



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bakonyi-Kiss Gyula
2022.10.25.



Outline

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

Executive Summary

In this final project the task was to predict if a rocket's first stage can land successfully.

By identifying and utilizing features that increase the chances of a successful landing, approximately a 100 million dollars can be saved per rocket launch, if the first stage lands successfully.

To predict landing success, data science methodologies and machine learning algorithms are used. This was realized in the following phases: Data Gathering, Data Wrangling, Exploratory Data Analysis, Visualization and finally Prediction.

During the analysis, we have found features that have a correlation with the landing success.

Introduction

As a newly recruited data scientist at a private space launch company, we have to pick on the big guy SpaceX, to gain some market.

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used by our company if we want to bid against SpaceX for a rocket launch.

So our main objective (or question) is if we can find useful information or feature amongst the launch data, from which we can predict the landing success of a rockets first stage.

Section 1

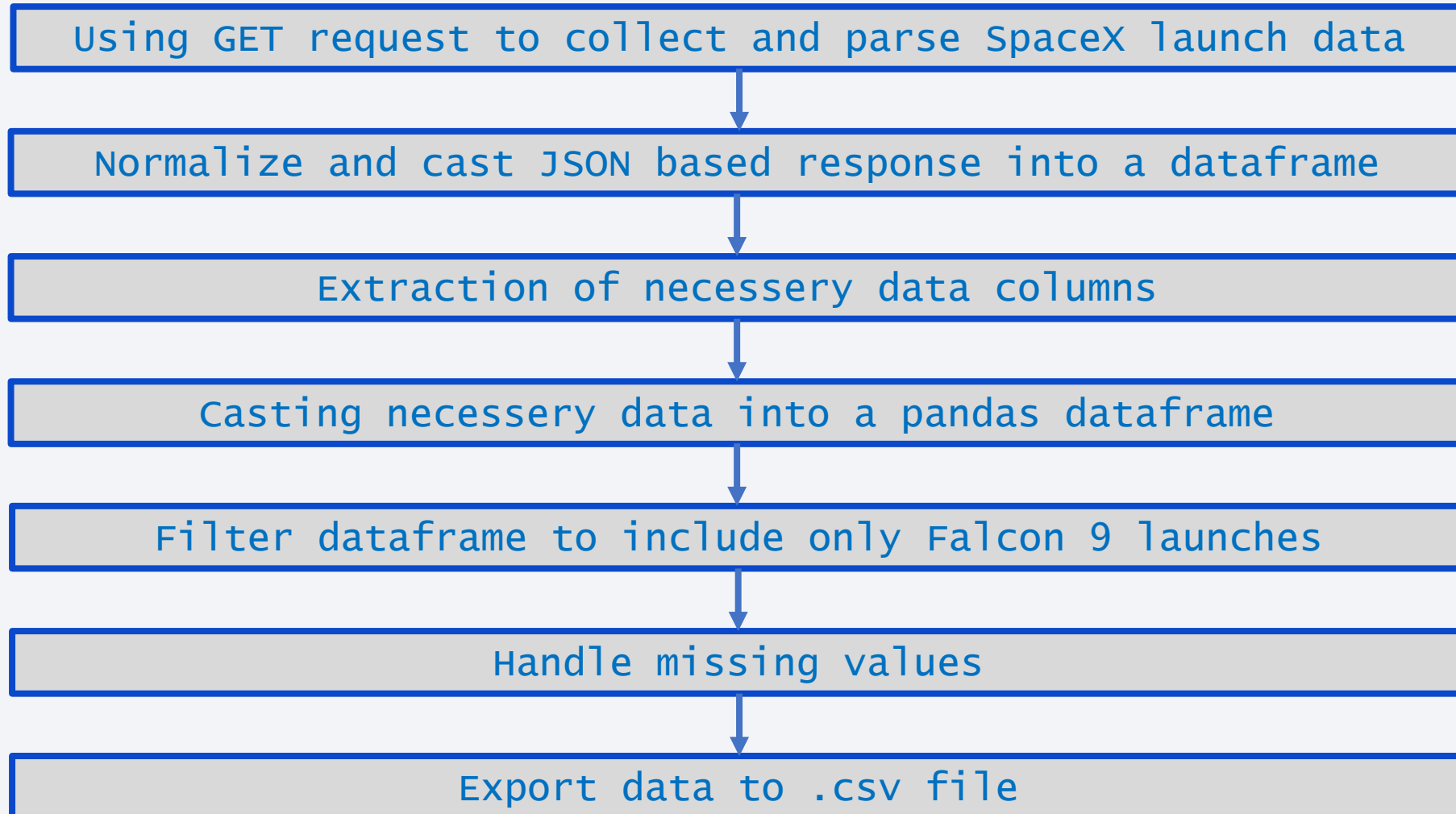
Methodology

Methodology

Executive Summary

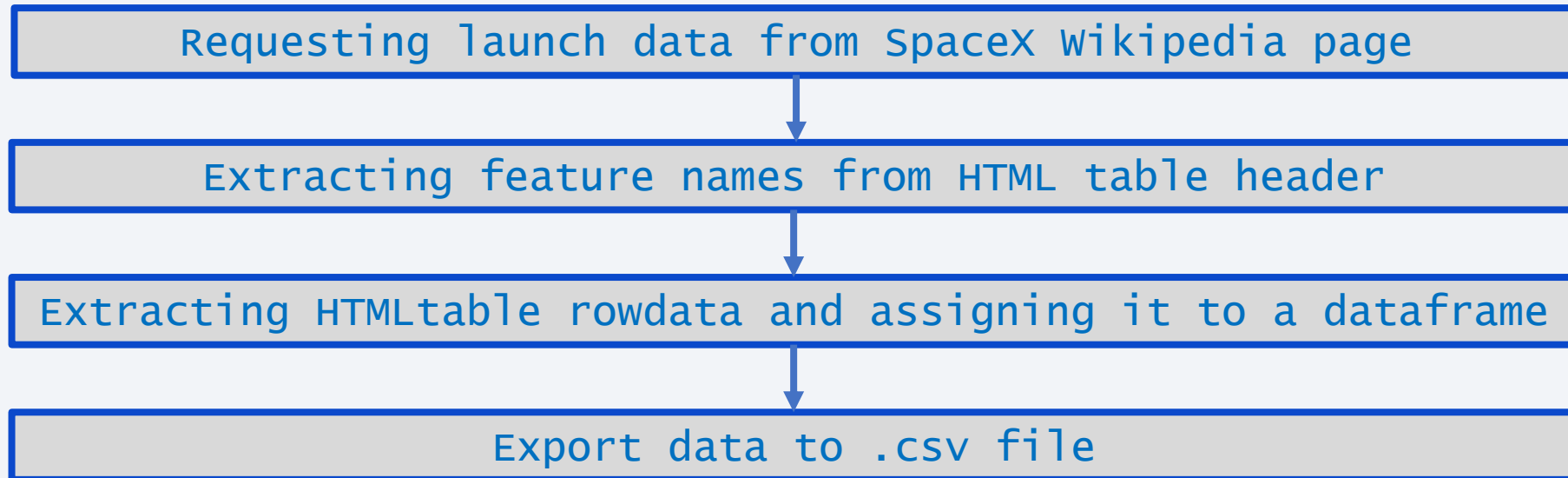
- Two ways were used to gather data. The first was using API requests from SpaceX, the other was webscraping from the wikipedia page of SpaceX.
- The next step was to clean data using Data wrangling techniques.
- After this data conditioning, data analysis (EDA) was performed. The main focus was on the visualization of the data, because "seeing" your data intuitively helps you to figure out connections not seen in pure numbers. Another way of analyzing data was to request ordered data using SQL queries.
- Data analysis ended with visualization of data using Folium, and supplemented by an interactive dashboard created in Plotly Dash.
- The final step was the predictive analysis using four types of classification models. Namely logistic regression, support vector machines, k-nearest neighbour and decision tree classifiers. Data was separated to train and teach datasets, and after that each model was trained, tuned and evaluated to find the best one.

Data Collection

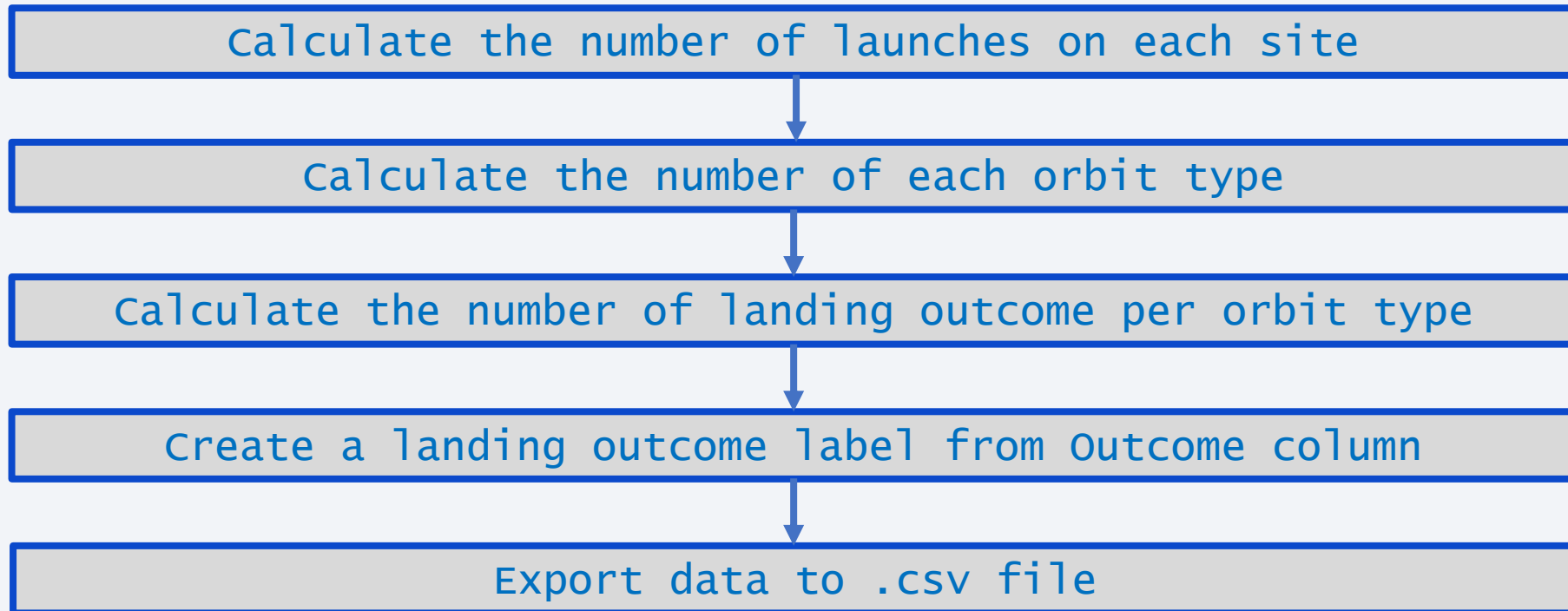


Detailed solution described here: [API guideline](#)

Data Collection - Scraping



Data Wrangling



EDA with Data Visualization

There are several types of charts/graphs implemented in Python's Seaborn package. It is important to choose the proper chart type for our data visualization, as it can highlight or erode the essence of what we want to show. During the EDA phase, 3 chart types were used:

- Line charts for showing data trends over time, like SuccessRate vs. NumberOfYears
- BarCharts to compare values of datagroups. X axis for categories and Y axis for values, like in the case of SuccessRate vs. OrbitTypes
- ScatterPlots for check the relationship between different sets of features, like FlightNumber vs. PayloadMass(kg)

EDA with SQL

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string „KSC”
- Displaying the total payload mass carried by boosters launched by NASA
- Displaying average payload mass carried by booster version F9 v1.1,
- Listing the date when the first successful landing outcome in drone ship was achieved
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the successful landing outcomes in ground pad, their booster versions, and launch site names for in year 2017
- Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

Build an Interactive Map with Folium

Different kind of objects were created and added to a Folium map:

- Circle objects were added to show all launch sites on the map
- Markers were added to show successful/failed launches for each site on the map.
- Line objects were used to calculate the distances between a launch site to its proximities

Examining these objects, we can answer the following questions:

- Are launch sites in close proximity to railways? Yes
- Are launch sites in close proximity to highways? Yes
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

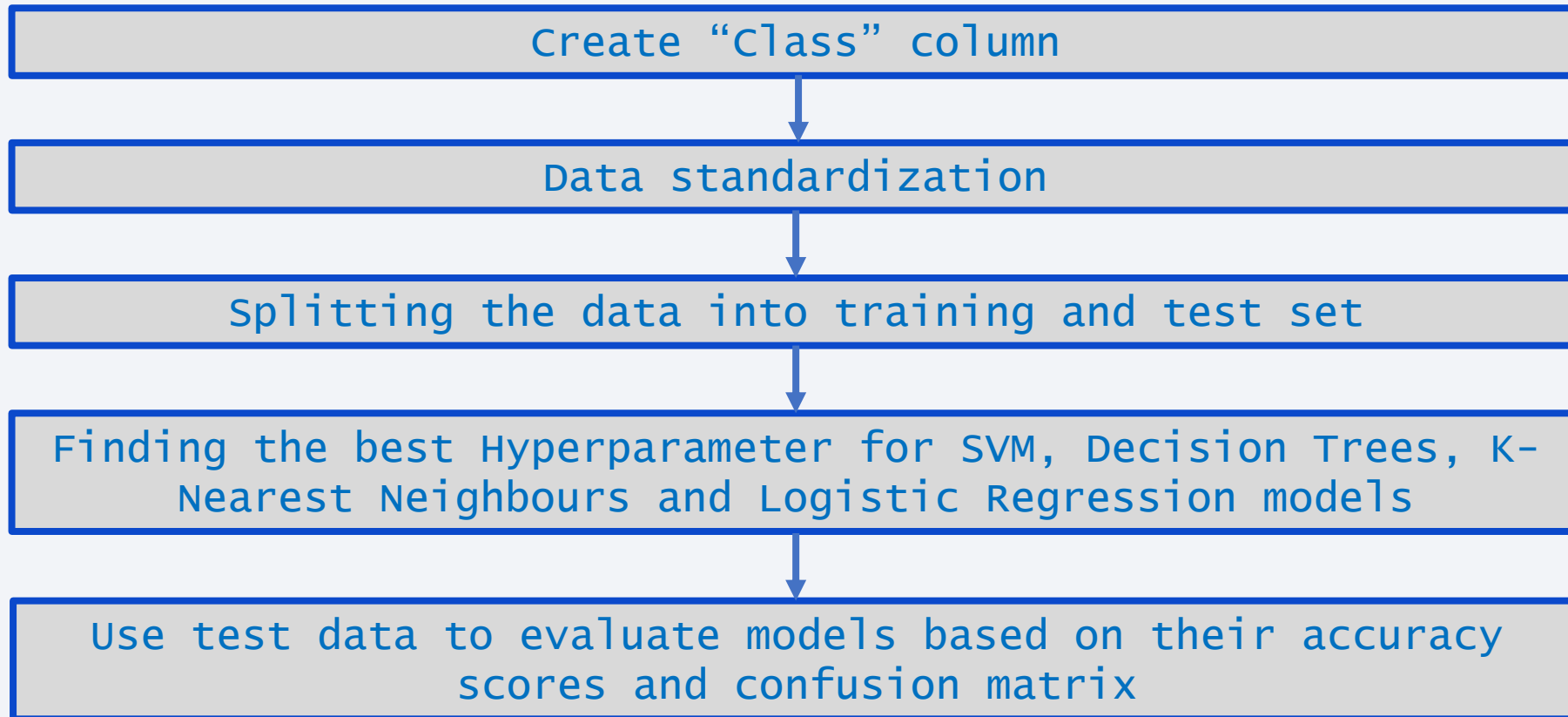
Build a Dashboard with Plotly Dash

The dashboard contains the following elements:

- **Dropdown list:**
 - It allows to choose between launch sites
 - It can also select all launch sites
 - Dividing data into smaller parts can help during interpretation
- **Pie chart:**
 - Shows the successful launch by each site, or all sites.
 - One can visualize the distribution of landing outcomes across all launch sites
 - It shows the success rate of launches on individual sites.
- **Slider:**
 - It allows to select a minimum and a maximum size for payload mass, which can influence the outcome of the landing
- **Scatter plot:**
 - it shows the relationship between landing outcomes and the payload mass of different boosters.
 - Easy to understand visualization how different payload mass affects the landing outcomes

Detailed solution described here: [SpaceX Dash](#), [Task1](#), [Task2 1](#), [Task2 2](#), [Task3](#), [Task4](#)

Predictive Analysis (Classification)



Results

EDA analysis results show, that the success rate of the landings was 66.66%

The predictive analysis results showed that all of the classification algorithms resulted in an accuracy of 83,3%

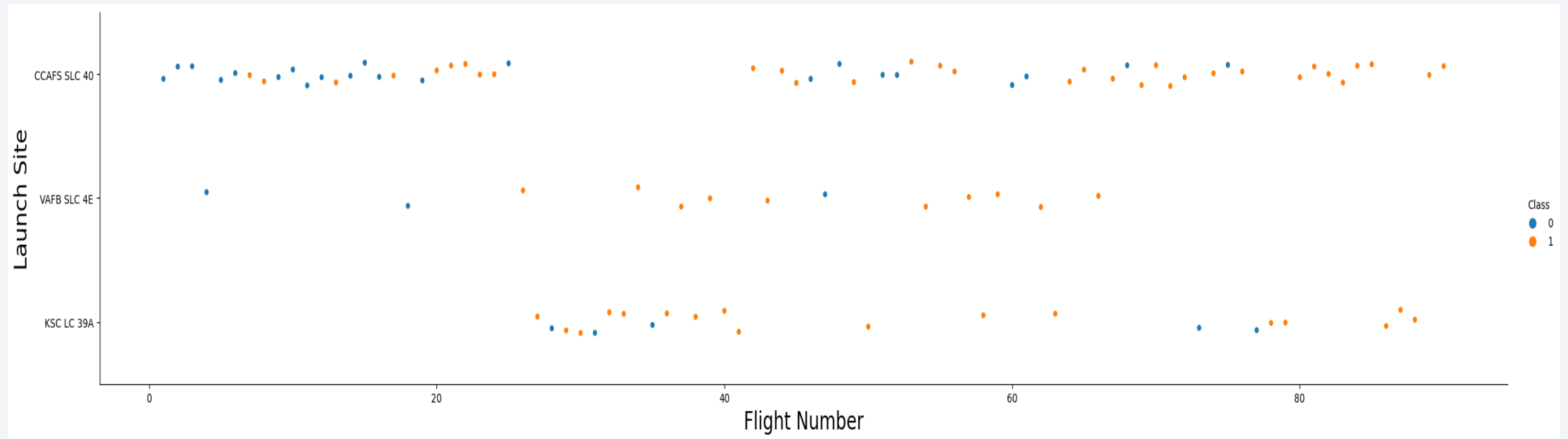
Interactive analytics (Dash) screenshots: [Task1](#), [Task2_1](#), [Task2_2](#), [Task3](#), [Task4](#)

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

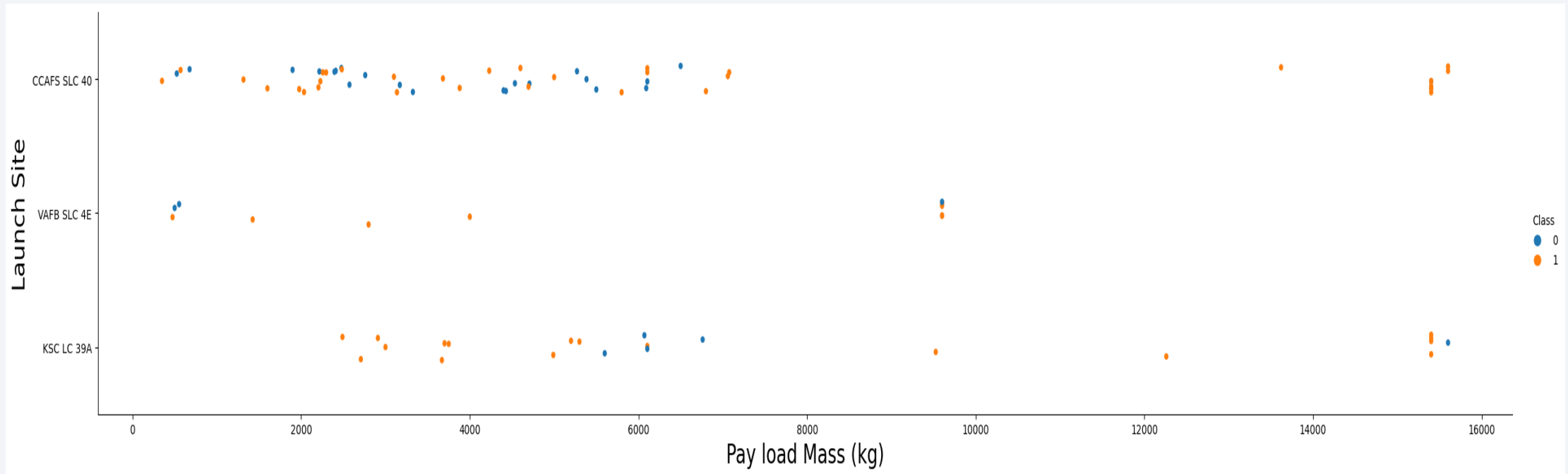
Insights drawn from EDA

Flight Number vs. Launch Site



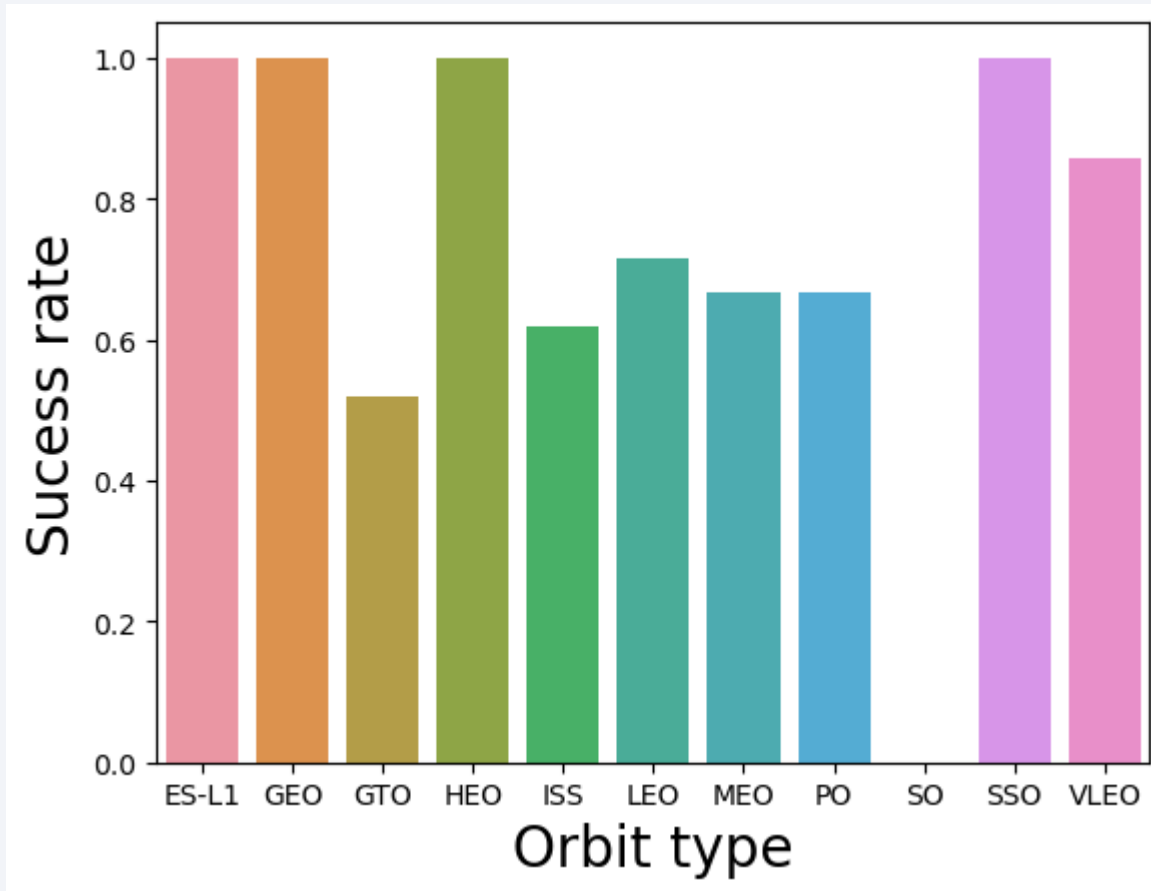
- Blue dots represent the failed landings, while the red dots represent the successful ones.
- The plot shows that the success rate increased as the number of flights increased.
- It can be seen, that the first 6 landings failed, while there are no failed landings after the 80th Flight number
- There seems to be an increase in successful flights after the 40th launch.

Payload vs. Launch Site



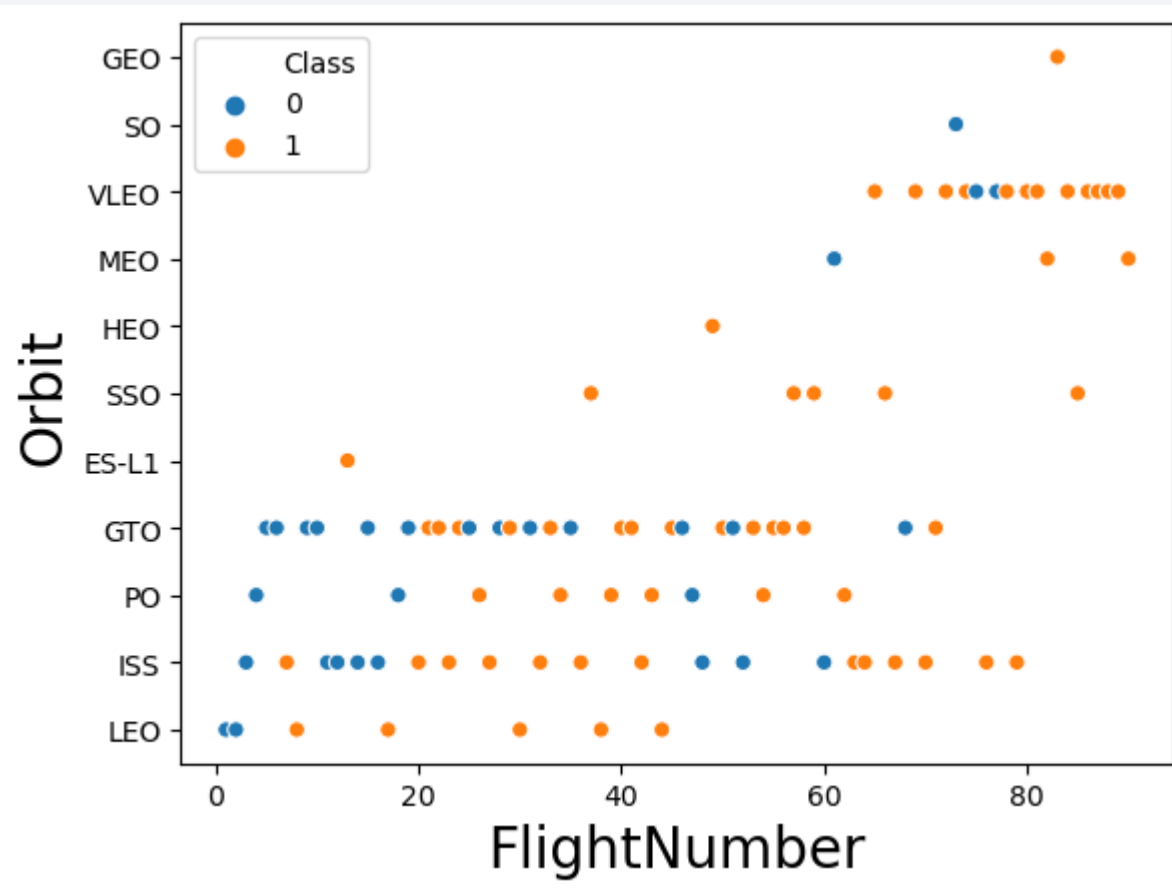
- Blue dots represent the failed landings, while the red dots represent the successful ones.
- The plot shows that heavy payloads (mass > 10000 kg) have a very high successful landing rate.
- For the VAFB-SLC launchsite there are no rockets launched with heavy payload mass
- For small (mass < 10000 kg) payload mass there is only weak correlation between „launch site” and „pay load mass (kg)” features, so this metric is difficult to use for prediction purposes

Success Rate vs. Orbit Type



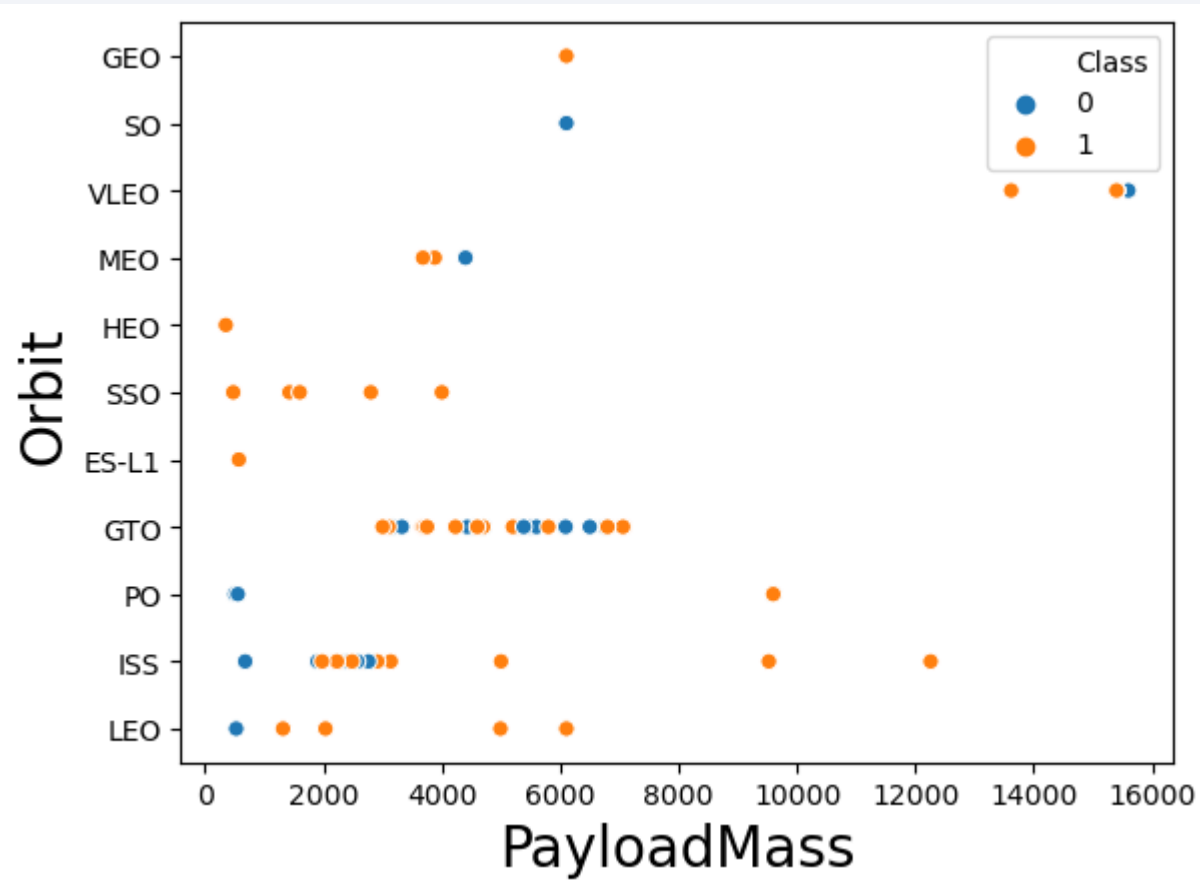
- Orbit types ES-L1, GEO, HEO, SSO have 100% landing success rates.
- Orbit type SO did not have any successful landings at all.

Flight Number vs. Orbit Type



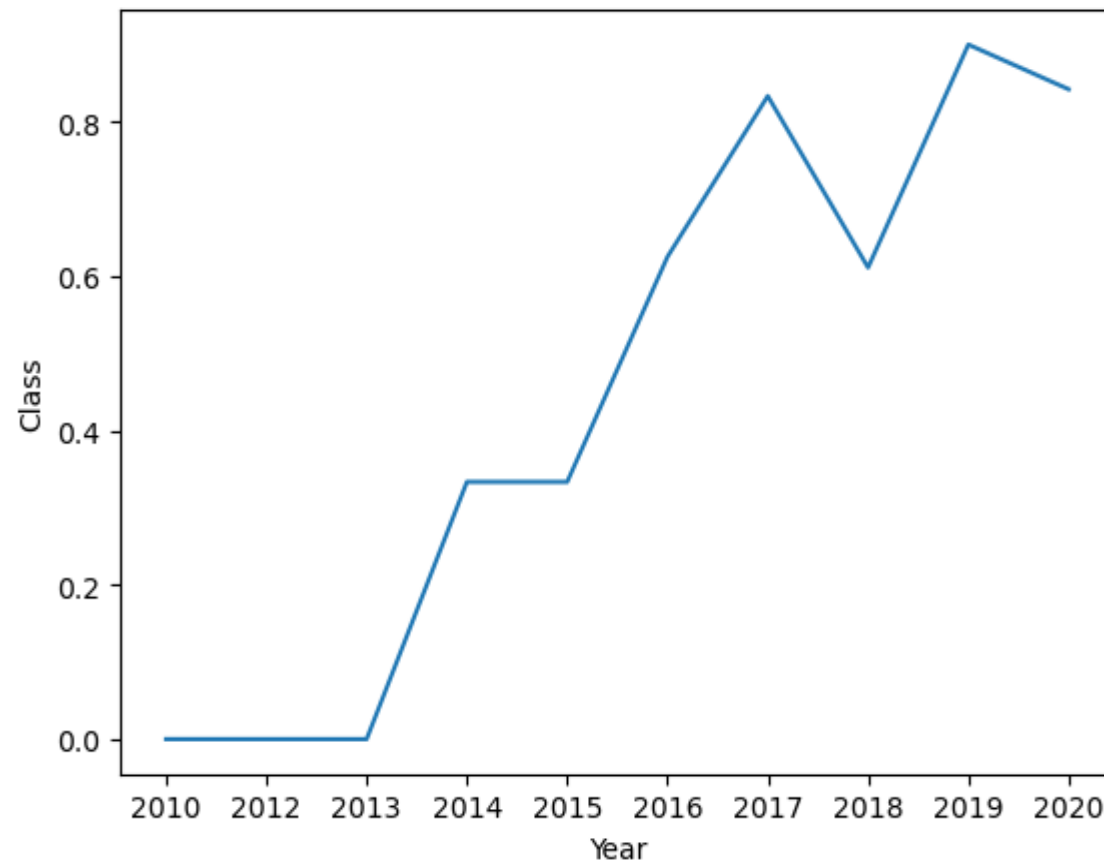
- Orbit types ES-L1, GEO, HEO, SSO have 100% landing success rates, but the number of flights are very few compared to other orbit types.
- Orbit type SO did not have any successful landings at all, but the number of flights are very few compared to other orbit types.
- In the case of Orbit Type LEO there is a positive correlation between success rate and flightnumber
- In the case of Orbit Type GTO, there is no correlation between success rate and flightnumber

Payload vs. Orbit Type



- This examination holds no new information, the statements are the same as in the case of Slide 18 (Payload vs. Launch Site) and Slide 20 (Flight Number vs. Orbit Type).
- In the case of Orbit Types LEO, PO there is a positive correlation between orbit type and payload mass

Launch Success Yearly Trend



- Landing success rate is increasing as time passes.
- There are two years (2018 and 2020), where there is a fallback in success rate.

All Launch Site Names

launch_site

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

- SQL queries were used to get the launch site names from the database
- „Distinct” keyword was used, to remove duplicate launch site names, and list only unique names
- The 4 unique launch site names can be seen on the picture to the left

Launch Site Names Begin with 'KSC'

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

- „Like” keyword and „%” sign was used to search for database rows, where the launch site name started with the „KSC” phrase
- „Limit” keyword was used to show only the first five results

Total Payload Mass



1
45596

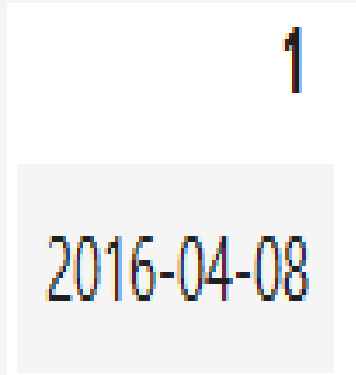
- The total payload mass measured in kg can be seen in the picture to the left
- This is the result of using the „SUM” function on the „payload_mass_kg” column of the dataset

Average Payload Mass by F9 v1.1



- The average payload mass measured in kg can be seen in the picture to the left
- This is the result of using the „AVG” function on the „payload_mass_kg” column in case the „Booster_version” column was equal to „F9 v1.1”.

First Successful Ground Landing Date



1
2016-04-08

- The dates of the first successful landing on drone ship can be seen in the picture to the left
- This is the result of using the „MIN” function on the „Date” column in case the „where” function used on the „Landing_outcome” column equals „Success (droneShip)”

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- The names of booster types which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 can be seen in the picture to the left
- The "Between" keyword was used to get only datarows of payload mass greater than 4000 but less than 6000.
- The „where" keyword was used to filter out the results to include only boosters which successfully landed on drone ship

Total Number of Successful and Failure Mission Outcomes



- The the total number of successful and failure mission outcomes can be seen in the picture to the left.
- The „Count” and „Like” keywords were used to count the number of occurrences of different mission outcomes.

Boosters Carried Maximum Payload

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 names of the boosters which have carried the maximum payload mass can be seen in the picture to the left.
- A subquery was created using the „Max” keyword to retrieve the maximum payload mass
- The „where keyword was used to get the list of boosters where the payload weight was the maximum payload mass

2017 Launch Records

month_name	landing_outcome	booster_version	launch_site
FEBRUARY	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
MAY	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
JUNE	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
AUGUST	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
SEPTEMBER	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
DECEMBER	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- The records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017 can be seen in the picture to the left.
- The „Select” and „where” keywords were used to get multiple columns from the table
- The „To_Char” keyword was used to convert the months of dates into string format
- The „Like” keyword was used to filter dates containing 2017

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

DATE	COUNT
2015-12-22	1
2016-04-08	1
2016-05-06	1
2016-05-27	1
2016-07-18	1
2016-08-14	1
2017-01-14	1
2017-02-19	1

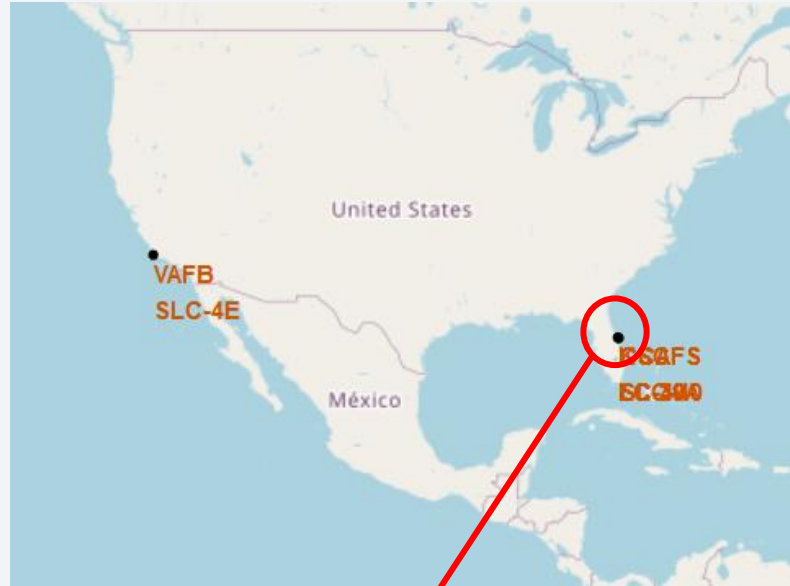
- The ranking of the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order can be seen in the picture to the left.
- The „Count” keyword was used to count the different landing outcomes.
- The „where” and „Between” keywords filtered the results to only include dates between 2010-06-04 and 2017-03-20.
- The „Groupby” keyword ensured that the counts were grouped by their date.
- The „Orderby” and „Desc” keywords were used to sort the results by descending order

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

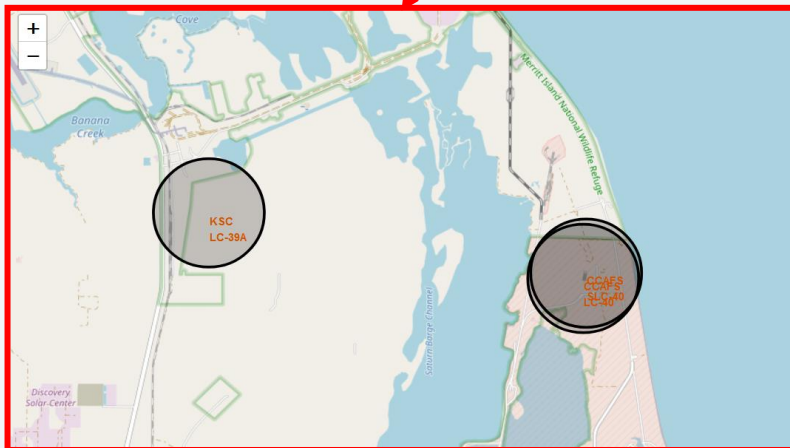
Section 3

Launch Sites Proximities Analysis

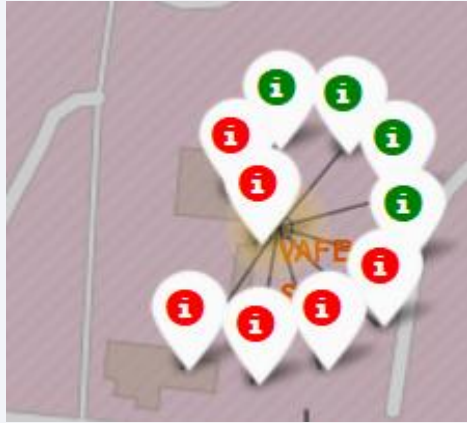
Launch Sites



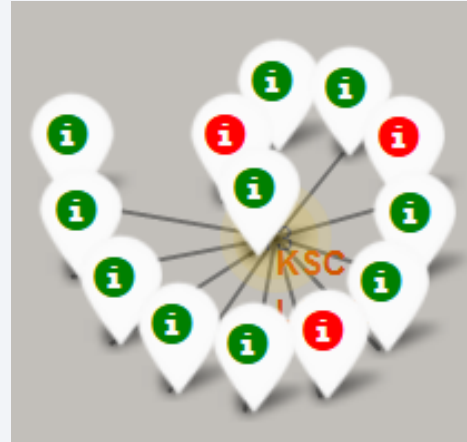
- The black markers indicate the locations of all SpaceX launch sites in the USA.
- The launch sites are near the coast.
- There are launch sites on the both sides (East coast / West coast) of the country.



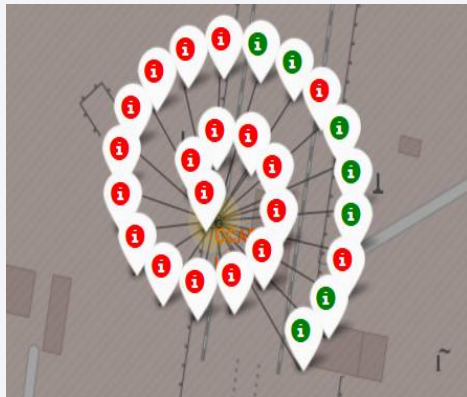
Landing outcomes of launch sites



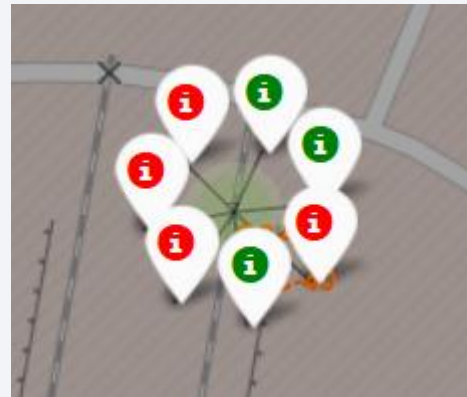
VAFB SLC-4E



KSC LC-39A



CCAFS-LC40



CCAFS-SLC40

- Green markers show successful landing outcomes
- Red markers show failed landing outcomes

Launch site's proximities



Examining launch site proximities, we can conclude that:

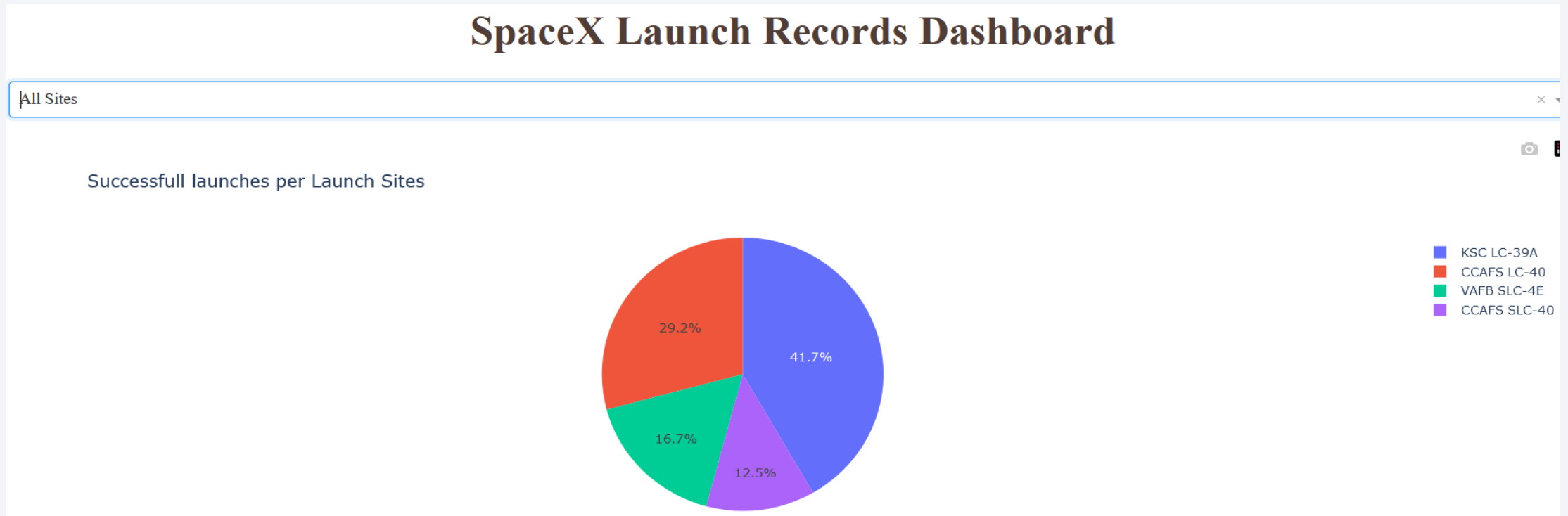
- Launch sites are in close proximity to railways (for effective transportation of materials and goods)
- Launch sites are in close proximity to highways (for effective transportation of human resources)
- Launch sites are close proximity to coastline (for easy reclamation of returned rocket first stages)
- Launch sites keep distance from cities (to not to disturb local citizens)



Section 4

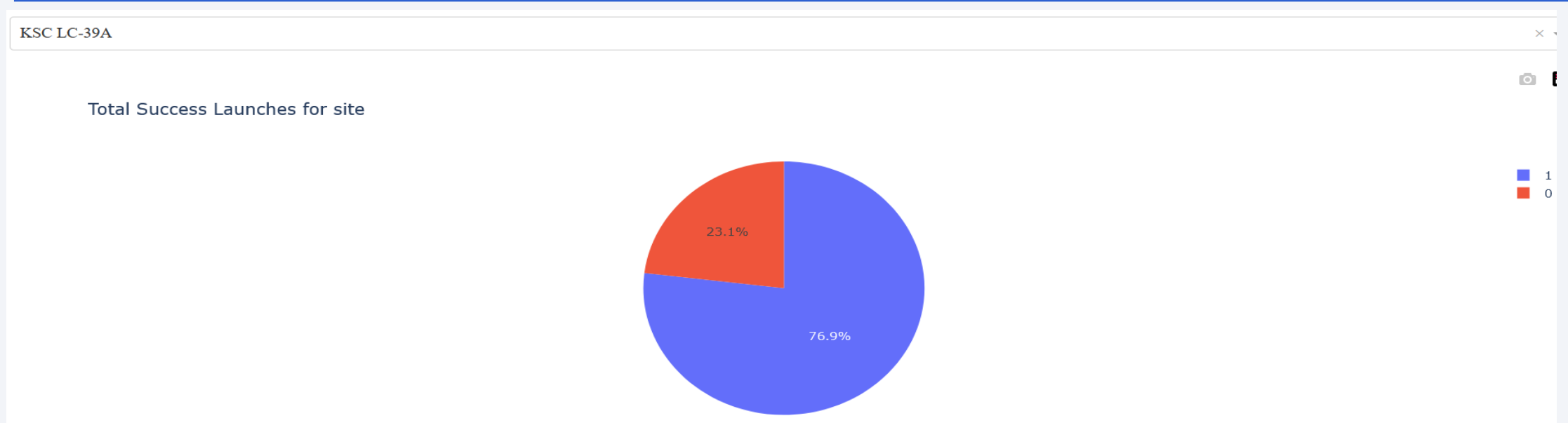
Build a Dashboard with Plotly Dash

Successfull landings per launch sites



- KSC LC-39A Launch site has the most successful landings (10 in total)
- All four launch sites had successful landings

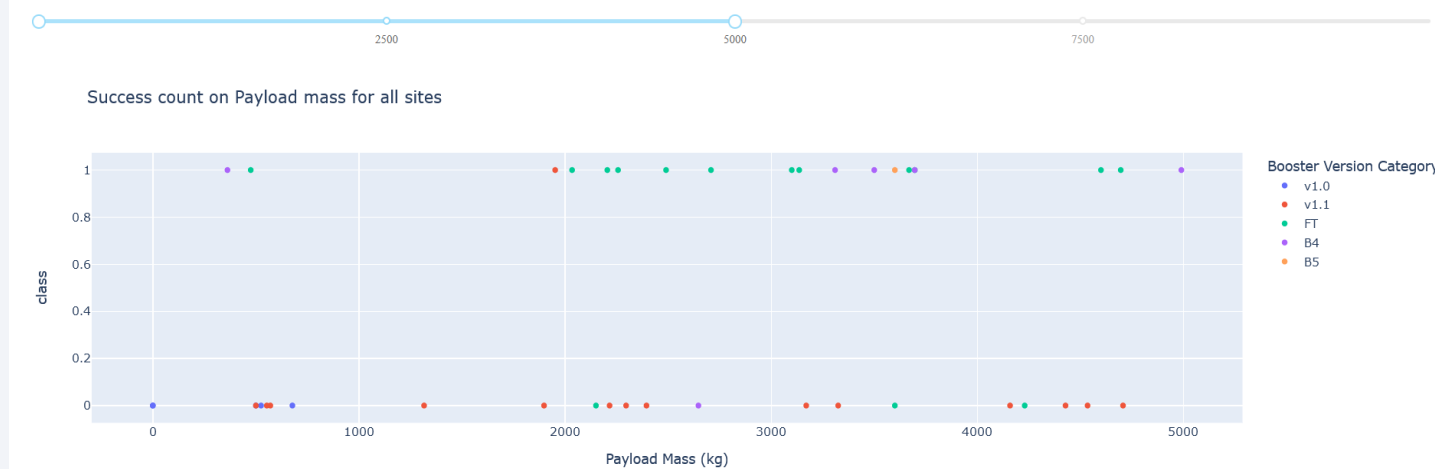
Most successfull launch site



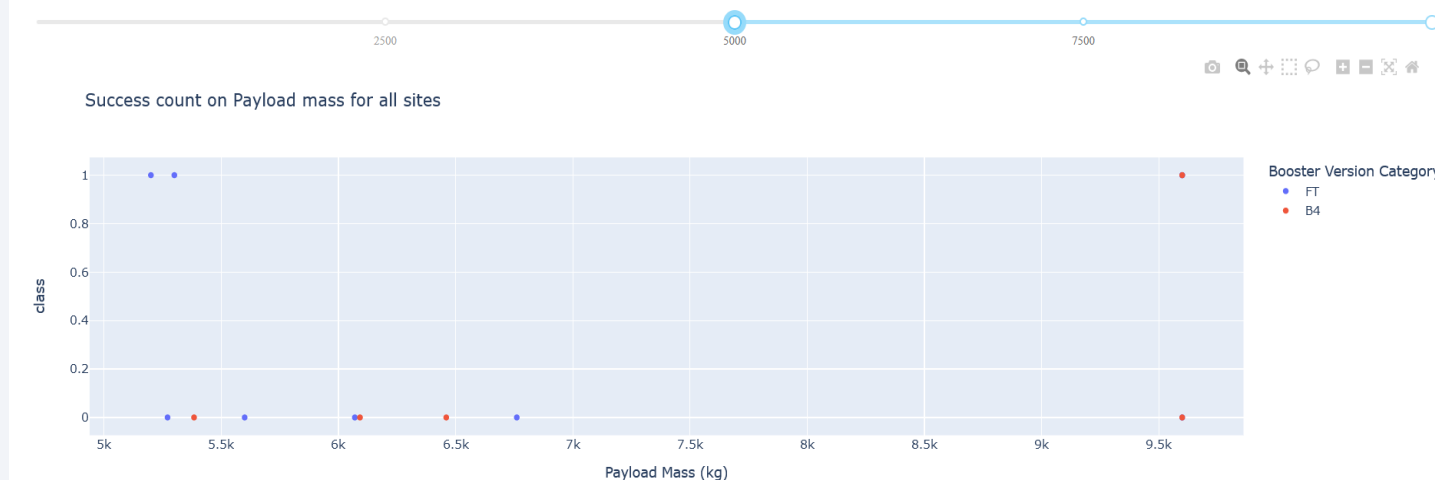
- The most successfull launch site was the KSC LC-39A Launch site with a 76.9 % success ratio.

Payload vs. Launch Outcome

Payload range (Kg):



Payload range (Kg):

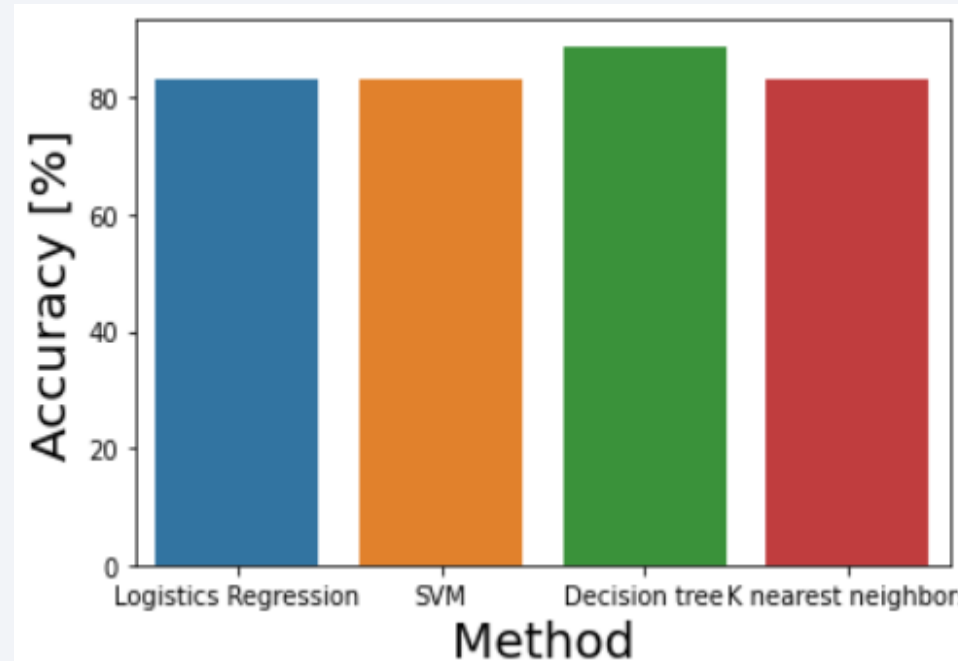


- The launch success rate for payloads 0-5000 kg is much higher than that of payloads 5000-10000 kg.
- Sadly this higher success ratio is still under 50% 😞
- Not only success ratio, but flightnumber count is also much higher for the 0-5000kg payloads than that of 5000-10000kg.
- The booster version that has the largest success rate in both of the ranges is the FT.

Section 5

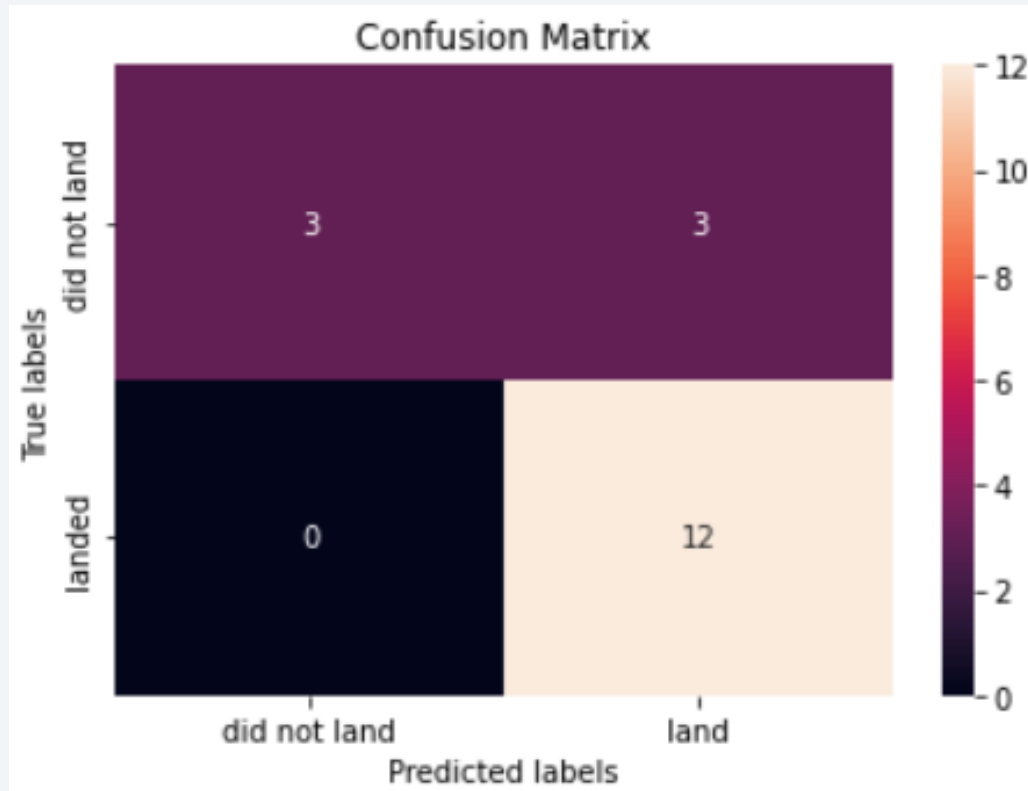
Predictive Analysis (Classification)

Classification Accuracy



- Accuracy for Logistics Regression method: 0.8333333333333334
 - Accuracy for Support Vector Machine method: 0.8333333333333334
 - Accuracy for Decision tree method: 0.8888888888888888
 - Accuracy for K nearest neighbors method: 0.8333333333333334
-
- The Decision Tree classifier had the best accuracy at 88.8%.

Confusion Matrix



- The model predicted 12 successful landings when the True label was successful (True Positive) and 3 unsuccessful landings when the True label was failure (True Negative).
- The model also predicted 3 successful landings when the True label was unsuccessful landing (False Positive).
- The model generally predicted successful landings.

Conclusions

- The analysis showed that there is a positive correlation between number of flights and success rate as the success rate has improved over the years.
- There are certain orbits like SSO, HEO, GEO, and ES-L1 where launches were the most successful.
- Success rate can be linked to payload mass as the heavier payloads generally proved to be more successful than the lighter payloads.
- The launch sites are strategically located near highways and railways for transportation of personnel and cargo, but far away from cities for safety.
- The best predictive model to use is the Decision Tree Classifier as it had the highest accuracy with 88.8%.

Appendix

- GitHub Repository: <https://github.com/bkgy/PythonCapstoneProject>

Thank you!

