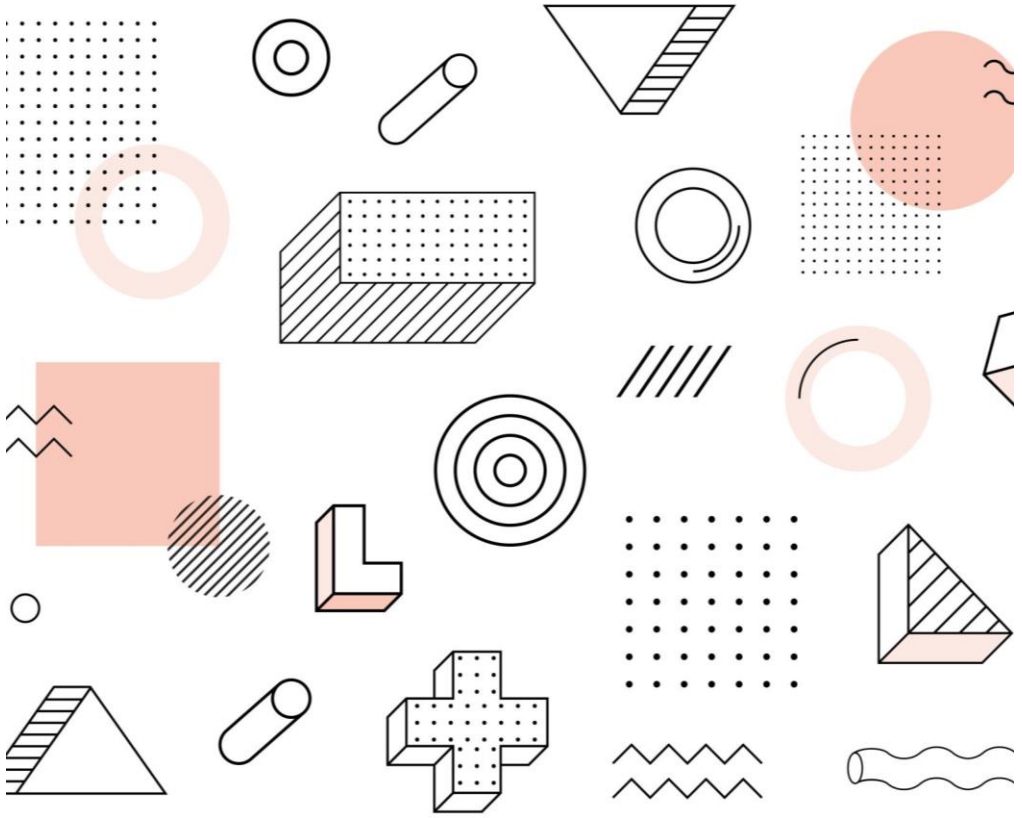
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

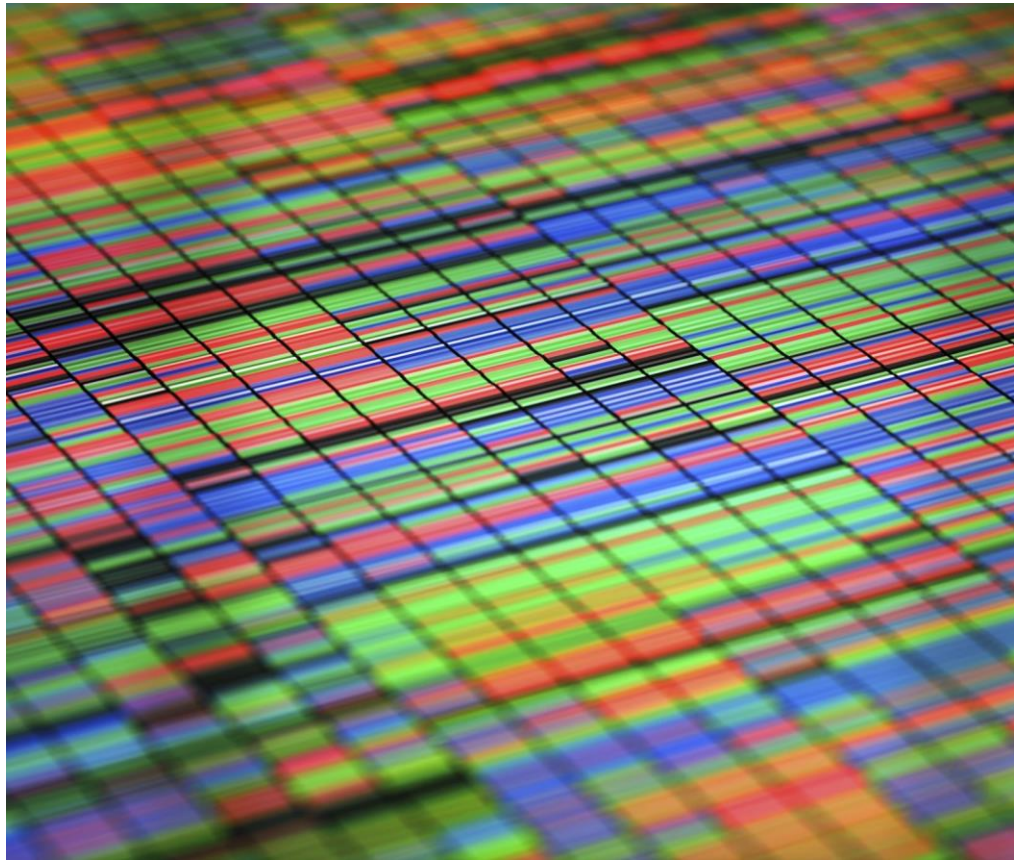
Thomas York
17/09/2023

Contents



Executive Summary

- Summary of methodologies
  - Data collection.
  - Data wrangling.
  - Exploratory data analysis with data visualization.
  - Exploratory data analysis with SQL.
  - Build an Interactive Map with Folium.
  - Build a Dashboard with Plotly Dash.
  - Predictive Analysis (Classification).
- Summary of results
  - Able to train a set of models to ~ 83.3% accuracy.
  - Exploratory data analysis results.
  - Interactive analytics demo in screenshots.
  - Predictive analysis results.



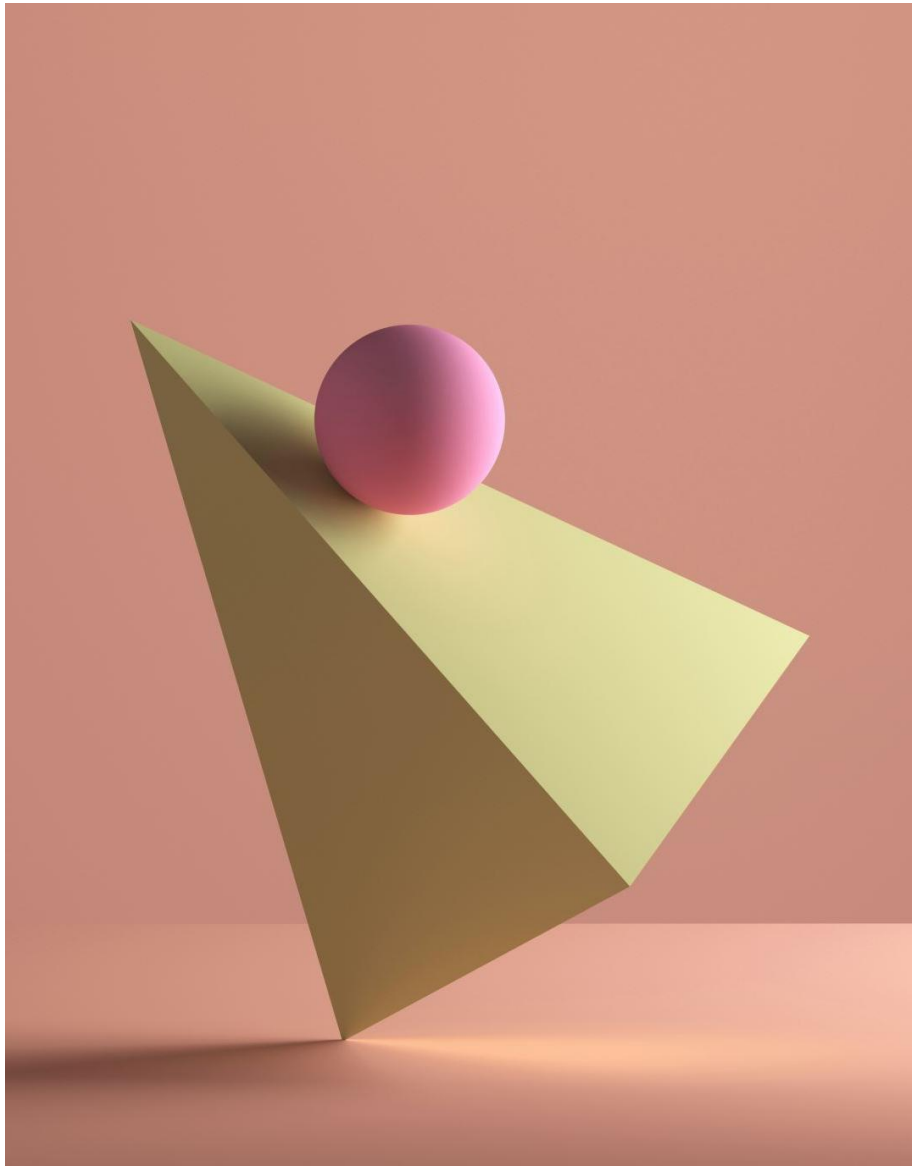
Introduction

Background and context

We aim to predict if SpaceX will reuse the first stage Falcon 9 boosters. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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.

Questions to be answered

- ❑ What is the likelihood of a Falcon 9 booster landing successfully?
- ❑ Which factors, such as orbit type and payload mass, have a significant correlation with landing success?
- ❑ Which classification model can provide the most accurate predictions?



Section 1

Methodology

Methodology

Executive Summary

- Data collected via:
  - Request from the SpaceX Rest API.
  - Web-scraping of the SpaceX Wikipedia page.
- Performed data wrangling
  - Missing or null records were removed, and remaining data was filtered for relevance.
  - Employed one hot encoding on remaining categorical data points.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Employed, Evaluated and compared a set of classification models to establish the most accurate.

Data Collection

- To obtain complete information on the launches, we engage in a combination of get requests and web scraping.
- We request a response from the API: https://api.spacexdata.com/v4/launches/past. We proceed with a series of get requests to receive the information detailed to the right.
- To complete our collection, we scrape the Wikipedia page https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches. We extract Falcon 9 launch records, with specific data points listed to the right.

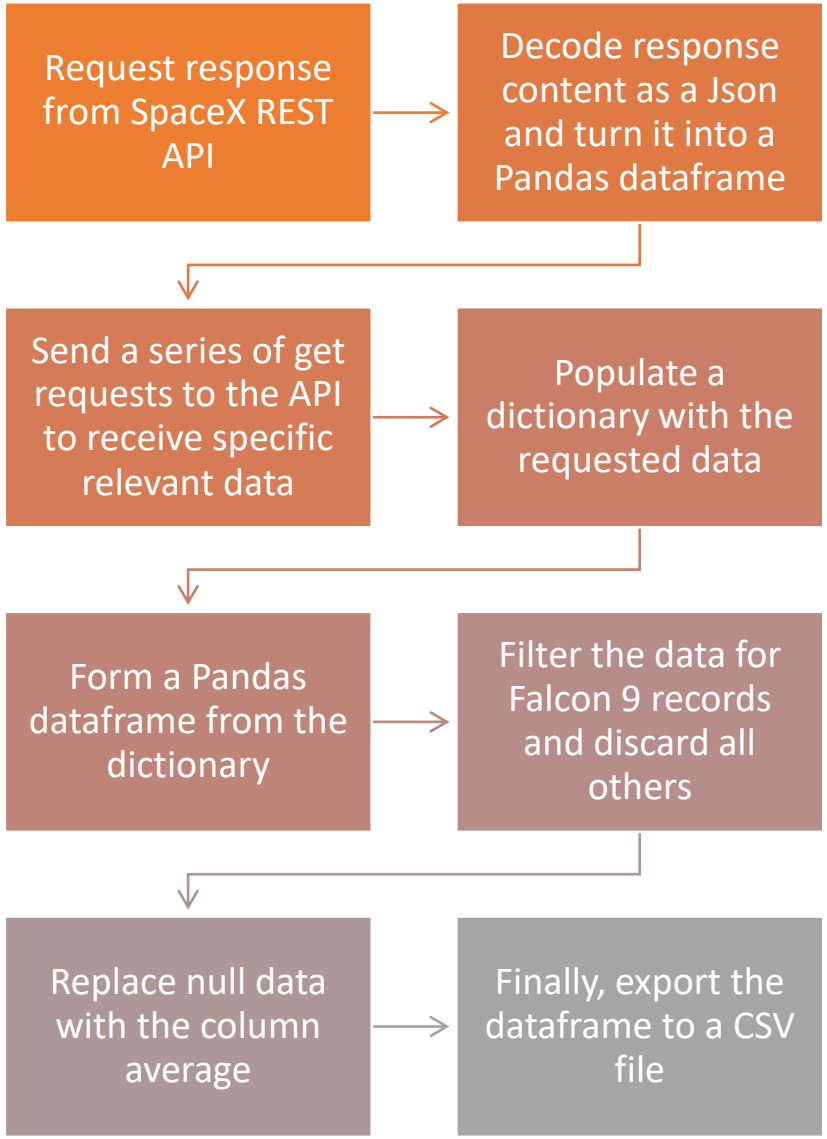
Data collected via SpaceX REST API

- Rocket (booster name)
- Payload (mass, orbit)
- Launchpad (site name, latitude and longitude)
- Cores (landing outcomes, landing type, landing pad, reused cores, gridfin usage, leg usage, core block, core serial number)

Data collected via Web Scraping

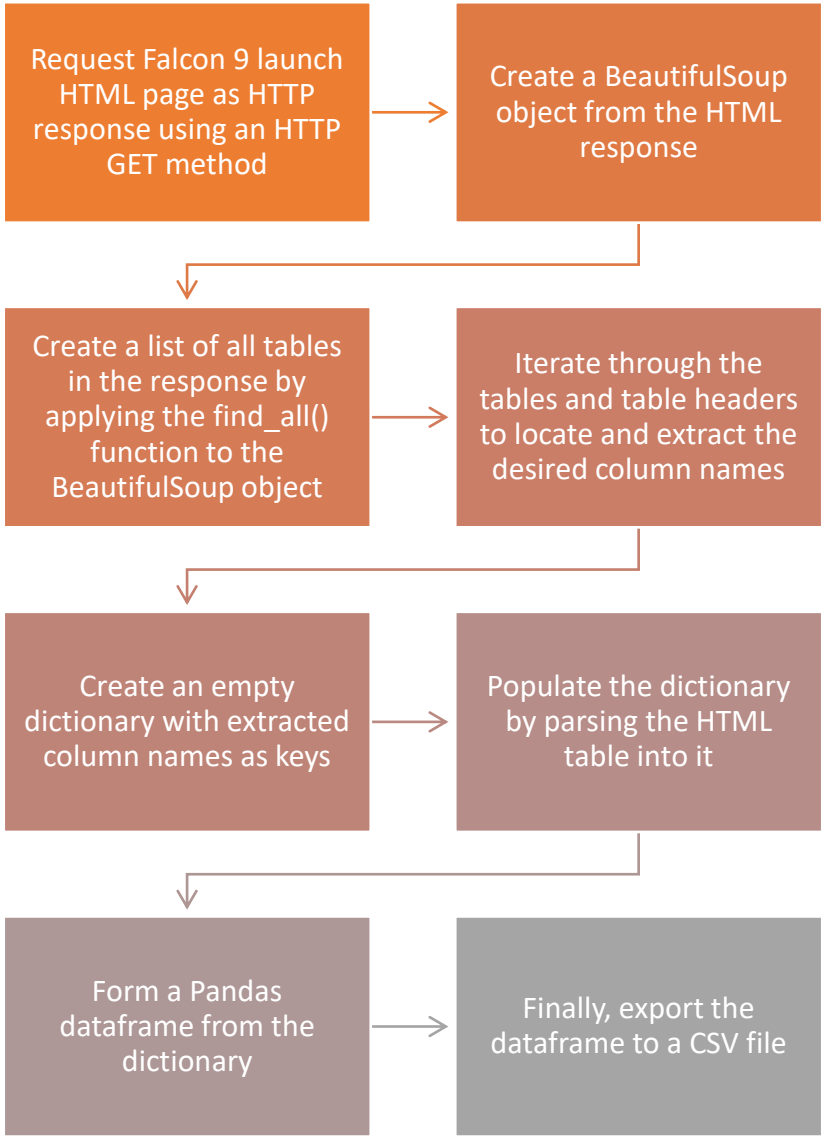
- Flight number, date and time
- Payload and mass
- Launch site, launch outcome
- Orbit
- Customer
- Booster version, booster landing

Data Collection – SpaceX API

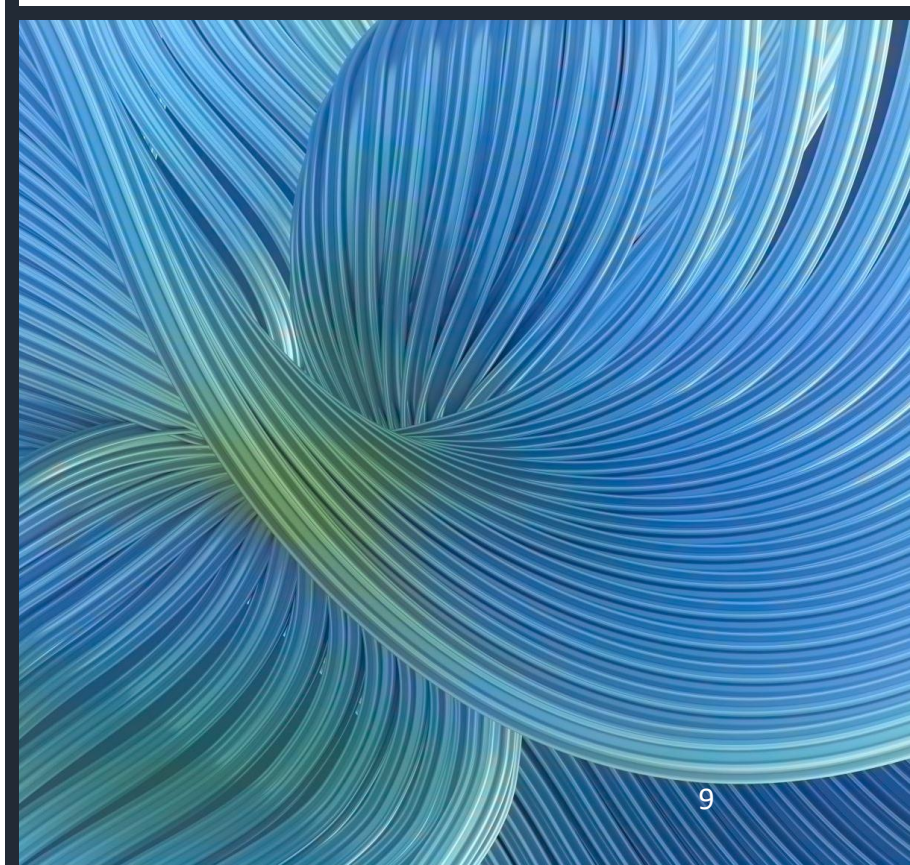


Github URL: Data Collection SPACEX REST API

Data Collection - Web Scraping

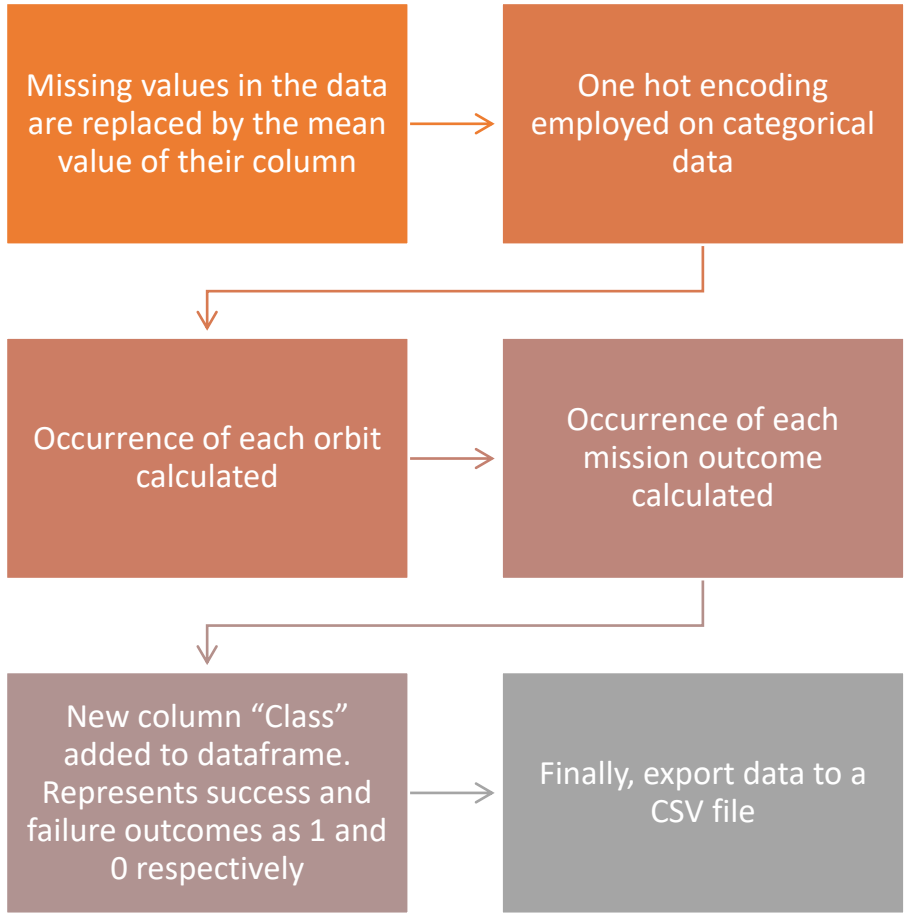


[Github URL: Data Collection Web Scraping]



Data Wrangling

- The flow chart is straightforward and typical for data wrangling, but it's worth expanding on the new "Class" column created for our dataframe.
- The data collection yielded a set of launch records with a variety of landing outcomes such as "True Ocean", "True RTLS", "False ASDS". These specify both where the landing was attempted, such as on the ocean or land, and if it was a success. We are only interested in the outcome of a landing, not where it was attempted so we map every success to integer 1 and every failure to 0 and store this result in the column "Class".

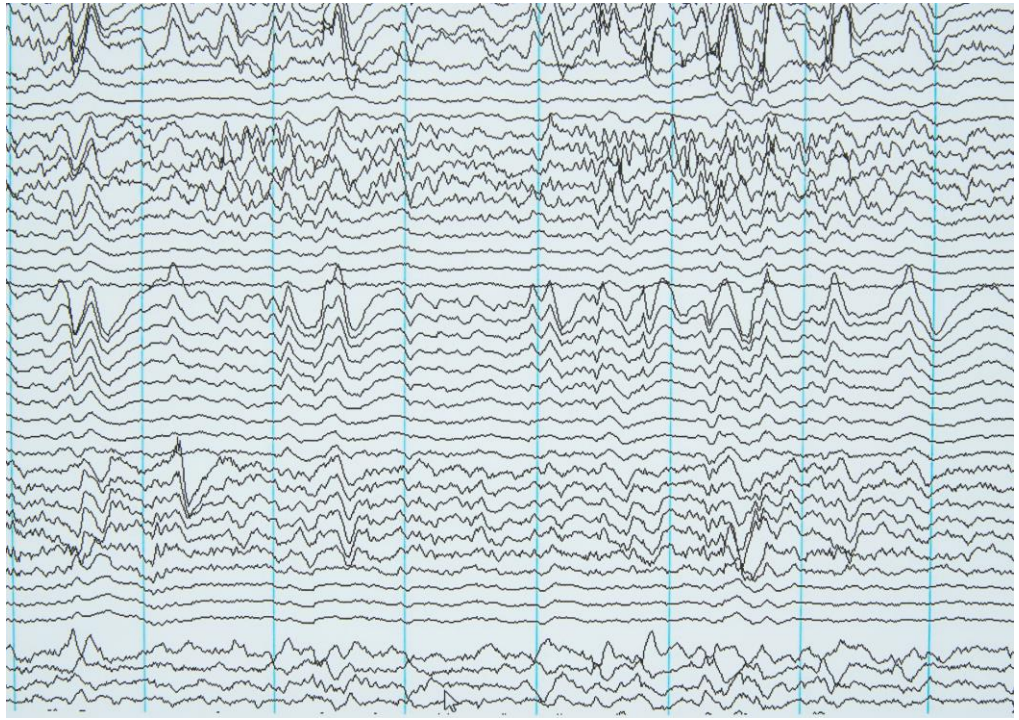


Github URL: Data Wrangling

EDA with Data Visualisation

Charts Plotted

- Scatter plots:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Flight Number vs. Orbit
  - Payload Mass vs. Orbit
- Bar chart:
  - Orbit vs. Success Rate
- Line chart:
  - Year vs. Success Rate

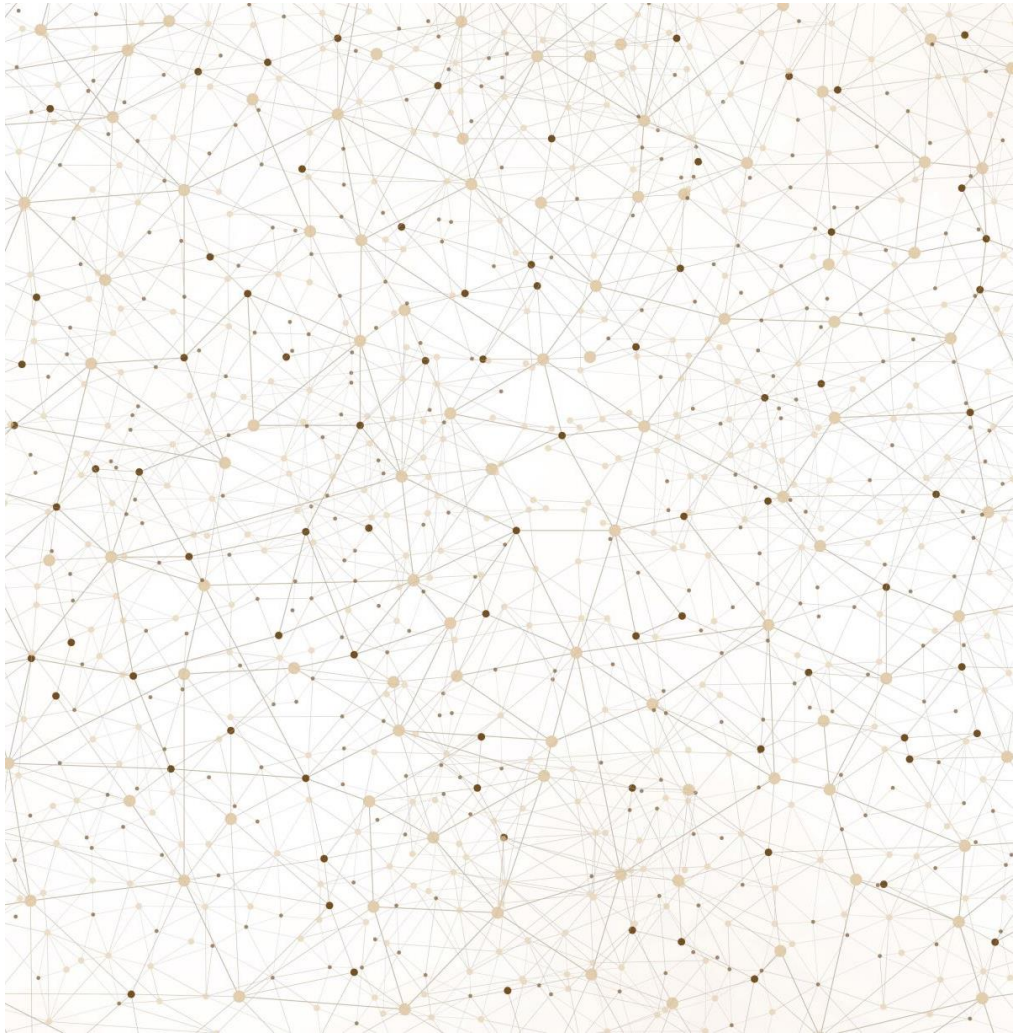


Github URL: EDA Visualisation

EDA with SQL

SQL Queries

- Display the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin with the string '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 records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months 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: EDA SQL

Build an Interactive Map with Folium

Map Objects - We instantiate a folium Map object with the NASA Johnson Space Centre coordinates.

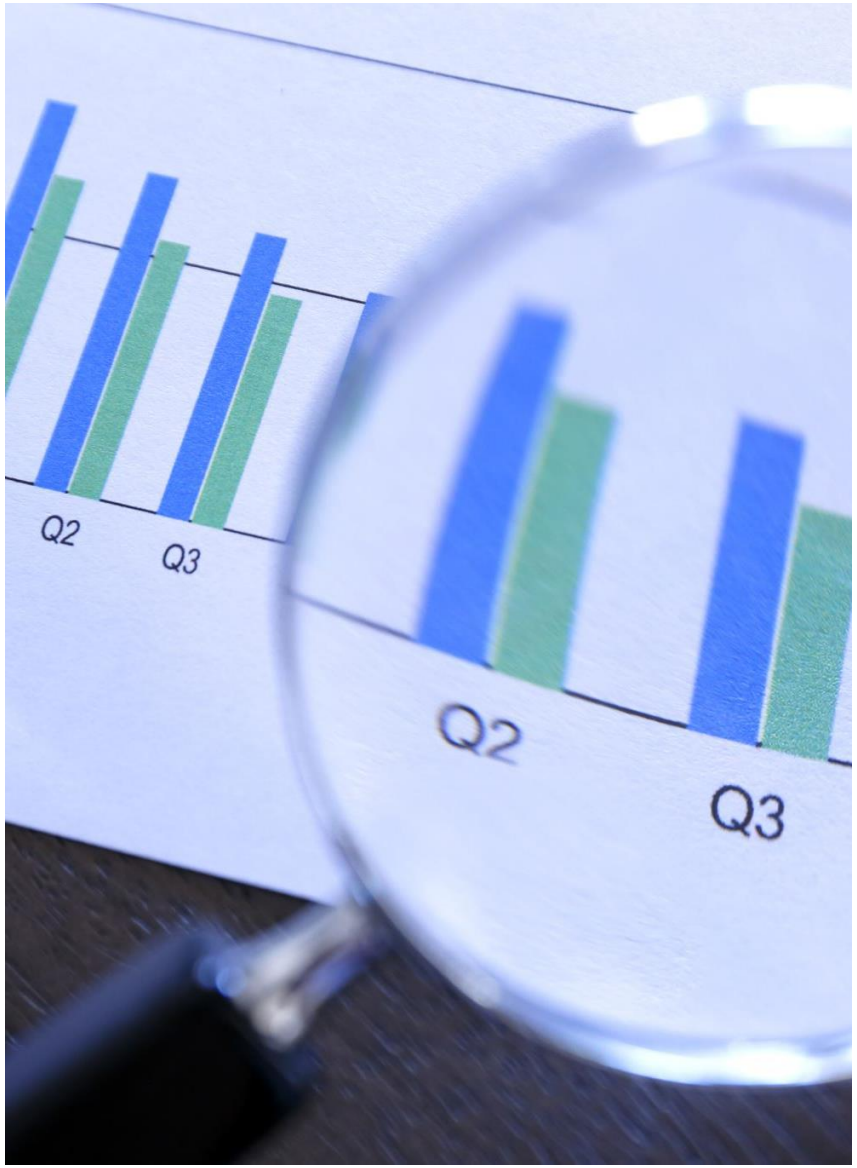
- Circle launch sites:
  - Added a folium.Circle and folium.Marker to each of the 4 launch sites.
  - This allows us to easily focus in on a particular site.
- Marker cluster outcomes:
  - Added clusters of coloured markers to each launch site indicating the outcome of a launch.
  - This allows us to easily visualise the successes and failures at each site with green markers indicating success, red marking failure.
- Marked launch site proximities:
  - Calculated the distance between launch site centre and closest points of interest, such as coastline and railway.
  - Added lines between launch site centre and its proximities.

Github URL: Visualisation Folium

Build a Dashboard with Plotly Dash

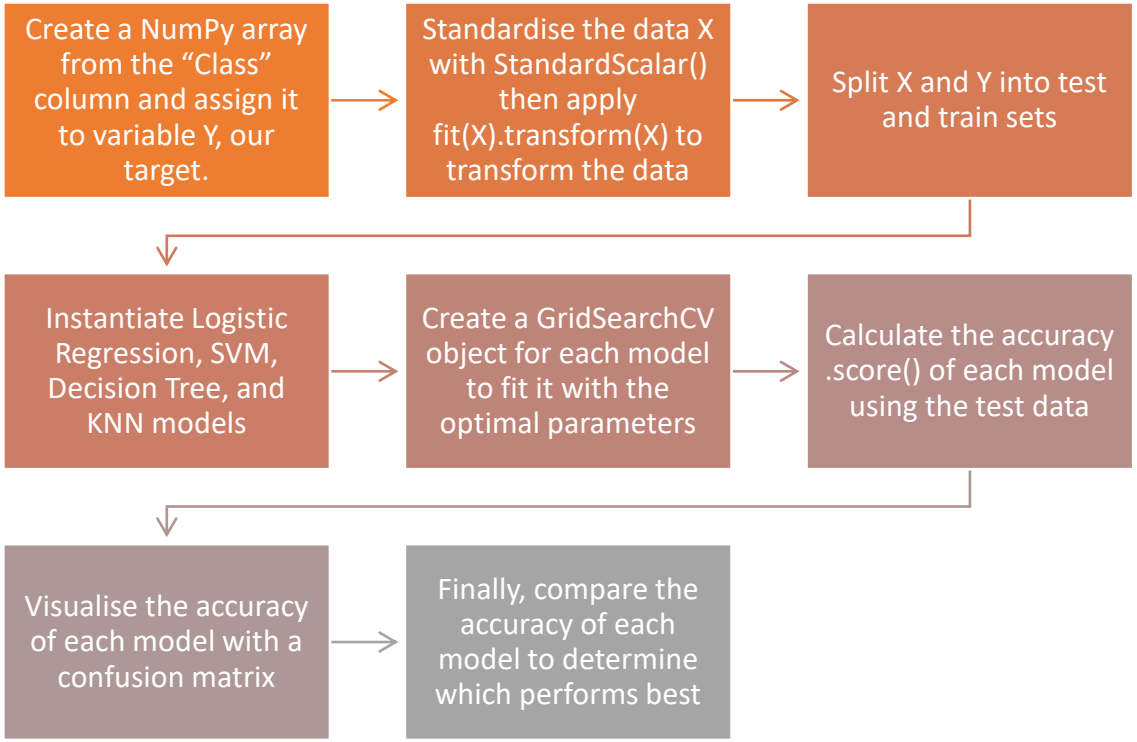
Dashboard Components - Plots and Graphs

- Interactivity:
  - Included a dropdown list to enable launch site specification. The plots and graphs update to user input.
  - Included a sliding scale to allow user to specify a mass range. Plots update according to user input.
- Pie chart:
  - Interactive chart to display proportion of successful launches across all launch sites or a site as specified by user.
  - This allows for detailed comparison of success across the four sites.
- Scatter plot:
  - Interactive scatter plot to visualise success and failure across varying payload mass for different booster versions.
  - This allows us to spy correlation between these variables across a user specified mass range.



Github URL: Spacex Dash App

Predictive Analysis (Classification)



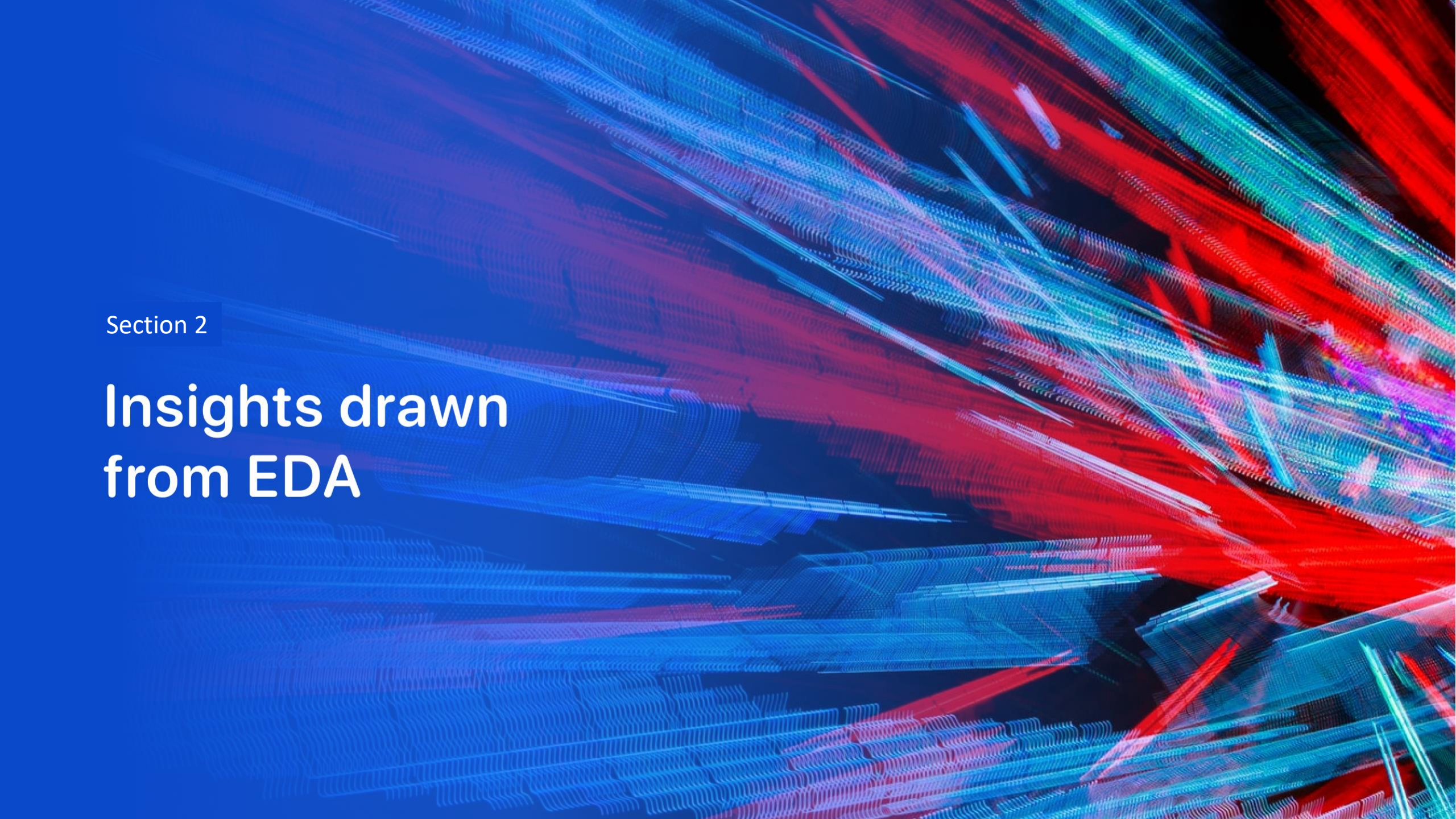
[Github URL: Machine Learning Prediction]

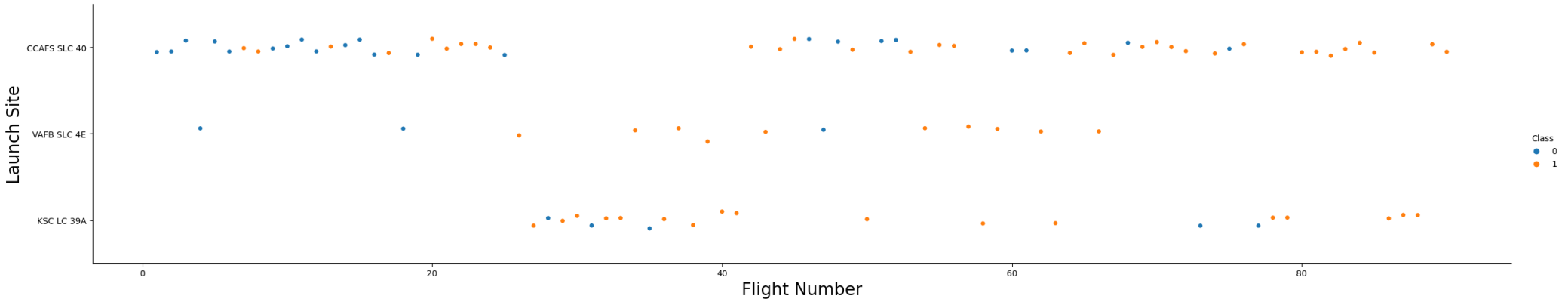
Results

Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

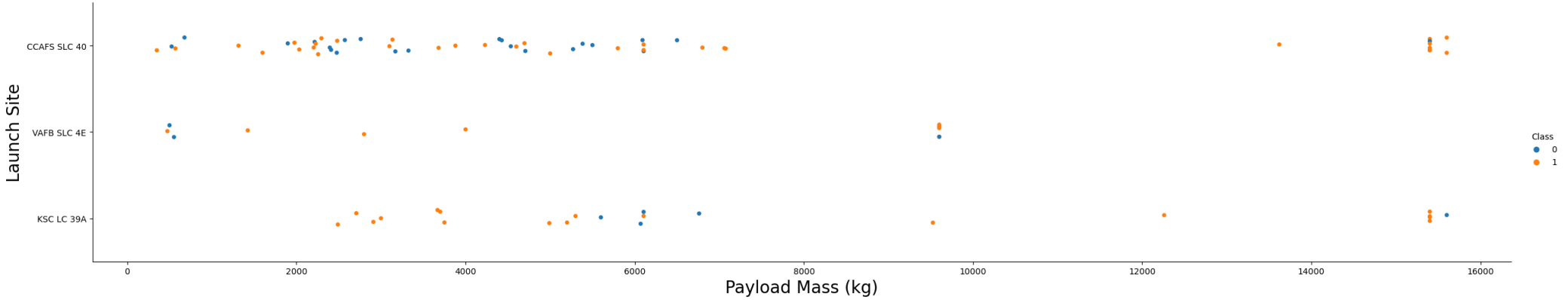

Section 2
Insights drawn from EDA



Flight Number vs. Launch Site

Insights

- Across all launch sites, the latest flights have all been successful and the rate of success has increased over time.
- Recalling our definition of Class, we see successful flights in orange and failures in blue.



Payload vs. Launch Site

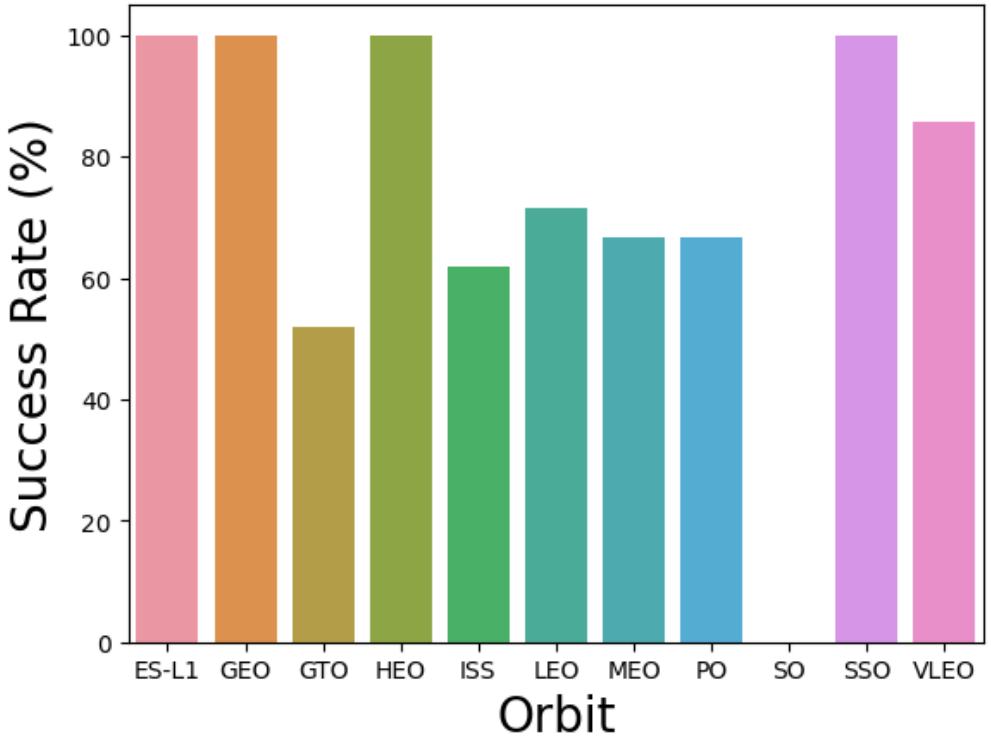
Insights

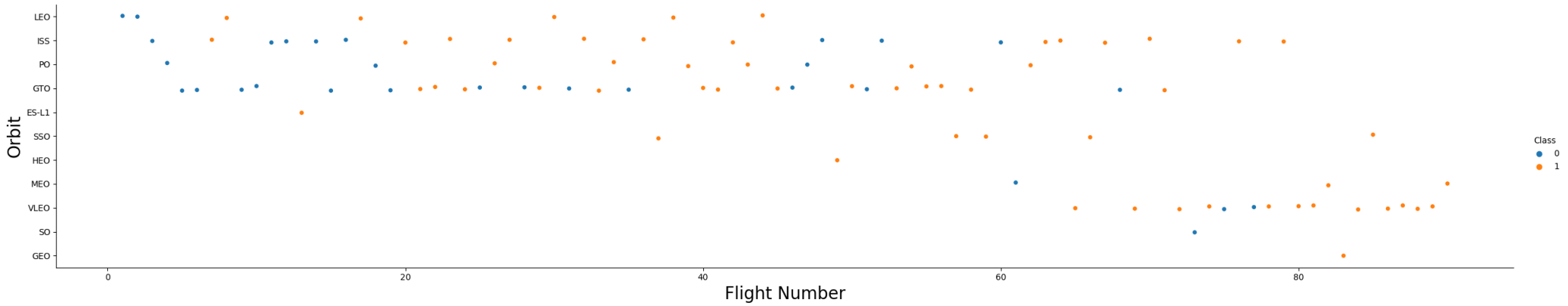
- Across each launch site we observe that a higher payload mass is correlated with a higher success rate. Most flights above 7000kg were a success.
- The KSC LC 39A launch site, however, is very successful with low payload mass (below 5500kg).

Success Rate vs. Orbit Type

Insights – Success Rates

- ❑100%: ES-L1, GEO, HEO, SSO.
- ❑50%-85%: GTO, ISS, LEO, MEO, PO, VLEO.
- ❑0%: SO.

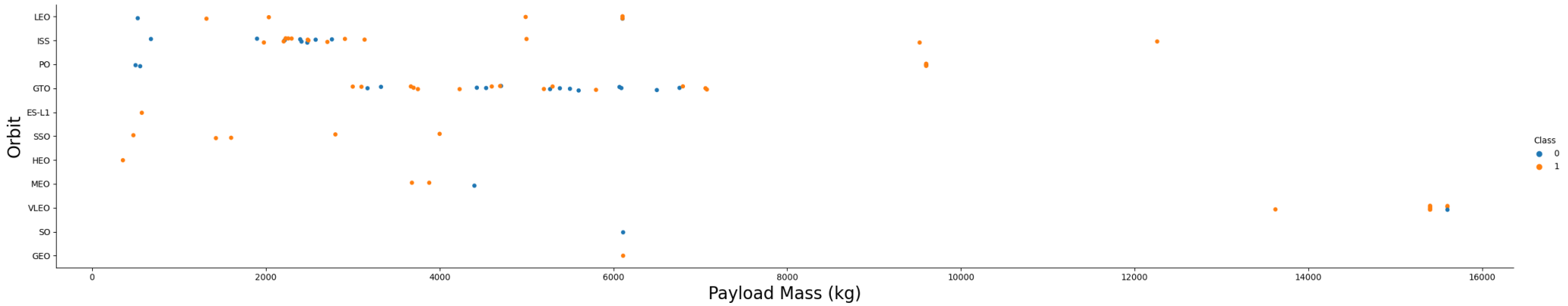




Flight Number vs. Orbit Type

Insights

- Across all orbits we observe that the most recent flights are the most successful. The strongest correlation is seen in the LEO orbit, however, there appears to be very little correlation for the GTO orbit.
- The SSO, HEO, GEO and ES-L1 orbits have no failures at all, however each has many fewer total flights than the other orbits.



Payload vs. Orbit Type

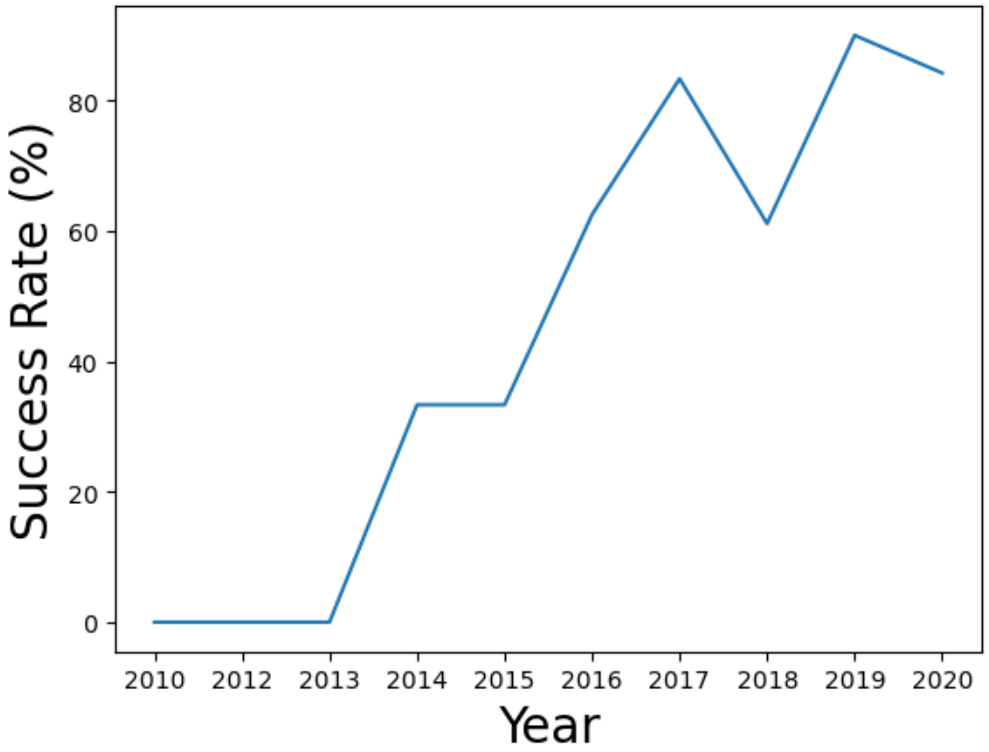
Insights

- Across all orbits we observe that the heavier payloads, say above 7000kg, are highly successful, although the sample of flights above this payload is small.
- The correlation between the two variables appears to negatively impact success for GTO orbits whereas Leo orbits increase in success as payload increases.

Launch Success Yearly Trend

Insights – Success Rates

- Success rates have seen a net increase since 2013 through 2020.
- 2018 saw a dip in success and again in 2020.
- 2019 is the year with the highest observed success rate.



All Launch Site Names

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

Query

```
%%sql
SELECT DISTINCT Launch_Site
FROM SPACEXTABLE;
```

Explanation

- Displays the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Query

```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE Launch_Site LIKE "CCA%"
LIMIT 5;
```

Explanation

- Displays 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

TOTAL_PAYLOAD_NASA
48213

Query

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_NASA
FROM SPACEXTABLE
WHERE Customer LIKE "%NASA (CRS)%";
```

Explanation

- Displays the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

AVG_PAYLOAD_F9
2534.6666666666665

Query

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_F9
FROM SPACEXTABLE
WHERE Booster_Version LIKE "%F9 v1.1%";
```

Explanation

- Displays average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

FIRST_SUCCESSFUL_LANDING
2015-12-22

Query

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESSFUL_LANDING
FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (ground pad)";
```

Explanation

- Displays the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Query

```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE LANDING_OUTCOME = "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

Explanation

- Displays the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	Total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Query

```
%%sql
SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Total
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

Explanation

- Displays the total number of success and failure mission outcomes.
- (Seems as though an error in the data has caused one mission success to take a unique entry in the table)

Boosters Carried Maximum Payload

Booster_Version
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

Query

```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
                           FROM SPACEXTABLE);
```

Explanation

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

2015 Launch Records

Month	Landing_Outcome	Booster_Version	Launch_Site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Query

```
%%sql
SELECT CASE SUBSTR(Date, 6, 2) WHEN '01' THEN 'January' WHEN '02' THEN 'Febuary'
       Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTABLE
WHERE SUBSTR(Date, 1, 4) = "2015" AND Landing_Outcome = "Failure (drone ship)";
```

Explanation

- Lists the records which will display the month names, failure landing outcomes in drone ship, booster versions and launch site for the months in year 2015.

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

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

Query

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Total
FROM SPACEXTABLE
WHERE Date BETWEEN "2010-06-04" AND "2017-03-20"
GROUP BY Landing_Outcome
ORDER BY Total DESC;
```

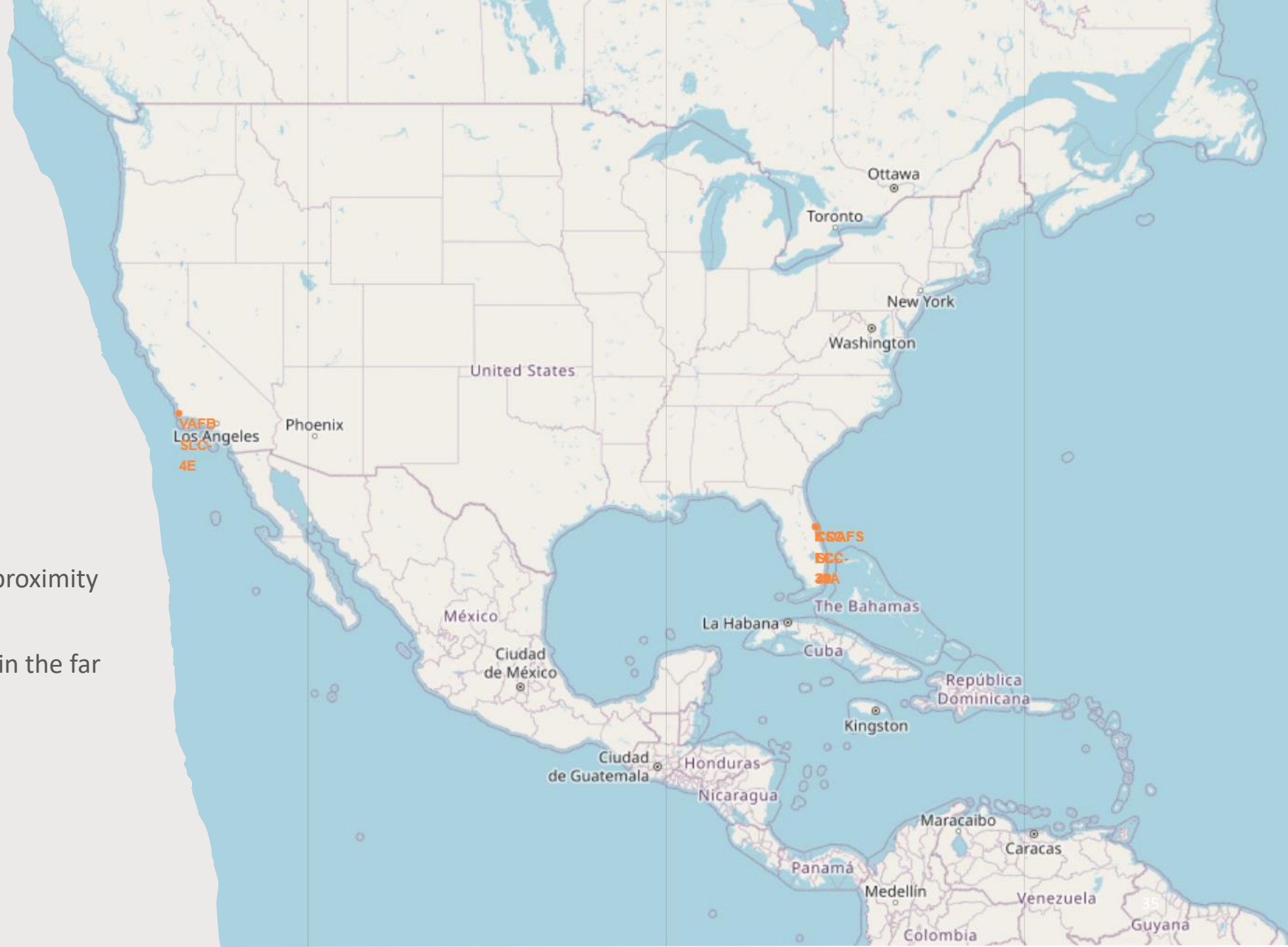
Explanation

- Ranks 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.

Section 3

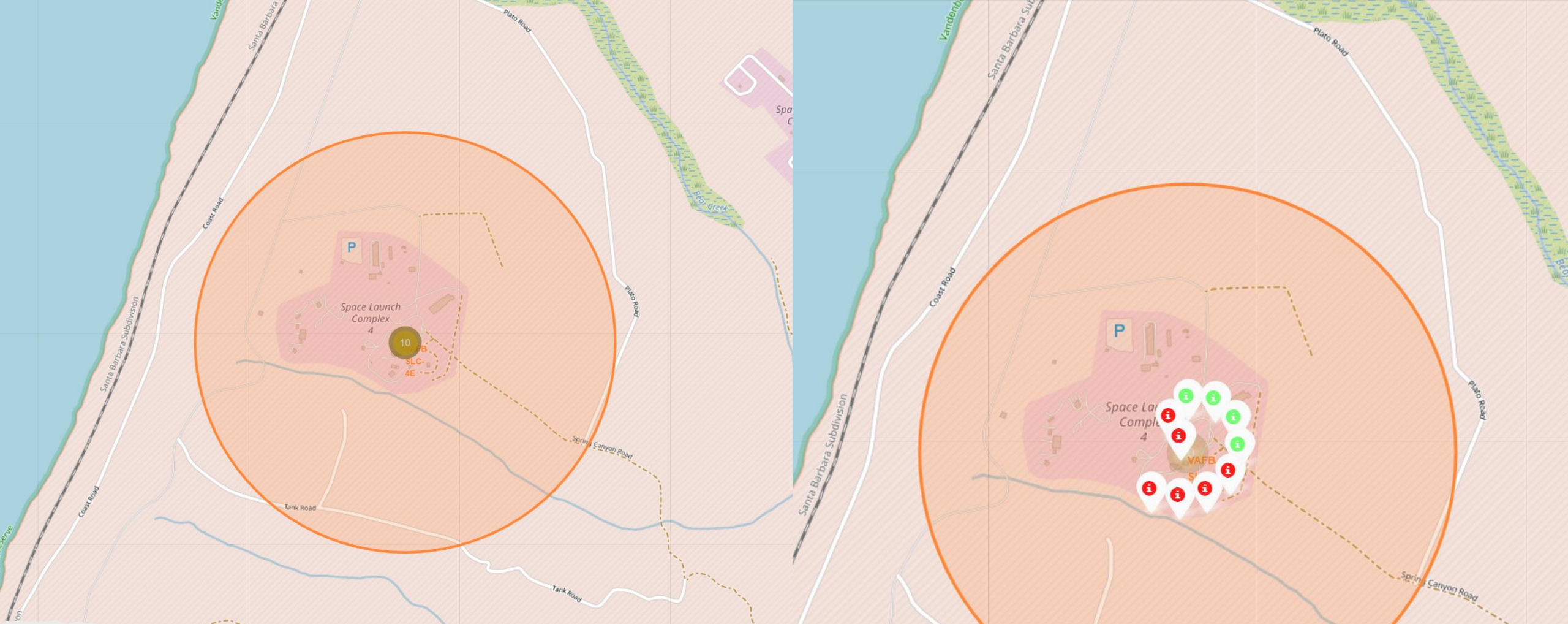
Launch Sites Proximities Analysis

Launch Site Locations



Observations

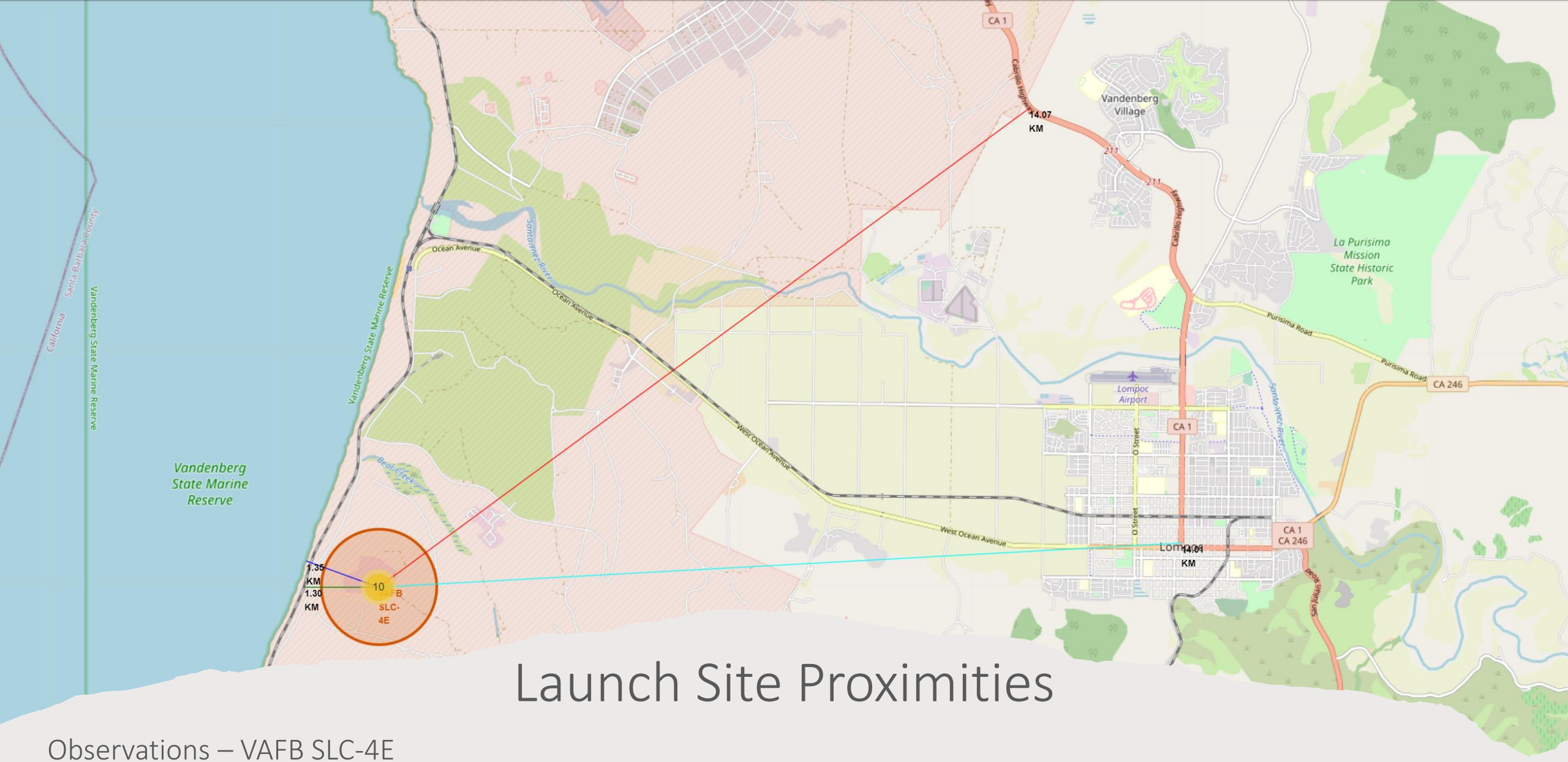
- Each launch site is in close proximity to the coastline.
- Each launch site is situated in the far south of the US, minimizing proximity to the equator.



Colour-Labeled Launch Outcomes

Observations – VAFB SLC-4E

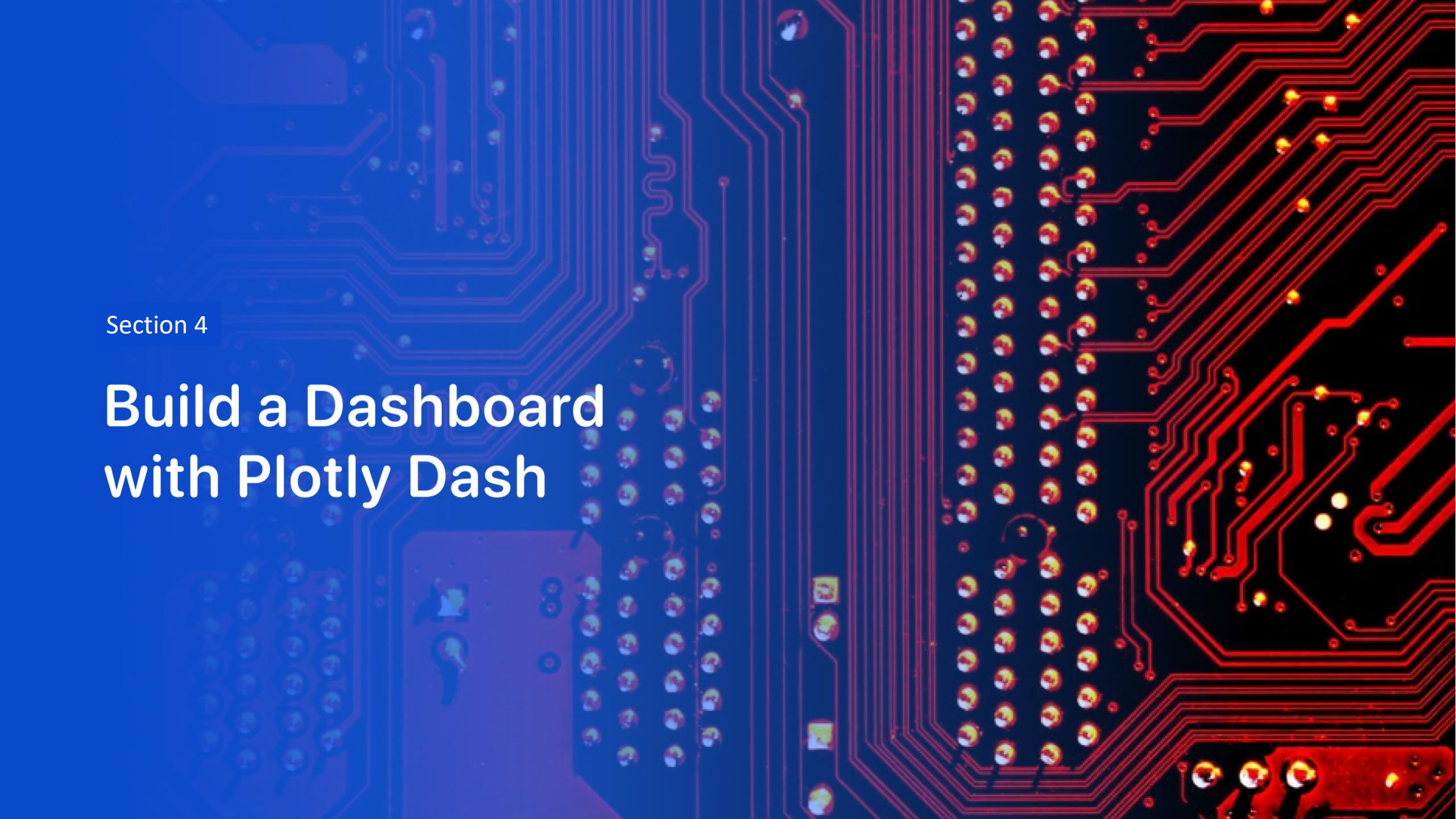
- ❑ We easily identify success and failures in green and blue respectively for each site.



Launch Site Proximities

Observations – VAFB SLC-4E

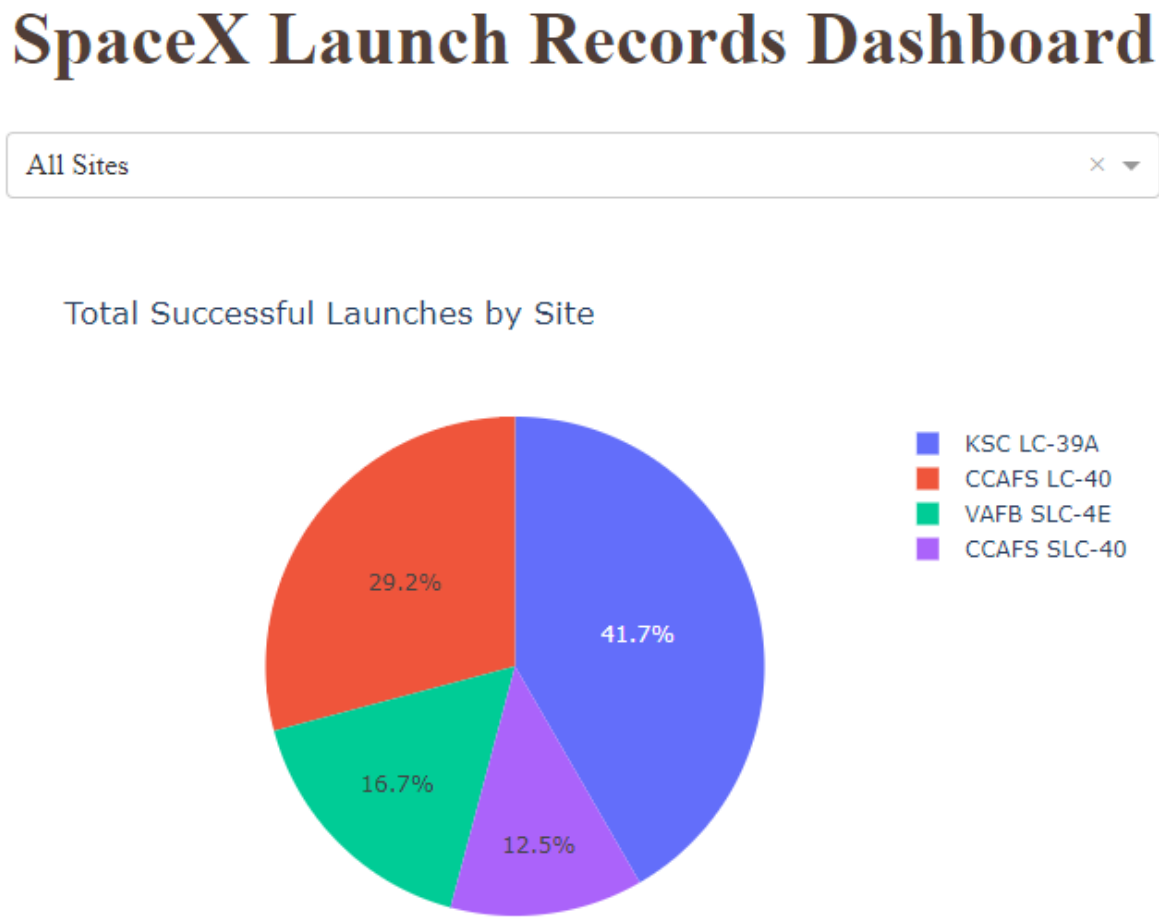
- We can easily identify the nearest coastline (blue, 1.35km), railway (green, 1.3km), City Lompoc (cyan, 14.01km) and major highway (red, 14.07km).


Section 4
Build a Dashboard with Plotly Dash

Landing Success by Launch Site

Explanation

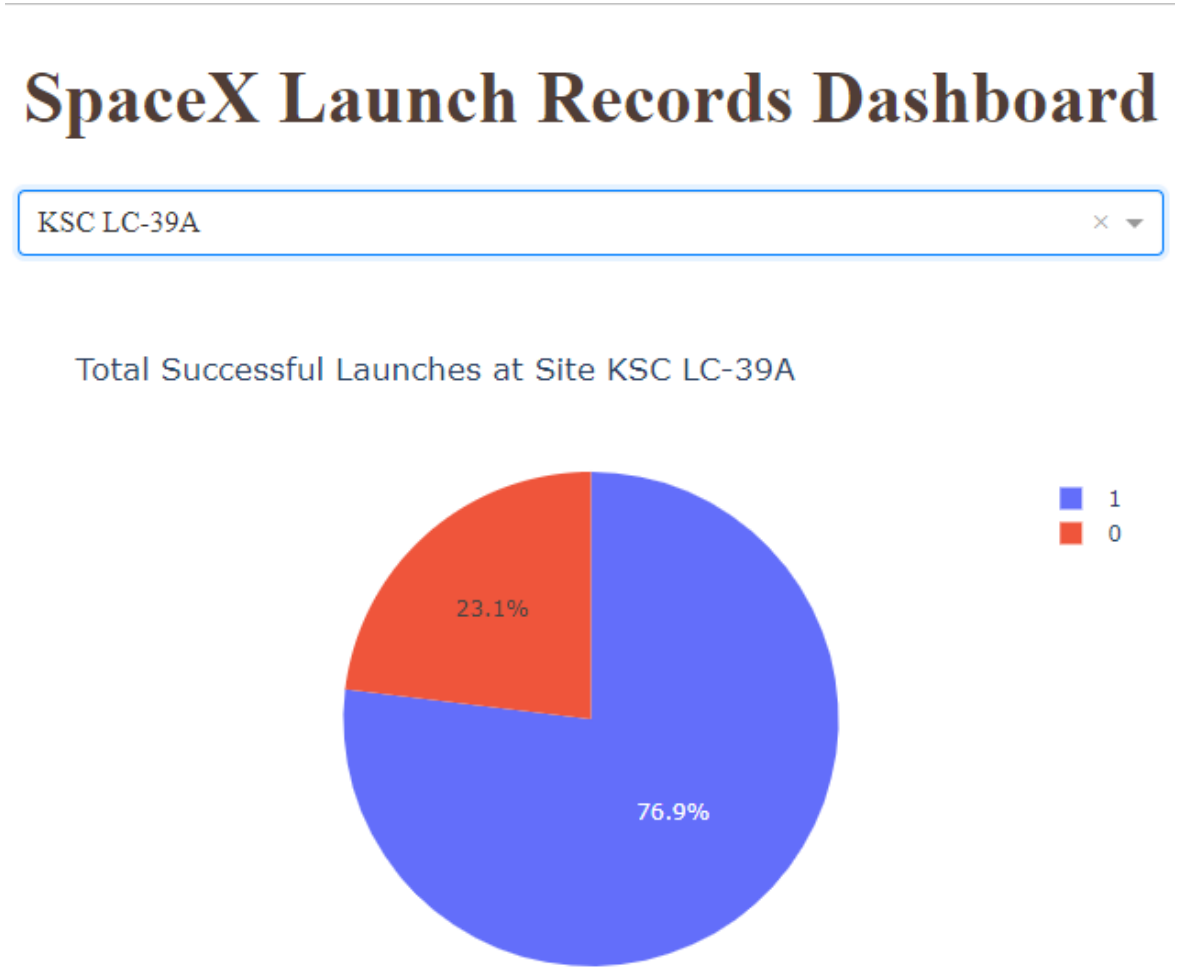
- We select “All Sites” from our Dashboard dropdown to render the displayed pie chart.
- From the chart we see that the most successful landings came from flights launched from the KSC LC-39A site with 41.7% of all successful launches.
- The least successful site is CCAFS SLC-40 with only 12.5% of total successful flights.



Most Successful Launch Site

Explanation

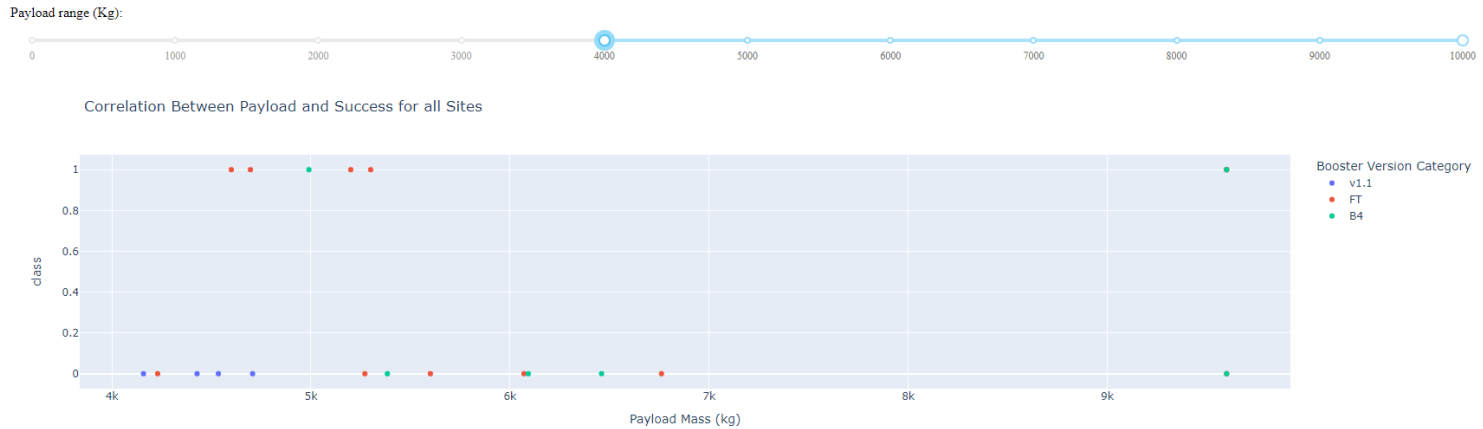
- We select “KSC LC-39A” from our Dashboard dropdown to render the displayed pie chart.
- In the previous slide, we observed that this was the site with the most successful landings, however, this doesn’t give us the full picture of the success rate. Perhaps this site launches a vast majority of total flights and has many more failures than successes.
- Visaulising the data for KSC LC-39A alone, we see that the success rate here is 76.9%.



Payload Mass vs. Success

Explanation

- We select “All Sites” from our Dashboard dropdown and we select a mass range from the “Payload range (kg)” slider.
- Displayed are two examples for the range (0, 6000) and (4000, 10000).
- Booster version is colour coded so we may identify which boosters are successful for our range of payload mass.
- Observe that payloads in the range (0, 6000) see the most success and the most successful booster category is FT.

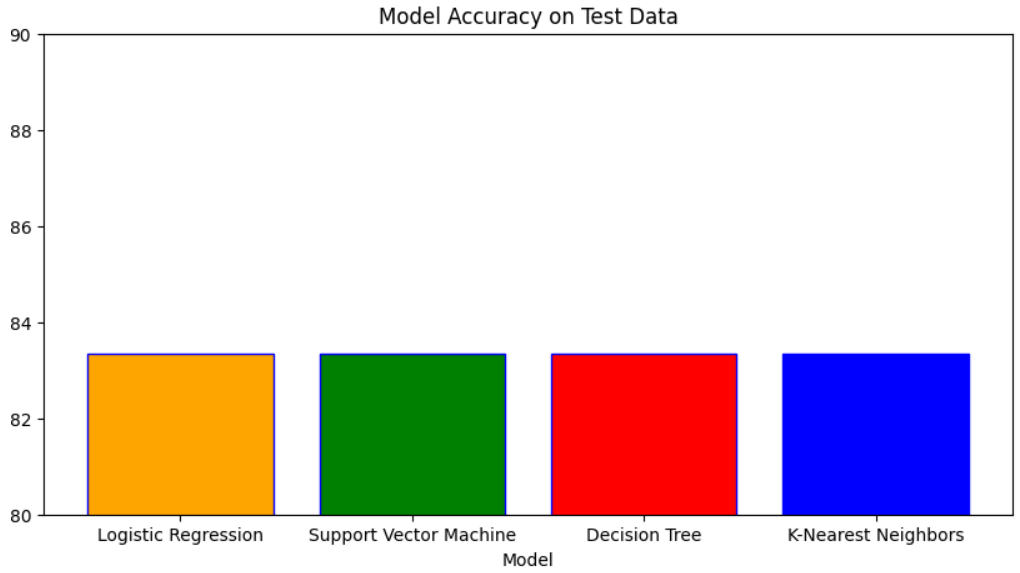


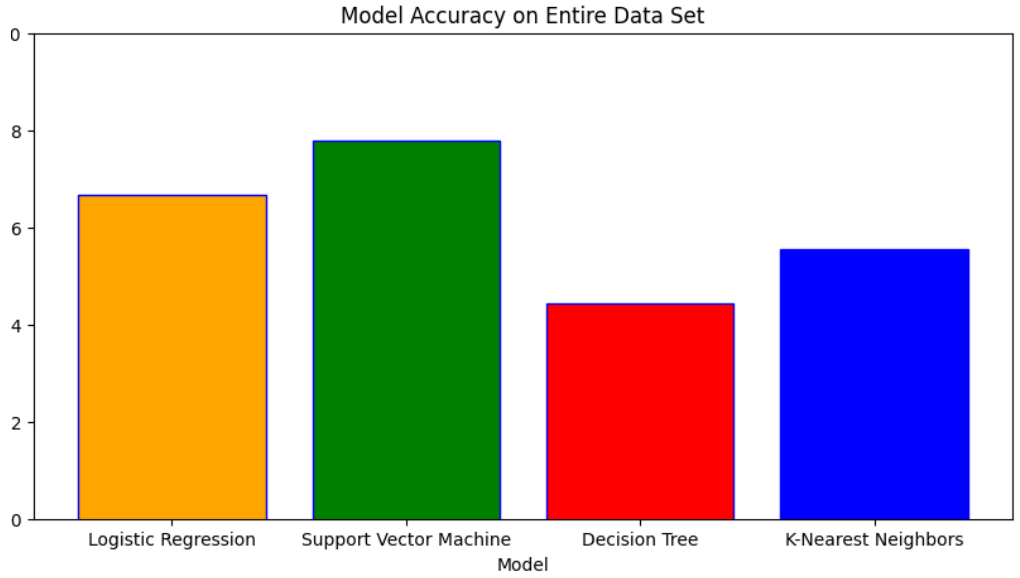
Section 5

Predictive Analysis (Classification)

Launch Success Yearly Trend

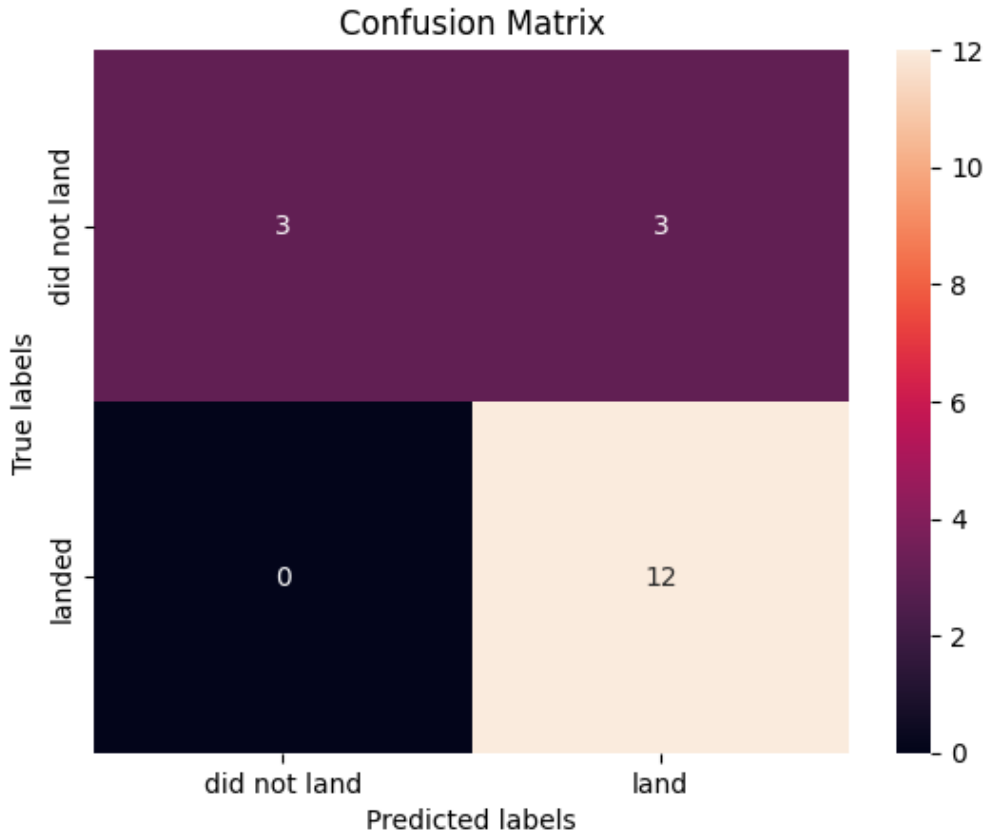
- We see that each model has the same accuracy score of 83.33% on the test data.
- This is likely due to our limited set of test data of only 18 samples.
- Expanding our test set to include the entire data set, we see that the support vector machine model yields the highest accuracy (87.8%).





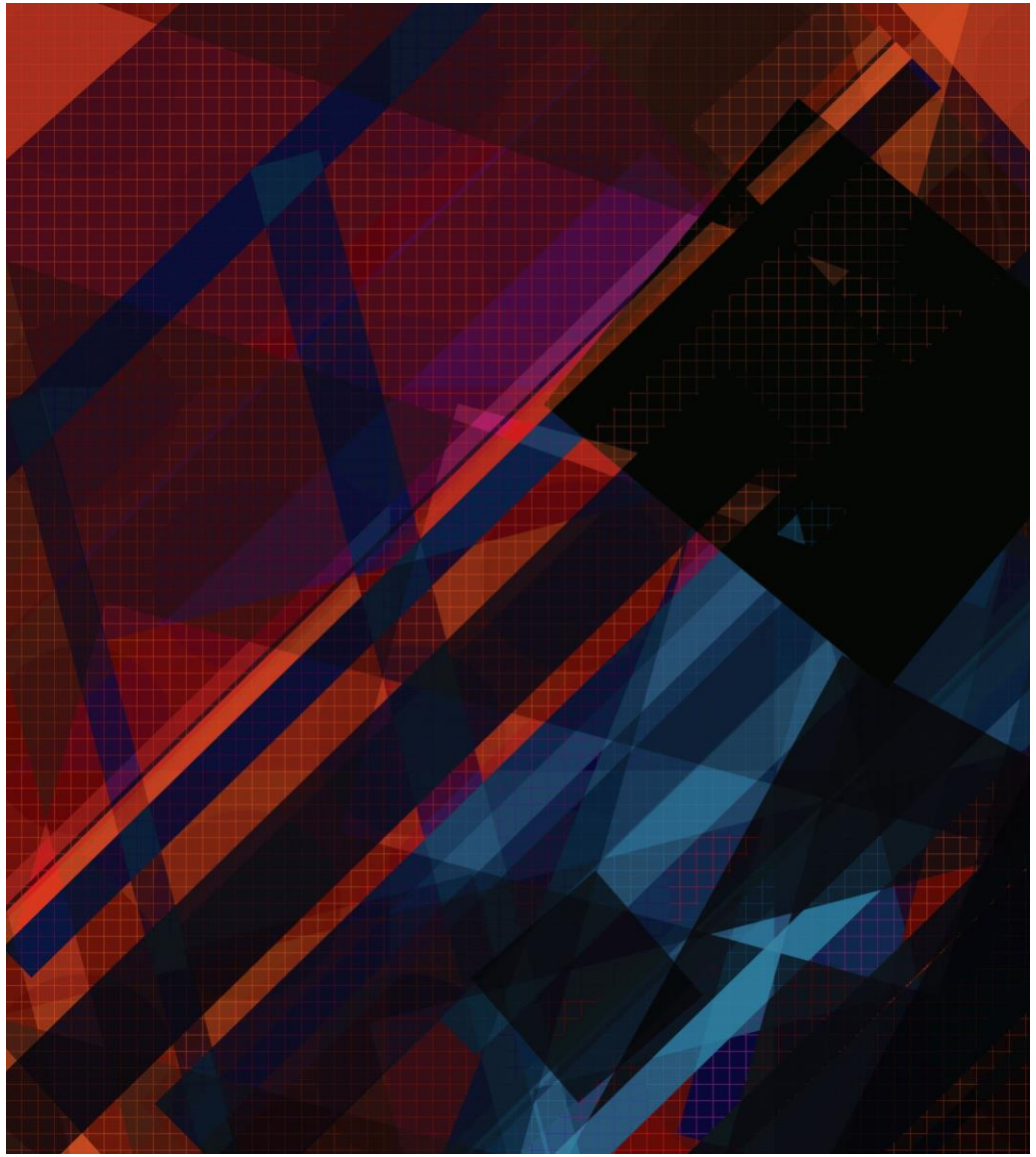
Confusion Matrix

- Each model produced the same confusion matrix on the test data.
- Observe that all failures were successfully predicted.
- Observe the top right panel tells us that the models produce 20% false positives on the test data.



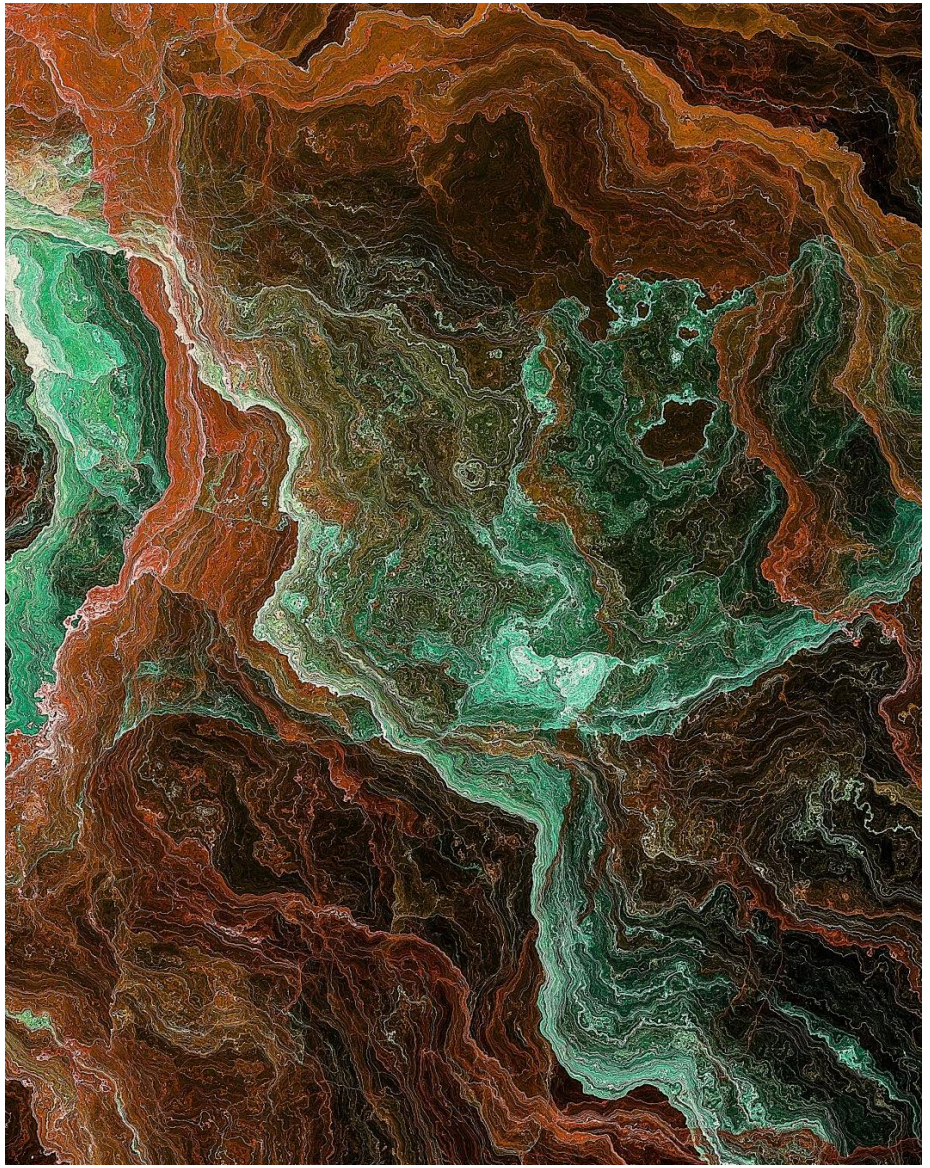
Conclusions

- We found that success rate broadly has increased over time across all launch site.
- ES-L1, GEO, HEO, SSO orbits had 100% success rate.
- The F9 FT series of boosters were most successful for payload in the range of 4000-6000 kg.
- The KSC LC-39A launch site saw the most successful landings and had a success rate of 76.9%.
- We trained four models to predict landing outcomes, each achieving an accuracy of 83.3% on our test data. After expanding the test set to include the whole data set, we found that the support vector machine model had the greatest accuracy, with 87.8%.
- Model could be improved by training on a larger data set. Potentially an issue with 20% false positive predictions.



Appendix

- ❑Many thanks to IBM Coursera instructors!
- ❑Special thanks to Microsoft stock art.
- ❑Github URL



Thank you!