IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Margarita
8/17/2022

Outline

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

Executive Summary

- Public information from SpaceX api and Wikipedia was used to train a machine learning model and use it to predict if SpaceX reuses the first stage of the Falcon 9 rocket (if the launch is successful or not).
- Results of data analysis are visualized with graphs, maps and interactive dashboard.
- The best orbits and launch site were defined.

Introduction

- 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.
- Therefore if we can determine if the first stage will land, we can determine the cost of a launch.
- In this project a model was built to predict, if the Falcon 9 first stage lands successfully.

Section 1

Methodology

Methodology

Executive Summary

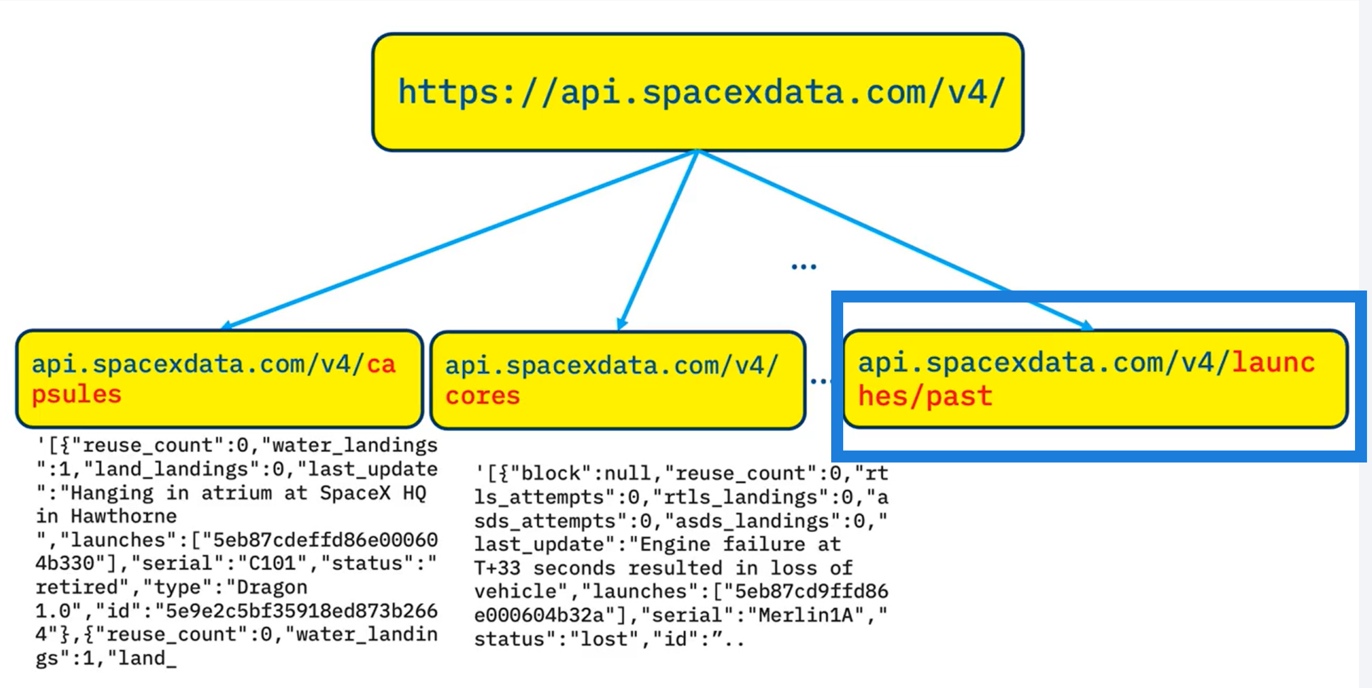
- Data collection methodology:
  - Data was collected by the Request to the SpaceX API;
  - Date was collected by web scraping to collect Falcon 9 historical launch records from a Wikipedia page;
- Exploratory data analysis (EDA) using visualization and SQL;
- Interactive visual analytics using Folium and Plotly Dash;
- Predictive analysis using classification models:
  - Best Hyperparameter for SVM, Classification Trees, Logistic Regression and KNN were found;
  - The best result was achieved with Classification Trees.

Data Collection

- 2 methods of data collection were used:
  - From SpaceX api
  - As Web Scrapping from Wikipedia

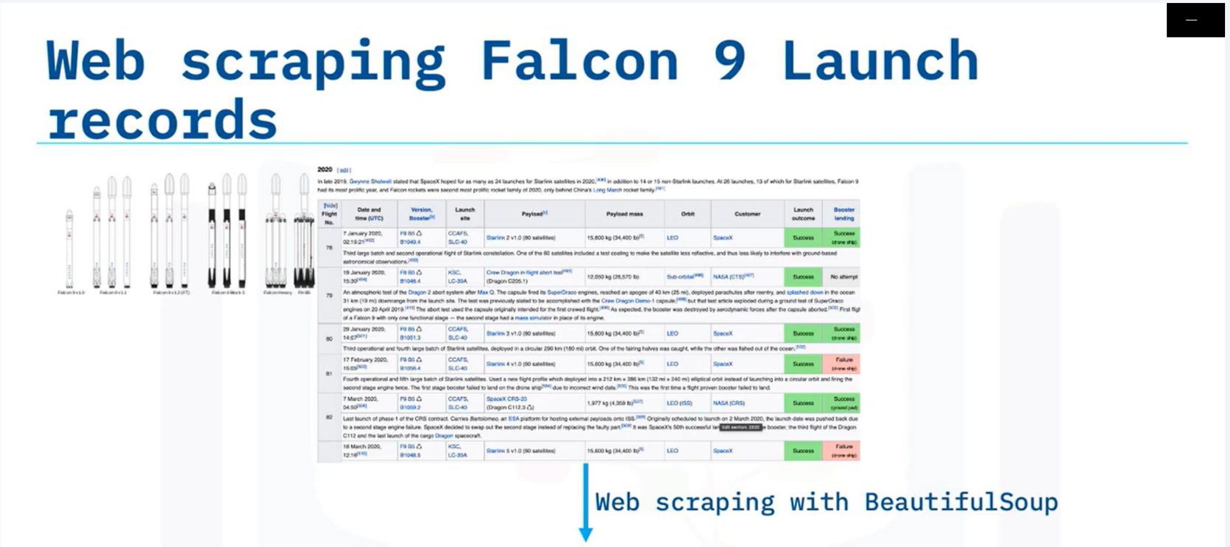
Data Collection – SpaceX API

- Api.spacexdata.com/v4/launches/past was used;
- Get request was performed with a list of .json objects as result ;
- Converted into Pandas Data Frame with normalize function;
- GitHub URL of the completed SpaceX API calls notebook



Data Collection – Web Scraping

- Web scraping with BeautifulSoup package from Wikipedia (static version url from at 11:39, 9 June 2021);
- Parced and converted data into Pandas Data Frame (id numbers were replaced by data from API, null values were replaced by mean values)
- Falcon 1 data was removed from the table
- GitHub URL of the completed web scraping notebook



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCAFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Data Wrangling

- Exploratory Data Analysis (EDA) was performed to find some patterns in the data and determine what would be the label for training supervised models.
- In the data set, there are several different cases where the booster did not land successfully:
  - True Ocean means the mission outcome was successfully landed to a specific region of the ocean ;
  - 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.
- Those outcomes were converted into Training Labels y with 1 means the booster successfully landed 0 means it was unsuccessful.

- GitHub URL of the completed data wrangling related notebook

EDA with Data Visualization

- Scatter Plot, Bar chart and Line graph were used:
  - Scatter Plot showed the distribution of successful and unsuccessful launches and the relation between order of launches or payload mass and Launch Site, and Orbit.
  - Bar chart showed the success rate for each orbit.
  - Line graph was built to show the trend for success rate over the years 2010-2020.

- GitHub URL of the completed EDA with data visualization notebook

EDA with SQL

- SQL queries performed:
  - Unique Launch Sites were found using DISTINCT query;
  - Total payload was calculated using SUM query;
  - Average payload, which was carried by specific booster was found with AVG query;
  - First specific successful landing outcome was found using ORDER BY and ASC query;
  - Several queries were done using subquery, GROUP BY and taking into account DATE restrictions.

- GitHub URL of the completed EDA with SQL notebook

Build an Interactive Map with Folium

- The launch success rate may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.
- Finding an optimal location for building a launch site certainly involves many factors and we could discover some of the factors by analyzing the existing launch site locations.
    - Markers and circles for Successful and Unsuccessful launches were used to help to choose the location with highest success rate.
    - Using lines and we could indicate proximities with Neighbour objects such as roads and railways.
- GitHub URL of the completed interactive map with Folium map

Build a Dashboard with Plotly Dash

- This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
- Pie chart was chosen to show the rate of successful and unsuccessful launches.
- Scatter plot with payload range slider helps to choose the payload range with higher success rate.

- GitHub URL of the completed Plotly Dash
- GitHUB URL for Dashboard

Predictive Analysis (Classification)

- Machine learning pipeline was built to predict if the first stage of the Falcon 9 lands successfully. It included:
  - Preprocessing to standardize the data,
  - Train_test_split to split the data into training and testing data,
  - Train the model and perform Grid Search to find the hyperparameters that allow a given algorithm to perform best,
  - Using the best hyperparameter values to determine the model with the best accuracy using the training data.
- Were trained: Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors.
- Based on accuracy and the confusion matrix the best algorithm was chosen :
  - **Decision Tree Classifier.**
- GitHub URL of the completed predictive analysis

Results

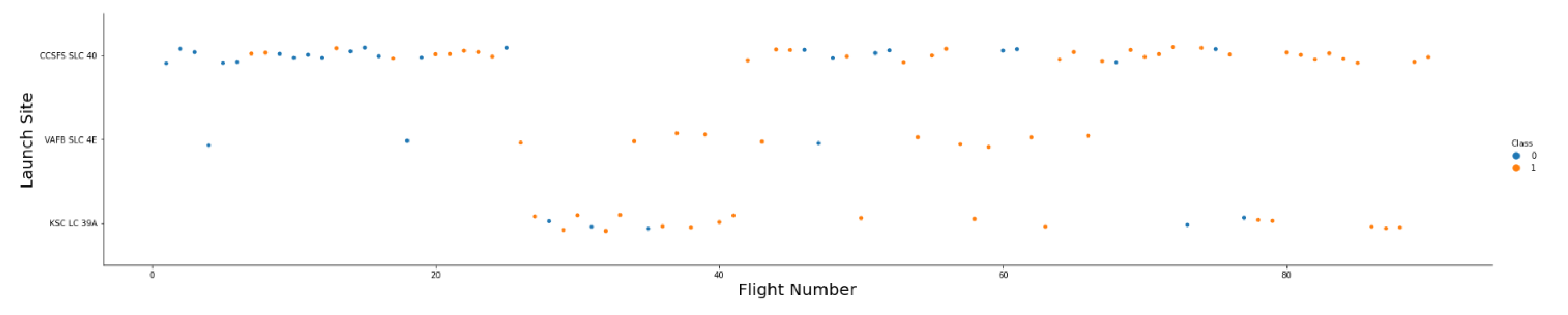
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site
- In this graph we can see that the first launches and the majority of launches were done from CCAFS SLC-40, the latest launches were successful from all Sites.



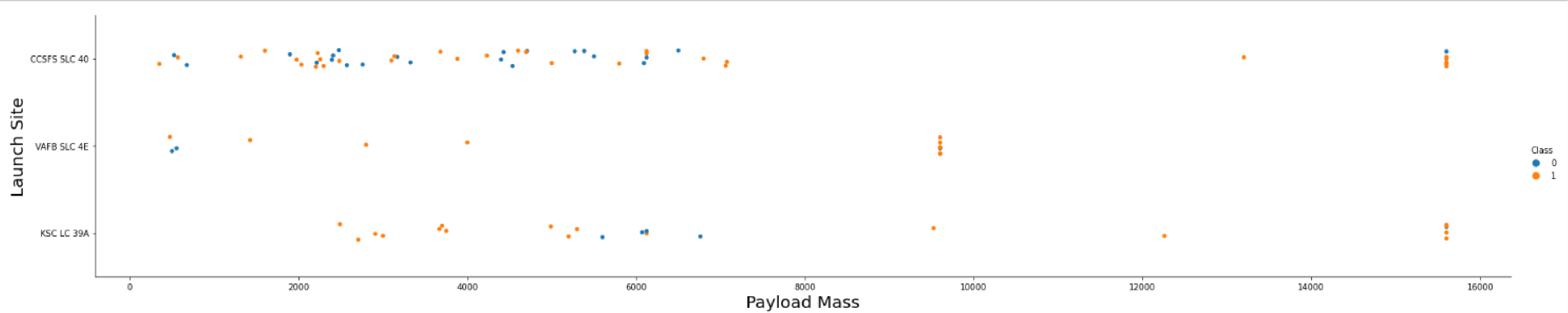
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

We can see that from the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Also the majority of heavy payload launches were successful.

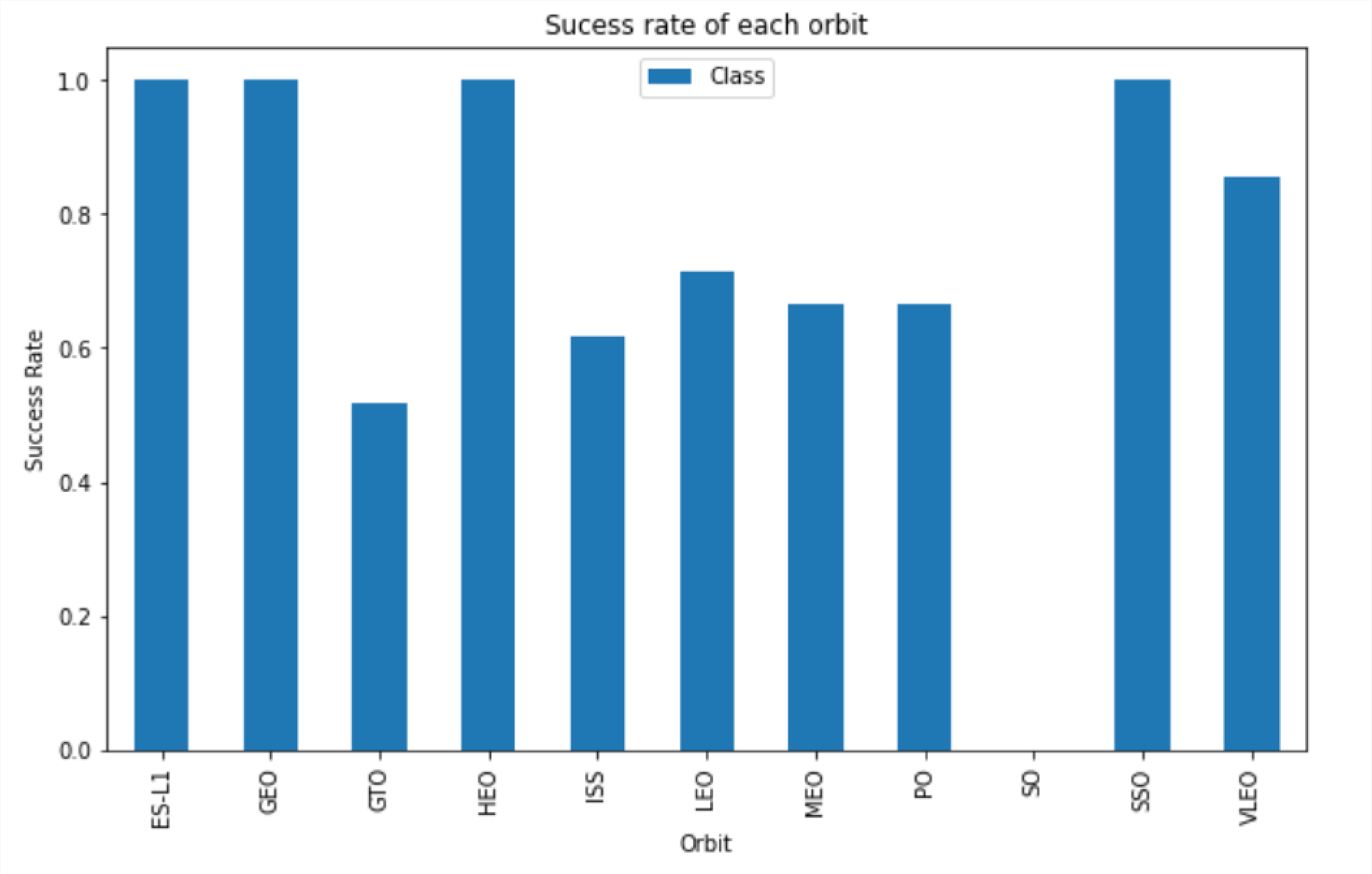
There seems to be no relationship between payload mass and success rate for CCAFS SLC-40 Site.



Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type.
- Orbit types are described in the Appendix.

- The highest success rates have ES-L1, GEO, HEO and SSO orbits.
- The lowest one – GTO.
- And SO orbit has no rate (most probably it's a mistype of SSO)

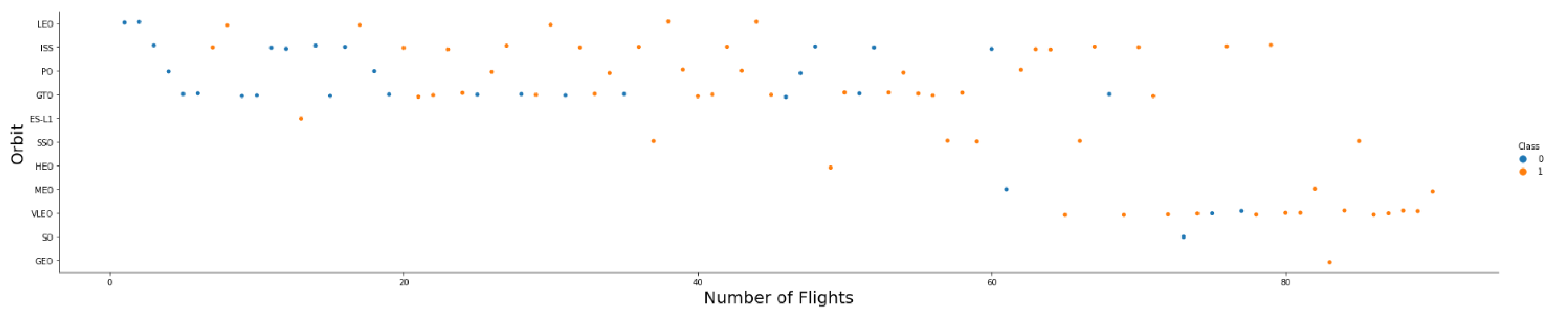


Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type

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

Some orbits have a very limited number of launches (such as GEO with 1 launch) - so it would be impossible to use this data for predictions for these orbits.

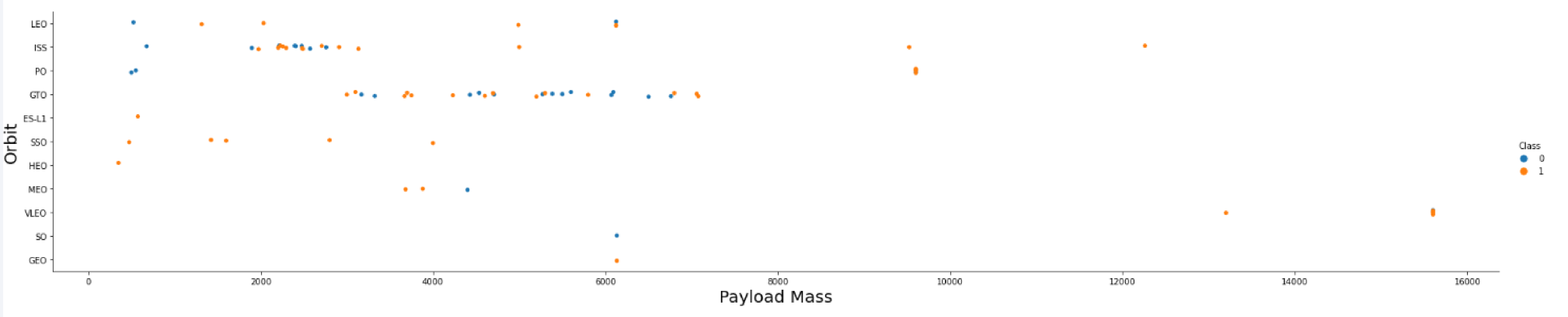


Payload vs. Orbit Type

- Scatter point of payload vs. orbit type

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

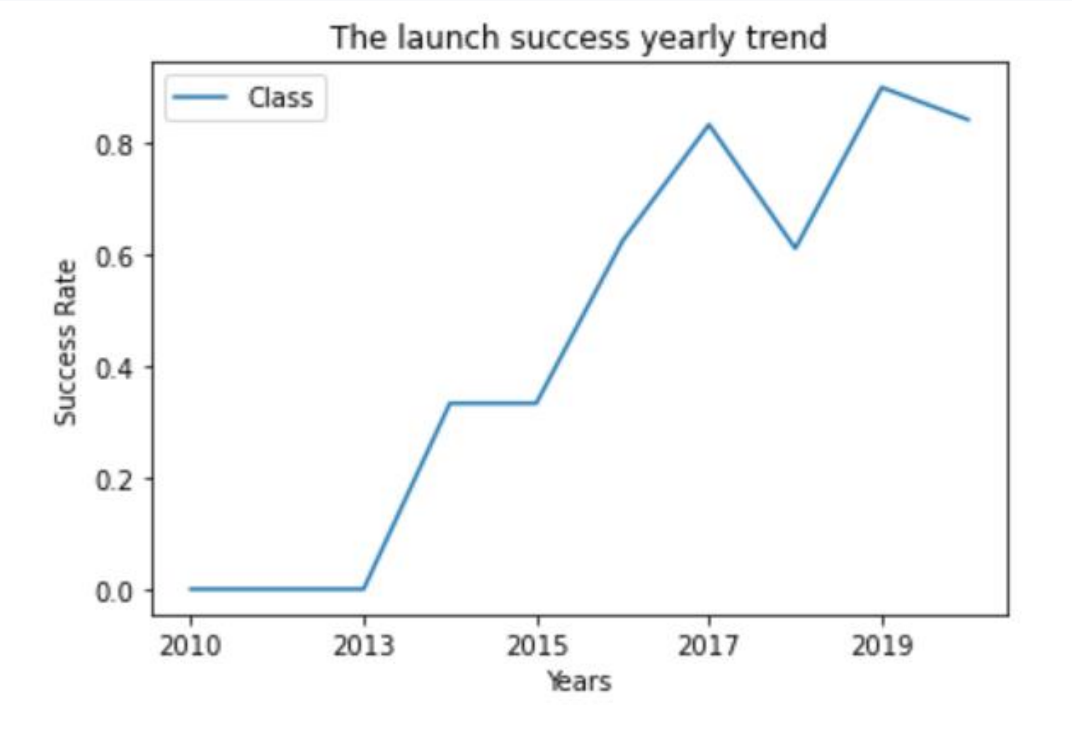
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are there .



Launch Success Yearly Trend

- Line chart of yearly average success rate between 2010 and 2020.

- The success rate since 2013 keeps increasing, which is well explained with the growth of the maturity level of the SpaceX researches and experience.



All Launch Site Names

There are 4 unique Launch Sites in our table:

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

```
%sql SELECT DISTINCT("Launch_Site") FROM SPACEXTBL;
```

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

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

The result is an SQL query via SQLite in Python.

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`:

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE '%CCA%' LIMIT 5;
```

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

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)
08-12-2010	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)
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

The result is an SQL query via SQLite in Python.

Total Payload Mass

The total payload carried by boosters from NASA: 45596 kg.

```
%sql SELECT "Customer", SUM("PAYLOAD_MASS__KG_") AS "Total_Payload" FROM SPACEXTBL WHERE "Customer" LIKE 'NASA (CRS)';
```

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

Customer	Total_Payload
NASA (CRS)	45596

The result is an SQL query via SQLite in Python.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1: 2928.4 kg.

```
%sql SELECT "Booster_Version", AVG("PAYLOAD_MASS__KG_") AS "Average_Payload" FROM SPACEXTBL WHERE "Booster_Version" LIKE 'F9 v1.1
```

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

Booster_Version	Average_Payload
F9 v1.1	2928.4

The result is an SQL query via SQLite in Python.

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad:

22nd of December 2015

```
# !!! Landing Outcome has space in the name, that's why it's mandatory to use `` and not "" or ''  !!!
%sql SELECT "Date", `Landing _Outcome` FROM SPACEXTBL WHERE `Landing _Outcome` LIKE 'Success%' ORDER BY date("Date") ASC LIMIT 1;
```

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

Date	Landing _Outcome
22-12-2015	Success (ground pad)

The result is an SQL query via SQLite in Python.

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: 4 names

```
%sql SELECT "Booster_Version", "PAYLOAD_MASS__KG_", `Landing _Outcome` FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_">4000 AND "PAYLOAD
```

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

Booster_Version	PAYLOAD_MASS__KG_	Landing _Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

The result is an SQL query via SQLite in Python.
Full code:
%sql SELECT "Booster_Version", "PAYLOAD_MASS__KG_", `Landing _Outcome` FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_">4000 AND "PAYLOAD_MASS__KG_"<6000 AND `Landing _Outcome` LIKE 'Success (drone ship)';

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes:

There is just 1 failure and 100 successful missions

```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") AS "Total_number" FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

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

The result is an SQL query via SQLite in Python.

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass:

```
%sql SELECT "Booster_Version", "PAYLOAD_MASS__KG_", `Landing _Outcome` FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_"=(SELECT MAX("PAYL
```

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

Booster_Version	PAYLOAD_MASS__KG_	Landing _Outcome
F9 B5 B1048.4	15600	Success
F9 B5 B1049.4	15600	Success
F9 B5 B1051.3	15600	Success
F9 B5 B1056.4	15600	Failure
F9 B5 B1048.5	15600	Failure
F9 B5 B1051.4	15600	Success
F9 B5 B1049.5	15600	Success
F9 B5 B1060.2	15600	Success
F9 B5 B1058.3	15600	Success
F9 B5 B1051.6	15600	Success
F9 B5 B1060.3	15600	Success
F9 B5 B1049.7	15600	Success

The result is an SQL query via SQLite in Python, subquery was used:

%sql SELECT "Booster_Version","PAYLOAD_MASS__KG_", `Landing _Outcome` FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_"=(SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015:
  - There 2 cases:

```
%sql SELECT "Date", substr("Date",4,2) AS "Month" , "Booster_Version", "Launch_Site",  `Landing _Outcome` FROM SPACEXTBL WHERE `L
```

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

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

The result is an SQL query via SQLite in Python.
Code:
%sql SELECT "Date", substr("Date",4,2) AS "Month" , "Booster_Version", "Launch_Site", `Landing _Outcome` FROM SPACEXTBL WHERE `Landing _Outcome` LIKE 'Failure%' and substr("Date",7,4) LIKE '2015' ORDER BY date("Date") DESC

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

- Landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:
  - There are 34 Successful outcomes in this period of time.
  - There 9 Unsuccessful outcomes in the same period.
- The result is an SQL query via SQLite in Python using COUNT, GROUP BY and ORDER BY:

```
#Successful outcomes:
%sql SELECT `Landing _Outcome`, COUNT(`Landing _Outcome`) AS "Count" FROM SPACEXTBL WHERE `Landing _Outcome` LIKE 'Success%' \
AND substr("Date",1,2)||substr("Date",4,2)||substr("Date",7,4) BETWEEN '04062010' AND '20032017' GROUP BY `Landing _Outcome` \
ORDER BY "Count" DESC;
```

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

Landing _Outcome	Count
Success	20
Success (drone ship)	8
Success (ground pad)	6

```
#Unsuccessful outcomes:
%sql SELECT `Landing _Outcome`, COUNT(`Landing _Outcome`) AS "Count" FROM SPACEXTBL WHERE `Landing _Outcome` LIKE 'Failure%' \
AND substr("Date",1,2)||substr("Date",4,2)||substr("Date",7,4) BETWEEN '04062010' AND '20032017' GROUP BY `Landing _Outcome` \
ORDER BY "Count" DESC;
```

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

Landing _Outcome	Count
Failure (drone ship)	4
Failure	3
Failure (parachute)	2

Section 3

Launch Sites Proximities Analysis

Launch Sites on the map

- 4 Launch Sites were displayed on the map using Follium.Map():

#	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

- We can see that VAFB SLC-4E Site is located far away from 3 other Sites grouped together.

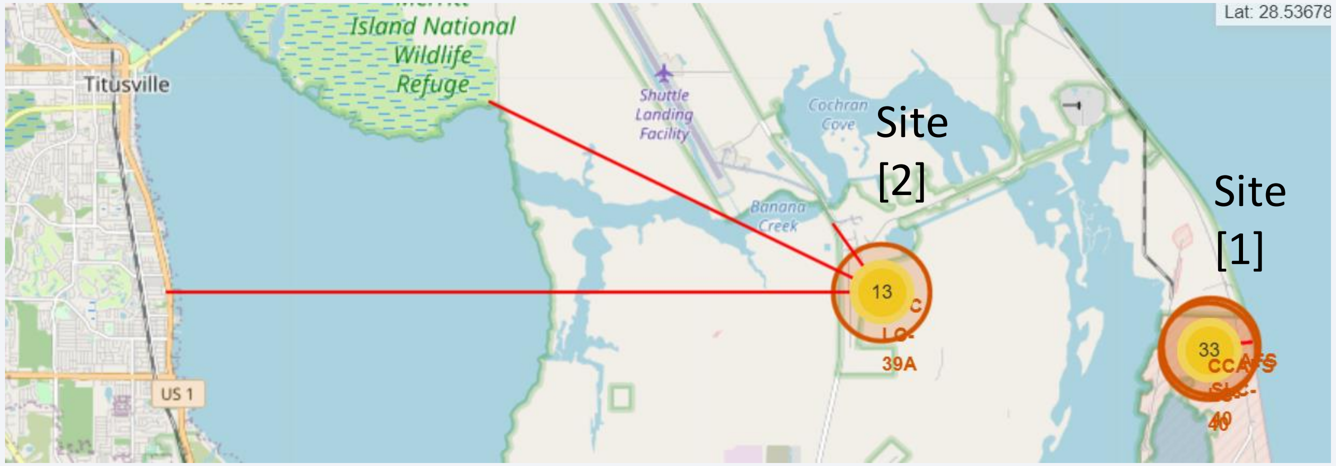
Launch outcomes for each site

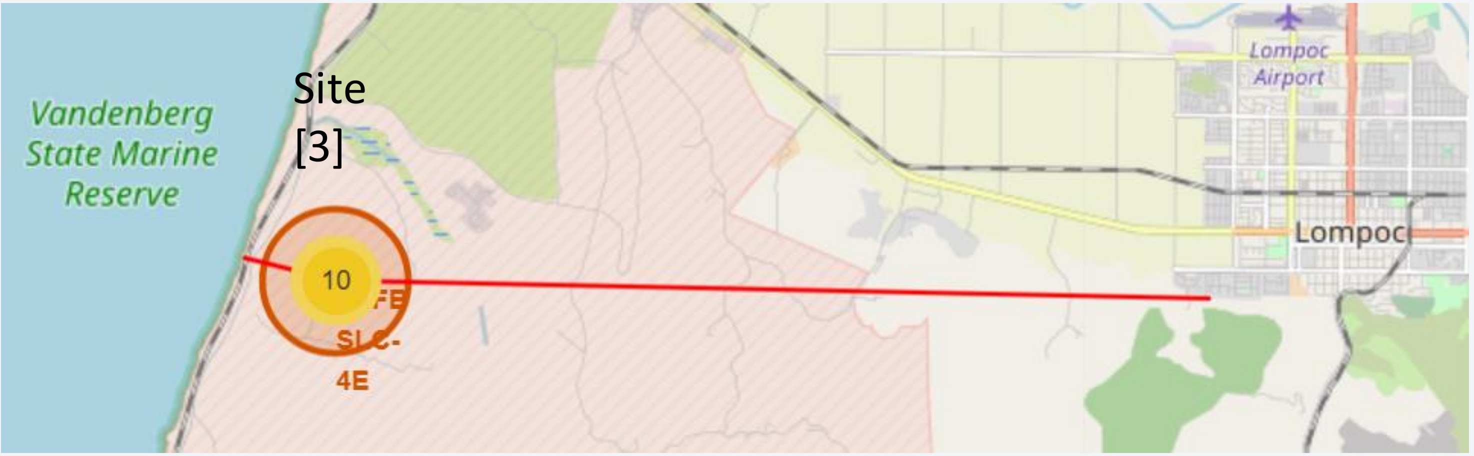
- Enhanced map by addition the launch outcomes (information from Class column) for each site shows that KSC LC-39A site has the highest success rates.
- Color green/red was chosen according to 1/0 Class.

The distances between a launch site to its proximities

- Plugins.MousePosition was added to the Folium.Map for fast identefication of coordinates of neighbour objects.
- Site [3] is close to coastal line and Marine Reserve (1,2km) and Lompoc city (12km).
- Site [2] is close to Wildlife Refege (9km), city Titusville with its railway, highways and airport (15km), Banana Creek river (1,7km). But Lompoc city is smaller than Titusville.
- Site [1] is the closest to coastline (0,8km) but further from other surroundings than Site [2] and [3].

Site [2] = KSC LC-39A is the closest to surrounding infrastructure.



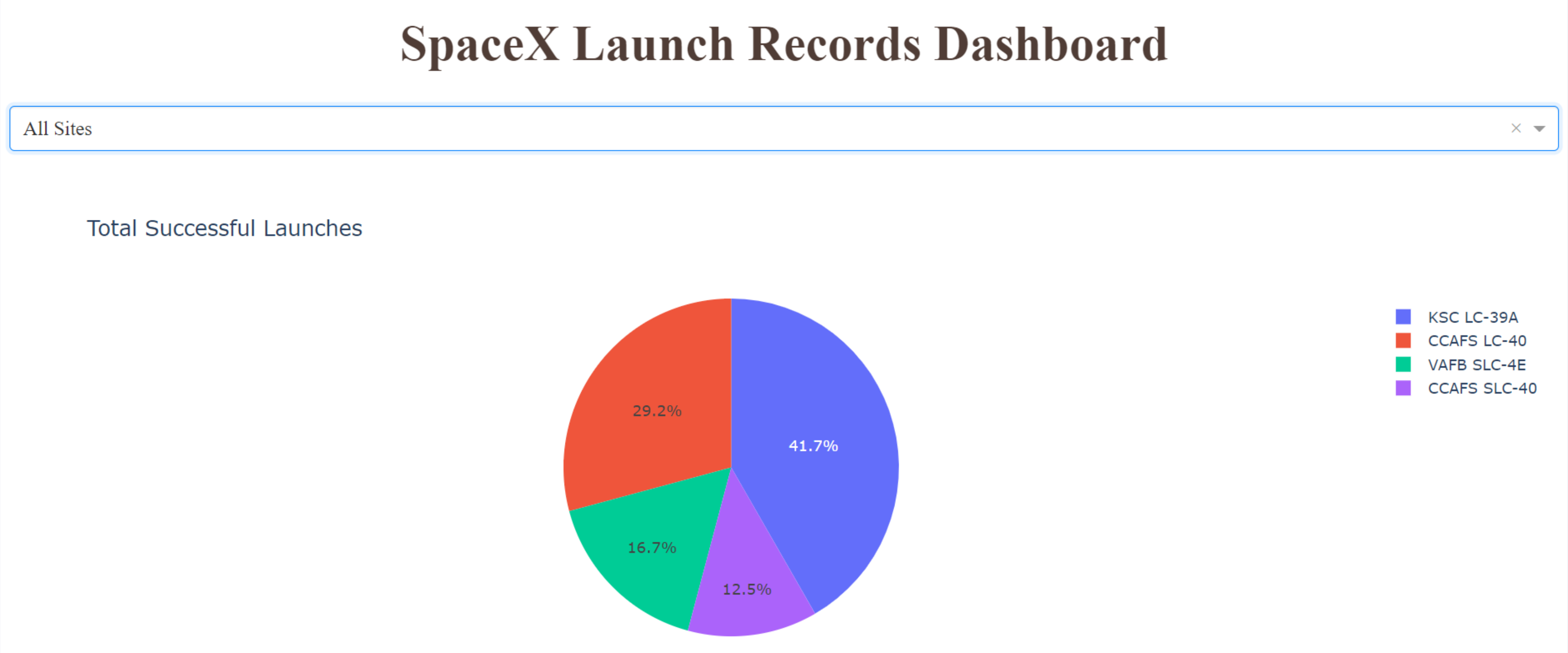


Section 4

Build a Dashboard with Plotly Dash

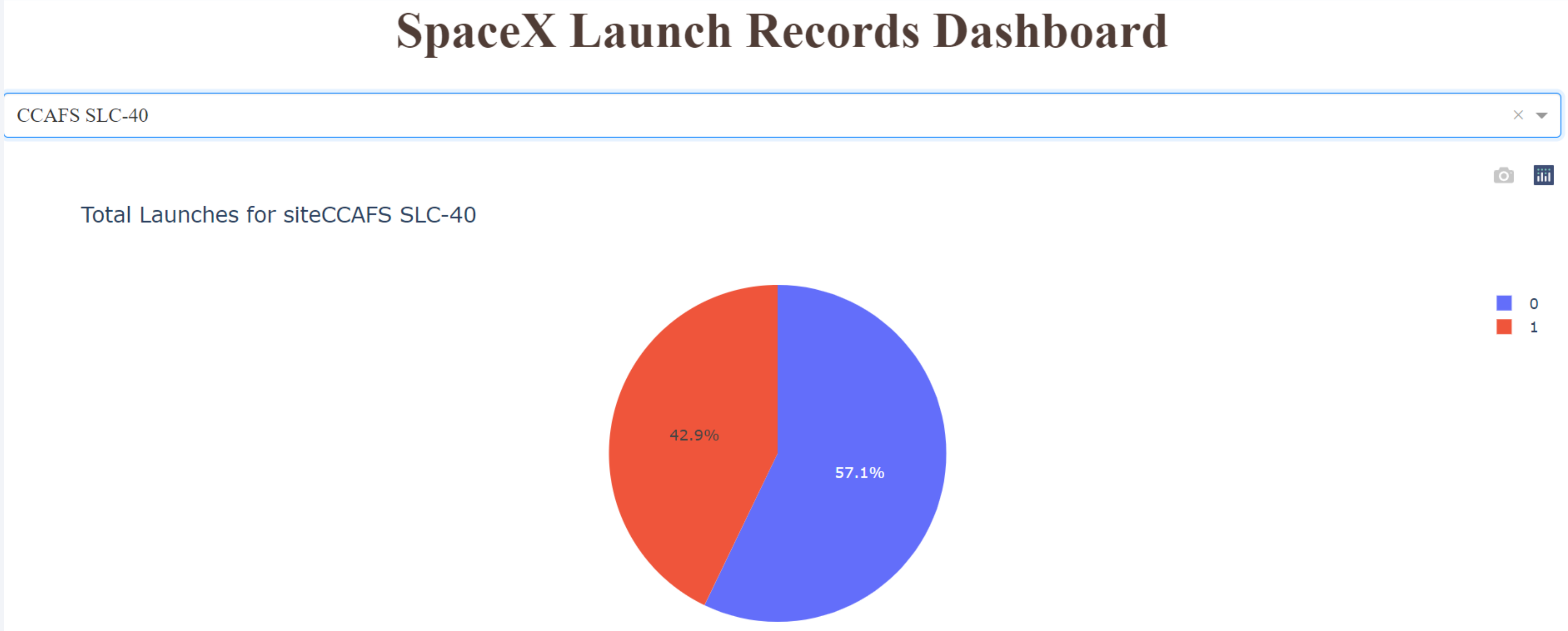
Dashboard with All Sites selection

- We can see from the Pie Chart, that the biggest number of successful launches has KSC LC-39A Site and the lowest number – CCAFS SLC-40.



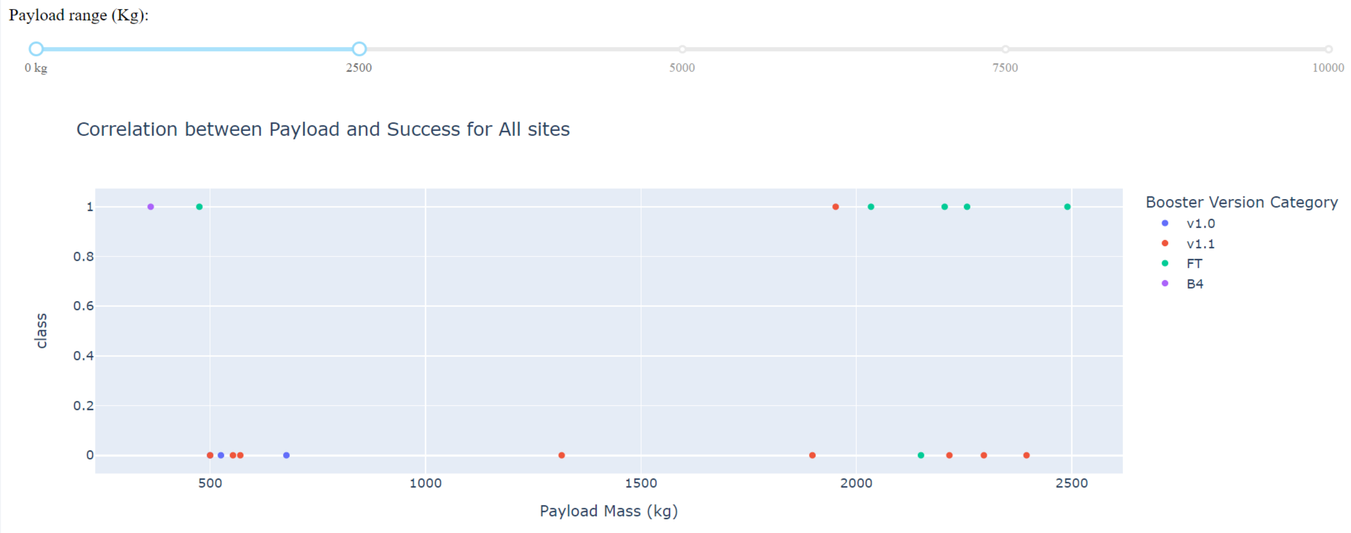
Dashboard with CCAFS LC-40 selection

- The Site CCAFS LC-40 has the highest ratio of Success Launches



Dashboard with Payload distribution

- Based on the scatter plots for All Sites we see, that FT Booster Version has the highest success rate.
- Also higher success rate have launches with smaller payload between 2 and 6k kg.



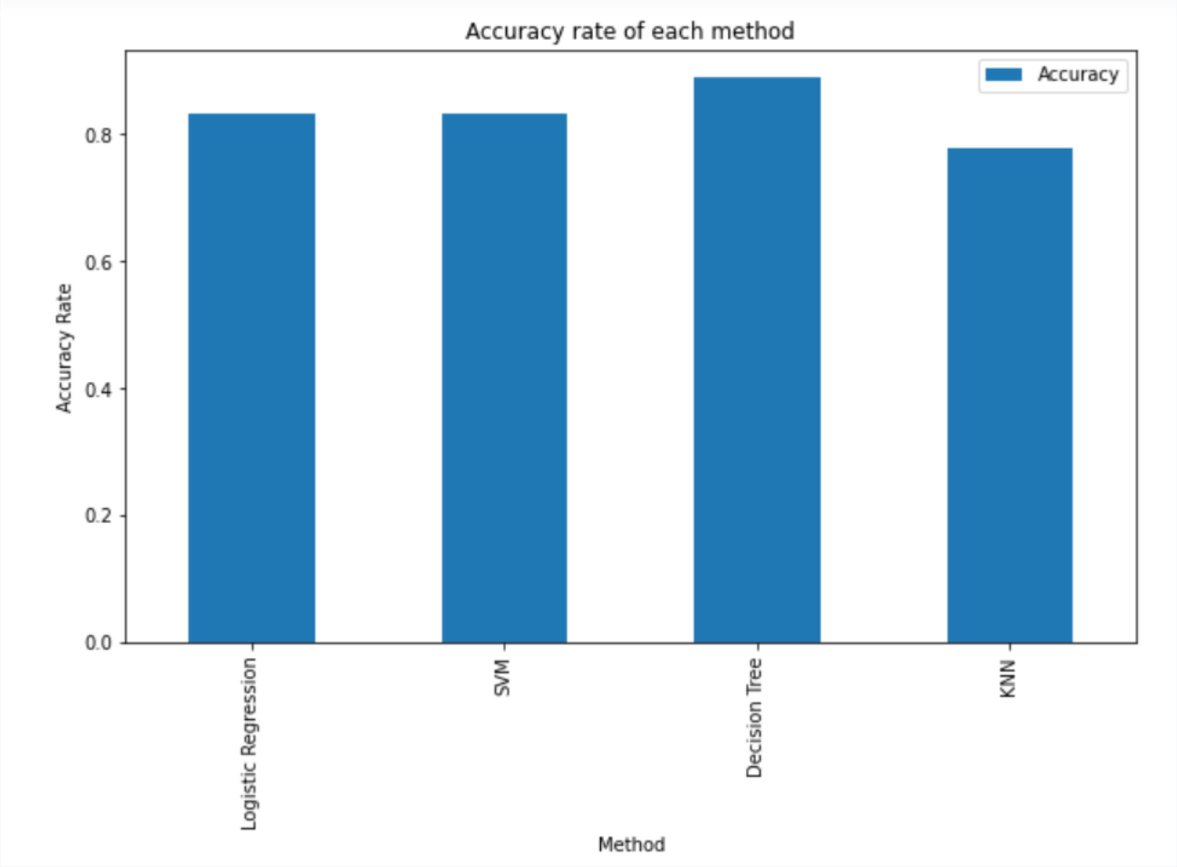
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The highest accuracy has Decision Tree Classificator - 0.888889
-

	Method	Accuracy
0	Logistic Regression	0.833333
1	SVM	0.833333
2	Decision Tree	0.888889
3	KNN	0.777778



Confusion Matrix

- The best performing model - **Decision Tree Classifier** - can distinguish between the different classes. It found just 1 false positive and 1 false negative (which are the less critical in terms of this project).

Conclusions

- The highest success rates have ES-L1, GEO, HEO and SSO orbits.
- KSC LC-39A site has the highest success rates among the sites, but in the same time this site is the closest one to civil infrastructure.
- The best prediction for launches result can be done with Decision Tree classifier with following accuracy:
  - `Jaccard Score:  0.846`
  - `F1 Score:  0.9167`
  - `Accuracy:  0.889`

Appendix

Orbits description:

- **LEO**: Low Earth orbit (LEO)is an Earth-centred orbit with an altitude of 2,000 km (1,200 mi) or less (approximately one-third of the radius of Earth),[1] or with at least 11.25 periods per day (an orbital period of 128 minutes or less) and an eccentricity less than 0.25.[2] Most of the manmade objects in outer space are in LEO [1].
- **VLEO**: Very Low Earth Orbits (VLEO) can be defined as the orbits with a mean altitude below 450 km. Operating in these orbits can provide a number of benefits to Earth observation spacecraft as the spacecraft operates closer to the observation[2].
- **GTO** A geosynchronous orbit is a high Earth orbit that allows satellites to match Earth's rotation. Located at 22,236 miles (35,786 kilometers) above Earth's equator, this position is a valuable spot for monitoring weather, communications and surveillance. Because the satellite orbits at the same speed that the Earth is turning, the satellite seems to stay in place over a single longitude, though it may drift north to south," NASA wrote on its Earth Observatory website [3] .
- **SSO (or SO)**: It is a Sun-synchronous orbit also called a heliosynchronous orbit is a nearly polar orbit around a planet, in which the satellite passes over any given point of the planet's surface at the same local mean solar time [4] .
- **ES-L1** :At the Lagrange points the gravitational forces of the two large bodies cancel out in such a way that a small object placed in orbit there is in equilibrium relative to the center of mass of the large bodies. L1 is one such point between the sun and the earth [5] .
- **HEO** A highly elliptical orbit, is an elliptic orbit with high eccentricity, usually referring to one around Earth [6].
- **ISS** A modular space station (habitable artificial satellite) in low Earth orbit. It is a multinational collaborative project between five participating space agencies: NASA (United States), Roscosmos (Russia), JAXA (Japan), ESA (Europe), and CSA (Canada) [7]
- **MEO** Geocentric orbits ranging in altitude from 2,000 km (1,200 mi) to just below geosynchronous orbit at 35,786 kilometers (22,236 mi). Also known as an intermediate circular orbit. These are "most commonly at 20,200 kilometers (12,600 mi), or 20,650 kilometers (12,830 mi), with an orbital period of 12 hours [8]
- **HEO** Geocentric orbits above the altitude of geosynchronous orbit (35,786 km or 22,236 mi) [9]
- **GEO** It is a circular geosynchronous orbit 35,786 kilometres (22,236 miles) above Earth's equator and following the direction of Earth's rotation [10]
- **PO** It is one type of satellites in which a satellite passes above or nearly above both poles of the body being orbited (usuallya planet such as the Earth [11]

Thank you!