IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sri Yenigalla
11/19/2022

Outline

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

Executive Summary

- In this assignment, we will predict if the Falcon 9 first stage will land successfully. SpaceX 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 due to the fact that SpaceX can reuse the first stage. We analyzed the success rate of Falcon9 launches with data collected from public SpaceX API and SpaceX Wikipedia page. Explored data using SQL, visualization, folium maps, and Plotly dashboards. Relevant columns were used as features by changing all categorical variables into binary using one hot encoding. Standardized data and used GridSearchCV to find best parameters for machine learning models.
- Following machine learning models were built and tested: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. All produced similar results with accuracy rate of about 83.33%. All models over predicted successful landings. More data is needed for better model determination and accuracy.

Introduction

Context

- SpaceX made major leaps in efficient rocket launches using Falcon 9
- Major savings with a cost of 62 million dollars per launch vs other providers costing 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.

Problem

- Space Y wants to use machine learning techniques to predict the success of Stage 1 recovery

Section 1

Methodology

Methodology

Executive Summary

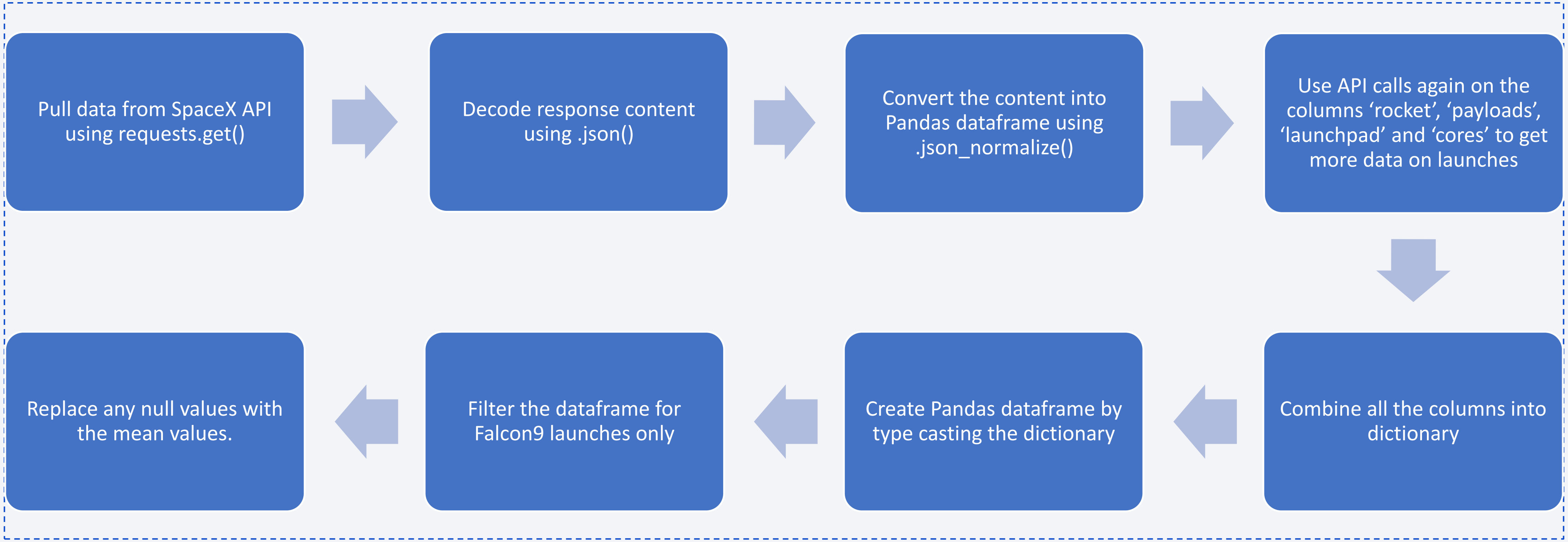
- Data collection methodology:
  - API calls were made to Space X public API and web scraping from Space X's Wikipedia.
- Perform data wrangling
  - Training labels were used to convert categorical data into numeric for regression analysis along with creation of adding columns 'Mission Outcome' and 'Landing Location'
- 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

- API requests were made to Space X public API and tables were sourced from Wikipedia HTML page using web scraping.
- Requested data was cleaned, parsed and converted into a Pandas dataframe with columns as shown below from each source.

Space X API	Web scraping from Wikipedia
BoosterVersion	Flight No.
PayloadMass	Launch Site
Orbit	Payload
LaunchSite	Payload mass
Outcome	Orbit
Flights	Customer
GridFins	Launch outcome
Reused	Version Booster
Legs	Booster landing
LandingPad	Date
Block	Time
ReusedCount	
Serial	
Longitude	
Latitude	

Data Collection – SpaceX API



Pull data from SpaceX API using requests.get()

Decode response content using .json()

Convert the content into Pandas dataframe using .json_normalize()

Use API calls again on the columns ‘rocket’, ‘payloads’, ‘launchpad’ and ‘cores’ to get more data on launches

Combine all the columns into dictionary

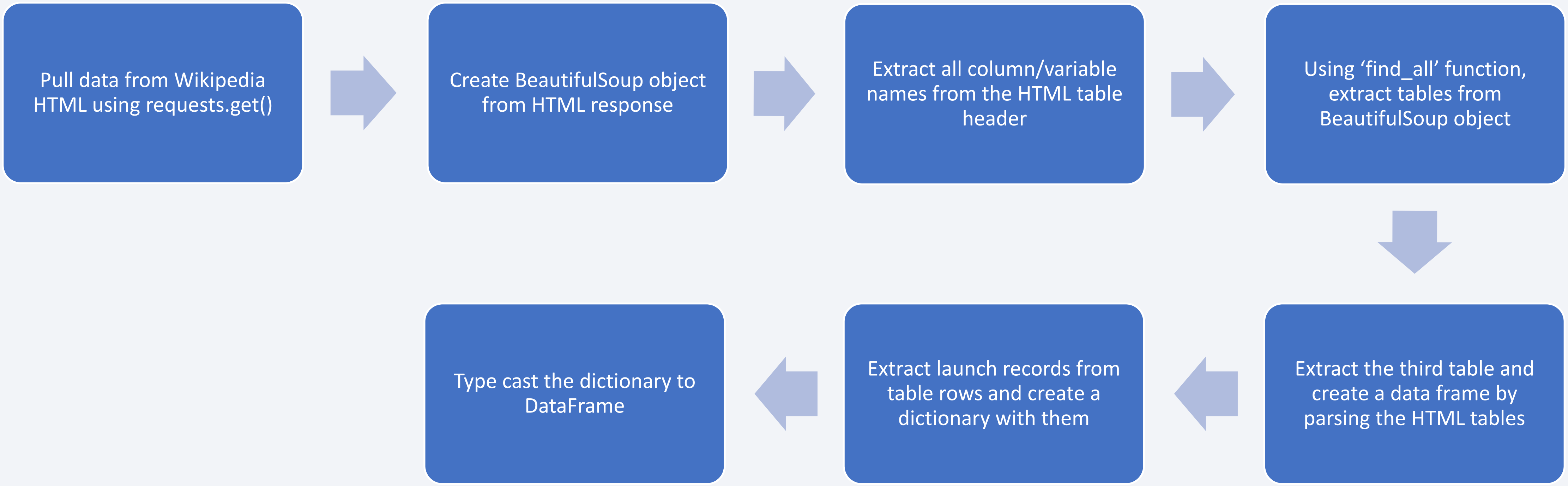
Create Pandas dataframe by type casting the dictionary

Filter the dataframe for Falcon9 launches only

Replace any null values with the mean values.

GitHub URL: https://github.com/Raghava33/DataScience-Certification-Capstone-project/blob/799293333451cabb741c6e5de768d1d4bd4c1620/Week%201/Data%20Collection%20API.ipynb

Data Collection - Scraping



Pull data from Wikipedia HTML using requests.get()

Create BeautifulSoup object from HTML response

Extract all column/variable names from the HTML table header

Using ‘find_all’ function, extract tables from BeautifulSoup object

Extract the third table and create a data frame by parsing the HTML tables

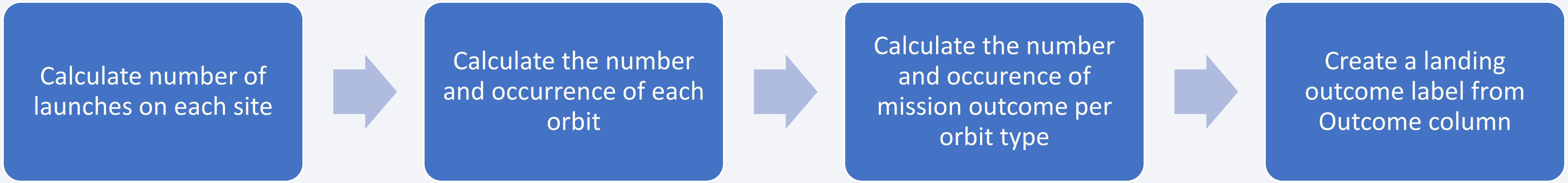
Extract launch records from table rows and create a dictionary with them

Type cast the dictionary to DataFrame

GitHub URL: https://github.com/Raghava33/DataScience-Certification-Capstone-project/blob/799293333451cabb741c6e5de768d1d4bd4c1620/Week%201/Data%20collection%20web%20scraping.ipynb

Data Wrangling

- In the data set there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, **True Ocean** means the mission outcome was successfully landed to a specific region of the ocean while **False Ocean** means the mission outcome was unsuccessfully landed to a specific region of the ocean. **True RTLS** means the mission outcome was successfully landed to a ground pad **False RTLS** means the mission outcome was unsuccessfully landed to a ground pad. **True ASDS** means the mission outcome was successfully landed on a drone ship **False ASDS** means the mission outcome was unsuccessfully landed on a drone ship.
- In this section, we will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.



Calculate number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome per orbit type

Create a landing outcome label from Outcome column

GitHub URL: https://github.com/Raghava33/DataScience-Certification-Capstone-project/blob/799293333451cabb741c6e5de768d1d4bd4c1620/Week%201/EDA%20lab.ipynb

EDA with Data Visualization

Exploratory Data Analysis with Visualization was performed on: Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

Following variable relationships were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend

Scatter plots, line charts, and bar plots were used to explore relationships between variables to see if they are fit for training the machine learning models

GitHub URL: https://github.com/Raghava33/DataScience-Certification-Capstone-project/blob/799293333451cabb741c6e5de768d1d4bd4c1620/Week%202/EDA%20with%20Visualization.ipynb

EDA with SQL

- Load data into IBM database and query the data into the notebook using the IBM DB2 credential information and implement queries using SQL magic. Following queries were performed to understand the data through EDA.
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass, using subquery
- 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
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

GitHub URL: https://github.com/Raghava33/DataScience-Certification-Capstone-project/blob/6ed5158dd8d3a85502aaebbd36c8edc6494970a4/Week%202/EDA%20with%20SQL%20skillslab.ipynb

Build an Interactive Map with Folium

- Circles were used to mark the launch site locations. Markers were used to display more labels once we goto a particular launch site. In case of multiple sites in nearby locations, Marker cluster object was used to display markers as we zoom into the map.
- Marker color was used to distinguish between successes and failures linked to launch sites.
- MousePosition was used to interactively locate nearby sites such as railway lines, coastlines, etc.
- PolyLine was used to draw lines between two points and distance between those points was labeled using an attribute of Marker.

The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors, so we added above objects to discover some of the factors by analyzing the existing launch site locations.

GitHub URL: https://github.com/Raghava33/DataScience-Certification-Capstone-project/blob/111a33a5ea3d41ba4fc9185a5659196c7982b050/Week%203/Interactive%20map%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Pie chart and scatter plot were used in the dashboard.

Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.

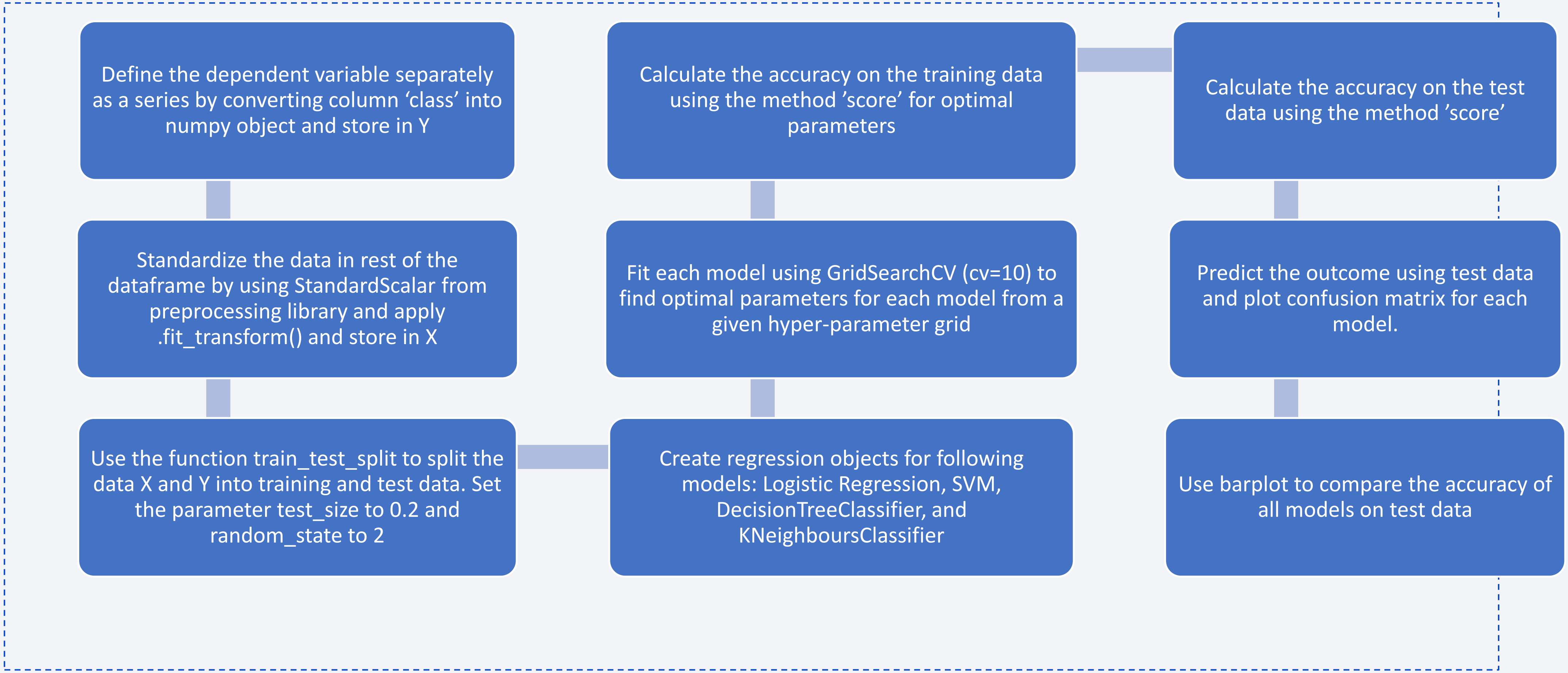
Scatter plot takes two inputs: Site (All or individual) and payload mass

Scatter plot also limits the data based on the slider selection between 0 and 10000 kg.

The pie chart visualizes the launch site success rate, whereas the scatter plot shows how success varies across launch sites, payload mass, and booster version category.

GitHub URL: https://github.com/Raghava33/DataScience-Certification-Capstone-project/blob/111a33a5ea3d41ba4fc9185a5659196c7982b050/Week%203/spacex_dash_app.py

Predictive Analysis (Classification)



Define the dependent variable separately as a series by converting column ‘class’ into numpy object and store in Y

Standardize the data in rest of the dataframe by using StandardScalar from preprocessing library and apply .fit_transform() and store in X

Use the function train_test_split to split the data X and Y into training and test data. Set the parameter test_size to 0.2 and random_state to 2

Create regression objects for following models: Logistic Regression, SVM, DecisionTreeClassifier, and KNeighboursClassifier

Fit each model using GridSearchCV (cv=10) to find optimal parameters for each model from a given hyper-parameter grid

Calculate the accuracy on the training data using the method ’score’ for optimal parameters

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

Predict the outcome using test data and plot confusion matrix for each model.

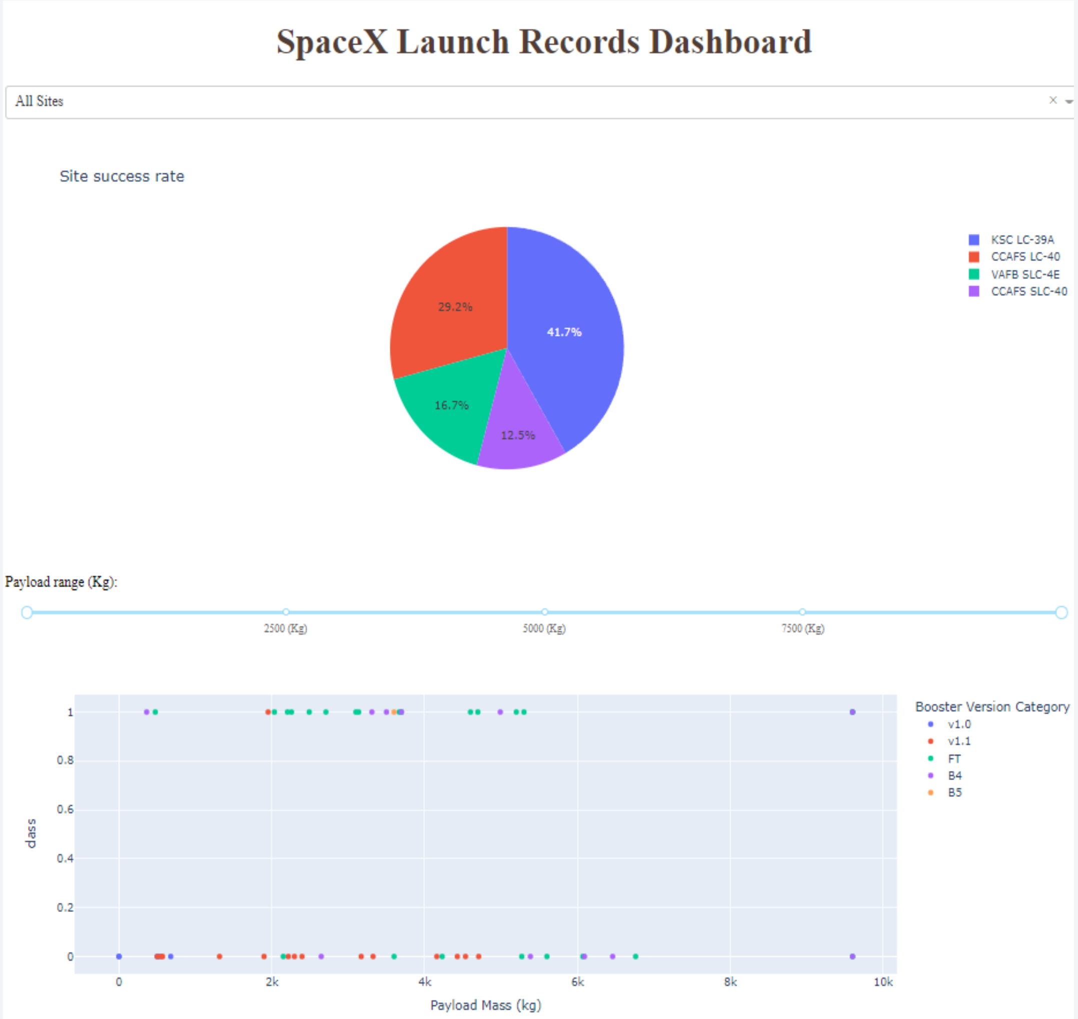
Use barplot to compare the accuracy of all models on test data

GitHub URL: https://github.com/Raghava33/DataScience-Certification-Capstone-project/blob/111a33a5ea3d41ba4fc9185a5659196c7982b050/Week%204/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

Results for EDA and all other sections are shown in following slides and insights are shared in conclusions slide

Model	Accuracy
Logistic Regression	83.3%
Support Vector Machine	83.3%
Decision tree classifier	77.8%
K Nearest Neighbours	83.3%



SpaceX Launch Records Dashboard

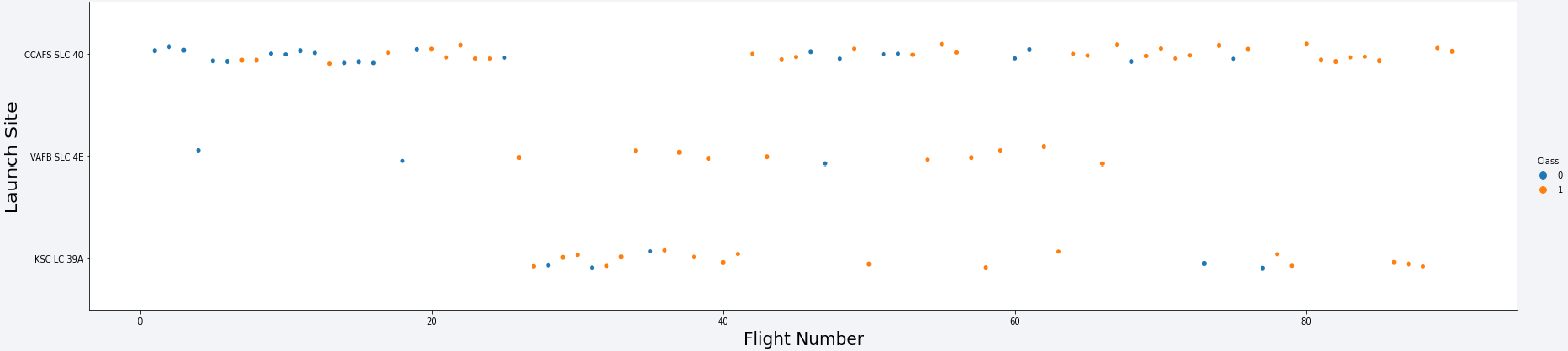
All Sites

Payload range (Kg):

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

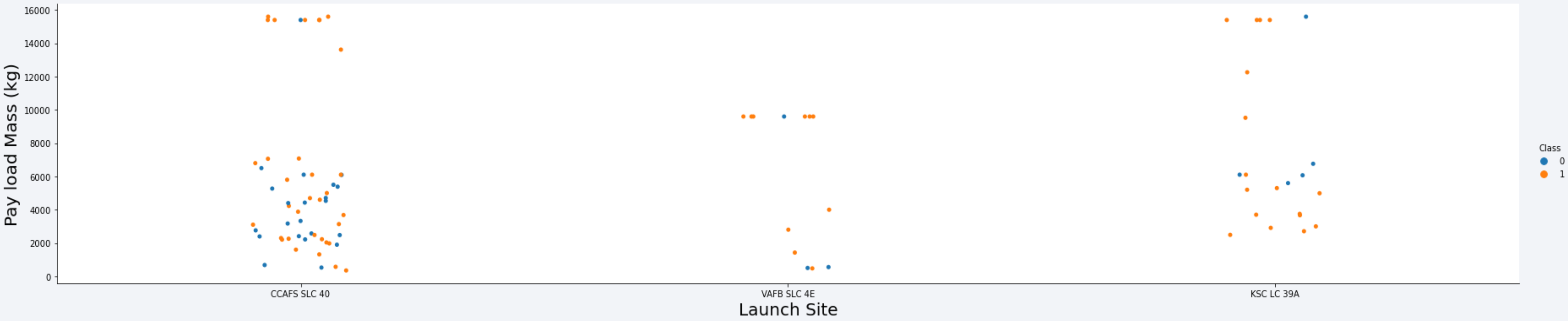


Successes increased gradually with major regime change after around flight #25. Seems like the techniques were refined until flight #80 whereafter there were no failures

Most of the flights were launched from ‘CCAFS SLC 40’, except the flight numbers between 25-40, where new launch sites were tried.

Most of the flights between 25-40 were launched from ‘KSC LC 39A’ and most of them were success. ‘VAFB SLC 4E’ had 100% success rate for these flights.

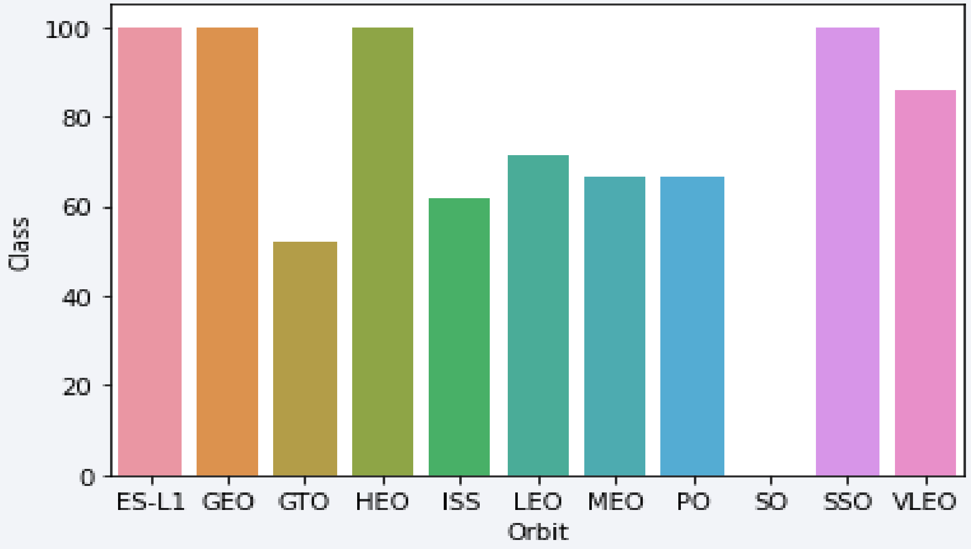
Payload vs. Launch Site



Success rate is higher for pay load mass above 8000kg for all launch sites.

Launch sites ‘VAFB SLC 4E’ and ‘KSC LC 39A’ have higher success rate than ‘CCAFS SLC 40’. This could be because most of the early launches were from ‘CCAFS SLC 40’.

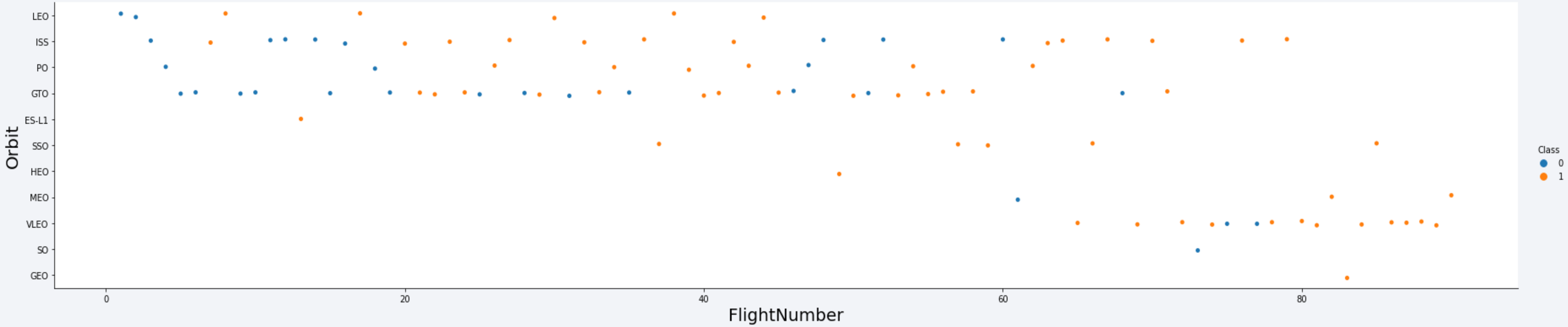
Success Rate vs. Orbit Type



Above bar chart shows the success as a percentage.

Only four orbits have 100% success rate and one has 0%. Most of the remaining orbits have around 60% success rate

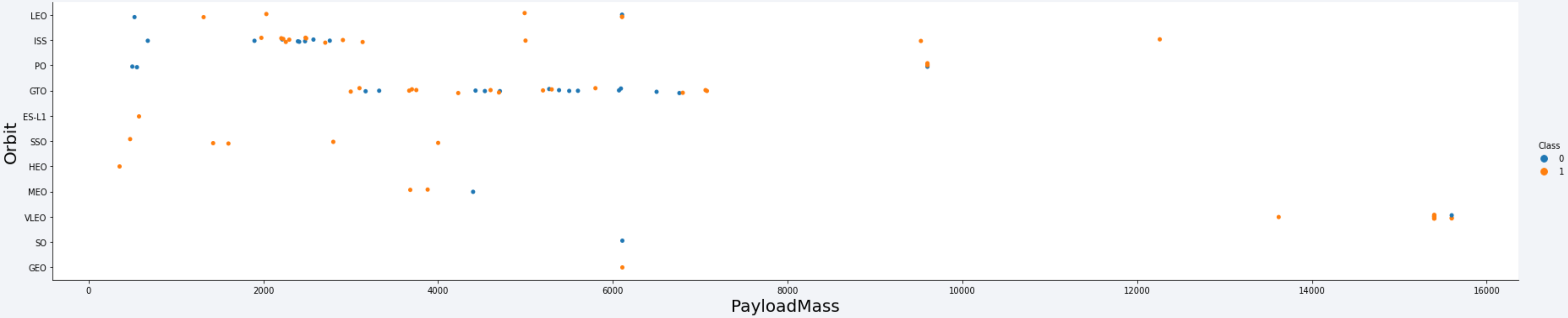
Flight Number vs. Orbit Type



Most launches were made to the orbits ‘LEO’, ‘ISS’, ‘PO’, ‘GTO’, with gradually expanding to other orbits with major trend change after flight #60.

Successes also increased gradually after flight #50 and entering into new orbits. There weren’t as many failures early in the new orbits unlike ones below flight #20. Seems like Space X adopted to the new orbits well from past experience and refinement.

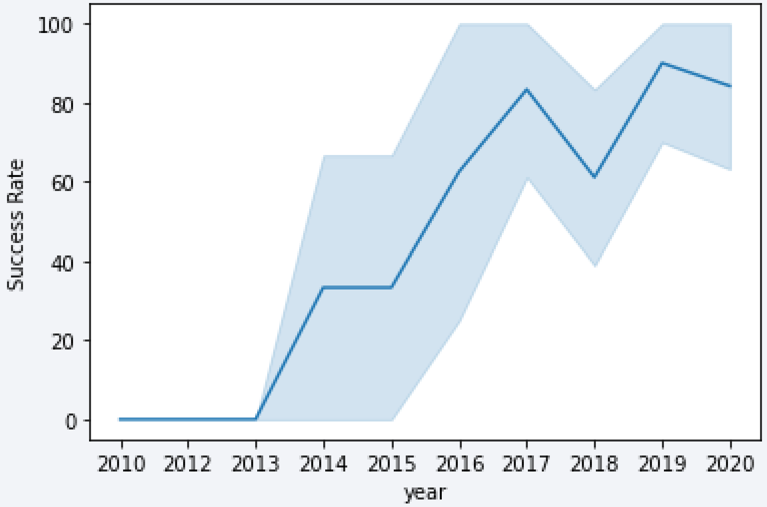
Payload vs. Orbit Type



Orbits are related to Payload mass. For example, SSO, MEO and HEO have payload masses below 5000kg, where VLEO has a mass only above 13,000kg.

GTO's payload is below 7000kg

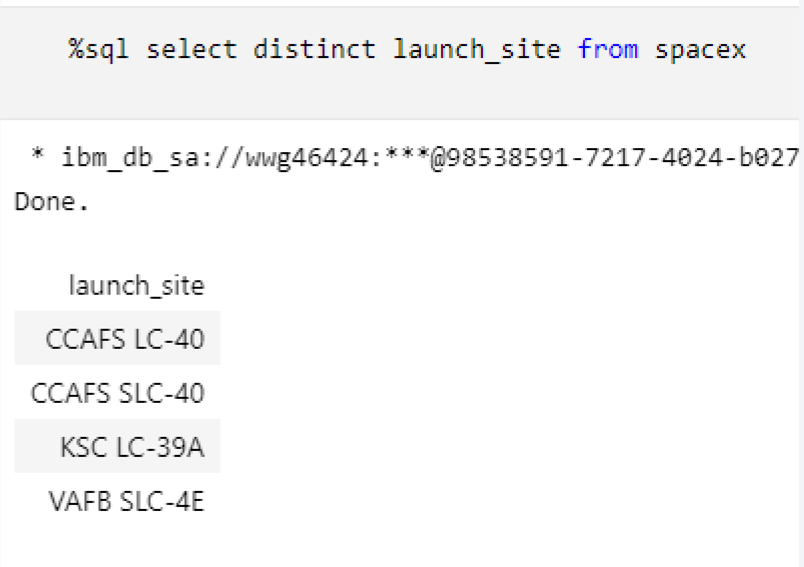
Launch Success Yearly Trend



Success rate kept increasing since 2013, until 2020 with a small dip in 2018

All Launch Site Names

- CCAFS SLC-40, CCAFS LC-40, KSC LC-39A, VAFB SLC-4E
- Use distinct function

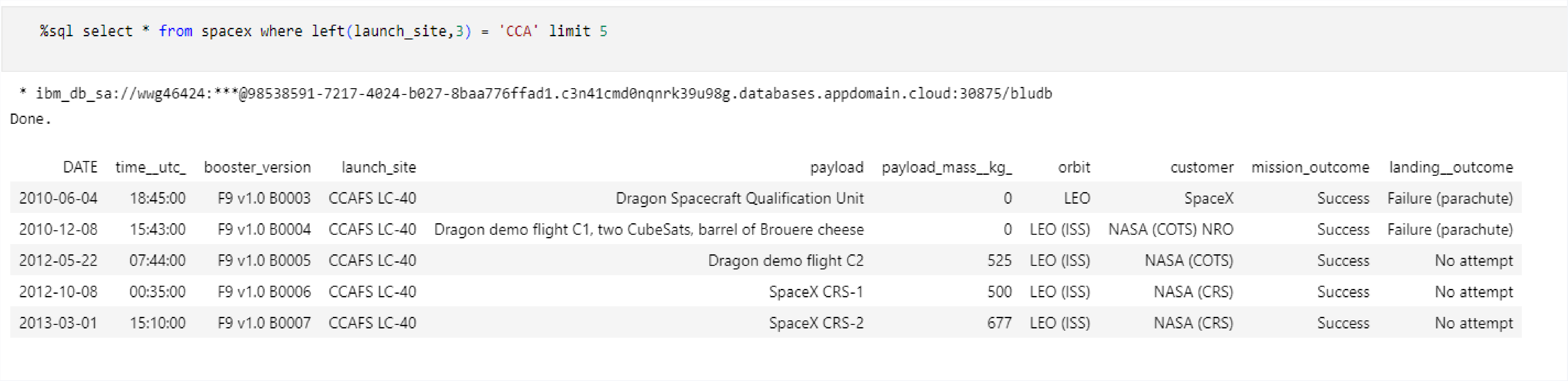


```
%sql select distinct launch_site from spacex
```

```
 * ibm_db_sa://wwg46424:***@98538591-7217-4024-b027
Done.

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

Launch Site Names Begin with 'CCA'



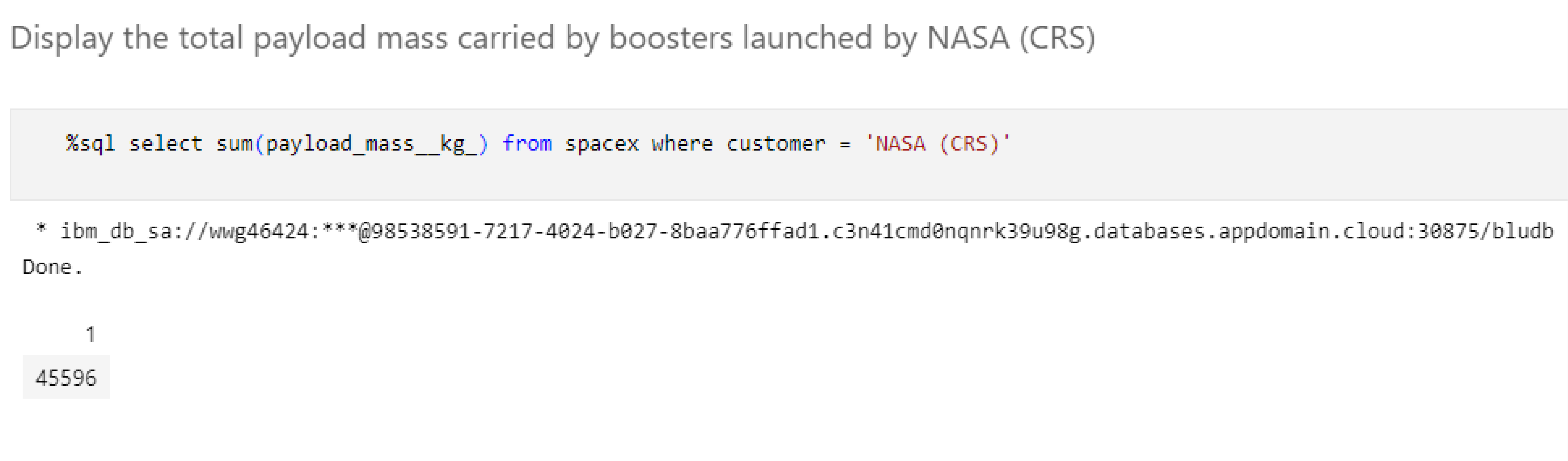
```
%sql select * from spacex where left(launch_site,3) = 'CCA' limit 5
```

```
 * ibm_db_sa://wwg46424:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
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
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- First 5 records starting with CCA

Total Payload Mass



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

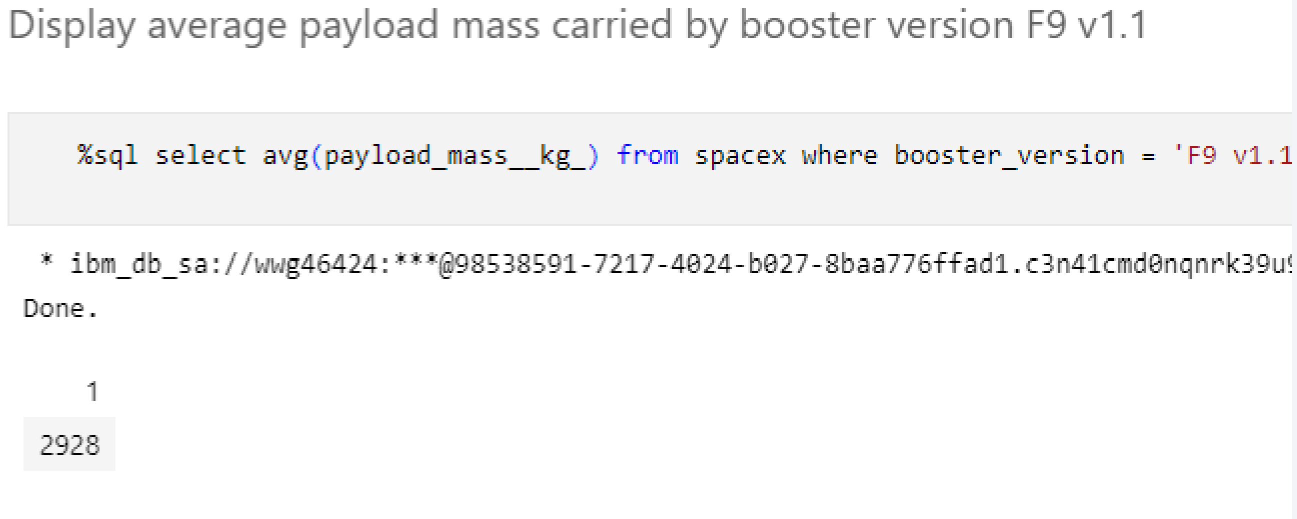
```
%sql select sum(payload_mass__kg_) from spacex where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://wwg46424:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.

    1
45596
```

Total payload mass from NASA(CRS) is 45,596kg

Average Payload Mass by F9 v1.1



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

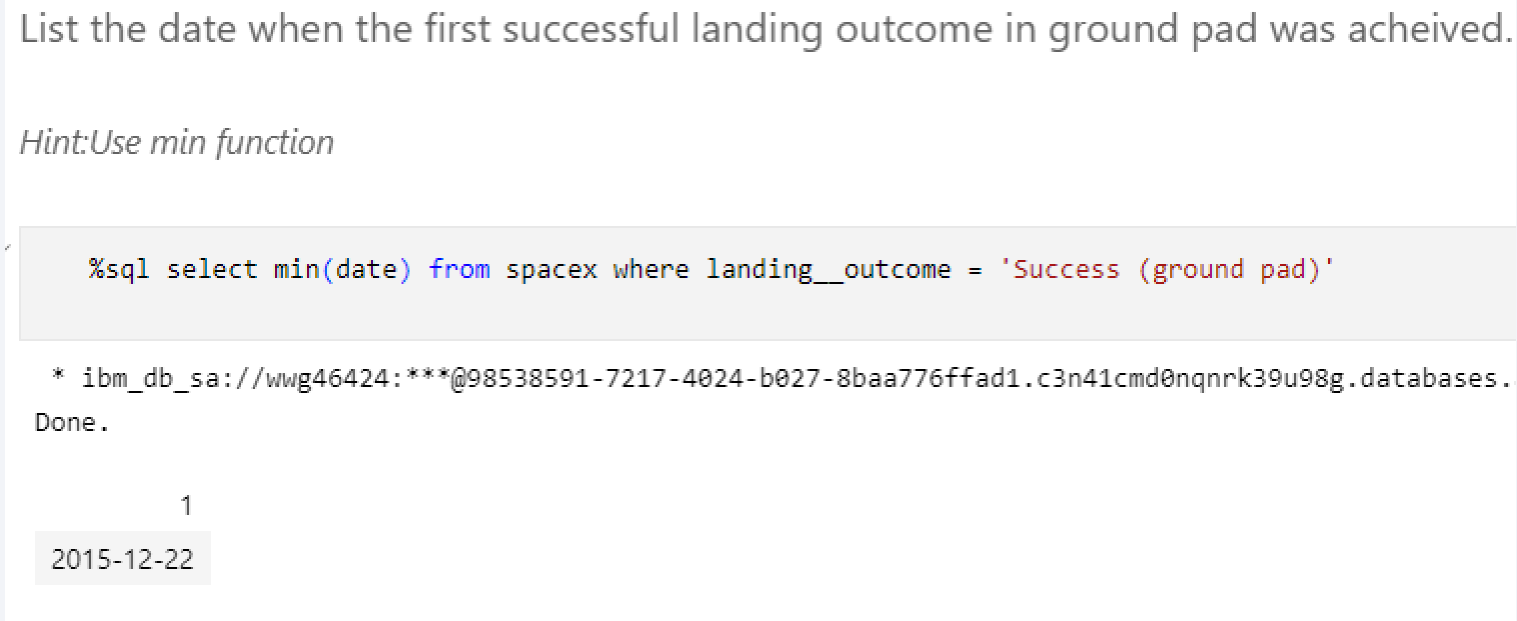
```
%sql select avg(payload_mass__kg_) from spacex where booster_version = 'F9 v1.1
```

```
 * ibm_db_sa://wwg46424:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u
Done.

   1
 2928
```

Average Payload mass from booster version ‘F9 v1.1’ is 2,928kg

First Successful Ground Landing Date



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

Hint:Use min function

```
%sql select min(date) from spacex where landing__outcome = 'Success (ground pad)'
```

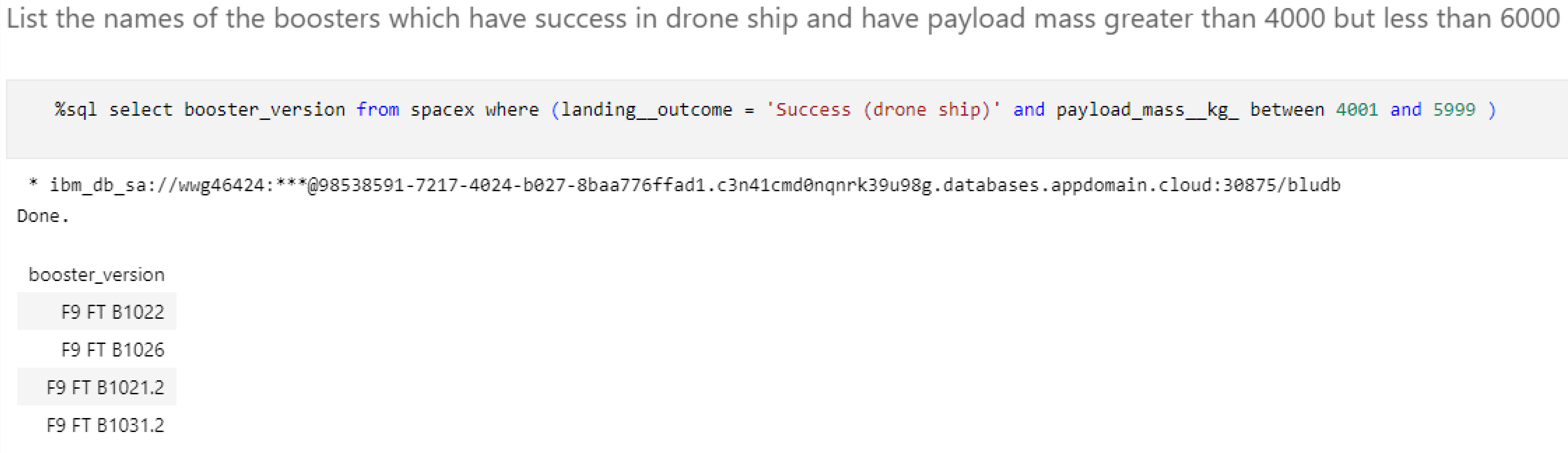
```
 * ibm_db_sa://wwg46424:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.
Done.
```

1
2015-12-22

First successful ground landing date is 22nd Dec 2015. I got a different date in Github when running it outside Watson Studio as I ran out of subscription (for last two tasks), but I have both versions of the files. Please see the Watson version below for this result. Skills lab version was used in main slide.

Github URL: https://github.com/Raghava33/DataScience-Certification-Capstone-project/blob/799293333451cabb741c6e5de768d1d4bd4c1620/Week%202/EDA%20with%20SQL.ipynb

Successful Drone Ship Landing with Payload between 4000 and 6000



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

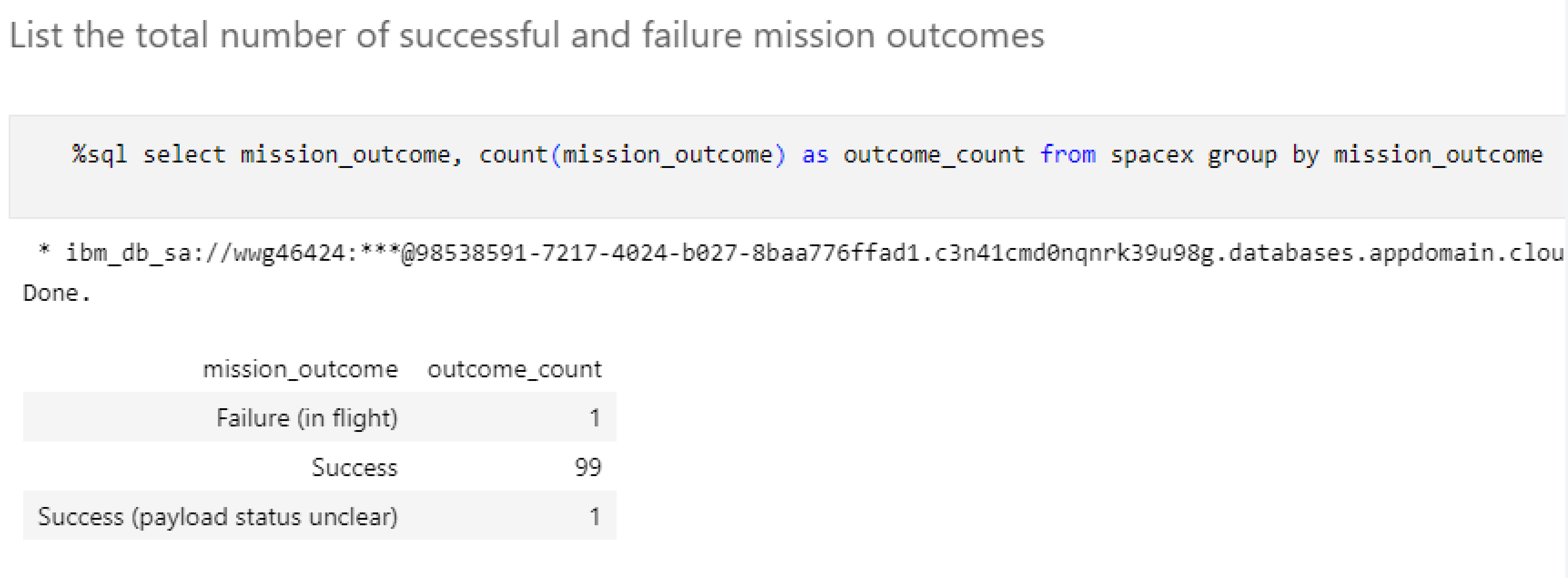
```
%sql select booster_version from spacex where (landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4001 and 5999 )
```

```
 * ibm_db_sa://wwg46424:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

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

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 all start with ‘F9 FT B10’

Total Number of Successful and Failure Mission Outcomes



List the total number of successful and failure mission outcomes

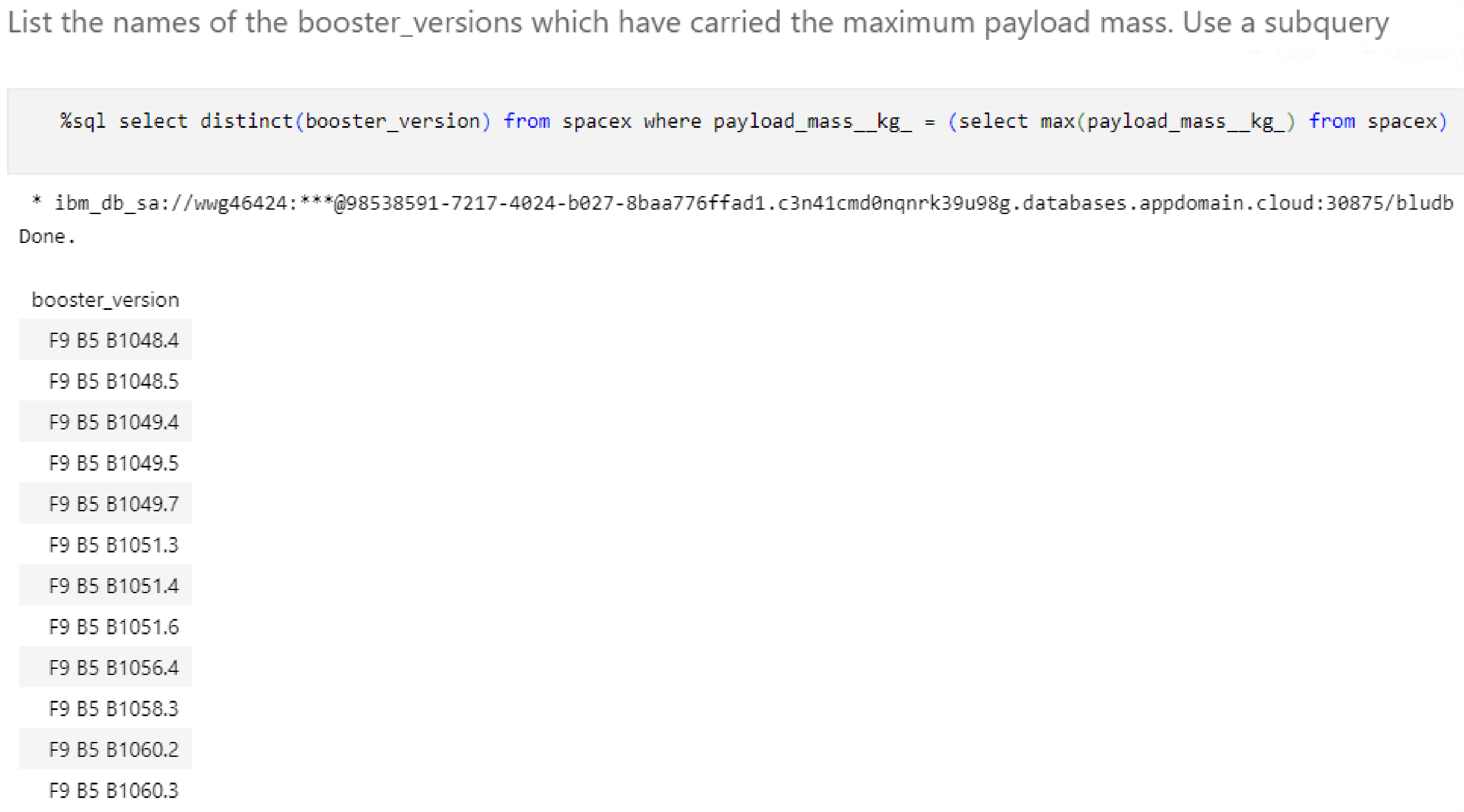
```
%sql select mission_outcome, count(mission_outcome) as outcome_count from spacex group by mission_outcome
```

```
 * ibm_db_sa://wwg46424:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.clou
Done.
```

mission_outcome	outcome_count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

SpaceX has 98% success rate

Boosters Carried Maximum Payload



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

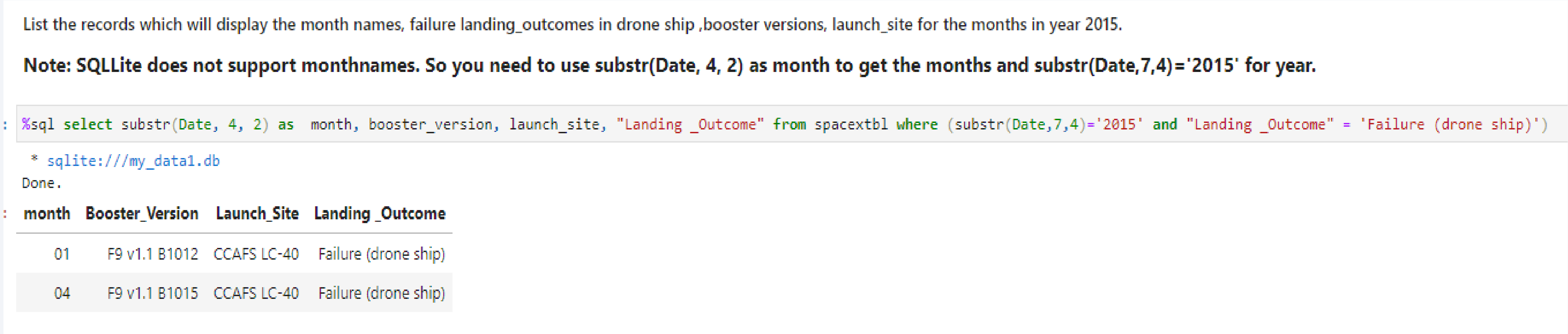
```
%sql select distinct(booster_version) from spacex where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex)
```

```
 * ibm_db_sa://wwg46424:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
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

List of boosters with Maximum payload mass

2015 Launch Records



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, booster_version, launch_site, "Landing _Outcome" from spacextbl where (substr(Date,7,4)='2015' and "Landing _Outcome" = 'Failure (drone ship)')
```

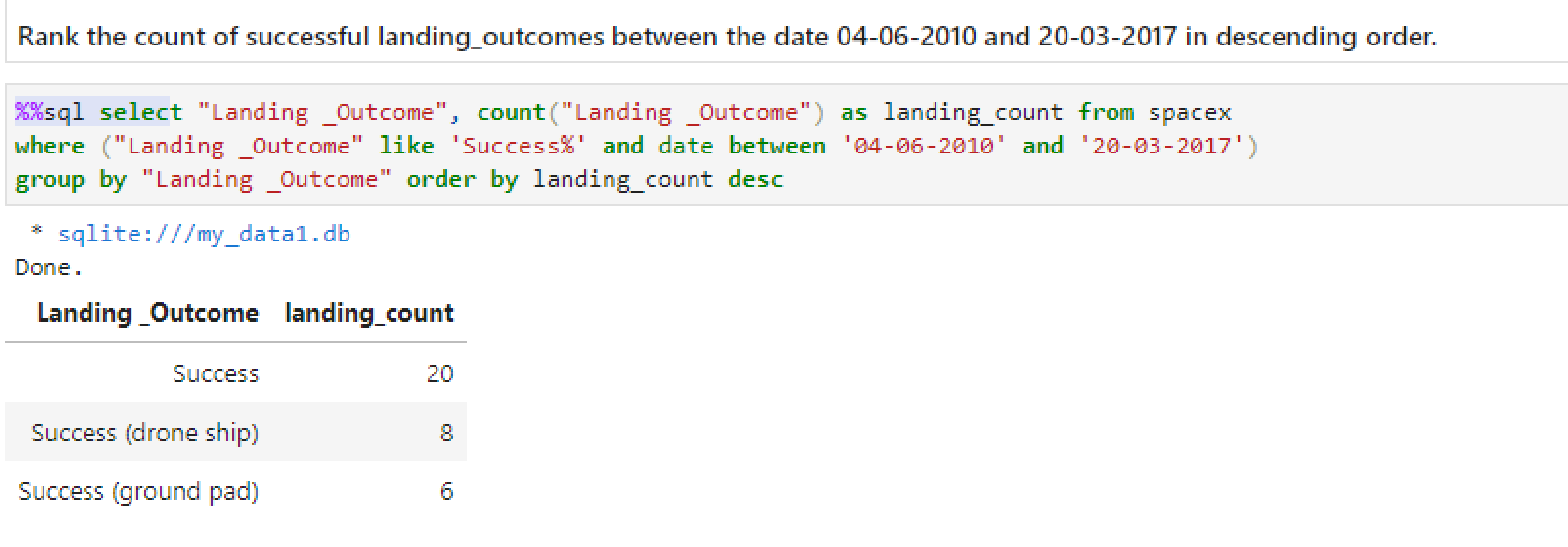
```
 * sqlite:///my_data1.db
Done.
```

month	Booster_Version	Launch_Site	Landing _Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

We have 2 failed landing outcomes in January and April

As shown in the screenshot, Skillslab doesn’t support monthname, it is instead extracted as a number using the formula provided by IBM.

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



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("Landing _Outcome") as landing_count from spacex
where ("Landing _Outcome" like 'Success%' and date between '04-06-2010' and '20-03-2017')
group by "Landing _Outcome" order by landing_count desc
```

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

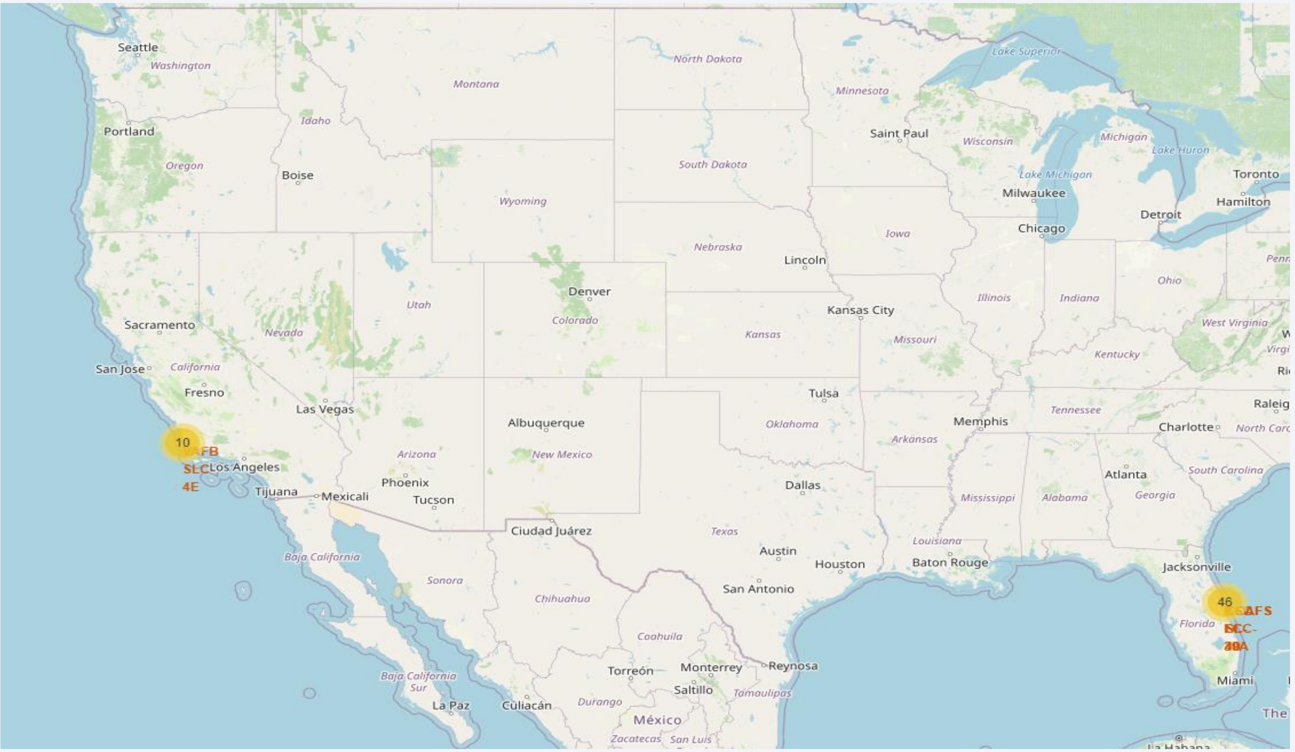
Landing _Outcome	landing_count
Success	20
Success (drone ship)	8
Success (ground pad)	6

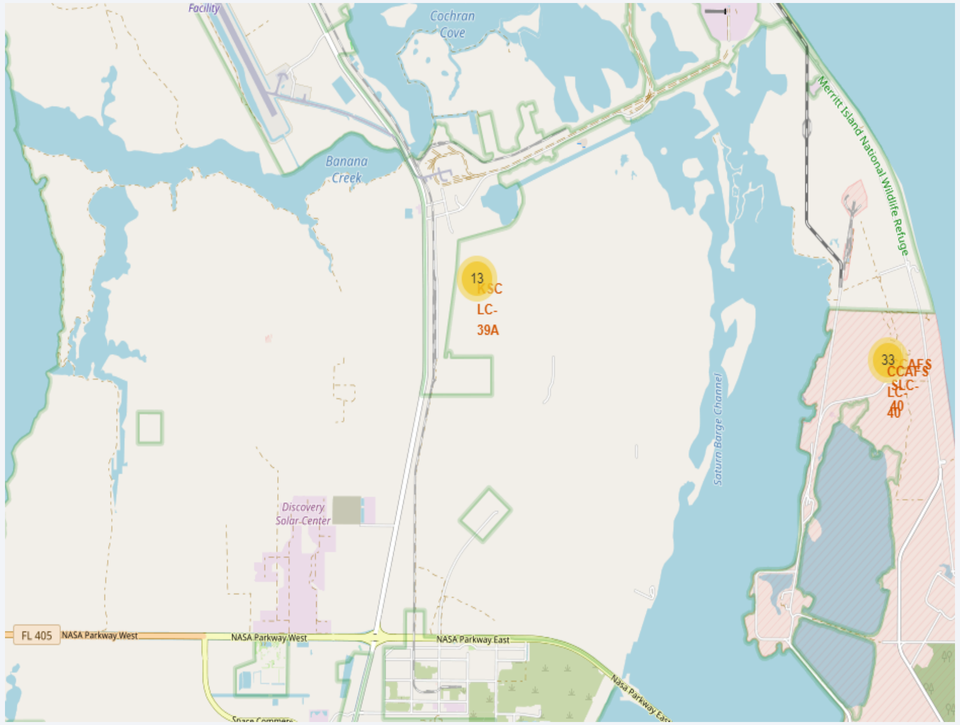
There were a total of 34 successful landings, out of which 8 are on drone ship and 6 are on ground pad

Section 3

Launch Sites Proximities Analysis

Launch site markers

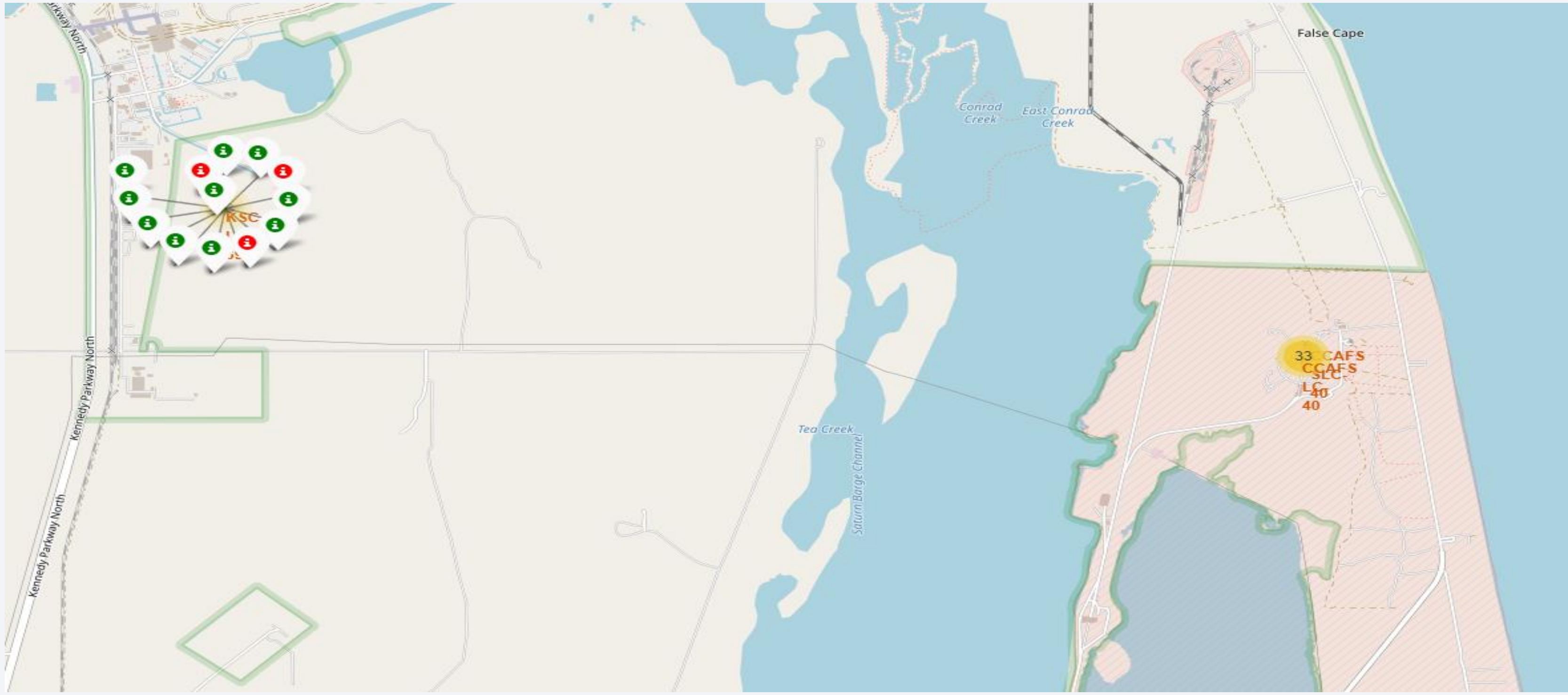




On the left is the USA map, on right is only the east coast.

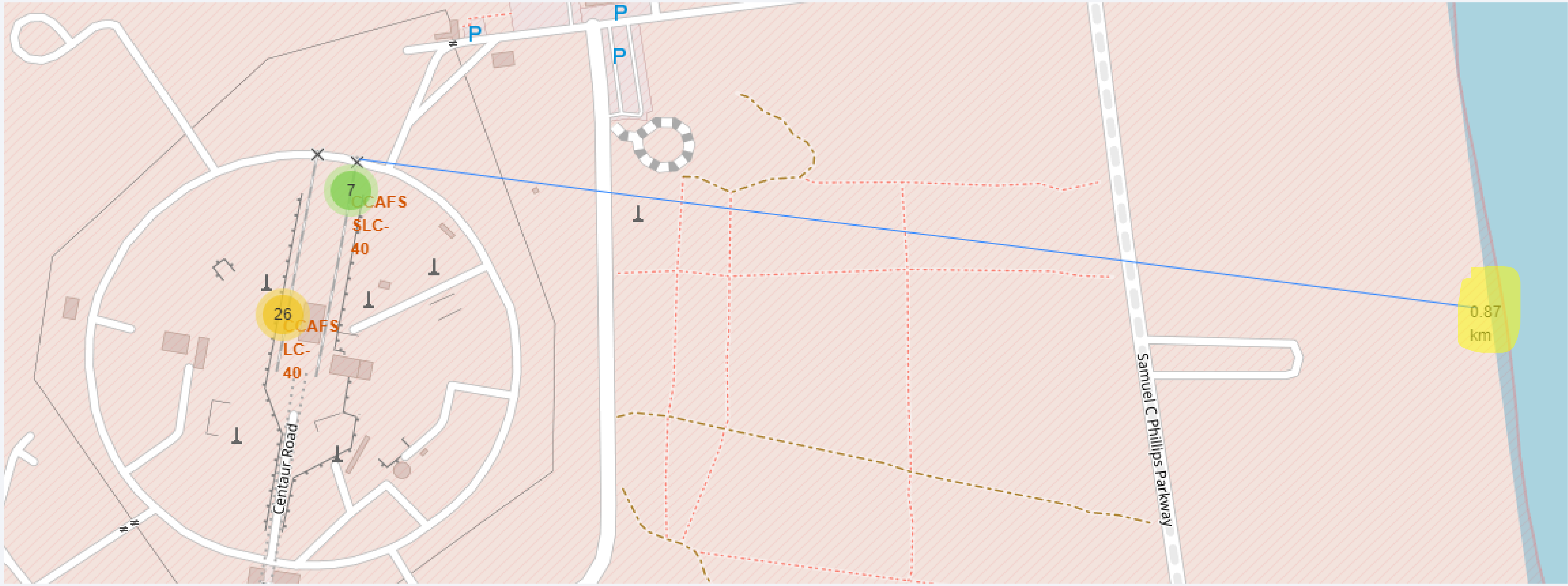
All launch sites are close to the ocean, with most of them located on east coast

Map of markers with success rate



Above map shows the east coast KSC LC 39A launch site success markers. Green is success and red is failure.

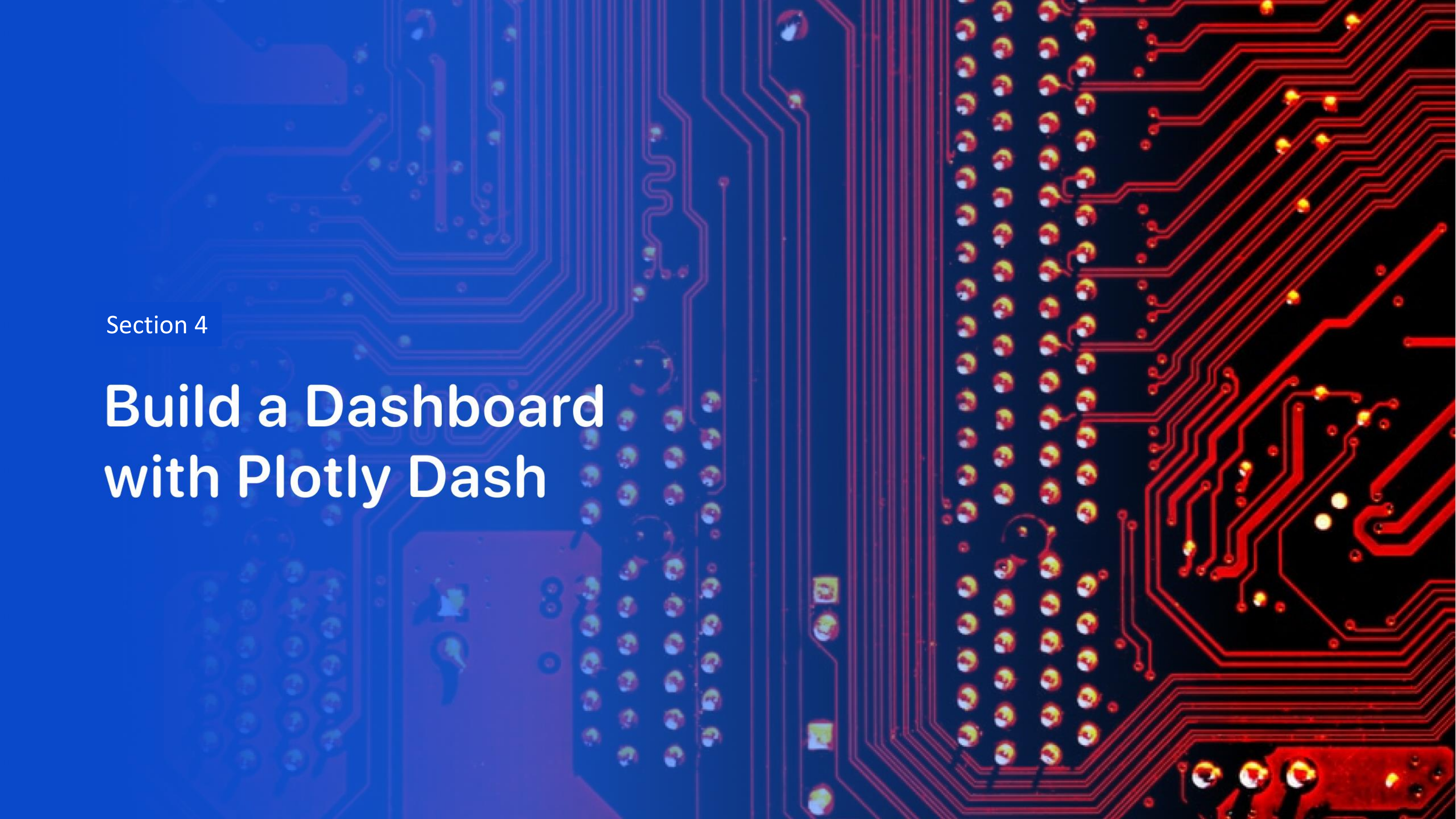
Distance to east coastline



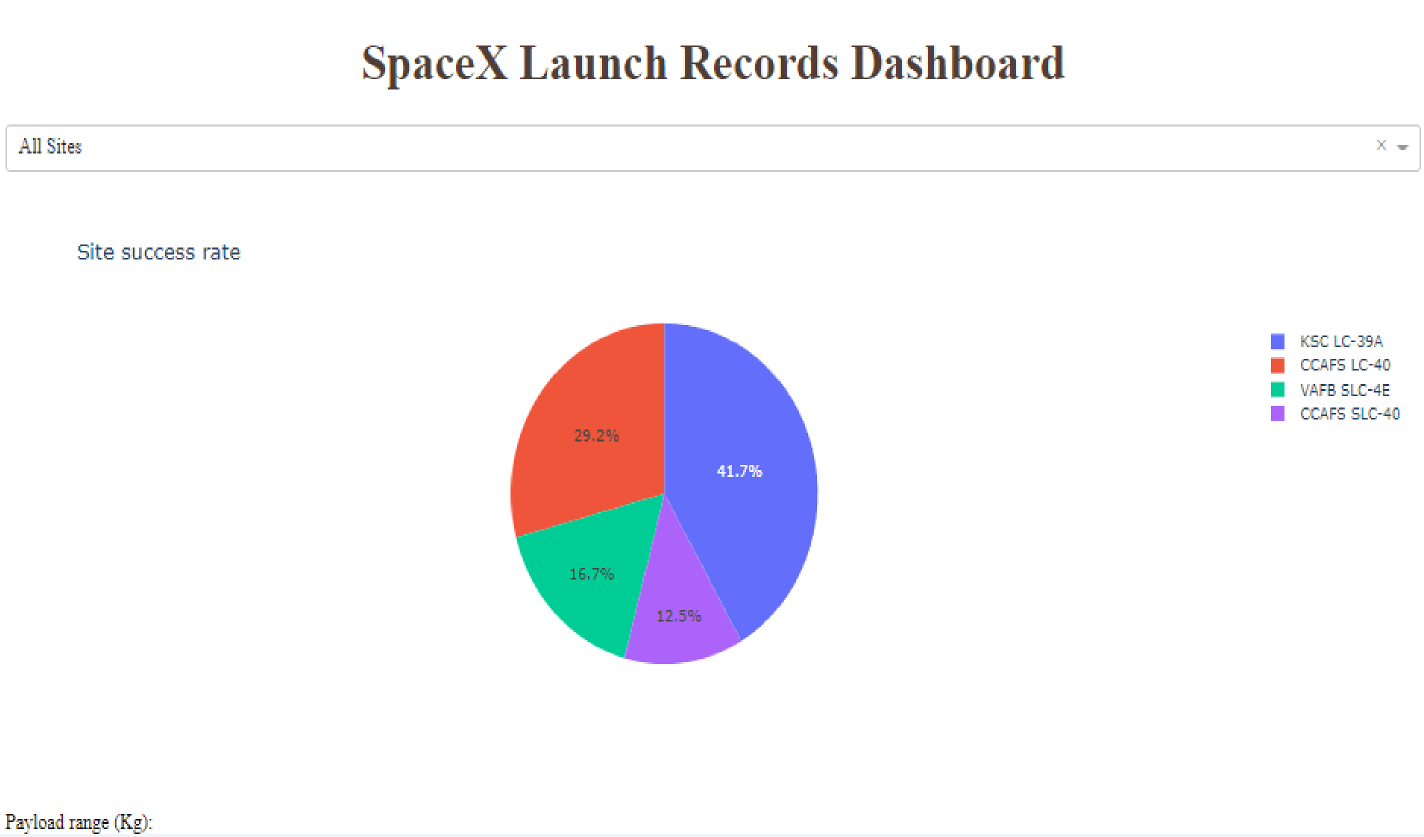
Distance from CCAFS SLC 40 is calculated as 0.87km (highlighted in yellow) and shown on the line (blue). Launch sites are located close to the coast to facilitate water landings and to keep population away from the explosions. They are also usually close to the transportation like railways.

Section 4

Build a Dashboard with Plotly Dash



Success rate across all launch sites



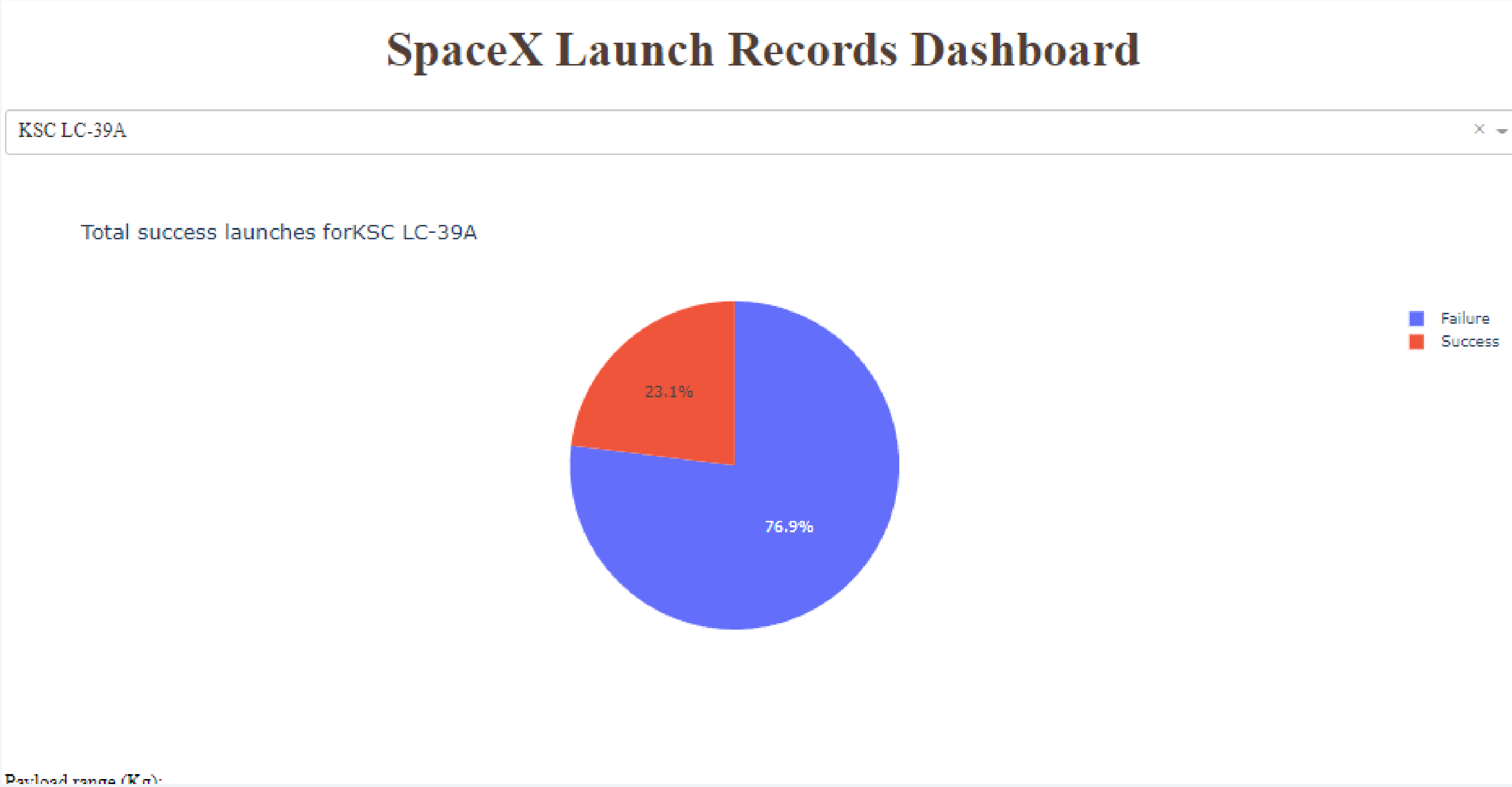
SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

KSC LC 39A has the highest success rate, followed by CCAFS LC-40 and VAFB SLC 4E

Success pie chart for KSC LC-39A



SpaceX Launch Records Dashboard

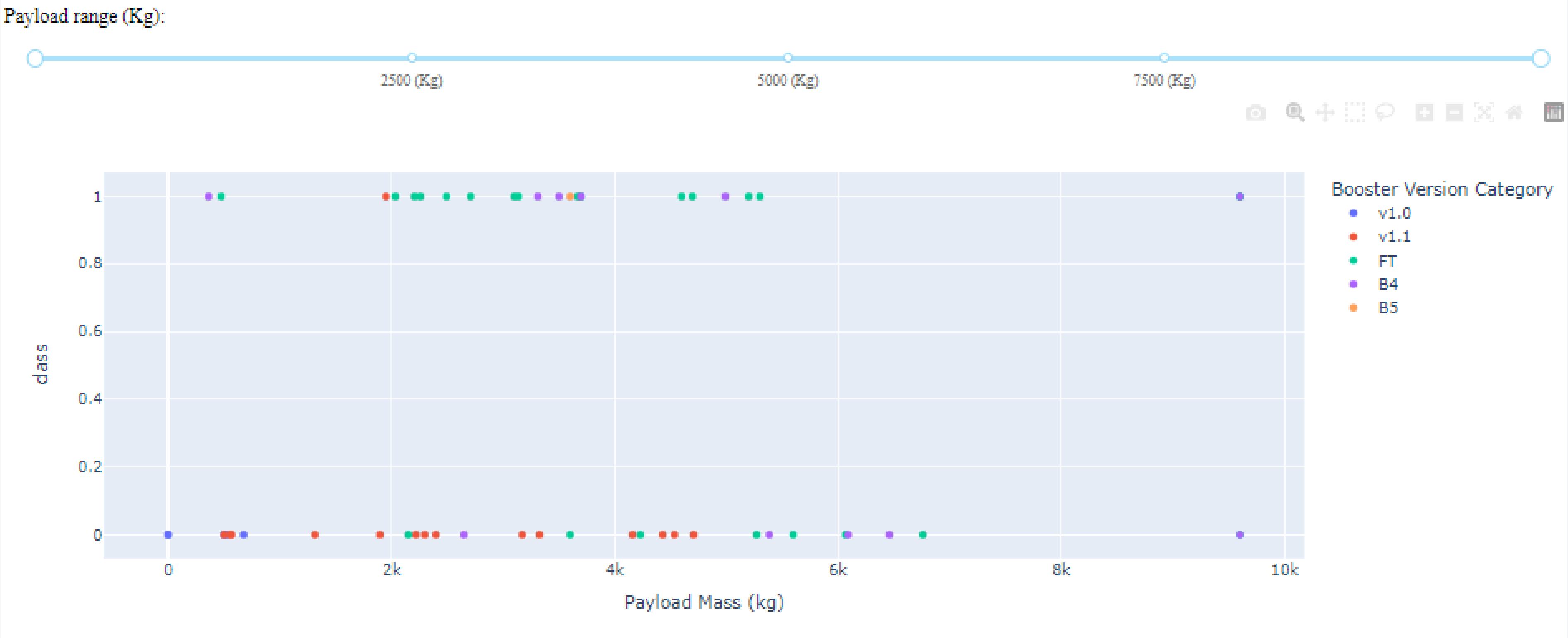
KSC LC-39A

Total success launches forKSC LC-39A

Payload range (Kg):

Although this site has the highest success rate, it is only 23.1% which does not seem impressive. This shows how tough the space launches are.

Success rate from Payload mass vs Launch Outcome

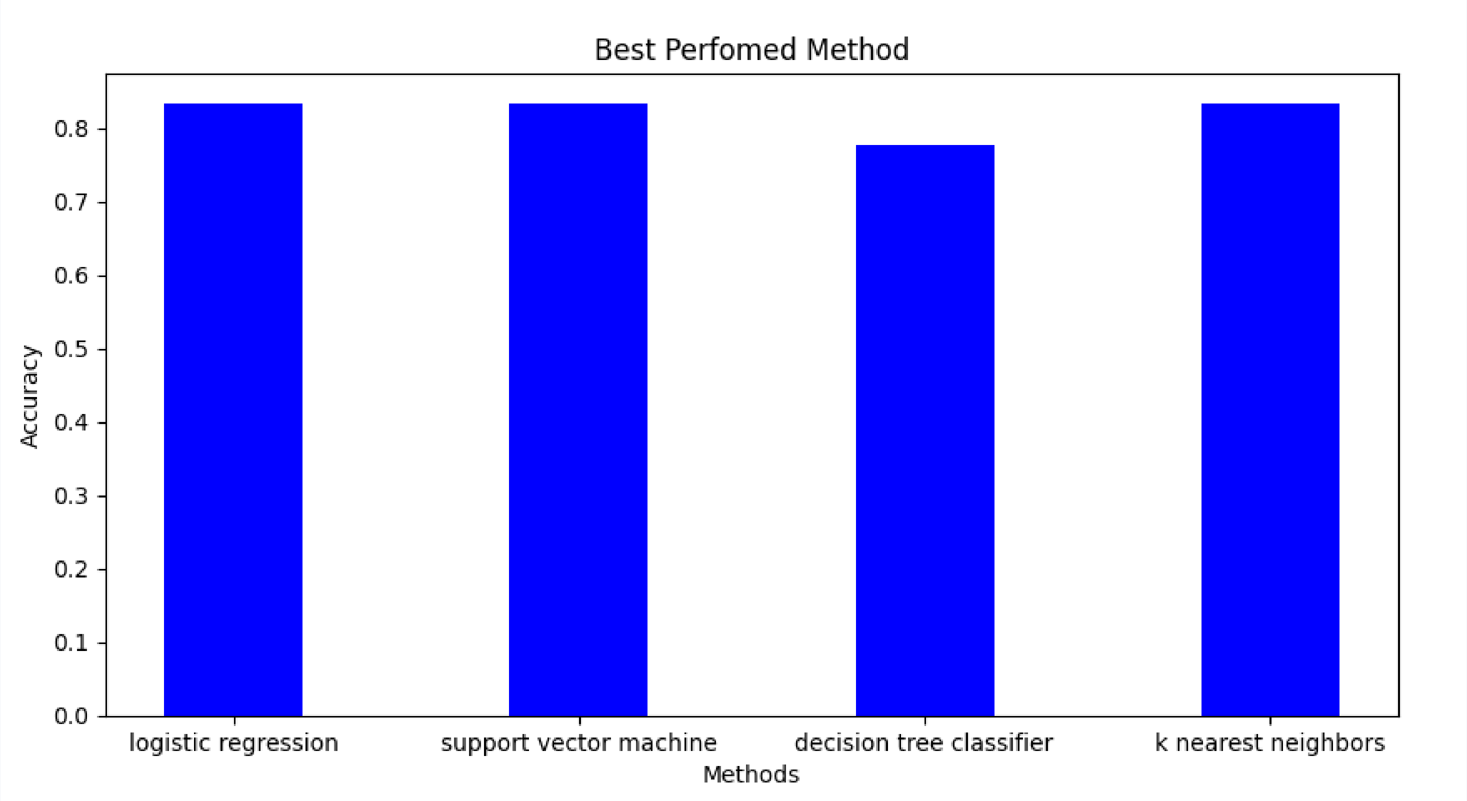


Plotly has a Payload range selector which is set from 0-10000. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, there were two failed landings with payloads of zero kg.

Section 5

Predictive Analysis (Classification)

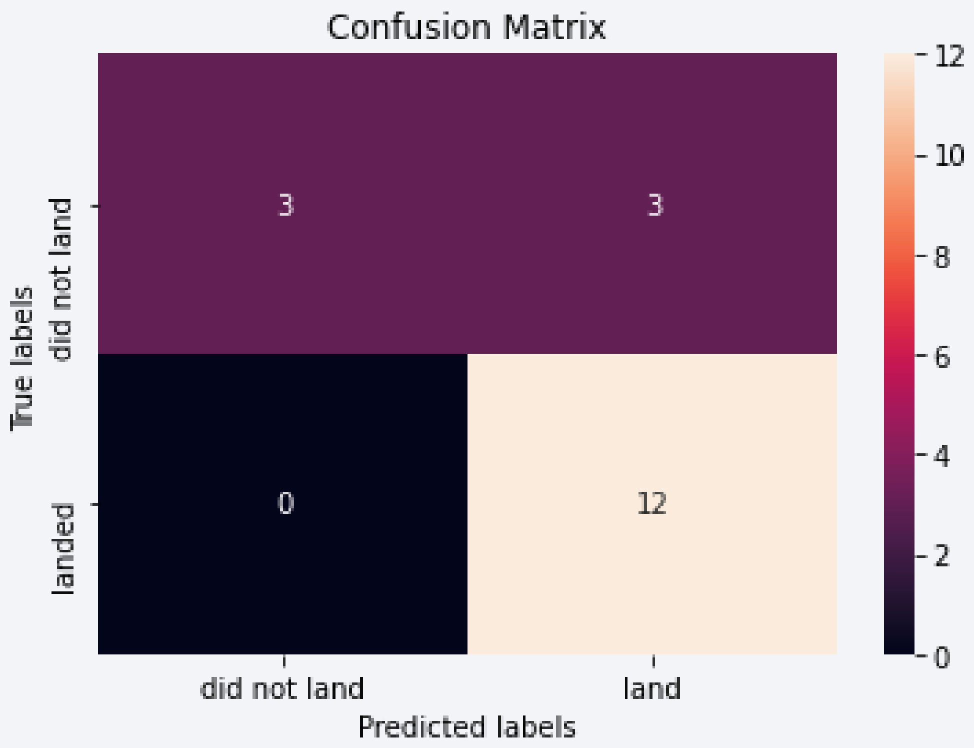
Classification Accuracy



All models have an accuracy of 83.3% except the Decision tree classifier which has an accuracy of 77.78%.

Only way to break the tie between the models is to use try and remove some of the independent variables, possibly using PCA (Principal Component Analysis)

Confusion Matrix



Confusion matrix is mostly same across all models.

The models predicted 12 successful landings when the true label was a successful landing.

The models predicted 3 unsuccessful landings when the true label was an unsuccessful landing.

The models predicted 3 successful landings when the true label was unsuccessful landings.

Models seem to over predict successful landings.

Conclusions

- EDA helped establish relationships between various variables which seem to be correlated. So, it is a good idea to conduct PCA (Principal Component Analysis) to improve explanatory power of remaining variables.
- Overall, the success rate is very high even when using k-fold validation for testing dataset.
- Stage 1 landing is consistent and hence we can recommend Space Y that Space X is doing well.
- Since most of the model predicted same accuracy, it is suggested to use PCA and extract more information from each model.
- …

Appendix

Repo link:

https://github.com/Raghava33/DataScience-Certification-Capstone-project

Thank you!