IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mateusz Molenda
10.06.2022

Outline

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

Executive Summary

- Methodologies used:
  - Data collection using SpaceX API and data wrangling
  - Exploratory data analysis using SQL
  - Interactive data visualization using folium and plotly
  - Machine learning prediction
- Summary of all results:
  - Visualizations showing different insights to the data
  - Machine learning model

Introduction

- Project background:

Using the available data from SpaceX, we want to build a model for the company SpaceY, that will allow them to minimize the cost of launching the rockets. Moreover we want to find correlates the most with the possibily of reusing the same rocket.

- Problems you want to find answers
  - What correlates the most with successful landing of the rocket?
  - What launching site is the best option?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from the Wikipedia page about Falcon 9 launches and SpaceX API
- Perform data wrangling
  - The data was processed using one-hot encoding
- 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

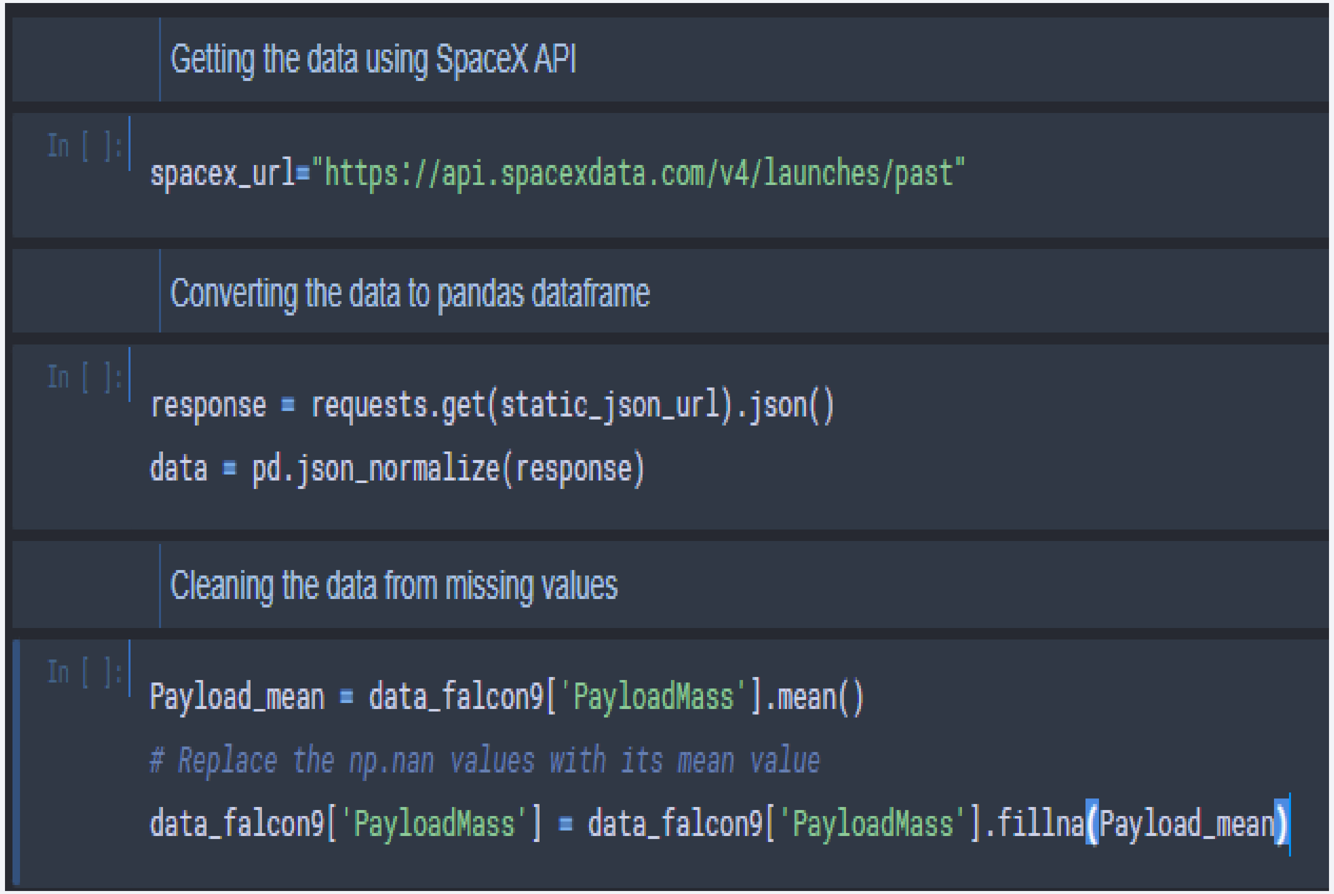
Data Collection

- The data set was downloaded from 2 sites: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922, https://api.spacexdata.com/v4/launches/past
- Then it was saved in the pandas dataframe and cleared from all the missing data
- A BeautifulSoup was used for extracting the necessary data from the Wikipedia page

Data Collection – SpaceX API

- The SpaceX API was used to collect, clean and format the data.

Source code:
https://github.com/molendziak/IBM_DataScienceCapstone/blob/main/Data_Collection_API.ipynb



Getting the data using SpaceX API

```
In [ ]: spacex_url="https://api.spacexdata.com/v4/launches/past"
```

Converting the data to pandas dataframe

```
In [ ]: response = requests.get(static_json_url).json()
        data = pd.json_normalize(response)
```

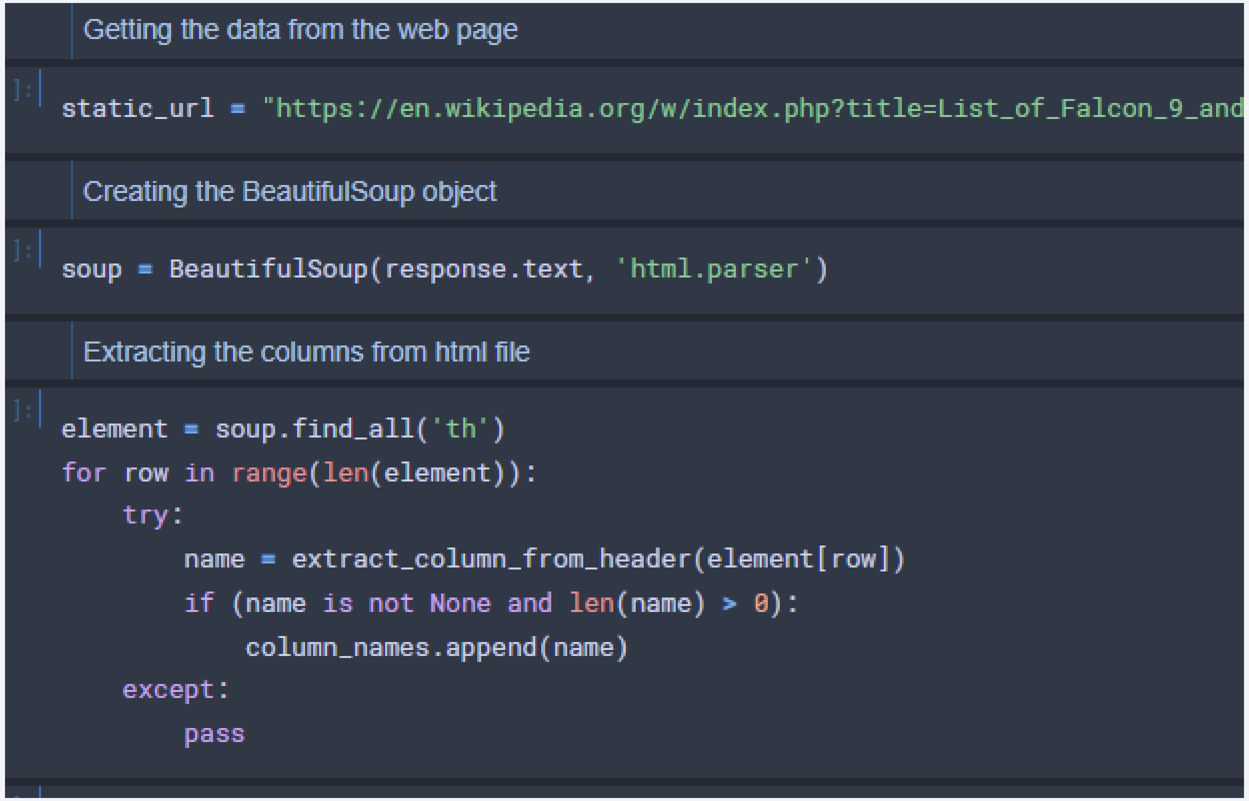
Cleaning the data from missing values

```
In [ ]: Payload_mean = data_falcon9['PayloadMass'].mean()
        # Replace the np.nan values with its mean value
        data_falcon9['PayloadMass'] = data_falcon9['PayloadMass'].fillna(Payload_mean)
```

Data Collection - Scraping

- The webscraping was performer from the Wikipedia and with the help of BeautifulSoup.

Source code: https://github.com/molendziak/IBM_DataScienceCapstone/blob/main/DataCollectionAndWebScraping.ipynb



Getting the data from the web page

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and
```

Creating the BeautifulSoup object

```
soup = BeautifulSoup(response.text, 'html.parser')
```

Extracting the columns from html file

```
element = soup.find_all('th')
for row in range(len(element)):
    try:
        name = extract_column_from_header(element[row])
        if (name is not None and len(name) > 0):
            column_names.append(name)
    except:
        pass
```

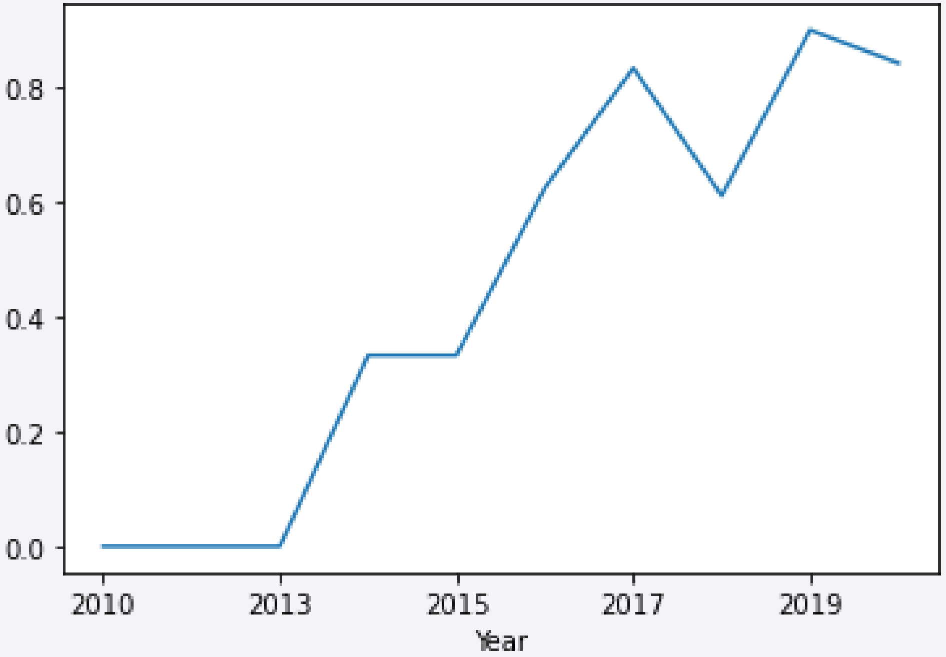
Data Wrangling

- We identified the amount of null values
- We performer initial exploratory data analysis determining the amount of launching sites, orbits, and landing outcomes
- We encoded the outcome into binary-like behaviour

Source code: https://github.com/molendziak/IBM_DataScienceCapstone/blob/main/DataWrangling.ipynb

EDA with Data Visualization

- Here, we were looking for pairs of variables that, based on the graph, could depend on eachother
  - The pairs are: Flight Number x Payload Mass, Payload Mass x Launch Site, Flight Number x Orbit, Payload Mass x Orbit, Yearly success rate



Source code:
https://github.com/molendziak/IBM_DataScienceCapstone/blob/main/EDA_DataViz.ipynb

EDA with SQL

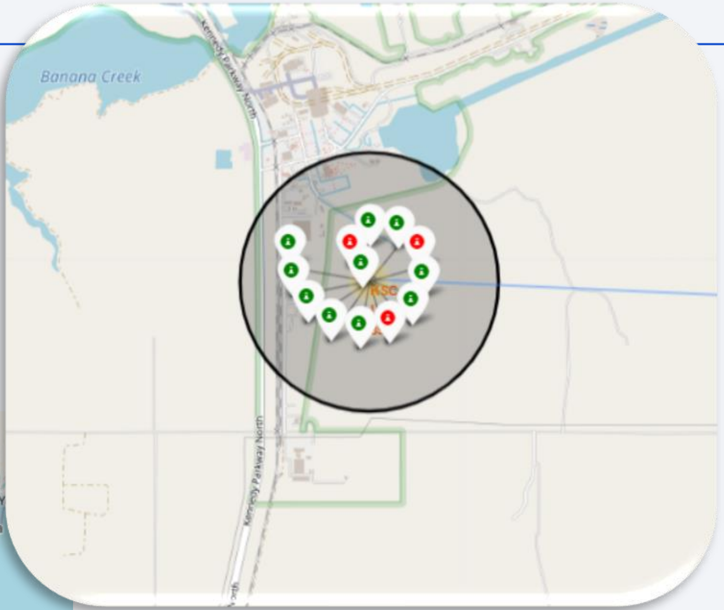
The SQl queries:

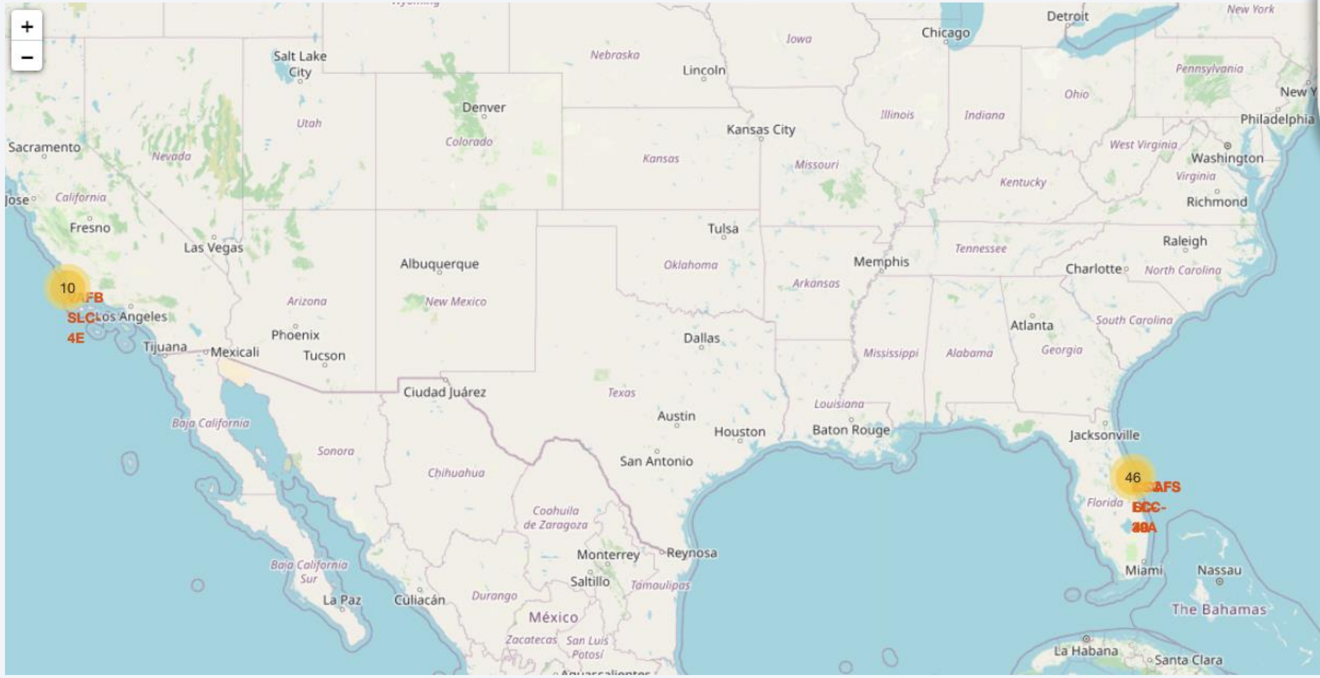
- Launch Site names
- Launch Sites with „CCA” in name
- Total payload mass carried by boosters
- Average payload mass carried by booster version F9 v1.1
- Date of the first successful landing
- Boosters with successful drone ship landing with payload between 4000 and 6000 kg
- Total numbaer of successful and failed missions
- Boosters which carried the maximum payload
- List of failed landing outcomes in year 2015
- Count of landing outcomes between 04/06/2015 and 20/03/2017

Source code:
https://github.com/molendziak/IBM_DataScienceCapstone/blob/main/EDA_SQL.ipynb

Build an Interactive Map with Folium

In Folium we wanted to mark on the map all sites locations, with the markes showing if the landing was successful or not, and to determine distance from various objects like: cities, roads and coastlines.





The purpose was to see which sites are the most succesful and how other objects might contrubite to that.

Source code: https://github.com/molendziak/IBM_DataScienceCapstone/blob/main/Analysis_with_folium.ipynb

Build a Dashboard with Plotly Dash

- We wanted to create and interactive Plotly app with graphs
- The plots included pie charts with total numer of launches by site
- The interactive plots included scatter graphs of Outcome vs. Payload Mass for different booster versions.

Source code:
https://github.com/molendziak/IBM_DataScienceCapstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- We wanted to check which of the four models (logistic regression, SVM, decision tree, k-nearest neighbour) would perform the best.
- To do that we used the train_test_split method and the GridSearchCV to determine best parameter value
- Then we caluculated accuracy, jaccard index and F2 score for each of the methods.

Source code:
https://github.com/molendziak/IBM_DataScienceCapstone/blob/main/ML_Prediction.ipynb

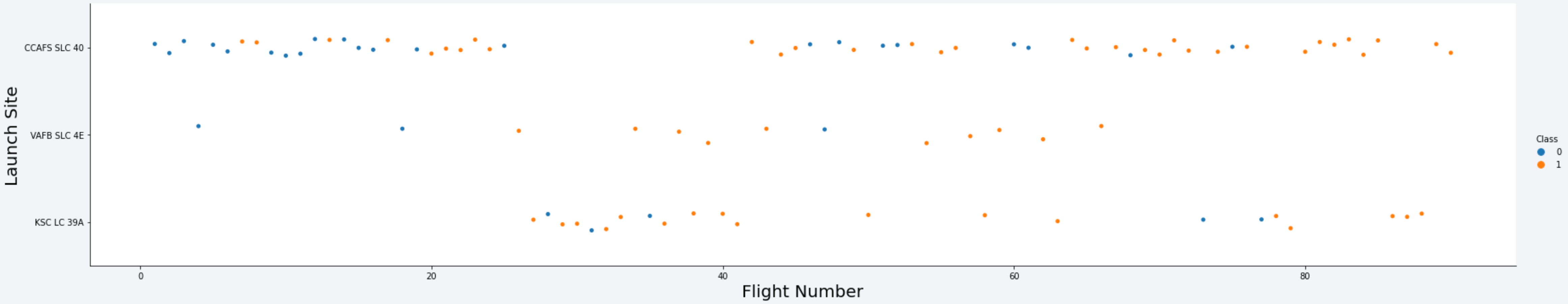
Results

- Exploratory data analysis results
  - The succes rate increases rapidly from 2013
  - There are 4 most succesful orbits
  - SpaceX uses 4 launching sites
  - Fisrt successful landing happened in 2015
  - Almost all missions were succesful
- Interactive analytics demo in screenshots
- Predictive analysis results
  - Predictive analysis showed that the Decision tree is the best classifiacation model with the accuracy around 92%.

Section 2

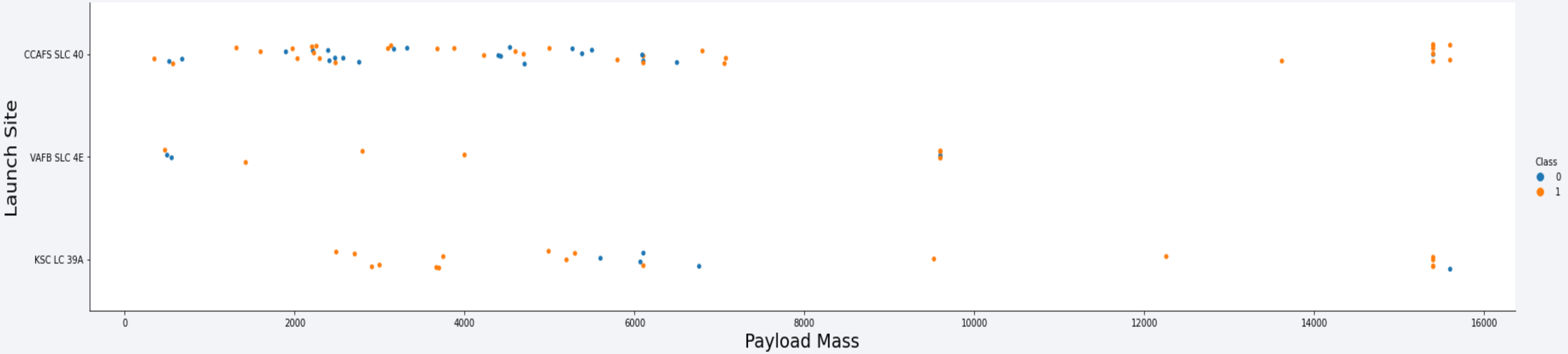
Insights drawn from EDA

Flight Number vs. Launch Site



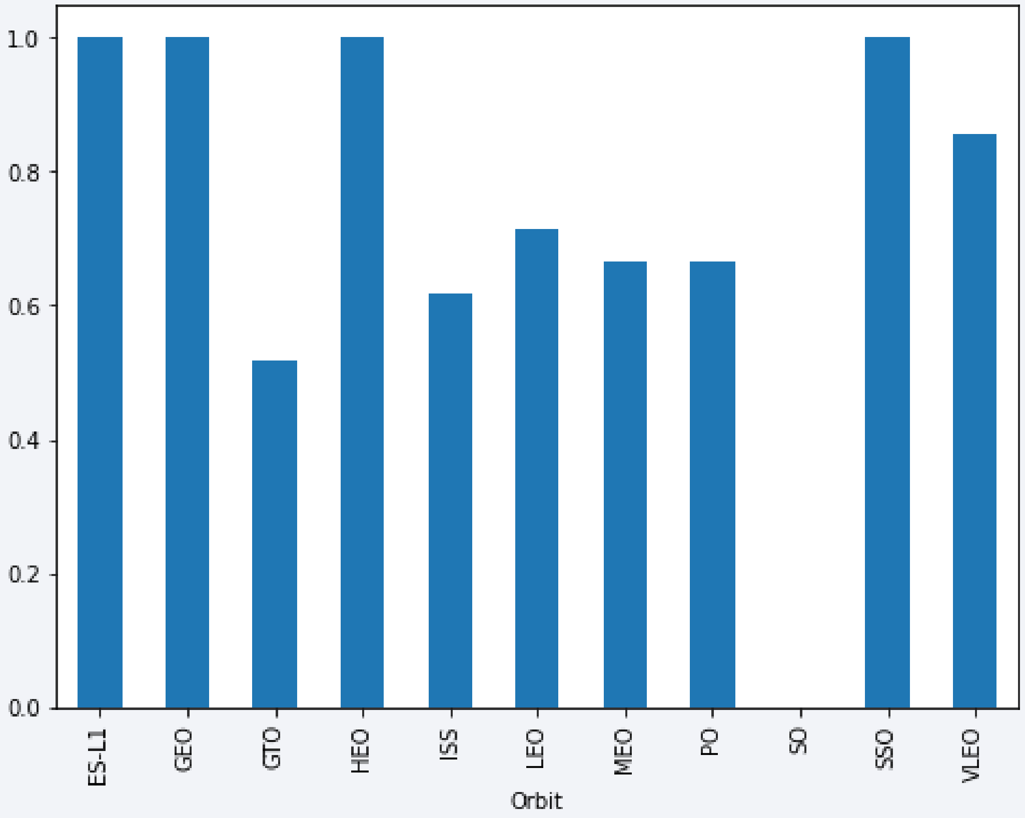
- We see that the success rate increases with flight numer
- The graph shows that the best launch site is CCAFS SLC 40

Payload vs. Launch Site



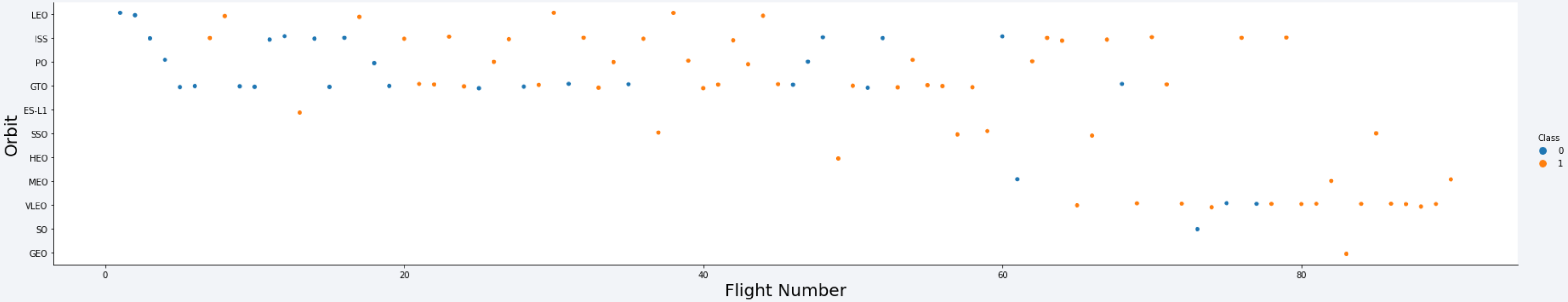
- We see that there is not many flight for payloads between 8000 and 14000 kg
- The CCAFS SLC 40 is most successful when doing heavy payloads
- The VAFB SLC 4E site did not do any heavy payloads

Success Rate vs. Orbit Type



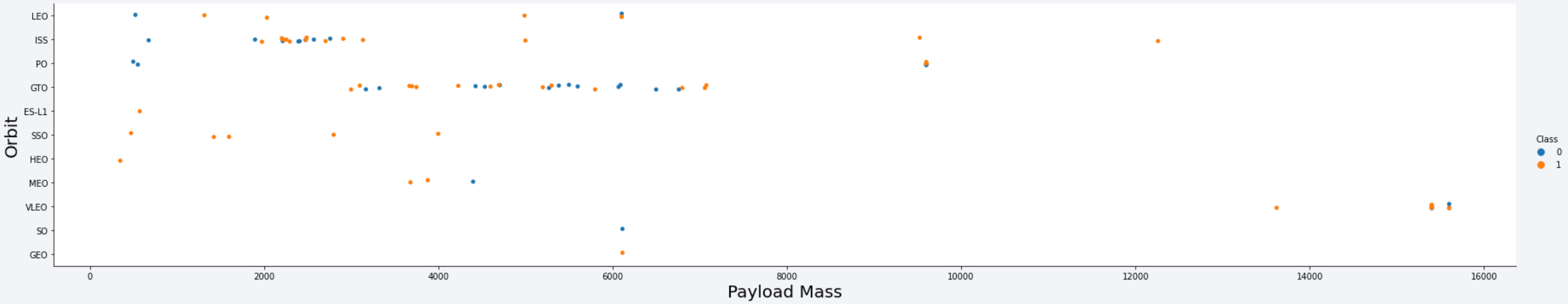
- There are 4 most succesful orbits: ES-L1, GEO, HEO, SSO.

Flight Number vs. Orbit Type



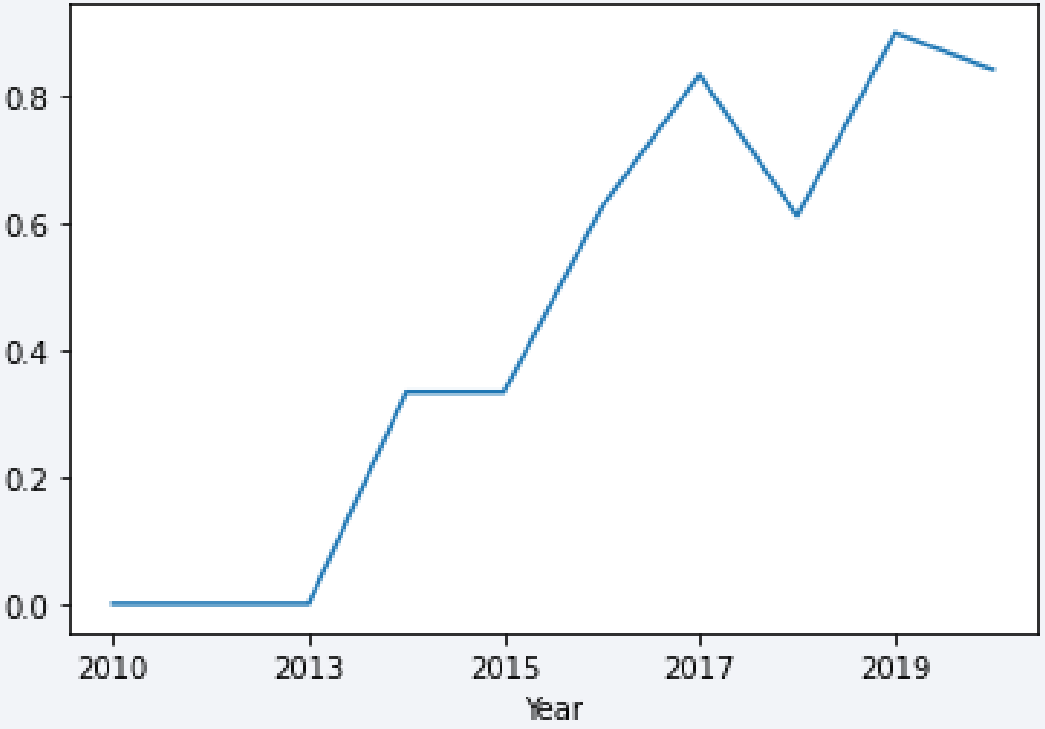
- The success rate increases over time
- Only GTO’s orbit success seems to not increase

Payload vs. Orbit Type



- No clear correlation visible
- SSO orbit has only successful missions

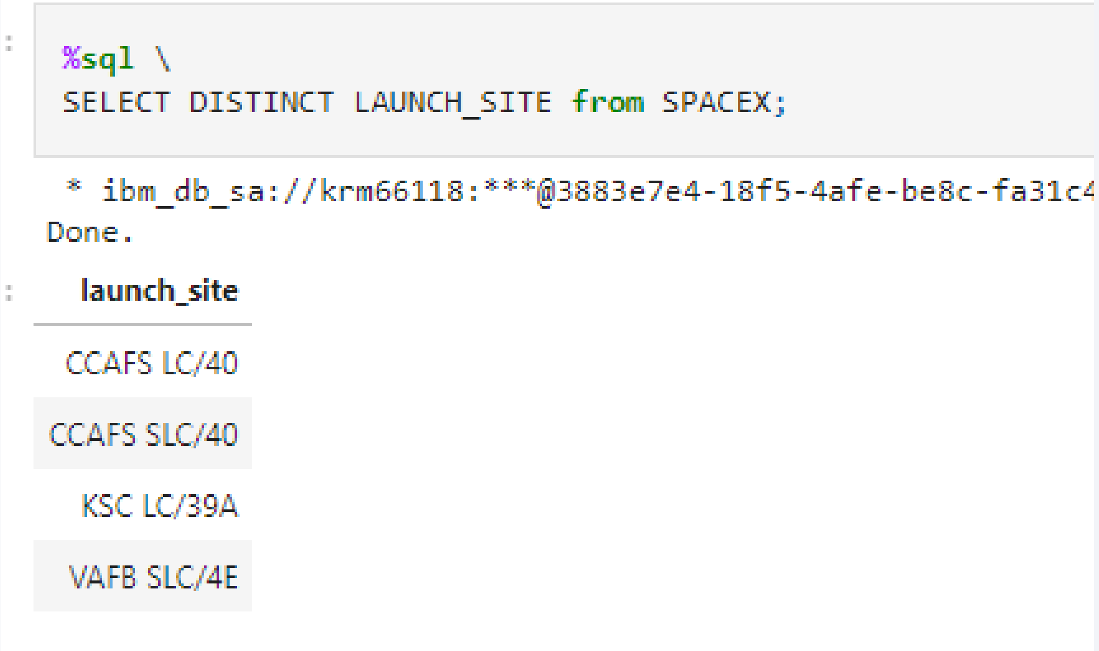
Launch Success Yearly Trend



- We see that from the year 2013 the success rate starts to rapidly increase
- In 2018 there is small decrease

All Launch Site Names

- According to the data there are 4 distinct launch sites



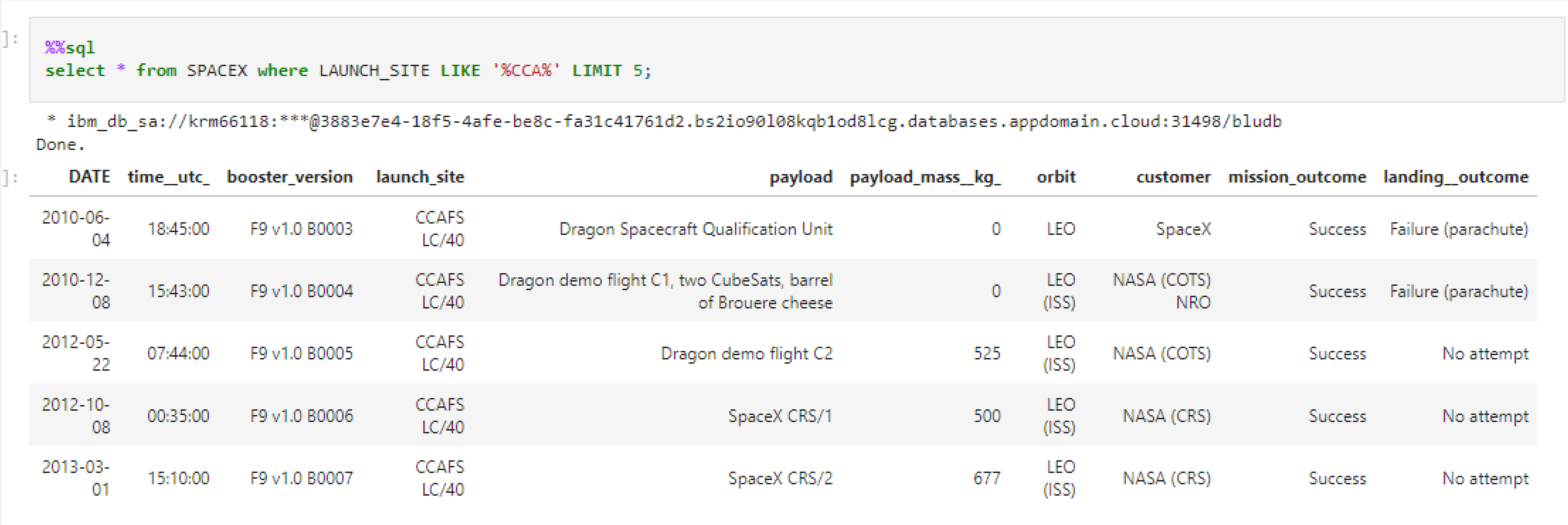
```
%sql \
SELECT DISTINCT LAUNCH_SITE from SPACEX;
```

```
 * ibm_db_sa://krm66118:***@3883e7e4-18f5-4afe-be8c-fa31c4
Done.
```

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

Launch Site Names Begin with 'CCA'

- Launch sites with „CCA”



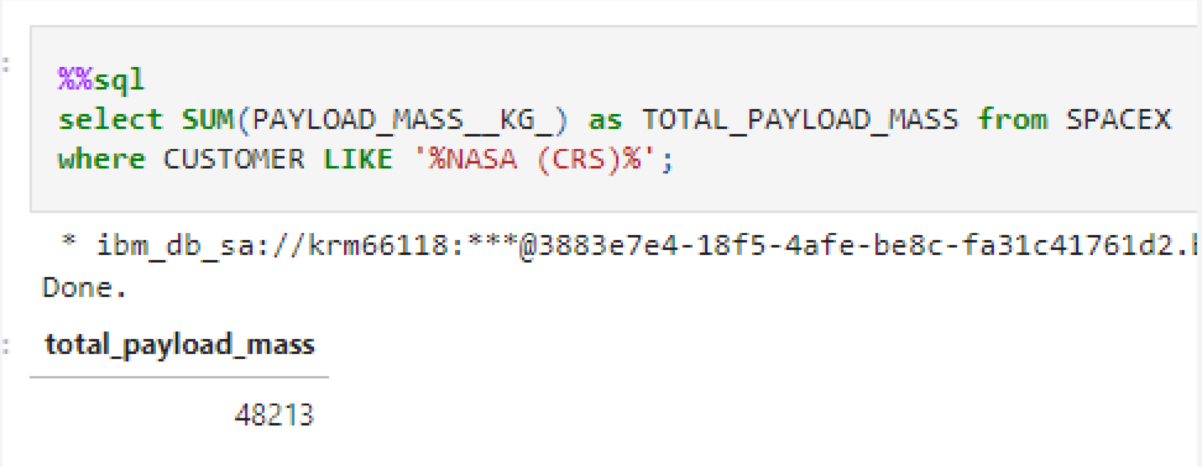
```
%%sql
select * from SPACEX where LAUNCH_SITE LIKE '%CCA%' LIMIT 5;
```

```
 * ibm_db_sa://krm66118:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/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

Total Payload Mass

- Total payload mass of boosters launched by NASA



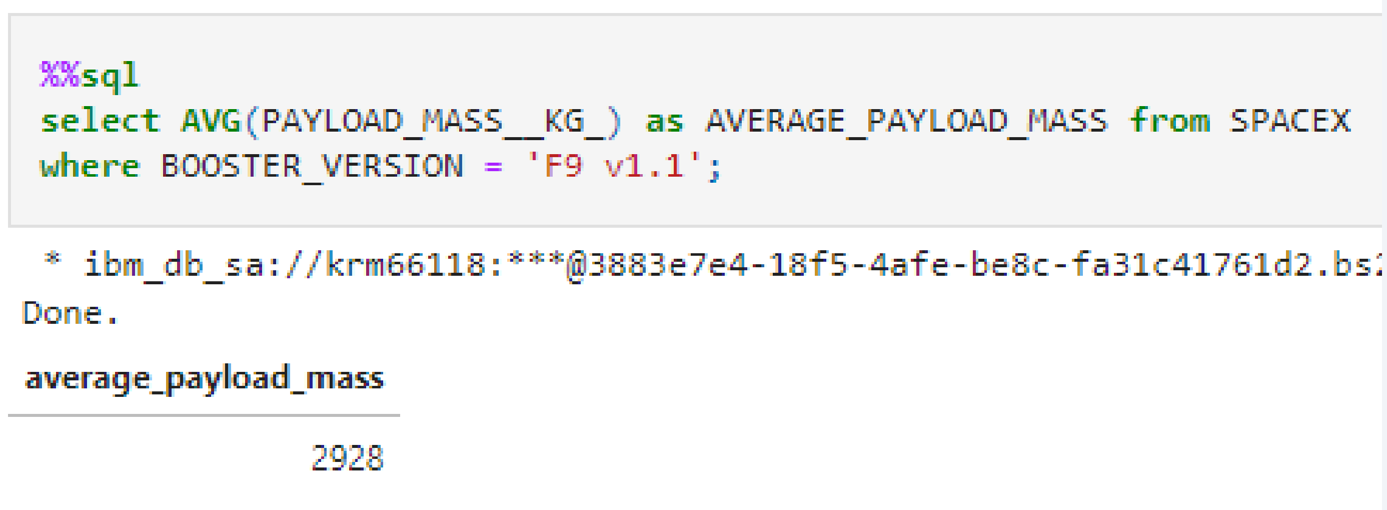
```
%%sql
select SUM(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD_MASS from SPACEX
where CUSTOMER LIKE '%NASA (CRS)%';
```

```
 * ibm_db_sa://krm66118:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.
Done.
```

total_payload_mass
48213

Average Payload Mass by F9 v1.1

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



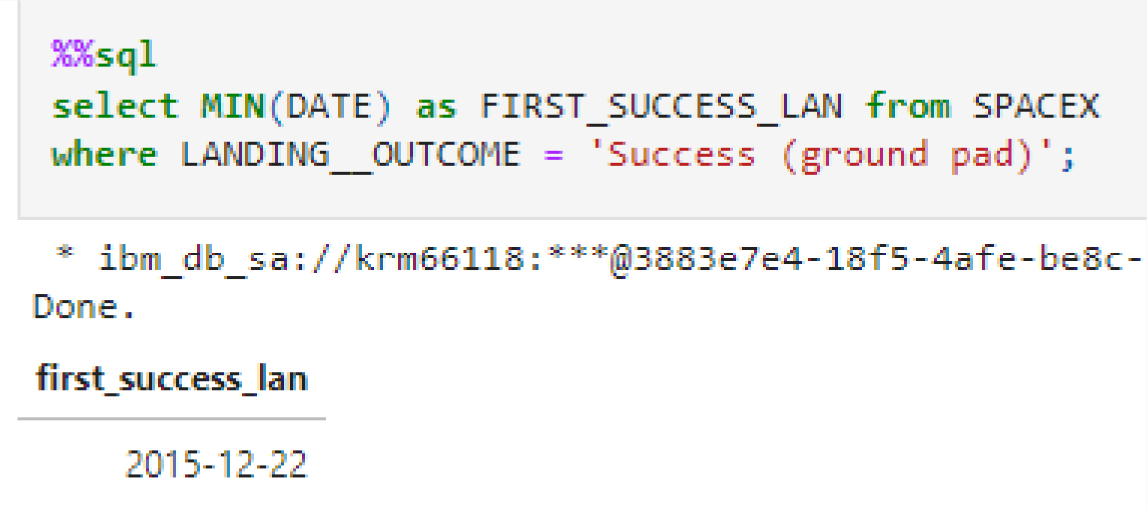
```
%%sql
select AVG(PAYLOAD_MASS__KG_) as AVERAGE_PAYLOAD_MASS from SPACEX
where BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://krm66118:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs:
Done.
```

average_payload_mass
2928

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad



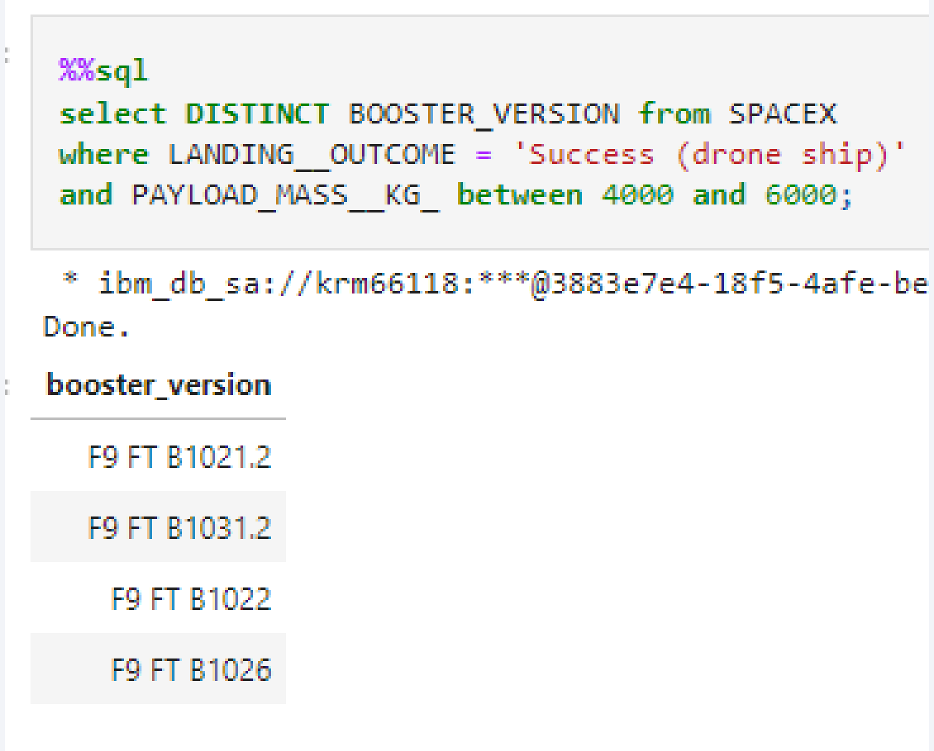
```
%%sql
select MIN(DATE) as FIRST_SUCCESS_LAN from SPACEX
where LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://krm66118:***@3883e7e4-18f5-4afe-be8c-
Done.
```

first_success_lan
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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



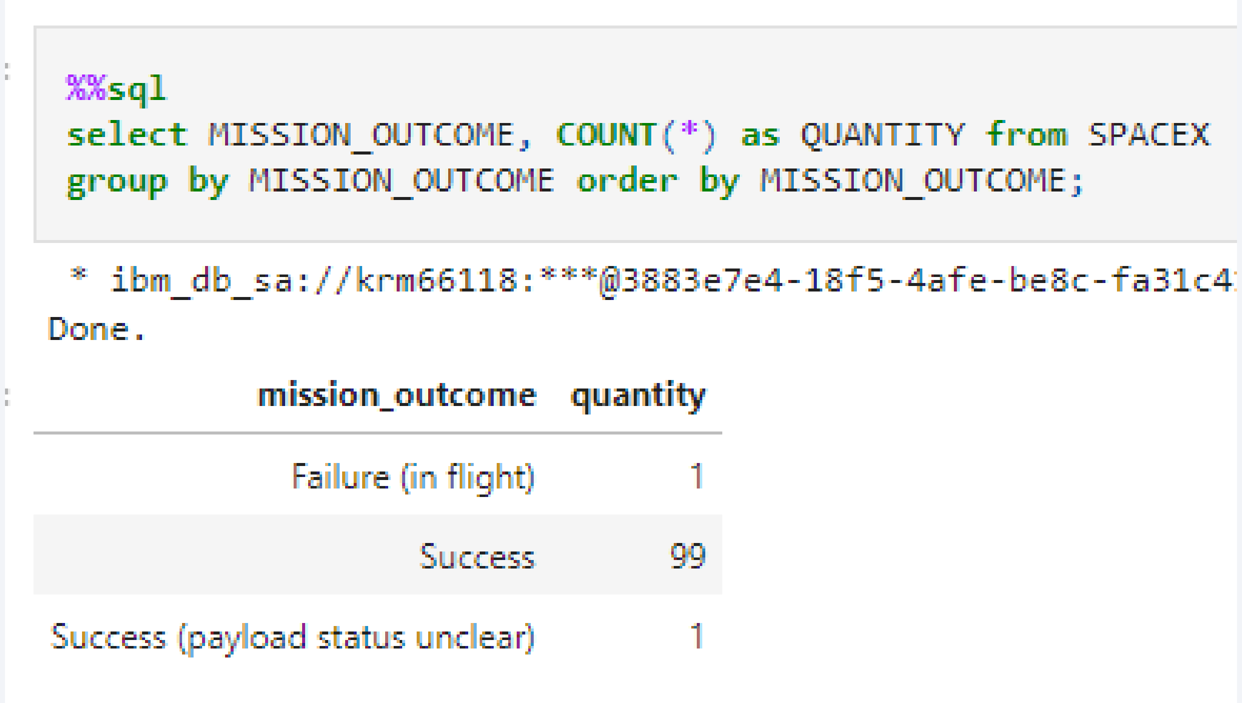
```
%%sql
select DISTINCT BOOSTER_VERSION from SPACEX
where LANDING__OUTCOME = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

```
 * ibm_db_sa://krm66118:***@3883e7e4-18f5-4afe-be
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes



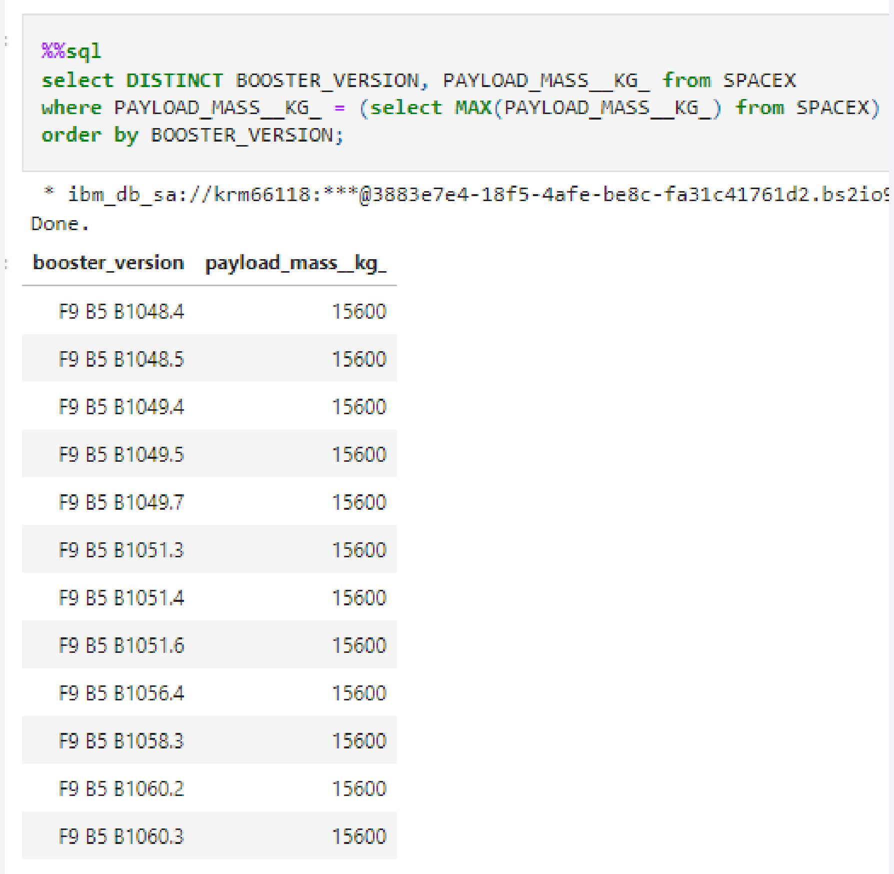
```
%%sql
select MISSION_OUTCOME, COUNT(*) as QUANTITY from SPACEX
group by MISSION_OUTCOME order by MISSION_OUTCOME;
```

```
 * ibm_db_sa://krm66118:***@3883e7e4-18f5-4afe-be8c-fa31c4
Done.
```

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

Boosters Carried Maximum Payload

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



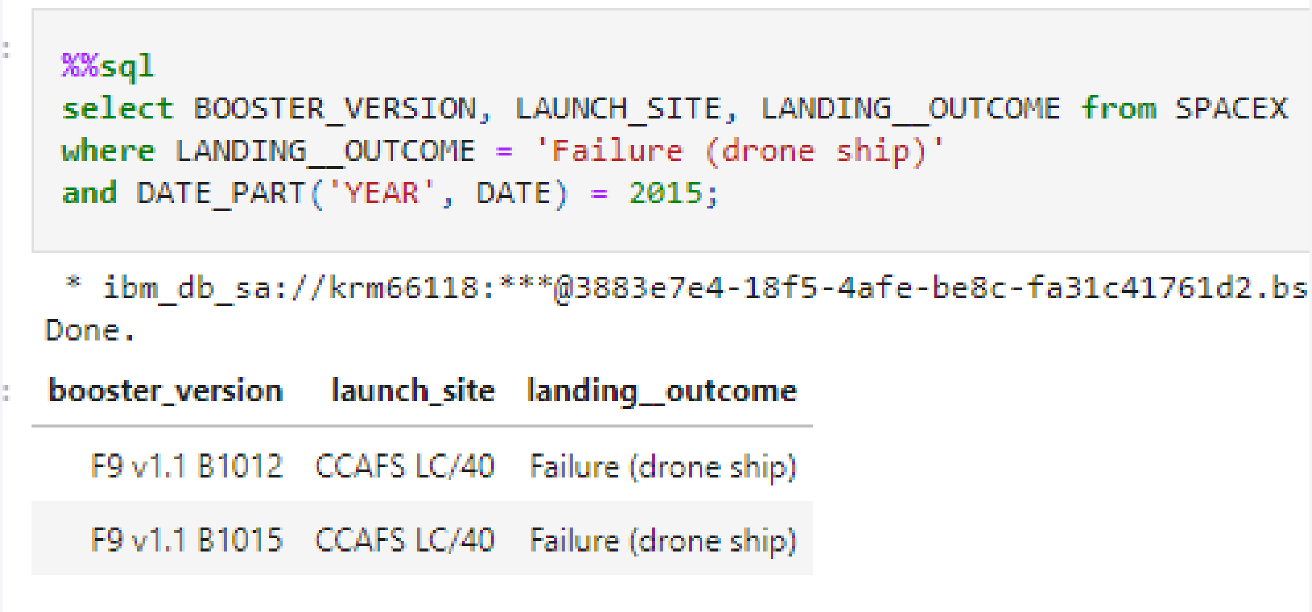
```
%%sql
select DISTINCT BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEX
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEX)
order by BOOSTER_VERSION;
```

```
 * ibm_db_sa://krm66118:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

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



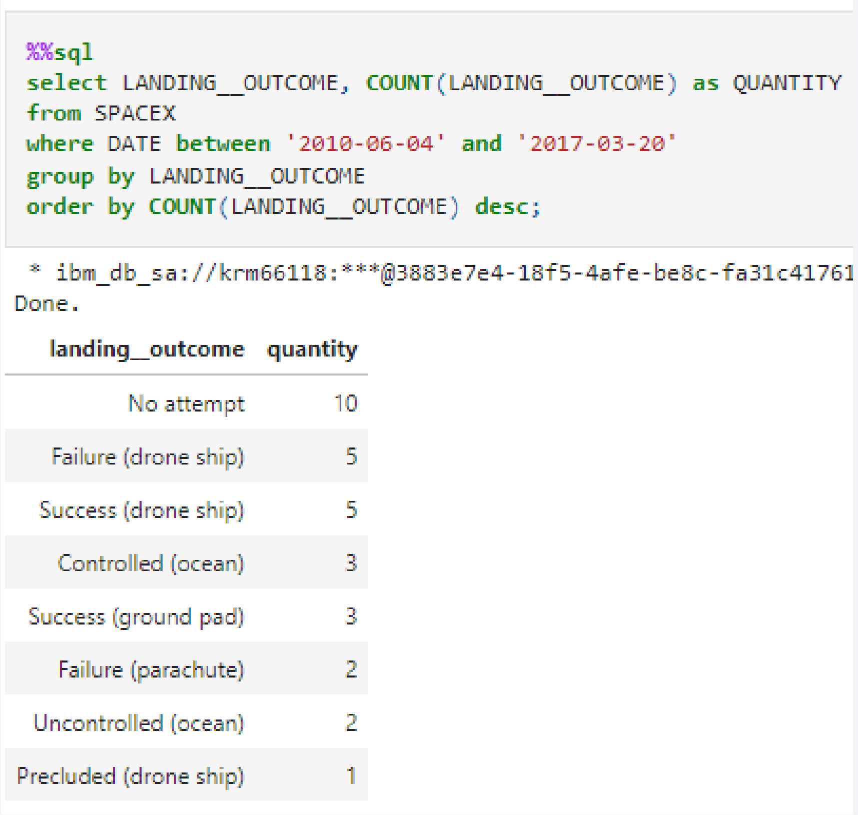
```
%%sql
select BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME from SPACEX
where LANDING__OUTCOME = 'Failure (drone ship)'
and DATE_PART('YEAR', DATE) = 2015;
```

```
 * ibm_db_sa://krm66118:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs
Done.
```

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC/40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC/40	Failure (drone ship)

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

- 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, in descending order



```
%%sql
select LANDING__OUTCOME, COUNT(LANDING__OUTCOME) as QUANTITY
from SPACEX
where DATE between '2010-06-04' and '2017-03-20'
group by LANDING__OUTCOME
order by COUNT(LANDING__OUTCOME) desc;
```

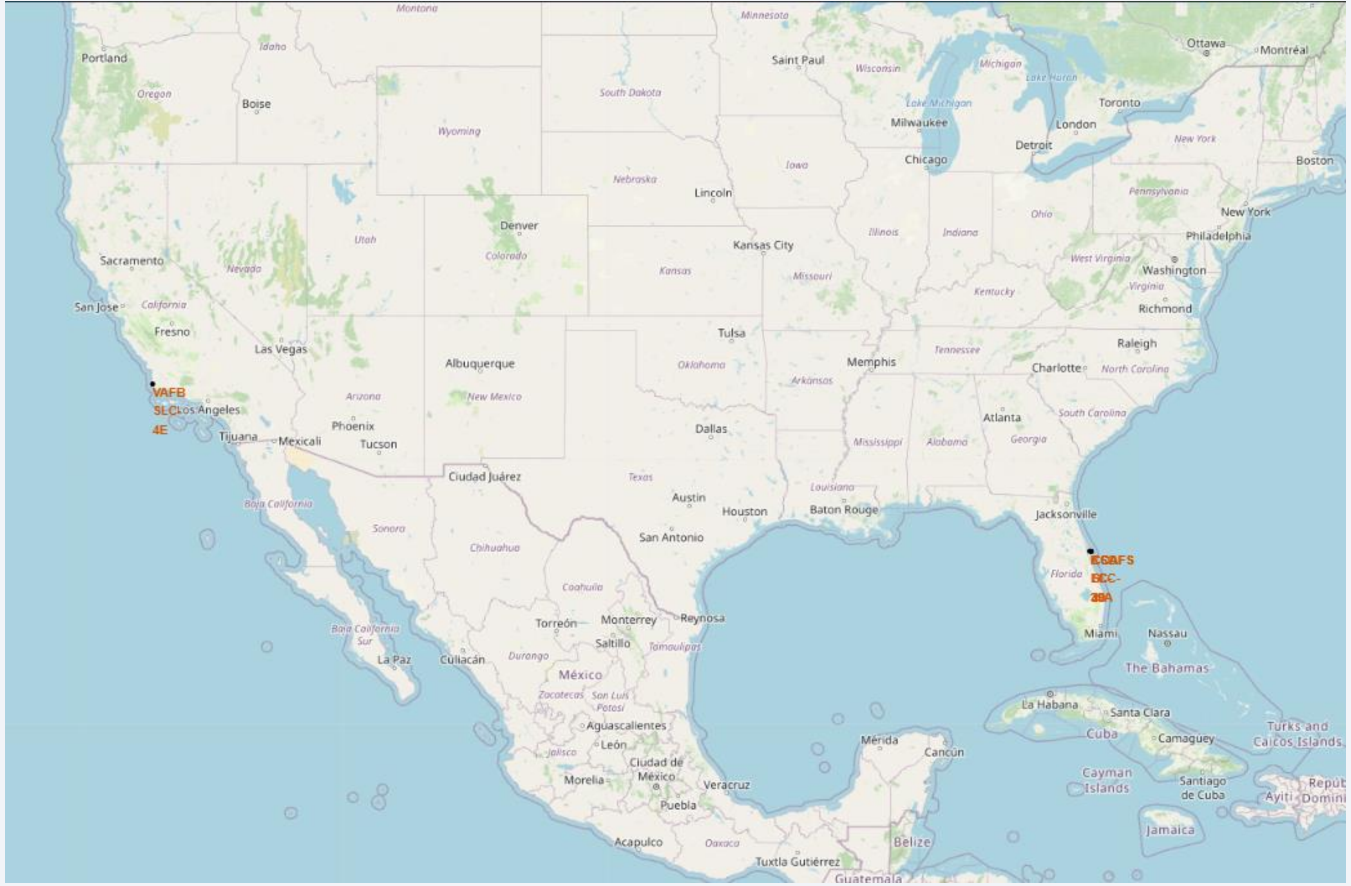
```
 * ibm_db_sa://krm66118:***@3883e7e4-18f5-4afe-be8c-fa31c41761
Done.
```

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

Section 3

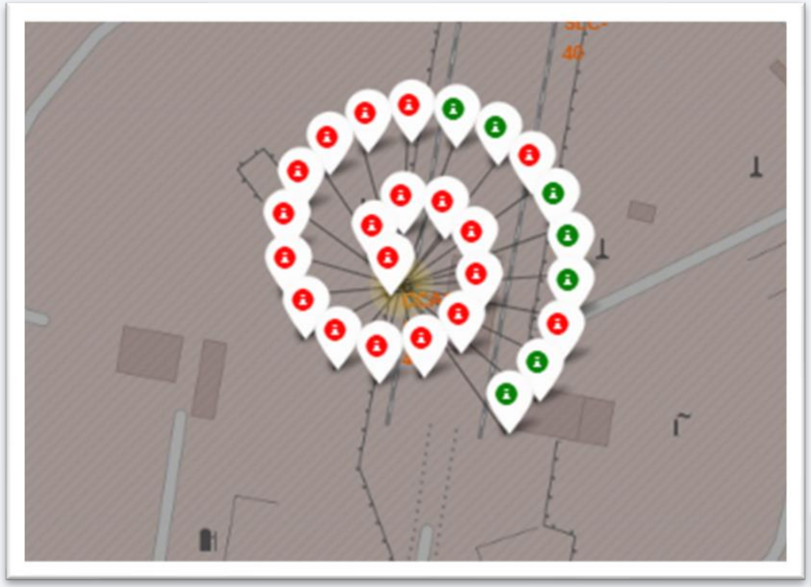
Launch Sites Proximities Analysis

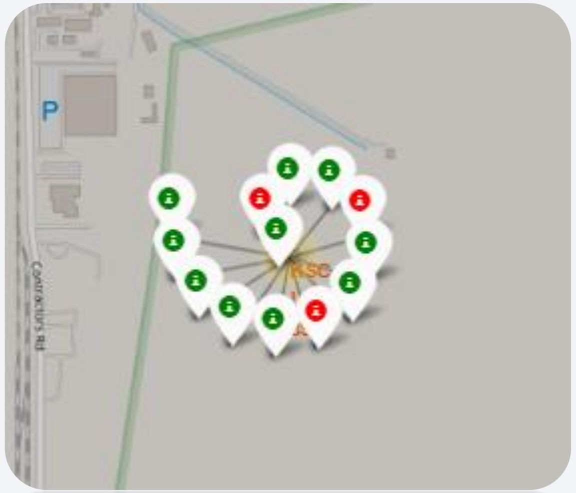
All launch sites

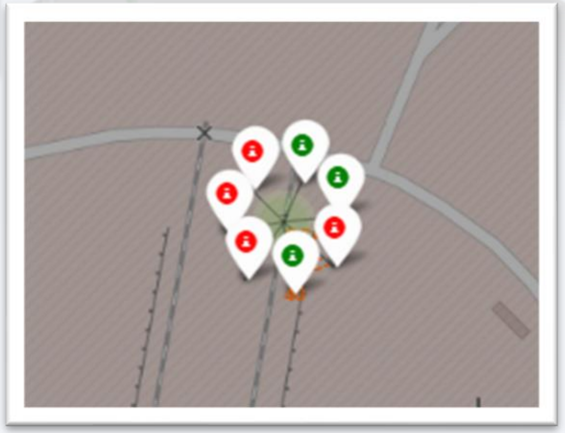


- All launch sites are located in the US in proximity to the ocean

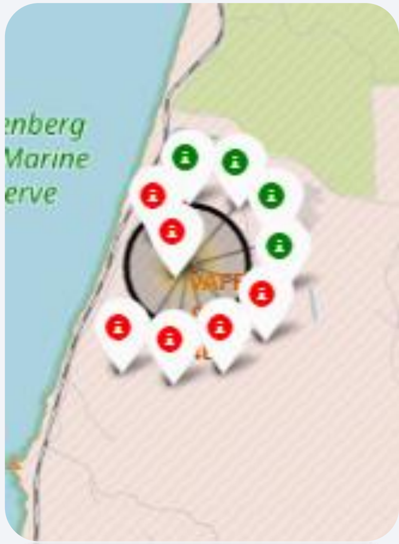
Launch sites with colored markers





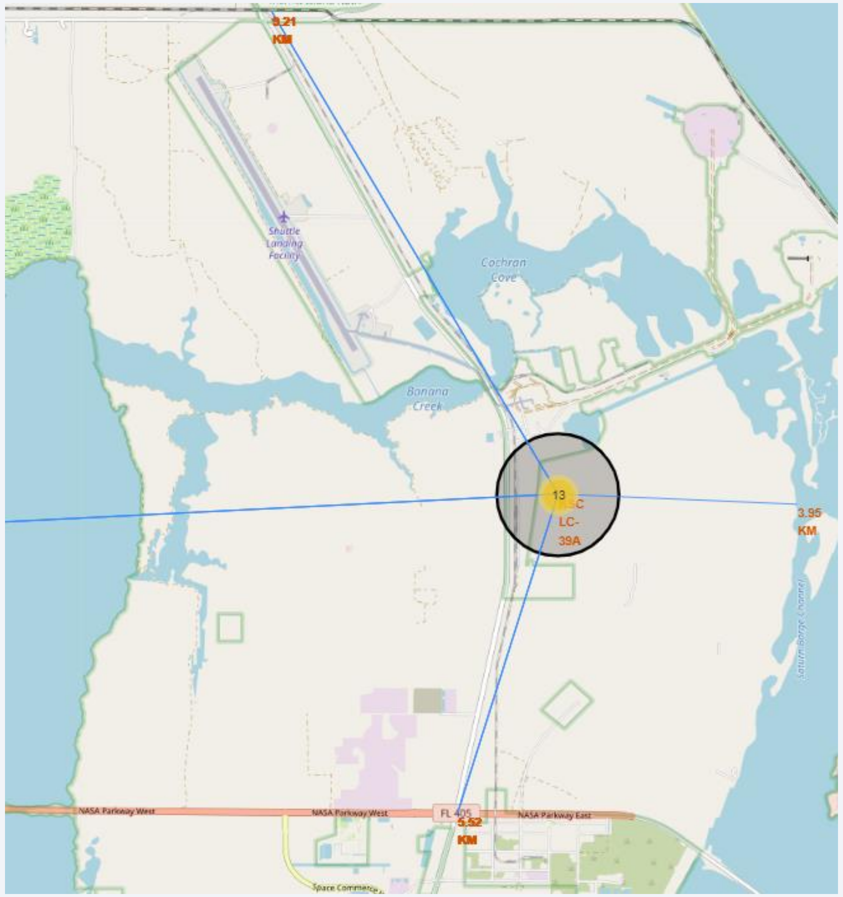


Three Florida launch sites



California launch site

Launch site distance from various landmarks

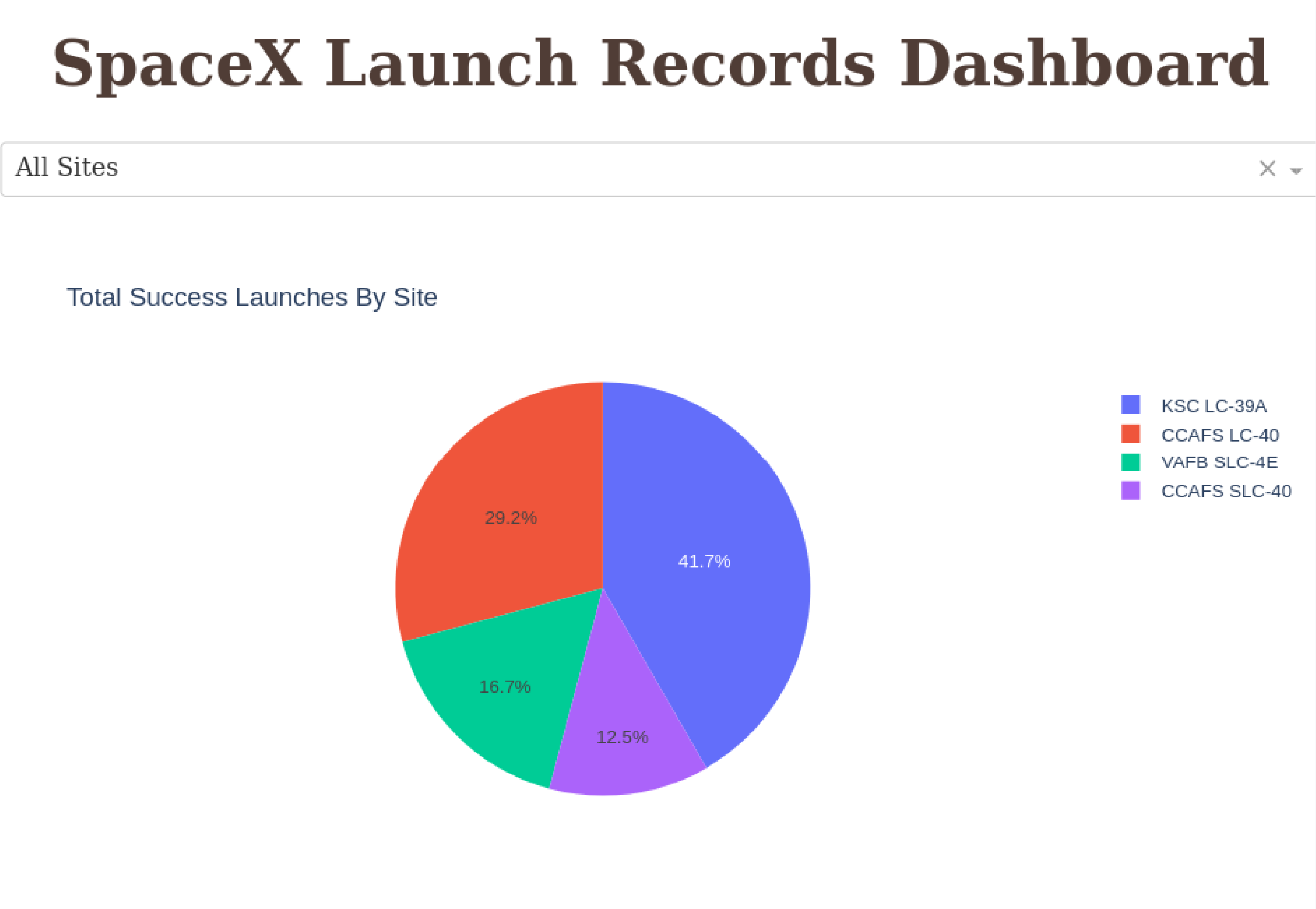


The launch site is relatively close to highways or railroads but rather far away from any cities, also in each case the ocean is near the site.

Section 4

Build a Dashboard with Plotly Dash

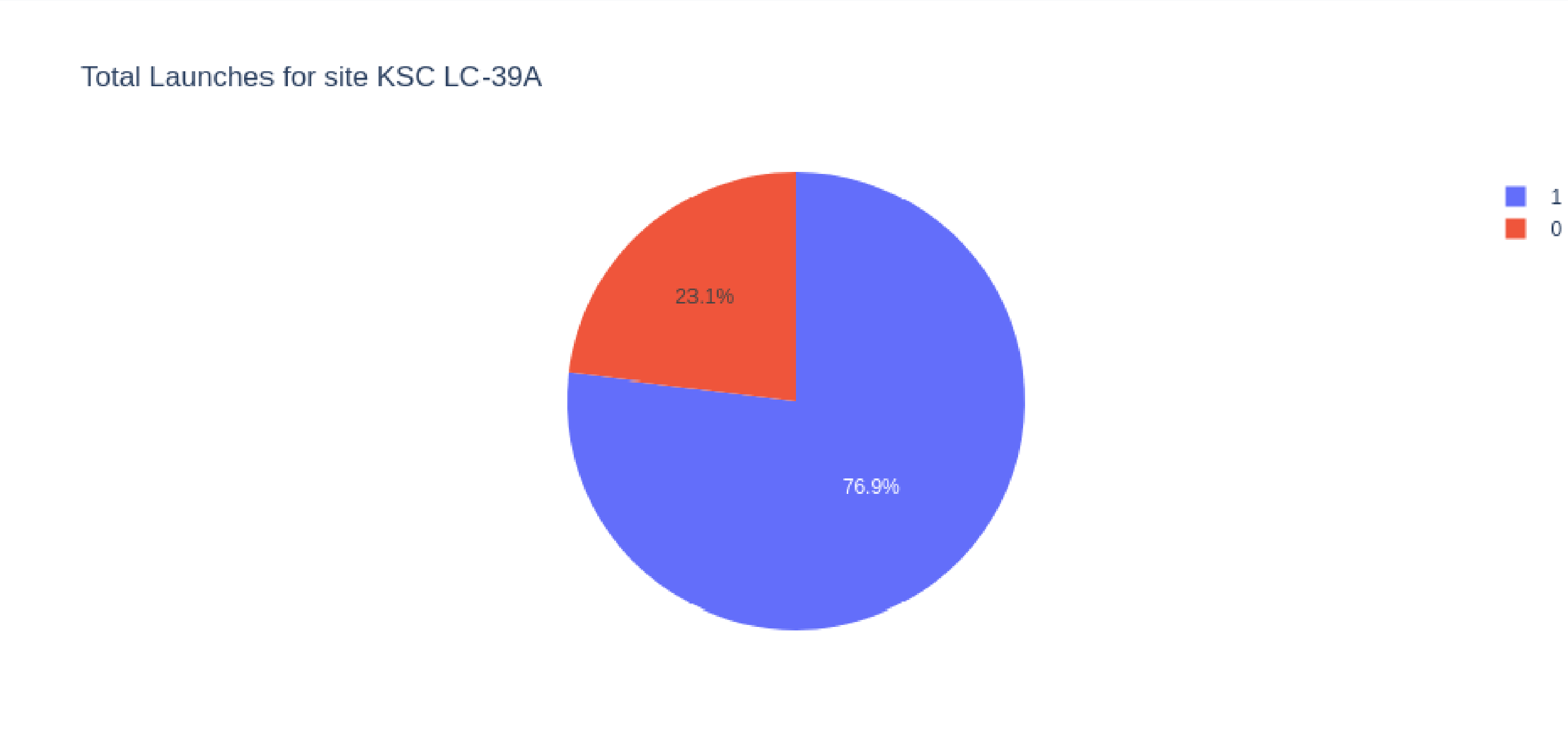
Total launches by site



SpaceX Launch Records Dashboard

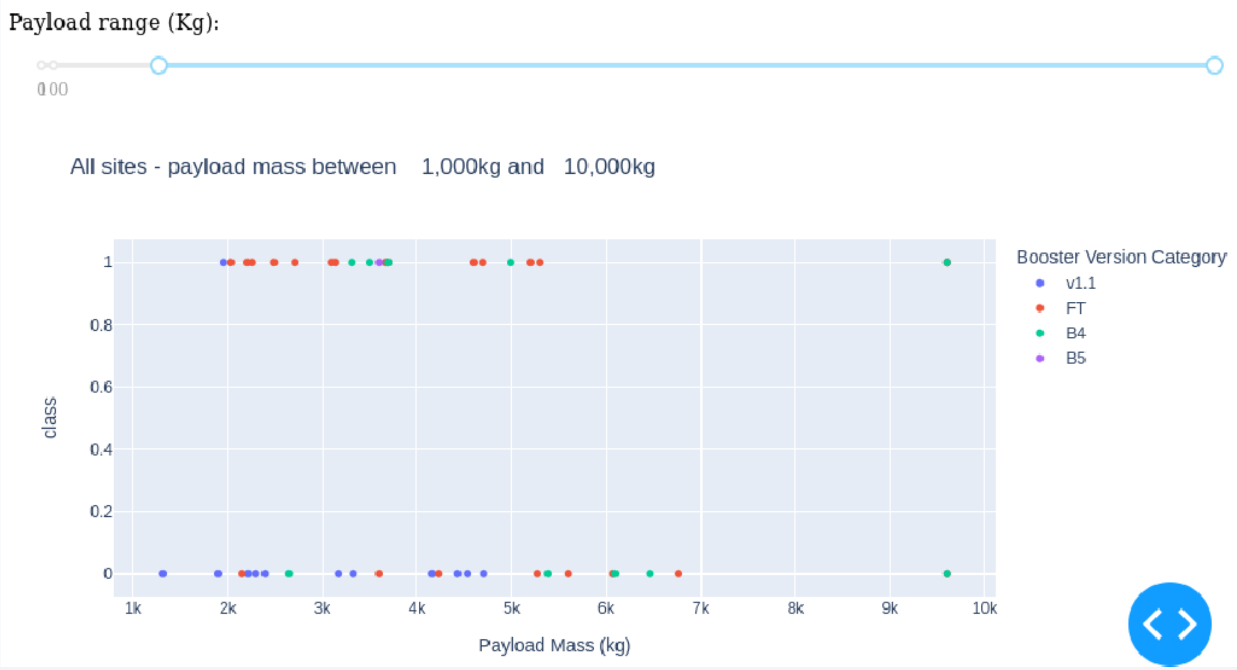
On the pie chart we can see that the most often used site is KSC LC-39A.

The most successful launch site



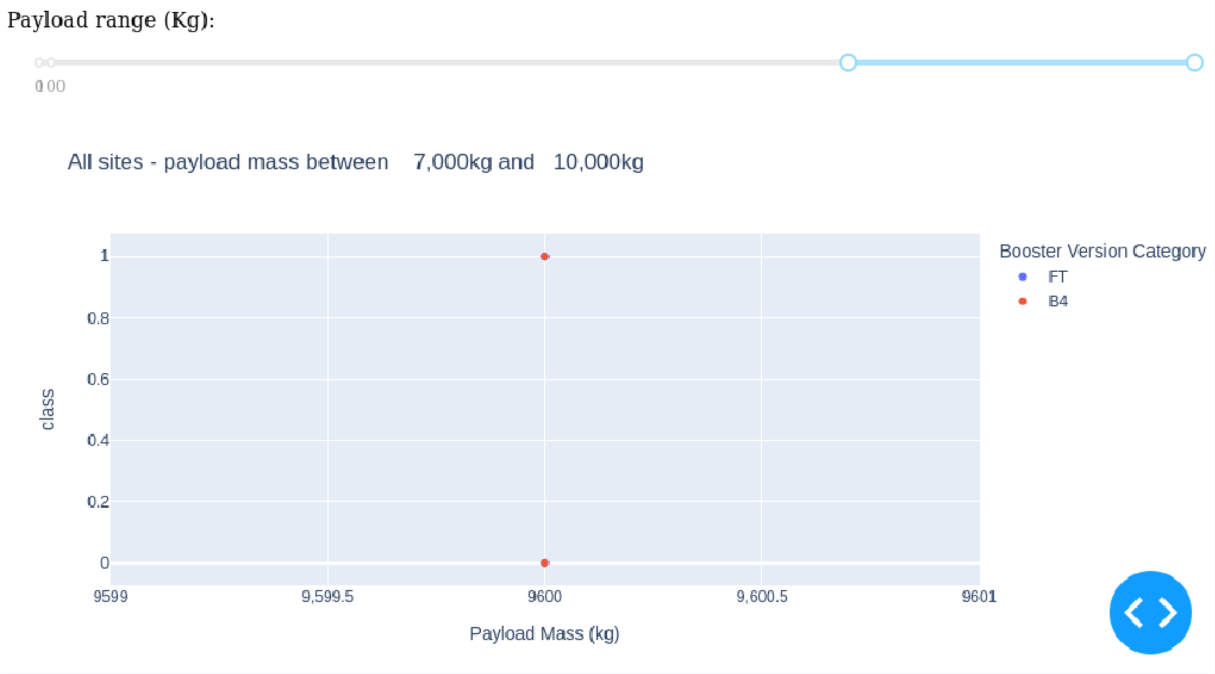
The procentage of success here is 76.9%.

Plots of Payload vs. Launch Outcome



Payload range (Kg):

All sites - payload mass between 1,000kg and 10,000kg



Payload range (Kg):

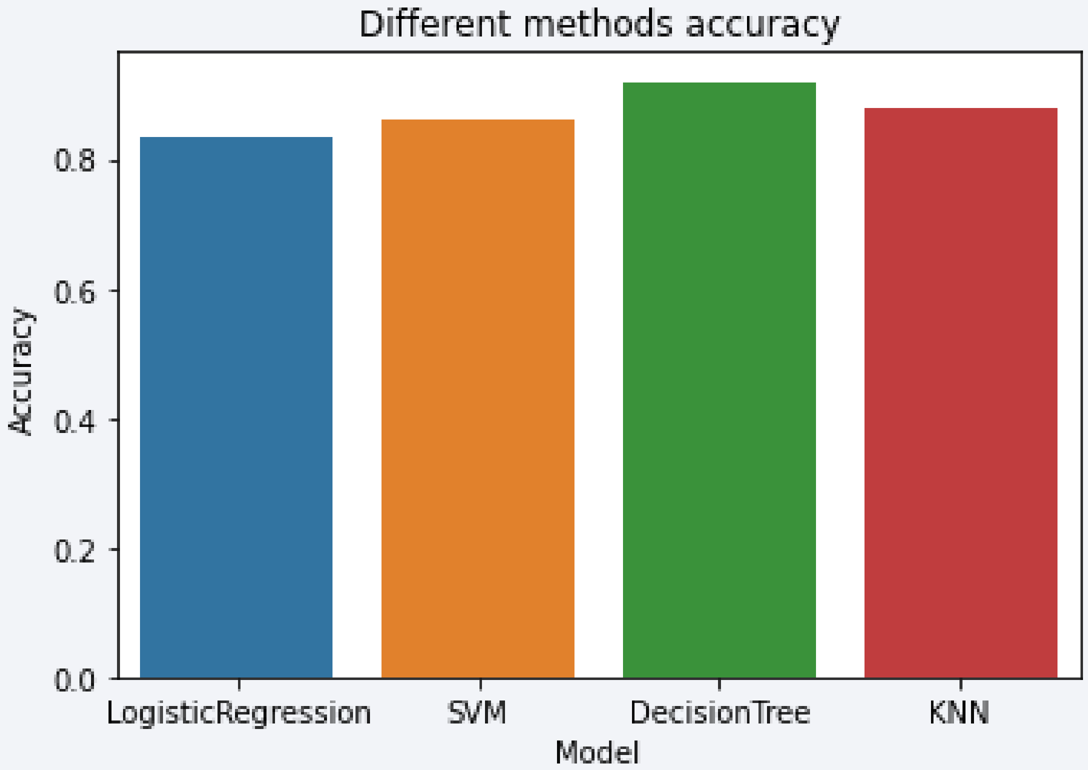
All sites - payload mass between 7,000kg and 10,000kg

We can clearly see that better launch outcome for payloads between 1000kg and 7000kg, when going into higher payloads the positive outcome decreases.

Section 5

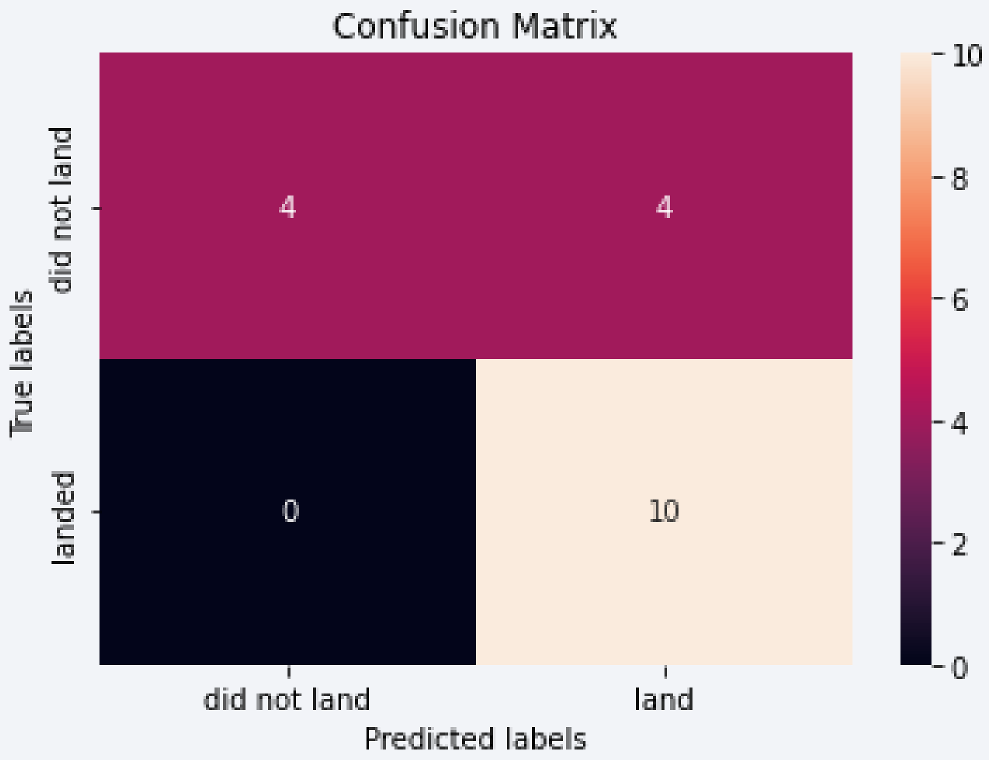
Predictive Analysis (Classification)

Classification Accuracy



In the bar plot wee see that all methods perform well but the best one is decision tree, with the accuracy of aorund 92%.

Confusion Matrix



This is the confusion matrix for the decison tree method. Good thing is that we have zero false negatives, but the worrying thing is false positives, which are landing outcomes marked as succesful by the classifier which in fact are unsuccessful.

Conclusions

Through the analysis we got to know that:

- There are 4 most successful orbits : ES-L1, GEO, HEO, SSO
- The most successful launching site is KSC LC-39A
- Launching sites are located near the ocean
- The best prediction model is Decision Tree model

Thank you!