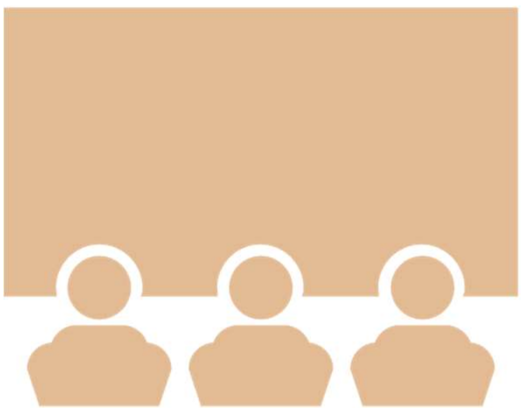
DATA SCIENCE CAPSTONE PROJECT

RUSHIKESH JAGDALE

https://github.com/Rushi717171/DataScience

03/18/2024

Outline



Executive Summary

- I collected SpaceX data, labeled successful landings, explored using SQL, visualization, and maps. Selected features, encoded variables, standardized data, optimized models, and visualized accuracy scores

- Four ML models (Logistic Regression, SVM, Decision Tree, KNN) achieved ~83.33% accuracy but tended to over-predict successful landings. More data required for improved accuracy.

Introduction

SpaceX Falcon 9 Rocket – The Verge

- **Background:**
- Commercial Space Age is Here
- Space X has best pricing ($62 million vs. $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

Problem:

- Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery

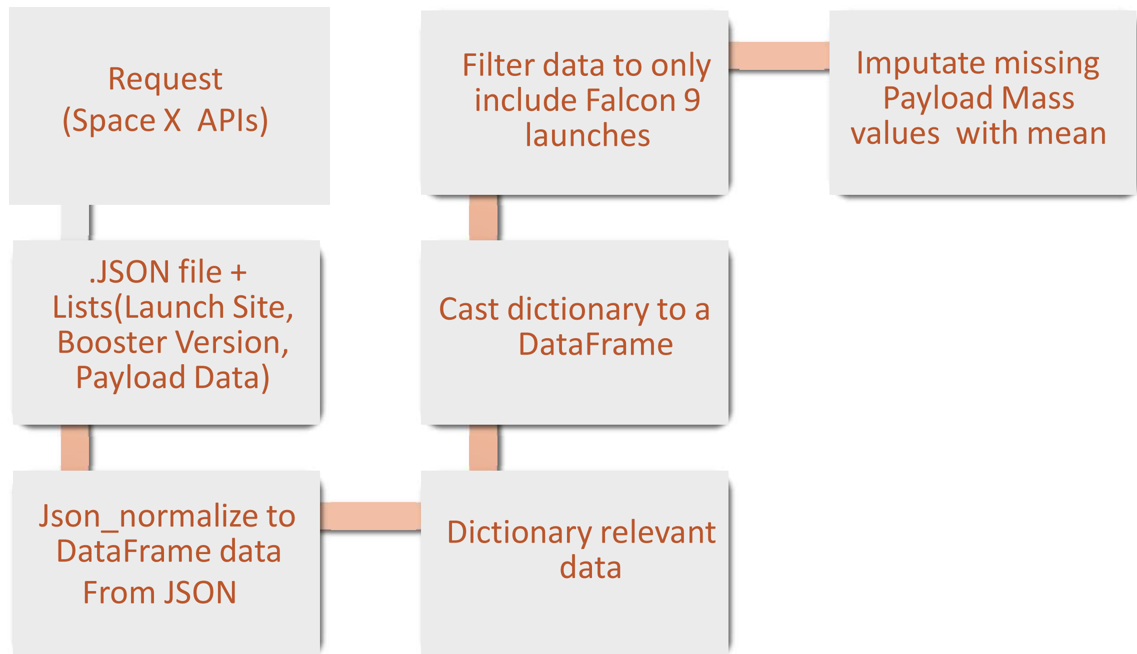
Methodology

- Data collection methodology: Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling- Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models- Tuned models using GridSearchCV

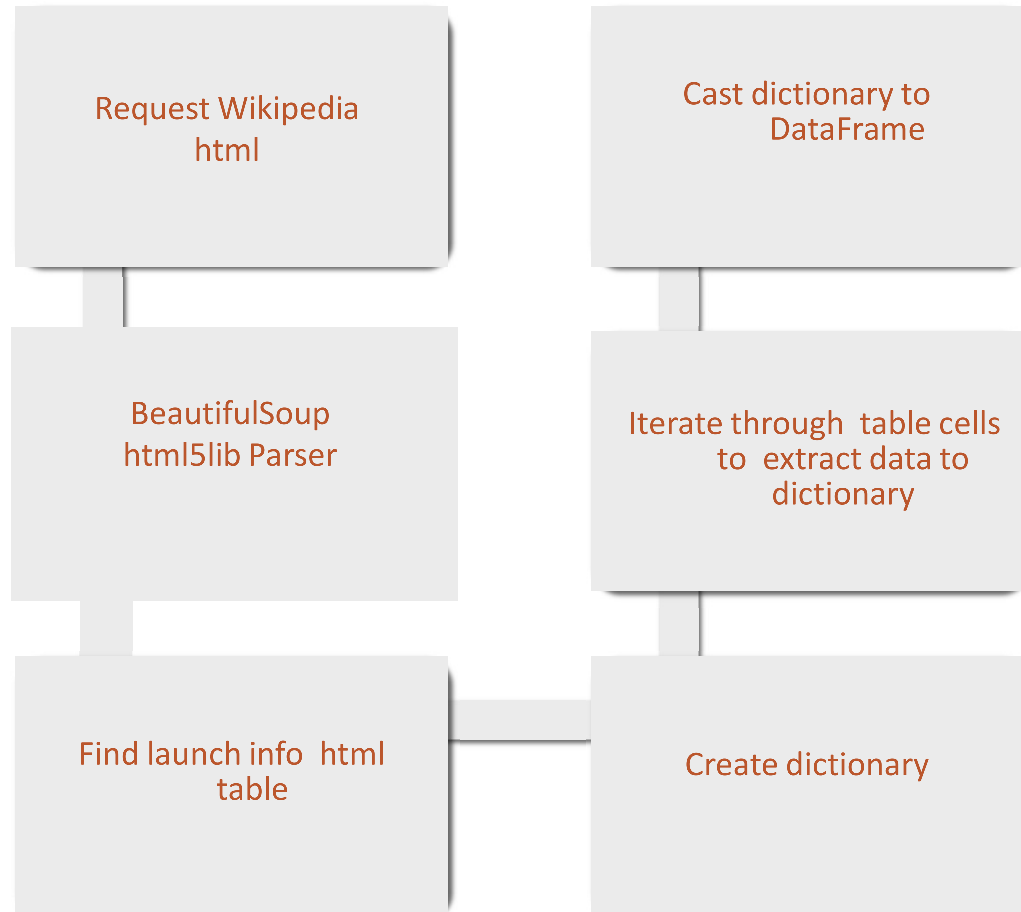
Data Collection Overview

- Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.
- The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from webscraping.
- Space X API Data Columns:
- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Wikipedia Webscrape Data Columns: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Data Collection – Web Scraping

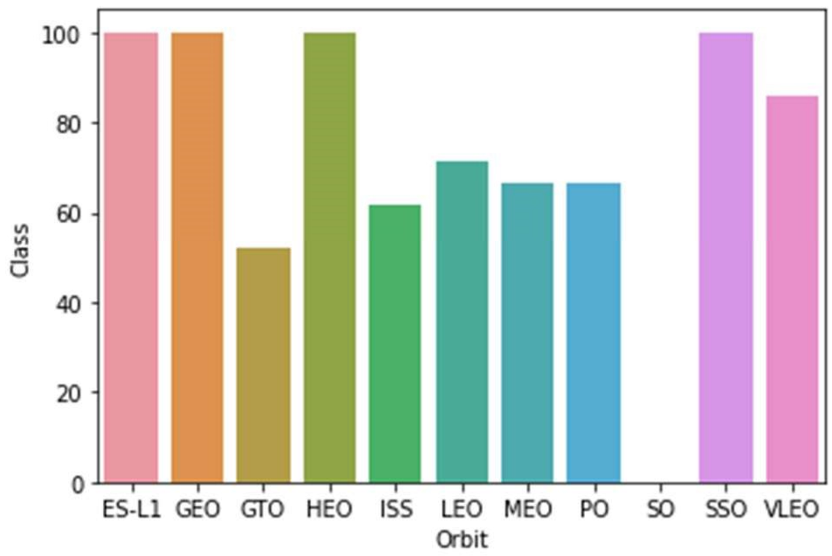


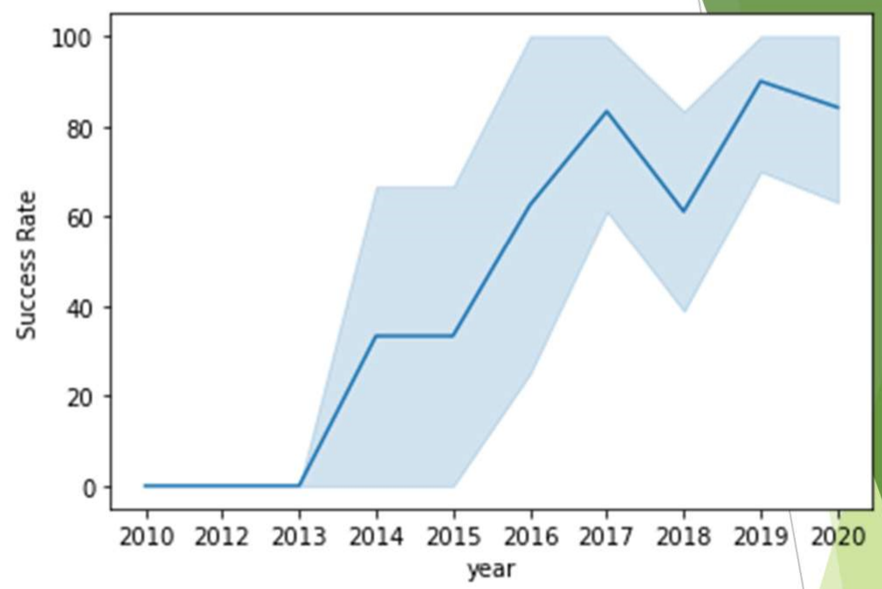
Data Wrangling

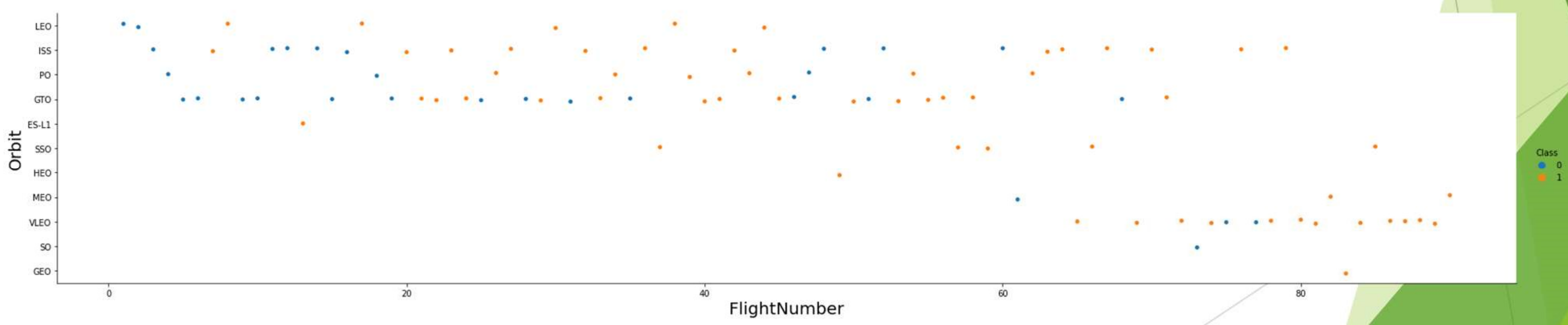
- Create a training label with landing outcomes where successful = 1 & failure = 0.
- Outcome column has two components: 'Mission Outcome' 'Landing Location'
- New training label column 'class' with a value of 1 if 'Mission Outcome' is True and 0 otherwise. Value Mapping:
- True ASDS, True RTLS, & True Ocean – set to -> 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

EDA with Data Visualization

- Conducted thorough analysis on Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year variables.
- Used diverse plots (scatter, line, bar) to explore relationships.
- Focused on key relationships like Flight Number vs. Payload Mass, Launch Site, and Orbit.
- Aiming to identify meaningful patterns for machine learning model training.
- Aimed to identify meaningful patterns for machine learning model training.
- Ensured data readiness for effective model development.


100
80
60
40
20
0
Class
ES-L1 GEO GTO HEO ISS LEO MEO PO SO SSO VLEO
Orbit


100
80
60
40
20
0
Success Rate
2010 2012 2013 2014 2015 2016 2017 2018 2019 2020
year


LEO
ISS
PO
GTO
ES-L1
SSO
HEO
MEO
VLEO
SO
GEO
Orbit
0
20
40
60
80
FlightNumber
Class
0
1

EDA with SQL

Loaded data set into IBM DB2 Database.

Queried using SQL Python integration.

Queries were made to get a better understanding of the dataset.

Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but les

```
%sql select booster_version from SPACEXDATASET where (mission_outcome like 'Success')
AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome like 'Success (drone ship)')
```

```
* ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdom
LUDB
Done.
```

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

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site
from SPACEXDATASET where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

```
* ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/
LUDB
Done.
```

MONTH	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Task 10

Rank the count of successful landing_outcomes between the date 2010-06-04 and 20

```
%sql select landing__outcome, count(*) as count from SPACEXDATASET
where Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP by landing__outcome ORDER BY count Desc
```

```
* ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgt
LUDB
Done.
```

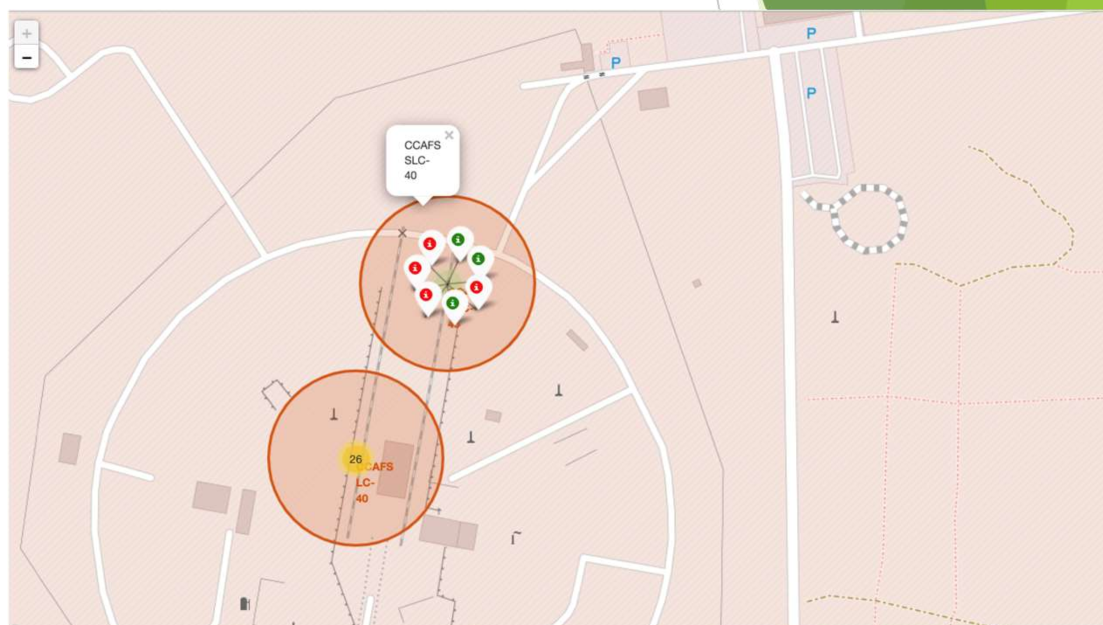
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Build an interactive map with Folium:

Folium maps display Launch Sites, successful and unsuccessful landings, and proximity examples to key locations: Railway, Highway, Coast, and City.

The maps provide insight into the rationale behind the selection of launch site locations.

They also visualize successful landings in relation to their geographic location, aiding in location analysis.



Build a Dashboard with Plotly Dash

The dashboard features a pie chart and a scatter plot.

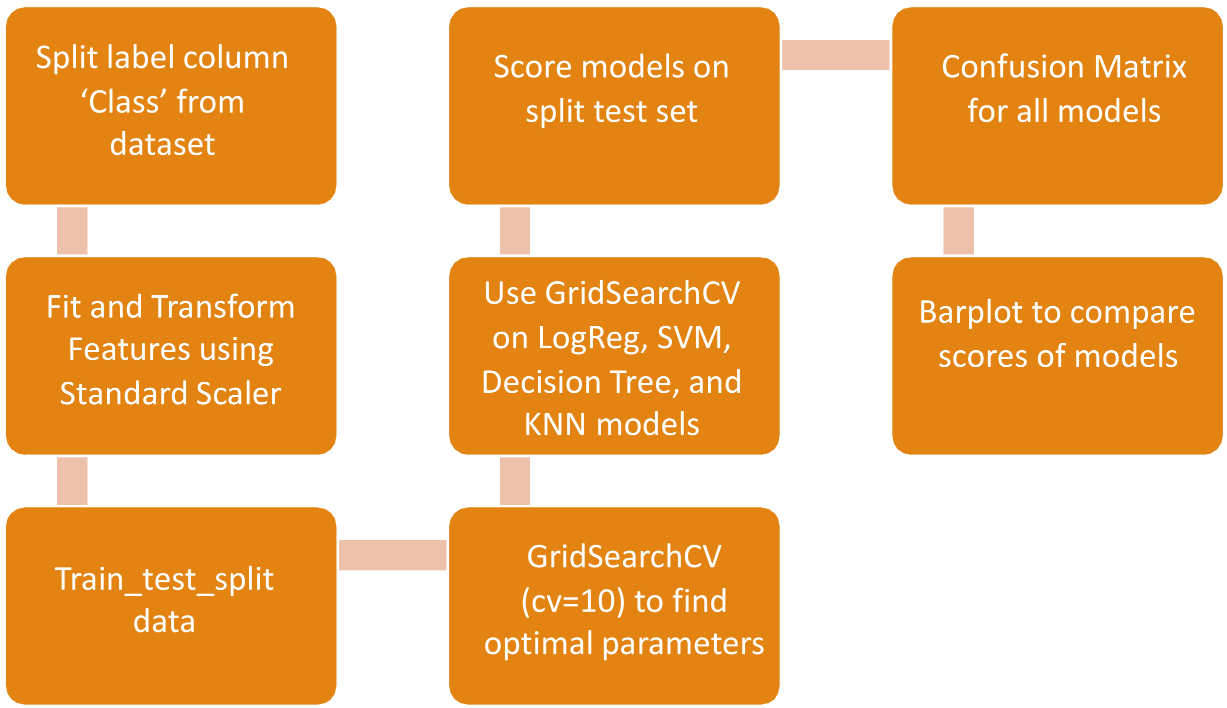
The pie chart allows selection to display the distribution of successful landings across all launch sites or individual launch site success rates.

The scatter plot allows selection of either all sites or individual sites, with a slider for payload mass ranging from 0 to 10000 kg.

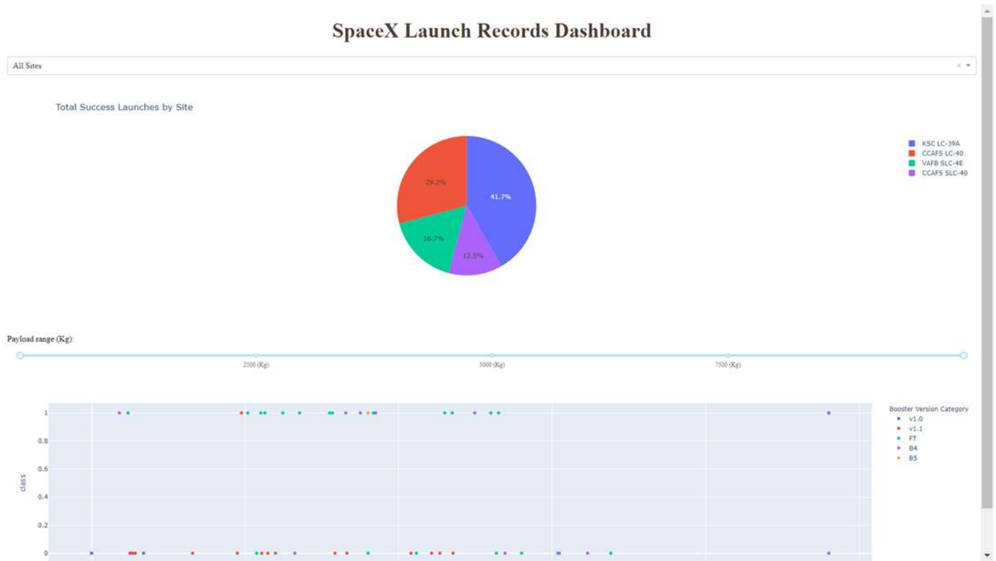
The pie chart visualizes launch site success rates.

The scatter plot helps in observing variations in success rates across launch sites, payload mass, and booster version categories.

Predictive analysis (Classification)



Results

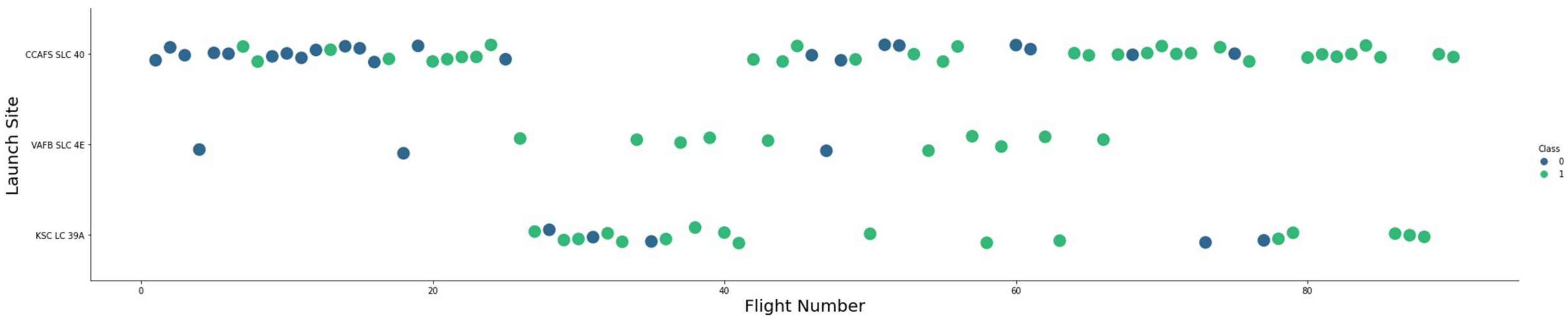


This preview showcases the Plotly dashboard, which includes the outcomes of Exploratory Data Analysis (EDA) through visualization and SQL, an Interactive Map using Folium, and the results of our model with an accuracy of approximately 83%

EDA with Visualization

- EXPLORATORY DATA ANALYSIS WITH SEABORN PLOTS

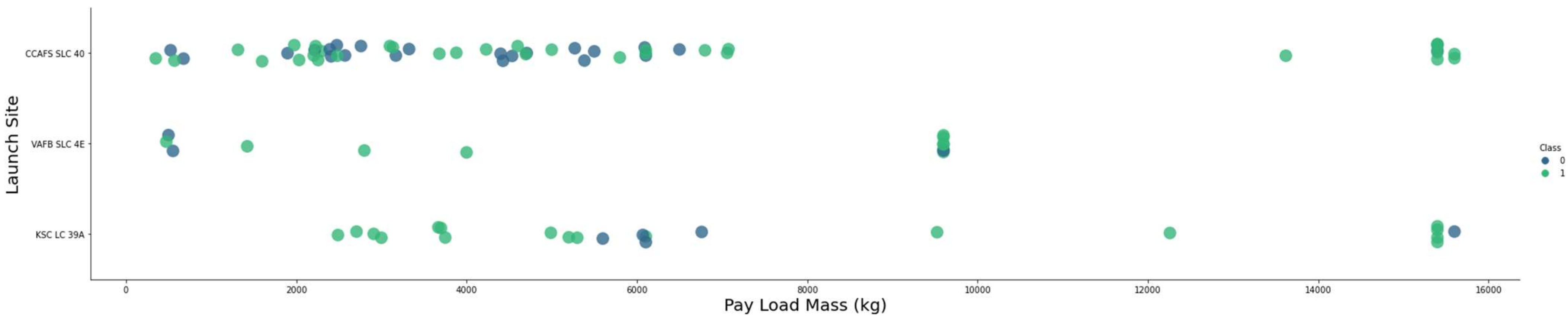
Flight Number Vs Launch Site



Green indicates successful launch; Purple indicates unsuccessful launch.

The graph indicates a gradual rise in success rates over time, as depicted by the increase in Flight Number. There seems to be a notable breakthrough around flight number 20, leading to a significant spike in success rates. Additionally, it is evident that CCAFS is the primary launch site, given its higher volume compared to others.

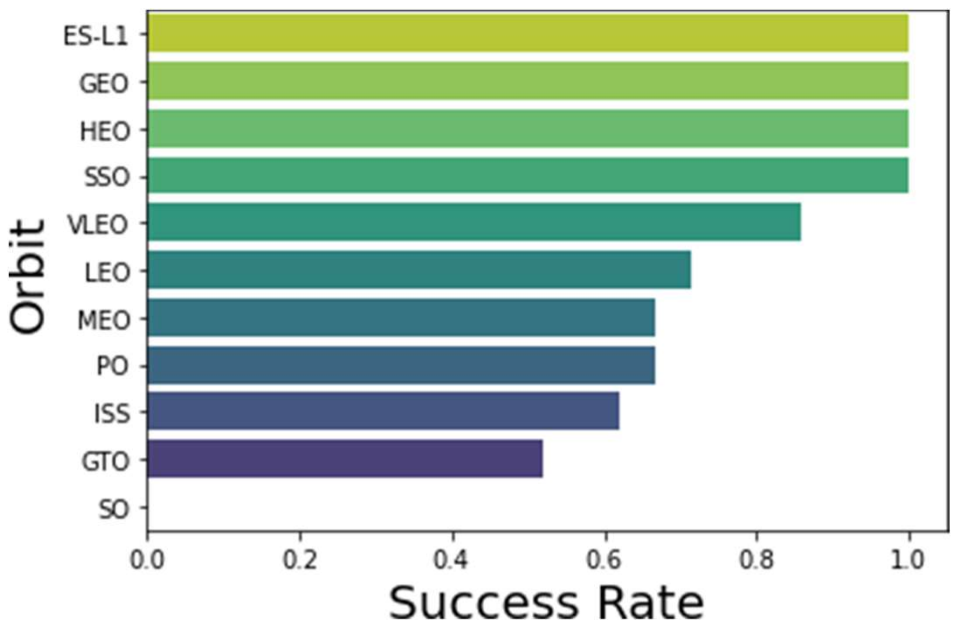
Payload Vs Launch Site



Green indicates successful launch; Purple indicates unsuccessful launch.

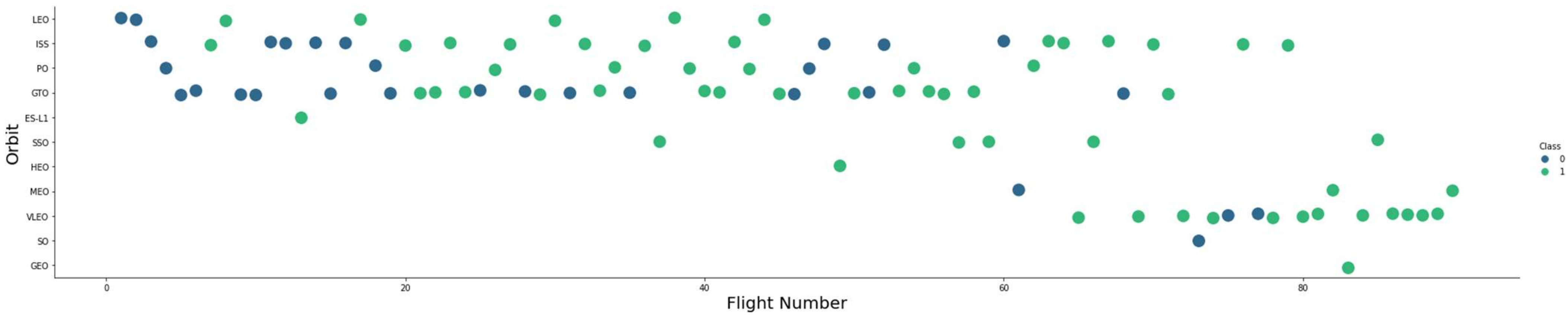
The analysis reveals that the majority of payload masses fall within the range of 0-6000 kg. Furthermore, it appears that different launch sites employ varying payload masses for their missions.

Success rate Vs Orbit type



ES-L1, GEO, and HEO, each with a sample size of 1, achieved a 100% success rate. Similarly, SSO, with a sample size of 5, also attained a 100% success rate. VLEO, with 14 attempts, demonstrated a respectable success rate. In contrast, SO, with a single attempt, experienced a 0% success rate. Notably, GTO, with the largest sample size of 27, achieved a success rate of approximately 50%

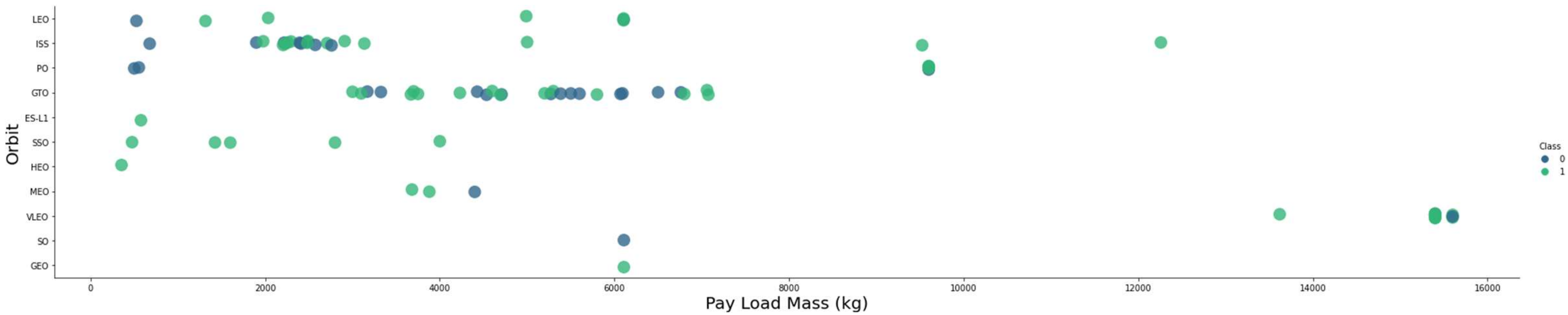
Flight Number Vs Orbit type



Green indicates successful launch; Purple indicates unsuccessful launch.

SpaceX's launch orbit preferences have changed over time, correlating with Flight Number and Launch Outcome. Initially focusing on LEO missions, success rates were moderate. Recent launches show a return to VLEO, suggesting better performance in lower or Sun-synchronous orbits.

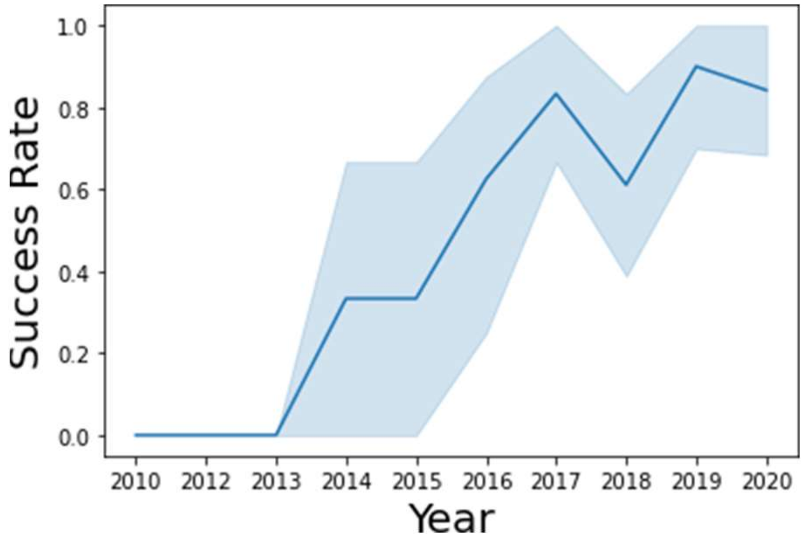
Payload Vs Orbit type



Green indicates successful launch; Purple indicates unsuccessful launch.

Payload mass appears to correlate with the orbit type, with Low Earth Orbit (LEO) and Sun-Synchronous Orbit (SSO) exhibiting relatively low payload masses. Conversely, the Very Low Earth Orbit (VLEO), which is among the most successful orbits, tends to have payload mass values at the higher end of the range.

Launch Success Yearly Trend



95% confidence interval (light blue shading)

Success rates have generally shown an upward trend since 2013, albeit with a slight dip observed in 2018. In recent years, success rates have stabilized at around 80%.

EDA with SQL

EXPLORATORY DATA ANALYSIS WITH SQL DB2

INTEGRATED IN PYTHON WITH SQLALCHEMY

All Launch Site Names

In [4]:
```
%%sql
SELECT UNIQUE LAUNCH_SITE
FROM SPACEXDATASET;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f
Done.
```

Out[4]:

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

- Retrieve unique launch site names from the database.
- There are likely data entry errors, as "CCAFS SLC-40" and "CCAFSSLC-40" likely represent the same launch site.
- "CCAFS LC-40" was the previous name. Therefore, there are likely only three unique launch site values:
- "CCAFS SLC-40", "KSC LC-39A", and "VAFB SLC-4E".

Launch Site Names Beginning with `CCA`

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Retrieve the first five entries from the database where the Launch Site name begins with "CCA".

Total Payload Mass from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

This query sums the total payload mass in kg where NASA was the customer.

CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-8
Done.
```

avg_payload_mass_kg
2928

This query calculates the average payload mass or launches which used booster version F9 v1.1

Average payload mass of F9 1.1 is on the low end of our payload mass range

First Successful Ground Pad Landing Date

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

This query returns the first successful ground pad landing date.

First ground pad landing wasn't until the end of 2015.

Successful landings in general appear starting 2014.

Successful Drone Ship Landing with Payload Between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively.

Total Number of Each Mission Outcome

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

This query returns a count of each
mission outcome.

SpaceX appears to achieve its mission outcome nearly 99% of the time.

This means that most of the landing
failures are intended.

Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

Boosters that Carried Maximum Payload

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

This query returns the booster versions that carried the highest payload mass of 15600 kg.

These booster versions are very similar and all are of the F9 B5 B10xx.x variety.

This likely indicates payload mass correlates with the booster version that is used.

2015 Failed Drone Ship Landing Records

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

There were two such occurrences.

Ranking Counts of Successful Landings Between 2010-06-04 and 2017-03-20

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

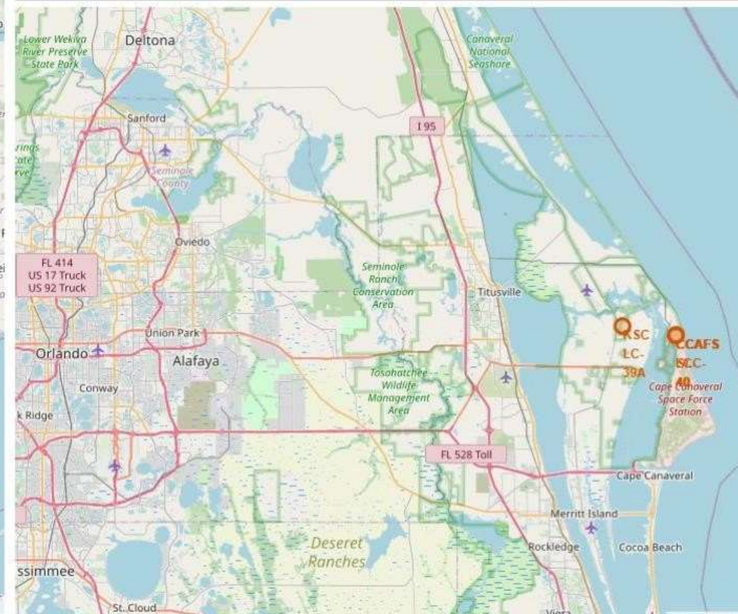
This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.

There are two types of successful landing outcomes: drone ship and ground pad landings.

There were 8 successful landings in total during this time period

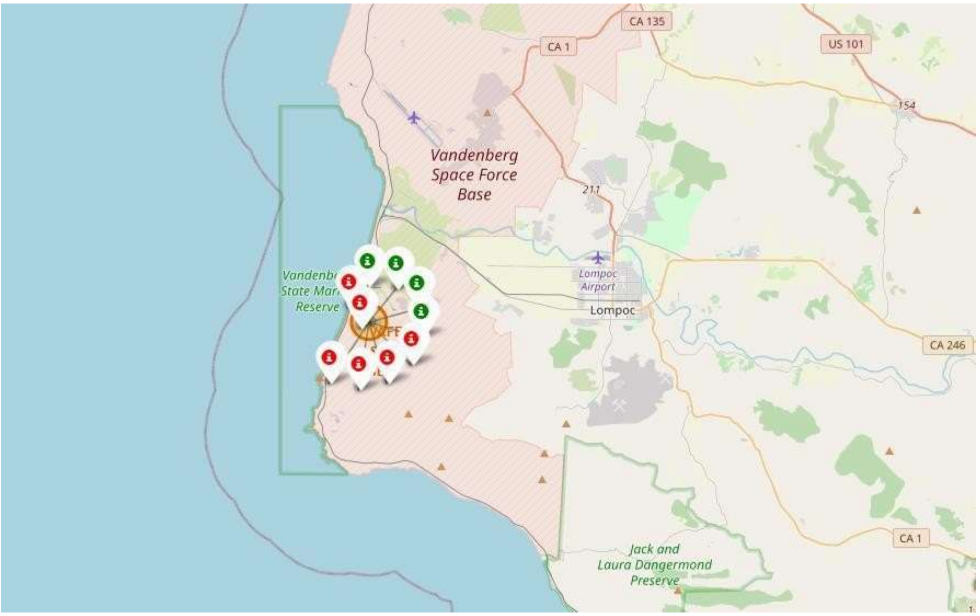
Interactive Map with Folium

Launch Site Locations



The left map shows all launch sites relative to the US map, while the right map focuses on the two Florida launch sites due to their proximity, all of which are situated near the ocean

Color-Coded Launch Markers



Clusters on the Folium map are clickable, revealing individual successful (green icon) and failed (red icon) landings. For instance, VAFB SLC-4E exhibits 4 successful landings and 6 failed landings in this example.

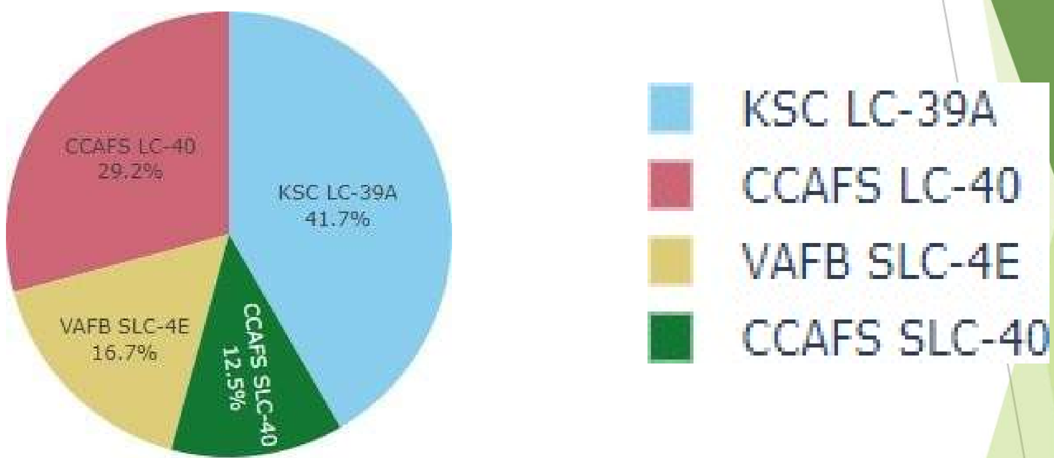
Key Location Proximities

Launch sites like KSC LC-39A are positioned near railways for supply transportation, highways for accessibility, and coastlines to mitigate risks of launch failures in densely populated areas.

Build a Dashboard with Plotly

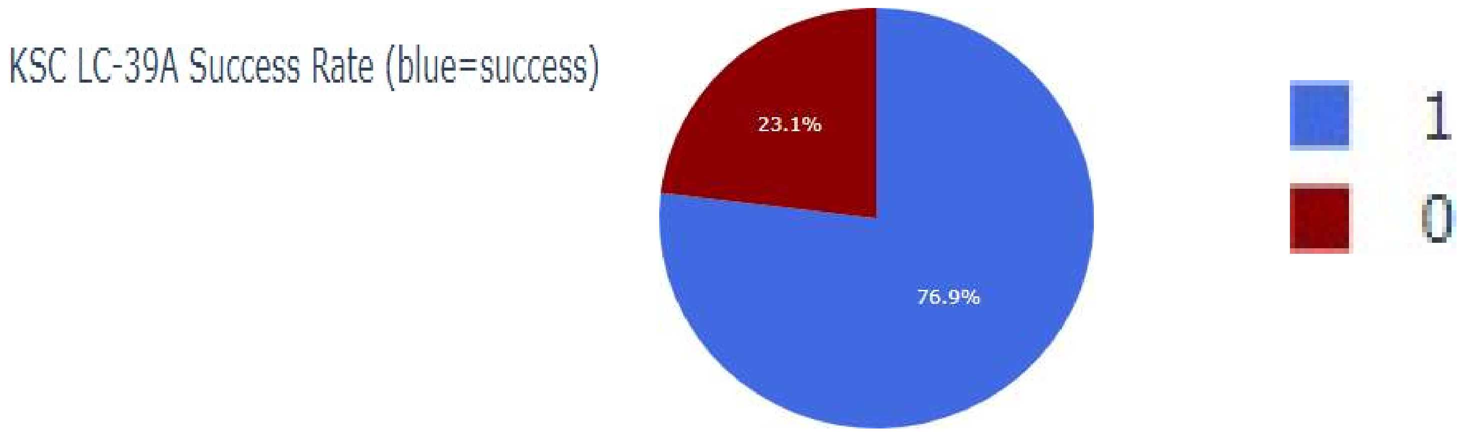
Dash

Successful Launches Across Launch Sites



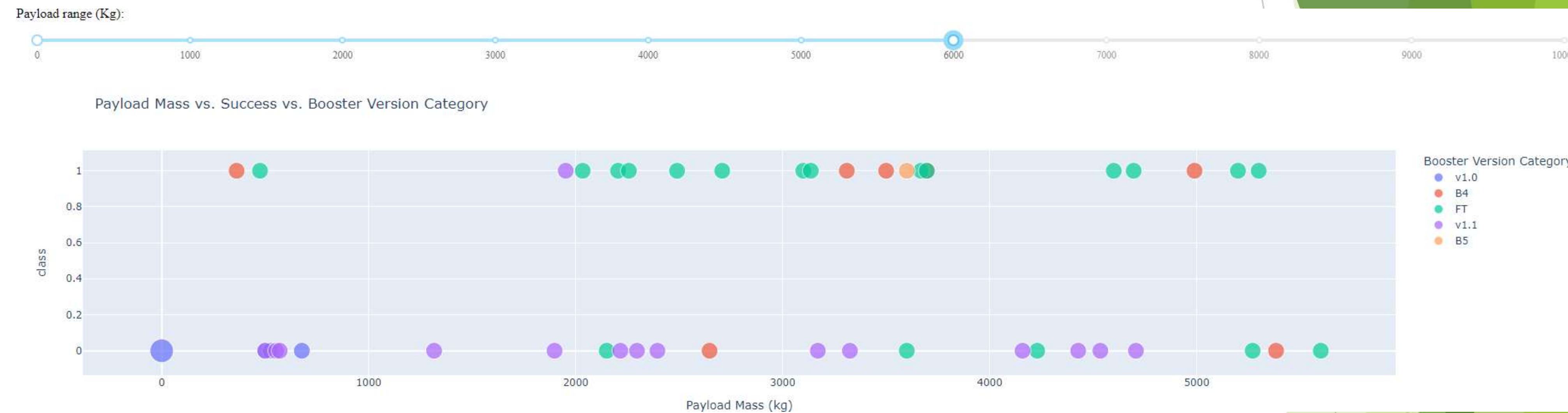
CCAFS LC-40, now known as CCAFS SLC-40, and KSC have an equal number of successful landings. However, most successful landings at CCAFS occurred before the name change.
VAFB has the smallest proportion of successful landings, possibly due to a smaller sample size and the increased difficulty of launching on the west coast.

Highest Success Rate Launch Site



KSC LC-39A boasts the highest success rate, with 10 successful landings and 3 failed landings.

Payload Mass vs. Success vs. Booster Version Category

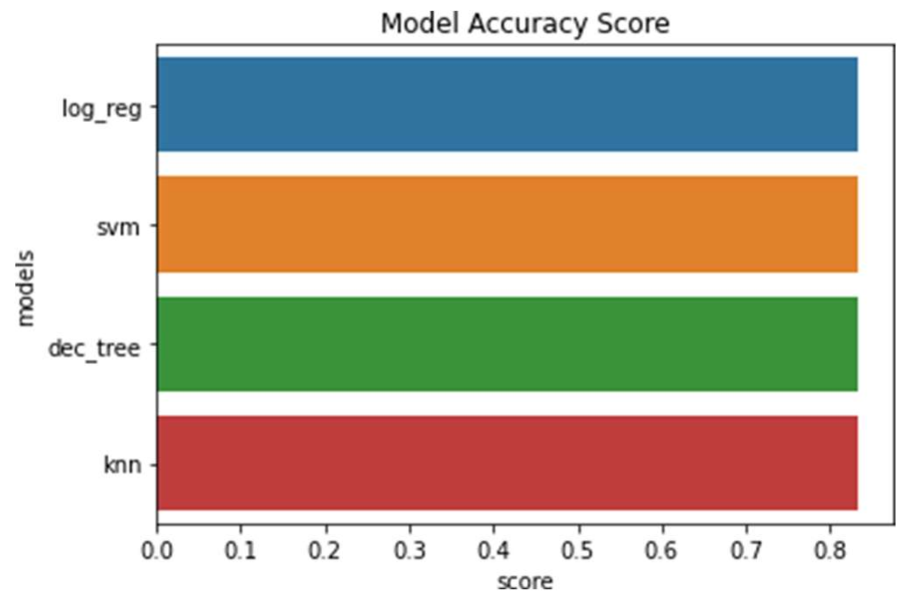


- The Plotly dashboard's payload range selector is set from 0 to 10000 instead of the max payload of 15600.
- "Class" indicates 1 for successful landings and 0 for failures.
- The scatter plot includes booster version category as color and number of launches as point size.
- Within the 0 to 6000 payload range, two failed landings with payloads of zero kg are observed, which is noteworthy.

Predictive Analysis (Classification)

GRIDSEARCHCV(CV=10) ON
LOGISTIC REGRESSION, SVM, DECISION TREE, AND KNN

Classification Accuracy



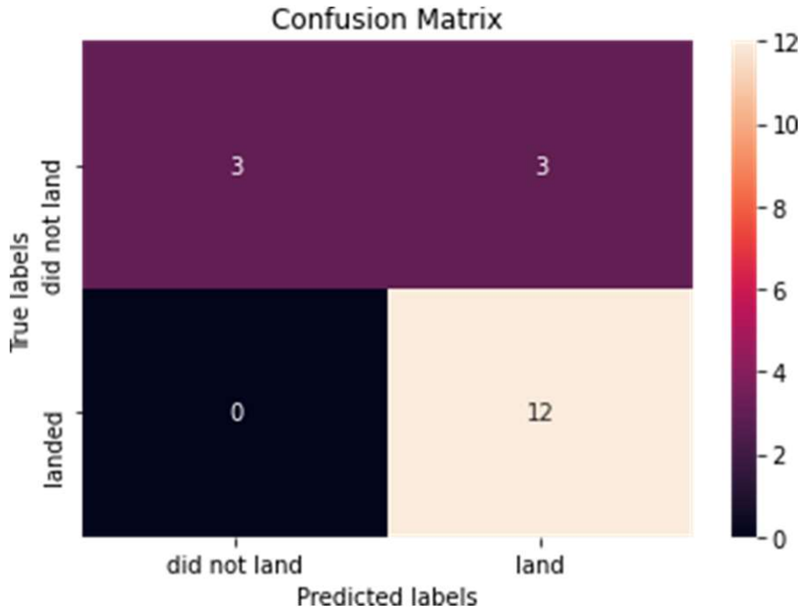
All models achieved an 83.33% accuracy rate on the test set.

However, the test set's small size, consisting of only 18 samples, can result in considerable accuracy variance.

The Decision Tree Classifier model showed notable variability in accuracy across repeated runs.

To reliably determine the best model, acquiring more data is recommended.

Confusion Matrix



Correct predictions are on a diagonal from top left to bottom right.

Observation: The models tend to overpredict successful landings, as indicated by the false positives.

Confusion Matrix Analysis:

True Positive (TP): Models predicted 12 successful landings correctly.

True Negative (TN): Models predicted 3 unsuccessful landings correctly.

False Positive (FP): Models incorrectly predicted 3 successful landings when the true label was unsuccessful landings.

False Negative (FN): Not specified in the provided text.

CONCLUSION

- Task: Develop a machine learning model for Space Y to predict successful Stage 1 landings and potentially save $100 million USD.
- Data Sources: Utilized data from a public SpaceX API and web scraping SpaceX Wikipedia page.
- Data Handling: Created data labels and stored data into a DB2 SQL database.
- Visualization: Developed a dashboard for visualization purposes.
- Machine Learning Model: Built a machine learning model with an accuracy of 83%.
- Application: Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch, aiding in decision-making.
- Future Improvements: Suggested collecting more data to further improve accuracy and determine the best machine learning model.

APPENDIX

GitHub repository URL: https://github.com/Rushi717171/DataScience

Instructors: Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo

Special Thanks to All Instructors:

https://www.coursera.org/professional-certificates/ibm-data-science?#instructors