IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Iuliia Shal
25/05/2022

# Outline

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

# Executive Summary

- Summary of methodologies:
  - Data collection and wrangling using SpaceX API and Wikipedia Web Scrapping
  - Exploratory Data analysis
  - Interactive visualization (map with Folium and dashboard with Plotly Dash)
  - Predictive analysis (classification models)

- Summary of all results
  - Exploratory data analysis results - Section 2
  - Interactive analytics demo in screenshots - Section 3
  - Predictive analysis results - Section 4

# Introduction

- **Context**

As a new provider in Space Transportation, we would like to beat SpaceX.

But how can we win the competition with the cheapest provider in the market (on average 3 times cheaper than others)?

SpaceX saved that much money due to reuse of the first stage of the rocket.

- **Goal**

Predict if the Falcon 9 first stage will land successfully and understand what are key factors of success

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - SpaceX API
  - Wikipedia page Web Scraping
- Perform data wrangling
  - NAs eliminated, Data types check, Dependent variable (Y) created
- 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

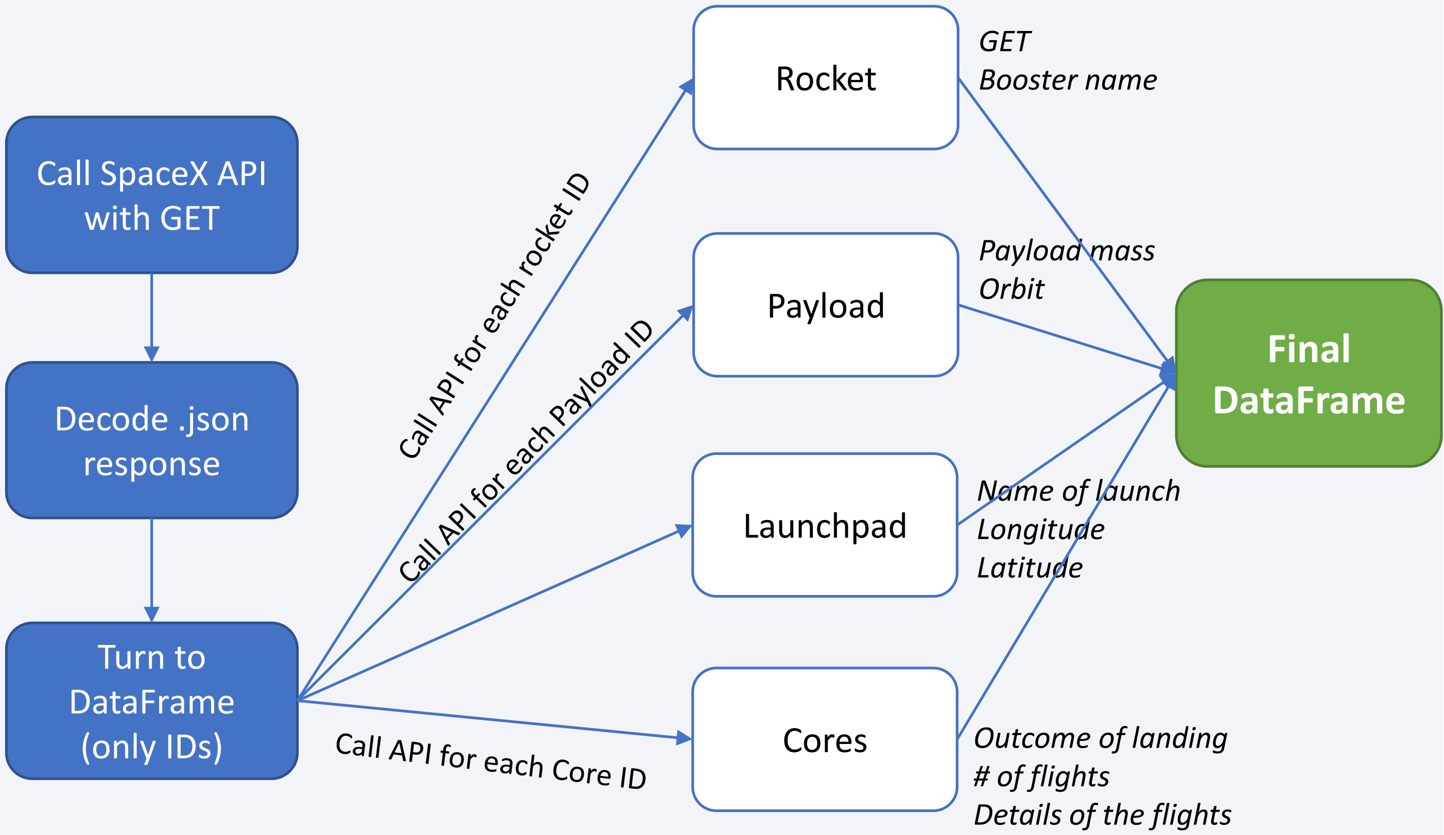
- Data for this project were collected from 2 sources:
  - Official SpaceX REST API . Please see GitHub page for more details
  - Wikipedia page

# Data Collection – SpaceX API

- To collect data from SpaceX API we first collect the roster of all launches, but only with IDs in columns

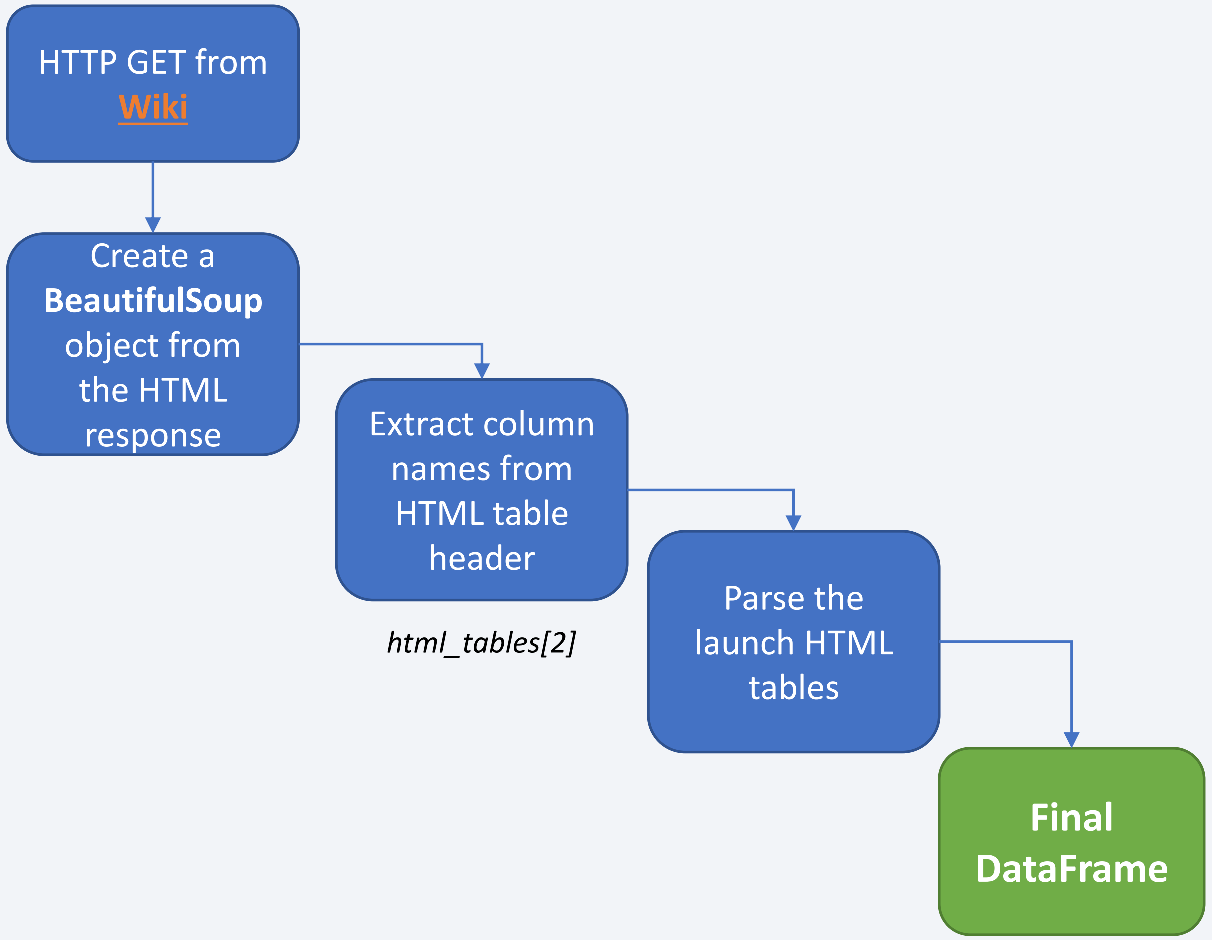
- Next we restore all missing descriptive columns by calling API for each ID (example, requests.get("https://api.spacexdata.com/v4/**rockets/rocket_id**).json()

- [GitHub URL]()



# Data Collection - Scraping

- Web Scrapping HTML table from [Wiki page] to get data about Falcon 9 and Falcon Heavy launches
- Parse HTML tables and convert it to the DataFrame
- GitHub URL



# Data Wrangling

- After getting raw data it should be preprocessed for exploratory data analysis and modeling in future.
- Key procedures applied:
  - NULLs check
  - Data types check
  - # records per launch site, orbit and landing outcome
  - Y variable/Class created to determine Success/Failure of launch (dependent variable)

- GitHub URL

# EDA with Data Visualization

- After Data Wrangling stage we’re ready to play with data.
- There are different questions we wanted to answer with this step:
  - Relationship between Payload Mass and FlightNumber (by launch sites) to see if anything changes with more tests and technology development
  - Relationship between FlightNumber and Launch site
  - Relationship between PayloadMass and Launch site to understand if there are difference success rate between sites with different Payload Mass
  - Success rate by Orbit type to see how far into space we can launch the rocket successfully
  - Relationship between PayloadMass and Orbit type to see how far into space we can launch the rocket with a heavy mass successfully
  - Yearly trend of launch success to see the progress in time



- GitHub URL

# EDA with SQL

To check the data structure and aggregated statistics we used the following SQL queries:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'KSC’
- 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 where the first successful landing outcome in drone ship was achieved
- List the names of the boosters which have success in ground pad 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
- List the records which will display the month names, successful landing outcomes in ground pad ,booster versions, launch site for the months in year 2017
- Rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.



[GitHub URL]

# Build an Interactive Map with Folium

- 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 which we are going to find by analyzing the existing launch site locations.
- Folium will help us create a Map where we mark:
  - All launch sites (Circle and Marker with name to easily identify these locations)
  - Success/failed launches for each site as Cluster Marker to visually assess success rate
  - Distances between a launch site to its proximities (closest city, highway, railway, coast) as line with marker (km) to understand strategic positions of these sites
- GitHub URL

# Build a Dashboard with Plotly Dash

- Building an interactive dashboard and charts is a very convenient and easy way to present data to potential stakeholders.
- With Plotly Dash we showed:
  - Pie chart of success launches by sites together with success rate for each site to get a quick answer about most/least successful site
  - Scatter plot of launch outcome by Payload Mass and Booster version to visualize if launch outcome depends on there 2 variables

GitHub URL

# Predictive Analysis (Classification)

The following steps were performed to do a high-quality predictive analysis:

1. Create a column for the Launch Outcome **(Y)**: Success (1) / Failure (0)
2. Standardize the data using **StandardScaler()**
3. Split into training data and test data
4. Find best Hyperparameter for SVM, Classification Trees and Logistic Regression using **GridSearchCV** using **train** data
5. Find the method performs best using **test** data and calculating **Score** and **Confusion Matrix**

GitHub URL

# Results

- Exploratory data analysis results - Section 2
- Interactive analytics demo in screenshots - Section 3
- Predictive analysis results - Section 4

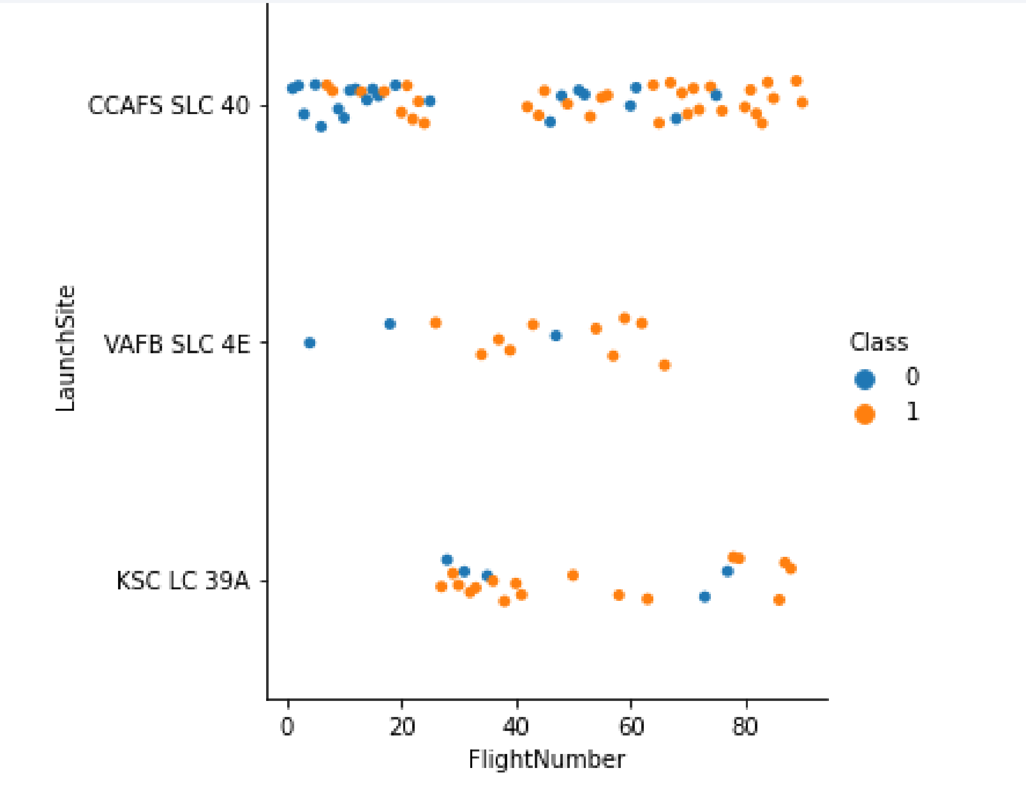
Section 2

# Insights drawn from EDA

# Flight Number vs. Launch Site

- First 20 attempts made mainly from Cape Canaveral were mostly not successful (try and error)

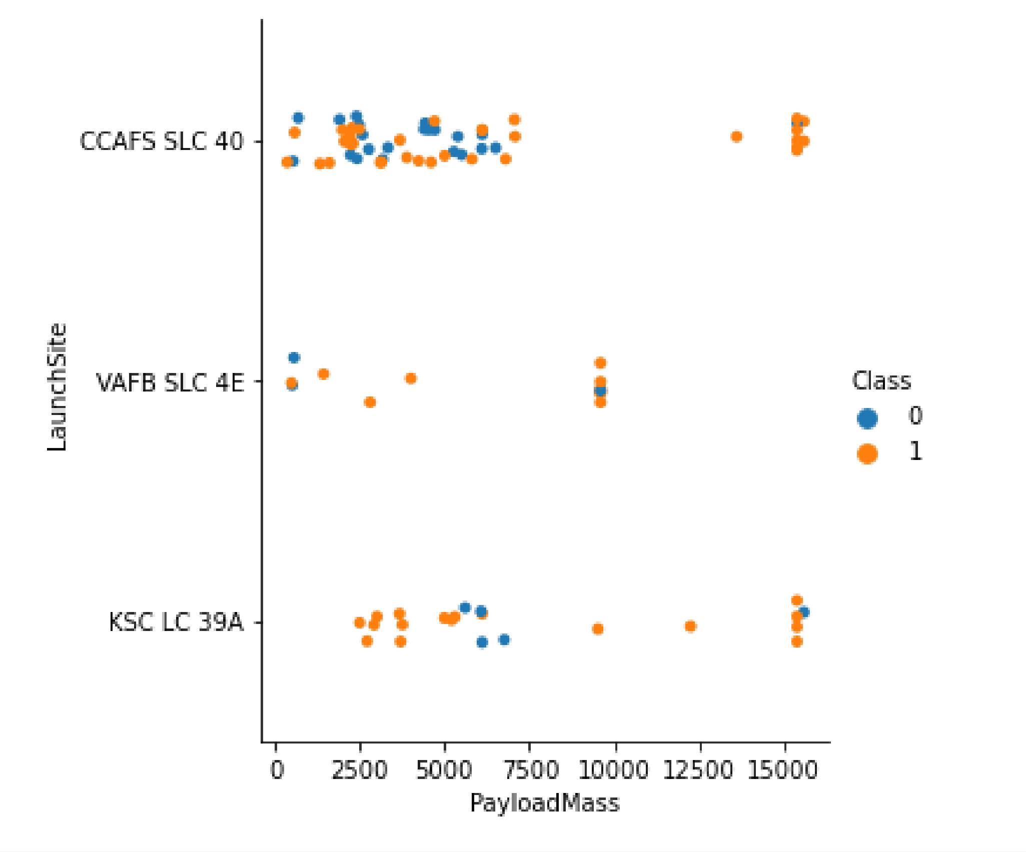
- But after 21-25th launch we can observe quite high success rate in all sites



# Payload vs. Launch Site

- CCAFS site doesn't show good stable result in payload range 0-7,500kg. However, it succeeds with high loads > 12,500 kg

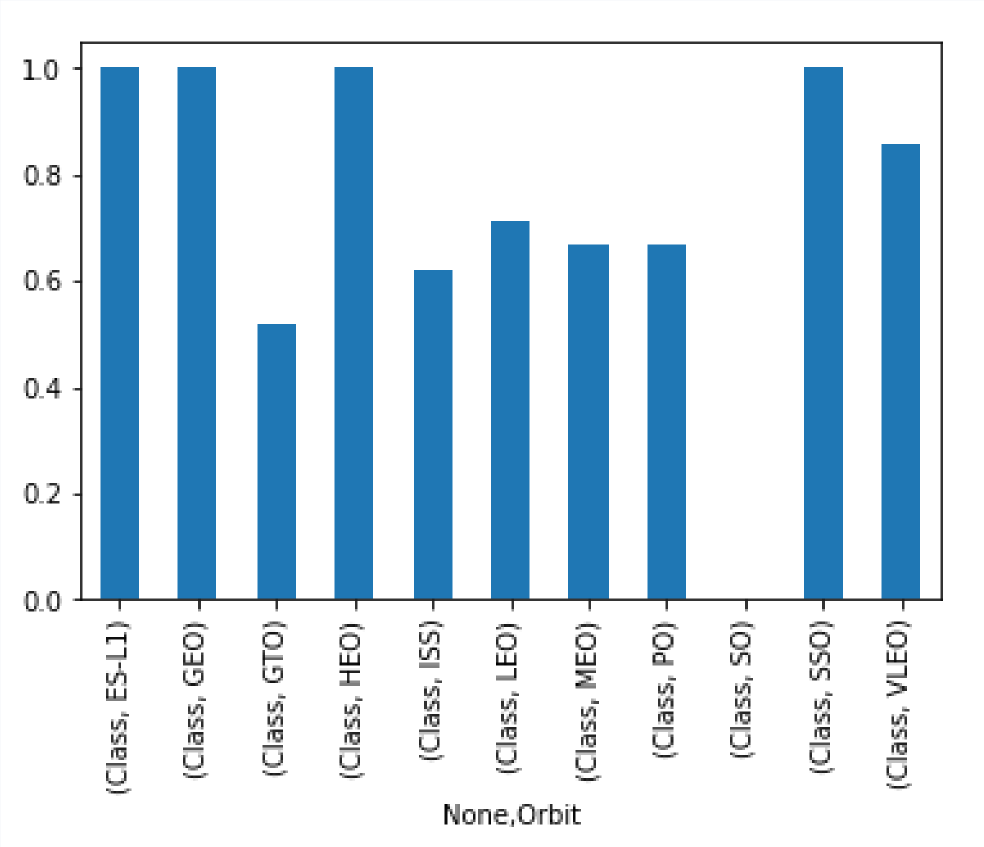
- Meanwhile, KSC LC showed 100% success rate for lower payloads (2,500-5,000 kg) as well as high ones (>10,000 kg)



# Success Rate vs. Orbit Type

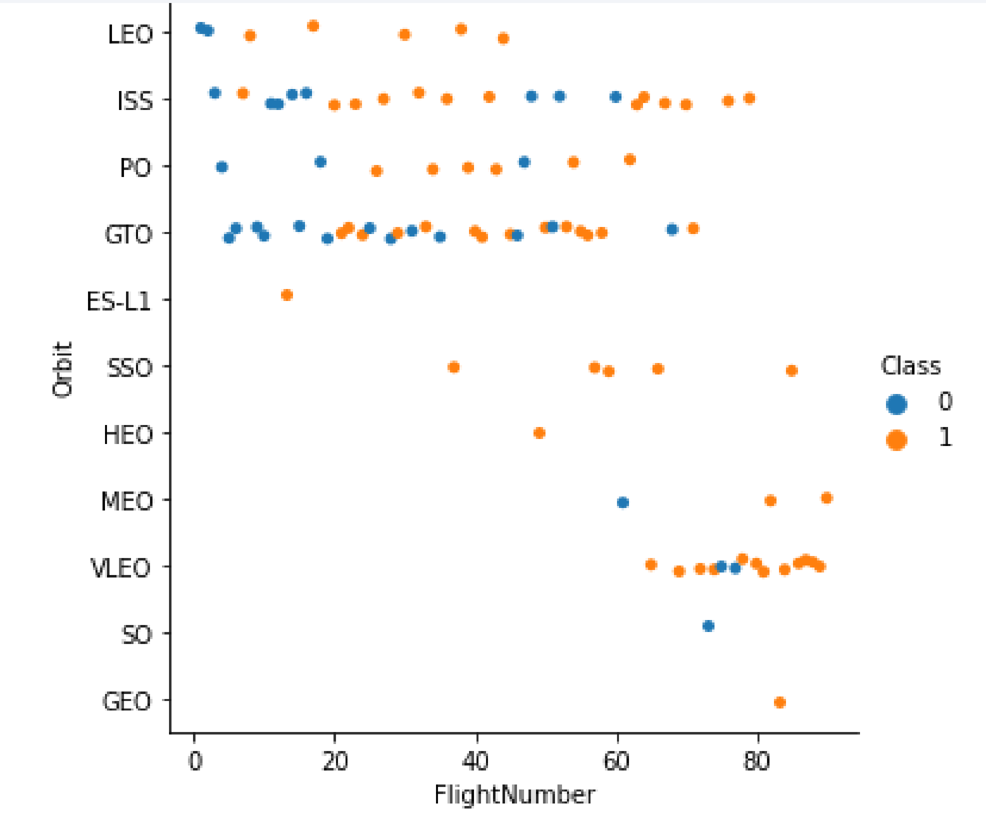
- Highest success rate is observed for orbits:
  - ES-L1
  - GEO (a circular geosynchronous orbit 35,786 kilometers (22,236 miles) above Earth's equator)
  - HEO (Geocentric orbits above the altitude of geosynchronous orbit (35,786 km or 22,236 miles))
  - SSO (Sun-synchronous orbit, a nearly polar orbit around a planet)

- Lowest success rate is observed for orbits:
  - SO (error in the db – should have been combined with SSO as it's the same)
  - GTO
  - ISS (International Space Station)
  - MEO (most commonly at 20,200 kilometers (12,600 mi) or 20,650 kilometers (12,830 mi))
  - PO (poles orbit)
  - LEO (Low Earth orbit is an Earth-centred orbit with an altitude of 2,000 km or lower)



# Flight Number vs. Orbit Type

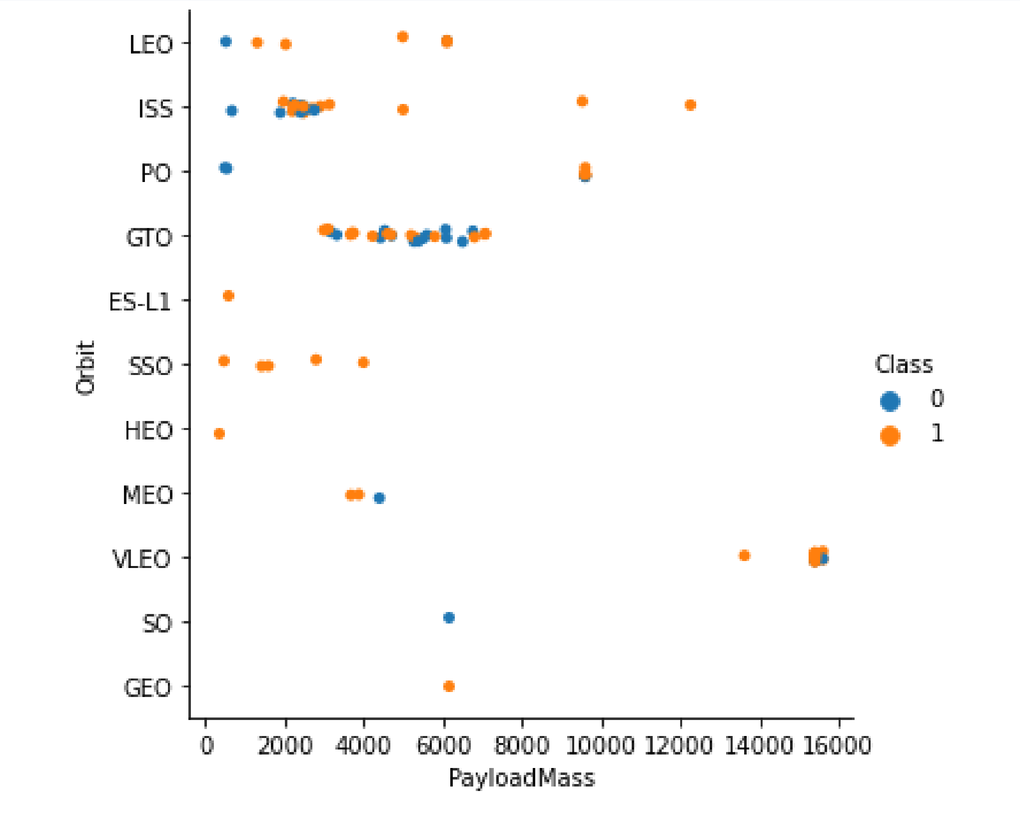
- First attempts on low orbit types were not successful
- However, with more tries and errors almost all orbits launches became successful
- Exception:
  - GTO has many launches and still around 50% of success rate



# Payload vs. Orbit Type

- Very heavy payloads were launched only to low orbits (and successfully):
  - VLEO (very low earth orbit)
  - ISS (international Space Station)
  - PO (poles orbit)

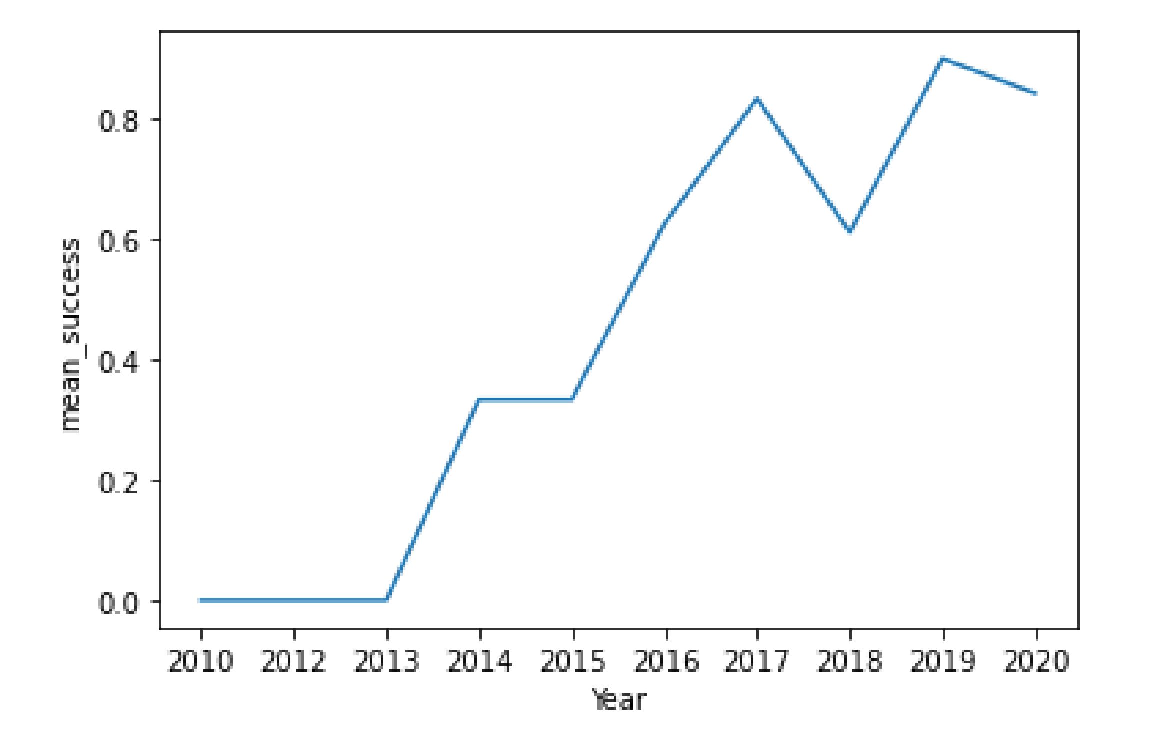
- Payloads <5,000 kg were effectively launch in 100% of cases to SSO, HEO and ES-L1



# Launch Success Yearly Trend

With development of new booster versions and space technologies, the success rate has been gradually increasing and has achieved

**> 80%** in 2019



# All Launch Site Names

- There are 4 launch sites:
    - 3 on East Coast  (CCAFS x2 and KSC)
    - 1 on West Coast of US (VAFB SLC-4E)

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

# Launch Site Names Begin with 'KSC'

- Below we can see first 5 records of the launch site KSC LC-39A, the site with the highest success rate.

| DATE | time__utc_ | booster_version | launch_site | payload | payload_mass__kg_ | orbit | customer | mission_outcome | landing__outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2017-02-19 | 14:39:00 | F9 FT B1031.1 | KSC LC-39A | SpaceX CRS-10 | 2490 | LEO (ISS) | NASA (CRS) | Success | Success (ground pad) |
| 2017-03-16 | 06:00:00 | F9 FT B1030 | KSC LC-39A | EchoStar 23 | 5600 | GTO | EchoStar | Success | No attempt |
| 2017-03-30 | 22:27:00 | F9 FT B1021.2 | KSC LC-39A | SES-10 | 5300 | GTO | SES | Success | Success (drone ship) |
| 2017-05-01 | 11:15:00 | F9 FT B1032.1 | KSC LC-39A | NROL-76 | 5300 | LEO | NRO | Success | Success (ground pad) |
| 2017-05-15 | 23:21:00 | F9 FT B1034 | KSC LC-39A | Inmarsat-5 F4 | 6070 | GTO | Inmarsat | Success | No attempt |

# Total Payload Mass

- In total rockets carried over 45,000 kg from NASA

| total_payload |
| --- |
| 45596 |

# Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 equals to 2,534 KG

| avg_payload |
| --- |
| 2534 |

# First Successful Ground Landing Date

- The first successful landing outcome on ground pad happened on 8 Apr 2016

2016-04-08

# Successful Drone Ship Landing with Payload between 4000 and 6000

- The following 3 boosters have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

| booster_version |
| --- |
| F9 B4 B1040.1 |
| F9 B4 B1043.1 |
| F9 FT B1032.1 |

# Total Number of Successful and Failure Mission Outcomes

- Almost all mission were successful

| mission_outcome | 2 |
| --- | --- |
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

# Boosters Carried Maximum Payload

- The following booster versions carried the maximum payload mass
- All have version F9 B5…

| 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 |

# 2017 Launch Records

- In 2017 most of successful landing on ground pad happened from KSC LC-39A launch site

| month_name | landing__outcome | booster_version | launch_site |
|---|---|---|---|
| February | Success (ground pad) | F9 FT B1031.1 | KSC LC-39A |
| May | Success (ground pad) | F9 FT B1032.1 | KSC LC-39A |
| June | Success (ground pad) | F9 FT B1035.1 | KSC LC-39A |
| August | Success (ground pad) | F9 B4 B1039.1 | KSC LC-39A |
| September | Success (ground pad) | F9 B4 B1040.1 | KSC LC-39A |
| December | Success (ground pad) | F9 FT B1035.2 | CCAFS SLC-40 |

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

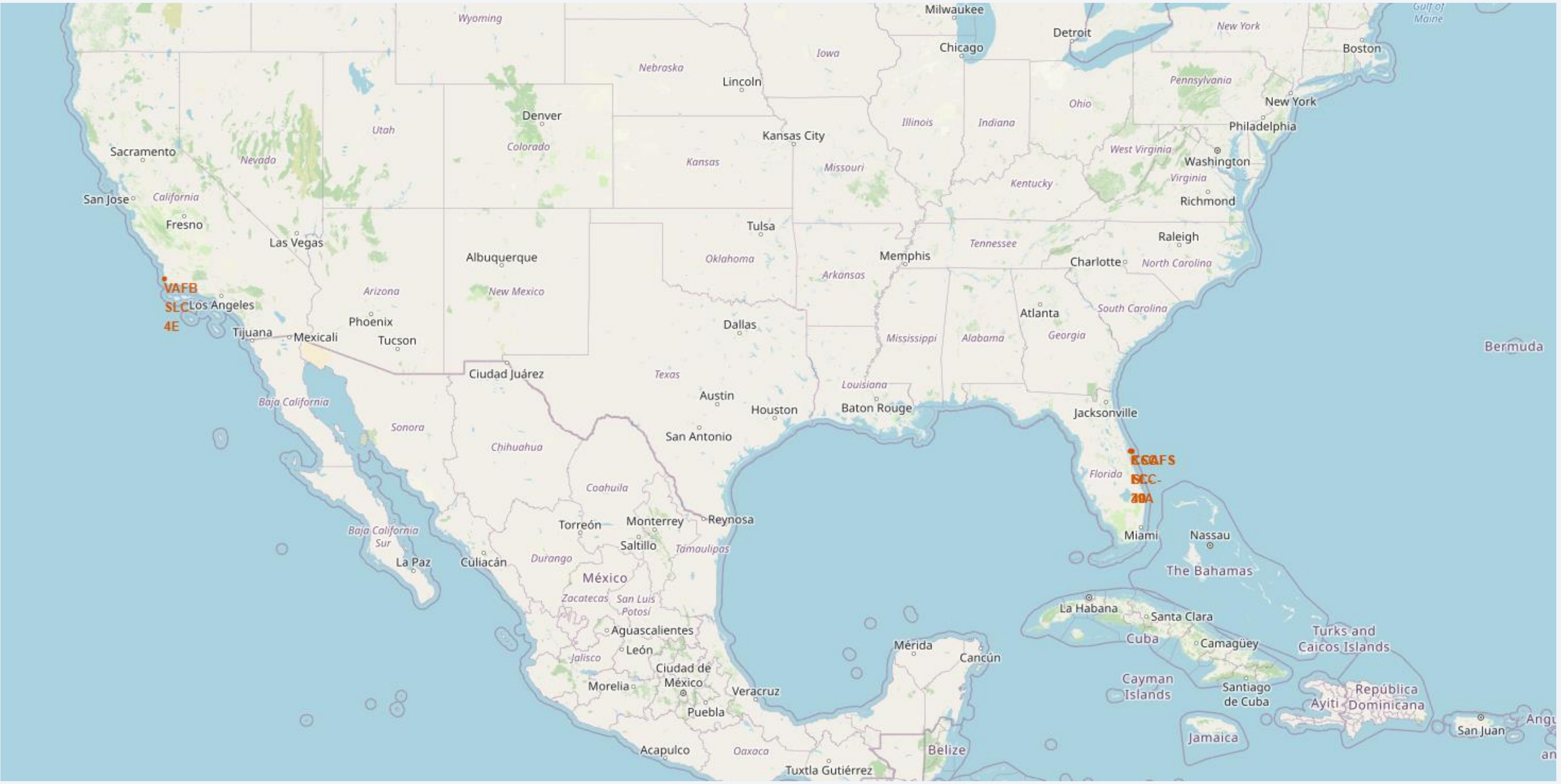
- There were only 8 successful landing outcomes between the date 2010-06-04 and 2017-03-20
  - 5 on drone ship
  - 3 on ground pad

| landing_outcome | cn |
|---|---|
| Success (drone ship) | 5 |
| Success (ground pad) | 3 |

Section 3

# Launch Sites Proximities Analysis

# Launch sites' location

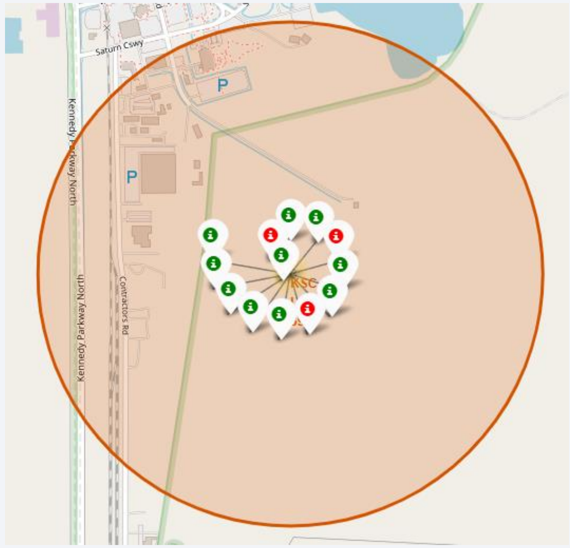



- All launch sites are very close to the coast
- All launch sites are as close to the equator as possible on the US territory

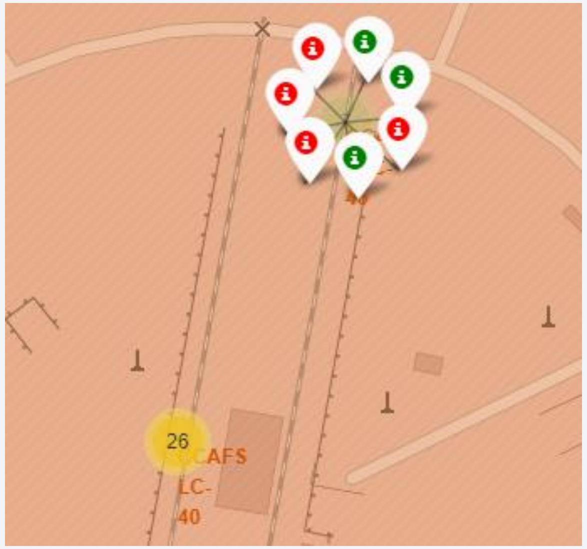
# Launch outcomes

VAFB SLC-4E (West)

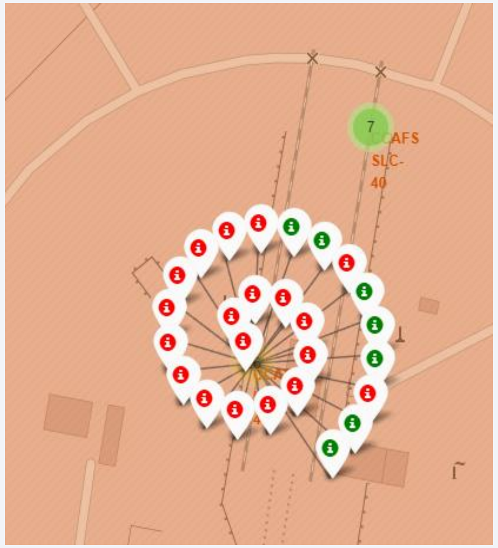

KSC LC-39A (East)




CCAFS SLC-40 (East)



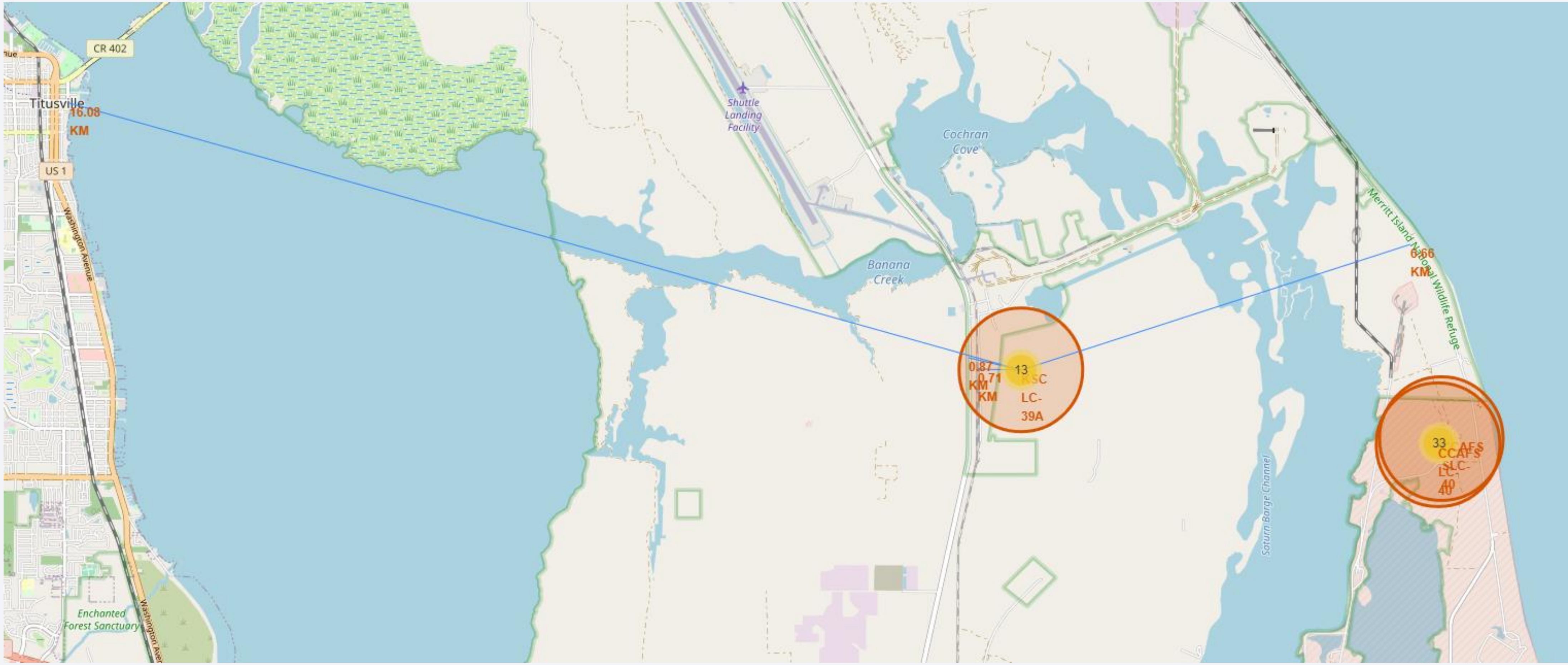

CCAFS LC-40 (East)




- KSC LC-39A launch site has the highest success rate among all sites
- Interesting that it’s the only site which is located further from the ocean coast than other sites (by almost 7 km)

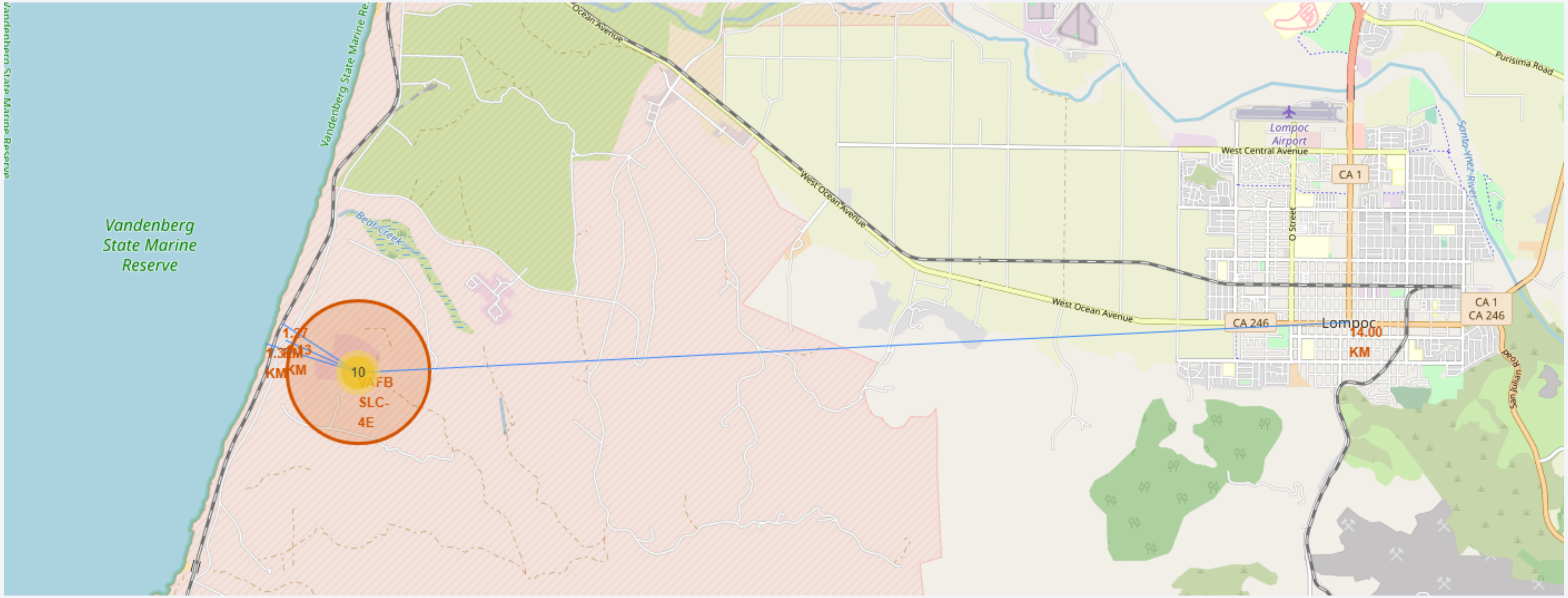
# Site location and its proximities

KSC LC-39A (East)



# Site location and its proximities

VAFB SLC-4E (West)



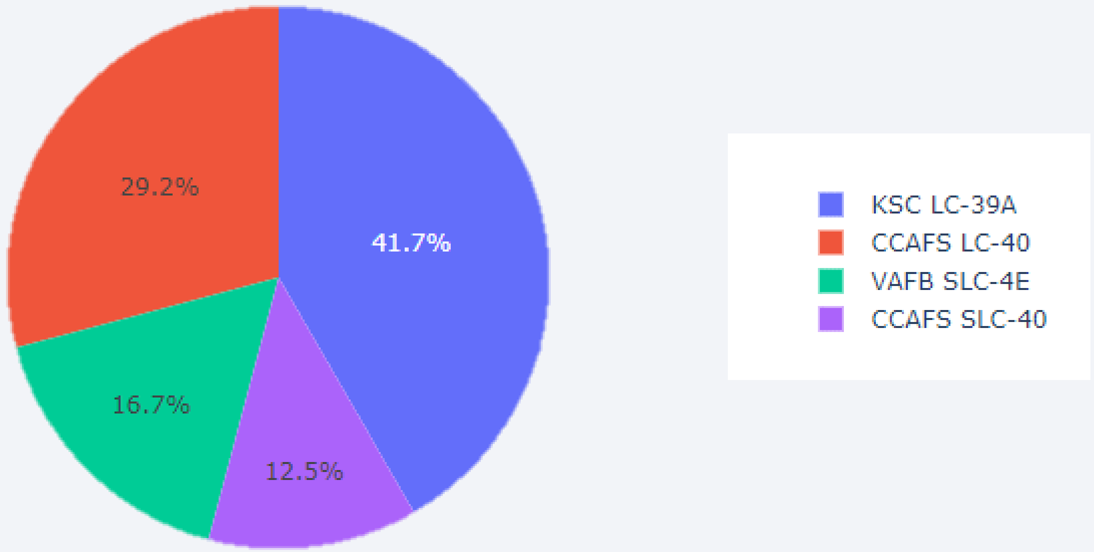

- All launch sites are in close distance to the railway and highway (0.5-1.5 km), 3 sites - to the coast as well (except KSC LC-39A as mentioned previously)
- However, they are quite far from the nearest city (>10 km)

Section 4

# Build a Dashboard with Plotly Dash

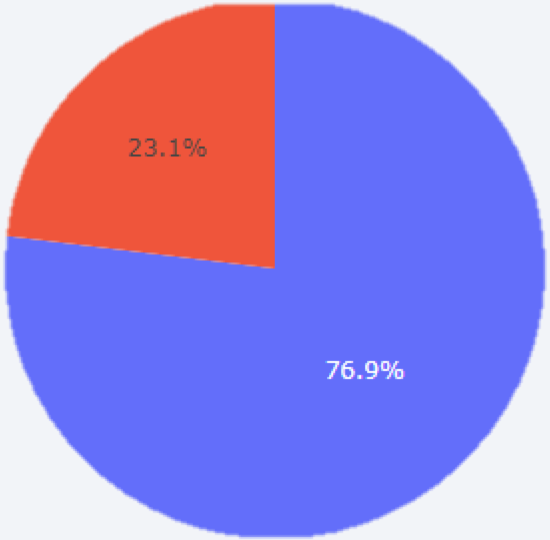
# Launch success by sites

- KSC LC-39A has the highest number of success launches (41.7% out of all successes)
- 2 place is for CCAFS LC-40 (29.2%)



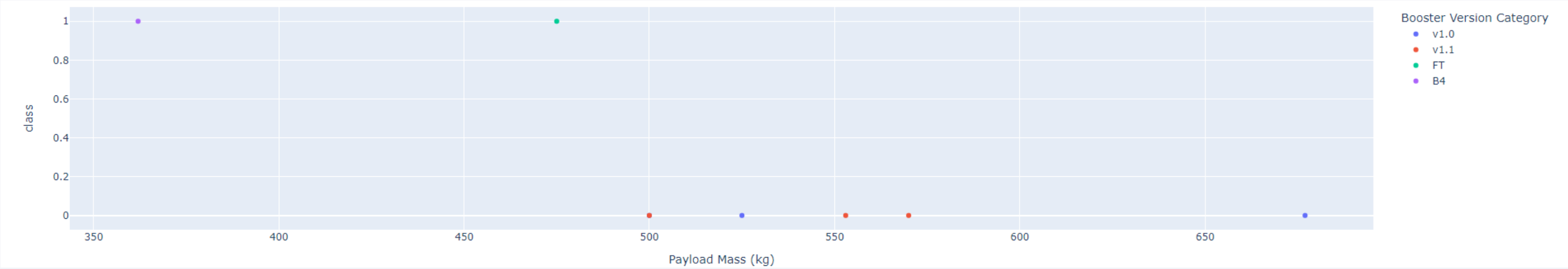
# Launch success rate – KSC LC-39A

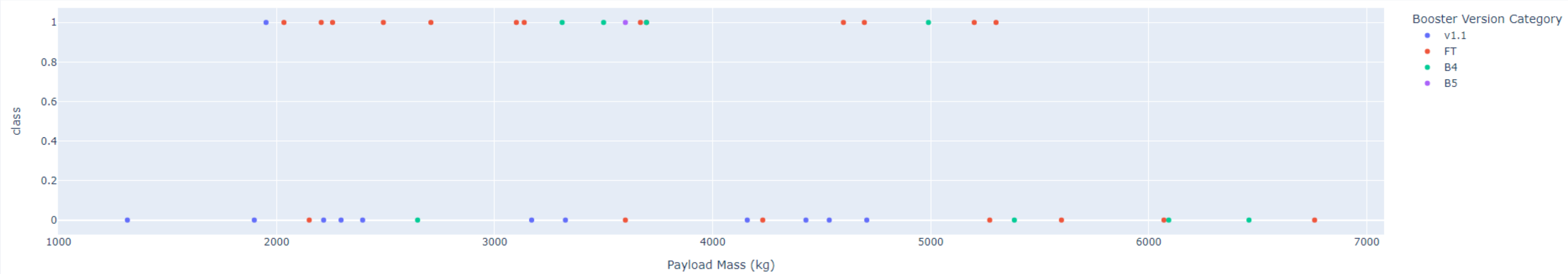
As mentioned previously, KSC LC-39A has also the highest success rate – **76.9%**



# Payload vs Launch Outcome

- Payloads < 1,000 kg and > 6,000 kg has the lowest launch success rate
- Payloads between 2,000 kg and 4,000 kg has the highest launch success rate
- Booster version v1.1 has the lowest success rate in the whole payload range
- Booster version FT has the highest success rate but only for 2,000-5000 kg payload



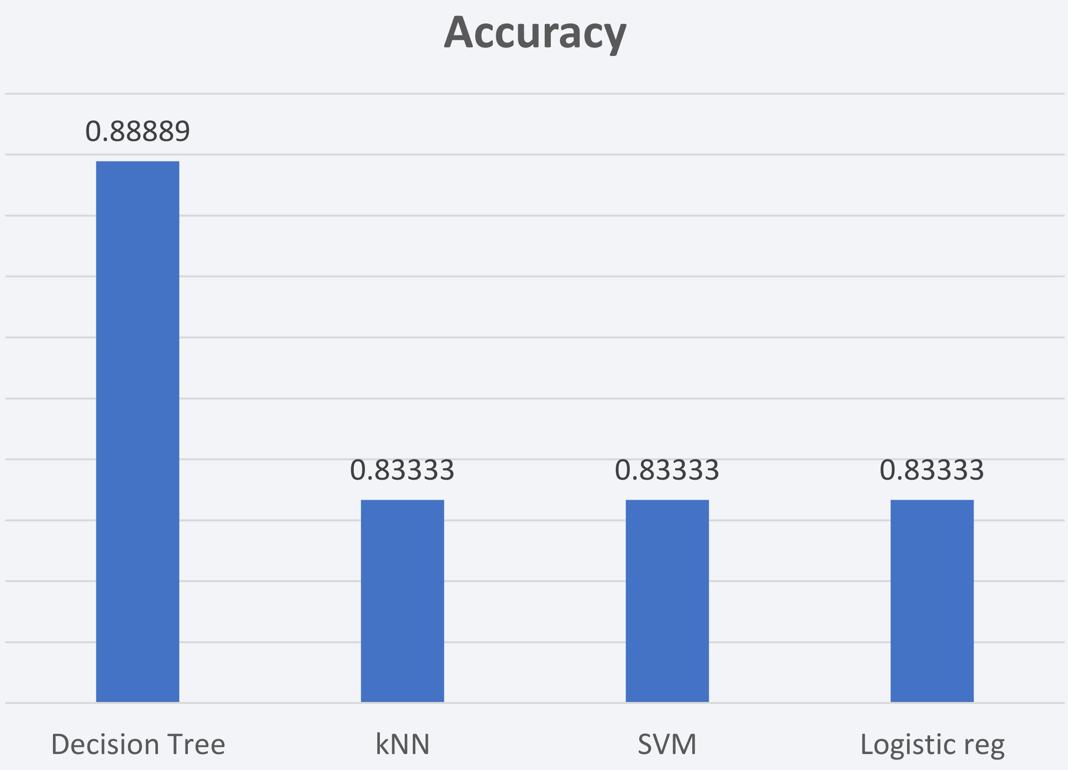



Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

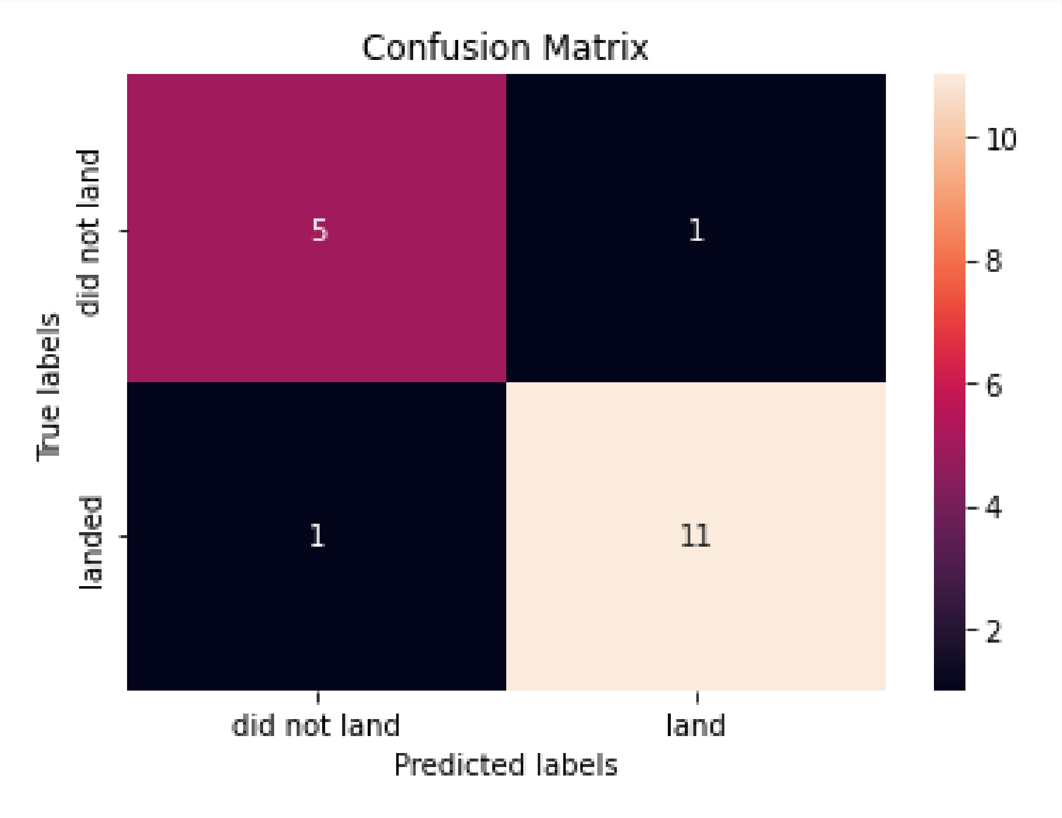
- Decision Tree model showed the highest accuracy amongst other 3 models tested (kNN, SVM and Logistic regression)
- Best parameters chosen for the model are:
    - Criterion: entropy
    - Max Depth: 12
    - Min Samples leaf: 4
    - Min Sample split: 10



# Confusion Matrix – Decision Tree

- On the graph below we can see that the model made only 2 mistakes:
  - 1 case which landed as didn't land
  - 1 case – vice verse.

- 11 out of 12 landed and 5 out of 6 didn't land were predicted accurately

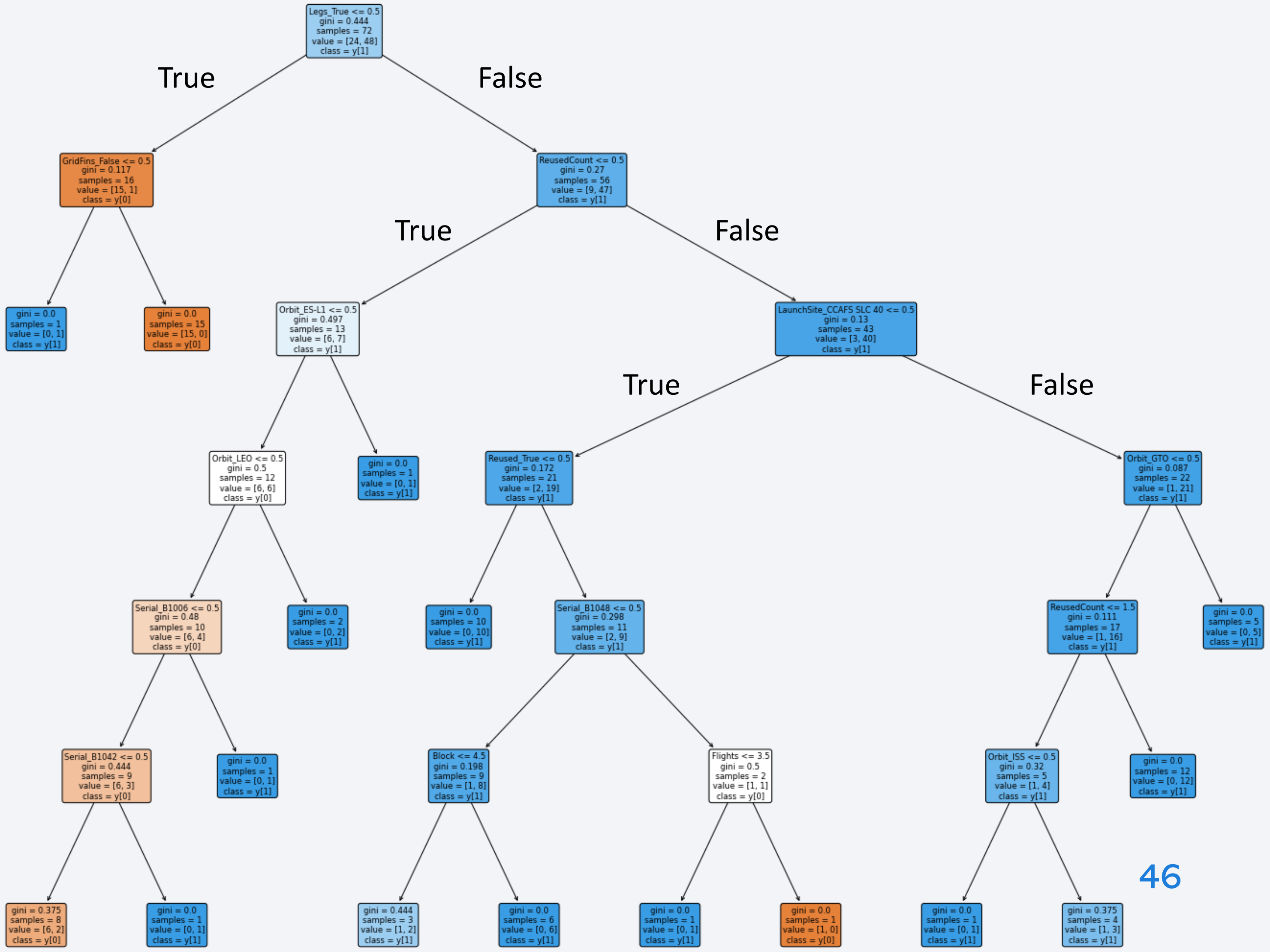


# Decision Tree

- In this illustration the analysis was done on the original, **not Scaled** Data as it'd be hard to interpret the results of the model with relative values

- The highest level of the tree says that if legs and grid fins were not used during landing (<=0.5) , then the rocket will crash.

- If we go to False branch, we can have a success if:
  - using legs during landing
  - The rocket is not re-used
  - Orbit is either LEO or ES-L1



# Conclusions

To successfully launch a new rocket and reuse the first stage with highest probability we need:

- Set up site closer to equator and not too close to the coast (around 6 km)
- Rocket must use legs during landing
- Choose closer orbit such
- Rocket should be re-used no more than 3 times

# Appendix 1.1 – SQL used for analysis

## Task 1

Display the names of the unique launch sites in the space mission

In [14]:
```
%sql select distinct launch_site from spacextbl
```

```
 * ibm_db_sa://bpp32620:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[14]:

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

## Task 2

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

In [15]:
```
%sql select * from spacextbl where launch_site like 'KSC%' limit 5
```

```
 * ibm_db_sa://bpp32620:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[15]:

| DATE | time__utc_ | booster_version | launch_site | payload | payload_mass__kg_ | orbit | customer | mission_outcome | landing__outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2017-02-19 | 14:39:00 | F9 FT B1031.1 | KSC LC-39A | SpaceX CRS-10 | 2490 | LEO (ISS) | NASA (CRS) | Success | Success (ground pad) |
| 2017-03-16 | 06:00:00 | F9 FT B1030 | KSC LC-39A | EchoStar 23 | 5600 | GTO | EchoStar | Success | No attempt |
| 2017-03-30 | 22:27:00 | F9 FT B1021.2 | KSC LC-39A | SES-10 | 5300 | GTO | SES | Success | Success (drone ship) |
| 2017-05-01 | 11:15:00 | F9 FT B1032.1 | KSC LC-39A | NROL-76 | 5300 | LEO | NRO | Success | Success (ground pad) |
| 2017-05-15 | 23:21:00 | F9 FT B1034 | KSC LC-39A | Inmarsat-5 F4 | 6070 | GTO | Inmarsat | Success | No attempt |

# Appendix 1.2 – SQL used for analysis

## Task 3

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

In [20]:
```
%sql select sum(payload_mass__kg_) as total_payload from spacextbl where customer='NASA (CRS)'
```

```
 * ibm_db_sa://bpp32620:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[20]:

| total_payload |
|---|
| 45596 |

## Task 4

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

In [23]:
```
%sql select avg(payload_mass__kg_) as avg_payload from spacextbl where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://bpp32620:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[23]:

| avg_payload |
|---|
| 2534 |

## Task 5

List the date where the first succesful landing outcome in drone ship was acheived.

*Hint:Use min function*

In [25]:
```
%sql select min(DATE) from spacextbl where landing__outcome='Success (drone ship)'
```

```
 * ibm_db_sa://bpp32620:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[25]:

| 1 |
|---|
| 2016-04-08 |

# Appendix 1.3 – SQL used for analysis

### Task 6

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

In [27]:
```
%sql select distinct booster_version from spacextbl where landing__outcome='Success (ground pad)' and payload_mass__kg_>4000 and payload_mass__kg_<600
```

```
 * ibm_db_sa://bpp32620:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[27]:

| booster_version |
|---|
| F9 B4 B1040.1 |
| F9 B4 B1043.1 |
| F9 FT B1032.1 |

### Task 7

List the total number of successful and failure mission outcomes

In [29]:
```
%sql select mission_outcome,count(*) from spacextbl group by mission_outcome
```

```
 * ibm_db_sa://bpp32620:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[29]:

| mission_outcome | 2 |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

### Task 8

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

In [32]:
```
%sql select booster_version from spacextbl s,  (select max(payload_mass__kg_) as max_payload from spacextbl)m where s.payload_mass__kg_=m.max_payload
```

```
 * ibm_db_sa://bpp32620:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

# Appendix 1.4 – SQL used for analysis

## Task 9

List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

In [46]:
```
%sql select sysfun.monthname(DATE) as month_NAME, landing__outcome,booster_version, launch_site from spacextbl where landing__outcome like 'Success (g
```

```
 * ibm_db_sa://bpp32620:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[46]:

| month_name | landing__outcome | booster_version | launch_site |
|---|---|---|---|
| February | Success (ground pad) | F9 FT B1031.1 | KSC LC-39A |
| May | Success (ground pad) | F9 FT B1032.1 | KSC LC-39A |
| June | Success (ground pad) | F9 FT B1035.1 | KSC LC-39A |
| August | Success (ground pad) | F9 B4 B1039.1 | KSC LC-39A |
| September | Success (ground pad) | F9 B4 B1040.1 | KSC LC-39A |
| December | Success (ground pad) | F9 FT B1035.2 | CCAFS SLC-40 |

## Task 10

Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

In [36]:
```
%sql select landing__outcome, count(*) as cn from spacextbl where landing__outcome like 'Success%' and DATE between '2010-06-04' and '2017-03-20' grou
```

```
 * ibm_db_sa://bpp32620:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[36]:

| landing__outcome | cn |
|---|---|
| Success (drone ship) | 5 |
| Success (ground pad) | 3 |

# Appendix 2.1 – Data preparation for modeling

1. Standardize X variables:

```
transform = preprocessing.StandardScaler()
X=transform.fit(X).transform(X)
```

2. Split into Train and Test sets:

```
X_train, X_test, Y_train, Y_test=train_test_split(X,Y,test_size=0.2, random_state=2)
```

# Appendix 2.2 – Parameters tuning

**GridSearchCV** is used for parameters tunning. Example of Logistic Regression:

```
lr=LogisticRegression()
logreg_cv=GridSearchCV(lr,parameters,cv=10)
logreg_cv.fit(X_train, Y_train)

print("tuned hyperparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

Thank you!