IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ashkan Zand
14 December 2024

Outline

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

Executive Summary

- Summary of methodologies
  - SpaceX Dada Collection using SpaceX API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization and SQL
  - Space-X Launch Sites Analysis with Folium-Interactive Visual Analytics and Ploty Dash
  - Machine Learning Landing Prediction
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive Visual Analytics and Dashboards
  - Predictive analysis results

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches 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

  We will predict if the Falcon 9 first stage will land successfully using the data from Falcon 9 rocket launches advertised on its website

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

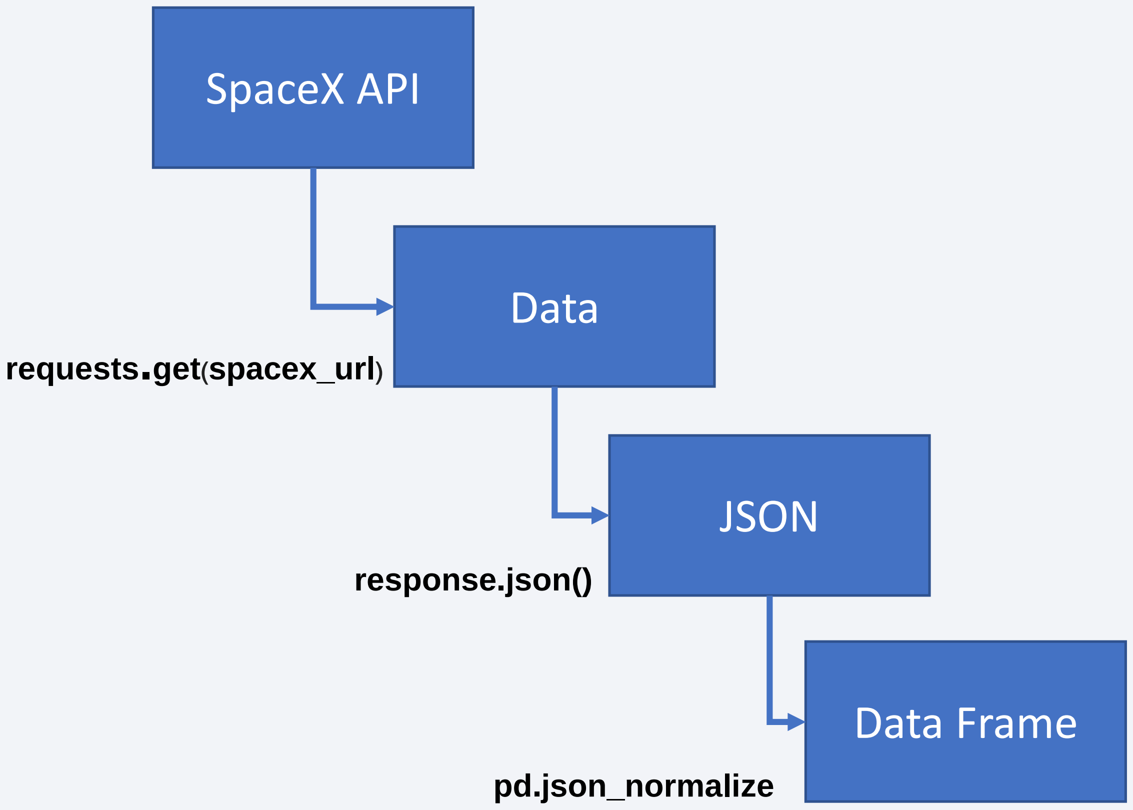
Data Collection

- Data is collected using SpaceX API (a RESTful API)
- Also, Web Scraping is performed to collect Falcon 9 historical records from the Wikipedia page titled List of falcon 9 and Falcon Heavy launches.

Data Collection – SpaceX API

- Collecting Rocket launch data from SpaceX API, by sending a get request.
- Decoding the response content as a JSON
- Turing JSON results into a Pandas Data Frame.

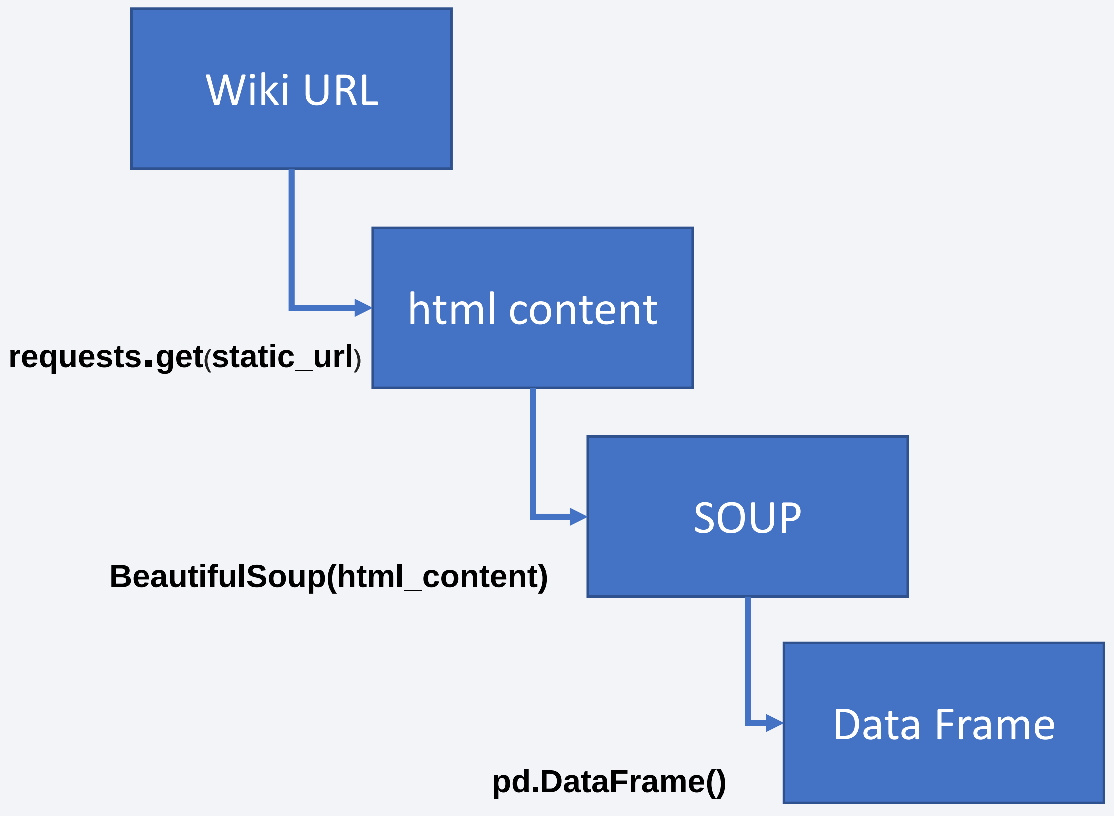
- HERE the GitHub URL of the completed SpaceX API calls notebook: https://github.com/AshkanZand/Projects/blob/a8434070144e71307f1cc15d38022552e389a4a1/SpaceX%20Falcon%209/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection – Web Scraping

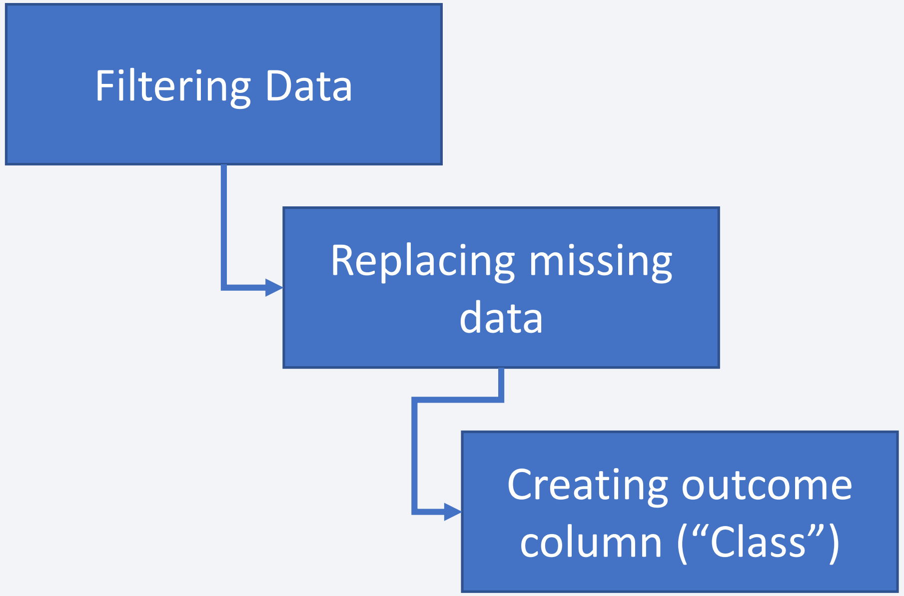
- Performing an HTTP GET method to request the Falcon9 Launch HTML page
- Creating a BeautifulSoup object from the HTML response
- Creating a data frame by parsing the launch HTML tables

- HERE is the GitHub URL of the completed Web Scraping notebook: https://github.com/AshkanZand/Projects/blob/a8434070144e71307f1cc15d38022552e389a4a1/SpaceX%20Falcon%209/jupyter-labs-webscraping.ipynb



Data Wrangling

- Filtering created Data Frame to only keep Falcon 9 lunches
- Replacing missing data values in the LandingPad and PayloadMass columns by mean value of the column
- Creating a landing outcome label (0 or 1) from Outcome column
- [HERE](https://github.com/AshkanZand/Projects/blob/a8434070144e71307f1cc15d38022552e389a4a1/SpaceX%20Falcon%209/labs-jupyter-spacex-Data%20wrangling.ipynb) is the GitHub URL of the completed Data Wrangling notebook: https://github.com/AshkanZand/Projects/blob/a8434070144e71307f1cc15d38022552e389a4a1/SpaceX%20Falcon%209/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Using scatter plots to visually check if there are any relationship between
  - Flight Number and Payload Mass
  - Flight Number and Launch Site
  - Payload Mass and Launch Site
  - Flight Number and Orbit type
  - Payload Mass and Orbit type
- Using Bar chart to Visualize the relationship between success rate of each orbit type
- Using Line plot to Visualize the launch success yearly trend
- HERE is the GitHub URL of the completed EDA with Data Visualization notebook: https://github.com/AshkanZand/Projects/blob/a8434070144e71307f1cc15d38022552e389a4a1/SpaceX%20Falcon%209/edadataviz.ipynb

EDA with SQL

EDA were performed using following SQL queries:

- Display the names of the unique launch sites in the space mission

```
%sql select  distinct("Launch_Site") from SPACEXTABLE
```

- Display 5 records where launch sites begin with the string 'CCA’

```
%sql select * from SPACEXTABLE where "Launch_Site" like "CCA%" limit 5
```

- Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum("PAYLOAD_MASS__KG_") from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

```
%sql select avg("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" = "F9 v1.1"
```

EDA with SQL

EDA were performed using following SQL queries:

- List the date when the first succesful landing outcome in ground pad was acheived.

```
%sql select min("Date") from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)"
```

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql  select "Booster_Version" from SPACEXTABLE where ("Landing_Outcome" = "Success (drone ship)" and 4000<"PAYLOAD_MASS__KG_"<6000)
```

- List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
```

EDA with SQL

EDA were performed using following SQL queries:

- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
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.

```
%%sql
select substr(Date, 6,2),"Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where (
substr(Date,0,5)='2015' and Landing_Outcome like "Failure (drone ship)")
```

EDA with SQL

EDA were performed using following SQL queries:

- 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.

```
%%sql
select "Landing_Outcome", count("Landing_Outcome") from SPACEXTABLE where ("2010-06-04"<"Date"< "2017-03-20")
group by "Landing_Outcome"
ORDER BY count("Landing_Outcome") desc
```

- HERE is the GitHub URL of the completed Web Scraping notebook: https://github.com/AshkanZand/IBM-Data-Science-Professional/blob/969feeda7844a92d3d68133dcbc0ab2d6b4881e1/SpaceX-Falcon-9/jupyter-labs-eda-sql-coursera_sqllite.ipynb

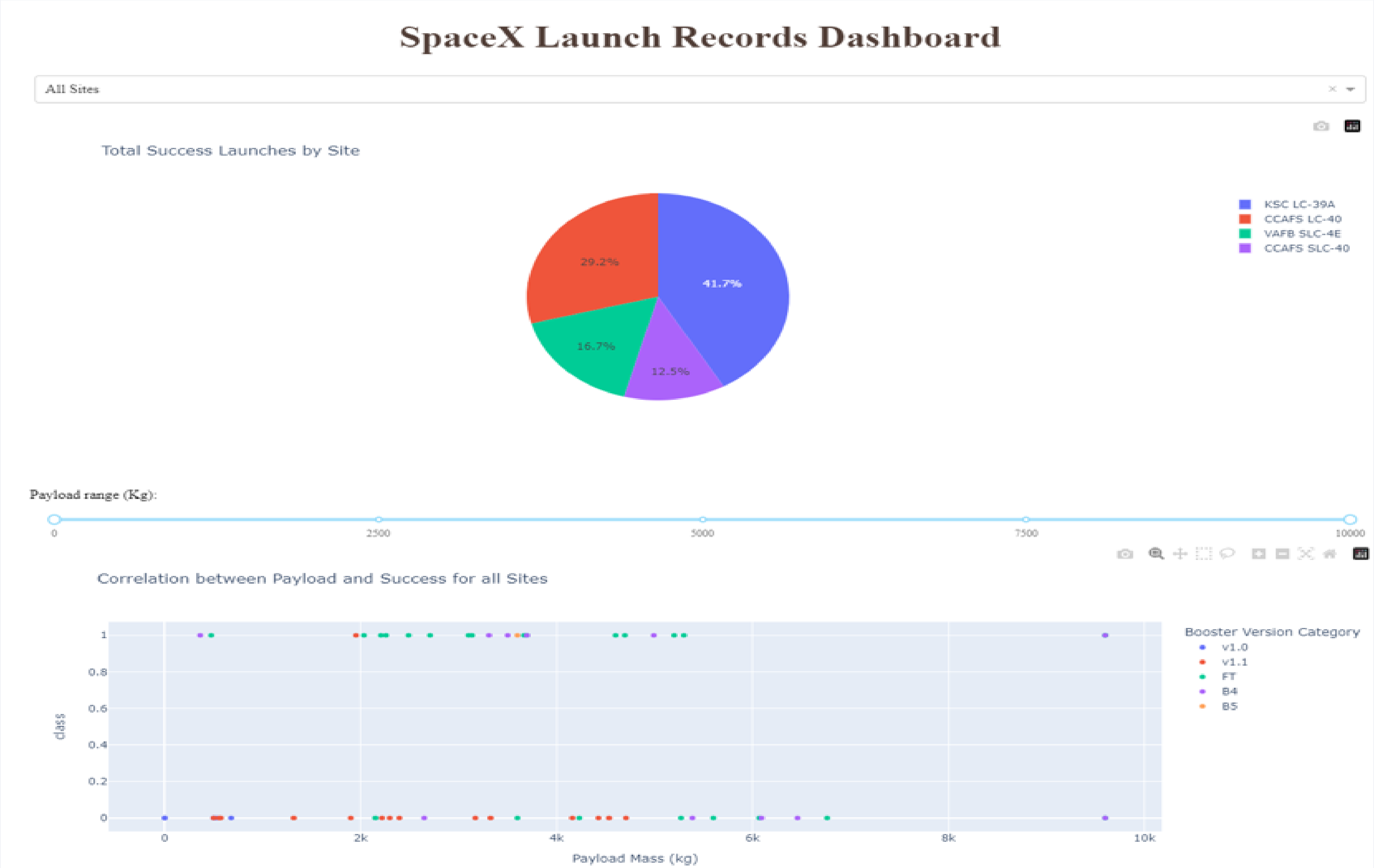
Build an Interactive Map with Folium

- Creating folium map to marked all the launch sites
- Creating map objects such as markers, circles, lines to mark the success or failure of launches for each launch site on the map.
- Creating a marker with distance to a closest city and drawing a line between the marker to the launch site
- HERE is the GitHub URL of the completed Web Scraping notebook: https://github.com/AshkanZand/IBM-Data-Science-Professional/blob/969feeda7844a92d3d68133dcbc0ab2d6b4881e1/SpaceX-Falcon-9/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Built an interactive dashboard application with plotly dash by:
  - Adding a Launch Site Drop-down Input Component
  - Adding a Callback Function to render Success-Pi-Chart based on selected Site Dropdown
  - Adding a Range Slider to Select Payload
  - Adding a Callback Function to render the Success-Payload-Scatter-Chart scatter plot
- HERE is the GitHub URL of the completed Web Scraping notebook: https://github.com/AshkanZand/IBM-Data-Science-Professional/blob/969feeda7844a92d3d68133dcbc0ab2d6b4881e1/SpaceX-Falcon-9/spacex_dash_app.py

Build a Dashboard with Plotly Dash



Predictive Analysis (Classification)

- Loading the data as a Pandas Dataframe and performing exploratory Data Analysis and determining Training Labels Using Sklearn Library

- Finding the best ML Model/method using the Test Data between SVM, Classification Trees, K nearest neighbors and Logistic Regression

create a column for the class -> data["Class"].to_numpy()

Standardize the data -> preprocessing.StandardScaler()

Split into training data and test data ->
train_test_split(X, Y, test_size=0.2, random_state=2)

Creating a GridSearchCv object for each model ->
GridSearchCV(estimator=, param_grid, cv=10)

Finding the best Hyperparameter and best Score using the validation data

Using the method Score to calculate the accuracy on the Test Data

Plotting a confusion matrix using the Test and prediction Outcomes

Predictive Analysis (Classification)

- Compering the test data accuracy score for each Classification Algorithm to which one performed the best

- HERE is the GitHub URL of the completed Web Scraping notebook: https://github.com/AshkanZand/IBM-Data-Science-Professional/blob/969feeda7844a92d3d68133dcbc0ab2d6b4881e1/SpaceX-Falcon-9/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

	Classification Algorithm	Accuracy Score
0	logreg	0.833333
1	KNN	0.833333
2	SVM	0.833333
3	Decision Tree	0.833333

Results

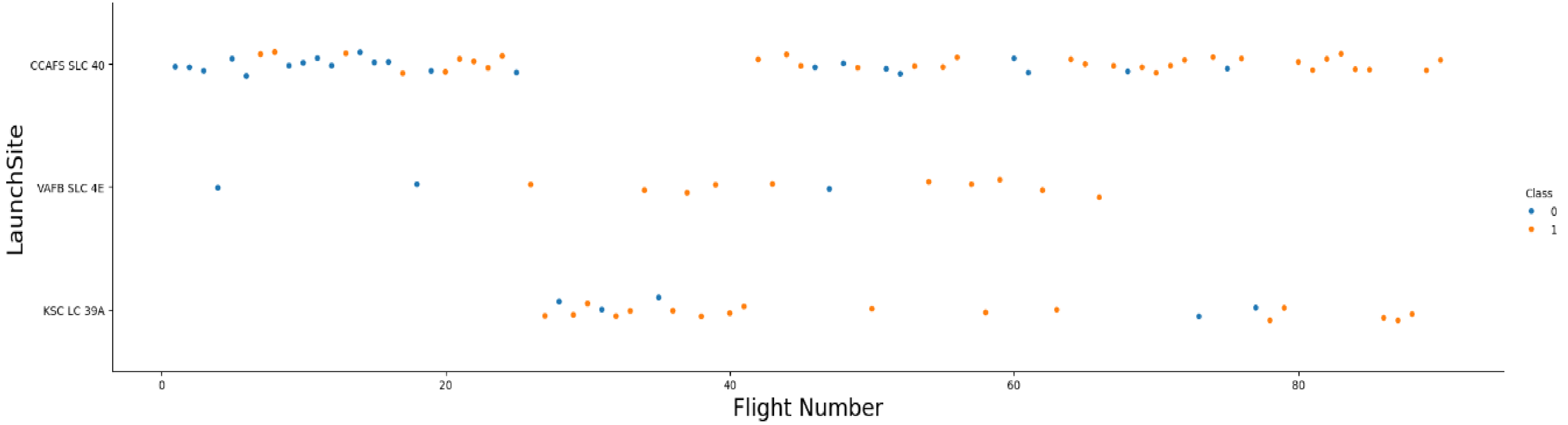
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Scatter plot of Flight Number vs. Launch Site with explanation

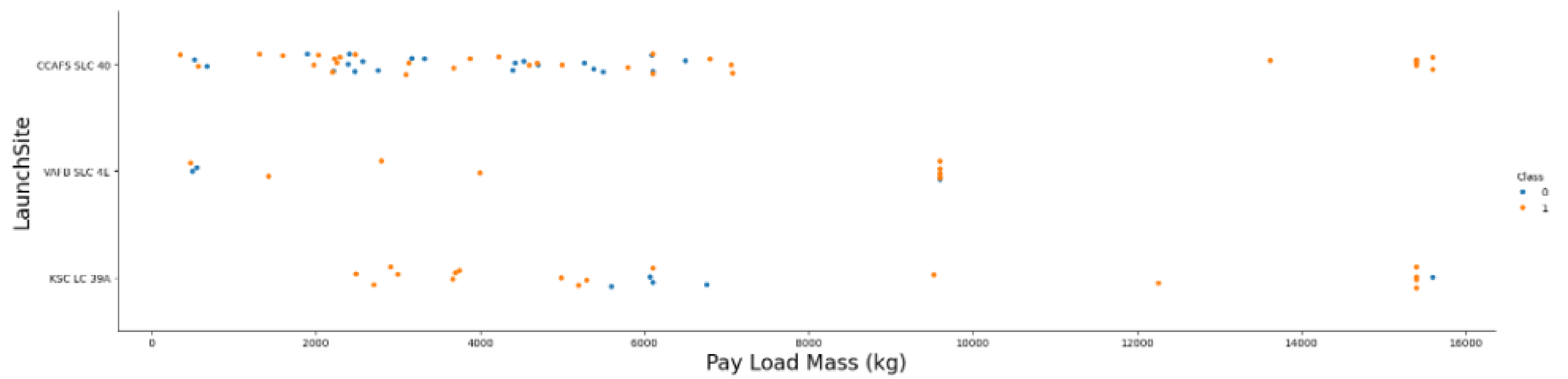


Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

We can observe that as the flight number increases at each of the three launch sites, the success rate also rises. For example, both CCAFS SLC 40 and KSC LC 39A achieve a 100% success rate after the 80th launch, while VAFB SLC 4E reaches a 100% success rate after the 50th launch

Payload vs. Launch Site

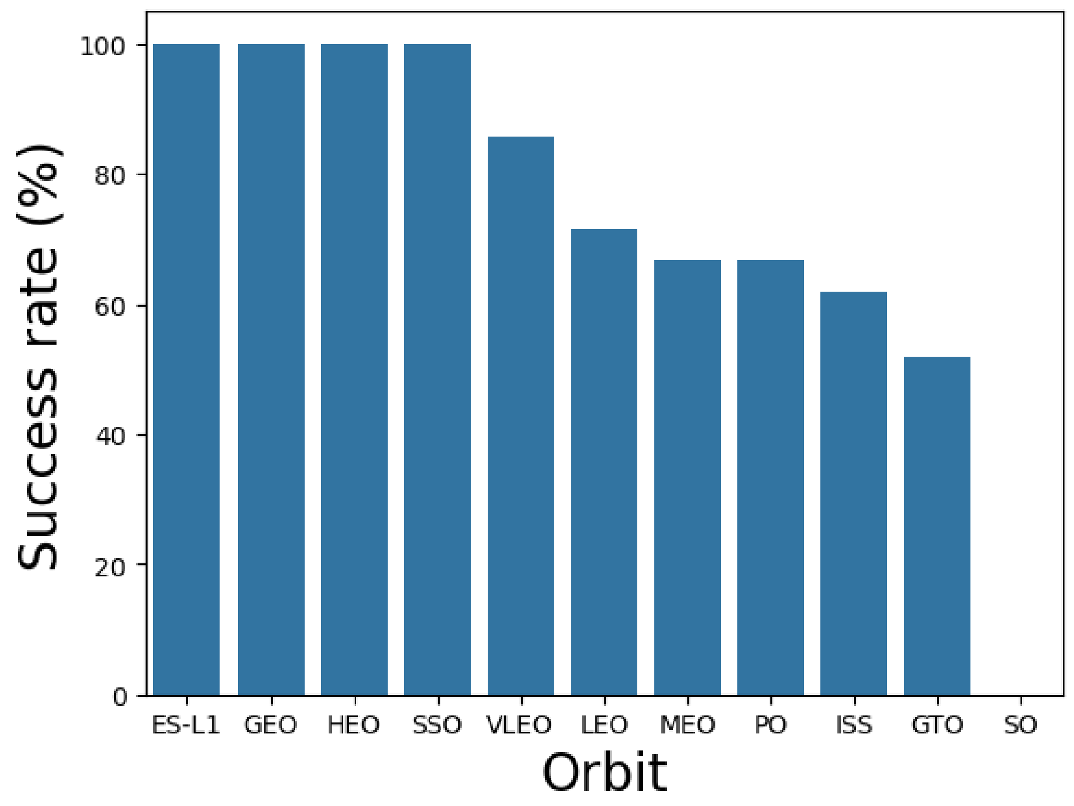
Scatter plot of Payload vs. Launch Site with explanation



Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

Barchart for the success rate of each orbit type with explanation

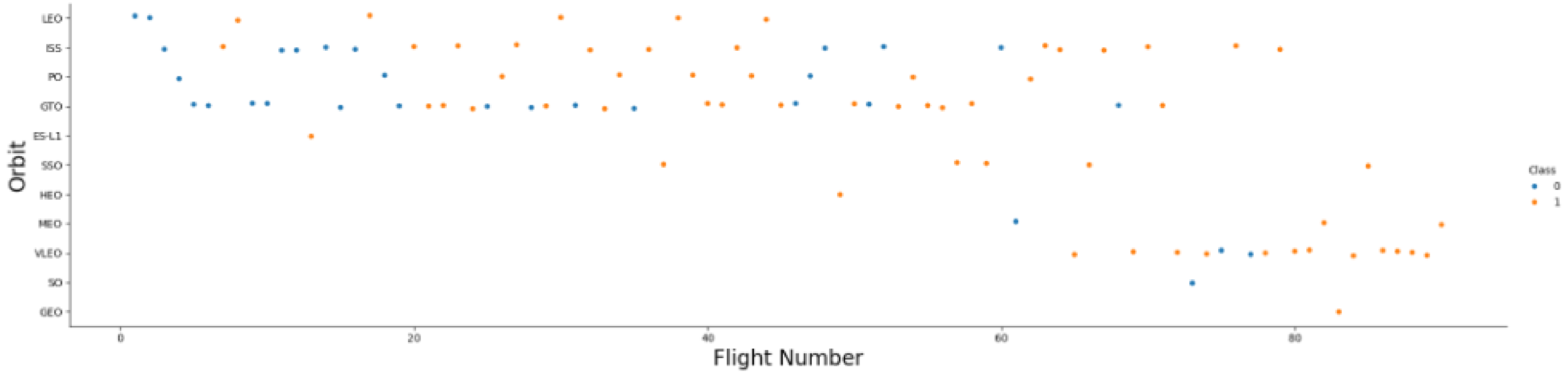


Analyze the plotted bar chart to identify which orbits have the highest success rates.

Orbits ES-L1, GEO, HEO & SSO have the highest success rates at 100%. Orbit SO has 0% success rate

Flight Number vs. Orbit Type

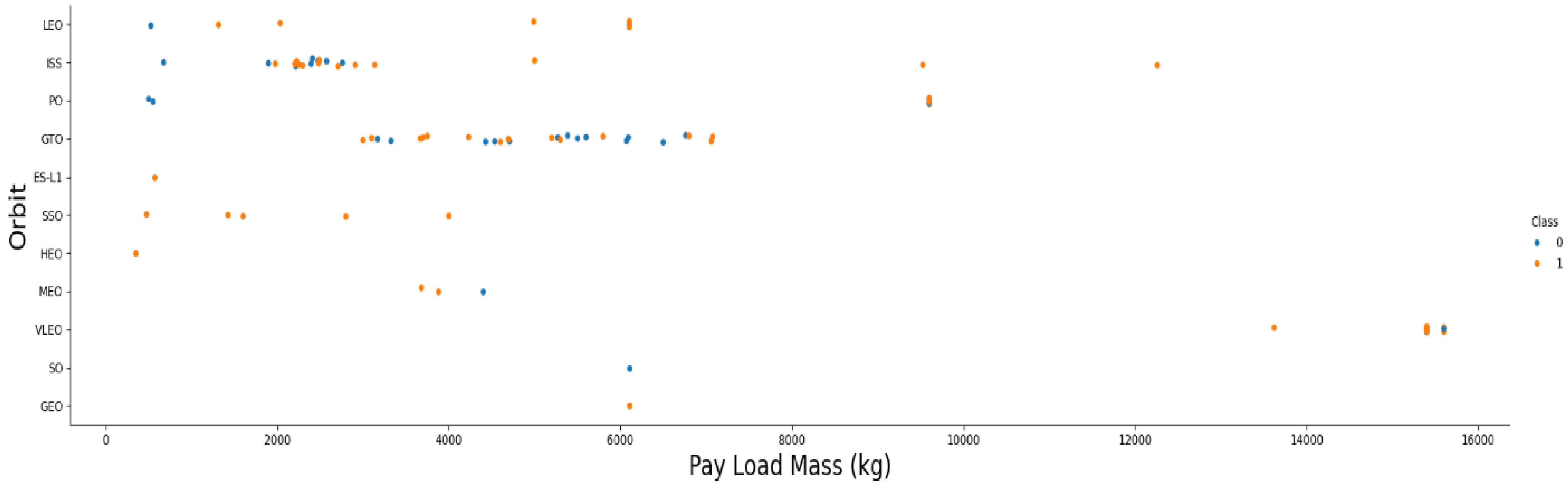
Scatter plot of Flight number vs. Orbit type with explanations



You 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 with explanations

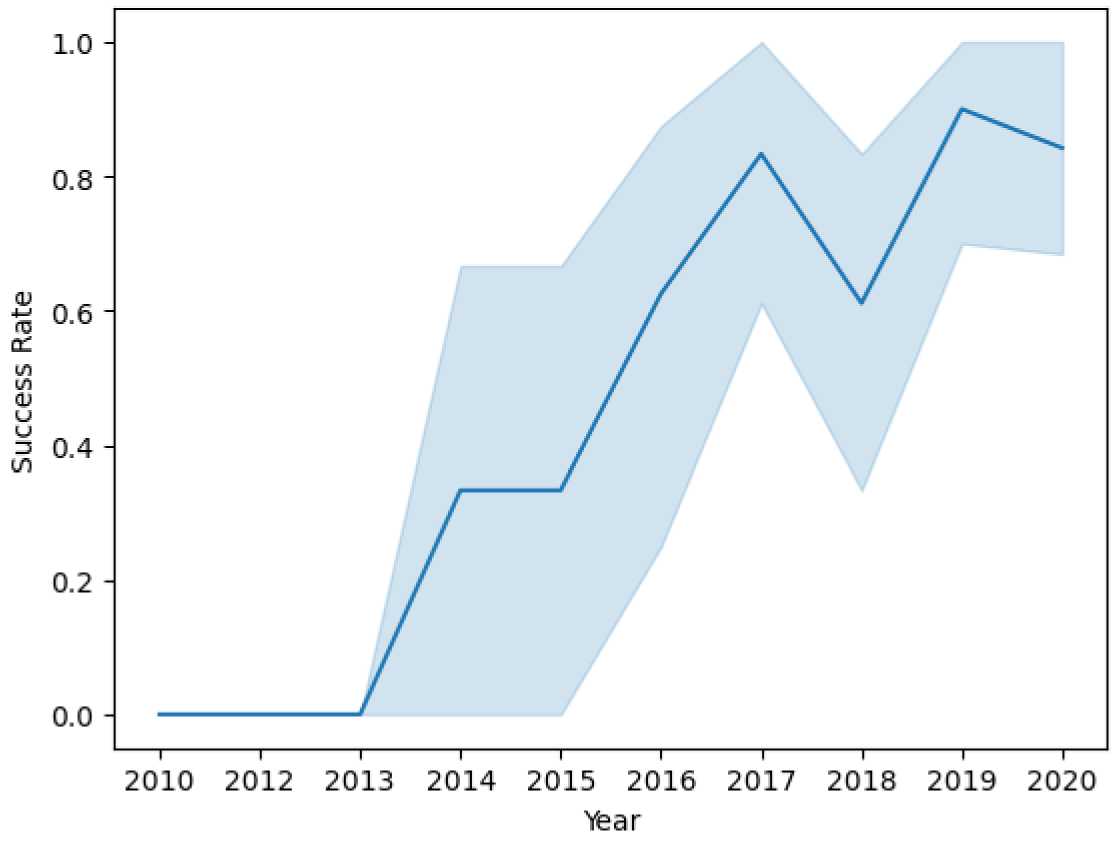


With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

Line chart of yearly average success rate with explanations



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Using “SELECT DISTINCT“ to return only the unique launch sites from the “LAUNCH_SITE“ column of the “SPACEXTBL“ table

Display the names of the unique launch sites in the space mission

```
%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

Launch Site Names Begin with 'CCA'

Used `**LIKE**` command with `**%**` wildcard in `**WHERE**` clause to select and display a table of all records where launch sites begin with the string `CCA`

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTABLE 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

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
```

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

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Using the `**SUM()**` function to return and display the total sum of `**PAYLOAD_MASS_KG**` column for customer `**NASA(CRS)**`

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg("PAYLOAD_MASS__KG_") as "AVG PAYLOAD MASS (KG)", Booster_Version as "Booster Version" from SPACEXTABLE where "Booster_Version" like "F9 v1.1%"
```

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

AVG PAYLOAD MASS (KG)	Booster Version
2534.6666666666665	F9 v1.1 B1003

Using `**AVG()**` function to return and display the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min("Date") from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)"
```

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

min("Date")
2015-12-22

Using the `**MIN()**` function to return and display the first (oldest) date when first successful landing outcome on ground pad `**Success (ground pad)**` happend

Successful Drone Ship Landing with Payload between 4000 and 6000

Using `**SELECT**` statement to return and list the names of boosters with payload **between 4000 and 6000** with landing outcome of **`Success (drone ship)`**

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select "Booster_Version", "Payload" from SPACEXTABLE where ("Landing_Outcome" = "Success (drone ship)"  and 4000<"PAYLOAD_MASS__KG_"<6000)
```

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

Booster_Version	Payload
F9 FT B1021.1	SpaceX CRS-8
F9 FT B1022	JCSAT-14
F9 FT B1023.1	Thaicom 8
F9 FT B1026	JCSAT-16
F9 FT B1029.1	Iridium NEXT 1
F9 FT B1021.2	SES-10
F9 FT B1029.2	BulgariaSat-1
F9 FT B1036.1	Iridium NEXT 2
F9 FT B1038.1	Formosat-5
F9 B4 B1041.1	Iridium NEXT 3
F9 FT B1031.2	SES-11 / EchoStar 105
F9 B4 B1042.1	Koreasat 5A
F9 B4 B1045.1	Transiting Exoplanet Survey Satellite (TESS)
F9 B5 B1046.1	Bangabandhu-1

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(Mission_Outcome) as TOTAL from SPACEXTABLE group by Mission_Outcome
```

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

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

Using **`COUNT`** together with the **`GROUP BY`** statement to return total number of missions outcomes

Boosters Carried Maximum Payload

Using a Subquerry to return and pass the Max payload and used it list all the boosters that have carried the Max payload.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%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

2015 Launch Records

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql
select substr(Date, 0,5) as "Year", substr(Date, 6,2) as "Month","Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE
where ( substr(Date,0,5)='2015' and Landing_Outcome like "Failure (drone ship)")
```

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

Year	Month	Landing_Outcome	Booster_Version	Launch_Site
2015	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Using the \`**subsrt()**\` in the select statement to get the month and year from the date column where **substr(Date,0,5)=‘2015‘** for year and **Landing_outcome** was \`Failure (drone ship)\` and return the records matching the filter.

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

Ranking the count of landing using ‘**COUNT**’ and grouping the ‘**Landing_Outcome**’ and filtering the date between 210-06-04 and 2017-03-20, in descending order

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.

```
%%sql
select "Landing_Outcome", count("Landing_Outcome") from SPACEXTABLE where ("2010-06-04"<"Date"< "2017-03-20")
group by "Landing_Outcome"
ORDER BY count("Landing_Outcome") desc
```

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

Landing_Outcome	count("Landing_Outcome")
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

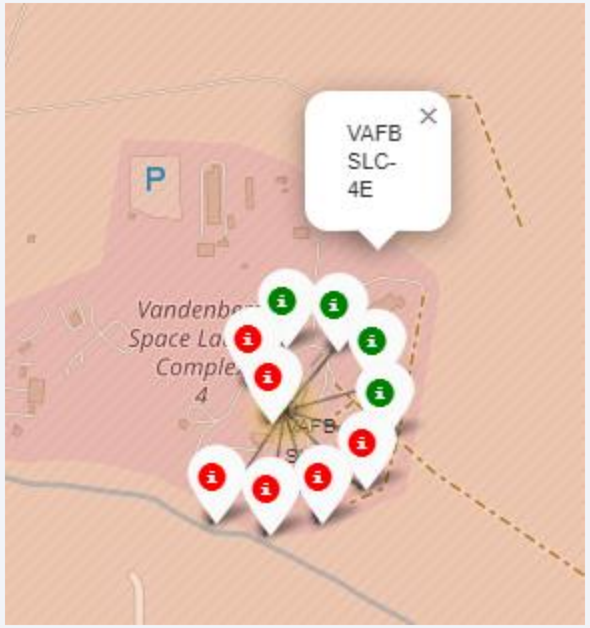
Section 3

Launch Sites Proximities Analysis

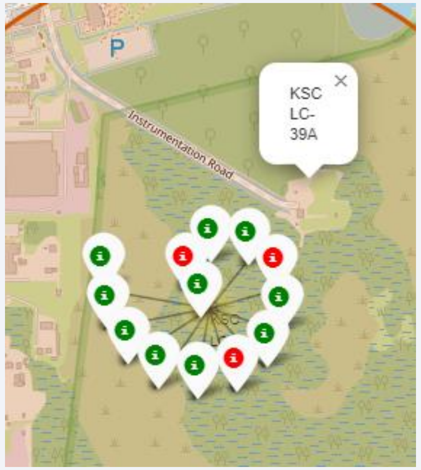
Markers of all Launch Sites on Global Map

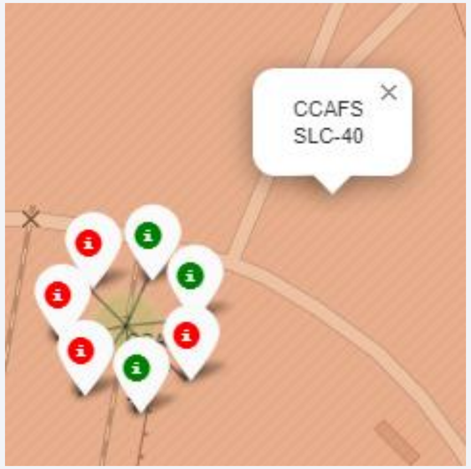
- All launch sites are in proximity to the Equator line
- All launch sites are in very close proximity to the coast (located southwards of USA).

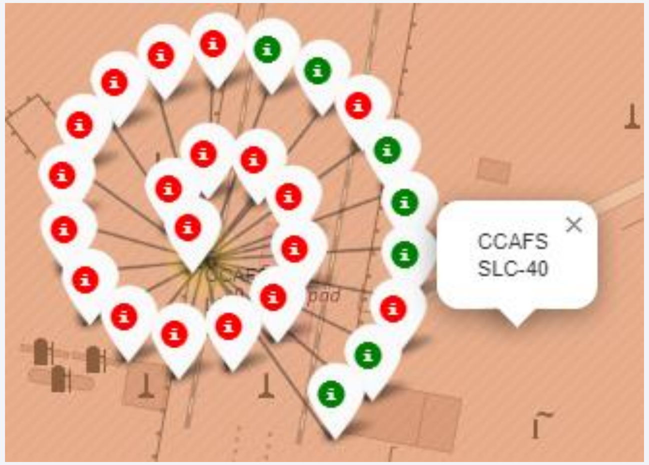
Launch Outcomes for each Site on the Map



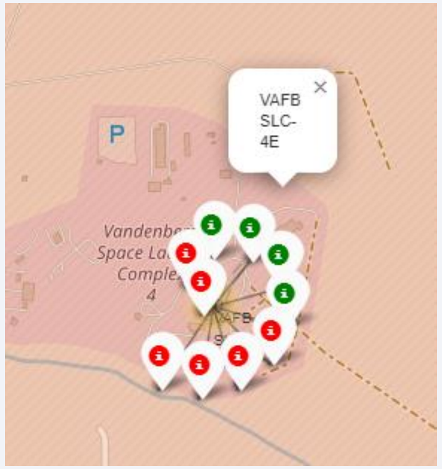
Launch Outcomes for each Site on the Map







East Coast Florida Sites



West Coast California

- In the East Coast (Florida), KSC LC-39A has relatively high success rates (77%) compared to CCAFS SLC-40 (43%) and CCAFS LC-40 (27%).

- In the West Coast (California), VAFB SLC-4E has relatively lower success rates (40%) compared to KSC LC-39A launch site in the East Coast.

Distance between a launch site to its proximities

Launch site CCAFS SLC-40 proximity to coastline is 0.86 KM

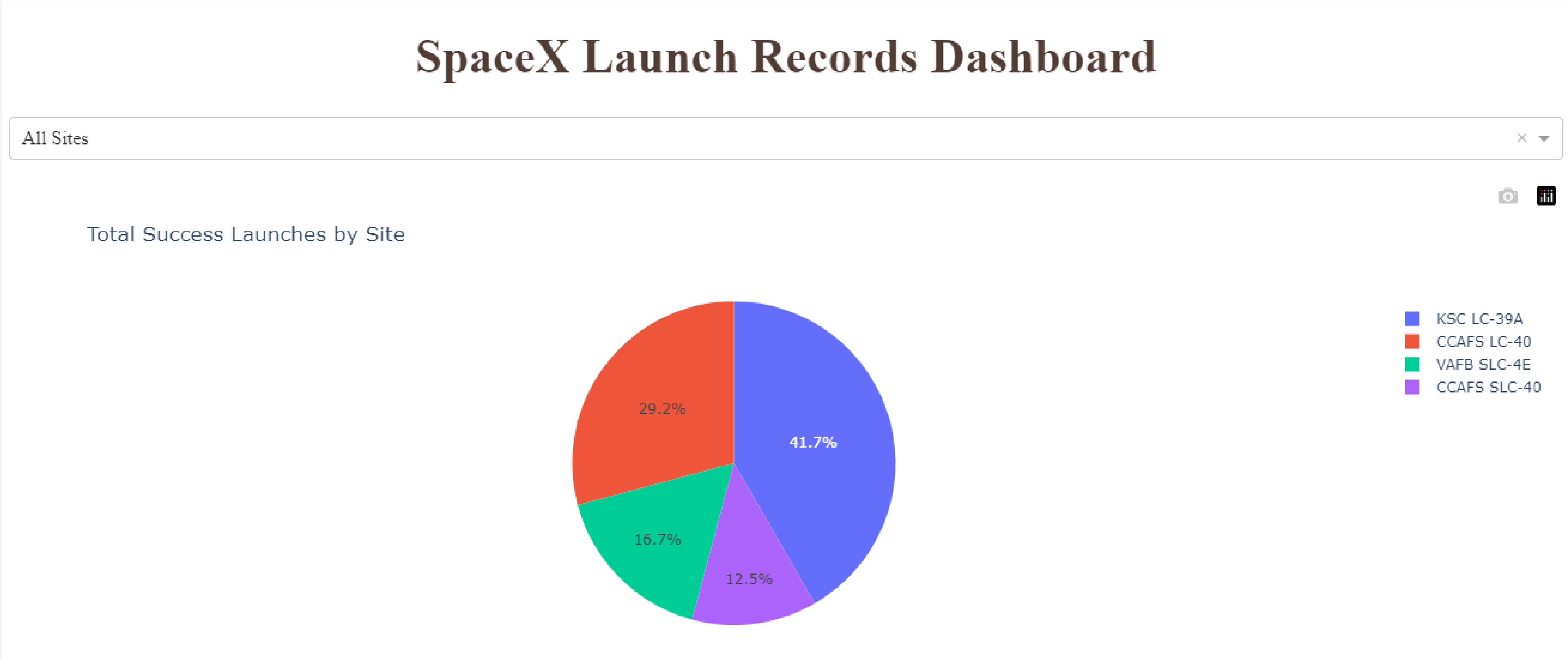
Distance between a launch site to its proximities

Launch site CCAFS SLC-40 proximity to the city Cape Canaveral is 19.69 KM

Section 4

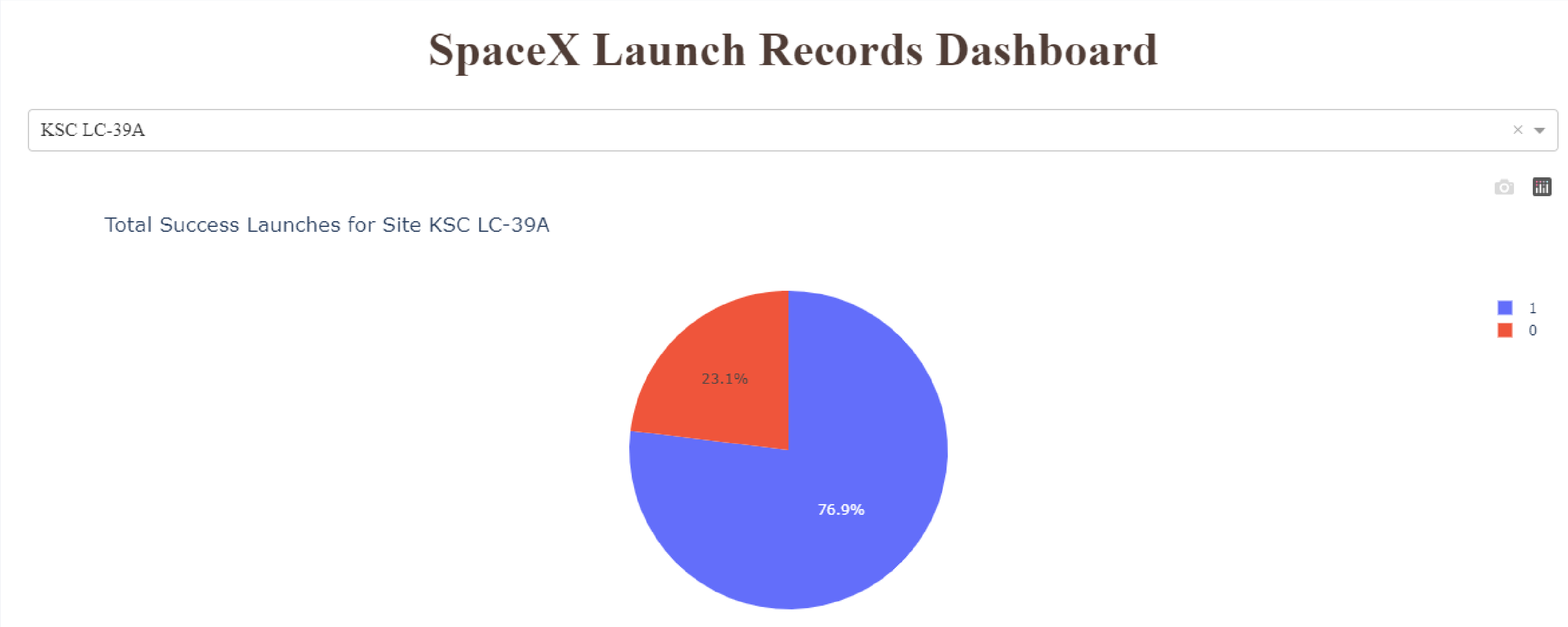
Build a Dashboard with Plotly Dash

Pie-Chart for Launch success count for all sites



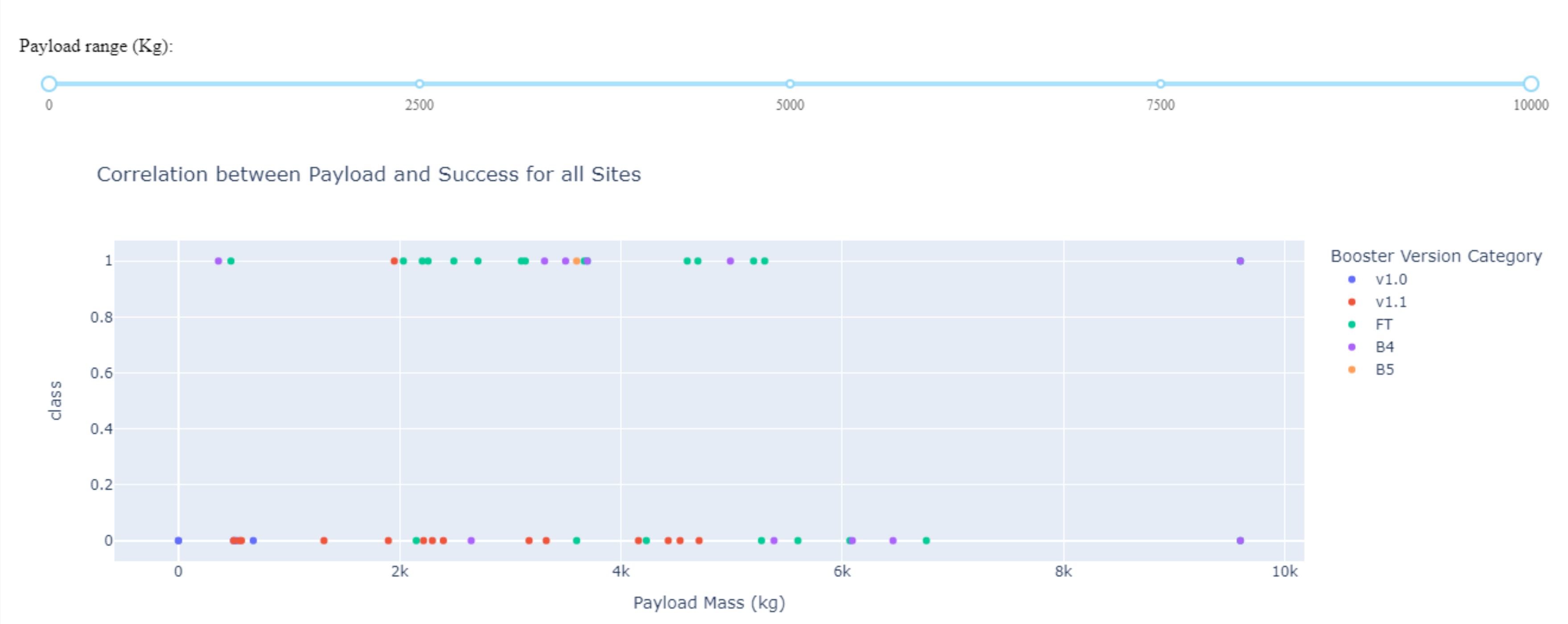
Launch site KSC LC-39A has the highest launch success rate at 42% followed by CCAFS LC-40 at 29%, VAFB SLC-4E at 17% and lastly launch site CCAFS SLC-40 with a success rate of 13%.

Pie-Chart for the Launch site with highest success rate



Launch site KSC LC-39A had the highest success ratio of 73% success against 27% failed launches

Payload vs. Launch outcome scatter plot for all sites



F9 Booster version FT has the highest launch success rate

Payload vs. Launch outcome scatter plot for all sites

- According to the plots the Payload range of (2500 to 5000) has the highest launch success rate (55%).
- The payload in the range of (5000 to 7500) has the lowest launch success rate (22%).
- For the payload in the range of (7500 to 10000) there was only two lunches, one successed, one failed.

Section 5

Predictive Analysis (Classification)

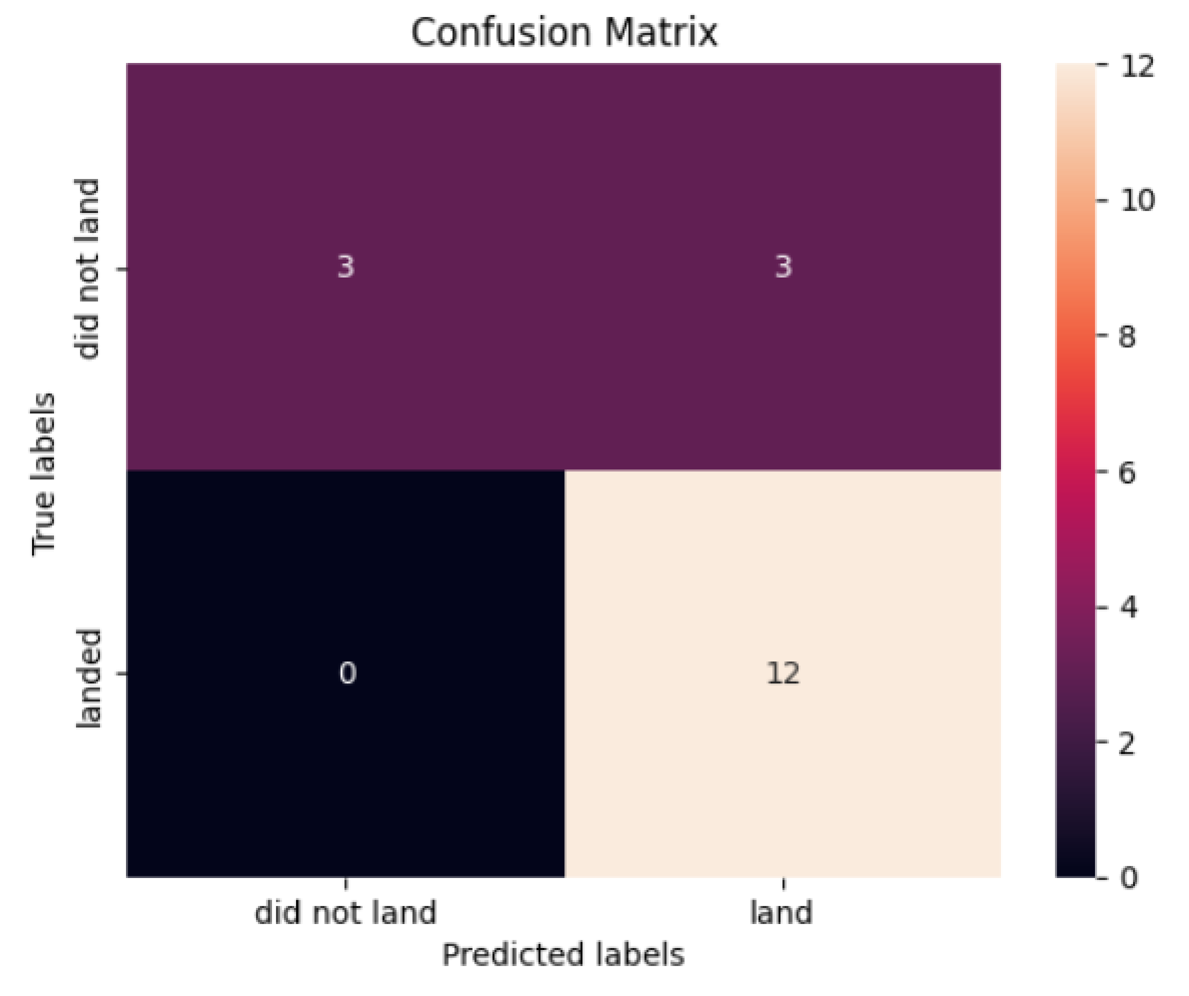
Classification Accuracy

	Classification Algorithm	Accuracy Score
0	logreg	0.833333
1	KNN	0.833333
2	SVM	0.833333
3	Decision Tree	0.833333

All the methods perform equally on the test data, they all have the same accuracy of 0.833 on the test Data.

Confusion Matrix

All the 4 classification model had the same confusion matrixes and were able equally distinguish between the different classes. The major problem is false positive for all the models.



Conclusions

- Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60%, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- We can conclude that as the flight number increases at each of the three launch sites, the success rate also improves. For example, the VAFB SLC 4E launch site achieves a 100% success rate after 50 flights, while both KSC LC 39A and CCAFS SLC 40 reach a 100% success rate after 80 flights.
- If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass greater than 10000.
- The orbits ES-L1, GEO, HEO, and SSO boast the highest success rates at 100%, while GTO orbit has the lowest success rate, around 50%, and a 0% success rate for Orbit SO.
- In LEO orbit, success seems to be linked to the number of flights; however, in GTO orbit, there doesn't appear to be any correlation between the flight number and success.
- For heavy payloads, the successful landing rate is higher for Polar, LEO, and ISS missions. However, in the case of GTO, it's difficult to discern a clear pattern, as both successful and unsuccessful landings are present.
- The success rate steadily increased from 2013 until 2020.
- By applying various machine learning classification algorithms, we can predict the success or failure of landings with an accuracy of 0.833.

Thank you!