IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Roman Oppermann
08.03.2022

Outline

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

Executive Summary

- Goal of the Project: Estimate if Falcon 9 stage 1 would land successfully or not
  - This fact plays major role in estimate the price of its launch
  - Data for the estimation was sourced from SpaceX REST API and the Falcon 9 Wikipedia Page
- Process to achieve the Goal after collecting the data:
  - Perform some data wrangling to determine the best predictors for our outcome
  - Exploratory Data Analysis and feature scaling to visualize the data and get an overview of it
  - Calculating and evaluation of four different machine learning models to find the best one for the goal of the project
- The results shows that the outcome depents on various parameters
  - Launch factors: Launch site, payload mass, target orbit
  - Technical factors: gridgins, cores

Introduction

- Space X is able to reduce the cost of launch of the Falcon 9 Rocket to 62 million dollars due to reuse the Stage 1
  - Other competitors need 165 million dollars for a rocket launch
  - Space X is by this fast in the moment the leader of the space race
- The goal of this project is to build up a machine learning pipeline to predict, if the first stage will land successfully
  - To Predict, if the first stage will successfully, all factors and their impacts need to be defined
  - All interactions between the factors need to be find
  - From Results of the machine learning pipeline can be say which job can be accomplished safely

Section 1

Methodology

Methodology

- Data collection methodology
  - Diretly from SpaceX's REST API and with web scrapping from the Falcon 9 Wikipedia
- Perform data wrangling
  - Handling of missing values (drop or replaced) and adding column for summarize if landing was successfully as Preparation for the classification 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
  - Data got prepared to use with four Classification models (SVM, logistic regression, tree classifier and k-nearest neigbours) and got evaluate with several performance indicators like Score, Jaccard, F1-score and LogLoss

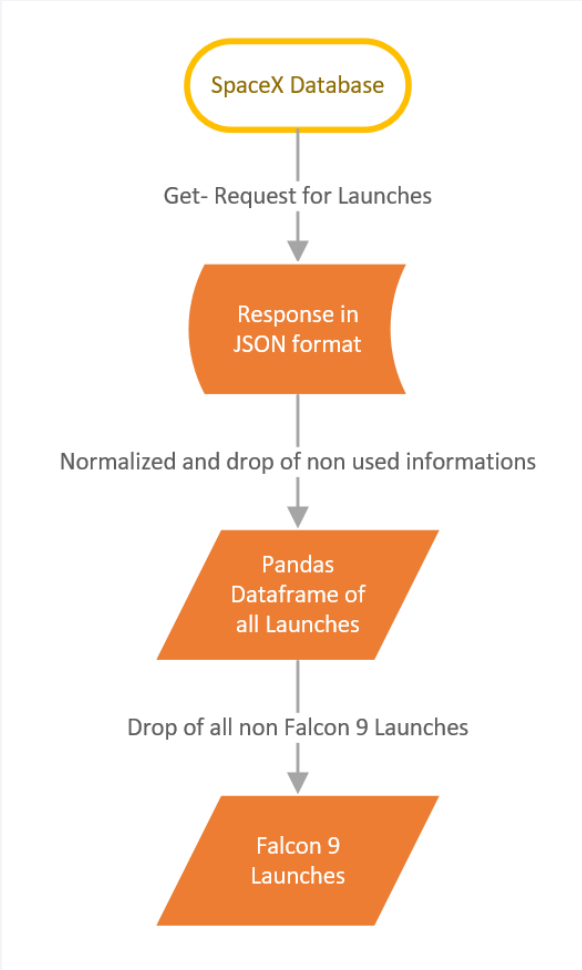
Data Collection

- The dataset was collected on two different ways
  - SpaceX Rest API: All Data about rockets, launchpad and payload got requested and then reduced to all Information to Falcon 9 rockets
  - Webscrapping of Falcon 9 Wikipedia Page: Collected all Data of Falcon 9 Wikipedia which stand there in Tables

Data Collection – SpaceX API

- The left side presented flowchart show the data collecting process from the Space X REST API
- The Data get requested from the API, normalize and converted to a Pandas dataframe

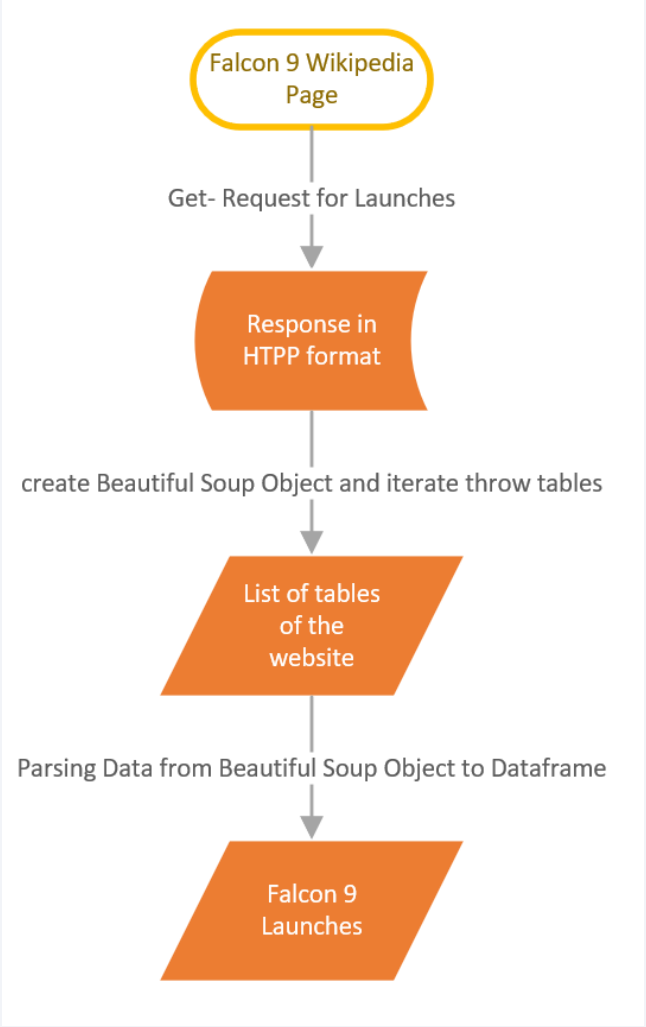
- Jupyter Notebook on GitHub



Data Collection - Scraping

- Using webscrapping to get informations from the launch tables of the Falcon 9 Wikipedia Pages

- [Jupyter Notebook on GitHub]



Data Wrangling

- The collected data needed to transferred in the right format, p. a. the date format or missing information need to get replace by the mean value of that column, p. a. the payload of the rockets.
- Also new columns need to get create to get a new basis to achive the goal of the goal of the work. So it needed to create a column which contains the information, if a landing was successful displayed by 0 and 1 in Preparation for the machine learning models
- Link: Replacement missing information by the mean of the column
- Link: Creating column of successful landing in Preparation for machine learning

EDA with Data Visualization

- Used Scatter point charts to see, if there is an relationship between Launch Site and Flightnumber or Payload Mass and the Outcome of Landing → Scatter point chart allow to visualize three information an the same time
- Used a bar chart to visualize the outcome of the landing and the orbit where the payload got transport to
- Used a line chart to visualize the development of successful outcomes over years. The line chart is easy to interpret the information to see a trend
- All charts and the codes are here

EDA with SQL

- Following SQL request got performed to get deeper in the Data:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - 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
- All results are to find there

Build an Interactive Map with Folium

- Added markers and line
  - Markers: Launch Sites, Launches with the outcome of the landing, distances to the shoreline, the next highway, city, railways
  - Line: show the distance to the shoreline, the next highway, city, railways
- The object get display to show location of the launch site, their impact of the landing outcome and there location to the normal infrastructure to show that these are special places on special zone on the earth
- The Map and the code are to find here

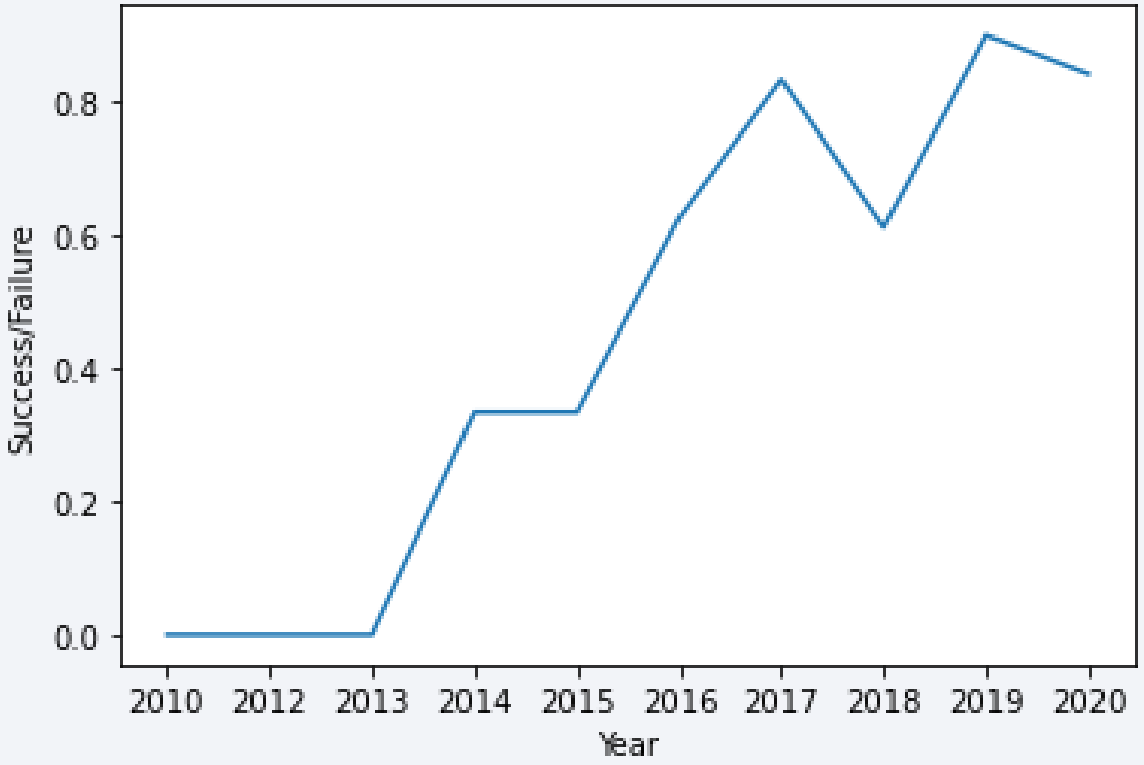
Build a Dashboard with Plotly Dash

- Following plots/graphs and interactions got added:
  - Drop-down menu to select the launch site that the user can separate the launch sites
  - Range-slider to choose the range of payload that the user can select the payload range what he needs and give him with the interactive pi and the scatter plot more depth according to his needs
  - Interactive pie chart showing the success rate of all launch sites
  - Scatter plot showing launch outcomes of all sites according to their payload
- Python File is to find here

Predictive Analysis (Classification)

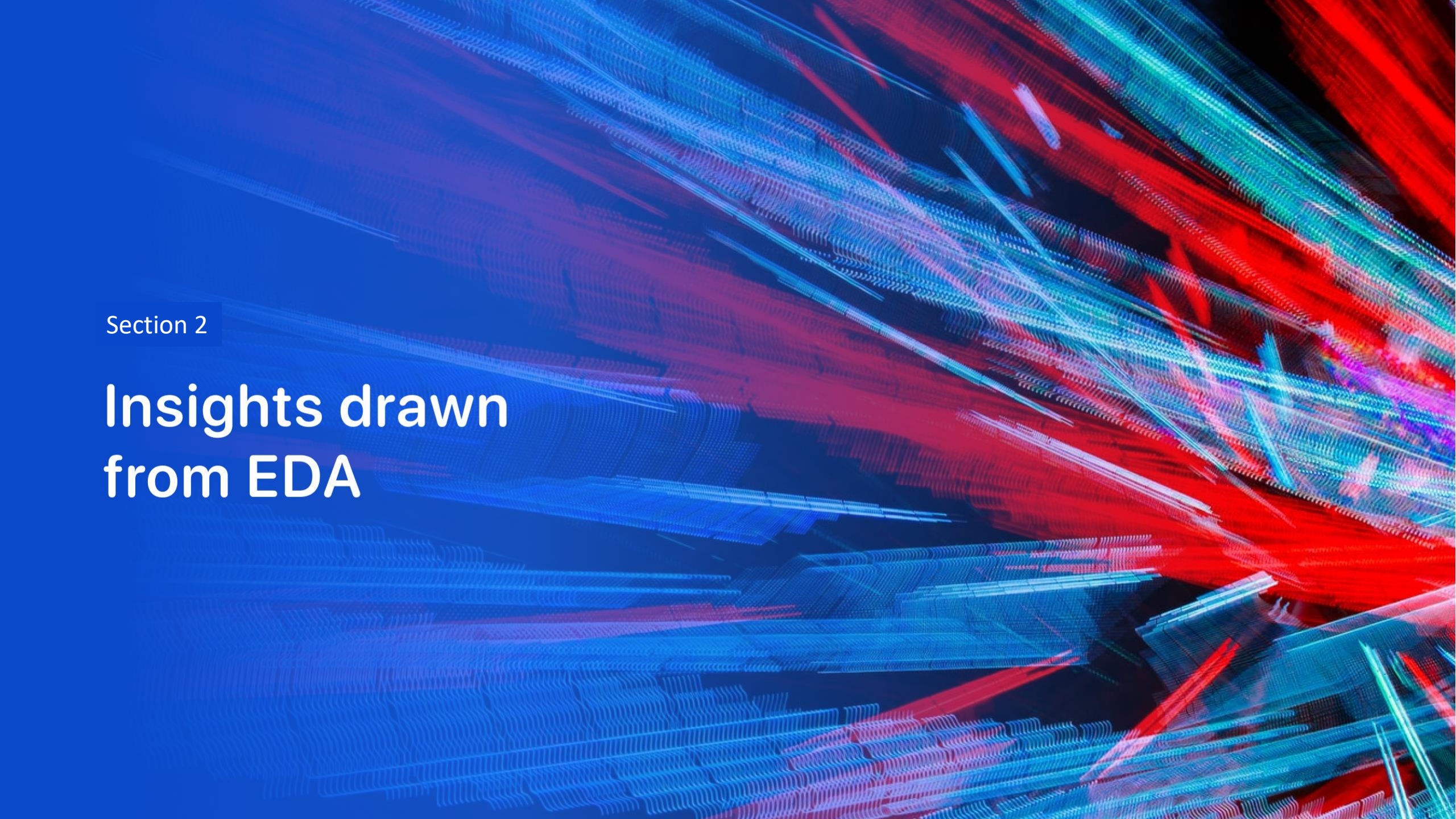
- The dataset got split in a train and test set, with the test set the machine learning models got created and evaluated with the test set
- Following ML models were used: logistic regression, SVM, decision tree and KNN
- For the evaluation following methos were used: Score, Jaccard, F1 and LogLoss (which only can use for logistic regression)
- The best ML models were logistic regression, SVM and KNN with a Score of 0.83
- All Results and the Code is here

Results

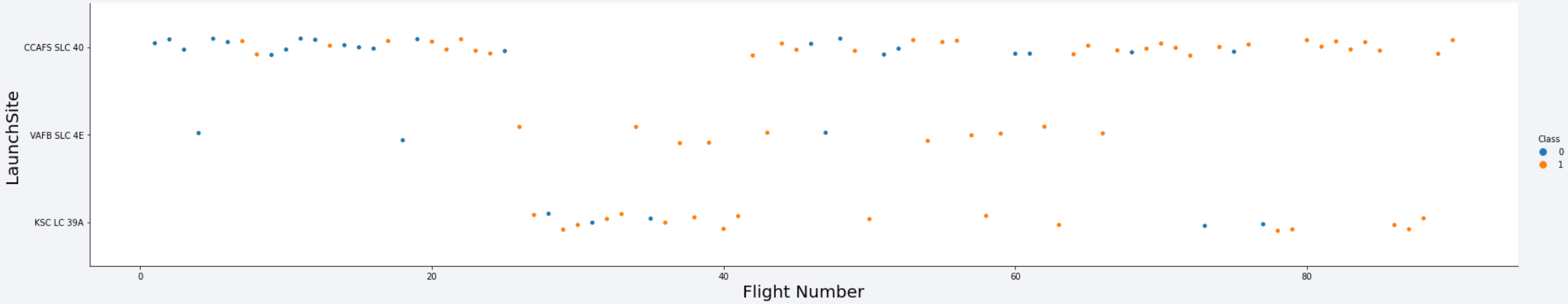


Success / Failure over the Years

Confusion Matrix of KNN ML model

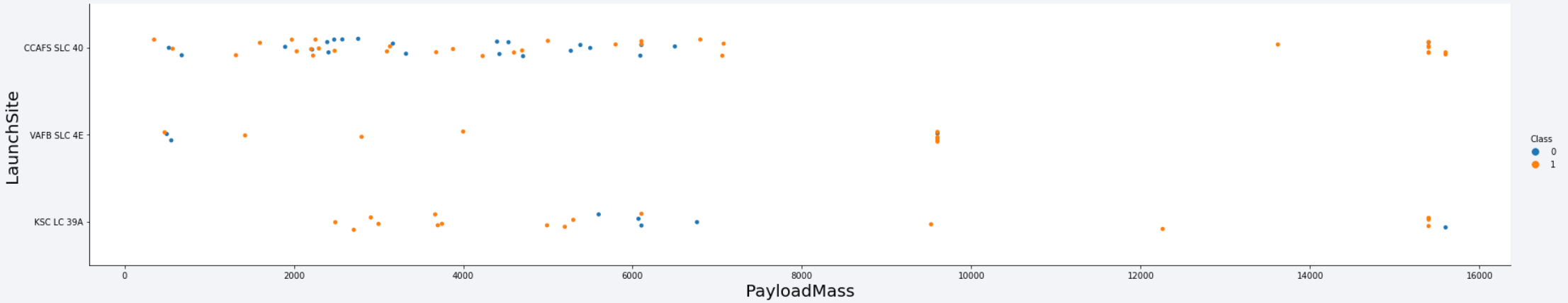

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAFS SLC 40 most used Launch Site and got used nearly the whole Time
- VAFB SLC 4E just used rarely but have a good success/ failure rate
- KSC LC 39A got intruced later and is also not that often in use
- With increasing number of starts the success rate is rising

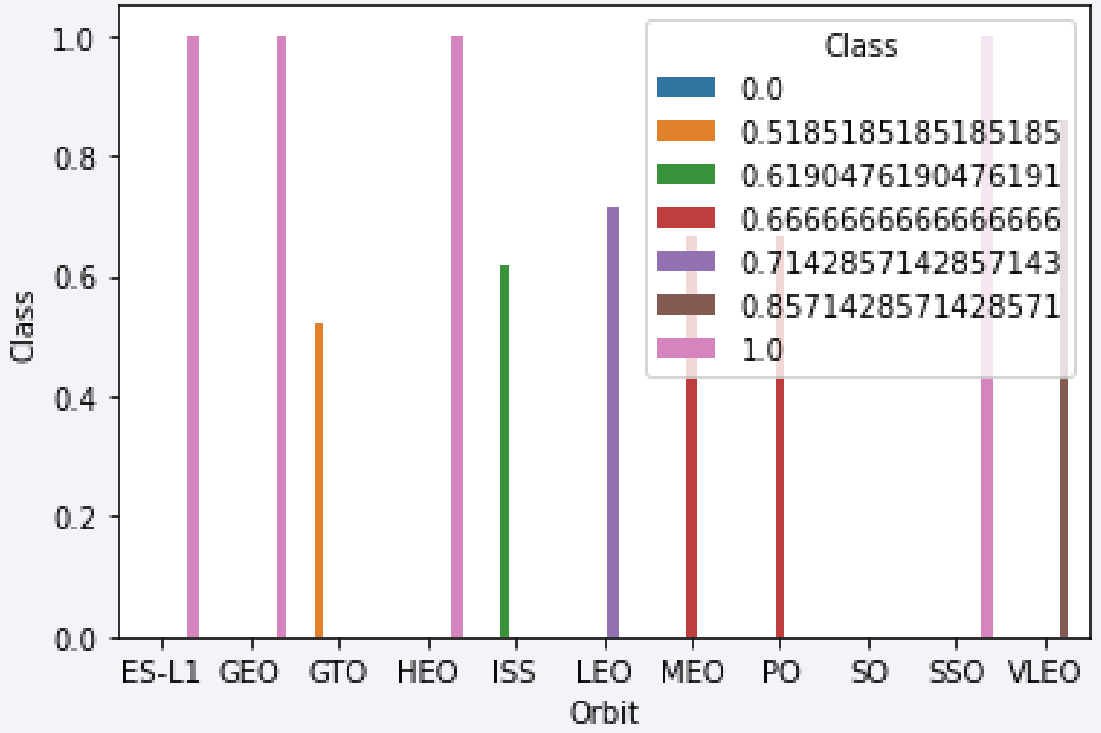
Payload vs. Launch Site



- CCAFS SLC 40 and KSC LC 39A are used for Launches with the highest payload mass
- The maximum payload mass of VAFB SLC 4E is under 10 000
- With higher payload mass the success rate increase, which also indicates that higher payload mass came with increasing flight numbers

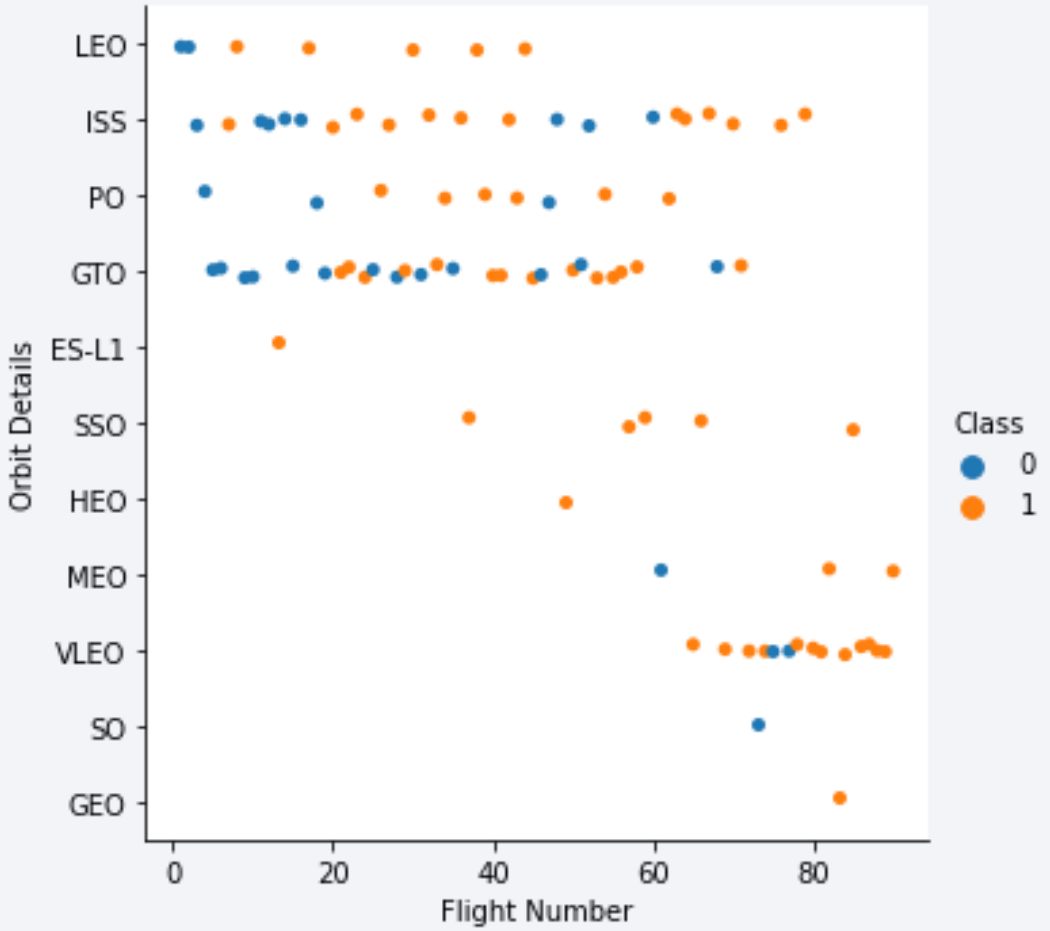
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have a high chance of success



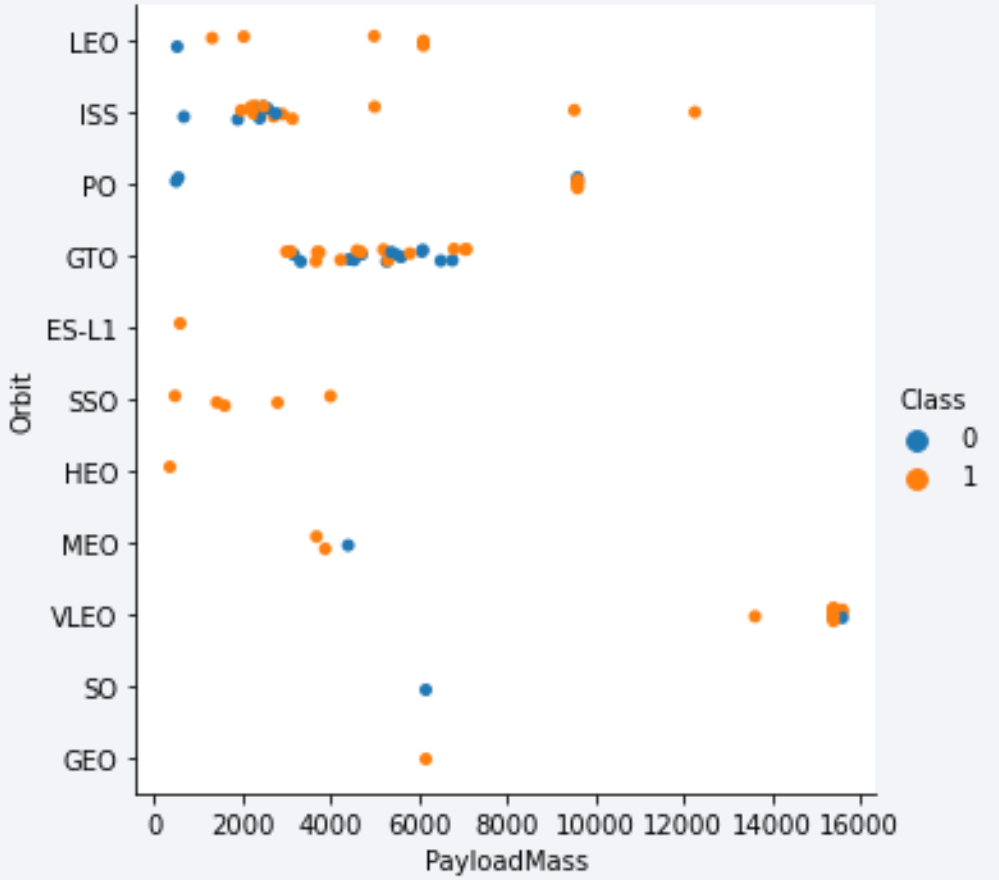
Flight Number vs. Orbit Type

- The most flights were to GTO, ISS and VLEO
- ES-L1, SO and GEO got achived only once each
- ES-L1, SSO, HEO, GEO have a 100% success rate



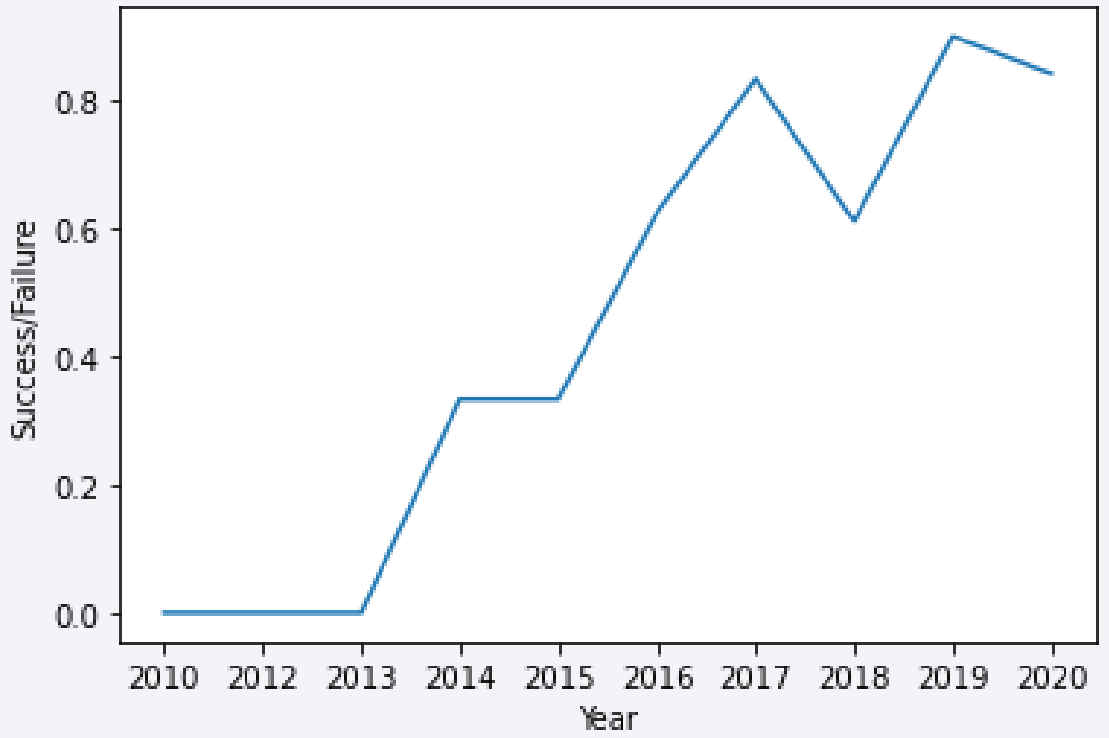
Payload vs. Orbit Type

- Highest payload mass got send to VLEO with good success rate
- Most flights to the ISS had a payload mass between 2000 and 4000
- Most flights to FO had a payload mass between 2000 and 8000



Launch Success Yearly Trend

- In general: By Time the succes rate increase
- Between 2014 and 2015 it stays equal
- There is a big Drop in 2018 and a small in 2020



All Launch Site Names

- The names of the all launch sites are:

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

- used DISTINCT parameter to get unique results of the column launch_site from the Table SPACEXTBL:

```
1 %sql SELECT DISTINCT launch_site from SPACEXTBL
```

Launch Site Names Begin with 'CCA'

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

- Table showing the first 5 starts from a launch_site that starts with “CCA”

Total Payload Mass

- The Sum of payload Mess carried by boosters launched by NASA (CRS):

1
45596

- The Query to get this result look like this:

```
%sql SELECT SUM(payload_mass__KG_) FROM SPACEXTBL WHERE customer = 'NASA (CRS)'
```

Average Payload Mass by F9 v1.1

- The average payload mass carriad by booster version F9 v1.1 is 2928:

1
2928

- The Query to get this result look like this:

```
%sql SELECT AVG(payload_mass__KG_) FROM SPACEXTBL WHERE booster_version LIKE 'F9 v1.1'
```

First Successful Ground Landing Date

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

1
2015-12-22

- The Query to get this result look like this:

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE landing__outcome = 'Success (ground pad)'
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of booster names, which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

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

- The Query to get this result look like this:

```
%sql SELECT booster_version FROM SPACEXTBL WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ >4000 AND payload_mass__kg_ < 6000
```

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes:

1
101

- The Query to get this result look like this:

```
%sql SELECT COUNT(mission_outcome) FROM SPACEXTBL
```

Boosters Carried Maximum Payload

- List of the boosters, which have carried the maximum payload mass

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

- The Query to get this result look like this:

```
%sql SELECT DISTINCT a.booster_version, b.payload_mass__kg_ FROM SPACEXTBL a, SPACEXTBL b WHERE a.booster_version = b.booster_version ORDER BY b.payloa
```

2015 Launch Records

- List of the failed landing_outcomes in drone ship with the name of their booster versions, and launch site names for in year 2015:

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

- The Query to get this result look like this:

```
%sql SELECT landing__outcome, booster_version, launch_site, DATE FROM SPACEXTBL WHERE landing__outcome = 'Failure (drone ship)' AND DATE BETWEEN '2015-
```

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

- Rank of the categories 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:

landing__outcome
Uncontrolled (ocean)
Success (ground pad)
Success (drone ship)
Precluded (drone ship)
No attempt
Failure (parachute)
Failure (drone ship)
Controlled (ocean)

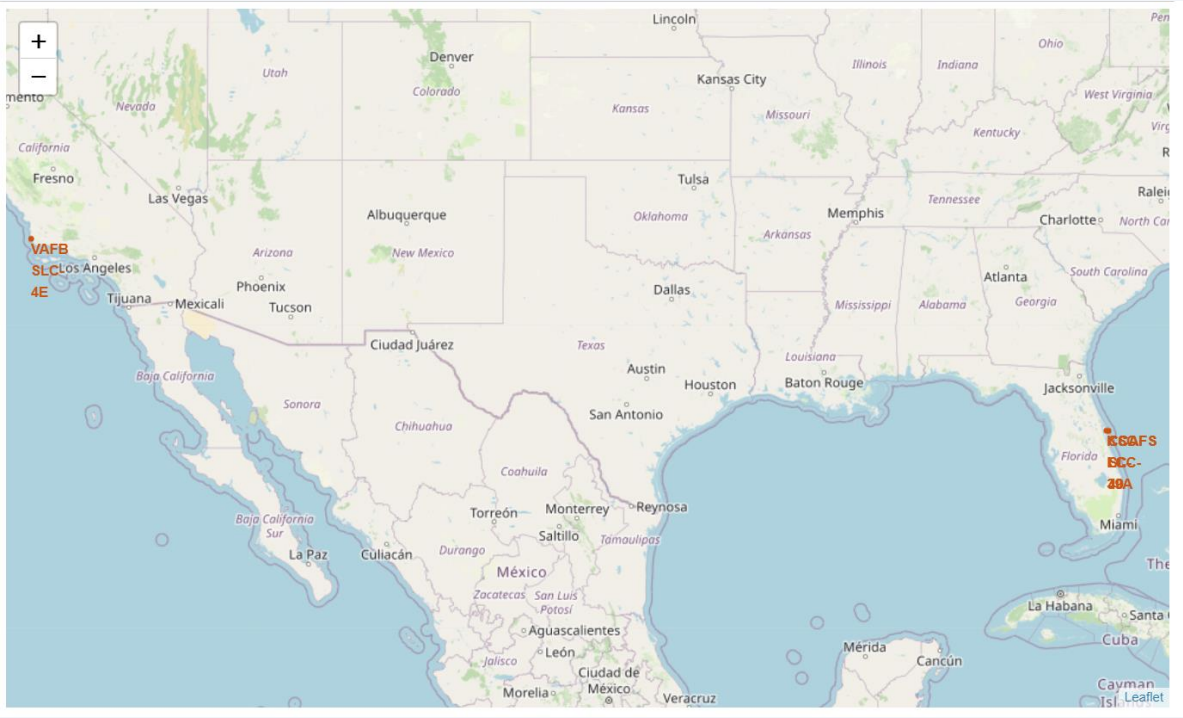
- The Query to get this result look like this:

```
1 SELECT DISTINCT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY LANDING__OUTCOME DE
```

Section 3

Launch Sites Proximities Analysis

Launch Sites of Space X

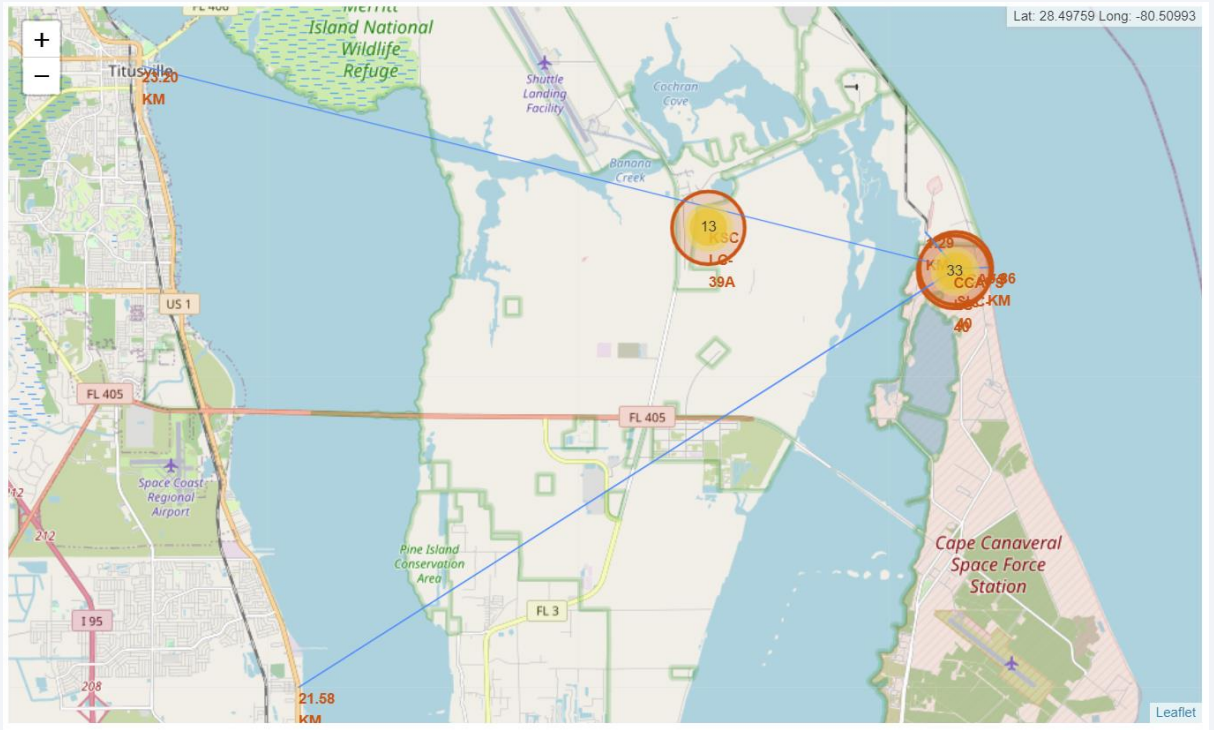


- On the map above are the launch sites displayed which Space X used to launch the Falcon 9 rockets, three are on the east coast in a sub-tropic area and on is at the west coast
- SpaceX is so more flexibily due to bigger weather events

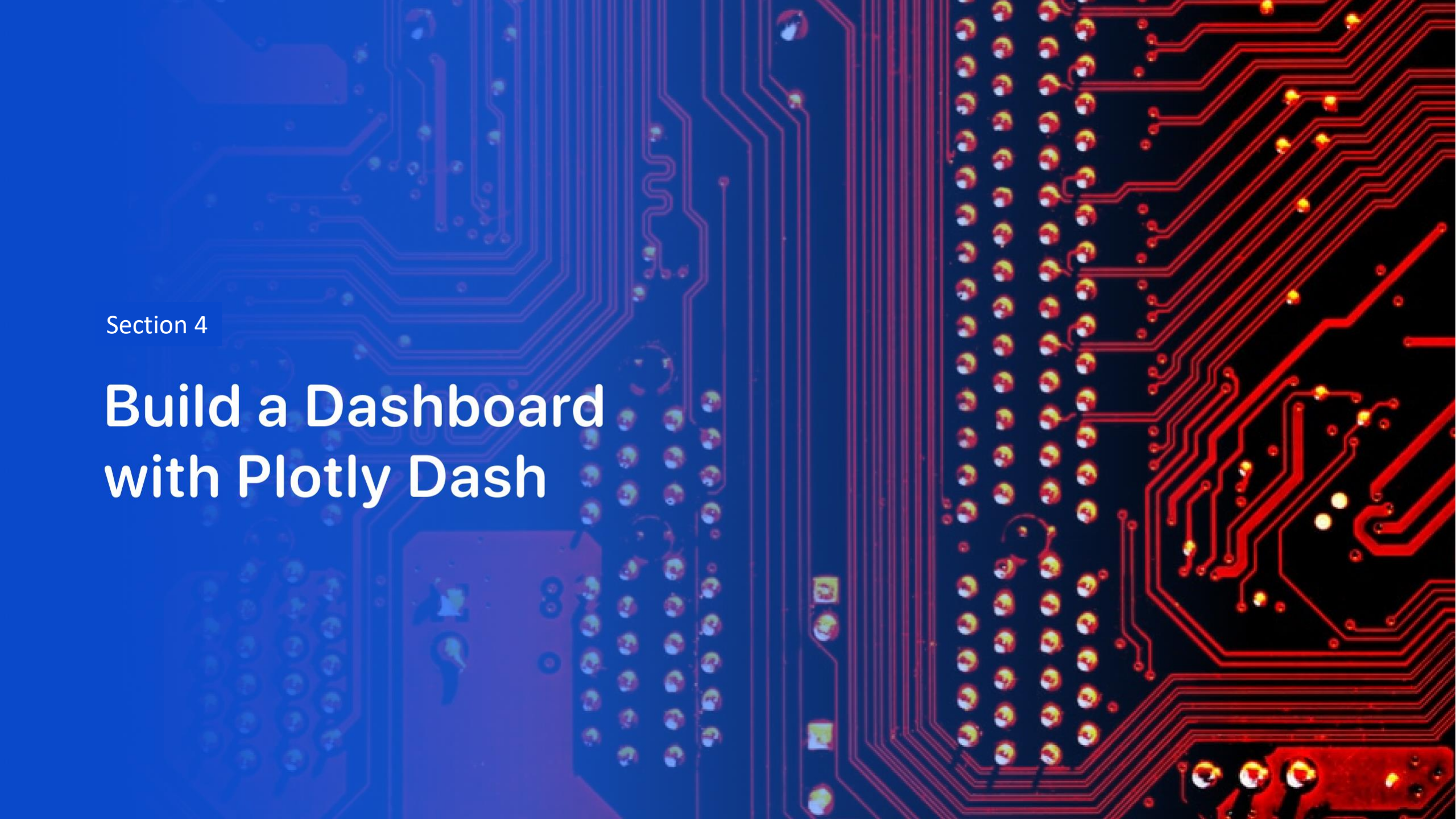
<Folium Map Screenshot 2>

- The two Screenshots show the both launch sites in Florida which are located very close next to each other
- The northern one was less used than the southern one

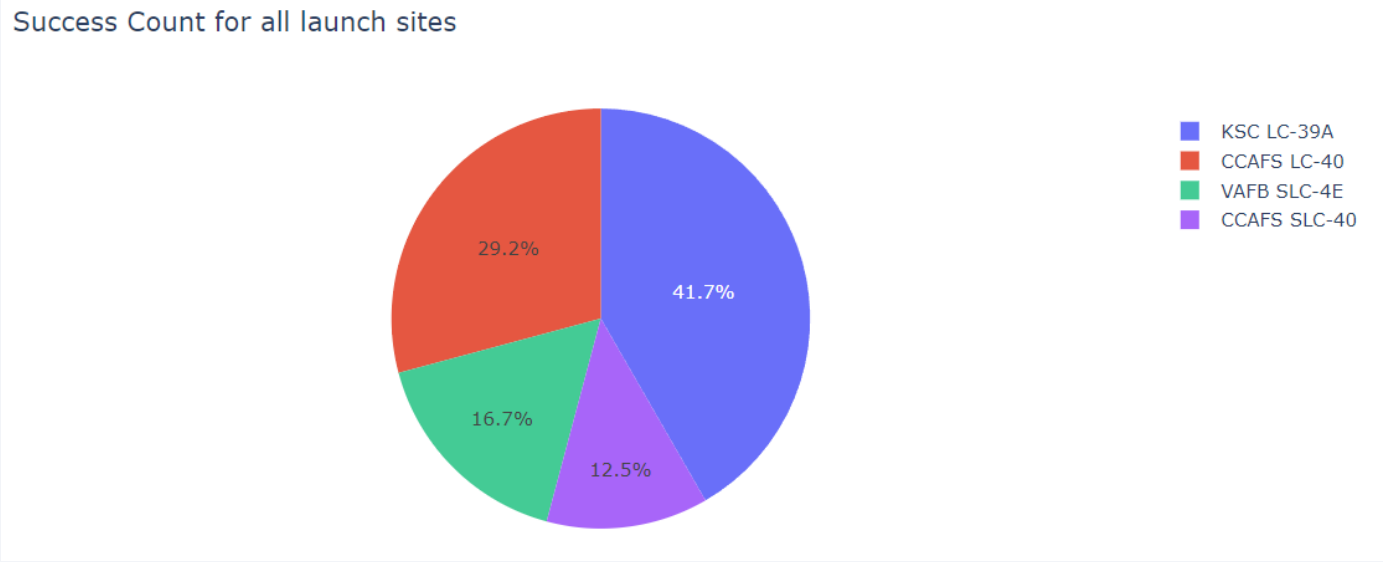
Proxmities of a launch site



- The launch site are very near to the ocean and have a greater distance to highways, railways and cities
- We bigger distance to highway, railways and cities are planned due to chance of a failure and the wish to keep uninvolved people safe

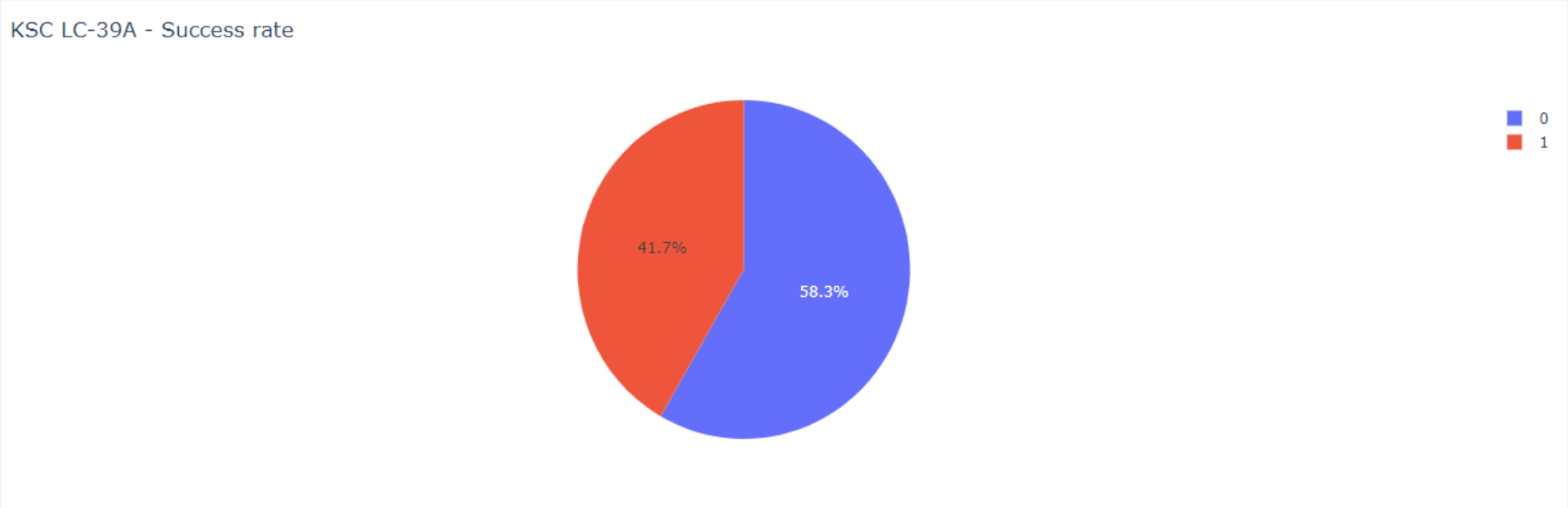

Section 4
Build a Dashboard with Plotly Dash

Launch success rate



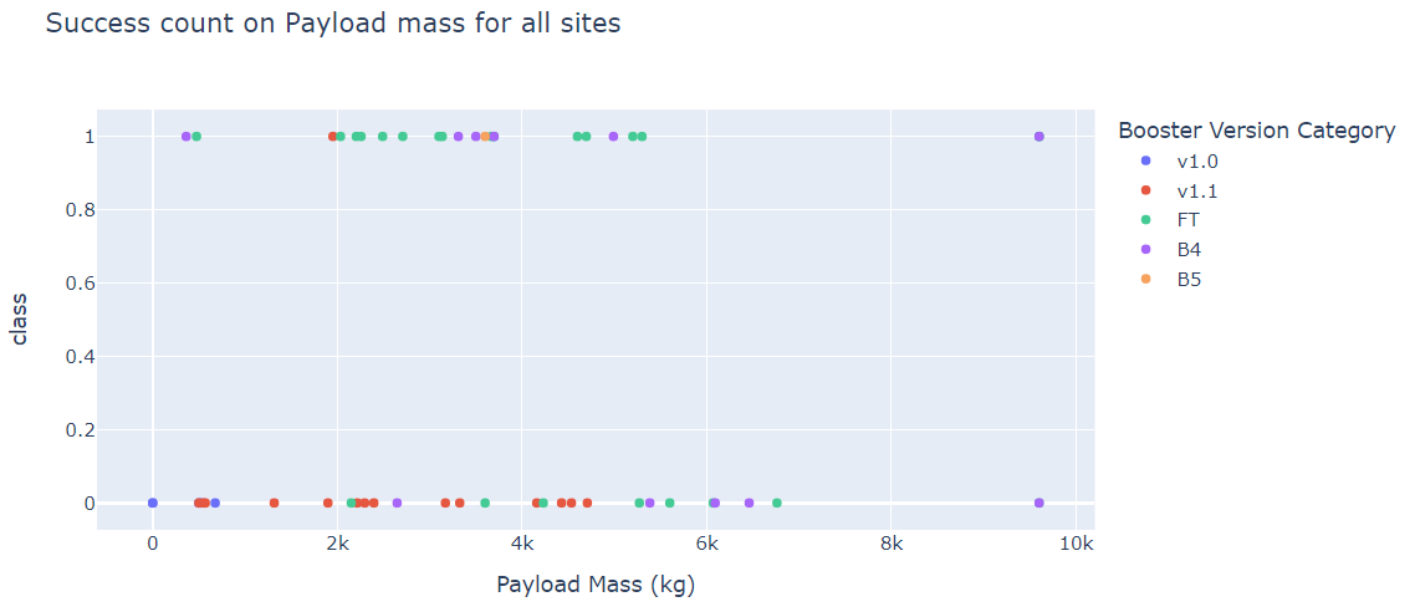
- The pie chart show the success count for all launch sites
- KSC LC-39A have the biggest part of the success count and CCAFS SLC-40 habe the smallest part

Launch site with highest launch success ratio



- KSC LC-39A have the highest launch success ratio
- 41.7 % of the starts have a positive outome

Payload vs. Launch Outcome of all Launch Sites

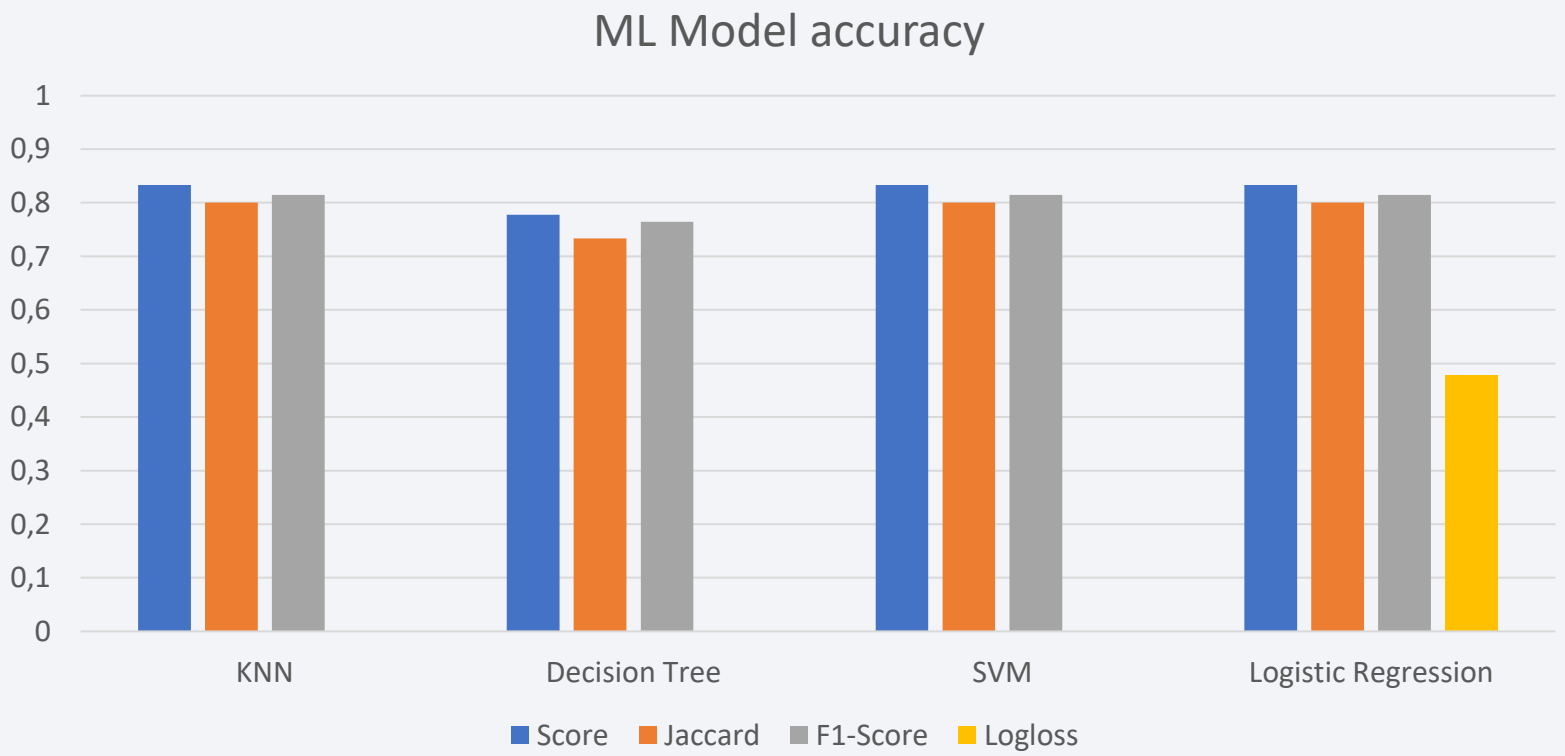


- The Booster Version FT have the highest success rate
- The Booster Version v1.0 have no positive success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- KNN, SVM and Logistic Regression have the highest accracy
- The result of logLoss is not comparable with the other scores

Confusion Matrix

- That is the KNN Confusion Matrix

Conclusions

- There is a of progress in positive launch outcomes with increasing number of launches each year
- The launch sites are next to ocean and have a huge distance to the daily live, the KSC LC-39A have the highest success rate
- Launches with a high payload mass have in average a higher success rate than launches with lower payload mass
- The success rate was also dependent on the orbit and payload mass, we saw that ISS and VLEO orbits had a good success rate.
- Support Vector Machine was a suitable model to predict if the stage one would land or not, it had an accuracy of 83%

Thank you!