IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marek Sokołowski

25/06/2023

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
  - Using SpaceX API
  - Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis (EDA)
  - Using SQL
  - Using Visualization
    - Folium
    - Ploty Dash
- Predictive Analysis (Machine Learning Predictions)

Summary of all results

- Relevant data about SpaceX has been collected.
- EDA using Databases and Visualization techniques have been performed and with it, best features for the predictive model have been found, where the best features are: {FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude, Class}
- Random Forest Regressor has been established as the best Predictive Model with an accuracy of 83.33%.

Introduction

Project background and context

In this project, a prediction about the first stage of the Falcon 9 successful landing has been made. On its website, SpaceX promotes Falcon 9 rocket launches for 62 million dollars; other providers charge upwards of 165 million dollars for each launch. A large portion of the savings is due to SpaceX's ability to reuse the first stage. Therefore, if one can figure out whether the first stage will land, then one can figure out how much a launch will cost.

- Based on Lab 1 Introduction

Problems I want to find answers

In order to build a SpaceX rival, I want to gather information on the price of rocket launches as well as the conditions that must be met in order for successful lunch.

Section 1

Methodology

Methodology

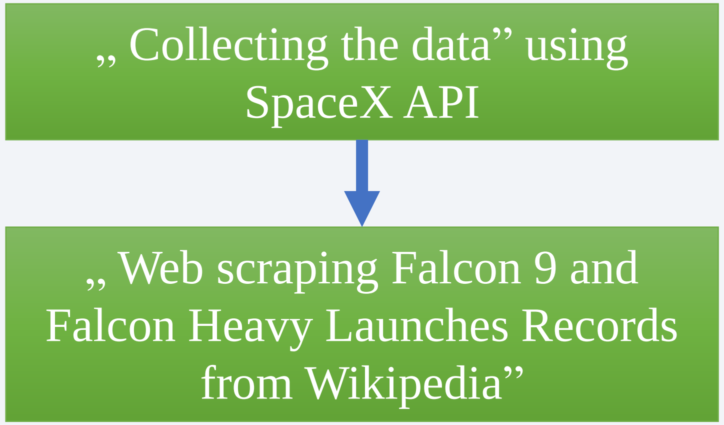
Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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 from https://api.spacexdata.com/v4 using multiple Python 3 libraries and from https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- Data collection process:



„ Collecting the data” using SpaceX API

„ Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia”

- Citations from Lab 1 & 2

- Data from Space X API have been collected using URL requests via the Python request library.
- Data from Wikipedia have been collected using BeautifulSoup via the Python BeautifulSoup library.
- Collected datasets have been properly wrangled, cleaned, and formatted for the further stages with Python libraries (Pandas, NumPy) and additionally written Auxiliary Functions for that purpose have been written and used.

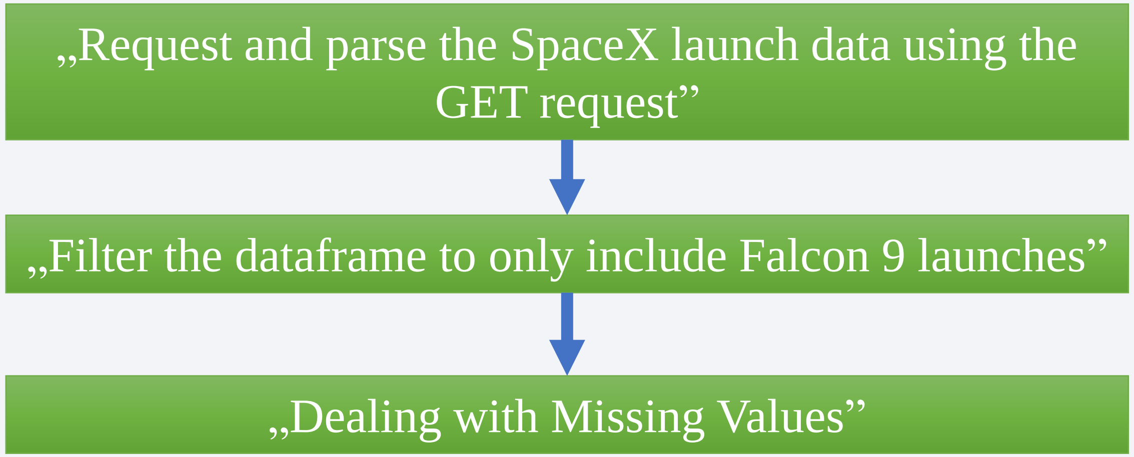
- Based on Lab 1 & 2

Data Collection – SpaceX API

The conclusions presented in this slide are based on:
(must include completed code cell and outcome cell)

- Data from Space X API have been collected using URL requests via the Python request library.
- Collected datasets have been properly wrangled, cleaned, and formatted for further stages with Python libraries (Pandas, NumPy), and additionally written Auxiliary Functions for that purpose have been written and used.

- Based on Lab 1



„Request and parse the SpaceX launch data using the GET request”

„Filter the dataframe to only include Falcon 9 launches”

„Dealing with Missing Values”

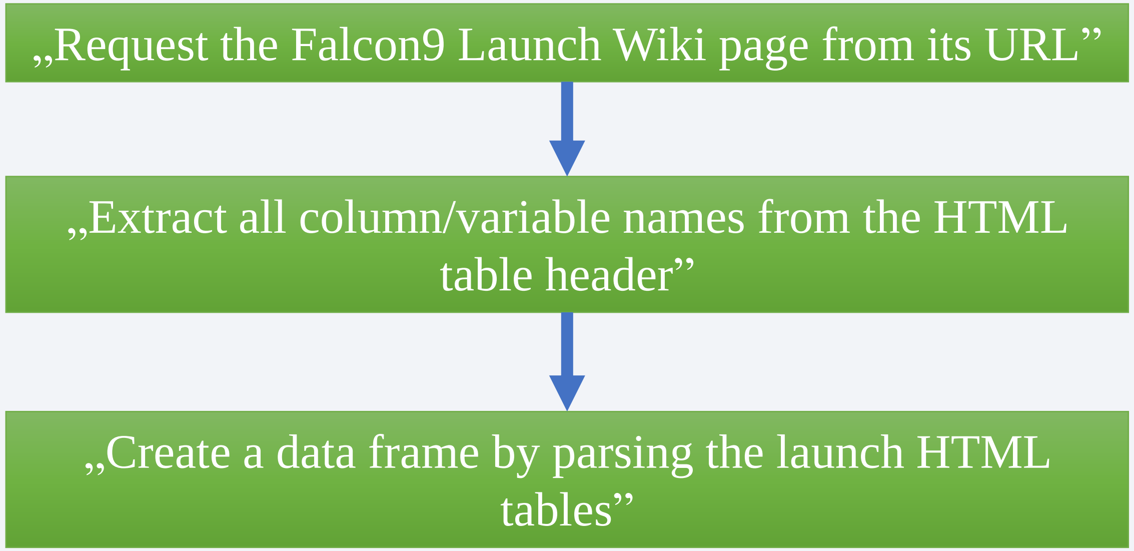
- Citations from Lab 1

Data Collection - Scraping

The conclusions presented in this slide are based on:
(must include completed code cell and outcome cell)

- Data from Wikipedia have been collected using BeautifulSoup via the Python BeautifulSoup library.
- Collected datasets have been properly wrangled, cleaned, and formatted for further stages with Python libraries (Pandas, NumPy) and additionally, written Auxiliary Functions for that purpose have been written and used.

- Based on Lab 2



„Request the Falcon9 Launch Wiki page from its URL"

„Extract all column/variable names from the HTML table header"

„Create a data frame by parsing the launch HTML tables"

- Citations from Lab 2

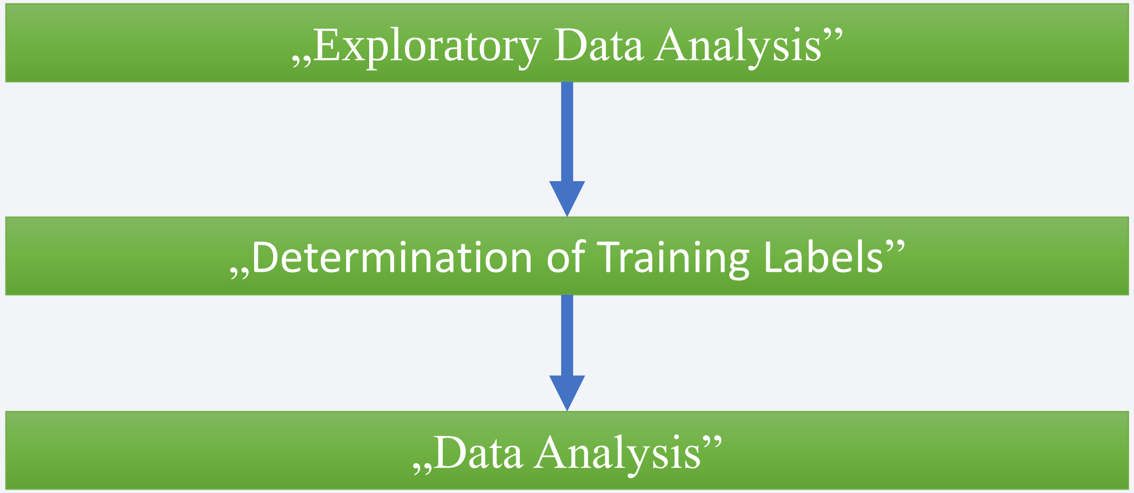
Data Wrangling

The conclusions presented in this slide are based on:
(must include completed code cell and outcome cell)

In order to determine Training Labels I've Calculated the:

- number of launches on each site
- number and occurrence of each orbit
- number and occurrence of mission outcome per orbit type
- and I've created a landing outcome label from Outcome column

- Based on Lab 3


„Exploratory Data Analysis”
„Determination of Training Labels”
„Data Analysis”

- Citations from Lab 3

EDA with Data Visualization

The conclusions presented in this slide are based on:
(must include completed code cell and outcome cell)

During EDA following graph has been created to visually observe a relationship between features:

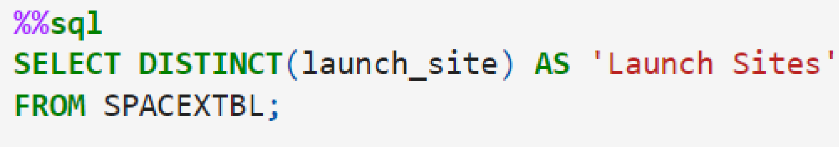
- The scatter plot of Flight Number vs. Launch Site to analyze the relationship between Flight Number and Launch Site.
- The scatter plot of Payload vs. Launch Site to analyze the relationship between Payload and Launch Site.
- Bar chart for the success rate of each Orbit Type to analyze the relationship between success rate and each Orbit Type.
- The scatter plot of Flight number vs. Orbit Type to analyze the relationship between Flight Number and each Orbit Type.
- The scatter plot of payload vs. Orbit Type to analyze the relationship between payload and each Orbit Type.
- Line plot of year vs the average launch success trend to analyze yearly average launch success trend.

EDA with SQL (part 1/4)

The conclusions presented in this slide are based on:
(must include completed code cell and outcome cell)

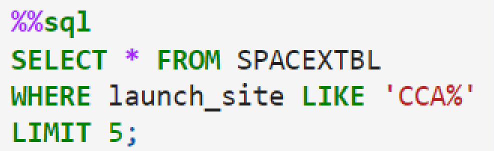
EDA using SQL was performed with the following queries (listed as ‚Task’)

- Task 1 – „Display the names of the unique launch sites in the space mission”
  - Code:



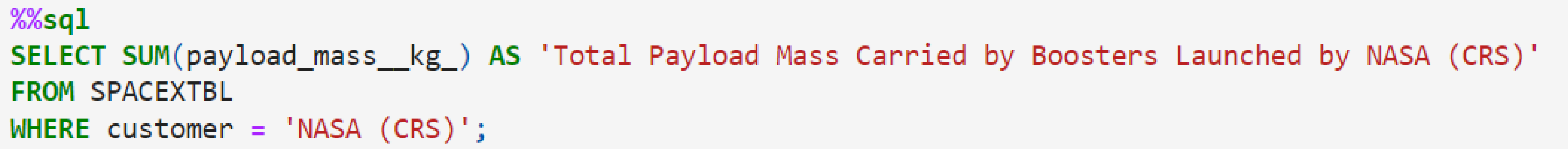
```
%%sql
SELECT DISTINCT(launch_site) AS 'Launch Sites'
FROM SPACEXTBL;
```

- Task 2 – „Display 5 records where launch sites begin with the string 'CCA’”
  - Code:



```
%%sql
SELECT * FROM SPACEXTBL
WHERE launch_site LIKE 'CCA%'
LIMIT 5;
```

- Task 3 – „Display the total payload mass carried by boosters launched by NASA (CRS)”
  - Code:



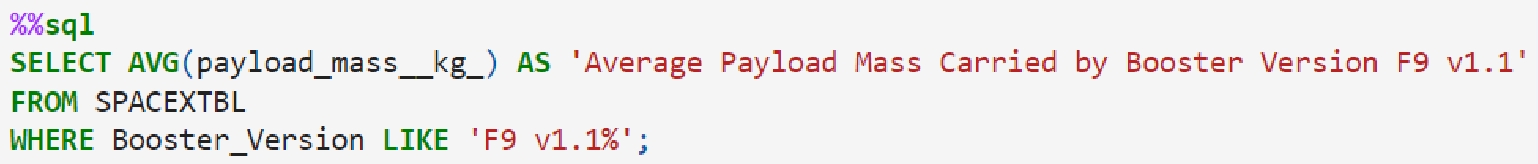
```
%%sql
SELECT SUM(payload_mass__kg_) AS 'Total Payload Mass Carried by Boosters Launched by NASA (CRS)'
FROM SPACEXTBL
WHERE customer = 'NASA (CRS)';
```

- Citations from Lab 4

EDA with SQL   (part 2/4)

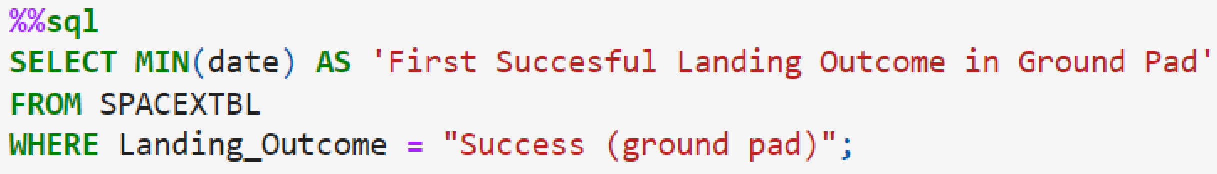
The conclusions presented in this slide are based on:
(must include completed code cell and outcome cell)

- Task 4 – „ Display average payload mass carried by booster version F9 v1.1”
  - Code:



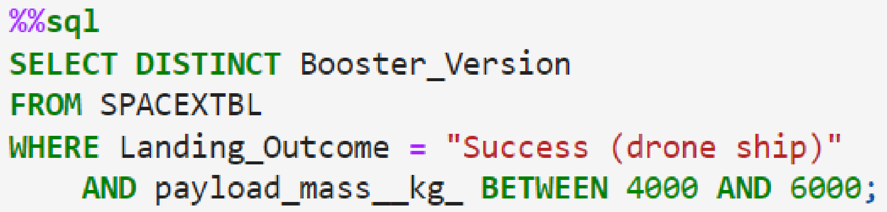
```
%%sql
SELECT AVG(payload_mass__kg_) AS 'Average Payload Mass Carried by Booster Version F9 v1.1'
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1%';
```

- Task 5 – „ List the date when the first succesful landing outcome in ground pad was acheived.”
  - Code:



```
%%sql
SELECT MIN(date) AS 'First Succesful Landing Outcome in Ground Pad'
FROM SPACEXTBL
WHERE Landing_Outcome = "Success (ground pad)";
```

- Task 6 – „ List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000”
  - Code:



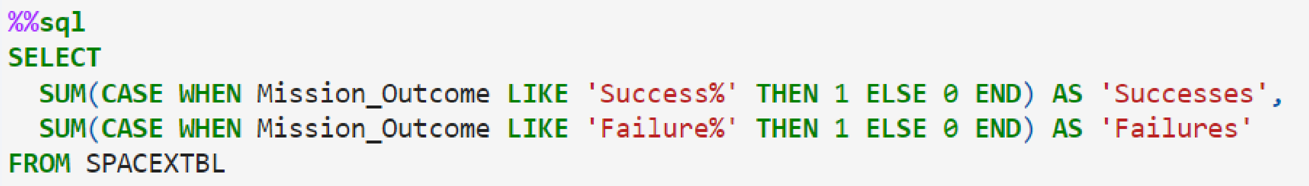
```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome = "Success (drone ship)"
    AND payload_mass__kg_ BETWEEN 4000 AND 6000;
```

- Citations from Lab 4

EDA with SQL (part 3/4)

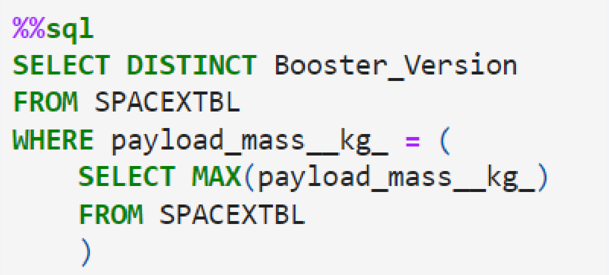
The conclusions presented in this slide are based on:
(must include completed code cell and outcome cell)

- Task 7 – „ List the total number of successful and failure mission outcomes”
    - Code:



```
%%sql
SELECT
  SUM(CASE WHEN Mission_Outcome LIKE 'Success%' THEN 1 ELSE 0 END) AS 'Successes',
  SUM(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 ELSE 0 END) AS 'Failures'
FROM SPACEXTBL
```

- Task 8 – „ List the names of the booster_versions which have carried the maximum payload mass. Use a subquery”
    - Code:

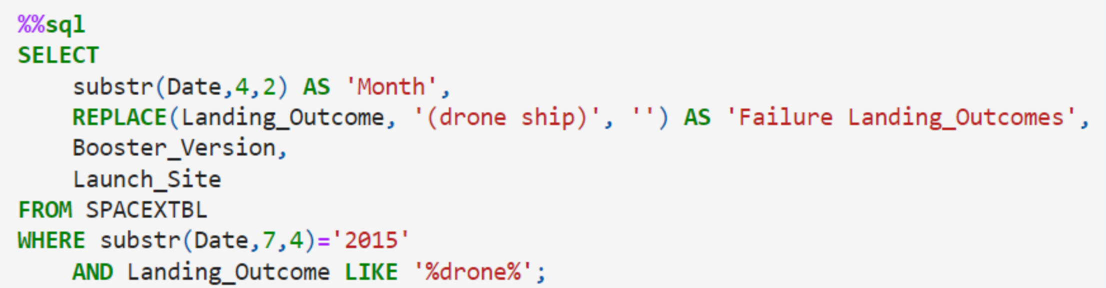


```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTBL
WHERE payload_mass__kg_ = (
    SELECT MAX(payload_mass__kg_)
    FROM SPACEXTBL
    )
```

EDA with SQL   (part 4/4)

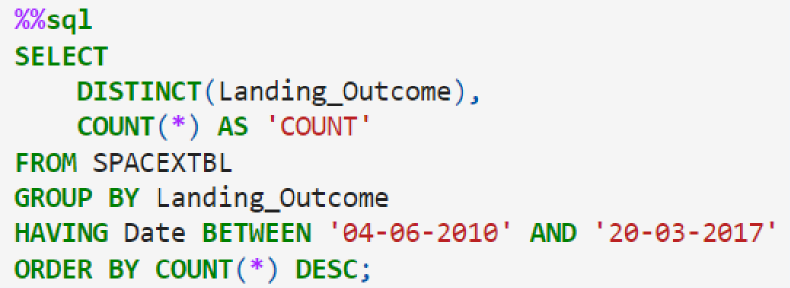
The conclusions presented in this slide are based on:
(must include completed code cell and outcome cell)

- Task 9 – „ List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.”
  - Code:



```
%%sql
SELECT
    substr(Date,4,2) AS 'Month',
    REPLACE(Landing_Outcome, '(drone ship)', '') AS 'Failure Landing_Outcomes',
    Booster_Version,
    Launch_Site
FROM SPACEXTBL
WHERE substr(Date,7,4)='2015'
    AND Landing_Outcome LIKE '%drone%';
```

- Task 10 – „ 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.”
  - Code:



```
%%sql
SELECT
    DISTINCT(Landing_Outcome),
    COUNT(*) AS 'COUNT'
FROM SPACEXTBL
GROUP BY Landing_Outcome
HAVING Date BETWEEN '04-06-2010' AND '20-03-2017'
ORDER BY COUNT(*) DESC;
```

- Citations from Lab 4

Build an Interactive Map with Folium

The conclusions presented in this slide are based on:
(must include completed code cell and outcome cell)

Using Folium following folium features have been created and added to presented maps:

- Markers and circles – adding visible points with Lunching location and with pop-up descriptions to localize what specific point is diagnosed.
- MarkerCluster – to dynamically see how points are clustering
- PolyLine – displaying lines between selected two points to visualize the distance between them.

Build a Dashboard with Plotly Dash

The conclusions presented in this slide are based on:
(must include completed code cell and outcome cell)

Using Dashboard following plots/graphs have been created and added to a dashboard:

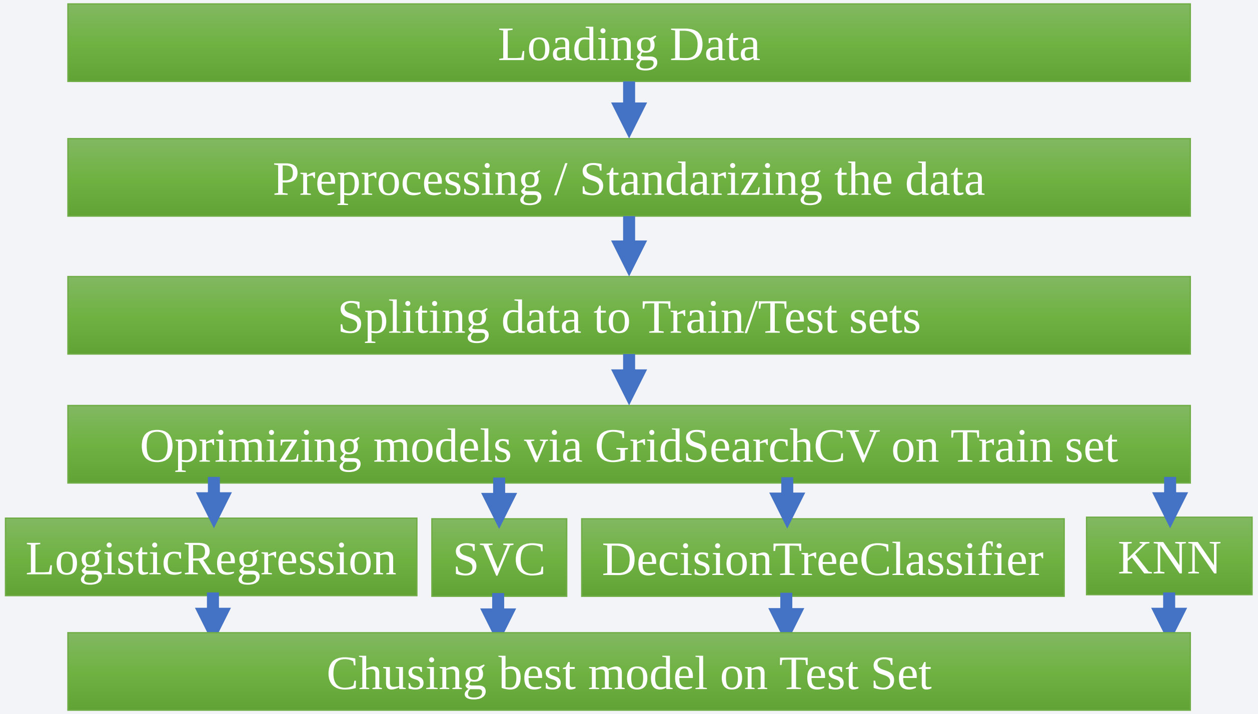
- Piechart allows dynamically selecting sites for which the success count of launch has been presented.
  - Dropdown lists have been used allowing the selection of certain sites or all of them at once.
- Scatter plot allows dynamically selecting a range of payload on a graph Payload vs. Launch which allows analyzing Payload vs. Launch
  - RangeSlider has been used to allow the selection of a desired payload range.

Predictive Analysis (Classification)

The conclusions presented in this slide are based on:
(must include completed code cell and outcome cell)

- Data have been loaded via pandas (90 records)
- Data have been standardized using StandardSlicer
- Data have been split into Train/Test split with proportion (Train (80%) / Test (20%))
- Models (LogisticRegression, SVC, DecisionTreeClassifier, and KNeighborsClassifier) have been optimized via GridSearch with cv=10
- Confusion matrices and accuracy for each optimized model have been calculated
- Accuracy using Test data have been calculated for each model.
- The best model has been selected as Logistic Regression with a test accuracy of 83.33 %.

Breef flowchart describing model development:



Loading Data

Preprocessing / Standarizing the data

Spliting data to Train/Test sets

Oprimizing models via GridSearchCV on Train set

LogisticRegression

SVC

DecisionTreeClassifier

KNN

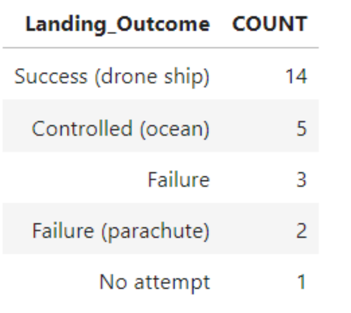
Chusing best model on Test Set

- Based on Lab 8

Results

• Exploratory data analysis results

- We can see that as flight number progresses, so does the success rate
- Orbits with the highest success rate (of 100%) are: ES-L1, GEO, HEO, and SSO
- the success rate of Launch Success Yearly Trend to increase from 2013 till 2020
- The total payload mass carried by boosters launched by NASA (CRS) is 45596.000 kg
- The average payload mass carried by booster version F9 v1.1 is 2543.667 kg
- The total number of successful and failed SpaceX missions outcomes: 100 successful and 1 failure.
- The names of the booster which have carried the maximum payload mass are F9 B5 B10(48.4, 49.4, 51.3, 56.4, 48.5, 51.4, 49.5, 60.2, 58.3, 51.6, 60.3, 49.7)
- The count of landing outcomes between the dates 2010-06-04 and 2017-03-20:



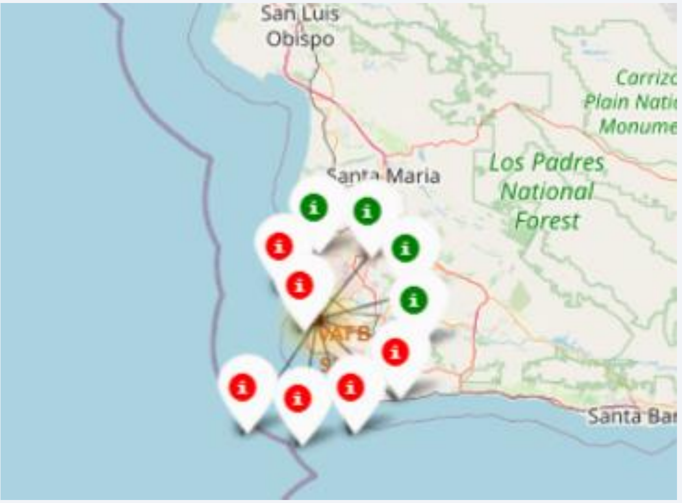
Landing_Outcome	COUNT
Success (drone ship)	14
Controlled (ocean)	5
Failure	3
Failure (parachute)	2
No attempt	1

Results

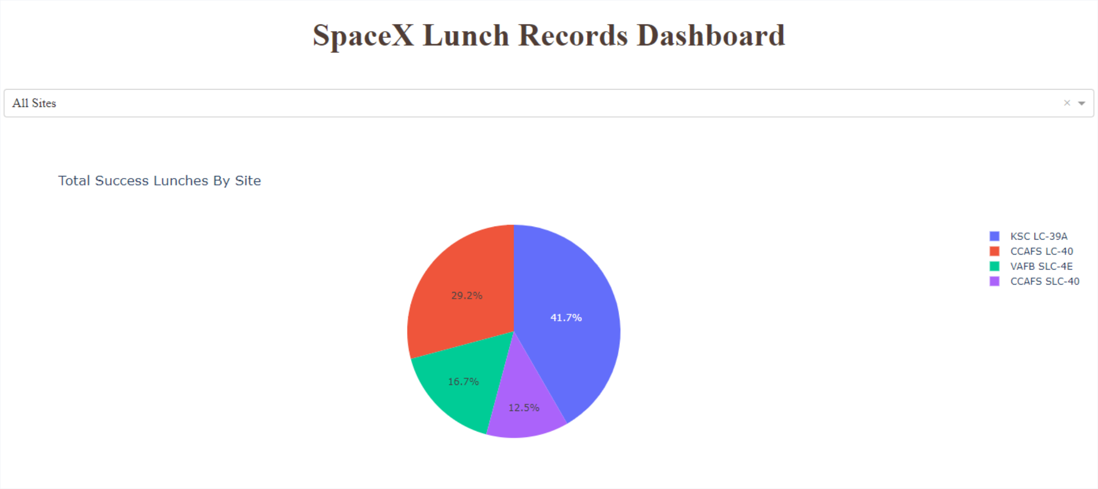
• Interactive analytics demo in screenshots

In order to boost analyses interactive analysis has been presented, e.g.

- Interactive maps:



- Interactive graphs:



SpaceX Lunch Records Dashboard

All Sites

Total Success Lunches By Site

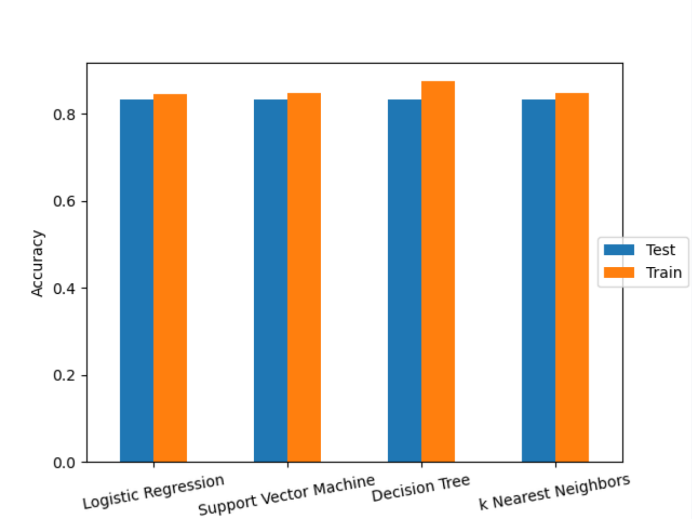
Results

• Predictive analysis results

As an outcome of predictive analysis, a predictive model predicting the successfulness of a rocket landing has been developed.

The best-found model is the Decision Tree Classifier with an accuracy of 83.33%, a specificity of 50%, and sensitivity 100%.

The accuracy of each tested model is presented below:

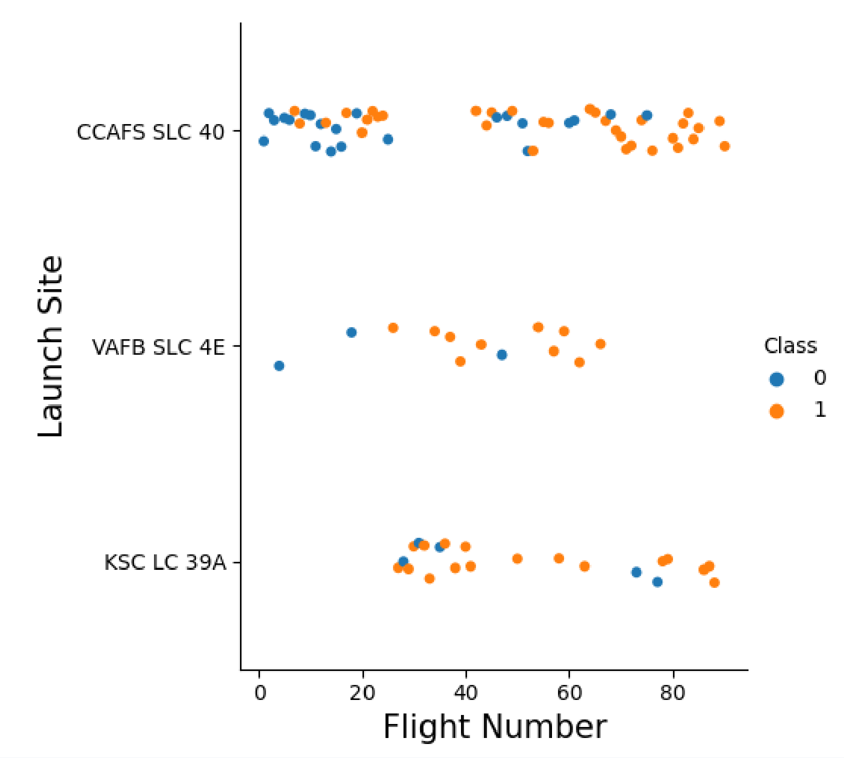


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- We can see that as Flight number progresses, so does the success rate (class 1 starts to dominate over class 0)
- We can see that CCAFS SLC 40's most recent launches were the most successful followed by KSC LC 39A and lastly, the worst is VAFB SLC 4E
- We can see that CCAFS SLC 40 has a majority of points so choosing the above proportions or KSC LC 39A and VAFB SLC 4E might be insufficient.

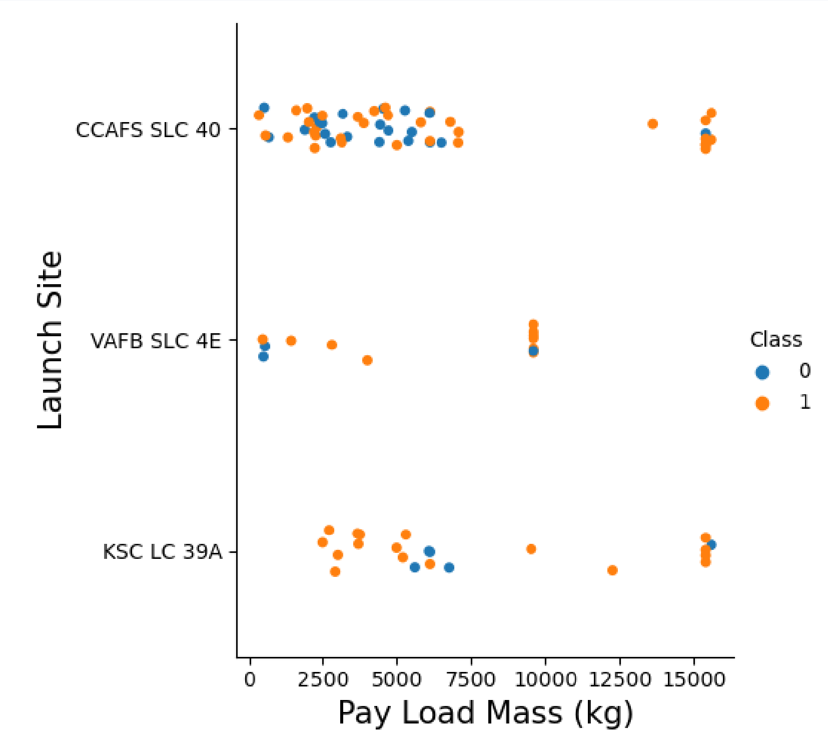


Scatter plot of Flight Number vs. Launch Site.

Payload vs. Launch Site

- We can see that ,, for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).”
- We can also observe that presented points tend to cluster.

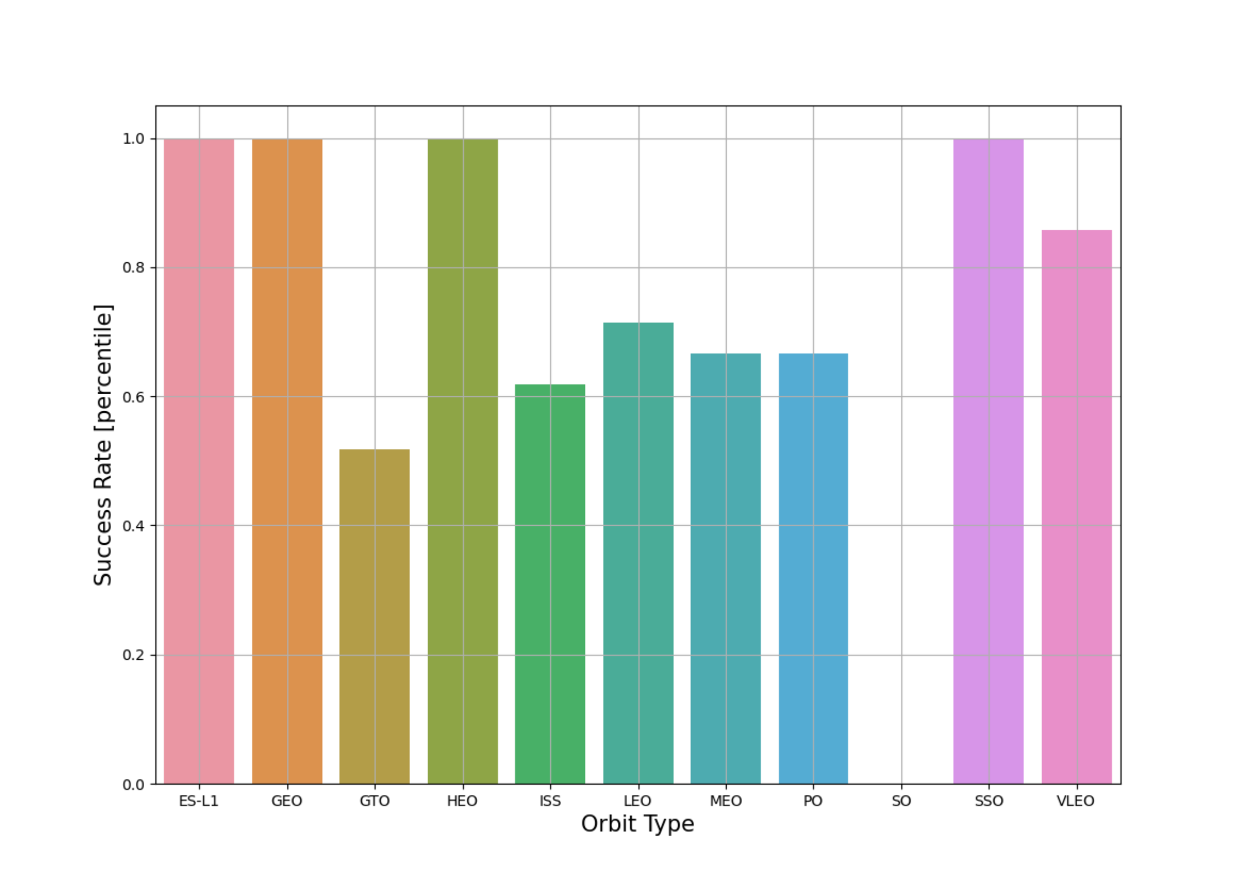
- Citations from Lab 5



Scatter plot of Payload vs. Launch Site.

Success Rate vs. Orbit Type

- We can see that the **orbits with the highest success rate (of 100%) are: ES-L1, GEO, HEO, and SSO.**
- The second success rate, followed by 100% is VLEOs’ 70%.
- We can see that near 55% success rate orbits of type GTO, IS, LEO, MEO, and PO.
- And lastly, we can see that SO has a success rate of 0%.

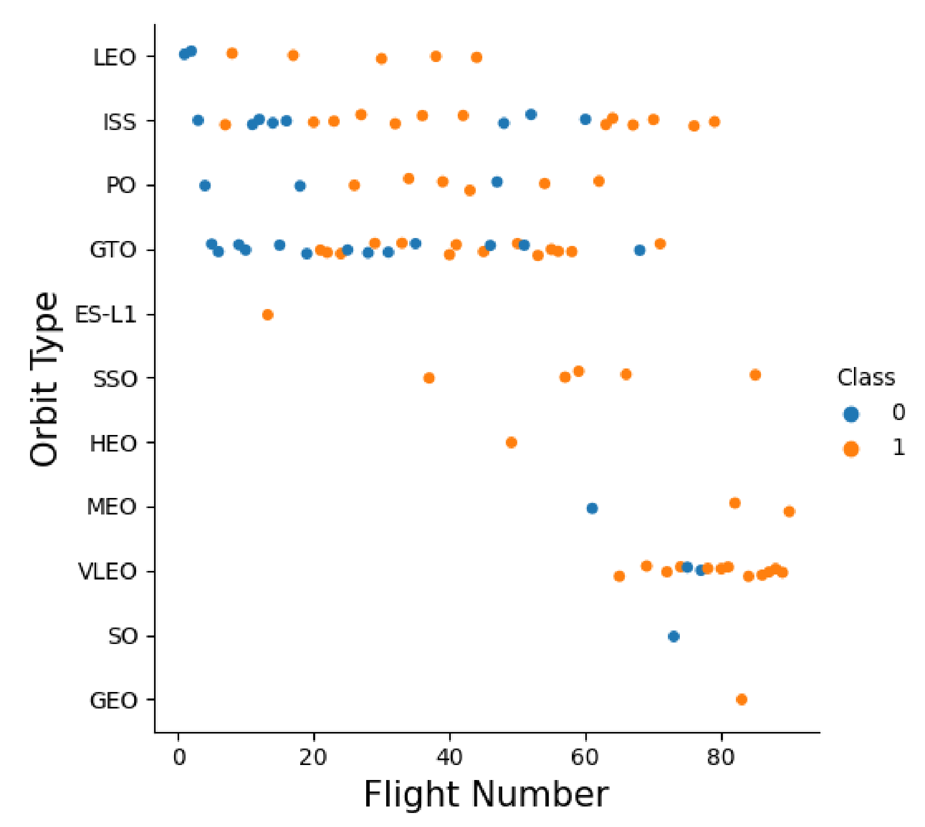


Bar chart for the success rate of each orbit type.

Flight Number vs. Orbit Type

- We can „see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit”.
- For the last couple of attempts (Flight Number), we can see that LEO, ISS, MEO, and VLEO seem to look promising – the trend of success.
- We can see that points locally tend to group within the same class.
- The shown graph looks like projected onto lower-dimensional space.

- Citations from Lab 5

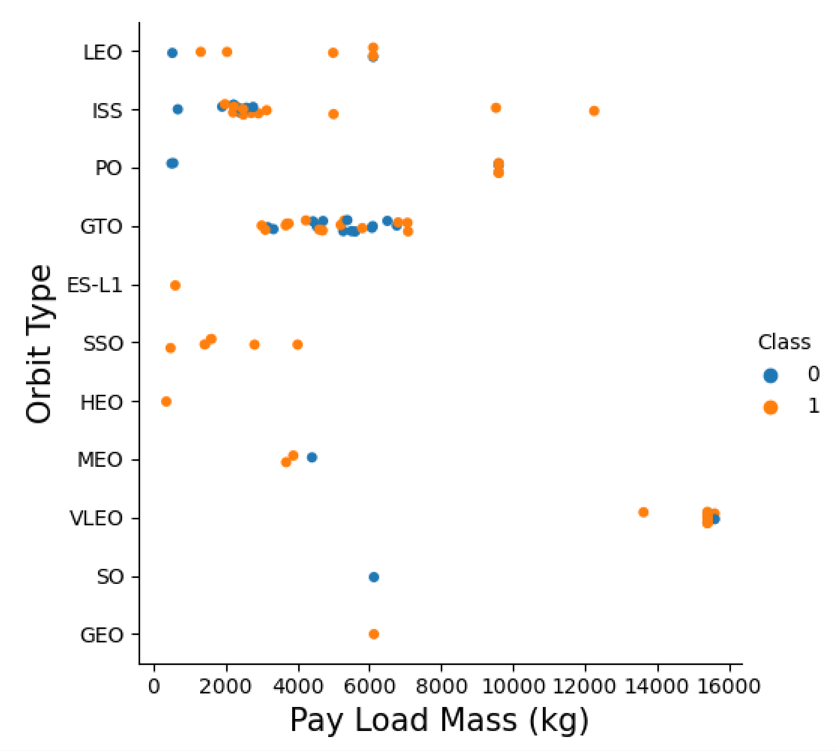


Scatter plot of Flight number vs. Orbit type.

Payload vs. Orbit Type

- We can see that for almost all orbit types payload mass aggregates into clusters (similar masses were used), but within those clusters both success and failure are indistinguishable.
- „With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.”
- „However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.”

- Citations from Lab 5

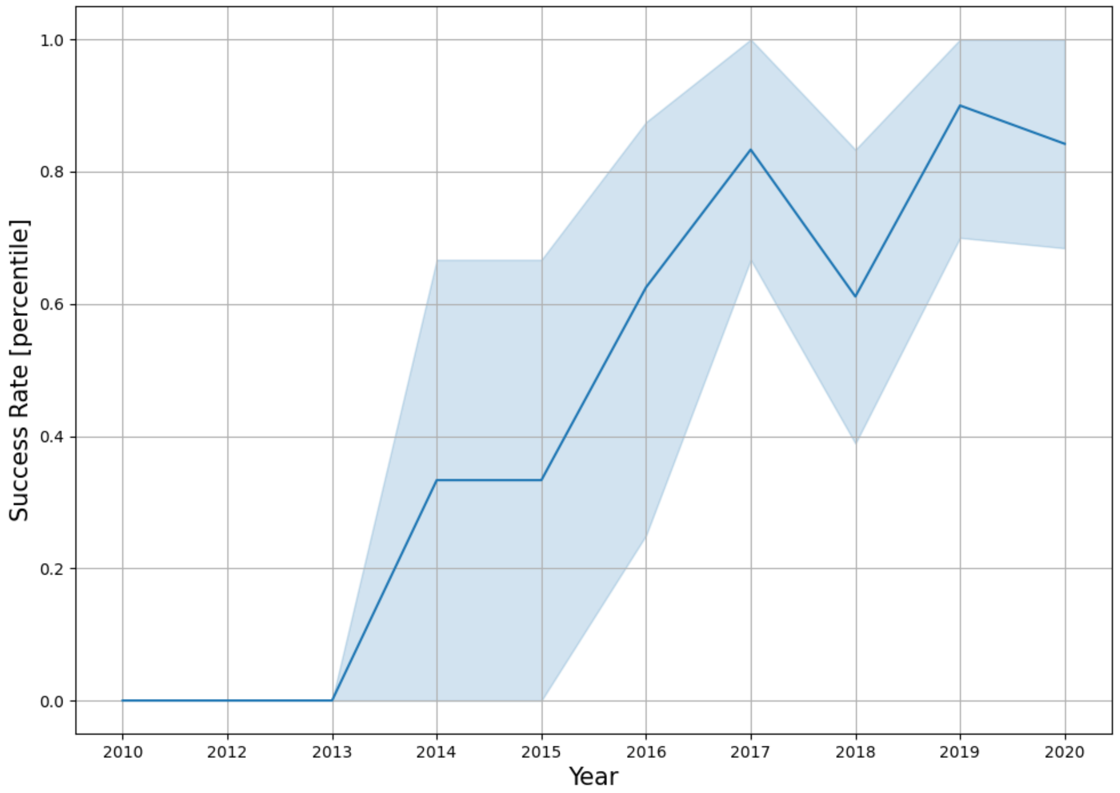


Scatter plot of payload vs. Orbit type.

Launch Success Yearly Trend

- One „can observe that the sucess rate since 2013 kept increasing till 2020”
- We can see relatively large errors that have been established. In my opinion, it would be good to say that SpaceX launched so many rockets that we can observe a relatively small percentage of failures, and on average Sucess Rate increased.

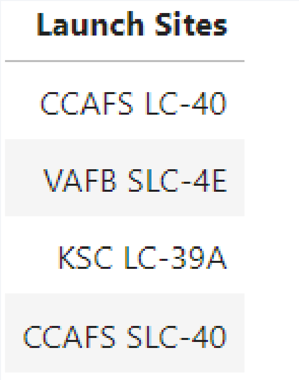
- Citations from Lab 5



Line plot of year vs the average launch success trend.

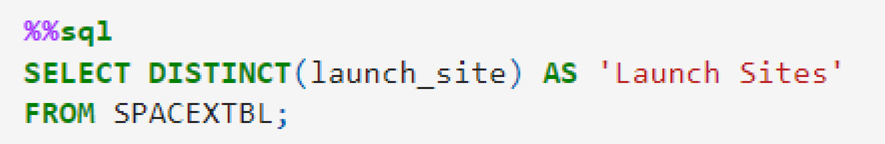
All Launch Site Names

The names of the unique launch sites are:



Launch Sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

My query:

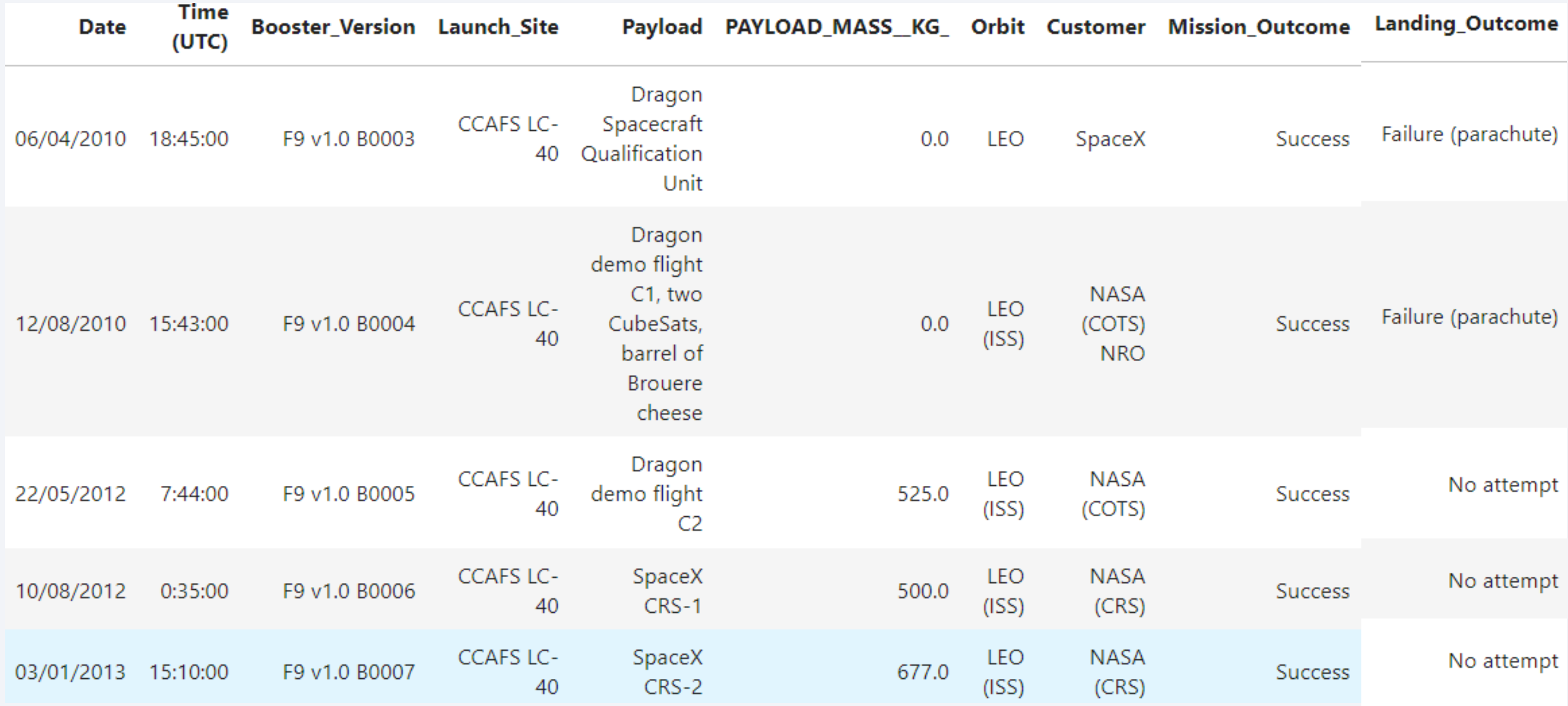


```
%%sql
SELECT DISTINCT(launch_site) AS 'Launch Sites'
FROM SPACEXTBL;
```

By using the DISTINCT function on launch sites, I select only those columns that do not repeat - they are unique launch sites.

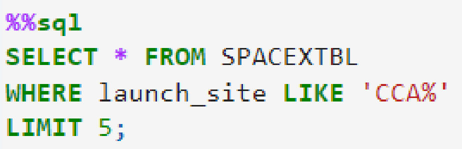
Launch Site Names Begin with 'CCA'

The 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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

My query:

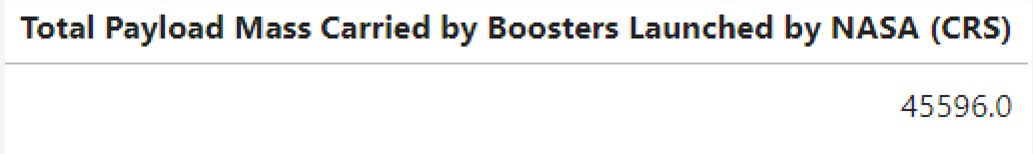


```
%%sql
SELECT * FROM SPACEXTBL
WHERE launch_site LIKE 'CCA%'
LIMIT 5;
```

By using ’*’ I’m selecting a whole row, by using LIKE ’CCA%’ I’m selecting all lunch_sites’ that start with ’CCA’, by using LIMIT 5, I’m only displaying 5 rows.

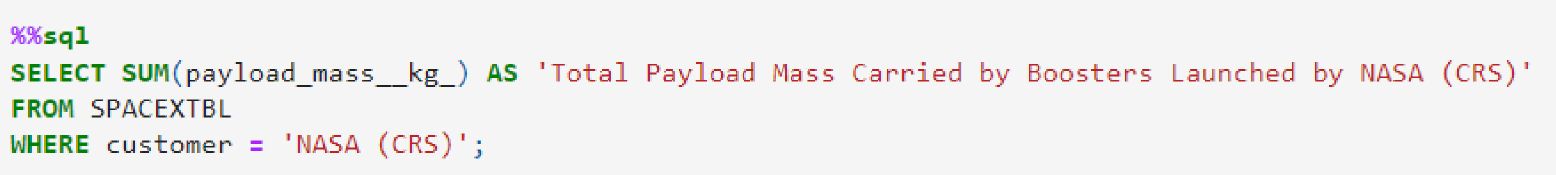
Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS):



Total Payload Mass Carried by Boosters Launched by NASA (CRS)
45596.0

My query:



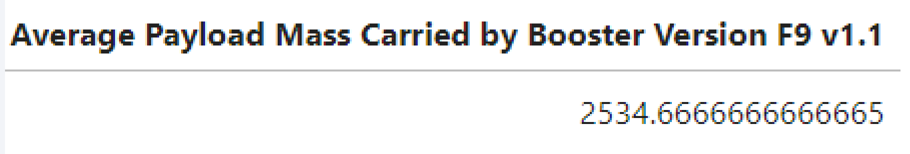
```
%%sql
SELECT SUM(payload_mass__kg_) AS 'Total Payload Mass Carried by Boosters Launched by NASA (CRS)'
FROM SPACEXTBL
WHERE customer = 'NASA (CRS)';
```

By using the SUM function I’m adding all accounted payload_mass__kg_ and

by using ”WHERE costumer = ’NASA (CRS)’” I’m selecting only costumers that are NASA (CRS).

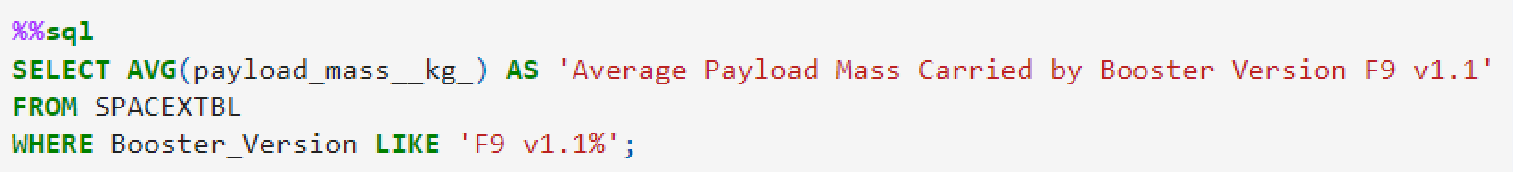
Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1



Average Payload Mass Carried by Booster Version F9 v1.1
2534.6666666666665

My query:



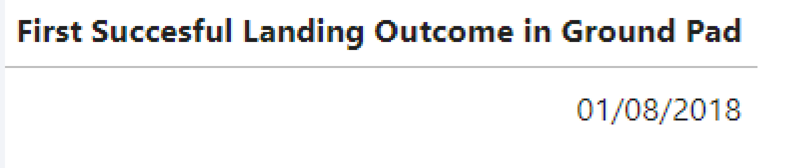
```
%%sql
SELECT AVG(payload_mass__kg_) AS 'Average Payload Mass Carried by Booster Version F9 v1.1'
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1%';
```

By using the AVG function I’m taking an average of all accounted payload_mass__kg_ and

by using ”WHERE Booster_Version LIKE ’F9 v1.1%’” I’m selecting only Booster_Versions that starts with the ’F9 v1.1%’ – selecting all F9 v1.1.

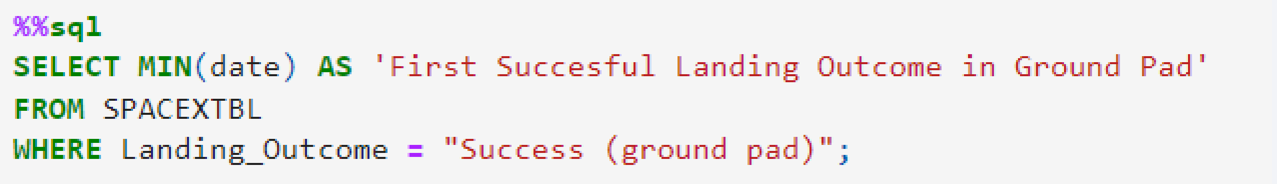
First Successful Ground Landing Date

The date of the first successful landing outcome on the ground pad:



First Succesful Landing Outcome in Ground Pad
01/08/2018

My query:

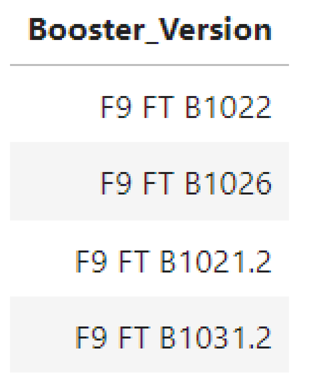


```
%%sql
SELECT MIN(date) AS 'First Succesful Landing Outcome in Ground Pad'
FROM SPACEXTBL
WHERE Landing_Outcome = "Success (ground pad)";
```

By selecting MIN(DATE) I’m selecting only the lowest date in DATE set, by using ’ Landing_Outcome = "Success (ground pad) ’ I’m selecting only Landing Outcome on the ground pad that is successful.

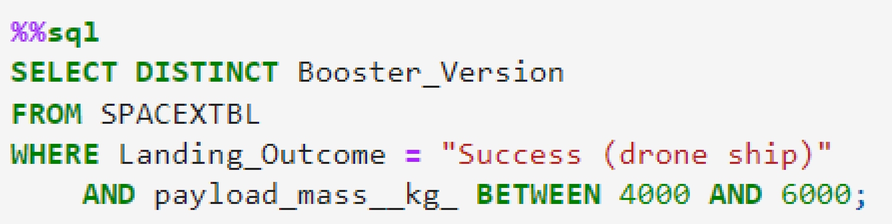
Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters that have successfully landed on drone ships 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

My query:



```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome = "Success (drone ship)"
    AND payload_mass__kg_ BETWEEN 4000 AND 6000;
```

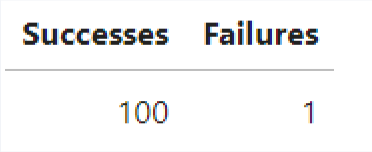
By using the DISTINCT function on Booster_Version I’m selecting only columns that do not repeat – they are unique.

By using ” Landing_Outcome = "Success (drone ship)" ” I’m selecting only Landing_Outcomes that are "Success (drone ship)„.

By using ” payload_mass__kg_ BETWEEN 4000 AND 6000” I’m selecting only payload_mass__kg_ from the range [4000, 6000] kg.

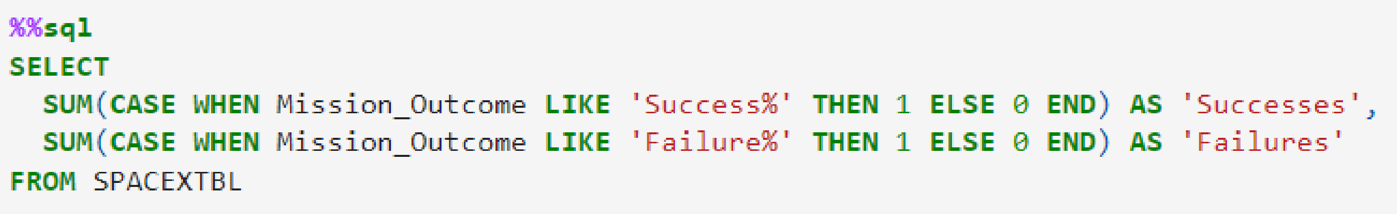
Total Number of Successful and Failure Mission Outcomes

The total number of successful and failed mission outcomes :



Successes	Failures
100	1

My query:



```
%%sql
SELECT
  SUM(CASE WHEN Mission_Outcome LIKE 'Success%' THEN 1 ELSE 0 END) AS 'Successes',
  SUM(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 ELSE 0 END) AS 'Failures'
FROM SPACEXTBL
```

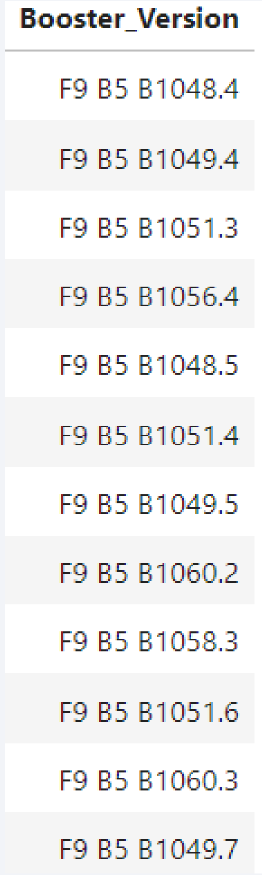
By using AS function I’m using an alias on a selected column,

By using SUM(CASE WHEN Mission_Outcome LIKE 'Success%' THEN 1 ELSE 0 END) I’m counting all Mission_Outcome values that start with ’Success’ and similarly by using SUM(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 ELSE 0 END) I’m counting all Mission_Outcome values that start with ’ Failure’.

In SQL a CASE function is similar to Pythons’ ’elif ’ function where WHEN Mission_Outcome LIKE… is a statement and if it is correct then it will return 1 or else 0 is returned. All returned values are summed up via SUM and thus it works as a counter of values started with some phrase.

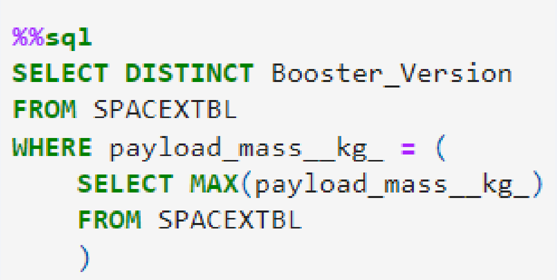
Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass:



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

My query:



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

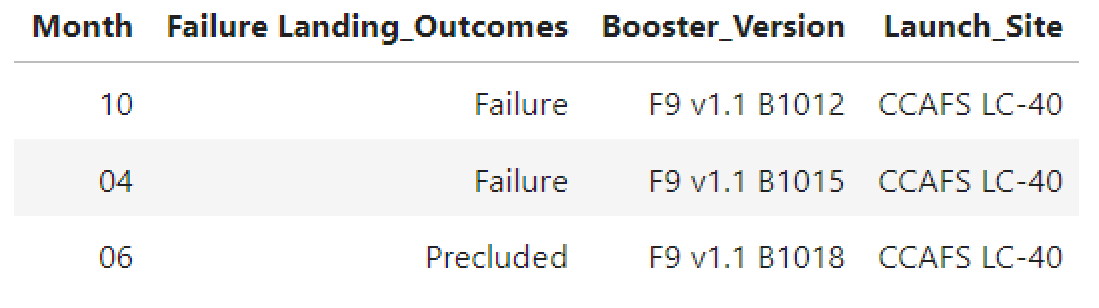
By using the DISTINCT function on Booster_Version I’m selecting only columns that do not repeat – they are unique.

With (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL) I’m selecting maximal payload_mass__kg_.

By combining the above features only Booster Versions that are unique and have the biggest payload_mass__kg_ are selected.

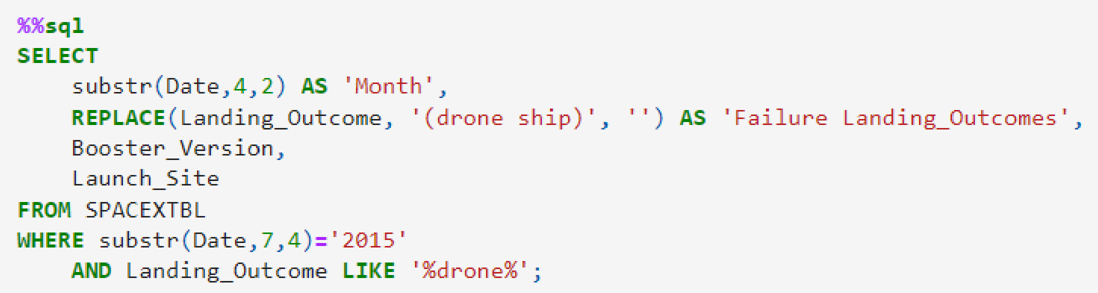
2015 Launch Records

The failed landing_outcomes in drone ships, their booster versions, and launch site names for the year 2015:



Month	Failure Landing_Outcomes	Booster_Version	Launch_Site
10	Failure	F9 v1.1 B1012	CCAFS LC-40
04	Failure	F9 v1.1 B1015	CCAFS LC-40
06	Precluded	F9 v1.1 B1018	CCAFS LC-40

My query:



```
%%sql
SELECT
    substr(Date,4,2) AS 'Month',
    REPLACE(Landing_Outcome, '(drone ship)', '') AS 'Failure Landing_Outcomes',
    Booster_Version,
    Launch_Site
FROM SPACEXTBL
WHERE substr(Date,7,4)='2015'
    AND Landing_Outcome LIKE '%drone%';
```

By using WHERE substr(Date,7,4)='2015’ only the Date that is 2015 is selected.

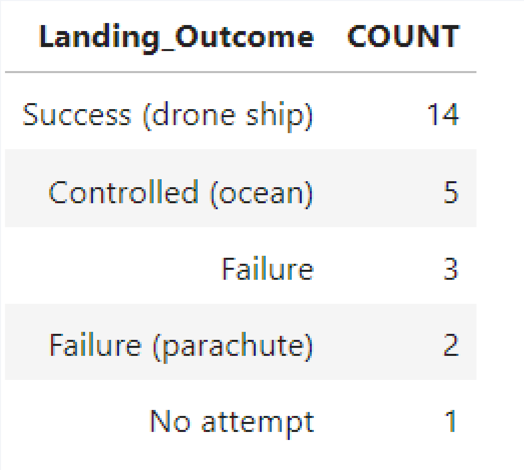
By using Landing_Outcome LIKE '%drone%' only Landing_Outcome that contains the phrase 'drone' is selected.

By using REPLACE(Landing_Outcome, '(drone ship)', ,’) phrase '(drone ship)’ in Landing_Outcome is not displayed.

By using substr(Date,4,2) a month from Date is selected.

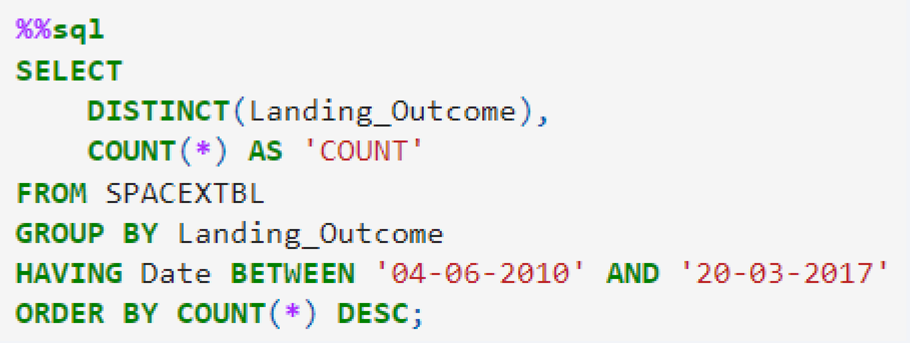
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

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



Landing_Outcome	COUNT
Success (drone ship)	14
Controlled (ocean)	5
Failure	3
Failure (parachute)	2
No attempt	1

My query:

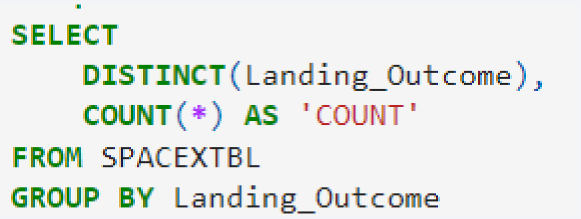


```
%%sql
SELECT
    DISTINCT(Landing_Outcome),
    COUNT(*) AS 'COUNT'
FROM SPACEXTBL
GROUP BY Landing_Outcome
HAVING Date BETWEEN '04-06-2010' AND '20-03-2017'
ORDER BY COUNT(*) DESC;
```

By using ORDER BY COUNT(*) DESC we can see the list in descending order.

By using HAVING Date BETWEEN '04-06-2010' AND '20-03-2017’ only dates between '04-06-2010’ and '20-03-2017’ are considered.

By using landing outcomes are counted.

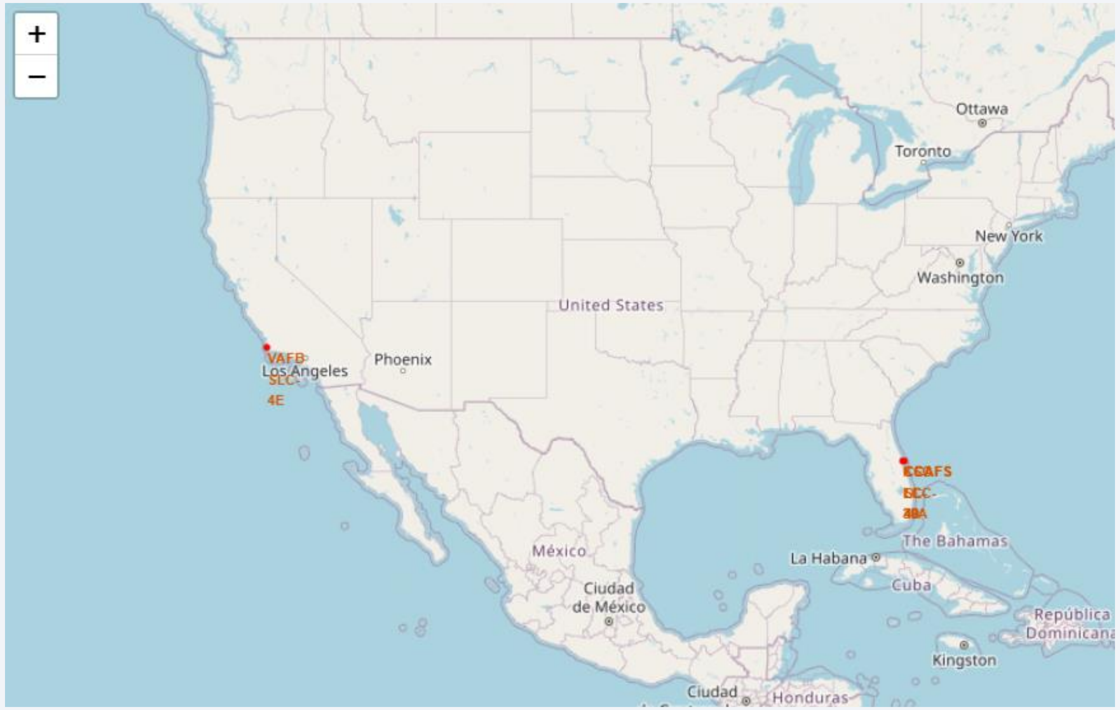


```
SELECT
    DISTINCT(Landing_Outcome),
    COUNT(*) AS 'COUNT'
FROM SPACEXTBL
GROUP BY Landing_Outcome
```

Section 3

Launch Sites Proximities Analysis

SpaceX Rockets Lunch Locations

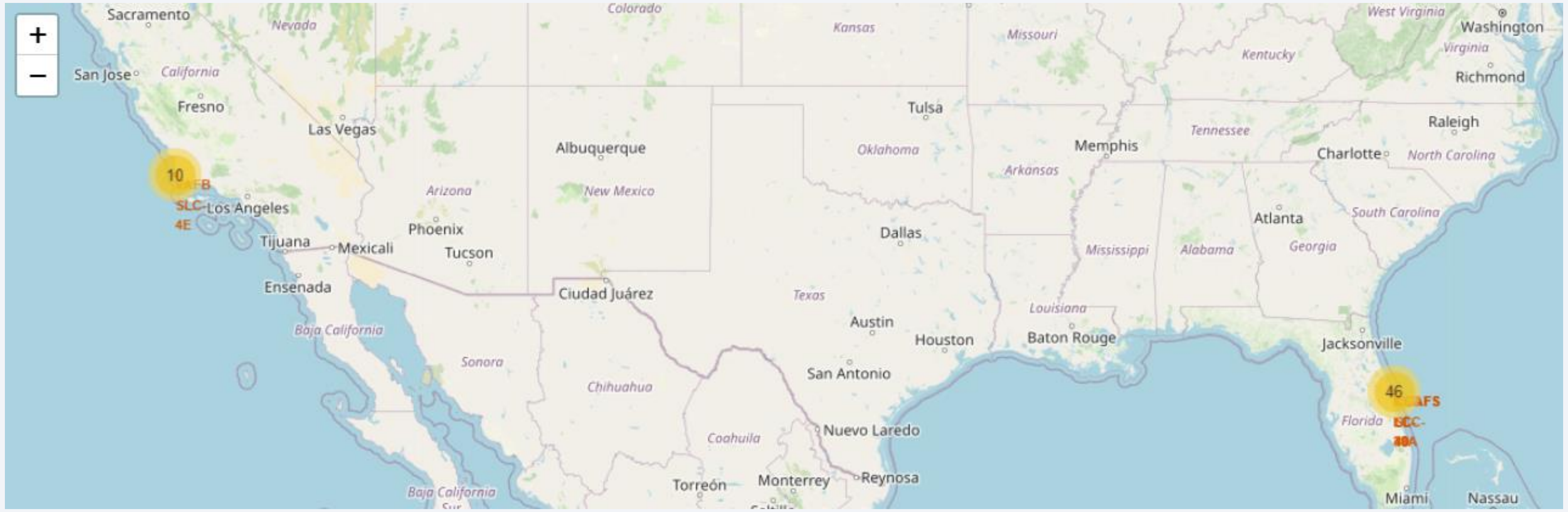


Map of SpaceX launch sites' location markers on a global map are pointed by red dots.

We can see those chosen locations are far away, which probably allows us to change locations e.g. if bad weather conditions etc have been observed.

We also can see that both locations are near the ocean which can allow us to point damaged rockets into a sea and not to a populated location.

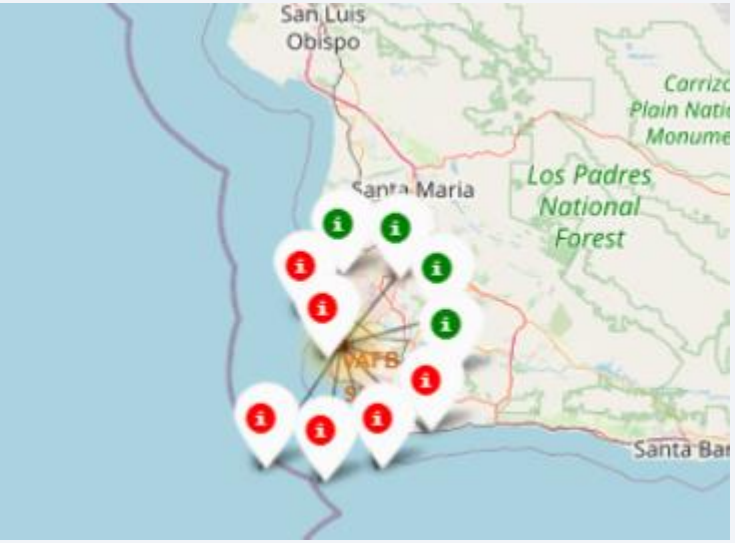
Launch Outcomes (part 1/2)

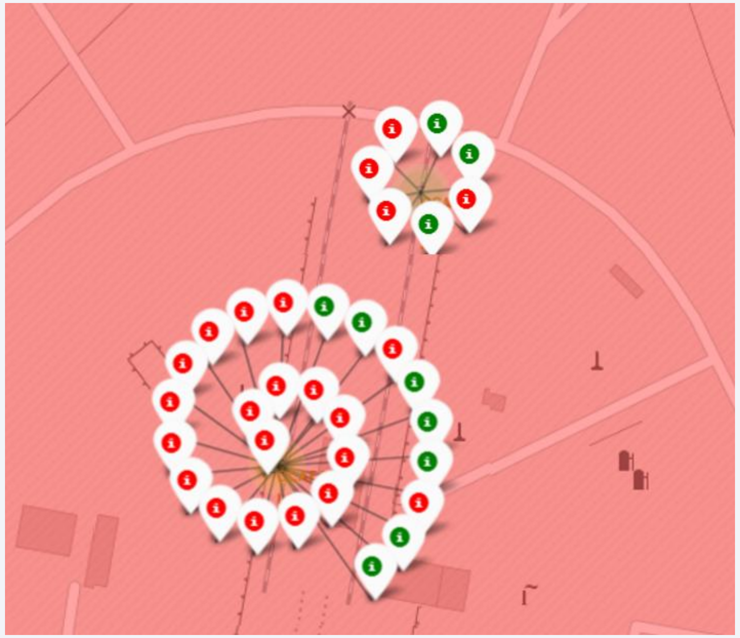


Map of SpaceX launch sites' location markers on a global map where numbers on yellow circles represent the number of lunches according to locations left (VAFB SLC-4E) had 10 lunches and right of the map (CCAFS (SLC-40, LC-40) had 46 lunches.

The right launching location is closer to the NASA base which may be a cause higher amount of lunches at that location.

Launch Outcomes (part 2/2)





The above maps represent SpaceX launch sites’ locations left (VAFB SLC-4E) and right of the map (CCAFS (SLC-40, LC-40) where red markers represent failer and green successful lunch.

We can see that each location is with nearly the same proportion of failed to succeed.

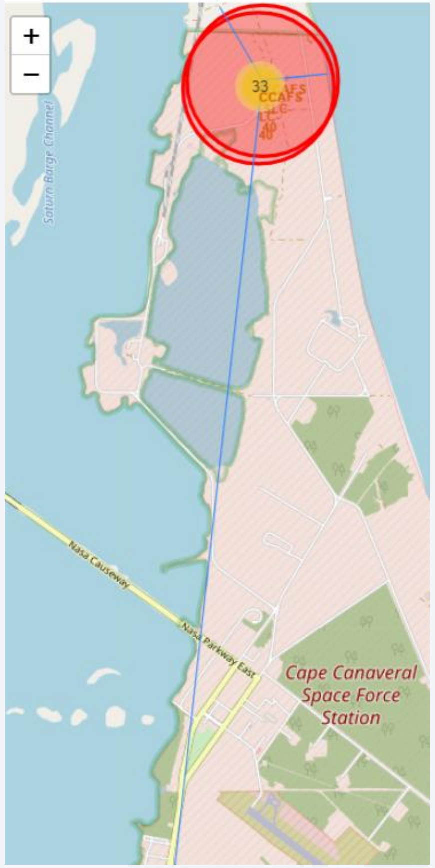
We can see that 2 out of 3 lunching locations are similarly looking to Archimedes Spirale.

Importance of lunching location.

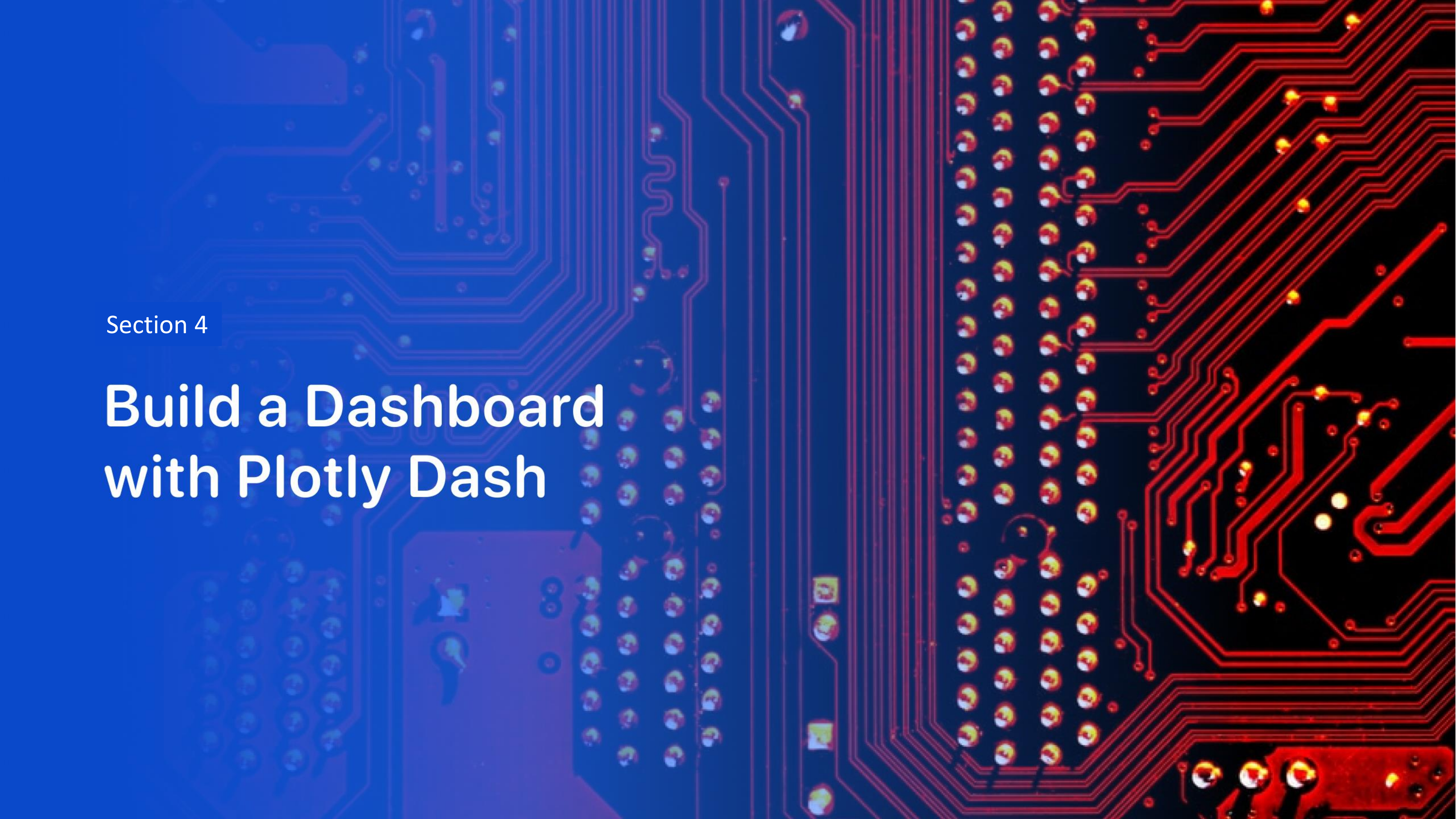
Distance from Launch Site to the nearest:

- Coastline is 0.86 km.
- Rairoad (NASA Rairoad) is 1.14 km.
- Pkwy (Samuel C Philips Pkwy) is 0.58 km.
- City (Melburne) is 51.30 km.

We can see that CCAFS lunch location is relativly close to the ocean, roads (rairoad, pkwy) and far from the city.



The above map represents SpaceX launch locations of the CCAFS (SLC-40, LC-40) as a yellow dot from which blue lines are provided to the nearest coastline point, Railroad, Pkwy, City.

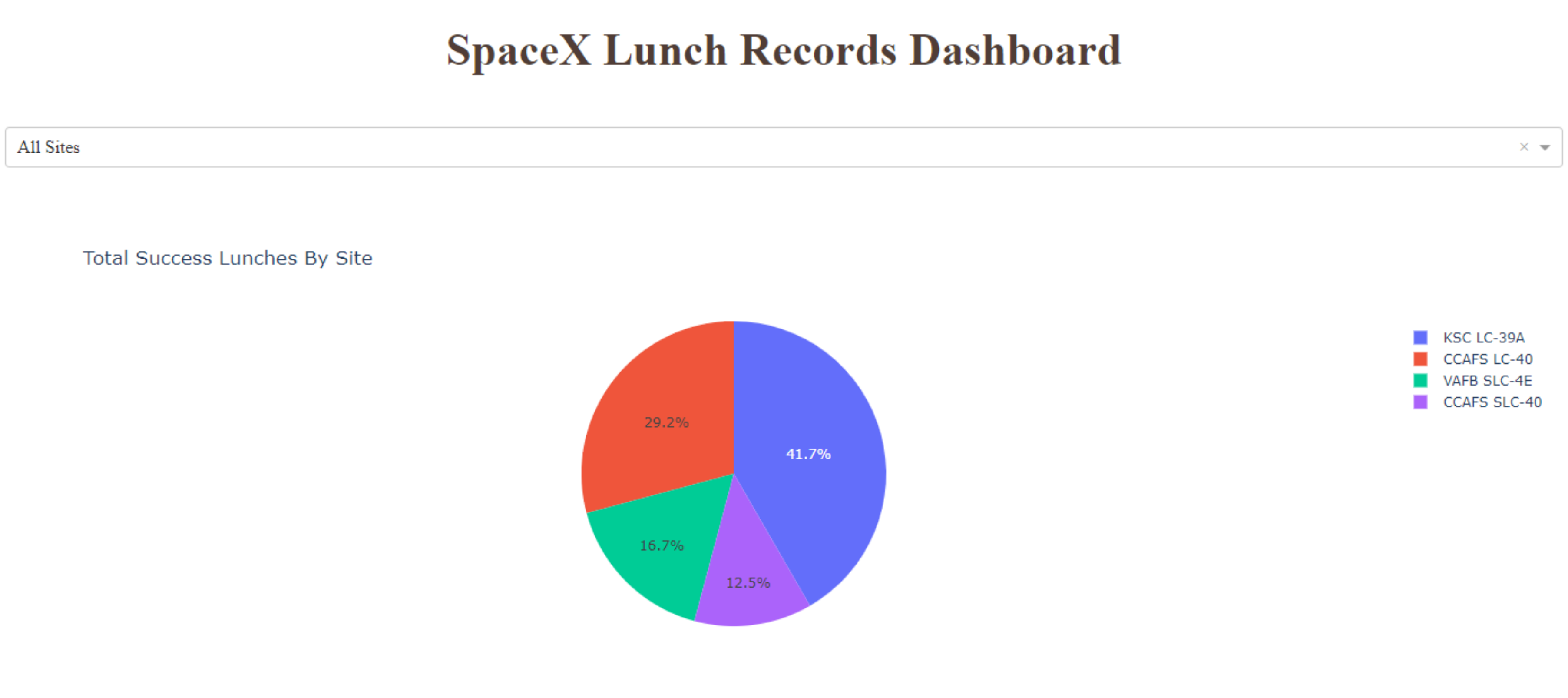


Section 4

Build a Dashboard with Plotly Dash

Quantitative analysis of launches success in all sites

- As we can see on the pie chart the majority of 41.7% of successful lunches were at KSC LC-39A followed by CCAFS LC-40 with 29.2% of successful lunches.
- The bottom two are VAFB SLC-4E (16.7%) and CCAFS SLC-40 (12.5%)



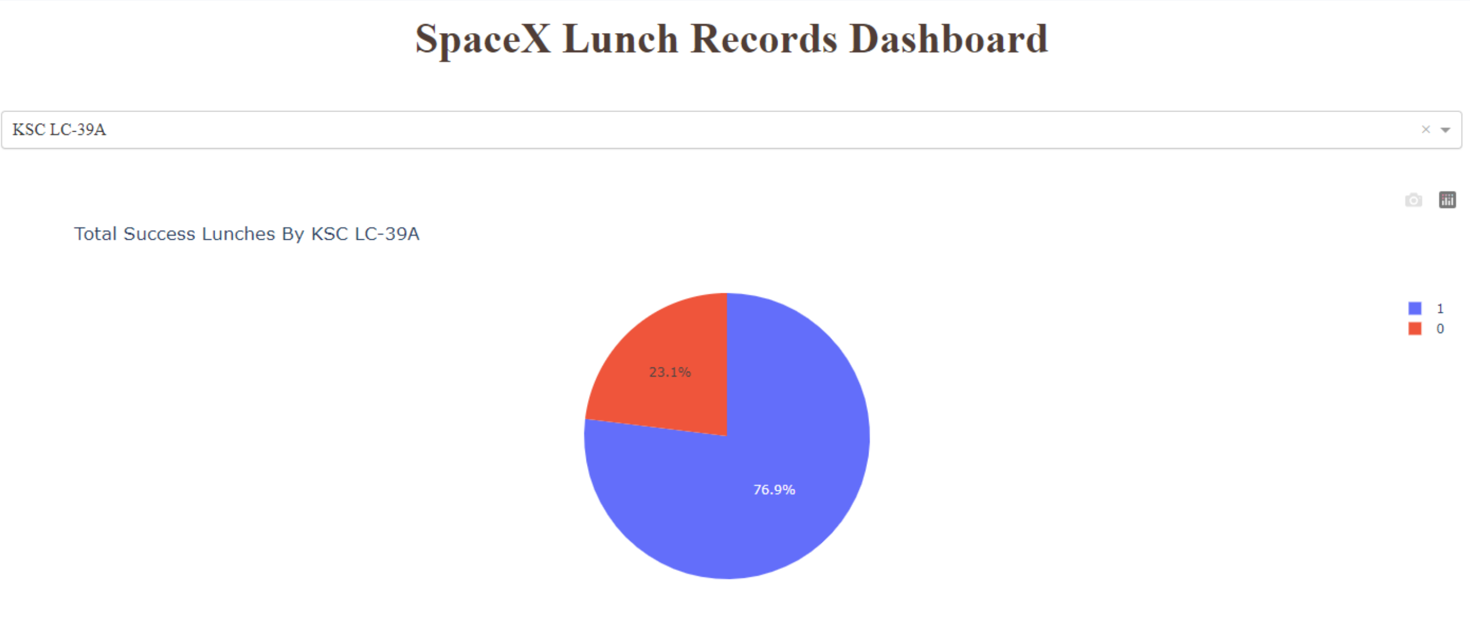
SpaceX Lunch Records Dashboard

All Sites

Pie Chart of launch success counts for all sites.

Analysis of successful launches at KSC LC-39A

As we can see on the graph on the right, the success rate is more than 75%, which knowing that KSC LC-39A has the most records of successful lunches (41.7% on the previous slide) makes KSC LC-39A the best candidate for launching a rocket and it is worth investigating what makes that specific place different from the others, making such high success rate difference.

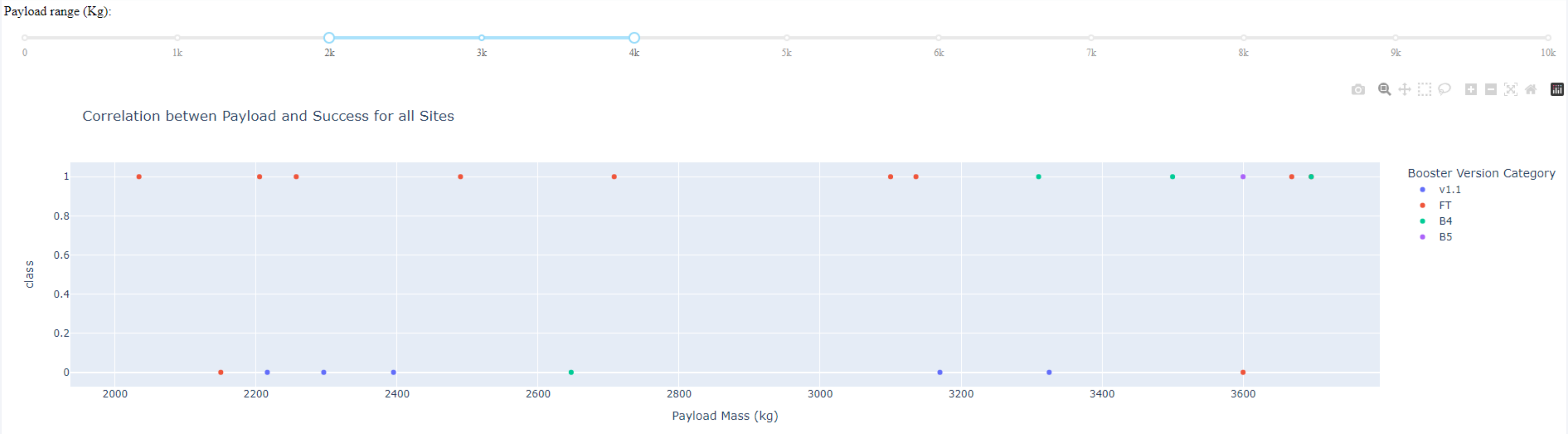


SpaceX Lunch Records Dashboard

KSC LC-39A

Piechart of the launch site with highest launch success ratio

Payload vs. Launch Outcome with different payload

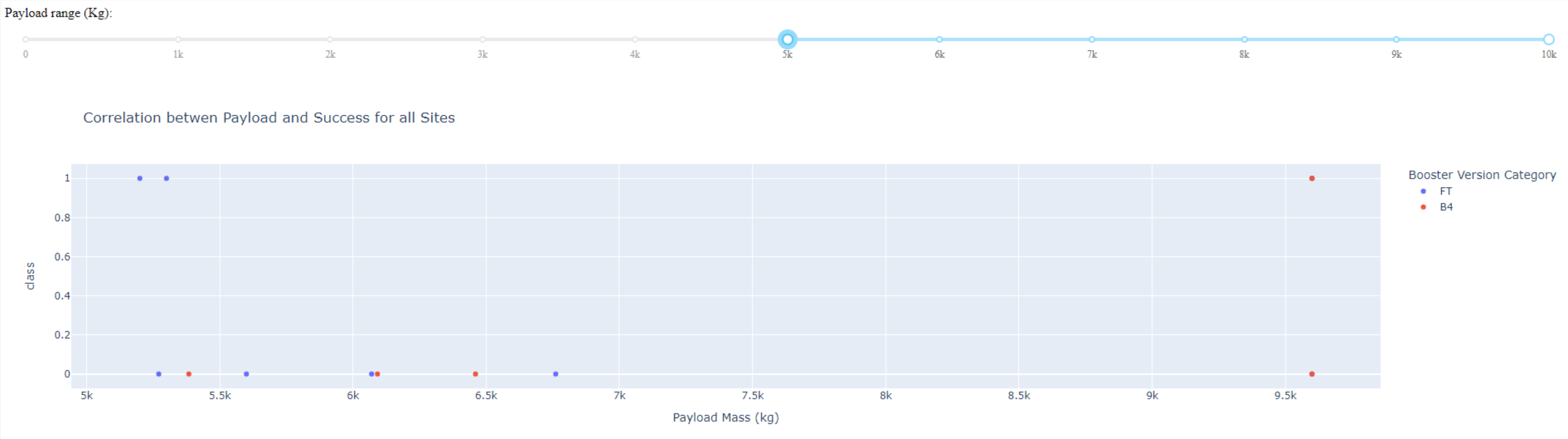


Scatter plot of Payload vs. Launch Outcome for all sites, with payload between 2,000 and 4,000 USD.

We can see that in the selected range almost only Booster Version v1.1 is failing, thus despite Booster Version v1.1 Boosters Ft, B4, and B5 have high success yield.

We can see that in the selected range mostly FT has been explored and within that range, FT has the highest success rate.

Payload vs. Launch Outcome with different payload



Scatter plot of Payload vs. Launch Outcome for all sites, with payload between 5,000 and 10,000 USD.

We can see that in the selected range there are only two Boosters tested FT and B4.

We can see that in the range of 5,000 to 5,500 USD FT performs with a 75% higher success rate than B4.

We can see that for the range of 5,500 to 7,000 USD all Boosters failed.

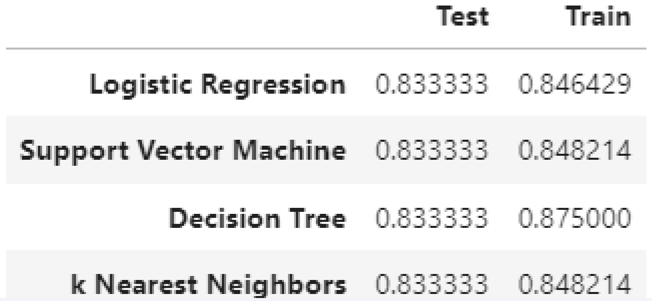
For 10,000 only one Booster has been used – B4 with a 50% success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

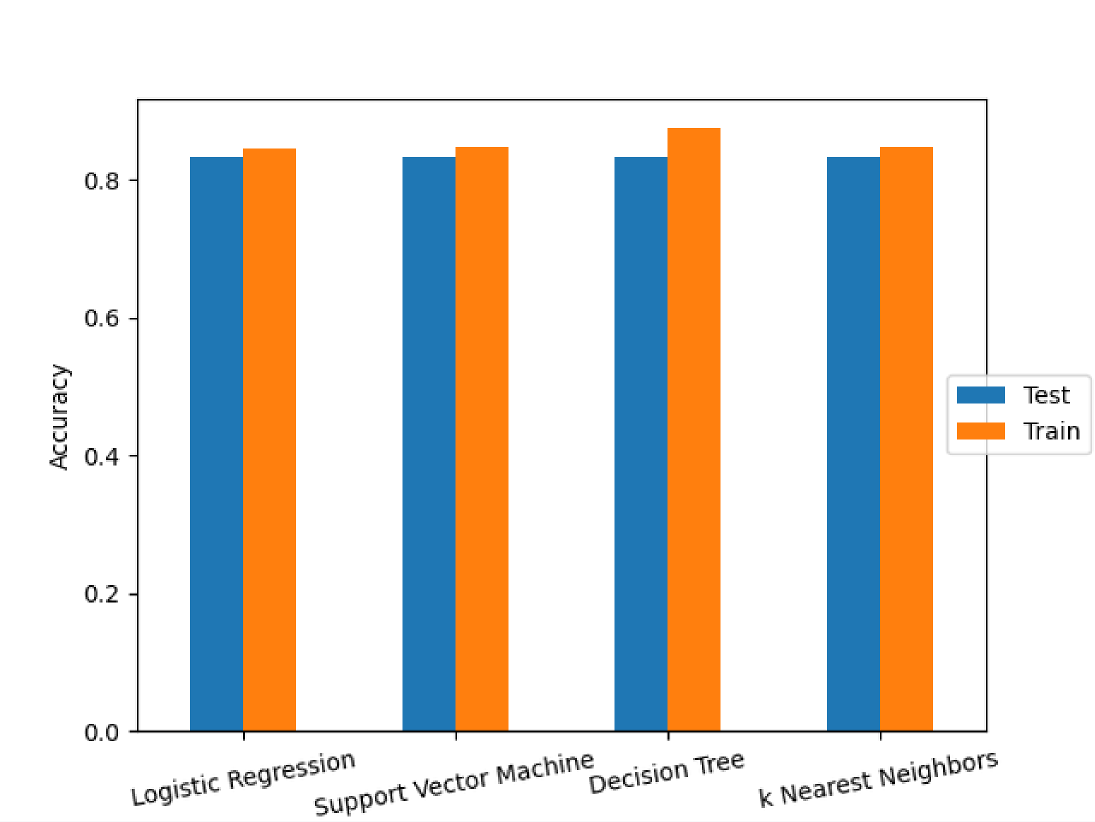
Data for the graph on the right:



	Test	Train
Logistic Regression	0.833333	0.846429
Support Vector Machine	0.833333	0.848214
Decision Tree	0.833333	0.875000
k Nearest Neighbors	0.833333	0.848214

We can see that all models have the same test accuracy of 83.33% so the model pointed by train accuracy is Decision Tree with 87.50%.

It is worth pointing out that we do not know the loss curves so we do not know whether or not models are overtrained/undertrained.



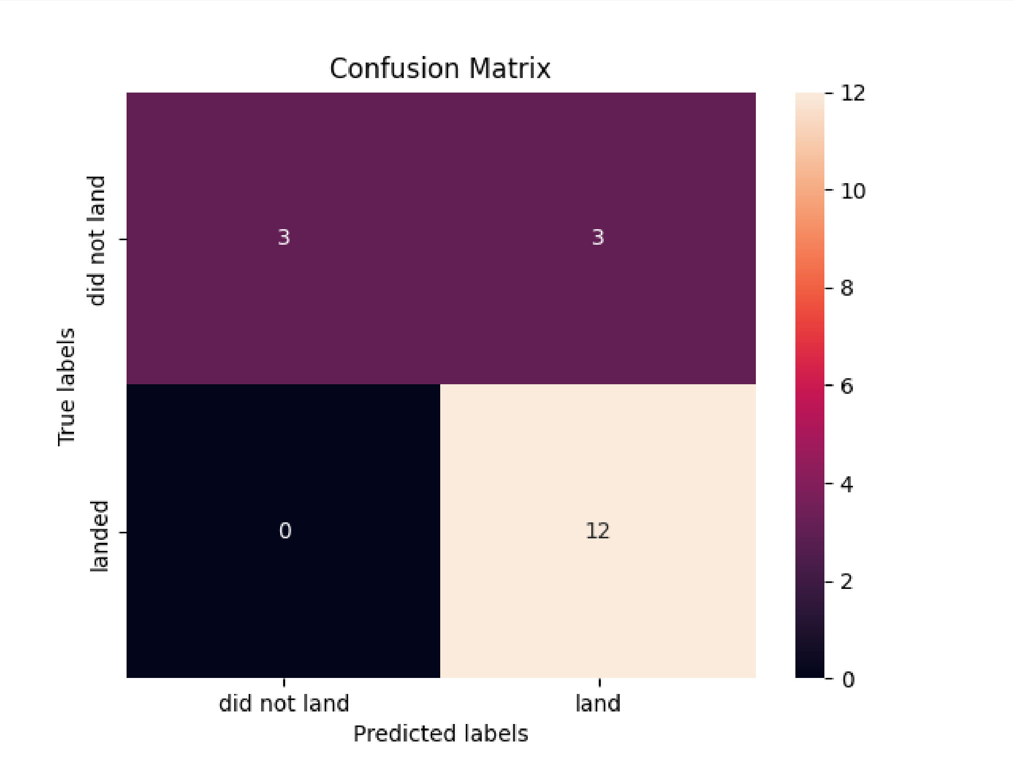
Bar chart of model accuracy for all built classification models.

Confusion Matrix

We can see that model specificity is 50% (type 1 errors are 50% when predicting ’did not land’ data) which may suggest that model is overfitted.

We can see that model sensitivity is 100% (the amount of correctly predicted true positives is 100% when predicting on ’land’ data).

As we worked with a small amount of data adding points by data augmentation might help to overfit.



Confusion matrix of the Decision Tree model.

Conclusions

- We can see that with the increasing amounts of lunches for the specific rocket, the success rate increases which might be interpreted as a correlation to „experience with building rockets”.
- The highest probability of success is at NASAS Lunching CCAFS (SLC-40, LC-40).
- Rocket with Booster Version FT from payload range between 2,000 and 4,000 USD have the highest success probability.
- The best predictive model predicting the success of the rocket landing with an accuracy of 83.33% has been found, thus four other analyses are recommended.

Appendix

As trained models seem overfitted plotting of loss function has been recommended.

As we are using a small amount of data, to prevent overfitting data augmentation has been recommended.

It is worth knowing that trained models have been created with a small range of parameters, thus a larger range/smaller seed might be used to find more accurate model parameters.

Thank you!