IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- DATA COLLECTION:

a. Public SpaceX API and SpaceX Wikipedia. Labels column 'class' is used to classify successful landings of the first stages. Data exploring thru SQL

b. Visualization, Folium Maps and dashboards.

c. Gathering of relevant columns to be considered and used as features

d. Changing all categorical variables to binary by encodingStandarisation of data and usage og GridSearchCV to optimize Machine Learn Parameters, and show a visual scoring of models.

- RESULTS SUMMARY:

Four ML models were produced; Linear Regression, Support Vector Machine, Decision Tree Classsifier and KNN. All them prodiced similar results around 83,33%. Clear overprdediction of successful landings, this indicates we need more data to improve accuracy.

Introduction

Background

- Commercial companies have arrived to Space Business
- Space X is the dominant competitor
- Its competitive advantage is the ability to recover the most expensive part ot the rocket, the first stage.
- Space Y wants to get into this market.

Problems you want to find answers

Space Y wants from us to train a machine learning model in order to predict successful recoveries of Stage 1

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classify true landings as successful or unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Usage of GridSearchCV

Data Collection

Data collection process involved a combination of API requests from Space X public API and webscraping data from a table in Space X Wikipedia entry.

The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from webscraping.

Space X API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,

Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Wikipedia Webscrape Data Columns:

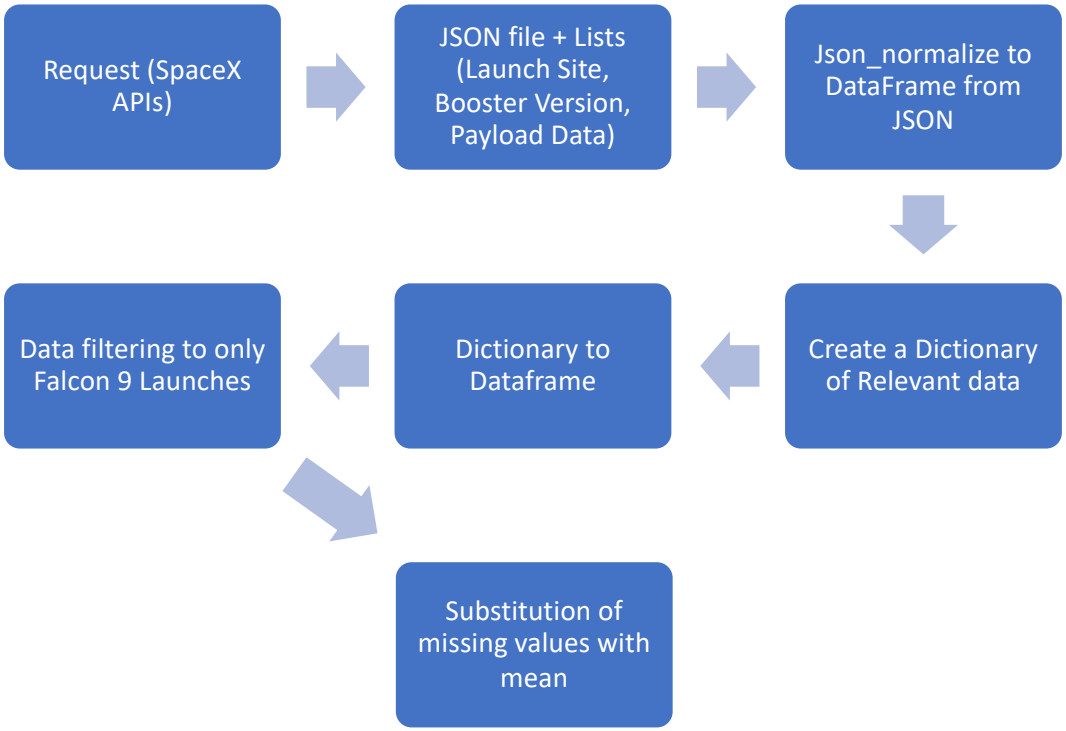
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

.

Data Collection – SpaceX API

Data Collection

SPACEX API

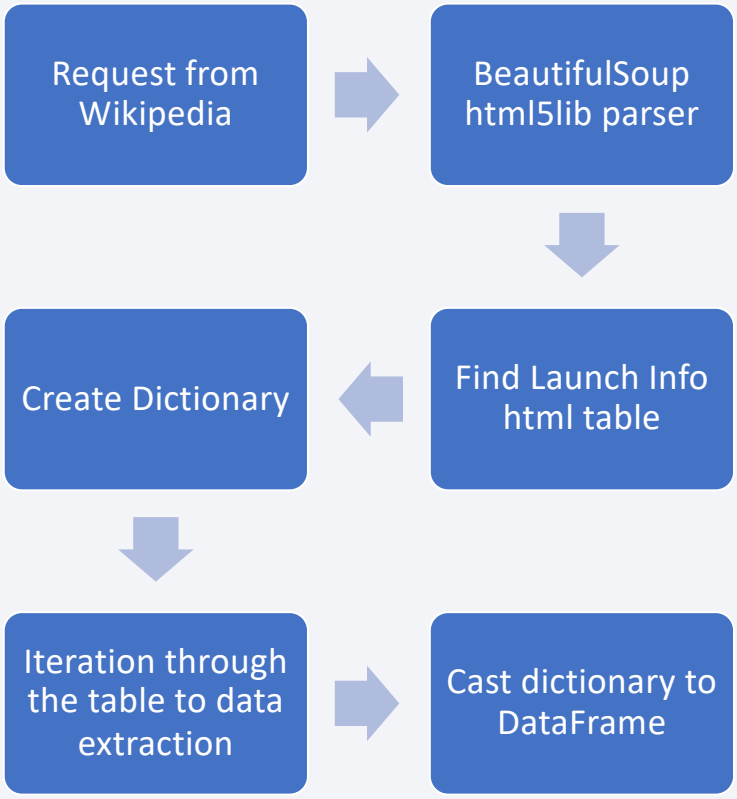


Data Collection - Scraping

Data Collection

WEB SCRAPING

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose



Data Wrangling

- Training label creation with landing outcomes; Success=1, Failure=0
- Outcome column has two components: <Mission Outcome> and <Landing Location>
- The training label column “class” with value 1 if Mission Outcome is True, and 0 otherwise-
- VALUE MAPPING:

  True ASDS, True RTLS and True Ocean set to 1

  None None, False or None ASDS, False or None RTLS or None or False Ocean set to 0

GitHub URL
https://github.com/JCCHECA1966/Professional_Certificate_Data_Science/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- EDA performed on variables Flight number, Payload Mass, Launch Site, Orbit, Class and Year
- PLOTS USED: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend, Scatter plots, line charts, and bar plots were used to compare relationships between variables to deduct the existing relations that could be used in training the machine learning mode

GitHub URL:

https://github.com/JCCHECA1966/Professional_Certificate_Data_Science/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

- Data loaded to IBM DB2 Database
- Usage of SQL Python integrated queries
- The queries had the purpose of a better understanding of the dataset
- These queries were about launch site names, mission outcomes, various payload sizing from customers and landing outcomes.

GitHub URL:

https://github.com/JCCHECA1966/Professional_Certificate_Data_Science/blob/main/EDA%20con%20SQL%20LAB.ipynb

Build an Interactive Map with Folium

- Folium maps marking launch sites, successful and unsuccessfullandings, and an example of key locations: Railroads, Highways, Coast or City.
- The aim is to show and understand the reason for the launch site election and also the relation between launch site election and the misión outcome.

GitHub url:
https://github.com/JCCHECA1966/Professional_Certificate_Data_Science/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

Dashboard includes a pie chart and a scatter plot.

Pie chart gives the option to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.

Scatter plot gathers two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.

The pie chart is used to visualize launch site success rate.

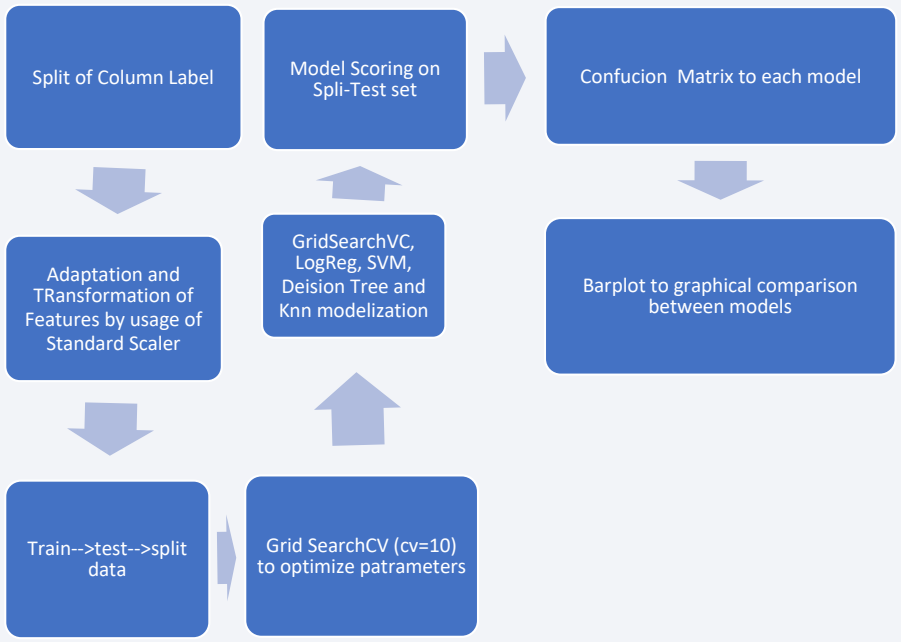
The scatter plot can help us see how success varies across launch sites, payload mass, and

booster version category.

GitHub URL:

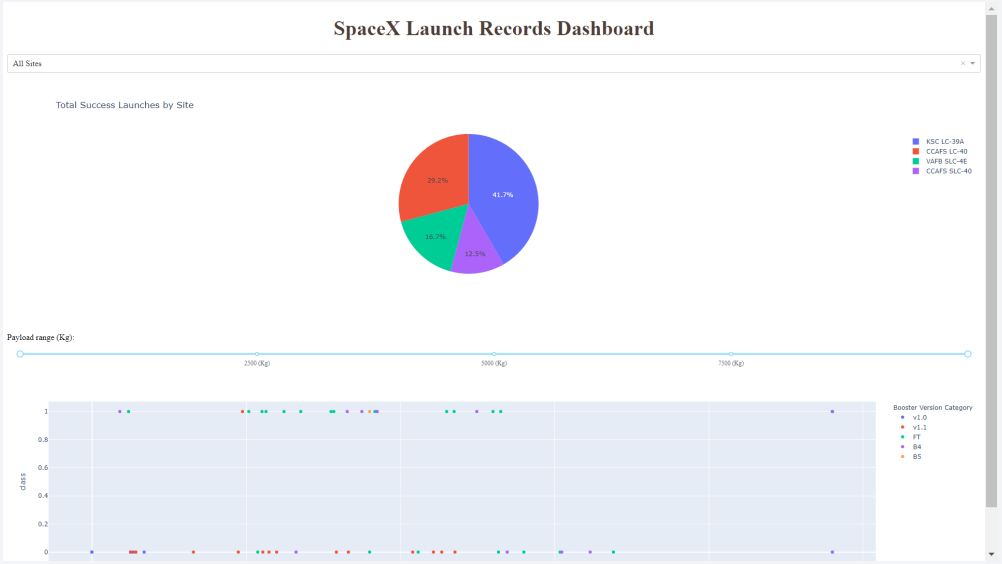
https://github.com/JCCHECA1966/Professional_Certificate_Data_Science/blob/main/Interactive%20Dashboard%20with%20Ploty%20Dash.ipynb

Predictive Analysis (Classification)



https://github.com/JCCHECA1966/Professional_Certificate_Data_Science/blob/main/Machine%20Learning%20Prediction%20lab.ipynb

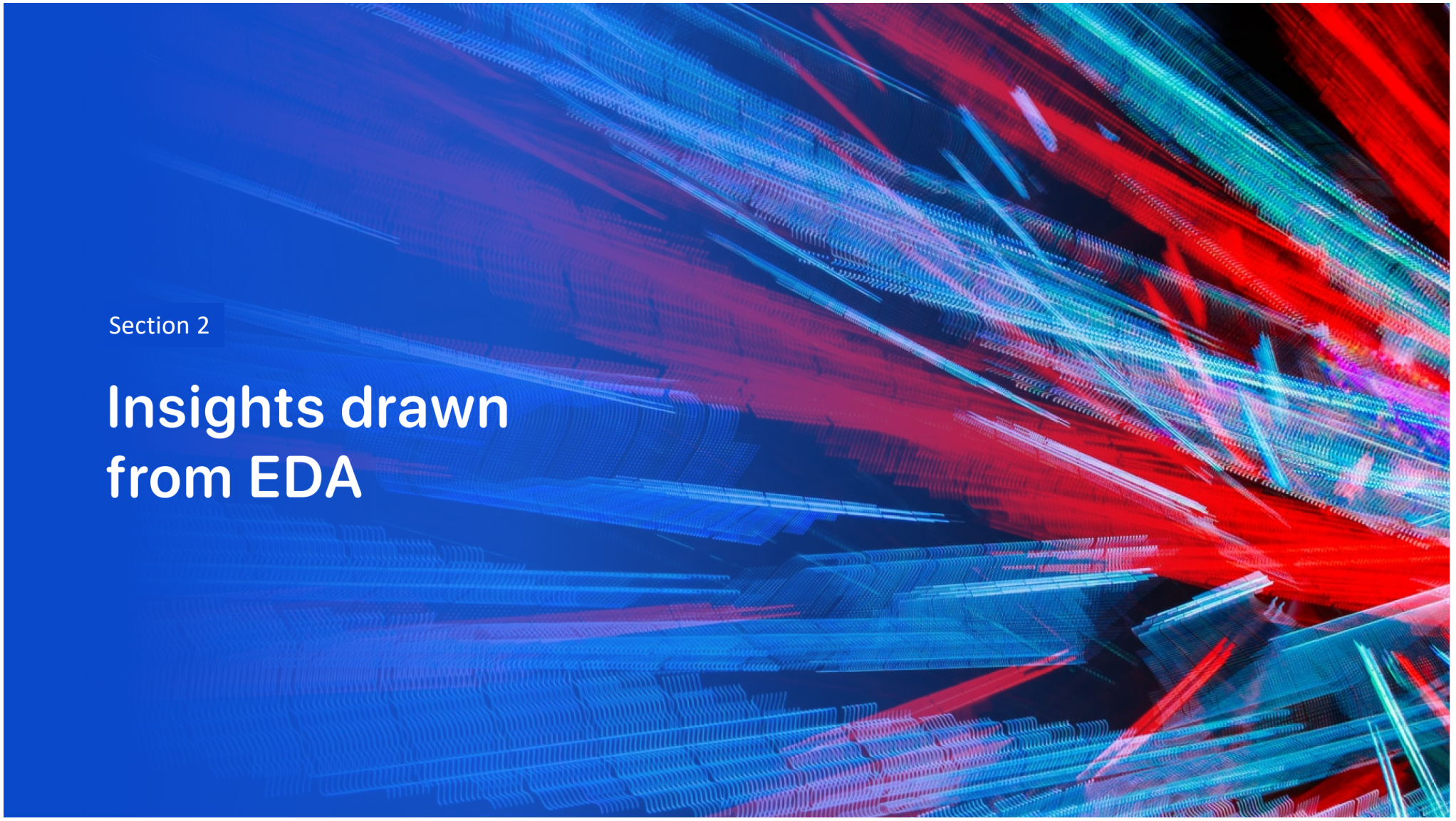
Results



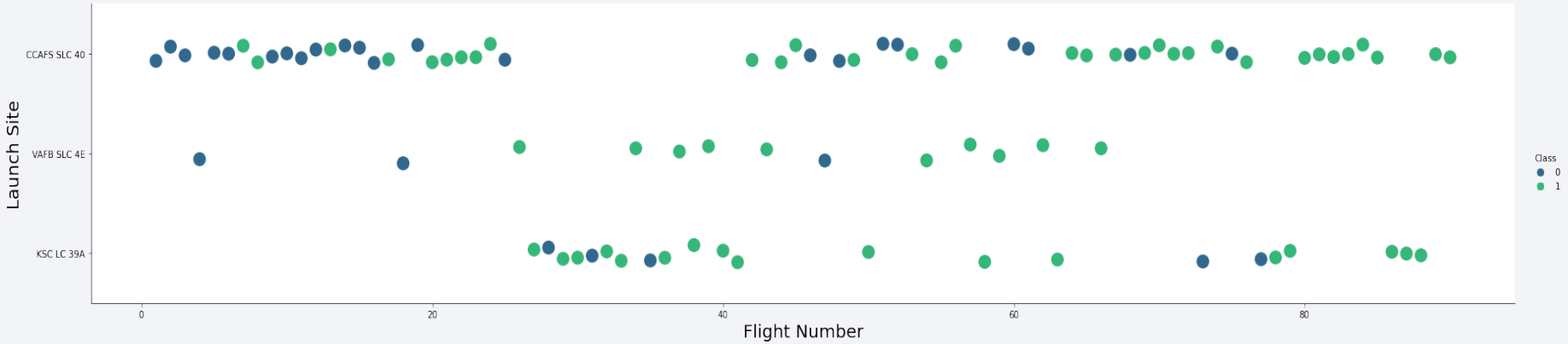
This is the preview of Interactive Dashboard. Next slis will show the results with EDA with Visualization, with SQL, Maps with Folium and the results of the model (Accuracy around 83%

Section 2

Insights drawn from EDA



Flight Number vs. Launch Site



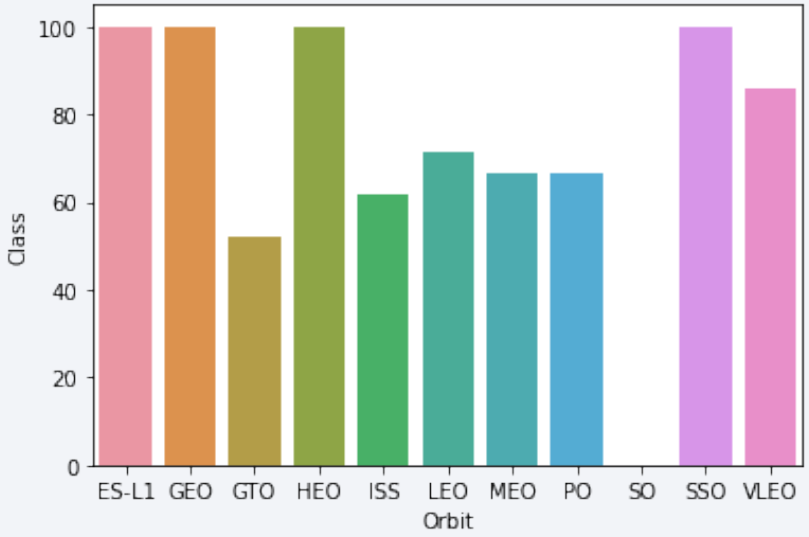
- Green indicates Sucessful Launches, Blue shows are uncsuccessful ones.
- CCAFS appears to be the launch site with the highest success rate; the trend also suggests that success rate increases in time (flight number can be associated to time)

Payload vs. Launch Site

- Green =Success; Blue =Failure
- All sites launch similar in the same interval of payload mass
- Payload mass interval is 0-6000 Kgs

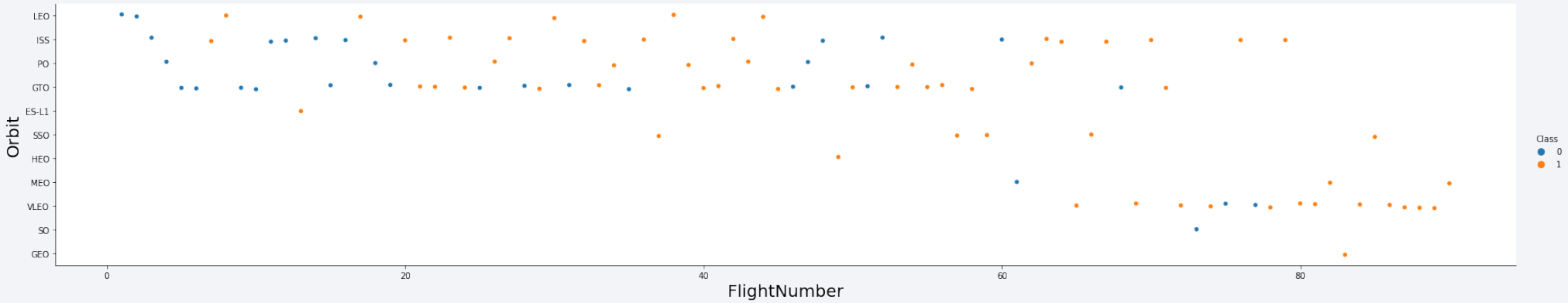
Success Rate vs. Orbit Type

- ES-L1, GEO1, HEQ have a 100% success rate but sample is only one flight.
- SSO with 5f lights has 100% success rate
- GTO has the biggest number of fights and a success rate 50%
- SO has 0% and only one flight



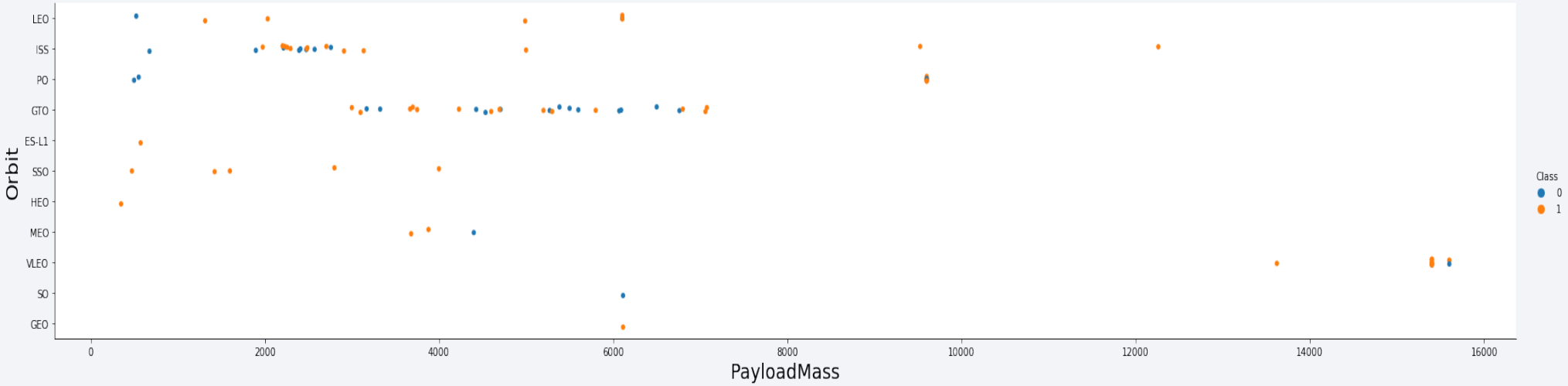
Success Rate Scale with 0 as 0%
0.6 as 60% 1 as 100%

Flight Number vs. Orbit Type



We seet hat in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

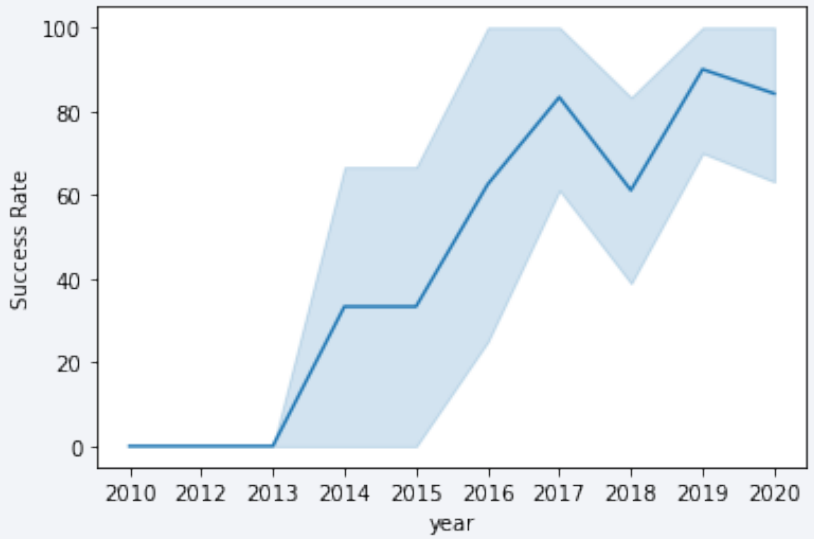
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



Success rate is increasing since 2013, till 2020 around 80%

All Launch Site Names

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

SQL QUERY:
%sql select DISTINCT LAUNCH_SITE from SPACEX

CCAFS SLC-40 and CCAFSSLC-40 both represent the same

launch site with data entry errors.; CCAFS LC-40 was the previous name. Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Begin with 'KSC'

```
%sql select * from SPACEX where launch_site like 'KSC%' limit 5
```

```
 * ibm_db_sa://zyv08444:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/BLUDB
Done.
```

]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16-03-2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
[illegible]	[illegible]	FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

ick to scroll output; double click to hide

Total Payload Mass

```
In [8]: %sql select sum(payload_mass__kg_) as sum from SPACEX where customer like 'NASA (CRS)'
 * ibm_db_sa://zyv08444:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/BLUDB
Done.
Out[8]:
SUM
45596
```

The SQL query sums the total payload mass in kgs where the customer is NASA

CSR indicates that the destination of payload was the ISS

Average Payload Mass by F9 v1.1

```
[25]: %sql select avg(payload_mass__kg_) as Average from SPACEX where booster_version like 'F9 v1.1%'
 * ibm_db_sa://zyv08444:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/BLUDB
Done.
Out[25]:
average
2534
```

- The SQL query calculates the average payload mass launchec with bosster F9 v1.1
- This average is in the low range of our payload mass interval.

First Successful Ground Landing Date

```
%sql select min(date) as Date from SPACEX where mission_outcome like 'Success'AND (landing__outcome like 'Success (ground pad)')
```

```
 * ibm_db_sa://zyv08444:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/BLUDB
Done.
```

DATE
01-05-2017

- This query returns the first successful ground pad landing date.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively.

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

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

- Query returns a count of mission outcomes
- Apparently 99% of times result is achieved.

Boosters Carried Maximum Payload

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- The SQL query returns the booster versions that carried the highest payload mass of 15600 kg.
- the F9 B5 B10xx.x variety.
- This shows that payload mass correlates with the booster version that is used.

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

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

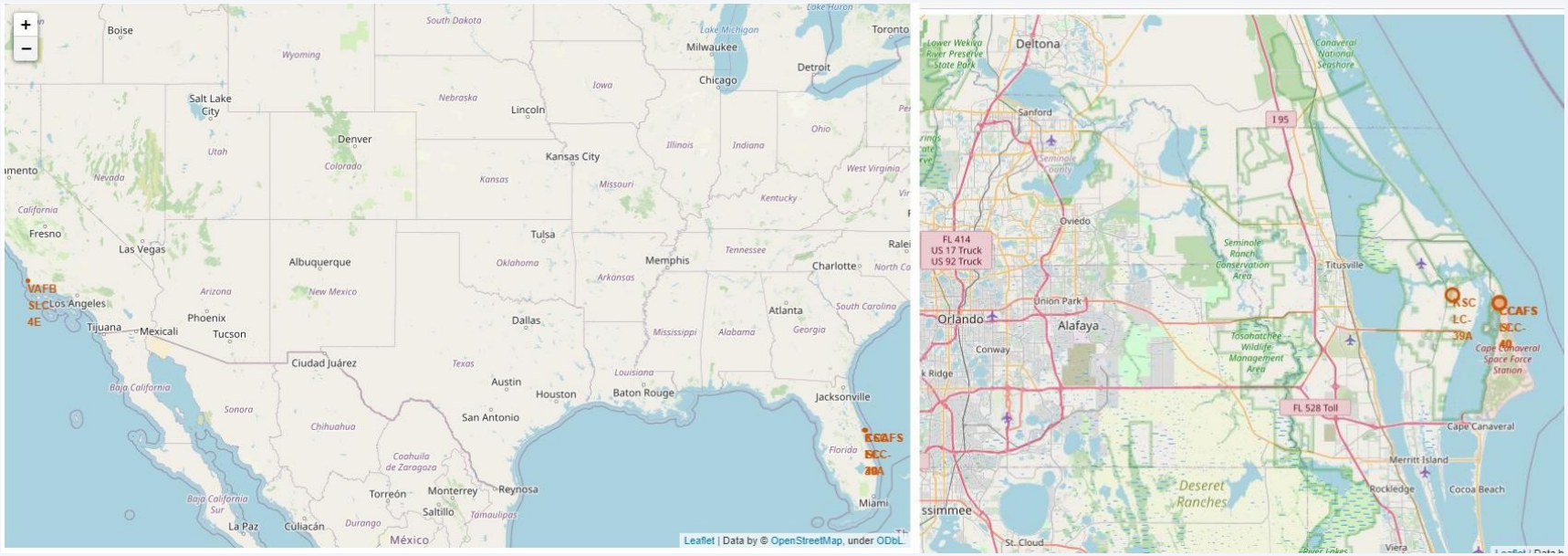
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

- This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.
- There are two types of successful landing outcomes: drone ship and ground pad landings.
- There were 8 successful landings in total during this time period

Section 3

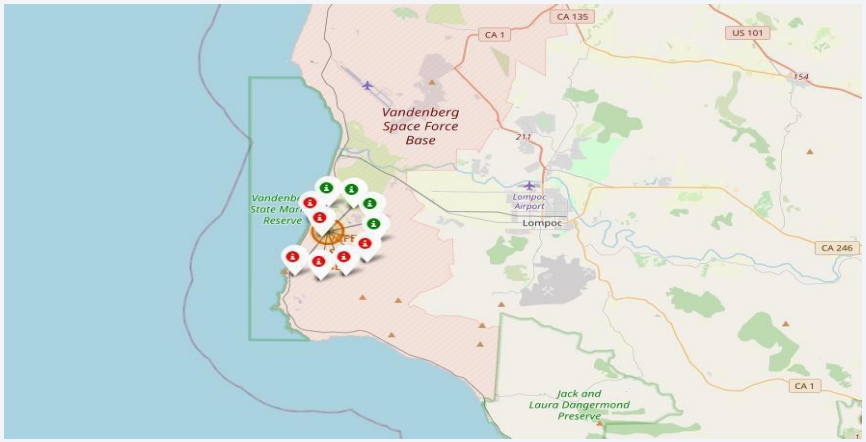
Launch Sites Proximities Analysis

Launch site Locations



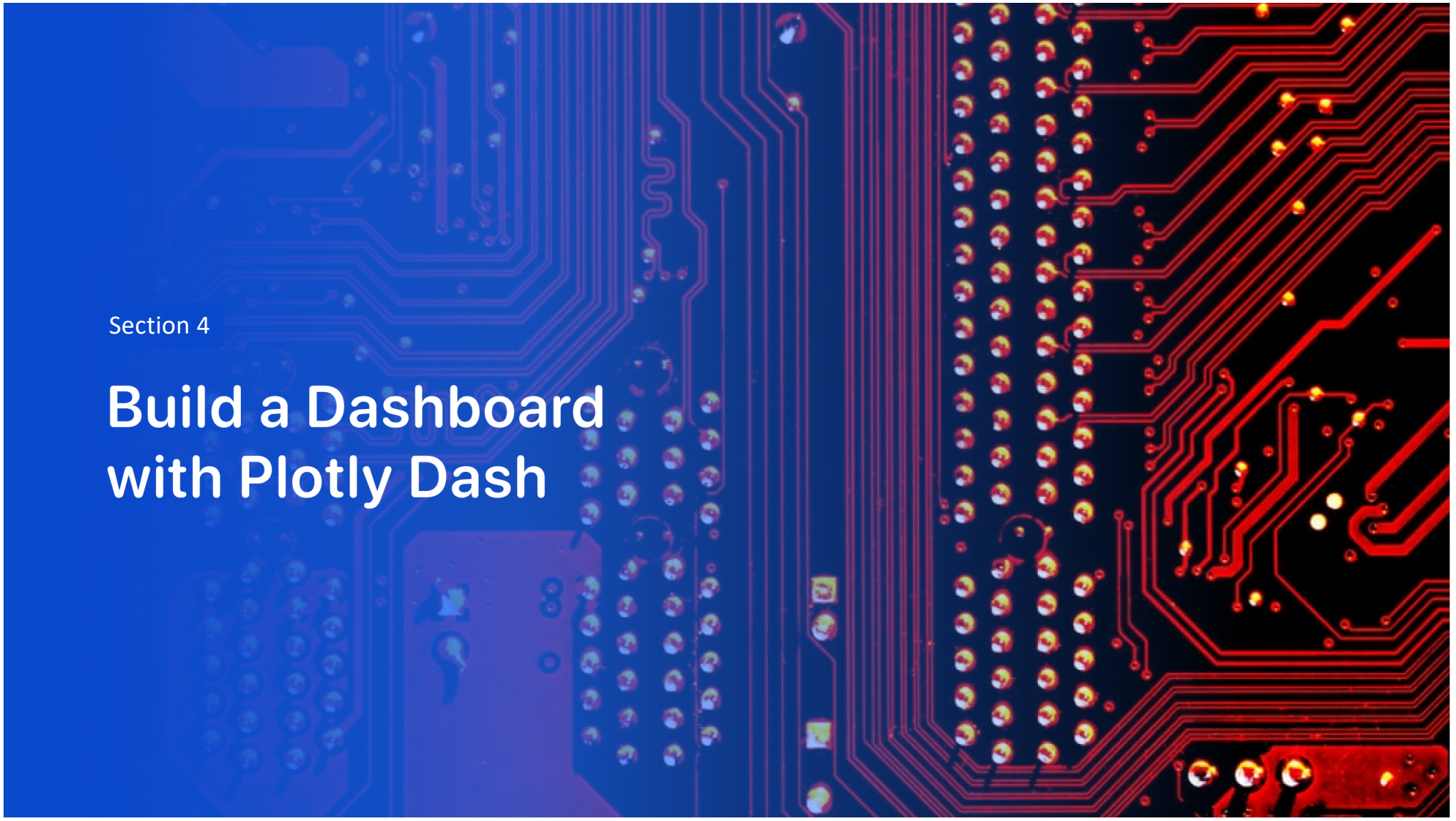
The left map shows all launch sites within an US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are close to the ocean.

Color Coded Launch Markers

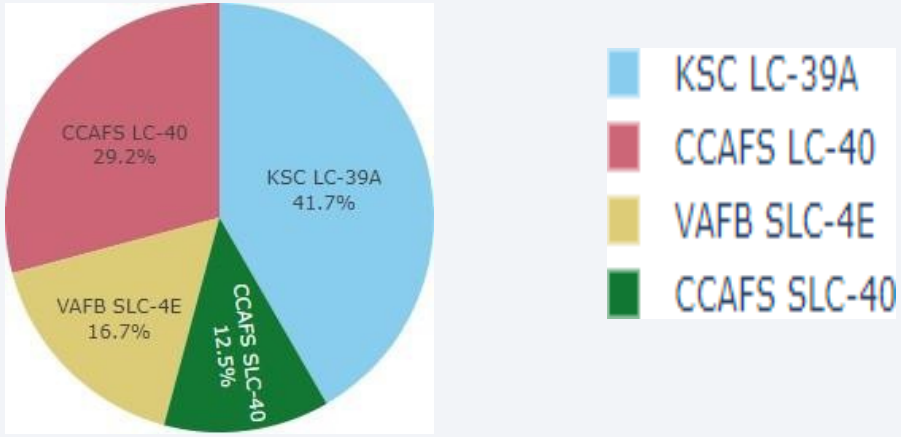


In this image VAFB SLC 4E shows

- 4 successful missions,
- 6 Unsuccessful missions


Section 4
Build a Dashboard with Plotly Dash

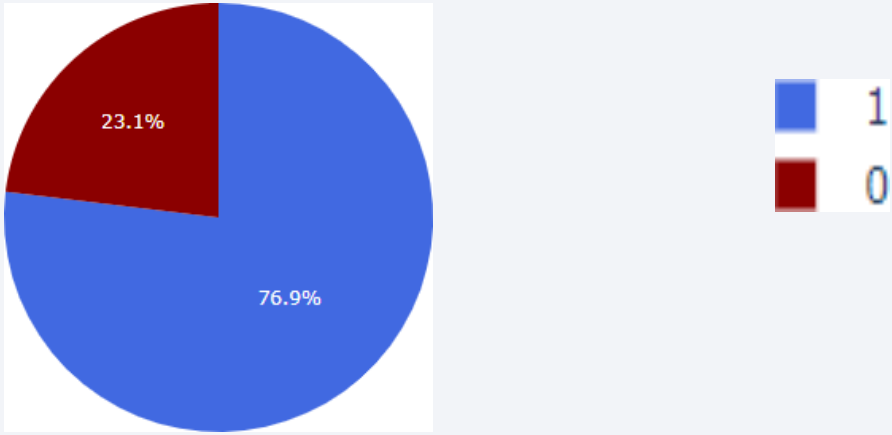
Successful Launches Across Launch sites



Distribution of successful launches across all launch site. Actually, CCAFS SLC-40 and CCAFS LC-40 are the same site. The lowest share is for VAFB, but the sample is smaller…

Highest Success Rate Launch Site

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



KSC LC 39A has the highest rate with 10 over 13 landings

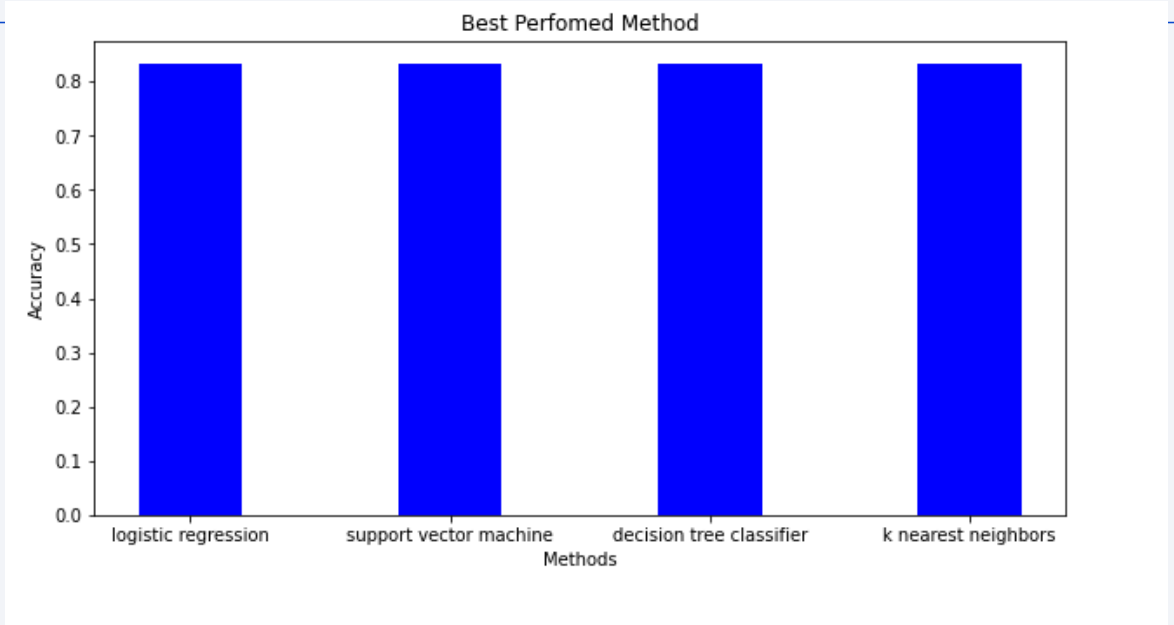
Payload Mass vs Success vs Booster Version

Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- All models offer the same accureacy level, around 83%
- As the sample is small, in statistical terms we can affirm tha tht lack of data can cause an over prediction of successes.

Confusion Matrix

- Correct Predictions are within the diagonal top left – bottom right
- All models offer exactly the same prediction, and from my point of view lack of data in the sample induces a, over prediction of successes.

Conclusions

- Task: Develop a ML model for Space Y, trying to compete with Space X
- The first goal is to predict when Stage 1 will successfully land to save the most expensive art of the spaceship
- Used Space X API and webscrapping techniques
- Dashboard for result interactive visualization
- ML accuracy is 83%
- The company can use the ML model to predict with an acceotable level of accuracy if a Satage 1 will successfully land and decide the go / no-go of launches.
- IT would be better to have a bigger sample to depurate ML accuracy

Appendix

- GITGHUB Repository:

https://github.com/JCCHECA1966/Professional_Certificate_Data_Science

Thank you!