IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Agnė Griškevičiūtė

2022-07-31

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using API, Web scraping
  - Data wrangling
  - Exploratory Data Analysis (EDA) with data visualization
  - EDA with SQL
  - Interactive map analysis with Folium
  - Dashboard with Plotly dash
  - Predictive analysis (classification)
- Summary of all results
  - EDA results
  - Interactive map and dashboard results
  - Predictive analysis results

Introduction

- Companies, such as SpaceX, are making space travel affordable for everyone. SpaceX can do this by launching the rocket relatively inexpensively. The company advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each. Much of the savings is because SpaceX can reuse the first stage. Therefore, if it is possible to determine if the first stage will land, the cost of a launch will be determined.
- **The goal** of this project is to build a machine learning model in order to determine if SpaceX will reuse (land successfully) the first stage.
- The **problems/questions** are the following:
  - What are the factors that determine successful/failed launches?
  - What are the relationship between different variables that determine successful/failed launches?

Section 1

Methodology

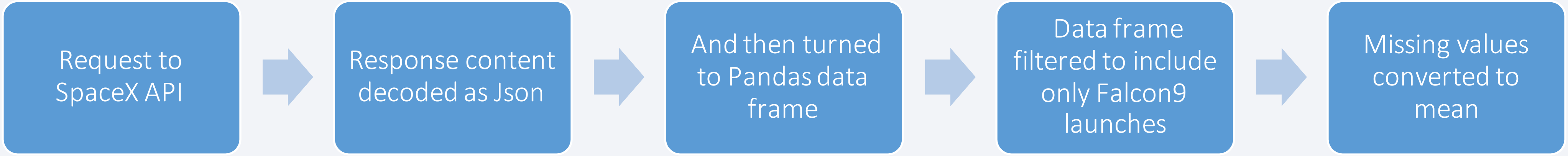
Methodology

Executive Summary

- Data collection methodology:
  - Data gathered from SpaceX API and by web scraping launch records from Wikipedia
- Perform data wrangling
  - EDA was performed; mission outcomes converted to 0 and 1
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data standardization, split to training data and test data
  - Different models were trained and hyperparameters were selected using the function `GridSearchCV`. The most accurate method was calculated on test data using method `score`.

Data Collection

- SpaceX launch data was gathered from the SpaceX REST API and converted to Pandas data frame. In addition, Falcon 1 launches have been removed and missing values have been converted to mean.



Request to SpaceX API

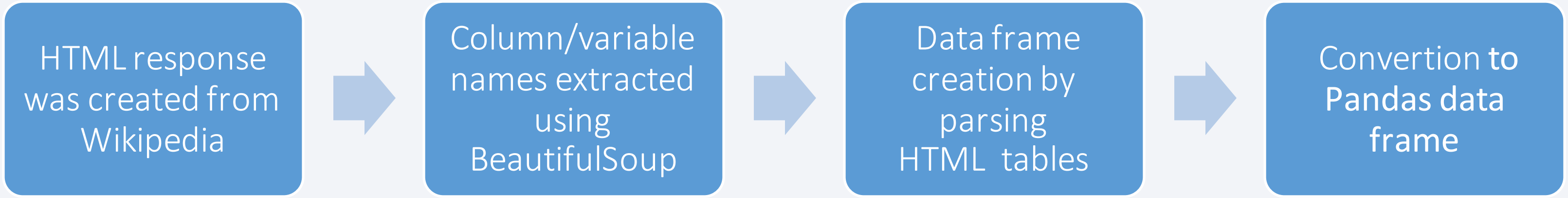
Response content decoded as Json

And then turned to Pandas data frame

Data frame filtered to include only Falcon9 launches

Missing values converted to mean

- Falcon 9 launch records were web scraped from Wikipedia and the HTML table was parsed and converted to Pandas data frame for further visualization and analysis.



HTML response was created from Wikipedia

Column/variable names extracted using BeautifulSoup

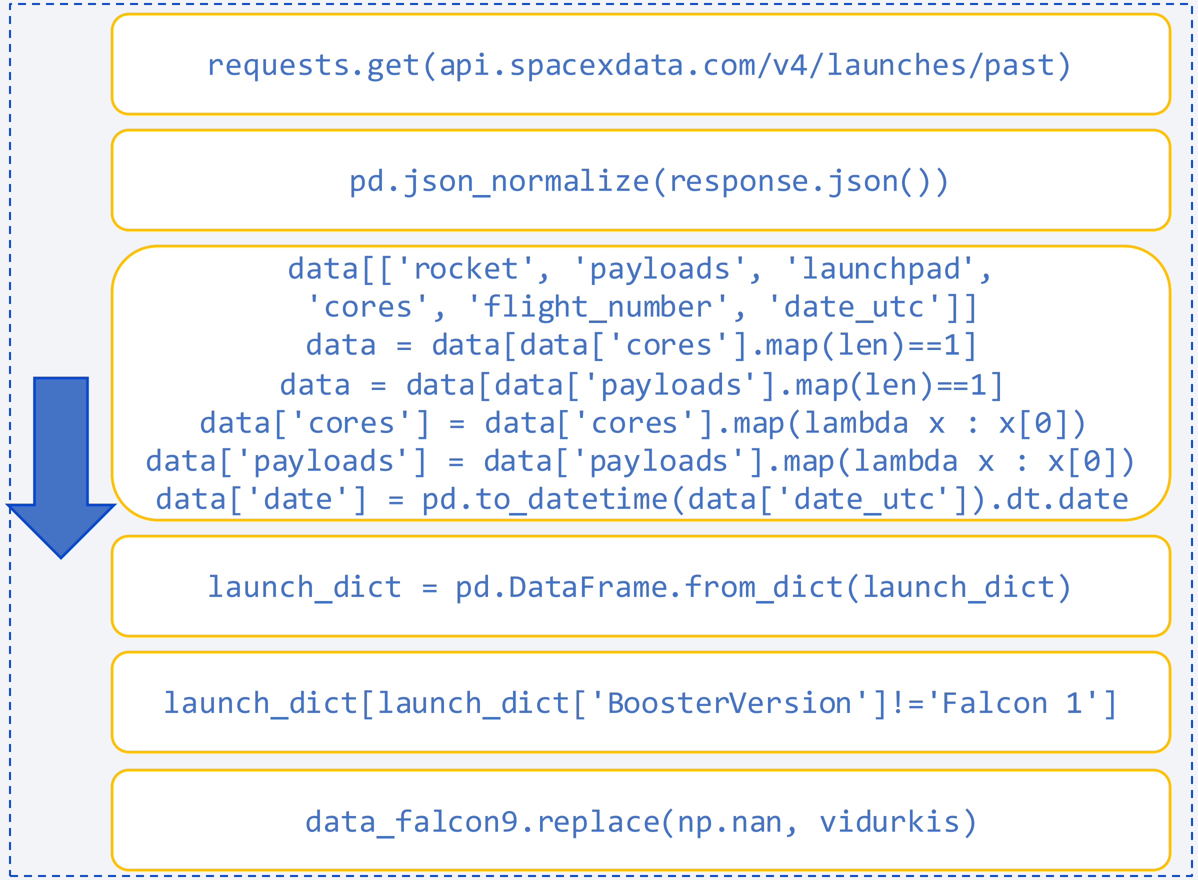
Data frame creation by parsing HTML tables

Convertion to Pandas data frame

Data Collection – SpaceX API

1. Rocket launch data was requested from SpaceX API with the following URL: api.spacexdata.com/v4/launches/past.
2. Response content was decoded as a Json using `.json()`. To convert JSON into a Pandas data frame, `.json_normalize()` function was used.
3. Rows with multiple cores were removed. For payloads and cores, the single value has been extracted in a list and replaced the feature. Date utc converted to datatype.
4. Columns combined into a dictionary (`launch_dict`) and Pandas data frame created.
5. Falcon 1 launches were removed keeping only the Falcon 9 launches.
6. Missing values converted to mean.

GitHub URL



```
requests.get(api.spacexdata.com/v4/launches/past)
```

```
pd.json_normalize(response.json())
```

```
data[['rocket', 'payloads', 'launchpad',
'cores', 'flight_number', 'date_utc']]
data = data[data['cores'].map(len)==1]
data = data[data['payloads'].map(len)==1]
data['cores'] = data['cores'].map(lambda x : x[0])
data['payloads'] = data['payloads'].map(lambda x : x[0])
data['date'] = pd.to_datetime(data['date_utc']).dt.date
```

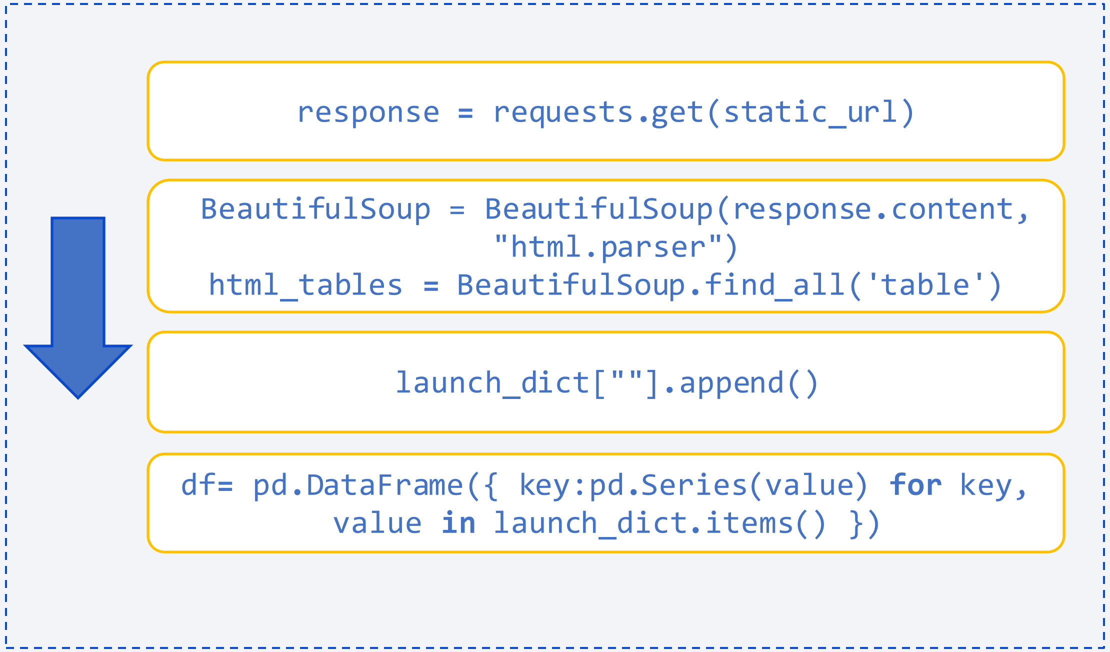
```
launch_dict = pd.DataFrame.from_dict(launch_dict)
```

```
launch_dict[launch_dict['BoosterVersion']!='Falcon 1']
```

```
data_falcon9.replace(np.nan, vidurkis)
```

Data Collection - Scraping

1. HTTP GET method was used to request Falcon9 Launch HTML page, as an HTTP response.
2. `BeautifulSoup` package used to web scrape HTML tables.
3. An empty dictionary (`launch_dict`) was created and filled up with launch records by extracting them from table rows.
4. Dictionary was then converted into Pandas data frame for further visualization and analysis.



```
response = requests.get(static_url)
```

```
BeautifulSoup = BeautifulSoup(response.content,
                "html.parser")
html_tables = BeautifulSoup.find_all('table')
```

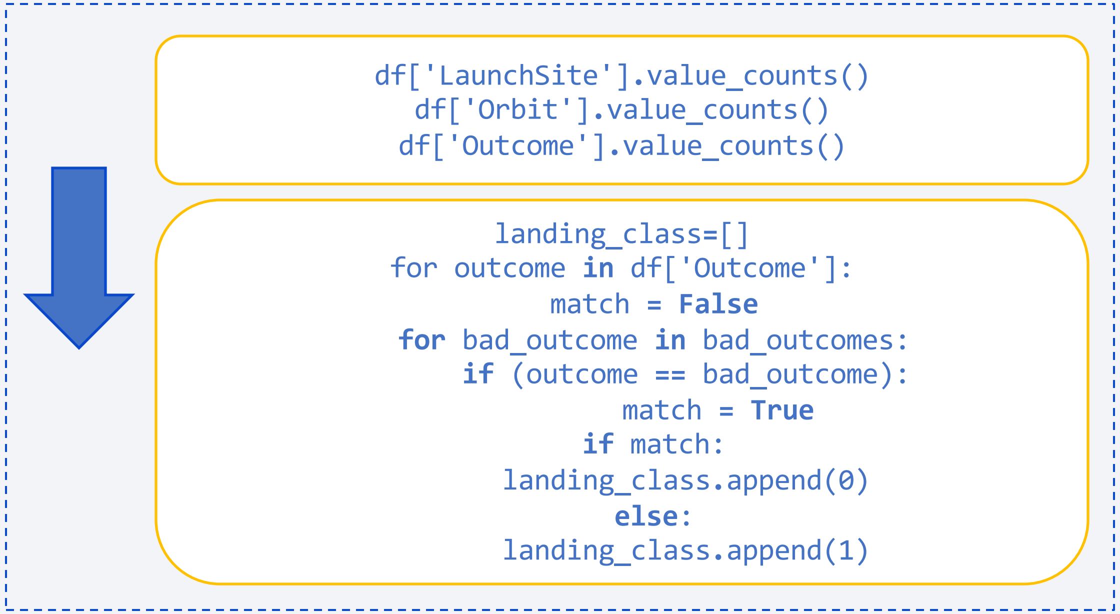
```
launch_dict[""].append()
```

```
df= pd.DataFrame({ key:pd.Series(value) for key,
    value in launch_dict.items() })
```

GitHub URL

Data Wrangling

1. Exploratory Data Analysis was performed in order to find patterns in the data:
   - the number of launches on each site were calculated
   - the number and occurrence of each orbit was calculated
   - the number and occurrence of mission outcome per orbit type was calculated
2. Landing outcome label of each launch was created (0 if bad outcome, 1 otherwise).



```
df['LaunchSite'].value_counts()
df['Orbit'].value_counts()
df['Outcome'].value_counts()
```

```
landing_class=[]
for outcome in df['Outcome']:
    match = False
    for bad_outcome in bad_outcomes:
        if (outcome == bad_outcome):
            match = True
    if match:
        landing_class.append(0)
    else:
        landing_class.append(1)
```

GitHub URL

EDA with Data Visualization

- The following charts have been plotted to visualize the relationship between:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Success rate of each Orbit type
  - Flight number and Orbit type
  - Payload and Orbit type
  - Launch success yearly trend
- These charts used to determine what attributes are correlated with successful landings, how each important variable would affect the success rate, select the features to be used in success prediction.

GitHub URL

EDA with SQL

The following SQL queries have been executed for detailed analysis of different variables/relations:

- Display the names of the unique launch sites in the space mission and 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 successful landing outcome in ground pad was achieved;
- List the names of the boosters which have success in drone ship and have payload mass > 4000 but < 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;
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

GitHub URL

Build an Interactive Map with Folium

- Different map objects have been created and added to a folium map in order to give some insights if the launch success rate may also depend on the location and proximities of a launch site to railway, highway, coastline, etc..
- The following objects were created:
  - highlighted circle area showing launch site area,
  - markers with text label showing launch site name,
  - color-labeled markers in marker clusters (green color showing successful launch, red – failed),
  - distance lines.

GitHub URL

Build a Dashboard with Plotly Dash

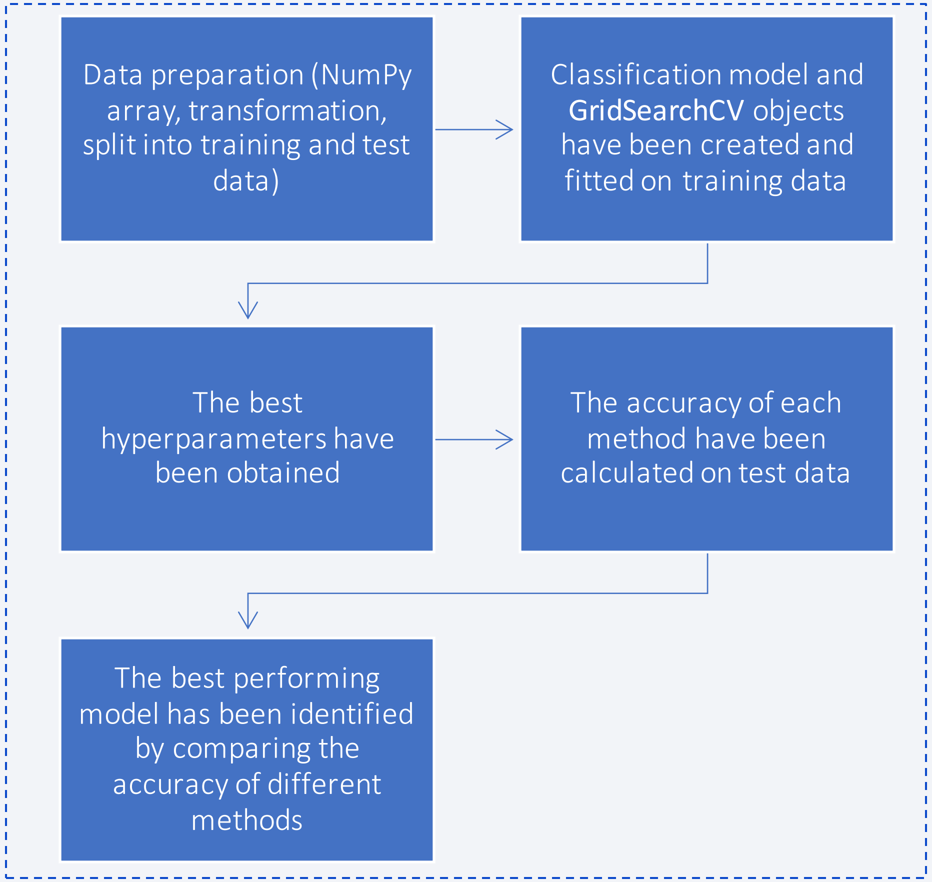
- Interactive dashboard was build using Python Plotly Dash package to find more insights from the SpaceX data.
- The following charts have been built in the dashboard:
  - Pie chart showing success rates for specific launch site and for total launches,
  - Scatter point chart showing relationship between payload mass (kg) and launch outcome for different booster versions,
- The dashboard also contains input components, such as dropdown list and range slider that allows to interact with scatter point and pie charts.

GitHub URL

Predictive Analysis (Classification)

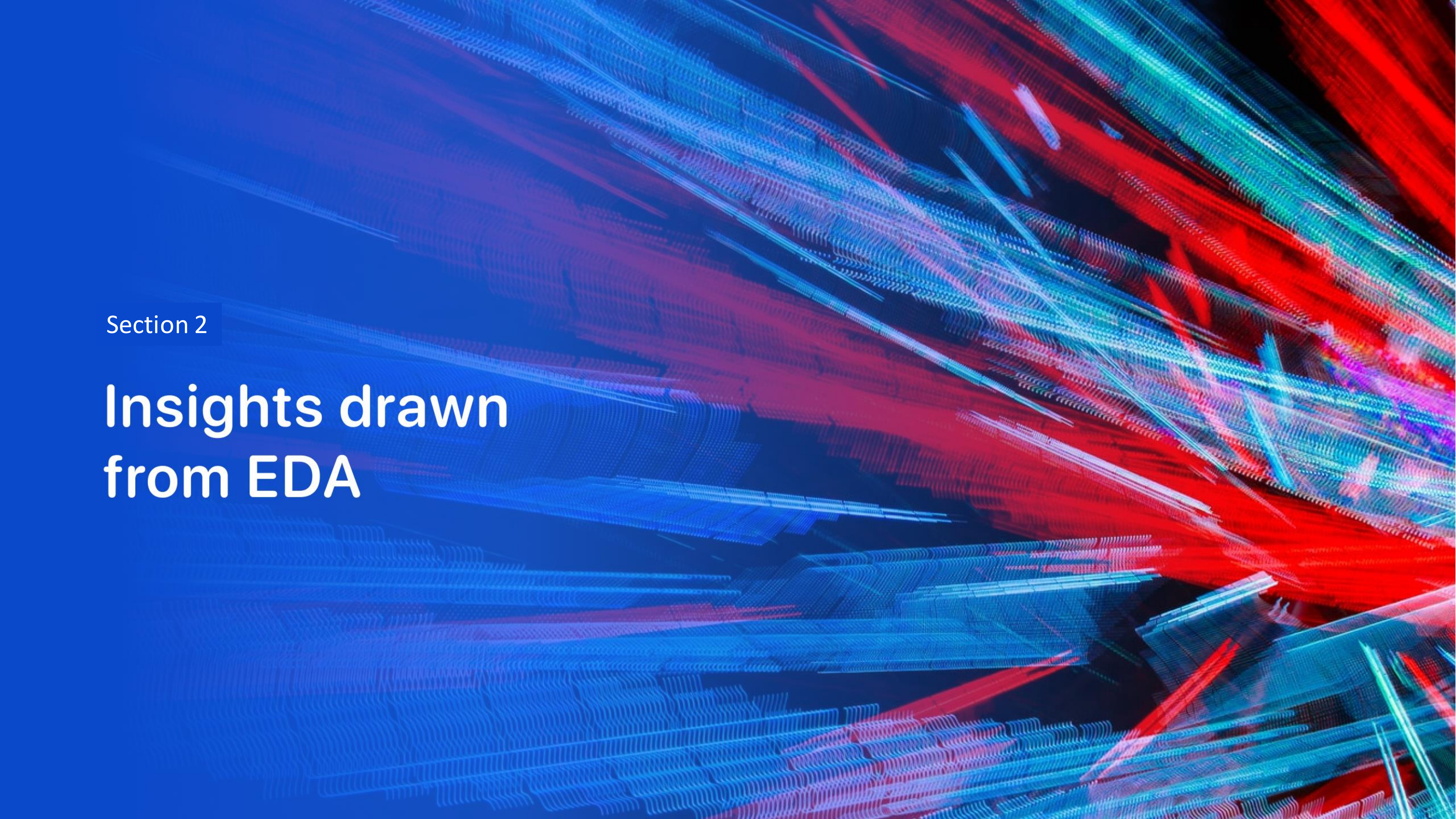
- Data was loaded and prepared for the modeling by creating NumPy array for dependent variable, transforming data and splitting into training and test data using function `train_test_split`.
- SVM, KNN, Classification Trees and Logistic Regression methods have been built and different hyperparameters have been obtained using `GridSearchCV` function.
- The accuracy of each method has been calculated on test data using `score` method and confusion matrix visualization.
- The best performing model has been identified by comparing the accuracy of each model.

GitHub URL


Data preparation (NumPy array, transformation, split into training and test data)
Classification model and **GridSearchCV** objects have been created and fitted on training data
The best hyperparameters have been obtained
The accuracy of each method have been calculated on test data
The best performing model has been identified by comparing the accuracy of different methods

Results

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

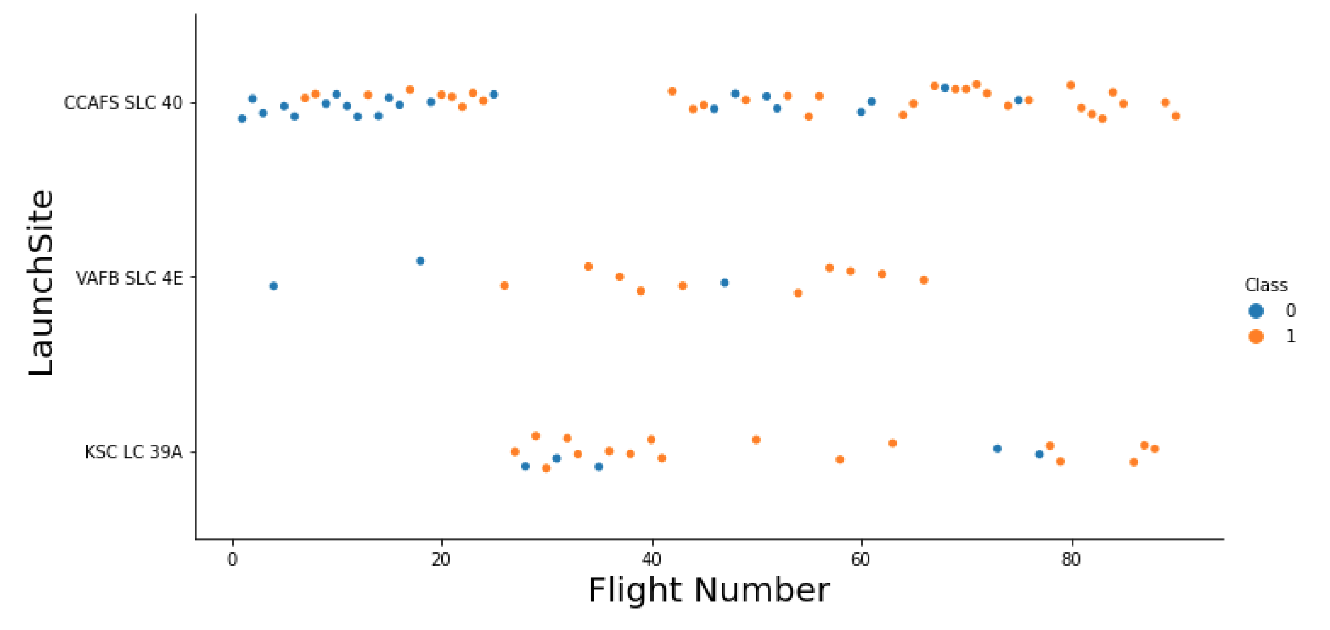


Section 2

Insights drawn from EDA

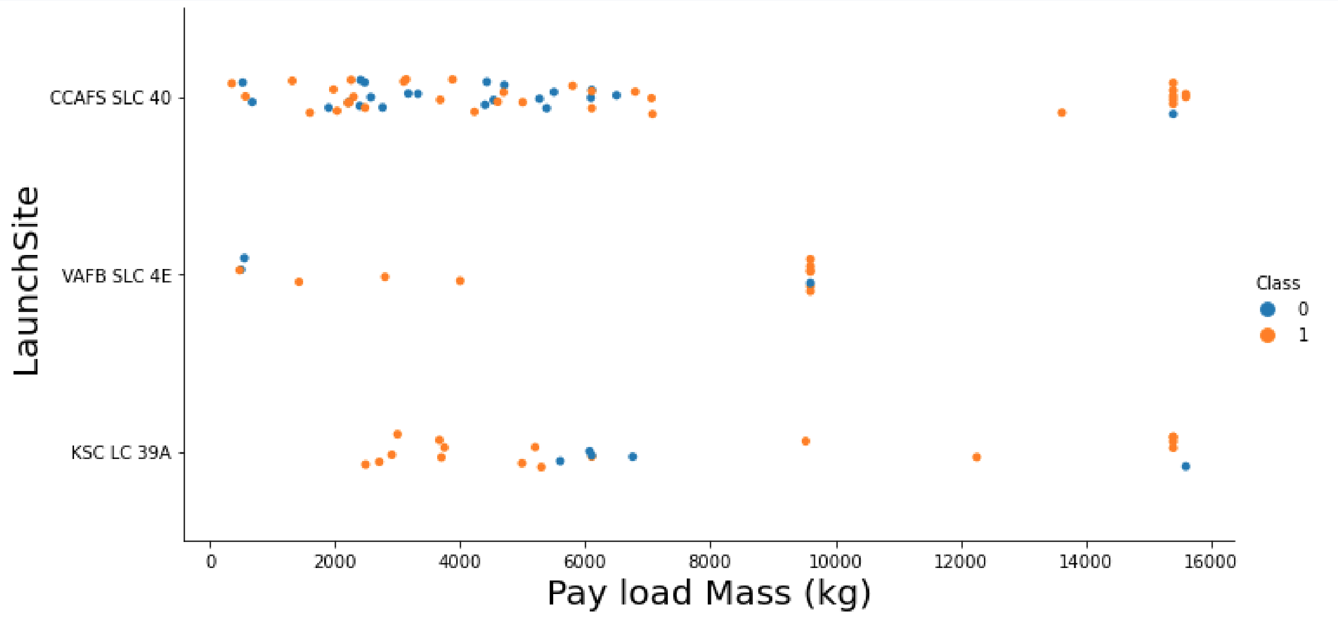
Flight Number vs. Launch Site

- It was found that CCAFS SLC 40 launch site has lower success rates compared to other launch sites (VAFB SLC 4E and KSC LC 39A).
- It can be observed from the plot that as the flight number increases, the more likely it is to land successfully.
- KSC LC 39A site had only higher flight numbers.



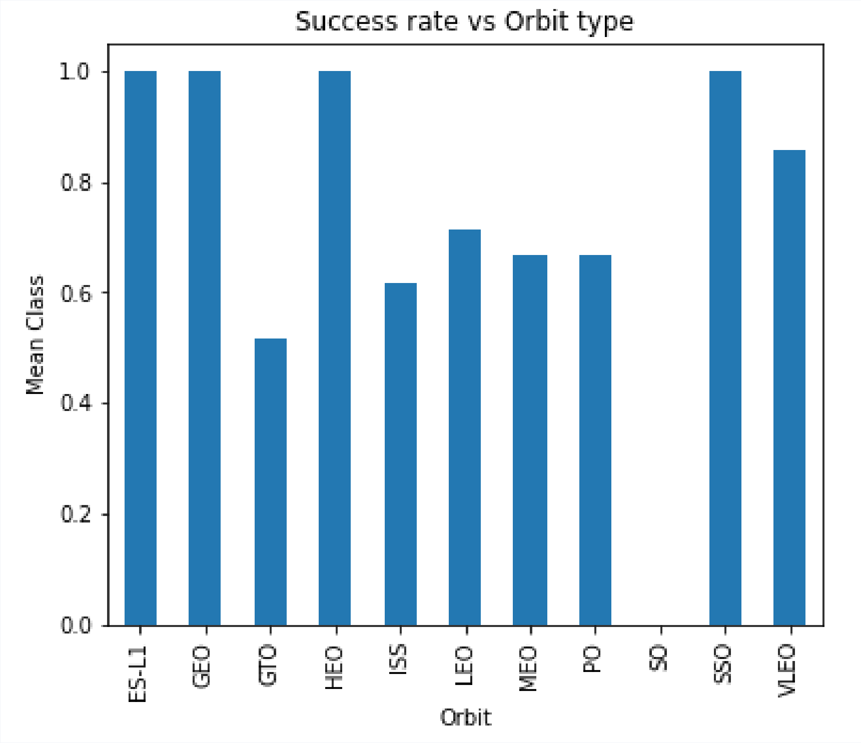
Payload vs. Launch Site

- Success rates in both heavy and light payloads were mixed.
- In the VAFB SLC 4E launch site there weren't any heavier payload mass rockets.



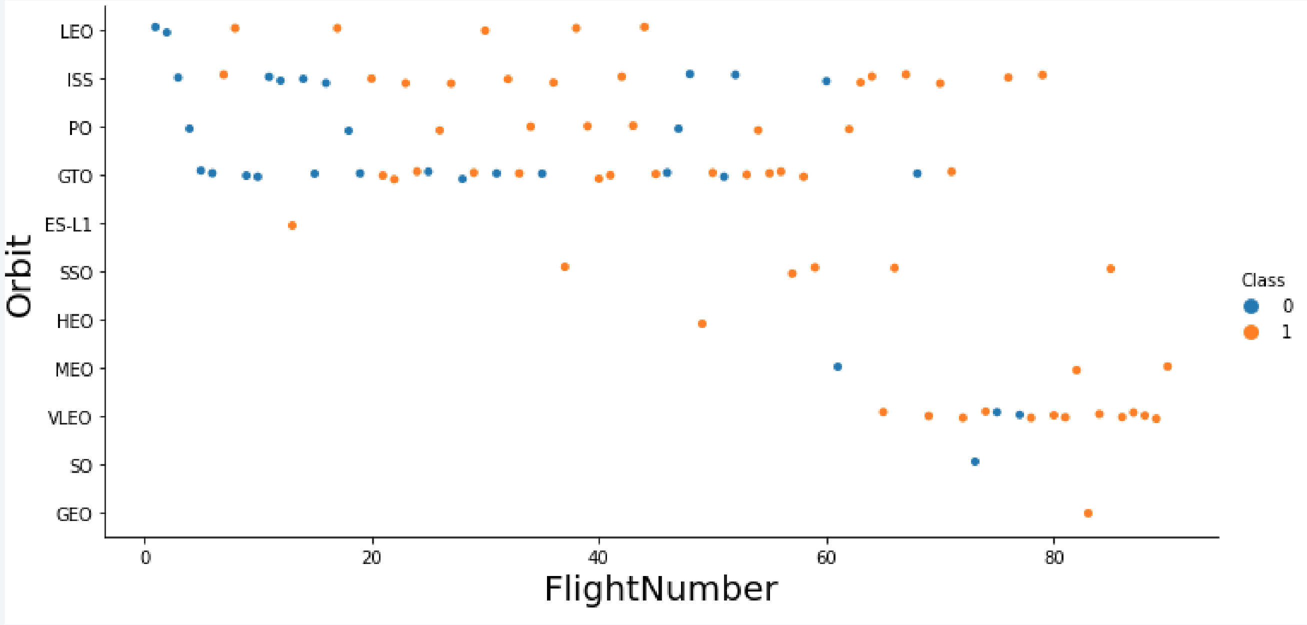
Success Rate vs. Orbit Type

- From the chart it is seen that ES-L1, GEO, HEO and SSO orbits have highest success rates (100%).
- SO orbit has the lowest success rate (0%).



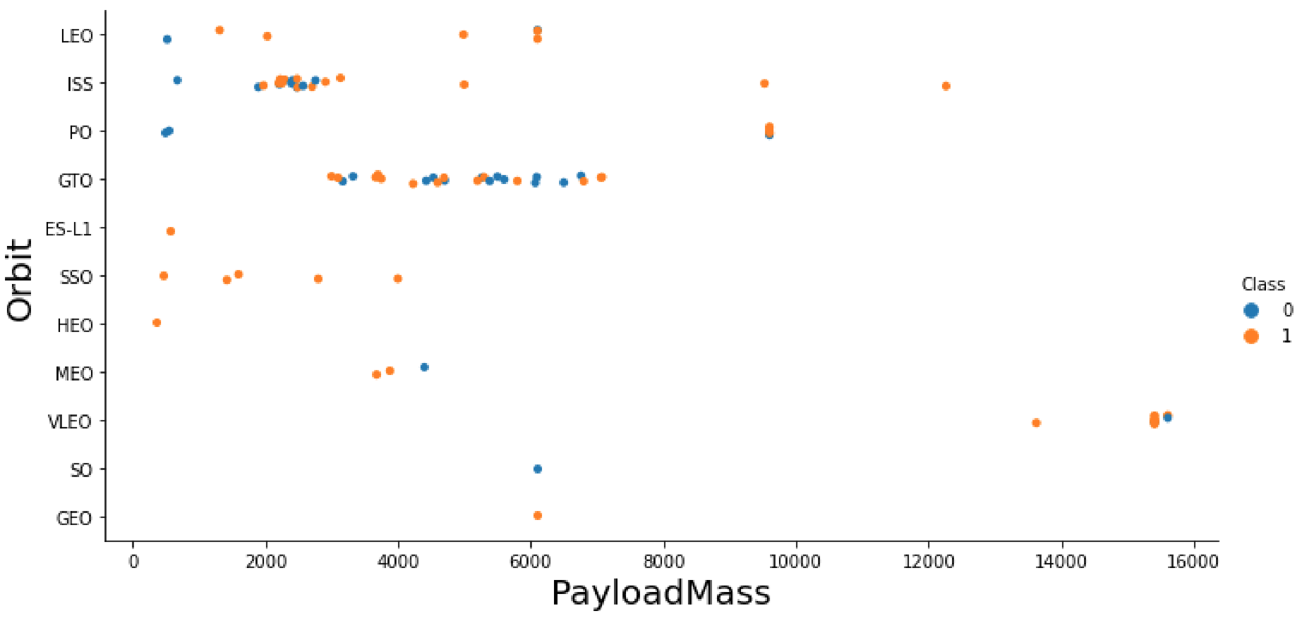
Flight Number vs. Orbit Type

- GEO, SO, VLEO, MEO orbits had only high flight numbers (greater than 60).
- In LEO orbit there seems to be a link between flight number and success rate (i.e. as flight number increases, the first stage is likely to land successfully). However, in other orbits such link is not observed.



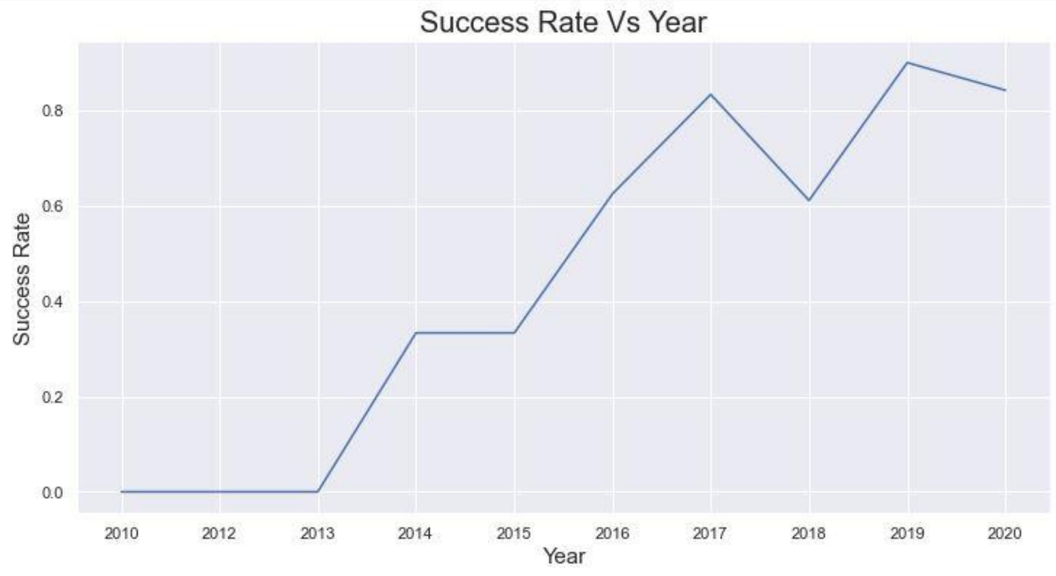
Payload vs. Orbit Type

- More successful landings can be observed with heavier payloads in LEO, ISS, PO orbits, however, no such clear pattern can be observed in GTO orbit, where both heavy and light payloads had positive and negative landing rates.



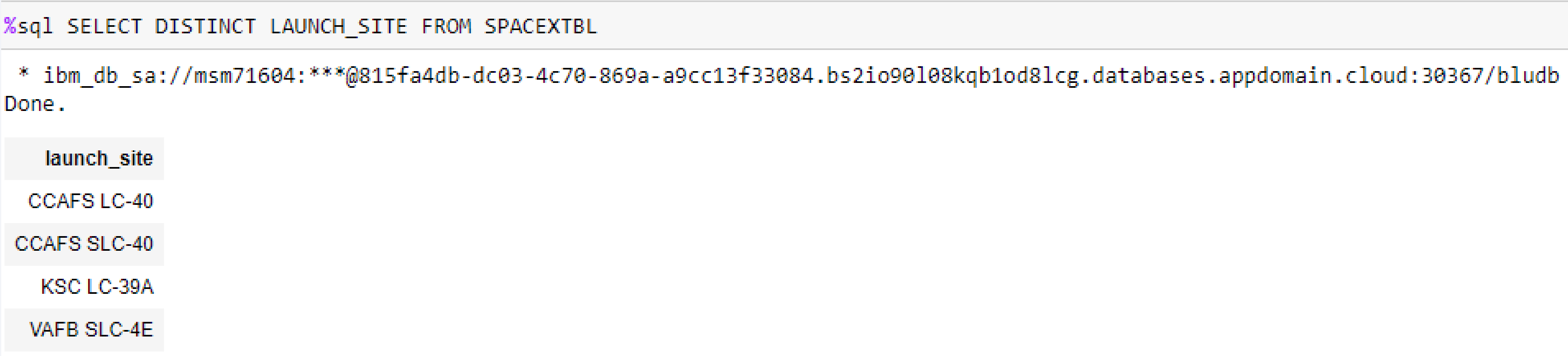
Launch Success Yearly Trend

- Line chart indicates that 2010-2013 period was unsuccessful. However, starting from 2013 success rate started to increase.



All Launch Site Names

- Using DISTINCT in the query, it was found that in the space mission there were four unique launch sites:



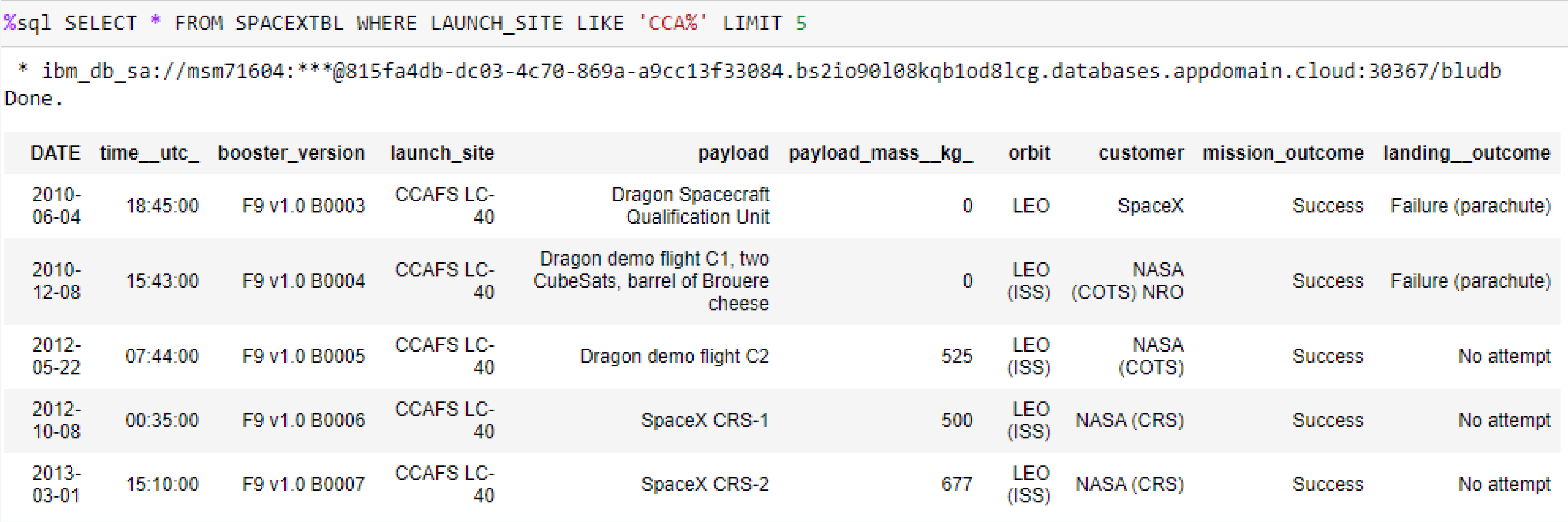
```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

```
 * ibm_db_sa://msm71604:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

- The following screenshot shows 5 records WHERE launch sites begin with string `CCA`:



```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

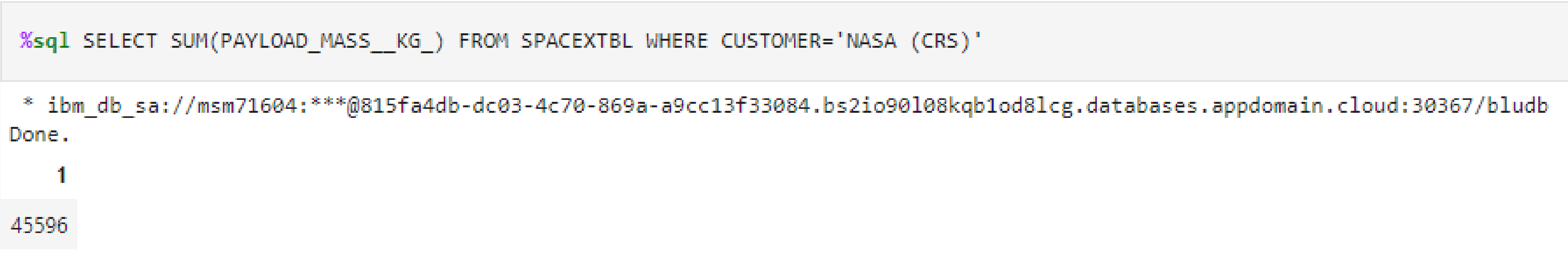
```
 * ibm_db_sa://msm71604:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

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

- Using LIKE helped to filter launch sites that contain string 'CCA' and LIMIT allows to display only specific number of records.

Total Payload Mass

- Using SUM for the payload mass (kg), it was found that the total payload mass carried by boosters launched by NASA (CRS) is equal to 45596:

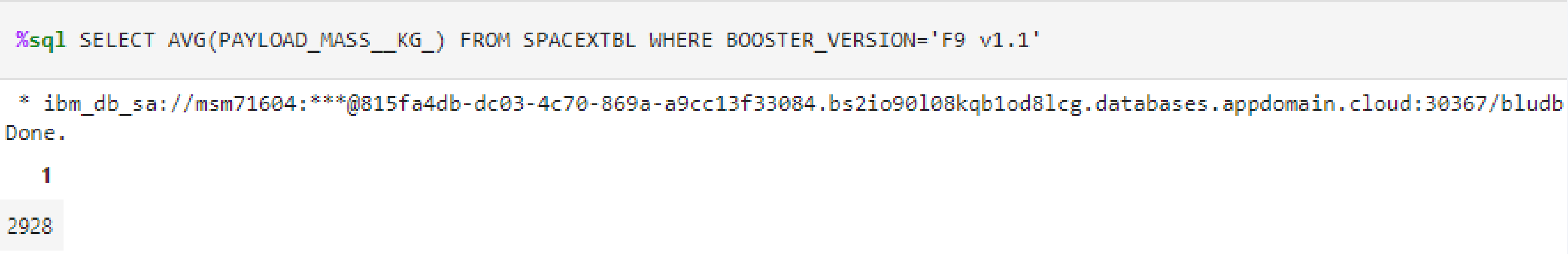


```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

```
 * ibm_db_sa://msm71604:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
    1
45596
```

Average Payload Mass by F9 v1.1

- Using AVERAGE for the payload mass (kg), it was found that the average payload mass by booster version F9 v1.1 is equal to 2928:

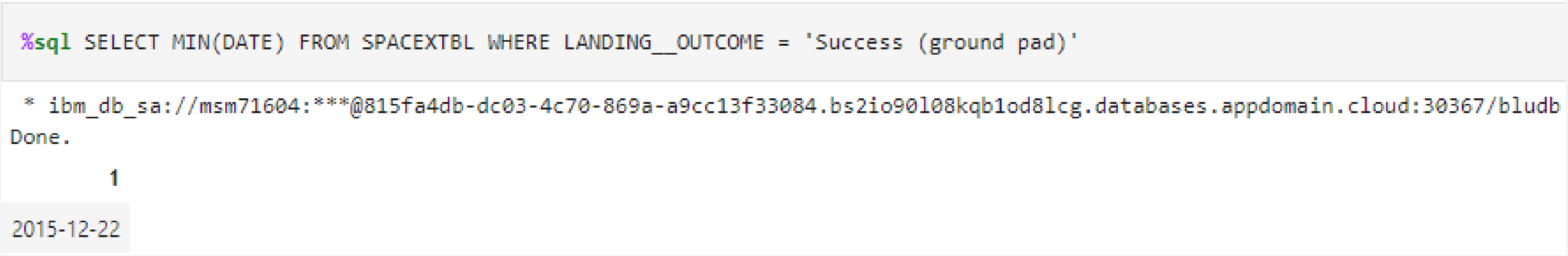


```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1'
```

```
 * ibm_db_sa://msm71604:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
   1
2928
```

First Successful Ground Landing Date

- From execution of MIN(DATE) function it was found that the earliest date of the first successful landing outcome on ground pad was 22nd of December, 2015:



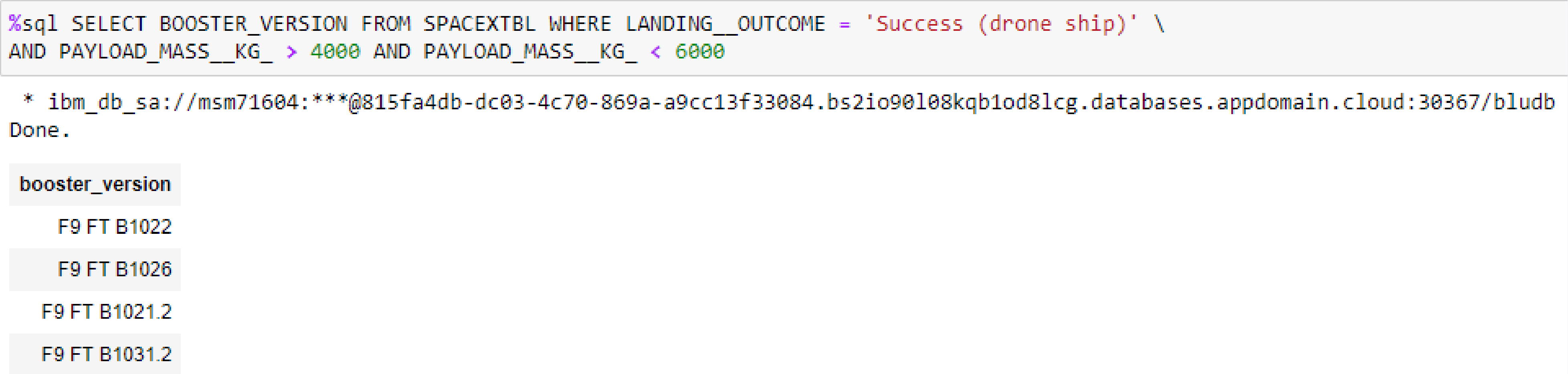
```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://msm71604:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The following 4 names of the boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 have been identified:



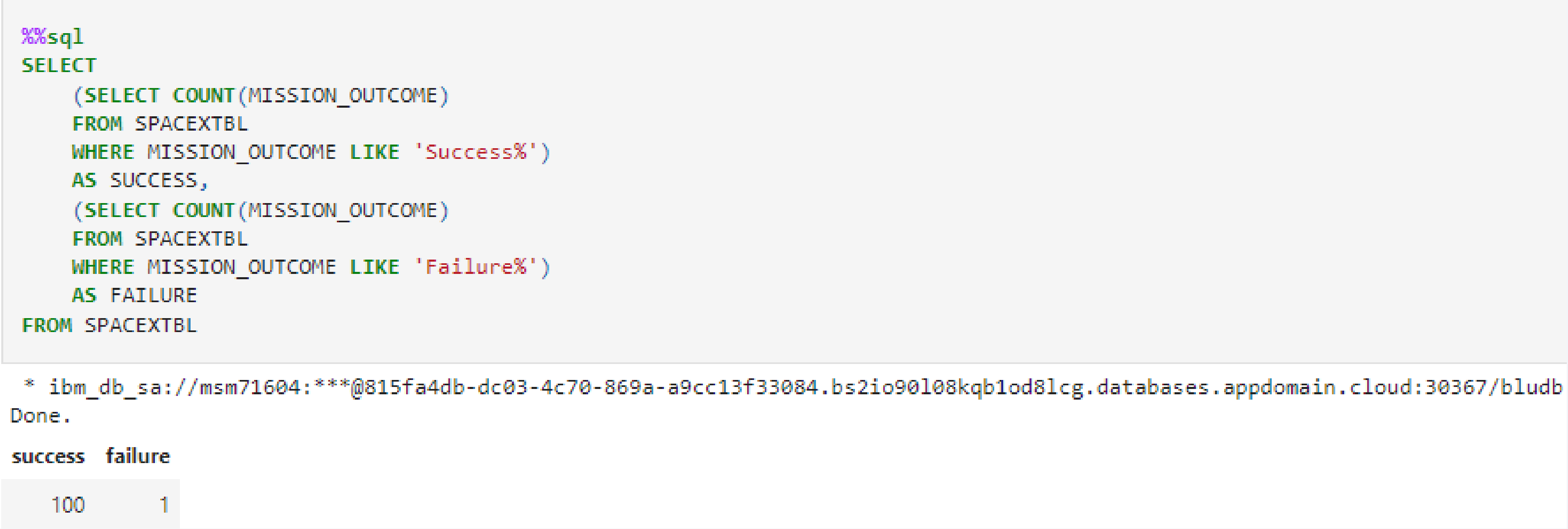
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

```
 * ibm_db_sa://msm71604:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
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

- Using two subqueries it was found that in total there were 100 successful and 1 failure mission outcomes:



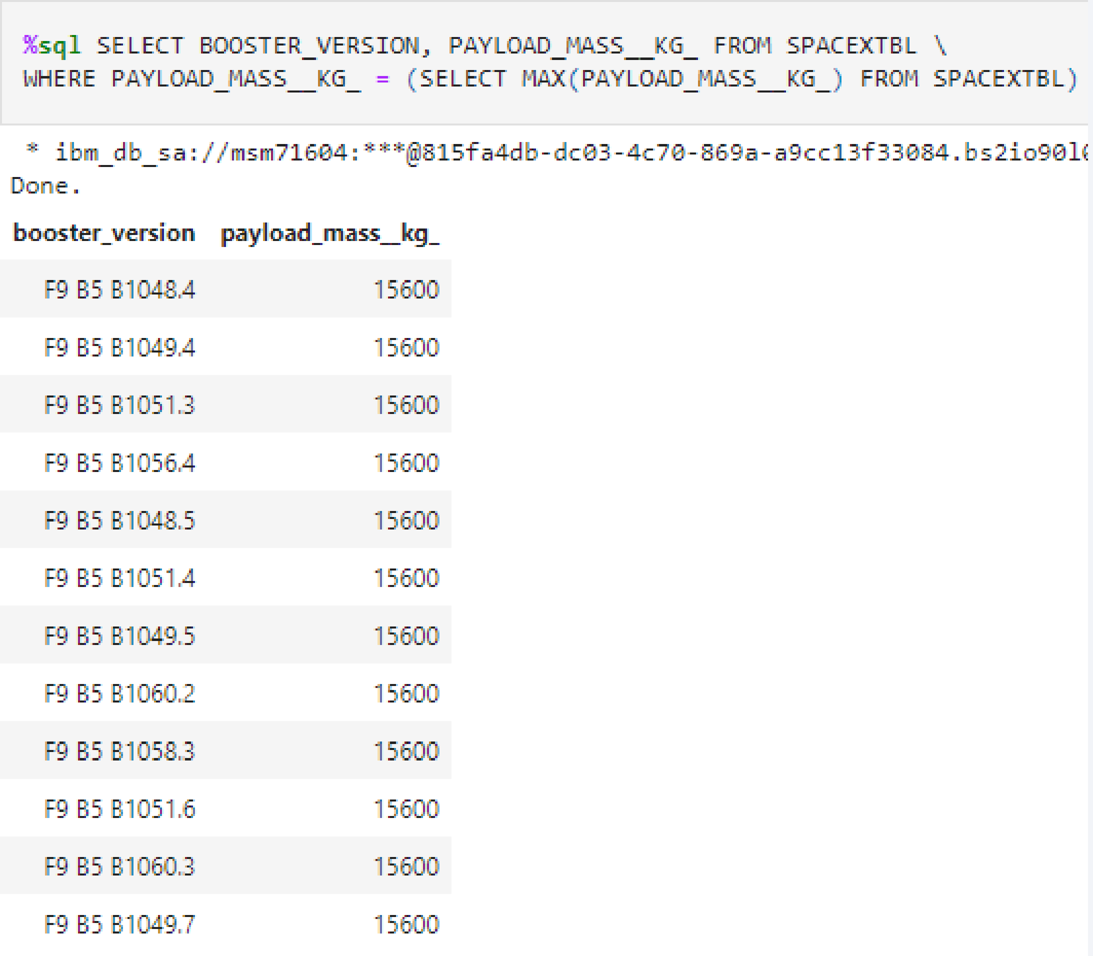
```
%%sql
SELECT
    (SELECT COUNT(MISSION_OUTCOME)
    FROM SPACEXTBL
    WHERE MISSION_OUTCOME LIKE 'Success%')
    AS SUCCESS,
    (SELECT COUNT(MISSION_OUTCOME)
    FROM SPACEXTBL
    WHERE MISSION_OUTCOME LIKE 'Failure%')
    AS FAILURE
FROM SPACEXTBL
```

```
 * ibm_db_sa://msm71604:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

success	failure
100	1

Boosters Carried Maximum Payload

- Using a subquery to select only maximum payload mass (kg), it was found that 12 boosters carried the maximum payload mass (15600 kg).



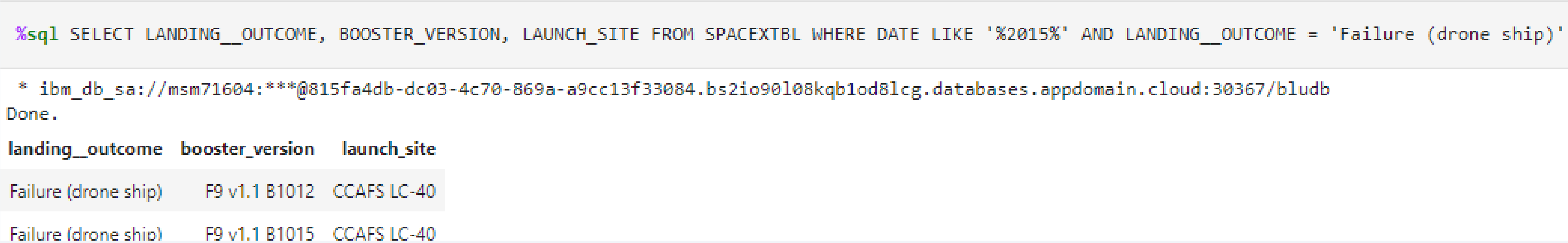
```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * ibm_db_sa://msm71604:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l
Done.
```

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

2015 Launch Records

- The query bellow uses WHERE clause to find failed landing outcomes in drone ship for year 2015 with their corresponding booster versions:



```
%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE DATE LIKE '%2015%' AND LANDING__OUTCOME = 'Failure (drone ship)'
```

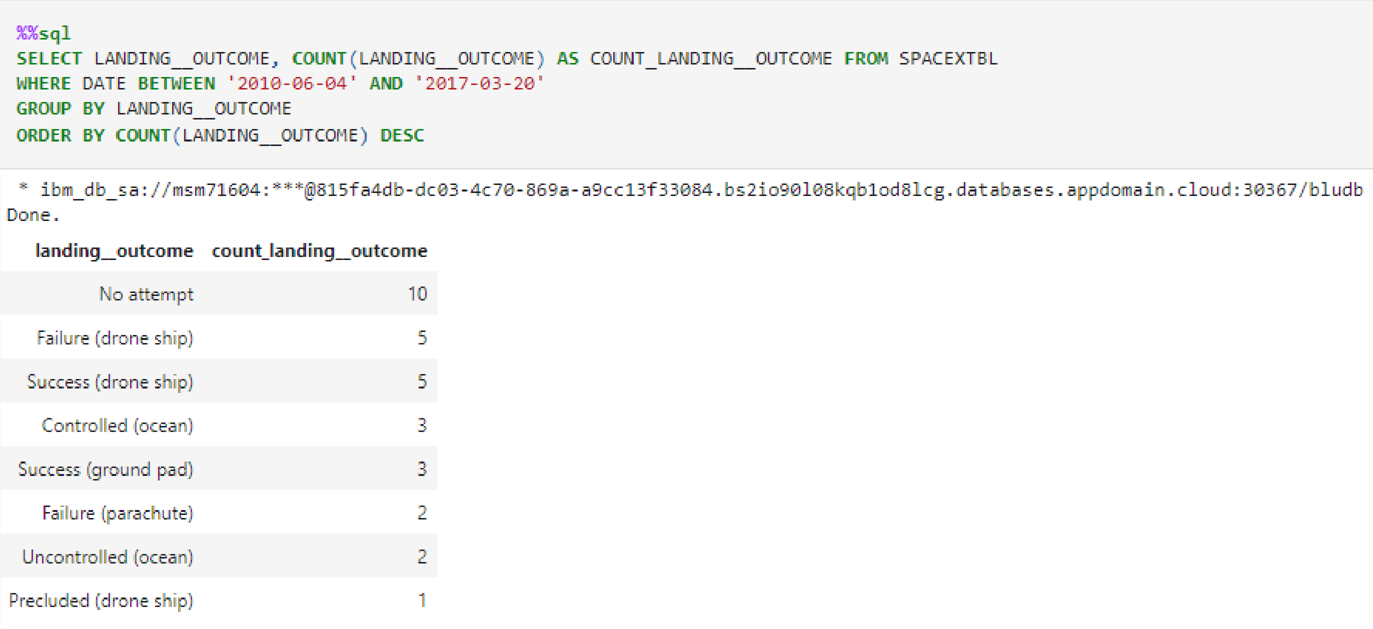
```
 * ibm_db_sa://msm71604:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

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

- It was found that in 2015 there were 2 failed landing outcomes. The booster versions of these outcomes were F9 v1.1 B1012 and F9 v1.1 B1012 and launch site CCAFS LC-40:

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

- After ranking the COUNT of landing outcomes in period 2010-06-04 – 2017-03-20, it was identified that significant part of such outcomes (32%) had landing outcomes 'no attempt' and for the drone ship successful and failure outcomes have been the same (5):



```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS COUNT_LANDING__OUTCOME FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY COUNT(LANDING__OUTCOME) DESC
```

```
 * ibm_db_sa://msm71604:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

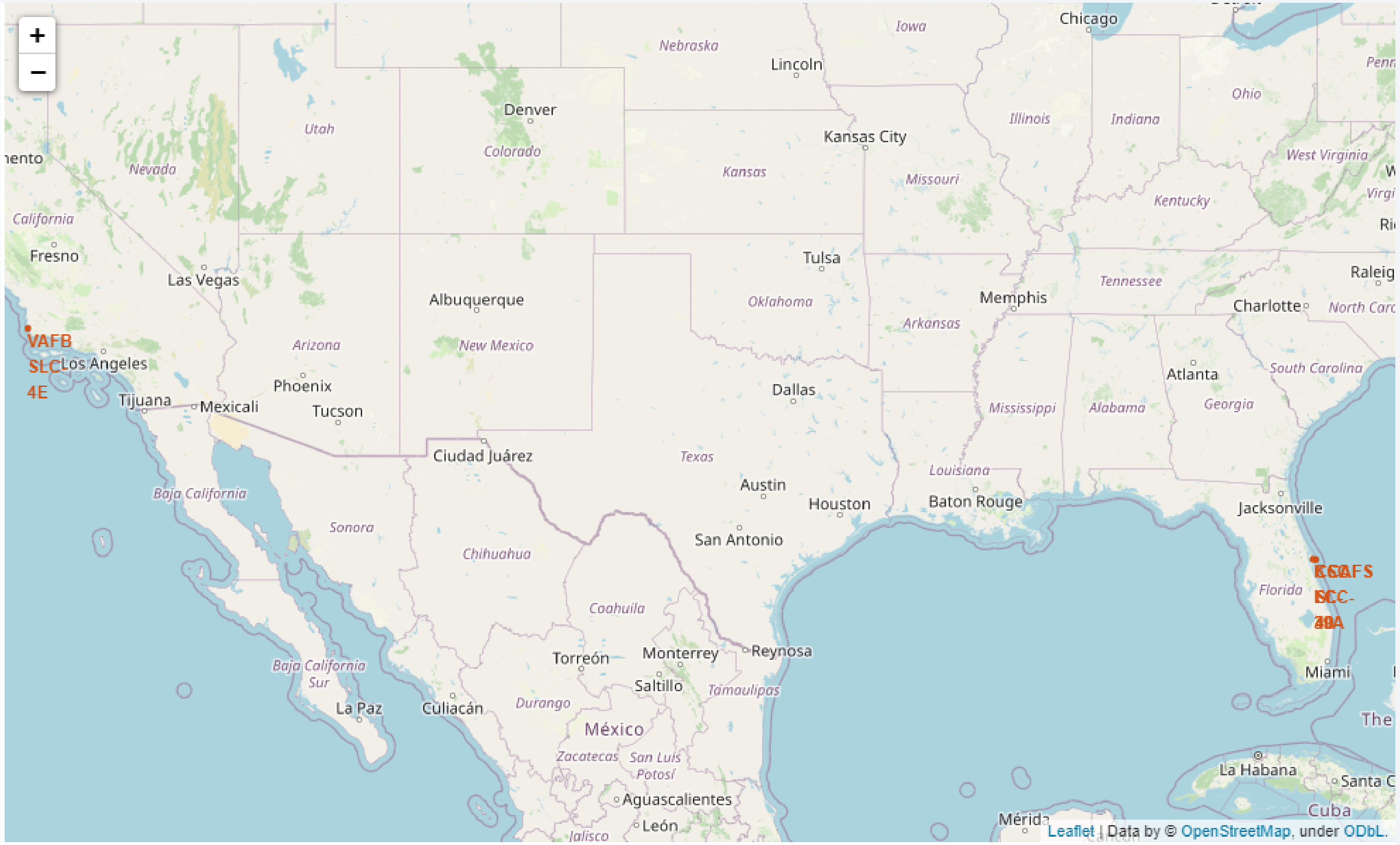
landing__outcome	count_landing__outcome
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

All launch sites on a Folium Map

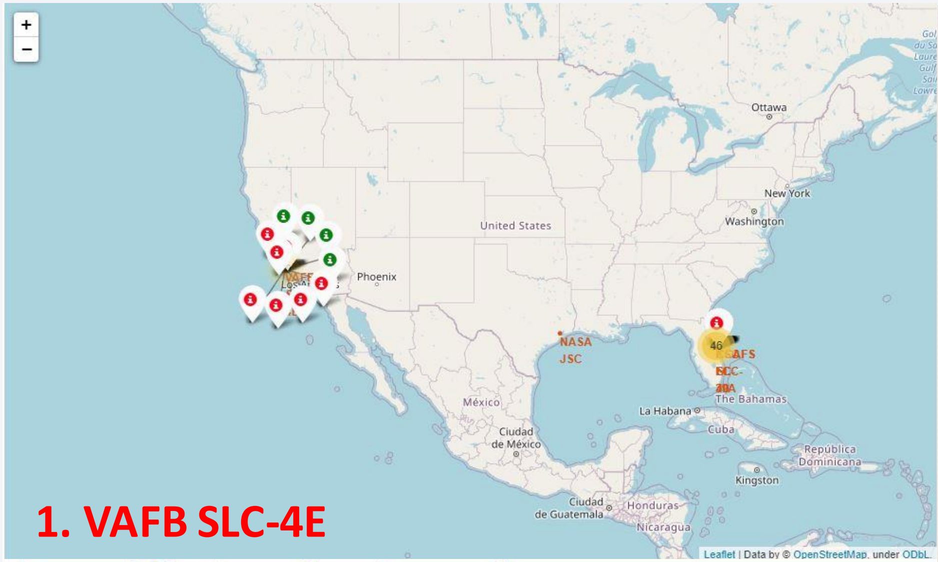
- After addition of launch sites' location markers on a global map, it is seen that all launch sites are located in USA and in close proximity to the coast.
- 3 out of 4 launch sites are located in Florida, the remaining launch site - in California.

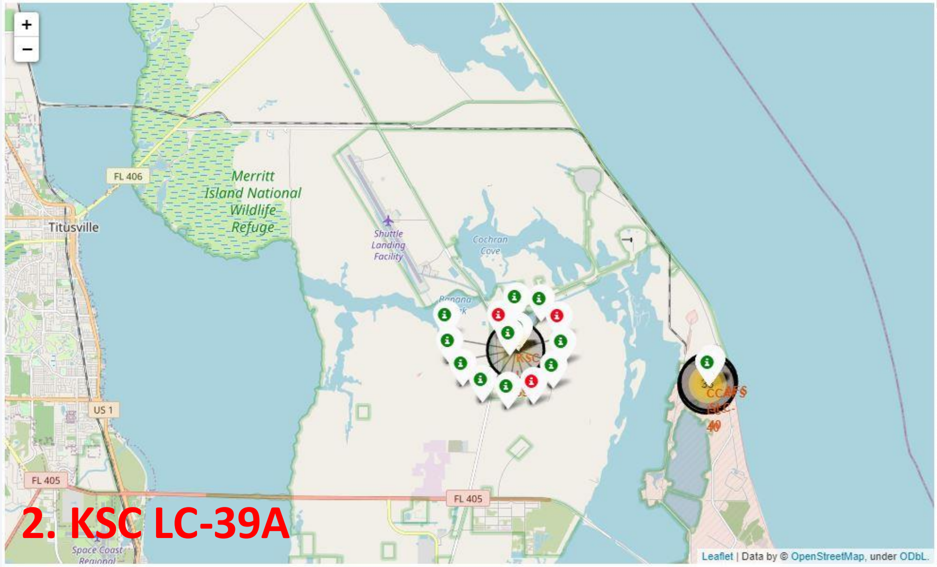


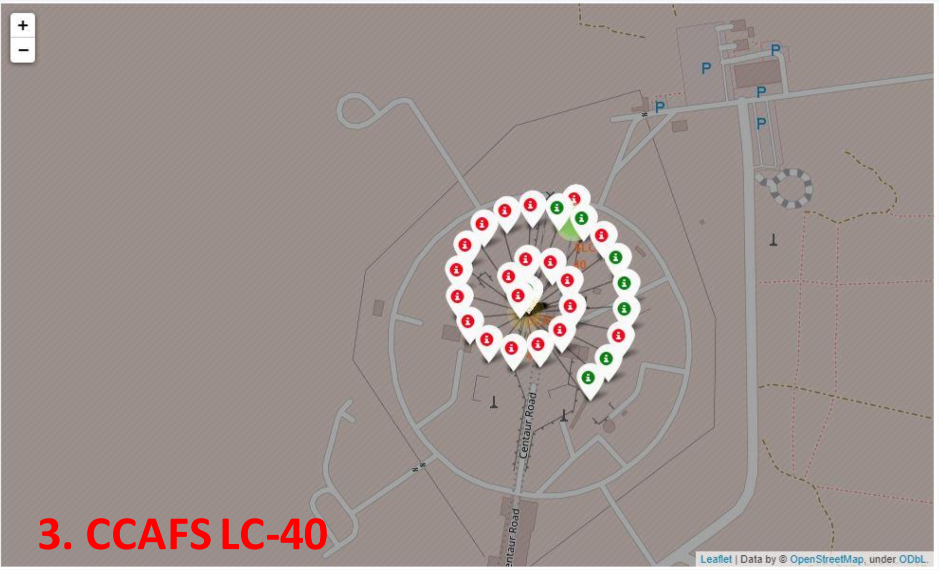
Success/failed launches for each site*

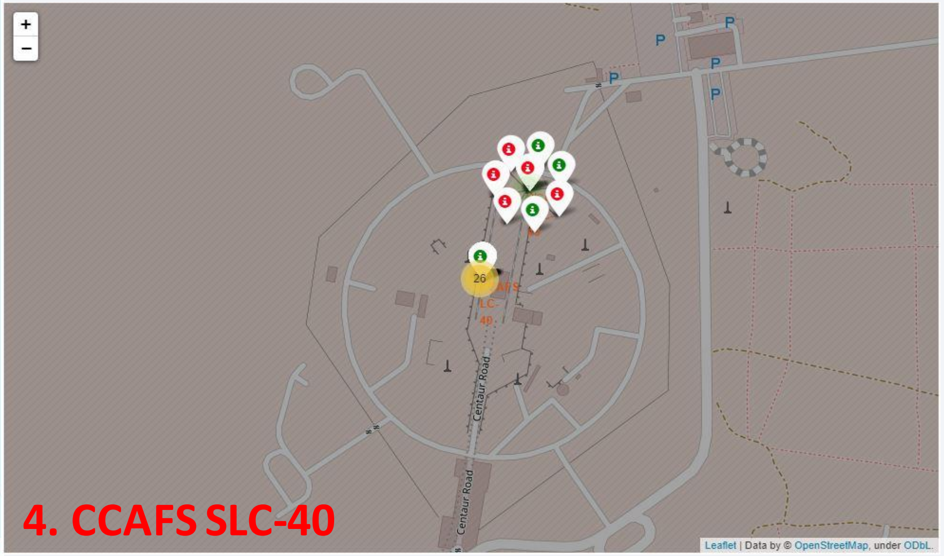
- From the folium maps it can be seen that KSC LC-39A site have relatively high success rates (as %, of total launches)
- CCAFS LC-40 site has the lowest success rates.

* green marker shows success launches, red marker shows failed launches


1. VAFB SLC-4E

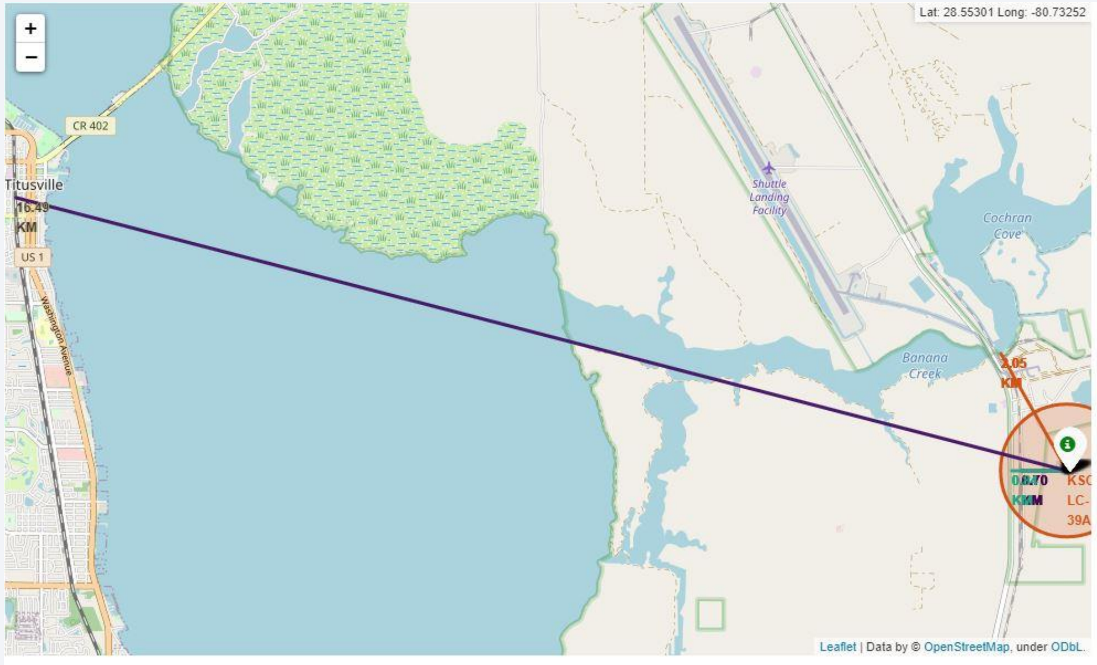

2. KSC LC-39A
+
−
FL 406
Titusville
Merritt Island National Wildlife Refuge
Shuttle Landing Facility
Cochran Cove
US 1
FL 405
CCAFS
Leaflet | Data by © OpenStreetMap, under ODbL.

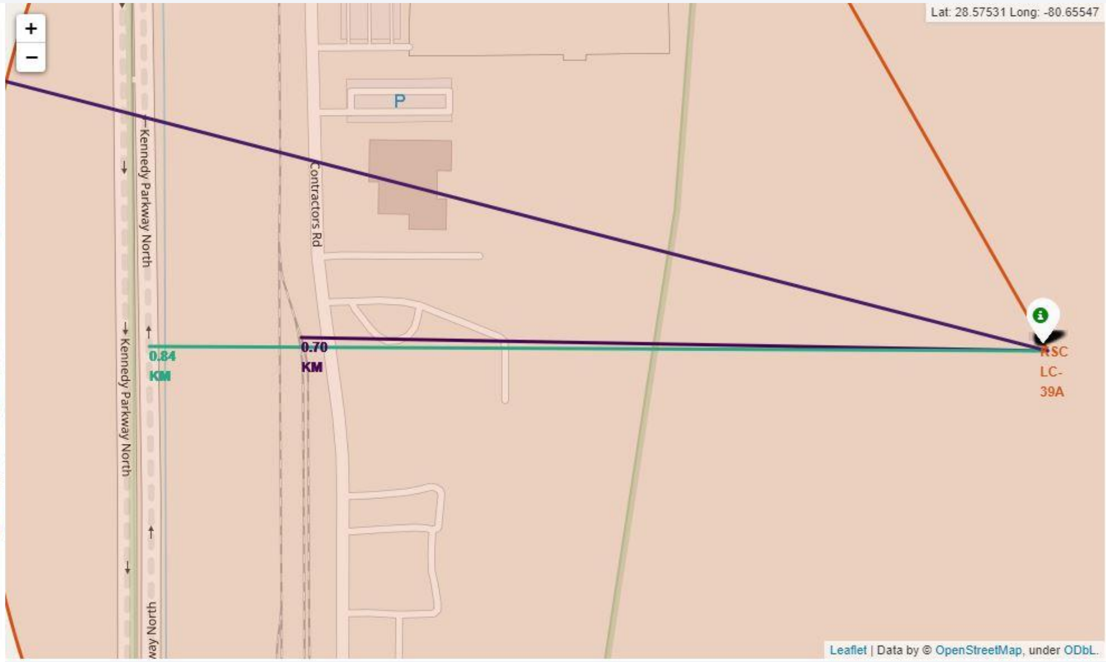

3. CCAFS LC-40


4. CCAFS SLC-40

Distances between launch site to its proximities

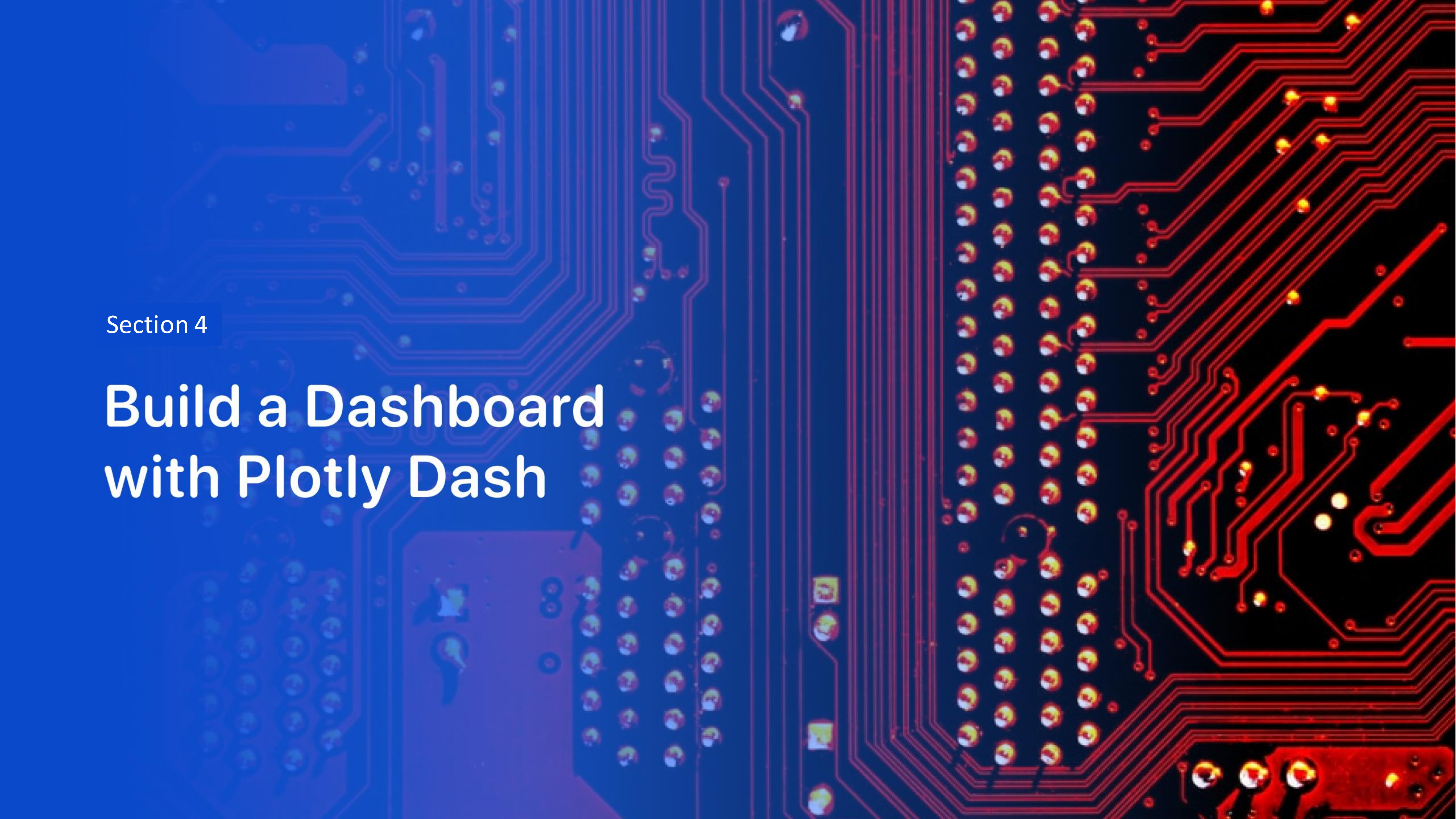
- It was found that launch sites are located close to railways, highways and not very far from the coastline. However, there is a certain distance between launch sites location and cities.
- An example bellow (for KSC LC-39A launch site) shows that the distance to the closest city is 16.5 km, while the distance to the railway and highway is < 1 km, and distance to the closest coastline is 2 km.




Lat: 28.57531 Long: -80.65547
+
−
P
Kennedy Parkway North
Contractors Rd
0.84 KM
0.70 KM
KSC LC-39A
Leaflet | Data by © OpenStreetMap, under ODbL.

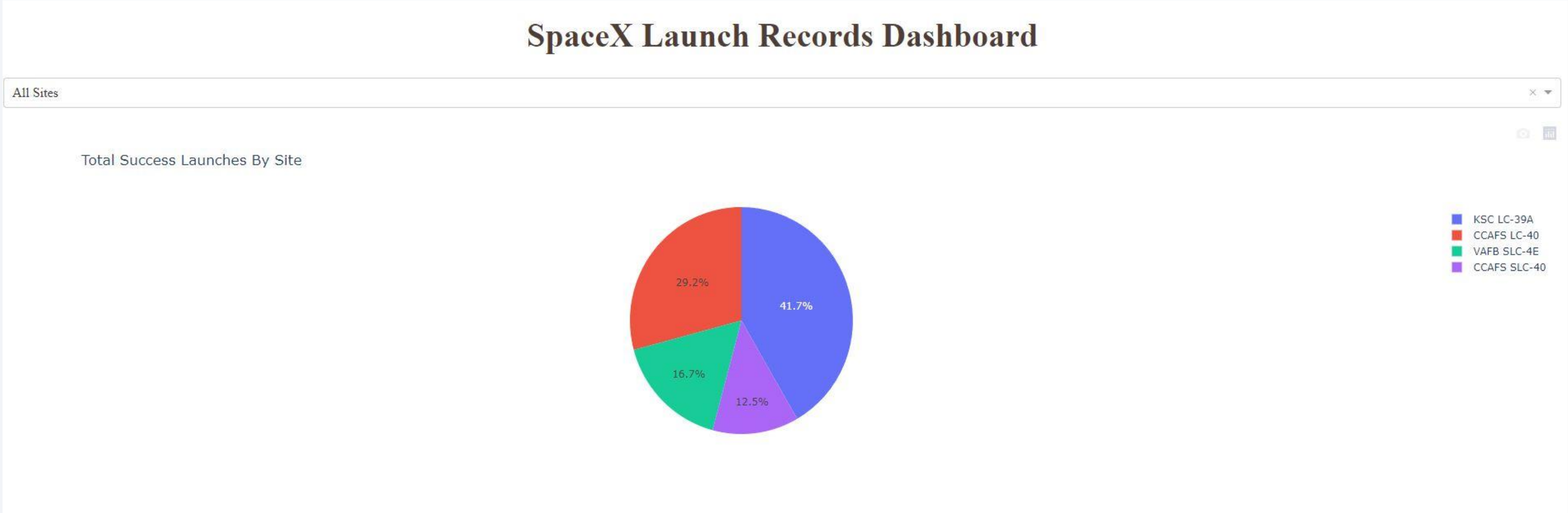
Section 4

Build a Dashboard with Plotly Dash



SpaceX launch success records by site

- The pie chart below indicates that the largest successful launches has been observed in KSC LC-39A launch site. The least success rates are seen in VAFB SLC-4E and CCAFS SLC-40 sites.

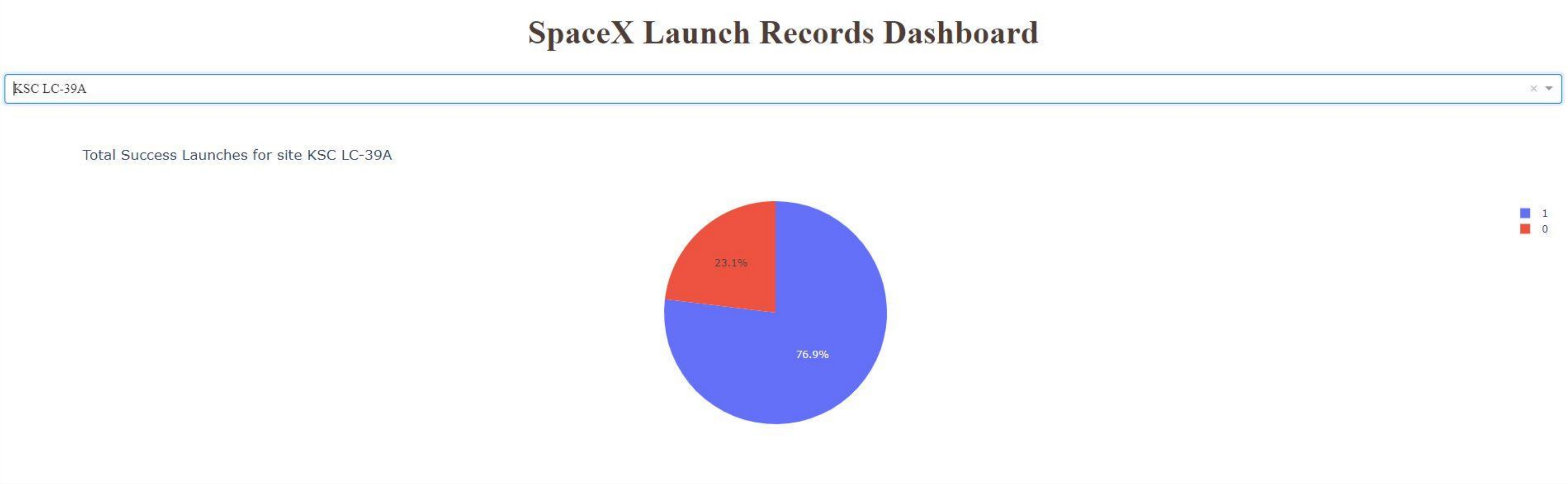


SpaceX Launch Records Dashboard

All Sites

Launch success/failure for site KSC LC-39A

- The pie chart below shows the launch site (KSC LC-39A) with the highest launch success ratio (76.9%).



SpaceX Launch Records Dashboard

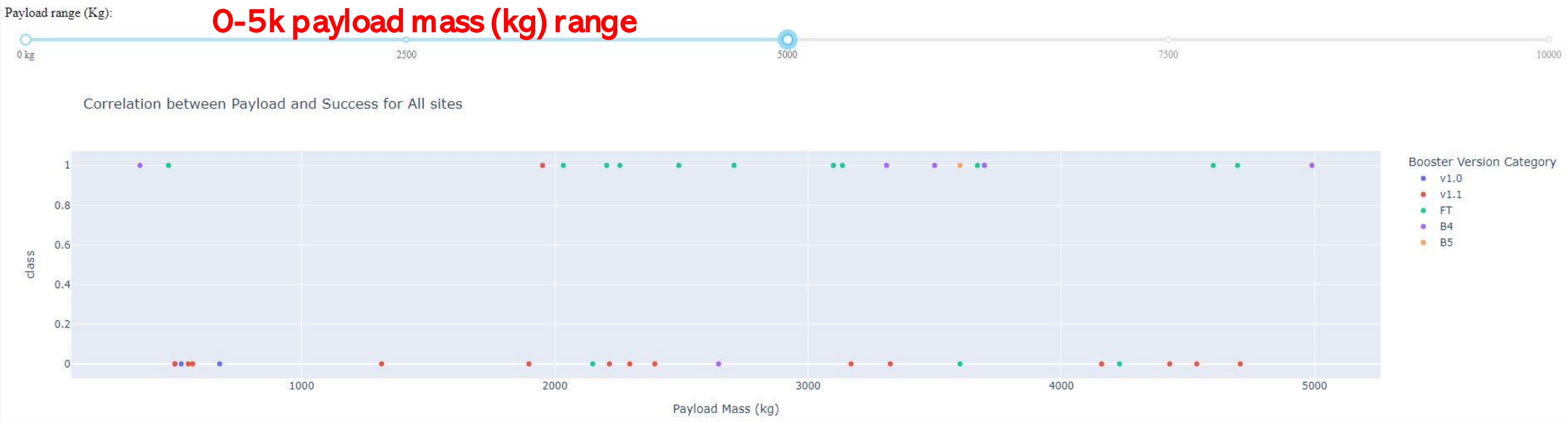
KSC LC-39A

Total Success Launches for site KSC LC-39A

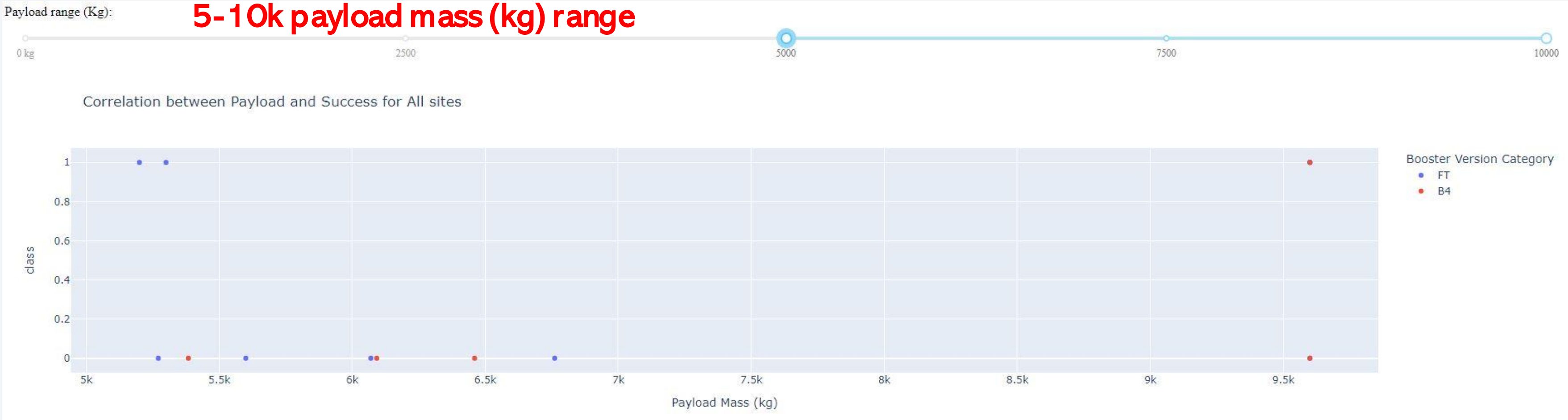
Correlation between payload and launch outcome

- From the comparison of different payload ranges (i. e. 0-5k kg and 5-10k kg) vs. launch outcomes, it was found that a lower weighted payload mass has higher success rates compared to heavier payload mass (47% and 27% respectively).

- Overall, booster versions B5 and FT had the highest launch success rates (100 % and 65% respectively).



0-5k payload mass (kg) range



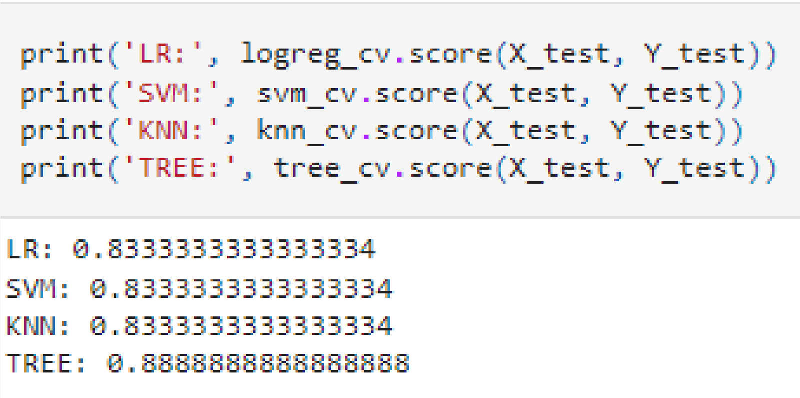
5-10k payload mass (kg) range

Section 5

Predictive Analysis (Classification)

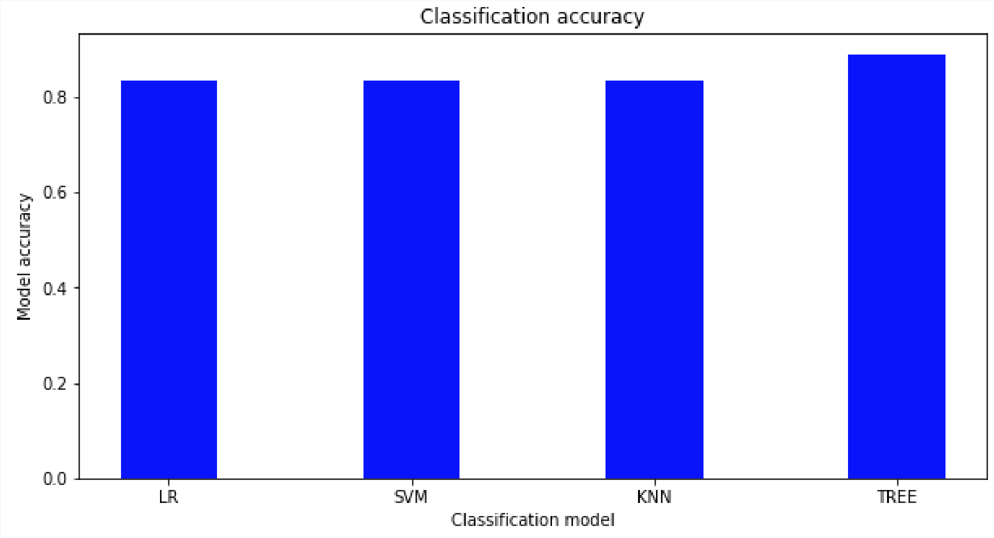
Classification Accuracy

- After identifying the best hyperparameter and calculating the accuracy of each method, it was found that the best performing method using test data was Classification Trees (accuracy rate = 0.89 vs 0.83 for other methods).



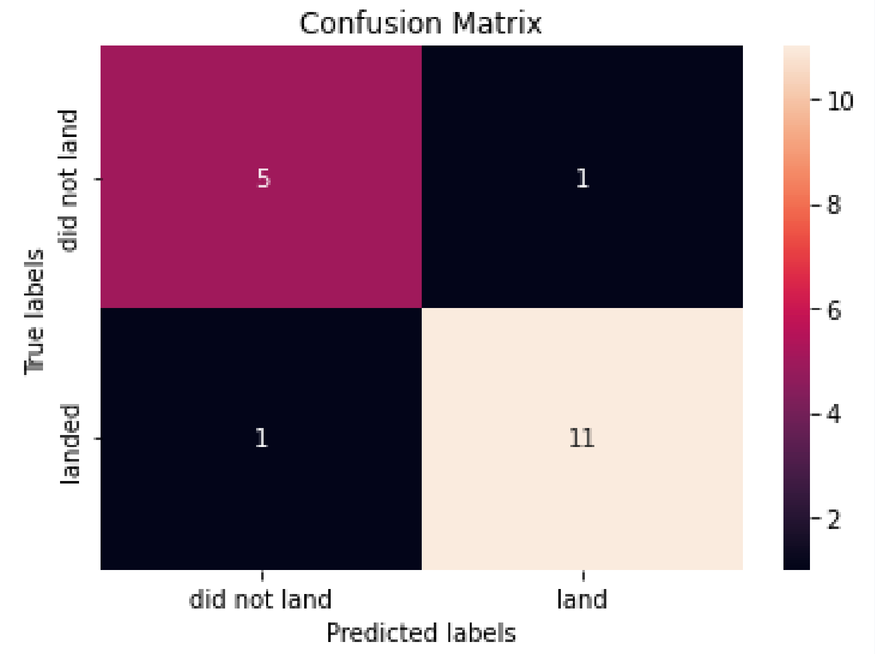
```
print('LR:', logreg_cv.score(X_test, Y_test))
print('SVM:', svm_cv.score(X_test, Y_test))
print('KNN:', knn_cv.score(X_test, Y_test))
print('TREE:', tree_cv.score(X_test, Y_test))
```

```
LR: 0.8333333333333334
SVM: 0.8333333333333334
KNN: 0.8333333333333334
TREE: 0.8888888888888888
```



Confusion Matrix

- From the confusion matrix of Classification Trees method it is seen that 16 out of 18 cases have been classified correctly (accuracy rate = 0.89).



Conclusions

- From EDA analysis it was found that ES-L1, GEO, HEO and SSO orbits have highest success rates (100%).
- Overall, it was seen that a lower weighted payload mass has higher success rates compared to heavier payload mass (47% and 27% respectively), however, there also seems to be a dependency on the orbit, as more successful landings were observed with heavier payloads in LEO, ISS, PO orbits, however, no such clear pattern was observed in GTO orbit.
- From the folium maps, also Plotly dash, it was found that KSC LC-39A site has highest success rates. This can be explained by factors: this site had only higher flight numbers, as well majority of the booster versions in the side were B5 and FT which had the highest launch success rates (overall, 100 % and 65% respectively). However, additional data is needed to confirm this.
- Results of the estimated Classification models showed that the best performing model was Classification Trees with highest accuracy rate (0.89).

Thank you!