



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

M. Görkem Ulutürk
15/11/2024



Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Methodologies
 - Data was collected using SpaceX API, and by web scraping using bs4.BeautifulSoup
 - Data wrangling revolved around creating new features
 - Data visualization tools were used to display relations between variables
 - SQL was used for advanced querying for data exploration purposes
 - Interactive maps were built with folium for discovering similarities between launch sites
 - A dashboard was built with Plotly Dash for interactive plotting, further data exploration
 - Different models were built for predictive analysis. Models were trained, tested, refined, and scored to find the best performing one
- Results
 - SQL queries and data wrangling revealed nontrivial insights
 - Data visualization showed clear relations between some of the features
 - Maps helped finding the similarities of launch sites
 - Classification models performed the same on the test data but had different results on the entirety of the dataset

Introduction

- As a new competitor in the space industry, we inspected SpaceX' data to understand the profitability of space launches
- We explored, visualized, and analyzed the data to obtain a mechanism which would predict the success of a mission where we would return the launched rocket
- SpaceX has relatively lower costs and higher profit margins compared to others because of their reusable rockets; therefore, being able to predict the outcome based on features gives a competitive edge that would increase profits even further
- To achieve our goal, we collected, cleaned, explored, and visualized various data concerning payloads, launch stations, orbits, etc. to develop a model

Section 1

Methodology

Methodology

Executive Summary

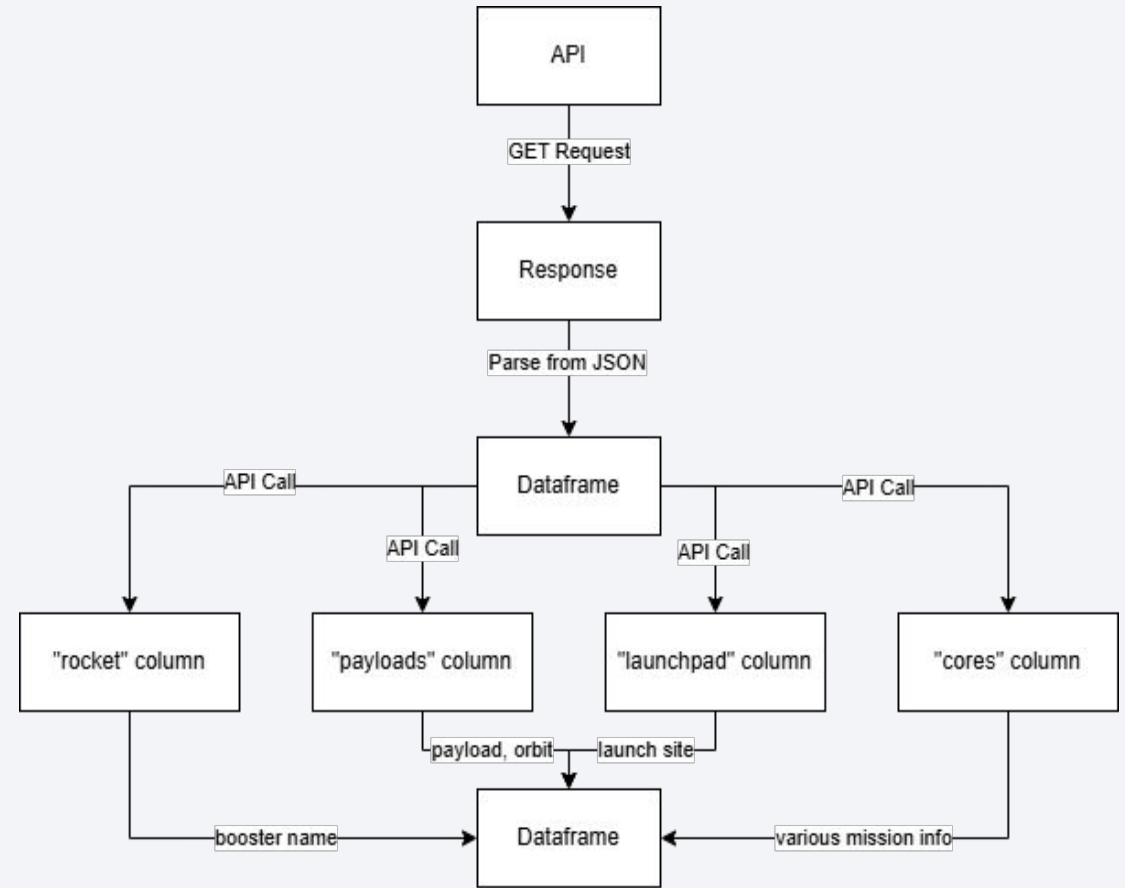
- Data collection methodology:
 - Data was collected using [SpaceX API](#) and by web scraping
- Perform data wrangling
 - Data was processed using Pandas
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - Models were created and tuned using Scikit-Learn

Data Collection

- Before data collection began, we started by identifying the data needs
- To develop a model, we needed past launch data consisting of various features
- Payload mass, launch site information, booster version, mission outcome, etc. are examples of such features
- The first stage of data collection: SpaceX API
- The second stage of data collection: Web Scraping
- Data was cleaned and properly formatted during the collection phase
- Data was then parsed and merged to a pandas DataFrame
- Data was later used to create an SQL database for advanced querying

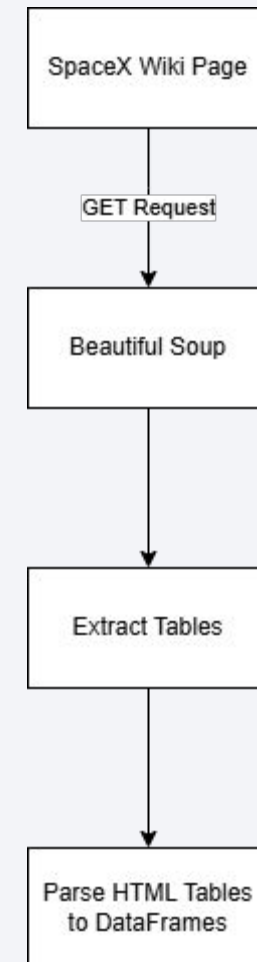
Data Collection – SpaceX API

- Past launch data was collected using SpaceX API GET requests
- Response is a string with a JSON structure
- Data was then parsed to JSON, then to pandas dataframe
- Data such as mission info, outcomes, etc. were obtained later via GET requests
- Responses merged to the main dataframe were cleaned and formatted
- [Data Collection API Notebook](#)



Data Collection - Scraping

- SpaceX Wikipedia page was used to collect data
- Orbit and customer information, flight number, payload information, etc. are examples of data collected via web scraping
- Webpage was obtained via an HTML GET request
- Response was parsed to html and processed by bs4.BeautifulSoup
- Tables from the webpage were extracted and processed
- Created a dictionary from the obtained data
- Dictionary was later parsed to a pandas DataFrame object
- [Data Collection with Web Scraping Notebook](#)



Data Wrangling

- Data was cleaned and formatted during the collection phase
- Data was explored to gain insights
- For example, there are 3 launch sites, Cape Canaveral Space Launch Complex 40 with the most launches
- There are 11 orbits with GTO being the most frequent orbit
- Different landing outcomes were categorized into Success and Fail
- Success rate was calculated as 66.67%
- [Data Wrangling Notebook](#)

EDA with Data Visualization

- Scatter, bar, and line plots were used to explore the data
- Scatter charts were used to compare the relationships between variables
- For example, payload masses by launch sites were plotted using scatter charts to demonstrate that heavier payloads were launched from particular launch sites
- Bar plots were used to compare rates
- For example, mission success rate of different orbits were compared by bar plots
- Line plots were used to discover trends
- For example, line plots were used to show the rise in success rate over the years
- [EDA with Data Visualization Notebook](#)

EDA with SQL

- Following SQL queries were performed for gaining insights about the data
 - Names of the unique launch sites
 - Records of Cape Canaveral Space Launch Complex
 - Total payload mass carried for NASA
 - Average payload mass carried by F9 v1.1
 - The date of the first successful landing
 - Boosters that successfully landed in drone ships
 - Total number of successful and failed missions
 - Boosters that carried the maximum payload
 - Months where drone ship landings resulted in failures in the year 2015
 - Landing outcomes between 2010 and 2017
- [EDA with SQL Notebook](#)

Build an Interactive Map with Folium

- Many maps were created using Folium to gain insights about launch sites
- Added markers, lines, and circles to the map
- Circles were used to indicate launch sites
- Markers were used to indicate mission outcomes per launch site
- Lines were used to show nearest shoreline, railroad, etc.
- [Interactive Visual Analytics with Folium Notebook](#)

Build a Dashboard with Plotly Dash

- Plotly Dash was used to build an interactive dashboard
- Dropdown menu added for launch site selection
- Range slider added for selecting payload mass range
- A pie chart was used for showing proportions of successful missions by launch site
- A scatter chart was used for showcasing the relationship between payload mass and mission outcomes
- [SpaceX Dash App Script](#)

Predictive Analysis (Classification)

- Scikit-Learn was used for preparing the data, creating models, and evaluation
- Logistic Regression, SVM, Decision Tree, and kNN models were used
- Used `train_test_split` function for creating training and testing data
- Improved each model with `GridSearchCV` using various parameters
- Tested the accuracy of the models using the testing set and various metrics
- Compared the results of different metrics for each model to find the best performing
- [SpaceX Machine Learning Prediction](#)

Results

- EDA revealed insights on relations between features, such as some orbits having more successful missions than others
- EDA showcased the increase in success rates in a yearly basis
- Maps showed similarities between launch sites, such as proximities to shores, etc.
- Interactive dashboard showed the proportion of successful missions by launch sites, and the relationship between payload, booster version, and success
- Predictive analysis produced models that performed fairly well on both the test data and the whole data set
- We compared different models to find the best performing one, which turned out to be SVM

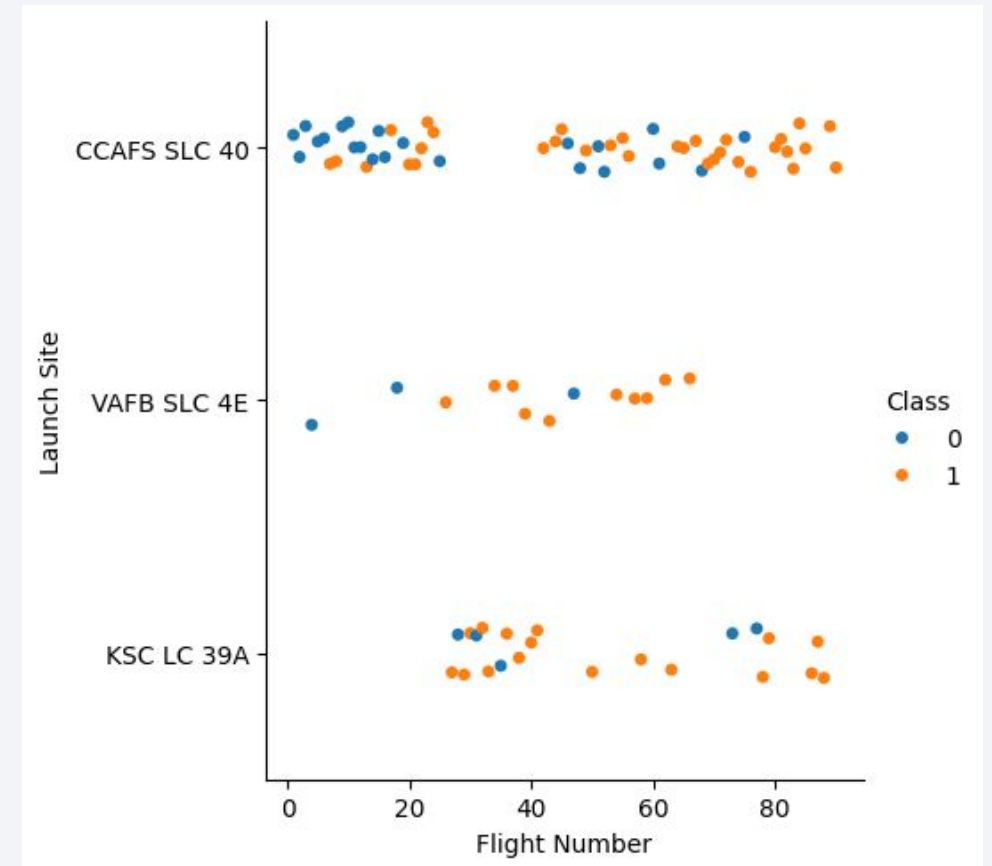
The background of the slide is an abstract composition. It features a solid blue area on the left side, which transitions into a dynamic pattern of diagonal streaks in shades of blue, red, and cyan on the right. These streaks are layered over a faint, grid-like texture, creating a sense of depth and movement.

Section 2

Insights drawn from EDA

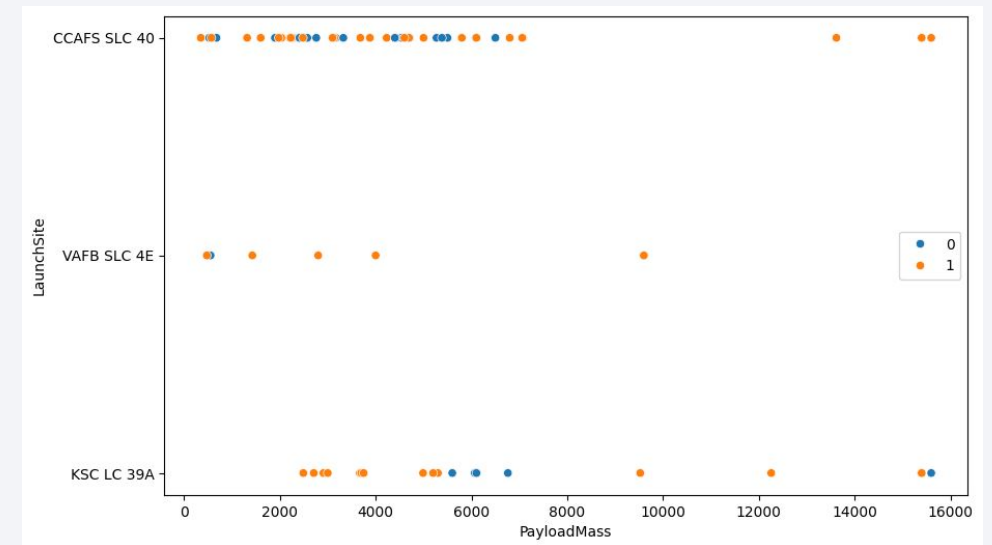
Flight Number vs. Launch Site

- Most flights occurred in CCAFS
- KSCLC wasn't used initially
- CCAFS is the most frequent entry
- CCAFS carried most failures



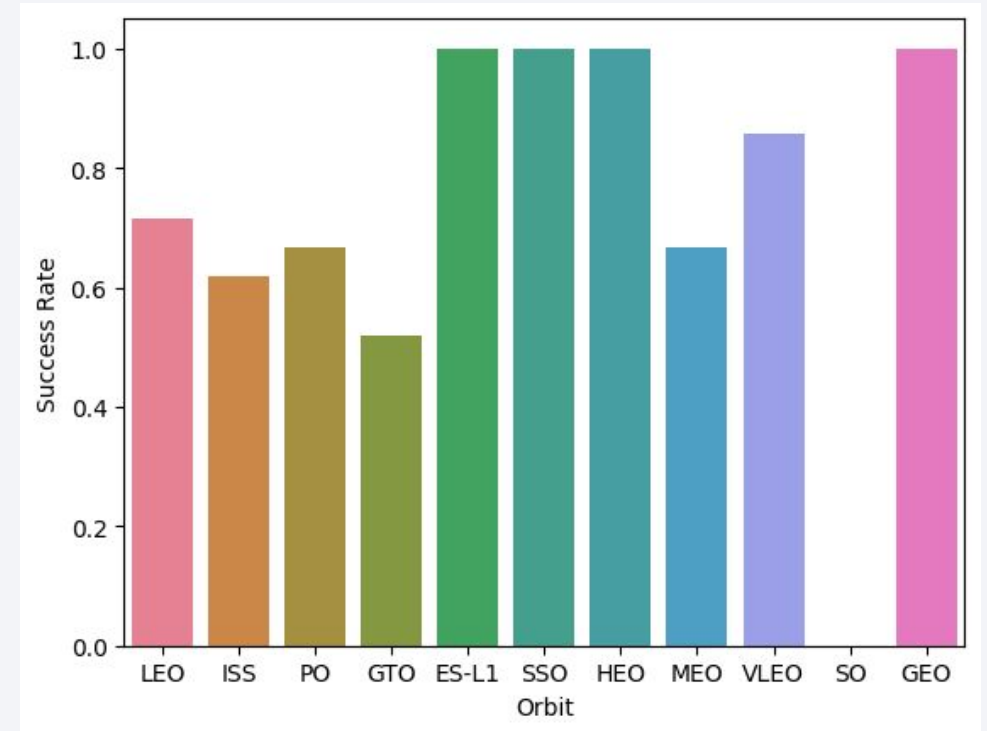
Payload vs. Launch Site

- Heavier payloads aren't launched from VAFB
- KSCLC is not preferred for lighter payloads
- Missions with payloads heavier than 8000 kg mostly succeeded
- Heavy payloads are less frequent than light payloads



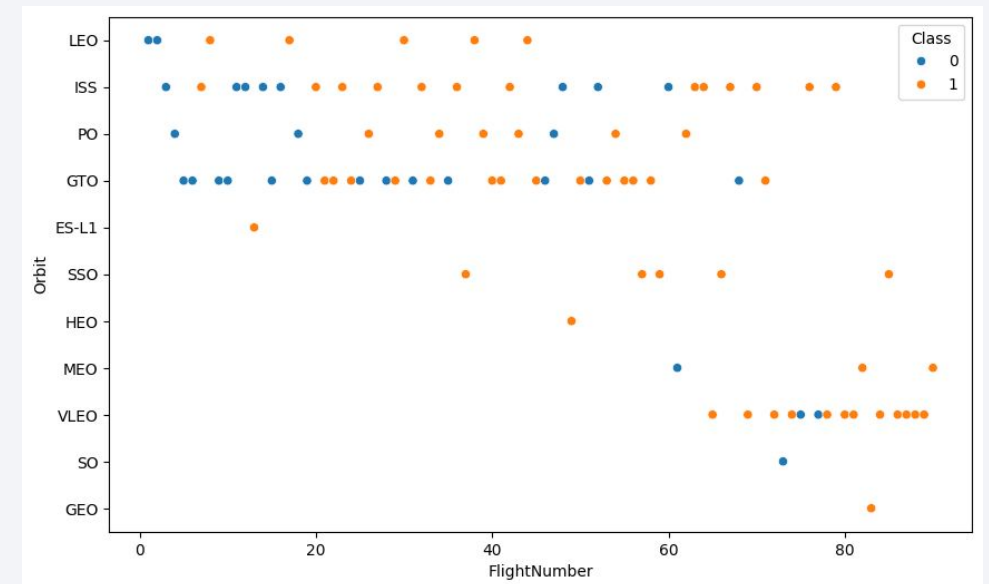
Success Rate vs. Orbit Type

- GTO and SO has the least success rate
- ES-L1, SSO, HEO, GEO has 100% success rate
- Most of the orbits have less than 80% success rate
- Almost all orbits have more than 50% success rate
- GEO, HEO, and ES-L1 are the orbits with highest success rate and highest from the surface



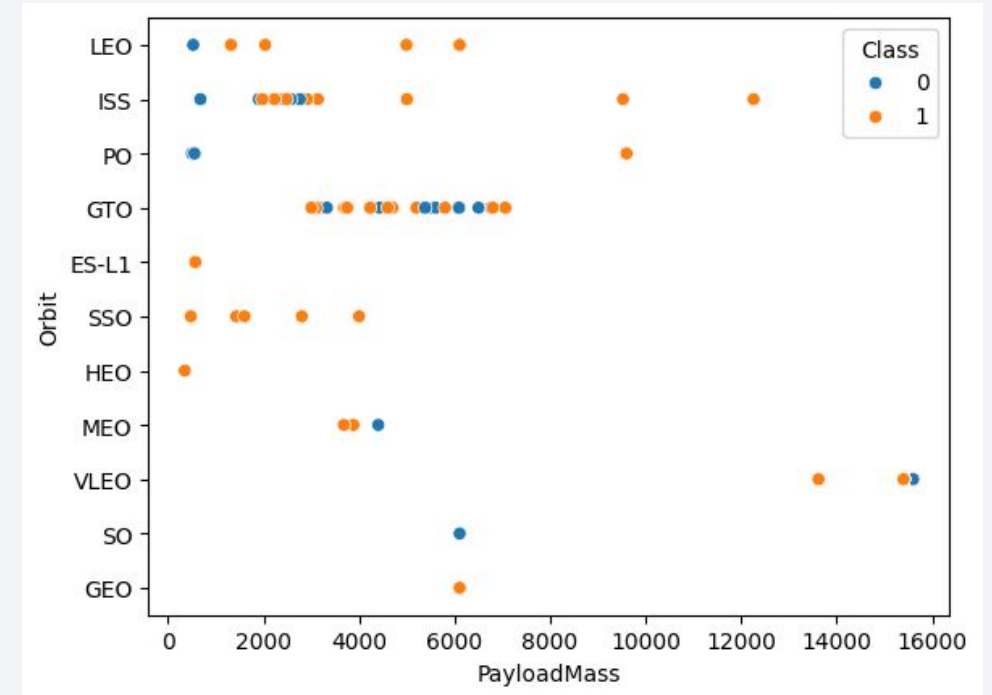
Flight Number vs. Orbit Type

- Some orbits such as ES-L1 and HEO missions are very rare
- ISS, GTO, and VLEO missions are most frequent
- Success increased with flight number for all orbits
- All orbits with 100% success rate had very few missions
- SO, only orbit with 0% success only had 1 mission



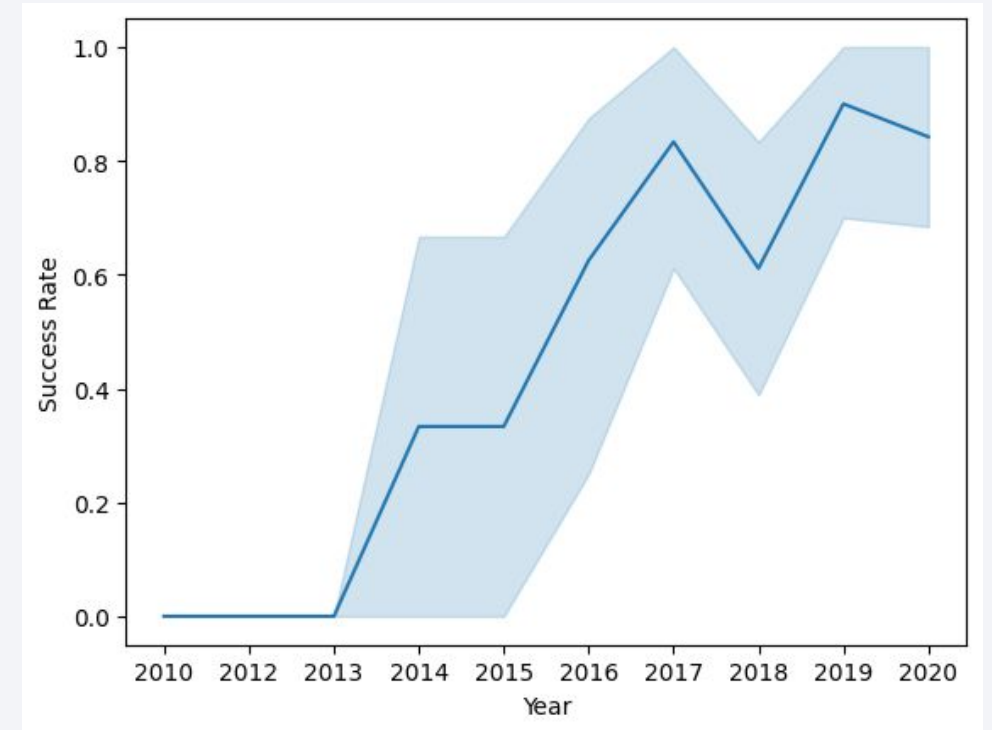
Payload vs. Orbit Type

- Some orbits such as LEO and SSO didn't receive heavier payloads
- VLEO received heaviest payloads and no light payload
- Success rates increased with payload mass for LEO, ISS, and PO
- Orbits with 100% success rate received payloads of 6000 kg or less



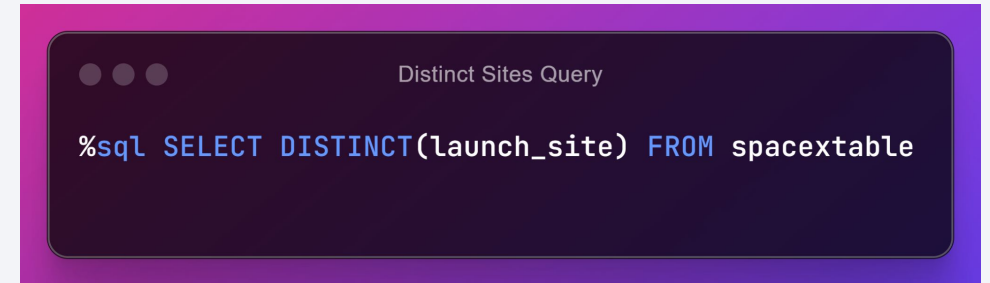
Launch Success Yearly Trend

- First successful mission was in 2014
- Success rate increased consistently until 2017
- No improvement from 2014 to 2015
- 2018 had a big drop in success
- More than 90% of missions in 2019 resulted in success
- Success rate has an overall positive trend from 2010 to 2020



All Launch Site Names

- Unique launch sites found using the query on right:
 - CCAFS LC-40 (Cape Canaveral Space Force Station Launch Complex 40)
 - CCAFS SLC-40
 - KSC LC-39A (Kennedy Space Center Launch Complex 39A)
 - VAFB SLC-4E (Vandenberg Space Force Base Space Launch Complex 4 East)



```
%sql SELECT DISTINCT(launch_site) FROM spacetable
```

Launch Site Names Begin with 'CCA'

- 5 records from the launch site CCAFS found using the query on right:

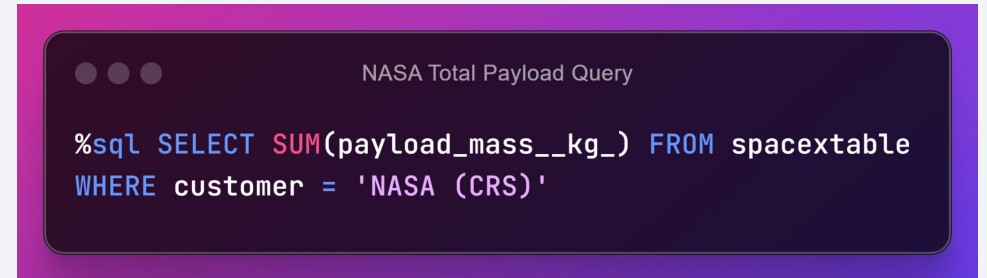
```
CCA Site Query

%sql SELECT * FROM spacetable WHERE launch_site
LIKE 'CCA%' LIMIT 5
```

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

- Total payload mass carried for NASA found using the query on right is 45,596 kg

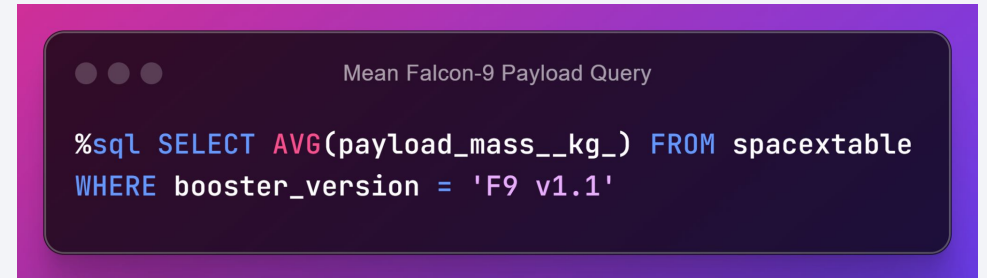
A terminal window with a dark background and a pink-to-purple gradient border. The title bar reads "NASA Total Payload Query". The terminal contains a SQL query in a monospaced font with syntax highlighting: "%sql SELECT SUM(payload_mass__kg_) FROM spacetable WHERE customer = 'NASA (CRS)'".

```
NASA Total Payload Query

%sql SELECT SUM(payload_mass__kg_) FROM spacetable
WHERE customer = 'NASA (CRS)'
```

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 found using the query on right is 2928.4 kg

A terminal window with a dark background and a pink-to-purple gradient border. The title bar reads "Mean Falcon-9 Payload Query". The terminal contains a SQL query:

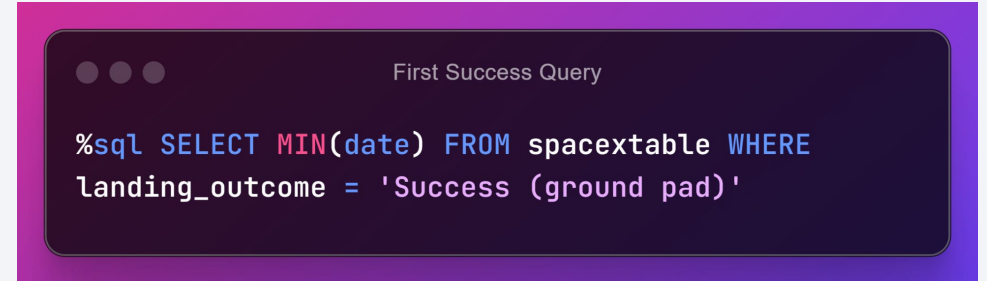
```
%sql SELECT AVG(payload_mass__kg_) FROM spacetable WHERE booster_version = 'F9 v1.1'
```

```
Mean Falcon-9 Payload Query

%sql SELECT AVG(payload_mass__kg_) FROM spacetable
WHERE booster_version = 'F9 v1.1'
```

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad found using the query on right is 2015-12-22
- This date is close to the first successful landing which was in 2014

A terminal window with a dark background and a pink border. The title bar at the top right says "First Success Query". The terminal contains a SQL query in a syntax-highlighted font:

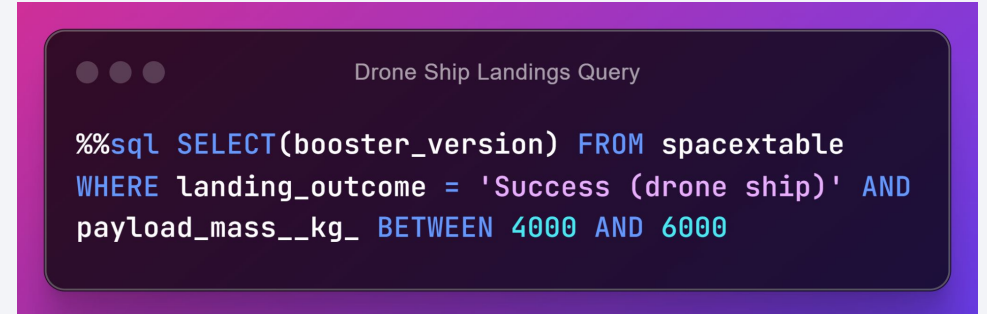
```
%sql SELECT MIN(date) FROM spacetable WHERE  
landing_outcome = 'Success (ground pad)'
```

```
First Success Query

%sql SELECT MIN(date) FROM spacetable WHERE
landing_outcome = 'Success (ground pad)'
```


Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 found using the query on right:
 - F9 FT B1022
 - F9 FT B1026
 - F9 FT B1021.2
 - F9 FT B1031.2

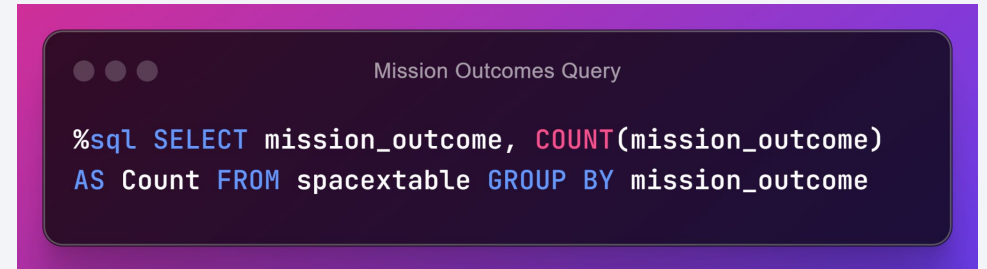
A terminal window with a dark background and a pink-to-purple gradient border. The title bar reads 'Drone Ship Landings Query'. The terminal contains a SQL query in a monospaced font with syntax highlighting: '%%sql SELECT(booster_version) FROM spacetable WHERE landing_outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000'.

```
Drone Ship Landings Query

%%sql SELECT(booster_version) FROM spacetable
WHERE landing_outcome = 'Success (drone ship)' AND
payload_mass__kg_ BETWEEN 4000 AND 6000
```

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes found using the query on right
- Almost all missions in the dataset was success
- Only 1 failure was recorded
- Note that we look at the mission outcomes rather than landing outcomes

A terminal window with a dark background and a pink border. The title bar says "Mission Outcomes Query". The SQL query is displayed in a monospaced font with syntax highlighting: %sql SELECT mission_outcome, COUNT(mission_outcome) AS Count FROM spacetable GROUP BY mission_outcome.

```
%sql SELECT mission_outcome, COUNT(mission_outcome)
AS Count FROM spacetable GROUP BY mission_outcome
```

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass found using the query on right:

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

```
Maximum Payload Boosters Query

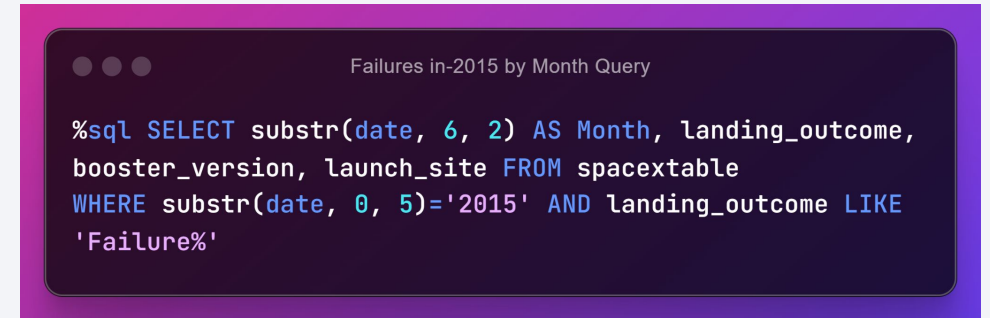
%%sql SELECT booster_version FROM spacetable
WHERE payload_mass__kg_ = (SELECT
MAX(payload_mass__kg_) FROM spacetable)
```

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 found using the query on right:

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

- Only two failures recorded, which came in January and April



```
%sql SELECT substr(date, 6, 2) AS Month, landing_outcome,
booster_version, launch_site FROM spacetable
WHERE substr(date, 0, 5)='2015' AND landing_outcome LIKE
'Failure%'
```

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

- 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 found using the query on right:

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

```
Mission Outcomes Query

%sql SELECT landing_outcome,
COUNT(landing_outcome) AS Count FROM spacetable
WHERE date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing_outcome ORDER BY
COUNT(landing_outcome) DESC
```

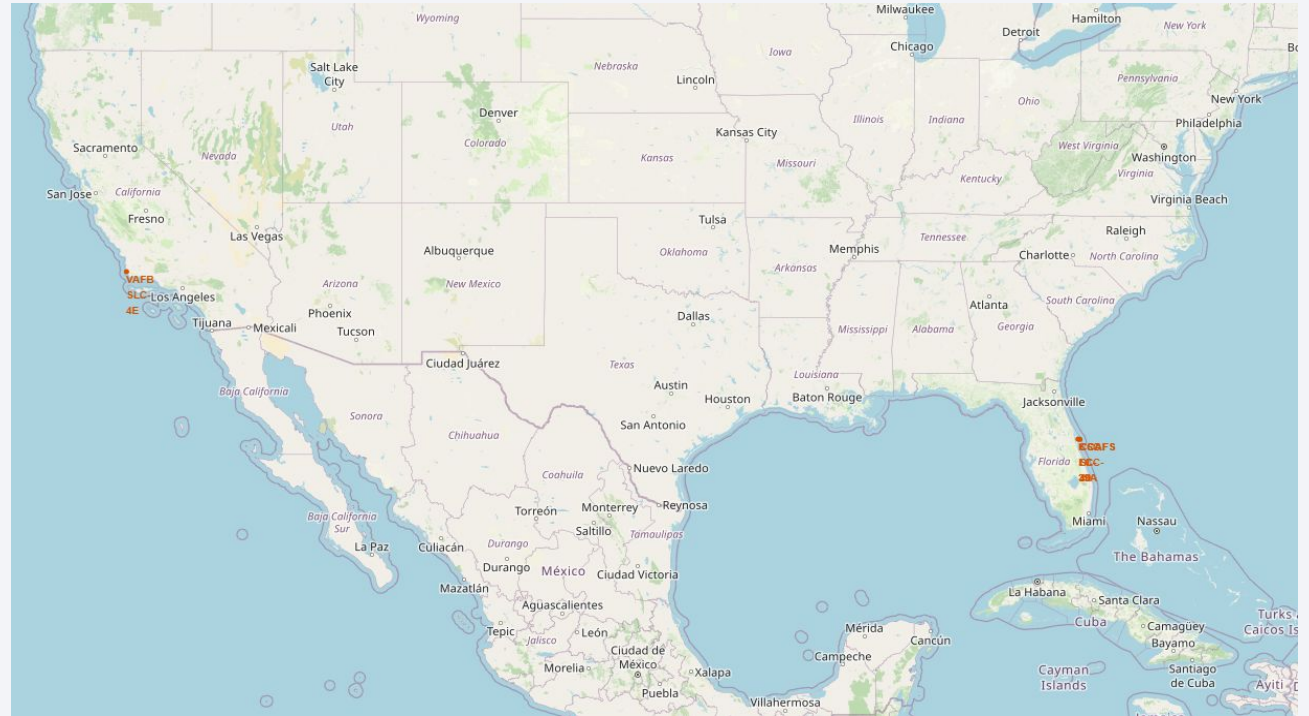
A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a dark blue sky with stars and a view of the Earth's surface from space. The Earth's surface is mostly dark, with a dense network of yellow and orange lights representing city lights at night. The lights are concentrated in a few large, bright clusters, likely representing major metropolitan areas. The overall image has a high-contrast, high-resolution appearance, typical of satellite imagery.

Section 3

Launch Sites Proximities Analysis

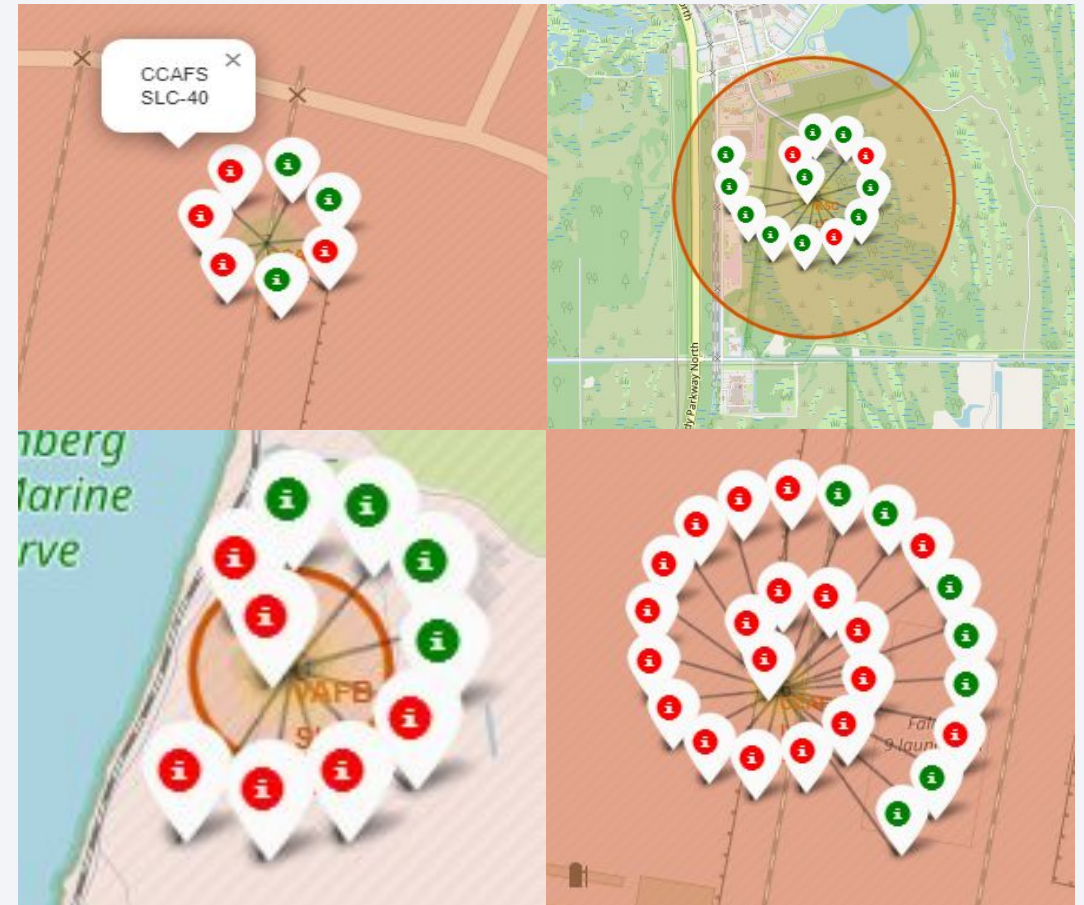
Launch Site Locations

- All launch sites were in USA
- One site is in California while the other three are in Florida
- All sites are southern regions of the country, as close as possible to the equator
- All sites are close to the ocean



Mission Outcomes of Launch Sites

- CCAFS LC-40 had the most failures while KSC had the most success
- CCAFS SLC-40 and VAFB had fewest launches and had a success rate just above 40%



Proximities of Launch Sites

- All sites were in close proximity of shorelines, railways, highways, and further away from cities
- For example, KSC had a distance of
 - 0.89 km to closest shoreline
 - 0.67 km to closest highway
 - 0.68 km to closest railway
 - 16.70 km to closest city



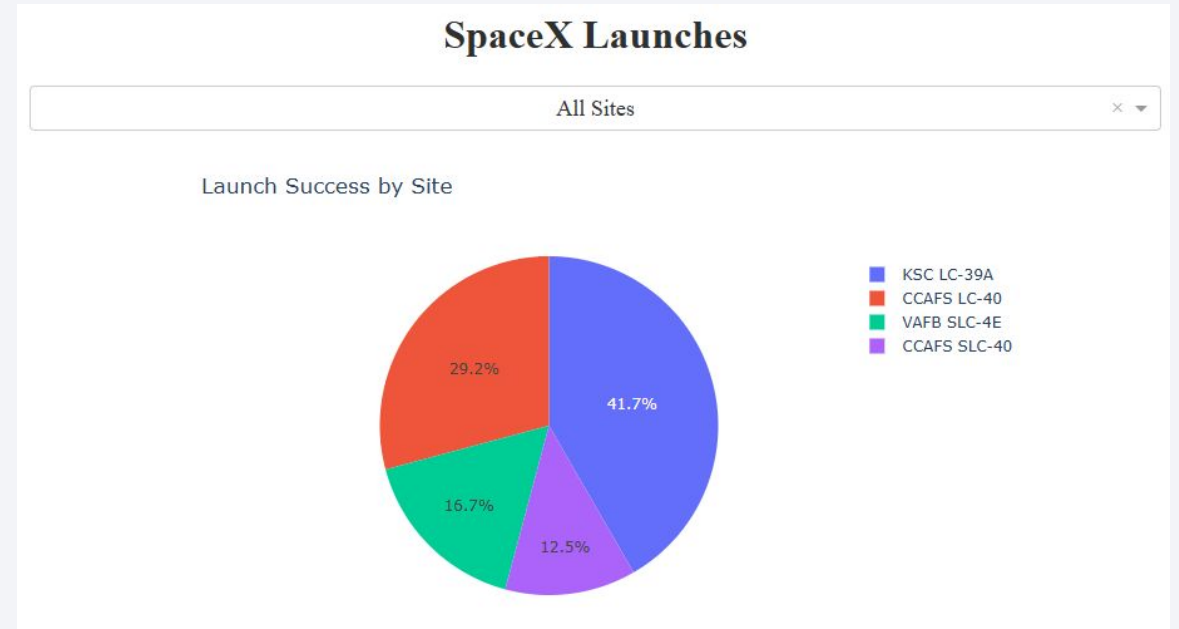


Section 4

Build a Dashboard with Plotly Dash

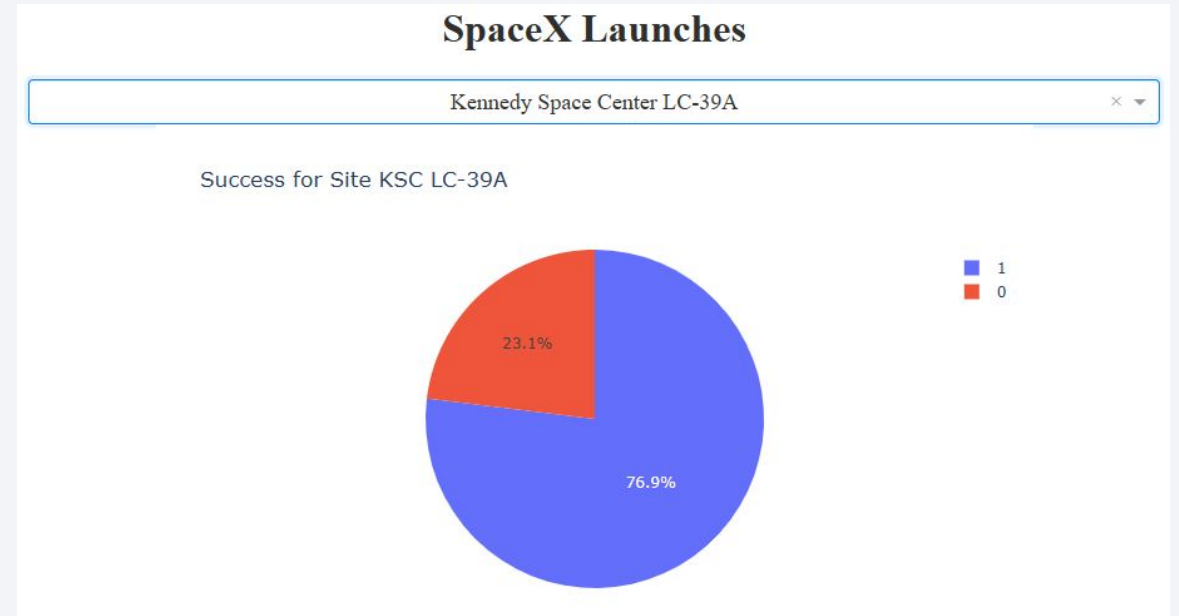
Launch Success of All Sites

- KSC had the most number of successful launches with 41.7%, followed by CCAFS LC-40, VAFB, and finally CCAFS SLC-40 with the lowest number of successful launches at 12.5%.



Launch Success by Site

- KSC had the highest success rate with 76.9%, followed by CCAFS SLC-40 at 42.9%, VAFB at 40%, and CCAFS LC-40 at 26.9%
- KSC had 10 successes and 3 failures



Effect of Payload Mass on Success

- Heaviest payloads were carried by B4 and FT boosters
- Payloads between 2000 and 4000 kg range had the most number of successful missions
- Payloads between 4000 and 8000 kg range has very low success rate
- Payloads between 0 and 2000 kg range has very low success rate

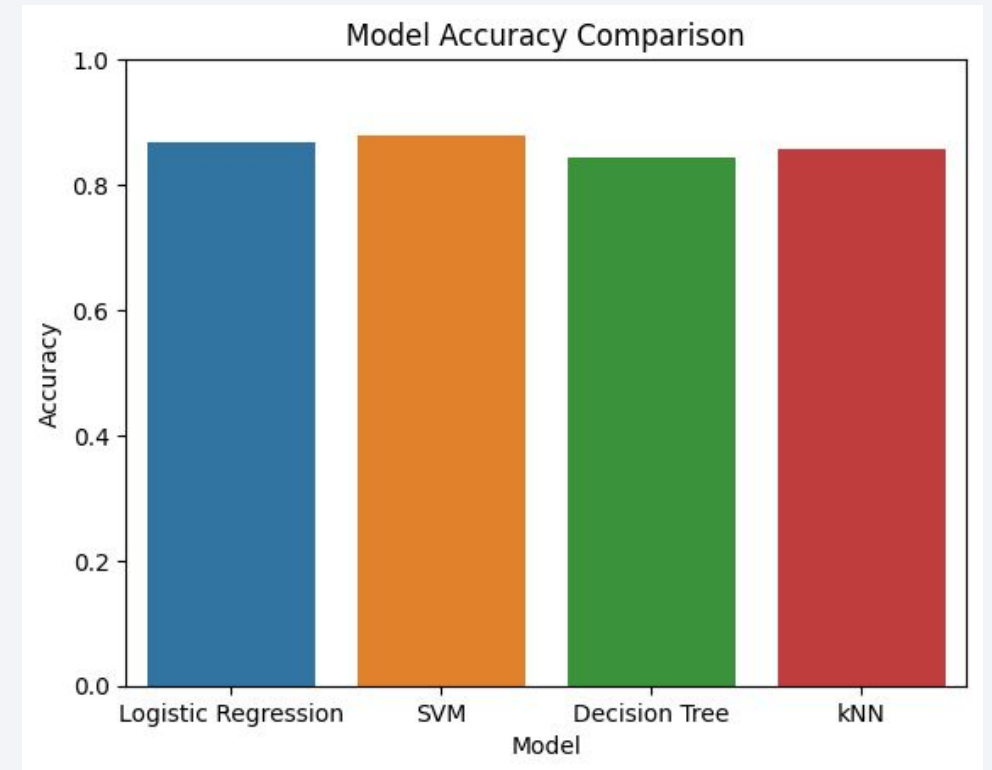


Section 5

Predictive Analysis (Classification)

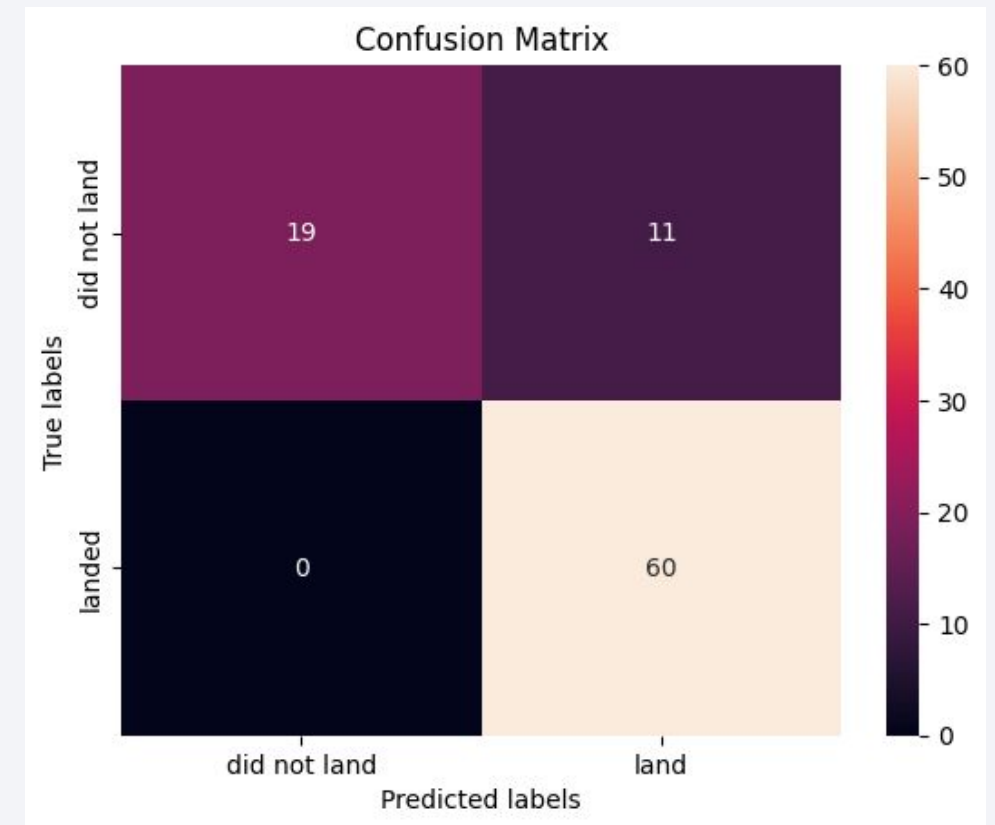
Classification Accuracy

- When tested only on the test set, all models performed the same at 83.33%
- When tested on the entire dataset, SVM had the highest accuracy with 87.78%, followed by Logistic Regression at 86.67%, kNN at 85.56%, and Decision Tree at 84.44%.



Confusion Matrix

- The model's performance on the whole dataset is on the right
- 60 True Positives, 19 True Negatives, 11 False Positives, and 0 False Negatives were recorded
- 11 FP is a bit high so the model may need further refinement
- This gives an 0.88 accuracy, 0.85 precision, 1.0 recall, and 0.92 F1 score



Conclusions

- All models performed the same on the test data and SVM performed a bit better on the whole dataset
- However, even SVM had high FP, the model may need refinement
- Dataset was relatively small considering the number of different categorical values for each feature, e.g., orbits
- Thus, more data may increase the models' performance drastically
- Some data may be irrelevant so feature engineering may be examined for which features are actually needed

Conclusions

- Any data related to estimated costs and revenue would help determining which path to take
- Overall, the data we have would help for making informed decisions but more accurate decisions require further refinements to model, and more and better data for its development
- At the end, we have gained valuable insights about launch sites, payloads, success rates, boosters, orbits, etc.

Thank you!

