IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rupak Ghawghawe

11/04/2022

Outline

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

Executive Summary

- Summary of methodologies
  - The following methodologies were used to analyze data:
    - ➢ Data Collection using web scraping and SpaceX API
    - ➢ Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics
    - ➢ Machine Learning Prediction
- Summary of all results
    - ➢ It was possible to collect valuable data from public sources
    - ➢ EDA allowed to identify which features are the best to predict success of launchings
    - ➢ Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

- Project background and context
  - ➢ The objective is to evaluate the viability of the new company Space Y to compete with Space X.
- Problems you want to find answers
  - ➢ The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets
  - ➢ Best place to launch the rockets

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - ➢ Space X API (https://api.spacexdata.com/v4/rockets/)
    - ➢ WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

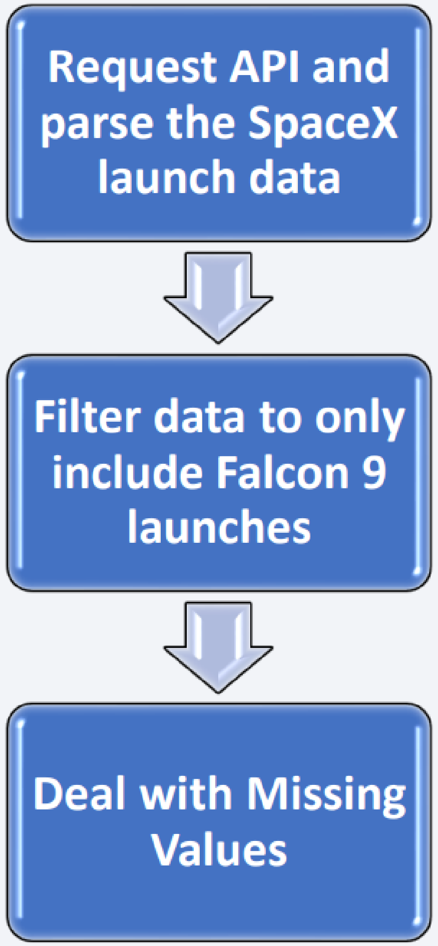
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models. The accuracy of each model evaluated using different combinations of parameters.

Data Collection

- Data sets were collected from
  - Space X API (https://api.spacexdata.com/v4/rockets/) and
  - Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used
- This API was used according to the flowchart beside and then data is persisted.
- **Source Code:** dscproj/SpaceX_Lab.ipynb at master · rupakghawghawe/dscproj (github.com)



Request API and parse the SpaceX launch data

Filter data to only include Falcon 9 launches

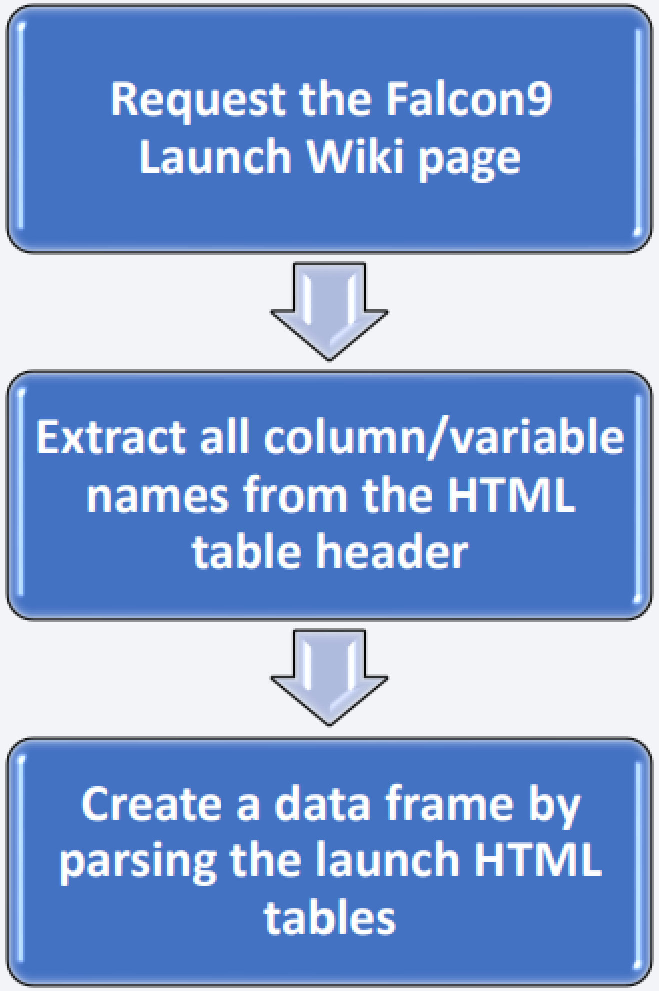
Deal with Missing Values

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia

- Data are downloaded from Wikipedia according to the flowchart and then persisted.

- **SOURCE CODE:** dscproj/Data Collection with Web Scraping.ipynb at master · rupakghawghawe/dscproj (github.com)


Request the Falcon9 Launch Wiki page

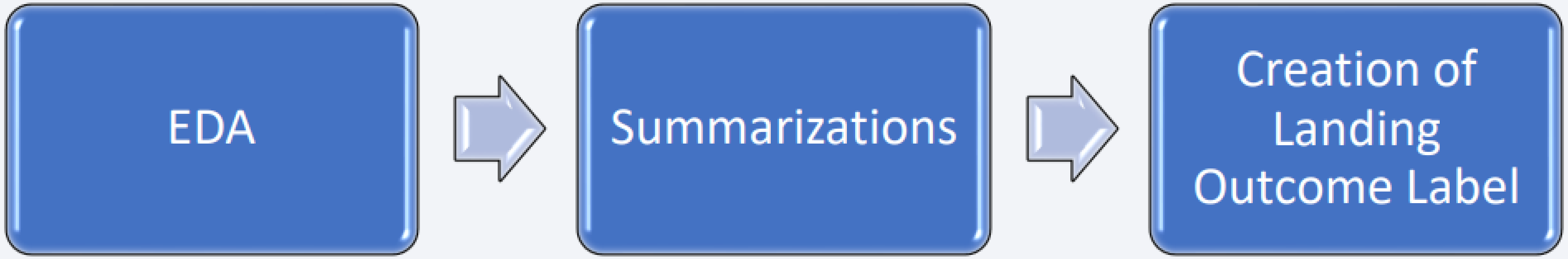
Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.

- Source Code: https://github.com/rupakghawghawe/dscproj/blob/master/Data%20Wrangling.ipynb



EDA

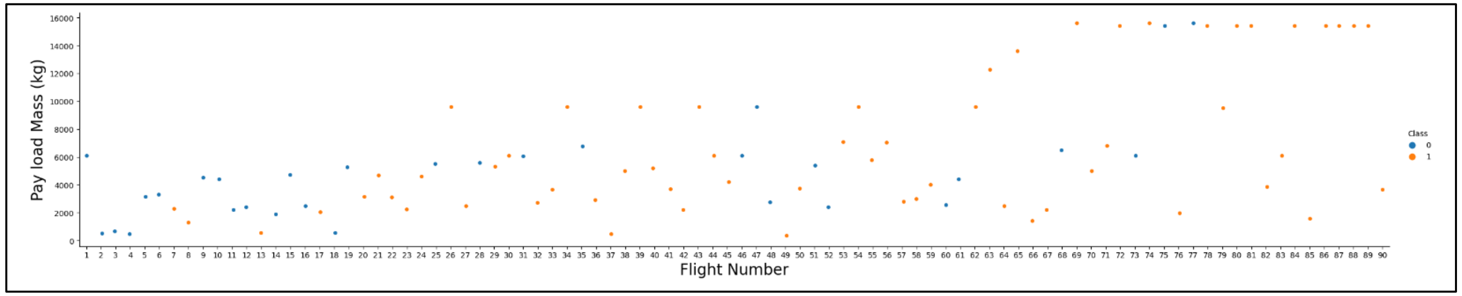
Summarizations

Creation of Landing Outcome Label

EDA with Data Visualization

- To explore data, scatterplots and bar plots were used to visualize the relationship between pair of features:
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

- **Source Code:** dscproj/EDA with Data Visualization.ipynb at master · rupakghawghawe/dscproj (github.com)



EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission
  - Top 5 launch sites whose name begin with the string 'CCA’
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

- Source Code: https://github.com/rupakghawghawe/dscproj/blob/master/EDA_SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site
  - Lines are used to indicate distances between two coordinates.

- Source Code: [dscproj/lab_jupyter_launch_site_location.ipynb at master · rupakghawghawe/dscproj (github.com)]

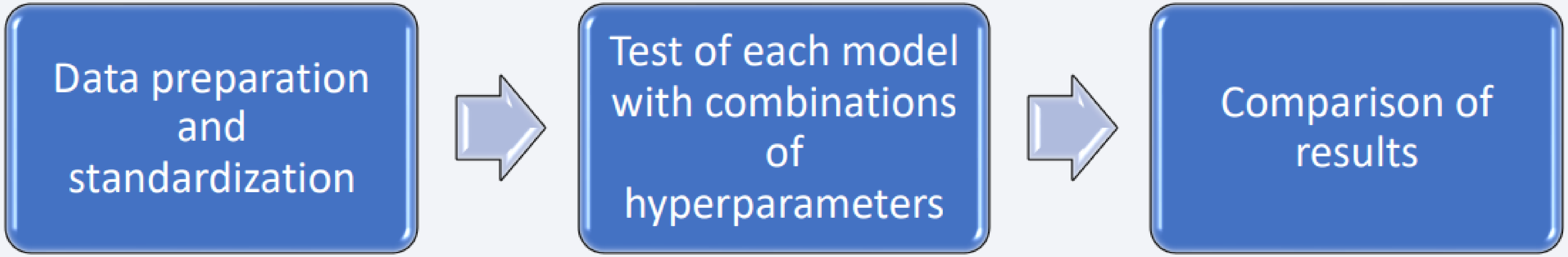
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
  - This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.
- Source Code: dscproj/spacex_dash_app.py at master · rupakghawghawe/dscproj (github.com)

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.

- Source Code: https://github.com/rupakghawghawe/dscproj/blob/master/Machine%20Learning%20Prediction.ipynb



Data preparation and standardization

Test of each model with combinations of hyperparameters

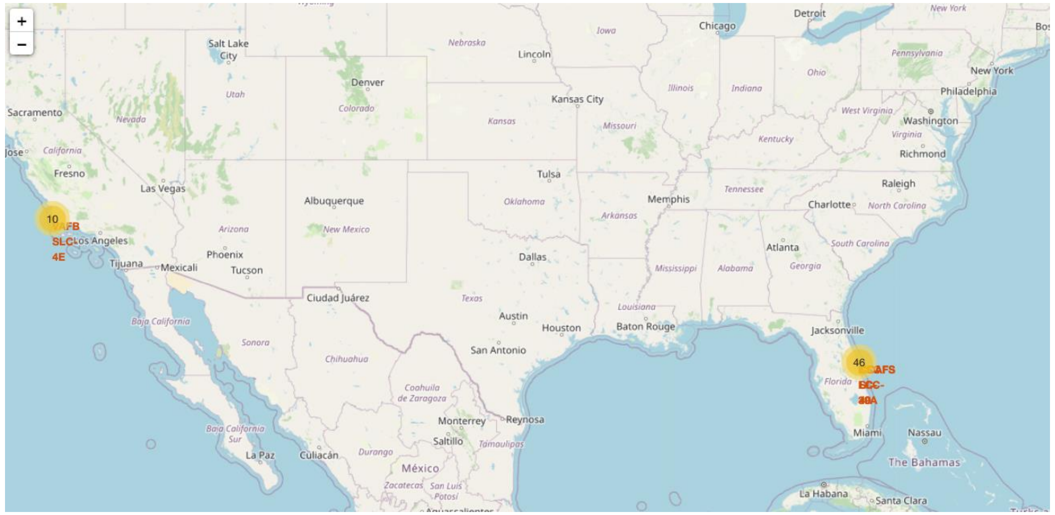
Comparison of results

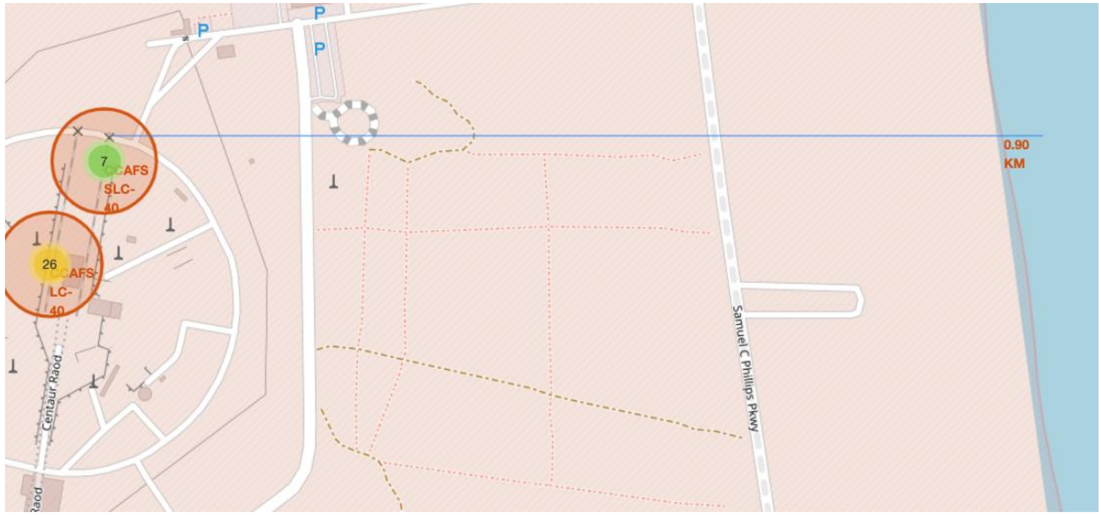
Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites
  - The first launches were done by Space X itself and NASA
  - The average payload of F9 v1.1 booster is 2,928 kg
  - The first successful landing outcome happened in 2015 fiver year after the first launch
  - Many Falcon 9 booster versions were successful at landing drone ships having payload above the average
  - Two booster versions failed at landing drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
  - The number of landing outcomes became better as years passed.

Results

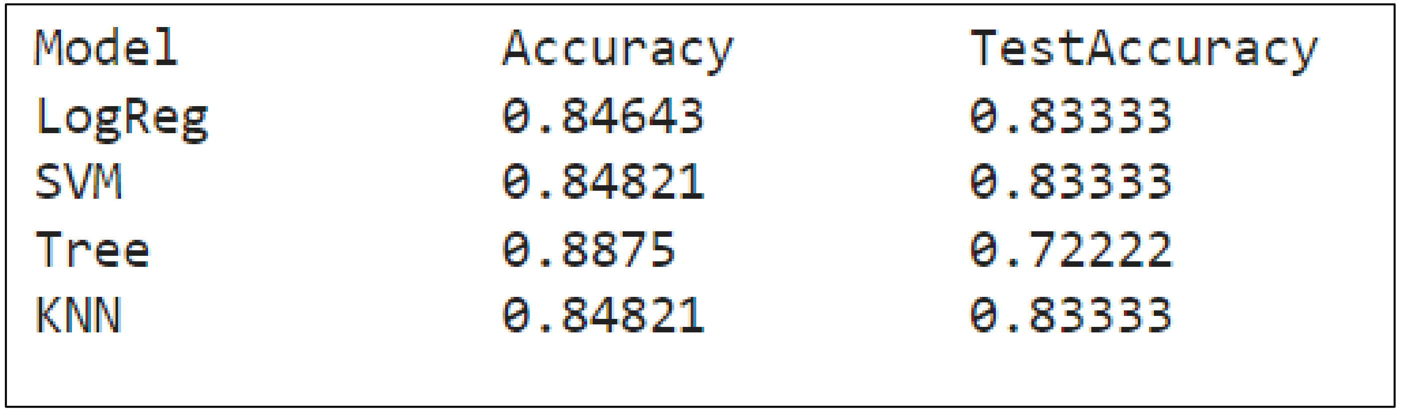
- Interactive analytics demo in screenshots
  - Using interactive analytics, it was possible to identify launch sites in safe places that can be used such as near sea and have a good logistic infrastructure.
  - Most launches happens at east cost launch sites.





Results

- Predictive analysis results
  - Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 89% and accuracy for test data over 72%.

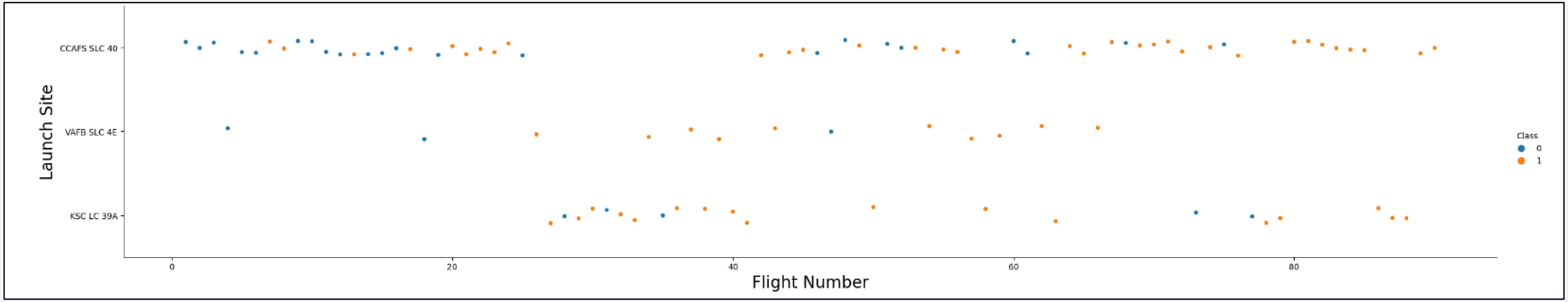


```
Model           Accuracy        TestAccuracy
LogReg          0.84643         0.83333
SVM             0.84821         0.83333
Tree            0.8875          0.72222
KNN             0.84821         0.83333
```

Section 2

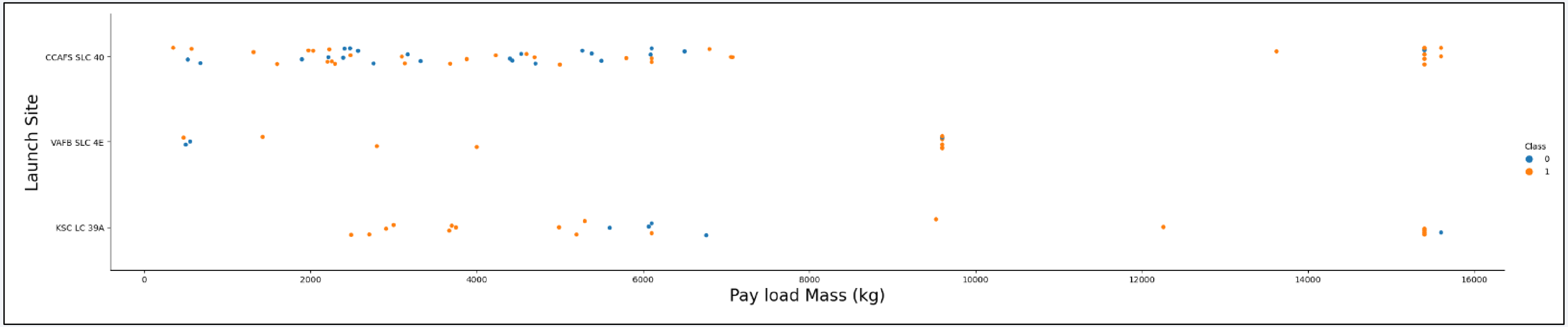
Insights drawn from EDA

Flight Number vs. Launch Site



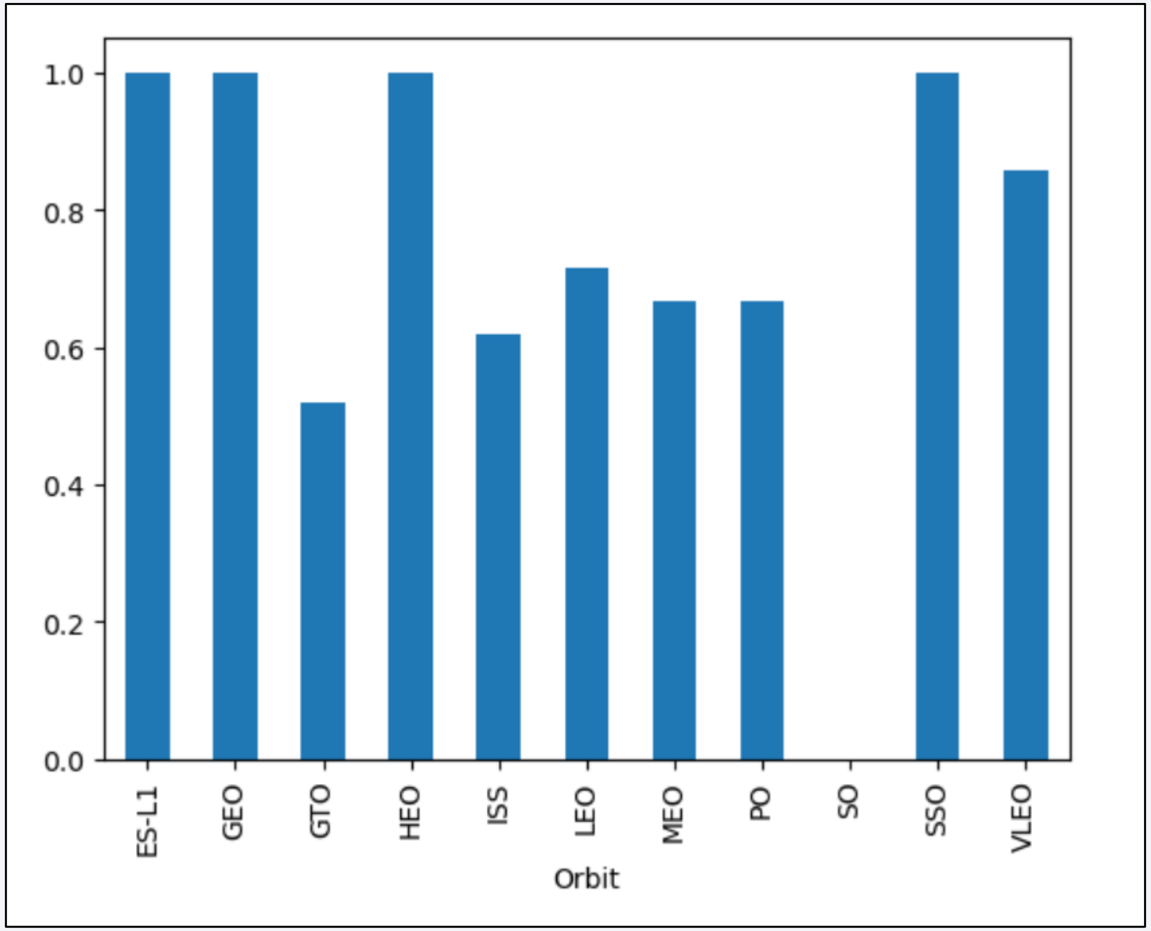
- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful. In second place VAFB SLC 4E and third place KSC LC 39A
- It's also possible to see that the general success rate improved over time.

Payload vs. Launch Site



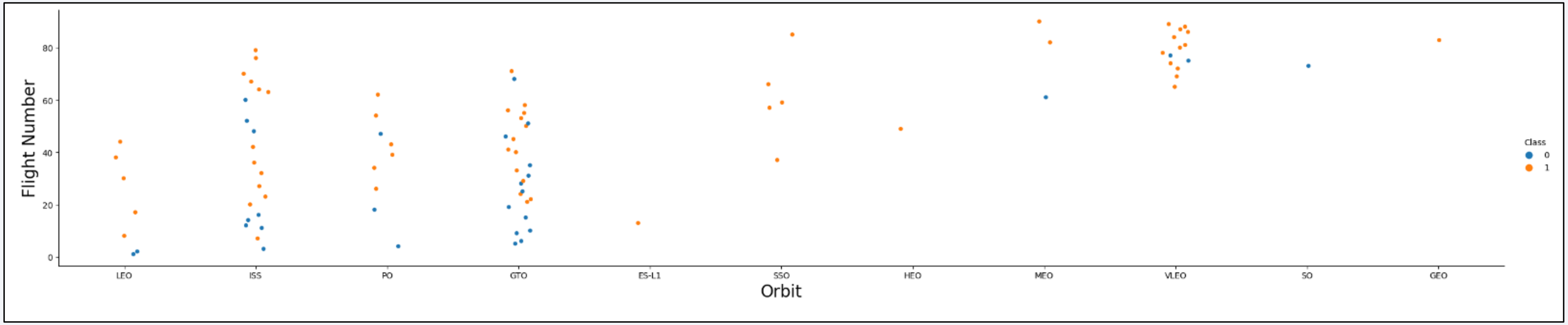
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites

Success Rate vs. Orbit Type



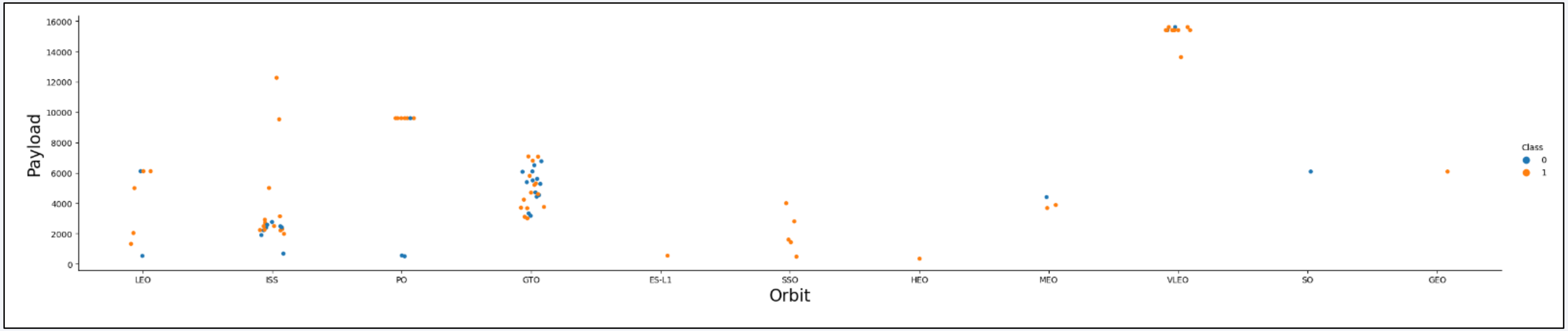
- The biggest success rates is for orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Followed by:
  - VLEO (above 80%)
  - LFO (above 70%)

Flight Number vs. Orbit Type



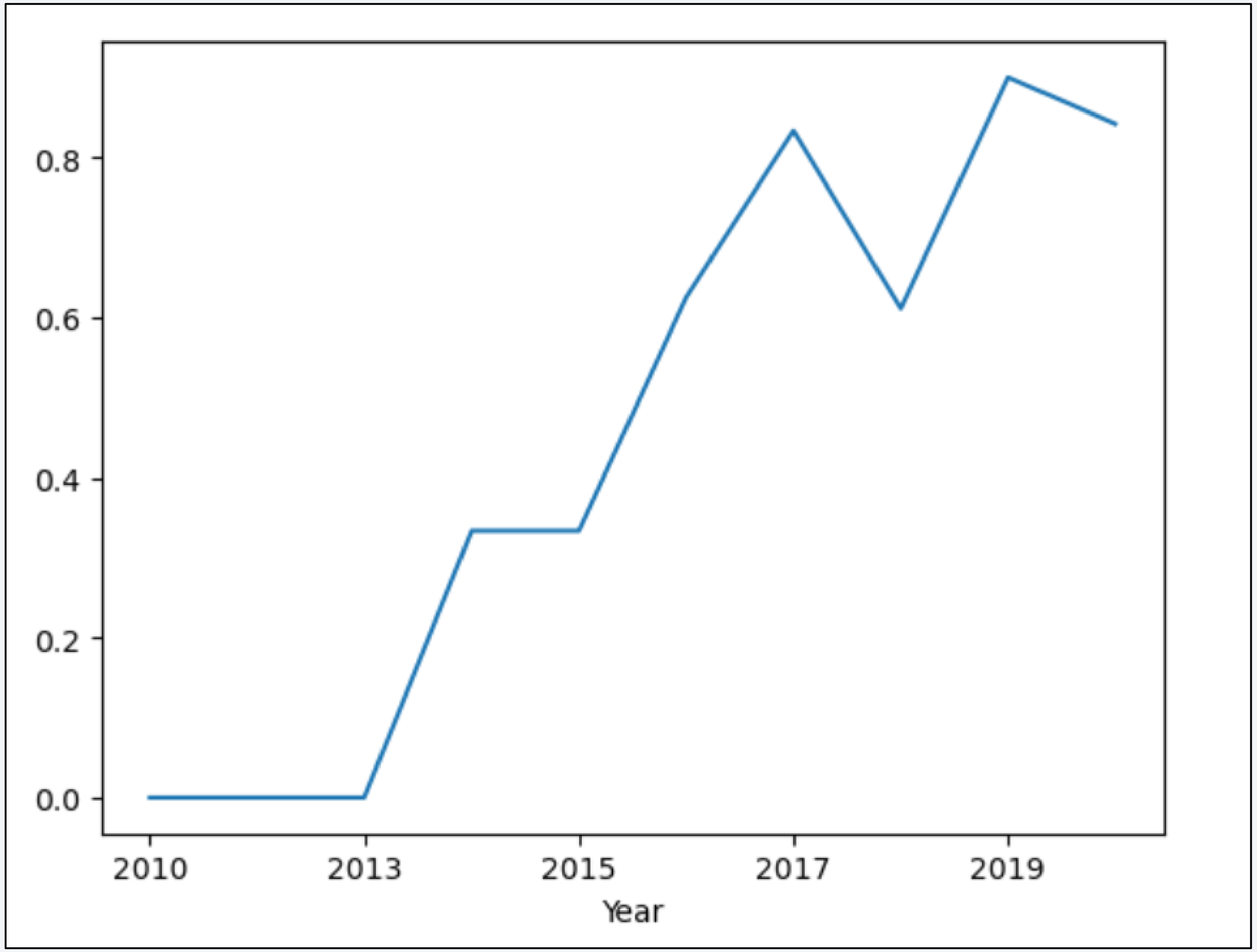
- Success rate improved over time to all orbits
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency

Payload vs. Orbit Type



- There is no relation between payload and success rate to orbit GTO
- ISS orbit has the widest range of payload and a good rate of success
- There are few launches to the orbits SO and GEO.

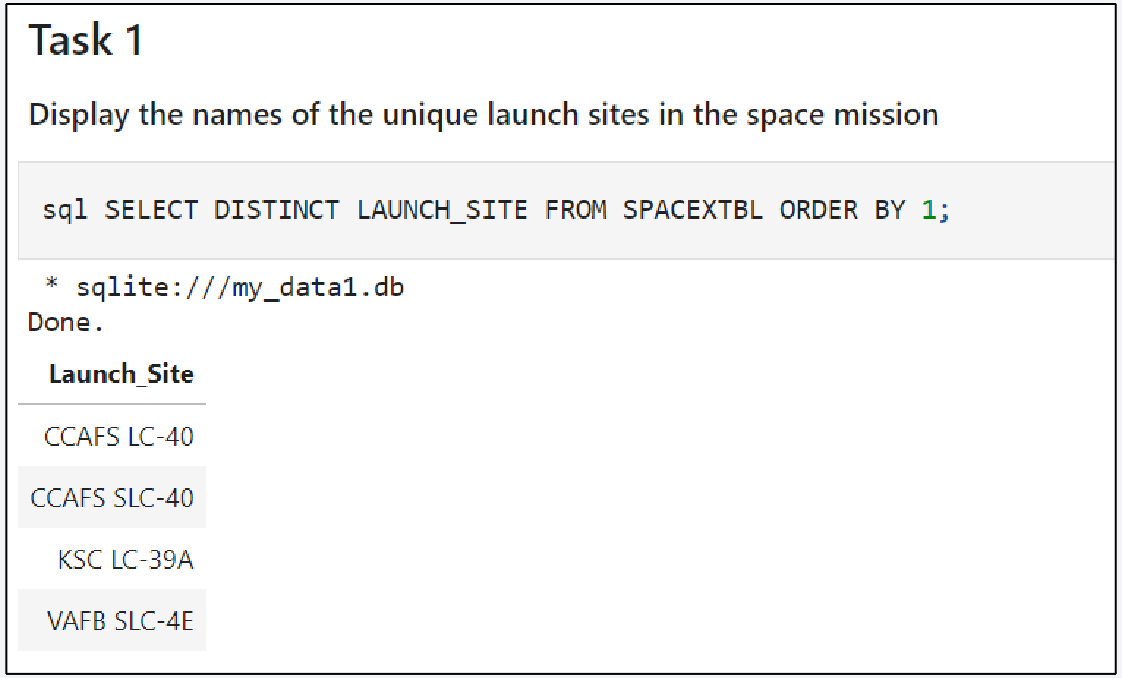
Launch Success Yearly Trend



- Success rate started increasing in 2013 and kept increasing until 2020
- First three years seem to be a period of adjustments and improvement in technology

All Launch Site Names

- According to data, there are four launch sites:



Task 1

Display the names of the unique launch sites in the space mission

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

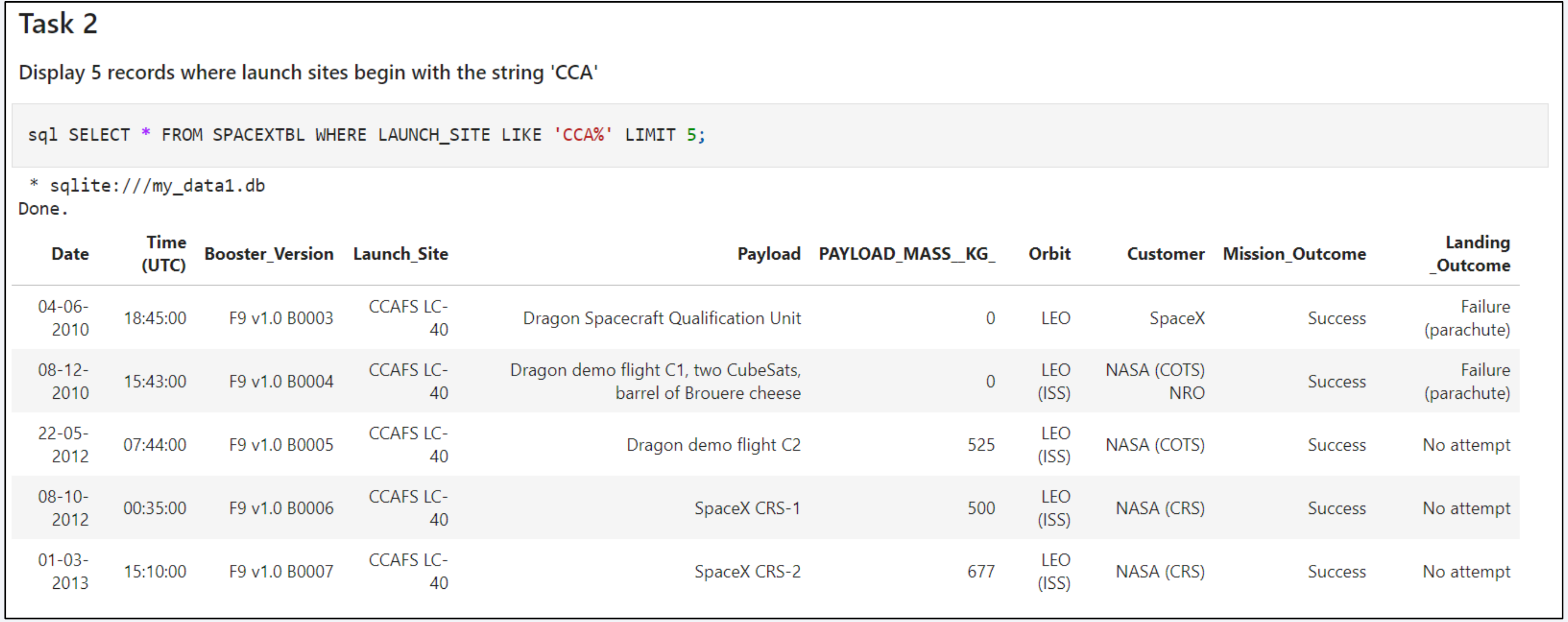
```
 * sqlite:///my_data1.db
Done.
```

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

- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`



Task 2

Display 5 records where launch sites begin with the string 'CCA'

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

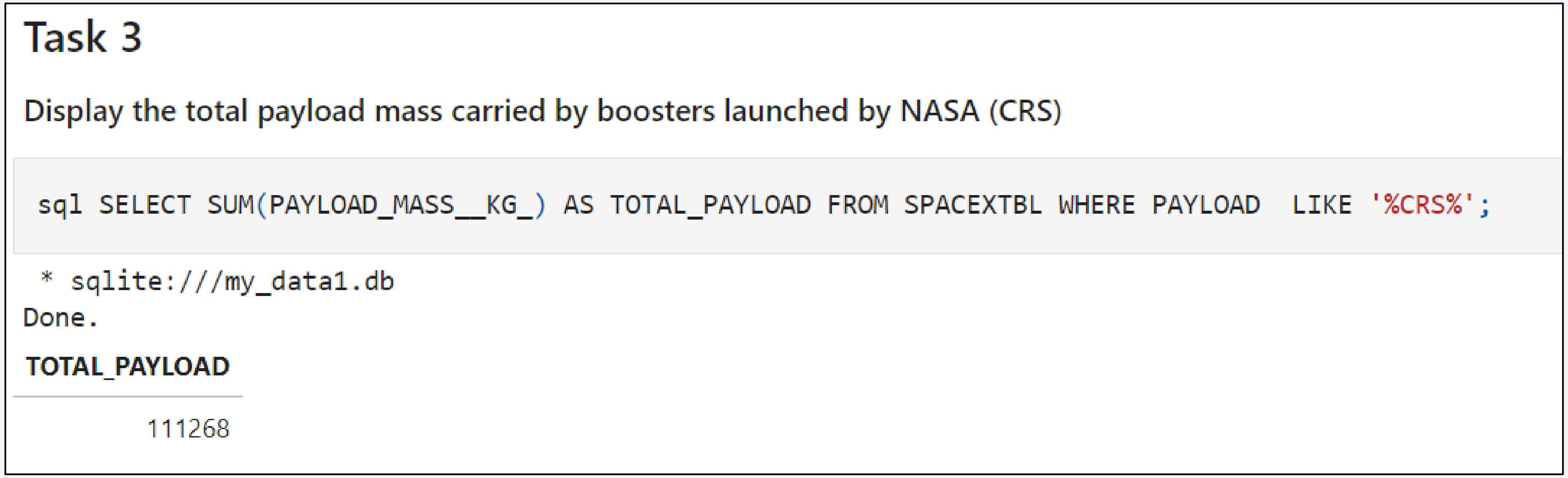
```
 * sqlite:///my_data1.db
Done.
```

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

- In the table, we can see 5 records of launched from Cape Canaveral (CCA)

Total Payload Mass

- The total payload carried by boosters from NASA = 111,268Kg.



Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

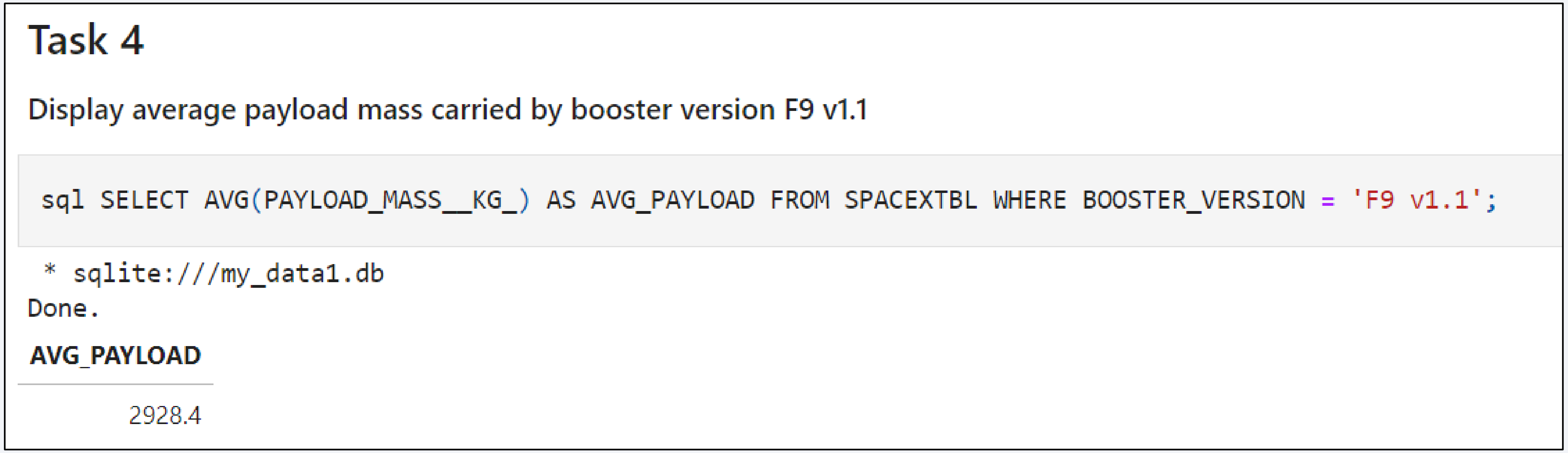
```
 * sqlite:///my_data1.db
Done.
```

TOTAL_PAYLOAD
111268

- This is calculated by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 = 2,928.4Kg
- This result can be obtained by filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928.4 kg.



Task 4

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

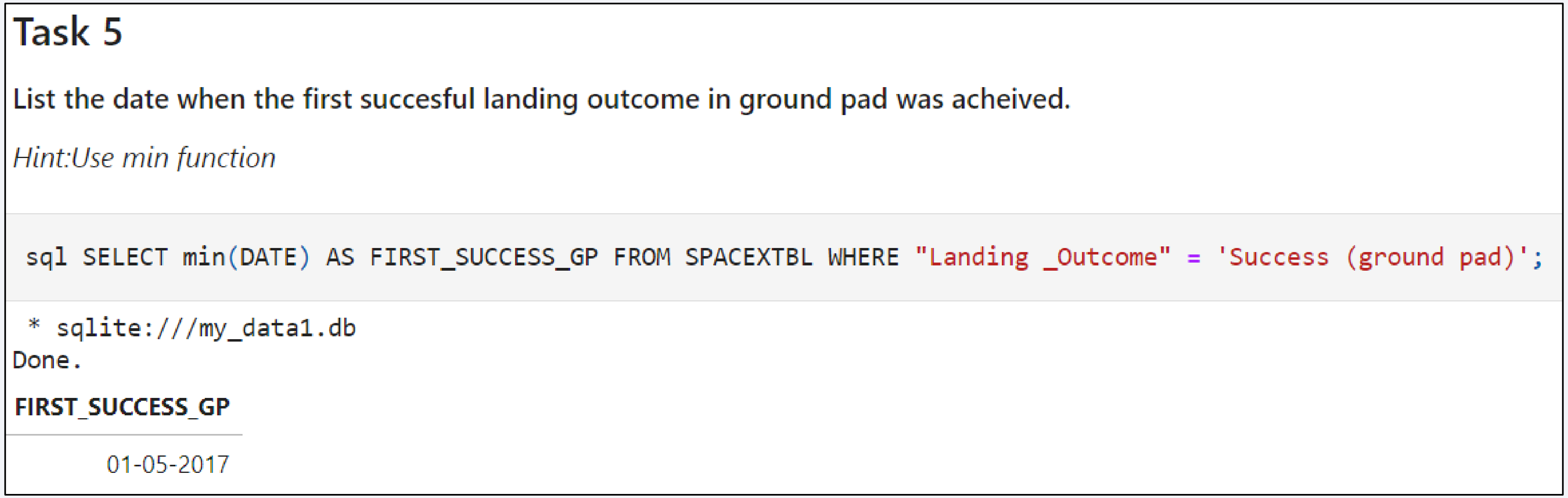
```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad: 01-05-2017
- This result is obtained by filtering data by successful landing outcome on ground pad and getting the minimum value for date



Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

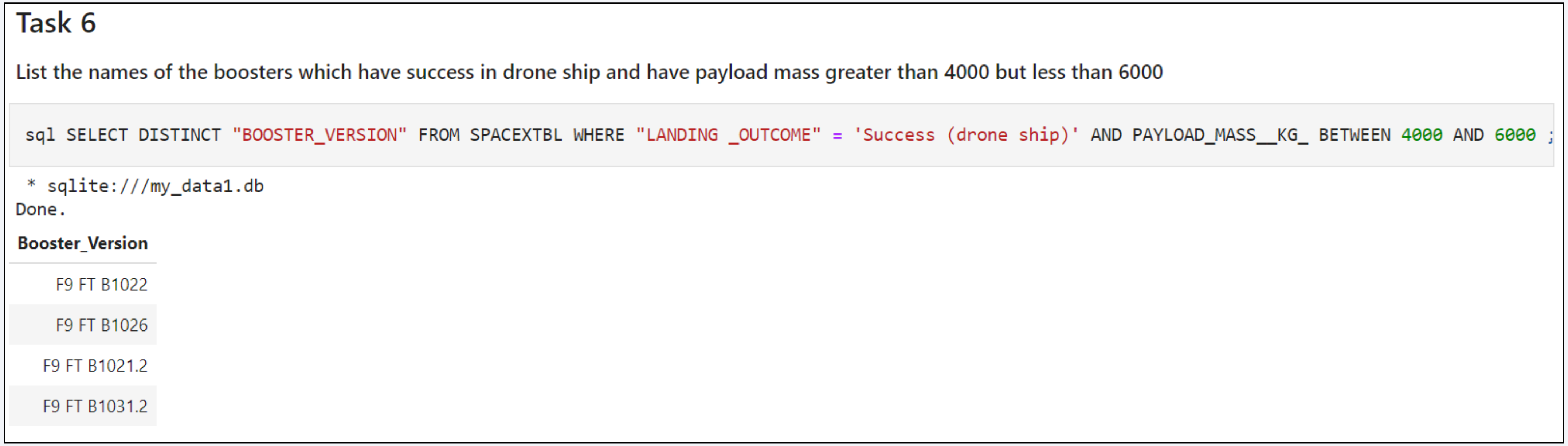
```
sql SELECT min(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)';
```

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

FIRST_SUCCESS_GP
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

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



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

```
sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 ;
```

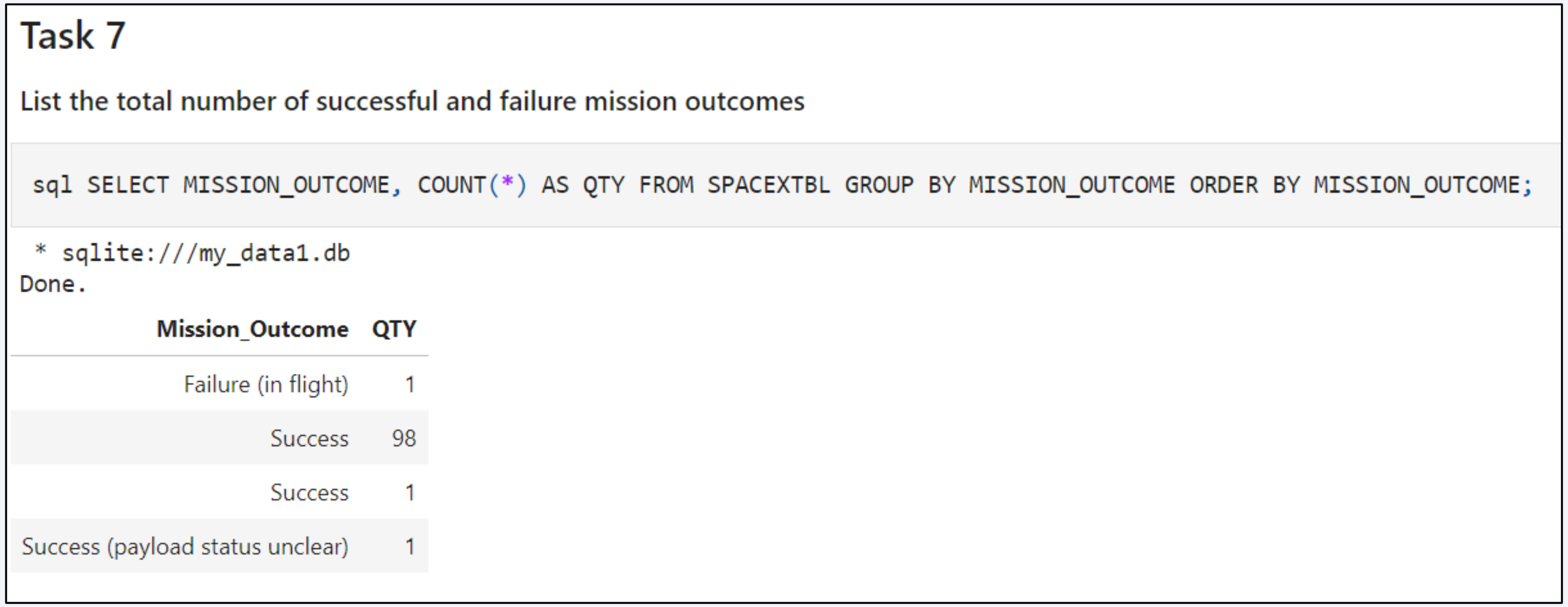
```
 * sqlite:///my_data1.db
Done.
```

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

- This result is obtained by filtering the booster version acc to the filters required

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes



Task 7

List the total number of successful and failure mission outcomes

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

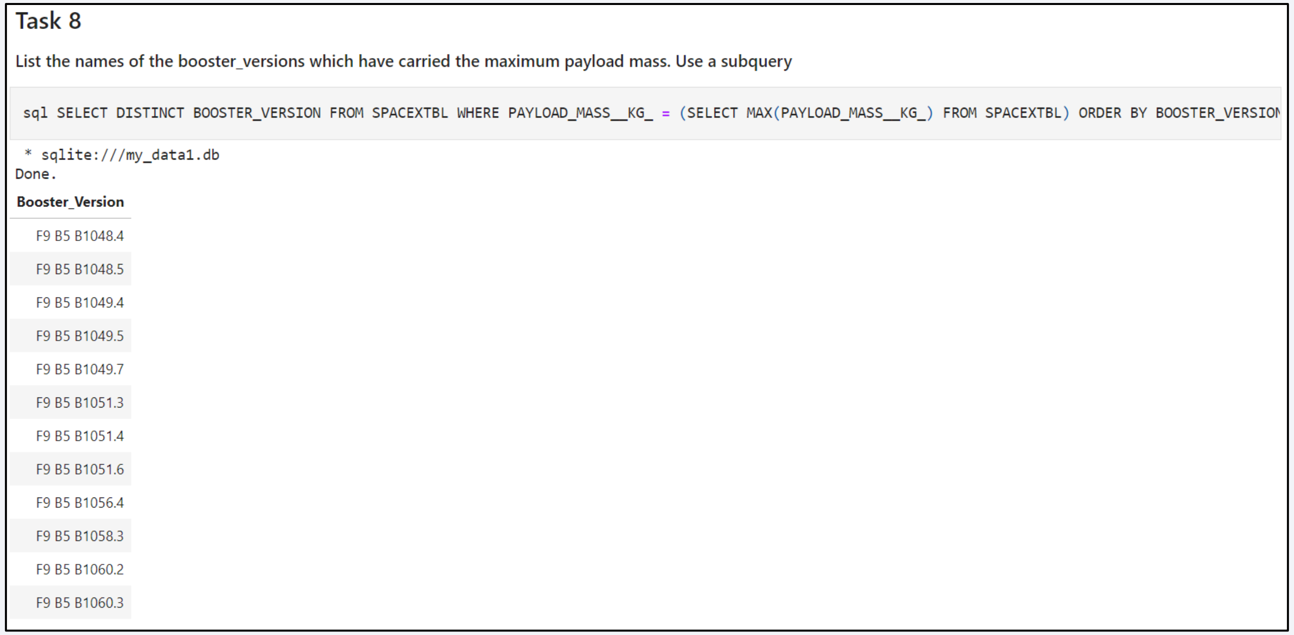
```
 * sqlite:///my_data1.db
Done.
```

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

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass



Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION
```

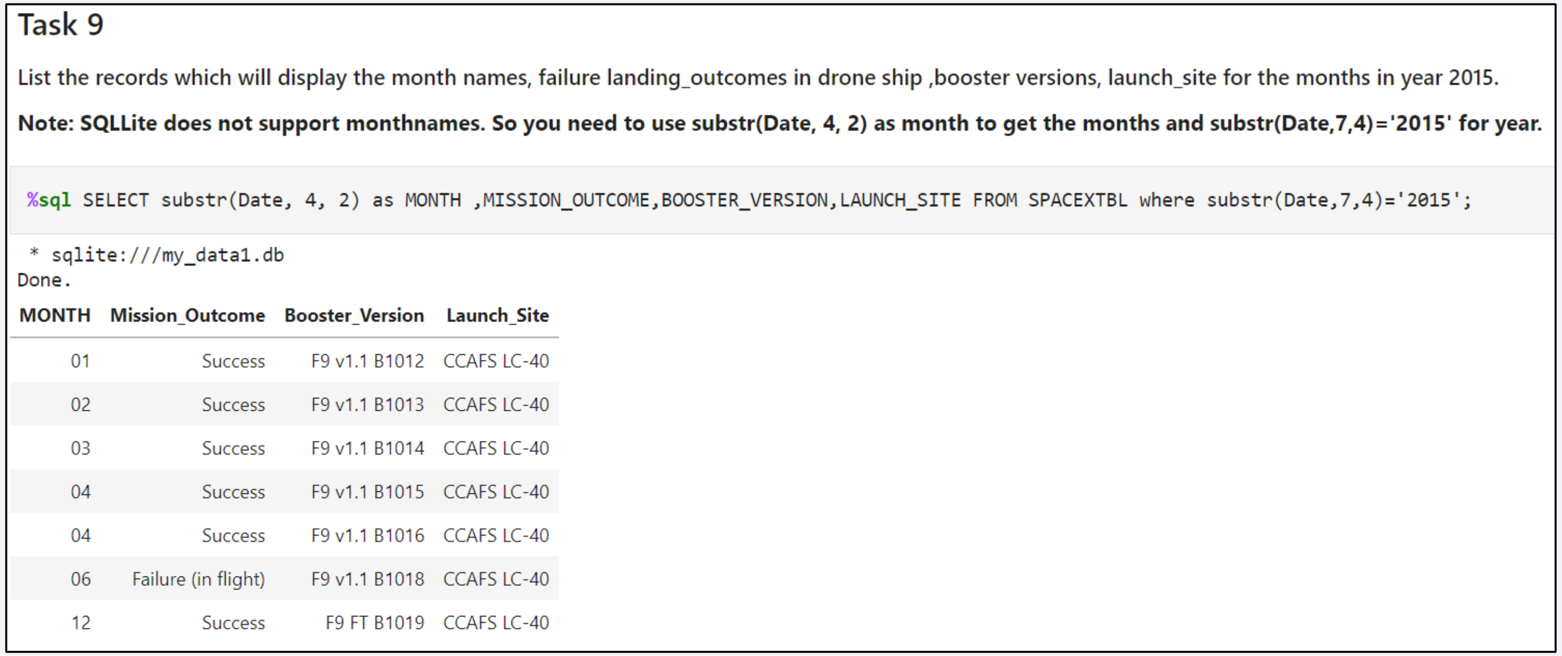
```
 * sqlite:///my_data1.db
Done.
```

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

- Grouping the data by payload mass and using subquery to determine max payload

2015 Launch Records

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



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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql SELECT substr(Date, 4, 2) as MONTH ,MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where substr(Date,7,4)='2015';
```

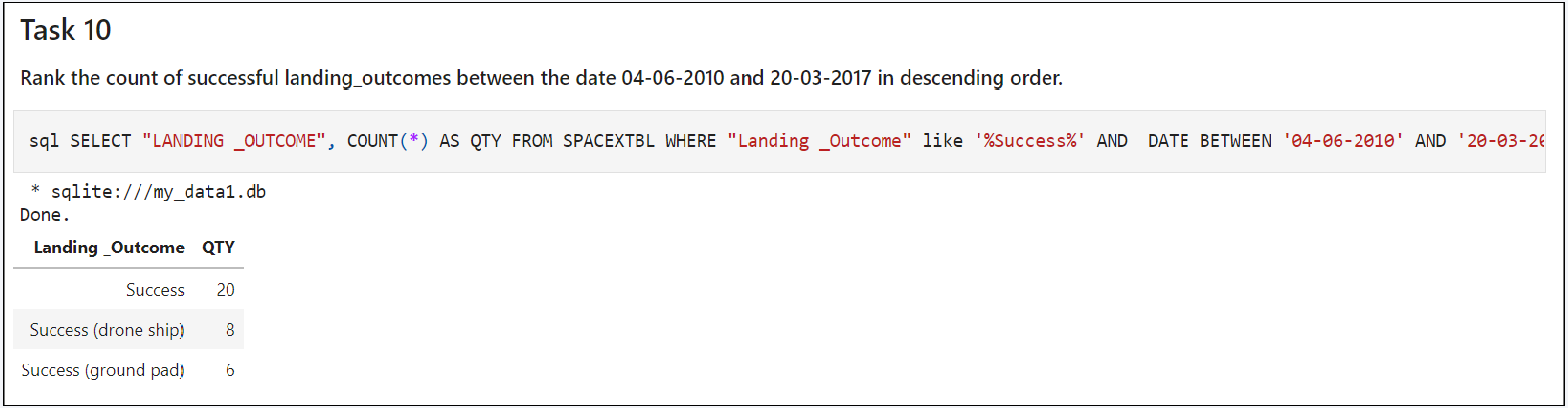
```
 * sqlite:///my_data1.db
Done.
```

MONTH	Mission_Outcome	Booster_Version	Launch_Site
01	Success	F9 v1.1 B1012	CCAFS LC-40
02	Success	F9 v1.1 B1013	CCAFS LC-40
03	Success	F9 v1.1 B1014	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40
04	Success	F9 v1.1 B1016	CCAFS LC-40
06	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

- The query was created as displayed in the picture

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

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



Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
sql SELECT "LANDING _OUTCOME", COUNT(*) AS QTY FROM SPACEXTBL WHERE "Landing _Outcome" like '%Success%' AND  DATE BETWEEN '04-06-2010' AND '20-03-2
```

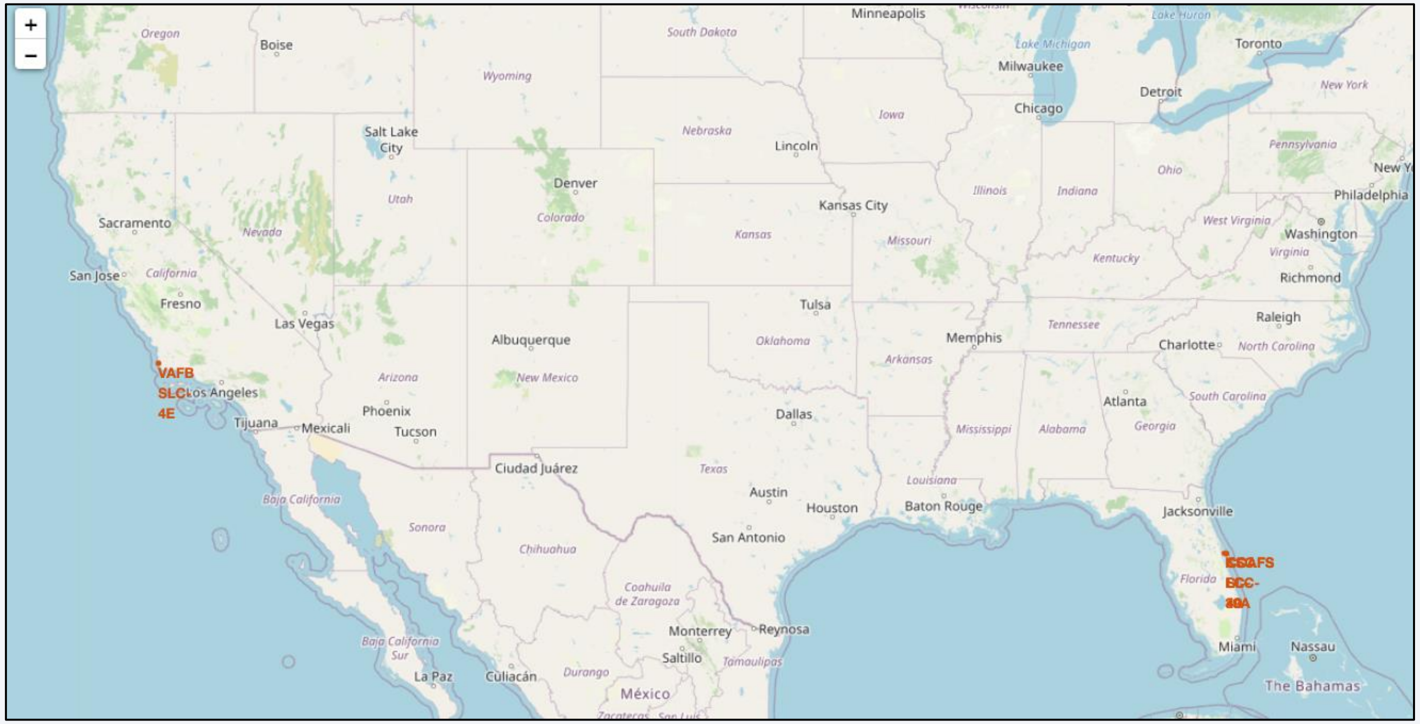
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	QTY
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

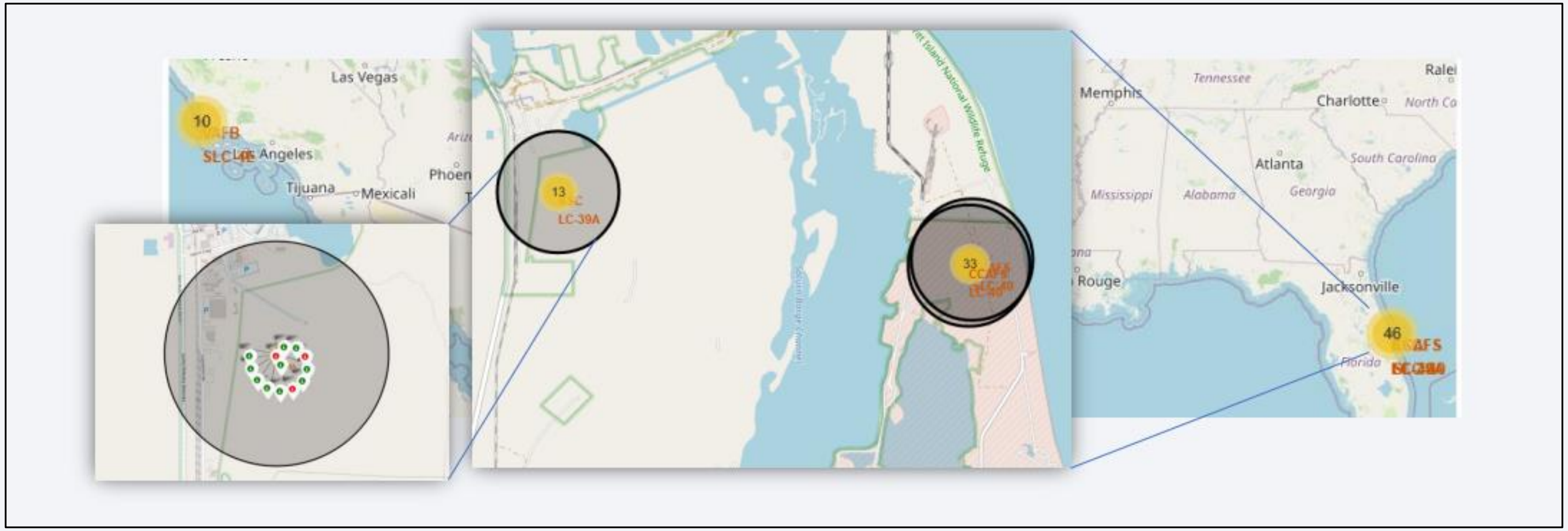
All Launch Sites



- Launch sites are near sea, probably by safety, but not too far from roads and railroads

Launch Outcome by Site

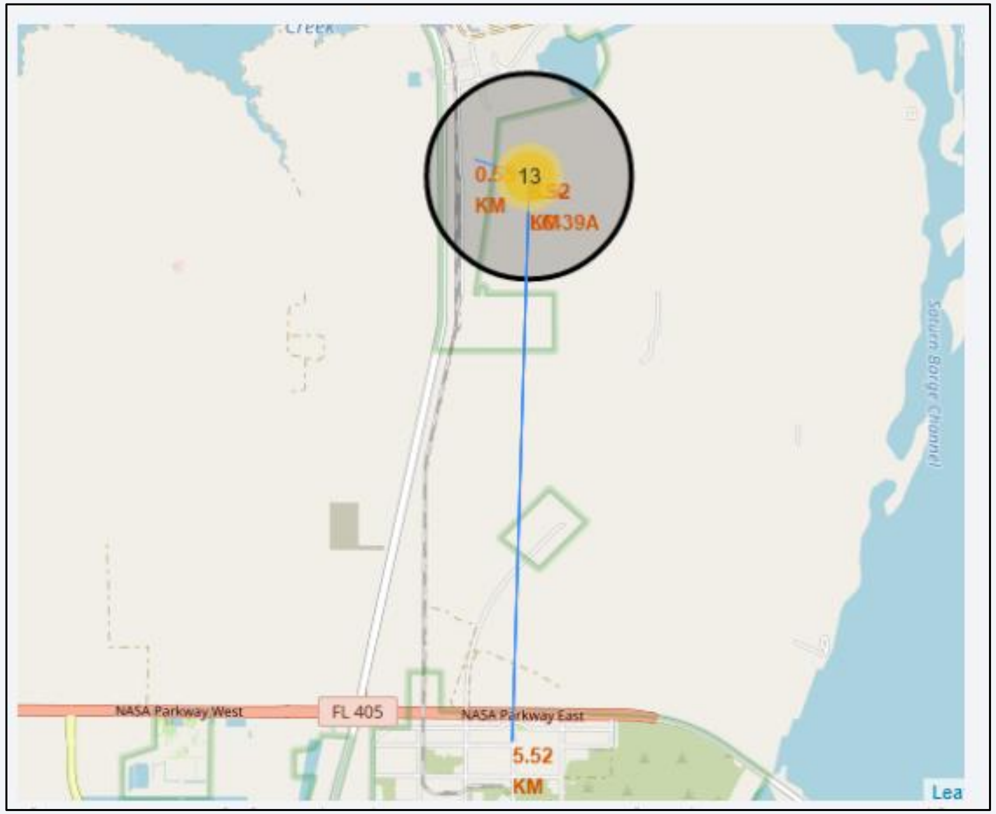
- Example of KSC LC-39A launch site launch outcomes

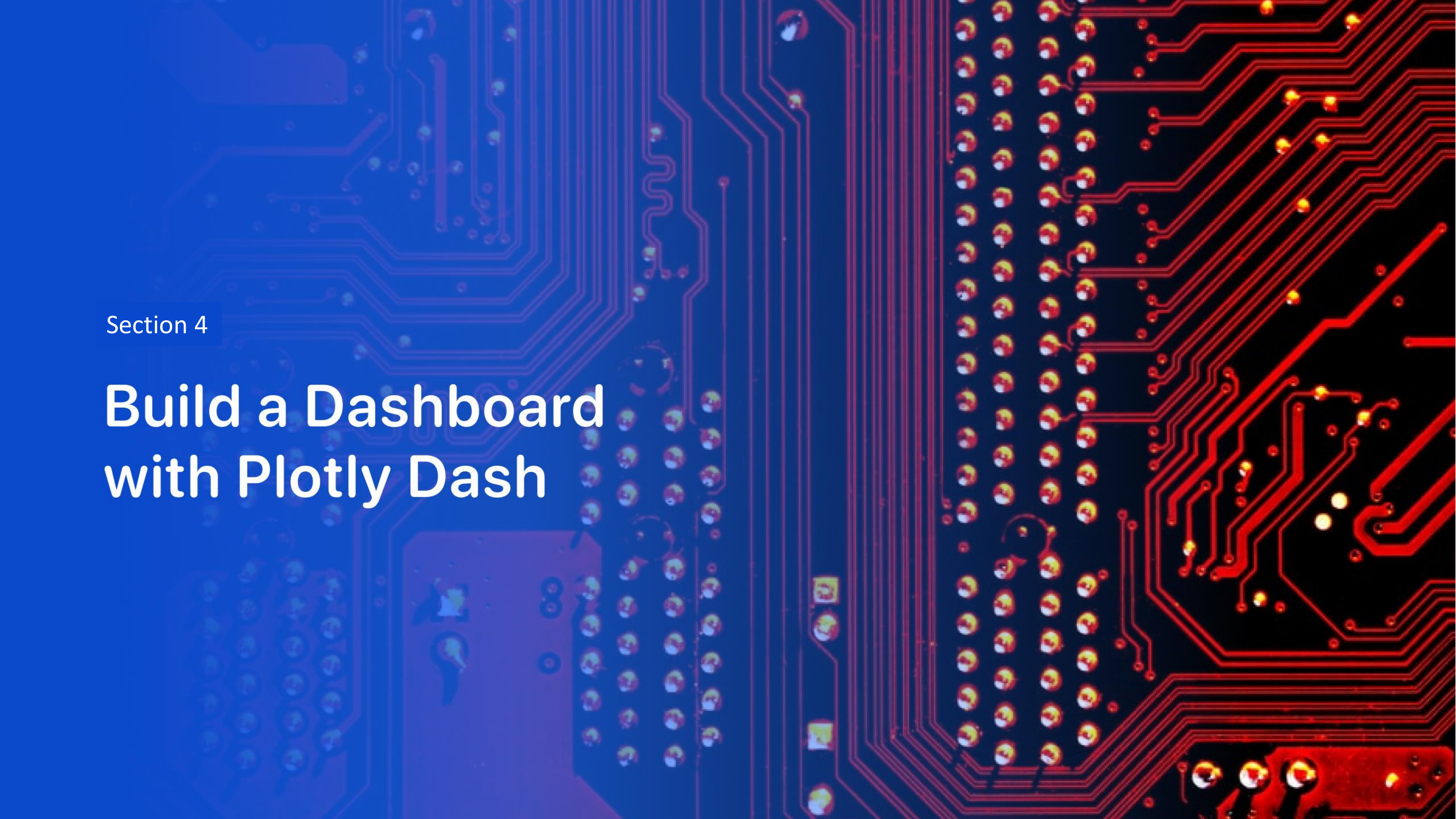


- Green markers indicate successful and red ones indicate failure.

Logistics and Safety

- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

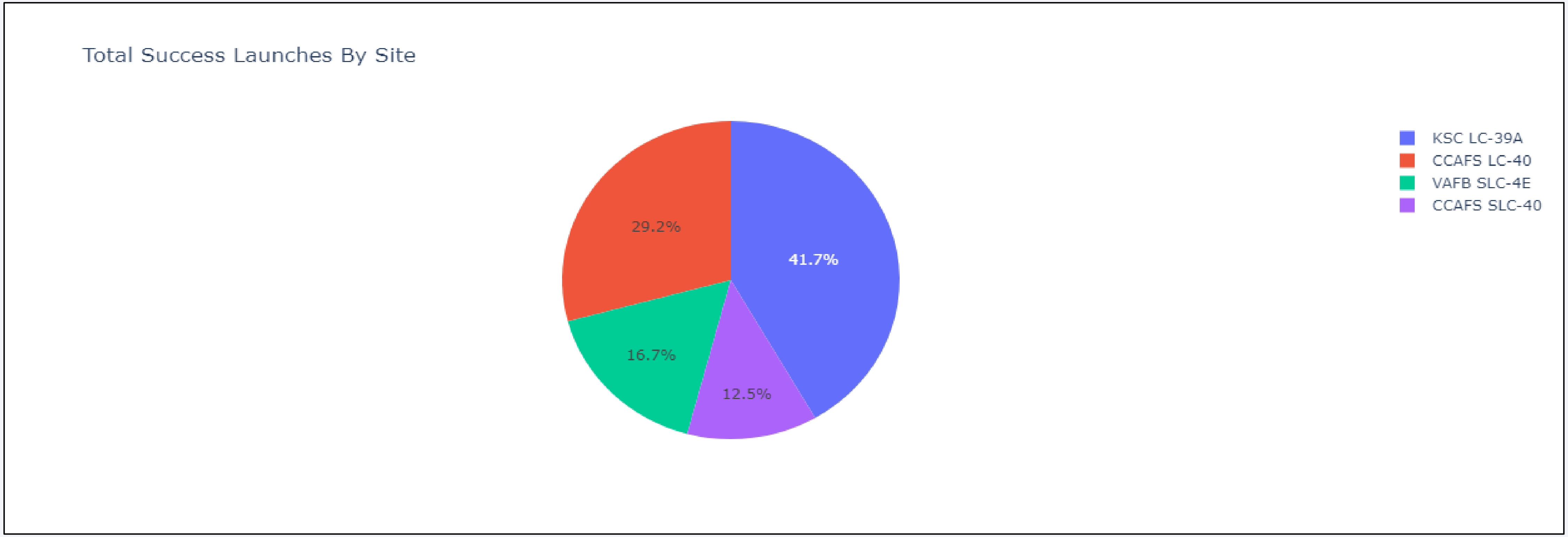




Section 4

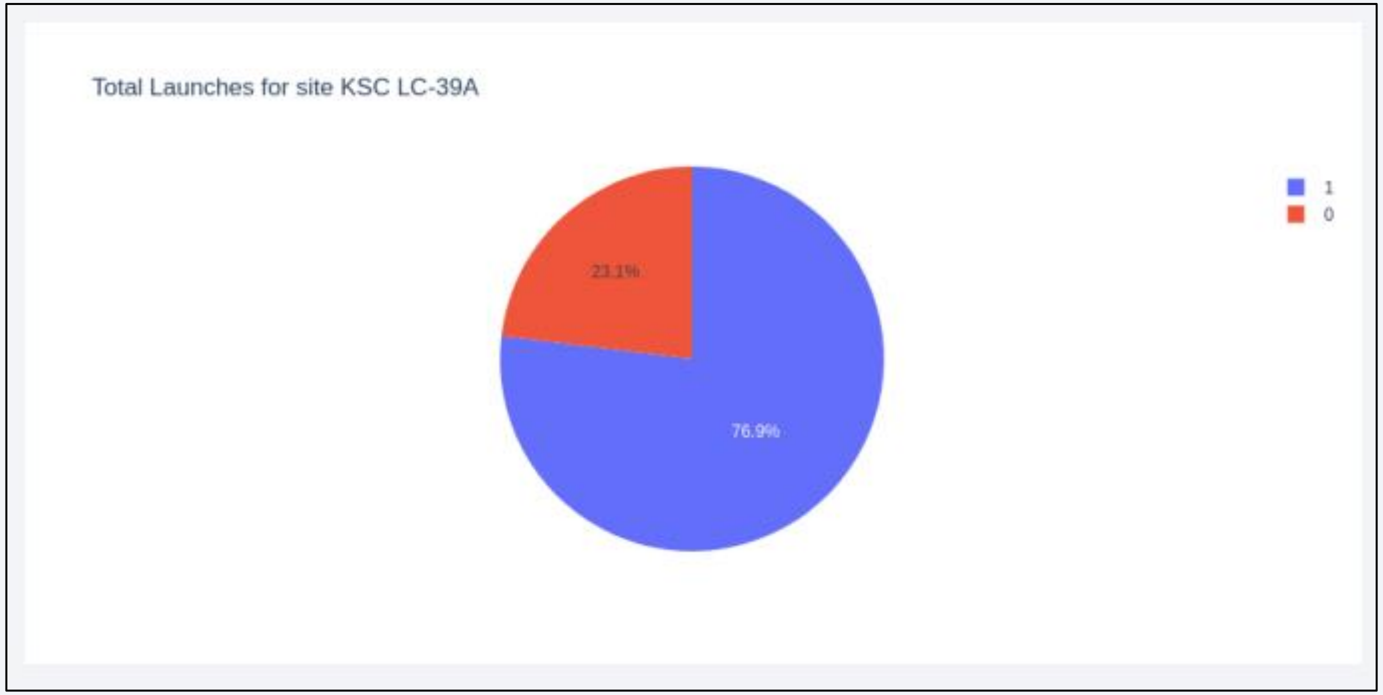
Build a Dashboard with Plotly Dash

Successful Launches by Site



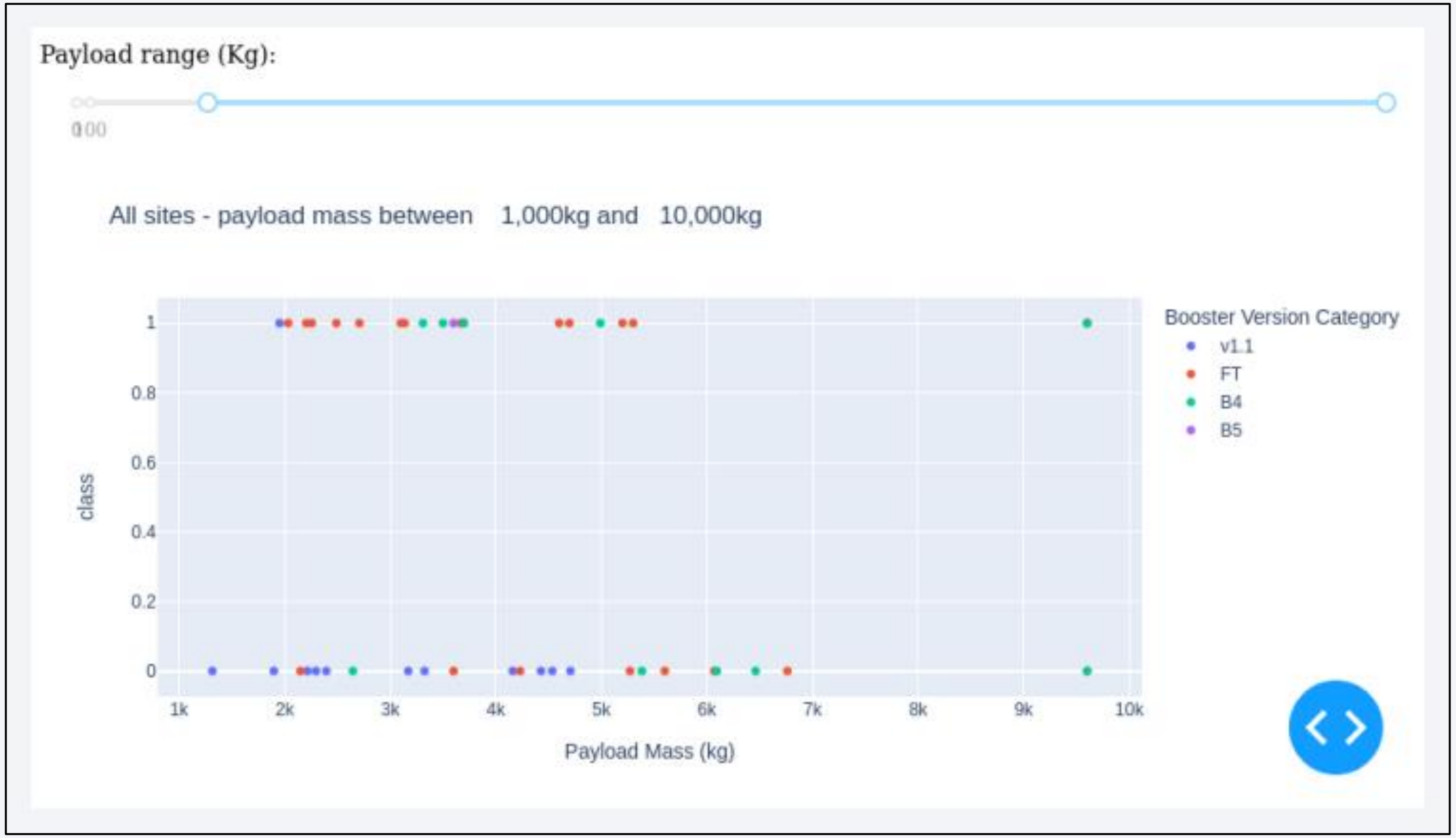
- The place of launch is a critical factor in the success of launch with KSC LC – 39A site having highest success %

Launch Success Ratio for KSC LC-39A



- 76.9% success rate at KSC LC-39A

Payload vs. Launch Outcome



Payload range (Kg):

- Payloads under 6,000kg and FT boosters are the most successful combination.

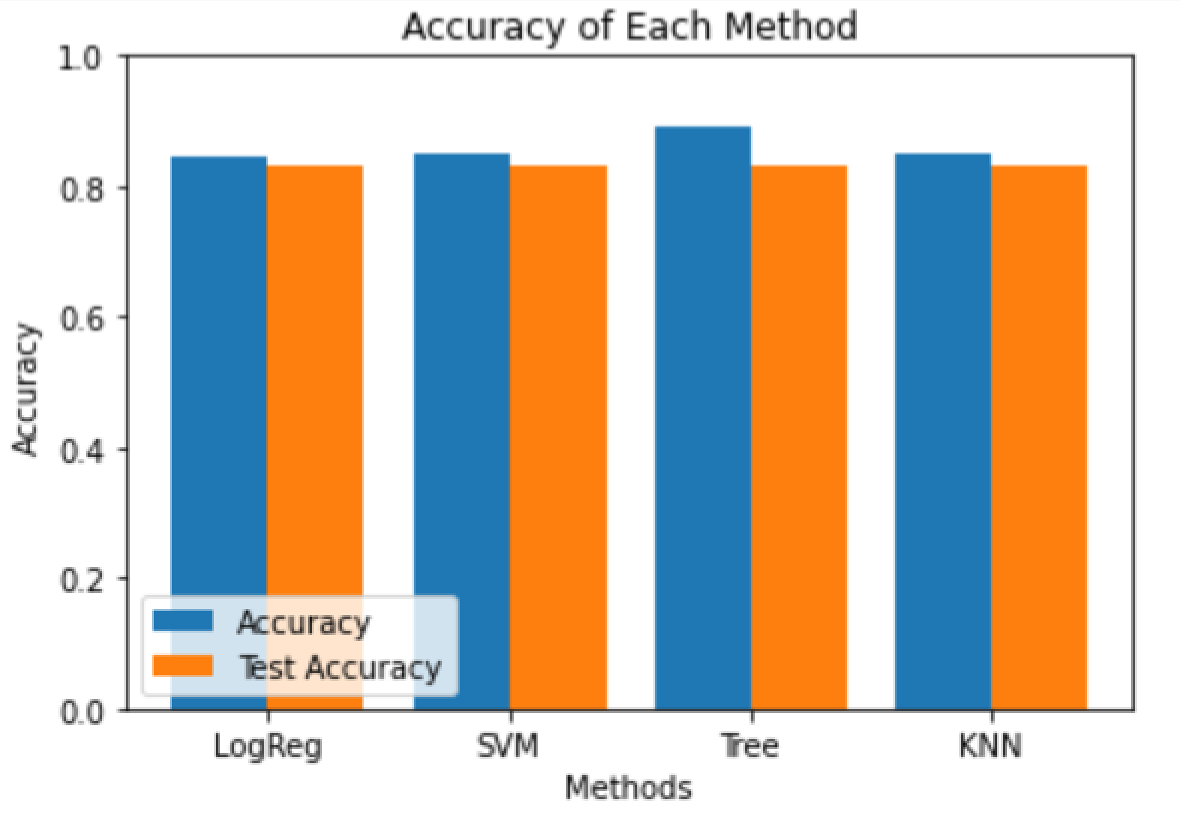
Section 5

Predictive Analysis (Classification)

Classification Accuracy

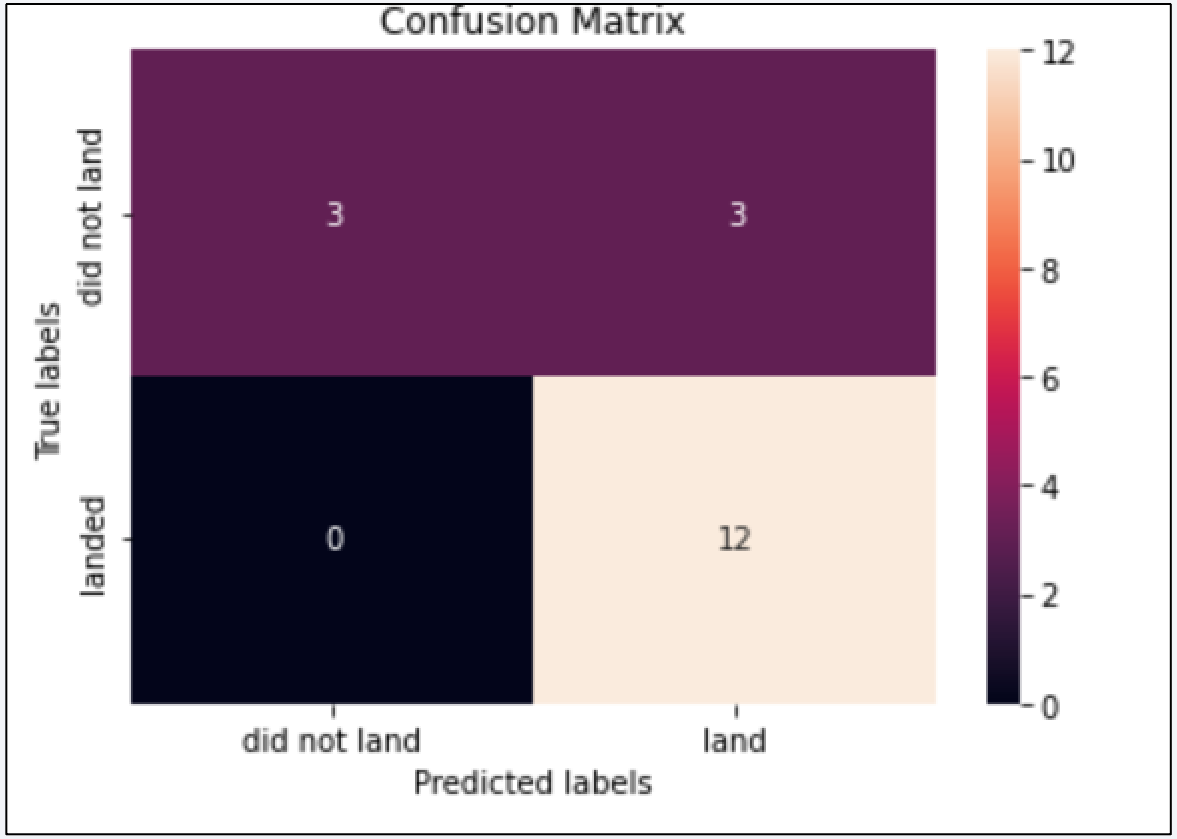
- Visualize the built model accuracy for all built classification models, in a bar chart

- Decision tree has the highest accuracy with prediction accuracy of 88.9% and test accuracy of 83.3%



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.



Conclusions

- Different data sources were analyzed and conclusions were refined along the process
- The best launch site is KSC LC-39A
- Launches above 7,000kg are less risky
- Although most of mission outcomes are successful; successful landing outcomes seem to improve over time, according the evolution of processes and rockets
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- GitHub link for all code:
- rupakghawghawe/dscproj at master (github.com)

Thank you!