the best parameters using GridSearchCV, all algorithms had the same accuracy of 0.8333



Confusion Matrix

- The algorithms all correctly predicted all 12 positive outcomes, with zero false positives
- The algorithms were slightly worse at predicting negative outcomes, with accuracy being only 50%

Conclusions

- If optimised parameters are used for different classification algorithms , they will tend to give similar results
- We only had 18 data points in our testing set, which would be too small to detect any divergences in the classification algorithms, resulting in similar accuracy scores
- Our models seemed to be biased towards predicting positive outcomes, with a high false positive rate, this might reflect the size of our dataset
- 83% success rate is still quite high, so we can use it to make predictions about whether a launch will be successful or not, but more data and more optimisation is required

Thank you!