
IBM Developer
SKILLS NETWORK
Winning Space Race
with Data Science
John Pauline Pineda
October 14, 2023

Outline

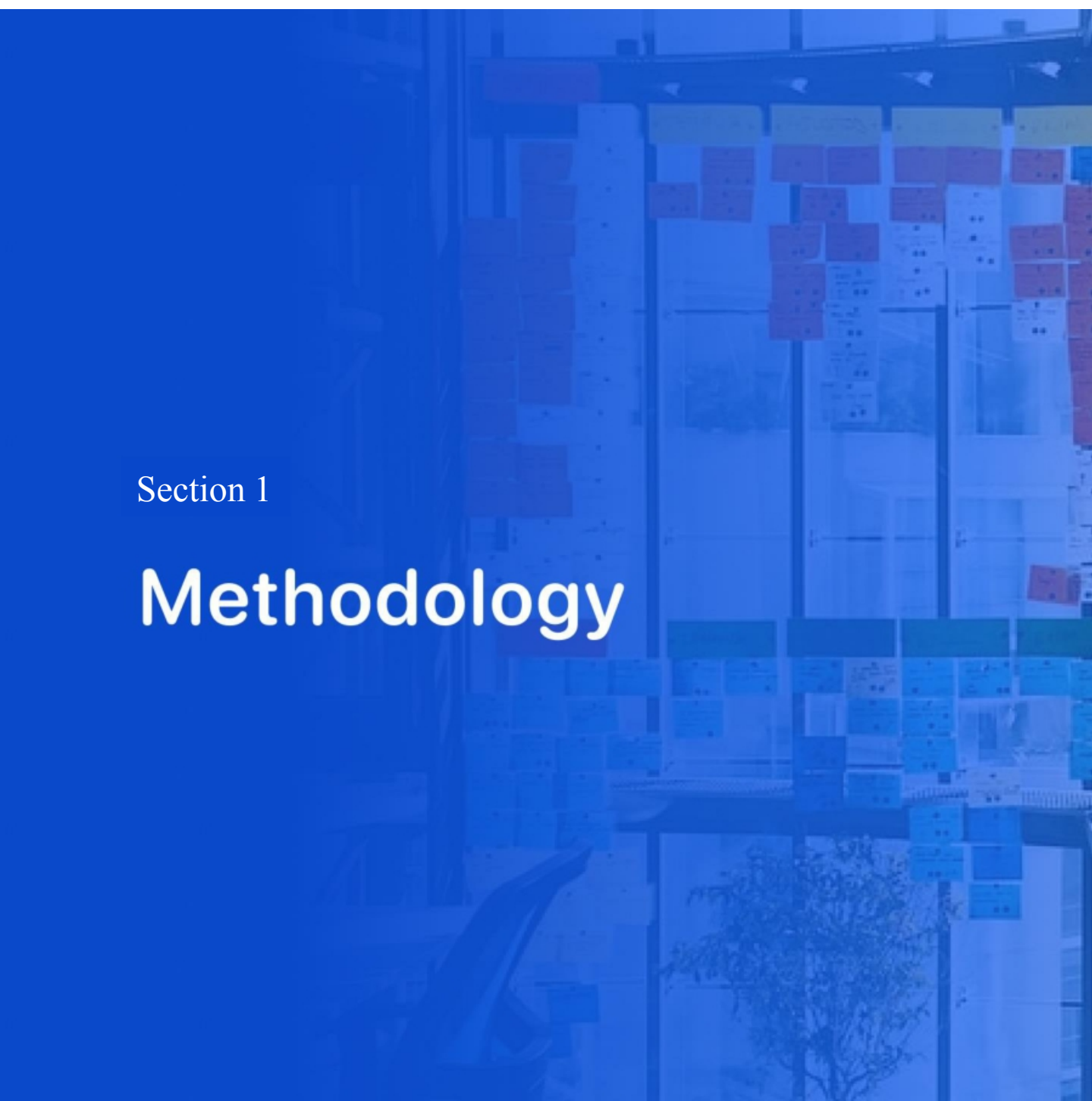
- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Methodologies explored in the study included the fol
  - Data collection using REST API and web scraping
  - Pre-processing methods
  - Exploratory data analysis using visualization and SQL
  - Interactive visual analytics using Folium and Plotly Dash
  - Predictive analysis encompassing model development, e
- Study results provided the following:
  - Complete and pre-processed data appropriate for analys
  - Determined effects of features to successful landing outc
  - Determined effects of spatial features to successful laun
  - Determined effects of features to successful launch outc
  - Classification model for predicting landing outcomes wit

Introduction

- Space exploration has traditionally been associated w
  under the leadership of Elon Musk, has emerged as a
- Notably, SpaceX has revolutionized the space explora
  cost savings and cost-efficiency that sets it apart fro
- SpaceX pioneered the reusable rocket technology. Th
  reusable first stage, allows multiple launches with the
  portions of their launch vehicles, SpaceX dramaticall
  overall mission expenses.
- This capstone project generally aims to investigate th
  of SpaceX's first stage through exploratory data anal
  thereby allowing reuse for multiple launches and ach
  - In particular, a classification model will be formulated w
    of the Falcon 9 rocket's successful landing outcome fro
    predictors, while delivering accurate predictions when a



Section 1

Methodology

Methodology

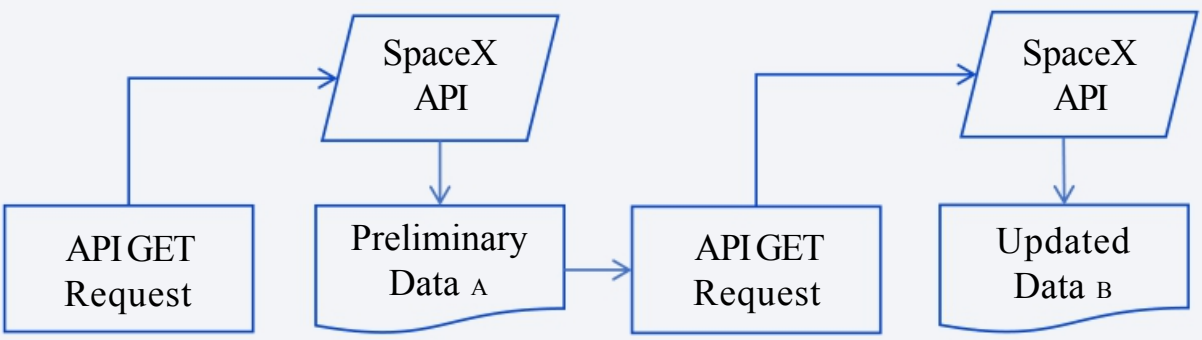
Executive Summary

- Data collection methodology
  - Sources: SpaceX REST API + web scraping
- Perform data wrangling
  - Pre-Processing: row + column filtering, missing data imp
- Perform exploratory data analysis (EDA) using visuali
- Perform interactive visual analytics using Folium and
- Perform predictive analysis using classification model
  - Classification Models: decision tree, k-nearest neighbors
  - Train-Test Ratio: 80% train and hyperparameter tuning
  - Model Performance Evaluation: accuracy on train data w

Data Collection

- SpaceX launch data was gathered from 2 sources:
  - Source 1: SpaceX REST API containing data about Space
    rocket used, payload delivered, launch specifications, lan
    among others.
  - Source 2: web scraping of Falcon 9 launch data from relev
- Data collection process flowchart is presented as foll

Option 1: SpaceX REST API

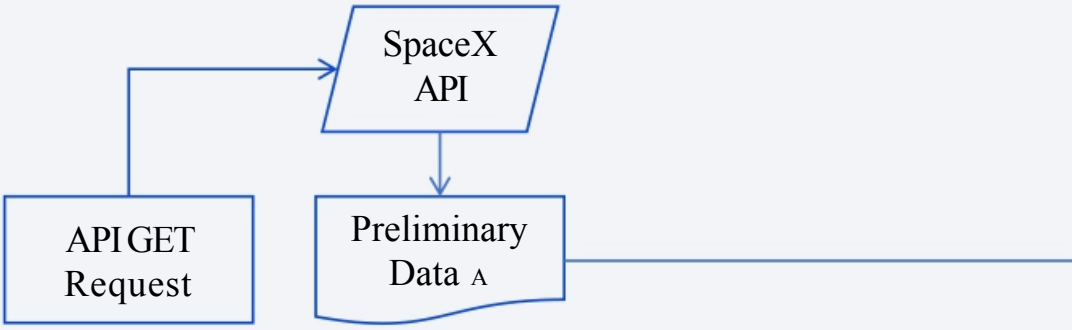


Data: A Rocket Launch B Rocket Launch + Booster + Launch Pad + Payload +

Data Collection – SpaceX API

- Data collection process flowchart with SpaceX REST

Option 1: SpaceX REST API



API GET Request Steps

- Define SpaceX URL for Rocket Launch data
- Apply API GET request on SpaceX URL
- Decode the API response as JSON object
- Transform JSON object to a Pandas data frame
- Perform necessary feature sub-setting and row filtering
- Generate preliminary data from filtered data frame
- Create additional empty global variable list

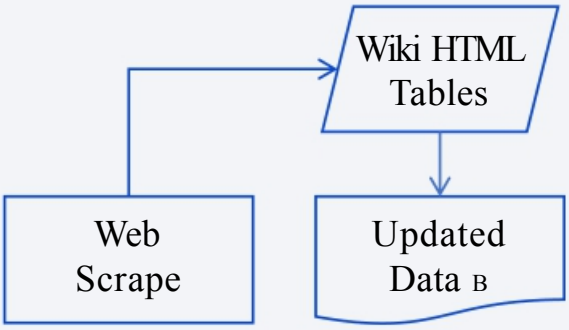
Data: A Rocket Launch B Rocket Launch + Booster + Launch Pad + Payload +

- GitHub URL of the completed SpaceX API calls Pyth

Data Collection – Scraping

- Data collection process flowchart with web scraping

Option 2: World Wide Web



Web Scrape Steps

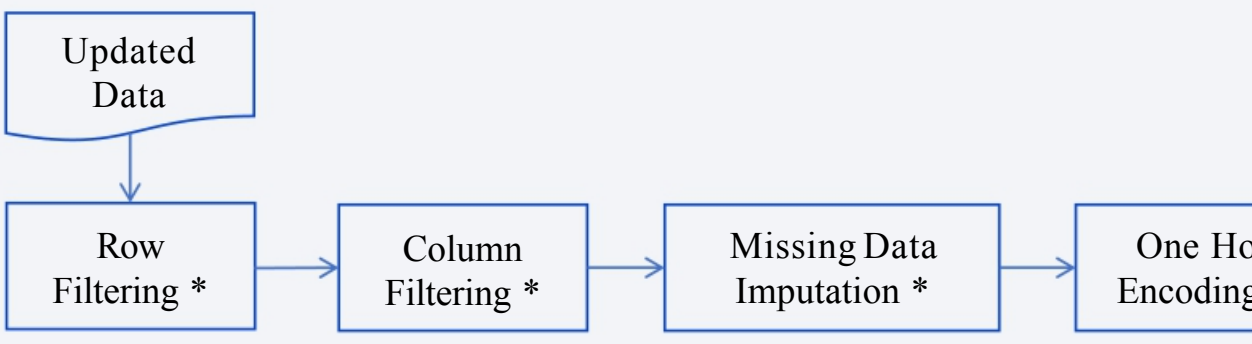
- Define WWW URL for Wiki HTML tables for Booster + Launch Pad + Payload + Core data
- Apply API GET request on WWW URL
- Decode the API response as BeautifulSoup object
- Extract all column names and row information from the BeautifulSoup object HTML tables
- Create a Pandas data frame from the parsed HTML tables

Data: в Rocket Launch + Booster + Launch Pad + Payload + Core

- GitHub URL of the completed web scraping Python n

Data Wrangling

- Data wrangling was applied on the updated data set w
  and transforming raw data into a format suitable for
- Data wrangling process flowchart is presented as foll



* Data wrangling steps which were incorporated under the data collection Py

- GitHub URL of the data collection Python notebook w
  is provided here.
- GitHub URL of the completed data wrangling Python

EDA with Data Visualization

- Exploratory data analysis involved formulating the a
  and visualize the effects of the different features on l
  - Categorical Plot: Effect of payload mass and flight numb
  - Categorical Plot: Effect of launch site and flight number
  - Categorical Plot: Effect of payload mass and launch site
  - Categorical Plot: Effect of flight number and orbit to lan
  - Categorical Plot: Effect of payload mass and orbit to lan
  - Bar Plot: Effect of orbit to landing success rate
  - Line Plot: Effect of year to landing success rate
- GitHub URL of the completed EDA with data visualiz

EDA with SQL

- Exploratory data analysis involved drill-down investig
  - Distinct Selection: SELECT statement + DISTINCT expres
  - String Patterns: SELECT statement + WHERE clause + LI
  - Functions: SELECT statement + SUM|AVG |MIN function
  - Multiple Conditions: SELECT statement + WHERE clause
  - Grouping Result Sets: SELECT statement + COUNT functi
  - Embedded Subquery: SELECT statement + WHERE clause
  - Substring Extraction: SELECT statement + SUBSTR funct
  - Range Conditions: SELECT statement + COUNT function
  - Sorting Result Sets: SELECT statement + COUNT function
- GitHub URL of the completed EDA with SQL Python

Build an Interactive Map with

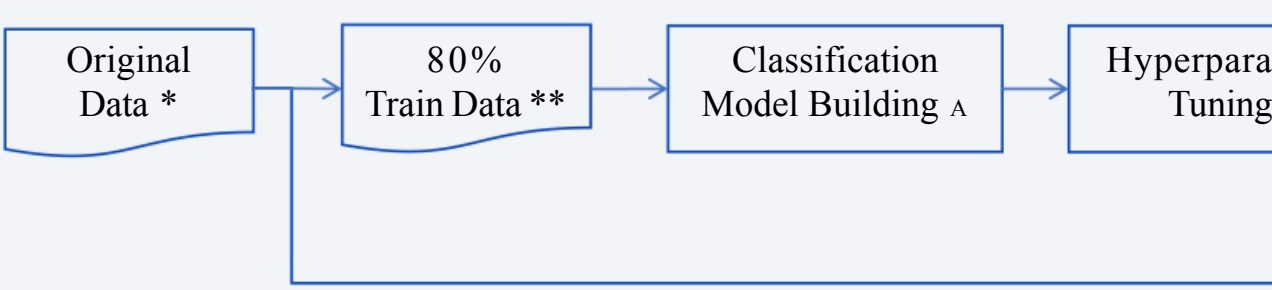
- Proximity analysis of launch data was conducted usin
  - Circles: Indicates the launch sites
  - Markers: Indicates the names of the launch sites as a po
  - Marker Cluster: Indicates individual launch record for ea
  - Mouse Position: Indicates latitude and longitude informa
  - Lines: Connects the launch sites to identified landmarks (
  - Distance Circle: Indicates the distance between the launc
- GitHub URL of the completed interactive Folium map
- NOTE: Despite enabling the 'Trust Notebook' setting the uploaded python notebook is viewed through Gi and map outputs were provided here as an alternativ

Build a Dashboard with Plotly

- Interactive visual analytics involved formulating the
  investigate the individual and combined effects of di
  - Pie Chart: Contribution of each launch site on the combi
  - Pie Chart: Distribution of the launch outcomes for each l
  - Categorical Plot: Effect of payload mass and booster ver
  - Range Slider: Effect of changing the payload mass range
- GitHub URL of the completed Plotly Dash Python co

Predictive Analysis (Classificati

- Predictive analysis involved the process steps describ
  - Model Development: Included classification model deve
  - Model Evaluation: Involved both internal and external ev
  - Model Selection: Involved the identification of the best



* 90 observations + 83 features | ** 72 observations + 83 features | *** 18
A 4 candidate models built using the Logistic Regression, Support Vector Mach
B Hyperparameters of 4 candidate models fine-tuned and internally evaluated
C 4 candidate models with optimal hyperparameters were externally evaluated
D Best model among 4 optimal candidate models determined based on interna

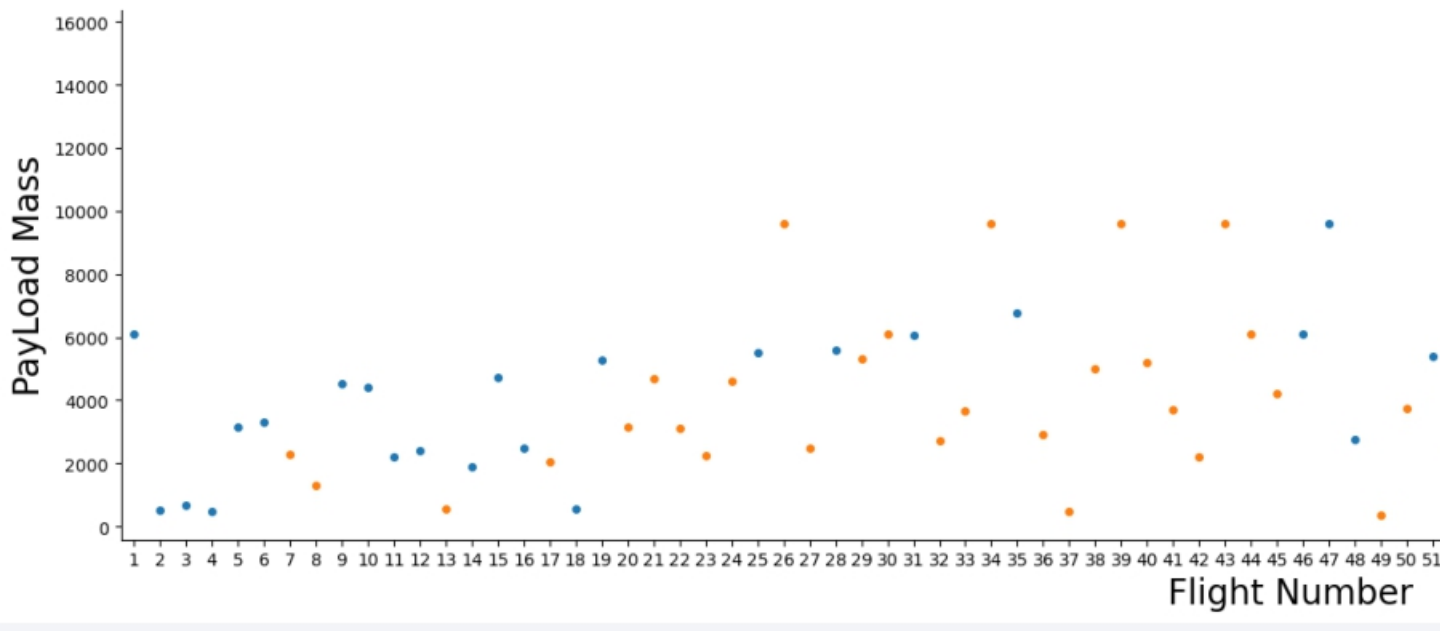
- GitHub URL of the completed Predictive analysis Pyt

Results

- Exploratory data analysis results
  - Insights Drawn from EDA
- Interactive analytics demo in screenshots
  - Launch Sites Proximities Analysis
  - Build a Dashboard with Plotly Dash
- Predictive analysis results
  - Predictive Analysis (Classification)

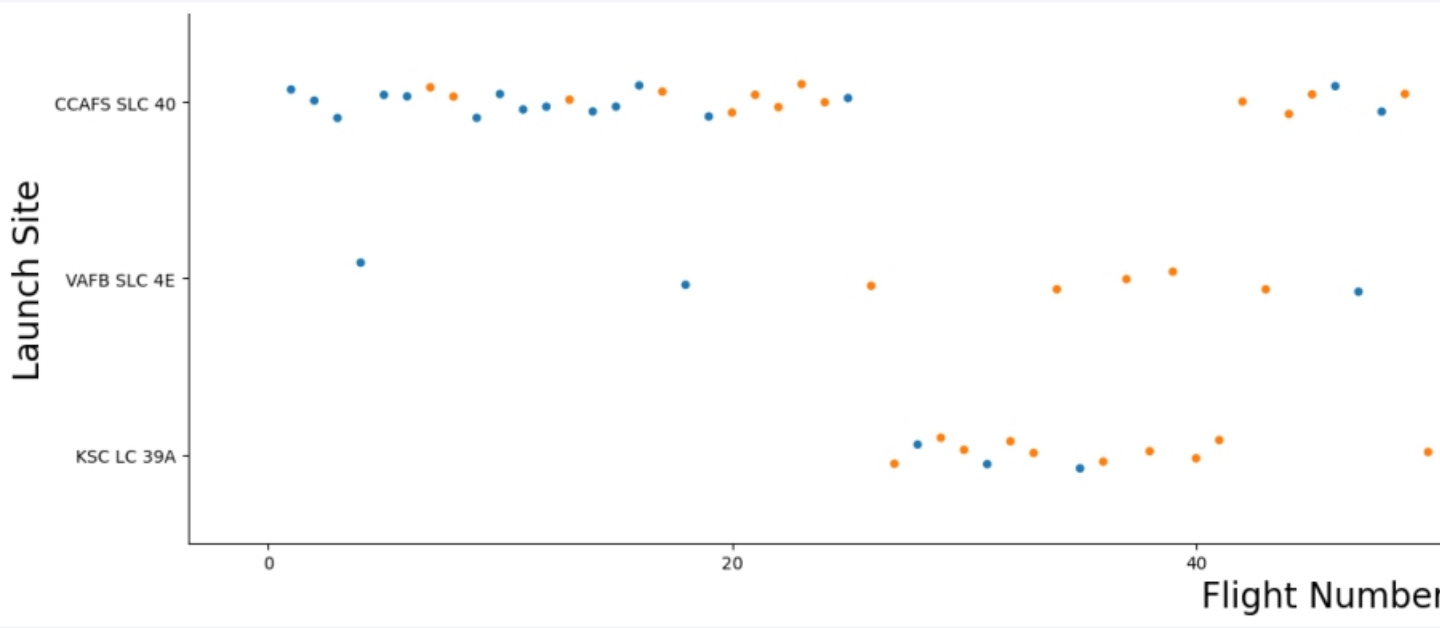

Section 2
Insights drawn
from EDA

Flight Number vs. Payload



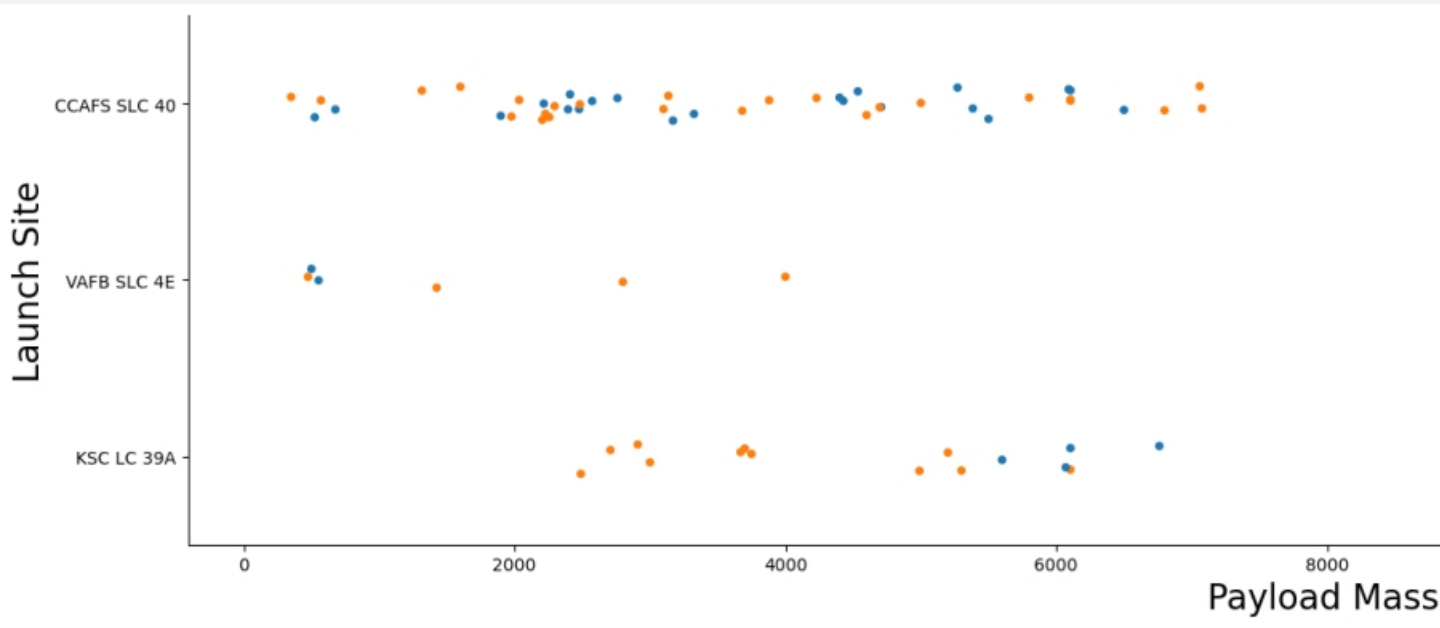
- Categorical Plot: Effect of Payload Mass and Flight N
  - More successful landings generally observed as flight nu
  - Successful landings observed across varying ranges of p

Flight Number vs. Launch Site



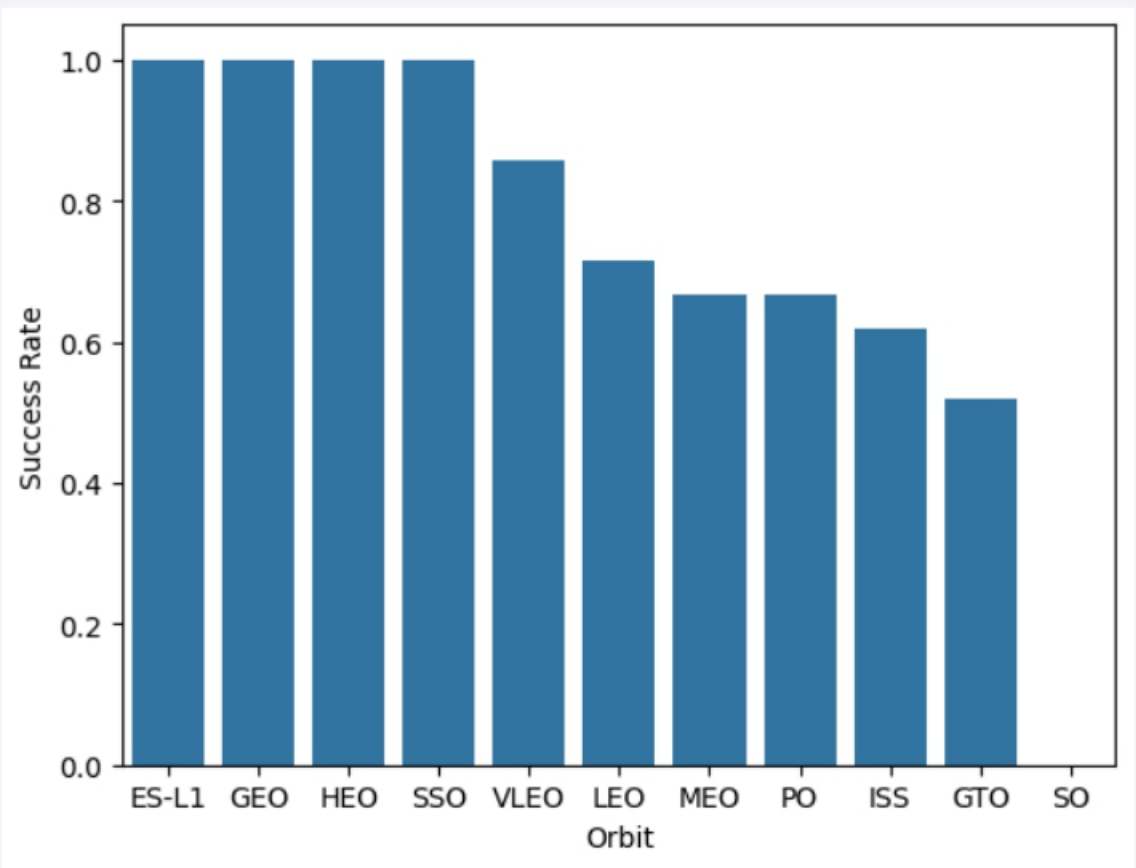
- Categorical Plot: Effect of Launch Site and Flight Nu
  - More successful landings generally observed for VAFB SL
  - Increased flight numbers only observed for CCAFS SLC 4

Payload vs. Launch Site



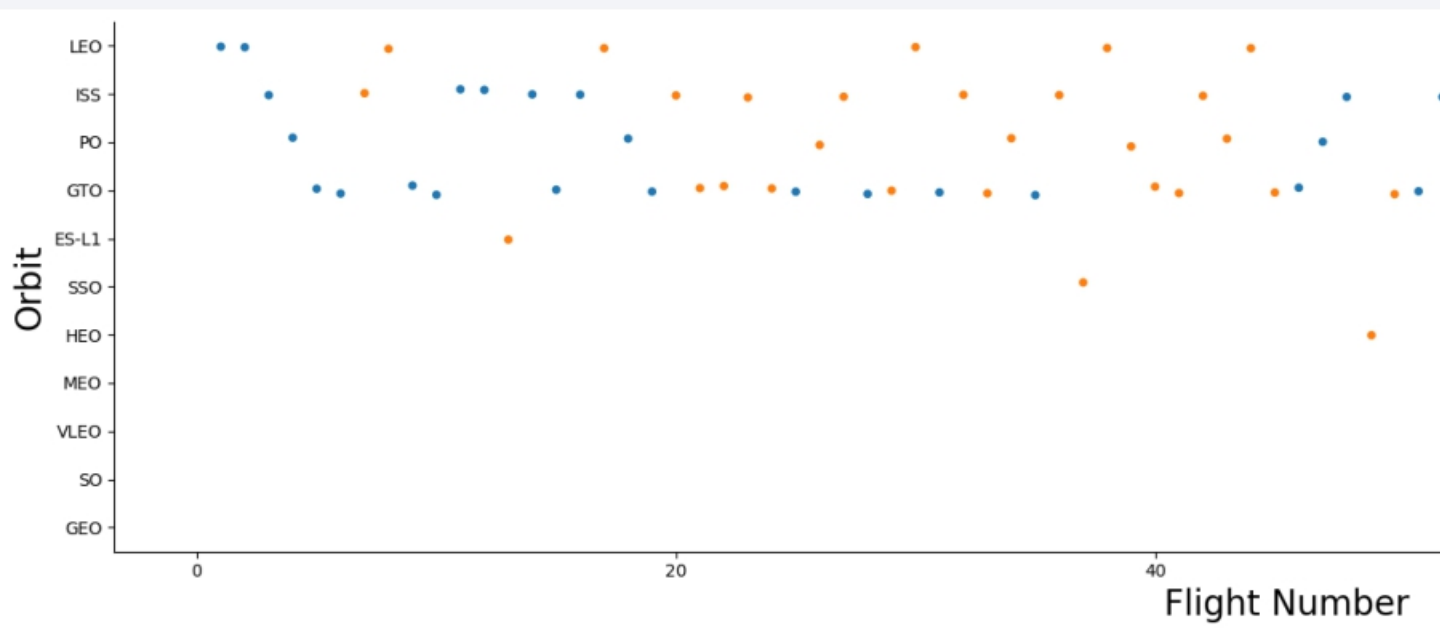
- Categorical Plot: Effect of Payload Mass and Launch
  - Successful landings observed across varying ranges of p
  - Higher payload masses >10K only observed for CCAFS S

Success Rate vs. Orbit Type



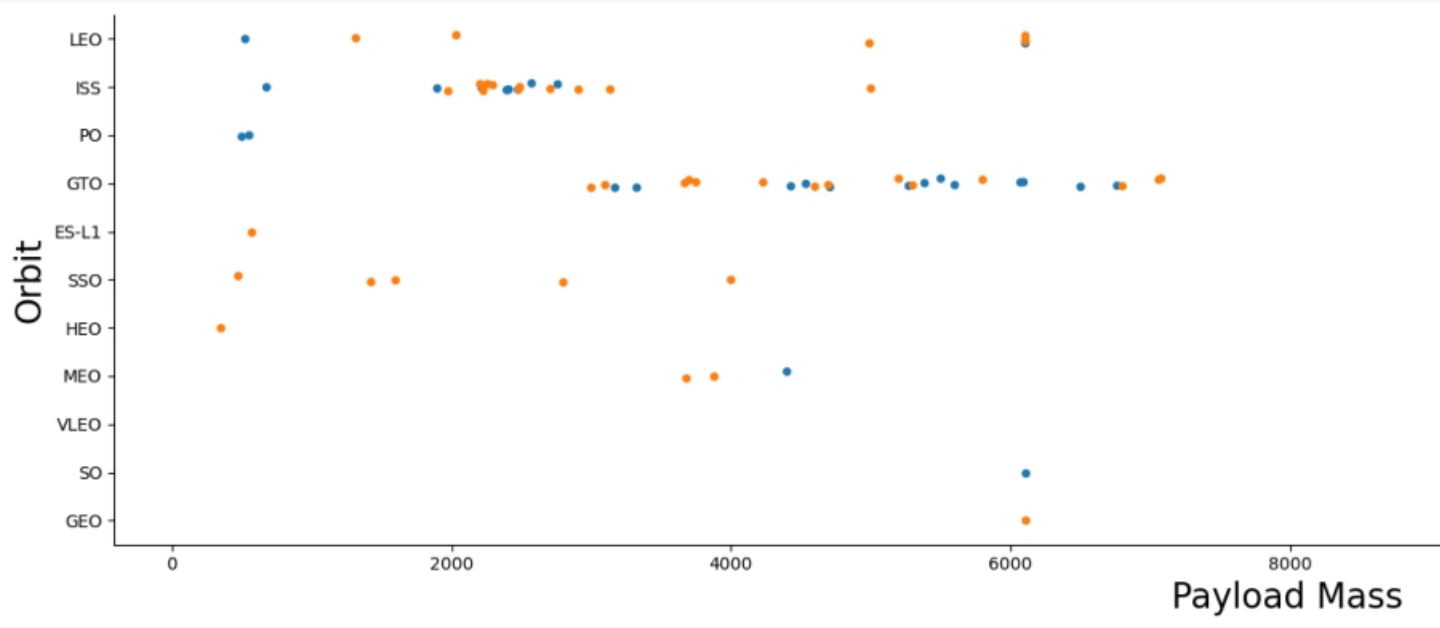
- Bar Plot:
  - High
    GEO

Flight Number vs. Orbit Type



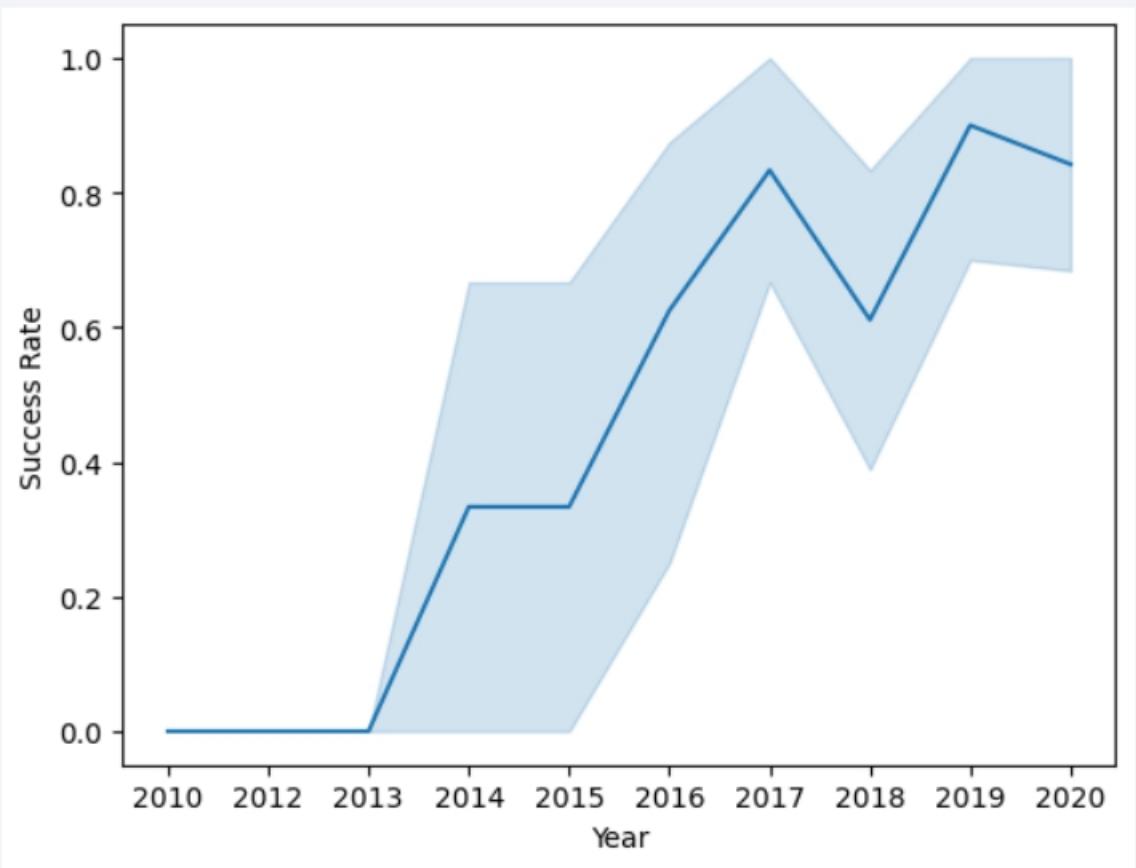
- Categorical Plot: Effect of Flight Number and Orbit to
  - Previous observation on higher landing success rate for E
  - Increased flight numbers only observed for the VLEO orb
  - Given a considerable number of flights, LEO, SSO and V

Payload vs. Orbit Type



- Categorical Plot: Effect of Payload Mass and Orbit to
  - Higher payload masses >9K only observed for ISS, PO ar
  - Successful landings observed across varying ranges of p

Launch Success Yearly Trend



- Line Plo
- High
  2013

All Launch Site Names

```
%%sql SELECT DISTINCT Launch_Site
FROM SPACEXTABLE;
```

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

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

- Distinct Selection: SELECT statement + DISTINCT exp
  - There were 4 unique launch sites included the space missi KSC LC-39A and CCAFS SLC-40.

Launch Site Names Begin with

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOA
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	

- String Patterns: SELECT statement + WHERE clause +
  - 5 records of launch sites beginning with the string 'CCA'

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
AS "TOTAL PAYLOAD MASS", Customer
FROM SPACEXTABLE
WHERE Customer LIKE "%NASA (CRS)%";
```

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

TOTAL PAYLOAD MASS	Customer
48213	NASA (CRS)

- Functions: SELECT statement + SUM function + WHE
  - The total payload mass carried by boosters launched by N

Average Payload Mass by F9 v

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
AS "AVERAGE PAYLOAD MASS", Customer, Booster_Version
FROM SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

AVERAGE PAYLOAD MASS	Customer	Booster_Version
2534.6666666666665	MDA	F9 v1.1 B1003

- Functions: SELECT statement + AVG function + WHE
  - The average payload mass carried by booster version F

First Successful Ground Landi

```
%%sql
SELECT MIN(DATE) FROM SPACEXTABLE
WHERE "Landing_Outcome"
LIKE "Success (ground pad)"
```

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

MIN(DATE)
2015-12-22

- Functions: SELECT statement + MIN function + WHE
  - The first successful landing outcome in a ground pad w

Successful Drone Ship Landing between 4000 and 6000

```
%%sql
SELECT * FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (drone ship)" AND  PAYLOAD_MASS__KG_>4000  AND PAYLOAD_MASS__
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbi
2016-06-05	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GT
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GT
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GT
2017-11-10	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GT

- Multiple Conditions: SELECT statement + WHERE cla
  - There were 4 boosters launched on the dates June 5, 20 November 10, 2017 which had success landing in dron 4000 but less than 6000.

Total Number of Successful an Failure Mission Outcomes

```
%%sql
SELECT Mission_Outcome, COUNT(Mission_Outcome) AS "Total Number of Mission Outcomes"
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

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

Mission_Outcome	Total Number of Mission Outcomes
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Grouping Result Sets: SELECT statement + COUNT fu
  - There were 100 successful mission outcomes and 1 faile

Boosters Carried Maximum Pa

```
%%sql
SELECT Booster_Version,Payload, PAYLOAD_MASS__KG_
FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

- Embedde
  WHERE
  - The b
    maxi
    F9 B
    F9 B
    F9 B
    F9 B

2015 Launch Records

```
%%sql
SELECT SUBSTR(Date, 6, 2) AS MONTH, Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTABLE
WHERE SUBSTR(Date, 0, 5) = "2015" AND Landing_Outcome = 'Failure (drone ship)';
```

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

MONTH	Booster_Version	Launch_Site	Landing_Outcome
10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Substring Extraction: SELECT statement + SUBSTR fu
  - There were two records for the months of April and Oct month names, failure landing outcomes in drone ship, b

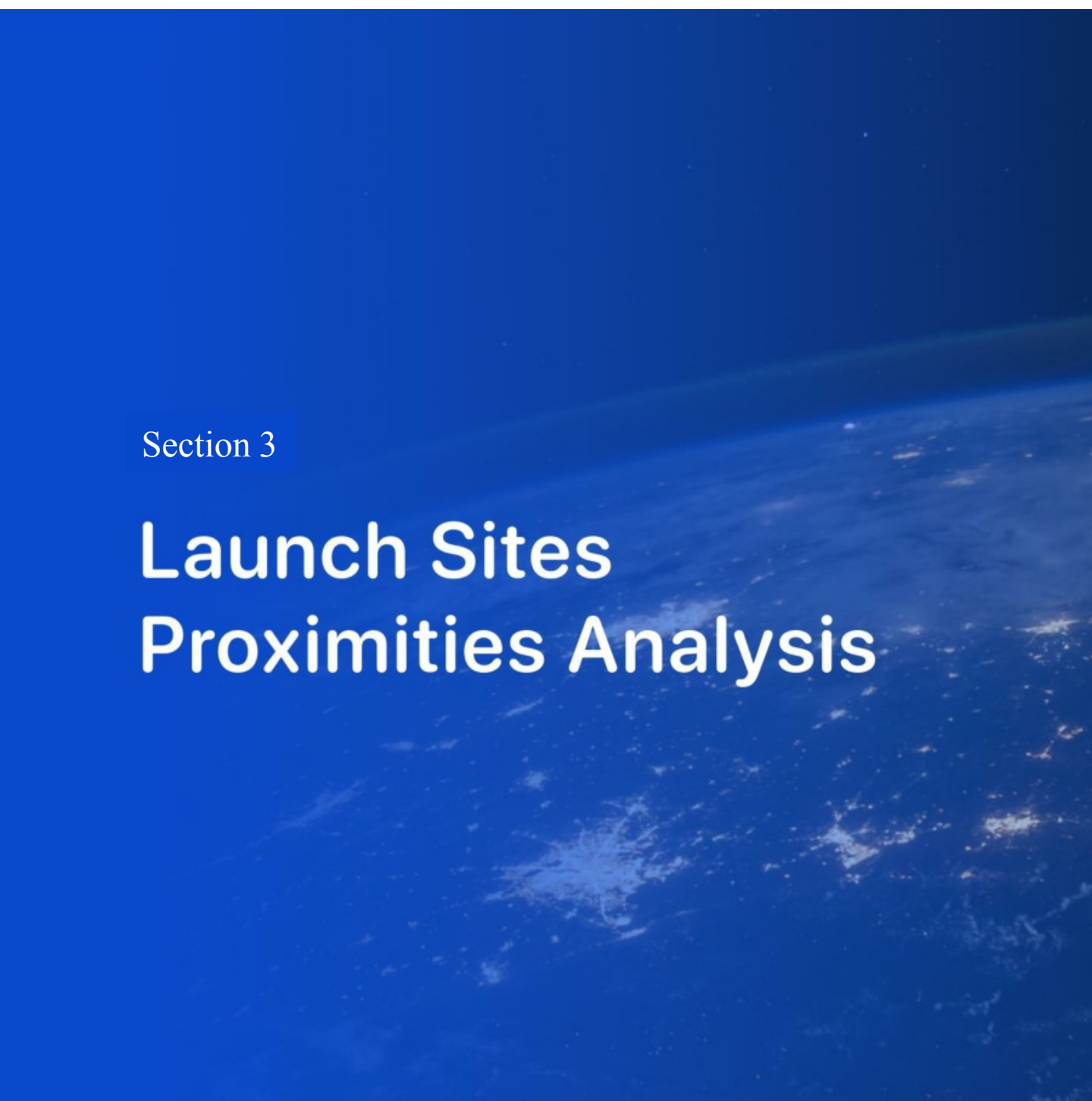
Rank Landing Outcomes Betw
2010-06-04 and 2017-03-20

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

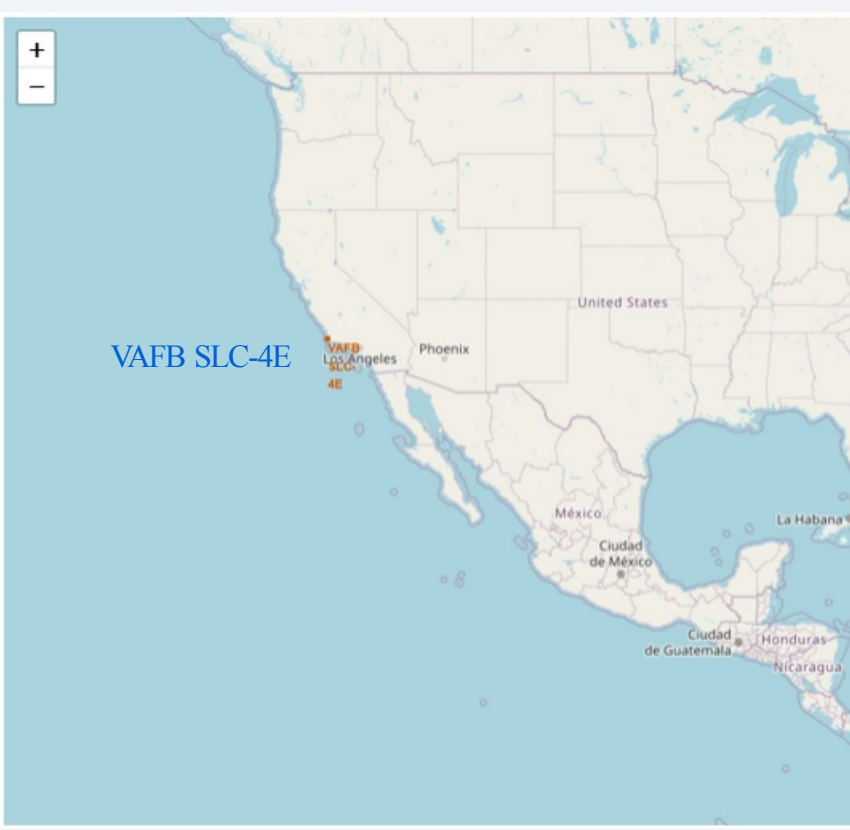
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Landing Outcome Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

- Range C
  SELECT
  clause +
  BY | OR
  - There
    the d
    whic

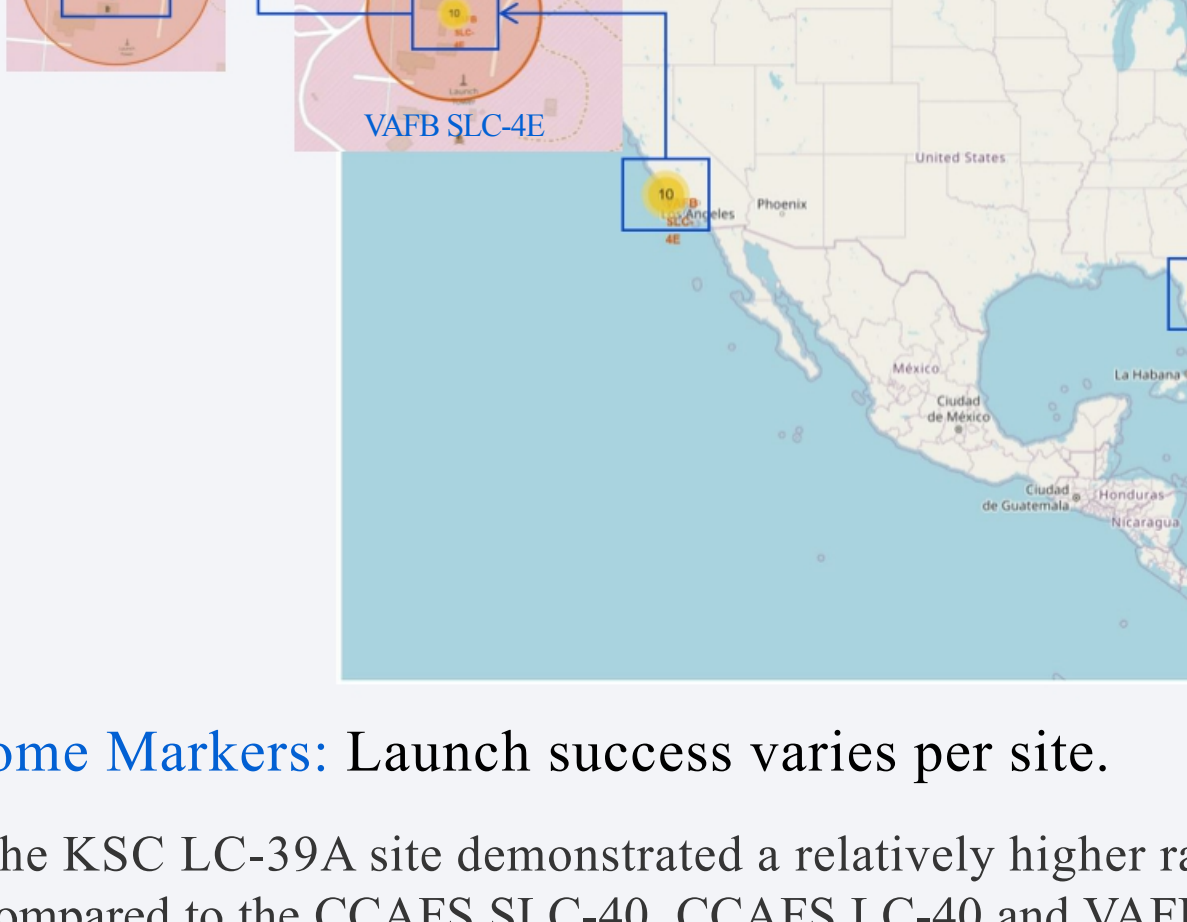

Section 3
Launch Sites
Proximities Analysis

Location Markers of Launch S



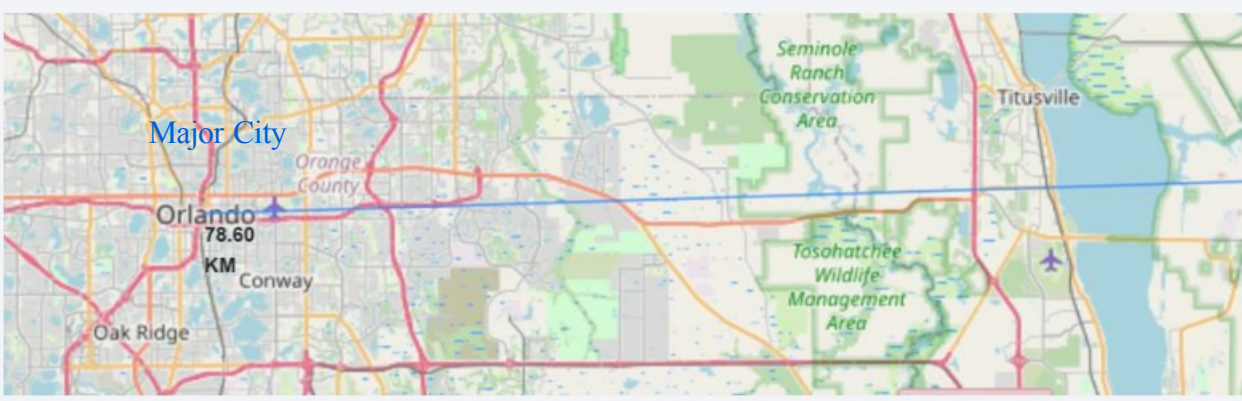
- Location Markers: US launch sites were normally situ
  - West Coast: VAFB SLC-4E
  - East Coast: KSC LC-39A + CCAFS SLC-40 + CCAFS LC-

Color-Labeled Launch Outcom

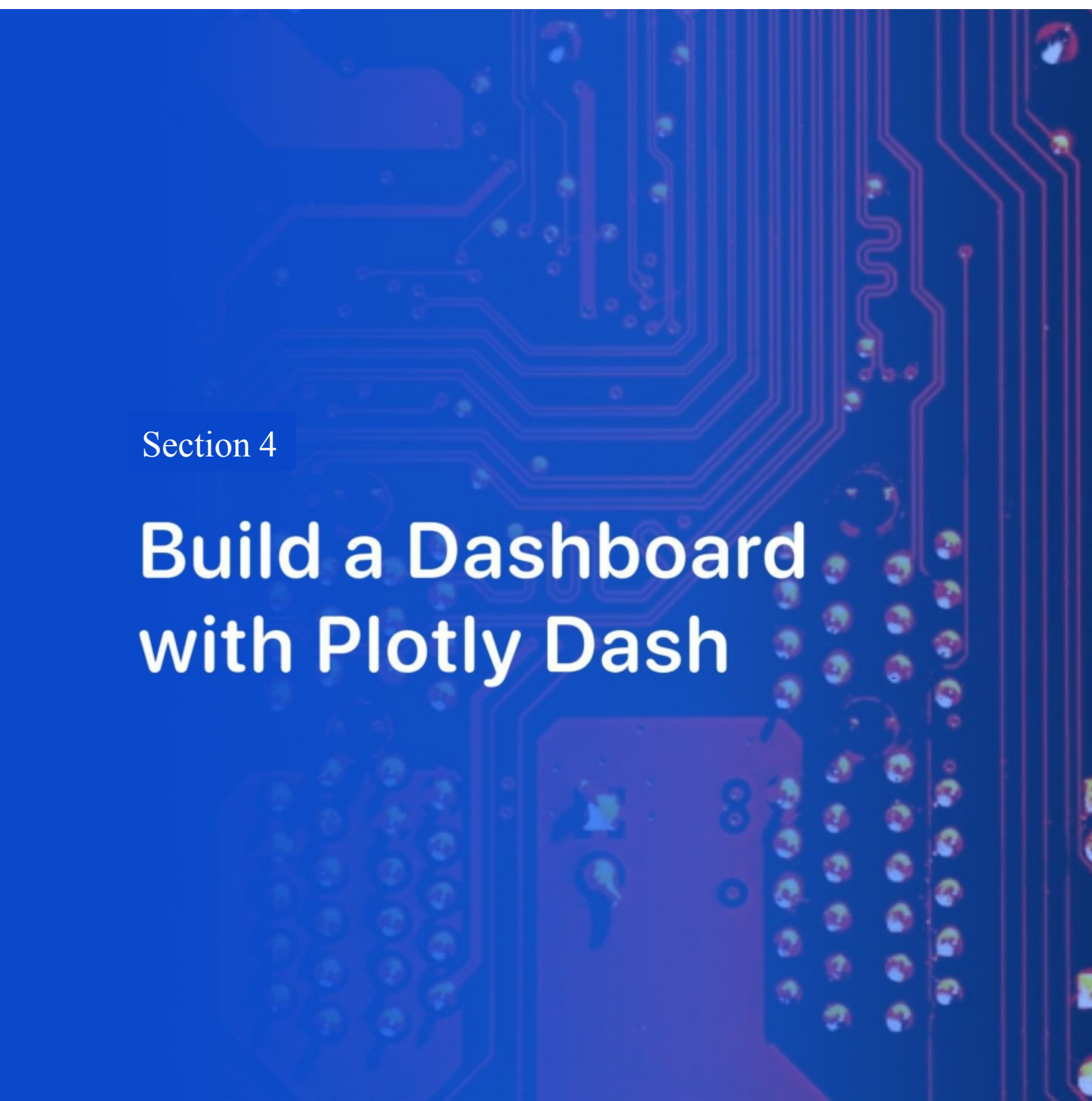


- Outcome Markers: Launch success varies per site.
  - The KSC LC-39A site demonstrated a relatively higher ra
    compared to the CCAFS SLC-40, CCAFS LC-40 and VAF

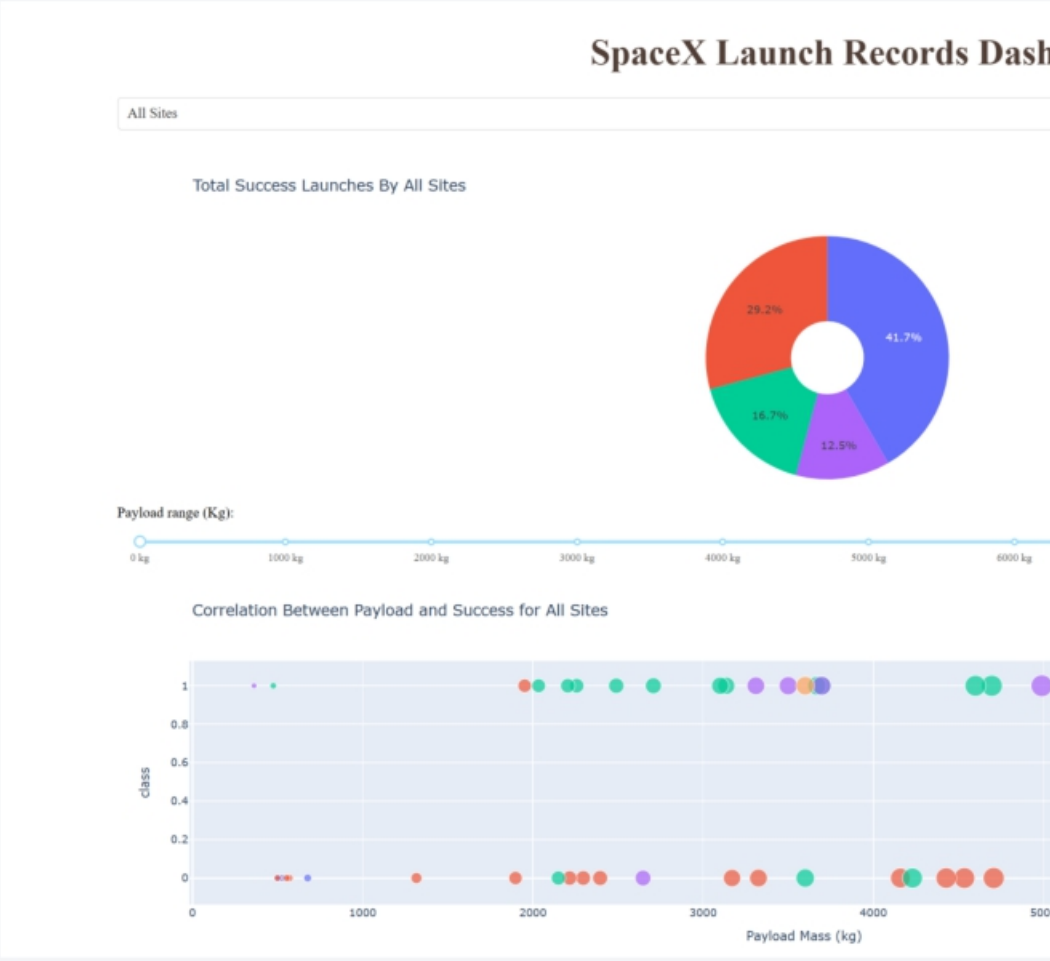
Proximity Analysis of Launch S
Coastline, Highway, Railway ar



- Distance Markers: Launch sites were located far from coastlines, highways and railroads.
  - Taking CCAFS SLC-40 as an example, its approximate d Orlando was 78.60 Km.
  - The approximate distance of the CCAFS SLC-40 launch
  - The approximate distance of the CCAFS SLC-40 launch
  - The approximate distance of the CCAFS SLC-40 launch

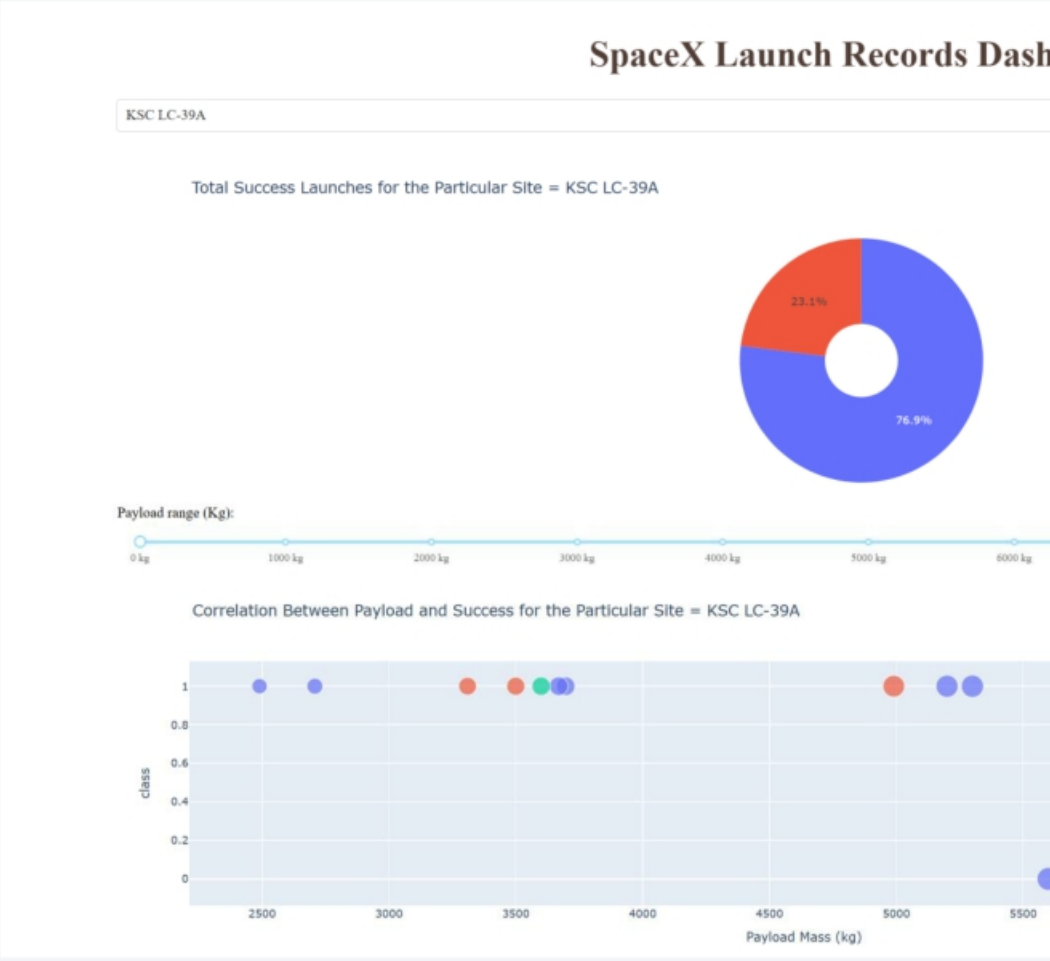

Section 4
Build a Dashboard with Plotly Dash

Combined Launch Success Co



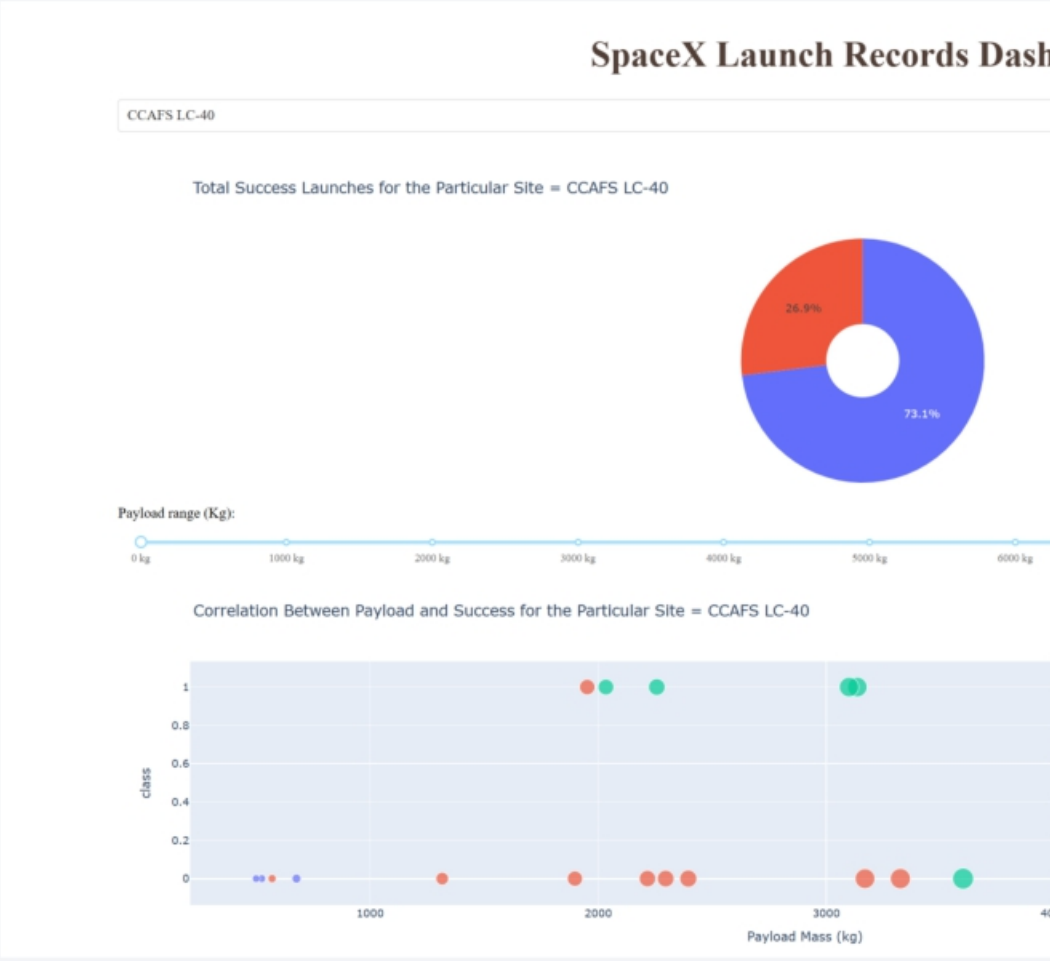
· Launch Success Count: The most number of successfu
KSC LC-39A contributing 41.7% to the overall launch

Launch Success Ratio Per Indi



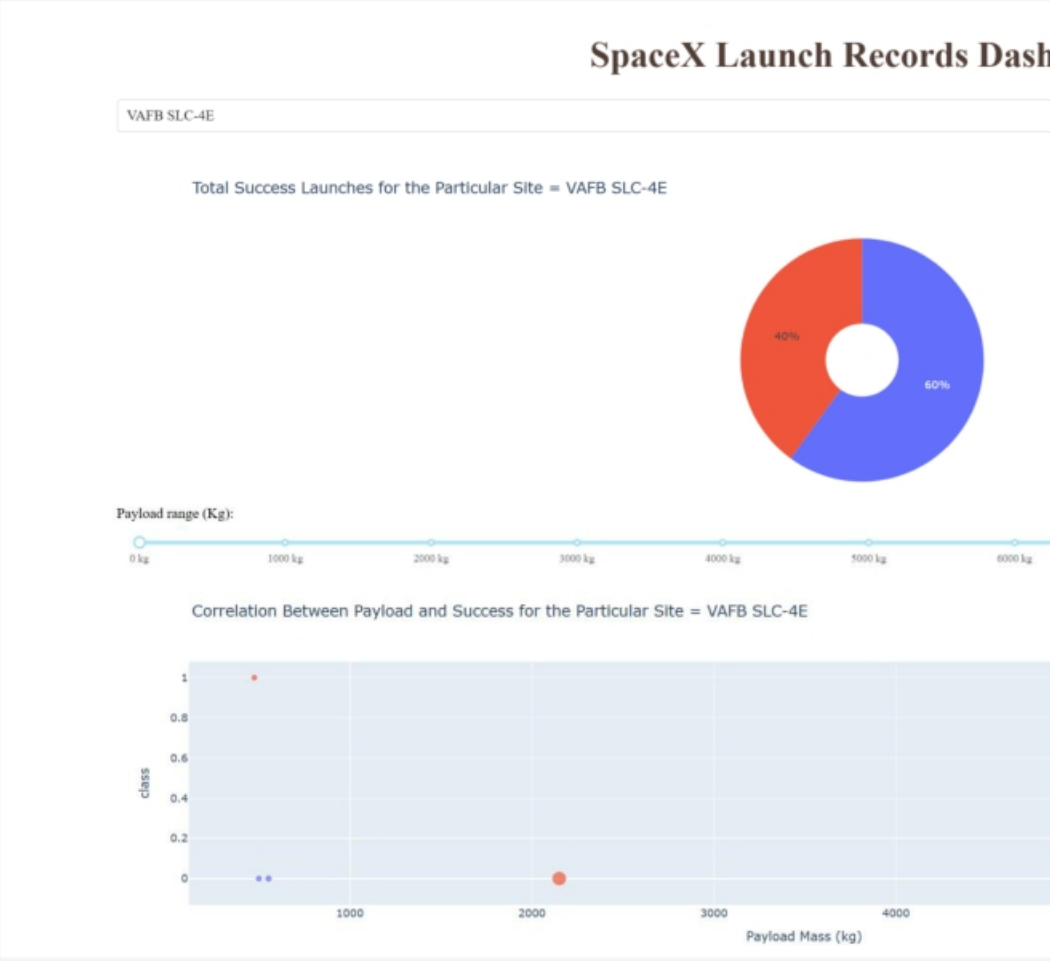
· Launch Success Ratio: Among all sites, the launch succ
ranked first. 76.9% of all launches from KSC LC-39A

Launch Success Ratio Per Indi



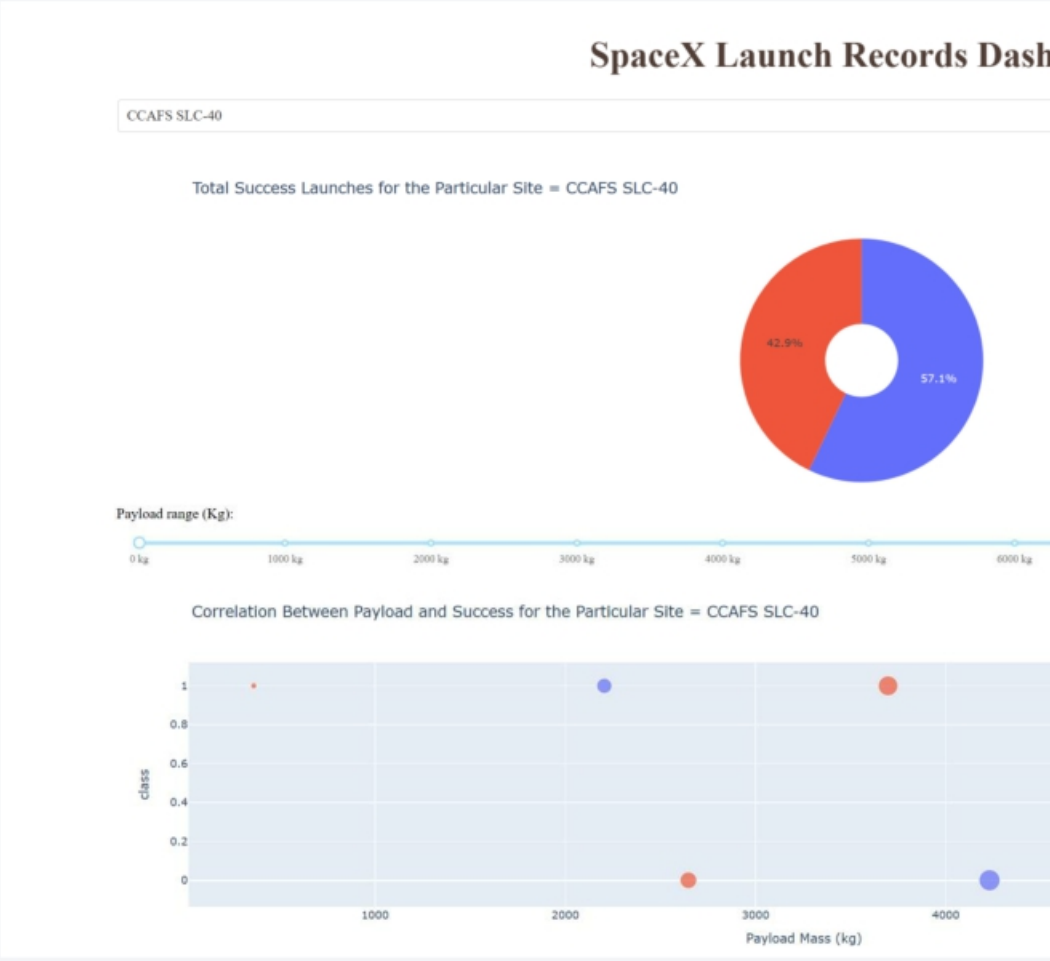
· Launch Success Ratio: Among all sites, the launch succ
ranked second. 73.1% of all launches from CCAFS LC

Launch Success Ratio Per Indi



· Launch Success Ratio: Among all sites, the launch succ
ranked third. 60.0% of all launches from VAFB SLC-4

Launch Success Ratio Per Indi

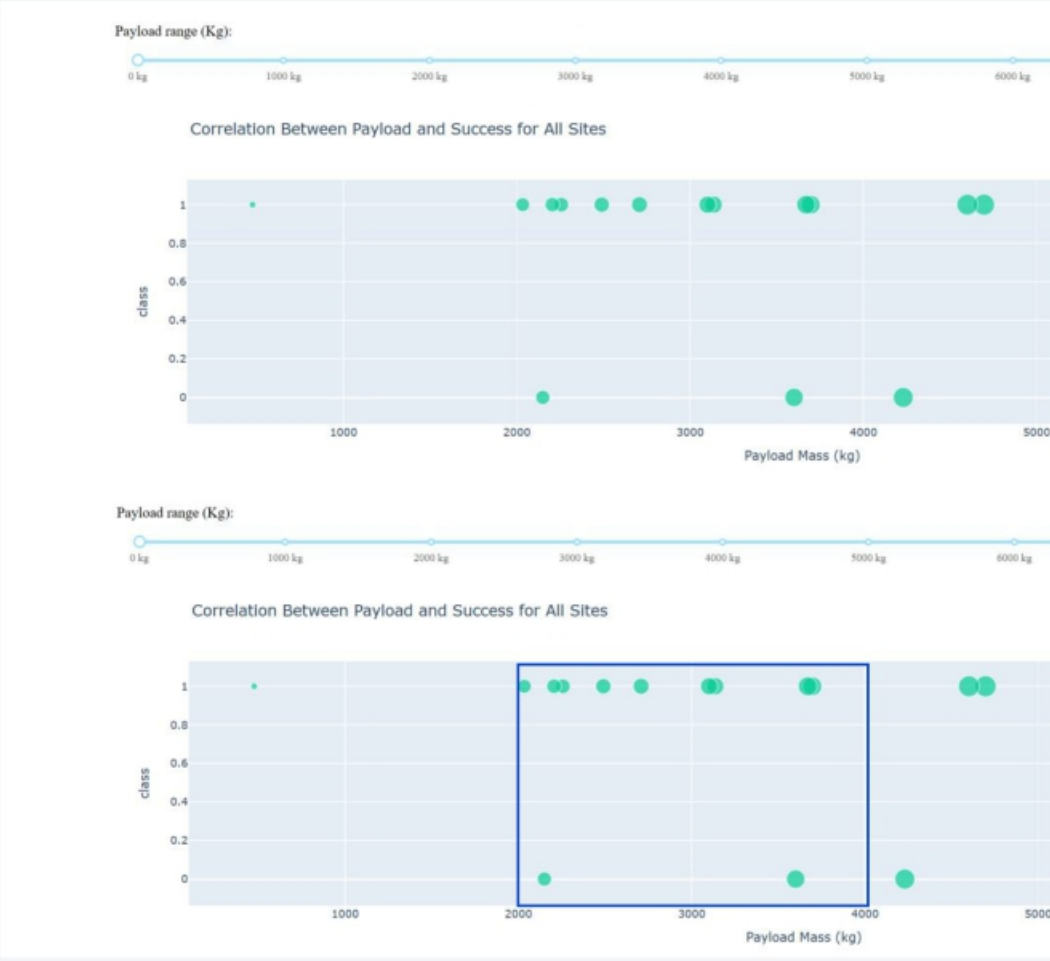


· Launch Success Ratio: Among all sites, the launch succ
was ranked third. 57.1% of all launches from CCAFS S

Launch Outcomes By Payload

· Launch Outcomes By Payload: At a payload range of
ratio is at 60.0%. In contrast, launch success ratio is

Launch Outcomes By Booster

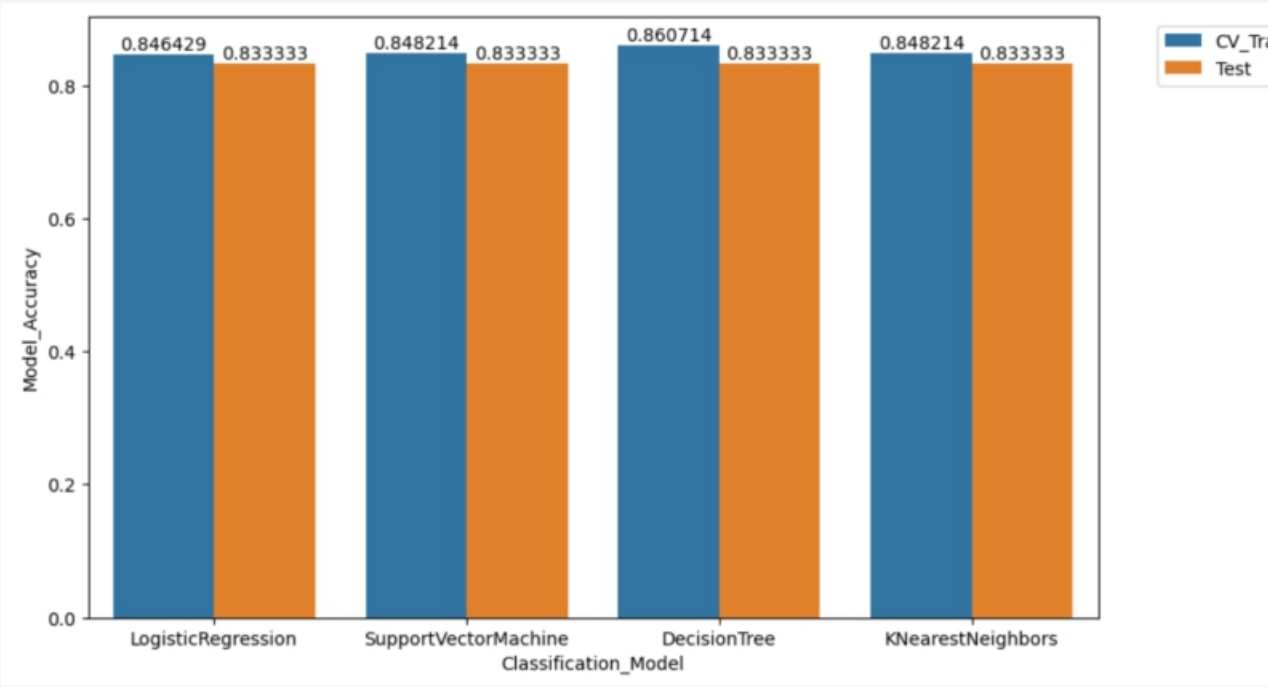


· Launch Outcomes By Booster Version: The highest lau
booster FT version at 65.0%. This increases to 81.8%

Section 5

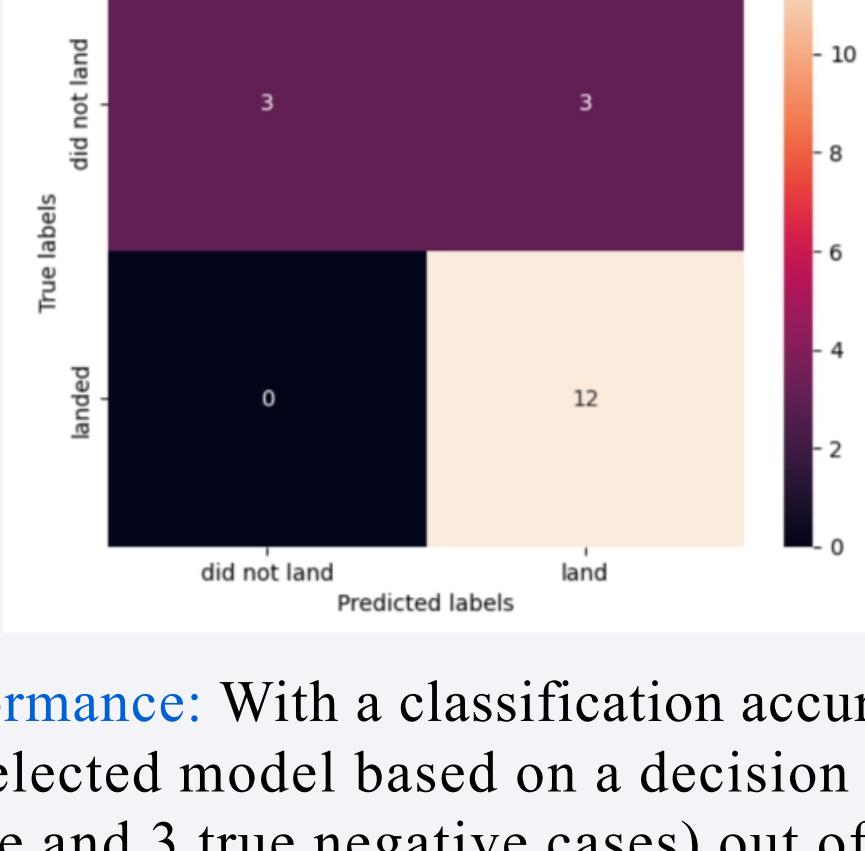
Predictive Analysis (Classification)

Classification Accuracy



- Internal + External Model Evaluation: All 4 candidate
  83.33% accuracy on the test set. Higher accuracy was
  tree model at 86.07% based on internal evaluation us
  - Cross-Validated Accuracy on Train: 86.07% (Decision Tree)
  - Accuracy on Test: 83.33% (LR, SVM, Decision Tree, KNN)

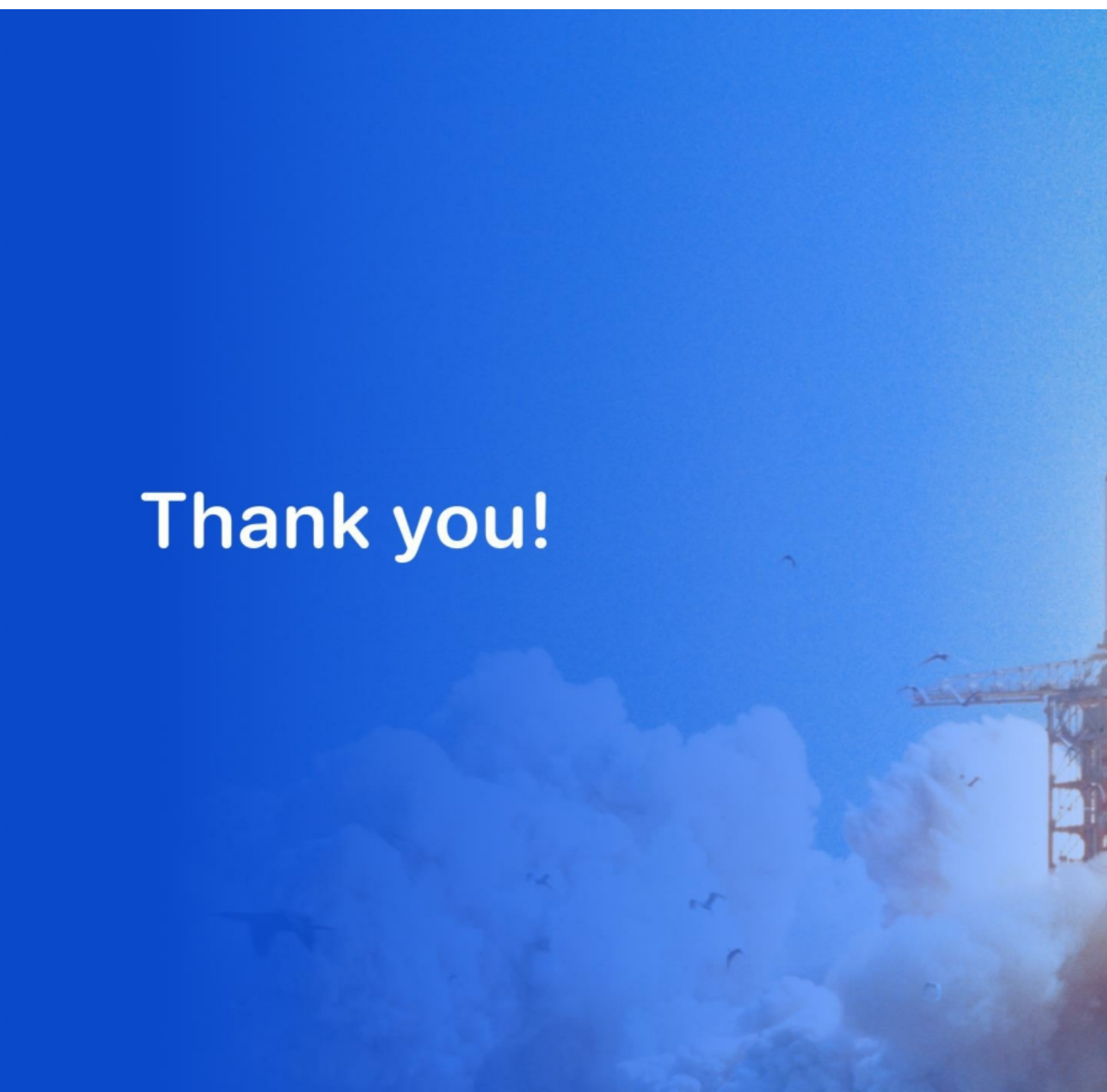
Confusion Matrix



- Classification Performance: With a classification accu
  test set, the final selected model based on a decision
  15 (12 true positive and 3 true negative cases) out of
  cases were noted which the model predicted as succes

Conclusions

- Data collection for the analysis involved SpaceX REST
- Appropriate pre-processing methods including row a imputation, one-hot encoding, target creation were a subsequent analysis and modeling.
- EDA using visualization and SQL demonstrated the e flight number, launch site and orbit to successful lan
- Interactive visual analytics using Folium demonstrate proximities to a coastline, highways, railways and ma
- Interactive visual analytics using Plotly Dash demons launch site, payload mass and booster version to suc
- A decision tree classification model provided a robus landing outcomes with a cross-validated accuracy o accuracy of 83.33%.


Thank you!