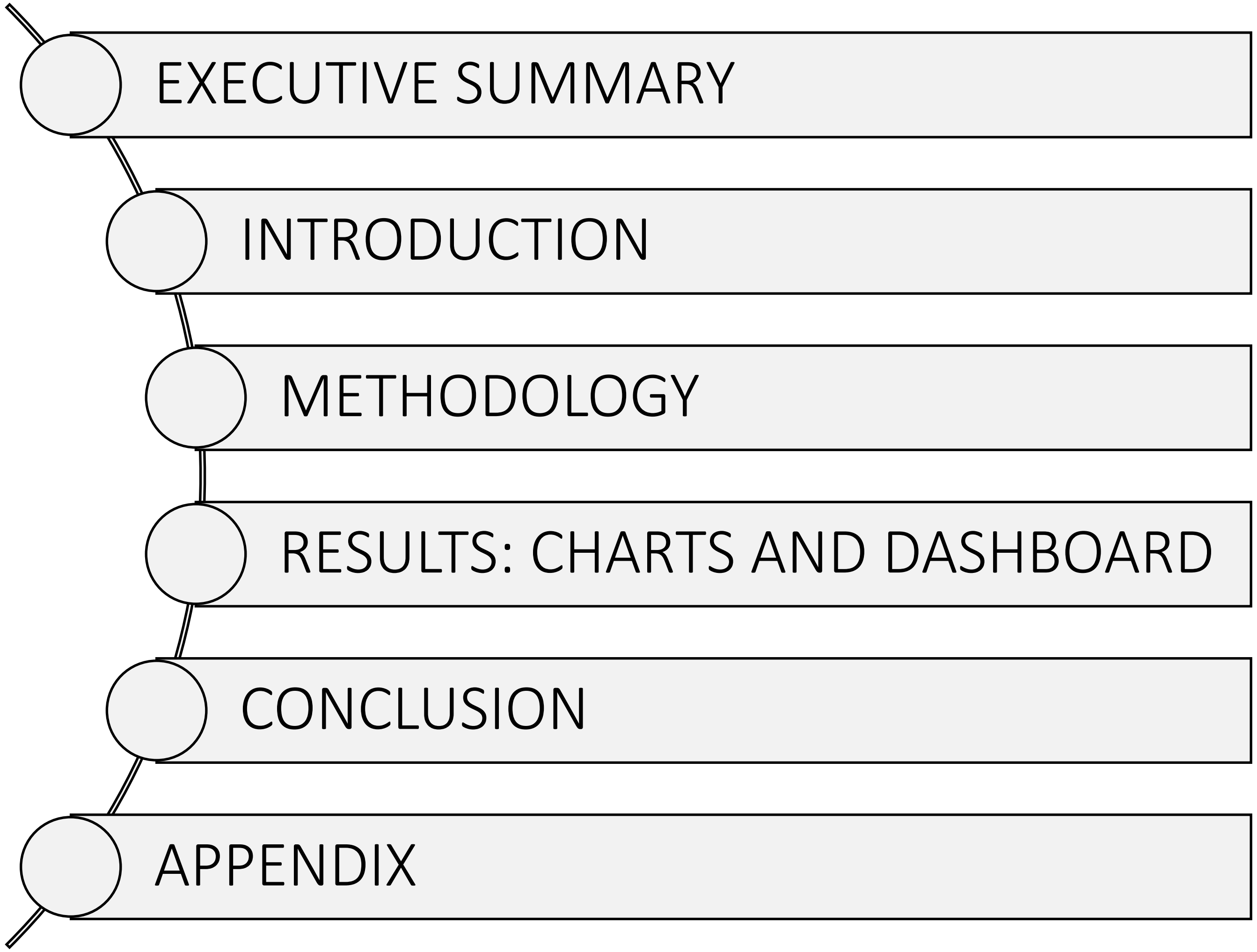
WINNING SPACE RACE WITH DATA SCIENCE

Ekaterina Takmakova
02.08.2023

OUTLINE



EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS: CHARTS AND DASHBOARD

CONCLUSION

APPENDIX

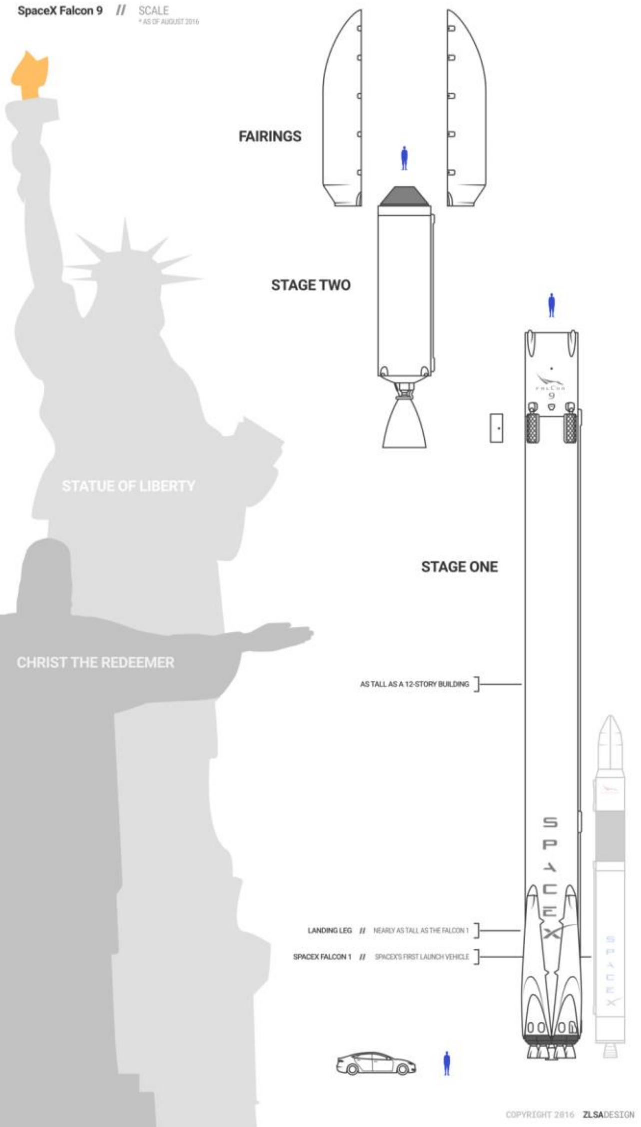
EXECUTIVE SUMMARY

In this project, records of Falcon 9 rocket launches performed by SpaceX were collected and analyzed in order to predict the likelihood of a successful first stage rocket landing.

- SpaceX API and web scraping were used for data collection, followed by data wrangling and exploratory data analysis. The most important variables were selected to train four different machine learning classification algorithms.

- Optimal hyperparameters for logistic regression, support vector machine, decision tree, and k-nearest neighbors algorithms were identified and their prediction accuracy was compared to find the model performing the best.

INTRODUCTION


SpaceX Falcon 9 // SCALE
* AS OF AUGUST 2016
FAIRINGS
STAGE TWO
STAGE ONE
STATUE OF LIBERTY
CHRIST THE REDEEMER
AS TALL AS A 12-STORY BUILDING
LANDING LEG // NEARLY AS TALL AS THE FALCON 1
SPACEX FALCON 1 // SPACEX'S FIRST LAUNCH VEHICLE
FALCON 9
SPACEX
SPACEX
COPYRIGHT 2016 ZLSADESIGN

The commercial space age is here, companies are making space travel affordable for everyone. SpaceX advertises Falcon 9 rocket launches with a cost of $62 million. Other providers cost upward of $165 million for a launch.

Falcon 9 rocket has the payload enclosed in the fairings, the second stage that helps bring the payload to orbit, and the first stage that does most of the work and is quite large and expensive.

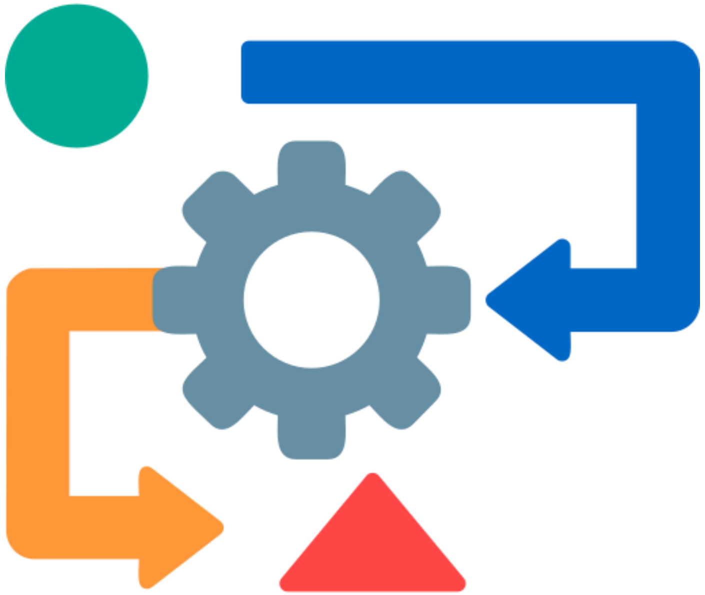
SpaceX can reuse the first stage boosters and achieve significant savings.

INTRODUCTION

If we can determine if the first stage will land, we can determine the cost of a launch.

- In this project, we want to predict if the Falcon 9 first stage will land successfully. We train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

This information can be used to predict the cost of a rocket launch in case an alternate company wants to bid against SpaceX.

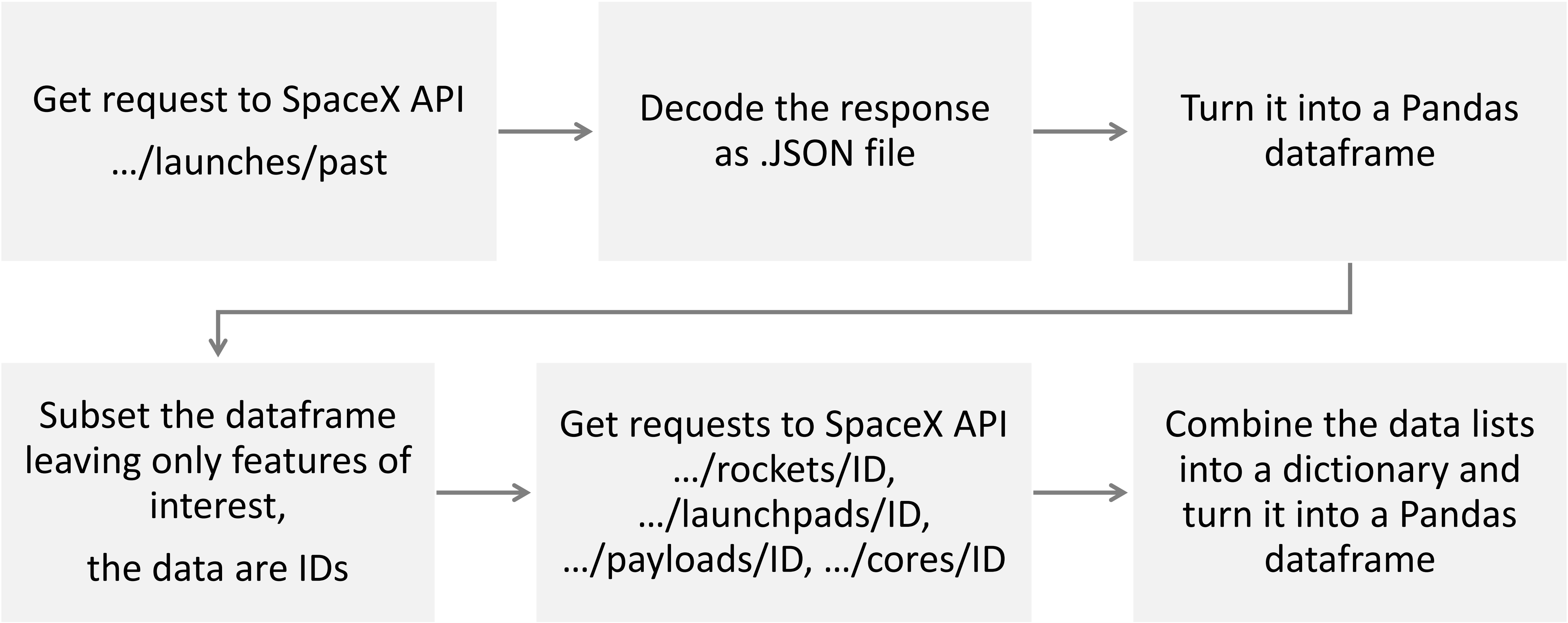


SECTION 1 – METHODOLOGY

METHODOLOGY OUTLINE

1. Data collection
   - using the SpaceX API
   - web scraping Falcon 9 launch records from Wikipedia
2. Data wrangling
   - Missing values were identified and replaced by a mean
   - Target variable “Class” was created (0 – unsuccessful and 1 - successful landing)
3. Exploratory data analysis (EDA) using SQL and visualization
4. Interactive visual analytics using Folium and Plotly Dash
5. Predictive analysis using classification models
   - Selected categorical variables were one hot encoded
   - Four models were build on train dataset using the best hyperparameters
   - The models were evaluated on the test dataset using confusion matrix

Data Collection – SpaceX API



Get request to SpaceX API

…/launches/past

Decode the response as .JSON file

Turn it into a Pandas dataframe

Subset the dataframe leaving only features of interest,

the data are IDs

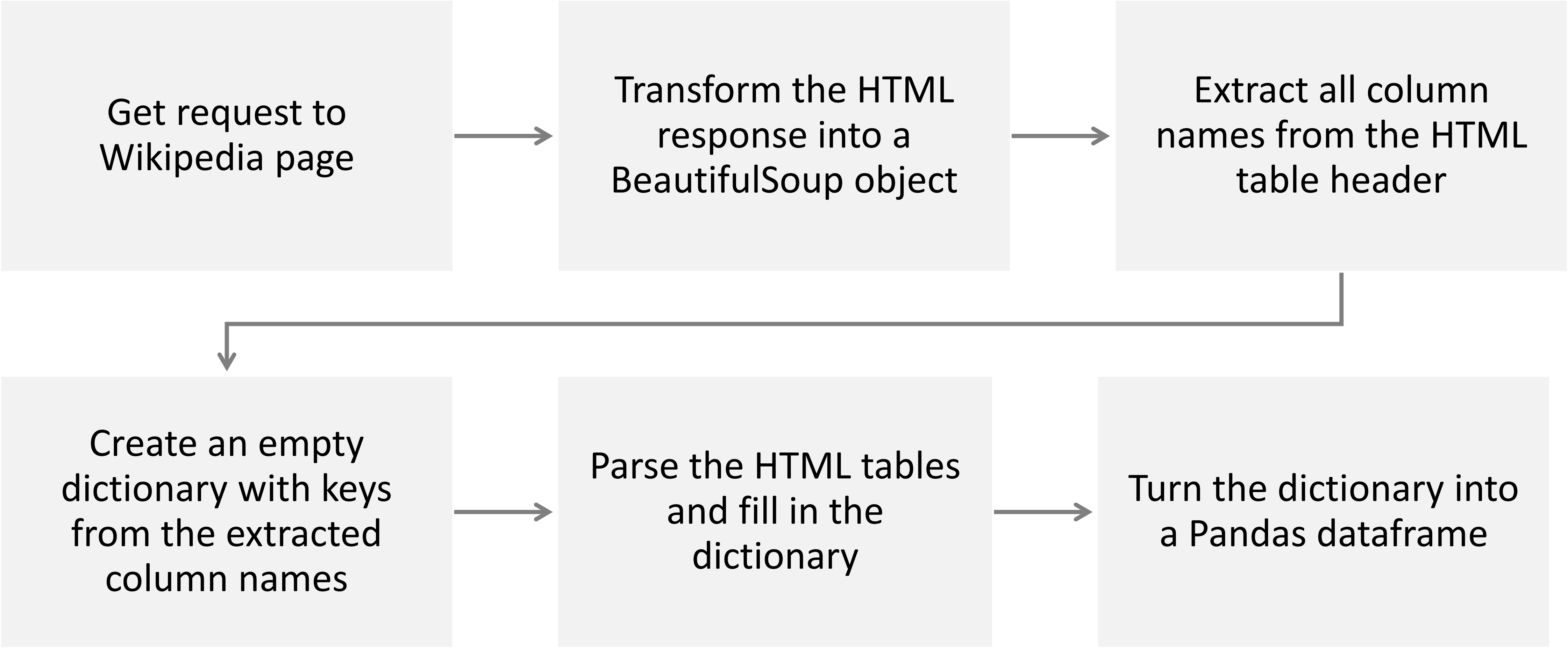
Get requests to SpaceX API …/rockets/ID, …/launchpads/ID, …/payloads/ID, …/cores/ID

Combine the data lists into a dictionary and turn it into a Pandas dataframe

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

Link to the SpaceX API notebook

Data Collection – Web Scraping



Get request to Wikipedia page

Transform the HTML response into a BeautifulSoup object

Extract all column names from the HTML table header

Create an empty dictionary with keys from the extracted column names

Parse the HTML tables and fill in the dictionary

Turn the dictionary into a Pandas dataframe

Link to the Web Scraping notebook

Data Wrangling

SpaceX API dataset:

- Data was filtered using the Booster Version column to keep only Falcon 9 launches
- The Flight Number column was reset to the natural order
- Missing values were identified for the column Payload Mass and replaced with the mean value of this column

  Link to the SpaceX API notebook

- Number of launches were calculated for each site and orbit
- The launch outcome column was used to create a “class” column, where 0 means the first stage did not land successfully and 1 – successful landing

  Link to the Data Wrangling notebook

EDA with SQL

SQL series were performed:

- To reveal the unique launch sites
- To search for launch sites which name begin with 'CCA'
- To calculate the total payload mass
- To calculate average payload mass carried by booster version F9 v1.1
- To find the date when thee first successful landing happened
- To check which boosters have success in drone ship with payload mass 4000 - 6000 kg
- To calculate the total number of successful and failure mission outcomes
- To check the booster versions which have carried the max. payload mass
- To look into the records of year 2015
- To rank the count of landing outcomes between 4.06.2010 and 20.03.2017

Link to EDA with SQL notebook

EDA with Data Visualization

The following charts were plotted using Matplotlib and Seaborn:

- Cat plots to observe any relationship between two parameters
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Flight Number vs Orbit type
  - Payload Mass vs Orbit type
- Bar chart to compare the success rate for different orbit types
- Line plot to see the success rate trend over the years

Link to the EDA with visualization notebook

Interactive Map with Folium

The following map objects were created and added to a Folium map:

- folium.Circle() to highlight a circle area around launch sites
- folium.Marker() to add a text lable on a map
- MarkerCluster() object for multiple markers with the same location
- MousePosition() to get coordinates for a mouse point on the map
- folium.PolyLine() to draw a line between a launch site and a selected point

Link to the Folium Map notebook

Dashboard with Plotly Dash

Interactive dashboard was created with Dash and several Plotly.Express graphs were added to it. Several components was added to the dash app:

- Dropdown list enabling Launch Site selection
- Pie chart with callback function showing the total count of successful launches for the site (selected with a help of the dropdown list).
- Slider allowing to select a payload range
- Scatter plot with callback function showing the correlation between the payload (selected with a help of the slider) and launch success for the site (selected with a help of the dropdown list).

Link to the Dashboard notebook

Predictive Analysis (Classification)

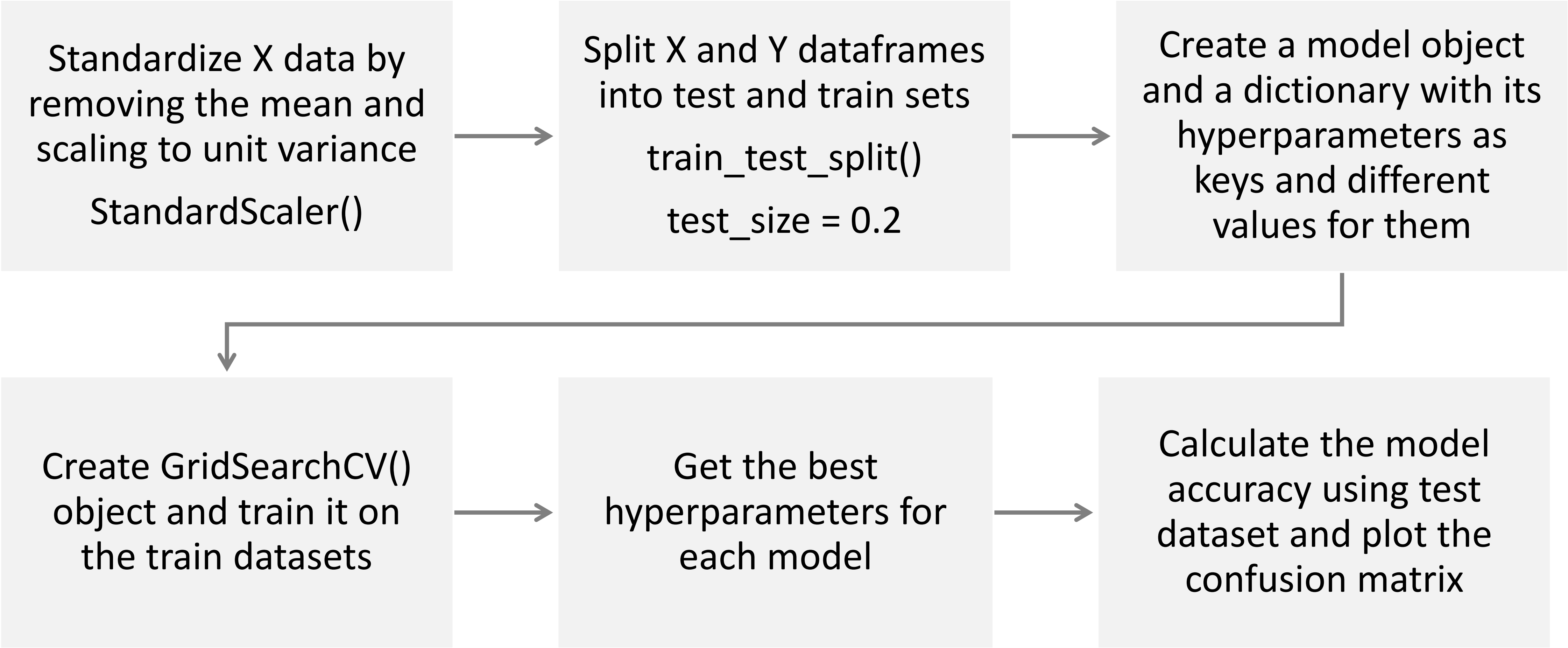
- Based on the insight from EDA, the most important variables were selected as features for the prediction model. The features were one hot encoded.

Features Engineering part

- The “class” column (launce outcome) was assigned to a target variable Y
- The one hot encoded features were assigned to independent variables X
- Four different algorithms were used to build the model: logistic regression, support vector machine, decision tree classifier, and k nearest neighbors.

Link to the Classification notebook

Predictive Analysis (Classification)



Standardize X data by removing the mean and scaling to unit variance

StandardScaler()

Split X and Y dataframes into test and train sets

train_test_split()

test_size = 0.2

Create a model object and a dictionary with its hyperparameters as keys and different values for them

Create GridSearchCV() object and train it on the train datasets

Get the best hyperparameters for each model

Calculate the model accuracy using test dataset and plot the confusion matrix

RESULTS – OUTLINE

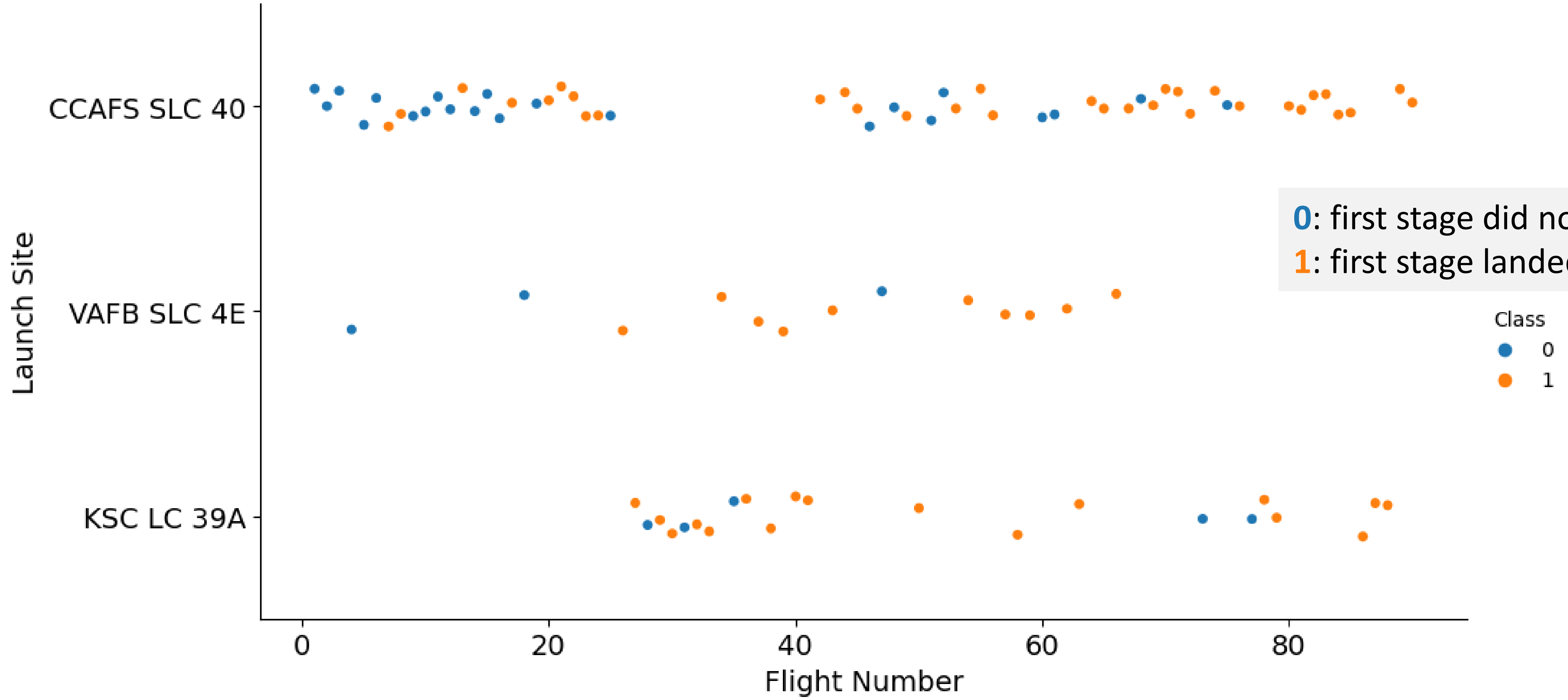
1. Exploratory data analysis results
   - using visualization tools
   - using SQL queries

2. Interactive analytics in screenshots
   - Dashboard with a pie chart and a scatter plot
   - Folium map with launch site locations

3. Predictive analysis results
   - logistic regression
   - support vector machine
   - decision tree classifier
   - K-nearest neighbors

SECTION 2
INSIGHTS DRAWN FROM EDA

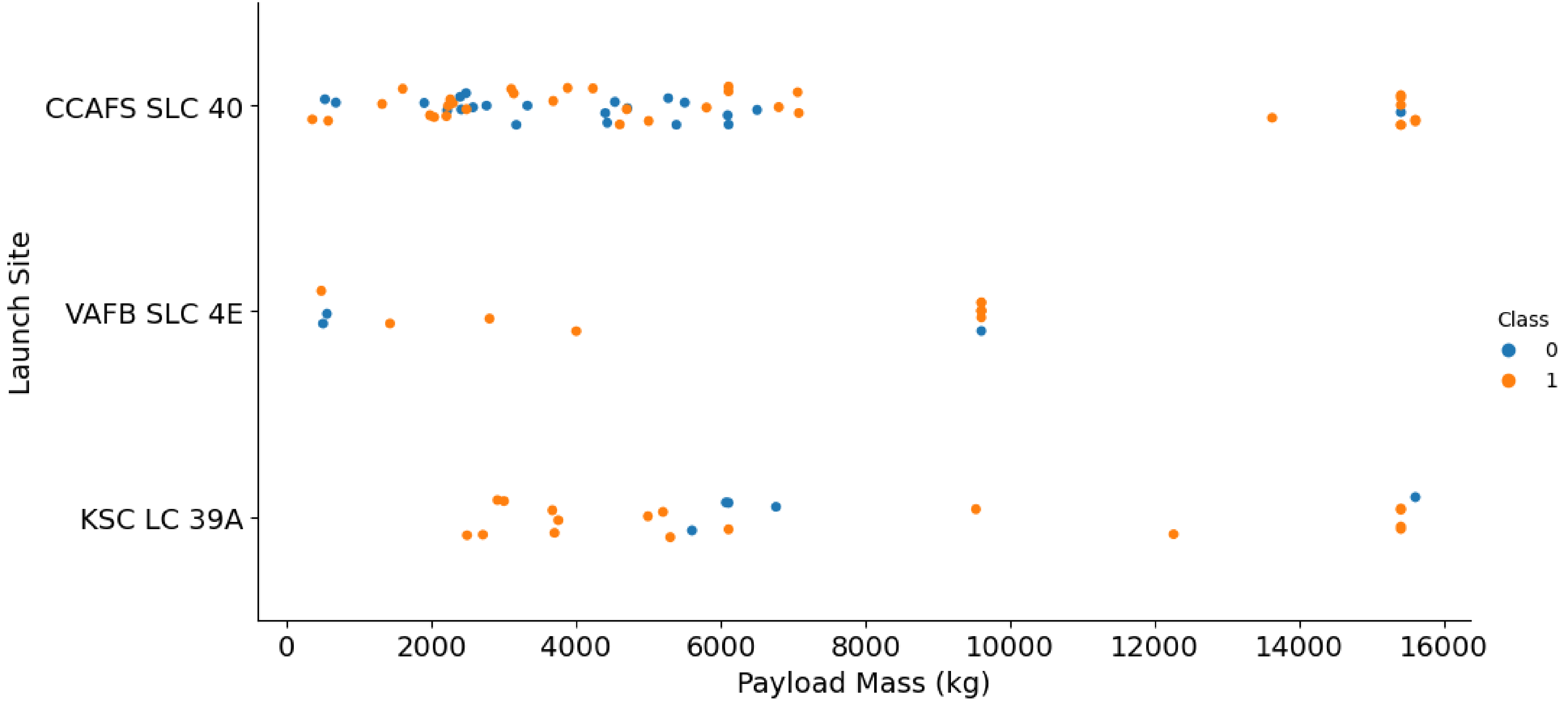
Flight Number vs. Launch Site



0: first stage did not land
1: first stage landed

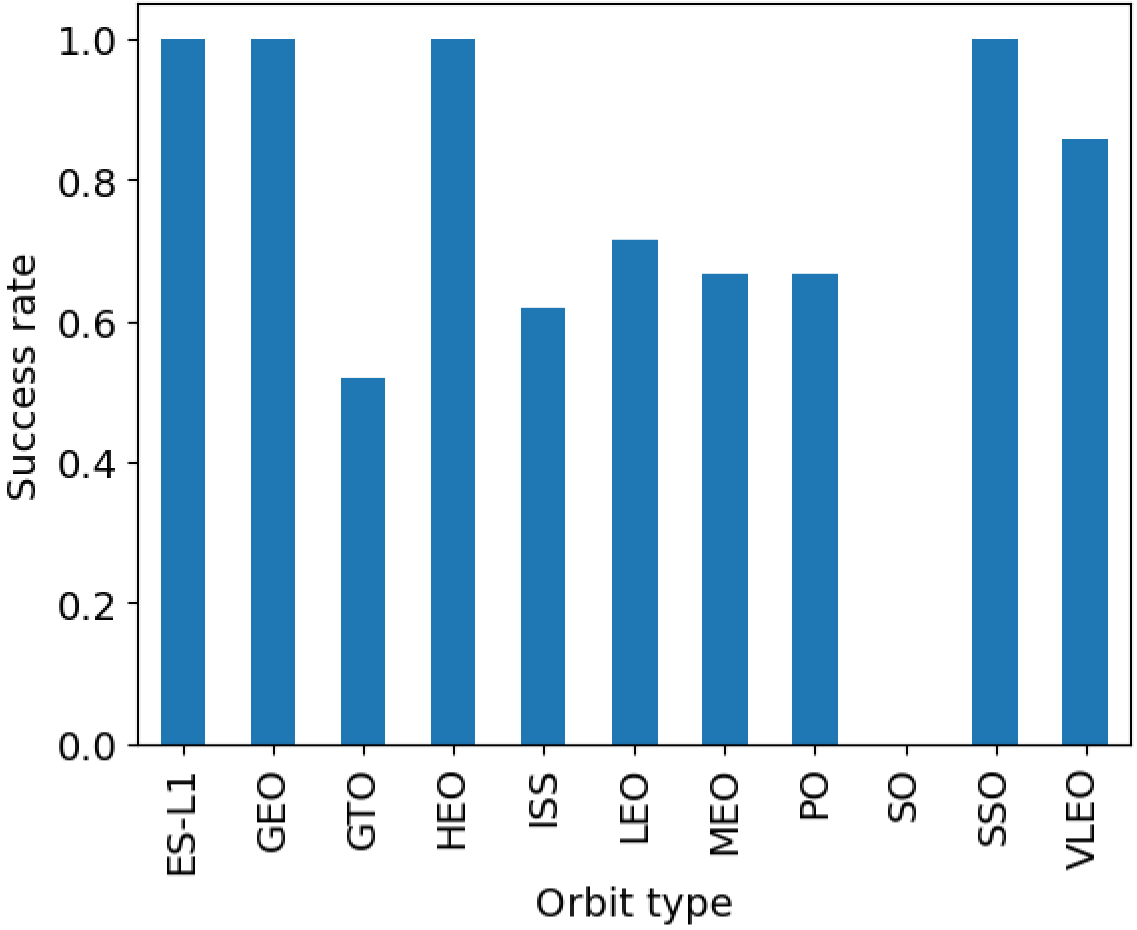
- The majority of launches were performed from CCAFS LC-40 site. There were a lot of failures in the range of flight numbers 1 - 25 with a success rate of 39%, but later success rate increased to 75% for flight numbers 26 - 90. Total success rate for KSC LC 39A is 77% and for VAFB SLC 4E – 77%.

Payload Mass vs. Launch Site



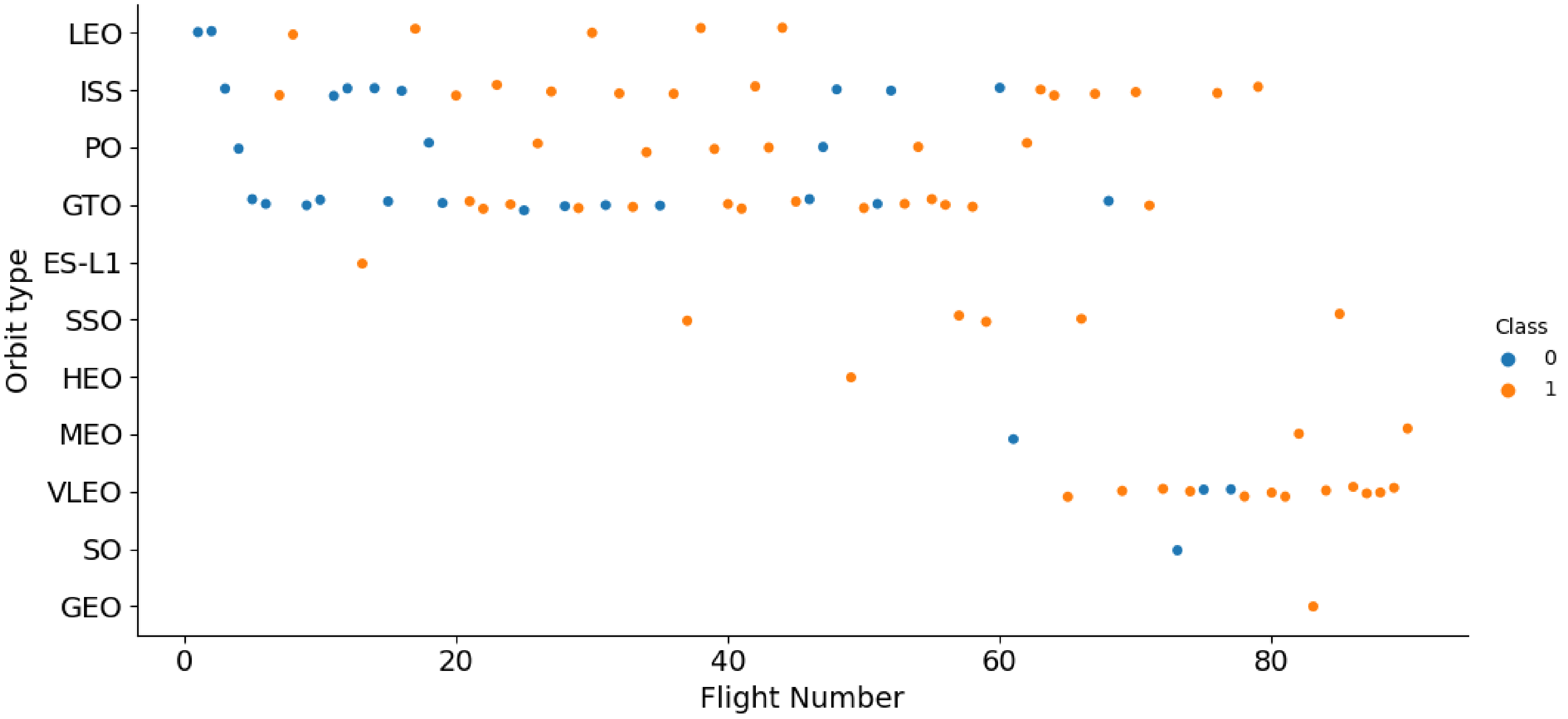
➢ For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000). The majority of launches with lower payload mass (0 – 8000 kg) were done from CCAFS SLC 40 site.

Success Rate vs. Orbit Type



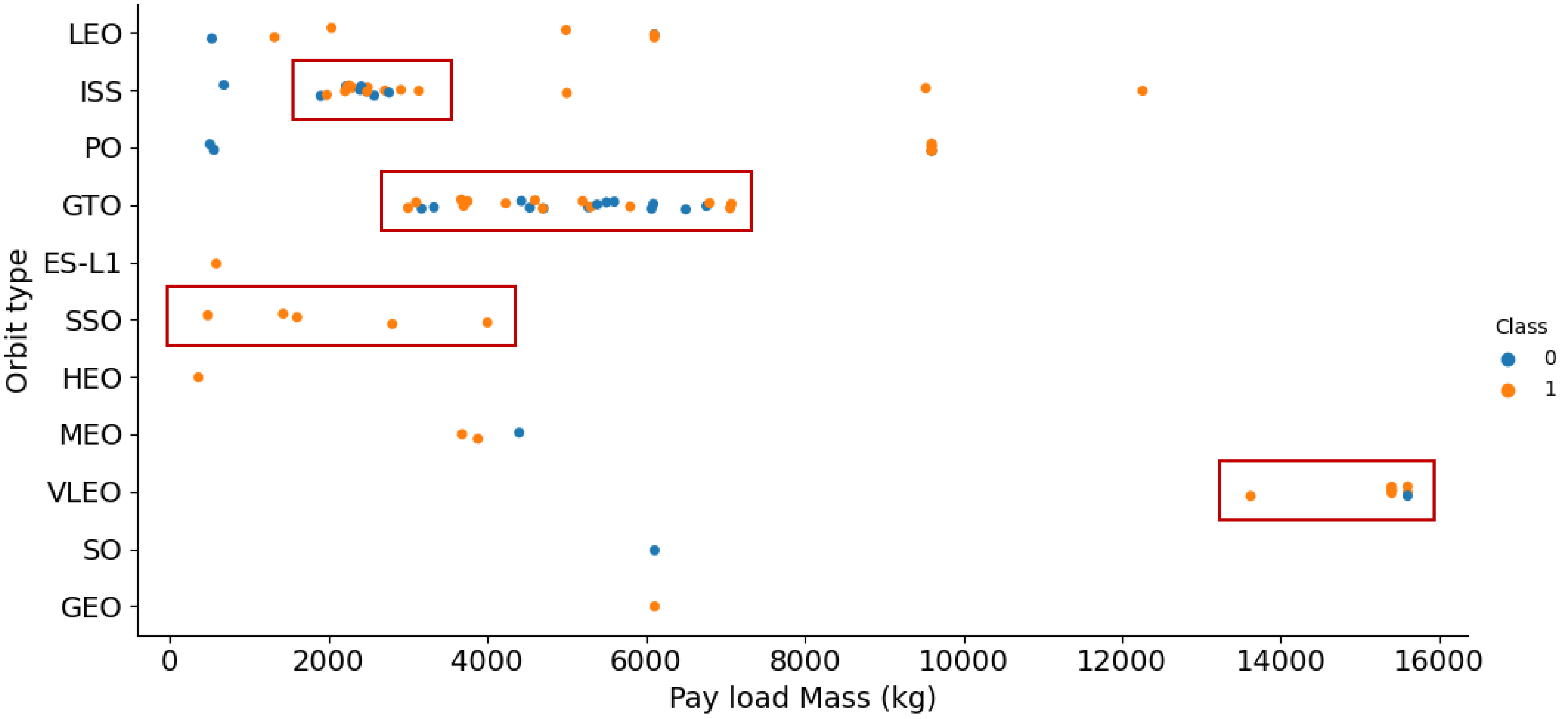
- ➢ ES-L1, GEO, HEO, SSO, VLEO orbits have high success rates, while GTO, ISS, LEO, MEO, PO have success rate only about 60%. SO orbit does not have successful launches.
- ➢ It is important to look into the number of launches per each orbit, since 0 and 100% success rates could by due to a single data point.

Flight Number vs. Orbit Type



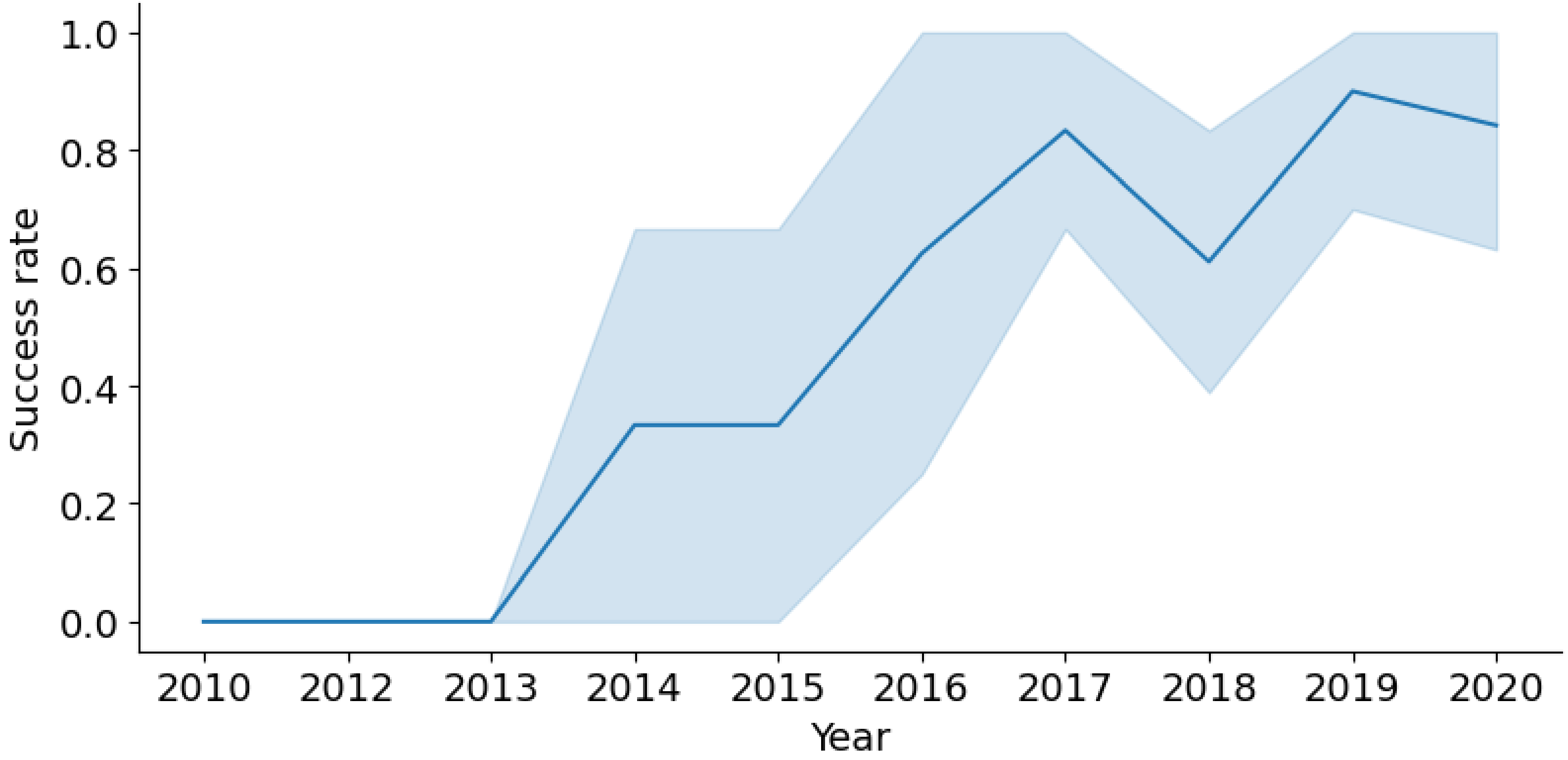
➢ Among orbits with the highest success rate ES-L1, GEO, HEO had just 1 attempt, SSO – 5 attempts and all successful, VLEO – 12 successful out of 14 total. Launches to the VLEO orbit started in the resent years. GTO and ISS orbits have the highest number of launches and the lowest success rate.

Payload vs. Orbit Type



- Only heavy payload mass (>13000 kg) was launched to the VLEO orbit. In contrast, only low payload mass (< 5000 kg) was used in case of SSO orbit. The majority of launches to ISS were done with 2000 – 4000 kg payload mass and to GTO – with 3000 – 7000 kg.

Launch Success Yearly Trend



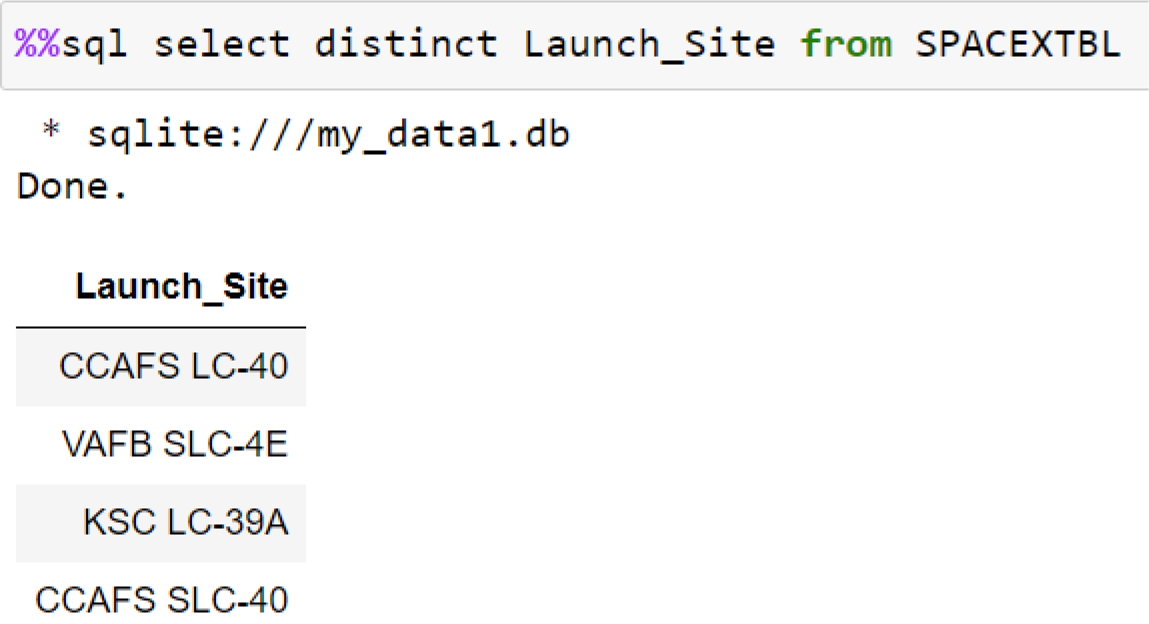
- ➢ Overall, since 2013 the launch success was significantly increasing till 2019. It stabilized in the recent years.

All Launch Site Names

➢ Find unique launch sites

○ DISTINCT clause was used for the column Launch_Site

✓ There are 4 different launch sites



```
%%sql select distinct Launch_Site from SPACEXTBL
```

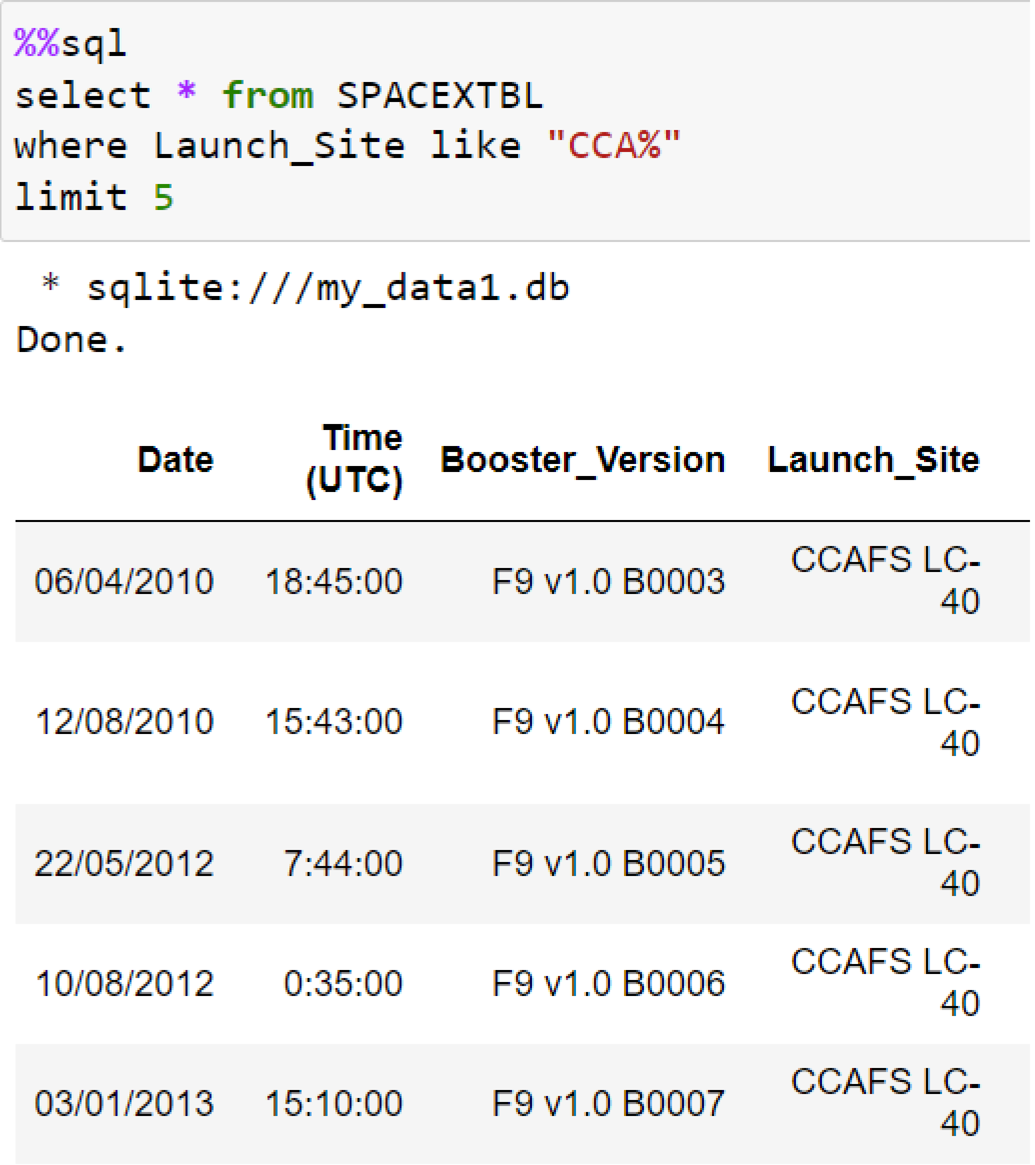
```
 * sqlite:///my_data1.db
Done.
```

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`

- ○ WHERE clause, LIKE operator and LIMIT clause were used for the query



```
%%sql
select * from SPACEXTBL
where Launch_Site like "CCA%"
limit 5
```

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

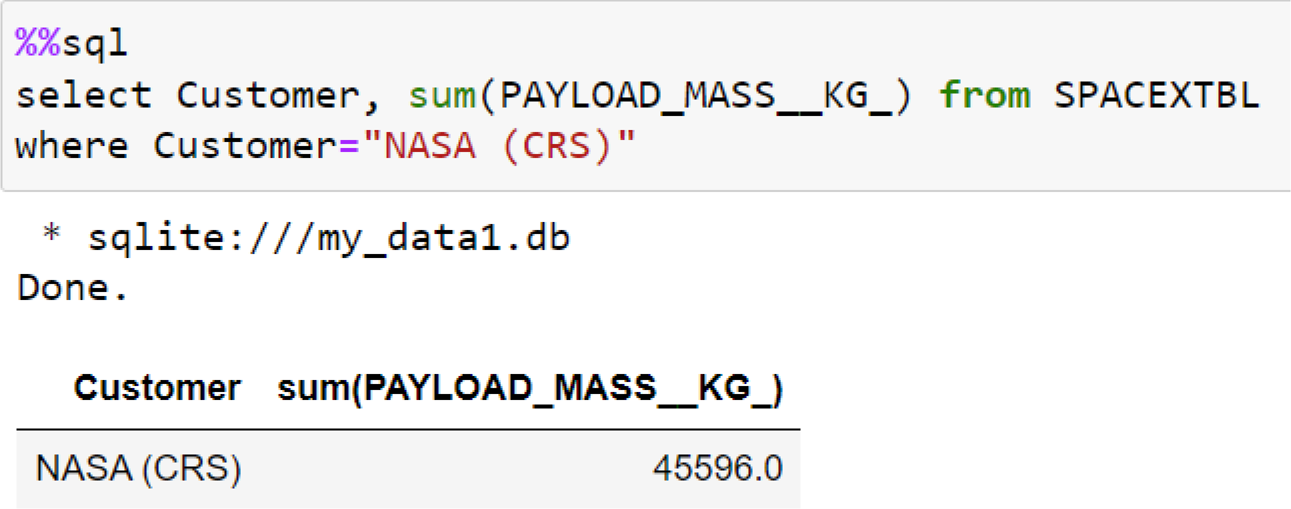
Date	Time (UTC)	Booster_Version	Launch_Site
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40

Total Payload Mass

➢ Calculate the total payload carried by boosters from NASA

○ SUM() function and WHERE clause were used for the query

✓ 45596 kg



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

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

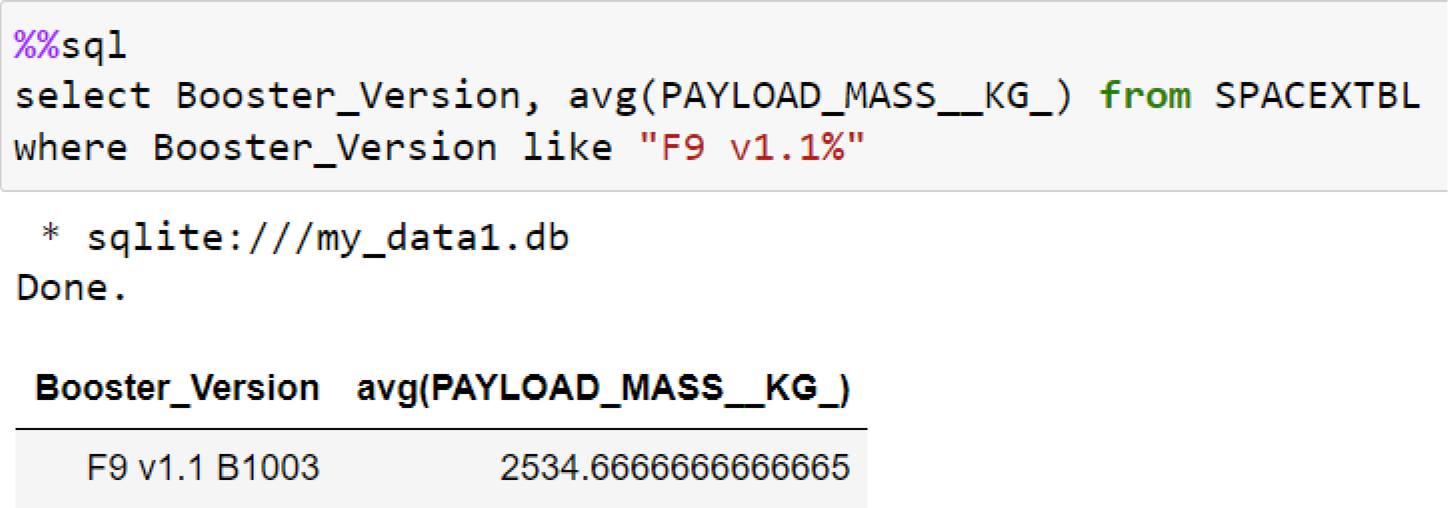
Customer	sum(PAYLOAD_MASS__KG_)
NASA (CRS)	45596.0

Average Payload Mass by F9 v1.1

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

○ AVG() function, WHERE clause and LIKE operator were used for the query

✓ 2535 kg



```
%%sql
select Booster_Version, avg(PAYLOAD_MASS__KG_) from SPACEXTBL
where Booster_Version like "F9 v1.1%"
```

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

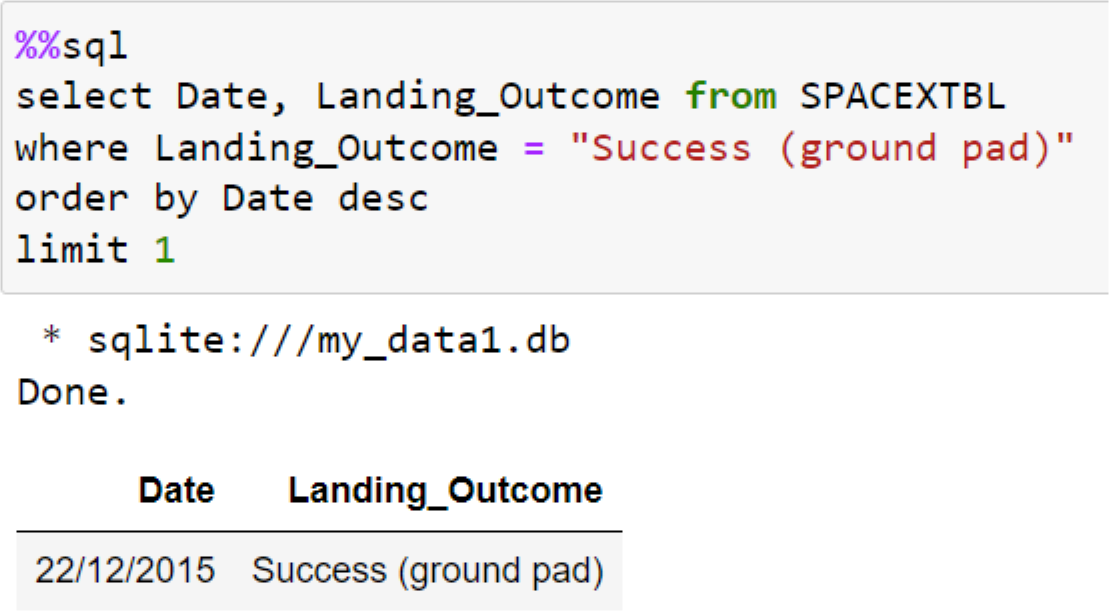
Booster_Version	avg(PAYLOAD_MASS__KG_)
F9 v1.1 B1003	2534.6666666666665

First Successful Ground Landing Date

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

○ WHERE clause, ORDER BY statement, and LIMIT clause were used for the query

✓ 22.12.2015

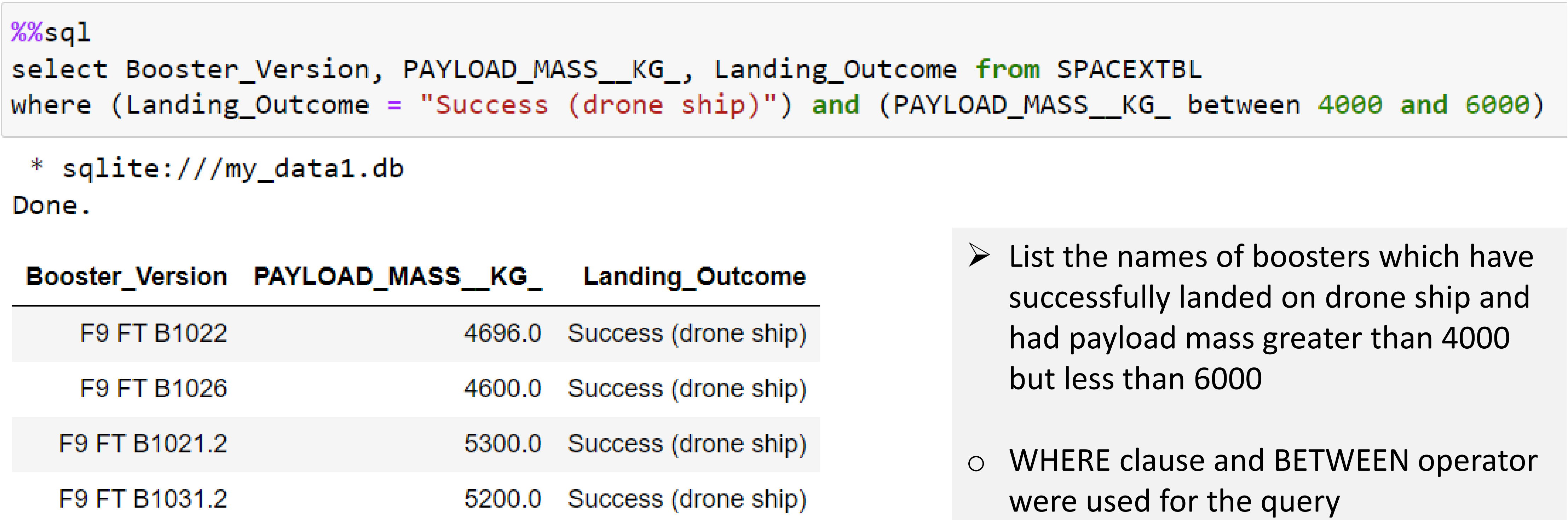


```
%%sql
select Date, Landing_Outcome from SPACEXTBL
where Landing_Outcome = "Success (ground pad)"
order by Date desc
limit 1
```

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

Date	Landing_Outcome
22/12/2015	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000 kg



```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_, Landing_Outcome from SPACEXTBL
where (Landing_Outcome = "Success (drone ship)") and (PAYLOAD_MASS__KG_ between 4000 and 6000)
```

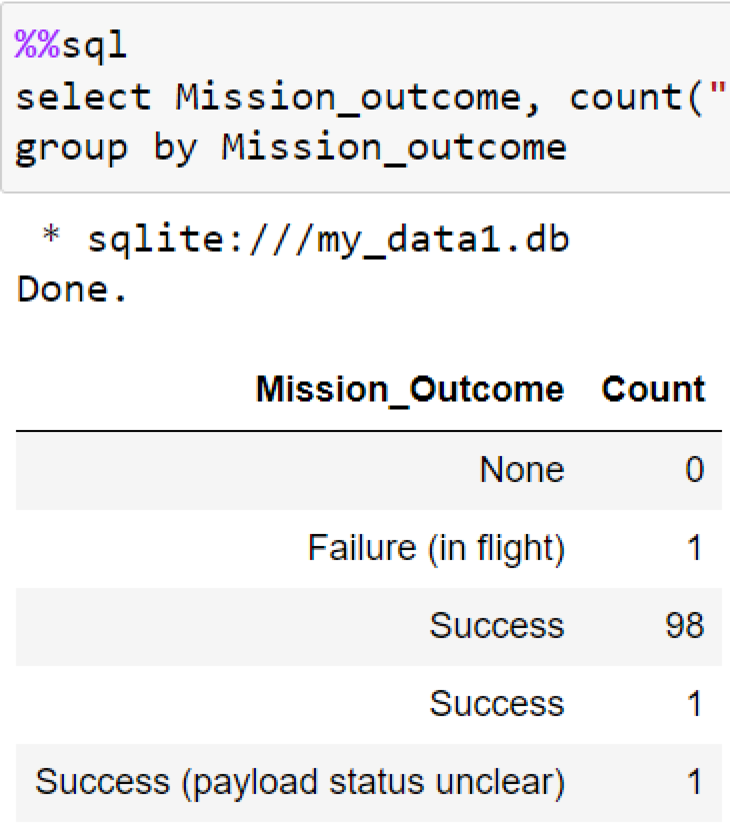
```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696.0	Success (drone ship)
F9 FT B1026	4600.0	Success (drone ship)
F9 FT B1021.2	5300.0	Success (drone ship)
F9 FT B1031.2	5200.0	Success (drone ship)

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

- WHERE clause and BETWEEN operator were used for the query

Total Number of Successful and Failure Mission Outcomes



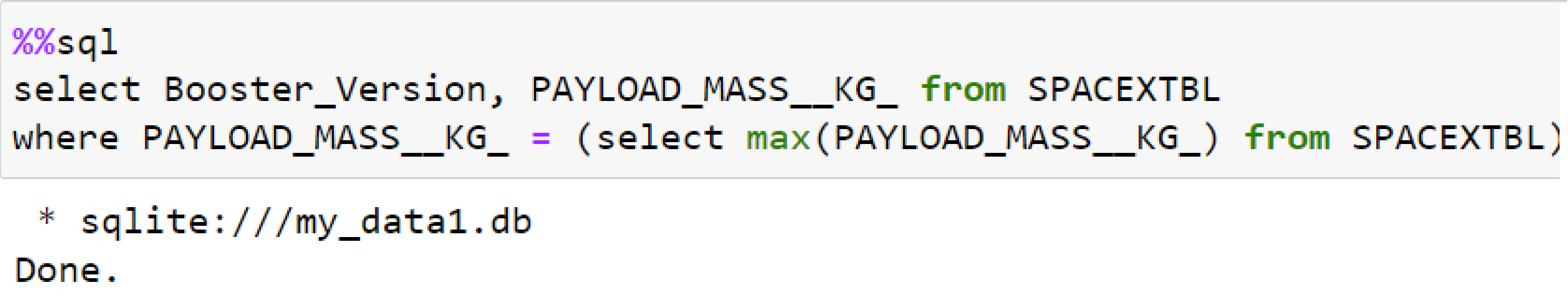
```
%%sql
select Mission_outcome, count("Mission_outcome") as Count from SPACEXTBL
group by Mission_outcome
```

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

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

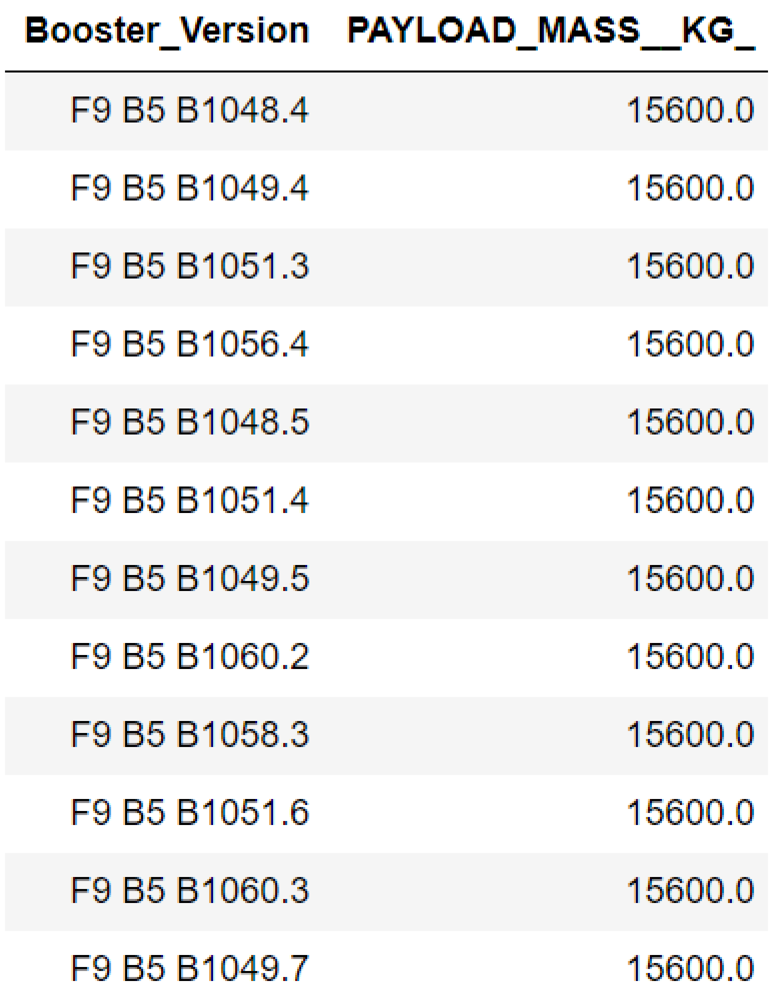
- ➢ Calculate the total number of successful and failure mission outcomes
- ○ COUNT() function and GROUP BY clause were used for the query
- ✓ Only 1 failure mission outcome

Boosters Carried Maximum Payload



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

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



Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

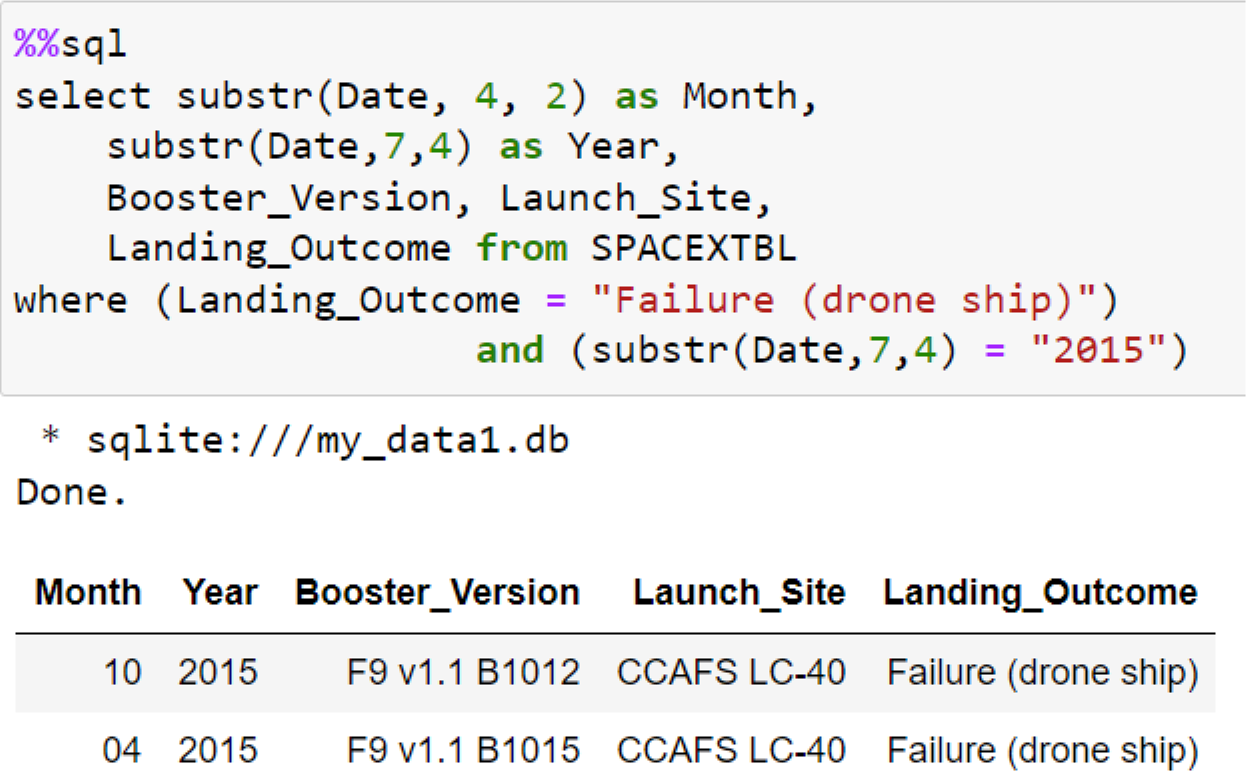
- ➢ List the names of the booster which have carried the maximum payload mass

- ○ WHERE clause and sub-query with MAX() function were used for the query

2015 Launch Records

- ➢ List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

- ○ SUBSTR() function and WHERE clause were used for the query

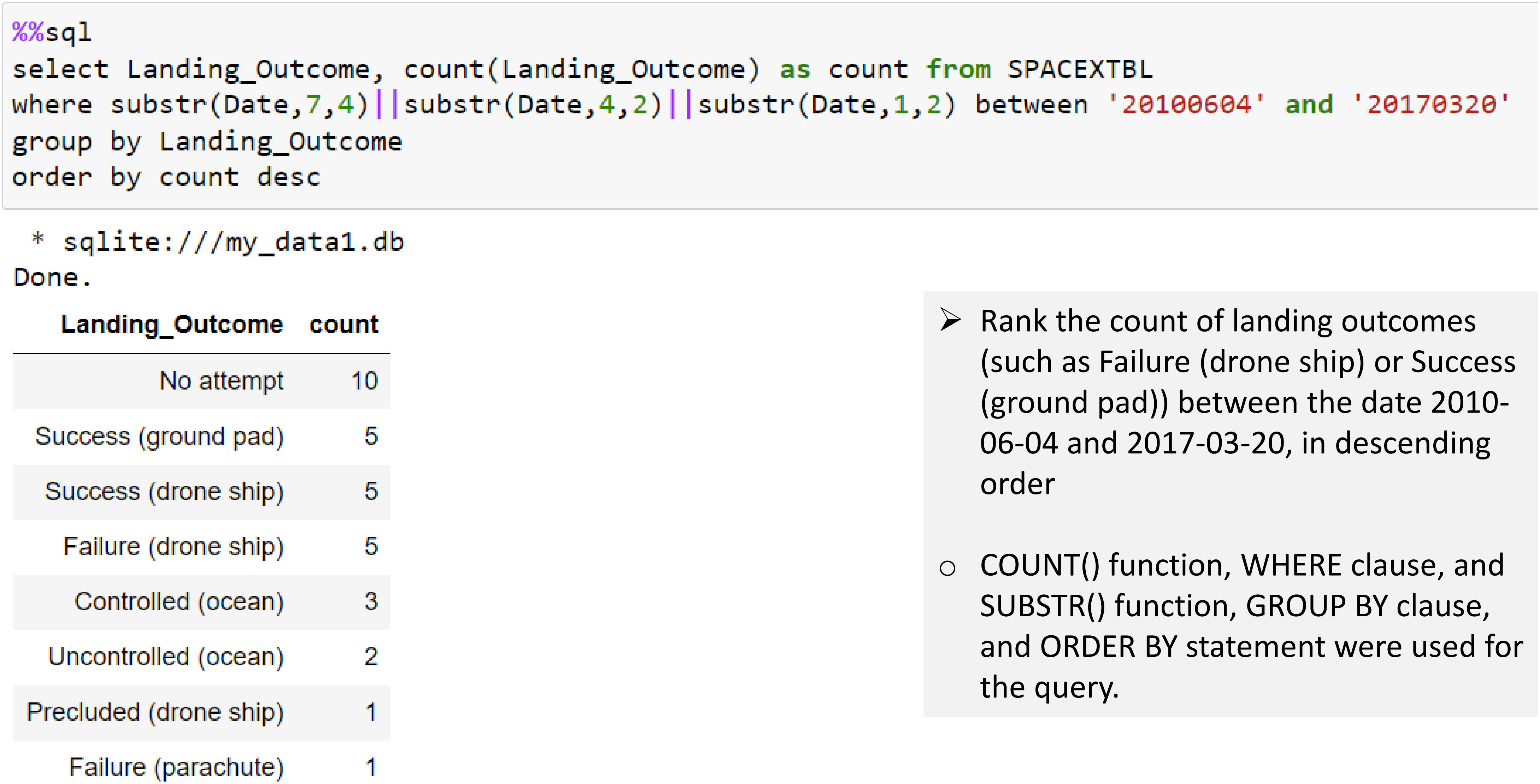


```
%%sql
select substr(Date, 4, 2) as Month,
    substr(Date,7,4) as Year,
    Booster_Version, Launch_Site,
    Landing_Outcome from SPACEXTBL
where (Landing_Outcome = "Failure (drone ship)")
                        and (substr(Date,7,4) = "2015")
```

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

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



```
%%sql
select Landing_Outcome, count(Landing_Outcome) as count from SPACEXTBL
where substr(Date,7,4)||substr(Date,4,2)||substr(Date,1,2) between '20100604' and '20170320'
group by Landing_Outcome
order by count desc
```

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

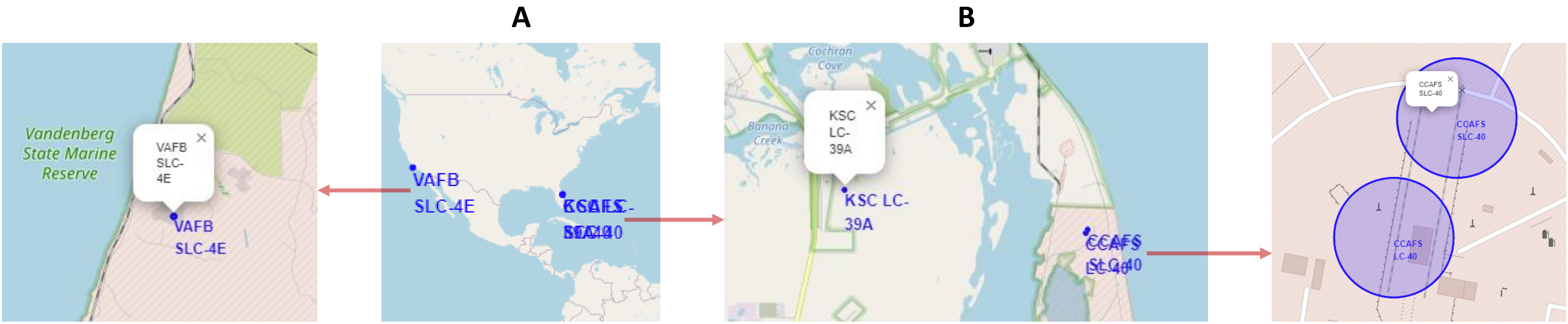
Landing_Outcome	count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

- ➢ 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

- ○ COUNT() function, WHERE clause, and SUBSTR() function, GROUP BY clause, and ORDER BY statement were used for the query.

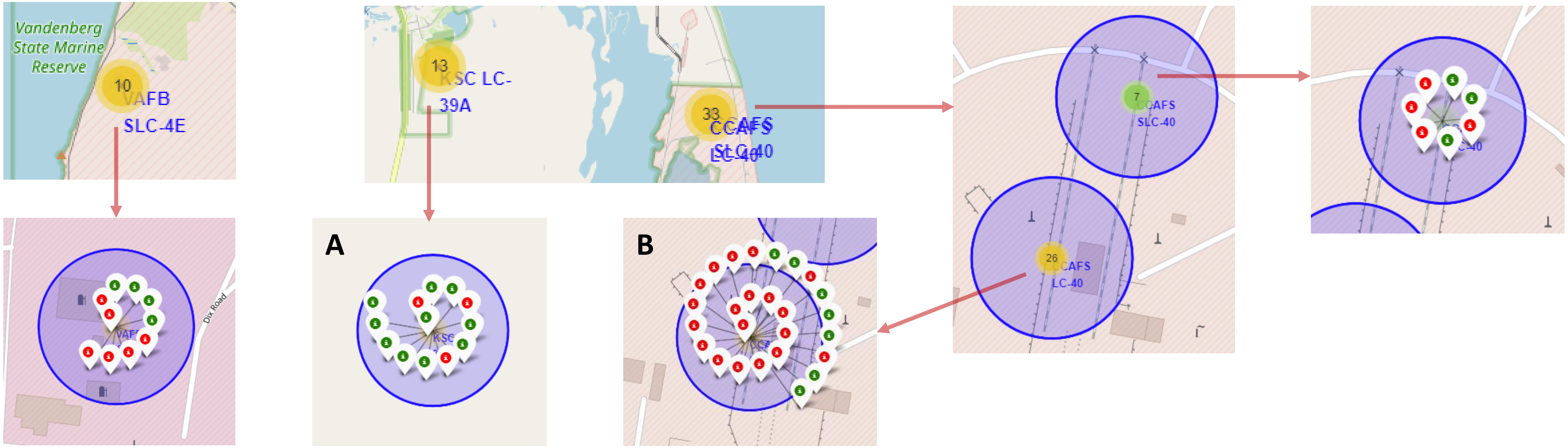
SECTION 3
LAUNCH SITE PROXYMITIES ANALYSIS

All Launch Sites on a Map



- (A) The launch sites are located in the south of the US and are in a close proximity to the coast.
- (B) Three of them (KSC LC-39A, CCAFS LC-40, and CCAFS SLC-40) are located on the coast with water to the east, which is beneficial in terms of safety and speed.
- The launches are usually eastward and located closer to the equator to take advantage of a boost to rocket velocity from the rotation of the earth.
- It is desirable for the rocket path to be over uninhabited areas due to a crash chance. The sea is considered an uninhabited area at least of permanent buildings.

Success / Failed Launches for each site

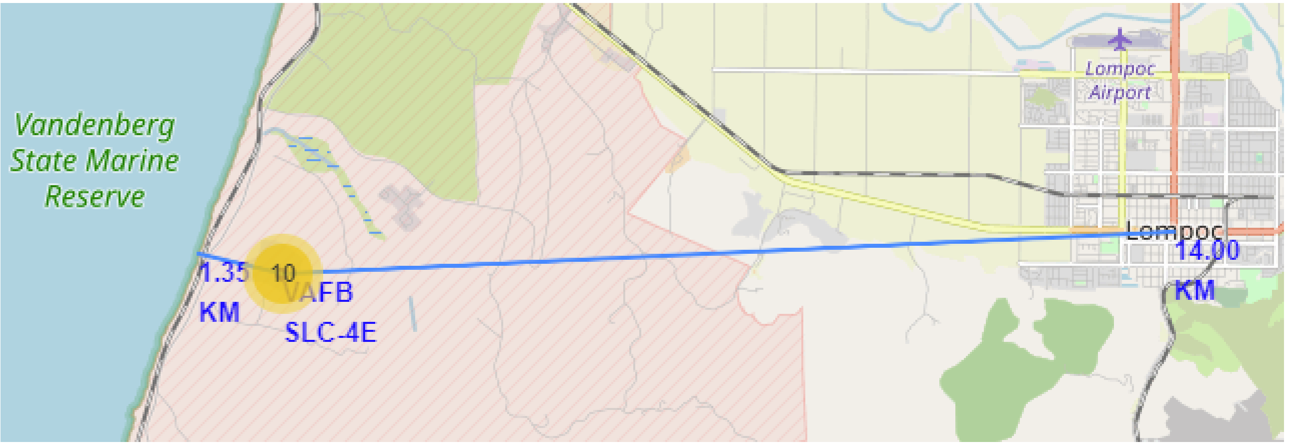


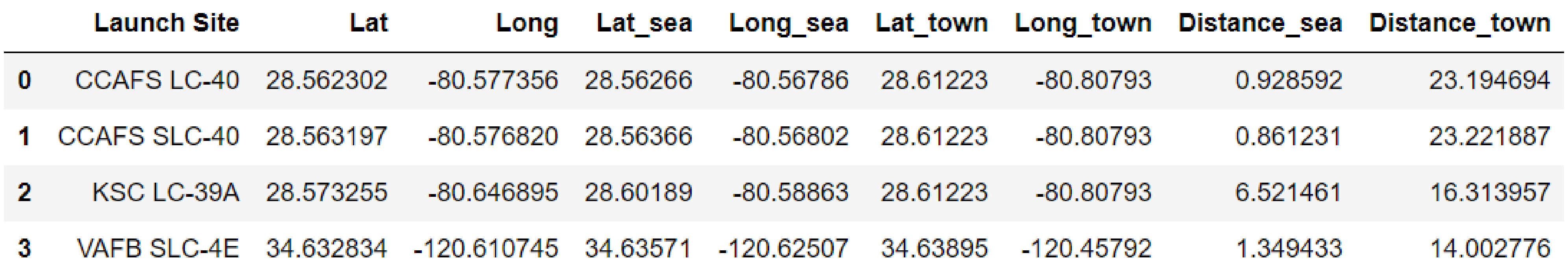
- Markers were created for all launch records. If a launch was successful, a green marker was used. If a launch was failed, a red marker was used.
- Many launch records have the exact same coordinate. Therefore, clusters of markers were used to simplify a map.
- (A) KSC LC-39A launch site have the highest success rate.
- (B) The highest amount of launches were done from CAFS LC-40 site (26 attempts).

Launch Site Proximities

- There are railways and highways in launch site close proximities (within 1 km), which might be important for supply deliveries

- The launch sites maintain a safe distance to near cities (14 – 23 km)

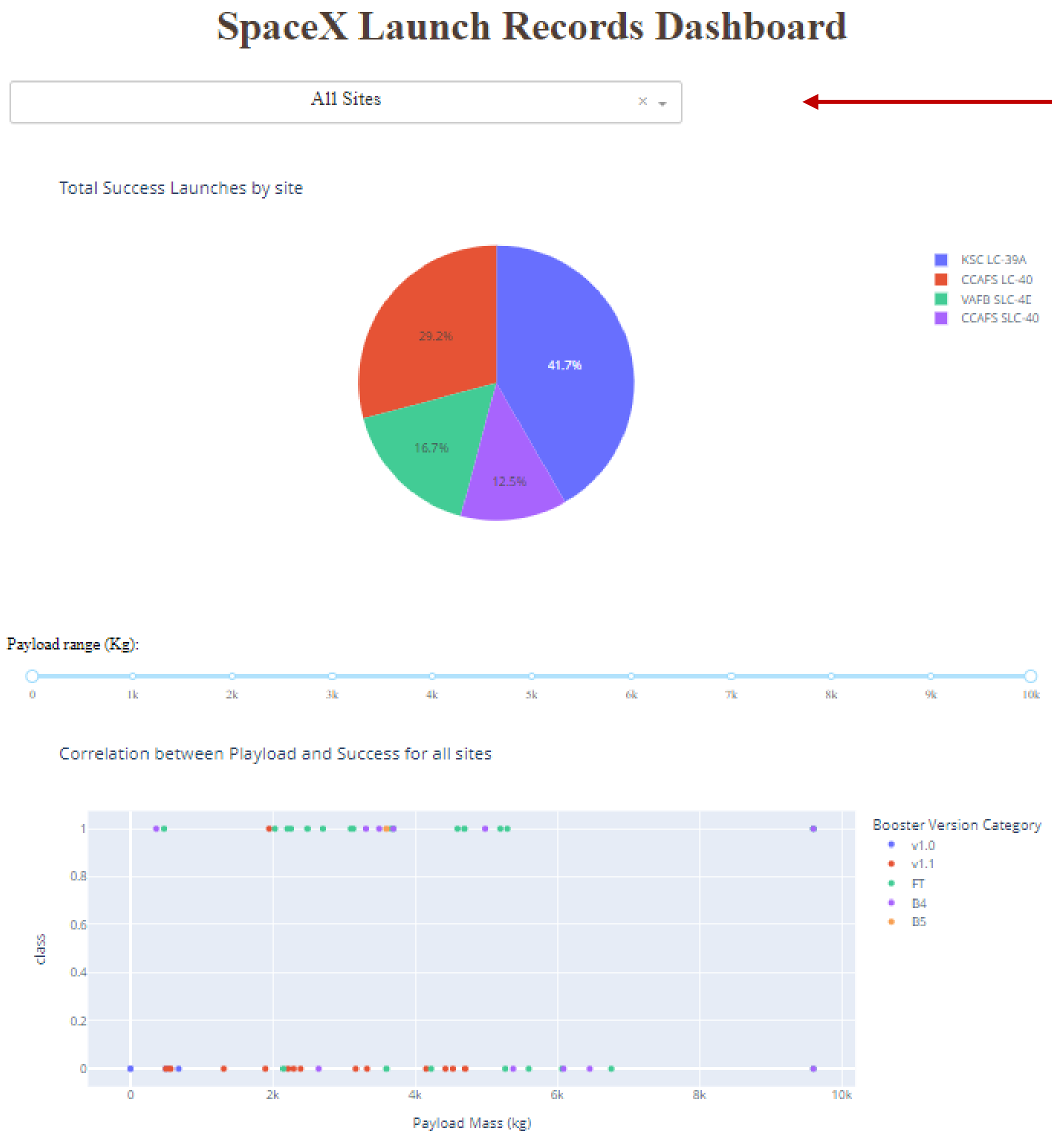




	Launch Site	Lat	Long	Lat_sea	Long_sea	Lat_town	Long_town	Distance_sea	Distance_town
0	CCAFS LC-40	28.562302	-80.577356	28.56266	-80.56786	28.61223	-80.80793	0.928592	23.194694
1	CCAFS SLC-40	28.563197	-80.576820	28.56366	-80.56802	28.61223	-80.80793	0.861231	23.221887
2	KSC LC-39A	28.573255	-80.646895	28.60189	-80.58863	28.61223	-80.80793	6.521461	16.313957
3	VAFB SLC-4E	34.632834	-120.610745	34.63571	-120.62507	34.63895	-120.45792	1.349433	14.002776

SECTION 4
BUILD A DASHBOARD WITH PLOTLY DASH

Interactive Dashboard – Overview



SpaceX Launch Records Dashboard

All Sites

Total Success Launches by site

Payload range (Kg):

Correlation between Playload and Success for all sites

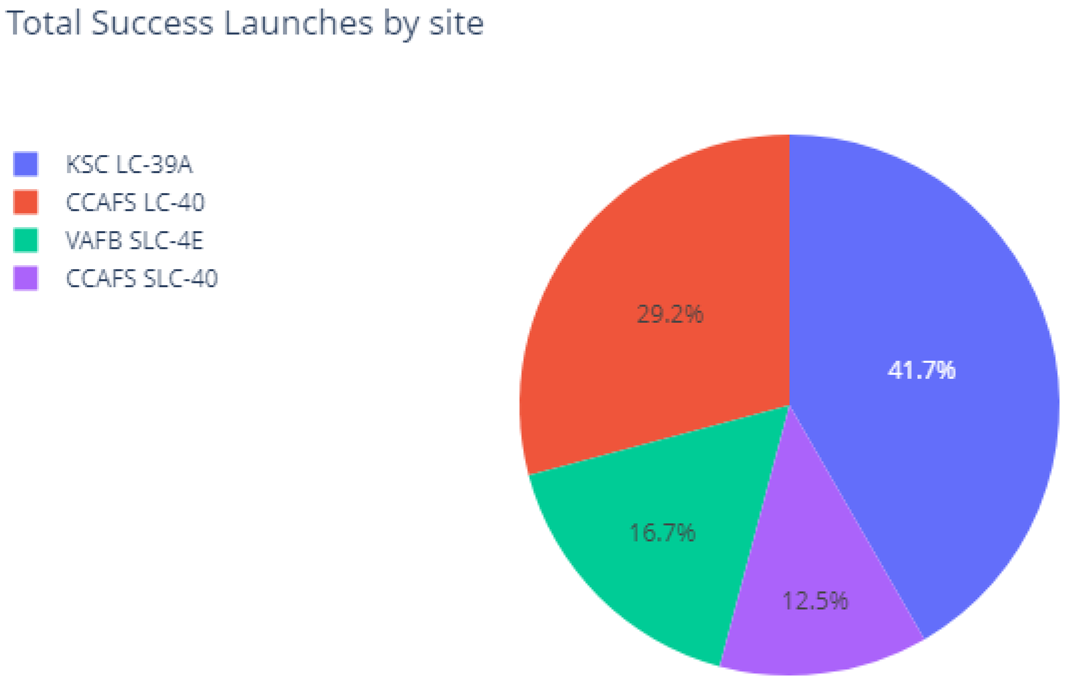
Dropdown list to select launch sites

Pie chart showing success rate of launches

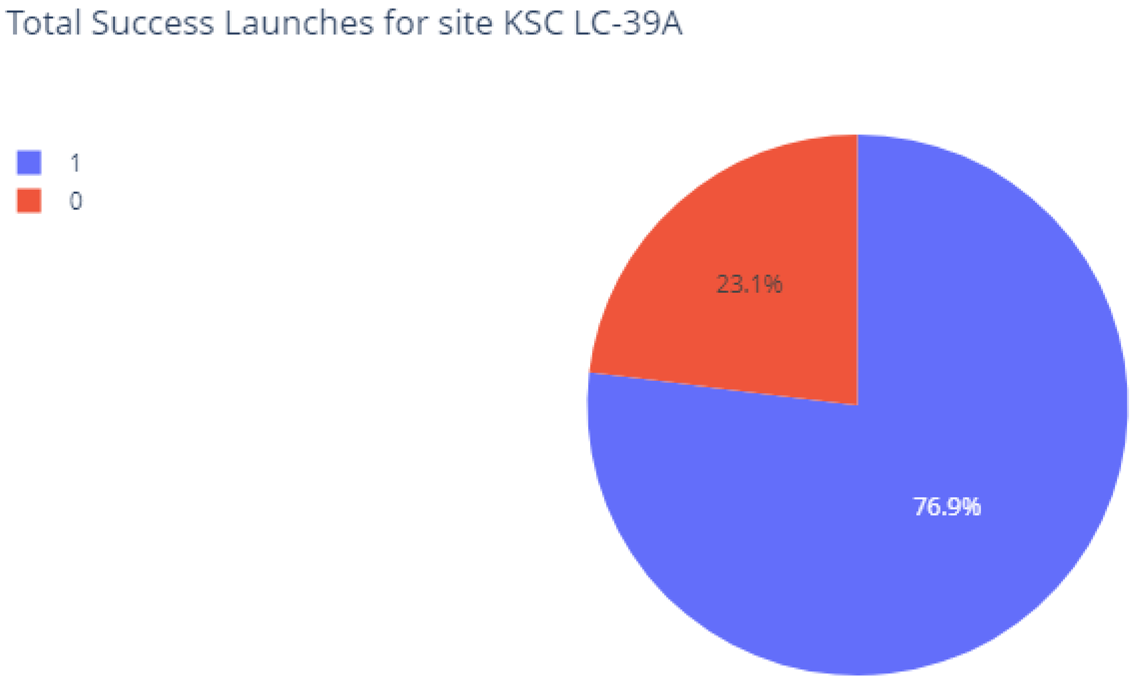
Slider to select a payload range

Scatter plot showing a correlation between success rate of launches and used payload mass

KSC LC-39A site has the highest success rate

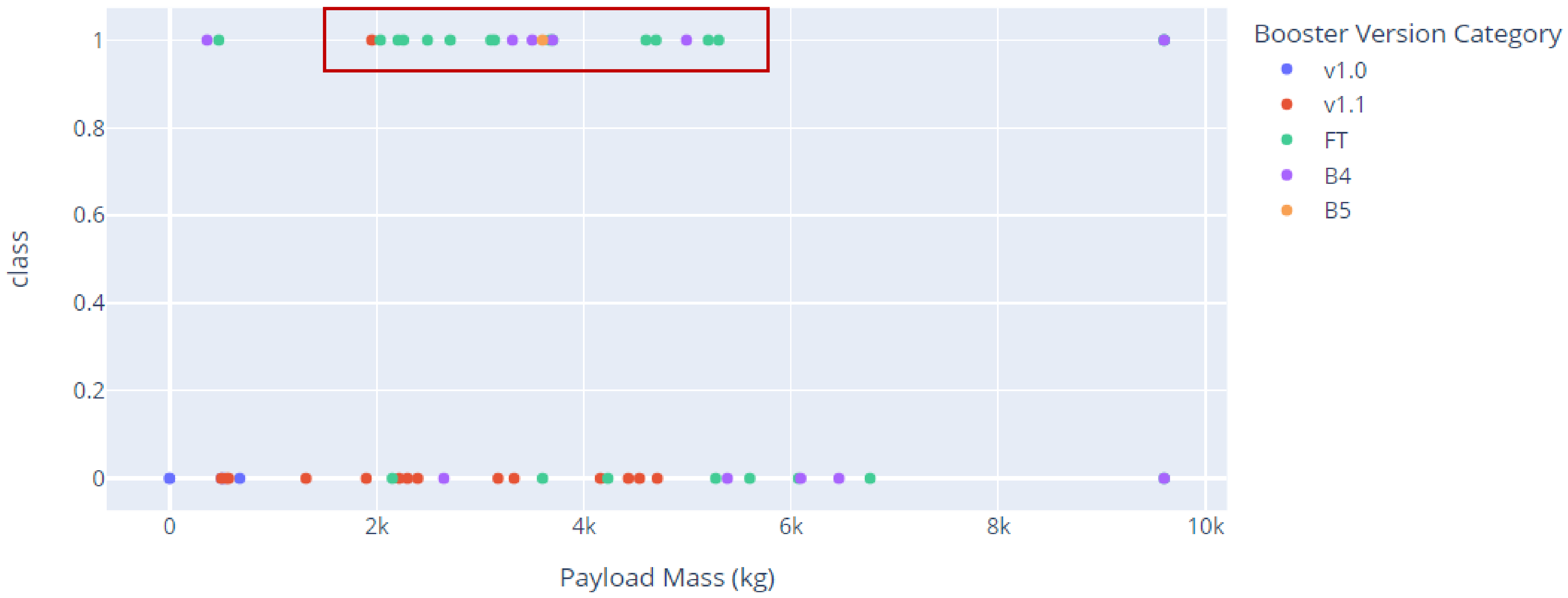


Total Success Launches by site



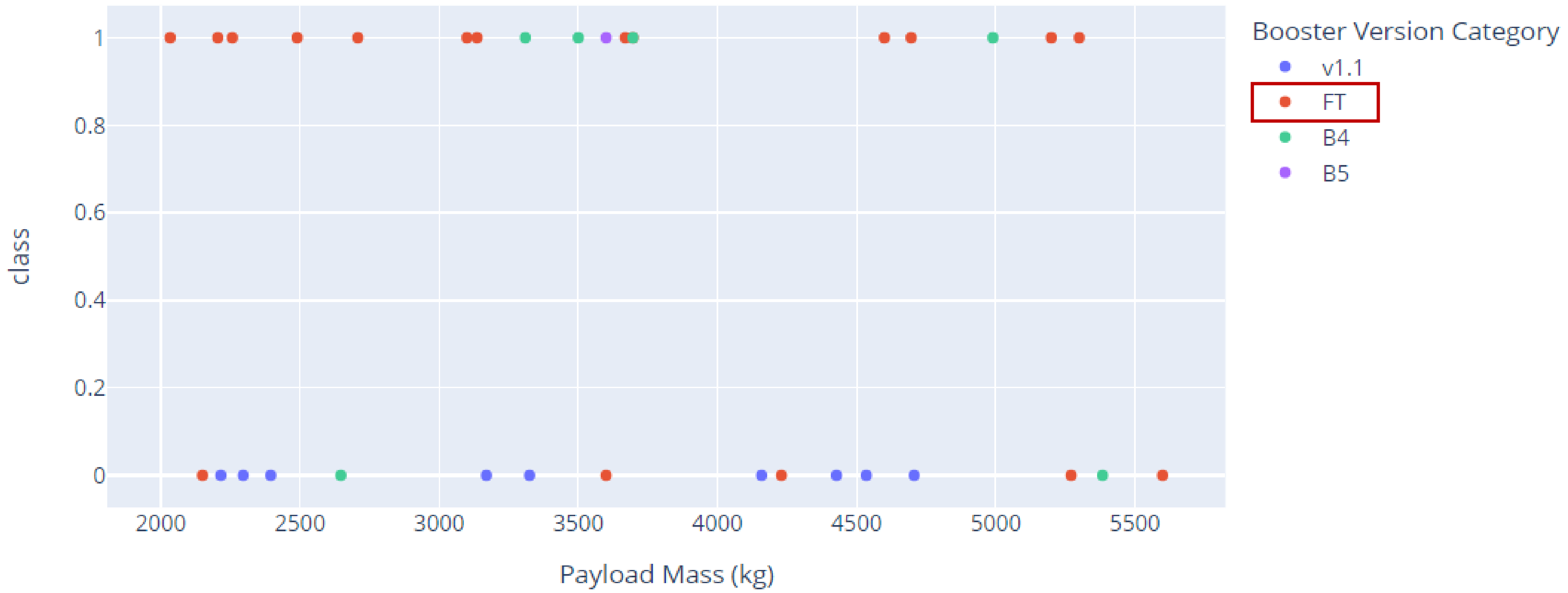
Total Success Launches for site KSC LC-39A

Payload vs. Launch Success for all sites



- Payload range of 6.000 – 10.000 has the lowest success rate
- Only 3 (out of 11) successful launches with payload 0 – 2.000 kg
- Only 3 (out of 9) successful launches with payload 6.000 – 10.000 kg

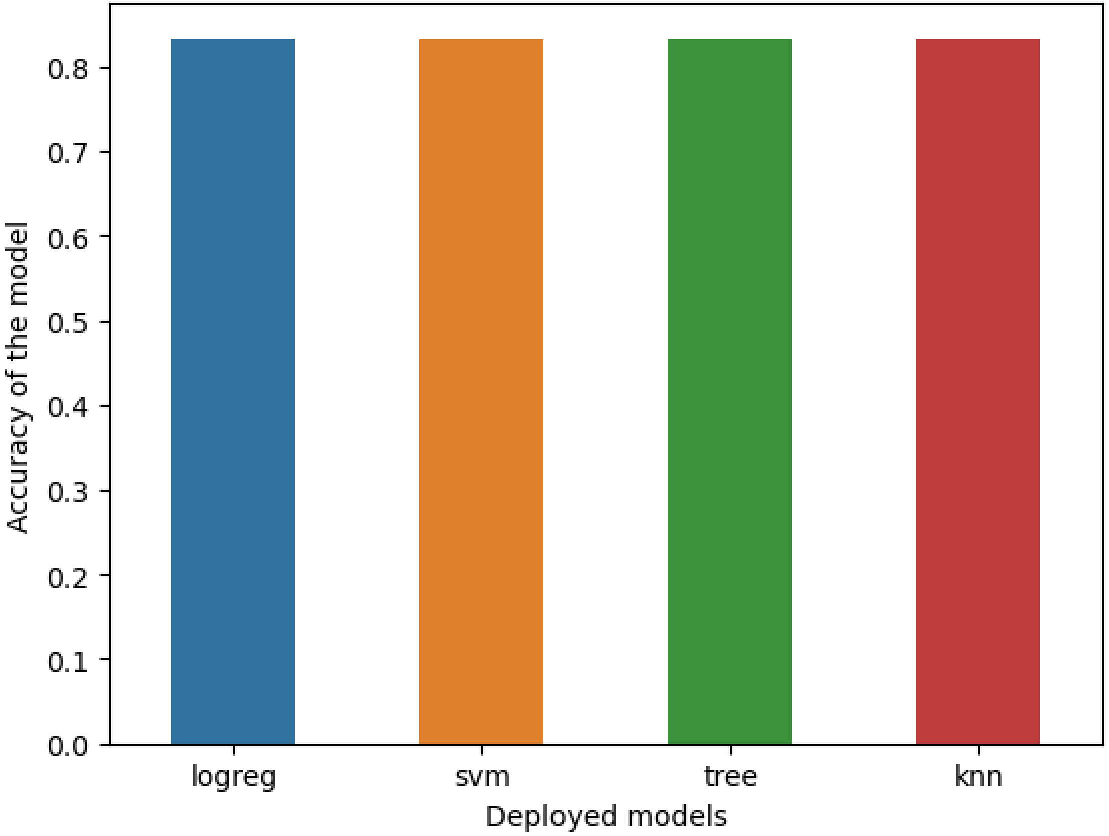
Payload vs. Launch Success for all sites



- Booster version FT (16 out 24) has the highest launch success rate
- v1.1 (1 out 15) and v1.0 (0 out 5) have the lowest success rate

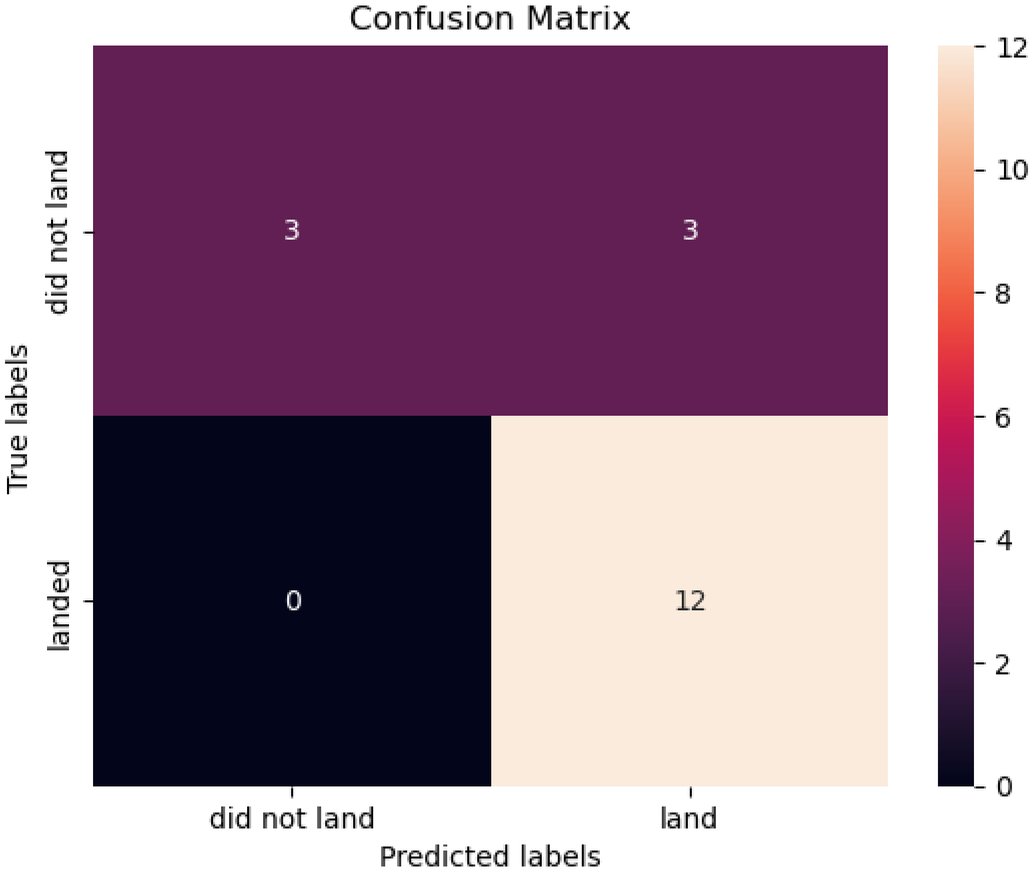
SECTION 5
PREDICTIVE ANALYSIS (CLASSIFICATION)

Classification Accuracy



- The accuracy of all tested models was the same (about 0.83)
- All used algorithms perform practically the same

Confusion Matrix



- Confusion matrix is the same for all four algorithms.
- We can distinguish between the different outcomes, but the major problem is false positives.
- Logistic regression model can predict the outcome (land vs did not land) but also give the probability of this happening.
- Best hyperparameters for LogReg model are C = 1.0, gamma = 0.0316, kernel = 'sigmoid'.

CONCLUSION

- Records of Falcon 9 rocket launches done by SpaceX were collected by using SpaceX API and web scraping.

- Over the years SpaceX managed to increase successful landings of the rocket first stage.

- After exploratory data analysis, four different classification models were build.

- Logistic regression model can not only predict the outcome (land vs did not land), but also infer the probability of this happening.

APPENDIX

All notebooks are stored in the GitHub repository:

https://github.com/KatjaTak/coursera_data_science_capstone_project