DATA SCIENCE CAPSTONE

Den Bych
30.05.2023

Outline

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

Executive Summary

Summary of methdologies:

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of results:

- Exploratory Data Analysis results
- Interactive Visual Analytics
- Prediction Analysis

Introduction

- In this capstone, 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 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. We are going to analyze public information and determine which landing is success and which are not. We will use exploratory, visual and predictive analysis for this purpose.
- What influences a successful landing? What are the rocket and environmental requirements?

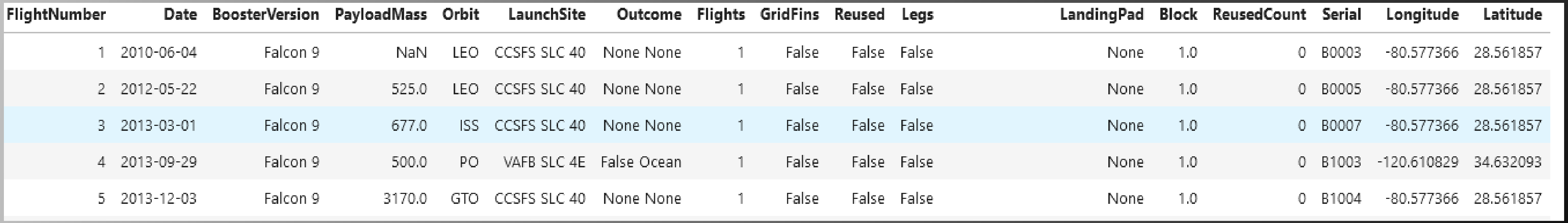
Methodology

All steps of project:

1. Data Collection and Wrangling(sources: SpaceX API, Web Scraping from the Wikipedia)
2. Exploratory Data Analysis(EDA), using:
    - SQL
    - Pandas
3. Data Visualization, using:
    - Matpoltlib and Seaborn
    - Folium
    - Dash
4. Machine learning, using:
    - Logistic regression
    - SVM
    - Decision Tree
    - KNN

Data Collection – SpaceX API

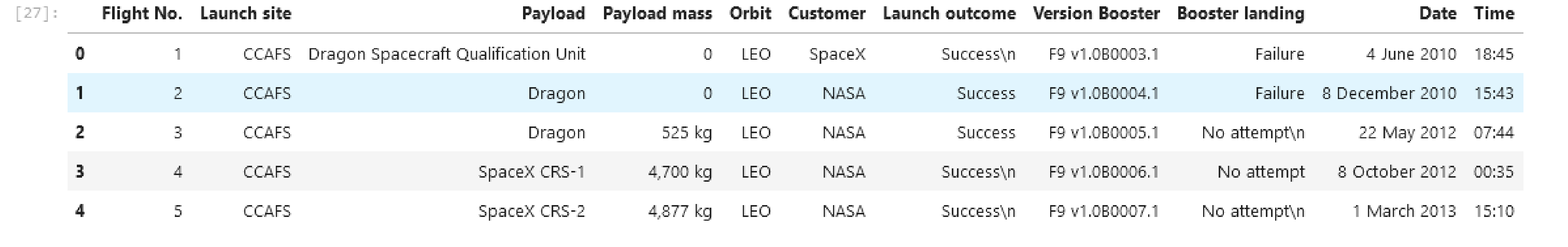
- One way to collect data is to use the SpaceX API. Like this: https://api.spacexdata.com/v4/rockets/
- Data are filtered to include only Falcon 9 rockets. Missing values are replaced by the mean of the corresponding column.



FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Data Collection - Scraping

- Another way is web scraping from different websites. We will use this link now: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches?utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01
- Data contain only Falcon 9 launches.



[27]:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

- ✓ At this stage we try to understand what we have and what we will do with the data. We are introduced to the data set and its characteristics.
- ✓ How do we analyze the data and create a new feature such as Class to understand what is a successful landing or not.

EDA with SQL

Performed SQL queries:

- 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 achieved
- 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
- 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

EDA with Data Visualization

- ✓ Charts were plotted: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend. Scatter plots show the relationship between variables. If a relationship exists and has a good correlation, it could be used in a machine learning model.
- ✓ Bar charts show comparisons between discrete categories. The aim is to show the relationship between the specific categories being compared and a measured value.
- ✓ Line charts show trends in data over time (time series).

Build an Interactive Map with Folium

- ✓ Added circle, pop-up and text markers for all launch pads with their latitude and longitude coordinates to show their geographic location and proximity to the equator and coastline.
- ✓ Coloured markers for successful (green) and unsuccessful (red) launches have been added using a cluster of markers to identify which launch pads have a relatively high success rate.
- ✓ Coloured lines have been added to show the distances between launch sites and their immediate surroundings, such as railways, highways, coastlines and the nearest town.

Build a Dashboard with Plotly Dash

- ✓ Added a dropdown list to the Launch Site selection.
- ✓ Added a pie chart to show the total number of successful launches for all sites and the number of successes vs. failures for the site when a specific launch site has been selected.
- ✓ Added a slider to select the payload range.
- ✓ Added a scatter plot to show the correlation between payload and launch success.

Predictive Analysis (Classification)

The scikit-learn library is used to develop predictive models, evaluate them, select the best and preprocess the data.

In this step used:

1. Data standardization
2. Creation of a training and test set from the original data set
3. Building and comparing logistic regression, SVM, decision tree and KNN models

Results

The results are divided into 4 parts:

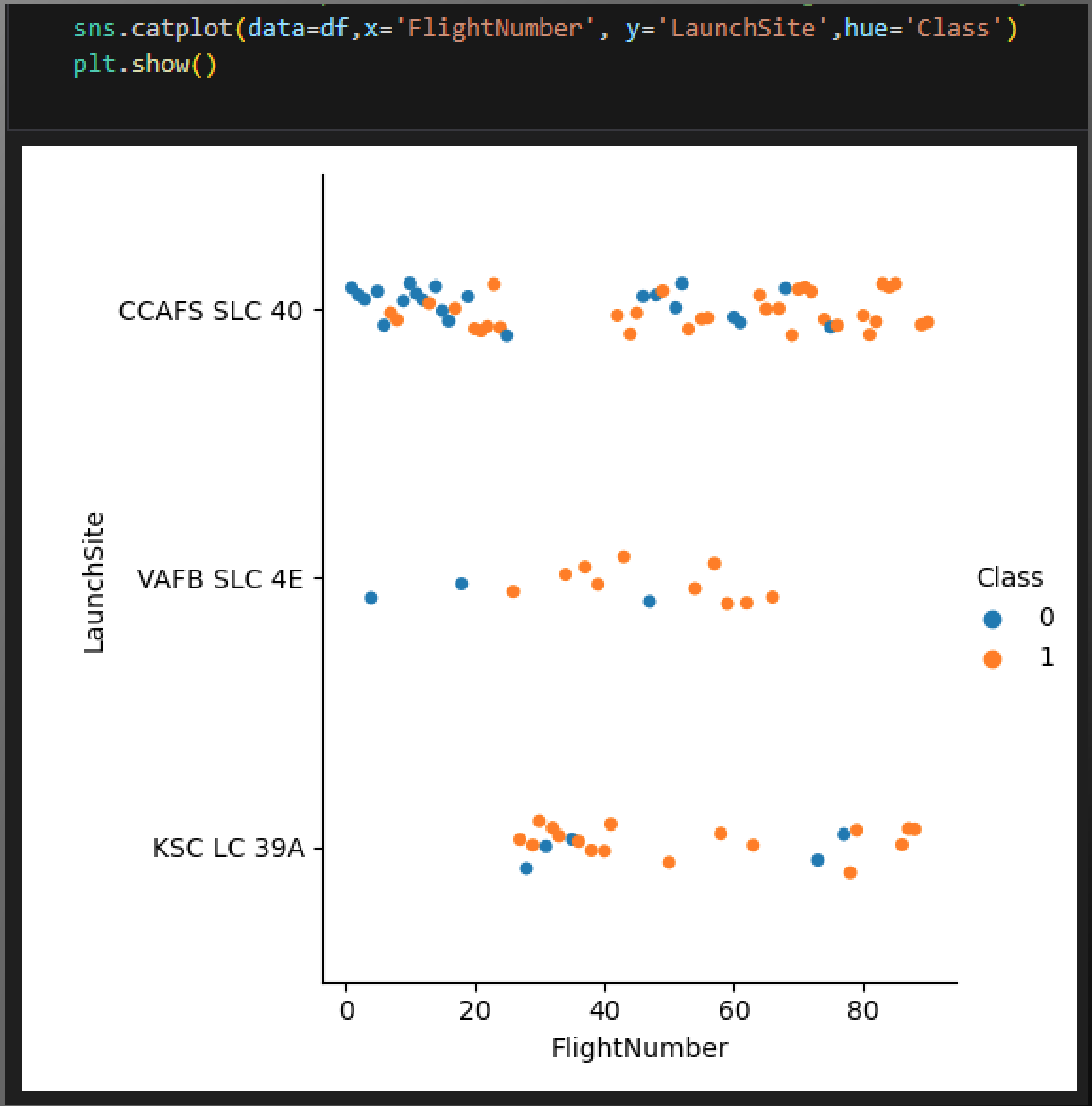
- Exploratory data analysis results(with plots and maps and SQL queries)
- Interactive analytics demo in screenshots(dashboard)
- Predictive analysis results(4 predictive models)

EDA VIA VISUALIZATION

Flight Number vs. Launch Site

Explanation:

- As the number of launches increases, so does the likelihood of success.
- CCAFS SLC 40 is the most popular Launch Site

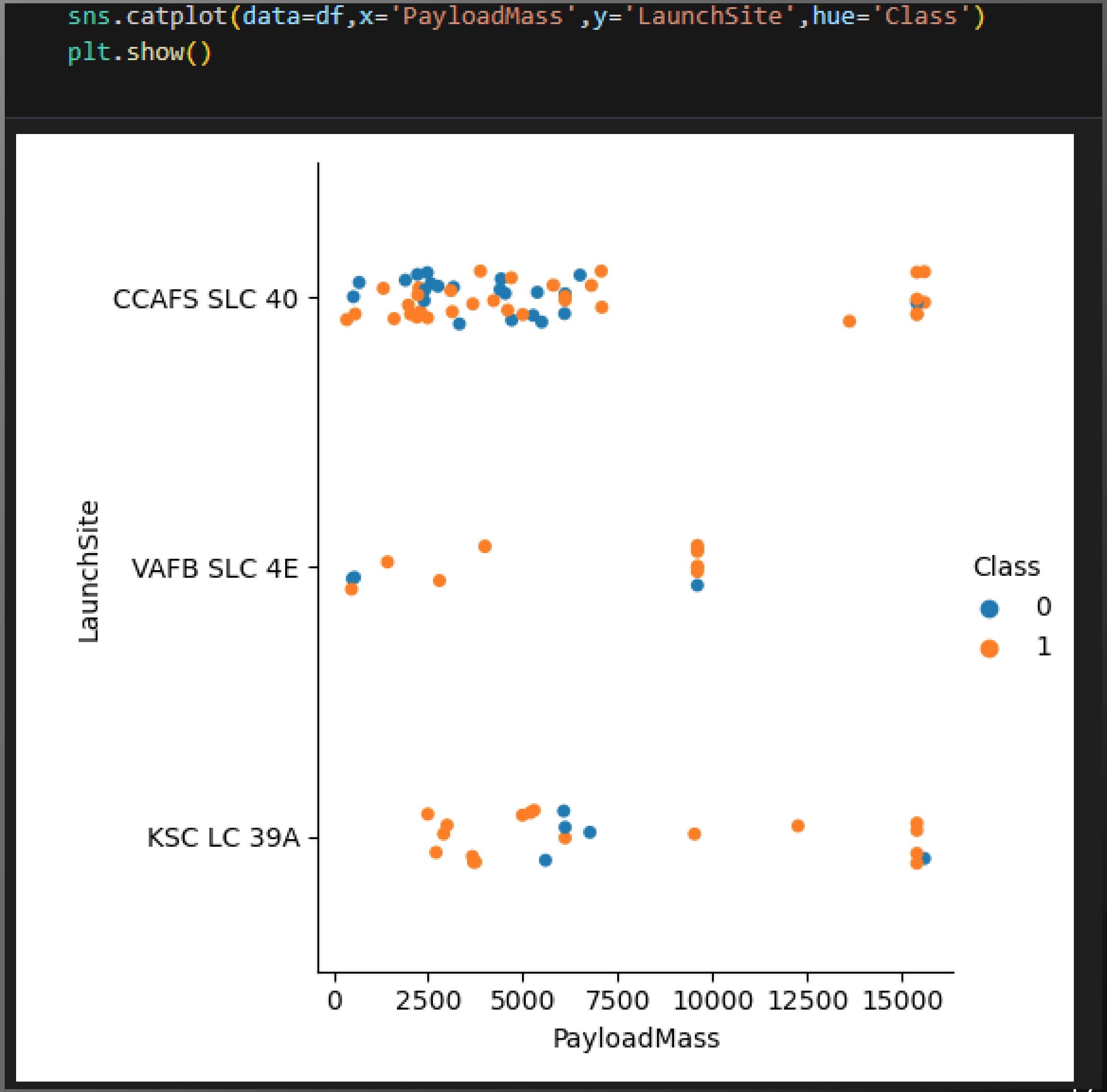


```
sns.catplot(data=df,x='FlightNumber', y='LaunchSite',hue='Class')
plt.show()
```

Payload vs. Launch Site

Explanation:

- Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- CCAFS SLC 40 is good choice for extra heavy rockets(>10000)
- KSC LC 39A has 100% success rate for payload mass under 5000

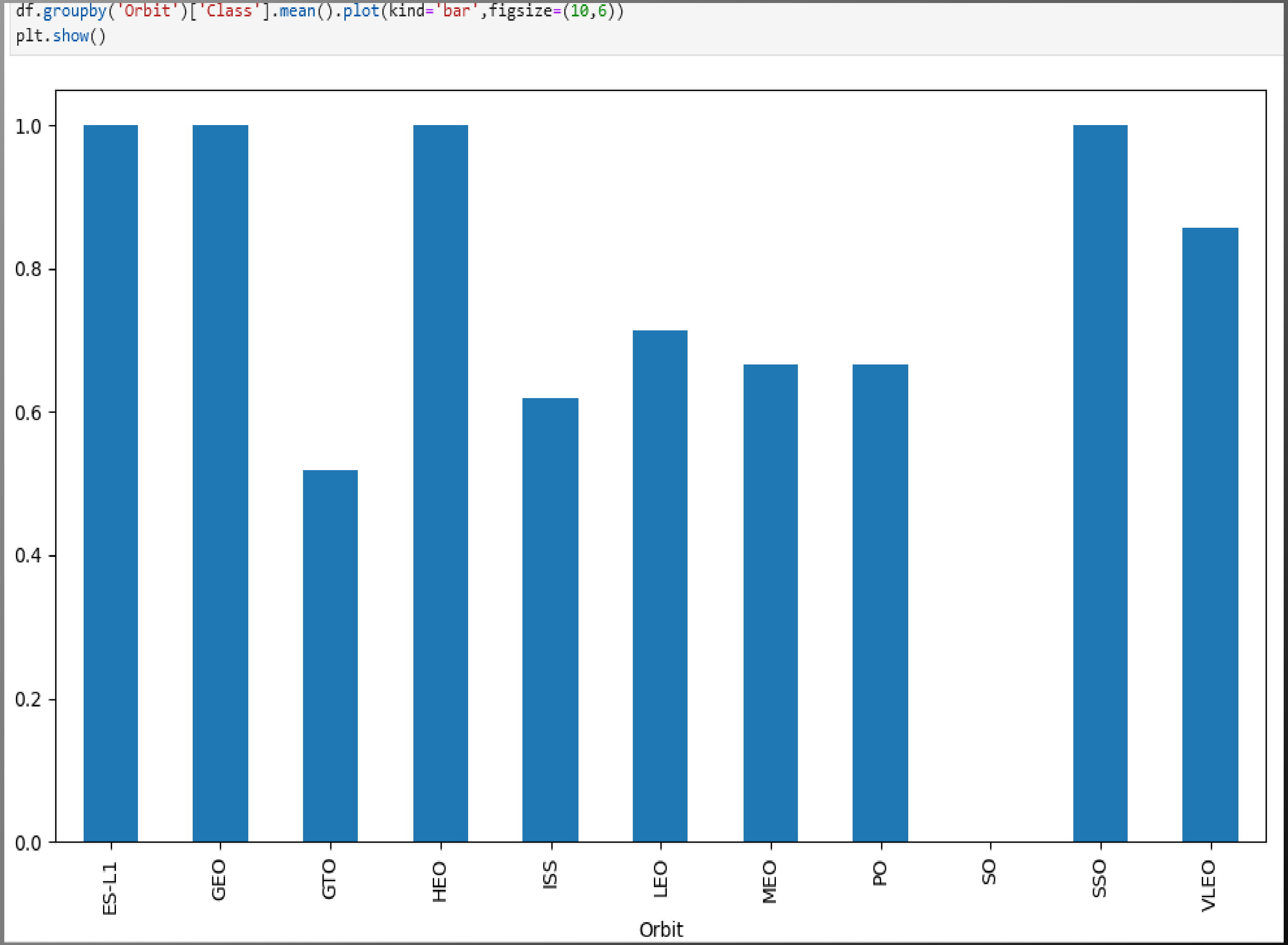


```
sns.catplot(data=df,x='PayloadMass',y='LaunchSite',hue='Class')
plt.show()
```

Success Rate vs. Orbit Type

Explanation:

- The ES-L1, GEO, HEO,SSO has 100% success rate
- The SO orbit has 0% rate, but it was just one launch

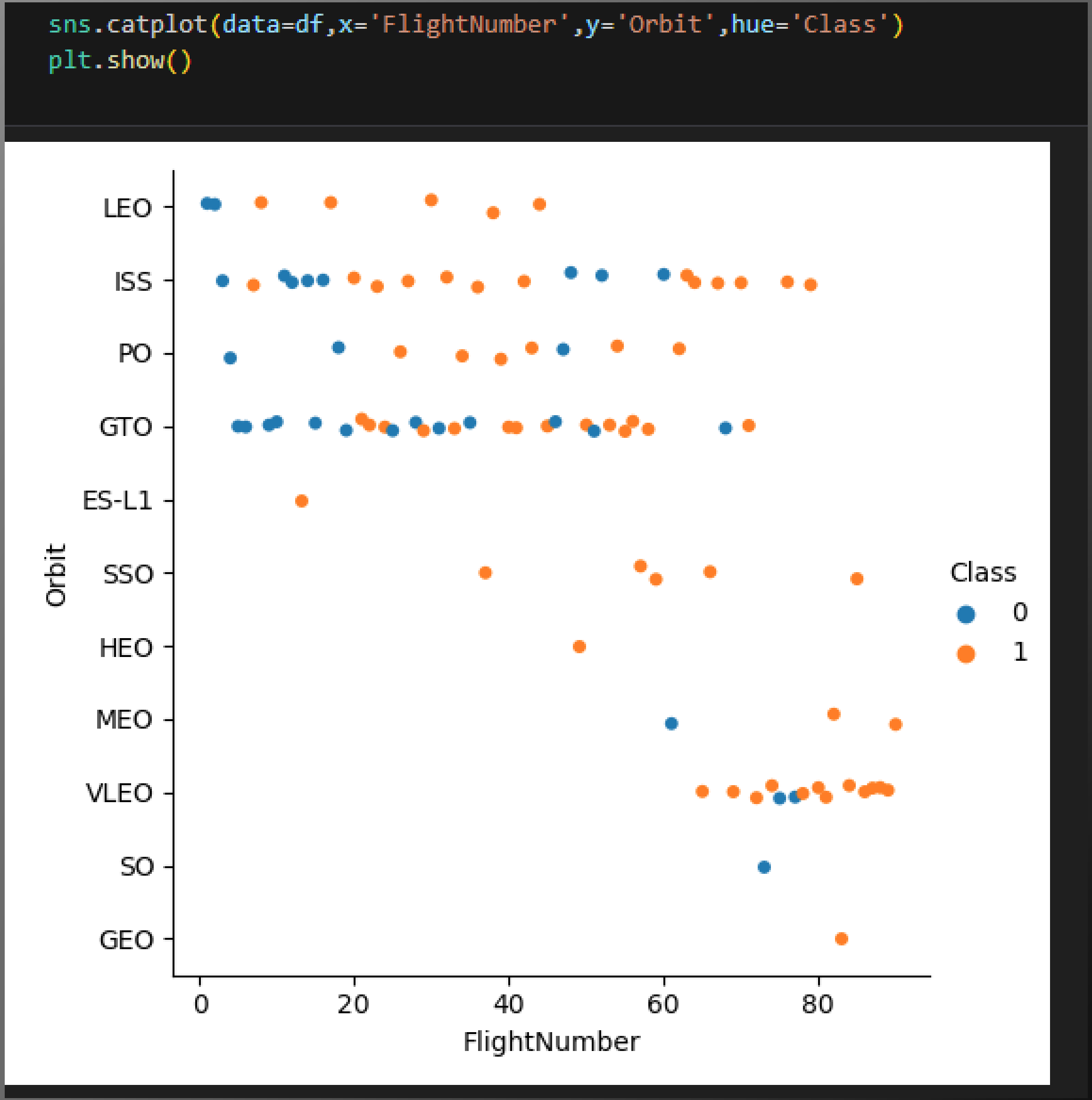


```
df.groupby('Orbit')['Class'].mean().plot(kind='bar',figsize=(10,6))
plt.show()
```

Flight Number vs. Orbit Type

Explanation:

- The LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

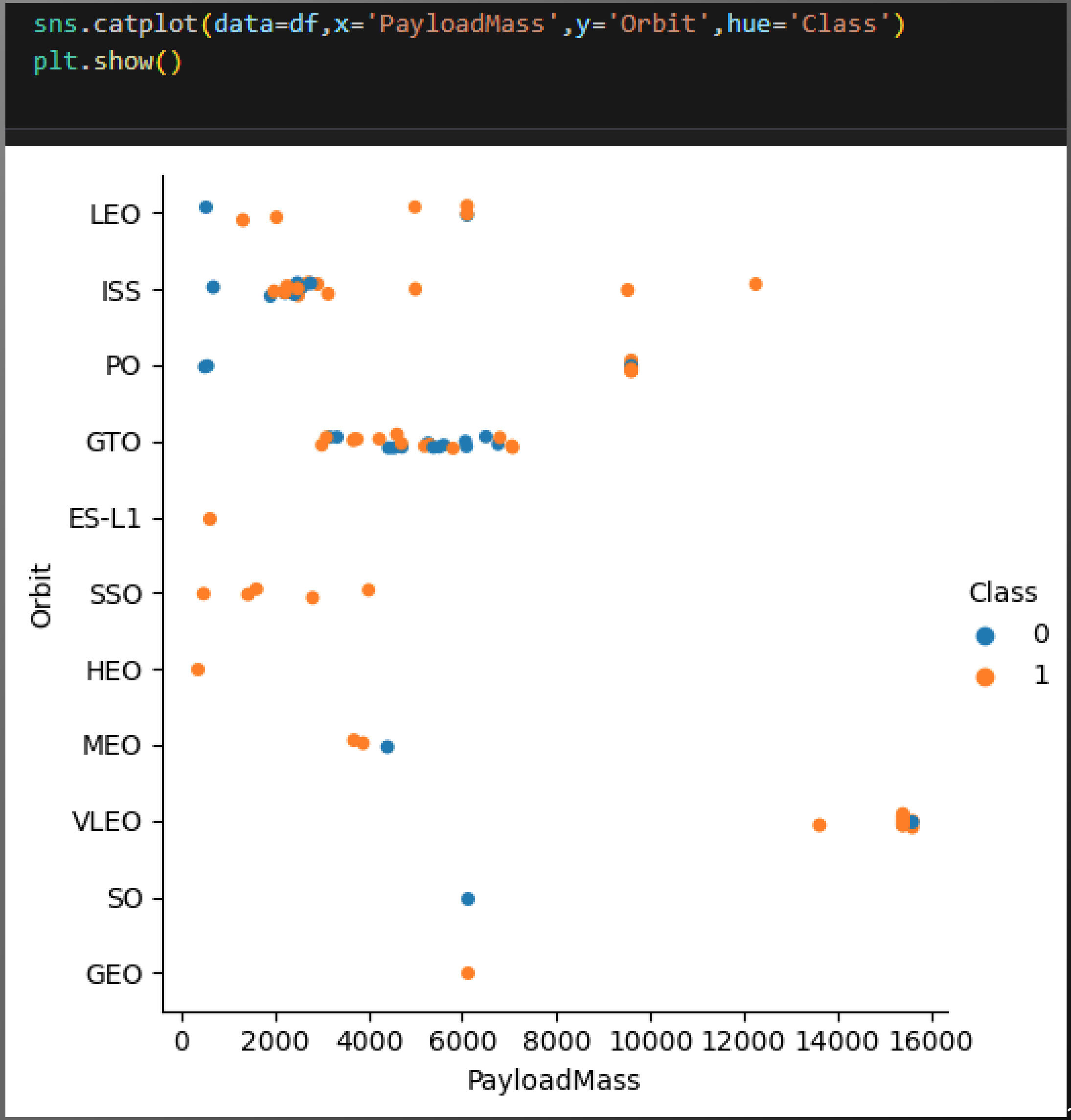


```
sns.catplot(data=df,x='FlightNumber',y='Orbit',hue='Class')
plt.show()
```

Payload vs. Orbit Type

Explanation:

- 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 both positive landing rate and negative landing(unsuccessful mission) are both there here.

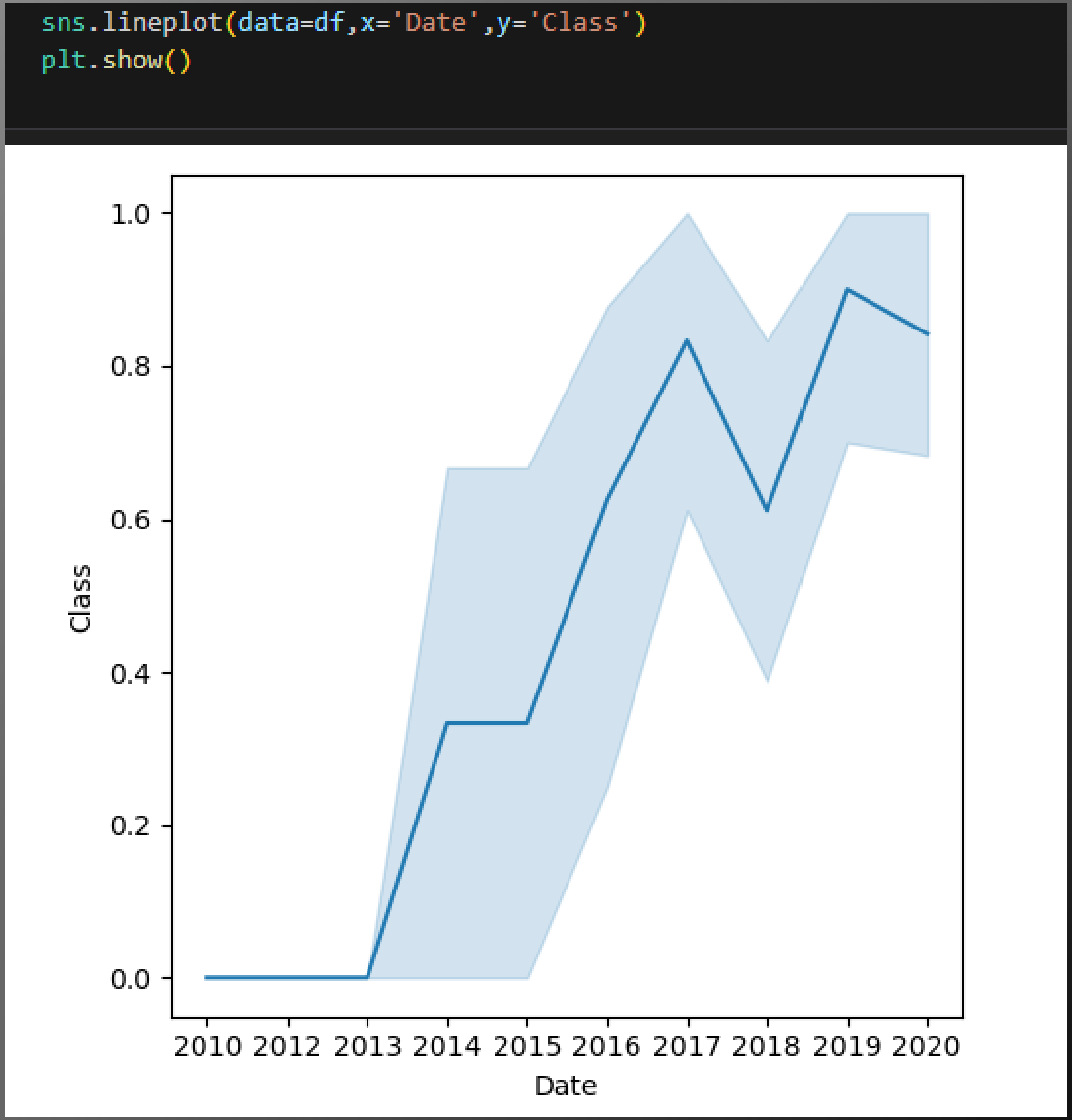


```
sns.catplot(data=df,x='PayloadMass',y='Orbit',hue='Class')
plt.show()
```

Launch Success Yearly Trend

Explanation:

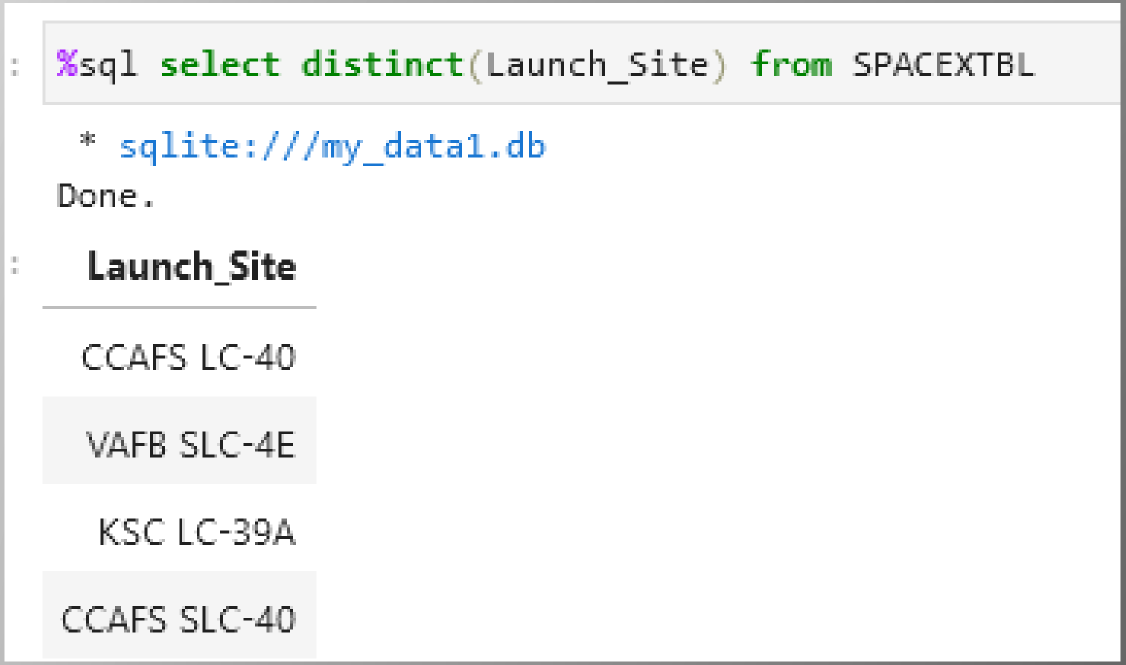
- The success rate since 2013 kept increasing till 2020.



```
sns.lineplot(data=df,x='Date',y='Class')
plt.show()
```

EDA with SQL

All Launch Site Names



```
%sql select distinct(Launch_Site) from SPACEXTBL
```

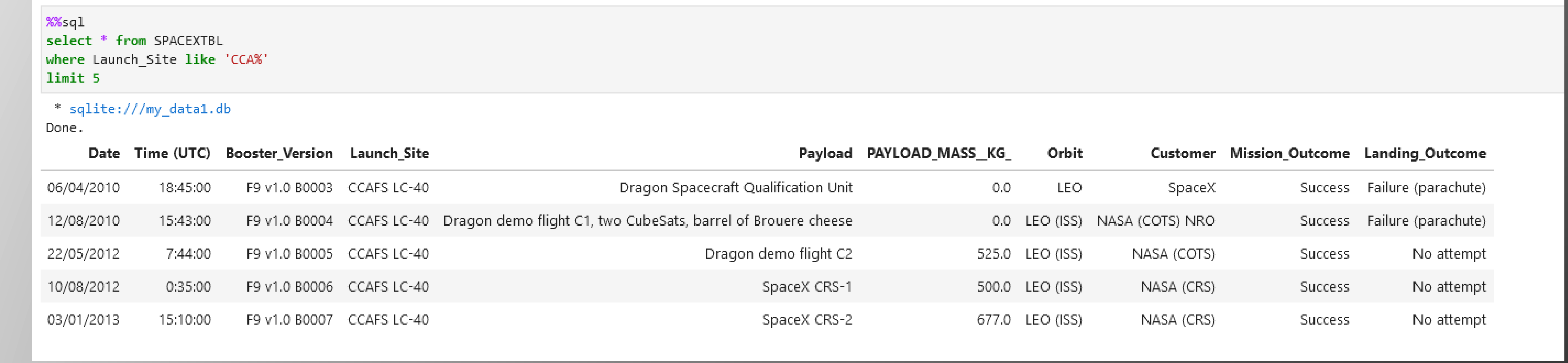
```
 * sqlite:///my_data1.db
Done.
```

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

Explanation:

- All unique launch site names.

Launch Site Names Begin with 'CCA'



```
%%sql
select * from SPACEXTBL
where Launch_Site like 'CCA%'
limit 5
```

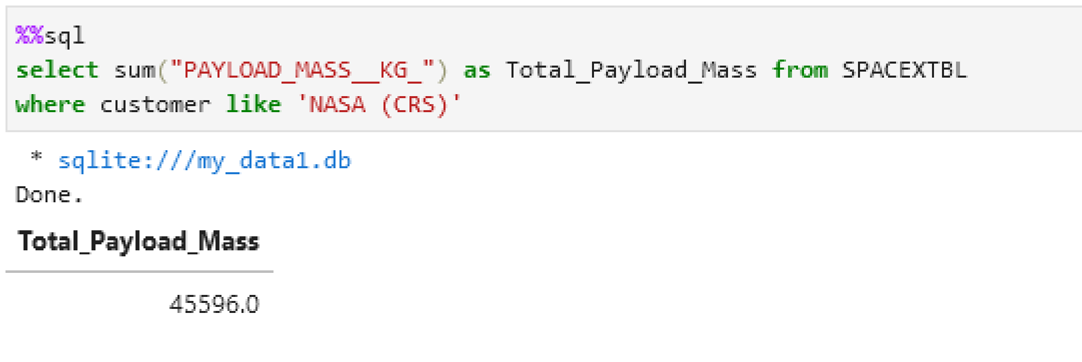
```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation:

- Displaying 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass



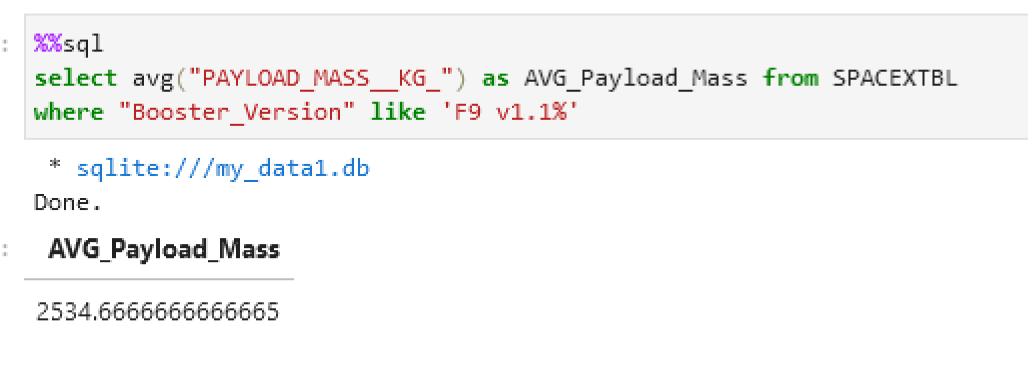
```
%%sql
select sum("PAYLOAD_MASS__KG_") as Total_Payload_Mass from SPACEXTBL
where customer like 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
Total_Payload_Mass
           45596.0
```

Explanation:

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

Average Payload Mass by F9 v1.1



```
%%sql
select avg("PAYLOAD_MASS__KG_") as AVG_Payload_Mass from SPACEXTBL
where "Booster_Version" like 'F9 v1.1%'
```

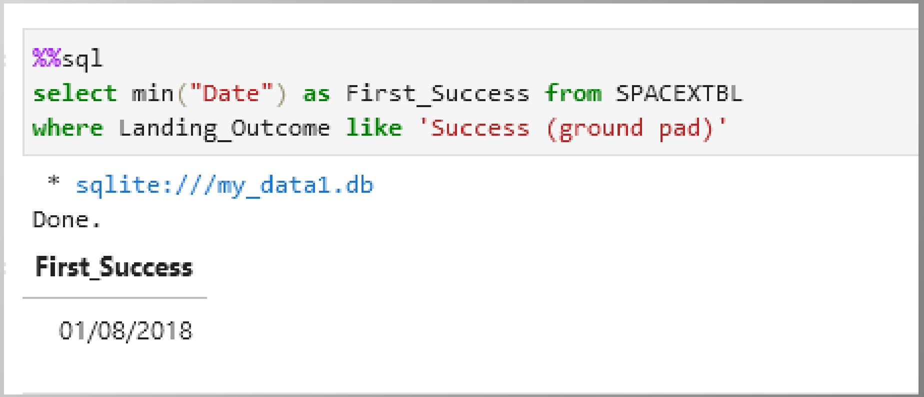
```
 * sqlite:///my_data1.db
Done.
```

AVG_Payload_Mass
2534.6666666666665

Explanation:

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

First Successful Ground Landing Date



```
%%sql
select min("Date") as First_Success from SPACEXTBL
where Landing_Outcome like 'Success (ground pad)'
```

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

First_Success
01/08/2018

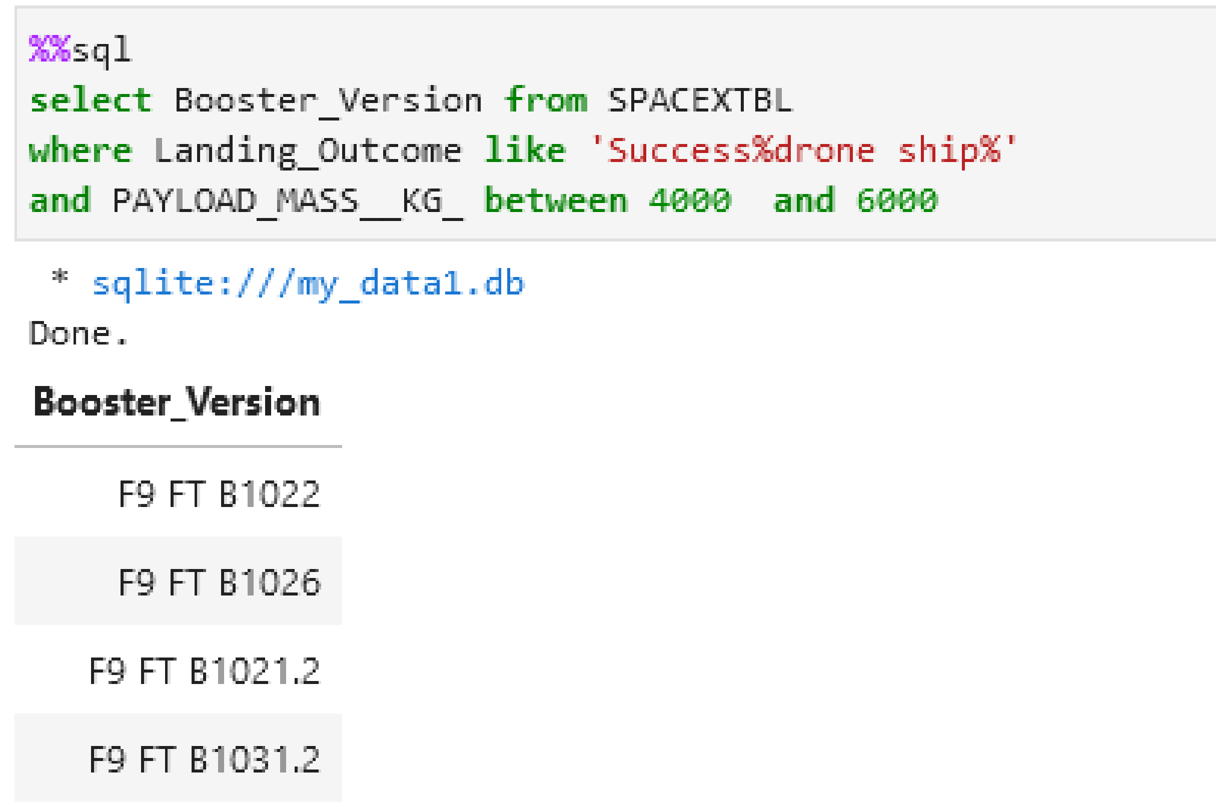
Explanation:

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

Successful Drone Ship Landing with Payload between 4000 and 6000

Explanation:

- **List of 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 SPACEXTBL
where Landing_Outcome like 'Success%drone ship%'
and PAYLOAD_MASS__KG_ between 4000  and 6000
```

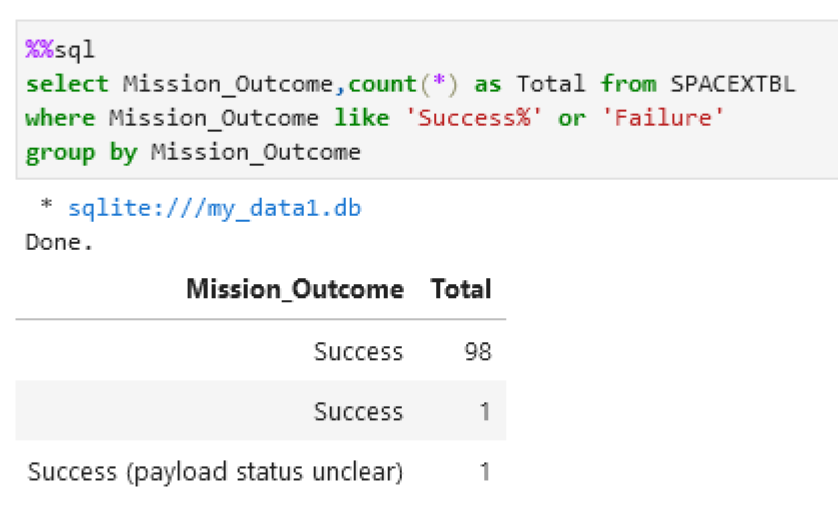
```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

Explanation:

- **List of the total number of successful and failure mission outcomes**



```
%%sql
select Mission_Outcome,count(*) as Total from SPACEXTBL
where Mission_Outcome like 'Success%' or 'Failure'
group by Mission_Outcome
```

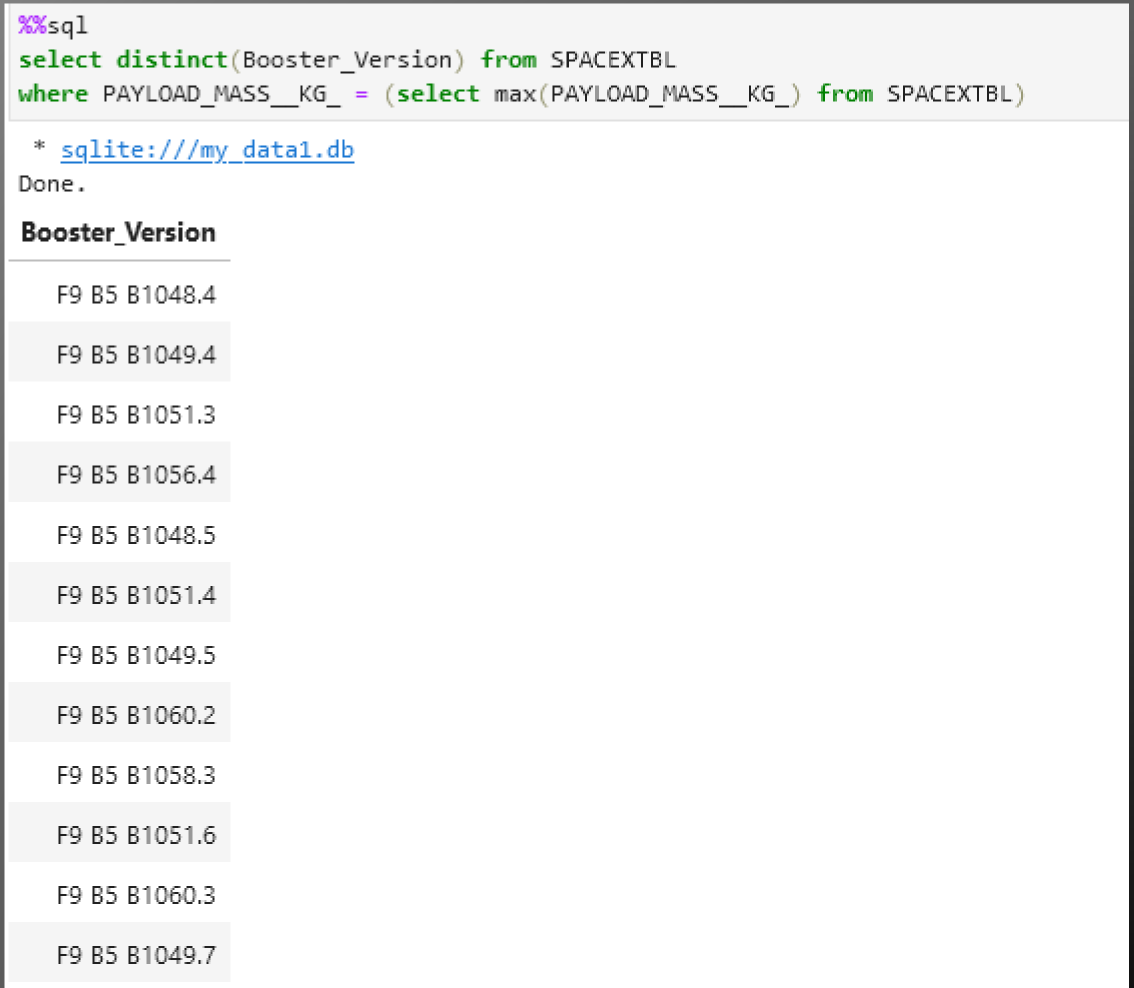
```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	Total
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Explanation:

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

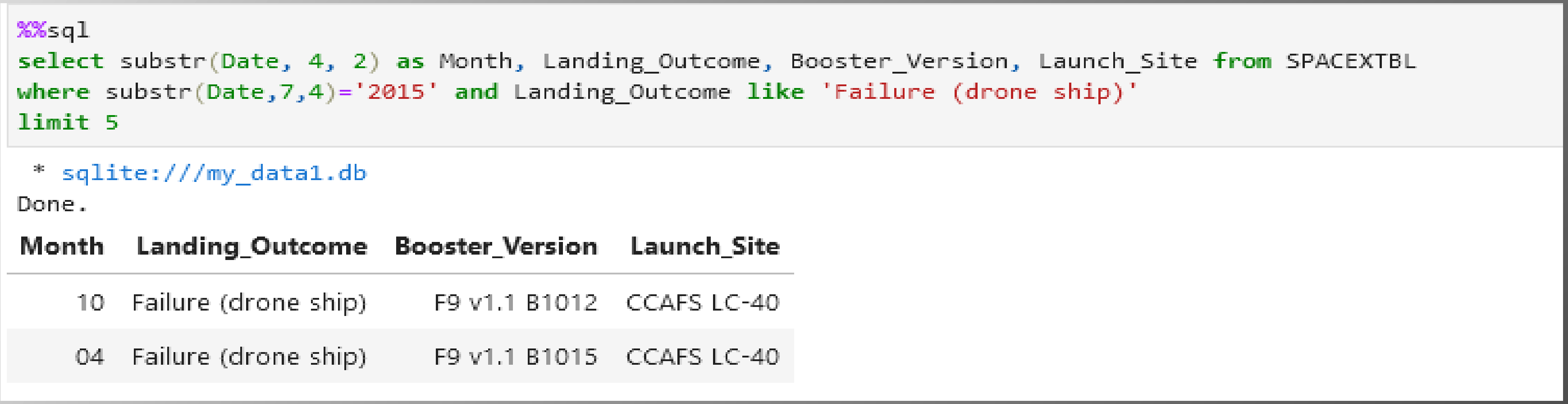


```
%%sql
select distinct(Booster_Version) from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

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

2015 Launch Records



```
%%sql
select substr(Date, 4, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL
where substr(Date,7,4)='2015' and Landing_Outcome like 'Failure (drone ship)'
limit 5
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

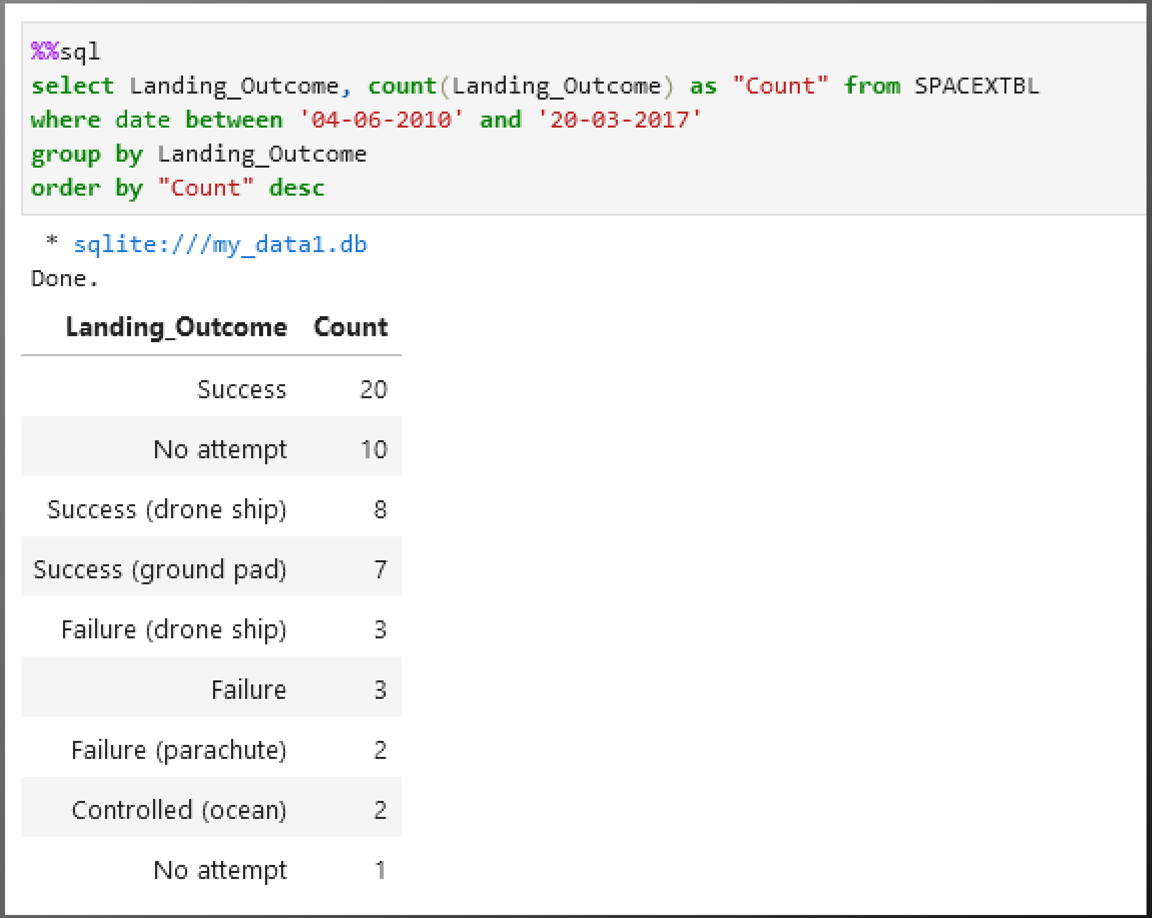
Explanation:

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

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

Explanation:

- 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



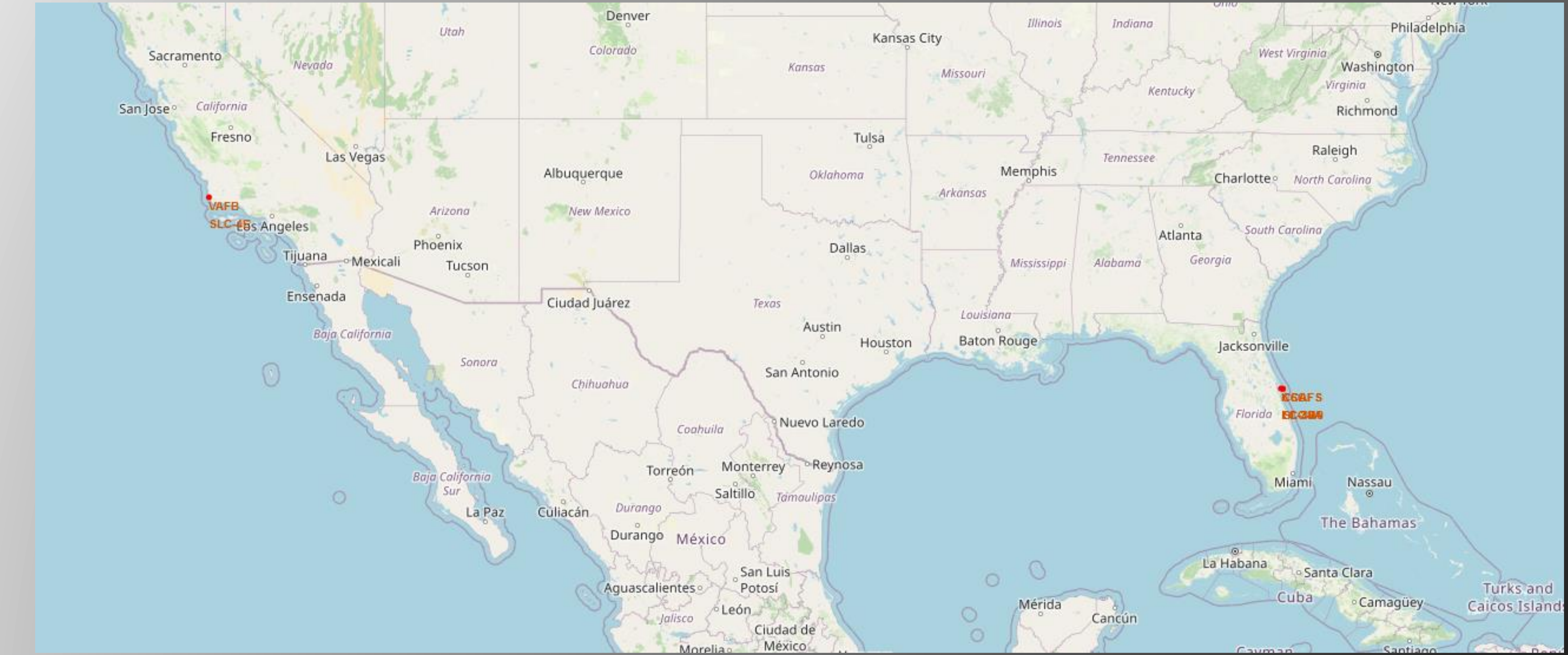
```
%%sql
select Landing_Outcome, count(Landing_Outcome) as "Count" from SPACEXTBL
where date between '04-06-2010' and '20-03-2017'
group by Landing_Outcome
order by "Count" desc
```

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

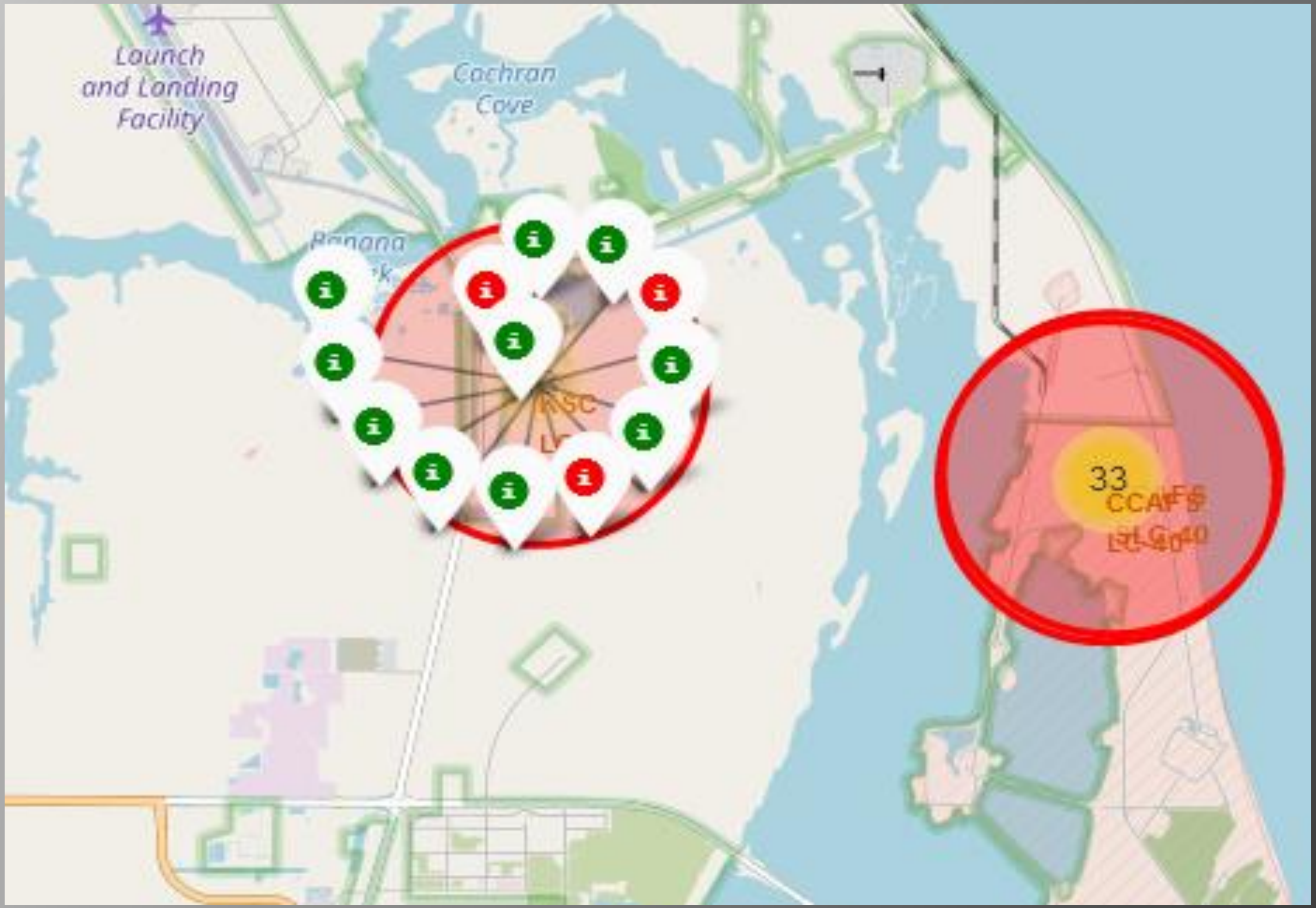
Landing_Outcome	Count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Interactive map with Folium

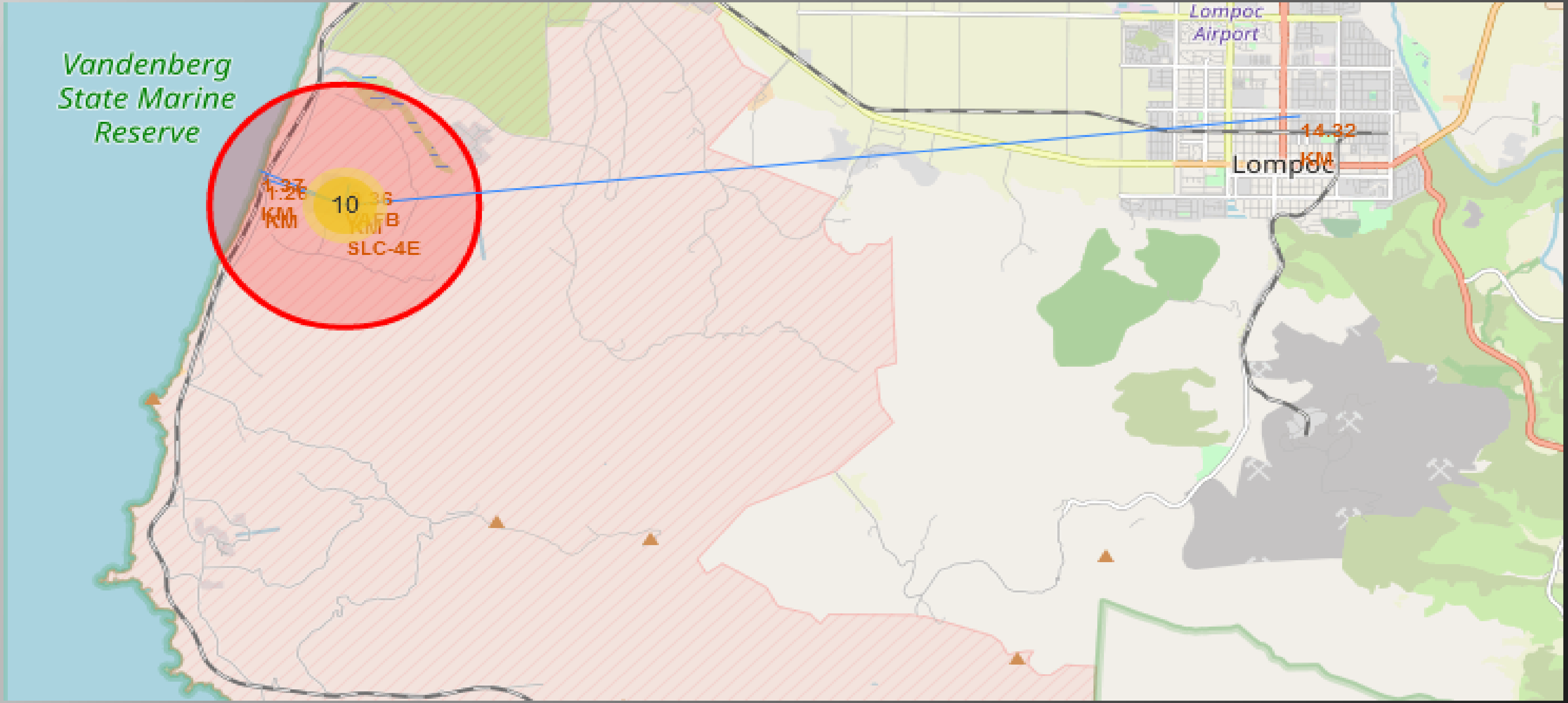
All launch sites on a map



The success/failed launches for each site on the map

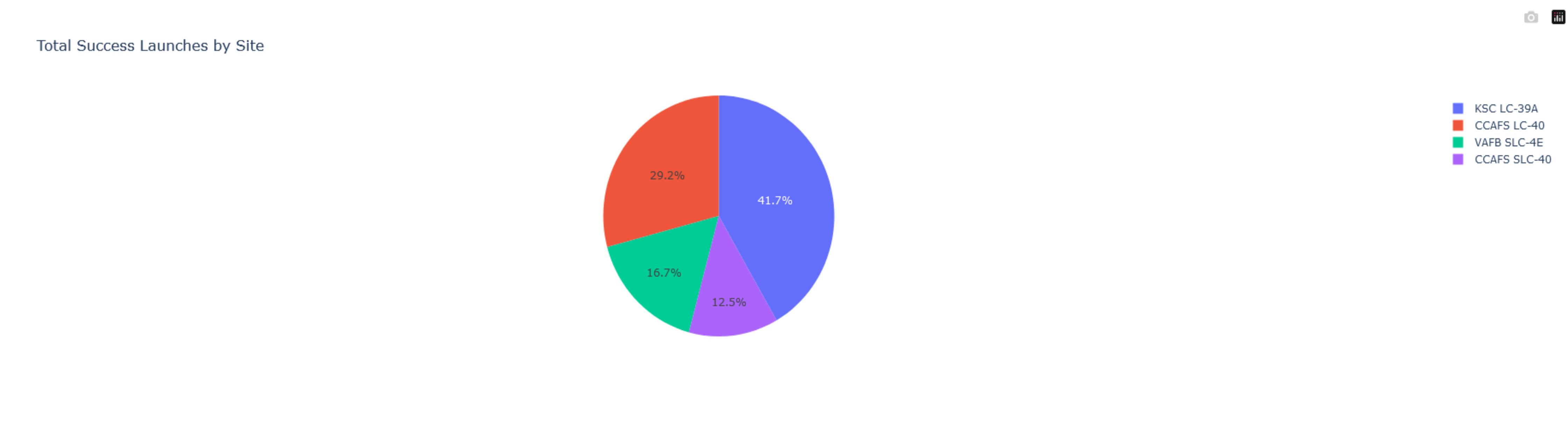


The distances between a launch site to its proximities



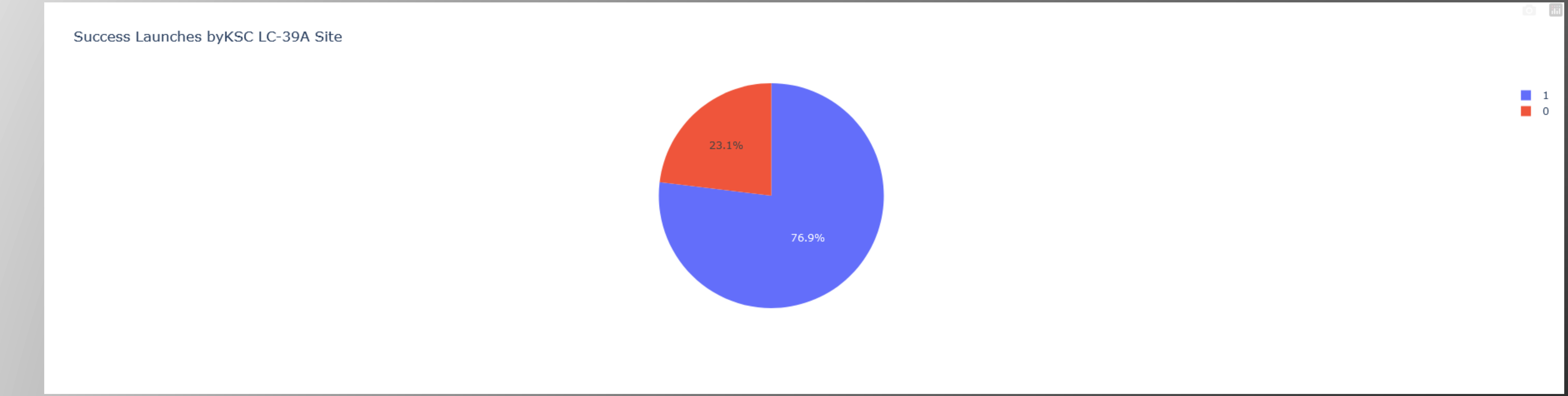
Dashboards

Launch success count for all sites



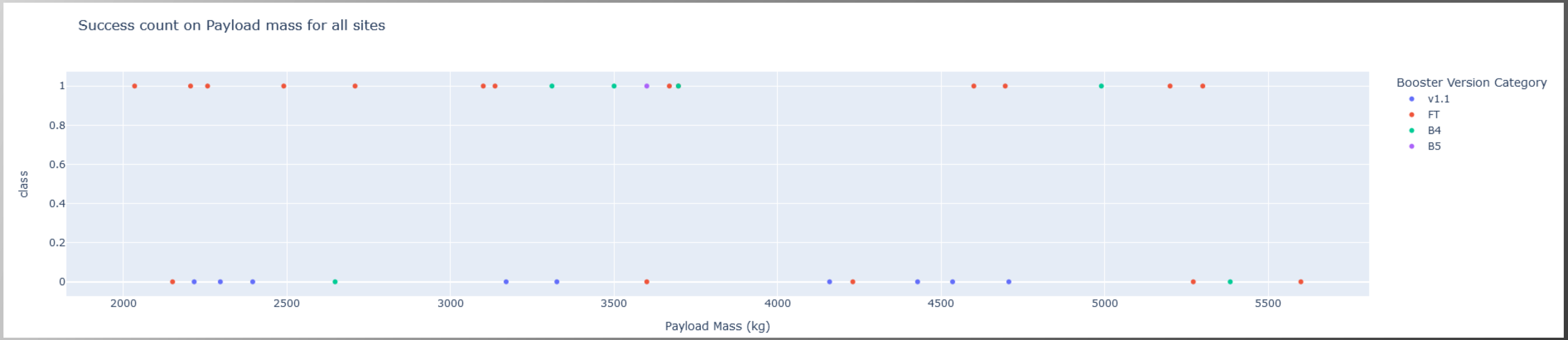
The pie chart shows that from all the sites, KSC LC-39A has the most successful launches.

The launch site with highest launch success ratio



KSC LC-39A has 76,9% success rate.

Payload vs. Launch Outcome scatter plot for all sites

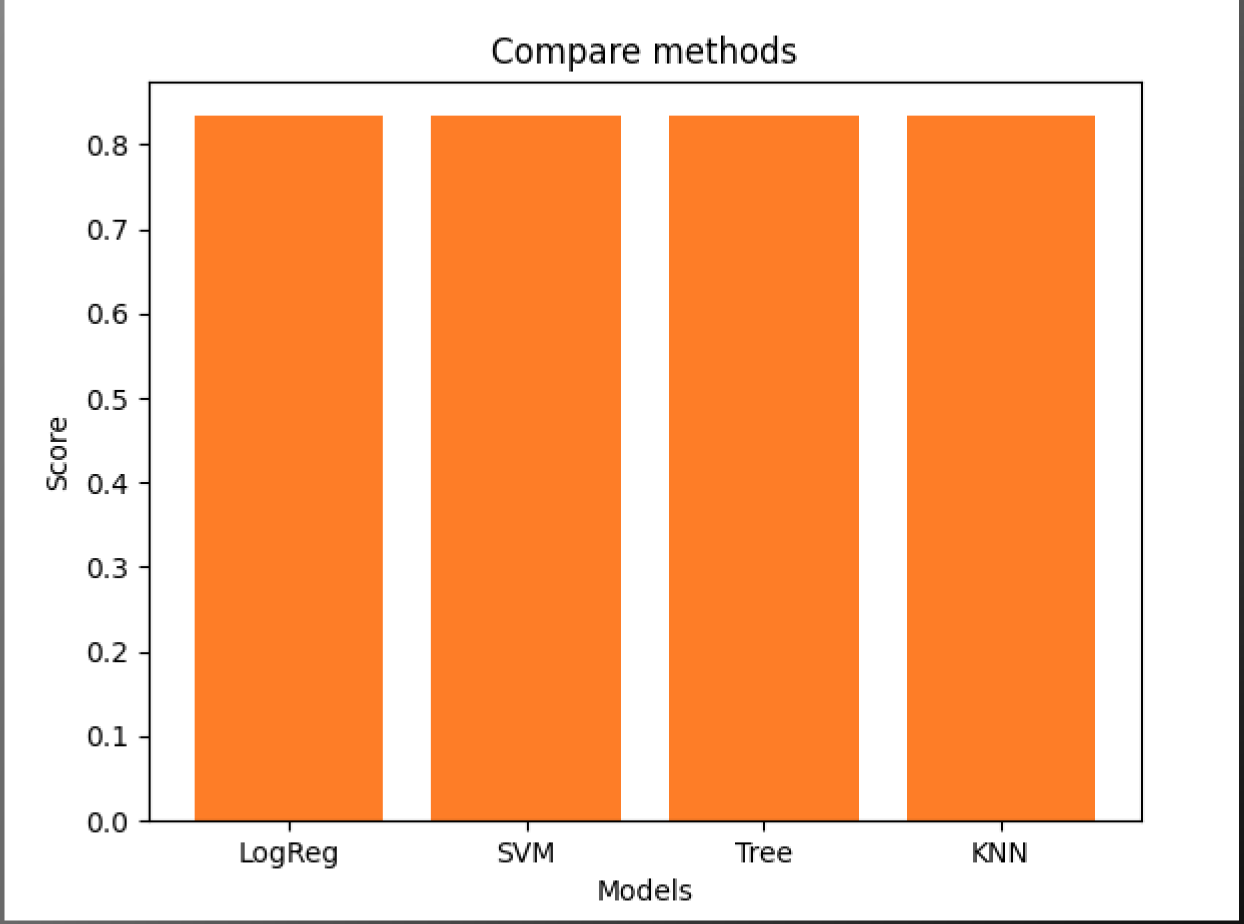


The best results are located between 2000 and 5000 payload mass(kg).

Machine Learning Algorithms

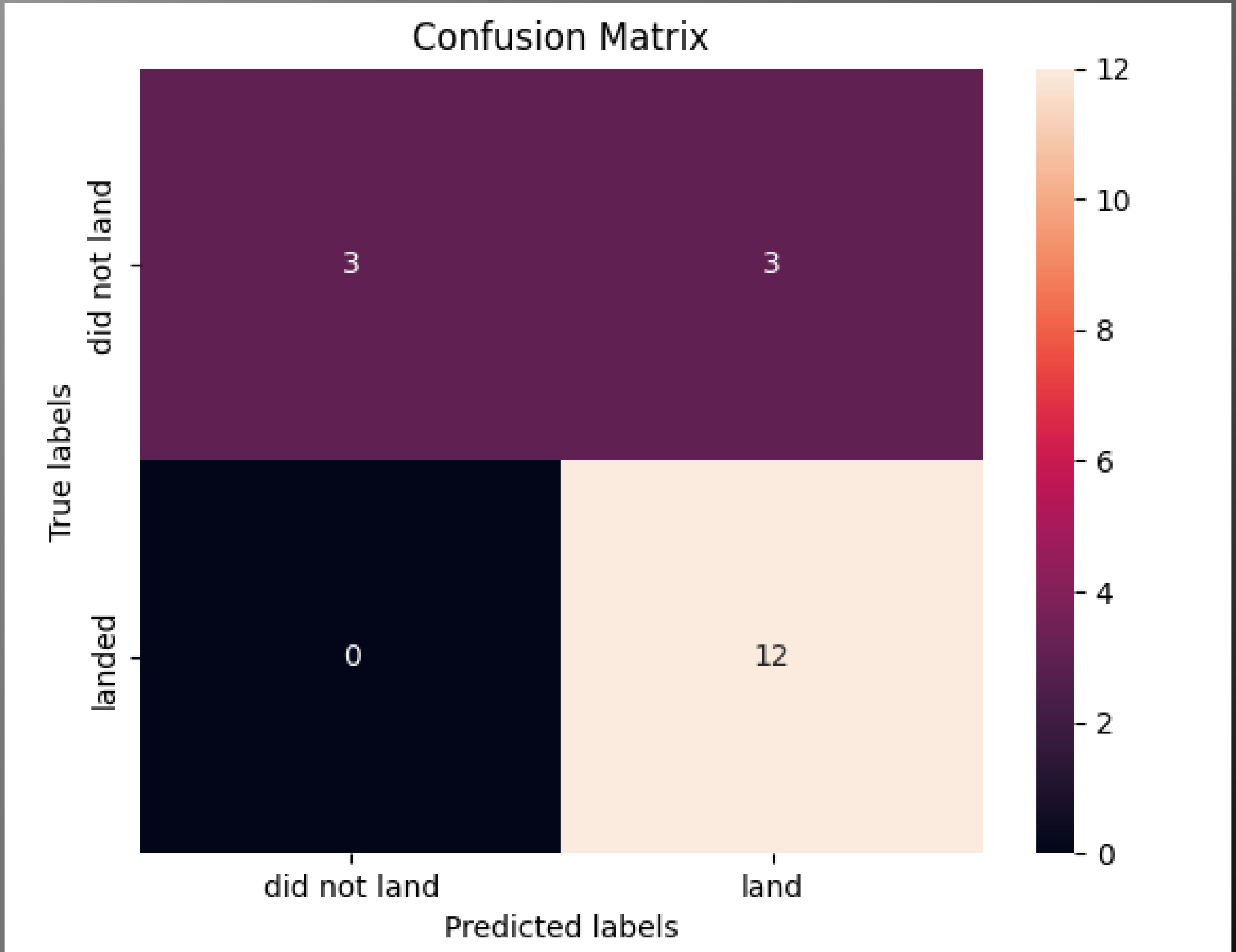
Classification Accuracy

The standard evaluation method(‘model’.score()) of GridSearchCV is used. It showed that all 4 classification models can be used on this type of data.



Confusion Matrix

- We can see the distinction between the different classes on the confusion matrix. We can see that the main problem is false positives.



Conclusions

- The ES-L1, GEO, HEO,SSO has 100% success rate.
- The success rate since 2013 kept increasing till 2020.
- CCAFS SLC 40 is good choice for extra heavy rockets(>10000)
- KSC LC 39A has 100% success rate for payload mass under 5000
- Most of the launch sites are close to the equator and all are very close to the coast.
- All 4 predictive models have the same results on the test set. So we can use any algorithm for new inputs.

Appendix

- https://api.spacexdata.com/v4/rockets/ - SpaceX api where stored data about launches.
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches?utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01 - SpaceX api where stored data about only Falcon 9 launches.