

# Winning Space Race with Data Science

Naraen Palanikumar  
09/22/2025



# Outline

---

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

# Executive Summary

---

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis
    - Data Visualization
    - SQL
  - Building an interactive map with Folium
  - Building a dashboard with Plotly Dash
  - Predictive Analysis comparing 4 classification algorithms
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive component screenshots
  - Predictive Analysis relative performance results

# Introduction

---

- Project background and context
  - SpaceX is a well-known commercial space company that specializes in saving costs for customers seeking to deploy satellites and other space-related things into different orbits. This cost saving is largely achieved through a reusable first stage rocket. Therefore, whether or not this first stage will properly land is vital in determining the cost of a launch. This is the main focus of this project.
- Problems you want to find answers to
  - How do variables associated with launches, such as payload mass and launch site, affect the success of the first stage landing?
  - Does the rate of successful landings increase over the years?
  - What is the best algorithm we can utilize to predict whether a first stage will land given the features of interest associated with it?

Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX API as well as scraped from an archived Wikipedia page relating to past SpaceX launches.
- Perform data wrangling
  - Data was standardized and merged across datasets, and one-hot encoding was applied to the launch data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four different classification models were utilized and their relative predictive performance was compared.

# Data Collection

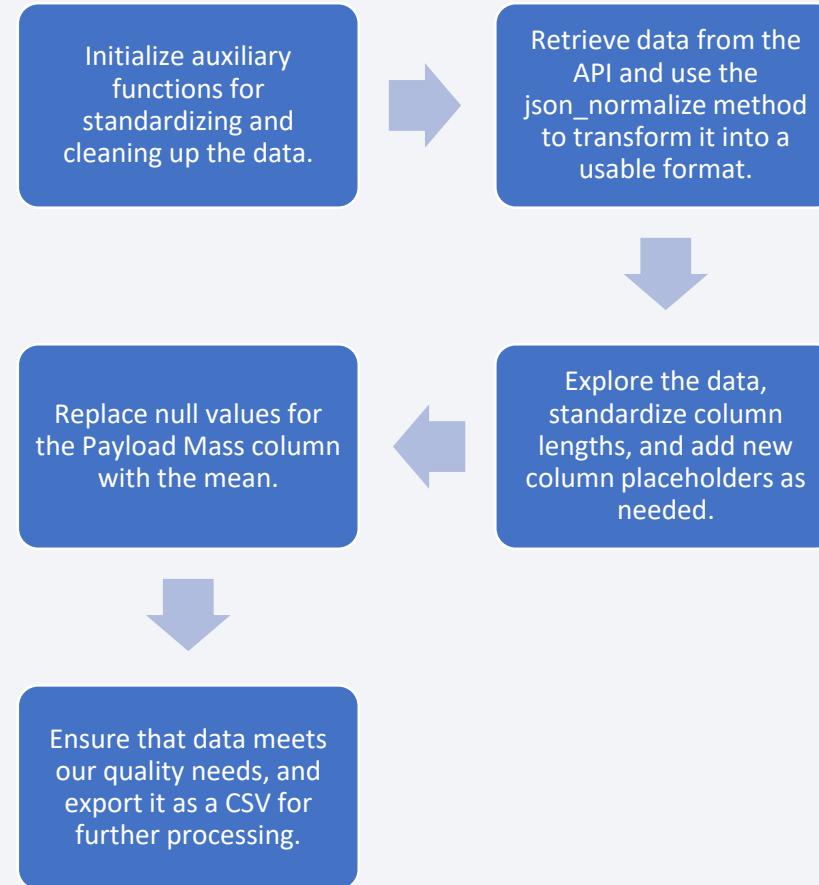
---

- Sources of data collection:
  - Datasets were collected from two sources, the SpaceX API as well as an archived Wikipedia page containing past launches.
- The next two slides will go into further detail using flowcharts to detail how this data was collected and processed for further use.

# Data Collection – SpaceX API

---

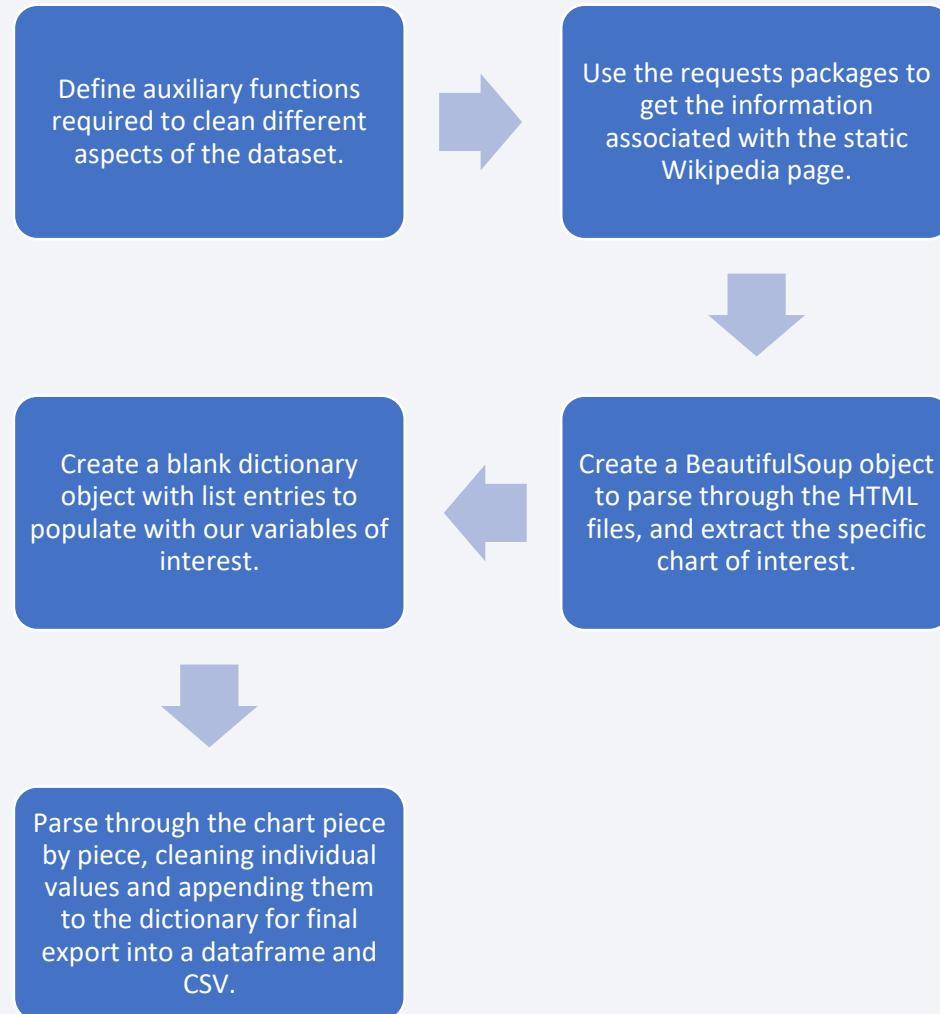
- On the right, you can view the general flowchart outlining how data was collected from and processed from the SpaceX API.
- [https://github.com/naraenp/course\\_ra\\_portfolio/blob/main/applied\\_ds\\_capstoneproj/ds\\_capstone.ipynb](https://github.com/naraenp/course_ra_portfolio/blob/main/applied_ds_capstoneproj/ds_capstone.ipynb)



# Data Collection - Scraping

---

- On the right, you can view the general flowchart outlining how data was collected from and processed from the Wikipedia Page.
- [https://github.com/naraenp/coursera\\_portfolio/blob/main/applied\\_ds\\_capstoneproj/ds\\_capstone.ipynb](https://github.com/naraenp/coursera_portfolio/blob/main/applied_ds_capstoneproj/ds_capstone.ipynb)



# Data Wrangling

---

- Data was first explored through basic commands like `head()` and `value_counts()` to determine the makeup of the data and the presence of null values.
- Landing outcomes were put into a single list and a new column was created to classify landings as successes (1) or failures (0).
- [https://github.com/naraenp/coursera\\_portfolio/blob/main/applied\\_ds\\_capstoneproj/ds\\_capstone.ipynb](https://github.com/naraenp/coursera_portfolio/blob/main/applied_ds_capstoneproj/ds_capstone.ipynb)

# EDA with Data Visualization

---

- Scatter plots comparing payload mass and flight number with hues indicating success were used to determine the influence of these variables on a successful landing.
- Similar scatter plots looking at variables like launch site and orbit type were made. A bar plot of success rate by orbit type was also made.
- These sorts of visualizations help us narrow down what features actually affect the success rates.
- [https://github.com/naraenp/coursera\\_portfolio/blob/main/applied\\_ds\\_capstone\\_eproj/ds\\_capstone.ipynb](https://github.com/naraenp/coursera_portfolio/blob/main/applied_ds_capstone_eproj/ds_capstone.ipynb)

# EDA with SQL

---

- Attached is an image containing all the SQL queries performed
- [https://github.com/naraenp/coursera\\_portfolio/blob/main/applied\\_ds\\_capstone/proj/ds\\_capstone.ipynb](https://github.com/naraenp/coursera_portfolio/blob/main/applied_ds_capstone/proj/ds_capstone.ipynb)

```
...
SQL Commands for Previous Section:
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE "CCA%" LIMIT 5
%sql SELECT SUM(PAYLOAD_MASS_KG_) FROM SPACEXTABLE WHERE Customer LIKE "NASA%"
%sql SELECT AVG(PAYLOAD_MASS_KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE "F9 v1.1%"
%sql SELECT DISTINCT Landing_Outcome FROM SPACEXTABLE
%sql SELECT * FROM SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)" ORDER BY Date ASC LIMIT 1
%sql SELECT DISTINCT Landing_Outcome FROM SPACEXTABLE
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS_KG_" BETWEEN 4000 AND 6000
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS TotalMissions FROM SPACEXTABLE GROUP BY Mission_Outcome
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS_KG_" = (SELECT MAX("PAYLOAD_MASS_KG_") FROM SPACEXTABLE)
%sql SELECT * FROM SPACEXTABLE LIMIT 5
%sql SELECT DISTINCT Landing_Outcome FROM SPACEXTABLE
%sql SELECT substr(Date, 6,2) AS "Month", "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' AND "Landing_Outcome" = 'Failure (drone ship)'
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS TotalMissions FROM SPACEXTABLE WHERE Date BETWEEN "2010-06-04" AND "2017-03-28" GROUP BY Mission_Outcome
...  
...
```

# Build an Interactive Map with Folium

---

- Circles indicating launch sites were added to Folium, along with marker clusters attached to these launch sites indicating the number of successful and failed landings there.
- Lines were also plotted from a single launch site to the nearest coast, highway, city, and railroad to annotate the distances of facilities like these from a standard launch site.
- [https://github.com/naraenp/coursera\\_portfolio/blob/main/applied\\_ds\\_capstoneproj/ds\\_capstone.ipynb](https://github.com/naraenp/coursera_portfolio/blob/main/applied_ds_capstoneproj/ds_capstone.ipynb)

# Build a Dashboard with Plotly Dash

---

- The main inputs of the Plotly app help narrow down what site to select to visualize and the ranges of payloads to visualize as well.
- The main outputs are a Pie chart showing successful launches by site and a scatter plot showing the correlation between payload size and success.
- [https://github.com/naraenp/coursera\\_portfolio/blob/main/applied\\_ds\\_capstone\\_proj/ds\\_capstone.ipynb](https://github.com/naraenp/coursera_portfolio/blob/main/applied_ds_capstone_proj/ds_capstone.ipynb)

# Predictive Analysis (Classification)

---

- Four different classification models (Logistic, SVM, Decision Tree, and KNN) were used with an 8:2 training test split to evaluate model performance.
- The best model was then selected based on relative performance with the testing data and related confusion matrix.
- [https://github.com/naraenp/coursera\\_portfolio/blob/main/applied\\_ds\\_capstone\\_proj/ds\\_capstone.ipynb](https://github.com/naraenp/coursera_portfolio/blob/main/applied_ds_capstone_proj/ds_capstone.ipynb)

# Results

---

- Exploratory data analysis results
  - Greater flight number and higher payload mass seem to be correlated with more successful launches.
  - The site KSC LC 39A seems to have a higher number of successful launches, but this may be because it was only used in more recent launches.
  - ES-L1, SSO, HEO, GEO orbits had the highest success rates, but they also had the least number of associated launches.
- Interactive analytics demo in screenshots
  - Will be show in coming slides
- Predictive analysis results
  - The predictive analysis models compared showed that a decision tree classifier performed the best with an 88.9 percent accuracy on testing data.

The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that curves and twists across the frame, resembling a three-dimensional space or a network of data points. The overall effect is futuristic and dynamic.

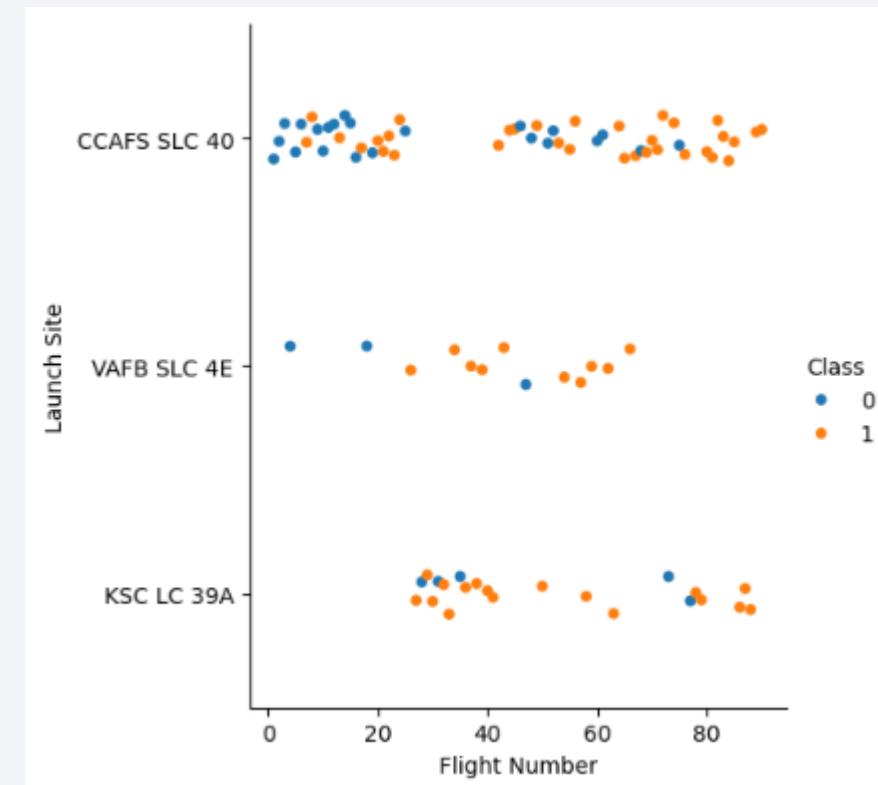
Section 2

## Insights drawn from EDA

# Flight Number vs. Launch Site

---

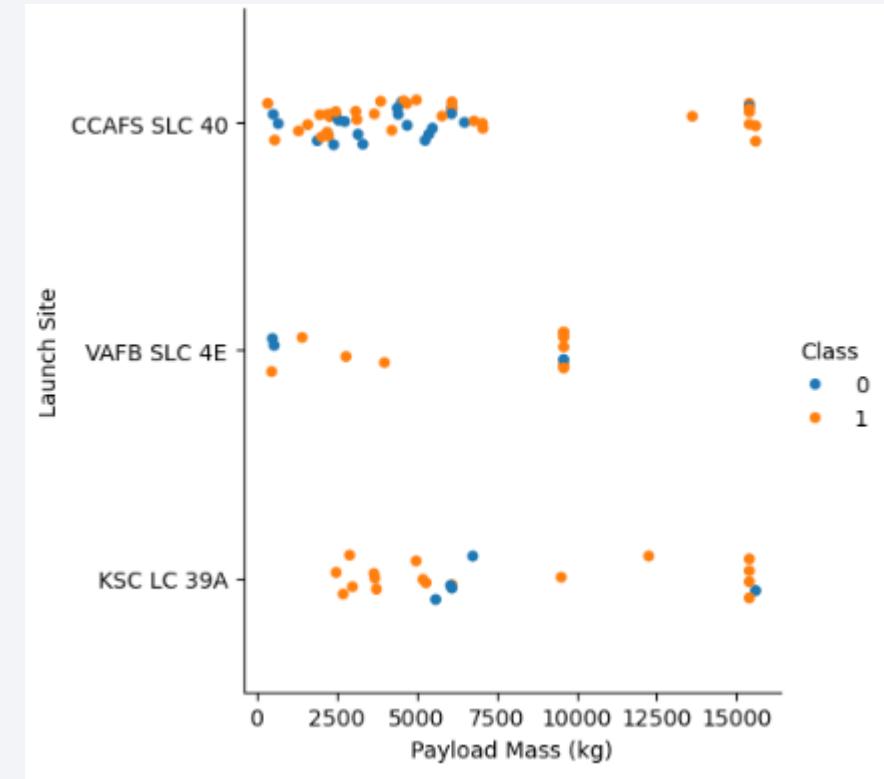
- Higher flight numbers seem correlated with higher success rate. The Launch site seems less correlated compared to flight number. This makes sense as greater experience and data collected from previous launches should ensure that future ones go better.



# Payload vs. Launch Site

---

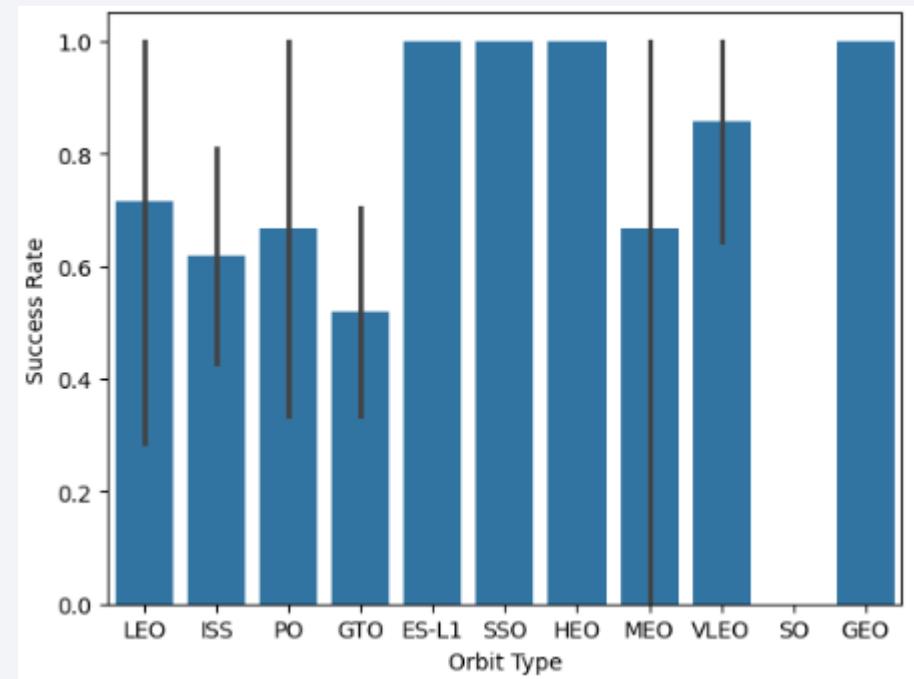
- It seems that CCAFS and KSC launch sites are preferred for the very highest payload launches where VAFB is preferred for more intermediate sized launches. Higher payloads also tend to be more successful, maybe due to a more stringent QC on these possibly more important launches.



# Success Rate vs. Orbit Type

---

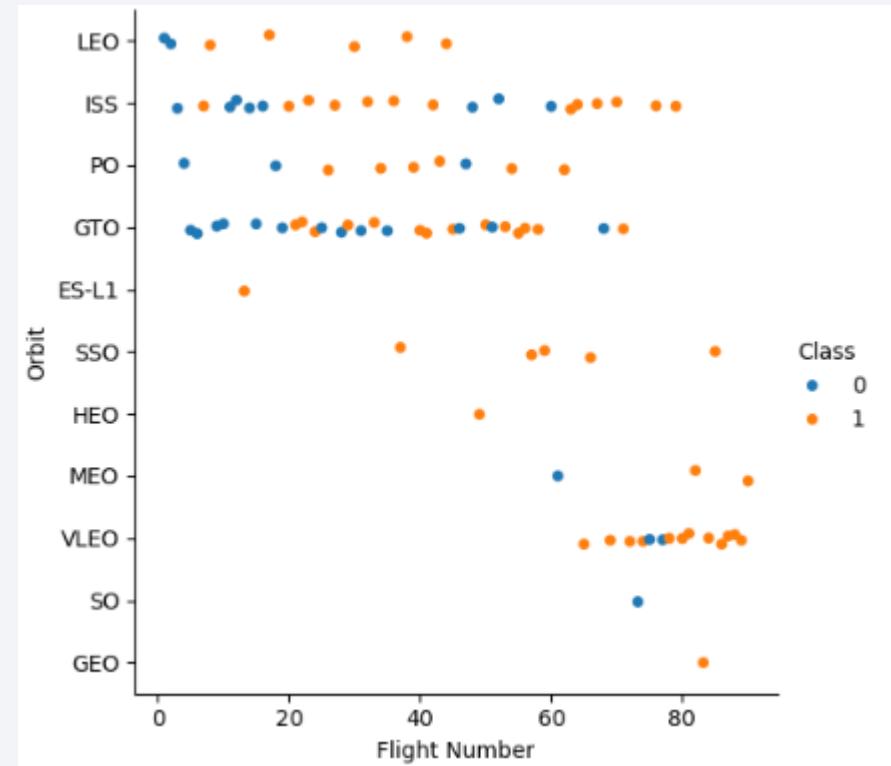
- This chart seems to be better at showing which orbits have had the most and latest launches than the actual success rate of each orbit type. The orbits with the highest success rate also have had the fewest launches associated with them and have been relatively more recent launches.



# Flight Number vs. Orbit Type

---

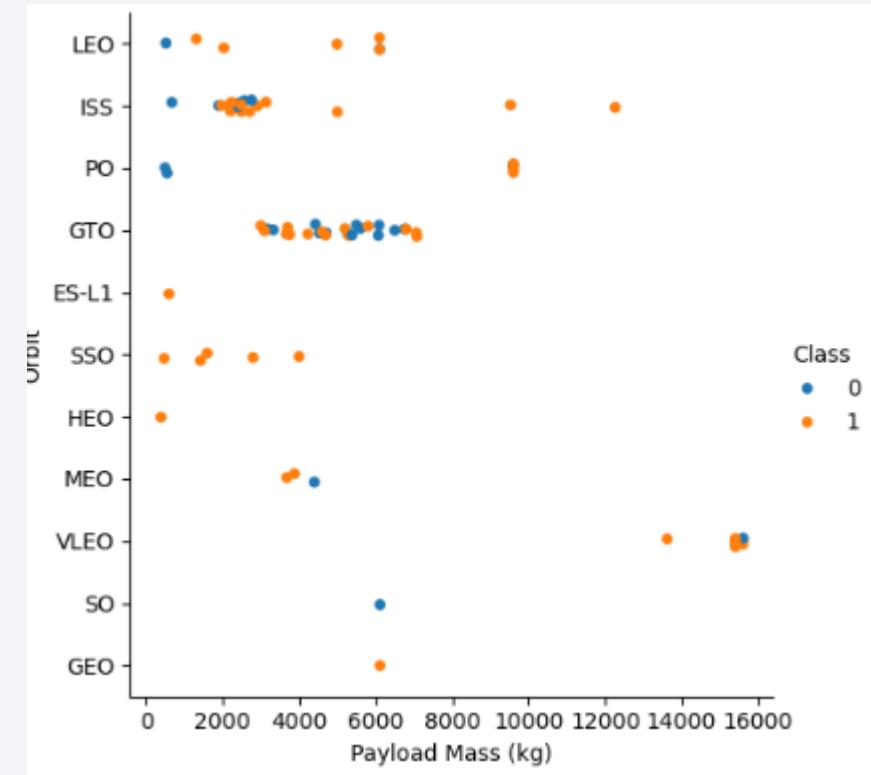
- It seems that certain orbit configurations are only recently being tested and have been quite successful. This chart seems to show a story of growth from testing launches with certain orbits to becoming successful with higher scale/more complex orbital launches.



# Payload vs. Orbit Type

---

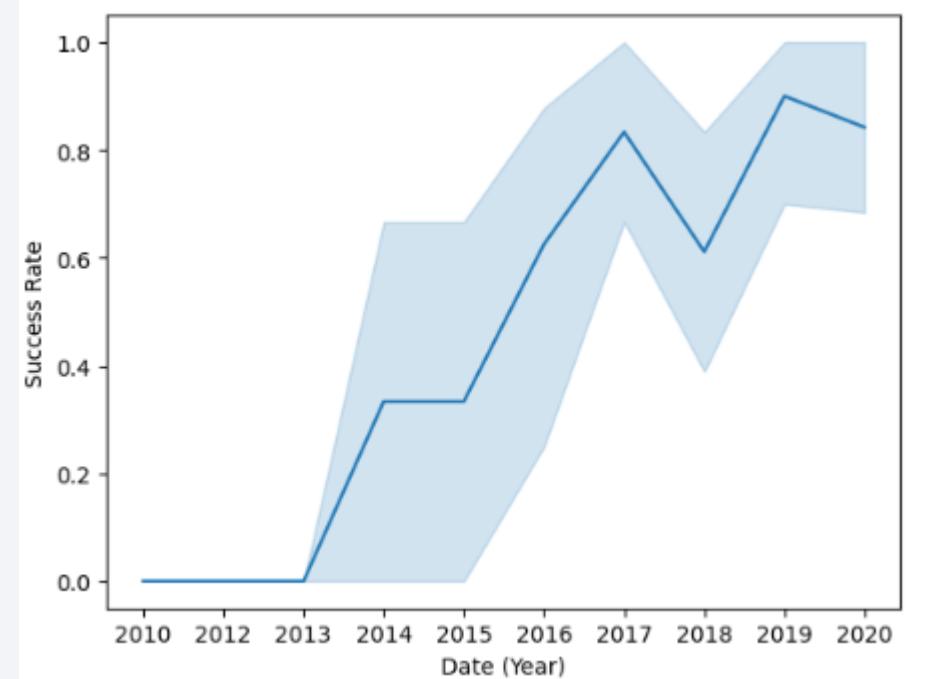
- This plot seems to show that the higher payload masses are usually reserved for certain orbital configurations, especially VLEO. There are also some heavy payloads going to the ISS which intuitively makes sense as it has to be frequently resupplied.



# Launch Success Yearly Trend

---

- This line chart clearly shows that the success rate of launches has been increasing with every year.



# All Launch Site Names

---

Launch\_Site

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

# Launch Site Names Begin with 'CCA'

---

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

---

`SUM(PAYLOAD_MASS_KG)`

---

`99980`

# Average Payload Mass by F9 v1.1

---

AVG(PAYLOAD\_MASS\_KG )

---

2534.66666666666666665

# First Successful Ground Landing Date

---

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

# Successful Drone Ship Landing with Payload between 4000 and 6000

---

**Booster\_Version**

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

# Total Number of Successful and Failure Mission Outcomes

---

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

# Boosters Carried Maximum Payload

---

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

# 2015 Launch Records

---

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

---

Mission_Outcome	TotalMissions
Failure (in flight)	1
Success	30

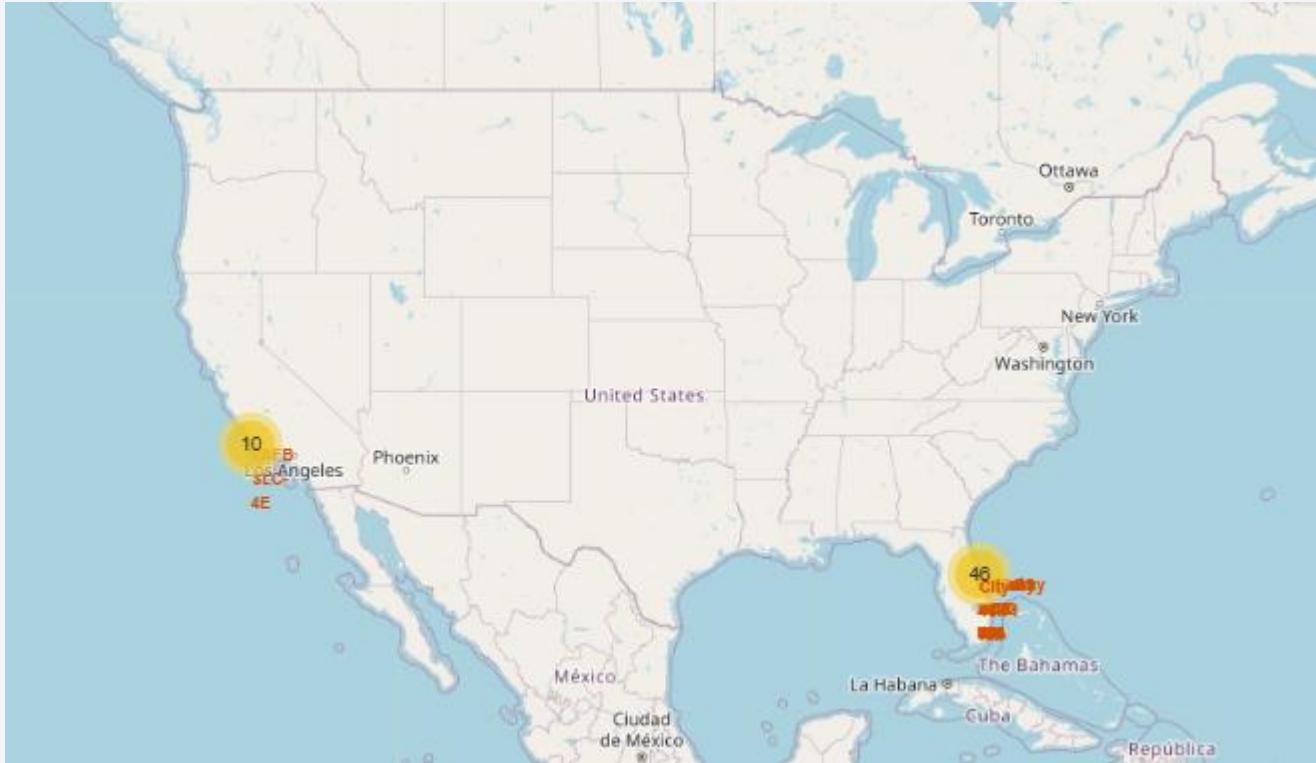
The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against the dark void of space. City lights are visible as numerous small white and yellow dots, primarily concentrated in the lower right quadrant where the United States appears. In the upper left quadrant, the green and blue glow of the aurora borealis (Northern Lights) is visible in the upper atmosphere.

Section 3

# Launch Sites Proximities Analysis

# Global Folium Map

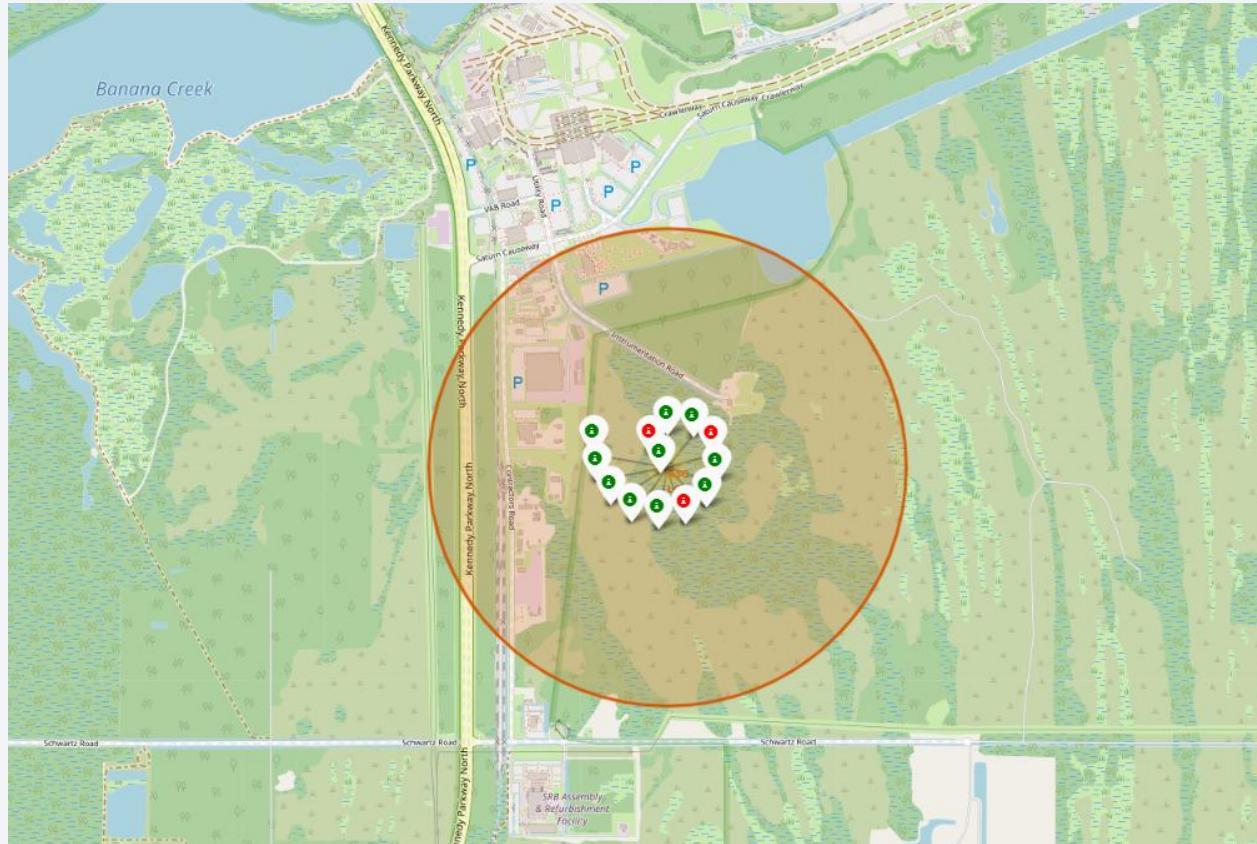
---



- Folium map showing the main launch site areas represented in the data set. We can see that they seem to be clustered in the southern coastal regions of the US. Which makes sense as launch sites closer to the equator can save rockets a good amount of the energy required to leave the atmosphere.

# Folium Map Marker

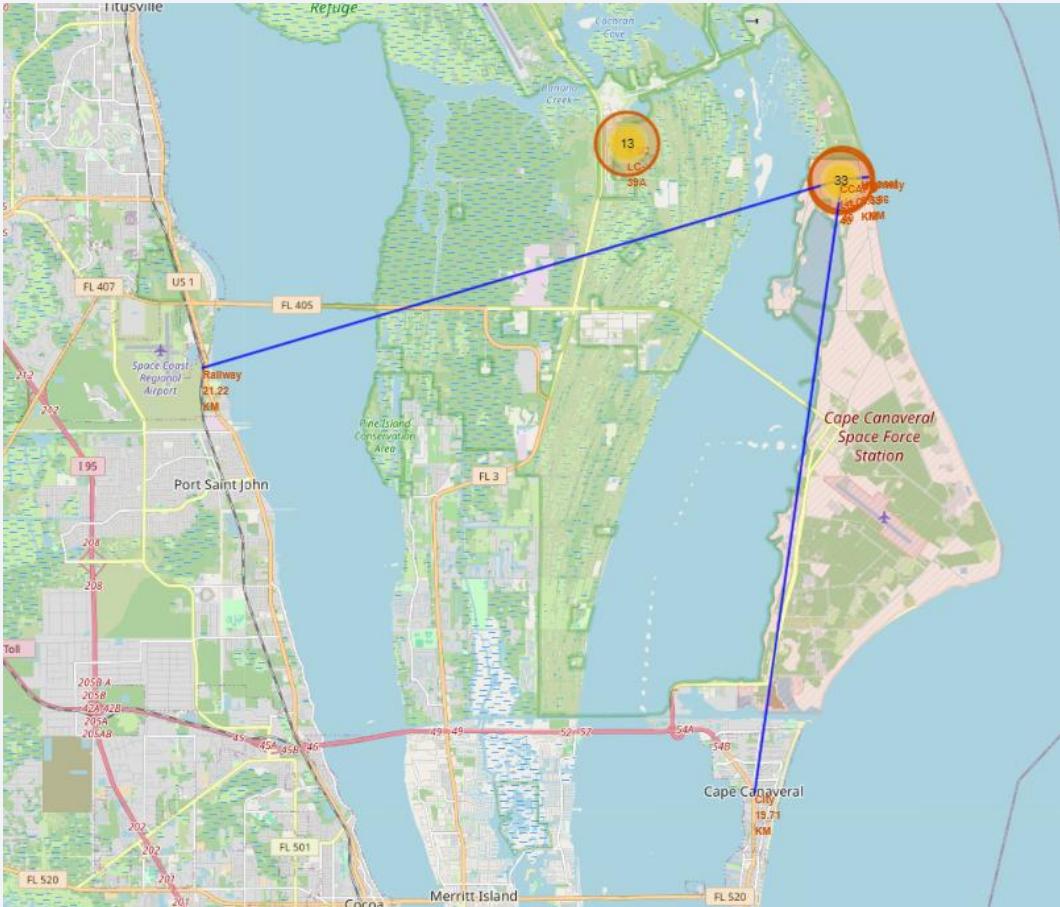
---



- Folium map marker showing expanded site details for launch site KSC LC-39A. These markers represent successful and failed launches represented in green and red respectively.

# Folium Map Distances

---



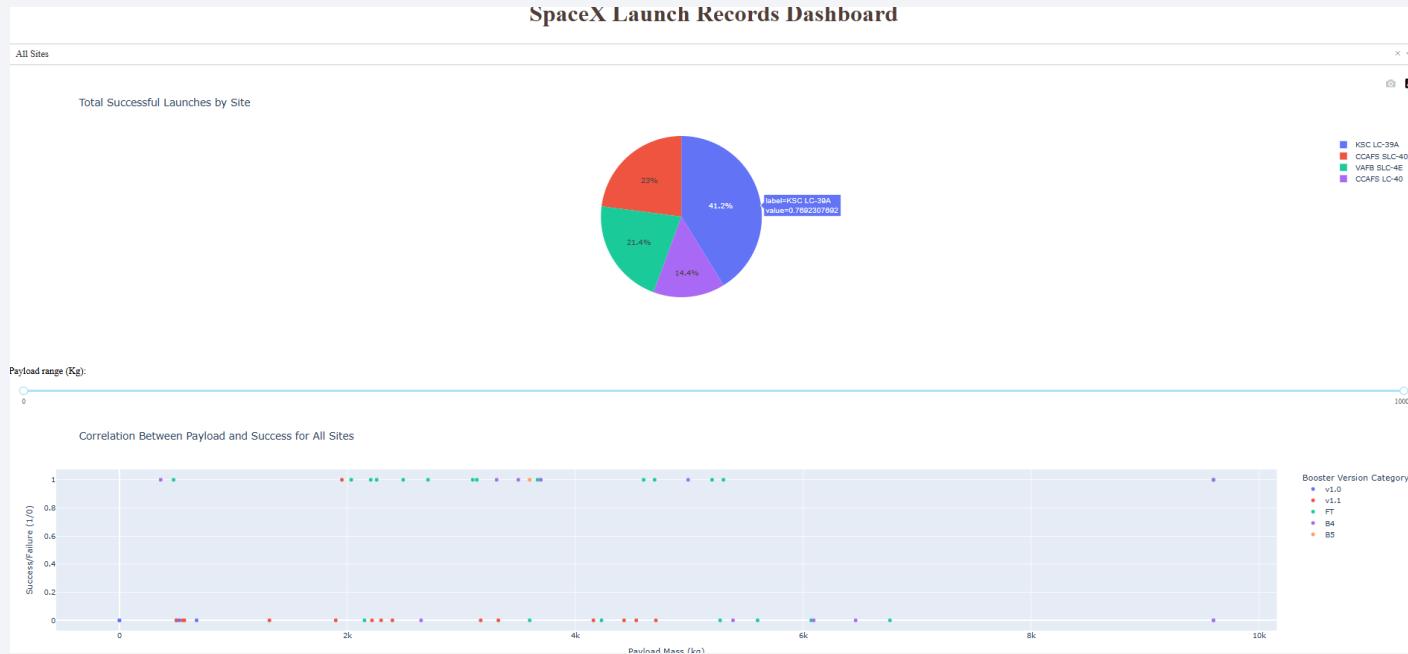
- Folium map lines showing the distances from the CCAFS SLC-40 launch site to the nearest coast (0.86 km), highway (0.59 km), railroad (21.22 km), and city (19.71 km). These distances tell us that launch sites are often in close proximity to coasts and highways (which makes sense for transport and recovery logistics). The high distance from cities and railroads is likely because launch sites should be somewhat isolated from civilian infrastructure in case things go wrong.

Section 4

# Build a Dashboard with Plotly Dash



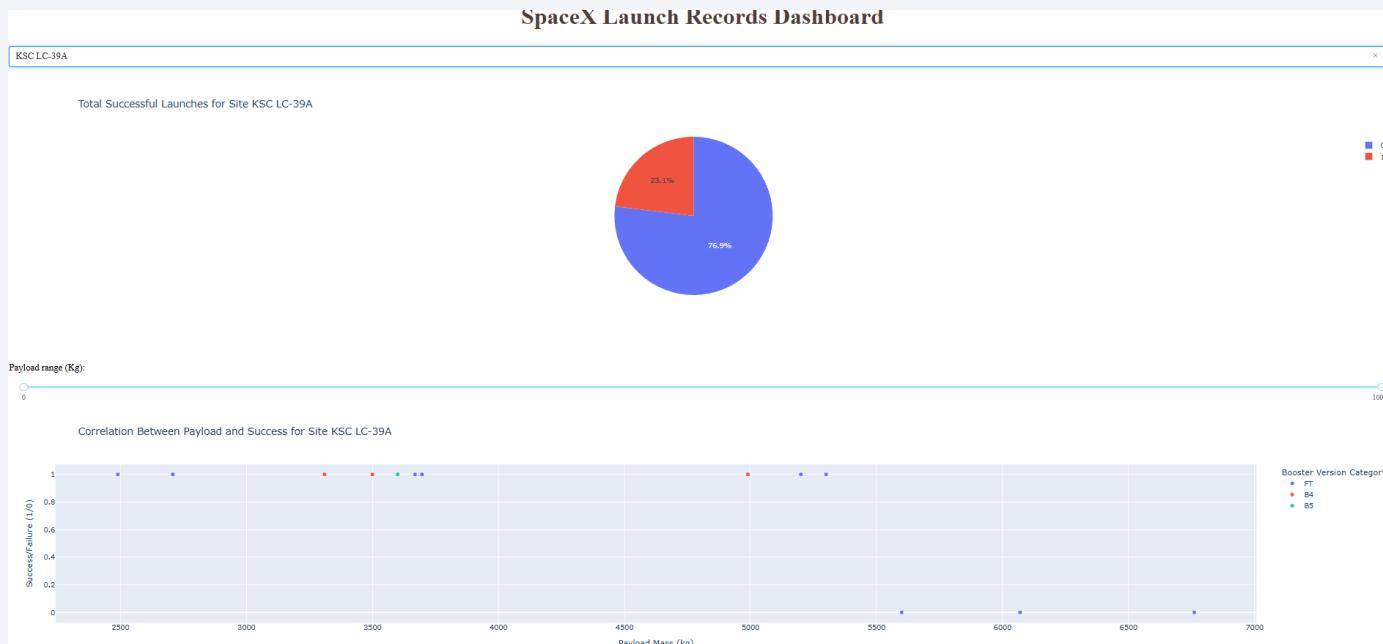
# Dashboard (All Sites)



- The pie chart within the dashboard showing all sites tells us that the majority of successful launches have come from the site KSC LC-39A.
- The scatter plot honestly doesn't do that good of a job of representing any sort of relation between payload mass and success rate. It seems like intermediate range masses from 2 to 6 thousand kg have a higher rate of demand.

# Dashboard (KSC LC-39A)

---

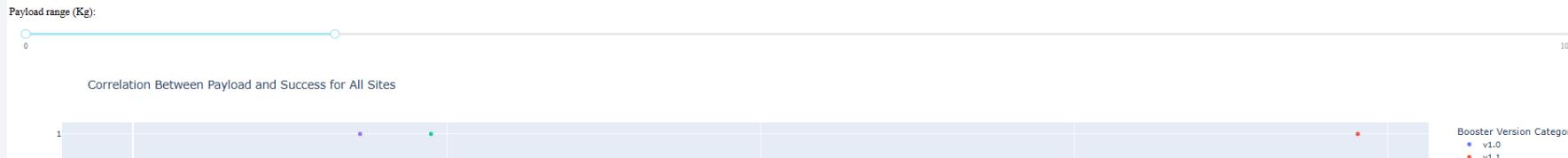


- The pie chart here tells us that around 77 percent of all launches at site KSC LC-39A have been successful.
- The scatter plot tells us that larger payloads (i.e. greater than 5500 kg) are the ones that tend to fail more often.

# Payload vs Success Rate



- Heavy payloads greater than 6 thousand kg seem to have the lowest success rate.
- Light payloads follow up with a similarly low success rate.
- Intermediate payloads between 2 and 6 thousand kg both have a higher demand and a higher success rate.



The background of the slide features a dynamic, abstract design. It consists of several thick, curved lines in shades of blue and yellow, creating a sense of motion and depth. The lines curve from the bottom left towards the top right, with some lines being more prominent than others. The overall effect is reminiscent of a tunnel or a high-speed journey through a digital space.

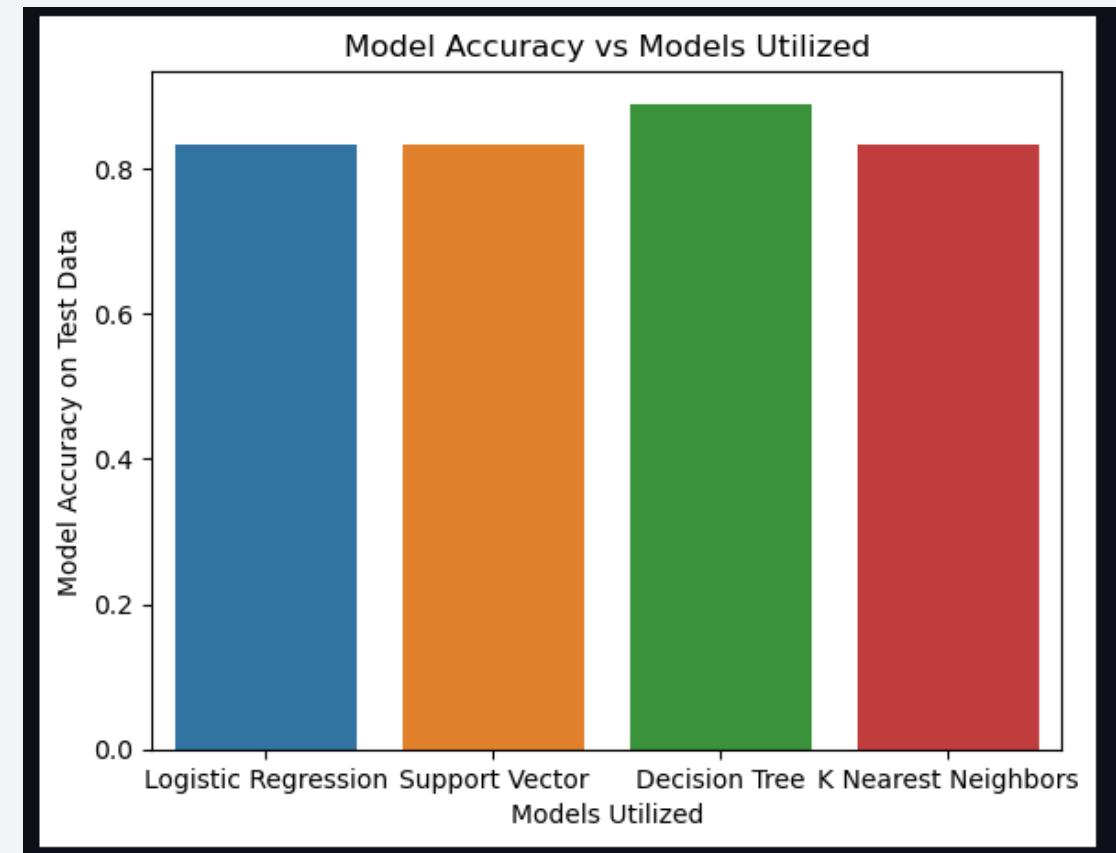
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

---

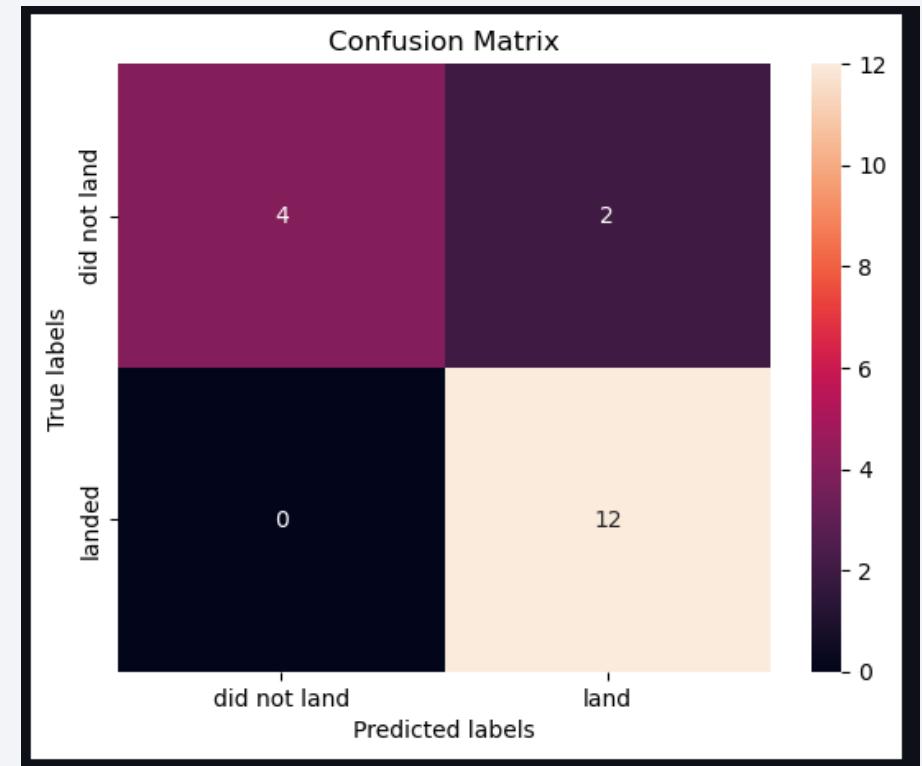
- From this graphic, it is evident that the decision tree classifier has the highest accuracy on the testing data. Therefore, if we were to deploy a model to predict first stage landing success, this is the model that should be utilized.



# Confusion Matrix

---

- This is the confusion matrix for the best performing decision tree classifier. We see 12 true positive, 4 true negative results. Along with this, 2 false positive and 0 false negative results. This gives us an accuracy of  $(12 + 4)/(12 + 4 + 2 + 0) = 88.9\%$  accuracy.



# Conclusions

---

- The success rate of launches have increased over the years.
- Most of the launch sites are close to the equator and the coast.
- Launches with intermediate payload mass seem to perform the best.
- KSC LC-39A has the highest success rate of launches of all sites.
- The decision tree model is the best classifier to move forward with the data.

# Appendix

---

- Thank you to Coursera, IBM, and the instructors for providing this course!

Thank you!

