IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vinita Mamarde
13 November 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data was gathered through web scraping techniques and by leveraging the SpaceX API.
  - Conducted thorough Exploratory Data Analysis (EDA), which encompassed data cleaning, visualization, and the use of interactive visual analytics tools.
  - Applied Machine Learning models to predict outcomes based on the analyzed data.
- Summary of all results
  - Successfully acquired valuable datasets from publicly accessible sources.
  - EDA helped pinpoint the most influential features for forecasting launch success.
  - Machine Learning identified the optimal predictive model, highlighting the critical factors that drive successful launches by utilizing the comprehensive dataset.

Introduction

- This analysis aims to determine whether the emerging company Space Y can effectively challenge the established Space X.
- Key questions to address include:
  - ✓How can we most accurately forecast the overall launch expenses by assessing the likelihood of first-stage rocket recovery?
  - ✓What locations offer the optimal conditions for conducting rocket launches?

Section 1

Methodology

vin2winter

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API's and Web scraping techniques
- Perform data wrangling
  - Collected data was curated the landing outcome label based on the outcome data after summarising and analysing features
- Perform exploratory data analysis (EDA) using visualisation and SQL
  - Analyse the data with SQL to calculate the Total payload, payload range for successful launch and total success and failure outcomes

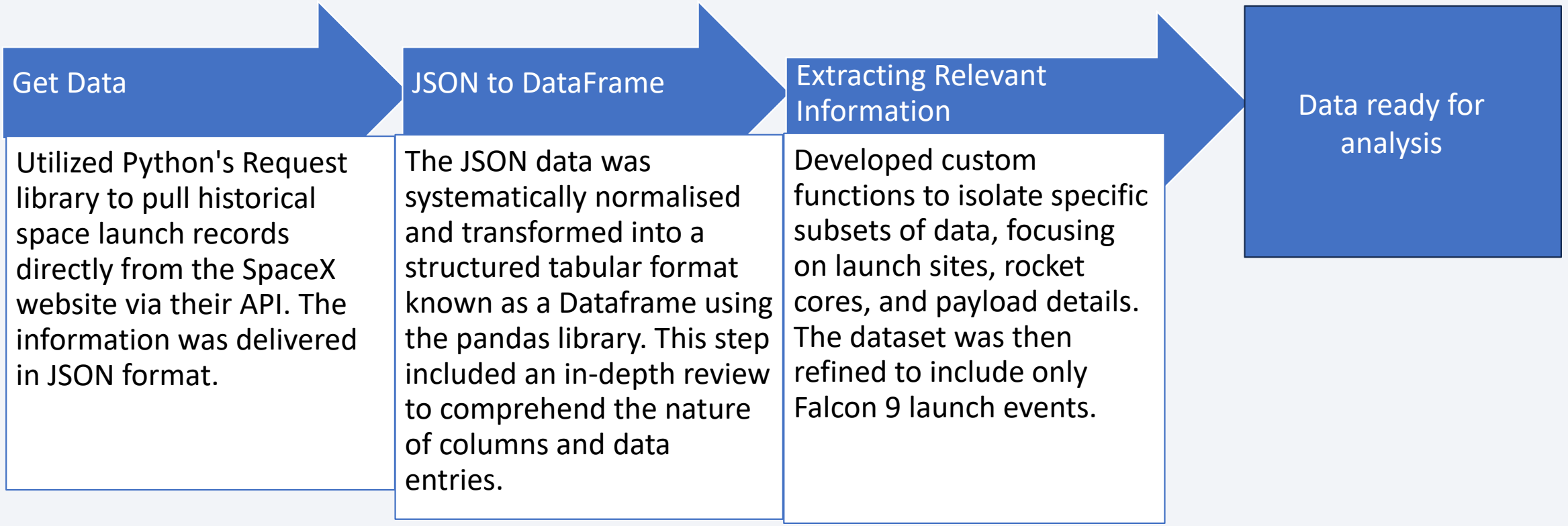
Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
    - Explore the statistics of successful landing across different payloads, over the years.
- Perform predictive analysis using classification models
    - The dataset gathered up to this point was standardized, then split into training and testing subsets. Four distinct classification models were applied, with their accuracy assessed across multiple parameter configurations.

Data Collection

Git Link



Data Collection – SpaceX API

Git Link

From the rocket column we would like to learn the booster name.

https://api.spacexdata.com/v4/rockets/)

From the launchpad we would like to know the name of the launch site being used, the logitude, and the latitude.

((https://api.spacexdata.com/v4/launchpads)

From the payload we would like to learn the mass of the payload and the orbit that it is going to.

(https://api.spacexdata.com/v4/payloads/)

From cores we would like to learn the outcome of the landing, the type of the landing, number of flights with that core, whether gridfins were used, wheter the core is reused, wheter legs were used, the landing pad used, the block of the core which is a number used to seperate version of cores, the number of times this specific core has been reused, and the serial of the core.

(https://api.spacexdata.com/v4/cores/)

```
# Show the head of the dataframe
df_launch.head()
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0

Task 2: Filter the dataframe to only include `Falcon 9` launches

Finally we will remove the Falcon 1 launches keeping only the Falcon 9 launches. Filter the data dataframe using the `BoosterVersion` column to only keep the Falcon 9 launches. Save the filtered data to a new dataframe called `data_falcon9`.

```
# Hint data['BoosterVersion']!='Falcon 1'
# Filter to keep only Falcon 9 launches
data_falcon9 = df_launch[df_launch['BoosterVersion'] != 'Falcon 1']
data_falcon9.head()
```

Python

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0

Data Collection - Scraping

Git Link

To keep the lab tasks consistent, you will be asked to scrape the data from a snapshot of the `List of Falcon 9 and Falcon Heavy launches` Wikipage updated on `9th June 2021`

First, use HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response. Create a `BeautifulSoup` object from the HTML `response`

Extract all column/variable names from the HTML table header. Starting from the third table is our target table, which contains the actual launch records.

Iterate through the `<th>` elements and apply the provided `extract_column_from_header()` to extract column name one by one

Create a data frame by parsing the launch HTML tables. After you have fill in the parsed launch record values into `launch_dict`, you can create a dataframe from it. Export the data to CSV file.

Data Wrangling

Git Link

- Utilized the `.value_counts()` function to tally launches from each site, revealing CCAFS SLC 40 as the top launch location.
- Analyzed the number of launches targeting various orbits, with the Geosynchronous Transfer Orbit (GTO) leading at 27 missions.
- Landing results were categorized into "success" and "failure," then encoded as 1 (success) and 0 (failure) within a "class" column in the launch data dataframe.
- The majority of successful landings occurred on drone ships, achieving 41 successes out of 47 attempts.
- Examined the dataset for missing entries and verified the data types of each feature to ensure integrity and readiness for analysis.

EDA with Data Visualization

Git Link

Categorical Scatter Plot (Seaborn)

- Investigated Feature Interactions with Landing Outcomes
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site

Scatter Plot (Matplotlib)

- Matplotlib scatter plots helped analyze the frequency of successful landings across various orbits, considering their relationship with payload mass and flight number.
  - Very Low Earth Orbit (VLEO) shows the highest success rates, especially with increased flight numbers.
  - Sun-Synchronous Orbit (SSO) exhibits favorable success when payload mass is relatively low.

Line Plot

- The line graph reveals a clear upward trend in launch success over the years, with a remarkable surge post-2013 and 2019

Bar Chart

- Bar charts were utilized to depict the success rates across different orbit categories, Highlighting variations in landing outcomes

EDA with SQL

Git Link

- Counted unique launch sites involved
- Identified launch sites containing the keyword "CCA"
- Computed the total payload mass for NASA (CRS) missions
- Display average payload mass carried by booster version F9 v1.1
- Display the date when the first succesful landing outcome in the ground pad was achieved.
- Successful booster landings on drone ships with payloads ranging between 4000 and 6000 kg
- Tallied total successful and failed landing attempts
- Determined boosters that carried the heaviest payloads
- Analysed monthly landing outcome trends
- Filtered data for launches in 2015
- Ranked landing outcomes occurring between June 4, 2010, and March 20, 2017

Build an Interactive Map with Folium

Git Link

- **Marker**
  - Used to pinpoint the precise locations of launch sites on the map.
- **Circle**
  - Employed to depict the surrounding area around each launch site visually.
- **PolyLine**
  - Drawn to represent distances from the launch sites to the nearest coastline, railway, city, and highway.
- **Marker Cluster**
  - Utilised to group markers that indicate different landing outcomes, enhancing map readability.

Build a Dashboard with Plotly Dash

Git Link

- **Dropdown Menu**
  - Choose a specific launch site to update the visualizations, including the pie chart and scatter plot, based on the selected location.
- **Pie Chart**
  - Displays the ratio of successful versus failed launches for the chosen site, providing a quick overview of outcomes.
- **Range Slider**
  - Allows users to filter the data by selecting a payload mass range, which dynamically adjusts the scatter plot visualization.
- **Scatter Plot**
  - Visualises landing results categorised by launch site and payload mass, helping to identify patterns and correlations.

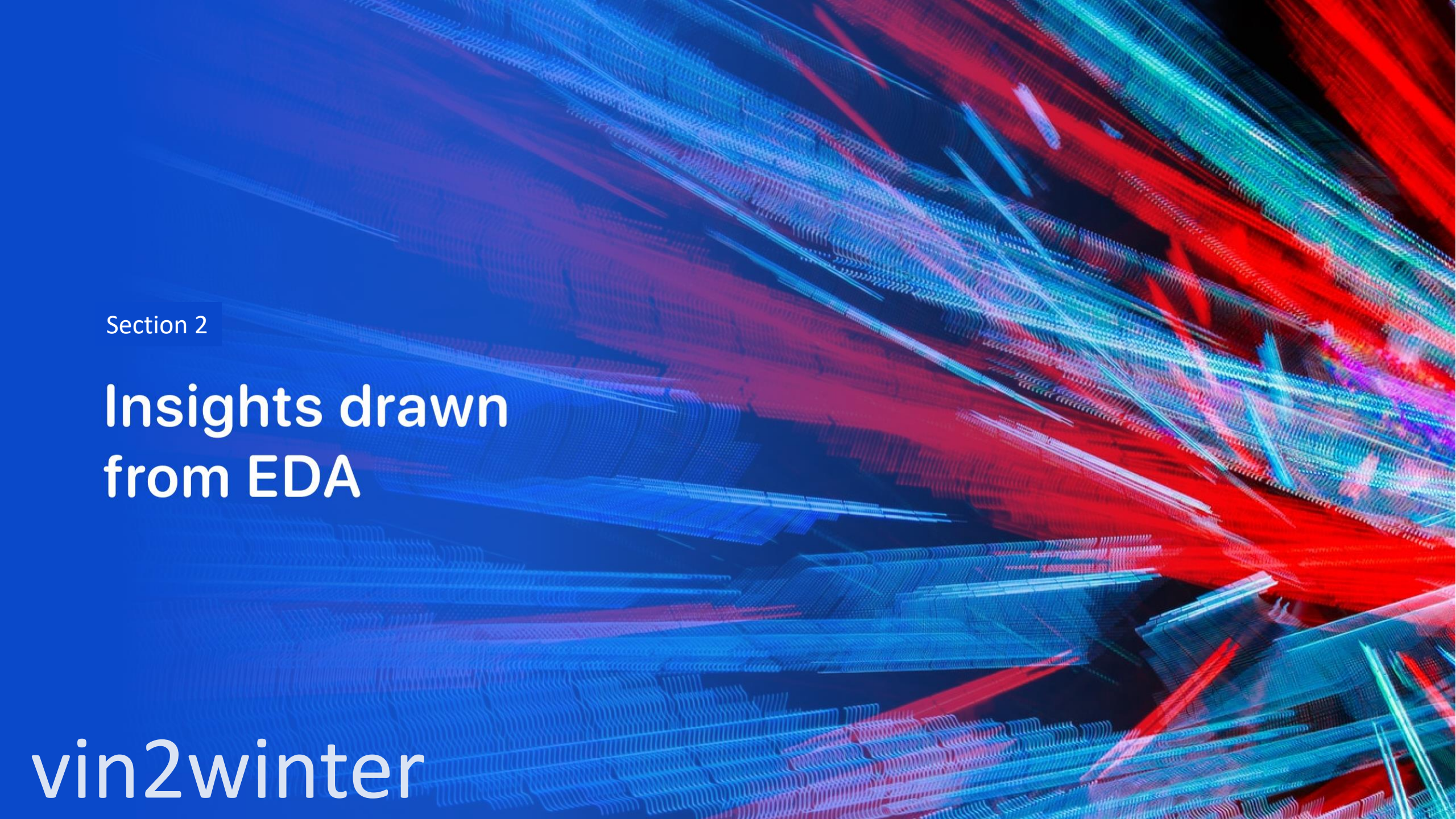
Predictive Analysis (Classification)

Git Link

- Evaluated four distinct classification algorithms: logistic regression, support vector machine (SVM), decision tree, and k-nearest neighbours (k-NN).
- Data was meticulously prepared and standardized to ensure consistency.
- Data was split 80:20 ratio for training and testing, respectively, using the train_test_split() function
- Each model underwent rigorous testing across various hyperparameter configurations.
- Results from all models were systematically compared to identify the best performer.

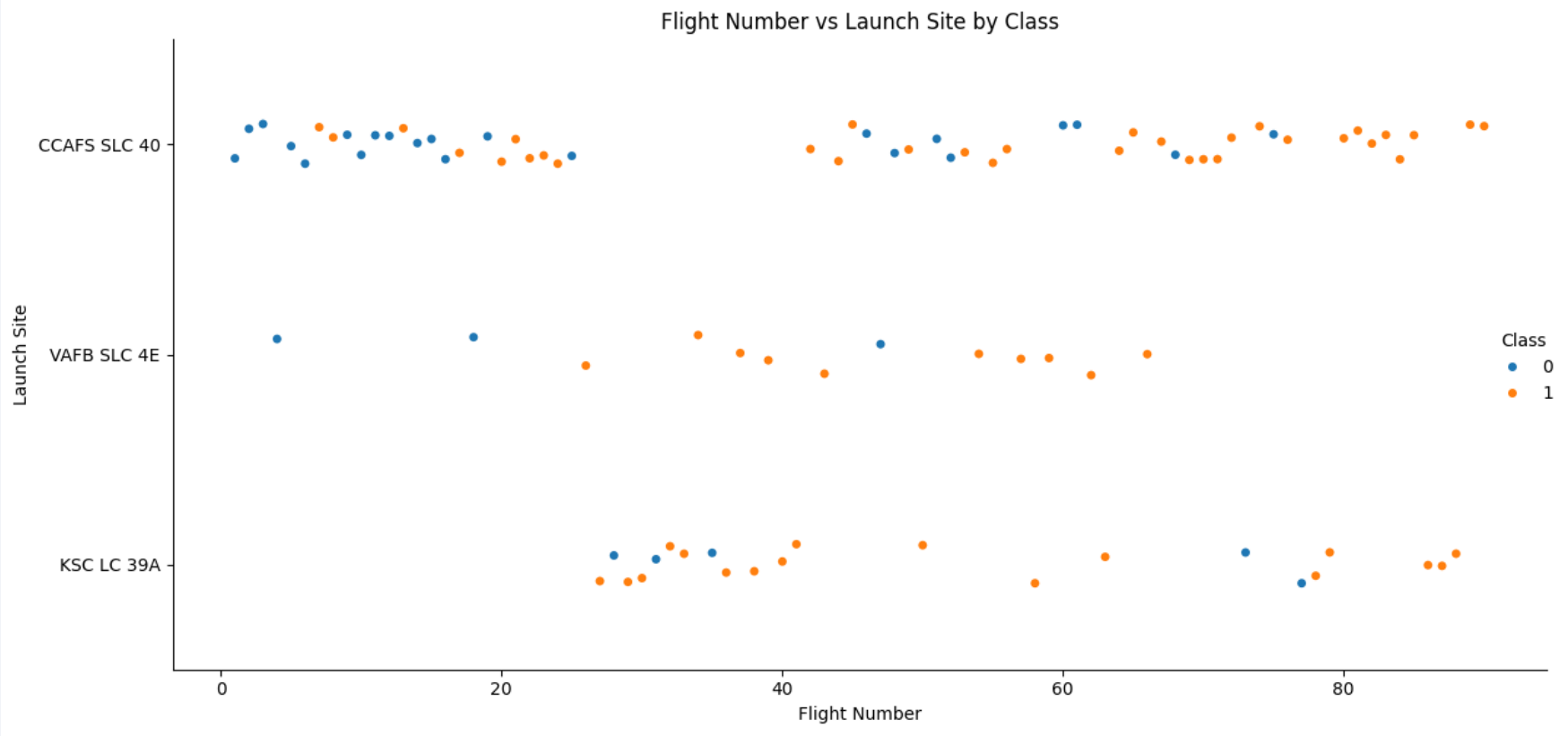
Results

- SpaceX operates launches from four distinct sites
- Initial missions were conducted for SpaceX and NASA
- The Falcon 9 v1.1 booster carries an average payload of 2,928 kg
- The first successful booster landing occurred in 2015, five years after the inaugural launch
- Multiple Falcon 9 booster variants have successfully landed on drone ships, often carrying payloads exceeding the average
- Nearly all missions achieved successful outcomes
- Two Falcon 9 v1.1 boosters, B1012 and B1015, failed to land on drone ships in 2015
- Landing success rates have improved steadily over the years.


Section 2
Insights drawn from EDA
vin2winter

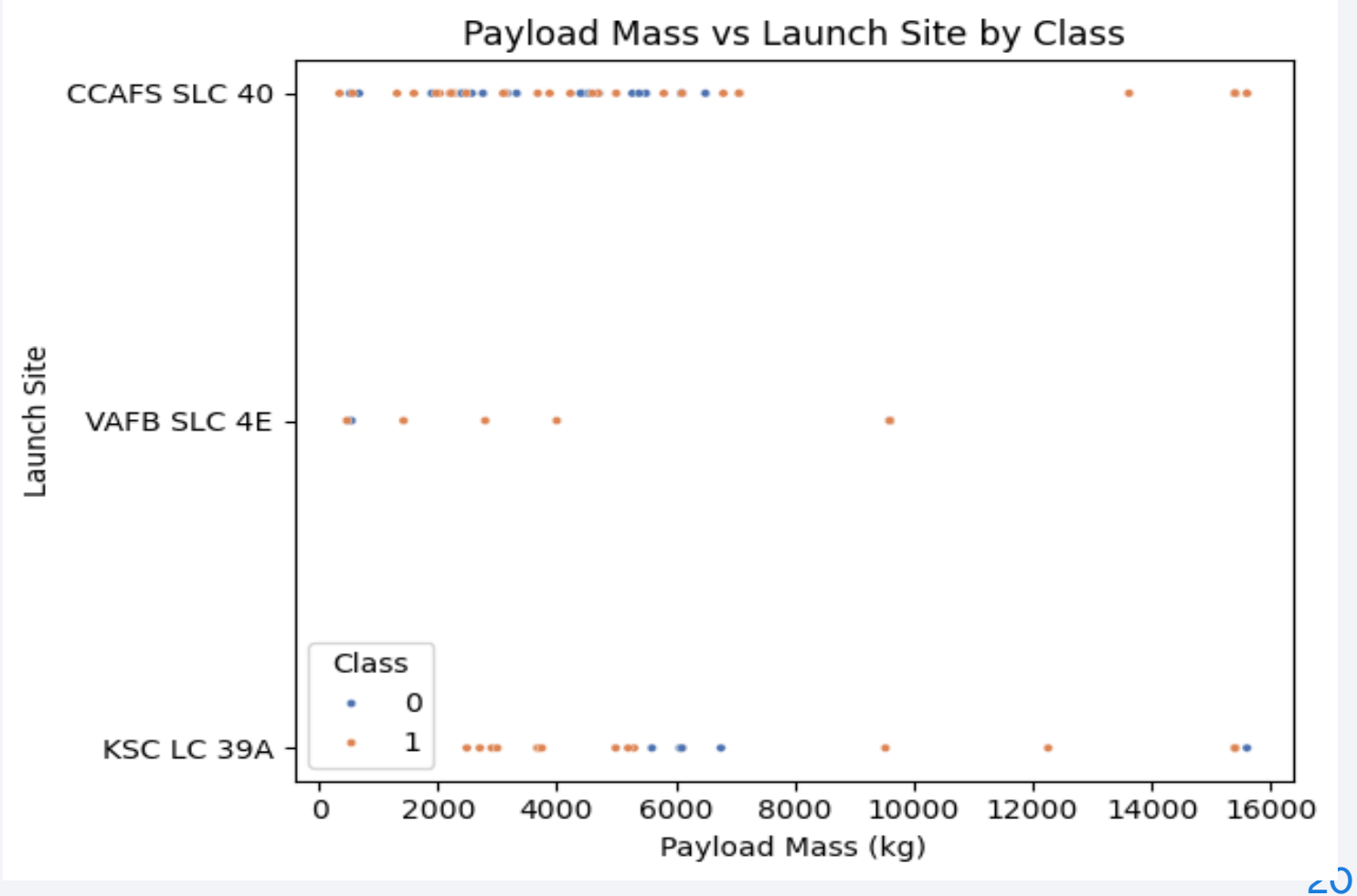
Flight Number vs. Launch Site

- The chart above clearly indicates that CCAF5 SLC 40 currently stands out as the premier launch site, boasting the highest number of successful recent missions.



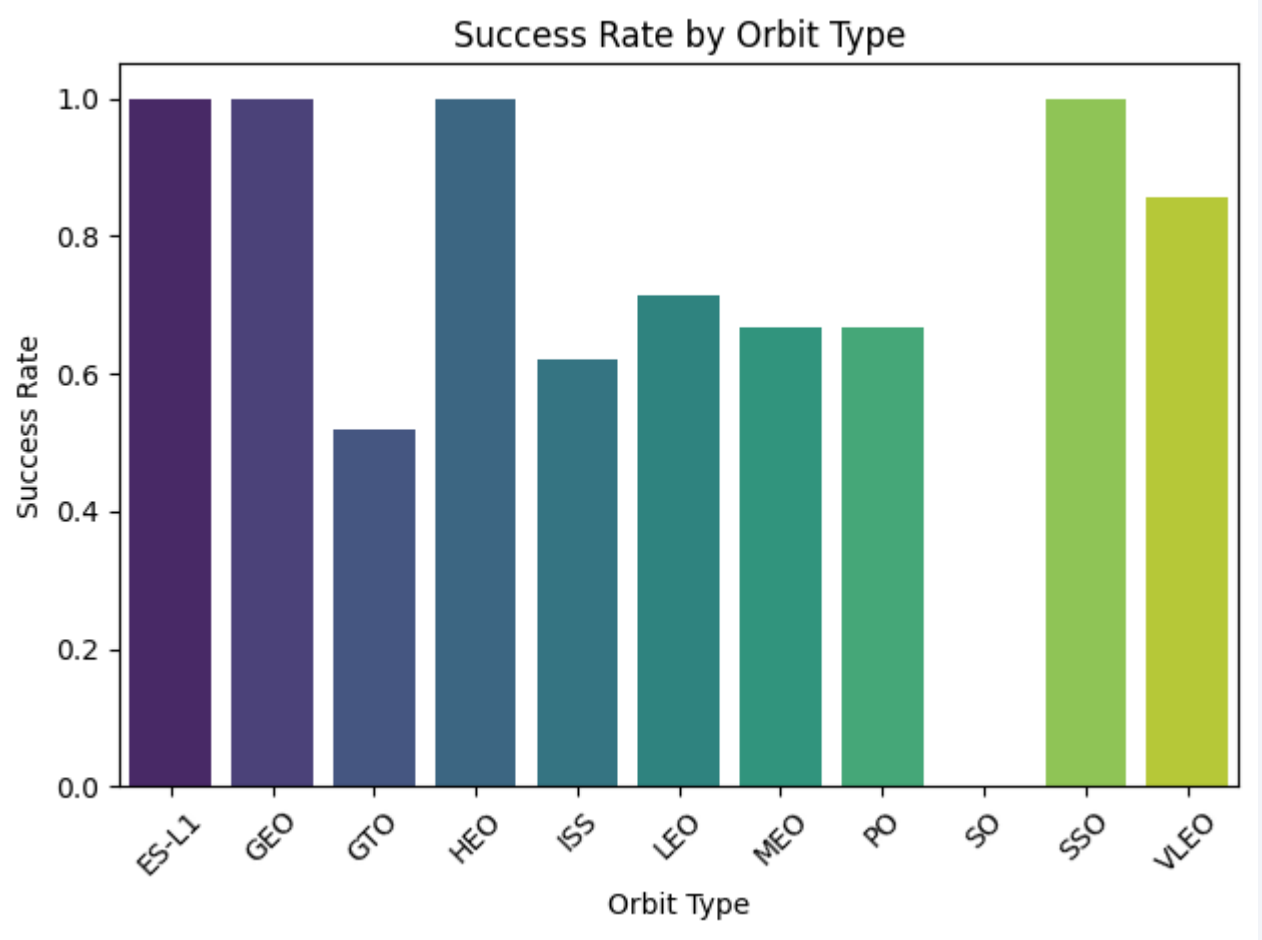
Payload vs. Launch Site

- At Cape Canaveral Space Launch Complex 40 (CCAFS SLC 40), landing outcomes vary significantly until the payload weight surpasses 14,000 kg.
- Vandenberg Air Force Base Space Launch Complex 4E (VAFB SLC 4E) consistently achieves successful landings for payloads exceeding 500 kg.
- Kennedy Space Centre Launch Complex 39A (KSC LC 39A) demonstrates reliable landing success within a payload range of approximately 600 kg to 2,000 kg.



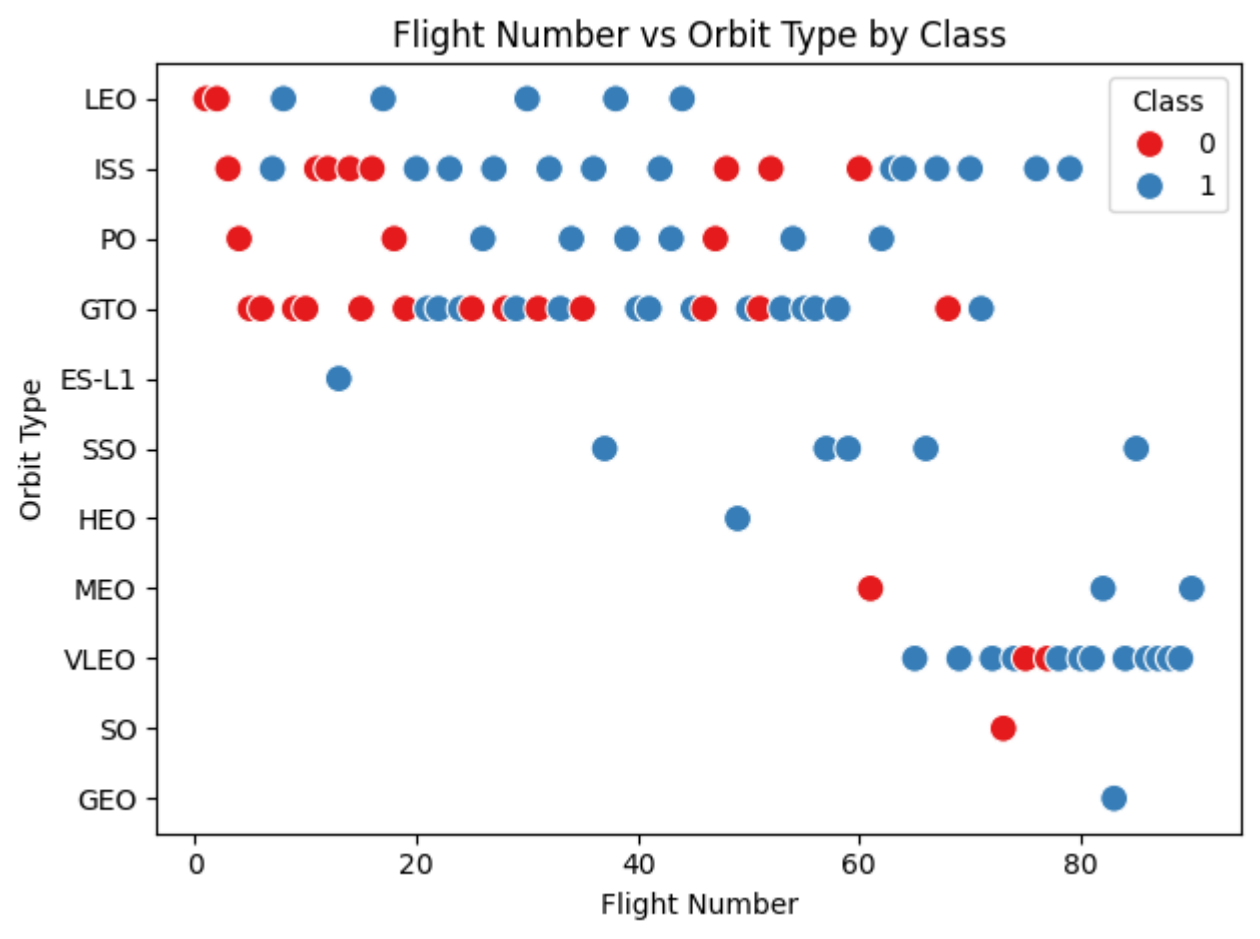
Success Rate vs. Orbit Type

- The success rate is all time high for SSO, HEO, GEO and ES-L1
- There is zero success launch for SO



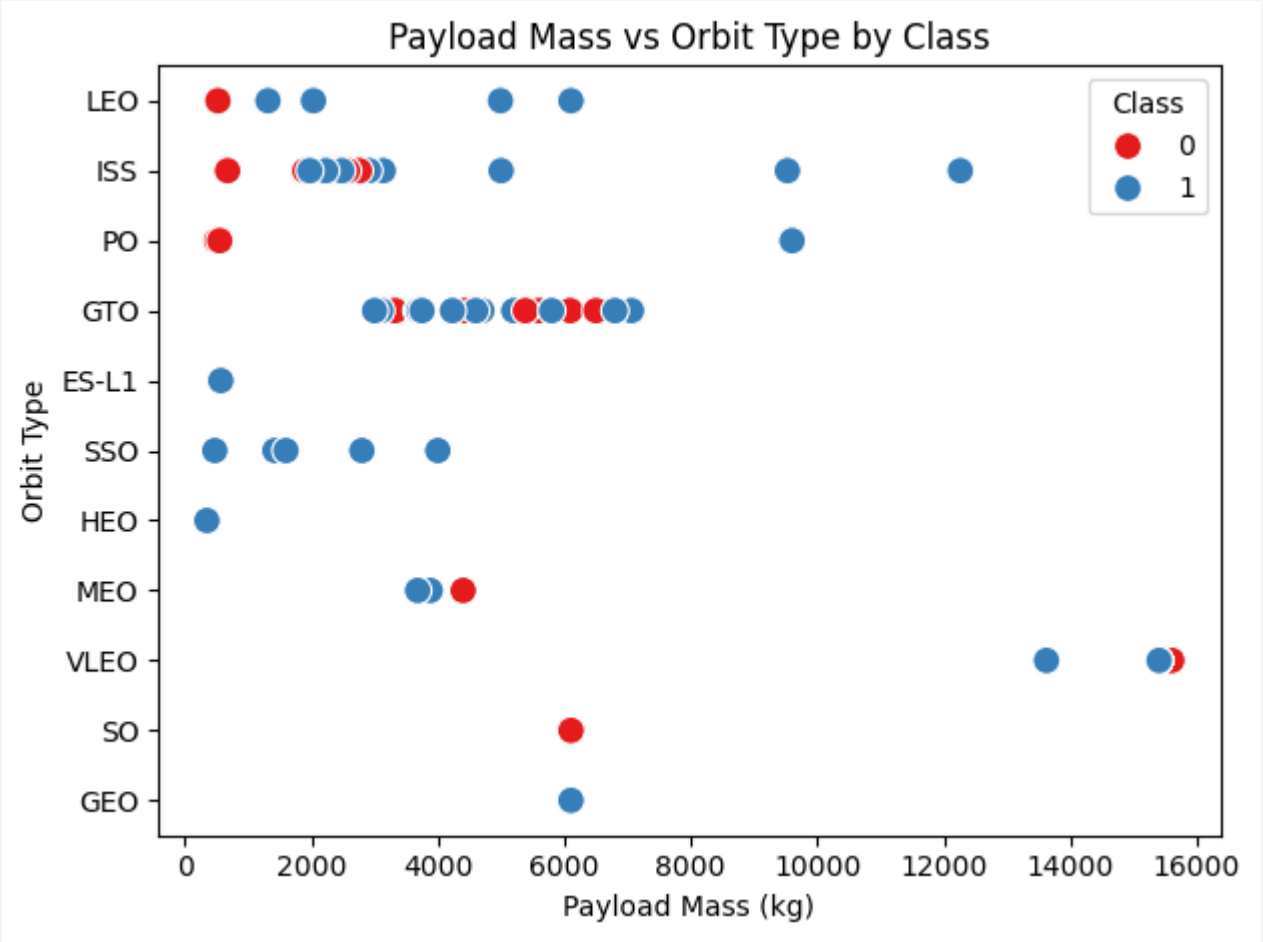
Flight Number vs. Orbit Type

- You can observe that in the LEO orbit, success seems to be related to the number of flights.
- Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



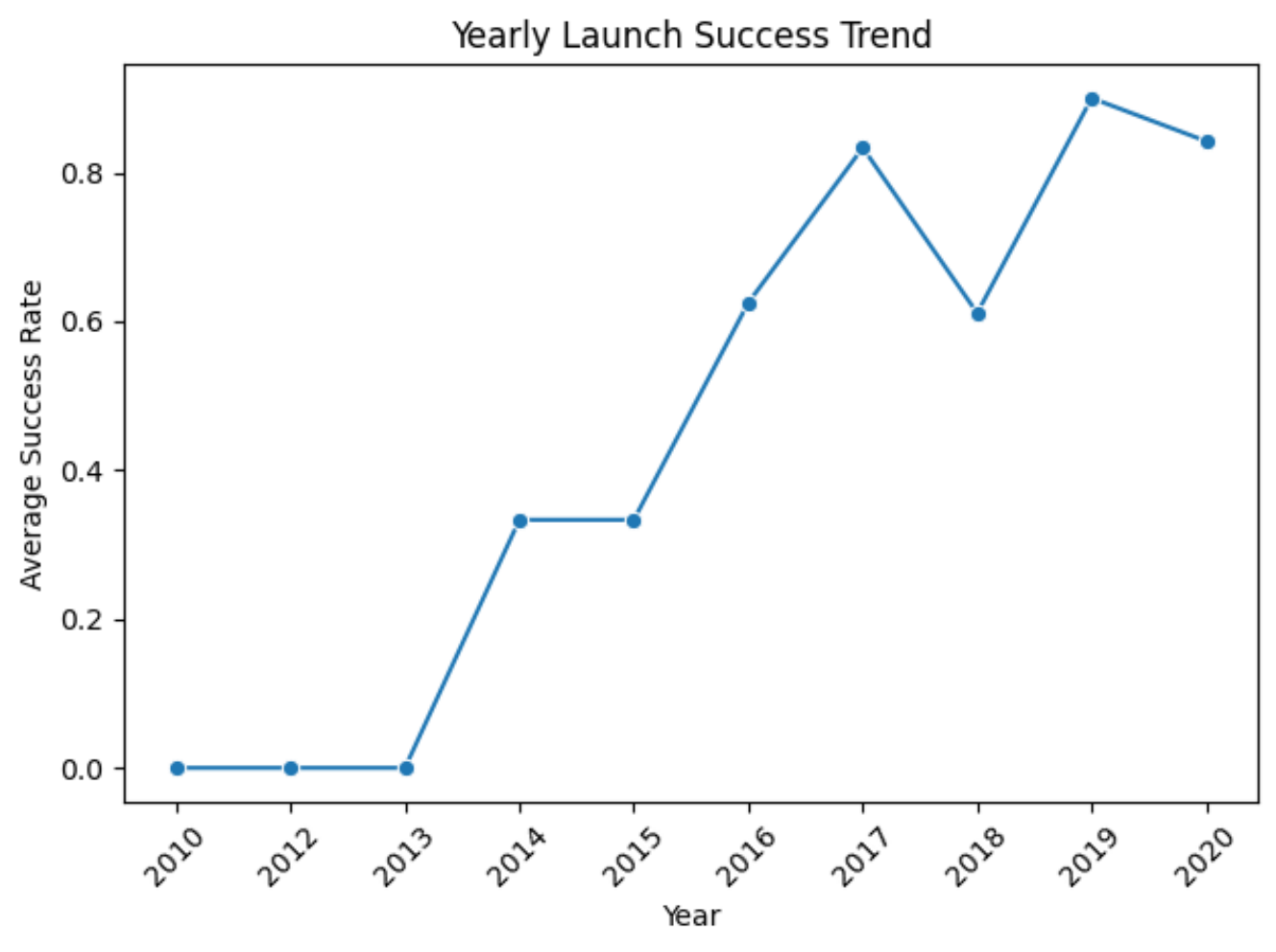
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

- you can observe that the sucess rate since 2013 kept increasing till 2019



All Launch Site Names

- Find the names of the unique launch sites
- Libraries: csv, sqlite3, prettytable
- Use DISTINCT to get unique "Launch Site" names from SPACEXTABLE.

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

```
 * 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'

- Find 5 records where launch sites begin with `CCA

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

Python

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

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

- Calculate the total payload carried by boosters from NASA

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
```

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

Total_Payload
45596

Average Payload Mass by F9 v1.1

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

Display average payload mass carried by booster version F9 v1.1

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

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

Average_Payload
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql SELECT MIN("Date") AS First_Ground_Pad_Landing FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

First_Ground_Pad_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD
```

Python

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

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

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

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

Python

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

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

2015 Launch Records

- List the failed landing_out comes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
SELECT
    substr("Date", 6, 2) AS Month,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE
    "Landing_Outcome" LIKE 'Failure (drone ship)' AND
    substr("Date", 0, 5) = '2015';
```

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

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

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

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

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

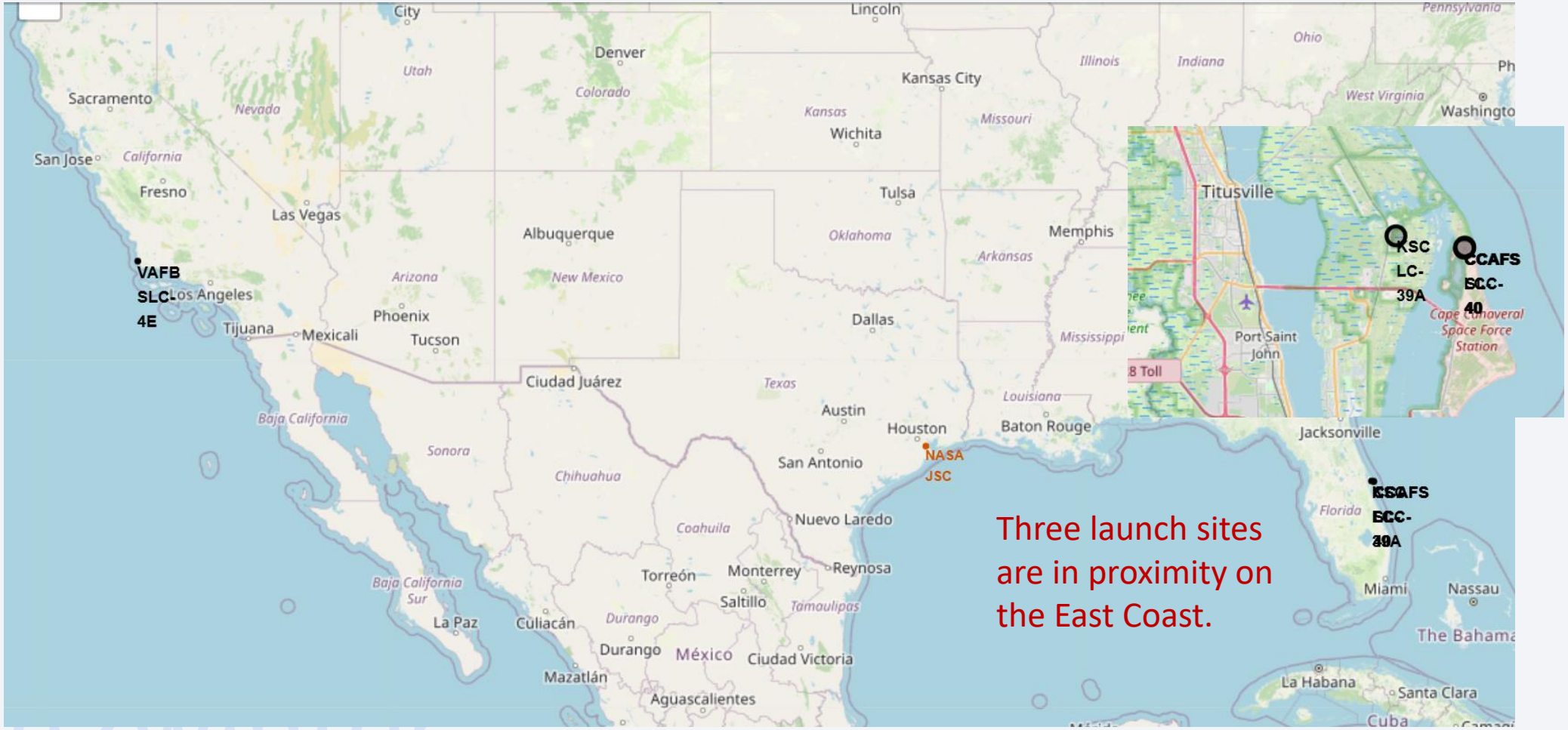
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

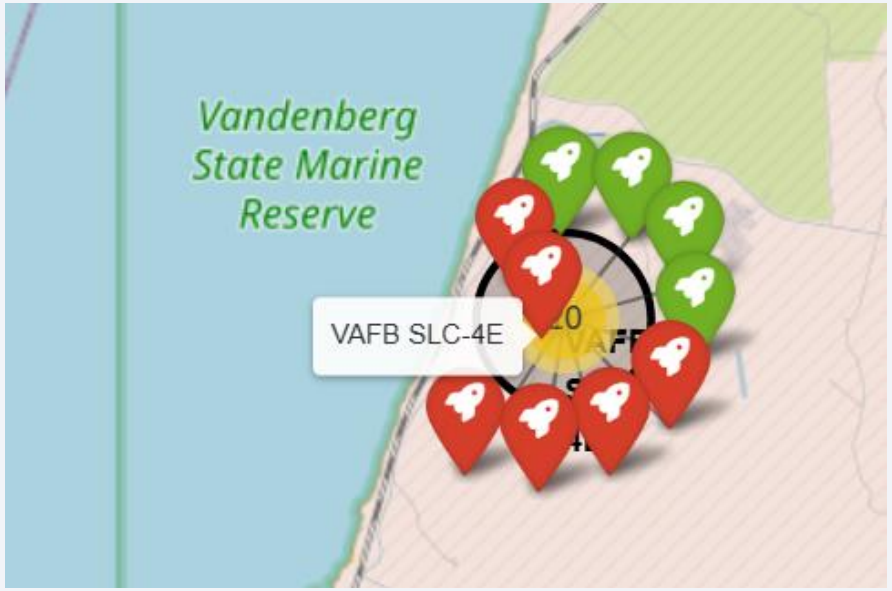
Launch Sites Proximities Analysis

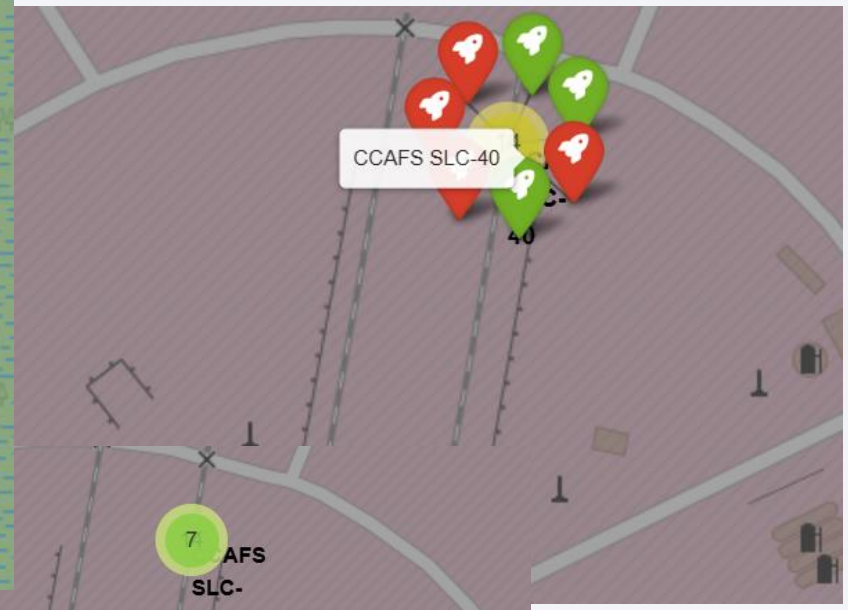
vin2winter

Launch Sites



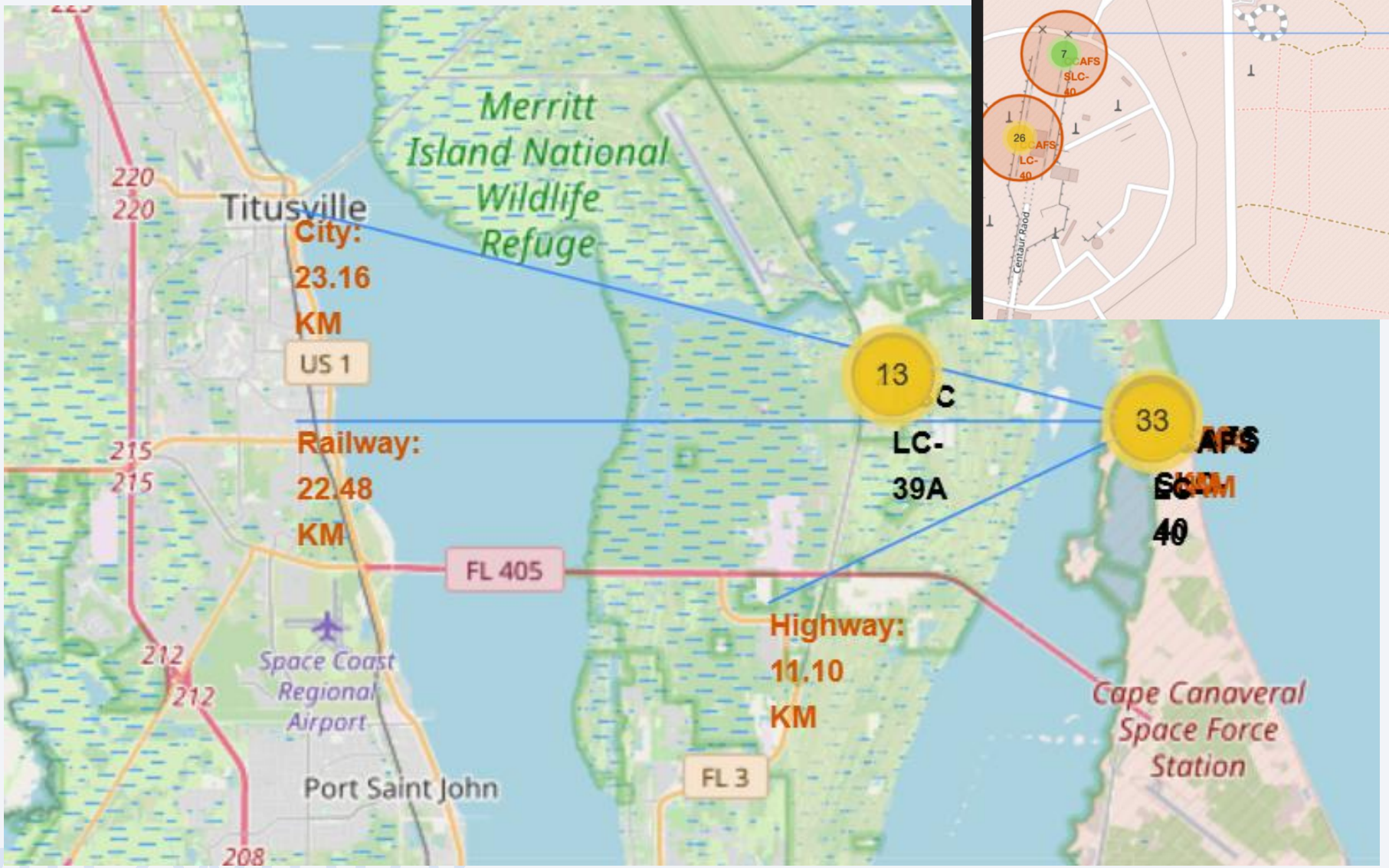
Landing Outcome of Each Location Sites

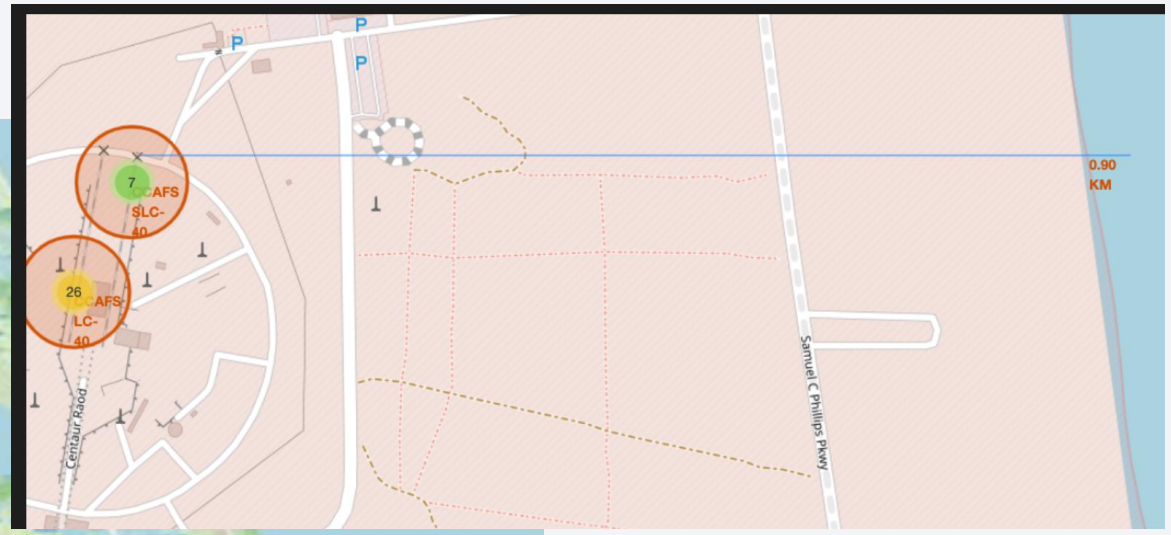




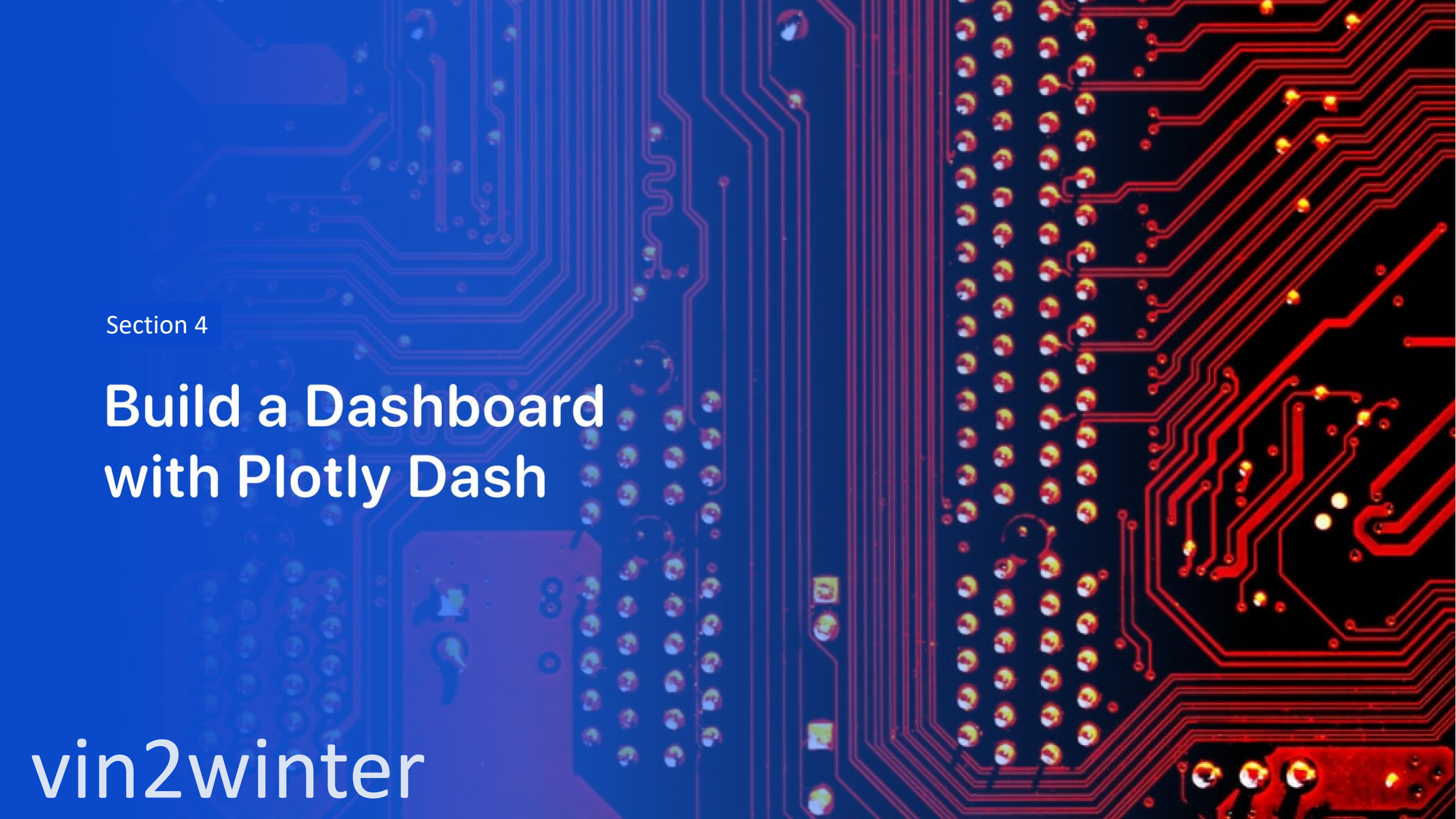
The success is marked with green marker and failure with red marker for better visualization.

Location Characteristics

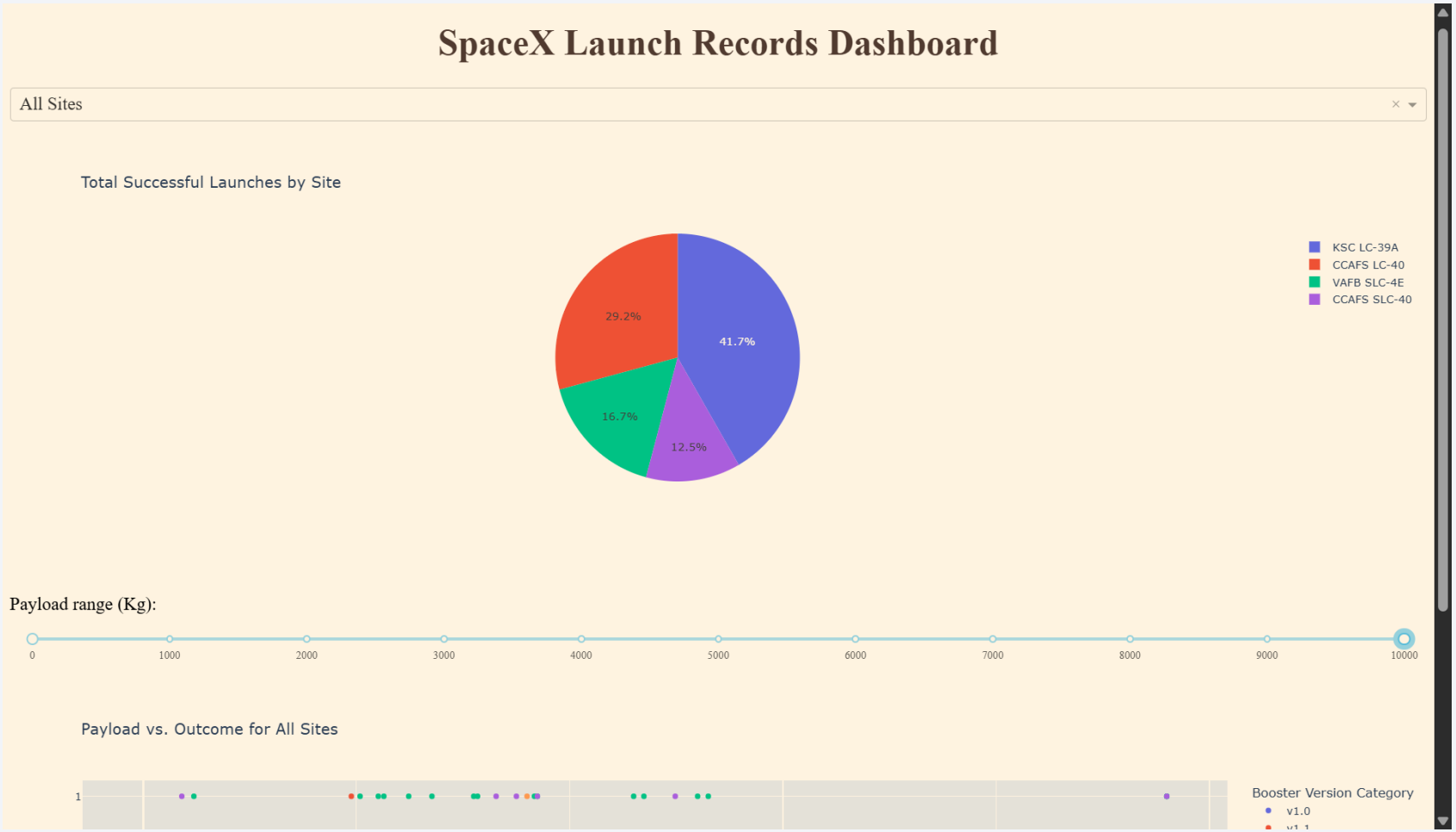




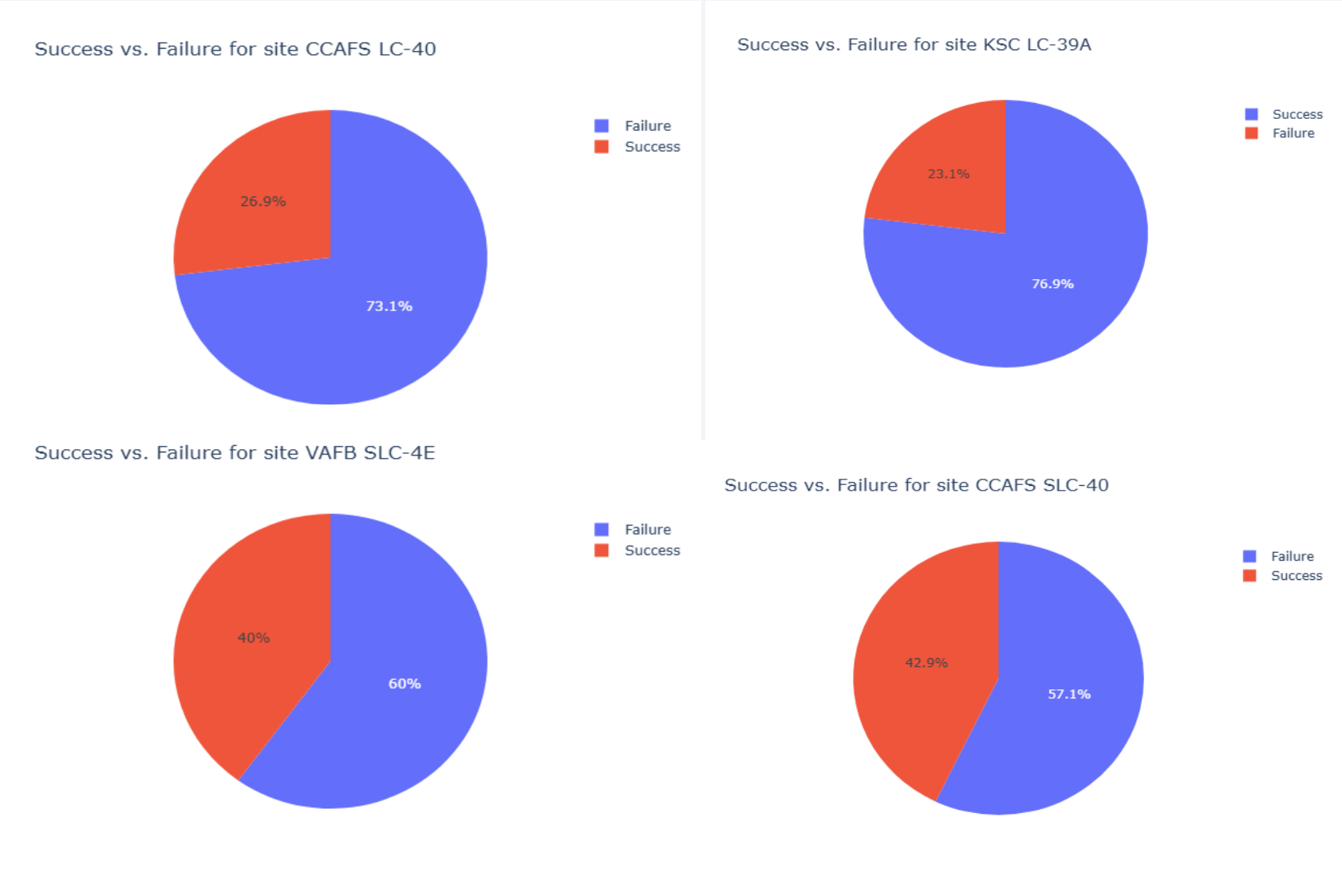
- All locations are close to coastal lines.
- Launch sites are far from the city centre but close to the Highway for logistics


Section 4
Build a Dashboard with Plotly Dash
vin2winter

Dashboard : Total Success rate of all sites



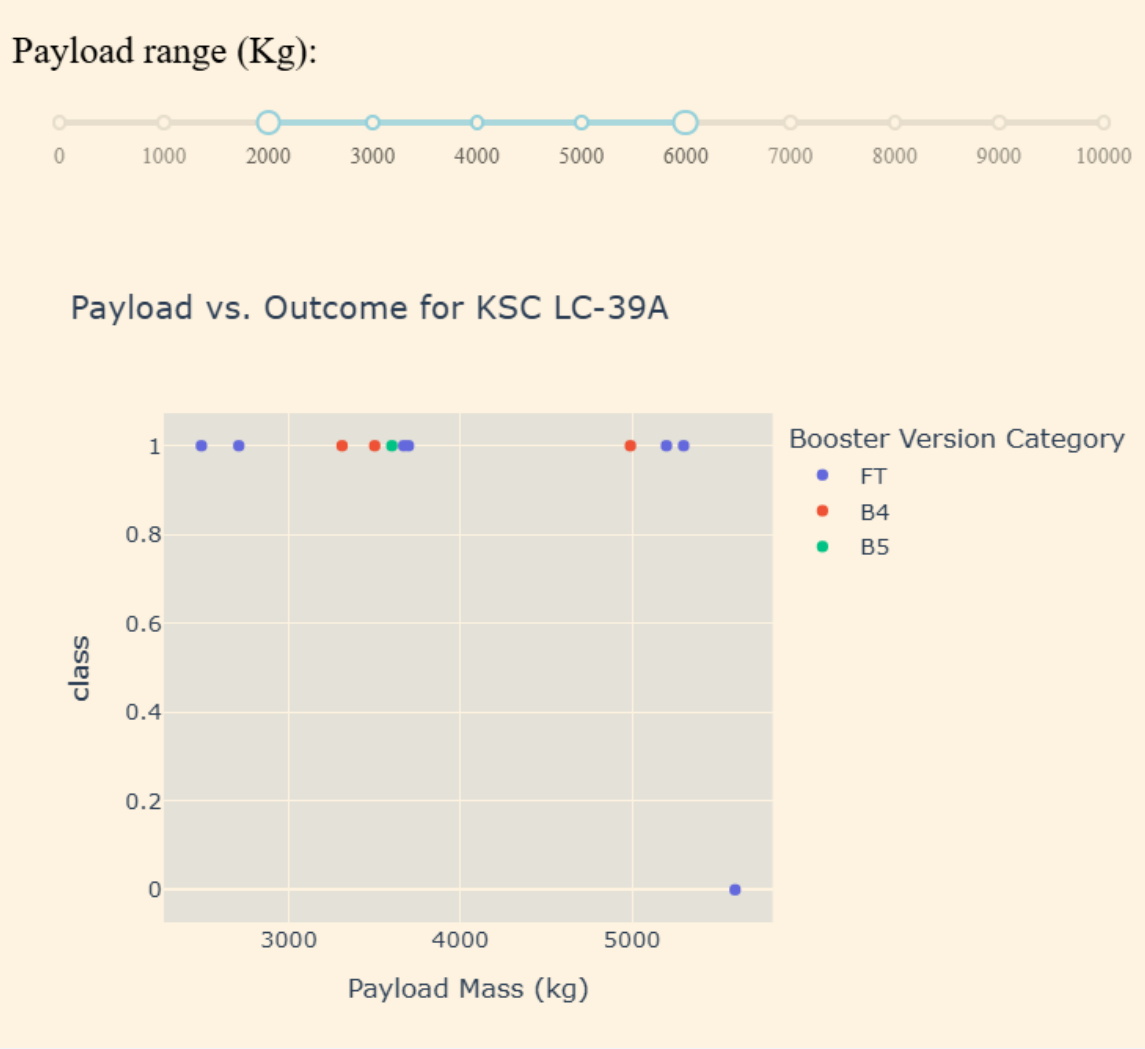
Dashboard : Success rate of all sites



Location KSC LC-39A has shown the highest success rate

Payload range vs outcome success for KSC LC-39

Missions deploying payloads ranging from 2,000 to 6,000 kilograms tend to achieve higher success rates.
Success rate increases especially for FT Type Booster for the 2,000 to 6,000 payload range.



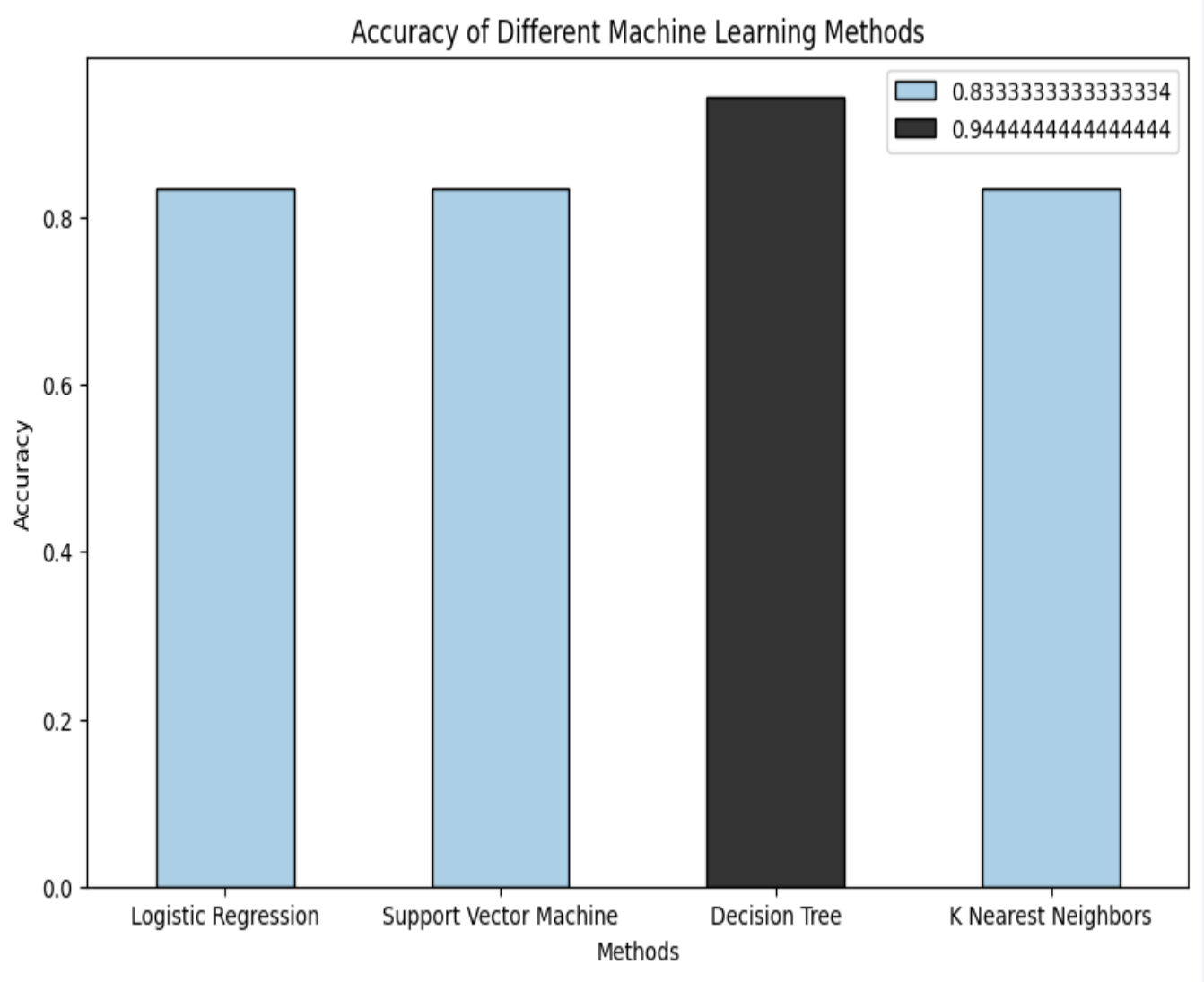
Section 5

Predictive Analysis (Classification)

vin2winter

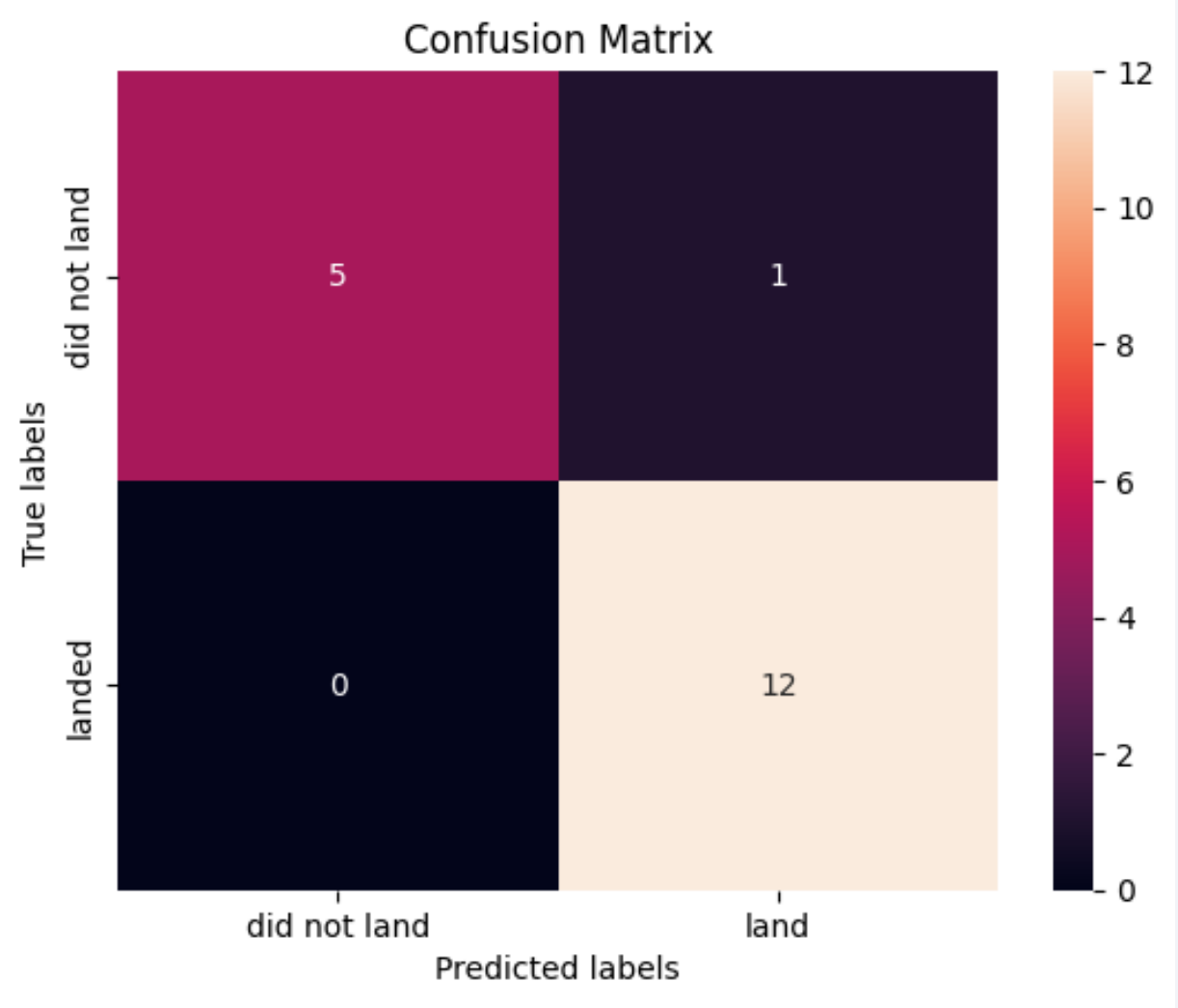
Classification Accuracy

- We had a very small data set to train and test all Machine Learning models.
- This shows the same accuracy for all models except for the Decision Tree Model.
- Decision Tree Shown accuracy of 94%



Confusion Matrix

- Confusion matrix indicate only one false-negative.
- It has shown the highest accuracy in prediction compared to other models



Conclusions I

Summary of Findings

- **Launch Site Effectiveness:** The "KSC LC-39A" site stands out with the highest success rate, playing a crucial role in the overall achievement of successful landings.
- **Booster Model Performance:** Among the boosters, FT and B4 versions led in success, whereas the F9 V1.1 model lagged, underscoring how booster design critically impacts mission success.
- **Impact of Payload Weight:** Missions carrying payloads between 2,000 kg and 6,000 kg consistently showed superior success rates, indicating this range as optimal for dependable launches.
- **Annual Success Patterns:** Since 2013, successful landings have risen markedly, with significant surges observed in 2017 and 2019.
- **Predictive Modeling Accuracy:** Logistic Regression proved to be the most reliable method for forecasting landing results, boasting a 94.44% accuracy rate and minimal errors, as reflected in the confusion matrix.

Conclusions II

If we want to compete with SpaceX with SpaceY point to take notice

- To build successful launch sites, we need to be selective with our location, which is close to the coastline but away from human centres of activity and highways closed for logistics.
- We can focus on the booster that gives more success to the failure ratio to attract investors.
- We also have a prediction model to track the SpaceX development curve, which indicates a promising future in space.
- It also indicates that we have well passed the primary hurdle in technology testing, which will further decrease the investment needed by SpaceY to compete with SpaceX

Appendix I

GitHub: https://github.com/vinwin10/Data_Science_Capstone_Project.git
Data: https://api.spacexdata.com/v4/launches/past Python
Libraries:

- requests: fetch SpaceX API data
- pandas: data cleaning and analysis
- matplotlib: static plots
- seaborn: advanced visualizations
- folium: interactive maps
- plotly: interactive dashboards
- sqlite3: SQL queries and local database
- numpy: numerical operations

Appendix II

Python Libraries and Their Uses:

- sklearn:
  - SVC: Support Vector Classification
  - DecisionTreeClassifier: Tree models
  - KNeighborsClassifier: K-Nearest Neighbors
  - train_test_split: Split data into train/test
  - GridSearchCV: Tune hyperparameters
  - preprocessing: Standardize/transform data
- Other Tools:
  - prettytable: Display query results in tables
  - csv: Manage CSV files for saving/exporting

Appendix III

SQL statements summary:

- SELECT DISTINCT: Got unique launch site names.
- LIKE and LIMIT: Filtered sites starting with "CCA", limited to 5.
- SUM(): Total payload mass for NASA (CRS).
- AVG(): Average payload for booster F9 V1.1.
- MAX(): Max payload and related booster versions.
- MIN(): Date of first successful ground landing.
- BETWEEN: Filtered landing outcomes by date.
- CASE and SUM(): Counted successful and failed missions.
- substr(): Extracted month/year to filter 2015 launches.
- GROUP BY and ORDER BY: Ranked landing outcomes by count in date range.

Thank you!

vin2winter