IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vinh Tran
21/02/2025

Outline

- Executive Summary
- Introduction
- Methodology
- Conclusion
- Appendix

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

Executive Summary

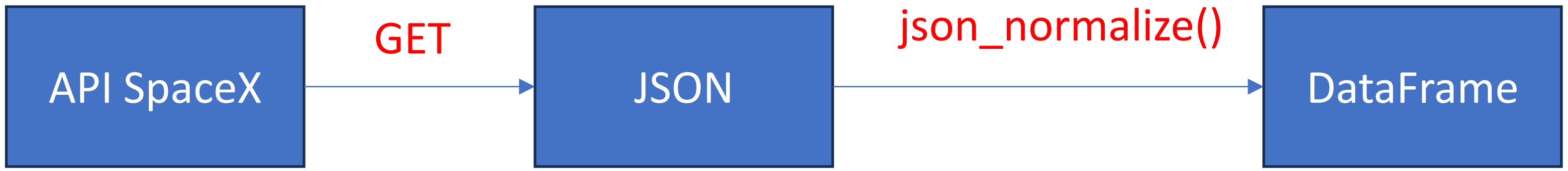
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Diagram of how to collect data

Data Collection

- Data acquisition



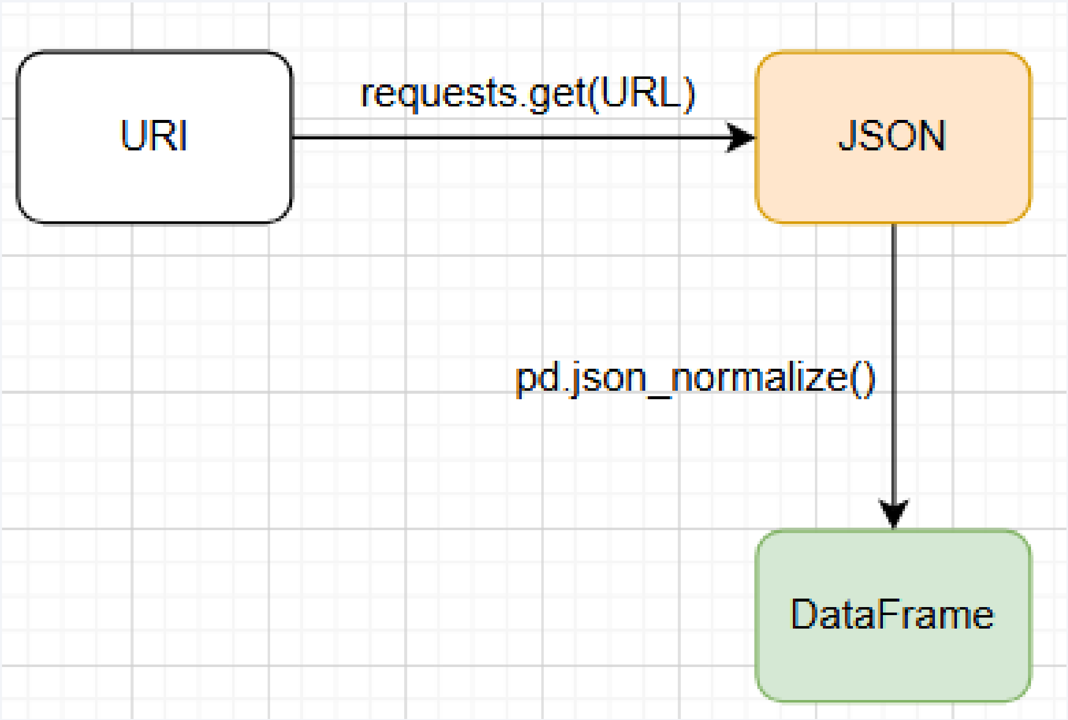
- Data wrangling:
  - Some data columns contain only IDs, use the API and these IDs to get more detailed information

Data Collection

- Sampling
  - Eliminate unnecessary data including Falcon 1 rocket launches
- Handling null values
  - PayloadMass: Calculate the average value of the PayloadMass column and replace NULL values with this average value
  - LandingPad: NULL value in this column is meaningful, handle it using one-hot encoding method

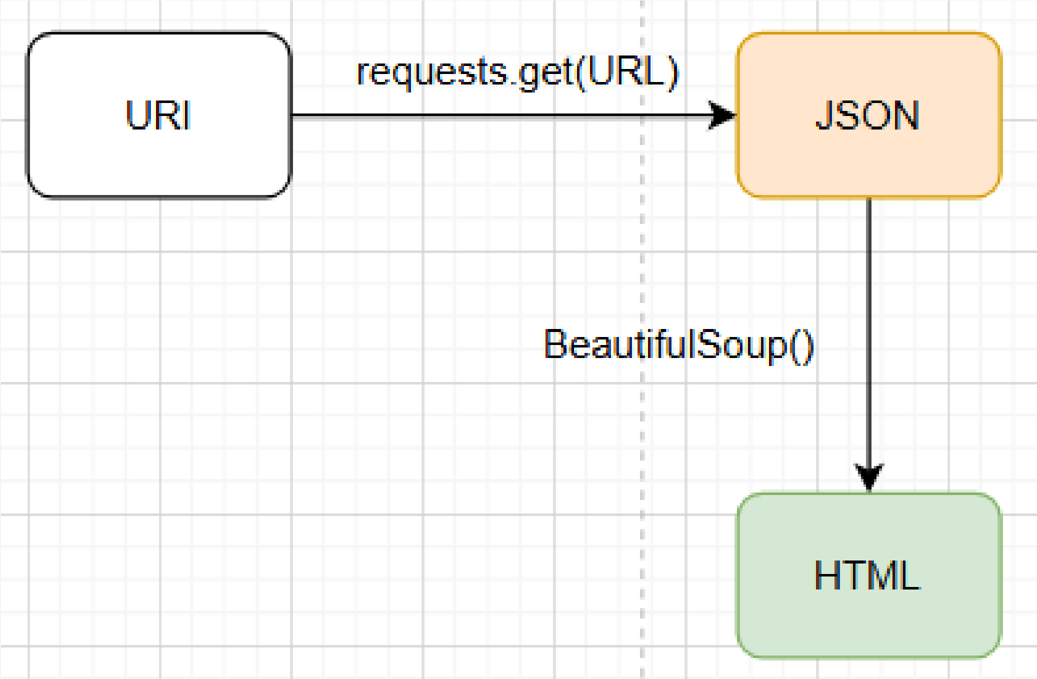
Data Collection – SpaceX API

- GitHub URL https://github.com/Kro-V/respo-V/blob/Child-Branch/jupyter-labs-spacex-data-collection-api.ipynb



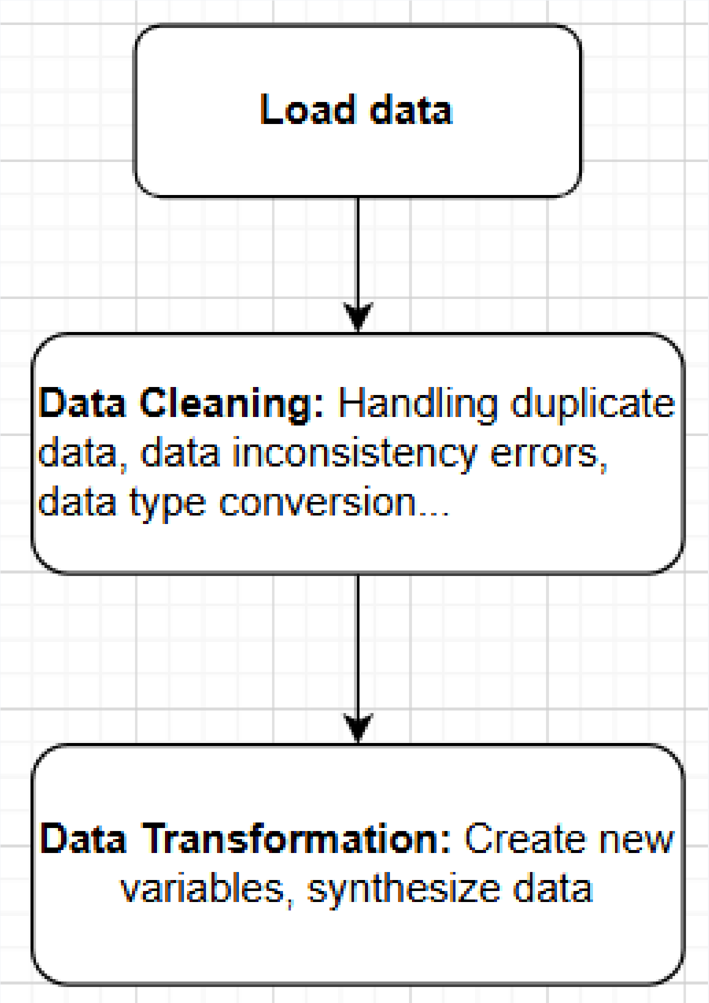
Data Collection - Scraping

- GitHub URL
  https://github.com/Kro-V/respo-V/blob/Child-Branch/jupyter-labs-webscraping.ipynb



Data Wrangling

- GitHub URL
  https://github.com/Kro-V/respo-V/blob/Child-Branch/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter point chart:
  - Visualize the relationship between Flight Number, Payload Mass and Launch Site
  - Visualize the relationship between FlightNumber, Payload Mass and Orbit type
- Bar chart:
  - Visualize the relationship between success rate of each orbit type
- Line chart:
  - Visualize the launch success yearly trend
- GitHub URL https://github.com/Kro-V/respo-V/blob/Child-Branch/edadataviz.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
  - SELECT DISTINCT "Launch_Site“ FROM SPACEXTABLE;
- Display 5 records where launch sites begin with the string 'CCA’
  - SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%'LIMIT 5;
- Display the total payload mass carried by boosters launched by NASA (CRS)
  - SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)'
- Display average payload mass carried by booster version F9 v1.1
  - SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1'

EDA with SQL

- List the date when the first succesful landing outcome in ground pad was acheived
  - SELECT "Date“ FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)’ ORDER BY "Date" ASCLIMIT 1;
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - SELECT "Booster_Version“ FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND 'PAYLOAD_MASS__KG_' > 4000 AND 'PAYLOAD_MASS__KG_' > 6000;
- List the total number of successful and failure mission outcomes
  - SELECT "Mission_Outcome", COUNT(*) AS TotalCount FROM SPACEXTABLE GROUP BY "Mission_Outcome";

EDA with SQL

- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - SELECT "Booster_Version"FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
- 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
  - SELECT substr("Date", 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site“ FROM SPACEXTABLE WHERE substr("Date", 1, 4) = '2015' AND "Landing_Outcome" LIKE 'Failure (drone ship)';

EDA with SQL

- 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
  - SELECT "Landing_Outcome", COUNT(*) AS OutcomeCount FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome"ORDER BY OutcomeCount DESC;
- GitHub URL https://github.com/Kro-V/respo-V/blob/Child-Branch/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

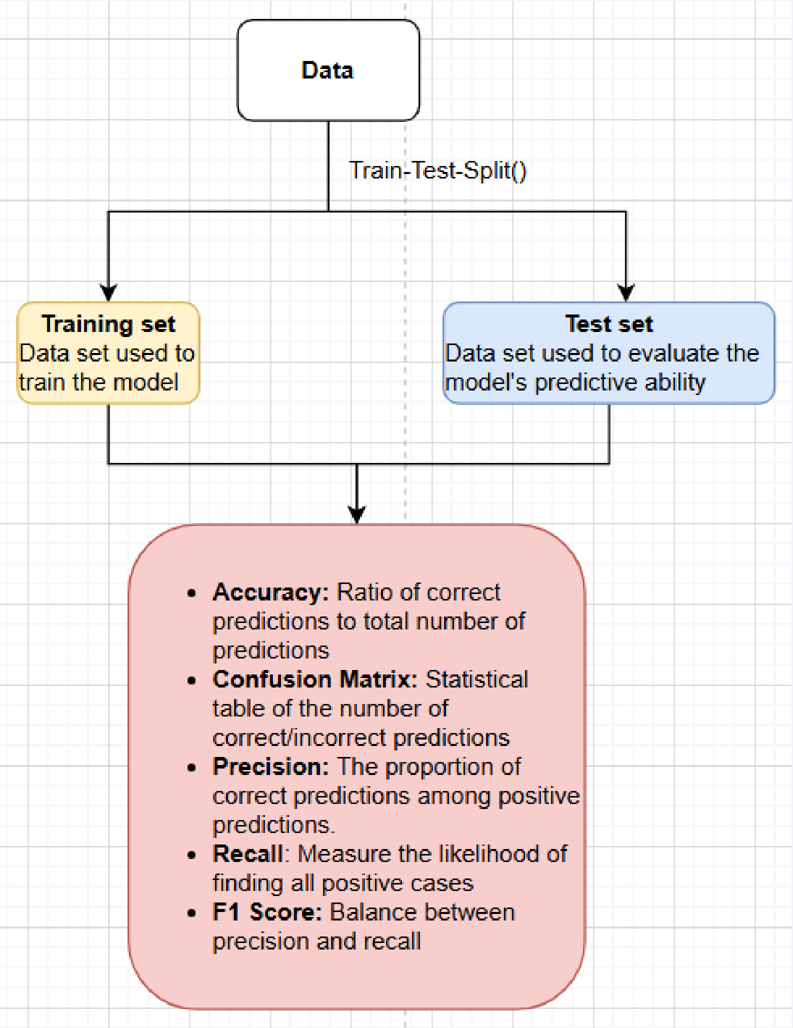
- Folium.Circle() is used to circle the pointsFolium.Marker() is used to mark each point
- Explain why you added those objects
- GitHub URL https://github.com/Kro-V/respo-V/blob/Child-Branch/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie chart
  - Select all launch locations or a specific launch location
  - Displays launch success rate based on selected area/all areas
- Scatter chart
  - Select all launch locations or a specific launch location
  - Shows the relationship between Payload Mass and launch success rate by specific region/all regions
- GitHub URL https://github.com/Kro-V/respo-V/blob/Child-Branch/spacex_dash_app.py

Predictive Analysis (Classification)

- GitHub URL https://github.com/Kro-V/respo-V/blob/Child-Branch/Evaluating%20Classification%20Models.ipynb

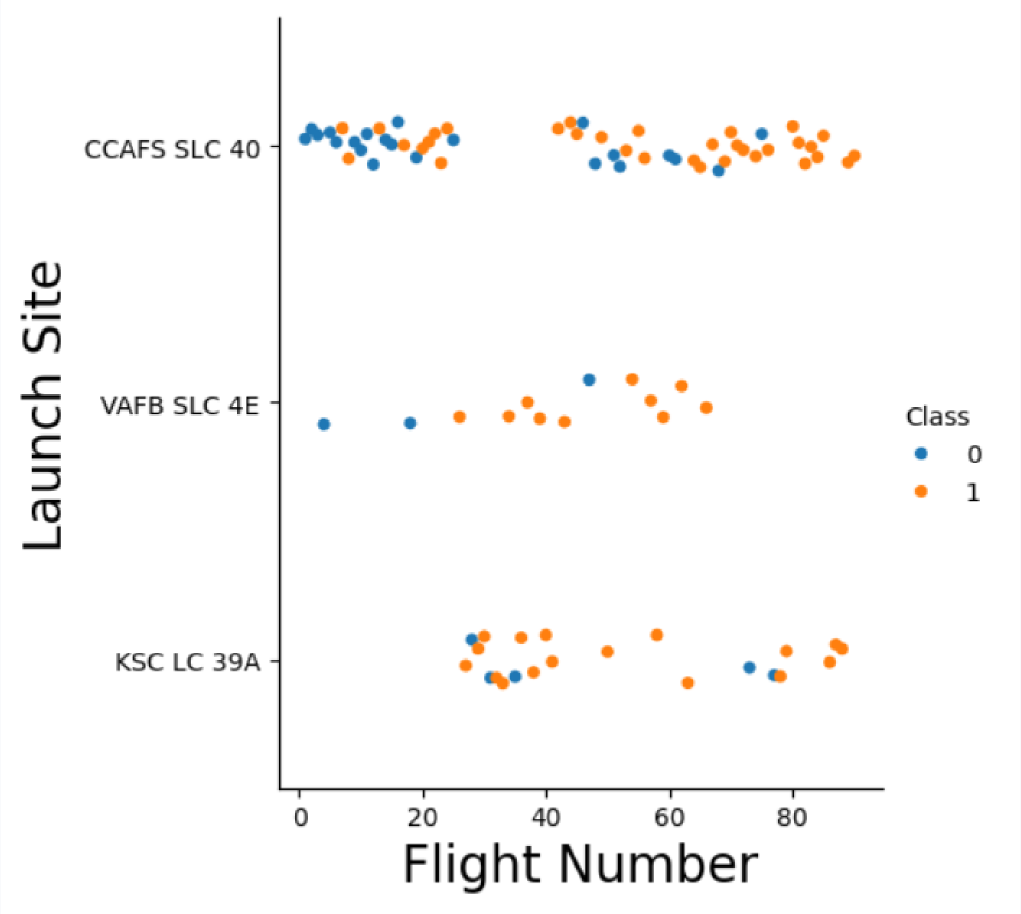


Section 2

Insights drawn from EDA

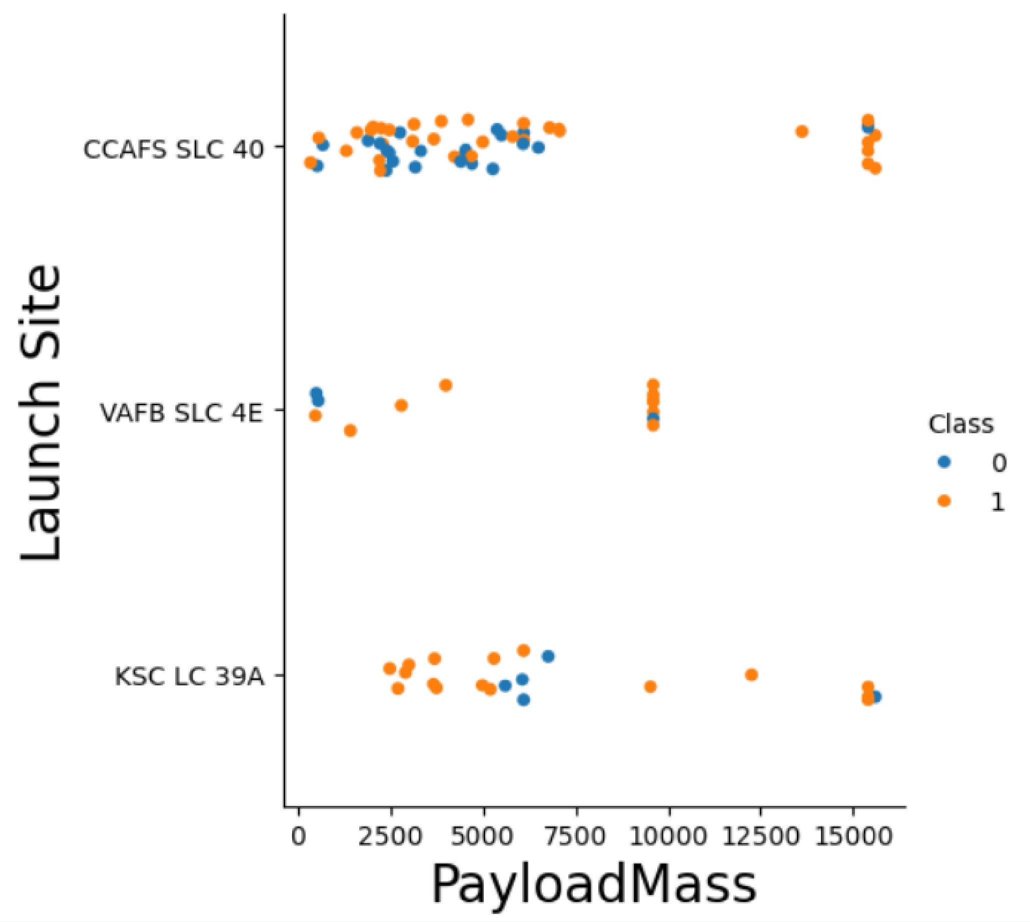
Flight Number vs. Launch Site

- CCAFS SLC 40 has the most number of flights, followed by KSC LC 39A and VAFB SLC 4E with quite similar number of flights.
- CCAFS SLC 40 has a high number of launch failures, while VAFB SLC 4E and KSC LC 39A appear to have a lower failure rate.



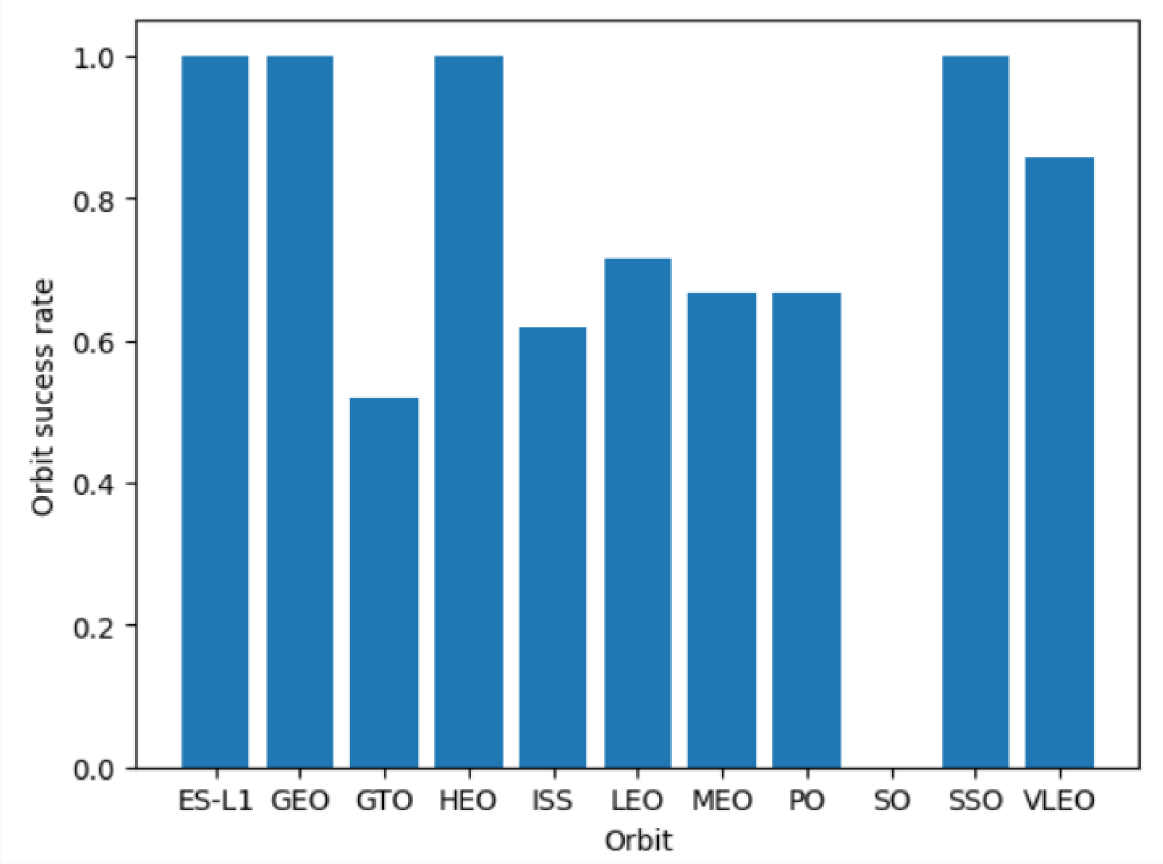
Payload vs. Launch Site

- Payload Mass is mainly in the range of 2500 kg to 7000 kg



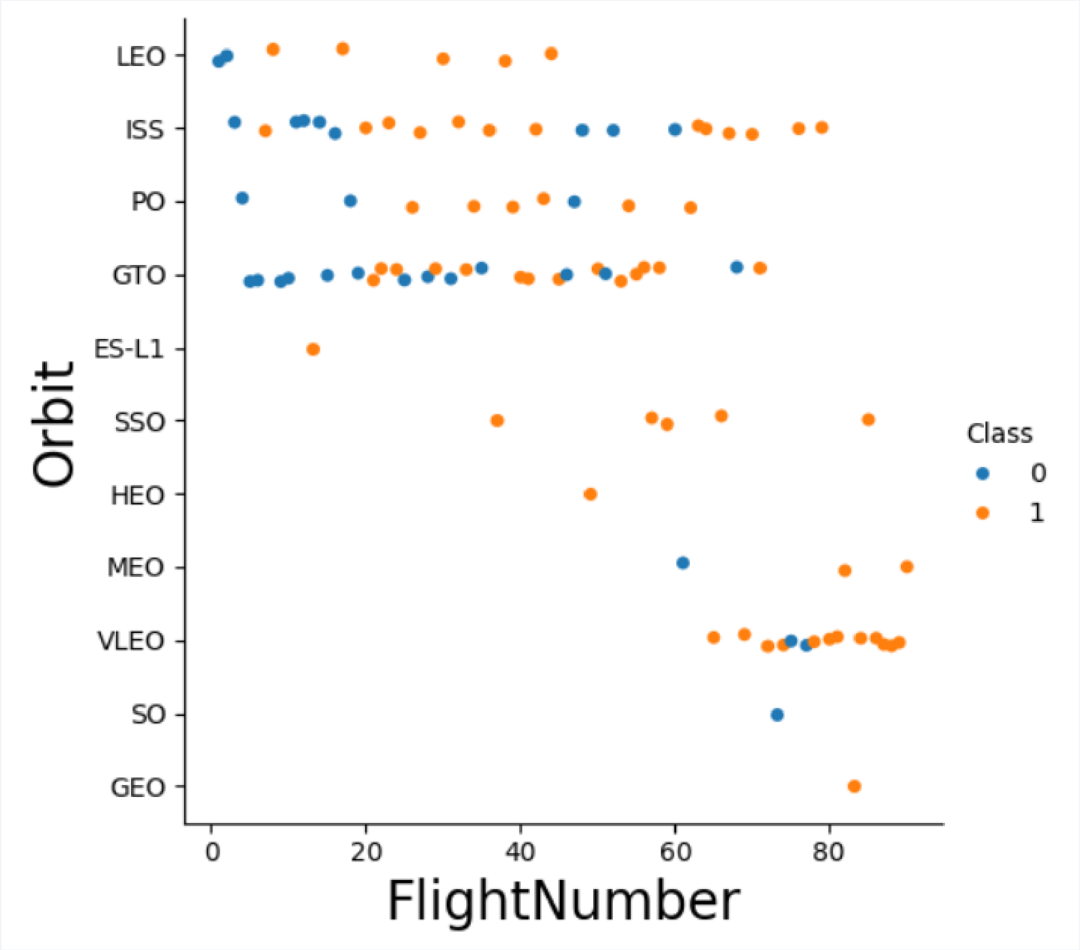
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO have 100% success rate
- VLEO has a success rate of around 90%.
- GTO has the lowest success rate of around 50%
- The remaining orbits have success rates between 60% and 70%.



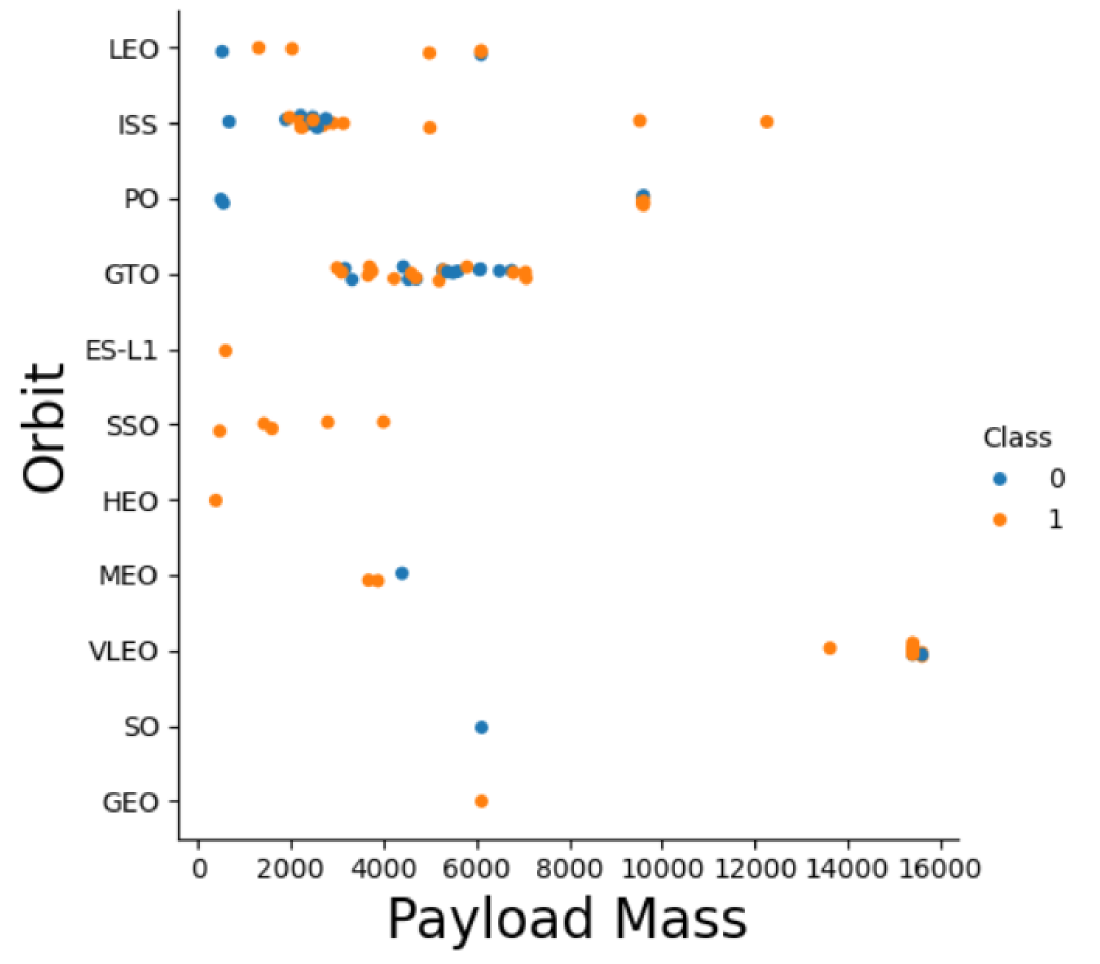
Flight Number vs. Orbit Type

- GTO, ISS, VLEO have the highest number of flights
- Next are LEO and PO
- The remaining orbits have the least number of flights



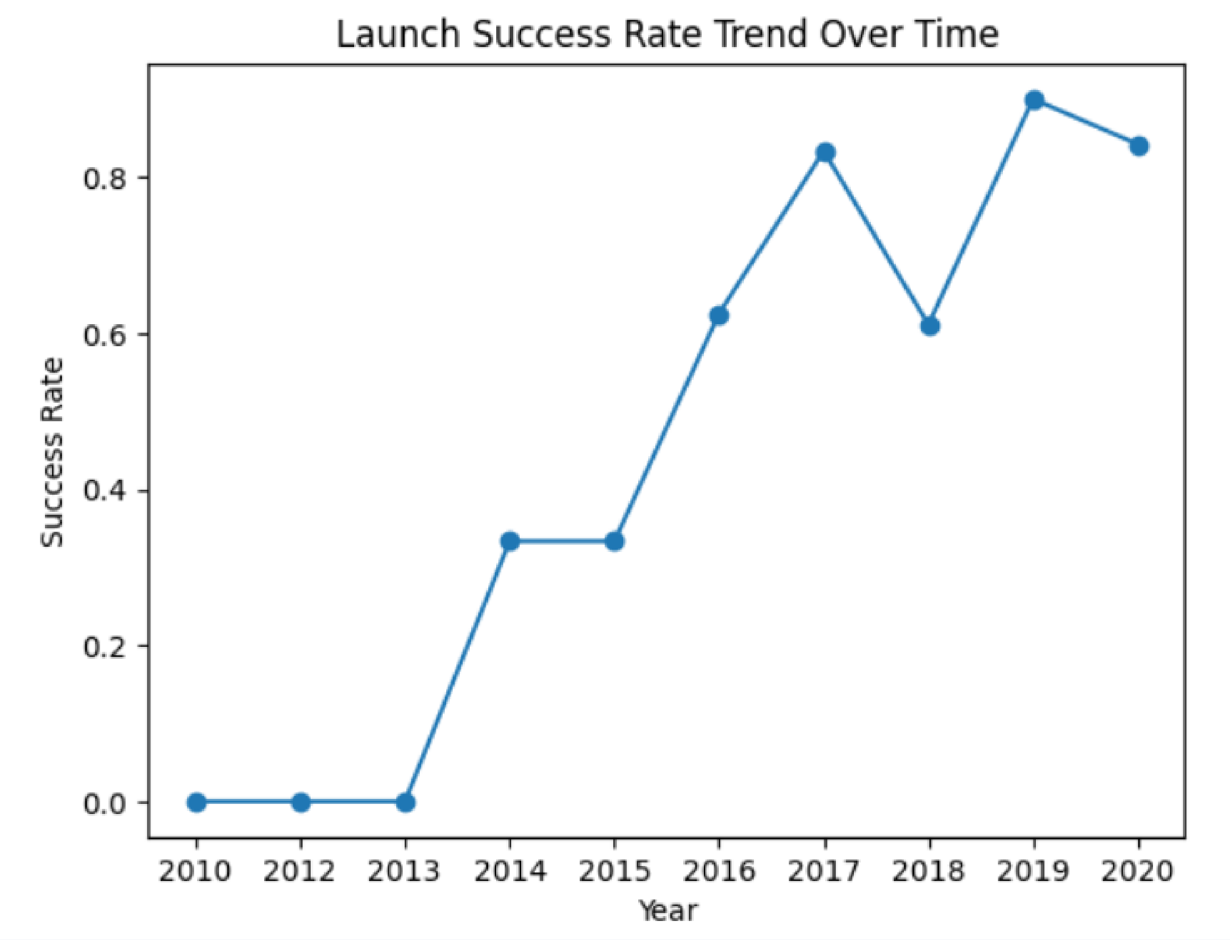
Payload vs. Orbit Type

- The ISS has flights with payloads in the range of 2000 kg to 3000 kg
- GTO has flights with payloads ranging from 2500 kg to 7000 kg



Launch Success Yearly Trend

- 2010 - 2013 has a low success rate
- 2013 - 2014 success rate increased to about 40%
- 2015 - 2017 success rate increased to about 80%
- 2019 has the highest success rate of about 90%



All Launch Site Names

- SELECT DISTINCT "Launch_Site“ FROM SPACEXTABLE;

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

Launch Site Names Begin with 'CCA'

- SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%’ LIMIT 5;

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

Total Payload Mass

- SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)'

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1'

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- SELECT "Date“ FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)’ ORDER BY "Date" ASCLIMIT 1;

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- SELECT "Booster_Version“ FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND 'PAYLOAD_MASS__KG_' > 4000 AND 'PAYLOAD_MASS__KG_' > 6000;

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- SELECT "Mission_Outcome", COUNT(*) AS TotalCount FROM SPACEXTABLE GROUP BY "Mission_Outcome";

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

Boosters Carried Maximum Payload

- SELECT "Booster_Version“ FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- SELECT substr("Date", 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site“ FROM SPACEXTABLE WHERE substr("Date", 1, 4) = '2015' AND "Landing_Outcome" LIKE 'Failure (drone ship)';

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

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

- SELECT "Landing_Outcome", COUNT(*) AS OutcomeCount FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20’ GROUP BY "Landing_Outcome“ ORDER BY OutcomeCount DESC;

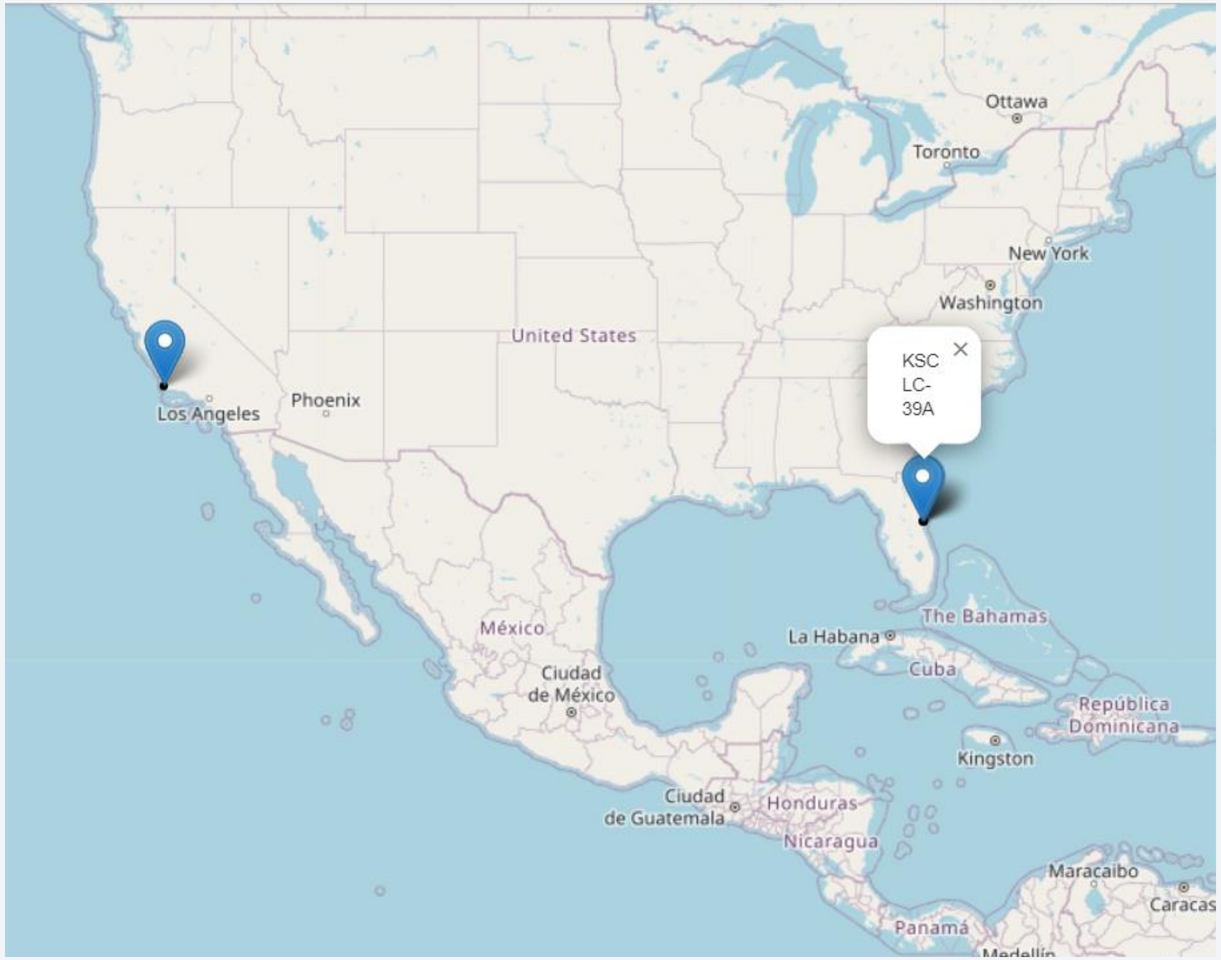
Landing_Outcome	OutcomeCount
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

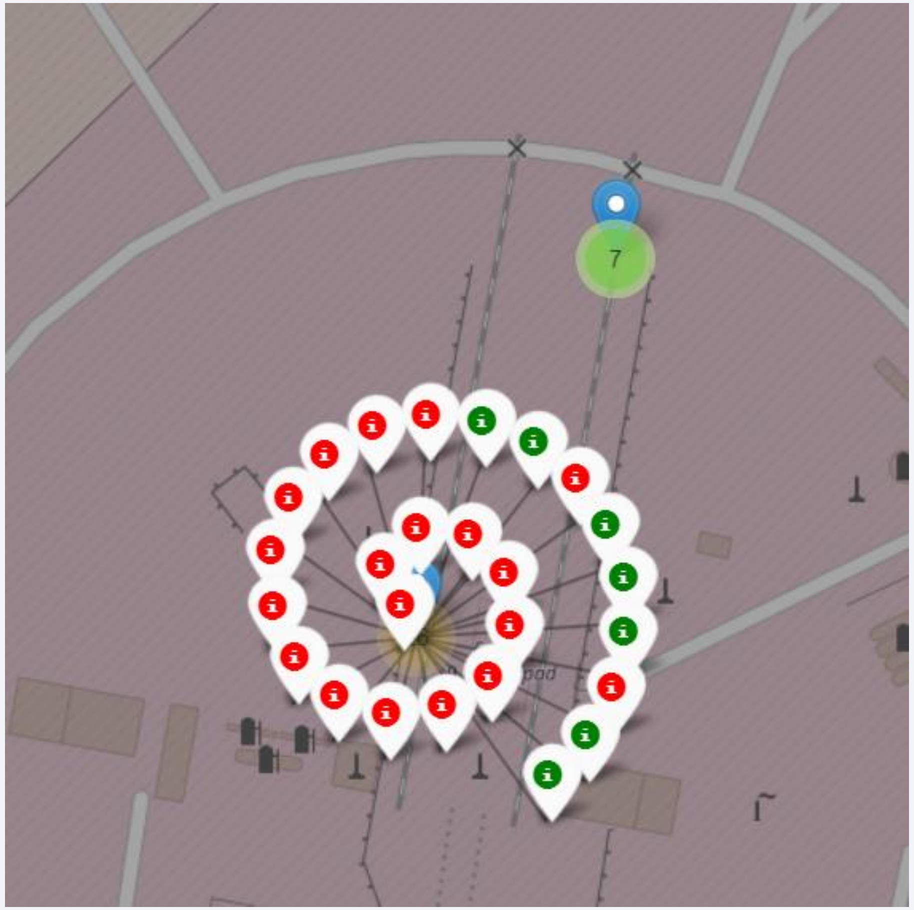
Mark launch locations

- Launch locations are marked on the map



Displays launch success/failure results for each launch location

- Launch sites are divided into clusters, each cluster consisting of launch sites
- Each launch location shows information about the number of successful/failed launches, blue is successful launches, red is failed launches

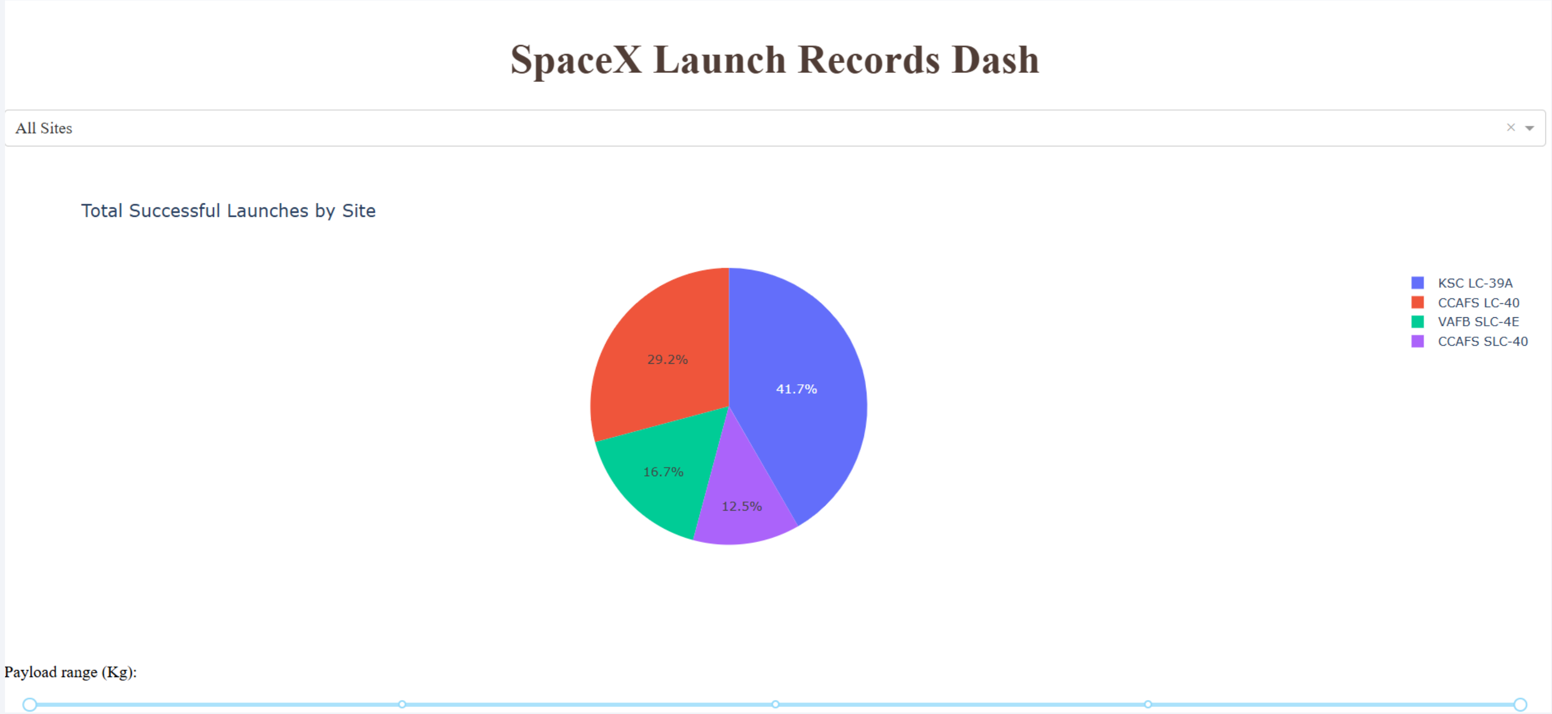


Section 4

Build a Dashboard with Plotly Dash

Success rates for all launch sites

- KSC LC 39A has the highest successful launch rate of 41.7%



Section 5

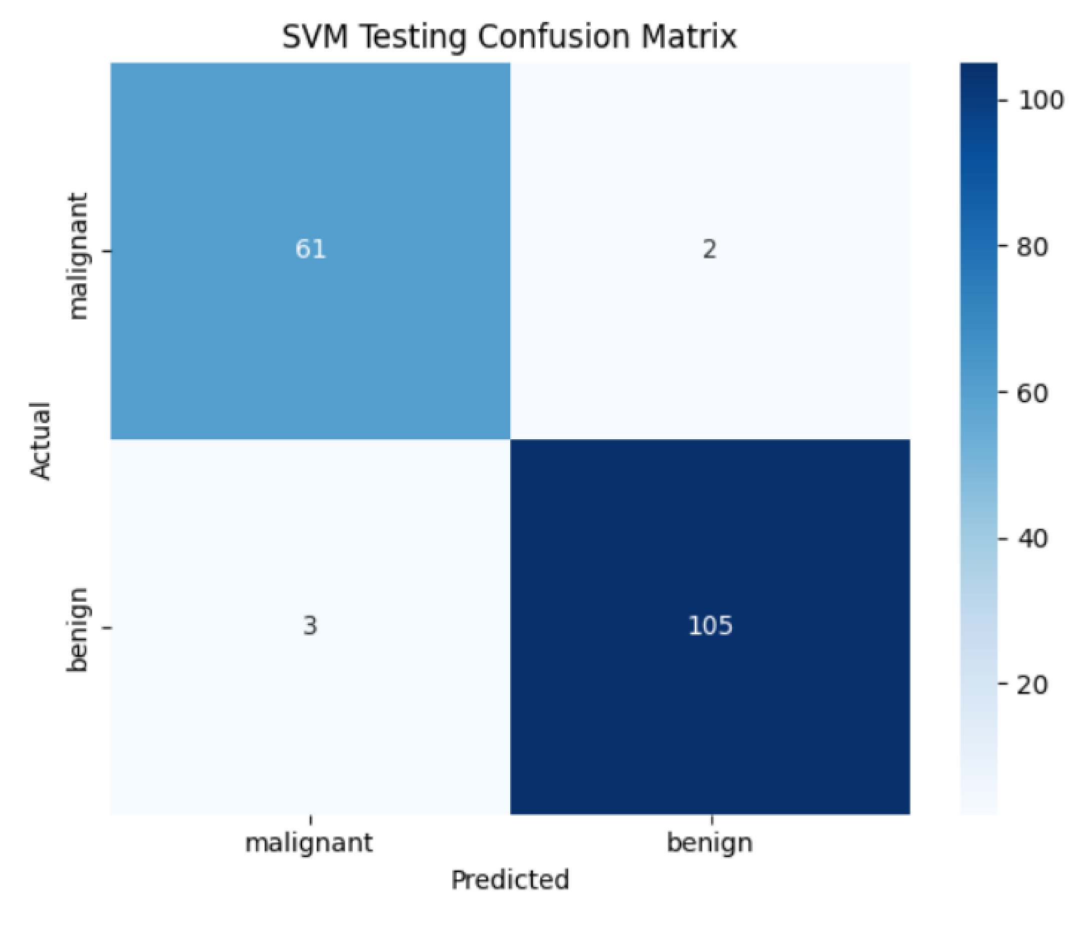
Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- Find which model has the highest classification accuracy

Confusion Matrix

- SVM models have better prediction ability with Accuracy = 0.971
- There were 2 tumors that were actually benign but the model incorrectly predicted they were malignant
- There were 3 tumors that were actually malignant but the model incorrectly predicted they were benign



Thank you!