IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pratik Ashok Jawade

Outline

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

Executive Summary

We found some models to be able to reliably predict if the Falcon9 first stage will safely land and therefore we can estimate the cost of the launch much more accurately than if we didn’t have this information.

We found some variable to be more predictive than others

In the future spaceXs ability to land the first stage will likely continue to grow so we might need to recalibrate the model every few new flights to keep it accurate.

Introduction

- This project is the final project for the IBM Data Science course issued by Coursera. It is a full data science journey starting from data collection and finishing with training machine learning models to help us answer our questions
- We will try to asnwer if the first stage of the Falcon 9 rocket from SpaceX will land successfully based on other factors of the launch

This paper tries to predict the probability of a successful landing of rockets from Spacex based on past data with machine learning and other data science tools. SpaceX has far cheaper rockets than the competition, much of this is attributed to SpaceX being able to reuse the first stage of the Falcon 9 rocket, therefore if we can determine if the first stage will successfully land we can estimate the cost of a lauch, which might be useful to other companies betting against SpaceX

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: •The data was collected from the SpaceX API
- Perform data wrangling
  - The data was processed to better suit our needs
  - Some missing values were replaced by the mean for that column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

- We used logistic regression, decision trees, KNN, svm, all hypertuned with grid search to find the best parameters

Data Collection

Data sets were collected using python requests library which allowed me to make HTTP requests.

Then I turned it into a Pandas library to make the data easier to work with

We filtered the dataframe to only include Falcon9 launches as the others were not representative of the sample

Data Collection – SpaceX API

- Data was downloaded from this URL

https://api.spacexdata.com/v4/launches/past

- This is my Jupyter notebook for reference with all the work

https://github.com/simonf2004/capstone-project/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb

Don’t know what to do honestly
I have searched through the whole course and didn’t find how to do this

Data Collection - Scraping

- We webscraped some data from

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

To collect historical launch records of Falcon9. Using the beautiful soup library.

This is my Jupyter notebook for reference with all the work:

https://github.com/simonf2004/capstone-project/blob/main/jupyter-labs-webscraping%20(1).ipynb

Don’t know what to do honestly

Data Wrangling

We performed some Exploratory Data Analysis to find patterns in the data.
We noted missing values and replaced some of them
Tried to find which features of the data might be useful for future training of models. We calculated the number of launches on each site. Created a landing outcome label from outcome column.

This is my jupyter notebook for reference:
https://github.com/simonf2004/capstone-project/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb

EDA with Data Visualization

Here we used mostly scatter point plots because they illustrate the point the best we used seaborn and matplotlib package to visualis the relationships between different variables to help us understand how they would affect the launch outcome. I also payed attention to what variable affect the outcome the most and where is the strongest correlation so I know which variables to train the models on.

This is my jupyter notebook for reference:

https://github.com/simonf2004/capstone-project/blob/main/jupyter-labs-edadataviz%20(1).ipynb

EDA with SQL

We performed sql commands to better uderstand the data such as:

Displaying the unique lauch sites

Displaying the average payload mass carried by booster version

F9 v1.1

We did it in sql because sql is much better language at exploring data than python in my opinion

You can see all the sql queries in this jupyter notebook:

https://github.com/simonf2004/capstone-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

We marked all launch sites using folium package also marked all the launches from those sites. We also calculated the distances between lauch sites to its proximities and this helped us to understand the dataset and see if for example the launch sites are in proximity to the equator line or if they are in close proximities to the coast.

Note: I had some problems with this. The lab environment couldn't start and I had to run it locally and make some changes that didn't always work. I did my best

This is my jupyter notebook where you can find all the exercises:

https://github.com/simonf2004/capstone-project/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

We tried to explain if the launch sites have any impact on the succes rate of the flight. We added a dropdown menu where you can chose from all the different launch sites and a pie chart showing the success rates based on the launch site. There is also a scatter plot showing the relationship between the payload mass and if the launch was successful. The dashboard shows that all and interactively.

This is my jupyter notebook where you can see it all:

https://github.com/simonf2004/capstone-project/blob/main/dash%20board%20capstone.ipynb

Predictive Analysis (Classification)

The last step was training the models and try to answer the original question:

Can we predict if the first stage will land succefully or not?

We used logistic regresion, knn, svm, and decision trees. We optimised their parameters with the grid search. Then based on confussion matrix and the jaccard score index we choose the best model. The attempt was successful and we found models that are able to predict the outcome. I will talk about the models on section 5

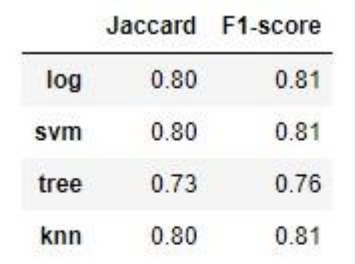
Jupyter notebook for further reference:

https://github.com/simonf2004/capstone-project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

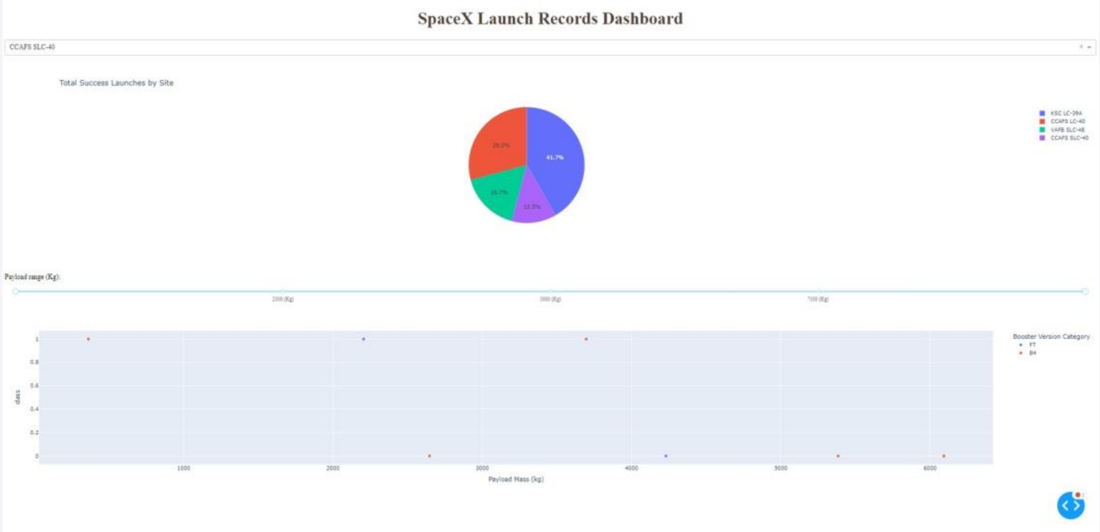
Results

- In the LEO orbit the success rate is related to the number of flights but in the GTO urbit there is no such a relationship.
- The success rate has been increasing since 2013
- The launch site KSC LC 39A has been the most successful so far

- From the predictive analysis the logistic regression, svm, knn have the same performance metrics and they all appear as a good choice
  here is a report of all the models with their metrics



	Jaccard	F1-score
log	0.80	0.81
svm	0.80	0.81
tree	0.73	0.76
knn	0.80	0.81



SpaceX Launch Records Dashboard

CCAFS SLC-40

dashboard

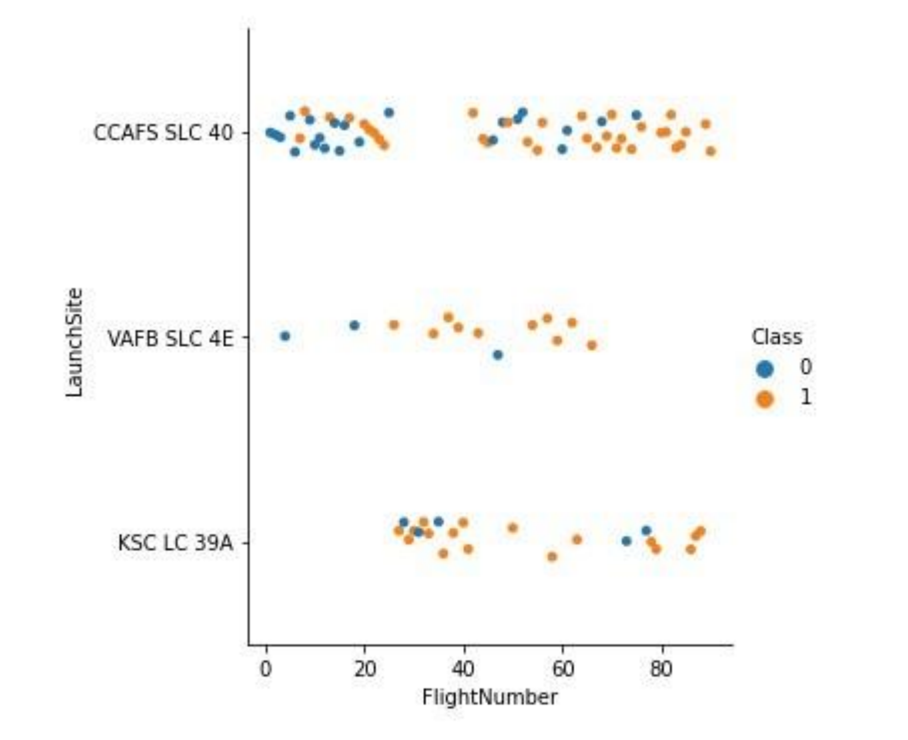
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

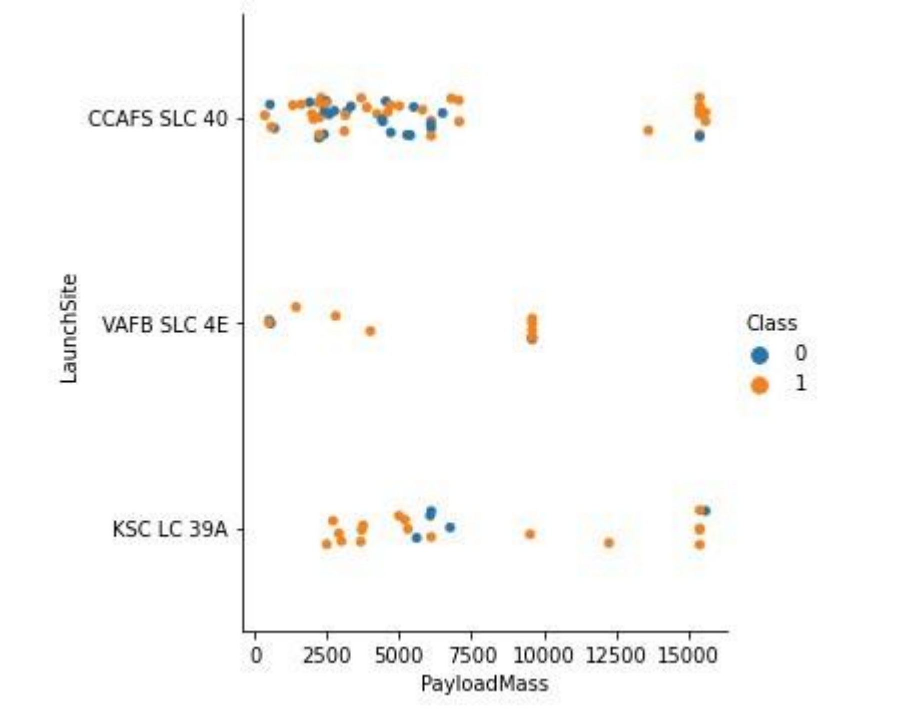
- As you can see SpaceX uses the CCAFS SLC 40 launch site the most and it seems that the VAFB SLC 4E launch site has not been used in some time

- Class represents if the first stage of the rocket landed safely



Payload vs. Launch Site

- You can see that SpaceX uses either CCAFS or KSC launch sites for their biggest rockets.
- VAFB has been used a lot for rockets with a payload mass of 10000. it seems that VAFB is only for testing purposes as the rockets launched there are all very small.



Success Rate vs. Orbit Type

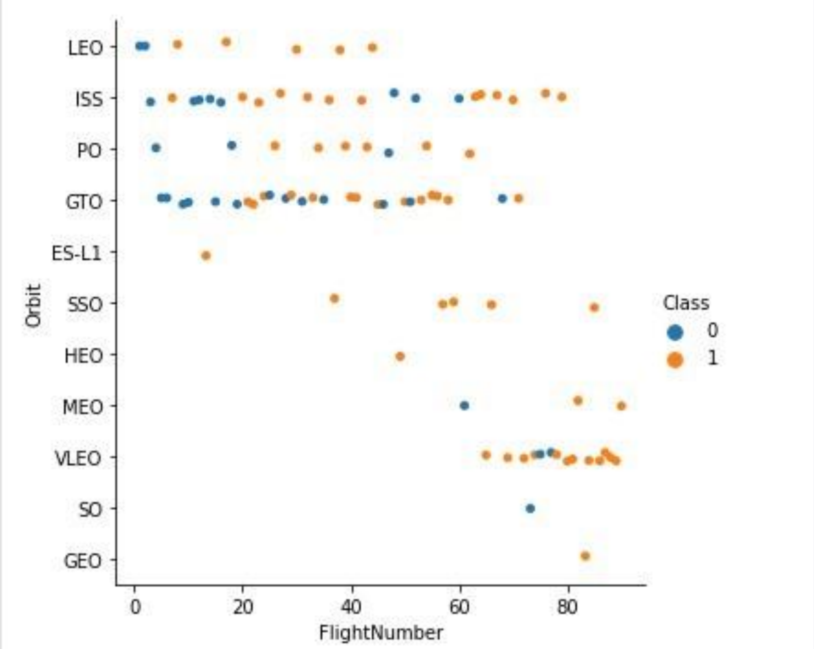
- I wasnt able to complete this one due to difficulties with the jupyter lab

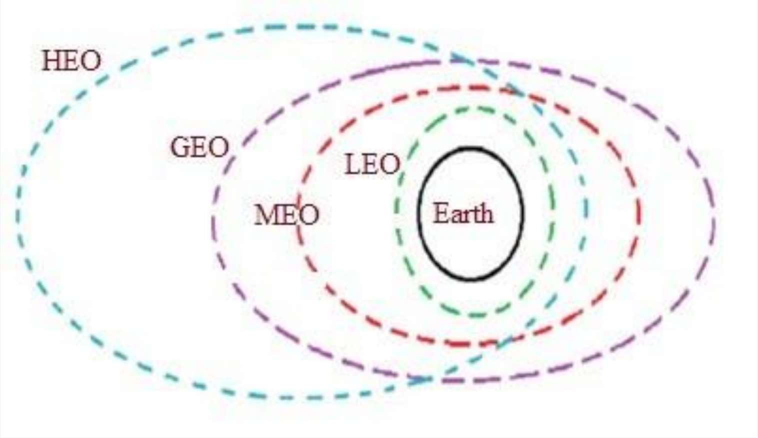
- I would assume the highest obitals have the smallest success rate

Flight Number vs. Orbit Type

- We can see that SpaceX makes flights to lower orbits as the flight number progresses. That is unexpected to me. They have been making a lot of flights to VLEO(very low earth orbital) I would assume those are just test flights
- The success rate does seem to increase as the flight number increases indicating that SpaceX is getting better

- Here are some of the orbit types to better understand this plot

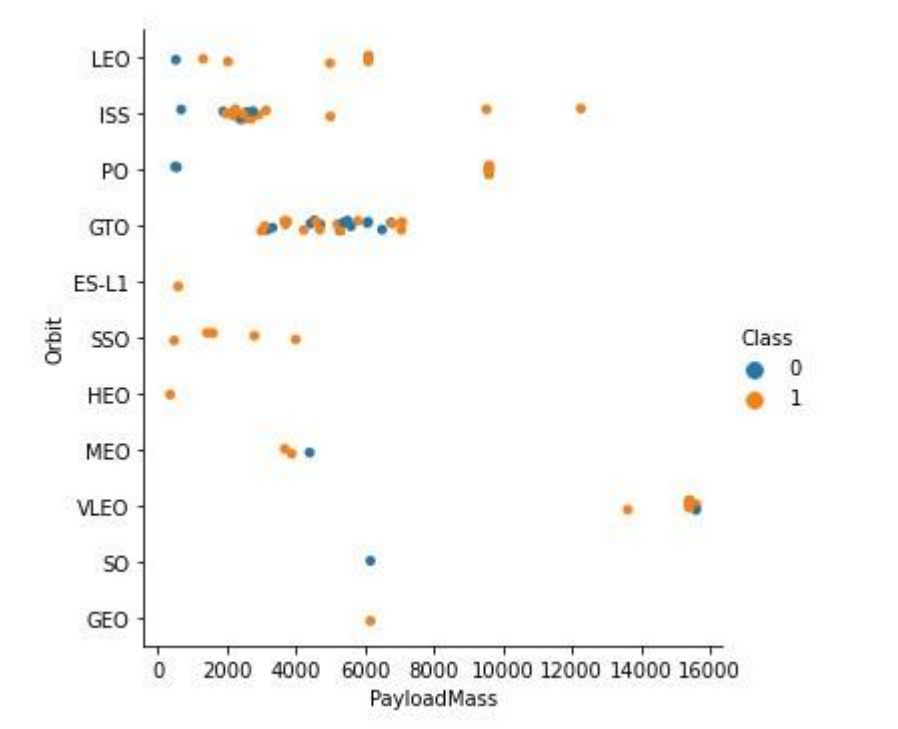




Payload vs. Orbit Type

- Here you can see that the largest rockets fly almost exclusively to the VLEO.

  That might be because they are only testing the bigger prototypes and based only on the data I would assume they will fly to higher orbitals with the heavier rockets in the future



Launch Success Yearly Trend

- I wasnt able to do this one

- Show the screenshot of the scatter plot with explanations

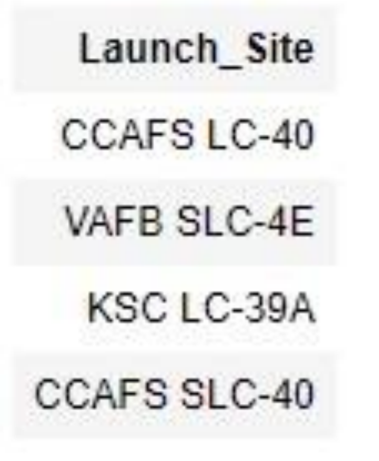
All Launch Site Names

Find the names of the unique launch sites:

These are the names of the launch sites

The names are: Cape Canaveral Space Launch Complex CCAFS (SLC-40), Vandenberg Space Force Base Space Launch Complex VAFB (SLC-4E), Kennedy Space Center Launch Complex KSC (LC-39A),

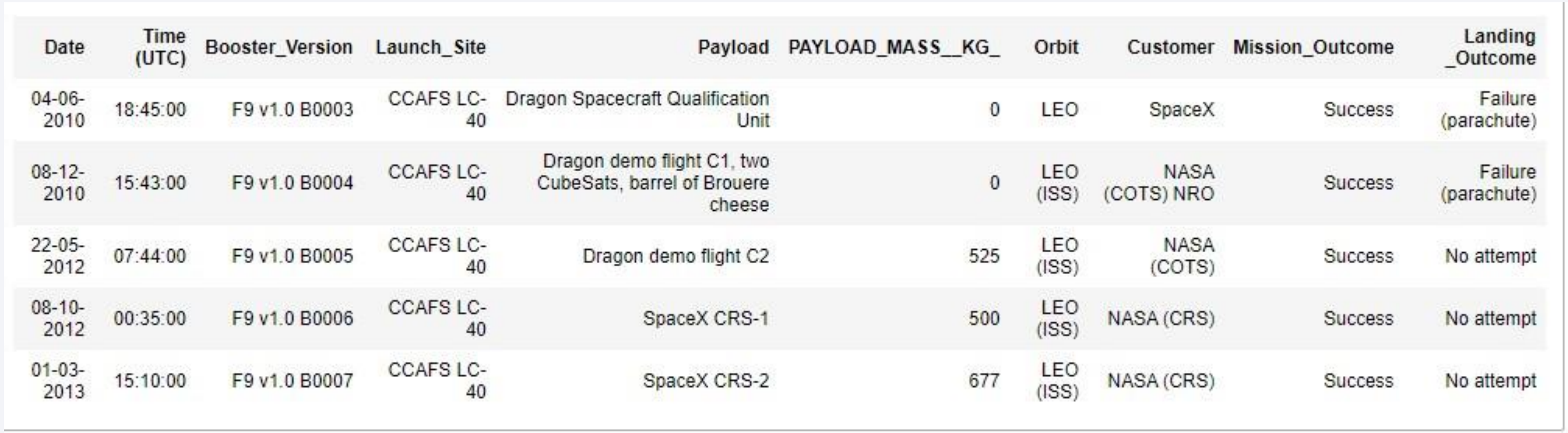
and Brownsville South Texas Launch Site CCAFS (LC-40)



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 begin with `CCA`

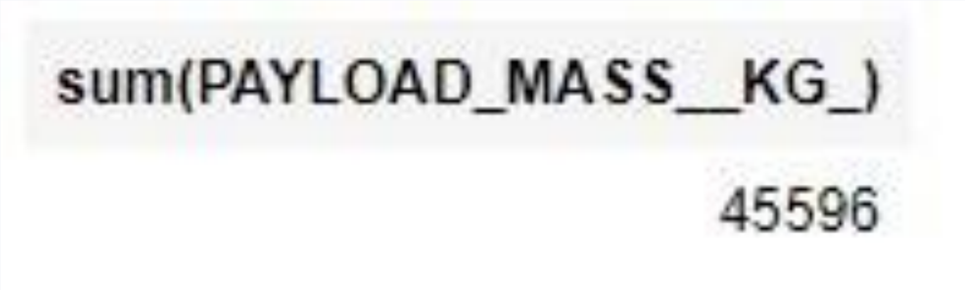


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

Total Payload Mass

Calculate the total payload carried by boosters from NASA

The total payload carried by nasa boosters

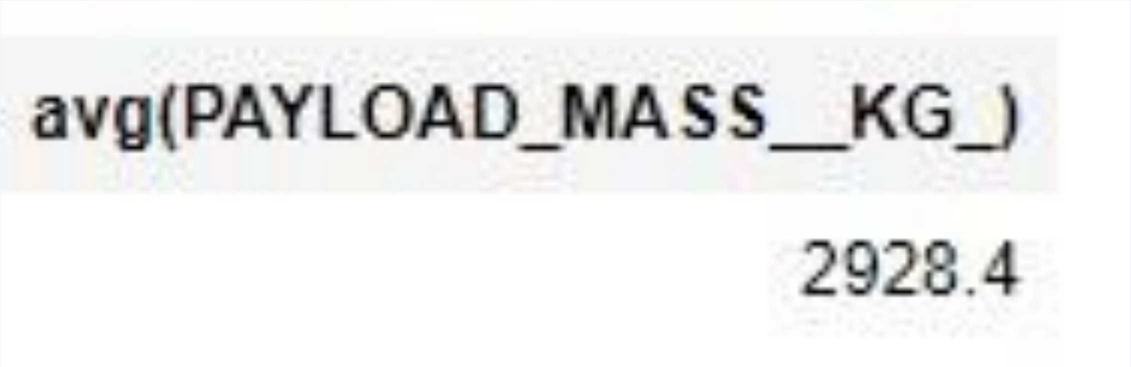


sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

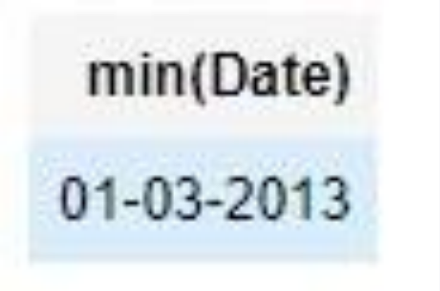
Here is F9 v 1.1



avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

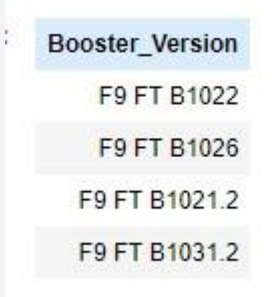
Find the dates of the first successful landing outcome on ground pad



min(Date)
01-03-2013

Successful Drone Ship Landing with Payload between 4000 and 6000

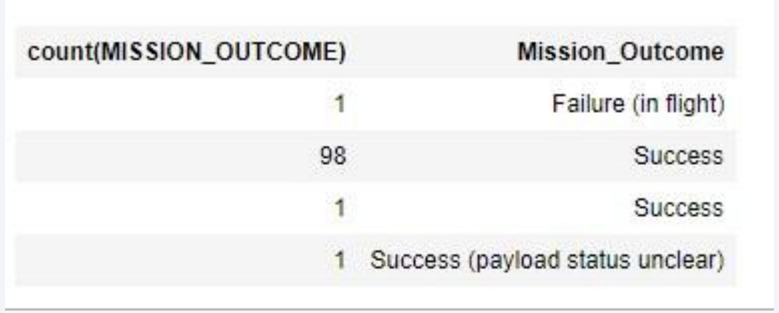
List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



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

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

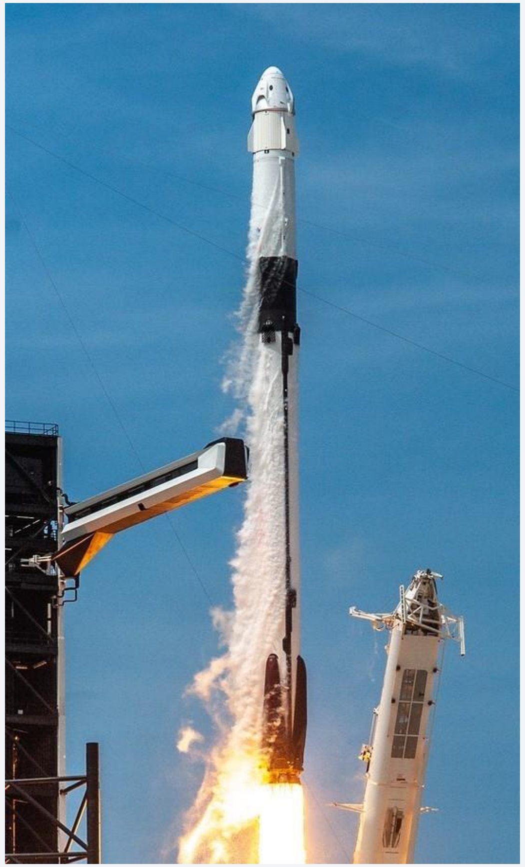


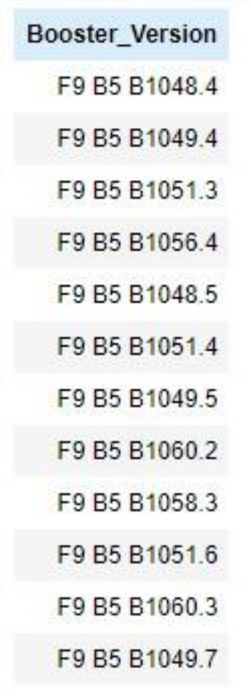
count(MISSION_OUTCOME)	Mission_Outcome
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

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

Heres the F9 block 5 rocket

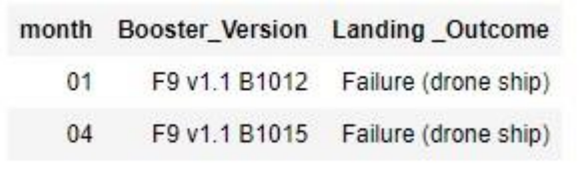




Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

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



month	Booster_Version	Landing _Outcome
01	F9 v1.1 B1012	Failure (drone ship)
04	F9 v1.1 B1015	Failure (drone ship)

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

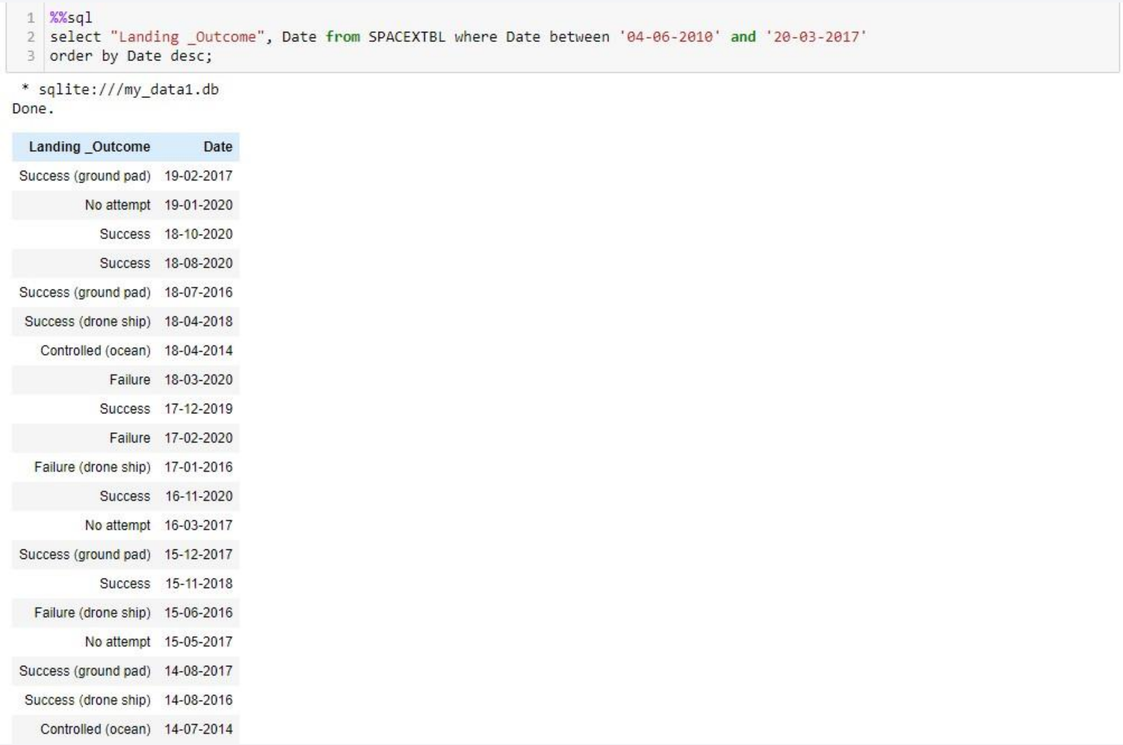
This query doesn’t work for me

I have tried hard but it only

shows the difference between the days.

The data type in the database

might be wrong I think



```
%%sql
select "Landing _Outcome", Date from SPACEXTBL where Date between '04-06-2010' and '20-03-2017'
order by Date desc;
```

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

Landing _Outcome	Date
Success (ground pad)	19-02-2017
No attempt	19-01-2020
Success	18-10-2020
Success	18-08-2020
Success (ground pad)	18-07-2016
Success (drone ship)	18-04-2018
Controlled (ocean)	18-04-2014
Failure	18-03-2020
Success	17-12-2019
Failure	17-02-2020
Failure (drone ship)	17-01-2016
Success	16-11-2020
No attempt	16-03-2017
Success (ground pad)	15-12-2017
Success	15-11-2018
Failure (drone ship)	15-06-2016
No attempt	15-05-2017
Success (ground pad)	14-08-2017
Success (drone ship)	14-08-2016
Controlled (ocean)	14-07-2014

Section 3

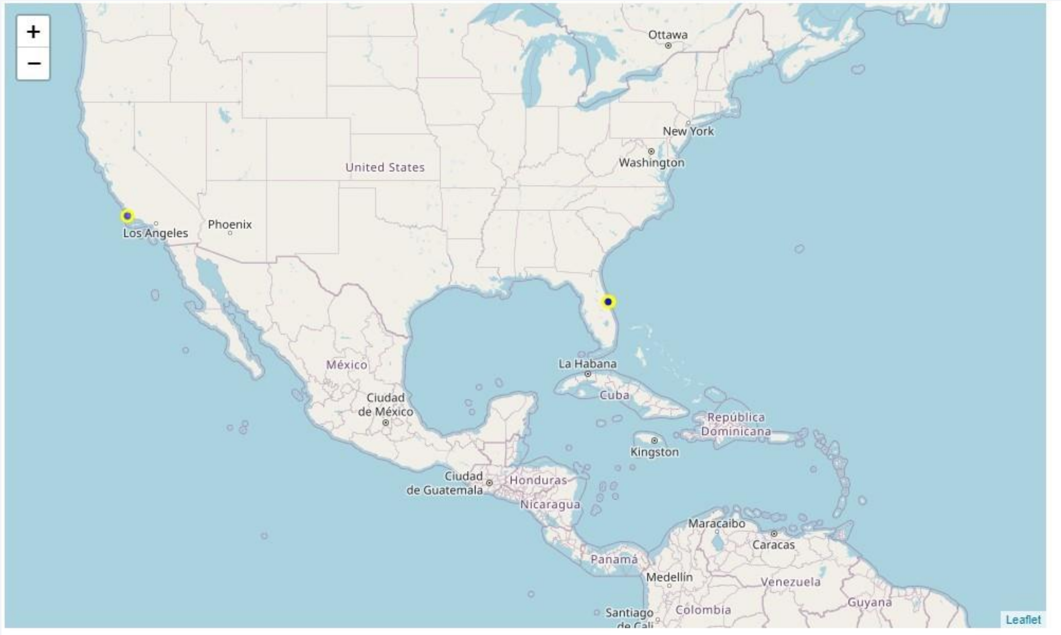
Launch Sites Proximities Analysis

Launch sites locations

Spacex launch sites located on a map

As you can see they are in similar locations the distance from the coastline and the distance from the equator are very similar

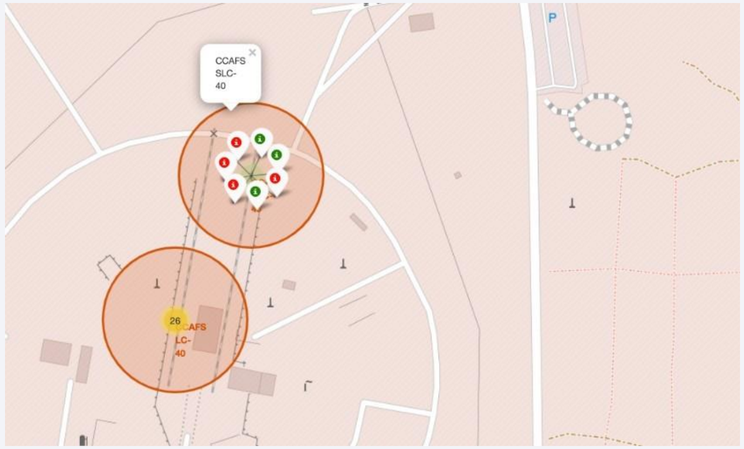
I would say its because of weather conditions



Launch outcomes from each site on a map

Here you can see the successful and failed landings of the first stage based on the launch site

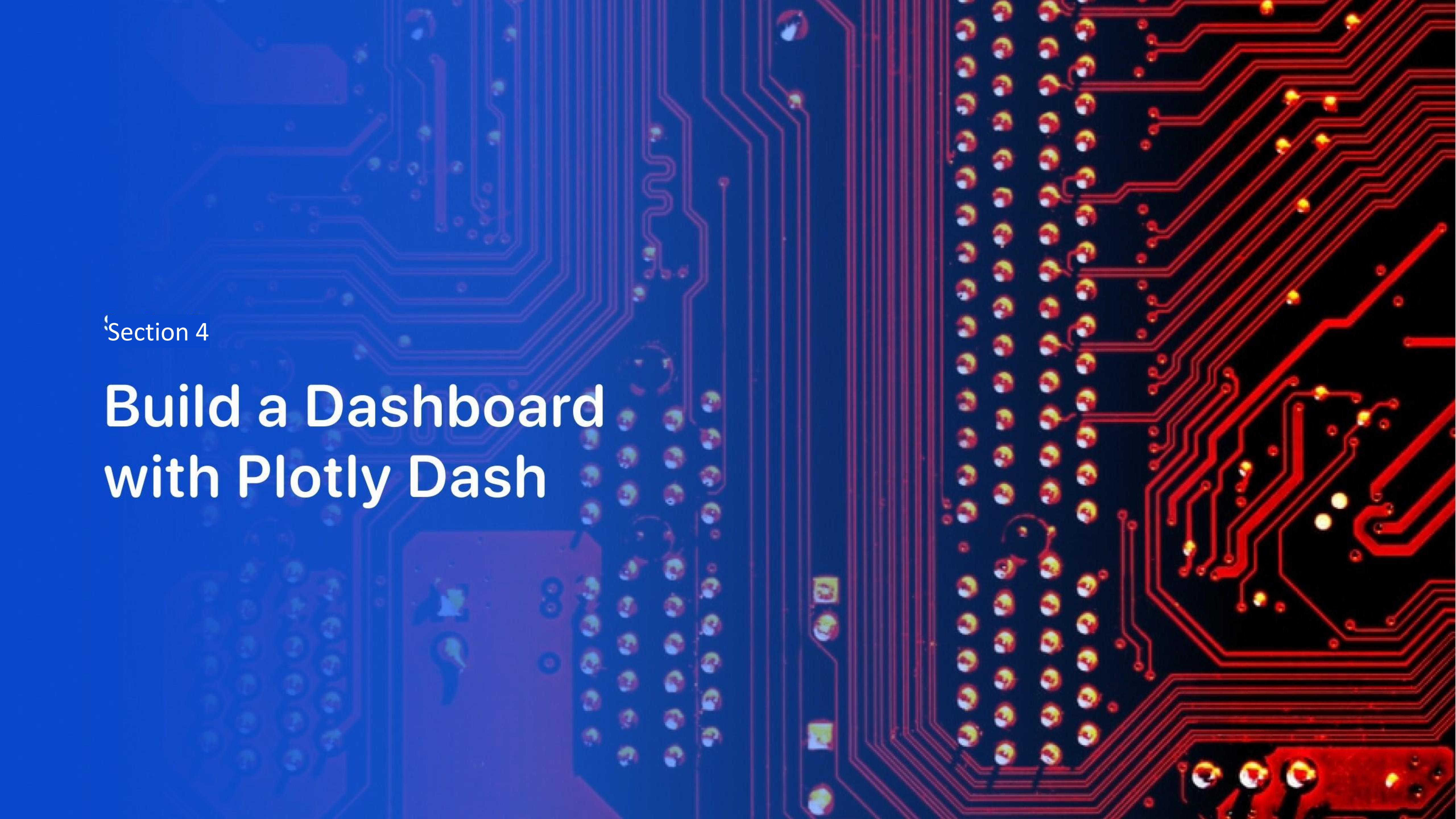
There seems to be no correlation between the launch site and the landing outcome



Distances between launch sites to its proximities

I was unable to finish this one, I had a lot of problems with installing folium and some functions just don’t work no matter what I do it seems

Launch sites have very similar distance to its proximities mainly because its good to have similar weather on all the launch sites

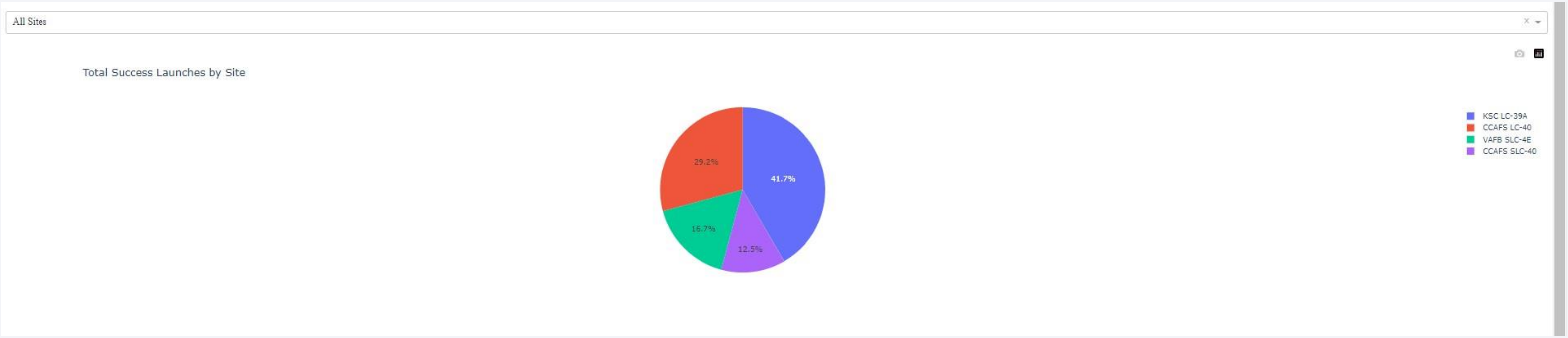


Section 4

Build a Dashboard with Plotly Dash

Dashboard: pie chart

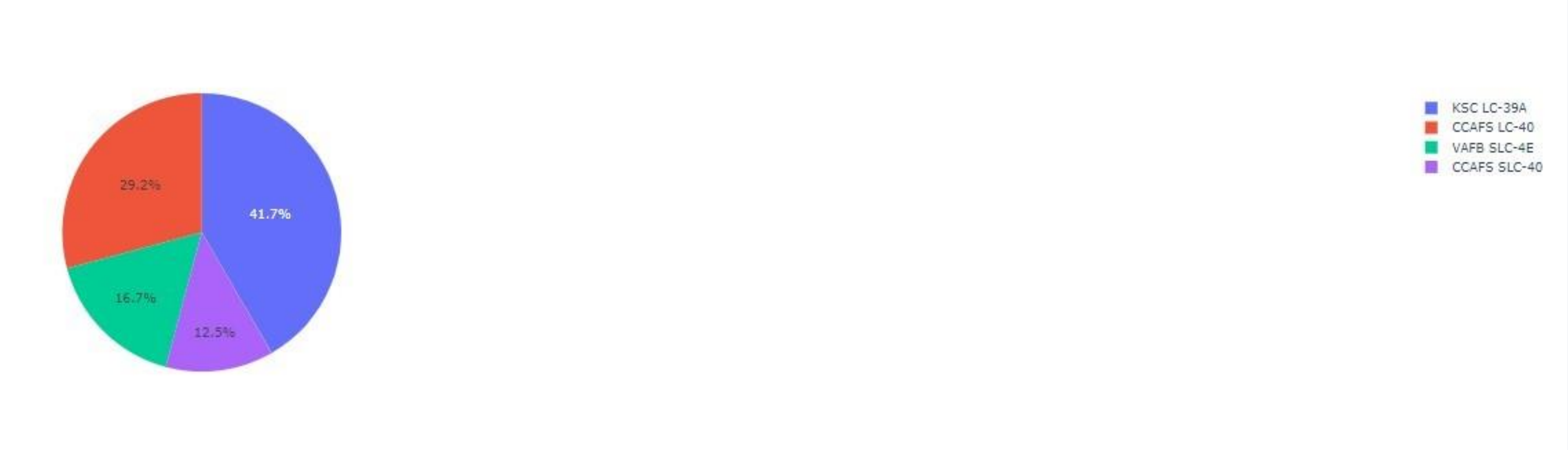
You can see that most of the succesful landing of the first stage come from KSC however I would attribute that more to the fact that most of the seirous launches are there



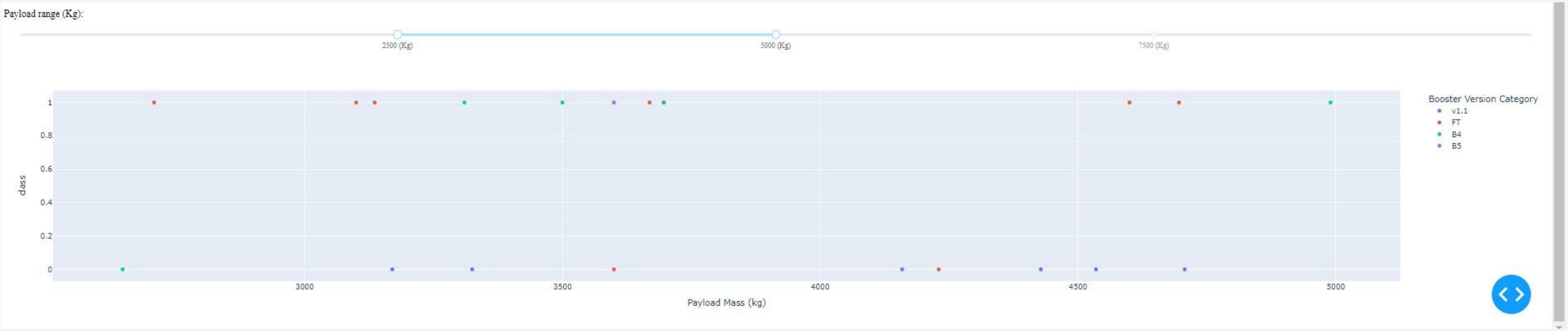
All Sites

Pie chart of successful landings

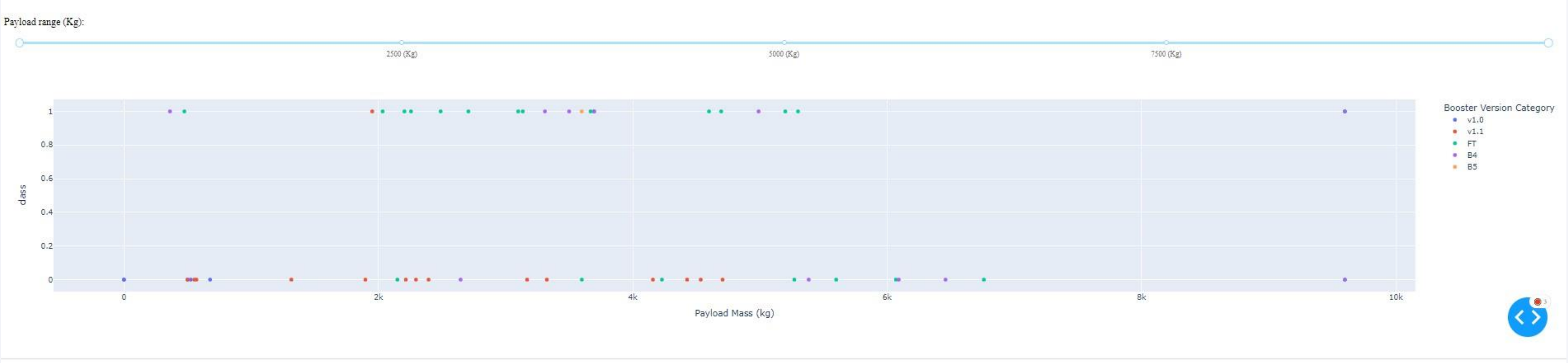
The pie chart for the total success launches by site



Payload vs. Launch Outcome scatter plot



2500-5000kg payload range seems to have the largest success rate and booster FT seems to have to best success rate in all ranges.



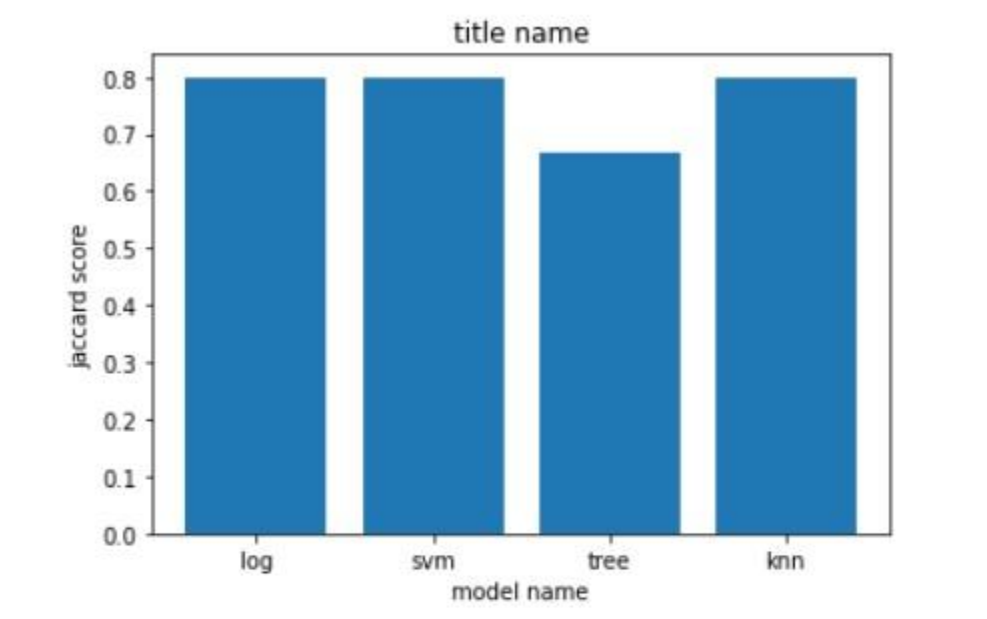
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart:

- I found 3 models to have the same accuracy I would attribute that to small test sample, based on my opinion I would trust the knn the most



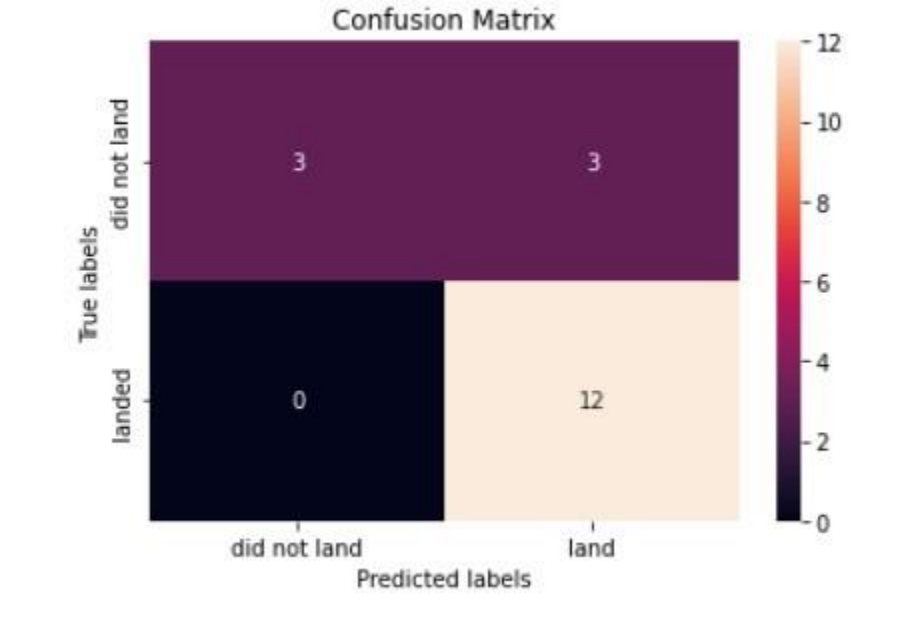
Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation
- Here you can see the cofusion matrix for the knn
- There is no false negatives but it does make some

  false positives
- So sometimes the model might predict that the rocket

  would land but in reality I wont, which means that

  we would expect the launch to cost less than it in reality wi



- In my opinion as SpaceX becomes better this will be closer to reality

Conclusions

-I found multiple models with the same score, I attribute this to small sample size we had only 18 rows in the test data set and that is not enough in my opinion. Of course that is because of the nature of the project there is not thousands of launches so the data sample will always be limited

- svm, logistic regression and k nearest neighbor can all reliably predict if the first stage will land safely or not, and based on that we could estimate the cost of the launch

-i think that we would need to train the model again or calibrate it every few new flights because SpaceX ability to land the first stage will likely increase so the model might grow redundant.

Appendix

I have to say I had a lot of technical problems during this project. The labs almost never worked so I had to download all the notebooks and do it locally which resulted in environment problems and I couldn’t complete some tasks.

I used anaconda to run my notebooks locally.

Thanks very much for rating my presentation I did my best

Thank you!