IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jimmy Bachir
November 2021

Outline

Executive Summary

The ability to reduce the cost of launching rockets into space will be a major factor in determining the success of a commercial space company.

Being able to predict if a first stage booster rocket can be reused helps determine the cost of a launch.

This report summarises the analysis conducted on SpaceX booster rocket survivals from 2010 to 2020.

Summary of Methodologies & Approaches Used

- Data collection via REST APIs and Web Scraping
- Data wrangling and processing to prepare for analysis
- Exploratory Data Analysis, using visualisation and SQL
- Interactive Data Analytics using Folium and Plotly Dash
- Predictive Analysis, using classification models

Summary of Results

- Launch Success is dependent on multiple variables, with later missions and smaller payload weights generally leading to greater success
- Launch sites are sited near the coastline, near a railway line, but further from highways and population centres
- All four classification models were equally able to predict launch success / failure with an accuracy of 0.833 on the test dataset

Introduction

Project background and context

- Many companies are developing rockets to enable commercial access to space
- SpaceX has developed a reusable first stage rocket which allows it to reduce the cost of launches from ~$165m to around $63m[1]
- If we can reliably determine whether a first stage rocket lands successfully, we can better predict the launch cost and hence be competitive

Key questions to answer

- Can we predict the success / failure of a first stage rocket landing based on key factors such as payload weight, launch site?
- Which factors have the greatest influence on launch rocket landing success?

Source: 1. Project Scenario & context

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected via a REST API and Web Scraping
- Perform data wrangling
  - Analyse data and derive a new variable to represent the mission outcome
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune, and evaluate classification models to predict mission outcome

Data Collection

- Data on SpaceX was collected using:
  1. SpaceX API (https://api.spacexdata.com/v4/launches/past)
  2. Web Scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

- The following slides explain the process followed to collect the data

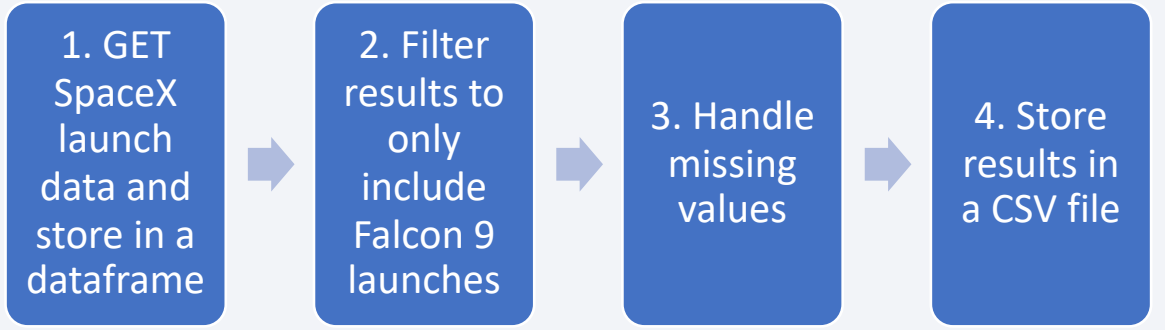
Data Collection – SpaceX API

Objective:

- Perform a GET request on the SpaceX API to return data about:
  - Launch Sites
  - Payload Data
  - Core Data

Github URL for workbook:

https://github.com/jbachir/Applied-Data-Sciences-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Flight Number
Date
Booster Version
Payload Mass
Orbit
Launch Site
Outcome
Flights
Grid Fins
Reused
Legs
Landing Pad
Block
Reused Count
Serial
Longitude & Latitude

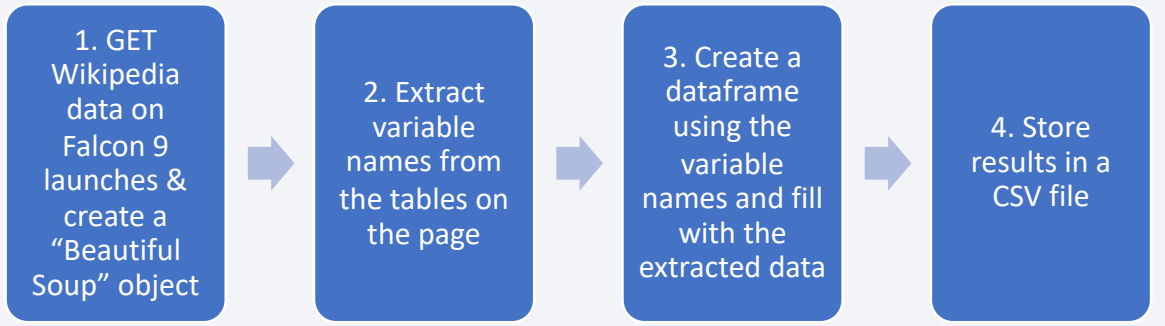
Use the mean value for Payload Mass

Retain 'None' values in 'Landing Pad' to show when no landing pad was used

Data Collection – Scraping

Objective:

- Scrape data from the SpaceX Falcon 9 Wikipedia page and use Beautiful soup to extract data



Flight Number
Launch Site
Payload
Payload Mass
Orbit
Customer
Launch Outcome
Version Booster
Booster landing
Date
Time

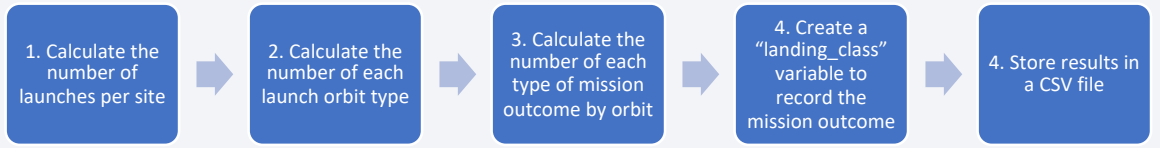
Github URL for workbook:

https://github.com/jbachir/Applied-Data-Sciences-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Objective

- Analyse & process data to make easier for further analysis



landing_class:
0 = failure
1 = success

Github URL for workbook:

https://github.com/jbachir/Applied-Data-Sciences-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling%20with%20count%20updates.ipynb

EDA with Data Visualization

Objective

- Conduct exploratory data analysis using to visualise the relationships between key variables and gain insights into how these variables affect the mission success rate
- This enables us to select the features to use to predict mission success

GitHub URL:

https://github.com/jbachir/Applied-Data-Sciences-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

Features analysed:

- Relationship between Flight Number and Launch Site
- Relationship between Payload and Launch Site
- Relationship between Orbit Type and Success Rate
- Relationship between Flight Number and Orbit Type
- Relationship between Payload and Orbit Type
- Launch Success yearly trend

EDA with SQL

Objective

- Use SQL to understand the data

Queries conducted:

- Display the names of the unique launch sites
- Display 5 records where launch sites begin with 'CCA'
- Display 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 achieved.
- 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

GitHub URL:

https://github.com/jbachir/Applied-Data-Sciences-Capstone/blob/main/EDA_SQL.ipynb

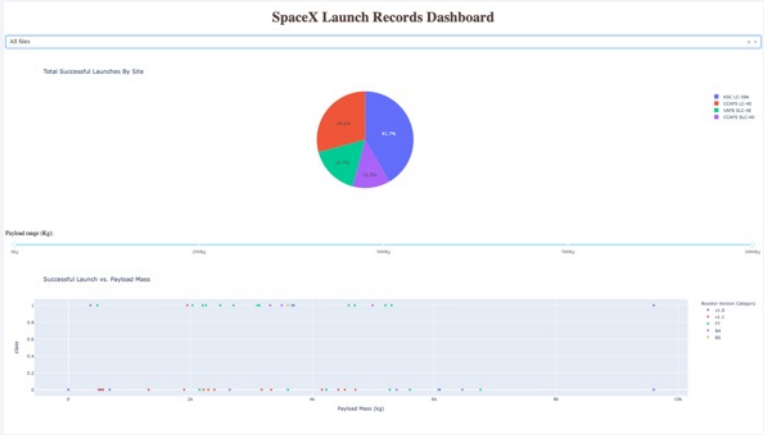
Build an Interactive Map with Folium

To better understand geographic data, various features were added to an interactive Folium map, as described below

Interactive Map Added Element	Why Added
Circles	Show the location of launch sites and other features of interest (cities, railways, highways, nearest coastline)
Markers	Show additional information about each marked location, and the ability to drill down to see more detail
Lines	Mark the distance between launch sites and other features of interest

Build a Dashboard with Plotly Dash

- A Dashboard was created to show:
  - Total Successful Launches
  - Successful Launches vs. Payload Mass, indicating the Booster Version
- For each displayed graph, the data could be shown per site and for all sites
- For the “Successful Launches vs. Payload Mass” chart, the payload mass could be specified in a range between 0 and 10,000kg



- These charts allowed us to easily see:
  - Which sites have the most successful launches and the highest launch success rate
  - Which payload range(s) have the highest and lowest success rates
  - Which F9 Booster version has the highest success rate

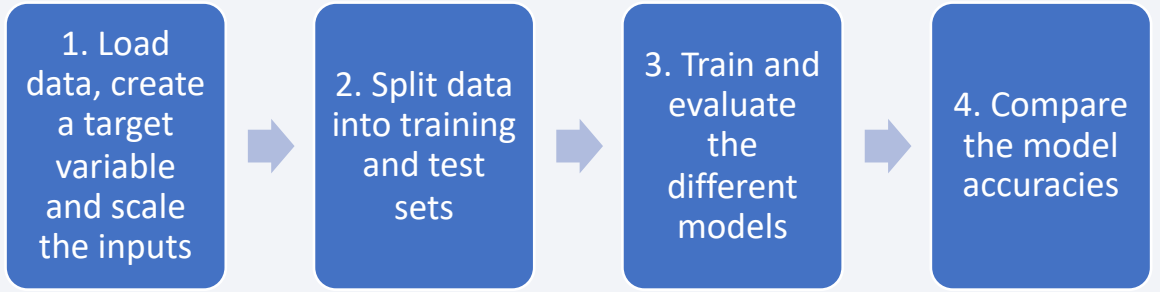
Predictive Analysis (Classification)

Objective:

- Determine which classification model can most accurately predict a mission outcome
- Four models were compared:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K Nearest Neighbours (KNN)
- In each case, a grid search was used to find the optimum model parameters

Github URL for workbook:

https://github.com/jbachir/Applied-Data-Sciences-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



- Load from CSV
- Generate a numpy array 'Y' with the target variable representing the mission outcome
- Standardise the input variables using scikit-learn StandardScaler

- Use an 80 / 20 split

Four models:

1. Logistic Regression
2. Support Vector Machine
3. Decision Tree
4. K Nearest Neighbours

Use a Grid Search to find the best model parameters

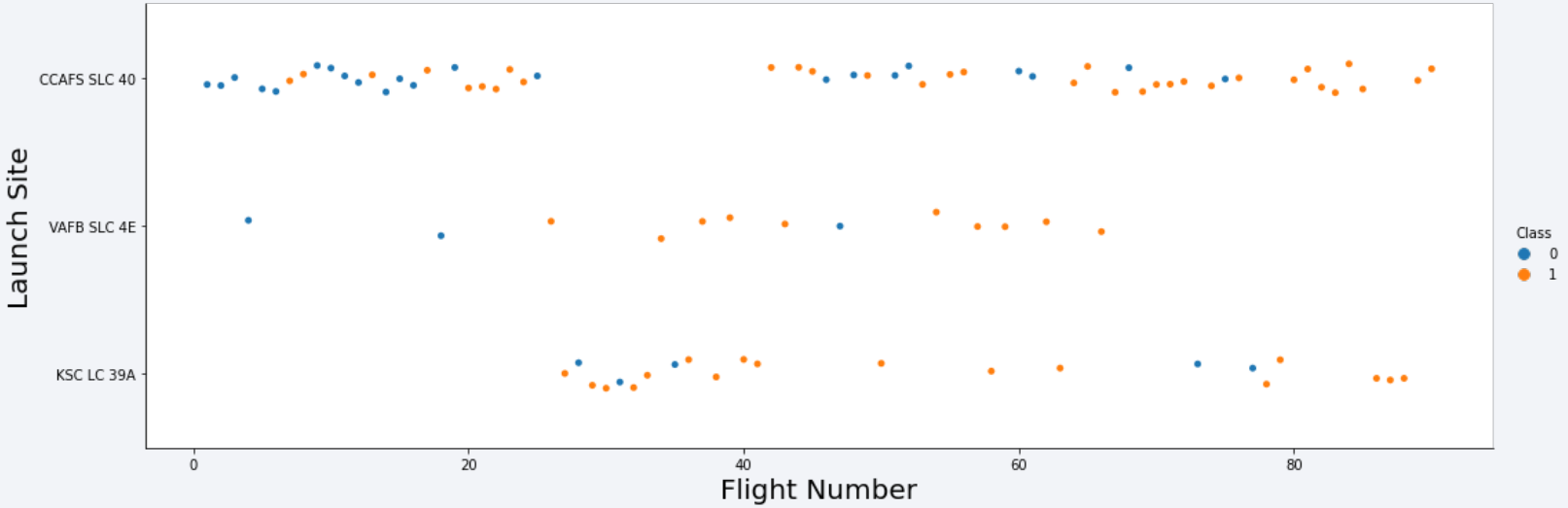
Measure the accuracy and show the confusion matrix for each model

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Flight Number vs Launch Site



Classes: 0 is mission failure; 1 is mission success

Launch sites:

CCAFS SLC 40: Cape Canaveral Space Launch Complex 40

VAFB SLC 4E: Vandenberg Space Force Base Space Launch Complex 4E

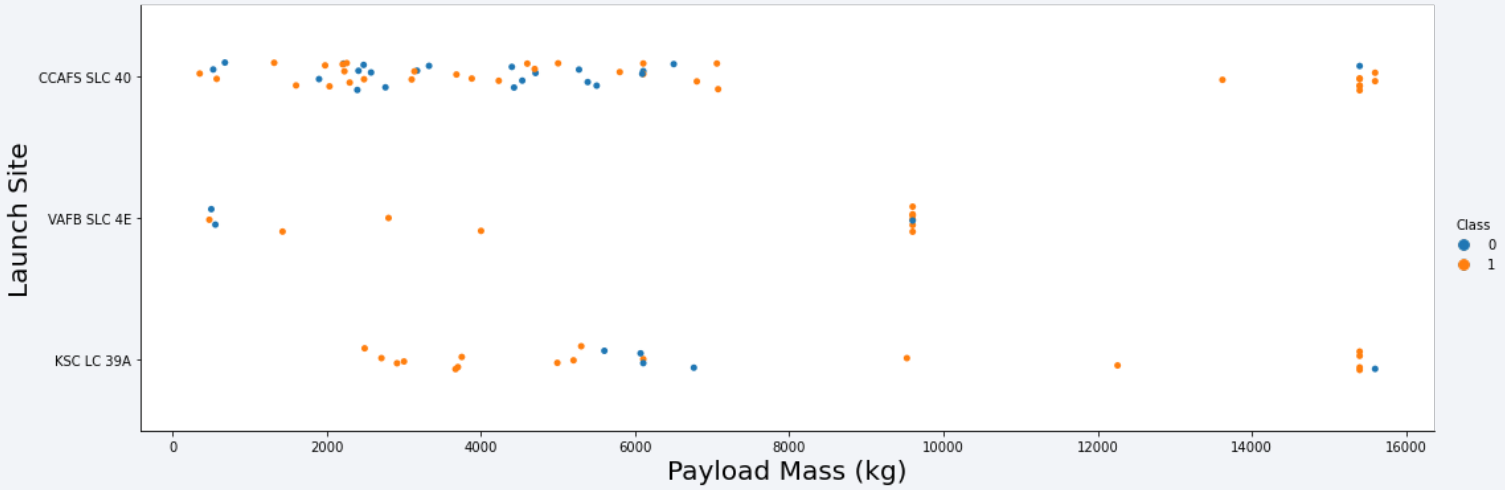
KSC LC 39A: Kennedy Space Center Launch Complex 39A

- For all launch sites, as the flight number increases, there is a greater probability of mission success
- VAFB and KSC launches have a 77% success rate
- CCAFS launches have a 60% success rate
- 61% of launches were from the CCAFS site (14% from VAFB, 24% from KSC)[1]

Note: 1. total does not add to 100% due to rounding

Payload vs. Launch Site

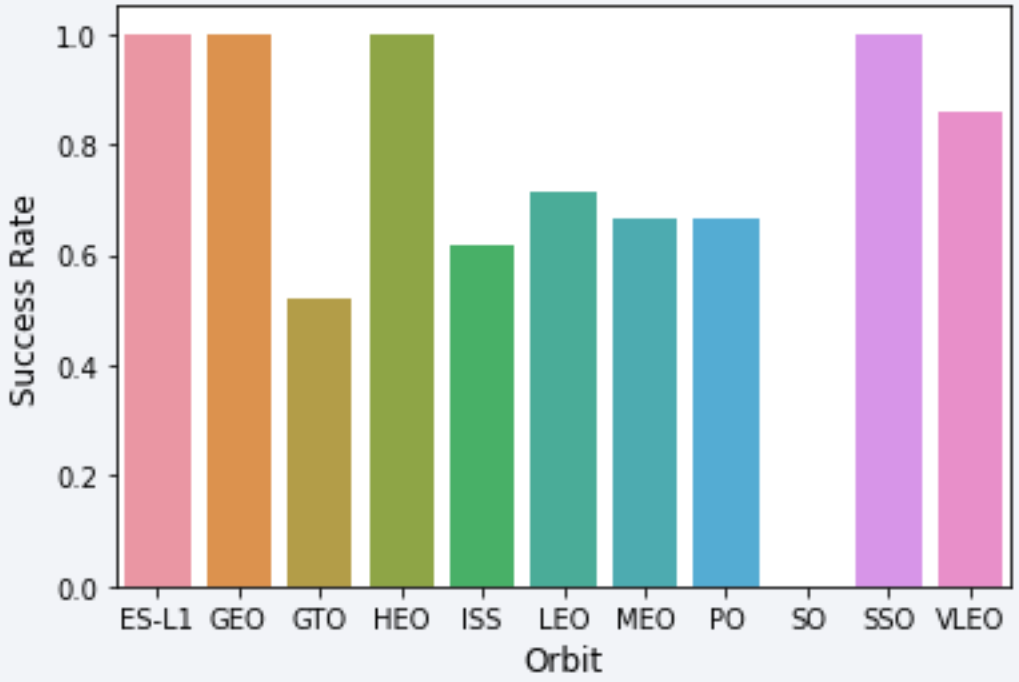
Payload Mass (kg) vs Launch Site



- There does not appear to be a strong correlation between payload mass and mission outcome
- There are no launches from VAFB with a payload of more than 10,000kg
- Over 85% of launches have a payload over 10,000kg

Success Rate vs. Orbit Type

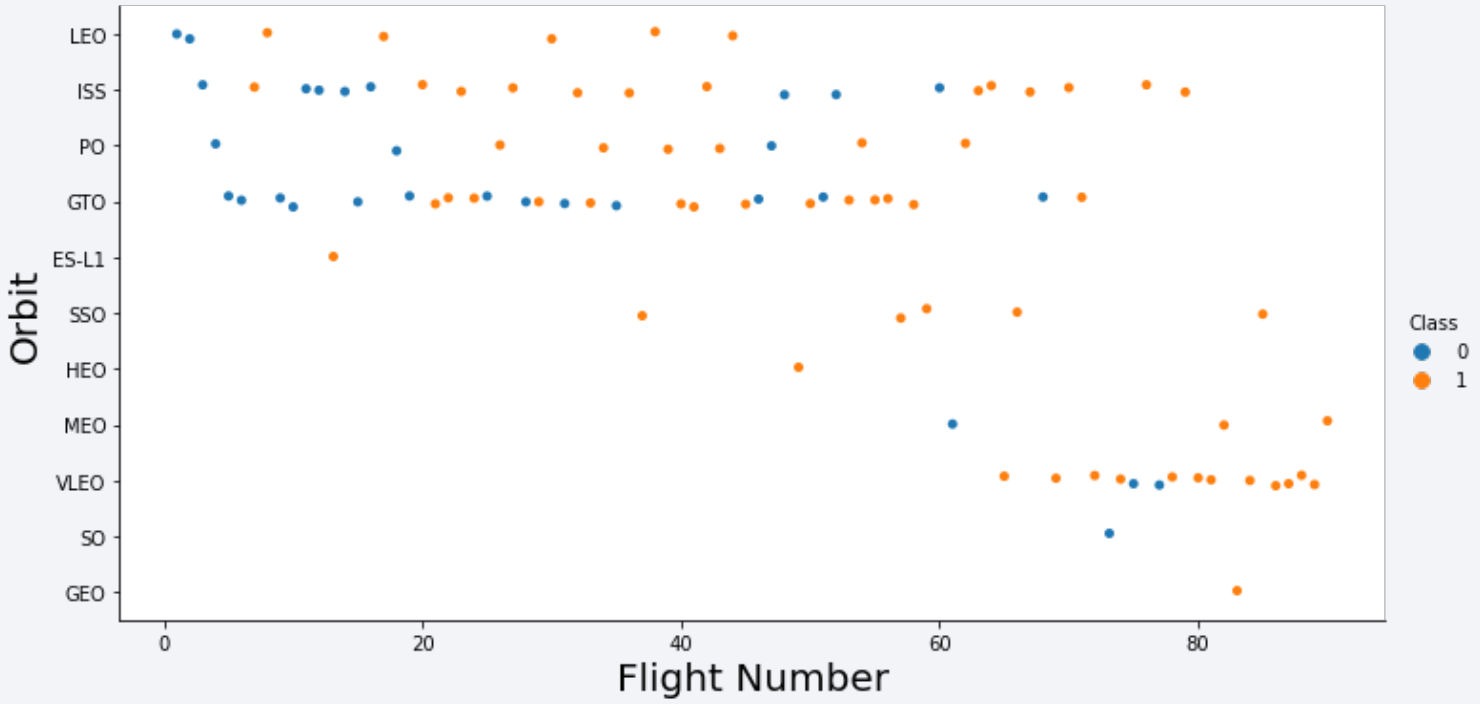
Success Rate for each Orbit Type



- The following orbit types have a 100% success rate
  - ES-L1, GEO, HEO, SSO
  - However, the ES-L1, GEO and HEO orbit types only had one mission each
- The SO orbit type has a 100% failure rate, however only one mission attempted this orbit type
- By itself, Orbit Type does not seem a strong predictor of mission success

Flight Number vs. Orbit Type

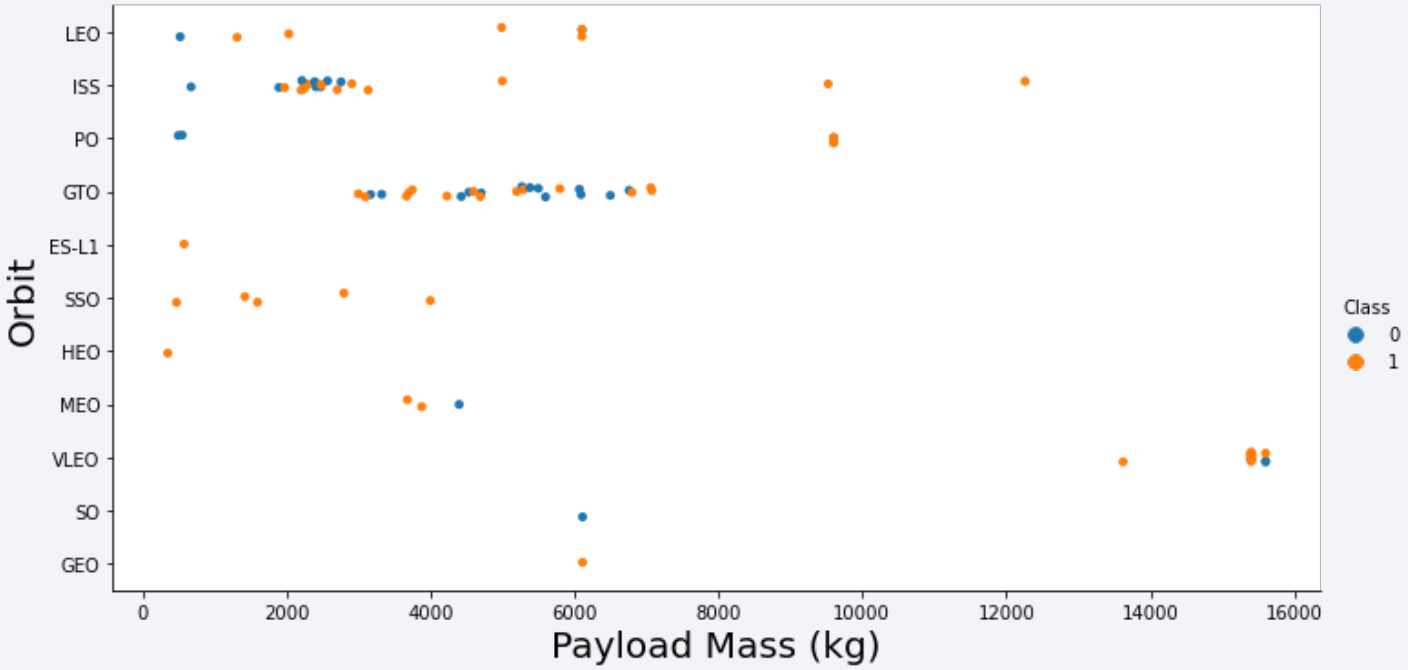
Flight Number vs. Orbit Type



- For all orbit types, there is a general correlation between flight number and success rate, i.e. later missions have a higher success rate
- The LEO orbit type has the clearest correlation with flight number (the first 2 missions failed; the subsequent 5 all succeeded)

Payload vs. Orbit Type

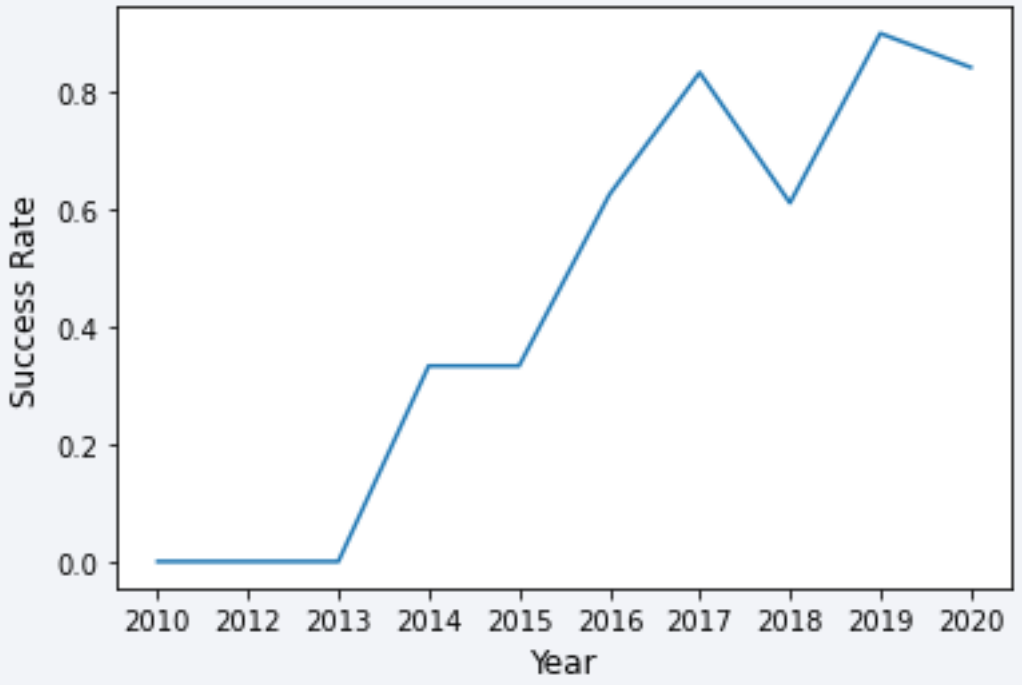
Payload Mass (kg) vs. Orbit Type



- The PO, LEO and ISS orbit types seem to have a more successful rate with heavier payloads
- For the GTO orbit type, there does not appear to be a connection between payload mass and success rate

Launch Success Yearly Trend

Success Rate yearly trend



Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
# Missions	1	0	1	3	6	6	8	18	18	10	19

- From 2013 to 2020 the success rate has been increasing
- The “dips” in success rate in 2018 and 2020 could be partially explained by more missions being carried out in those years than other years, although 2017 had the same number of missions as 2018

All Launch Site Names

- There are four SpaceX launch sites
  1. CCAFS LC-40 (Cape Canaveral Space Launch Complex 40)
  2. CCAFS SLC-40 (Cape Canaveral Space Launch Complex 40)
  3. KSC LC-39A (Kennedy Space Center Launch Complex 39A)
  4. VAFB SLC-4E (Vandenberg Space Force Base Space Launch Complex 4E)

- These were found using the following SQL query on the dataset:

```
select distinct launch_site from spacexdataset;
```

- Using the “distinct’ keyword means only the unique launch_site names are selected

Launch Site Names Begin with 'CCA'

- Below are 5 records where the launch site begins with "CCA"

- They were selected using the following query:

```
select * from spacexdataset where launch_site like 'CCA%' limit 5;
```

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

- The “like” keyword selects rows that start with ‘CCA’ and “limit 5” returns only 5 rows

Total Payload Mass

- The total payload carried by boosters from NASA was 99,980kg

- This was found using the following query:

```
select sum(payload_mass__kg_) from spacexdataset where customer
like 'NASA%';
```

- As there are multiple NASA variants, we need to use the “like” keyword to select them all

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2,928kg

- This was found using the following query:

```
select avg(payload_mass__kg_) from spacexdataset where
booster_version = 'F9 v1.1';
```

- The “avg” keyword calculates the average of the result

First Successful Ground Landing Date

- The date of the first successful landing outcome on a ground pad was:
  - 22nd December 2015 (2015-12-22)

- This was found using the following query:

```
select min(date) from spacexdataset where landing__outcome =
'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The following boosters have successfully landed on a drone ship and had payload mass greater than 4000kg but less than 6000kg;
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

- This was found using the following query:

```
select booster_version from spacexdataset where landing__outcome =
'Success (drone ship)'and payload_mass__kg_ > 4000 and
payload_mass__kg_ < 6000;
```

Total Number of Successful and Failure Mission Outcomes

- There were 100 successful and 1 failure mission outcomes
  - Of the successful mission outcomes, 1 was marked as “Success (payload status unclear)”
- Calculate the total number of successful and failure mission outcomes

- This was found using the following query:

```
select mission_outcome, count(mission_outcome) from spacexdataset
group by mission_outcome;
```

Boosters Carried Maximum Payload

- The following boosters have carried the maximum payload mass (15,600kg):
  - 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 found using the following query:

```
select booster_version from spacexdataset
where payload_mass__kg_ = (select max(payload_mass__kg_) from spacexdataset)
order by booster_version;
```

2015 Launch Records

- In 2015, the failed landing outcomes on a drone ship, their booster versions, and launch site names were:

LANDING__OUTCOME	BOOSTER_VERSION	LAUNCH_SITE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- This was found using the following query:

```
select landing__outcome, booster_version, launch_site from
spacexdataset

where year(date) = 2015 and landing__outcome like 'Failure%';
```

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

- The adjacent table is the count of landing outcomes between the dates 2010-06-04 and 2017-03-20, 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

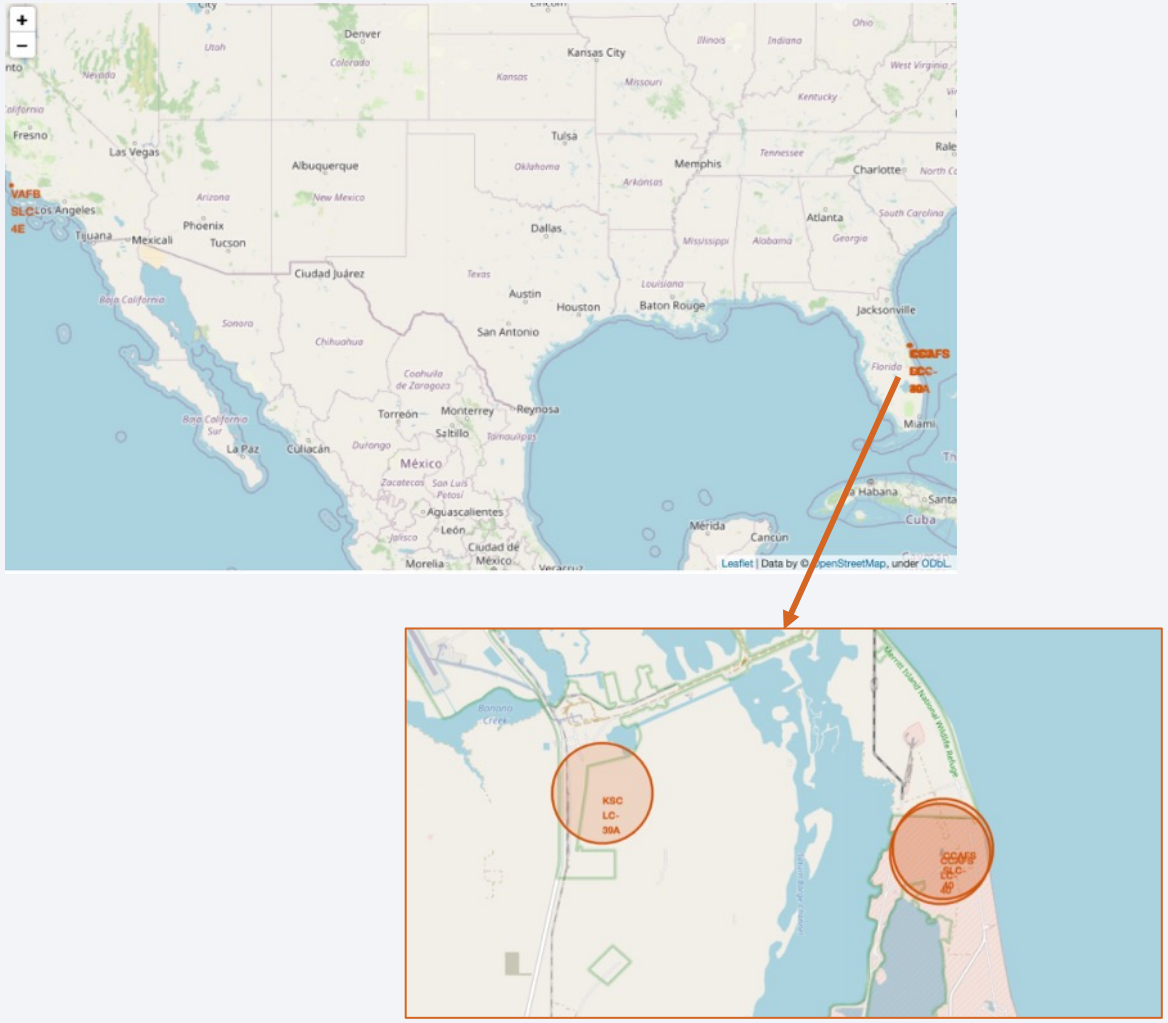
- This was found using the following query:

```
select landing__outcome,
count(landing__outcome) from spacexdataset

where date between '2010-06-04' and '2017-
03-20' group by landing__outcome

order by count(landing__outcome) desc;
```

Section 4

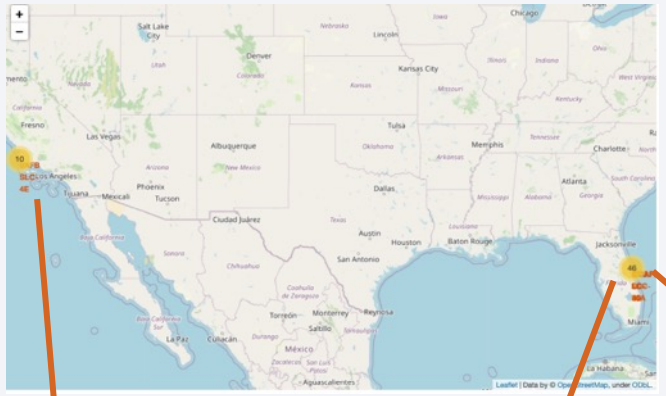
Launch Sites Proximities Analysis

Launch Sites Map

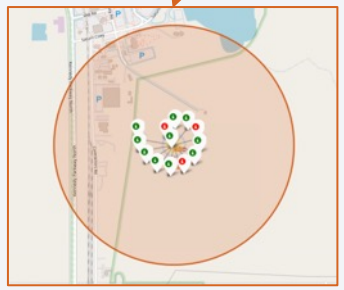


- This map shows the launch sites' locations
- There are 4 sites shown: 3 in Florida and 1 in California
- As the 3 Florida sites are near each other, they are expanded in the lower map
- All the sites are near the coast

Map showing launch outcomes at each site



Expanded view of launch outcomes at the VAFB SLC-4E site



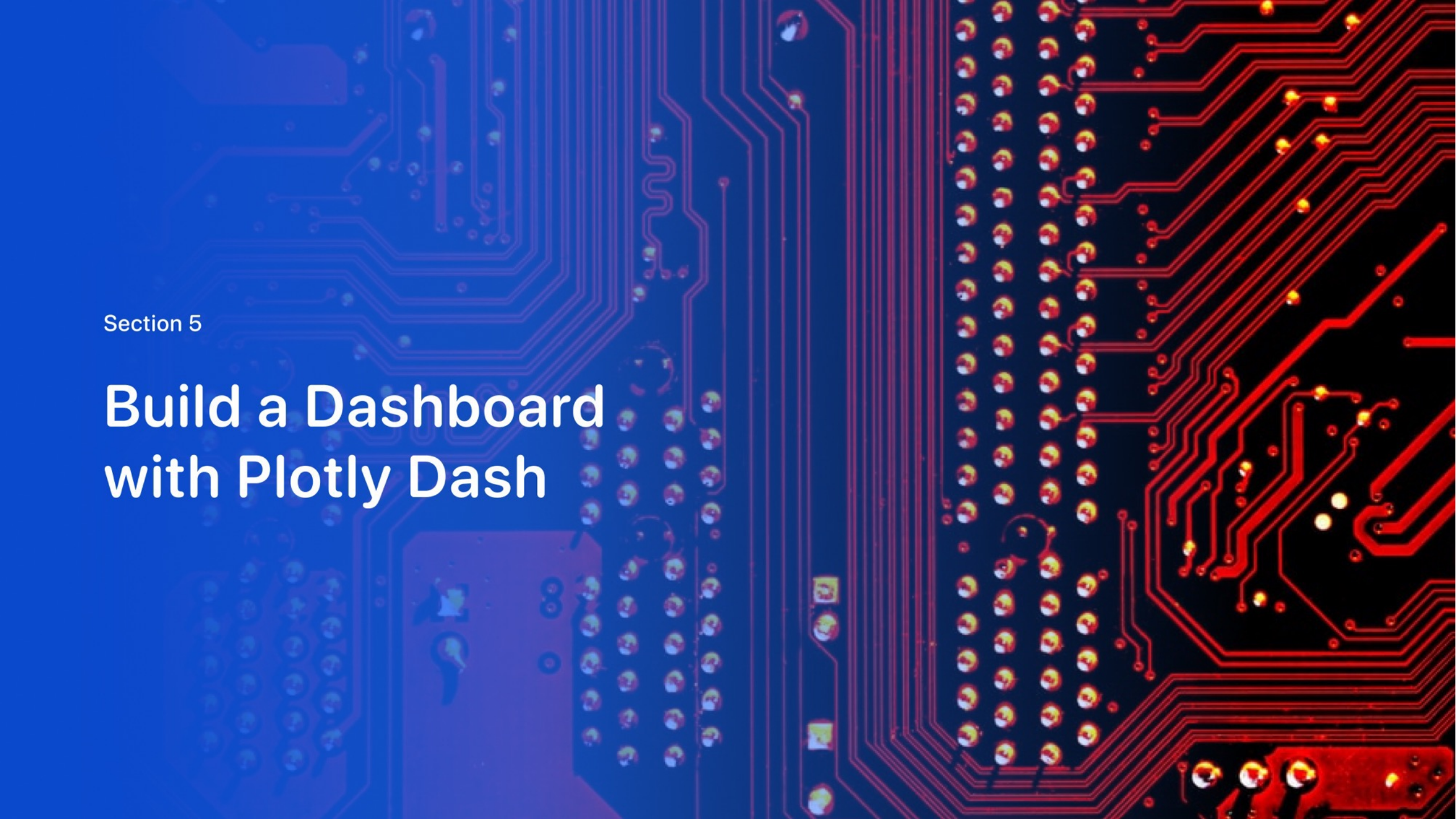
Expanded view of launch outcomes at the KSC LC-29A site

Expanded view of launch outcomes at the CCAFS LC-40 and CCAFS SLC-40 sites

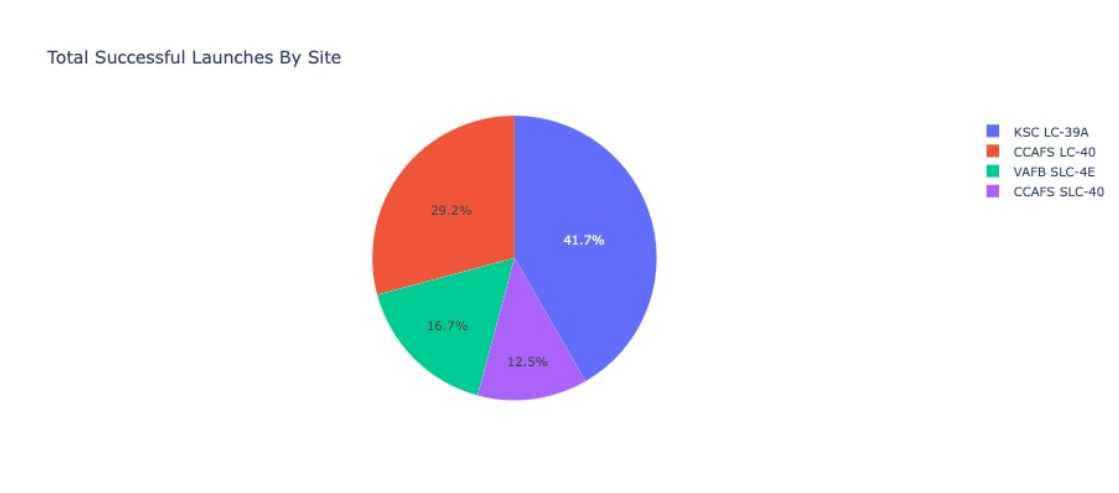
- The adjacent maps show launch sites with markers for success / failure launches
- At the full map level, the launch outcomes are grouped, displaying the total number of launches and colour coded based on success / failure
- When zoomed in, individual launches are colour coded for success / failure

Map of VAFB launch site with distance to nearby locations

- The adjacent map shows the distance from the VAFB SLC-4E launch site to the key landmarks
  - Coastline is 1.36km away
  - Railway line: 1.31km
  - Highway: 14.02km
  - Nearest town (Lompoc): 14.07km
- This is a similar pattern for the other launch sites
- We can infer that the launch sites are
  - Near the coastline in case of misfiring launches
  - Near a railway line for transportation of materials
  - Away from major population centres and highways, to minimise risk to the population

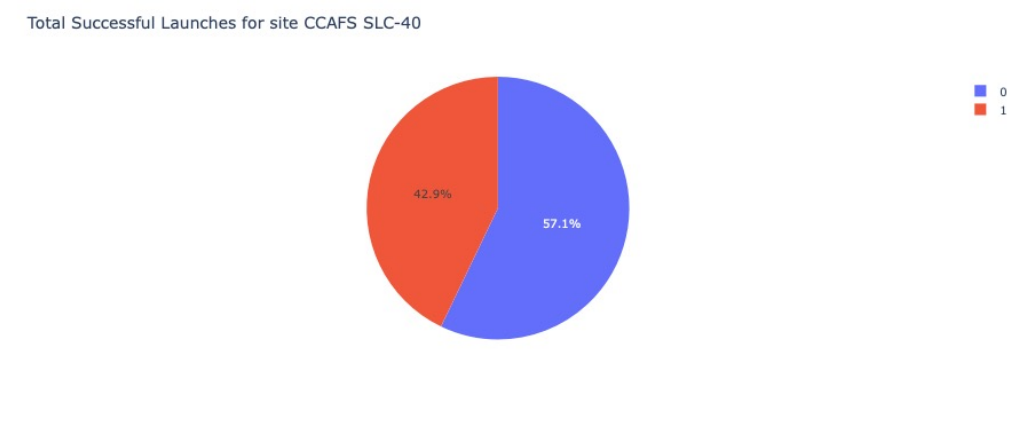

Section 5
Build a Dashboard with Plotly Dash

Launch Success For All Sites



- The pie chart on the left shows the breakdown of successful launches by site
- The KSC LC-39A site has the most successful launches (41.7%)
- The site with the smallest proportion of successful launches is the CCAFS SLC-40 site

The site CCAFS SLC-40 has the highest proportion of successful launches

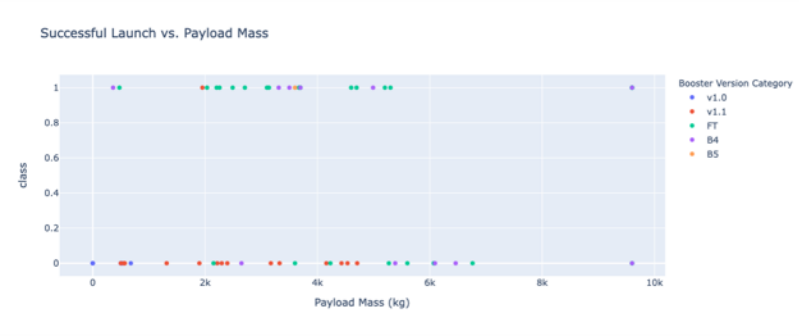


- The site CCAFS SLC-40 has the highest proportion of successful launches (42.9% success rate)
- As noted on the previous slide, this is also the site with the fewest number of overall successful launches compared to the other sites

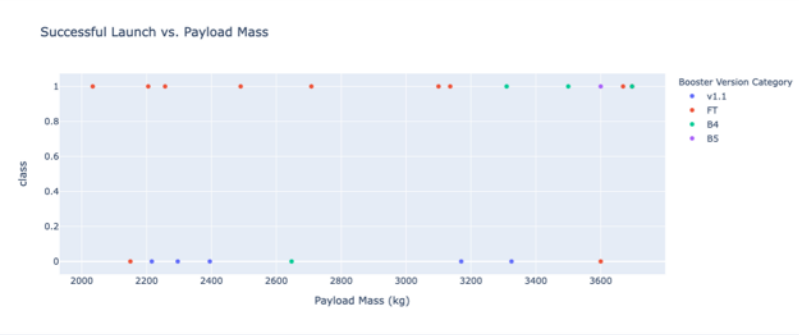
Key:
0 – failed launch
1 – successful launch

Successful Launch vs. Payload Mass

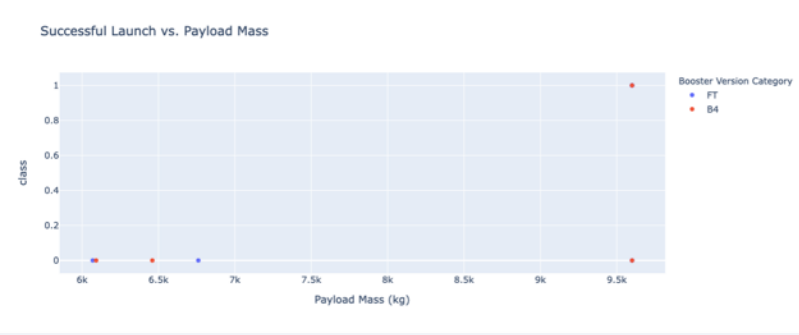
All Payload Masses
(0 – 10,000kg)



Highest Success Rate
(2,000 – 4,000kg)



Lowest Success Rate
(6,000 – 10,000kg)



- The adjacent charts show the success rate for different payload masses, color coded by booster version

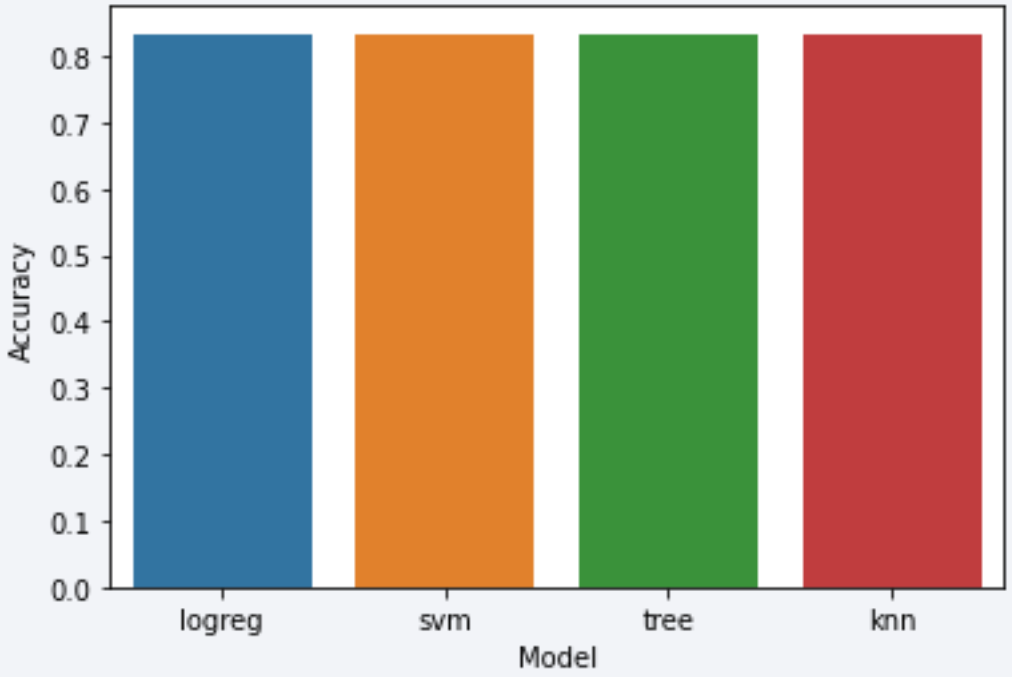
- The payload range 2,000 – 4,000 has the highest success rate
- The payload range 6,000 – 10,000 has the lowest success rate[1]
- The booster type ‘FT’ has the highest success rate

1. Note: range chosen to include at least 1 successful launch

Section 6

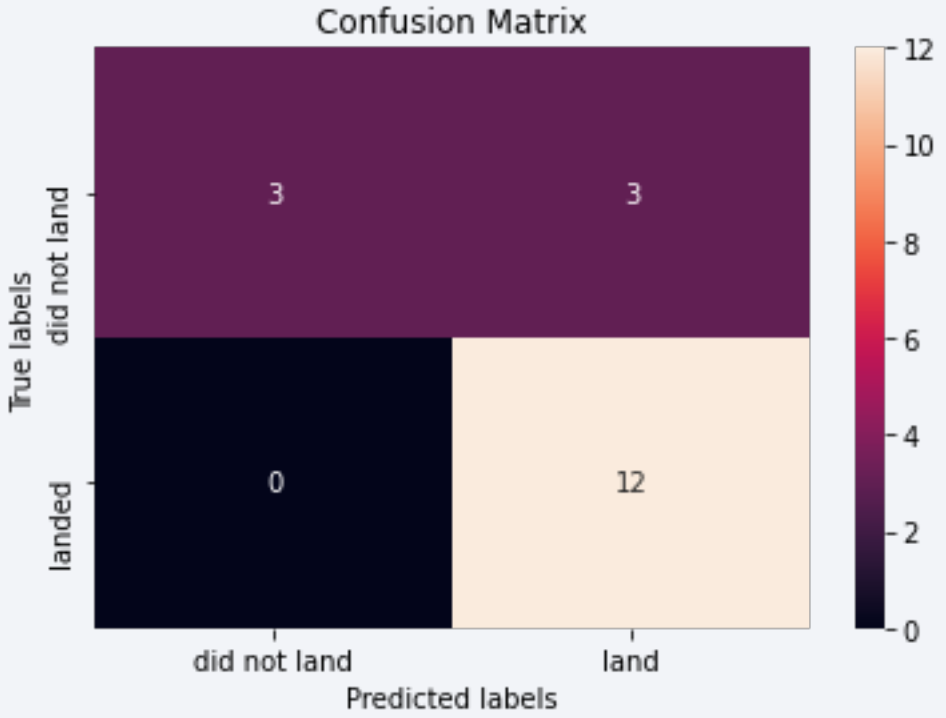
Predictive Analysis (Classification)

Classification Accuracy



- The adjacent bar chart shows the model accuracy for all the classification models:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K Nearest Neighbours
- For all models, the classification accuracy on the test data set is the same (0.833)

Confusion Matrix



- The confusion matrix is the same for all models
- It shows the True & False Positive and Negative classifications
- In each case, there are:
  - 12 True Positive Classifications
  - 3 True Negative Classifications
  - 0 False Positive Classifications
  - 3 False Negative Classifications

Conclusions

- Launch Success is dependent on multiple variables
  - Later missions have higher success rates
  - Lower payload weights generally have higher success rates
  - Site KSC LC-39A has the highest success rate
- Launch sites are sited near the coastline, near a railway line, but further from highways and population centres
- All four classification models were equally able to predict launch success / failure
  - An accuracy of 0.833 was obtained on the test dataset

Thank you!