Winning Space Race with Data Science

Eduardo Catarino
08/02/2020

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection: Web Scrapping and Webservices REST API Calls
  - Data wrangling: Cleaning, Processing, Transforming, Enriching and Storing
  - EDA: SQL analysis, Graphic Analysis, Ploty interactive, Folium Maps
  - Predictive Analysis: Best estimator discovery with Scikit Learn
- Summary of all results
  - Best estimator found: Decision Tree

Introduction

Falcon 9 is a reusable, two-stage rocket designed and manufactured by SpaceX for the reliable and safe transport of people and payloads into Earth orbit and beyond.

Falcon 9 is the world's first orbital class reusable rocket.

Reusability allows SpaceX to refly the most expensive parts of the rocket, which in turn drives down the cost of space access.

https://www.spacex.com/vehicles/falcon-9/

- The goal of this project is, with the database of Falcon 9 space flight information, predict if a new mission will land successfully based on its current parameters.
- All the methodologies and processes used are under a Data Science approach.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web Scrapping and REST API calls
- Perform data wrangling
  - Cleaning, Processing, Transforming, Enriching and Storing with Python, Pandas
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Search of best estimator with SciKit Learn GridSearchCV for the estimator: Log Reg, SVM, Decision Tree, K Nearest Neighbors

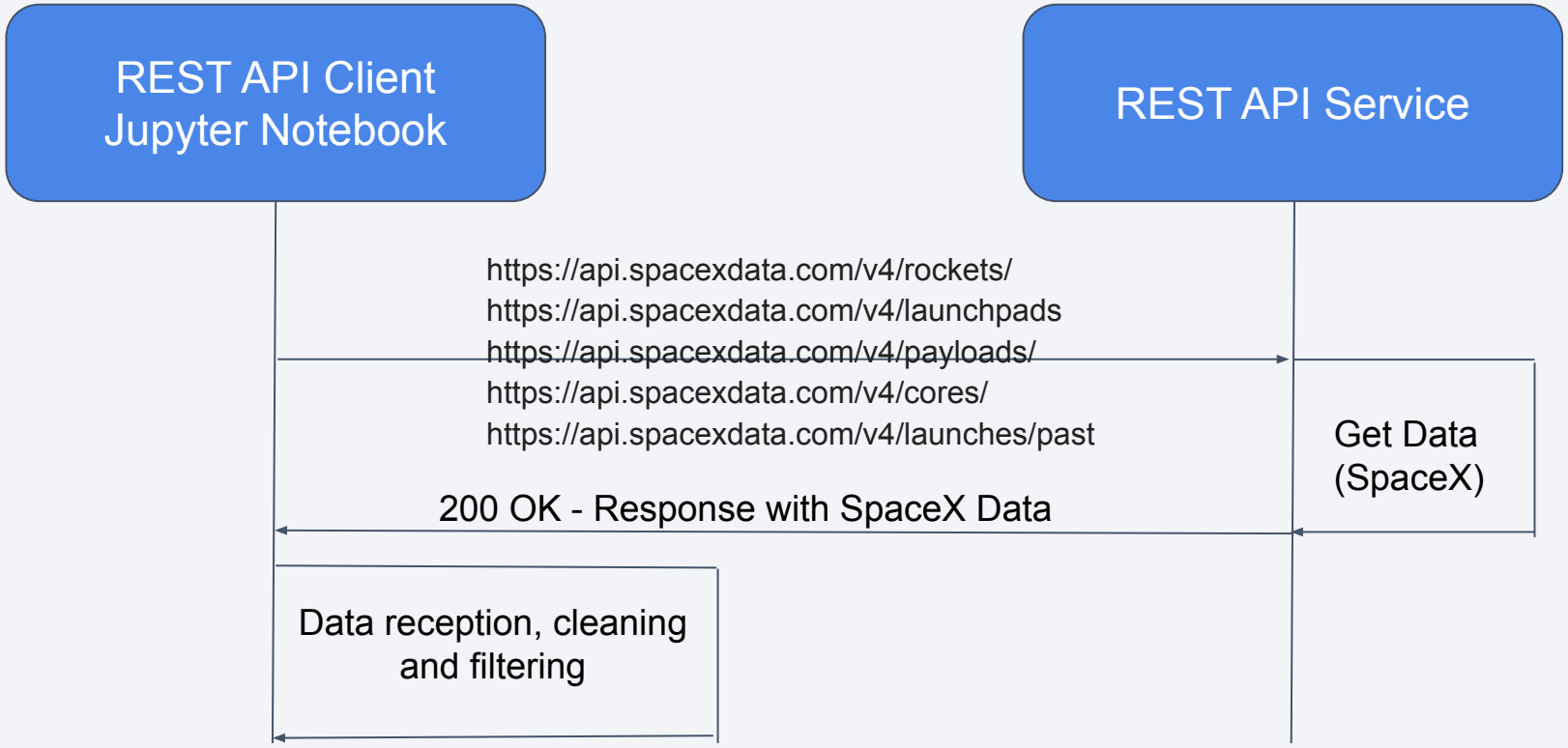
Data Collection

For this project data was collected by the following methods:

- Through calls to the SpaceX REST API
- Web Scraping of SpaceX information web pages

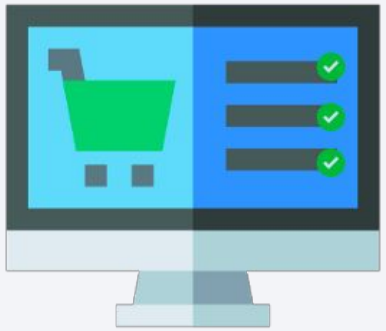
Data Collection – Calls to SpaceX REST API

Data collection procedure description - SpaceX Rest API



Jupyter Notebook Location on GitHub:
https://github.com/smartlearningci/capstone_ibm_data_science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



Targeting the website for data collection

Extract and analyse the source code

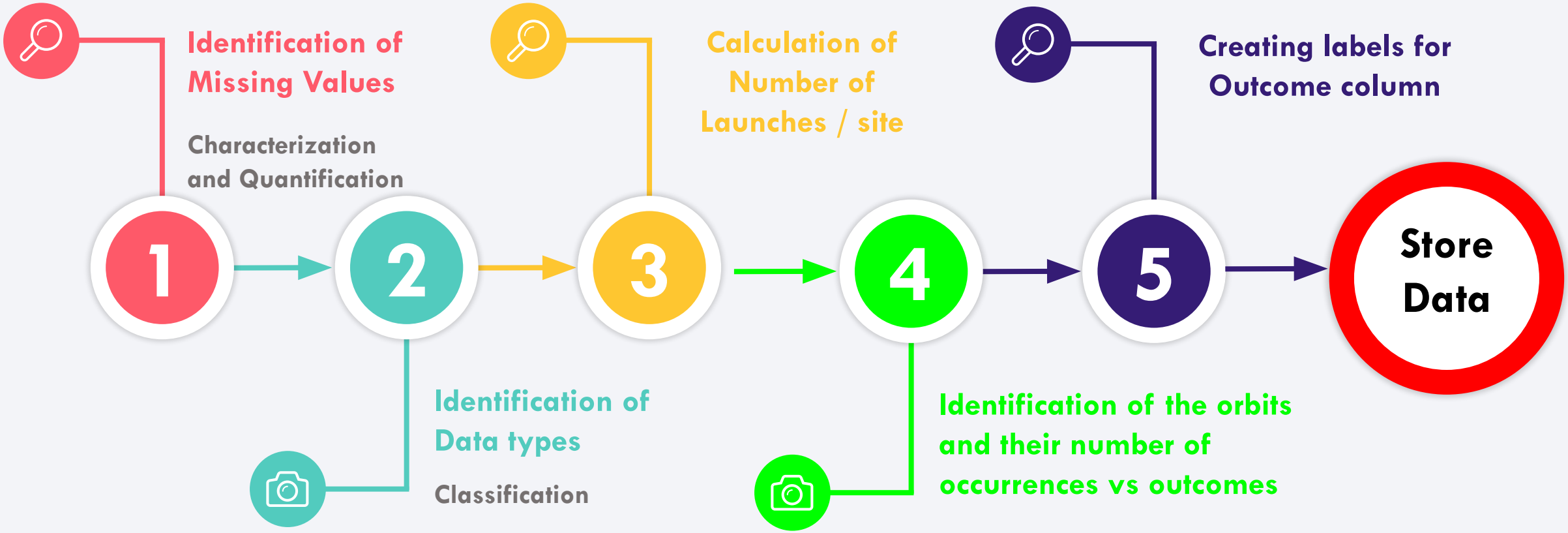
Scrape the data with Python and BeautifulSoup

Data cleaning and filtering

[Wikipedia Falcon 9 Heavy Launches]

Jupyter Notebook Location on GitHub:
https://github.com/smartlearningci/capstone_ibm_data_science/blob/main/jupyter-labs-webscraping%20(1)%20(1).ipynb

Data Wrangling



Jupyter Notebook Location on GitHub:
https://github.com/smartlearningci/capstone_ibm_data_science/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

FlightNumber vs. PayloadMass

EDA with Data Visualization

For Exploratory Data Analysis purposes where plotted the following charts:

Flight Number vs PayLoad Mass
To discover the evolution of the payload mass in time
1

Flight Number vs Launch Site
To understand the placement of the launches in time
2

PayLoad Mass vs Launch Site
To understand the influence of the payload mass on the localization of the launches
3

Success Rate vs Orbit type
To understand dependence of orbit type on the launch success
4

Flight Number vs Orbit Type
To understand the evolution of orbit types in time
5

PayLoad Mass vs Orbit type
To understand the influence or restrictions of the payload mass on the flight orbits
6

Success Trend
To understand how the success of the launches evolution in time
7

Jupyter Notebook Location on GitHub:
https://github.com/smartlearningci/capstone_ibm_data_science/blob/main/jupyter-labs-eda-dataviz%20(1)%20(1).ipynb

EDA with SQL

For Exploratory Data Analysis purposes the following charts were performed:

- Task 1: “select DISTINCT LAUNCH_SITE FROM NQV27042.SPACEXTBL”
- Task 2: “select * FROM NQV27042.SPACEXTBL where LAUNCH_SITE LIKE 'CCA%' Limit 5”
- Task 3: “select SUM(PAYLOAD_MASS__KG_) FROM NQV27042.SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'”
- Task 4: “select AVG(PAYLOAD_MASS__KG_) FROM NQV27042.SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
- Task 5: “select min(DATE) FROM NQV27042.SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)'”
- Task 6: “select DISTINCT BOOSTER_VERSION FROM NQV27042.SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 400 AND 6000;
- Task 7: “select (select COUNT(MISSION_OUTCOME) FROM NQV27042.SPACEXTBL where MISSION_OUTCOME = 'Success') AS NUMBER_OF_SUCCESS,(select COUNT(MISSION_OUTCOME) FROM NQV27042.SPACEXTBL where MISSION_OUTCOME != 'Success') AS NUMBER_OF_FAILURES FROM NQV27042.SPACEXTBL LIMIT 1”
- Task 8: “SELECT DISTINCT BOOSTER_VERSION FROM NQV27042.SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM NQV27042.SPACEXTBL)”
- Task 9: “SELECT BOOSTER_VERSION, LAUNCH_SITE FROM NQV27042.SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' and YEAR(DATE) = 2015”
- Task 10: “SELECT LANDING__OUTCOME, COUNT(*) FROM NQV27042.SPACEXTBL where DATE Between '2010-06-04' and '2017-03-20' GROUP BY LANDING__OUTCOME ORD”

Jupyter Notebook Location on GitHub:
https://github.com/smartlearningci/capstone_ibm_data_science/blob/main/jupyter-labs-eda-sql-coursera%20(3)%20(1).ipynb

Build an Interactive Map with Folium

- Map markers are helper objects to help to understand the problem better adding to a map contexted information.
- In this project was used the Folium Map to describe the site launch locations
- The Folium map markers used for this project were:
    - Circles
    - Lines
    - Icons
    - Cluster Markers
- All markers were added to the map to help to:
    - understand the location of the launch sites
    - characterize the success of launches on each site
    - understand and characterize the proximities characteristics

Jupyter Notebook Location on GitHub:
https://github.com/smartlearningci/capstone_ibm_data_science/blob/main/lab_jupyter_launch_site_location%20(2).ipynb

Build a Dashboard with Plotly Dash

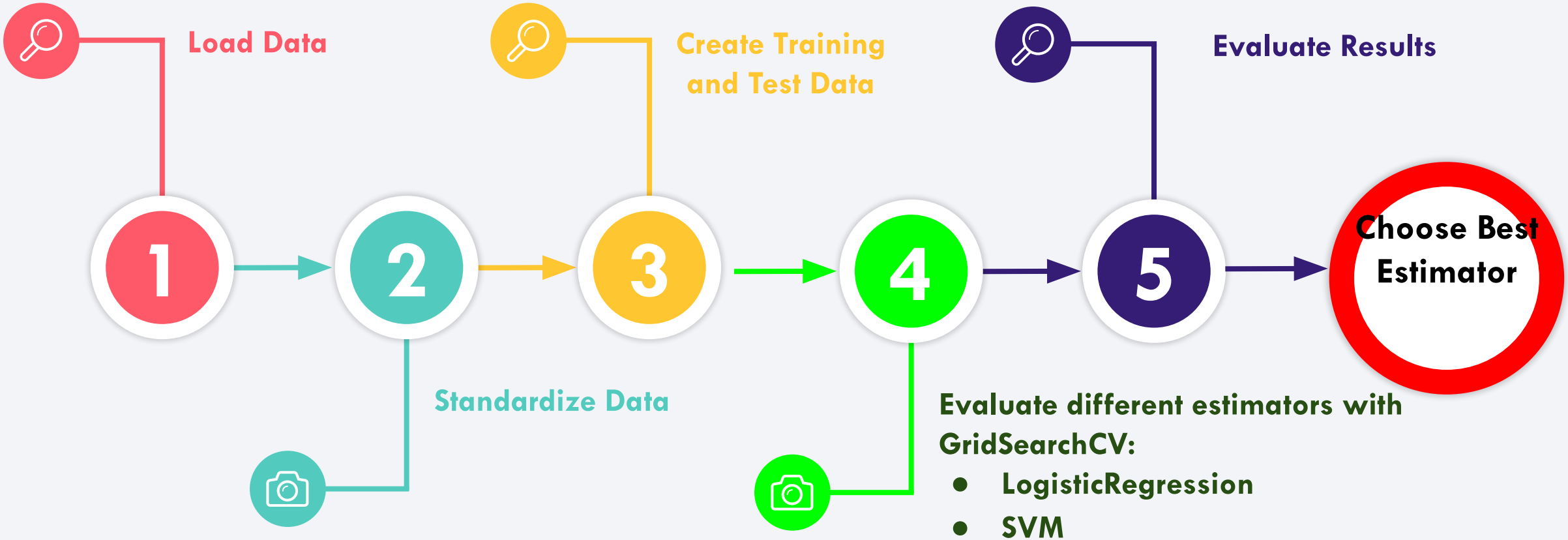
To understand and clarify some of the existing relations between a Plotly was designed. With this purpose the following interactive plots were constructed:

- Total number (%) of success launches per site. With this graph is possible to evaluate the overall performance of launches by site.

- Total of success launches per value of payload mass (with the payload mass range configurable by slider). With this graph is possible to evaluate per site the performance of launches per payload mass. With the configurable slider is possible to the user have a finer tuning over the input parameters.

Jupyter Notebook Location on GitHub:
https://github.com/smartlearningci/capstone_ibm_data_science/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



Jupyter Notebook Location on GitHub:
https://github.com/smartlearningci/capstone_ibm_data_science/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(2).ipynb

Results

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

EDA Results

Here you can add some text as explanation and consider replacing these demo text with your one

Interactive Analytics

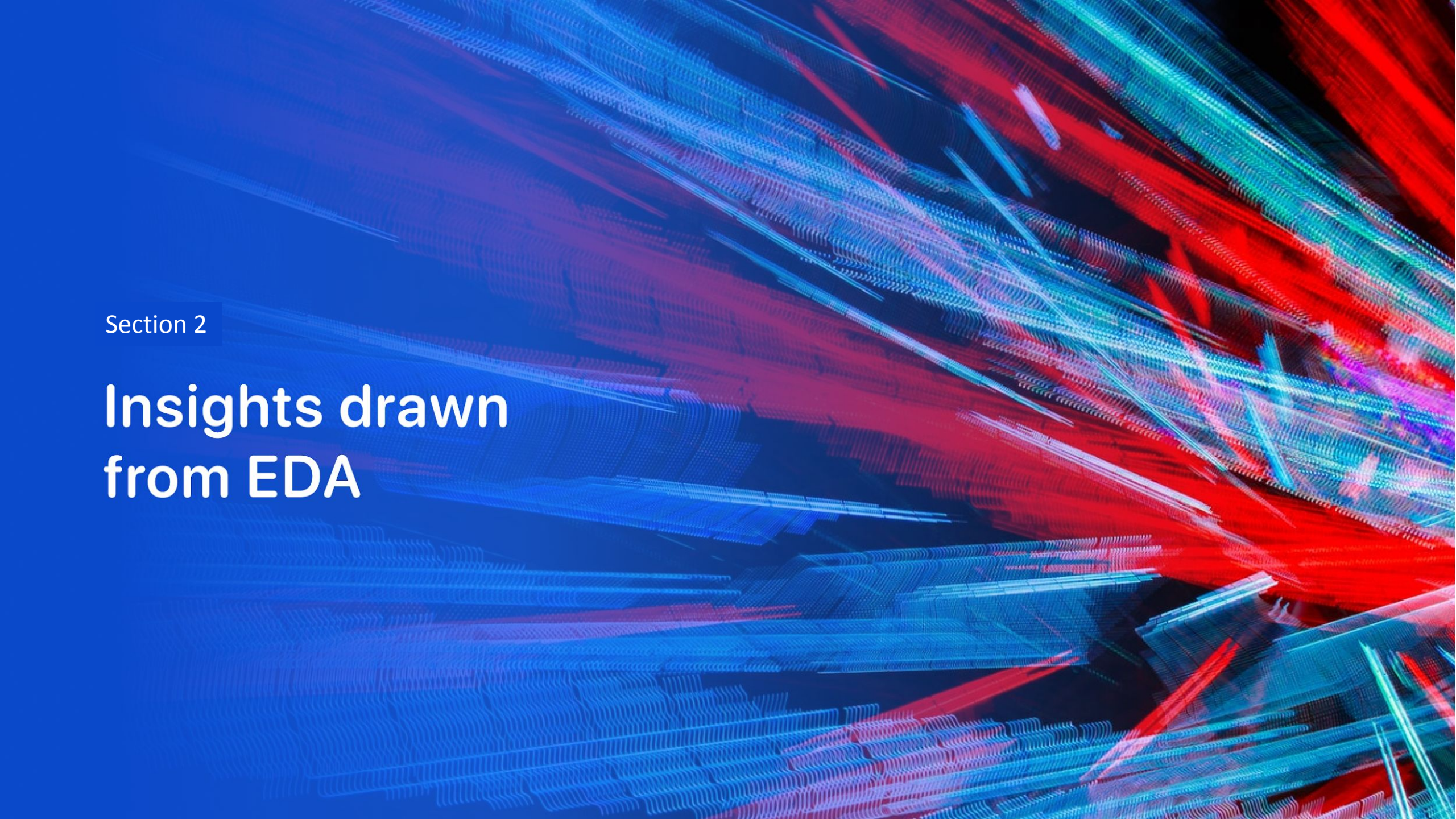
Here you can add some text as explanation and consider replacing these demo text with your one

Predictive Analysis Results

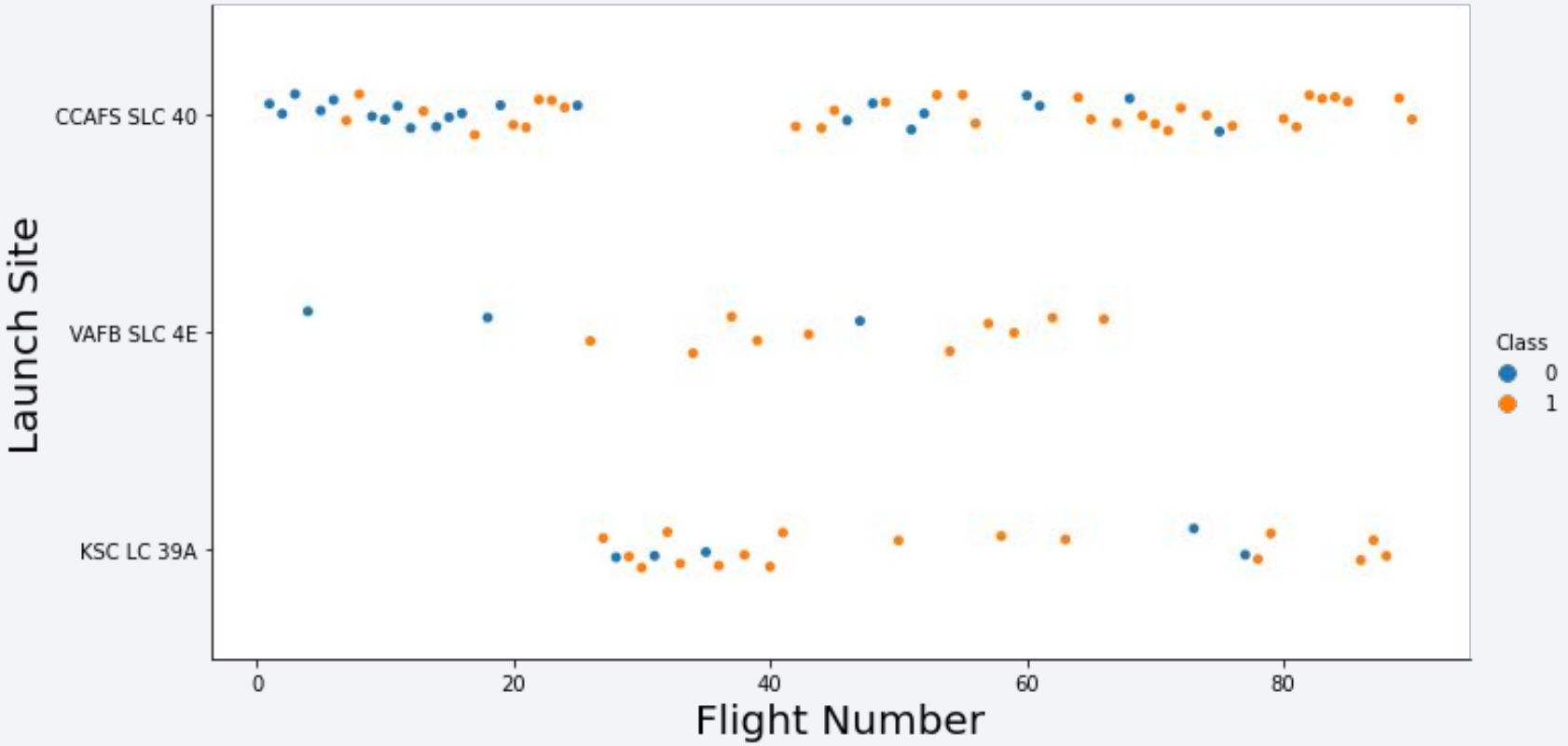
Here you can add some text as explanation and consider replacing these demo text with your one

Section 2

Insights drawn from EDA



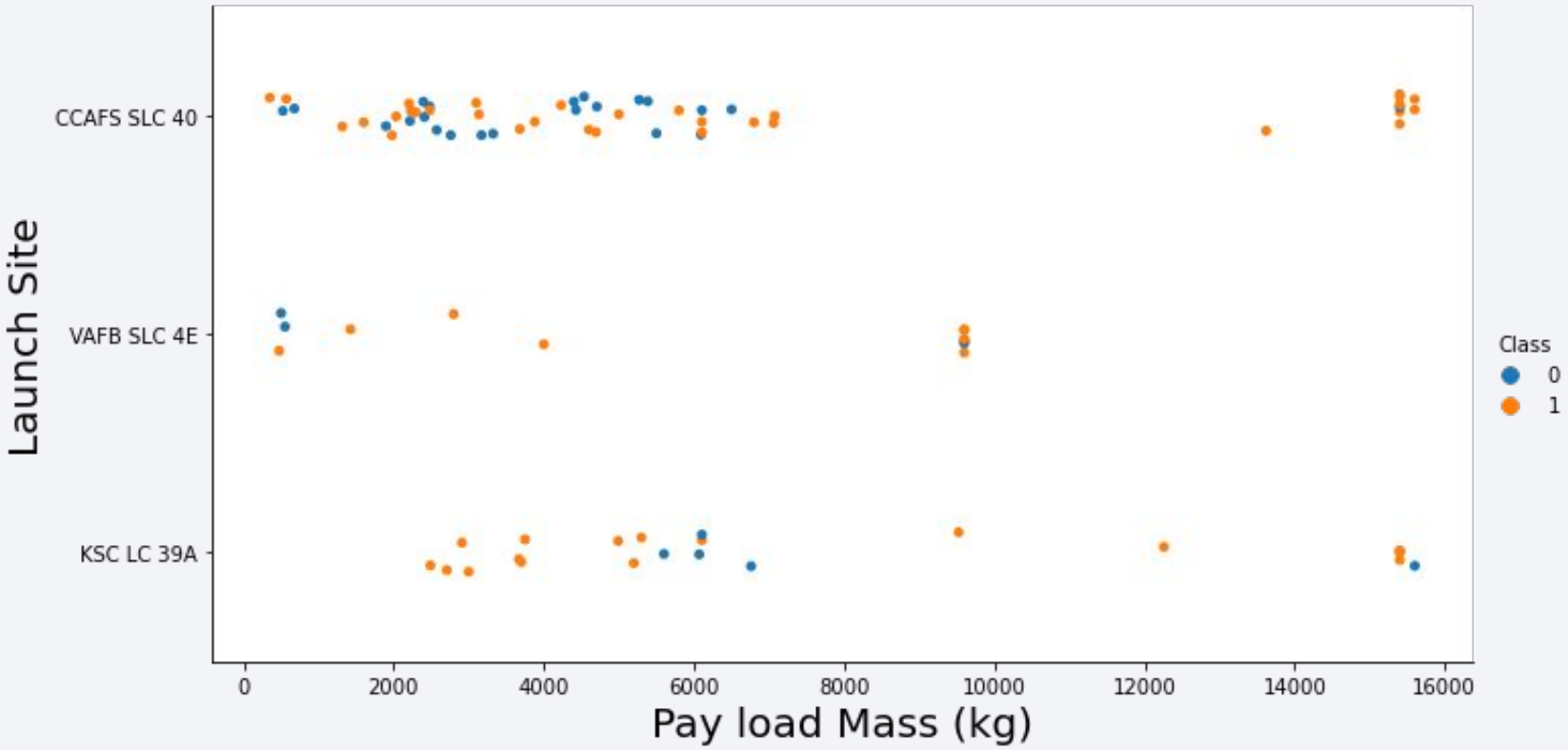
Flight Number vs. Launch Site



From the plot drawn is possible to conclude that:

- CCAFS SLC 40 is the main launch site used
- VAFB SLC 4E is the leat site used
- CCAFS SLC 40 has been used during the all launch process
- VAFB SLC 4E has been used for the first 3 quarters of the launching process
- KSC LC-39A has been used for last 3 quarters of the launching process

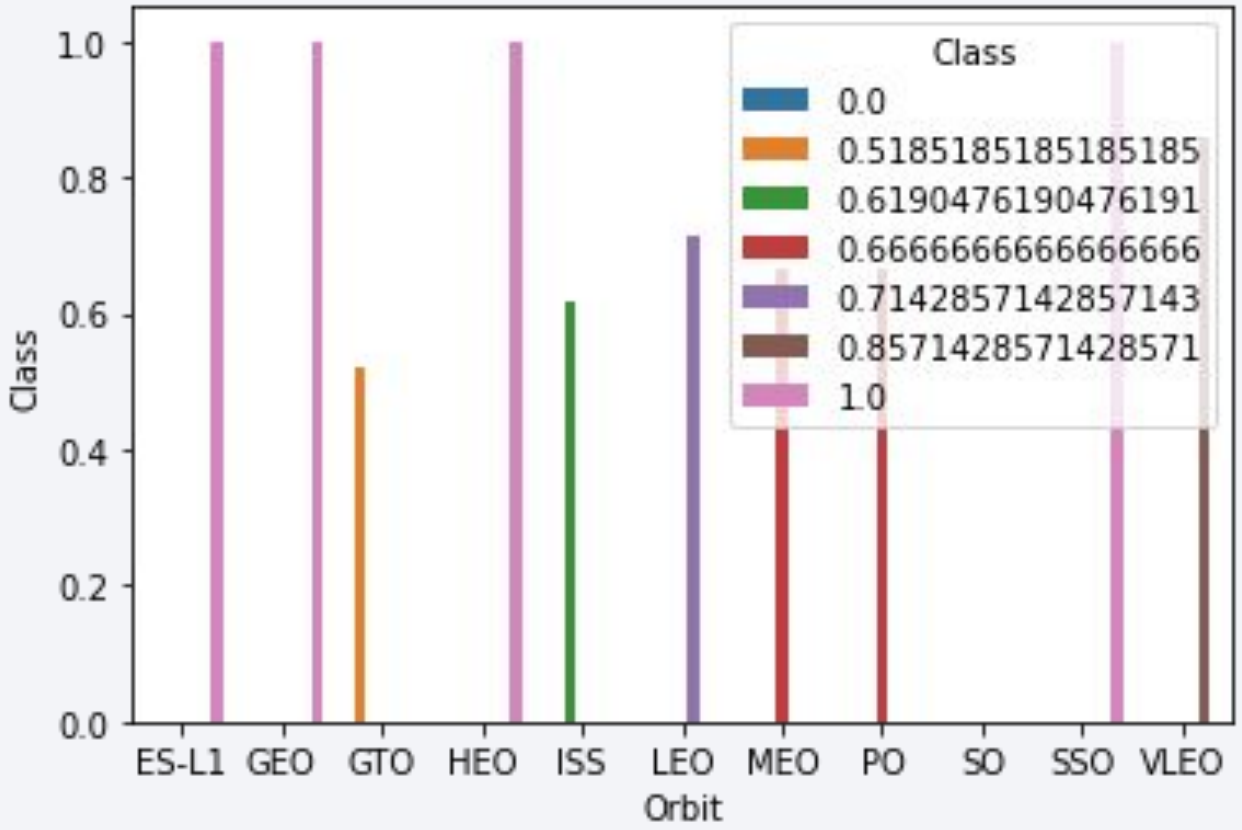
Payload vs. Launch Site



From the plot drawn is possible to conclude that:

- Most of the launches where done with less than half the maximum value of payload mass used
- CCAFS SLC 40 is the main launch site used
- VAFB SLC 4E is used with mainly for low values of payload mass
- KSC LC-39A has been used less times than CCAFS SLC 40 but with approximately the same profile of the CCAFS SLC 40 site

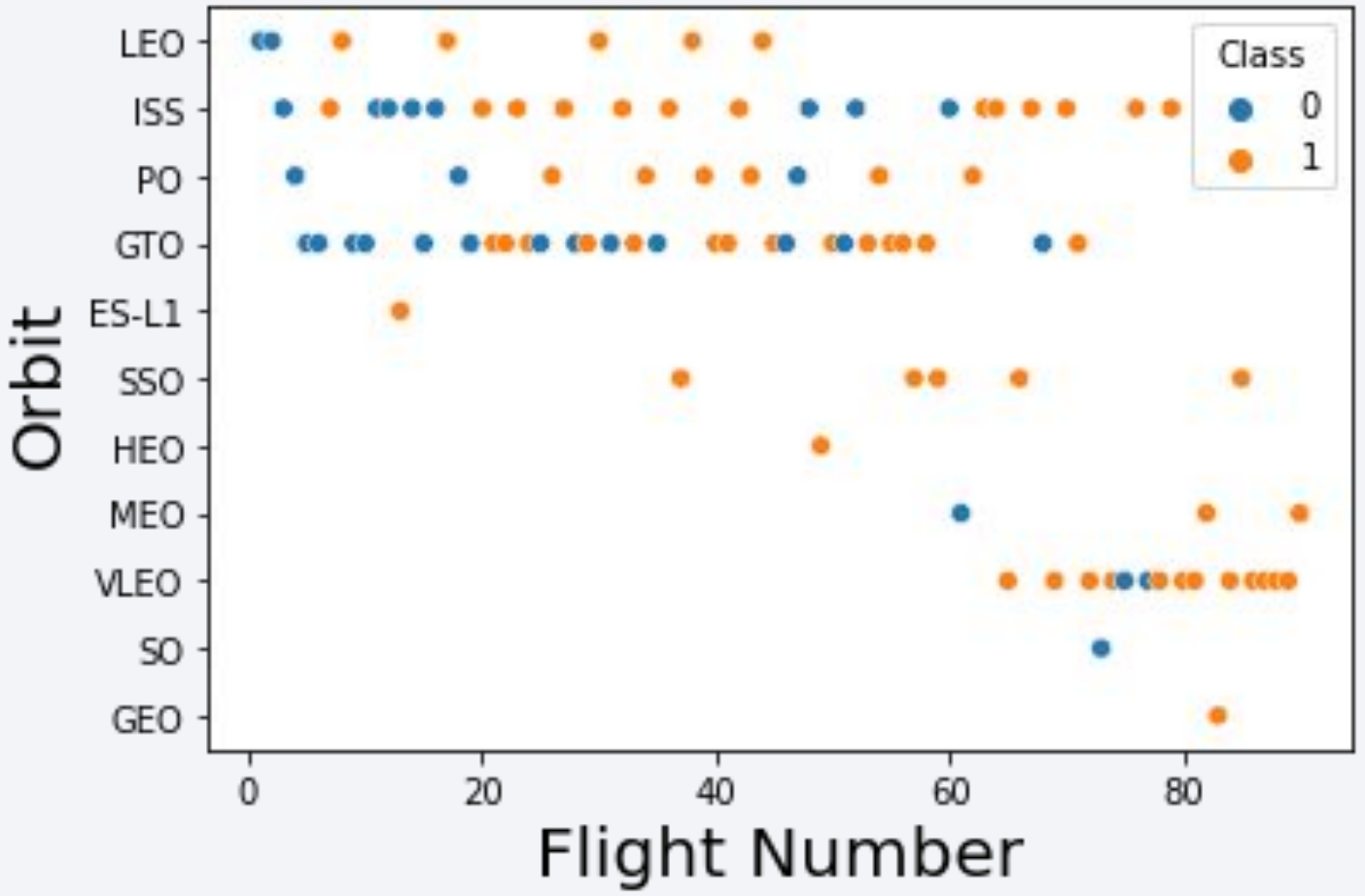
Success Rate vs. Orbit Type



From the plot drawn is possible to conclude that:

- ES-L1, GEO, HEO and SSO orbit types have the best performance
- SO has the worst performance

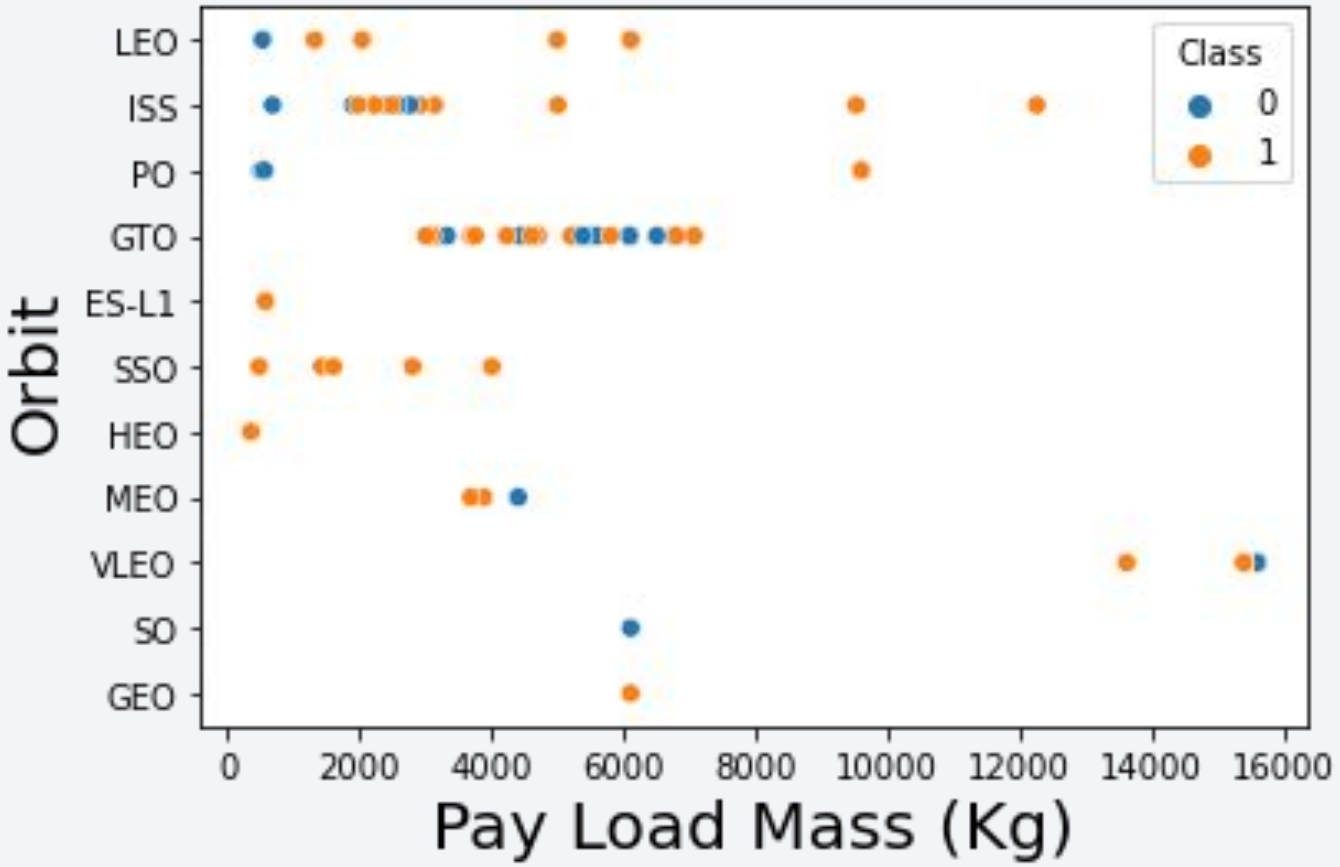
Flight Number vs. Orbit Type



From the plot drawn is possible to conclude that:

- LEO orbit has been used for the first half of the launches
- ISS orbit has been used uniformly for launches during the analysed period
- PO and GTO orbits were used for the first 70% of the analysed period
- ES-L1, SSO, HEO, MEO, SO, GEO orbits may be considered outliers since than have been used only less than 5 times
- VLEO orbit have been used only for the last quarter of the analysed period

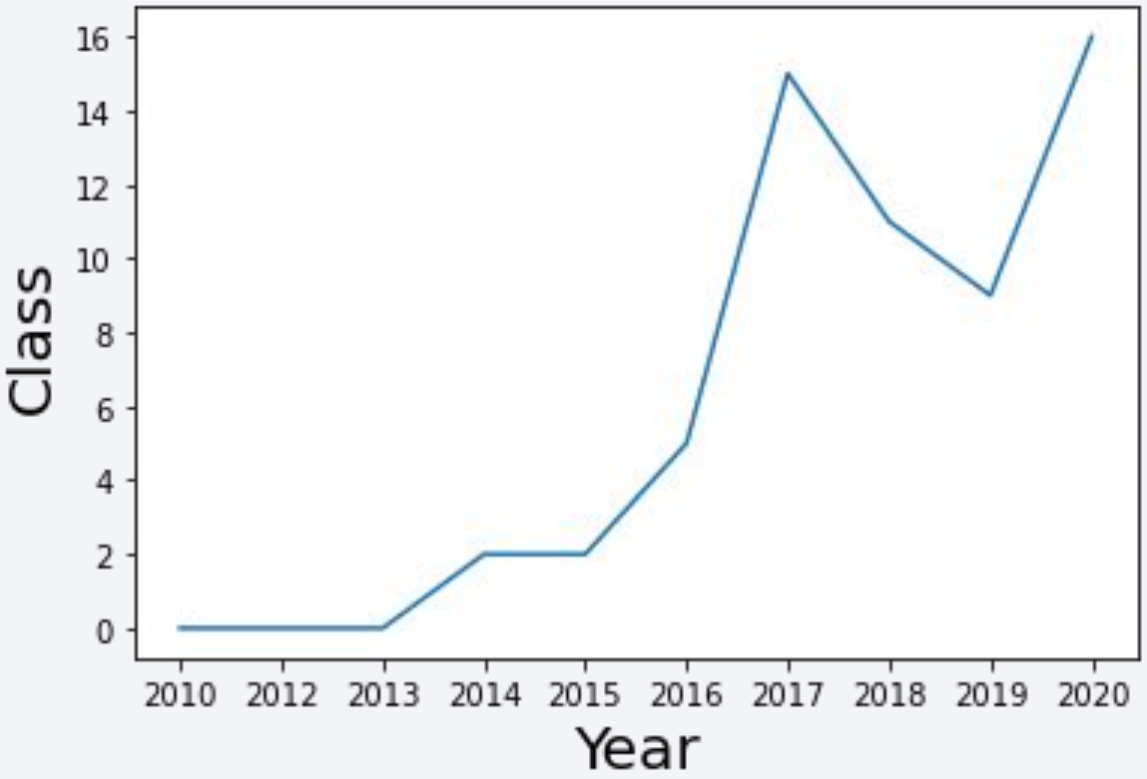
Payload vs. Orbit Type



From the plot drawn is possible to conclude that:

- Only six launches had the payload mass higher than the half of the maximum value.(2 - ISS, 1 - PO, 3 - VLEO)
- The VLEO orbit have been used for the higher values of the payload mass

Launch Success Yearly Trend



From the plot drawn is possible to conclude that:

- The success of the launches have significantly improved in the last 5 years, however after a small drawback between 2017 and 2019, the performance have now recovered.

All Launch Site Names

To retrieve all site names the following sql query was performed:

“

“select DISTINCT LAUNCH_SITE FROM NQV27042.SPACEXTBL”

```
%sql  select DISTINCT LAUNCH_SITE FROM NQV27042.SPACEXTBL
 * ibm_db_sa://nqv27042:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

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

Note: The figure is self explanatory

Launch Site Names Begin with 'CCA'

To retrieve all site names beginning with ‘CCA’ the following sql query was performed:

“

“select * FROM NQV27042.SPACEXTBL where LAUNCH_SITE LIKE 'CCA%' Limit 5”

Task 2

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

```
%sql  select * FROM NQV27042.SPACEXTBL where LAUNCH_SITE LIKE 'CCA%' Limit 5
```

```
 * ibm_db_sa://nqv27042:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

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

Note: The figure is self explanatory

Total Payload Mass

To retrieve the total payload mass the following sql query was performed:

“

“select SUM(PAYLOAD_MASS__KG_) FROM NQV27042.SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'”

”

Task 3

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

```
%sql  select SUM(PAYLOAD_MASS__KG_) FROM NQV27042.SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://nqv27042:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

1
45596

Note: The figure is self explanatory

Average Payload Mass by F9 v1.1

To retrieve the average of payload mass vy F9 v1.1 the following sql query was performed:

“

“select AVG(PAYLOAD_MASS__KG_) FROM NQV27042.SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'”

”

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%sql  select AVG(PAYLOAD_MASS__KG_) FROM NQV27042.SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://nqv27042:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

1
2928

Note: The figure is self explanatory

First Successful Ground Landing Date

To retrieve the first successful ground landing date the following sql query was performed:

“select min(DATE) FROM NQV27042.SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)'”

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql  select min(DATE) FROM NQV27042.SPACEXTBL WHERE  LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://nqv27042:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

1
2015-12-22

Note: The figure is self explanatory

Successful Drone Ship Landing with Payload between 4000 and 6000

To retrieve the successful drone ship landing with 4000 <= Payload <= 6000 the following sql query was performed:

“select DISTINCT BOOSTER_VERSION FROM NQV27042.SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 400 AND 6000;”

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql  select DISTINCT BOOSTER_VERSION FROM NQV27042.SPACEXTBL WHERE  LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BET
```

```
 * ibm_db_sa://nqv27042:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

booster_version
F9 B4 B1042.1
F9 B5 B1046.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1031.2
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1038.1

Note: The figure is self explanatory

Total Number of Successful and Failure Mission Outcomes

To retrieve the total number of successful and failure mission outcomes the following sql query was performed:

“

“select (select COUNT(MISSION_OUTCOME) FROM NQV27042.SPACEXTBL where MISSION_OUTCOME = 'Success') AS NUMBER_OF_SUCCESS,(select COUNT(MISSION_OUTCOME) FROM NQV27042.SPACEXTBL where MISSION_OUTCOME != 'Success') AS NUMBER_OF_FAILURES FROM NQV27042.SPACEXTBL LIMIT 1 “

Task 7

List the total number of successful and failure mission outcomes

```
%sql select (select COUNT(MISSION_OUTCOME) FROM NQV27042.SPACEXTBL where MISSION_OUTCOME = 'Success') AS NUMBER_OF_SUCCESS,(select COUNT
```

```
 * ibm_db_sa://nqv27042:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

number_of_success	number_of_failures
99	2

Note: The figure is self explanatory

Boosters Carried Maximum Payload

To retrieve the nooster carried maximum payload the following sql query was performed:

“

“SELECT DISTINCT BOOSTER_VERSION FROM NQV27042.SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM NQV27042.SPACEXTBL) “

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM NQV27042.SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM NQV27042.SPAC
```

```
 * ibm_db_sa://nqv27042:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
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

Note: The figure is self explanatory

2015 Launch Records

To retrieve the 2015 Launch records the following sql query was performed:

“

“SELECT LANDING__OUTCOME, COUNT(*) FROM NQV27042.SPACEXTBL where DATE Between '2010-06-04' and '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY COUNT(*) DESC“

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM NQV27042.SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' and YEAR(DATE) = 2015
```

```
 * ibm_db_sa://nqv27042:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Note: The figure is self explanatory

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

To retrieve Rank Landing Outcomes the following sql query was performed:
“SELECT LANDING__OUTCOME, COUNT(*) FROM NQV27042.SPACEXTBL where DATE Between '2010-06-04' and '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY COUNT(*) DESC

Task 10

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

```
%sql SELECT LANDING__OUTCOME, COUNT(*) FROM NQV27042.SPACEXTBL  where DATE Between '2010-06-04' and '2017-03-20'  GROUP BY LANDING__OUTC
```

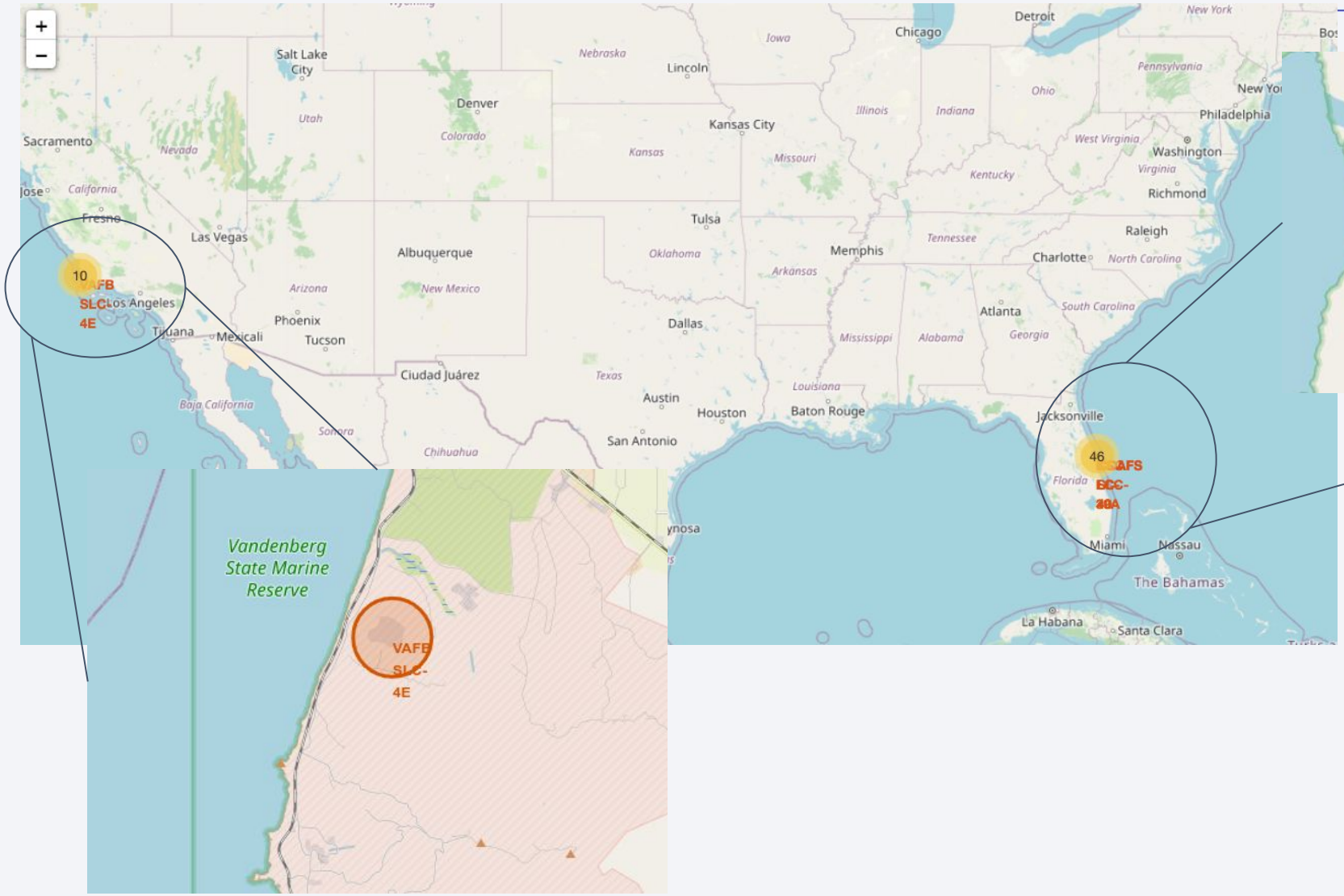
```
 * ibm_db_sa://nqv27042:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

landing__outcome	2
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

Note: The figure is self explanatory


Section 3
Launch Sites
Proximities Analysis

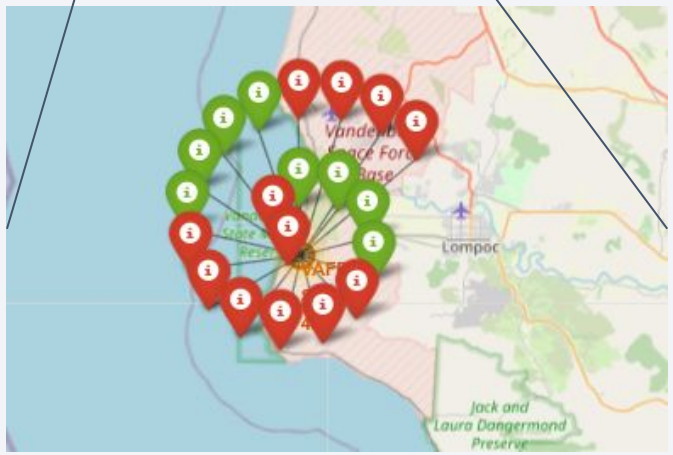
SpaceX Falcon 9 site launch locations

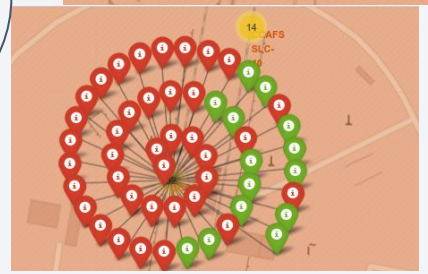


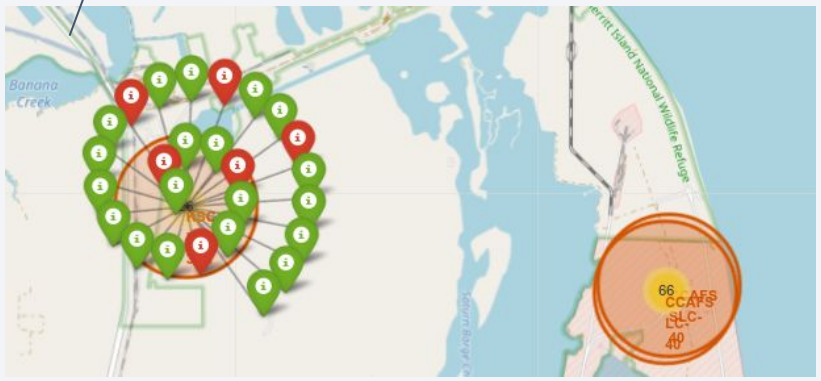
From the Folium Map drawn is possible to conclude that:

- All site locations are in the coast near to the Atlantic Ocean and Pacific Ocean

Success/Failure Launches per site



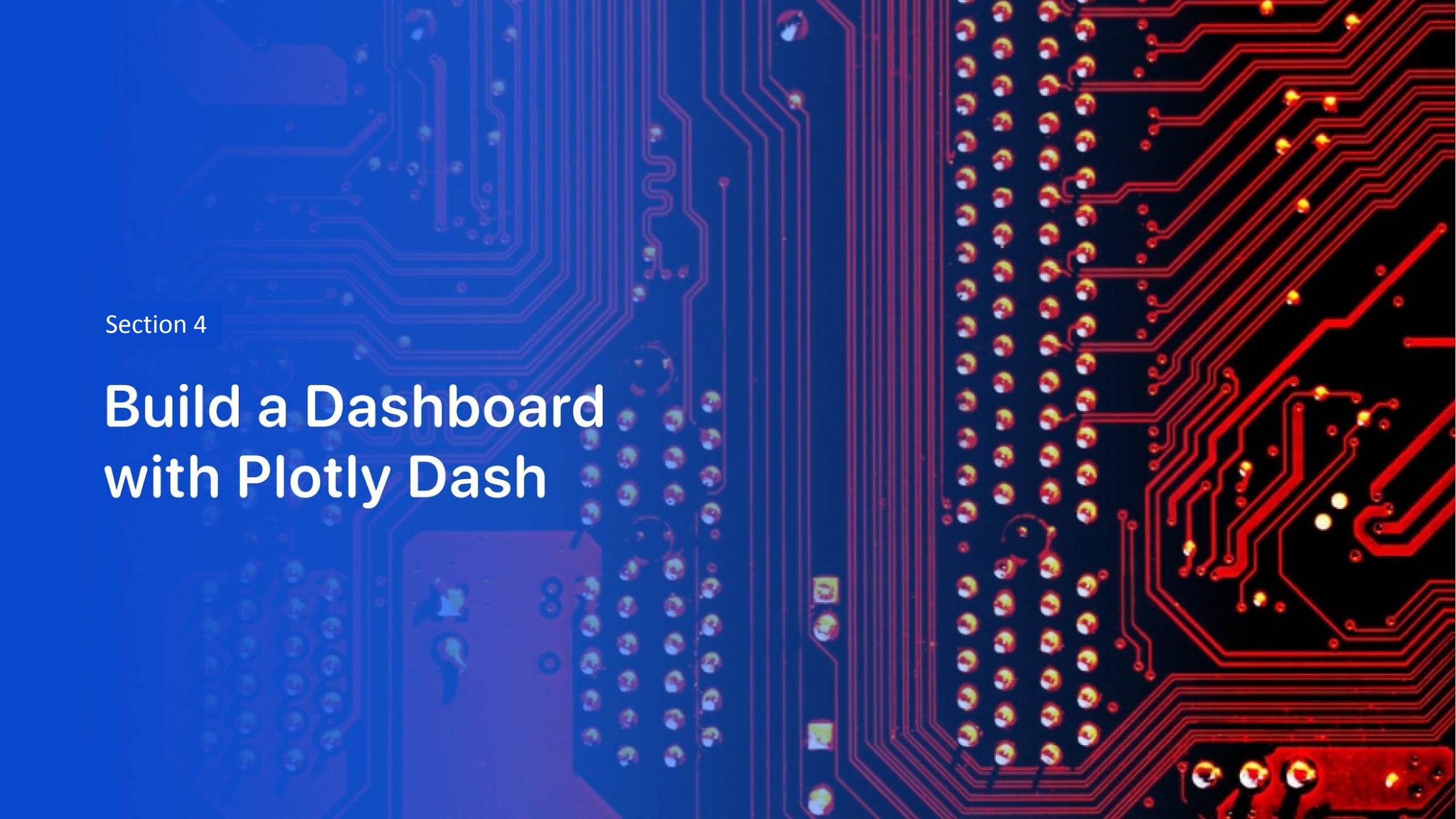




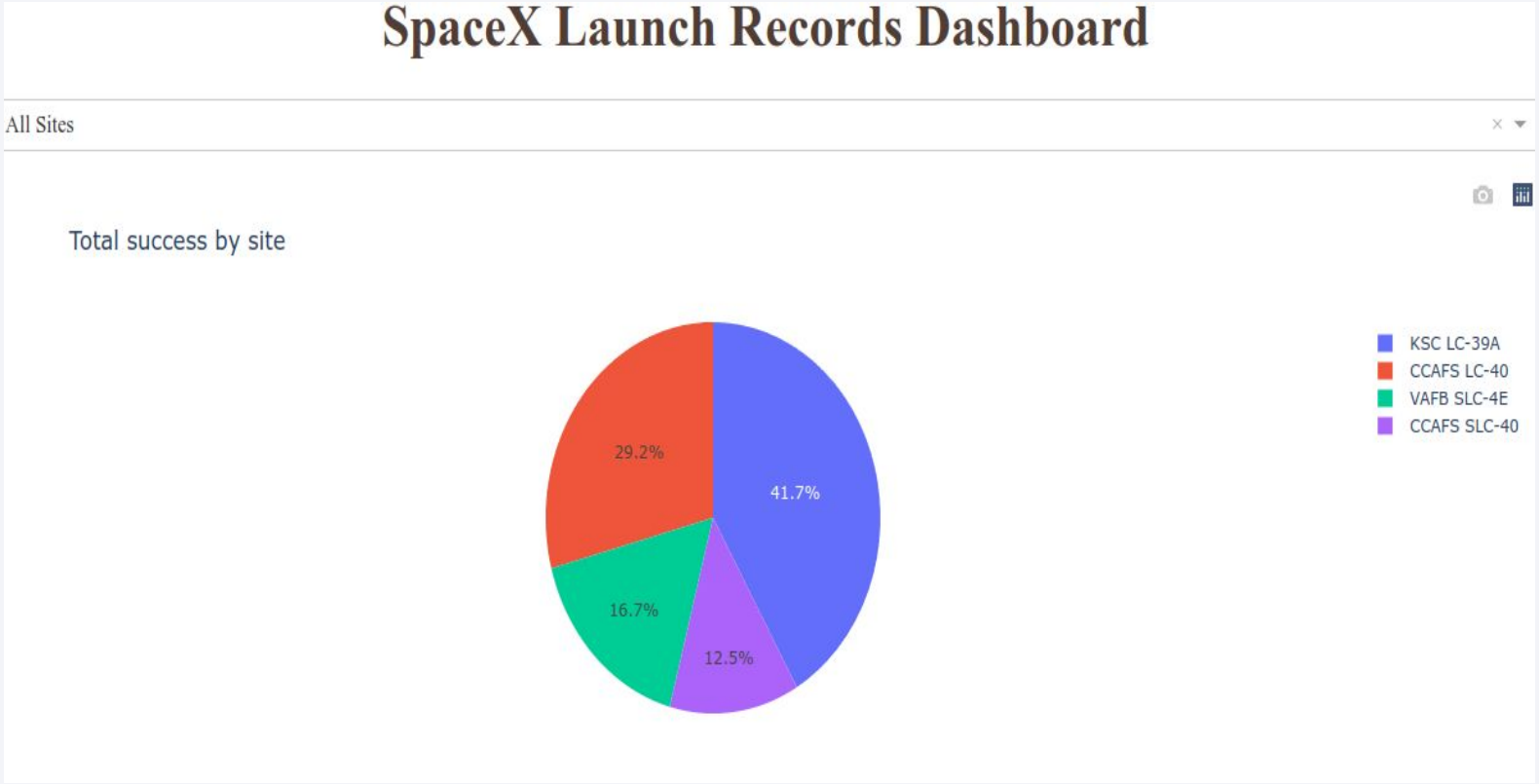
Distance from launch site to the ocean

From the Folium Map drawn is possible to conclude that:

- With the appropriate coordinate information it is easily demonstrated that is possible to calculate the distance of launch sites to different points of interest


Section 4
Build a Dashboard with Plotly Dash

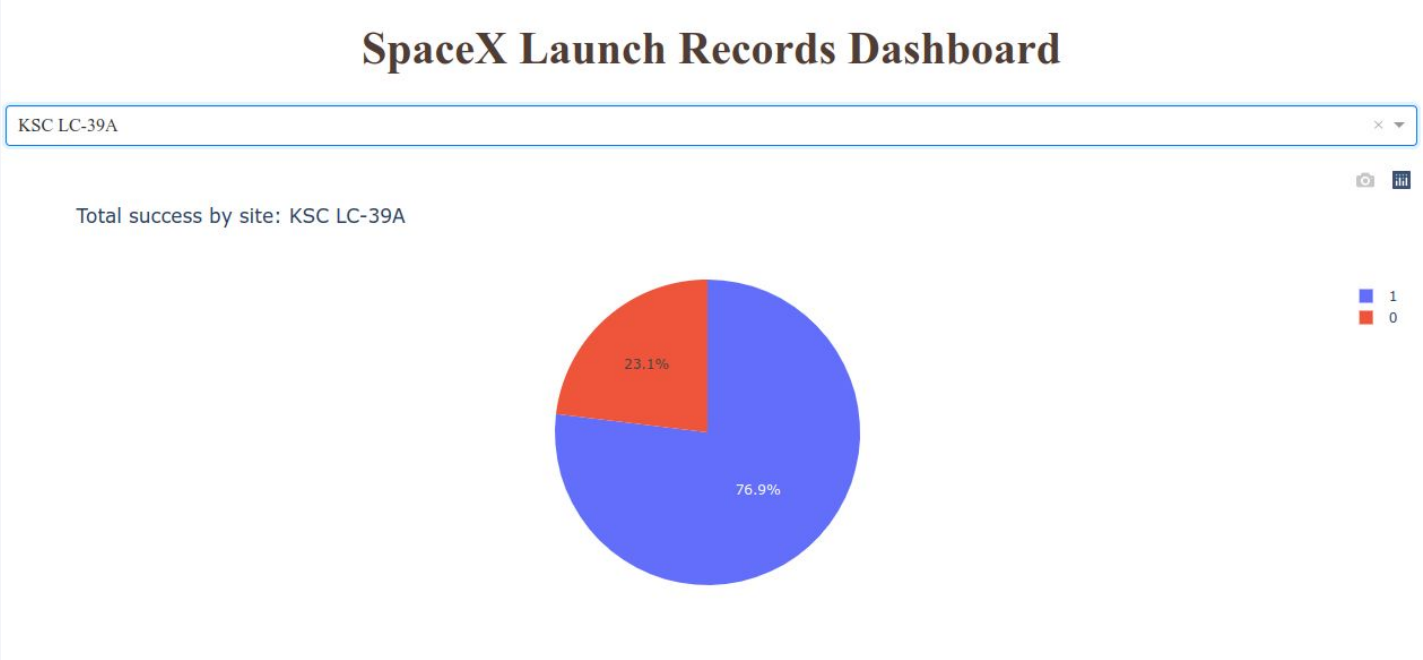
Launch Success Performance per site



From the Pie Chart drawn is possible to conclude that:

- The site with best performance is KSC LC-39A
- The site with worst performance is CCAFS SLC-40

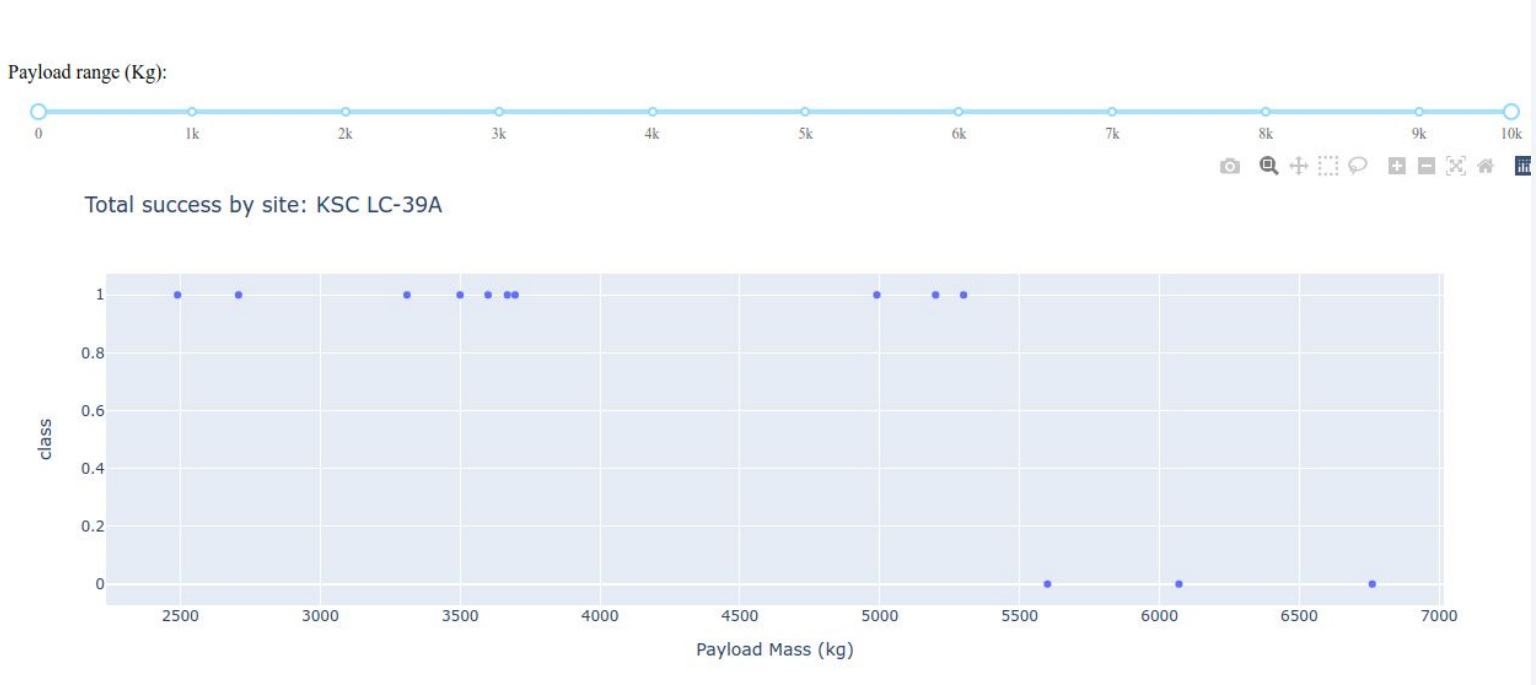
Launch performance KSC LC-39A



From the Pie Chart drawn is possible to conclude that:

- Total of success ratio = 73.1%

Site Launch Performance per Payload Mass



From the Chart drawn is possible to conclude that:

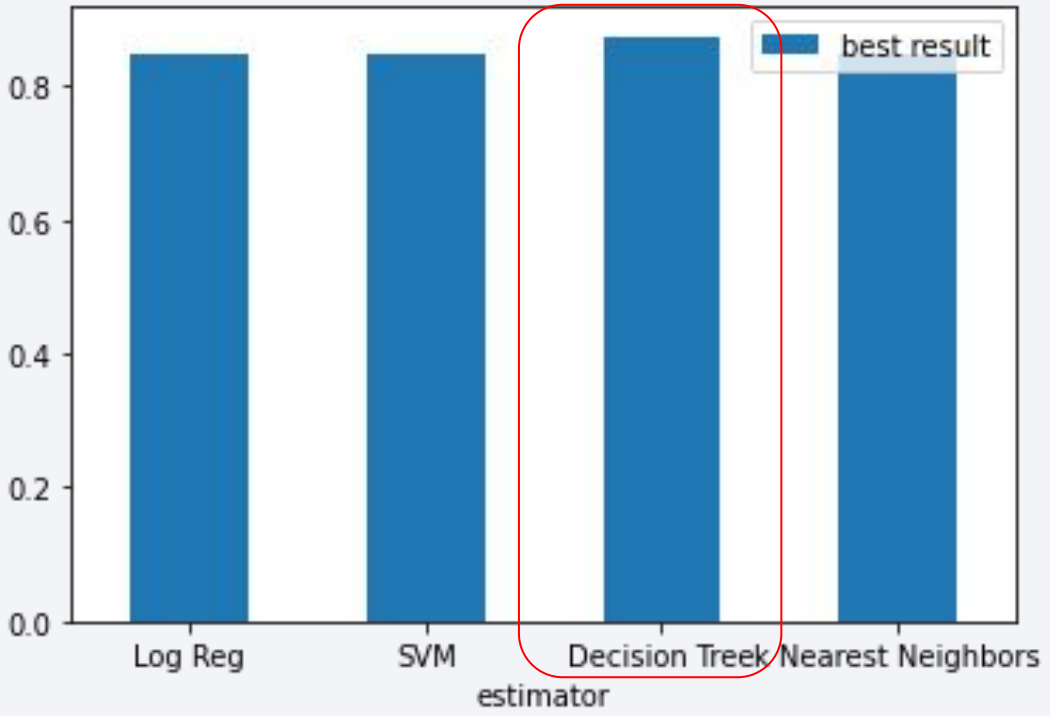
- The interactive chart provides filters:
  - Site
  - Payload Mass
- The filters allow the user to explore data in a more efficient manner

Section 5

Predictive Analysis (Classification)

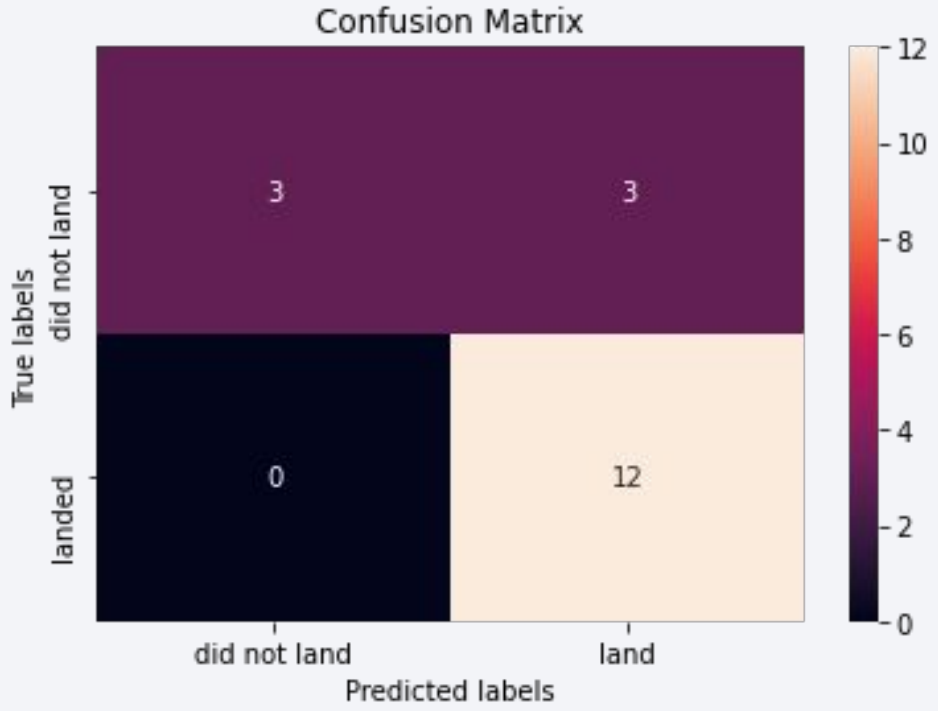
Classification Accuracy

	estimator	best result
0	Log Reg	0.846429
1	SVM	0.848214
2	Decision Tree	0.875000
3	k Nearest Neighbors	0.848214



Best result was found with the estimator Decision Tree

Confusion Matrix



From the Confusion Matrix is possible to conclude that:

- The decision tree classifier has an accuracy of 100% for landed rockets
- The decision tree classifier has an accuracy of 50% for not landed rockets

Conclusions

- For the 4 estimators tested and from the perspective of the best results found there were no significant differences between them all.
- The best estimator found was Decision Tree
- The precision for landed rockets is 100% for the test data
- The precision for not landed rockets is 50% for the test data

Appendix

- GitHub Location: https://github.com/smartlearningci/capstone_ibm_data_science

Thank you!