IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Murat Benbanaste

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Summary of methodologies

- Data collection
- EDA (Data wrangling)
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

SpaceX is arguably the most successful private company trying to make affordable space travel possible. Due to the technology they have, their Falcon 9 rocket launches cost 62 million dollars, while for other providers it costs upward of 165 million dollars each. The technology that makes SpaceX have this advantage concerns re-useability of the first stage. As such, if whether the first stage will land or not could be determined, the costs of a launch could also be determined. Based on public information on previous launches and machine learning models, we are going to predict whether the next Falcon 9 will land successfully, i.e., whether SpaceX will reuse the first stage.

The questions to be answered include:

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used in this case?

Section 1

Methodology

Methodology

- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform exploratory data analysis (EDA), i.e., data wrangling
  - Filtering the data
  - Handling with missing values
  - One Hot Encoding to prepare the data to a binary classification
- Perform EDA using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

Data collection process involved API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

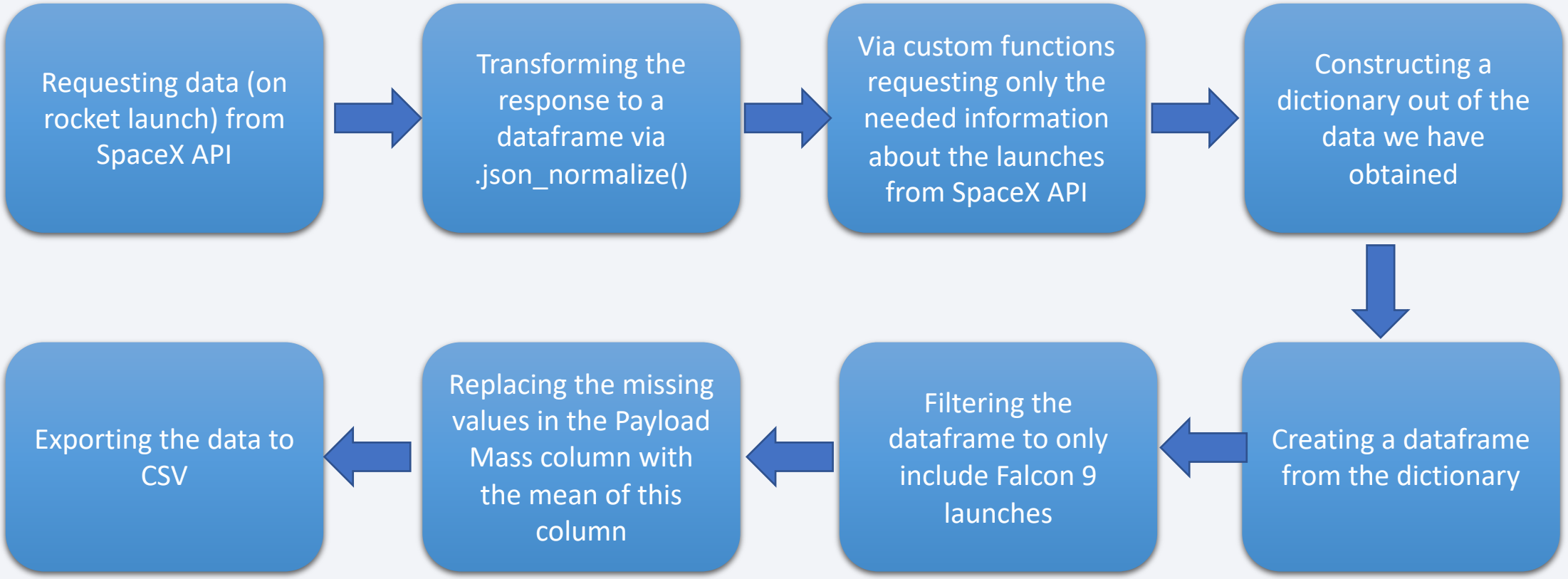
Data Columns obtained via SpaceX REST API

FlightNumber	Reused
Date	Legs
BoosterVersion	LandingPad
PayloadMass	Block
Orbit	ReusedCount
LaunchSite	Serial
Outcome	Longitude
Flights	Latitude
GridFins	

Data Columns obtained via Web Scraping Wikipedia

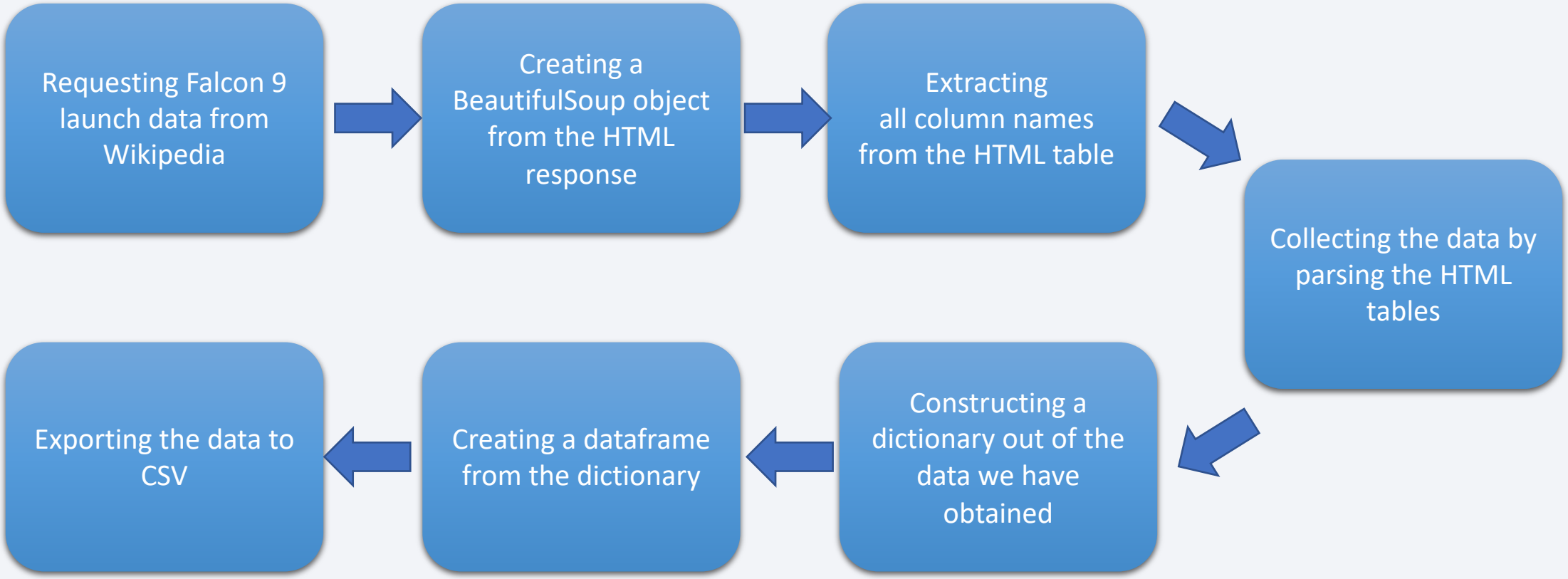
Flight No.	Launch outcome
Launch site	Version Booster
Payload	Booster landing
PayloadMass	Date
Orbit	Time
Customer	

Data Collection – SpaceX API



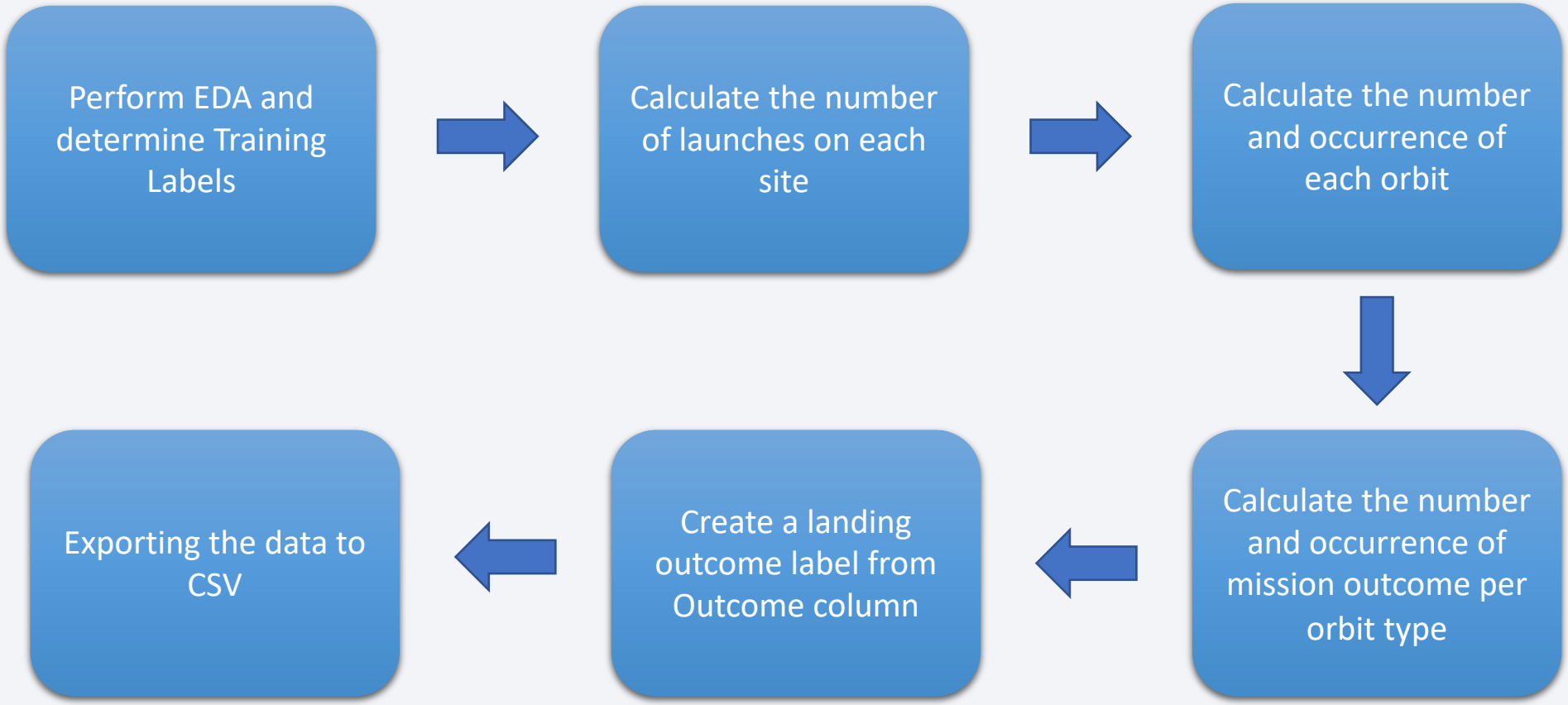
GitHub URL: Data Collection API

Data Collection – Web Scraping



GitHub URL: Data Collection with Web Scraping

EDA - Data Wrangling



GitHub URL: EDA (Data Wrangling)

EDA - Data Wrangling

In several cases, the booster did not land successfully due to varying reasons, e.g., an accident; different outcome types reflect this variance.

`True Ocean` means that the rocket successfully landed on the ocean.

`False Ocean` means that the rocket unsuccessfully landed on the ocean.

`True RTLS` means that the rocket successfully landed to a ground pad.

`False RTLS` means that the rocket unsuccessfully landed to a ground pad.

`True ASDS` means that the rocket successfully landed on a drone ship.

`False ASDS` means that the rocket unsuccessfully landed on a drone ship.

We convert these differing outcomes into training labels, where “1” means that the booster successfully landed, “0” means that it was unsuccessful.

EDA with Data Visualization

Charts plotted were:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs Orbit Type
- Success Rate Yearly Trend

- Scatter plots depict the relationship between continuous variables.
- Bar charts depict comparisons among discrete categories.
- Line charts show trends in data over time (time series).

GitHub URL: EDA with Data Visualization

EDA with SQL

SQL queries performed:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

GitHub URL: EDA with SQL

Build an Interactive Map with Folium

Markers of all Launch Sites:

Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

Colored Markers of the launch outcomes for each Launch Site:

Added colored Markers of successful (**Green**) and failed (**Red**) launches using Marker Cluster to identify the success rates of the launch sites.

Distances between a Launch Site to its proximities:

Added a line to show the distance between the Launch Site VAFB SLC-4E (as an example) and its proximity to the ocean

GitHub URL: Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

Added a pie chart to depict the total successful launches count for all sites and, if a specific Launch Site was selected, the Success vs. Failed counts for the selected site.

Slider of Payload Mass Range:

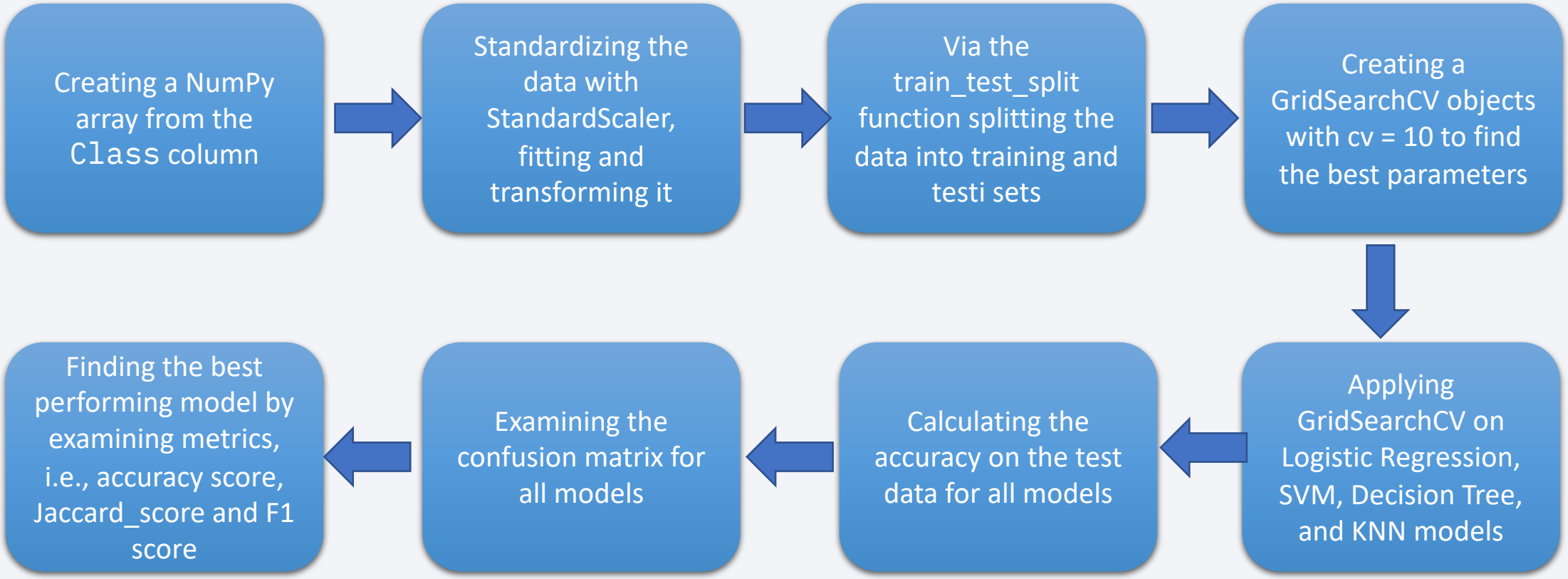
Added a slider to select Payload mass range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

Added a scatter chart to depict the correlation between Payload mass and Launch Success.

GitHub URL: SpaceX Dash App

Predictive Analysis (Classification)



GitHub URL: Machine Learning Prediction

Results

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

Section 2

Insights drawn from EDA

EDA with Data Visualization

Flight Number vs. Payload Mass

Flight Number vs. Launch Site

Payload Mass vs. Launch Site

Orbit Type vs. Success Rate

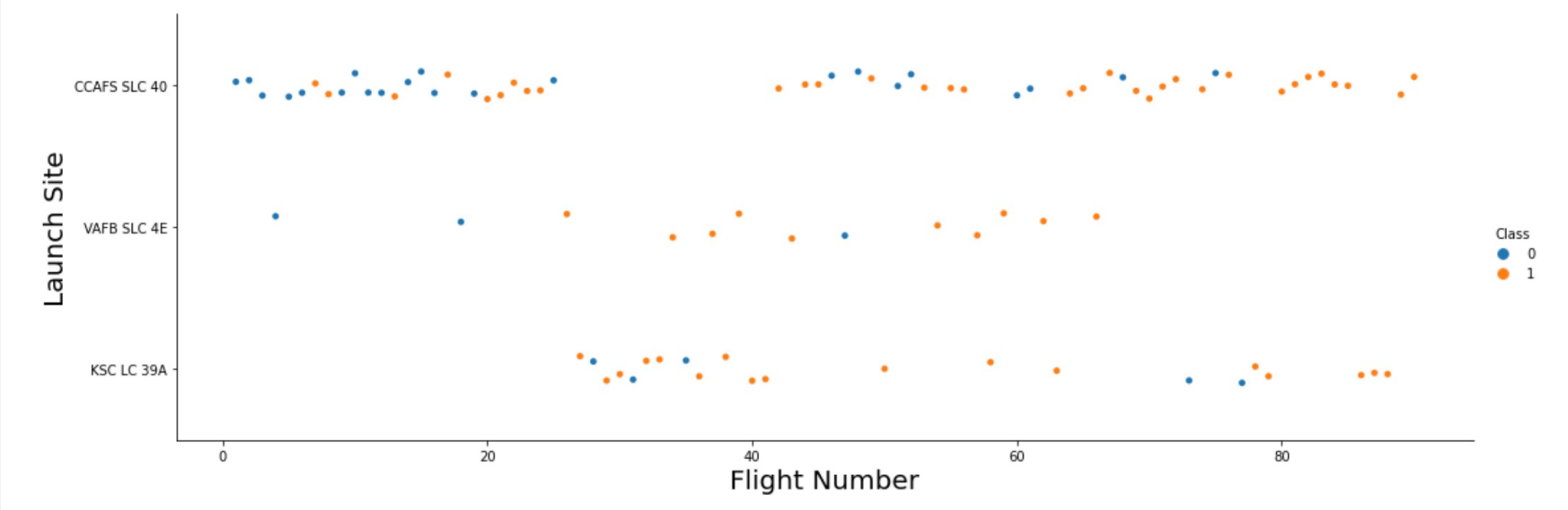
Flight Number vs. Orbit Type

Payload Mass vs Orbit Type

Success Rate Yearly Trend

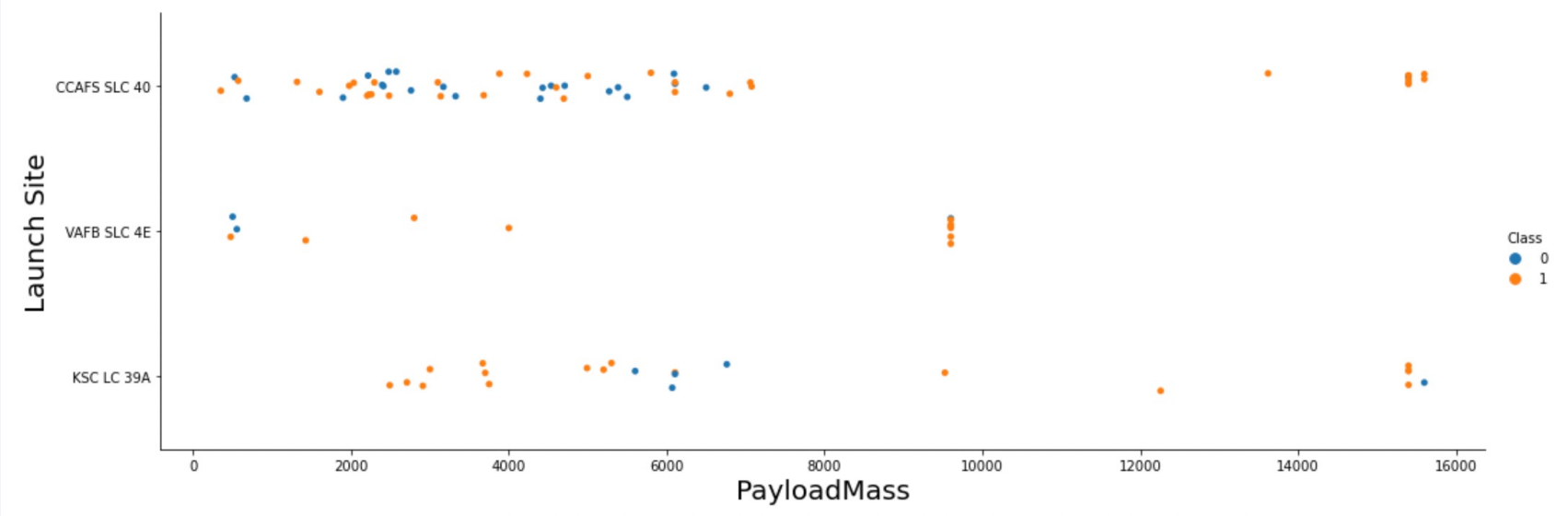
Flight Number vs. Launch Site

- The earliest flights mostly failed; the success rate increased in later flights.
- Almost half of all launches were at the CCAFS SLC 40 launch site.
- Due mostly to the high failure rate at the earliest flights, CCAFS SLC 40 has lower success rates relative to VAFB SLC 4E and KSC LC 39A.



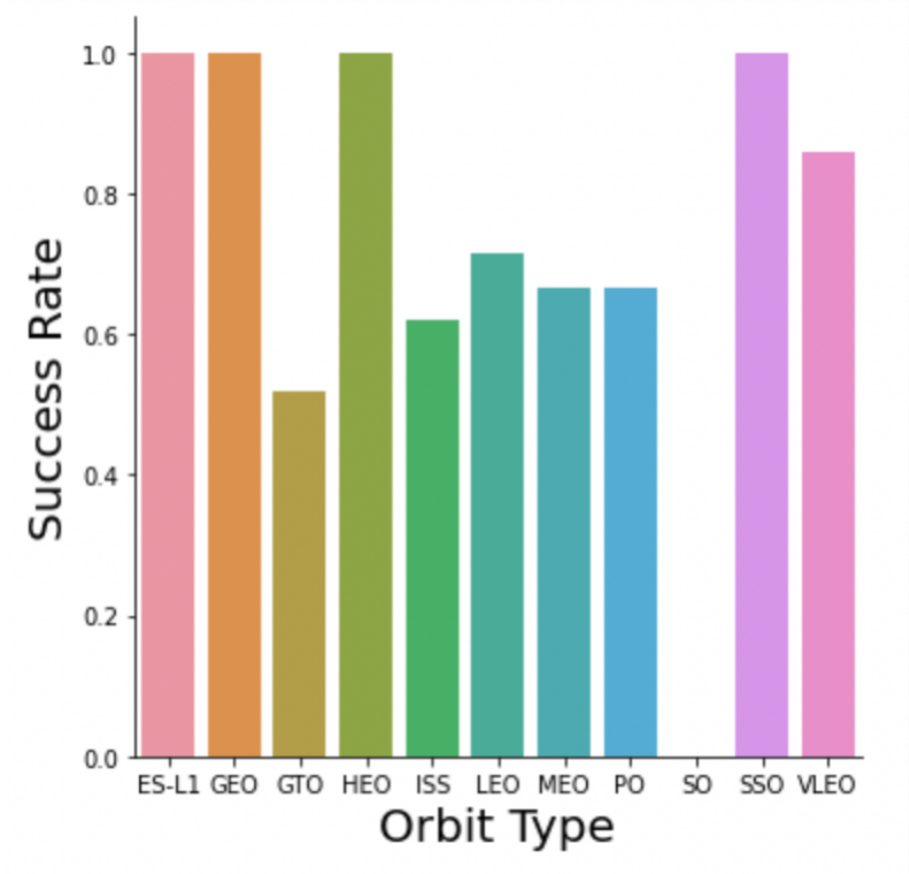
Payload vs. Launch Site

- Higher the payload mass corresponds to higher the success rate in every launch site.
- Success rate is significantly high where the payload mass is over 7000 kg.
- Noticeably, though, KSC LC 39A has a perfect success rate where payload mass is under 5500 kg.



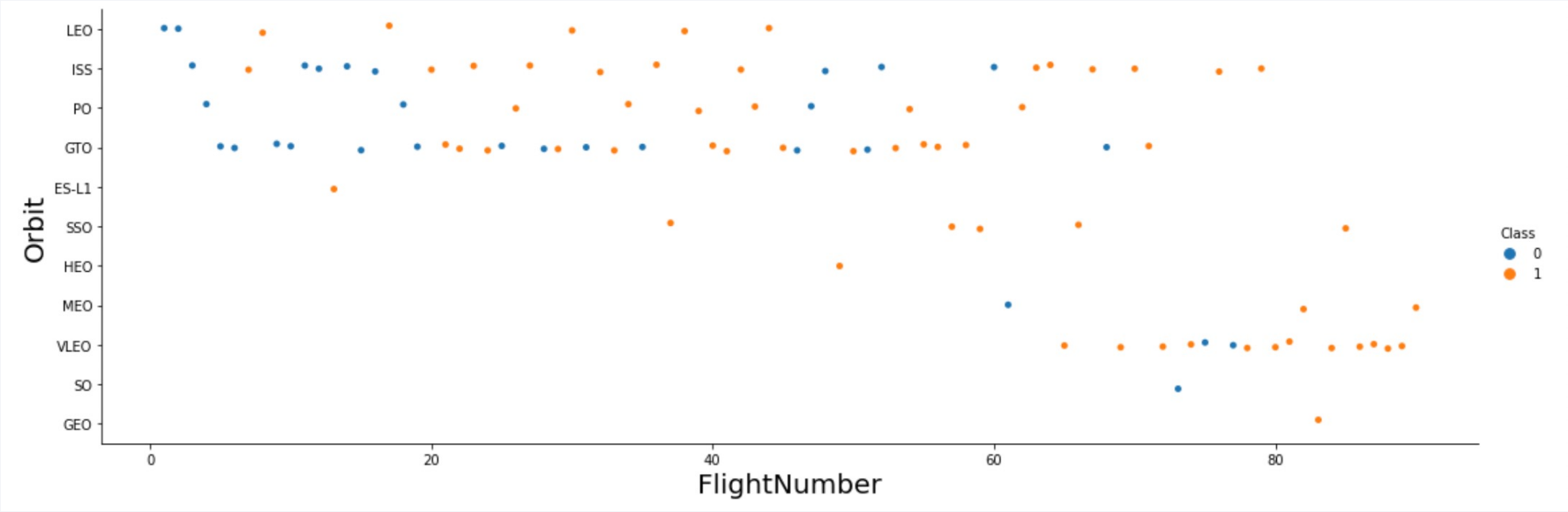
Success Rate vs. Orbit Type

- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO
- Orbits with 0% success rate:
  - SO



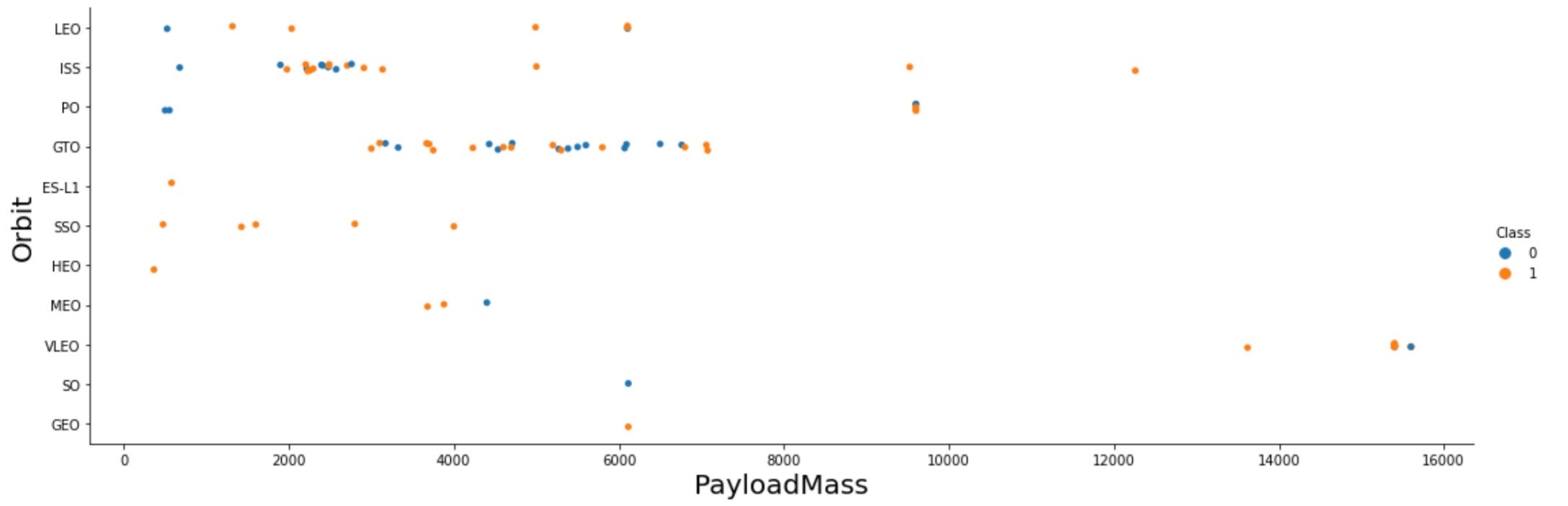
Flight Number vs. Orbit Type

- GTO seems to exhibit no pattern based on flight number.
- Relatively most apparent pattern is exhibited by LEO orbit where the success rate increases as the flight number increases.



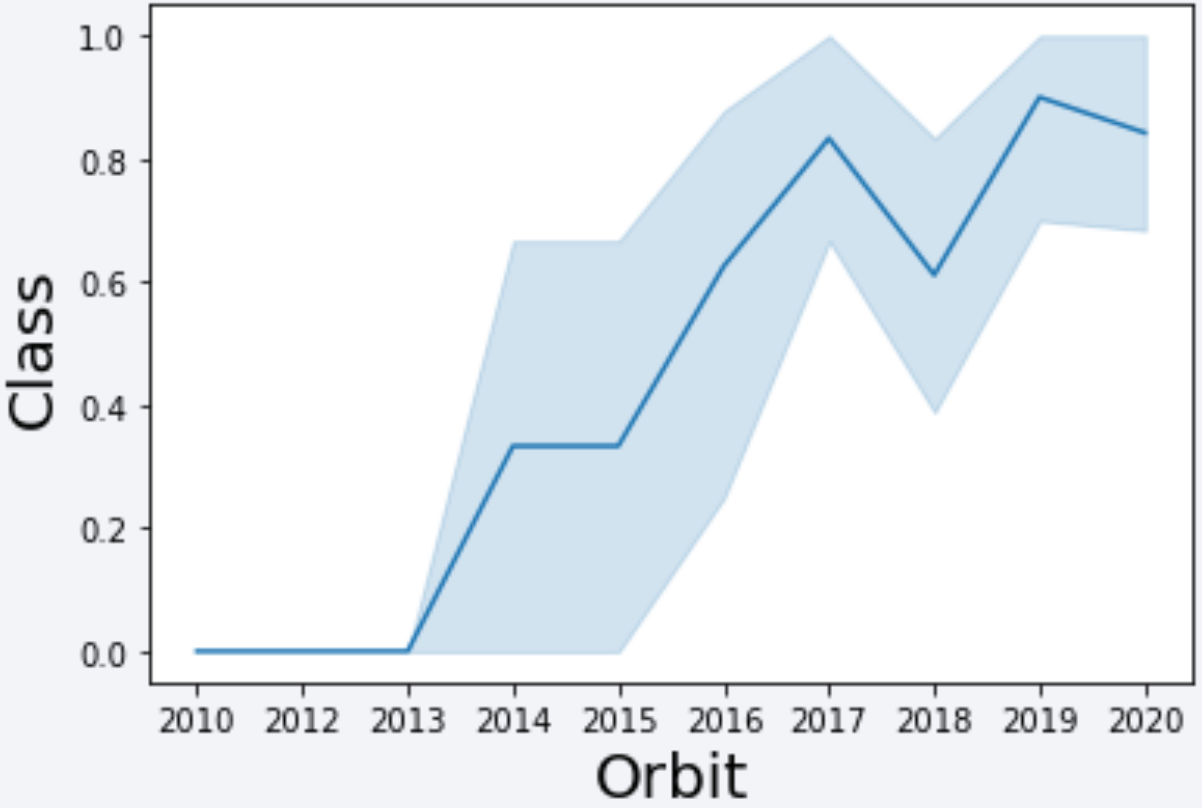
Payload vs. Orbit Type

- Arguably, GTO orbit seems to be influenced most negatively by heavier payload masses.
- LEO and ISS orbits seem to increase their success rates as the payload mass increases.



Launch Success Yearly Trend

Since 2013, the success rate has been increasing overall – despite brief relative decreases.



EDA with SQL

SQL queries performed:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

All Unique Launch Site Names

In [5]:
```
%%sql

SELECT DISTINCT (launch_site) FROM SpaceX
```

```
 * ibm_db_sa://pyq14834:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
1/bludb
Done.
```

Out[5]:

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

5 Launch Site Names Beginning with 'CCA'

In [6]:
```
%%sql

Select * From SpaceX Where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://pyq14834:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
1/bludb
Done.
```

Out[6]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 carried by Boosters launched

```
In [7]: %%sql

Select sum(payload_mass__kg_) from SpaceX where customer = 'NASA (CRS)';

 * ibm_db_sa://pyq14834:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
1/bludb
Done.

Out[7]:
```

1
45596

Average Payload Mass by Booster Version F9 v1.1

In [17]:
```
%%sql

Select avg(payload_mass__kg_) as average_payload_mass from SpaceX where booster_version like '%F9 v1.1%'
```

```
 * ibm_db_sa://pyq14834:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
1/bludb
Done.
```

Out[17]:

average_payload_mass
2534

First Successful Ground Landing Date

In [9]:
```
%%sql

select min(date) from spaceX where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://pyq14834:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
1/bludb
Done.
```

Out[9]:

1
2015-12-22

Boosters with Successful Drone Ship Landing (Payload between 4000-6000)

In [19]:
```
%%sql

select booster_version from spacex where landing__outcome ='Success (drone ship)' and payload_mass__kg_ > 4000
and payload_mass__kg_<6000
```

```
 * ibm_db_sa://pyq14834:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
1/bludb
Done.
```

Out[19]:

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

Total Number of Successful and Failure Mission Outcomes

In [11]:
```
%%sql

select mission_outcome, count(mission_outcome) as count from spacex group by mission_outcome
```

```
 * ibm_db_sa://pyq14834:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
1/bludb
Done.
```

Out[11]:

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters that carried Maximum Payload

In [12]:
```
%%sql

select booster_version from spacex where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex)
```

```
 * ibm_db_sa://pyq14834:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
1/bludb
Done.
```

Out[12]:

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 Failed Drone Ship Landing Records

In [20]:
```
%%sql

select date, booster_version,launch_site from spacex where landing__outcome like 'Failure (drone ship)'
and year(date) = 2015
```

```
 * ibm_db_sa://pyq14834:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
1/bludb
Done.
```

Out[20]:

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

In [14]:
```
%%sql

select landing__outcome, count(*) as count from spacex where date between '2010-06-04'
and '2017-03-20' group by landing__outcome order by count desc;
```

```
 * ibm_db_sa://pyq14834:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
1/bludb
Done.
```

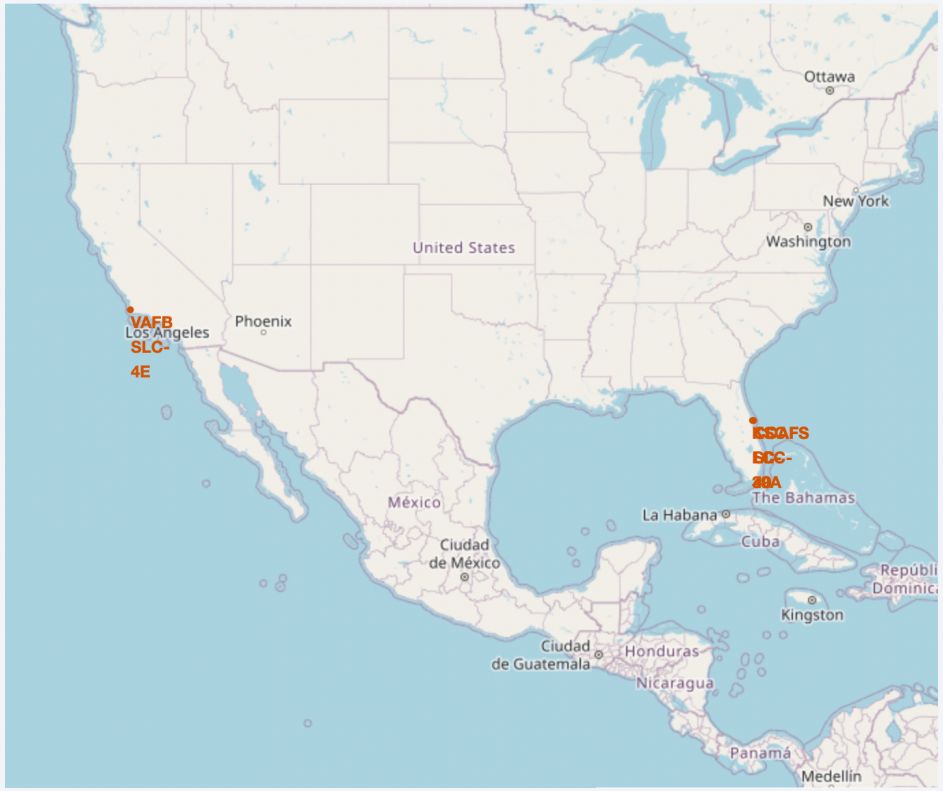
Out[14]:

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

Section 4

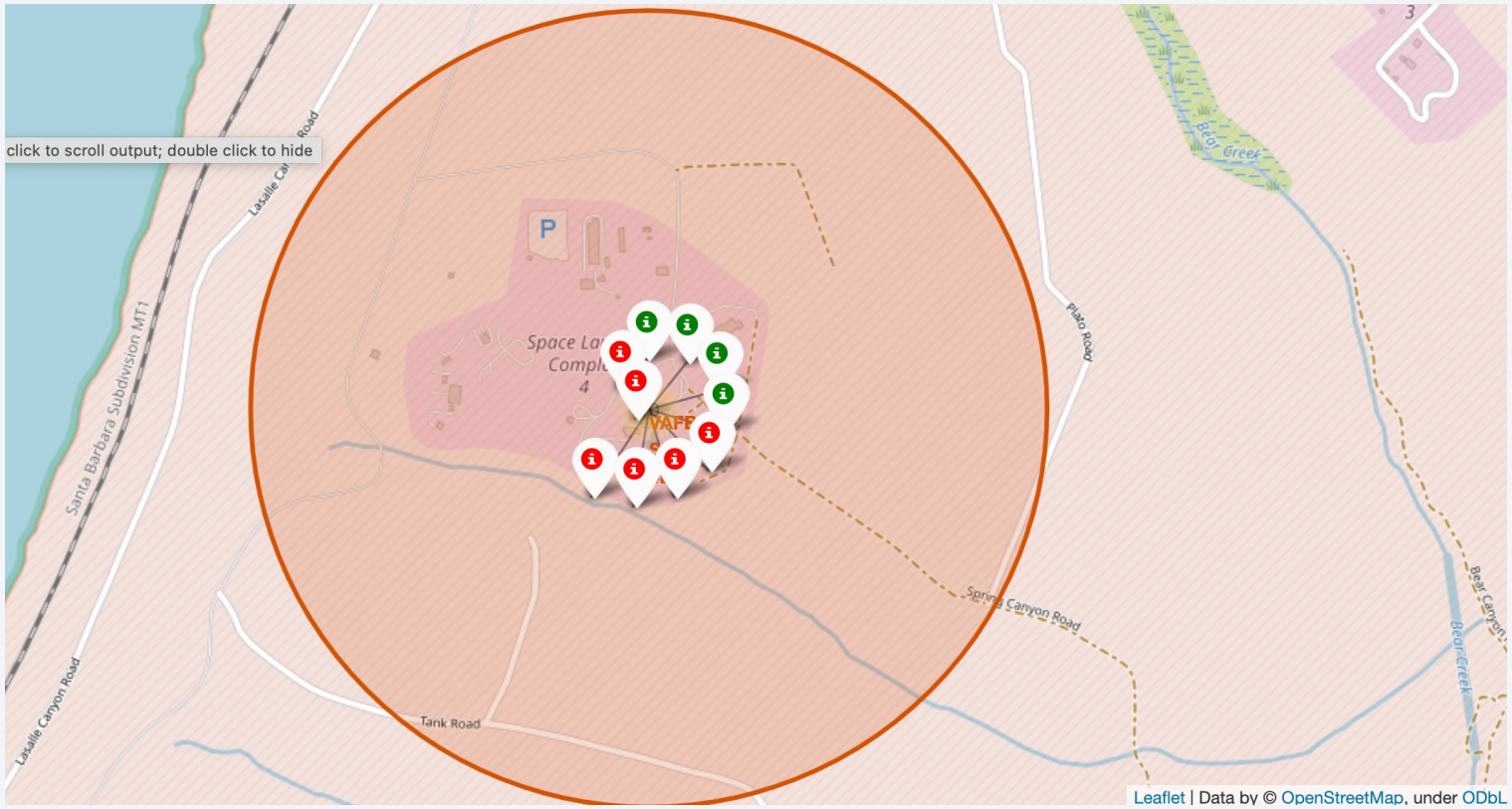
Launch Sites Proximities Analysis

All launch sites' location markers on a global map



- Presumably, since launching of rockets carry great risks for their immediate contexts, all launching sites are close to the ocean in order to minimize any such potential risks.

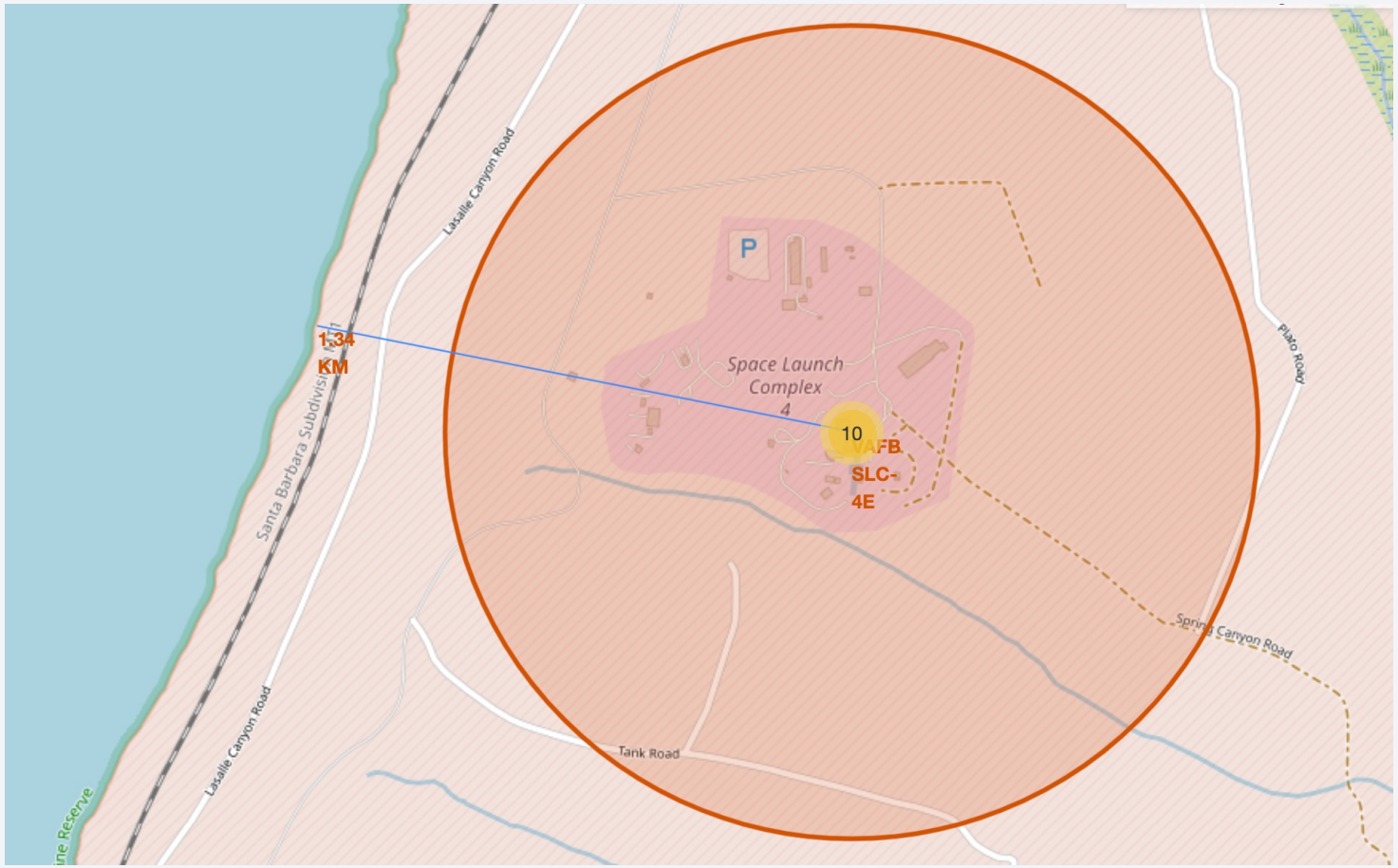
Color-labeled launch records on the map



Green Marker = Successful Launch Red Marker = Failed Launch

Launch Site VAFB SLC-4E has a mixed Success Rate.

Distance from the launch site VAFB SLC-4E to the ocean

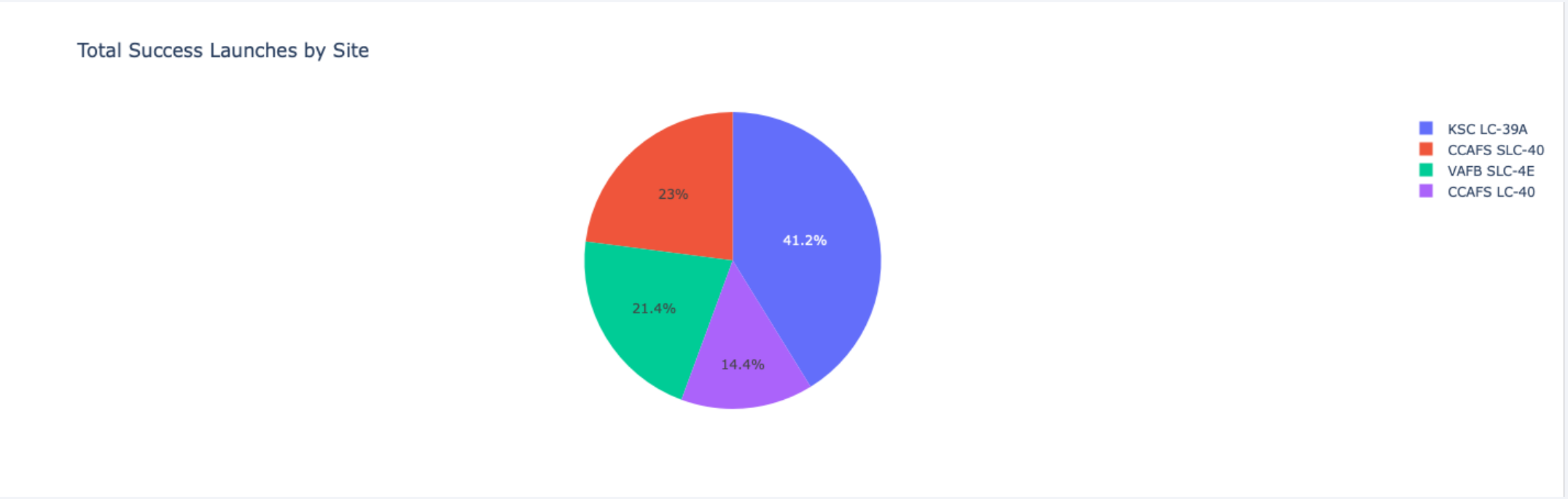


The launch site VAFB SLC-4E is very close to the ocean (1.34.km).

Section 5

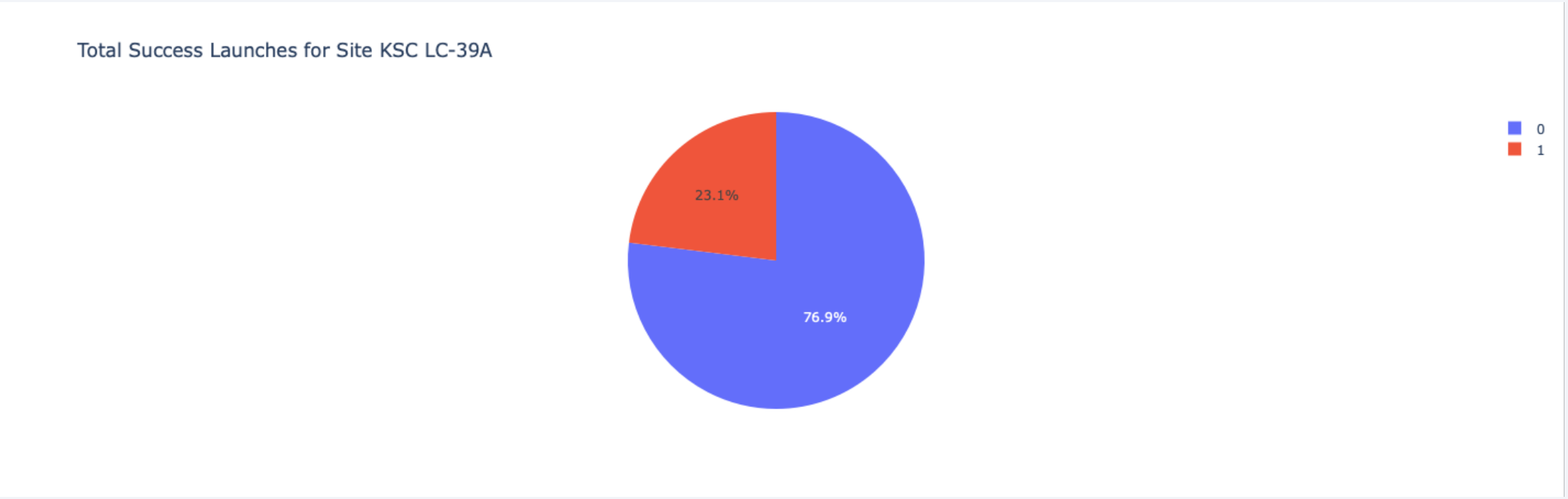
Build a Dashboard with Plotly Dash

Launch success count for all sites



KSC LC-39A has the most successful launches.

Launch site with highest launch success rate



With 10 successful landings out of a total 13, launch site KSC LC-39A has the highest success rate, i.e., 76.9%.

Payload Mass vs. Launch Outcome for all sites

Payloads between 2000 and 5500 kg have the highest success rate.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

Despite trying different metrics, scores on the test set gave no significant information on which model is the best as, for all types of scores, the models gave the same results; this invariance might be due to the test sample size, i.e., 18, which is relatively small.

	training accuracy scores	test accuracy scores	test jaccard scores	test f1 scores
Logistic Reg	84.821429	83.333333	80.0	81.481481
SVM	86.250000	83.333333	80.0	81.481481
DTree	90.000000	83.333333	80.0	81.481481
KNN	84.821429	83.333333	80.0	81.481481

When we take into account the training accuracy, the Decision Tree Model has the highest score and hence the best model.

Conclusions

- Decision Tree Model is the best algorithm, albeit not significantly.
- In general, launches with relatively lower payload mass have higher success rate.
- The success rate of launches increases over the years, especially since 2013.
- Launch site KSC LC-39A has the highest success rate, i.e., 76.9%.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- All launch sites are in very close proximity to the coast.

Thank you!