IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yanran Chen
Sep 2nd, 2025

Outline

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

© 2025 Yanran Chen. All rights reserved.

Executive Summary

This project aims to predict the landing success of SpaceX's Falcon 9 first stage by applying supervised machine learning techniques to historical launch data to identify patterns and key predictors of successful landings.

The dataset was collected from SpaceX REST API, including features such as launch site, payload mass, orbit type, booster version, and flight number, among others. After data cleaning, exploratory data analysis and visualization, four classification models were developed and evaluated: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Logistic Regression. Each model was trained using cross-validation, with hyperparameters tuned for optimal performance. Evaluation metrics included accuracy, precision, recall, and F1-score.

All four models achieved the same accuracy of 83.3% on testing set, demonstrated strong generalization on unseen data. These results suggest that machine learning can be a powerful tool for forecasting rocket landing success, potentially informing future launch strategies and significantly reducing launch costs.

© 2025 Yanran Chen. All rights reserved.

Introduction

SpaceX has revolutionized space transportation by pioneering reusable rocket technology, with the Falcon 9 first stage playing a central role in this innovation. Successful landings of the Falcon 9 booster are essential to reducing launch costs and increasing mission frequency. According to SpaceX, Falcon 9 rocket launches on its website with a cost of $62 million, whereas other providers cost upward of $165 million each. This dramatic cost difference underscores the financial importance of booster recovery. If we can predict whether the first stage will successfully land, we can effectively estimate the cost of a launch and assess mission risk. This project leverages publicly available launch data from the SpaceX REST API to build predictive models that classify whether a Falcon 9 first stage landing will succeed or fail.

The primary objective of this project is to answer the following questions:

- **Can machine learning models accurately predict the success of Falcon 9 first stage landings based on pre-launch and flight data?**
- **Which features are most influential in determining landing outcomes?**
- **How do different classification algorithms – K-Nearest Neighbors, Support Vector Machines, Decision Trees, and Logistic Regression – compare in terms of predictive performance?**
- **Can these models generalize well to unseen launches, and what are their limitations?**

© 2025 Yanran Chen. All rights reserved.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Dataset was collected from Space X Rest API and web scraping Wikipedia pages
- Perform data wrangling
  - Identified missing values and imputed missing Payload Mass with its mean value.
  - Convert landing outcomes into binary classes: Success / Failure
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was split into training and testing set
  - Cross validation was used to tune hyperparameters for optimal performance.
  - Evaluation metrics included accuracy, precision, recall, and F1-score.

© 2025 Yanran Chen. All rights reserved.

Data Collection

Two data collection methods were employed:

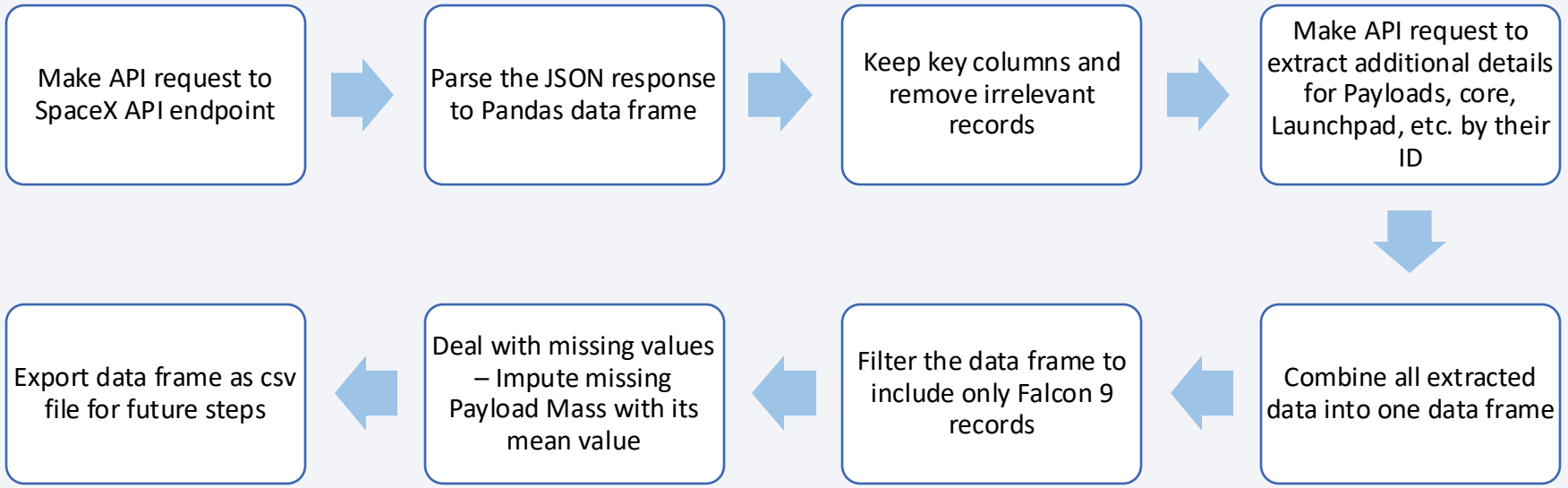
1. Launch records were scraped from the SpaceX Wikipedia page by making an HTTP request, parsing the HTML with BeautifulSoup, extracting table headers and row data, and compiling the results into a structured pandas DataFrame.

2. A more robust approach utilized the official SpaceX REST API to retrieve launch data in JSON format. Additional API calls were made to gather detailed information on payloads, cores, and launchpads using their unique IDs in the master JSON

Final datasets from both methods were exported as a CSV files for downstream analysis and model development.

© 2025 Yanran Chen. All rights reserved.

Data Collection – SpaceX API



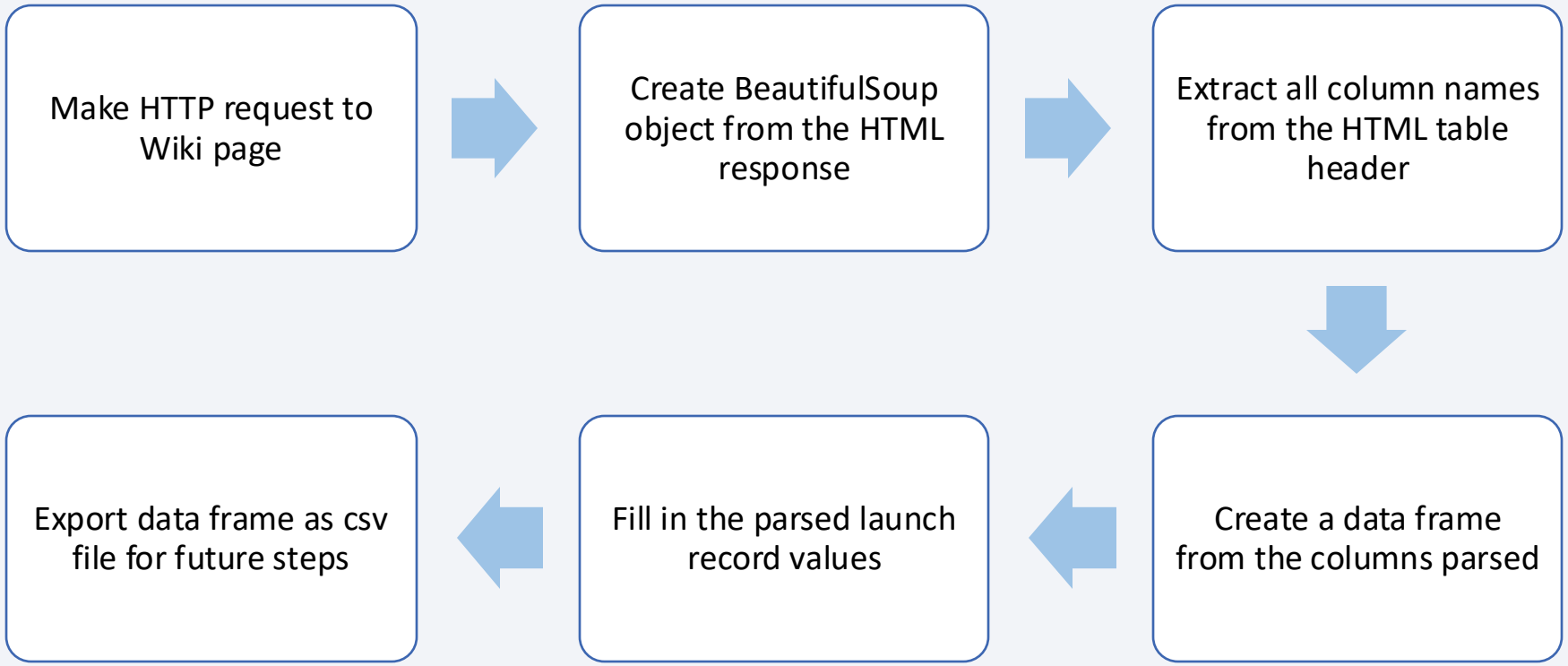
What the dataset looks like after completing this step

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

See complete notebook on GitHub:
https://github.com/YanranChen/IBM-Data-Science/blob/main/C10_Applied_Data_Science_Capstone/M1_Intro/M1Lab1_spacex-data-collection-api.ipynb

© 2025 Yanran Chen. All rights reserved.

Data Collection - Scraping



See complete notebook on GitHub:
https://github.com/YanranChen/IBM-Data-Science/blob/main/C10_Applied_Data_Science_Capstone/M1_Intro/M1Lab2_webscraping.ipynb

© 2025 Yanran Chen. All rights reserved.

Data Wrangling

Load the data set prepared in the Data Collection step → Identify numerical and categorical columns → Calculate the number of occurrence of each value for key categorical columns → Group different Outcomes to create binary Success/Failure label → Export data frame as csv file for future steps

What the dataset looks like after data wrangling

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

See complete notebook on GitHub:
https://github.com/YanranChen/IBM-Data-Science/blob/main/C10_Applied_Data_Science_Capstone/M1_Intro/M1Lab3_spacex-data-wrangling.ipynb

© 2025 Yanran Chen. All rights reserved.

EDA with Data Visualization

To uncover patterns and relationships within the SpaceX launch dataset, 6 targeted visualizations were created to address, each addressing a specific analytical question and reveal insights about launch outcomes, payloads, and orbital strategies.

1. **Flight Number vs. Launch Site:** A scatter plot was used to examine how launch sites varied across sequential launch attempts (represented by flight number and how this progression correlated with success rates. This helped highlight the technological maturity and reliability trends over time at each location.
2. **Payload Mass vs. Launch Site:** A scatter plot explored how payload weight distributions differed by site and whether certain locations were associated with heavier or lighter missions. Overlaying launch outcomes provided insight into how mass and site jointly influenced success.
3. **Success Rate by Orbit Type:** A bar chart was used to compare the success rates across different orbital destinations (e.g., LEO, GTO, SSO). This visualization helped identify which orbit types were more reliably reached and which posed greater challenges.
4. **Flight Number vs. Orbit Type:** A scatter plot illustrated how the choice of orbit evolved over time, and how it affected success rate.
5. **Payload Mass vs. Orbit Type:** A scatter plot examined how payload weight varied across orbit types, helping to assess whether heavier payloads were more likely to be sent to specific orbits and how that impacted success.
6. **Yearly Success Rate Trend:** A line plot showed how launch success rates changed year by year, providing a clear view of SpaceX's operational improvements and reliability growth over time.

Complete notebook for EDA can be found on GitHub:
https://github.com/YanranChen/IBM-Data-Science/blob/main/C10_Applied_Data_Science_Capstone/M2_EDA/M2Lab2_edadataviz.ipynb

© 2025 Yanran Chen. All rights reserved.

EDA with SQL

Additional SQL queries were performed in exploration of the below information:

- Names of the unique launch sites in the space mission
- The total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000
- Total number of successful and failure mission outcomes: 61 Success vs 10 Failure
- Count of landing outcomes between the date 2010-06-04 and 2017-03-20

Complete notebook for EDA with SQL can be found on GitHub:
https://github.com/YanranChen/IBM-Data-Science/blob/main/C10_Applied_Data_Science_Capstone/M2_EDA/M2Lab1_eda_sqllite.ipynb

© 2025 Yanran Chen. All rights reserved.

Build an Interactive Map with Folium

Using Python's Folium library, we created an interactive map to visualize the geographic distribution and performance of SpaceX launch activities. The map includes the following elements:

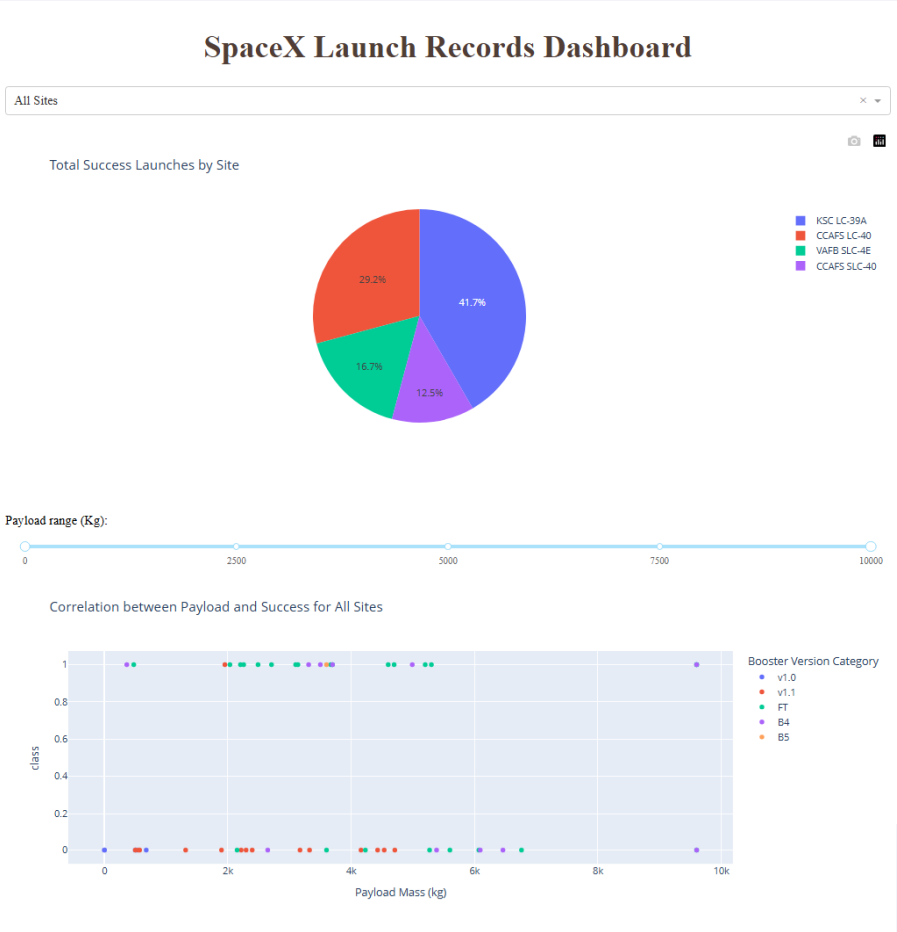
- **Initial Center Location:** The map is centered on NASA's Johnson Space Center in Houston, Texas, serving as the starting viewpoint for users.
- **Launch Site Markers:** Each of the 4 SpaceX launch sites is marked with a labeled icon and surrounded by a colored range circle, providing spatial context and allow users to identify key locations at a glance.
- **Launch Outcome Markers:** Individual launch outcomes are represented by colored circles placed at the corresponding launch site – green for successful landings and red for failed attemps. This color-coded system offers an intuitive way to assess performance across sites.
- **Proximity Analysis:** For illustrative purposes, the CCAFS SLC-40 launch site includes straight-line markers connecting it to nearby features such as the coastline, railway, highway, and city. Each line is labeled with its calculated distance, offering insight into logistical and environmental factors.

This visualization offers an intuitive and interactive overview of SpaceX launch activities, highlighting geographic locations and launch outcomes at each site. By combining spatial markers, outcome indicators, and proximity analysis, the map enables users to explore patterns in launch success and site utilization with ease.

Complete notebook for the Interactive Map can be found on GitHub: https://github.com/YanranChen/IBM-Data-Science/blob/main/C10_Applied_Data_Science_Capstone/M3_Interactive_Analytics_and_Dashboards/M3Lab1_launch_site_location.ipynb

© 2025 Yanran Chen. All rights reserved.

Build a Dashboard with Plotly Dash



To enable real-time interactive exploration of SpaceX launch data, a Plotly Dash application was developed to enable visual analytics through dynamic charts and filters.

This dashboard application contains two input components:

1. **Launch Site Selector**: A dropdown menu to choose a specific launch site or view data across all sites
2. **Payload Mass Filter**: A range slider to narrow results based on payload weight.

The dashboard features two interactive charts :

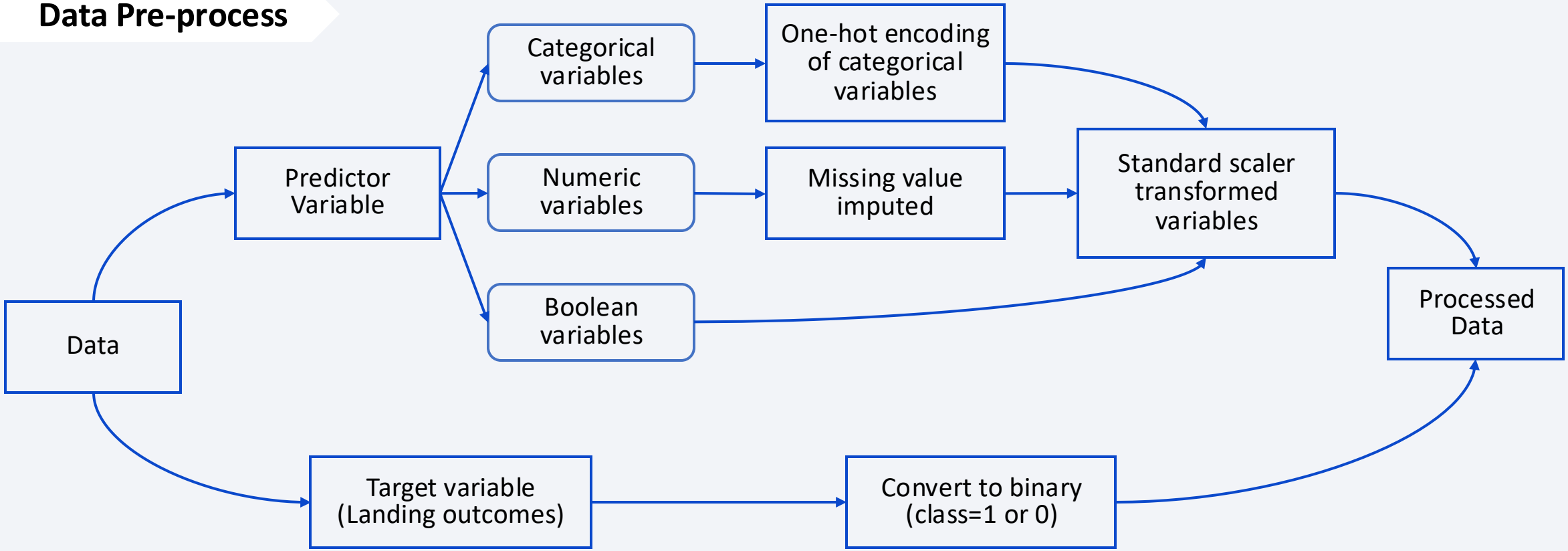
1. **Pie chart:** Displays the count of successful and failed launches for the selected site. When “All Sites” is selected, it shows the proportion of successful launches across different sites.
2. **Scatter plot:** Illustrates the relationship between payload mass and launch outcome, with booster categories represented through color overlays for deeper insight.

Complete code for the Dashboard can be found on GitHub: https://github.com/YanranChen/IBM-Data-Science/blob/main/C10_Applied_Data_Science_Capstone/M3_Interactive_Analytics_and_Dashboards/M3Lab2_spacex-dash-app.py

© 2025 Yanran Chen. All rights reserved.

Predictive Analysis (Classification)

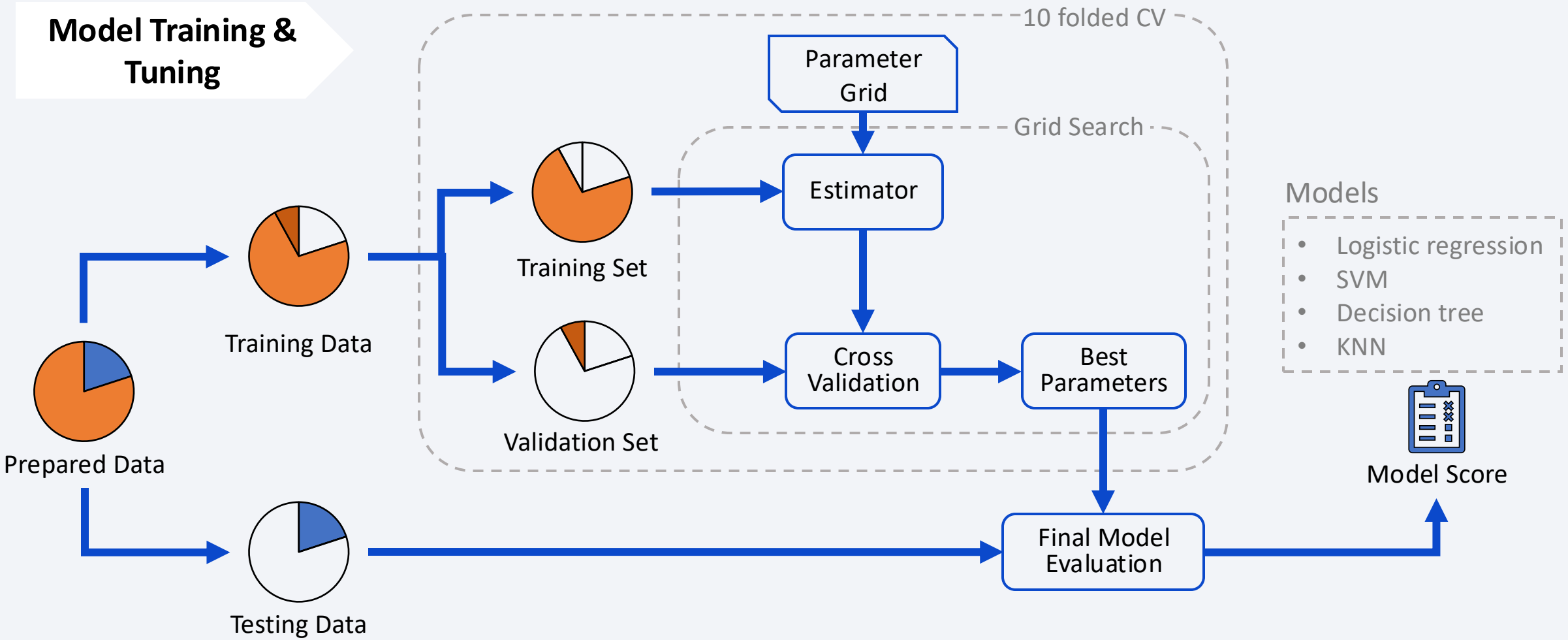
Data Pre-process



Complete code for the Dashboard can be found on GitHub: https://github.com/YanranChen/IBM-Data-Science/blob/main/C10_Applied_Data_Science_Capstone/M4_Predictive_Analysis/M4Lab1_SpaceX_Machine_Learning_Prediction.ipynb

© 2025 Yanran Chen. All rights reserved.

Predictive Analysis (Classification)

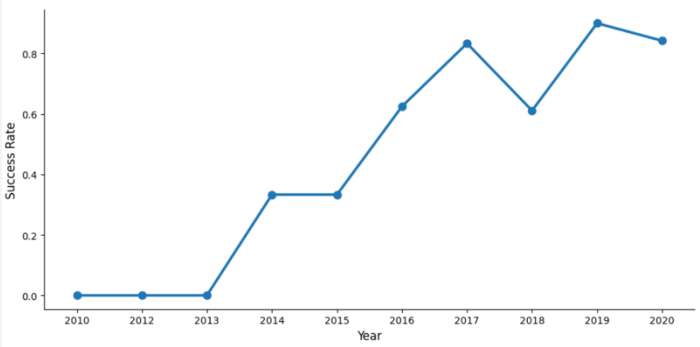


Complete code for the Dashboard can be found on GitHub: https://github.com/YanranChen/IBM-Data-Science/blob/main/C10_Applied_Data_Science_Capstone/M4_Predictive_Analysis/M4Lab1_SpaceX_Machine_Learning_Prediction.ipynb

© 2025 Yanran Chen. All rights reserved.

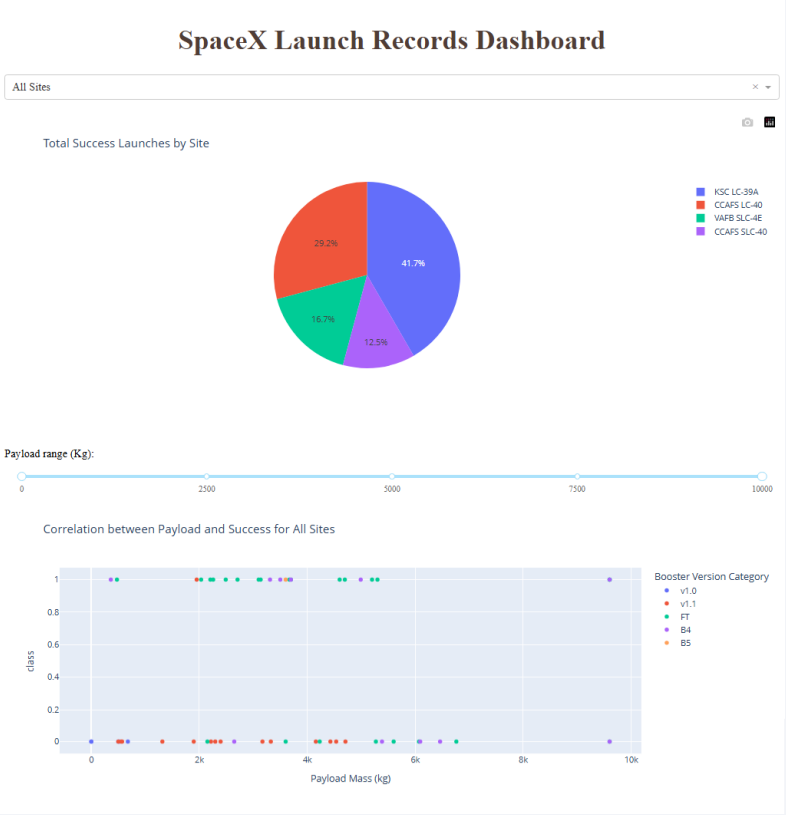
Results

EDA

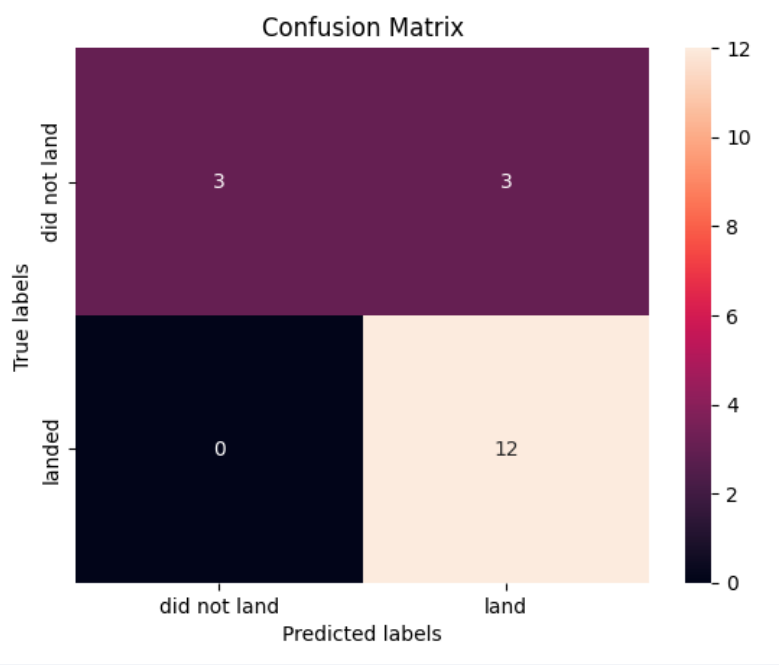


Interactive Map

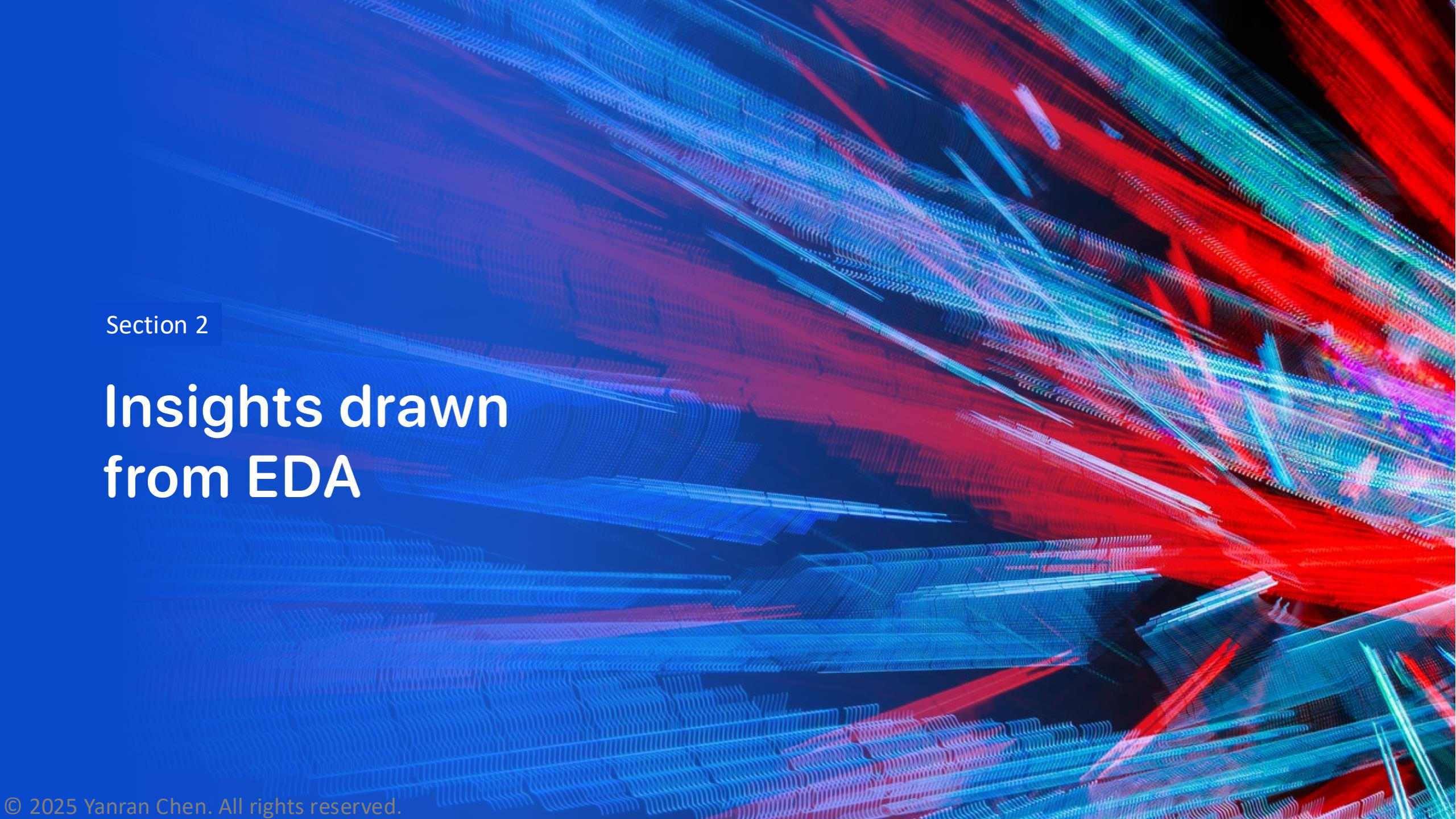
Dashboard



Predictive outcome



© 2025 Yanran Chen. All rights reserved.

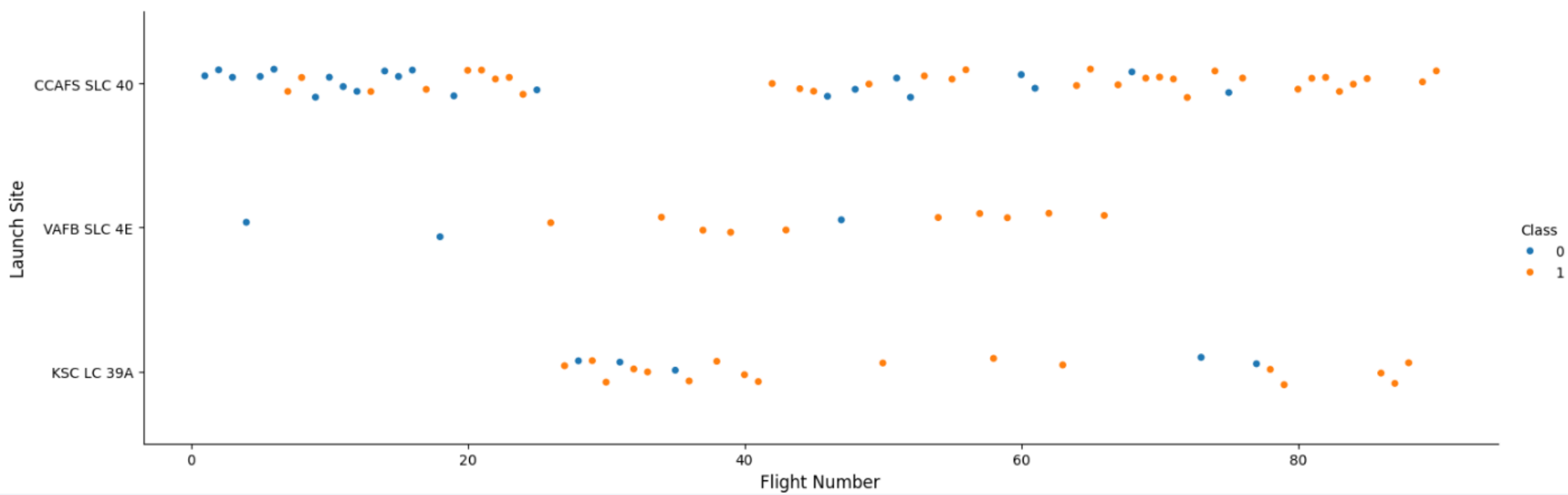


Section 2

Insights drawn from EDA

© 2025 Yanran Chen. All rights reserved.

Flight Number vs. Launch Site

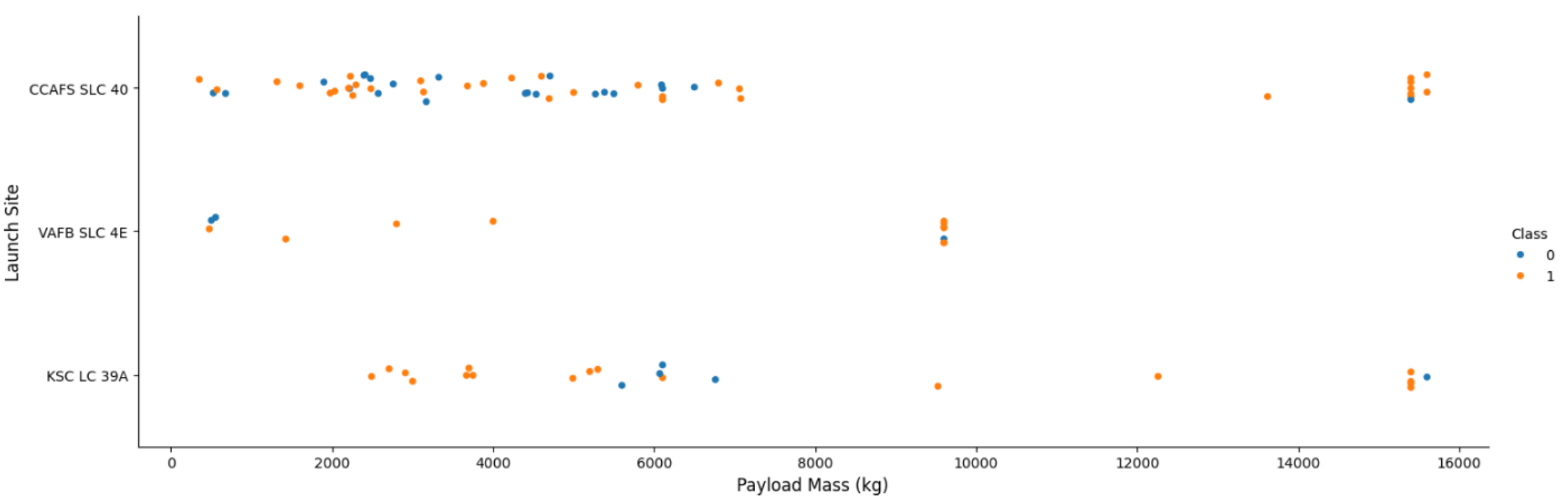


Observations:

1. Early launches were almost all from CCAFS SLC-40 and had low success rates.
2. Success rates improved with higher flight numbers, likely reflecting technological maturity over time.
3. Similar patterns appeared at the other two launch sites, suggesting that launch site location is not a major driver of landing success.

© 2025 Yanran Chen. All rights reserved.

Payload vs. Launch Site

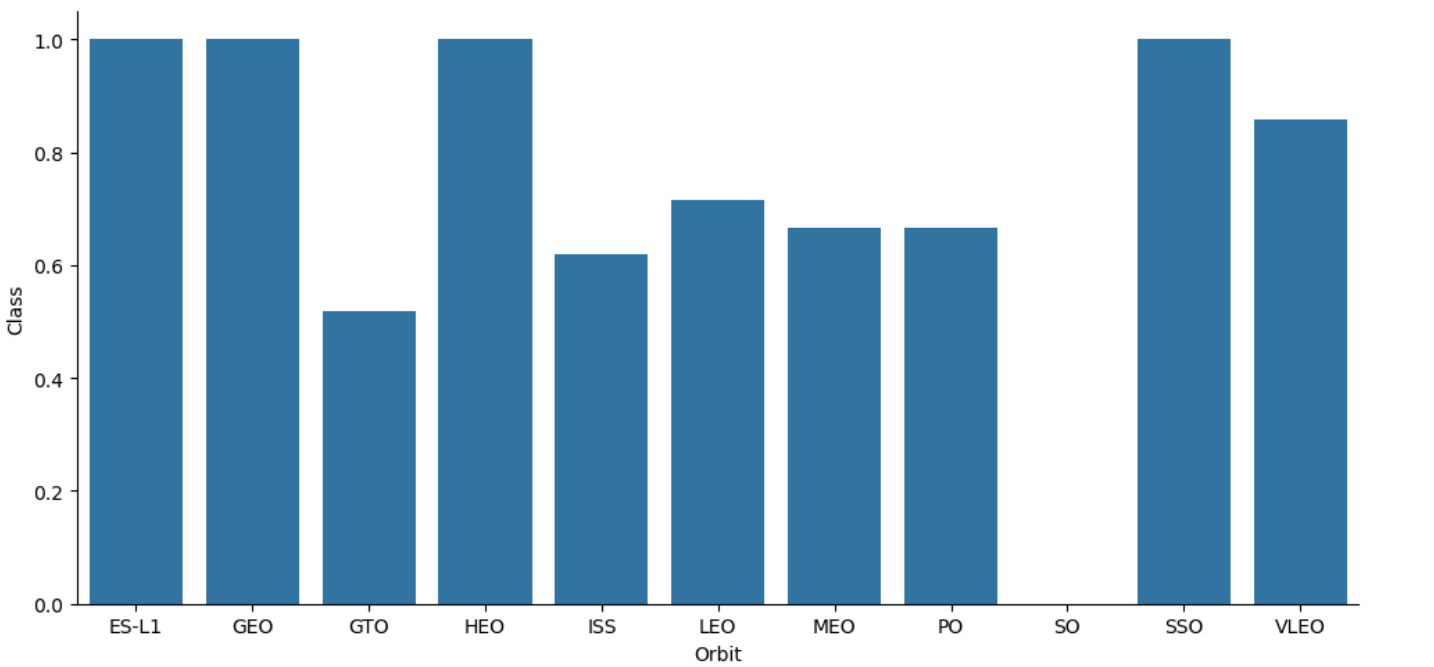


Observations:

1. Heaviest payload masses were mostly launched at CCAFS SLC-40 and KSC LC-39A sites
2. VAFB-SLC launch site was frequently used when payload mass was right below 10,000kg but never used when payload mass is beyond 10,000kg.

© 2025 Yanran Chen. All rights reserved.

Success Rate vs. Orbit Type



Launches of each orbit

```
Orbit
GTO     27
ISS     21
VLEO    14
PO       9
LEO      7
SSO      5
MEO      3
HEO      1
ES-L1    1
SO       1
GEO      1
Name: count, dtype: int64
```

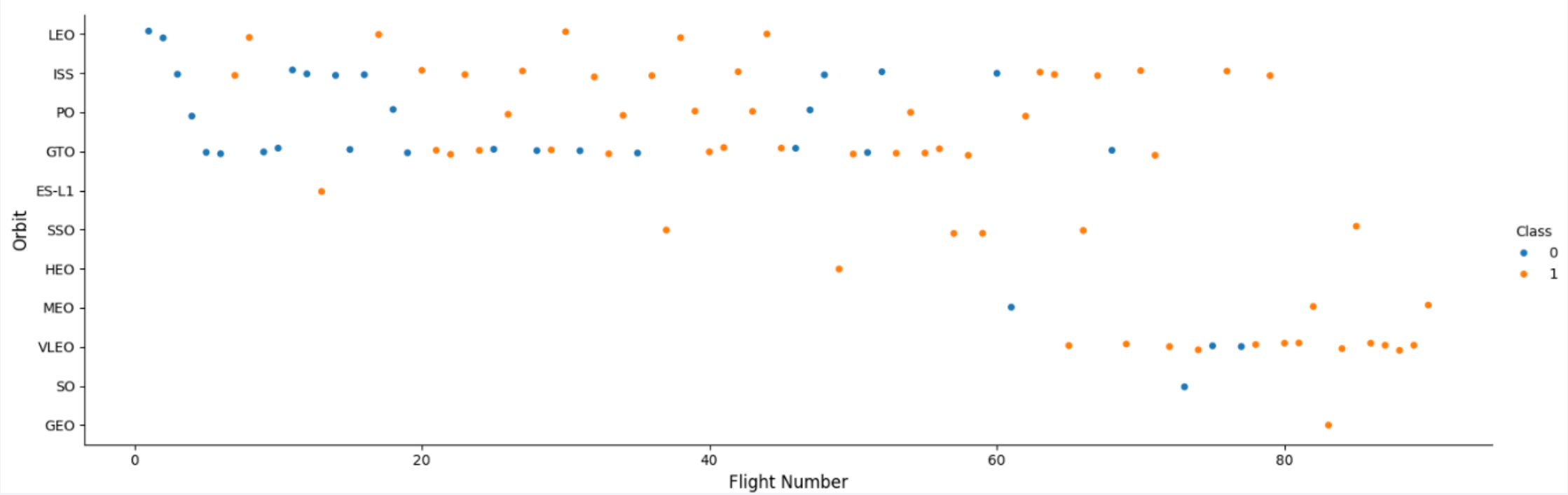
Observations:

1. Exclude orbits that had only 1 launch, SSO (5 launches) has the highest success rate (100%), VLEO (14 launches) also has relatively good success rate (above 80%)

*Note: GTO is a transfer orbit and not itself geostationary.

© 2025 Yanran Chen. All rights reserved.

Flight Number vs. Orbit Type

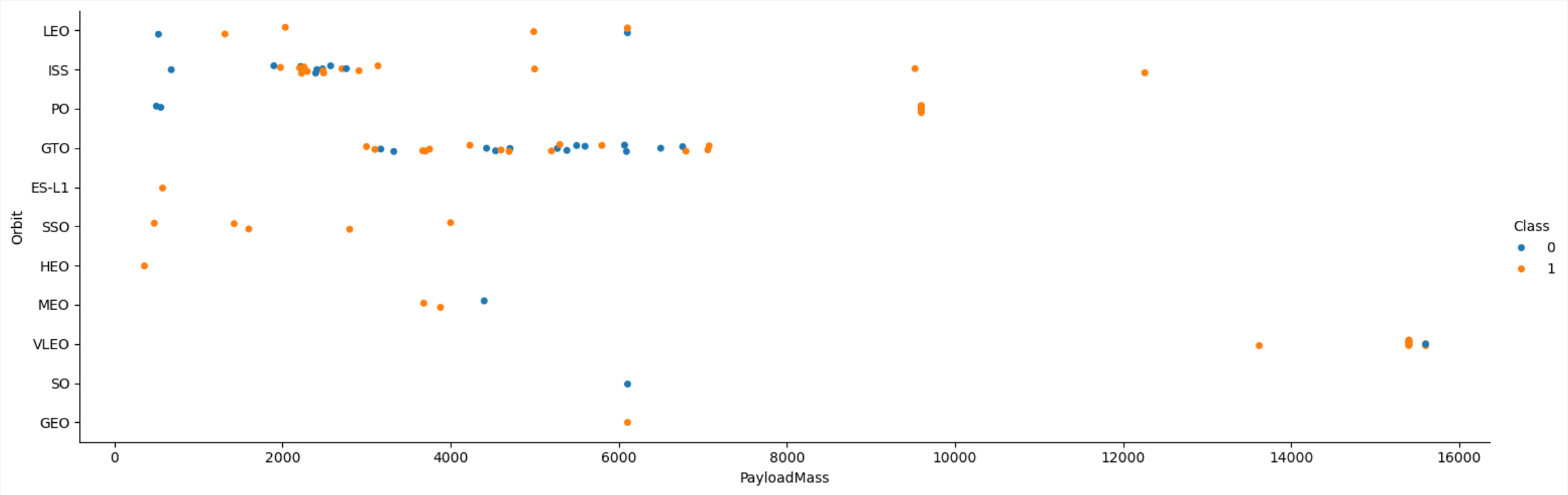


Observations:

1. In the LEO orbit, success seems to be related to the number of flights
2. VLEO orbit was frequently used for later launches (Flight Number >60), which, when combined with the previous observation of its high success rate, still indicate the maturity of the program seem to be a major driver.

© 2025 Yanran Chen. All rights reserved.

Payload vs. Orbit Type

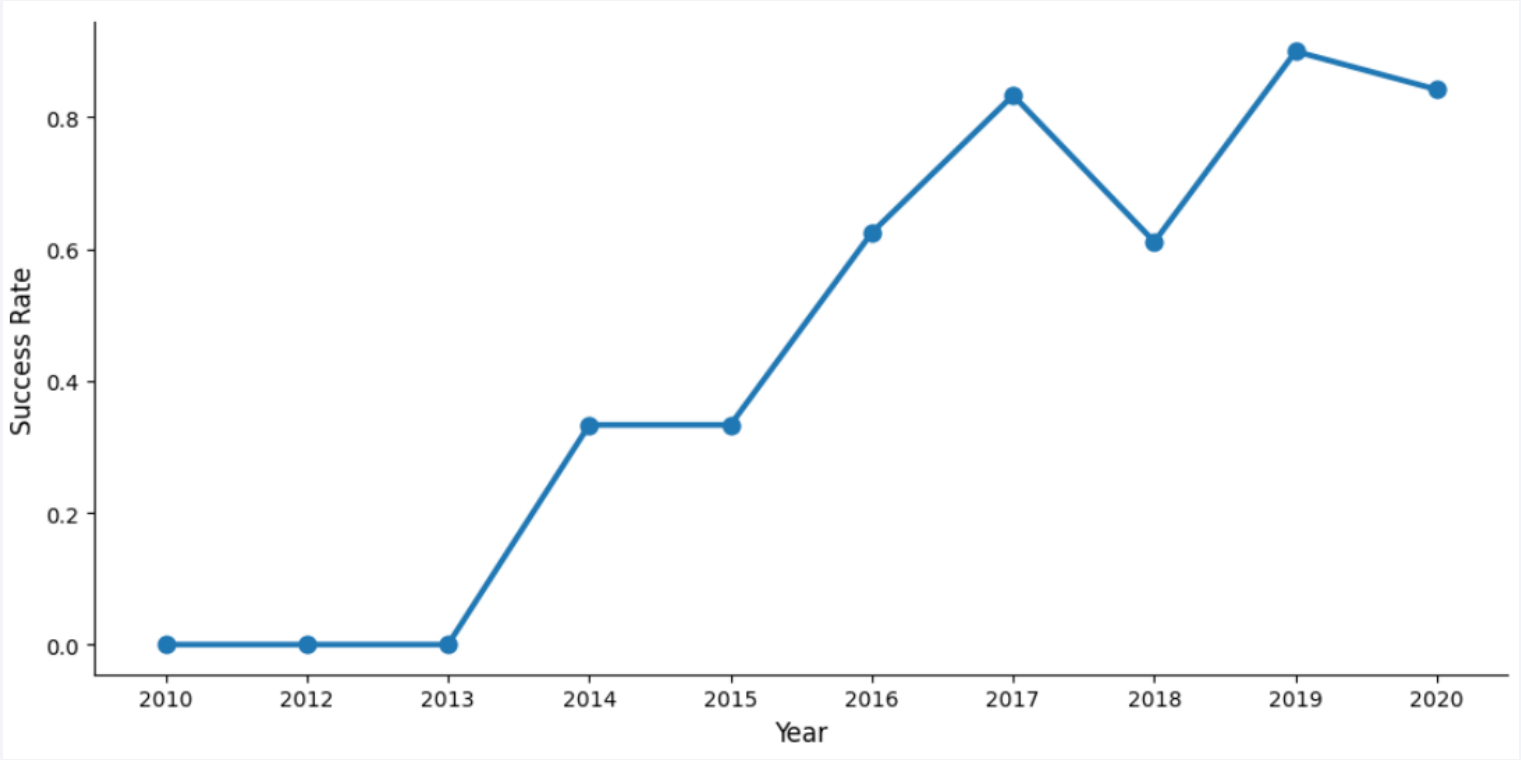


Observations:

1. With heavy payloads the successful landing or positive landing rate are more for PO and ISS.
2. SSO had 100% success rate with light payloads.

© 2025 Yanran Chen. All rights reserved.

Launch Success Yearly Trend



Observation:

As it has been repeated observed that Flight Number seems to be the biggest factor for success rate, our final chart here explores the success rate yearly trend. And we can see that success rate kept improving since 2013 through 2017, with a slight dip in 2018, and rebounded in 2019.

© 2025 Yanran Chen. All rights reserved.

All Launch Site Names

4 unique launch sites can be found in this dataset: CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A

```
%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

© 2025 Yanran Chen. All rights reserved.

Launch Site Names Begin with 'CCA'

Below is 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site like 'CCA%' LIMIT 5;
```

```
 * 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

© 2025 Yanran Chen. All rights reserved.

Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS): 45,596 KG

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as [Total payload mass] FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

Total payload mass
45596

© 2025 Yanran Chen. All rights reserved.

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1: 25,34.67 KG

```
%sql SELECT Avg(PAYLOAD_MASS__KG_) as [Avg payload mass] FROM SPACEXTABLE WHERE Booster_Version like 'F9 v1.1%'
```

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

Avg payload mass
2534.6666666666665

© 2025 Yanran Chen. All rights reserved.

First Successful Ground Landing Date

Date when the first successful landing outcome in ground pad was achieved: 2015-12-22

```
%sql SELECT min(Date) FROM SPACEXTABLE where Landing_Outcome='Success (ground pad)'
```

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

min(Date)
2015-12-22

© 2025 Yanran Chen. All rights reserved.

Successful Drone Ship Landing with Payload between 4000 and 6000

Four boosters had success in drone ship and have payload mass between 4000 and 6000, they are: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

```
%%sql
SELECT distinct Booster_Version FROM SPACEXTABLE
where Landing_Outcome='Success (drone ship)'
    and PAYLOAD_MASS__KG_>4000
    and PAYLOAD_MASS__KG_<6000
```

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

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

© 2025 Yanran Chen. All rights reserved.

Total Number of Successful and Failure Mission Outcomes

Among all mission outcomes, there are 61 Success and 10 Failure

```
%%sql
SELECT sum(case when Landing_Outcome  like 'Success%' then 1 else 0 end) as Success
    , sum(case when Landing_Outcome  like 'Failure%' then 1 else 0 end) as Failure
FROM SPACEXTABLE
```

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

Success	Failure
61	10

© 2025 Yanran Chen. All rights reserved.

Boosters Carried Maximum Payload

Below is a list of all the booster_versions that have carried the maximum payload mass

```
%%sql
SELECT distinct Booster_Version
FROM SPACEXTABLED
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

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

© 2025 Yanran Chen. All rights reserved.

2015 Launch Records

There are two failed drone shop landing in year 2015, below table shows their booster version, launch_site.

```
%%sql
SELECT substr(Date,0,5) as [Year]
    , case when substr(Date, 6,2) = '01' then 'January'
           when substr(Date, 6,2) = '02' then 'February'
           when substr(Date, 6,2) = '03' then 'March'
           when substr(Date, 6,2) = '04' then 'April'
           when substr(Date, 6,2) = '05' then 'May'
           when substr(Date, 6,2) = '06' then 'June'
           when substr(Date, 6,2) = '07' then 'July'
           when substr(Date, 6,2) = '08' then 'August'
           when substr(Date, 6,2) = '09' then 'September'
           when substr(Date, 6,2) = '10' then 'October'
           when substr(Date, 6,2) = '11' then 'November'
           when substr(Date, 6,2) = '12' then 'December'
    end as [Month]
    ,  Landing_Outcome, Booster_Version,Launch_Site
FROM SPACEXTABLE
where substr(Date,0,5) = '2015'
    and Landing_Outcome = 'Failure (drone ship)'
```

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

Year	Month	Landing_Outcome	Booster_Version	Launch_Site
2015	January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

© 2025 Yanran Chen. All rights reserved.

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

Below is the count of landing outcomes between the date 2010-06-04 and 2017-03-20, ranked in descending order.

```
%%sql
SELECT Landing_Outcome, count(*)
FROM SPACEXTABLE
where Date>='2010-06-04' and Date <='2017-03-20'
group by Landing_Outcome
order by 2 desc
```

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

Landing_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

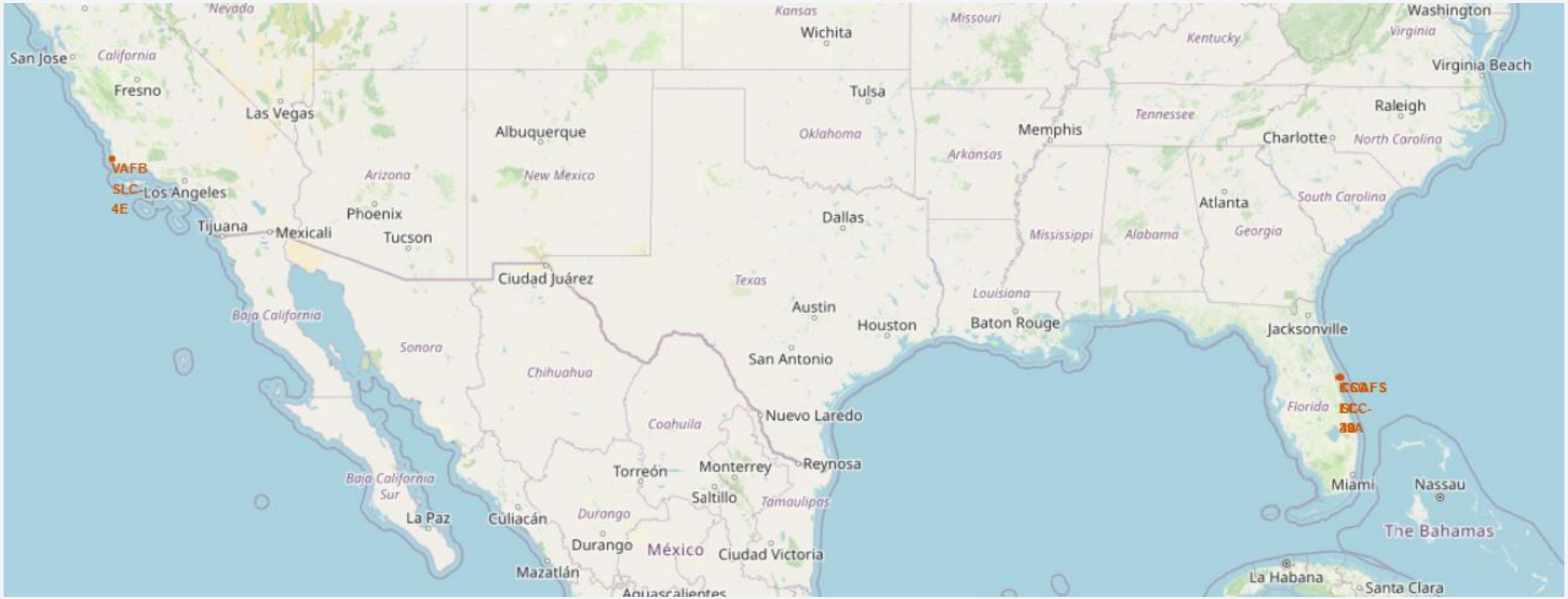
© 2025 Yanran Chen. All rights reserved.

Section 3

Launch Sites Proximities Analysis

© 2025 Yanran Chen. All rights reserved.

SpaceX Launch Site Locations



Key Insights:

- All SpaceX launch sites are strategically located near the Equator, optimizing fuel efficiency and orbital mechanics for missions.
- Each site is situated in close proximity to the coastline, which provides a safe trajectory over open water and facilitates logistical access for recovery operations.

© 2025 Yanran Chen. All rights reserved.

Launch Outcomes for Each Site

Key Insights:

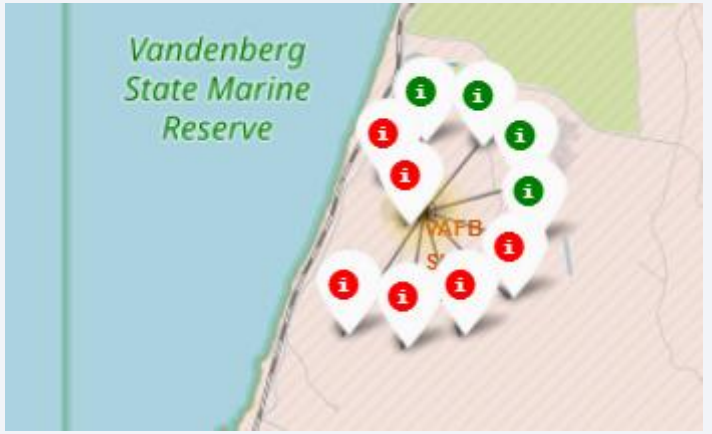
Mapping SpaceX launch outcomes reveals a clear performance distinction between sites:

- KSC LC-39A stands out with the highest success rate, underscoring its reliability for critical missions.
- In contrast, CCAFS LC-40, while hosting the highest number of launches, shows a comparatively lower success rate.
- This suggests that LC-40 may be used for higher-frequency, lower-risk missions, while LC-39A is reserved for high-stakes operations requiring proven dependability.

Launch Site CCAFS SLC-40

Launch Site KSC LC-39A

Launch Site CCAFS LC-40



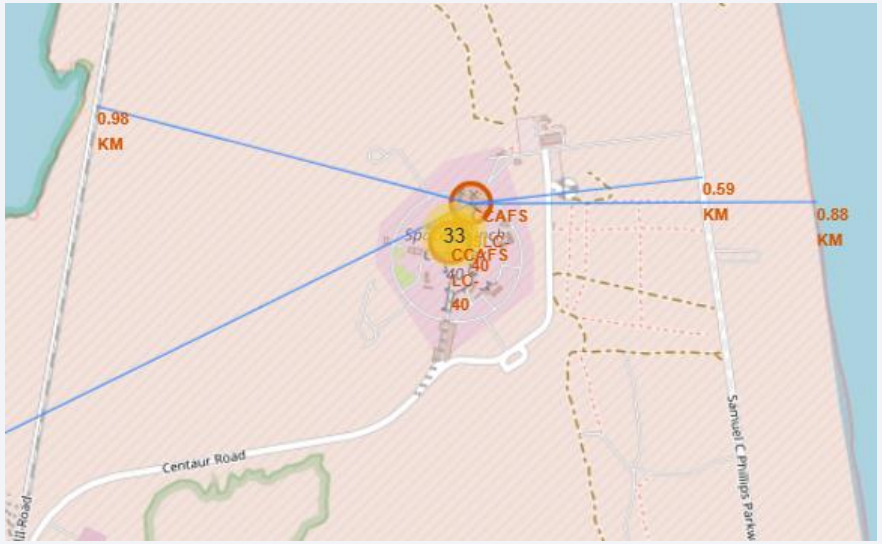
Launch Site VAFB SLC-4E

© 2025 Yanran Chen. All rights reserved.

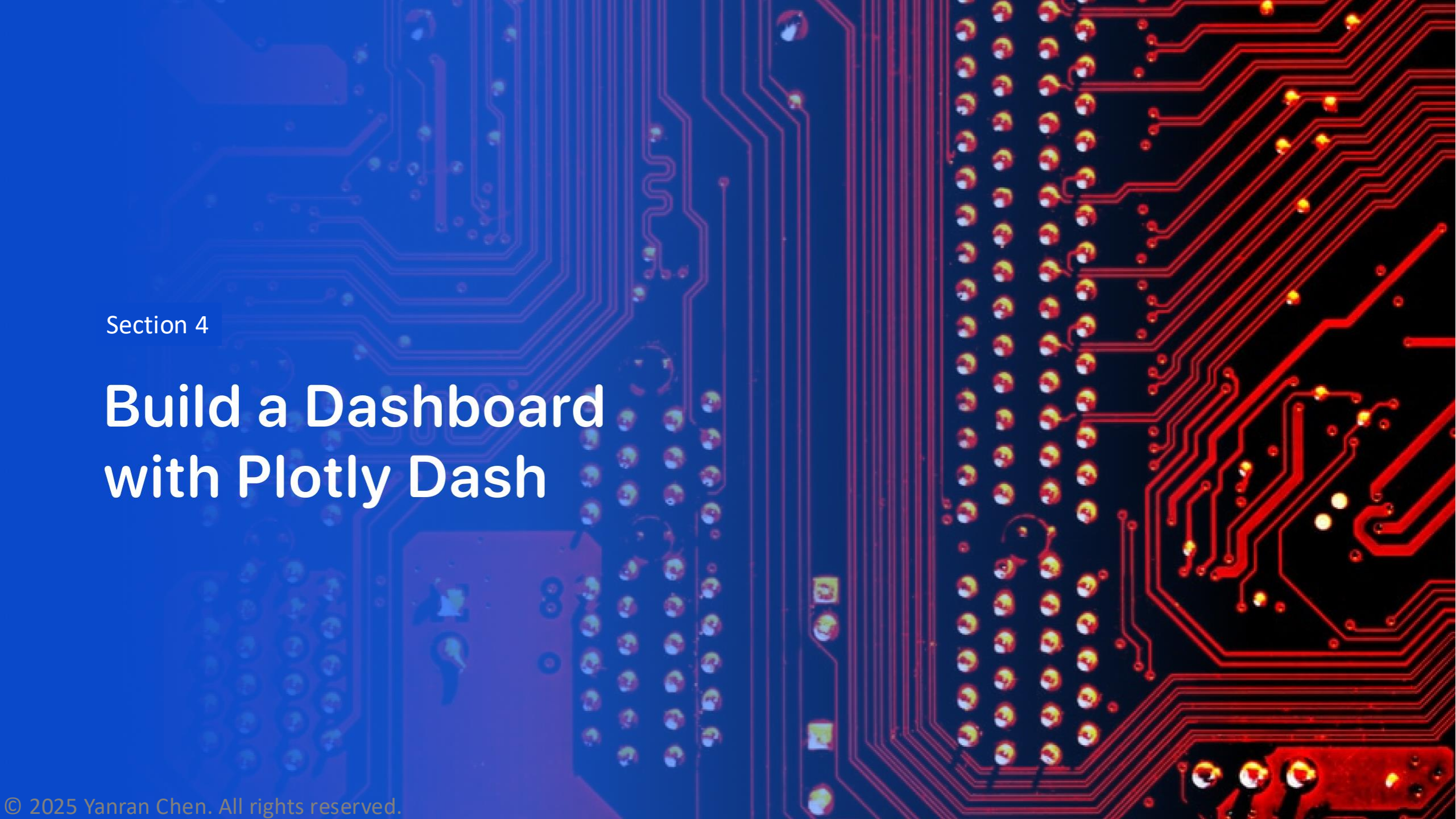
Proximities of Launch Sites

Key Insights:

- All launch sites are strategically located within close proximity – typically less than 1km – to railways, highways, and coastlines, facilitating logistical efficiency and safety. However, they are deliberately situated at a greater distance from populated cities to mitigate risk and minimize disruption.
- The screenshots below illustrated the proximities of the CCAFS SLC-40 launch site. It is located within 1 km of the nearest railway, highway, and coastline, while roughly 22 km away from the closest city.



© 2025 Yanran Chen. All rights reserved.

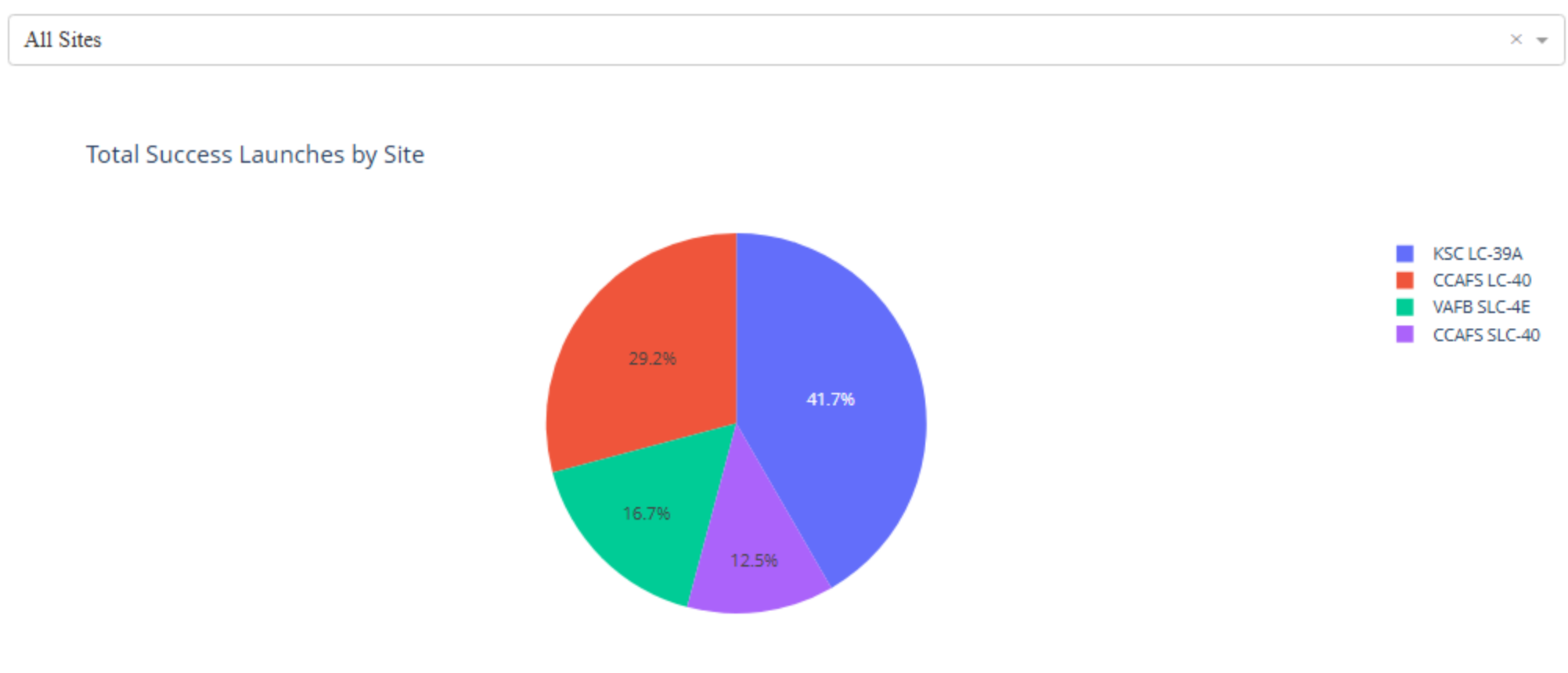


Section 4

Build a Dashboard with Plotly Dash

© 2025 Yanran Chen. All rights reserved.

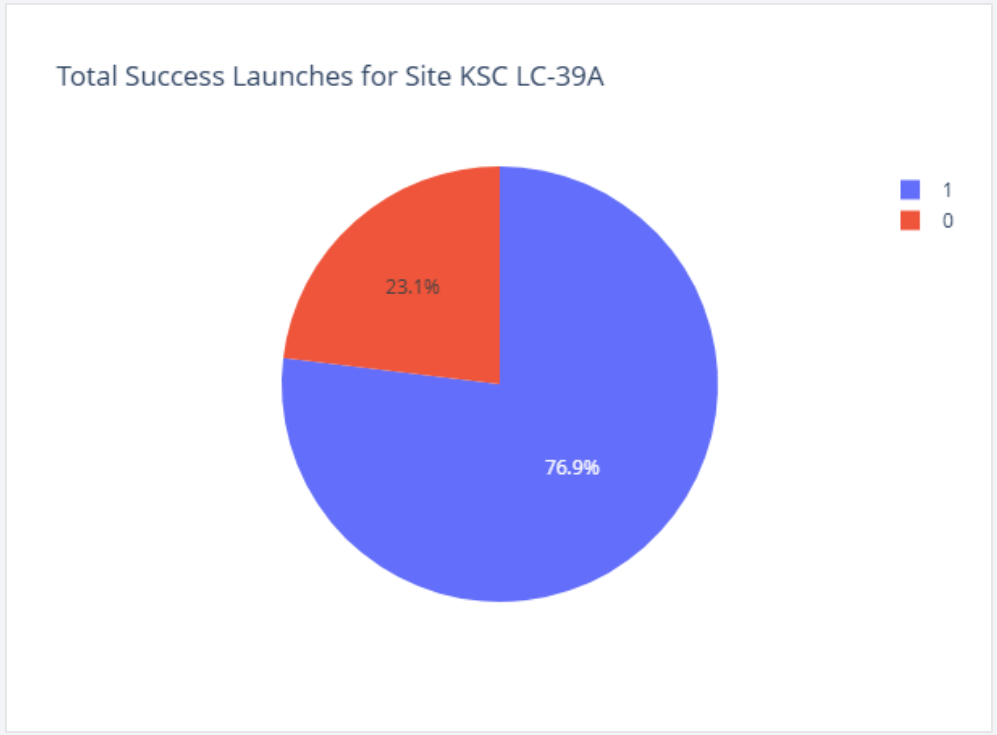
Share of Success Launches by Site



Key Insights: Launch Site KSC LC-39A leads in successful missions, contributing 41.7% of all SpaceX launch successes. This highlights its role as the most productive launch site in the dataset.

© 2025 Yanran Chen. All rights reserved.

Launch Site with Highest Success Rate



Key Insights:

Launch Site KSC LC-39A not only leads in total successful launches but also boasts the highest launch success rate at 77%. This underscores its operational reliability and strategic importance within SpaceX's launch infrastructure. CCAFS LC-40 follows closely with a 73% success rate.

© 2025 Yanran Chen. All rights reserved.

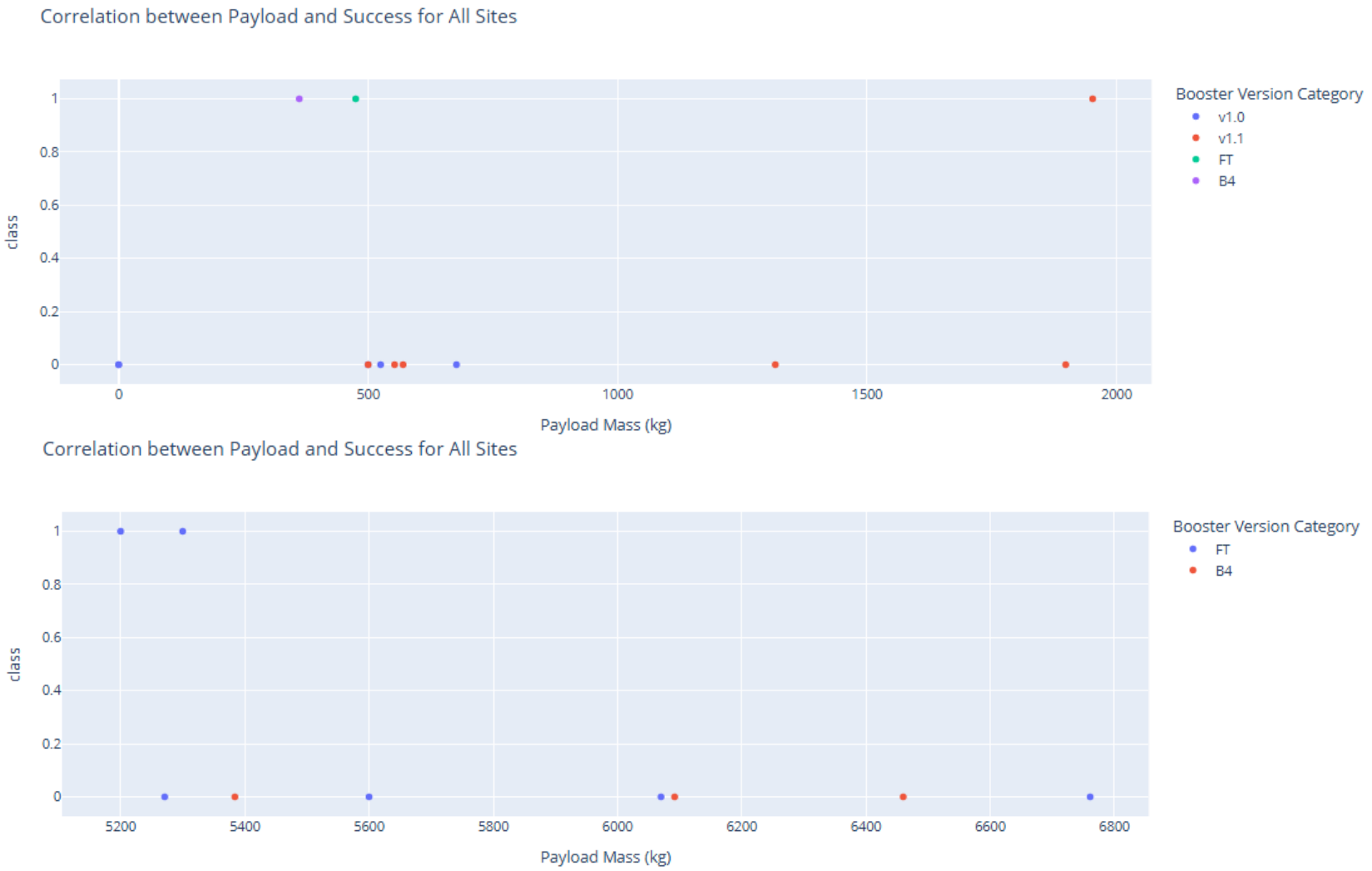
Payload Range with the Highest Success Rate

Key Insight:

- Launches with payloads between 3,000 kg and 4,000 kg achieved the highest success rate at 73% across 11 missions
  - Note: some launches within this payload range share identical payload masses, causing overlapping data points on the chart.

© 2025 Yanran Chen. All rights reserved.

Payload Ranges with the Lowest Success Rates

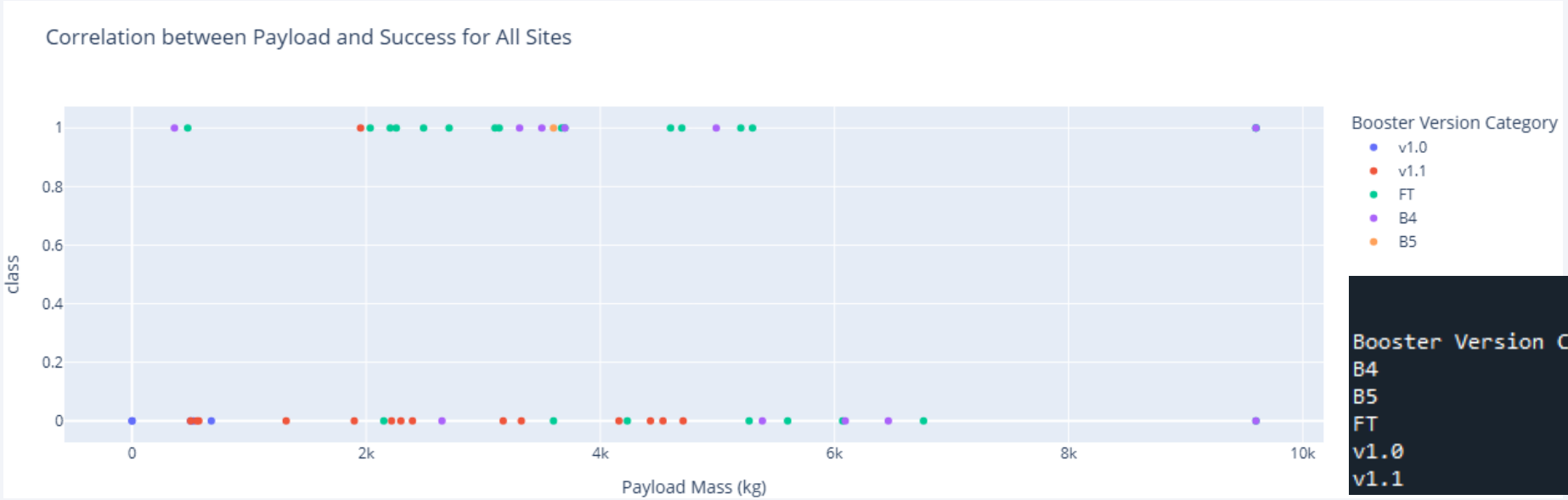


Key Insights:

- Launches carrying payloads between 0 kg and 2,000 kg exhibited a low success rate of just 23% across 13 missions, suggesting potential challenges with lighter payload configurations.
- The 5,000 kg to 7,000 kg range saw 2 successful launches out of 9 attempts, a success rate of 22%.

© 2025 Yanran Chen. All rights reserved.

Success Rates by Booster



Booster Version Category	class count	mean
B4	11	0.545455
B5	1	1.000000
FT	24	0.666667
v1.0	5	0.000000
v1.1	15	0.066667

Key Insight:

- FT appears to be the most reliable and frequently used booster, with a solid 66.7% success rate across 24 launches.
- B5 shows perfect performance, but with only one launch, it's too early to draw conclusions.
- B4 has moderate reliability at 54.5% over 11 launches.
- Meanwhile, v1.0 and v1.1 stand out as underperformers, with extremely low success rates (0.0% and 6.7%, respectively), suggesting they may need redesign or retirement. Overall, FT seems the strongest candidate for continued use and improvement.

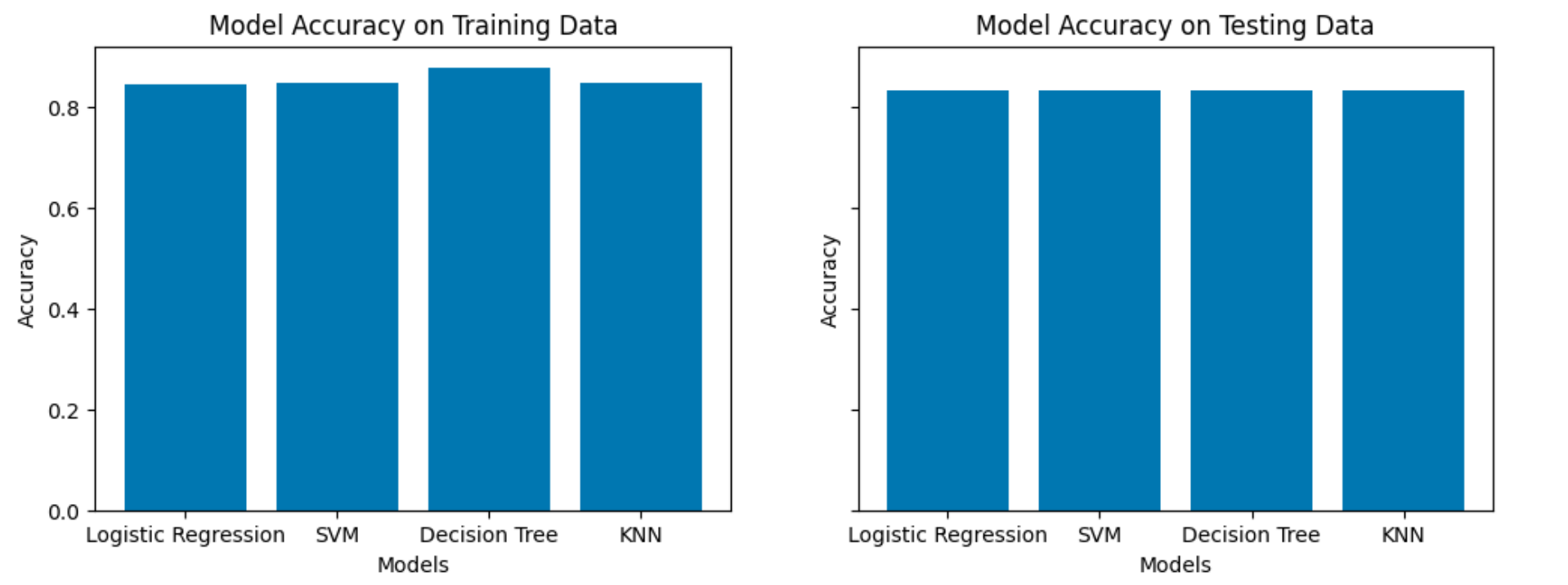
© 2025 Yanran Chen. All rights reserved.

Section 5

Predictive Analysis (Classification)

© 2025 Yanran Chen. All rights reserved.

Classification Accuracy

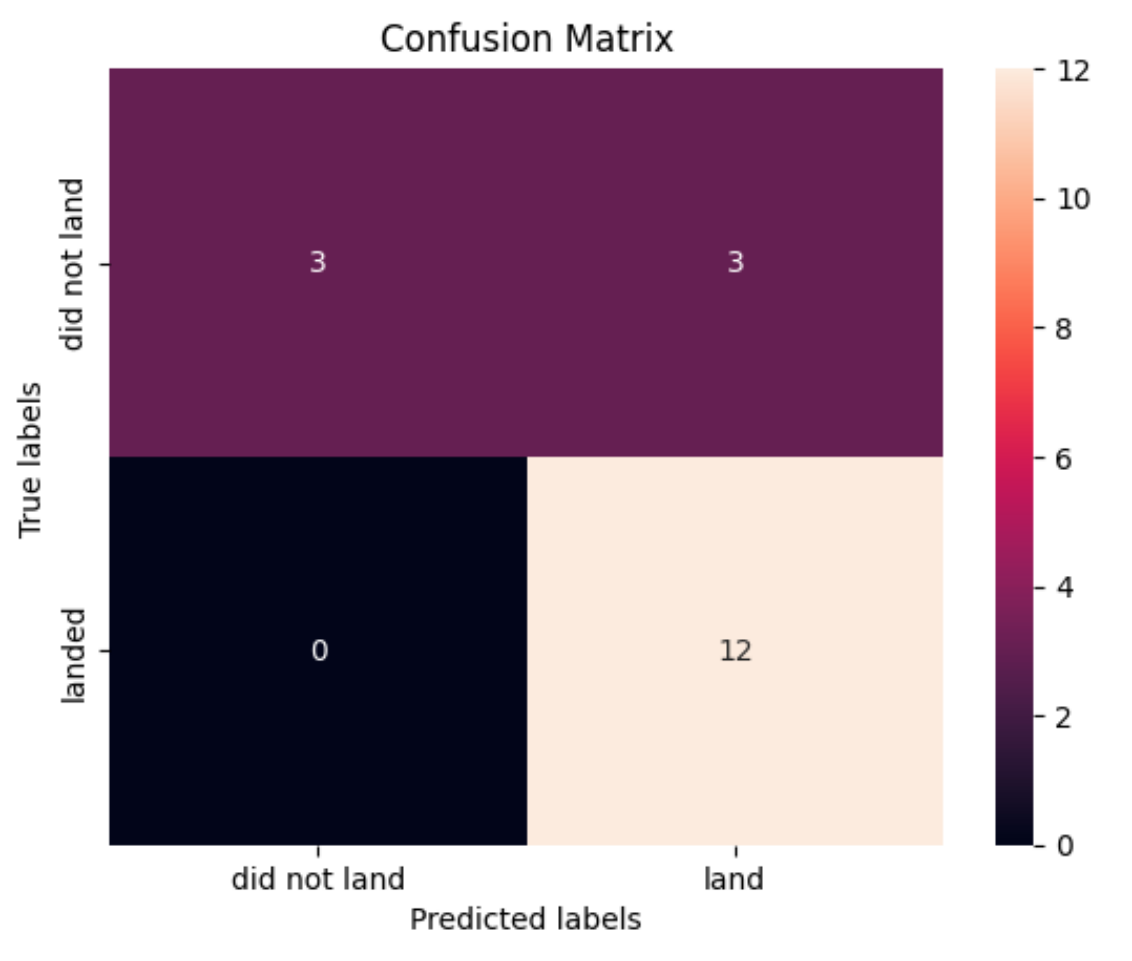


Key Insights:

- Trained four models (Logistic Regression, SVM, Decision Tree, KNN) using 10-fold CV with optimized hyperparameters.
- **Decision Tree** had the highest training accuracy (87.7% vs ~85% for others).
- On testing, all models achieved the same accuracy (83.3%).
- Test set is small (18 records), which limits confidence in performance differences.

© 2025 Yanran Chen. All rights reserved.

Confusion Matrix



Key Insights:

All four models produced identical confusion matrices.

- True Positives (TP): 12 – correctly predicted “landed.”
- False Positives (FP): 3 – predicted “landed” when true label was “not landed.”
- True Negatives (TN): 3 – correctly predicted “not landed.”
- False Negatives (FN): 0.

Models show strong performance in identifying “landed” cases, but some “not landed” cases are still misclassified. Limited test size (18 records) reduces confidence in generalization.

© 2025 Yanran Chen. All rights reserved.

Conclusions

- All four models (Logistic Regression, SVM, Decision Tree, KNN) achieved similar performance on the test set, with an accuracy of 83.3%.
- The Decision Tree model performed best on the training set (87.7%), but this advantage did not carry over to the test set.
- Identical confusion matrices across models suggest consistent predictive behavior: strong detection of “landed” cases, but occasional misclassification of “not landed” cases.
- Due to the very small data set (90 records), results should be considered preliminary. Larger datasets are needed to validate and differentiate model performance.

Recommendations for improvement:

- **Expand the dataset** to improve statistical reliability and better assess model generalization.
- **Explore ensemble methods** for more stable predictive performance.

© 2025 Yanran Chen. All rights reserved.

Appendix

All notebooks, including code and charts, can be found on GitHub:

https://github.com/YanranChen/IBM-Data-Science/tree/main/C10_Applied_Data_Science_Capstone

© 2025 Yanran Chen. All rights reserved.

Thank you!