IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Natalia Kniazeva
21.03.3032

Outline

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

Executive Summary

- Summary of methodologies

1. Data Collection Methodology
   - ✓ Using SpaceX Rest API
   - ✓ Using Web Scrapping from Wikipedia
2. Performed data wrangling
   - ✓ Filtering the data
   - ✓ Dealing with missing values
   - ✓ Using One Hot Encoding to prepare the data to a binary classification

- Summary of all results

  - ✓ Launches with a low payload mass show better results than launches with a larger payload mass.
  - ✓ Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
  - ✓ The rate of successful landings increases over the years.
  - ✓ KSC LC-39A has the highest success rate of the launches from all the sites.
  - ✓ Orbits ES-L1, GEO, HEO and SSO have the most success rate.
  - ✓ Decision Tree Model is the best algorithm for this dataset.

Introduction

- SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.
- Problems to find answers to:

✓ How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

✓ Does the rate of successful landings increase over the years?

✓ What is the best algorithm that can be used for binary classification in this case?

Section 1

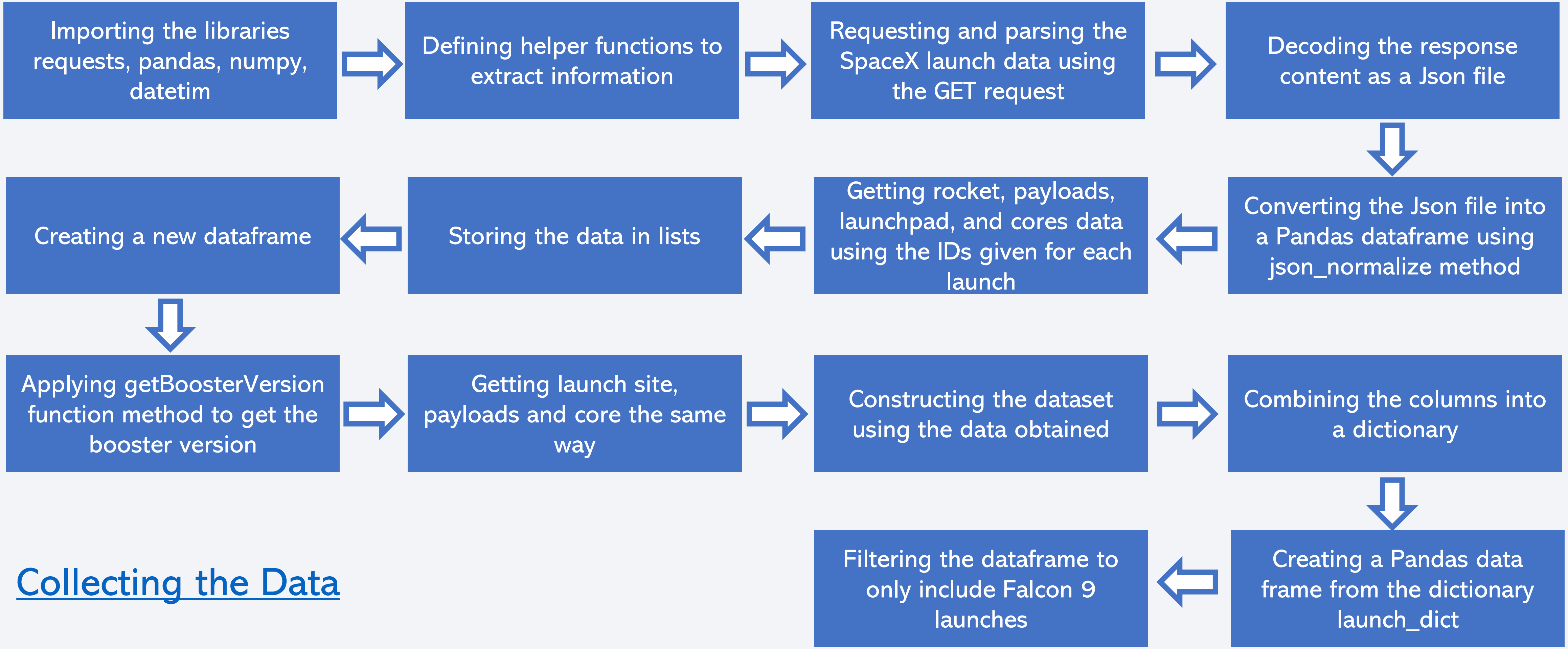
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX REST API and web scraping from a table in SpaceX's Wikipedia page.
- Perform data wrangling
  - Data was processed with encoding to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection



Importing the libraries requests, pandas, numpy, datetim

Defining helper functions to extract information

Requesting and parsing the SpaceX launch data using the GET request

Decoding the response content as a Json file

Converting the Json file into a Pandas dataframe using json_normalize method

Getting rocket, payloads, launchpad, and cores data using the IDs given for each launch

Storing the data in lists

Creating a new dataframe

Applying getBoosterVersion function method to get the booster version

Getting launch site, payloads and core the same way

Constructing the dataset using the data obtained

Combining the columns into a dictionary

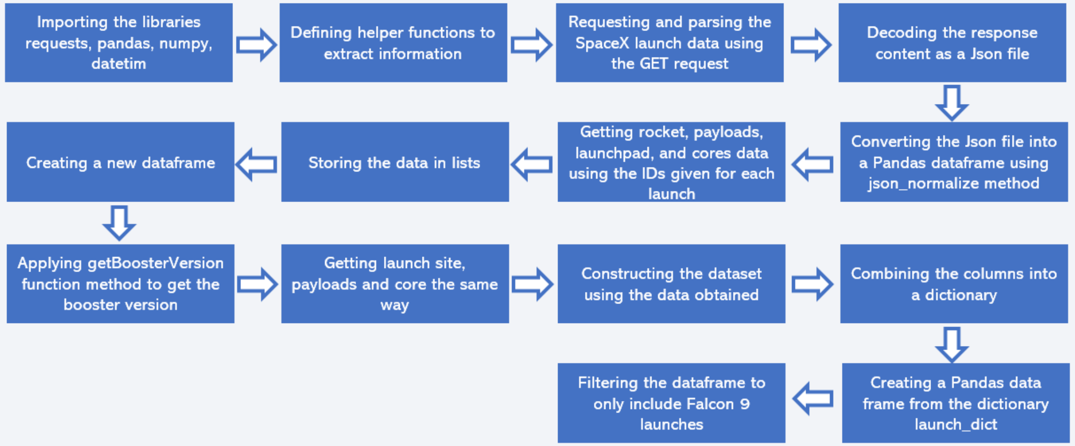
Creating a Pandas data frame from the dictionary launch_dict

Filtering the dataframe to only include Falcon 9 launches

Collecting the Data

Data Collection – SpaceX API

- Requesting to the SpaceX API
- Cleaning the requested data
- Collecting the Data



Importing the libraries requests, pandas, numpy, datetim

Defining helper functions to extract information

Requesting and parsing the SpaceX launch data using the GET request

Decoding the response content as a Json file

Converting the Json file into a Pandas dataframe using json_normalize method

Getting rocket, payloads, launchpad, and cores data using the IDs given for each launch

Storing the data in lists

Creating a new dataframe

Applying getBoosterVersion function method to get the booster version

Getting launch site, payloads and core the same way

Constructing the dataset using the data obtained

Combining the columns into a dictionary

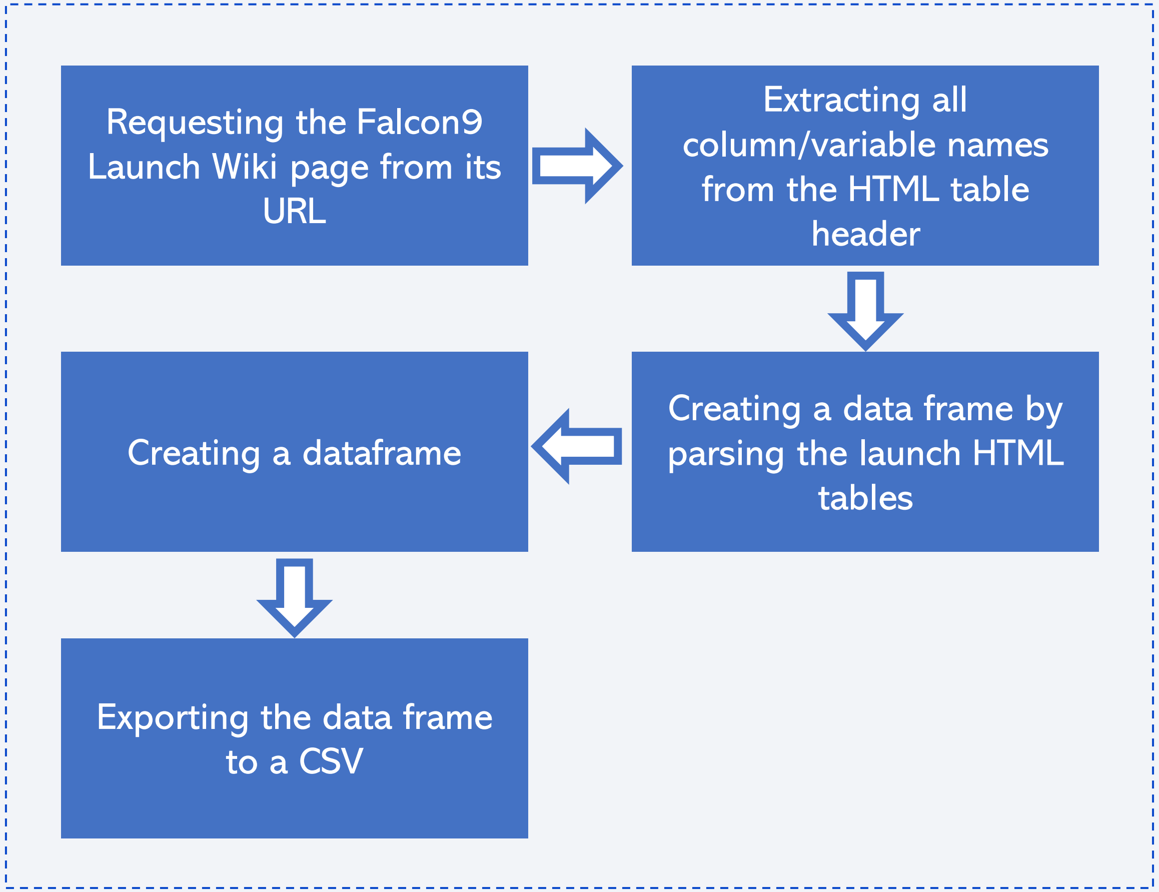
Creating a Pandas data frame from the dictionary launch_dict

Filtering the dataframe to only include Falcon 9 launches

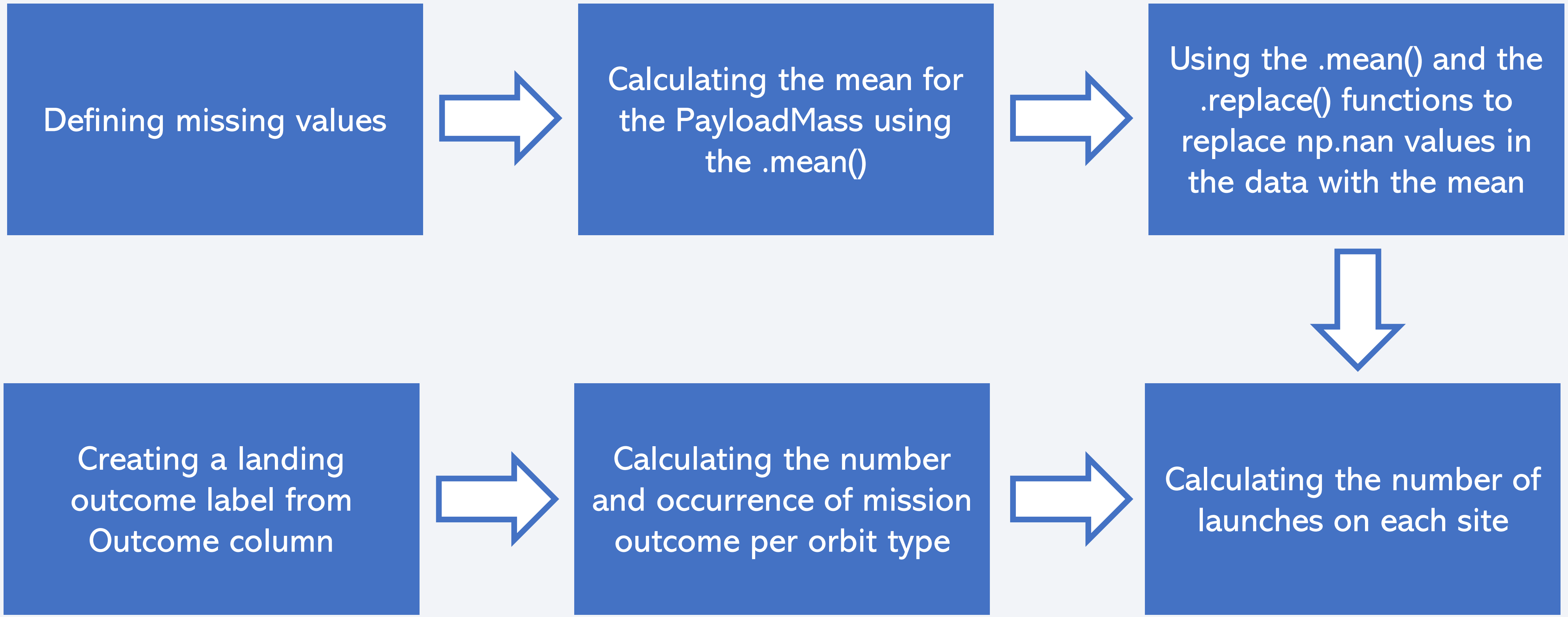
See flowcharts in detail on the page 7

Data Collection - Scraping

- Web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled **List of Falcon 9 and Falcon Heavy Launches**
- Web Scraping


Requesting the Falcon9 Launch Wiki page from its URL
Extracting all column/variable names from the HTML table header
Creating a data frame by parsing the launch HTML tables
Creating a dataframe
Exporting the data frame to a CSV

Data Wrangling



Defining missing values

Calculating the mean for the PayloadMass using the .mean()

Using the .mean() and the .replace() functions to replace np.nan values in the data with the mean

Calculating the number of launches on each site

Calculating the number and occurrence of mission outcome per orbit type

Creating a landing outcome label from Outcome column

- Collecting the Data
- Data Wrangling

EDA with Data Visualization

- The charts plotted reveal the following:
  - ✓ how the number of launch attempts and payload affect the launch outcome (catplot)
  - ✓ the relationship between launch attempts and launch site (catplot)
  - ✓ the relationship between payload and launch site (catplot)
  - ✓ the relationship between success rate and each orbit type (barplot)
  - ✓ the relationship between number of launch attempts and orbit type (catplot)
  - ✓ the relationship between payload and orbit type (catplot)
  - ✓ launch success yearly trend (lineplot)
- The charts can help us to predict if the Falcon 9 first stage will land successfully
- Exploring and Preparing Data

EDA with SQL

- The SQL queries performed during analysis:
  - ✓ Displaying the names of the unique launch sites in the space mission
  - ✓ Displaying 5 records where launch sites begin with the string 'CCA'
  - ✓ Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - ✓ Displaying average payload mass carried by booster version F9 v1.1
  - ✓ Listing the date when the first successful landing outcome in ground pad was acheived
  - ✓ Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - ✓ Listing the total number of successful and failure mission outcomes
  - ✓ Listing the names of the booster versions which have carried the maximum payload mass
  - ✓ Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - ✓ Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- SQL Notebook

Build an Interactive Map with Folium

- The following map objects were created and added to a folium map:
  - ✓ a marker with Circle, Popup Label and Text Label of NASA Johnson Space Center
  - ✓ markers with Circle, Popup Label and Text Label of all launch sites
  - ✓ coloured markers of the launch outcomes for each launch site
  - ✓ coloured lines to show the proximities of the KSC LC-39A launch site to a railway, a highway, a coastline, and a nearest city
- These map objects were created to identify:
  - ✓ the launch sites geographical locations, their proximity to Equator and coasts
  - ✓ which launch sites have relatively high success rates
  - ✓ distances between a Launch Site to its proximities
- Folium Maps

Build a Dashboard with Plotly Dash

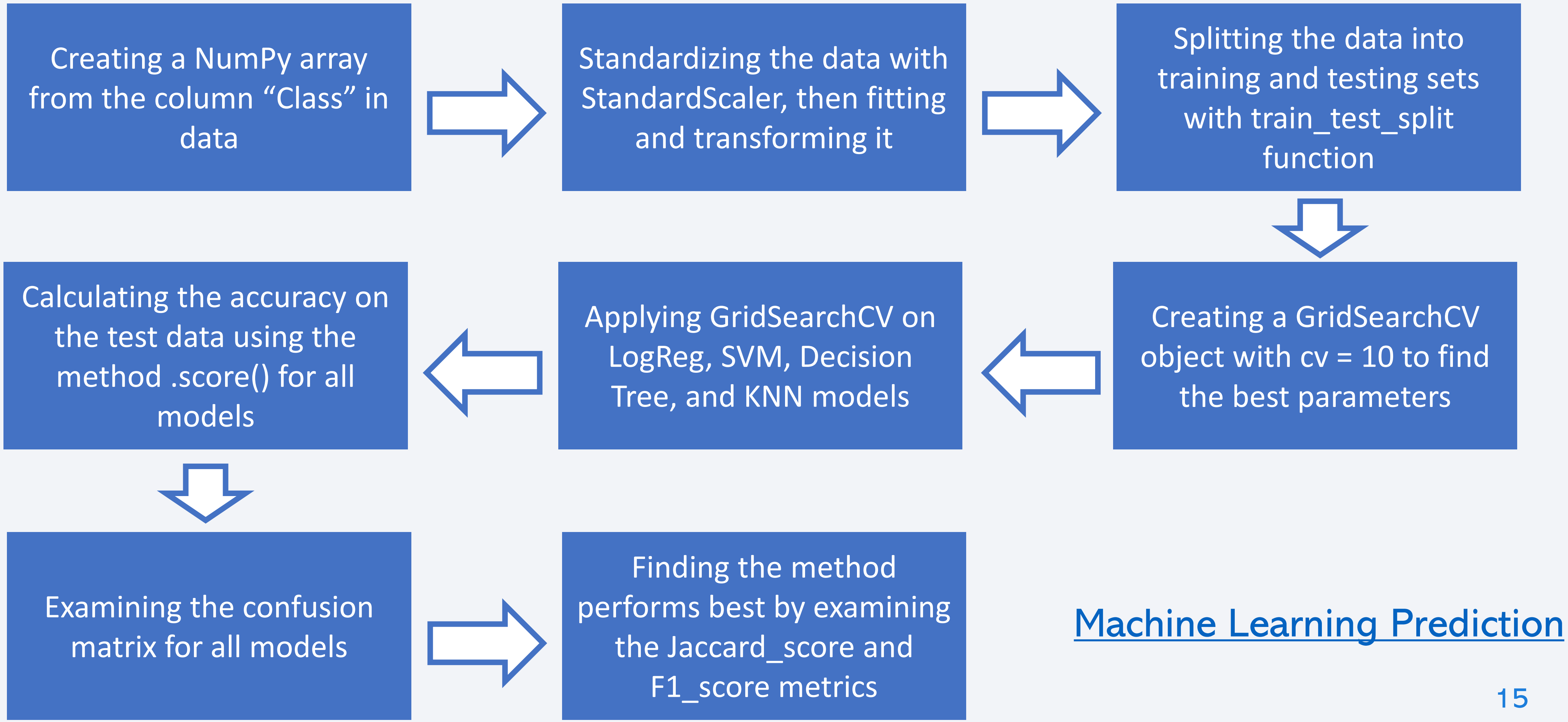
- The following plots were created:

  ✓Total Success Launches by Site (piechart)
  ✓Success Launches for site KSC LC-39A (piechart)
  ✓Payload vs. Launch Outcome for all sites (two scatter chart)

- The correlations between variables display launchings with which payloads and from which sites were more successful.

- Plotly Dash lab

Predictive Analysis (Classification)



Creating a NumPy array from the column “Class” in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy on the test data using the method .score() for all models

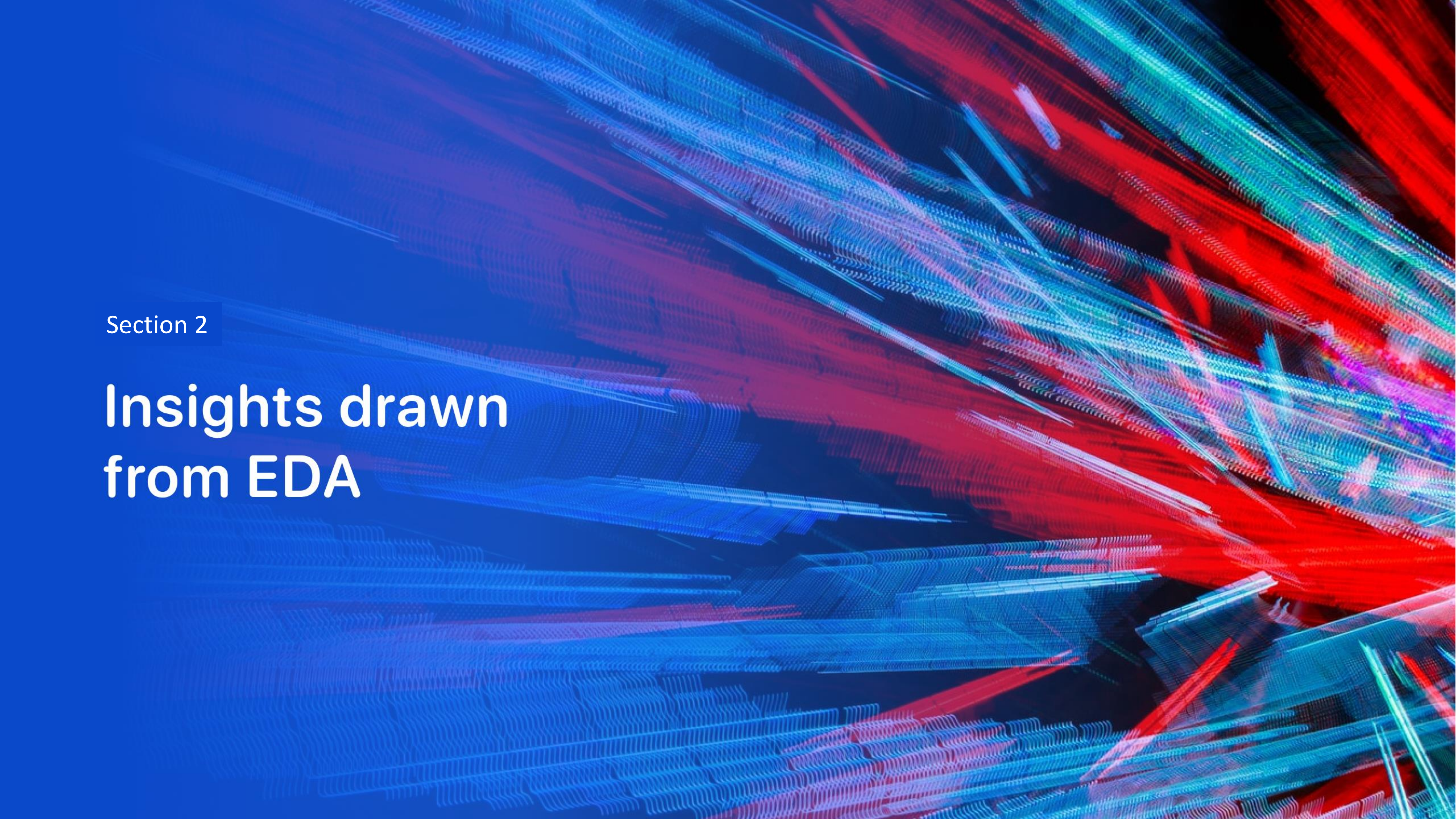
Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

Machine Learning Prediction

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



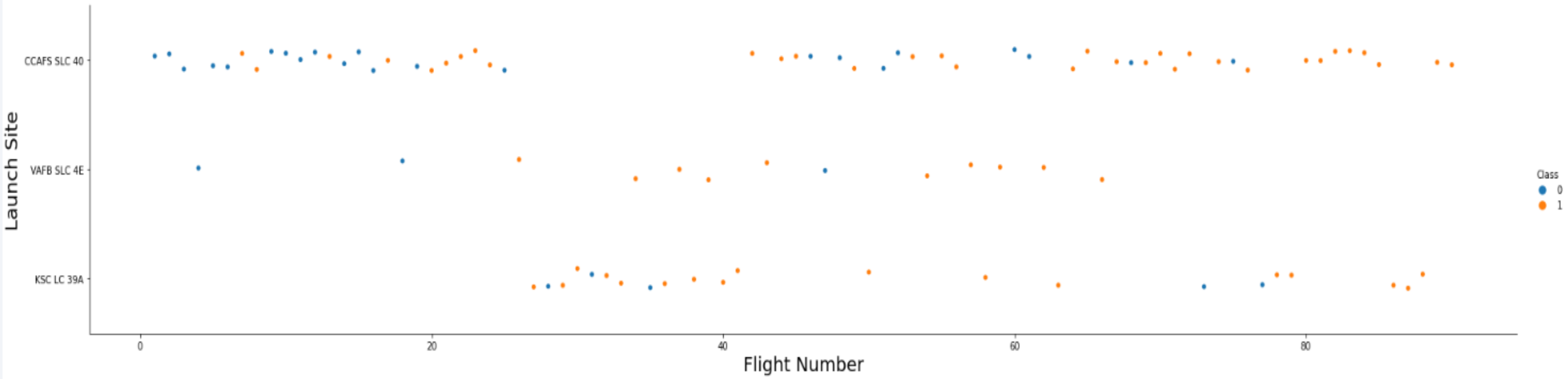
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- The launching performance improves with the number of flights for each site.

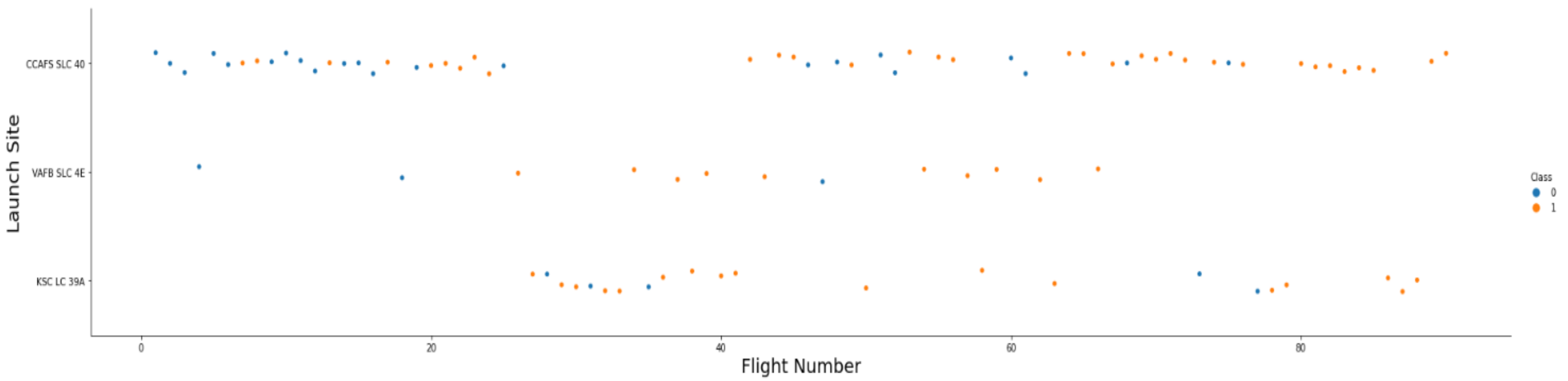
- Successful Launch
- Failed Launch



Payload vs. Launch Site

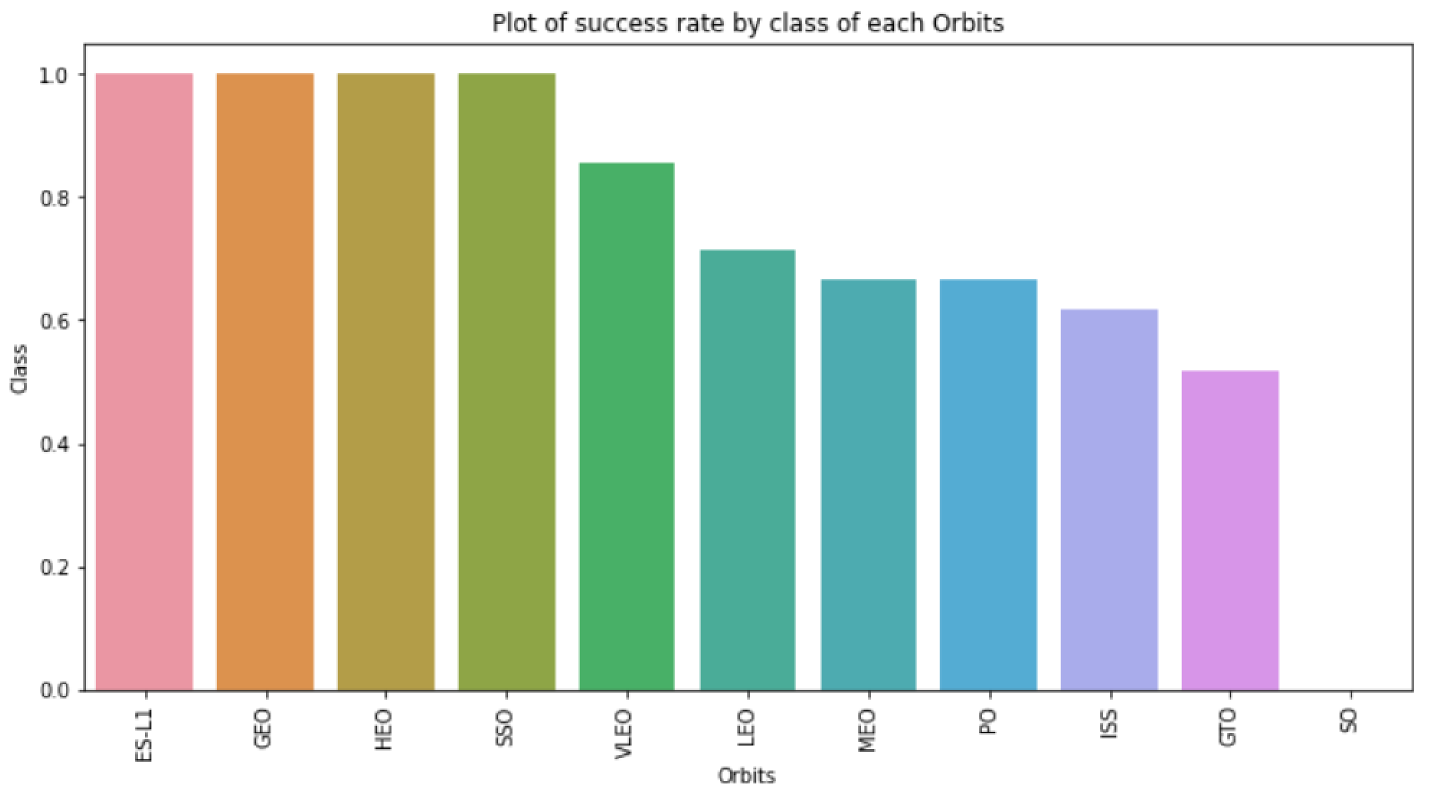
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000 kg).

- Successful Launch
- Failed Launch



Success Rate vs. Orbit Type

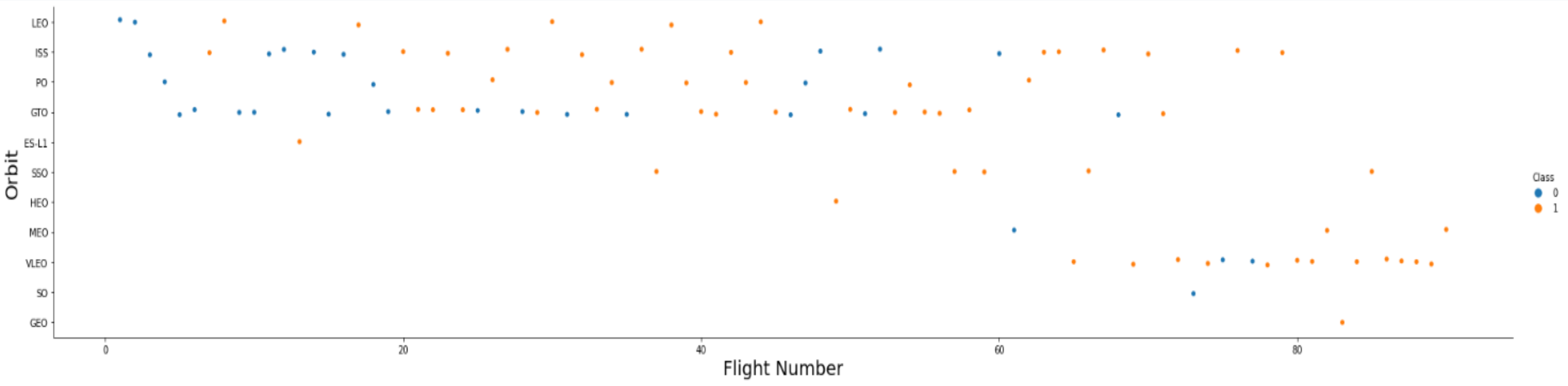
- ES-L1, GEO, HEO, SSO, and VLEO are the orbits that have high success rate.
- The SO has the least success rate amongst the orbits.



Flight Number vs. Orbit Type

- A relationship between Flight Number and Orbit type is not explicit.
- For the LEO orbit, the success appears related to the number of flights.
- There seems to be no relationship between flight number in GTO orbit.

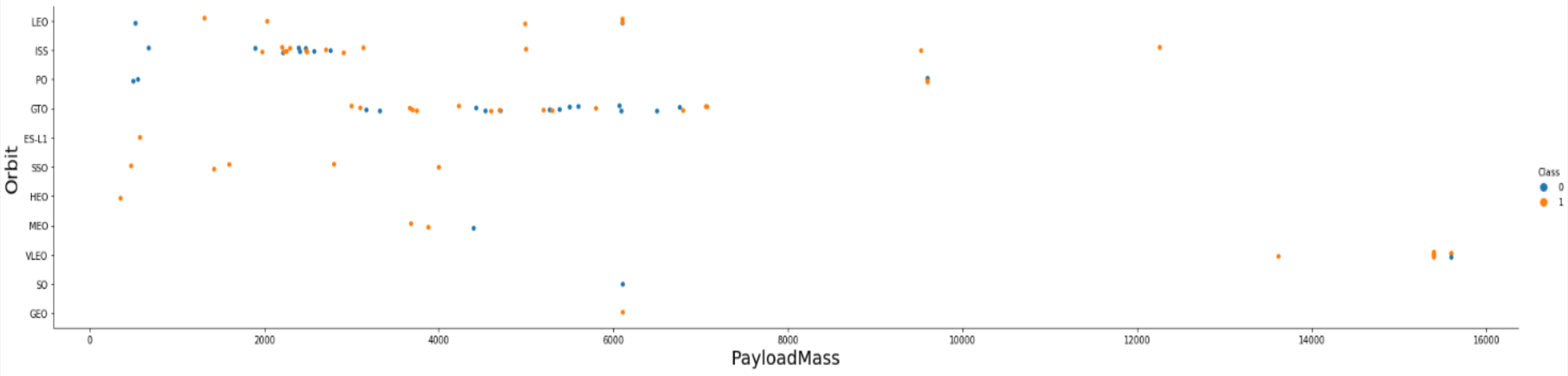
- Successful Launch
- Failed Launch



Payload vs. Orbit Type

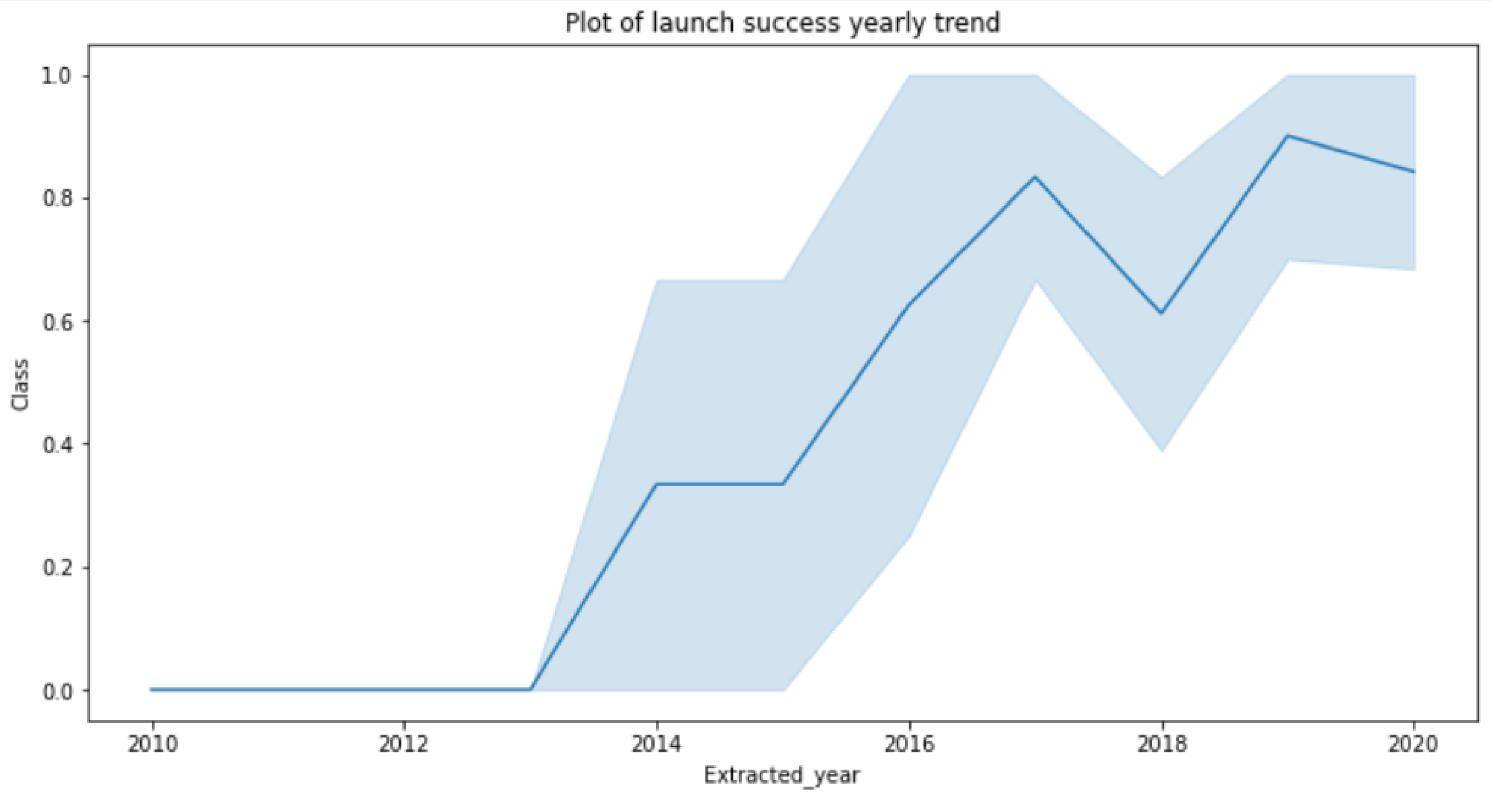
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as positive landing rate and negative landing (unsuccessful mission) are both there here.

- Successful Launch
- Failed Launch



Launch Success Yearly Trend

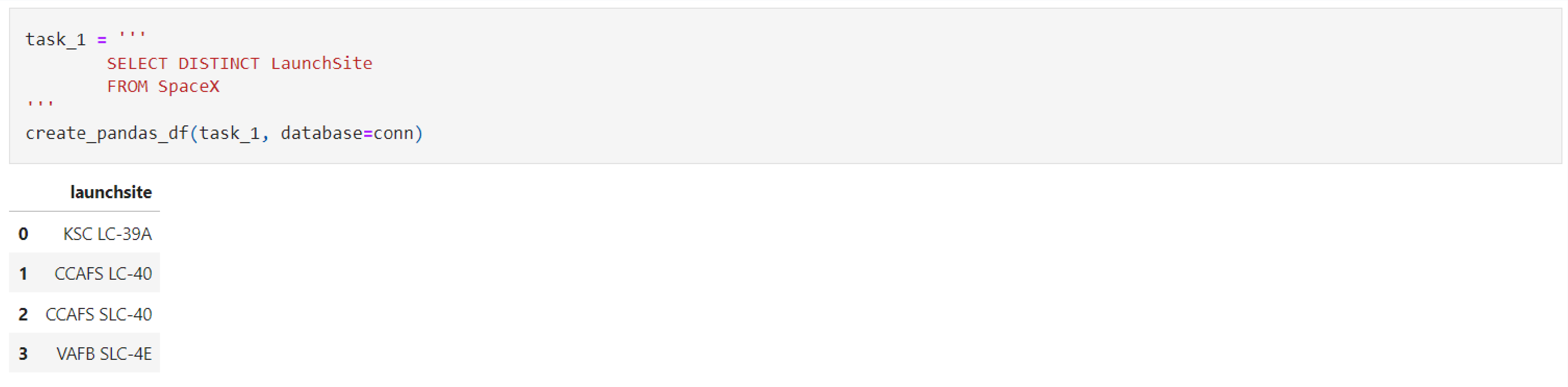
- The sucess rate since 2013 kept increasing till 2020.



All Launch Site Names

- There are four unique launch sites found through SQL request:

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

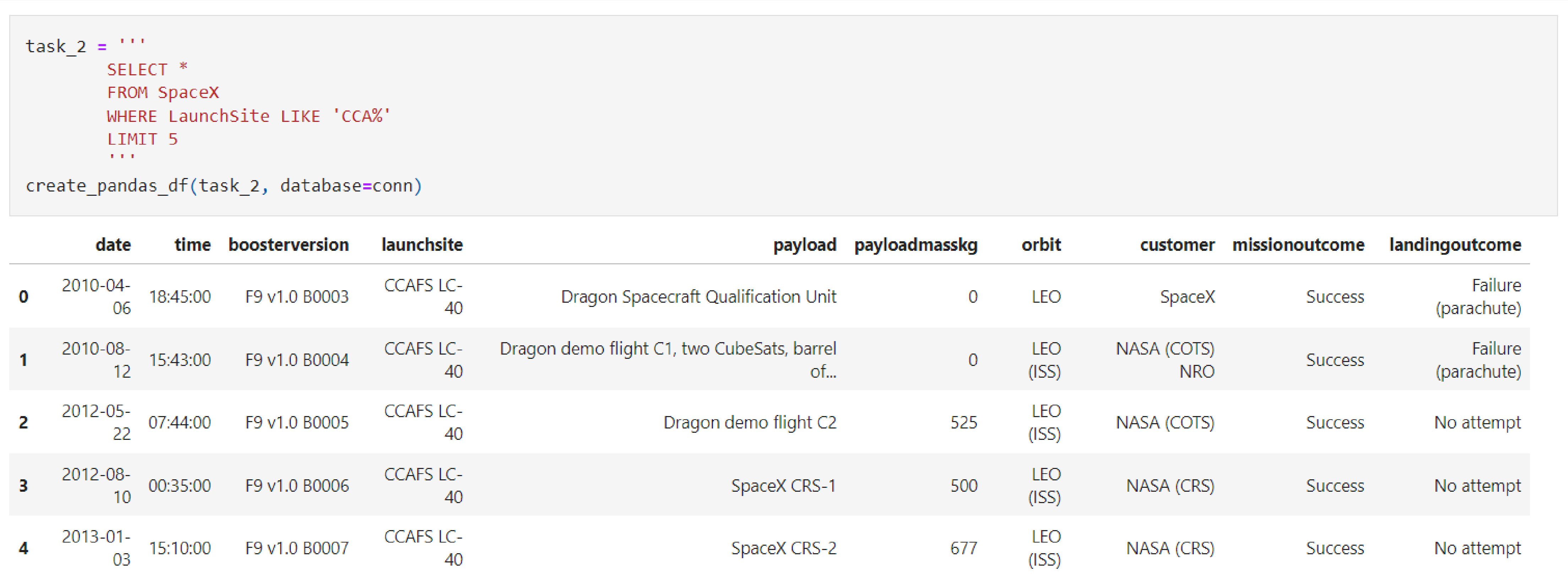


```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:

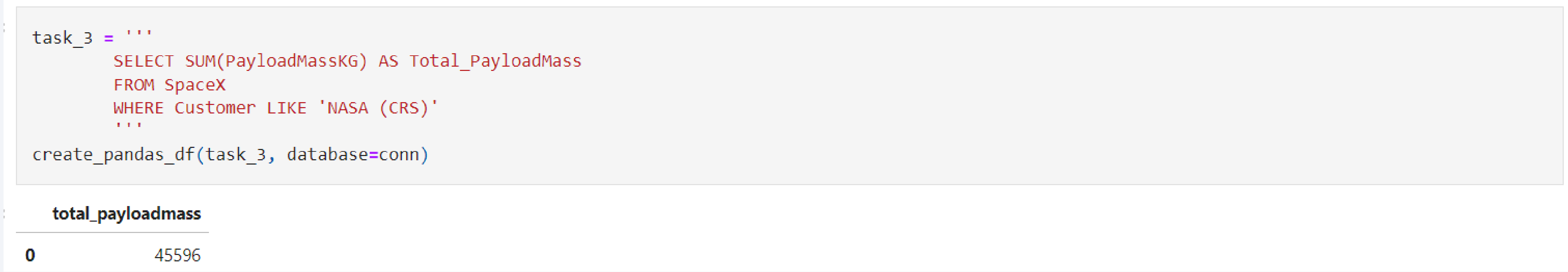


```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

	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

Total Payload Mass

- The total payload carried by boosters from NASA is 45596 kg:

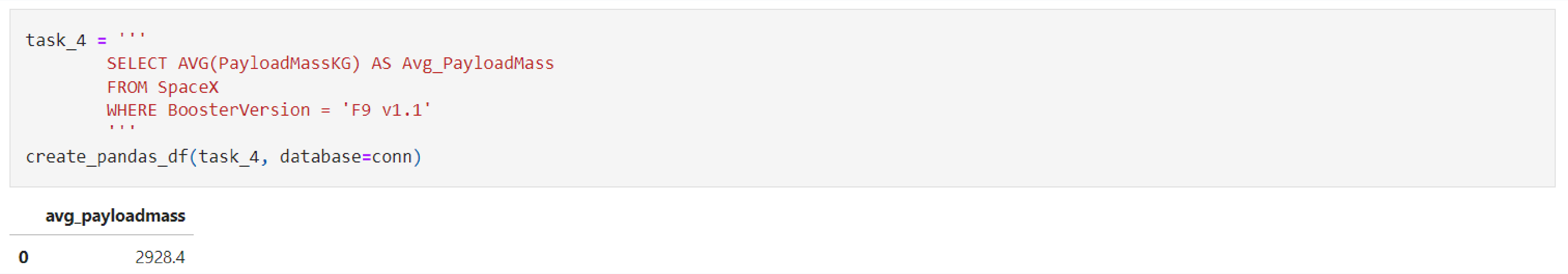


```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

```
   total_payloadmass
0              45596
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg.

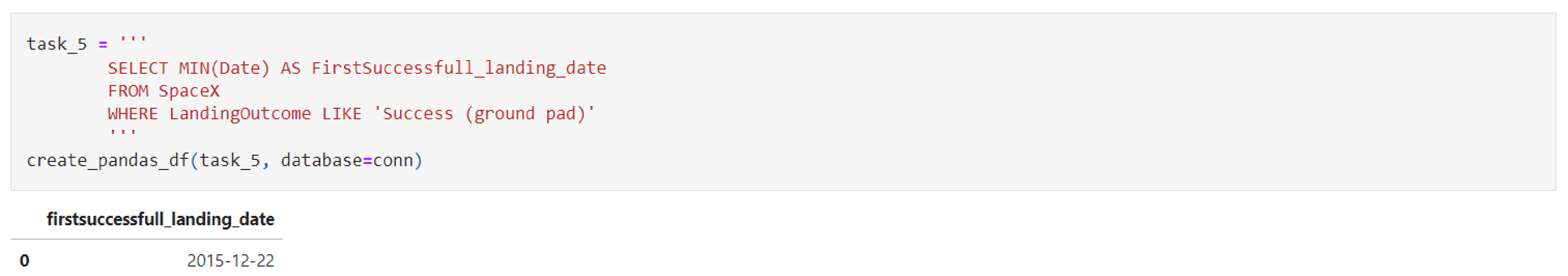


```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

- On 22 Dec, 2022 the first successful landing outcome on ground pad was achieved.

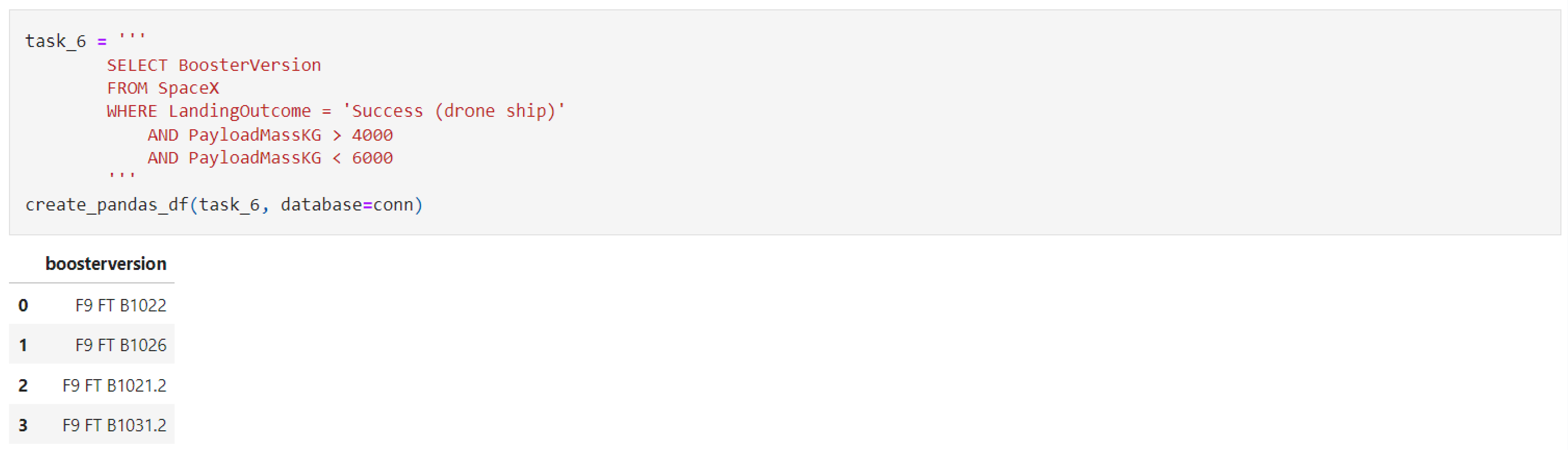


```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

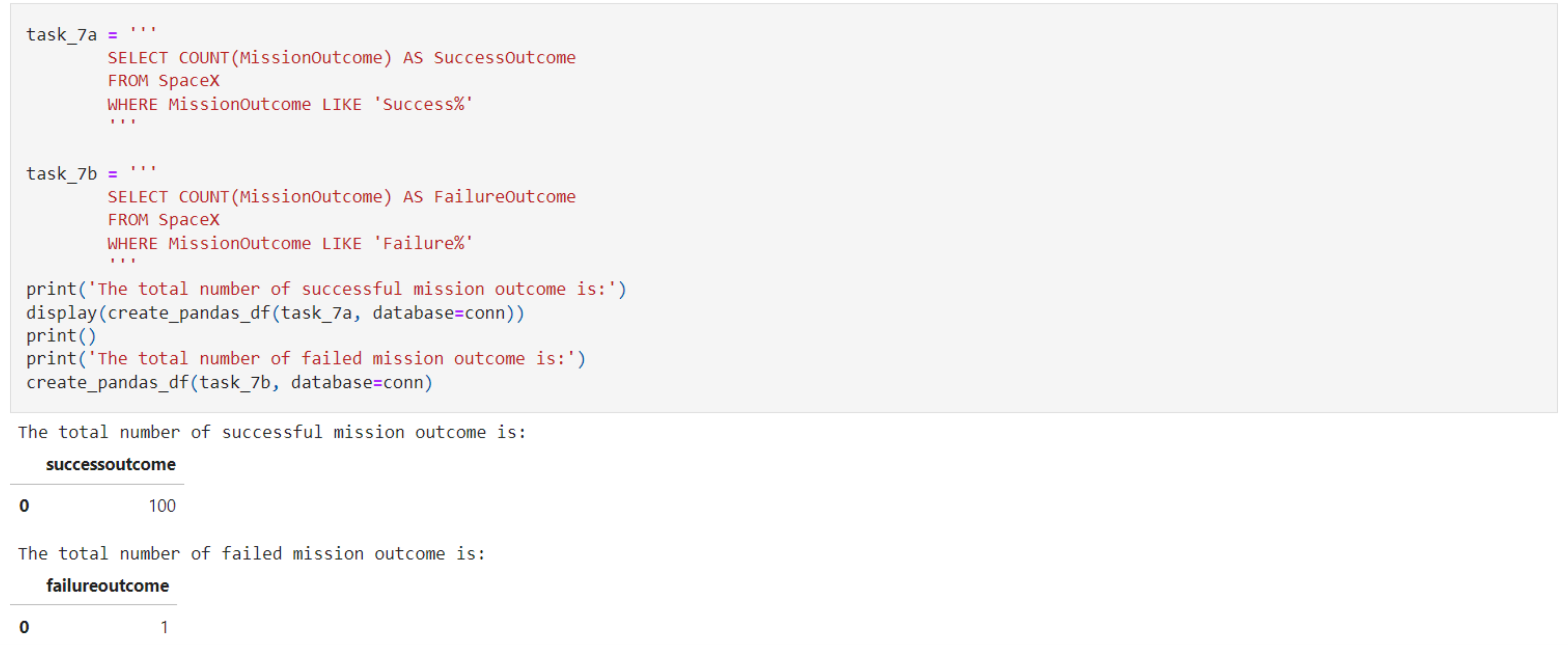


```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

	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

- The total number of successful outcomes is 100, and failure mission outcomes is 1.



```
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:
```

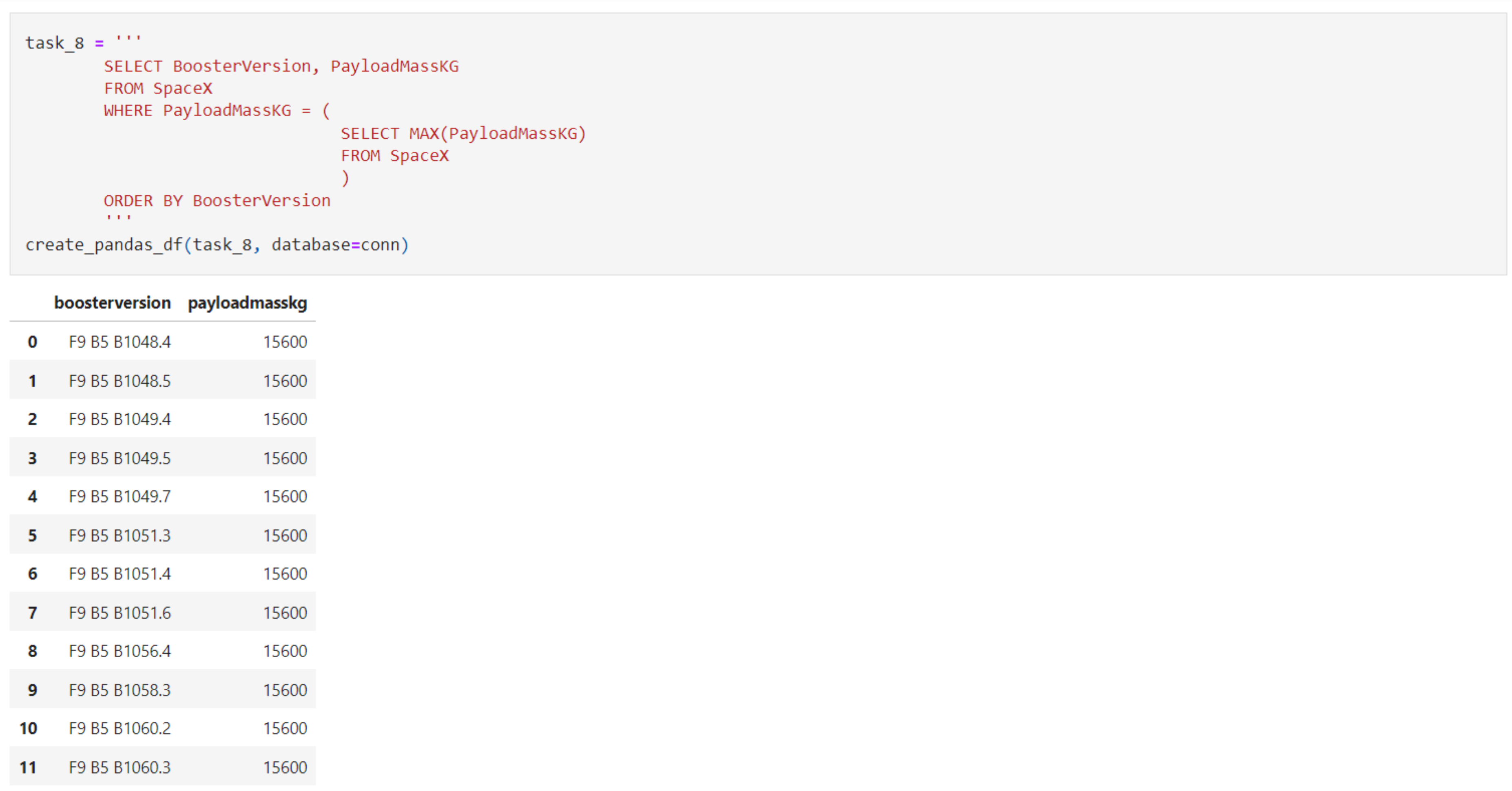
	successoutcome
0	100

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

	failureoutcome
0	1

Boosters Carried Maximum Payload

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

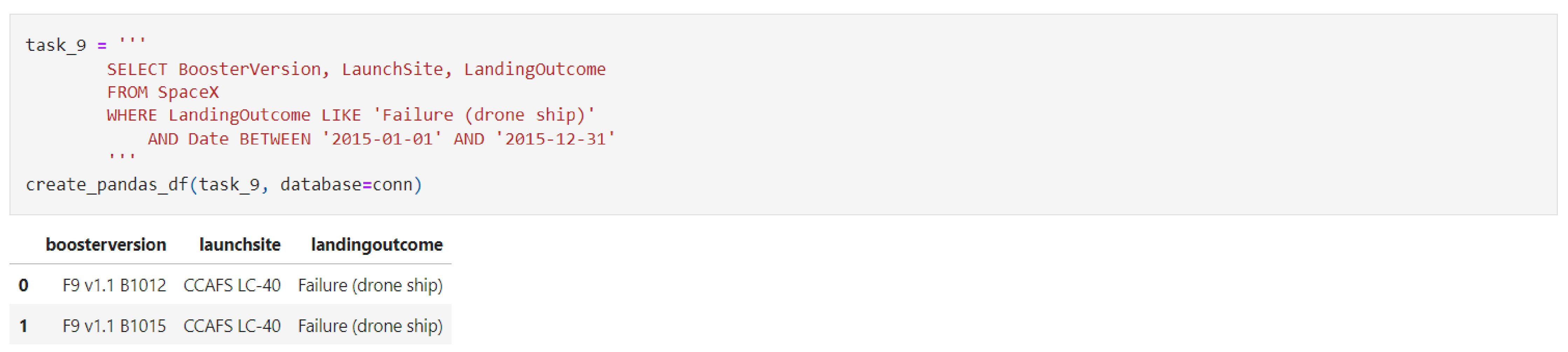


```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8,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

2015 Launch Records

- The failed landing outcomes in drone ship with booster versions, and launch site names for year 2015

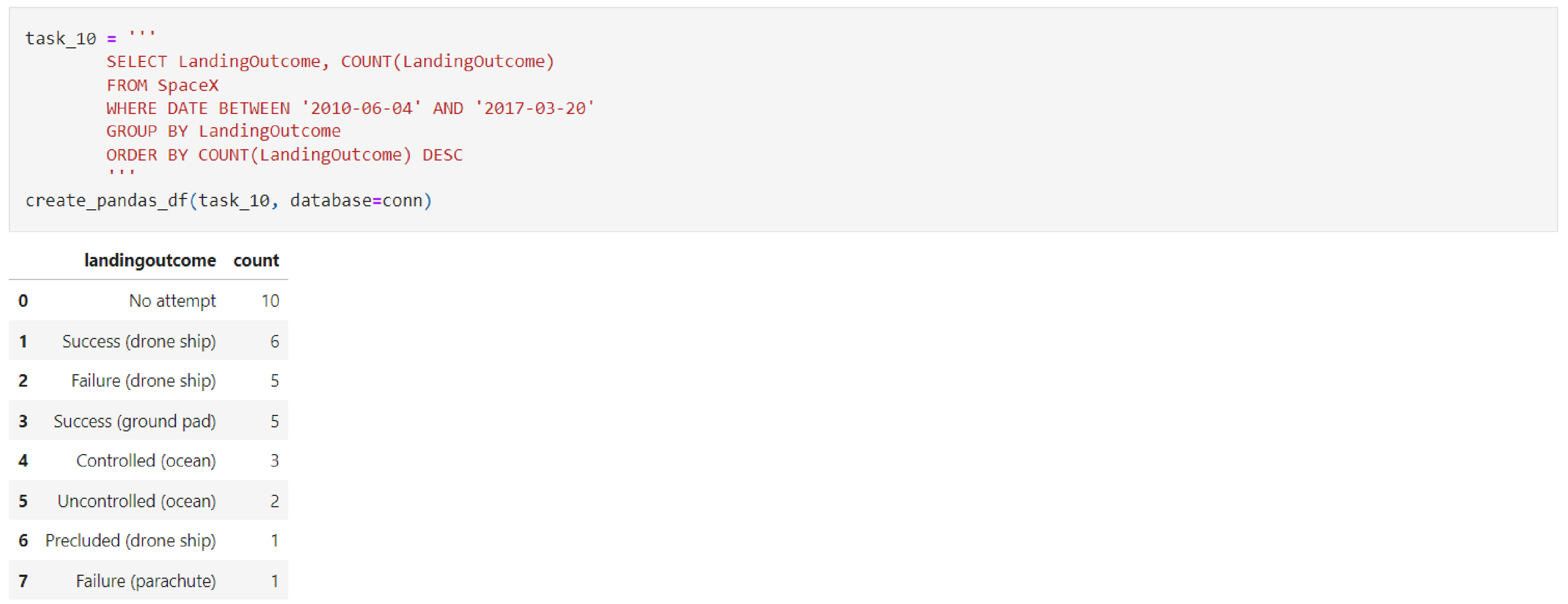


```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

	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

- The rank of landing outcomes 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)
```

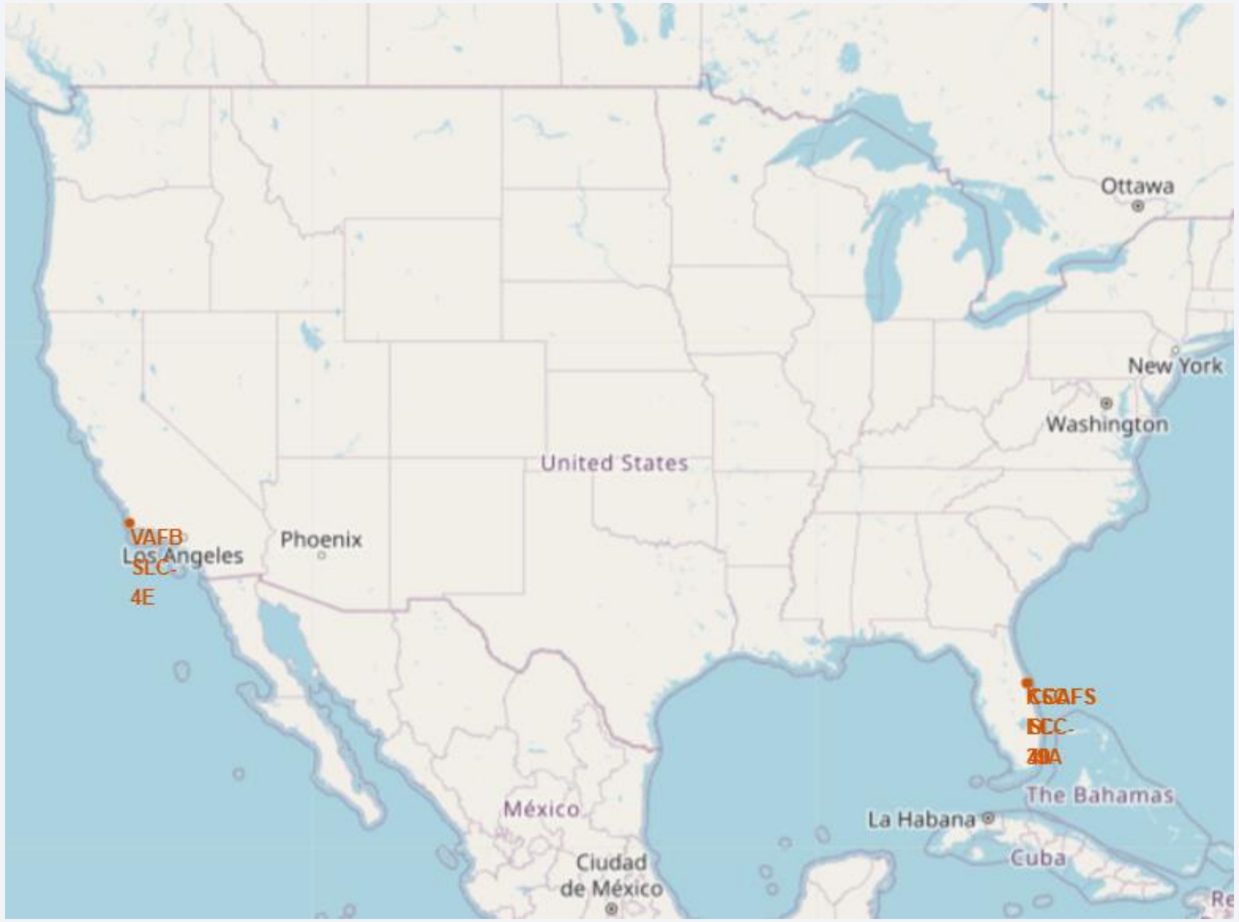
	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

All Launch Sites on the Map

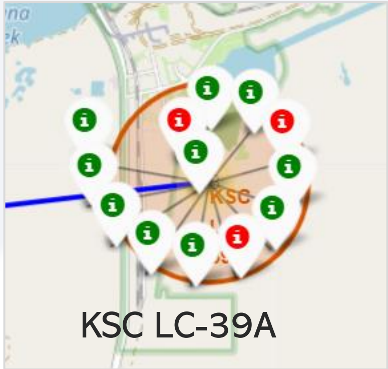
- Launch sites are close to equator, as it reduces fuel consumption when launching.
- As a launched rocket follows the trajectory of Earth's rotation from west to east, in case of a failed launch it drops on the national territory or in the ocean.
- All launch sites are close to the ocean. An ocean launch reduces risks related to launching over populated areas, providing better safety to third parties.
- The launch sites are located equidistant from the NASA's main space center in Texas.

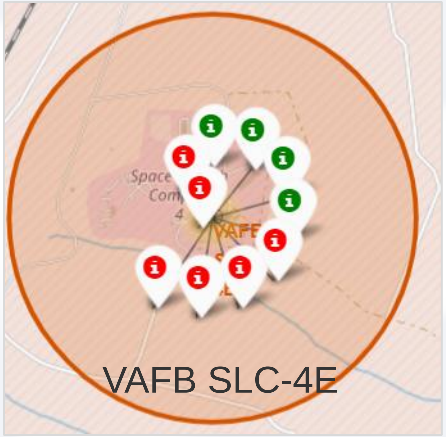


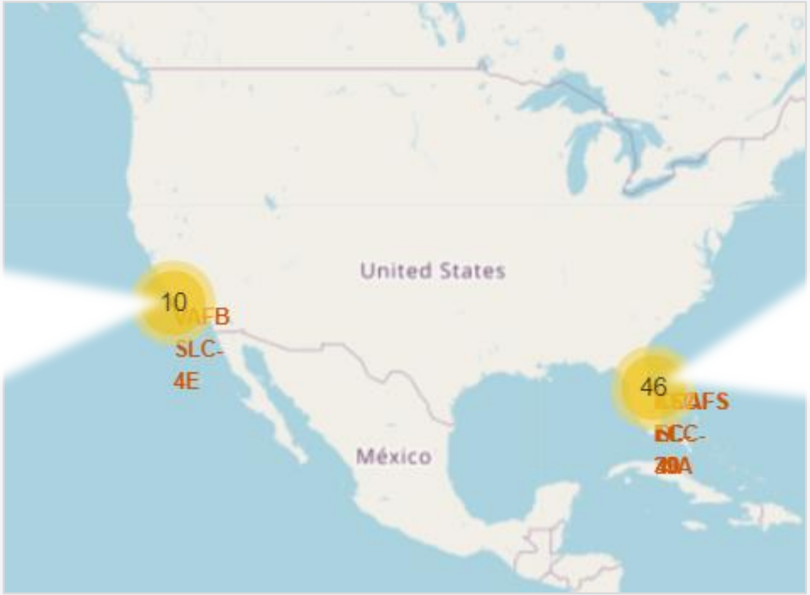
Launch Outcomes on the Map

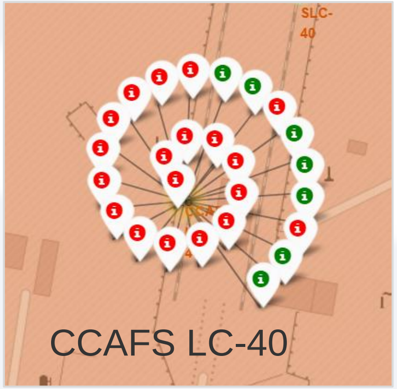
- The colour-labeled launch outcomes on the map represent the amount of successful and unsuccessful launches for each site.
- We can identify which launch sites have relatively high success rates.
- KSC LC-39A launch site has the highest success rate.

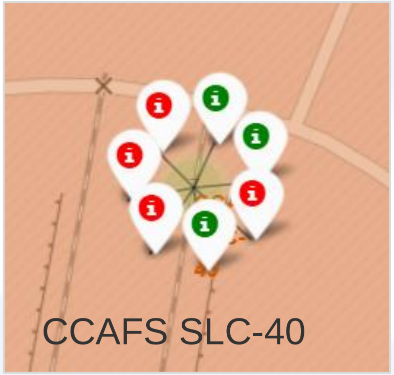
🟢 Successful Launch
🔴 Failed Launch











KLC LC39A Launch Site Proximities

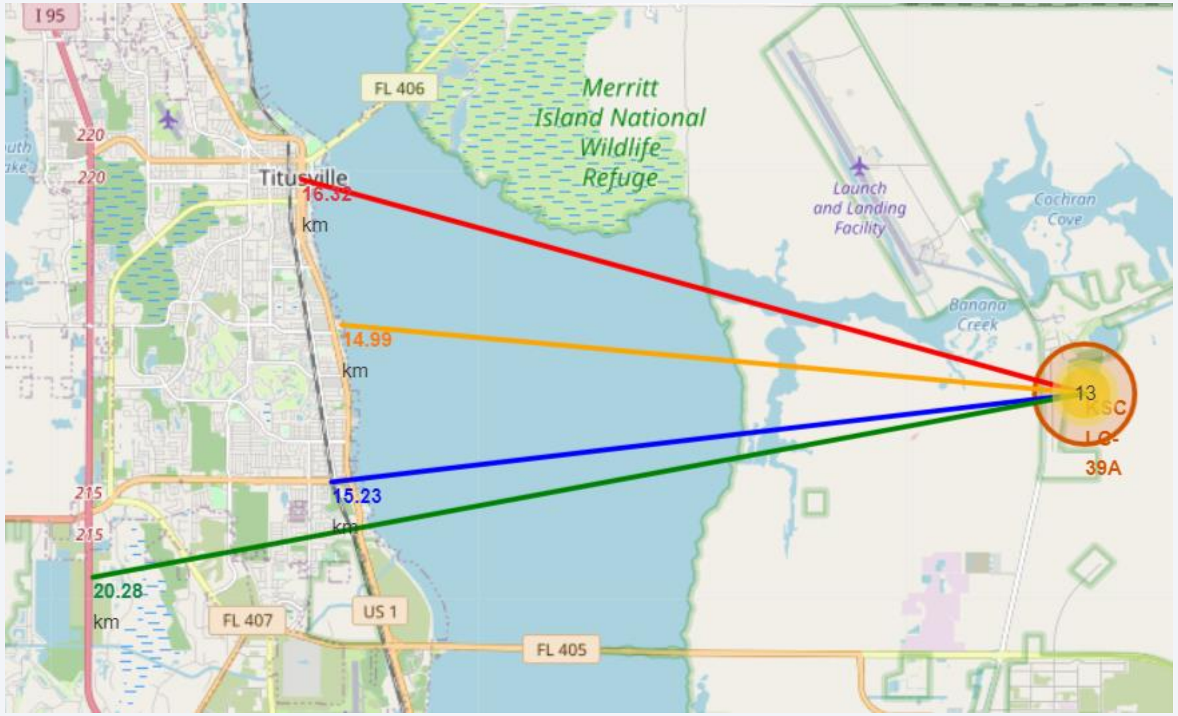
- The proximity of 15-20 km to infrastructure, public transportation and a city may be not sufficient, because of the possible wreckage.

🟥 Proximity to the nearest city (16.32 km)

🟨 Proximity to a coastline (14.99 km)

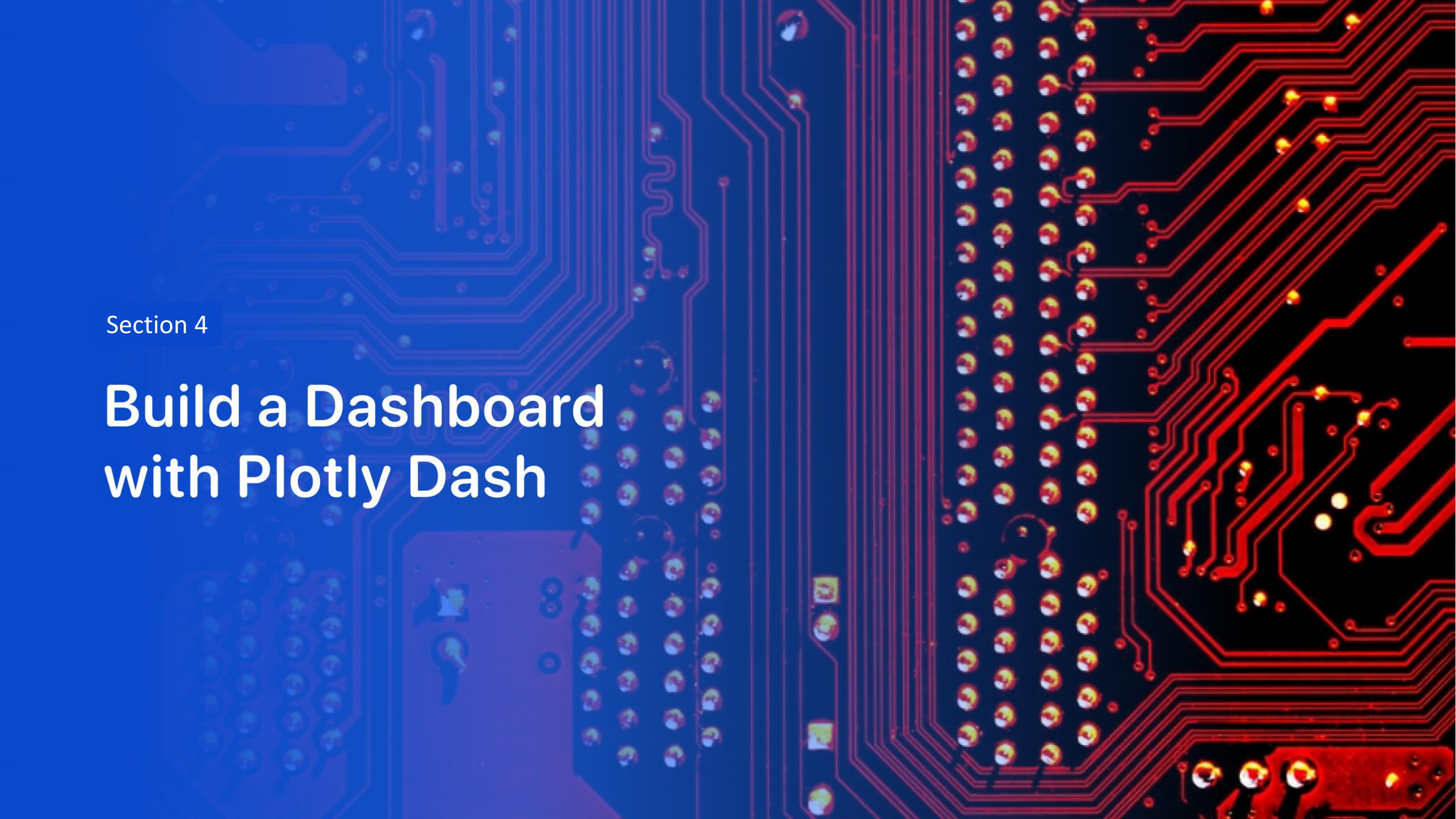
🟦 Proximity to railways (15.23 km)

🟩 Proximity to a highway (20.28 km)



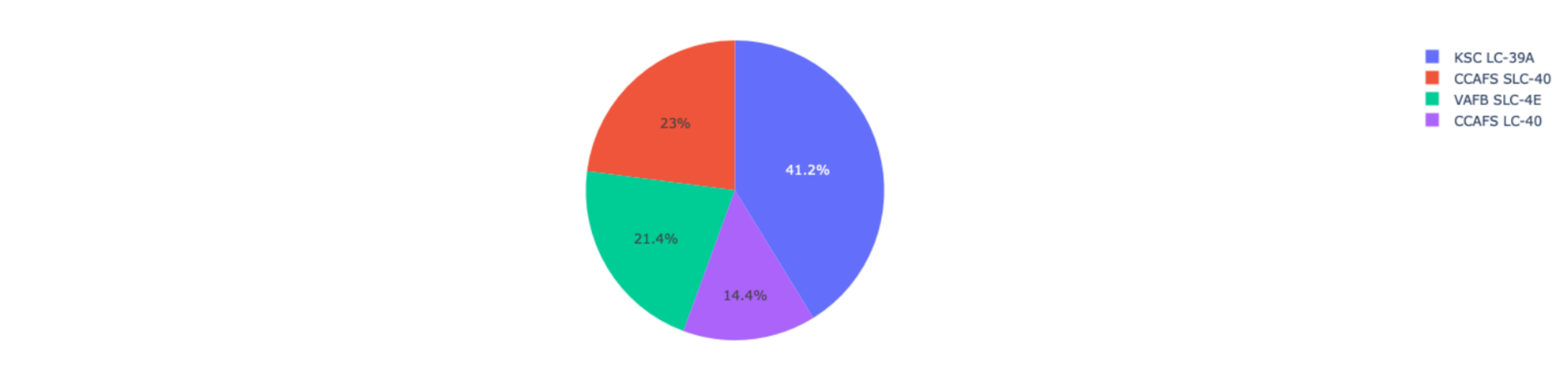
Section 4

Build a Dashboard with Plotly Dash



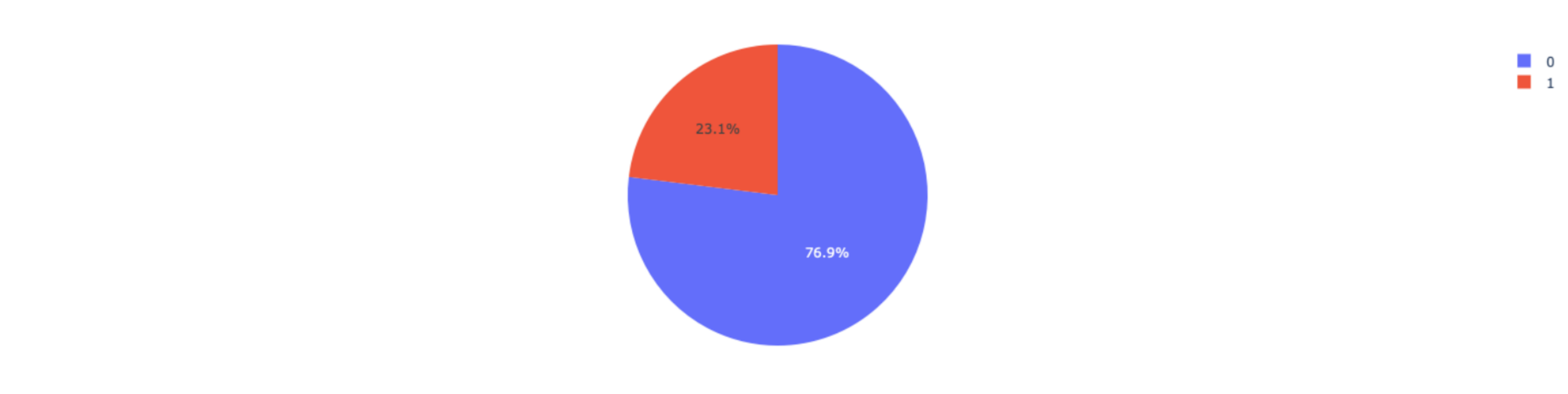
Total Success Launches by Site

- The highest percentage of successful lanches falls in Launch Complex 39 at NASA's Kennedy Space Center in Florida, KSC LC-39A (41.2%).
- The site with less success launches is Space Launch Complex 40 located at Cape Canaveral Space Force Station in Florida, CCAFS LC-40 (14.4%).



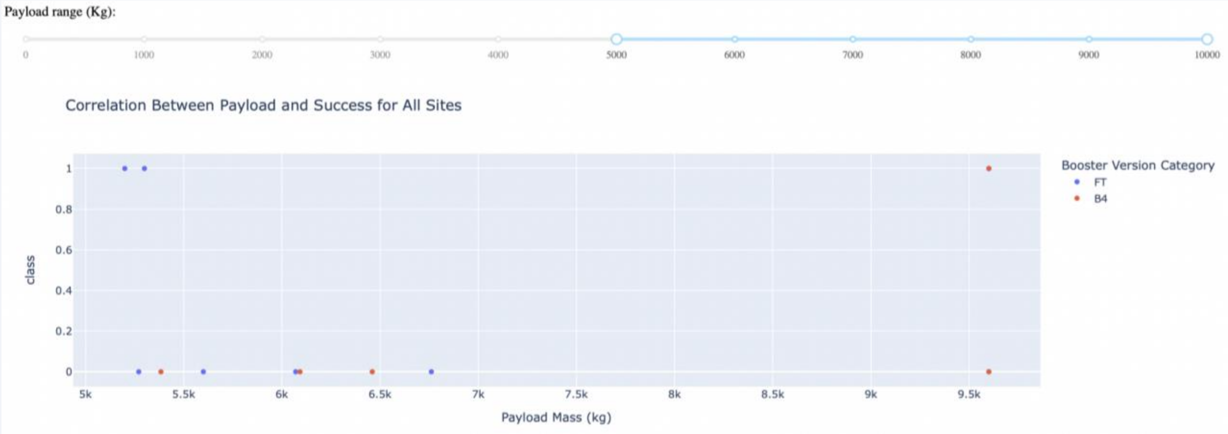
Success Launches for site KSC LC-39A

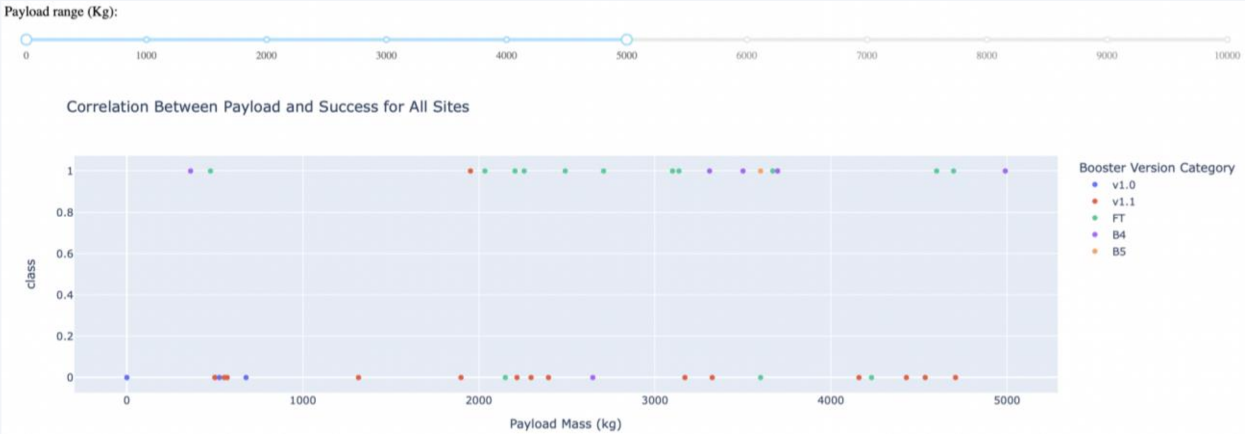
- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.



Payload vs. Launch Outcome for all sites

- Success rates for heavy weighted payloads are lower than for the low weighted payloads.
- In the range of 1800-5000 kg payload, launches with FT Booster Version Category have the highest success rate, while launches with v1.0 Booster Version Category have the lowest success rate.



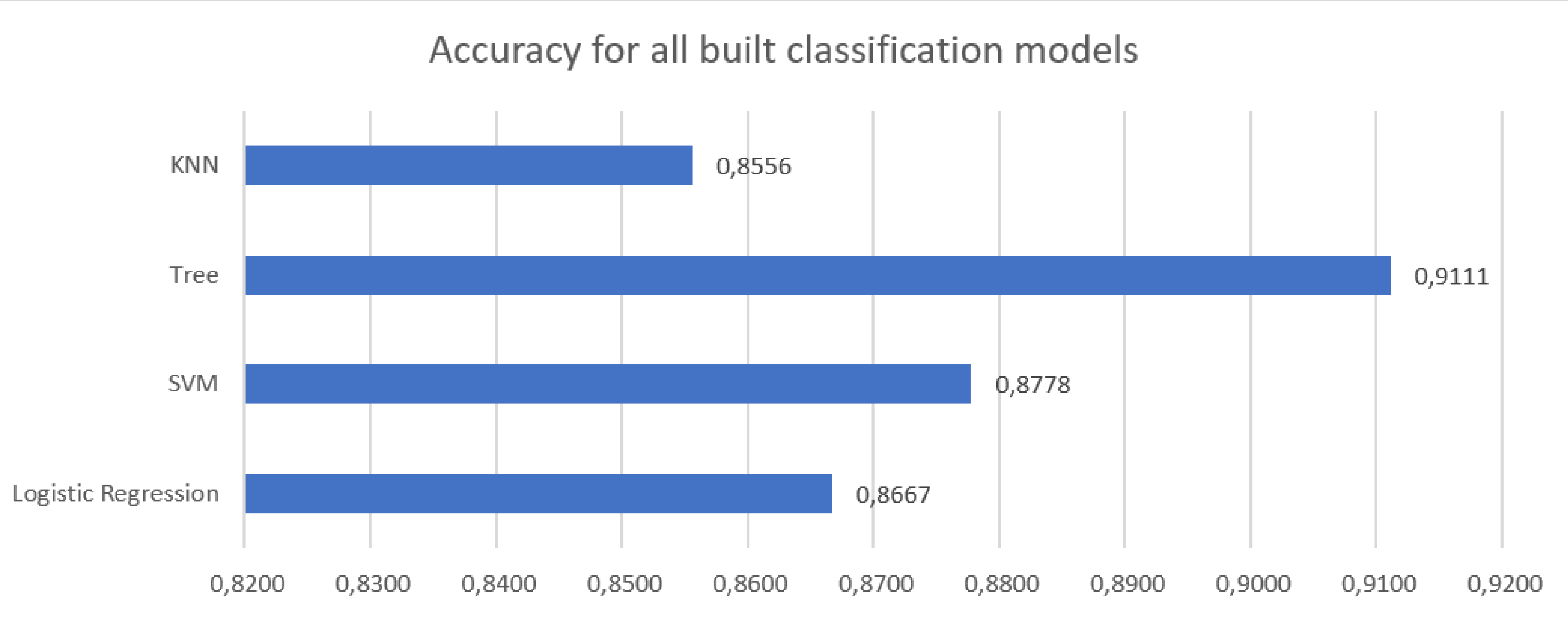


Section 5

Predictive Analysis (Classification)

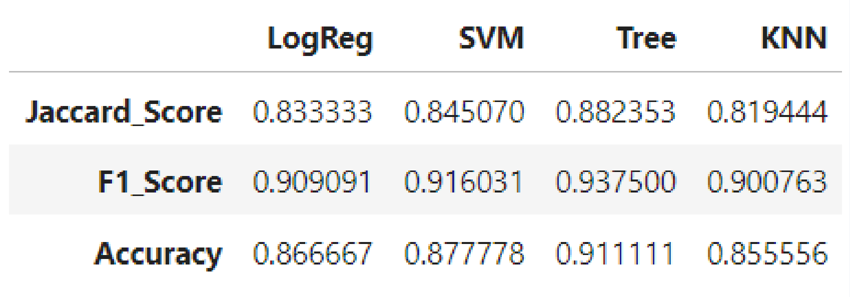
Classification Accuracy

- All built classification models have high classification accuracy.
- The Decision Tree Model is the model with the highest classification accuracy.



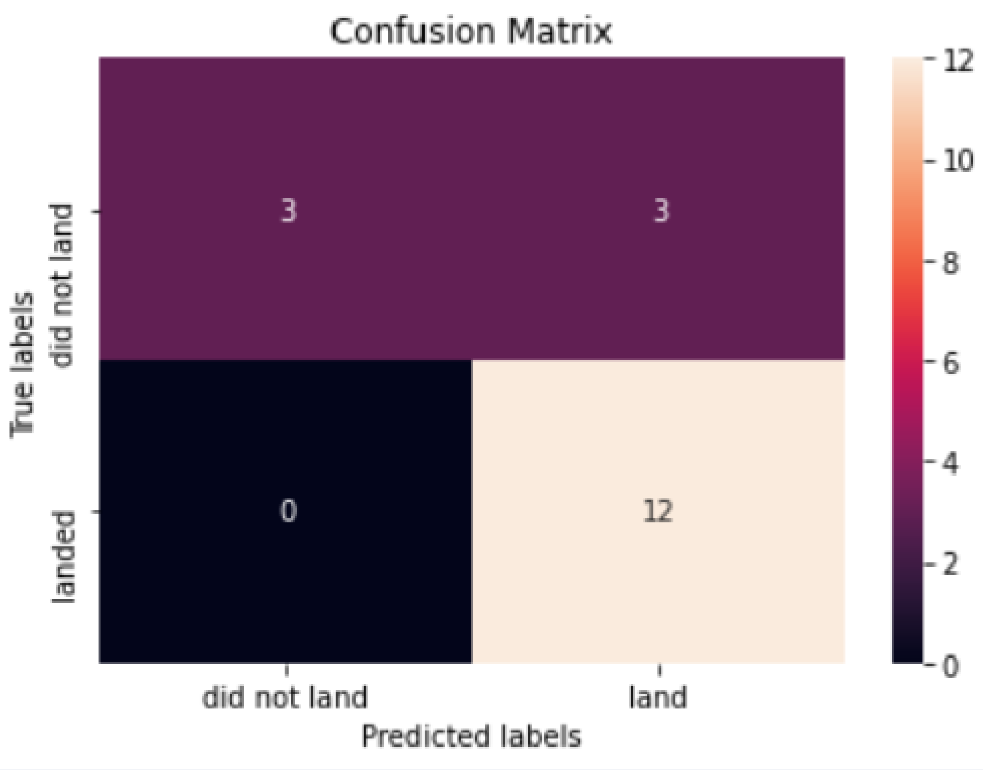
Confusion Matrix of the best performing model

- The Decision Tree Model is the best one by all the parameters.



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The rate of successful landings increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have the most success rate.
- Decision Tree Model is the best algorithm for this dataset.

Appendix

A big thank you to the course team:

Primary Instructors: Joseph Santarcangelo, Yan Luo

Project Lead: Rav Ahuja

Instructional Designer: Lakshmi Holla

Lab Authors: Joseph Santarcangelo, Yan Luo, Azim Hirjani, Lakshmi Holla

Technical Advisor: Yan Luo

Publishing: Grace Barker, Rachael Jones

Project Coordinators: Kathleen Bergner

Narration: Bella West

Video Production: Simer Preet, Lauren Hall, Hunter Bay, Tanya Singh, Om Singh

Teaching Assistants and Forum Moderators: Malika Singla, Duvvana Mrutyunjaya Naidu, Lakshmi Holla, Anita Verma

Thank you!