IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Sergio Fernández Acosta>
<11th Dec 2024>

Outline

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

Executive Summary

- Publicly available data has been collected, processed and analyzed to build and validate multi-variable predictive classification models for the outcome of space missions.
- The **selected model proved to be reliable with an 85% accuracy predicting the outcome of a mission** landing the booster for future reuse.
- Also, additional insights were gained during the process, including exploratory data analysis and visualization. Some examples are:
  - **Success rate has a strong correlation with the flight number**. This is an unequivocal trend showing that there is a steep learning curve in this field. **This is crucial information for us, Space Y, as a new player in this market** and will help us overcome this challenges.
  - We know which **launch/landing sites (KSC LC-39A)** and combination of **booster versions/architectures (FT) and payload masses (<5000kg)** have the highest success rate
  - We also know which orbits are more demanded by potential customers and their associated risks

Introduction

- Space Y is a new company aiming to revolutionize the field of space exploration and exploitation using cost-efficient, reliable, reusable and sustainable rockets.
- We will **leverage from existing public data** of our competitors and governmental space agencies to overcome our late arrival in the market.
- In order to prepare our first strategic plan, we will rely on out departments of Data Science and Design to help **predict launch outcomes based on multivariable analysis**.
- This derived insights will be used to adapt our plans accordingly
  - Development: select specific Booster design options that are more reliable
  - Deployment: select the best possible launch sites
  - Funding: estimate costs throughout the lifecycle of the project

Section 1

Methodology

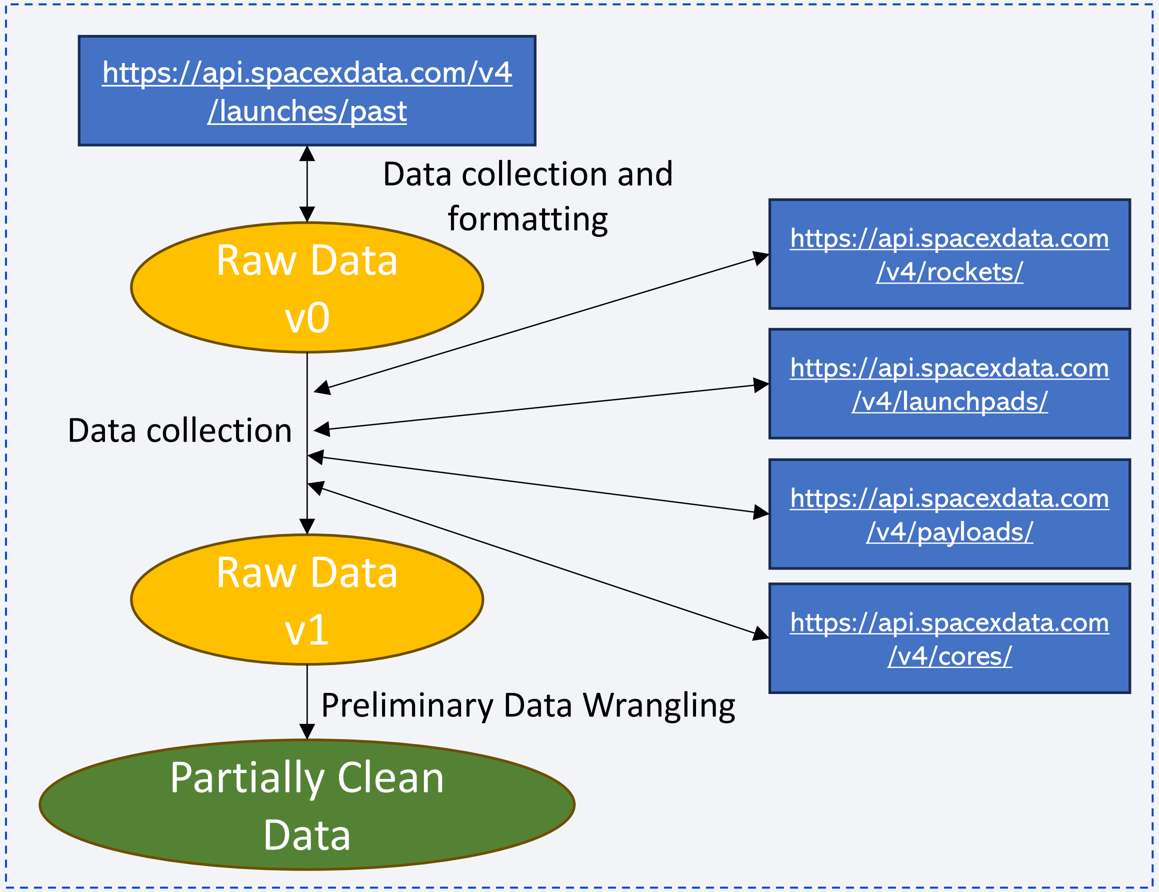
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using web scrapping techniques and data APIs from publicly available sites obtaining a large enough data set with multiple variables.
- Perform data wrangling
  - Data was processed and cleaned using the best practices and methodologies to ensure data completeness (e.g., replacing missing data) and consistency, and proper format and precision (e.g., converting categorical variables and then also numerical variables to double precision floats) required for the analysis and modeling techniques we used.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Several models were developed and tuned using cross-validation techniques. Finally, the deployed model was validated using specific test data to avoid overfitting and estimate its predictive accuracy.

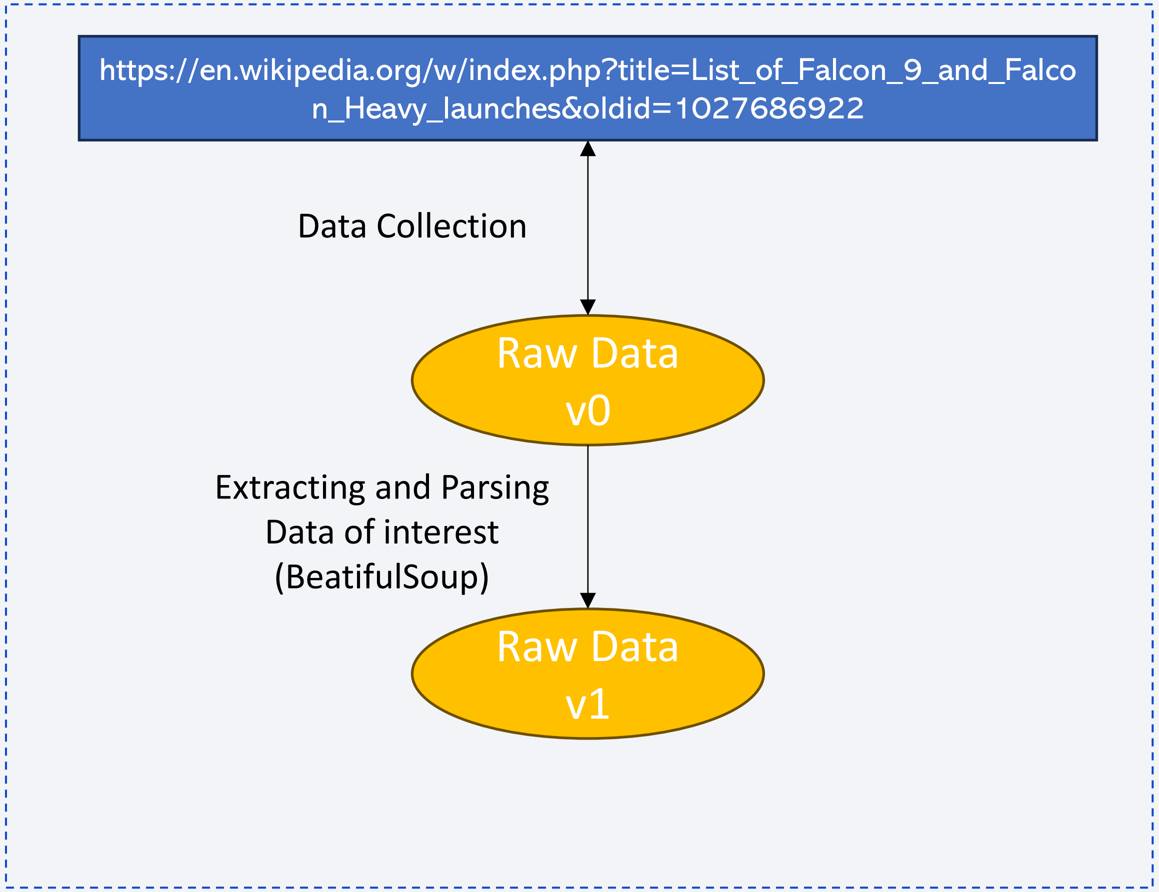
Data Collection – SpaceX API

- Launch raw data obtained with **REST API** (requests.get method) at https://api.spacexdata.com/v4/launches/past
- JSON response converted to Pandas DataFrame
- IDs in raw data used to **gather additional useful information** like Booster Versions, Payload and Cores data, and Launch Sites names and coordinates through specific REST interfaces (*launchpads*, *rockets*, *payloads* and *cores*)
- Data was then filtered to only investigate **Falcon 9** historical launches.
- Basic **Data Wrangling** performed to replace missing Payloads with the overall average value.
- **GitHub URL:** https://github.com/uvenabla/public_IBM_DataScience_Certificate/blob/411a3535c913346b13262311f6d7fa92d6592ac1/jupyter-labs-spacex-data-collection-api-v2.ipynb


https://api.spacexdata.com/v4/launches/past
Data collection and formatting
Raw Data v0
https://api.spacexdata.com/v4/rockets/
https://api.spacexdata.com/v4/launchpads/
Data collection
https://api.spacexdata.com/v4/payloads/
https://api.spacexdata.com/v4/cores/
Raw Data v1
Preliminary Data Wrangling
Partially Clean Data

Data Collection - Scraping

- **Raw Data collected** from HTML tables from Wikipedia **using web scraping** technique (*requests.get*)
- Iteratively Extracting and **Parsing the information** from the tables (*BeatifulSoup*)
- **Converting to** a *Pandas* ***DataFrame*** object.
- GitHub URL: https://github.com/uvenabla/public_IBM_DataScience_Certificate/blob/7e5c112e262c153ca3cc948f3db5d8002d4816e8/jupyter-labs-webscraping.ipynb



https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
Data Collection
Raw Data v0
Extracting and Parsing Data of interest (BeatifulSoup)
Raw Data v1

Data Wrangling

- Data was **checked for consistency** (missing data, data format, etc.)
- Launch Sites and Types of Orbits basic statistics were explored
- **Outcomes of the mission were explored**. This column combines outcome and attempted landing pad.
- We **create new label** as we are only interested in the **landing outcome**. It is a **binary** metric named "*Class*": 0 for failures and 1 for successful landings.
- At a later stage (EDA), we also created **dummy variables** with OneHotEncoder **to convert categorical into binary** (numeric) variables for modelling.
- **GitHub URL:** https://github.com/uvenabla/public_IBM_DataScience_Certificate/blob/3668815d78e748f2a8f212e9d7a2805e6f3ddc0b/labs-jupyter-spacex-Data%20wrangling-v2.ipynb

EDA with Data Visualization

- The following Charts were developed as we want to **explore if there are any correlations between the different variables**.
  - "*Relationship between Flight Number and Launch Site*"
  - "*Relationship between Payload and Launch Site*"
  - "*Relationship between success rate of each orbit type*"
  - "*Relationship between Flight Number and Orbit type*"
  - "*Relationship between Payload and Orbit type*"
  - "*Launch success yearly trend*"
- **GitHub URL:** https://github.com/uvenabla/public_IBM_DataScience_Certificate/blob/f9869f04df34ad94495061eea887689a7615c514/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- The following SQL queries were performed using %sql magic commands:
  1. Unique launch sites: ***select distinct Launch_Site from SPACEXTBL limit 100***
  2. 5 records where launch sites begin with the string 'CCA’: ***select * from SPACEXTBL where Launch_Site LIKE 'CCA%' limit 5***
  3. total payload mass carried by boosters launched by NASA (CRS): ***select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer ='NASA (CRS)'***
  4. average payload mass carried by booster version F9 v1.1: ***select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1’***
  5. date for the first successful landing outcome in ground pad: ***select min(Date) from SPACEXTBL where Landing_Outcome not like '%Failure%'***
  6. names of the boosters which have success in drone ship and payload mass greater than 4000 but less than 6000: ***select distinct Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000 limit 10***
  7. total number of successful and failure mission outcomes: ***select distinct Mission_Outcome from SPACEXTBL***
  8. names of the booster_versions which have carried the maximum payload mass: ***select Booster_Version,max(PAYLOAD_MASS__KG_) from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL) limit 1***
  9. month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015: ***select substr(Date, 6,2),Booster_Version, Launch_Site from SPACEXTBL where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'***
  10. Rank the count of landing outcomes: ***select Landing_Outcome,count(Landing_Outcome) as n from SPACEXTBL where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by n desc***

- **GitHub URL:** https://github.com/uvenabla/public_IBM_DataScience_Certificate/blob/06b675fda181ef274c9c6fd39957f97150133295/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Circles and Markers** were created at the **NASA Headquarter and all Launch Sites** to **show their location**.
- Lines and additional Markers were created **to show distance from one selected Launch Site to the nearest Railway/Highway, Coastline and City** as we want **to explore logistics and safety aspects** of the sites.
- **Markers within Marker Clusters were created to show** the **total number of launches and their outcome from each site** and a color used to represent if it was successful (green) or a failure (red) since we want **to understand if a there is any correlation between Location/site and success rate**.
- **GitHub URL:** https://github.com/uvenabla/public_IBM_DataScience_Certificate/blob/514799a75da540b3114a0fdbce8c2bc0d178b050/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- A **dropdown box** was used to select the **Site of interest**. You can select any of the different Launch Sites or All if you want to analyze global information.
- **This dropdown menu interacts with and control two different plots**:
  - A **Pie Chart to show the success rate** for the selected site or the proportion of successful launches for all the sites
  - **A Scatter plot of Outcome (class) vs Payload Mass in kg, differentiating with colors the Booster Version Category**.
    - In order to analyze different Payload Mass ranges for the scatter plot, we created and **interactive Slider where you can pick the minimum and maximum mass of interest**.
- We are interested in **understanding if the Launch Site has any influence in the success rate** and also whether the **Booster Version Category and Payload Mass also correlates with success**.
- **GitHub URL:** https://github.com/uvenabla/public_IBM_DataScience_Certificate/blob/4acc45c9fe59db0f7d442e54897d7e1dc534df7c/spacex_dash_app.py

Predictive Analysis (Classification)

- We implemented **4 different classification models**: Logistic Regression, Support Vector Machines, Tree and K-Neighbors.
- **Data was split into Train (80%) and Test (20%) data**
- **Preselected HyperParameters were optimized** for each model to obtain the best performance ("accuracy") using Train data and 10 folds with Cross Validation (*GridSearchCV*)
- **Accuracy Scores were obtained** for model performance with train and test data.
- **Confusion Matrix** then examined on test data to confirm the model is sufficiently generalized.
- **GitHub URL**:
- Executed on Cloud Environment => https://github.com/uvenabla/public_IBM_DataScience_Certificate/blob/fe28b4ba770aaf03c9a1a4d7d4945c8467d98255/SpaceX_Machine%20Learning%20Prediction_Part_5_cloud.ipynb
- Executed Locally with other parameters => https://github.com/uvenabla/public_IBM_DataScience_Certificate/blob/fe28b4ba770aaf03c9a1a4d7d4945c8467d98255/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

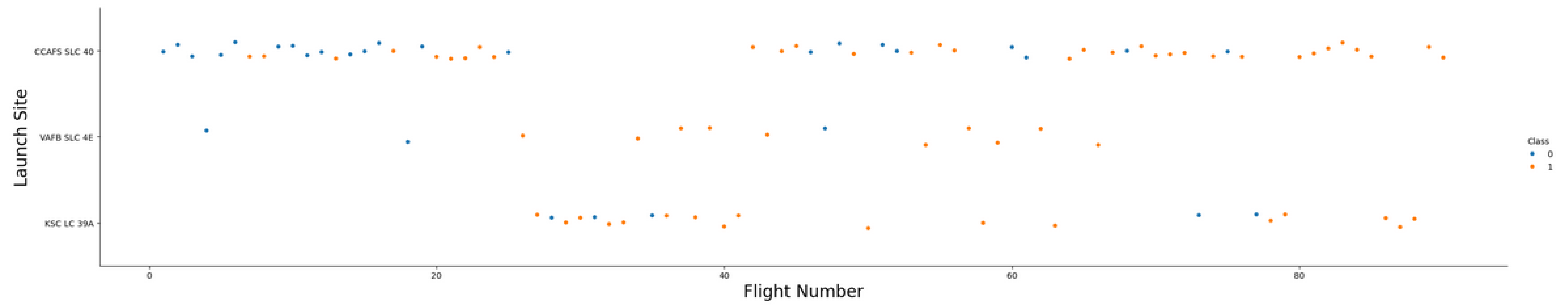
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

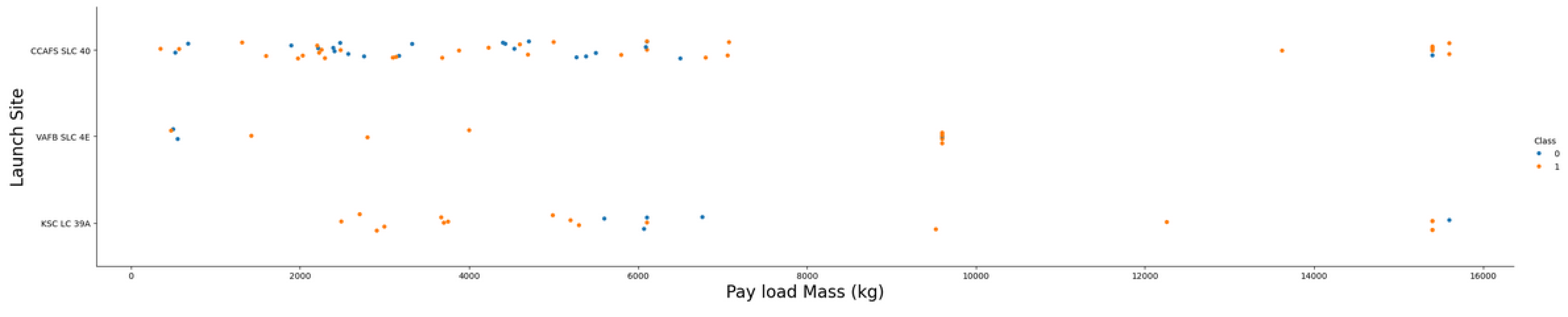
Flight Number vs. Launch Site

- **Site CCAFS SLC-40 and VAFB SLC 4E were used for the first 20+ launches with not too good results**.
- **Site CCAFS SLC-40 wasn't used for a period of time (flights 25-40),** but used again since flight number 40 and there after with much better outcomes.
- **Site VAFB SLC 4E has been much less used than the other sites**.
- **Site KSC LC-39A** was probably **not available until flight 25, and used intensively until flight ~41**, and more sporadically there after, but with good overall outcomes.

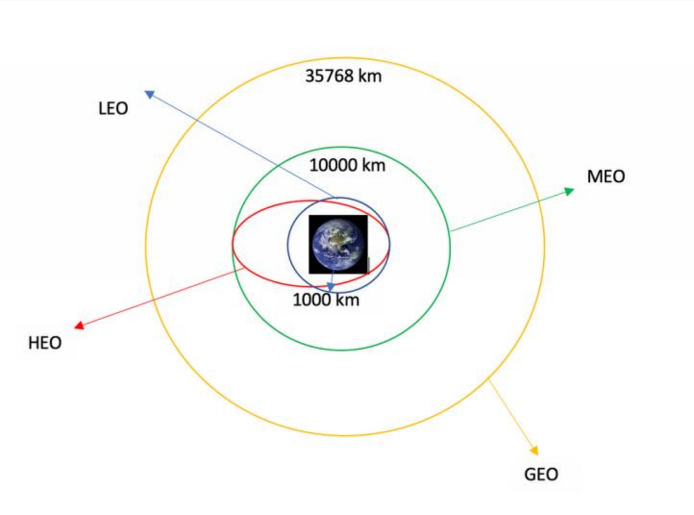


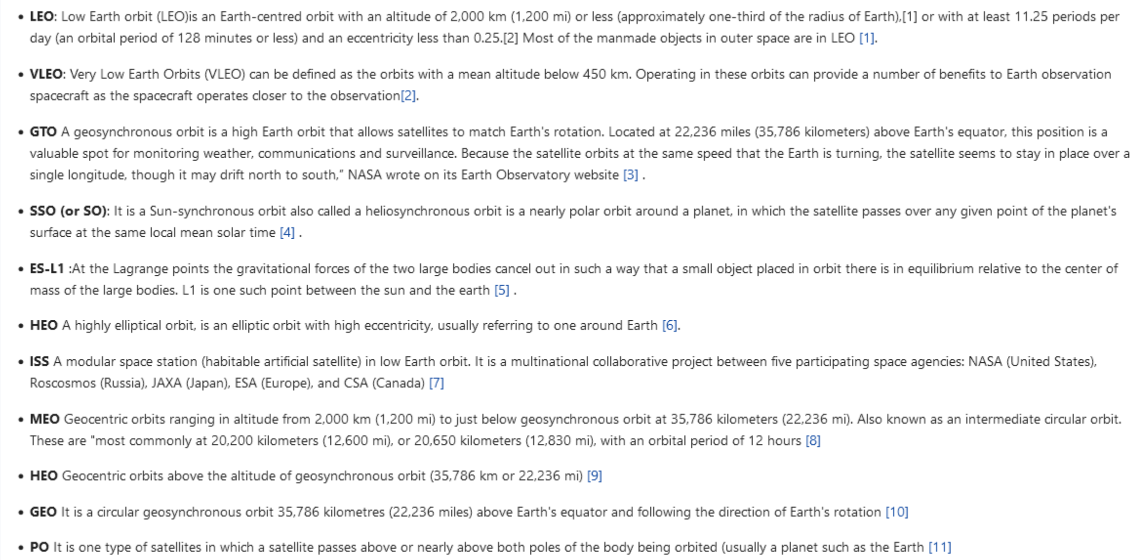
Payload vs. Launch Site

- There is **not any clear correlation between the explored variables**. However, we notice the following
  - Site VAFB SLC 4E has never been used for payloads larger than 10000kg
  - Site CCAFS SLC-40 has been used for low and large payload masses with modest success rate
  - Site KSC LC-39A has been used for intermediate and large payloads, with very good success rate.



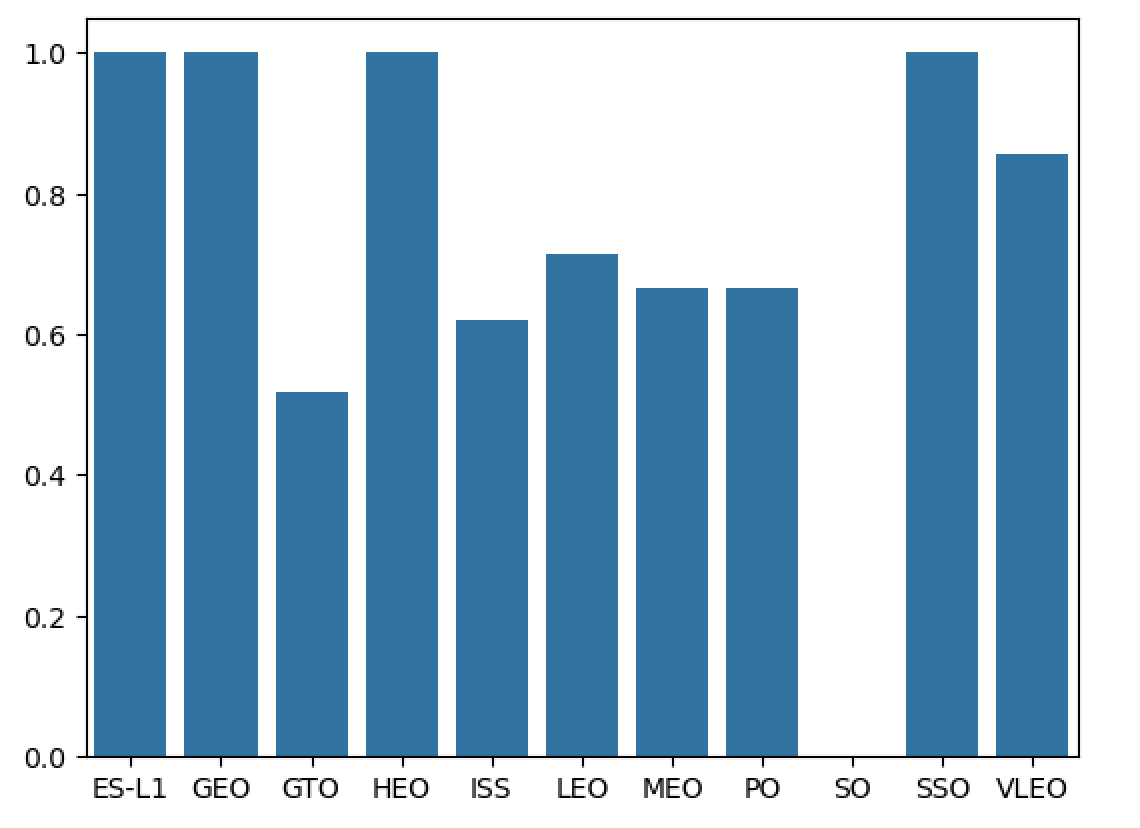
Success Rate vs. Orbit Type





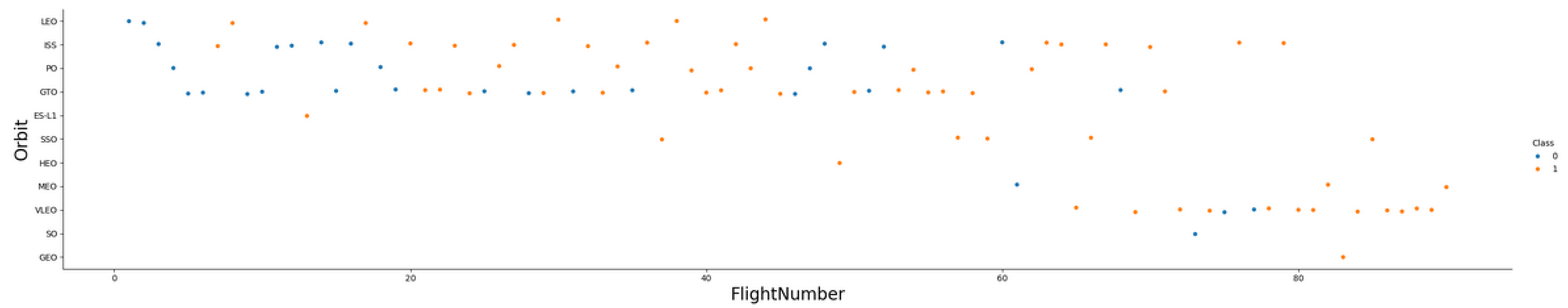
- **LEO**: Low Earth orbit (LEO)is an Earth-centred orbit with an altitude of 2,000 km (1,200 mi) or less (approximately one-third of the radius of Earth),[1] or with at least 11.25 periods per day (an orbital period of 128 minutes or less) and an eccentricity less than 0.25.[2] Most of the manmade objects 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 (usually a planet such as the Earth [11]

- The orbits with the **highest success rates are ES-L1, GEO, HEO and SSO**.
- The orbits with the lowest success rates are SO and GTO.



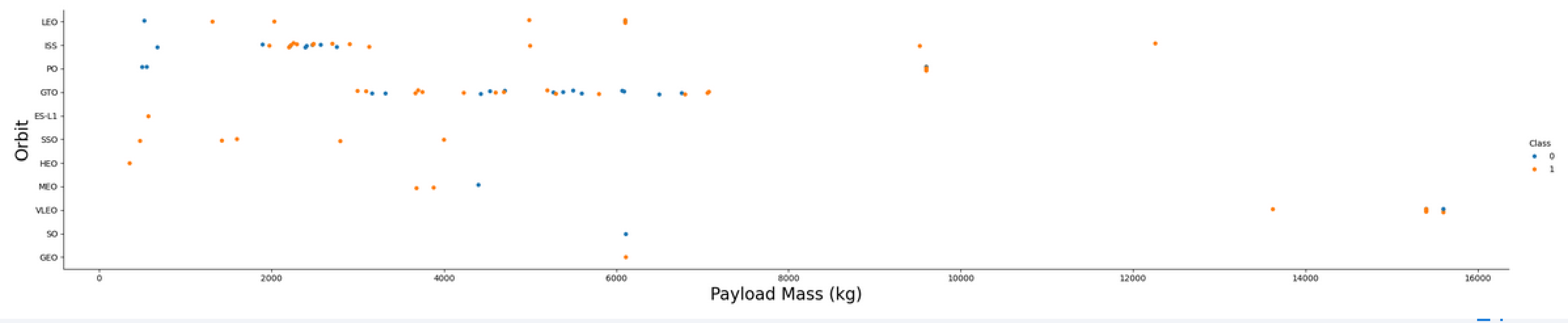
Flight Number vs. Orbit Type

- **Certain orbits were only exploited or accesible after a certain number of flights**. This might be due to technological limitations or strategical decisions. For example, Very Low Earth Orbits became popular and successful since flight number ~62
- **GTO geosynchronous and Low Earth Orbits (LEO) were very frequent at the beginning**, probably aiming for communication services like Starlink. Also, the ISS to service the station in agreement with NASA.



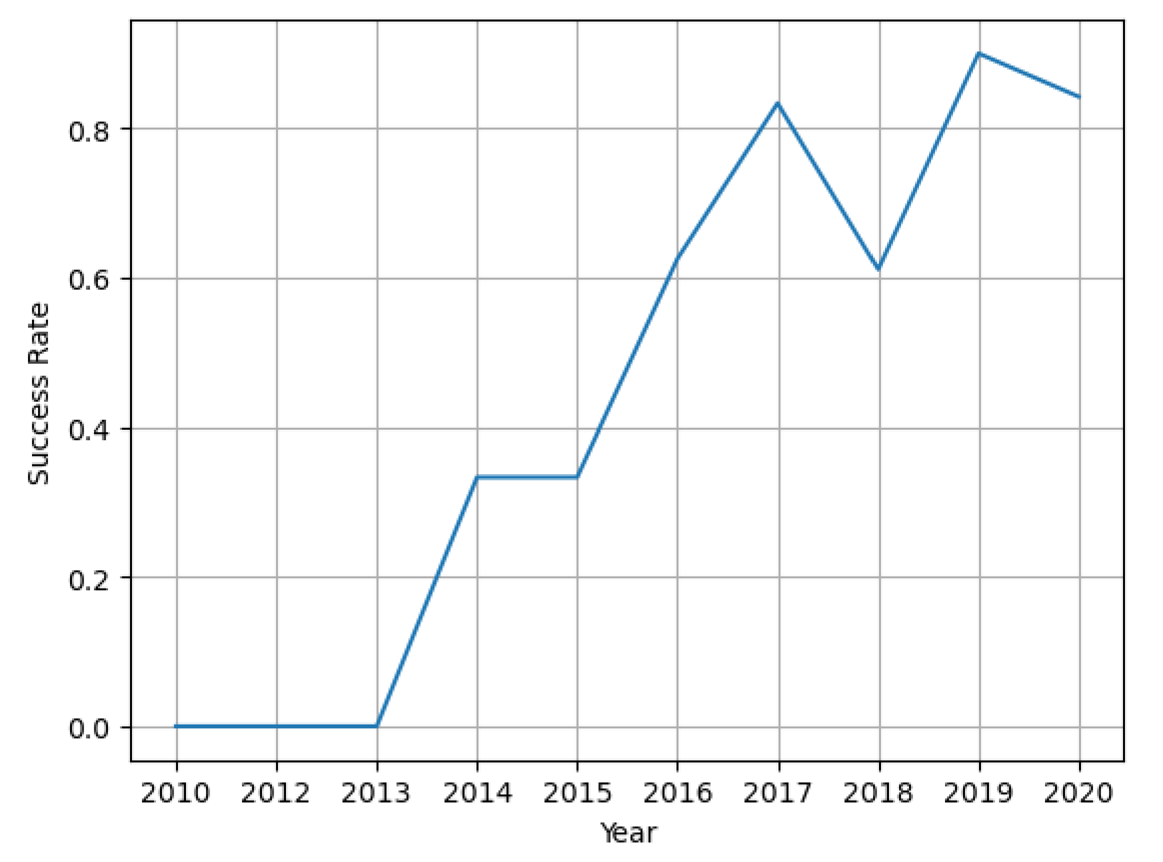
Payload vs. Orbit Type

- Large Payloads (masses beyond 10000kg) are not frequently launched and only to specific orbits
- A Payload Mass vs Flight Number/Date chart would help understand certain trends better (i.e. mass limitations on previous Booster versions)



Launch Success Yearly Trend

- Success rate was initially zero until 2013.
- Success rate **since 2013 kept increasing till 2017** (stable in 2014) probably because of the lessons learned the previous years.
- It dropped in 2018, maybe due to new developments, and **raised again in 2019, remaining very high (>80%) thereafter.**



All Launch Site Names

- There are 4 different Launch Sites:

%sql select distinct Launch_Site from SPACEXTBL limit 100

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

Launch Site Names Begin with 'CCA'

- We get the required 5 records with all its associated data with the following statement:

*%sql 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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- The total payload carried by boosters from NASA was 45596 kg

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer ='NASA (CRS)’
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was on **May 22nd, 2012**

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome not like '%Failure%’
```

min(Date)
2012-05-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are 4 boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000. All of them type “F9 FT” as seen below:

```
%sql select distinct Booster_Version from SPACEXTBL where Landing_Outcome = 'Success
(drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000 limit 10
```

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

Total Number of Successful and Failure Mission Outcomes

- Below you can see the total number of successful and failure mission outcomes.

```
%sql select Mission_Outcome,count(*) from SPACEXTBL group by Mission_Outcome
```

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

Boosters Carried Maximum Payload

- Booster **F9 B5 B1048.4** have carried the maximum payload mass of **15600kg**

```
%sql select Booster_Version,max(PAYLOAD_MASS__KG_) from SPACEXTBL where
PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL )
```

Booster_Version	max(PAYLOAD_MASS__KG_)
F9 B5 B1048.4	15600

2015 Launch Records

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

```
%sql select substr(Date, 6,2) as Month ,Booster_Version, Launch_Site from SPACEXTBL
where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'
```

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

- Ranking of landing outcomes between the date 2010-06-04 and 2017-03-20

```
%sql select Landing_Outcome,count(Landing_Outcome) as n from SPACEXTBL where Date
between '2010-06-04' and '2017-03-20'  group by Landing_Outcome order by n desc
```

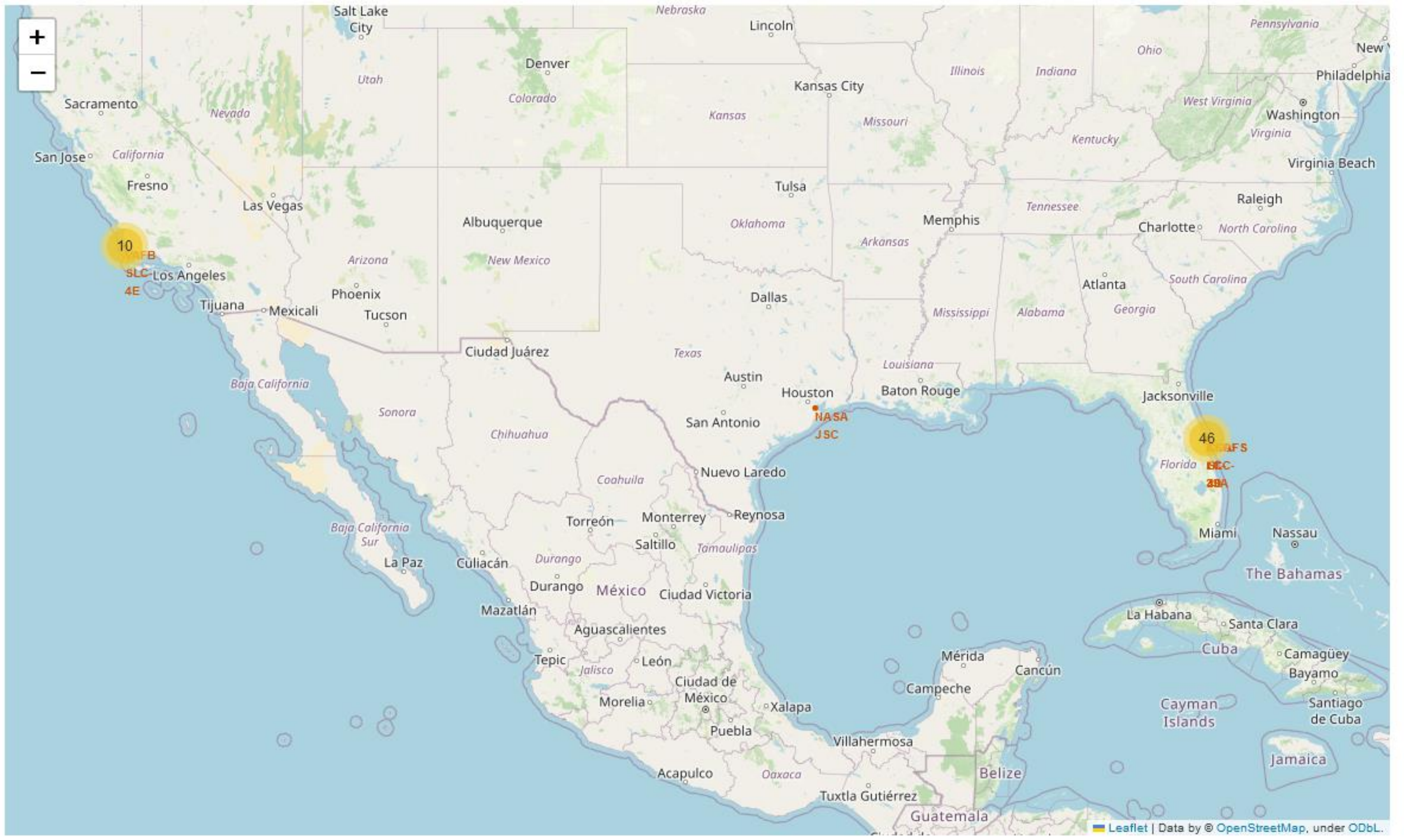
Landing_Outcome	**n**
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

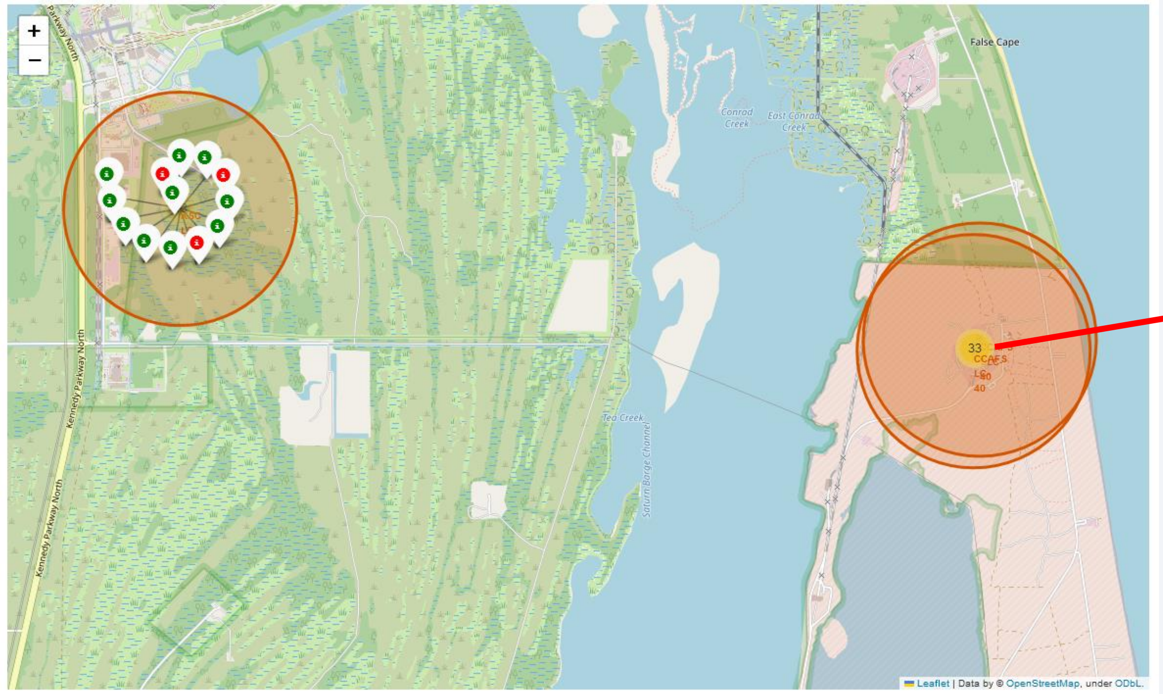
Launch Sites Locations

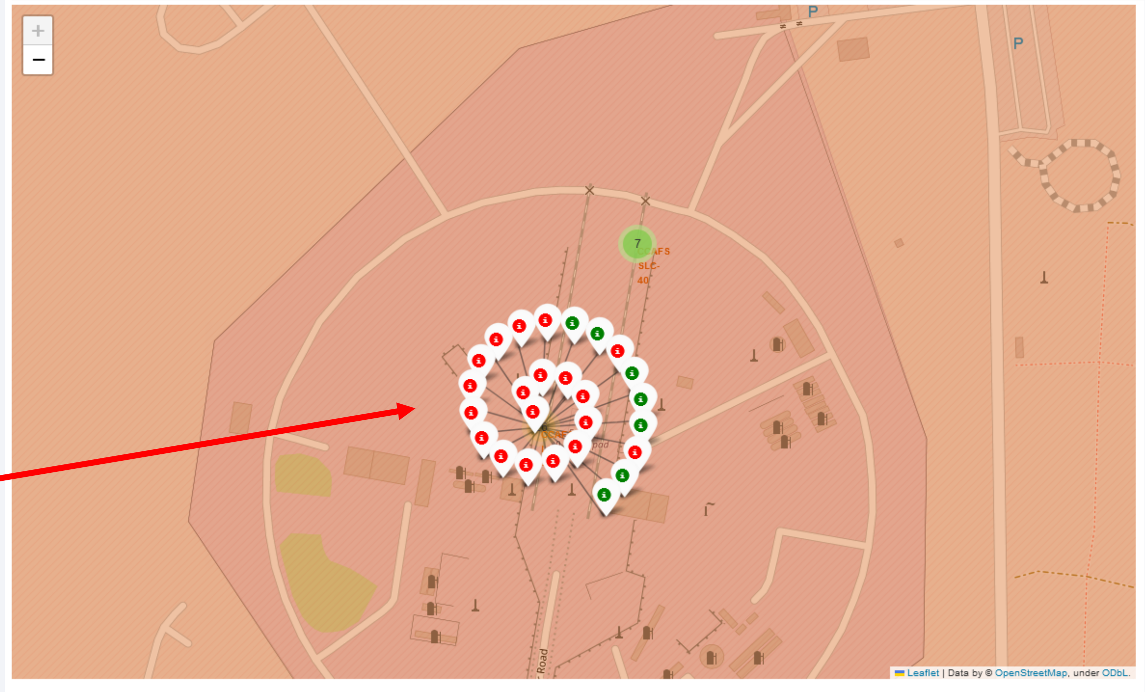
- Launch Sites are shown and are located near coastline in Florida and California
- They provide alternative launch windows within the USA.
- NASA Headquarters are located in Houston



Success/Failed launches for each site

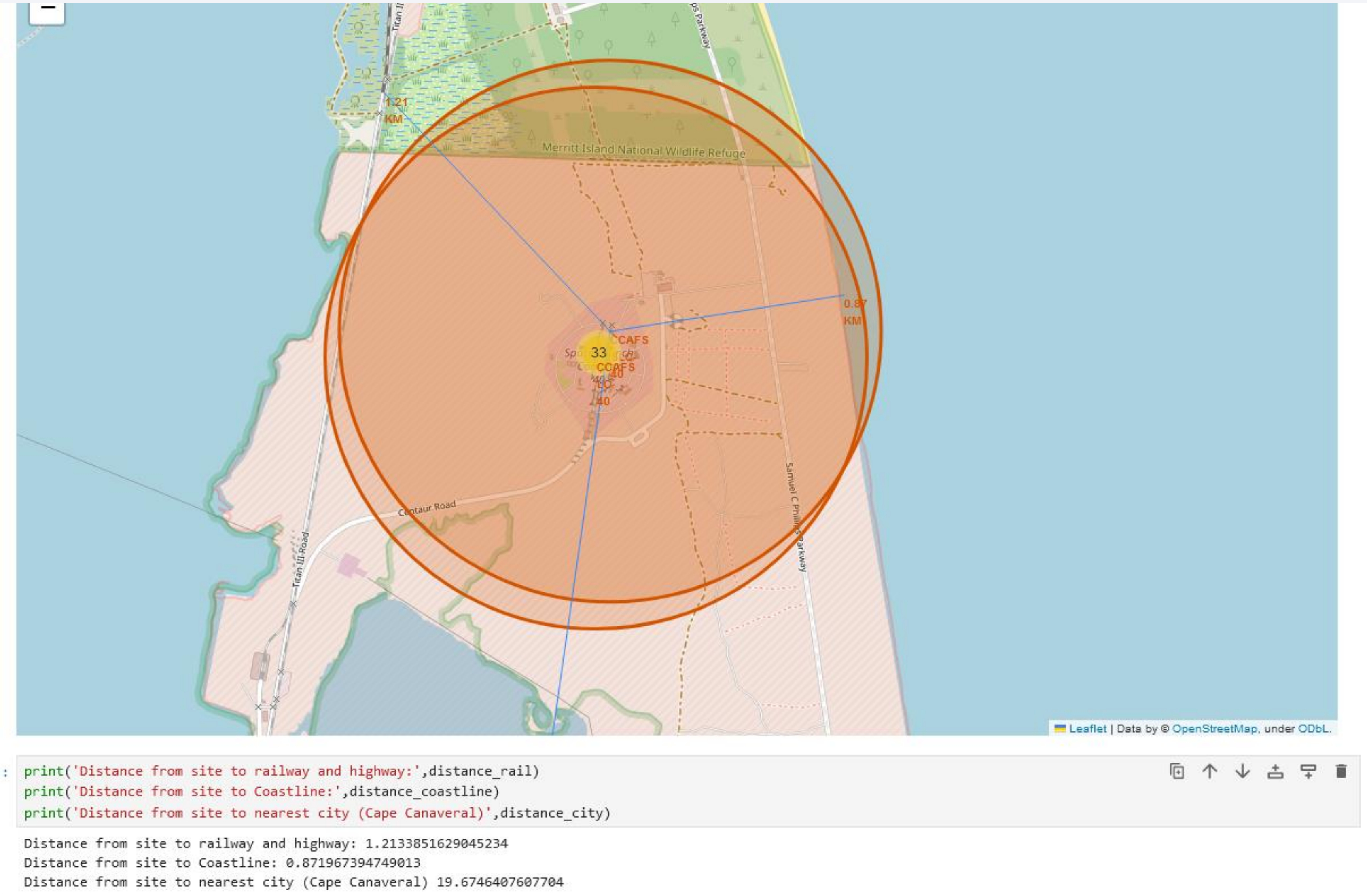
- Two nearby sites in Florida are shown with their total number of successful (green) and failed (red) launches
  - KSC LC-39A has the best rate
  - CCAFS SLC-40 has the wors rate





Distances between a launch site to its proximities

- Distance markers and lines are shown from selected site to nearest:
  - Highway and Railway
  - Coastline
  - City (Cape Canaveral)
- Launch sites are within 1 km of highways and railways because of logistics aspects.
- For safety considerations, they are close to the coast (~1.2km) and far enough from cities (~20km).

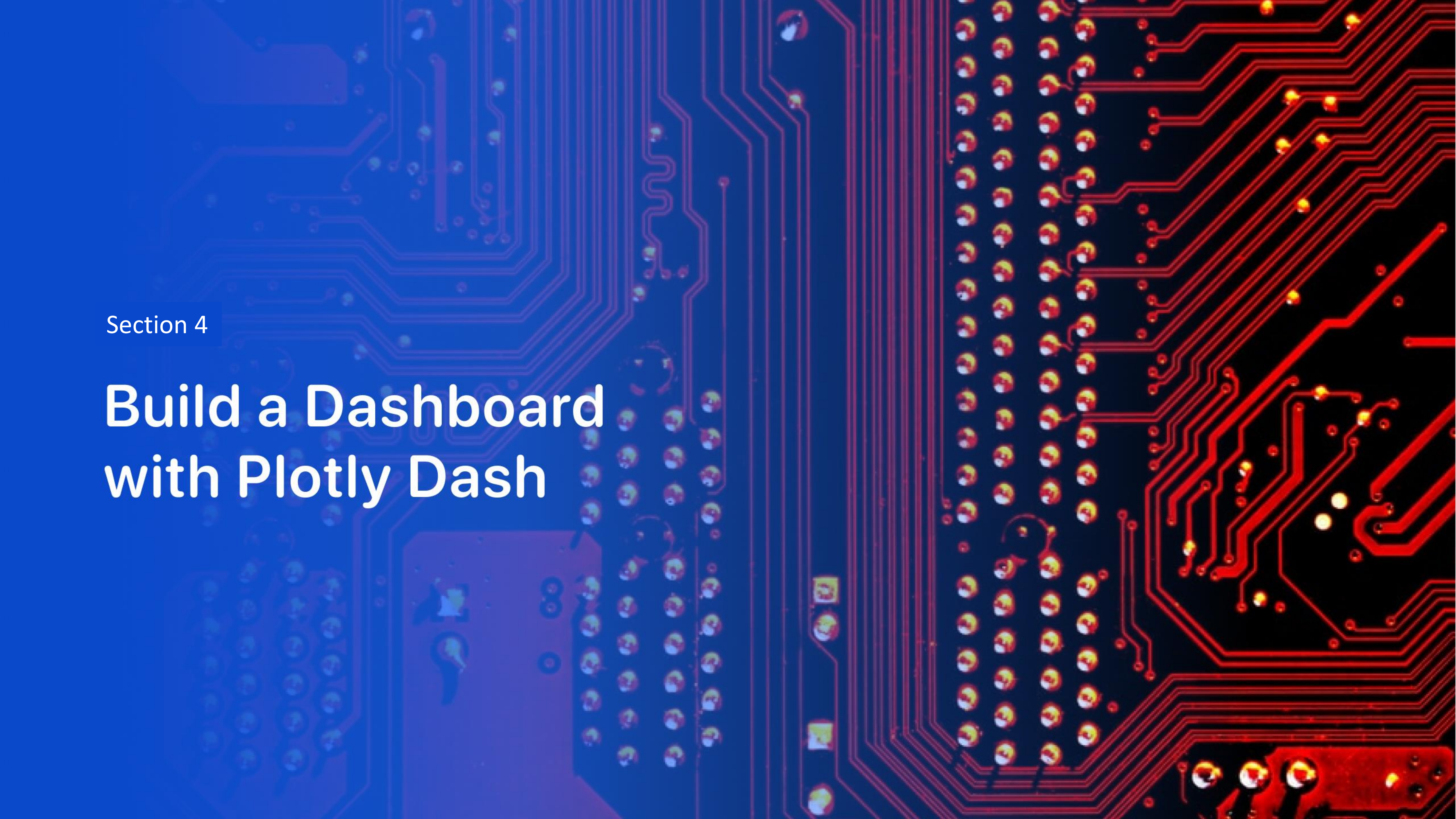

1.21 KM
Merritt Island National Wildlife Refuge
0.87 KM
CCAFS
33
CCAFS
Centaur Road
Titan III Road
Samuel C Phillips Parkway
Leaflet | Data by © OpenStreetMap, under ODbL.

```
print('Distance from site to railway and highway:',distance_rail)
print('Distance from site to Coastline:',distance_coastline)
print('Distance from site to nearest city (Cape Canaveral)',distance_city)
```

```
Distance from site to railway and highway: 1.2133851629045234
Distance from site to Coastline: 0.871967394749013
Distance from site to nearest city (Cape Canaveral) 19.6746407607704
```

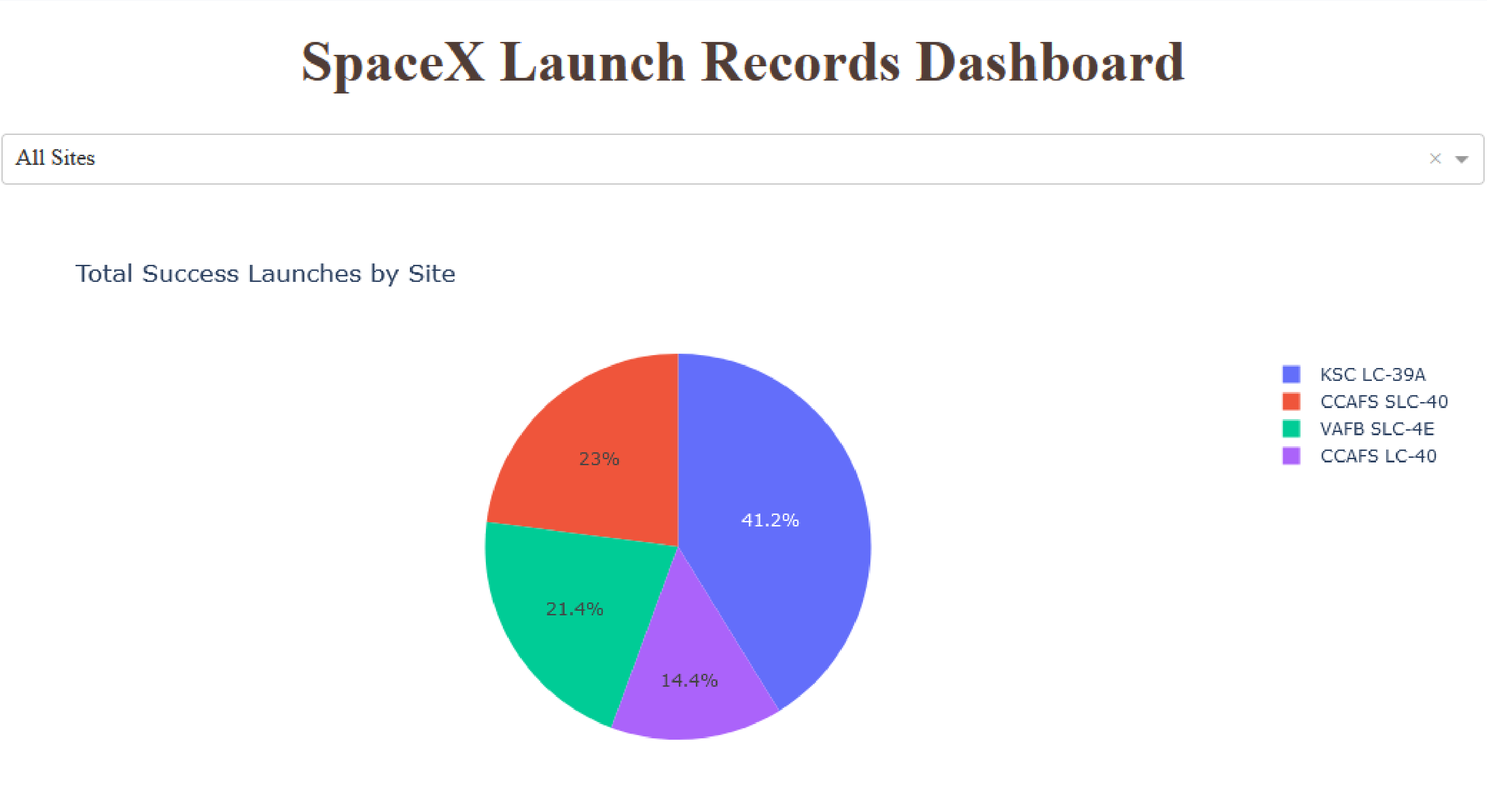
Section 4

Build a Dashboard with Plotly Dash



Total Success Launches by Site

- “KSC LC-39A” is the site with more successful launches overall



SpaceX Launch Records Dashboard

All Sites

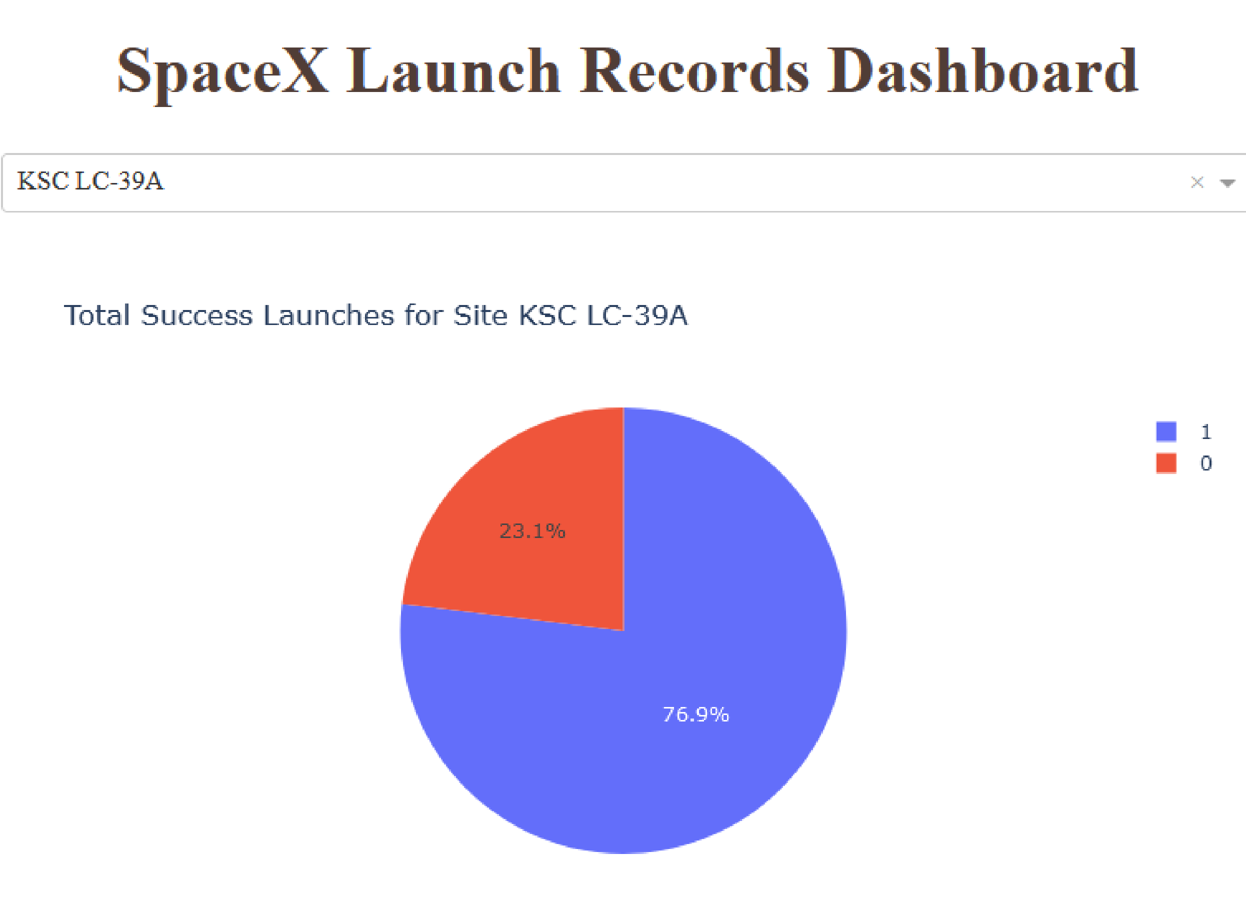
Disclaimer:

It is important to note that this does not represent the success rate for each individual site, as this is not normalized by the number of launches in each site.

See next slide for clarification.

Sites with the best Success Rate

- “KSC LC-39A” is not only the site with more successful launches overall, but also the one with the highest success rate with ~77% positive outcomes.

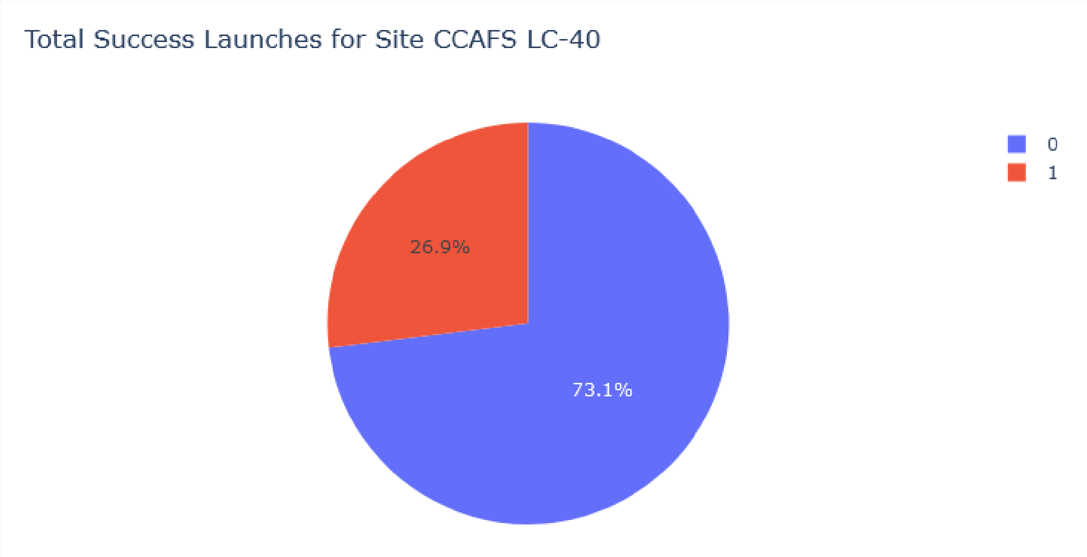


SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for Site KSC LC-39A

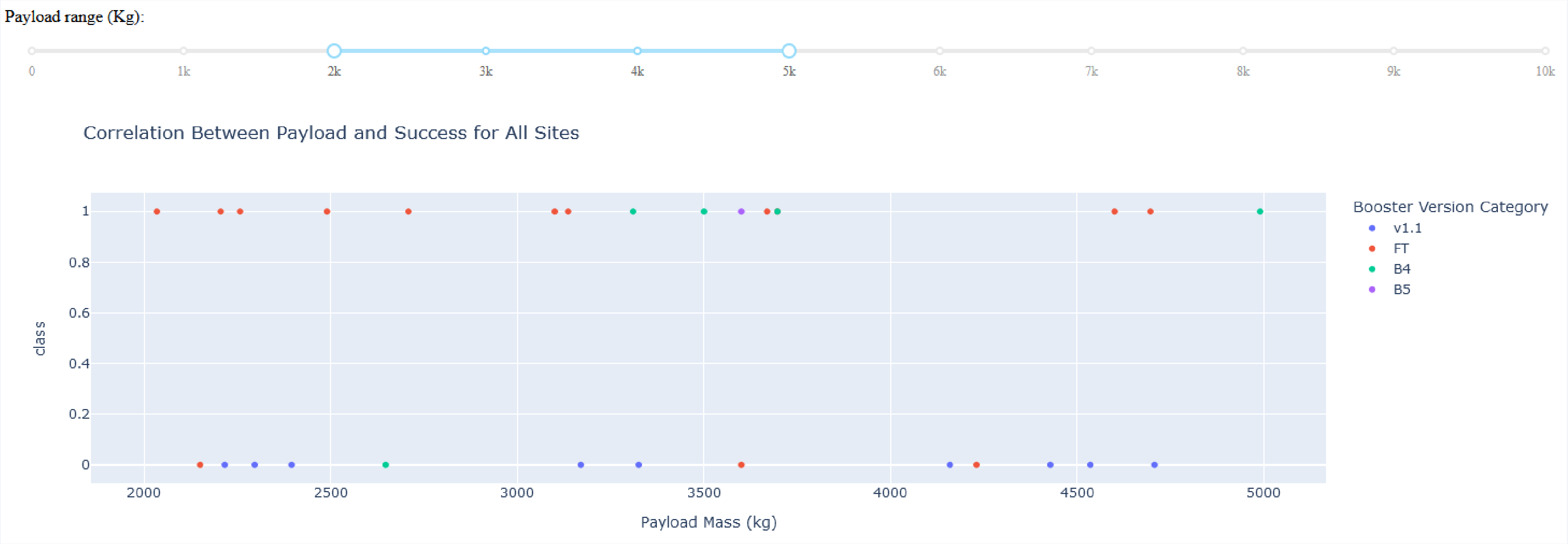
“CCAFS LC-40” had the lowest overall number of successes but it is due to a low number of attempts, and its rate is very high (though statistically not significant).



Total Success Launches for Site CCAFS LC-40

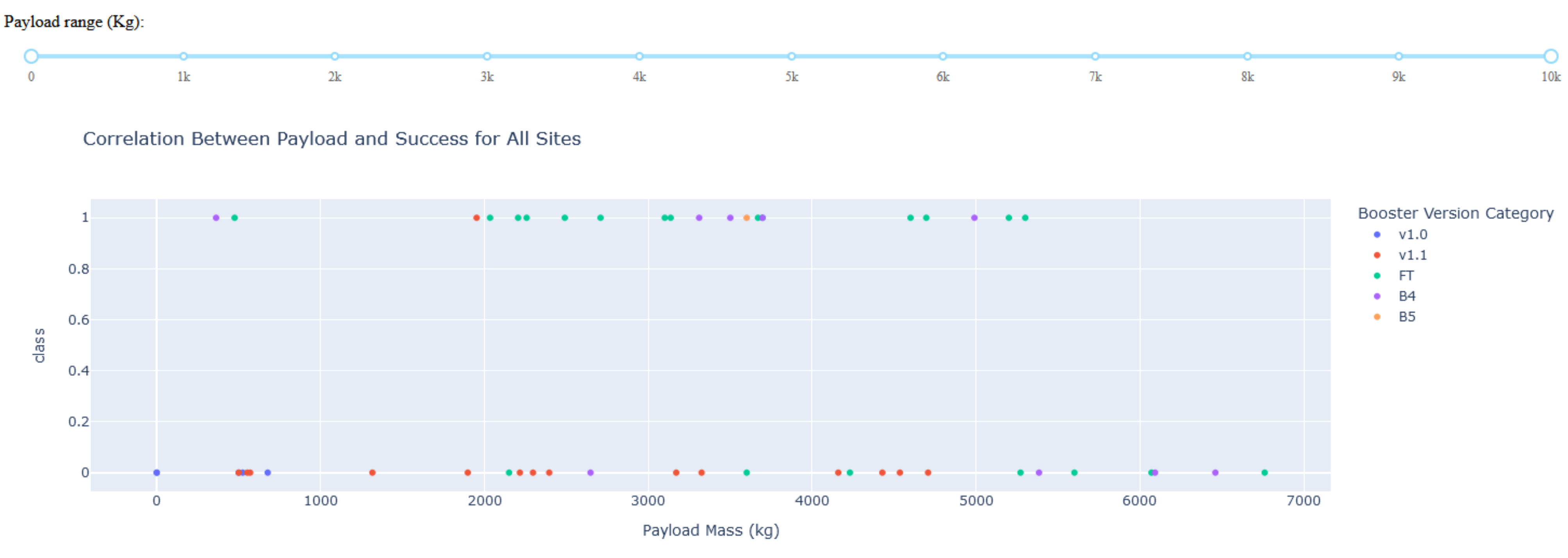
Correlation between Payload/Booster and Success (1)

- We can see that for Payloads between **2000 and 5000 kg**, **Booster category FT was very successful**



Correlation between Payload/Booster and Success (2)

- Overall, Booster version 1.1 is the least successful one.
- For payloads larger than 5500 kg, all attempts had failed

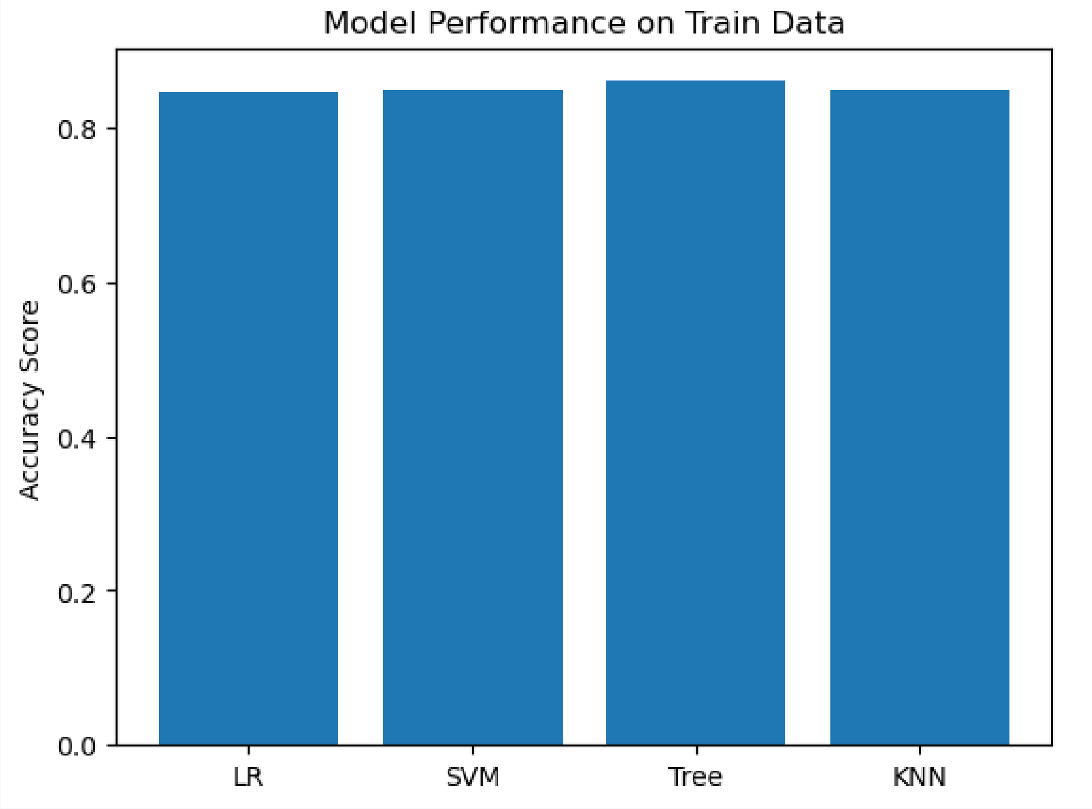


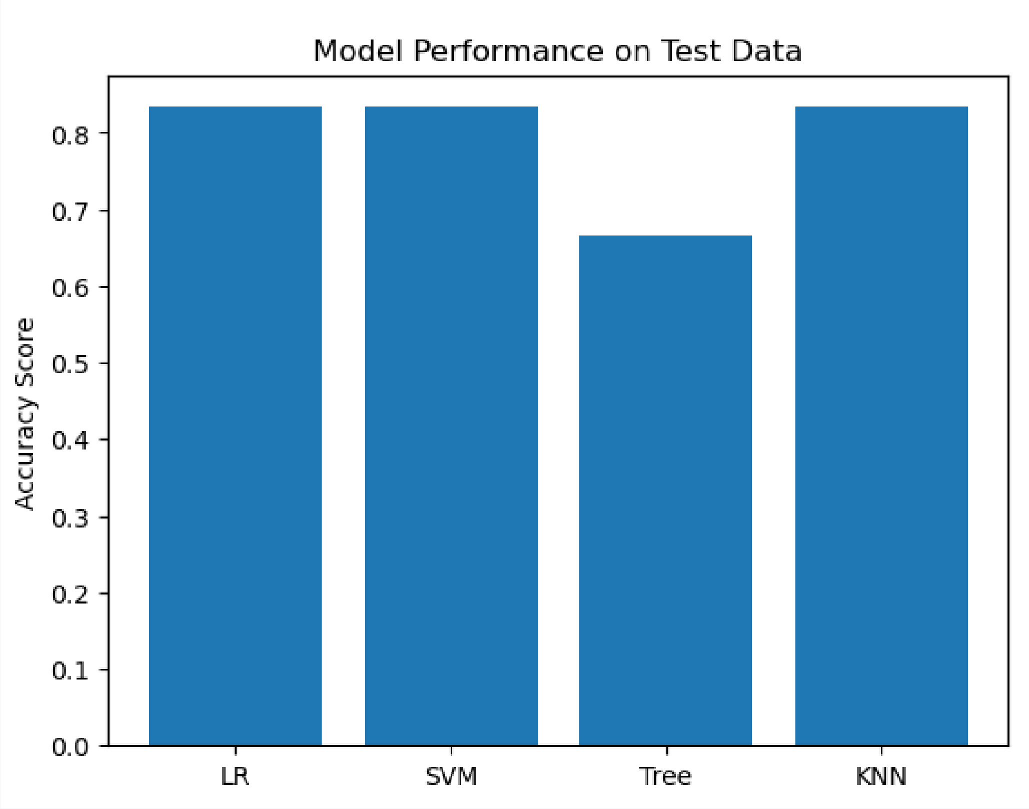
Section 5

Predictive Analysis (Classification)

Classification Accuracy

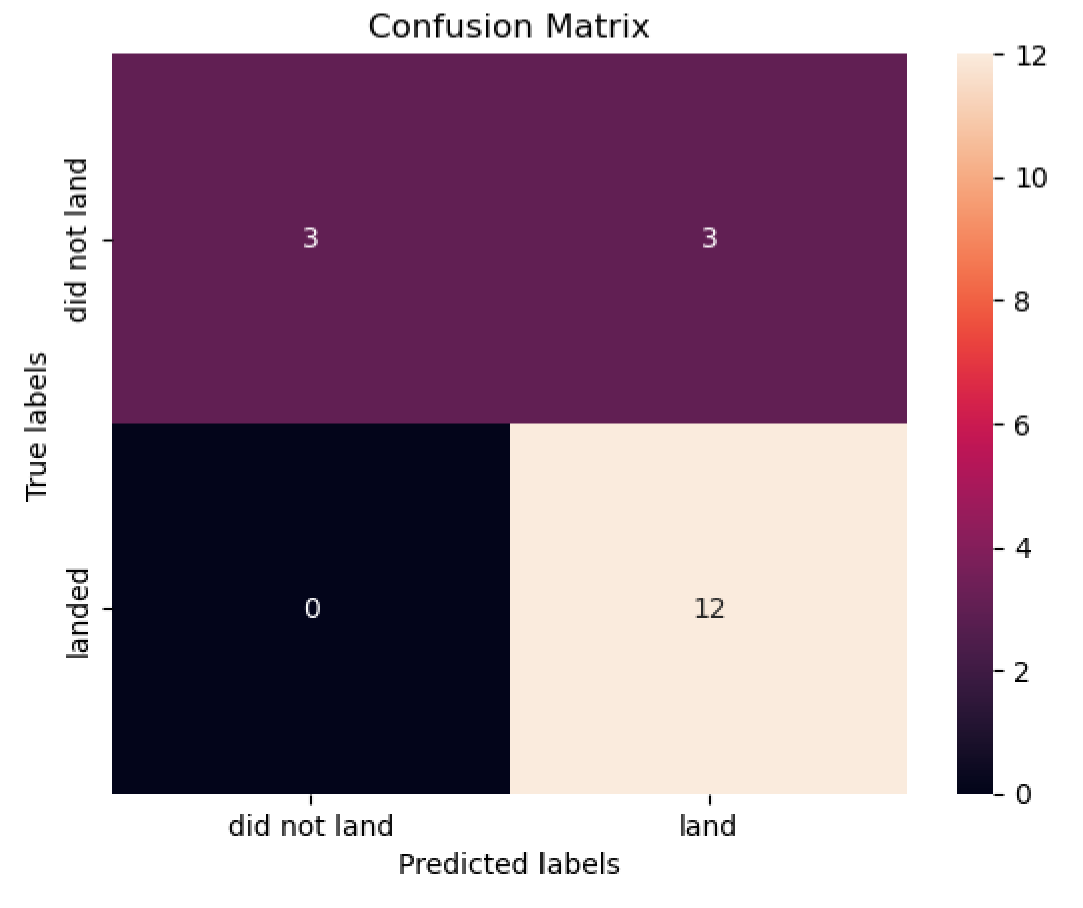
- Classification Tree suffers from overfitting with the extended parameters tested.
- All other models perform roughly the same (85% accuracy)





Confusion Matrix

- **Our prediction models** based on classification algorithms **performed well with an 85% accuracy**.
- However, there is a **risk of obtaining False Positives** on untrained data, that is, predicting a positive outcome when it will actually fail.
- Further analysis and modelling is recommended to minimize this issue.



Conclusions

- **Success Rate has a strong correlation with flight number**. This is a consequence of the **steep learning curve required in this field**, literally "rocket science".
  - This is a **very important conclusion that must be conveyed to Space Y management team and stakeholders**.
  - Strategic plans must take this into account to be able to compete despite our late arrival to the market.
- The selection of our headquarters and launch and landing sites is also crucial, not only for logistics and safety considerations, but also because they affect the success rate. **Studying the specific characteristics of site KSC LC-39A could yield important factors to understand its high success rate**.
- **Booster category FT** carrying **payloads up to 5000kg maximum** has the highest success rate of all.
- Our **classification model can predict the outcome of a mission with an 85% accuracy**.
  - Some false positives predictions can happen.
  - Additional investigation, including new data and modelling techniques is recommended to update and improve the quality of our models.

Appendix

- Source code, and this presentation/report, is available here:
    - https://github.com/uvenabla/public_IBM_DataScience_Certificate

Thank you!