IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Stefanos Nikolaou

October 2021

Outline

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

Executive Summary

- The global activity of the rocket launch industry grows rapidly and Data Science can play a significant role in optimizing the attempts by uncovering various insights
- As this project relies on Data Science, a set of related methodologies were followed for collecting, manipulating and analysing data
- Exploratory Data Analysis was performed for exploring the data and revealing insights
- Data Visualizations with Matplotlib, Interactive Maps with Folium and Interactive Dashboards with Plotly were created
- Finally Machine Learning was used for creating classifiers that predict the mission outcome and thus allowing us to make recommendations
- Generally, results show that there is an interaction between the outcome of a mission and features such as the payload mass

Introduction

- This report is the final submission of the Applied Data Science Capstone which is the final course of the IBM Data Science Professional Certificate
- The objective of the report is to analyse the SpaceX past launches records to reveal insights regarding successful launches
- According to different providers, rocket launches cost up to 165 million dollars
- SpaceX reuses the first stage of the rocket and therefore can save a lot of money. Advertisements show costs of 62 million dollars
- The main objective of this project is to determine if the first stage of the rocket will land successfully and eventually the cost of a launch
- This information can be used if other companies want to bid against SpaceX for a rocket launch

Section 1

Methodology

Methodology

Executive Summary

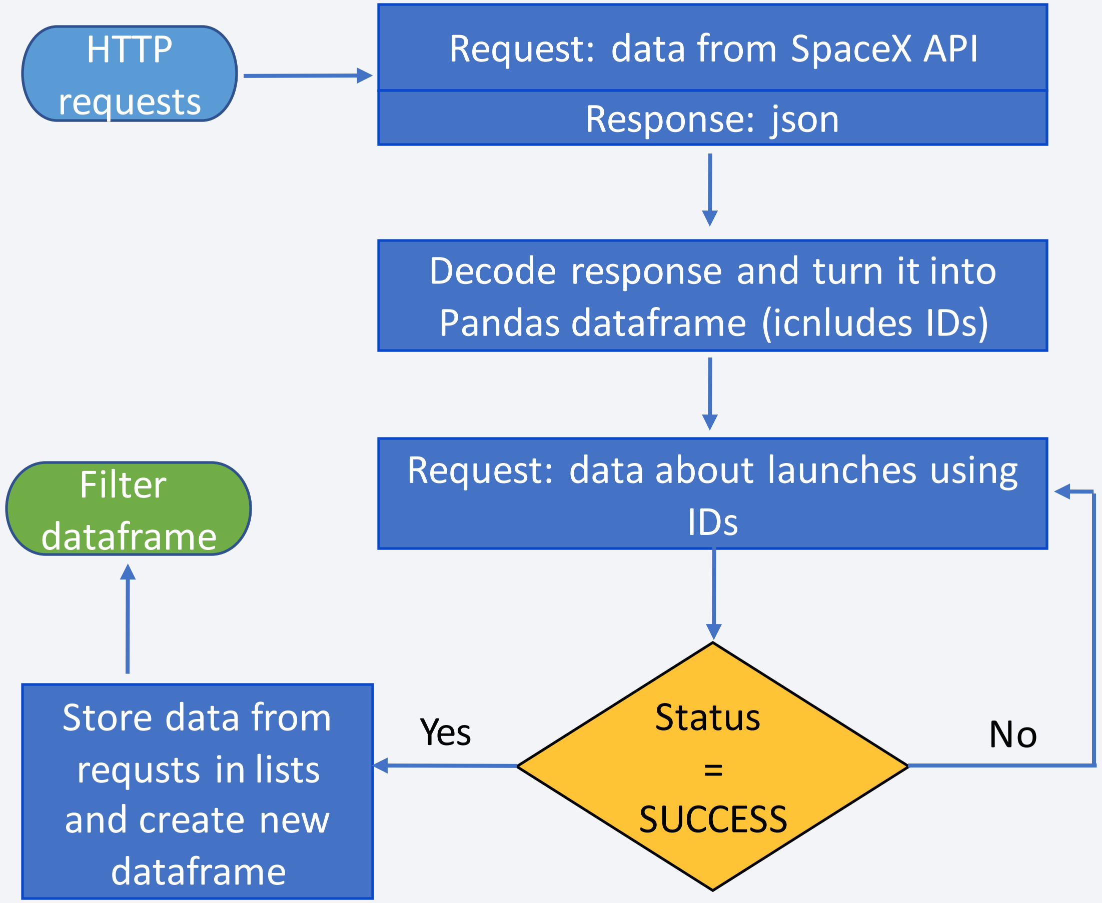
- Data collection methodology:
  - Collect data from the SpaceX API by making HTTP requests
  - Collect data by performing web scraping and using Python BeautifulSoup
- Perform data wrangling
  - Perform EDA to find patterns in the data and determine what would be the label for training supervised machine learning models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize data, split into training and testing sets and find the best hyperparameters

Data Collection

- For this project we are interested in the SpaceX Falcon 9 rocket and therefore the data used is related to Falcon 9 past launches only.
- Data sets were collected in two stages:
  - Data collection with SpaceX API HTTP requests:
    - Create Pandas dataframe and save it as a csv file
  - Data collection with web scraping HTML Wikipedia page:
    - Create Pandas dataframe and save it as a csv file

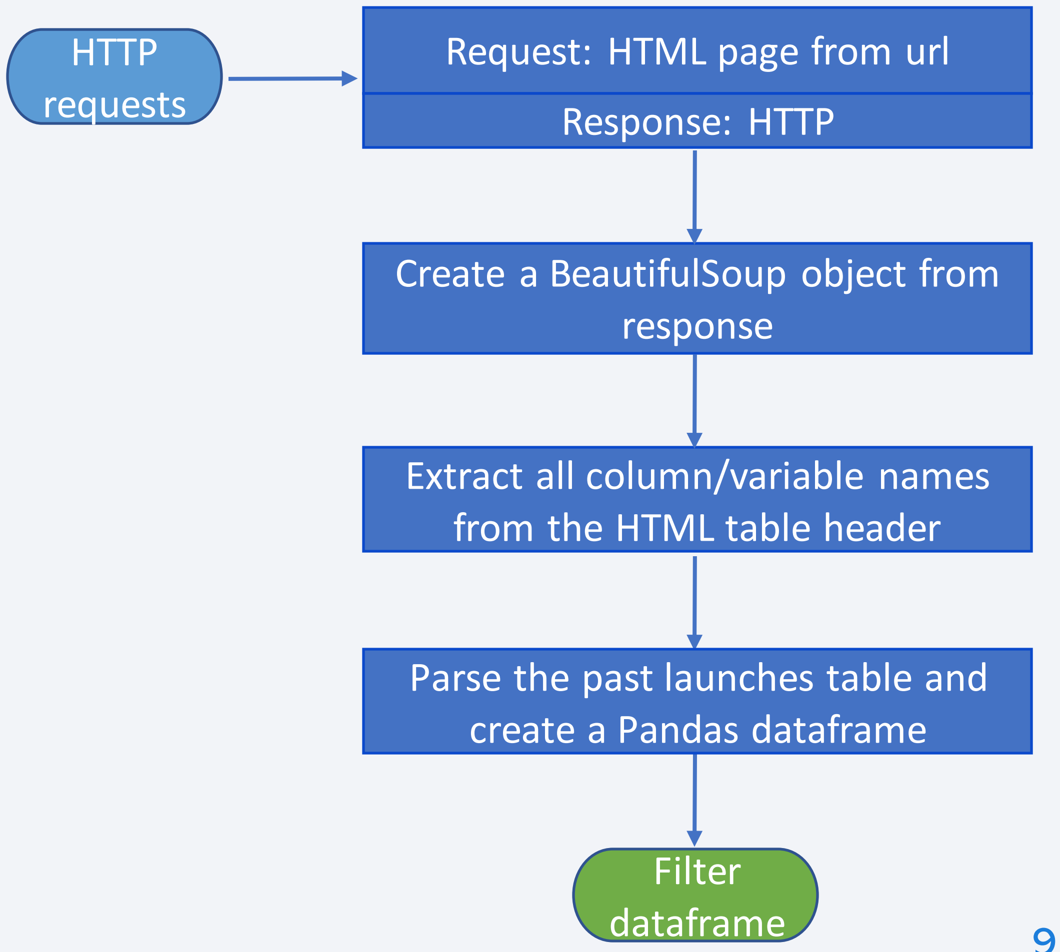
Data Collection – SpaceX API

- Request and parse the SpaceX launch data using the get request
- Manipulate the response so that it can be turned into a Pandas dataframe:
  - Decode the response content as a json
- Using the API again, get information about the launches using the IDs
- Data wrangling by dealing with missing values
- Code is available on Github right [here]



Data Collection – Scraping

- Request the List of Falcon 9 launches Wikipedia page using the HTTP get method
- Web scrap Falcon 9 launch records using Python library BeautifulSoup
- Extract the Falcon 9 past launches HTML table
- Parse the table and convert it into a Pandas dataframe
- Code is available on Github right here



Dataframes

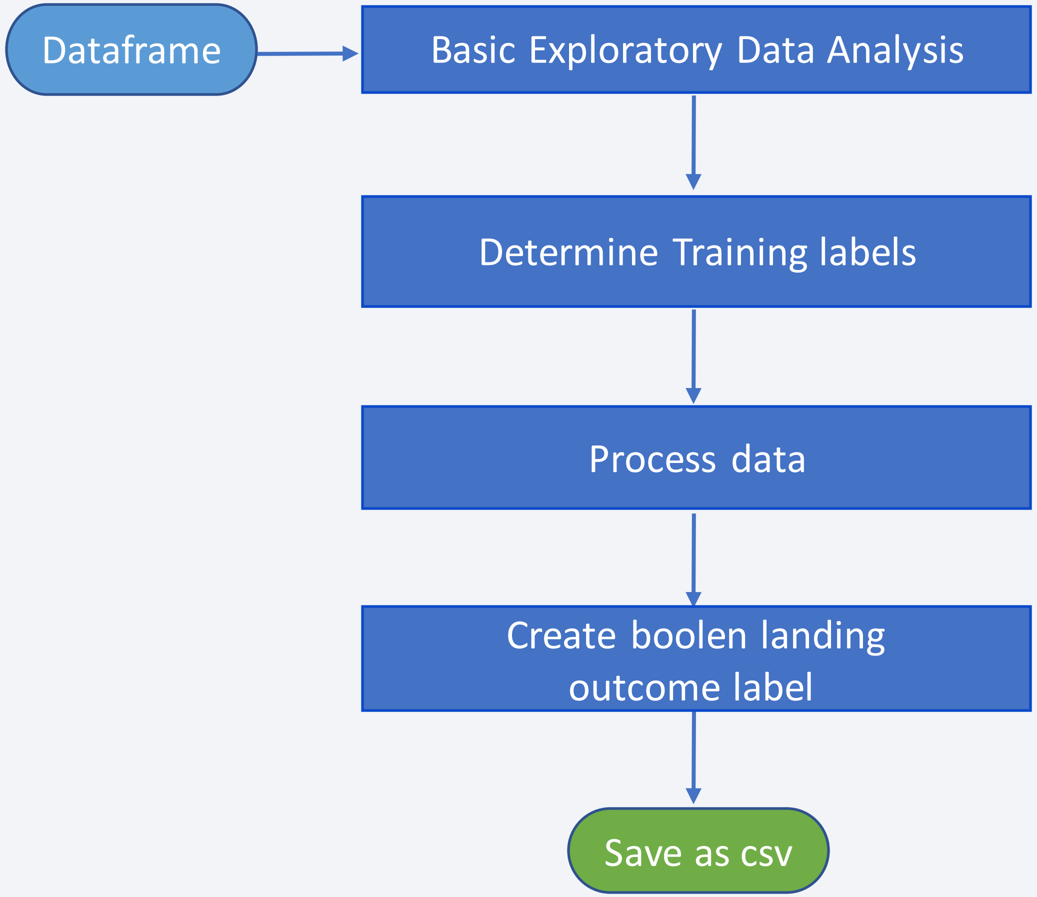
The two Pandas dataframes obtained can be seen below, the first being the first 5 lines of the result of API requests and the second being the result of web scraping

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA (COTS)\nNRO	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	SpaceX Capella Space and Tyvak	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2	Success	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA (CRS)	Success\n	F9 B5B1067.1	Success	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success	6 June 2021	04:26

Data Wrangling

- The main objective of this part is to perform some EDA on the first dataframe and determine the training labels
- Process data by calculating the number of occurrences of attributes, e.g. the number of launches on each site
- Create a Boolean landing outcome label from Outcome column
- Code is available on Github and can be found right [here]



EDA with Data Visualization

- Perform Exploratory Data Analysis and Feature Engineering with Data Visualization using Pandas, Matplotlib and Seaborn
- First we want to see how attributes affect each other and essentially how they affect the launch outcome
- Seaborn provides a high – level interface for drawing informative statistical graphics and therefore we plot scatter plots to display the above relationships
- Scatter plots allow us to get insights by showing how two different attributes affect the target variable. This filtering is done by setting the "hue" parameter accordingly
- Barplots and lineplots were also used as they visualize the relationship between different attributes and the success rate and the yearly average launch success trend
- Code is available on Github and can be found by clicking here

EDA with SQL

- Load the SpaceX dataset into the corresponding table in a DB2 database and establish a connection with the database
- Execute SQL queries to retrieve the desired information and solve tasks, e.g.:
  - Display the names of the unique sites in the space mission
  - Display average payload mass carried by booster version F9 v1.1
  - List the names of boosters which have success on drone ship and have payload mass greater than 4000 and less than 6000
  - List the names of booster versions which have carried the maximum payload mass
  - Rank the count of different landing outcomes between 2010 and 2017
- Click here to view to the code

Build an Interactive Map with Folium

- Mark all launch sites on map by creating a folium map object and folium.Circle to add a highlighted circle around each site and a popup label using their coordinates
- Add a Marker object to show the name of each site within their respective circle
- Mark the success/failed launches for each site on the map with colors, green indicating successful launch outcome and red indicating failed outcome
- Create MarkerCluster to group markers of the same coordinates
- Add a MousePosition object for finding coordinates of points of interest
- Calculate the distances between a launch site and its proximities, i.e. coastline, railway, highway, city and add a Marker object to show the distance along with a Polyline object to draw this distance as a line
- Code is available on Github right here but as interactive maps are not shown on Github you can copy the notebook link and view it on nbviewer

Build a Dashboard with Plotly Dash

- Interactive dashboard application that allows users to perform interactive visual analytics on SpaceX launch data in real-time and comprises the following:
  - Launch Site dropdown list as input component that allows users to select a Launch Site
  - The above interacts with a pie chart that depending on the user selection shows the total success launches by site or the total success launches rate for a specific site
  - Range Slider as input component that allows users to select the Payload
  - A scatter plot that interacts with the range slider and shows the correlation between the Payload and success for sites
- Callback functions are used to render the charts based on the user input selections
- Click here for viewing the code

Predictive Analysis (Classification)

- The main objective of this part is to determine if the first stage of the rocket will land successfully
- This will be done using Machine Learning Classification algorithms such as k Nearest Neighbor, Support Vector Machine, Decision Trees and Logistic Regression
- Standardize and split the data into training and testing sets
- Create a GridSearchCV object for each model, fit the training data, find the best hyperparameters and essentially find the method that performs best
- Calculate the accuracy on the testing data
- Make a prediction using the model built and create a confusion matrix to compare between the different classes
- Click here for viewing the code

Results

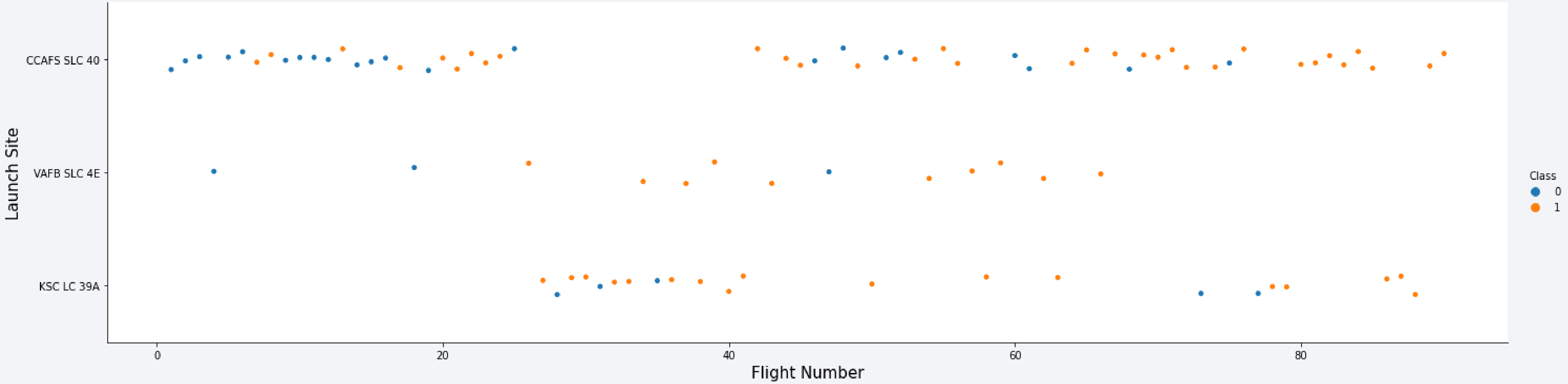
- In the next sections the results obtained are shown and analysed
- Exploratory data analysis results include scatter plots, line plot and bar plot showing relationships among features
- Information acquired using SQL queries
- Interactive analytics with Folium Maps
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

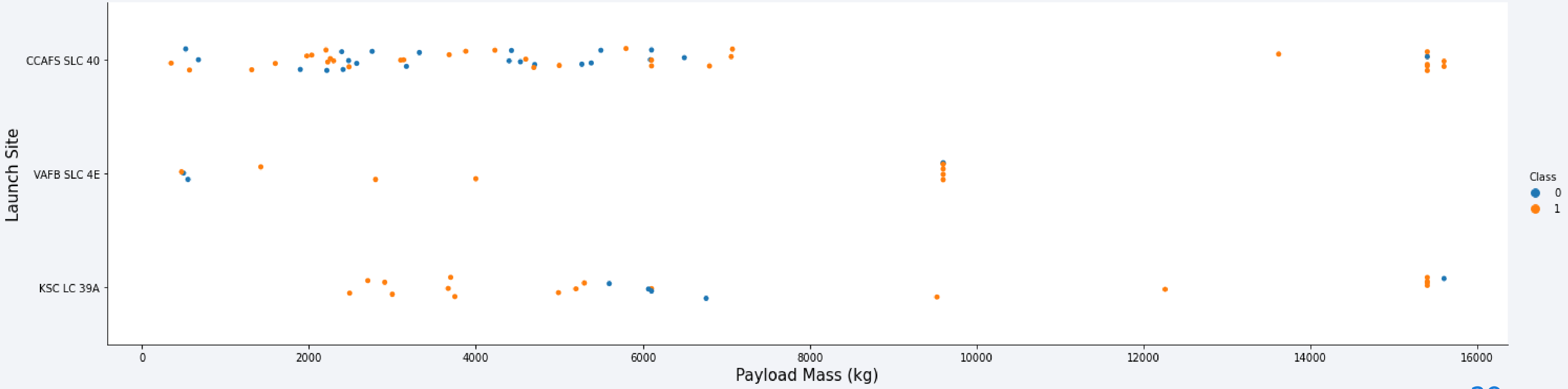
- As a general insight from the scatter plot below, it can be seen that as flight number increases the success rate increases
- For the CCAFS SLC 40 launch site, results are quite mixed and cannot give a clear answer, although from flight number 62 and above success rate seems to increase



Note: Class 0 denotes failed launch outcome and Class 1 denotes successful launch outcome

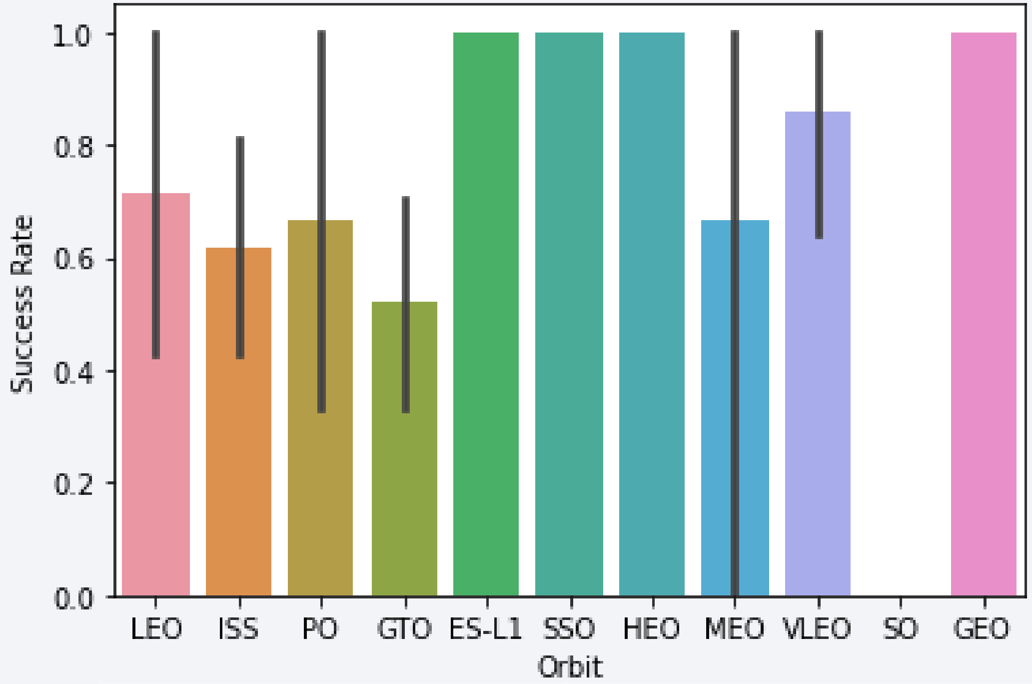
Payload vs. Launch Site

- Results are quite mixed for this approach, especially for the CCAFS SLC 40 launch site. It seems to have better success rate at higher payloads
- VAFB SLC 4E results are quite close in both low and high payloads
- For the KSC LC 39A we see that the less the payload the more likely the first stage is to return



Success Rate vs. Orbit Type

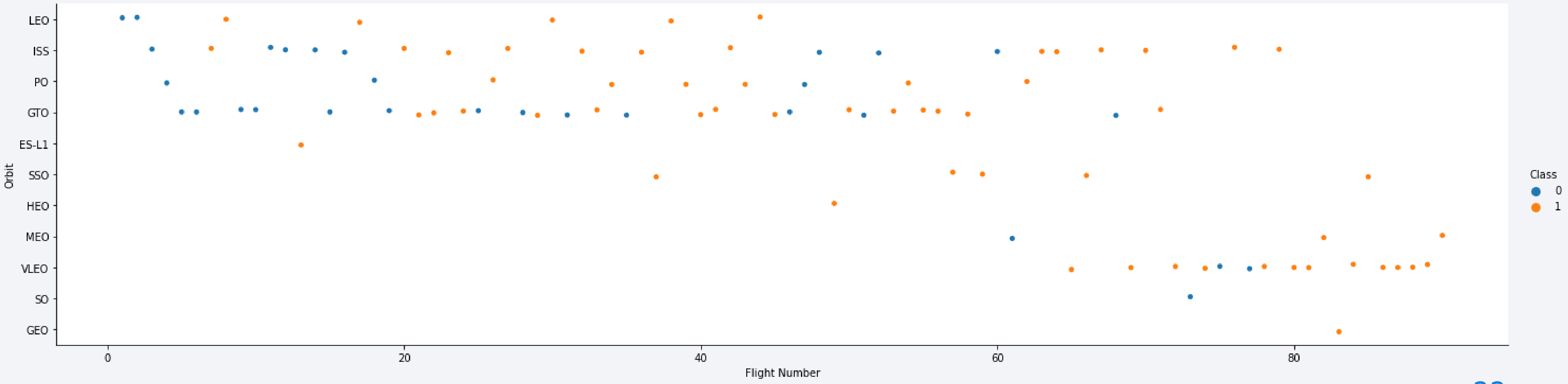
- For orbits that are very far away from the Earth's surface, the success rate is around 83%
- Low Earth Orbits (≤2000km) the success rate is around 77%
- Very Low Earth Orbits (≤450km) the success rate is around 73%
- The above results show that the farther the orbit is from the Earth's surface, the better the success rate is. However, the records we have available concerning these orbits are insignificant and therefore we cannot give a clear answer



Altitude (km)	Orbit
≤ 450	VLEO, ISS
≤ 1,000	PO, SSO
≤ 2,000	LEO
≤ 35,786	GTO, MEO, HEO, GEO
> 1,000,000	ES-L1

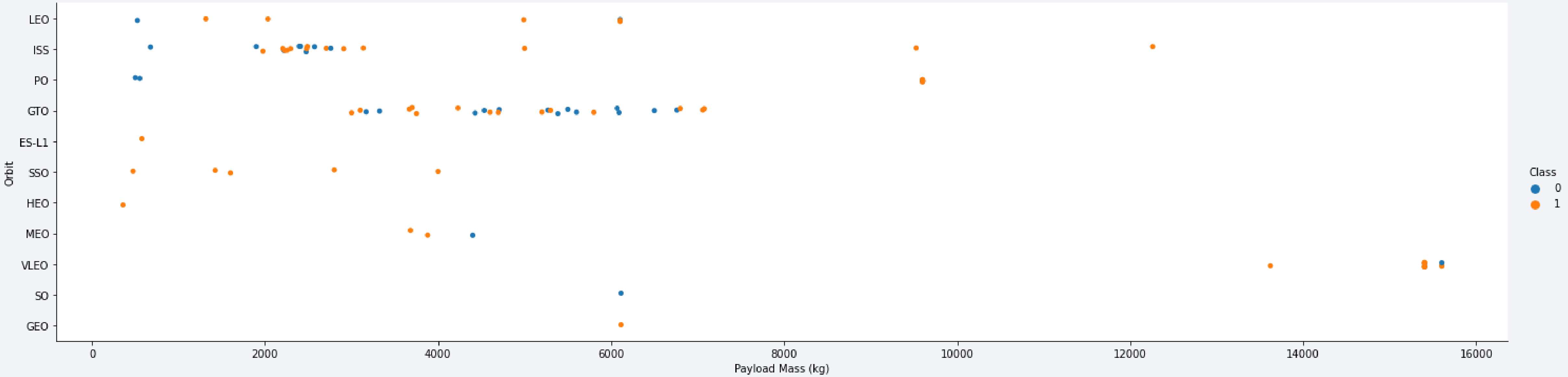
Flight Number vs. Orbit Type

- As can be seen from the scatter plot below, for some orbits the success rate is related to the number of flights, e.g.: LEO, VLEO
- On the other hand in GTO orbit there seems to be no relationship
- Orbits such as ISS and PO can partly show a connection but not to the extent of satisfaction



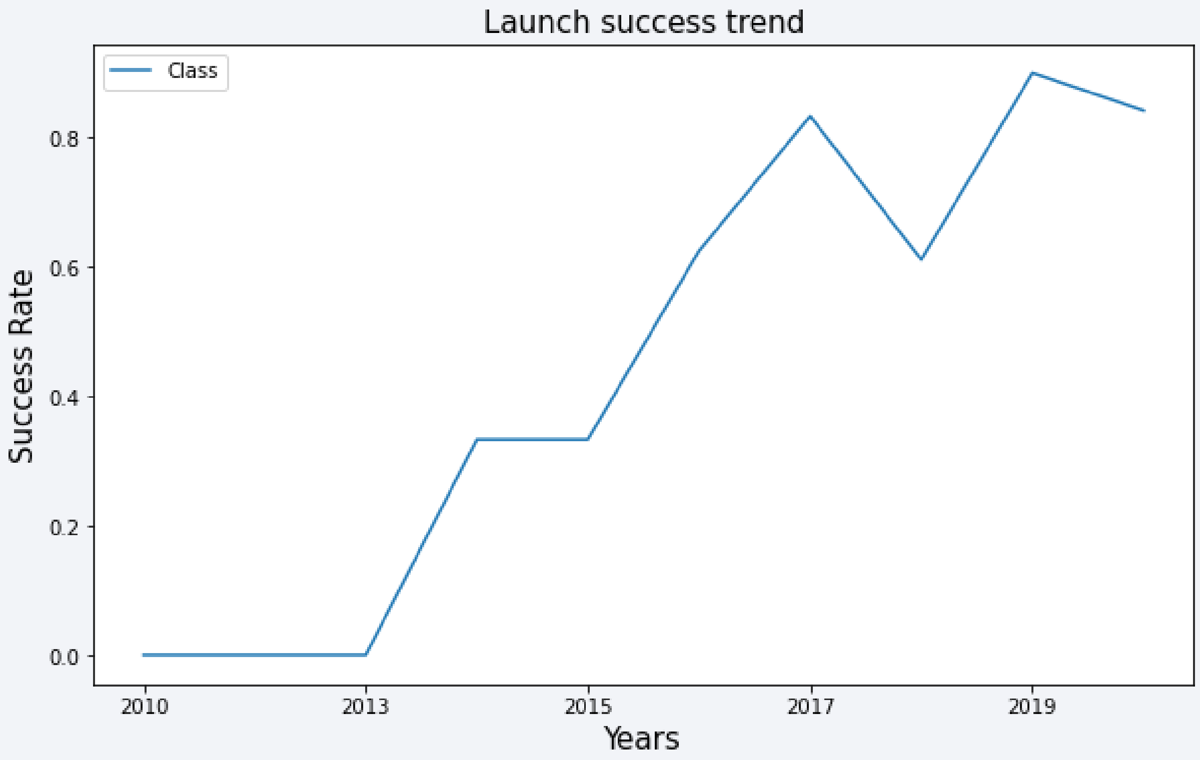
Payload vs. Orbit Type

- We observe that heavy payloads have a positive influence on orbits such as LEO, ISS, PO and negative influence on GTO
- For the rest of orbits there doesn't seem to be a very strong relationship between these two features



Launch Success Yearly Trend

- This line chart shows a clear upward trend in success rate from 2013 and onwards
- In 2017 and 2019 the success rate reached its peak with values of >80%
- However, it appears that there was a decline in both cases right after success rate reached its peak values and that constitutes the exception in the upward trend
- Specifically, 2017-2018 recorded decline of ~20% and 2019-2020 decline of ~5% after retrieving the losses in 2019



All Launch Site Names

- The following table shows the names of all Launch Sites as retrieved from database after using the SQL query:

  %sql select DISTINCT(LAUNCH_SITE) from SPACEX

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

Launch Site Names Begin with 'CCA'

- The following table shows 5 records where the Launch Site name starts with 'CCA' along with their respective feature records. Retrieved from database using the query:

```
%sql select * from SPACEX 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

- Calculate the total payload carried by boosters from NASA (CRS)
- The following SQL query was used and gave the result shown in the table

```
%sql select SUM(PAYLOAD_MASS__KG_) as "total kg" from SPACEX where CUSTOMER = 'NASA (CRS)'
```

total kg
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,928kg as shown
- The following SQL query was used for calculating the above

```
%sql select AVG(PAYLOAD_MASS__KG_) as "average kg" from SPACEX where BOOSTER_VERSION = 'F9 v1.1'
```

average kg
2928

First Successful Ground Landing Date

- Find the date of the first successful landing outcome on ground pad
- By using the following SQL query the date of the first successful ground landing is 2015-12-22
- Here the MIN function is used to retrieve the first record

```
%sql select MIN(DATE) from SPACEX where LANDING__OUTCOME = 'Success (ground pad)'
```

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- The SQL query for the above statement is shown below along with the result which shows 4 versions of boosters

```
%sql select BOOSTER_VERSION from SPACEX where LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6
000
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- The following SQL query was used for retrieving the total number for each case
- 99% of total mission outcomes is successful

```
%sql select DISTINCT(MISSION_OUTCOME), count(*) as "total" from SPACEX group by MISSION_OUTCOME
```

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- The table shows that the maximum payload mass for all booster versions is the same and is 15,600kg
- The following SQL query was used for retrieving the above information

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

```
%sql select DISTINCT(BOOSTER_VERSION), PAYLOAD_MASS__KG_ from SPACEX where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) fr
om SPACEX)
```

2015 Launch Records

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015
- As can be seen there are 2 booster versions, both from the same launch site
- The following SQL query was used to retrieve the above

```
%sql select BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME from SPACEX where LANDING__OUTCOME = 'Failure (drone ship)' and YEAR
(DATE) = '2015'
```

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- As observed most attempts for landing were made at the ocean, either on a drone ship or controlled
- Out of total outcomes, 36% was successful, 29% was failure and for 35% no attempt was made
- The table shown was obtained by using the SQL query below

landing__outcome	total
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

```
%sql select LANDING__OUTCOME, count(LANDING__OUTCOME) as "total" from SPACEX where DATE BETWEEN '2010-06-04' and '2017-03-20' gr
oup by LANDING__OUTCOME order by count(LANDING__OUTCOME) DESC
```

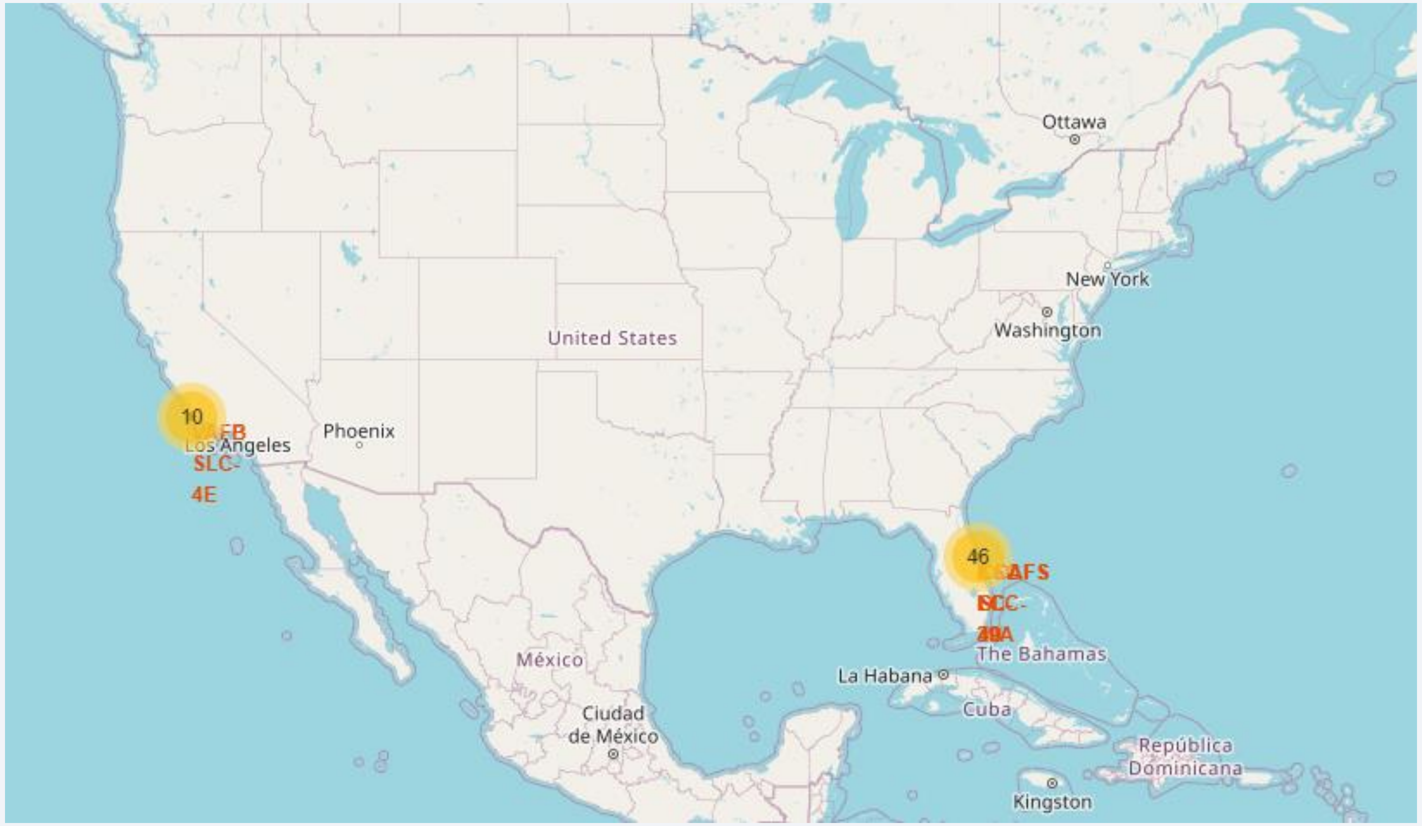
Note: Controlled (ocean) means controlled atmospheric entry and vertical splashdown on ocean's surface at near zero velocity. Boosters were destroyed at sea.

Section 4

Launch Sites Proximities Analysis

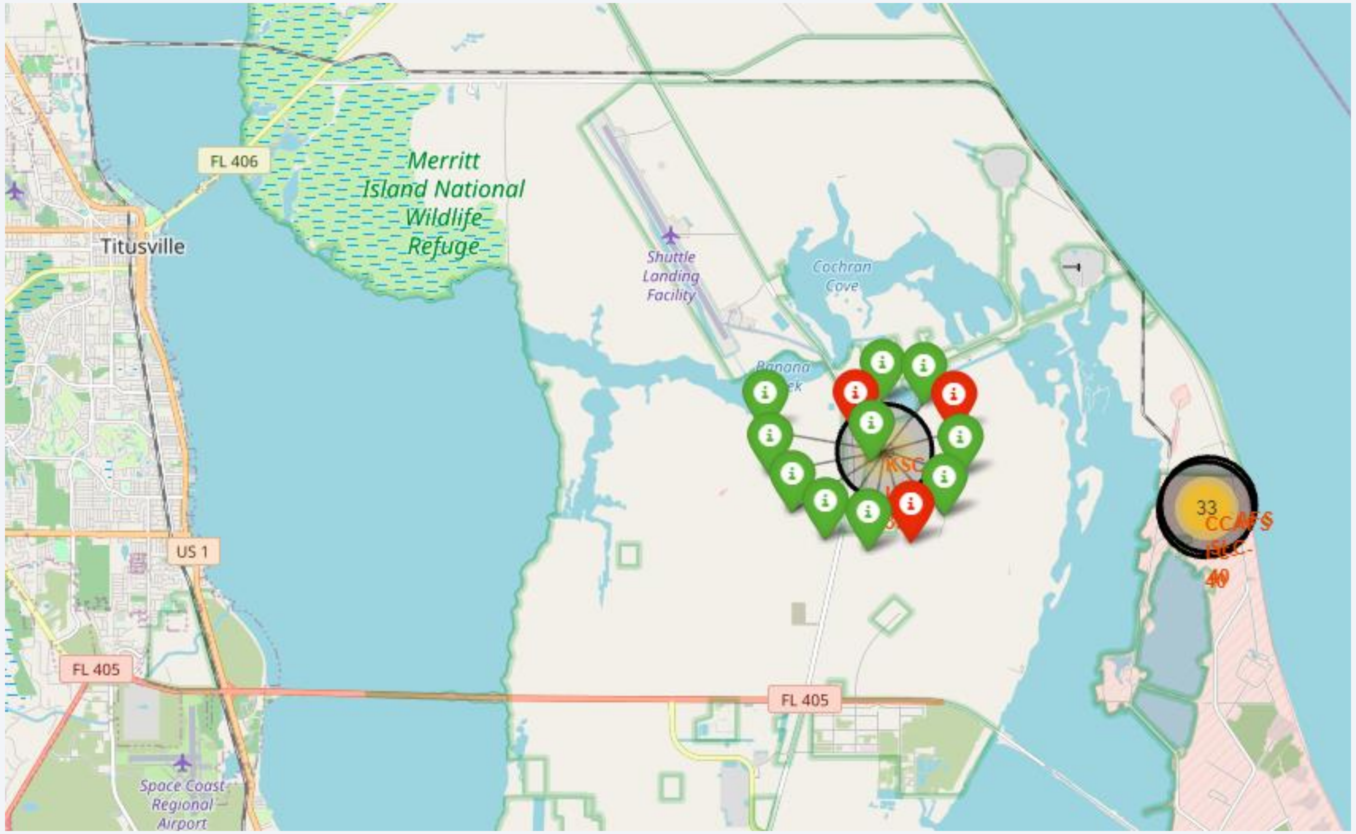
Location of Launch Sites on Folium Map

- The location of each site has been added to the map by using their coordinates and marker clusters were created for records of similar coordinates
- 3 launch sites are located in the state of Florida with total 46 mission outcome records
- 1 launch site is located in the state of California with 10 mission outcome records
- We can observe that the launch sites tend to be close to the ocean and to the equator



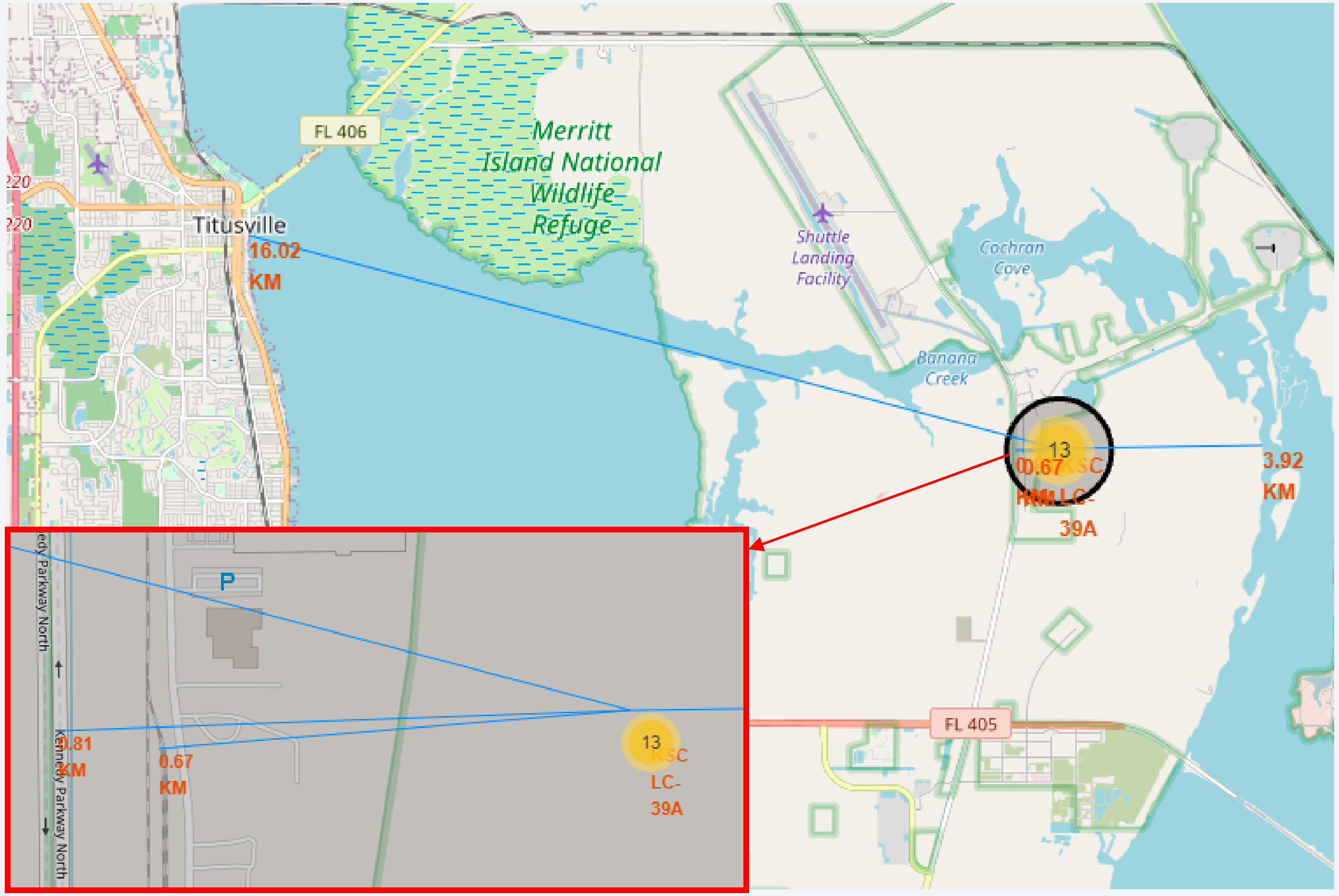
Launch Records

- Use color markers to indicate the outcome of missions, green for successful outcome and red for failed outcome
- Now it is easy to identify which launch sites have relatively high success rates
- Here we see that the KSC LC-39A launch site has a success rate of 77%



Proximities of Launch Sites

- Analyze and explore the proximities of launch sites
- Here the distances of the KSC LC-39A launch site to its proximities are shown, i.e. the distance between the launch site and the closest coastline, city, railway and highway
- We can observe that the launch sites are in close proximity to coastline, railways and highways. On the other hand a certain distance is kept away from cities

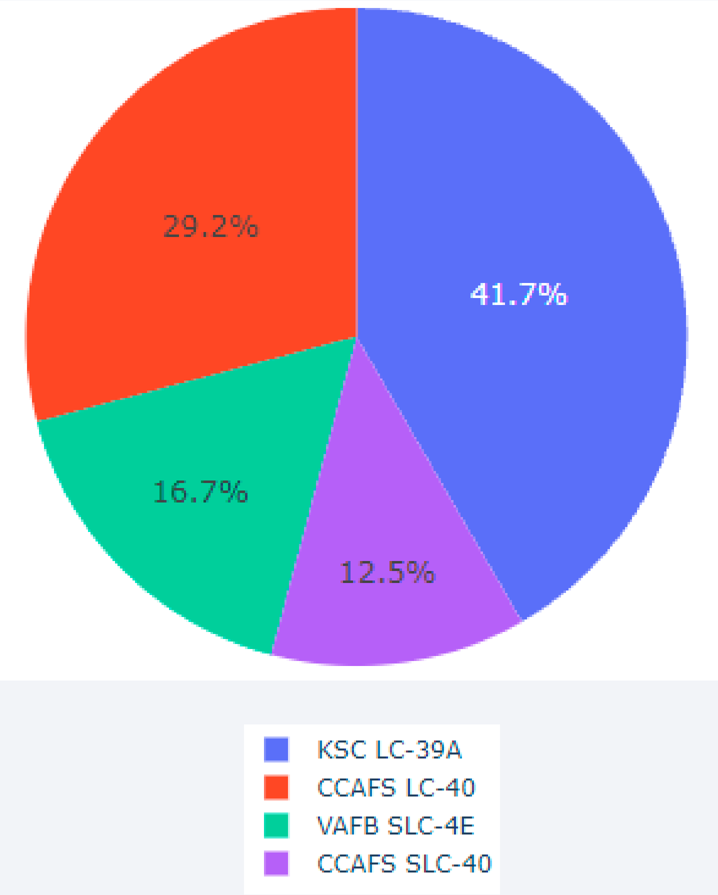


Section 5

Build a Dashboard with Plotly Dash

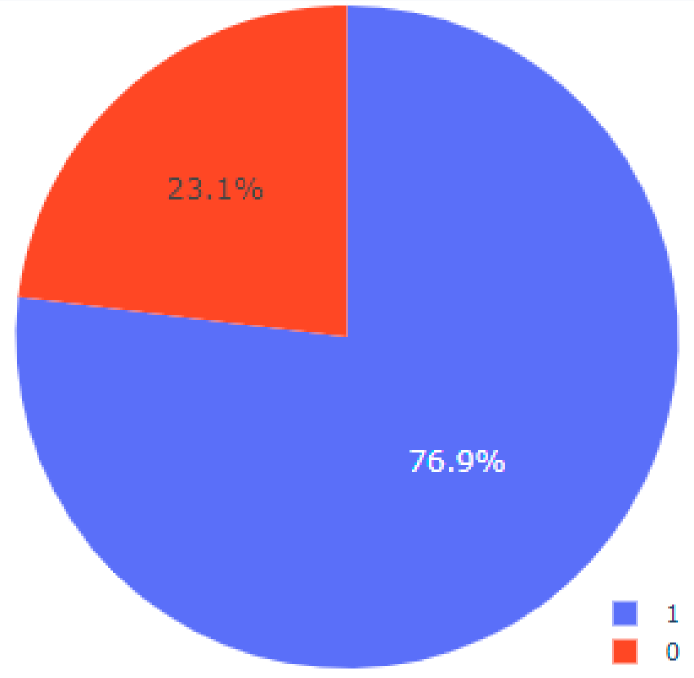
Launch Success For All Sites

- The adjacent figure shows the proportion of launch success rate of each launch site
- It is clear that the most successful launch site is KSC LC-39A and the least successful is CCAFS SLC-40
- It is worth mentioning that the most rocket launches have taken place on the CCAFS LC-40



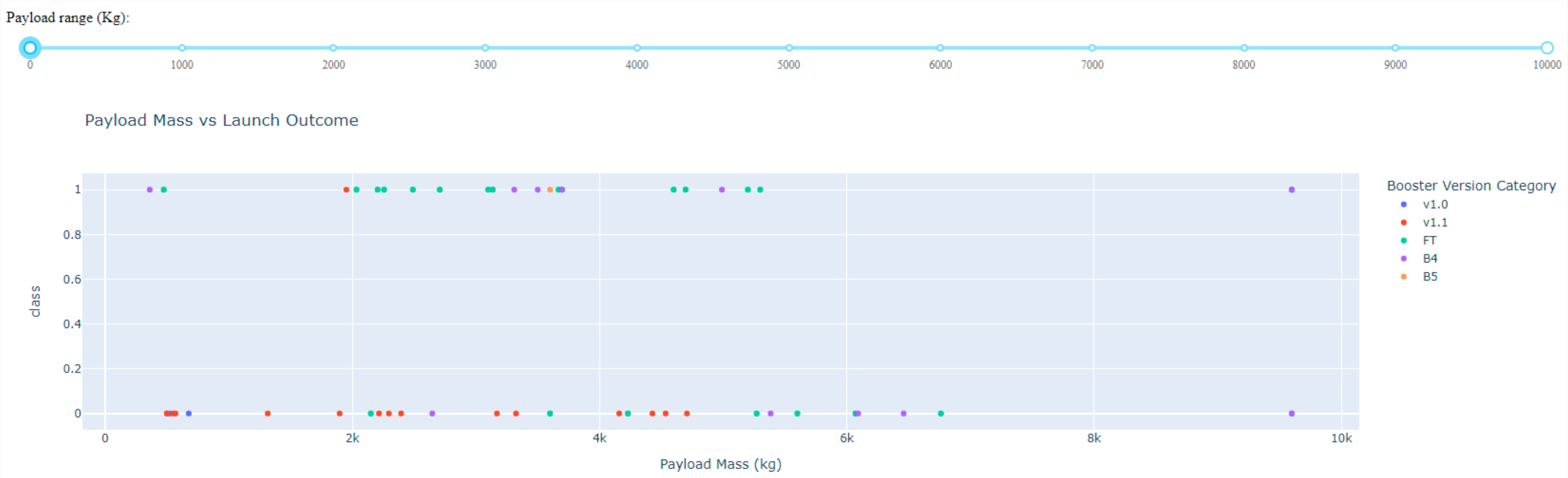
Most Successful Launch Site – KSC LC-39A

- The KSC LC-39A is the most successful launch site as mentioned before
- 10 out of the 13 launch records have had a positive outcome forming a 76.9% success rate
- 3 missions have had a negative outcome



Payload Mass vs Launch Outcome

- As observed from the scatter plot below, the most rocket launches have occurred with a payload mass of 2,000kg – 4,000kg, recording a success rate of 60%
- For payloads above 5,000kg the success rate is at its lowest at 33%
- The most successful booster version is the FT and the least successful is the v1.1

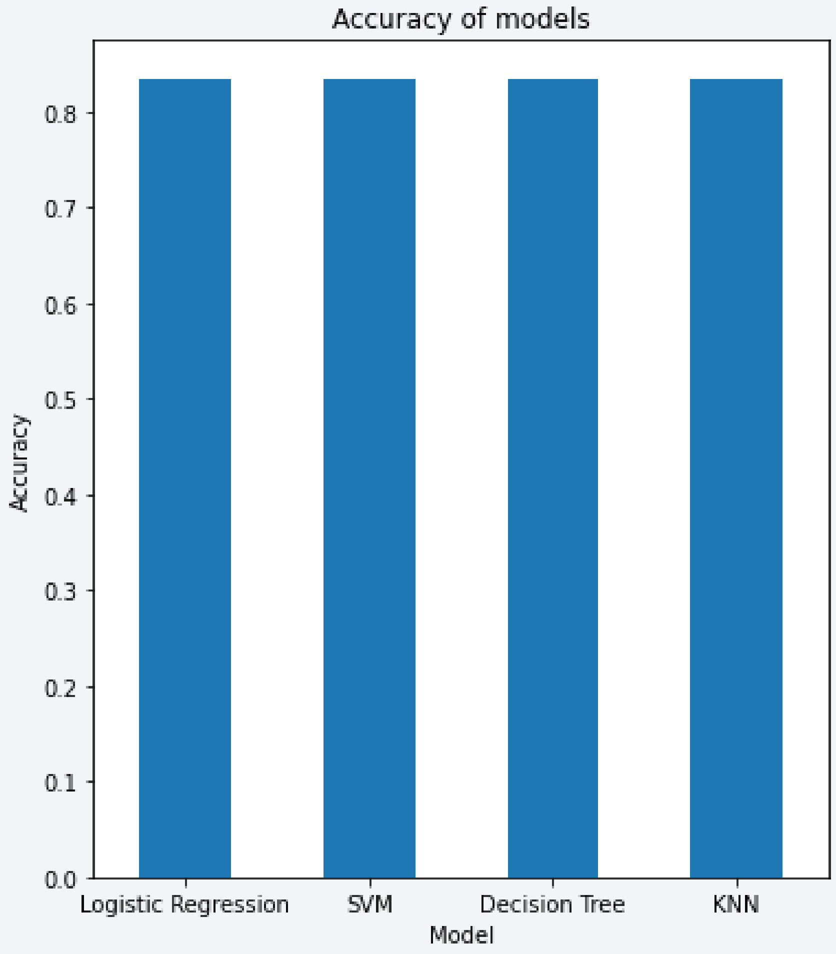


Section 6

Predictive Analysis (Classification)

Classification Accuracy

- Obtain the accuracy of each classification model on the test data
- As can be seen all models practically perform the same giving an accuracy of 83%
- The accuracy of the models might be good enough but the number of records we have available is insignificant, as we only have 18 records on our testing set after the 0.8/0.2 split on data



Confusion Matrix

- Since all models perform the same, the confusion matrix is the same
- It can be seen that the models can distinguish between the different classes and that the major issue is the False Positives
- The landed labels shows that the models have predicted by 100% correctly the actual values as False Negatives are 0

Conclusions

- The more massive the payloads are, the less likely the first stage is to return
- It appears that the safest payload range to operate is 2,000kg – 4,000kg, mission success rate at this range is 60% and any payload above 5,000kg is considered risky
- The most successful booster version is the FT, having a success rate of 80% when operating at the safest payload range and an overall success rate of 65%
- Launch sites are in close proximity to the equator and to coastline. The first stage is more likely to attempt landing on the ocean, either on a drone ship or splashing on ocean
- Concerning orbits, although it was obtained that farther orbits give better results, it is not very clear that there is a relationship as the records we have for these kind of orbits are very few compared to the Low Earth Orbit records
- As stated before, the prediction models perform quite well but the number of records we have available at this moment are not enough to give a safe answer about whether the first stage of the rocket will safely return

Appendix

- Project Github link: Project
- Interactive Dashboard: Dashboard
- SQL Queries: SQL notebook
- Datasets: the datasets obtained can be found right here

Thank you!