Winning Space Race with Data Science

Hakan Alkaya
29.10.2022

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- EDA with SQL
- EDA with Vizualization
- Visual Analytics with Folium
- Interactive Dashboard with Ploty Dash

Summary of all results

- Results from EDA
- Interactive Analytics
- Machine Learning Prediction

Introduction

Project Background

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other competitors cost upward of 165 million each
- This is mainly due to the reusability of the first stage of the Falcon 9 rockets

Problems to find answers for

- We as SpaceY therefore want to determine if and with what parameters the first stage will land -> so that we can determine the cost of a launch

Section 1

Methodology

Methodology

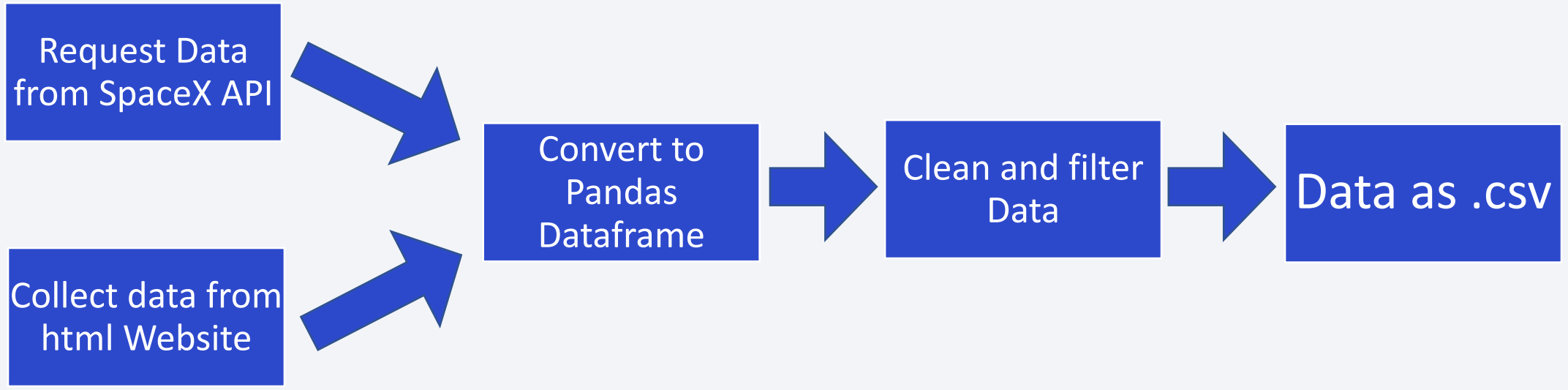
Executive Summary

- Data collection methodology:
  - Requests from SpaceX Rest API
  - Web Scraping from Wikipedia entry
- Perform data wrangling
  - Data preparation for ML algorithm, mainly One Hot Encoding and Data cleaning
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Two different methods where used:

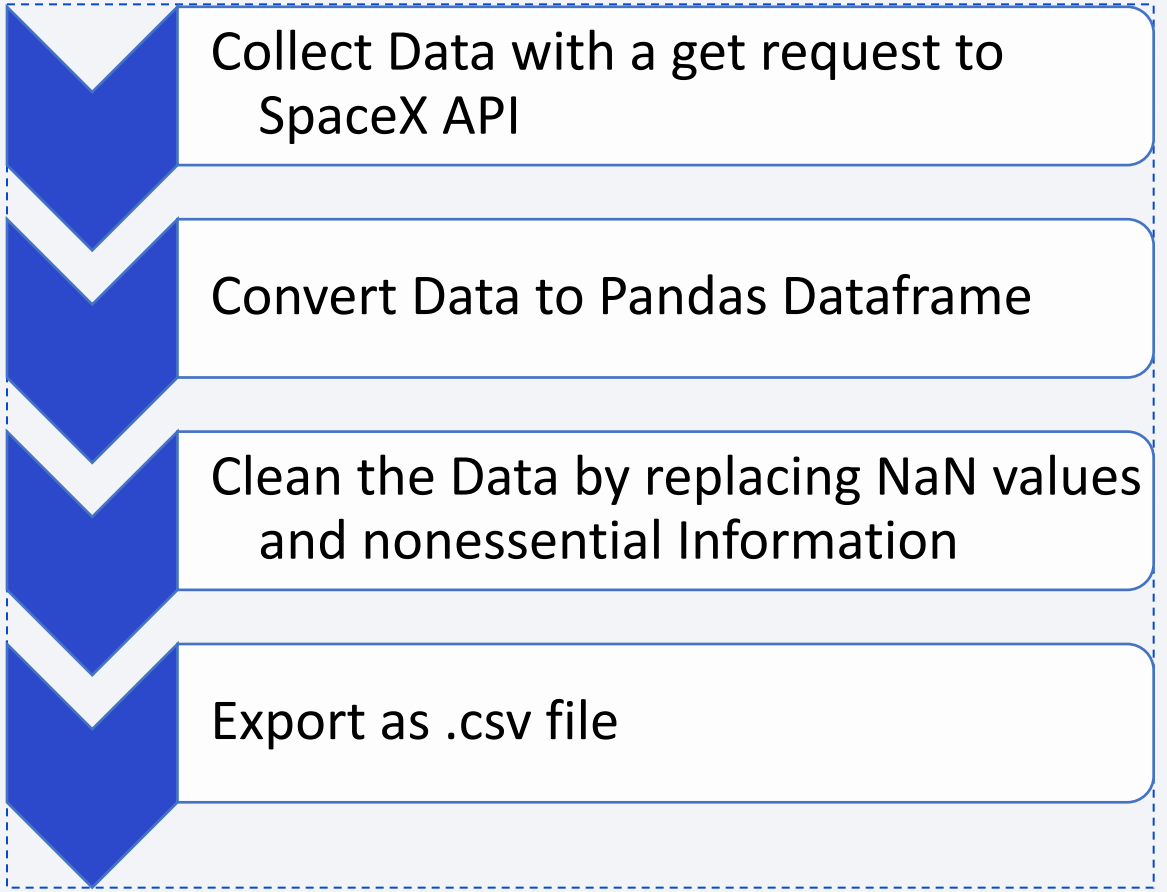
  Get requests to SpaceX API and **Web Scraping** from Wikipedia entry



Data Collection – SpaceX API

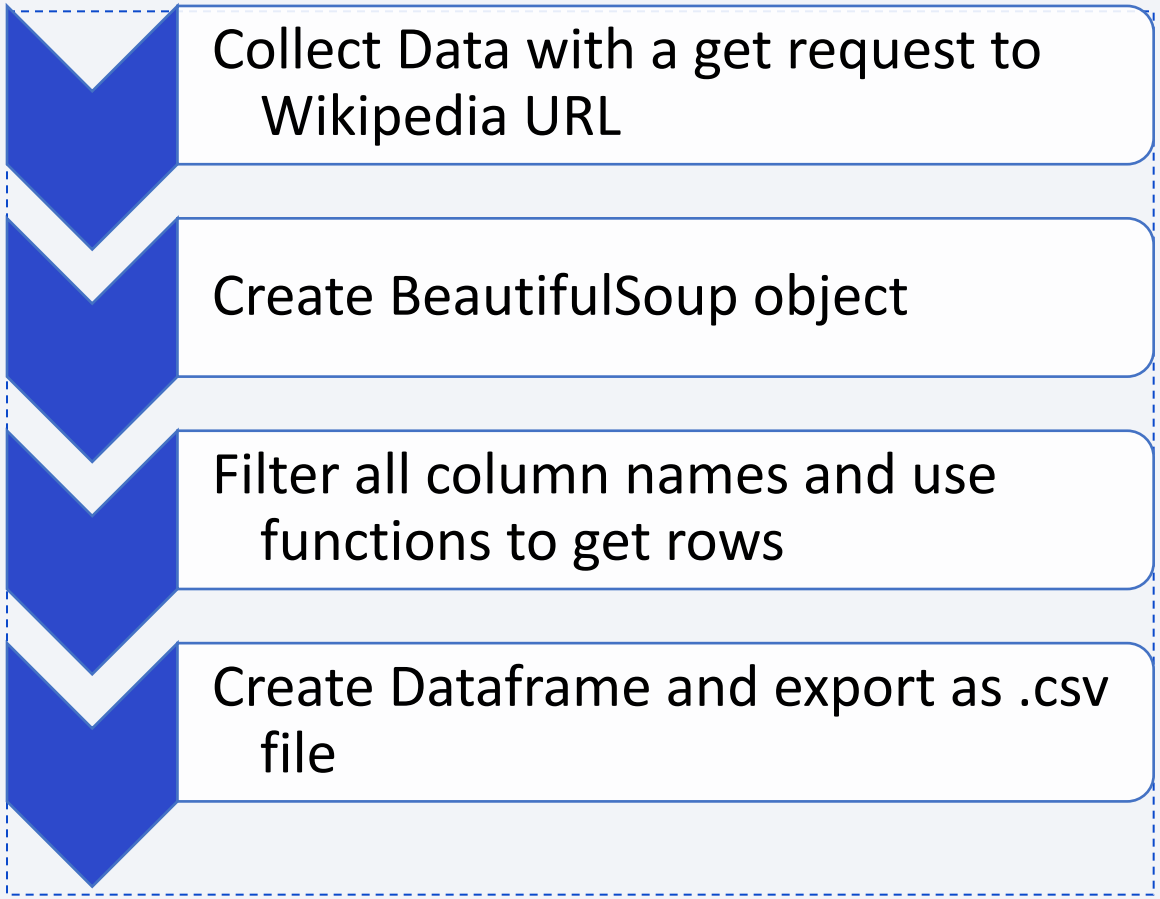
- The Data was collected by making a get request to the SpaceX API
- The noisy data is visualized and cleaned

- GitHub URL: https://github.com/HakHak34/master/blob/main/jupyter-labs-spacex-data-collection-api (1).ipynb



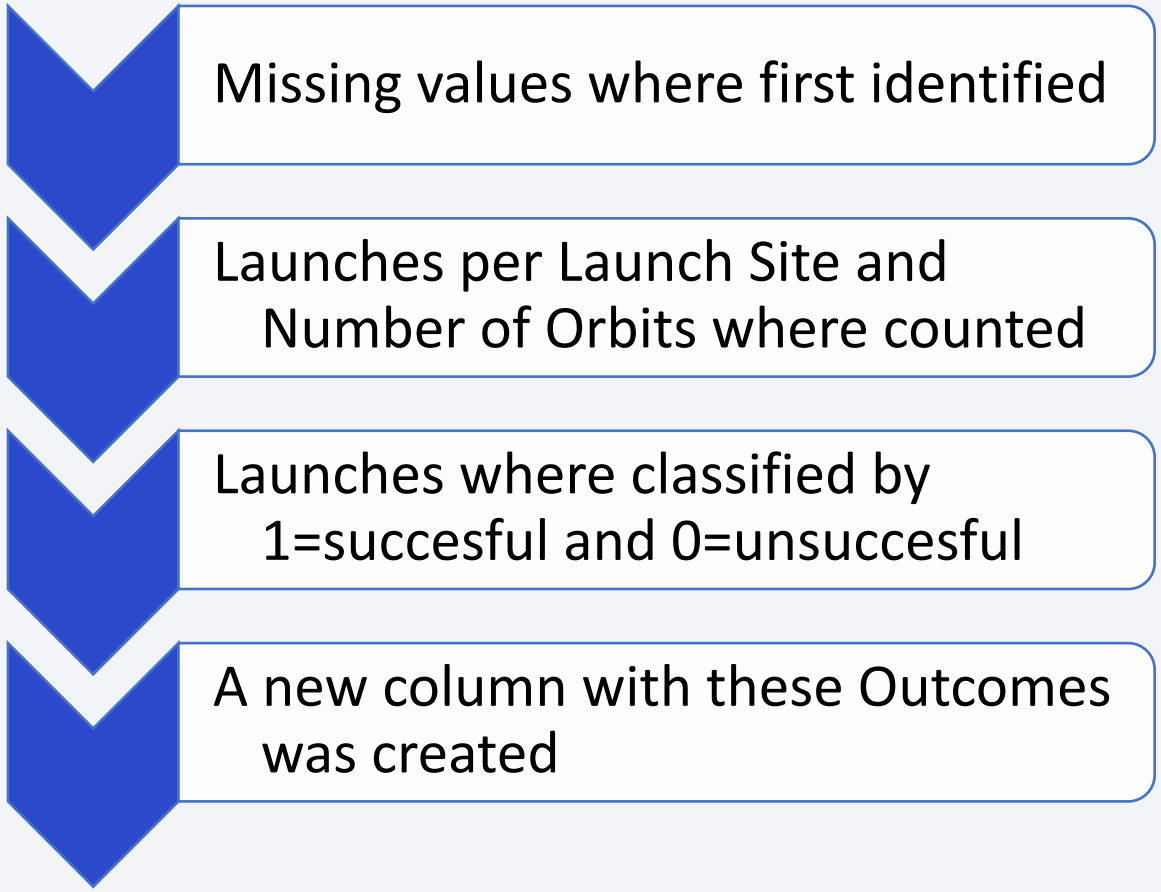
Data Collection - Web Scraping

- Data from the Wikipedia page of Falcon 9 and Falcon Heavy Launches gets collected via Web Scraping
- BeautifulSoup is used to filter html text
- GitHub URL: https://github.com/HakHak34/master/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- GitHub URL: https://github.com/HakHak34/master/blob/main/labs-jupyter-spacex-Data wrangling.ipynb



EDA with SQL

- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'KSC'
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- Listed the date where the succesful landing outcome in drone ship was acheived
- Listed the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed the names of the booster_versions which have carried the maximum payload mass
- Listed the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Ranked the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

- GitHub URL: https://github.com/HakHak34/master/blob/main/jupyter-labs-eda-sql-edx_sqllite.ipynb

EDA with Data Visualization

Summary of what charts were plotted

- Flight number vs. Payload Mass was plotted as a categorical plot to see if Payload rises with number of flights
- Flight number vs. Launch Site was also plotted as catplot to see if rocket launches shift towards certain locations
- Payload Mass vs. Launch Site was plotted to see payload limits of certain locations
- Success rate of different Orbits was plotted as a pie chart to compare them
- Flight Number vs. Orbit was plotted as catplot to see which Orbits get prioritized with time
- Succes Rate over Date was plotted as a scatter plot to see trends over time

- GitHub URL: https://github.com/HakHak34/master/blob/main/jupyter-labs-eda-dataviz.ipynb

Build an Interactive Map with Folium

- Markers with circles where added to the map to locate Launch Sites
- Showing Longitude and Latitude when hovering with mouse was added to get exact coordinates of relevant object
- Using this coordinates the distances where calculated using geometrical formulas
- Lines where added to visualize these distances

- GitHub URL: https://github.com/HakHak34/master/blob/main/lab_jupyter_launch_site_location.ipynb

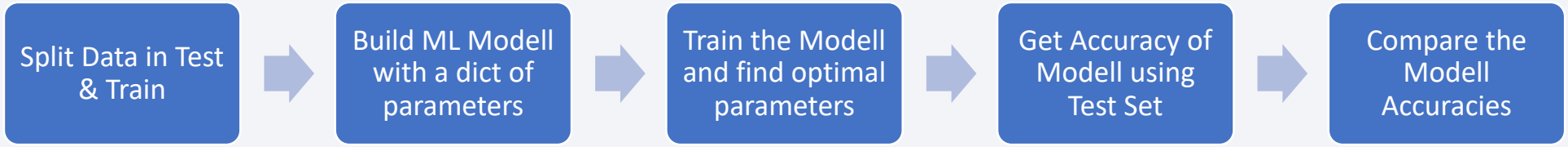
Build a Dashboard with Plotly Dash

- An interactive pie chart was created to visualize the Success rates of a given Launch Site or all Sites
- Next an interactive Scatter Plot was created to analyze the Outcome based on different ranges of the payload

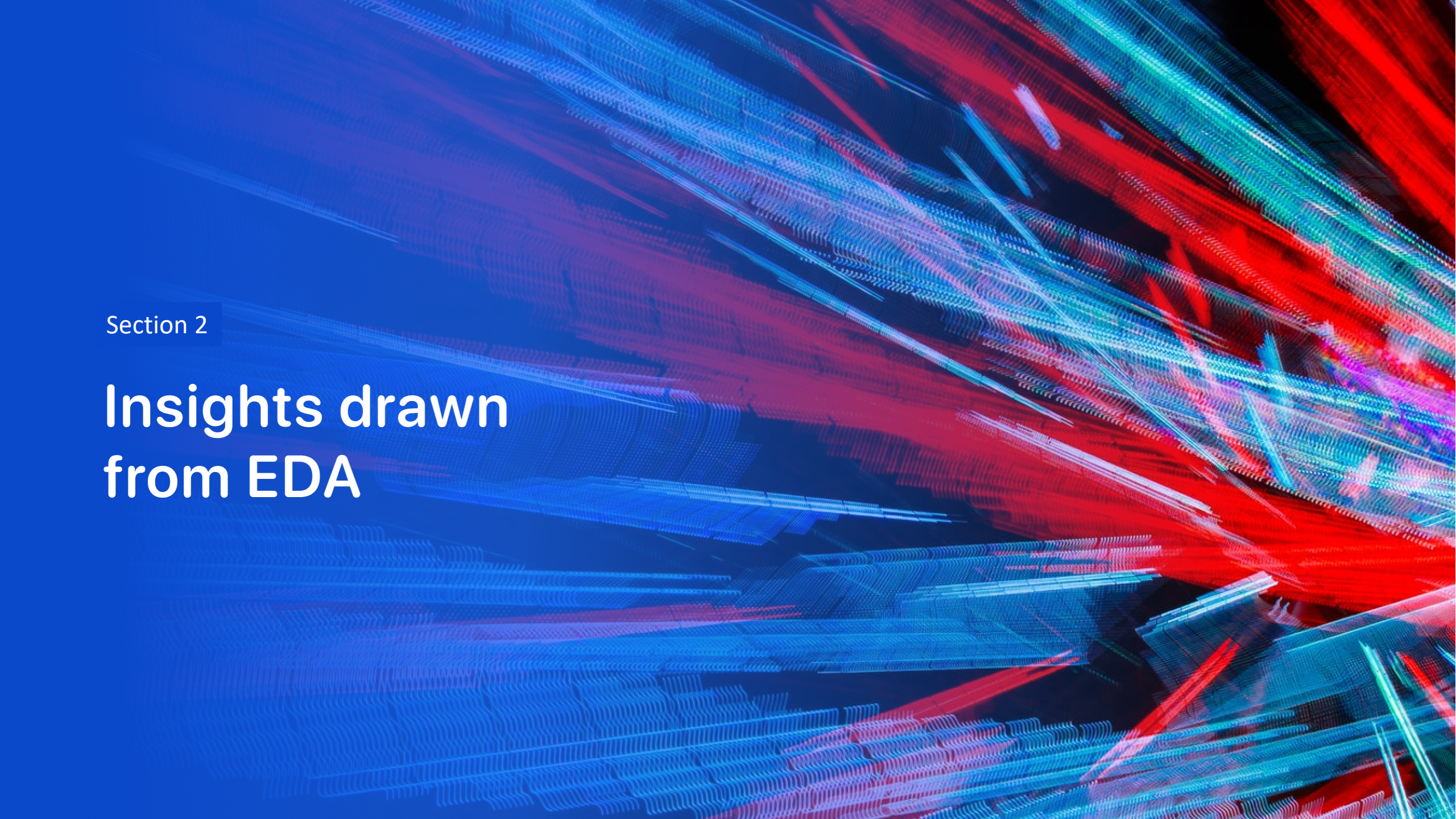
- GitHub https://github.com/HakHak34/master/blob/main/lab_jupyter_launch_site_location.ipynb

Predictive Analysis (Classification)

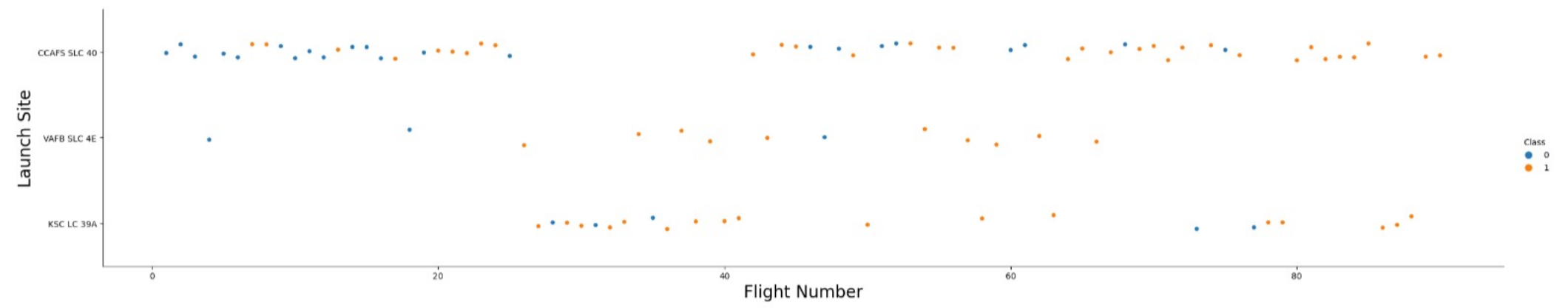
- The prepared Data was split into Training and Test Data
- Different ML-Models where built and the parameters where updated during the Training process using the Grid Search algorithm
- With these parameters the Modell was validated by getting predictions on the Test Data
- The Accuracy was compared between all the Models (SVC, KNN, Logistic Regression and Decision Tree



- GitHub URL: https://github.com/HakHak34/master/blob/main/SpaceX_Machine Learning Prediction_Part_5.ipynb

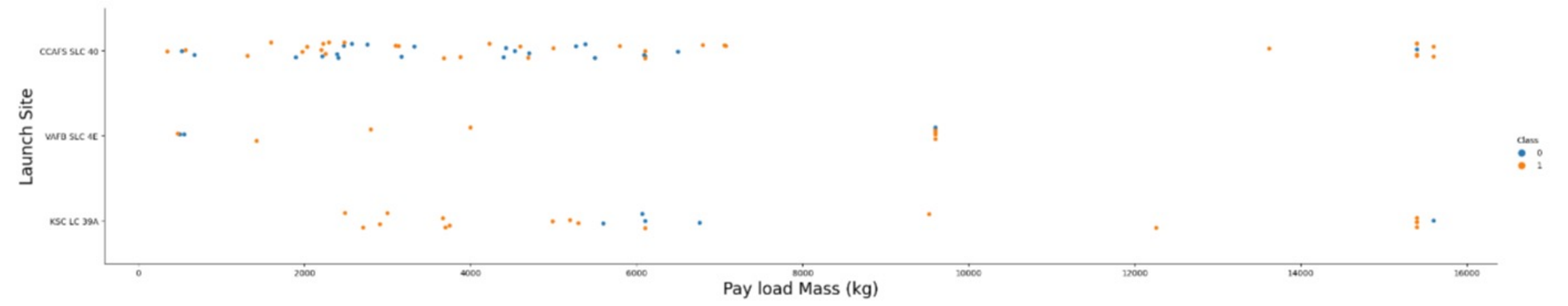

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



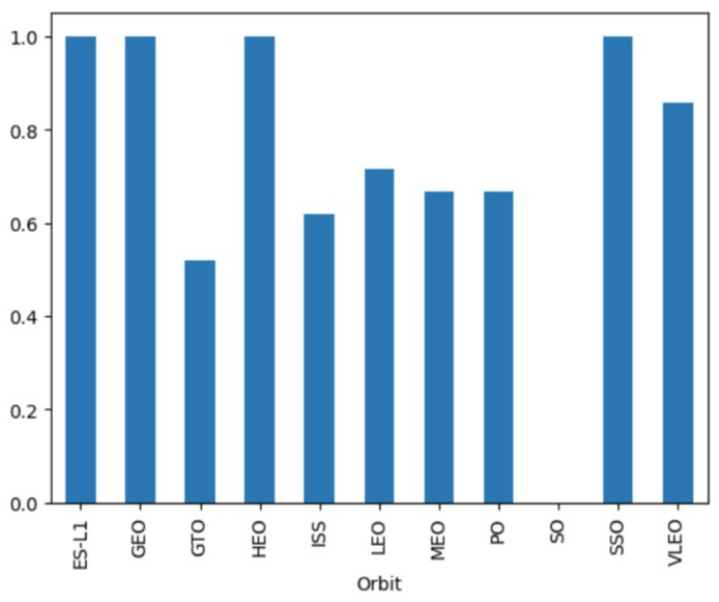
- The first Site used is CCAFS SLC 40 with many failed launches
- After aprox. 25 flights the KSC LC 39A is introduced
- Not many launches from VAFB SLC 4E but high success rate
- With rising flight numbers generally the success rate goes up

Payload vs. Launch Site



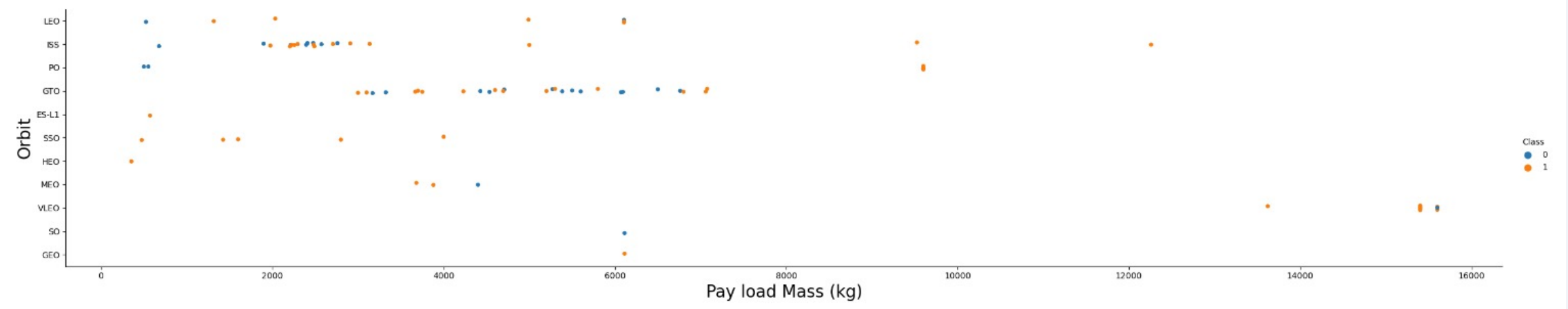
- CCAFS SLC 40 seems to be used for smaller payloads up to aprox. 7000kg
- KSC LC 39A is used for all ranges and the max. payload of aprox 13000kg
- VAFB SLC 40 no more than 10000kg
- Payload does not seem correlated to launch success

Success Rate vs. Orbit Type



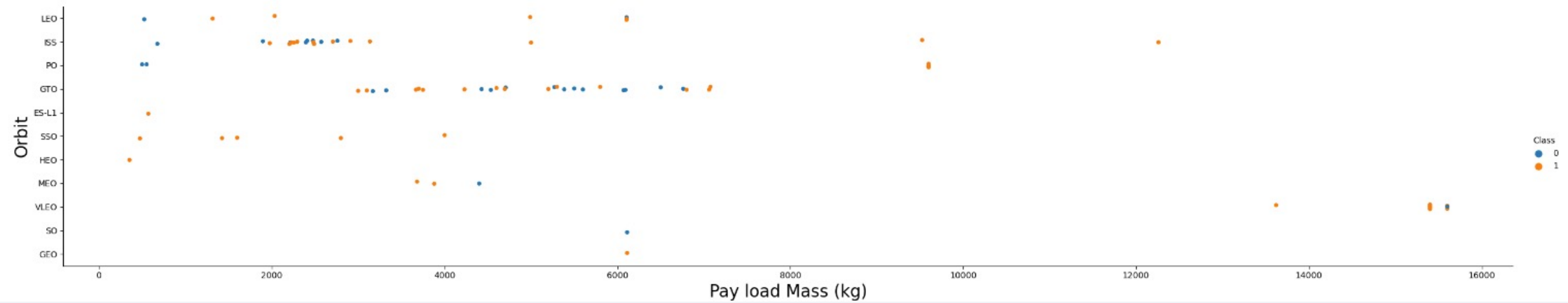
- The Sun-synchrous (SO) has never been successfull
- Whereas ES-L1, GEO, HEO, SSO all have 100% success rate
- Different orbits are in between with success rates > 50%
- Success rate does not seem correlated with distance to earth

Flight Number vs. Orbit Type



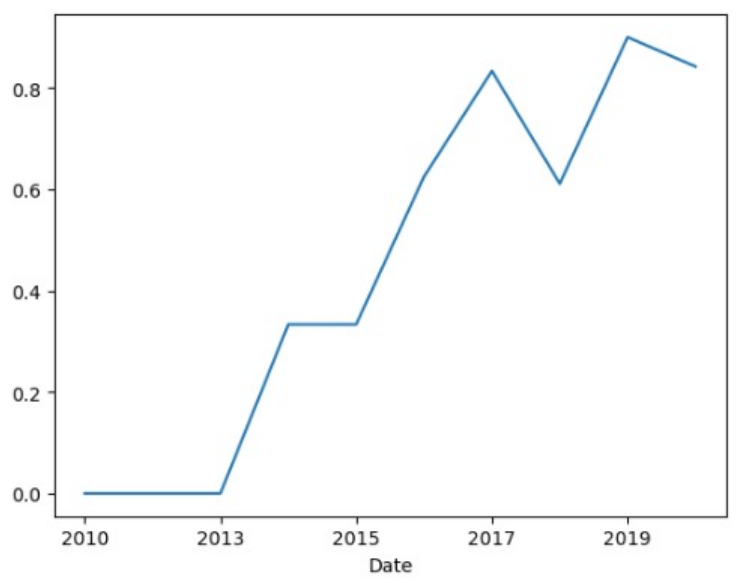
- VLEO gets prioritized for heavy payloads with high success rate
- SO only had one launch so 0% success rate not significant
- Most payloads between a few hundred kg up to 7000kg for most Orbits
- With heavy payloads the positive landing rate are often for Polar,LEO and ISS

Payload vs. Orbit Type



- Big range of payloads for ISS Orbit
- Light payloads for SSO, HEO, ES-L1 and HEO
- Heavy payloads for VLEO Orbit
- It appears that most failures happen with very light payloads smaller 1000 kg and payloads from 4000 kg up to 7000 kg

Launch Success Yearly Trend



- A clear trend over time evident
- Succes rate gets better for years although there was a outlier in 2018
- Reasons: Technology gets better, SpaceX learns from trial and error, NASA cooperations helped them gain knowledge

All Launch Site Names

- Find the names of the unique launch sites

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with CCA'

- Find 5 records where launch sites' names start with `CCA`

```
%sql SELECT * FROM SPACEXTBL WHERE (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- With * we get all columns from the Database called SPACEXTBL who meet the given condition
- Limitation of output with the LIMIT command

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER LIKE 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
SUM(PAYLOAD_MASS__KG_)
                 45596
```

- The SUM command sums up the values of all rows who fullfill the customer condition

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

- Here the AVG command calculates the mean of all payload mass columns where the Booster Version is F9v1.1

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on drone ship.

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE [Landing _Outcome] LIKE 'Success (drone ship)'
 * sqlite:///my_data1.db
Done.
```

MIN(DATE)
06-05-2016

- The condition gets applied to the Outcome column and only the first date gets printed out

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE [Landing _Outcome] LIKE 'Success (ground pad)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

- Here 3 conditions are used in sequence

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE '%Success%' OR Mission_Outcome LIKE '%Failure%'
```

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

COUNT(*)
101

- The OR condition is used to get both successful and failure outcomes

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * 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

- With the use of a subquery the max payload is determined to then be used as a condition for the Booster Versions

2015 Launch Records

- List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

```
%sql SELECT SUBSTR(DATE, 4, 2) AS Month_Name, [Landing _Outcome], Booster_Version, Launch_Site FROM SPACEXTBL WHERE SUBSTR(DATE,7,4)='2017' AND [Landing _Outcome]='Success (ground pad)'
 * sqlite:///my_data1.db
Done.
```

Month_Name	Landing _Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- The month is extracted using the SUBSTR command and the date column is being output as month_name

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

- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

```
%sql SELECT [Landing _Outcome], COUNT([Landing _Outcome]) AS COUNT FROM SPACEXTBL WHERE DATE >= \
'04-06-2010' AND DATE <= '20-03-2017' AND [Landing _Outcome] LIKE 'Success%' GROUP BY [Landing _Outcome] \
ORDER BY COUNT([Landing _Outcome]) DESC
```

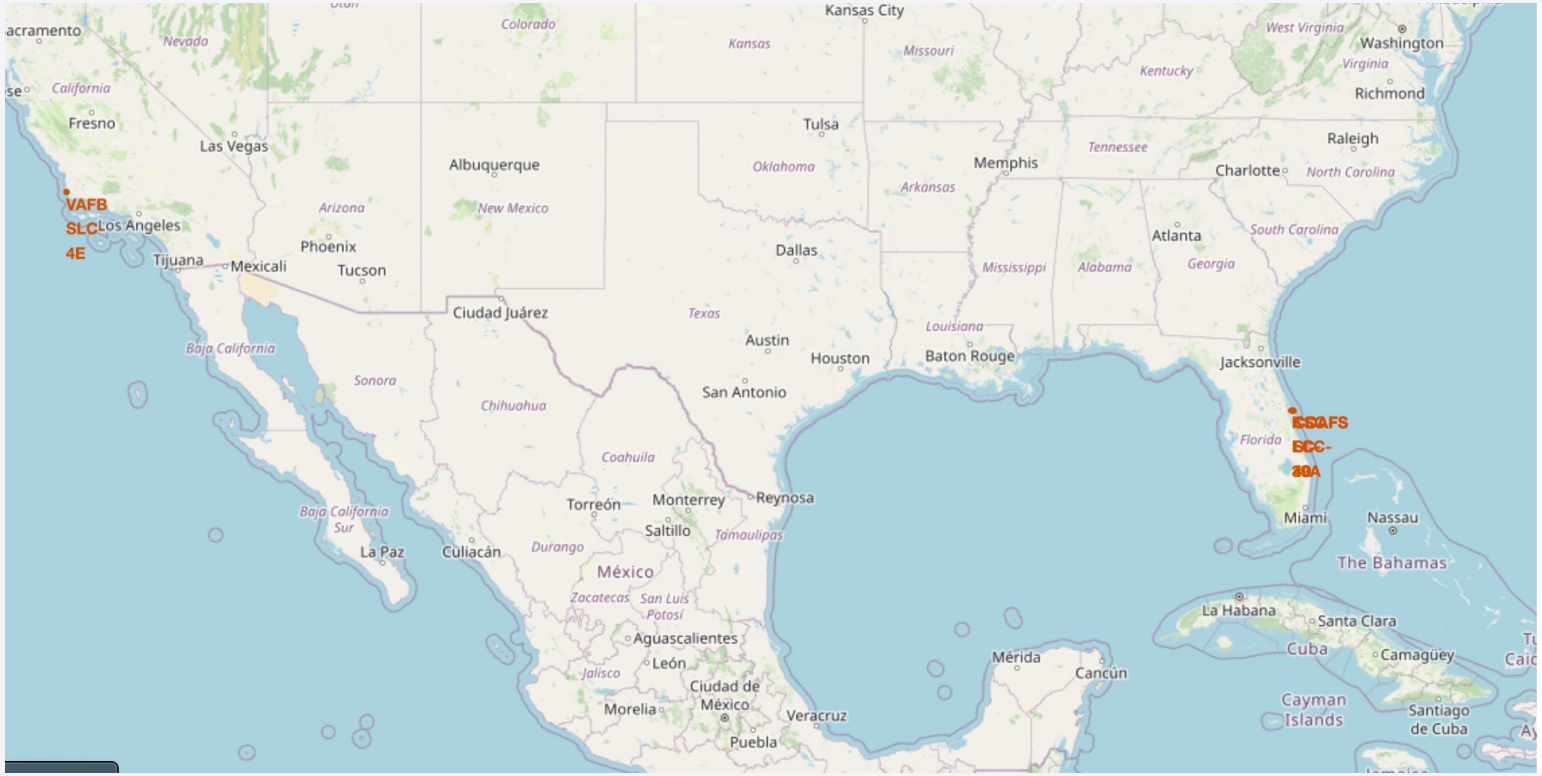
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	COUNT
Success	20
Success (drone ship)	8
Success (ground pad)	6

- The results are ranked using the RANK and DESC commands

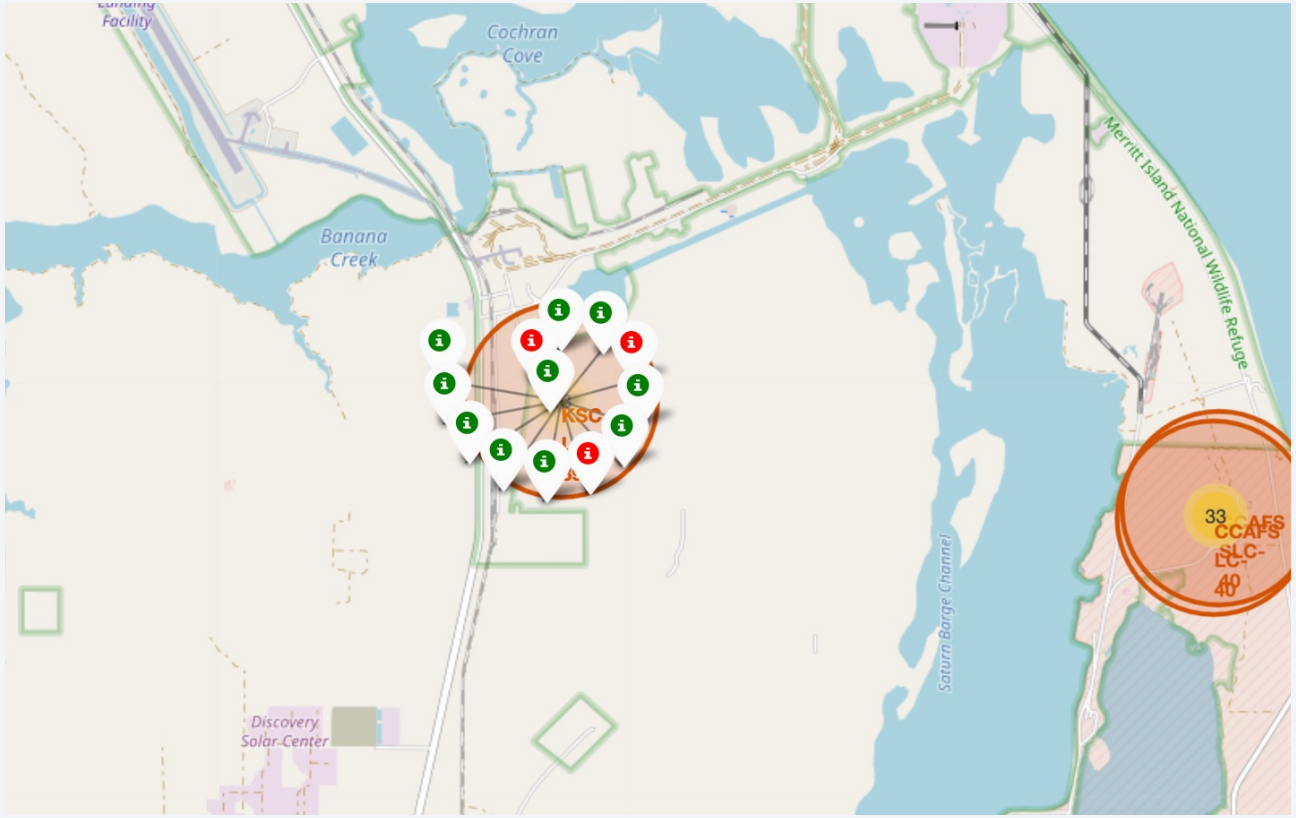

Section 3
Launch Sites
Proximities Analysis

Map with all Launch Sites



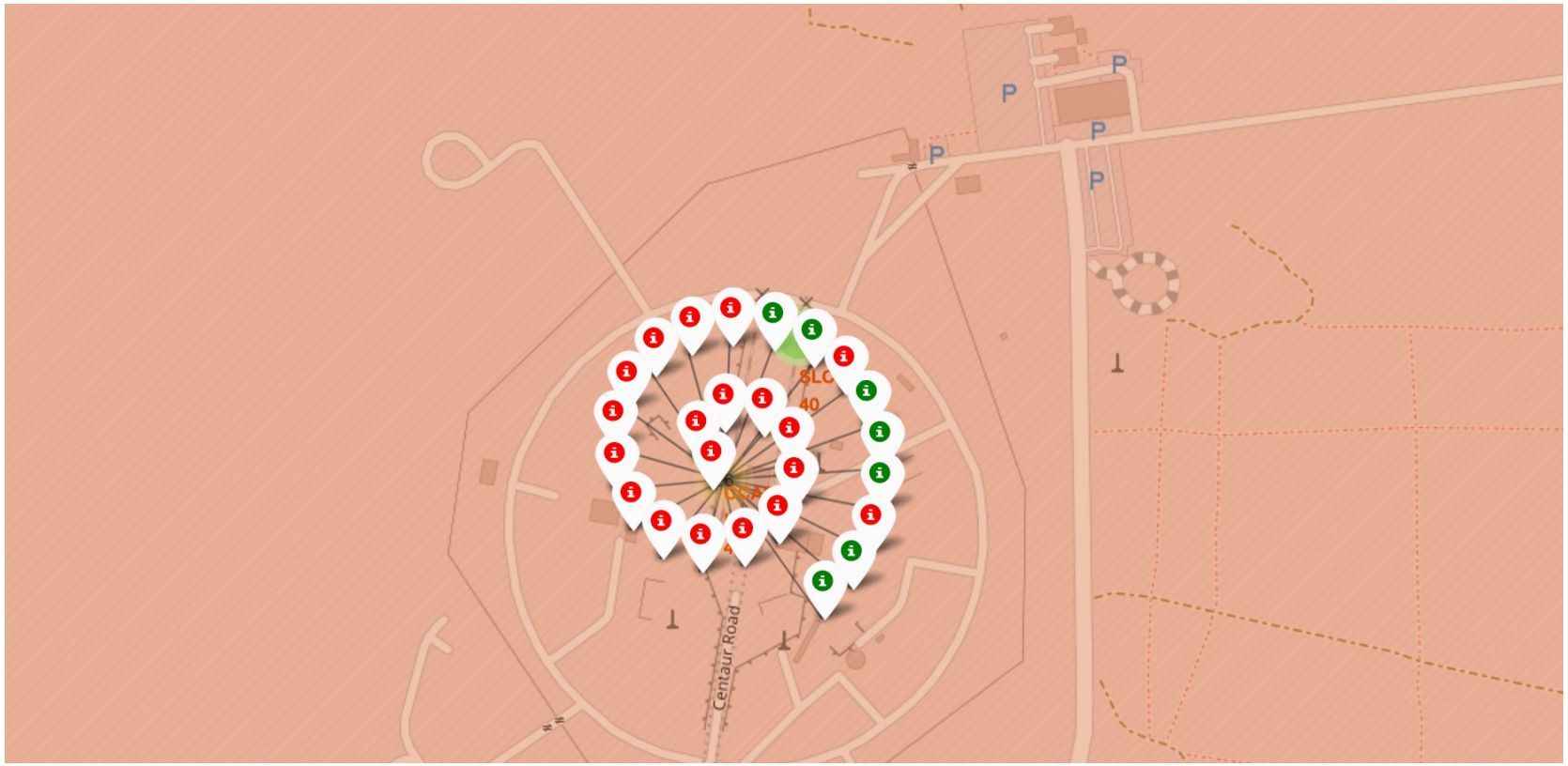
- All Launch Sites are located on coasts and near the earth equator
- By doing this one can take advantage of the earth's rotational speed

Color-labeled Outcomes for KSC



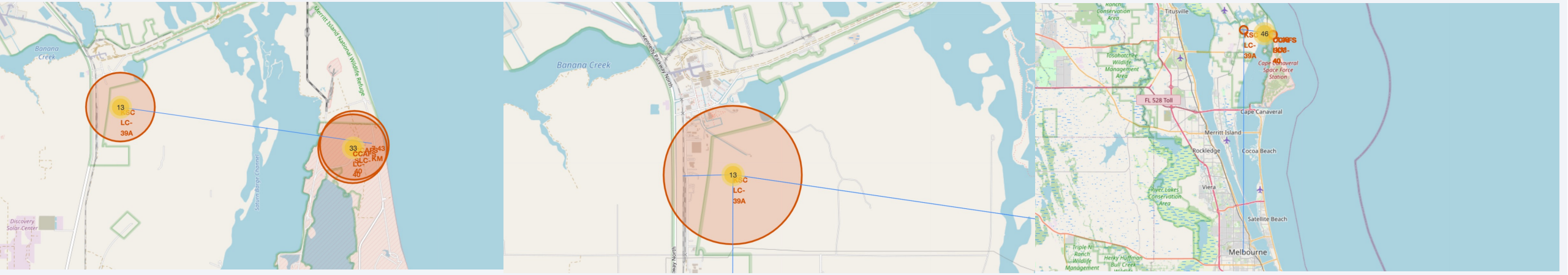
- Many successful launches from KSC with high Success rate
- This may be because the NASA with all it's experience and know-how is involved

Color-labeled Outcomes for CCAFS Sites

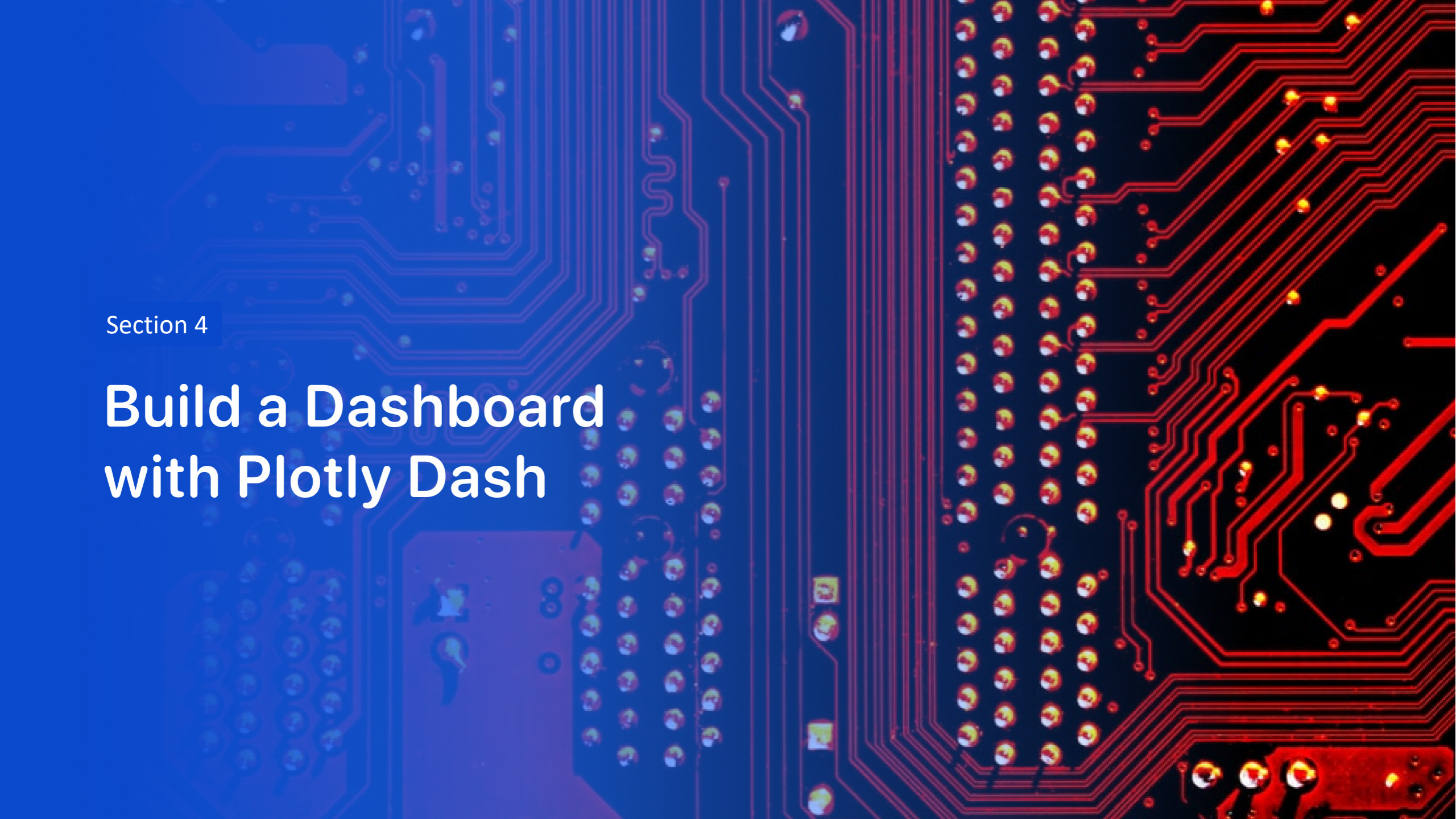


- Wheres the two CCAFS with close proximity to KSC have many failures
- A possible explanation is the inexpierence of SpaceX when launching from their own buil sites

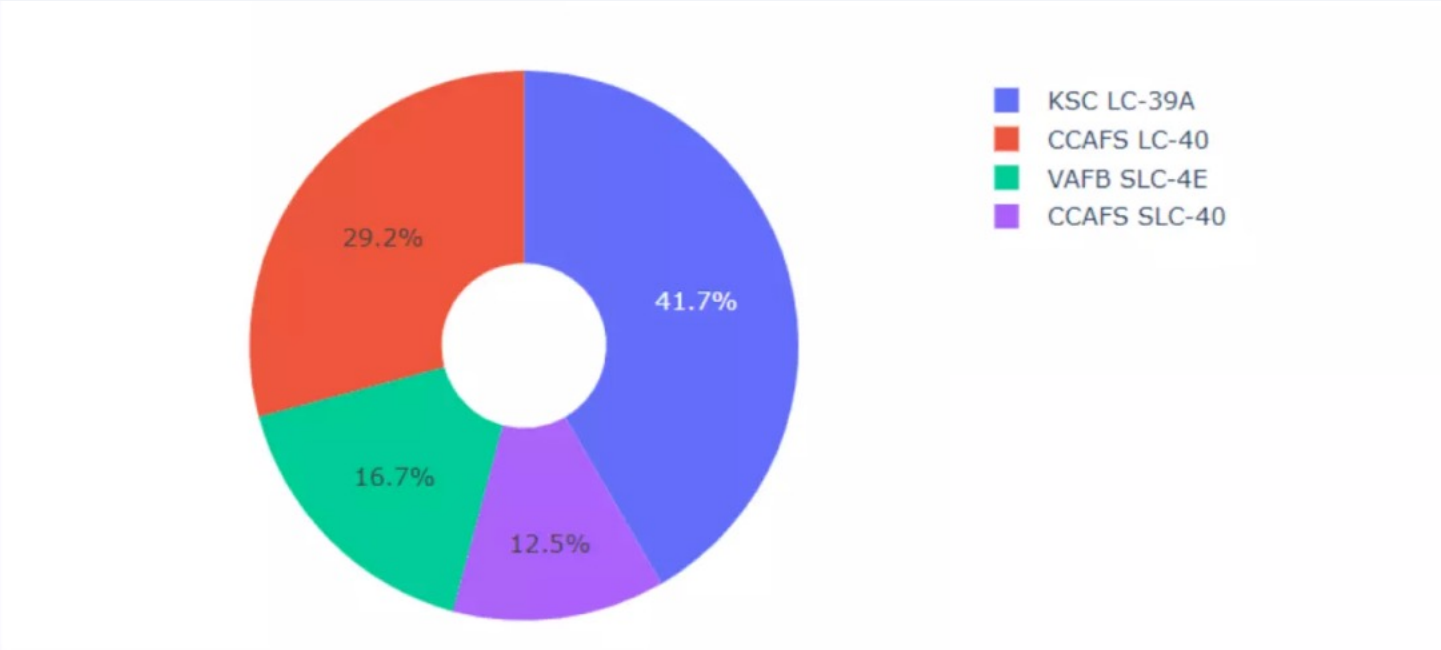
Proximities of KSC to nearby places



- Distance to coastline 7,4 km to crash in ocean in case of emergency
- Distance to railway only 0,7 km to facilitate shipment of large parts
- Distance to next bigger city 52 km due to safety reasons

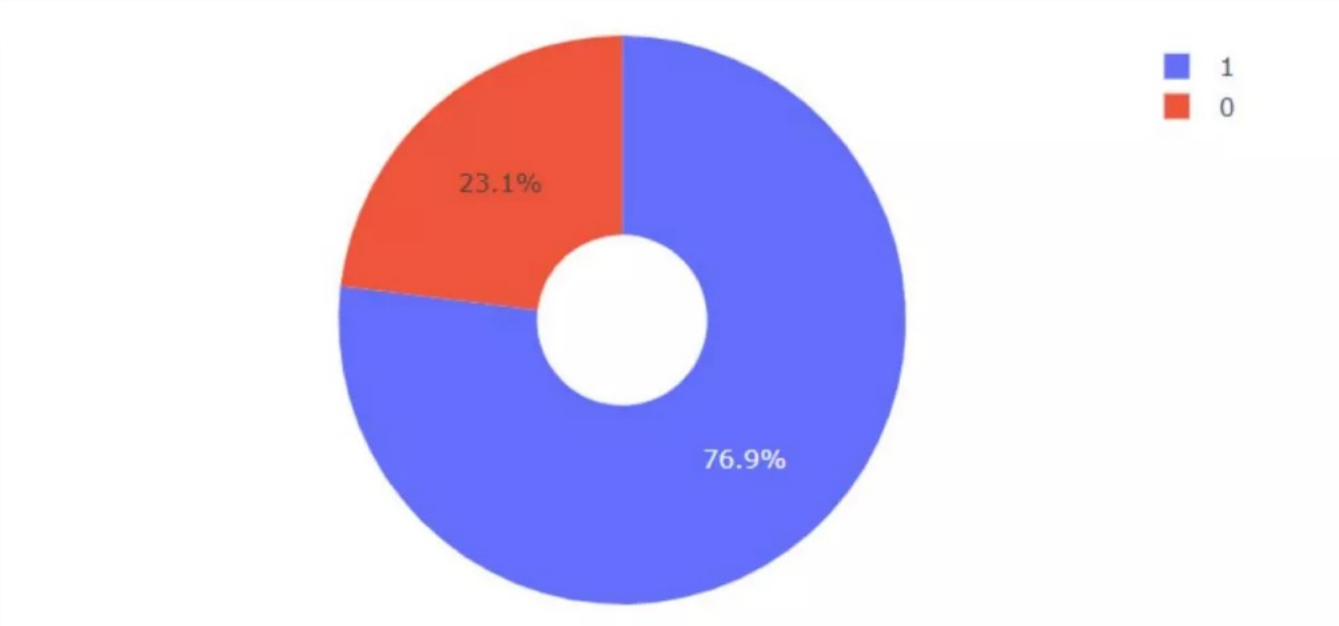

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by all Sites



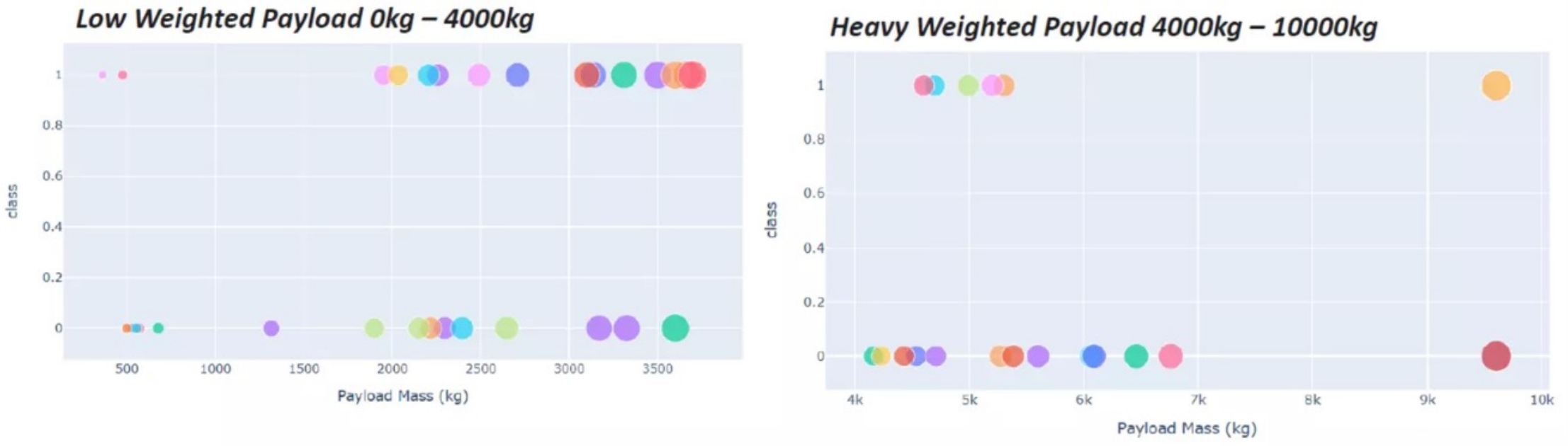
- KSC has to biggest share of Succes Launches and CCAFS SLC-40 the smallest

Success rate of KSC



- While having the biggest share of success launches KSC also has the highest success rate
- This means that KSC is overall the safest and best launch sites
- As already mentioned this is most likely because the Kennedy Space Center is a government facility of NASA with decades of expierence and the worlds best scientists and engineers

Payload vs. Outcome

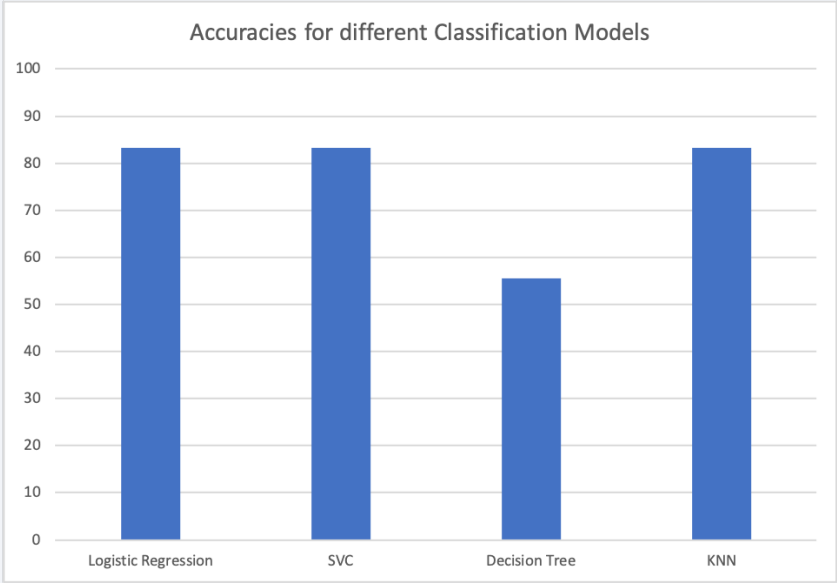


- Majority of payloads in the range between 2000 kg and 7000kg
- Its seems like payloads between 2000 kg and 4000 kg are correlated with a successful outcome
- Whereas payloads between 4000 kg and 7000 kg often result in failed outcomes

Section 5

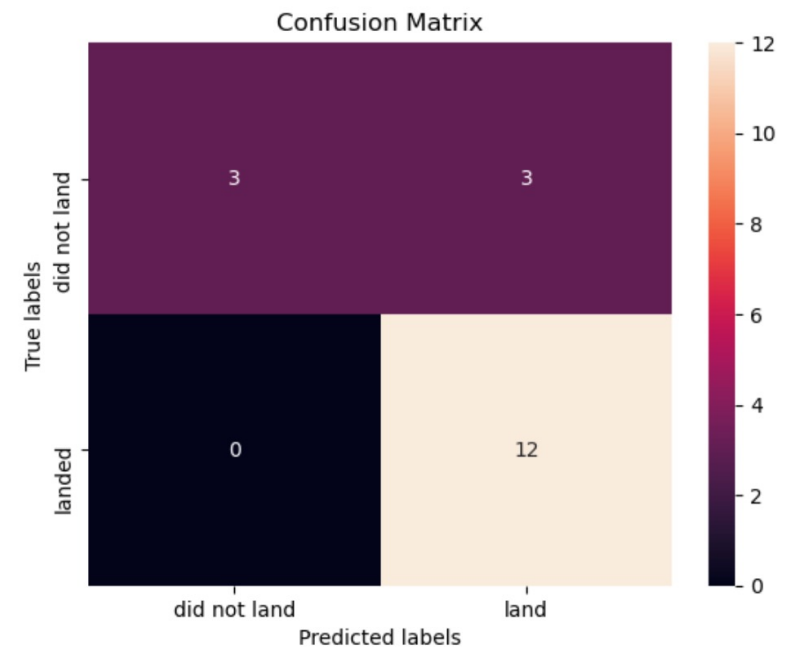
Predictive Analysis (Classification)

Classification Accuracy



- Logistic Regression, SVC and KNN all have the accuracy of 83,33%
- Decision Tree algorithm not well suited for this problem since accuracy is only 55%

Confusion Matrix for LR, SVC and KNN



- The modell is good in predicting the true positives e.g. the rocket landed
- It does relatively poorly when the rocket did not land. It got 3 false outputs and predicted that the rocket landed

Conclusions

- KNN, SVC and Logistic Regression all are equally suited for the classification task
- Accuracy is relatively low with 83,33% -> more training data needed
- Kennedy Space Center (KSC LC 39A) has the highest Success rate of all Sites
- SpaceX learns from trial and error and so the Success rate is rising with years
- Payloads heavier than 4000 kg are more difficult to launch successfully
- Orbits with highest Success rates are ES L1, GEO, SSO, HEO
- https://github.com/HakHak34/master

Appendix

- GitHub repository: https://github.com/HakHak34/master
- KSC image: https://www.google.de/search?q=kennedy+space+center&sxsrf=ALiCzsY022H8DFFtT0esZMKzTO9kKi9YVQ:1667131194128&source=lnms&tbm=isch&sa=X&ved=2ahUKEwiBvPrU84f7AhXdgf0HHbfWDigQ_AUoAnoECAIQBA&biw=1089&bih=694&dpr=2 - imgrc=CKE_vGGTGwVMiM

Thank you!