IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

A. Ziegler
20.07.2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection of past rocket launches with SpaceX REST API and Web Scrapping
  - Data wrangling to clean and filter data and to create a binary outcome variable
  - Explorative Data Analysis (EDA) using visualization and SQL
  - Interactive Visual Analytics using Folium and Plotly Dash
  - Predictive analysis using four classification models (Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors)
- Summary of key results
  - All launch sites are close to the equator line and the coast.
  - In the early years (2010-2013) there were only failed launches. Success rates since 2013 kept increasing.
  - All four tested classification models display the same confusion matrix and have the same accuracy ratio.

Introduction

- Project background and context
  - There is fierce competition in the rocket launch business.
  - SpaceX rocket launches are relatively inexpensive. It advertises Falcon 9 rocket launches with a cost of 62 million dollars, other providers cost more than 165 million dollars upwards.
  - Much of the savings is because SpaceX can reuse the first stage. Unlike other rocket providers, SpaceX's Falcon can recover the first stage. Thus, if we can determine if the first stage will land, we can determine the cost of a launch.
  - Based on publicly available information on past rocket launches from SpaceX and applying several classification models, we are going to predict if the first stage will land.
- Problems you want to find answers
  - How do variables such as e.g. the launch site location, payload mass, orbit type, booster version, etc. affect the success of the first stage landing?
  - Do the success rates increase over the years?
  - Which is the best binary classification model to predict the success of the first stage landing?

Section 1

Methodology

Methodology

Executive Summary

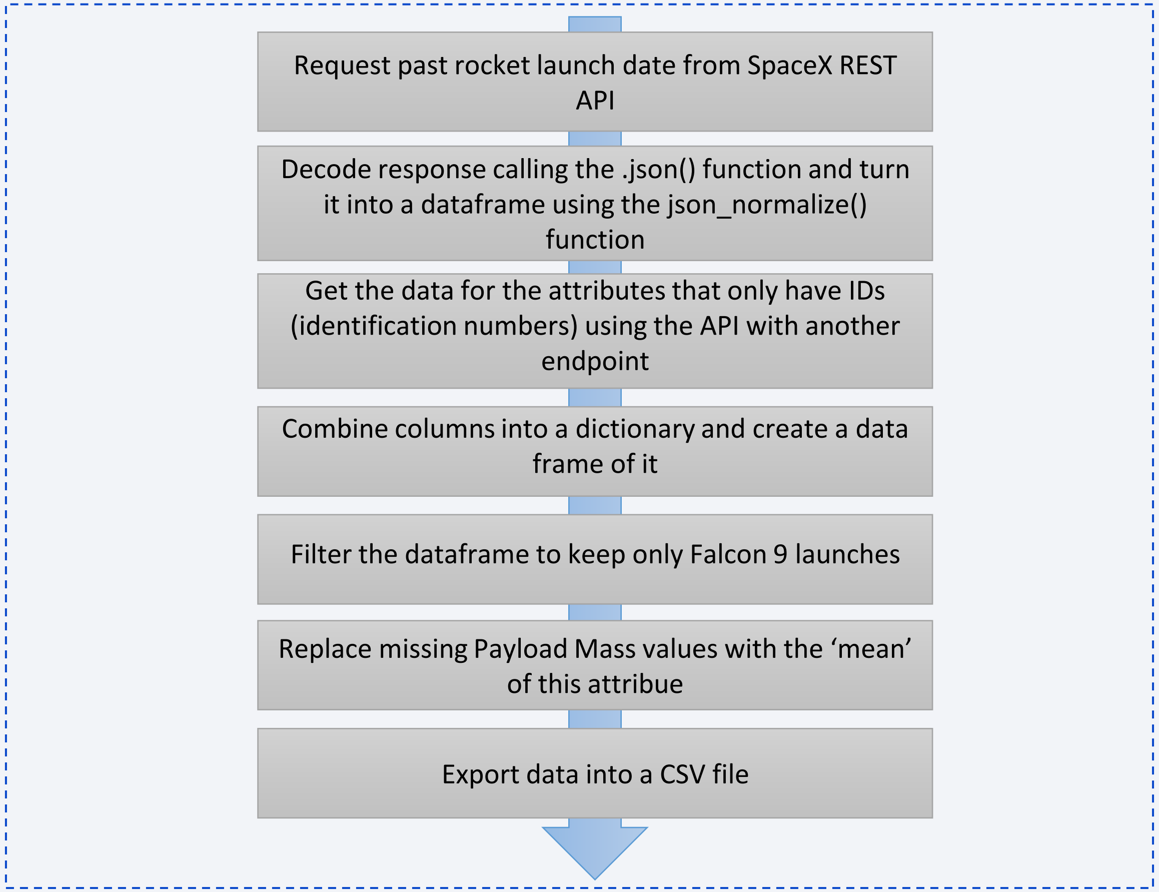
- Data collection methodology:
  - Via SpaceX Rest API and Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering data, dealing with missing values, creating a binary outcome variable (success/fail)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune and evaluate four classification models to get best performing model

Data Collection

- Data collection was performed based on two sources which contain valuable information about past launches that will be used to predict the success of the first stage landing:
    - Space X REST API (see slide 8)
    - Web Scrapping from Wikipedia (see slide 9)

Data Collection – SpaceX API

- On the right side the data collection with SpaceX REST API calls is being depicted. It is based on the following URL: https://api.spacexdata.com/v4/launches/past
- GitHub link to completed SpaceX API calls notebook: ->Link



Request past rocket launch date from SpaceX REST API

Decode response calling the .json() function and turn it into a dataframe using the json_normalize() function

Get the data for the attributes that only have IDs (identification numbers) using the API with another endpoint

Combine columns into a dictionary and create a data frame of it

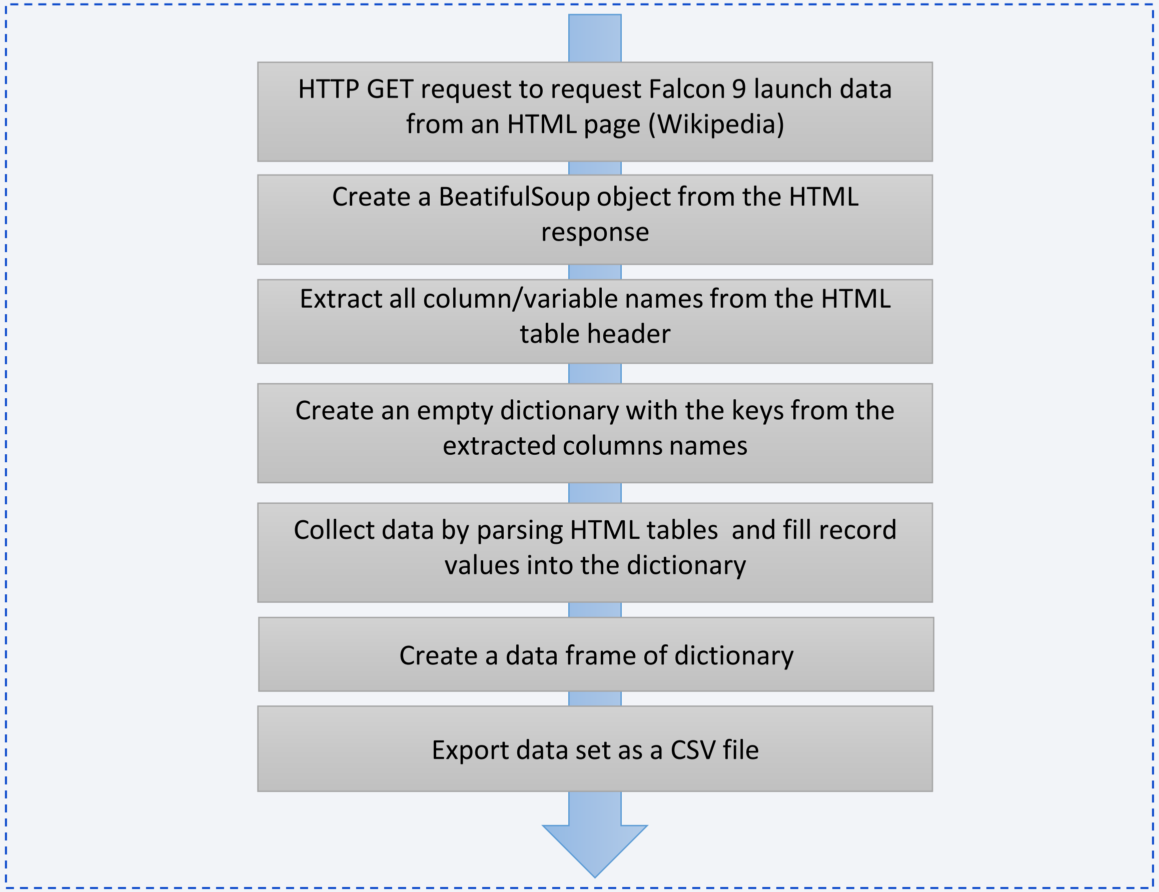
Filter the dataframe to keep only Falcon 9 launches

Replace missing Payload Mass values with the ‘mean’ of this attribue

Export data into a CSV file

Data Collection - Scraping

- On the right, the flow chart is presented how to get the Falcon 9 launch data from the following Wikipedia page by web scraping: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- GitHub link to completed web scraping notebook: ->Link



HTTP GET request to request Falcon 9 launch data from an HTML page (Wikipedia)

Create a BeatifulSoup object from the HTML response

Extract all column/variable names from the HTML table header

Create an empty dictionary with the keys from the extracted columns names

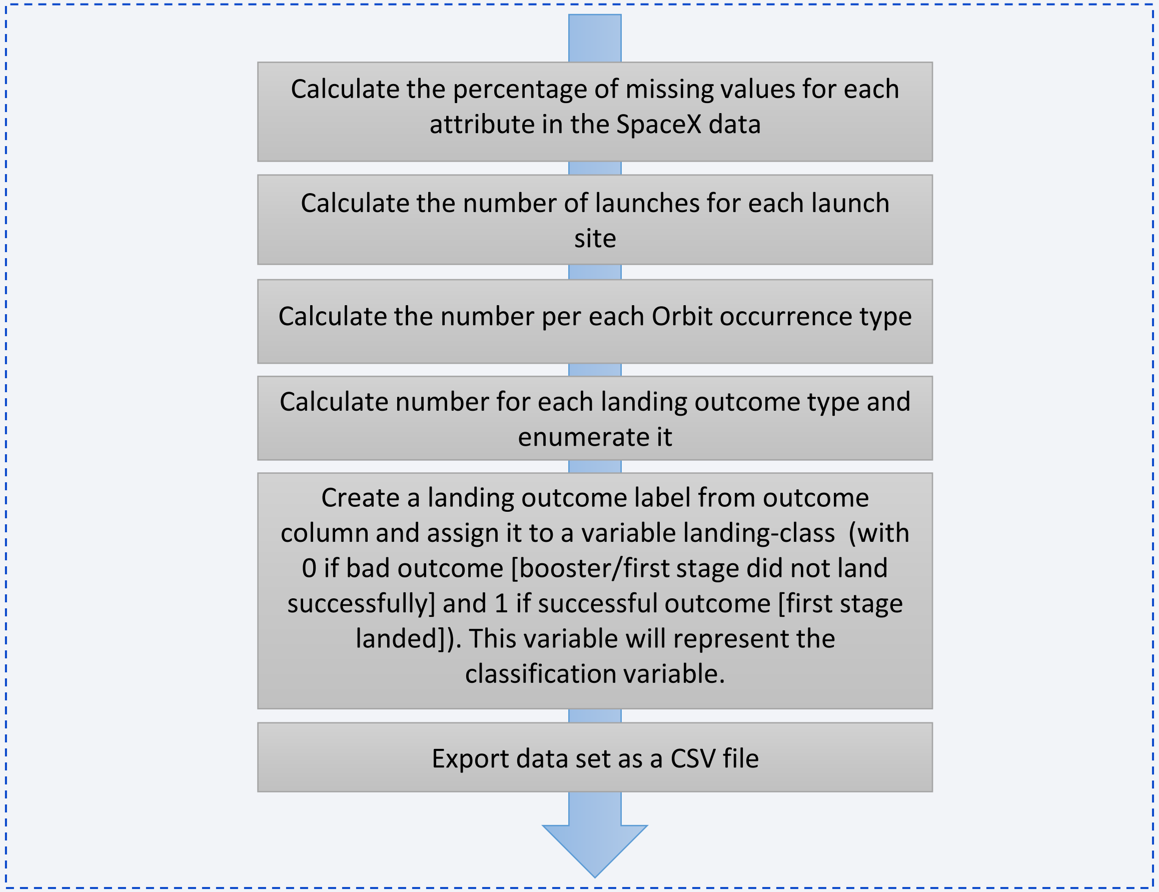
Collect data by parsing HTML tables and fill record values into the dictionary

Create a data frame of dictionary

Export data set as a CSV file

Data Wrangling

- Some Exploratory Data Analysis is performed on the data as outlined on the right. Additionally, the launch outcome variable is converted into a training label with 0 for bad outcomes and 1 successful outcomes
- GitHub link to completed data wrangling related notebook: ->Link



Calculate the percentage of missing values for each attribute in the SpaceX data

Calculate the number of launches for each launch site

Calculate the number per each Orbit occurrence type

Calculate number for each landing outcome type and enumerate it

Create a landing outcome label from outcome column and assign it to a variable landing-class (with 0 if bad outcome [booster/first stage did not land successfully] and 1 if successful outcome [first stage landed]). This variable will represent the classification variable.

Export data set as a CSV file

EDA with Data Visualization

- Plotted charts:
  - A scatter plot of Flight Number vs. Launch Site to see whether there is a connection to the outcome (success / fail)
  - A scatter plot of Payload mass (kg) vs. Launch sites to see whether there is a connection to the outcome (success / fail)
  - A bar chart for the success rate of each orbit type and a scatter plot vs. the outcome types to see the Orbits with highest success rates
  - A scatter plot of Flight Number vs. Orbit type to see whether there is connection to the outcome (success / fail)
  - A scatter plot of Payload mass (kg) vs. Orbit type to see whether there is an influence on outcome (success / fail)
  - A line chart of the yearly average success rate to verify whether there is a trend
- GitHub URL of the completed EDA with data visualization notebook: ->Link

EDA with SQL

- Performed SQL queries that are commented later in the presentation
  - Query displaying the names of unique launch sites in space mission
  - Query displaying 5 records where launch site begins with ‘CCA’
  - Query that displays the total payload mass carried by boosters launched by NASA (CRS)
  - Query that shows the average payload mass carried by booster version F9 v1.1
  - Query that yields the date when the first successful landing outcome in ‘ground pad’ was achieved
  - Query providing the names of the booster which have success in drone sop and have a payload mass greater than 4000 but less than 6000
  - Query providing the total number of successful and failed mission outcomes
  - Query listing the names of the booster_versions which have carried the maximum payload mass
  - Query listing the month names, failure landing_outcomes in drone ship, booster versions, and launch_site for the months in year 2015
  - Query providing the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, ranked in descending order
- GitHub link for the completed EDA with SQL notebook: ->Link

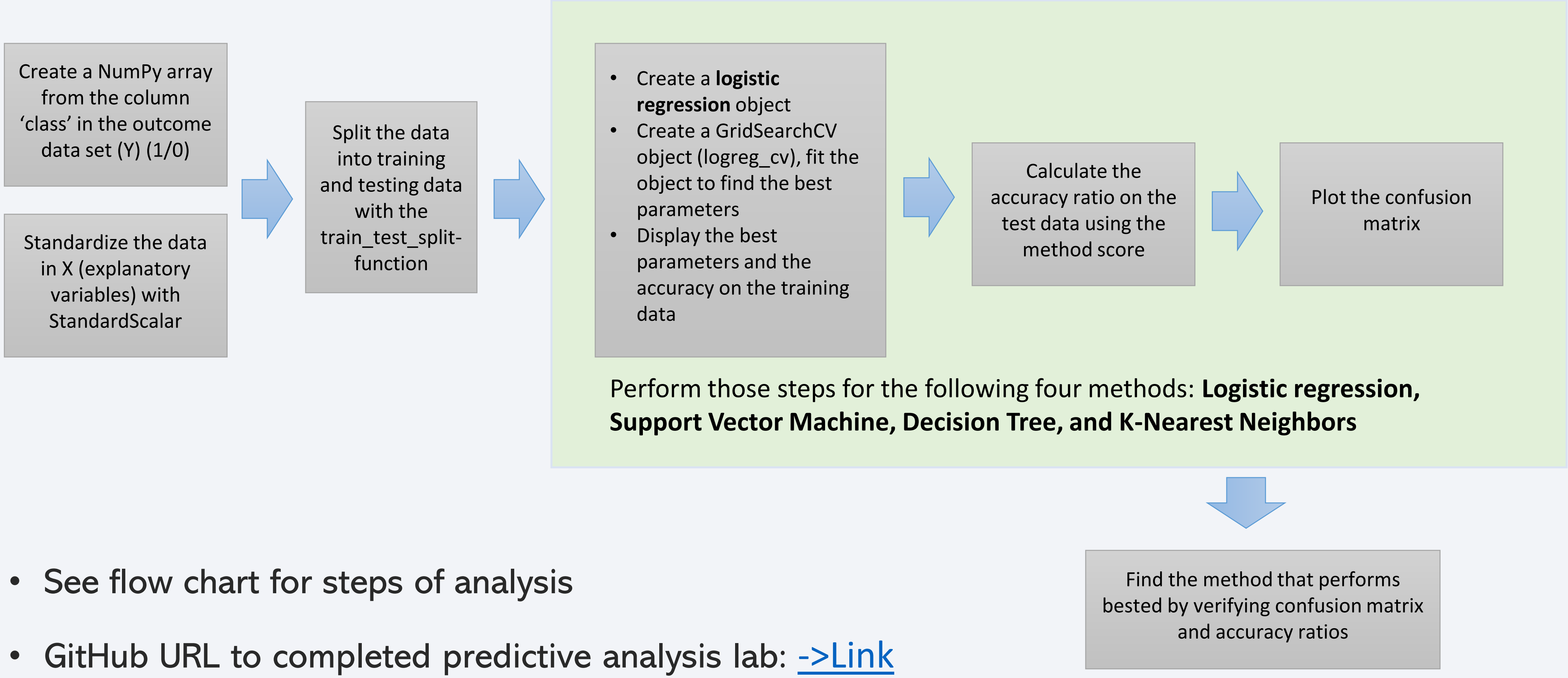
Build an Interactive Map with Folium

- A Circle at the coordinates of NASA Johnson Space Center with an icon showing its name was added to have an initial center location for the folium map
- A Circle marker and name at the coordinates of each launch site was added to the map to see where the launch sites are located
- Added colored markers at each launch site that show a green marker for a success and a red marker for a failure in a ‘marker cluster’ in order to see which sites have high success rates
- Added lines (including the distance in km) between the launch site CCAFS SLC-40 to its closest railway, coast, highway and city in order to see the proximities of the launch site to those places.
- GitHub URL of the completed interactive map with Folium map: ->Link

Build a Dashboard with Plotly Dash

- Launch Sites dropdown list in order the all sites together can be selected or each site individually for the investigation
- Pie chart in order to show the successful launches in percentage based on the launch site selection
- Slider with which one can choose the desired range of Payload mass (kg) in order to investigate effect of Payload mass on successes
- Scatter plot in order to show the Payload mass on one axis versus the success rate (success vs. fail) on the other axis for the different Booster Versions
- GitHub URL for the completed Plotly Dash lab: ->Link

Predictive Analysis (Classification)



Create a NumPy array from the column 'class' in the outcome data set (Y) (1/0)

Standardize the data in X (explanatory variables) with StandardScalar

Split the data into training and testing data with the train_test_split-function

- Create a **logistic regression** object
- Create a GridSearchCV object (logreg_cv), fit the object to find the best parameters
- Display the best parameters and the accuracy on the training data

Calculate the accuracy ratio on the test data using the method score

Plot the confusion matrix

Perform those steps for the following four methods: **Logistic regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors**

Find the method that performs bested by verifying confusion matrix and accuracy ratios

- See flow chart for steps of analysis
- GitHub URL to completed predictive analysis lab: ->Link

Results

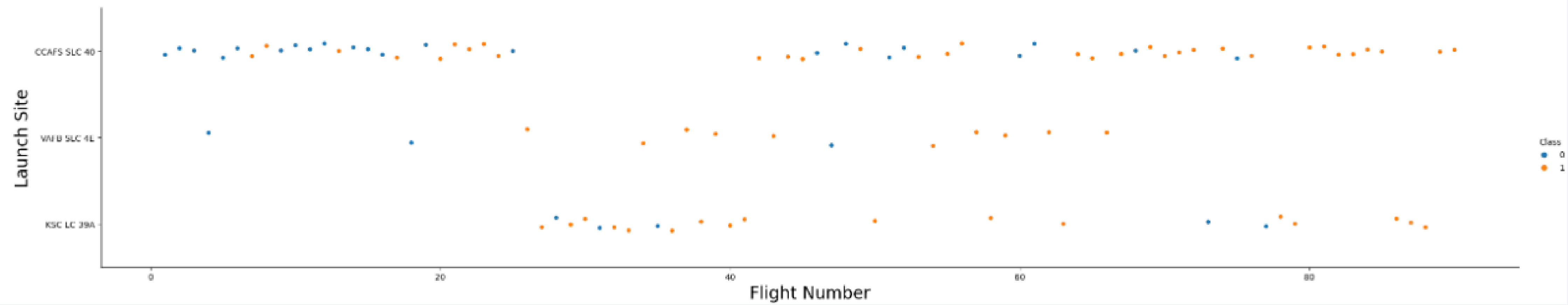
- Exploratory data analysis results (see section 2, slides 17-33)
- Interactive analytics demo in screenshots (see section 3/4, slides 34-42)
- Predictive analysis results (see section 5, slides 43-45)

Section 2

Insights drawn from EDA

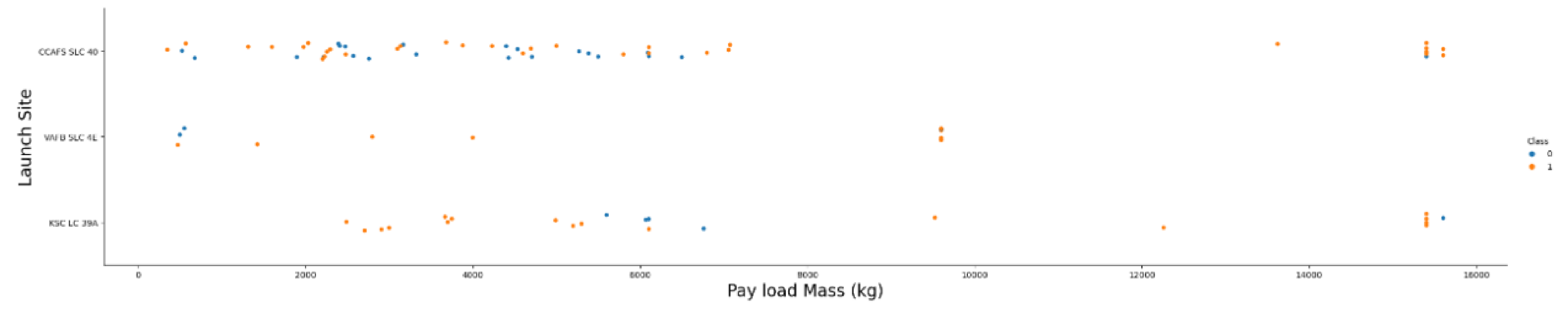
Flight Number vs. Launch Site

- A scatter plot of Flight Number vs. Launch Site is displayed below
- Explanations
  - Earlier launches mostly failed (0), while the last launches all succeeded (1)
  - CCAFS SLC 40 is the site with most launches followed by KSC LC 39A then VAFB SLC 4E
  - KSC LC 39A and VAFB SLC 4E have higher success rate



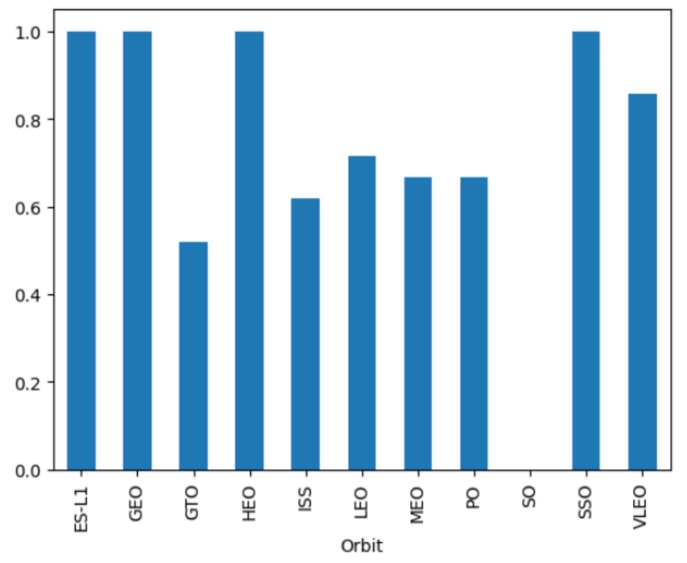
Payload vs. Launch Site

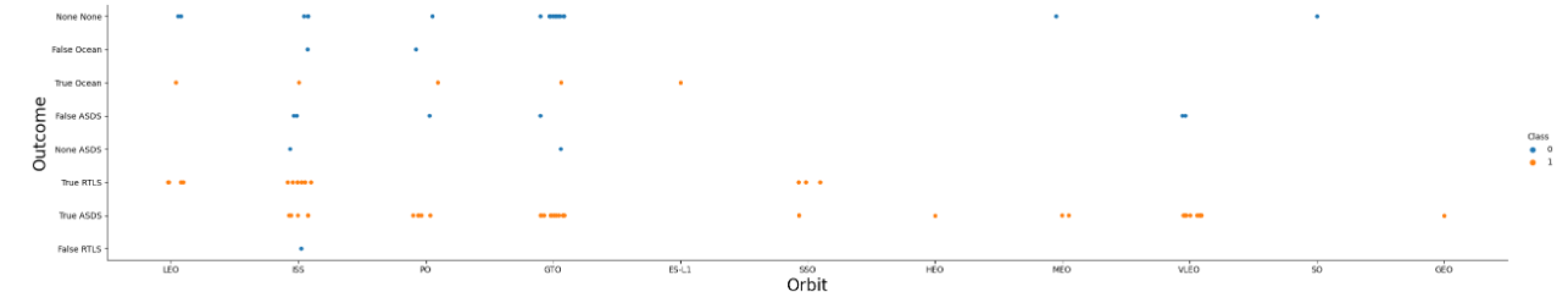
- Below scatter plot shows the Payload vs. Launch Site
- Explanations
    - For every launch site, the higher the Payload Mass the higher the success rate
    - Most of the launches with Payload Mass higher than 8000kg were successful
    - Launch site KSC LC-39A also shows successes with low Payload Mass



Success Rate vs. Orbit Type

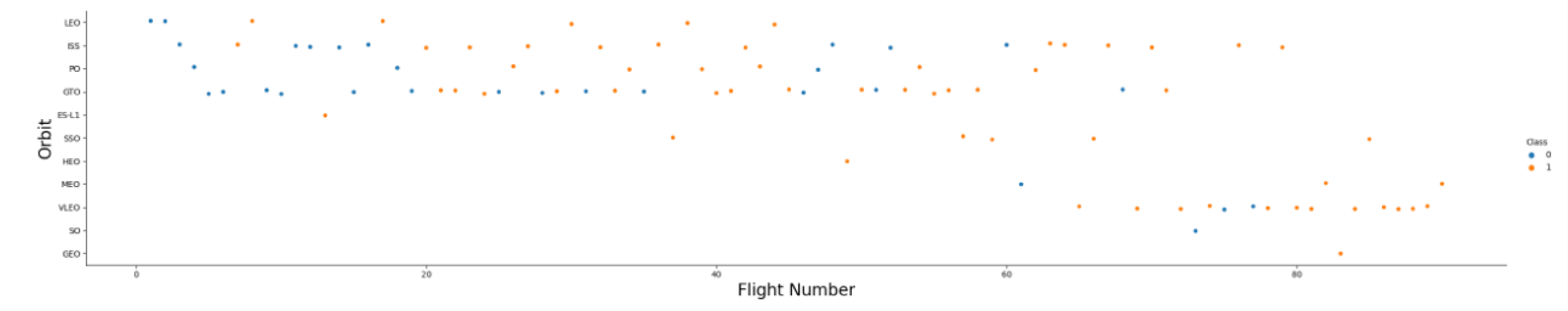
- A bar chart for the success rate of each orbit type is shown on the right and a scatter plot of the successes/fails at the bottom
- Explanations
  - Four Orbits have 100% success rates: ES-L1, GEO, HEO, SSO, VLEO
  - One Orbit has 0% success rate: SO
  - The other Orbits have roughly success rates between 50% and 90%





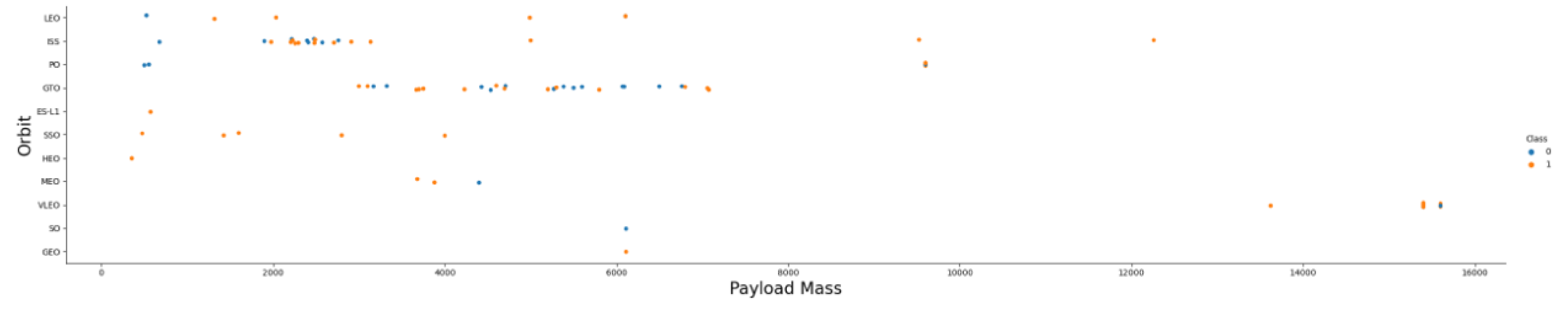
Flight Number vs. Orbit Type

- A scatter plot of Flight number vs. Orbit type is shown below
- Explanations
    - For Orbit LEO there seems to be a connection with flight number and successes, where in the beginning there were fails and then successes only
    - Other Orbits do not show a clear connection with flight number



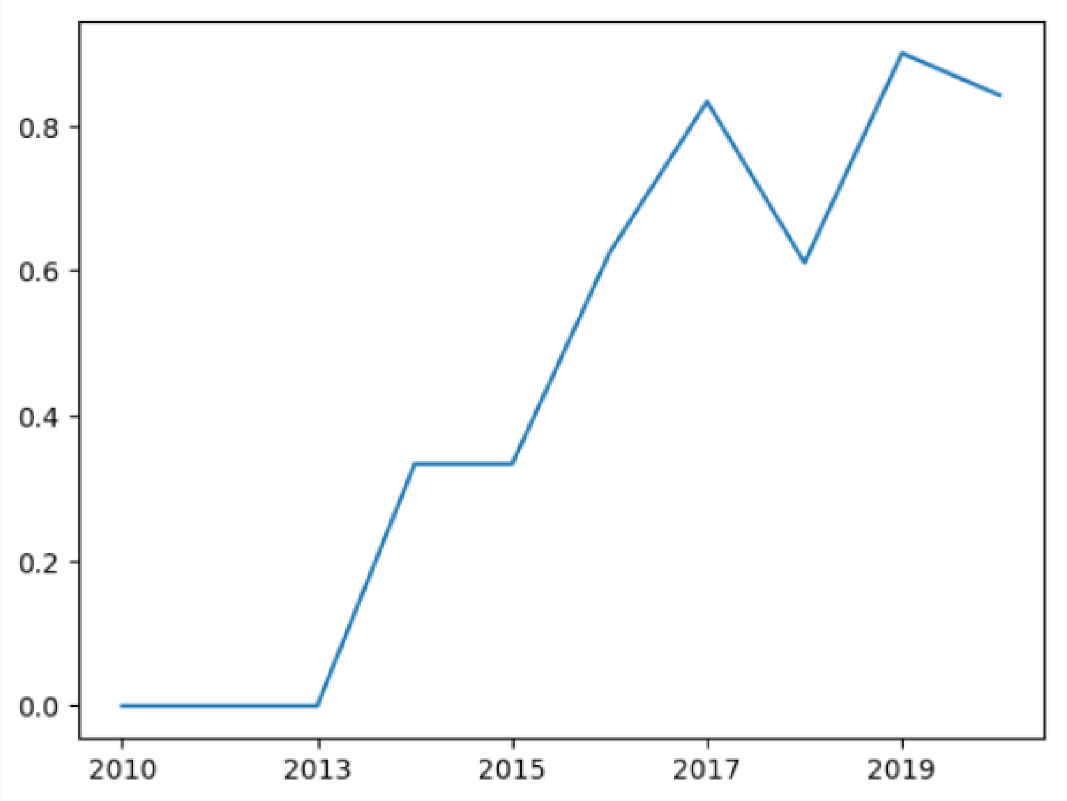
Payload vs. Orbit Type

- A scatter plot of payload vs. orbit type is shown below
- Explanations
  - Heavier Payload Mass show a positive influence (more success) for Orbits LEO and ISS
  - Orbit SOO with 100% success rate is all based on lower payload masses



Launch Success Yearly Trend

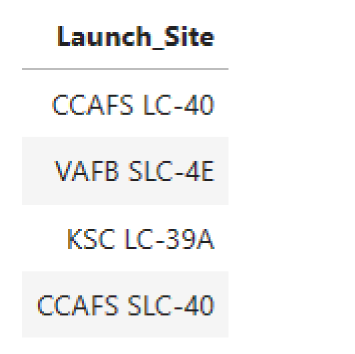
- A line chart of the yearly average success rate is shown on the right
- Explanations
  - Success rates since 2013 kept increasing till 2020, with just one dip in 2018
  - In the early years (2010-2013) there were only failed launches



All Launch Site Names

- The table on the right shows the names names of the unique launch sites
- Explanation:

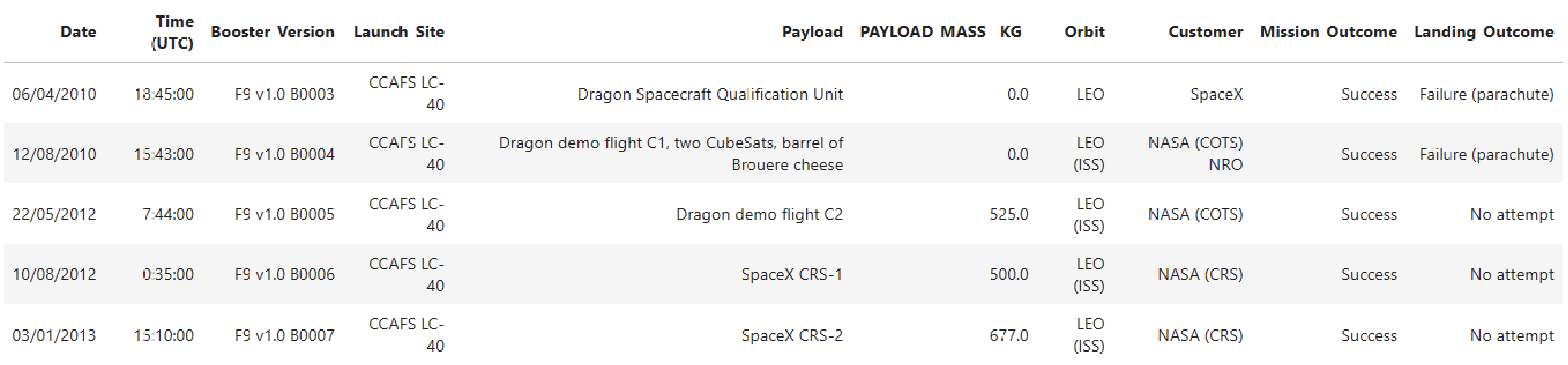
CCAFS LC-40	Cape Canaveral Launch Complex 40
VAFB SCL-4E	Vandenberg Space Launch Complex 4E
KSC LC-39A	Kennedy Space Center Launch Complex 39A
CCAFS SLC-40	Cape Canaveral Space Launch Complex 40



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

Launch Site Names Begin with 'CCA'

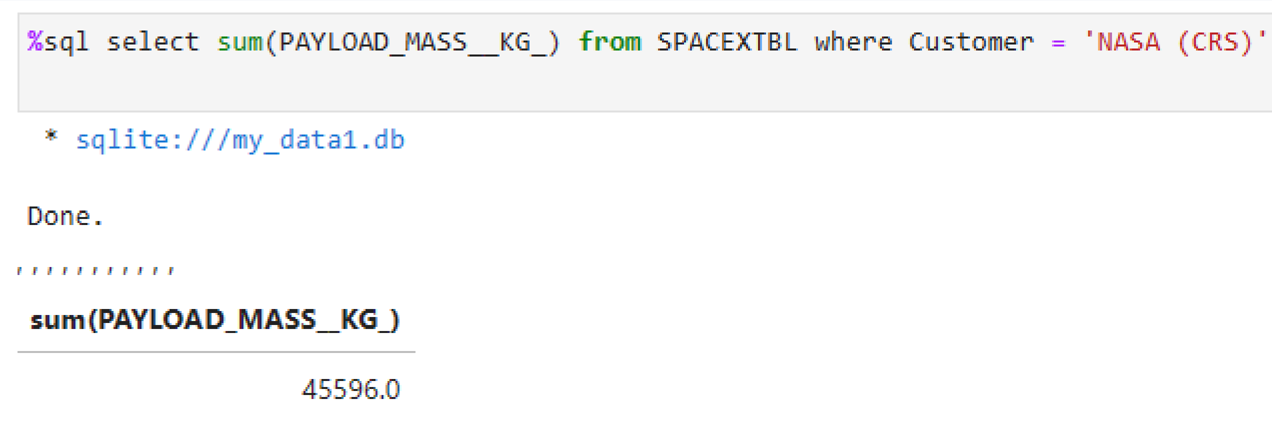
- The table below display 5 records where launch sites name begin with `CCA`
- CCA stands for the location Cape Canaveral



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

Total Payload Mass

- The number below shows the total payload mass carried by boosters launched by NASA



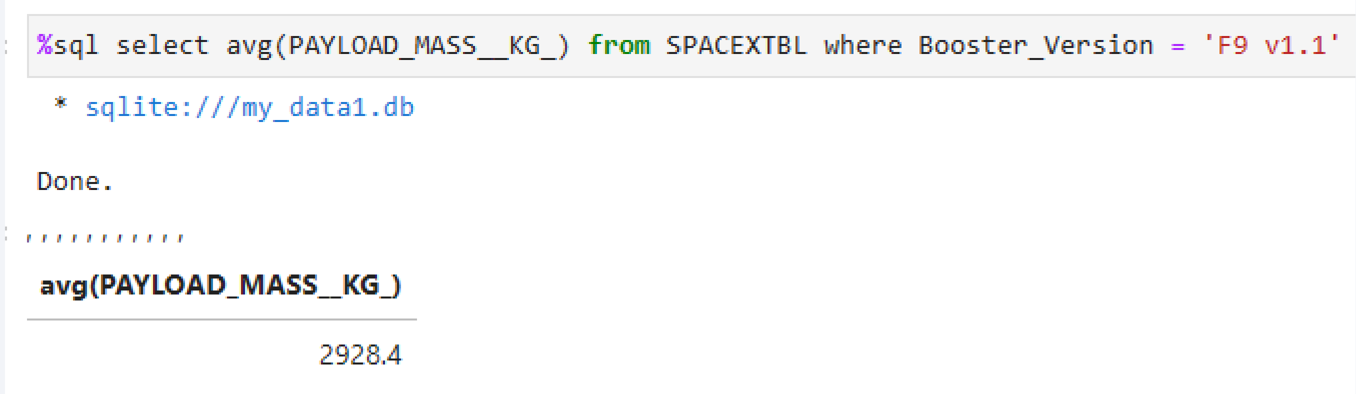
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

- The number below shows the average payload mass carried by booster version F9 v1.1



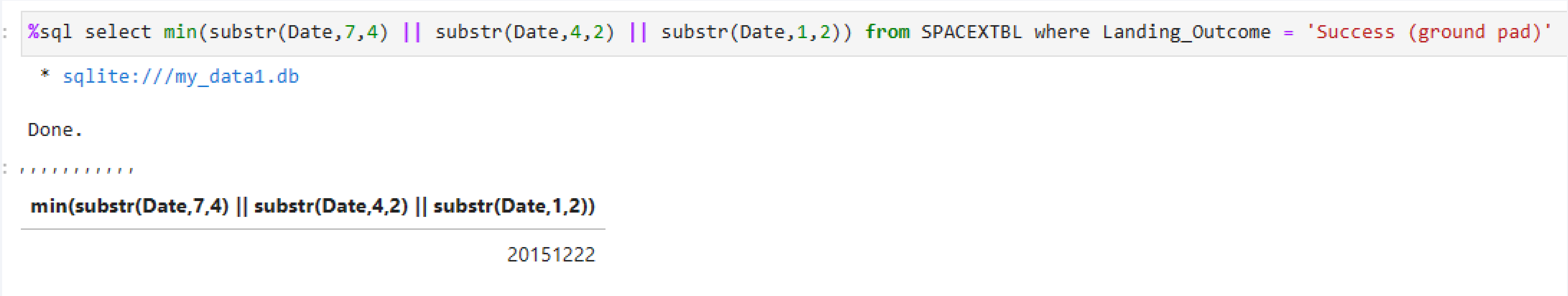
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The date displayed below shows the first successful landing outcome on ground pad



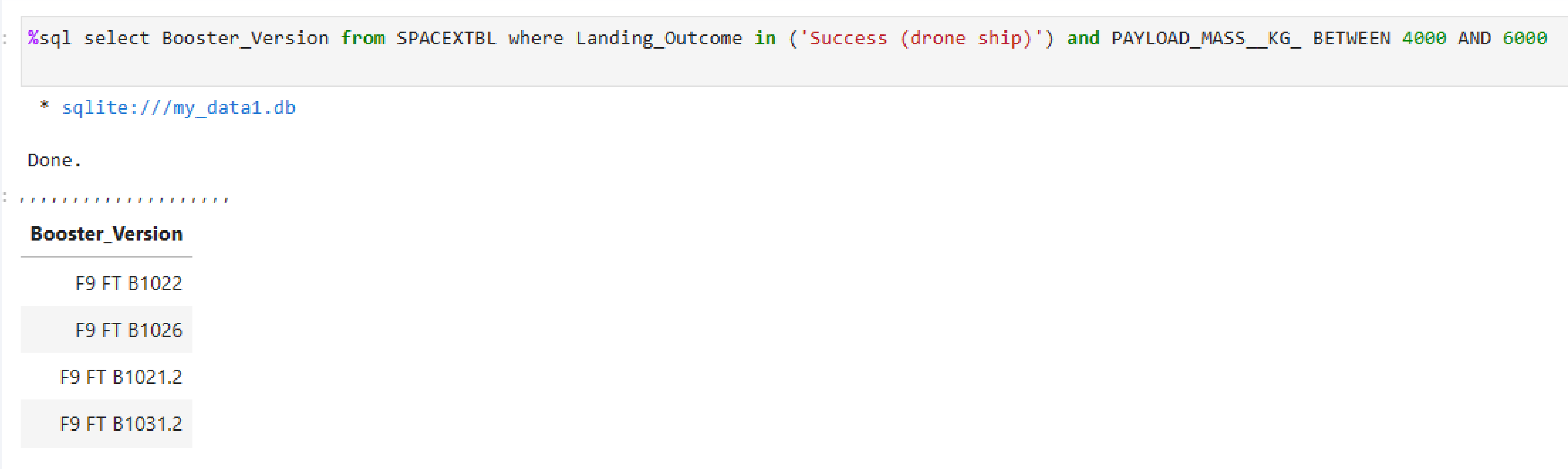
```
%sql select min(substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2)) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

min(substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2))
20151222

Successful Drone Ship Landing with Payload between 4000 and 6000

- The table below shows the names of the boosters (Booster_Version) which have successfully landed on drone ship and had a payload mass greater than 4000 but less than 6000



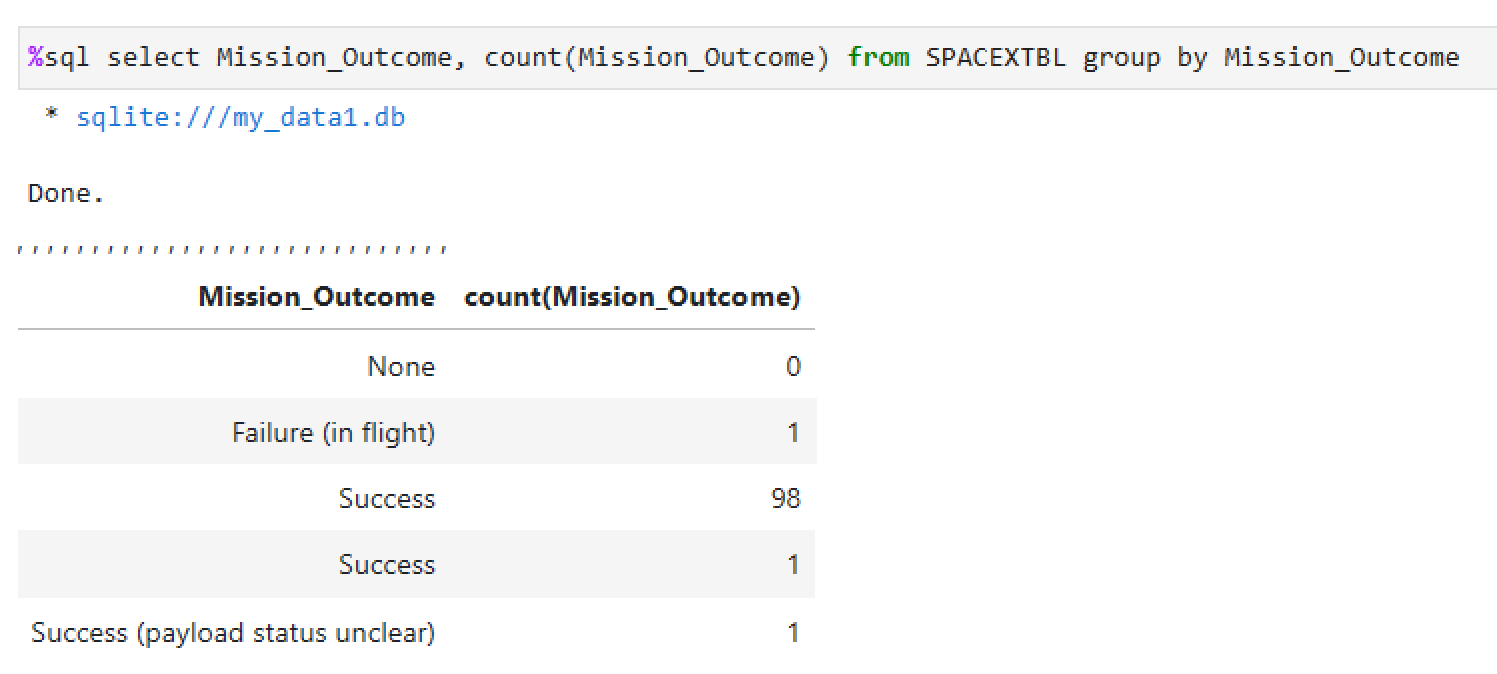
```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome in ('Success (drone ship)') and PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

- The table below shows the total number of successful and failure mission outcomes



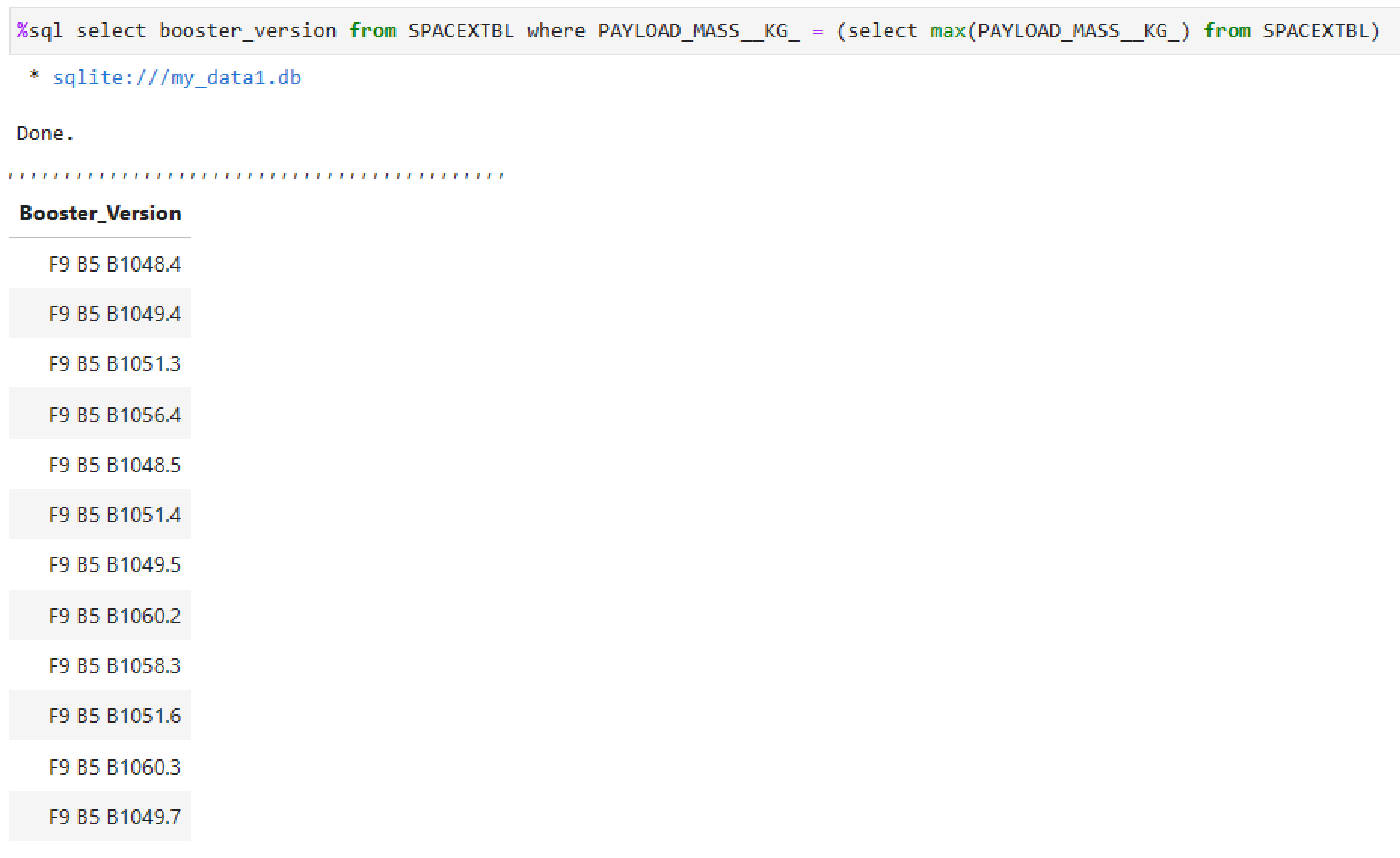
```
%sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTBL group by Mission_Outcome
```

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

Mission_Outcome	count(Mission_Outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The table on the right lists the names of the booster versions which have carried the maximum payload mass



```
%sql select booster_version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

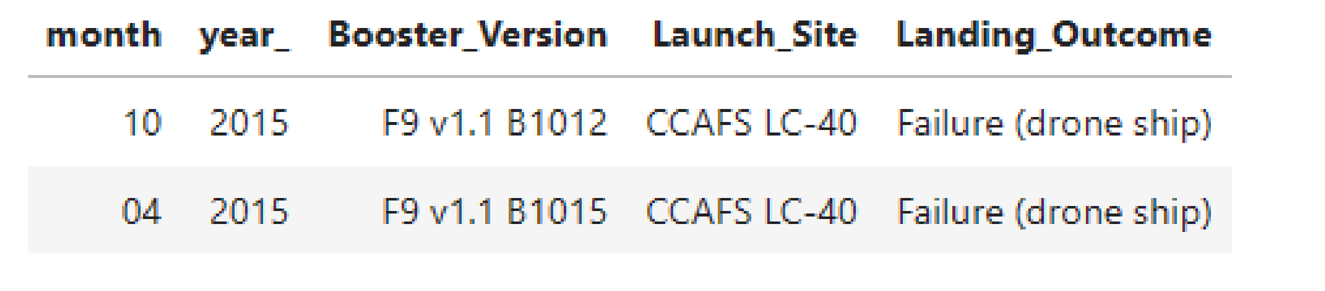
```
 * sqlite:///my_data1.db

Done.
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

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

2015 Launch Records

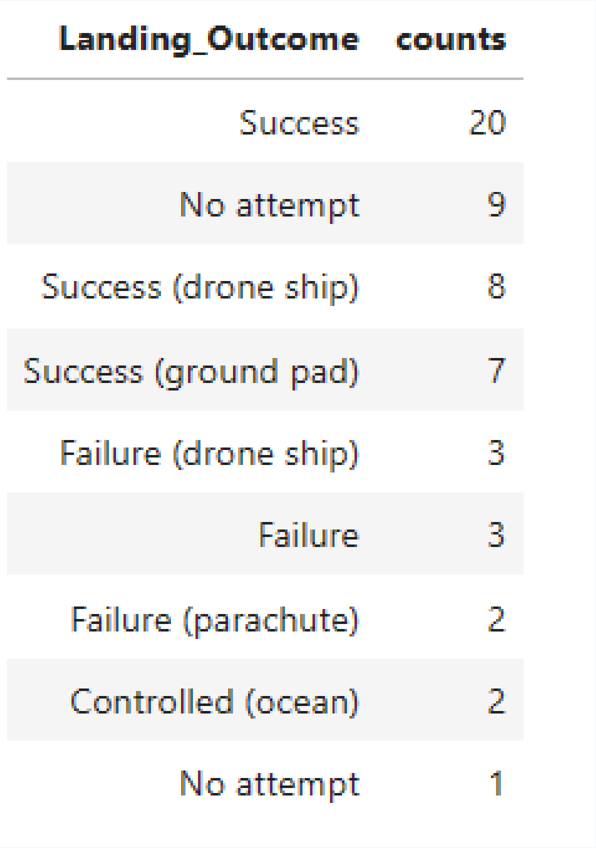
- The table below lists the failed landing outcomes in drone ship, their booster versions, and the launch site names for the months in year 2015



month	year_	Booster_Version	Launch_Site	Landing_Outcome
10	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- The table on the right shows the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- Success was the Landing_Outcome that happened most in this period.



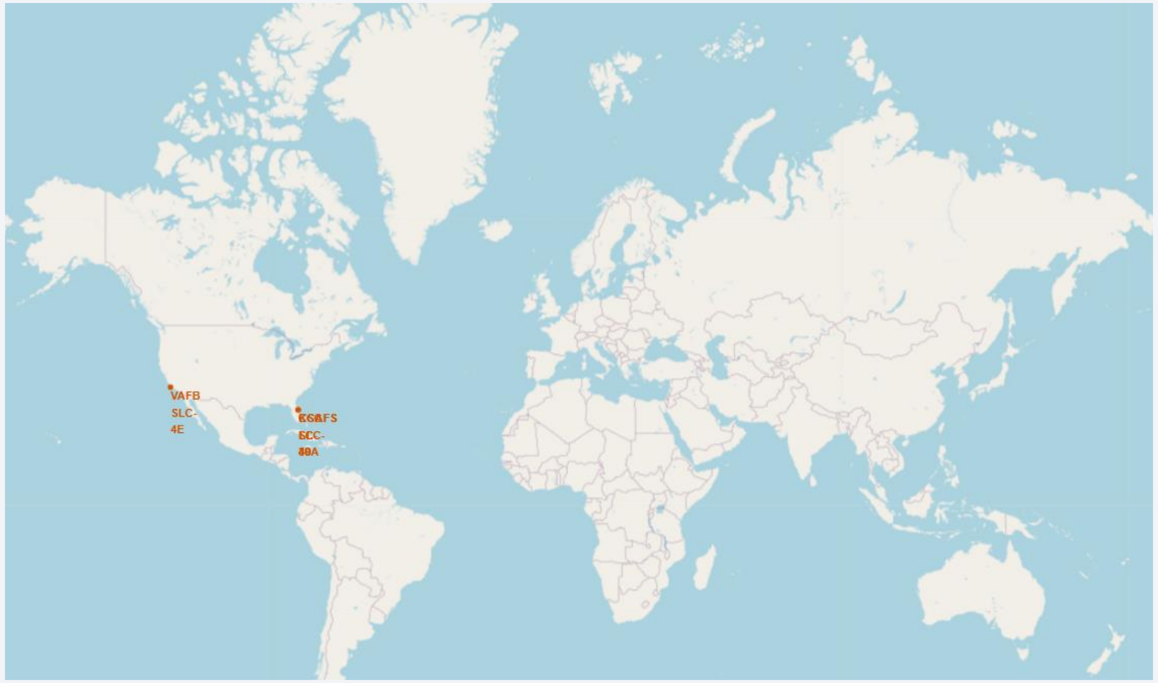
Landing_Outcome	counts
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

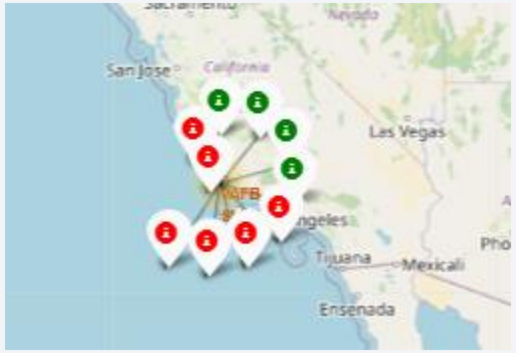
Launch sites displayed on global map

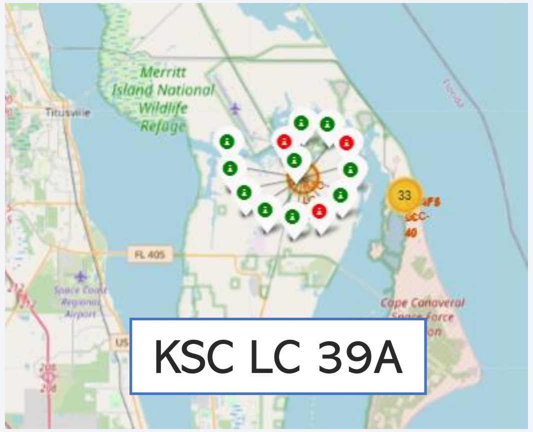
- This slide contains a screen shot of the folium map and displays all four launch sites' location with orange markers on this global map
- Explanations
  - Note that 3 of the 4 launch sites are very close together in Florida on the east coast; the fourth is in California, also close to the coast.
  - Launching rockets from the east coast to the east gives an additional booth for the rocket due to the earth's rotation. Additionally, the risk of damage in case of failures is minimized due to the proximity to the ocean.
  - Finally, all launch sites are rather close to the equator line which is additionally favorable for launching rockets.



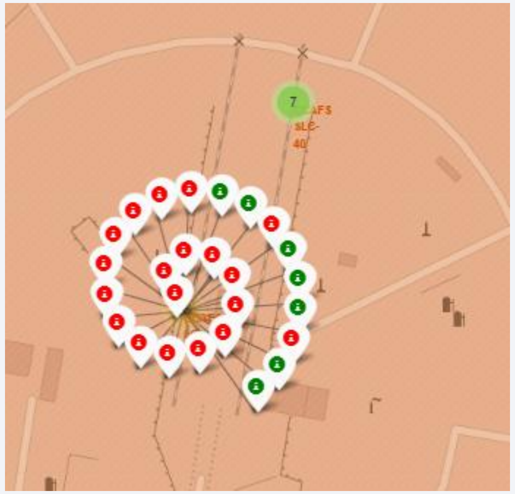
Launch outcomes with colored labels for each launch site

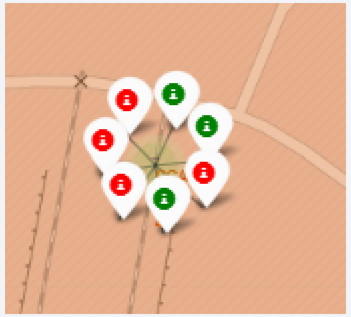
- The four pictures below from the folium map show the launch outcomes of each launch site with colored labels
  - Green marker means that it was a successful launch
  - Red marker means that it was a failed launch
- Launch site “KSC LC 39A” is the site with most successful launches




Merritt Island National Wildlife Refuge
Titusville
33
FL 405
Cape Canaveral Space Force

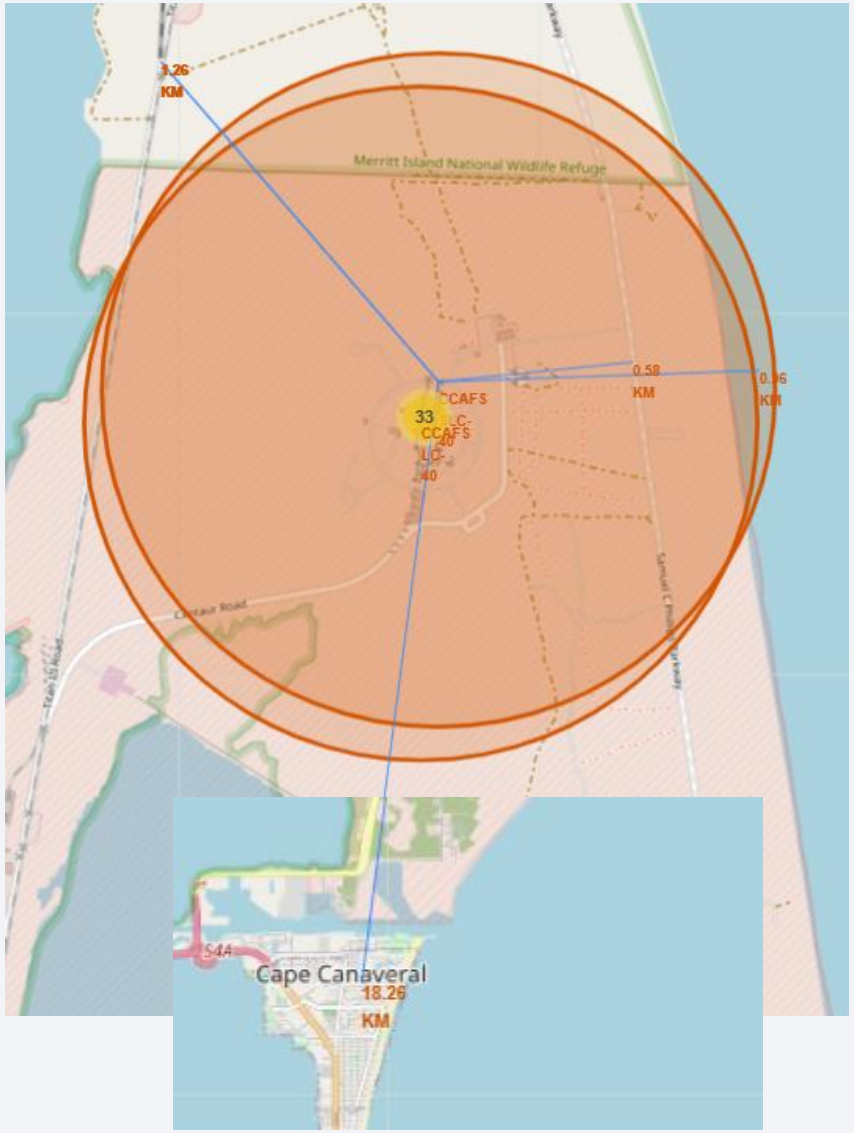
KSC LC 39A

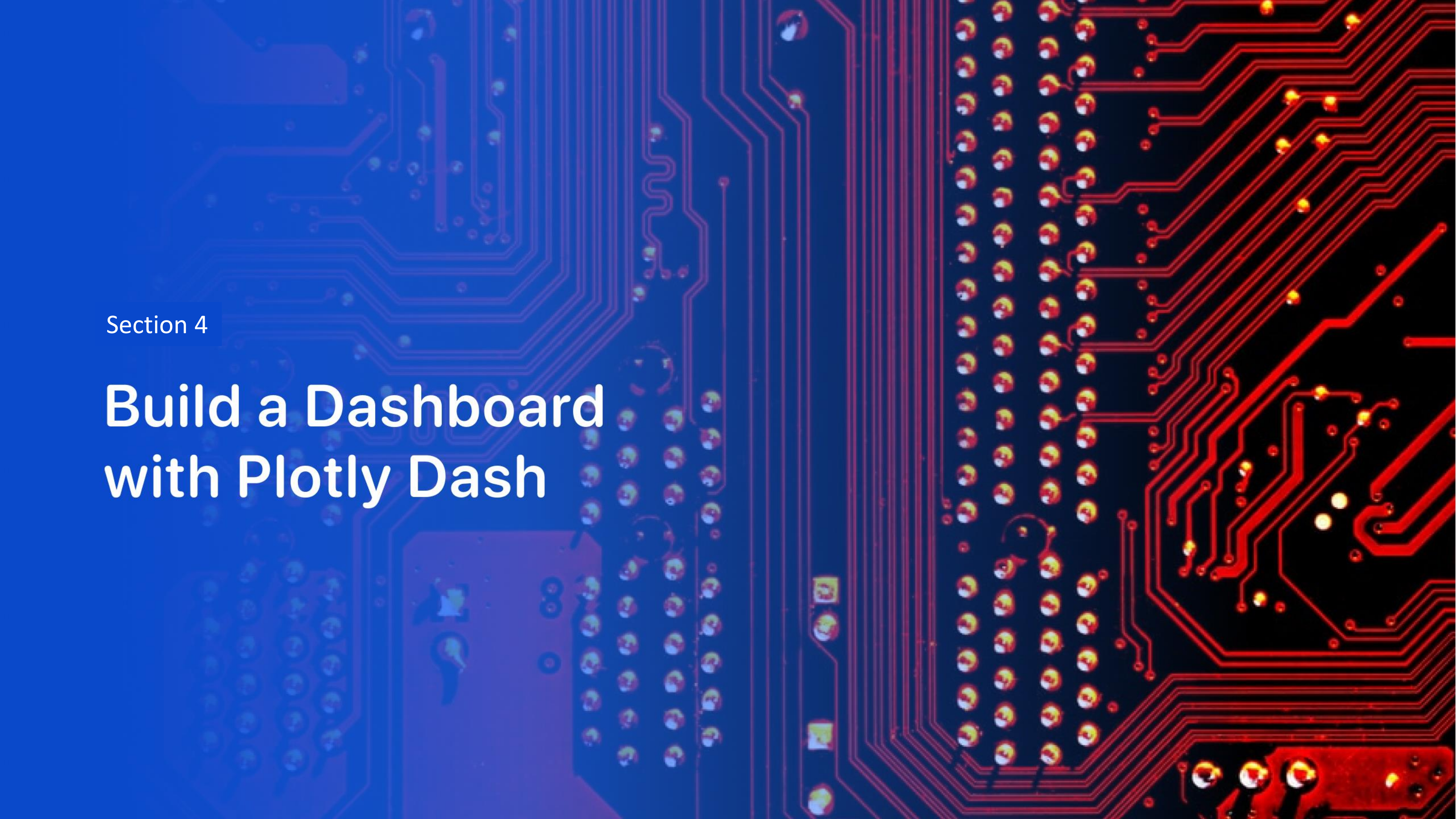




Proximities of launch site CCAFS SLC-40

- The picture on the right show the proximities of the launch site CCAFS SLC-40 to the railway, highway, coastline and a close city.
- The distances from the launch site are as follows:
  - To Railway: 1.26km
  - To Highway: 0.58km
  - To Coastline: 0.96km
  - To City: 18.26km
- Above confirms again that the launch site is very close to the coastline to minimize risks in case of failures. On the other hand, however, it is also very close to a railway, highway and a city (Cape Canaveral). Visual inspection confirms the same for the other launch sites.



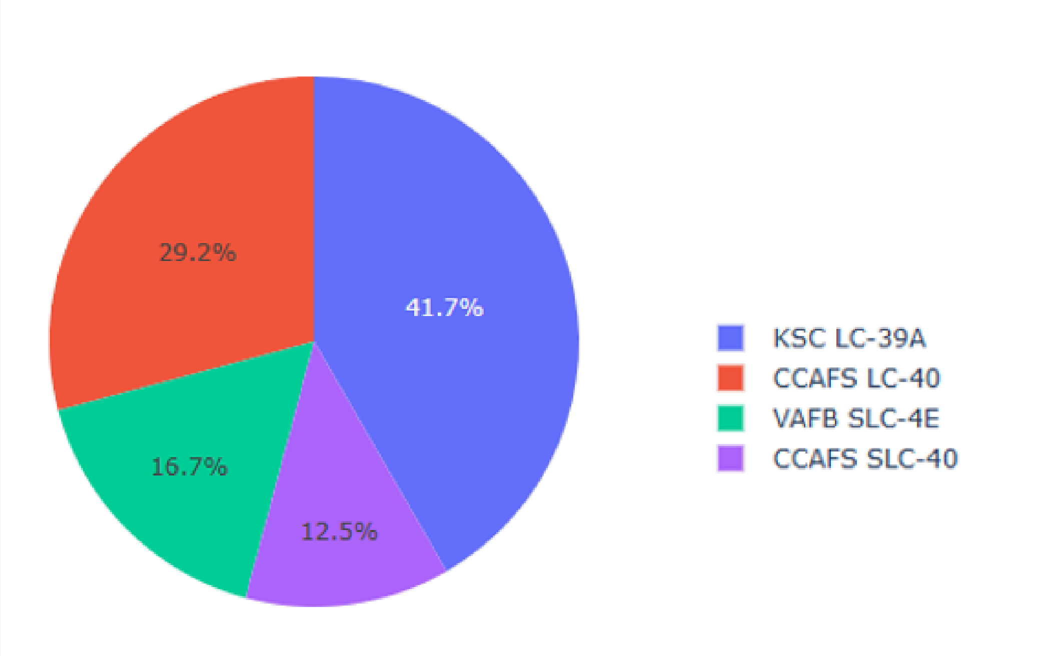


Section 4

Build a Dashboard with Plotly Dash

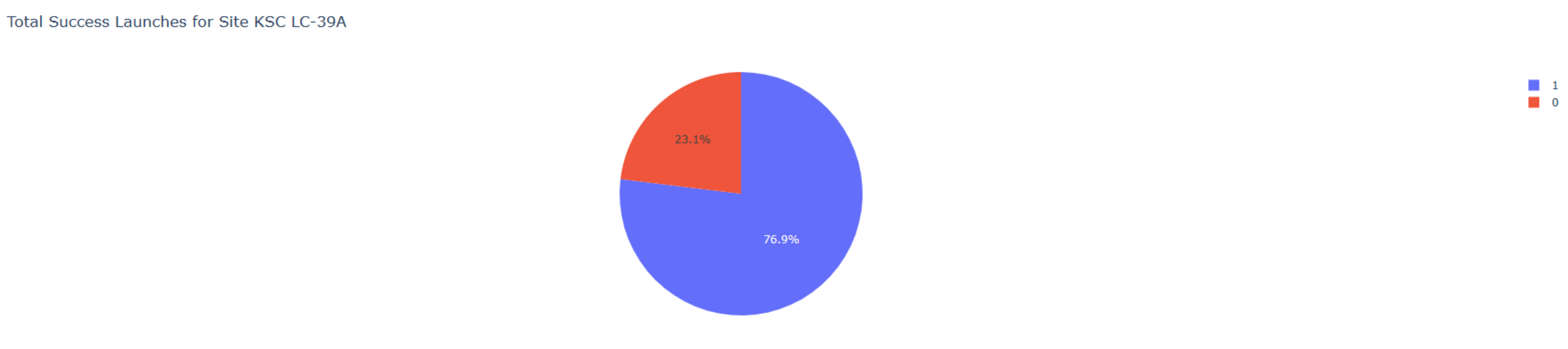
Launch successes by launch site

- The pie chart shows the launch success per each launch site
- The chart shows that the site "KSC LC-39A" has clearly the most successful launches followed by CCAFS LC-40; both are located on the east coast.



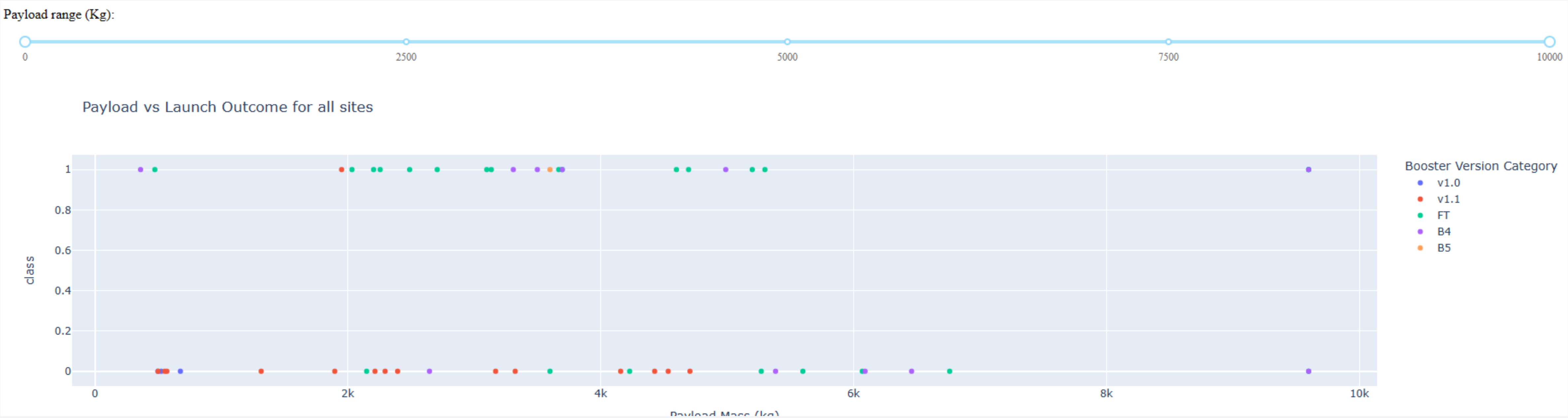
Launch site with highest launch success ratio

- The pie chart below shows the launch site “KSC LC-39A” that has the highest launch success ratio
- The site has a success rate of 76.9% with only 23.1% failed landings



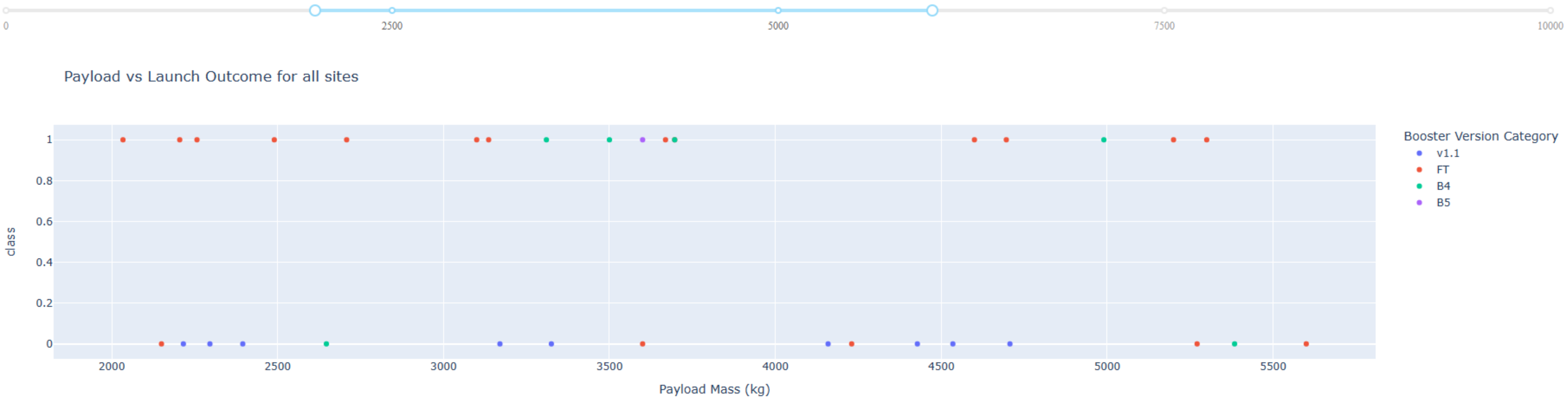
Booster version with highest success rate for all sites

- Booster version FT has the largest success rate



Payload range with highest success rate for all sites

- By visual inspection, a Payload range of circa 2000kg up to 6000kg seem to have the highest success *rate*.

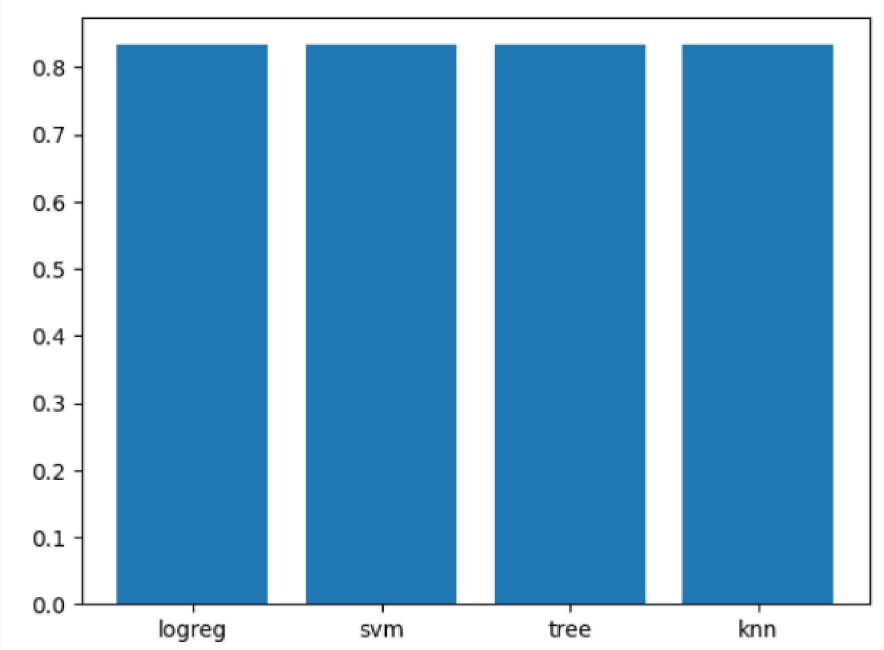


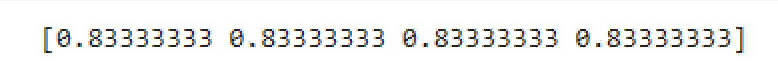
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The bar chart on the right shows the accuracy ratio for the four models
  - Logistic Regression (logreg)
  - Support Vector Machine (svm)
  - Decision Tree (tree)
  - K-Nearest Neighbors (knn)
- The accuracy ratio values are also displayed below the chart
- One can see the all four models have the same classification accuracy of 0.83333

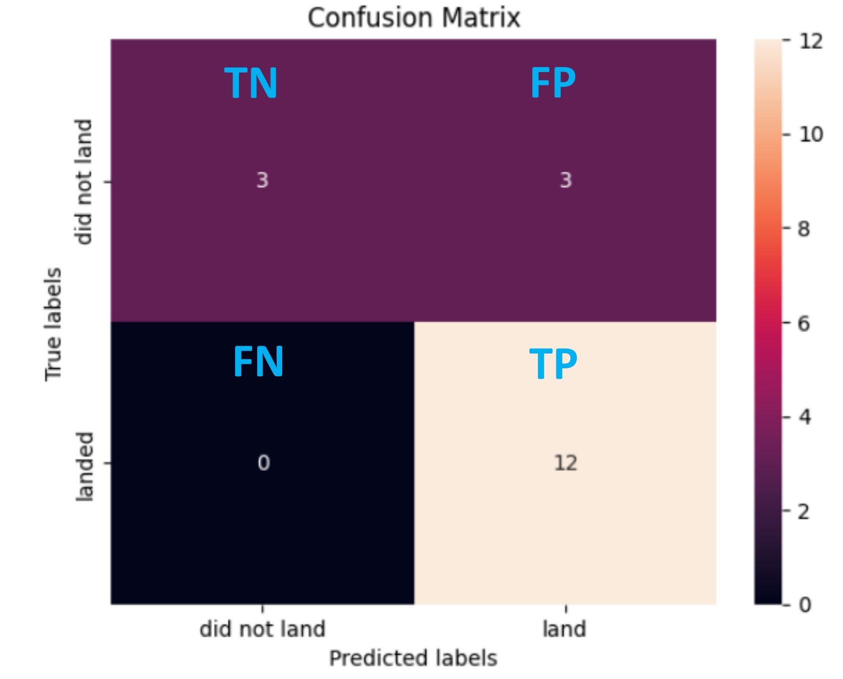




```
[0.83333333 0.83333333 0.83333333 0.83333333]
```

Confusion Matrix

- The confusion matrix of the logistic regression is displayed on the right
- In fact, the confusion matrix of all the other three models look the same and they are thus all best performing models judging based on the confusion matrix
- Explanation of the matrix
  - The models can distinguish between the different classes having many true negatives (TN) and true positives (TP)
  - The major problem is the false positives (FP) where the model predicted “land”, however the outcome was “did not land”



Conclusions

- Questions that we wanted to answer
  - How do different explanatory variables affect the success of the first stage landing? Do the success rates increase over the years? Which is the best binary classification model to predict the success of the first stage landing?
- Main Conclusions
  - Launch site “KSC LC-39A” has the highest success rate. All launch sites are rather close to the equator which is favorable for launching rockets. Additionally, all launch sites are very close to the coastline to minimize risks in case of failures. On the other hand, however, they are also very close to railways, highways and cities.
  - Four Orbits have 100% success rates: ES-L1, GEO, HEO, SSO, VLEO. One Orbit has 0% success rate: SO. Booster Version F has the highest success rate of all sites. A Payload range of circa 2000kg up to 6000kg seems to have the highest success rate.
  - In the early years (2010-2013) there were only failed launches. Success rates since 2013 kept increasing till 2020, with just one dip in 2018.
  - All four tested models (Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors) display the same Confusion Matrix and have the same accuracy ratio of 0.8333, hence they perform equally well based on those criteria.

Appendix

- Sources / References
  - Coursera, IBM Data Science Course
  - Link to Github containing notebooks and presentation: ->Link

Thank you!