Winning Space Race with Data Science

Julian Miranda
24 May 2025

Outline

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

Executive Summary

Summary of Methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an Interactive Map With Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis (Classification

Summary of All Results

- Exploratory Data Analysis Results
- Interactive Analytics Demo in Screenshots
- Predictive Analysis Results

Introduction

Project Background and Context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Problems You Want to Find Answers

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Has the rate of successful landings increased over the years?
- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

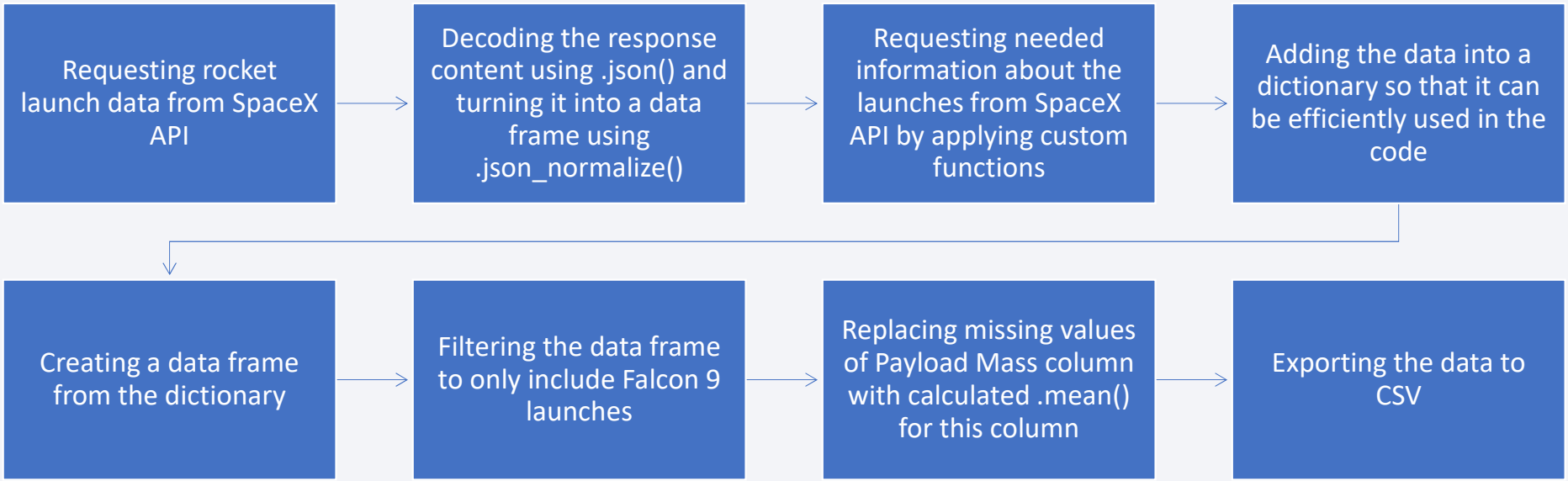
Executive Summary

- Data collection methodology:
  - Using SpaceX REST API
  - Using Web Scraping from Wikipedia
- Perform data wrangling
  - Filtering Data
  - Eliminating nans
  - Using One Hot Encoding to prepare the data for a Binary Classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Developed Machine Learning models with scikit too predict the best results.

Data Collection

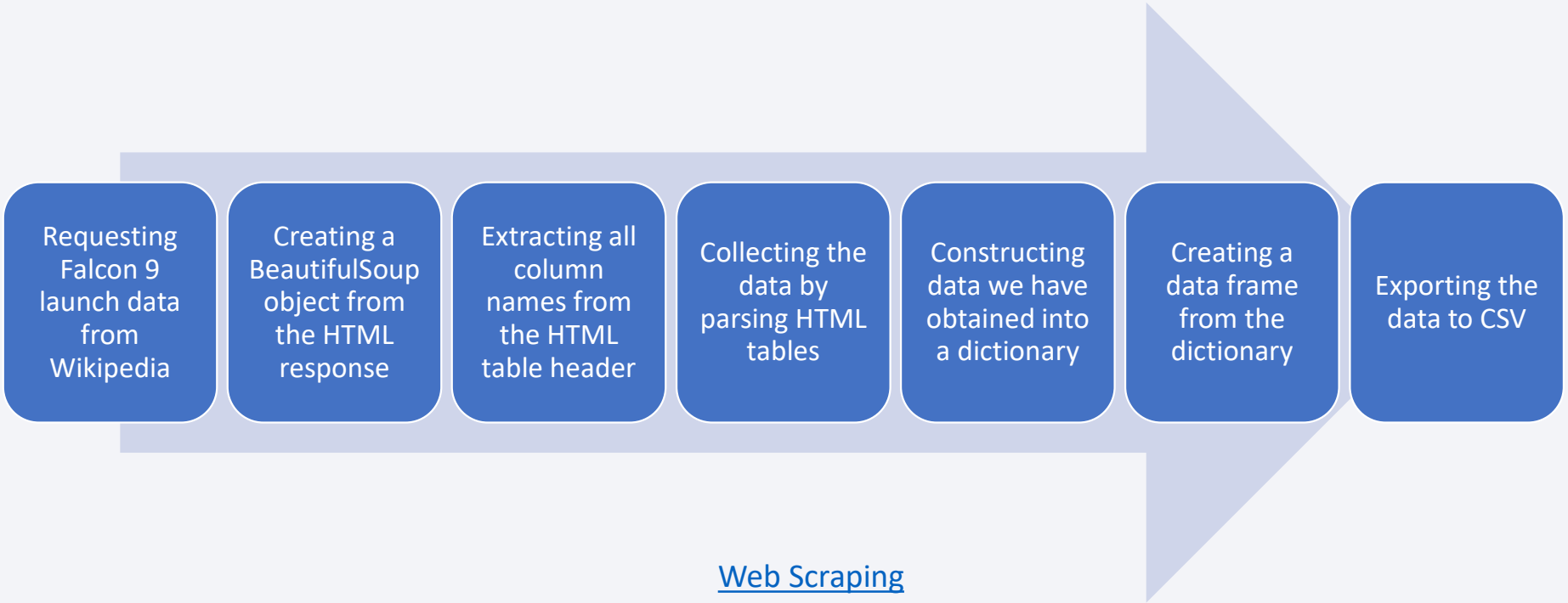
- Collected using a combination of API requests from SpaceX REST API and Web Scraping data from Wikipedia.
- Both methods were used in order to get complete information about the launches.
- **Data Columns Obtained from Wikipedia**
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time
- **Data Columns Obtained from SpaceX REST API**
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Collection – SpaceX API



API's

Data Collection - Scraping



Web Scraping

Data Wrangling

- In the Data Set, I used the values “04” and 1” to represent a successful or an unsuccessful landing. 1=Fail, 0=Success.
- There were six possible outcomes for the landing:
  - True RTLS: Successful landing on a ground pad.
  - False RTLS: Did not successfully land on a ground pad.
  - False Ocean: Did not successfully end up in a specific region of the Ocean
  - True Ocean: Unsuccessfully ended up in the specified region of the Ocean
  - True ASDS: Successfully landed on a drone ship
  - False ASDS: Unsuccessfully landed on a drone ship.

EDA with Data Visualization

- Scatter plots were used to show the relationship between variables.
- Bar charts were used to show comparisons among discrete categories.
- Line charts were used to show trends in data over time.
- Charts were plotted for the following:
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type, and Success Rate Yearly Trend
- EDA with Data Visualization

EDA with SQL

- I used SQL queries to execute several commands that could potentially give my project valuable insight. For example, I discovered that there have only been four successful drone ship landings, with the weights between 4,000 and 6,000 lbs.
- I was also able to discover the total payload mass for NASA and the average payload mass per booster.
- EDA with SQL

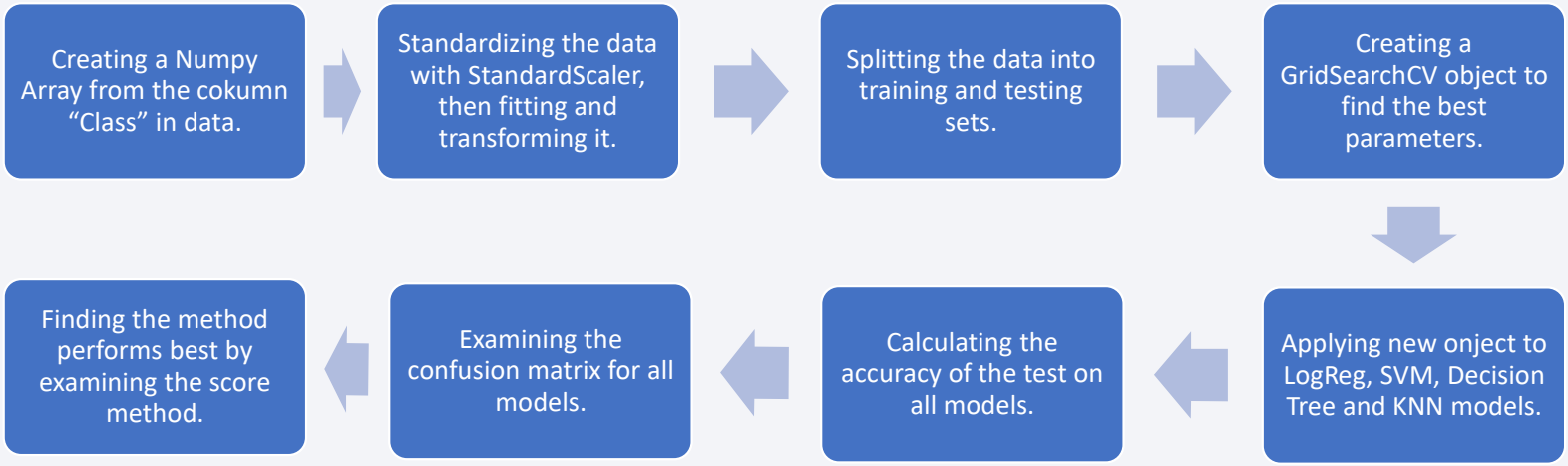
Build an Interactive Map with Folium

- When building the Folium map, the first thing I did was make sure that I had plotted all launch sites. I did this by using circle markers and pop-up labels.
- When the map is launched, the default location is set to the NASA Johnson Space Center using its latitude and longitude coordinates as its start location.
- When you zoom in close to a launch site, you will see either green or red markers to indicate either the success or failure of a rocket launch.
- Folium Map

Build a Dashboard with Plotly Dash

- My Plotly Dash application has three different components to it,
- Pie chart, which shows the landing success rate.
  - You can choose to see the overall success rate or the success rate at a particular site
- Payload Range Slider
  - Manually input the range of the data you would like to see.
- Scatter Chart of Correlation Between Payload and Success for All Sites.
  - Using the slider will directly affect what you see here.
  - The three listed Boosters: v1.1, FT, B4.
- Plotly Dash

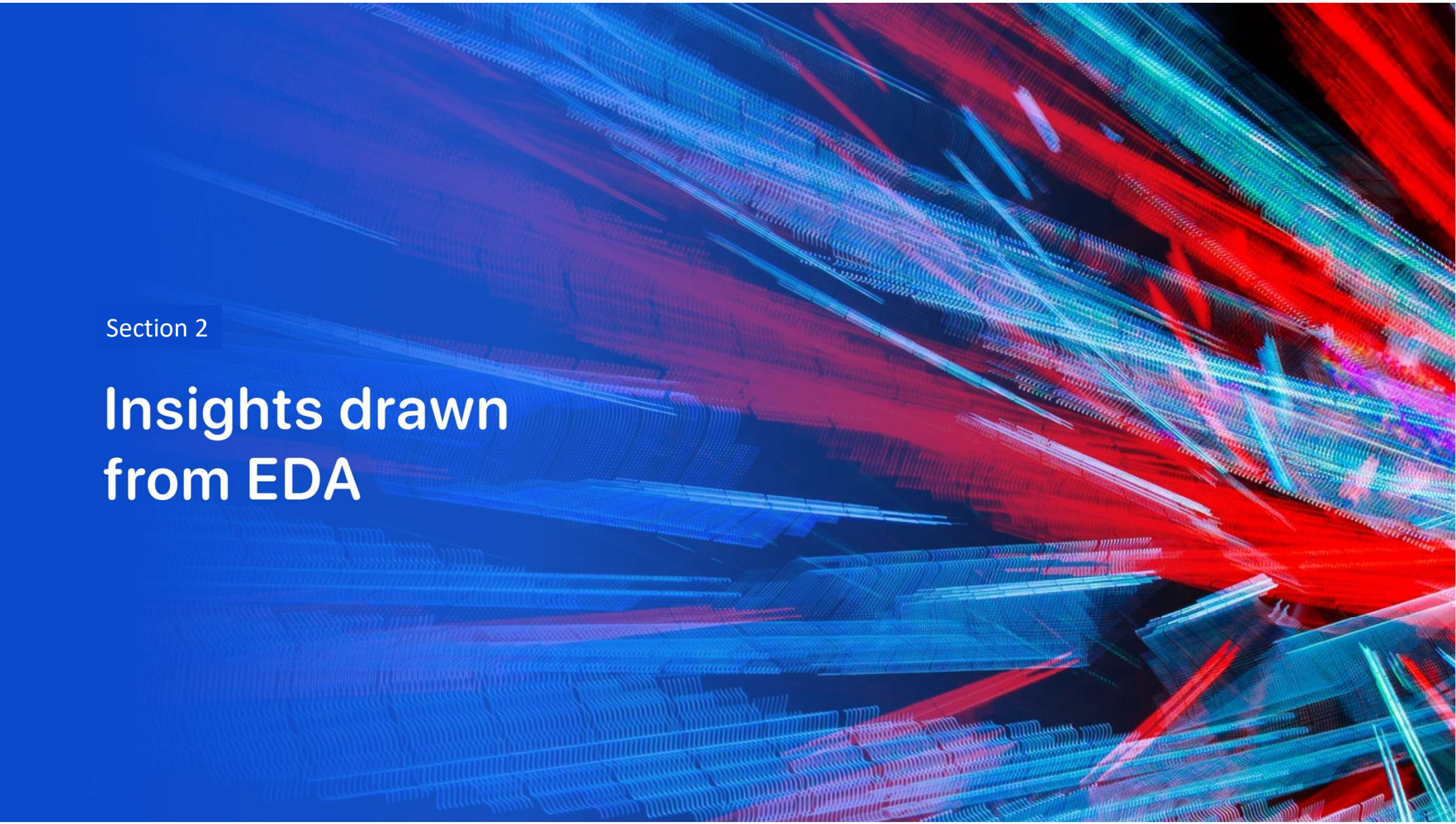
Predictive Analysis (Classification)



Predictive Analysis

Results

- EDA
  - 2019 saw the most successful year for Falcon 9 Launches, with a success rate of almost 95%.
  - After about 60 Launches, SpaceX began to focus on sending its rockets to VLEO.
  - Most rockets sent to VLEO had a payload of almost 16,000 lbs.
- Predictive analysis results
  - **The best-performing model on the test data was the Decision Tree, with a score of 88% accuracy.**
  - Logistic Regression Test Accuracy: 83%.
  - SVM Test Accuracy: 83%.
  - KNN Test Accuracy: 83%



Section 2

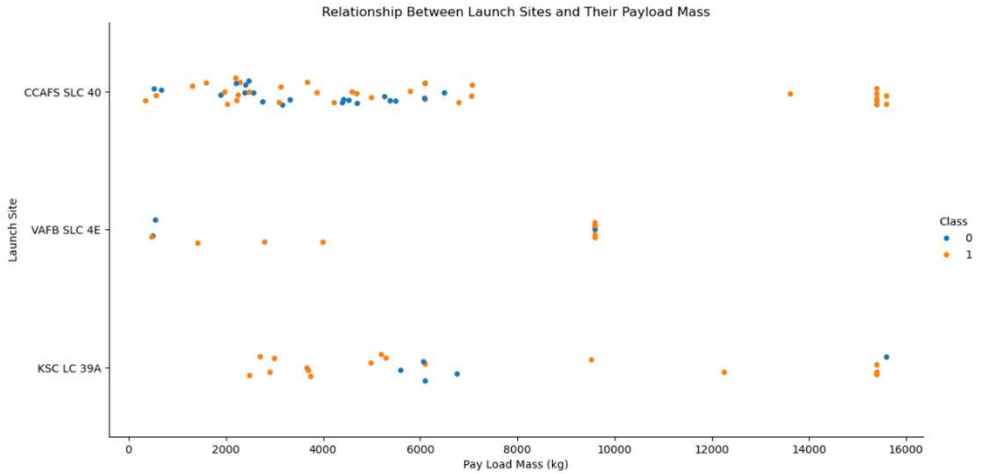
Insights drawn from EDA

Flight Number vs. Launch Site

- We can see that the later launches mostly occurred from Cape Canaveral, meaning that Vandenberg was most likely used for testing, and Cape Canaveral was used for commercial rocket launches.
- While success was abundant at the start of the Falcon 9 journey, SpaceX is probably trying new things, which is why there have been later launches that have not had as much success..

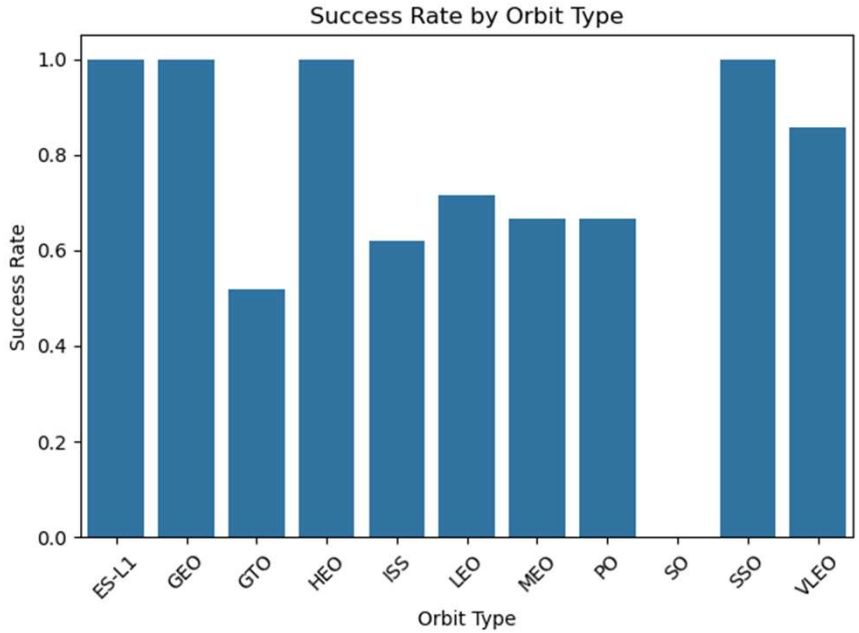
Payload vs. Launch Site

- We once again see that Vandenberg Space Launch Complex did not take a payload greater than 10,000 lbs, and most succesful launches with light payloads came from Cape Canaveral.
- Therefore, one could make the argument that Falcon 9 Rocket Launches began research at Cape Canaveral, underwent research and testing at Vandenberg, and were then redeployed to Florida for experimentation with heavier payloads.



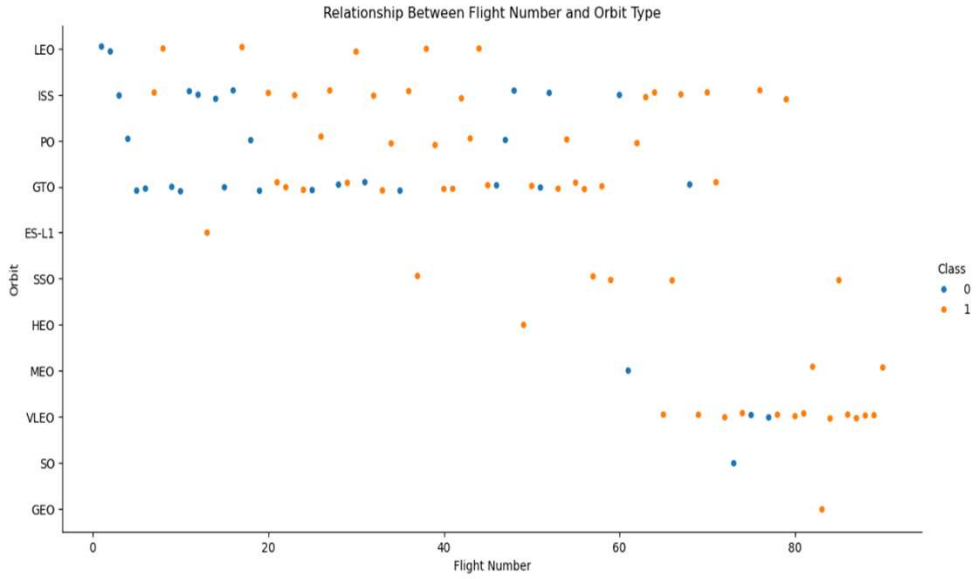
Success Rate vs. Orbit Type

- Most launches were intended to go to Geosynchronous orbit (GTO) or the International Space Station (ISS).
- The Quantity of each Orbit attempt:
  - GTO: 27
  - ISS: 21
  - VLEO: 14
  - PO: 9
  - LEO: 7
  - SSO: 5
  - MEO: 3
  - ES-L1: 1
  - HEO: 1
  - SO: 1
  - GEO: 1



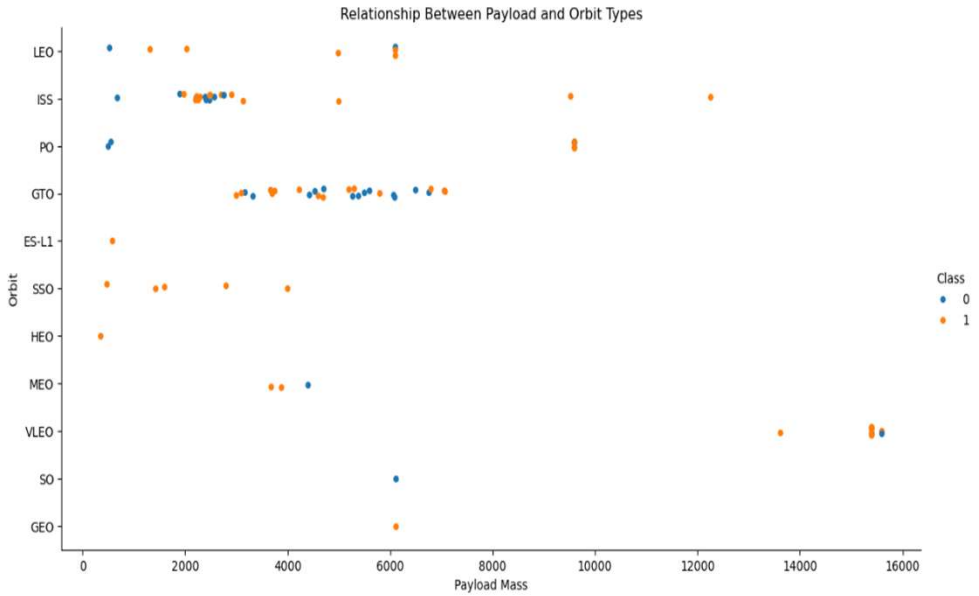
Flight Number vs. Orbit Type

- After about 60 launches, SpaceX focuses on sending its rockets to VLEO but isn't very successful.
- It is in the best interest of SpaceX to become better at this because satellites can efficiently operate in VLEO, and the main business model behind SpaceX is being able to put satellites in space on behalf of companies.



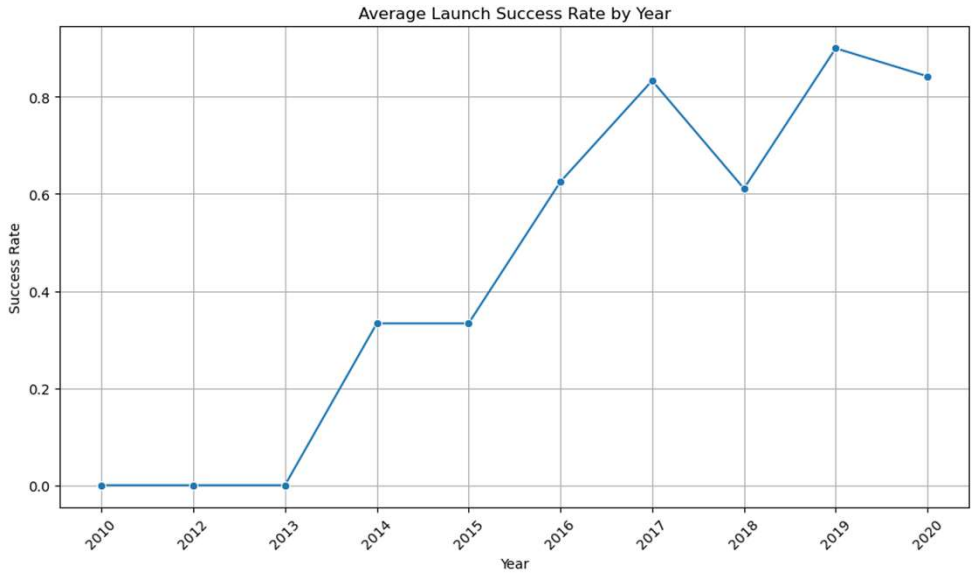
Payload vs. Orbit Type

- We can once again see that the heaviest payloads attempted were sent to VLEO, which serves as good evidence for the point made before.
- We can also see that many rockets were either sent to the ISS or GTO. These missions had fewer failures, probably because the cargo being taken to those places is lighter and therefore makes for a higher chance of success.



Launch Success Yearly Trend

- After 2013, rocket launches became increasingly more reliable, which led to them being commercialized.
- Success rate peaked in 2019 and took a small dip in 2020, likely due to the outbreak of COVID-19 limiting the companie's abilities.



All Launch Site Names

- %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL; → The SQL query used to determine the launch sites and their names, of which there were four.
- CCAFS LC-40 (Cape Canaveral Launch Complex 40).
- VAFB SLC-4E (Vandenberg Space Launch Complex 4).
- KSC LC-39A (Kennedy Space Center Launch Complex 39A).
- CCAFS SLC-40 (Cape Canaveral Space Launch Complex 40).

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

Launch Site Names Begin with 'CCA'

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

→ **This query retrieves 5 launch records where the launch site begins with “CCA.”**

Total Payload Mass

%%sql
SELECT SUM("Payload_Mass__kg_") AS Total_Payload_Mass
FROM SPACEXTBLWHERE "Customer" = 'NASA (CRS)'; → **This query retrieves the total payload mass carried by boosters launched by NASA (CRS).**

- The total payload mass is 45,496 lbs.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG("Payload_Mass__kg_") AS Avg_Payload_Mass
FROM SPACEXTBL
WHERE "Booster_Version" = 'F9 v1.1';
```

→ **This query displays the average payload mass carried by booster version F9 v1.1**

- The average payload mass for this booster is 2923.4 lbs.

First Successful Ground Landing Date

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

→ **This query retrieves when the first successful landing outcome in the ground pad was achieved.**

- It was on December 22nd, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE "Landing_Outcome" = 'Success (drone ship)'
   AND Payload_Mass__kg_ > 4000
   AND Payload_Mass__kg_ < 6000;
```

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

- This query retrieves the names of the boosters that have been successful in drone ship and have a payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

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

- This query lists the total number of successful and failed mission outcomes.

Cleaned_Outcome	Total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE Payload_Mass__kg_ = (
   SELECT MAX(Payload_Mass__kg_)
   FROM SPACEXTBL);
```

- This query lists all booster versions that have carried the maximum payload mass using a subquery.

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

2015 Launch Records

```
%%sql
SELECT
substr(Date, 6, 2) AS Month,
   "Landing_Outcome",
   "Booster_Version",
   "Launch_Site”
FROM SPACEXTBL
WHERE "Landing_Outcome" LIKE 'Failure (drone ship)’
AND substr(Date, 1, 4) = '2015’;
```

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

- This query lists the records that will display the month, landing outcome, booster version, and launch site for launches that happened in 2015.

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

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

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

- This query ranks the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the data 2010-05-04 and 2017-03-20, in descending order.

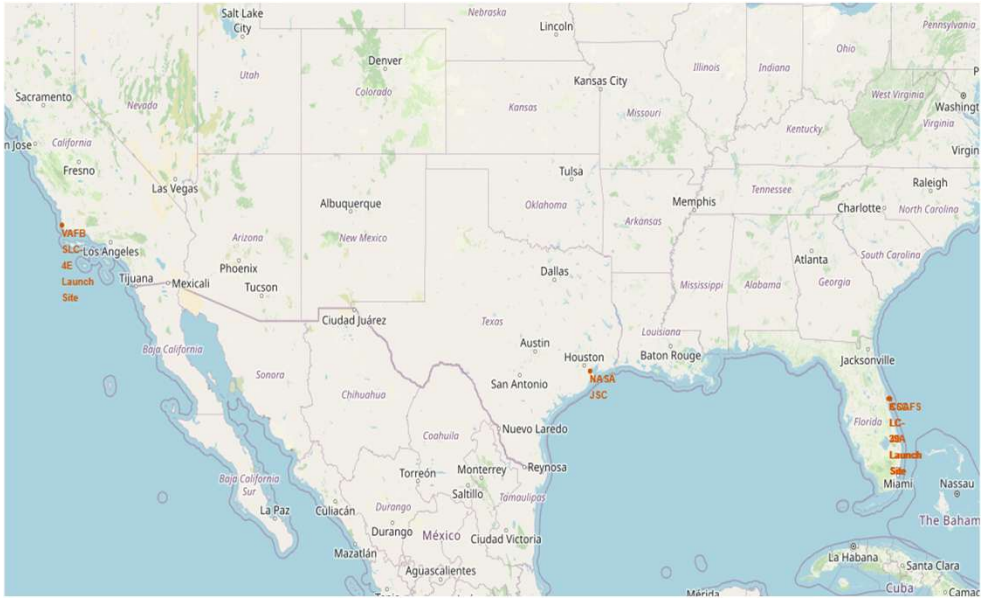
Section 3

Launch Sites Proximities Analysis

Launch Sites and Johnson Space Cetner

- Most rocket launch platforms are located near the equator because anything near the equator is already moving faster than any other place on Earth
  - If you are at the equator, you are already moving at 1670 km/hour.

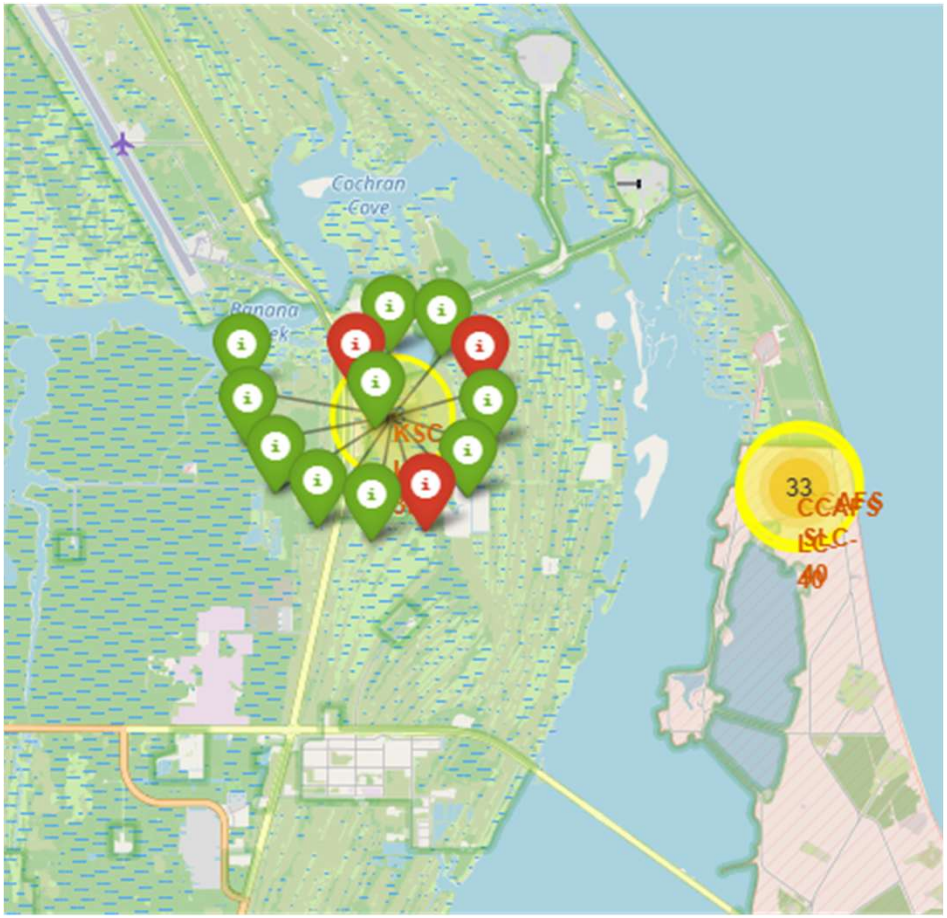
- Launch Sites are located near the coast to mitigate the risk of debris dropping or exploding near people.



Interactive Interface on Map

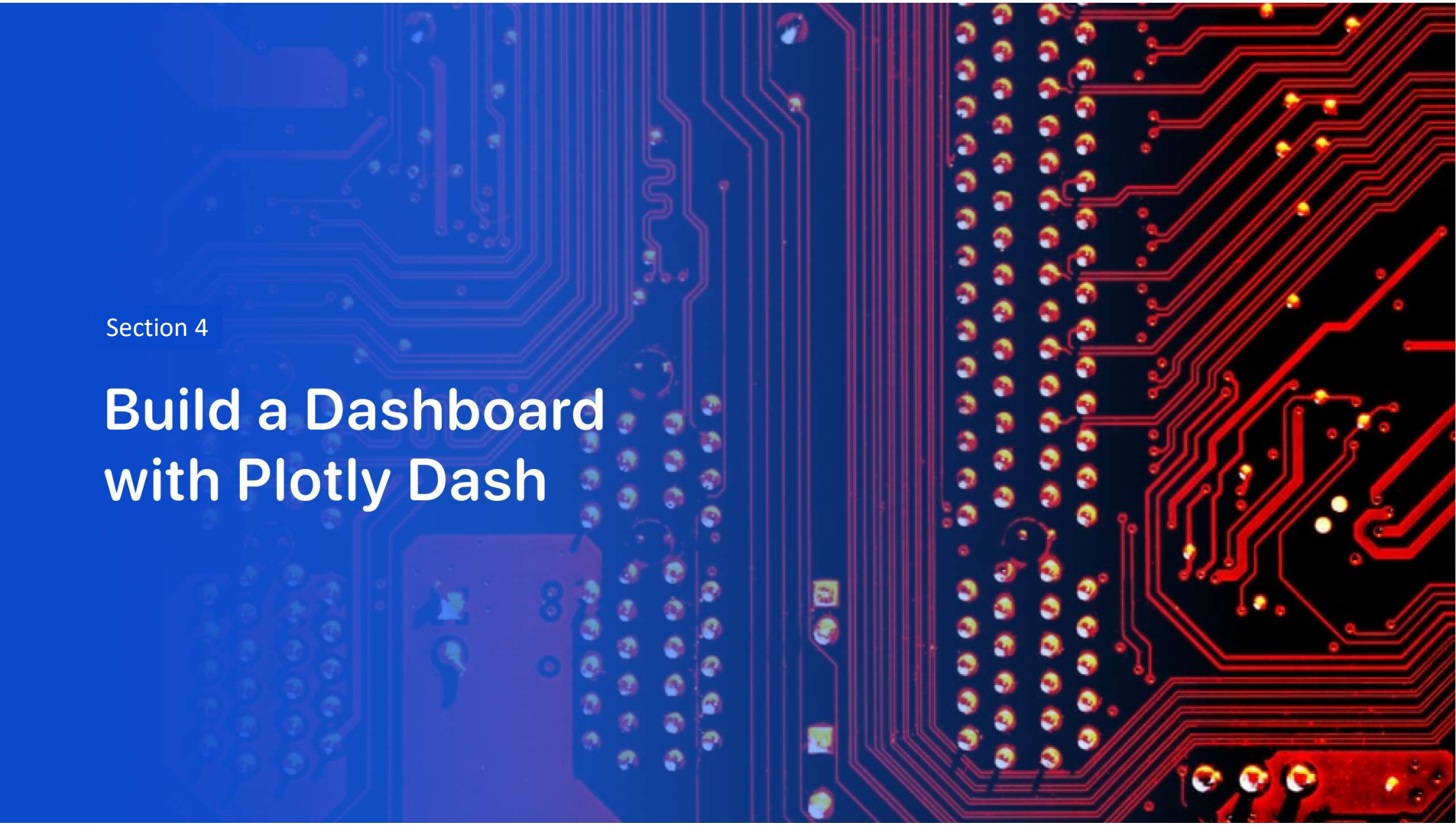
- The map gives the user the ability to see the launch outcome from any launch site simply by clicking on the launch pad. As seen here at Kennedy Space Center, there are only three unsuccessful launches.

- Of the 33 launches from Cape Canaveral, only ten were successful. However, it should be noted that the inclusion of a dedicated Falcon 9 rocket launch pad most likely indicates that many tests were done there.



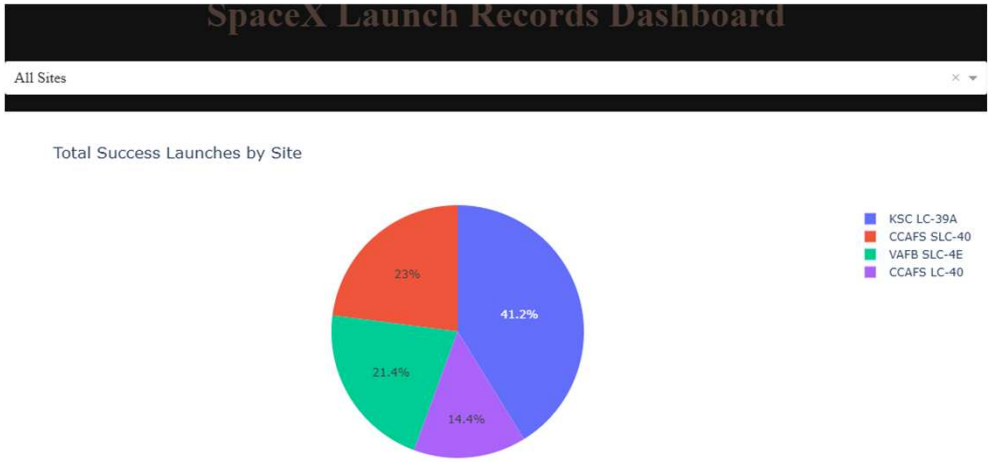
Distance From Launch Pad to Highway

- The map shows the distance from the launch pads to local landmarks, for example, the Cape Canaveral Space Launch Center is 0.58 km away from the nearest highway.
- The Kennedy Space Center is 16.32 km away from the nearest town, Titusville.



Section 4

Build a Dashboard with Plotly Dash



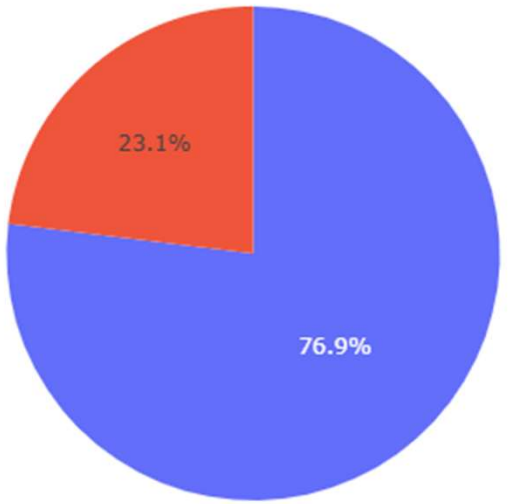
Launch Success for All Sites

- The Dash application shows the total successful launches by site.
- The user can either see all successful launches or pick which one they want to see

Best Launch Site?

- Kennedy Space Center Launch Center 39 has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.
- The runner-up was Cape Canaveral Launch Complex 40, with a success rate of 73.1%.
- VAFB SLC-4E: 60%.
- CCAFS SLC-40: 57.1%.

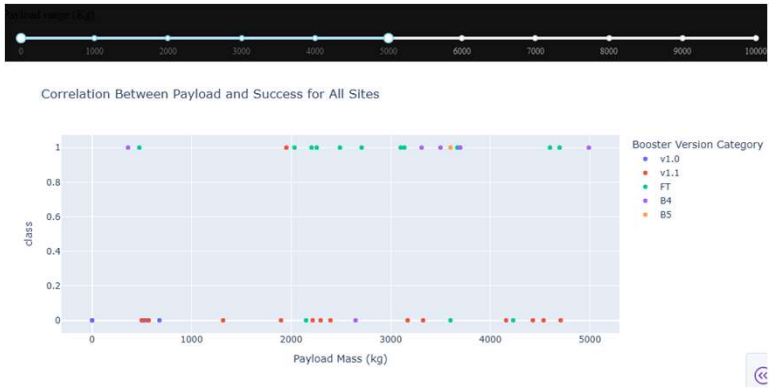
aunches for Site KSC LC-39A



Payload vs. Success Rate

- We can see from the chart that the booster success spread with rockets that have a payload between 0-5000lbs is more even

- Only the FT and B4 boosters were used from a range of 5000-10000lbs, and since these missions were likely more crucial, the failure rate is much lower.



Section 5

Predictive Analysis (Classification)

```
logreg_cv.score(X_test, Y_test)
svm_cv.score(X_test, Y_test)
tree_cv.score(X_test, Y_test)
knn_cv.score(X_test, Y_test)

print("Logistic Regression Test Accuracy:", logreg_cv.score(X_test, Y_test))
print("SVM Test Accuracy:", svm_cv.score(X_test, Y_test))
print("Decision Tree Test Accuracy:", tree_cv.score(X_test, Y_test))
print("KNN Test Accuracy:", knn_cv.score(X_test, Y_test))

best_model = "Decision Tree"

print(f"The best performing model on the test data is: {best_model}, with a score of {tree_cv.score(X_test, Y_test)}.")
```

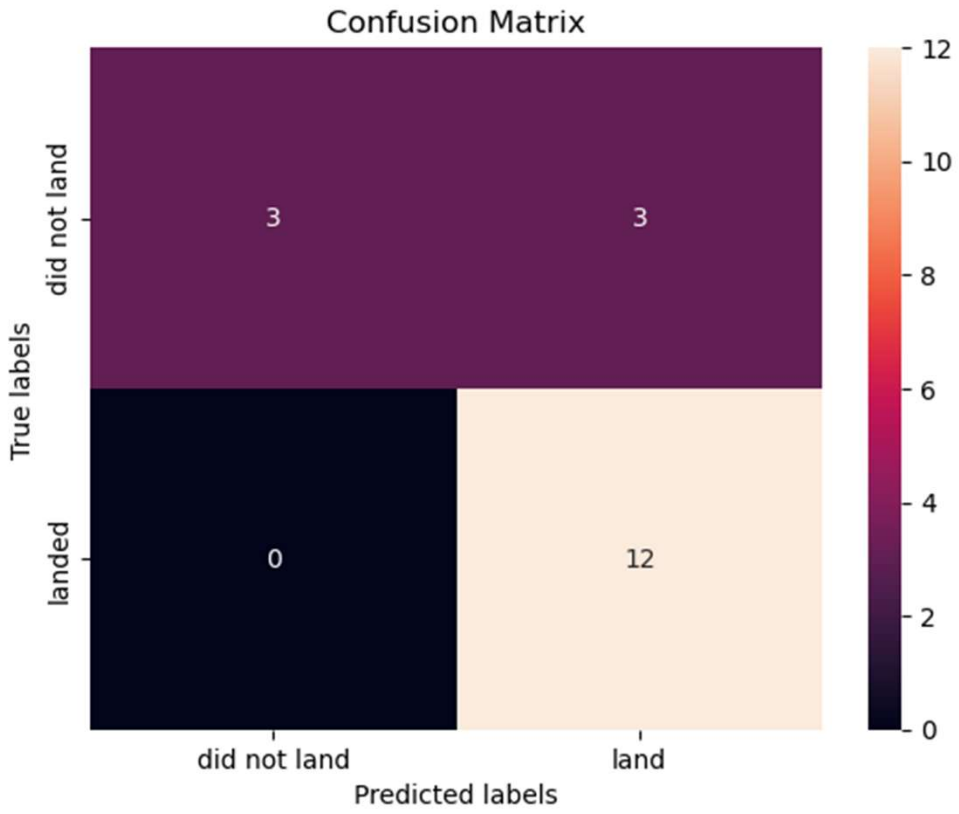
```
Logistic Regression Test Accuracy: 0.8333333333333334
SVM Test Accuracy: 0.8333333333333334
Decision Tree Test Accuracy: 0.8888888888888888
KNN Test Accuracy: 0.8333333333333334
The best performing model on the test data is: Decision Tree, with a score of 0.8888888888888888.
```

Classification Accuracy

- **Logistic Regression Test Accuracy:** 0.8333333333333334
- **SVM Test Accuracy:** 0.8333333333333334
- **Decision Tree Test Accuracy:** 0.8888888888888888
- **KNN Test Accuracy:** 0.8333333333333334
- **The best performing model on the test data is: Decision Tree, with a score of 0.8888888888888888.**

Confusion Matrix

- The Confusion Matrix for the Decision Tree is better than any other decision tree.
- The only problem is with the false positives (top right), which we see are three of them. However, there are no false negatives.



Conclusions

- SpaceX should conduct more research on how to land heavier rockets if it wants to become more profitable and commercialized.
- Based on this data, launches with lower payload masses are more successful than launches with high payload masses.
- The success rate of rocket launches does increase over the years.
- Orbits ES-L1, SSO, GEO, and HEO have a 100% success rate.
- The Decision Tree is the best algorithm for this data set.

Appendix

- Capstone_1
- Capstone_2
- Capstone_3
- Capstone_4
- Capstone_5
- Capstone_6
- Capstone_7
- Capstone_8

Thank you!