IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Danny Nasibu

June 1st , 2025

Outline

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

Executive Summary

This project aimed to predict the successful first-stage landing of SpaceX Falcon 9 rockets, a critical factor in the commercial space industry's cost reduction and expansion. Data was meticulously collected from the SpaceX API and Wikipedia, then meticulously wrangled to create a clean dataset, including transforming landing outcomes into a binary classification label.

Through extensive Exploratory Data Analysis (EDA) using visualizations and SQL queries, key insights were uncovered:

- Learning Curve**:** SpaceX demonstrated a significant learning curve, with early launches showing lower success rates, which drastically improved over time across all launch sites and orbit types.
- Launch Site Impact**:** CCAFS SLC 40 served as the primary testbed for reusability, while KSC LC 39A and VAFB SLC 4E consistently achieved high success rates from their introduction, even with varying payload masses. Launch sites are strategically located near oceans, railways, and highways.
- Orbit Versatility**:** Success rates are remarkably high across diverse orbits (LEO, GTO, SSO, ISS, etc.), indicating robust recovery capabilities for various mission profiles.
- Payload Management**:** While heavier payloads presented initial challenges, SpaceX has largely mastered recovery across a wide range of payload masses.

Finally, predictive analysis using classification models (Logistic Regression, SVM, Decision Tree, KNN) achieved an approximate **83% accuracy** in predicting landing outcomes. The K-Nearest Neighbors model, chosen for its simplicity, effectively identified successful landings (100% recall) but showed limitations in predicting failures (50% recall). This project provides valuable insights into the factors influencing Falcon 9's reusability, crucial for optimizing future commercial space operations.

Introduction

New Era: Private firms now lead space travel and access, expanding beyond traditional missions.

SpaceX's Impact: Falcon 9's reusable first stage is a game-changer, significantly lowering launch expenses.

Business Need: Accurate cost estimation requires predicting first-stage recovery success.

My Project: Build a predictive model using launch parameters to forecast recovery, enabling better resource planning and pricing in this dynamic industry.

Section 1

Methodology

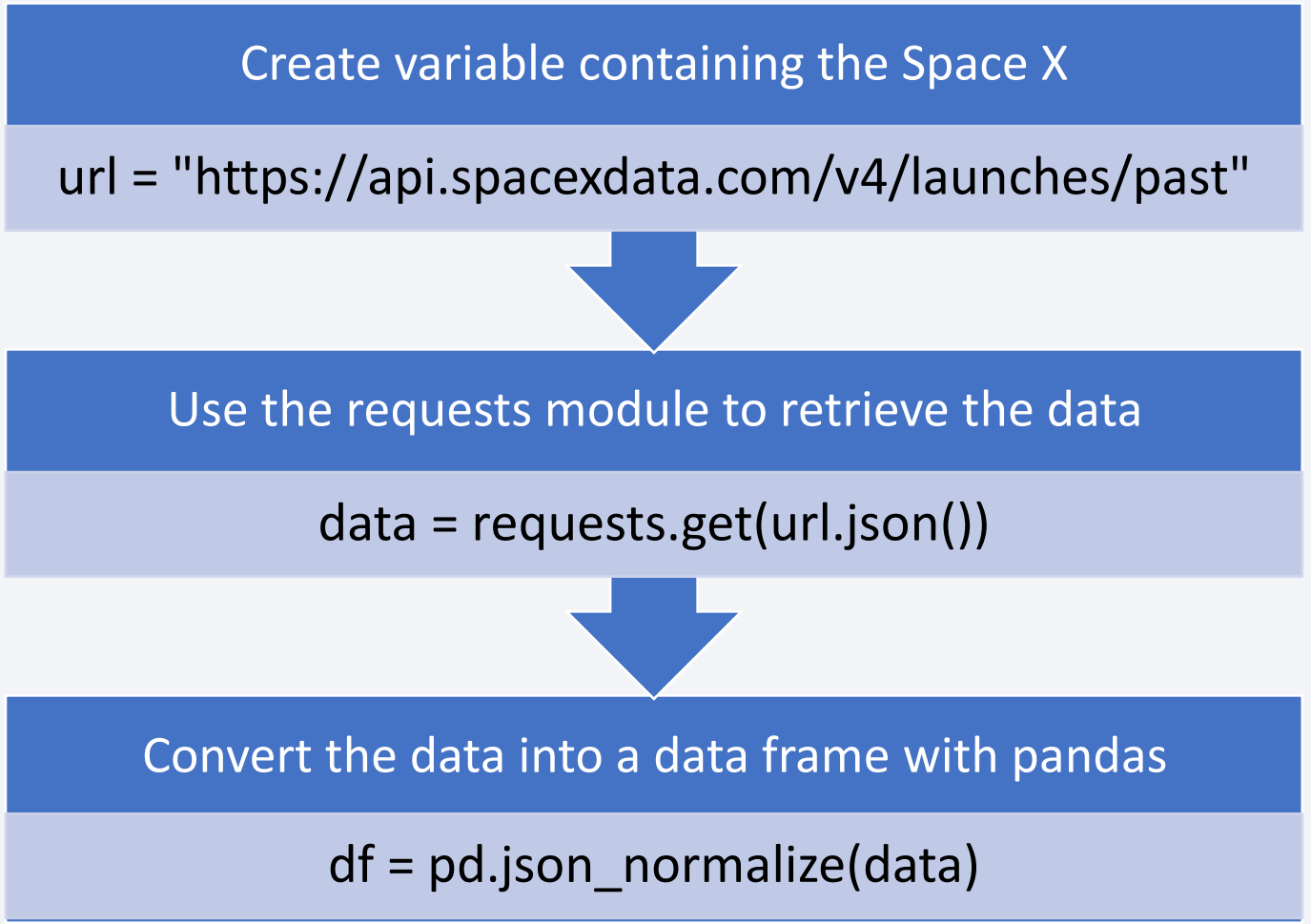
Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

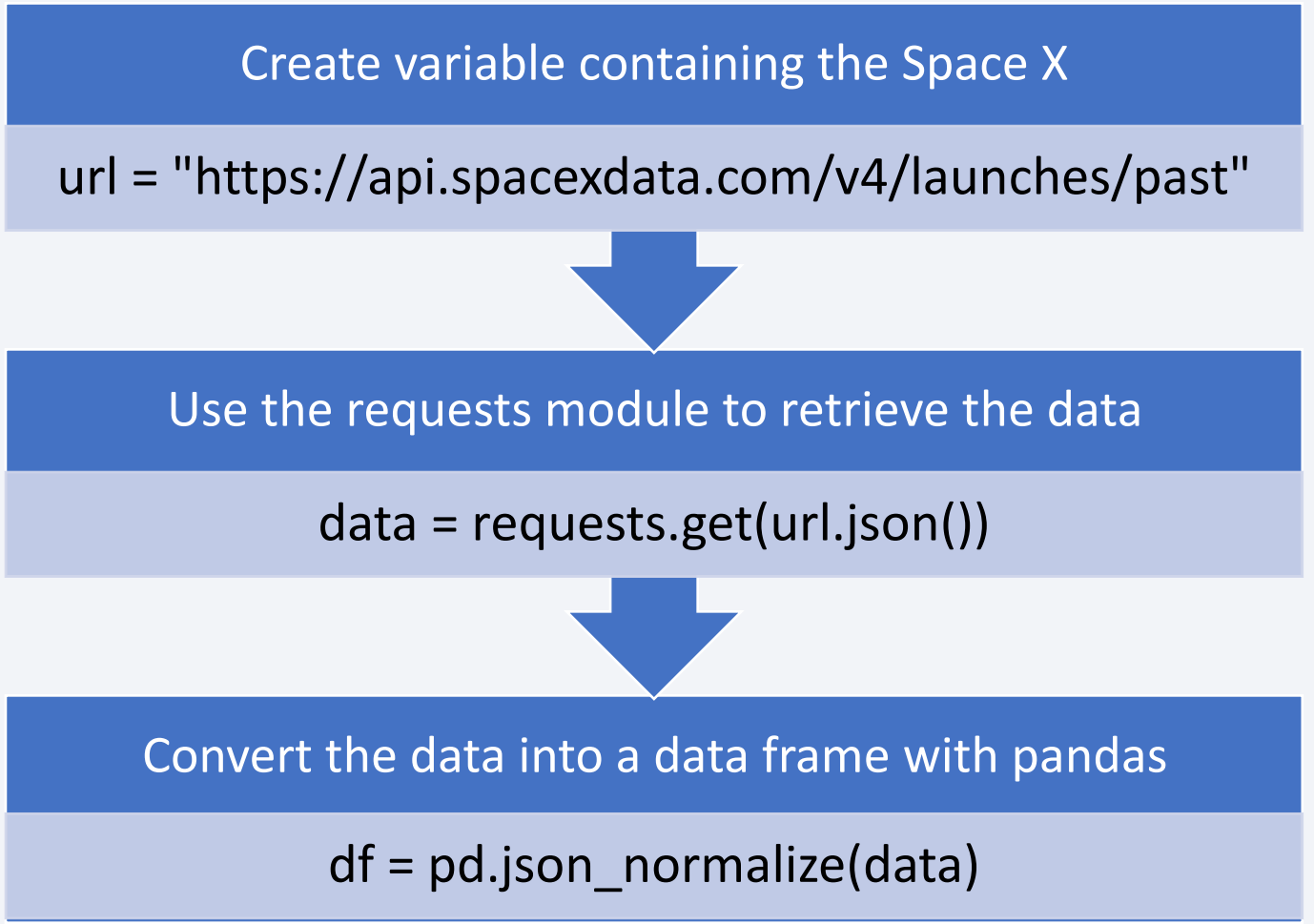
Data Collection – SpaceX API

- **API Endpoint:** Data was primarily sourced from *api.spacexdata.com/v4/launches/past*, which provides historical launch records.
- **Request Method:** a *GET* request, utilizing the *requests* Python library, was performed to retrieve the launch data.
- **Data Formatting**: the raw JSON data returned was then normalized into a flat tabular structure using the *json_normalize* function from the pandas library, facilitating future analysis.
- **GitHub URL**: click me to see the code



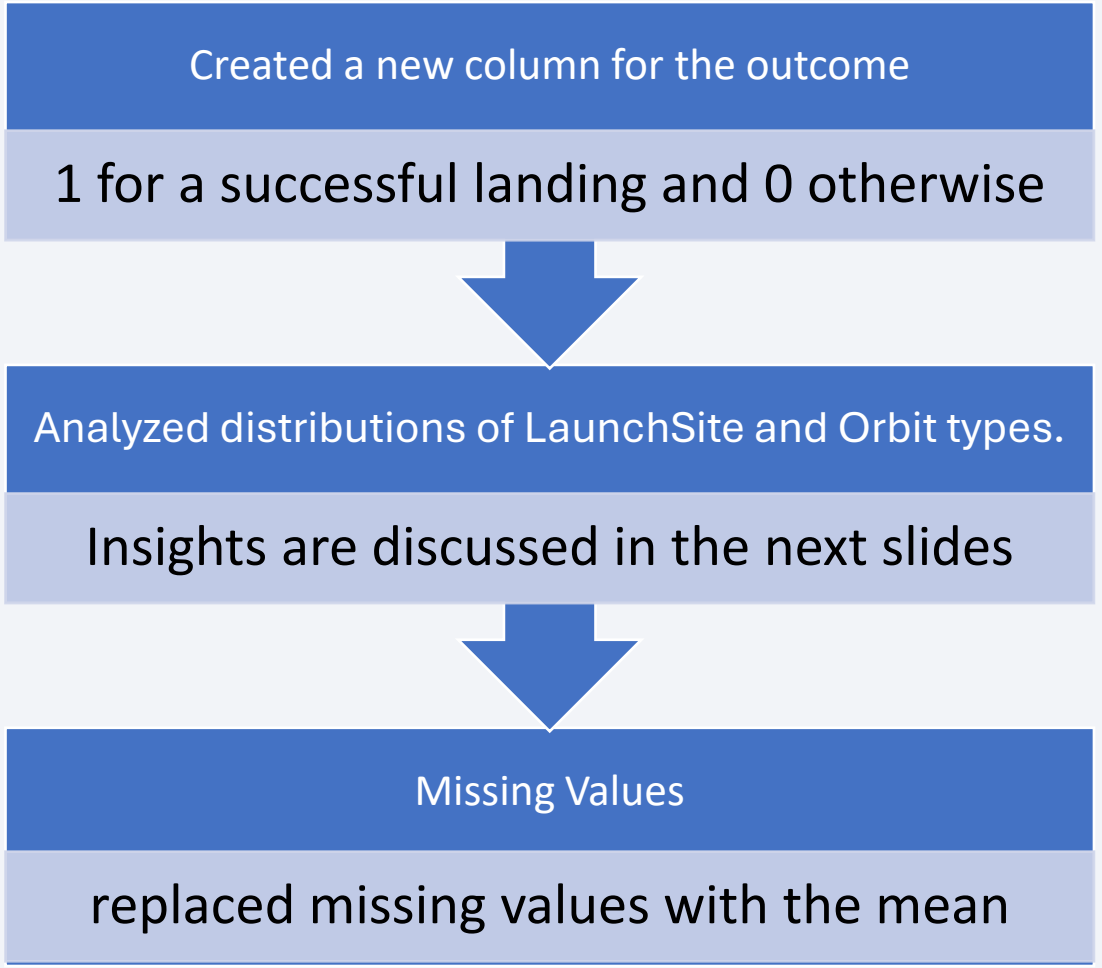
Data Collection – Scraping

- **Source:** The Wikipedia page titled List of Falcon 9 and Falcon Heavy launches was used as the data source.
- **Tool:** The BeautifulSoup Python package was employed to parse the HTML content of the Wikipedia page.
- **Process:** HTML tables containing valuable Falcon 9 launch records were identified and extracted. The raw data from these tables was then parsed and converted into a Pandas DataFrame for subsequent analysis and visualization.
- GitHub URL: click me to see the code



Data Wrangling

- **Purpose:** Clean and prepare SpaceX launch data for machine learning.
- **Attributes:** Focused on FlightNumber, PayloadMass, LaunchSite, Orbit, and Outcome.
- **Outcome Transformation:** Converted diverse Outcome strings (e.g., "True ASDS", "None None") into a binary Class variable:
  - 1: Successful first-stage landing.
  - 0: Unsuccessful first-stage landing.
- **Exploration:** Analyzed distributions of LaunchSite and Orbit types.
- **Missing Values:**
  - PayloadMass column nulls were imputed by replacing them with mean of existing data point.



GitHub URL click me to see the code

EDA with Data Visualization

- A scatter plot chart for the flight number against the launch sites was employed to visually explore how the continuous launch attempts and the launch site influence the success of the first stage landing.
- A scatter plot chart for the payload mass against the launch sites was employed to visually explore how the continuous launch attempts and the launch site influence the success of the first stage landing.
- Bar chart of the success rate of each orbit
- Scatter plot to study the relationship between flight number, the Orbit type, and the outcome of the launch
- Scatter plot to study the relationship between payload mass , the Orbit type, and the outcome of the launch
- Line plot to study the evolution of the rate of successful launches over the years
- GitHub URL: click me to see the code

EDA with SQL

- Query 1: displaying the names of the unique launch sites in the space mission
- Query2: Displaying thr 5 records where launch sites begin with the string 'CCA’
- Query 3: displaying the total payload mass carried by boosters launched by NASA (CRS)
- Query 4: Display average payload mass carried by booster version F9 v1.1
- Query 5: listing the date when the first successful landing outcome in the ground pad was achieved.
- Query 6: listing the names of the boosters that have been successful in drone ship and have a payload mass greater than 4000 but less than 6000
- Query 7: listing the total number of successful and failed mission outcomes
- Query 8: listing7 all the booster versions that have carried the maximum payload mass. Use a subquery.
- Query 9: listing the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
- Query 10: ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order
- GitHub URL: click me to see the code

Build an Interactive Map with Folium

- Created a site map centered at the NASA Johnson Space Center in Houston, Texas.
- Created a Circle object alongside a Marker object using Folium for each launching site:
  - Why: to easily locate any launch site on the map
- Created a marker for each launch record:
  - How: with an icon colored green if the outcome was successful and red otherwise
  - Why: to help analyze the success rate for each launching site
- Created multiple line objects from the launching site with the highest success rate:
  - Why: to analyze factors that influence the choice of the location of a launching site
- GitHub URL: click me to see the code

Build a Dashboard with Plotly Dash

- Created a dropdown menu to let the user select different launch sites or select “all” of them.
- Created a pie chart based on the input to the dropdown menu:
  - A pie chart summarizing the success count of each site if the option “all” was selected
  - A pie chart describing the success distribution of a specific site, if one were selected
- Added a slide range for the payload:
  - Why: to be able to select different payload ranges easily
- Plotted a scatter plot with the x axis to be the payload and the y axis to be the launch outcome:
  - Why: Identify how the selected variable payload range is correlated to the mission outcome
- GitHub Link: click me to see the code

Predictive Analysis (Classification)

1) Separated the features and label, then Standardized the features.

2) Splitted the data into training and testing set with a test size of 20%

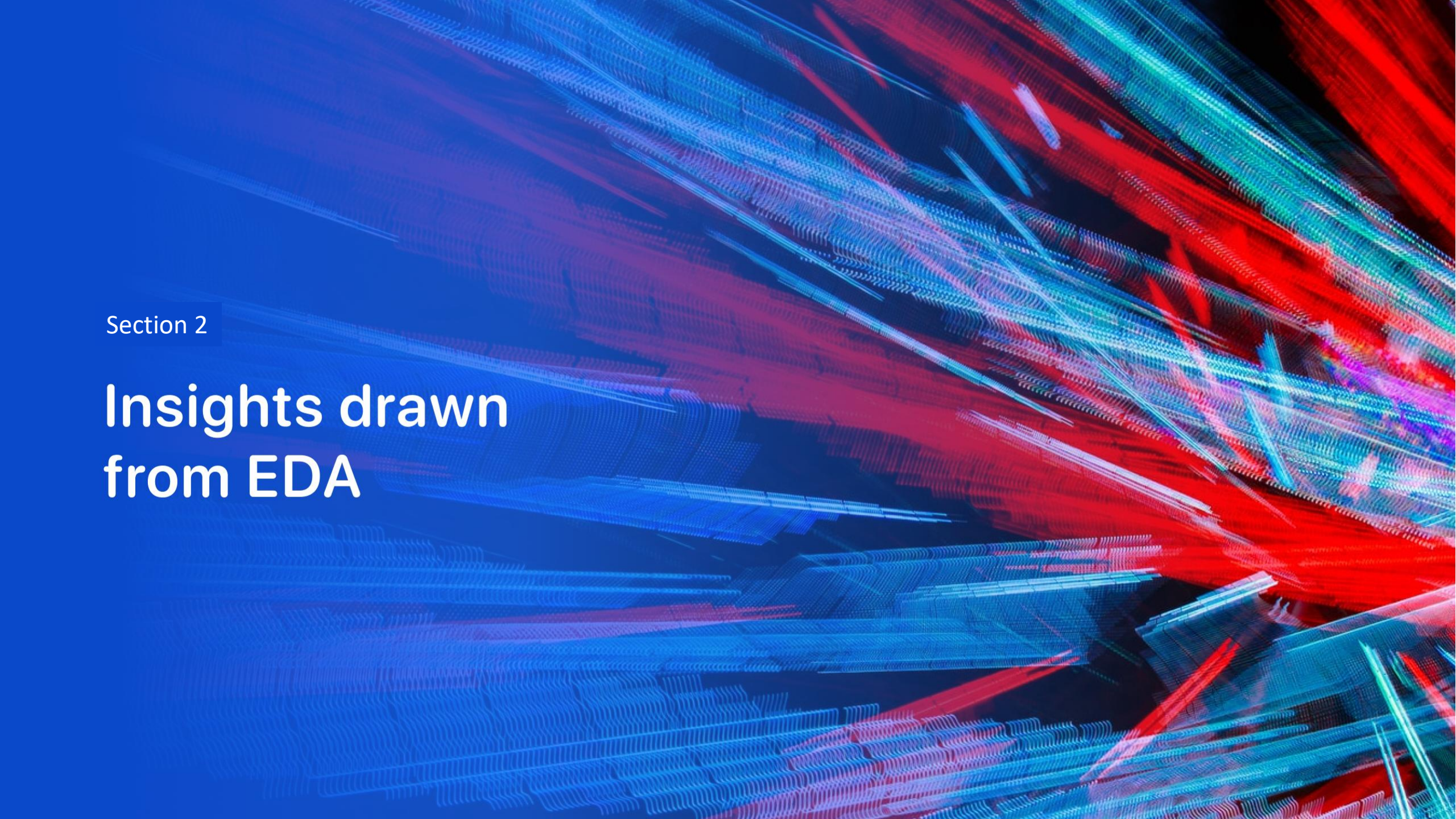
3) Fine tuned multiple ML algorithms on accuracy using a grid search

4) Plotted a bar plot of accuracy vs model to find the best model

GitHub URL: click me to see the code

Results

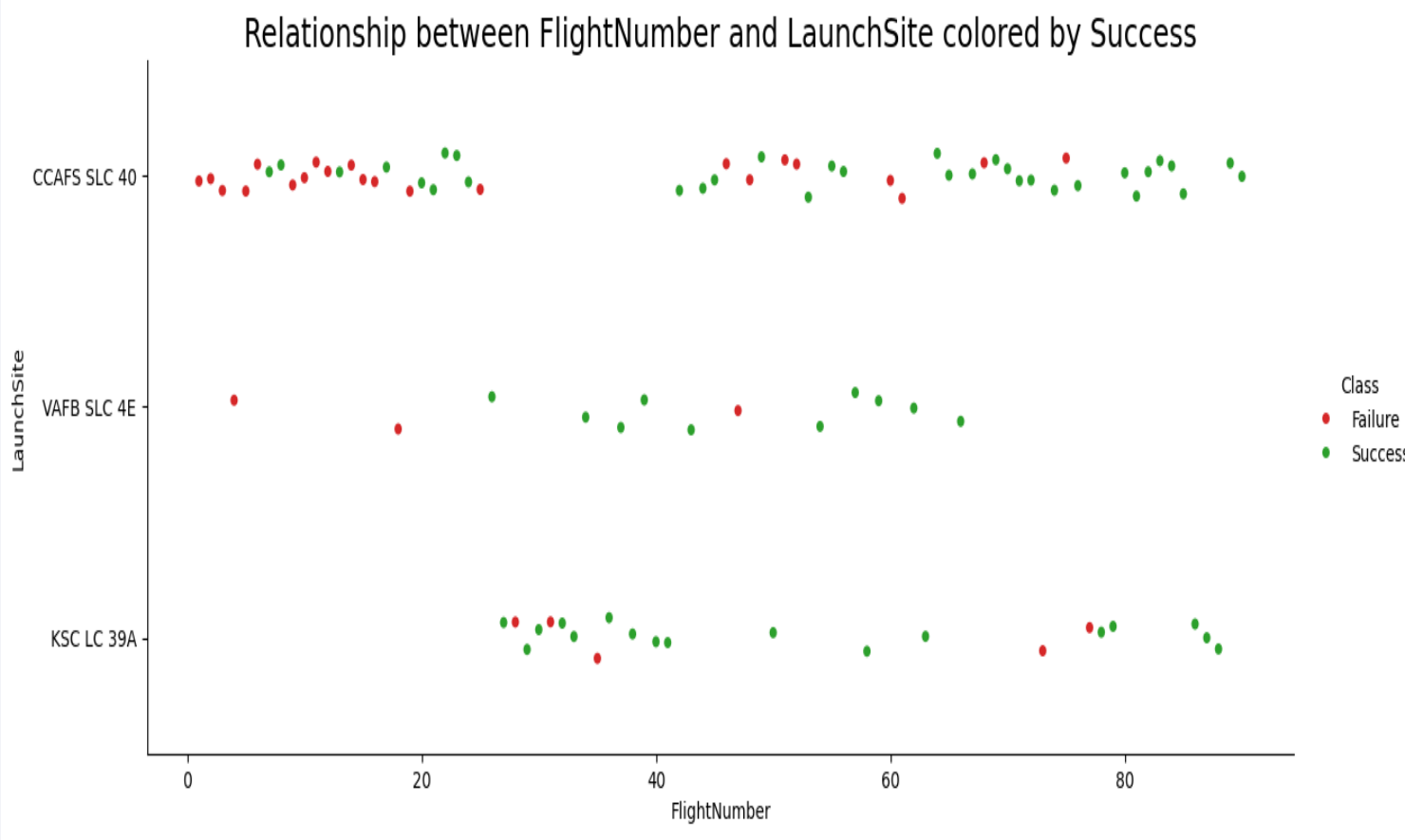
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Observations

- CCAFS SLC 40:
  - Site with the earliest launches
  - Initially shows a high proportion of failure
- KSC LC 39A:
  - Appeared later in the flight number
  - Has the fewest launches and a high success rate
- VAFB SLC 4E:
  - Latest to appear on the flight number
  - Has a high success rate



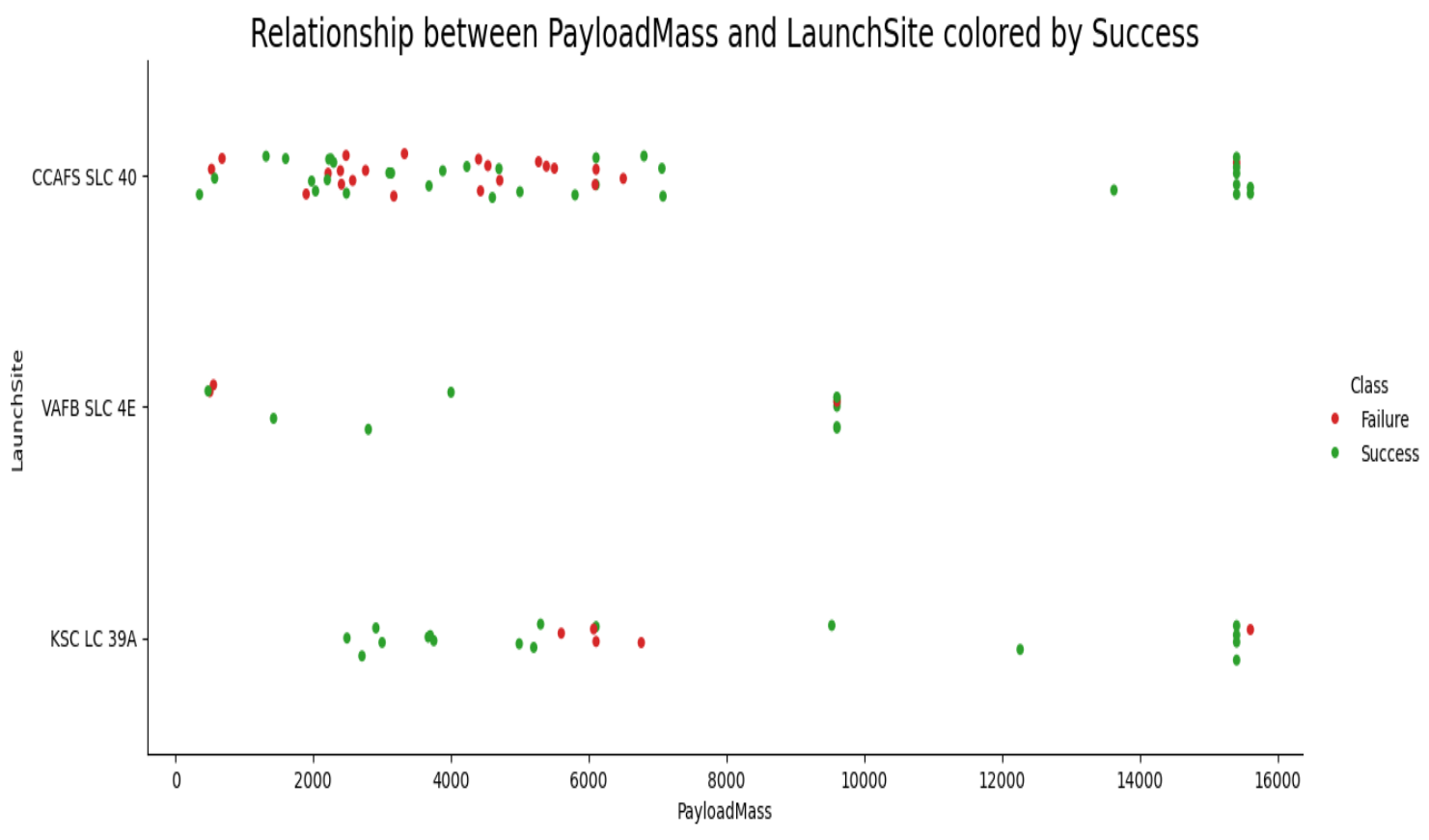
Key Takeaways

- CCAFS SLC 40 was crucial in the early development and testing phase of the Falcon 9's reusability.
- The later introduction of KSC LC 39A and VAFB SLC 4E for launches, coinciding with high success rates

Payload vs. Launch Site

Observations

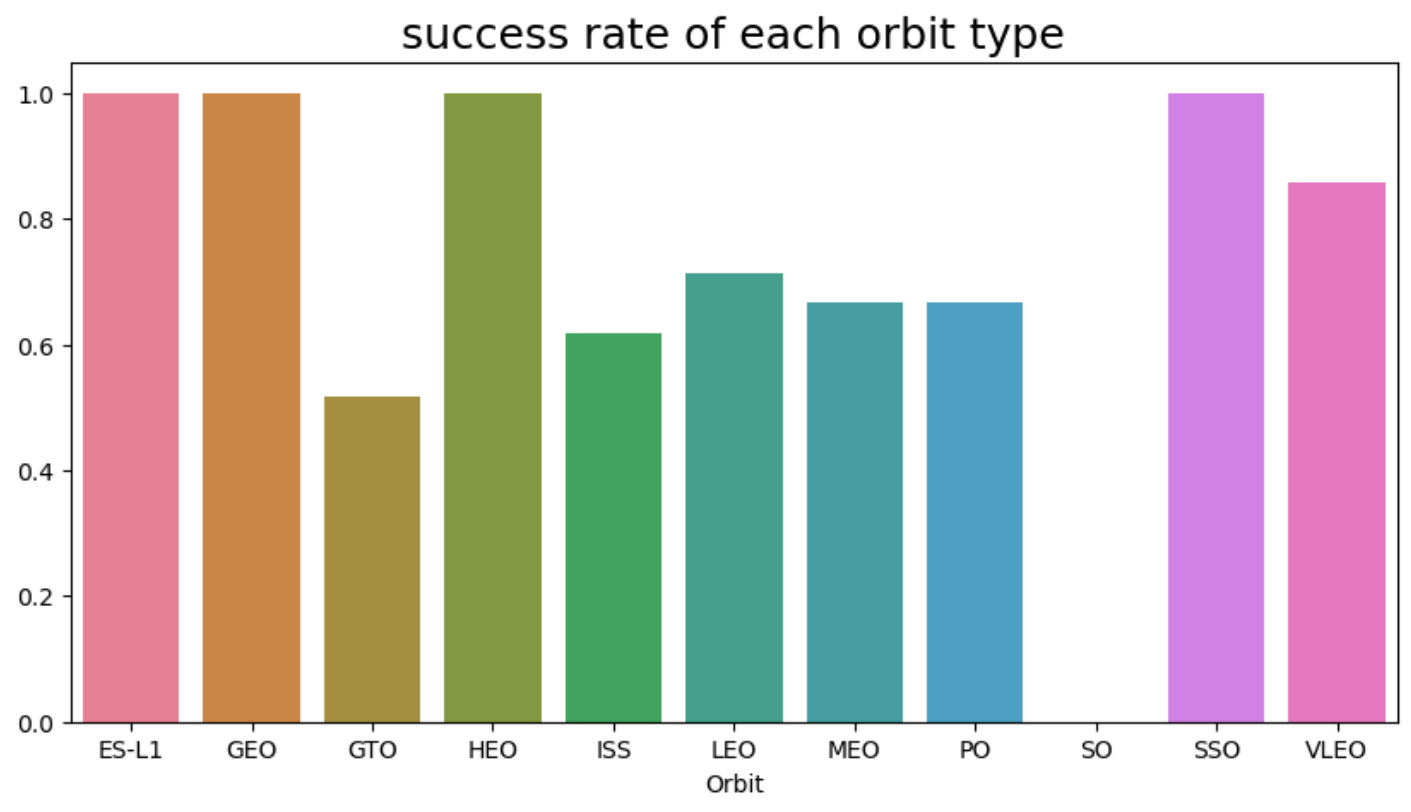
- CCAFS SLC 40:
  - Most launches do not exceed a payload mass of 800 kg
  - Mixed success rate in the lower tail, but very high success rate in the upper tail
- KSC LC 39A:
  - The recovery of stage 1 is almost always successful, regardless of the payload mass
  - Payload masses aren't as high as the two other sites
- VAFB SLC 4E:
  - Very high success rate for low and high payload masses



Key Takeaways

- The majority of payload masses are below 8000 kg across all sites
- The high failure rate at CCAFS SLC 40 reinforces the hypothesis that it was used in the earliest experimentations

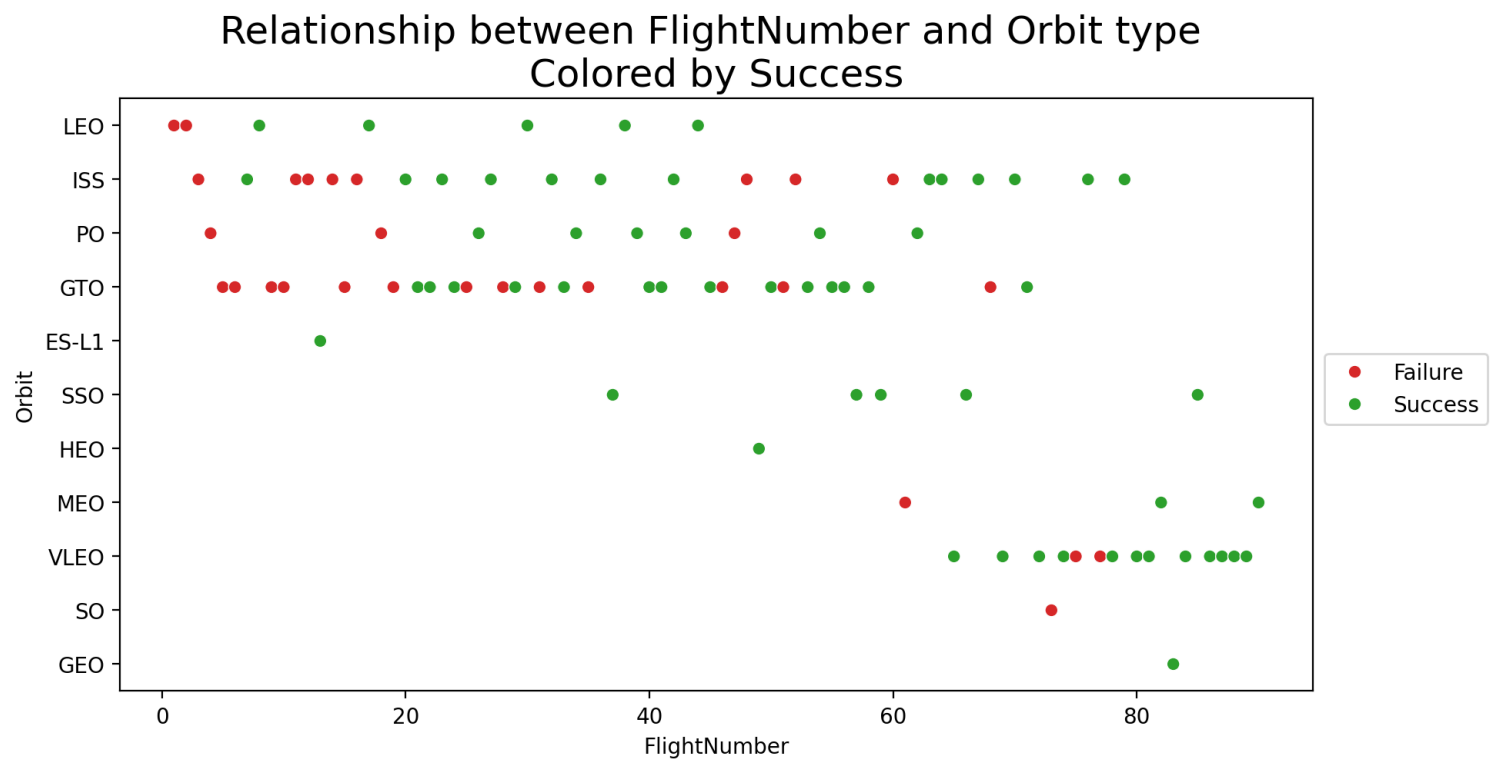
Success Rate vs. Orbit Type



Observations and Key Takeaways

- Almost half of the Orbit types showed a very high success rate, most even achieving a perfect success rate.
- The majority of the remaining orbits have a success rate between 45% and 65%.
- There is one outlier, which has a 0% success rate, corresponding to the orbit ‘SO’
- Warning: For extreme success rates, the interpretation of the graph can be misleading as it doesn’t indicate the number of launches per orbit.

Flight Number vs. Orbit Type



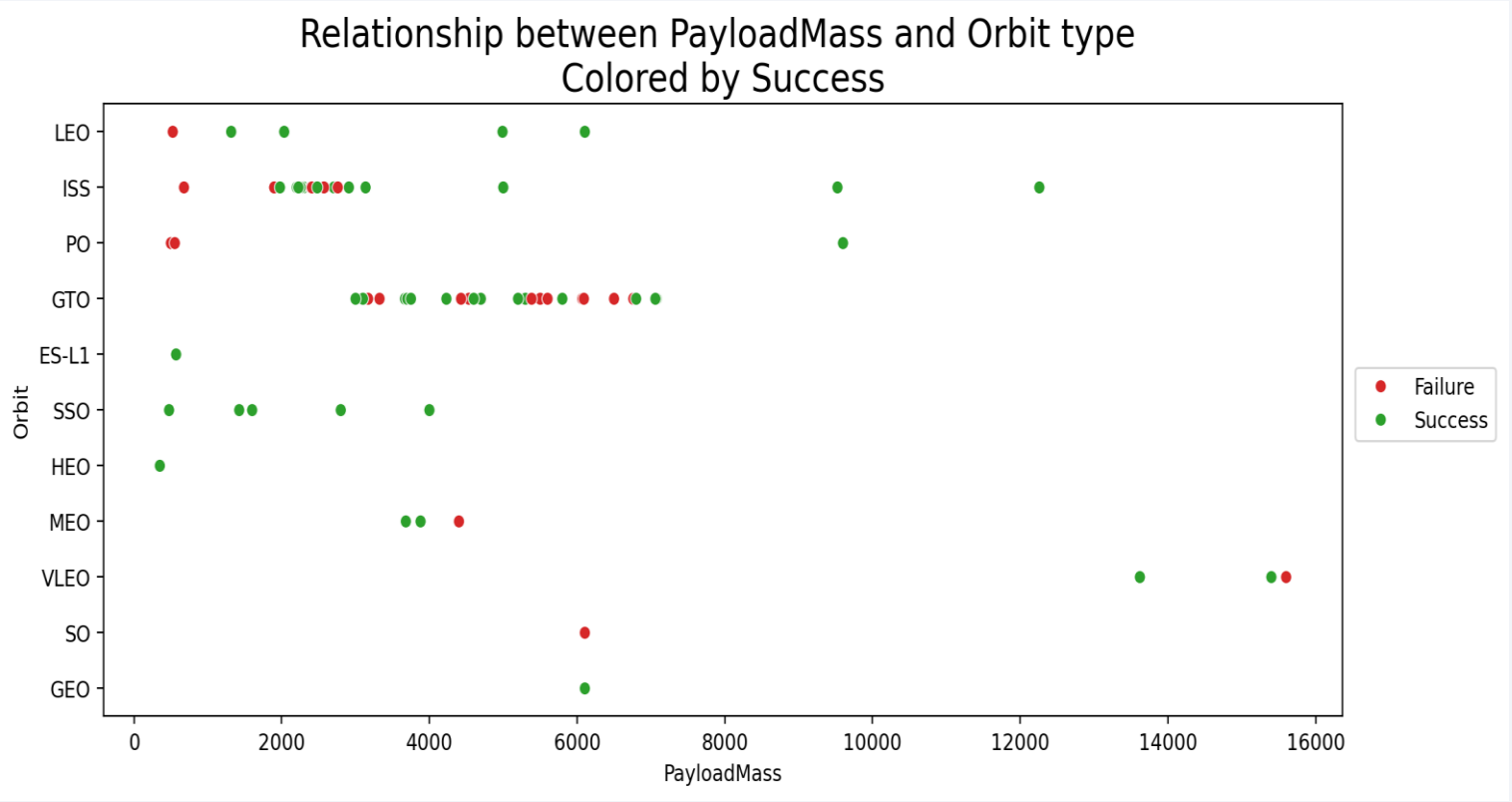
Observations

- The earliest flights primarily targeted LEO, ISS, PO, and the GTO orbit.
- For later flight numbers, it can be observed that the focus has shifted toward reaching new orbits such as VLEO and MEO
- The HEO, SO, GEO orbits have only one attempt

Key Takeaways

- There might be a hierarchical orbit importance that is reinforced by the observed number of attempts per orbit and a focus on some orbits during the earliest flights.
- HEO, SO, and GEO orbits: having only one launching attempt each:
  - their success rate should be interpreted carefully
  - suggest that they aren't a top priority for SpaceX

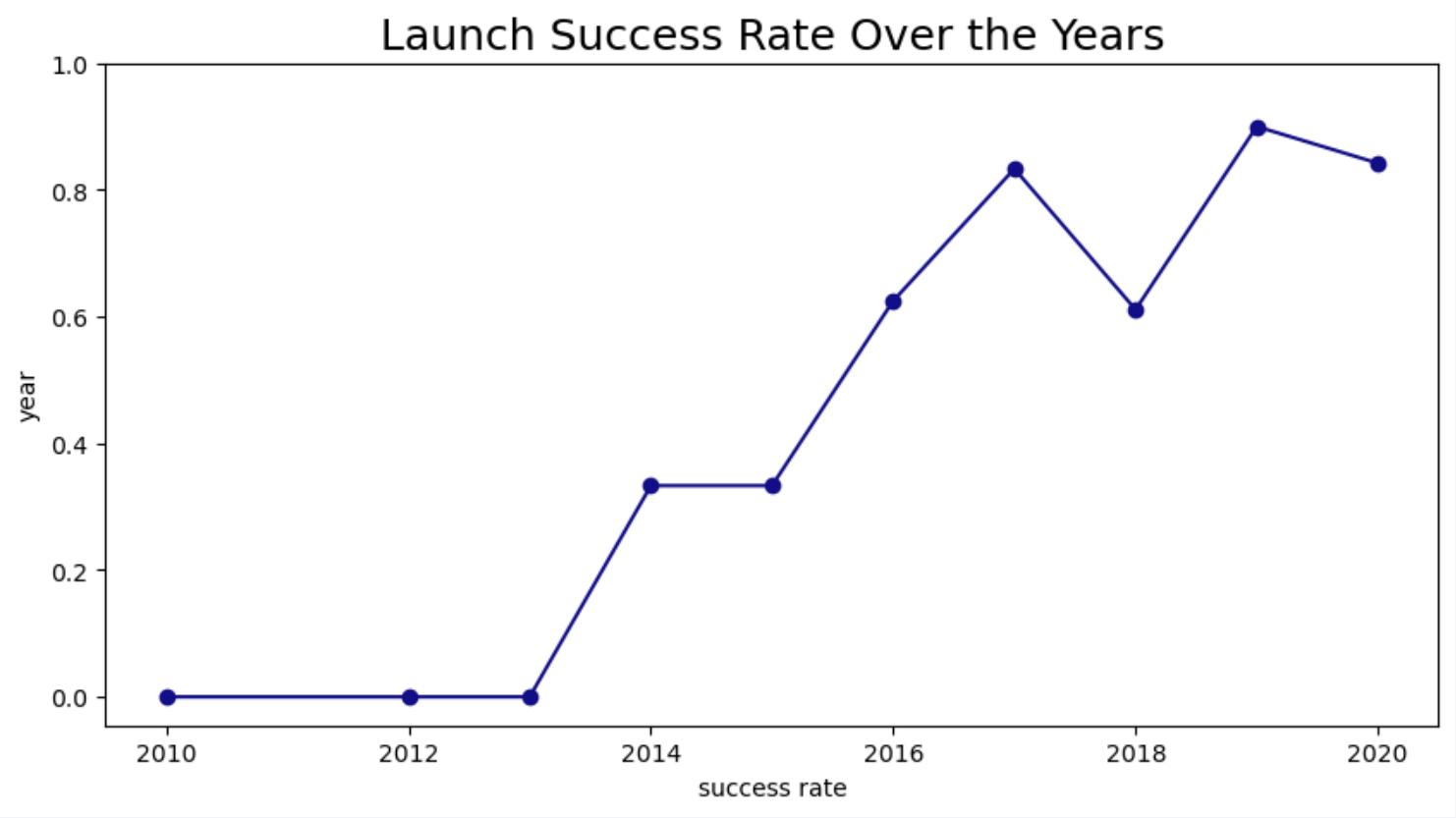
Payload vs. Orbit Type



Observations and Takeaways

- The LEO, ISS, and PO orbits have about the same success rate distribution, characterized by:
  - a successful outcome for larger payload masses
  - Unsuccessful outcome for smaller payload masses
- For the GTO orbit, the range of the payload mass is small (compared to other orbits), and the success distribution is approximately uniform.
- The SSO has a very small payload mass for each launch which all resulted in a successful recovery of the first stage.
- The HEO, MEO, SO, and GEO all have relatively small payload masses, which resulted in a successful outcome for most.
- VLEO is the orbit with the highest mean payload mass per launch and resulted mostly in a successful outcome.

Launch Success Yearly Trend



Observations and Key Takeaways

- From 2010 to 2013, no improvement was observed, and the success rate was 0%.
- In 2014, they significantly improved their success rate to almost 40%
- During 2015, there was almost no improvement in the success rate, which stayed below 40%.
- From 2016 to 2017, the success rate of recovering the first phase increased linearly up to 80%.
- In 2018, the success rate dropped to around 60%
- In 2019, their success rate reached an all time maximum of about 90%
- Finally, in 2020, their success rate dropped to about 85%
- Conclusion: The success rate has been increasing over time, emphasizing the learning curve

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql select distinct("Launch_Site") from SPACEXTBL;
```

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

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

Using SQL, it was found that there are

4 distinct launch sites, which are:

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL 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

Using the “where” and “like” clauses, it was possible to filter the data to retrieve the 5 records having their launch sites name beginning with “”CAA

Total Payload Mass

```
%sql select sum("PAYLOAD_MASS__KG_") from "SPACEXTBL" where "Customer" =  "NASA (CRS)";
```

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

sum("PAYLOAD_MASS__KG_")
45596

Using the built-in function “sum” and the “where” keyword, it was found that

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

is of 45596 Kg

Average Payload Mass by F9 v1.1

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

Using the built-in function “average” and the “where” keyword, it was found

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

is of 2928.4 Kg

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select Date from SPACEXTBL where Landing_Outcome = "Success (ground pad)" order by "Date", "Time (UTC)" limit 1;
```

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

Date
2015-12-22

Using the “where” keyword to filter and the “order” keyword to order the result, it was found that the first successful ground landing date is December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 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 SPACEXTBL
    where "PAYLOAD_MASS__KG_" between 4000 and 6000 and "Landing_Outcome" = "Success (drone ship)";
```

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

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

Taking advantage of the “where” and “between” statements to filter, it was found that the booster used in drone ship having a payload mass between 4000 and 600 are: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%%sql
select "successful mission" as "misssion outcome", count(*) as "count" from SPACEXTBL where Lower("Mission_Outcome") like "success%"
    UNION
select "unsuccessful mission" as "misssion outcome", count(*) as "count" from SPACEXTBL where Lower("Mission_Outcome") not like "success%";
```

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

misssion outcome	count
successful mission	100
unsuccessful mission	1

Taking full advantage of SQL, it was found that 100 missions resulted in a

successful outcome and only one resulted in a failure.

Boosters Carried Maximum Payload

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

```
%%sql

    select distinct(Booster_Version) from SPACEXTBL where "PAYLOAD_MASS__KG_" =
        (select max("PAYLOAD_MASS__KG_") from SPACEXTBL);
```

```
 * 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

Using nested queries, it was found that booster versions, carrying the maximum payload mass, are:

- 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

2015 Launch Records

```
%sql select substr(Date, 6,2), count(*), "Booster_Version", "Launch_Site" from SPACEXTBL
    where substr(Date,0,5)='2015' and Landing_Outcome = "Failure (drone ship)";
```

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

substr(Date, 6,2)	count(*)	Booster_Version	Launch_Site
01	2	F9 v1.1 B1012	CCAFS LC-40

Using an SQL query, it was found that the record having a failure landing outcome in the drone ship during the year 2015 occurred in the moth of January at the CCAFS LC-40 launch site with a booster version F9 v1.1 B1012

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 count from SPACEXTBL
    where "Date" between "2010-06-04" and "2017-03-20" group by "Landing_Outcome" order by count(*) 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

Between 2010-06-04 and 2017-03-20, the different landing outcomes and their respective counts are:

- "*No attempt*" with a count of 10
- "*Success drone ship*" with a count of 5
- "*Failure drone ship*" with a count of 5
- "*Success ground pad*" with a count of 3
- "*Controlled (ocean)*" with a count 3
- "*Uncontrolled (ocean)"* with a count of 2
- "*Failure (parachute)"* with a count of 2
- *"Precluded"* with a count of 1

Section 3

Launch Sites Proximities Analysis

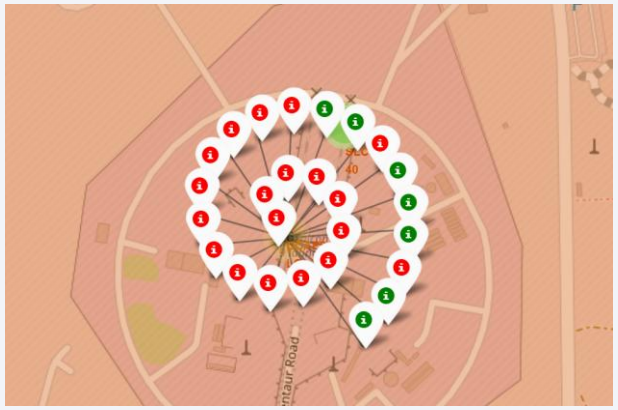
Locations of all launch sites

Observations

- 3 of the 4 launches are located on the East Coast
- The majority of launch sites are located in the state of Florida
- Only one launch site is located in the state of California
- All launch sites are located near an ocean

Success/failed launches for each site on the map

Site: CCAFS LC-40



Observations

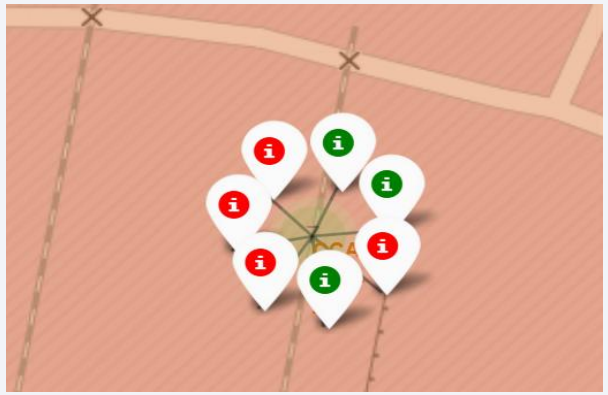
- A lot of launch attempts
- Low success rate

Site: KSC LC-39A

Observations

- A lot of launch attempts
- Highest success rate

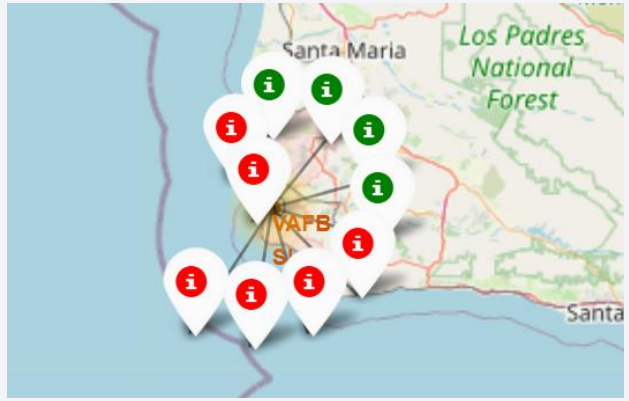
Site: CCAFS SLC-40



Observations

- Very few launch attempts
- Low success rate

Site: VAFB SLC-4E

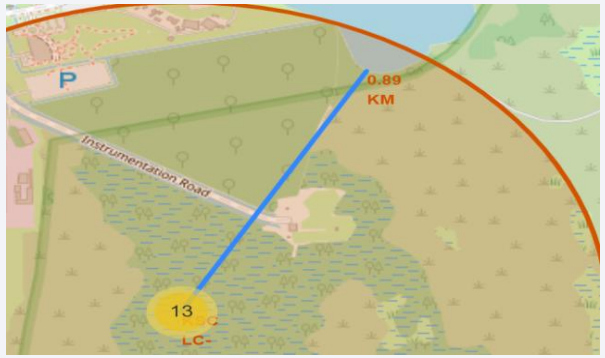


Observations

- Medium number of launch attempts
- Low success rate

Proximities of Site KSC LC-39A

Closets Coastline



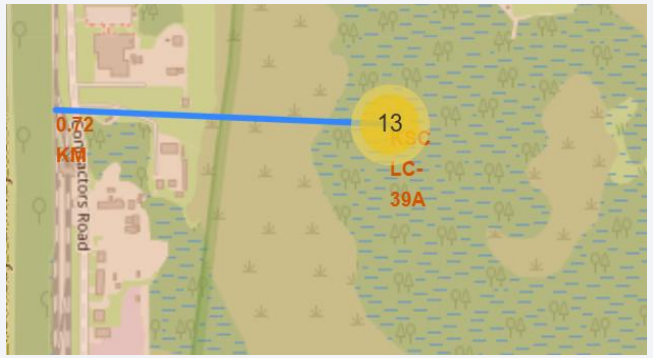
Observations

- The coastline is 0.89 km away from the launch site, which is pretty close

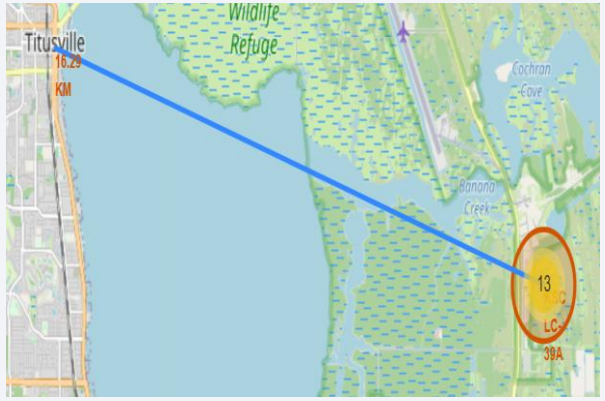
Observations

- The closest railway (Titusville) is located about 0.72 km away from the launching site

Closest Railway



Closest City



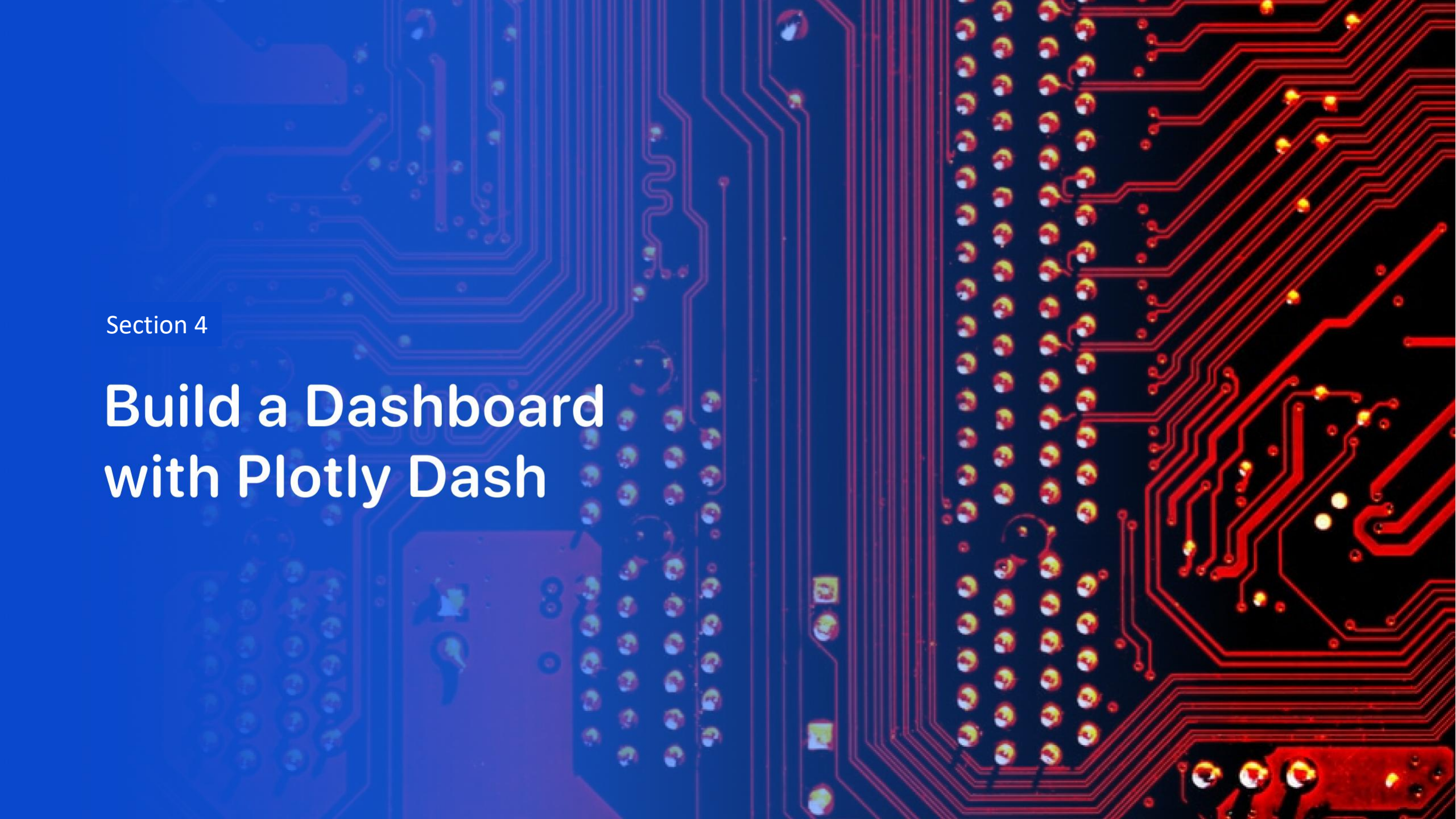
Observations

- The closest city (Titusville) is located about 16.29 km away from the launching site

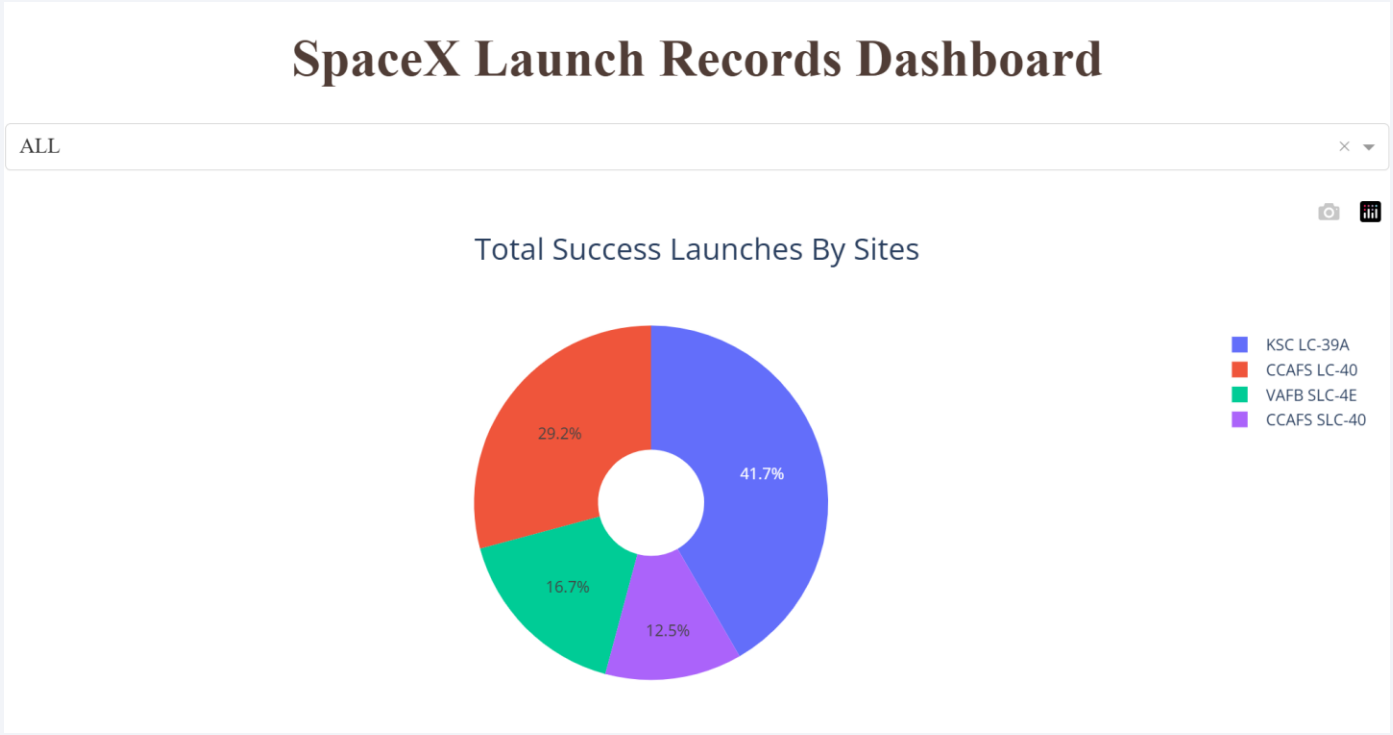
Observations

- The closest highway (contractors Road) is located 0.67 km away.

Closest Highway


Section 4
Build a Dashboard with Plotly Dash

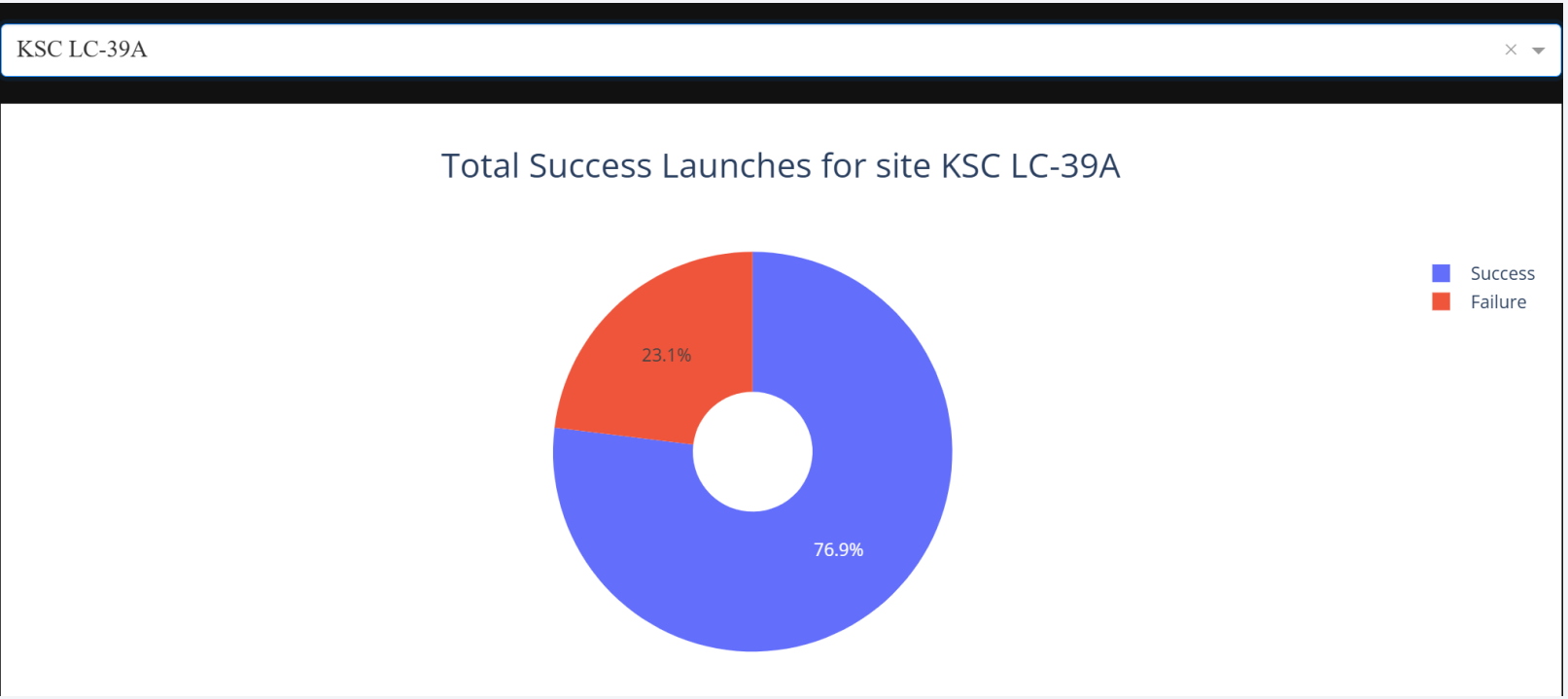
Total Successful launches by Sites



Observations

- KSC LC-39A has the highest total success, corresponding to 41.7% of the total success
- CCAFS LC-40 has the second highest total success, representing about 29.2%
- VAFB SLC-4E comes third, with about 16.7% of total success
- CCAFS SLC-40 comes last with only 12.5% of the total success.

Total Success Launches for the site KSC LC-39A



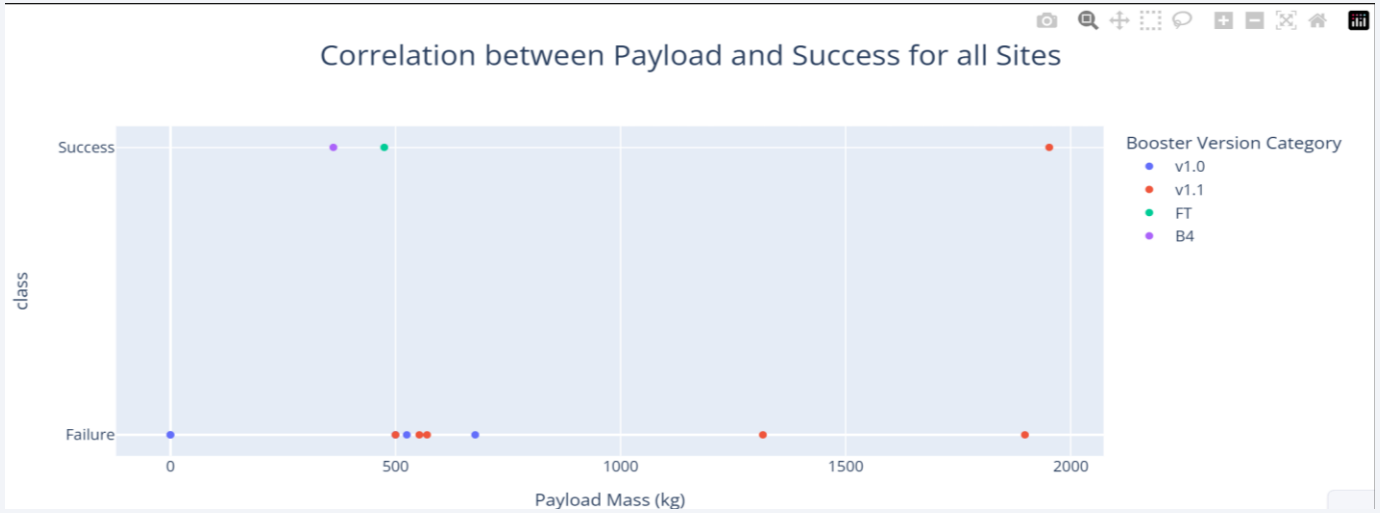
Observations

- KSC LC-39A has about 76.9% of its total launches that resulted in a successful outcome
- KSC LC-39A has about 23.1% of its total launches that resulted in a failure

Payload vs. Launch Outcome with different payload

This scatter plot indicates that:

- The majority of launches with a payload mass above 5000 kg result in a failure
- For a payload greater than 5000 kg, only the booster version categories FT and B4 are used.

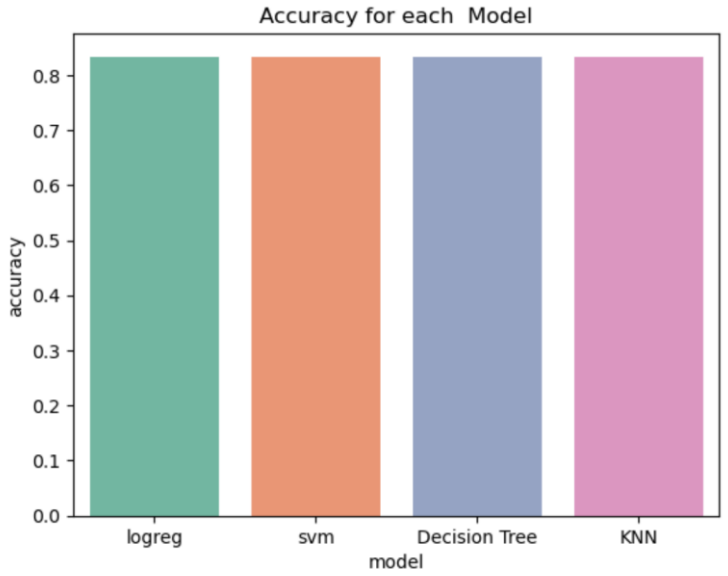


This scatter plot indicates that:

- The majority of launches for with a payload mass above 5000 kg result in a failure
- For a payload greater than 5000 kg, only the booster version categories FT and B4 are used.

Section 5

Predictive Analysis (Classification)

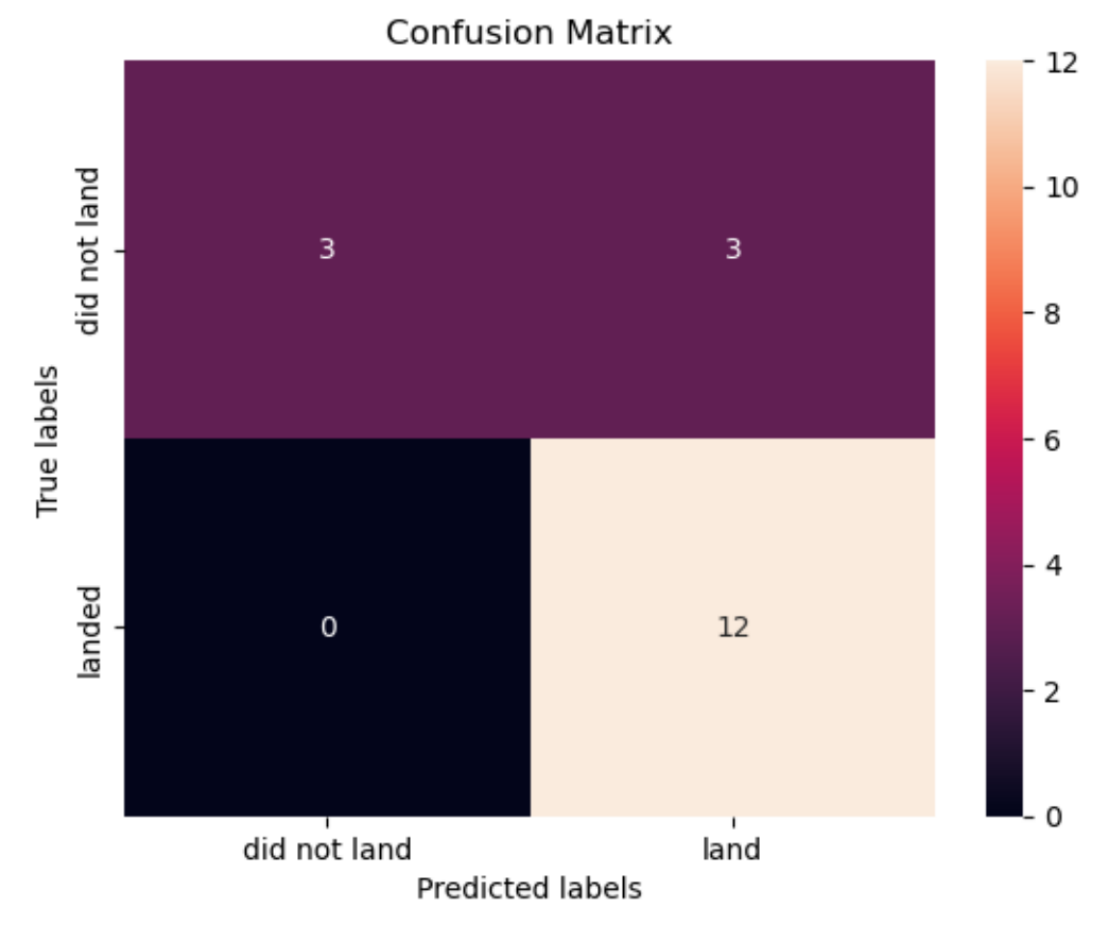


Classification Accuracy

- It was found that a classification accuracy of 0.8333 was about the same for all the machine learning models considered (logistic regression, support vector machine, decision tree, and K-Nearest Neighbors)

- Due to the simplicity of K-Nearest Neighbors, it was arbitrarily chosen as the best model for this project

Confusion Matrix of KNN model



Observations

- All launches that landed were correctly identified
- Half of the launches that didn't land (3 launches) were misclassified as successful landings
- "landed":
  - Recall Score = 100%
  - Precision Score = 80%
- "did not land":
  - Recall Score = 50%
  - Precision Score = 100%

Conclusions

- CCAFS SLC 40 was crucial in the early development and testing phase of the Falcon 9's reusability.
- The majority of payload masses are below 8000 kg across all sites
- The LEO, ISS, PO, and GTO orbits are the most important to SpaceX
- KSC LC-39A is the launch site with the highest success rate
- The launch success rate of SpaceX has been increasing over the years up to 95%
- The proximities across launching sites are similar: they are located near an ocean, a railway, and/or a highway. However, the closest city tends to be far away.
- The simplest model that best predicts the landing outcome was found to be a K Nearest Neighbors (with k=8):
    - The model has an accuracy of approximately 83%
    - The model did a great job at identifying launches with a successful outcome, however didn’t do great at predicting launches that will result in a failure

Appendix

KNN classification report

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.80	1.00	0.89	12
accuracy			0.83	18
macro avg	0.90	0.75	0.78	18
weighted avg	0.87	0.83	0.81	18

Thank you!