IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kyle Rattray
30th October 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of results
  - Exploratory Data Analysis results
  - Interactive Analytics in screenshots
  - Predictive Analytics results

Introduction

The aim of this project is to analyse, interpret and predict if the Falcon 9 first stage will successfully land given the publicly available data set.

SpaceX states on its website that a single Falcon 9 rocket launch costs US$62 million where other providers cost upwards of US$165 million each. This price difference is explained by the fact that SpaceX lands, recovers and re-uses the first stage. By determining the probability of the stage successfully landing, we can determine the true cost of a launch.

The major questions to be addressed are:

- What are the main characteristics of a successful or failed landing?
- What are the effects of each relationship of the rocket variables on the success or failure of a landing?
- What are the conditions which will allow SpaceX to achieve the best landing success rate?

Section 1

Methodology

Methodology

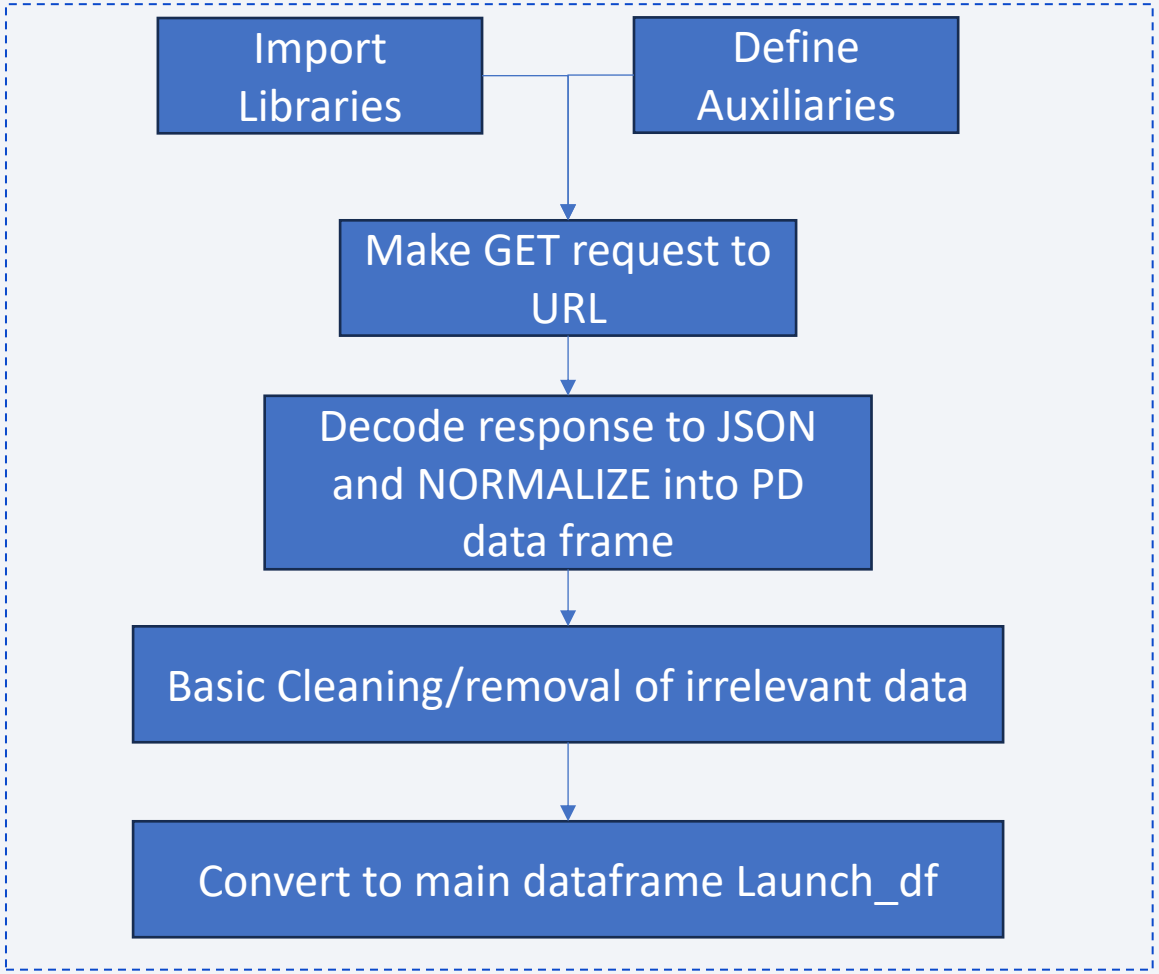
Executive Summary

- Data collection methodology:
  - SpaceX launch data was collected using the SpaceX REST API and subsequently converted from JSON format to data frame. Additional data was gathered via web scraping of Wiki HTML tables using BeautifulSoup and parsed into a dataframe.
- Perform data wrangling
  - Removal of NULL values from LandingPad records and replacing missing values in the PayloadMass data with mean values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Hyperparameter testing for SVM, Classification Tree and Logistic Regression and selection of best performer.

Data Collection – SpaceX API

1. Import relevant python libraries.
2. Define auxiliary functions to extract API data.
3. Make GET request from API.
4. Decode respone to JSON and normalize to a Pandas dataframe.
5. High level cleaning of data and push relevant details only to dictionary.
6. Conversion to Launch_df dataframe.

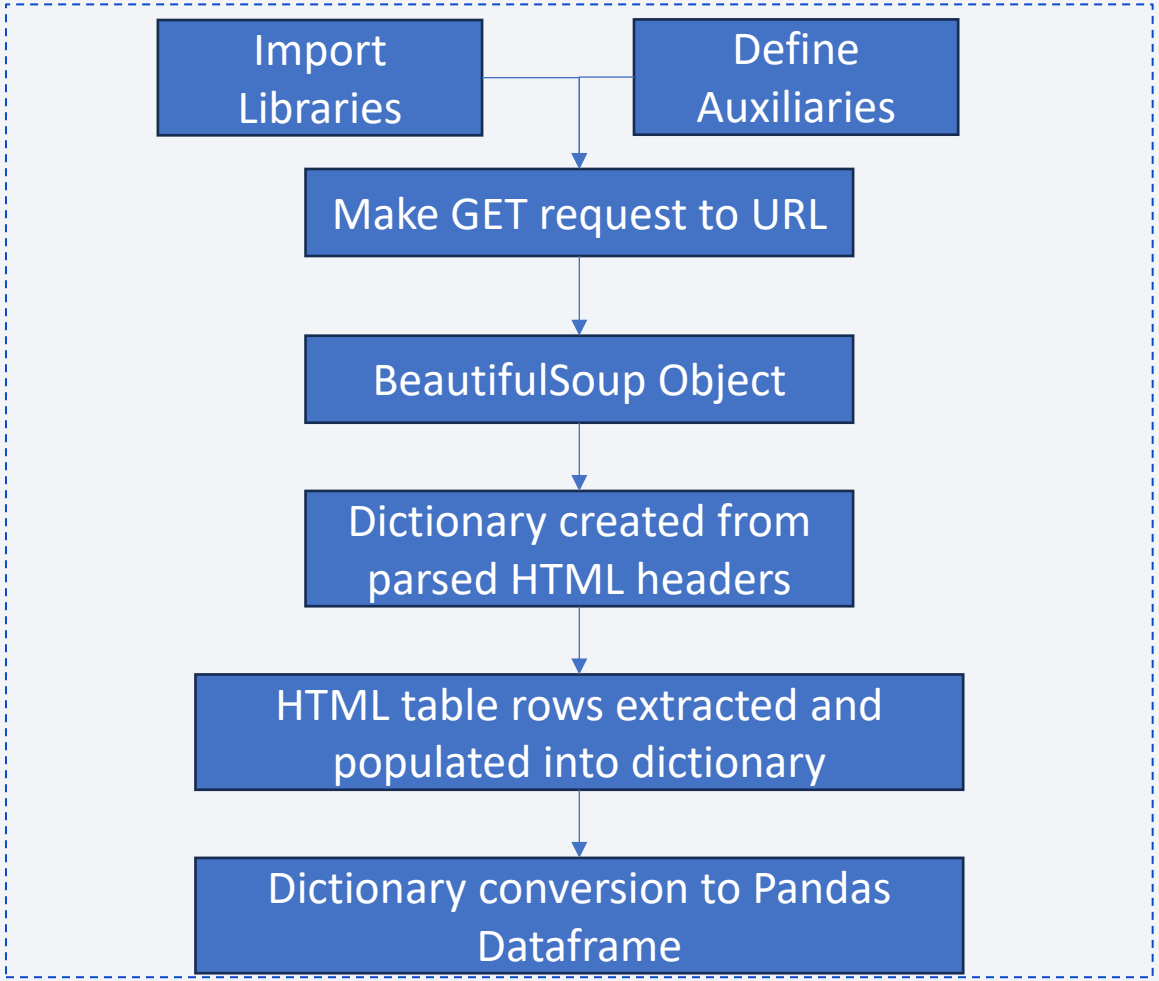
- GitHub Page For Data Collection



Data Collection - Scraping

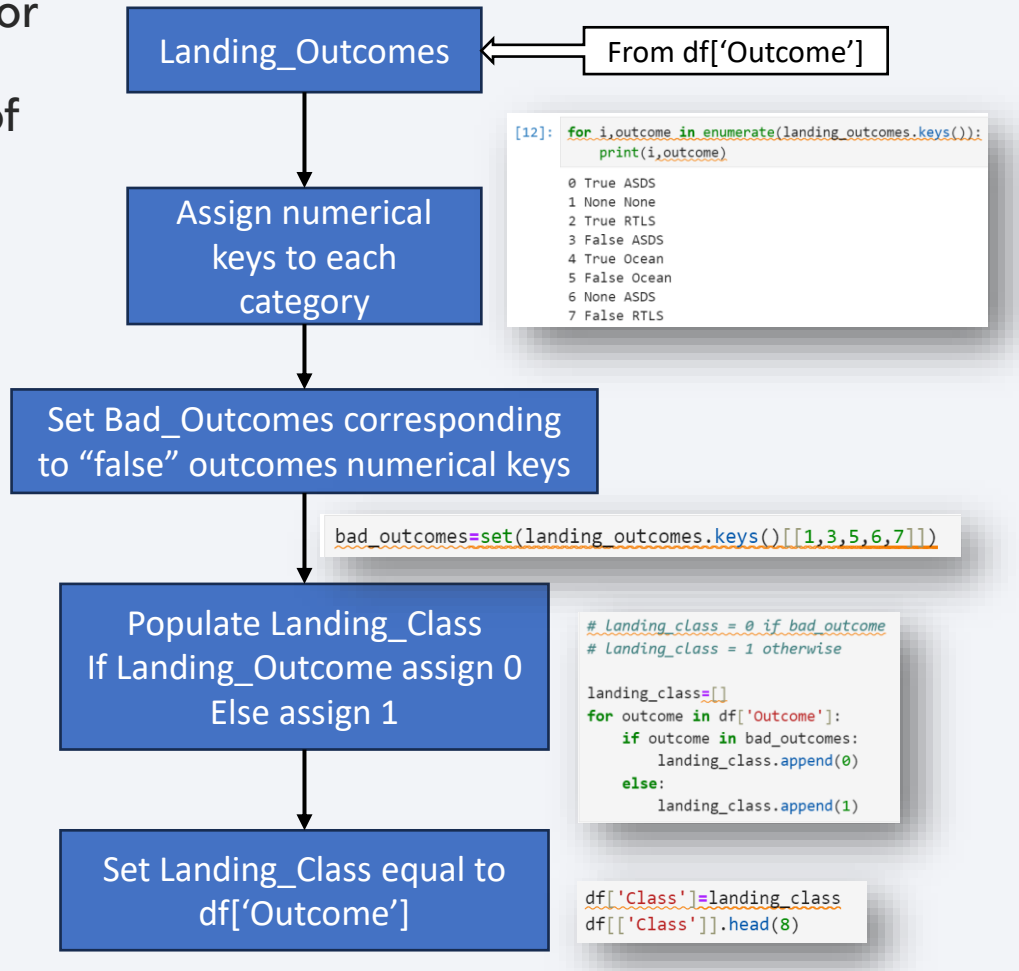
1. Import relevant Python packages
2. Define helper functions.
3. Make GET request and create a BeautifulSoup object from the response.
4. Extract HTML table header columns and parse to dictionary.
5. Extract table rows and populate to dictionary.
6. Create final Pandas DataFrame object

- GitHub URL for Web Scraping



Data Wrangling

- It is desirable to simplify the landing outcome into a success or failure category. The dataset contains six possible outcome categories depending on success/failure condition and type of landing location.
  - True/False Ocean
  - True/False RTLS
  - True/False ASDS
- These conditions can be simplified to a binary response:
  - 0 indicates failure
  - 1 indicates success
- Data Wrangling GitHub - Lab 2: Data wrangling



EDA with Data Visualization

EDA using data visualization provides graphical clues as to what factors may be further studied to develop the prediction model. Graphical plots produced are as follows:

1. Categorical Plot – Payload Mass Vs. Success Category
   - Does the payload mass impact the successful landing of the first rocket stage?
2. Categorical Plot – Flight Number Vs. Launch Site
   - How many flights are launched from each site?
3. Categorical Plot – Payload Mass Vs. Launch Site
   - What is the typical payload of flights by their launch site?
4. Bar Chart – Success Rate Vs. Orbit Type
   - How successful are launches to differing orbit types?
5. Categorical Plot – Payload Mass Vs. Orbit Type
   - What payload masses are sent to different orbit types?
6. Line Chart – Success Rate Vs. Year
   - How does success rate change over years of operation?

- EDA GitHub - Exploring and Preparing Data

EDA with SQL

Exploratory data analysis of the data set was performed via SQL query to address the following:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
9. List the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015.
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.

- SQL EDA GitHub - SQL Notebook

Build an Interactive Map with Folium

- Blue circles – Indicate the launch site locations
- Markers – Indicates the launch site name
- Marker Cluster – indicates successful or failed launches by colour
- Polylines – Drawn to indicate distance between launch sites and various nearby POI

- Folium Mapping GitHub - Launch Sites Locations Analysis with Folium

Build a Dashboard with Plotly Dash

- Plots and graphs added to the dashboard include a pie chart to demonstrate launch success and a scatter chart to demonstrate correlation between payload and launch success.
- Interactions included in the dashboard include a drop down list to select the launch site and a slider to filter the payload mass.

- Plotly Dashboard GitHub - Plotly Dash Code

Predictive Analysis (Classification)

Model Preparation [Start]

Import Libraries and Auxiliary function

Load data frame

Create Numpy array corresponding to launch success classification. [Variable = Y]

Standardise the data frame [Variable = X]

Separate the X and Y variable data into test and train sets using the train_test_split function. [test_size = 0.2, random_state = 2]

Model Preparation Complete [End]

Logistic Regression Object

```
logreg_cv = GridSearchCV(lr, parameters, cv = 10)
```

Support Vector Machine

```
svm_cv = GridSearchCV(svm, parameters, cv = 10)
```

Decision Tree Method

```
tree_cv = GridSearchCV(tree, parameters, cv=10)
```

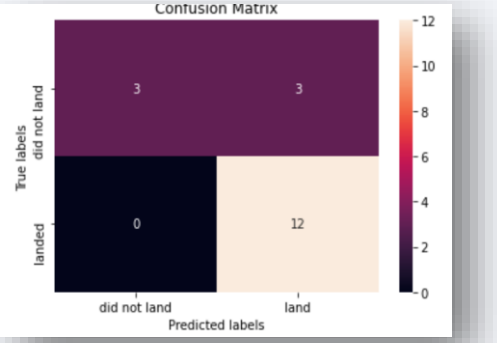
Kth Nearest Neighbour

```
knn_cv = GridSearchCV(KNN, parameters, cv=10)
knn_cv.fit(X_train,Y_train)
```

Return hyperparameters and return accuracy

Check test data accuracy with score method

Output confusion matrix



Summarise results to find best performing model

- Predictive Analysis GitHub - Predictive Analysis

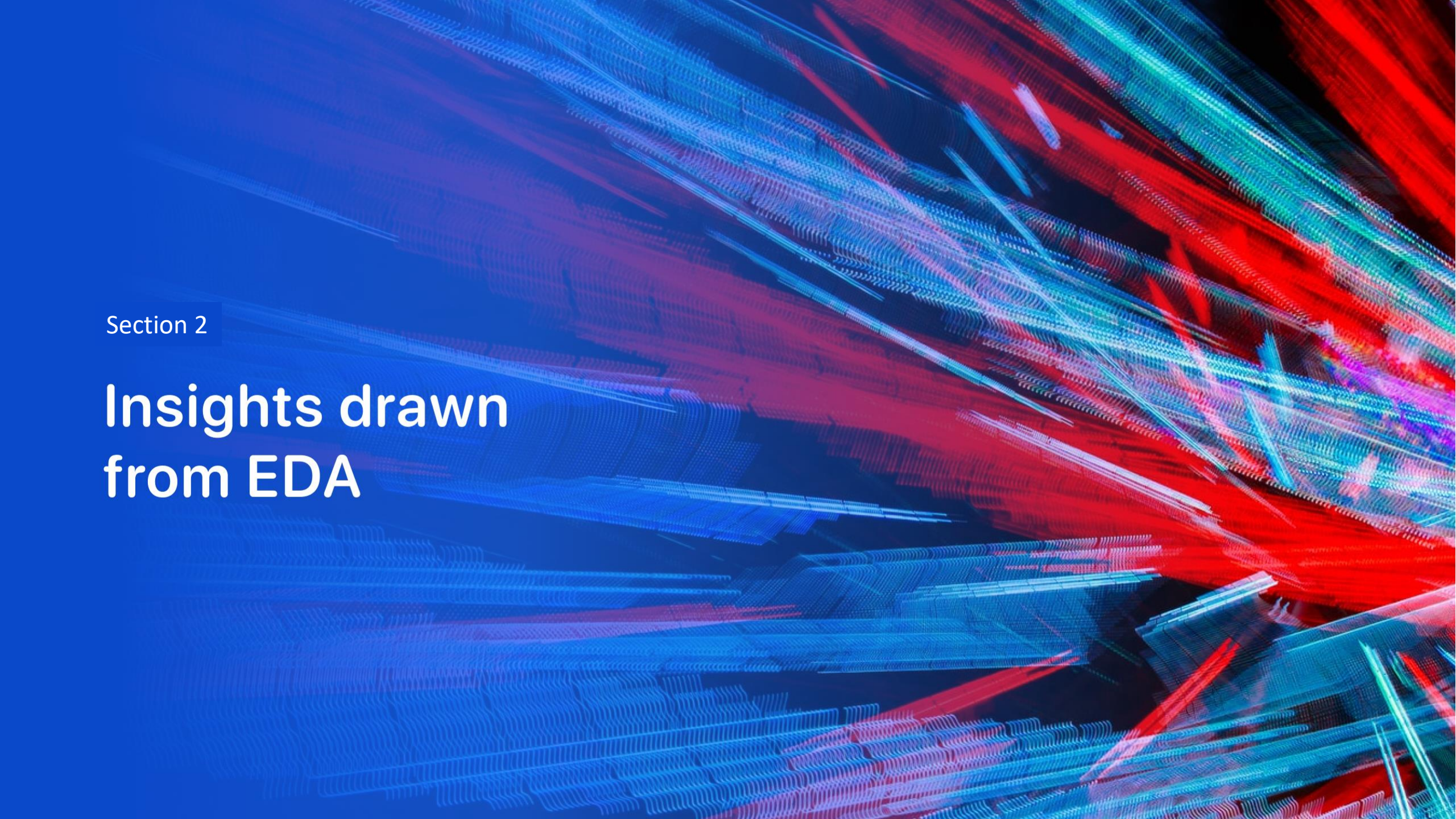
Results

Exploratory data analysis results

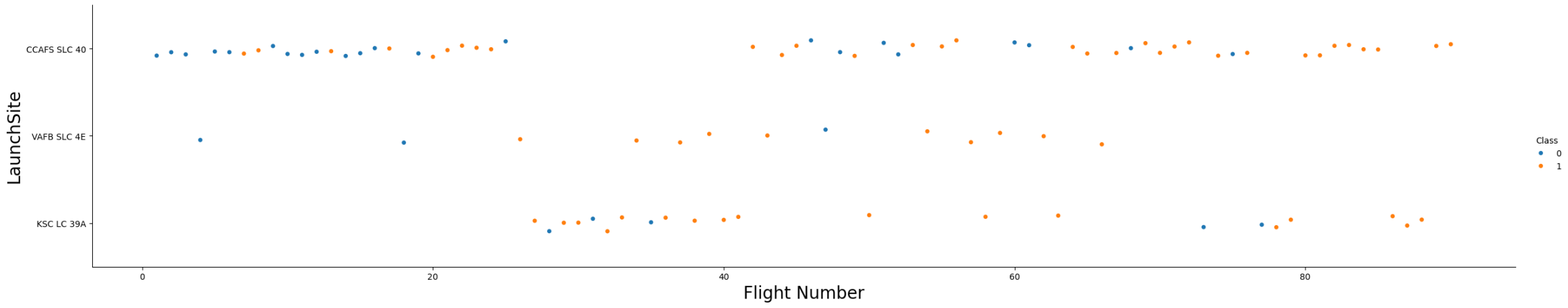
- Flight number Vs Payload Mass
  - Early flights with low payload have a higher failure rate. There is a trend of failure as payload increases . This trend drops away from around flight 45.
- Flight number Vs Launch site
  - CCAFS SCL 40 appears to be the home launch site and features a consistent flight schedule with exception of a period between flight 25 and 42 where launches were majority from KSC LC 39A and shows steady ongoing use. VAFB SLS 4E has steady use through flights 20 to 65 after which it is no longer utilised.
- Payload Mass Vs Launch Site
  - Payloads up to 10000kg can be launched from any site however larger payloads must be launched from CCAFS SLC 40 or KSC LC 39A. This indicates that VAFB SLC 4E may not be capable of launching high payload flights. There are significantly fewer launches above 10000kg but a small cluster sitting within the 15500kg to 16000kg range.
- Success Rate and Orbit Type
  - ES-L1, GEO, HEO and SSO orbit launches had 100% success rates. VLEO success rate was 90% and the remaining orbit types fell within the range of 50% to 75% success.
- Flight Number Vs Orbit Type
  - The majority of launches were to GTO, ISS and LEO orbits. The most risk orbit type appears to be LEO where failure rate is consistent throughout, LEO orbits have been mainly successful after the early flights. There is a cluster of VLEO orbit launches from flight 82 onwards. This type of orbit is the most common in the latter part of the dataset.
- Payload Mass Vs Orbit Type
  - There is a clear banding in the launch payload compared to orbit type. ISS launches fall between 1800kg and 3500kg, GTO launches fall between 3500kg and 7500kg.
- Launch Success Yearly Trend
  - There is a positive growth in success rate from 2013 through to 2020. There is a period between 2014 and 2015 where success rate was steady, a period between 2017 and 1018 where success rate declined and another period from 2019 to 2020 where success rate declines.

Results

- Interactive analytics demo in screenshots
- Predictive analysis results

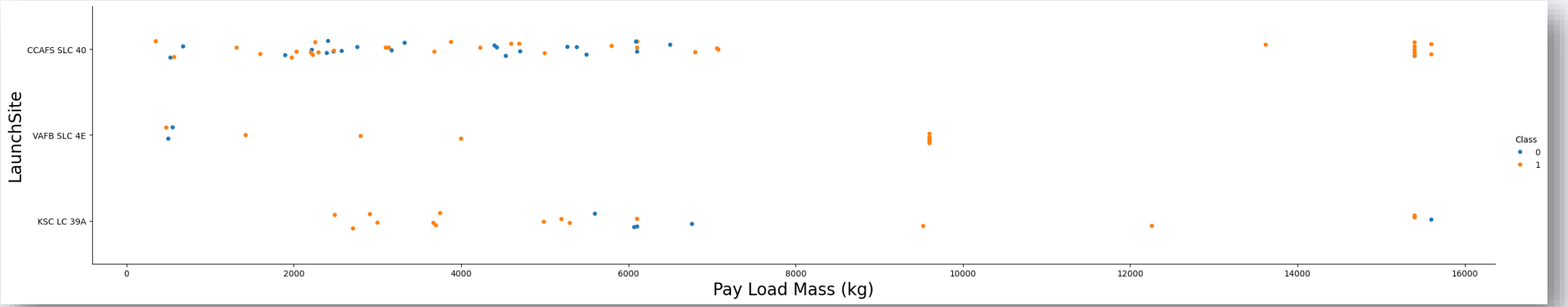

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



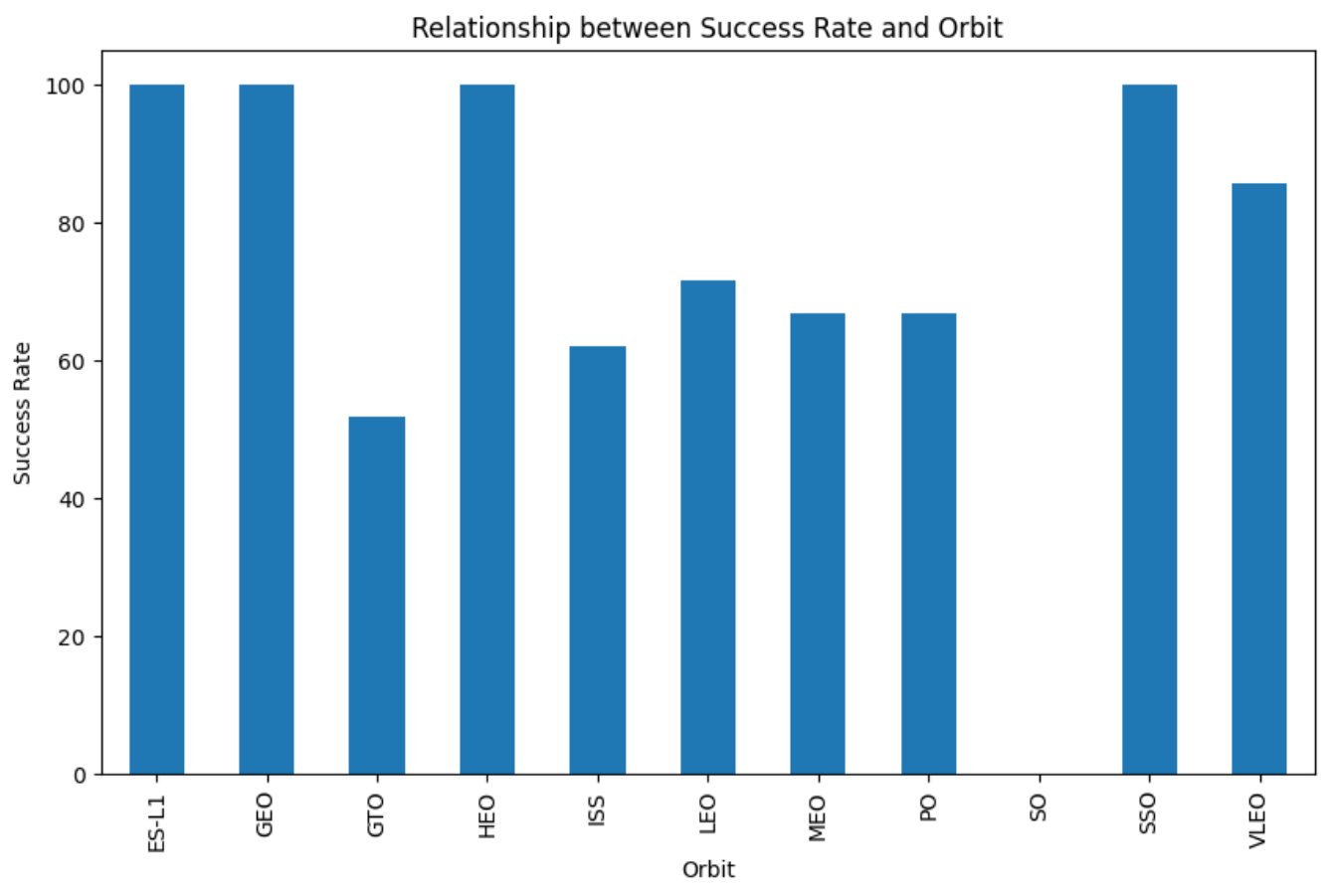
- CCAFS SCL 40 appears to be the home launch site and features a consistent flight schedule with exception of a period between flight 25 and 42 where launches were majority from KSC LC 39A and shows steady ongoing use. VAFB SLS 4E has steady use through flights 20 to 65 after which it is no longer utilised.

Payload vs. Launch Site



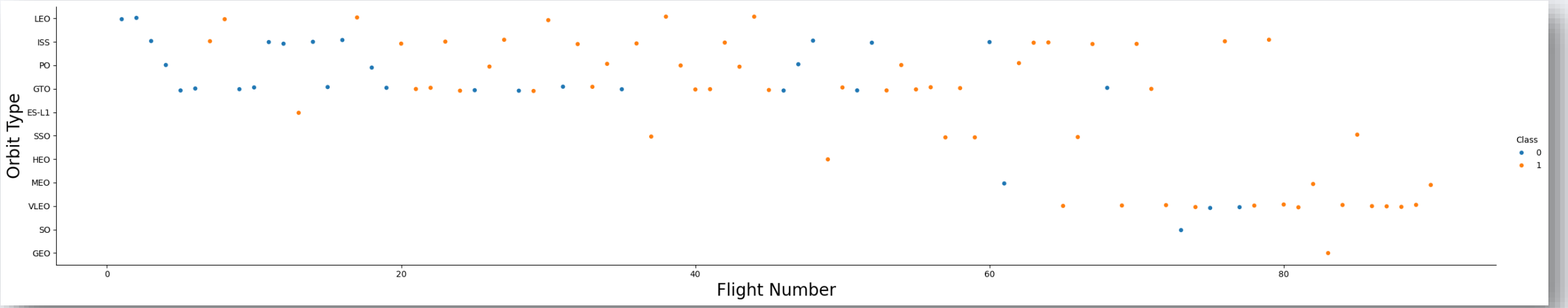
- Payloads up to 10000kg can be launched from any site however larger payloads must be launched from CCAFS SLC 40 or KSC LC 39A. This indicates that VAFB SLC 4E may not be capable of launching high payload flights. There are significantly fewer launches above 10000kg, but a small cluster sit within the 15500kg to 16000kg range.

Success Rate vs. Orbit Type



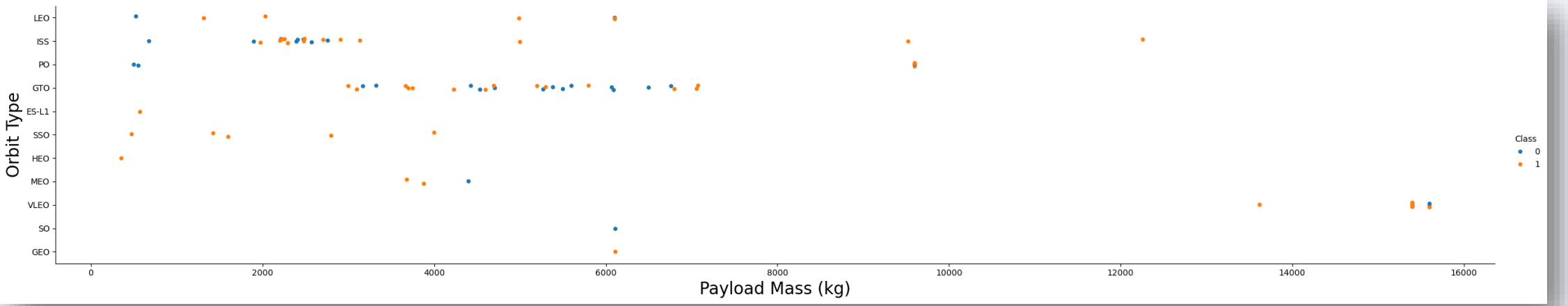
- ES-L1, GEO, HEO and SSO orbit launches had 100% success rates.
- VLEO success rate was 90%.
- The remaining orbit types fell within the range of 50% to 75% success.

Flight Number vs. Orbit Type



- The majority of launches were to GTO, ISS and LEO orbits. The riskiest orbit type appears to be LEO where failure rate is consistent throughout, LEO orbits have been mainly successful after the early flights. There is a cluster of VLEO orbit launches from flight 82 onwards. This type of orbit is the most common in the latter part of the dataset.

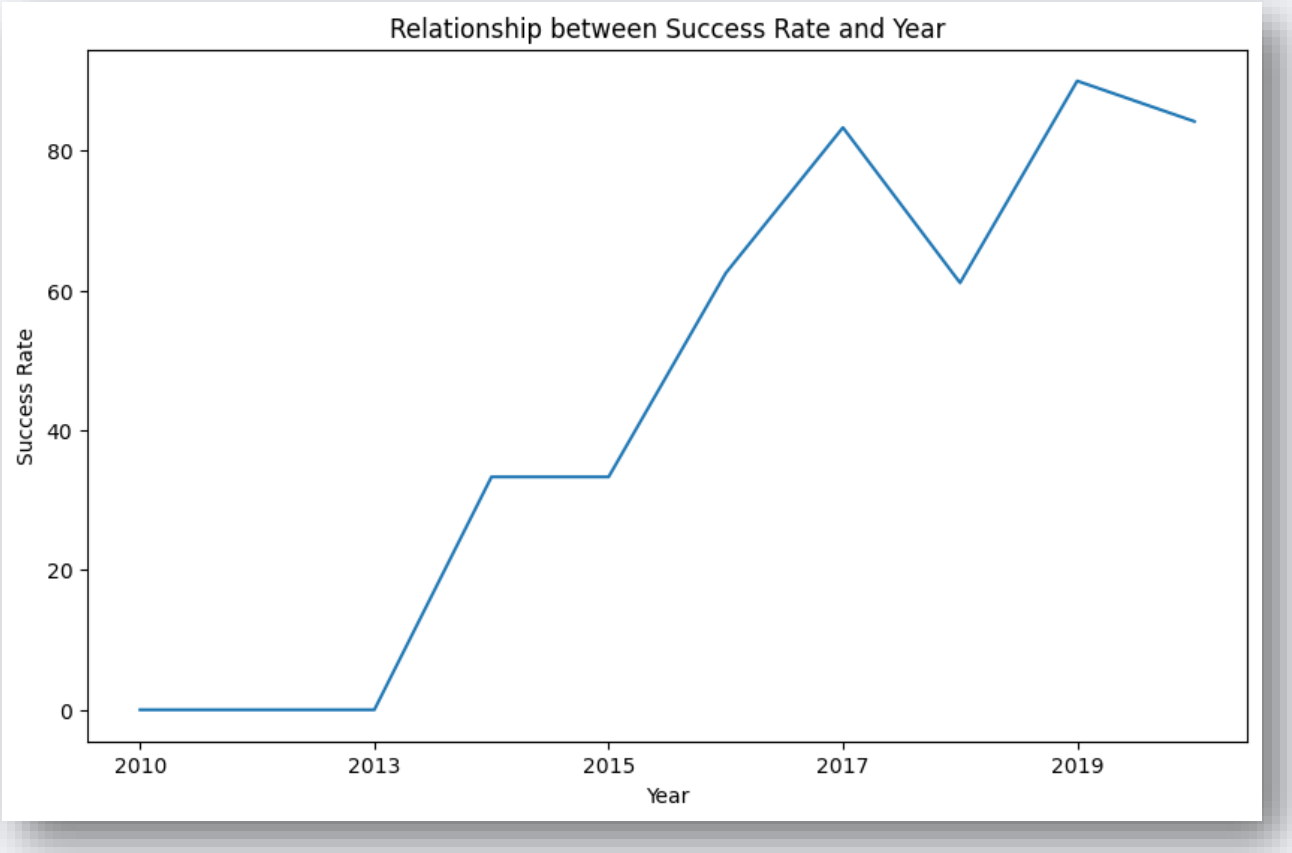
Payload vs. Orbit Type



- There is a clear banding in the launch payload compared to orbit type.
- ISS launches fall between 1800kg and 3500kg,
- GTO launches fall between 3500kg and 7500kg.

Launch Success Yearly Trend

- There is overall growth in success rate from 2013 through to 2020.
- There is a period between 2014 and 2015 where success rate was steady
- Success rate declined during a period between 2017 and 2018 and again between 2019 to 2020.



All Launch Site Names

```
%sql select distinct Launch_Site from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

- The SpaceX data frame was imported to a sqlite data base. This allows SQL queries to be run and important information to be extracted.
- The DISTINCT query only returns the unique values held within the Launch_Site column of the SPACEXTABLE

Launch Site Names Begin with 'CCA'

```
%sql select Launch_Site from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- Using the LIKE syntax allows returning partial or wildcard searches to be performed. In this case it's used to return only records beginning with CCA.
- The result here is limited to the first 5 records by addition of the limit clause.

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
45596

- The SUM function returns the total numeric value of the enclosed column. The WHERE clause is used to limit the summed records to only those where the customer is NASA (CRS).

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 V1.1%'
```

```
 * sqlite:///my_data1.db
Done.
```

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- The AVG function returns the average numeric value of the enclosed column. The WHERE clause is used to limit the summed records to only those where the Booster_Version contains “F9 V1.1”

First Successful Ground Landing Date

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome like 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
2015-12-22

- Using the MIN function to return the oldest date. The WHERE clause limits the results to Landing_Outcomes and the LIKE operator further limits results to only records containing “Success (Ground pad)”.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ between 4000 and 6000
```

Booster_Version	PAYLOAD_MASS__KG_
F9 v1.1	4535
F9 v1.1 B1011	4428
F9 v1.1 B1014	4159
F9 v1.1 B1016	4707
F9 FT B1020	5271
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1030	5600
F9 FT B1021.2	5300
F9 FT B1032.1	5300
F9 B4 B1040.1	4990
F9 FT B1031.2	5200
F9 B4 B1043.1	5000
F9 FT B1032.2	4230
F9 B4 B1040.2	5384
F9 B5 B1046.2	5800
F9 B5 B1047.2	5300
F9 B5 B1046.3	4000
F9 B5B1054	4400
F9 B5 B1048.3	4850
F9 B5 B1051.2	4200
F9 B5B1060.1	4311
F9 B5 B1058.2	5500
F9 B5B1062.1	4311

- SELECT Booster_Version and PAYLOAD_MASS_KG limits the results to these columns, the WHERE clause limits the results to only those records falling between 4000kg and 6000kg.

Total Number of Successful and Failure Mission Outcomes

```
%sql select Landing_Outcome, count(Landing_Outcome) as count from SPACEXTBL group by Landing_Outcome
```

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

- Results are restricted to the Landing_Outcome column and the number of those types of events [COUNT(Landing_Outcome)]. Outputing the types of outcome and the count of those event categories.

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- The top-level query draws the Booster_Version and PAYLOAD_MASS_KG columns from the database table SPACEXTBL and the sub query limits the returned results to only those where the PAYLOAD_MASS_KG is equal to the maximum value within the available range.

2015 Launch Records – Drone Ship Failed Landing

```
%%sql select substr(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL
where substr(Date,0,5)='2015' and Landing_Outcome like 'Failure%'
```

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

- Using the date substring to return month the Launch_Site, Booster_Version and Landing_Outcome are drawn from the table. The WHERE clause uses the date substring again to control the year to only those in 2015 and further restricts the results to those with a Landing_Outcome containing ‘Failure’

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

```
%%sql

SELECT Landing_Outcome, COUNT(*) AS COUNT_LAUNCHES FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT_LAUNCHES DESC;
```

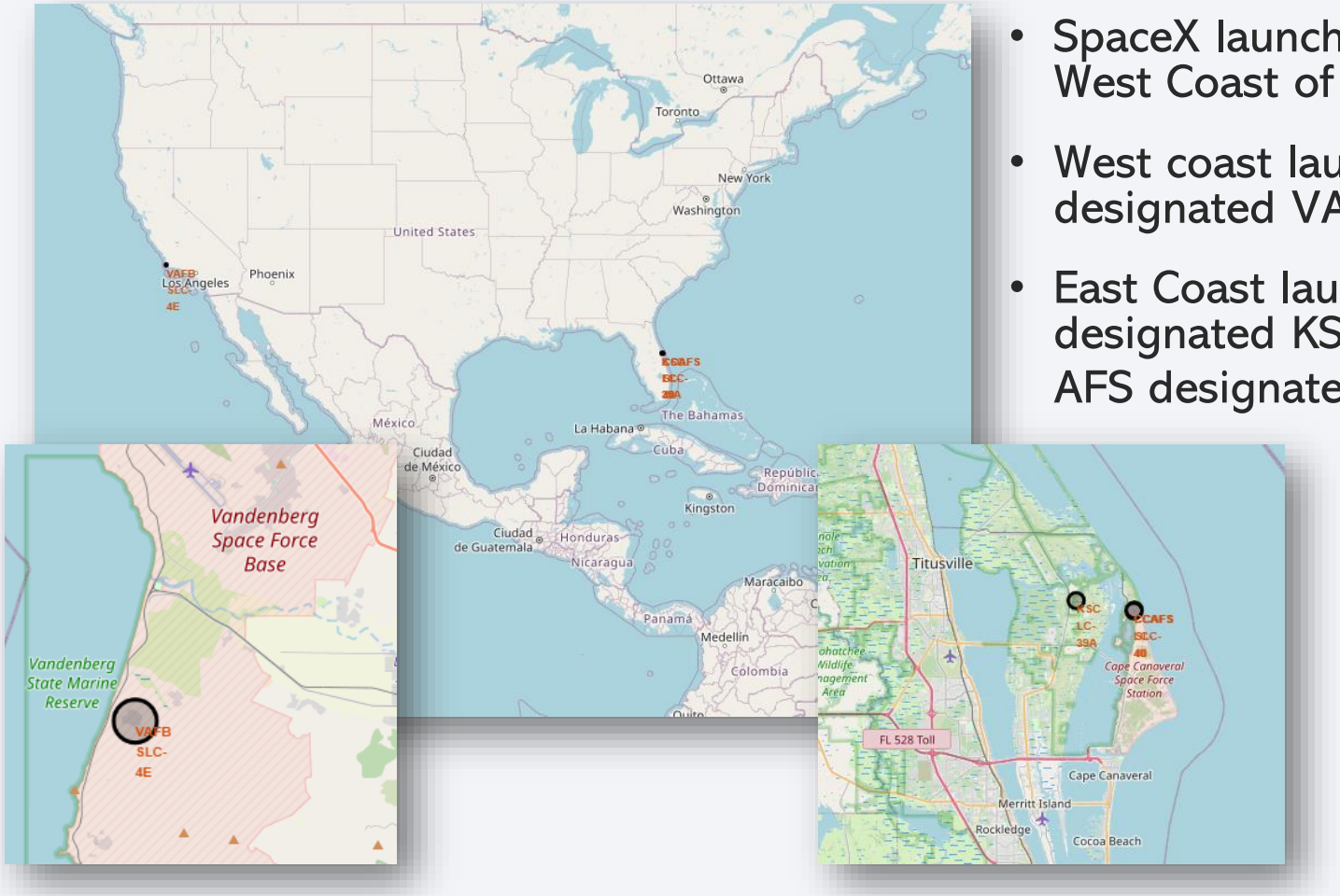
Landing_Outcome	COUNT_LAUNCHES
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Landing outcomes categories are grouped and presented in descending rank order.
- These results are limited to dates between 2010-06-04 and 2017-03-20

Section 3

Launch Sites Proximities Analysis

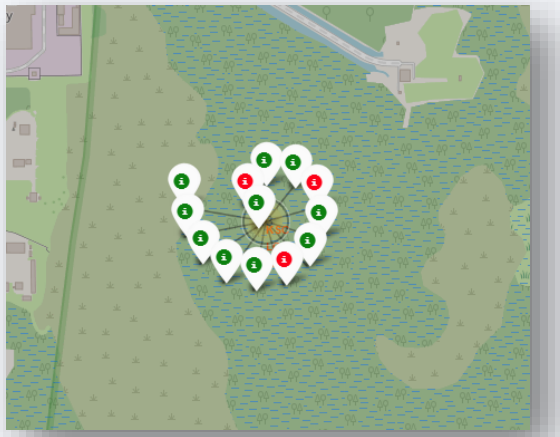
Launch Site Locations



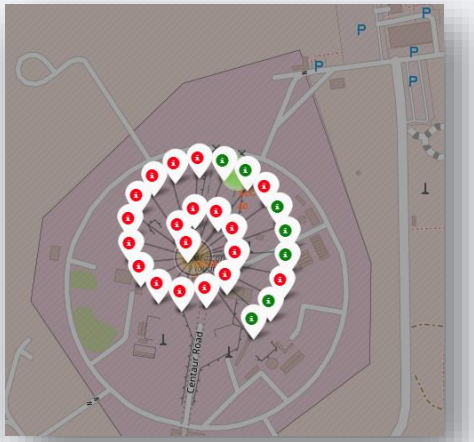
- SpaceX launch site are located on the East and West Coast of the USA.
- West coast launches are from Vandenburg AFB designated VAFB-SLC-4E
- East Coast launches are from Kennedy Space Centre designated KSC-LC-39A and from Cape Canavarel AFS designated CCAFS-SLC-40 & CCAFS-LC-40

Launch Site – Graphical Landing Success

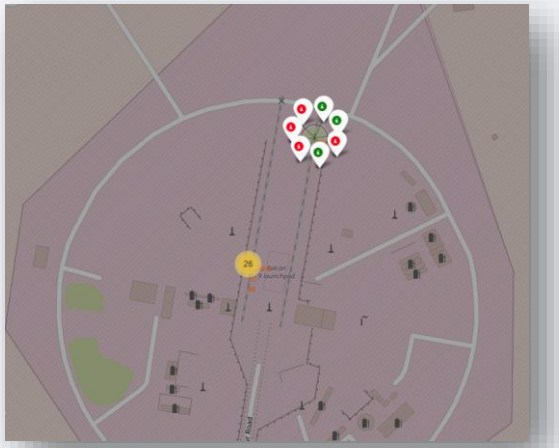
VAFB-SLC-4E



KSC-LC-39A



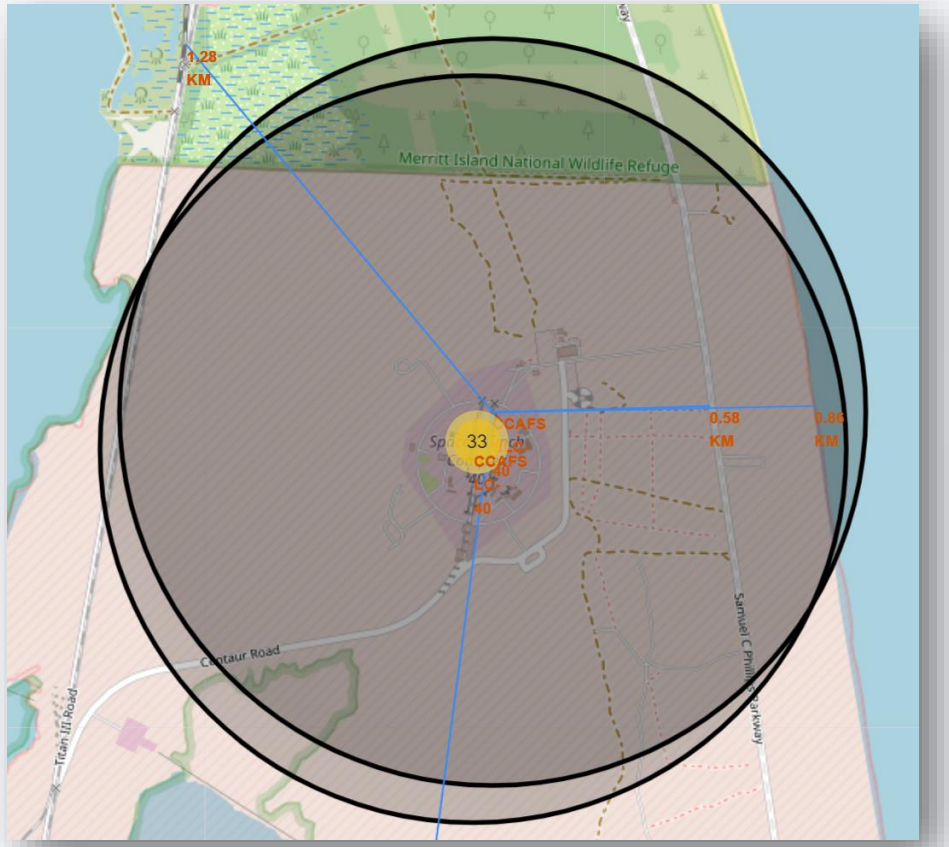
CCAFS-LC-40



CCAFS-SLC-40

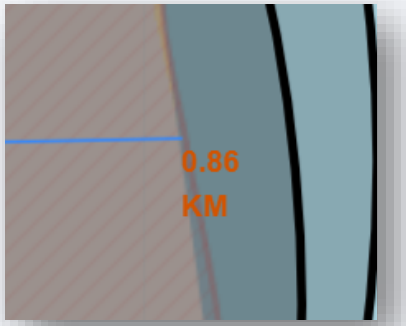
- Each of the individual launch sites are shown at their geographic locations. When the icon is selected the successful or unsuccessful launches are displayed with additional coloured icons, red indicating unsuccessful landings and green indicating successful landings.

Launch Sites – Distance to POI

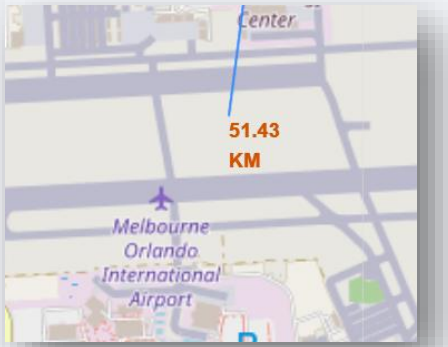


Nearest railway

Nearest highway

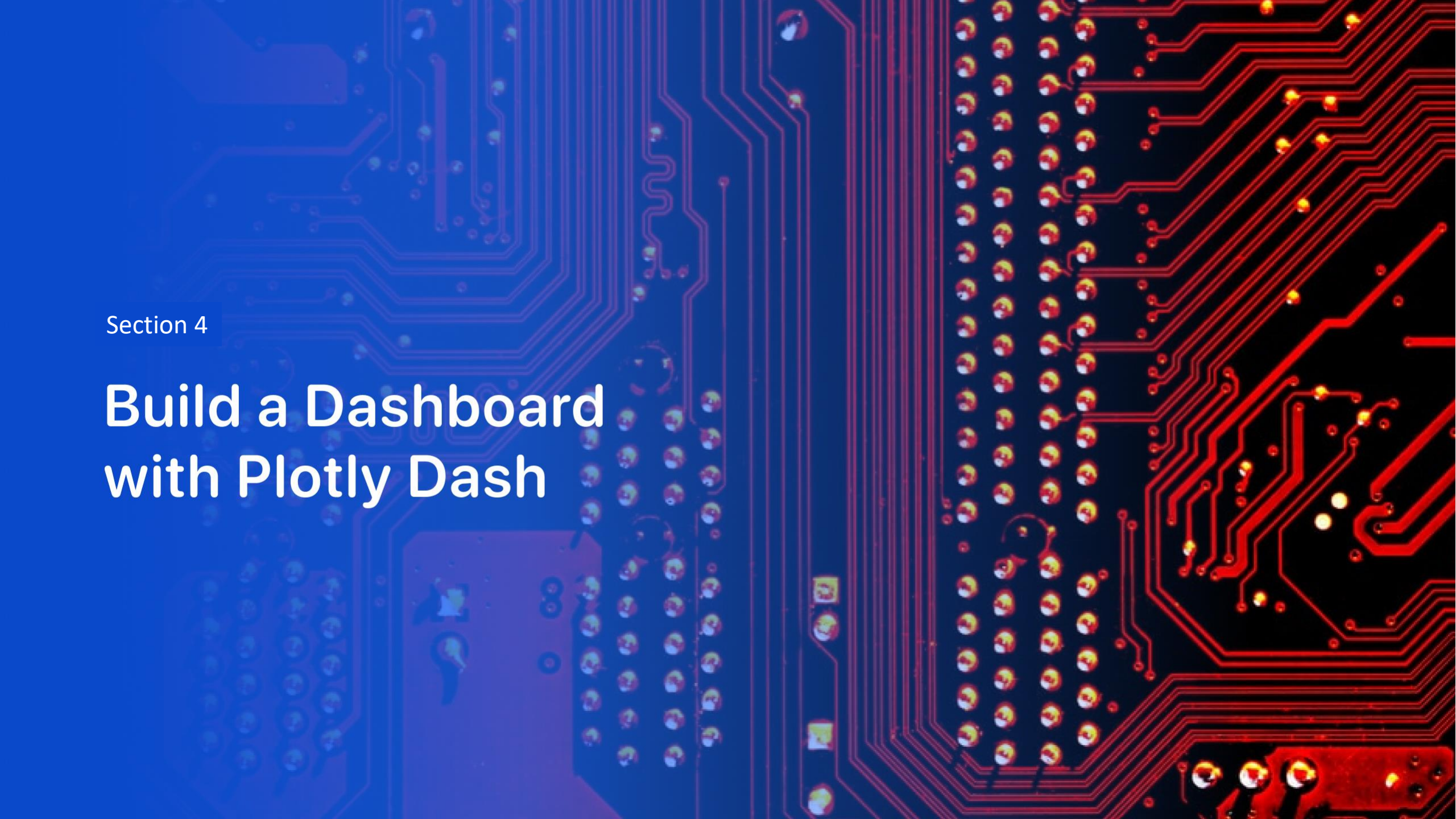


Nearest coastline

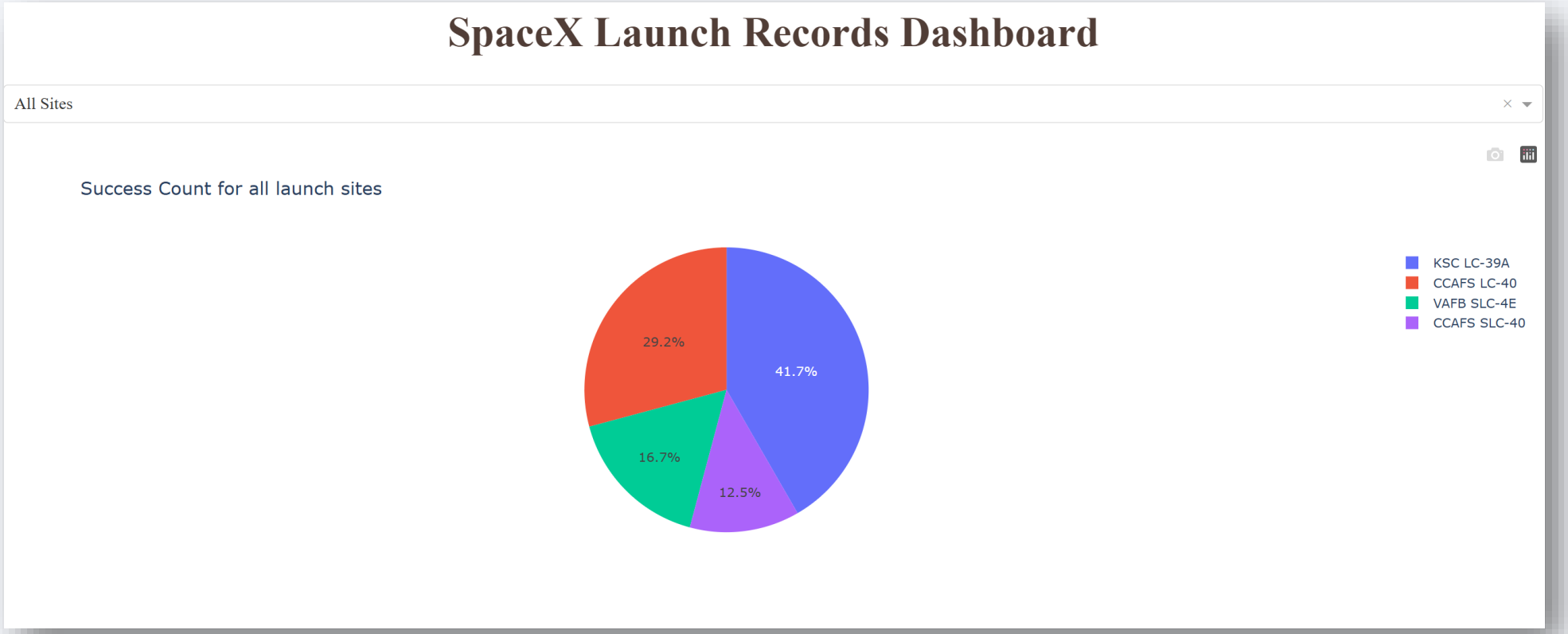


Nearest city

- Folium is used to calculate the distances to points of interest around the launch site. This is helpful in assessing the amenity of the launch site for possible preference.

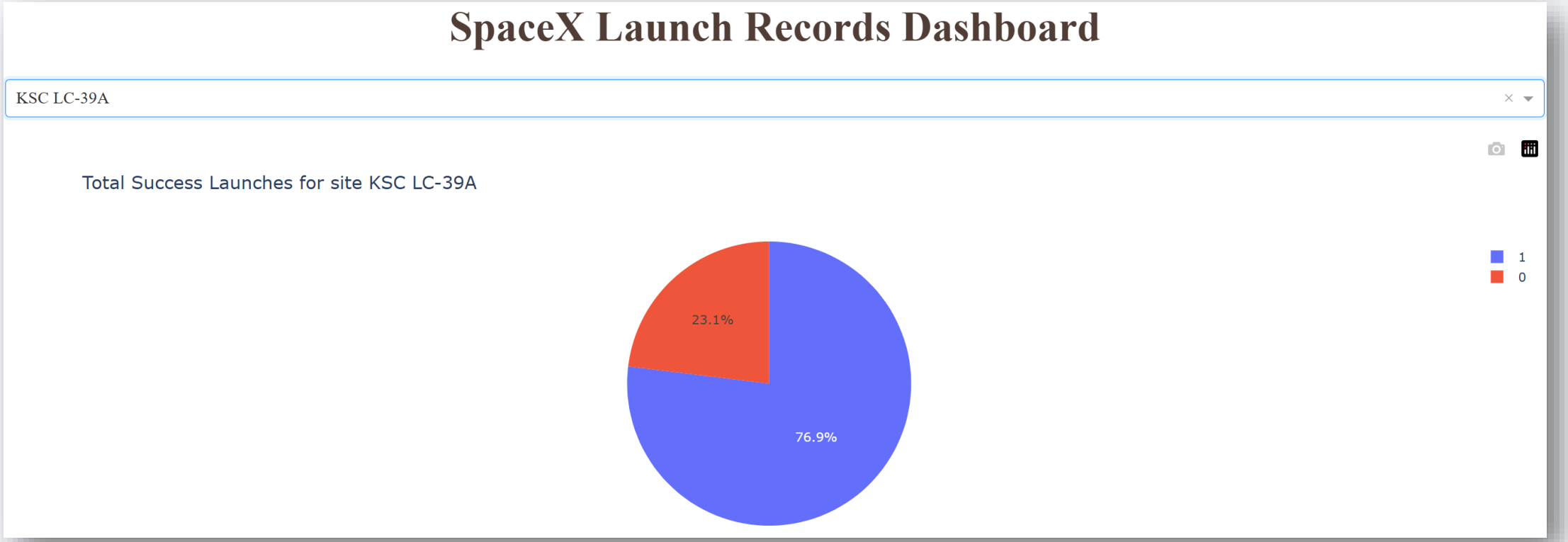

Section 4
Build a Dashboard with Plotly Dash

Dashboard – All Sites Success Count



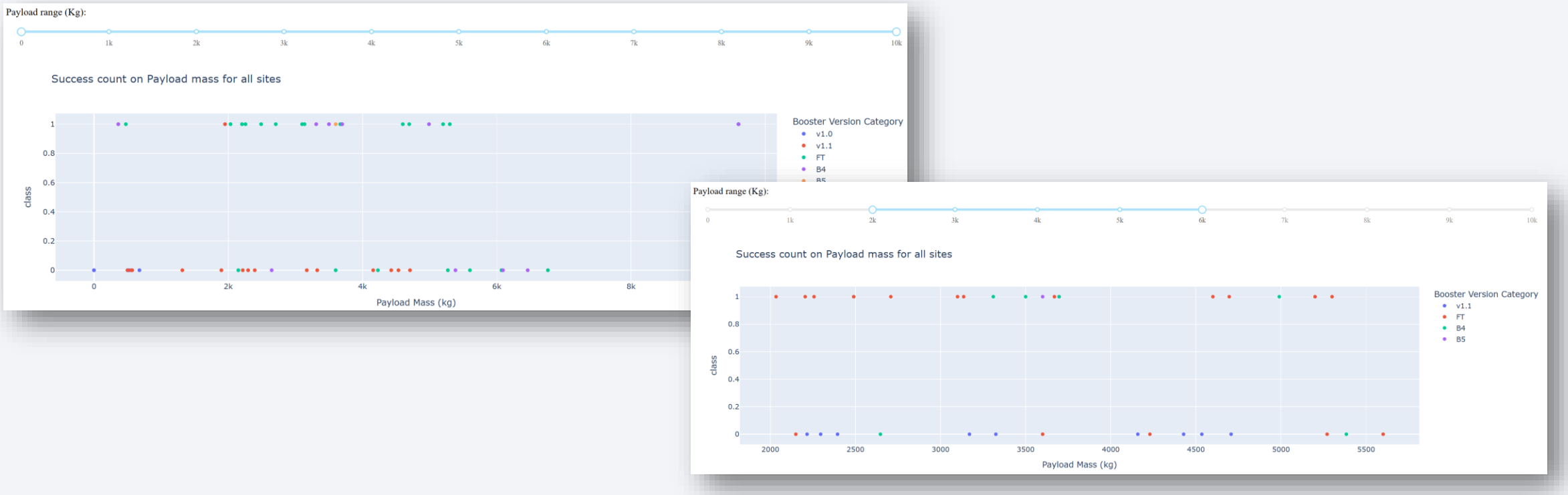
- The pie chart indicates graphically the success count for all launch sites.
- The chart shows a majority of successful launches from KSC LC-39A with the second most being from CCAFS LC-40.

<Dashboard Screenshot 2>



- The highest proportion of successful landings were launched from KSC LC-39A.
- Of those launched from this site 76.9% landed successfully and 23.1% were unsuccessful.

Dashboard – Success by Payload Mass

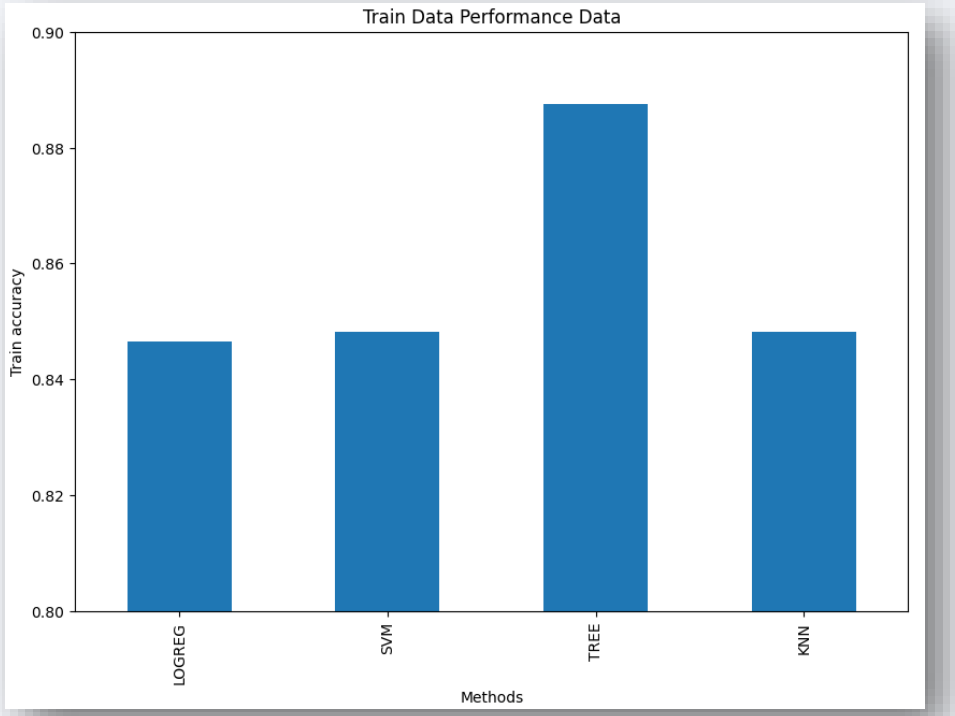


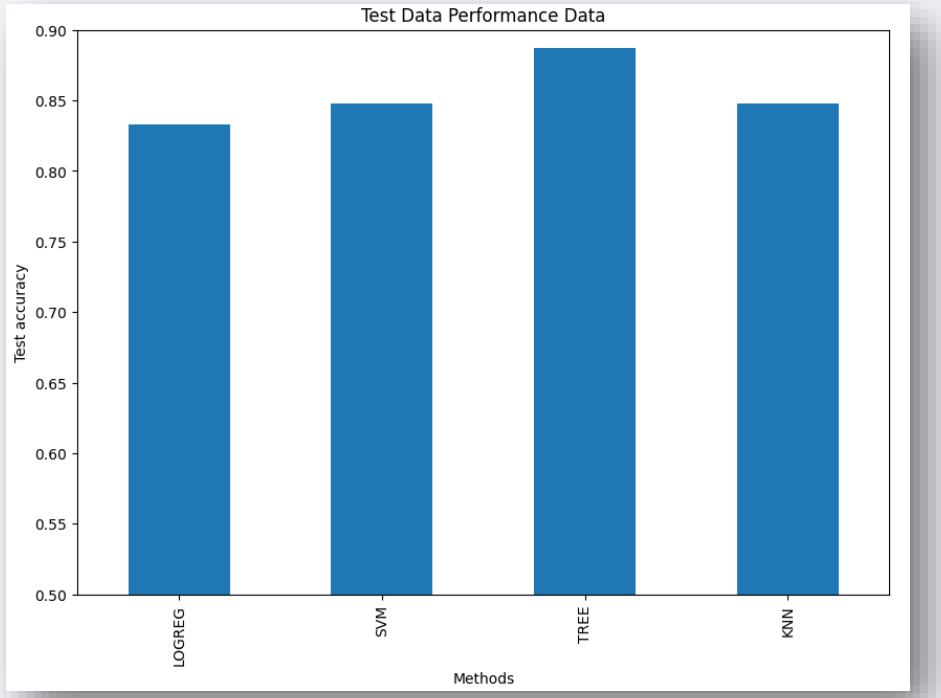
- The top image shows the success and failure of all records. It can be seen that the most common failure version is the v1.1.
- The lower image shows the payload range most likely to successfully land, this falls between 2000kg and 6000kg.

Section 5

Predictive Analysis (Classification)

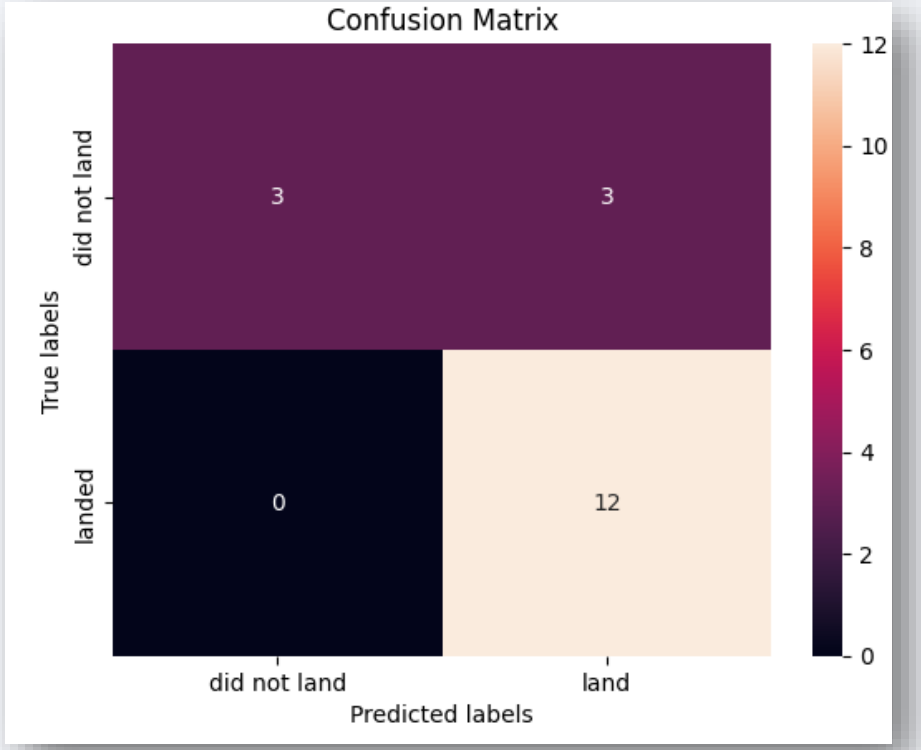
Classification Accuracy





- The train and test set accuracy is displayed in the above charts.
- The decision tree data performs best in both test and train models.

Confusion Matrix



- The decision tree confusion matrix is shown to the left.
- All tests had the same issue of false positive outputs.

Conclusions

- CCAFS SLC 40 is the most utilized launch site. However, landings here have a 60% success rate when launching from the site. This may be skewed due to the rapid development SpaceX employs. KSC LC-39A has the highest percentage of successful landings at 77.3% with VAFB SLC-4E having the lowest success rate of 76.9%.
- ES-L1, GEO, HEO and SSO orbit launches had 100% success rates, VLEO success rate was 90% and the remaining orbit types were between 50% and 75%.
- Success rate has increased exponentially from 2013 indicating learnings from failure being implemented.
- Launch payload is heavily concentrated in the zero to 7000kg range. There is a definite gap between the 7000kg and 16000kg where majority of payloads above this range are to VLEO.
- The best classification method for the date is the decision tree algorithm. This method performed significantly better than other methods with training data and marginally better with test data

Thank you!