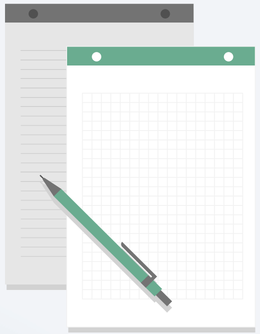
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sabri Ben Ayed
November 2021

Outline

Executive Summary



- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a dashboard with Ploty Dash
  - Predictive analysis

- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics with screenshots
  - Predictive analysis results

[GitHub URL to notebook]

Introduction

- Project background and context
  - In this capstone, we will predict if the Falcon 9 first stage will land successfully. 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch
- Problems you want to find answers
  - How do some characteristics of the launcher such as payload mass, number of flights and orbits affect the success of the first stage landing ?
  - Does the success rate increase over the years ?
  - What are the conditions that SpaceX needs to have to get the best results and ensure a successful landing ?

Section 1

Methodology

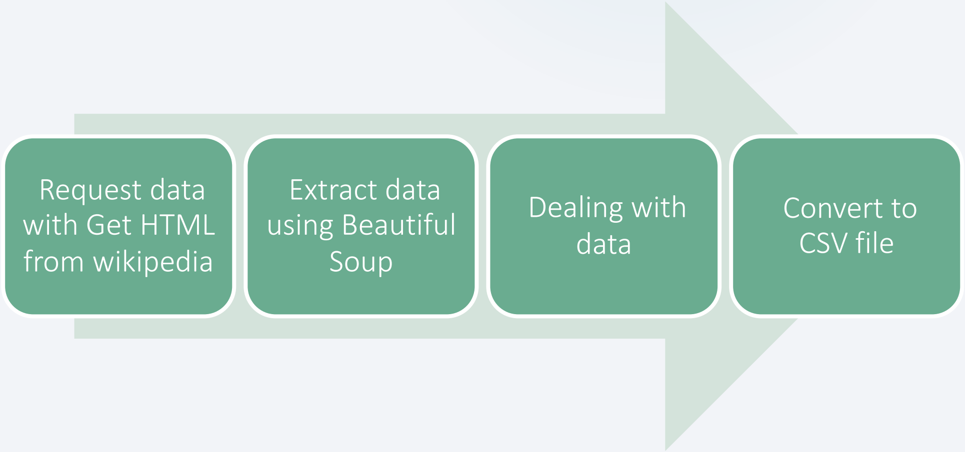
Methodology

Executive Summary

- Data collection methodology:
  - Data is collected from the SPACEX REST API and using Web scrapping from Wikipedia
- Perform data wrangling
  - Using pandas and numpy libraries we will explore the data and determine what would be the label for training supervised models
- 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

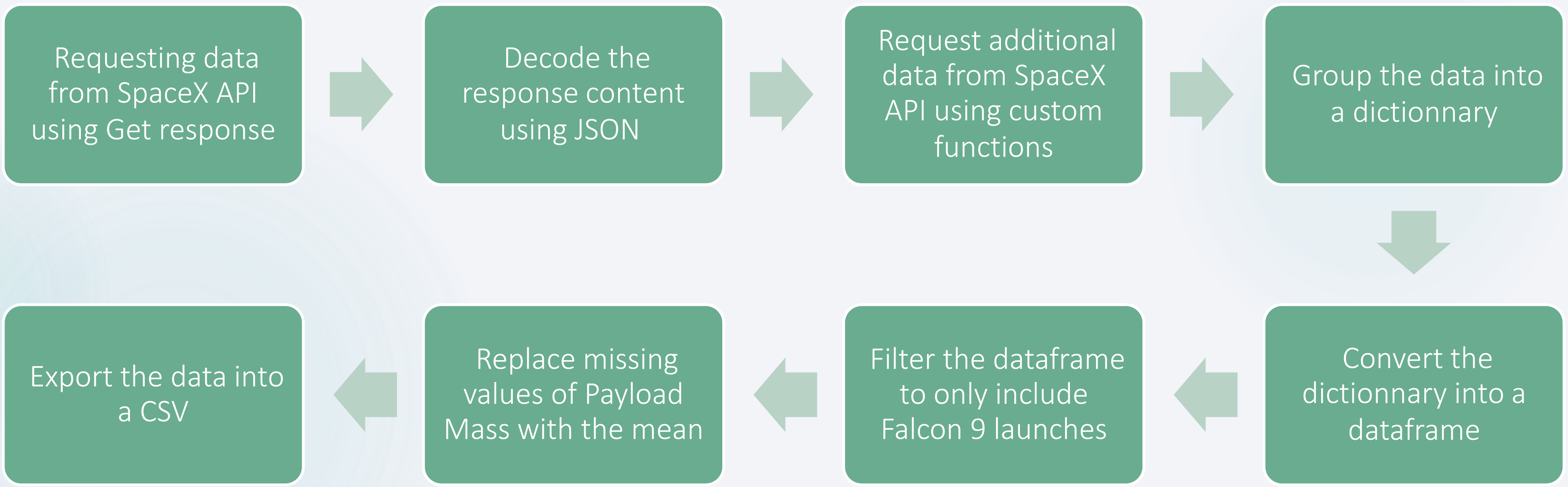
Data Collection

- Data was collected from multiple source to build the dataset
  - Data collected from SpaceX REST API:
    - https://api.spacexdata.com/v4/launches/past
    - `This link provides data about previous launches such as rocket type, payload mass, dates, success/failure`
  - Data collected from Wikipedia using web scrapping and Beautiful Soup



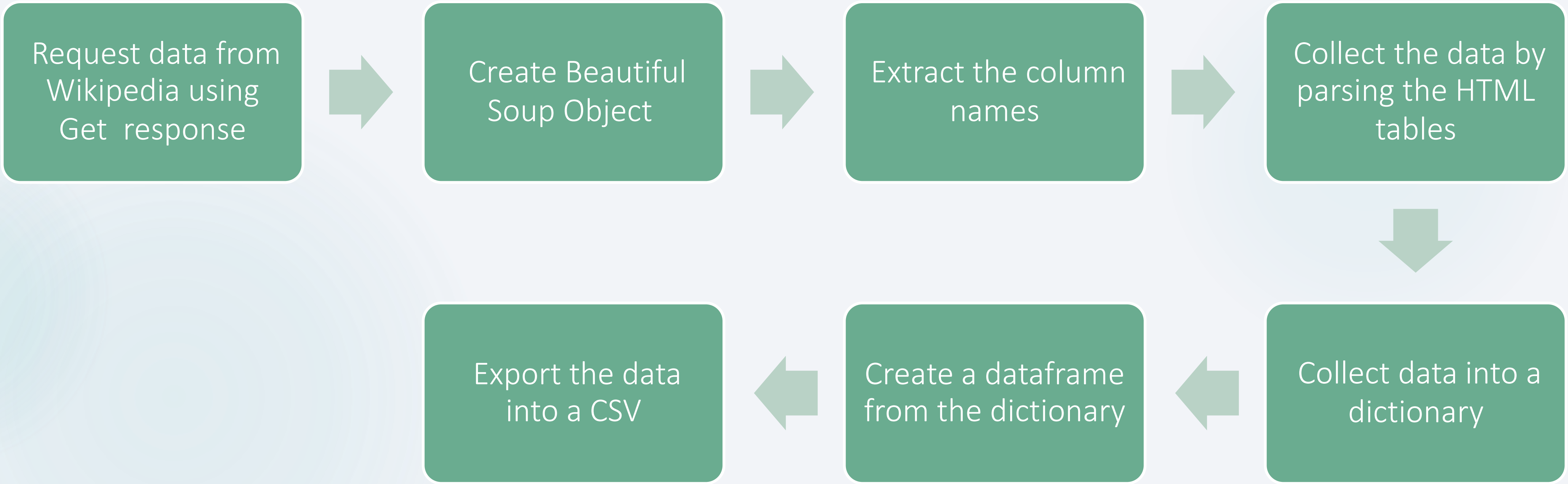
GitHub URL to notebook

Data Collection – SpaceX API



[GitHub URL to Notebook]

Data Collection - Scraping

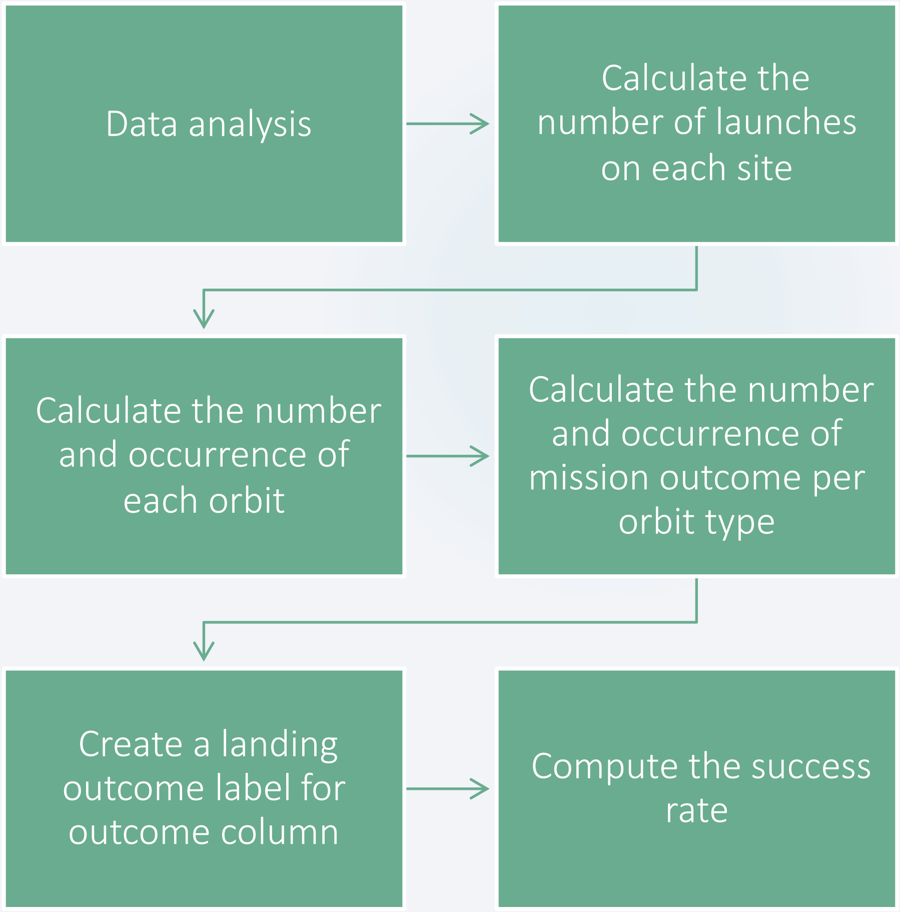


[GitHub URL to Notebook]

Data Wrangling

- We performed several checks on the data such as the percentage of missing values in each attribute and identify the data type
- Using the method value_counts() we computed:
  - Number of launches on each site
  - Number of missions
  - Number of Outcomes
- We created a landing outcome label from Outcome column having 1 as Success or 0 as Failure
- We computed the success rate

[GitHub URL to Notebook]



EDA with Data Visualization

- Using the library seaborn we plotted the following charts Flight number vs Payload Mass, Flight number vs Launch Site, Payload Mass vs Launch Site, Flight number vs Orbit type, Payload vs Orbit type
- Bar charts were also used to plot the success rate of the different orbit type
- We built a feature matrix of the most important attributes

GitHub URL to Notebook

EDA with SQL

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

[GitHub URL to Notebook]

Build an Interactive Map with Folium

- Mark all launch sites on a map
  - We created a folium `Map` object, with an initial center location to be NASA Johnson Space Center at Houston, Texas
  - We created a Circle for all Launch Sites using their latitude and longitude coordinates
- Mark the success (Green)/failed (Red) launches for each site on the map
- Calculate the distances between a launch site to its proximities to see if there was a relationship between success rate and distance to some objects such as cities or coastline

GitHub URL to Notebook

Build a Dashboard with Plotly Dash

The dashboard includes

- Pie chart
  - Display the total successful launches count for all sites and the success vs. Failed counts for the selected site
- Scatter plot
  - Shows the correlation between Payload and launches Success
  - Added a slider to select the payload range

[GitHub URL to Notebook]

Predictive Analysis (Classification)

Building the model

- Load dataset from csv to Pandas dataframe
- Transform the data
- Split the data into training and test data set
- Run the following ML algorithms: Logistic Regression, SVM, Decision Tree, KNN
- Train each model using GridSearchCV to optimize hyperparameters

Evaluating the models

- Check accuracy of the models
- Plot Confusion matrix

The model with the best accuracy score wins

GitHub URL to Notebook

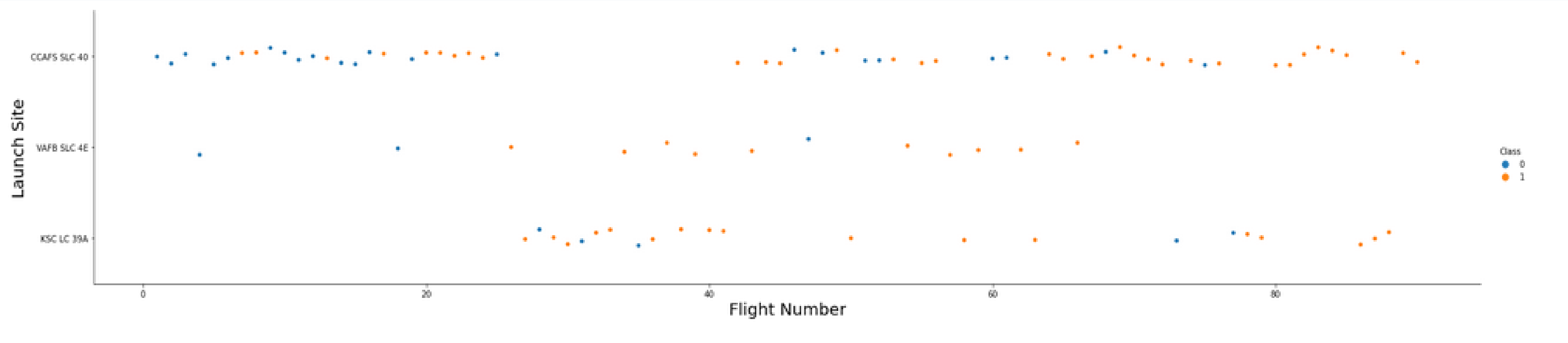
Results

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

Section 2

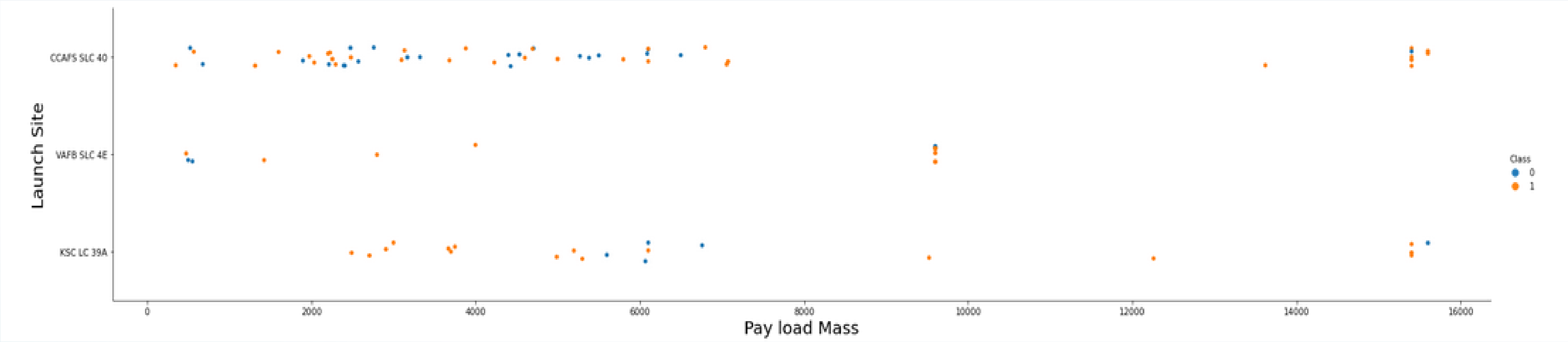
Insights drawn from EDA

Flight Number vs. Launch Site



- KSC LC 39A and VAFB SLC 4E have higher success rates
- Latest flights had a higher rate of success

Payload vs. Launch Site

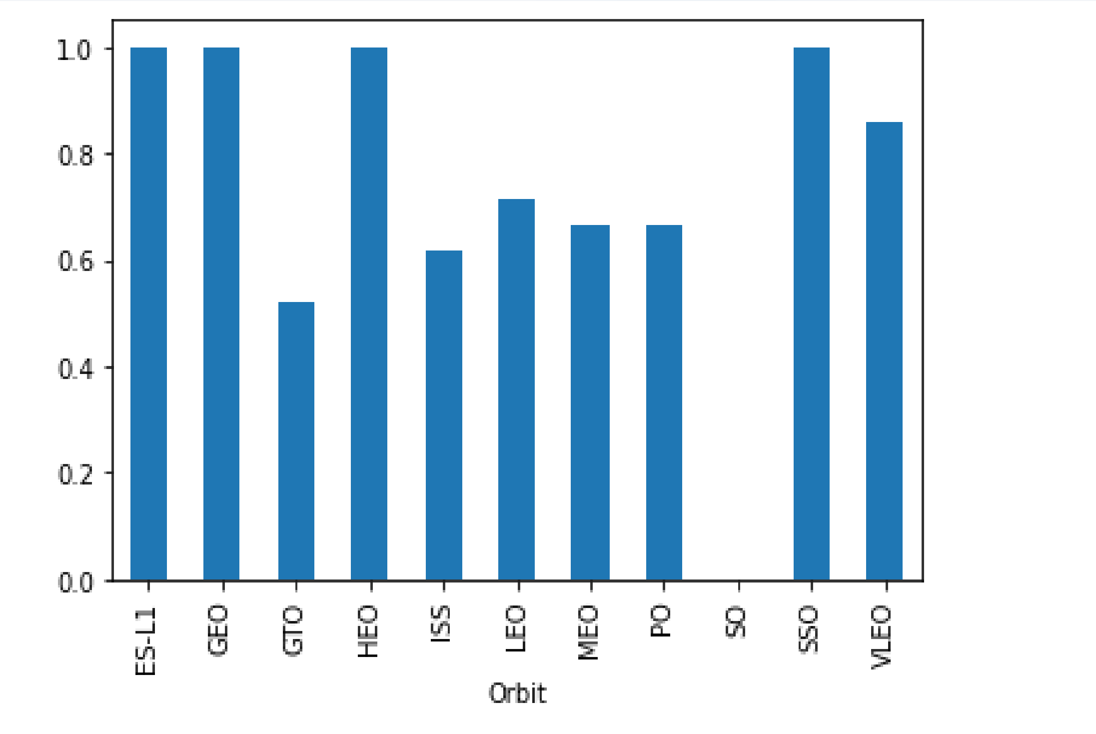


- Most of the launches with payload mass above 8000kg were successful
- For every launch site the higher the payload mass, the higher the success rate

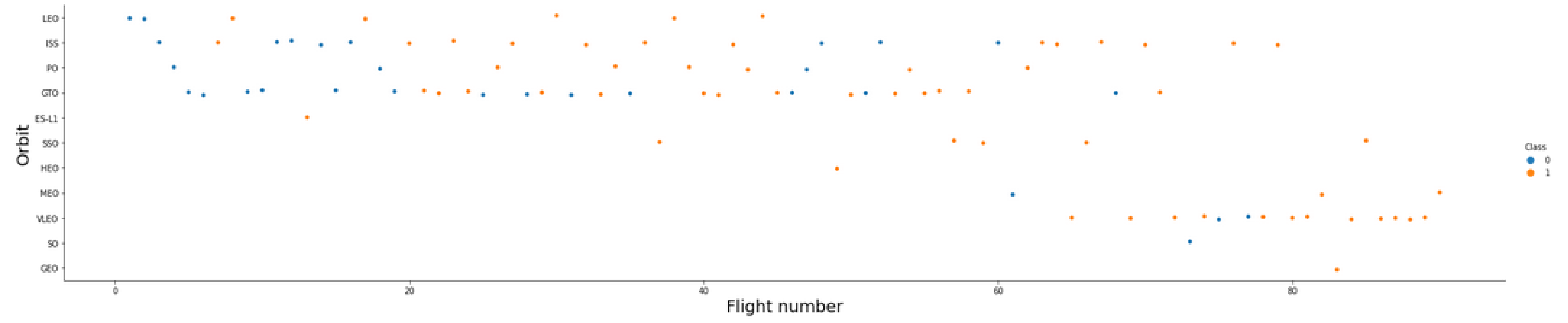
Success Rate vs. Orbit Type

- ES-L1, GEO HEO and SSO are the orbits with a 100% success rate

- SO orbit has a success rate of 0%

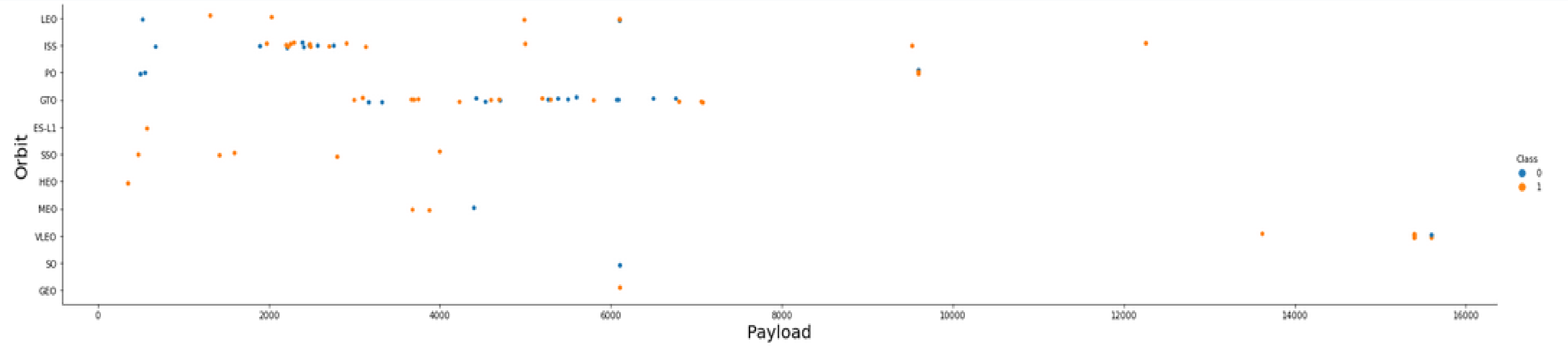


Flight Number vs. Orbit Type



- In the LEO orbit the success appears to be related to the number of flights, on the other hand, there seems to be no relationship between flight number when in GTO orbit

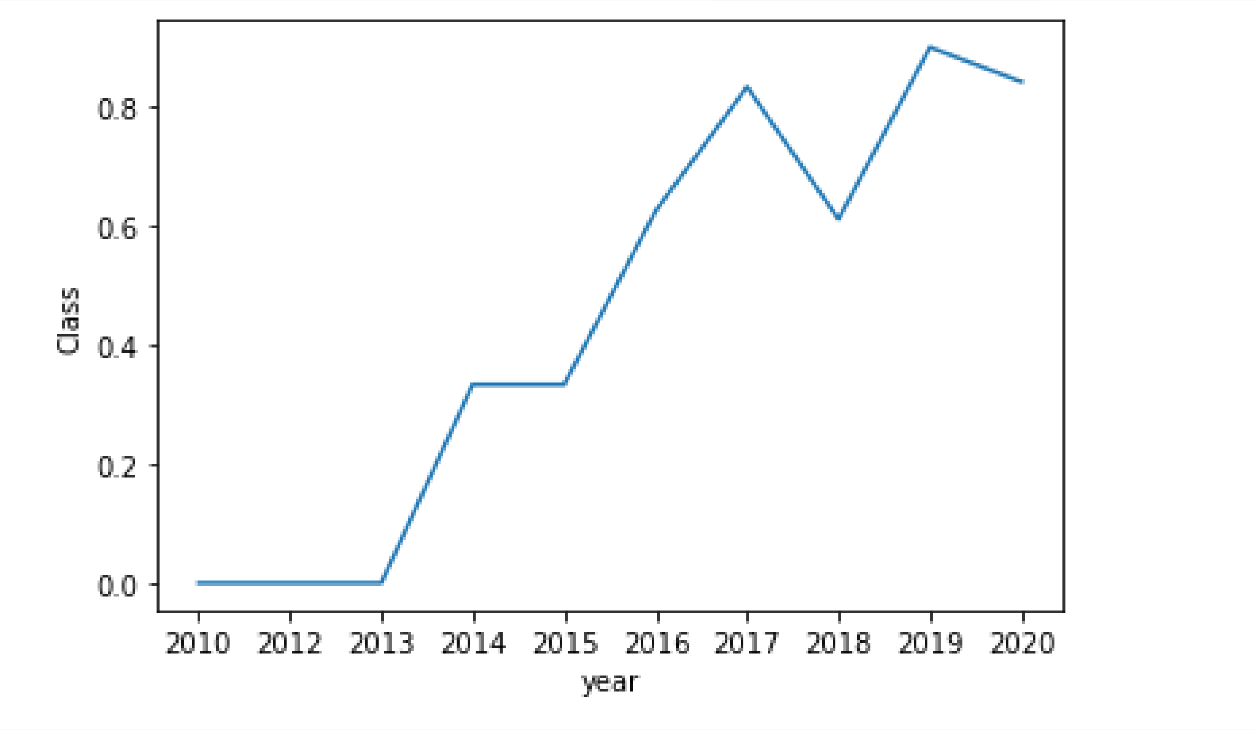
Payload vs. Orbit Type



- LEO, PO and ISS orbits have higher successful landing for heavy payloads

Launch Success Yearly Trend

- The success rate kept increasing till 2020



All Launch Site Names


```
In [5]: %sql SELECT distinct launch_site FROM SPACEXDATASET;

 * ibm_db_sa://hbh90814:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.

Out[5]:
```

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

- Displaying unique launch sites

Launch Site Names Begin with 'CCA'

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

```
In [6]: %sql SELECT * FROM SPACEXDATASET WHERE launch_site LIKE 'CCA%' limit 5;

 * ibm_db_sa://hbh90814:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/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

- Displaying 5 records where launch sites begin with `CCA`

Total Payload Mass


```
In [13]: %sql SELECT customer, sum(payload_mass__kg_) as sum FROM SPACEXDATASET WHERE customer = 'NASA (CRS)'GROUP BY customer;

 * ibm_db_sa://hbh90814:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[13]:

customer	SUM
NASA (CRS)	45596

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

Average Payload Mass by F9 v1.1

Entrée [13]:
```
%sql SELECT BOOSTER_VERSION, AVG(payload_mass__kg_) as average_payload_mass__kg FROM SPACEXDATASET WHERE BOOSTER_VERSION like '%F9 v1.1%' group by BOOSTER_VERSION ;
```

```
 * ibm_db_sa://hbh90814:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[13]:

booster_version	average_payload_mass__kg
F9 v1.1	2928
F9 v1.1 B1003	500
F9 v1.1 B1010	2216
F9 v1.1 B1011	4428
F9 v1.1 B1012	2395
F9 v1.1 B1013	570
F9 v1.1 B1014	4159
F9 v1.1 B1015	1898
F9 v1.1 B1016	4707
F9 v1.1 B1017	553
F9 v1.1 B1018	1952

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

First Successful Ground Landing Date


```
Entrée [14]: %sql SELECT MIN(DATE) as DATE FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (ground pad)';

 * ibm_db_sa://hbh90814:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.

Out[14]:
```

DATE
2015-12-22

- Listing the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000


```
Entrée [21]: %sql SELECT booster_version FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;

 * ibm_db_sa://hbh90814:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.

Out[21]:
```

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

- Listing the names of 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

Entrée [60]:
```
%sql SELECT mission_outcome, count(mission_outcome) as count FROM SPACEXDATASET GROUP BY  mission_outcome;
```

```
 * ibm_db_sa://hbh90814:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[60]:

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

- Listing the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

Entrée [73]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXDATASET WHERE payload_mass__kg_= (SELECT MAX(payload_mass__kg_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://hbh90814:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[73]:

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

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

2015 Launch Records


```
Entrée [67]: %sql SELECT DATE, booster_version, launch_site, landing__outcome FROM SPACEXDATASET WHERE landing__outcome = 'Failure (drone ship)' and YEAR(DATE) = '2015';

 * ibm_db_sa://hbh90814:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[67]:

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Listing the 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


```
Entrée [23]: #%sql SELECT DATE, landing__outcome FROM SPACEXDATASET WHERE  DATE > '2010-06-04' and DATE < '2017-03-20' ;
%sql SELECT LANDING__OUTCOME, COUNT (LANDING__OUTCOME) as count_Landing_Outcome FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04'AND '2017-03-20'  GROUP BY LANDING__OUTCOME;

 * ibm_db_sa://hbh90814:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[23]:

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

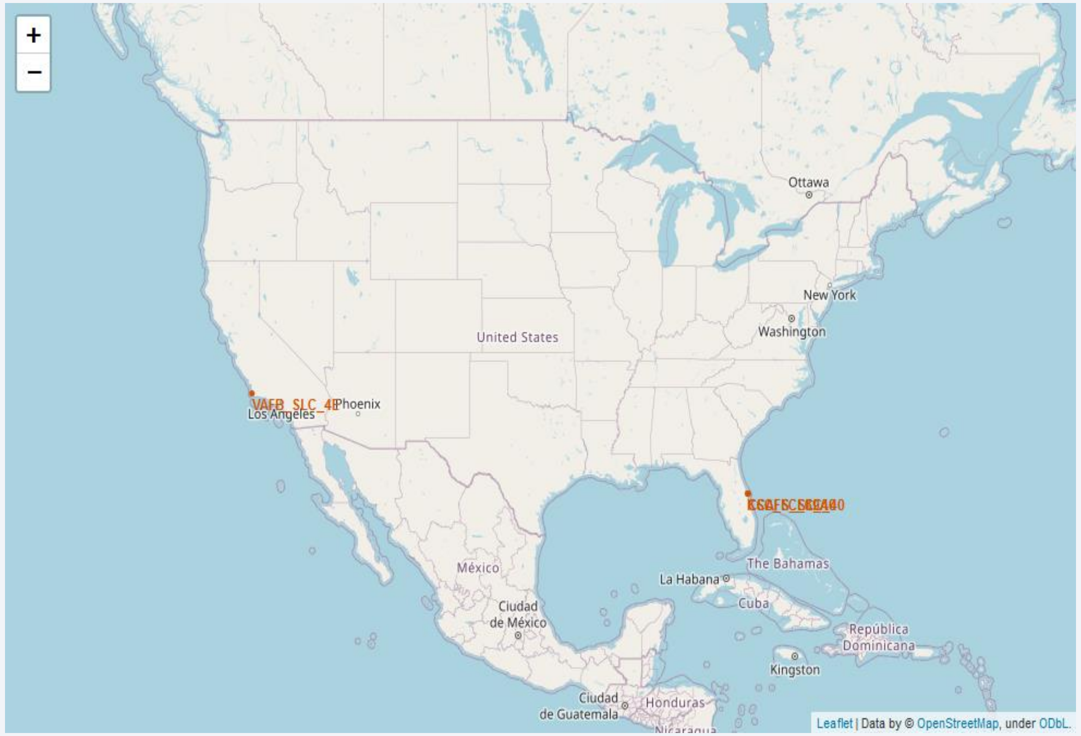
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 4

Launch Sites Proximities Analysis

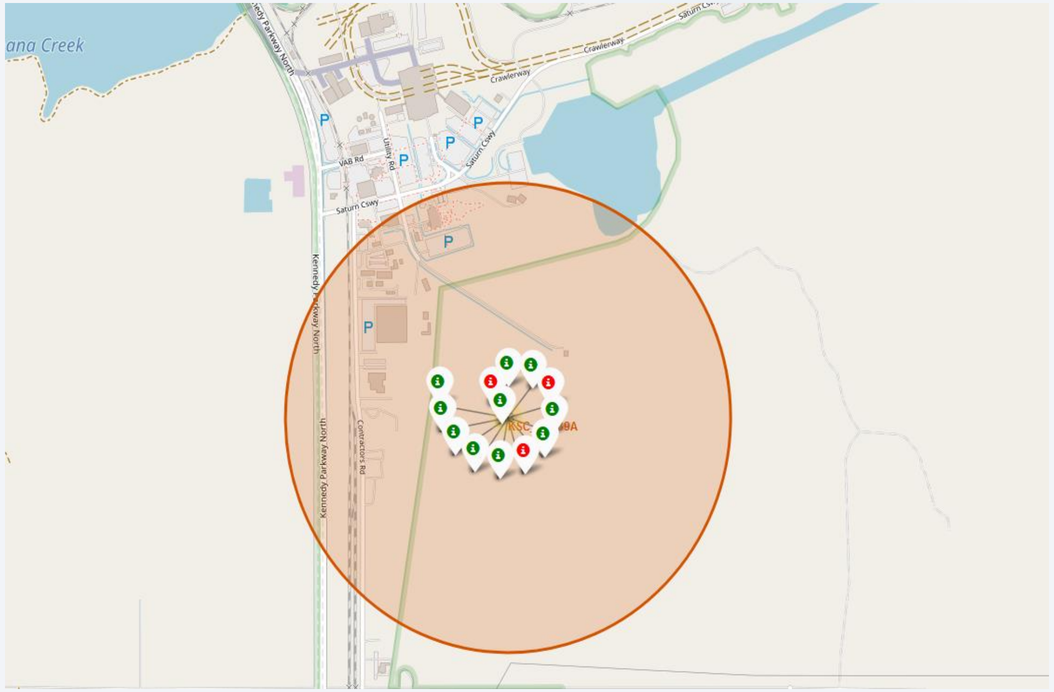
Launch Sites on a map

- All launch sites are in proximity to the Equator line and in close proximity to the coast line as it minimizes the risk of having accidents close to cities



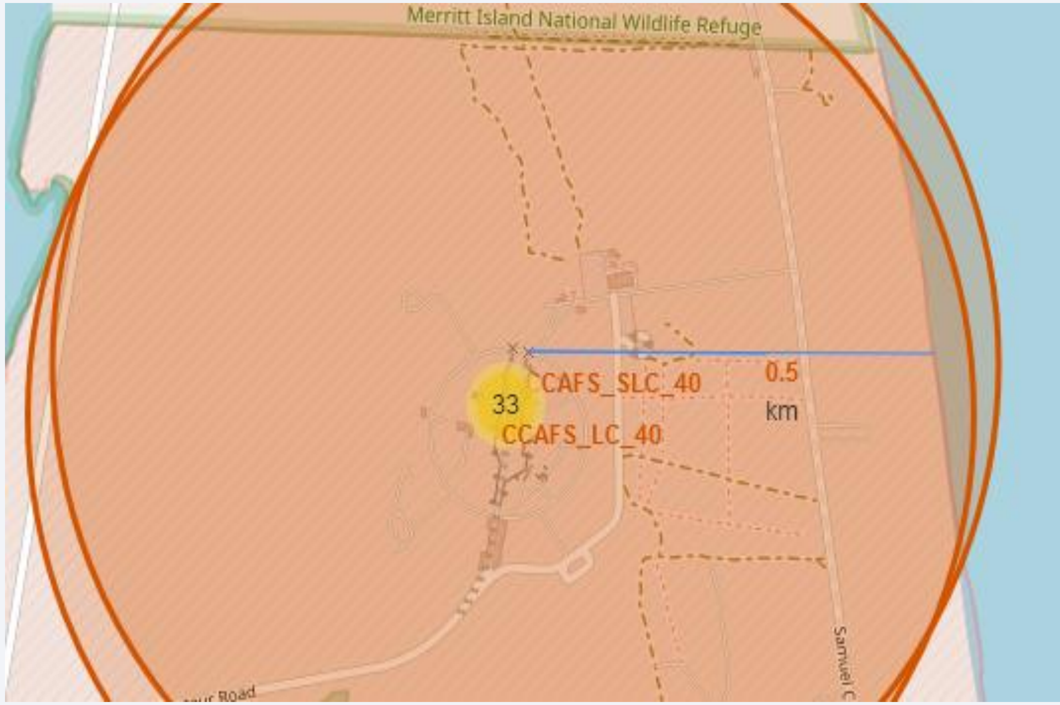
Color-labeled launch outcomes on the map

- From the color-labeled markers we can easily identify which launch sites have relatively high success rates
- Successful launches are marked in Green and Failed launches are marked in Red



Distance from the launch Site and Coastline

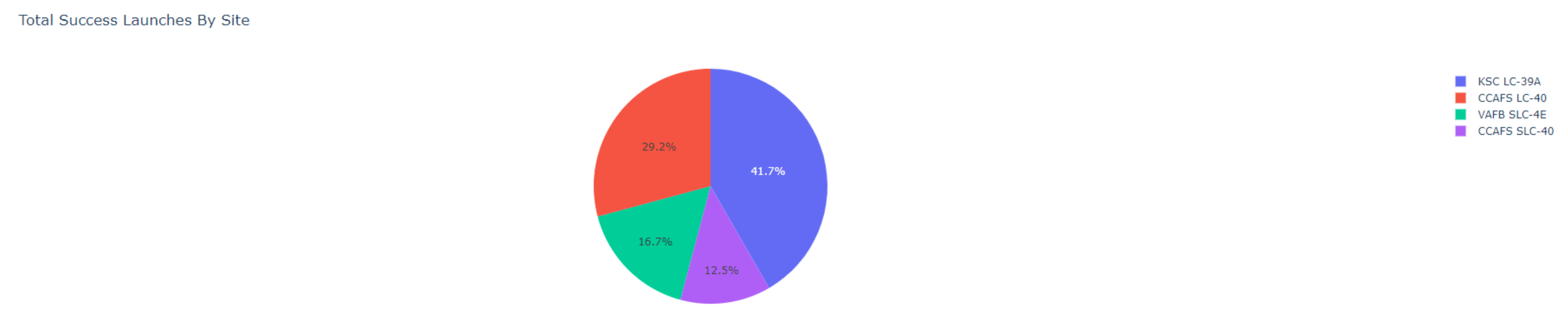
- From the map we can clearly see that the launch site is close to the coastline. The distance is displayed on the map



Section 5

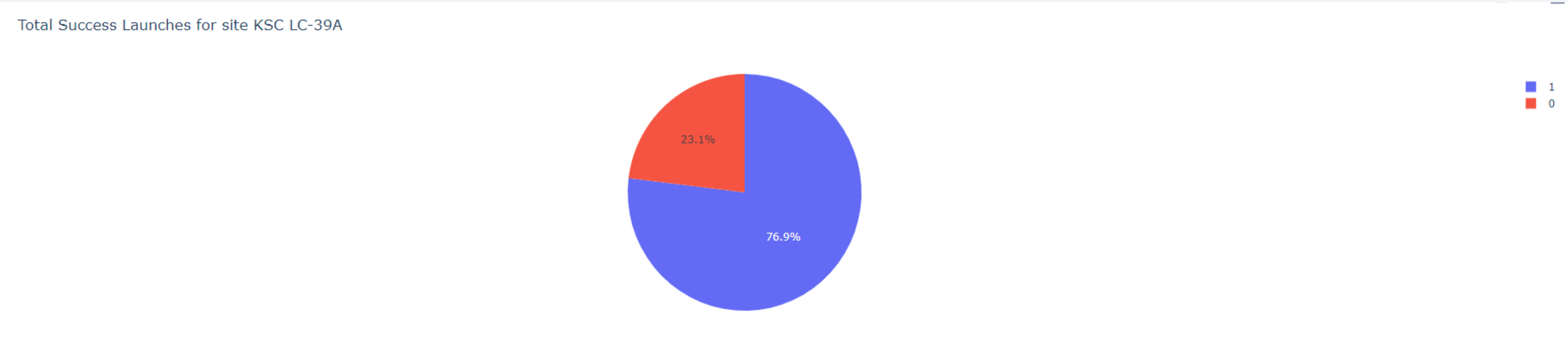
Build a Dashboard with Plotly Dash

Launch success for all sites



- The site KSC LC-39A hast the most successful launches

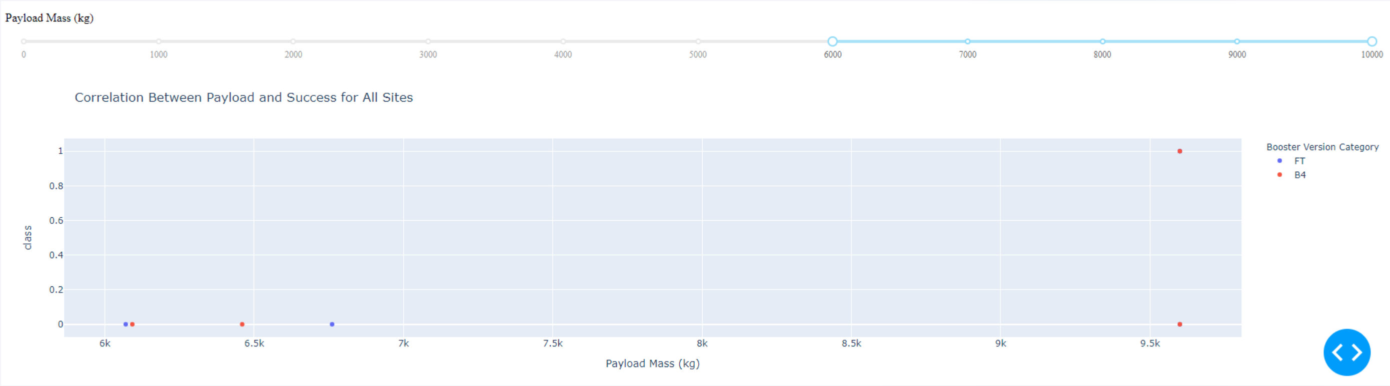
Launch Site with highest launch success ratio

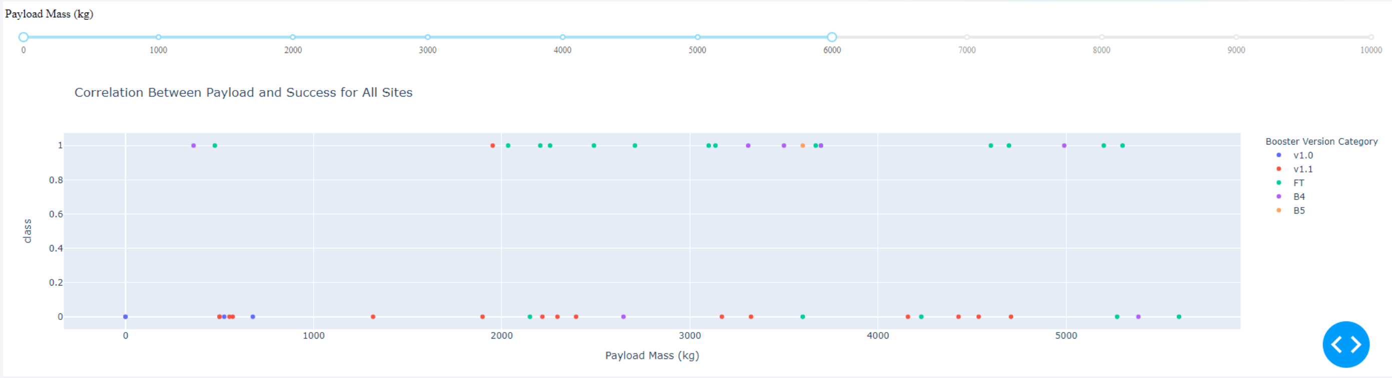


- KSC LC 39-A had the highest launch success rate (76.9%) with 10 successful and 3 failed

Payload Mass vs Launch outcome for all sites

- The highest successful launch rate was for payloads between 2,000 and 5,300 kg





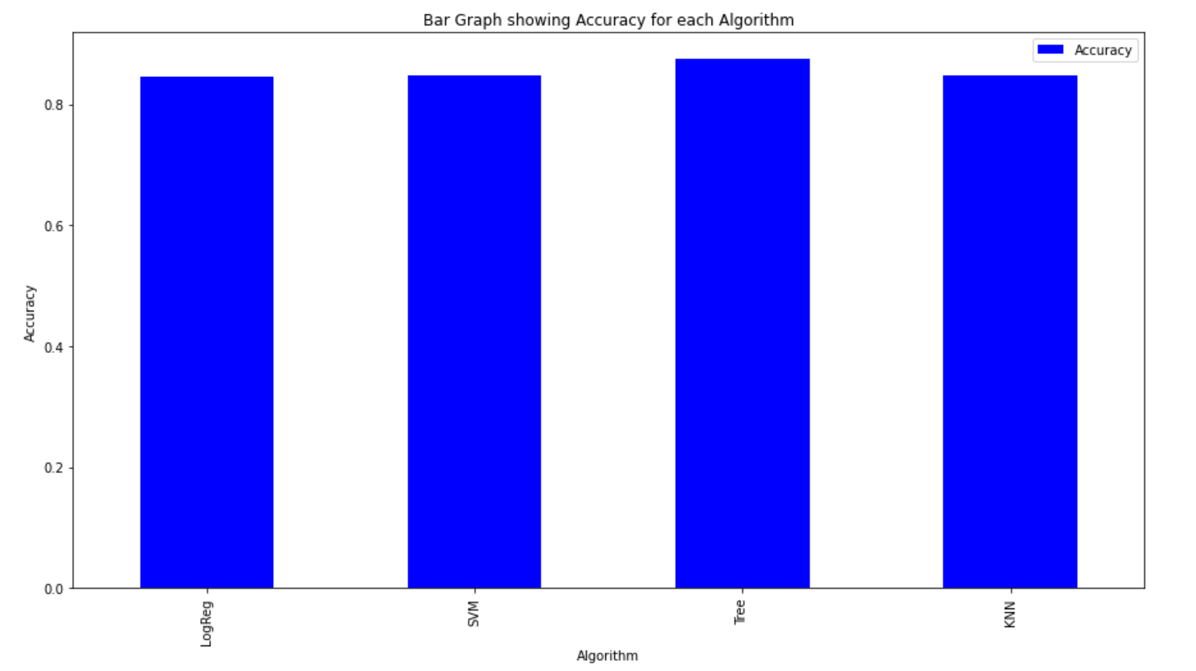
Section 6

Predictive Analysis (Classification)

Classification Accuracy

- As you can see on the Bar chart, all models have a close accuracy. However the Tree is the winner.

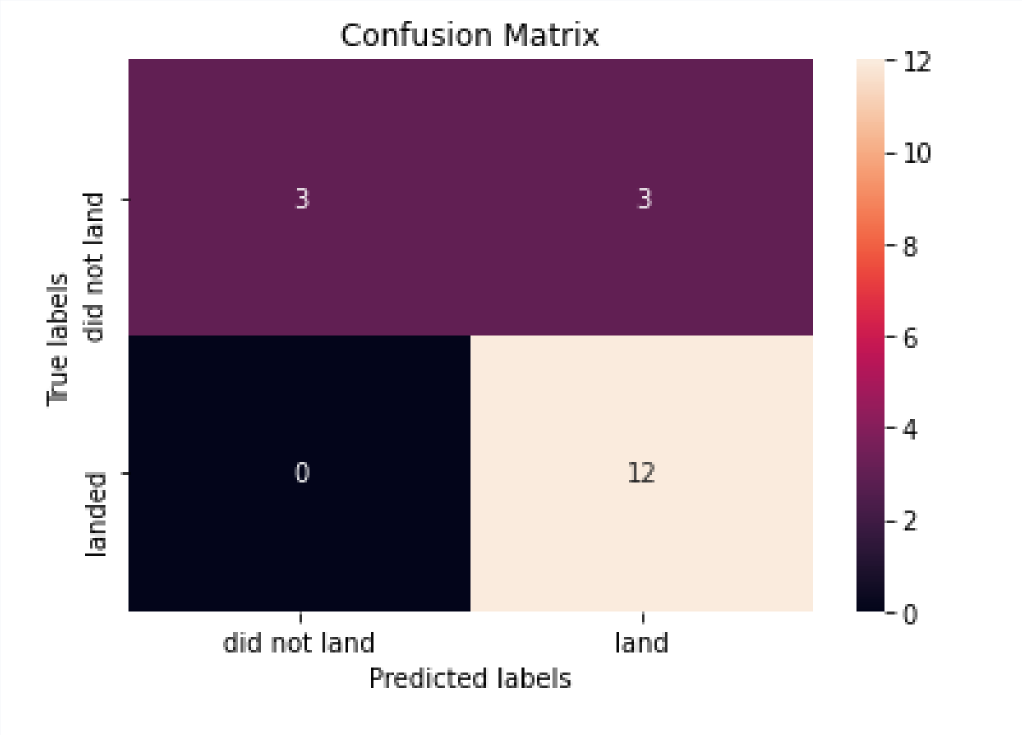
- Best model is the Tree algorithm with an accuracy of 0.875




```
Best Algorithm is Tree with a score of 0.875
Best Params is : {'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 5, 'splitter': 'random'}
```

Confusion Matrix

- Looking at the confusion matric, we see that the major problem of the Tree Classification is False positives



Conclusions

- The success rate for SpaceX launches increases over years
- Most of launch sites are situated close to the Equator line and in proximity to the coast
- The following launching sites KSC LC 39A and VAFB SLC 4E had higher success rates
- Orbit GEO, HEO, SSO, ES L1 had the highest Success Rate
- Decision Tree model is the best algorithm for this dataset

Appendix

Thanks to the Instructors, Coursera and IBM

Thank you!