IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Selena Liu>
<6/25/2025>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection & Wrangling: API, Wikipedia launch table, missing value replace with mean, filter, feature engineering (standardize numerical columns, one-hot categorical features).
  - Exploratory Data Analysis (EDA): strip plot, box plot, bar chart, line plot.
  - Geospatial Visualization: Folium Map, Launch Events, Proximity Analysis.
  - Machine-Learning Modeling: Train/Test Split, Logistic Regression. Support Vector Machine (sigmoid), Decision Tree, K-Nearest neighbors, Evaluation (LogReg/SVM, Tree/KNN).
- Summary of all results
  - All classifiers achieved perfect recall (no landed rocket was ever missed) but differed in false-positive rates.
  - Decision Tree performed best (≈ 94% accuracy), cutting false alarms to one.
  - Next steps could include threshold tuning, ensemble methods, or richer feature engineering to drive precision even higher.

Introduction

- Project background and context:

  SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

- Problems you want to find answers:

  In this capstone, we will predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium
- Perform predictive analysis using classification models

Data Collection

- Describe how data sets were collected.

❑ SpaceX REST APR – JSON – DataFrame

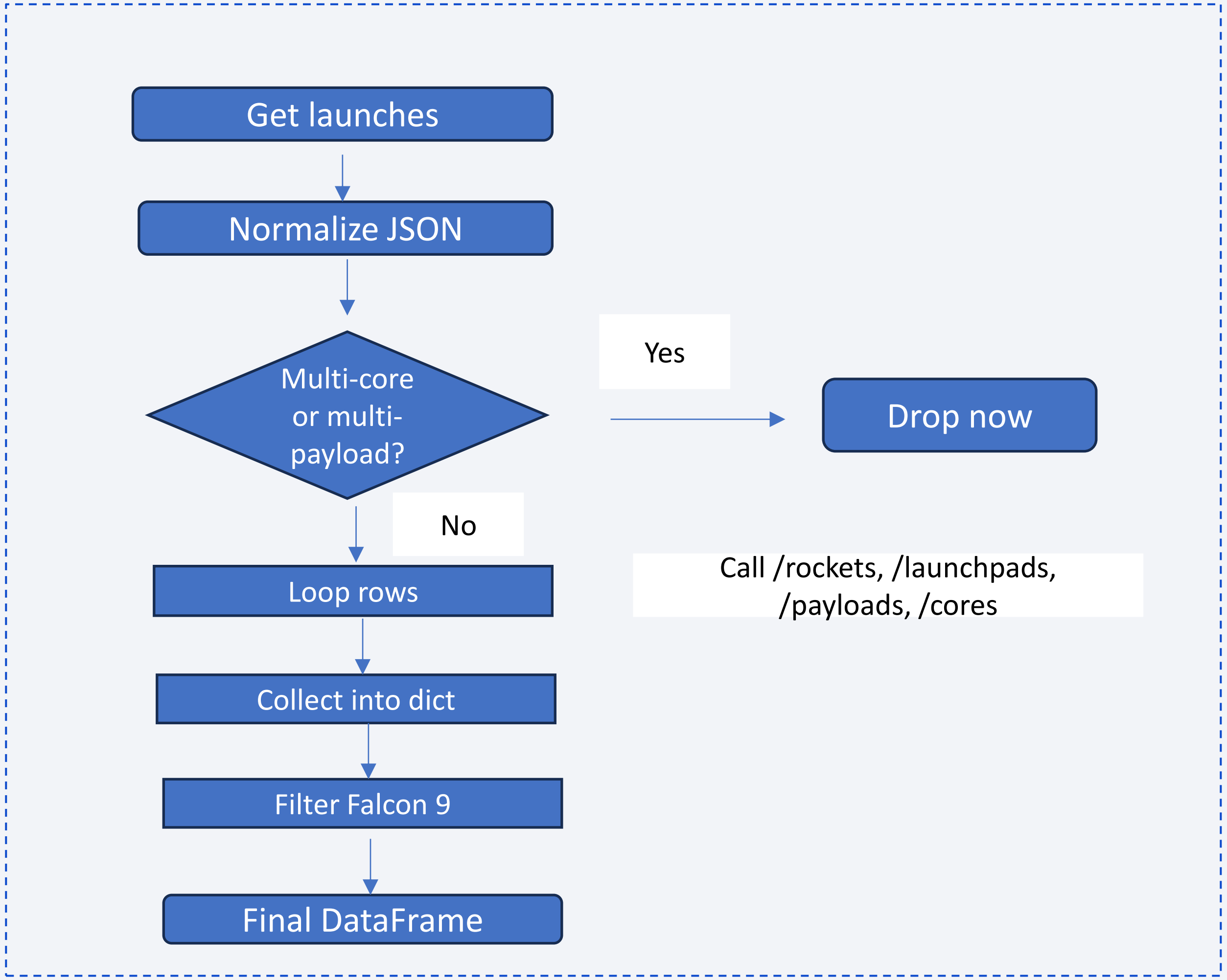
- ➢ Fetch launch records: use a “static JSON endpoint to lock in results.
- ➢ Filter & Unwrap: keep only the columns we need, drop multi entries, extracted single IDs for their one-element lists, and convert to a proper date and filter by cutoff
- ➢ Enrich with helper functions: looped over each ID column and call the corresponding v4 endpoint, store each new field in a global list, then built a final launch_dict and convert it to a DataFrame.
- ➢ Porst-processing: filtered to only Falco 9 launches, reset flight numbers, handled missing values.

❑ Wikipedia Web-scraping – BeautifulSoup – DataFrame

- ➢ Request the static Wiki page
- ➢ Locate the launch table: found all table elements, picked the table which contains the Falcon 9 launch records.
- ➢ Extract column names: iterated its table header cells, clean each header string into a flat list of column names.
- ➢ Parse each data row: loop over <tr> rows, detect those whose <th> is a digit. For each, pull out the <td> cells and applied helpers to get
    - Data & time
    - Booster version
    - Payload info
    - Landing status
- ➢ Append each filed into a launch_dict, then built a DataFrame.

Data Collection – SpaceX API

- Data Collection Flowcharts



Get launches

Normalize JSON

Multi-core or multi-payload?

Yes

Drop now

No

Loop rows

Call /rockets, /launchpads, /payloads, /cores

Collect into dict

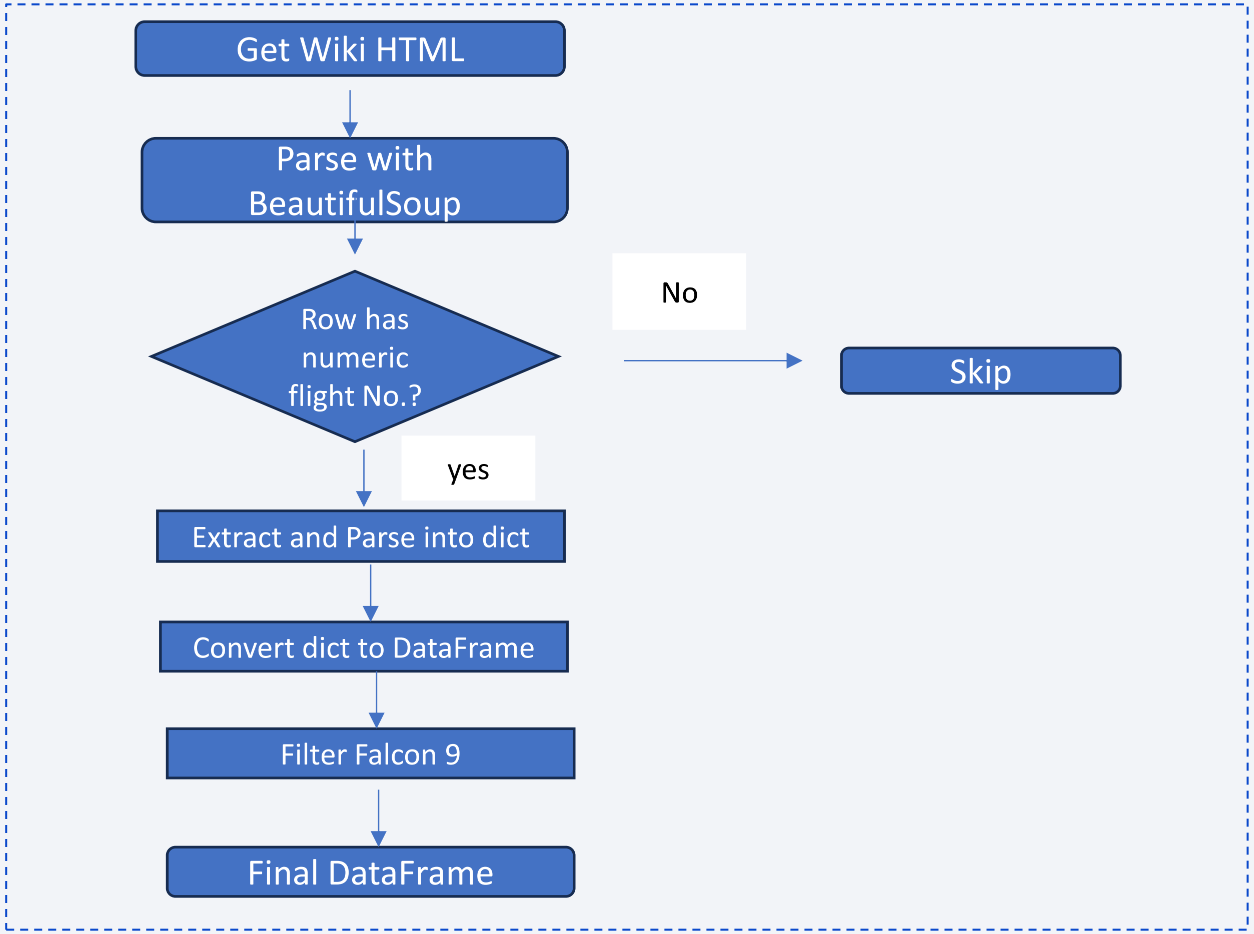
Filter Falcon 9

Final DataFrame

- GitHub URL: https://github.com/Selenaliu357/Data-Science-Capstone---First-Stage-of-Falcon-9-Landing-Prediction/blob/main/Collecting%20Data%20By%20API.ipynb

Data Collection - Scraping

- Web Scraping Flowcharts

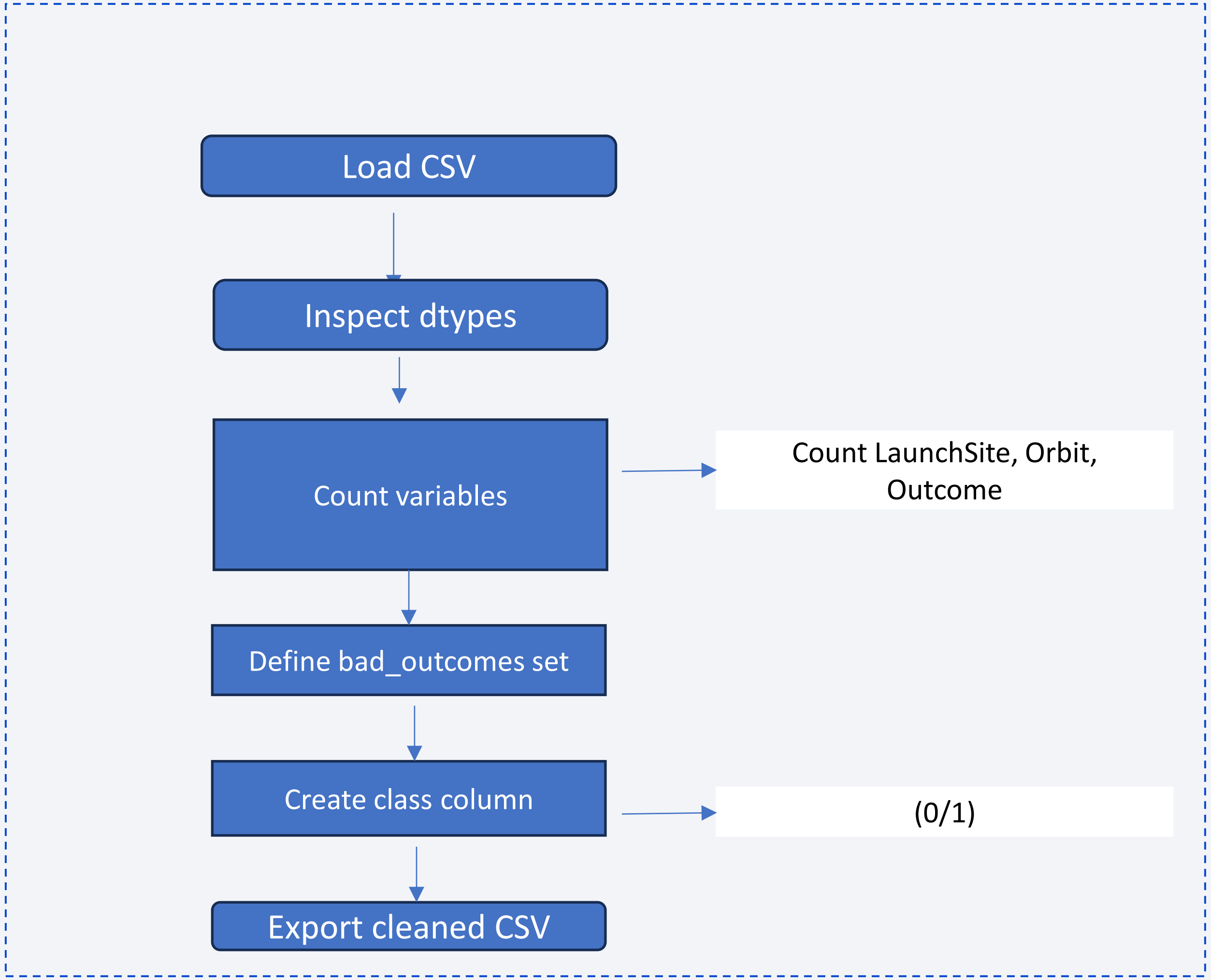

Get Wiki HTML
Parse with BeautifulSoup
Row has numeric flight No.?
No
Skip
yes
Extract and Parse into dict
Convert dict to DataFrame
Filter Falcon 9
Final DataFrame

- GitHub URL: https://github.com/Selenaliu357/Data-Science-Capstone---First-Stage-of-Falcon-9-Landing-Prediction/blob/main/Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- Describe how data were processed:
  - Load CSV - pd.read_csv
  - Missingness - % missing
  - Dtypes – df.dtypes
  - Values counts – LauchSite, Orbit and Outcome
  - Bad-outcome = pick indices [1,3,5,6,7] from landing_outcomes.keys()
  - Binary class
  - Success rate – df["Class"].mean( )
  - Export – df.to_csv

- GitHub URL: https://github.com/Selenaliu357/Data-Science-Capstone---First-Stage-of-Falcon-9-Landing-Prediction/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Scatter Plot: sns.catplot(x='Xvar', y='Yvar', hue='Category', data=df), we use scatter plot to show the **relationship** between two continuous variables at the individual-point level.
  - Box Plot: sns.boxplot(x='Category', y='Value', data=df), we use it to summarizes the **distribution** of a continuous variable for each category: median, quartiles, and outliers.
  - Bar Chart: sns.barplot(x='Category', y='Value', estimator='mean', data=df), it is best for **aggregate** comparisons—counts or summary statistics (mean, sum) per category.
  - Line Plot: sns.lineplot(x='TimeOrSeq', y='Metric', data=summary_df, marker='o'), it shows **trends** over an ordered or continuous variable (often time, flight number, etc.).

- GitHub URL: https://github.com/Selenaliu357/Data-Science-Capstone---First-Stage-of-Falcon-9-Landing-Prediction/blob/main/Exploring%20and%20Preparing%20Data%20(EDA).ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - SELECT DISTINCT
  - WHERE … LIKE
  - SUM(…), AVG(…), MIN(…), MAX(…)
  - GROUP BYORDER BY … DESC
  - Sub-query in WHERE (e.g. = (SELECT MAX(...)))
  - CASE … WHEN mapping for month names
  - substr(...) for date string slicing

- GitHub URL: https://github.com/Selenaliu357/Data-Science-Capstone---First-Stage-of-Falcon-9-Landing-Prediction/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

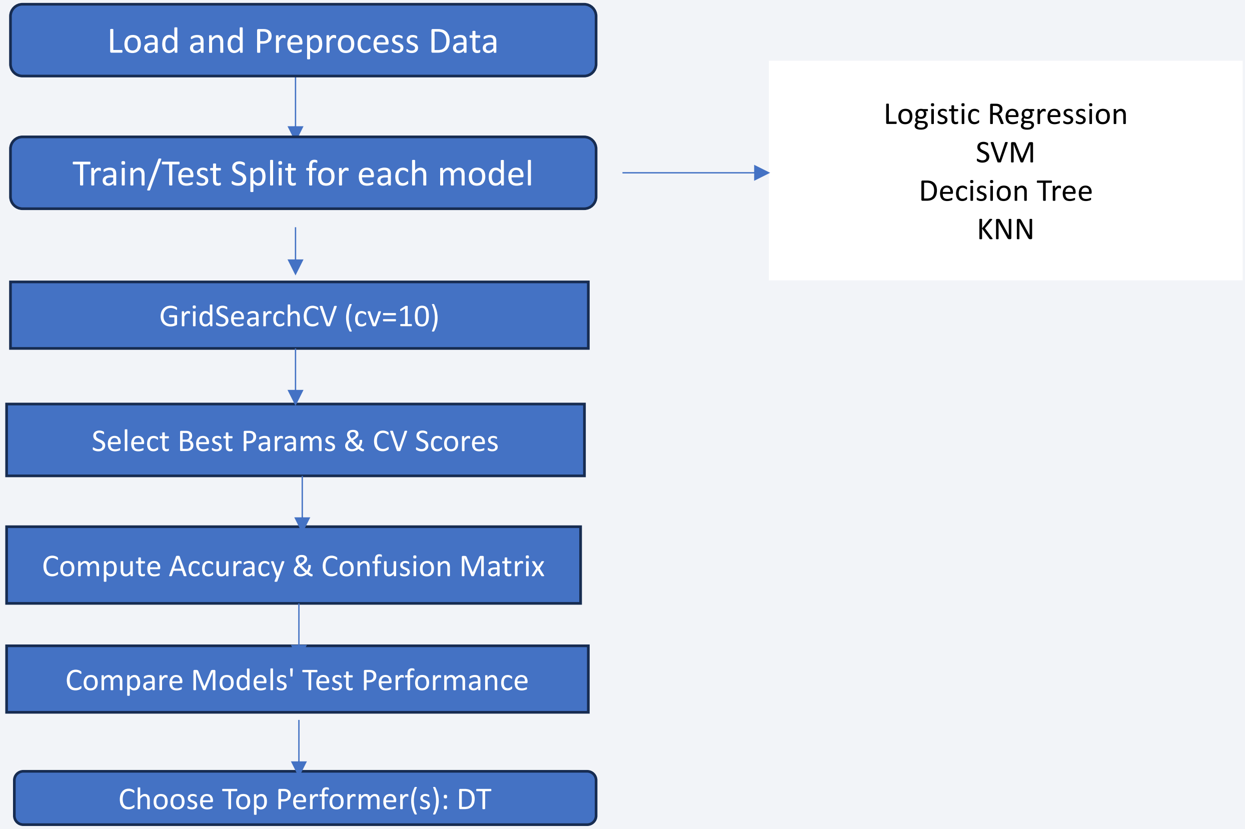
- Quick rundown of the Folium objects we added and their purposes:
  - **MousePosition:** Shows your cursor's latitude/longitude in real time—helps you pick exact points for coastlines, cities, highways, etc.
  - **MarkerCluster:** Groups many point markers into clusters at low zoom levels to reduce clutter, then expands them as you zoom in.
  - **Circle:** Draws a solid circle (e.g. 1 km radius) around each launch site to highlight its footprint or exclusion zone, with a popup for the site name.
  - **Marker + DivIcon:** Marks each individual launch event (clustered) and color-codes them green/red to indicate success or failure.
  - **PolyLine:** Connects two points (launch site ↔ coastline, city, highway, etc.) with a colored line, visually showing proximity relationships

  We add these to Annotate key locations, Encode outcomes, Communicate distances and Improve usability.

- GitHub URL: https://github.com/Selenaliu357/Data-Science-Capstone---First-Stage-of-Falcon-9-Landing-Prediction/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Predictive Analysis (Classification)

Model Development Process



Load and Preprocess Data
Train/Test Split for each model
Logistic Regression
SVM
Decision Tree
KNN
GridSearchCV (cv=10)
Select Best Params & CV Scores
Compute Accuracy & Confusion Matrix
Compare Models' Test Performance
Choose Top Performer(s): DT

GitHub URL: https://github.com/Selenaliu357/Data-Science-Capstone---First-Stage-of-Falcon-9-Landing-Prediction/blob/main/Machine%20Learning%20Prediction.ipynb

Results

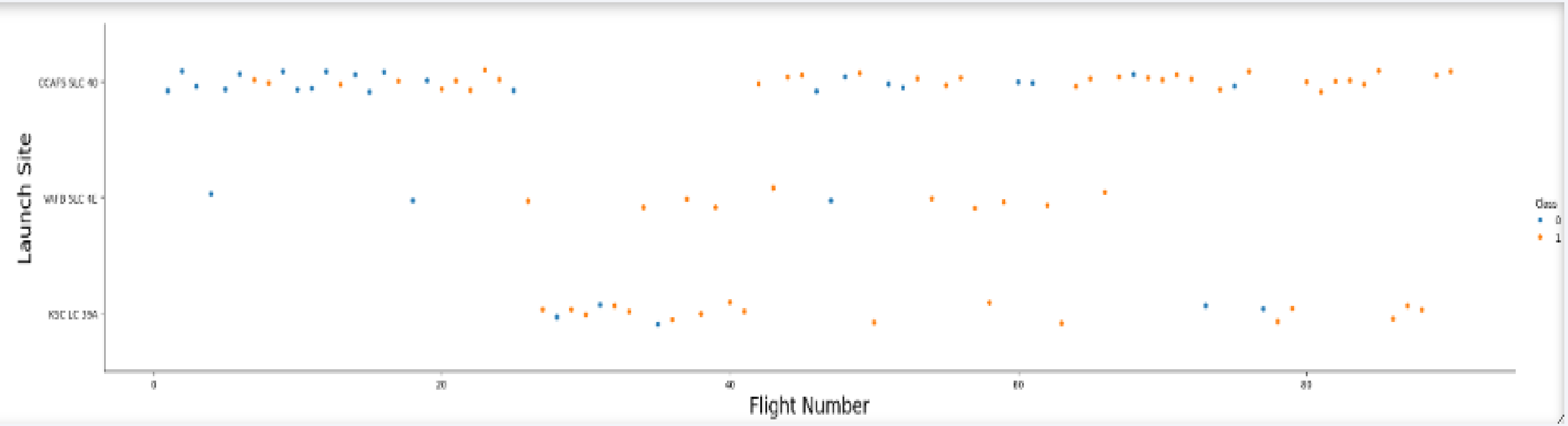
- Exploratory data analysis results
  - We can observe that the success rate since 2013 kept increasing till 2020.
  - More complex or higher-energy orbits (e.g. GTO, interplanetary SO) show lower landing success—reflecting the extra challenge of those mission profiles.
  - SpaceX's booster landing success rate improved steadily with each new flight, and all three pads ended up with very high success rates by around flight 50+.
- Interactive analytics demo in screenshots
  - The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site.
- Predictive analysis results
  - Decision Tree performed best (≈ 94% accuracy), cutting false alarms to one.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site



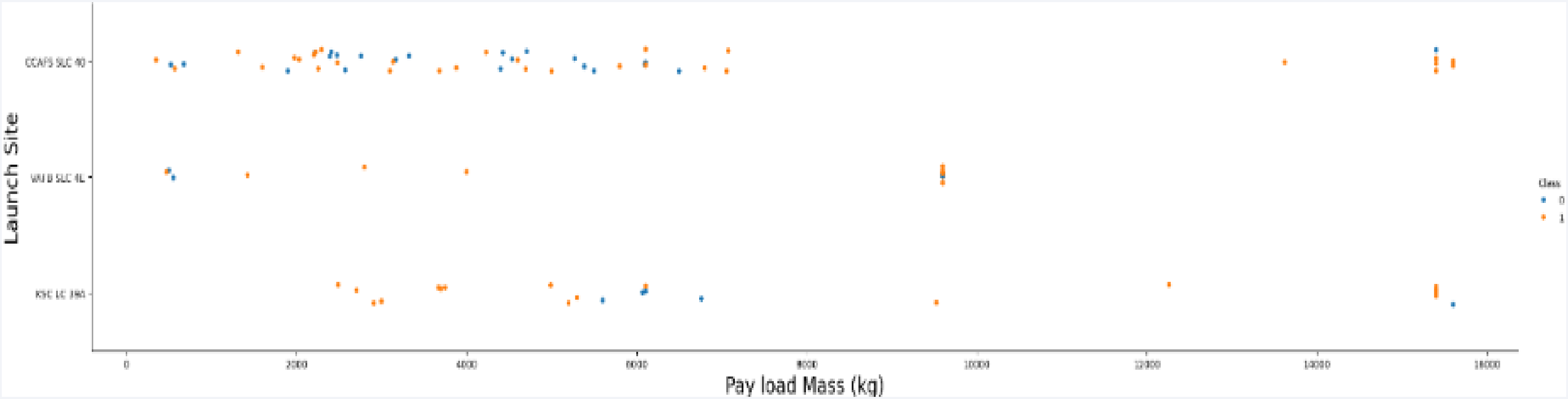
- Explanations:

  * SpaceX’s booster landing success rate improved steadily with each new flight,

  * and all three pads ended up with very high success rates by around flight 50+.

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site



- Explanations:
    - Payload Mass Vs. Launch Site scatter point chart for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- Explanations:

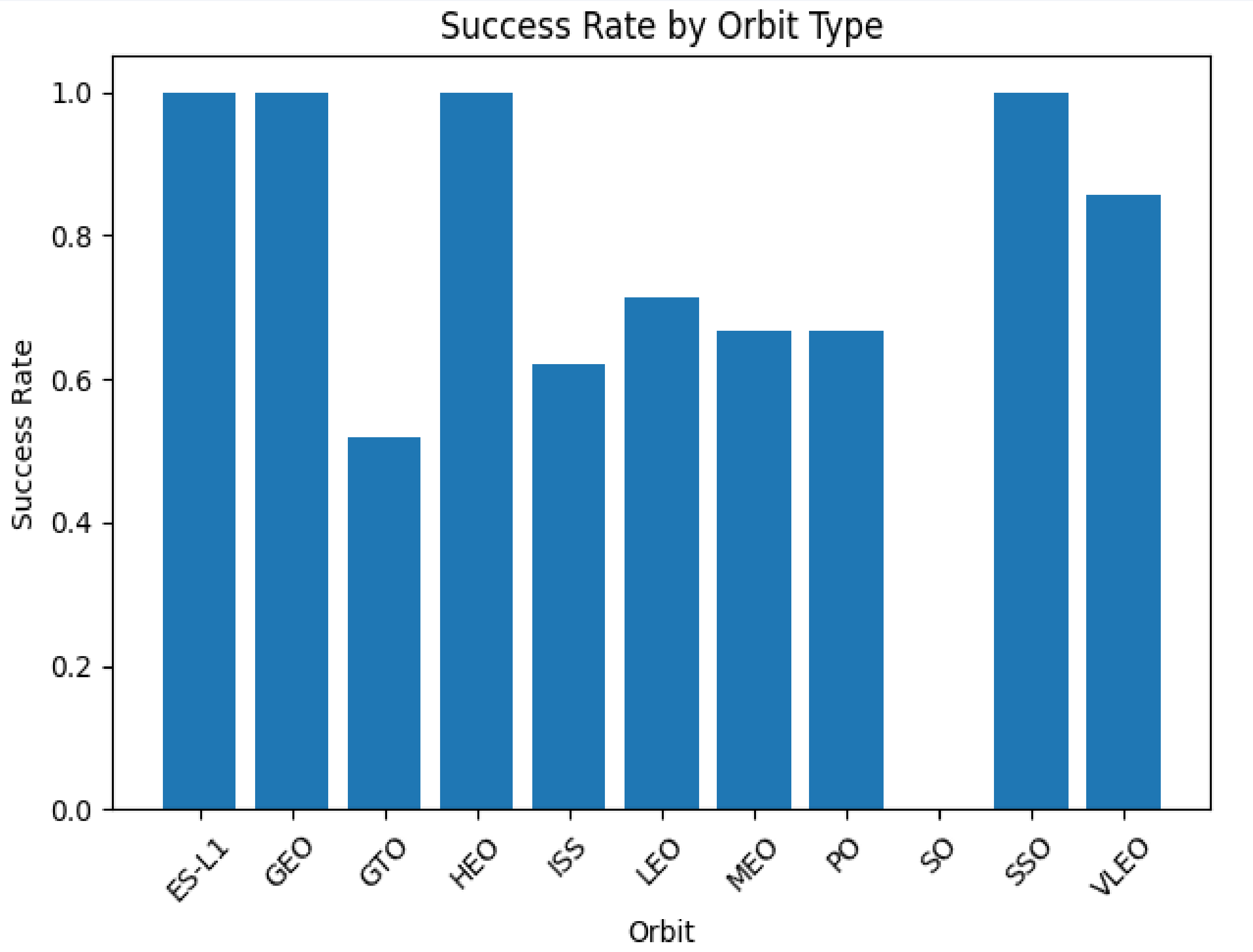
- Orbits with 100 % success: ES-L1, GEO, HEO, SSO

- High but not perfect: VLEO (~85 %) and LEO (~70 %)

- Moderate success: ISS (~62 %), MEO/PO (~67 %)

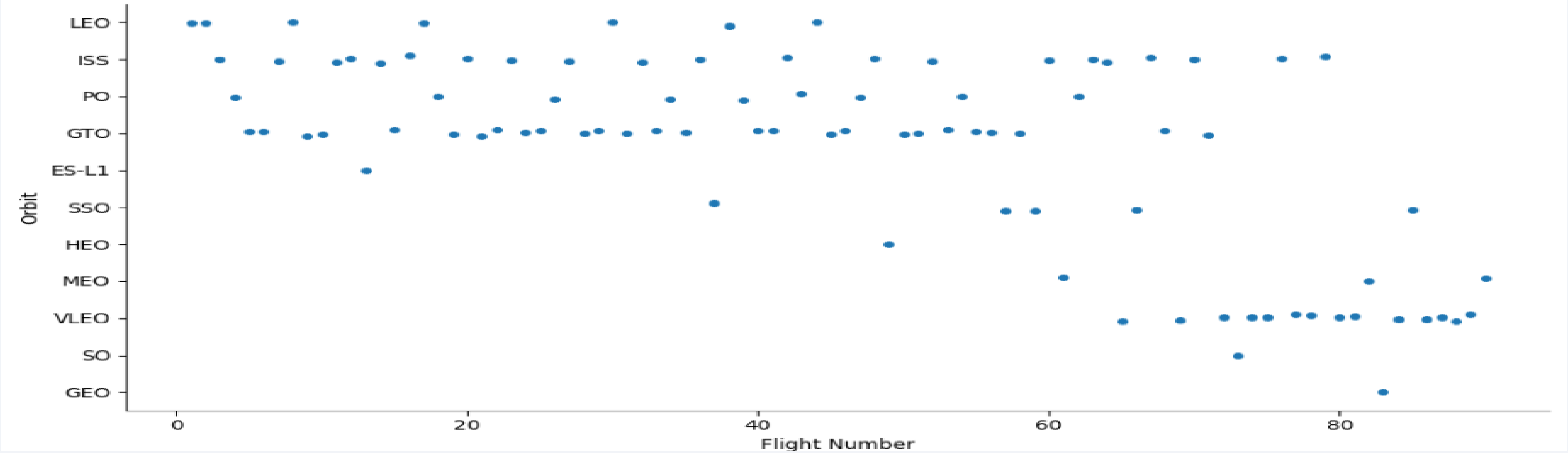
- Lowest success: GTO (~52 %) and SO (0 %)

- **Buttom Line**: More complex or higher-energy orbits (e.g. GTO, interplanetary SO) show lower landing success—reflecting the extra challenge of those mission profiles.



Flight Number vs. Orbit Type

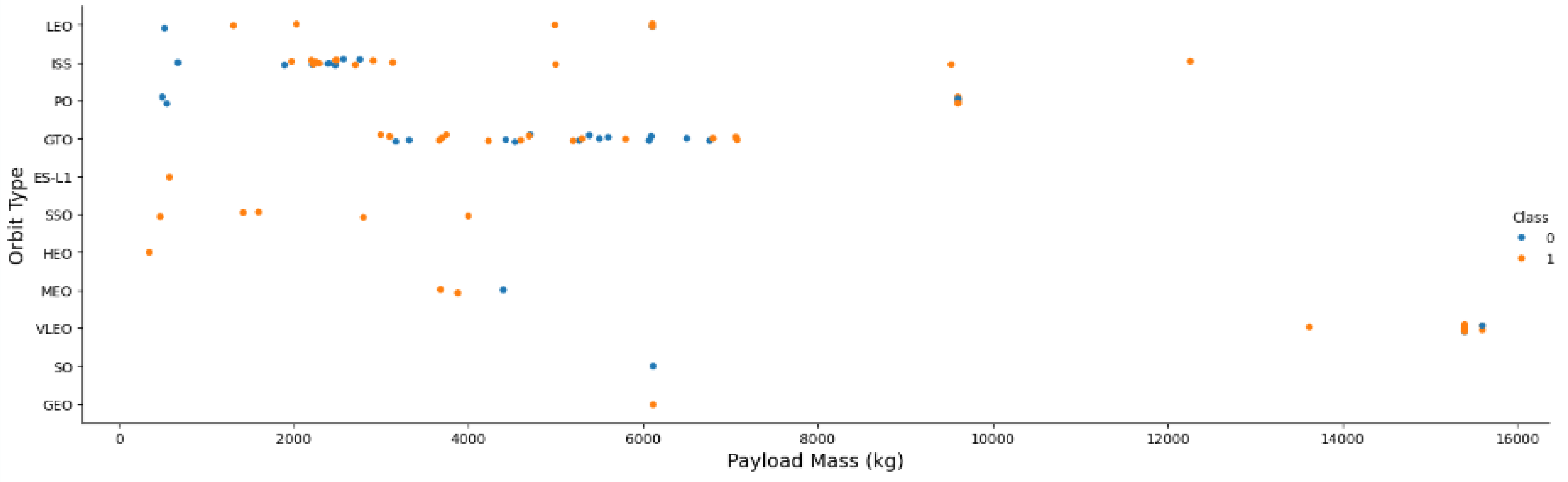
- Scatter point of Flight number vs. Orbit type



- Explanations:
  - We can observe that in the 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

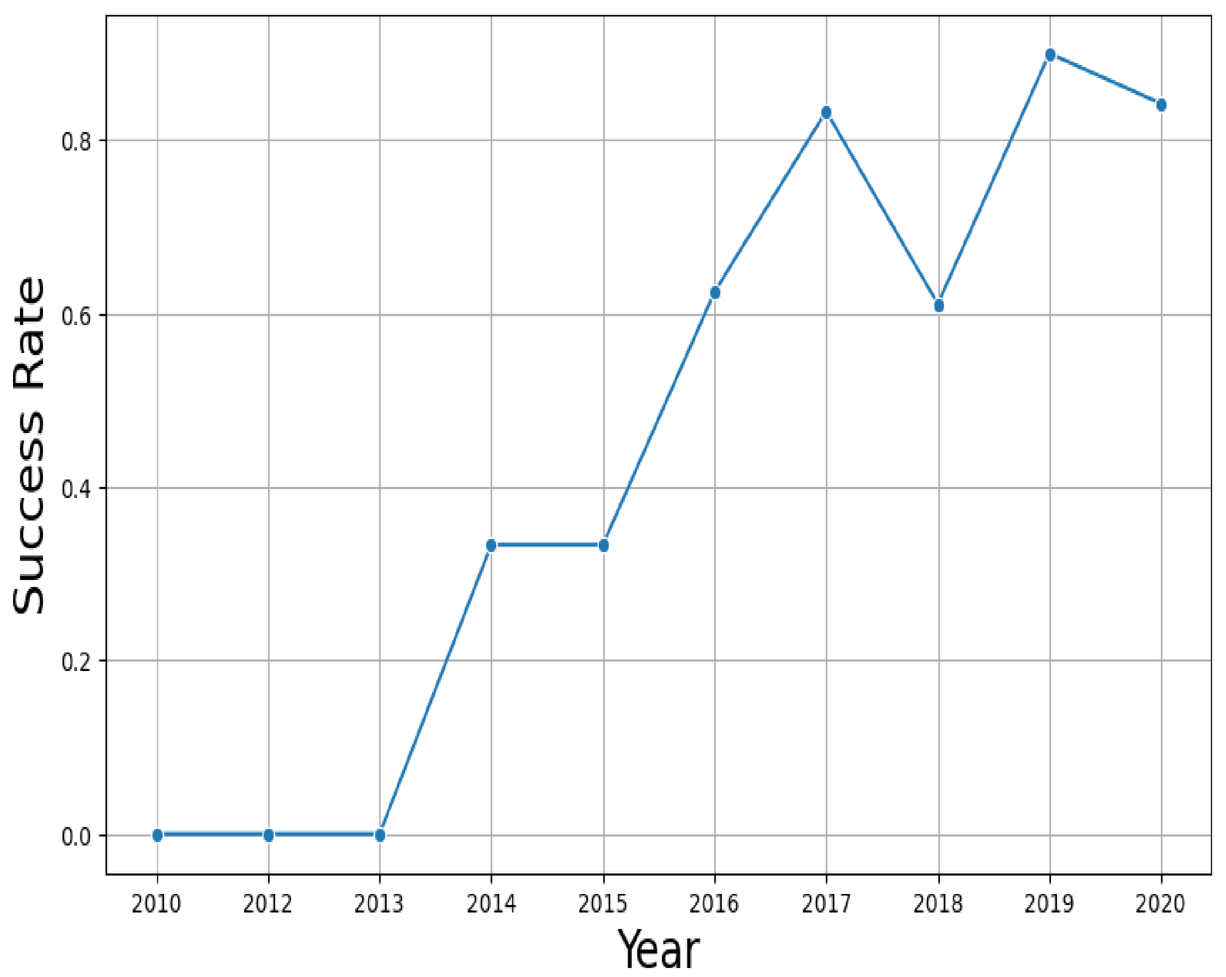
- Scatter point of payload vs. orbit type



- Explanations:
  - With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- Explanations:
  - We can observe that the success rate since **2013** kept increasing till 2020.

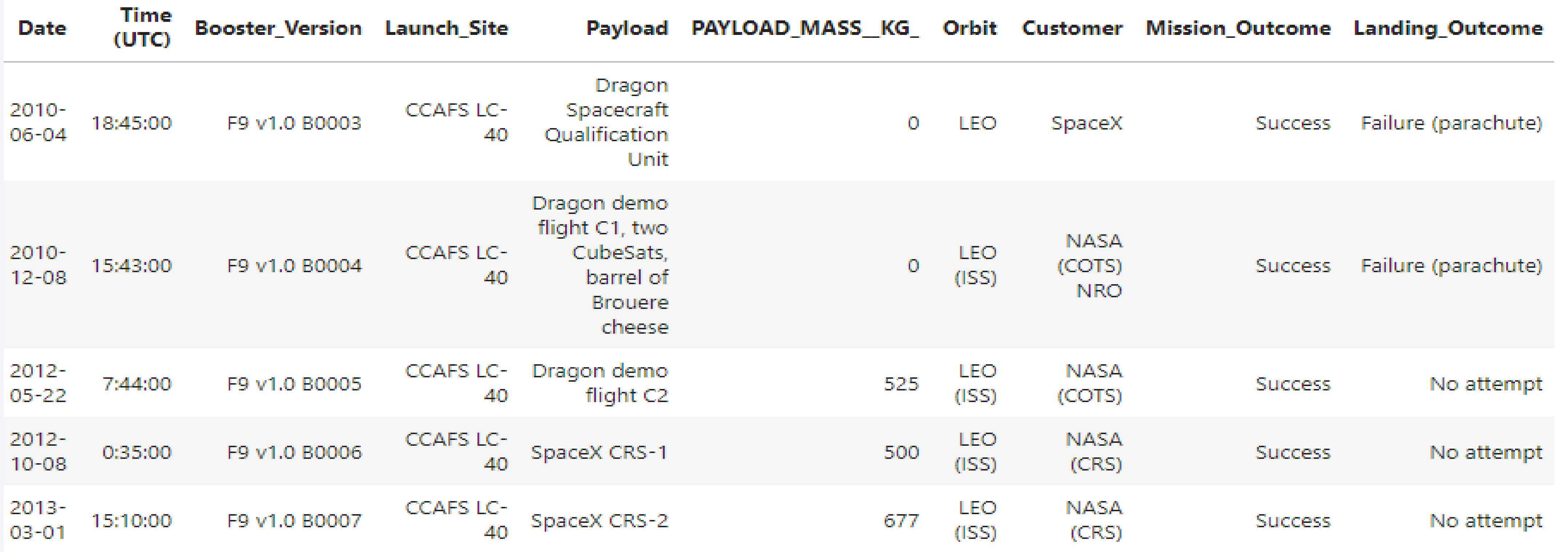


All Launch Site Names

- The names of the unique launch sites:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- Explanation:
  - There are four unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

- Five records where launch sites begin with `CCA`

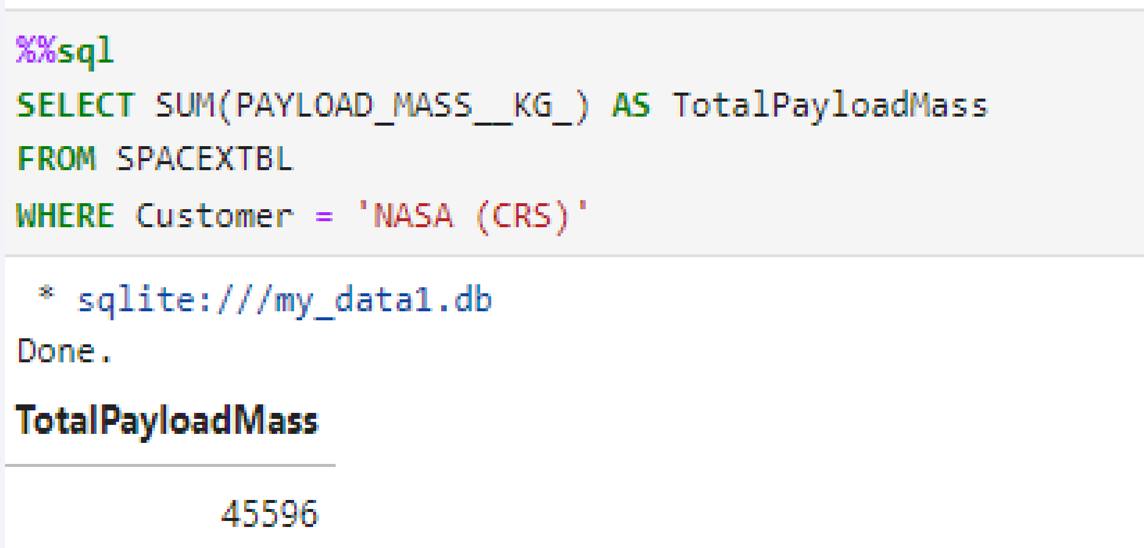


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Explanation:
  - There are more that 5 recorts where launch sites begin with ‘CCA’

Total Payload Mass

- Total payload carried by boosters from NASA:



```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayloadMass
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)'
```

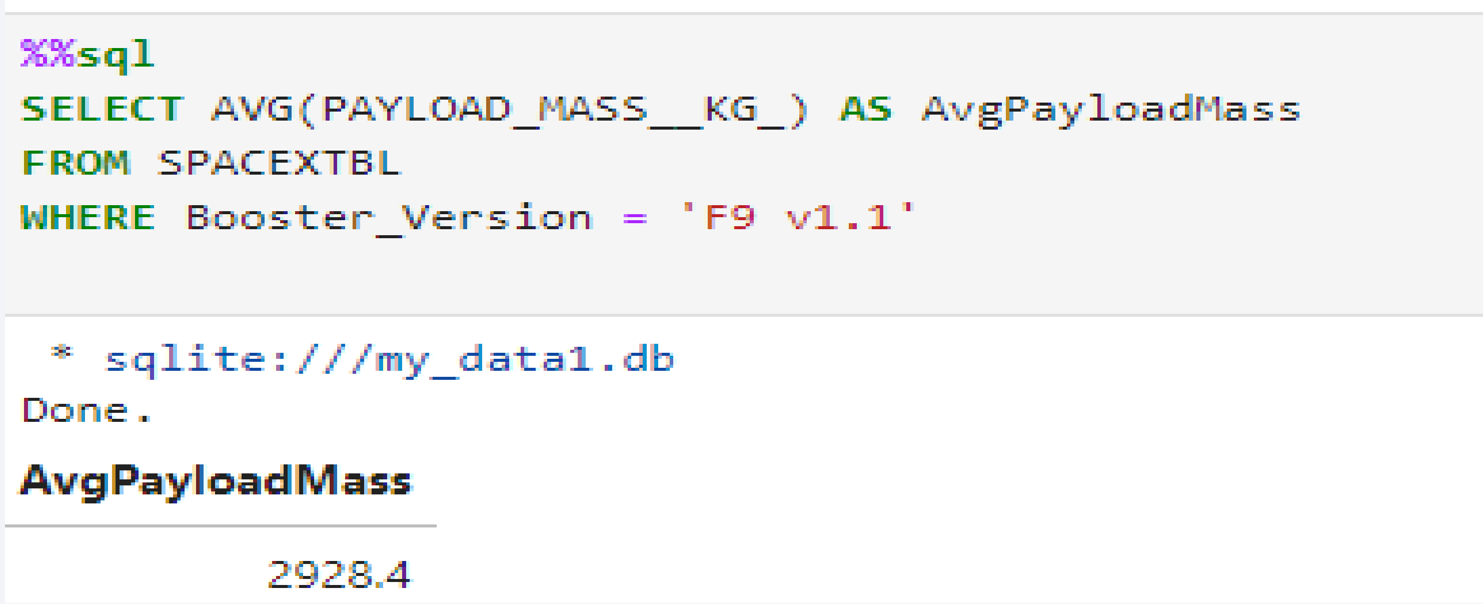
```
 * sqlite:///my_data1.db
Done.
```

TotalPayloadMass
45596

- Explanation:
  - The total payload carried by boosters from NASA is 45596.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1



```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AvgPayloadMass
FROM SPACEXTBL
WHERE Booster_Version = 'F9 v1.1'
```

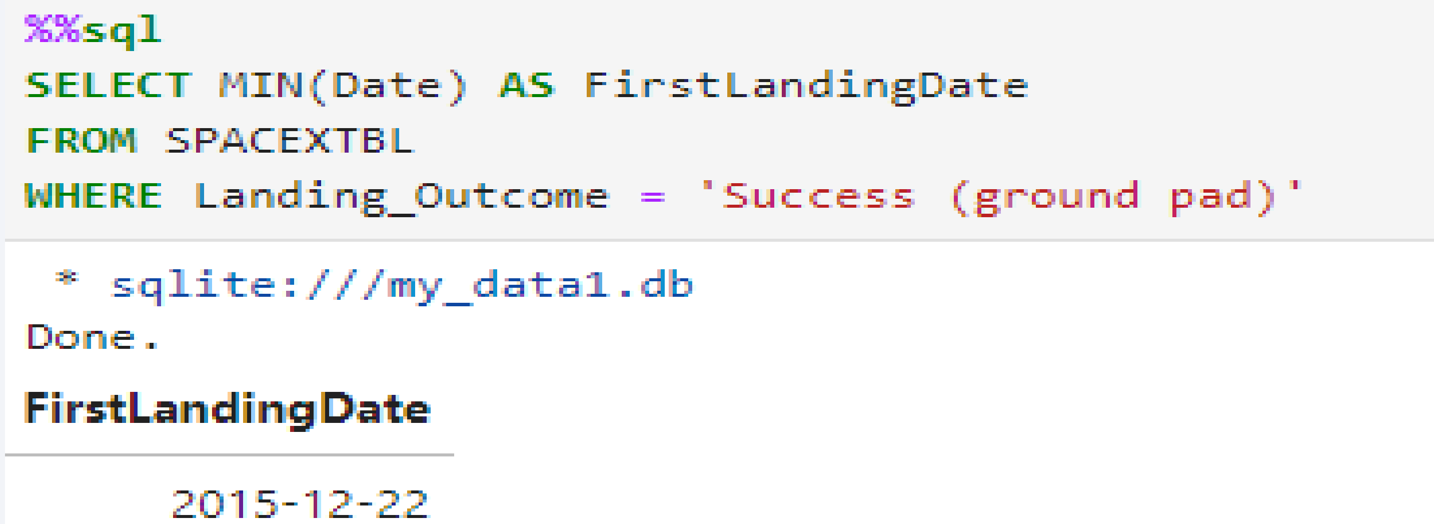
```
 * sqlite:///my_data1.db
Done.
```

AvgPayloadMass
2928.4

- Explanation:
  - The average payload mass carries by booster version F9 v1.1 is 2928.4.

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad



```
%%sql
SELECT MIN(Date) AS FirstLandingDate
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)'
```

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

FirstLandingDate
2015-12-22

- Explanation:
    - The date of first successful landing outcome in ground pad is 2015-12-22.

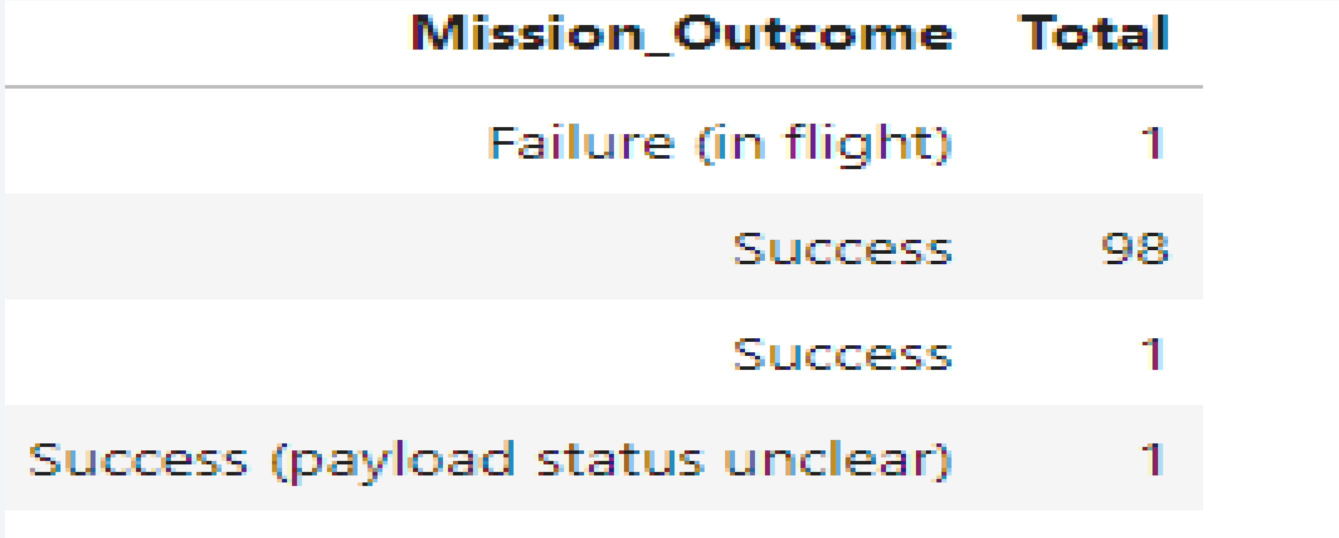
Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Explanation:
  - There are four boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

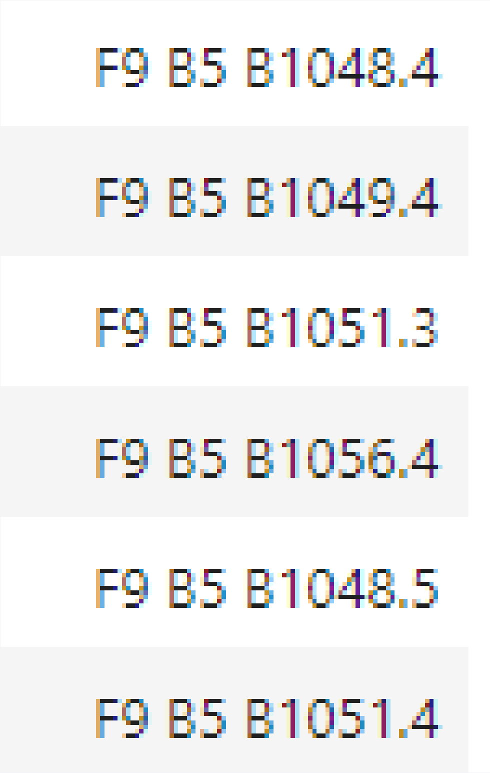


Mission_Outcome	Total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Explanation:
  - There are 4 distinct categories for Mission_Outcome.

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass



F9 B5 B1048.4

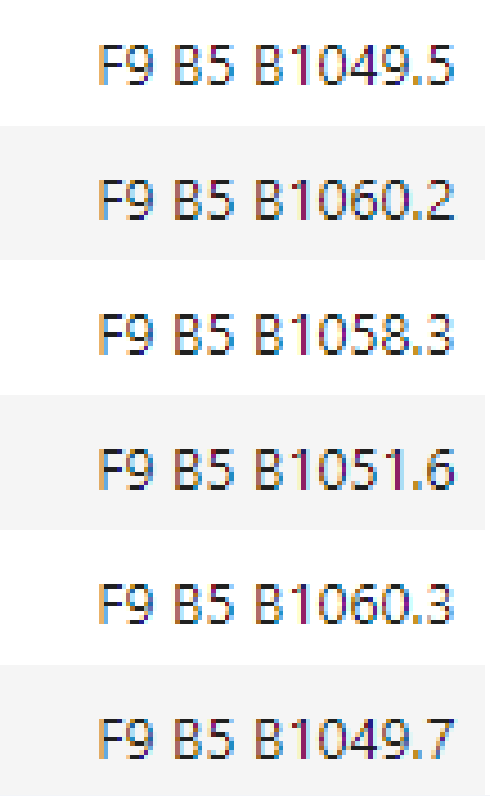
F9 B5 B1049.4

F9 B5 B1051.3

F9 B5 B1056.4

F9 B5 B1048.5

F9 B5 B1051.4



F9 B5 B1049.5

F9 B5 B1060.2

F9 B5 B1058.3

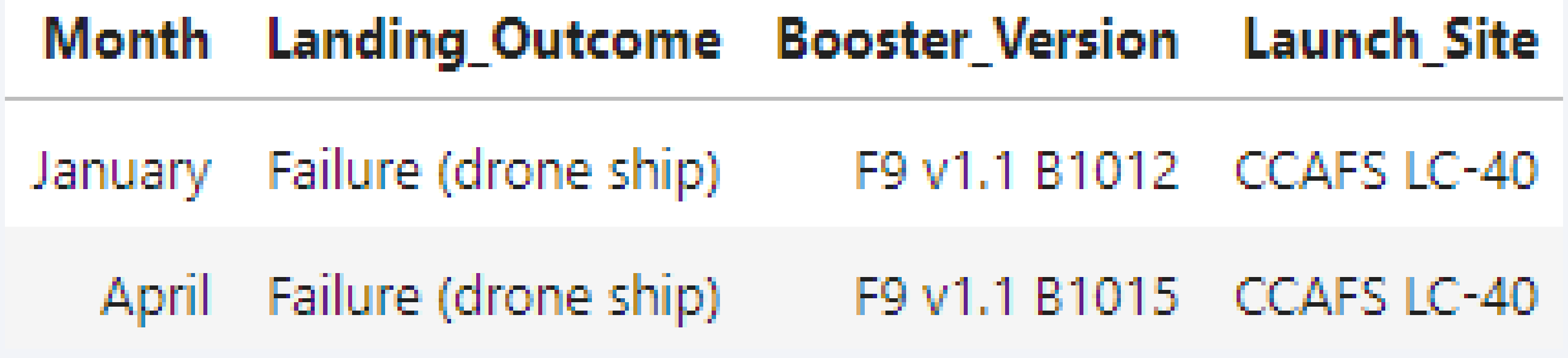
F9 B5 B1051.6

F9 B5 B1060.3

F9 B5 B1049.7

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

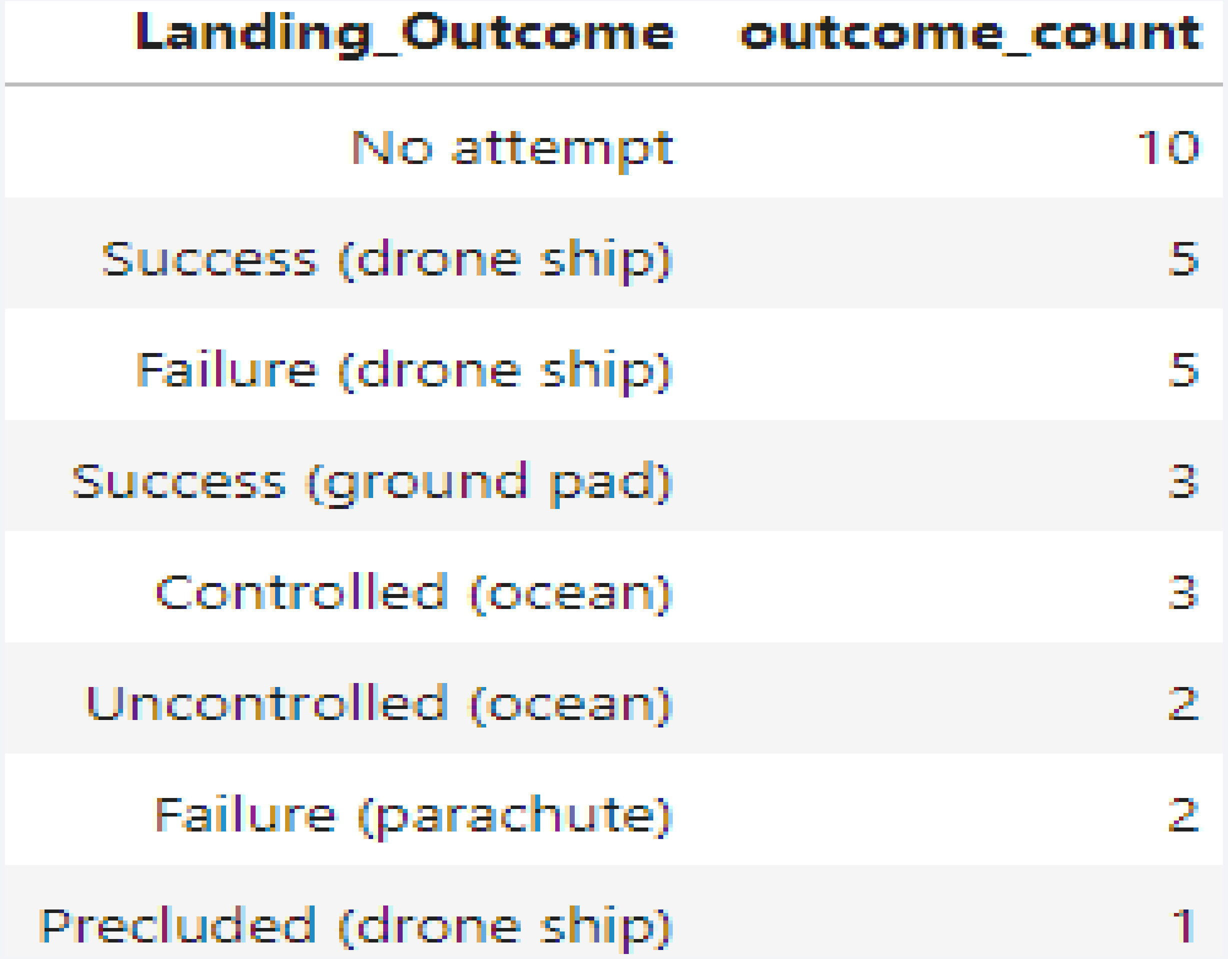


Month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Explanation:
  - There are two failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

- Rank 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

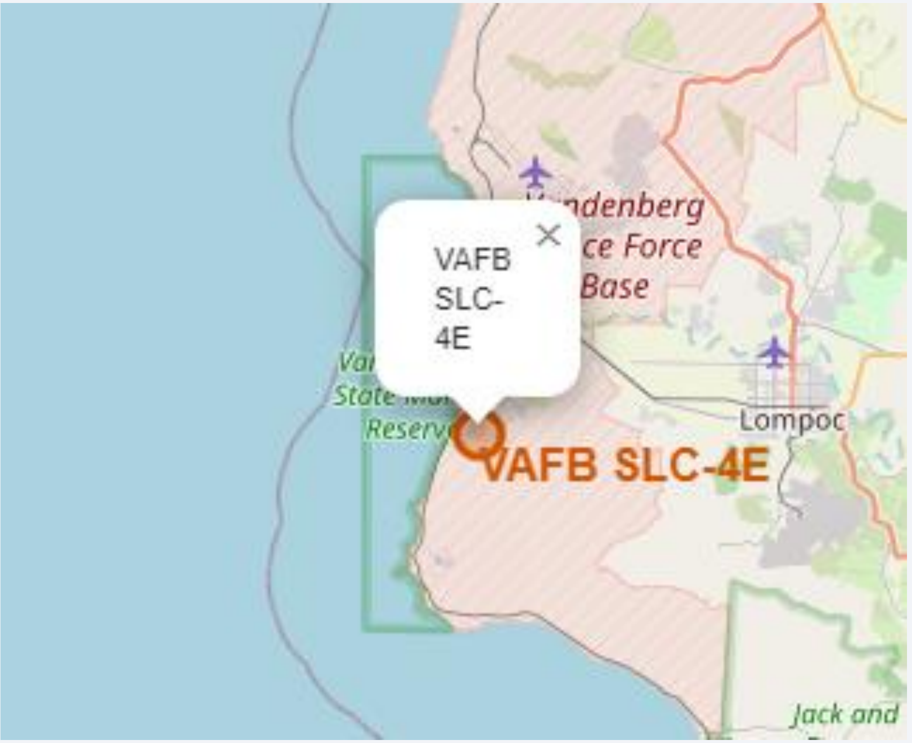


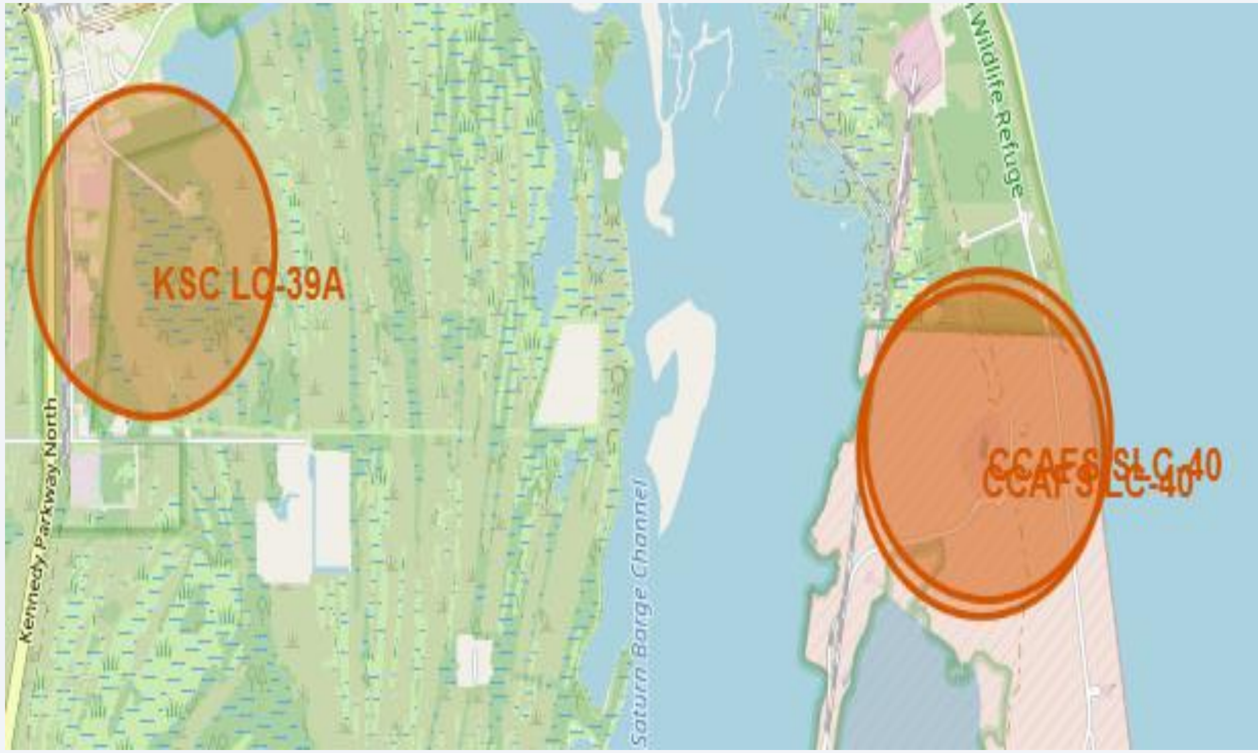
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

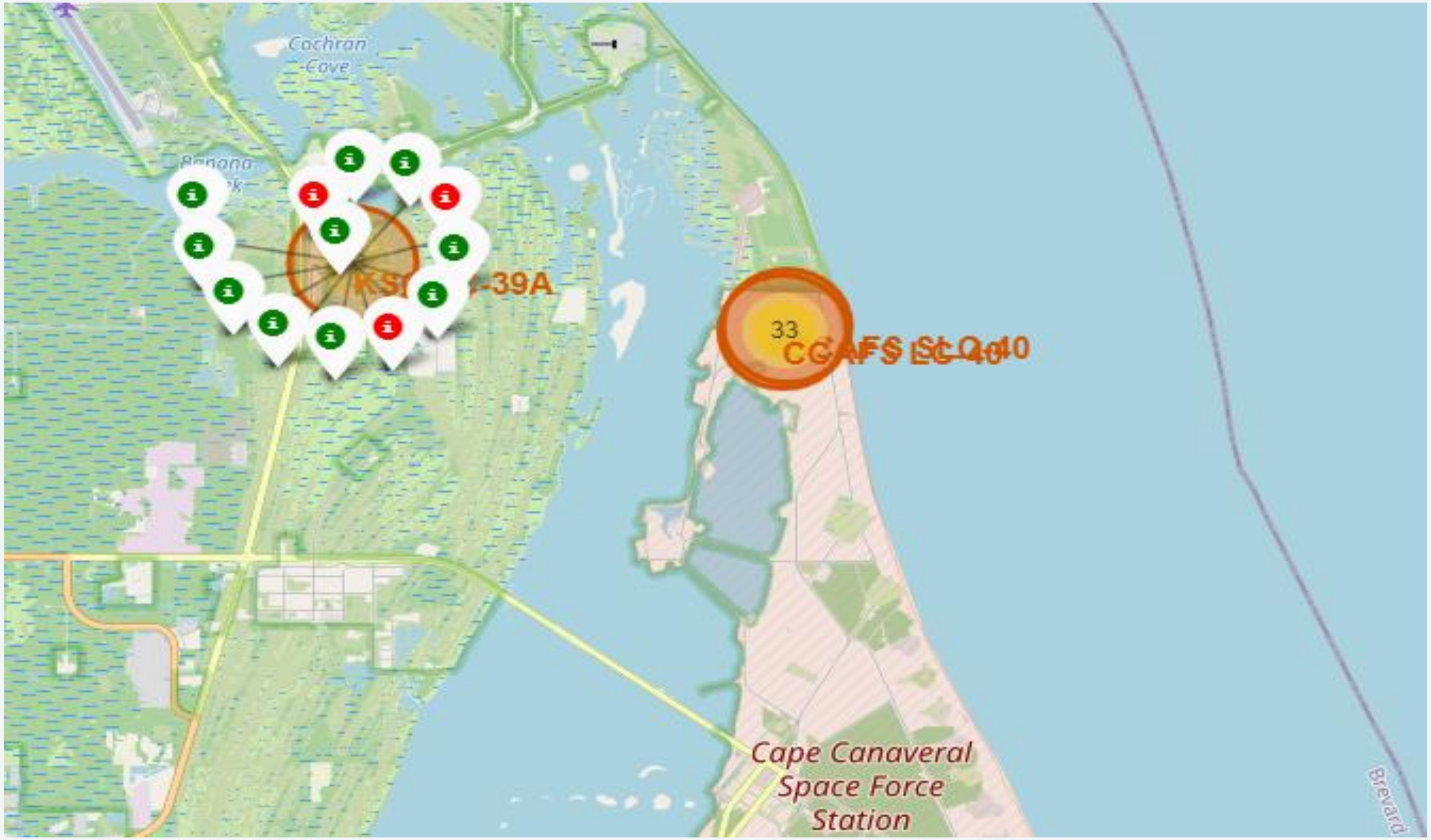
<Mark all launch sites on a map>





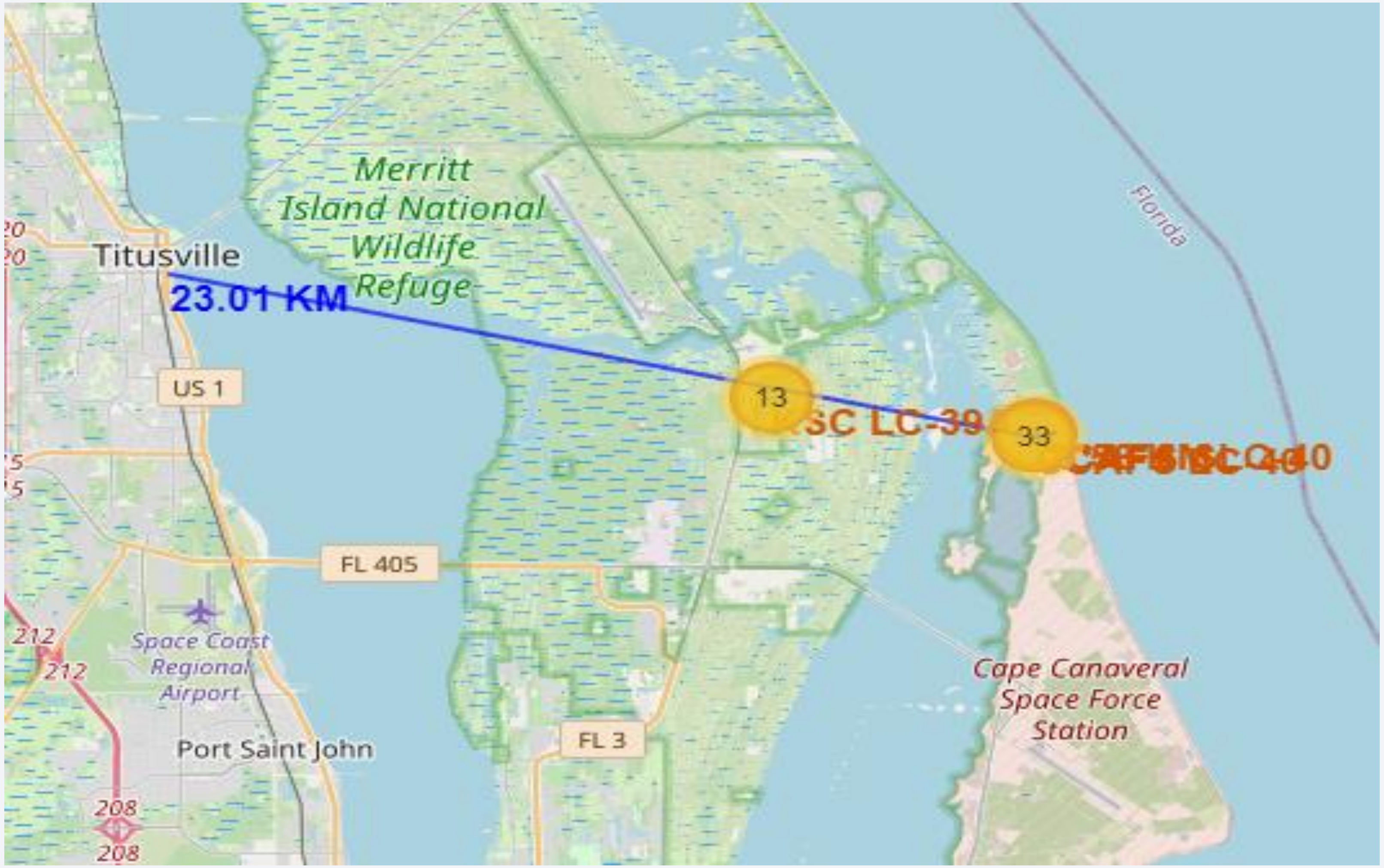
All four markers lie on or very near major bodies of water: CCAFS SLC-40 & KSC LC-39A on barrier islands beside the Atlantic; VAFB SLC-4E on the Pacific coast. This coastal siting is deliberate: rockets launch over open water and (for SpaceX) boosters can return to sea or shore without overflying populated areas.

< Mark the success/failed launches for each site on the map >



The map from left shows the launch sites have relatively high success rate.

< The distances between a launch site to Titusville, FL >


Merritt Island National Wildlife Refuge
Titusville
23.01 KM
US 1
13
SC LC-39
33
CCAFS SLC-40
Florida
FL 405
Space Coast Regional Airport
212
Port Saint John
208
FL 3
Cape Canaveral Space Force Station

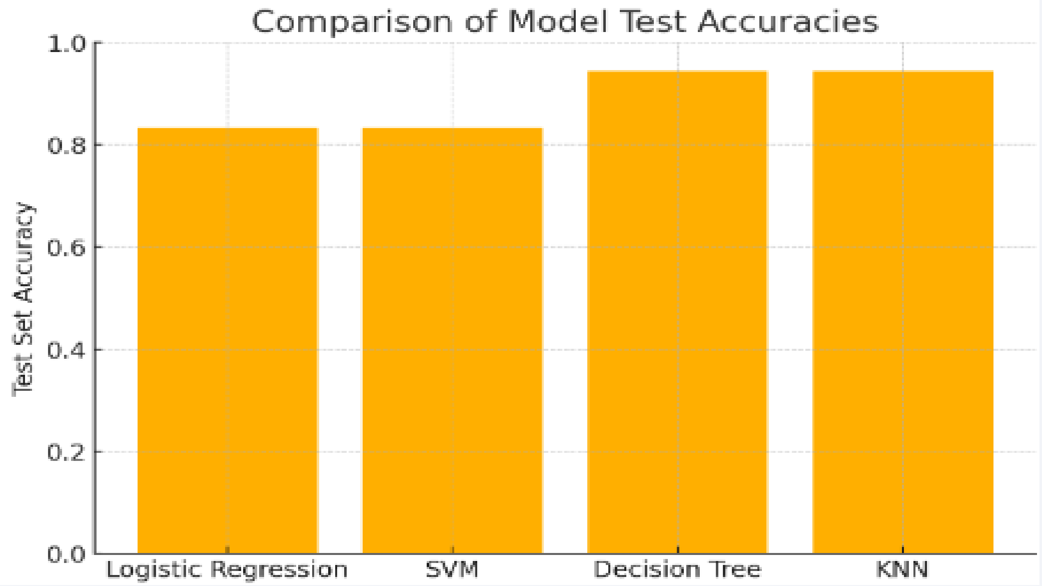
Map chart from left shows the distance between one of launch sites in close to city – Titusville, FL.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

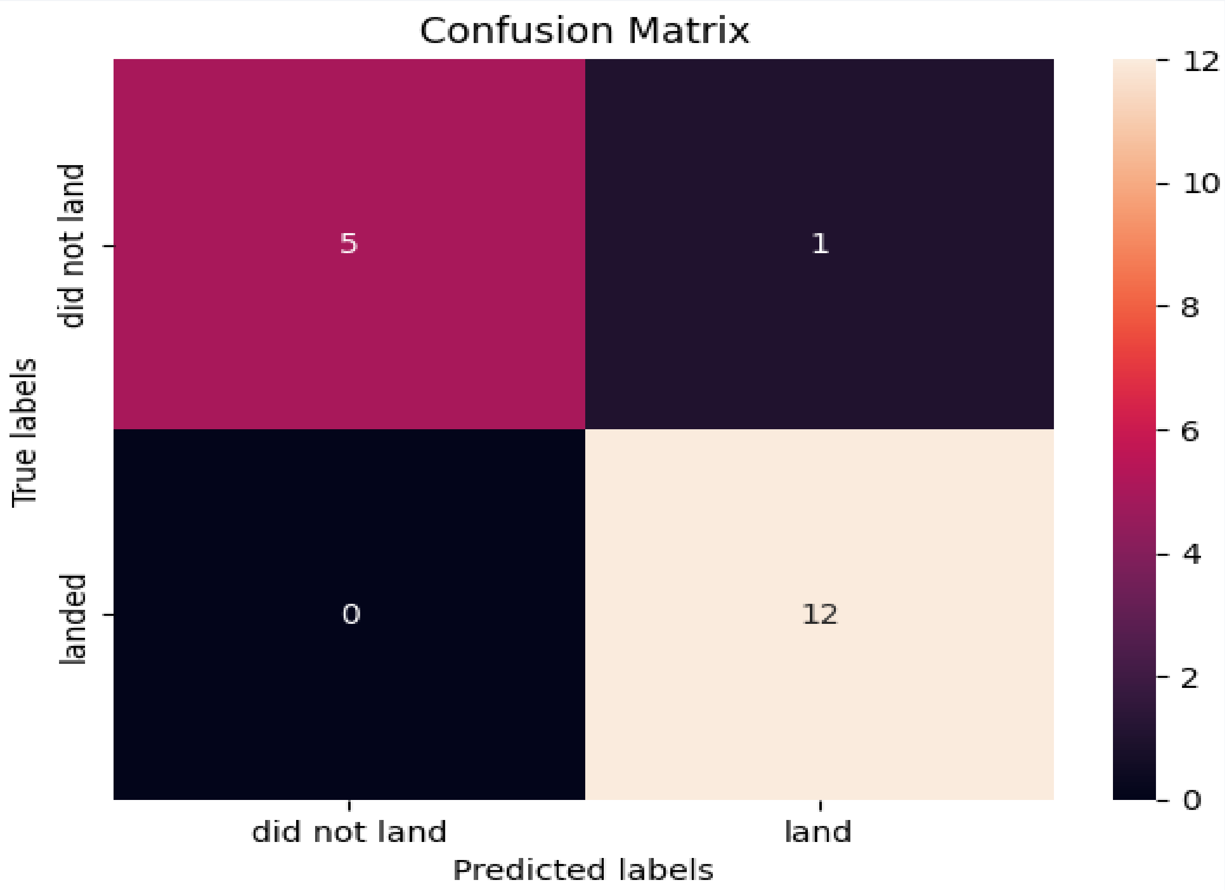
- Visualize the built model accuracy for all built classification models, in a bar chart



- Both Decision Tree and KNN model have the highest classification accuracy (~0.94)

Confusion Matrix

- Confusion matrix of the best performing model:



The Decision Tree model reduce the number of false alarms from 3 down to 1 (and still never miss a true landing), giving them the highest overall accuracy (≈94 %)

Conclusions

- **Logistic Regression**
  - Test accuracy ≈ 0.83, False positives = 3, false negatives = 0
- **SVM**
  - Test accuracy ≈ 0.83, False positives = 3, false negatives = 0
- **Decision Tree**
  - Test accuracy ≈ 0.94, False positives = 1, false negatives = 0
- **KNN**
  - Test accuracy ≈ 0.94, False positives = 1, false negatives = 0

Conclusion: Both the Decision Tree and KNN models reduce the number of false alarms from 3 down to 1 (and still never miss a true landing), giving them the highest overall accuracy (≈94 %). By that metric, they outperform Logistic Regression and SVM on this test set

Thank you!