IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - 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
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics
  - Predictive analysis results

Introduction

- Project background and context
  - SpaceX designs, manufactures and launches advanced rockets and spacecraft., making space travel affordable. Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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 variables such as payload mass, launch site, number of flights and orbits affect the success of the first stage landing?
  - With Time does the success rate improve?
  - Which is the best algorithm to use for the classification?

Section 1

Methodology

Methodology

Executive Summary

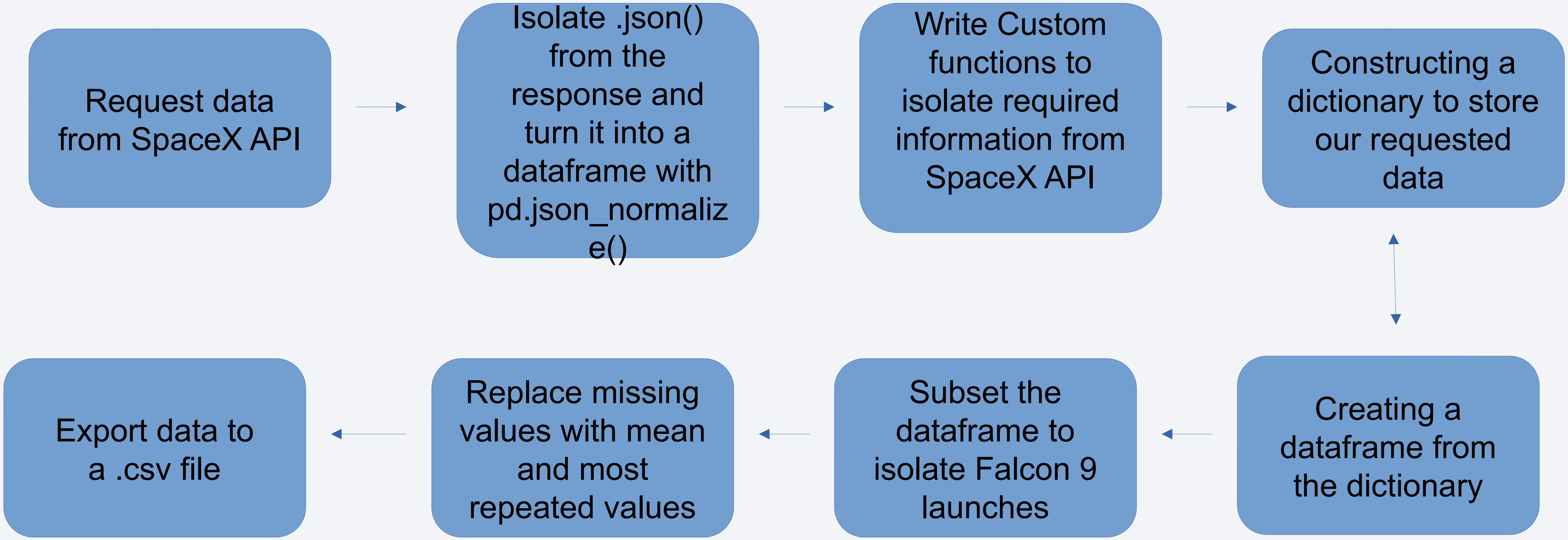
- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filter the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization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

Data Collection

Data was collected using an API request from SpaceX REST API and Web Scrapping data from a table in SpaceX's Wiki page. Using these two methods we collected the information about the launches of Falcon 9.

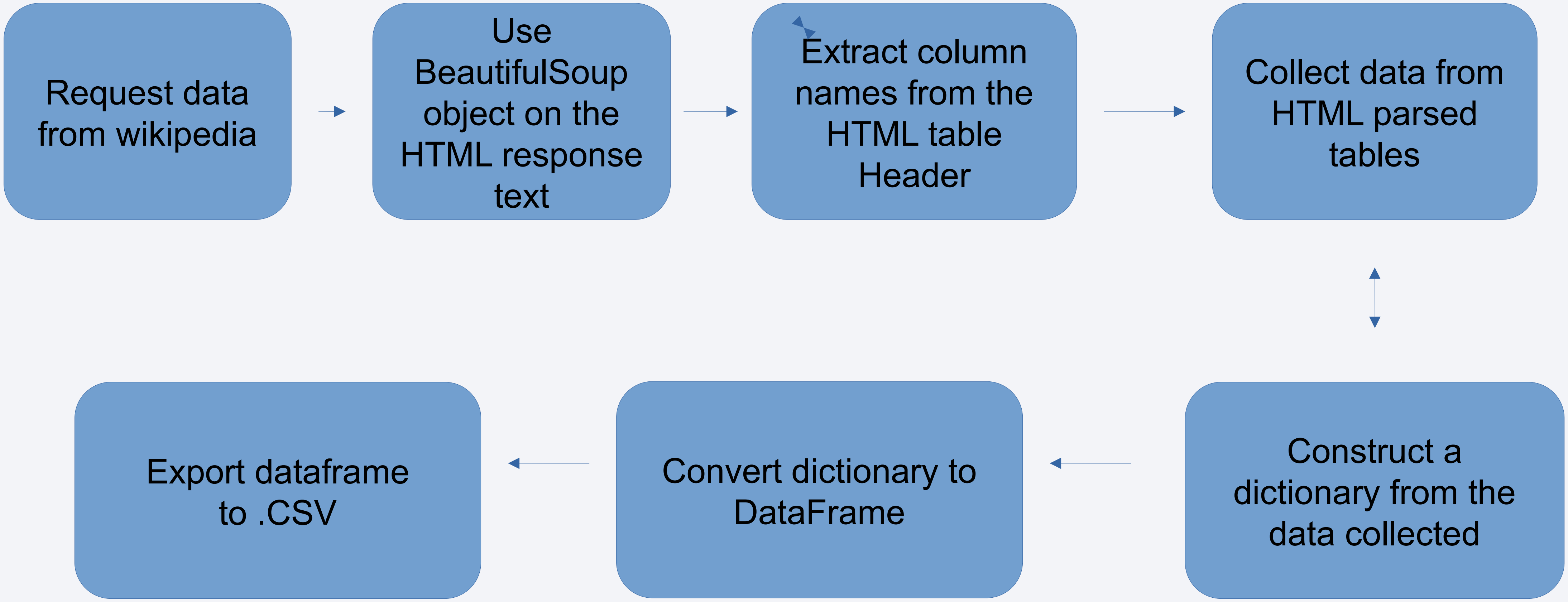
- Data Columns obtained using SpaceX REST API:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Web Scraping:
  - Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer, Launch, outcome, Version Booster, Booster Landing, Date, Time.

Data Collection – SpaceX API



GitHub– SpaceX API

Data Collection – Scraping



GitHub URL: Web Scraping

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. We convert those outcomes into training labels with ‘1’ meaning the booster landed successfully, ‘0’ unsuccessfully.

- Performing EDA and determine class labels
- Get the number of launches of Falcon9’s from each site
- Calculate the occurrences of each orbit
- Calculate the occurrences of mission outcomes per orbit
- Create a new feature variable class from landing outcomes and export the dataframe to .CSV

[GitHub Data Wrangling]()

EDA with Data Visualization

Charts Plotted

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

We used Scatter plots, Bar charts and Line charts to explain the above mentioned plots

Scatter plots show the relationship between two variable which we can observe and use later while designing an ML model

Bar charts explain categorical data, they show a relationship between a discrete categorical feature and a continuous measured value.

With Line charts we can observe trends in the data over time.

Github EDA with Data Visualization

EDA with SQL

SQL queries

- Displaying the names of the unique launch sires in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA’
- Display the total payload mass carried by booster launched by NASA(CRS)
- Displaying the average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad
- List the names of the boosters which have success in drone ship and have payload mass between 4000 and 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the max payload mass using a subquery.
- List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in the 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.

[GitHub EDA with SQL]()

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Add Marker with Circle, Popup Label and Text Label of NASA JSC using latitude and longitude coordinates.
- Add Markers with circle, popup label and text label of all launch sites using latitude and longitude

Colored markers of the launch outcomes for each launch sites:

- Add colored markers of success(Green) and failed(Red) launches using marker clusters to identify which launch sites have a high success rate

Distance between launch sites to its proximities:

- Add colored lines to show the distance between the launch site KSC LC -39A and its proximities like Railway, Highway, Coastline.

GitHub Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

Launch Sites Dropdown menu:

- Add a dropdown menu to select Launch Site

Pie Chart to display successful Launches:

- Add a pie chart to display successful launches for all sites and percentage of success/failure when specific sites are selected

A slider for payload mass:

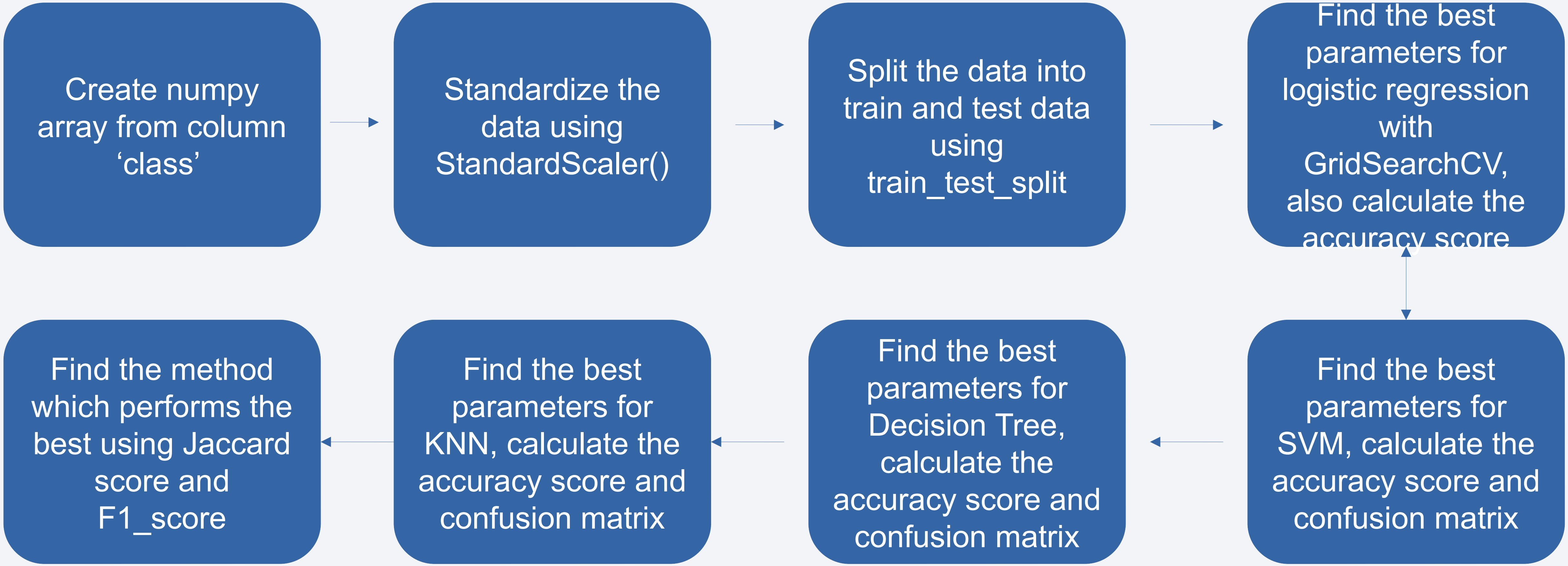
- Add a slidebar to select the Payload mass range.

Scatter chart of payload mass vs success rate

- Add a scatter chart for correlation between payload and launch success.

[Github SpaceX Dash App]

Predictive Analysis (Classification)



Github Machine Learning

Results

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

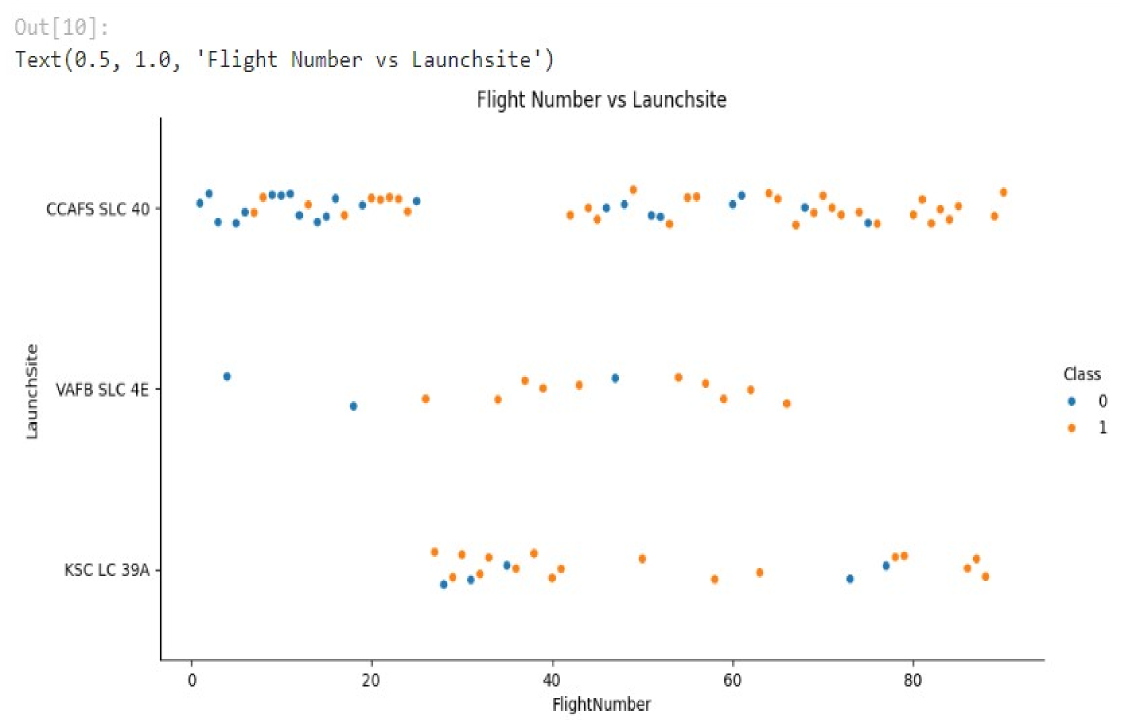
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

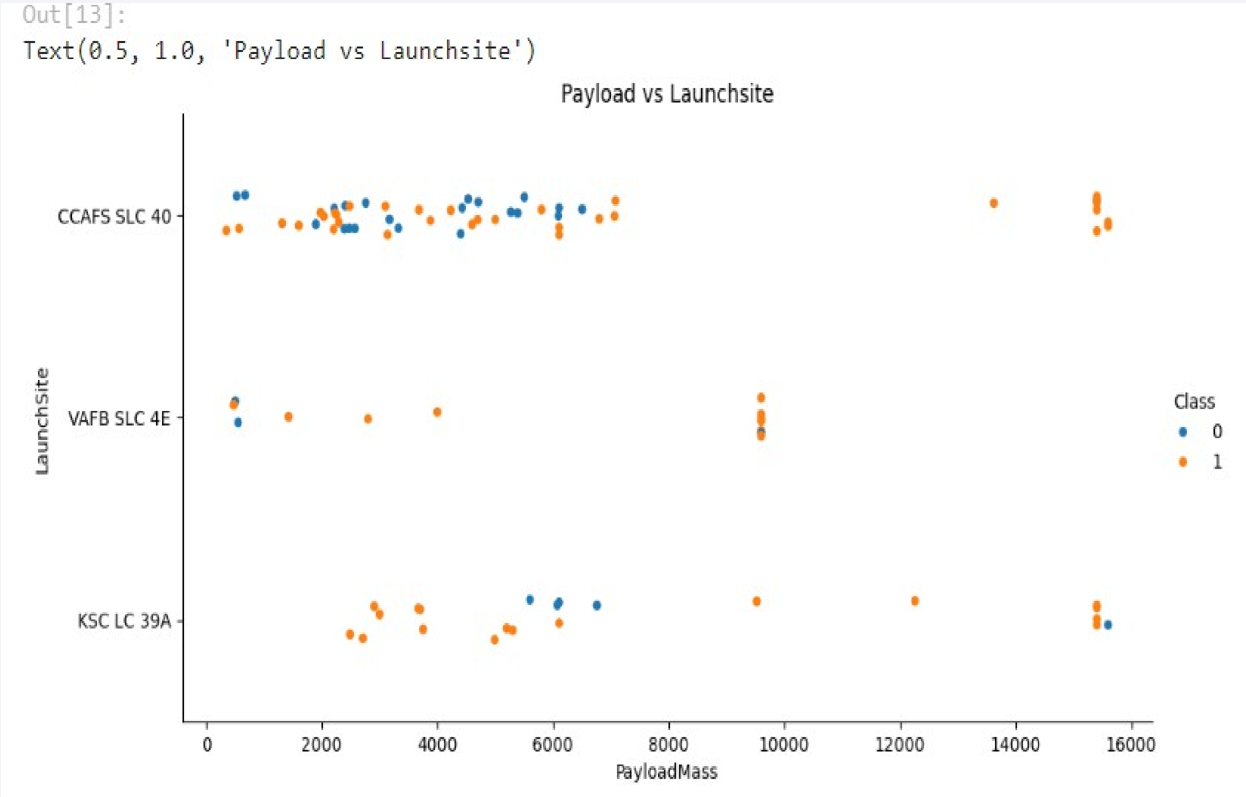
Explanation:

- The Flights early in the program failed while the latest ones are all successful
- Most of the flights have been launched from CCAFS SLC 40
- The overall success rates of the flights is high, specially the recent ones
- VAFB SLC 4E and KSC LC 39A sites have higher success rate.



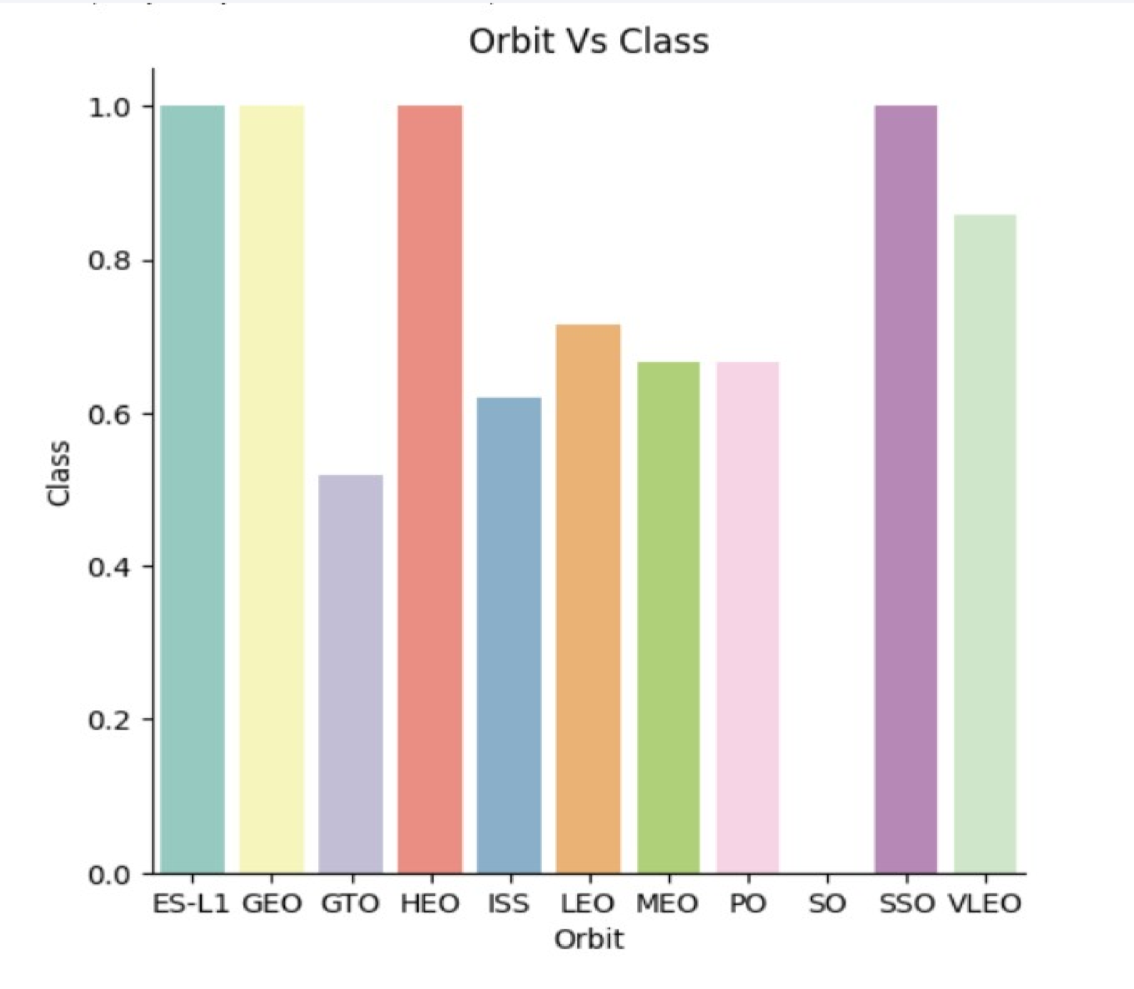
Payload vs. Launch Site

- Higher payload mass returns higher success rate of the success of the mission
- Payload mass of over 7000 Kgs returned better success rates to the mission
- KSC LC 39A shows a 100% success rate for payload mass of under 5000kgs.



Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO orbits have 100% success rate
- SO orbit has 0% success rate
- GTO, ISS, LEO, MEO, PO orbits have success rates between 50% and 85%

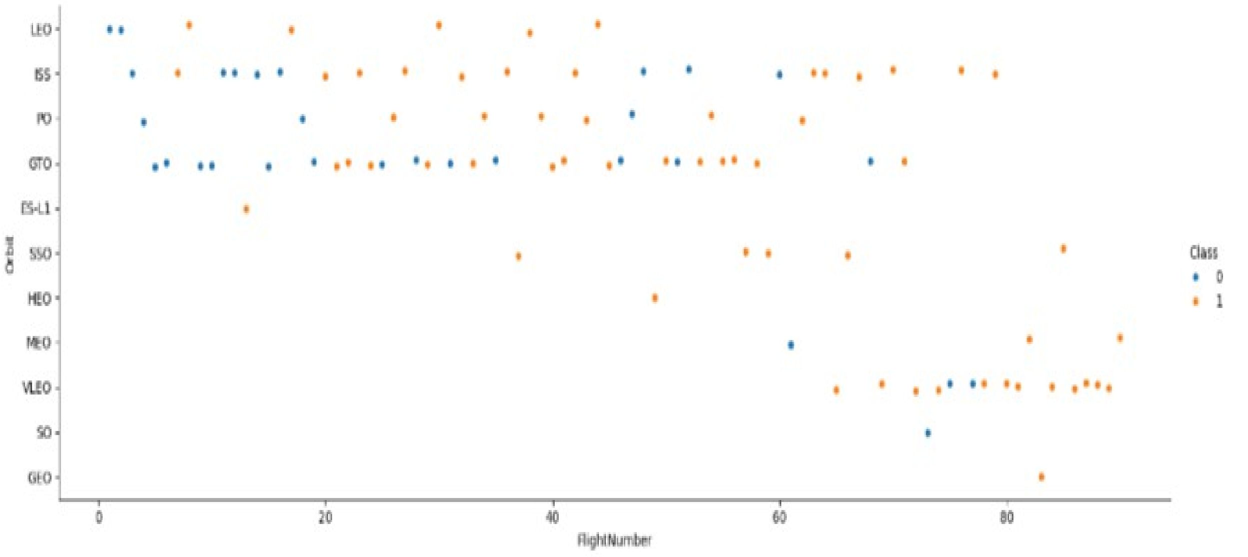


Flight Number vs. Orbit Type

- LEO has failures in the first couple of flights and then all have been successful, indicating early research problems.
- GTO orbit relationship with flight number is random,
- ES-L1, SSO has a 100% success rate
- Flights from VLEO have been successful offlate.

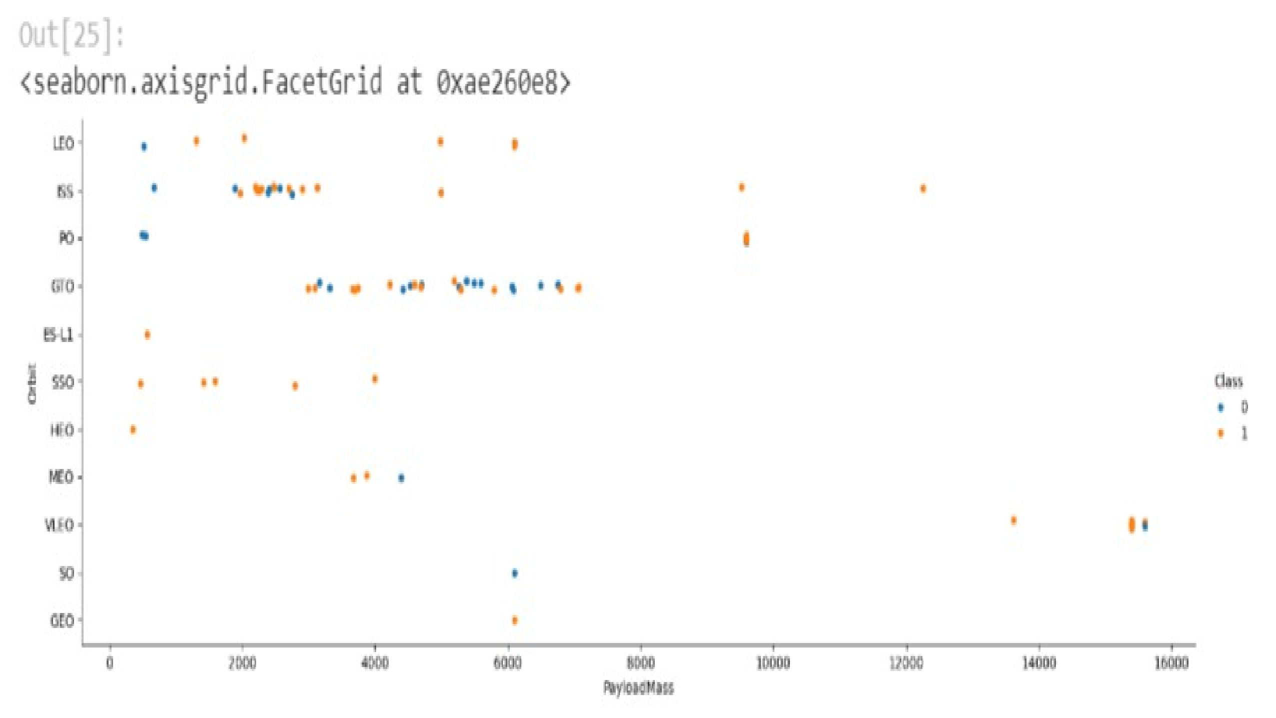
Out[23]:

<seaborn.axisgrid.FacetGrid at 0x697a428>



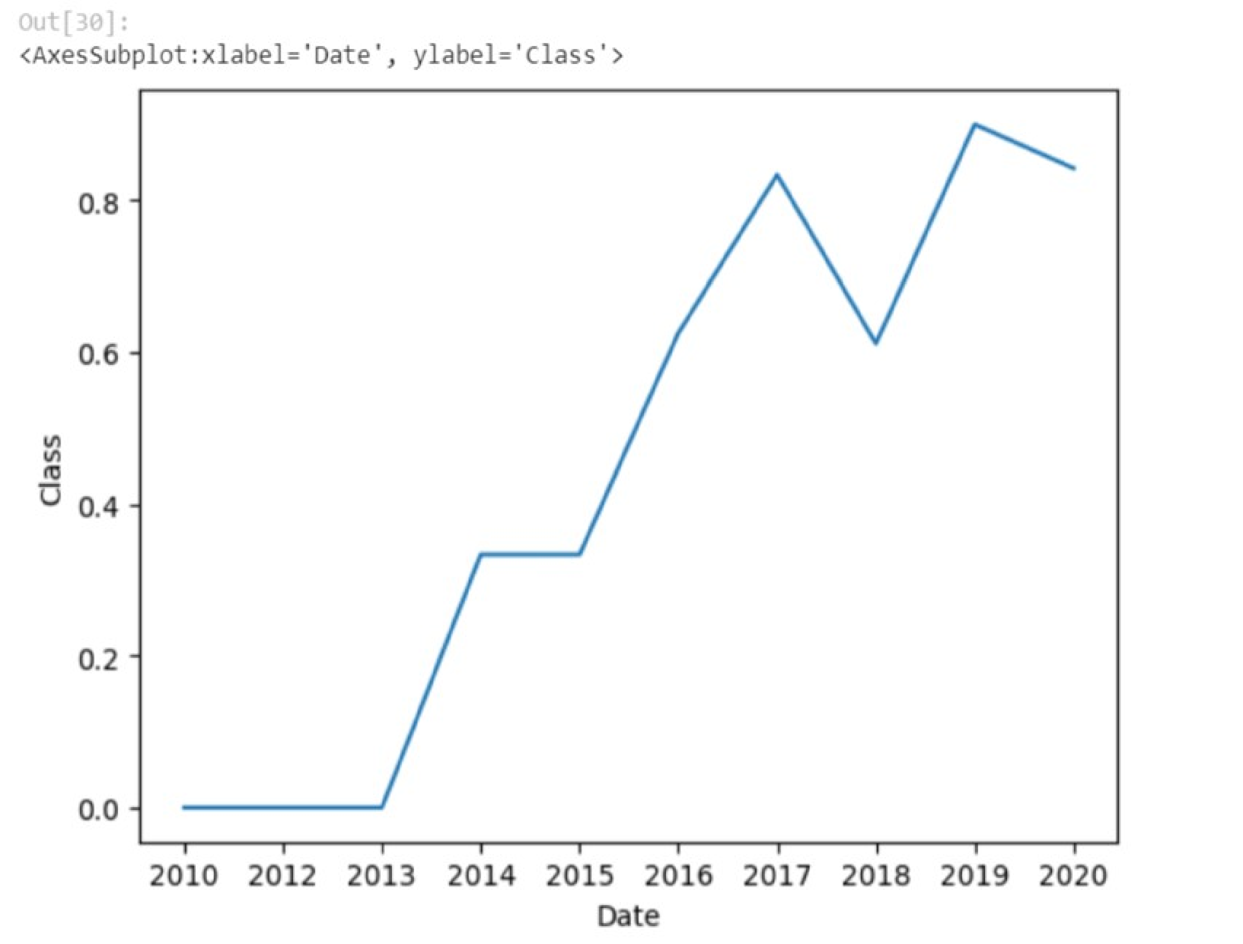
Payload vs. Orbit Type

- Heavy payloads have a negative correlation on GTO orbit and a positive correlation on ISS orbits.
- SSO has a payload limit at 4000 kg, with 100% success.



Launch Success Yearly Trend

- Success rates have increased since 2013 and saw a minor dip in 2018.
- 2019 and 2020 have peeked in the success rates.



All Launch Site Names

```
%%sql
Select DISTINCT "launch_site" From SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
Out[13]:
```

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

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE "launch_site" LIKE "%CCA%" LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
Out[14]:
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Displays the first 5 records where launch sites begin with ‘CCA’.

Total Payload Mass

In [15]:

```
%%sql
SELECT SUM(payload_mass__kg_) FROM SPACEXTABLE
WHERE "Customer" = 'NASA (CRS)'
```

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

Out[15]:

SUM(payload_mass__kg_)
45596

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

Average Payload Mass by F9 v1.1

In [16]:

```
%%sql
SELECT AVG(payload_mass__kg_) FROM SPACEXTABLE
WHERE Booster_Version LIKE '%F9 v1.1%'
```

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

Out[16]:

AVG(payload_mass__kg_)
2534.6666666666665

Displays the average payload mass carried by booster version 'F9 v1.1'

First Successful Ground Landing Date

```
In [17]:

%%sql
SELECT MIN("Date") FROM SPACEXTABLE
WHERE "Landing_Outcome" LIKE'%Success (ground pad)%'

 * sqlite:///my_data1.db
Done.
Out[17]:
```

MIN("Date")
2015-12-22

Lists the date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

In [18]:

```
%%sql
SELECT Booster_Version FROM SPACEXTABLE WHERE
Landing_Outcome LIKE '%Success (drone ship)%' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
```

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

Out[18]:

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

Lists 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

In [19]:

```
%%sql
SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE
GROUP BY MISSION_Outcome
```

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

Out[19]:

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

Lists the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

In [21]:

```
%%sql
SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS_
```

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

Out[21]:

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

List the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
In [26]:

%%sql
SELECT  substr("Date", 6, 2) as month, "Date", "Booster_Version", "Launch_Site", "landing_c
WHERE "Landing_Outcome" LIKE '%Failure (drone ship)%' and substr("Date", 0, 5) LIKE '%2015%

 * sqlite:///my_data1.db
Done.
Out[26]:
```

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

List 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

In [28]:

```
%%sql

SELECT "Landing_Outcome", count(*) FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' and '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY count(*) DESC
```

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

Out[28]:

Landing_Outcome	count(*)
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

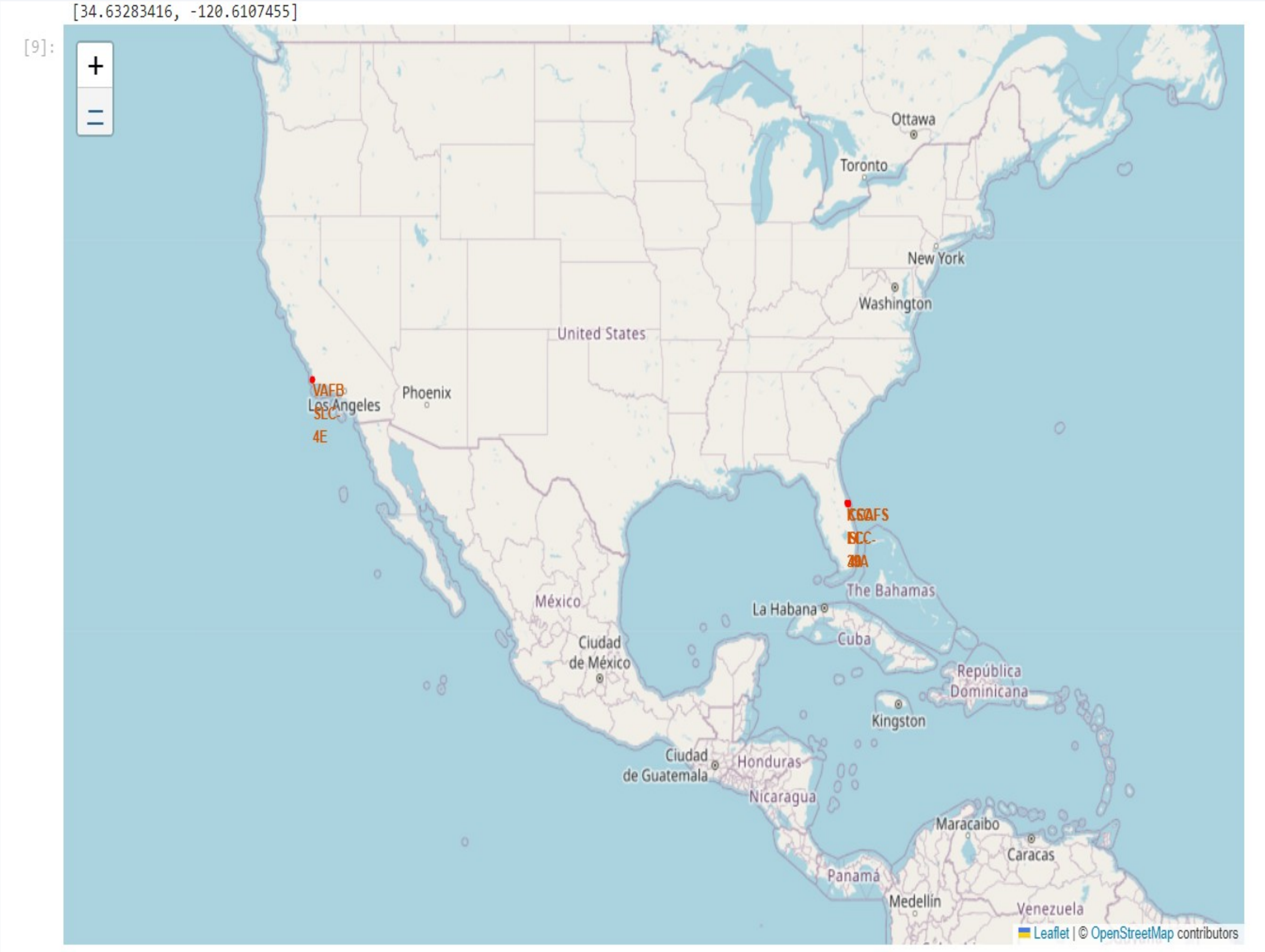
Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

Launch Sites Proximities Analysis

Launch Site location markers

- Most of the launch sites are around the east and west coast along the ‘Equator Line’, this is due to the low risk factor around the ocean if anything goes wrong. The launch sites are around the equator so that the rockets launched will be upto the speed of earth’s rotation(1670km/hr) quicker and less fuel/effort will be used



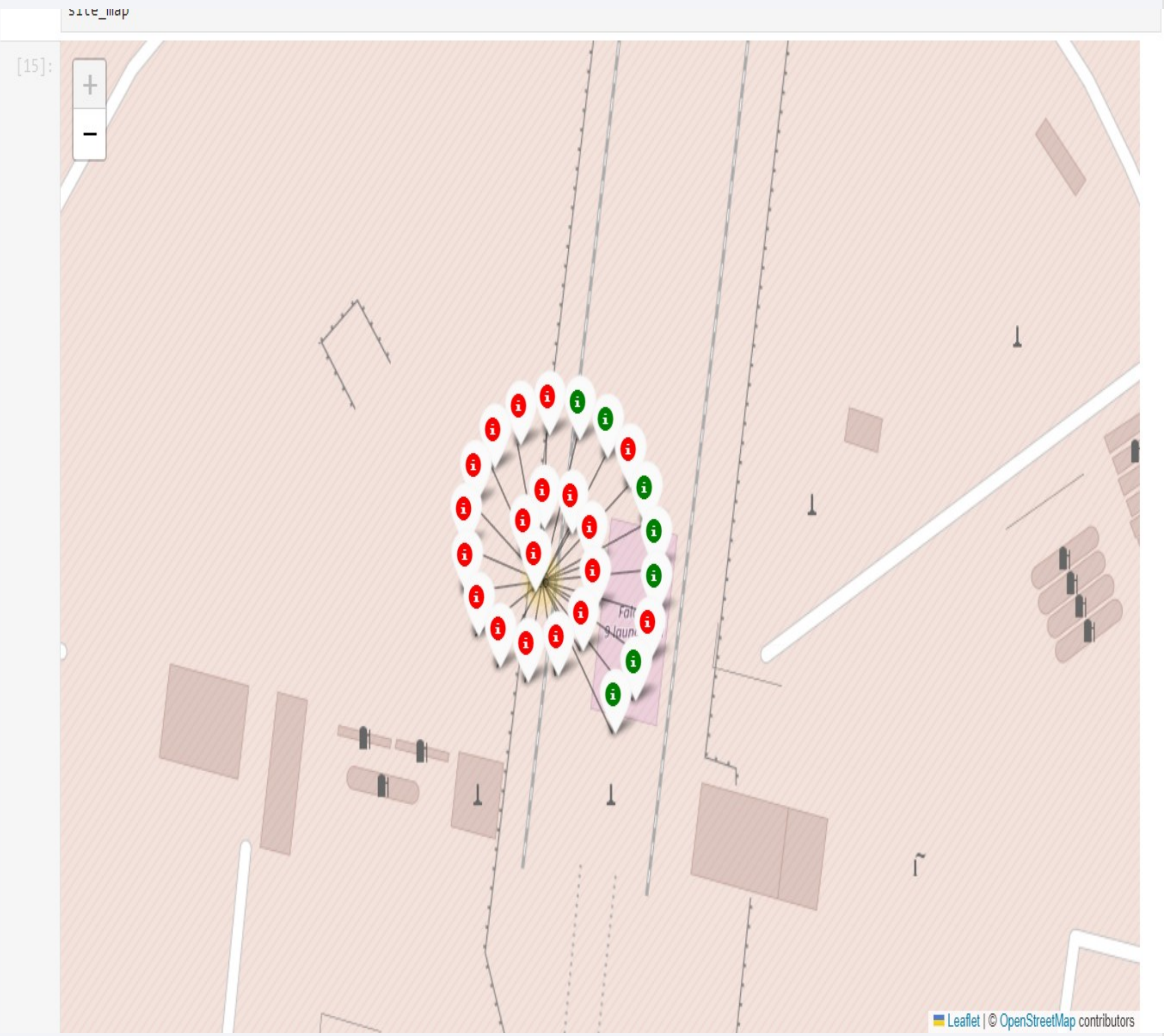
Launch records with color labels

Green Marker = Successful Launch

Red Marker = Failed Launch

From the color labeled markers we can easily identify which launch site has high or low success rates for the launch.

Launch Site CCAFS LC-40 has a very low success rate as can be seen in the image.

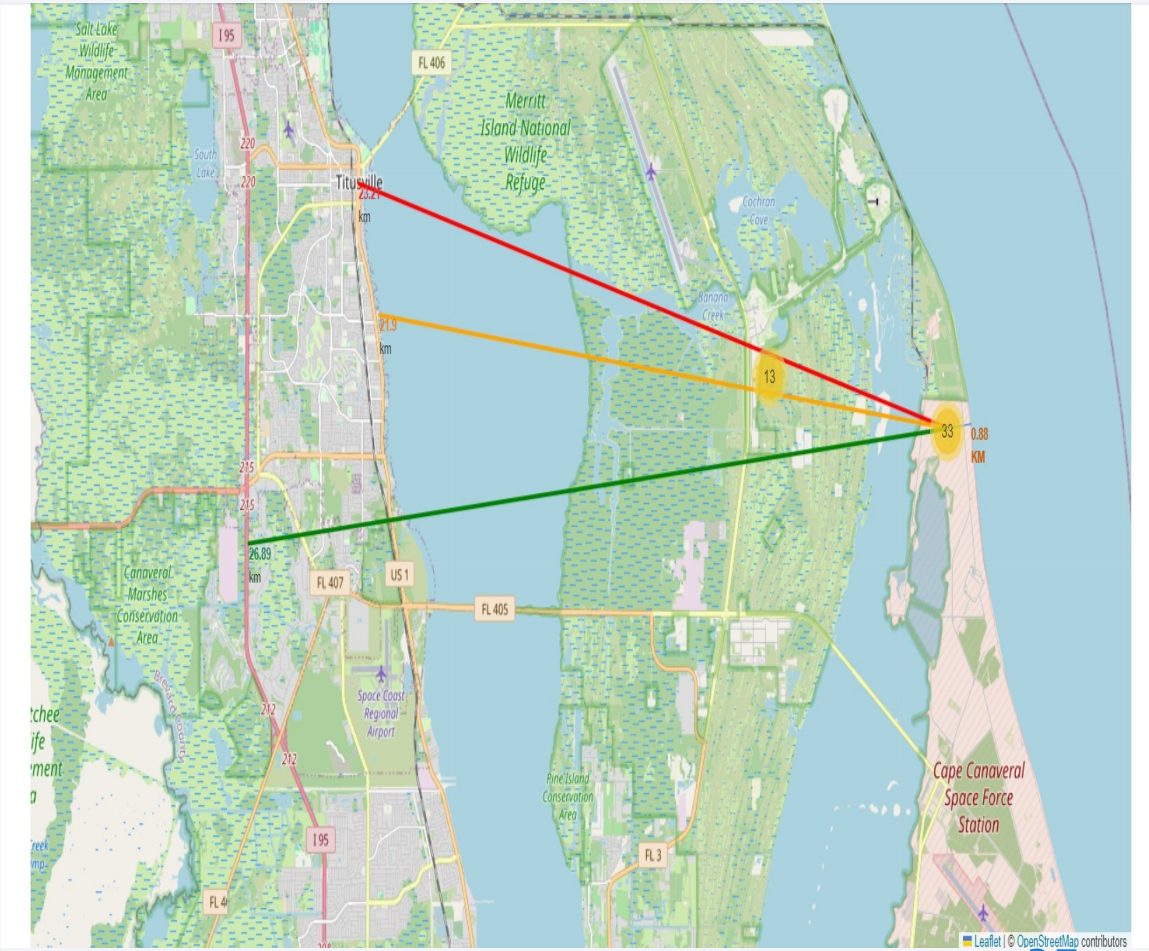


Distance to the proximities from CCAFC LC-40

From the launch site CCAFS LC-40, we have measured the distance to the proximities

- Relative proximity to railway (23.21km)
- Relative proximity to highway (28.9km)
- Relative proximity to coastline (21.9km)

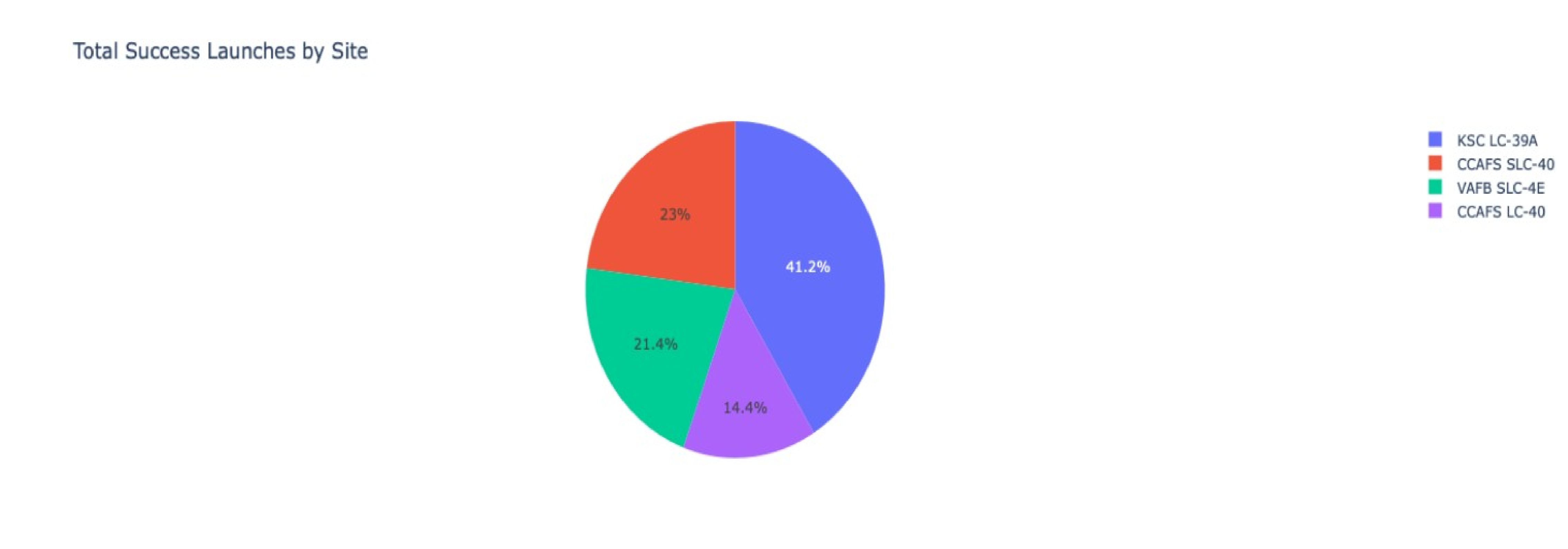
The launch site is 23km away from the closest city Titusville, which is relatively close as rockets can cover distance of 15-25km in just a few seconds. This could be potentially dangerous to populated area.



Section 4

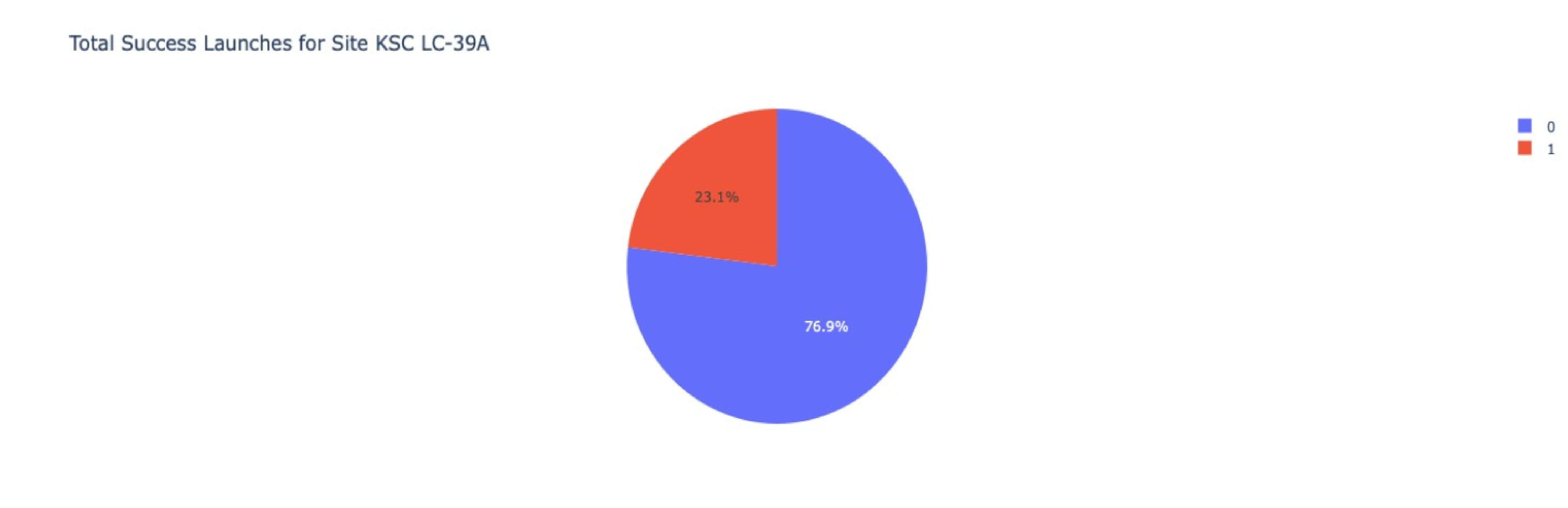
Build a Dashboard with Plotly Dash

Launch site success rate



The chart shows that launch site success rates, KSC LC-39A has the most successful launches among the launch sites

Launch Site with the highest success rate



From the previous chart we concluded that KSC LC-39A had the highest launch success rate, when observed on itself, we can see that it has a success rate of 76.9%.

Payload Mass vs Launch Site Outcomes

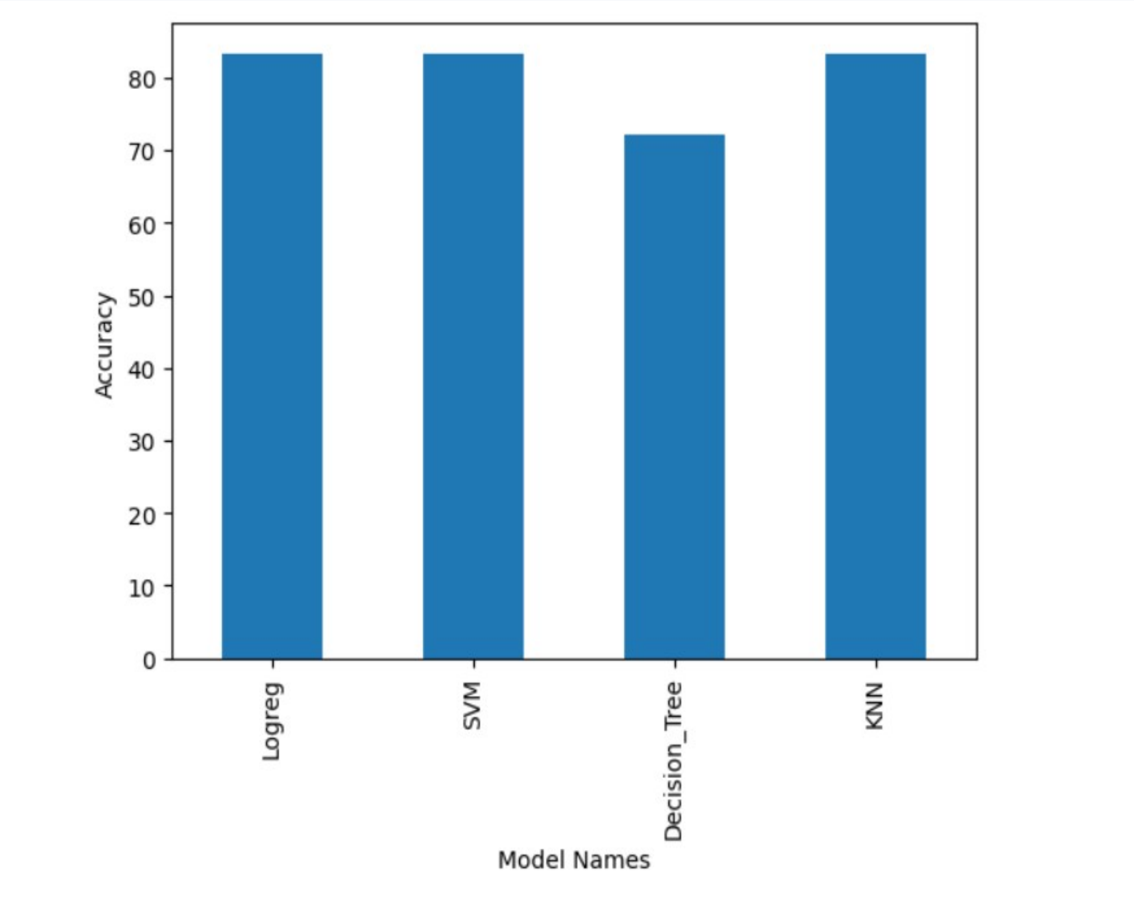
From the charts we can observe that payload mass between 2000 kg and 5300 Kg have the highest success rate

Section 5

Predictive Analysis (Classification)

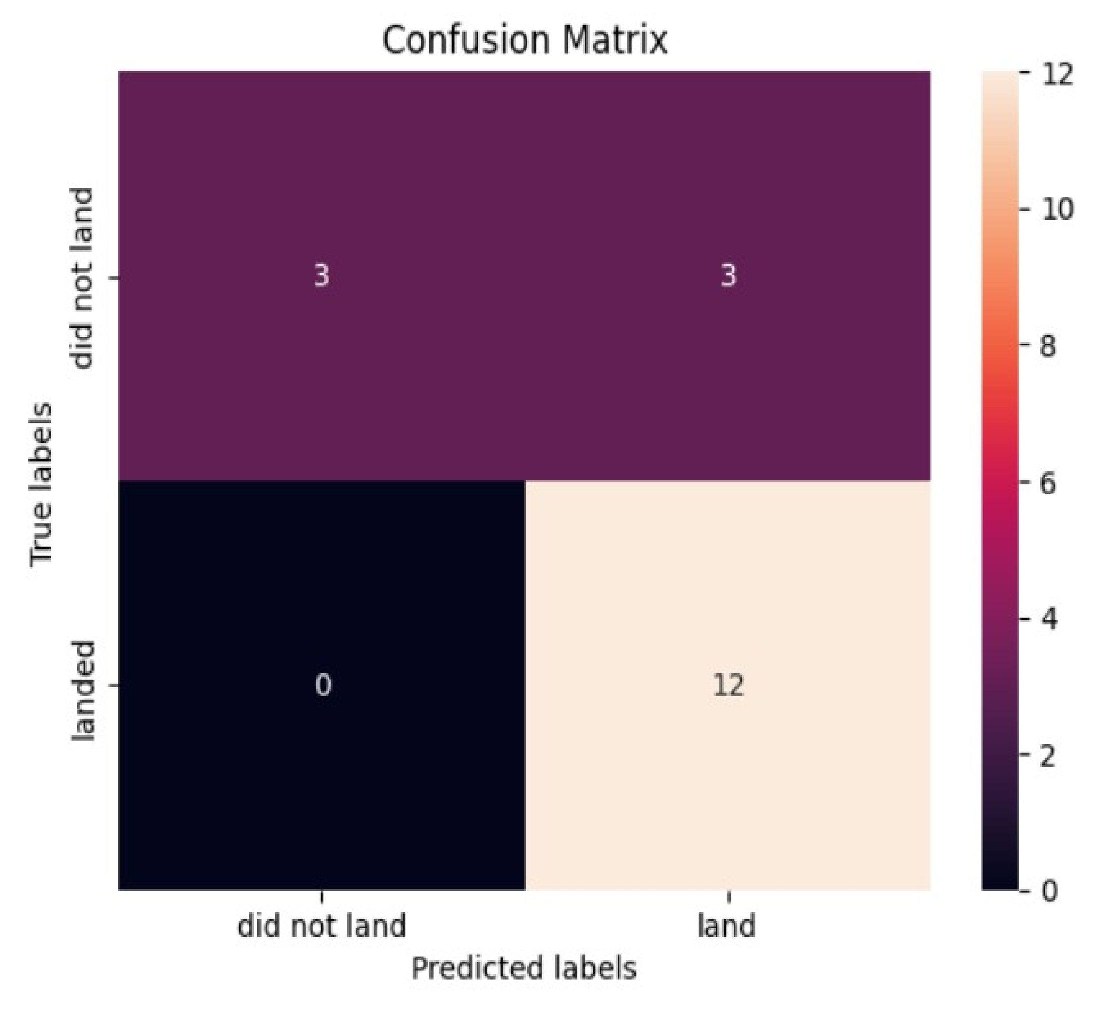
Classification Accuracy

- From the bar plot it can be seen that 3 of the 4 models are identical in accuracy of 83%, where as the decision tree model had an accuracy of 72%. Any of the three models, SVM, Logistic regression, KNN can be selected for classification as they perform similarly.



Confusion Matrix

- Examining the confusion matrix it can be stated that the logistic regression, SVM and KNN all can classify the landed label with 100% accuracy, but then couldn't properly classify the did not land label, it had an accuracy of 50% on the did not land category.



Conclusions

- KNN, SVM and Logistic regression models are all equally good and perform the best among the models selected for classification
- Launch with low payload mass showed better results than higher payloads.
- The first few launches were not successful, but with time the launches were successful and showed 90-100% success rates
- KSC LC-39A has the best launch success among the launch sites
- Orbits ES-L1, GEO, HEO, SSO all have 100% successful launches

Appendix

Special thanks to my family and friends who believed in me and a special mention to my mother who pushed me to procure this course offered by IBM, Thank you to the team at IBM and Coursera, and finally to all the Instructors involved in the course, a very heartfelt thank you.

Thank you!