IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jason Wan
2022-06-29

https://github.com/JaysonWan/DataScience_Capstone

Outline

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

Executive Summary

- Data was collected from the SpaceX API and SpaceX Wikipedia page.
- Explored data using SQL, visualization, folium maps, and dashboards.
- Gathered all variables to binary using one hot encoding.
- Standardized data and used GridSearchCV to find best parameters for machine learning models
- Visualized accuracy score of all models.
- Learning models of Logistic Regression, Support Vector Machine, Decision Tree Classifier and K Nearest Neighbors.
- All ML models produced similar results with the accuracy of 83%.
- More data is needed for an accurate model

Introduction

- SpaceX (Falcon 9) has the best pricing out of commercial space crafts
- Our goal is to recover the rocket after it has been launched.
- SpaceY wants to compete with SpaceX
- The challenge is to determine the price of each launch and to gather public information to create dashboards for the team
- Use the machine learning model to predict the successful recovery of stage 1.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tune models using GridSearchCV

Data Collection

- The Data was collected using API requests from Space X public API and web scraping data from the table in Space X’s Wikipedia page.
- The next slide will show the sequence on the processing data from the SpaceX public API and the one after will show sequence of processing the data from web scraping.

Data Collection – SpaceX API

1. Request (Space X APIs)
2. .JSON file + Lists(Launch site, Booster Version, Payload Data
3. Json_normalize to DataFrame data from JSON
4. Dictionary relevant data
5. Cast dictionary to a DataFrame
6. Filter data to only include Falcon 9 launches
7. Replace missing PayloadMass values with mean

https://github.com/JaysonWan/DataScience_Capstone/blob/main/Spacex%20data%20collection%20api.ipynb

Data Collection - Scraping

1. Request Wikipedia HTML
2. Beautiful Soup HTML5lib Parser
3. Find Launch info html table
4. Cast dictionary to DataFrame
5. Iterate through table cells to extract data to dictionary
6. Create dictionary

https://github.com/JaysonWan/DataScience_Capstone/blob/main/Spacex%20Webscraping.ipynb

Data Wrangling

- Created a training label with landing outcomes where successful = 1 and failure = 0. The outcome column will be either ‘Mission Outcome’ or ‘ Landing Location’ The column ‘class’ with value of 1 will be Mission Outcome if True and 0 otherwise.
- True ASDS, True RTLs, & True Ocean - set to 1
- None none, False ASDS, False RTLs, False Ocean - set to 0
- https://github.com/JaysonWan/DataScience_Capstone/blob/main/Data%20wrangling.ipynb

EDA with Data Visualization

- The plots used in data visualization were scatter plots, line charts and bar plots. These plots were used to compare the relationships between variables. These plots would later be used to determine if they are capable of being used for training the machine learning model.
- https://github.com/JaysonWan/DataScience_Capstone/blob/main/EDA%20with%20python.ipynb

EDA with SQL

- Loaded data set into IBM DB2 Database
- Queries were made to have a better understanding of the dataset.
- Made Queries with SQL python integration
- Query information included launch site names, mission outcomes, landing outcomes, payload sizes.

- https://github.com/JaysonWan/DataScience_Capstone/blob/main/EDA%20with%20sql.ipynb

Build an Interactive Map with Folium

- Locations analysis with Folium
- Made successful and unsuccessful landings using Folium maps through key locations of Railway, Highway, Coast and City.
- It allowed us to understand where the launch sites may be located where they are. Successfully visualized landings relative with the location.
- https://github.com/JaysonWan/DataScience_Capstone/blob/main/Launch%20site%20location%20with%20folium.ipynb

Build a Dashboard with Plotly Dash

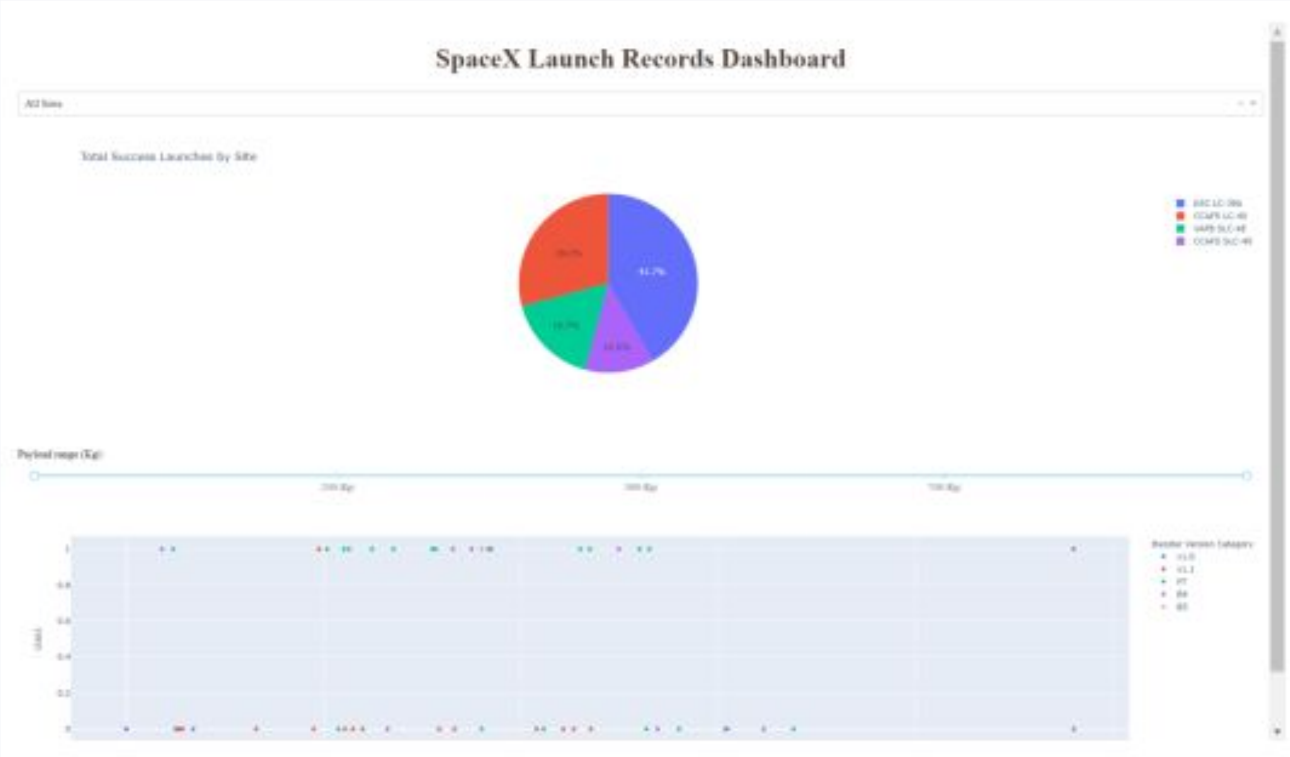
- There were various bar charts, pie charts, and scatter plots used to makes reports in failed launches and landing outcomes
- These interactions allowed users to access data from various reports from 2005
- https://github.com/JaysonWan/DataScience_Capstone/blob/main/Dashboard%20with%20Plotly%20Dash

Predictive Analysis (Classification)

- Split label column/Class’ from dataset
- Fit and transform the features using standard scaler
- GridSearchCV to find optimal parameters
- use GridSearchCV on LogReg, SVM, Decision Tree and KNN models
- Barplot to compare the scores of models.

https://github.com/JaysonWan/DataScience_Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

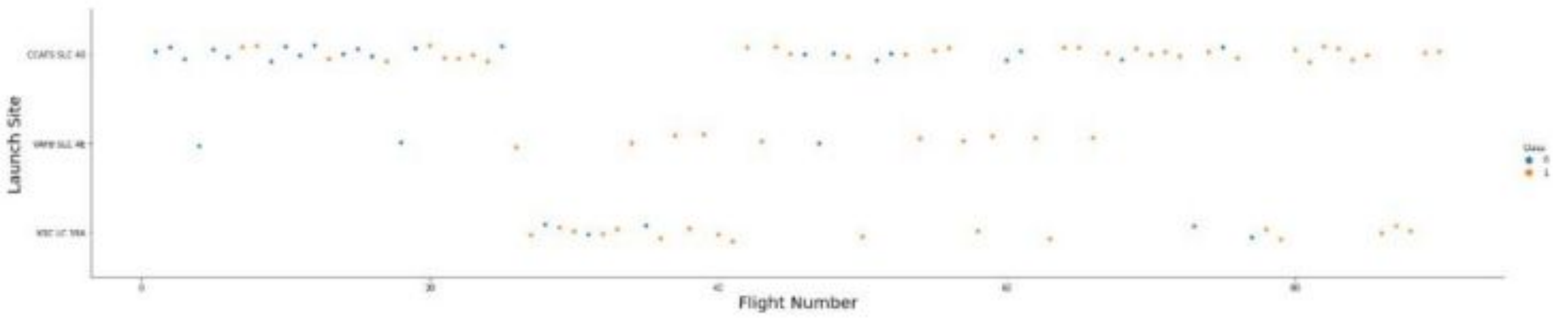


This is a preview of Plotly Dashboard. This is the result of EDA with visualization, EDA with SQL. The result wielded an accuracy of 83%.

Section 2

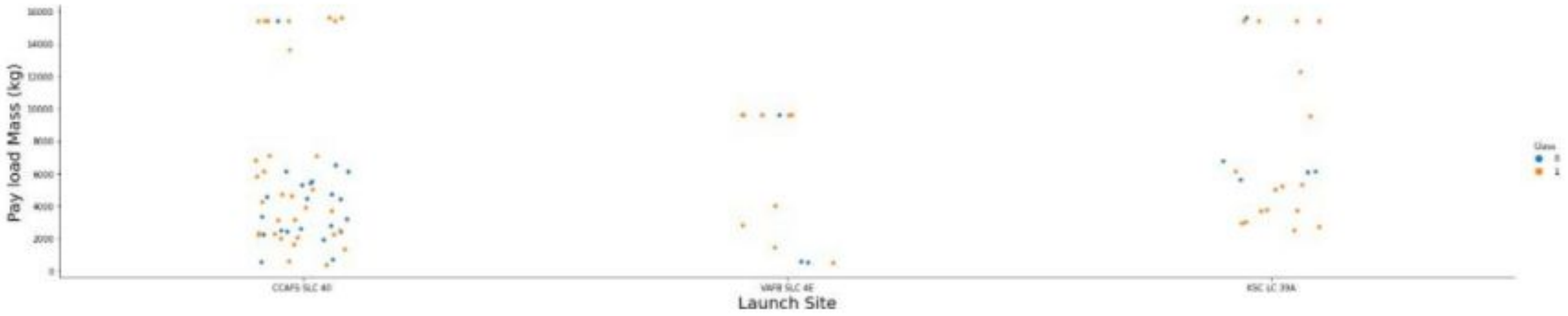
Insights drawn from EDA

Flight Number vs. Launch Site



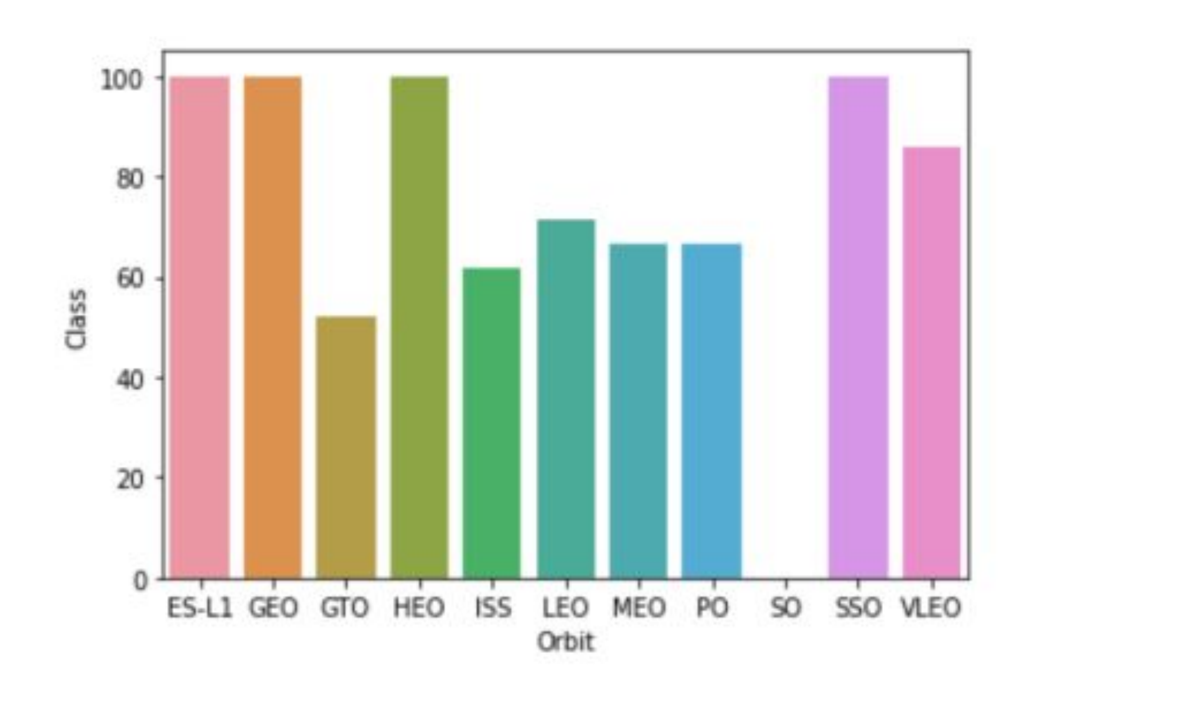
The graph indicates an increase in success rate over time. The breakthrough of flight 20 is likely the cause of the significant increase in success rate. The orange colour of the graph indicates successful launch and blue indicates an unsuccessful launch

Payload vs. Launch Site



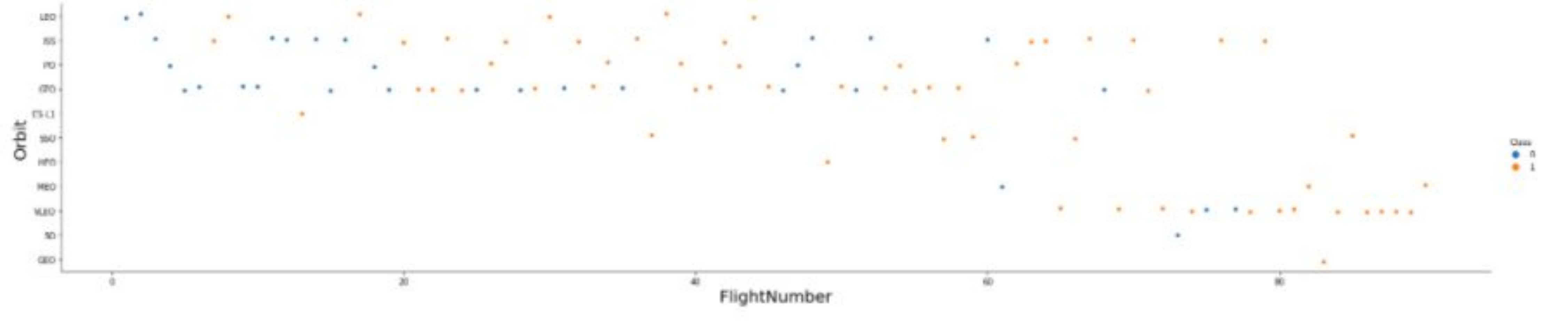
Same as the previous graph orange represents successful launch and blue representing unsuccessful launch. Different launch sites use different payload masses. The payload mass falls between 0-7000 kg.

Success Rate vs. Orbit Type



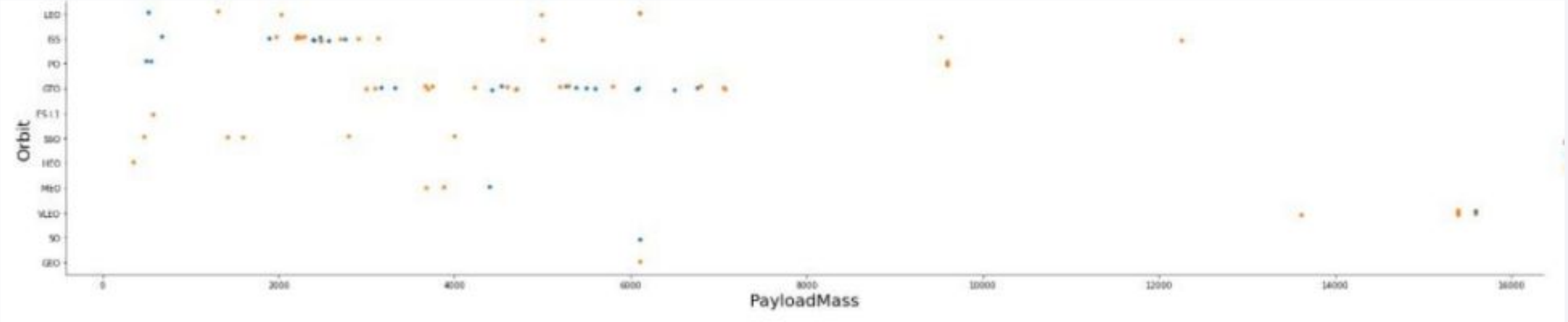
ES-L1, SSO, GEO, HEO have 100% success rate
VLEO has ok success rate
SO has 0% success rate
GTO and etc have around 50% success rate

Flight Number vs. Orbit Type



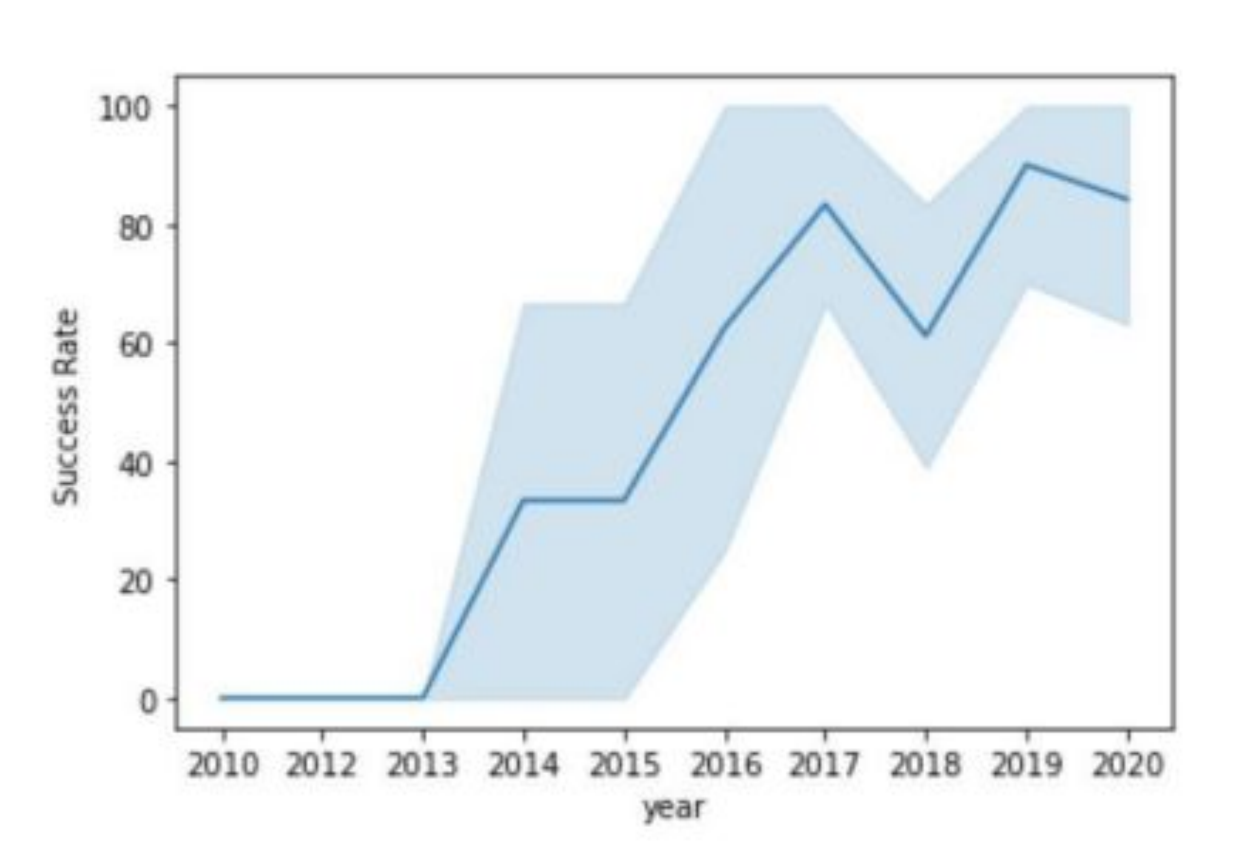
Launch orbit changed over flight number. Launch orbit also correlates with its preference. SpaceX performs better in low orbits or Sun synchronous orbits

Payload vs. Orbit Type



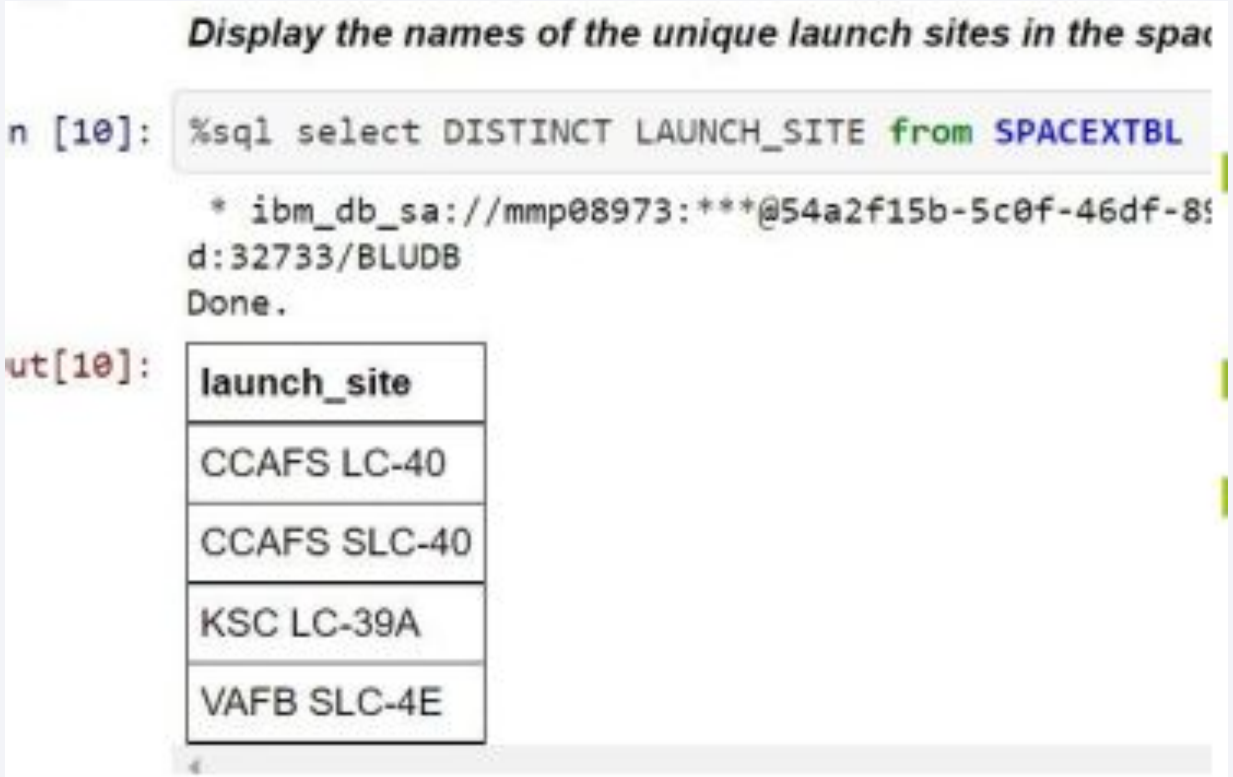
LEO and SSO have relatively low payload mass. The most successful orbit of VLEO has a far better payload mass.

Launch Success Yearly Trend



successful launches over the years 2010-2020. The blue fade indicates a confidence interval and shows the average success rate is around 80% in recent years

All Launch Site Names



Display the names of the unique launch sites in the spa

```
n [10]: %sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```

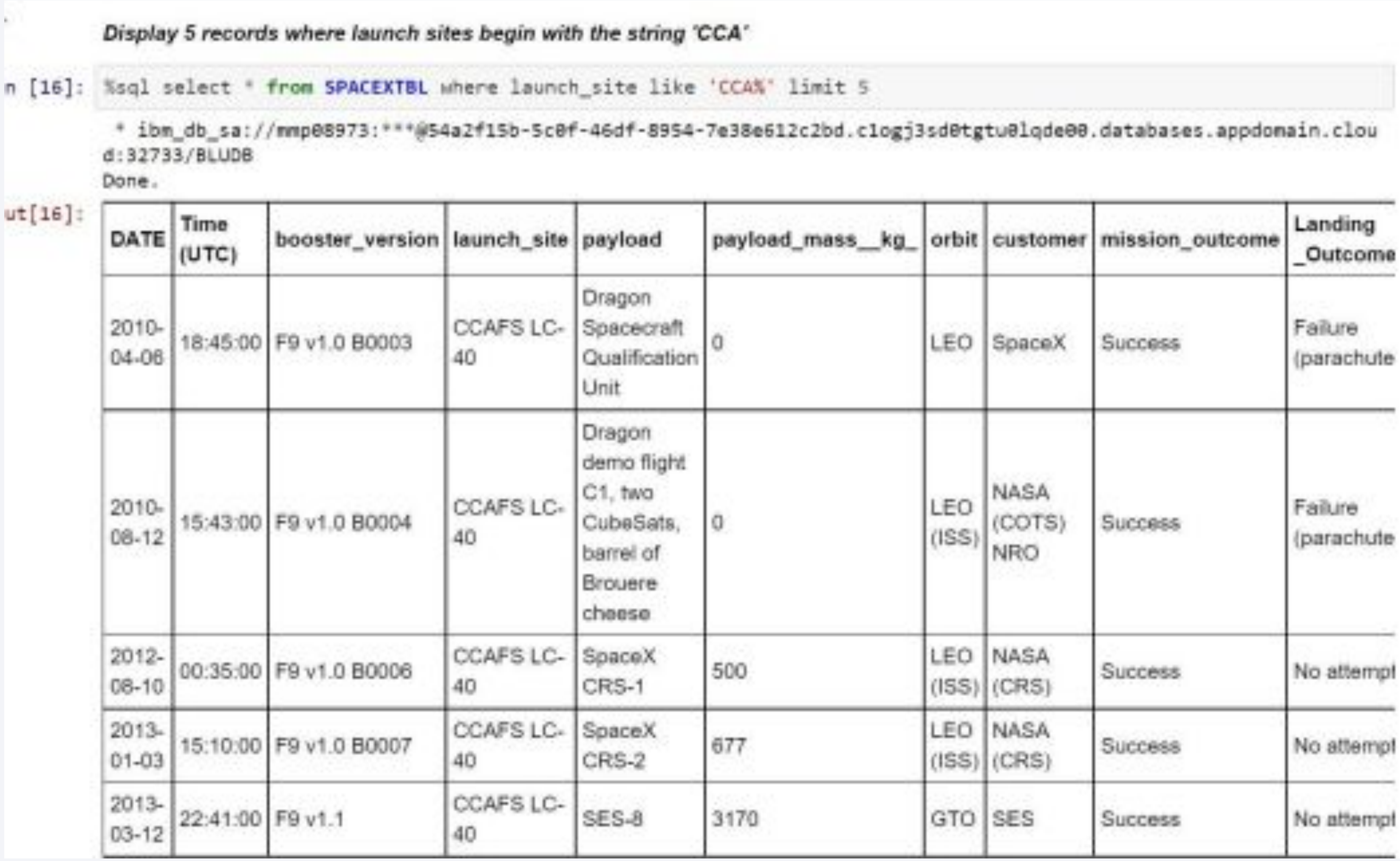
```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8
d:32733/BLUDB
Done.
```

ut[10]:

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

- Query unique launch site names from database
- CCAFS LC-40 and CCAFS SLC-40 are all represent the same
- CCAFS LC-40 was the previous name
- Launch site with data entry errors.

Launch Site Names Begin with 'CCA'



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

```
n [16]: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

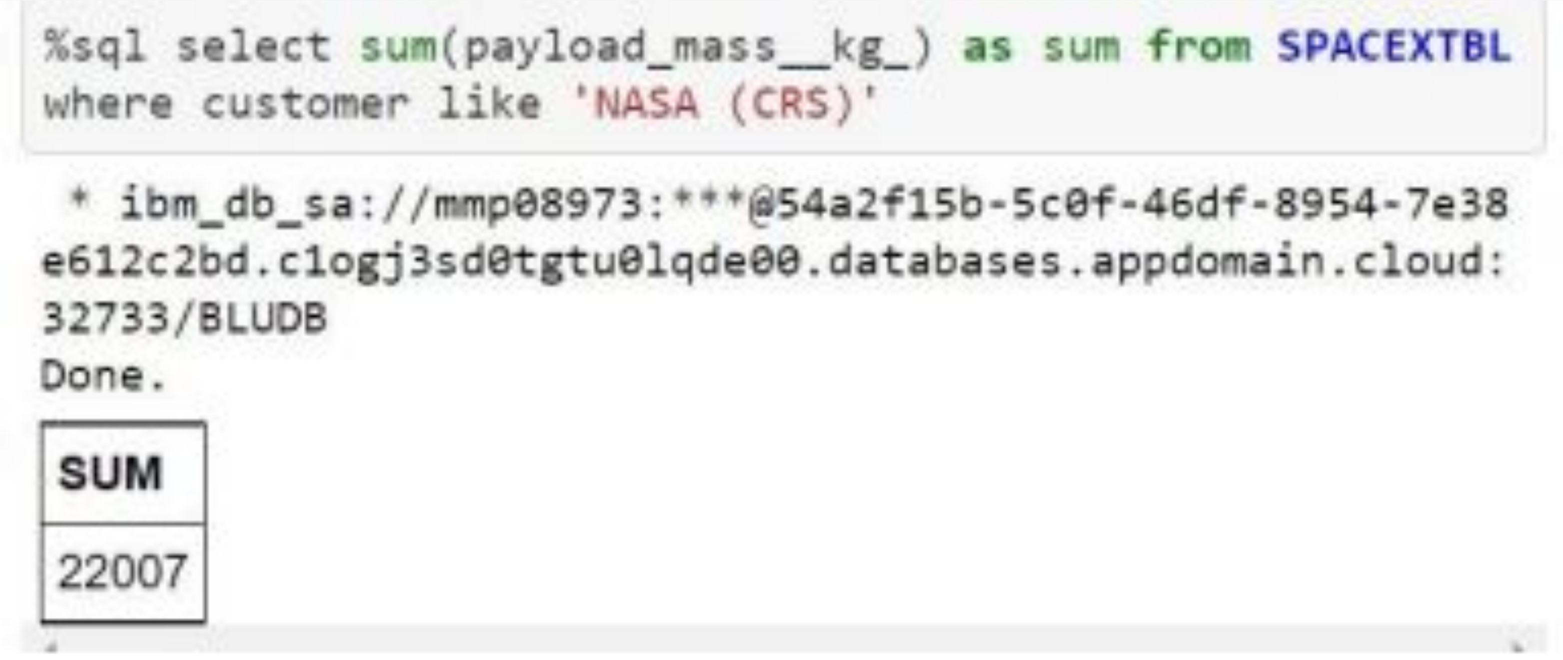
```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.clou
d:32733/BLUDB
Done.
```

ut[16]:

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

First 5 entries with launch site name beginning with CCA

Total Payload Mass



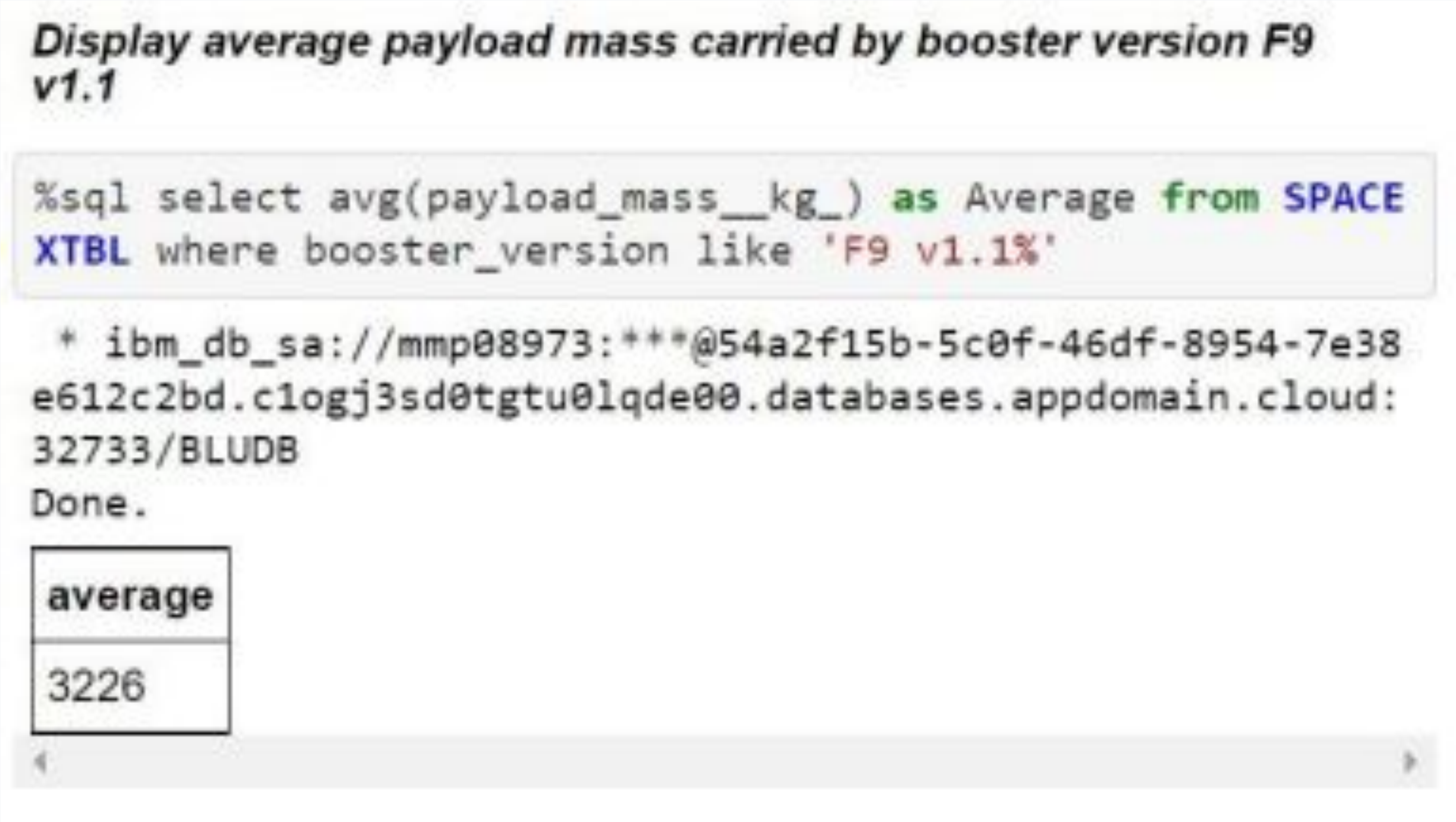
```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL
where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

SUM
22007

The query sums the total payload mass in kg. Nasa is the customer. CRS stands for commercial resupply services. These payloads were sent to the international space station

Average Payload Mass by F9 v1.1



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

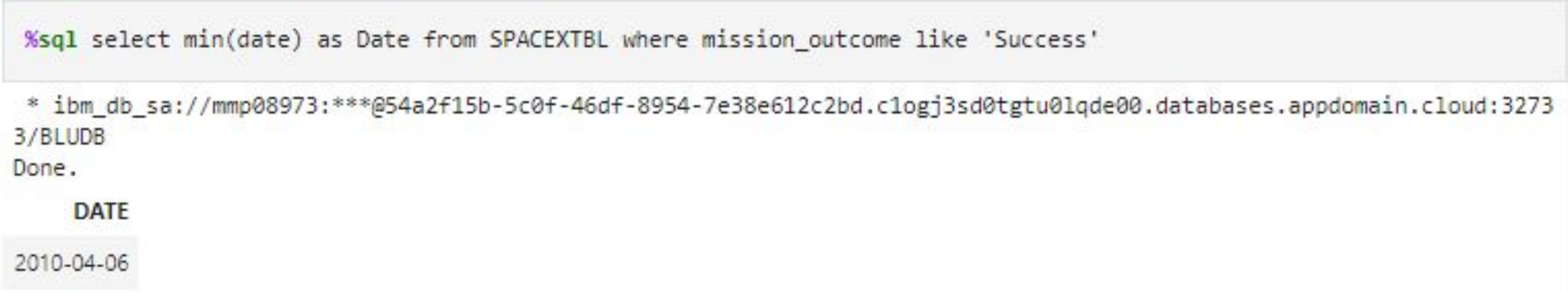
```
%sql select avg(payload_mass__kg_) as Average from SPACE
XTBL where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

average
3226

Average payload mass used booster version F9 v1.1

First Successful Ground Landing Date



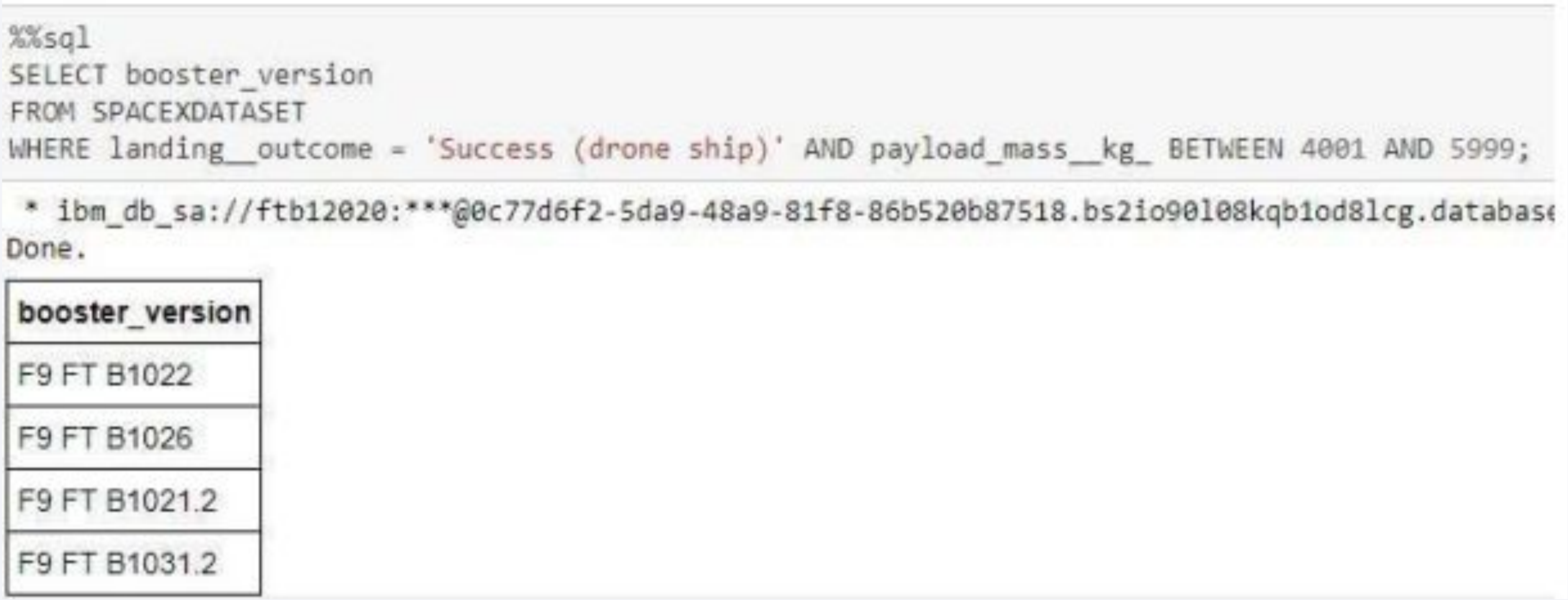
```
%sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/BLUDB
Done.
```

DATE
2010-04-06

2010-04-06 was the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000



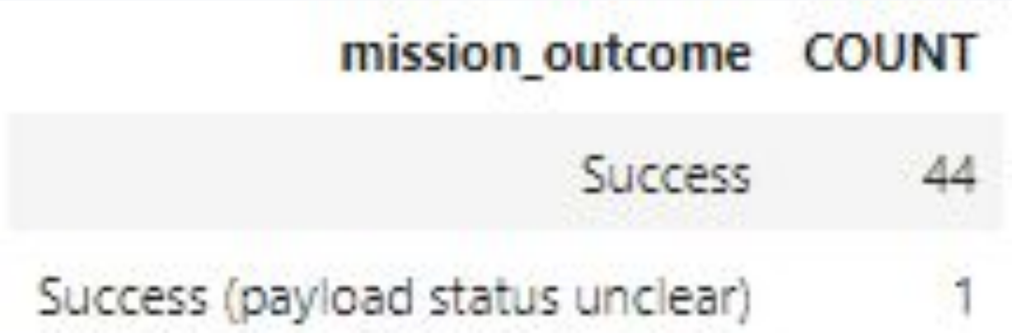
```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

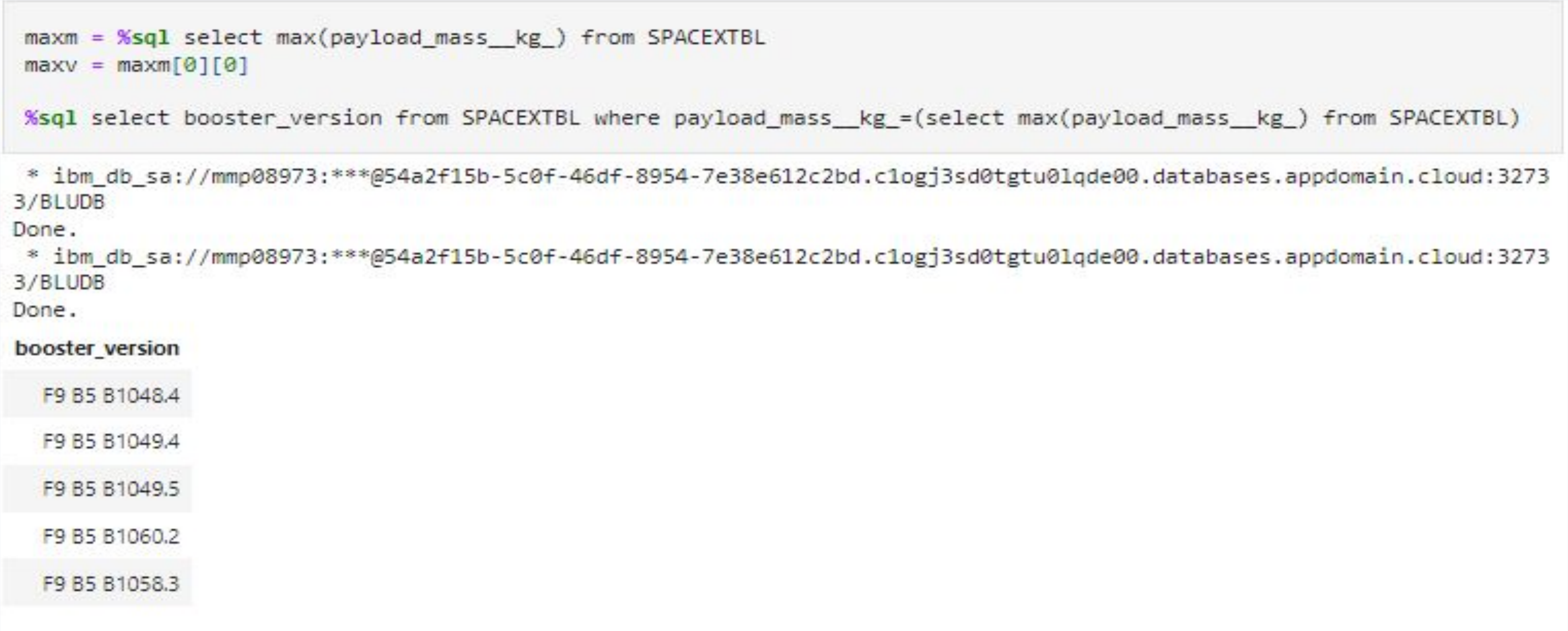
Total Number of Successful and Failure Mission Outcomes



mission_outcome	COUNT
Success	44
Success (payload status unclear)	1

The number of successful mission outcome is 44 and the number of payload status unclear is 1 so it averages around 98% to achieve the mission.

Boosters Carried Maximum Payload



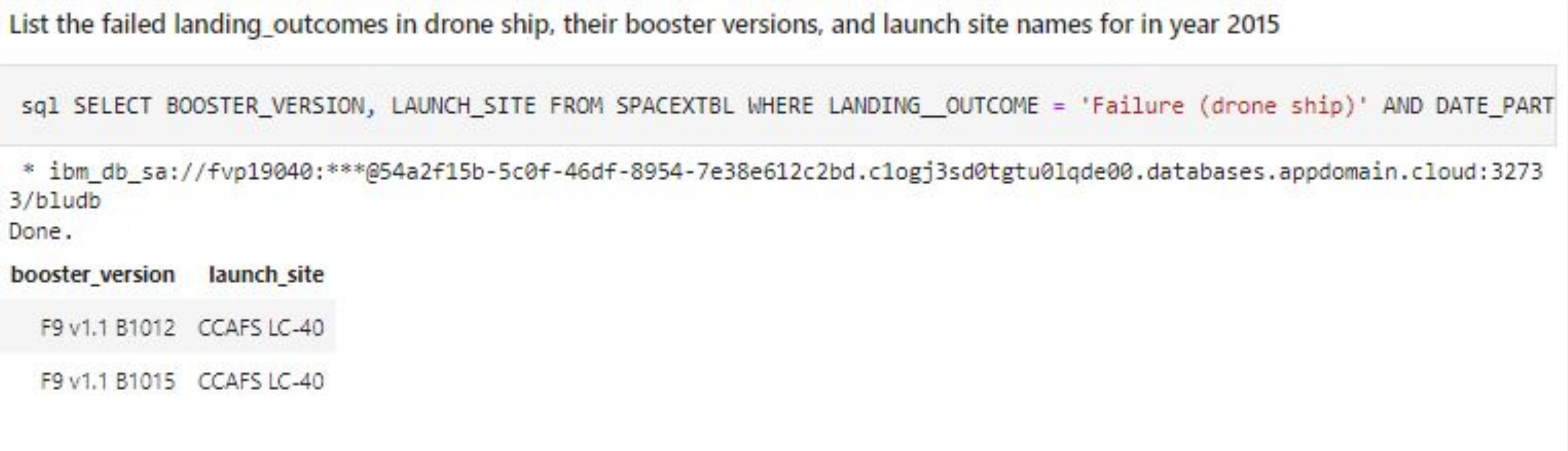
```
maxm = %sql select max(payload_mass__kg_) from SPACEXTBL
maxv = maxm[0][0]

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

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/BLUDB
Done.
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/BLUDB
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

The query returns the highest payload mass of 15600 kg for the booster versions. These are all F9 B5 B10 Variety and the payload correlates with the booster version that is used.

2015 Launch Records



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

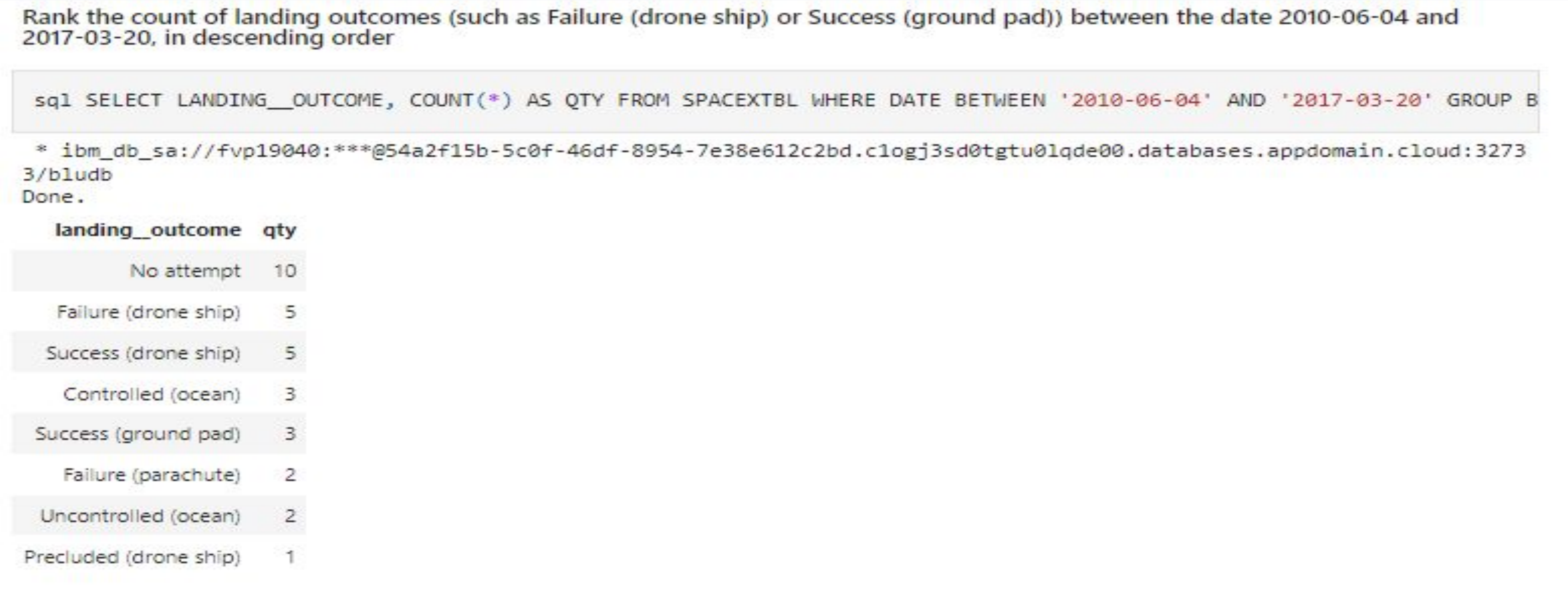
```
sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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



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(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP B
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/bludb
Done.
```

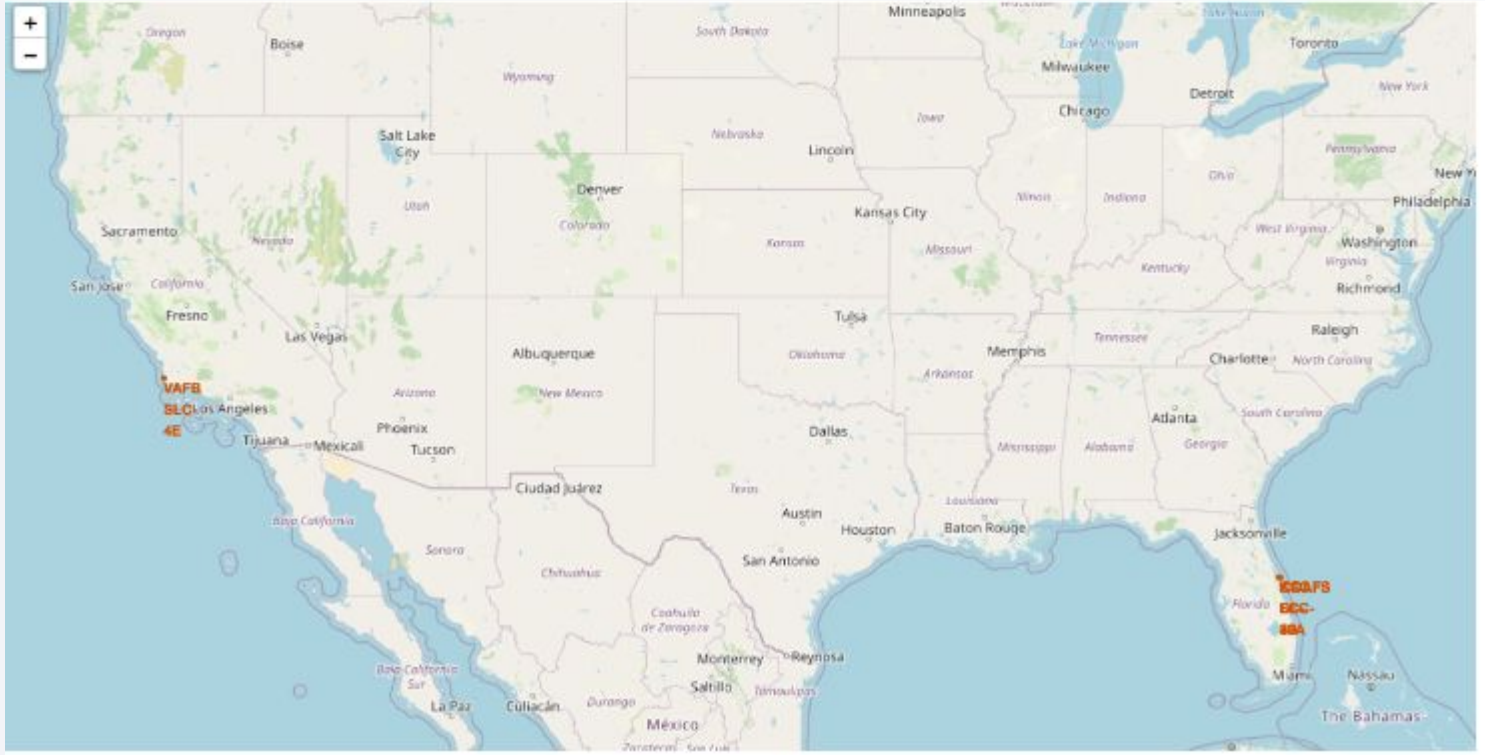
landing__outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

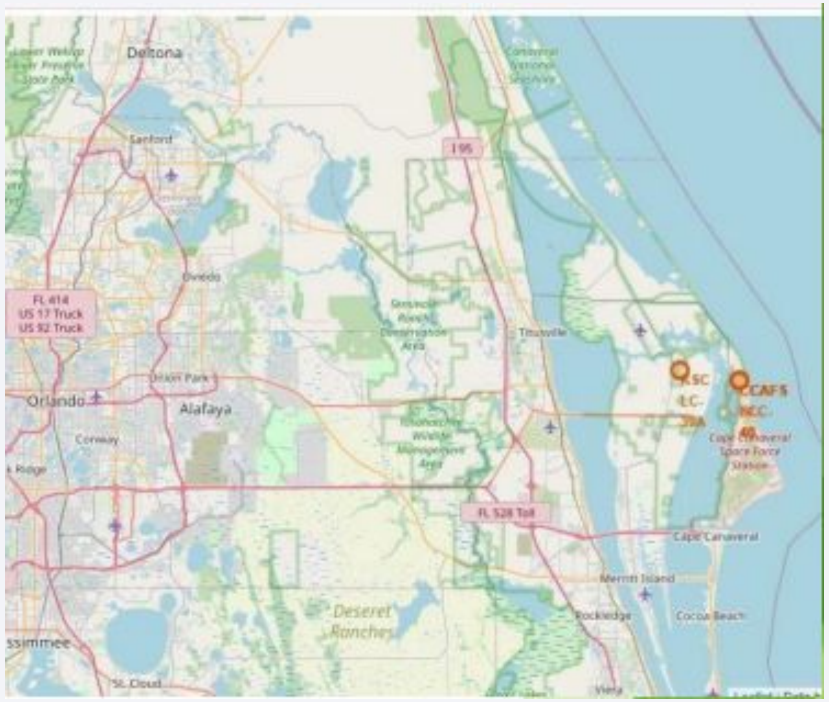
Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

Launch Sites Proximities Analysis

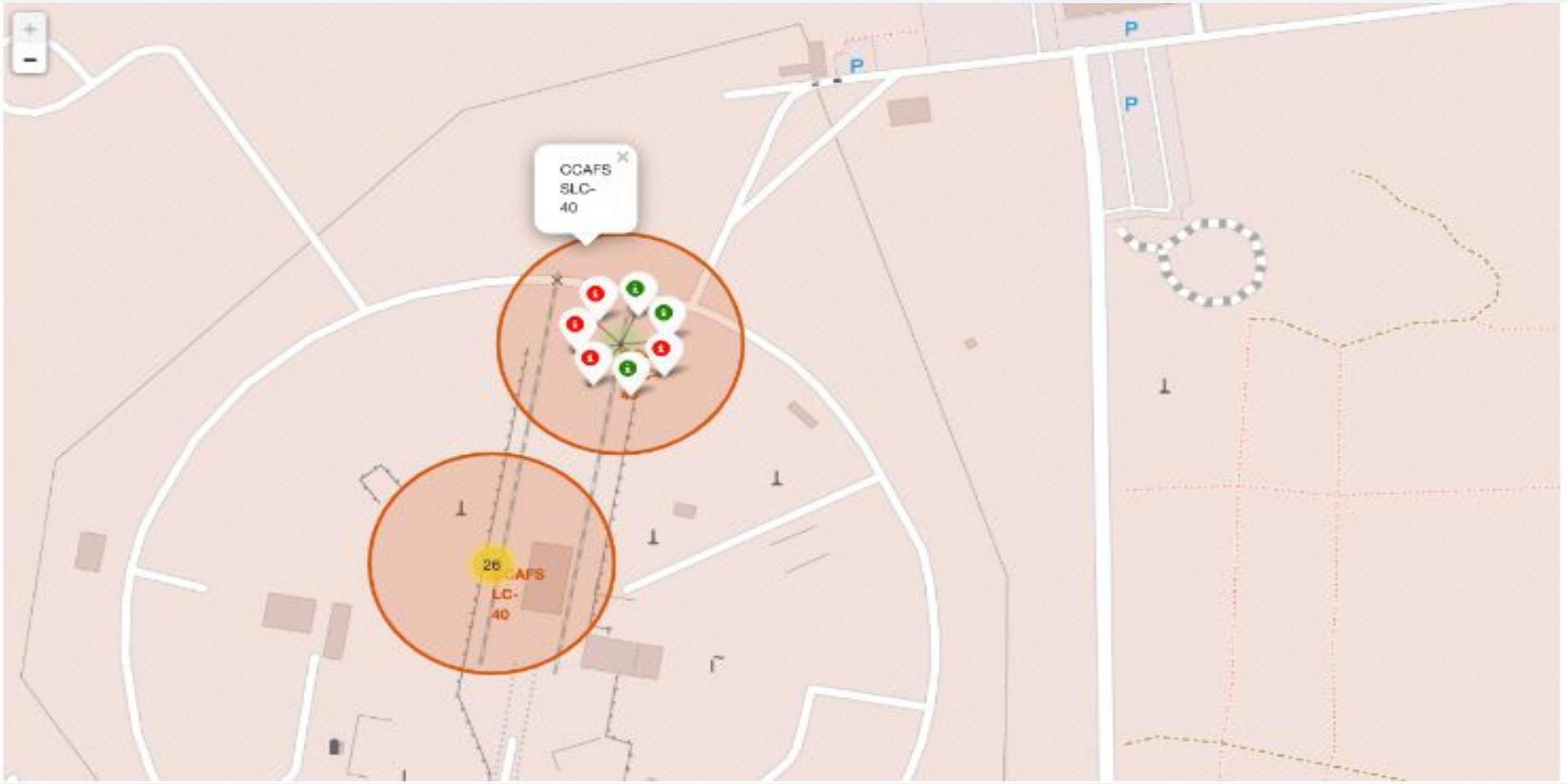
Launch Site Locations





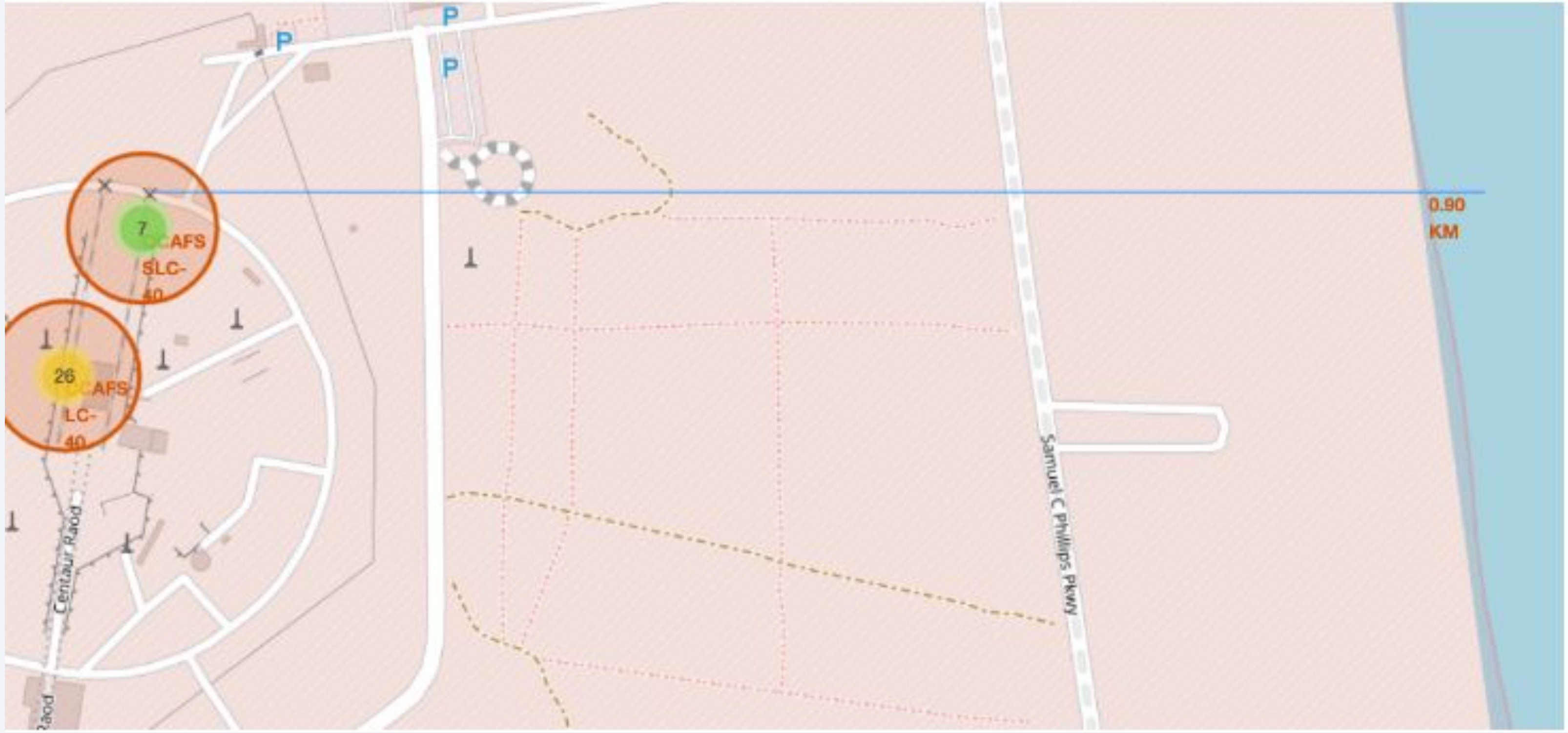
The left image shows the launch sites and the image on the right shows the 2 launch sites in Florida as they are really close to each other

Color-Coded Launch Markers

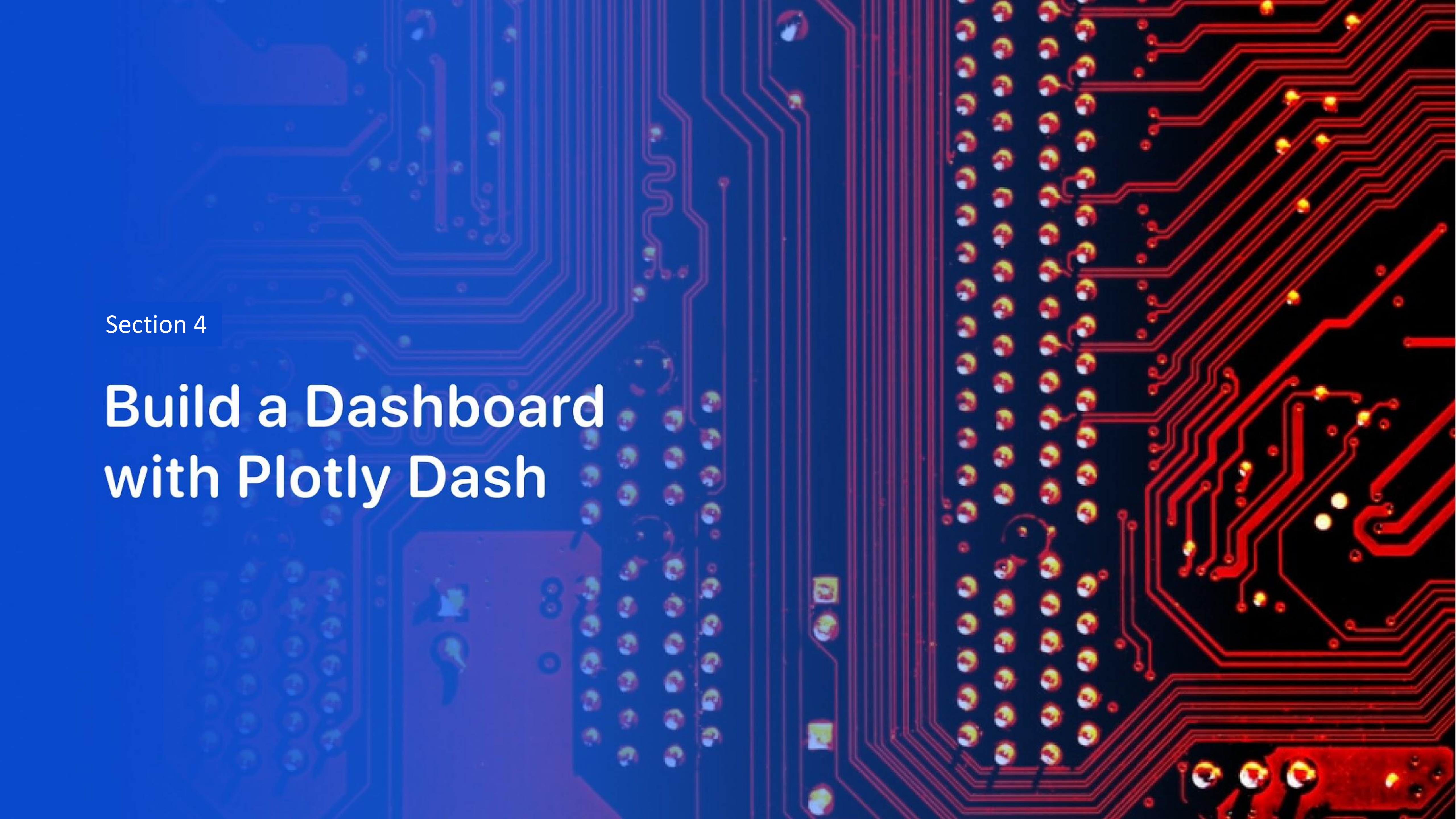


In this example we have a CCAFS SLC-40 completing 3 successful launches and 4 failed launches

Key Location Proximities



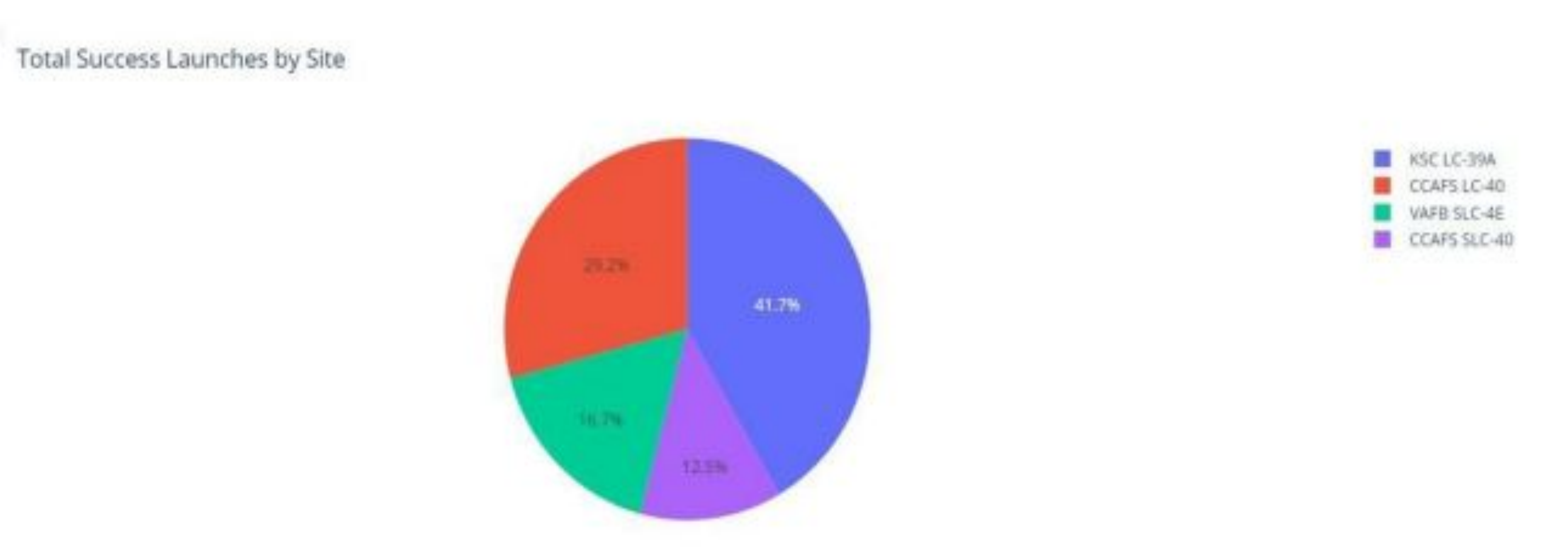
The image above is an example of CCAFS SLC-40 being close to highways, railways and coastlines.



Section 4

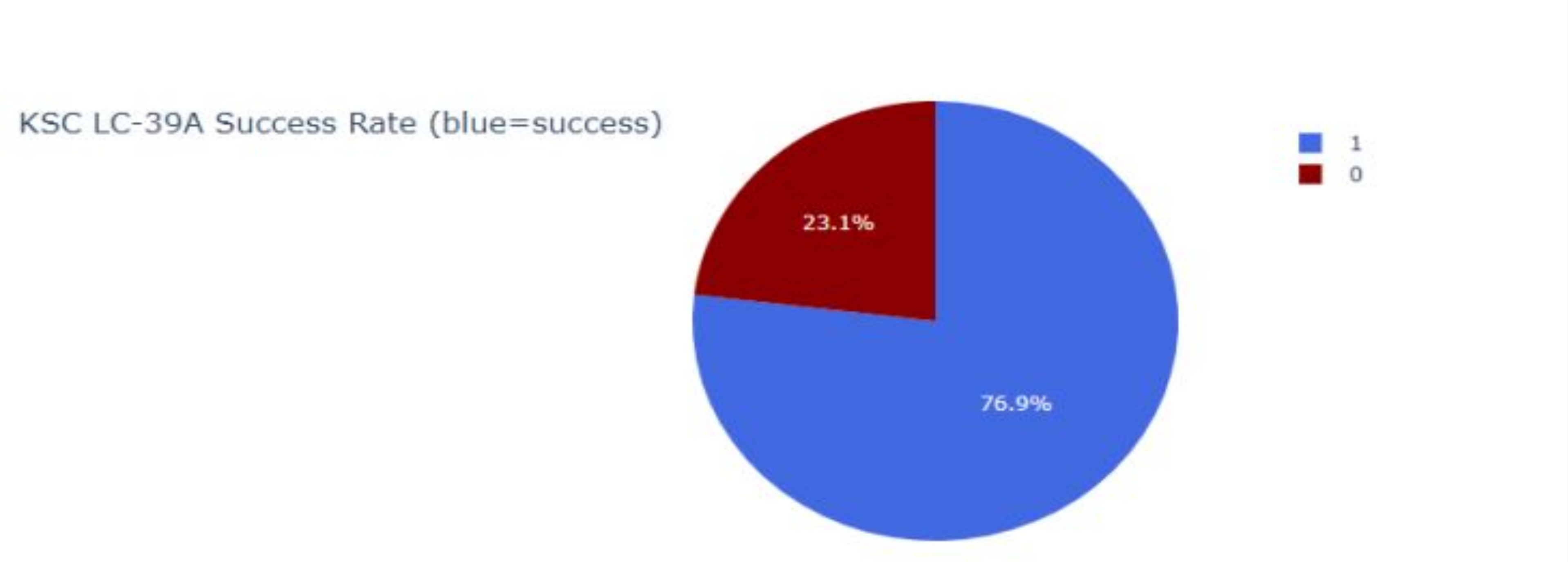
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



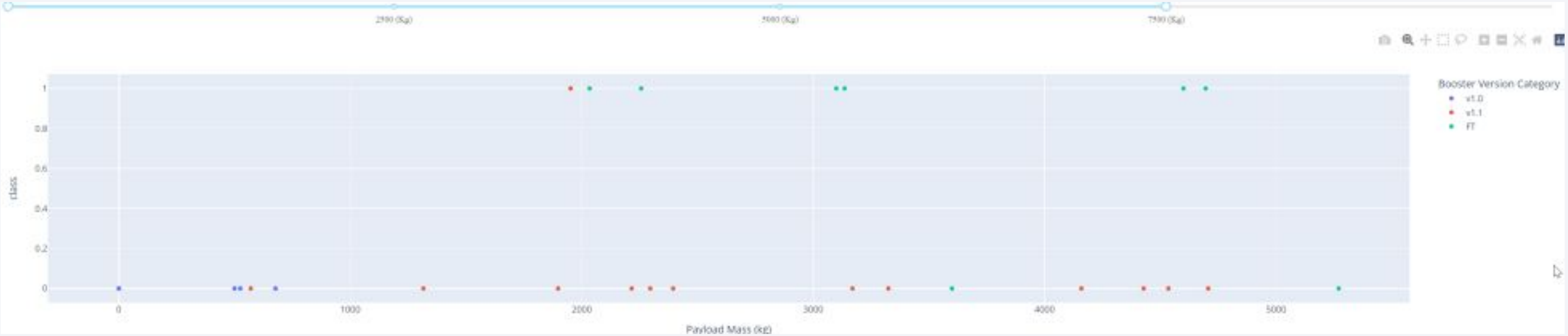
The pie chart represents a distribution of successful landings with different launch sites. VAFB SLC-4E and CCAFS SLC-40 has the lowest successful landings while the KSC LC-39A has the most successful landings. There are many environmental aspects when having a successful launch.

Highest Success Rate Launch Site



Highest success rate is the KSC LC-39A with 10 successful landings and 3 failed landings

Payload Mass vs Success booster Version Category

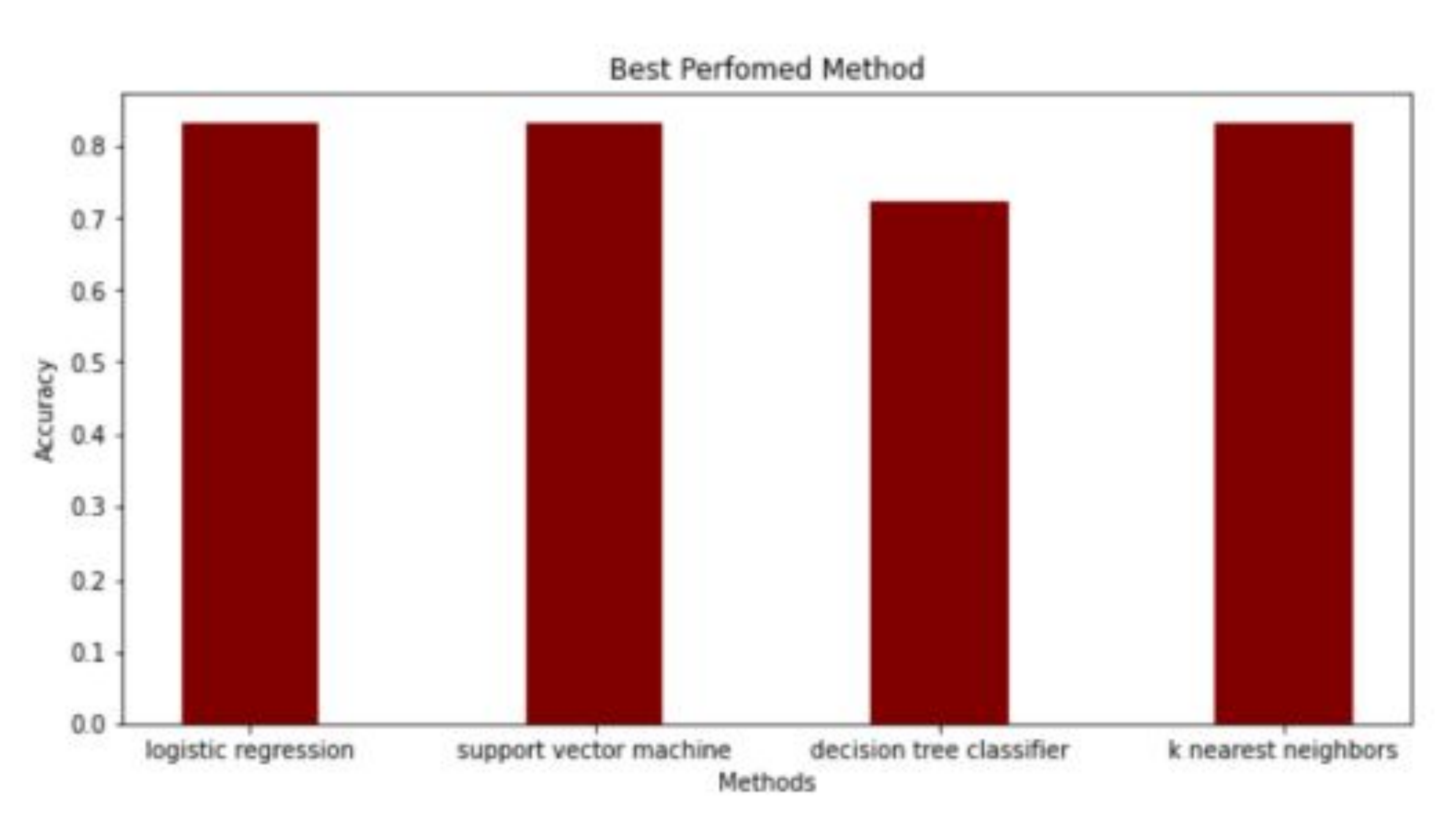


The dashboard has a payload range selector. The set above is a range of 0-10000 instead of the max 15600. If you can see it the 1 indicates a successful landing while 0 indicates unsuccessful. Interestingly from range 0-7500 there are 2 failed landings with payloads of 0 kg.

Section 5

Predictive Analysis (Classification)

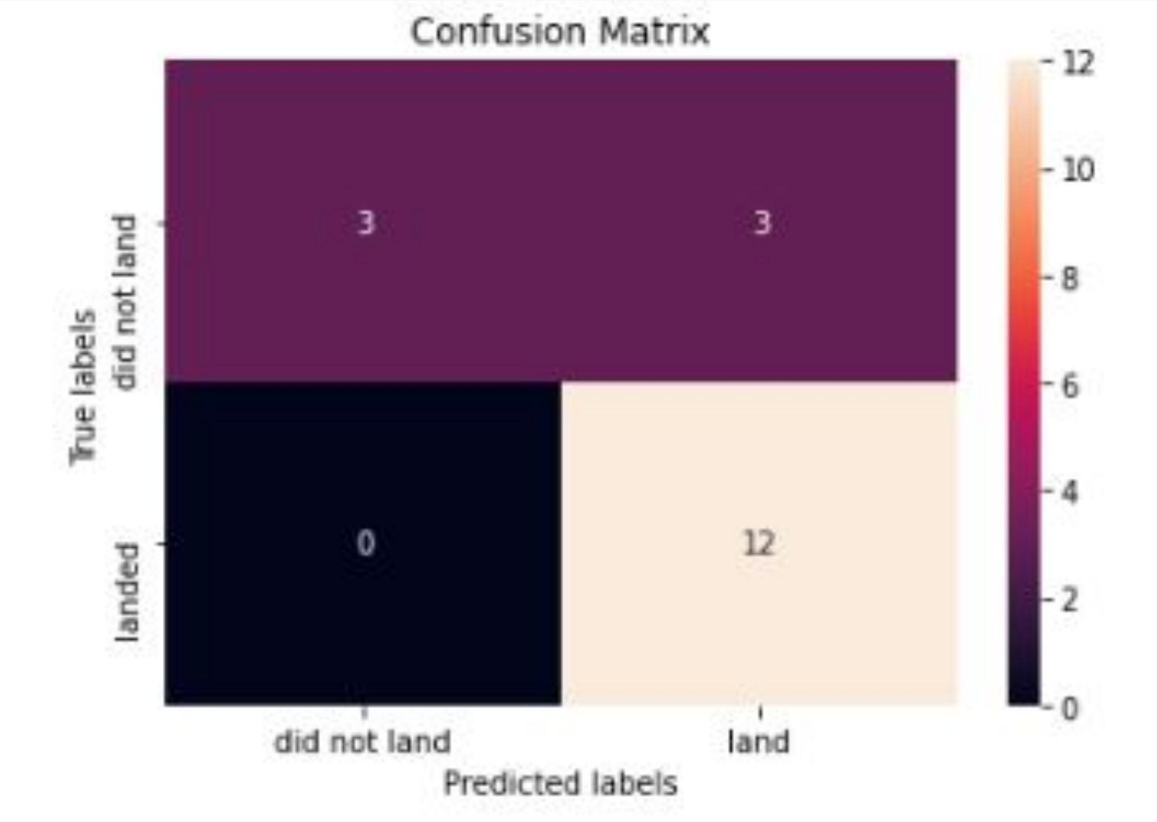
Classification Accuracy



The models have almost the same accuray of 83.33%. The only leading difference is the decision tree classifier

This test size is not done with a lot of samples with only a sample size of 18

Confusion Matrix



The model shows the same test set. The model predicted 12 successful landings. The models predicted 3 unsuccessful landing when the true label was unsuccessful. Some of them are false positives with 3 successful landings being true label as unsuccessful landing.

Conclusions

- The goal of the presentation was to develop machine learning for Space Y who wants to compete with SpaceX.
- The goal of the model is to predict if stage 1 will successfully save 100 million USD
- We were successful in using data from public SpaceX API and SpaceX Wikipedia page
- Successful in creating dashboard for visualization, creating a machine learning model with an accuracy of 83.33%. With this model we can successfully predict whether a launch could be successful or not.

Appendix

- https://github.com/JaysonWan/DataScience_Capstone
- SpaceX data
- Wikipedia

Thank you!