IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

D. Knecht

March 2024

Outline

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

Executive Summary

- Data Science methodologies were applied to analyze the success and failure criteria of Space X launches for commercial space travel.
- Data was gathered via API and web scraping, prepared as part of data wrangling and subsequently analyzed during an exploratory data analysis with SQL and hands-on visualizations.
- For your reference, findings have been visualized using Folium, too.
- Intermediate results have then been tested using machine learning algorithms and predictive analysis with train and test data. Various statistical methods were applied to elicit the most accurate predictions.
- As a result, it was found that success rate varies depending on launch site and payload mass.

Introduction

- Project background and context
  - Space X aspires to offer “affordable” space travel for private clients by applying re-usable technology.
  - The cost of a Space X launch depends heavily on whether or not such rocket components can be successfully recaptured, thus a rocket launch must be successful in the first place.
  - This project aims to predict the success rate of Space X rocket launches against certain variables such as launch site, payload and others.
- Problems you want to find answers
  - As per the above, the goal is to determine the key variable for a successful Space X launch.
  - The analysis shall depict the impact and co-dependencies of these variables using statistical methods.
  - Eventually, the results of the analysis shall provide the cornerstones for successful Space X launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API: https://api.spacexdata.com/v4/* (separate for rockets, payloads and launchsites)
  - Web scraping from Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Performed data wrangling
  - Collected data was enriched with landing outcome label based on outcome data
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models

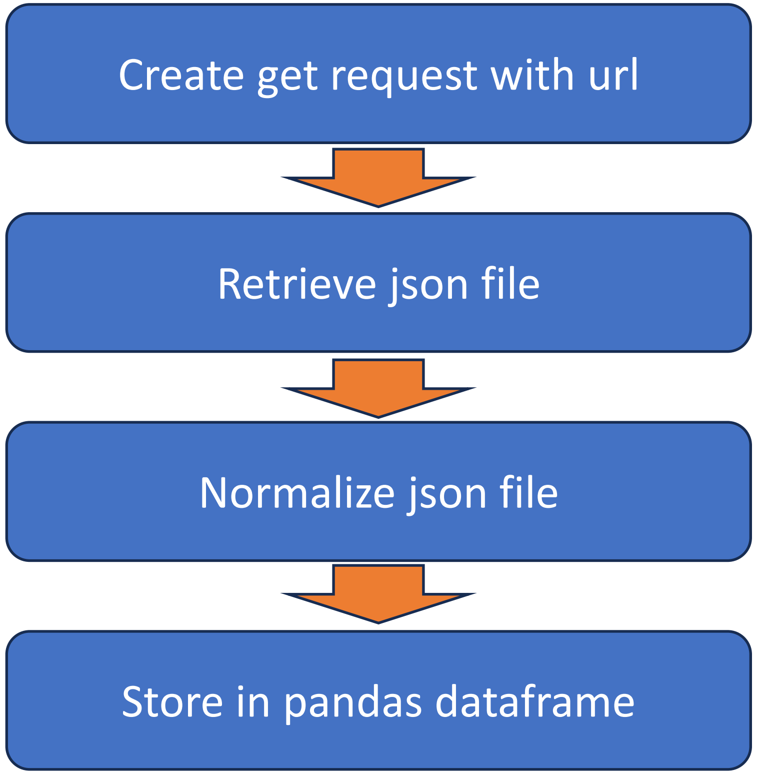
Data Collection

- Collection of data from two sources:
  - SpaceX API: https://api.spacexdata.com/v4/* (separate for rockets, payloads and launchsites)
  - Web scraping from Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- Process flow:
  - Use of GET request to obtain data from Space X API
    - Decoding of response to read .json() and restructure it to suit pandas dataframe with .json_normalize() command
    - Cleansing of data for missing or erroneous values
  - Performed web scraping from Wikipedia Falcon 9 launch records using BeautifulSoup library
    - Obtain launch records as HTML table to parse table and convert in separate pandas dataframe

Data Collection – SpaceX API

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose



Create get request with url

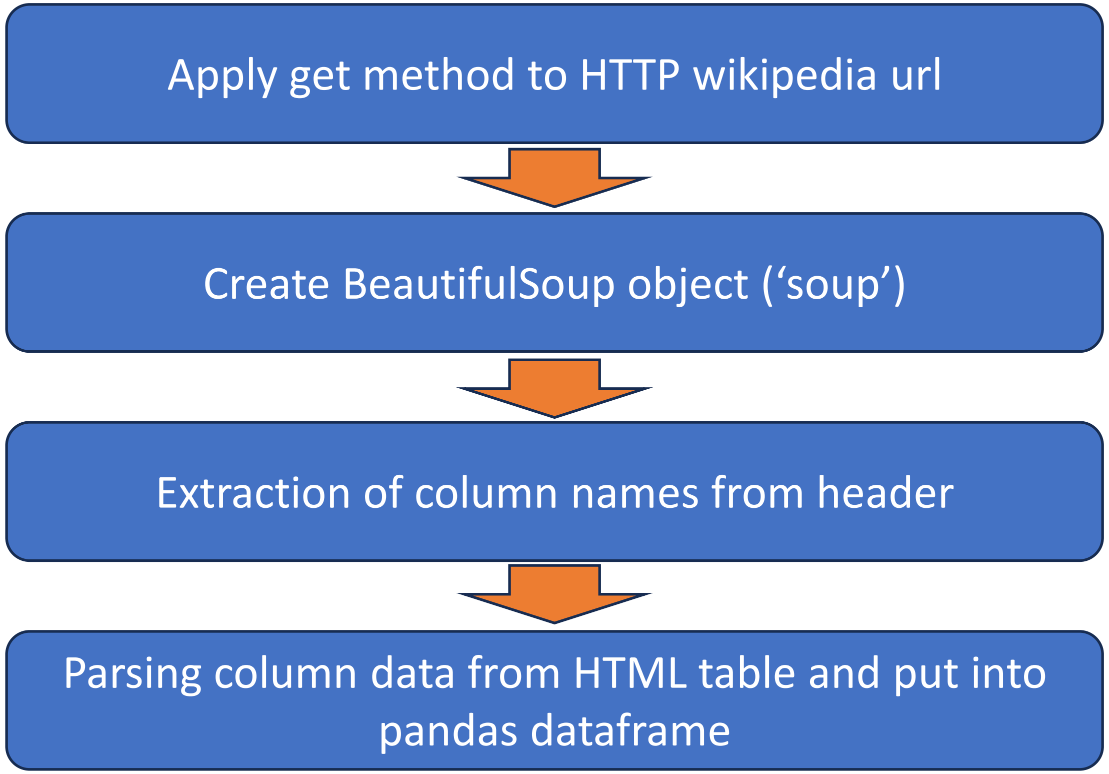
Retrieve json file

Normalize json file

Store in pandas dataframe

Data Collection - Scraping

- GitHub URL: https://github.com/knechtd/IBM-Data-Science/blob/main/01_jupyter-labs-spacex-data-collection-api.ipynb



Apply get method to HTTP wikipedia url

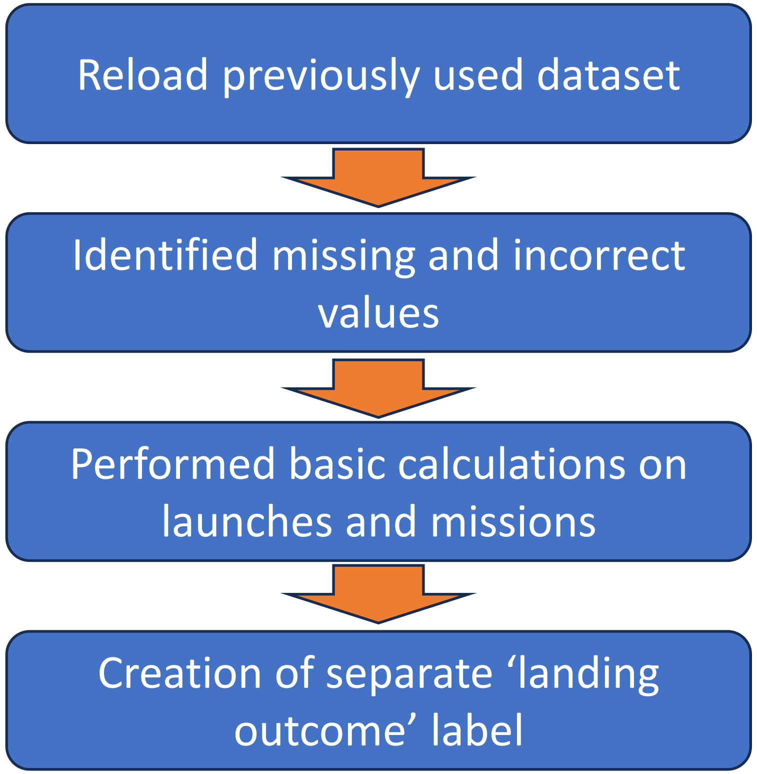
Create BeautifulSoup object (‘soup’)

Extraction of column names from header

Parsing column data from HTML table and put into pandas dataframe

Data Wrangling

- Loading the data from the previous exercise, the exercise first assessed the quality of the data to check whether any corrections had to be performed.
- Following some initial calculations and analyses, an additional variable ‘landing outcome’ was added for further use in future analyses
- GitHub URL: https://github.com/knechtd/IBM-Data-Science/blob/main/03_Data%20Wrangling.ipynb



Reload previously used dataset

Identified missing and incorrect values

Performed basic calculations on launches and missions

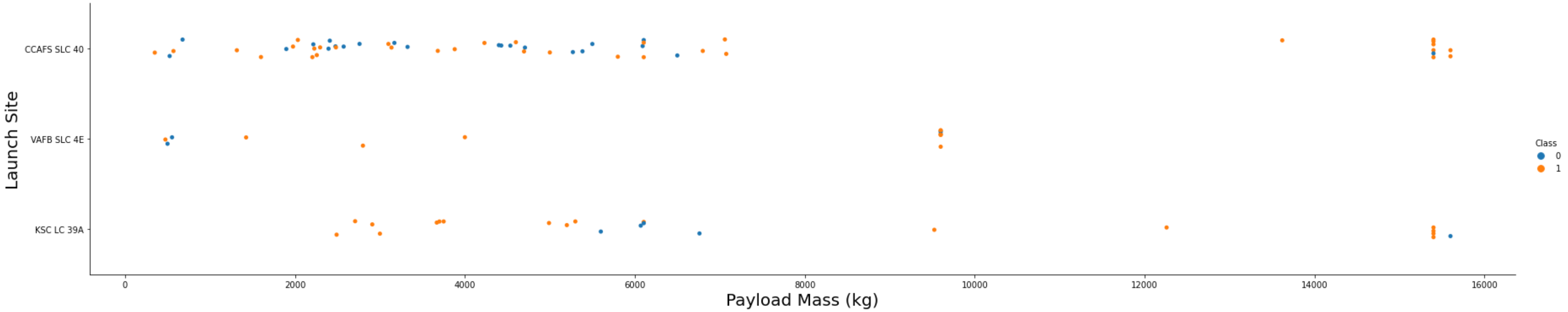
Creation of separate ‘landing outcome’ label

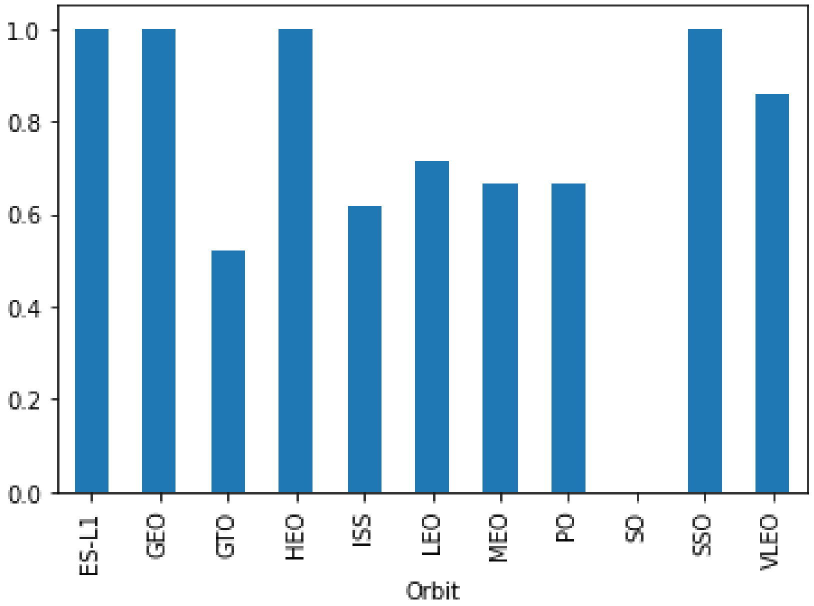
EDA with Data Visualization (1/2)

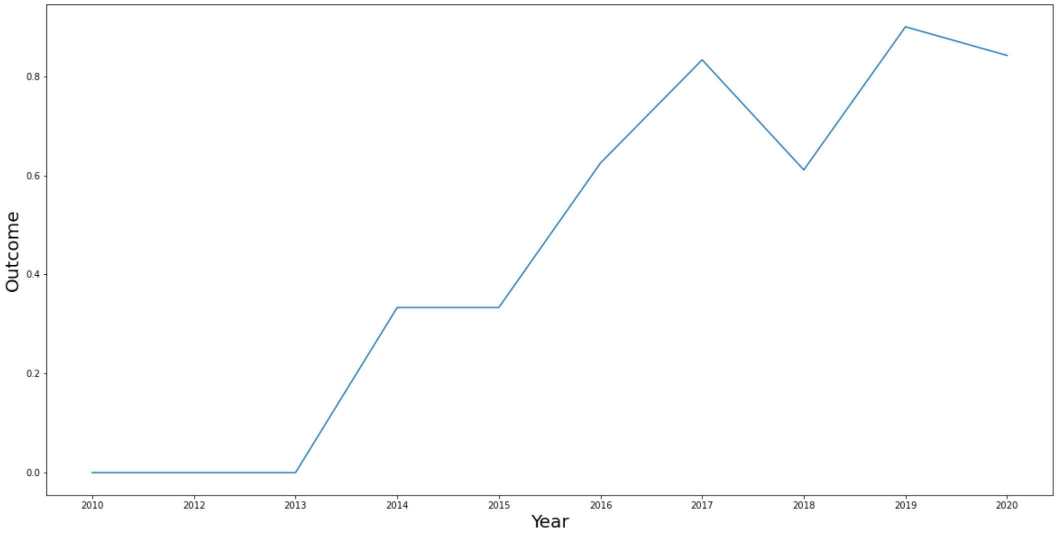
- Used charts (see next page for examples of graphs)
  - Scatter plots
    - Used to complete various comparisons such as launch sites vs. flight numbers, orbits vs. flight numbers etc.
  - Bar charts
    - Categorical visualization of orbits used by Space X to apply in analysis of success rate
  - Line charts
    - Allowed for comparisons over time to show increase of success rate after 2013 until 2020

- GitHub URL: https://github.com/knechtd/IBM-Data-Science/blob/main/05_Visualization%20Lab.ipynb

EDA with Data Visualization (2/2)







EDA with SQL

- SQL queries performed as part of EDA:
  - Distinct names of launch sites
  - Top 5 sites beginning with 'CCA’
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date of first successful landing on ground pad
  - Names of successful boosters with a respective payload mass between 4000 and 6000 kg
  - Count of successful and failed missions
  - Names of the booster version which have carried the maximum payload mass
  - Failed landing outcomes in drone ship, booster versions, and launch site names for 2015
  - Count of landing outcomes between the June 2010 and March 2017
- GitHub URL: https://github.com/knechtd/IBM-Data-Science/blob/main/04_jupyter-labs-Hands-on%20Lab%20Complete%20the%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

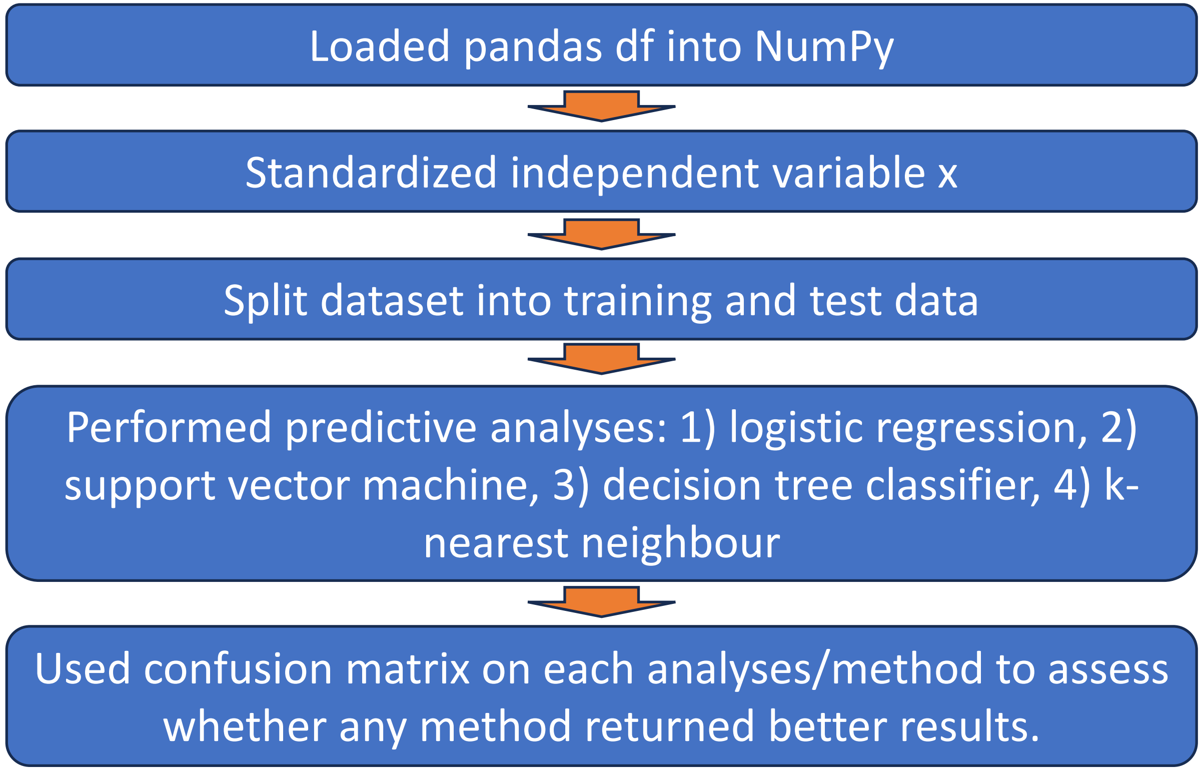
- Objects used in mapping visualization
  - Highlighted circle marker to show launch areas
  - Marker cluster and red/green markers to indicate success/failure of launches
  - Polylines used to indicate distances to objects of importance (highways, railroads, population centers, oceans)
- GitHub URL: https://github.com/knechtd/IBM-Data-Science/blob/main/06_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Plots used in dashboard:
  - Interactive dashboard built using Plotly Dash
  - Use of pie charts to indicate no. of launches per site
  - Use of scatter graph to depict relationship between outcome and payload mass per booster version.
- GitHub URL: https://github.com/knechtd/IBM-Data-Science/blob/main/07_spacex_dash_app.py

Predictive Analysis (Classification)

- As per the flowchart depicted on the right, the final result has shown that all methods are equally suited for this project. Decision tree training results performed minimally better than the rest.
- GitHub URL: https://github.com/knechtd/IBM-Data-Science/blob/main/08_Machine%20Learning%20Prediction.ipynb



Loaded pandas df into NumPy

Standardized independent variable x

Split dataset into training and test data

Performed predictive analyses: 1) logistic regression, 2) support vector machine, 3) decision tree classifier, 4) k-nearest neighbour

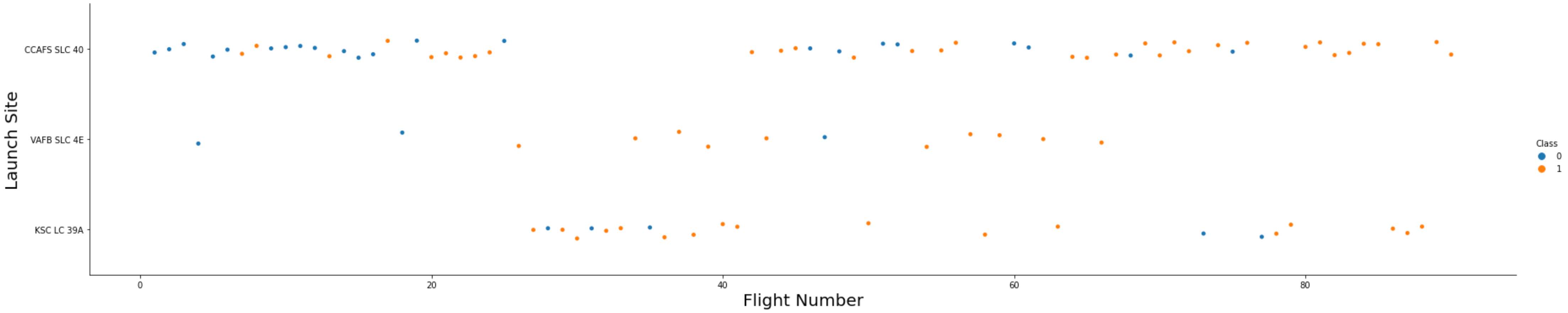
Used confusion matrix on each analyses/method to assess whether any method returned better results.

Section 2

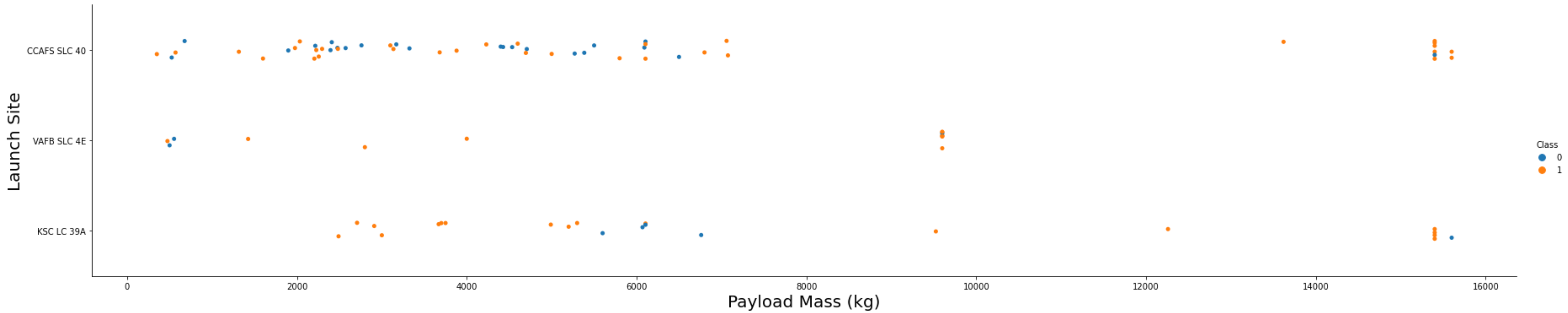
Insights drawn from EDA

Flight Number vs. Launch Site

- Flight number gives us a pseudo time-related picture on Space X attempts. We can see that most attempts were made for CCAF5 SLC 40.
- Attempts seem to become more successful the higher the flight number.



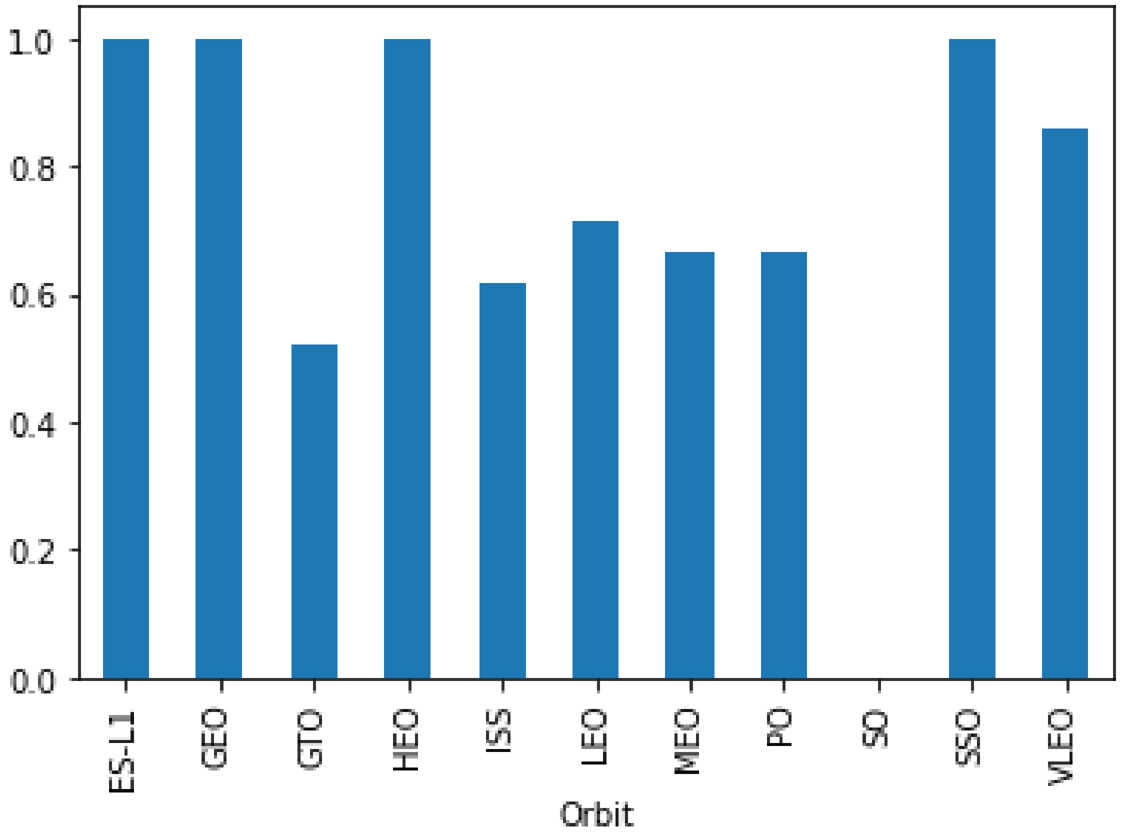
Payload vs. Launch Site



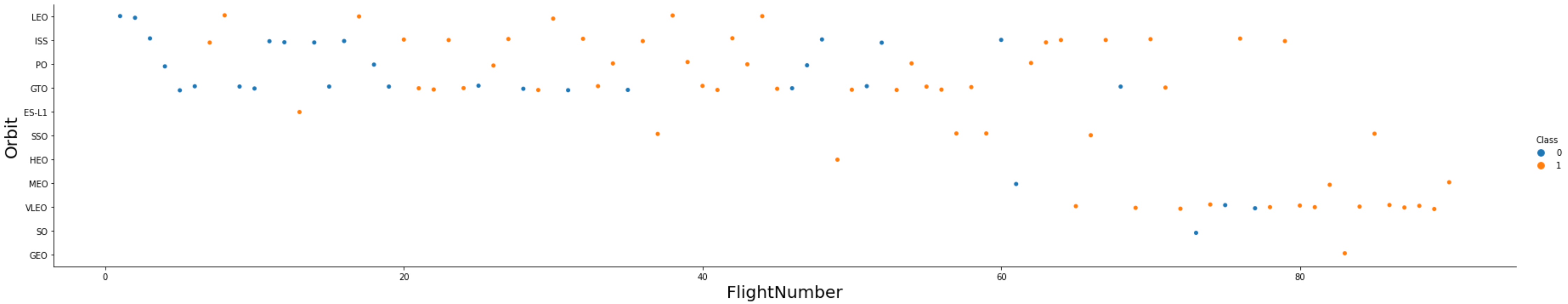
- Payload mass does not significantly differ between launch sites.
- It can be seen that most payload masses range between 0 and 7’000 kg for each launch site.

Success Rate vs. Orbit Type

- Data indicates that ES-L1, GEO, HEO and SSO feature the highest success rate.
- However, this does not give any indication on causation for the successes.

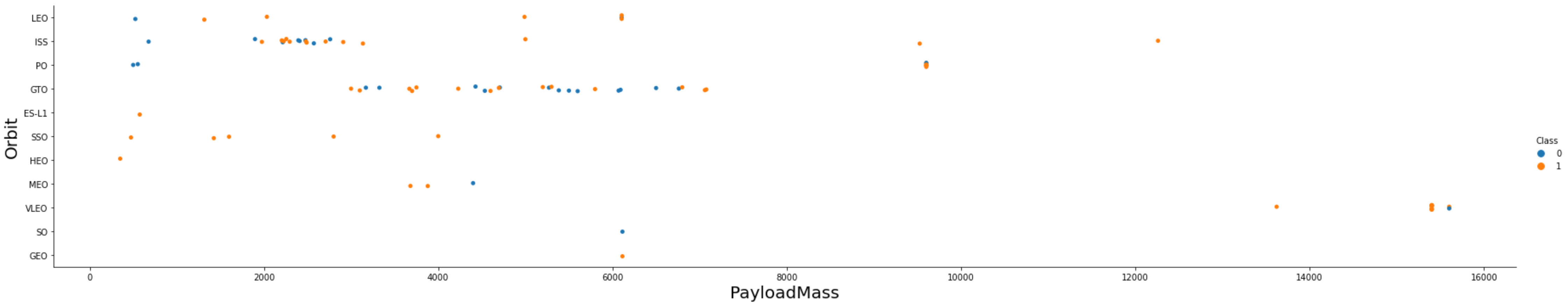


Flight Number vs. Orbit Type



- Earlier flights concentrated on orbitting LEO, ISS and GTO
- Flights in later stages focused more on VLEO instead

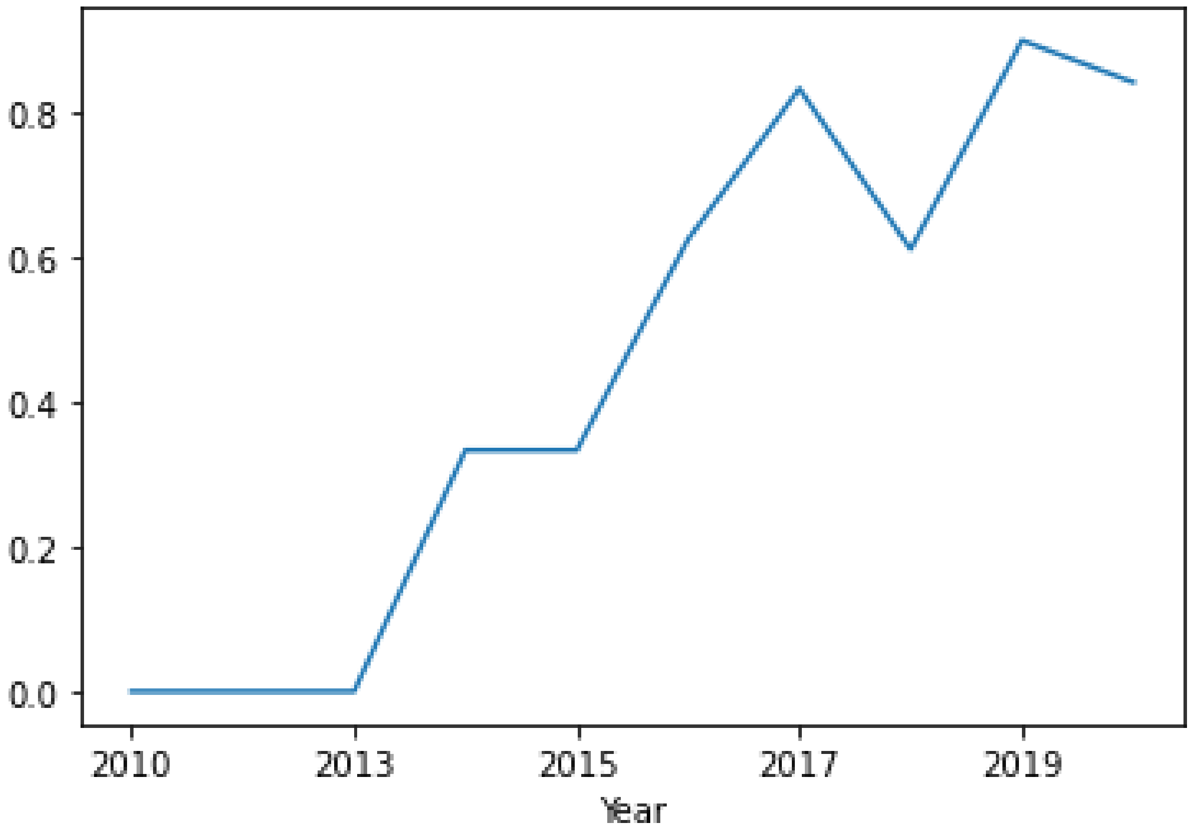
Payload vs. Orbit Type



- There is no clear indication from the data about a correlation between payload mass and orbit.
- The higher frequency of overall flights to LEO, ISS and GEO could give a misleading picture of the influence of payload mass on success and failure to reach a certain orbit.

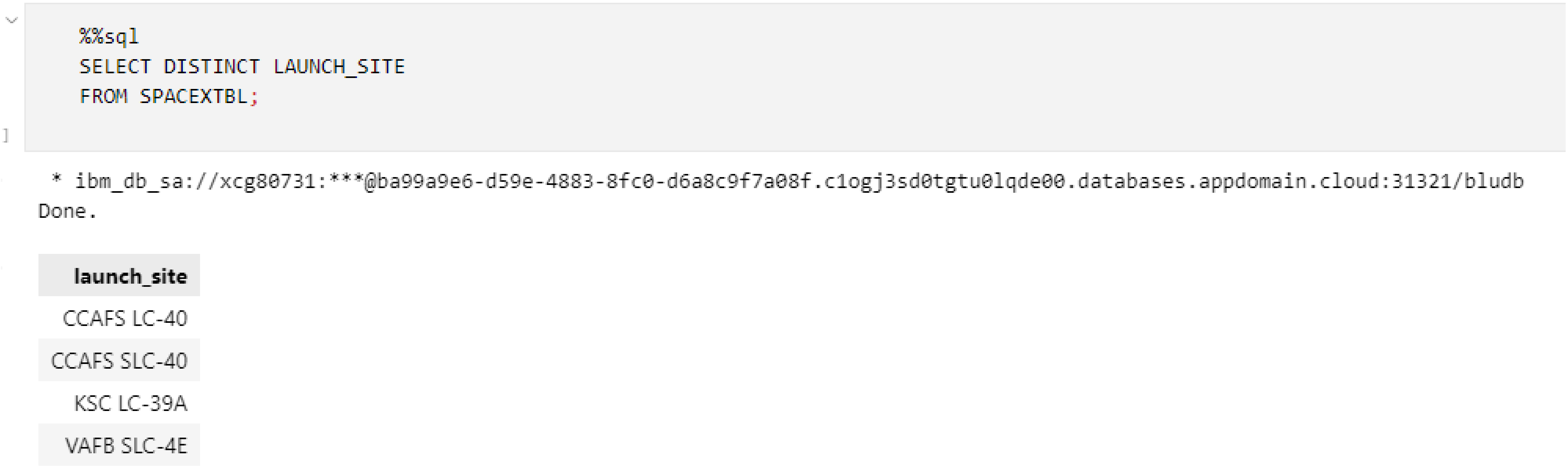
Launch Success Yearly Trend

- After an initial phase of failure (2010-2013), success rate has increased strongly until 2020, with a brief dip in 2018.



All Launch Site Names

- Launch sites have been retrieved from Space X flight data, using the DISTINCT command to de-duplicate entries to the four launch pads.

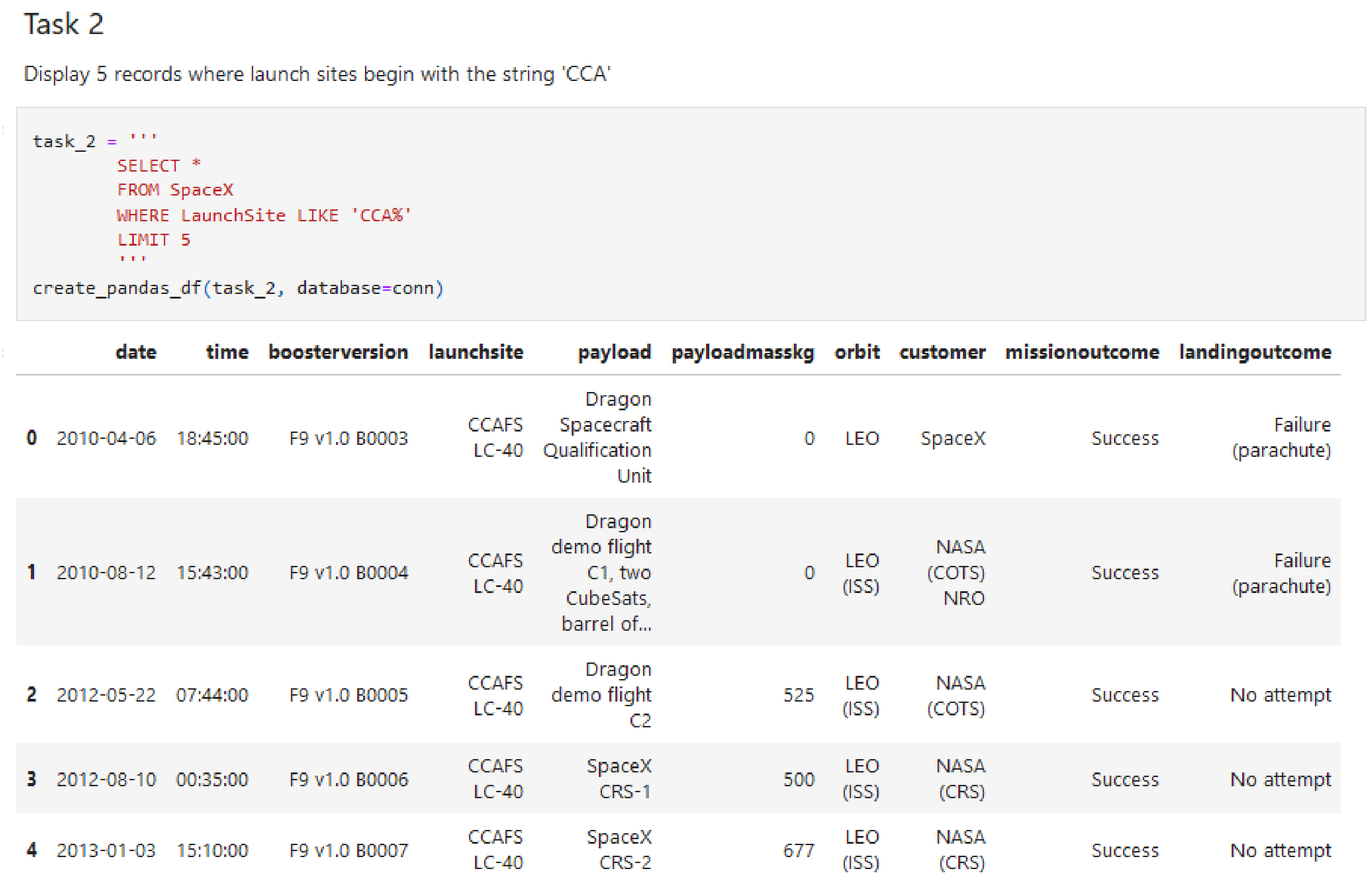


```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

```
 * ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'



Task 2

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

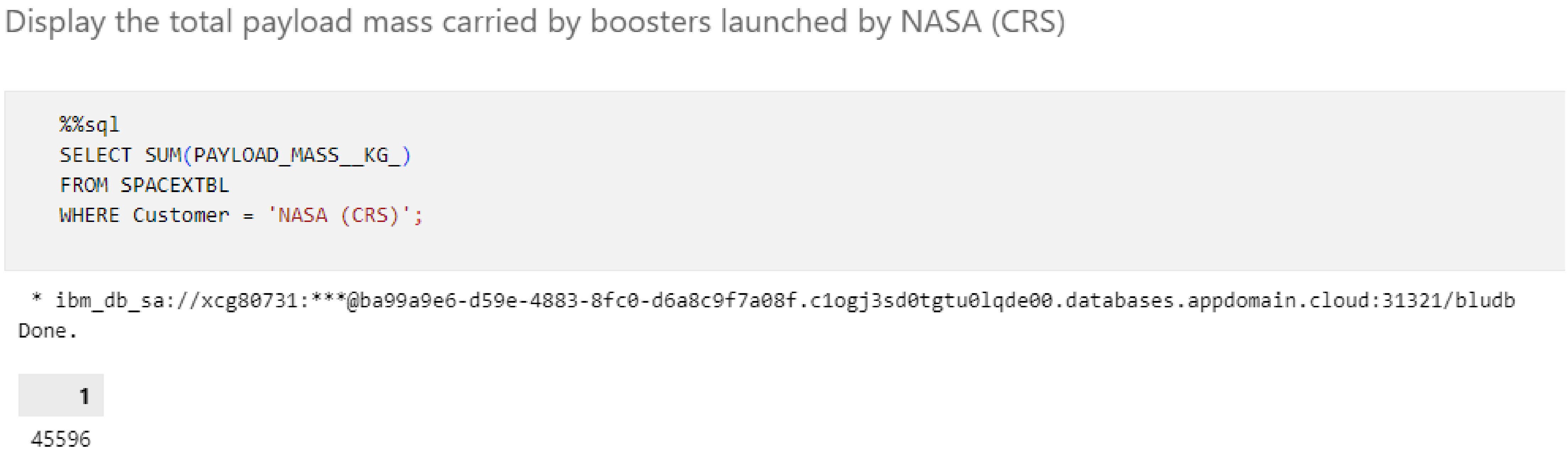
```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- As per the previous slide, the same dataset was used to identify and list the first five entries beginning with ‘CCA’. Applying LIMIT instead of DISTINCT, the entries show all the same launch site.

Total Payload Mass

- Total payload as per Space X dataset sums up to 45’596 kg.



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

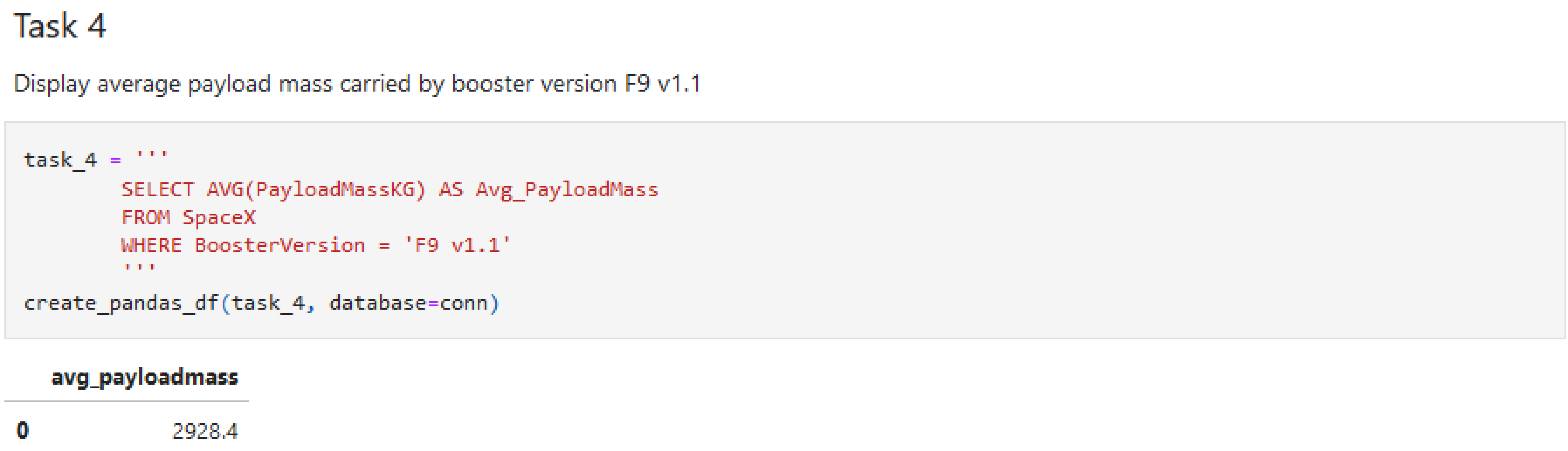
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

1
45596

Average Payload Mass by F9 v1.0%

- Booster payload mass entries per flight were averaged using a filter to focus on F9 v.1.1 only, resulting in 2’928.4 kg on average.



Task 4

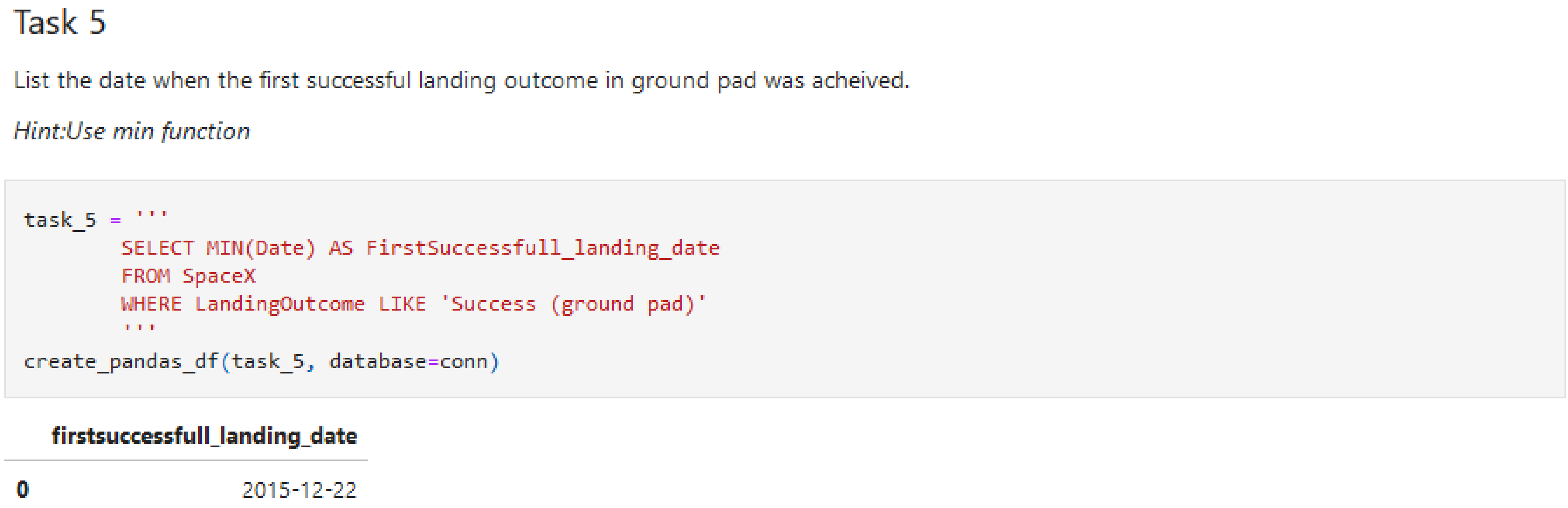
Display average payload mass carried by booster version F9 v1.1

```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

```
   avg_payloadmass
0           2928.4
```

First Successful Ground Landing Date

- Filtered Space X data set for landing outcomes, referencing minimum value due to oldest date having lower (numerical) value than newer dates.



Task 5

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

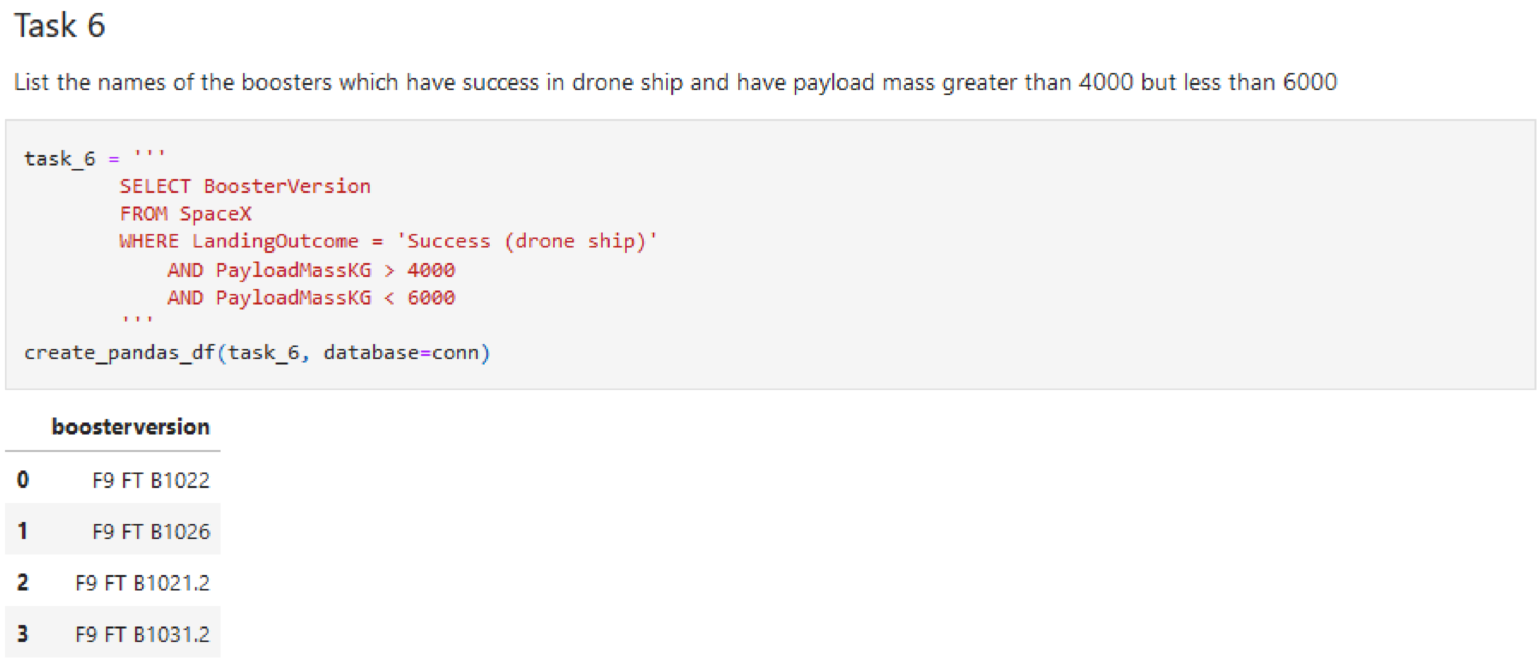
Hint:Use min function

```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000 kg

- There have been four successful drone ship landings by Falcon 9 with a payload mass between 4’000 and 6’000 kg



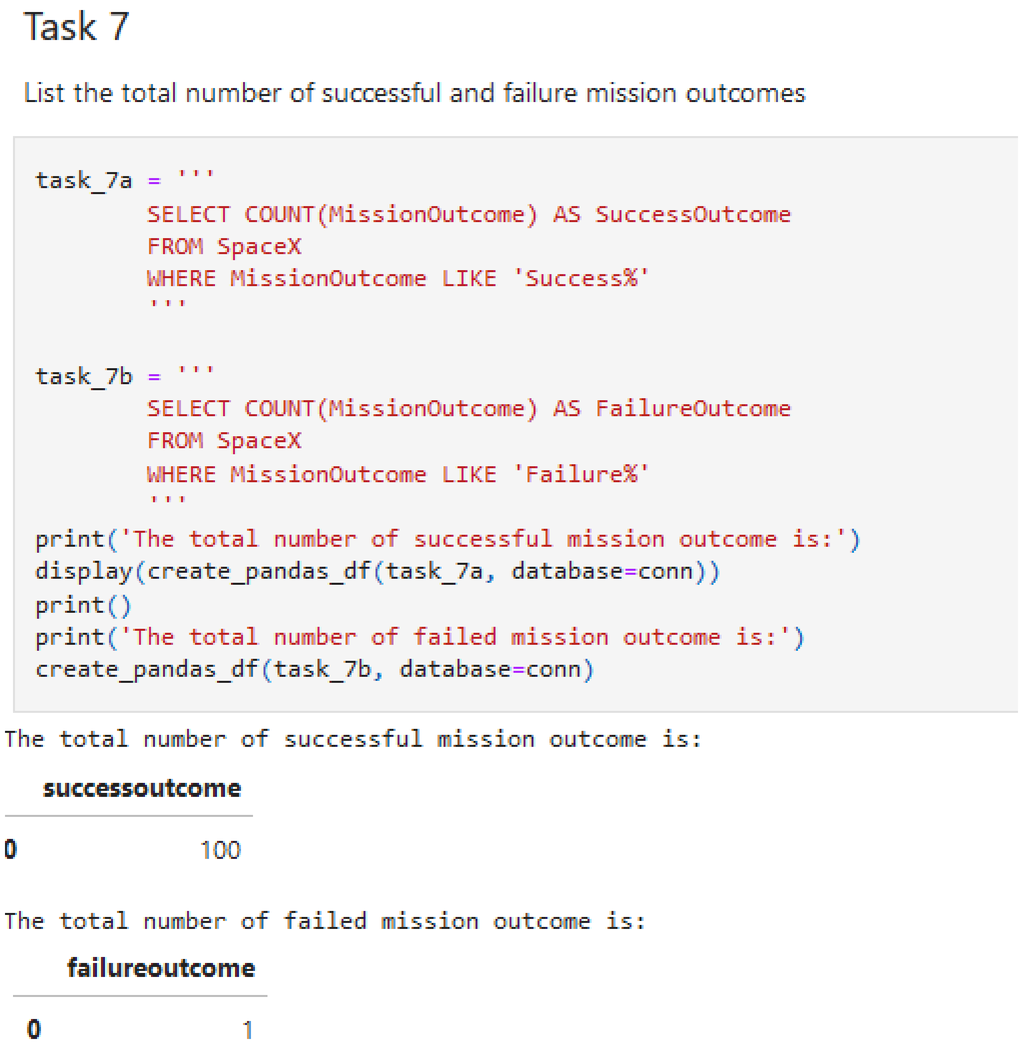
Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes



Task 7

List the total number of successful and failure mission outcomes

```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

```
The total number of successful mission outcome is:
```

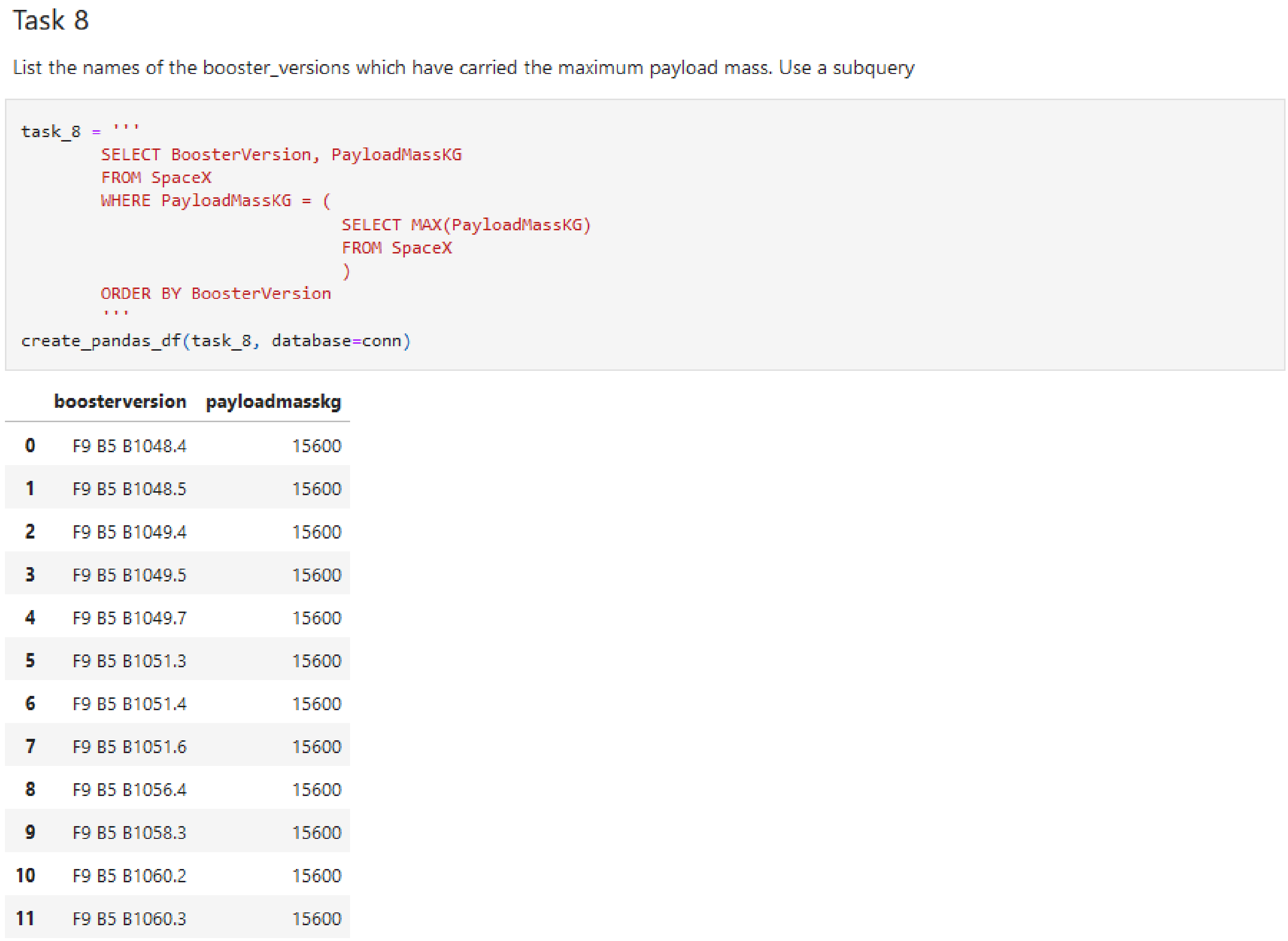
	successoutcome
0	100

```
The total number of failed mission outcome is:
```

	failureoutcome
0	1

- Applied count to mission outcome columns to elicit number of failures and successes separately.
- Use of WHERE to capture all outcomes that may contain ‘failure’ or ‘success’ wording.
- Resulted in a success to failure ratio of 100 to 1.

Boosters Carried Maximum Payload



Task 8

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

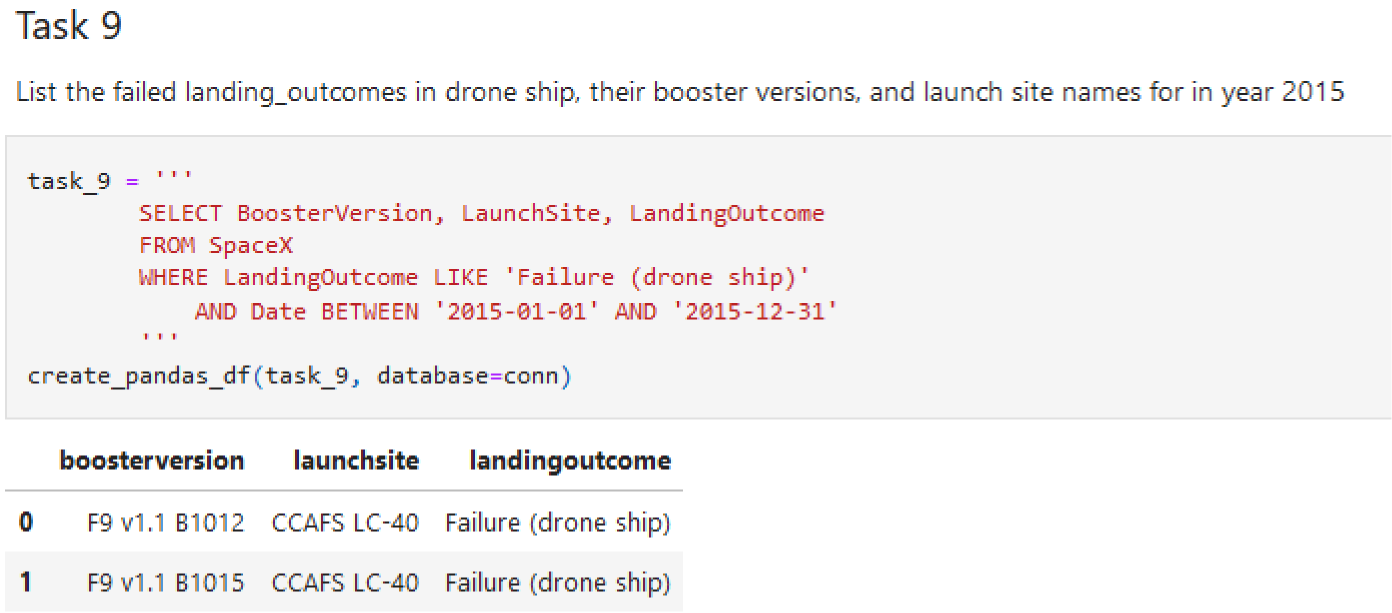
```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

- Using a select/where clause allows us to include a subquery to obtain results for the 11 booster versions with the highest payload mass in kg.
- As can be seen, each booster features a payload mass above 15’000 kg.

2015 Launch Records

- In 2015, two landing attempts failed. Both of them involved landing on a drone ship.



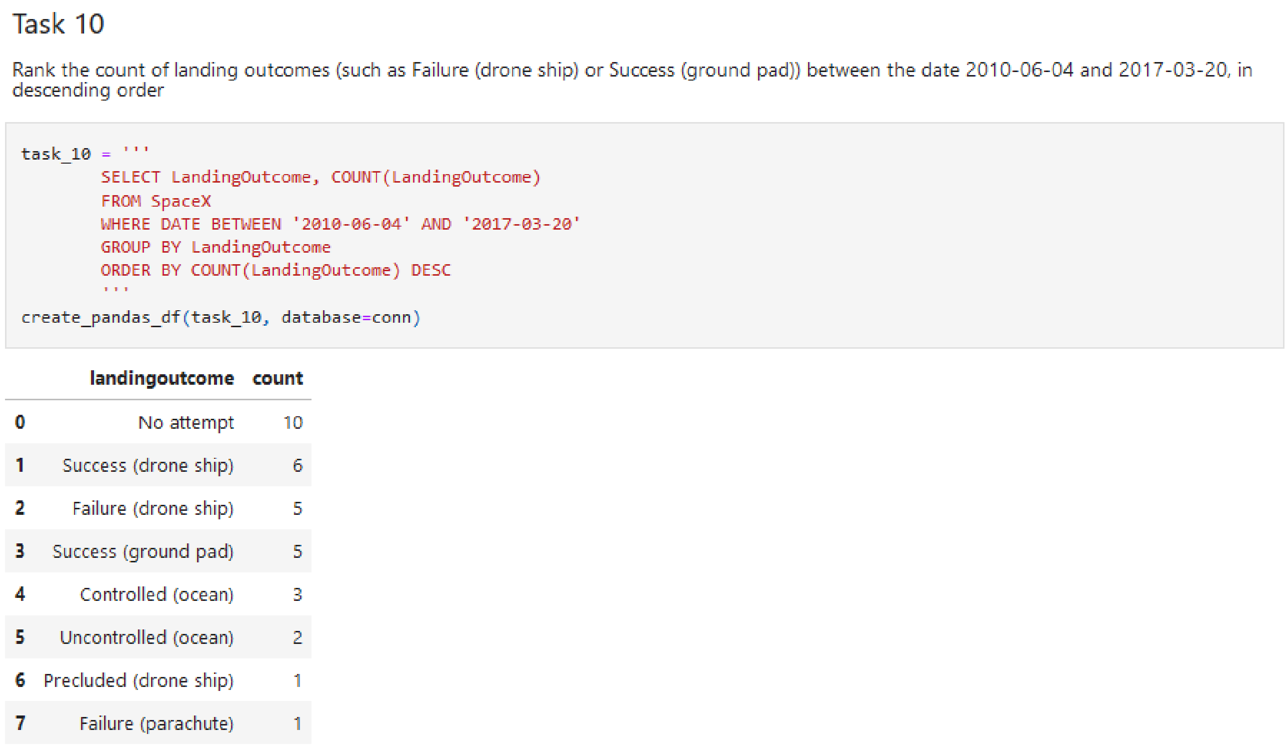
Task 9

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

```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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



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

```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

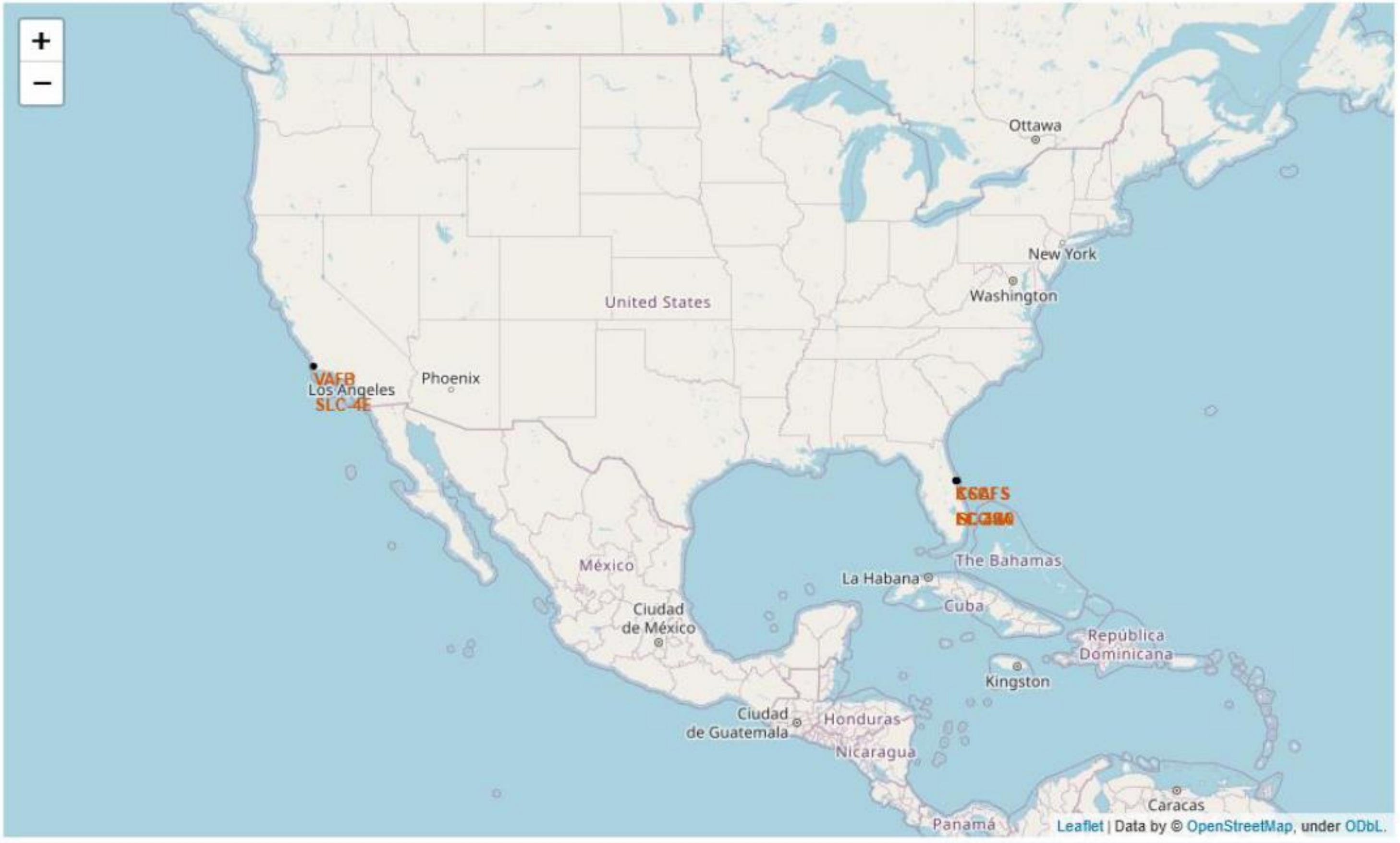
	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

- Indicating ‘No attempt’ as a failure, it has been the leading cause of failures.
- If we exclude ‘No attempt’ from our results, we can see that drone ship landings have been most successful in absolute numbers.

Section 3

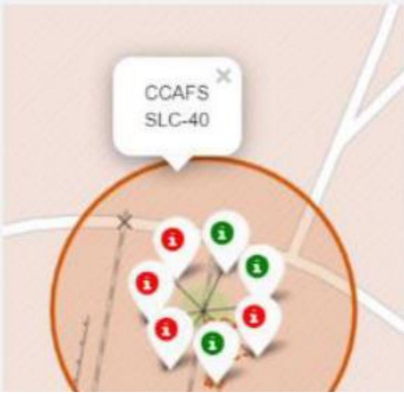
Launch Sites Proximities Analysis

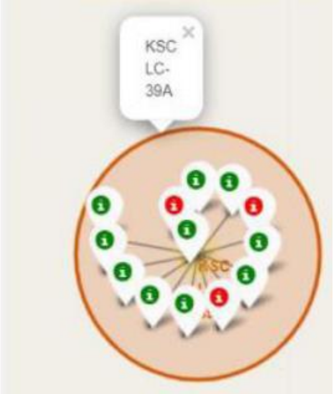
Total of successful launches per launch site

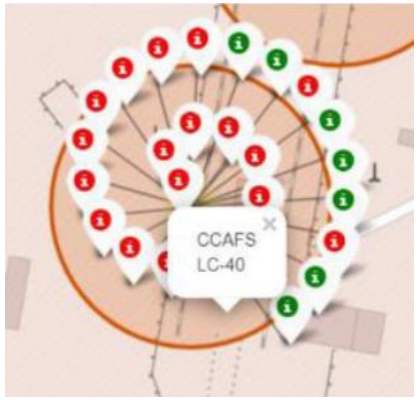


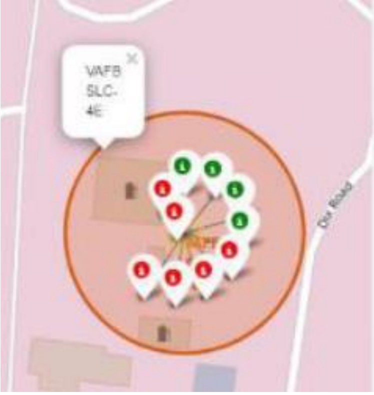
- Launch sites have been marked on both coasts of the United States (Florida and California)
- Sites are very close to the southern border, presumably due to favorably conditions to space travel

Visualization of success/failure by launch site



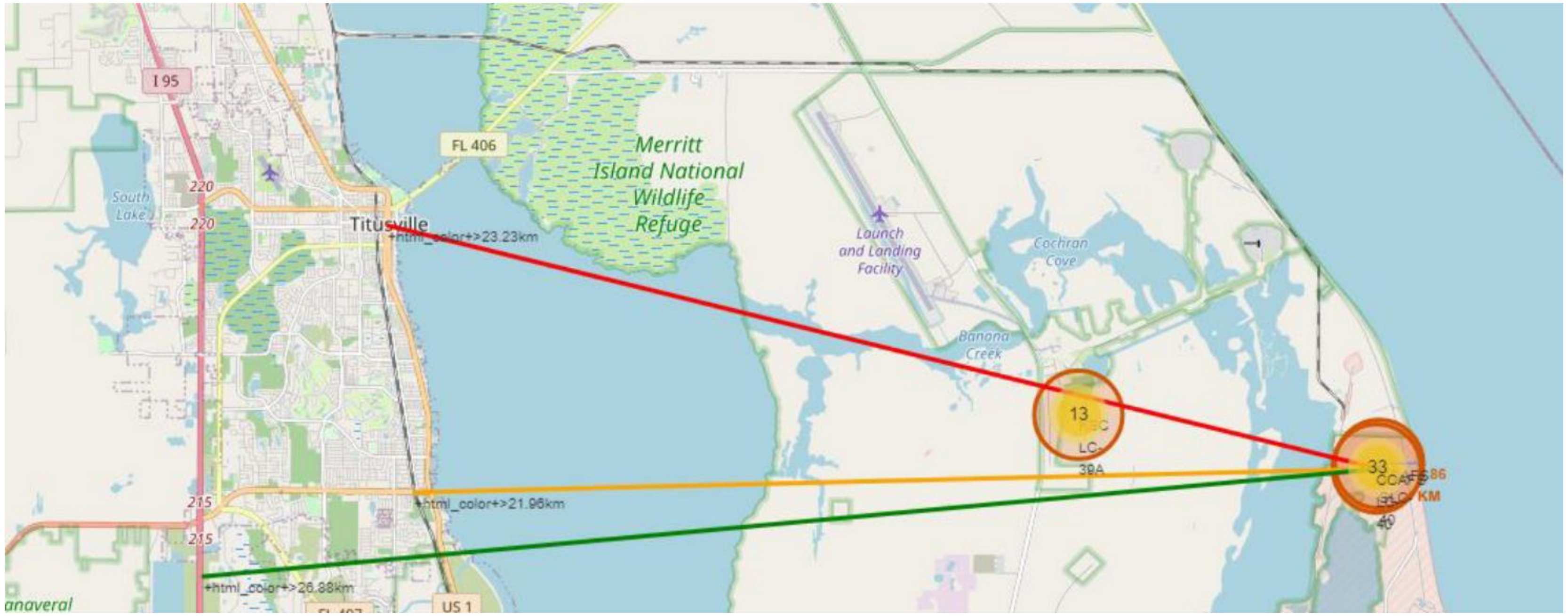






- Success and failure per launch site seem about equally distributed just from looking at it.
- However, KSC LC 39A seems to have more successful attempts than the other sites in relative proportions.

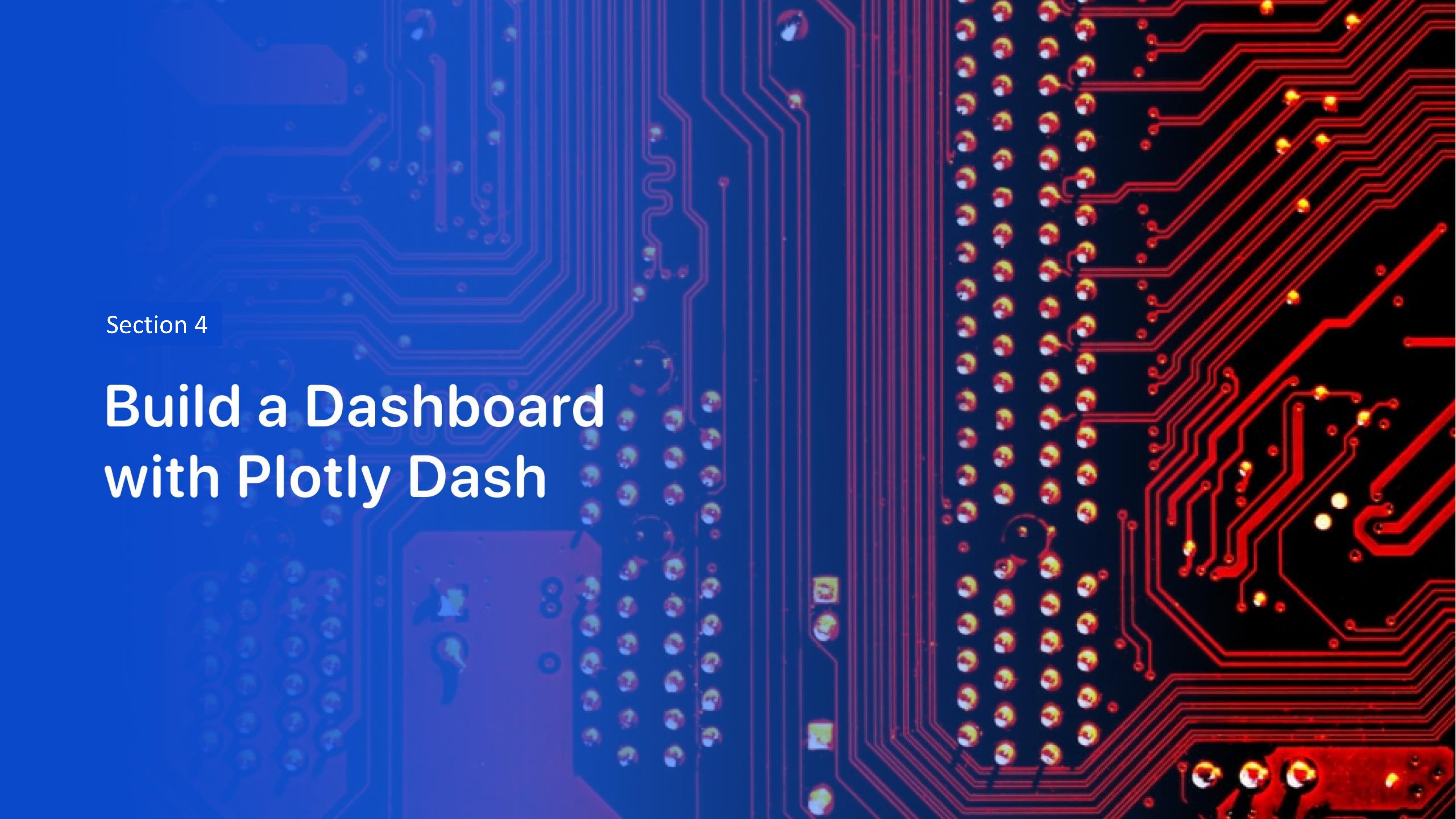
Proximity to nearby infrastructure for CCAFS



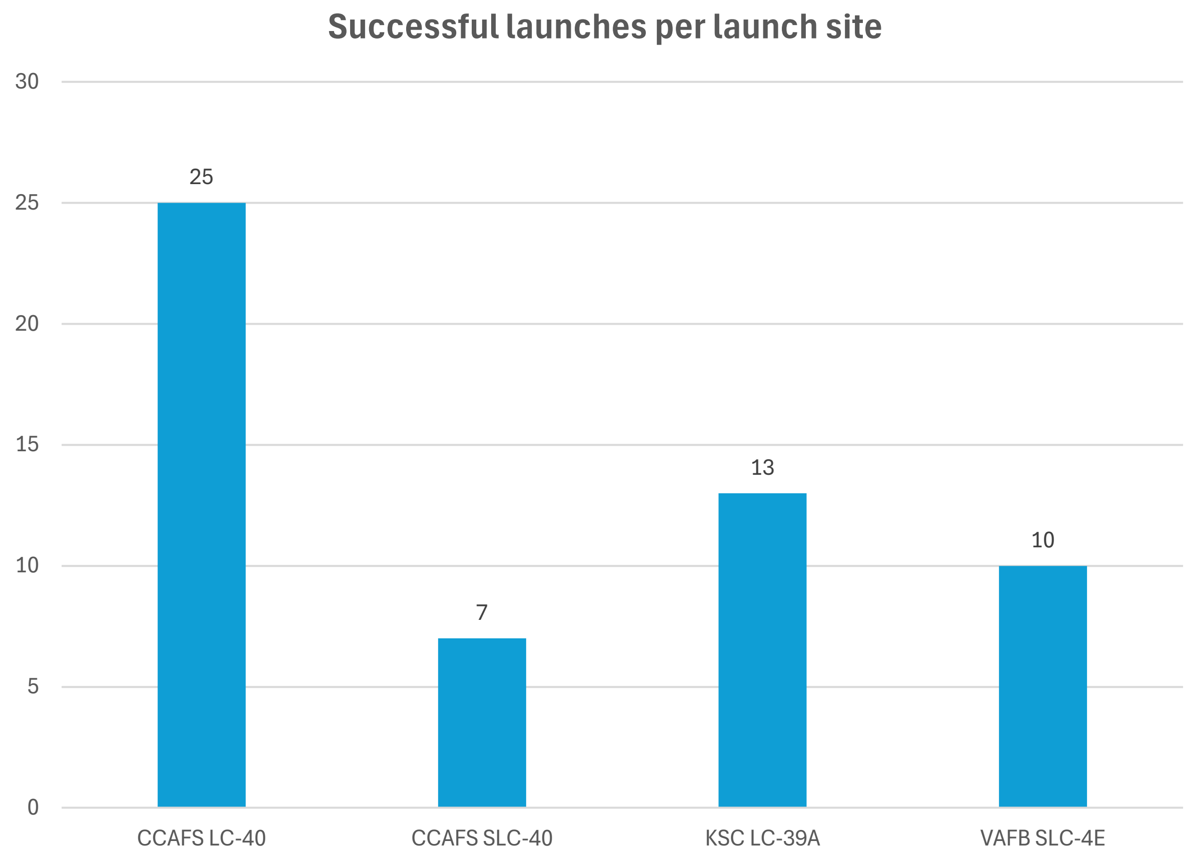
- CCAFS borders almost immediately to the coast line, where as it has a distance of around 20km to the nearest railway, city and highway.
- The distances to the shore and the nearest populous areas are most likely based on security concerns.

Section 4

Build a Dashboard with Plotly Dash

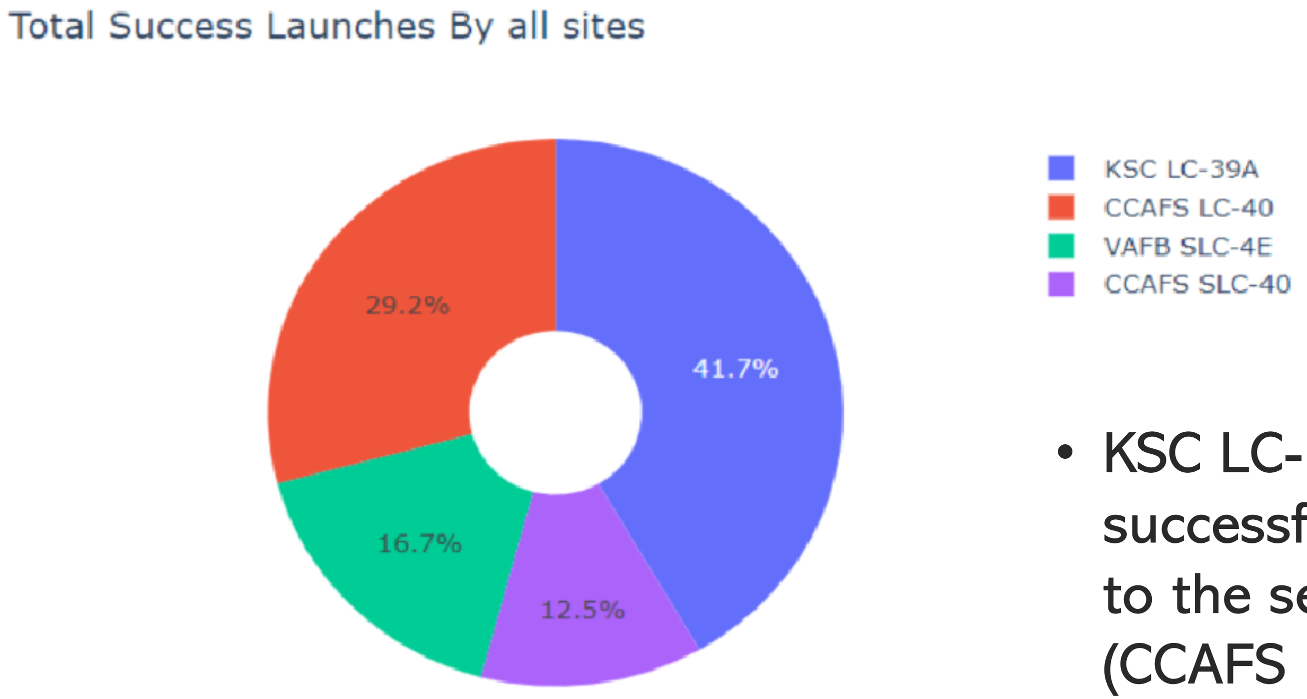


Number of successful launches per site



- The number of successful launches is by far the highest from CCAFS LC-40, followed by the other CCAFS site.
- However, this does not allow for conclusions about success without further clarifications.

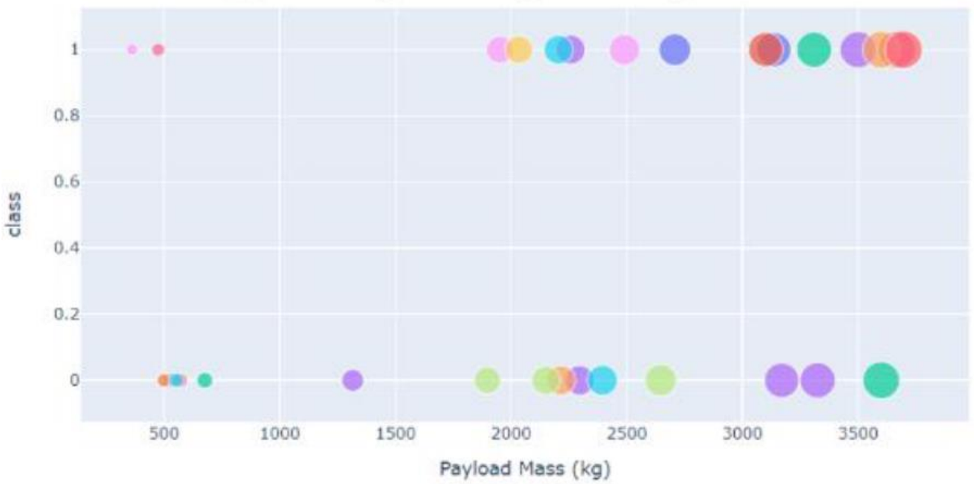
Percentage of successful launches per site



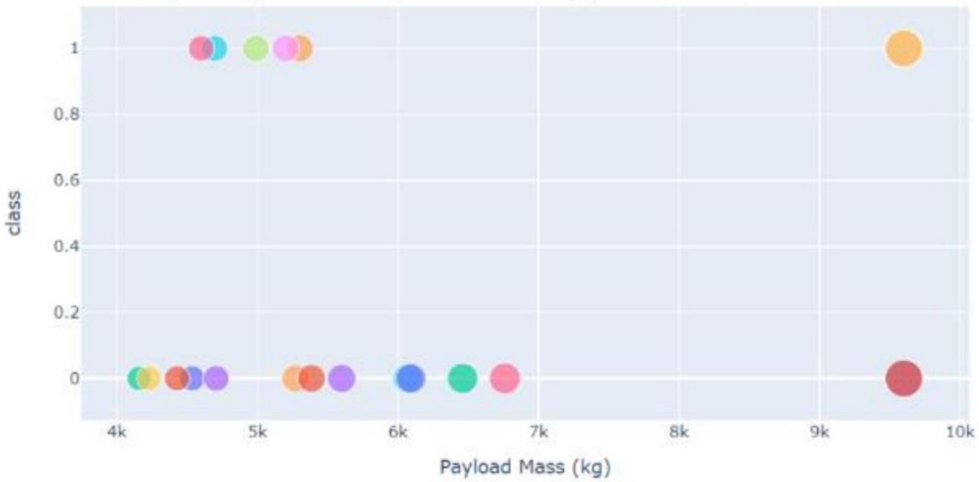
- KSC LC-39A shows by far the most successful launches with 41.7%, compared to the second most successful with 29.2% (CCAFS LC-40)
- Floridian launch pads appear to be better suited for space launches, but this hypothesis remains to be tested

Potential impact of payload on launch outcomes

Payload Mass 0-4’000 kg



Payload Mass 4’000-10’000 kg



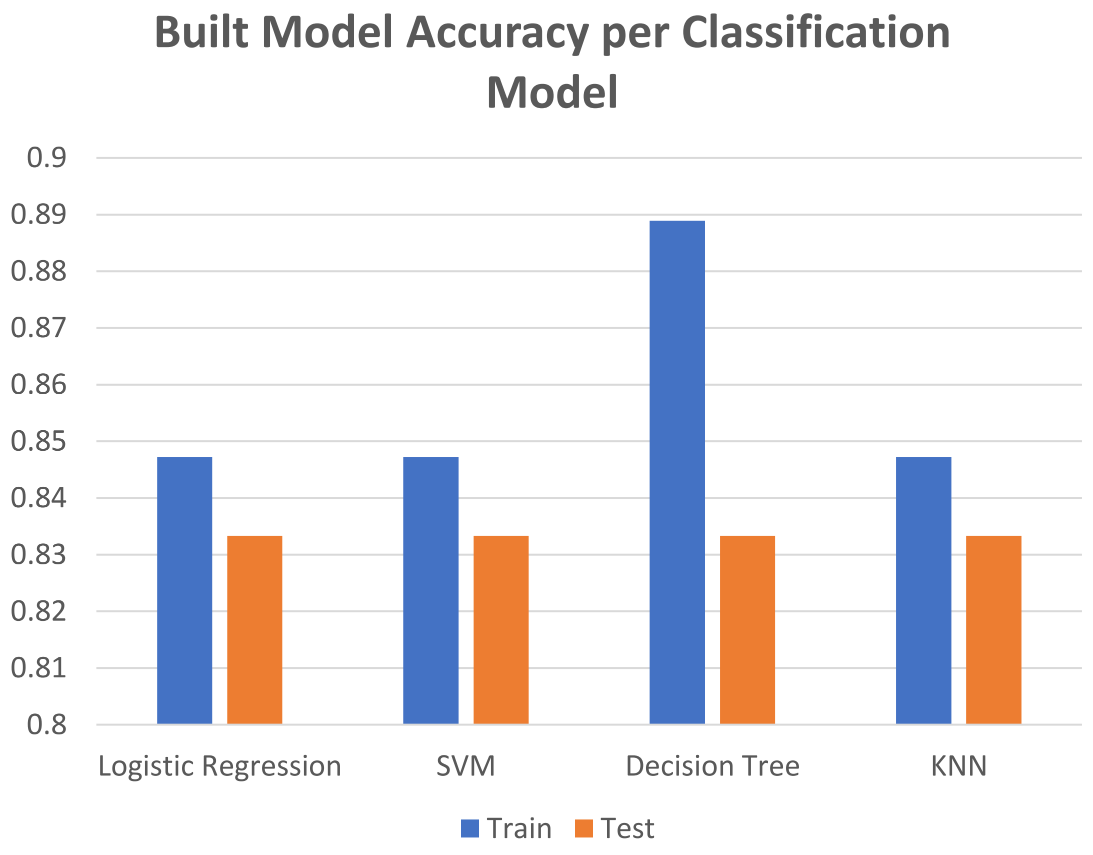
- Altogether, lower payload appears more successful. Statistical significance might be too low for solid conclusions, but current findings tend towards the lower end.

Section 5

Predictive Analysis (Classification)

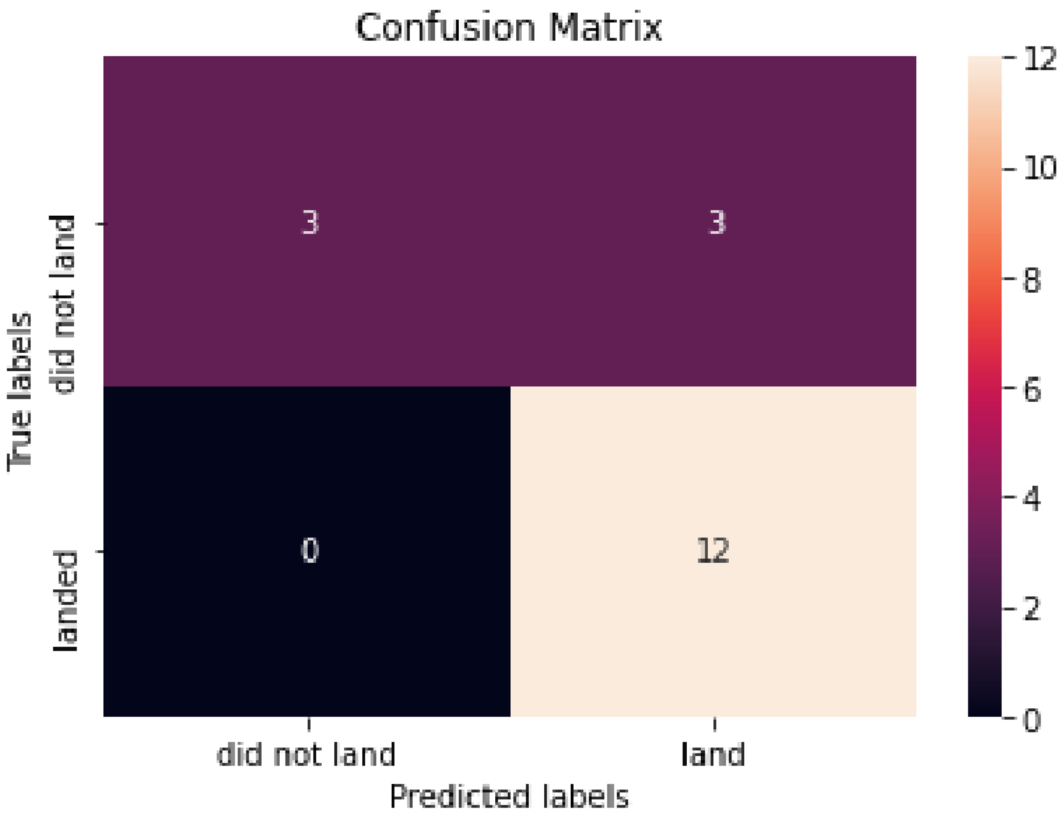
Classification Accuracy

- As per the training data accuracy, the decision tree model has a minimally higher accuracy compared to the others.
- Test scores for all four models is about the same.



Confusion Matrix

- The confusion matrix for the decision tree model show the best results



Conclusions

- The use of two data sources has delivered well-rounded results for this analysis. The use of Wikipedia as an unbiased source contrasts to Space X's own data.
- Space X technology must have improved over time, as the number and rate of successful landings has improved from 2013 until today.
- Launch site KSC LC-39a ('Kennedy Space Center') is deemed the best launch site for future attempts.
- Following various analyses, orbits ES-L1, GEO, HEO, SSO, VLE seem best suited for a successful attempt.
- Different predictive methods can be applied, but the decision tree classifier should lead to valid results.

Thank you!