Winning Space Race with Data Science

Dr Alasdair Brown

5th July, 2022

Outline

Executive Summary

- SpaceX data were collected and analysed
  - Mission profiles contained in future bid documents can be compared to data analysis of SpaceX launches to assess success potential within different scenarios:
    - Launch site requirement
    - Payload mass
    - Orbit required for payload
- A successful model was developed based on the publicly available data
  - The model predicts if a first stage landing will be successful with greater than 80% accuracy
- The model and data analysis can be used to develop a commercial bidding strategy to compete against SpaceX for launch missions

Introduction

- SpaceX data were analysed to yield input into a pricing proposal for a rival rocket launch
- The following analysis provides insight into:
  - Effect of launch site location on mission success
  - Success as a function of payload mass
  - Orbit requirement impact on success outcomes for landing the first stage
  - Logistical parameters around site location for rocket service
  - Training machine learning model to predict mission success

Section 1

Methodology

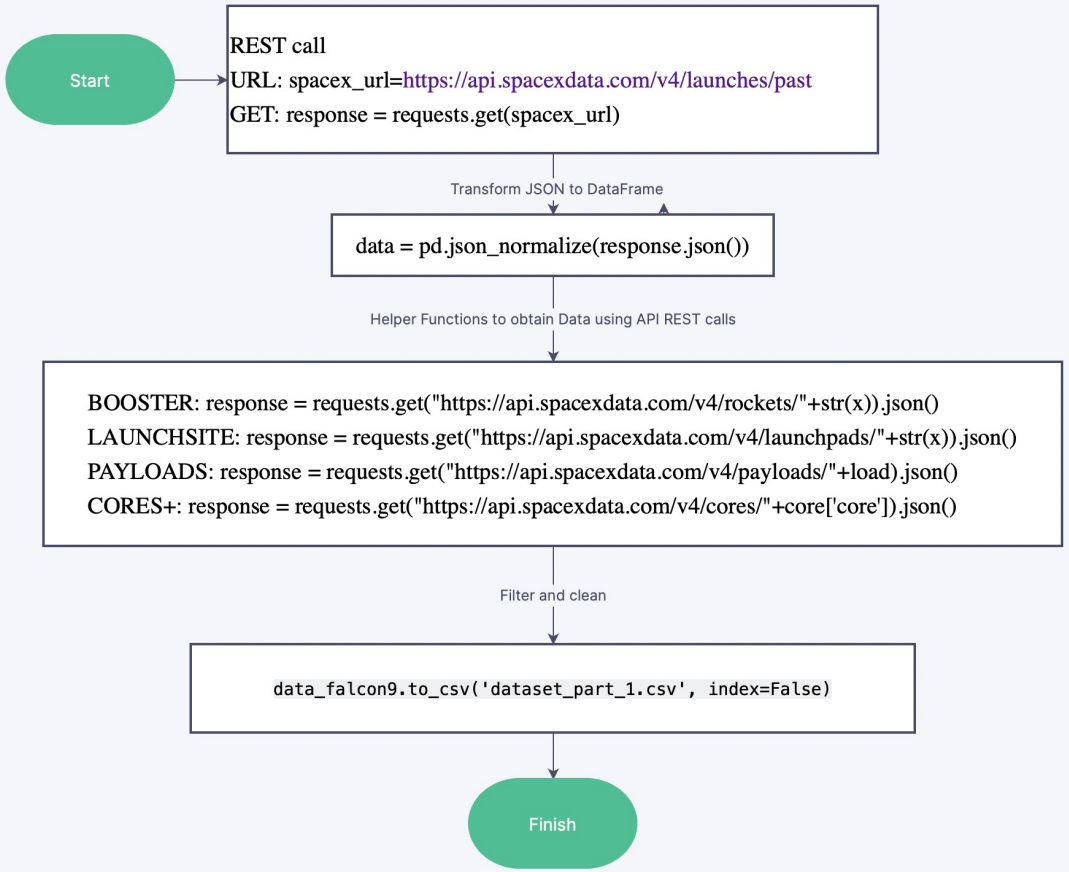
Methodology

Executive Summary

- Data collection methodology:
  - Publicly available data were gathered from Space X and internet sources
- Performed data wrangling
  - Data were profiled, irrelevant data discarded, missing data filled in using an appropriate strategy and an outcome variable defined for later modelling
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Four classification models were tested using the dataset and the best performing model chosen

Data Collection – SpaceX API

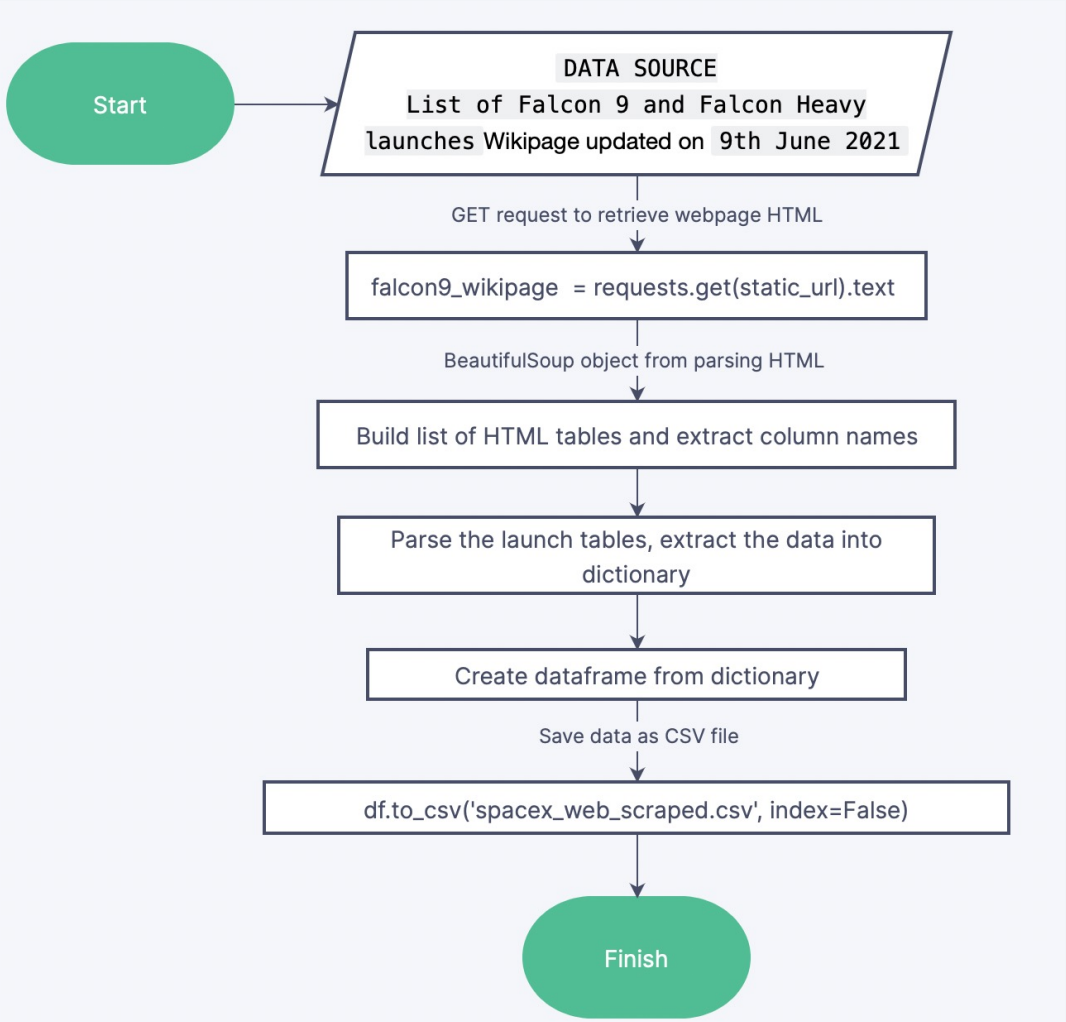
- Launch Data requested from spacex data website using REST API GET calls
- Resulting JSON format converted to pandas dataframe for processing
- Subsequent GET calls expanded details for booster type, payloads, &c.
- Dataframe missing values replaced by mean values and masked for Falcon 9 launches alone
- Final dataset saved to CSV file



https://github.com/AlasdairBrown/IBM-DataScience-Capstone-Project/blob/main/1%20-%20SpaceX%20Data%20Collection%20API%20(Final).ipynb

Data Collection - Scraping

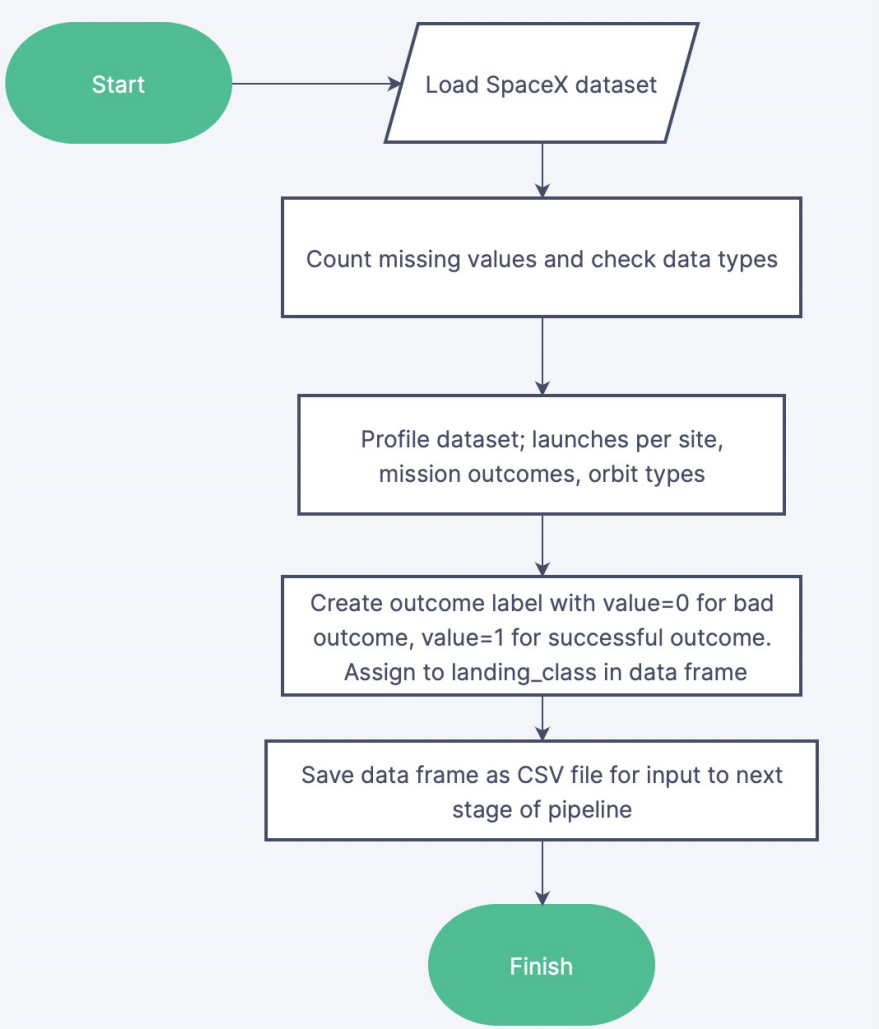
- Scraped launch data from archived Wikipedia page using GET request
- Created BeautifulSoup object from resulting JSON data
- Parsed the resulting tables to build dataframe to save to CSV file for next stage in data pipeline



https://github.com/AlasdairBrown/IBM-DataScience-Capstone-Project/blob/main/2%20-%20Space%20X%20Web%20Scraping%20(Final).ipynb

Data Wrangling

- Dataset from prior processing steps loaded in from CSV file and checked for missing values
- Data profiled to check sense of values in dataset and to get to know the data
- Created outcome label for success/failure of landing to use in training later ML model
- Saved dataframe to CSV to continue processing pipeline

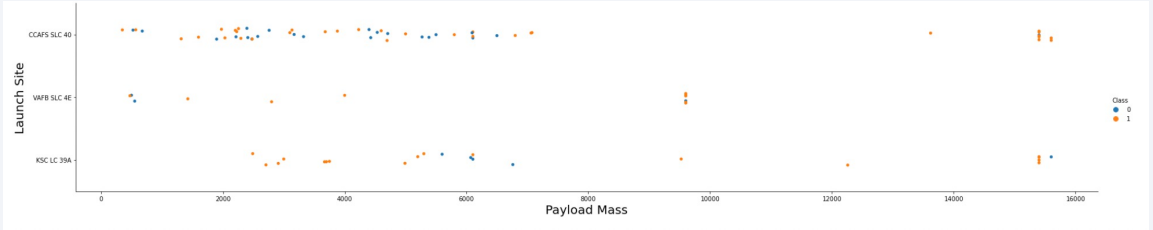


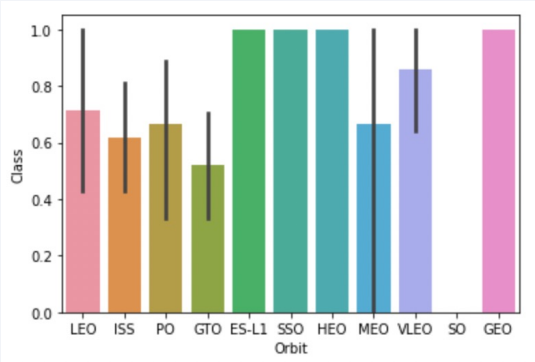
https://github.com/AlasdairBrown/IBM-DataScience-Capstone-Project/blob/main/3%20-%20SpaceX%20Data%20Wrangling%20(Final).ipynb

EDA with Data Visualization

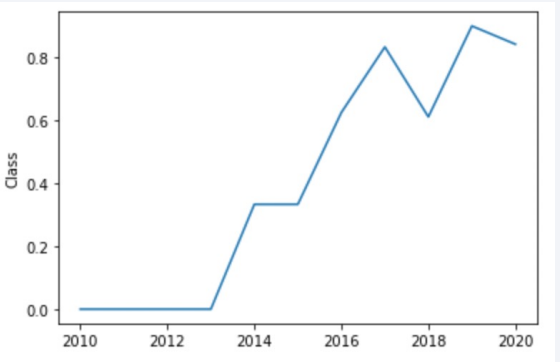
- Used three types of chart to visualise the data:
  - Scatter plots – used to visualise relationships between variables and to see any obvious correlation, e.g. payload mass and success rate
  - Histogram/Bar Chart – visualise categorical variables' data, e.g. success rate for different orbits
  - Line Plot – illustrate variation of parameter through time/sequence, e.g. success rate as a function of time/date

Scatter plot





Bar Chart



Line plot

https://github.com/AlasdairBrown/IBM-DataScience-Capstone-Project/blob/main/4%20-%20Space%20X%20EDA%20Data%20Visualisation%20(Final).ipynb

EDA with SQL

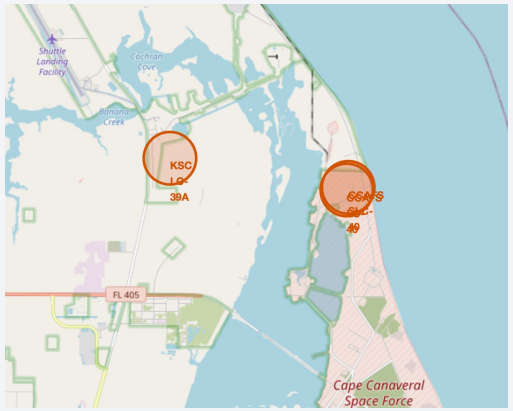
Summary of SQL Statements used in EDA for SpaceX data

1. select distinct launch_site from SPACEXDATASET;
2. select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
3. select sum(PAYLOAD_MASS__KG_) as "Payload Mass (kg)" from SPACEXDATASET where CUSTOMER like 'NASA (CRS)';
4. select avg(PAYLOAD_MASS__KG_) as "Ave Payload Mass (kg)" from SPACEXDATASET where BOOSTER_VERSION like '%F9 v1.1%';
5. select DATE as "First Successful Landing GroundPad" from SPACEXDATASET where "Landing _Outcome" like 'Success (ground pad)' order by DATE limit 1;
6. select BOOSTER_VERSION from SPACEXDATASET where (PAYLOAD_MASS__KG_ >= 4000 and PAYLOAD_MASS__KG_ < 6000) and ("Landing _Outcome" like 'Success (drone ship)');
7. select MISSION_OUTCOME as "Mission Outcome", count(*) as "Count" from SPACEXDATASET group by MISSION_OUTCOME;
8. select BOOSTER_VERSION as "Booster" from SPACEXDATASET where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEXDATASET);
9. select DATE as "Date", "Landing _Outcome", BOOSTER_VERSION as "Booster", LAUNCH_SITE as "Launch Site" from SPACEXDATASET where "Landing _Outcome" like 'Failure (drone ship)' and YEAR(DATE) = '2015';
10. select "Landing _Outcome" , count("Landing _Outcome") as "Count" from SPACEXDATASET where (DATE > '2010-06-04' and DATE < '2017-03-20') group by "Landing _Outcome" order by count desc;

https://github.com/AlasdairBrown/IBM-DataScience-Capstone-Project/blob/main/5%20-%20SpaceX%20EDA%20SQL%20(Final).ipynb

Interactive Map of Launch Site data

- As shown to the right
  - circles marked launch sites
  - markers marked launch success
  - lines marked offsets to nearby locations
- Different map elements visually show different information, hence the choice of each marker type



Circles

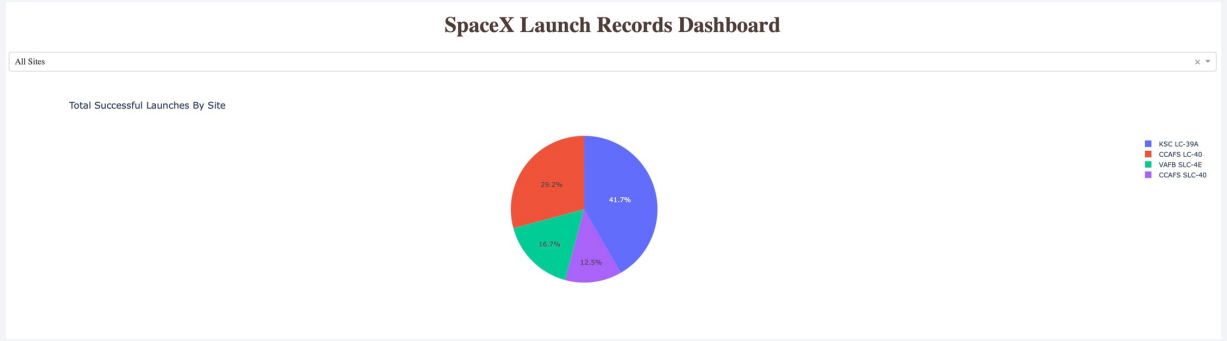
Markers

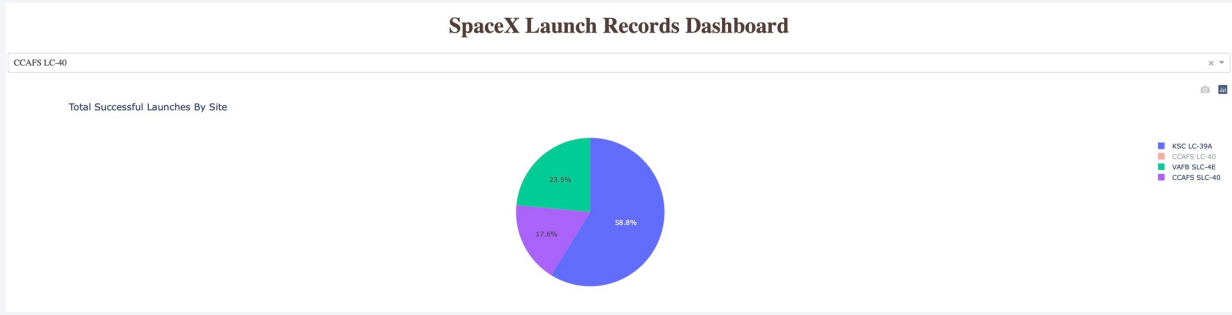
Lines

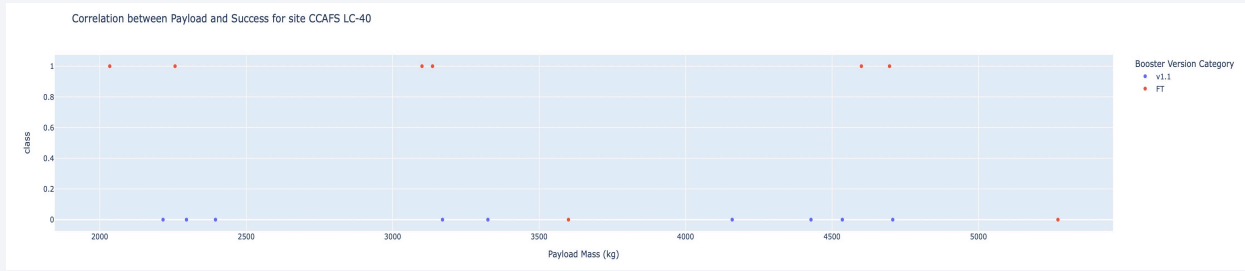
https://github.com/AlasdairBrown/IBM-DataScience-Capstone-Project/blob/main/6%20-%20SpaceX%20Folium%20Analysis%20(Final).ipynb

SpaceX mission Dashboard

- Dashboard includes pie chart to show launch success for all sites, or for chosen site
- Scatter plot of success (class) against payload mass for all sites, or chosen site
  - Slider used to select mass range for payload graph
- Pie charts are good for showing relative percentages quickly
- Scatter plot allows quick visual inspection of success for given (selectable) payload mass range



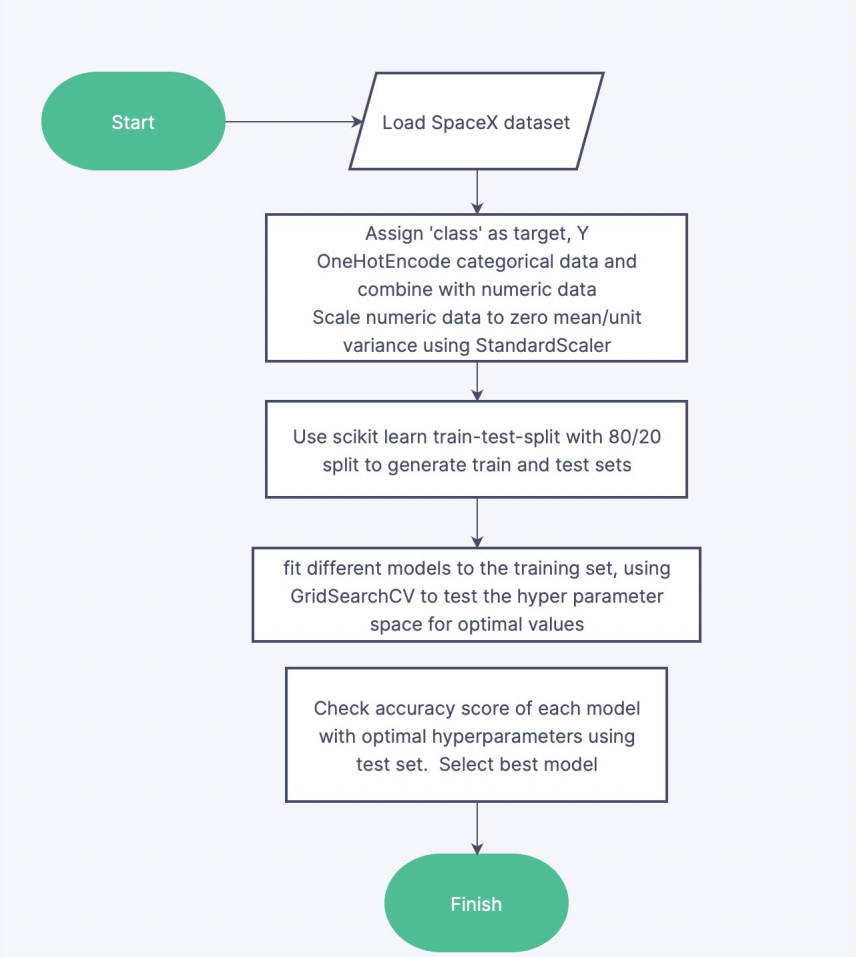




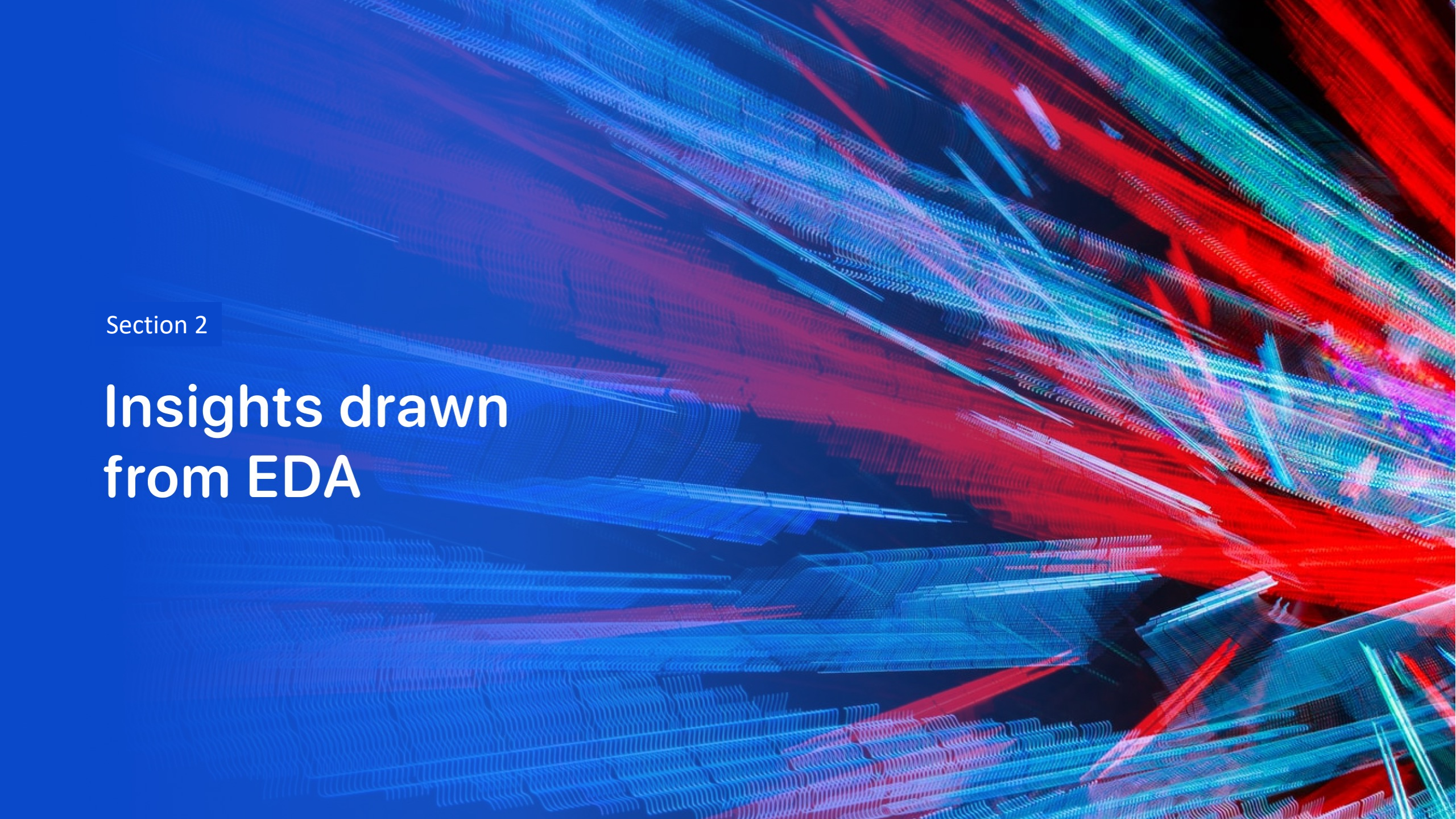
https://github.com/AlasdairBrown/IBM-DataScience-Capstone-Project/blob/main/7%20-%20Space%20X%20Plotly%20Dash%20dashboard%20(Final).py

Predictive Analysis (Classification)

- Dataset imported into pandas dataframe
- Categorical data encoded using OneHotEncoding, numerical data standardised to zero-mean/unit variance using StandardScaler
- Data split into training/test set with 20% holdback
- Tested different models and compared accuracy scores on test sets after fitting to select best model. Used GridSearchCV to test different hyperparameters and select optimal for each model.

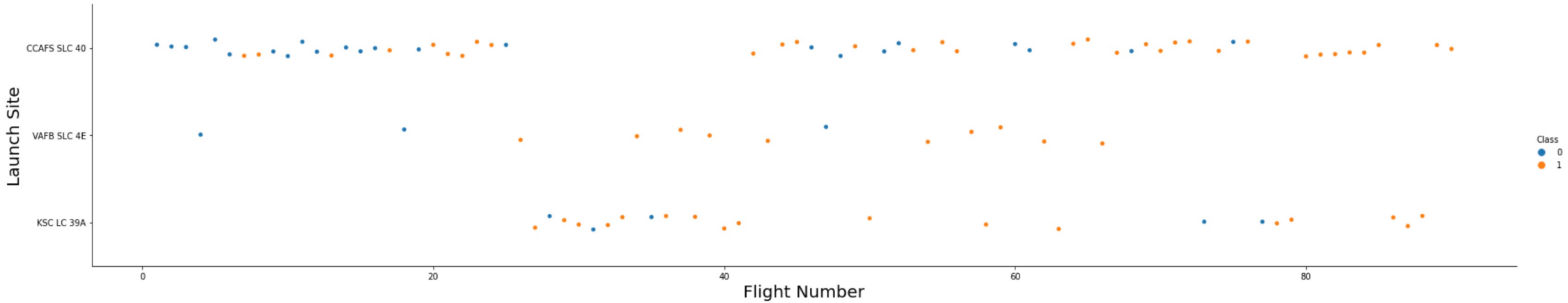


https://github.com/AlasdairBrown/IBM-DataScience-Capstone-Project/blob/main/8%20-%20SpaceX%20ML%20Prediction%20Model%20(Final).ipynb


Section 2
Insights drawn from EDA

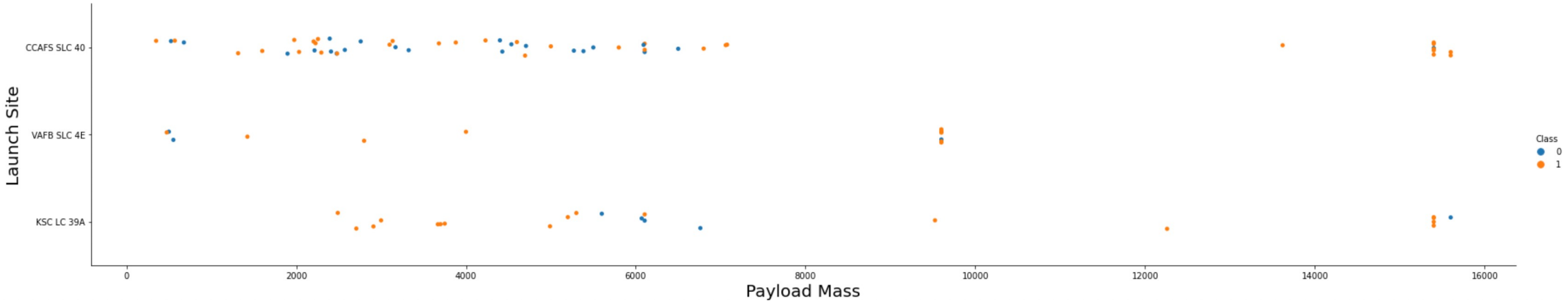
Flight Number vs. Launch Site

- The scatter plot clearly shows the improvement in outcome success with flight number, with the last portion of each launch sequence delivering successful landings every time.



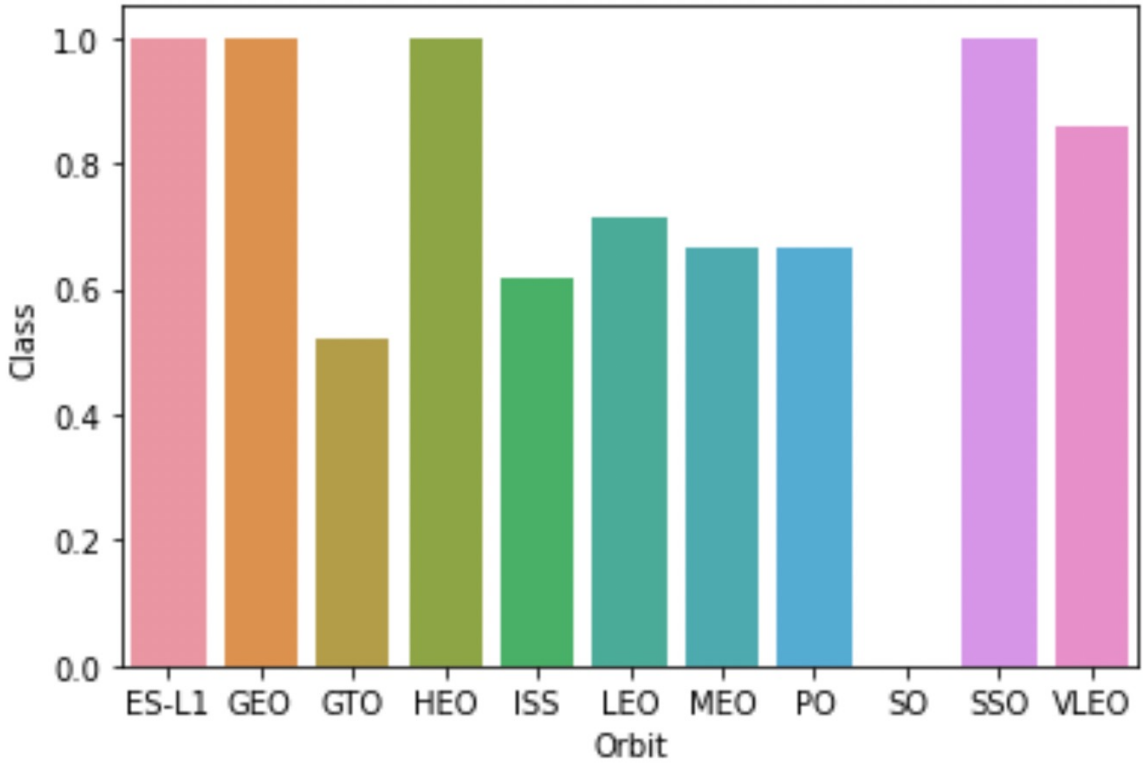
Payload vs. Launch Site

- CCAFS and KSC are capable of launching heavier payloads with successful outcomes than VAFB.
- VAFB appears to limit launches to under 10 tonnes in payload mass



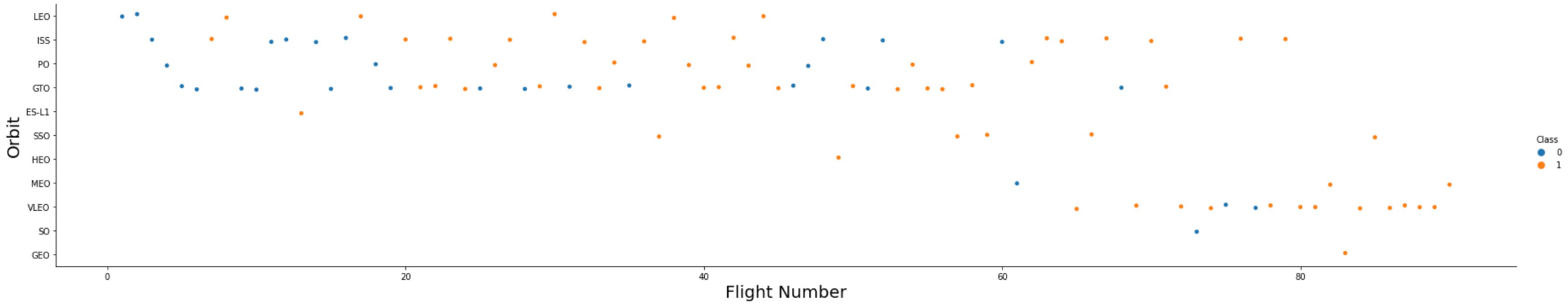
Success Rate vs. Orbit Type

- Four orbits have 100% success rates for landing the first stage successfully
- Further analysis is required to determine if these orbits have different launch profiles which would explain these differences



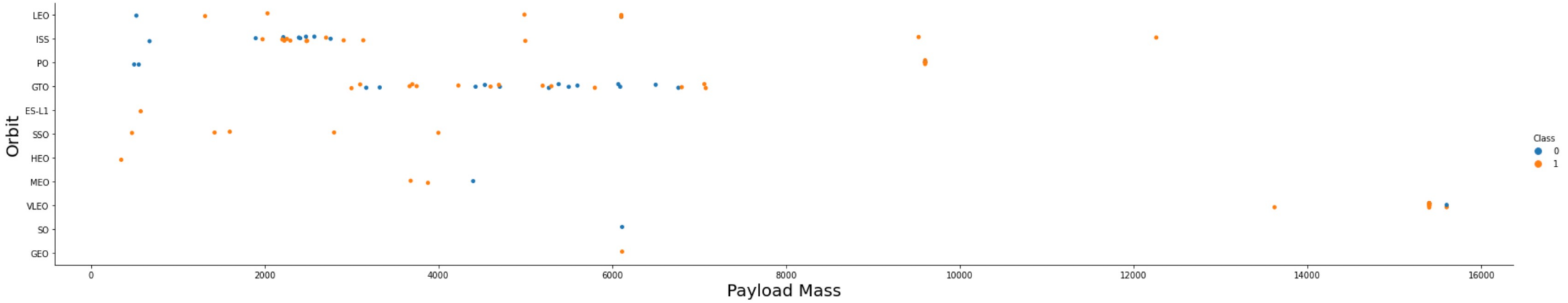
Flight Number vs. Orbit Type

- Over time (flight number), the orbits accessed have broadened in type
- SpaceX have an excellent record over the last 10-15 launches for recovering the first stage, no matter the orbit insertion required



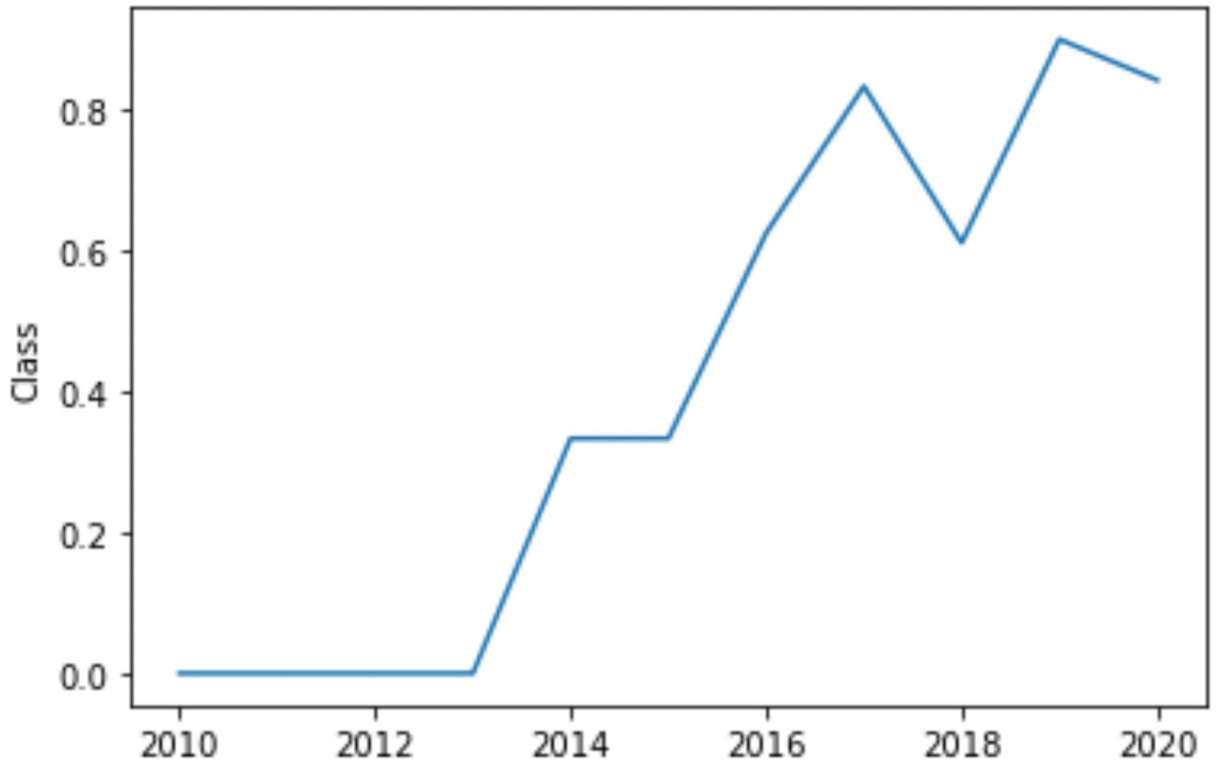
Payload vs. Orbit Type

- Most payloads are still under 7 tonnes. Heavier payloads are delivered exclusively to SO orbit



Launch Success Yearly Trend

- The line plot clearly shows a steady increase in launch/landing success
- SpaceX learn from their failures



All Launch Site Names

```
select distinct launch_site from SPACEXDATASET;
```

Out[17]:

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

Four launch sites output by SQL query

First 5 Launches from CCA sites

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

Out[18]:

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

SQL query pulls first five launches from CCA sites

Total Payload Mass

```
select sum(PAYLOAD_MASS__KG_) as "Payload Mass (kg)" from SPACEXDATASET where CUSTOMER like 'NASA (CRS)';
```

Out[19]:

Payload Mass (kg)
45596

NASA launches total about 45 ½ tonnes of payload mass

Average Payload Mass by F9 v1.1

```
select avg(PAYLOAD_MASS__KG_) as "Ave Payload Mass (kg)"  from SPACEXDATASET  where BOOSTER_VERSION like '%F9 v1.1%';
```

Out[20]:

Ave Payload Mass (kg)
2534

- Average payload mass lofted to orbit on F9 v1.1 booster is about 2½ tonnes

First Successful Ground Landing Date

```
select DATE as "First Successful Landing GroundPad" from SPACEXDATASET where "Landing _Outcome" like 'Success (ground pad)' order by DATE limit 1;
```

Out[21]:

First Successful Landing GroundPad
2015-12-22

First successful landing on a ground pad was fairly recent, in 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
select BOOSTER_VERSION from SPACEXDATASET
                where (PAYLOAD_MASS__KG_ >= 4000 and PAYLOAD_MASS__KG_ < 6000)
                and ("Landing _Outcome" like 'Success (drone ship)');
```

Out[22]:

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

- F9 FT booster shows good success, in this payload range, for drone ship landing

Total Number of Successful and Failure Mission Outcomes

```
select MISSION_OUTCOME as "Mission Outcome", count(*) as "Count" from SPACEXDATASET group by MISSION_OUTCOME;
```

Out[23]:

Mission Outcome	Count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- SpaceX data show their mission outcome success rate is very high

Boosters Carried Maximum Payload

```
select BOOSTER_VERSION as "Booster" from SPACEXDATASET where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEXDATASET);
```

Out[24]:

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

- The F9 B5 booster appears to be the only booster capable of successfully lifting the maximum payload to orbit

2015 Launch Records

```
select DATE as "Date", "Landing _Outcome", BOOSTER_VERSION as "Booster", LAUNCH_SITE as "Launch Site" from SPACEXDATASET
where "Landing _Outcome" like 'Failure (drone ship)' and YEAR(DATE) = '2015';
```

Out[25]:

Date	Landing _Outcome	Booster	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

- All failures to land on a drone ship in 2015 involved the F9 v1.1 booster

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

```
select "Landing _Outcome" , count("Landing _Outcome") as "Count" from SPACEXDATASET
                where (DATE > '2010-06-04' and DATE < '2017-03-20')
                group by "Landing _Outcome"
                order by count desc;
```

Out[70]:

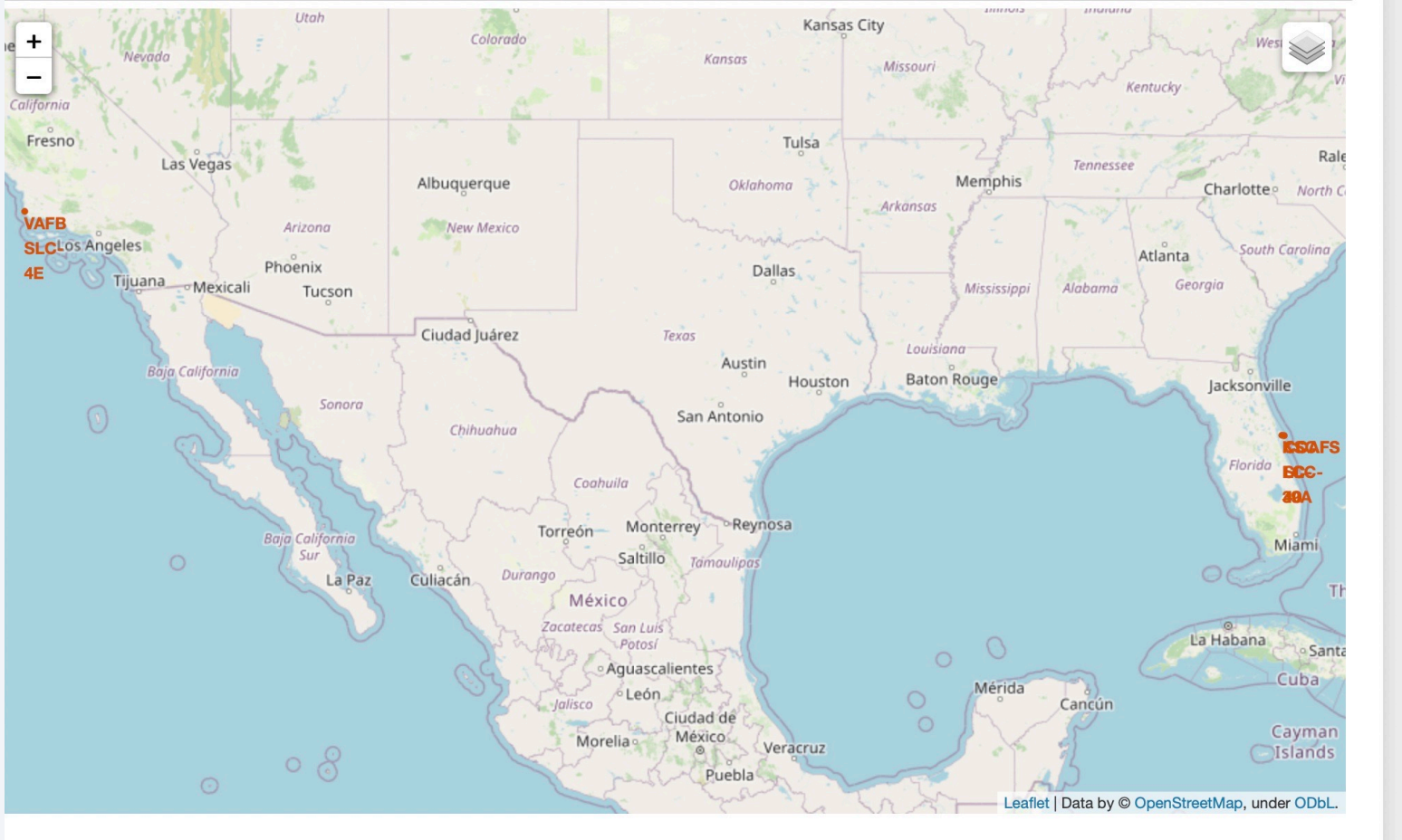
Landing _Outcome	Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

- Successful landing percentages in the period between 2010 and 2017 were poor


Section 3
Launch Sites
Proximities Analysis

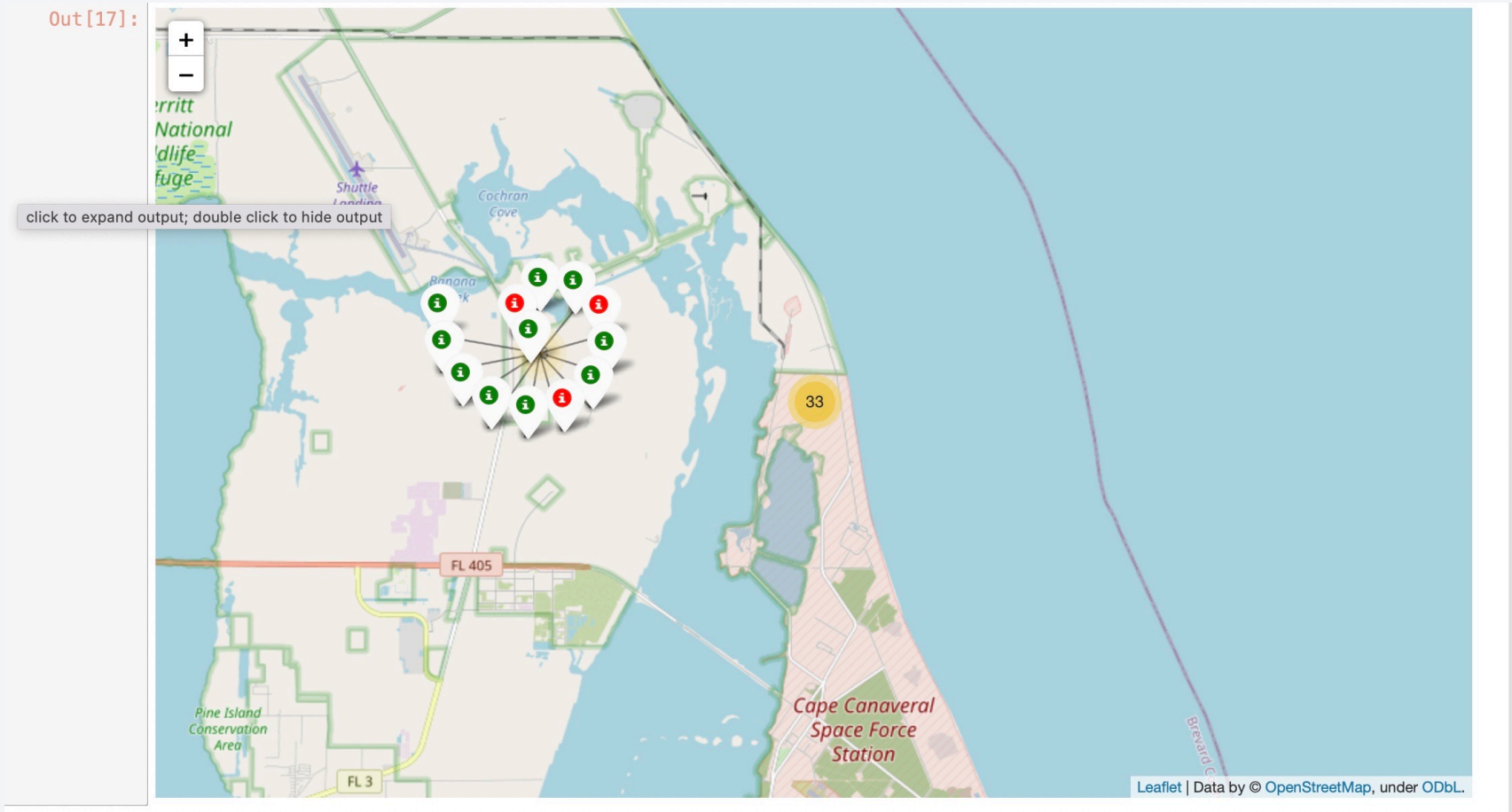
Launch Site Locations

- SpaceX launch sites located in US
- Sites are located on both East and West coasts at existing launch facilities

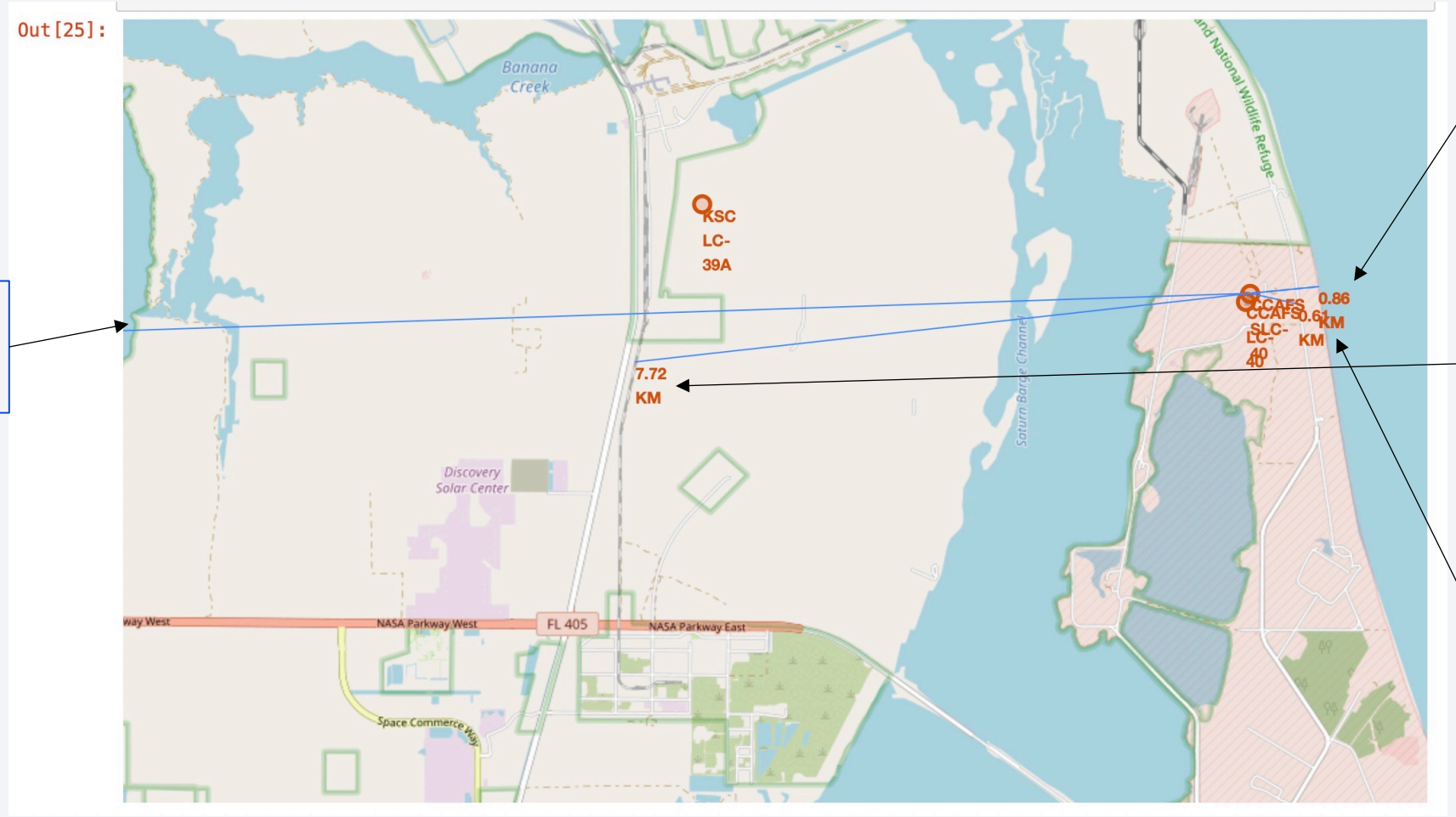


Launch Outcome Markers

- Markers show improving launch success at this site



Launch Site Proximity Analysis

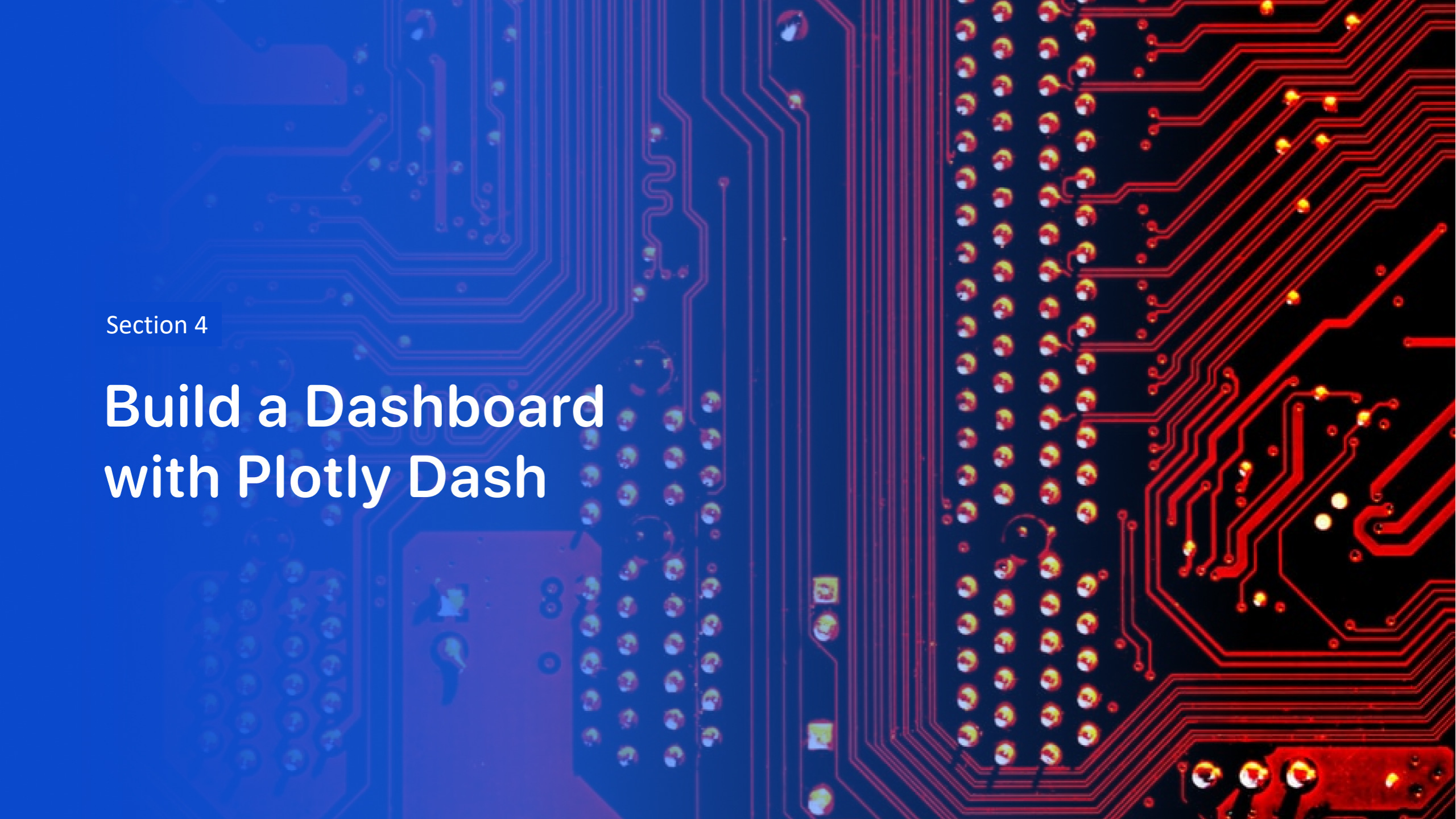


Orlando > 70km:
Population centre at safe distance

Coastline < 1km:
Safer location avoiding population centres

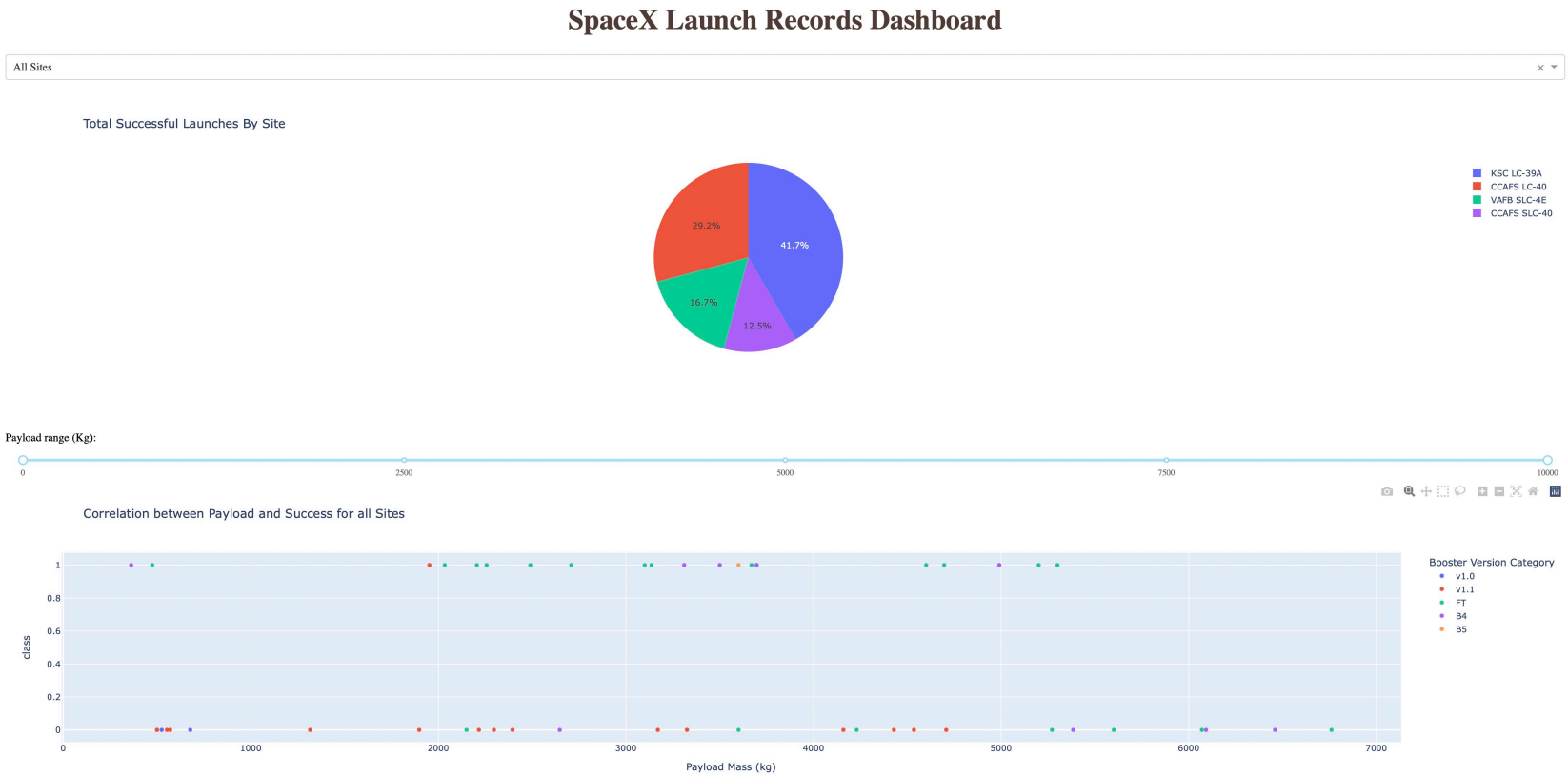
Railway < 10km:
Close access to heavy transportation option for rocket components

Road < 1km:
Close to transportation infrastructure for personnel access


Section 4
Build a Dashboard with Plotly Dash

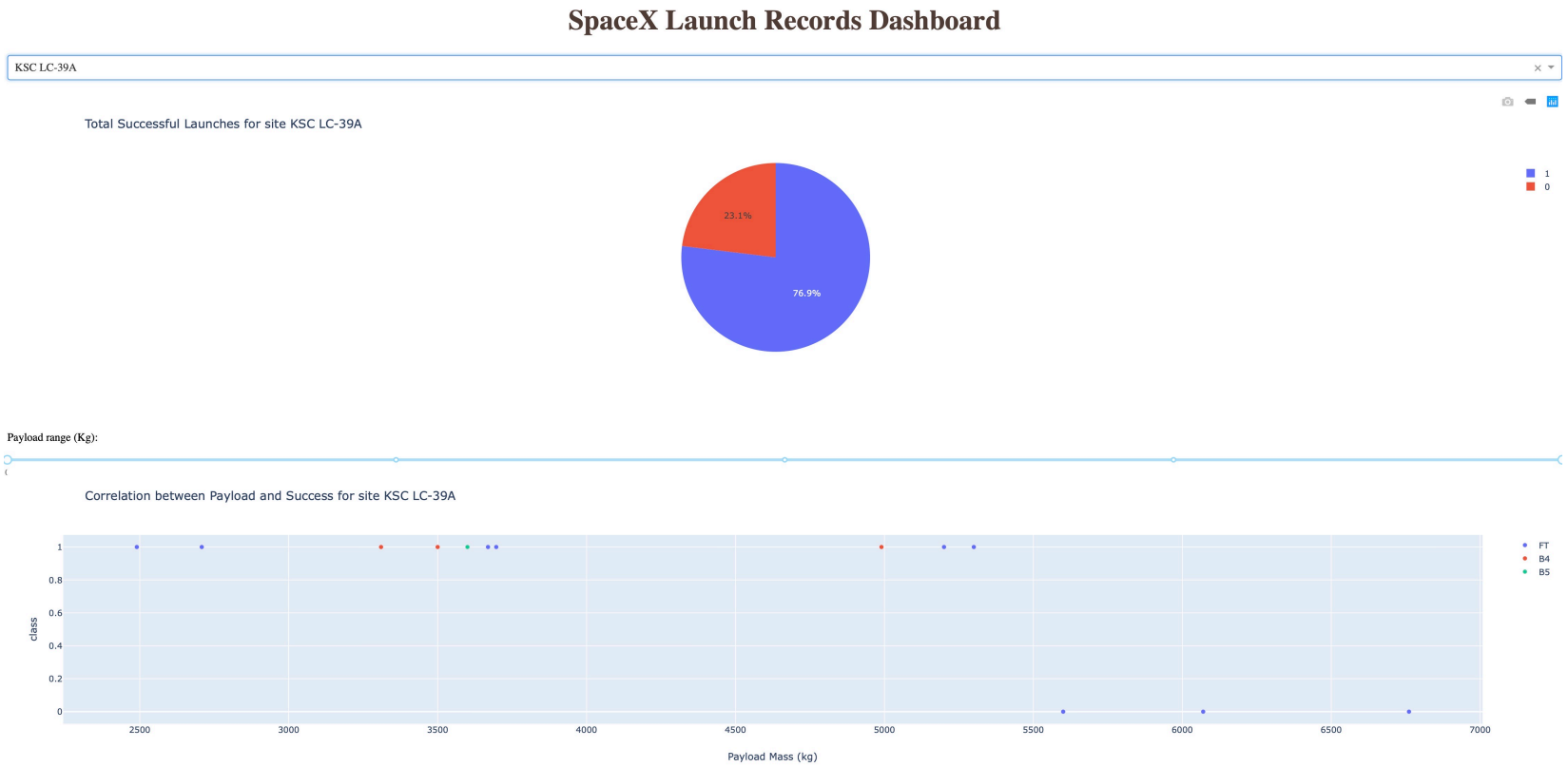
SpaceX Launch Record Overview

- Distribution of success rate as percentage of total launches shows KSC (Kennedy Space Centre) as most successful



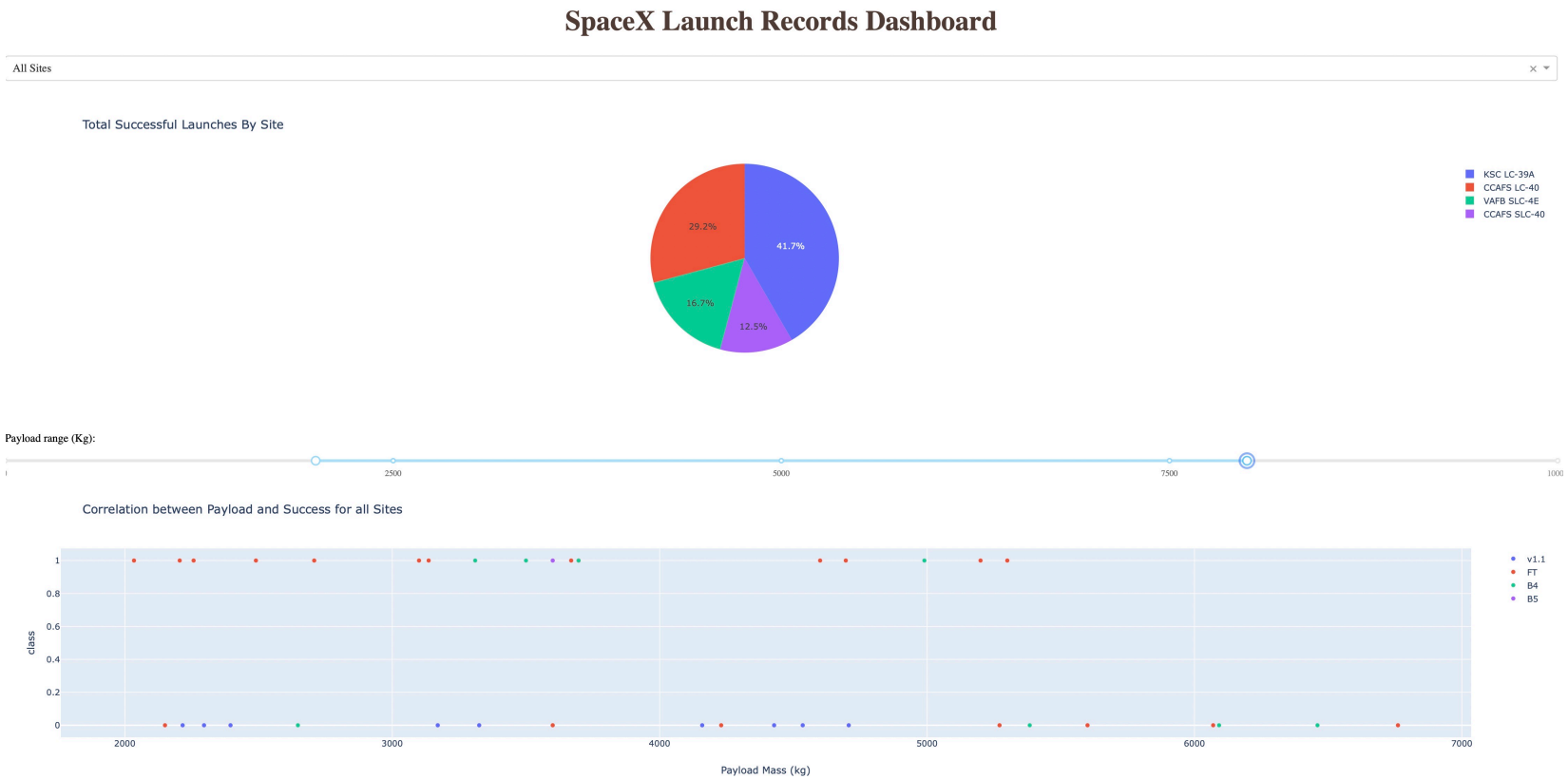
Highest Launch Success Site – KSC LC-39A

- Highest launch success site is Kennedy Space Centre LC-39A
- More than ¾ of launches from this site successfully stuck the landing
- Heavier payloads have been unsuccessful from this site compared to lighter ones



Payload Mass effect on Outcome Success

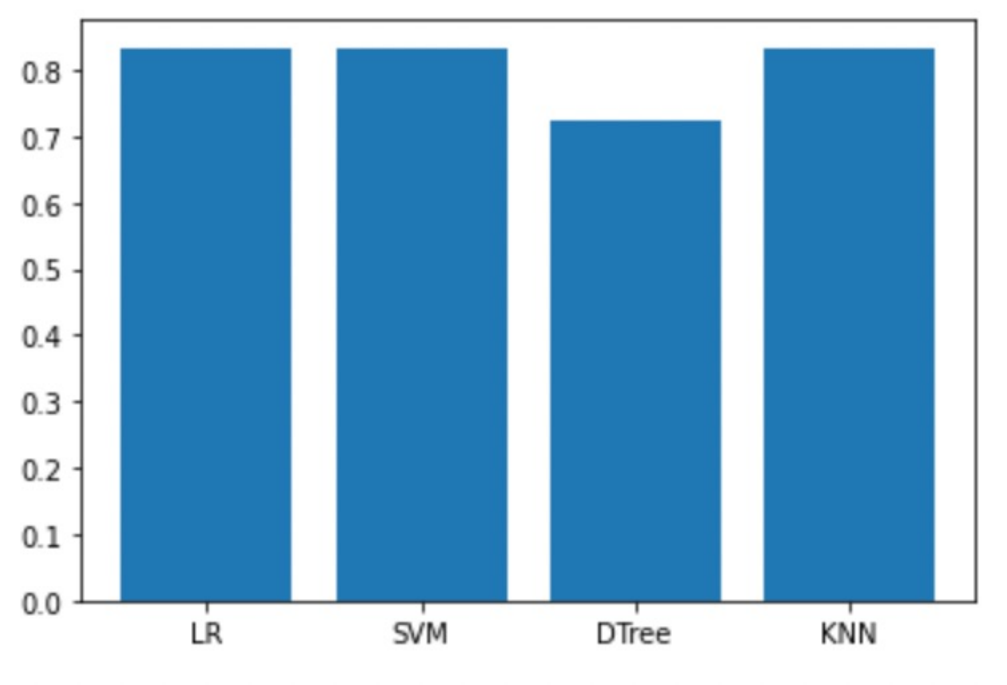
- Lighter payloads have a better success rate
- The FT Booster is the most successful



Section 5

Predictive Analysis (Classification)

Classification Accuracy

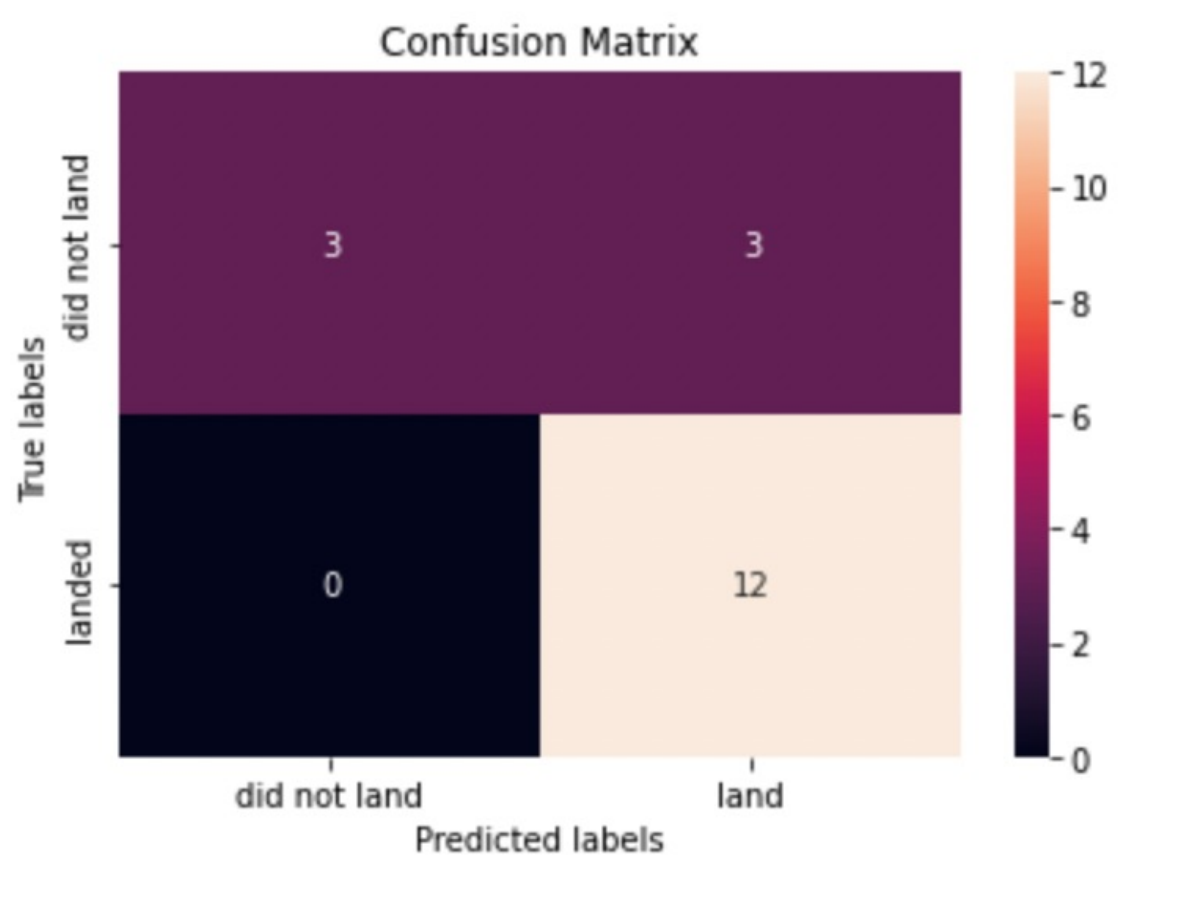


The Logistic Regression model has the highest accuracy score

Find the method performs best:

```
In [41]: max(acc_scores, key=lambda key: acc_scores[key])
Out[41]: 'LR'
```

Confusion Matrix



The model is accurate on prediction of landing success but has a false positive rate of 20%

Conclusions

- A logistic regression model trained on the launch data predicts landing success with greater than 83% accuracy

Thank you!