IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matthew Winder
February 2022

Outline

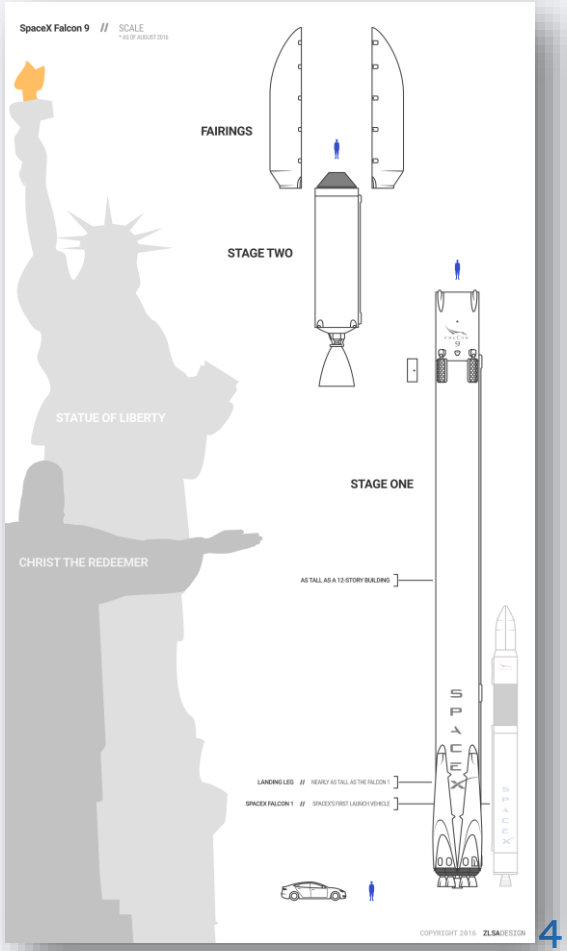
- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- In order to determine the price point of delivering a payload into orbit publicly available SpaceX launch data was used to develop predictive models which identified successful stage 1 booster landings using a variety of parameters including launch site, payload mass, orbit type and booster
- Analysis was carried out using a combination of programmatic queries and interactive maps and dashboards
- The most successful model generated from the data was a Decision Tree Classification model with a predictive accuracy of 88.8889%
- Assuming that the price of launch with a successful landing is $62m, the price without landing is $165m, and that SpaceY will only target launches with a successful landing, there is a mean launch price of $76.11m

Introduction – Scenario

- The commercialization of space travel is becoming viable with several organisations such as SpaceX, Virgin Galactic, Rocket Lab and Blue Origin, now providing payload delivery to customers
- SpaceY, founded by entrepreneur Allon Mask, is looking to join this industry
- Pricing from the industry incumbents ranges from ~$62m to ~ $165m with the costs heavily dependent on the reuse/functional recovery of the stage 1 booster



Introduction – Problem Statement

- SpaceY wishes to find a commercially viable price point to enter the industry
- Assuming that they will be able to produce rockets with approximately the same basic cost as SpaceX they need to determine the cost of a given launch
- The cost of a launch is highly dependent on reuse of stage 1, so our analysis will focus on understanding the launch factors that contribute to a successful landing
- Our prediction of success will be based on models generated and tested on the publicly available SpaceX flight and landing data

Section 1

Methodology

Methodology

Executive Summary

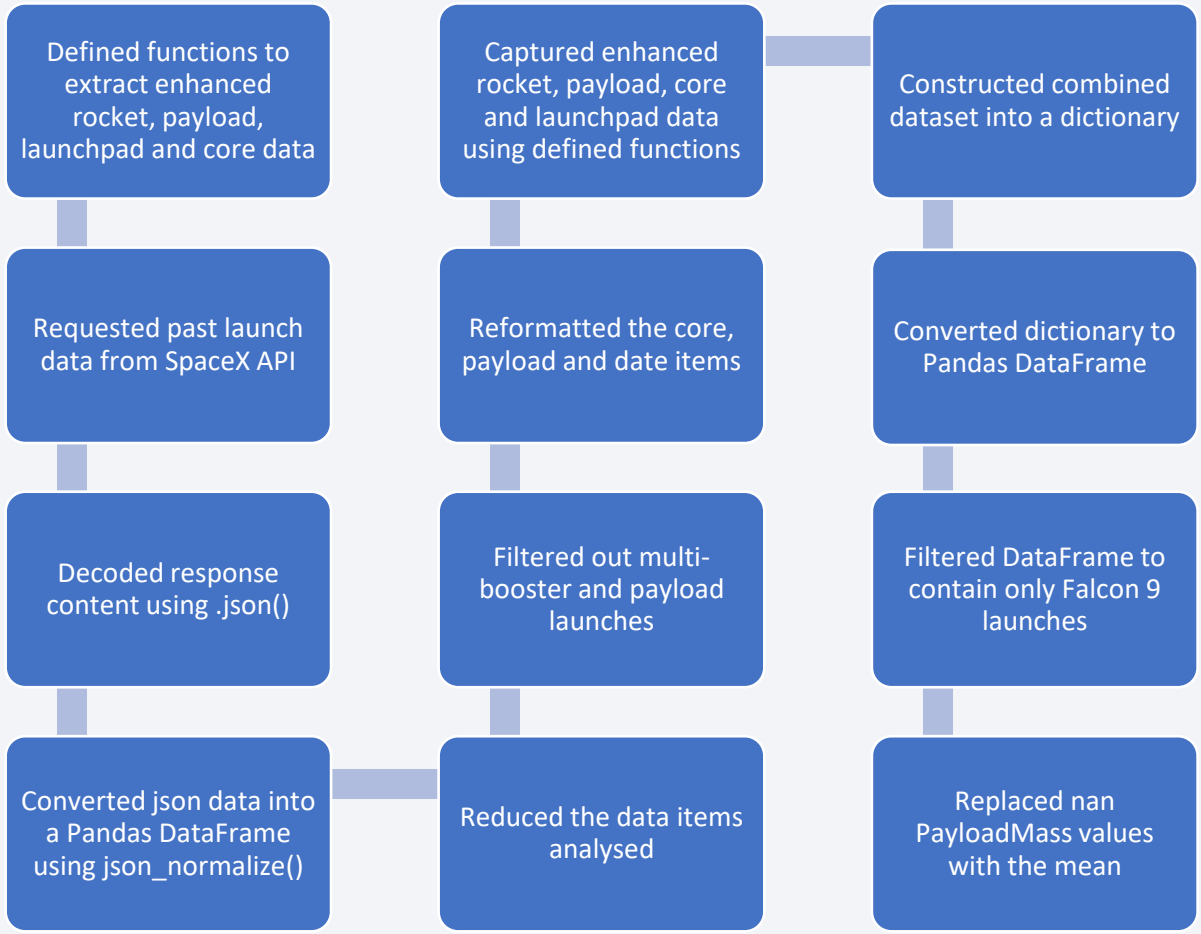
- Data collection methodology:
  - Data for the investigation was collected through the SpaceX API and through web scraping a table from SpaceX's Wikipedia page using the Requests and BeautifulSoup libraries
- Perform data wrangling
  - Data was cleaned, categorical data was parameterized, and a classification item added to enable predictive analysis
- 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 was standardized, split into training and testing sets, tuned using GridSearchCV, and multiple models evaluated for accuracy using score from Sklearn and visualized with confusion matrices

Data Collection

- Data was collected for the investigation from:
  - the SpaceX data API (https://api.spacexdata.com), and
  - the Wikipedia page detailing launches of the Falcon 9 and Falcon Heavy (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- The data items that were considered in the analysis were:
  - **API**: rocket, payload, launchpad, core, flight number, and date
  - **Wikipedia**: flight number, launch site, payload, payload mass, orbit, customer, launch outcome, version booster, booster landing, data and time

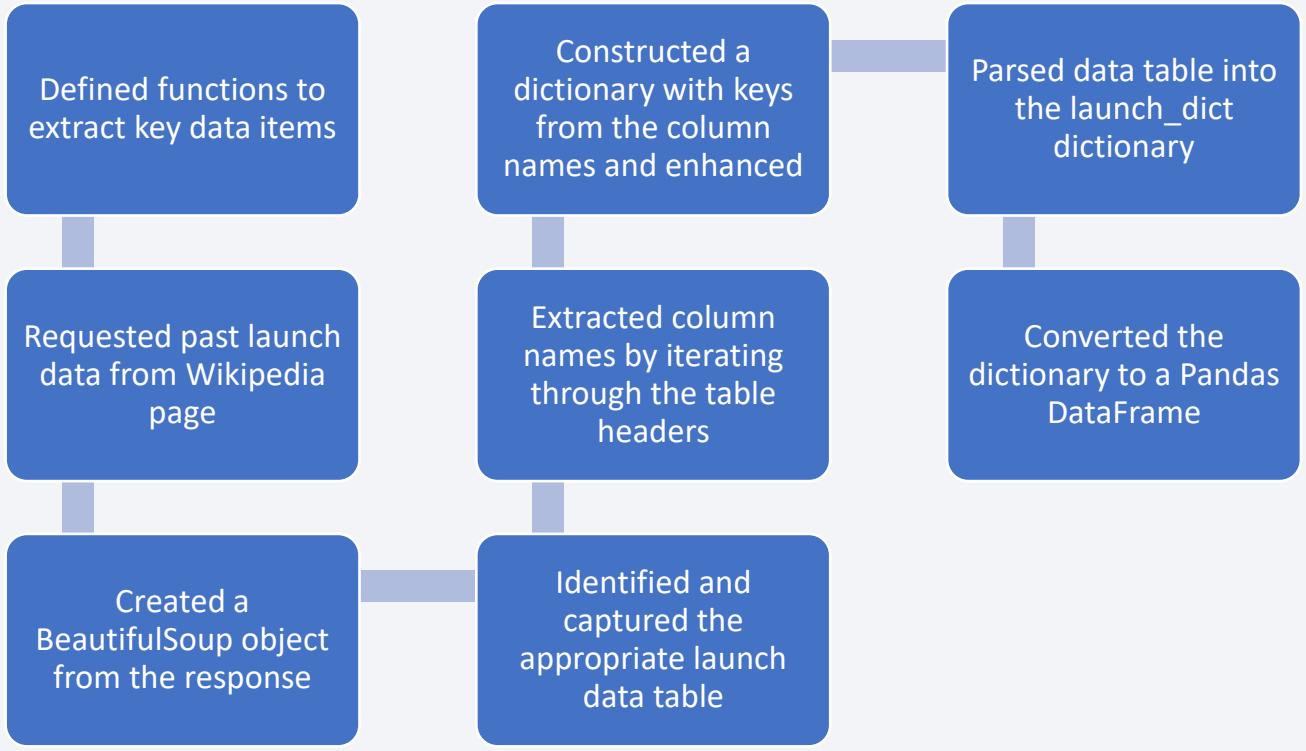
Data Collection – SpaceX API

- Base SpaceX API: https://api.spacexdata.com
- Enhancing data API calls:
  - https://api.spacexdata.com/v4/rockets/
  - https://api.spacexdata.com/v4/launchpads/
  - https://api.spacexdata.com/v4/payloads/
  - https://api.spacexdata.com/v4/cores/
- **API data items:** rocket, payload, launchpad, core, flight number, and date
- Completed SpaceX API calls notebook - GitHub



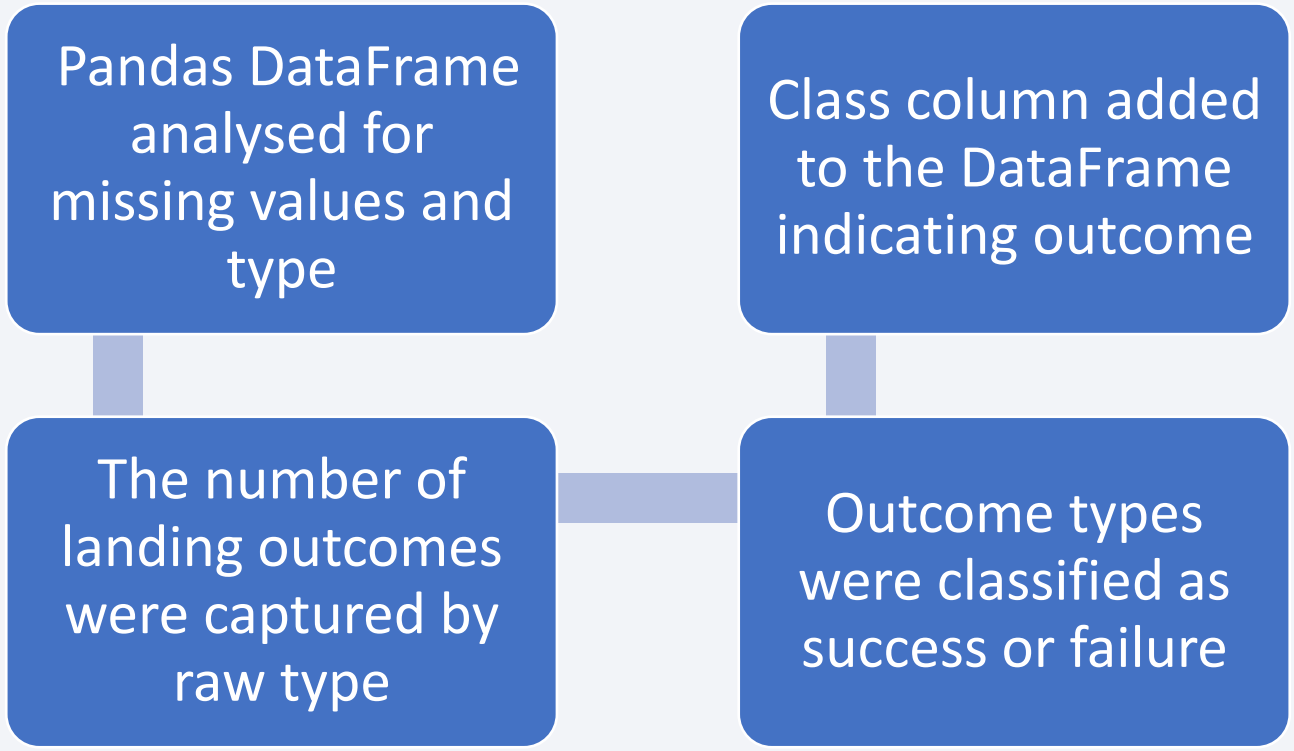
Data Collection - Scraping

- Wikipedia page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- **Web scraping data items:** flight number, launch site, payload, payload mass, orbit, customer, launch outcome, version booster, booster landing, data and time
- Completed data collection with web scraping notebook - GitHub



Data Wrangling

- A training classification (Class) of success as a single value was required
- Where a mission outcome is “True” Class is set to 1
- Mission Outcome mapping:
  - True ASDS, True RTLS, & True Ocean = 1
  - None None, False ASDS, None ASDS, False Ocean, False RTLS = 0
- Completed data wrangling notebook - GitHub



EDA with Data Visualization

- Exploratory Data Analysis was carried out to attempt to determine whether predictions could be made about the relationship with the following variables to a successful mission:
  - Payload mass, flight number (synonymous with time), launch site, and orbit type
- Categorical data was then converted using One Hot Encoding to generate a feature set for further analysis
- Scatter charts of flight number vs payload mass, flight number vs launch site, and payload mass vs launch site, flight number vs orbit type and payload vs orbit type were plotted to determine correlation between factors
- A bar chart of orbit type success and a line graph of the yearly success trend were also plotted as independent signals
- Completed EDA with data visualization notebook - GitHub

EDA with SQL

- Exploratory Data Analysis was carried out through the following SQL queries on the data set:
  - Distinct sites used during the period
  - Total payload mass carried for NASA
  - Average payload mass carried by the F9 v1.1 booster
  - First successful ground pad mission was launched
  - Successful boosters landed on a drone ship with a payload mass between 4,000 and 6,000kgs
  - Total number of successfully completed and failed missions
  - Booster versions which have carried the maximum payload mass
  - Booster versions with failed drone ship landings in 2015
  - Ranked the landing outcome frequency between 2010-06-04 and 2017-03-20
- Completed EDA with SQL notebook - GitHub

Build an Interactive Map with Folium

- The following objects were added to an interactive Folium map to aid analysis of location as a factor to success and required considerations of environment:
  - Circles and Markers to indicate the location of the launch sites
  - A line showing the location of the equator relative to the sites
  - Launch outcomes (success or failure) as a cluster to each launch site
  - Lines and Markers showing the launch site distance to the following items:
    - Coastline – Useful for parts transport and safety in case of launch/landing failure
    - Railway and main road – Useful for parts and ground crew transport
    - Nearest city – To ensure that debris does not harm civilians
- Completed interactive Folium map notebook - GitHub

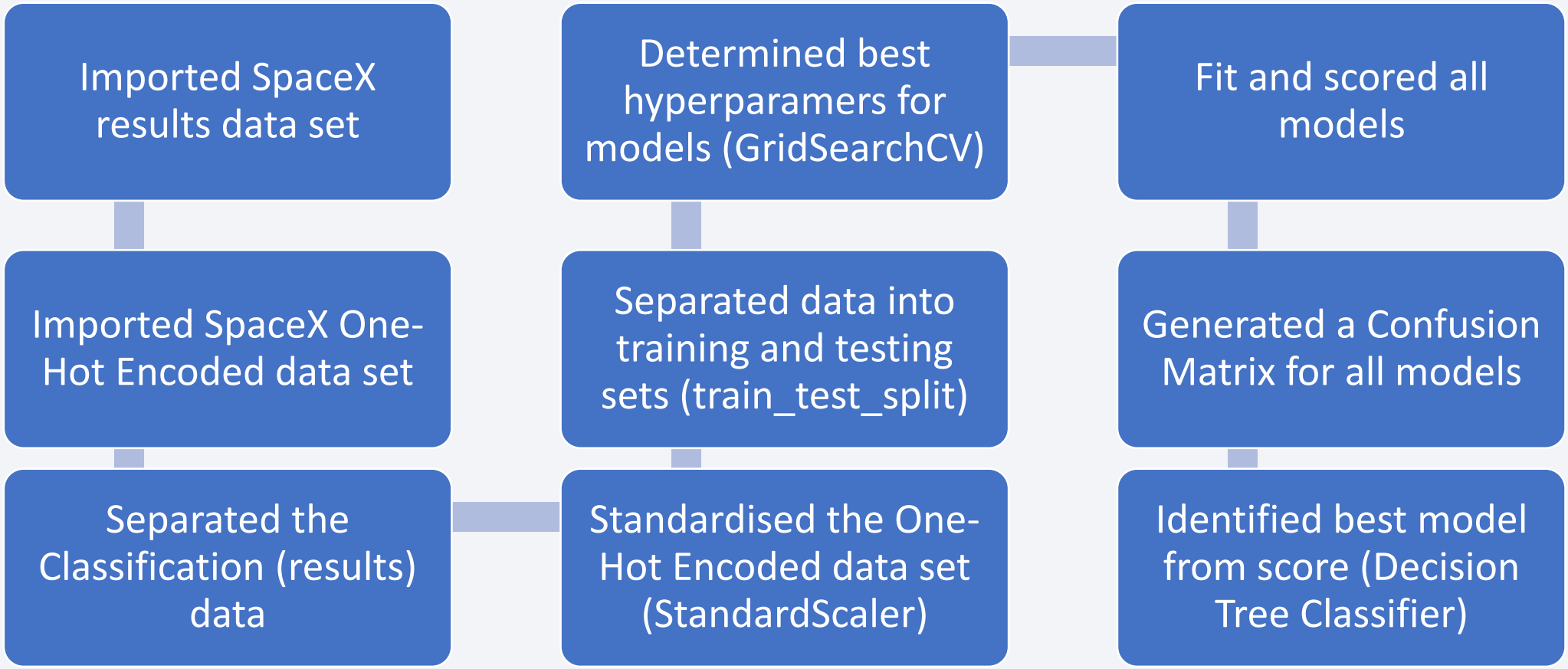
Build a Dashboard with Plotly Dash

- A dashboard was created with the following charts:
  - a pie chart to measure the launch success rates per site, and
  - a scatter plot displaying the correlation between success and payload mass
  - both graphics were filterable to display at a more granular level
- The pie chart was added to provide a general indication of success rates, with the scatter plot added to meet the need for more detailed analysis across payload levels
- Completed Plotly Dash lab notebook – GitHub

Predictive Analysis (Classification)

- 4 models were created from the SpaceX dataset seeking to determine the best predictor of landing success (Classification)
- Data was imported, standardised and split into training and test data sets, and then tested with multiple models, each of which were refit for the best performing hyperparamenter
- The methods assessed were Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbours
- Rerunning the notebook multiple times often returns an equal level of accuracy for all methods, however, where there is a different the method with the best performance is Decision Tree Classifier

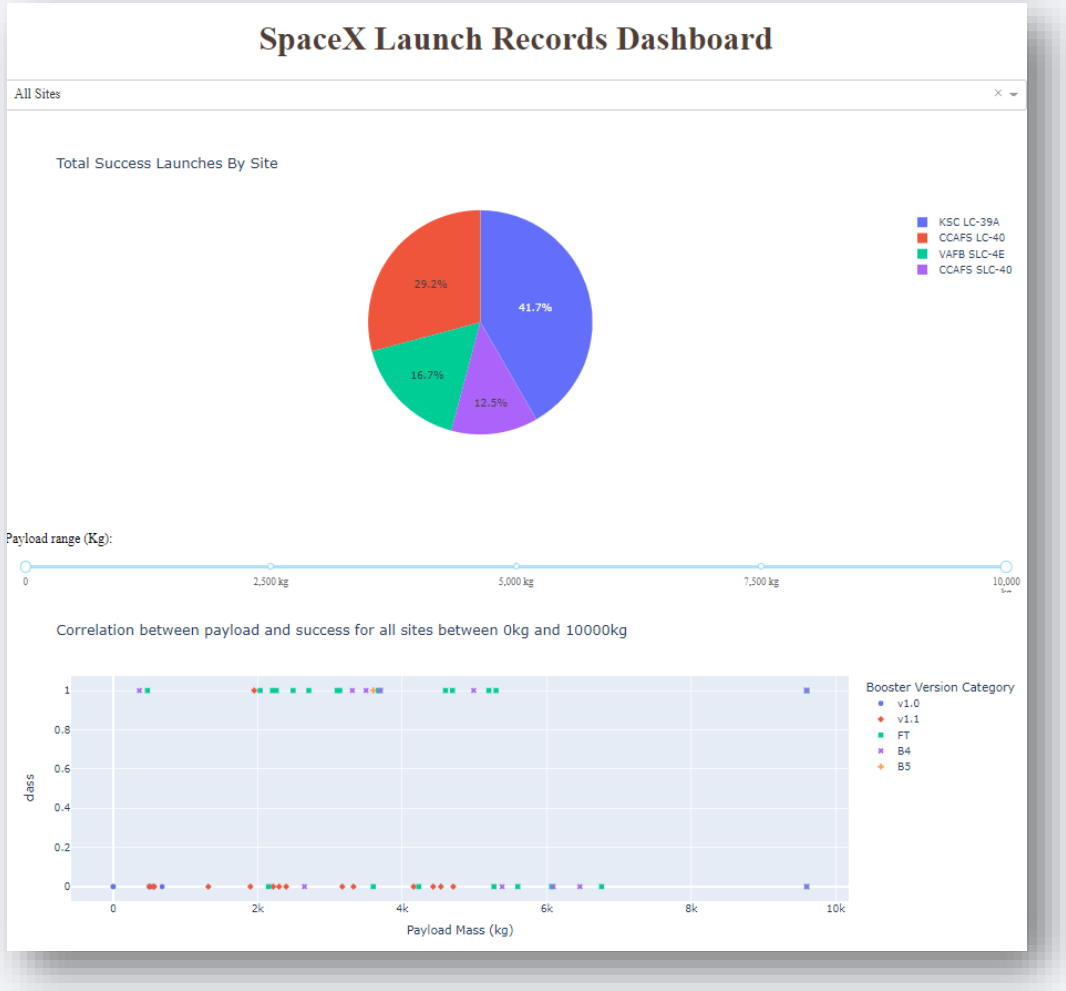
Predictive Analysis (Flow Chart)



- Completed predictive analysis lab - GitHub

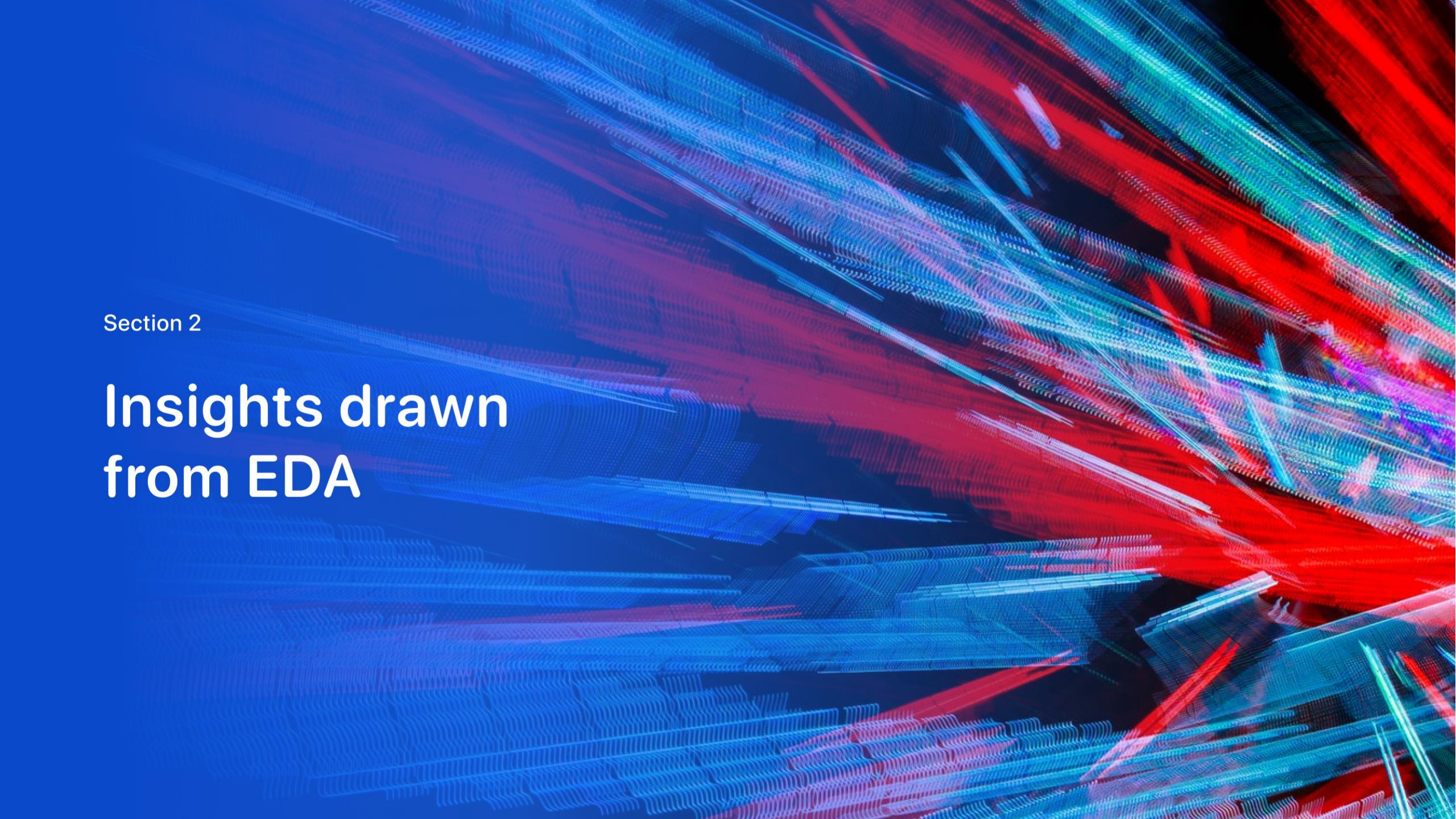
Results

- The slides in the following sections outline:
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

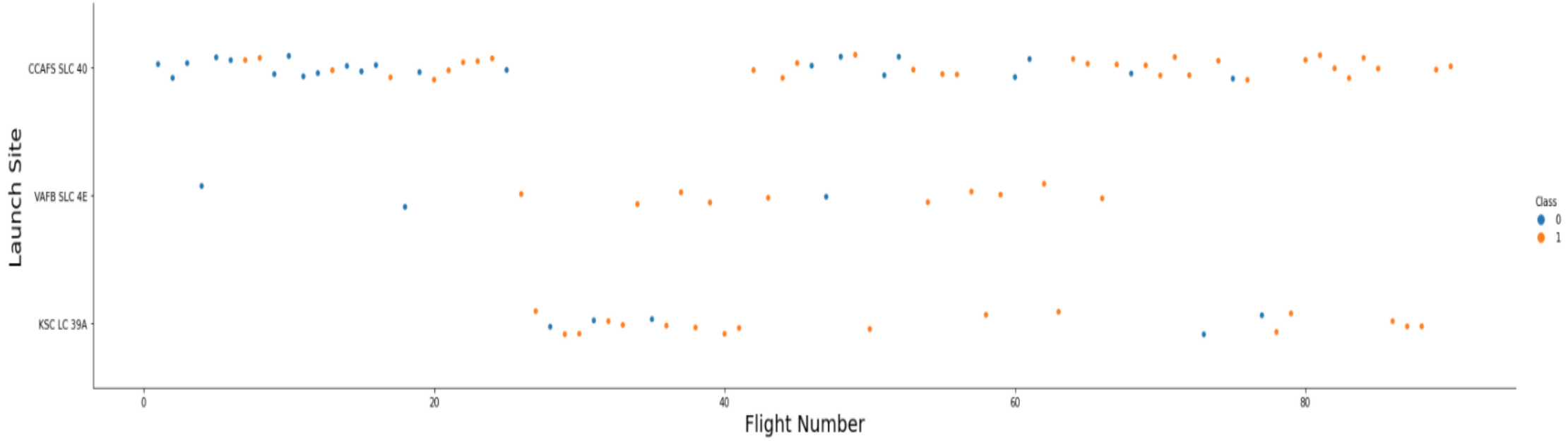


Section 2

Insights drawn from EDA



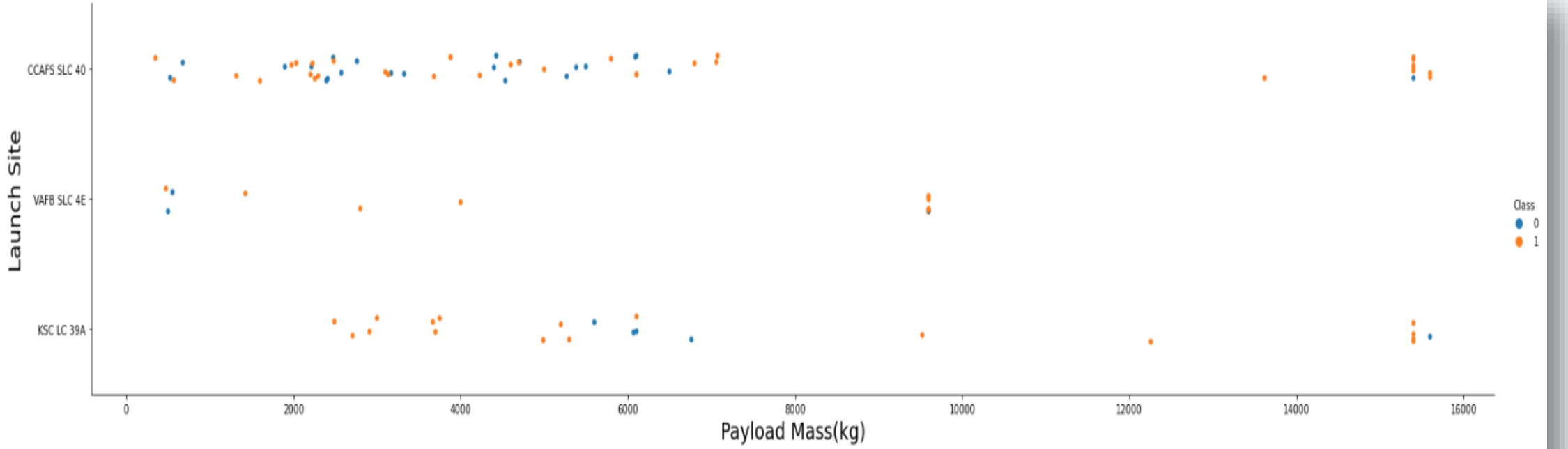
Flight Number vs. Launch Site



As the flight number increases we see a general improvement in the success rate of the launches.

While the overall rate of success of site CCAFS SLC 40 is the lowest of any of the sites, it should also be noted that the majority of the early tests were carried out here and learnings from those initial launches would hopefully have been applied to the launches from the other sites.

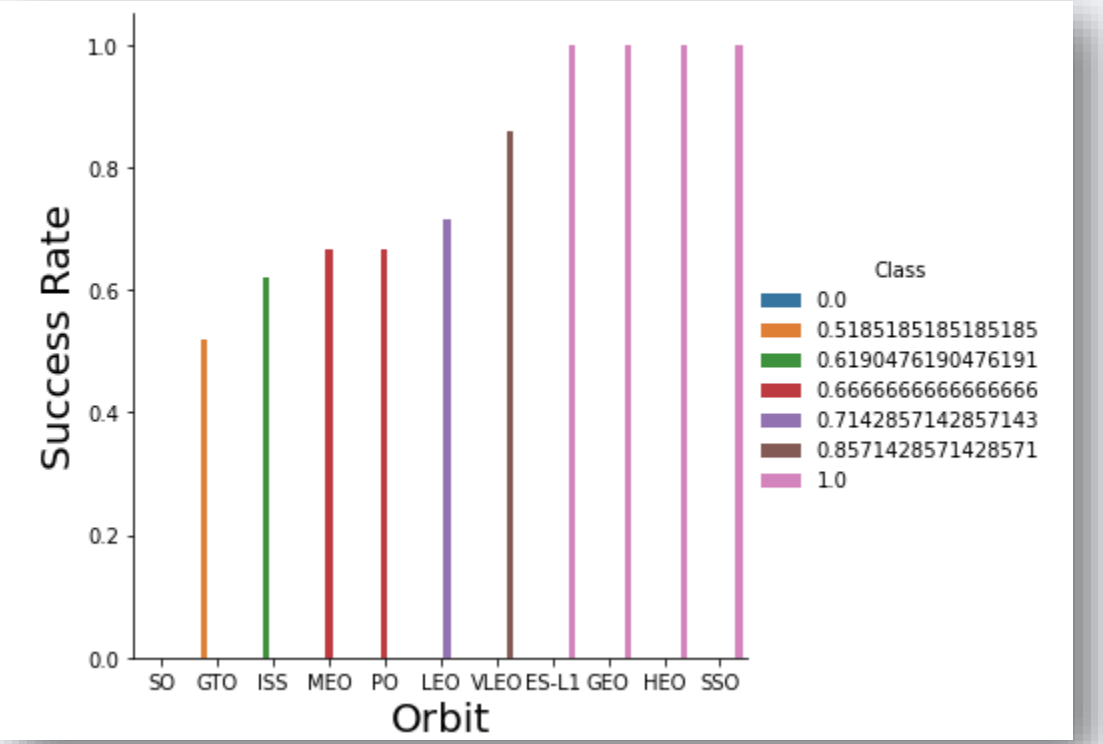
Payload vs. Launch Site



Now if you observe the Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for the heavy payload mass (greater than 10000).

There appears to be a relationship between the increasing payload mass and chance of a successful launch, with the breakthrough arriving at around the 7000kg mark.

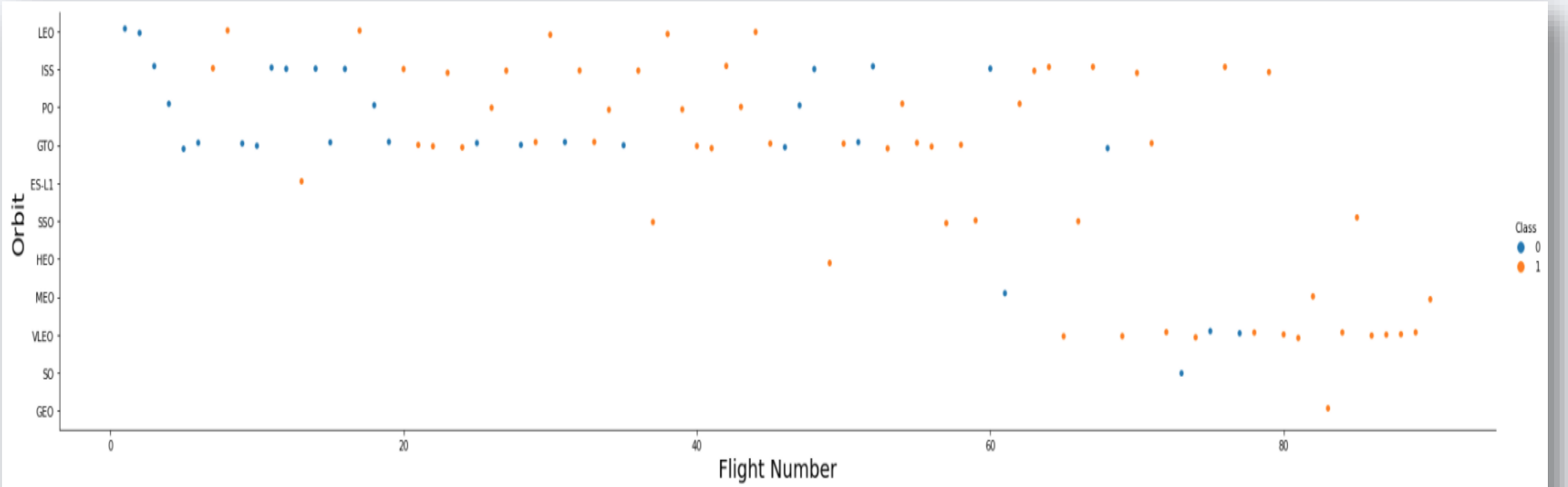
Success Rate vs. Orbit Type



The orbits with the highest success rates are:

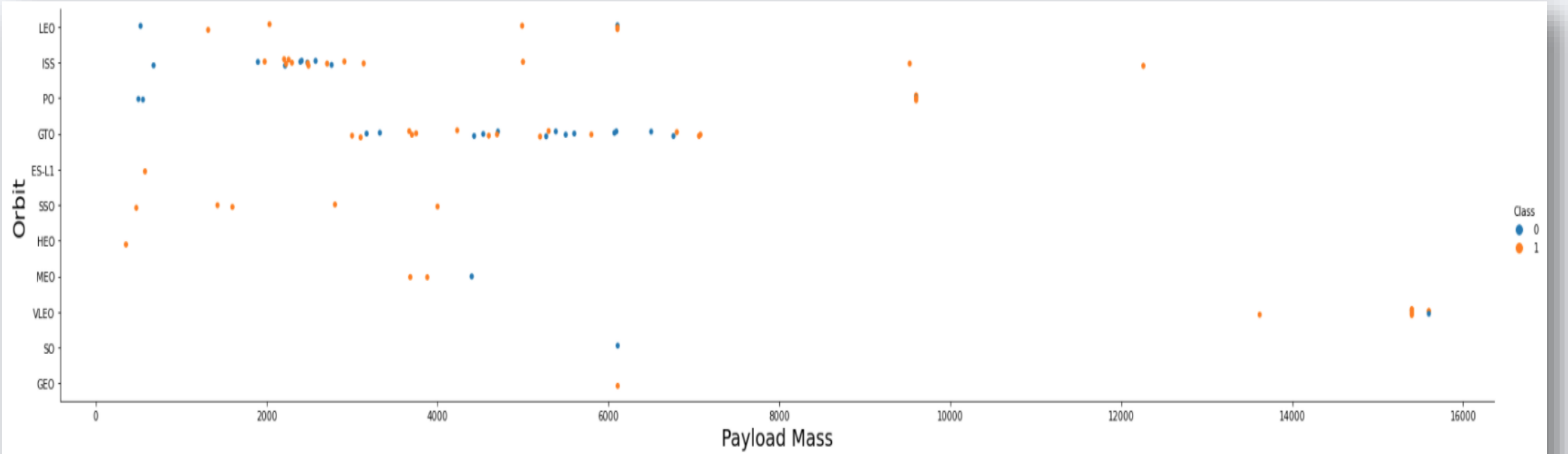
- ES-L1 - The Earth/Solar Lagrange point
- GEO - A high, circular geosynchronous equatorial orbit
- HEO - A highly elyptical orbit
- SSO - A sun-synchronous, nearly polar orbit

Flight Number vs. Orbit Type



We can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

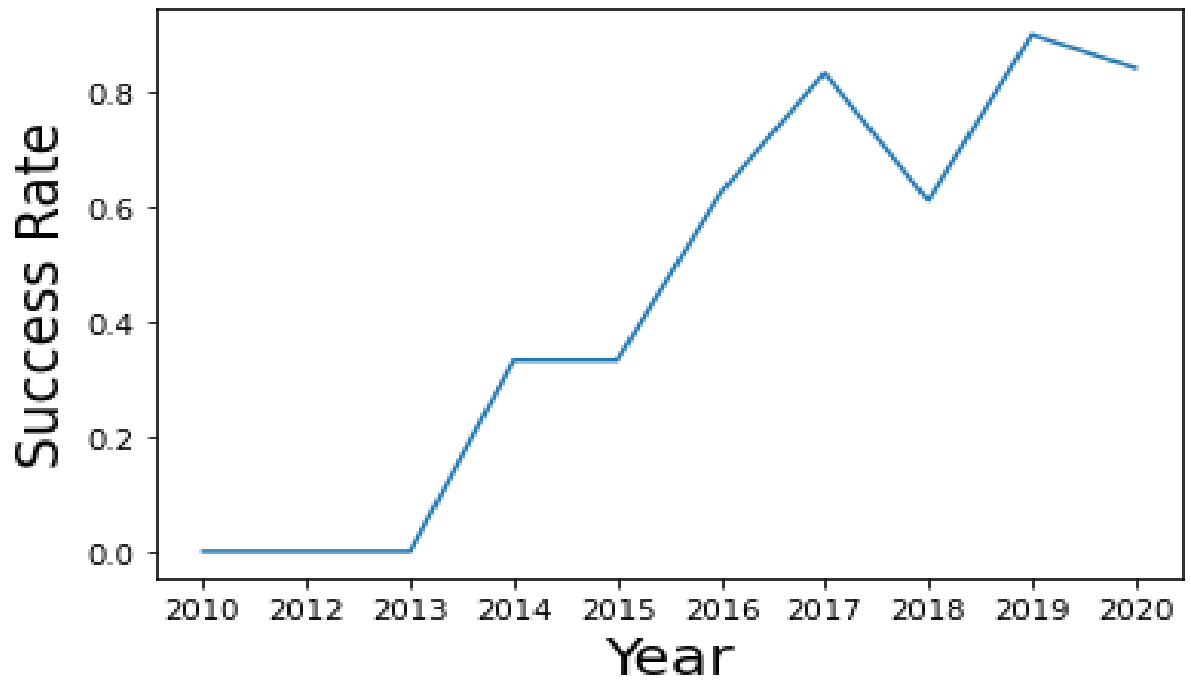
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are greater for Polar, LEO and ISS orbits.

However for GTO we cannot distinguish this clearly, as both positive landing and negative landings (unsuccessful mission), are present.

Launch Success Yearly Trend



As we saw from the first 2 scatter plots, the rate of success has generally increased since 2013, with a dip in 2018.

Section 3

Insights drawn from SQL

All Launch Site Names

- There are 4 unique launch sites:
  - CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E
- These were identified from the SpaceX DataSet using the following query:

  select distinct(launch_site) from SPACEXTBL

- This query is using the *distinct* statement to return a list of unique values from the launch_site column of the SPACEXTBL table which contains the SpaceX DataSet data

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- *select * from SPACEXTBL*
  where launch_site like 'CCA%'
  LIMIT 5

  - select all records from the SpaceX DataSet
  - use the % wildcard after CCA to allow for different values
  - return only the first 5 results to meet the criteria

Total Payload Mass

- The total payload carried by NASA (CRS) boosters was 45,596 kg
- This was identified from the SpaceX DataSet using the following query:

```
select sum(payload_mass__kg_)
from SPACEXTBL
where customer = 'NASA (CRS)'
```

- This query is using the *sum* function to sum all values in the payload_mass_kg_ column of the SPACEXTBL table which was filtered on the customer with the name of NASA (CRS)

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2,928 kg
- This was identified from the SpaceX DataSet using the following query:

```
select avg(payload_mass__kg_)
from SPACEXTBL
where booster_version like 'F9 v1.1'
```

- This query uses the avg function to find the mean of all payload mass values which were filtered by the booster version matching F9 v1.1

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was December 22nd 2015
- This was identified from the SpaceX DataSet using the following query:

```
select min(date)
from SPACEXTBL
where landing__outcome like 'Success (ground pad)'
```

- This query uses the min function to select the first (smallest) date which was filtered by the landing outcome of 'Success (ground pad)'

Successful Drone Ship Landing with Payload between 4,000 and 6,000 kg

- The following boosters have successfully landed on the drone ship and had a payload mass greater than 4,000 kg but less than 6,000 kg:

  F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2

- These were identified from the SpaceX DataSet using the following query:

```
select distinct(booster_version)
from SPACEXTBL
where payload_mass__kg_ between 4000 and payload_mass__kg_ < 6000
and landing__outcome like 'Success (drone ship)'
```

- This query uses the between keyword to select the payload mass range for the distinct 'Success (drone ship)' landing outcomes

Total Number of Successful and Failure Mission Outcomes

- There was 1 failed mission outcome, 99 successful mission outcome, with 1 mission successful, but with an unclear payload status
- This was identified from the SpaceX DataSet using the following query:

```
select count(mission_outcome),mission_outcome
from SPACEXTBL
group by mission_outcome
```

- This query uses the count aggregation method and has been grouped by the mission outcome with the outcome included to aid display

Boosters Carried Maximum Payload

- The following boosters have carried the maximum payload mass:

  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

- This was identified from the SpaceX DataSet using the following query:

```
select distinct(booster_version)
from SPACEXTBL
where payload_mass__kg_ =
   (select max(payload_mass__kg_) from SPACEXTBL)
order by booster_version
```

- This query uses a subquery to determine the maximum value of the payload mass, then uses that value to determine which boosters carried that mass

2015 Launch Records

- The following are the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- These were identified from the SpaceX DataSet using the following query:

```
select date, landing__outcome, booster_version, launch_site
from SPACEXTBL
where year(date) = 2015
and landing__outcome like 'Failure (drone ship)'
```

- This query uses the year function with the date field to return only those results from 2015

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

The table below shows the frequency of the landing outcomes between June the 4th 2010 and March the 20th 2017 in descending order:

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

This was identified from the SpaceX DataSet using the following query:

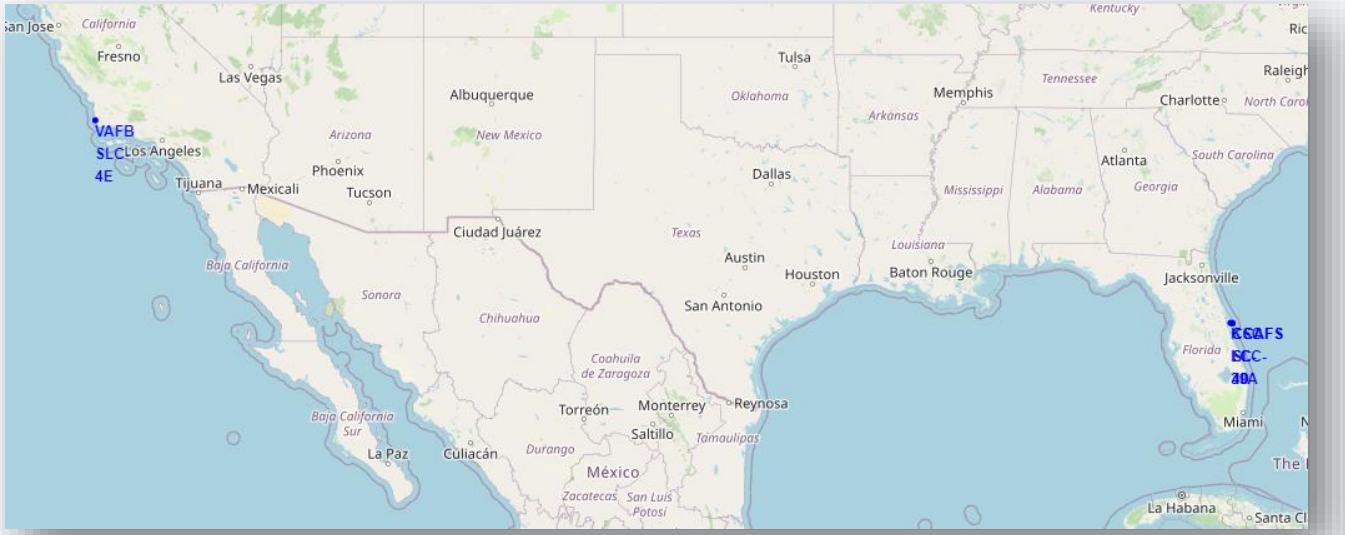
```
select landing__outcome as "Landing Outcome",
count(landing__outcome) as "Count"
from SPACEXTBL
where date between '2010-06-04' and '2017-03-20'
group by landing__outcome
order by 2 desc
```

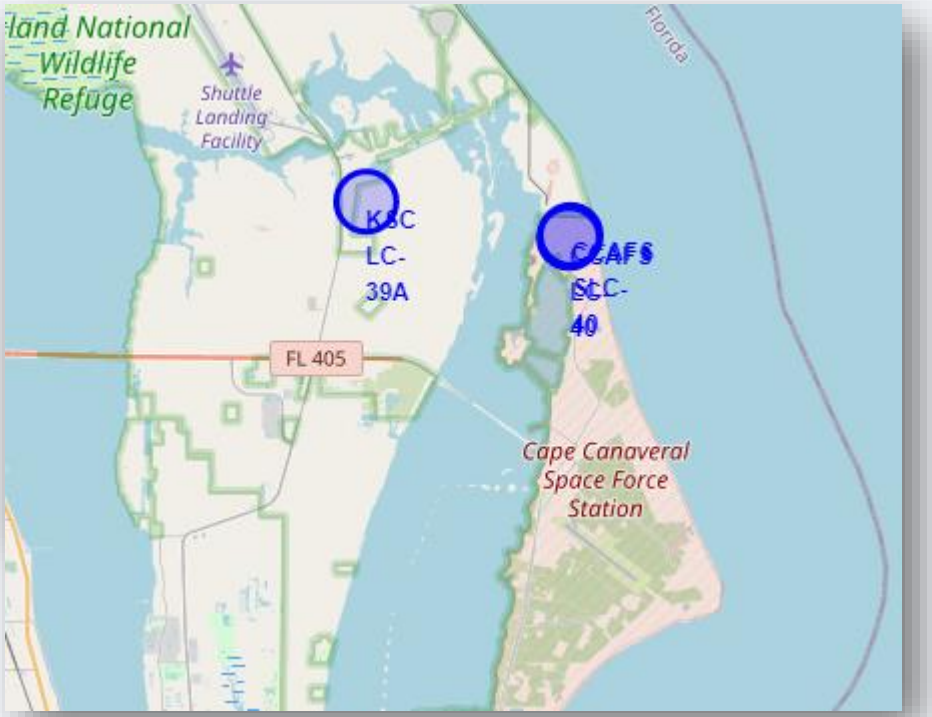
This query again makes use of the between keyword to specify the date range; the grouping is by the landing outcome; and we rank the frequency referring to the aggregate column by number in the result set (2).

Section 4

Launch Sites Proximities Analysis

Launch Site Locations

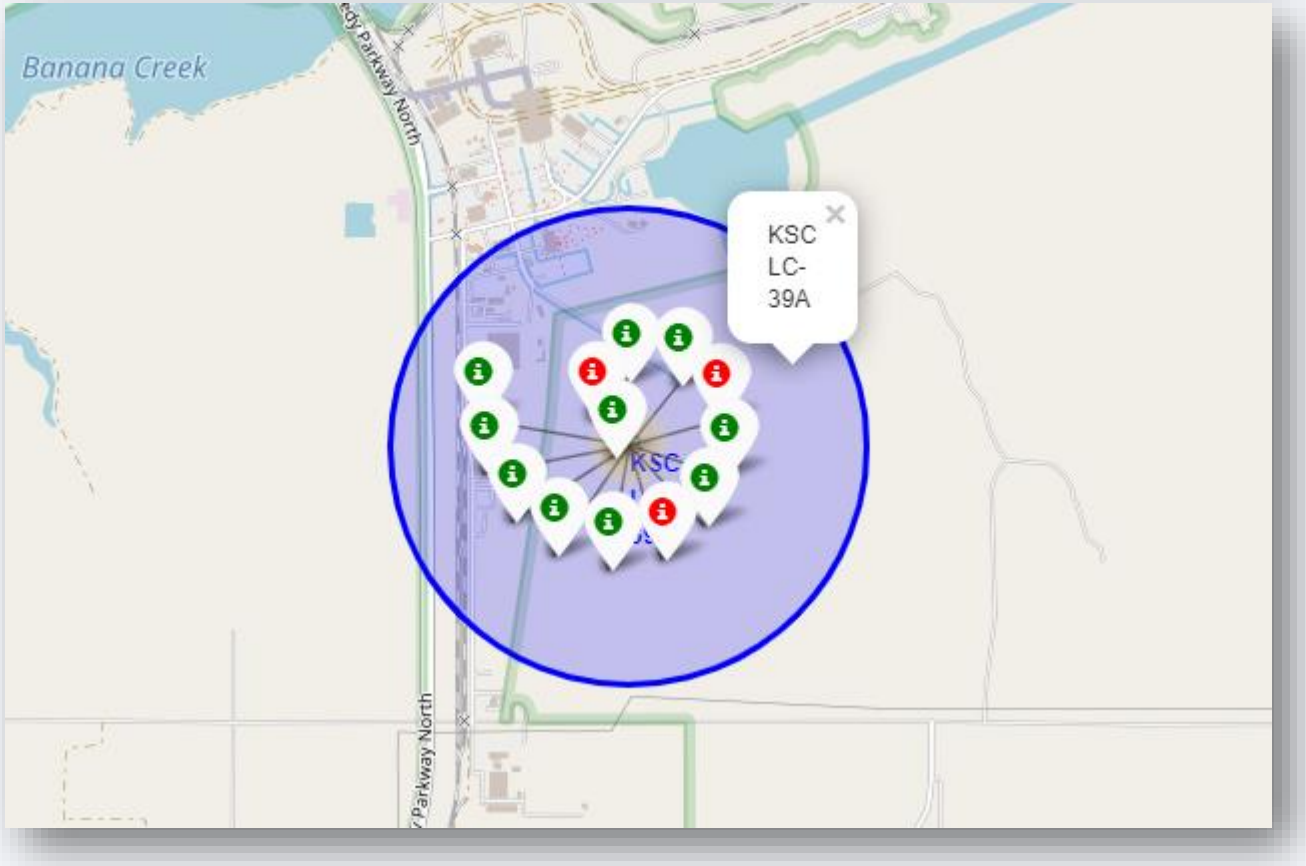




- The image above shows all launch site locations, with the image on the right expanded to show the eastern sites which are in close proximity to each other
- Notice that all sites are practically on the coast and as close to the equator as possible

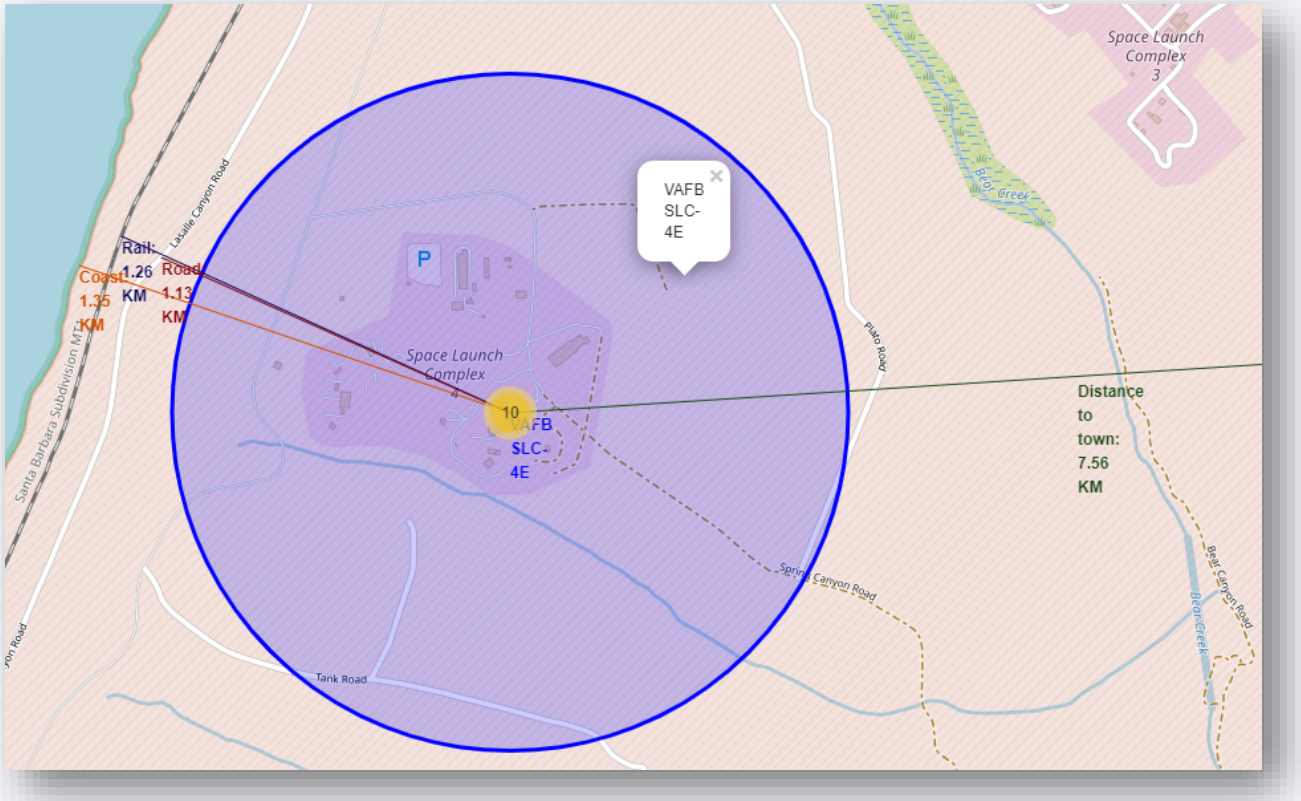
Successful Site Launches – KSC LC-39A

- This image provides a graphical representation of the launch successes for site KSC LC-39A
- Successful launches are marked in green, with failed launches marked in red



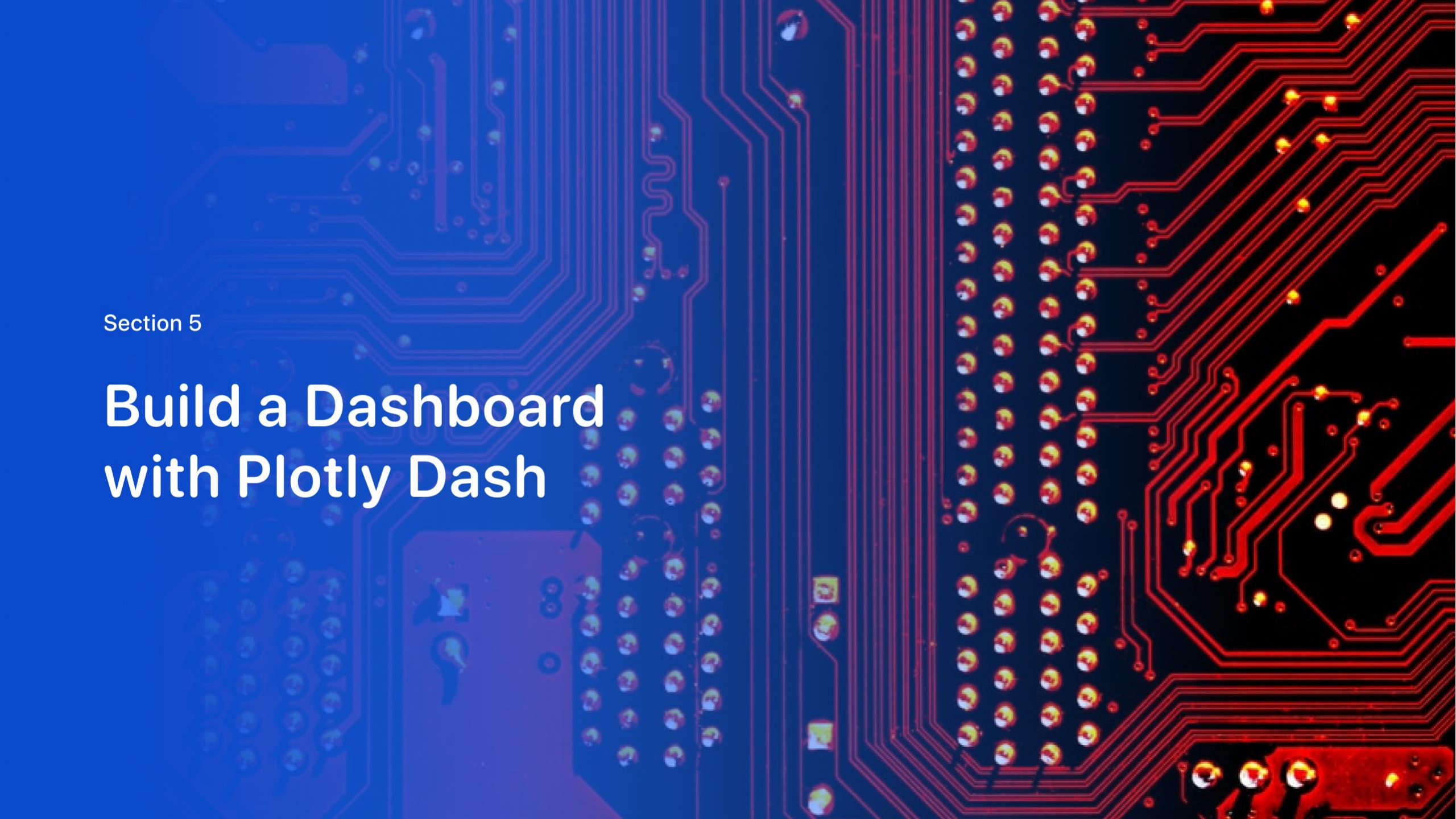
Distances to Items of Interest

- This image gives distances to several close points of reference including the coast, a road and a railway line
- The nearest town is far enough away that to include more than a reference to its location would render the other items illegible



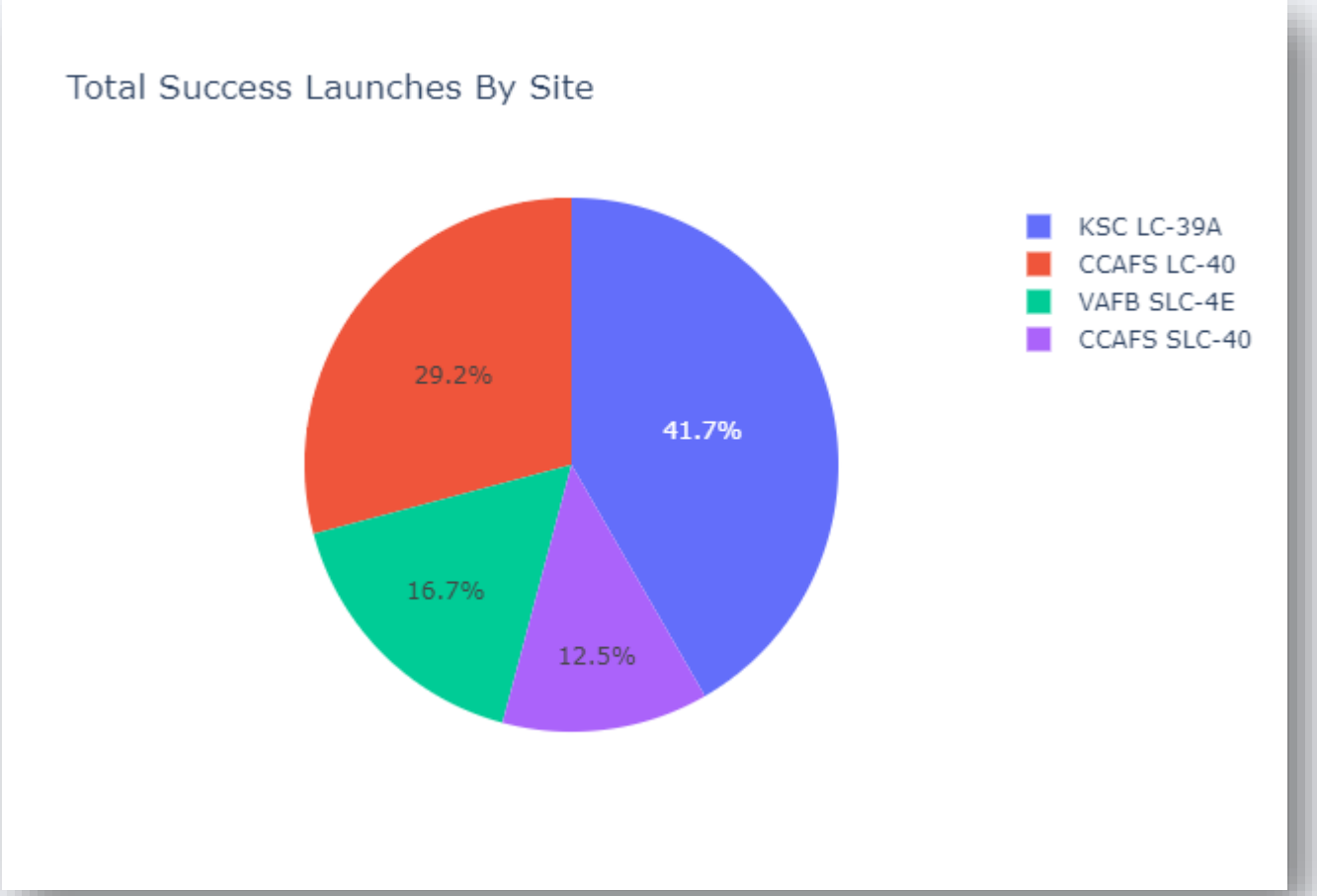
Section 5

Build a Dashboard with Plotly Dash



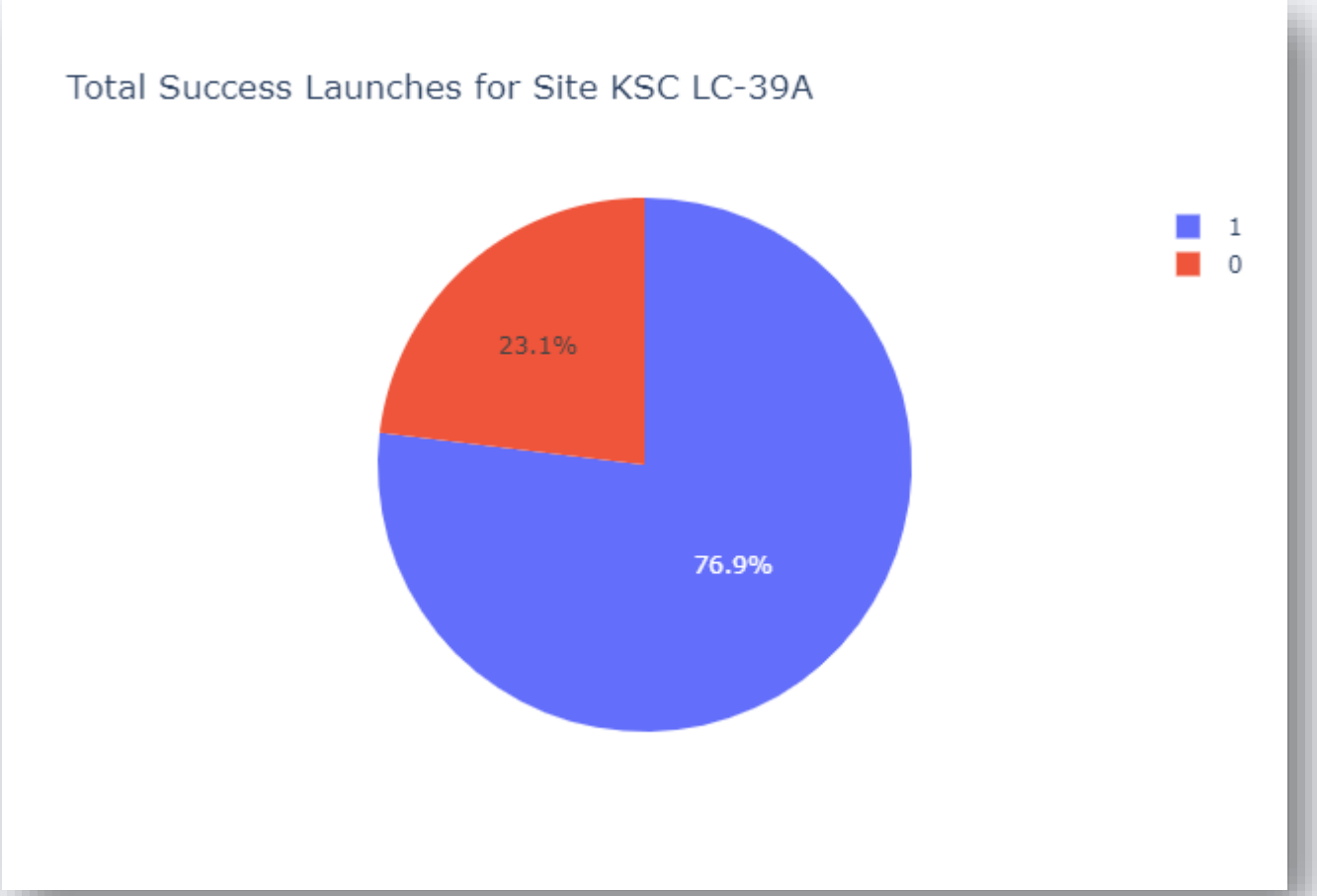
SpaceX Successful Launch % Per Site

- The figure opposite shows the successful launches by site as a % of total launches made during the period
- Three obvious limitations of this data are:
  - We do not know how many launches were made per site
  - The payload mass for each of the launches (determined later)
  - The date of each of the launches (later launches are typically more successful)

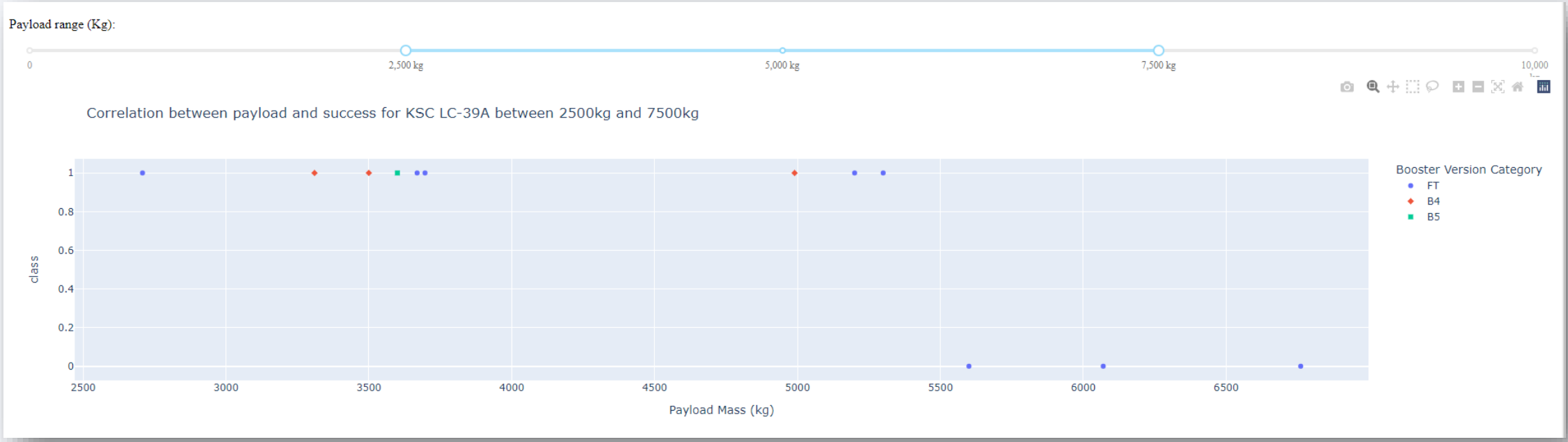


Successful Launch Ratio For Site KSC LC-39A

- Of the 4 sites, KSC LC-39A had the highest individual successful launch rate with 76.9%
- The figure opposite shows the breakdown by class with a 1 representing a successful launch



Analysis of Payload vs Success For Site KSC LC-39A



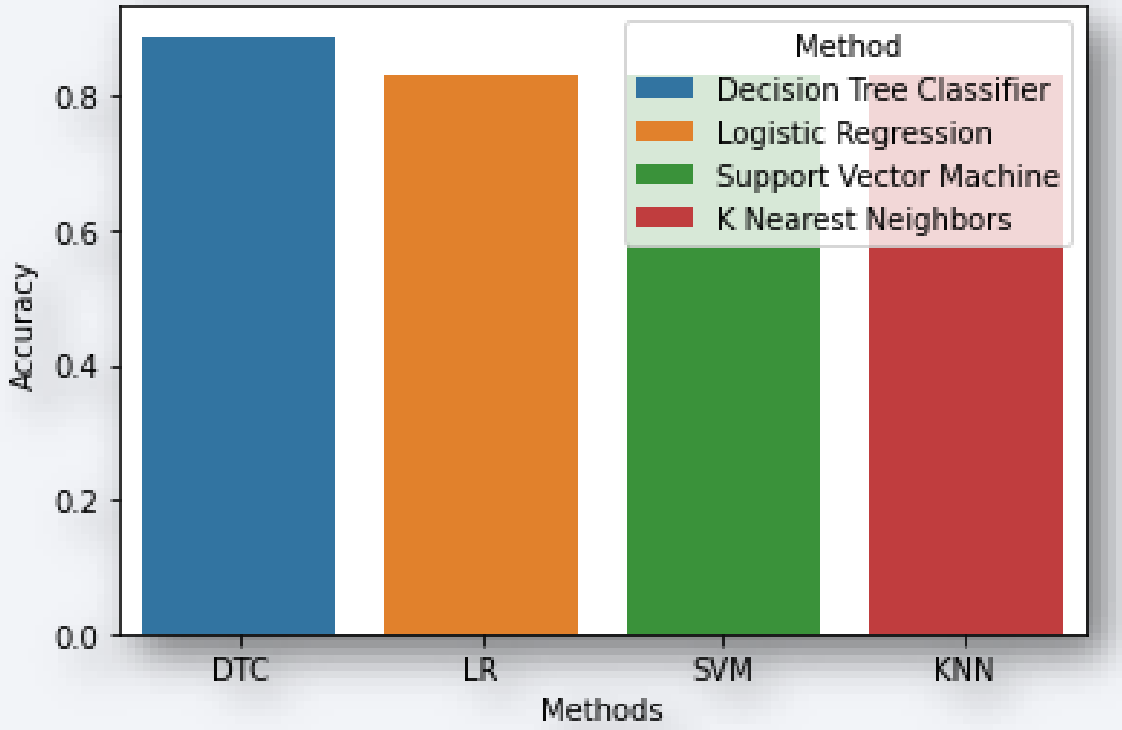
- The figure above shows the correlation of the payload mass to the success for site KSC LC-39A which has been taken from an interactive scatter graph generated using Plotly Dash
- As can be seen, the most successful payload mass for launches from the site during the period is between 2,500kg and 5,500kg

Section 6

Predictive Analysis (Classification)

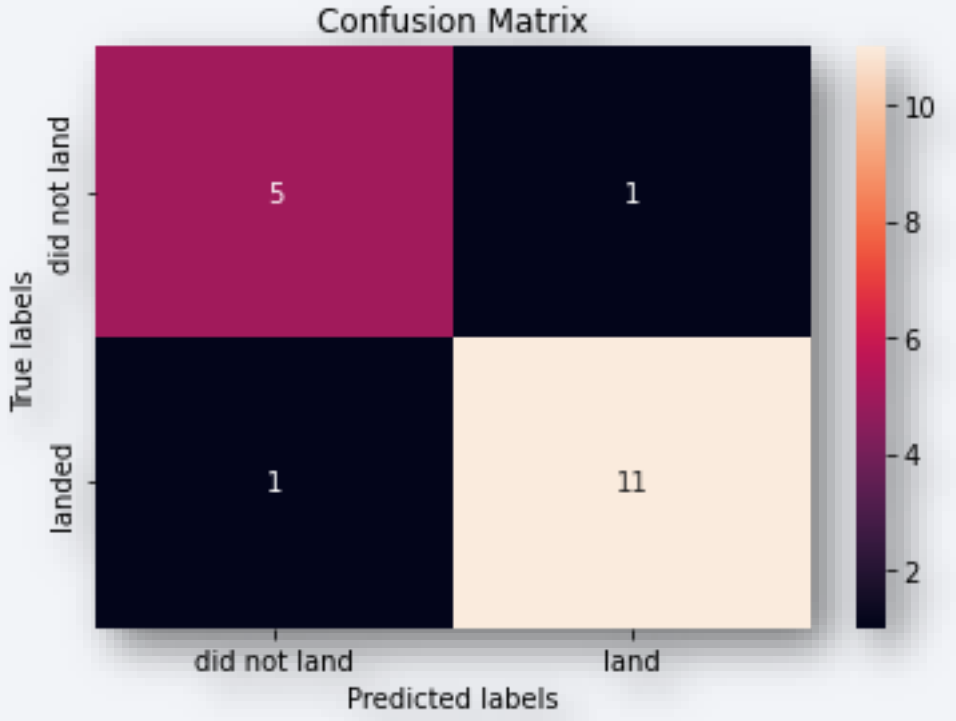
Classification Accuracy

- It should be noted that we only have a small number of test samples which likely gives rise to the variability between runs
- Where there is a most accurate model it is typically the Decision Tree Classifier that returned the highest accuracy, in this instance at 88.8889%



Confusion Matrix (DTC)

- The correct predictions are on the diagonal from top-left to bottom-right
- 5 of the 6 failed landings were correctly identified along with 11 of the 12 successful landings
- The other models all predicted a surplus of positive landings



Conclusions

- SpaceY wishes to find a commercially viable price point to enter the industry
- Assuming that they will be able to produce rockets with approximately the same basic cost as SpaceX they need to determine the cost of a given launch
- With the data available for modeling and the techniques used, there is an 88.8889% prediction rate of landing success based on the most accurate predictive model (DCT)
- Assuming that the price of launch with a successful landing is $62m, the price without landing is $165m, and that SpaceY will only target launches with a successful landing, there is a mean launch price of $76.11m
- Orbit type appears to display a more significant correlation with successful recovery than other factors. This is impacted by the nature of the payload (observation craft, satellite, ISS mission), and may inform a more granular pricing model
- This is a limited data set so a larger number of iterations than used during this analysis may be beneficial for further investigation if no further data can be gathered
- Launch sites need to be near a large body of water, a railway line and/or a major highway, and a reasonable distance from cities for safety

Appendix

- This presentation has been generated as part of the IBM Data Science Professional Certificate and has used the following resources:
  - SpaceX API – https://api.spacexdata.com
  - SpaceX Wikipedia page – https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Infographic by https://zlsadesign.com/
- All slides, notebooks and data sources are held at https://github.com/matthewrwinder/LaunchSites/tree/master

Thank you!