IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zekarias T. Birhanu
May 5, 2022

Outline

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

Executive Summary

- Competitor companies of SpaceX are at a disadvantage to offer a fair price on the rocket launch business.
- By cutting around one third of the price to launch a rocket, compared to other competitors, SpaceX seems to have won the race on launching rockets.
- This huge price saving can be gained by competitors had they implemented the functionality of reusing the first stage of the rocket.
- To address this, we will use collected data and prediction algorithms to determine if the first stage will land. Based on that, we will determine the cost for the launch.
- This can then be used help competitor companies to have a chance on competing with SpaceX.

Introduction

- Civilization on earth has enabled humans to enjoy space travel, however this comes with a super expensive cost. A company called SpaceX has done a very good job in cutting this cost of travel by around one third.
- But, other players in the field has found it hard to compete with SpaceX and we know why SpaceX comes to a lesser expensive solution. One of the reasons is reusing the first stage of the rocket launcher. Had competitor companies also apply the same methodology, we will have competition on the field and even cheaper space travel fee.
- So, why not study the free open data from SpaceX that can be found online and use predication algorithms to help and make the bid tougher.
- Using predication algorithms we will determine if the first stage will land, hence determine the cost of the launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Perform data wrangling
- 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 Collection

- In Data Science one of the things that should be done to answer our question is to collect data.
- For this task we have used two kinds of data collection methods:
  - The SpaceX REST API
  - Web scraping from Wikipedia

Data Collection – SpaceX API

- Basically two things happened on this step.
  - First the launch data was requested from SpaceX API using GET request.
  - The Data was cleaned
- GitHub URL:
  - https://github.com/lordakarias/IBM-DATA-SCIENCE-Coursera/blob/master/1.%20Data%20Collection%20API.ipynb

spacex_url="https://api.spacexdata.com/v4/launches/past"

response = requests.get(spacex_url)

- GET launch data

pd.json_normalize(response.json())

- convert the json result into a dataframe

pd.DataFrame()

- Create a Pandas data frame from the dictionary launch_dict.

data_falcon9 = df1[df1['BoosterVersion']!='Falcon 1']

- Filter only Falcon 9 launches

Data Collection - Scraping

static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"

- Python BeautifulSoup was used to scrap a wiki page to extract a table containing Falcon 9 launch records.
- GitHub URL:
  - https://github.com/lordakarias/IBM-DATA-SCIENCE-Coursera/blob/master/2.%20Web_Scraping.ipynb

response = requests.get(static_url)

- GET Falcon9 Launch HTML page

soup = BeautifulSoup(response.text, "…")

- Create a BeautifulSoup object

html_tables = soup.findAll("table")

- Find all tables on the wiki page

data_falcon9 = df1[df1['BoosterVersion']!='Falcon 1']

- Filter only Falcon 9 launches

Data Wrangling

“Data wrangling is **the process of cleaning and unifying messy and complex data sets for easy access and analysis” [1]**

- Convert mission outcomes into Labels 1 and 0 for successfully unsuccessful booster landing, respectively.
- GitHub URL:
  - https://github.com/lordakarias/IBM-DATA-SCIENCE-Coursera/blob/master/2.%20Web_Scraping.ipynb

df=pd.read_csv(...)

- Load Space X dataset

df['LaunchSite'].value_counts()

- Calculate the number of Space X launch facilities

df['Orbit'].value_counts()

- Calculate the number and occurrence of each orbit

landing_outcomes = df['Outcome'].value_counts()

- determine the number of landing_outcomes

df['Class']=landing_class

- **Create a landing outcome label**

EDA with Data Visualization

- Charts that were plotted for this task were:
  - **FlightNumber** vs. **PayloadMass**: to see the relationship between flight number and success of first stage landing. Payload Mass and success of first stage return.
  - Relation ship between **Flight Number** and **Launch Site** ¶
  - **Payload** Vs **Launch Site**
  - Success rate of each **orbit type**
  - **Flight Number** and **Orbit type**
  - **Payload** and **Orbit type**
- GitHub URL

EDA with SQL

- 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 achieved.
- 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 total number of successful and failure mission outcomes
- 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
- [GitHub URL]()

Build an Interactive Map with Folium

- Create a folium map with start location at NASA Johnson Space Center and add a highlighted circle area
- Add a circle for each launch site
  - To have a general view of the location of launch sites and start asking questions that can help us answer some of my initial questions.
- For each launch site on the map, add a folium.Marker
  - To easily identify which launch sites have relatively high success rates.
- Create a marker with distance to a closest costline, railway,
- [GitHub URL]

Build a Dashboard with Plotly Dash

- Plots added to the dashboard:
  - Pie-chart with drop-down to see launch success rates of launch sites
  - Scatter point chart with range slider to see which payload and booster versions haves a successful launch rate
- [GitHub URL]

Predictive Analysis (Classification)

Create a column for the class

Standardize the data

Split the data into training data and test data

Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

Find the method performs best using test data

- GitHub URL

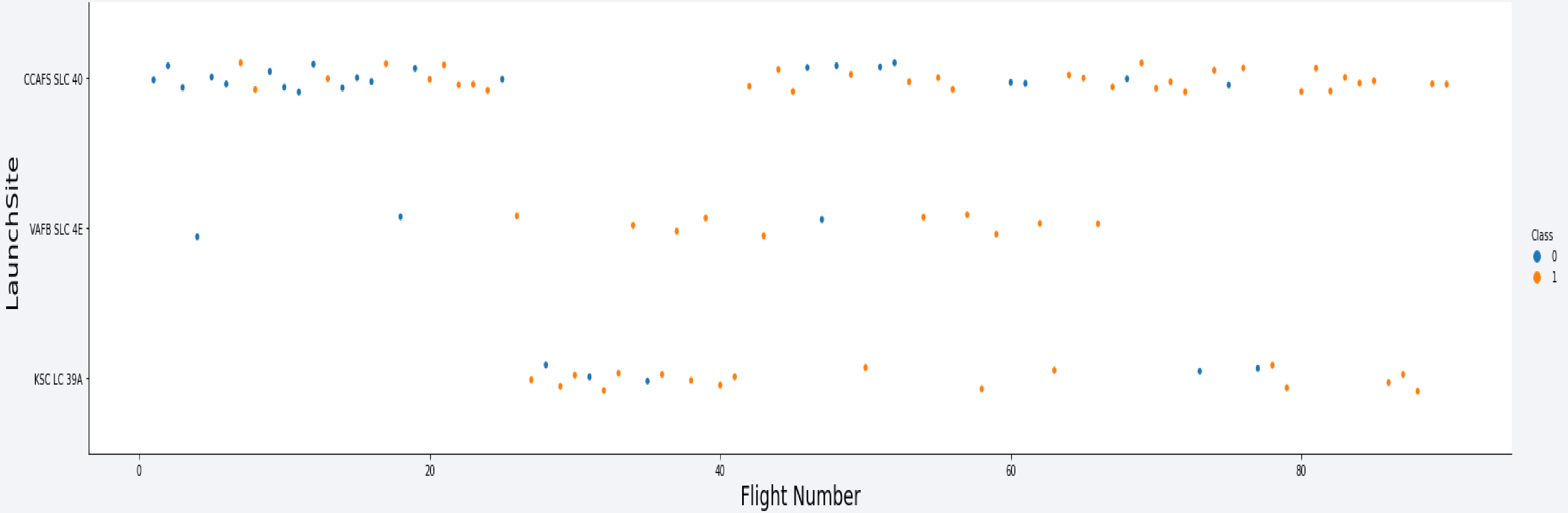
Results

- We have used data visualization and SQL queries to visualize the data and extract meaningful patterns to guide the modeling process.
- After that Interactive analytics was done on Launch Sites.
- Then used machine learning to model if the first stage of Falcon 9 will land successfully. Except the decision tree algorithm all the other three has an equal accuracy on prediction.

Section 2

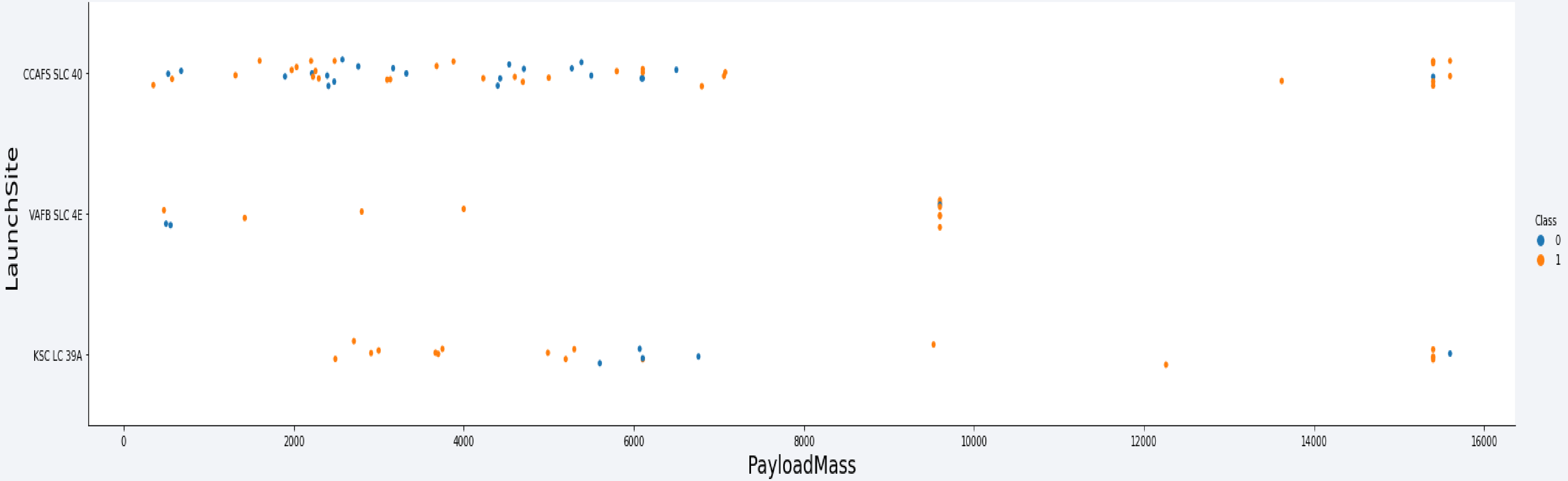
Insights drawn from EDA

Flight Number vs. Launch Site



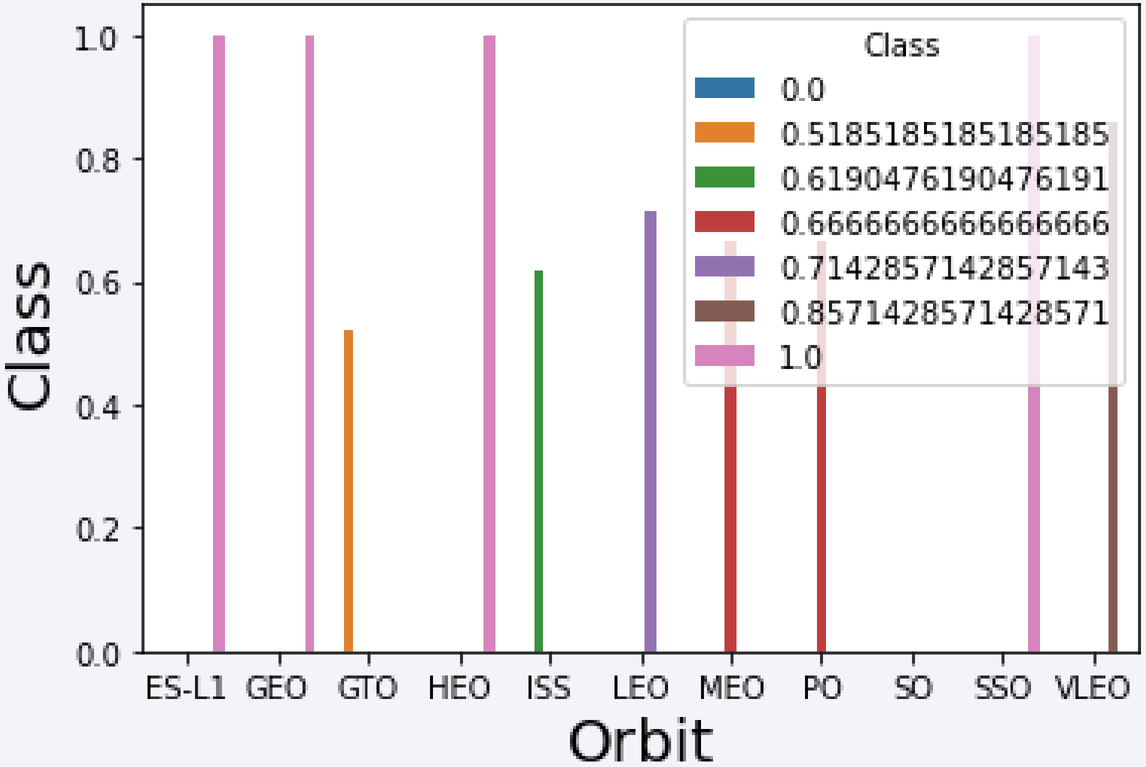
- The scatter plot shows, when the flight number increases the success rate of launches also increases on all launch sites.

Payload vs. Launch Site



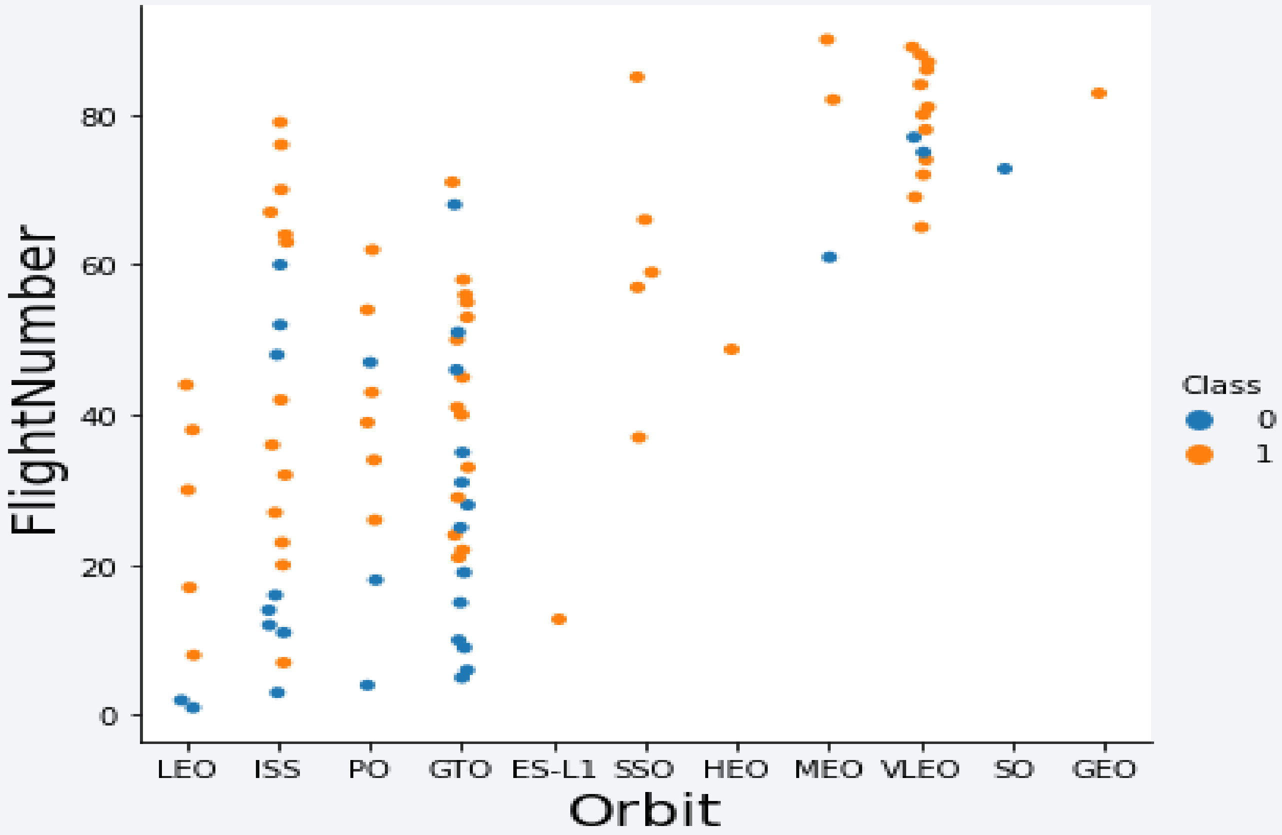
- As the PayloadMass increases the launch success rate seems to increase, but that might not be the whole story behind the success rate.
- Relatively VAFB SLC 4E launch site seems lower number of failure rate, and have lower number of launches.

Success Rate vs. Orbit Type



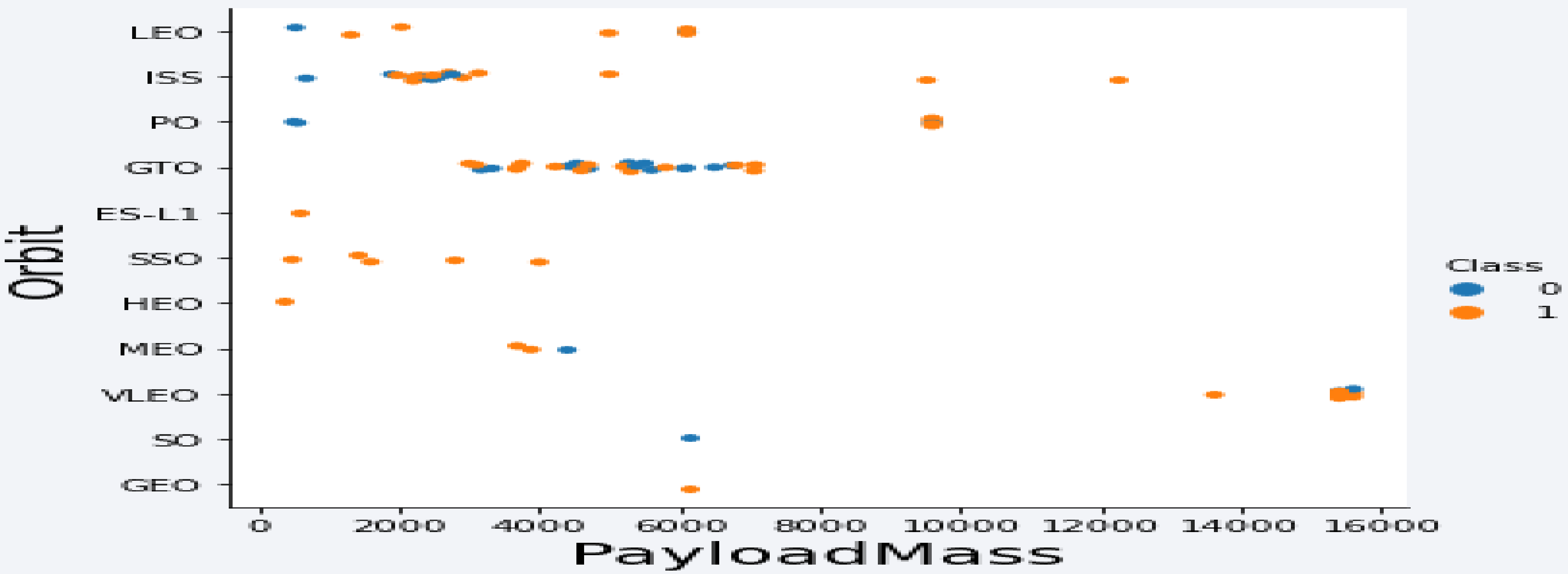
- Success rate at orbit SO and GTO is very low, followed by ISS, LEO and PO which also have a success rate less than 1.
- Orbits ES-L1, GEO, HEO, and SSO are orbits where success rate is very good and with additional more tests, we can say these can be the perfect orbits types SPACEX launches work at.

Flight Number vs. Orbit Type



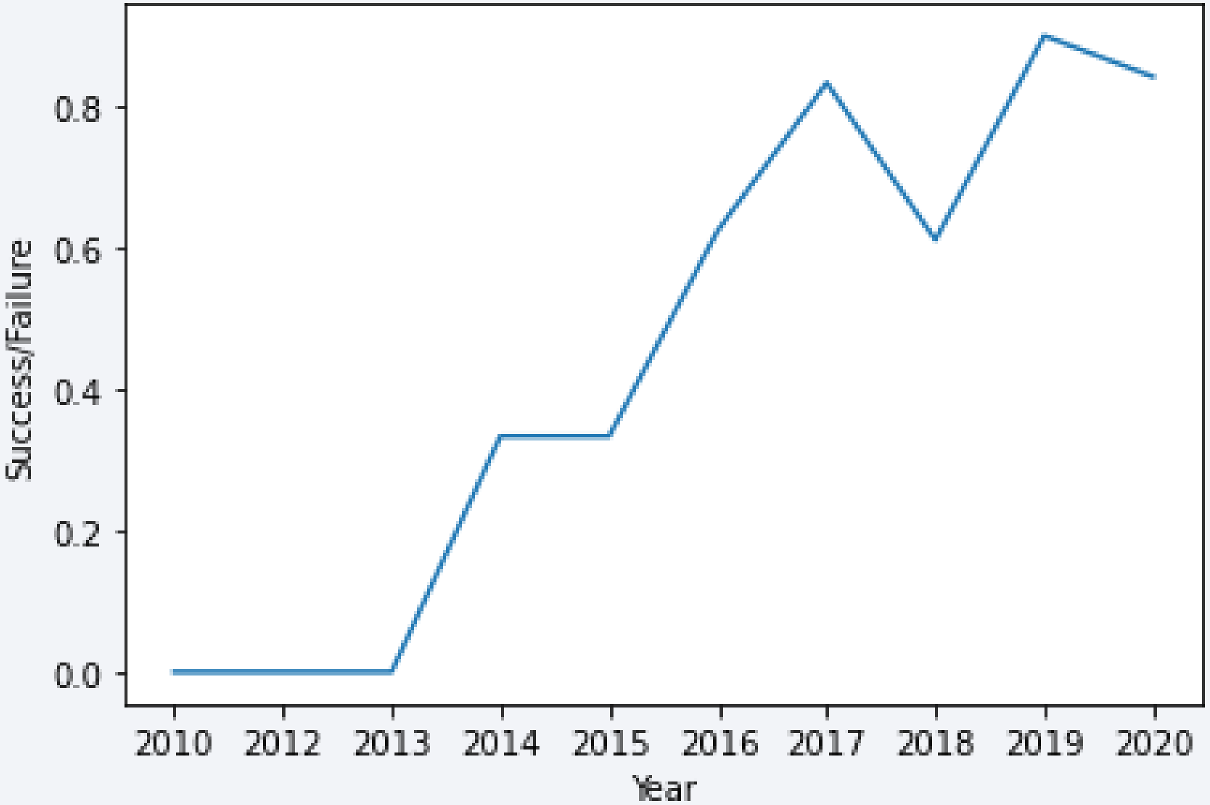
- It shows as the flight number increases, almost for every orbit type, success rate of launch increases. But, we might need more data to reach to conclusion if this is really the case for all the orbits, as some of the orbits have smaller number of tests.

Payload vs. Orbit Type



- Even if we need more data to confidently to react to conclusion, the bigger PayloadMass launches (>8000) have shown higher success rates.

Launch Success Yearly Trend



- Except the year 2018 (which needs further investigation), for most of the time, the success rate was increasing year after year.

All Launch Site Names

- We have four unique launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- Query: %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL2

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites 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

- Query: %sql select * from SPACEXTBL2 WHERE Launch_Site LIKE 'CCA%' LIMIT 5

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - The total payload carried by boosters from NASA is 45596 Kg
- Query: %sql select SUM(payload_mass__kg_) from Spacextbl2 where customer='NASA (CRS)'

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - The average payload mass carried by booster version F9 v1.1 is 2928 Kg
- Query: %sql select AVG(payload_mass__kg_) from Spacextbl2 where booster_version='F9 v1.1'

First Successful Ground Landing Date

- The first successful landing outcome on ground pad happened at 2015-12-22
- Query: %sql select min(DATE) from SPACEXTBL2 where landing__outcome='Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000

- Name of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are

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

- Query: %sql select booster_version from SPACEXTBL2 WHERE landing__outcome='Success (drone ship)' AND payload_mass__kg_ between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
  - Mission_outcome 2
  - Failure (in flight) 1
  - Success 99
  - Success (payload status unclear) 1

- Query: select mission_outcome,count(*) from SPACEXTBL group by mission_outcome;

Boosters Carried Maximum Payload

- Booster which have carried the maximum payload mass
  - 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
- %sql select booster_version,payload_mass__kg_ from SPACEXTBL2 WHERE payload_mass__kg_= (select MAX(payload_mass__kg_) from SPACEXTBL2)

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - booster_version   launch_site
  - F9 v1.1 B1012   CCAFS LC-40
  - F9 v1.1 B1015   CCAFS LC-40

- %sql select booster_version,launch_site from SPACEXTBL2 WHERE EXTRACT(YEAR FROM DATE)='2015' AND landing__outcome='Failure (drone ship)'

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

- Rank the count of landing between the date 2010-06-04 and 2017-03-20, in descending order
  - landing__outcome Count
  - 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 "Count" FROM SPACEXTBL2 WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'group by

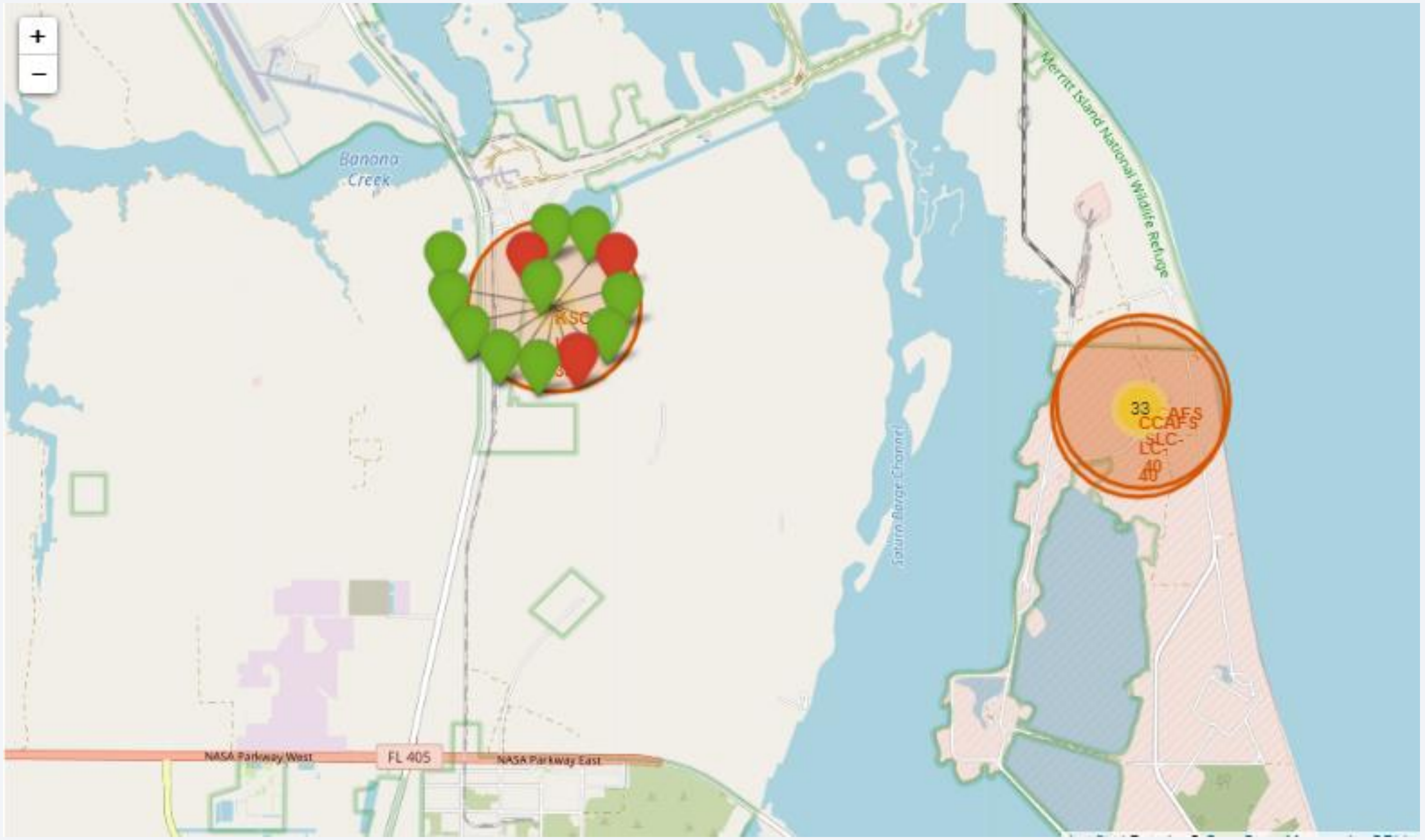
Section 4

Launch Sites Proximities Analysis

SpaceX launch sites

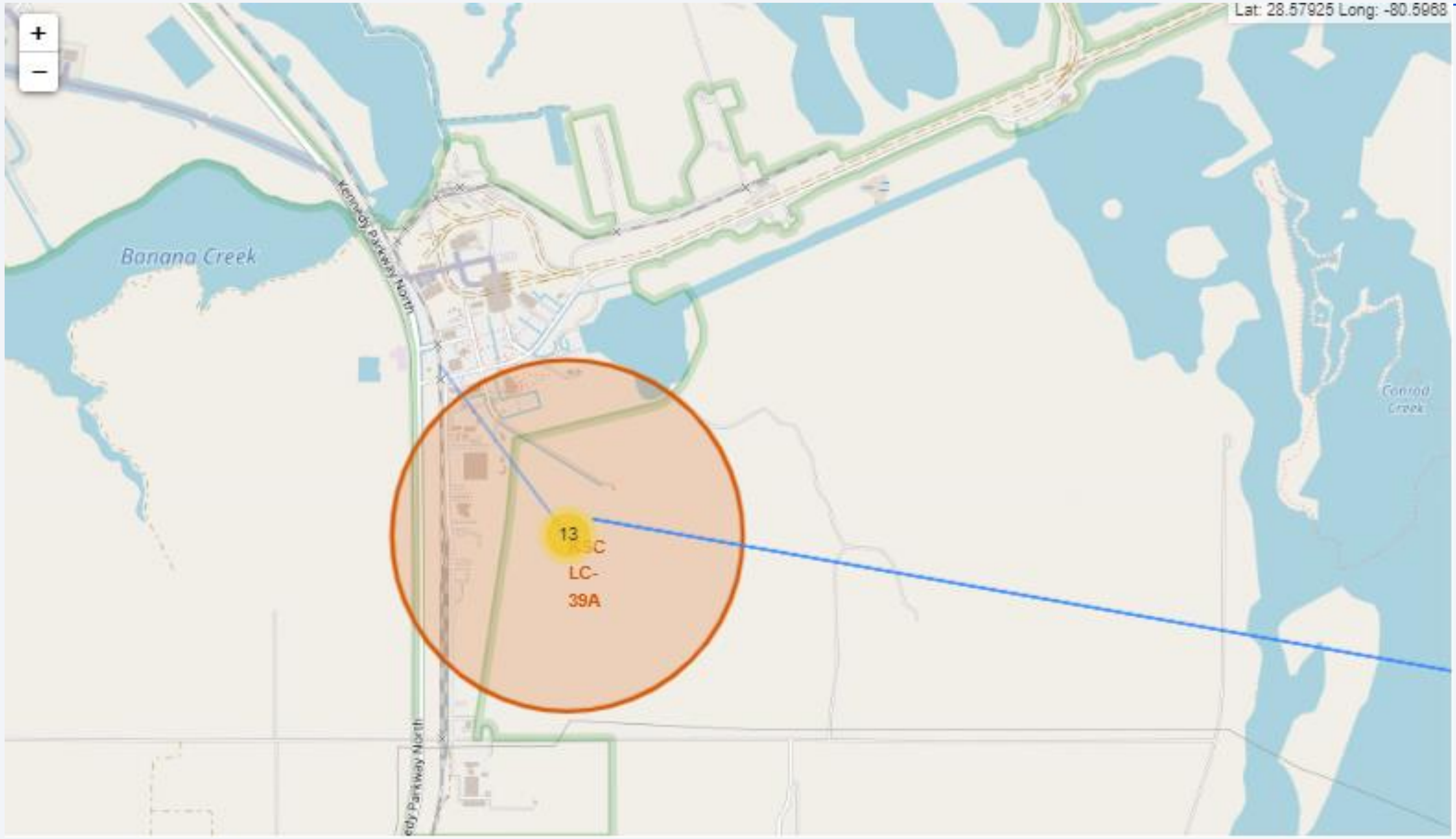
- All four Launch sites are located on Cost lines

Success/failed launches



- We have used here green markers to indicate success and red ones for failure.

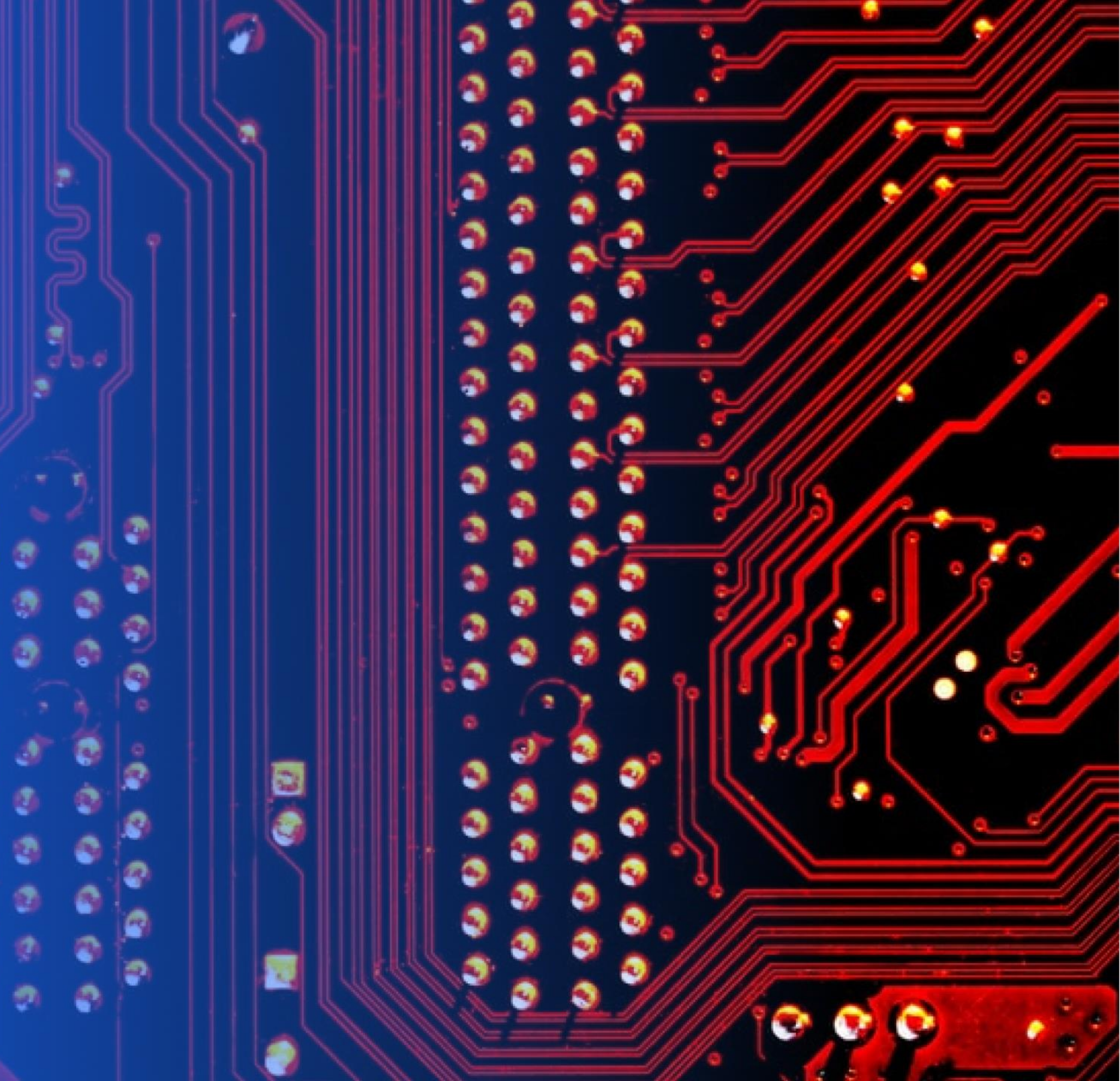
DISTANCE FROM LAUNCH SITE TO NEAREST RAILWAY STATION



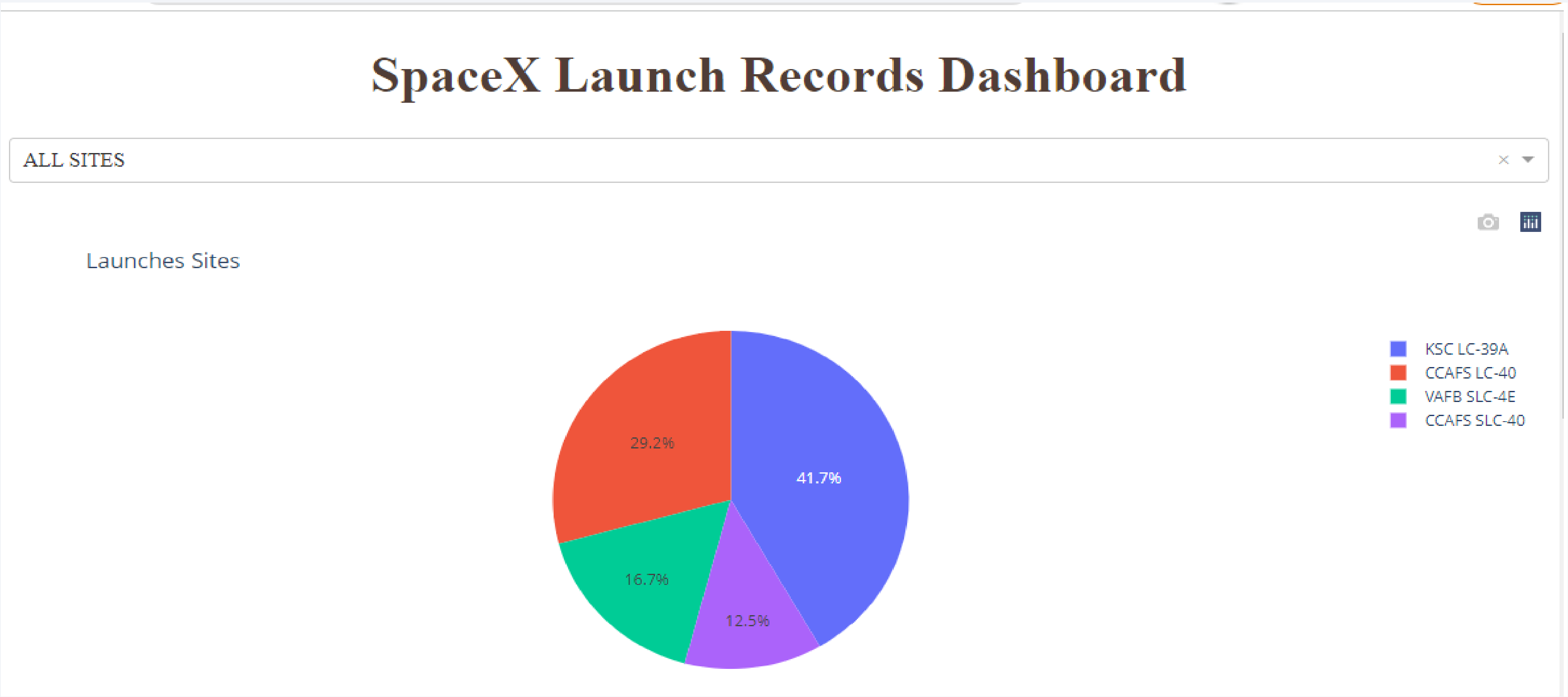
• Launch site to its proximities railway,

Section 5

Build a Dashboard with Plotly Dash

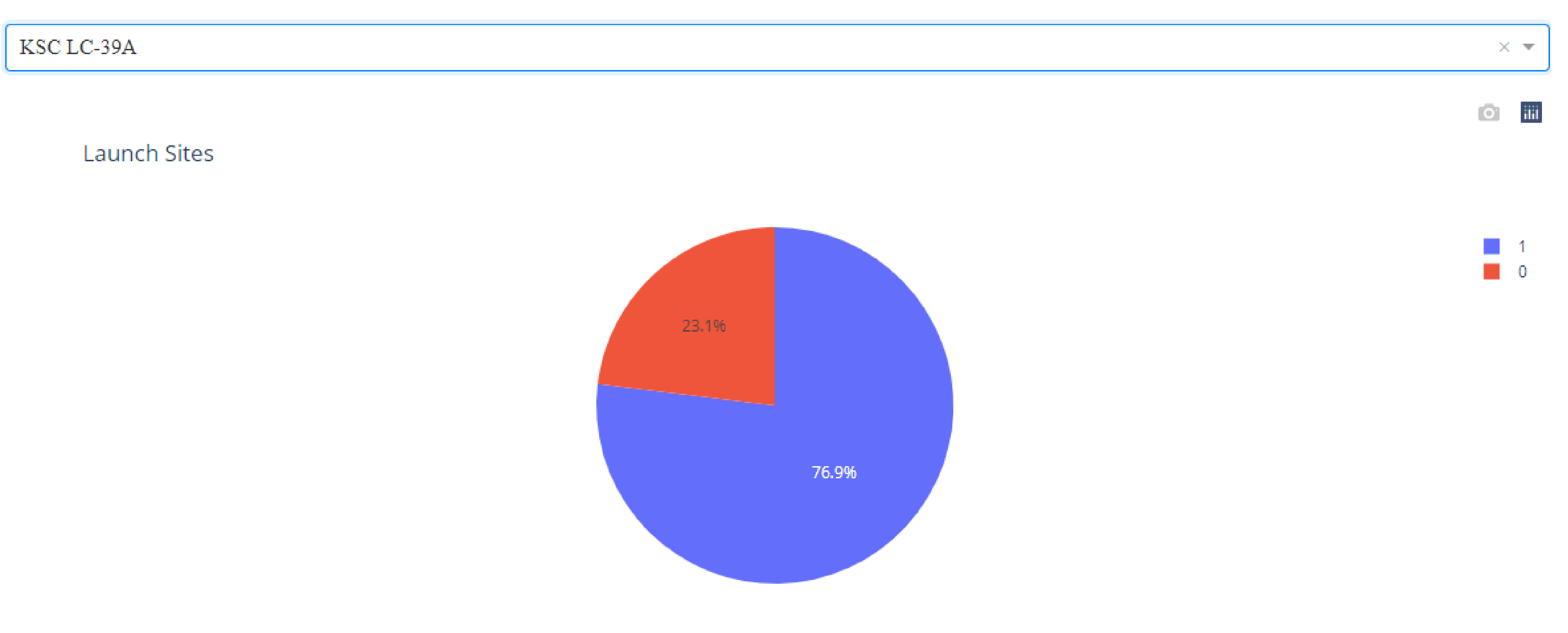


LAUNCHE SUCCESS BY SITE



- KSC LC has a highest launch success rate and CCAFS SLC has the least.

HIGHEST LAUNCH SUCCESS RATIO



- The KSC LC-39A has 76.9% launch success and a 22.1% failure ratio.

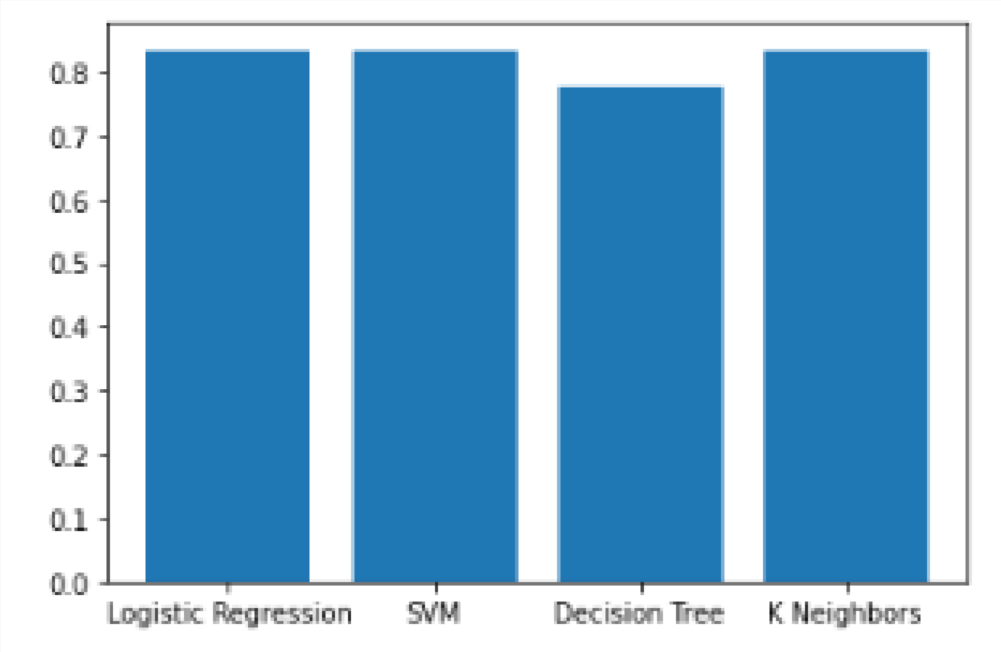
Payload vs. Launch Outcome

- From the result FT has the highest number of success launches and V1.1 the lowest

Section 6

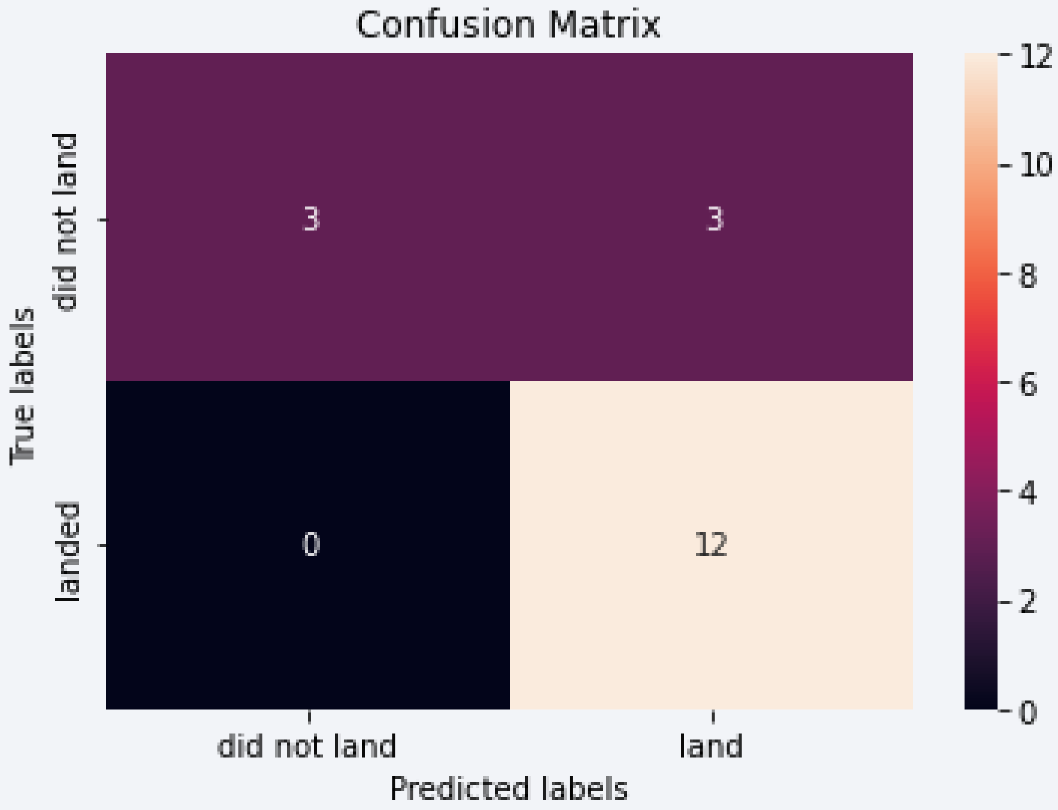
Predictive Analysis (Classification)

Classification Accuracy



- Expect the Decision tree all three have an accuracy score of 0.833

Confusion Matrix



- The best Confusion Matrix of the Classifier has the following results:
  - True Positive : 12, False Negative : 0, True Negative : 3, False Positive : 3
- For most of the cases the prediction is very good, except when it predicates successful landing but the true label indicates failure landing

Conclusions

- Our work has shown the powerful application of data and Machine Learning (ML). Very complicated and expensive tasks of such kind can be solved using ML.
- The more data we collect and use, the more confidence we will have on the algorithm predication.
- We have also identified different parameters like orbit type, launch site, PayloadMass, BoosterVersion and their impact on the success of the landing.
- Hence, using the data from SPACEX we can predict if the first stage can land successfully and determine the cost for the launch.

Reference

1. https://www.altair.com/what-is-data-wrangling

Thank you!