Winning Space Race with Data Science

Reinhardt Muehlhaeusser
May 2024

Outline

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

Executive Summary

- SpaceX advertises Falcon 9 rocket launches at $62 million. Other providers' rocket launches cost upwards of $165 million.
- SpaceX's cost savings are mainly due to reusing the first stage.
- The project’s objective is to derive a model that predicts if the Falcon 9 first stage will land successfully.

Executive Summary - Methodology

- **Data understanding and collection**
  - Identify and gather relevant data sources for SpaceX launches. Ensure the foundation of the analysis is based on accurate and complete data.
- **Data preparation and wrangling**
  - Clean and preprocess the data to ensure quality and usability for analysis..
- **Exploratory data analysis (EDA) and Interactive visual analytics**
  - Gain insights and understand data patterns, distributions and relationships.
  - Enable dynamic interaction with data visualizations to foster deeper insights.
- **Predictive analysis using classification models**
  - Develop models to predict future Falcon 9 landing outcomes

Executive Summary - Results

- The available data shows that the success rate for the last year of observations reached a range of more than 80% which explains the cost savings through reuse of the Falcon 9 first stage.
- We can see a clear positive trend in the success rate of first stage landings for the Falcon 9 over the years. Technological advancements like Grid Fins or Landing Legs paired with growing experience in operations support the trend and indicate sustained future cost advantages for SpaceX.
- With the derived Decision Tree model, we can predict future landing outcomes with an accuracy of more than 80%.
- The model should be continuously updated to take future outcomes and changes in technology into consideration.

Introduction

- Project background and context
- Questions to be addressed

Project background and Context

- SpaceX manufactures and operates space rockets and aims for reusability of the first stage of the rocket as competitive advantage.
- SpaceX advertises Falcon 9 rocket launches at $62 million. Other providers' rocket launches cost upwards of $165 million.
- SpaceX's cost savings are mainly due to reusing the first stage.
- Determining and understanding the influence factors for the first stage landing success can help determine launch costs. This information could aid alternate companies in bidding against SpaceX.
- The project’s objective is to derive a model that predicts if the Falcon 9 first stage will land successfully.

Questions to be addressed

- Can we see a success rate of landings for the Falcon 9 rocket Stage 1 that allows reuse and explains the cost advantage for SpaceX?
- How did the success rate of landings develop over time and is a trend visible that indicates sustained future cost advantages for SpaceX?
- What accuracy can we deliver with a ML model to predict the success of future landings of Falcon 9 first stage after launch?

Section 1

Methodology

Methodology

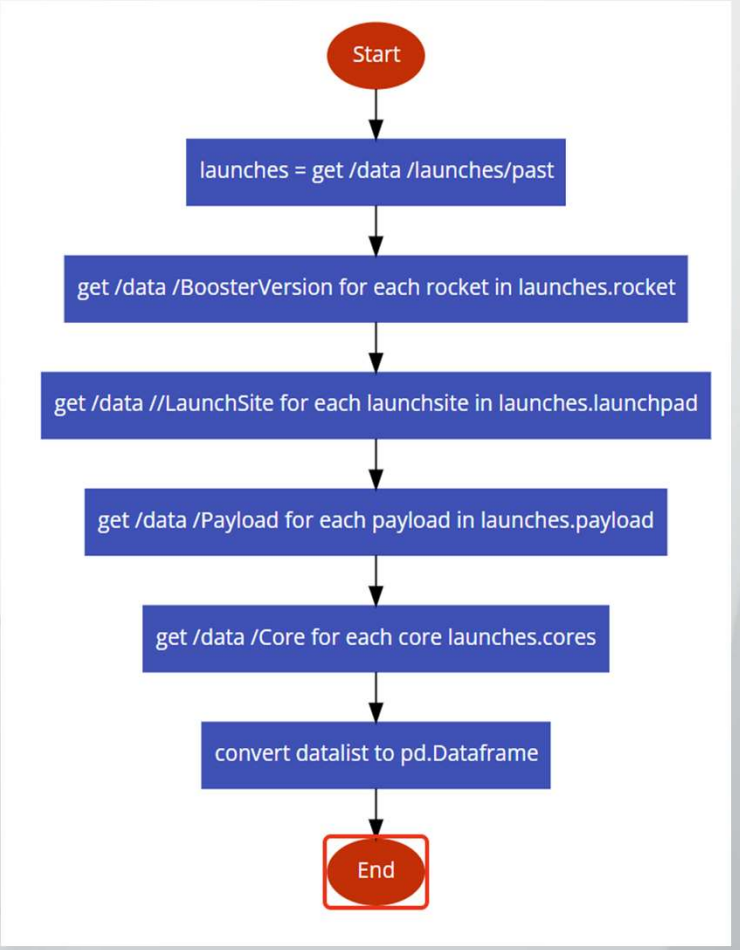
- Data understanding and collection
- Data preparation and wrangling
- Exploratory data analysis (EDA)
- Interactive visual analytics
- Predictive analysis using classification models
- Presentation of Results

Data understanding and collection

- We used two primary data sources:
  - The open available SpaceX API
  - Wikipedia as trusted open information source via Web Scraping
- The SpaceX API offers access to historic data for rocket launches and landings since 2010. It is implemented as REST web services that can be accessed with an URL. The successful response contains the data in JSON format.
- Wikipedia holds verified additional information for the SpaceX launches that can be extracted by Web scraping and joined with the SpaceX API data.

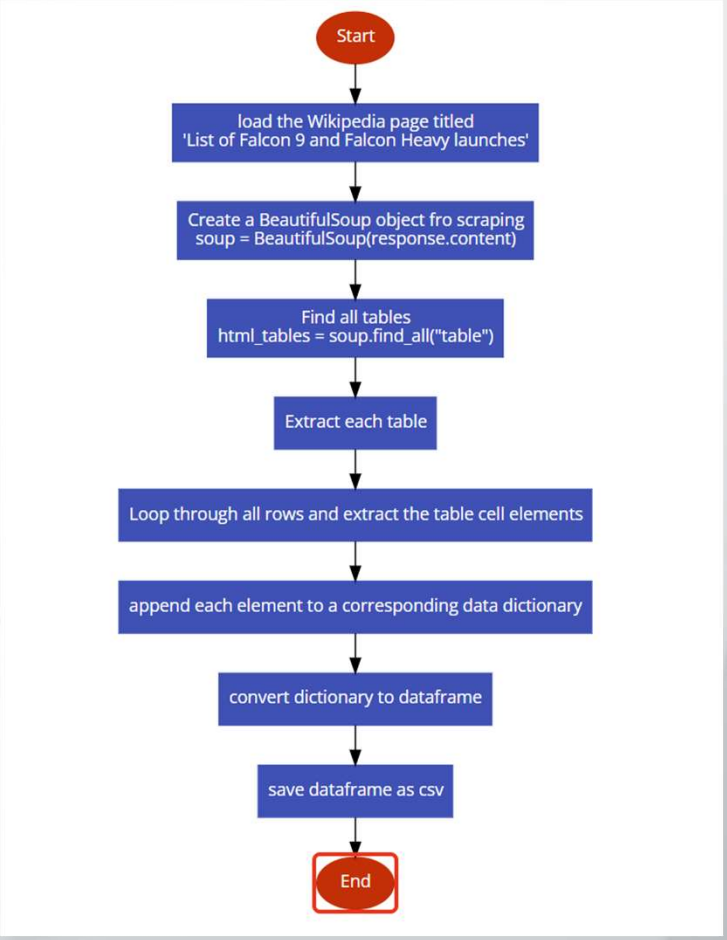
Data collection – SpaceX API

- The SpaceX API can be accessed by the base URL https://api.spacexdata.com/v4/
- There are different datasets available. For our purpose we used the following five:
  - /launches/ to get the list of all past launches
  - /rockets/ for the booster name
  - /launchpad/ for the launch site information
  - /payload/ for mass of payload and orbit
  - /cores/ for outcome and type of the landing, number of flights with that core, gridfins use, whether the core is reused, whether legs or landing pad were used, the block version of cores, number of reuse, and serial of the core.
- The Jupyter Notebook to extract the data can be found here: https://github.com/rmuehlhaeusser/CourseIBMProDataScience/blob/main/Capstone/jupyter-labs-spacex-data-collection-api.ipynb



Data collection – Web Scraping

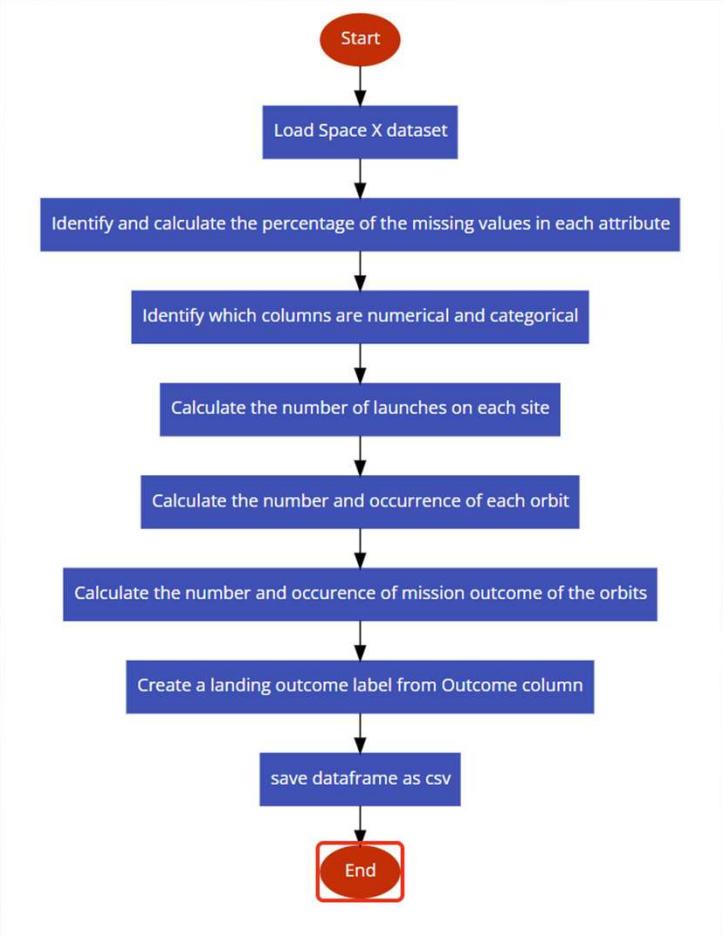
- Use of Web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled `List of Falcon 9 and Falcon Heavy launches`
- Web scrap Falcon 9 launch records with `BeautifulSoup`:
  - Extract a Falcon 9 launch records HTML table from Wikipedia
  - Parse the table and convert it into a Pandas data frame
- The Jupyter Notebook to extract the data can be found here: https://github.com/rmuehlhaeusser/CourseIBMProDataScience/blob/main/Capstone/jupyter-labs-webscraping.ipynb



Data preparation and wrangling

- Data wrangling was performed mainly for the SpaceX API data set. The steps are illustrated in the flowchart and included:
  - Identify and calculate the percentage of the missing values in each attribute
  - Identify which columns are numerical and categorical
  - Data aggregation and value counts
  - Creation of label to describe success or failure of the Falcong first stage landing

- The Jupyter Notebook can be found here https://github.com/rmuehlhaeusser/CourseIBMProDataScience/blob/main/Capstone/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- The following charts were prepared to explore the data and its relationships or dependencies
  - Payload Mass over the continuous Flightnumber at different dates as Scatterplot with indication of Success or Failure
  - Launch Site over the continuous Flightnumber at different dates as Scatterplot with indication of Success or Failure
  - Launch Site over Payload Mass as Scatterplot with indication of Success or Failure
  - Success Rate per Orbit type as Barplot
  - Orbit over the continuous Flightnumber at different dates as Scatterplot with indication of Success or Failure
  - Orbit over Payload Mass as Scatterplot with indication of Success or Failure
  - Success Rate for each year as Lineplot to visualize a trend
- The Jupyter Notebook can be found here:
  https://github.com/rmuehlhaeusser/CourseIBMProDataScience/blob/main/Capstone/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The following SQL queries were performed to explore the data and its relationships or dependencies further
  - Query the names of launch sites in the space mission
  - Query the average payload for Falcon9 v1.1
  - List the date of the first successful landing
  - List the Booster Versions with Successful landing on a drone ship
  - List the total number of of success and failures
  - List the Booster Versions that carried the maximum payload
  - List month, Launch Site and Booster Version for failed landings for 2015 on drone ships
  - Rank the landing outcomes between 2010-06-04 and 2017-03-20
- The Jupyter Notebook can be found here:

  https://github.com/rmuehlhaeusser/CourseIBMProDataScience/blob/main/Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Interactive Visualization with Folium

- To explore the locations of the different launch sites we created Folium maps and included the following objects
  - Circles to mark the four Launch Sites
  - Markers in a MarkerCluster for each successful or failed landing on the different Launch Sites
  - Lines and labels to indicate the distances of a selected Launch Site to the nearest coastline, highway, railway and closest city

- The Jupyter Notebook can be found here:
  https://github.com/rmuehlhaeusser/CourseIBMProDataScience/blob/main/Capstone/lab_jupyter_launch_site_location.ipynb

Interactive Dashboard with Plotly Dash

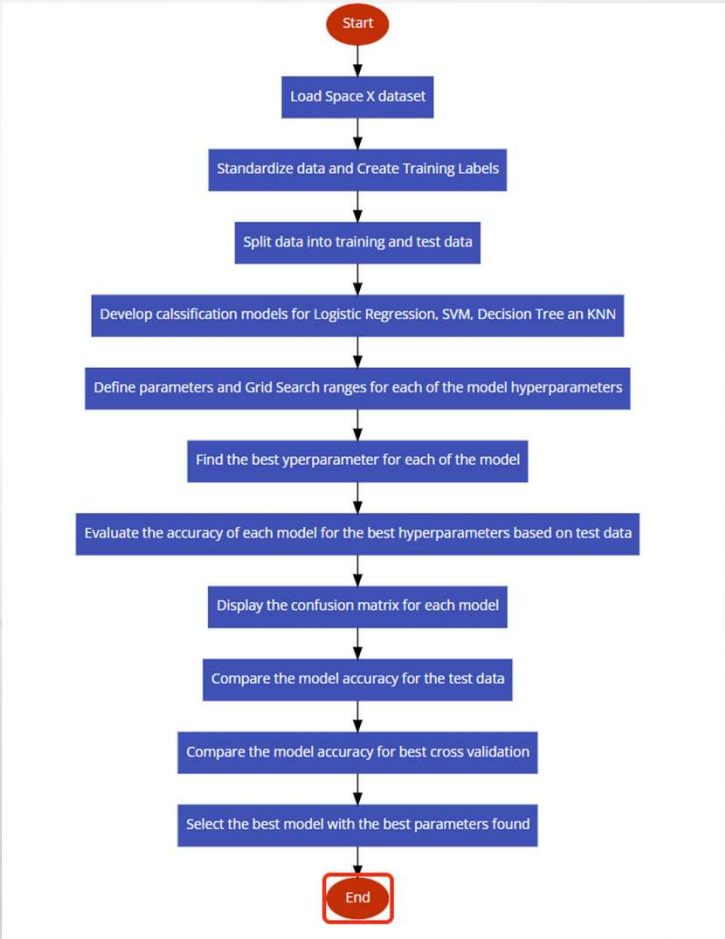
- A Dashboard with Plotly Dash has been implemented to perform interactive visual analytics on SpaceX launch data in real-time. Two Graphs are visualized:
  - A Pie Chart shows the success rates depending on the selected sites
  - A scatter plot shows the correlation between payload and success with a variation in the Booster Version Category, indicated by different colors. The underlying data can be filtered by site selection and payload range.
- The interactive dashboard filter for the data are:
  - A dropdown box to select a specific launch site or all sites
  - A range slider to define the payload range.
- The plots allow to understand the dependencies between launch site and success, as well as the impact of payload on the success for different Booster Versions.
- The App File can be found here:

  https://github.com/rmuehlhaeusser/CourseIBMProDataScience/blob/main/Capstone/spacex_dash_app.py

Predictive analysis using classification models

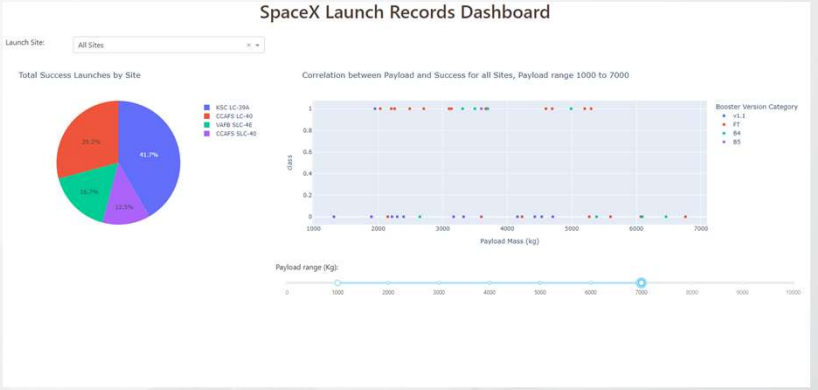
- To predictively analyze the SpaceX first stage landing success, we preprocessed the data, split the data into training and test data and developed four classification models:
  - A Logistic Regression Model,
  - a Support Vector Machine Model,
  - a Decision Tree Model and
  - a K Nearest Neighbor Model
- Each of the models were trained with the training data and tested with the test data set.
- The models were analyzed for accuracy by calculating the accuracy score and plotting the confusion matrix.
- The Jupyter Notebook can be found here:

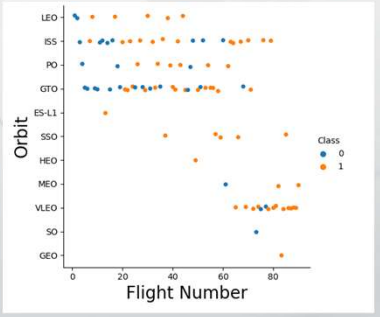
https://github.com/rmuehlhaeusser/CourseIBMProDataScience/blob/main/Capstone/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

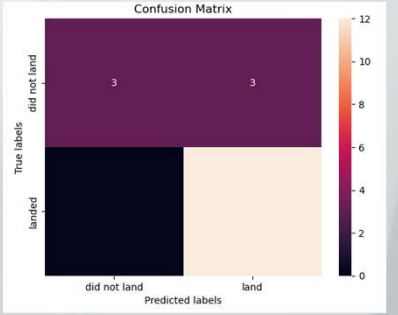


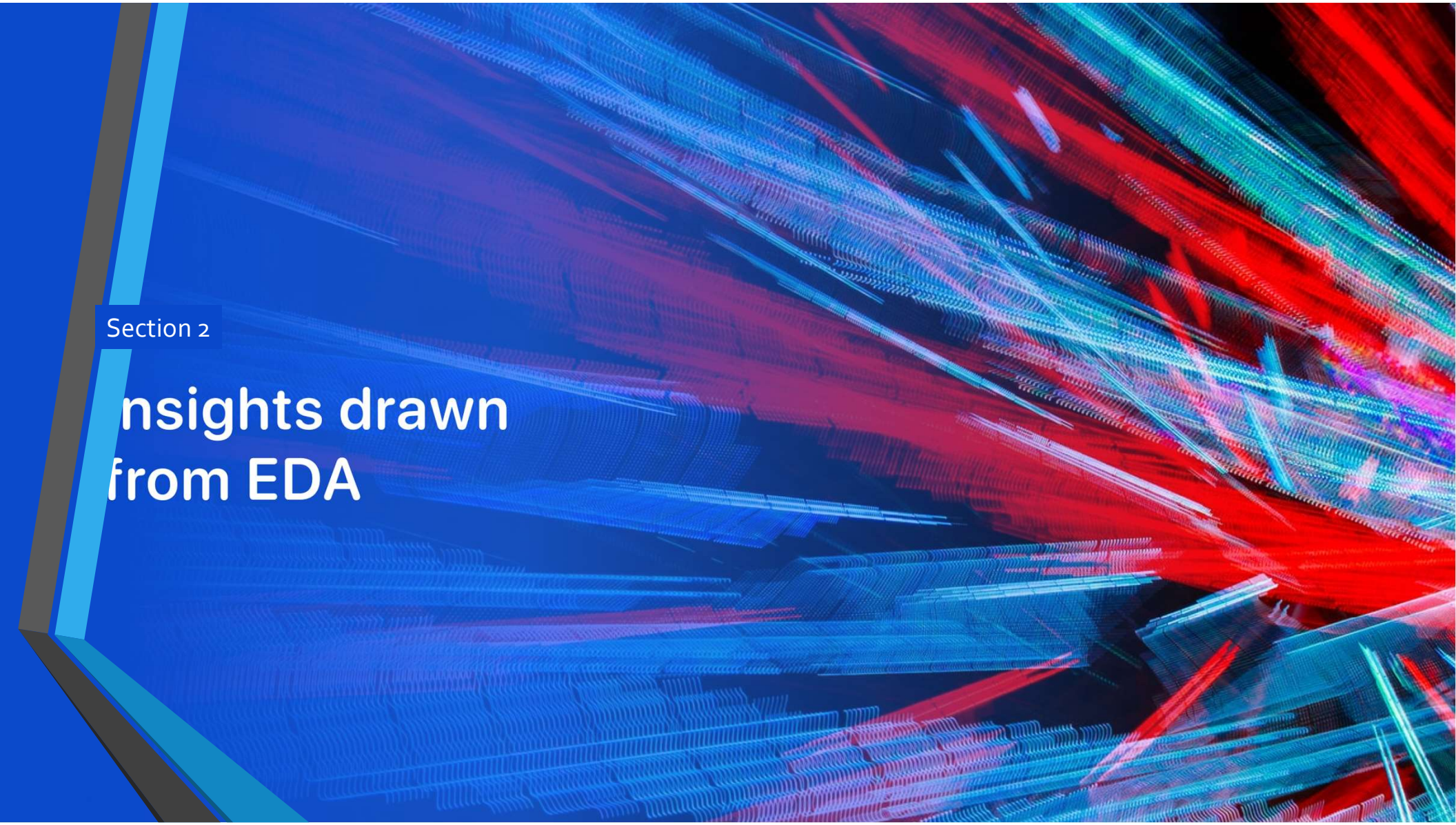
Presentation of Results

- The Results of the project will be presented in the final section and include:
  - Graphs and Plots created during Explorative Data Analysis,
  - Screenshots from the developed Interactive Plotly Dash App,
  - Accuracy tables and Confusion Matrix for the classification models

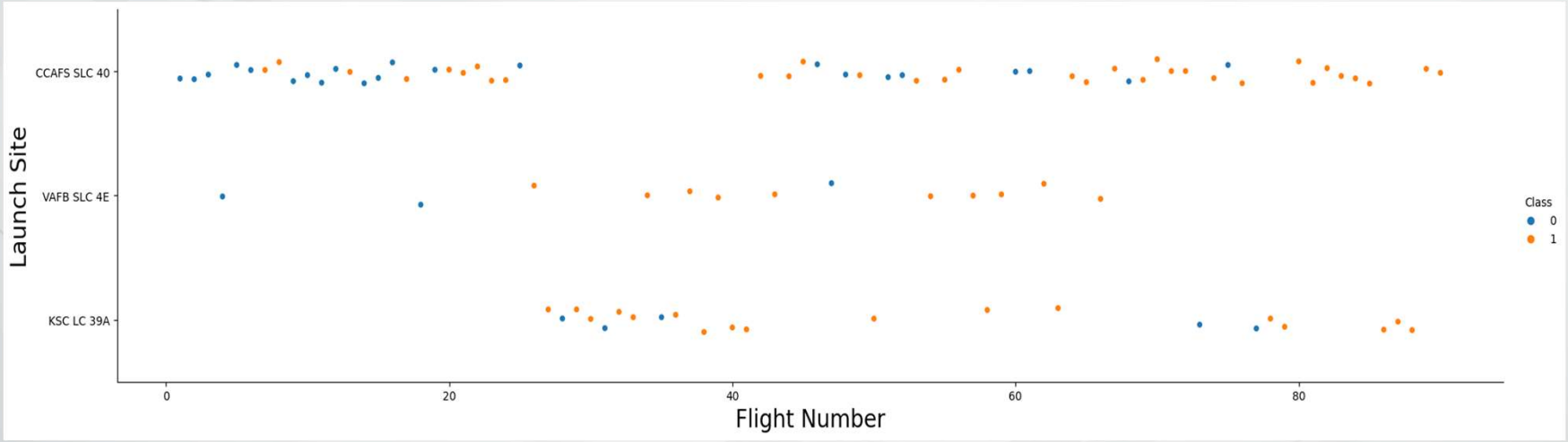






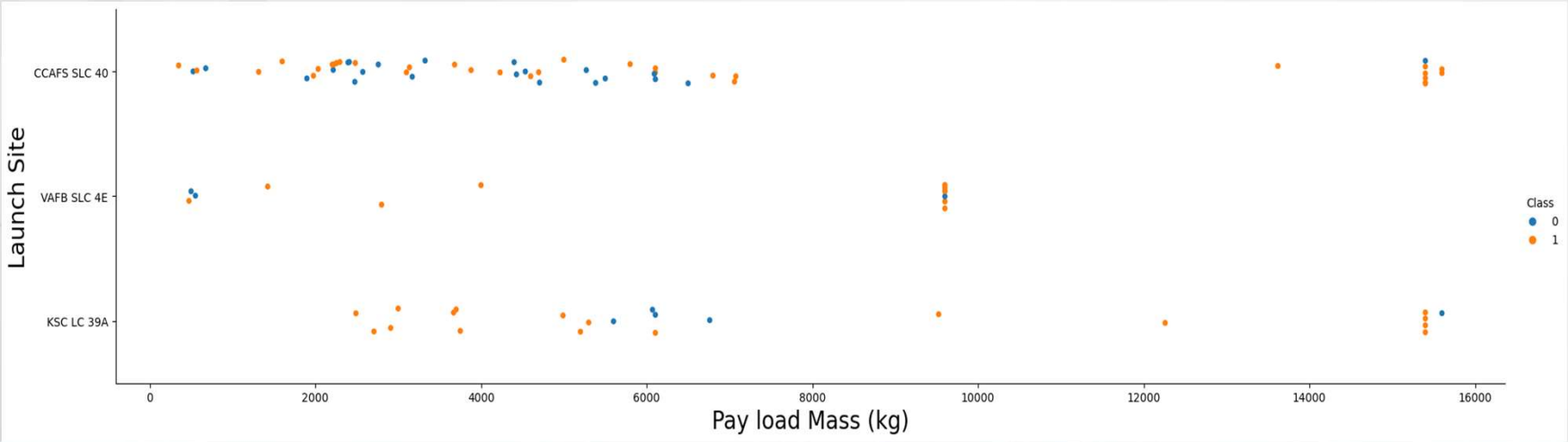

Section 2
nsights drawn
from EDA

EDA – Flight Number vs Launch Site



- No clear pattern that differentiates the different Launch Sites
- All Launch Sites show improvement in successful landing over time
- Launch Site CCAFS SLC 40 is the most frequently used Launch Site
- Launch Site VAFB SLC 4E is not as often used especially lately
- Launch Site KSC LC 39 A was introduced at a later stage

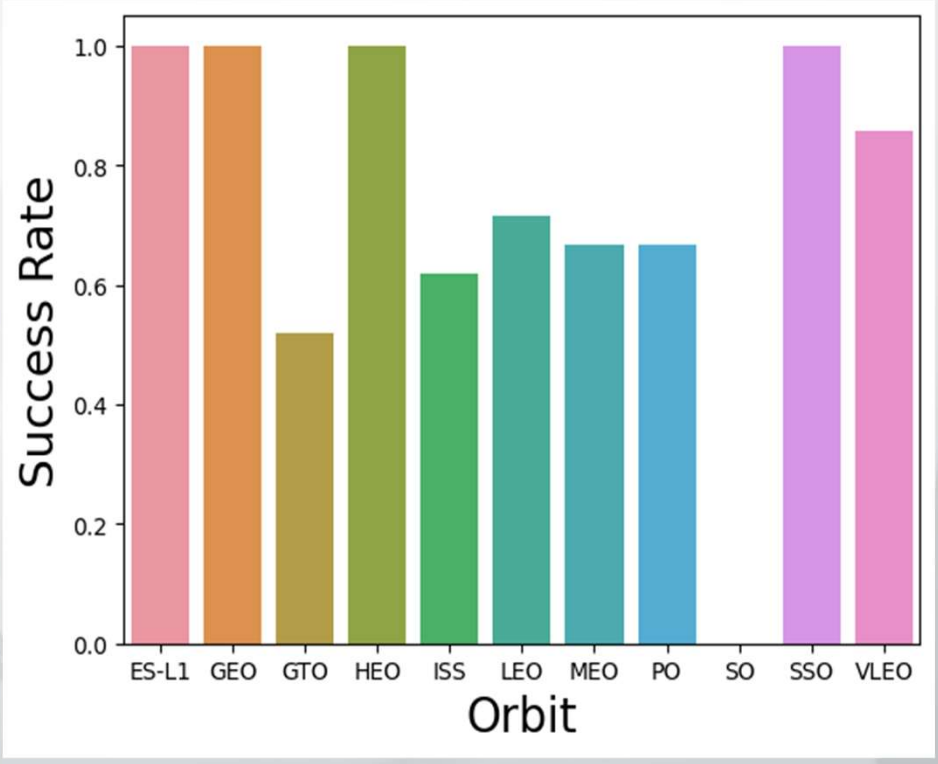
EDA – Payload vs Launch Site



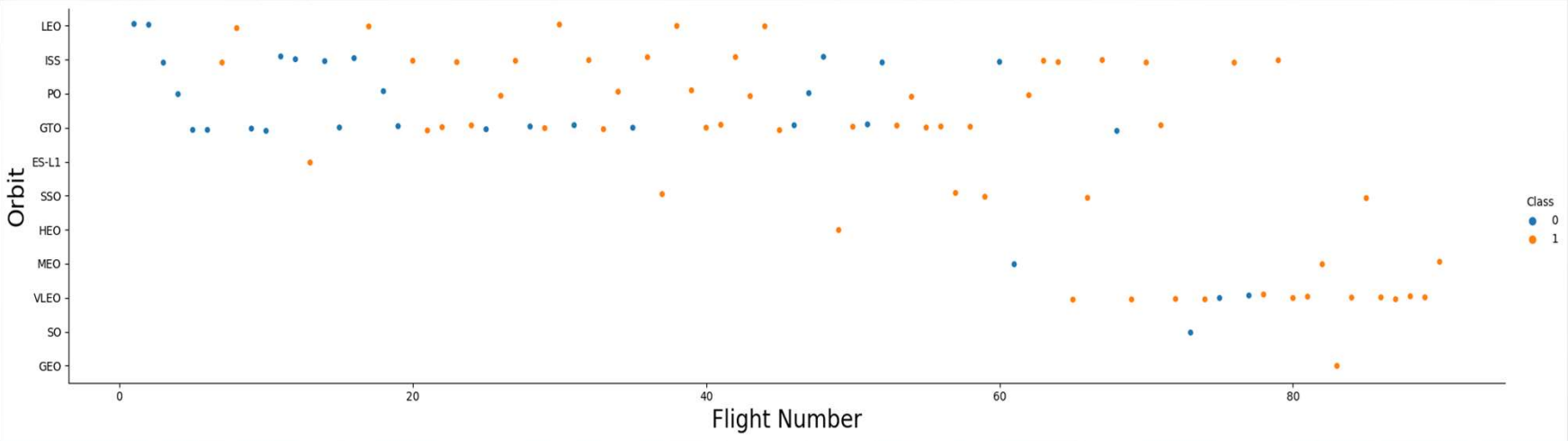
- Launch Site CCAFS SLC 40 was used for the whole range of payload and shows also success for high payloads
- Launch Site VAFB SLC 4E was not used extensively and did not launch rockets with high payload
- Launch Site KSC LC 39 A also show a wide range of payloads and successful launches for high payloads

EDA - Success Rate vs Orbit

- A barchart of the Orbits and their Success Rates for landing shows that there are some Orbit with perfect success.
- ES-L1, GEO, HEO and SSO have a 100% success rate
- VLEO is also showing better performance at 80%
- All otherOrbits are in the range of 50-70% with GTO at 50% and LEO at 70%
- The success rate does not take the number of launches into account! We used a second plot to understand the success of the Launch sites better on the next slide.

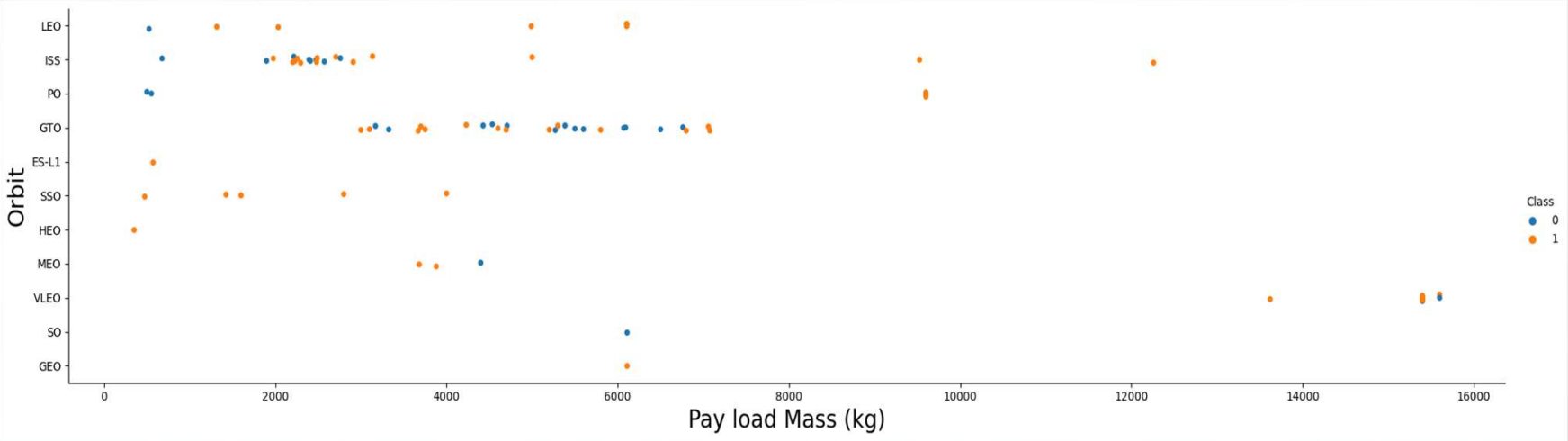


EDA – Orbit vs Flight Number



- The chart clarifies the results from the success rate
- The Orbits with 100% success were not targeted very often, some of them only once
- It looks like the VLEO orbit with its pretty high success rate was used only in the newer history, which can also explain better performance due to improved technology

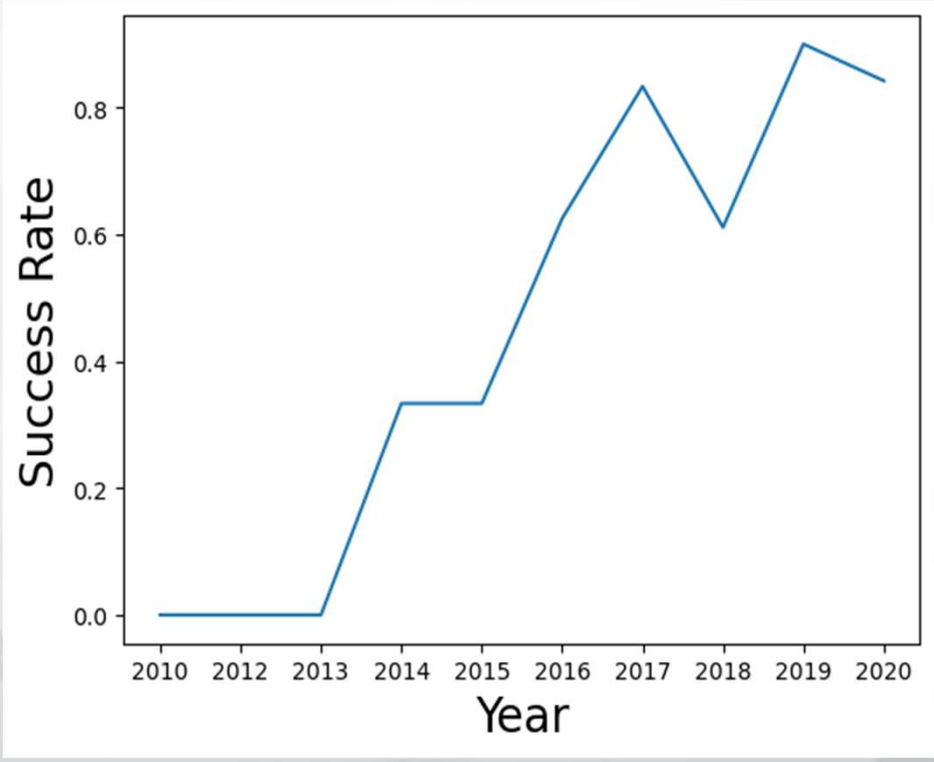
EDA – Orbit vs Payload



- The chart shows that the heaviest payloads were exclusively sent to Orbit VLEO
- It shows also that only big payloads were sent to VLEO
- ISS shows that a wide range of payloads was sent including some heavier payloads above 10000kg
- Orbit PO was targeted with either very low or very high payload, The high payloads were successful, the light ones failed.

EDA – Launch Success Yearly Trend

- We can see a clear trend of improving success rates over the years, even though it wasn't continuous (2018 and 2020 show reduced success rates compared to the previous year)
- The initial years were not successful at all
- 2016 was the first year that shows a success rate of more than 50%
- In general it looks like a good learning curve driven by gained insights and improved technology



EDA- SQL Launch Site Names

- On the right you can see the query to find the names of the Launch Sites

```
1  %sql select DISTINCT("Launch_Site") FROM SPACEXTABLE
```

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

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

EDA – SQL 5 Launch Site Names begin with CCA

```
1  %sql select * FROM SPACEXTABLE WHERE "Launch_Site" like 'CCA%' limit 5
```

Python

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

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

- Above 5 records of Launch Site names starting with CCA are shown

EDA – SQL Total Payload mass

```
1  %sql select sum("PAYLOAD_MASS__KG_") as Total_Payload from SPACEXTABLE  where "Customer" = 'NASA (CRS)'
```

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

Total_Payload
45596

- The total payload mass for NASA as customer is 45,596 kg

EDA – SQL Average Payload Mass for F9 v1.1

```
1  %sql select avg("PAYLOAD_MASS__KG_") as Avg_Payload from SPACEXTABLE  where "Booster_Version" like 'F9 v1.1%'
```

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

Avg_Payload
2534.6666666666665

- The average payload mass for Booster Version F9 v1.1 is 2,534.7 kg

EDA – SQL First Successful Ground Landing Date

```
1  %sql select min("Date") as FirstDate from SPACEXTABLE  where "Landing_Outcome" like 'Success%ground%'
```

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

FirstDate
2015-12-22

- The first successful ground landing was achieved on December 22nd 2015

EDA – SQL

Successful Drone Ship Landing with Payload between 4000 and 6000

```
1  %sql select "Booster_Version", "Landing_Outcome", "PAYLOAD_MASS__KG_" from SPACEXTABLE  where "Landing_Outcome" like 'Success%drone%' and "PAYLOAD_MASS__KG_" between 4000 and 6000
```

Python

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

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

- The list shows successful drone ship landings for payloads between 4000 and 6000

EDA – SQL Total number of success and failures

```
1  %sql select "Mission_Outcome", Count(*) from SPACEXTABLE group by "Mission_Outcome"
```

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

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

- The list shows the total numbers of successes and failures

EDA – SQL Boosters with maximum payload

```
1  %sql select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" in (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE )
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- The list shows the booster versions with the maximum payload

EDA – SQL 2015 drone ship landing failures

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

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

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The list shows the failed landing outcomes for drone ship, their booster versions, and launch site names for in year 2015

EDA – SQL

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

```
1  %sql select  "Landing_Outcome", Count(*) as Cnt
2  from SPACEXTABLE where "Date" between '2010-06-04' and '2017-03-20'
3                  group by "Landing_Outcome" order by Cnt desc
```

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

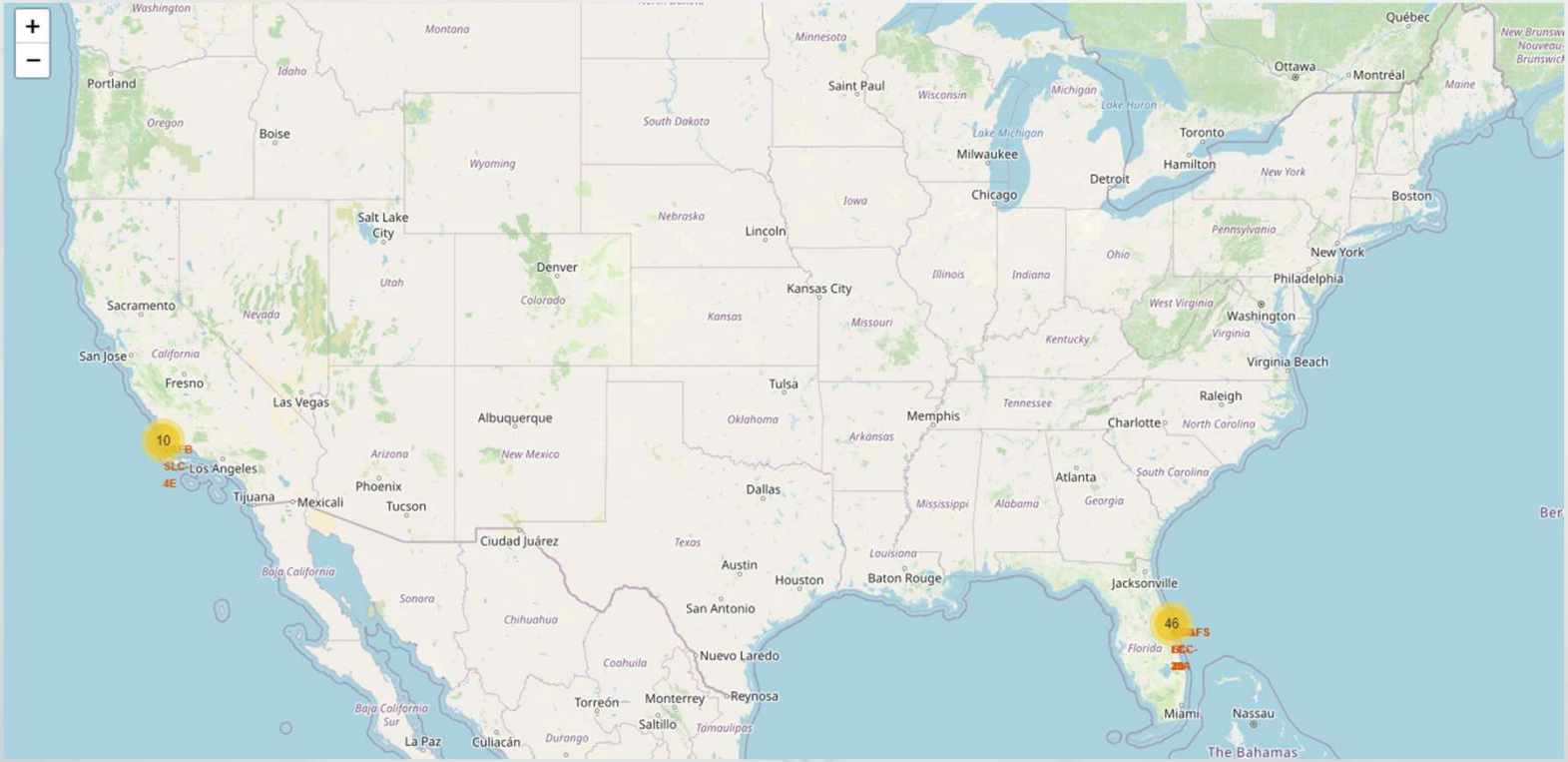
Landing_Outcome	Cnt
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- The list shows the ranked count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

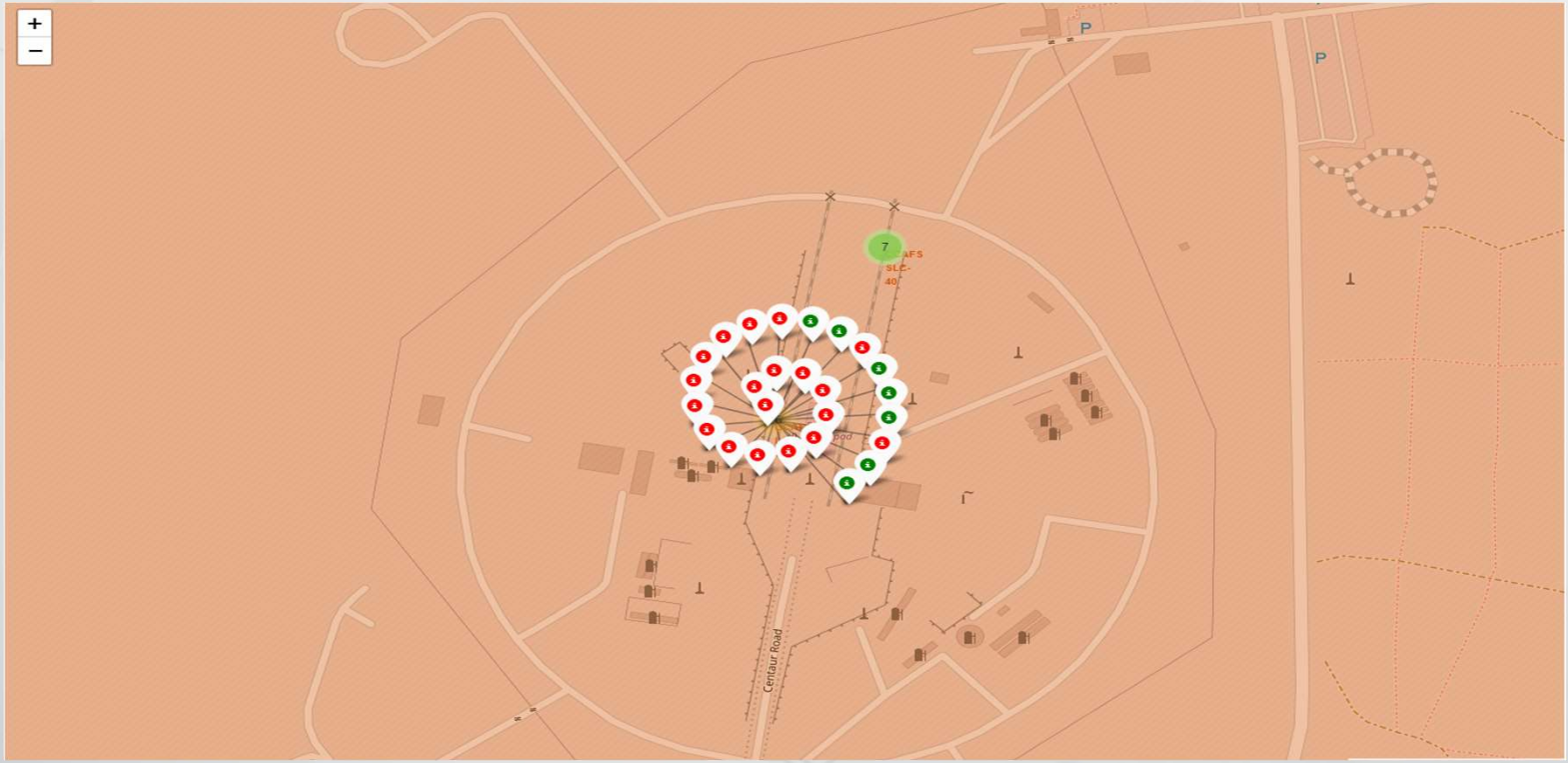
Launch Sites Proximities Analysis

Launch Sites Proximity Analysis



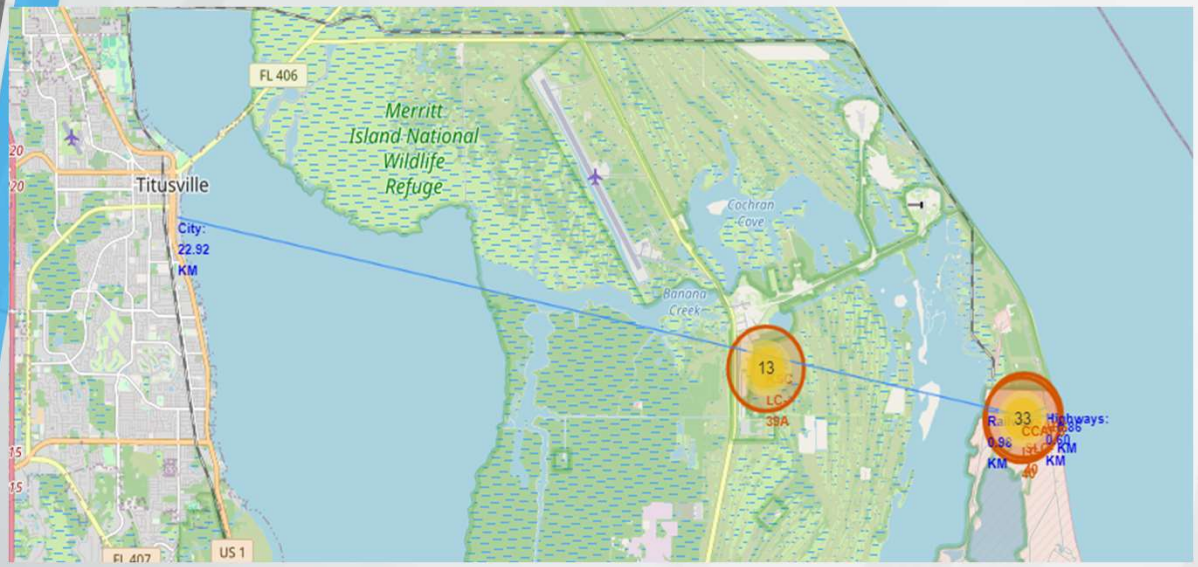
- Drawn on a map, we can see that the Launch Site locations are all close to the coastline.
- Circles, markers and labels have been added to interactively zoom into detail on the maps

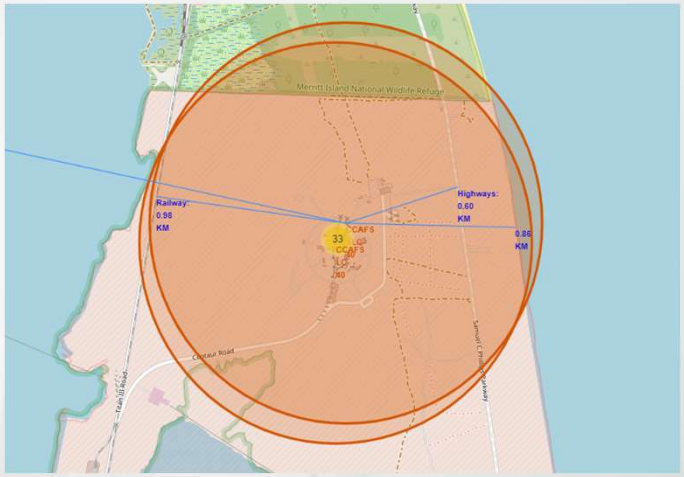
Launch Sites Proximity Analysis – Landing Outcomes



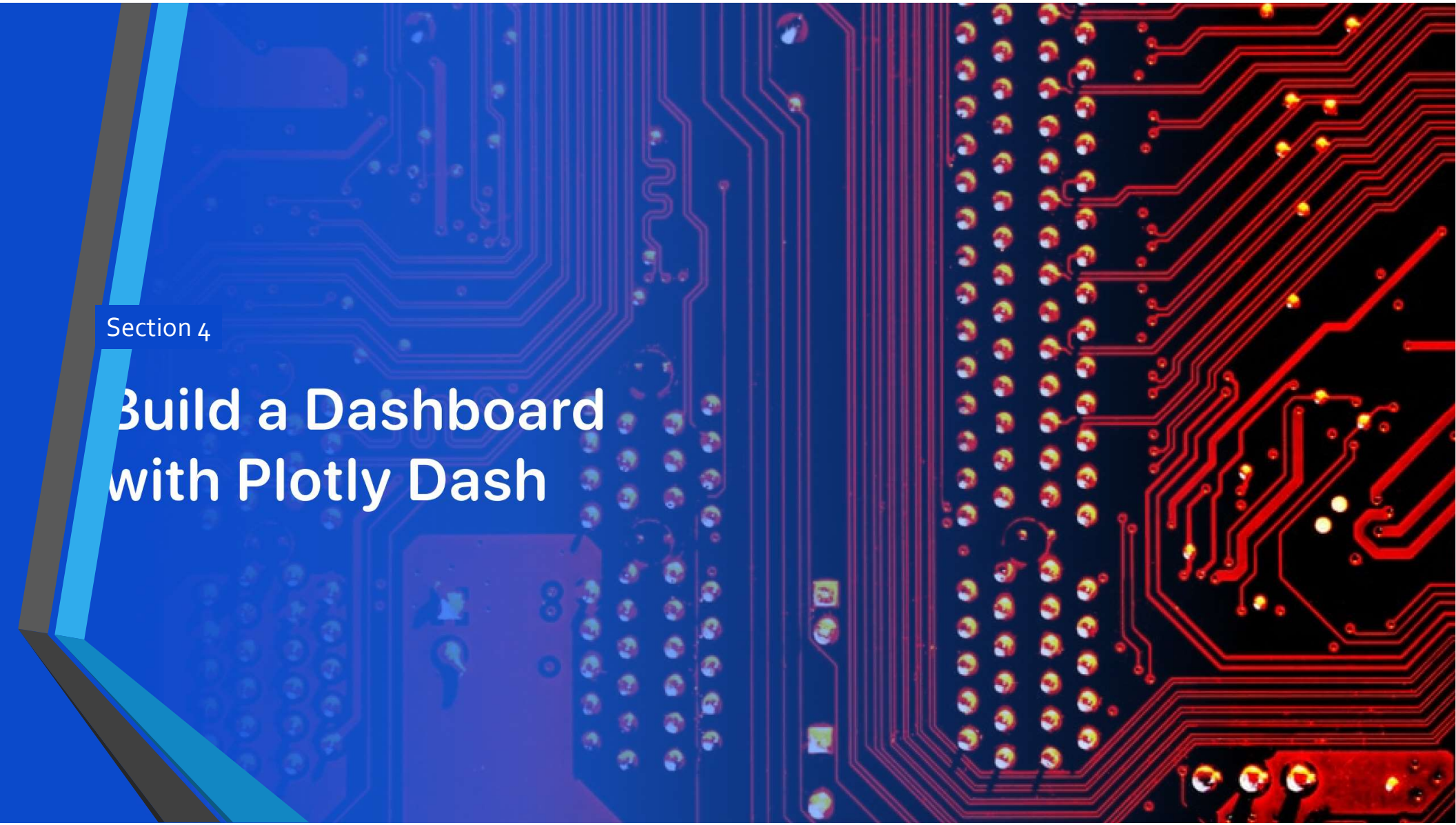
- For each landing attempt a marker has been drawn on the map to show the location of the attempt
- A red marker shows a failed landing, a green marker shows a success.

Launch Sites Proximity Analysis – Infrastructure

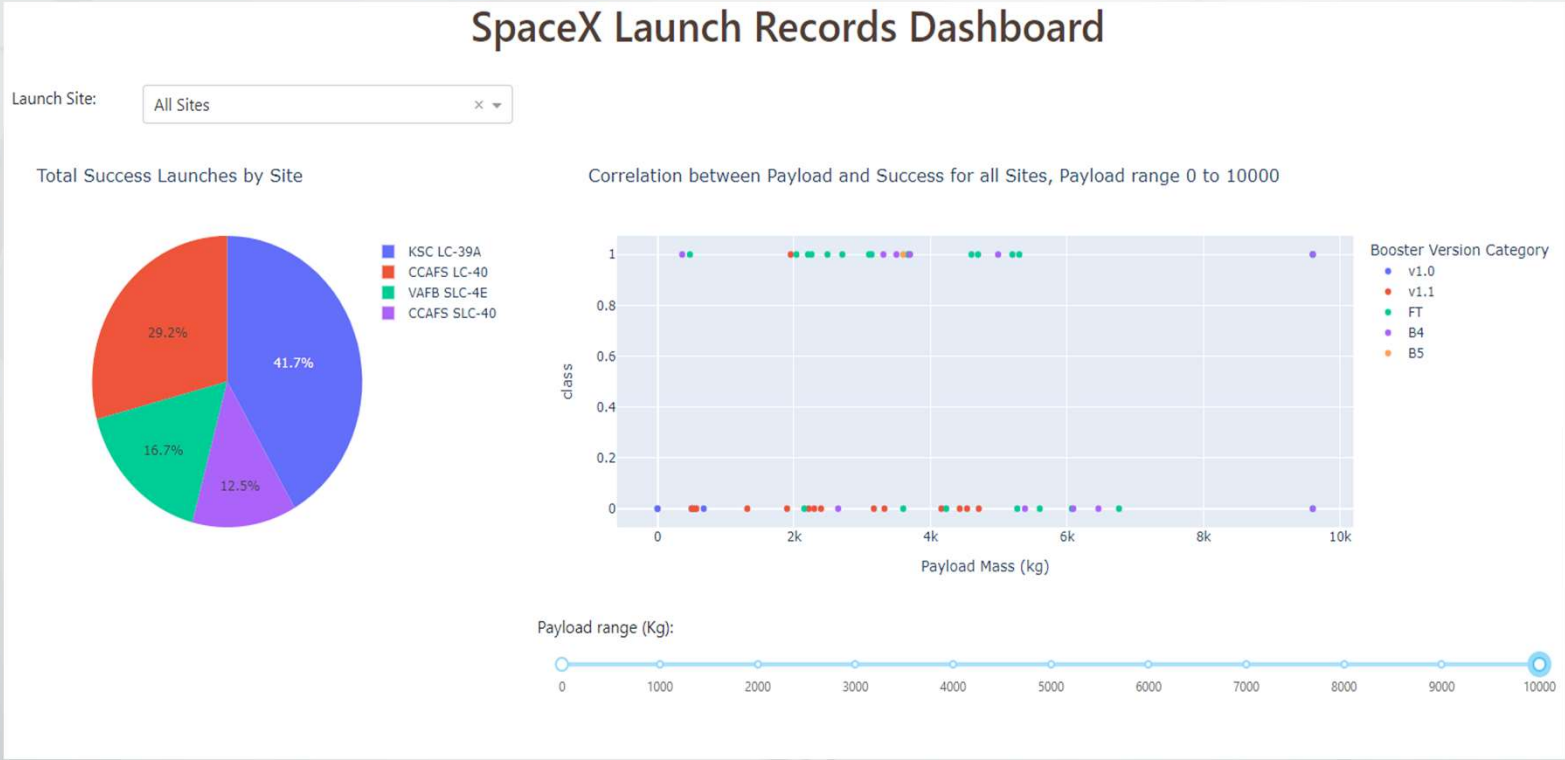




- For a launch site we indicate the proximities to infrastructure such as railway, highway, coastline and calculated and displayed the distance
- It shows that a launch site is typically located close to coastline, railway and highway but far away from cities or living space in general.

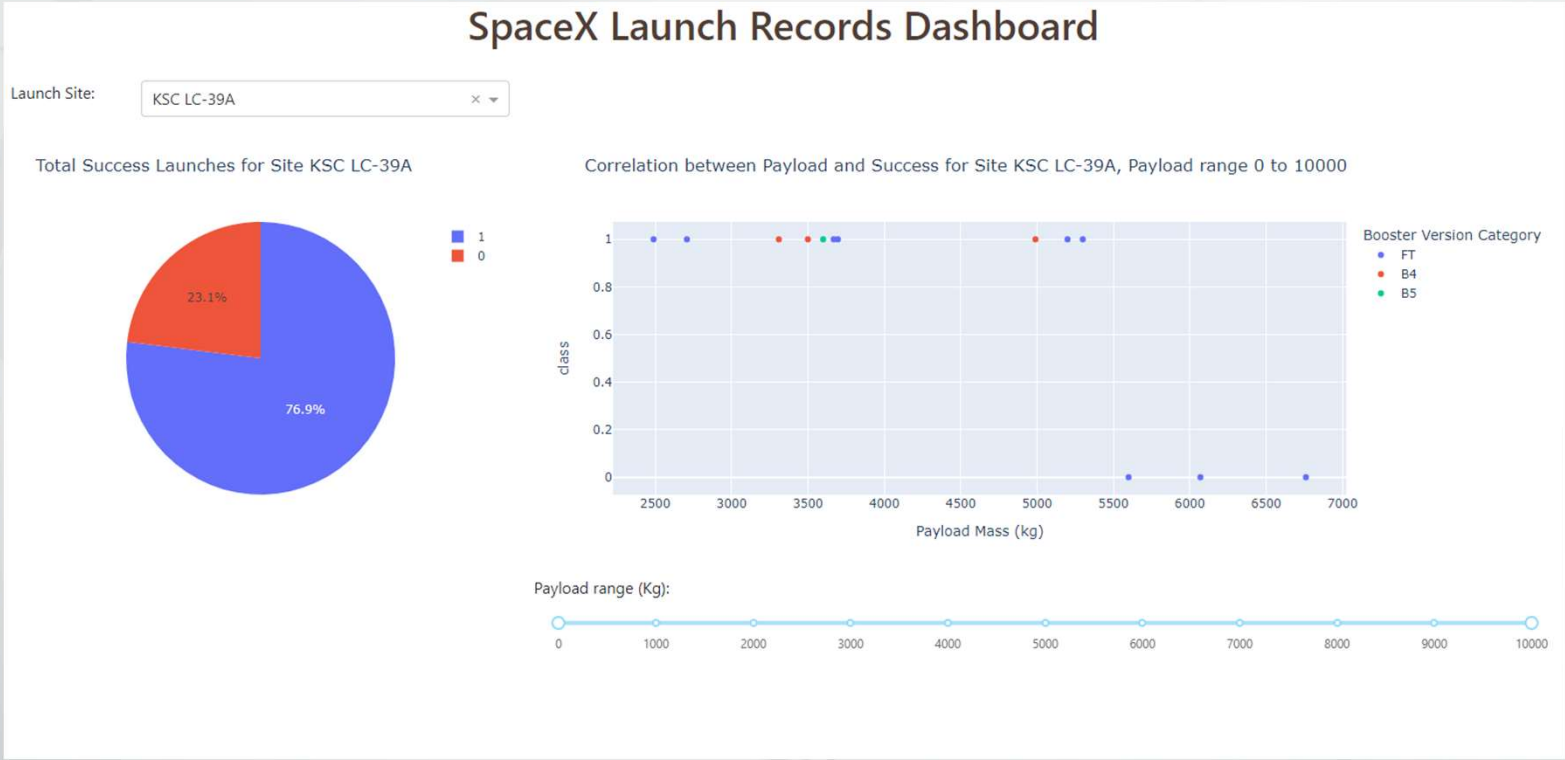

Section 4
Build a Dashboard
with Plotly Dash

Plotly Dash – Success rates for all Sites



- The dashboard allows to filter for Sites and payload range as displayed in the chart titles.
- Comparing the success rates, we can see that both KSC LC 39A and CCAFS LC-40 have the better success rates

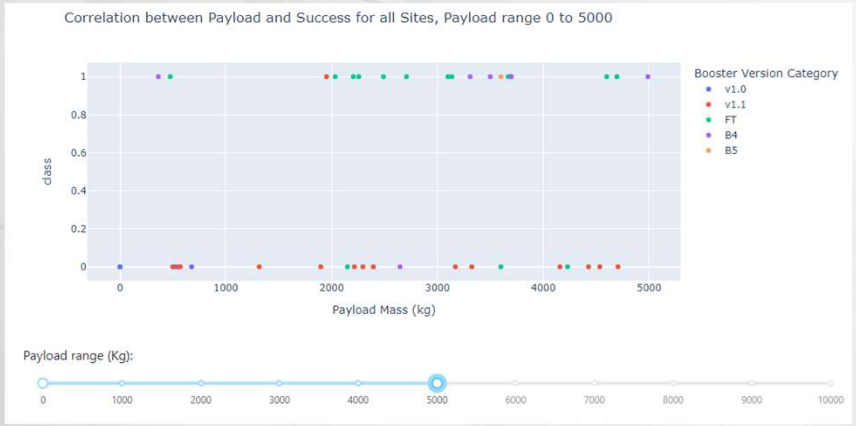
Plotly Dash – Success for KSC LC-39A



- Drilling down to KSC LC 39A we can see that the payload has a strong influence on the success for the site.
- Heavier payloads cannot be landed successfully here

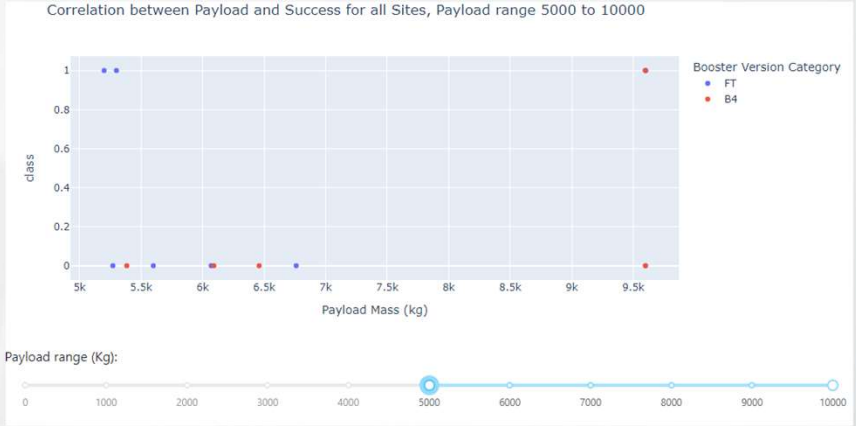
Plotly Dash – Payload and Booster Version

Low Payload



- For a payload range from 0-5000 kg we can see a mixed picture
- Early Booster versions like v1.1 have a high failure rate
- Newer Boosters like B4 and B5 show far better performance but also show some failures

High Payload



- For a payload range of more than 5000 kg we can see predominantly failures
- We also see that only two Booster version FT and B4 were used for higher payloads
- B4 looks like the preferred option for payloads above 6000 kg

Section 5

Predictive Analysis (Classification)

Prediction Analysis – Accuracy Score

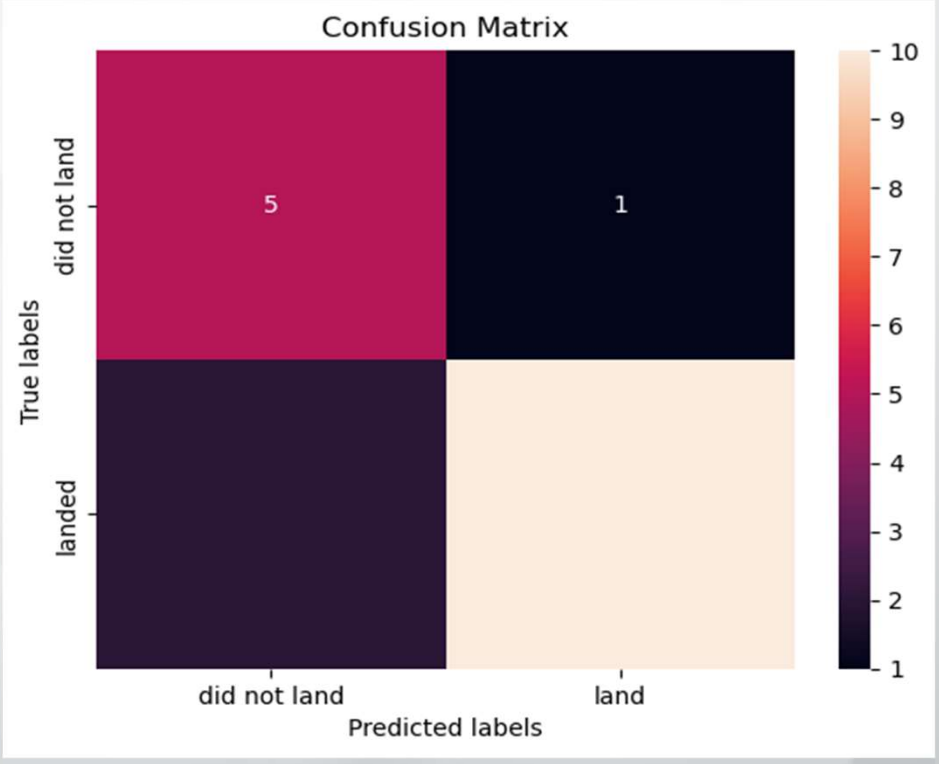
	Score	LR	SVM	Tree	KNN
0	Test Data	0.833333	0.833333	0.833333	0.833333
1	Best CrossVal	0.846429	0.848214	0.887500	0.848214
2	All Data	0.866667	0.877778	0.944444	0.855556

- The accuracy scores for the four developed models are identical for test data.
- If we extend the scoring for the best score achieved during cross validation and the score of the best tuned model for all data, test and training, we can see that the best scoring result is achieved by the Decision Tree model.
- This is confirmed by the different Confusion matrix results.

Footnote:
There might be a risk of overfitting in the Decision tree due to the small size of data compared to the node count.

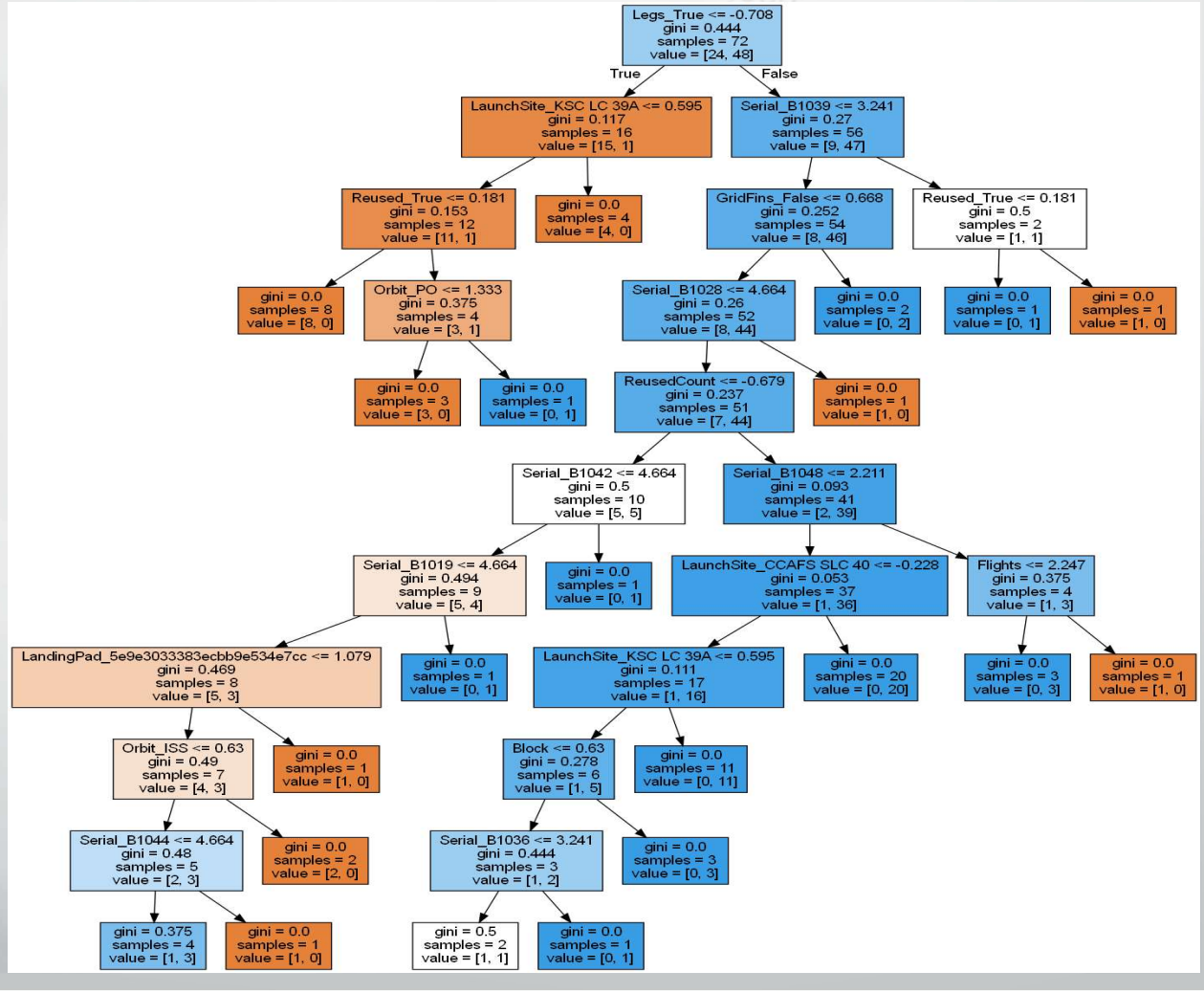
Predictive analysis – Confusion Matrix

- The confusion matrix for the best fitted decision tree model, shows good accuracy.
- Out of 18 test samples:
  - 15 are predicted correctly
  - 1 false positive, predicted as success but did not land
  - 2 false negative, predicted as failure but did land



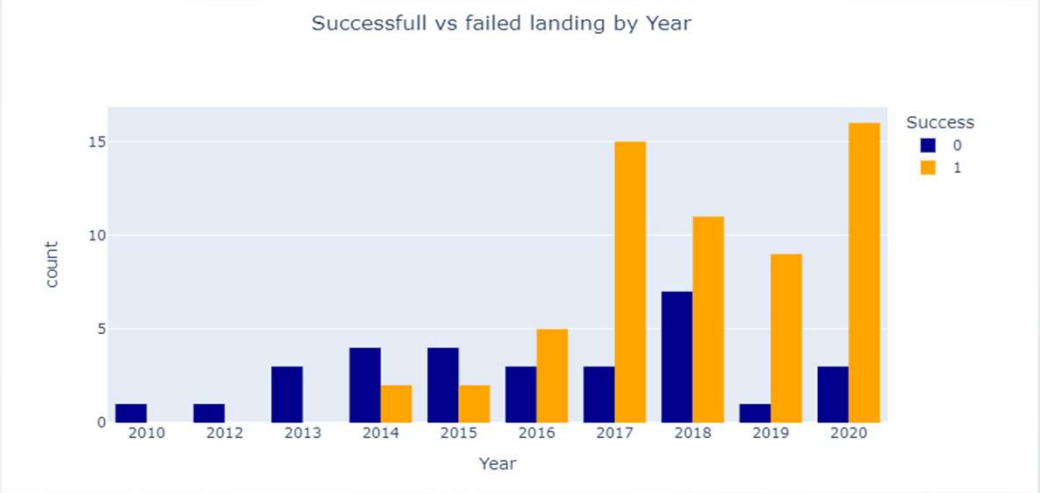
Predictive analysis – The tree model

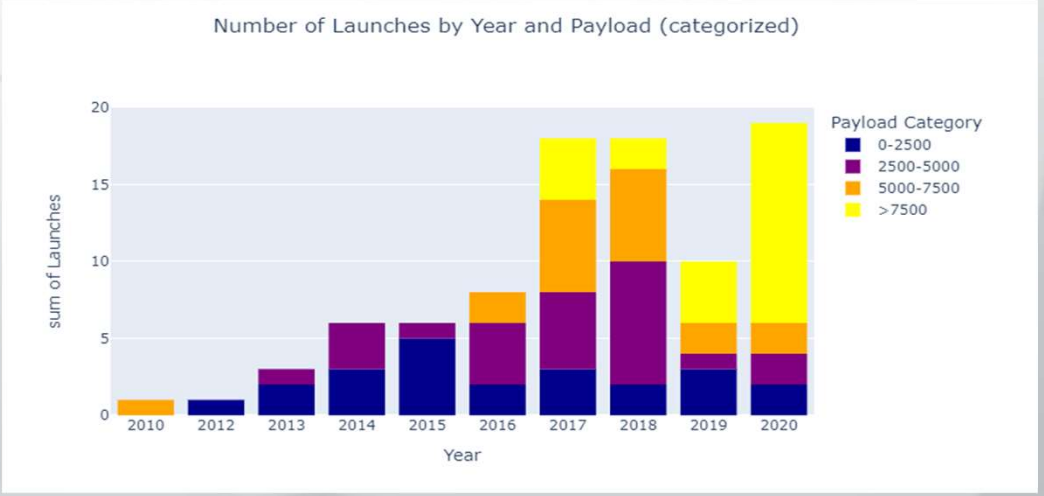
- The decision tree model gives some insight which features were discriminating.
- Two technical features were identified for further analysis and potential new insights:
  - Legs and
  - Grid Fins



Predictive analysis – Trends

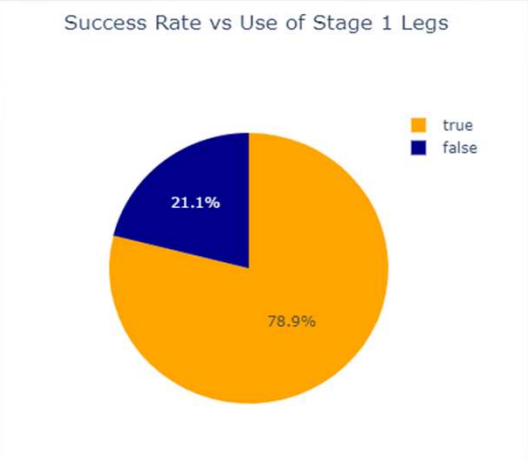
- To visualize the trend, we aggregated the data into years and payload categories.
- We can see that success rate, and also success count increase over time
- The ability to handle larger payloads improved also over the years.
- Payloads larger than 7500 kg were first introduced in 2017





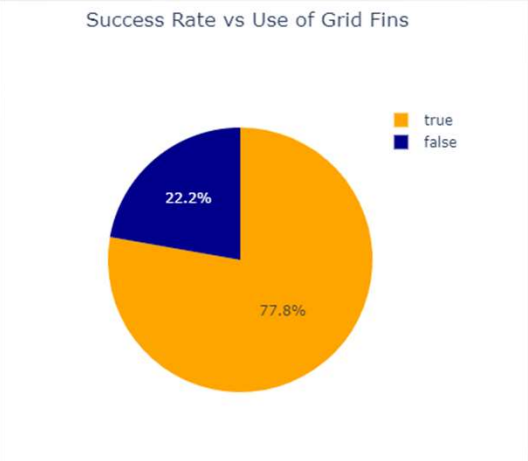
Predictive analysis – Use of Legs for Stage 1

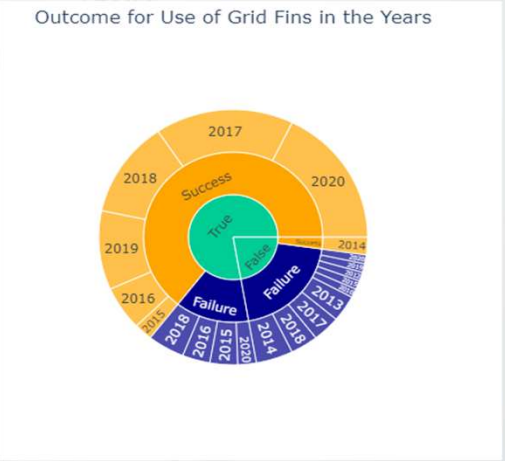
- Falcon 9's first stage legs are designed to retract along the side of the rocket and extend outward to facilitate stable landing on solid ground or drone ships after launch. They were first introduced 2014.
- We can see that the use of these deployable landing legs has a success rate of 78.9%.
- We can also see that they are more often used in later history and that they while not guaranteeing a success are deployed at increasing success rate.

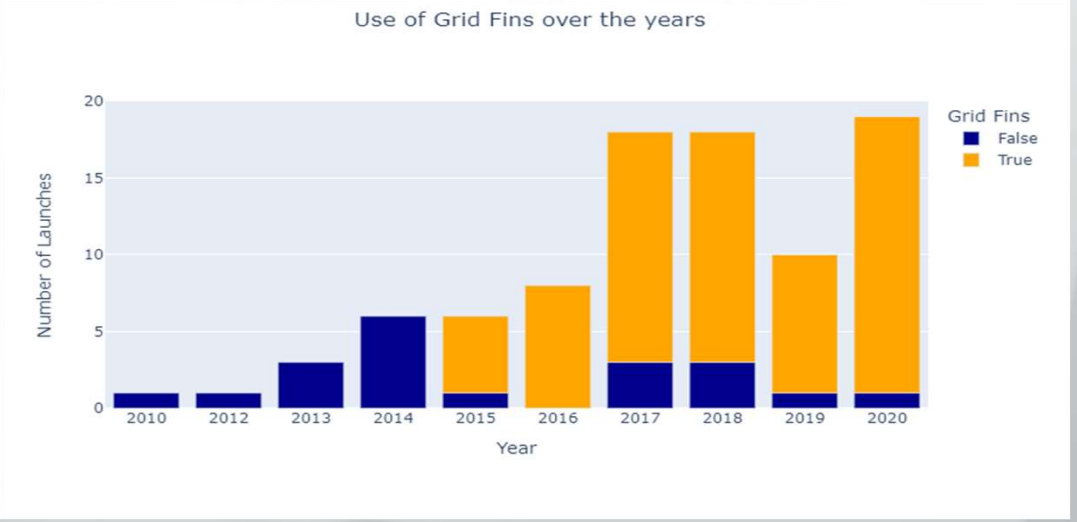


Predictive analysis – Use of Grid Fins

- Falcon 9's grid fins deploy during descent to provide precise aerodynamic control for steering and stabilizing the first stage as it returns to Earth. They were first introduced 2015.
- We can see that the use of these deployable landing legs has a success rate of 77.8%.
- As Stage 1 Legs they are more frequent used in later history. They do not guarantee a success but are deployed with increasing success rate.

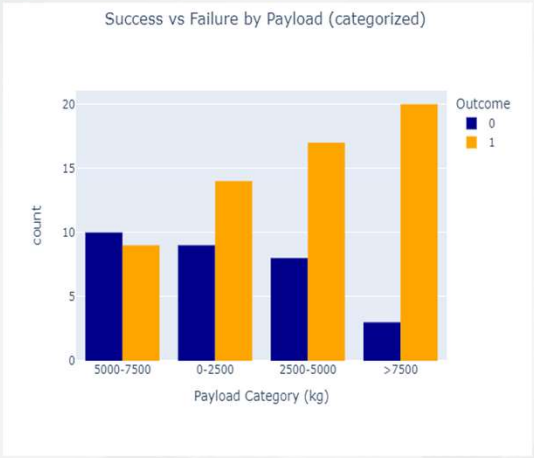


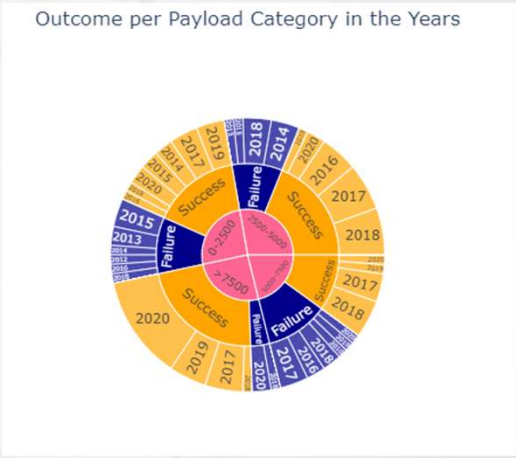


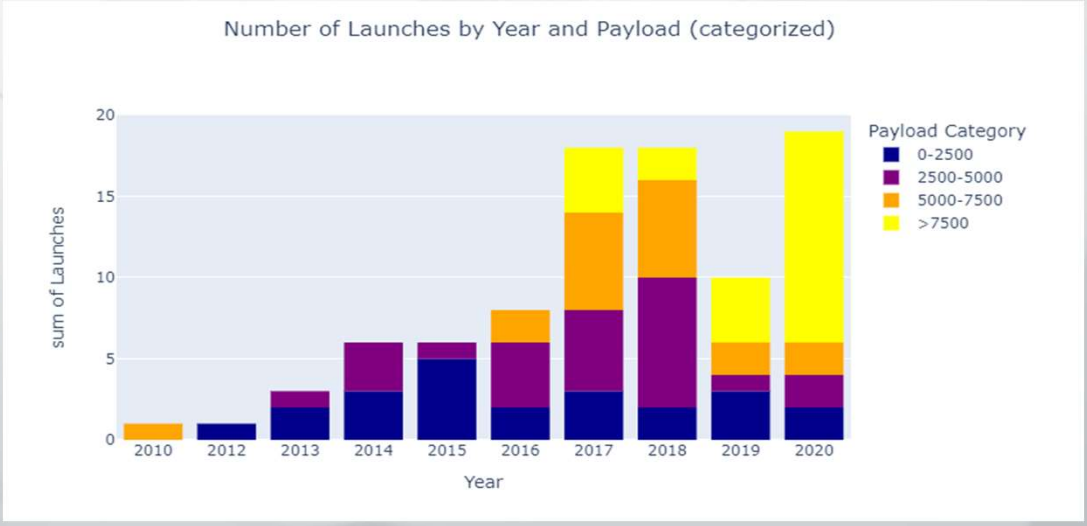


Predictive analysis – Payloads

- Payload seem not to have a decisive impact on the success if we analyse the data aggregated in categories.
- Higher payloads, later introduced in the launches, perform not worse than lighter payloads in an annual comparison
- The sunburst chart shows as example perfect success rates for the high payload category in the years 2019 and 2017.







Conclusions

- The available data shows that the success rate for the last year of observations reached a range of more than 80% which explains the cost savings through reuse of the Falcon 9 first stage.
- We can see a clear positive trend in the success rate of first stage landings for the Falcon 9 over the years.
- Technological advancements like Grid Fins and landing Legs contributed to the improvements. Certainly, experience in launching and landing support also the trend and indicate sustained and potentially increased future cost advantages for SpaceX.
- With the derived Decision Tree model, we can predict future landing outcomes with an accuracy of more than 80%.
- The model has to be continuously adjusted to take future outcomes and changes in technology into consideration.


Thank you!