



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ubong Ntekim
4/12/2023



Outline

▪ Executive Summary	3
▪ Introduction	4
▪ Methodology	5 - 16
▪ Insight drawn from EDA	17 - 33
▪ Launch Sites Proximities Analysis	34 - 37
▪ Build a Dashboard with Plotly Dash	38 - 41
▪ Predictive Analysis (Classification)	42 - 44
▪ Conclusion	45

Executive Summary

Summary of methodologies

The data in this report has been collected using public sources. Used Space X application interface (API) to get core data. Other data is obtained via web scraping Wikipedia. The data is explored using structured query language and visualization in the form of geographic maps. Subsequently, tables with columns containing important characteristics were selected. By means of one-hot encoding, the categorical variables are transformed into binary variables. A label that reflects successful landings of Space X launches was defined. After the data is standardized, Grid-Search was applied to find the best variables for Machine Learning models. The accuracy score of all models are included in this report.

Summary of all results

Four machine learning models were applied to the data: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbours. All produced similar results with accuracy rate of about **83.33%**. All models predicted successfully landings more data is needed for better model determination and accuracy.

Introduction

Project background and context

Space_X 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.

Space_N wants to determine if the Falcon 9 first stage will land successfully assuming that 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 statement

Can Space_N make use of data from available resources to train machine learning models that can predict a successful launch and recovered landing of the first stage.

Section 1

Methodology

Methodology

A. Data collection methodology:

Data was collected from two public resources Space X public API and Wikipedia.

B. Performed data wrangling:

Transformed data from categorical to binary variables to classify successful and unsuccessful Space X landings.

C. Performed exploratory data analysis (EDA) using visualization and SQL:

Used IBM Watson Studio DB2 cloud service to set up a Database. Created a Python integration and queried tables from database with Structured Query Language to draw insights.

D. Perform interactive visual analytics using Folium:

Folium is a library in Python that helped visualize Space X launch sites and key location.

E. Performed predictive analysis using classification models

Applied Logistic Regression, Support Vector Machine, Decision Tree and KNN, achieved increased machine learning model performance with Grid-Search.

Data Collection Overview

Data collection Process

Is a combination of a call requests to the public Space X API and web scraping Space X data on Wikipedia.

In the following slides flowcharts are used to visualize data collection process.

Data columns from Space X API

The data columns that were extracted from Space X API are;

- Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins,
- Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude

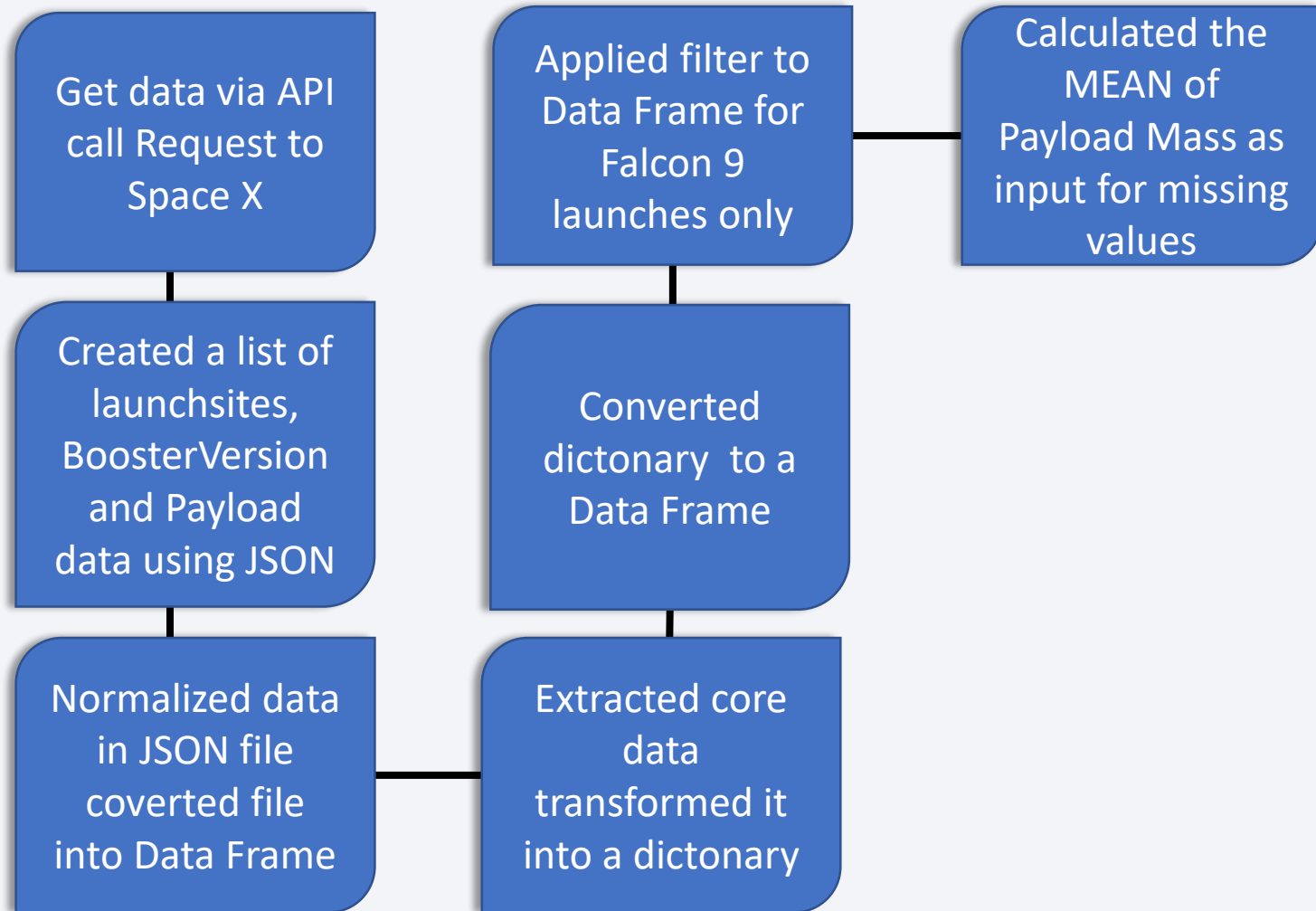
Data columns from Web scraping

- Flight No., Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection- SpaceX API

Data Collection
via Space X API

[https://github.com/ubongn/IBM Data Science Professional Certification/blob/master/Applied%20Data%20Science%20Capstone/Data%20Collection%20API.ipynb](https://github.com/ubongn/IBM_Data_Science_Professional_Certification/blob/master/Applied%20Data%20Science%20Capstone/Data%20Collection%20API.ipynb)



Data Collection- Web scrapping

Data Collection
with
Web Scraping

https://github.com/ubongn/IBM_Data_Science_Professional_Certification/blob/master/Applied%20Data%20Science%20Capstone/Complete%20the%20Data%20Collection%20with%20Web%20Scraping%20lab.ipynb.

Executed call
request to
Wikipedia HTML

Used
BeutifulSoup
library to parse
HTML data to
table

Detected launch
information in
HTML table

Cast HTML Table
to Data Frame

Extract data by
interating
through Data
Frame

Created a
dictionary from
extracted data

Data Wrangling

Data wrangling is the process of cleaning and unifying datasets ready for analysis.

During this process a label is created for training data sets with (1) for successful landing outcomes and (0) for failed landing outcomes.

Mission Outcome and Landing Location are target columns that are supportive to the landing outcomes.

Finally a new label column called Class is added with value (1) for successful and (0) for failed Mission Outcomes.

https://github.com/ubongn/IBM_Data_Science_Professional_Certification/blob/master/Applied%20Data%20Science%20Capstone/Data-Wrangling.ipynb

EDA with Data Visualization

EDA stands for Exploratory Data Analysis this process starts with initial investigation on the data.

The Flight Number column (indicates the continuous launch attempts) and Payload are variables that could affect the Landing Outcome. After plotting this two variables in a scatter plot observe that as Flight Number increases, the first stage is more likely to land successfully. It seems that as the Payload increases the less likely the first stage will return.

Other variables that were plotted and analysed are: *Flight Number and Launch Site, Payload and Launch Site, Orbit and Success Rate, Flight Number and Orbit, Payload and Orbit*, and a yearly trend of Succeed launches.

Obtained preliminary insights on the relationship of variables, and how they would effect the landing success rate.

Feature engineering refer to the process of selecting and transforming relevant variables from raw data when creating a predictive model. The features that are used to predict the success rate of launches are:

'FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial'.

[Github Url](#)

EDA with SQL

To gather some information about the dataset, some SQL queries were performed.

SQL queries performed are for :-

- 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 the average payload mass carried by booster version F9 v1.1
 - List the date when the first successful landing outcome on a ground pad was achieved.
 - List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg
 - List the total number of successful and failed mission outcomes.
 - List the names of the booster versions which have carried the maximum payload mass.
 - List the failed landing outcomes on drone ships, their booster versions, and launch site names for 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 Url](#)

Build an Interactive Map with Folium

The success rate of a launch depends on many factors such as payload, orbit type or flight number. It may also depend on location and launch site proximities.

Building a Launch Site involves many factors in this part we want to discover these factors by analysing the existing Launch Site locations.

Folium library is used to create geographic maps, launch sites are marked with a circle, based on launch records. The success rate of a launch location is marked with (1) green and (0) red for failed launches. A drawn line on the map displays the distances between Launch Site proximities.

The locations analysis gives better insights on the launch site and understanding of its proximities.

https://github.com/ubongn/IBM_Data_Science_Professional_Certification/blob/master/Applied%20Data%20Science%20Capstone/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Built an interactive dashboard with plotly dash by plotting a pie chart showing the total launches and sites. After selecting a drop down menu the distribution of successful landings and success rate of a Launch Site is shown a pie chart visual.

The Scatter plot takes two input variables, the selected Launch Site and Payload.

The Payload variable is captured in a slide bar for dynamic selection of mass between 0 and 10000 kilograms.

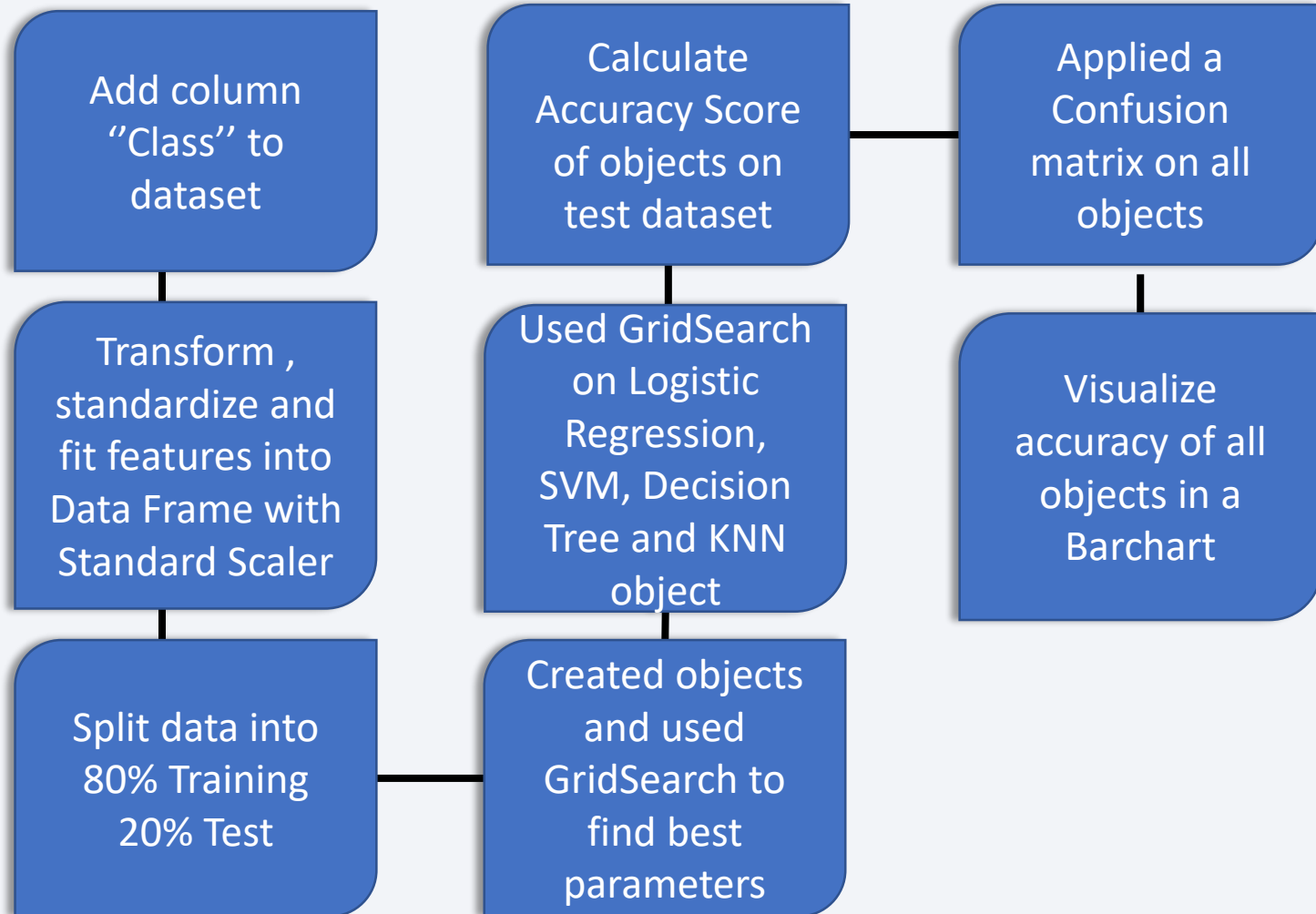
The Scatter plot visualize how successful landings (Class 0 and 1) vary across launch sites, payload and respective Booster version category.

https://github.com/ubongn/IBM_Data_Science_Professional_Certification/blob/master/Applied%20Data%20Science%20Capstone/spacex_dash_app.py

Predictive Analysis (Classification)

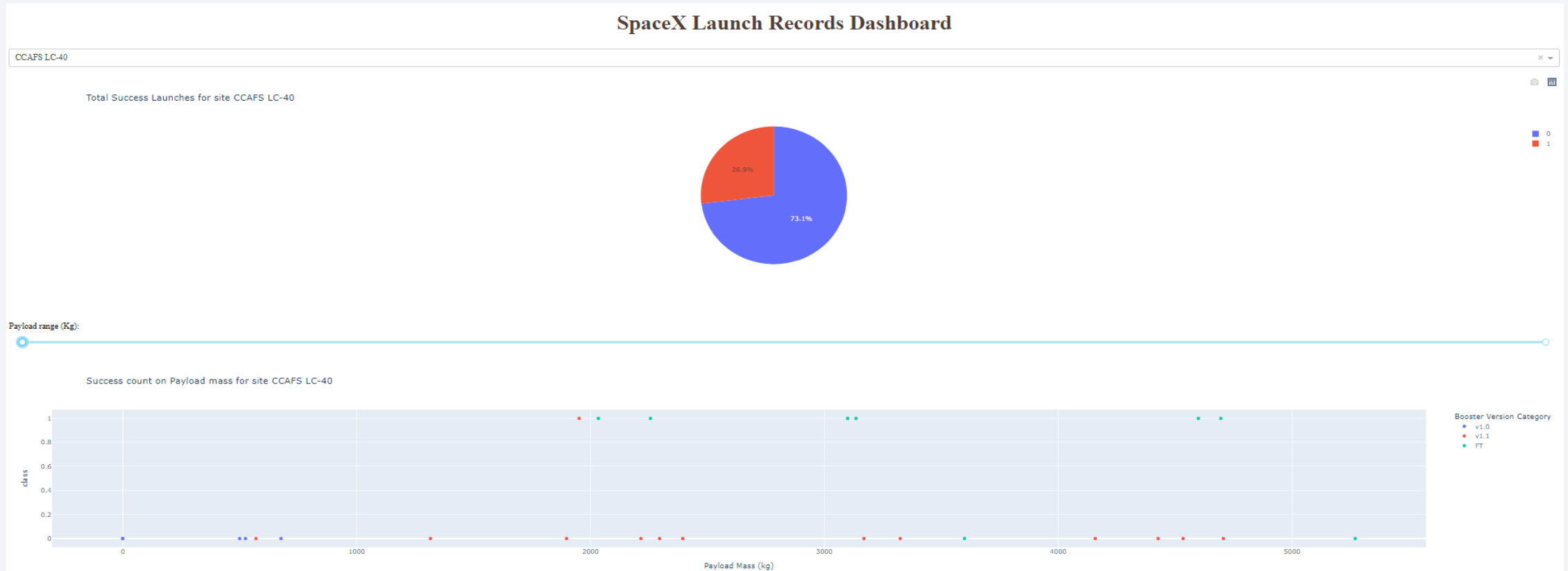
Perform exploratory Data Analysis and determine Training Labels

https://github.com/ubongn/IBM_Data_Science_Professional_Certification/blob/master/Applied%20Data%20Science%20Capstone/Machine%20Learning%20Prediction.ipynb



Results

A print screen of the final dashboard is shown below. At the central top there is a opportunity to select a Launch site. The slide bar with dynamic selection for Payload is shown in the middle of the dashboard. On the bottom a Scatter plot captures the success rate count in combination with Payload and respective.

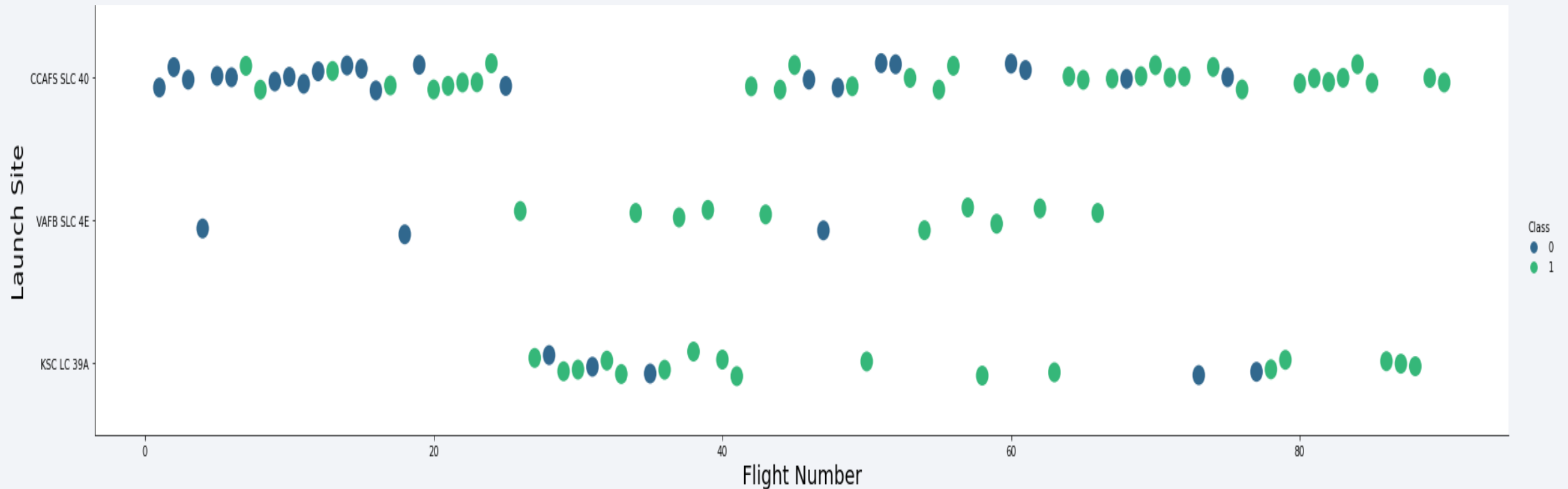


The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

Insights drawn from EDA

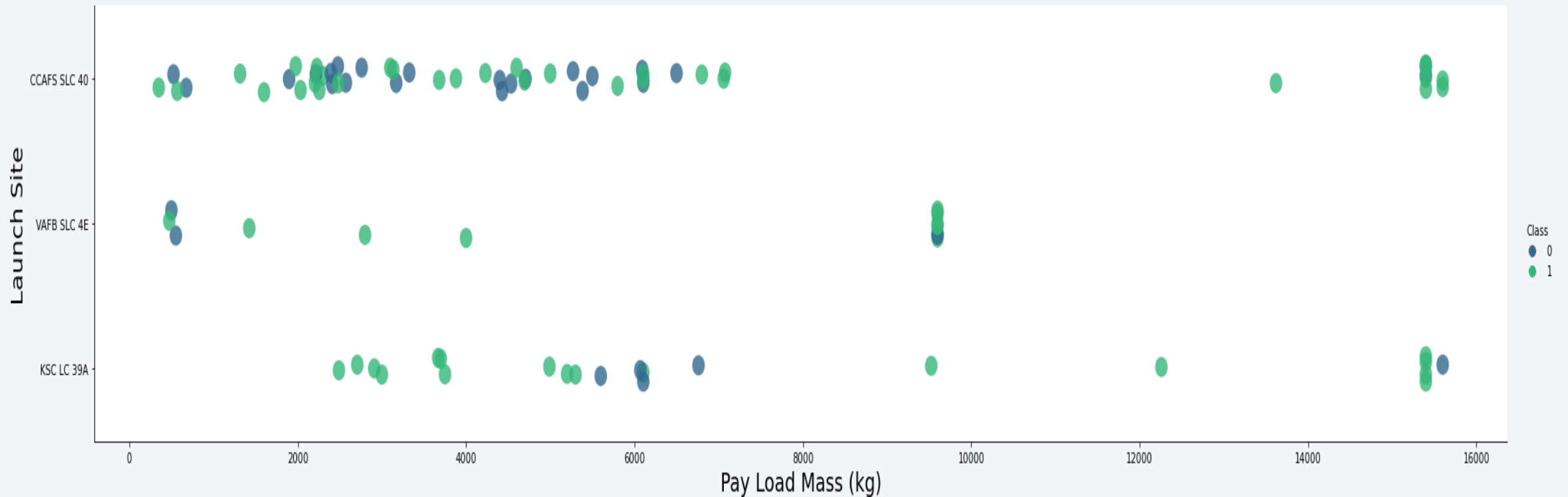
Flight Number vs. Launch Site



The Scatter Plot shows the relationship between Flight Number and Launch Site. The Green dots represent successful and Purple dots a unsuccessful launch.

- The plot shows that flights are launched from different sites. With an interval of approximal 20 flights per Launch Site.
- The Launch Sites have different success rates CCAFS LC-40 has a success rate of 66% while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Most Flights (number) are launched from site CCAFS LC-40 therefore appear to be a main Launch Site.

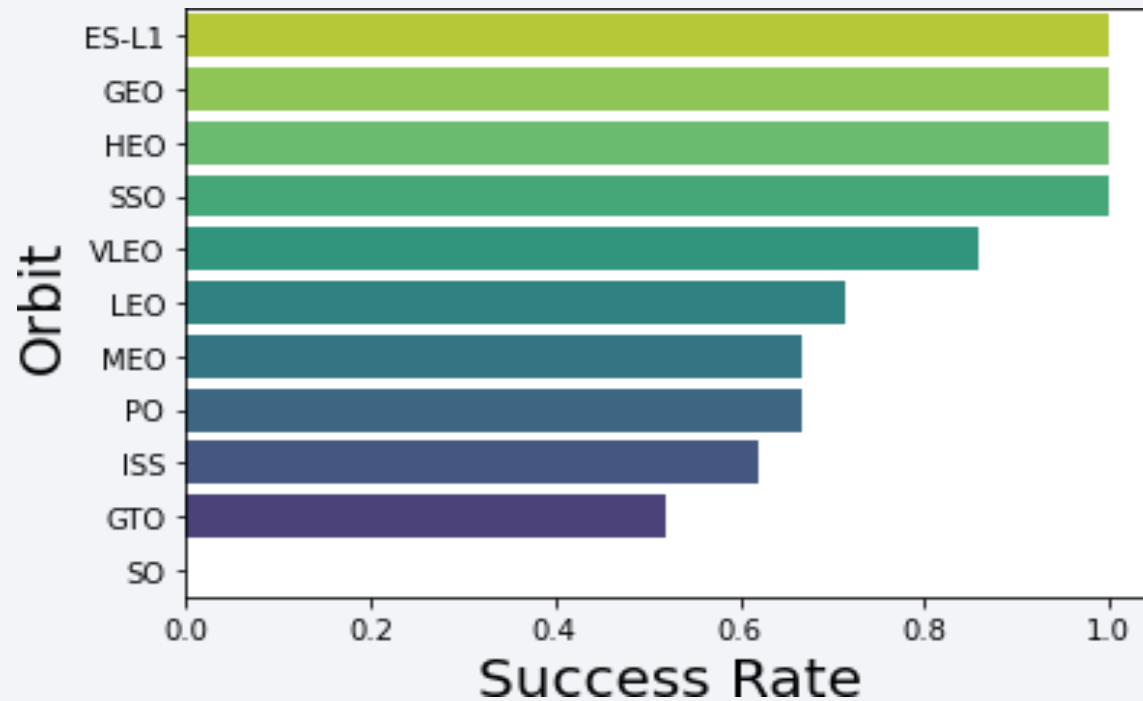
Payload vs. Launch Site



The Scatter Plot shows the relationship between Payload and Launch Site. The Green dots represent successful and Purple dots a unsuccessful launch.

- The plot shows that for Launch Site VAFB-SLC 4E there are no rockets launched for heavy payload greater than 10.000 kilograms.
- The Payload of the majority of rockets launched is between 0 and 6.000 kilograms.
- It seems that the Launch Sites have a variety of Payload used during launch.

Success Rate vs. Orbit Type

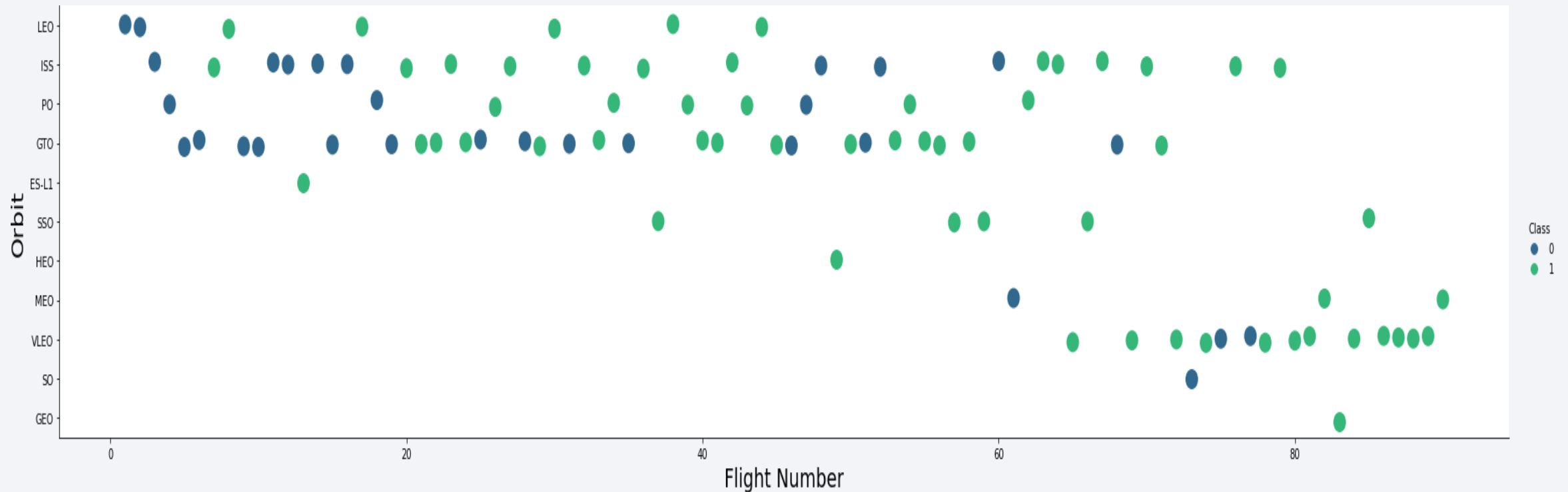


Success Rate Scale:
0 as 0%
0.6 as 60%
1 as 100%

The bar chart shows the relationship between Success Rate and Orbit type.

- Orbit ES-L1, GEO, and HEO have a success rate of 1 is equal to 100% with a sample size of (1) launch.
- SSO orbit have a success rate of 1 is equal to 100% with a sample rate of (5) launches.
- GTO orbit have a success rate of 50% with the largest sample rate of (27) launches.
- Finally SO not a successful orbit as it have an success rate of 0 equal to 0%.

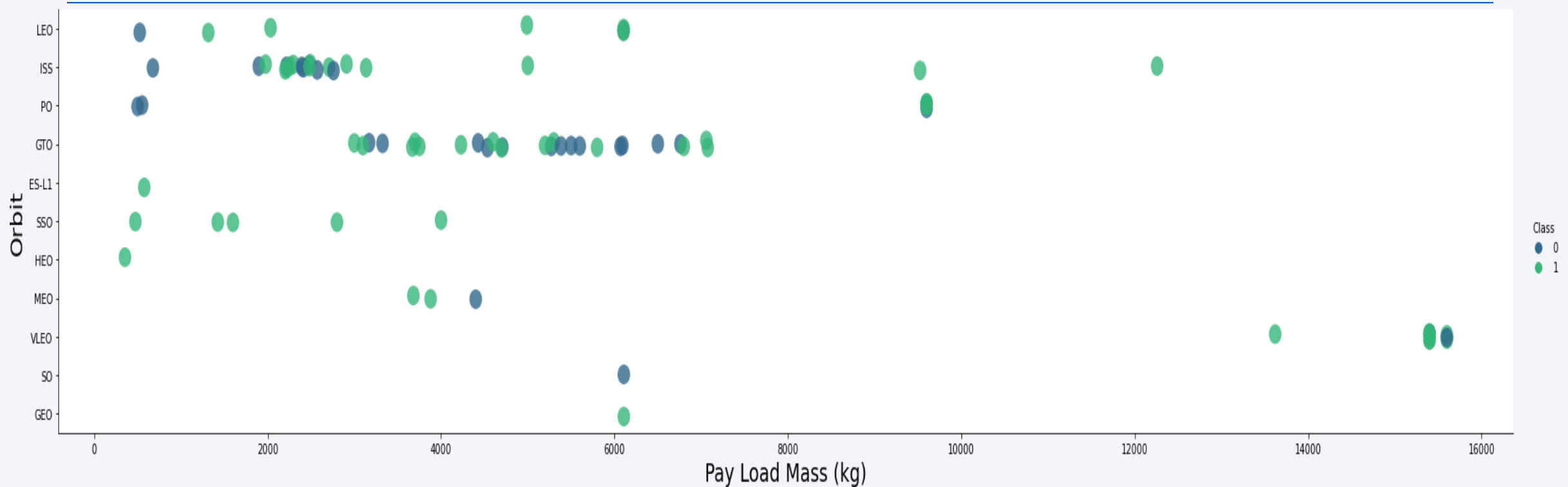
Flight Number vs. Orbit Type



The Scatter Plot shows the relationship between Flight Number and Orbit Type. The Green dots represent successful and Purple dots a unsuccessful launch.

- The plot shows that in LEO Orbit the Success appears related to flights; on the other hand, there seems to be no relationship between flight number when in GTO Orbit.
- There is a relationship between LEO and VLEO as SpaceX started with LEO and after success continued with VLEO.
- The launches of Space X perform better at low Orbits or Sun-synchronous Orbit meaning that the distance between Earth and Orbit is approx. 200 to 1,000 km.

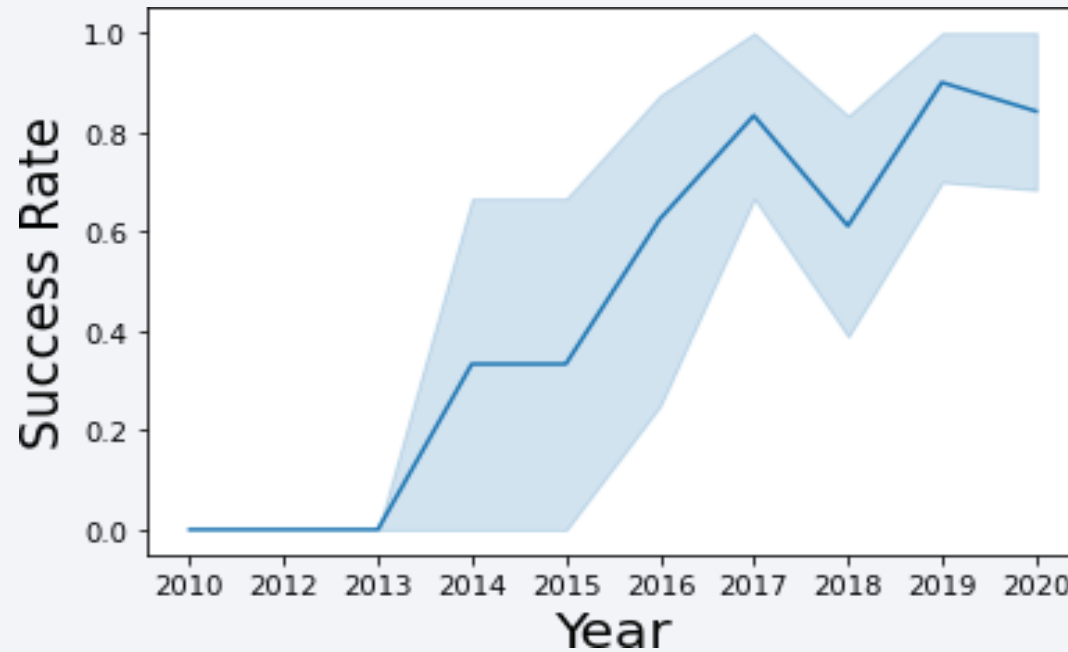
Payload vs. Orbit Type



The Scatter Plot shows the relationship between Payload and Orbit Type. The Green dots represent successful and Purple dots a unsuccessful launch.

- There is a clear relation between Payload and Orbit type.
- Launches with a heavy Payload will succeed with LEO, ISS and Polar (PO) Orbits.
- It is difficult to determine the success of GTO Orbit as it has positive and negative launches.

Launch Success Yearly Trend



The Line chart above shows the average launch trend on year basis.

- The success rate increases yearly as from 2013.
- In 2018 the success rate drops and return back to 80% average for the year after.

All Launch Site Names

```
In [8]: %sql select DISTINCT LAUNCH_SITE from SPACEXTBL
* ibm_db_sa://jqf40619:***@19af6446-6171-4641-8a
1qde00.databases.appdomain.cloud:30699/bludb
Done.
```

```
Out[8]:
```

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

The SQL query above is executed using DISTINCT to extract all site names from a SQL DBS cloud database created for this project.

- CCAFS SLC-40 and CCAFSSLC-40 represent launch site with data errors.
- CCAFS LC-40 was the previous name of site CCAFSSLC-40.
- Overall there are 3 distinct unique launch sites.

Launch Site Names Begin with 'CCA'

In [9]: `%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE '%CCA%' limit 5;`

* ibm_db_sa://jqf40619:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/bludb
Done.

Out[9]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
03-12-2013	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

The print screen a above is a SQL query that shows the launch sites in the database beginning with “CCA”.

Total Payload Mass

```
In [12]: %sql SELECT SUM(payload_mass__kg_) AS SUM_PAYLAOD_MASS FROM SPACEXTBL WHERE customer = 'NASA (CRS)';  
* ibm_db_sa://jqf40619:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.a  
ppdomain.cloud:30699/bludb  
Done.  
  
Out[12]: sum_paylaod_mass  
38856
```

The print screen a above is a SQL query that display the total payload mass carried by boosters where NASA is customer.

CRS stands for Commercial Resupply Services this indicates that the launches of NASA were sent to the International Space Station (ISS)

Average Payload Mass by F9 v1.1

```
In [15]: %sql SELECT AVG(payload_mass__kg_) AS average_payload_mass_F9_v11
|FROM SPACEXTBL WHERE booster_version like 'F9 v1.1%'

* ibm_db_sa:///jqf40619:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.a
ppdomain.cloud:30699/bludb
Done.

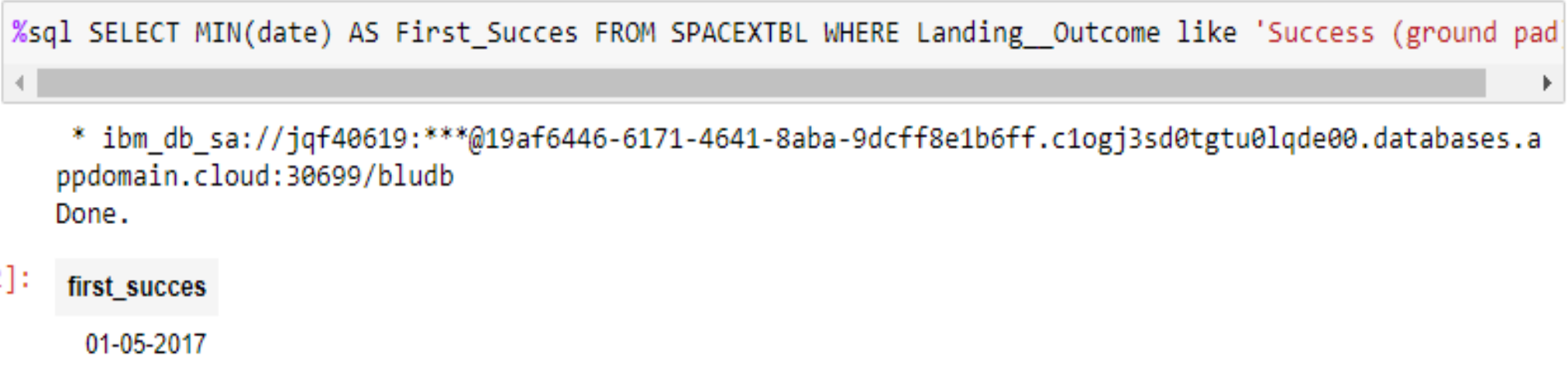
Out[15]: average_payload_mass_f9_v11
          2534
```

The print screen a above is a SQL query with a WHERE CLAUSE displaying the calculated average Payload Mass or Launches where Booster version F9 v1.1. was used.

The average weight of F9 v1.1 Payload Mass is low compared to the Payload Mass range.

First Successful Ground Landing Date

```
In [22]: %sql SELECT MIN(date) AS First_Succes FROM SPACEXTBL WHERE Landing__Outcome like 'Success (ground pad
```



```
* ibm_db_sa://jqf40619:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.a  
ppdomain.cloud:30699/bludb  
Done.  
Out[22]: first_succes  
01-05-2017
```

The print screen above is a SQL query that display the first successful Ground Landing date in history of Space X launches.

Based on the dataset successful landings appear starting from 2018.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [13]: %sql SELECT booster_version
FROM SPACEXTBL WHERE (mission_outcome like 'Success') AND (payload_mass__kg_ BETWEEN 4000 AND 6000)
AND (landing__outcome like 'Success (drone ship)')

* ibm_db_sa://jqf40619:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.a
ppdomain.cloud:30699/bludb
Done.
```

Out[13]:

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

The print screen above is a SQL query that display the four booster versions that had successful drone ship landings with a Payload Mass from 4,000 kilogram to 6,000 kilogram.

Total Number of Successful and Failure Mission Outcomes

```
In [38]: %sql SELECT COUNT(*) FROM spacextbl WHERE mission_outcome LIKE '%Success%'
* ibm_db_sa://jqf40619:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.a
ppdomain.cloud:30699/bludb
Done.

Out[38]: 1
        65

In [39]: %sql SELECT COUNT(*) FROM spacextbl WHERE mission_outcome LIKE '%Failure%'
* ibm_db_sa://jqf40619:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.a
ppdomain.cloud:30699/bludb
Done.

Out[39]: 1
        1
```

The print screen above is a SQL query that display a count of all mission outcomes.

- Space X achieves a success of 99% of the time.
- This could mean that most Failure launches are intended.

Boosters Carried Maximum Payload

```
In [47]: %sql select booster_version, payload_mass__kg_
         from SPACEXTBL where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTBL)

* ibm_db_sa://jqf40619:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.a
ppdomain.cloud:30699/bludb
Done.
```

```
Out[47]:
```

booster_version	payload_mass__kg_
F9 FT B1029.1	9600
F9 FT B1036.1	9600
F9 B4 B1041.1	9600
F9 FT B1036.2	9600
F9 B4 B1041.2	9600
F9 B5B1048.1	9600
F9 B5 B1049.2	9600

The print screen above is a SQL query that use the WHERE clause and MAX function to display the booster versions that carried the highest Payload.

- It seems that all F9 booster version can handle a Payload of maximum 9,600 kilogram.
- There is a likelihood that Payload relates to F9 booster versions.

2015 Launch Records

```
In [51]: %sql select (DATE) as Month_Year, landing__outcome, booster_version, payload_mass__kg_, launch_site
|FROM SPACEXTBL where DATE like '%2015%' AND landing__outcome like 'Failure (drone ship)'

* ibm_db_sa:///jqf40619:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.a
ppdomain.cloud:30699/bludb
Done.
```

```
Out[51]:
```

month_year	landing__outcome	booster_version	payload_mass__kg_	launch_site
10-01-2015	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
14-04-2015	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

The print screen above is a SQL query use the WHERE clause, LIKE AND, and BETWEEN to display the launch records of 2015 where Stage 1 failed to land on a drone ship.

- It seems that all F9 booster version can handle a Payload of maximum 9,600 kilogram.
- The query shows two records in 2015.

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

```
In [58]: %sql select landing__outcome, count(*) as Number_Outcome
        from SPACEXTBL WHERE landing__outcome LIKE '%Failure%' AND DATE BETWEEN '06-04-2010' AND '20-03-2017'
        GROUP BY landing__outcome ORDER BY Number_Outcome Desc
```

* ibm_db_sa://jqf40619:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.apdomain.cloud:30699/bludb
Done.

Out[58]:

landing__outcome	number_outcome
Failure (drone ship)	4

The print screen above is a SQL query that use the COUNT, WHERE clause, BETWEEN, GOUP BY and the ORDER BY clause to display the failed drone ship landings between 2010 and 2017 inclusively.

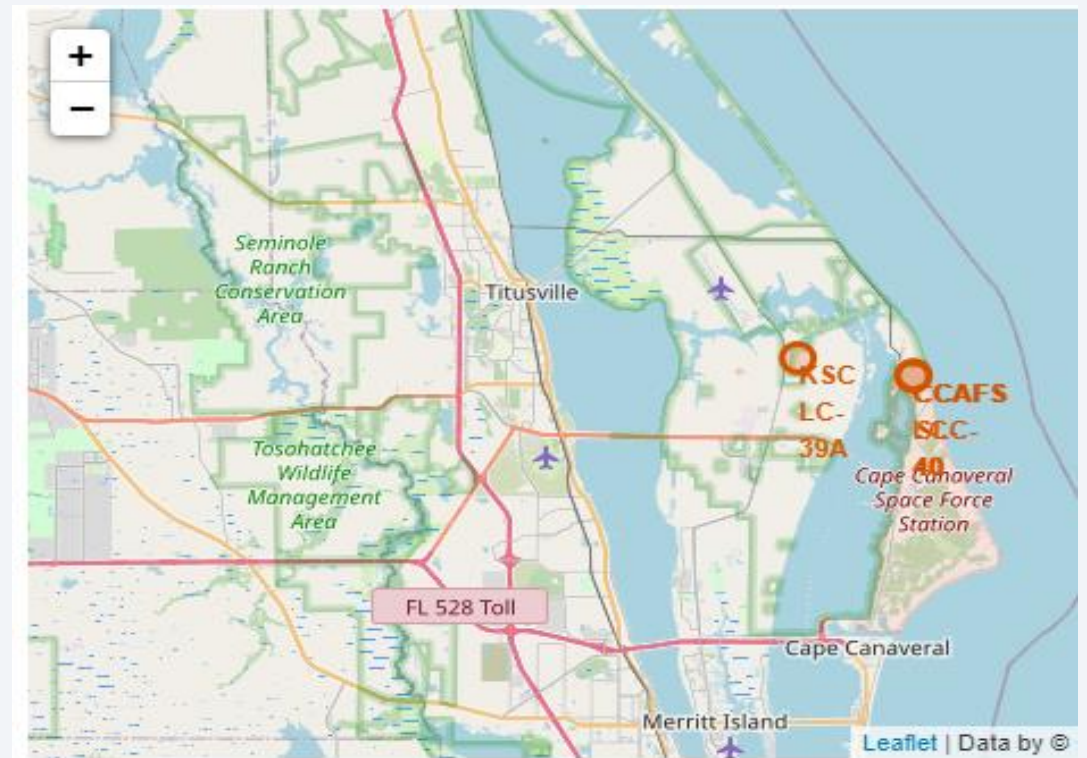
- It seems that there were four failed drone ship landings in that period.

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

Launch Sites Proximities Analysis

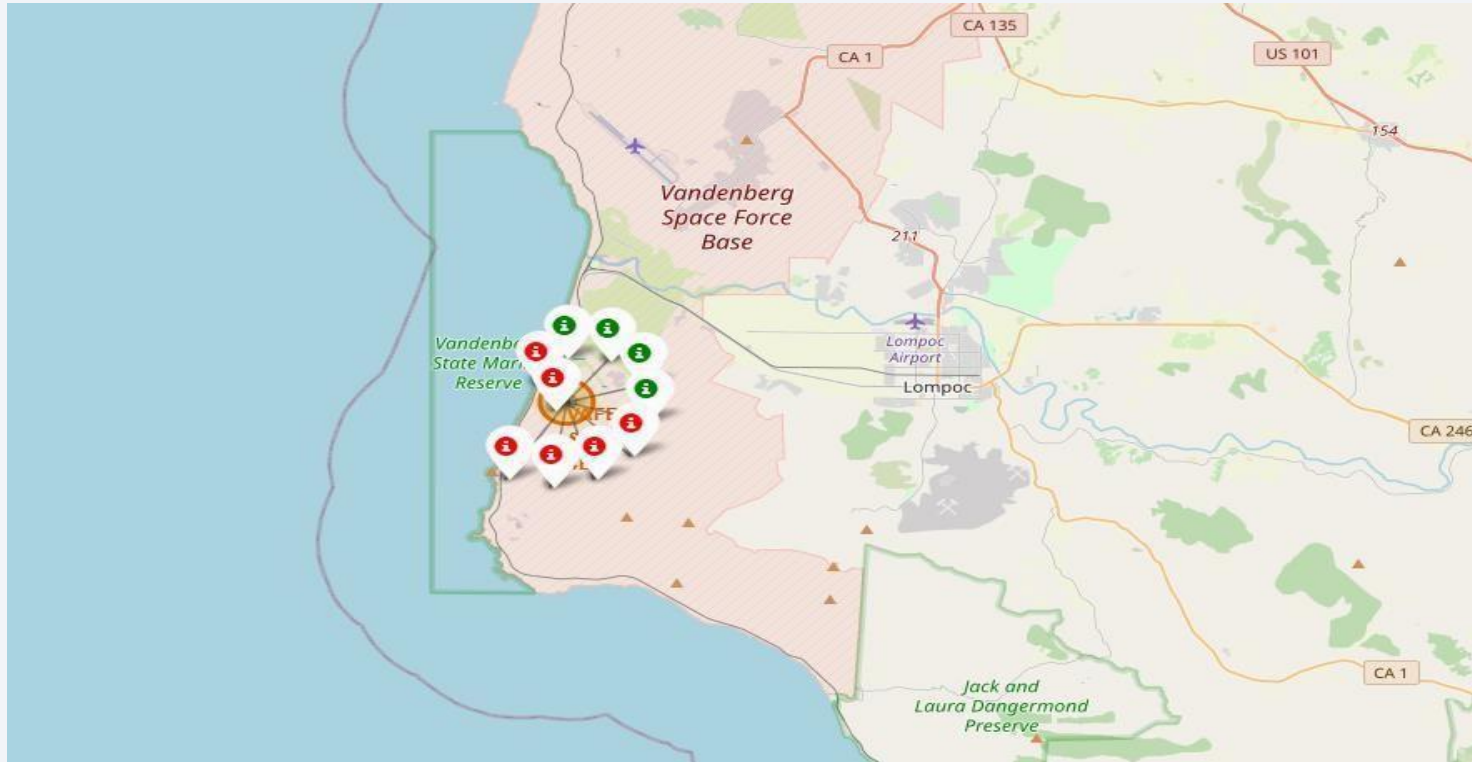
Launch Site Locations



The map on the left side is a global visualization of launch site locations in the United States, on the right is a zoomed visual of the sites in Florida (Orlando).

- All Space X locations in the dataset are in the United States.
- All launch site location are near the ocean.

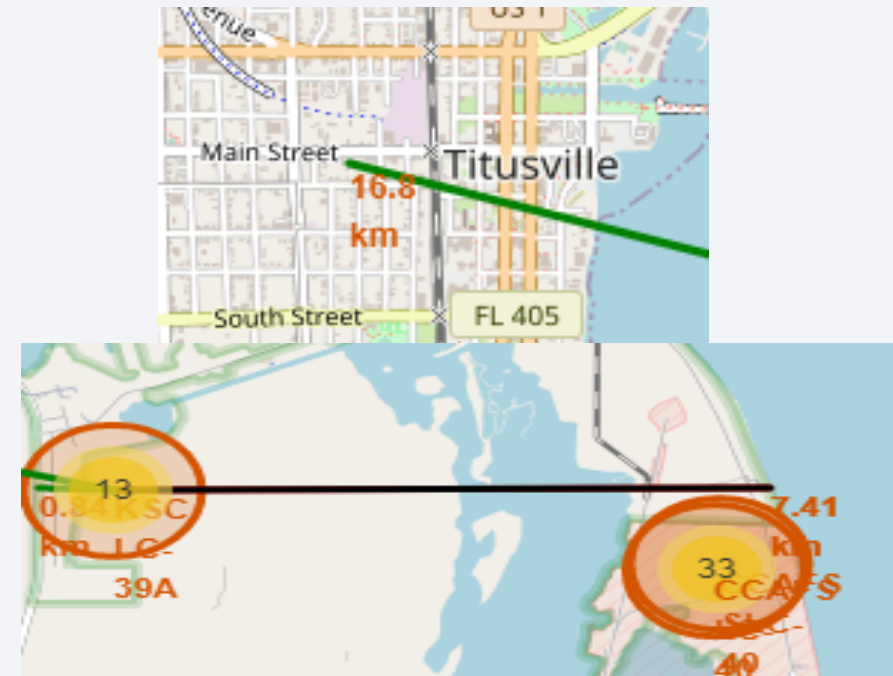
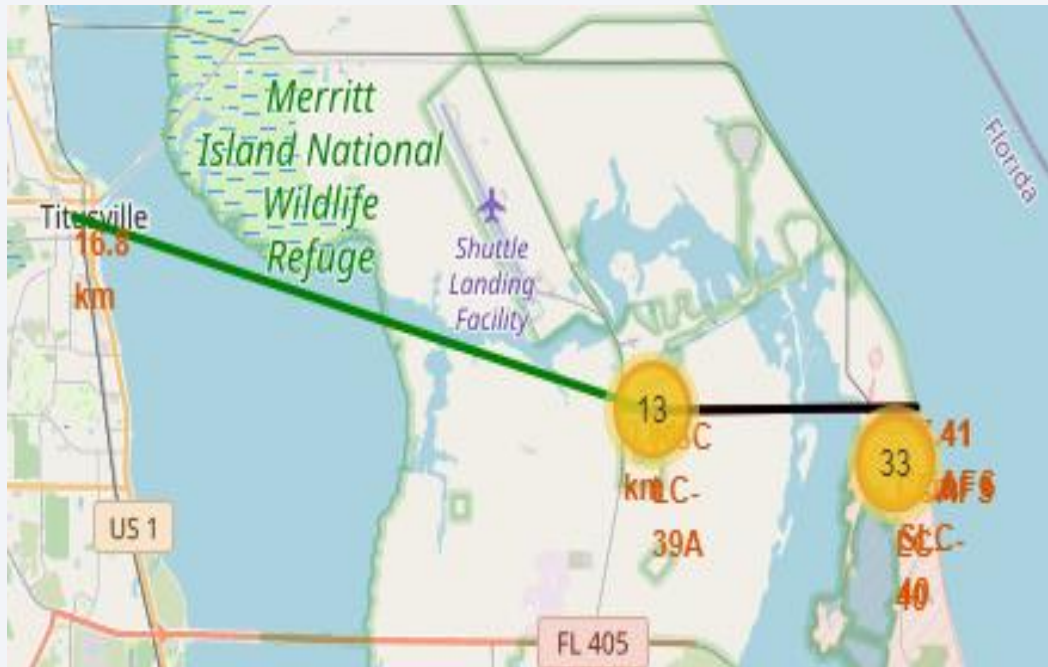
Color – Coded Launch Markers



The above is a visualization of a site including color coded launch markers.

- With clicks on the map it displays green and red icons indicating successful and unsuccessful landings.
- The map shows that launch site VAFB SLC – 4E has six failed and four successful landings.

Key locations and Proximities



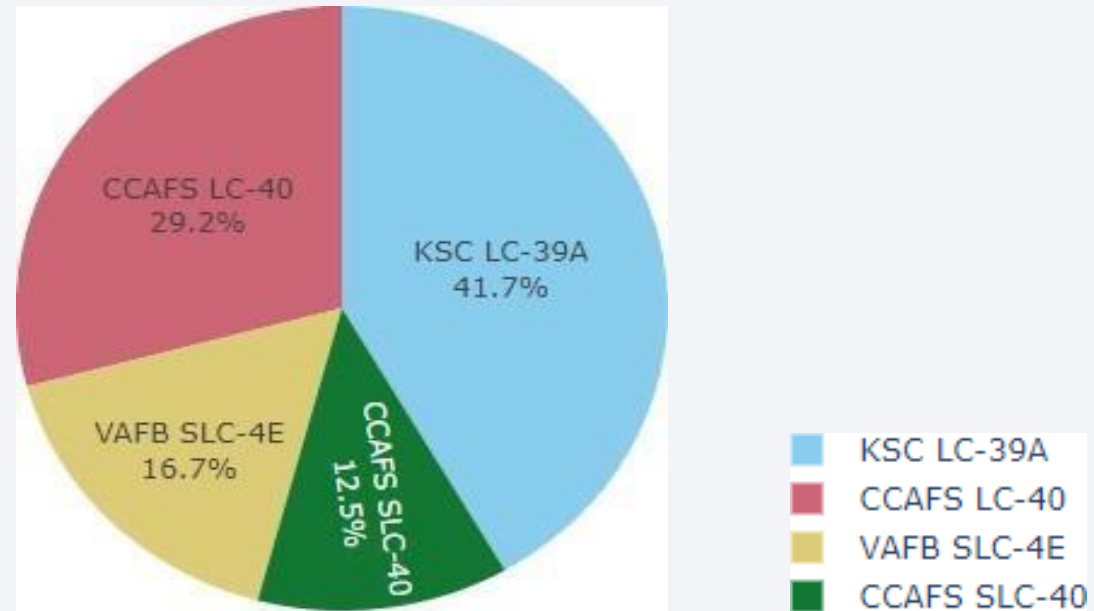
Observed that KSC LC-39A is a launch site that is close to a railway and close to the coastline. Supply transportation can take place by land via train. This launch site is also close to a highway for human and supply transports. In general launch sites are far away from cities so failure launches can safely land in the sea and not in populated areas.



Section 4

Build a Dashboard with Plotly Dash

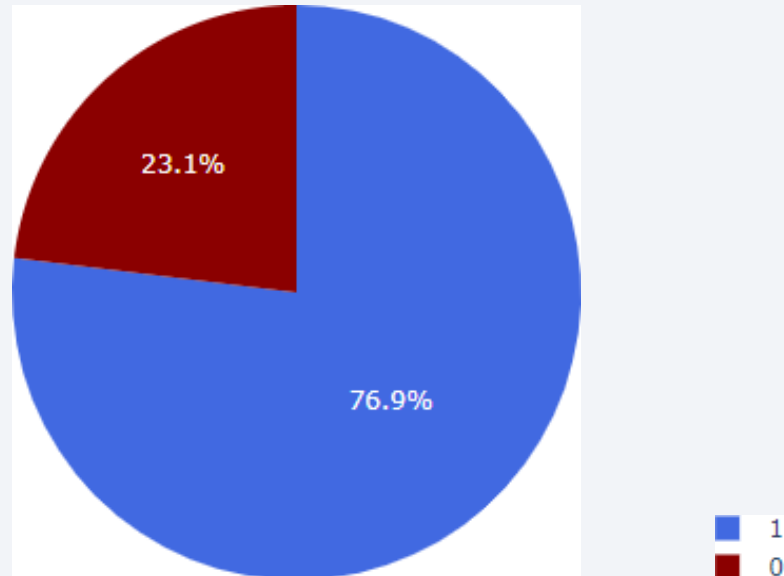
Successful Launches of all Launch Sies



Observed that KSC LC-39A had with (41%) the most successful launches from all launch sites.

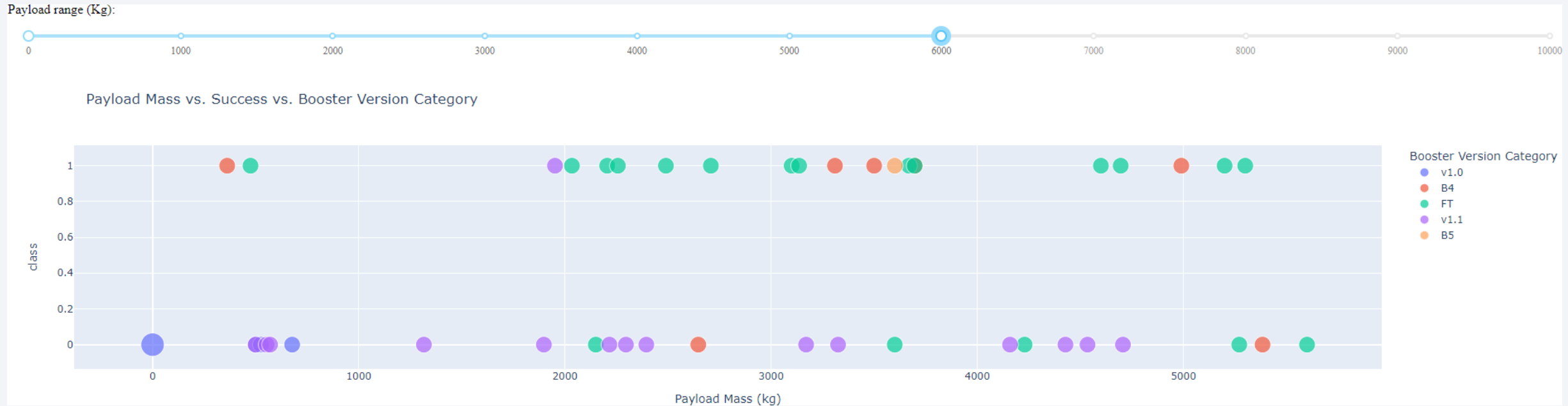
Highest Success rate of Launch Site

KSC LC-39A Success Rate (blue=success)



A close observation shows that launch site **KSC LC – 39A** achieved a success of 76.9% and a failure rate of 23.1%.

Paylaod Mass vs Succes vs. Booster Version Category



The Scatter Plot shows the relationship between Payload Mass, Success Rate and the Booster Version. Observe that the Success Rate of low weighted Payload is high compared to Booster Versions with a heavy weighted Payload.

Section 5

Predictive Analysis (Classification)

Predictive Analysis (Classification)

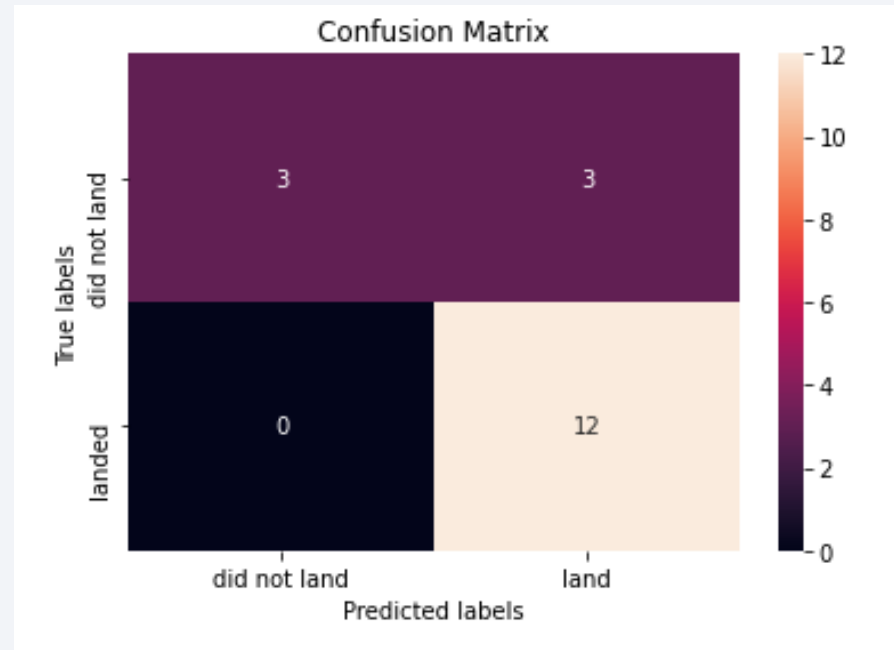
```
models = {'KNeighbors':knn_cv.best_score_,
          'DecisionTree':tree_cv.best_score_,
          'LogisticRegression':logreg_cv.best_score_,
          'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)

Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

There where four classification models analyzed; KNN, Decision Tree, Logistic Regression and Support Vector Machine. The Decision Tree classifier is the model with the highest (predictive) accuracy.

Confusion Matrix



The Confusion Matrix shows that when the Decision Tree classifier is used to predict. The classifier shows a clear differentiation between predicted classes. The matrix show three false positives that could lead to a error in predictions, because unsuccessful landings could be treated and marked as successful landings.

Conclusions

After completing the Data Science journey the following conclusion can be drawn.

- The flight analysis show that when the number of flights increases the effect is a greater success rate for that launch site.
- The Success Rate of Space X launches increased in the period 2013 till 2020.
- The Orbits with the highest Success Rate are; VLEO, GEO, HEO, SSO, ES-L1.
- Launch Site: KSC LC-39 A has the most success launches compared to all sites.
- The Decision Tree Classifier is the best performing machine learning algorithm.

Thank you!

