IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matthew Pinkerton
17/01/2022

Outline

Executive Summary

Summary of methodologies

- Data Collection, from public SpaceX API and SpaceX Wikipedia page.
- Data wrangling, including adding an indicator for successful landings.
- EDA, via SQL queries and various visualizations and data summaries.
- Explored and analyzed further, using Folium to generate interactive maps.
- Interactive dashboard developed with Plotly Dash.
- Machine learning models trained to predict successful landings through classification techniques.

Summary of all results

- The resulting models all produced similar results, with an accuracy rate of ~83.33% when tested with a test data set. The models tended to over predict successful landings. Training the model with more data could lead to improved accuracy.

Introduction

Project background and context

We will predict if the Falcon 9 first stage will land successfully. 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.

Problems you want to find answers

- To be able to predict the likelihood of a rocket successfully landing.
- Which factors influence if a rocket will successfully land?
- Optimal conditions to achieve a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX REST API and web scraping from SpaceX Wikipedia page.
- Performed data wrangling
  - Irrelevant fields removed, records with null values removed.
  - One hot encoding used to express landing success as a binary result.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Built, tuned, evaluated classification models

Data Collection

SpaceX REST API

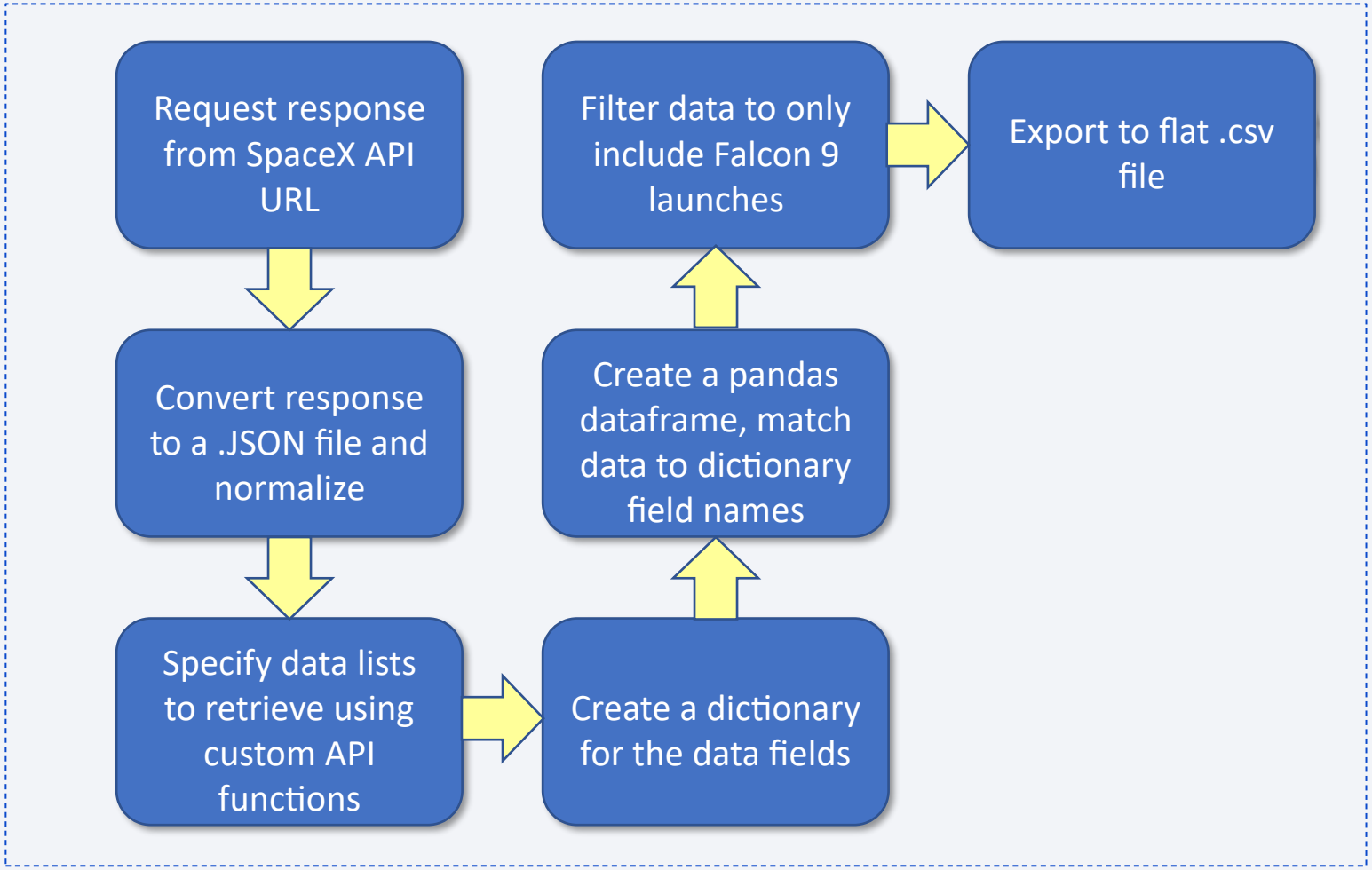
- Data was collected from the SpaceX REST API: api.spacexdata.com/v4/. More information can be found @ https://docs.spacexdata.com/. The data included information about the rocket used, payload, landing outcome, and other launch & landing specifications in the form of a .JSON file. Data is normalized into a flat .csv file.

Web Scraping

- Data was also collected by web scraping a table from the SpaceX Wikipedia page using the BeautifulSoup python package. Information can be appended to our dataset by using the rocket/flight id as a key. Data is normalized into a flat .csv file.

Data Collection – SpaceX API

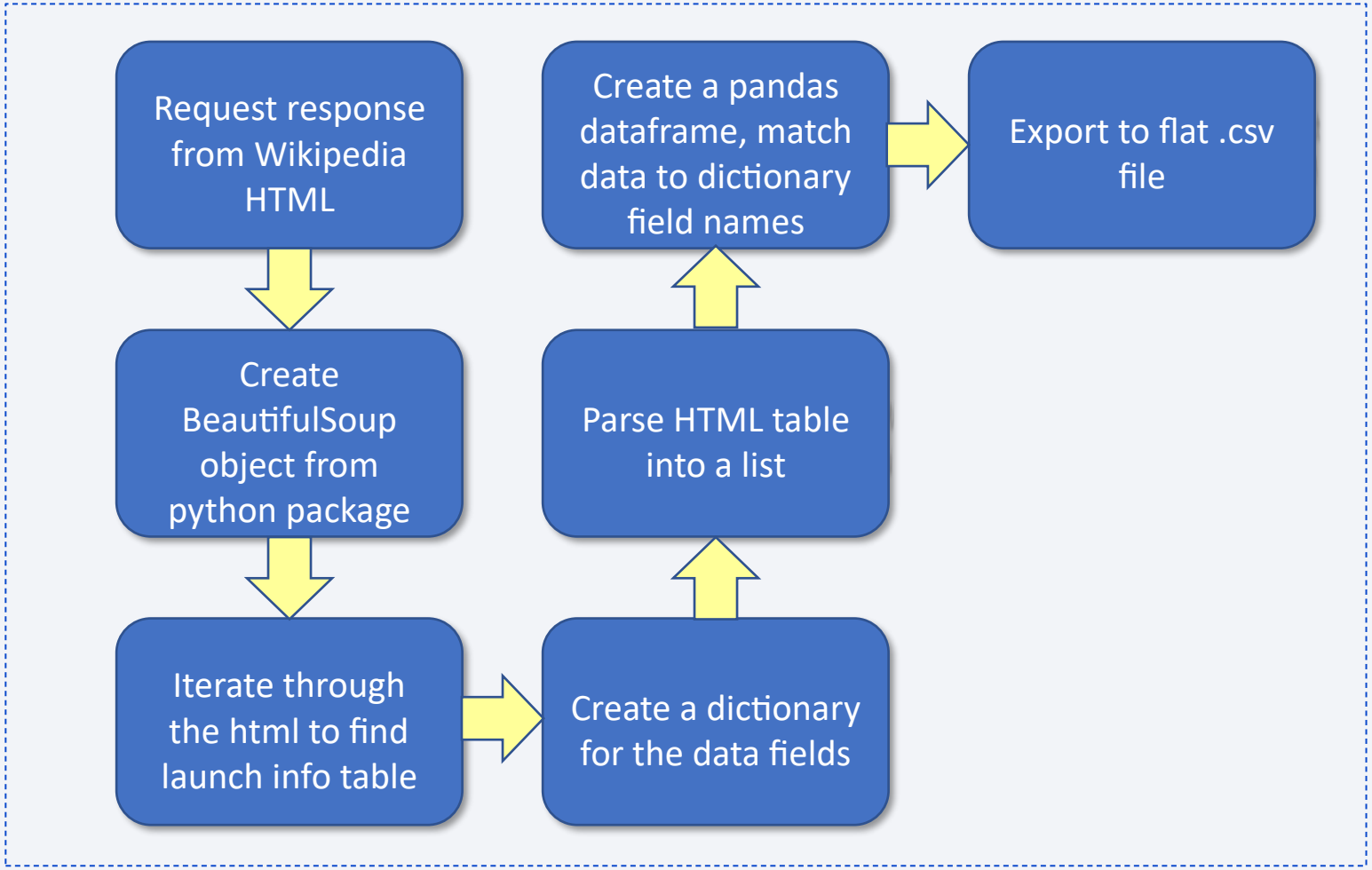
- Flowchart



GitHub URL to notebook

Data Collection – Web Scraping (Wikipedia)

- Flowchart



GitHub URL to notebook

Data Wrangling

- Training label
  - The outcome field details two components: ‘mission outcome’ and ‘landing location’
  - We want to create a training label ‘Class’ to indicate successful landing = 1; unsuccessful landing = 0
  - Value mapping:
    - Outcomes ‘True ASDS’, ‘True RTLS’, & ‘True Ocean’ – set Class to -> 1
    - Outcomes ‘None None’, ‘False ASDS’, ‘None ASDS’, ‘False Ocean’, ‘False RTLS’ – set Class to -> 0

GitHub URL to notebook

EDA with Data Visualization

Goal

- Exploratory Data Analysis carried out on the variables ‘Flight Number’, ‘Payload Mass’, ‘Launch Site’, ‘Orbit’, ‘Class’ and ‘Year’, to investigate relationships between variables.

Charts plotted

- Scatter charts: Flight Number VS. Payload Mass, Flight Number VS. Launch Site, Payload VS. Launch Site, Orbit VS. Flight Number, Payload VS. Orbit Type, Orbit VS. Payload Mass
- Bar charts: Mean VS. Orbit
- Line charts: Success Rate VS. Year

Github URL to notebook

EDA with SQL

- Goal
  - To better understand the data, the dataset is loaded into IBM DB2 Database and queried using SQL magic in python.
- Queries
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string ‘CAA’
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date where the successful landing outcome in drone ship was achieved
  - Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster_versions which have carried the maximum payload mass
  - Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2015
  - Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

GitHub URL to notebook

Build an Interactive Map with Folium

- Folium maps visualises the launch data onto an interactive map. Using the latitude and longitude coordinates of each launch site, we added labelled circle markers at each launch site. Using MarkerCluster(), we indicate successful outcomes with green markers, and unsuccessful outcomes with red markers. We can calculate the distance to key locations on the map and mark a line on the map to visualise this. E.g. distance to nearest railway, highway, coast, city.

Github URL to notebook

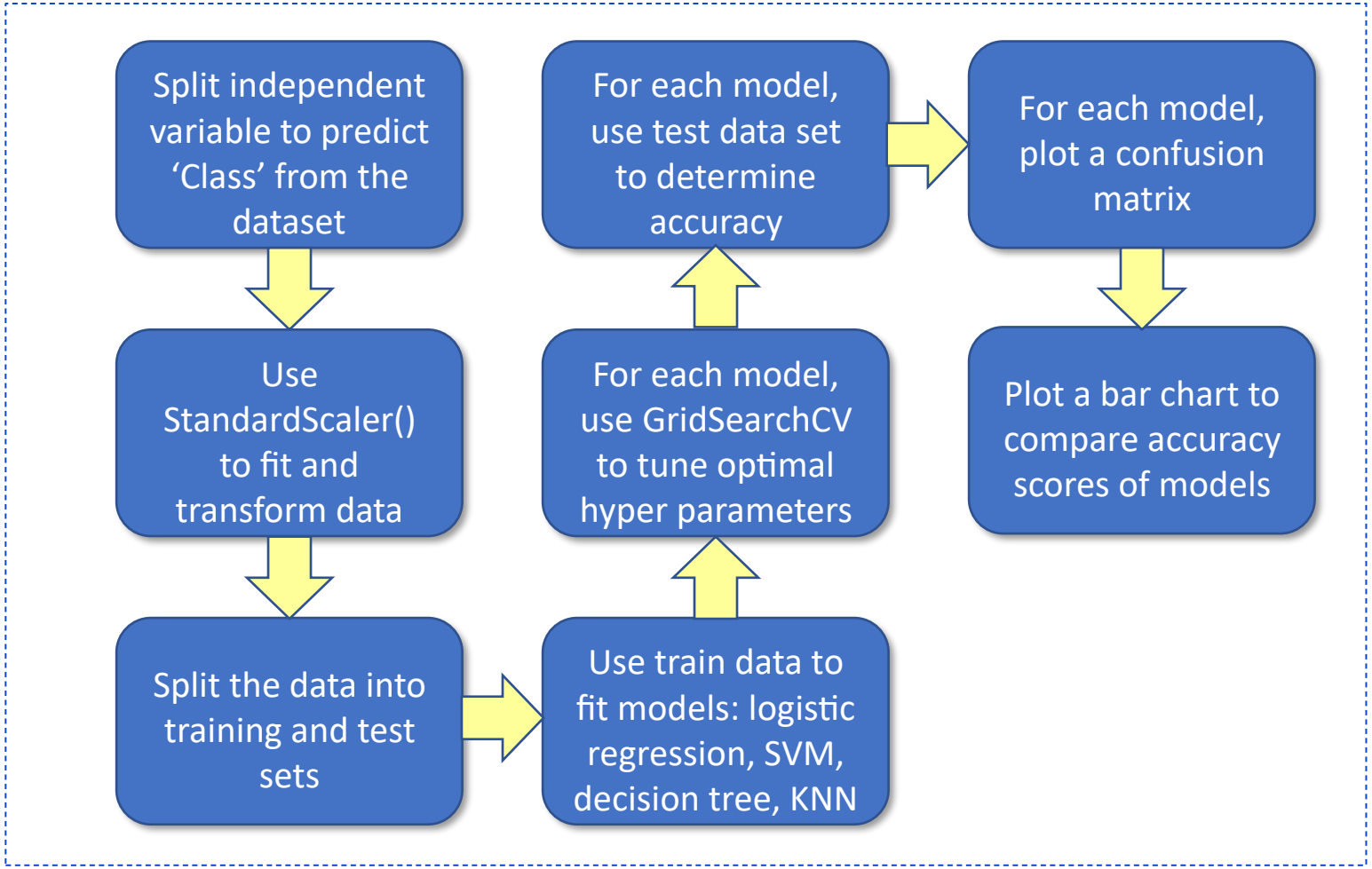
Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Interactive pie chart used to visualise launch site success rate; showing distribution of successful landings across all launch sites or distribution of successful landings for specific individual launch site.
- Scatter plot used to visualise how success varies dependent on payload mass and booster version category.

Github URL to notebook

Predictive Analysis (Classification)

- Flowchart



[GitHub URL to notebook]

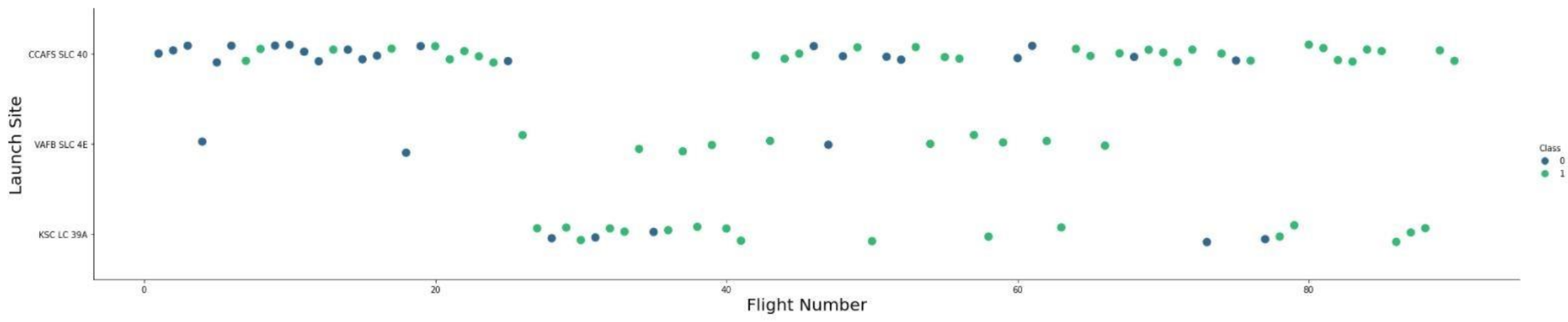
Results

- Results are displayed in the following slides in the form of:
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Section 2

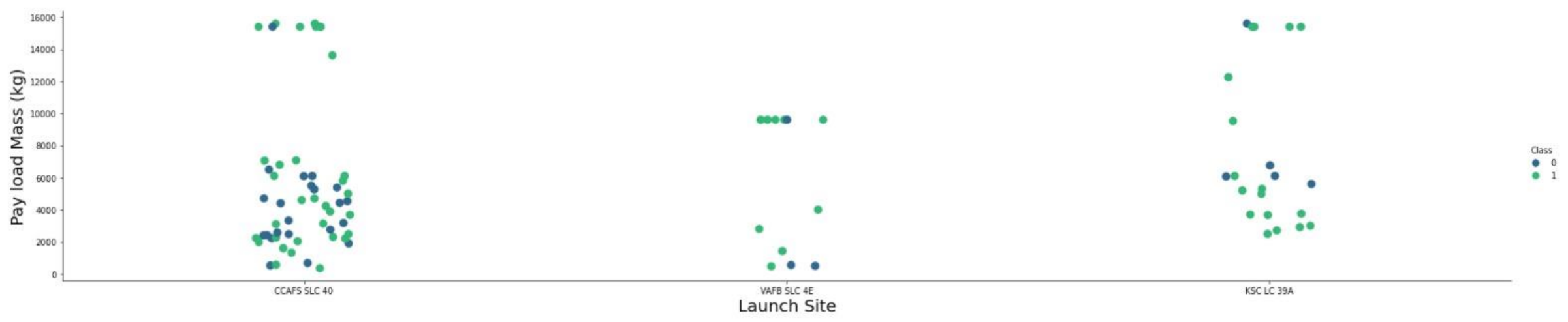
Insights drawn from EDA

Flight Number vs. Launch Site



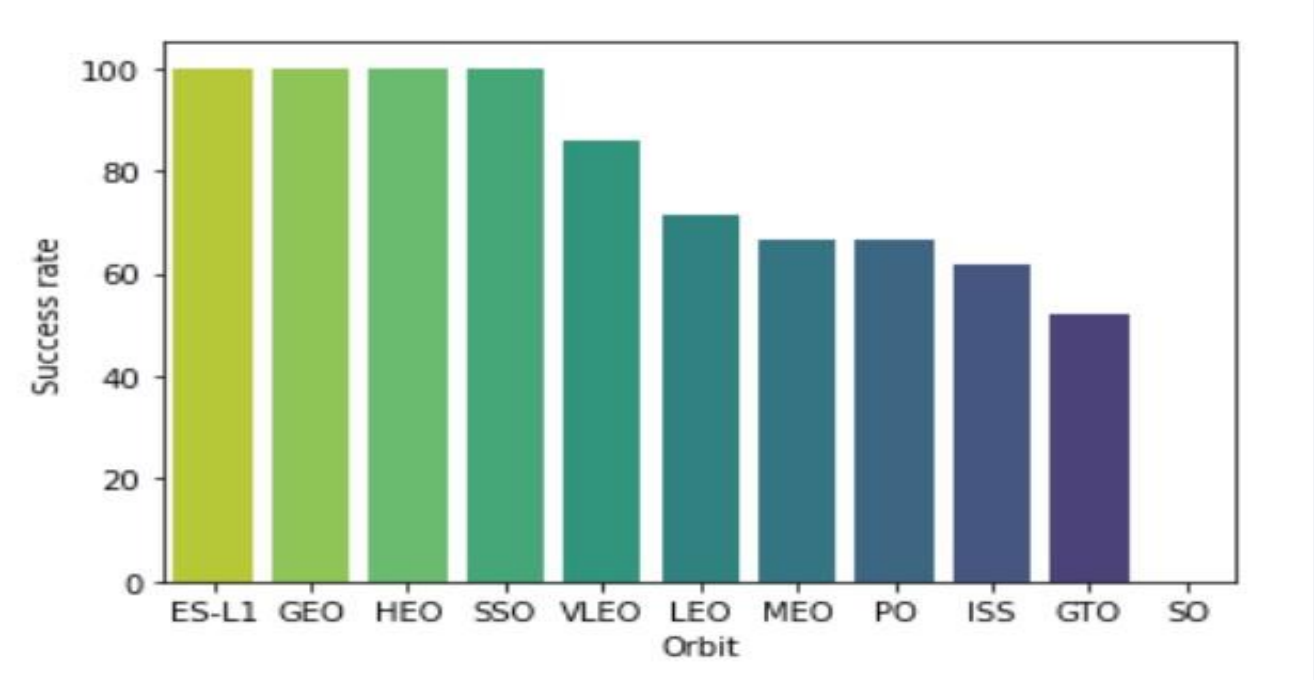
- Green indicates a successful launch. Purple indicates an unsuccessful launch.
- Unsuccessful launches were more frequent in the early flight numbers, success rate has improved for more recent flights.

Payload vs. Launch Site



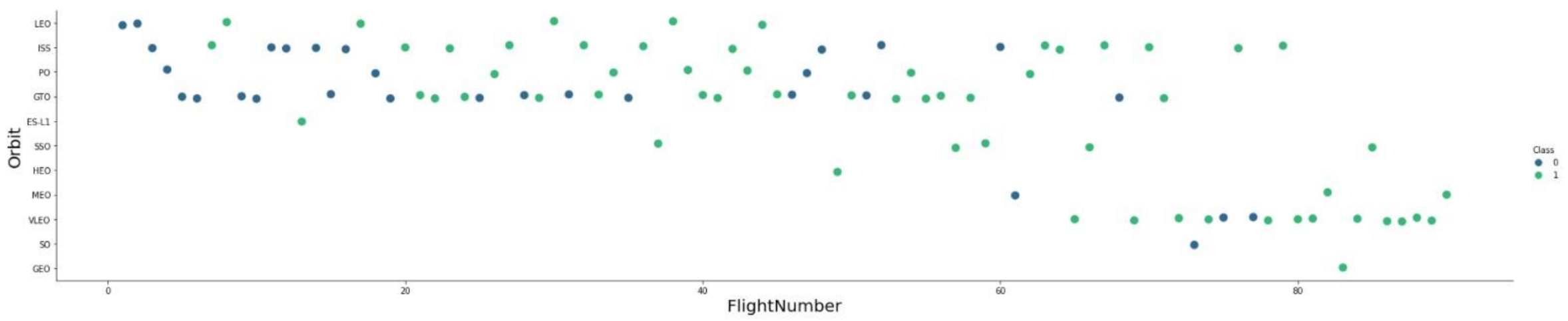
- Green indicates a successful launch. Purple indicates an unsuccessful launch.
- Unsuccessful launches are more frequent in flights with mid-lower pay load mass.

Success Rate vs. Orbit Type



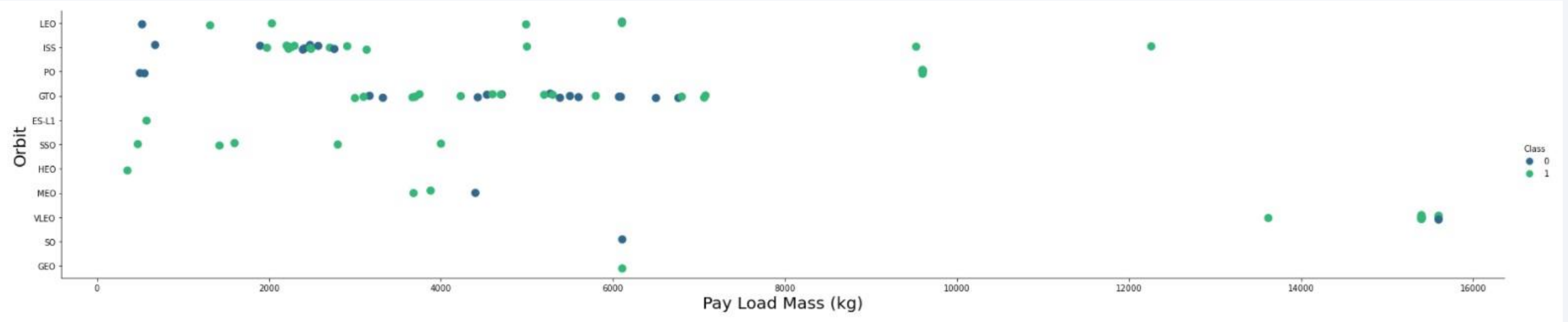
- ES-L1, GEO, HEO, SSO orbits have 100% successful launch rate.
- SO orbits have 0% successful launch rate.

Flight Number vs. Orbit Type



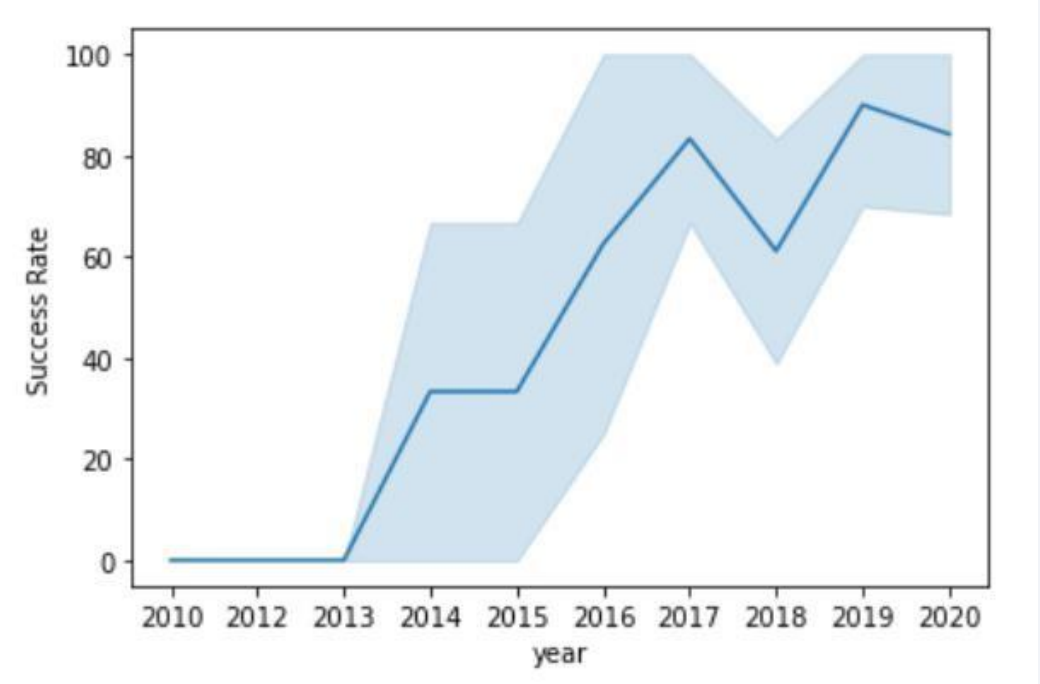
- Green indicates a successful launch. Purple indicates an unsuccessful launch.
- Orbit preference appears to have changed over time. We see some correlation of a higher success rate with the more recent orbit preferences.

Payload vs. Orbit Type



- Green indicates a successful launch. Purple indicates an unsuccessful launch.
- There doesn’t seem to be much correlation with pay load mass for GTO orbits. Some other orbits were more successful with heavier payloads.

Launch Success Yearly Trend



- We can see that success rate has generally increased from 2013-2020, with a slight decrease in 2018, and the highest success rate so far being observed in 2019.

All Launch Site Names

- SQL Query

```
%sql select DISTINCT LAUNCH_SITE from SPACEX_DATA
```

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

- Explanation
  - Returns unique launch site names from the database. The first three results are all likely the same launch site, with errors in the data entry.

Launch Site Names Begin with 'CCA'

- SQL Query

```
%sql select * from SPACEX_DATA 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

- Explanation
  - Returns first five records in the database where launch site name begins with “CCA”.

Total Payload Mass

- SQL Query

```
%sql select sum(payload_mass__kg_) as sum from SPACEX_DATA where customer like 'NASA (CRS)'
```

sum_payload_mass_kg
45596

- Explanation
  - Returns the sum of payload mass (kg) for all records where the customer name is “NASA (CRS)”.

Average Payload Mass by F9 v1.1

- SQL Query

```
%sql select avg(payload_mass__kg_) as Average from SPACEX_DATA where booster_version like 'F9 v1.1%'
```

avg_payload_mass_kg
2928

- Explanation
  - Returns the average of payload mass (kg) for records where the booster version name begins with “F9 v1.1”.

First Successful Ground Landing Date

- SQL Query

```
%sql select min(date) as Date from SPACEX_DATA where mission_outcome like 'Success'
```

first_success
2015-12-22

- Explanation
  - Returns the date of the first successful launch landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query

```
%sql select booster_version from SPACEX_DATA where (mission_outcome like 'Success')
AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome like 'Success (drone ship)')
```

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

Explanation

- Returns the booster versions of records which successfully landed a drone ship landing and where the payload was between 4000kg and 6000kg.

Total Number of Successful and Failure Mission Outcomes

 SQL Query

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEX_DATA GROUP by mission_outcome ORDER BY mission_outcome
```

mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

 Explanation

- Returns all types of mission outcome, and the number of occurrences. Most missions are successful, and is not an indicator of whether or not the landing of the first stage was successful.

Boosters Carried Maximum Payload

- SQL Query

```
maxm = %sql select max(payload_mass__kg_) from SPACEX_DATA
maxv = maxm[0][0]
%sql select booster_version from SPACEX_DATA where
payload_mass__kg_=(select max(payload_mass__kg_) from SPACEX_DATA)
```

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

- Explanation
  - Returns the names of the booster versions which have carried the maximum payload mass.

2015 Failed Drone Ship Landing Records

- SQL Query

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site
from SPACEX_DATA where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

- Explanation
  - Returns the month, landing outcome, booster version, payload mass (kg) and launch site of 2015 launches which failed to land a drone ship landing.

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

SQL Query

```
%sql select landing__outcome, count(*) as count from SPACEX_DATA
where landing_outcome like 'success' AND Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP by landing__outcome ORDER BY count Desc
```

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

Explanation

- Returns all landing outcome types and the number of occurrences for each, for successful landings between 04/06/2010 and 20/03/2017.

Section 4

Launch Sites Proximities Analysis

Folium Map: Launch site locations

Insights

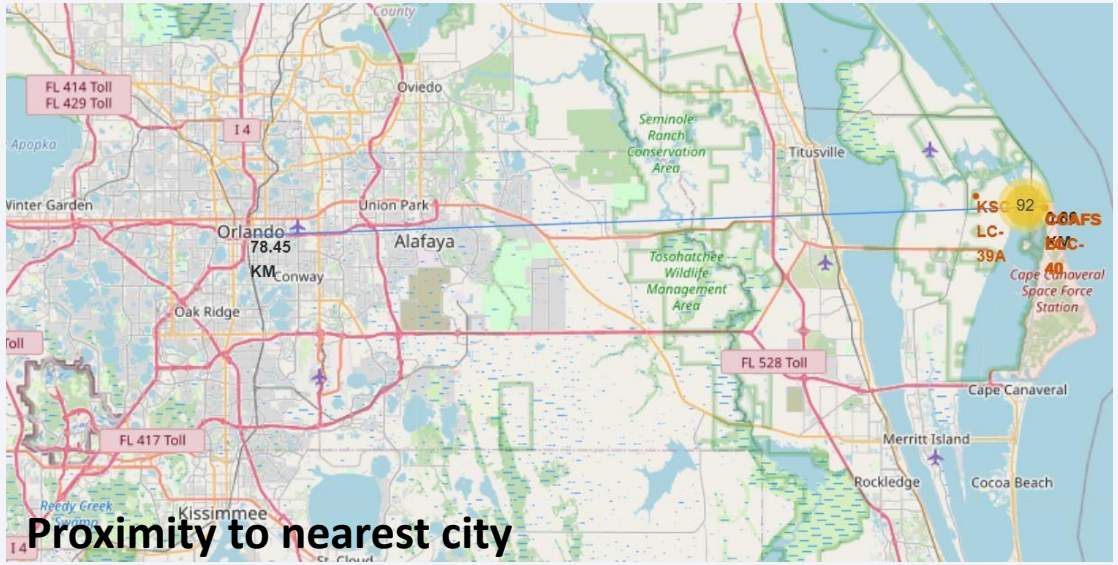
- We can see that all launch sites are located in North America and that all launch sites are located near to coastlines, specifically the coasts of Florida and California.

Folium Map: Color-labelled launch outcomes

Insights

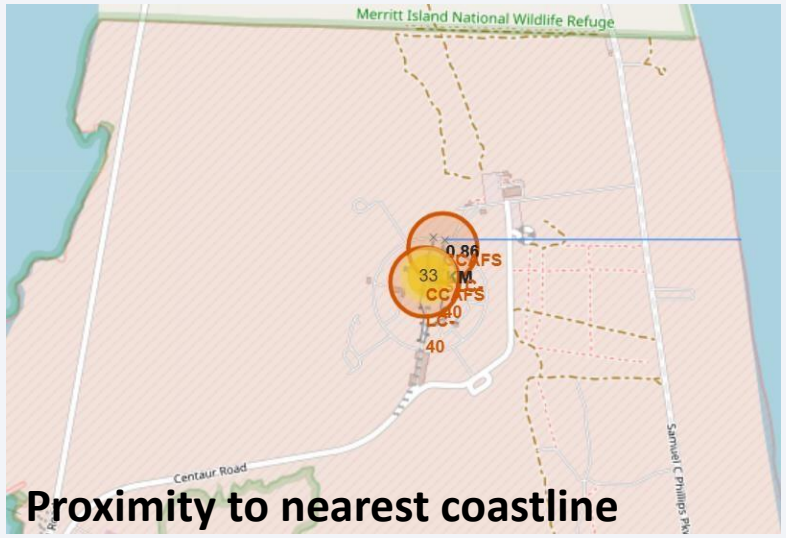
- The left screenshot tells us that 10 launch records are clustered at this launch location (VAFB SLC-4E). We can drill down by clicking on the cluster, expanding the image like shown in the right screenshot. This tells us that there were 4 successful landings (green) and 6 unsuccessful landings (red).

Folium Map: Proximity to key locations

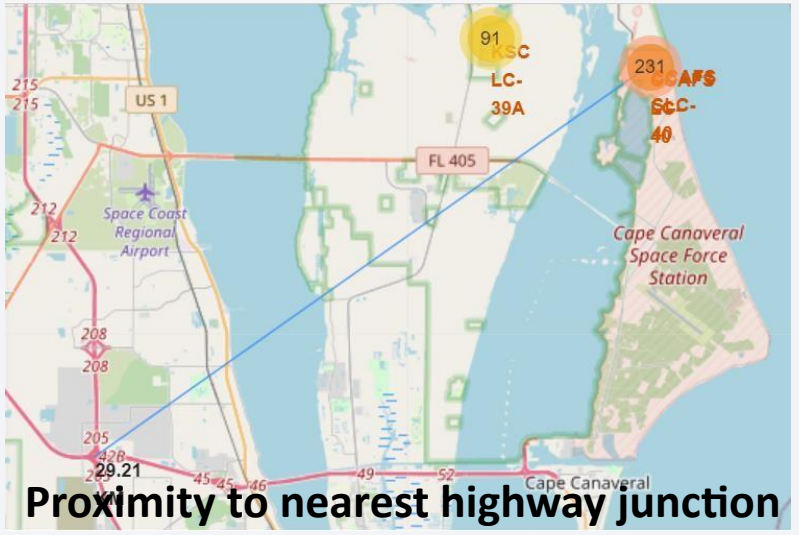


Proximity to nearest city

Proximity to nearest railway station



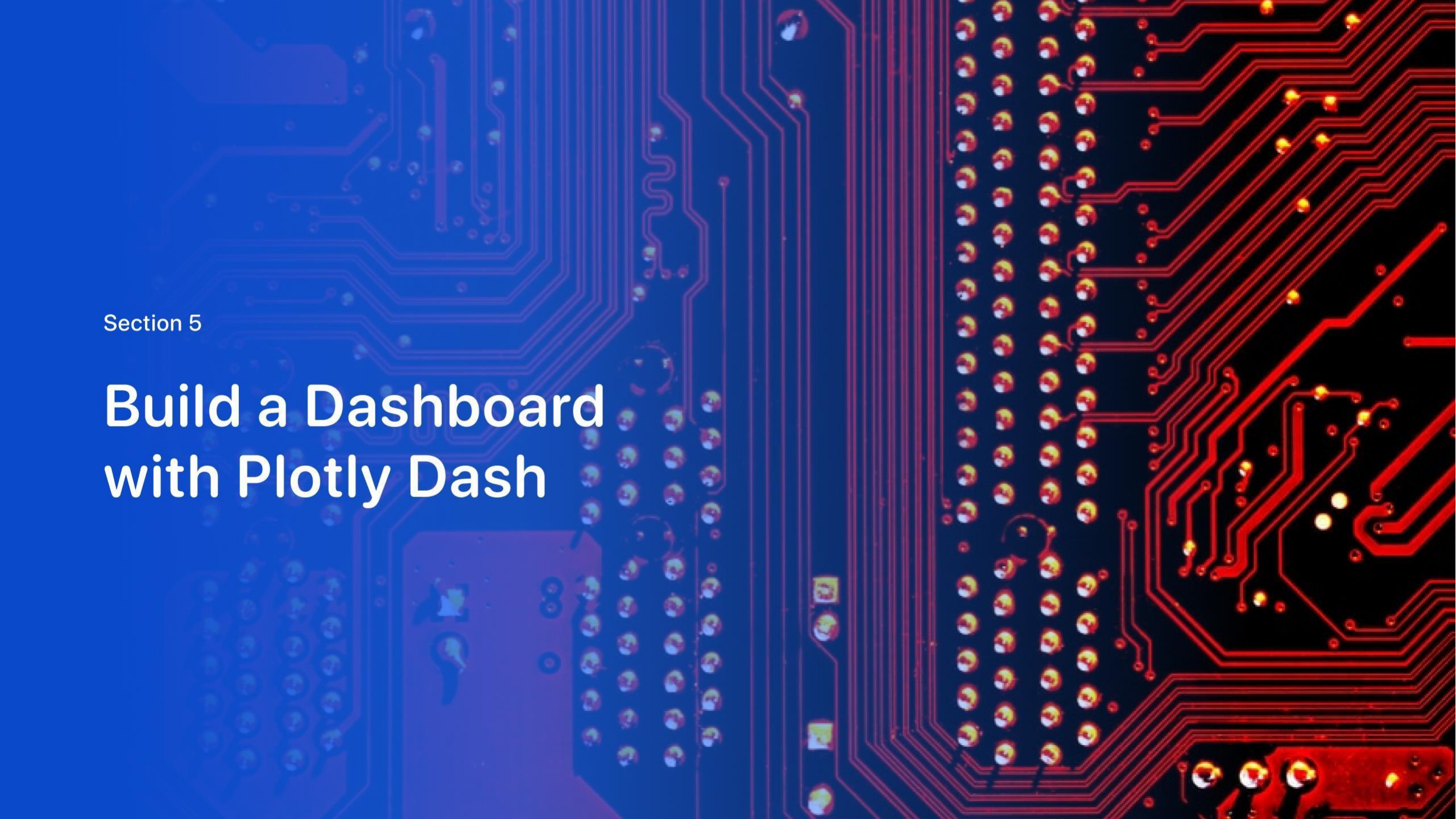
Proximity to nearest coastline



Proximity to nearest highway junction

Insights

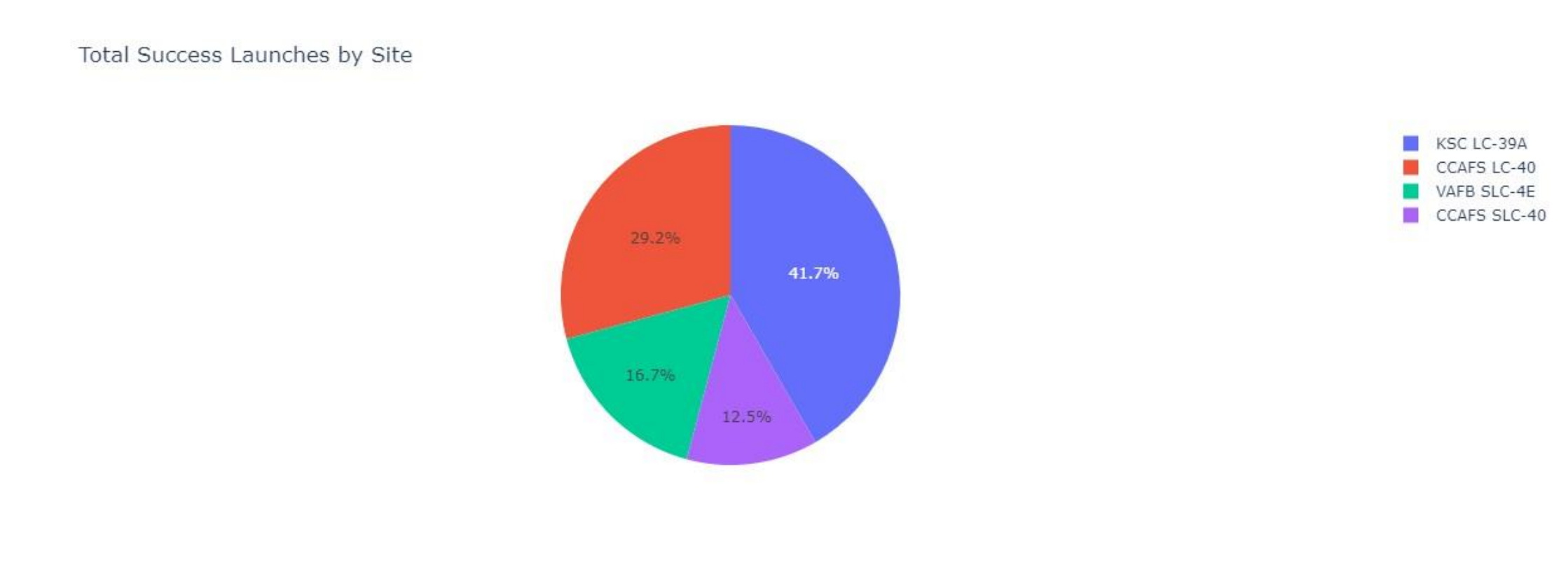
- Each screenshot shows proximity to key locations in km for launch site KSC LC-39A.



Section 5

Build a Dashboard with Plotly Dash

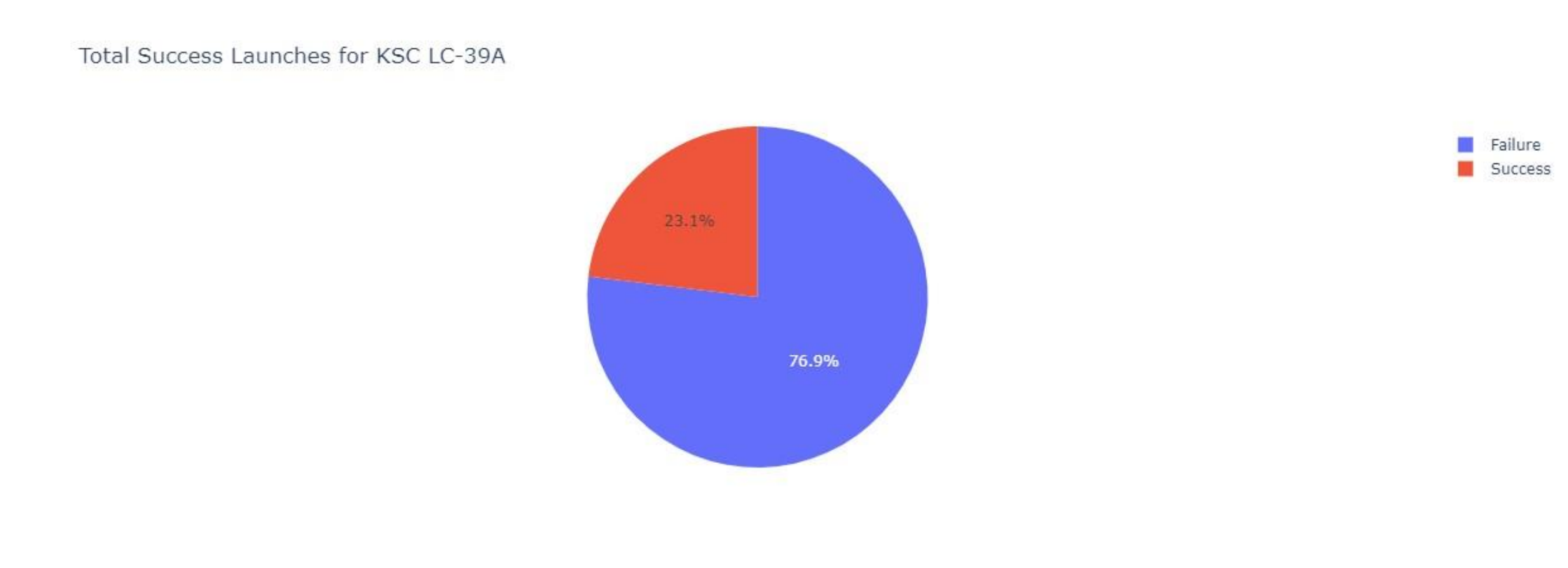
Plotly Dash: Successful Stage 1 Landings By Launch Site



- Insights

We can see that most successful landings were launches from KSC LC-39A. The least successful landings were launches from CCAFS SLC-40.

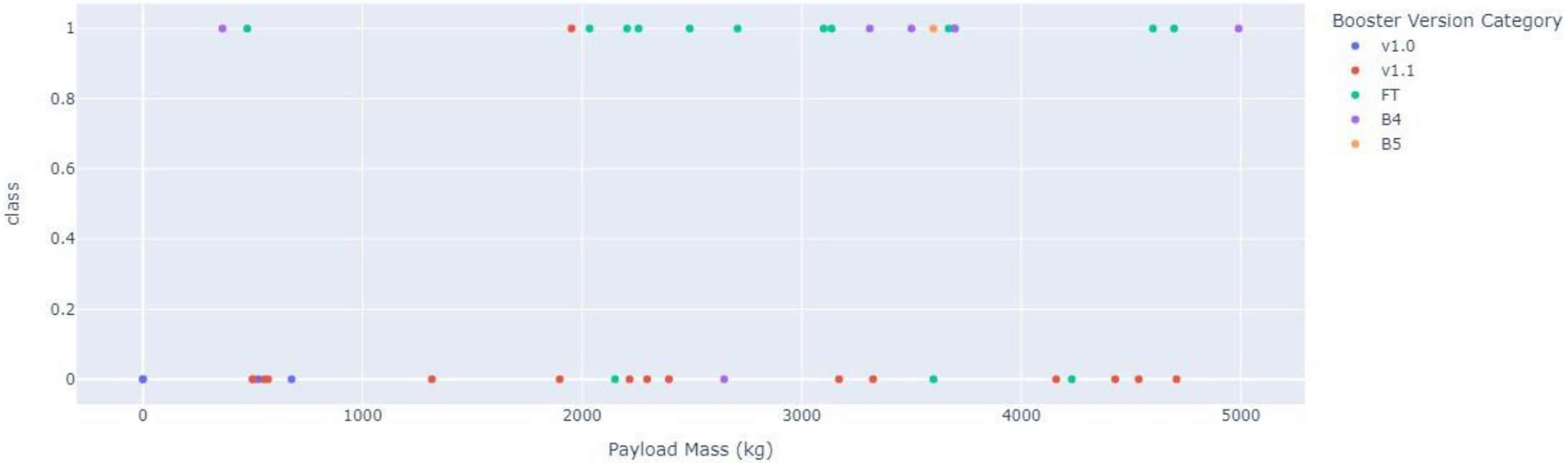
Plotly Dash: Successful Stage 1 Landings for KSC LC-39A



- Insights

Drilling into the most successful launch site, we can see the success vs. failure for KSC LC-39A. Even though many of the population successes are from this launch site, it actually has a low success rate.

Plotly Dash: Payload Mass vs. Success vs. Booster Version Category



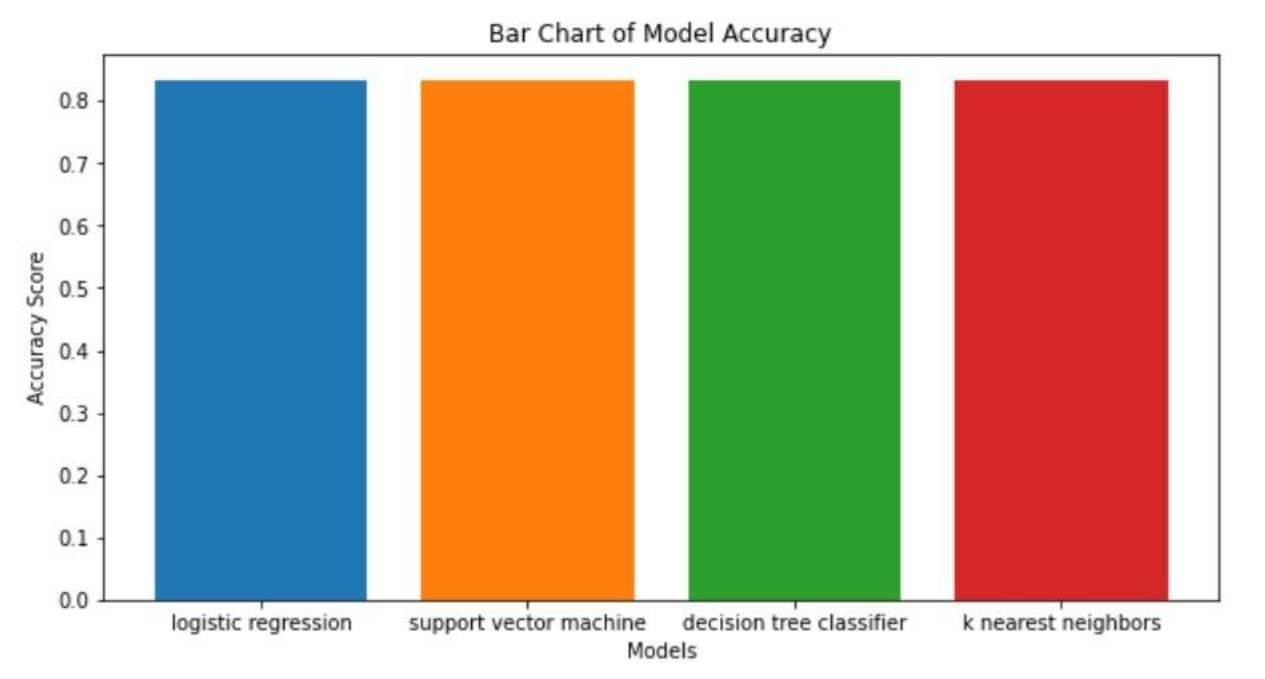
- Insights

We can see that booster version category FT has many successes and few failures. In contrast, v1.1 has many failures and few successes. There are recorded failures with payload mass of 0kg, this may be a data entry error.

Section 6

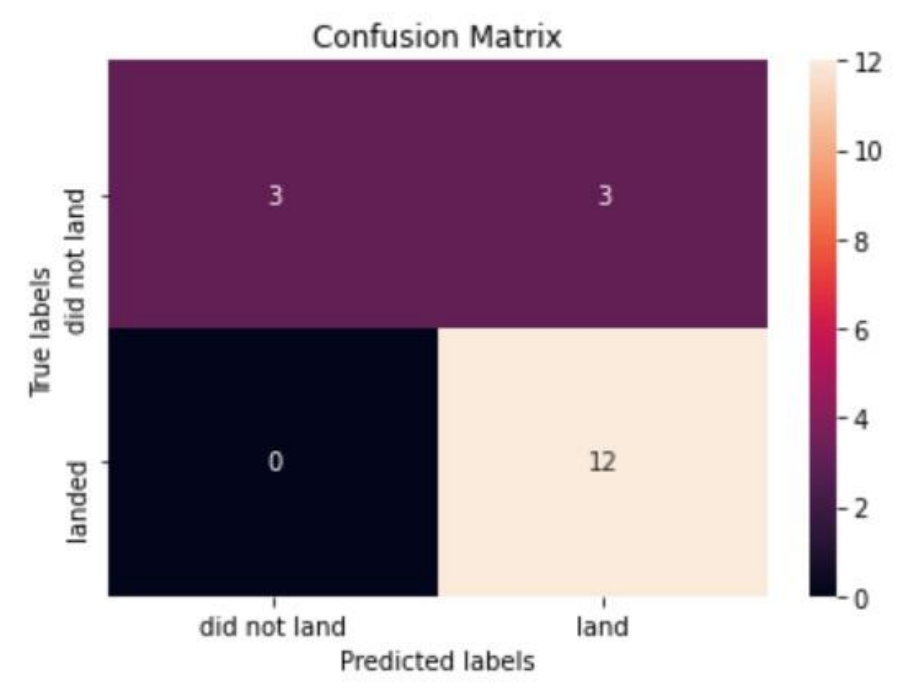
Predictive Analysis (Classification)

Classification Accuracy



- All models produced the same accuracy against the test data set (~83.33%).
- This is likely due to the limited data used, training and testing the models on larger data sets may produce more varied results.

Confusion Matrix



- All models generated the same confusion matrix.
- The models correctly predicted all successful landings as successful landings. The models only correctly predicted half of unsuccessful landings. The models over predict successful landings, they tend to give false positives.

Conclusions

- Our goal was to develop a machine learning model to predict if stage 1 will successfully land for a given launch.
- We developed four machine learning models, which all predicted successful landings with ~83.33% accuracy for some test data. The models tend to over predict successful landings, the models could be improved by using more data.
- In addition, we found that:
  - Success rate of stage 1 landings has improved over time.
  - ES-L1, GEO, HEO, SSO orbits have the best success rate.
  - Launch sites are typically located close to coastlines.

Appendix

- GitHub URL

- Special thanks to IBM instructors for delivering the IBM Data Science Professional Certificate courses!

- https://github.com/neelapala
- hema/data-collection

Thank you!