IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Patrik Antal
2022.04.27.

Outline

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

Executive Summary

- For predicting landing outcomes, I have collected data from more sources and done EDA for discovering features that do affect the success of landing. By building and optimizing classification algorithms using grid search, we were aiming for the highest possible accuracy.
- Throughout the process we were able to discover useful relationships between input features that can help increase the landing process's success. By training a SVM model on this data we could reach 1.0 accuracy.

Introduction

- Falcon 9 is a two-stage-to-orbit medium-lift launch rocket that is a product of SpaceX. The company is offering air transportation for commertional purposes with much less cost than its close competitors. This is mainly because its ability to land its first stage rocket part and reuse it for later launches.
- This unique capability is gaining the company undebatable advantages on the market. If we, or our company is planning to compete with SpaceX is would be important to get to know the key factors needed for landing our first-stage rocket part. However publicly available data can be accessed directly form SpaceX with detailed launch information. Using this we can identify success factors of rocket landing and we can even build machine learning models that can predict the possibility of successful landing outcomes, which comes handy for price calculation.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the SpaceX REST API with different endpoints. There were Wikipedia pages scraped as well to get a more detailed, therefore more reliable dataset.
- Perform data wrangling
  - The columns of interest were selected and cleaned from null values. Also, values were converted for easier processing.
- 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

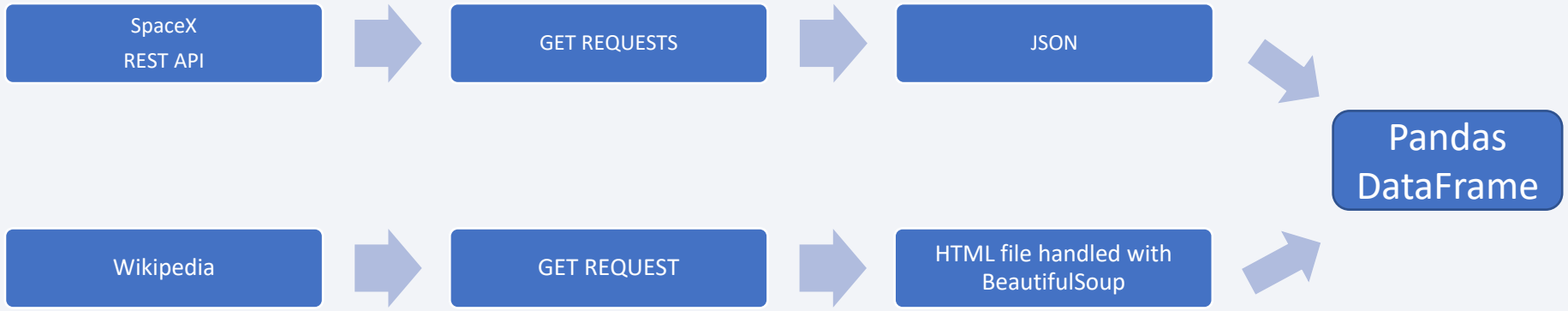
- Datasets were collected in two different ways.

a) SpaceX's **REST API** with **different endpoints**

- **main table's source:** https://api.spacexdata.com/v4/launches/past.

b) **Web scraping** with the BeautifulSoup library

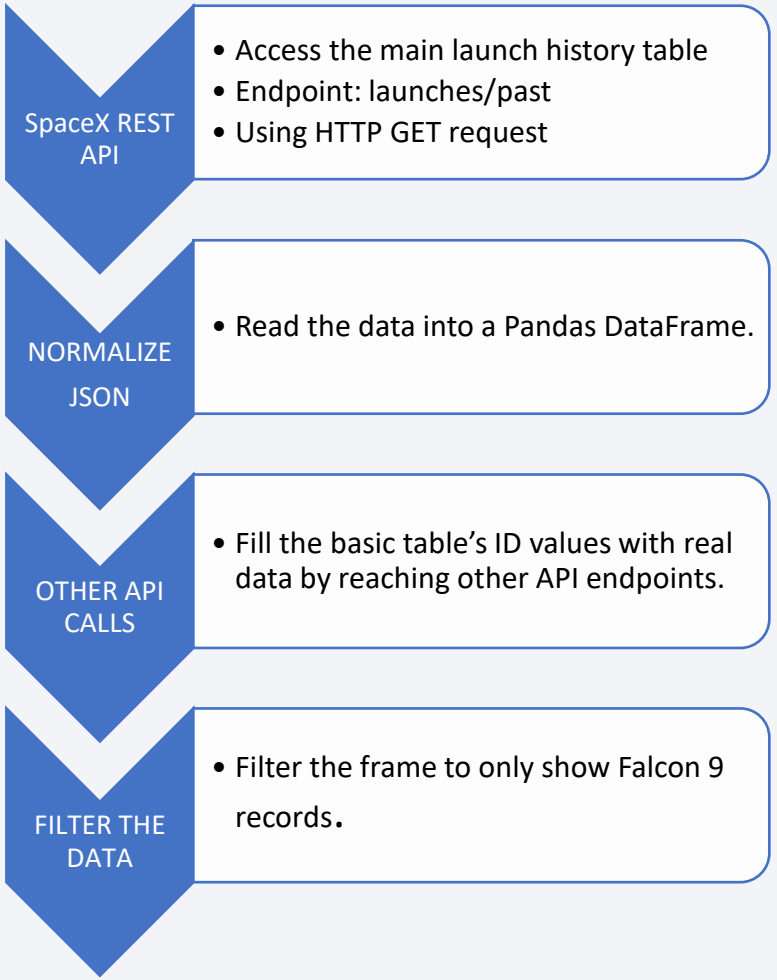
- **source:** https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches



Data Collection – SpaceX API

- SpaceX data can be accessed through several API endpoints
- Using GET requests we can derive data in JSON format
- JSON is easily convertible to Pandas DataFrame.

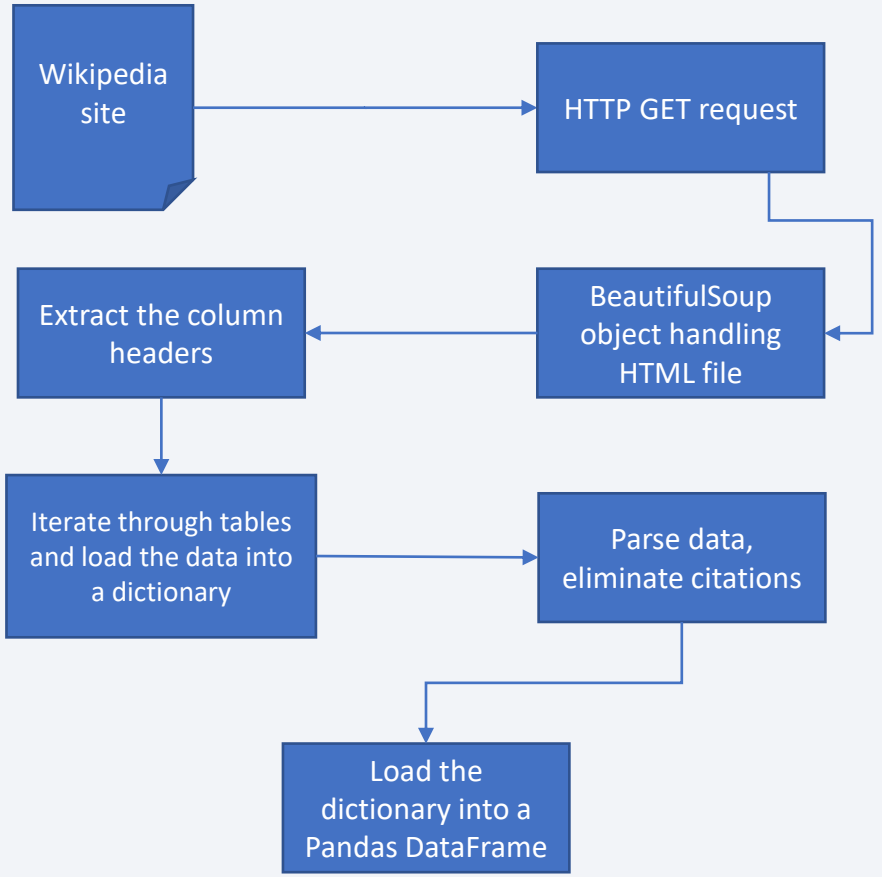
- For accessing the notebook click here: GitHub link



Data Collection - Scraping

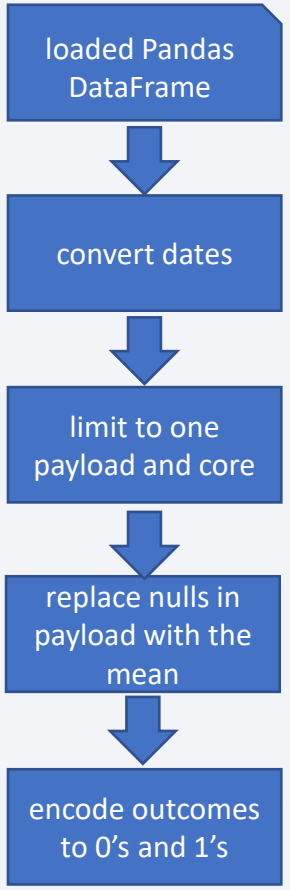
- Scraping Wikipedia gives another valuable data source.

- For accessing the notebook click here: GitHub link



Data Wrangling

- Data wrangling means converting and cleaning the data.
- In our case:
  - parse and convert values
  - filter for Flacon 9 launches
  - convert outcomes to 0's and 1's
  - eliminate nulls in PayloadMass
- For accessing the notebook click here: GitHub link



EDA with Data Visualization

- It is wise to make sense of the data before getting dirty with it.
- There were different graphs used for distinct purposes
  - to visualize two different variables affect on the outcome (also their relationship) **scatter plots** were used
  - to see the yearly trend of success rate we plotted a **line chart**
  - the success rate for each orbit could be compared best using **bar chart**
- For accessing the notebook click here: GitHub link

Feature Engineering

- Using the information gathered form EDA we can **select** and **convert** variables that we think is important for predicting landing outcome. Doing so will increase the performance of the desired ML algorithm.
  - **limit** the input range to 13 features → with presumed predictive power
  - **one-hot-encode** categorical variables
  - cast everything to **floating point numbers** → ML algorithms only accept numerical data

- For accessing the notebook click here: GitHub link

Build an Interactive Map with Folium

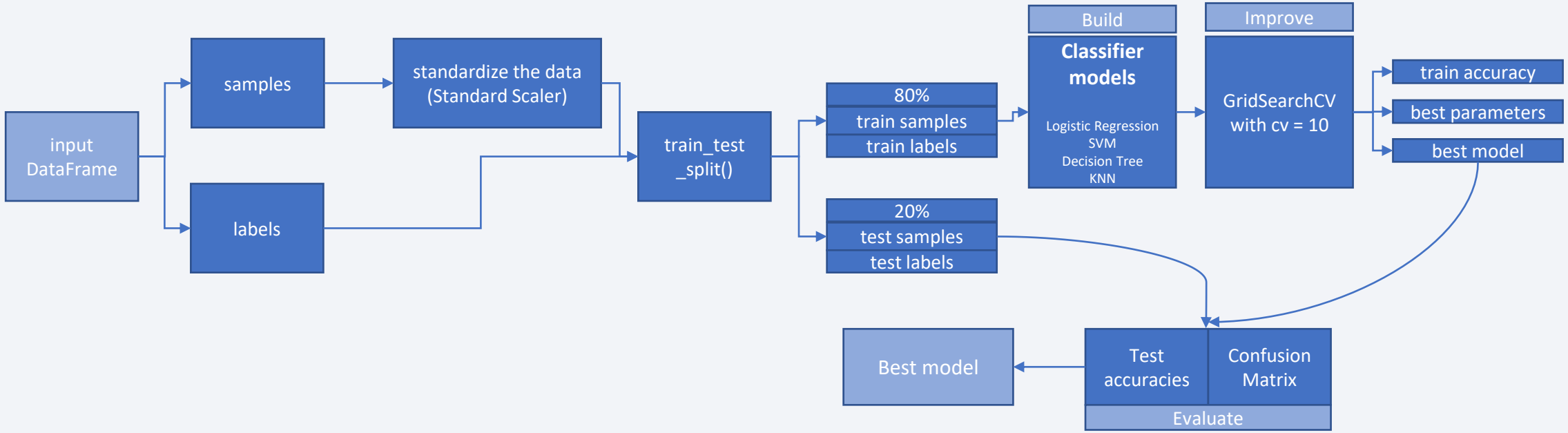
- Visually inspecting the surroundings of the launch sites could have further insights.
- Several objects of the Folium library were used:
  - Map - centered at NASA Johnson Space Center
  - Marker – added icon and label to the sites on the map
  - Circles – added to the map to make the site's surroundings more visible and to get a sense of distance on the map
  - Icon – an object making the map points visible, also added an information sign mark with respect to landing's success
  - MarkerCluster – made possible to map points with same coordinates
  - MousePosition – used for reading coordinates from a map
  - PolyLine – used for connecting launch sites to their proximities
- For accessing the notebook click here: GitHub link

Build a Dashboard with Plotly Dash

- Building a Dashboard makes it possible to analyze data that is modified in real-time.
- Architecture:
  - dropdown list – empowers the user to choose among launch sites, or select them all
  - range slider – allow us to select the payload range from 0 to 10.000 kg
  - pie chart - connected to the dropdown list
    - with all launch sides it shows the successful launches per site
    - in respect to only one site the success/failure rate is shown
  - scatter plot - input: dropdown list and range slider
    - the plot shows the class of the launches in respect to the payload mass (adjusted by the slider), colored by the booster version category
  - callback function
    - connecting the user input to the graphs
- For accessing the notebook or the dash app click here: notebook, dash app

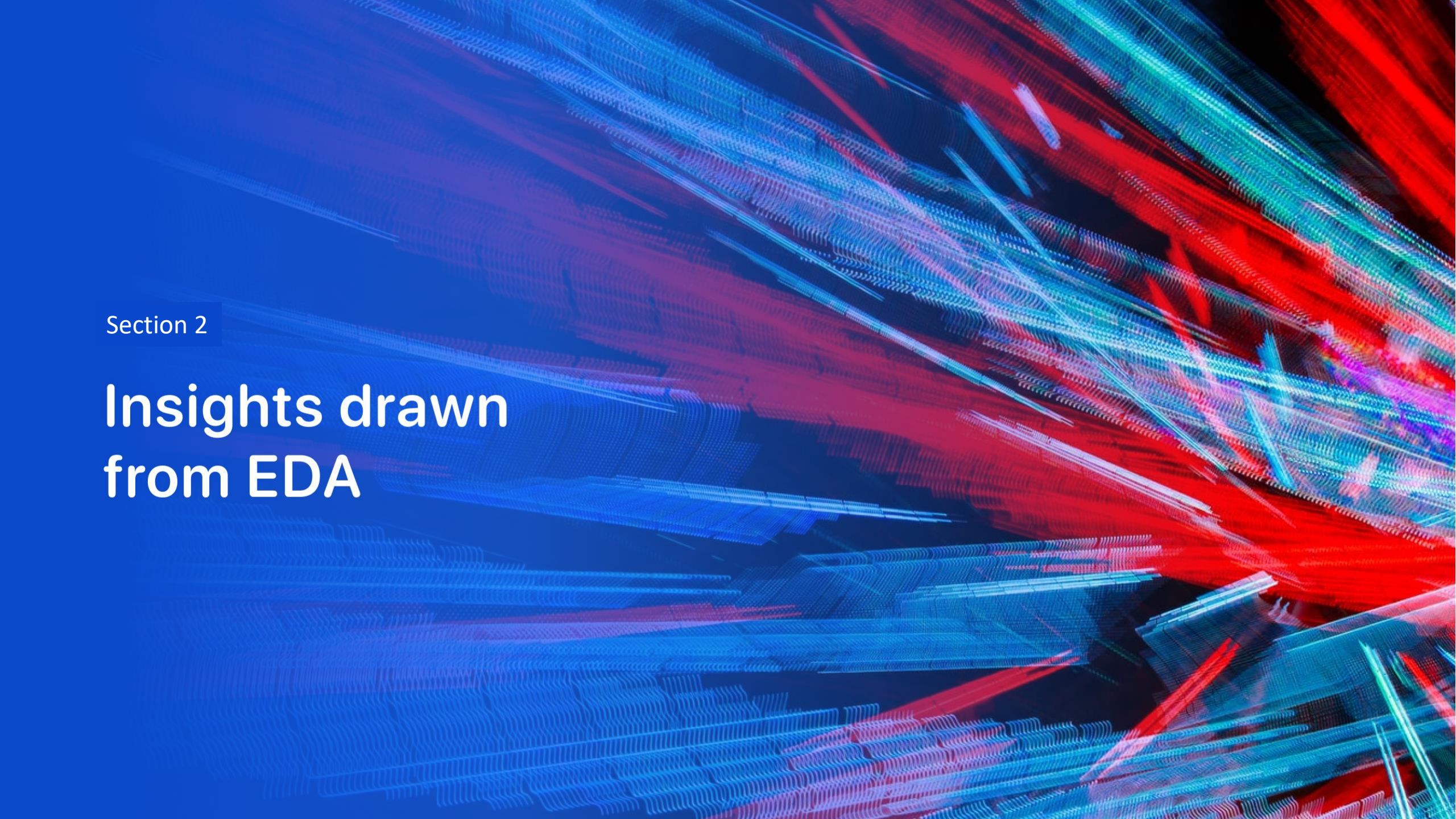
Predictive Analysis (Classification)

- I have trained many classification algorithms on training data using grid search for finding the best hyperparameters.
- For accessing the notebook click here: GitHub link

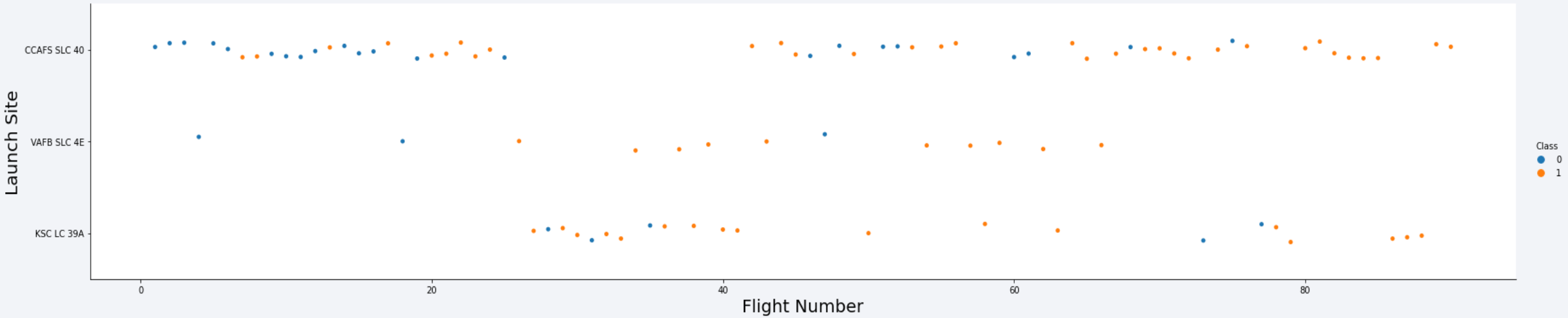


Results

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

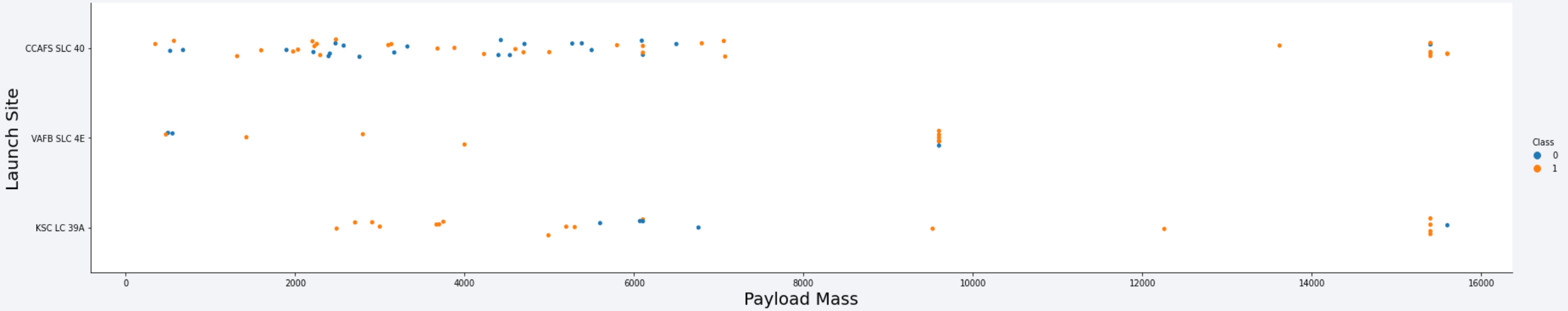

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- As we can see most launches were started from CCAFS SLC 40 site, and this platform is getting more successful with time.
- VAB SCL 4E is achieving good and stable results over time, but this station is somehow not preferred nowadays.
- KSC LC 39A was used intensively when the CCAFS platform were on break.

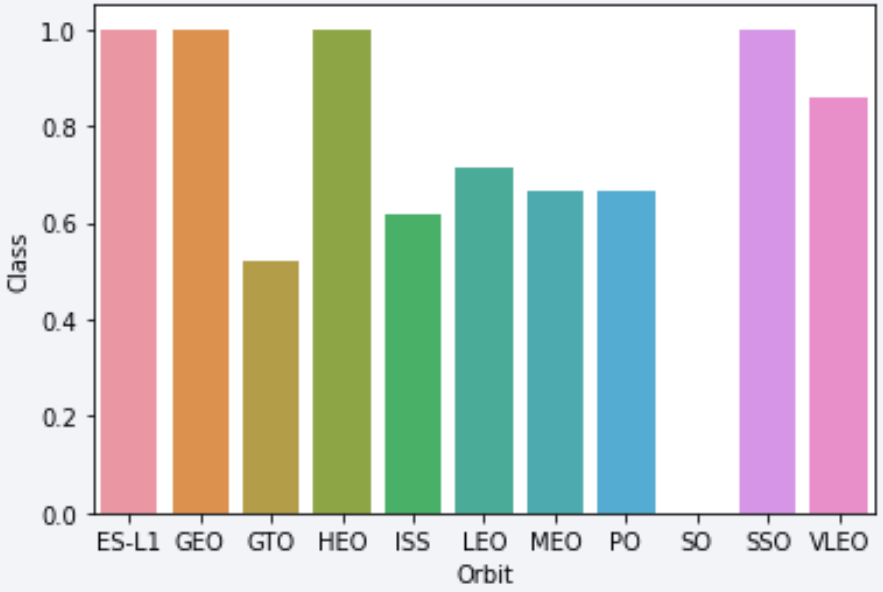
Payload vs. Launch Site



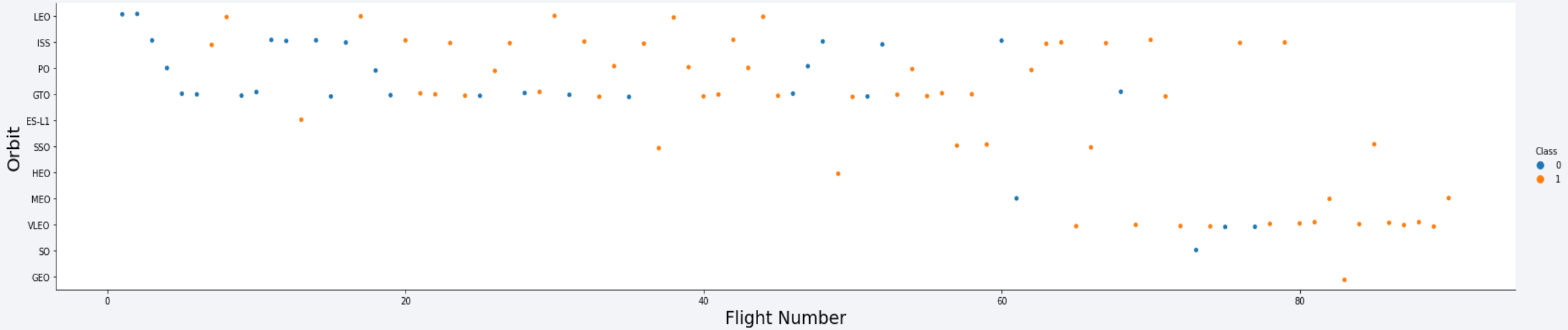
- It is interesting that rockets with high payloads have higher success rates.
- VAFB SLC 4E site is not launching rockets with higher payloads than around 10.000 kg.
- Most payloads are in the 500 to 7500 kg range.

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO related spaceflights didn't have any trouble with landing.
- There were no SO orbit flights yet.
- The height of the flight is not the only factor affecting success rate, as GEO, GTO orbits are both 35786 km above the equator.

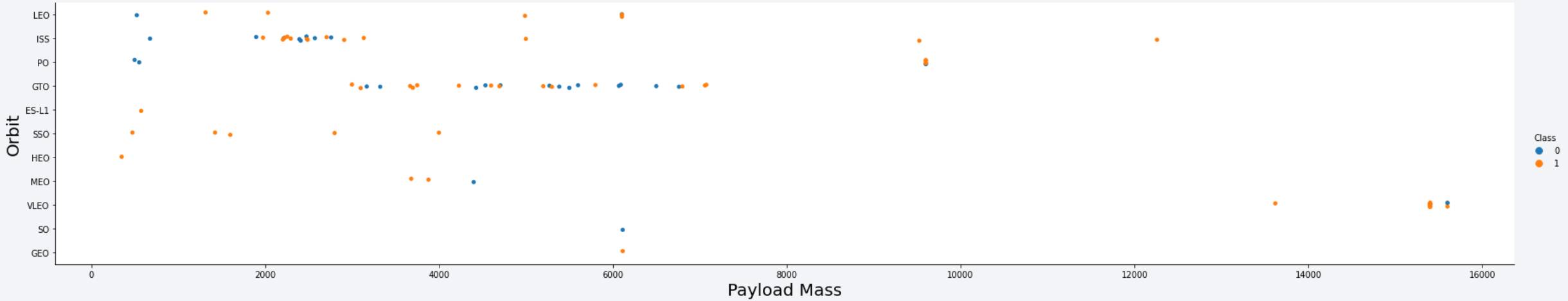


Flight Number vs. Orbit Type



- In some case having more flights helps gaining experience and yields in better results. This can be seen with the LEO orbit.
- In other cases, no linear improvement can be seen. (~GTO)
- ISS is presenting an interesting pattern. After many unsuccessful landings grater number of successful ones proceed. This might be a sign of trial periods, which is followed by improved performance.
- Nowadays VLEO flights are preferred, but LEO orbits are also frequent targets.

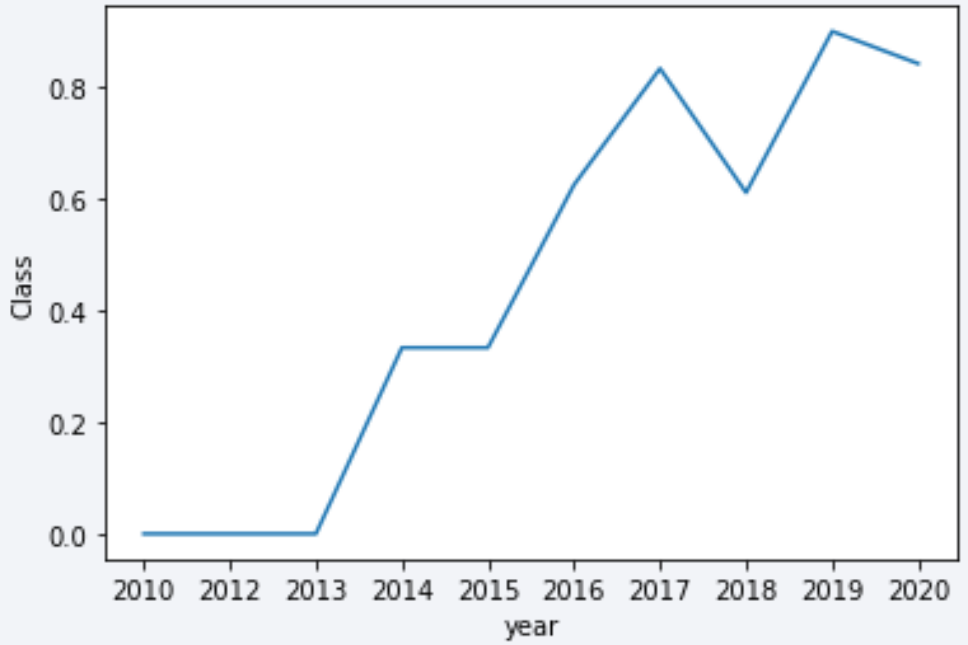
Payload vs. Orbit Type



- In some cases, the higher the payload the better the landing outcomes will be. This can be seen with PO, LEO and ISIS.
- VLEO orbit is only visited with high payloads (<13000 kg).

Launch Success Yearly Trend

- After an initial stagnation from 2010 to 2013 there is a steep linear increasement is success rate.
- As we can see there are some stagnations (2014-15) and fallbacks (2018), but the trend is clear.
- Consequently, time indeed is a good estimator of successful landing.



All Launch Site Names

- There were 4 unique launch site names identified in the dataset.
- Two of them have a very similar name (CCAFS LC-40, CCAFS SCL-40), suggesting they are somehow related (maybe in infrastructure of geographical location).

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

Launch Site Names Begin with 'CCA'

- With CCA string matches, in the first 5 rows I have got only records with the CCAFS LC-40 launch site. As dates are in an ascending order this suggests, that at the beginning of the project CCAFS SLC-40 site wasn't built/activated.

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload that was carried by Falcon 9 rockets for NASA (CRS) was 45596 kg. Comparing to our (0-16000 kg) payload range this might mean few heavy load flights or many with smaller ones.

sum_payload
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2928 kg (about the weight of an elephant).
- This metric is holding much more information than the sum, because we can state that the payloads are generally from the lower range. (Or course we need to keep in mind, that with low sample size a very low and a high value can yield this result, like 500 kg and 5500 kg in average equals 3000 kg…)

sum_payload
2928

First Successful Ground Landing Date

- By combining outcomes and dates we can easily acquire the first successful landing from the dataset. It was in 2015. 12. 22. just before Christmas. From the previous graphs we could see that there were attempts from 2010, so the engineers took 5 years to achieve a breakthrough.

1

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- In this payload rage there are 7 booster versions that could succeed. They are all denominated with F9 FT B10, and the rest is their specific version.
- We can assume that these versions were developed for this particular weight range.

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- There were 100 cases considered as success and only one which failed. 1 from the 100 was labeled with payload status unclear, it needs consideration whether to include it into our payload analysis.
- We could see that much more landings ended with failure so launch and mission outcomes are not closely related.

1	mission_outcome
1	Failure (in flight)
99	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

- There were 12 rockets carrying the maximum payload, which was 15600 kg.
- We can suggest that special booster families are responsible for heavy transportation, that are only differ in the last digit (e.g., B1048.4, B1048.5)

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- As we could see the first successful landing was performed in 2015, so now let's see the balance of the year. The query resulted with 2 failures that year, so 1/3 of the attempts were rewarding.
- They were trying distinct boosters with 3-month difference from the same launch site, but the drone ship landings weren't successful.

booster_version	launch_site	DATE	Landing _Outcome
F9 v1.1 B1012	CCAFS LC-40	2015-01-10	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	2015-04-14	Failure (drone ship)

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

- As we saw earlier in the initial period there were lots of successful missions, but no achieved landing. The table on the right can give us an explanation, as in most cases the landing wasn't even attempted.
- Most successful landings at that time period were achieved using drone ship.
- Landing in the ocean is not labeled with failure and success but rather by controlled and uncontrolled attempts.

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

Section 3

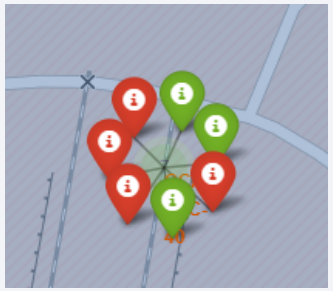
Launch Sites Proximities Analysis

SpaceX launching sites – a global view

- Launching centers are placed on the East and West Coast of the USA.
- They are placed very close to the equator, so this could be beneficial for launching a rocket.
- Also 3 of the 4 launch sites are located on the East Coast, maybe this is a better place for rocket science…

Landing success of individual launch sites

- On the map every red marker represents a failure of landing. Green markers symbolize the exact opposite.
- It is no question that the launch site can influence the outcome of the landing, and the absolute winner is KSC LC-39A in that respect. VAFB SCL-4E and CCAFS SCL-40 seems similar, and the worst performance relates to CCAFS LC-40
- We should be cautious with this findings as if we want to build a launch site on the exact location of KSC that could yield with worse performance. Simply it can be SpaceX's business policy to establish different sites for riskier missions.



CCAFS SLC-40

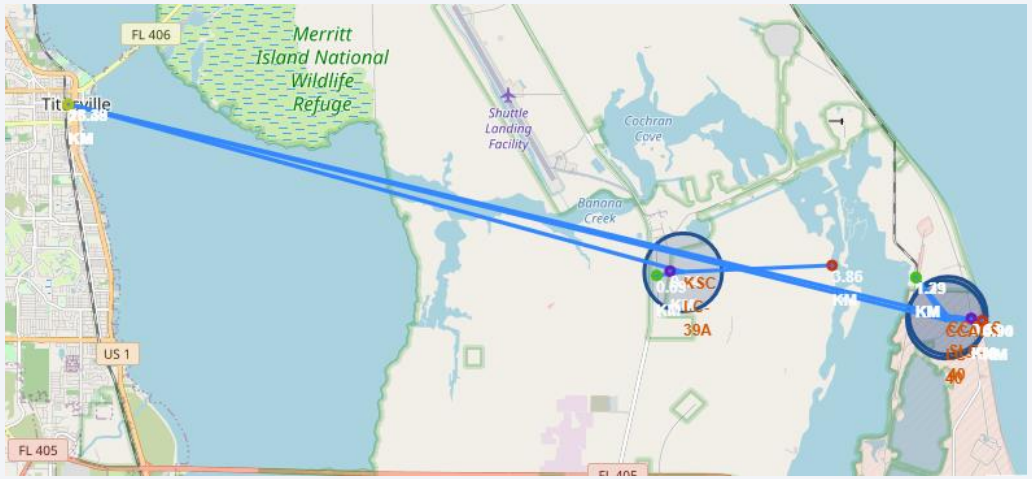
CCAFS LC-40

KSC LC-39A

VAFB SLC-4E

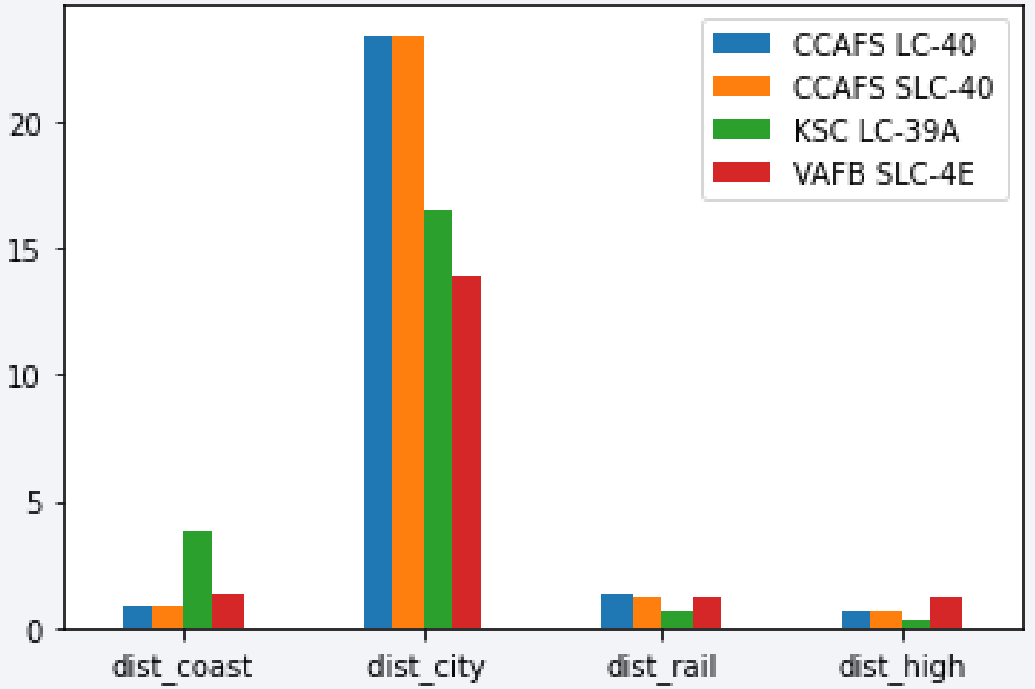
Explore the surroundings of the launch sites

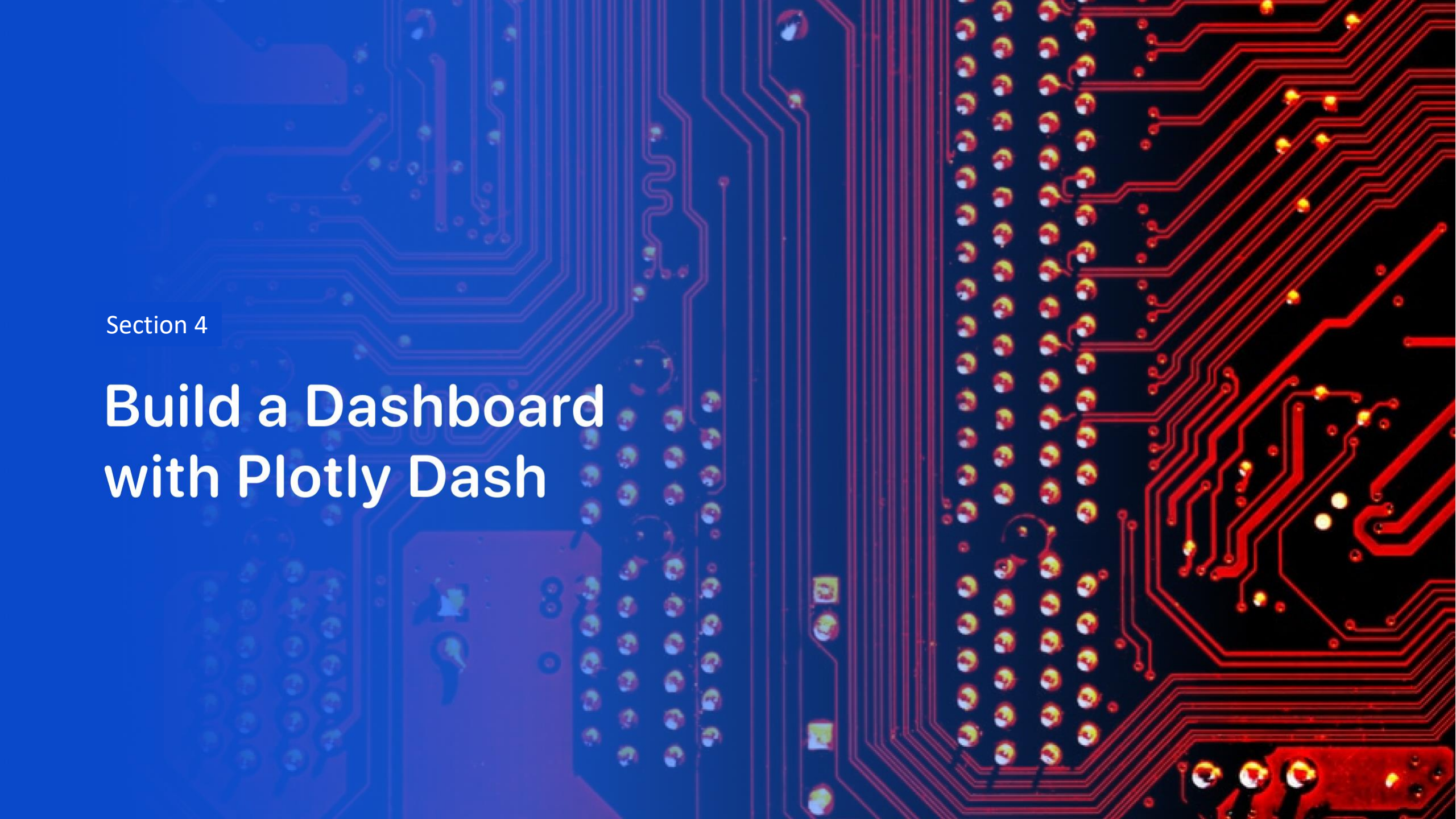
- On the first map we can see the East Coast launch sites and their proximities.
- It can be stated that transport options are very important to the sites as they are within 1 km from the complexes.
- The ocean can be crucial as well, due to it is not further than 4 km.
- However, the closest city is 15 km at best, so it is advisable to build launching sites far away from inhibited areas. Maybe this is legally regulated to limit noise pollution or being outside of electrical waves of the city is needed for successful communication with the rocket.



Dig deeper into proximity data

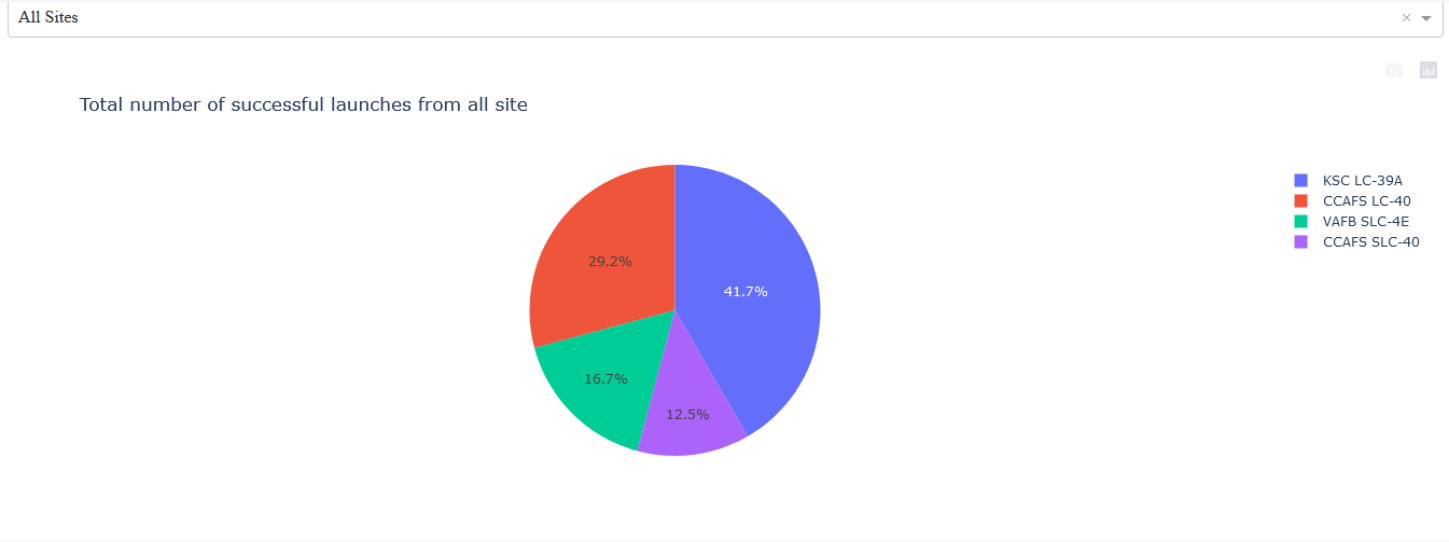
- By visualizing the gathered data on a bar plot, we can generate useful insights.
- Distances of the cities are outstanding, so if we want to buy a land for a lunching site it should be minimum 13 km away from the closest city.
- Although the given land should have good infrastructure regarding to transportation.
- Another important aspect is having the ocean within 5 km.
- The site with the best landing statistics (KSC) have a bit unusual geographical pattern. It is very close to transportation, and moderately close to the coast. It would be interesting to examine that this little bit of extra distance from the coast can have any significant effect on landing statistics. If landing is performed on a ground pad nearby the launch site, airflows from the ocean could have negative effects on coordination.




Section 4
Build a Dashboard with Plotly Dash

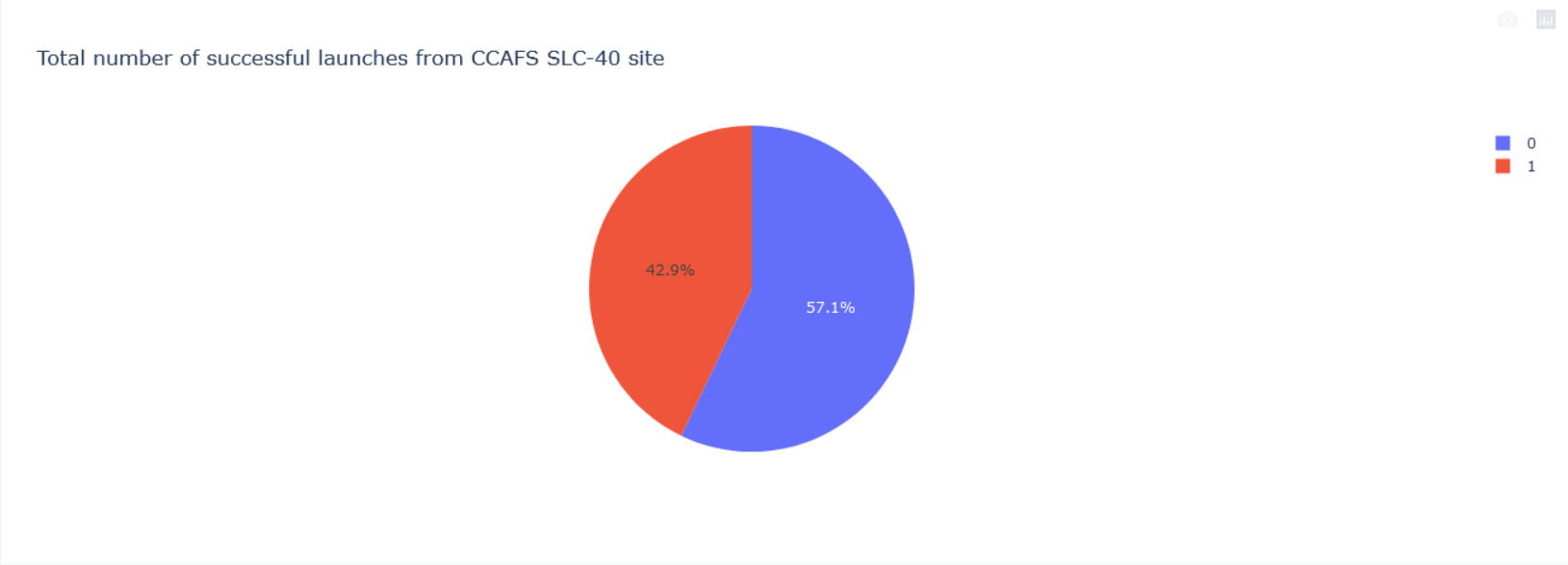
Share of launch sites from all successful launches

- From the pie chart it is clear most successful launches was achieved at KSC (10, 41.7%), while CCAFS SCL-40 has the least (3, 12.5%).
- However, number of landings have little descriptive power over the launch sites.



Individual launch site success rate

- A more descriptive statistic is to show the success/failure ratio, which could be visualized well on a pie chart.
- The image is showing the launch site with the highest launching success ratio, namely CCAFS SLC-40.
- It gives a great example of correctly implementing the results, as we could easily name the KSC site the most successful site of all based on the chart from the previous slide. But the highest number of successful launches relative to failures is related to CCAFS SCL-40. Meaning that if we suppose to launch a new rocket we should do it from the latter site, if we want to be surer in the successful outcome.



Which booster will carry my payload?

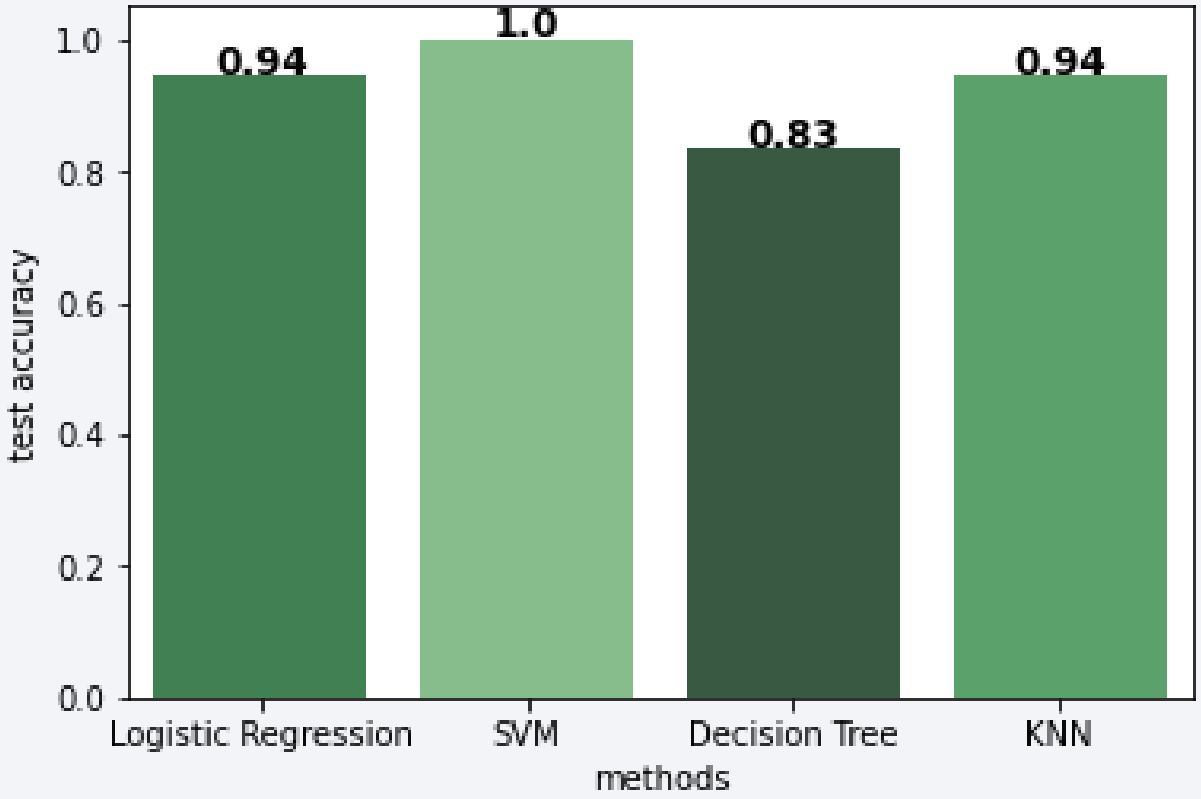
- By splitting the payload range into half, we can see that the lower half have all the booster version represented in much higher density than in the upper half, where there are only two types of booster used. Meaning, specialized booster versions are needed for heavier loads. Also, this suggests that heavier payload transportation is less likely, which can be explained by the low number of successful launches.
- The best launching ratio belongs to FT boosters, but their performance is also decreasing due to additional payload.
- There are boosters with very low sample size (~ v1.0) which is hard to evaluate.
- The worst performance belongs to v1.1 which had only one successful launch despite of the continuous attempts.

Section 5

Predictive Analysis (Classification)

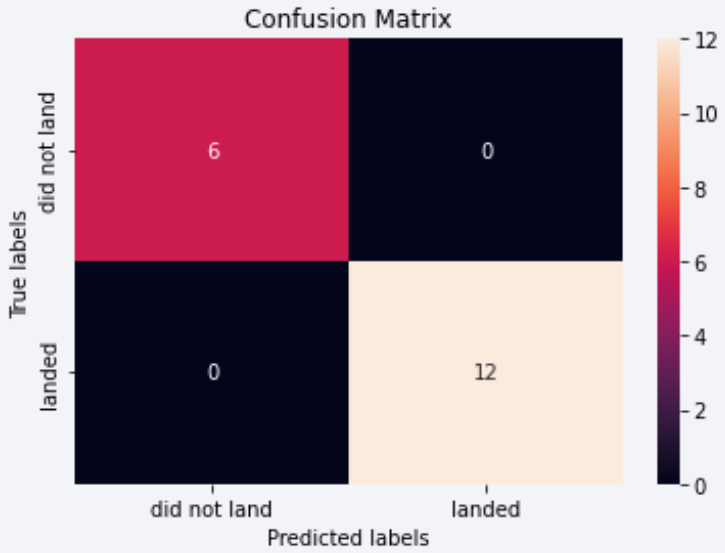
Classification Accuracy

- For the given models after 10 cross validation steps the highest test accuracy could be reached with SVM (support vector machine). This has 1.0 accuracy that means it could predict the class of never-before-seen data from the test set without a mistake.
- Caution: The test data were very little (only 18 samples) would be suitable to test the model on many more real-life samples to reliably compare them).



Confusion Matrix

- The figure on the right shows the confusion matrix of the SVM model.
- It shows that the test set was not balanced, there were more successful landings than unsuccessful ones. (Maybe should be tested on a balanced set, as nowadays this is how landing outcomes are distributed)
- Every sample is correctly classified which is the cause and consequence of the 1.0 accuracy.

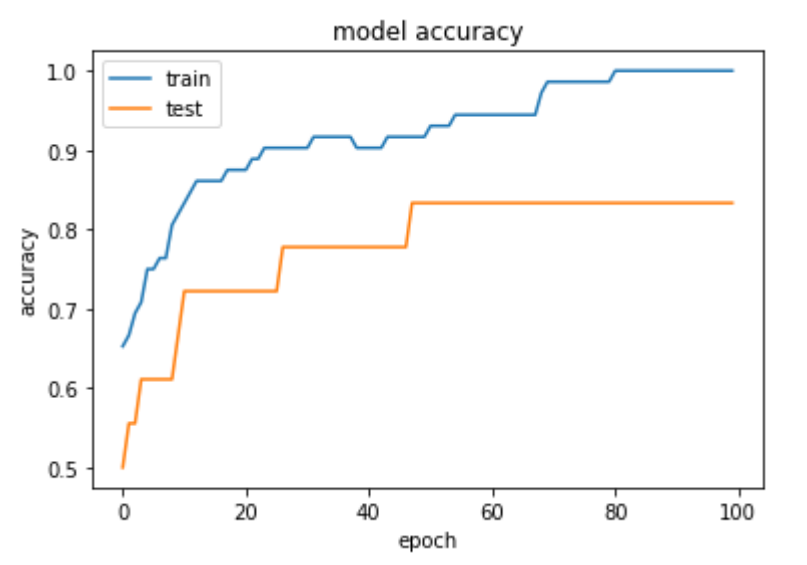


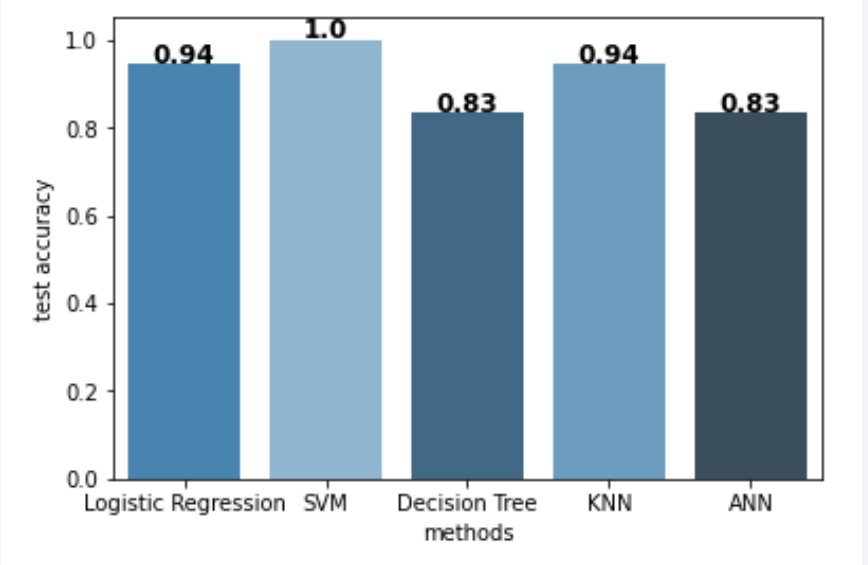
Conclusions

- On a given launch site the chance of having a better landing outcome rises with the number of flights sent from there.
- The success rate of landing is increasing with time.
- Heavy payload transport is a rather new phenomenon, but it can perform good landings with special boosters.
- Orbit types are good indicators for landing outcomes.
- Different orbit type missions, and launch sites used are showing trends over time.
- Special geographical considerations are needed to built a new launching site (close to coast, far from cities)
- With data on the space mission and the rocket (like flight number(~time), payload mass, grid fin usage, legs used, reused parts, launch site, landing pad, and booster version) we can predict the landing outcome with 1.0 accuracy using SVM classification.

Appendix

- The classification was done with a basic neural network architecture too, however it could not reach better performance than 0.83 on the test set.
- The training history shows that there is an overfitting, but the algorithm learns valuable features over time.
- Modifying the hyperparameters can affect the training, but as we have an algorithm with 1.0 accuracy, I won't include this analysis here.





Thank you!