IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Carson Parfitt
July 2024

Outline

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

Executive Summary

Methodologies

- Data Collection
  - API
  - Web Scrapping
- Data Wrangling
- Exploratory Data Analysis
  - SQL
  - Data Visualization (Pandas Matplotlib)
- Interactive Visual Analytics with Folium
- Predictive Analysis

Results

- Results of Predictive Analysis
- Results of Exploratory Data Analysis
- Interactive Data Visualization

Introduction

SPACE Y background and context

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. This projects goal is understand if SPACE Y can accurately predict if the first stage of the rocket launch will land successfully.

Problems you want to find answers

- What will cause the rocket to land successfully?
- How successful will the landing be?
- What conditions need to exist for a successful launch?
- What is the cost of a Launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from SpaceX API and web scraping from Wikipedia Pages
- Perform data wrangling
  - Data was collected and converted into Pandas data frame for visualization and analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using machine learning to address if the first stage of Falcon 9 landing will be successful.

Data Collection

The data was collected from SpaceX REST API and web scrapped from Wikipedia

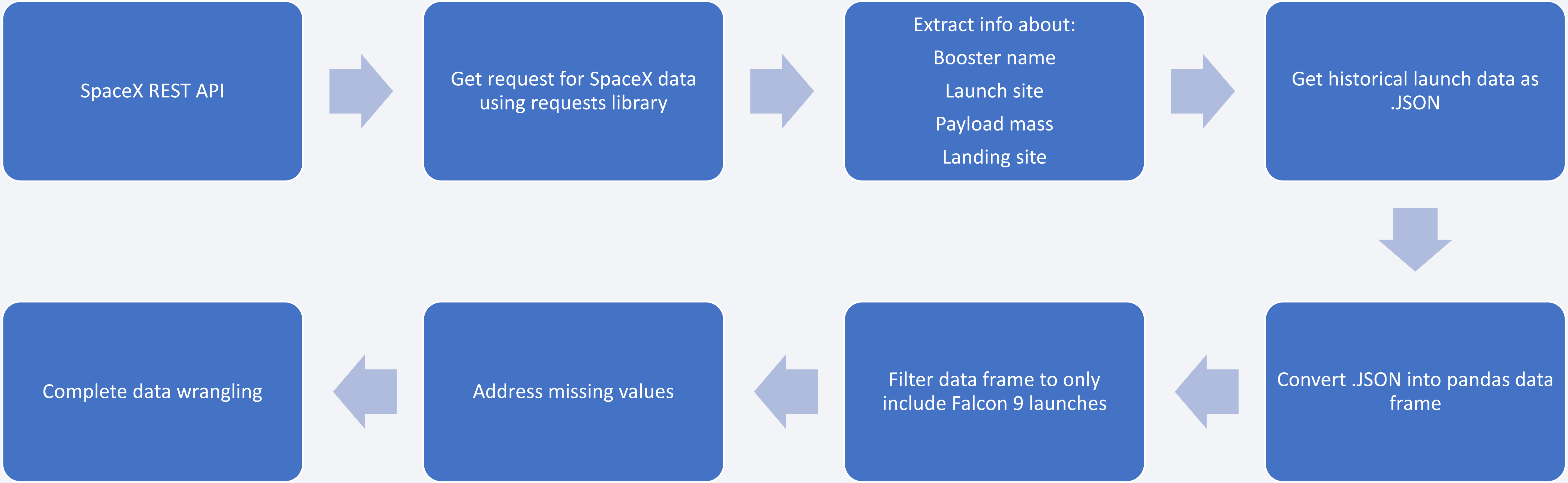
SpaceX REST API → Get request for SpaceX data using requests library → Receive data as .JSON file → Convert .JSON into Pandas data frame → Clean and check data

Web Scraping Falcon 9 launch records → Use Beautiful Soup to web scrape HTML tables → Extract HTML table → Parse data from tables → Convert table into pandas data frame

Data Collection – SpaceX API

Here the data was collected, cleaned and formatted

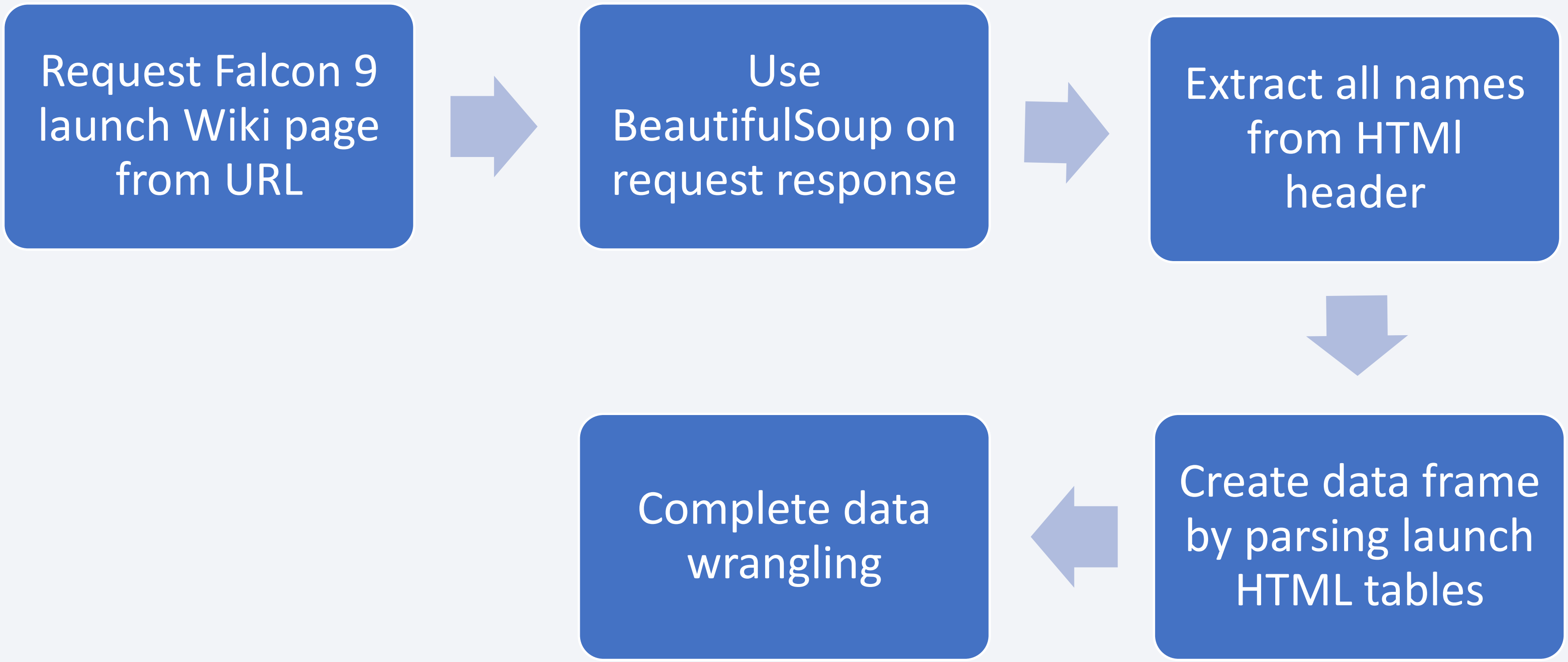
[Collection API notebook]



Data Collection - Scraping

Applied web scraping to collect Falcon 9 launch records from Wikipedia Pages

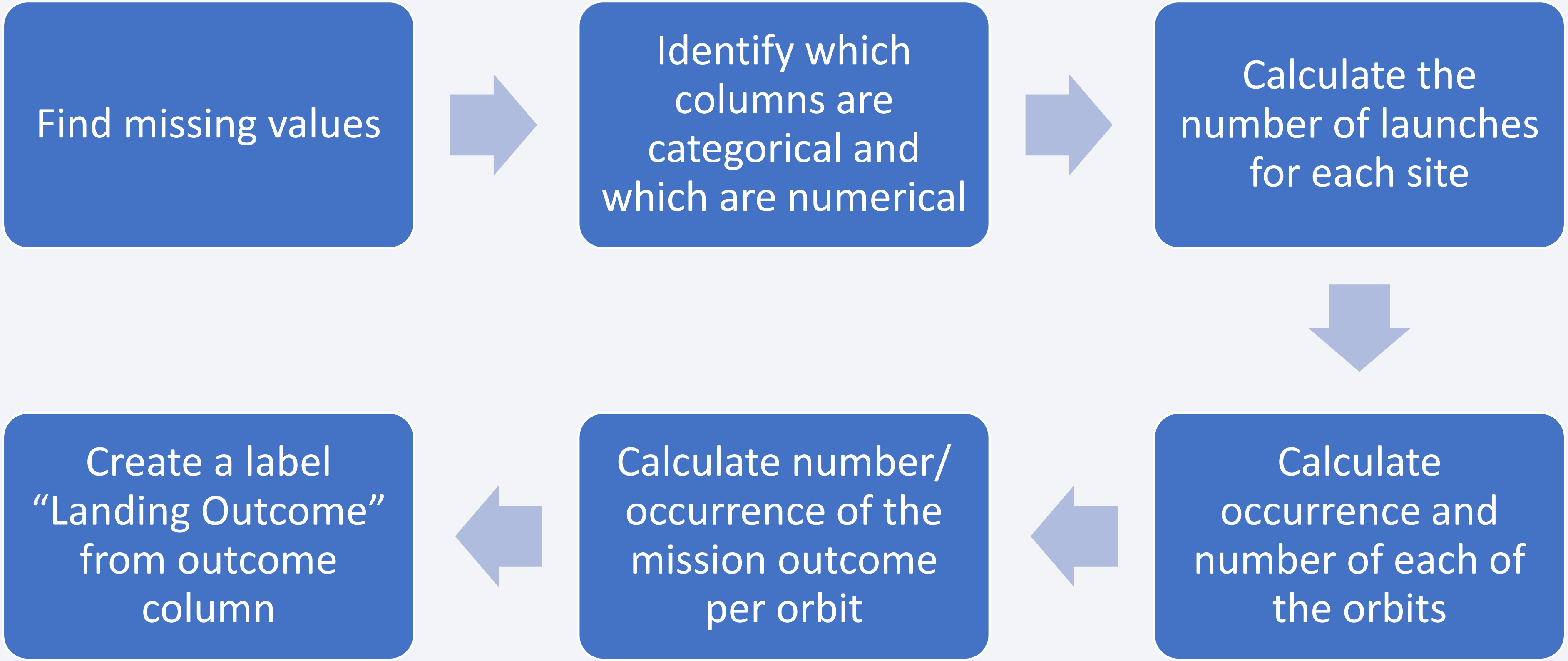
Web Scraping Notebook



Data Wrangling

Preformed Exploratory Data Analysis (EDA) to determine the labels needed to train dataset and to find patterns

[Data Wrangling Notebook]



EDA with Data Visualization

What charts were plotted and why

Most of the charts plotted are Catplot(scatter plot) to help visualize the relationship between a numerical value and a categorical value

- Flight number and Payload
- Flight number and Launch
- Payload and Launch Site
- Flight Number and Orbit Type
- Payload and Orbit type

One chart plotted was a Bar chart used to show a distribution of data/compare values

- Success rate and Orbit type

One chart plotted was a Line chart used to track changes over time

- Launch success yearly trend

EDA with Data Visualization Notebook

EDA with SQL

EDA was applied with SQL to help obtain information about the data.

- Launch sites starting with “CCA” were pulled out
- Total payload mass carried by boosters/boosters F9 v1.1were determined
- Successful landing(ground pad)/(drone ship) outcomes determined
- Total number of success and failures of missions
- The failed landing outcomes in drone ship, their booster version and launch site names

EDA with SQL Notebeook

Build an Interactive Map with Folium

All launch sites were marked and map objects (markers, circles, and lines) were added to the map to mark the success or failures of each launch. Additional features were added to help address questions as well.

- A highlighted circle area with a text label on a specific coordinate for each launch site on the site map was added
- Cluster objects were added to simplify the map that had many markers having the same coordinate
- Mouse position was added to have a pointer show up when hovering over the map
- Line were drawn between each launch site and its closest city, railway and highway.

Launch Sites Locations with Folium Map

Build a Dashboard with Plotly Dash

This dashboard was created to preform interactive analysis of SpaceX launch data in real time.

This dashboard contains various inputs

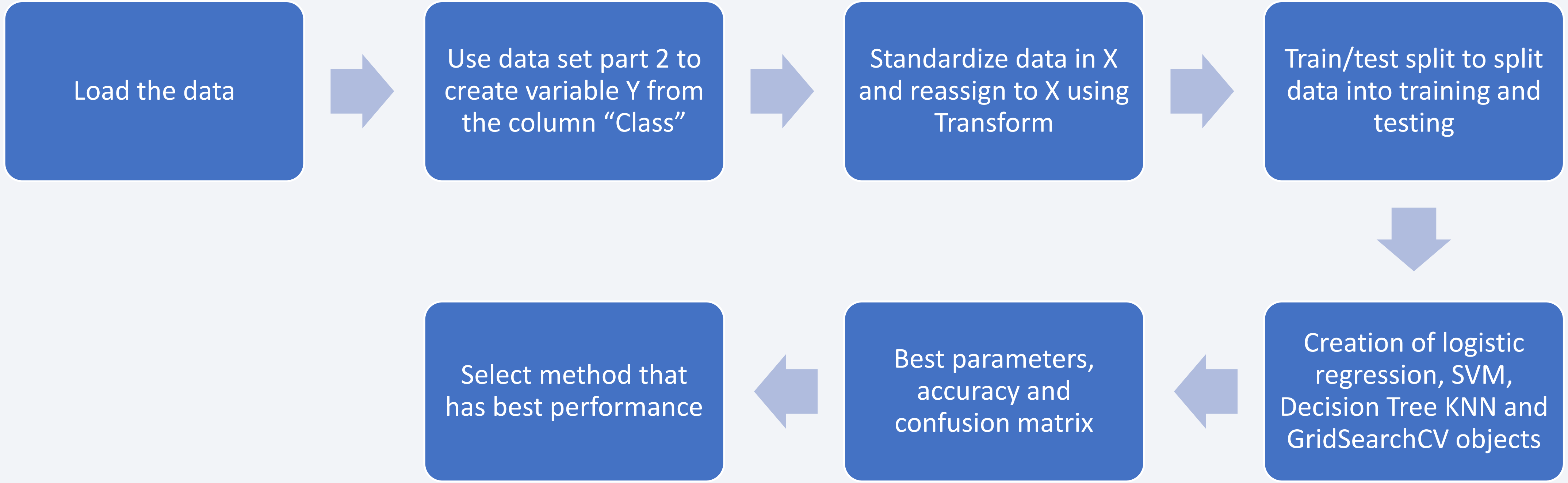
- Launch site drop down
- Callback function – this renders a success pie chart based on the selected site drop down.
- A range slider used to select payload

[Interactive Plotly Dashboard]()

Predictive Analysis (Classification)

Through the machine learning process the best hyperparameters for logistic regression, SVM, Decision Trees and KNN classifiers were able to be determined

Machine Learning Prediction Notebook



Results

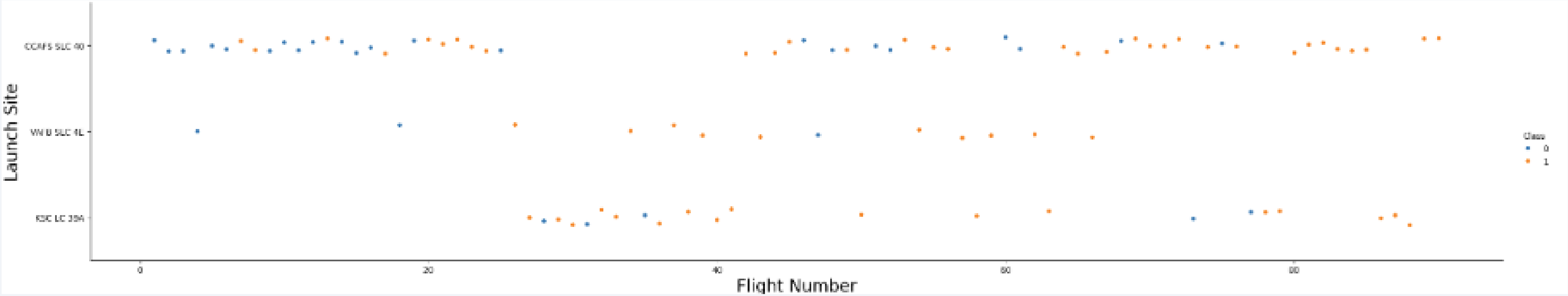
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

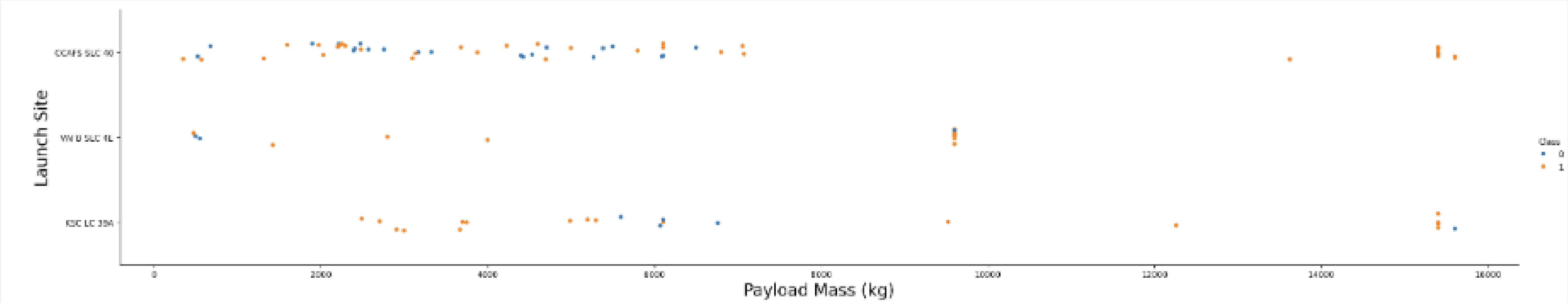
Flight Number vs. Launch Site

- Over time the success of the launch rates have increased at every launch site. Most notably at CCAFS SLC-40 where most launches take place
- VAFB and KSC represent 1/3 of the total launches



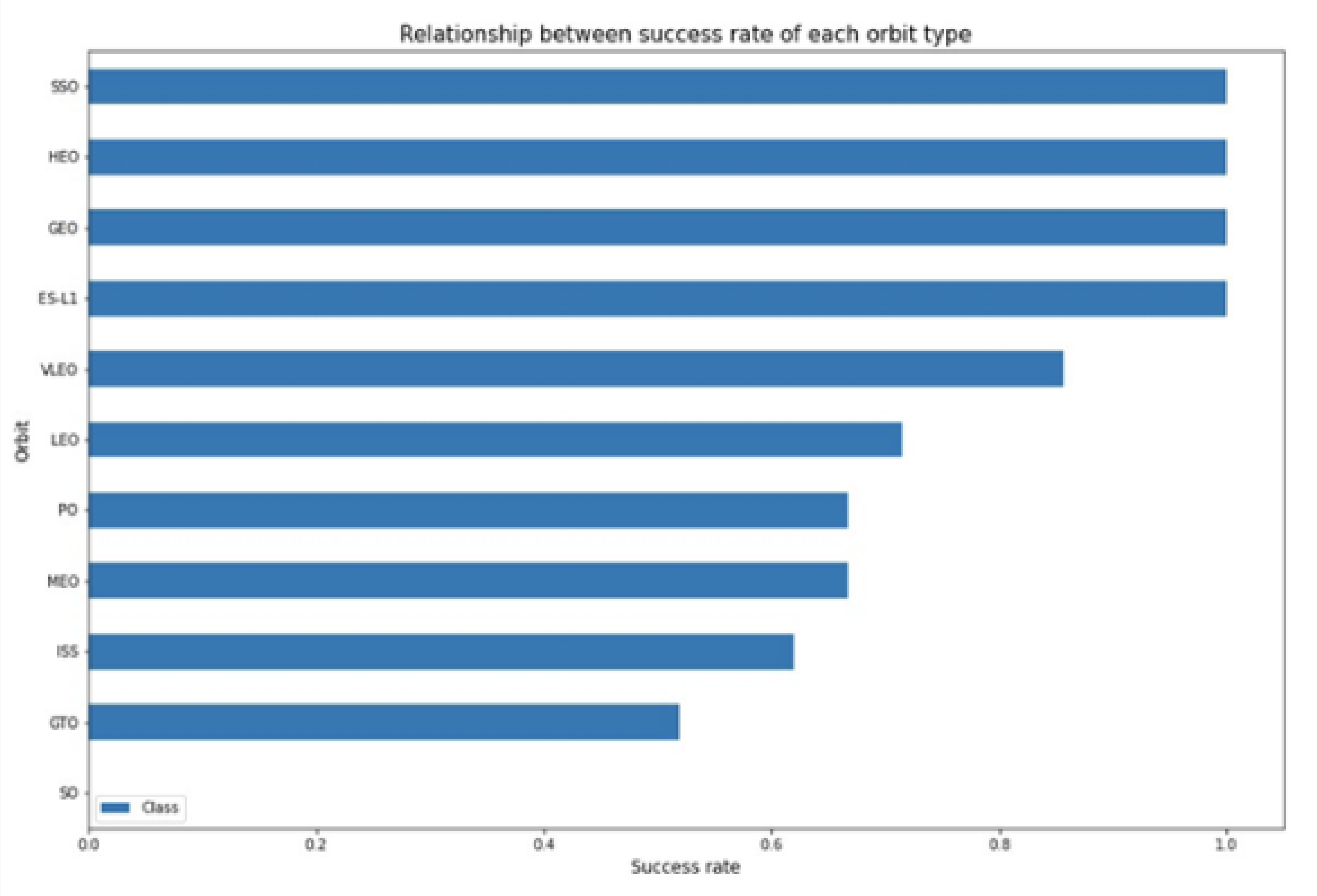
Payload vs. Launch Site

- VAFB does not launch rockets with a heavy payload mass (greater than 10000 kg)
- KSC does not launch rockets with a lower payload mass (less than 2500 kg)
- CCAFS has launched rockets less than 7500kg and more than 13000kg but does not launch rockets in-between that range



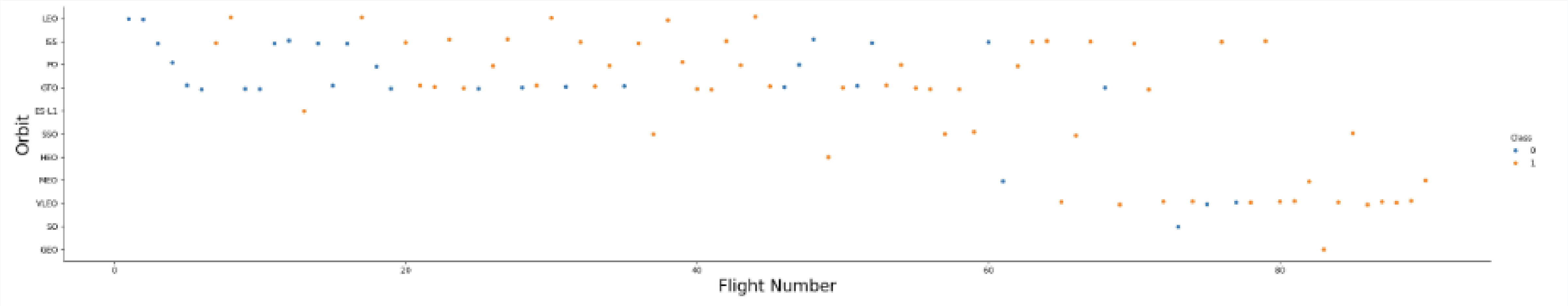
Success Rate vs. Orbit Type

- The first four (SSO, HEO, GEO, ES-L1) have the best success rate



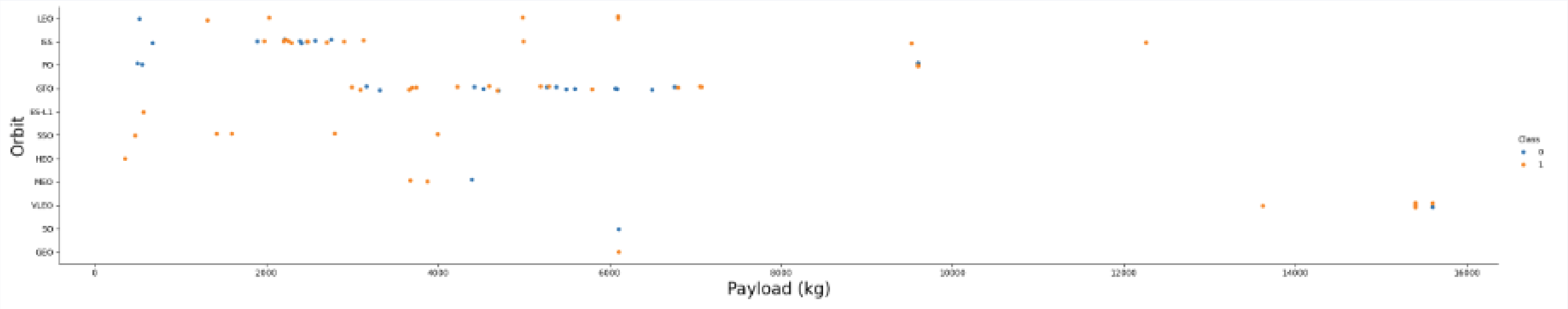
Flight Number vs. Orbit Type

- There are more failures at the beginning of launch series but as more progress the ratio improvs by the number of unsuccessful landings reducing
- GTO and ISS orbits have the highest concentration of launches with the lowest ratio of successful landings
- The orbits with a higher success rate have an overall fewer launches made



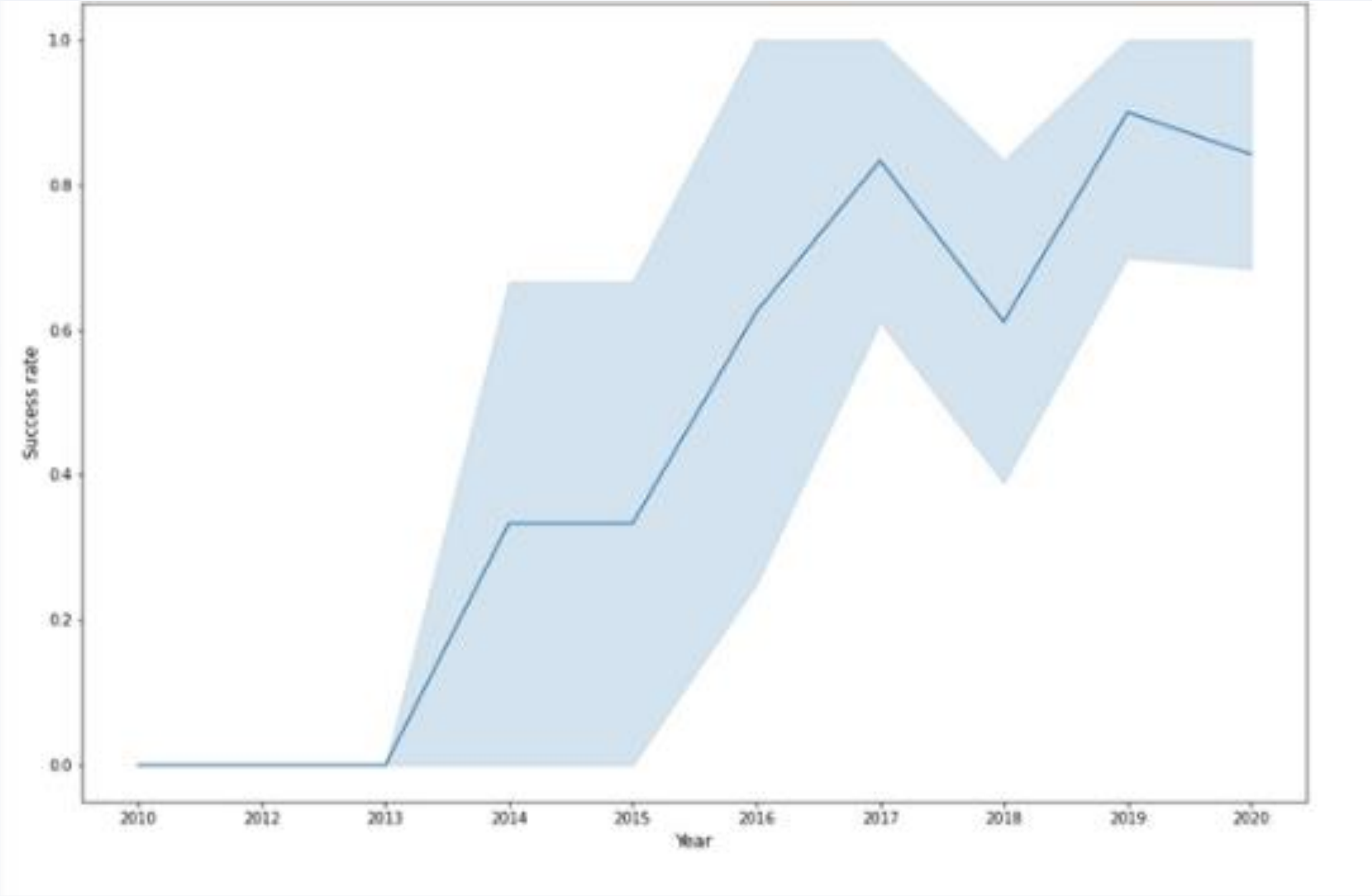
Payload vs. Orbit Type

- For heavy payloads PO, LEO and ISS have a higher landing success rate
- GTO has both positive and negative landing rates making it hard to determine overall success rate



Launch Success Yearly Trend

- The success rate continued to increase from 2013 - 2020



All Launch Site Names

- There are Four unique launch sites
- “DISTINCT” in SQL finds all unique values – in this case it was used in the launch column to pull the unique 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

Launch Site Names Begin with 'CCA'

- The query used WHERE LIKE LIMIT to find 5 records where launch sites begin with
- “CCA”

```
In [36]: %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' limit 5;

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

Out[36]:

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

Total Payload Mass

- Using SUM an d WHERE the total payload mass carried y booster launched by NASA was calculated

```
In [44]: %sql select SUM(Payload_Mass__KG_) AS Total_Payload_Mass_KG FROM SPACEXTBL Where Customer LIKE 'NASA (CRS)'

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

Out[44]:

Total_Payload_Mass_KG
45596

Average Payload Mass by F9 v1.1

- Using AVG() function the average payload mass carried by booster version F9 v1.1 was calculated

```
In [48]: %sql select AVG(Payload_Mass__KG_) AS AVG_Payload_Mass_KG FROM SPACEXTBL Where booster_version = 'F9 v1.1'

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

Out[48]:

AVG_Payload_Mass_KG
2928.4

First Successful Ground Landing Date

- This is the data from the first successful ground pad landing outcome. This was achieved using MIN

In [49]:
```
%sql select MIN(Date) As first_success_landing from SPACEXTBL where Landing_Outcome like 'Success (ground pad)'
```

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

Out[49]:

first_success_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000. This is found using WHERE and AND clause together

In [50]:
```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_  > 4000  ANI
```

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

Out[50]:

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes. This uses COUNT with GROUP BY

List the total number of successful and failure mission outcomes

```
%sql select (select COUNT(Mission_Outcome) AS SuccessOutcome FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Success%'), (select
```

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

(select COUNT(Mission_Outcome) AS SuccessOutcome FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Success%')	(select COUNT(Mission_Outcome) AS FailureOutcome FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Failure%')
100	1

Boosters Carried Maximum Payload

- These are the names of the booster versions which have carried the max payload mass, this was found using subquery

```
%sql SELECT DISTINCT(booster_version), (SELECT MAX(payload_mass__kg_) AS "maximum_payload_mass" FROM SPACEXTBL) FROM SPACEXTBL
```

booster_version	maximum_payload_mass
F9 B4 B1039.2	15600
F9 B4 B1040.2	15600
F9 B4 B1041.2	15600
F9 B4 B1043.2	15600
F9 B4 B1039.1	15600

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT landing_outcome, booster_version, launch_site, DATE FROM SPACEXTBL WHERE landing_outcome LIKE '%Failure (drone ship)%'
```

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

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

- Thecount 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. Query uses CPUNT, WHERE, BETWEEN and GROUP BY

```
%sql SELECT landing_outcome, COUNT(landing_outcome) AS "total" FROM SPACEXTBL WHERE (DATE BETWEEN '2010-06-04' AND '2017-03-20')
```

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

Section 3

Launch Sites Proximities Analysis

All Launch Sites

- Launch sites are close to the coast and within the US California and Florida

Success/Failed Launches per site

- The first map shows clusters for each launch site. The second map shows green markers if the launch was successful and red markers if it was not

Proximities to Launch Site

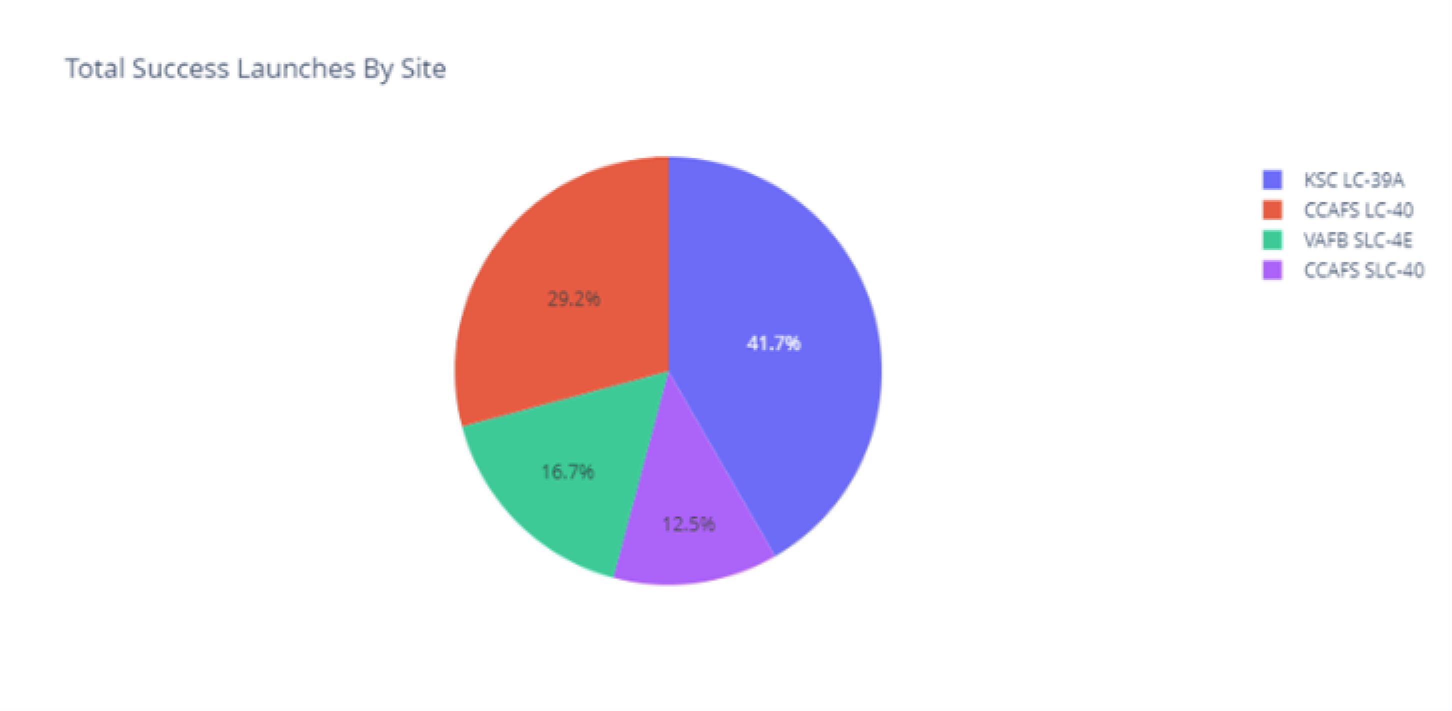
- This shows how close the launch sites are to railways, roads, coastlines, etc.

Section 4

Build a Dashboard with Plotly Dash

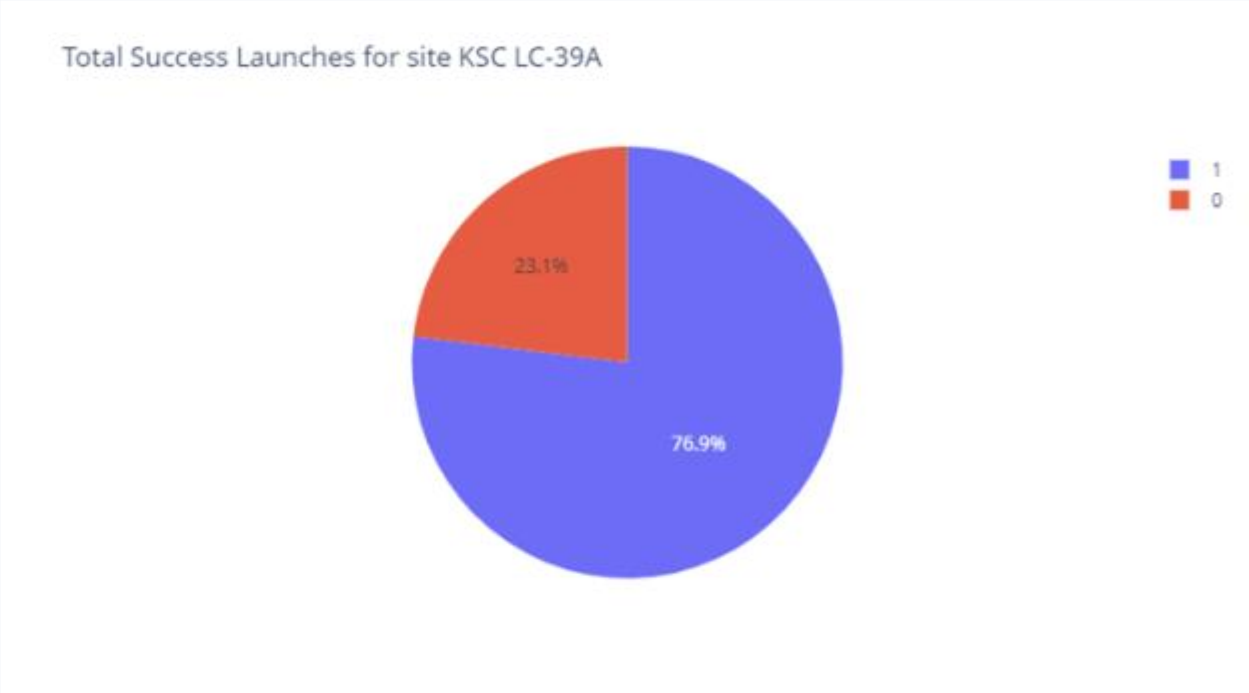
Total Success Launches by Site

- KSC has the highest successful launches followed by CCAFS



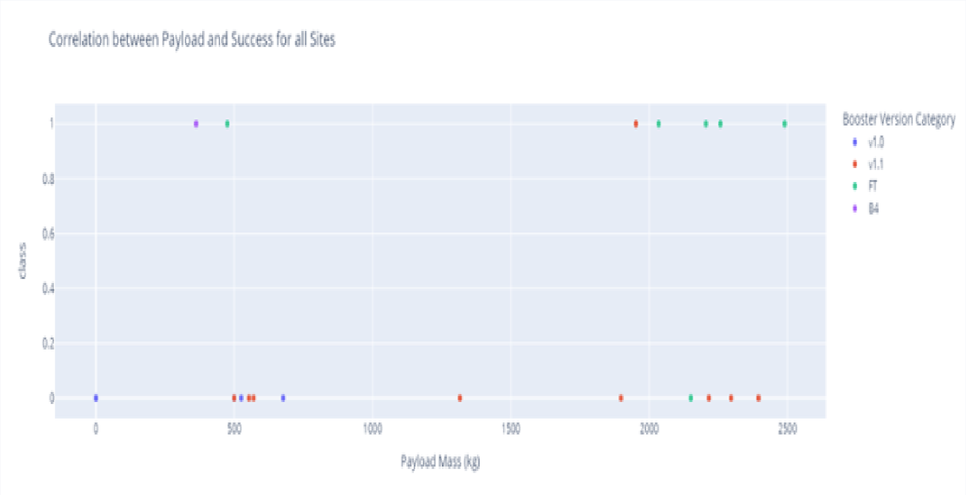
Highest Launch Success Ratio

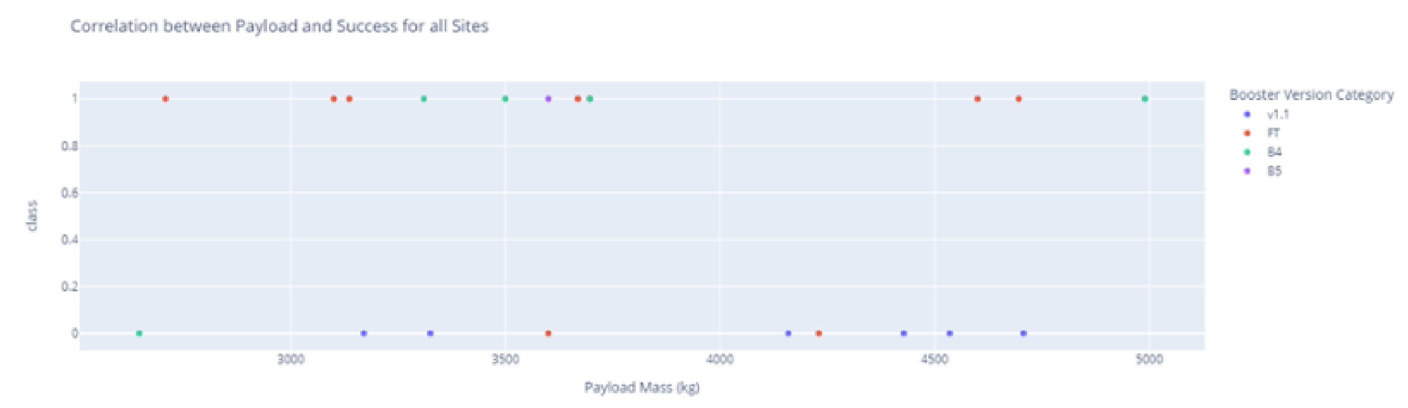
- This pie chart sows KSC as the site which the highest launch success ratio

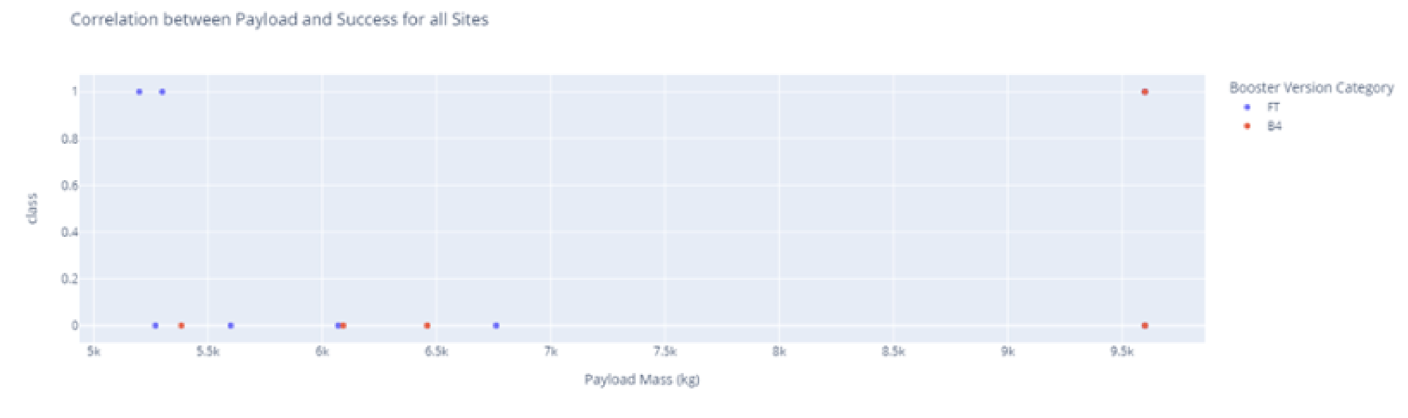


Payload VS Launch Outcome

- Scatter plot for all sites (2500kg, 5000kg and 10000kg) payload ranges
- Payload range 2500-5000kg have most of the successful launches





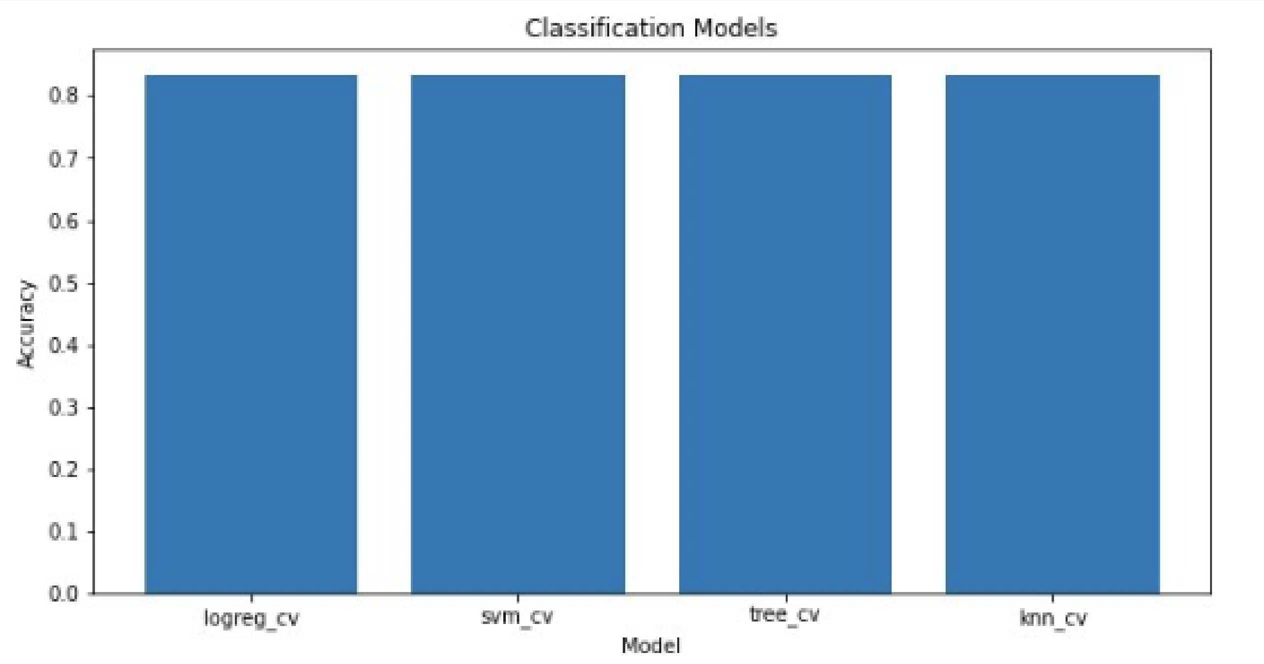


Section 5

Predictive Analysis (Classification)

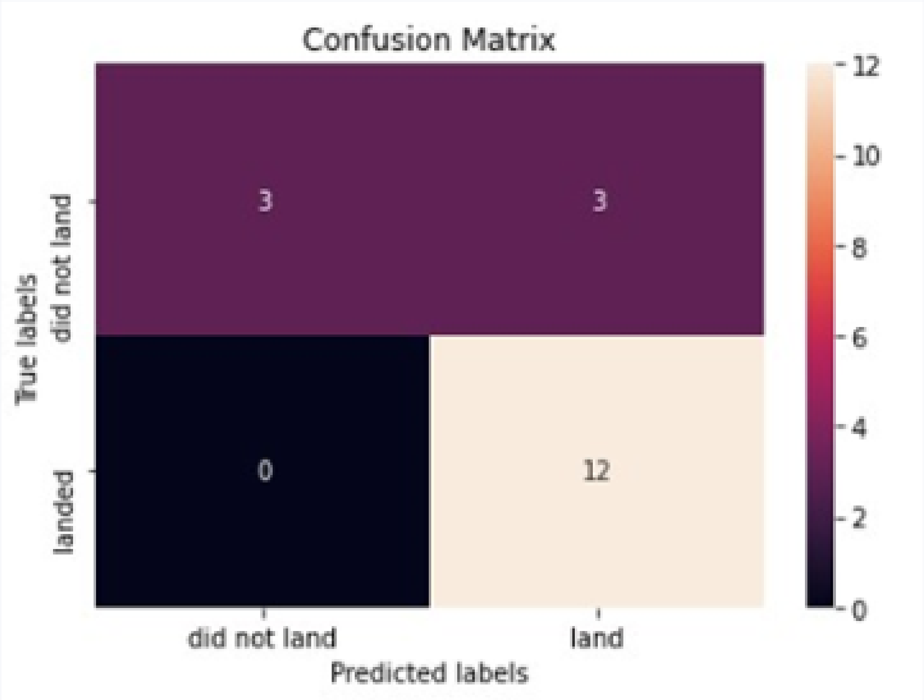
Classification Accuracy

- Accuracy is the **same** for all models



Confusion Matrix

- Confusion Matrix **same** for all models



Conclusions

- In conclusion:
  - The more flights that take place at a launch location the higher the chances of are of a successful launch
  - Successful launches increased from 2013-2020
  - KSC LC-39A had the most successful launches of any site
  - All machine learning models seem to have the same amount of accuracy

Appendix

- For additional information/ raw files of the above please go to this GitHub Repository

Thank you!