Winning Space Race with Data Science

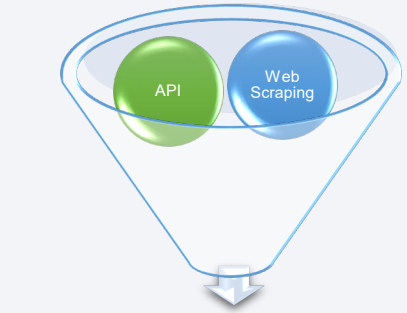
Zafer Ali Serbes
02.04.2025

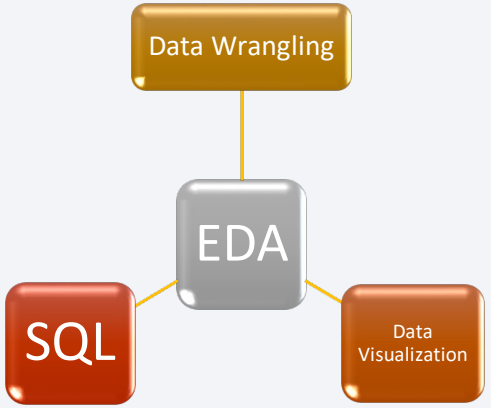
Outline

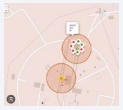
- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- CONCLUSION
- APPENDIX

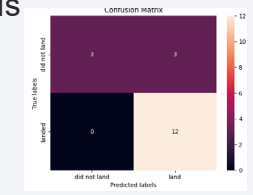
Executive Summary

- Methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis
  - Predictive Analysis









Introduction

- Project Background:

SpaceX's Falcon 9 costs $62M—over **60% cheaper** than competitors ($165M+) due to first-stage reusability.

- Problem Statement:

If we can determine "*whether the first stage will land*", we can accurately estimate the launch cost—critical for competitive bidding.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Request to the SpaceX API
  - Web scraping Falcon 9 launch records with BeautifulSoup
- Perform data wrangling
  - Dealing with Missing Values and export it to a CSV
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, SVM, Decision Tree, and KNN models
  - GridSearch cross validation, Accuracy calculation and confusion matrix

Data Collection – SpaceX API

1. • requesting rocket launch data from SpaceX API
   • URL: spacex_url="https://api.spacexdata.com/v4/launches/past"
2. • decode the response content as a Json using and turn it into a Pandas dataframe using
   • **data=pd.json_normalize(response.json())**
3. • The data from these requests will be stored in lists and used to create a new dataframe.
   • create a Pandas data frame from the dictionary launch_dict
4. • Filter the dataframe to only include Falcon 9 launches
   • **data_falcon9=launch_data[launch_data['BoosterVersion']!='Falcon 1']**

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

- GitHub URL:

https://github.com/zaferaliserbes/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api(2).ipynb

Data Collection – Scraping

1. • Web scrap Falcon 9 launch records with BeautifulSoup
   • URL: "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
2. • Request the Falcon9 Launch Wiki page from its URL
   • use requests.get() method with the provided static_url
3. • Extract all column/variable names from the HTML table header
   • **Create a data frame by parsing the launch HTML tables**
4. • **Exporting file**

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

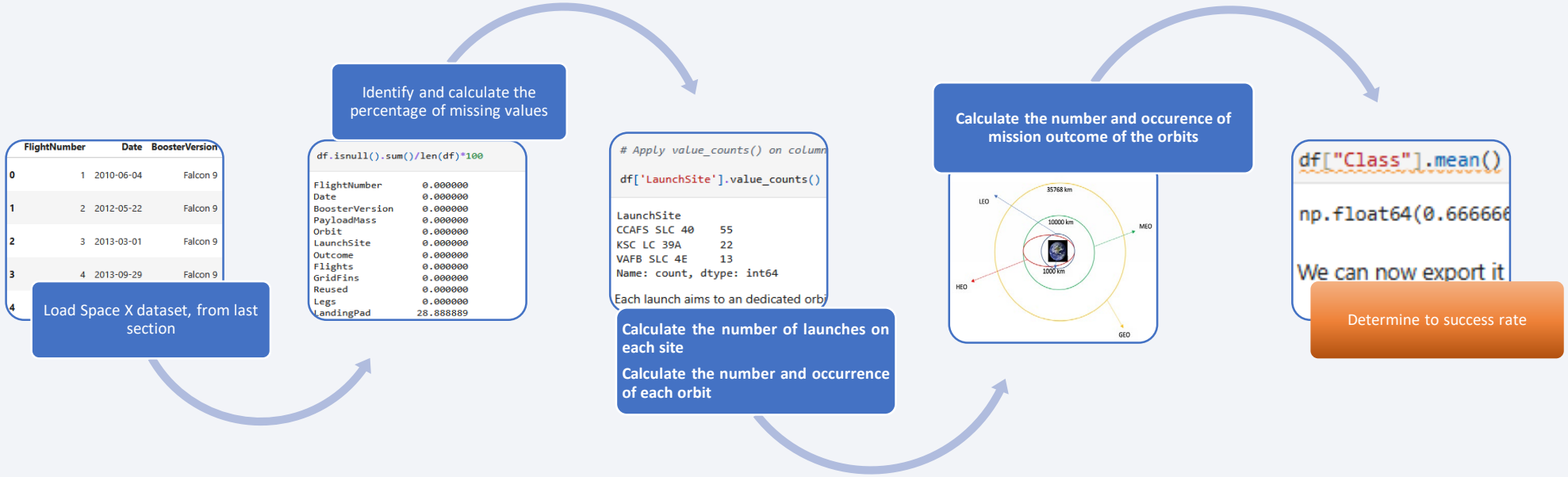
```
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

- GitHub URL:

https://github.com/zaferaliserbes/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-webscraping(1).ipynb

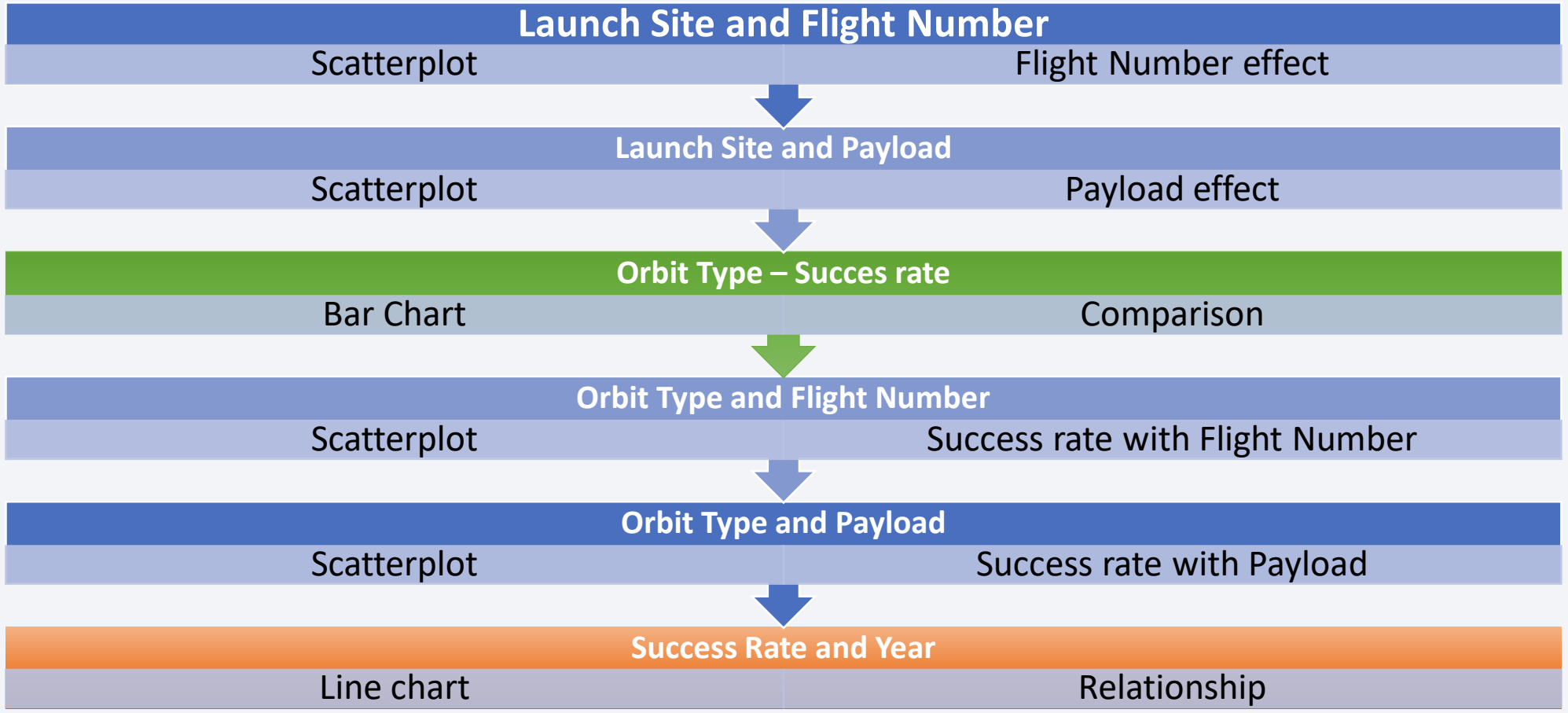
Data Wrangling



- GitHub URL:

https://github.com/zaferaliserbes/IBM-Data-Science-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb

EDA with Data Visualization



- GitHub URL:

https://github.com/zaferaliserbes/IBM-Data-Science-Capstone-Project/blob/main/edadataviz(1).ipynb

EDA with SQL

- **unique** launch sites in the space mission
- launch sites begin with the string '**CCA**'
- total payload mass carried by boosters launched by **NASA (CRS)**
- average payload mass carried by booster version **F9 v1.1**
- date when the first succesful landing outcome in ground pad was **acheived**
- names of the boosters which have success in drone ship and have payload mass **greater than 4000 but less than 6000**
- **total number** of successful and failure mission outcomes
- names of the booster_versions which have carried the **maximum payload mass**
- display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months **in year 2015**.
- Rank the count of landing outcomes **between** the date 2010-06-04 and 2017-03-20, in **descending order**

- GitHub URL:

https://github.com/zaferaliserbes/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite(1).ipynb

Build an Interactive Map with Folium

- Markers are used to indicate specific points on the map, such as rocket launch sites.
- Circles highlight areas around certain coordinates, like the NASA Johnson Space Center.
- Marker clusters group multiple events occurring at the same location, such as different rocket launches at a launch site.
- Lines represent distances between two coordinates or show connections between locations.

- GitHub URL:

https://github.com/zaferaliserbes/IBM-Data-Science-Capstone-Project/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

The Plotly Dash dashboard features a dropdown to select data from a single launch site or all sites:

- For one site, the pie chart shows the success and failure distribution of Falcon 9 landings.
- For all sites, it displays the success distribution across launch sites.
- A slider filters payload mass for the scatterplot.
- The scatterplot visualizes Falcon 9 landings by payload mass, mission outcome, and booster version.

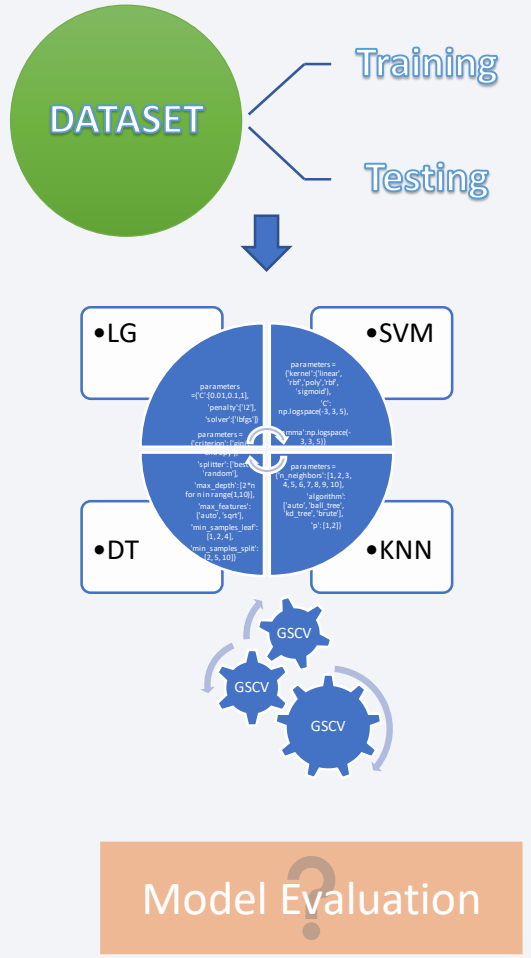
- GitHub URL:

https://github.com/zaferaliserbes/IBM-Data-Science-Capstone-Project/blob/main/spacex-dash-app%20(1).py

Predictive Analysis (Classification)

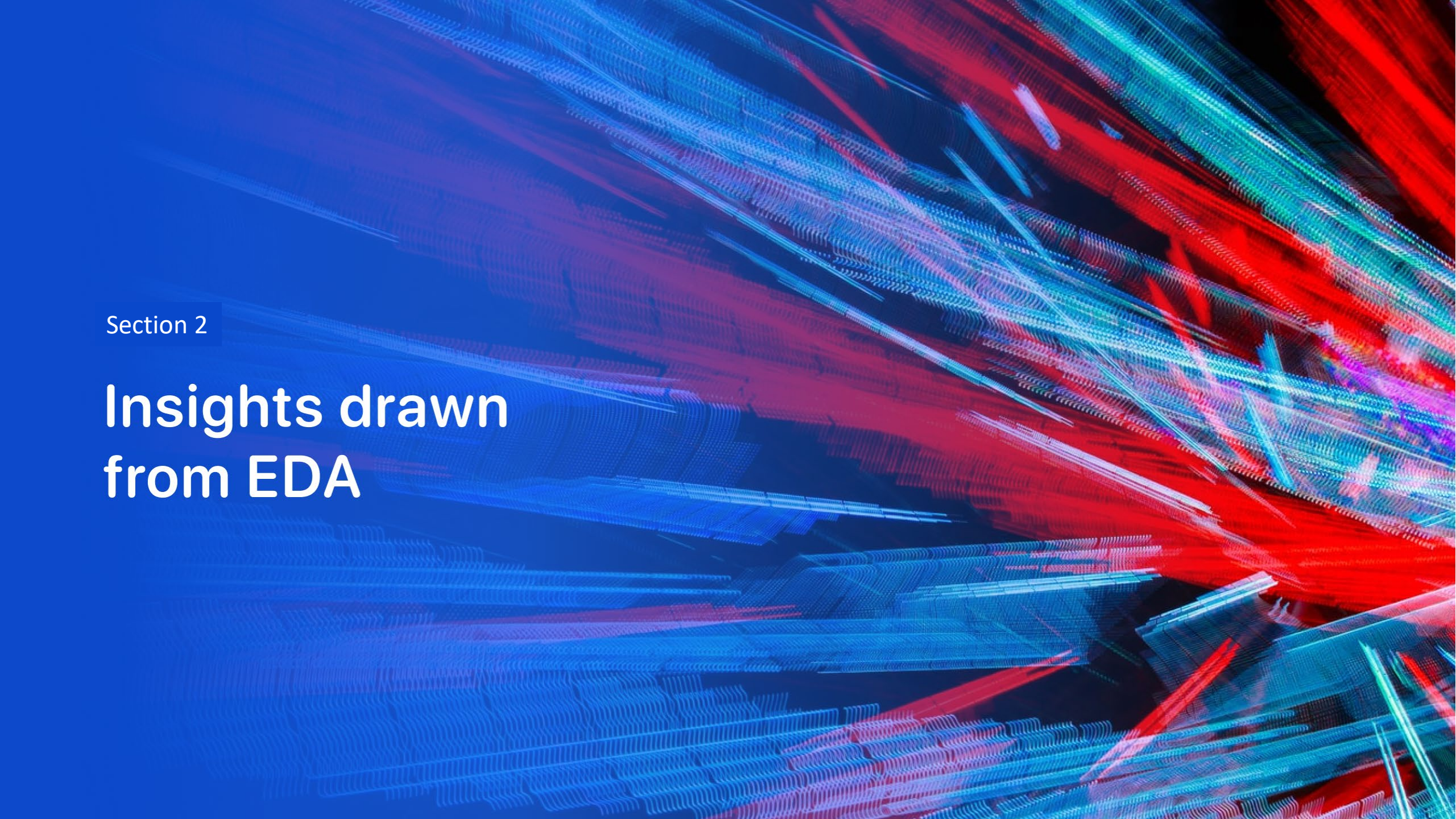
The dataset was split into training and testing sets:

- Logistic Regression, SVM, Decision Tree, and KNN models were trained on the training data.
- GridSearchCV() was used to tune hyperparameters, and the best ones were selected.
- The models were then evaluated on the test set for accuracy using the optimal hyperparameters.

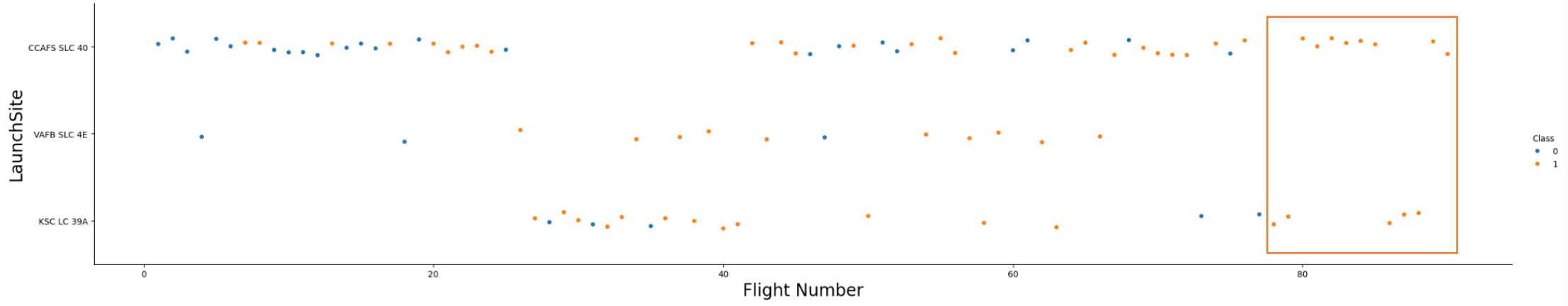


- GitHub URL:

https://github.com/zaferaliserbes/IBM-Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(2).ipynb


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



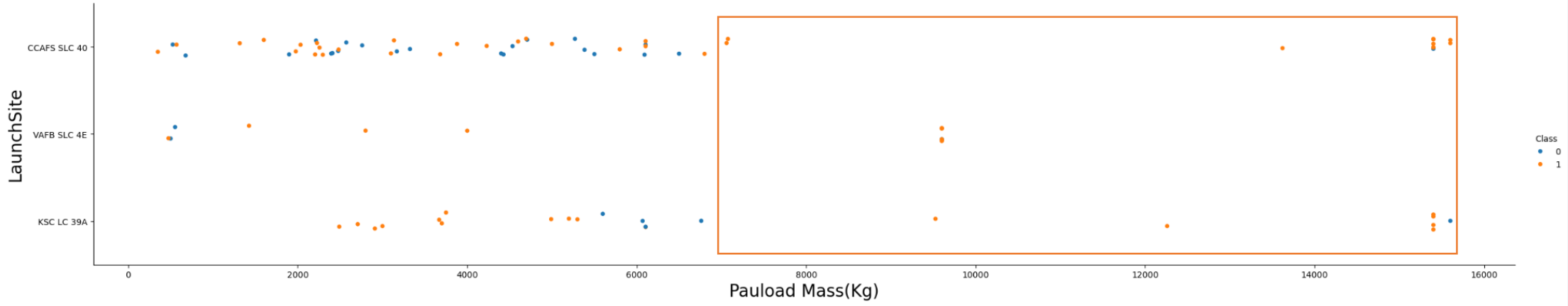
As the number of flights increases, the success rate increases. The success rate was 100% after 78th total flight number.

CCAFS SLC 40 : 46 and above flights,

VAFB SCL 4E : 9 and above flights

KSC LC 39A : 18 and above flights

Payload vs. Launch Site



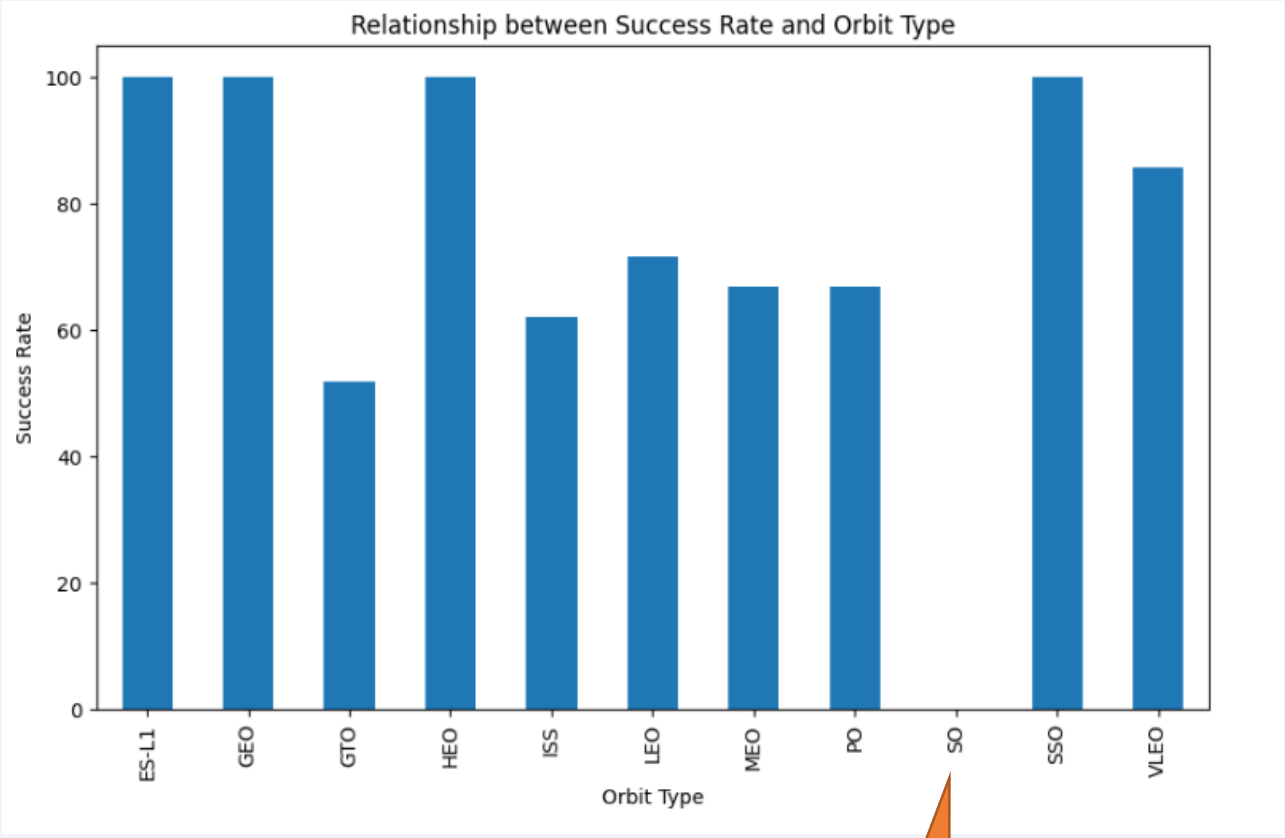
As the Payload Mass increases, the success rate is **almost** 100%,

Especially after 7000 for CCAFS SLC 40 and KSC LC 39A .

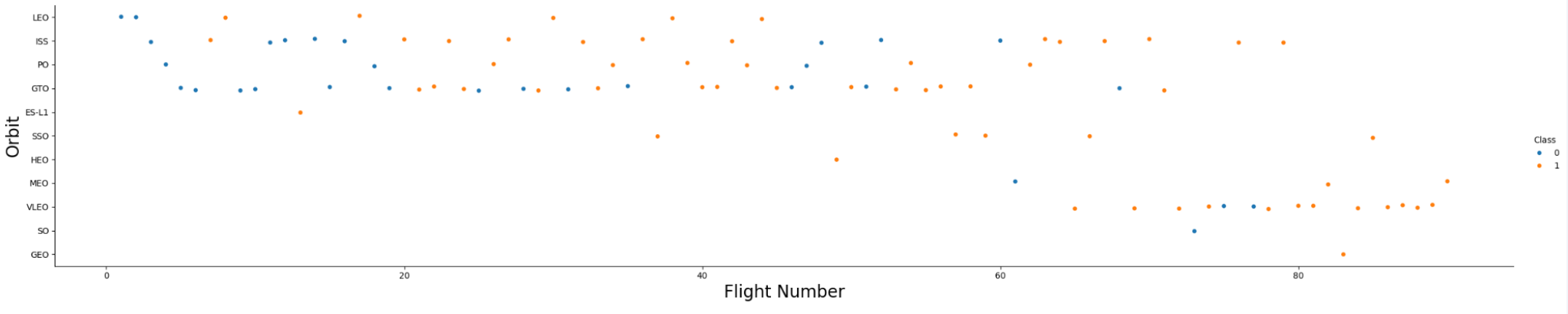
for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000), and success rate is very high.

Success Rate vs. Orbit Type

- All recorded launches targeting **ES-L1**, **SSO**, **HEO**, and **GEO** have completed first stage landings without failure.
- In contrast, **SO** orbit launches have yet to demonstrate a successful recovery of the first stage.



Flight Number vs. Orbit Type



- **LEO** orbit, success seems to be related to the number of flights.
- Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

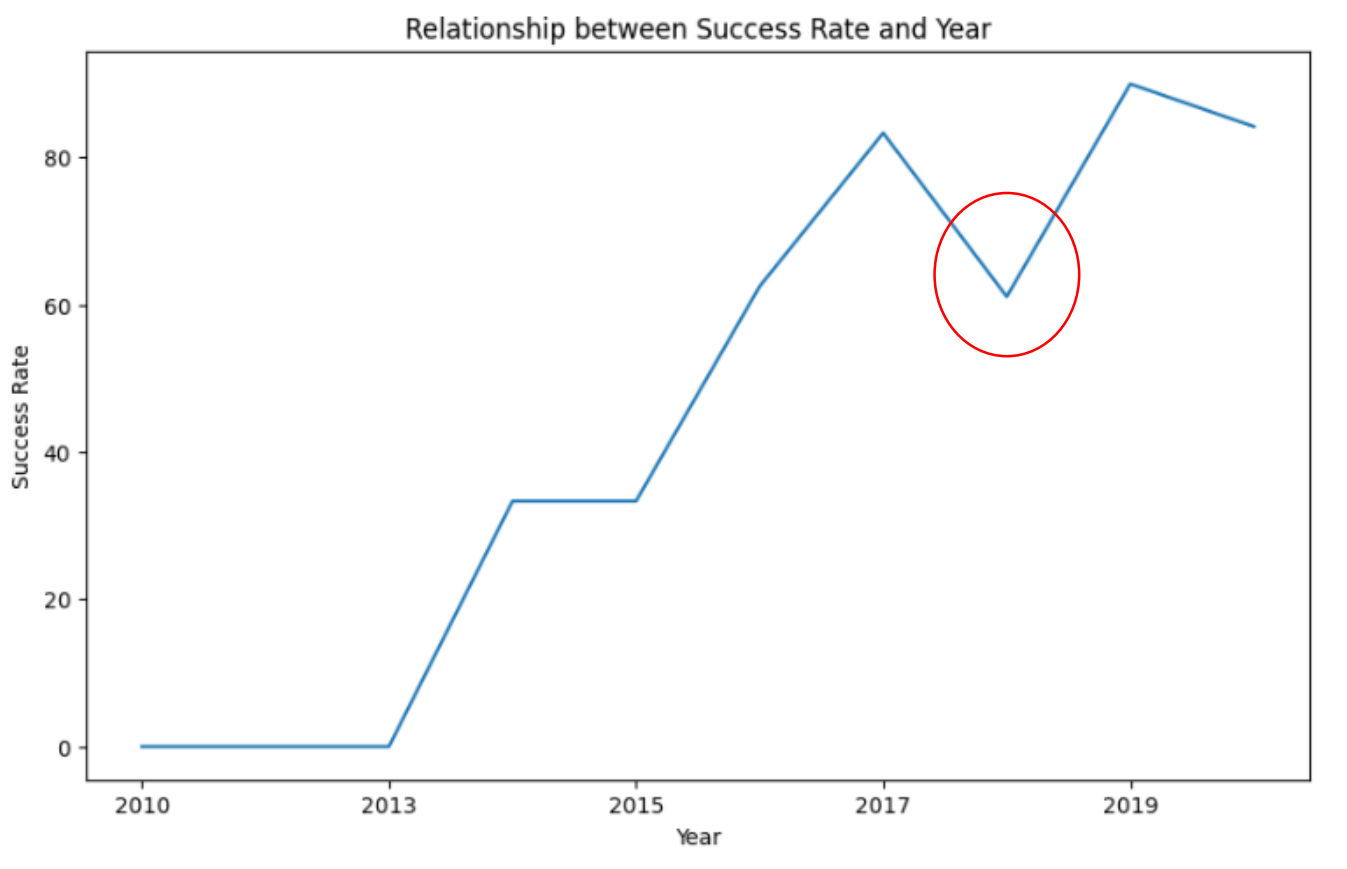
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for **Polar**, **LEO** and **ISS**.

- However, for **GTO**, it's <u>difficult to distinguish</u> between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- An upward trend in the success rate has been noted during the period from 2013 to 2020.
- A considerable drop in the success rate was identified in 2018, highlighting a deviation from the overall trend.



All Launch Site Names

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;
```

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

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

- A total of four unique launch sites are represented in the dataset, each serving as a point of origin for SpaceX missions.

Launch Site Names Begin with 'CCA'

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

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

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

- To gain an overview of the data distribution, a simple sampling method was applied to the database table.
- The SQL keyword ***LIKE*** is applied to filter records where the launch site name begins with the prefix "**CCA**".

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER="NASA (CRS)";
 * sqlite:///my_data1.db
Done.
SUM(PAYLOAD_MASS__KG_)
                 45596
```

- Based on the dataset, NASA boosters have transported a total payload of 45596 kg.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE "F9 V1.1%";
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- A mean payload mass of 2534.67 kg was computed for the F9 v1.1 booster using query-based data extraction.

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME="Success";
 * sqlite:///my_data1.db
Done.
```

MIN(DATE)
2018-07-22

- According to the dataset, the earliest documented instance of a successful ground pad landing took place on **December 22, 2015**.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000) AND (Landing_Outcome="Success (drone ship)");
```

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

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

- According to the query results, **four** distinct booster versions fulfilled the conditions of a successful drone ship landing and payload mass falling **between 4000 and 6000 kg**.

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(mission_outcome) from SPACEXTBL GROUP BY mission_outcome
```

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

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

- A total of 101 missions were evaluated, of which 100 resulted in success and 1 in failure.

Boosters Carried Maximum Payload

```
%sql select booster_version, payload_mass__kg_ from SPACEXTBL\
where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
```

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

- Using SQL's **MAX()** function in a **nested WHERE clause** allowed for the identification of the booster that transported the largest payload

2015 Launch Records

```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome FROM SPACEXTBL WHERE Landing_Outcome='Failure (drone ship)' AND substr(Date,0,5)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- A query was executed to list the failed drone ship landing outcomes, their corresponding booster versions, and launch site names for missions in 2015.

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

The query results indicate that:

- the most frequent landing outcome within the given date range was "**No attempt**," with 10 occurrences. Other notable outcomes include "**Success (drone ship)**" and "**Failure (drone ship)**," each with 5 occurrences.

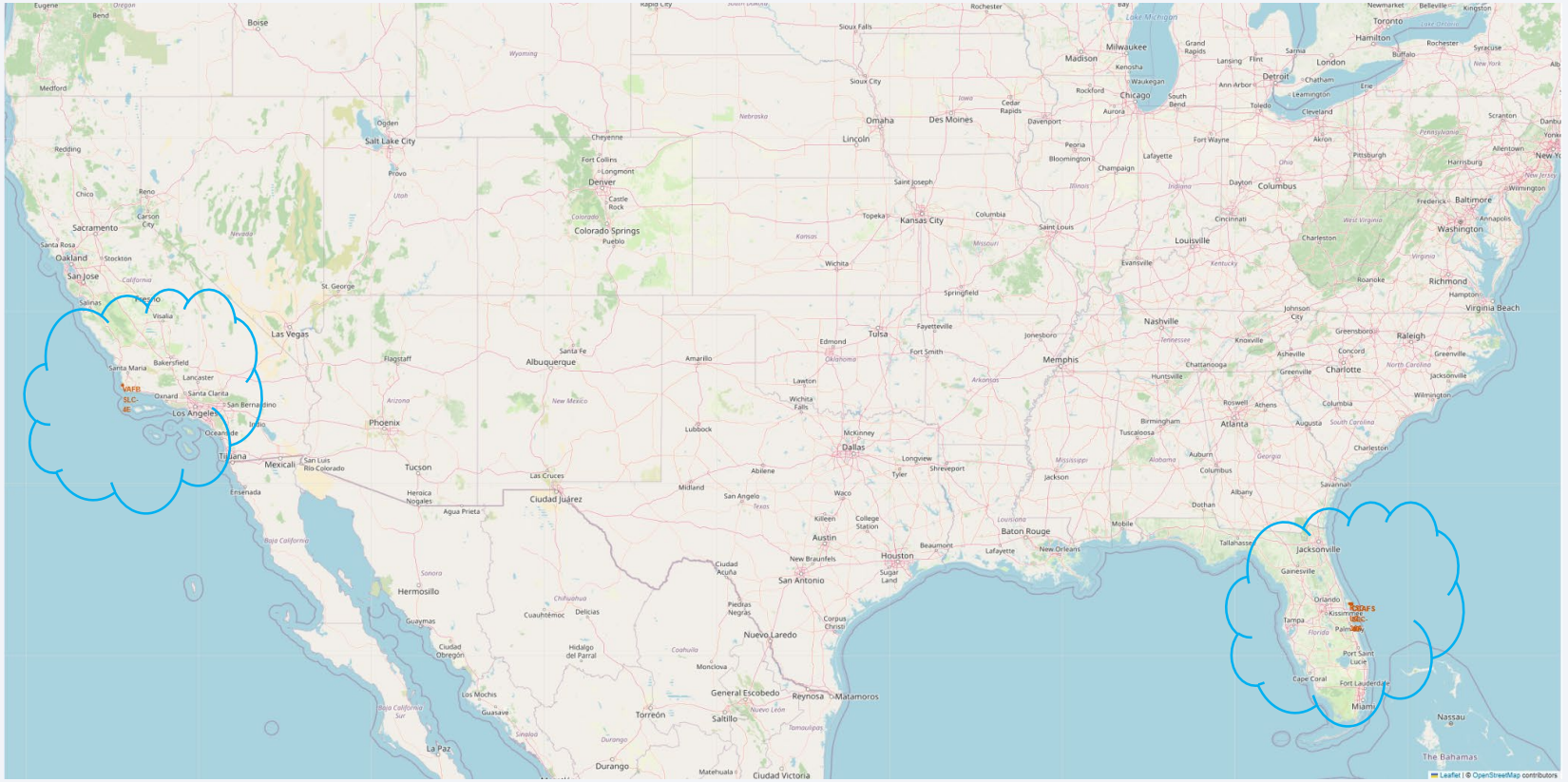
```
%sql select count(landing_outcome), landing_outcome from SPACEXTBL \
where DATE between '2010-06-04' and '2017-03-20' \
GROUP BY landing_outcome \
ORDER BY count(landing_outcome) desc
```

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

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

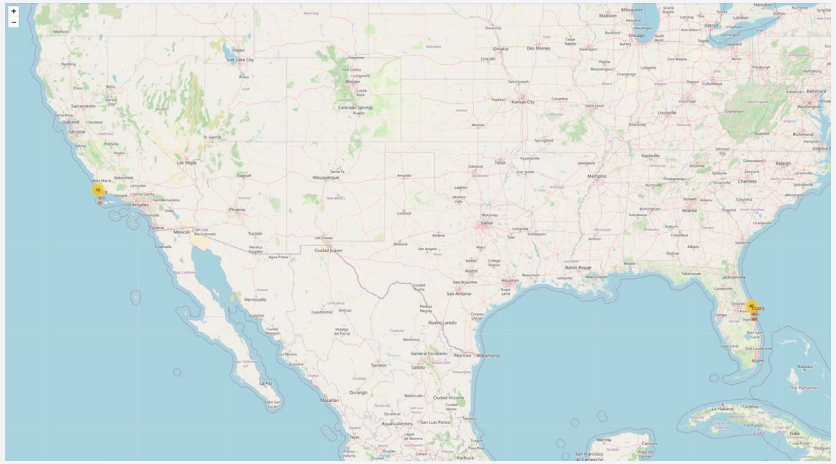

Section 3
Launch Sites
Proximities Analysis

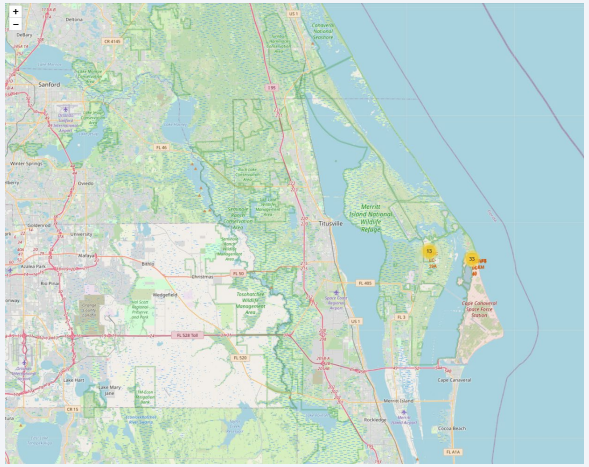
Global Map of Falcon 9 Launch Sites



- The launch site distribution reveals that VAFB SLC-4E is the only facility based in California, with the remainder positioned along the Florida coastline.

Visual Mapping of Launch Site Landing Outcomes



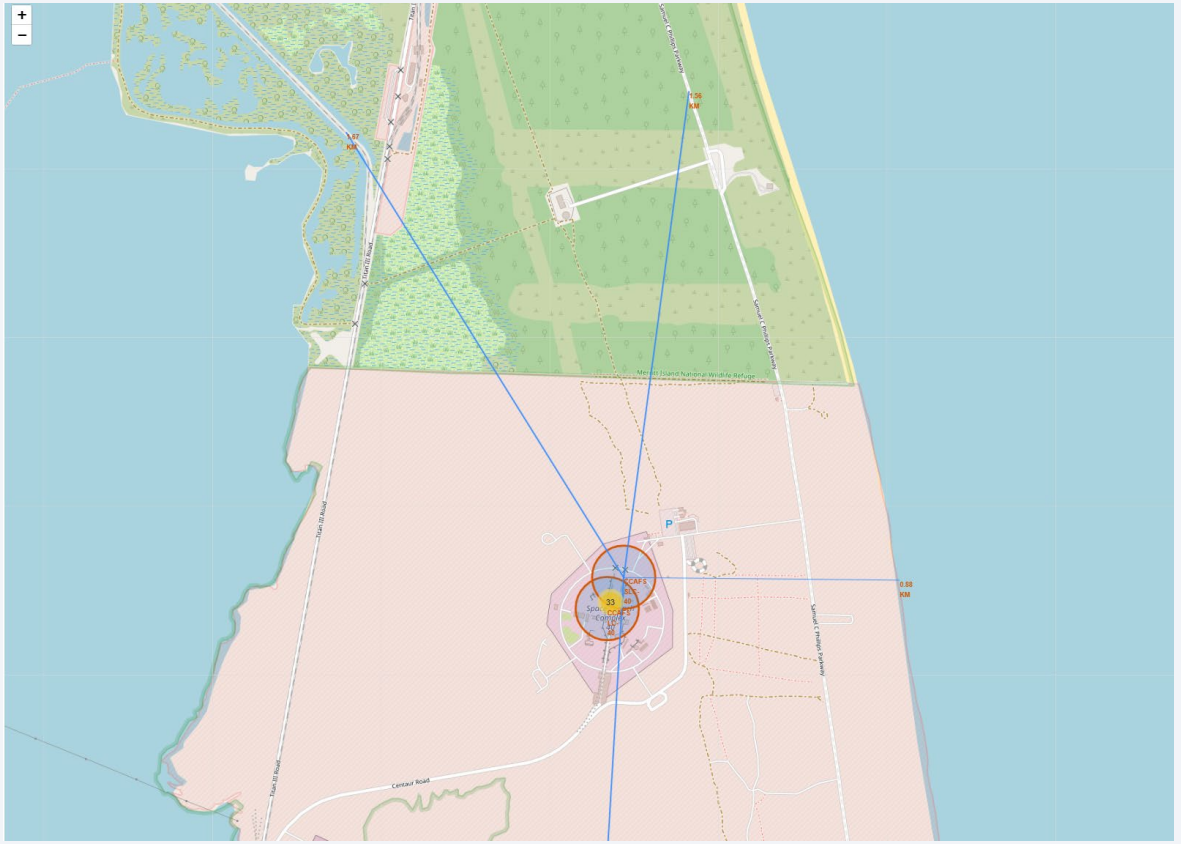


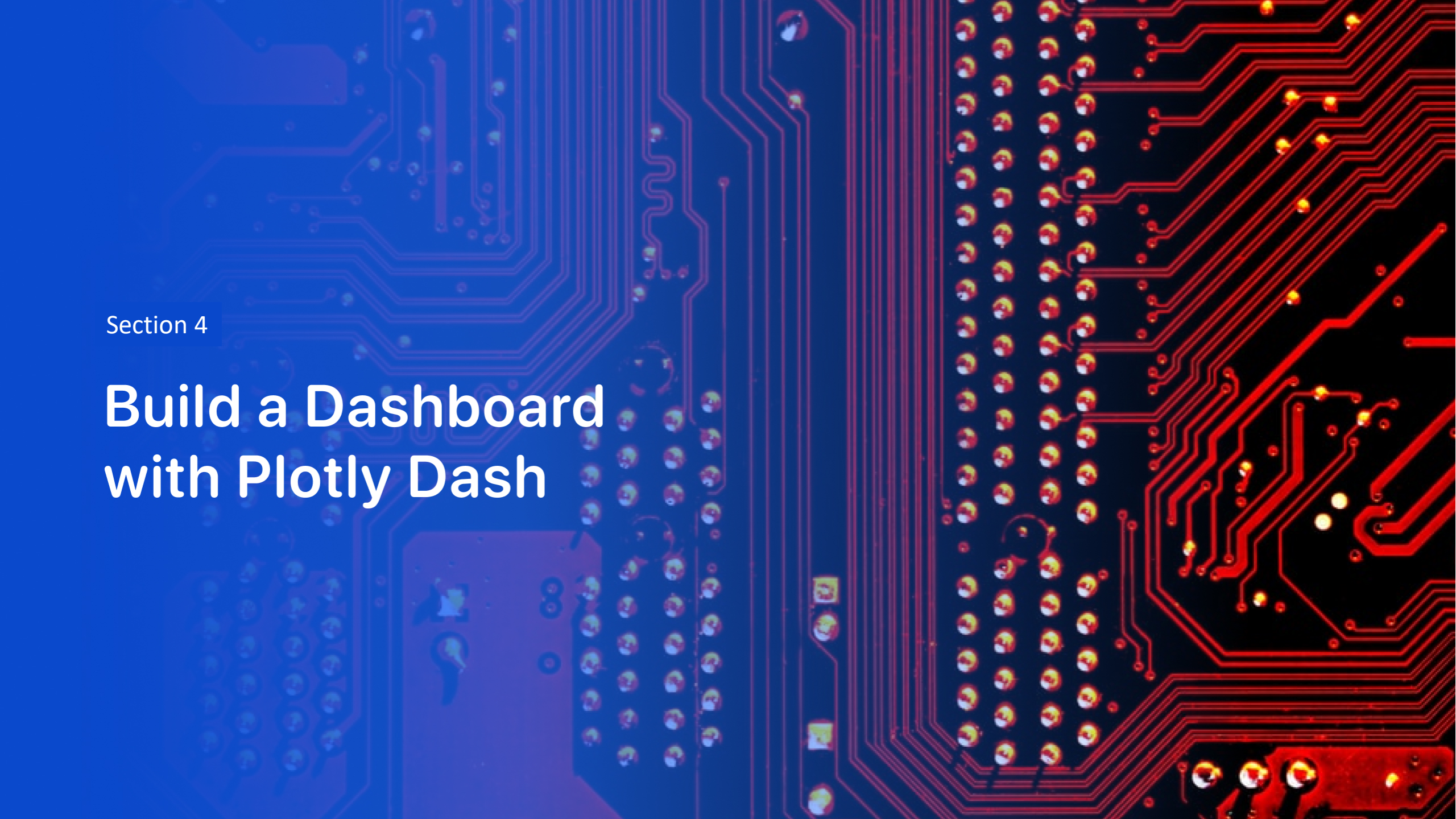
The mission outcomes for Falcon 9 first stage landings are represented using visual markers on the map, color-coded to indicate success (green) or failure (red).

These markers are geographically grouped based on the coordinates of the corresponding launch site.

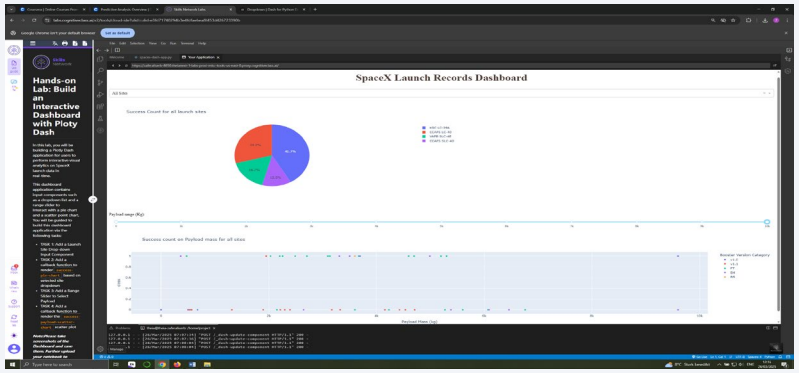
Geospatial Proximity of Launch Sites to Major Landmarks

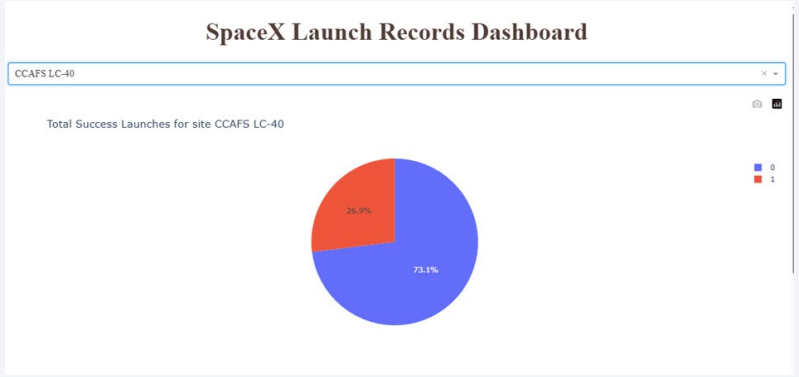
- The geographical positioning of the site—located approximately **1.67 km from a railway** supporting the transport of heavy infrastructure,
- **1.56 km from a highway** facilitating operational mobility,
- **0.88 km from the coastline** providing abort and recovery pathways over water,
- and **54.03 km away from the nearest urban area** to minimize civilian exposure—reflects a deliberate emphasis on operational efficiency and safety in launch site selection.

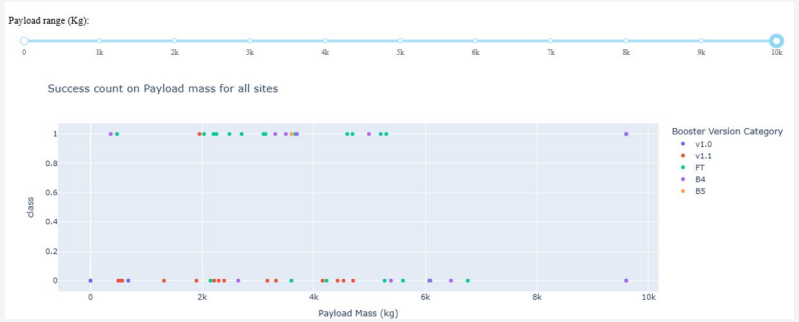



Section 4
Build a Dashboard with Plotly Dash

Success Rate Quantification Across SpaceX Launch Complexes

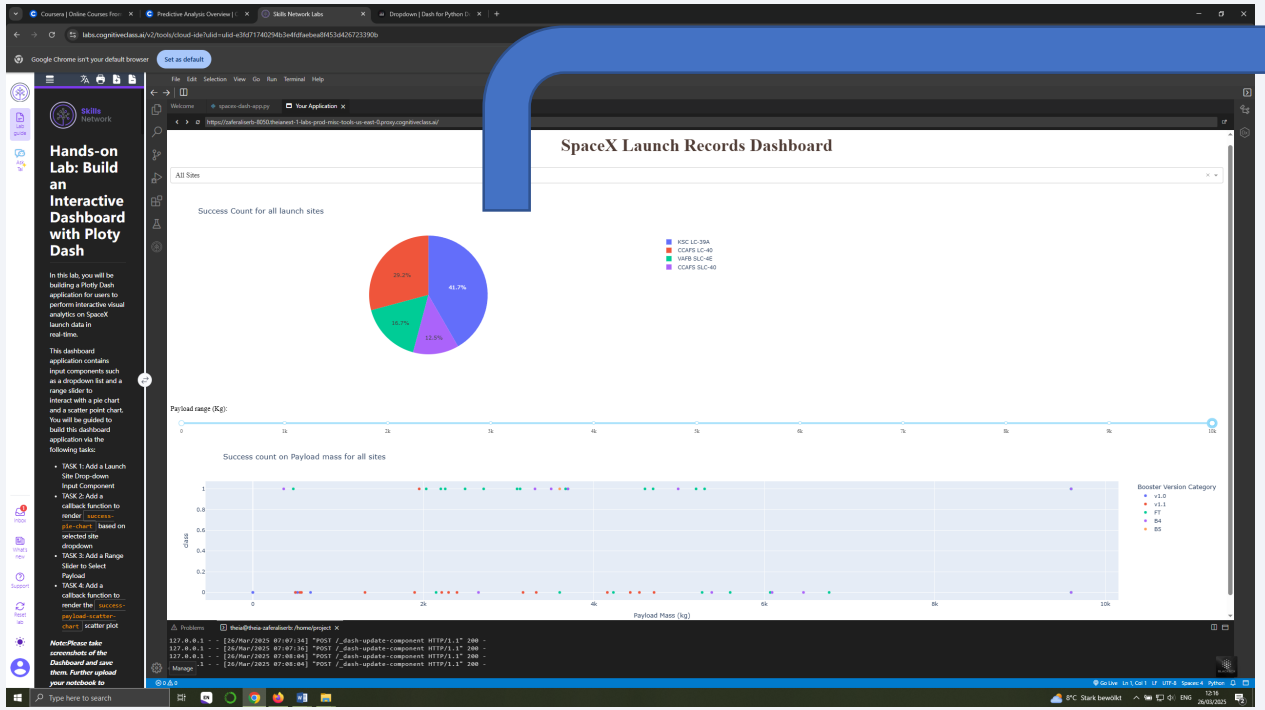


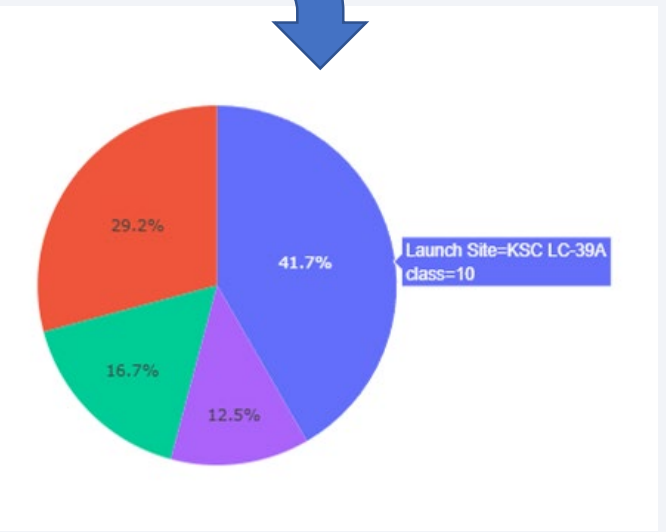




- **Dropdown Menu:** Allows users to select either all launch sites or specific individual sites for analysis.
- **Range Slider:** Provides the functionality to filter payload mass within a specified range, enabling detailed exploration of payload data.
- **Scatter Chart:** Visualizes the relationship between payload mass and landing outcomes, offering insights into the performance of each launch.
- **Pie Chart:** Illustrates the success rate of booster landings, providing a clear breakdown of landing successes and failures.
- **Color Coding by Booster Version:** Differentiates landing outcomes by booster version, enhancing the clarity of trends and comparisons.
- **Color Coding by Launch Site:** Distinguishes performance metrics based on the launch site, enabling site-specific analysis.

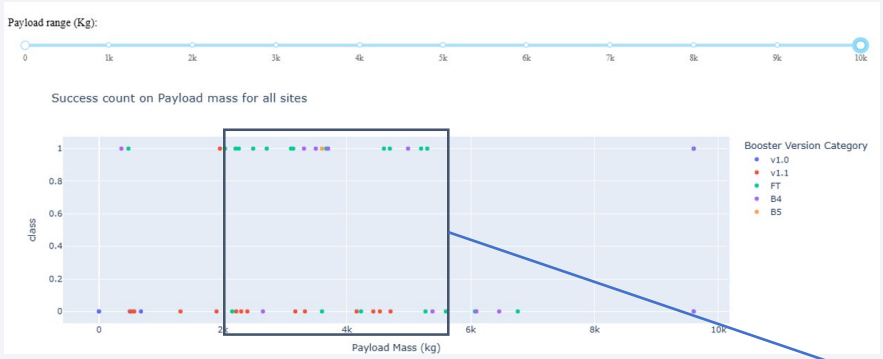
The highest proportion of successful Falcon 9 first stage landings

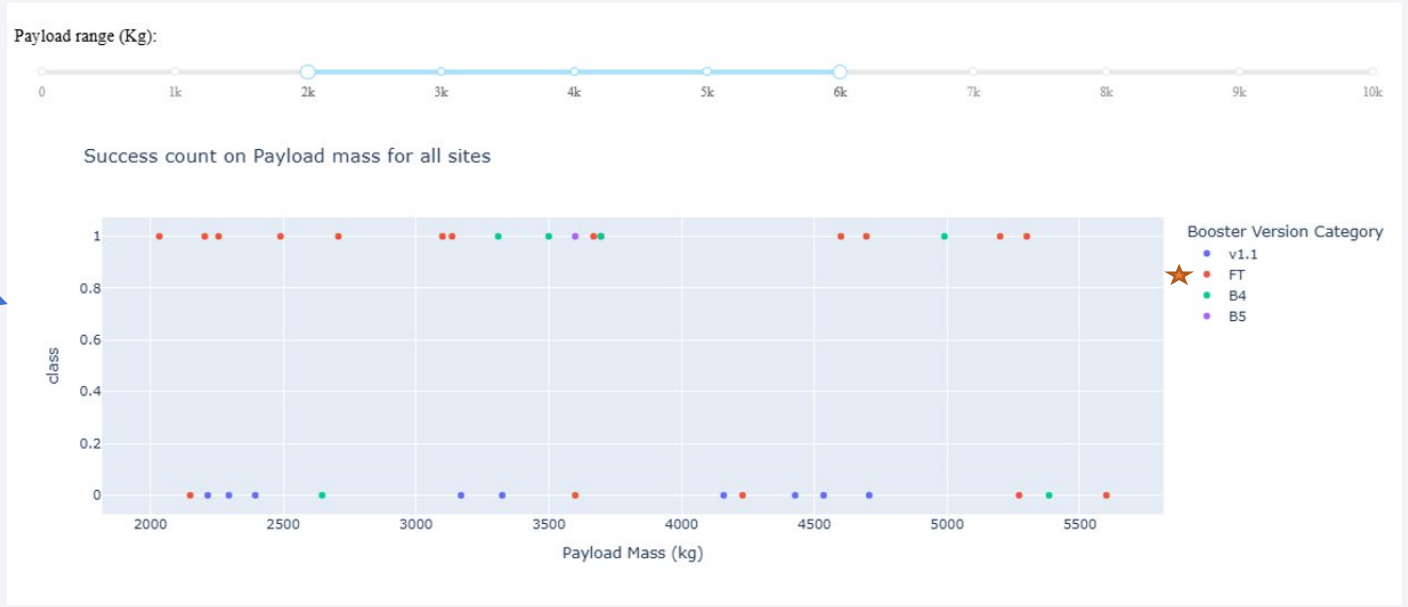




- The highest proportion of successful Falcon 9 first stage landings, accounting for **41.7%** of the total, was observed at **KSC LC-39A**.

Exploring the Link Between Payload Weight and Launch Succes





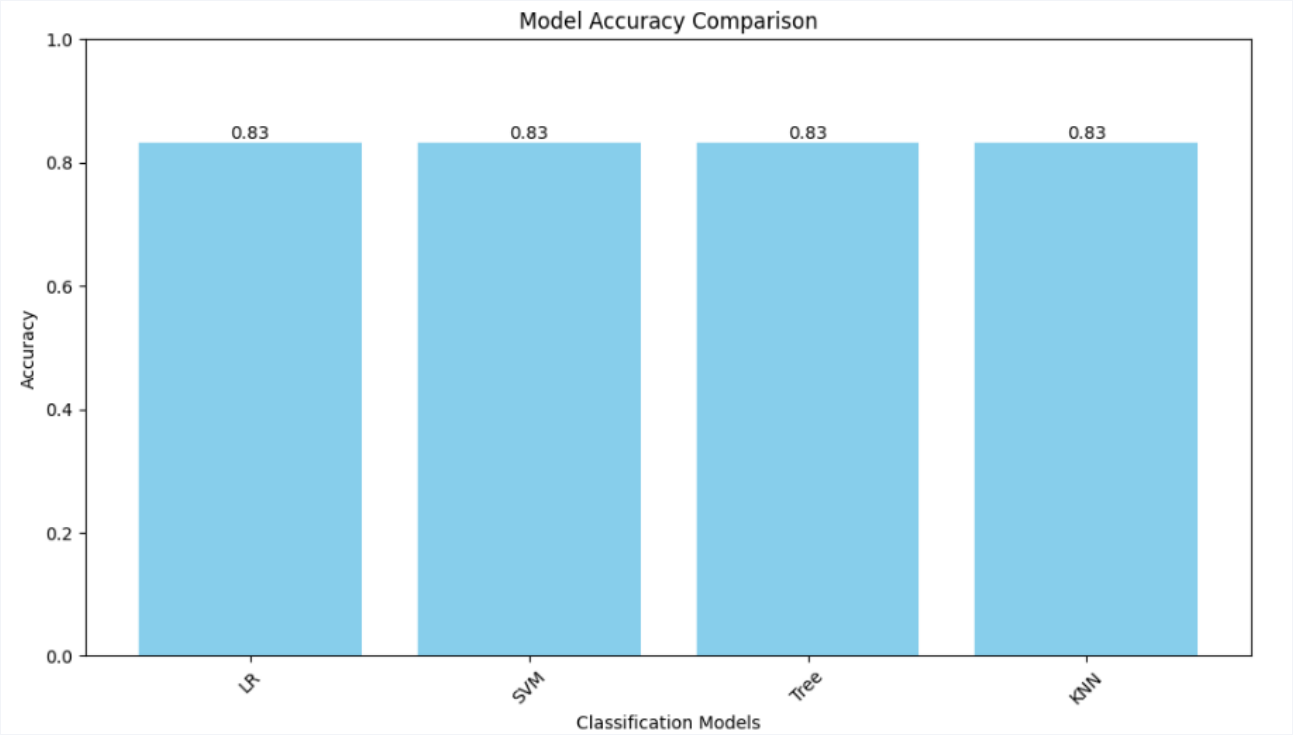
- "The '**FT**' booster version category demonstrates **the greatest success rate** among all variants."

- "The payload range between approximately **2000 kg and 5000 kg** exhibits the highest success rate."

Section 5

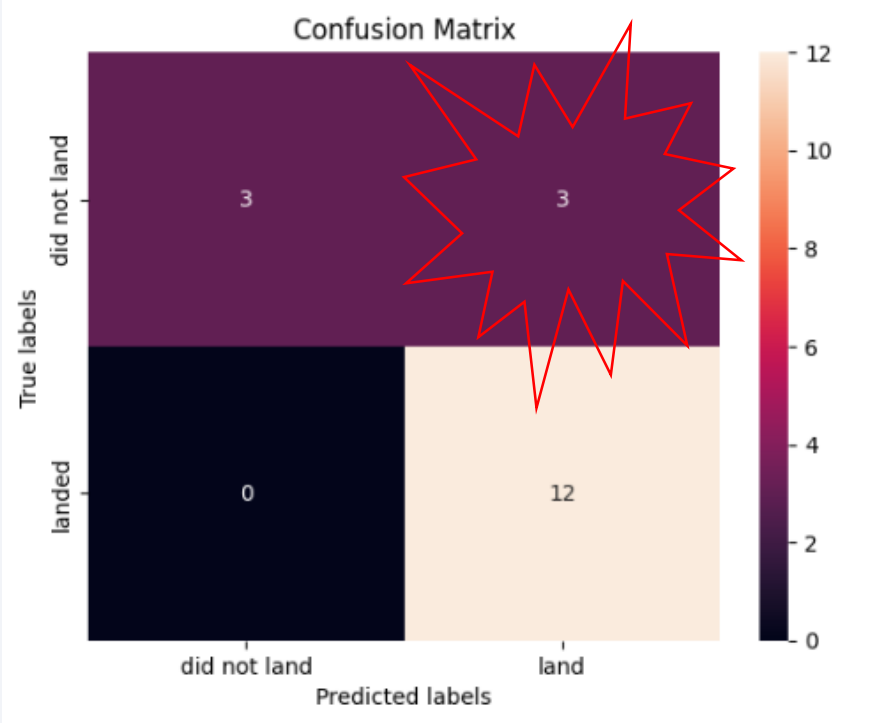
Predictive Analysis (Classification)

Classification Accuracy



- During the **testing** phase, all four models yielded identical accuracy scores of **83.33%.**

Confusion Matrix



- The confusion matrices for models are **identical**.
- The primary issue observed is the occurrence of **false positives**, as demonstrated by the models incorrectly predicting the first stage booster landing in **3** **out of 18** test samples.

Conclusions

- **Correlation between Launch Frequency and Success Rate:**

  A **higher number of flights** at a given launch site correlates with an **increased success rate** for that site.

- **Orbits with Highest Success Rates:**

  ES-L1, **SSO**, **HEO**, and **GEO** orbits exhibit the highest success rates for SpaceX launches.

- **Payload Weight and Success Rate:**

  For **heavier payloads**, success rates tend to improve for **Polar**, **LEO** and **ISS** orbits.

- **Launch Success Trends:**

  The launch success rate showed a steady **increase from 2013 to 2020**, with only minor fluctuations (*2018) during this period.

- **Top Performing Launch Site:**

  KSC LC-39A holds the highest launch success rate among SpaceX launch sites.

- **Prediction Accuracy with Models:**

  Using the models developed in this report, SpaceY can predict SpaceX's success in landing the first stage booster with **83.3%** accuracy.

Conclusions

Future Opportunities for Improved Bidding Predictions:

- Freeze the best-performing model and hyperparameters, and re-train it with the full dataset instead of just training data for better accuracy.
- Incorporate additional launch data as it becomes available to further refine predictions.

Appendix

- **SpaceX API (JSON):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- **Wikipedia (Webpage):** https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- **Respectfully, and with sincere appreciation;**

Yan Luo, Ph.D., Data Scientist and DeveloperIBM

Joseph Santarcangelo, Ph.D., Data Scientist at IBMIBM Developer Skills Network

- GitHub URL:

https://github.com/zaferaliserbes/IBM-Data-Science-Capstone-Project/tree/main

Thank you!