Analysis & modelling of SpaceX launch data

Stephen Plimmer

20 January, 2025

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Exploratory data analysis (SQL & trend)
  - Mapping analysis
  - Success rate analysis (via Dashboard)
  - Predictive model evaluation
- Discussion
  - Implications of launch analysis
  - Implications for future modelling
- Conclusions
- Appendices

EXECUTIVE SUMMARY

- To help SpaceY become competitive with Space X, Space X launch data from 2010-2023 was analysed, visualised and a predictive model was created:
- **Exploratory data analysis & data visualisation showed that…**
  - Success strongly depends on booster type and site, so these should be selected carefully based on our findings: The best booster choice varies by payload between B4, B5 and FT. (The success rate between the best and worst sites was a factor of 3)
  - Good success rates (>60%) can be achieved up to payloads of 6kkg, but not higher, suggesting payloads should remain below this value.
  - Space X success rates started low in 2010, but reached consistently high values (>60%) after about 20-30 launches, which took Space X 5-6years, suggesting a long term commitment is needed from SpaceY to optimise success rates.
- **Predictive modelling showed that…**
  - The choice of model type (out of logistical regression, decision trees, support vector machines and knn) mattered less than the optimal selection of parameters found from a grid search.
  - Good accuracy can be achieved on test data (0.83) using all models, but the results are prone to some false positives.
  - More data would make models more robust, particularly given the evolution of success over time and the small data set, which may render older data as less representative.

INTRODUCTION: Overview

- SpaceX have developed a competitive advantage In the commercial space market through offering lower cost launches compared to competitors.
- The advantage is largely gained through their ability to reuse and recycle first stage boosters.
- SpaceY would like to understand the success rate statistics of launches and the drivers of un/successful launches to find insights that can be used to develop strategies to increase competitiveness against Space X.
- Space Y has commissioned "SP" to carry out exploratory analysis and machine learning to visualize insights that help with their strategy, and create a predictive model that can predict future launch success based on the parameters of the launch.

INTRODUCTION: Objectives

There are several objectives from this project:

- Analyse launch data to find relevant insights
  - e.g. Relationship between key independent variables and launch success
- Provide a visualisation tool to share the insights
  - e.g. to show the impact of independent variables on launch success
- Provide launch data on mapping tool to show geographic insights
  - e.g. Proximity of launch sites to built up areas, transport sites etc
- Develop a predictive model for future launches
  - i.e. Evaluate different models to relate independent variables with launch success

METHODOLOGY: Overview

The methodology used follows much of the CRISP-data science methodology with the following stages:

1. **Business understanding:** Understand the types of insights that the client will find valuable.
2. **Data understanding:** Locate, collect and understand the data. Carry out preliminary exploration.
3. **Data preparation:** Handle null values and normalise the data prior to modelling.
4. **Modelling:** Identify & test different models and use the best to understand launch success
5. **Evaluate:** Evaluate model performance and the interpretability of results provided.

NB: Deployment: We did not deploy the model to a production environment in this project.

METHODOLOGY: 1. Business understanding

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage, unlike other commercial providers.

The first stage is most critical in the cost of the launch, being much larger than the second stage.

Several characteristics of launches influence the probability that a launch and landing are successful, which are therefore of most interest to SpaceY, e.g. launch site, booster type, payload.

Space Y's motivation is to understand SpaceX data so as to help inform SpaceY on how they can become competitive, with particular importance placed on the probability that first stages can be successfully landed and recovered.

METHODOLOGY: 2. Data understanding

Key data sets were used as below (See Annex A for sources). Data was explored using SQL and visualised using Folium (for mapping) and Dash (for relationships)

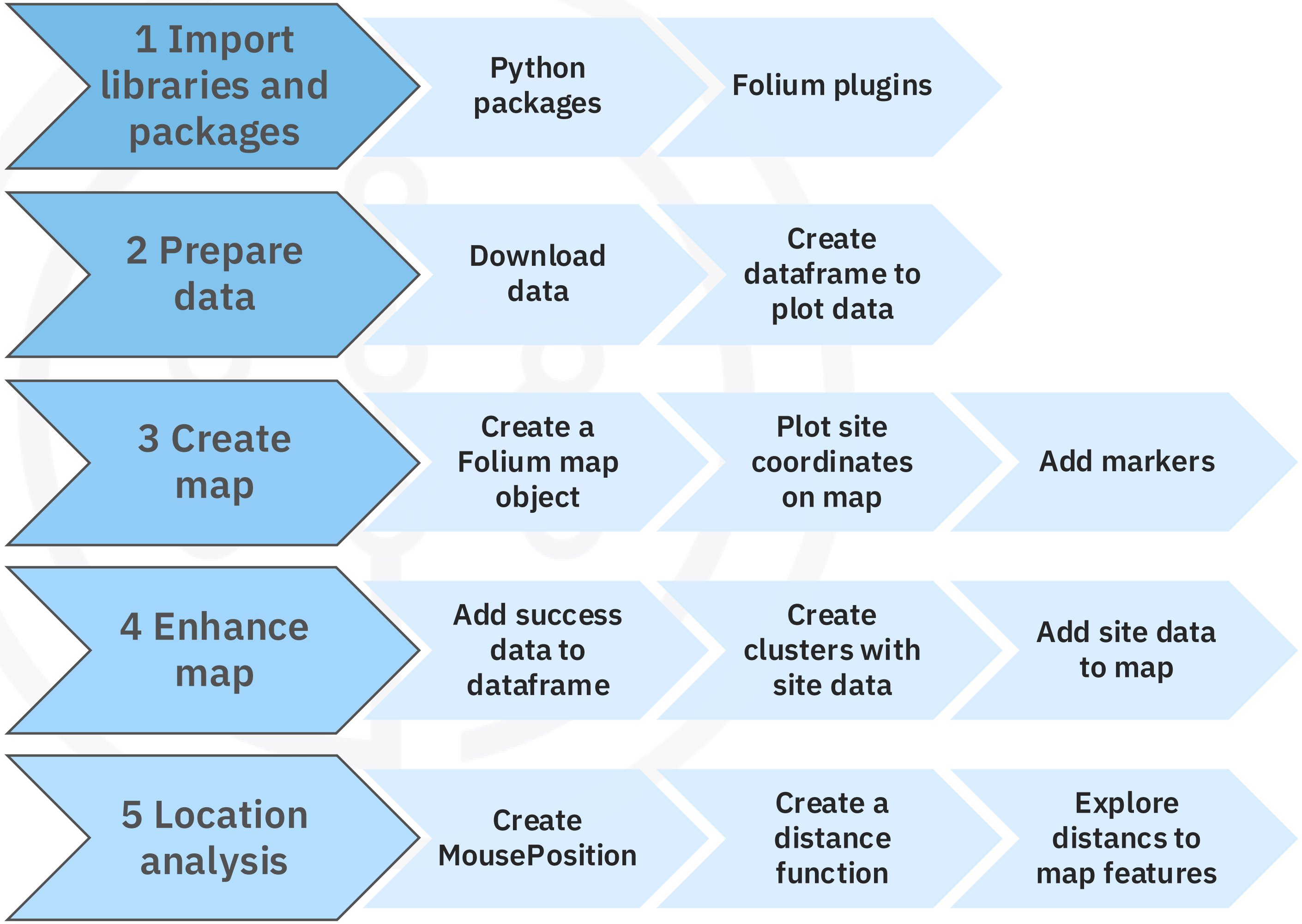
Data source type	API	Web scraping
Purpose	Launch data and associated characteristics by launch and site	Get data on historical launches from Wikipedia
Format	JSON	HTML
Contents	Data for each launch and site	Booster & site names, Payload data
#Records	90 launches x 17 variables	121 launches x 11 variables
Example variables	Mission date, Payload mass, Launch site, Locations, Reuse of stage 1 (Y/N), Mission success (Y/N).	Mission date, Payload mass, Launch date, Orbit, Launch outcome, Booster landing (Y/N)
Missing data	5x payload mass values – assumed as the average payload mass. (Some landing pad values were Null, but this represented that they were not used).	None

Data understanding *cont:* Visual analytics 1

Part of the EDA was to create a map using Folium to show **sites** and explore geography.

The main stages of this exercise are shown to the right.

Full code is available on Github.



1 Import libraries and packages → Python packages → Folium plugins

2 Prepare data → Download data → Create dataframe to plot data

3 Create map → Create a Folium map object → Plot site coordinates on map → Add markers

4 Enhance map → Add success data to dataframe → Create clusters with site data → Add site data to map

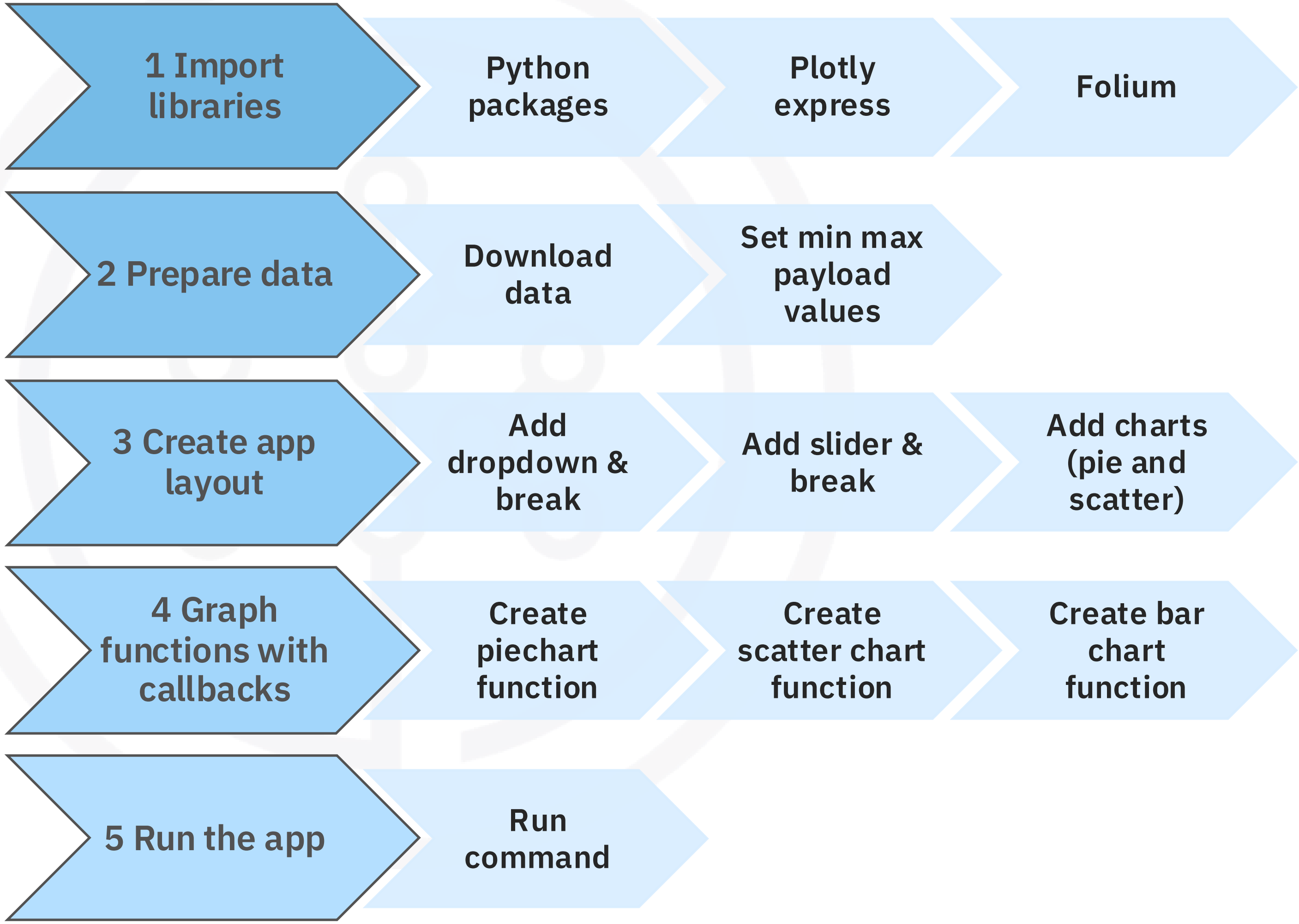
5 Location analysis → Create MousePosition → Create a distance function → Explore distancs to map features

Data understanding *cont:* Visual analytics 2

A second exercise was to create a dashboard using **Dash.**

The main stages of this exercise are shown to the right.

Full code is available on Github.



1 Import libraries → Python packages → Plotly express → Folium

2 Prepare data → Download data → Set min max payload values

3 Create app layout → Add dropdown & break → Add slider & break → Add charts (pie and scatter)

4 Graph functions with callbacks → Create piechart function → Create scatter chart function → Create bar chart function

5 Run the app → Run command

METHODOLOGY: 3. Data preparation/wrangling

Some steps were required to prepare the data for analysis and modelling, as shown below:

Extract	Assemble	Filter	Missing values	Web scrape	Extract/Export	Normalise
Use API to extract data from different columns covering booster version, launch site, payload and mission outcomes	Assemble JSON data from different columns into a new Pandas data frame, and look up values where an ID is given by the main APIs.	Filter the data to show only Falcon 9 results, that are of interest to this project. (Reset the ID in the dataframe)	Count missing values (payload data) and fill with the mean value from the other launches.	Gather further data on launches by using Beautiful Soup to webscrape data from the Wikipedia page on historic launches. (No missing values existed in this data set)	Relevant variables were extractred. Both API and Web scraping dataframes were exported to CSV files for evaluation and modelling.	All independent variables were normalised before modelling using StandardScalar() so that they had a mean of zero and standard deviation of 1.

METHODOLOGY: 4. Modelling

Several models were deployed to find the relationship between independent variables with launch success and to develop the best. In each case, the test data was 20% of the total data set and the GridSearch involved 10 cross-validation fits to the data.

Modelling parameters

Logistic regression	Support vector machines	Decision trees	K-nearest neighbours
parameters ={'C':[0.01,0.1,1,10,100], 'penalty':['l2'], 'solver':['lbfgs']	parameters = {'kernel':('linear', 'rbf','poly','rbf', 'sigmoid'), 'C': np.logspace(-3, 3, 5), 'gamma':np.logspace(-3, 3, 5)}	parameters = {'criterion': ['gini', 'entropy'], 'splitter': ['best', 'random'], 'max_depth': [2*n for n in range(1,10)], 'max_features': ['log2', 'sqrt'], 'min_samples_leaf': [1, 2, 4], 'min_samples_split': [2, 5, 10]}	parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'], 'p': [1,2]}

METHODOLOGY: 5. Evaluation

Several models were deployed to find the relationship between independent variables with launch success and to develop the best

The best models of each type were found through using the GridSearch method, to explore ranges of potential values for coefficients, and then the best overall model from the different types were found.

Models were ultimately compared using "Accuracy" versus the test data (See Annex A for definition of accuracy)

The nature of inaccuracies were shown using a confusion matrix

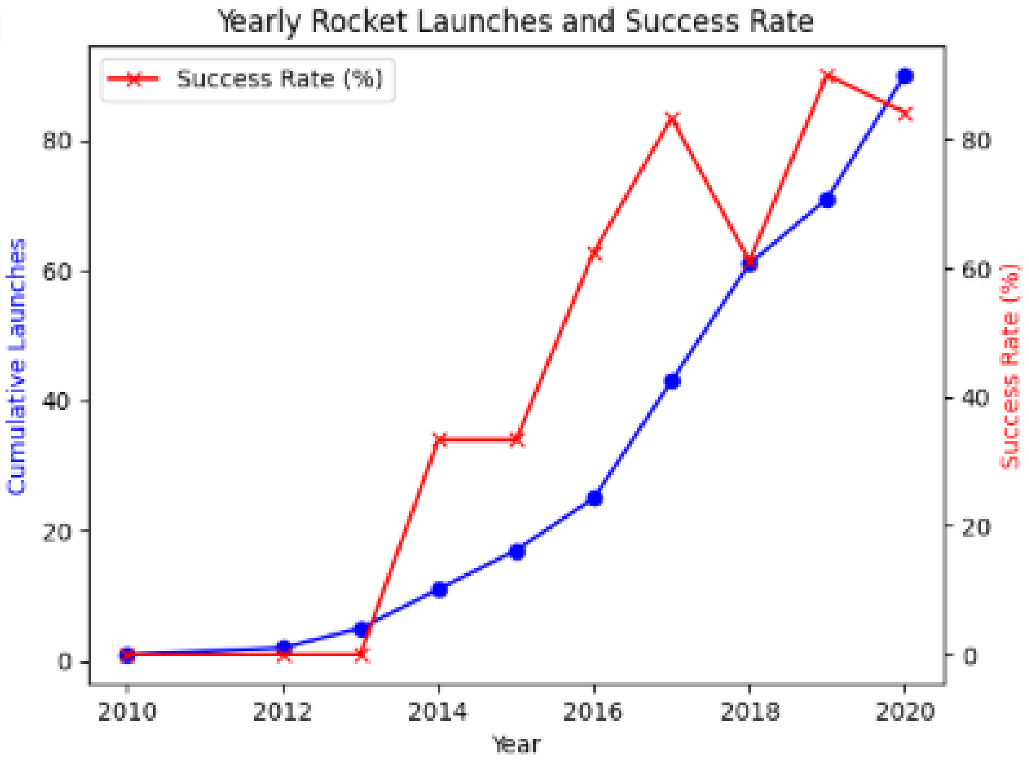
RESULTS: 1. Exploratory data analysis in SQL

Answers to specific SQL EDA queries that were raised:

Task	Question	Answer
1	Display the names of the unique launch sites	CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
2	Display 5 records where launch sites begin with the string 'CCA'	Thsese were the launches on the following dates: 2010-06-04, 2010-12-08, 2012-05-22, 2012-10-08, 2013-03-01. See Annex B for full record
3	Display the total payload mass carried by boosters launched by NASA (CRS)	45596
4	Display average payload mass carried by booster version F9 v1.1	2928.4
5	List date of the first successful landing outcome in ground pad	2015-12-22
6	List boosters which have success in drone ships & payload mass 4-6k kg	F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2
7	List total no. of successful & failure mission outcomes	100 (success) and 1 (fail)
8	List the names of the booster_versions which carried the max. payload mass.	F9 B5 B1048.4F9 B5 B1049.4F9 B5 B1051.3F9 B5 B1056.4F9 B5 B1048.5F9 B5 B1051.4F9 B5 B1049.5F9 B5 B1060.2F9 B5 B1058.3F9 B5 B1051.6F9 B5 B1060.3F9 B5 B1049.7
9	List the records with month names, failure landing_outcomes in drone ship ,booster versions, launch_site for months in year 2015.	These were the launches in Month 1 with F9 v1.1 B1012 and Month 4 with F9 v1.1 B1015. See Annex B for full record.
10	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.	The top value was Success (with 38 records), and No Attempt (with 21 records). See Annex B for full record

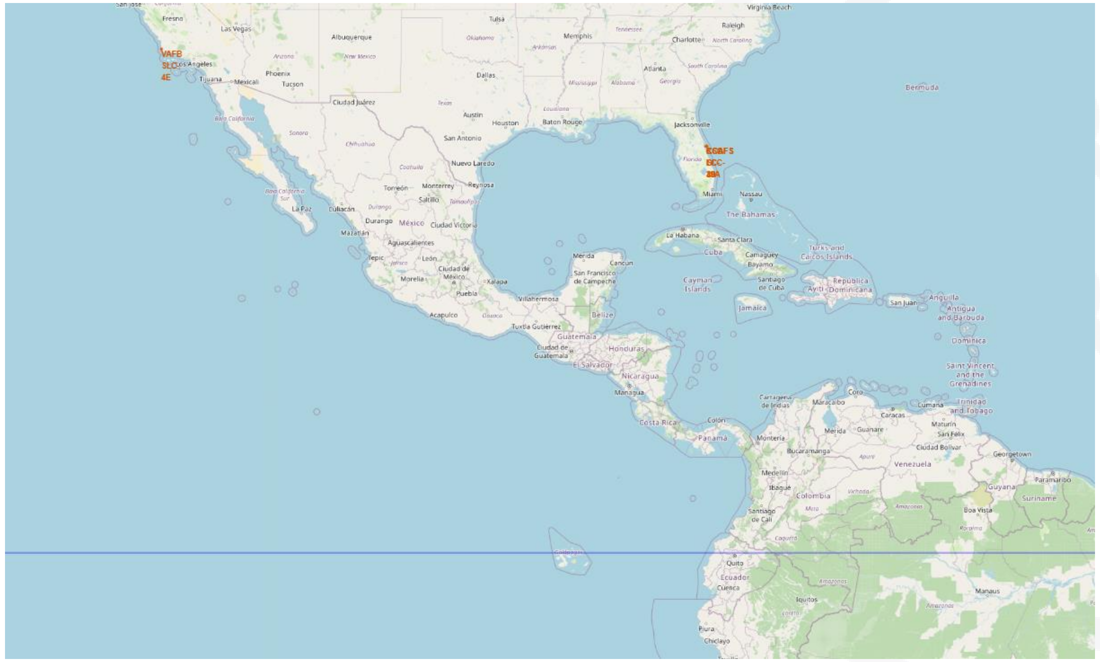
Exploratory data analysis *cont*

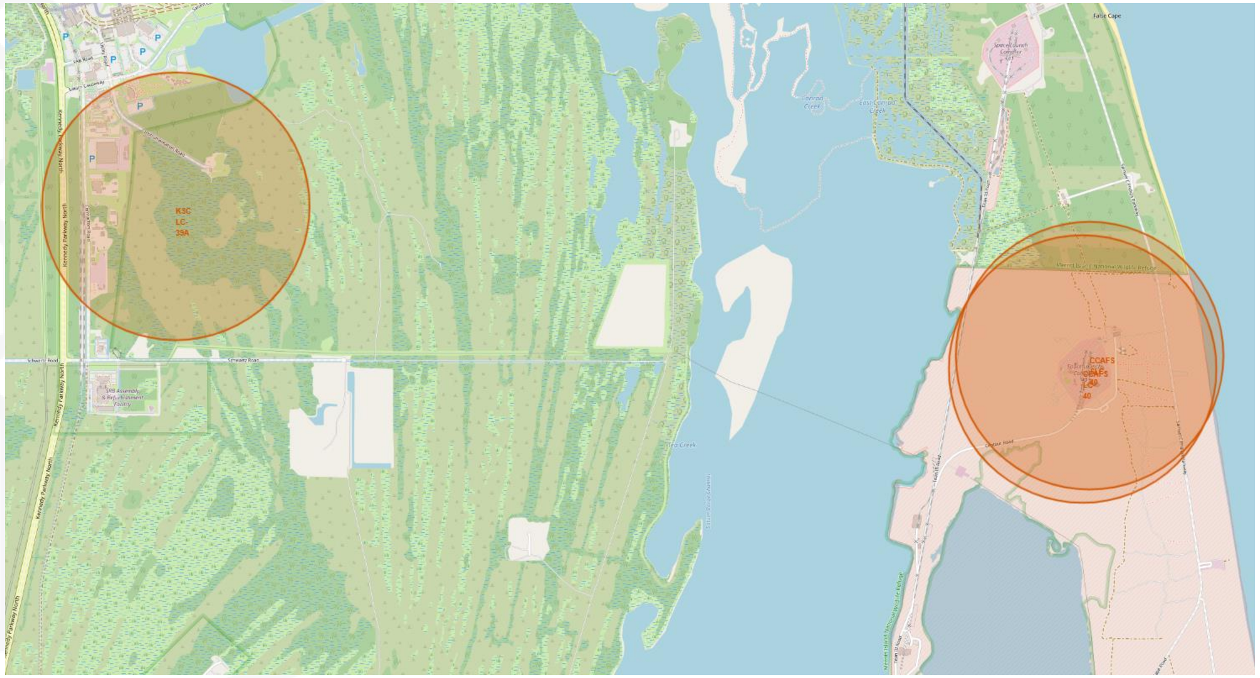
- Space X have increased their performance over time with increasing number of launches.
- It took over 6 years (2010-16) and about 20 launches to start to consistently achieve high levels of launch success (>60%).



RESULTS: 2. Mapping analysis

Bases were sited on East and West coasts (see plot left), well above the equator (blue line), with three East coast bases in very close proximity (see plot right).



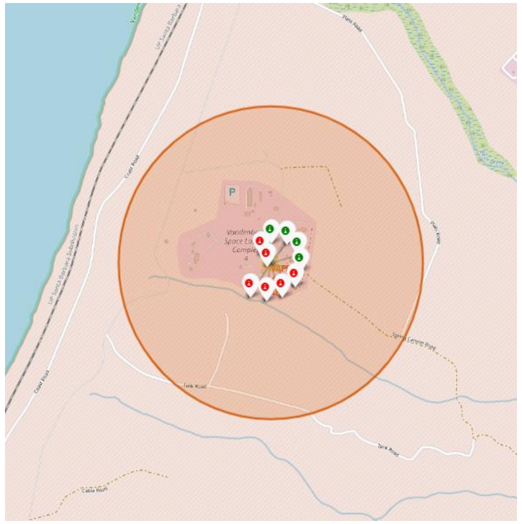


CCAFS LC-40 and CCAFS SLC-40 were both c0.5km from the coast, while KSC LC-39A was c7.5km away.

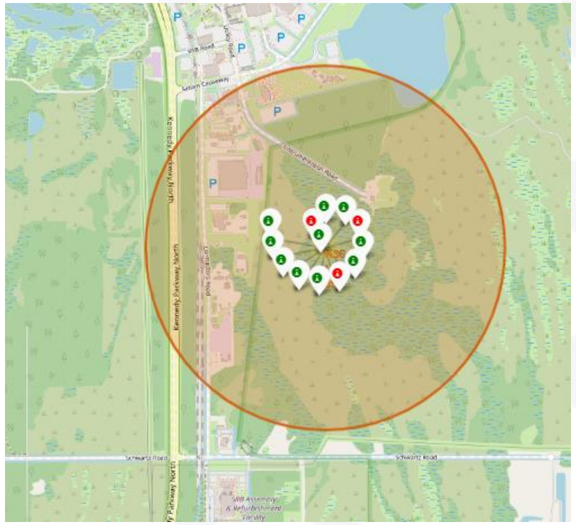
Mapping analysis *cont* – success by site

Red denotes fail, green denotes success.

VAFB SLC-4E

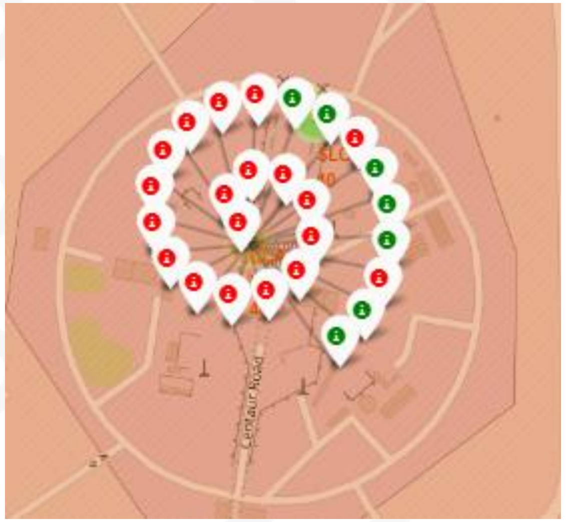


KSCLC-39A



Highest success rate

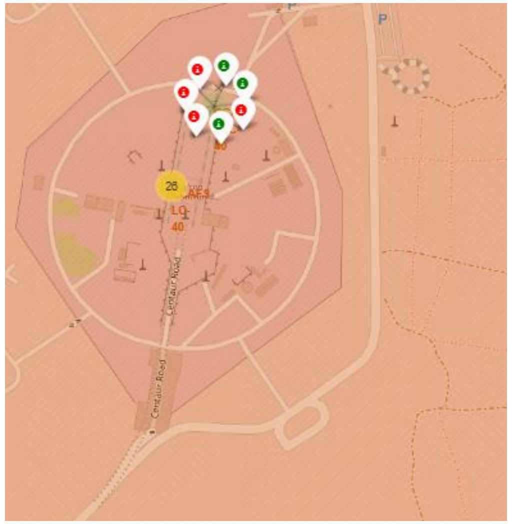
CCAFS-LC40



Lowest success rate

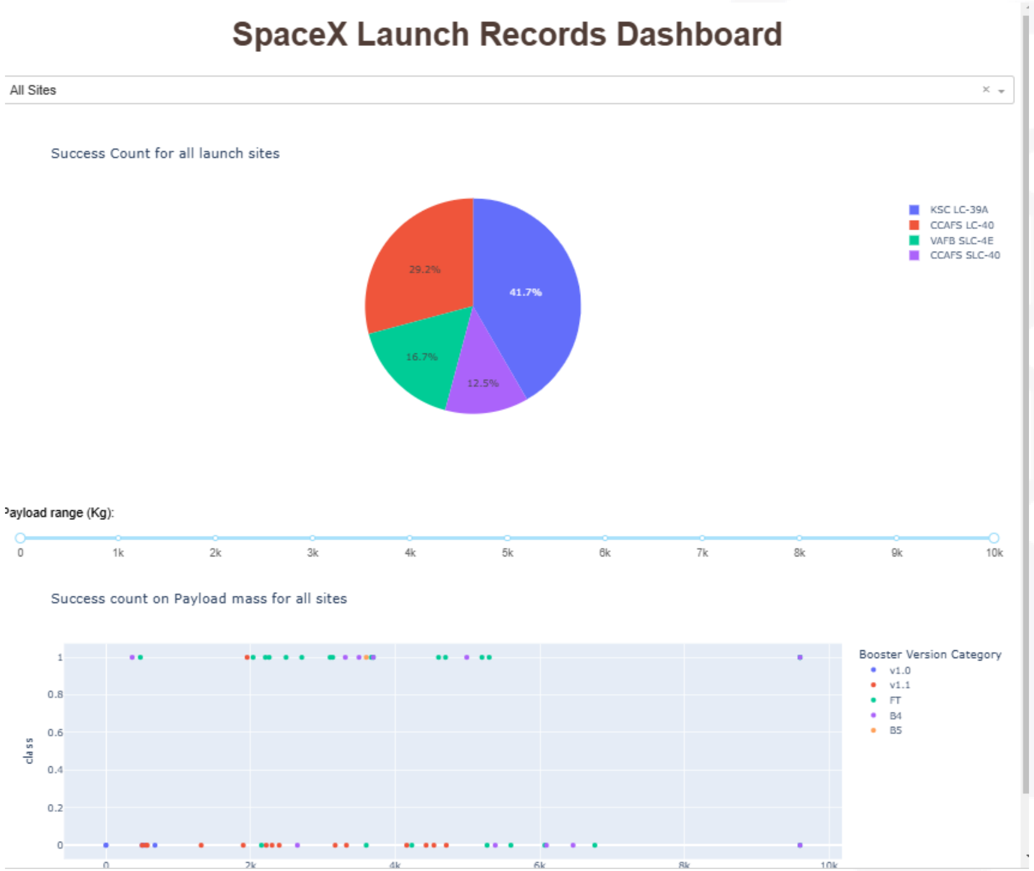
Most launches

CCAFS-SLC40



Fewest launches

RESULTS: 3. Visualisation



SpaceX Launch Records Dashboard

All Sites

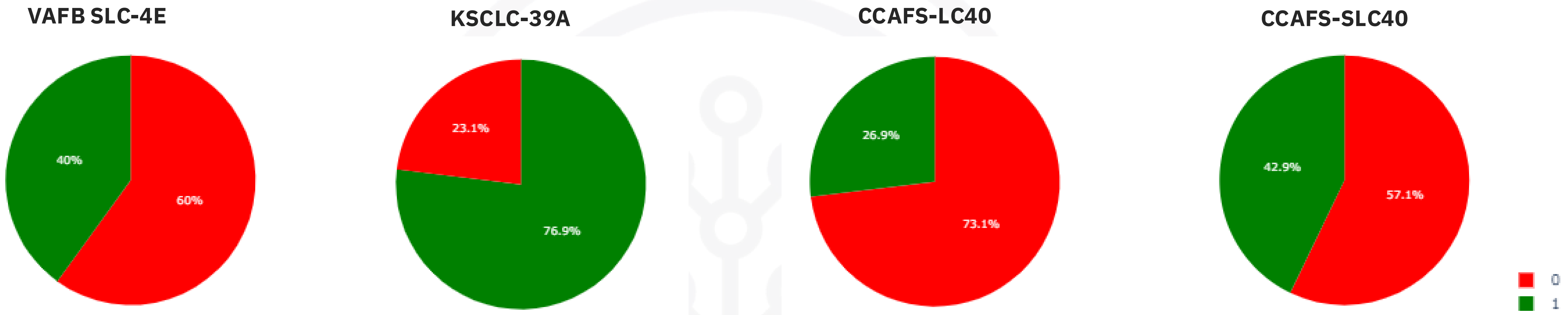
Success Count for all launch sites

Payload range (Kg):

Success count on Payload mass for all sites

- Important launch data was used to create a Dashboard (left), which allowed filtering for all or specific sites.
- The overall result (shown top left), which shows the distribution of successful launches.
- The highest volume of successful sites was from KSC LC-39A (41.9%).
- (A further plot was added to summarize the scatter chart results – see following slides)

RESULTS: 3. Success by site

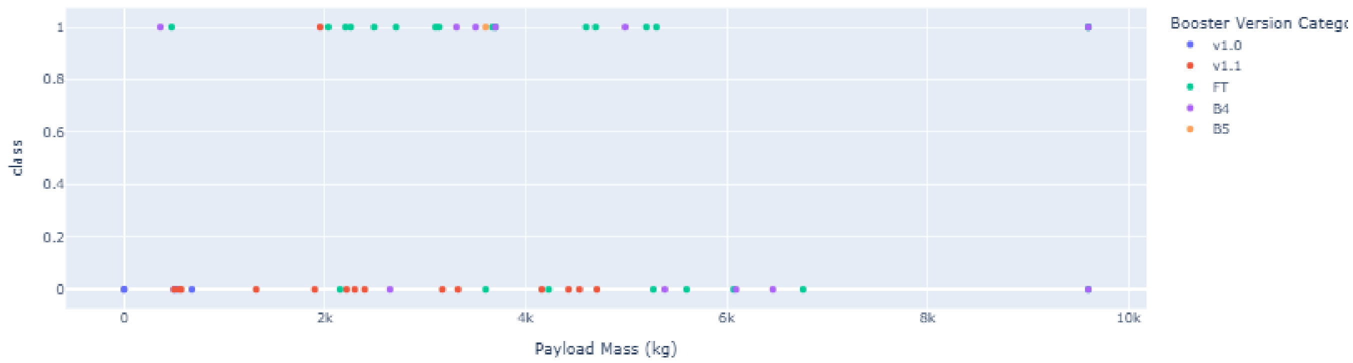


- (From the dashboard) we see that KSC LC-39A not only had the highest volume of successful sites, but the highest rate of success (76.9%).
- CCAFS-LC40 had the lowest success rate (26.9%).

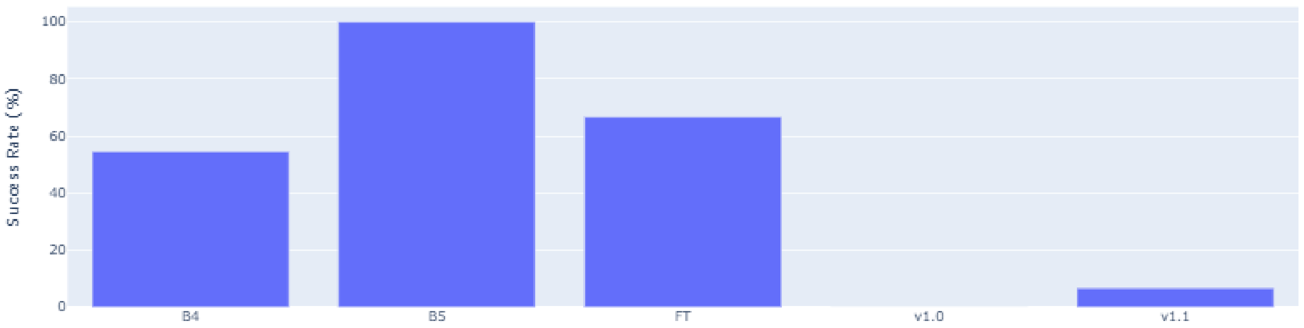
RESULTS: 3. Success by booster type

- We can now plot the success rates by booster type (right). The charts to the right show values for all payloads
- The scatter (top) chart shows that success also depends on payload.
- For all payloads, the summary bar chart (bottom) shows B5 has the highest success rate and v1.0 and v1.1 are zero and close to zero.
- The top scatter chart indicates most successes are between 2k and 6k payloads. (We look more closely in the next slide)

Success count on Payload mass for all sites

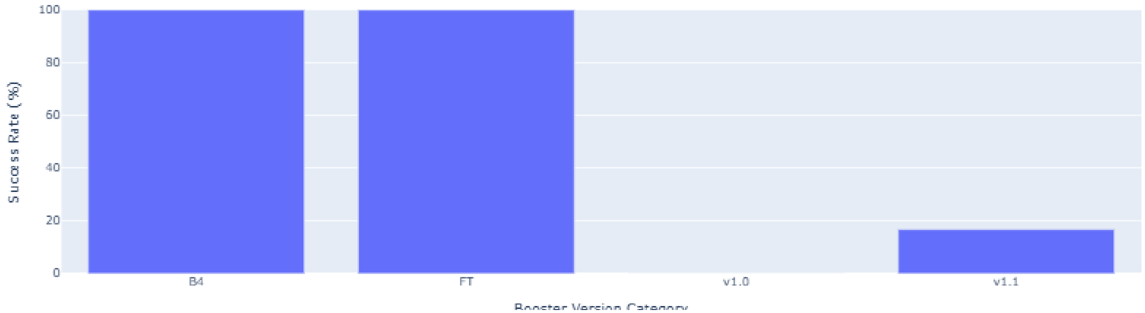


Success Rate by Booster Version Category



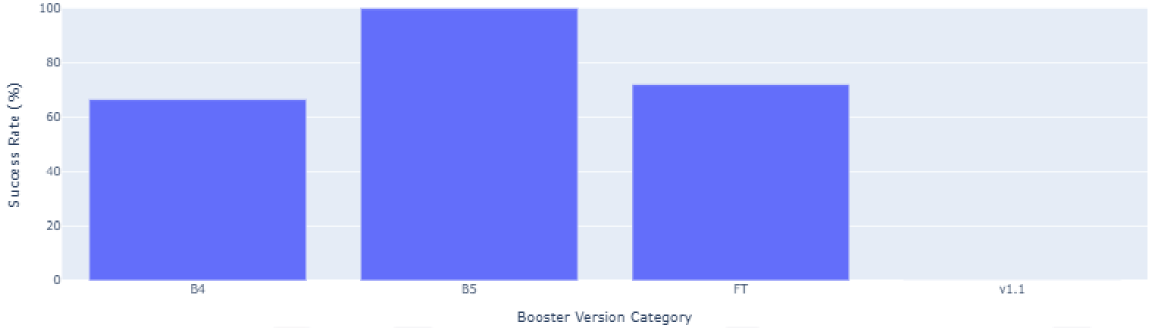
RESULTS: 3. Influence of payload & booster

Low payloads (<2k kg)



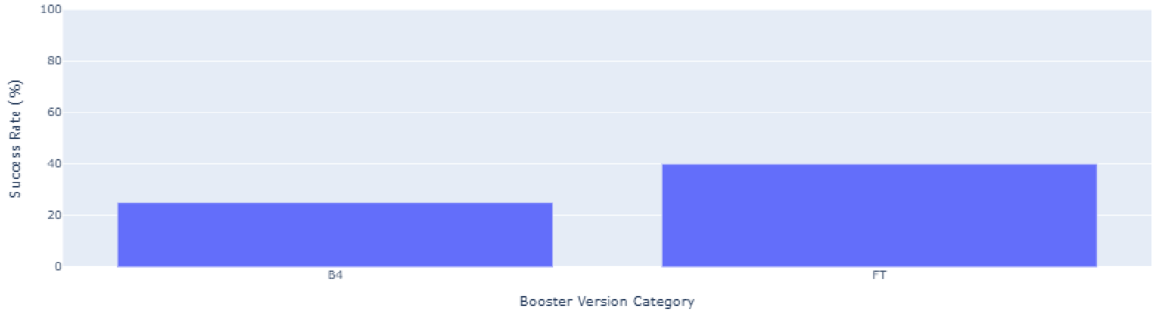
Both **B4 and FT** achieve 100% success.

Mid payloads (2k-6k kg)



B5 achieves 100% success, while FT and B4 success falls < 100%

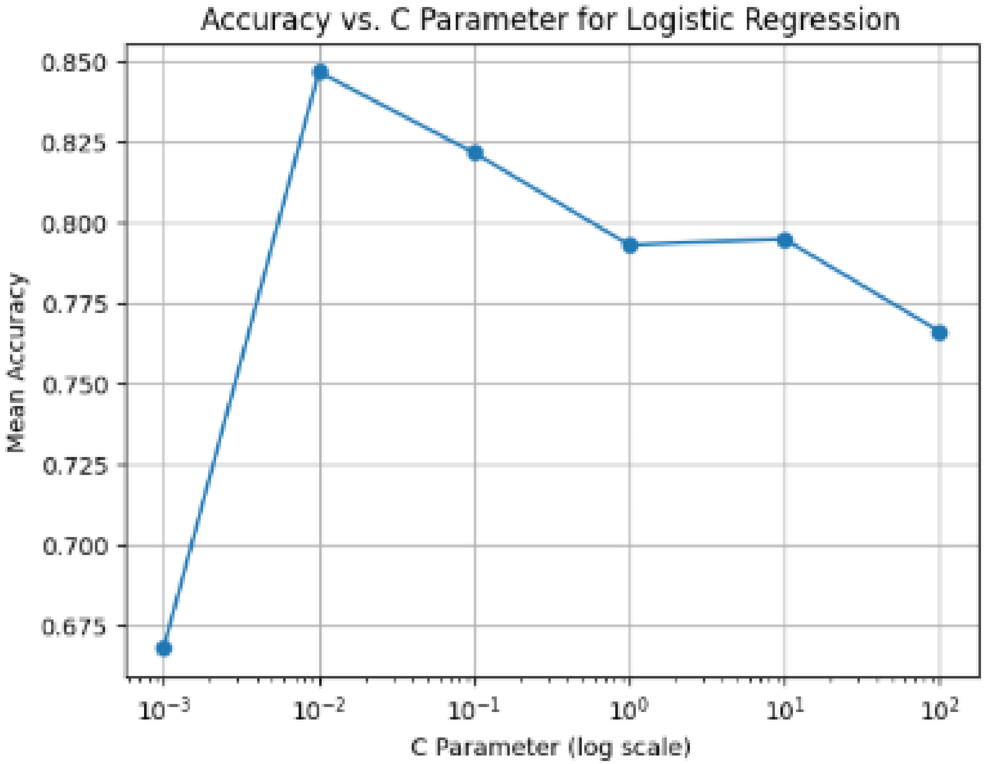
High payloads (>6k kg)



No boosters are successful, but **FT** offers the best chance

RESULTS: 4. Predictive model creation

- For each set of the parameters for the models, the accuracy value was calculated on training data
- An example is shown to the left where the C value of 0.01 (10^-2) gives the best fit, with accuracy approaching 0.85.
- The accuracy on the test data was then calculated using the best model....
- *(See Annex A for definitions of accuracy and C values and other parameters)*



RESULTS: 4. Model parameter results

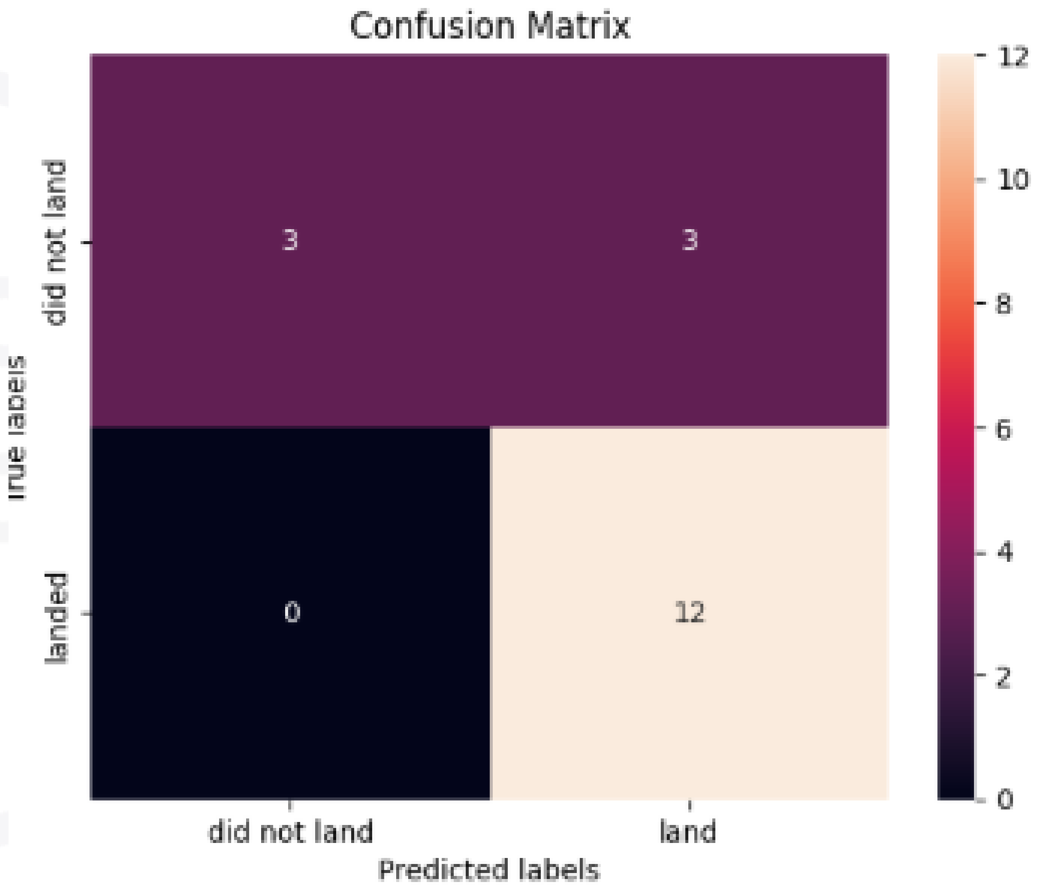
- The "best" parameters along with their performance given by accuracy scores for training data and test data are shown for each type of model below.
- Results show that all models ultimately work equally well (score=0.833 on the test data).

Modelling parameters for "best" models

	Logistic regression	Support vector machines	Decision trees	K-nearest neighbours
Parameters	{'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}	{'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}	{'criterion': 'gini', 'max_depth': 6, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'random'}	{'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
Accuracy	Training: 0.846 Test: 0.833	Training : 0.848 Test: 0.833	Training : 0.9 Test: 0.833	Training: 0.848 Test: 0.833

RESULTS: 4. Predictive model evaluation

- Confusion matrices were used to show more insights into the accuracy of the models when used to predict the test data.
- All four models produced the same confusion matrix (as per the right).
- There are 12 true positives and 3 true negatives, but 3 false positives which are the source of the inaccuracy – this reflects the accuracy score of 0.833 where 15 predictions were correct but 3 incorrect.



DISCUSSION: 1. Implications of analysis

Long term commitment is needed: SpaceX improved success rates, but gradually over about 5 years and a cumulative 40-50 launches, suggesting that it may also take SpaceY time to reach their maximum performance. They also deployed 14 booster types. Many launches may be needed to find the most successful boosters, launch sites and payloads. However, it was shown that a long term strategy can improve mission success significantly.

Boosters perform differently at different payloads: SpaceY will find that the optimal booster differs at different payload values. We found that **B4 and FT** achieve 100% success for <2k kg but **B5** achieves 100% success for payloads of between 2 and 6k kg. These are good options to start with.

Payloads of >6k kg perform poorly: No boosters were able to achieve more than 40% success rates at high payloads, and so new technologies may be needed to carry such payloads.

KSCLC-39A is the launch site with the best performance, and so represents the best bet for SpaceY to use with all other things being equal.

DISCUSSION: 2. Implications for future models

The choice of model doesn't affect results so far: Partly because of the small data set, we ultimately found that all models work as well as each other, generating accuracy scores of 0.83 on the test data.

Grid search is important: However, carrying out a grid search exercise is important, with significant differences in the accuracy found across ranges of parameters.

False positives, not false negatives are the problem: The confusion matrix shows that false positives rather than false negatives are causing inaccuracy. Therefore, we should expect that the model a small but significant percentage of positive outcomes that do not materialise.

Warnings on data limitations: The predictions are only found on 18 data points so more data would be useful. We also warn that the overall performance has changed over time, so older data may become a poor predictor of the future.

CONCLUSIONS

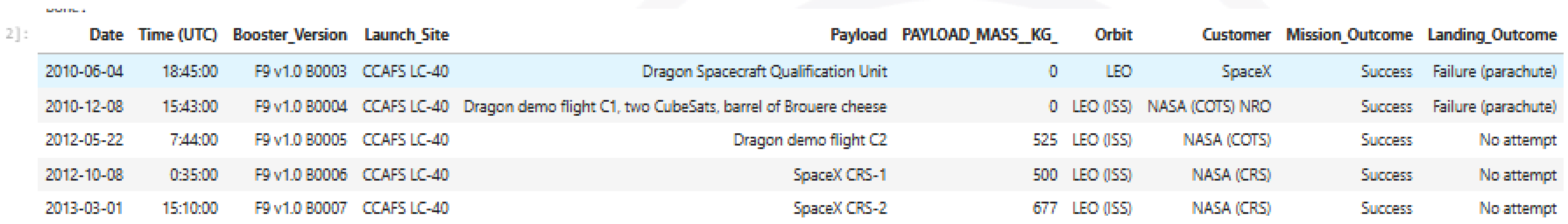
- Space X data provides many indicators for Space Y in their goal to increase competitiveness.
- Boosters should be selected differently for different payloads. (We shared the best choices for payloads of <2kkg and 2kkg-6kkg)
- At the moment, payloads of >6kkg should be avoided as success rates are very low.
- Performance can be expected to improve significantly with the number of launches carried out.
- Using the KSCLC-39A site offers the highest chance of success (with all else equal).
- When using predictive models in future, performance depends more on parameter selection than type. Models can predict success with reasonably high accuracy (83%) but can be prone to false positives.
- It would be desirable to collect more data in future to improve the robustness of the models and focus on more recent launch data, or including the number of cumulative launches as an independent variable to capture learning effects.

APPENDIX A - Definitions

- **Accuracy** is the number of correctly predicted values in the data set, defined as (TP+TN)/N where TP is true positive, TN is true negative and N is the size of the dataset.
- **C** is the inverse of regularization strength, and controls the trade-off between fitting the training data and avoiding overfitting. Higher values of C (lower regularization) lead to models that fit the training data more closely but may overfit and perform poorly on unseen data.
- **Entropy** parameters, as used in decision trees, calculate the different ways the impurity is measured once data is split into different branches.
- **Kernel:** These parameters are used in support vector machine calculations to transfer data to higher dimensional spaces, to separate the data.
- **Penalty** specifies the type of regularisation to use. Common options include 'l1' (LASSO regularization) and 'l2' (ridge regression).
- **Solver** specifies the algorithm used to fit the logistic regression model. Common options include 'lbfgs' (limited-memory Broyden-Fletcher-Goldfarb-Shanno) and 'liblinear' (suitable for large datasets with sparse features).

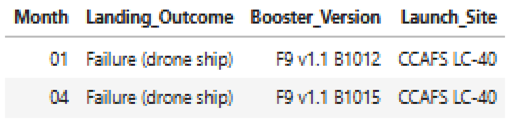
APPENDIX B - Miscell SQL queries

Task 2



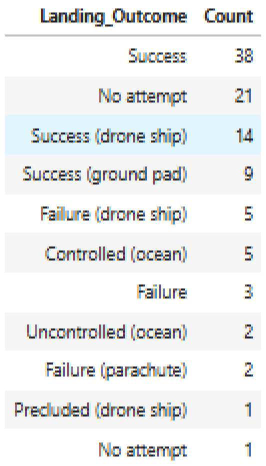
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Task 9



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

Task 10



Landing_Outcome	Count
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

APPENDIX C – Report requirements

- Uploaded the URL of your GitHub repository including all the completed notebooks and Python files (1 pt)
- Uploaded your completed presentation in PDF format (1 pt)
- Completed the required Executive Summary slide (1 pt)
- Completed the required Introduction slide (1 pt)
- Completed the required data collection and data wrangling methodology related slides (1 pt)
- Completed the required EDA and interactive visual analytics methodology related slides (3 pts)
- Completed the required predictive analysis methodology related slides (1 pt)
- Completed the required EDA with visualization results slides (6 pts)
- Completed the required EDA with SQL results slides (10 pts)
- Completed the required interactive map with Folium results slides (3 pts)
- Completed the required Plotly Dash dashboard results slides (3 pts)
- Completed the required predictive analysis (classification) results slides (6 pts)
- Completed the required Conclusion slide (1 pts)
- Applied your creativity to improve the presentation beyond the template (1 pts)
- Displayed any innovative insights (1 pts)