IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Balu Prakash
02-Oct-2021

Outline

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

Executive Summary

- **Objective:**
  Develop a classification model using Machine Learning, to predict the **landing outcome** (ie. success vs crash) of a Falcon 9 First Stage.

- **Approach / Methodology:**
  Data from 90 previous SpaceX Falcon 9 launches was collected and explored to determine significant relationships and the most relevant features (eg. Launch Site, Payload, etc), via EDA techniques including Folium, Plotly Dash, and various standard plots.

  Four different Machine Learning models were then developed: Logistic Regression, K-Nearest Neighbour, Support Vector Machine, and Decision Tree.
  Models were trained and optimised on the data after normalisation, using only the 'train' portion of a train/test split, before being tested on the 'test' set.

- **Results:**
  Three of the four models performed identically on the test set (n=18), achieving an 83% accuracy score for predicting the landing success. The other model (Decision Tree) only obtained 72% accuracy.
  The three 'winning' models tend to over-predict landing success – ie. For the failed landings in the test set (n = 6), it only correctly predicted the outcome for 50% (3/6).
  Whereas for all the successful landings (n=12), it correctly predicted the outcome for all of them.

Introduction

- **Project background and context:**

  We would like to start a competitor company ‘SpaceY’.
  A key factor in launch cost is whether or not the first stage booster can be successfully landed for later re-use.
  Therefore, if we can predict the chance of successful landing of the first stage booster, we may be able to out-bid SpaceX for launch contracts.

- **Problem to solve:**

  For any given launch scenario, we want to be able to predict whether the first stage booster will be successfully landed.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from two sources:
    - (i) SpaceX website, using HTTP requests through the SpaceX API
    - (ii) Wikipedia, using webscraping to extract tables from the HTML using BeautifulSoup
- Perform data wrangling
  - Data was processed by converting to a DataFrame, creating training labels for the data based on landing outcome, and examining summaries of each feature.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - 4 classification models (Logistic Regression, Support Vector Machine, K Nearest Neighbour, and Decision Tree) were built on the ‘train’ portion of the normalised dataset
  - Hyperparameters for each model were tuned & optimised using GridSearchCV with 10-fold cross-validation.
  - Models were then evaluated based on their predictions given the ‘test’ dataset. Confusion matrices and accuracy score were used for evaluation.

Data Collection

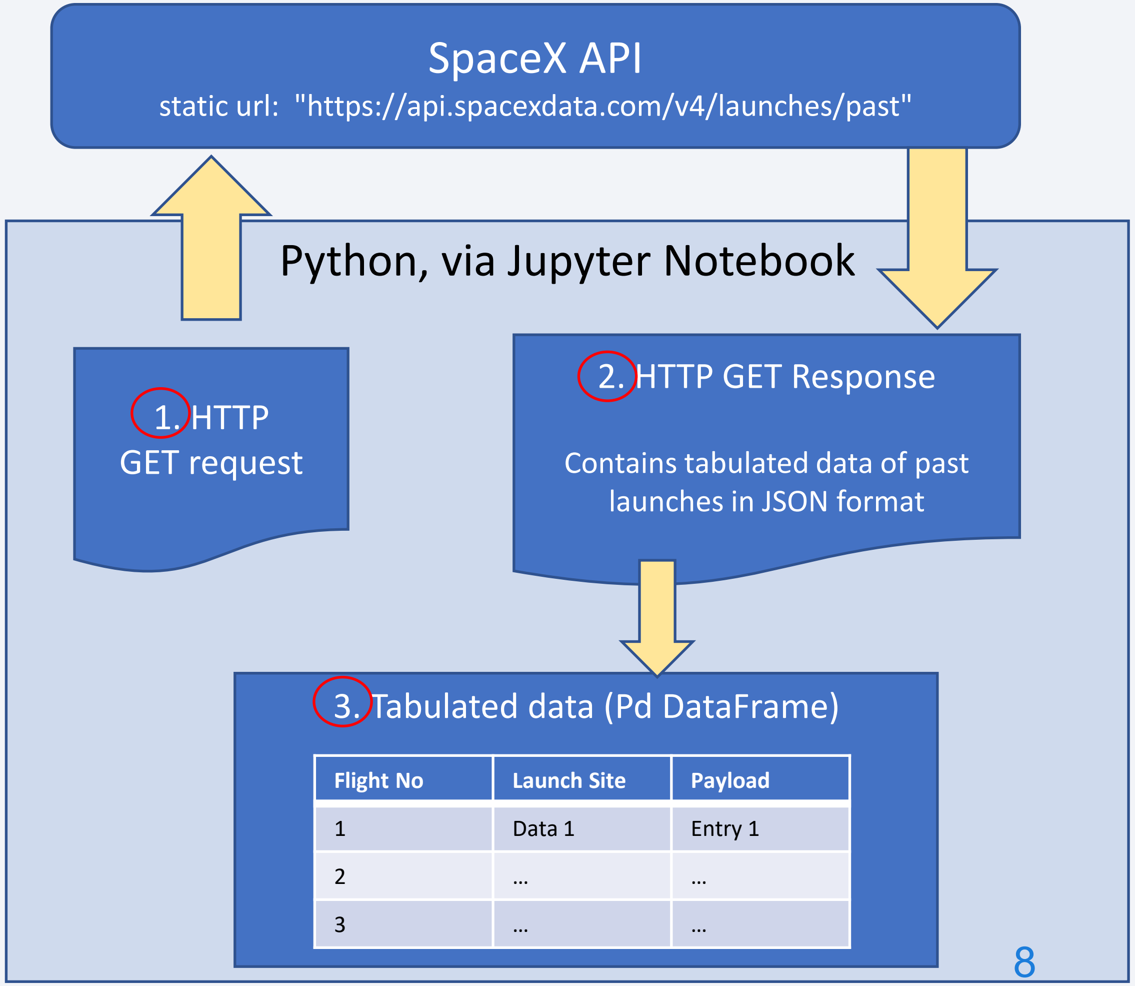
Data was collected from two sources:

1. SpaceX website, using HTTP requests through the SpaceX API
2. Wikipedia, using webscraping to extract tables from the HTML using BeautifulSoup

Data Collection – SpaceX API

- Past launch data was obtained using an HTTP GET Request from the SpaceX REST API
- The GET Response contained tabulated data of past launches in JSON format.
- The JSON format was parsed into a Pandas Dataframe.

- GitHub URL of the completed notebook: https://github.com/monsieur-le-git/IBM_DataSci_Course/blob/main/2.%20Data%20Collection%20with%20Web%20Scraping%20Lab.ipynb



SpaceX API

static url: "https://api.spacexdata.com/v4/launches/past"

Python, via Jupyter Notebook

1. HTTP GET request

2. HTTP GET Response

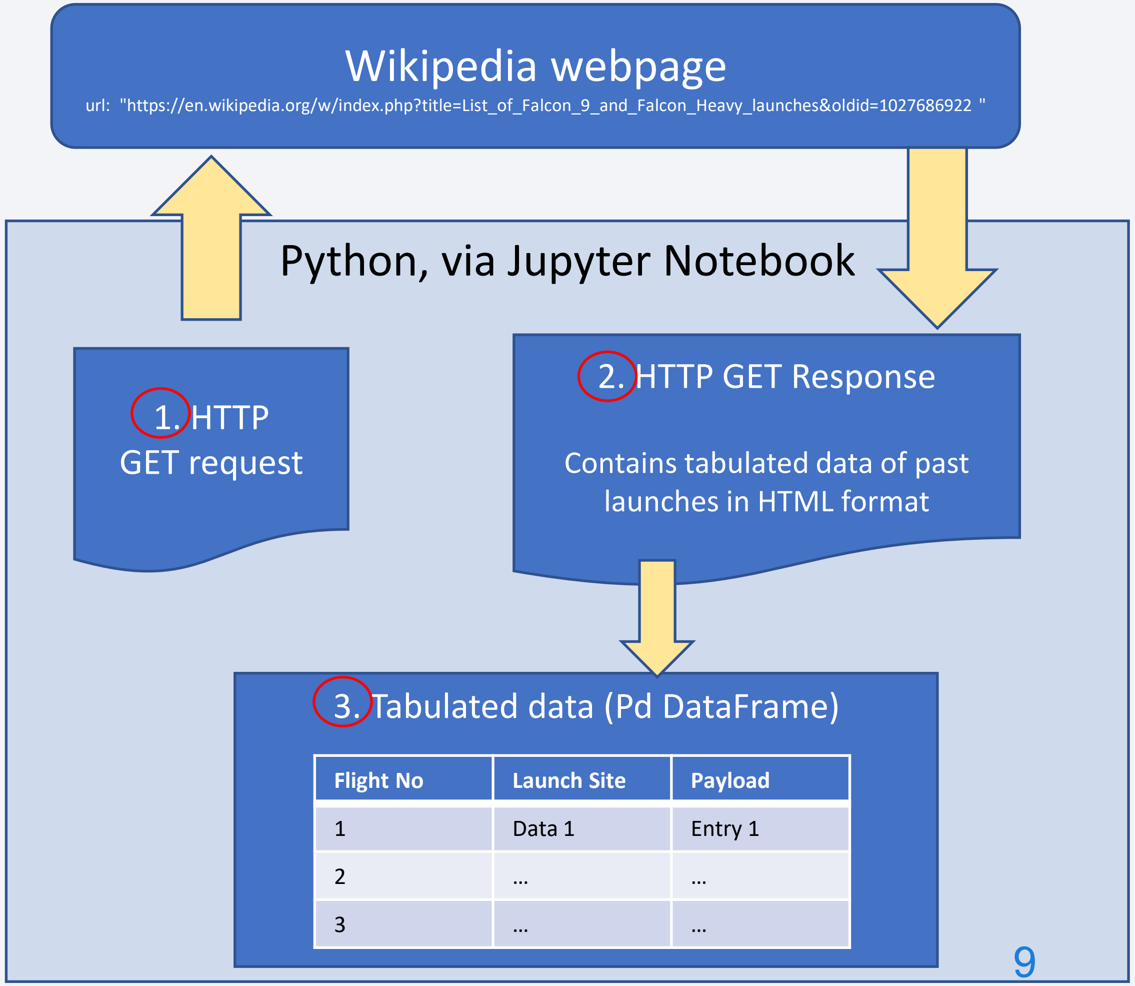
Contains tabulated data of past launches in JSON format

3. Tabulated data (Pd DataFrame)

Flight No	Launch Site	Payload
1	Data 1	Entry 1
2	…	…
3	…	…

Data Collection - Scraping

- Additional past launch data was obtained from a Wikipedia page, using Web Scraping (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- An HTTP GET Request was sent to the Wikipedia page, and the response contained tabulated data of past launches in HTML format.
- The HTML format was parsed into a Pandas Dataframe using the BeautifulSoup module within Python.
- GitHub URL of the completed notebook: https://github.com/monsieur-le-git/IBM_DataSci_Course/blob/main/2.%20Data%20Collection%20with%20Web%20Scraping%20Lab.ipynb



Wikipedia webpage

url: "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922 "

Python, via Jupyter Notebook

1. HTTP GET request

2. HTTP GET Response

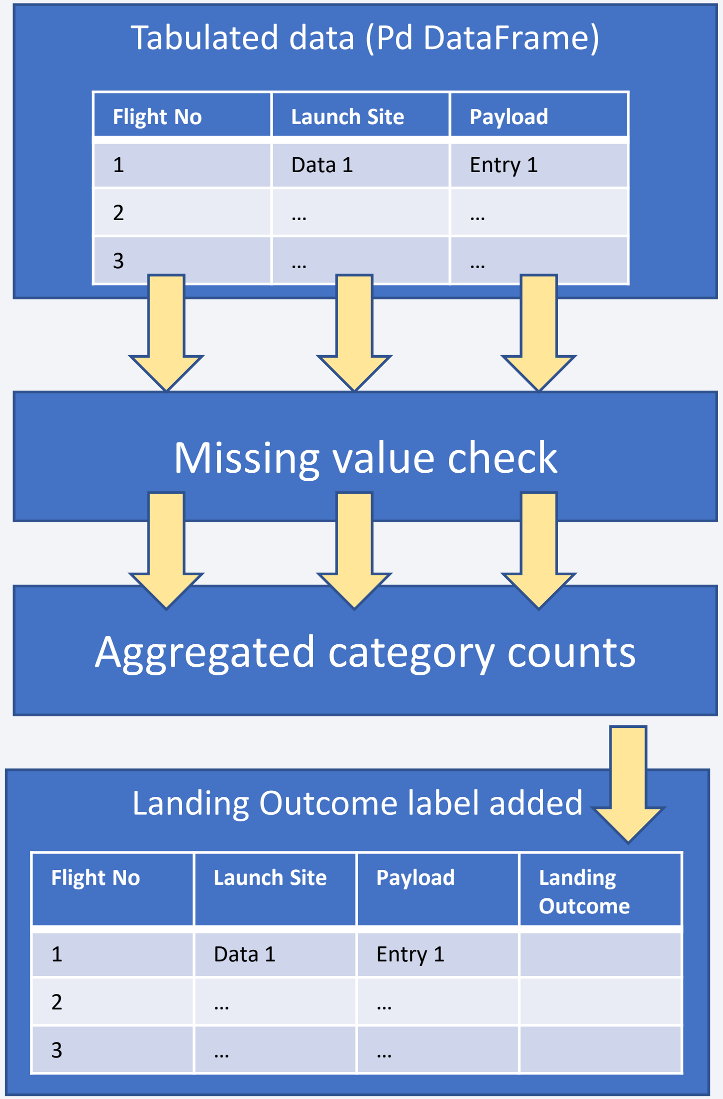
Contains tabulated data of past launches in HTML format

3. Tabulated data (Pd DataFrame)

Flight No	Launch Site	Payload
1	Data 1	Entry 1
2	…	…
3	…	…

Data Wrangling

- Features were examined for missing values, and for their data type (numerical / categorical)
- The following features were analysed for “number of occurrences in each category”, using value counts:
  - Launch site
  - Orbit type
  - Mission outcome
- A ‘landing outcome’ label was created from the Outcome column, to indicate landing success or not
- In later labs, the following data wrangling steps were performed:
  - One-hot encoding of categorical variables into dummy variables. (Orbit, Launch Site, Landing Pad, Booster Serial)
  - Conversion of non-numeric columns to numeric (float64)
  - Normalisation of the dataset
- GitHub URL of the completed notebook: https://github.com/monsieur-le-git/IBM_DataSci_Course/blob/main/3.%20Data%20Wrangling%20Lab.ipynb



Tabulated data (Pd DataFrame)

Flight No	Launch Site	Payload
1	Data 1	Entry 1
2	…	…
3	…	…

Missing value check

Aggregated category counts

Landing Outcome label added

Flight No	Launch Site	Payload	Landing Outcome
1	Data 1	Entry 1	
2	…	…	
3	…	…	

EDA with SQL

- SQL queries were performed to explore the dataset in the following ways:
  - Listing of launch sites – unique, and specific using LIKE, LIMIT, DISTINCT
  - Listing payload mass averages for different conditions (eg. NASA CRS, Booster Version) using AVG()
  - Dates for first successful landing using MIN()
  - Listing Booster names which met certain criteria (Success, with Payload 4k – 6k) using AND
  - Listing total numbers for success & failure – overall, and for specific criteria using GROUP BY and Subquery

- Some data wrangling was also performed in this lab – namely, conversion of categorical variables into dummy variables via one-hot encoding (Orbit, Launch Site, Landing Pad, Booster Serial).

- GitHub URL of the completed notebook: https://github.com/monsieur-le-git/IBM_DataSci_Course/blob/main/4.%20EDA%20with%20SQL%20Lab.ipynb

- PLEASE NOTE: The Notebook linked to above does not have the code output cells showing correctly, due to a known error reported on the Coursera Discussion Forums. So I completed the workbook in Skills Network Labs as suggested by Instructor Lakshmi Holla, where all queries ran correctly. Then I pasted the code back into the workbook in Watson Studio to publish to GitHub, but still cannot get it to run there due to the issue. Issue on Discussion Forums described at: https://www.coursera.org/learn/applied-data-science-capstone/discussions/weeks/2/threads/2kjcMxbxEeyzTxKP_hNrxQ

EDA with Data Visualization

- The following charts were plotted and analysed, to try to identify success factors in landing outcomes:
  - Scatter Plots
    - Flight Number vs Payload Mass
    - Flight Number vs Launch Site
    - Payload Mass vs Launch Site
    - Flight Number vs Orbit Type
    - Payload Mass vs Orbit Type
  - Bar plot (best for showing categorical data)
    - Orbit Type vs Landing Success Rate
  - Line plot (best for showing trend vs time)
    - Launch Success vs Year (ie. Yearly trend)
- Some data wrangling was also performed in this lab – namely, conversion of categorical variables into dummy variables via one-hot encoding (Orbit, Launch Site, Landing Pad, Booster Serial).
- GitHub URL of the completed notebook: https://github.com/monsieur-le-git/IBM_DataSci_Course/blob/main/5.%20EDA%20with%20Visualization%20Lab.ipynb

Build an Interactive Map with Folium

- Markers were added to a Folium map to show Launch Site geographic location, along with landing success outcomes.
  Why? So that the relationship between Launch Site and Landing Outcome could be explored visually.

- Markers added included:

  - Launch Site location: Icon, circle, and text for Launch Site Name.
    These were added so that we could see where the Launch sites were in relation to each other

  - Landing Success: Marker Cluster with multiple icons colored red & green for landing outcome
    These were added so that we could explore whether a certain geographic location had a higher rate of landing success

  - Launch Site proximities: Lines and text, to show distance from Launch Sites to nearby points of interest – railway yard, coastline
    These were added so we could see whether launch site proximities correlated with landing success

- GitHub URL of the completed notebook:
  https://github.com/monsieur-le-git/IBM_DataSci_Course/blob/main/6.%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

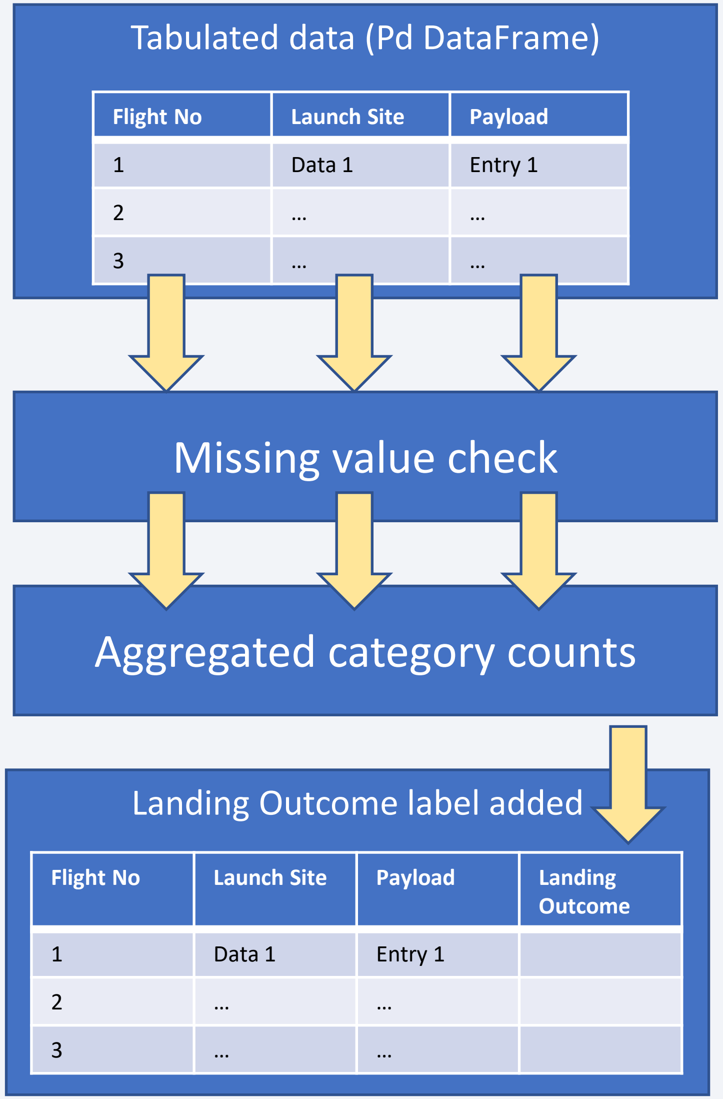
Build a Dashboard with Plotly Dash

- A dashboard was created to allow interactive exploration of the data.
  Why? To examine relationship between Launch Site, Landing Outcome, and Payload Mass.
- Dashboard included inputs:
  - Dropdown menu for selecting Launch Site – either individually, or ‘all’
  - Range slider for selecting Payload Mass
- Dashboard included outputs:
  - Pie chart showing landing success rates
  - Scatter plot showing Flight Number vs Payload Mass and Landing Outcome
- GitHub URL of the completed Python file:
  https://github.com/monsieur-le-git/IBM_DataSci_Course/blob/main/7.%20Interactive%20Visual%20Analytics%20with%20Plotly%20Dash.py

Predictive Analysis (Classification)

- Dataset was split into train/test, and models were developed using only the ‘train’ portion of the normalised dataset.
- 4 classification models (Logistic Regression, Support Vector Machine, K Nearest Neighbour, and Decision Tree) were built using the scikit learn module.
- Hyperparameters for each model were tuned & optimised using GridSearchCV with 10-fold cross-validation.
- Finally, to evaluate models and select the best one, all models were given the ‘test’ dataset to see how accurate their predictions were. Confusion matrices and accuracy score were used for evaluation.

- GitHub URL of the completed notebook: https://github.com/monsieur-le-git/IBM_DataSci_Course/blob/main/8.%20Machine%20Learning%20Prediction%20Lab.ipynb



Tabulated data (Pd DataFrame)

Flight No	Launch Site	Payload
1	Data 1	Entry 1
2	…	…
3	…	…

Missing value check

Aggregated category counts

Landing Outcome label added

Flight No	Launch Site	Payload	Landing Outcome
1	Data 1	Entry 1	
2	…	…	
3	…	…	

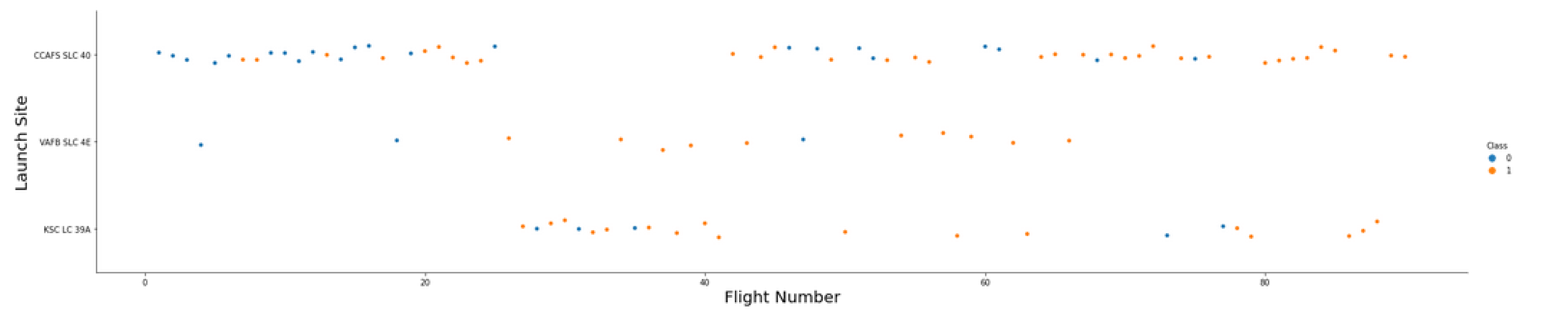
Results

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

Section 2

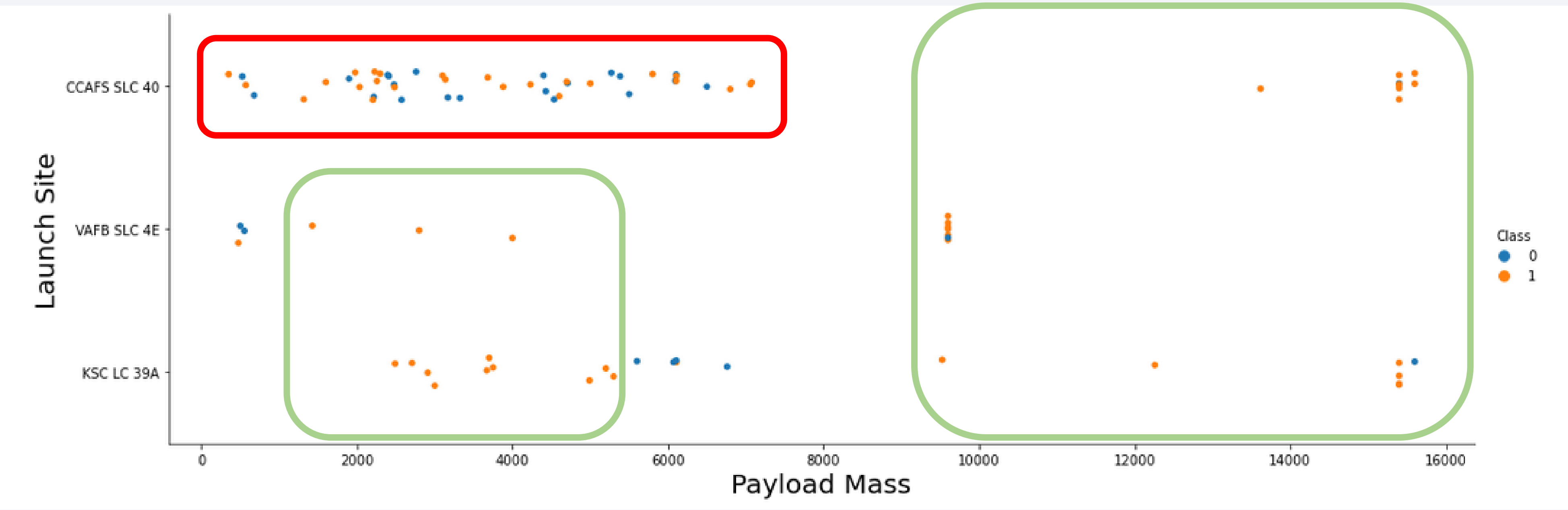
Insights drawn from EDA

Flight Number vs. Launch Site



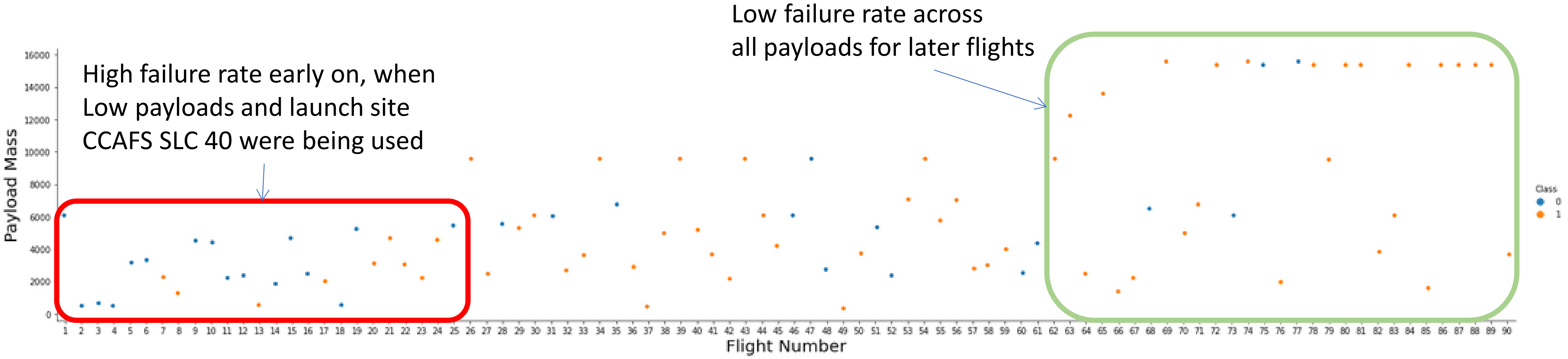
- Landing success becomes more frequent as flight number increases.
  This trend is seen overall, and for each Launch site.
  This may be due to the SpaceX team learning over time, making improvements, and becoming better at successfully landing.
- There does not seem to be a significant correlation between Launch Site and Landing success, once you remove the ‘confounder’ of CCAFS being used for the first 25 flights.
  Reasoning: For the last 10 launches at each site, there were only 1-2 landing failures for each.

Payload vs. Launch Site



- Initially, there *appears* to be a higher success rate for large payloads, ie. > 8000 kg, and for lower payloads (1500 – 5500 kg) for Launch Sites KSC LC 39A and VAFB SLC 4E.
- It also *looks like* there is a poor rate of success at the top-left, ie. For payloads up to 8000 kg launched from CCAFS SLC 40.
- HOWEVER – I decided to plot an EXTRA chart (see next slide) which reveals the real reason for this – it's because the EARLIER FLIGHT numbers used lower payloads, and launched from CCAFS SLC 40.

EXTRA CHART: Payload vs. Flight Number



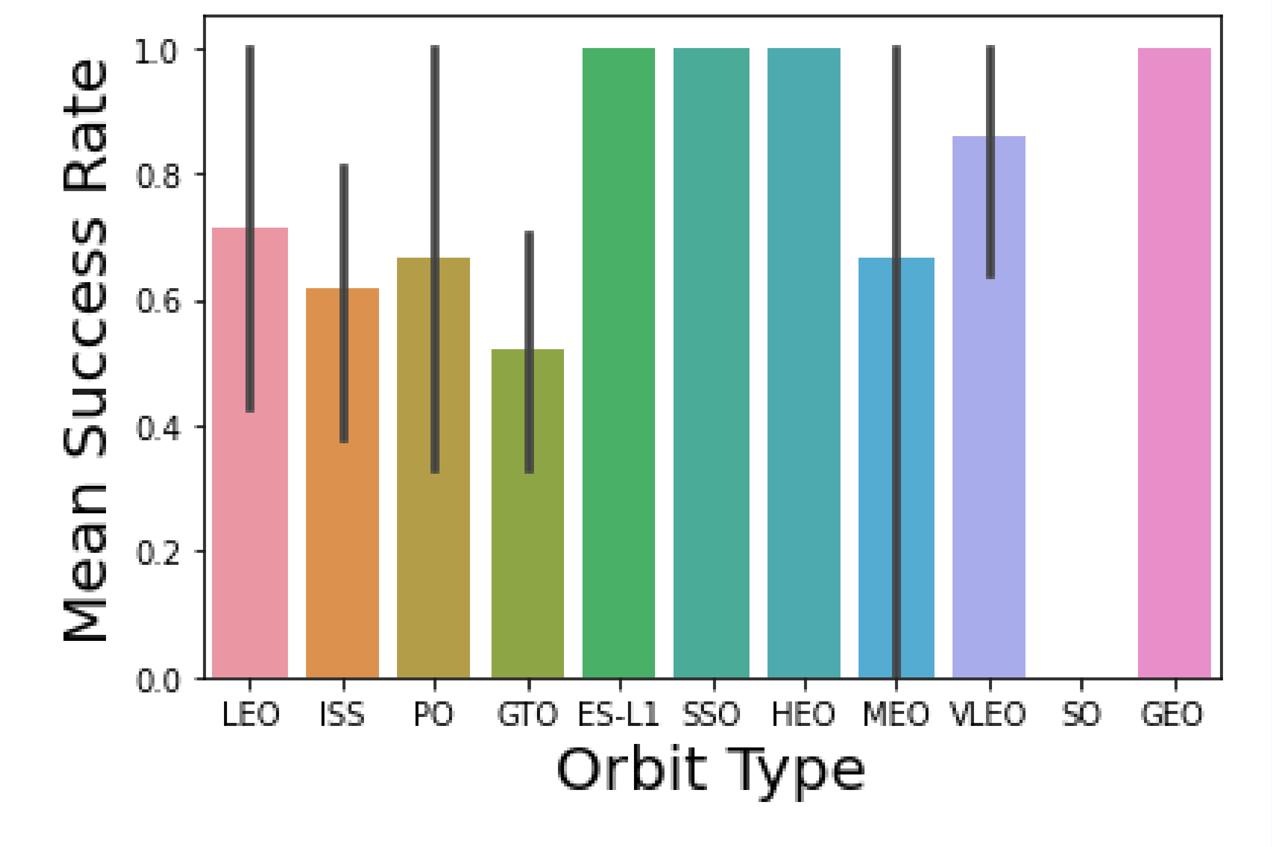
High failure rate early on, when Low payloads and launch site CCAFS SLC 40 were being used

Low failure rate across all payloads for later flights

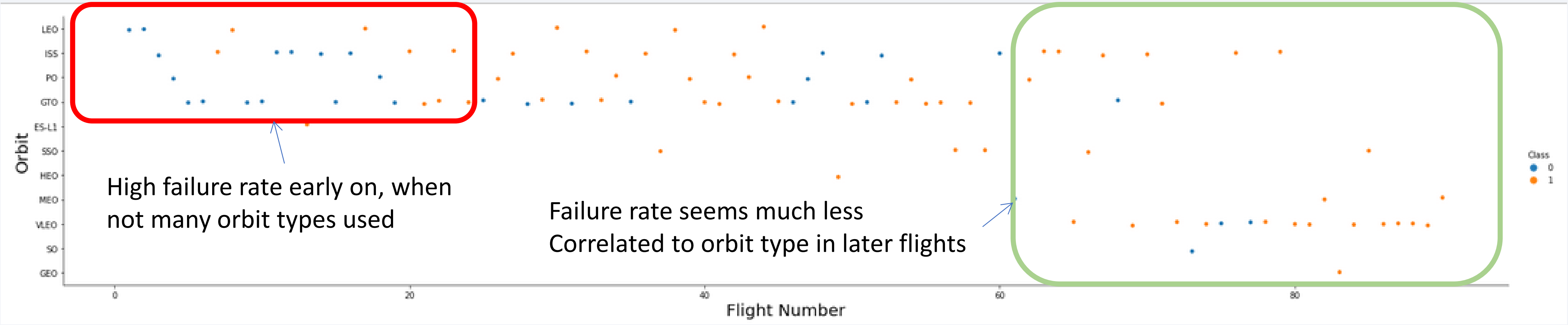
- As can be seen, there is a clear trend of Payload increasing with flight number.
  This means that the ‘apparent’ relationship between Landing Success and Payload on the previous slide, is most likely actually due to the trend of success increasing over time.
- Supporting this theory: for flights 62-90, where Payloads are much more evenly spread, there does not seem to be a high correlation between Payload and success rate.
- The first 25 flights contained many failures – but they were also conducted with low payloads, and almost all were from Launch Site CCAFS SLC 40.
- From this, I conclude that there is not likely any significant correlation between Payload Mass and Landing success, once you remove the confounder of the early flights using low Payloads.

Success Rate vs. Orbit Type

- Initially it appears that certain orbit types would be good predictors of success or failure
- HOWEVER – similar to the last slide, I realised this could be due to the improved success rate over time.
- Therefore the next slide examines Orbit Type vs Flight number



Flight Number vs. Orbit Type

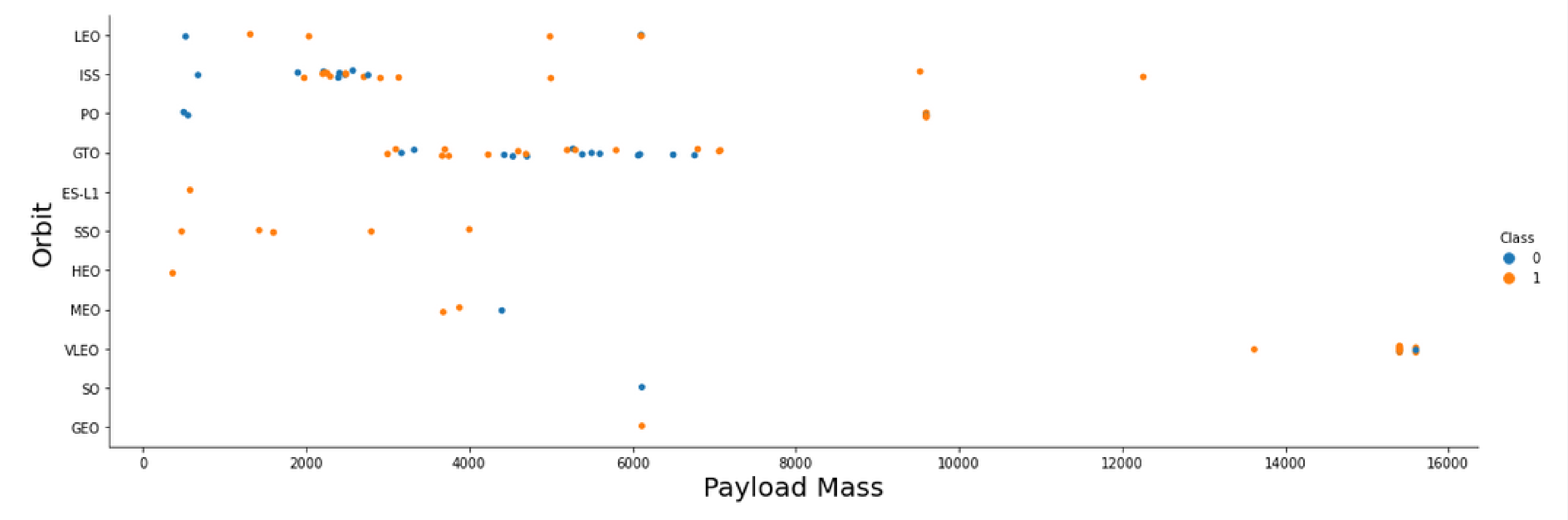


High failure rate early on, when not many orbit types used

Failure rate seems much less Correlated to orbit type in later flights

- As can be seen, it looks as though Orbit type is not a strong predictor of Landing Success, once you remove the confounding factor that success rate improves over time.
- In the early flights, there is a high failure rate, when only GTO, PO, LEO, and ISS orbits are being used.
- But as time goes on, the failure rate decreases across ALL orbit types.
- Therefore I conclude that Orbit type is not likely to be a strong predictor of landing success.

Payload vs. Orbit Type

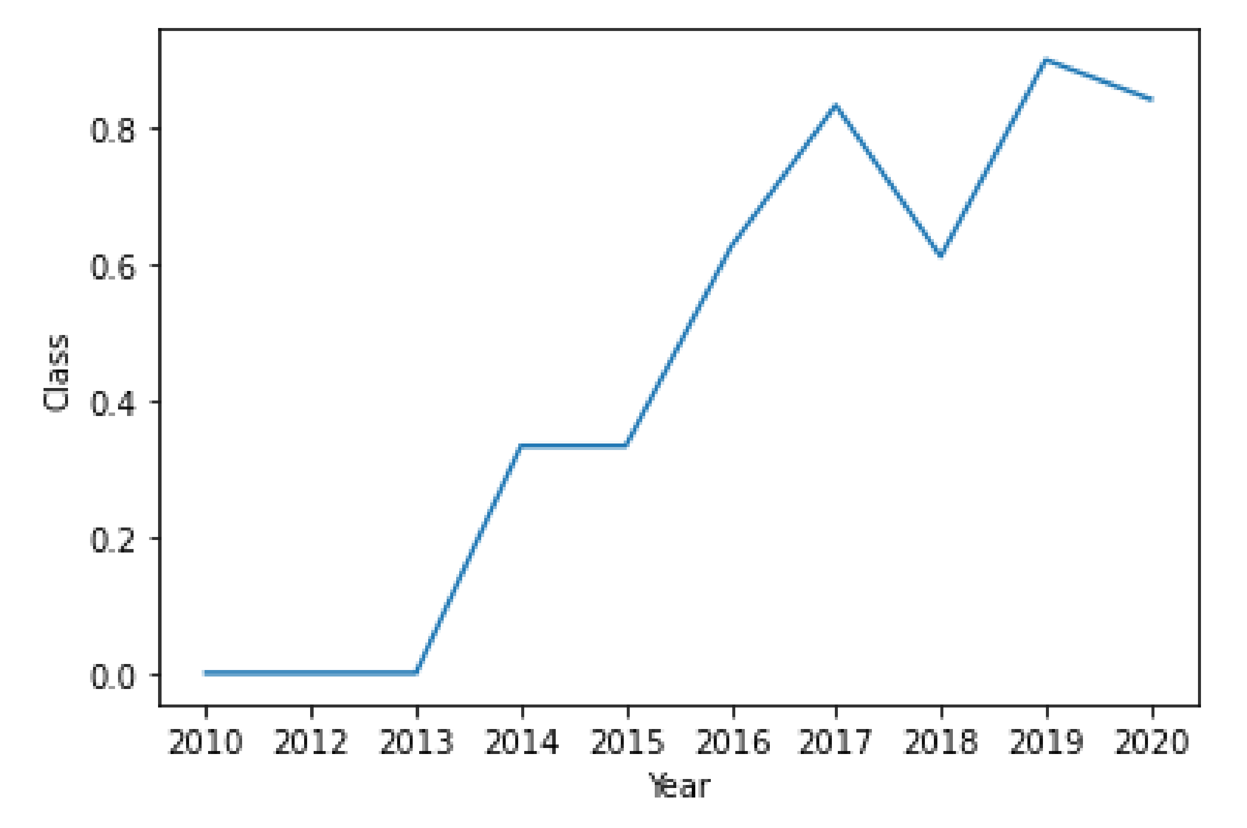


- It appears that heavier payloads have a negative influence on LEO and PO orbits, and a positive influence on GTO and ISS orbits.

- However – I would not conclude that there is a strong causative relationship here, because I suspect this trend is much more likely due to Payloads increasing over time. Also, with such low numbers in each orbit type, an element of chance may be playing a dominant role.

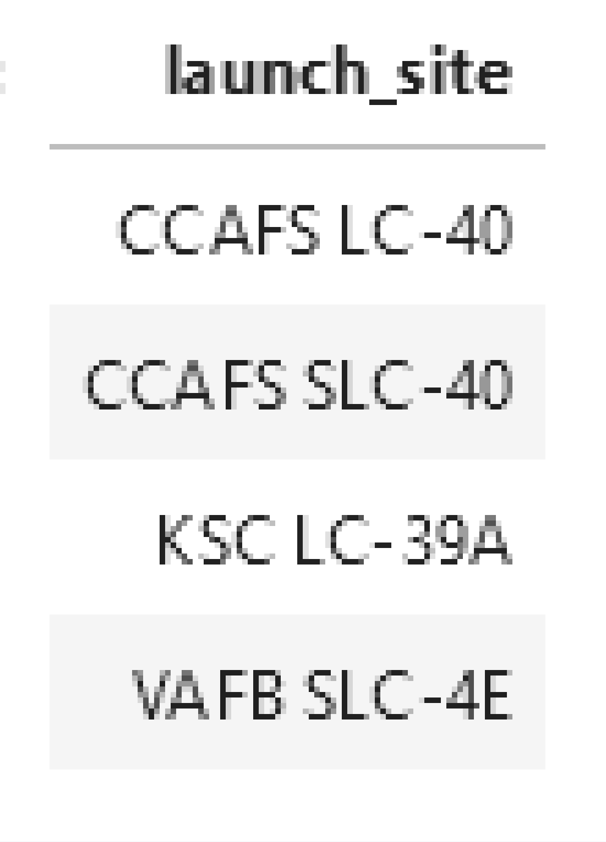
Launch Success Yearly Trend

- This appears to be the main predictor for success! Success rate improves dramatically with year of launch.
- This is most likely due to the SpaceX team improving their equipment, software, and techniques to optimise landing success.



All Launch Site Names

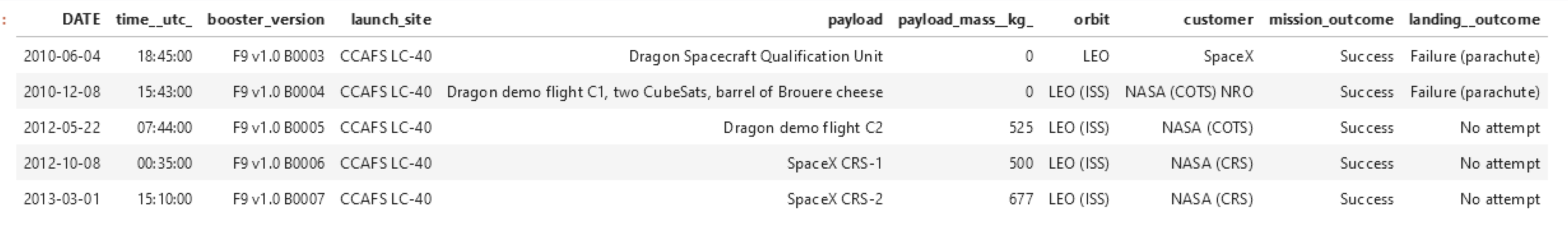
- Unique launch sites are listed at right
- There are four in total
- Later slides show where these are located geographically
- VAFB is on the USA West Coast
- The other 3 are quite close to each other on the East Coast



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

Launch Site Names Begin with 'CCA'

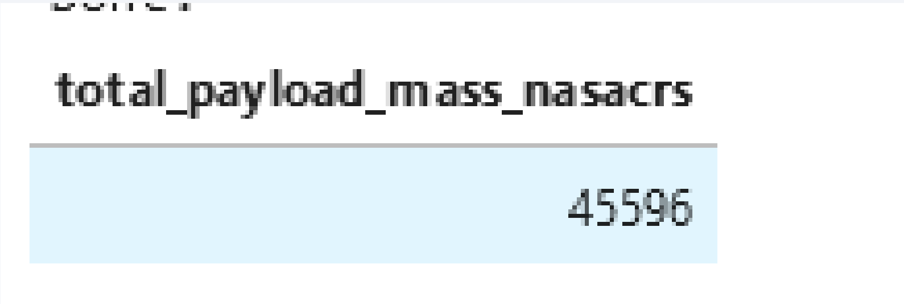
- Below shows 5 records where launch sites begin with `CCA`
- Columns include the date, booster version, launch site, payload, payload mass, orbit, customer, mission outcome, and landing outcome.
- Note that NASA are the customer for 4 of the 5 launches.



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

Total Payload Mass

- Total payload carried by boosters where client was NASA (CRS) is 45596 kg
- This is just a sum of all the payload carried from all NASA (CRS) flights

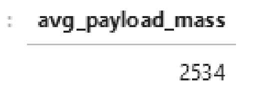


total_payload_mass_nasacrs
45596

Average Payload Mass by F9 v1.1

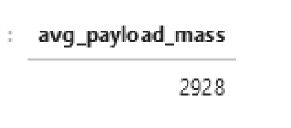
- The average payload mass carried by booster version F9 v1.1 is shown below.
- PLEASE NOTE: I wasn't sure whether the question was asking "Booster versions entered as ONLY 'F9 v1.1'", or "Booster versions which START WITH 'F9 v1.1'" (eg. F9 v1.1 B00005, F9 v1.1 B00004, etc).
- Therefore I did queries for both:

Boosters STARTING with F9 v1.1:



avg_payload_mass
2534

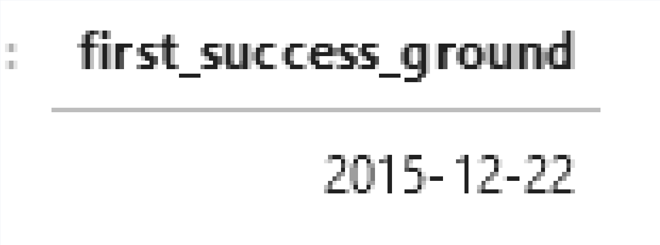
Boosters ONLY entered as F9 v1.1:



avg_payload_mass
2928

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad is shown below
- This is 22-Dec-2015 as shown below.

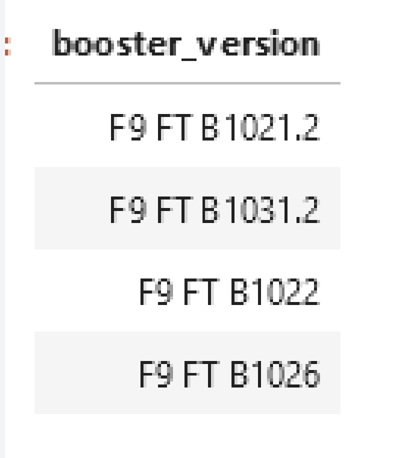


first_success_ground
2015-12-22

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 shown below

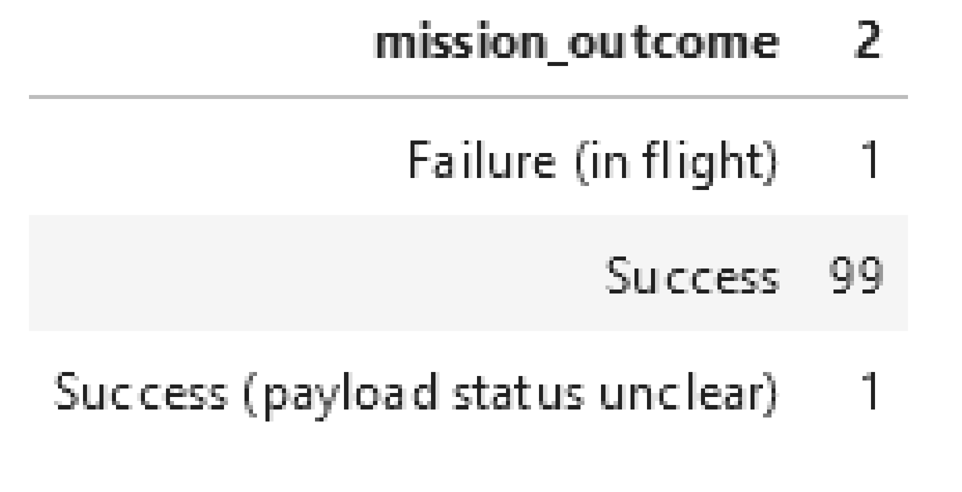
- Note they are all the “F9 FT” booster version. This version may be specialised to land on drone ships.



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

Total Number of Successful and Failure Mission Outcomes

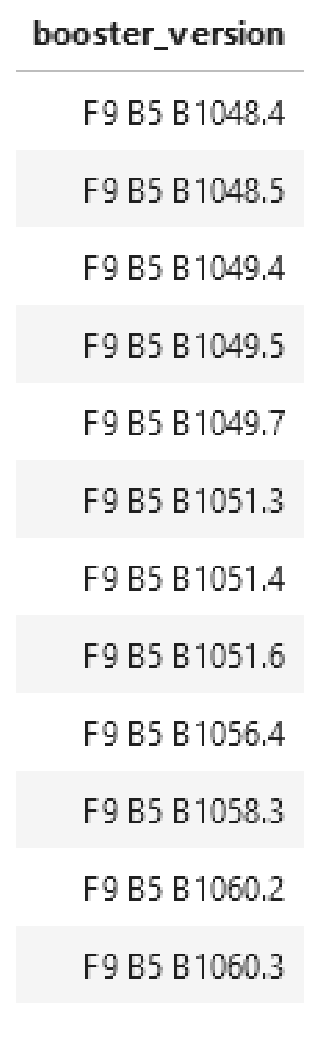
- The total number of successful and failure mission outcomes is shown below
- **This shows 100 successes in total, and 1 failure**



mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are listed at right
- Note that they are all the same parent booster version, ie. The F9 **B5**, just different variants.
- This suggests that the F9 B5 is the heavy-lifter booster version.

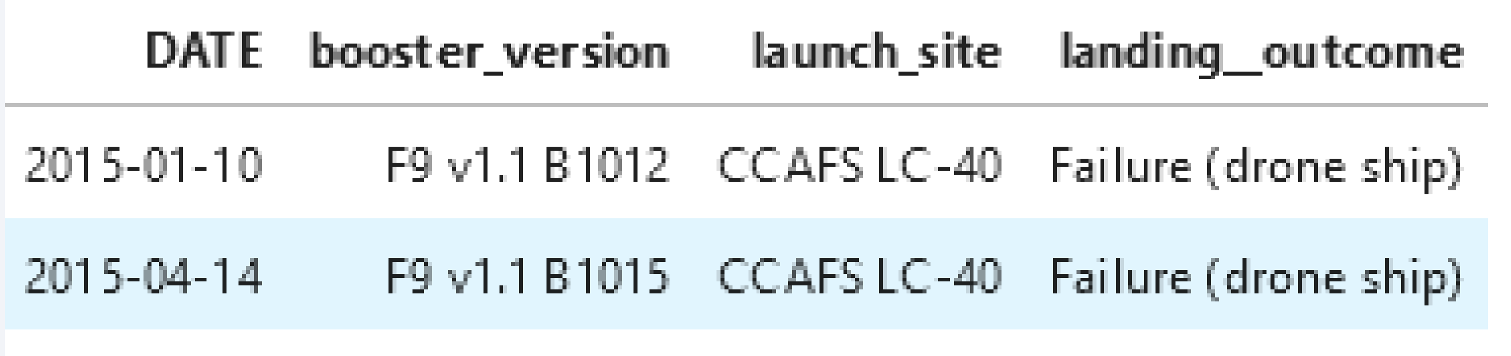


booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are shown below

- Both failures were from the same launch site, with the same parent booster version F9 v1.1 (though slightly different variants). They were also quite close in date, being only 3 months apart.

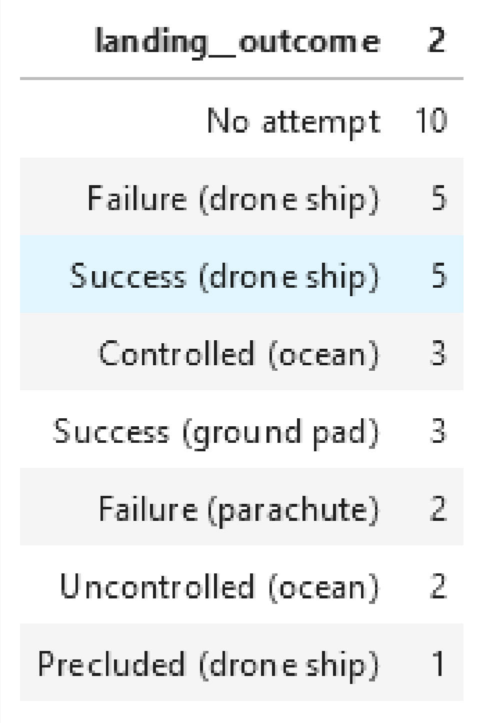


DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Below query result shows 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

- We can see that there was a mix of failures and success

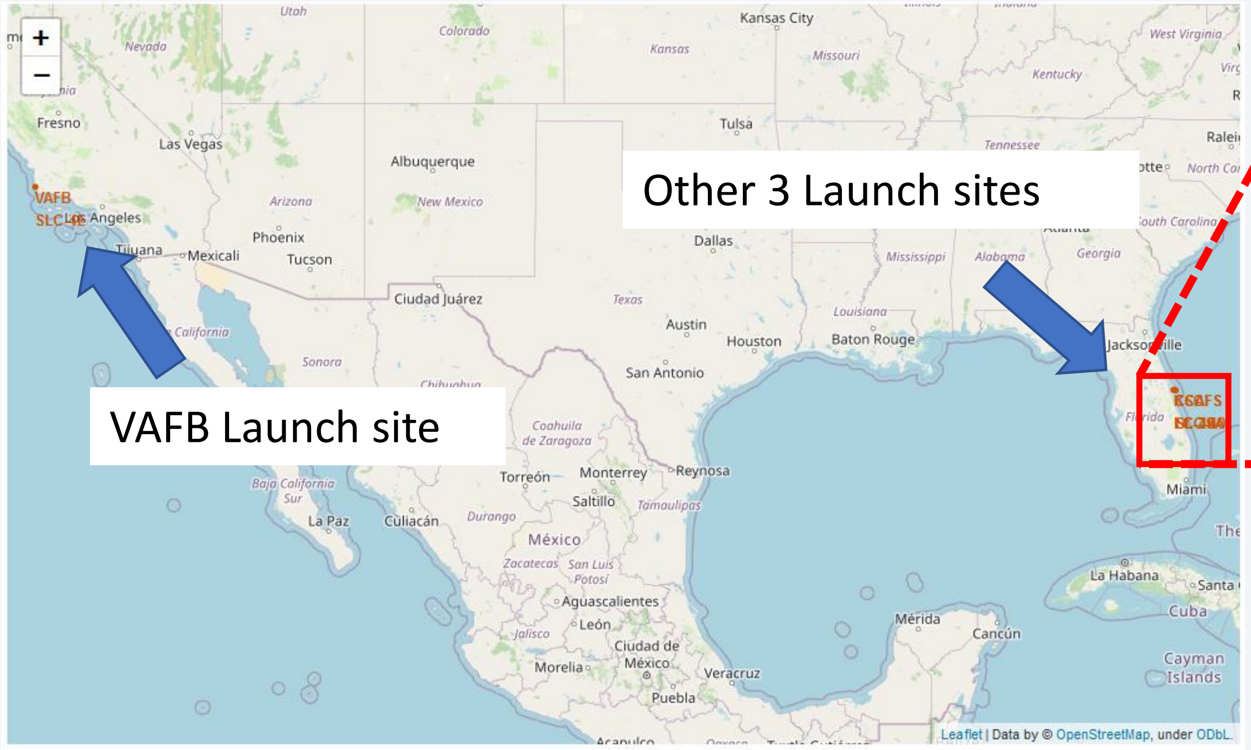


landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

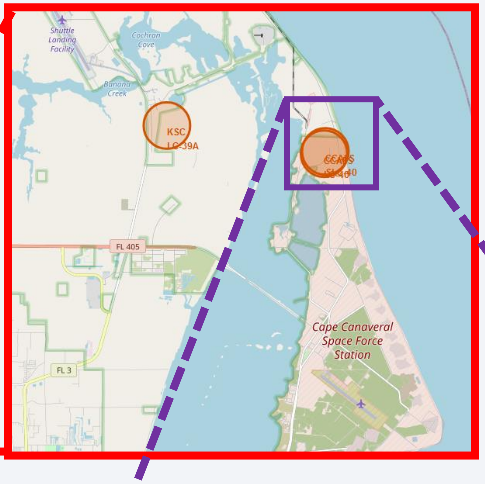
Launch Site Locations



Other 3 Launch sites

VAFB Launch site

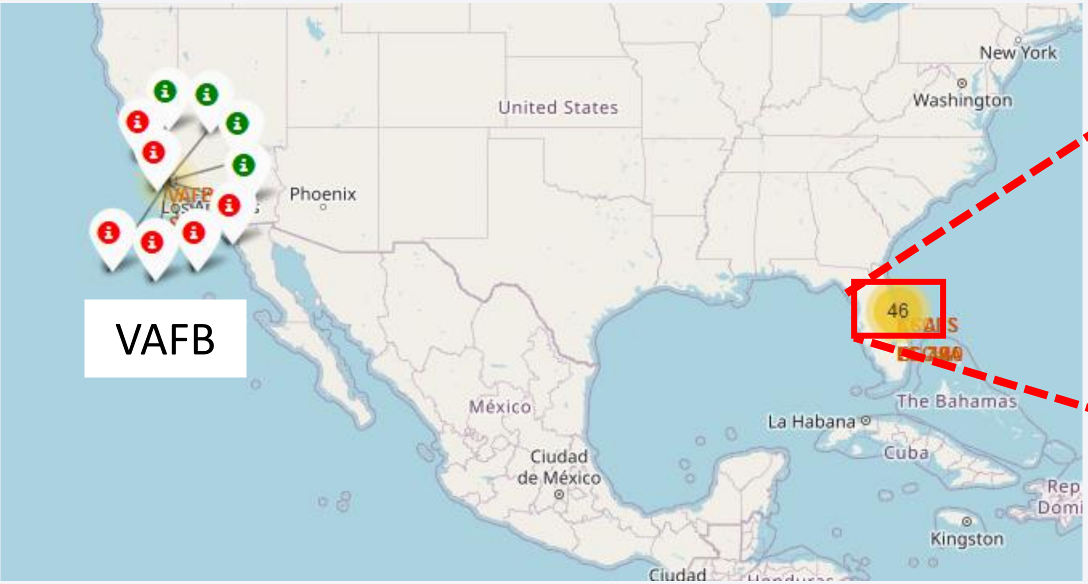
Leaflet | Data by © OpenStreetMap, under ODbL.

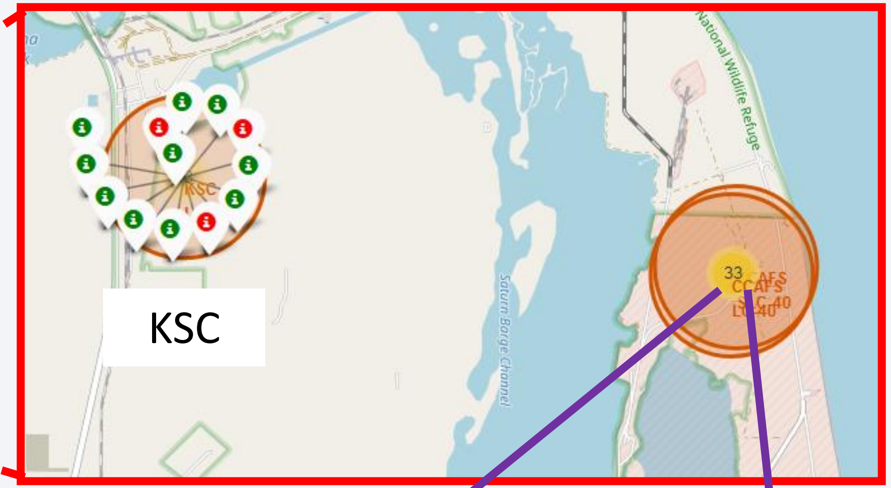


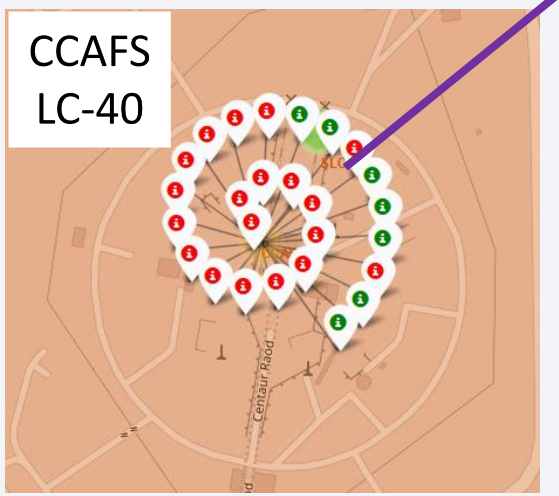
CCAFS
SLC-40
CCAFS
LC-40
Centaur Raod

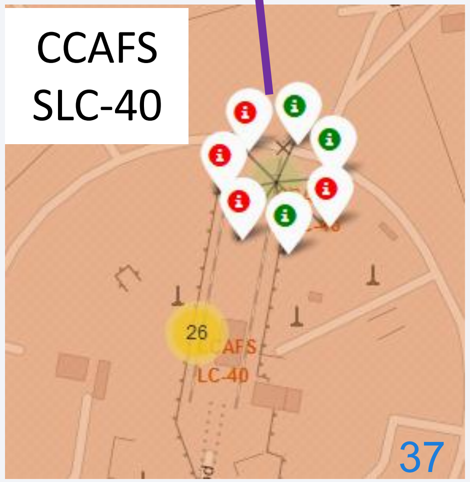
- Four Launch Sites in total
- VAFB is on the West Coast of the USA
- The other three are on the East Coast of the USA
- The two CCAFS LC-40 and CCAFS SLC-40 are at almost the same location

Landing Outcomes by Launch Site





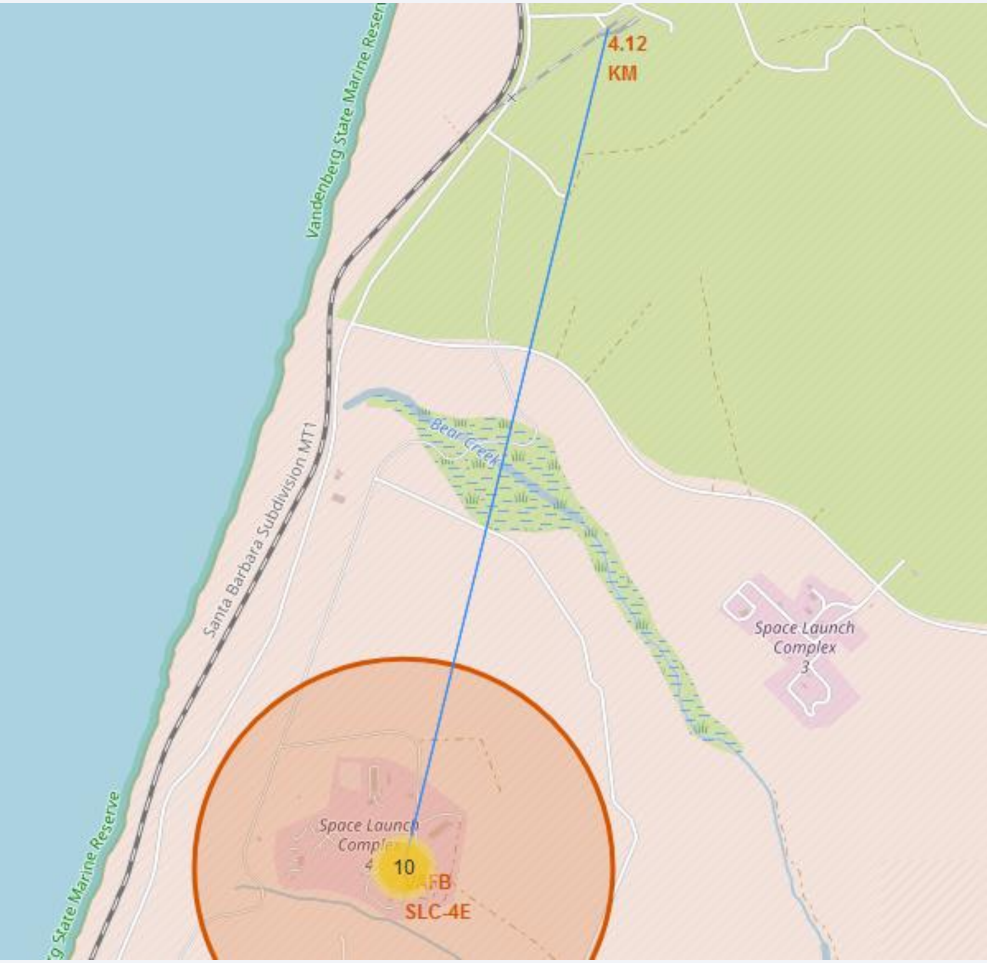




- Red dot = Failed landing, Green dot = Success Landing
- Many more launches at the East Coast launch sites
- **Highest landing success rates is at KSC**
- Much lower landing success rates at VAFB and CCAFS LC-40 & CCAFS LC-40

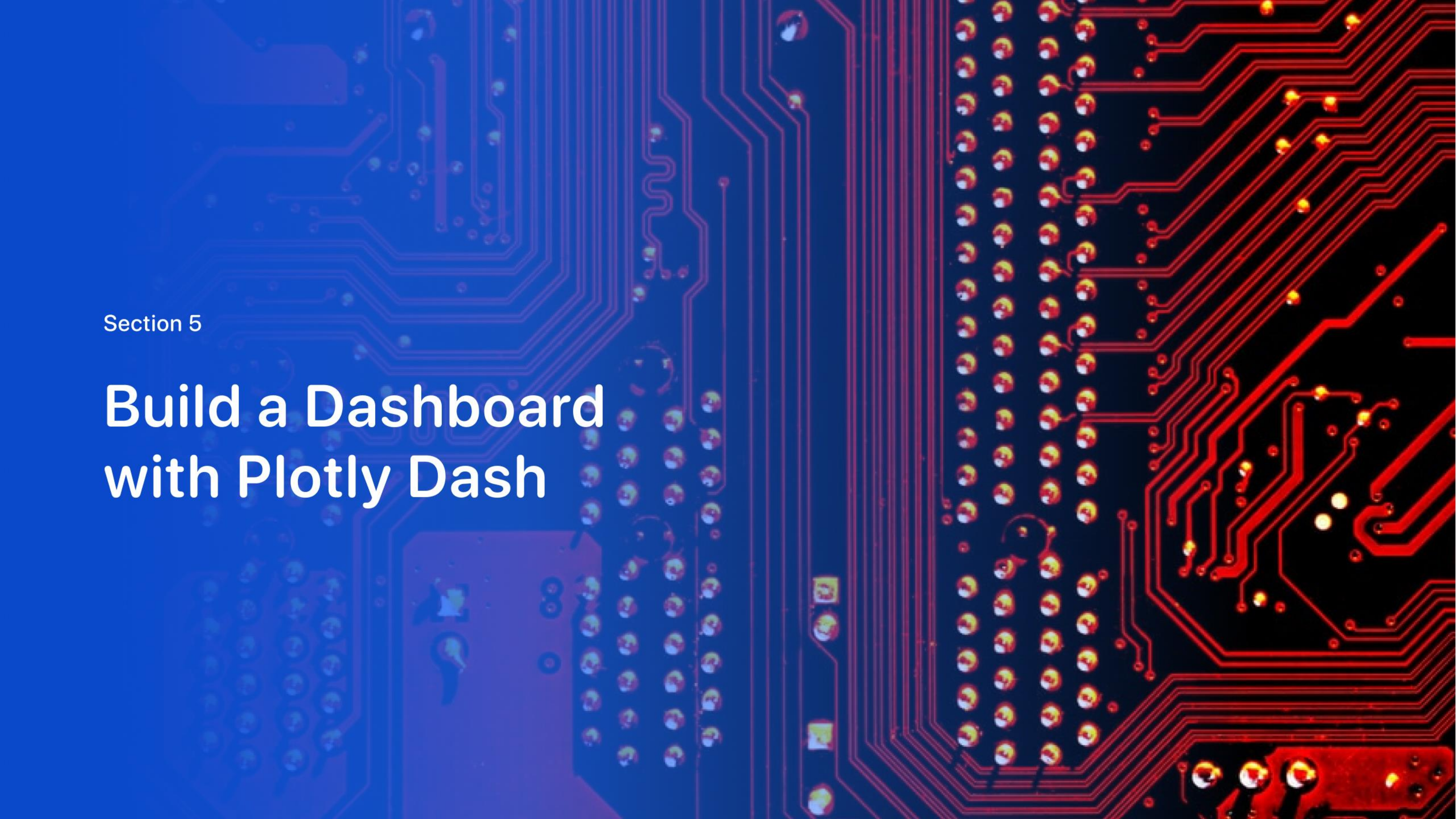
Launch Site Proximities

- VAFB Launch Site is within 4.12km of a rail unloading area
- Blue line on the map shows the distance from VAFB to rail terminus
- 4.12 KM is marked on the map


4.12
KM
Vandenberg State Marine Reserve
Bear Creek
Santa Barbara Subdivision MT1
Space Launch Complex 3
Space Launch Complex 4
10
VAFB
SLC-4E
Vandenberg State Marine Reserve

Section 5

Build a Dashboard with Plotly Dash

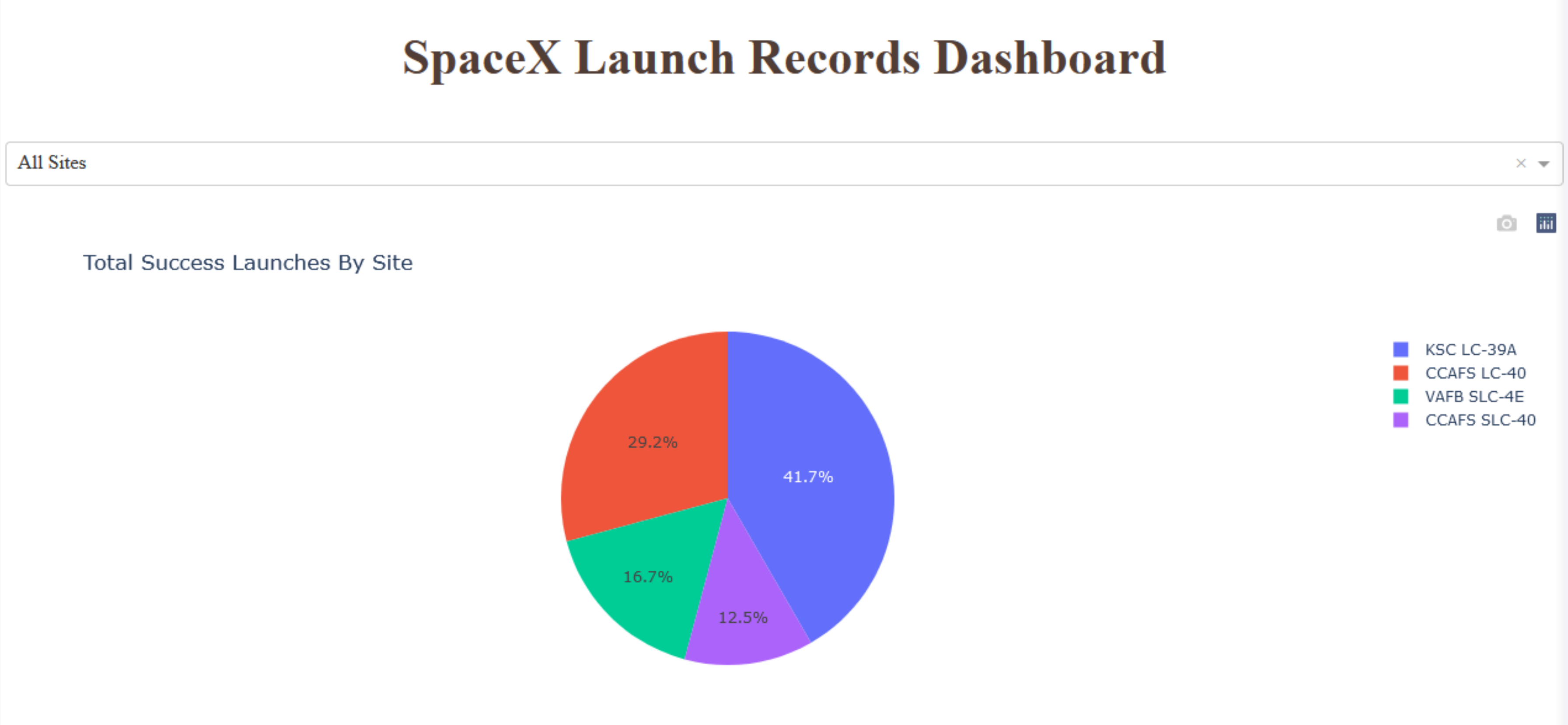


Insight note:

- This section appears to be a based on **<u>different target variable</u>** than the previous ones – it deals with LAUNCH OUTCOME, not LANDING OUTCOME.
  In other words, this section looks at: “Was the **mission** a success?”
  But previously we were looking at: “Did the First Stage **land** successfully?”
  Those are two very different things!

  The reason I think this is because the data in this section conflicts with earlier graphs.
  You can see the graph of Payload vs Launch success in this section, there is only one ‘success’ for payloads above 6000kg.
  Whereas the earlier scatter plots generated, showed many successes for payloads in this range.

Total Launch Successes by Site



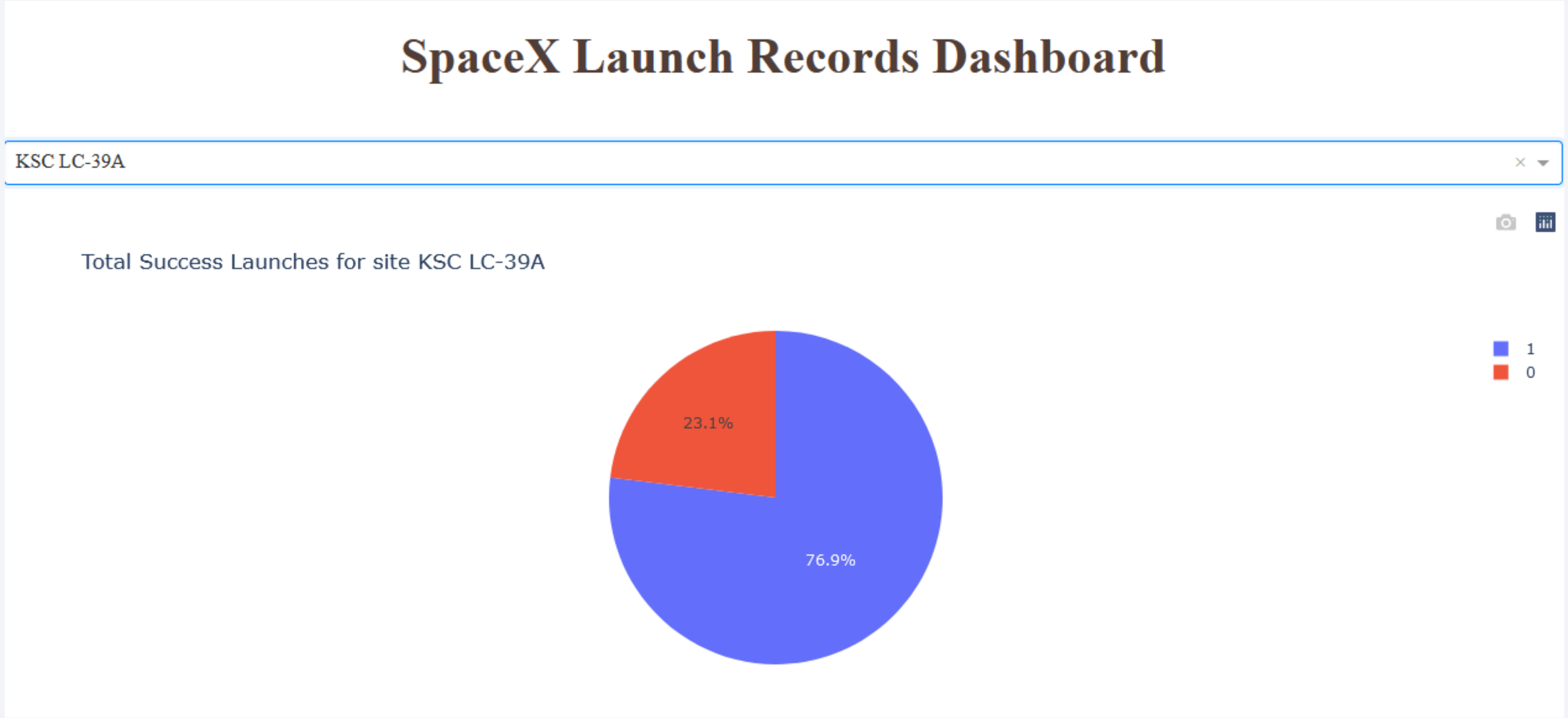
SpaceX Launch Records Dashboard

All Sites

Total Success Launches By Site

- KSC has the highest number of successes
- CCAFS LC-40 has the next-highest
- CCAFS SLC-40
- **INSIGHT: Note that this graph is not very useful, since it doesn't show success rate, only NUMBER of successes.**

Launch site with highest Success Ratio



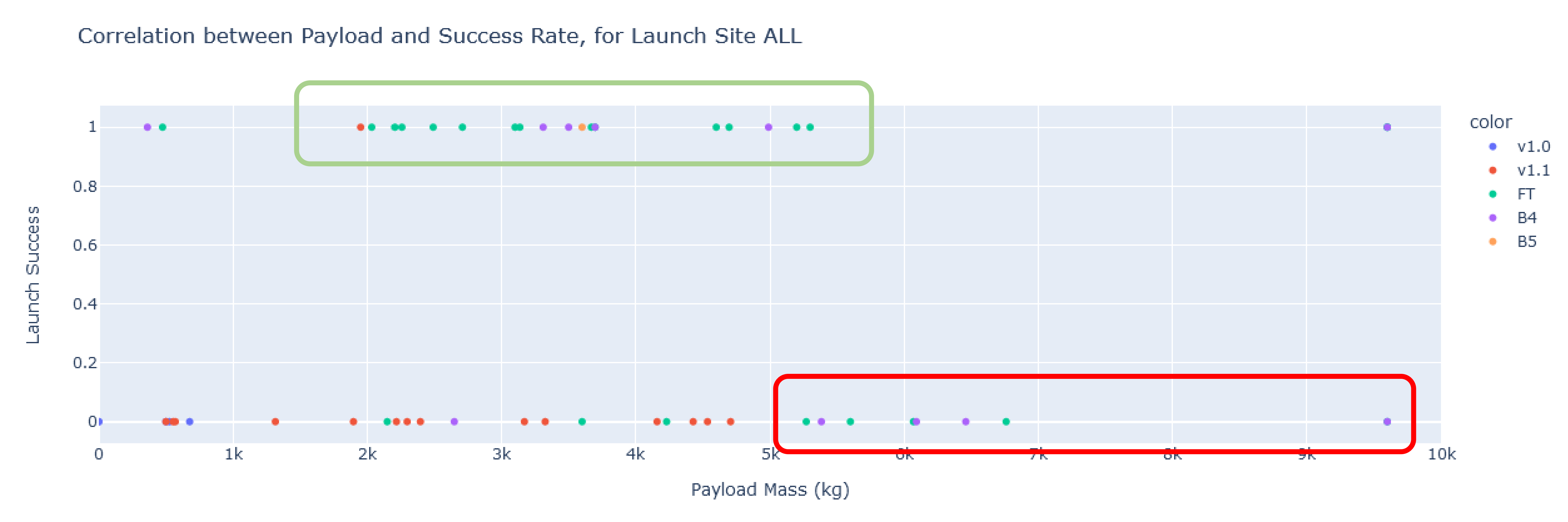
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

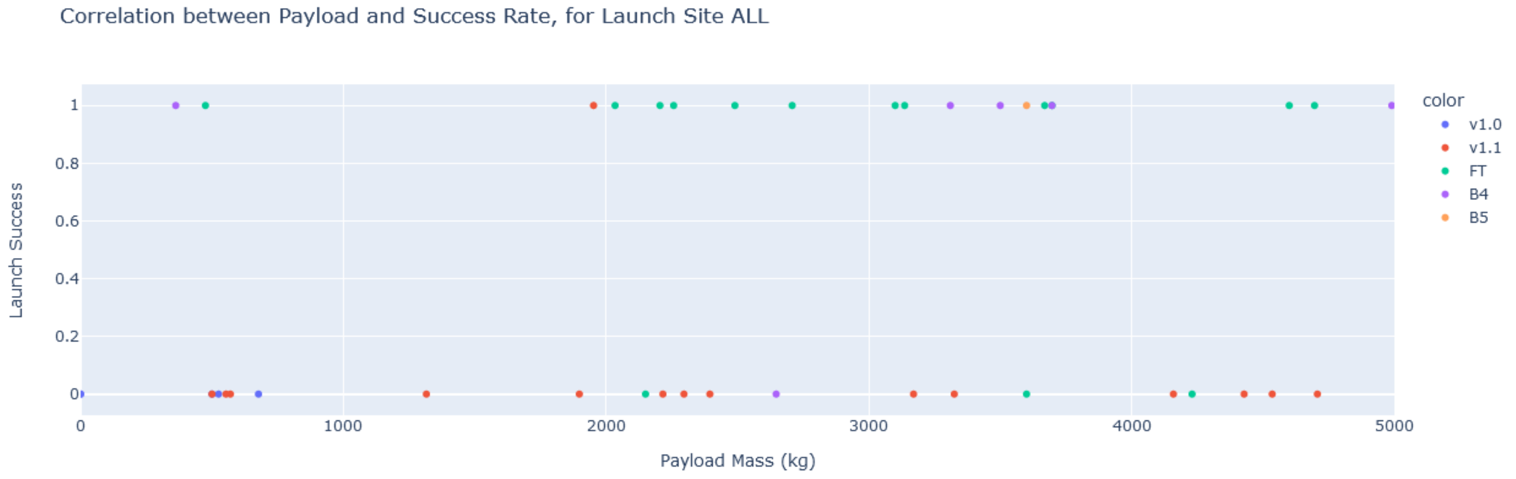
- When we look at all the Launch Sites individually, we make a key finding:
  KSC has the highest Launch Success Ratio, with 76.9% of its launches being successful

Payload vs Launch Outcome: Overview



- Higher payloads have lower launch success
- Booster versions FT and B4 have high success ratio
- Booster version v1.1 has very low success ratio!!
- Booster versions v1.0 and B5 do not have enough launches to conclude about success ratio

Payload vs Launch Outcome: Low vs High Payload



Range slider: Low Payloads

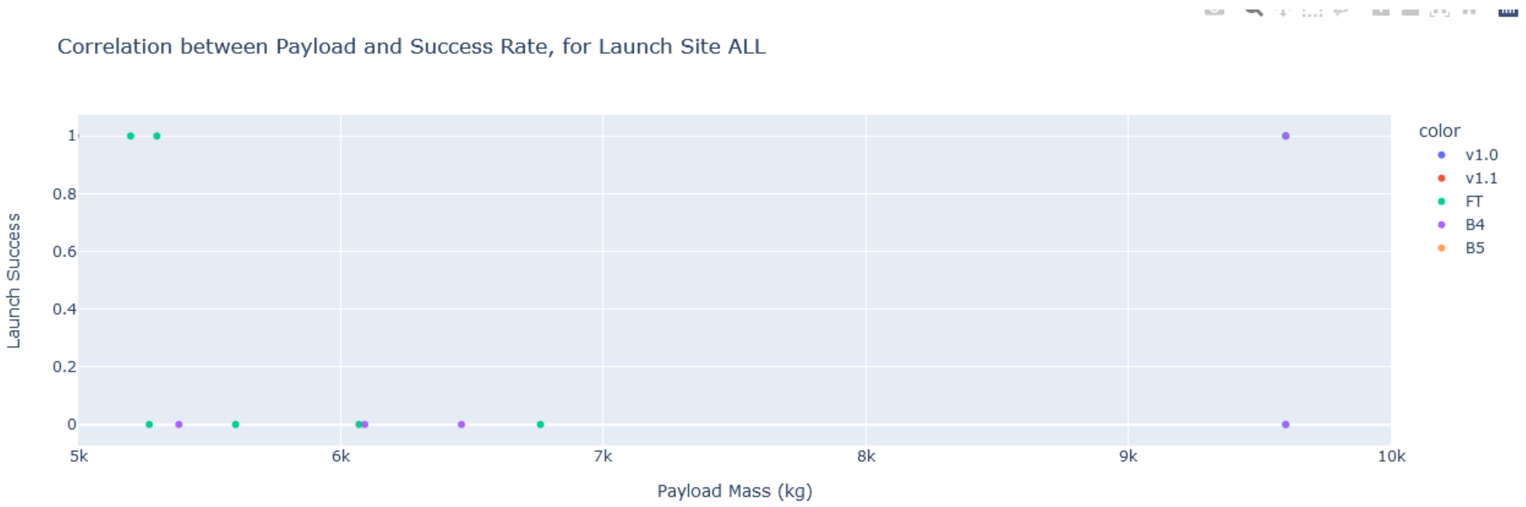
0 – 5000 kg

Best boosters: FT, B4

Range slider: High Payloads

5000 - 1000 kg

Very low success ratio



Section 6

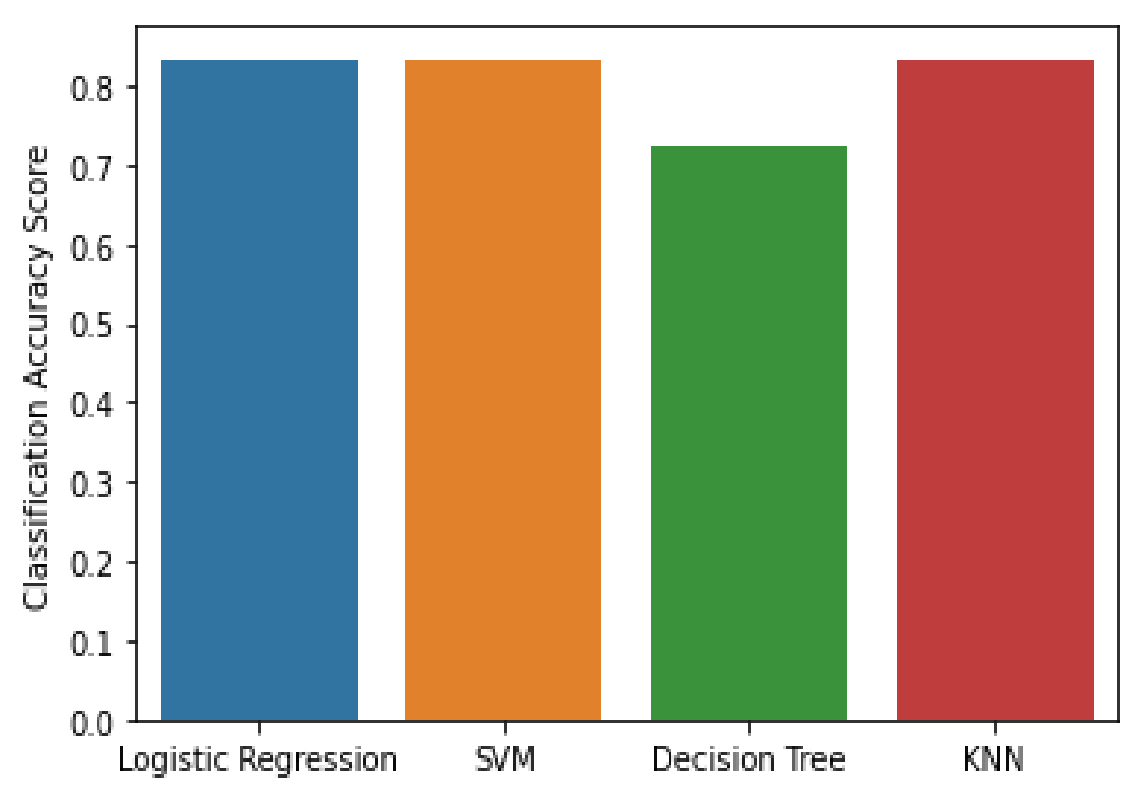
Predictive Analysis (Classification)

Classification Accuracy

- The bar chart at right shows the Classification Accuracy score of each model on the Test Set only (n=18).

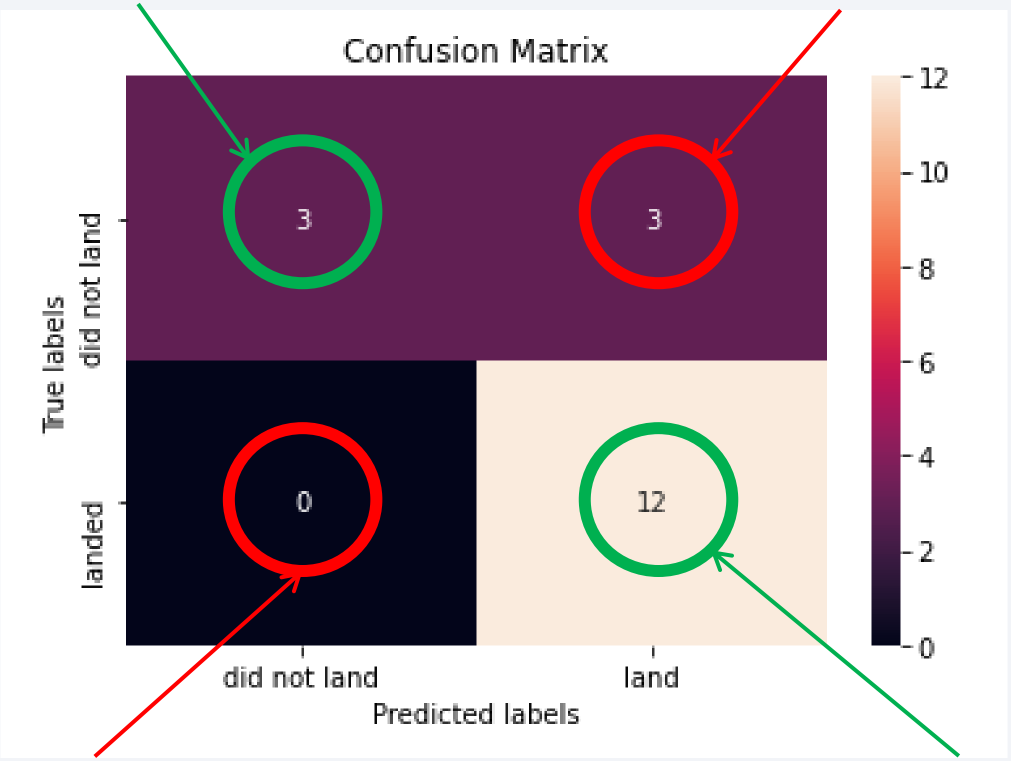
- Decision Tree performed the worst, all other models performed equivalent.

- Based on this I would select the **<u>Logistic Regression</u>** model to use, because it is able to output a probability along with classification prediction.



Confusion Matrix

- Confusion matrix of the selected model (Logistic Regression) is shown at right.
- Explanation is shown using text & arrows
- Out of the 18 records in the Test Set, the model **correctly predicted 83% of landing outcomes (15/18)**
- The model did not work well with true ‘did not land’ records, only predicting 50% (3/6) correctly.
- The model worked extremely well with true ‘land’ records, predicting 100% (12/12) correctly.



3 x Correct classification
“True Positive”

3 x Incorrect classification
“False Positive”

0 x Incorrect classification
“False Negative”

12 x Correct classification
“True Negative”

Conclusions

- **A 'Logistic Regression' classification model** has been built to predict the 'landing success' outcome from a given set of launch parameters.
- The model achieved **83% prediction accuracy** on the Test Set of 18 records, with the errors it made being 'False negatives' (ie. Model predicted landing success, but actually failed).
- EDA showed that **'Flight Number' is a strong predictor of success**, reflecting the learning & improvements that the SpaceX team made over time. Flight Number was included as a feature in the prediction model, hence the model will be leveraging this fact.
- The **predictive value of other factors is not clear** (Launch Site, Payload, Orbit, etc), because although there are apparent correlations, on closer inspection it was found that all of these factors changed over time also, and hence are 'confounded' by the strong relationship between time/Flight Number and success. We cannot conclude that they are causative.

Appendix

- Creativity to improve presentation beyond template:
  - Use of text, arrows, and coloured outlines to highlight areas of interest on graphs
  - Use of multiple-zoom levels for geographic maps, to show overview & detail on same slide

- Innovative Insights displayed:
  - Flight Number / time is the main predictor of success, and the other ‘apparent’ relationships are predominantly due to this
  - Logistic Regression model is the best model to move forward with, because it can output a probability / confidence, not just the predicted class

Thank you!