IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andrey Koleda

10 Nov 2021

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with visualization
  - EDA 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 demo in screenshot
  - Predictive analysis results

Introduction

- Project background and context

We will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

- Problems that needed solving

  - What factors influence the successful landing?
  - What parameters should be used for optimal landing?
  - What is the probability?

Section 1

Methodology

Methodology

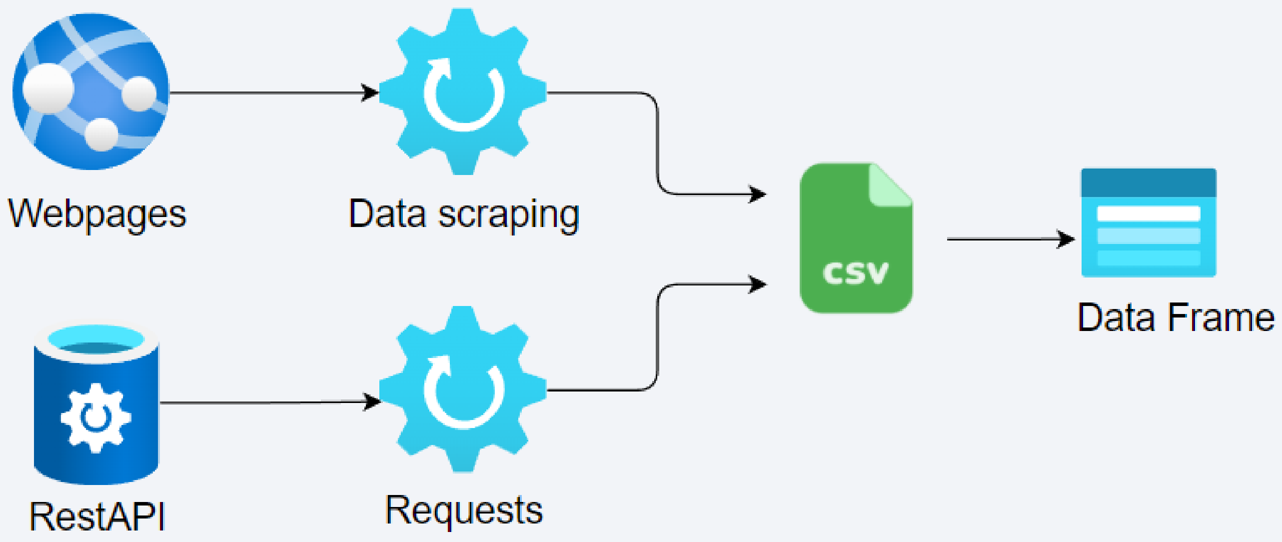
Executive Summary

- Data collection methodology:
  - RestAPI and Web Scrapping
- Perform data wrangling
  - Filling in the missing values and deleting unnecessary columns and
  - One Hot Encoding (OHE)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- How data sets were collected.

  We use the SpaceX API to collect data about rockets and their launches and landings. But there is not enough data. To enrich the dataset, we use data parsing from Wikipedia.

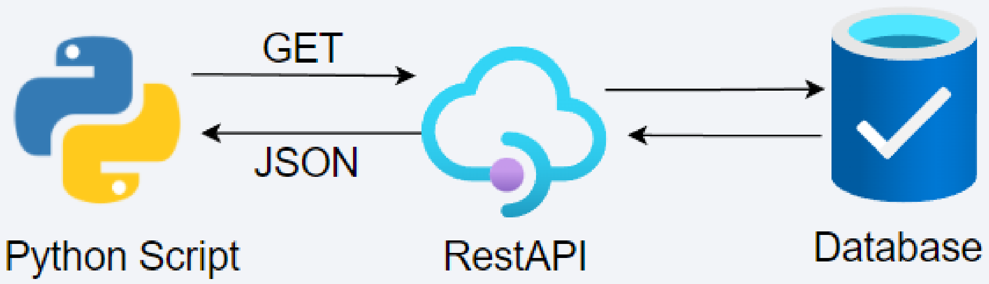


Data Collection – SpaceX API

Using Python, we access the SpaceX API and get data in JSON format and combine everything into one dataset

Base URL: https://api.spacexdata.com/v4/{endpoint}

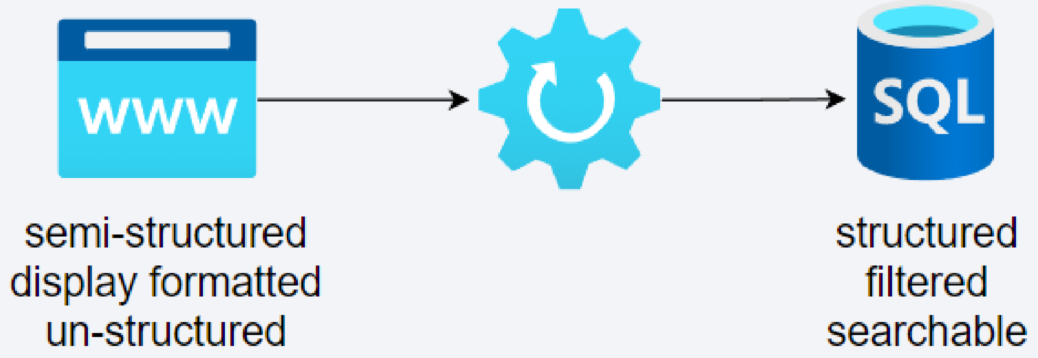
Endpoints: rockets, launchpad, payloads, cores, etc.



https://github.com/koav/IBM-DataScience-SpaceX-Project/blob/master/w01-01-data-collection-api.ipynb

Data Collection - Scraping

- Sending an HTTP GET request to the URL of the webpage that we want to scrape which will respond with HTML content (Wikipedia)
- Fetching and parsing the data using **Beautifulsoup**

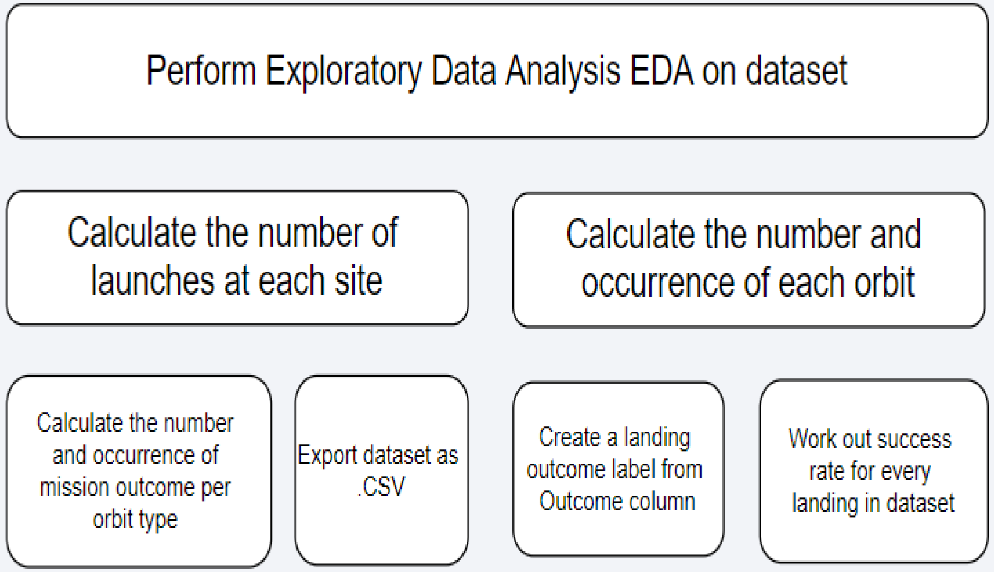


https://github.com/koav/IBM-DataScience-SpaceX-Project/blob/master/w01-01-data-collection-api.ipynb

Data Wrangling

In the dataset there is when the rocket landed not successfully. There are launches in our dataset on land (RTLS) or at sea (ASDS / Ocean).

We mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful



EDA with Data Visualization

Scatter Graphs being drawn:

- Flight Number VS. Payload Mass
- Flight Number VS. Launch Site
- Payload VS. Launch Site
- Orbit VS. Flight Number
- Payload VS. Orbit Type

Bar Graph being drawn

- Mean VS. Orbit

Line Graph being drawn

- Success Rate VS. Year

https://github.com/koav/IBM-DataScience-SpaceX-Project/blob/master/w02-02-eda-dataviz.ipynb

EDA with SQL

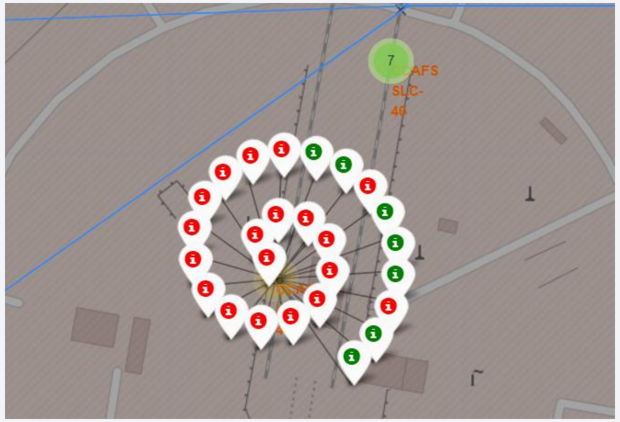
Performed SQL queries to gather information about the dataset

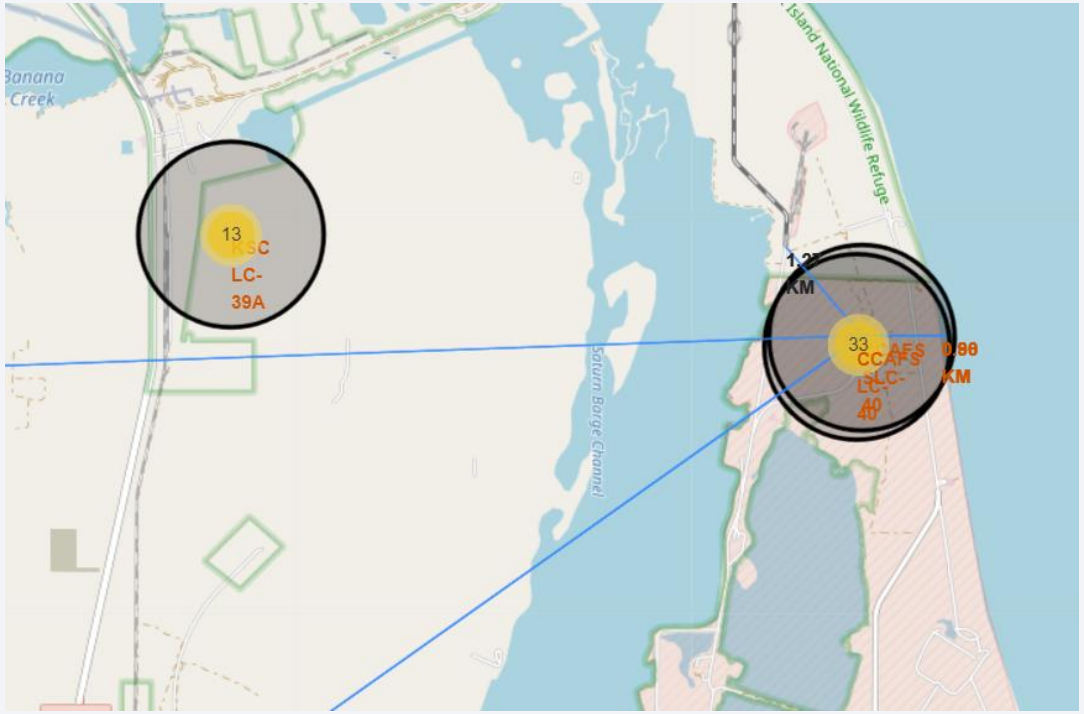
- Displaying the names of the unique launch, 'KSC’ launches, total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions

https://github.com/koav/IBM-DataScience-SpaceX-Project/blob/master/w02-01-eda-sql.ipynb

Build an Interactive Map with Folium

Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns



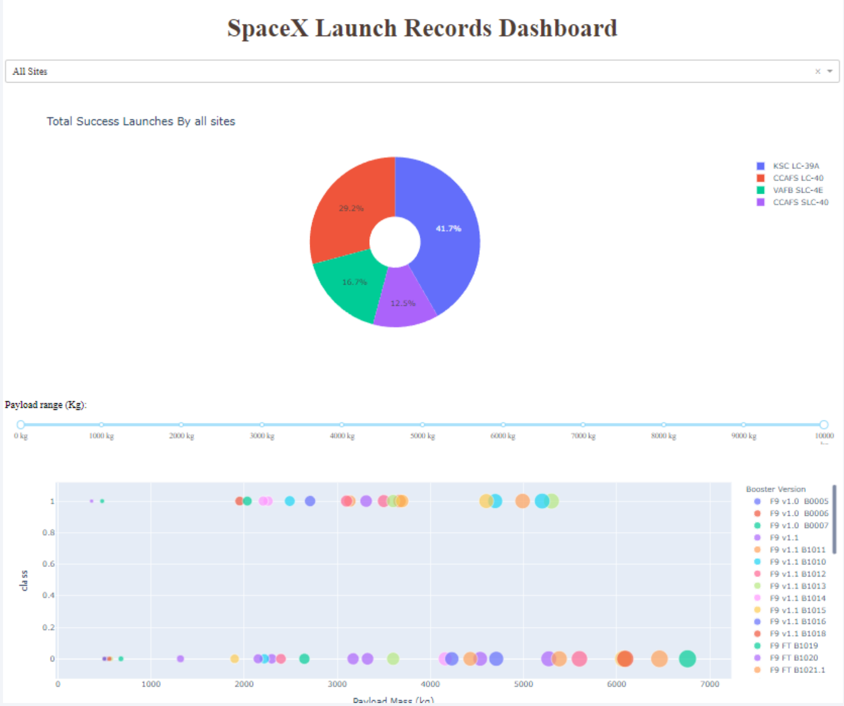


https://github.com/koav/IBM-DataScience-SpaceX-Project/blob/master/w03-01-launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Created an interactive dashboard with metrics using **Dash**

- Pie Chart showing the total launches by a certain site/all sites
- Scatter Graph showing the relationship with Outcome and Payload Mass (Kg) for the different Booster Versions

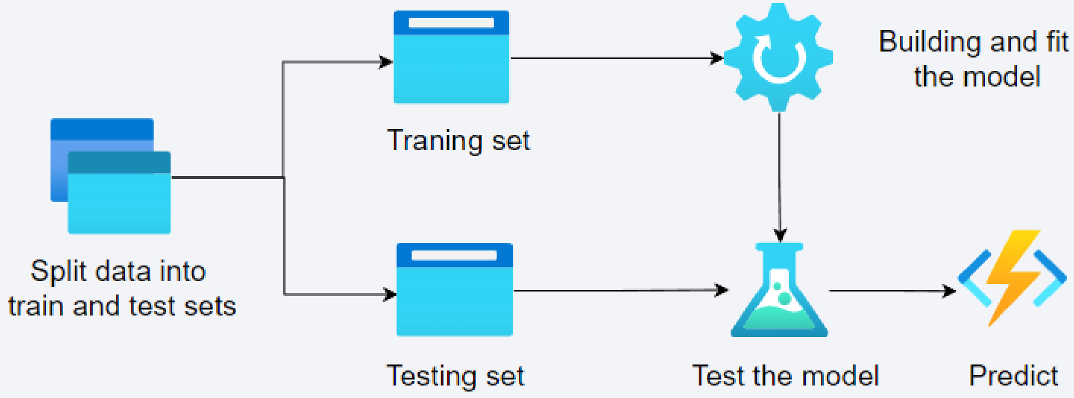


https://github.com/koav/IBM-DataScience-SpaceX-Project/blob/master/scripts/spacex_dash_app.py

Predictive Analysis (Classification)

How did we get a quality model

- Load our dataset
- Transform data
- Split our data into training and test data sets
- Use learning algorithms and parameters
- Fit and find best parameters with GridSearchCV
- Check the score for each mode
- Improve the model



https://github.com/koav/IBM-DataScience-SpaceX-Project/blob/master/w04-01-predictions.ipynb

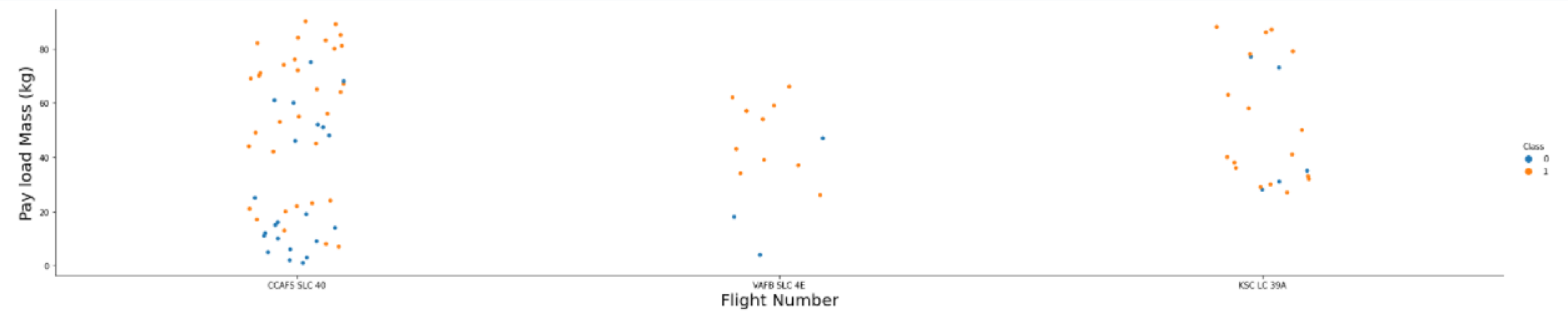
Results

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

Section 2

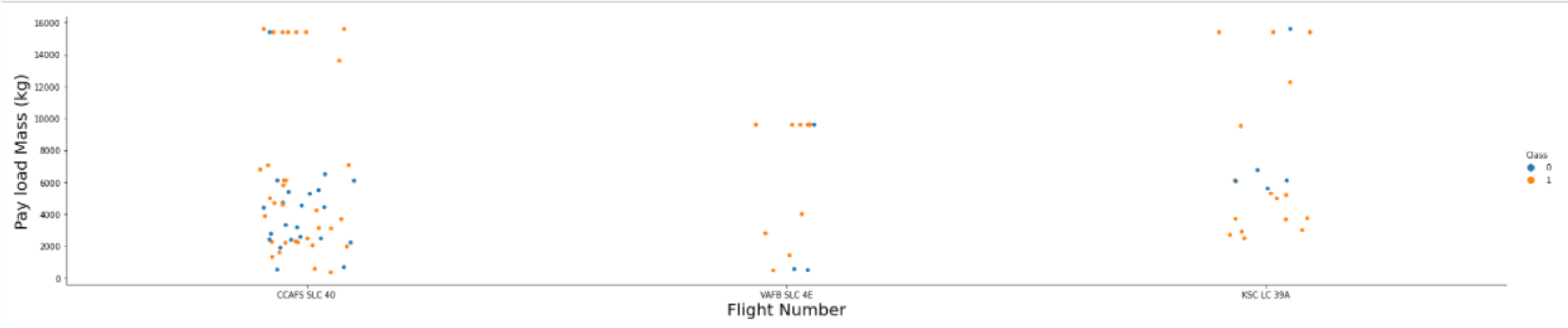
Insights drawn from EDA

Flight Number vs. Launch Site



The more amount of flights at a launch site the greater the success rate at a launch site

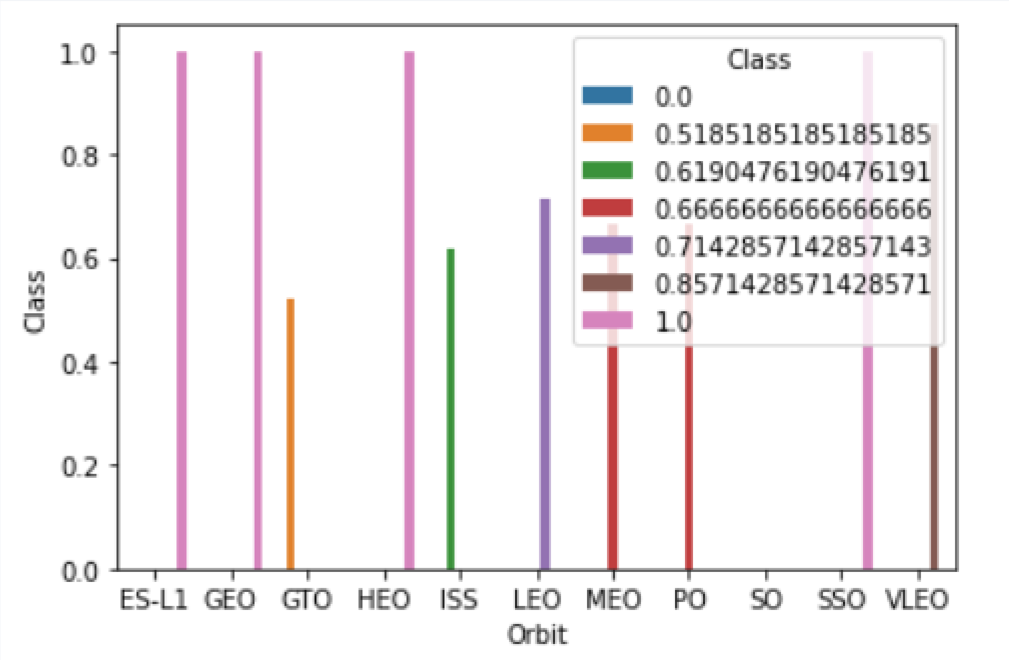
Payload vs. Launch Site



The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket

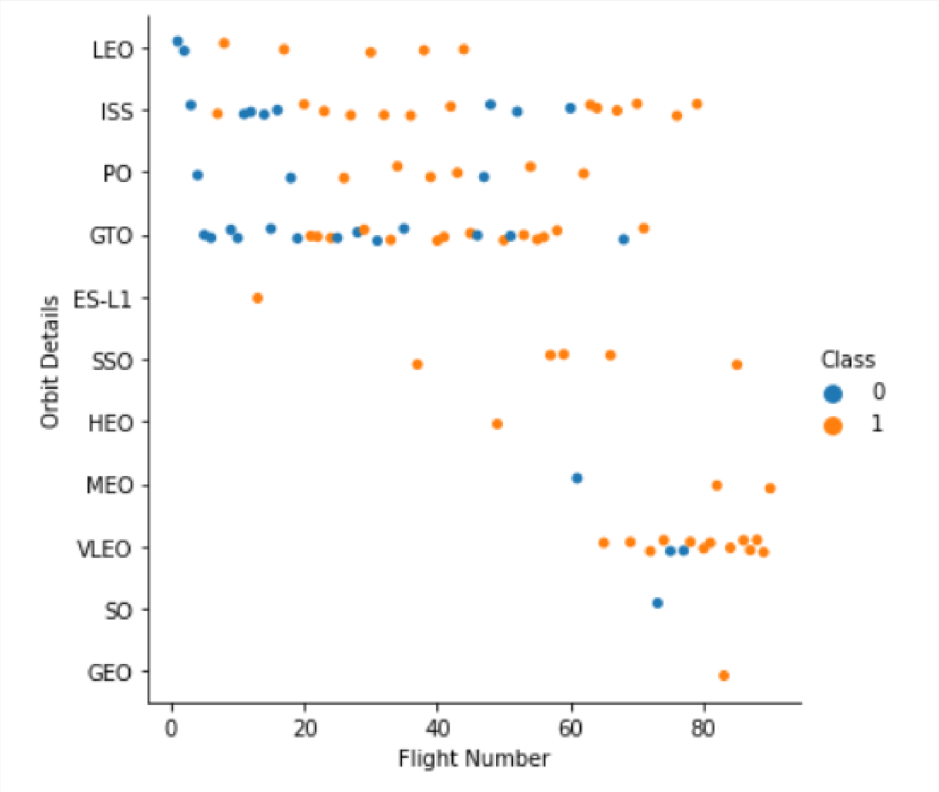
Success Rate vs. Orbit Type

Orbits ES-L1, GEO, HEO, SSO has the best Success Rate



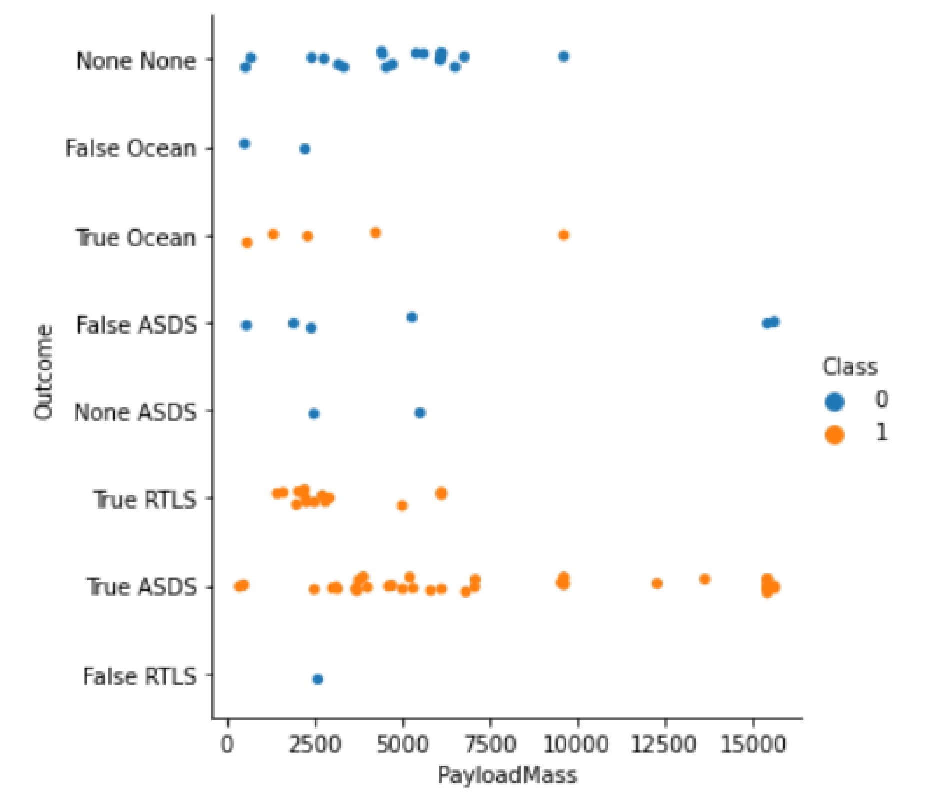
Flight Number vs. Orbit Type

You should see that 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.



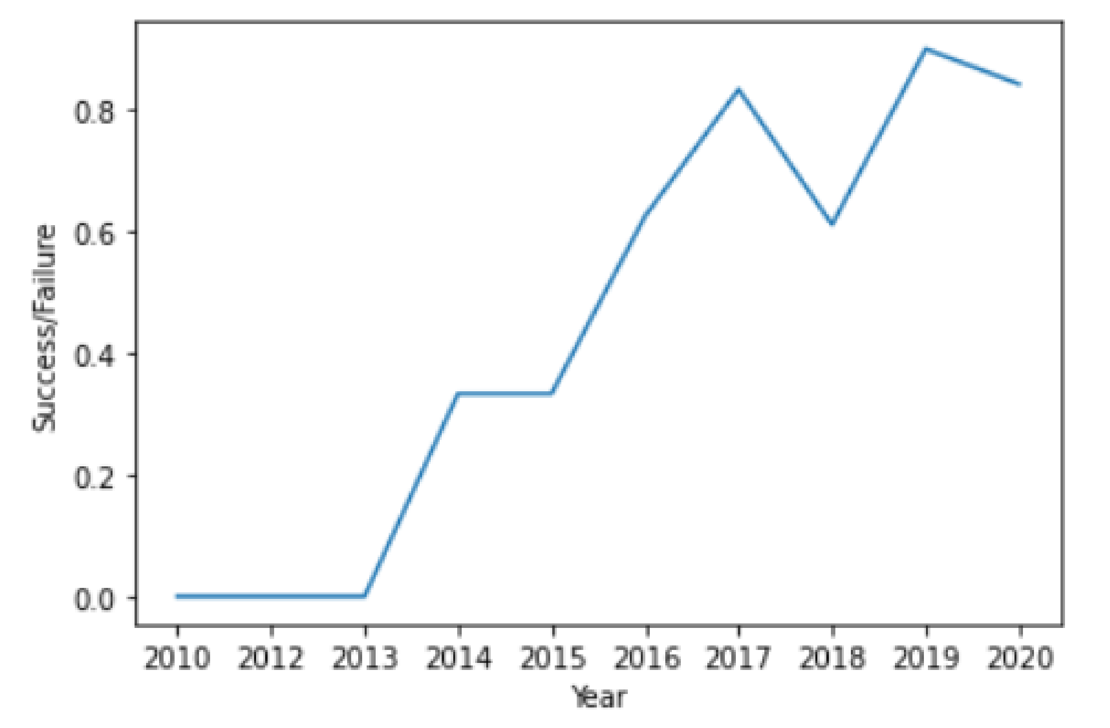
Payload vs. Orbit Type

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



Launch Success Yearly Trend

Here you can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

The names of the unique launch sites in the space mission

```
select DISTINCT Launch_Site from SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

```
select * from SPACEXTBL where Launch_Site like 'CCA%' LIMIT 5
```

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

Total Payload Mass

Total payload mass carried by boosters launched by NASA (CRS)

```
select SUM(PAYLOAD_MASS__KG_) Mass from SPACEXTBL where Customer = 'NASA (CRS)'
```

mass
45596

Average Payload Mass by F9 v1.1

Here you can see average payload mass carried by booster version F9 v1.1

```
select AVG(PAYLOAD_MASS__KG_) mass from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

mass
2928

First Successful Ground Landing Date

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

```
. select MIN(DATE) landing from SPACEXTBL where LANDING__OUTCOME = 'Success (drone ship)'
```

landing
2016-04-08

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
 select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME =
'Success (ground pad)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MA
SS__KG_ < 6000
```

booster_version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

```
 sql SELECT Count(*) Attempts FROM SPACEXTBL where MISSION_OUTCOME
LIKE '%Success%' UNION ALL SELECT Count(*) FROM SPACEXTBL where MI
SSION_OUTCOME LIKE 'Failure%'
```

attempts
1
100

Boosters Carried Maximum Payload

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

```
    select BOOSTER_VERSION as boosterversion from SPACEXTBL where
PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

2015 Launch Records

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

```
    SELECT MONTH(DATE),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE
FROM SPACEXTBL where EXTRACT(YEAR FROM DATE)='2015';
```

1	mission_outcome	booster_version	launch_site
1	Success	F9 v1.1 B1012	CCAFS LC-40
2	Success	F9 v1.1 B1013	CCAFS LC-40
3	Success	F9 v1.1 B1014	CCAFS LC-40
4	Success	F9 v1.1 B1015	CCAFS LC-40
4	Success	F9 v1.1 B1016	CCAFS LC-40
6	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

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

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

```
    SELECT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04'
AND '2017-03-20' ORDER BY DATE DESC;
```

No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Failure (drone ship)
Success (ground pad)
Precluded (drone ship)

No attempt
Failure (drone ship)
No attempt
Controlled (ocean)
Failure (drone ship)
Uncontrolled (ocean)
No attempt
No attempt
Controlled (ocean)
Controlled (ocean)
No attempt
No attempt
Uncontrolled (ocean)

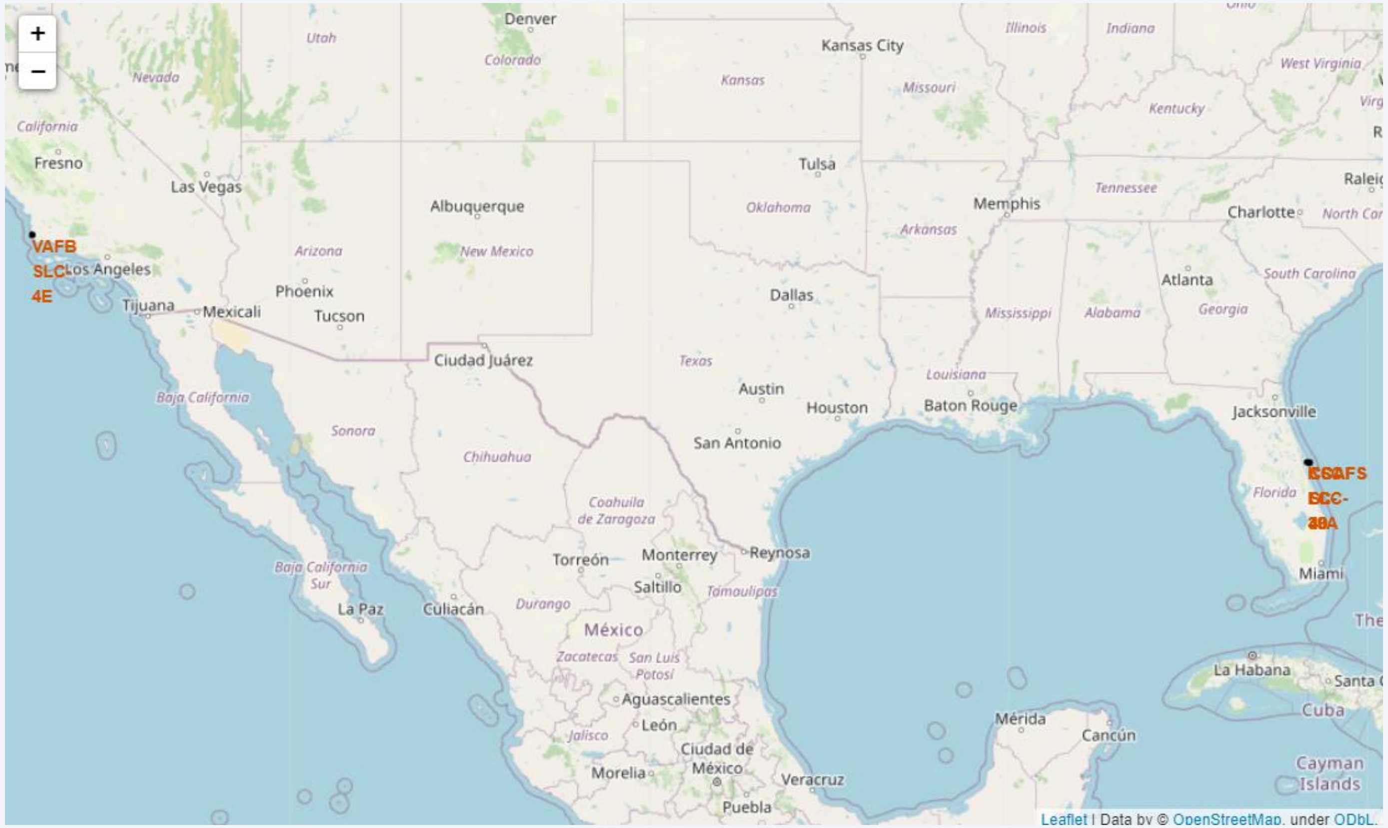
No attempt
No attempt
No attempt
Failure (parachute)
Failure (parachute)

Section 4

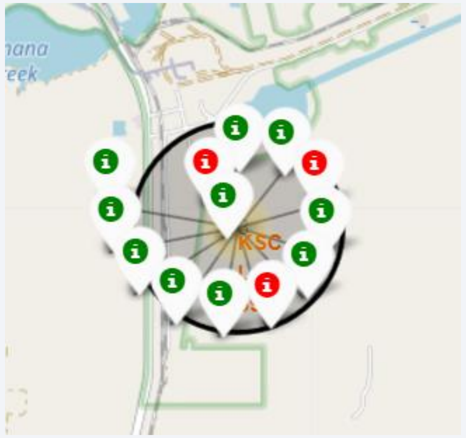
Launch Sites Proximities Analysis

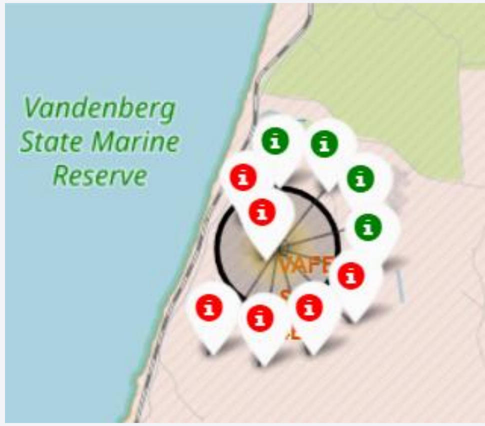
All launch sites

Here we can see that the SpaceX launch sites are in the USA coasts



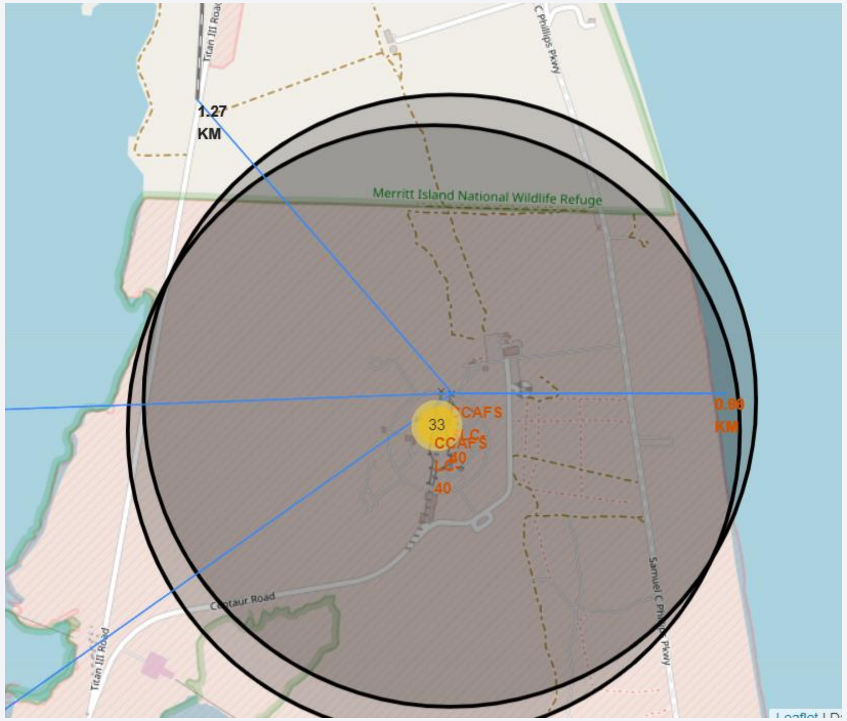
Colour Labelled Markers

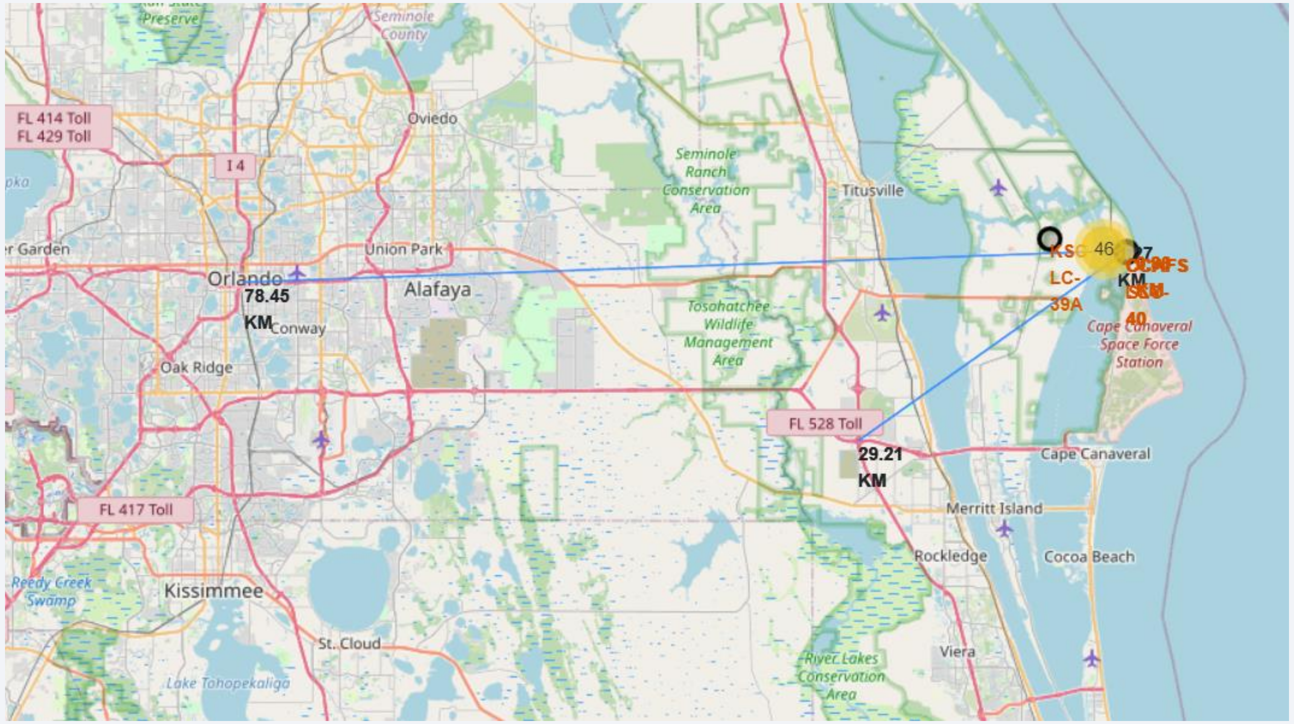




Green Marker shows successful Launches and Red Marker shows Failures

Launch Sites distance





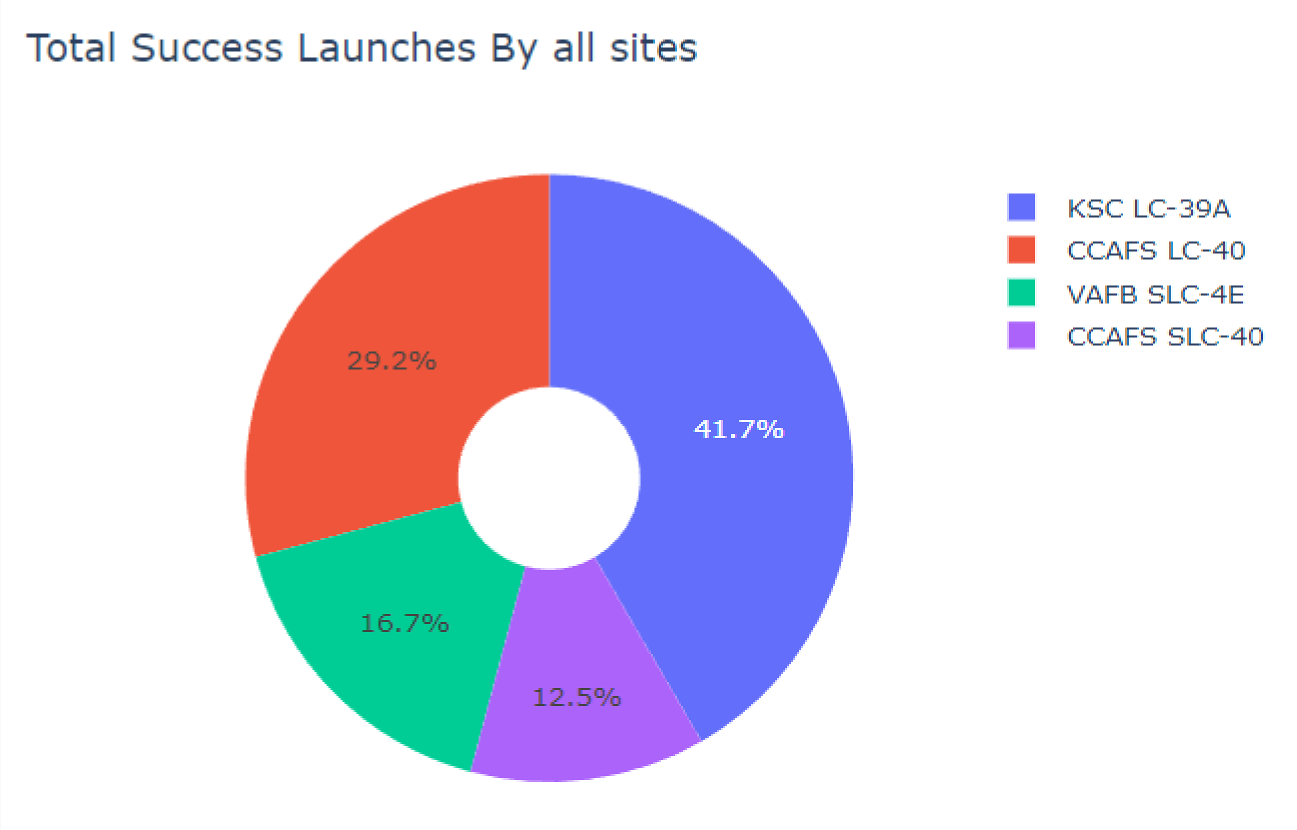
Find trends with Haversine formula using CCAFS-SLC-40 as a reference . 78,5 rm to City, 29 km to highway, 1 km to railway

Section 5

Build a Dashboard with Plotly Dash

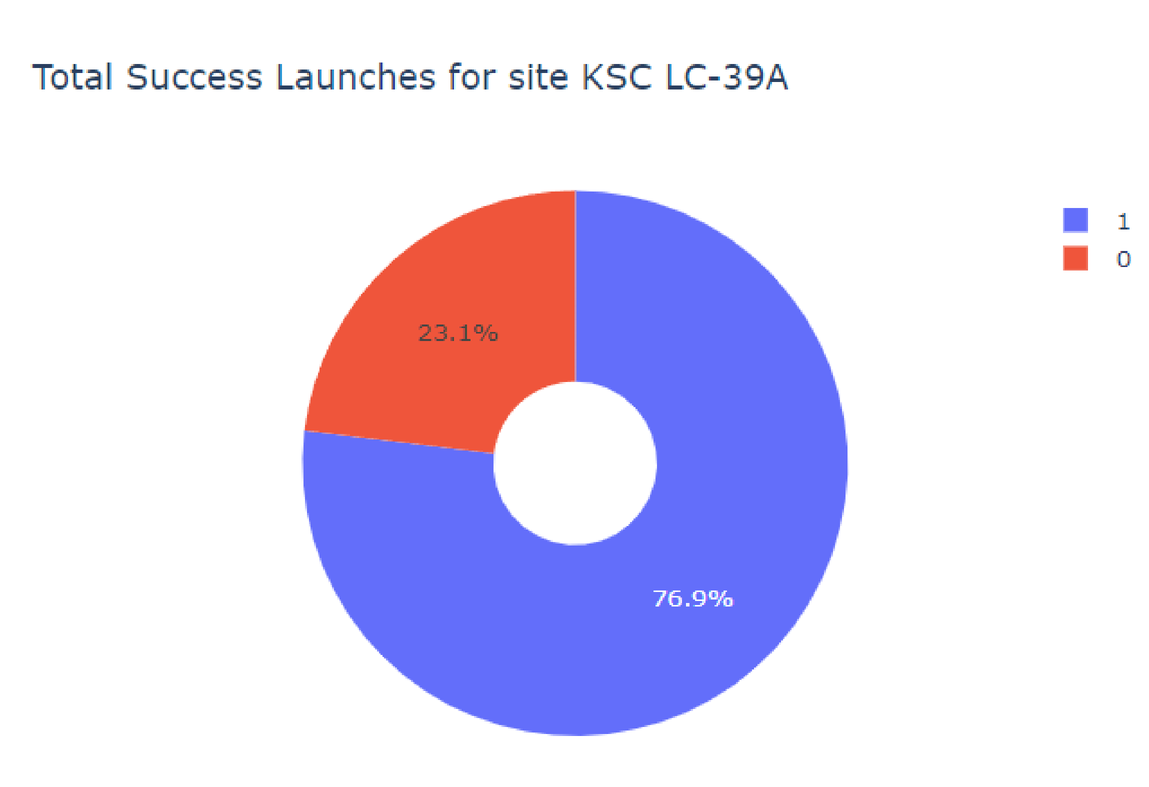
Success launches by all sites

KSC LC-39A had the most successful launches

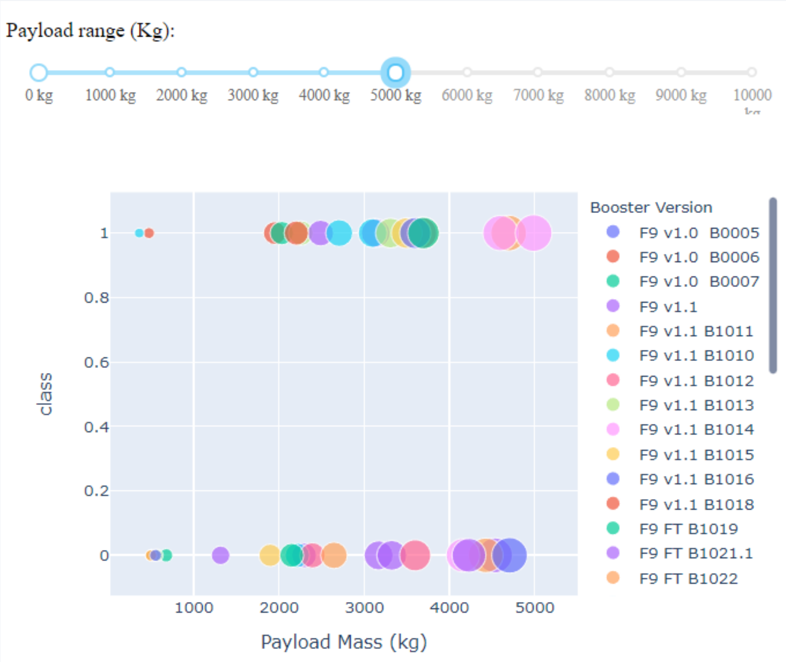


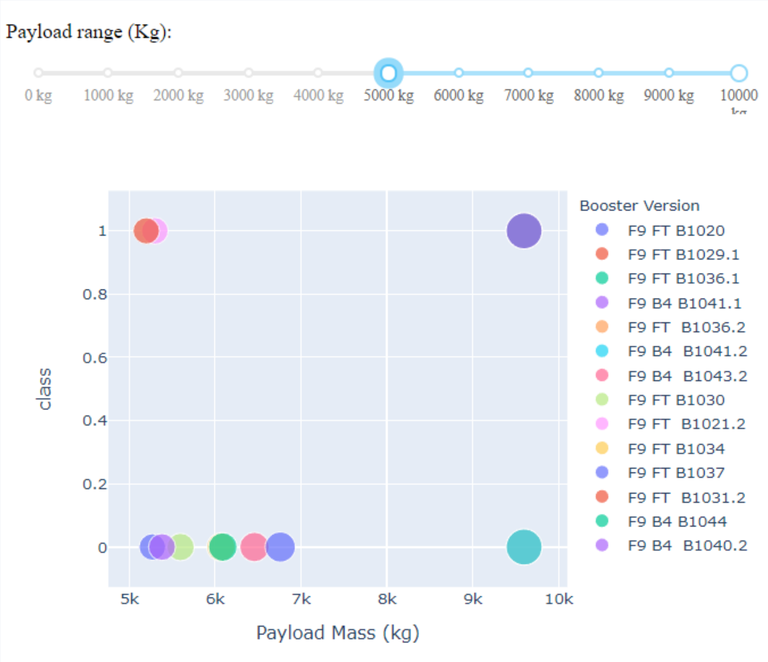
Launch site with highest launch success ratio

KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate



Payload vs. Launch Outcome



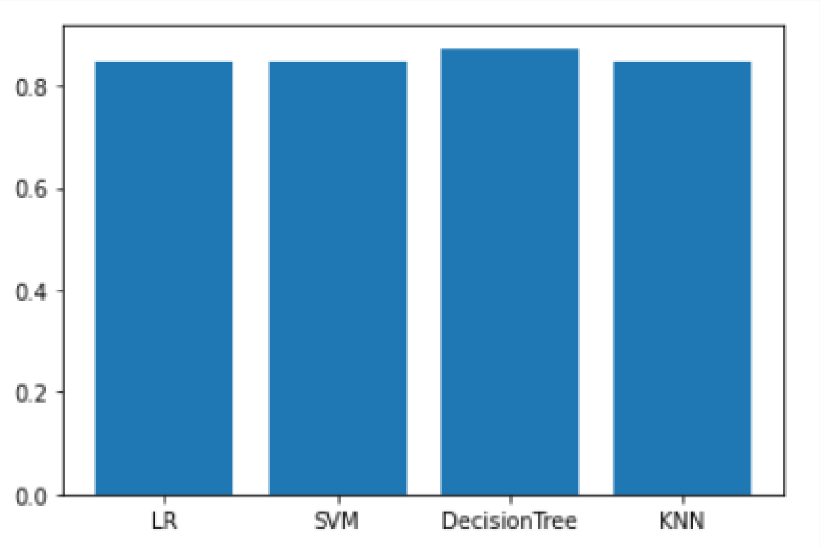


the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 6

Predictive Analysis (Classification)

Classification Accuracy




```
== Train set ==============
LR              0.8464285714285713
SVM             0.8482142857142856
DecisionTree    0.8732142857142857
KNN             0.8482142857142858

Best: DecisionTree  0.8732142857142857

== Test set =============
LR              0.8333333333333334
SVM             0.8333333333333334
DecisionTree    0.8888888888888888
KNN             0.8333333333333334

Best: DecisionTree  0.8888888888888888
```

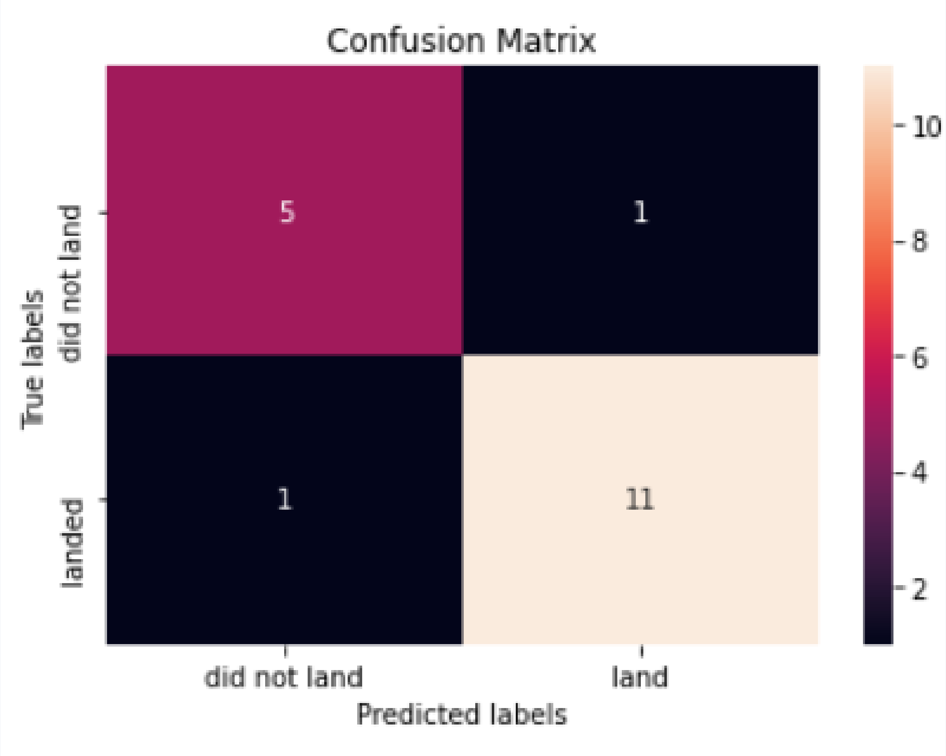
The best model is decision tree with tuned hyperparameters :

criterion: entropy, max_depth: 2, max_features: sqrt, min_samples_leaf: 4, min_samples_split: 10, splitter: best

Accuracy : 0.8732142857142857

Confusion Matrix

The confusion matrix is an indicator of classification success. We see that our model is quite accurate.



Conclusions

- KSC LC-39A had the most successful launch
- Orbit GEO,HEO,SSO,ES-L1 has the best result
- Low weighted payloads perform better than the heavier payloads
- The decision tree classifier is the best in our case (Accuracy : 0.88)

Appendix

- SpaceX API Documentation https://docs.spacexdata.com/
- Haversine formula https://en.wikipedia.org/wiki/Haversine_formula
- DASH https://dash.plotly.com/
- Github Repository https://github.com/koav/IBM-DataScience-SpaceX-Project

Thank you!