IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anh Tu Chau
05/01/23

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization and SQL
  - Built an interactive map and dashboard
  - Predictive analysis
- Summary of all results
  - Launch success keeps increasing with every new launch
  - Launch Sites are located near the Equator line, near the coast, and relatively far from the population.
  - Site KSC LC-39A has the highest number of successful launches with 41.7% of them
  - Decision tree is the best method to predict success or failure with almost 0.9 accuracy

Introduction

- Project background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, costing 62 million dollars; other providers cost upwards of 165 million dollars each, and 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- Problems you want to find answers
  - Can we predict the success of the first stage landing?
  - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
  - Which binary classification method to use for the best results?

Section 1

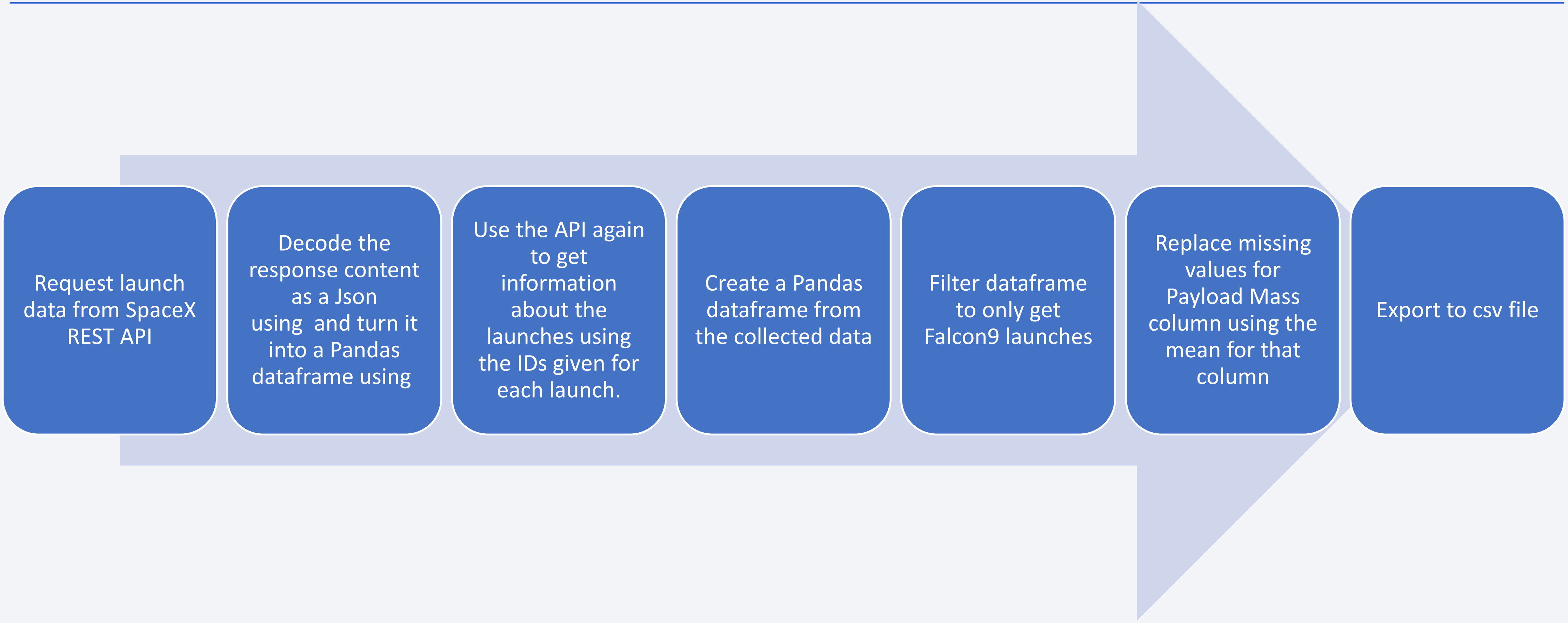
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Gather launch data from SpaceX REST API
  - Web scrapping related Wikipedia pages
- Perform data wrangling
  - Filter data
  - Replace missing values
- Perform exploratory data analysis (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 – SpaceX API



[GitHub URL for data collection – SpaceX API]

Data Collection - Scraping

- Perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response
- Create a BeautifulSoup object from the HTML
- **Extract all column/variable names from the HTML tables**
- **Create a data frame by parsing the launch HTML tables**
- Export to CSV file

[Git Hub URL for data collection - scraping]

Data Wrangling

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome per orbit type

Create a landing outcome label from Outcome column (If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully)

Git Hub URL for data wrangling

EDA with Data Visualization

- Scatter Plot: Show the relationship between variables
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
- Bar Chart: Show the relationship between categories
  - Success Rate vs Relationship
- Line Chart: Show progress over time
  - Average success rate over the years

[Git Hub URL for EDA with Data Visualization]

EDA with SQL

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

[Git Hub URL for EDA with SQL]()

Build an Interactive Map with Folium

- Circles and markers for each launch site:
  - Circle based on the launch site longitude and latitude, and a popup with the name of the launch site
  - Markers for each launch site based on the launch site's longitude and latitude with the name of the site
- Mark cluster with success/failed launch for each site on the map:
  - Green marker for a successful launch
  - Red marker for a failed launch
- Colored line from a launch site to nearby points of interest:
  - Colored line to closest railway, highway, coastline, and city with the distance included

[Git Hub URL for Interactive Map with Folium]

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown list: Allows us to choose all the sites or a specific site to show on the different visuals.
- Launch Success Pie Chart: Displays total successful launches if All sites or selected and displays success vs failed launches if a specific site is selected.
- Payload Mass range slider: Allows us to select a payload mass range.
- Scatter Chart of Payload Mass vs. Success Rate: Displays the correlation between the payload mass vs the success rate of launches for different booster version categories.

Git Hub URL for Plotly Dash

Predictive Analysis (Classification)

- Create a Numpy array from the column “Class” in our data
- Standardize the data with StandardScaler, and then fit and transform it.
- Use the function train_test_split to split the data X and Y into training and test data
- Create a GridSearchCV object with cv = 10 for the different classification methods
- Fit the different methods to find the best parameters
- Calculate accuracy for all methods using .score()
- Examine the confusion matrix
- Find the method that performs best by examining the accurcy

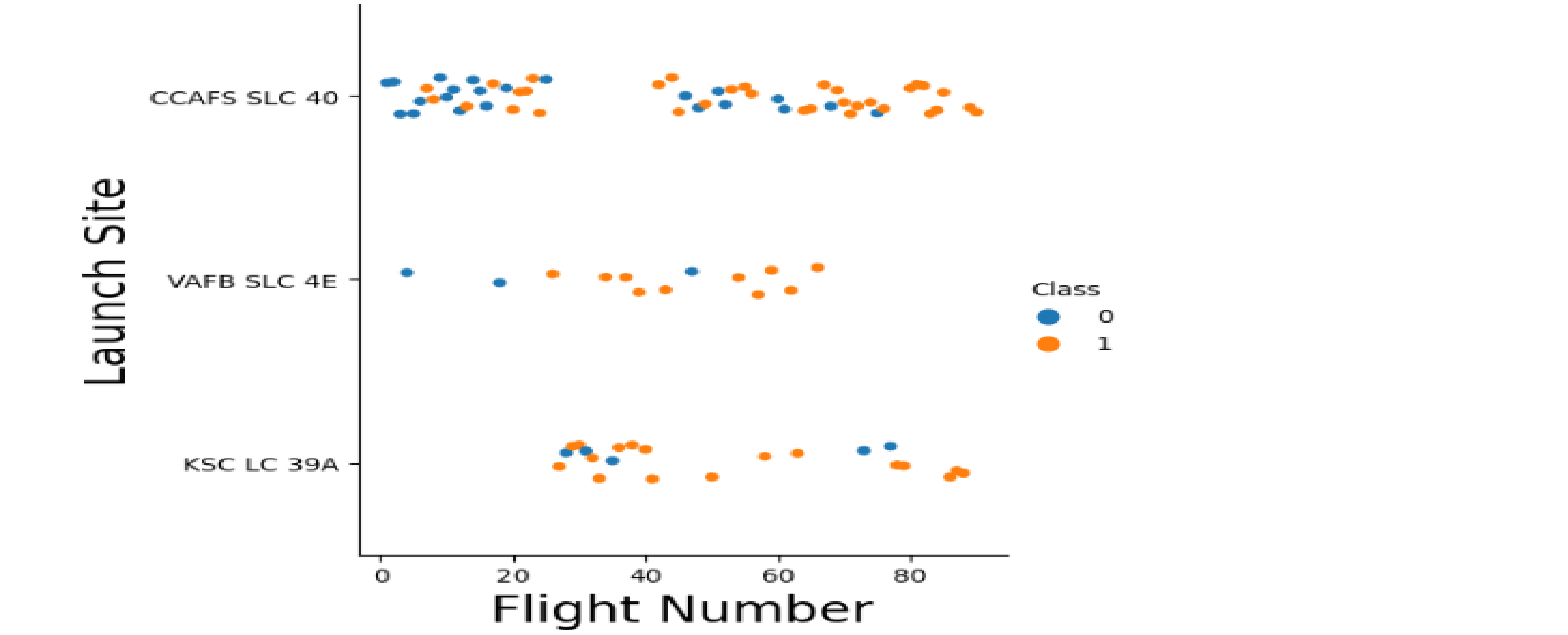
Results

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

Section 2

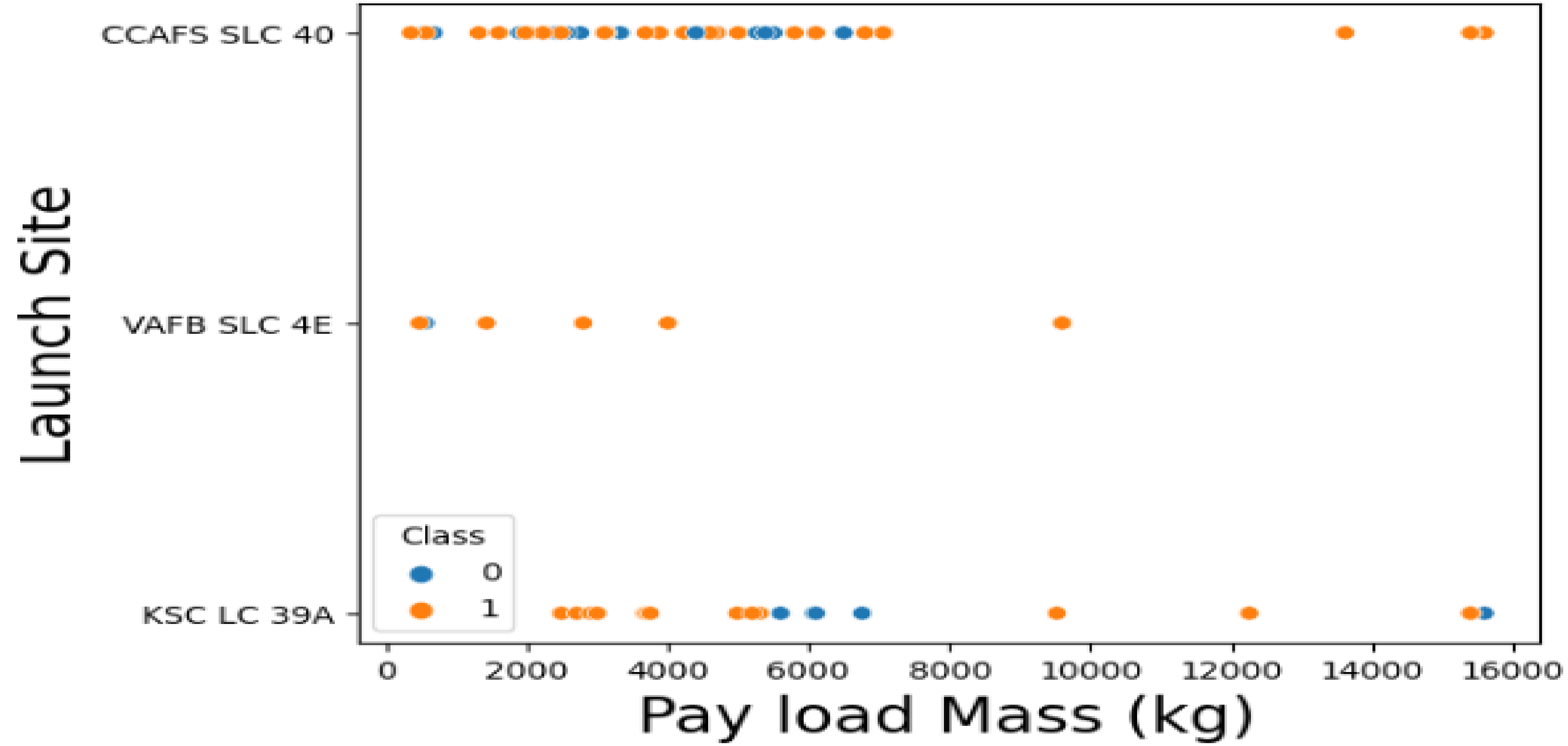
Insights drawn from EDA

Flight Number vs. Launch Site



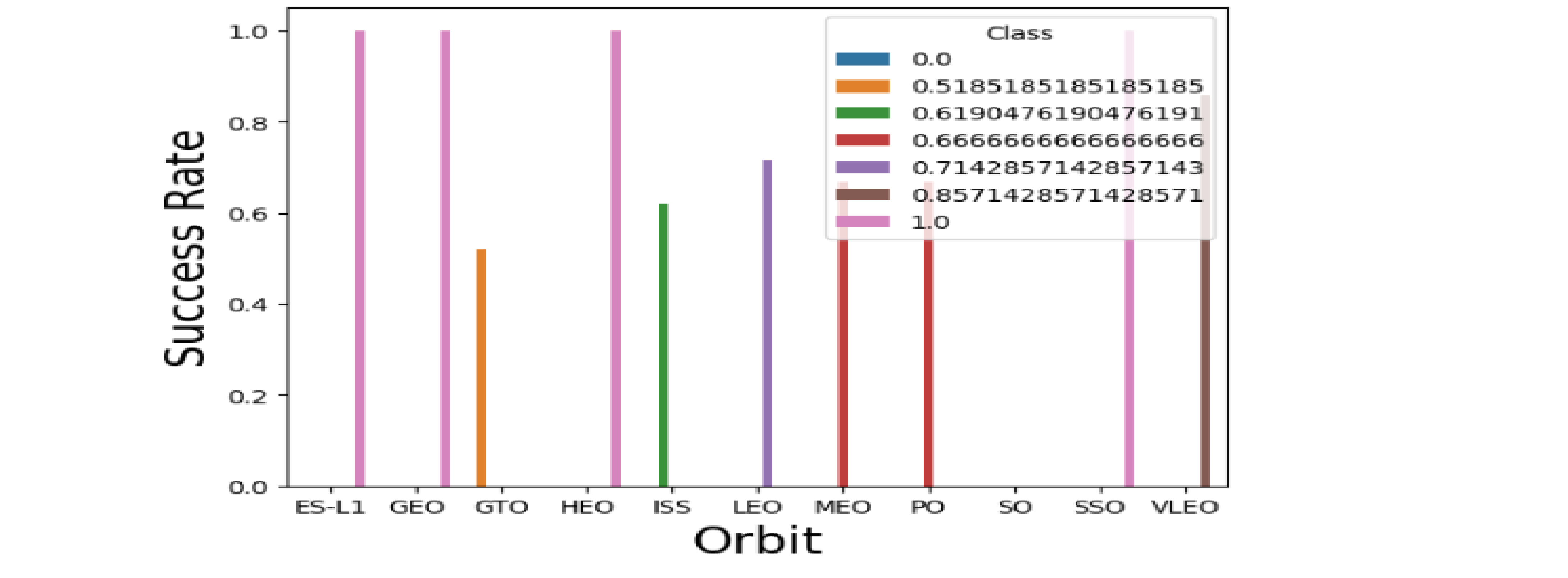
- The newer launches have a higher success rate than the older ones
- The success rate for CCAFS SLC 40 is lower because it has a lot of the “earlier” launches

Payload vs. Launch Site



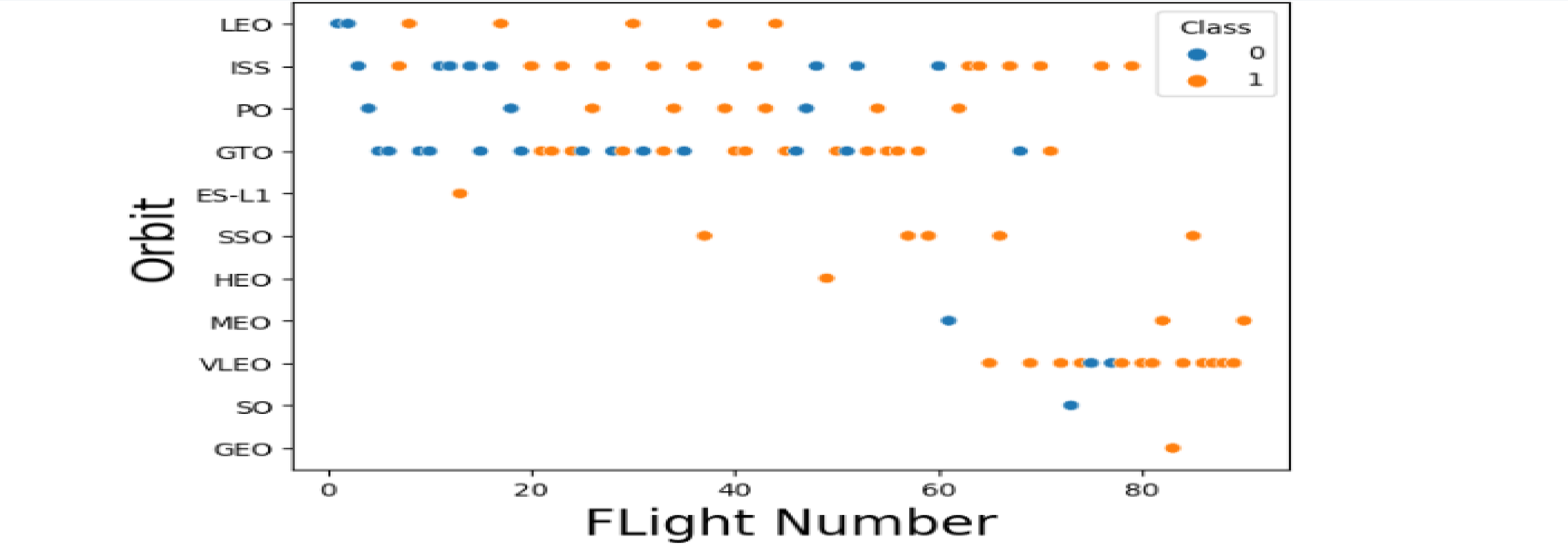
- Almost all the launches with heavy loads (> 8000 kg) were successful.
- KSC LC 39A has a lot of success with loads under 4000 kg

Success Rate vs. Orbit Type



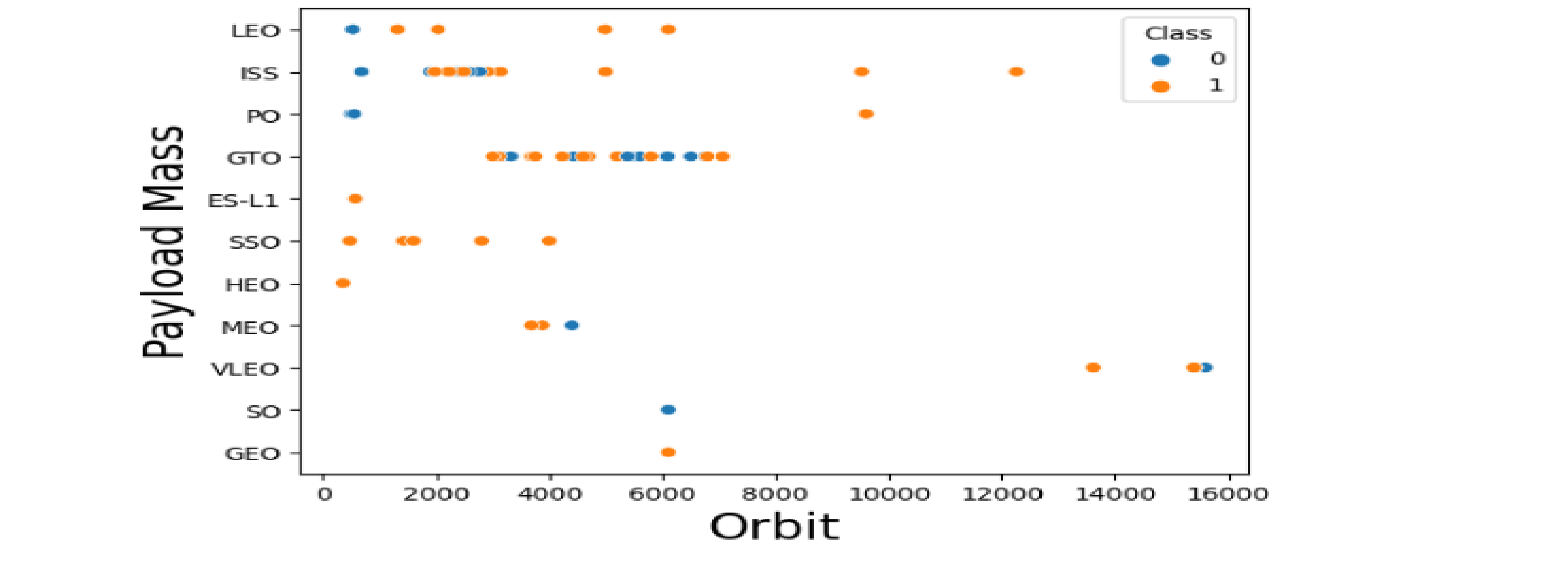
- Orbits of type ES-L1, GEO, HEO, and SSO have a 100% success rate.
- Orbits of type GTO, ISS, LEO, MEO, PO, and VLEO have a success rate between 50% and 85%.
- Orbit of type SO has a 0% success rate

Flight Number vs. Orbit Type



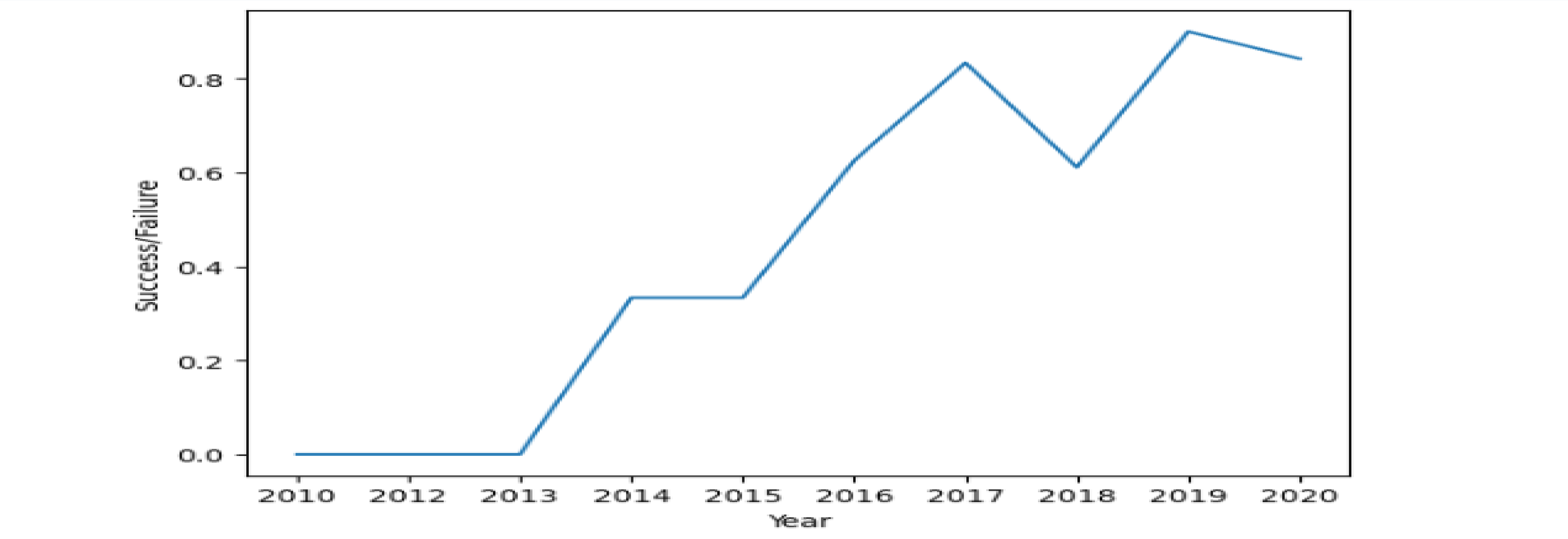
- There does not seem to be any correlation between the flight number and the orbit type

Payload vs. Orbit Type



- On orbit LEO, payload masses over 2000 kg have more success
- On orbit ISS, payload masses over 4000 kg have more success

Launch Success Yearly Trend



- The success rate has kept increasing since 2013

All Launch Site Names

```
%sql select distinct LAUNCH_SITE FROM SPACEX
```

```
 * ibm_db_sa://fnf27284:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/BLUDB
Done.
```

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

- The different launch sites

Launch Site Names Begin with 'CCA'

```
%sql select * from spacex where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://fnf27284:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/BLUDB
Done.
```

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- 5 records where launch sites begin with `CCA`

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as total_mass from spacex where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://fnf27284:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.datat
Done.
```

total_mass
45596

- Total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

```
%sql select sum(payload_mass__kg_) as avg_mass from spacex where booster_version = 'F9 v1.1'

 * ibm_db_sa://fnf27284:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databas
Done.
```

avg_mass
14642

- Average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql select min(date) as date from spacex where landing__outcome like 'Success%'
```

```
 * ibm_db_sa://fnf27284:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde
Done.
```

DATE
2015-12-22

- Dates of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select payload from spacex where payload_mass__kg_ between 4000 and 6000
```

```
 * ibm_db_sa://fnf27284:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.clogj3sd0tgtu0
Done.
```

payload
AsiaSat 8
AsiaSat 6
ABS-3A Eutelsat 115 West B
Turkmen 52 / MonacoSAT
SES-9
JCSAT-14
JCSAT-16
EchoStar 23
SES-10
NROL-76
Boeing X-37B OTV-5
SES-11 / EchoStar 105
Zuma
GovSat-1 / SES-16
SES-12
Merah Putih
Es hail 2
SSO-A
GPS III-01
Nusantara Satu, Beresheet Moon lander, S5
RADARSAT Constellation, SpaceX CRS-18
GPS III-03, ANASIS-II
ANASIS-II, Starlink 9 v1.0
GPS III-04 , Crew-1

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
 * ibm_db_sa://fnf27284:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/BLUDB
Done.
```

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

- Number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql select distinct booster_version from spacex where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex)
```

```
 * ibm_db_sa://fnf27284:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/BL
Done.
```

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

- Names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%sql select landing__outcome, booster_version, launch_site from spacex where date like '2015%' and landing__outcome like 'Failure%'
```

```
 * ibm_db_sa://fnf27284:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/BLUDB
Done.
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- 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

```
%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://fnf27284:***@6667d8e9-9d4d-4ccb-ba32-21da3bb5aafc.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30376/BLUDB
Done.
```

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

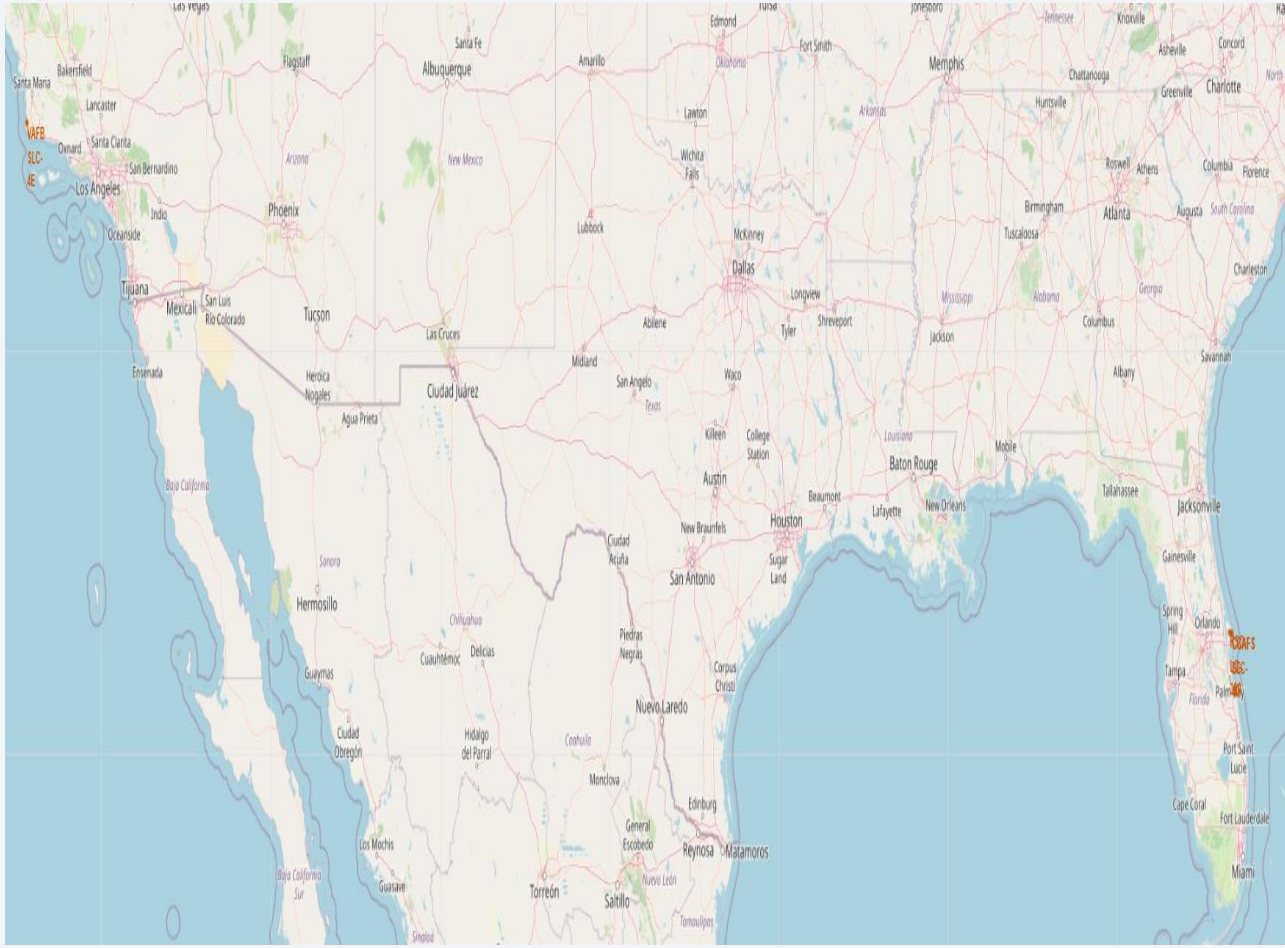
- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

Launch Sites Proximities Analysis

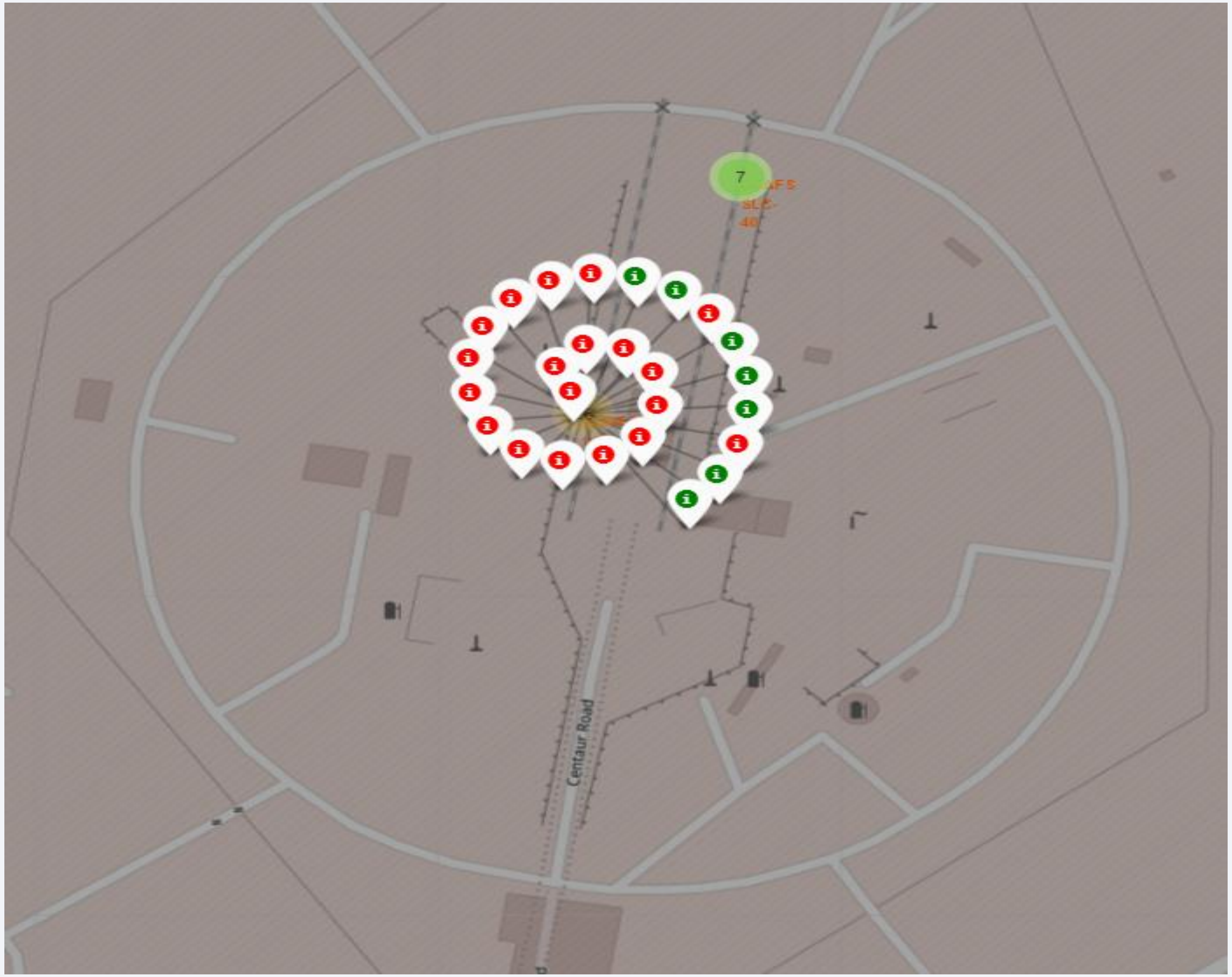
Launch Sites Location on the map

- All the launch sites are located near the Equator line.
  - This is to take advantage of the Earth's substantial rotational speed.
  - Earth's rotation relative to its center is at its highest near the Equator.
  - equator-launched vehicles need less propellant
- All the launch sites are near the coast
  - Launching rockets toward the ocean is better for people's safety



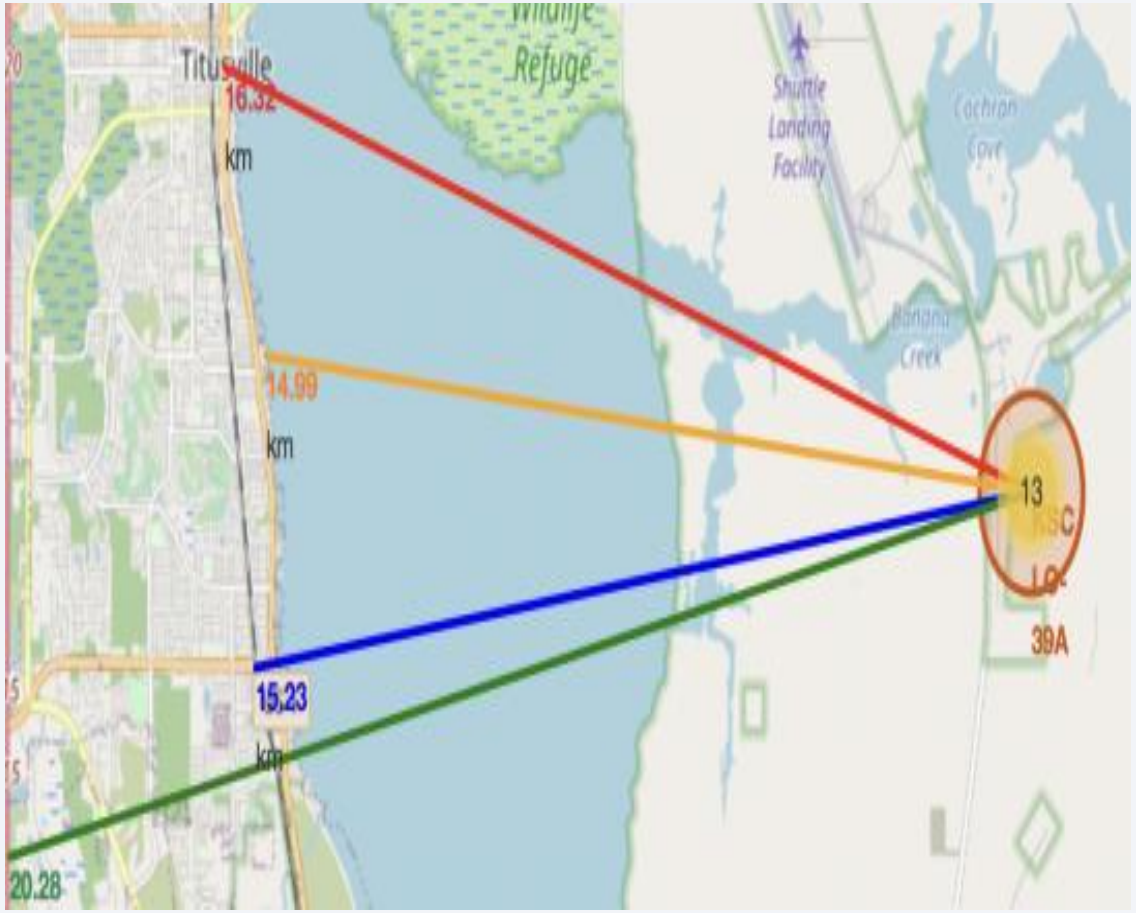
Colored Marker Cluster of Success and Failures

- Clusters of colored markers for each launch site so we can easily the success rate for each site.
- A green marker means success and a red marker means failure.
- As we can see in the image, the success rate of the site CCAFS LC-40 is bad since there are a lot more red markers than green markers.



Proximity to nearby points of interest

- As we can see in the image on the right, the launch sites are relatively far from any highway, railway, coastline, or city.
- They are over 15 km in distance from the launch site.
- This is so the Rocket won't fly over populated areas in case of an accident.

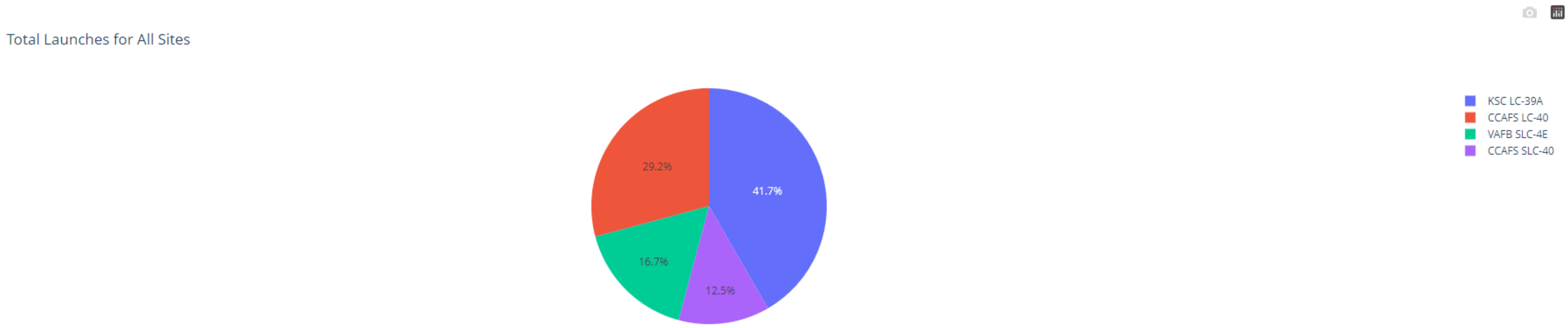


Section 4

Build a Dashboard with Plotly Dash

Launch Success for all sites

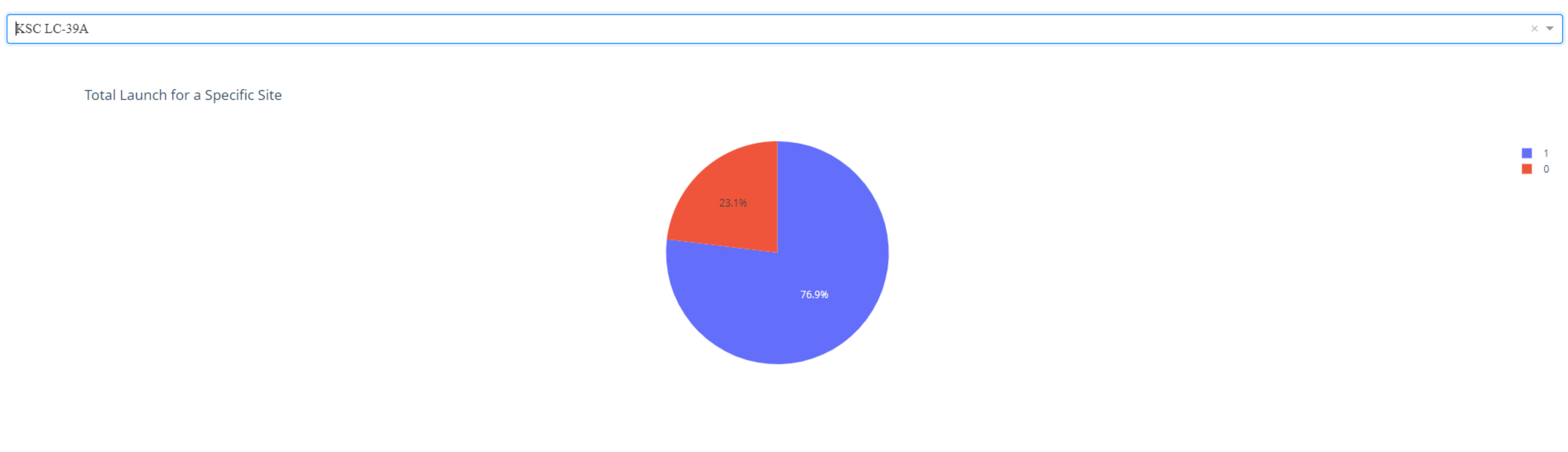
ALL SITES



Payload range (Kg):

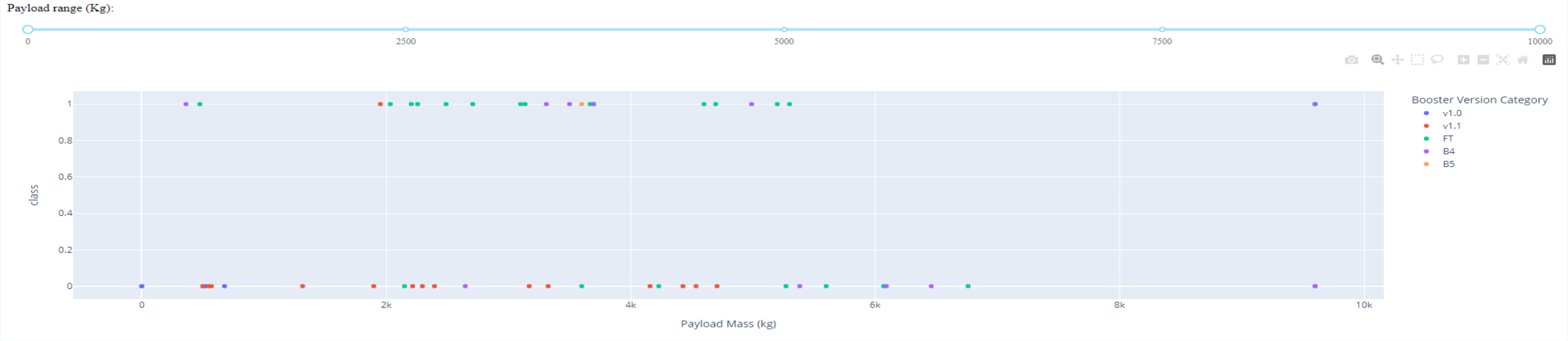
- We can see that site KSC LC-39A has the highest number of successful launches with 41.7% of them
- Site CCAFS SKC-40 has the lowest number with 12.5%

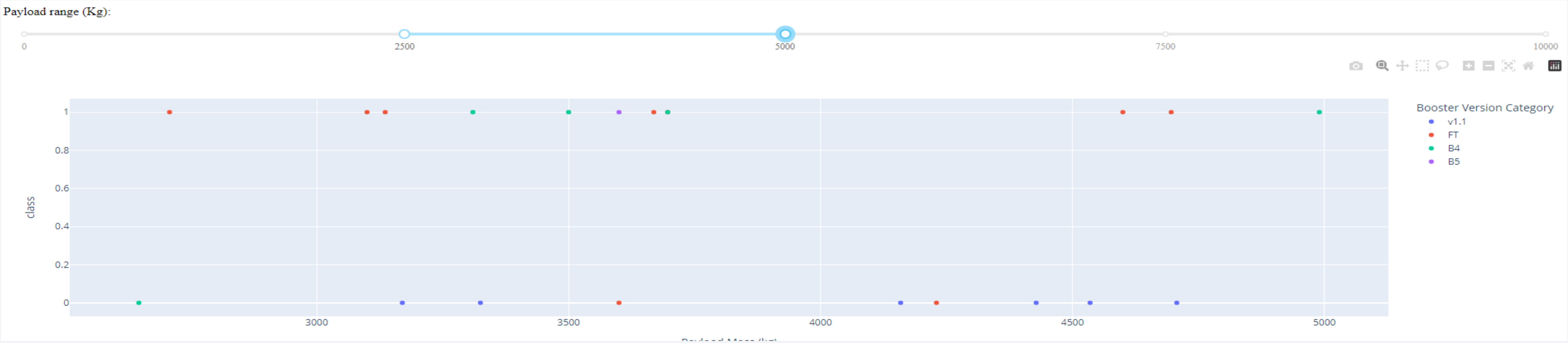
Launch Site with Highest Success Ratio



- Site KSC LC-39A has the highest success ratio with 76.9% success rate

Payload vs Launch Outcome Scatter Plot





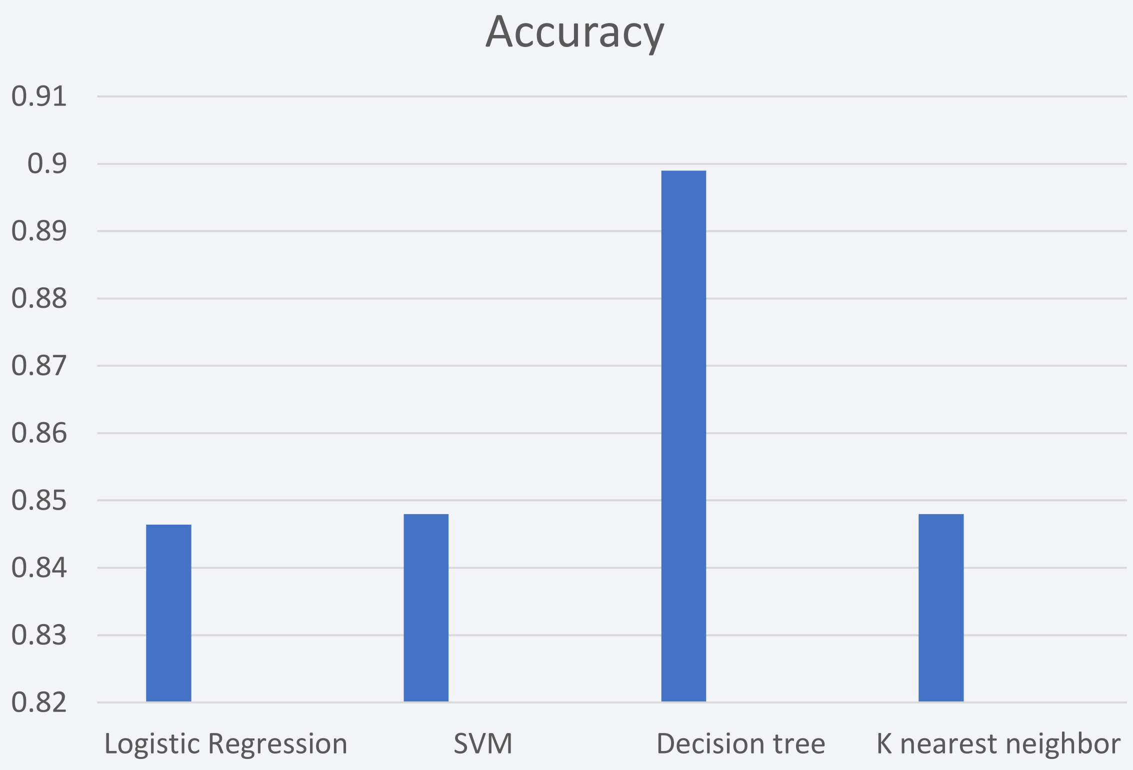
- We can see on the image that for all payload mass the success rate is around even
- The range with the highest success rate is between 2500kg and 5000kg

Section 5

Predictive Analysis (Classification)

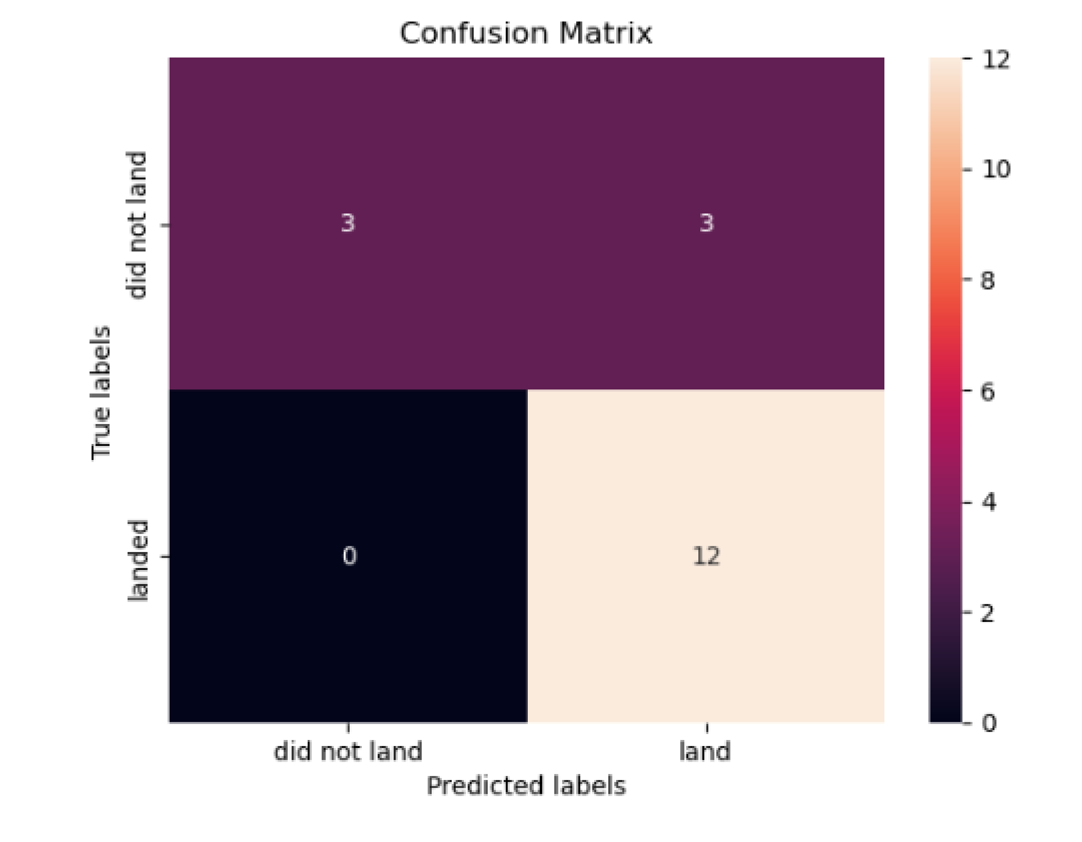
Classification Accuracy

- On the bar chart on the right, we can see the accuracy of the different classification methods.
- We can see that the decision tree is the most accurate with an accuracy of nearly 0.9
- The other methods are all sitting around 8.5



Confusion Matrix

- We can see from the confusion matrix for the decision tree method that almost every case is getting classified correctly.
- There are only three cases that were misclassified. They have been falsely labeled as successful.



Conclusions

- Launch success keeps increasing with every new launch
- Orbits of type ES-L1, GEO, HEO, and SSO have a 100% success rate.
- Launch Sites are located near the Equator line, near the coast, and relatively far from the population.
- Site KSC LC-39A has the highest number of successful launches with 41.7% of them
- Decision tree is the best method to predict success or failure with almost 0.9 accuracy

Thank you!