DATA SCIENCE CAPSTONE PROJECT

BUSINESS PRESENTATION

Phan Pham Quynh Hoa

19 Jun 2024

TABLE OF CONTENT

Data Science Capstone Project

EXECUTIVE SUMMARY

- Collected data from public SpaceX API and SpaceX Wikipediapage.Created labels column ‘class’ which classifies successful landings.Explored data using SQL, visualization, foliummaps, anddashboards.Gathered relevant columns to be used as features. Changedall categorical variables to binary using one hot encoding. Standardizeddata and used GridSearchCV to find best parameters for machinelearning models. Visualize accuracy score of all models.

EXECUTIVE SUMMARY

Four machine learning models were produced: Logistic Regression,Support Vector Machine, Decision Tree Classifier, and KNearestNeighbors. All produced similar results with accuracy rateof about83.33%. All models over predicted successful landings. Moredataisneeded for better model determination and accuracy.

INTRODUCTION

BACKGROUND

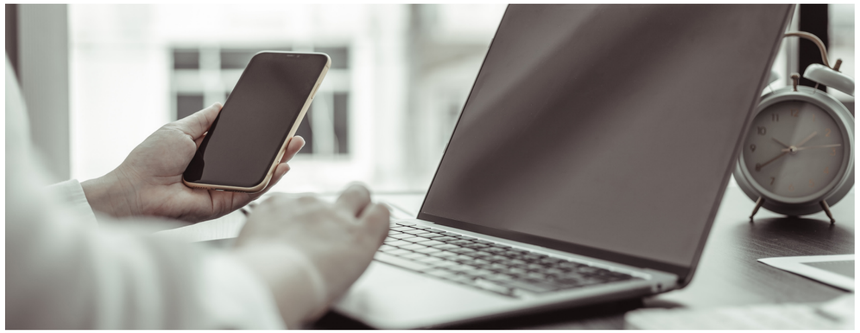
SpaceXisthe mostsuccessful company of the commercial space industry, making space travel affordable.The companyadvertises Falcon9 rocket launcheson itswebsite, with acost of 62 milliondollars; other providerscost upward of 165 million dollars each, much ofthe savingsis becauseSpaceX can reusethe first stage. Therefore, ifwe can determinethe first stage will land, we can determinethe cost of a launch. Based on public information and machine learning models, we are goingtopredict if SpaceX willreuse the first stage.

PROBLEMS

01 Space Y tasks us to train a machinelearning model to predict successful Stage 1 recovery

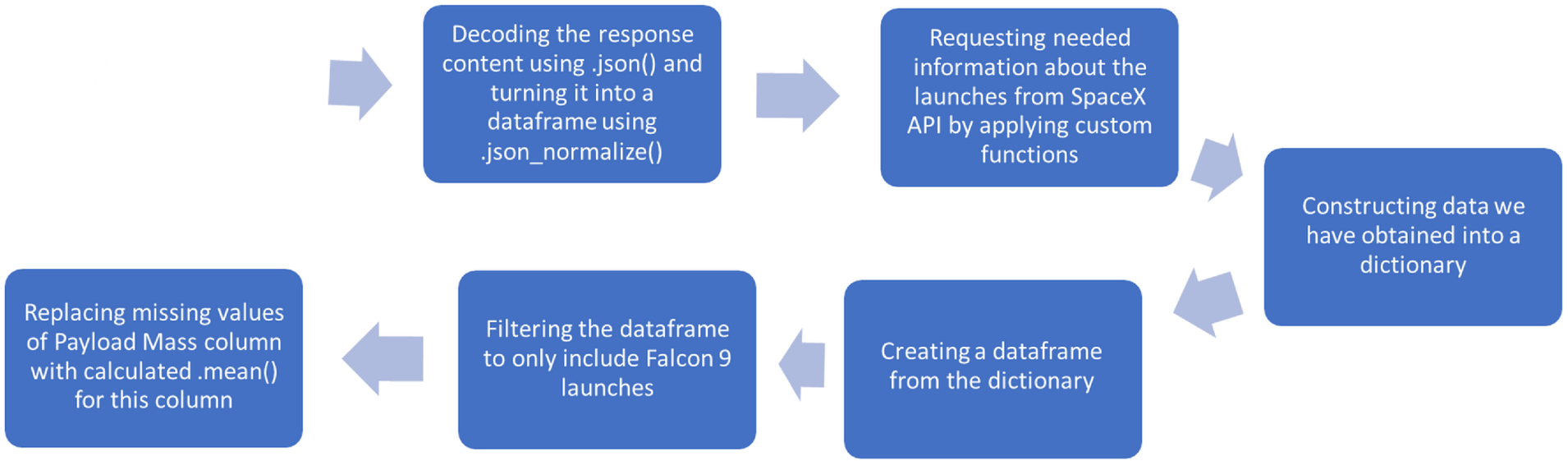
02 What is the best Classificationalgorithm that can be used for theprediction of the reuseof a first stage?

METHODOLOGY

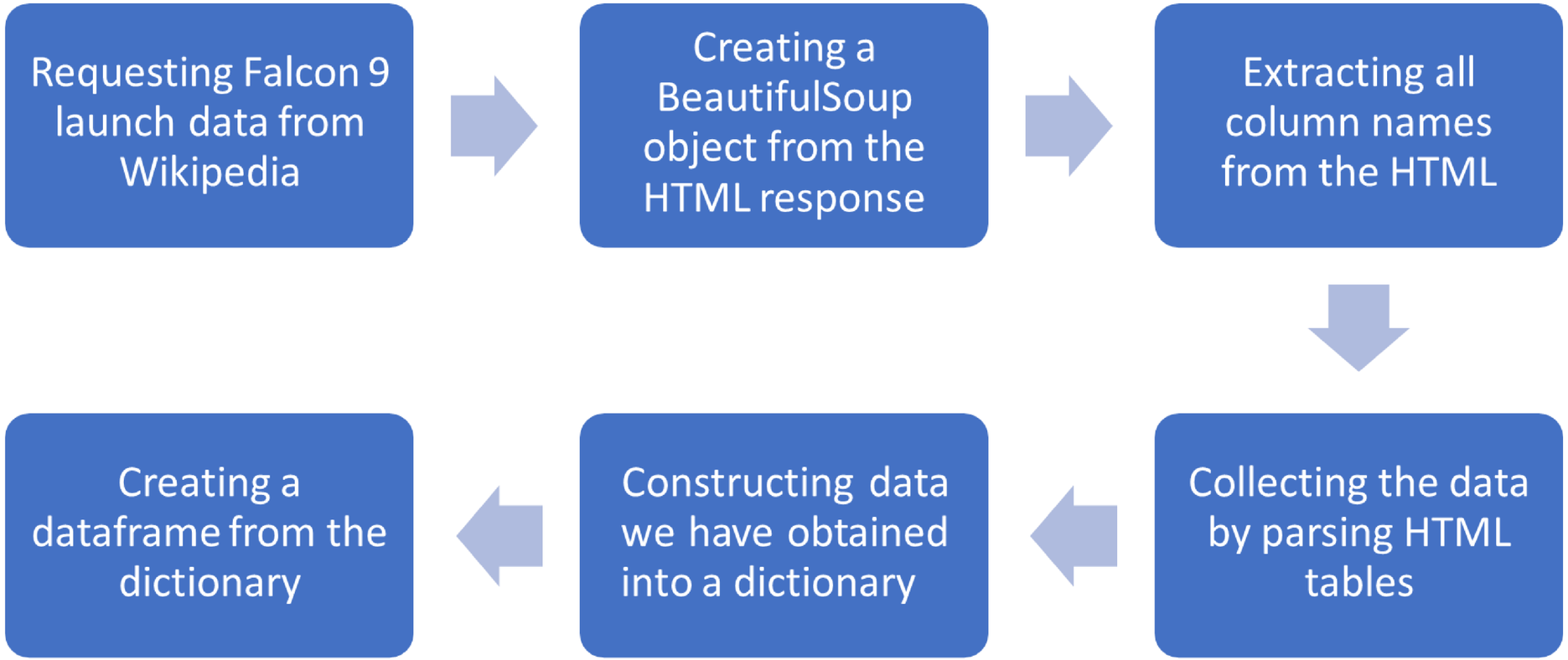


- Data collection methodology: Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling: Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using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

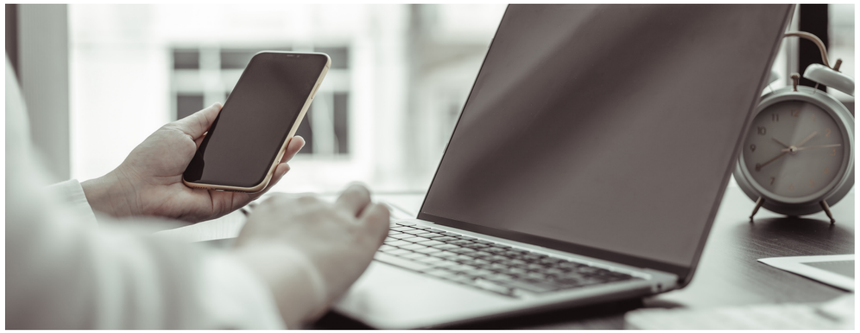
DATA COLECTION– SPACEXAPI



DATA COLECTION– WEBSCRAPING



DATA WRANGLING



- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean and others.
- We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.

EDA WITH DATA VISUALIZATION

Charts 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 OrbitType.
- Success Rate Yearly Trend.

ChartType and theirUse Cases:

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

Bar charts show comparisons among discrete categories. The goal 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).

EDA WITH SQL

PerformedSQLqueries:

•Displaying the names of the unique launch sites in thespace mission
•Displaying 5 records wherelaunch sites begin withthe string ‘CCA'
•Displaying the total payload mass carried by boosters launched by NASA(CRS)
•Displaying average payload mass carried by boosterversion F9 v1.1
•Listing the date when the first successful landing outcome in ground pad was achieved
•Listing the names of theboosters 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 namesof thebooster versions which have carried themaximum payload mass
•Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year2015
•Ranking the countof landing outcomes (such as Failure (drone ship) or Success (ground pad))between thedate 2010-06-04 and 2017-03-20 in descending order

BUILD AN INTERACTIVE MAP WITH FOLIUM

Markers of all Launch Sites

Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

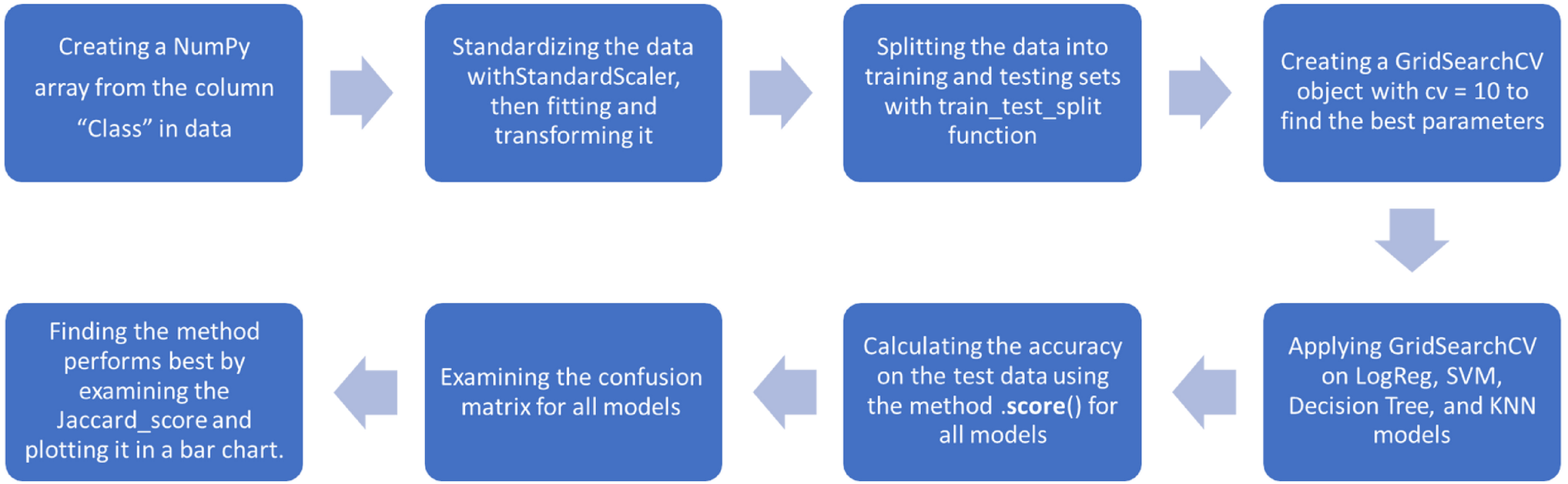
Distances between a Launch Site to its proximities:

Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

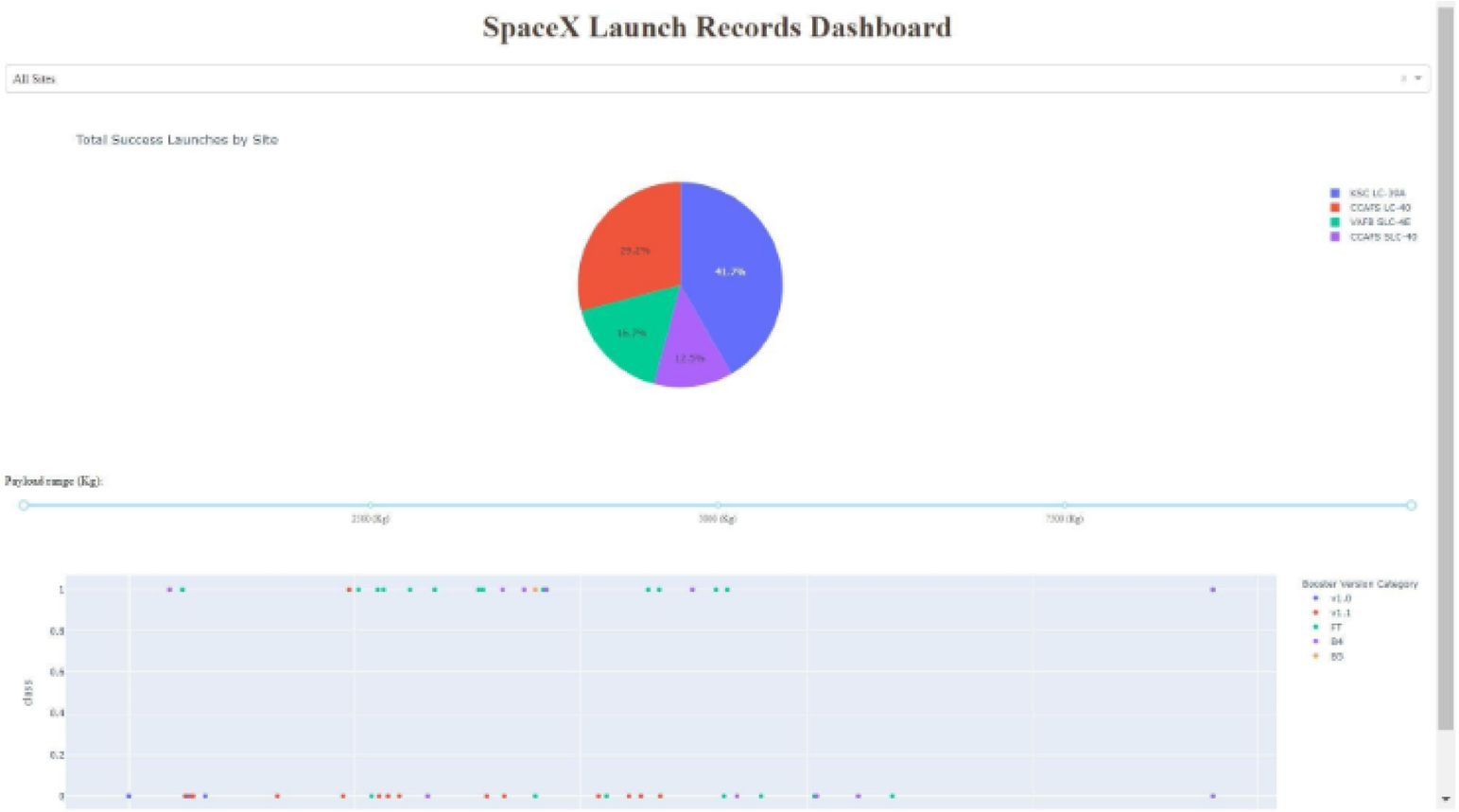
BUILD A DASHBOARD WITH PLOTLY DASH

Dashboard includes a pie chart and a scatter plot. Pie chart can be selected to show distribution of successful landings acrossall launchsites and can be selected to show individual launch site success rates. Scatter plot takes two inputs: All sites or individual site and payloadmassonasliderbetween 0 and 10000 kg. The pie chart is used to visualize launch site success rate. The scatter plot can help us see how success varies across launch sites, payloadmass, and booster version category.

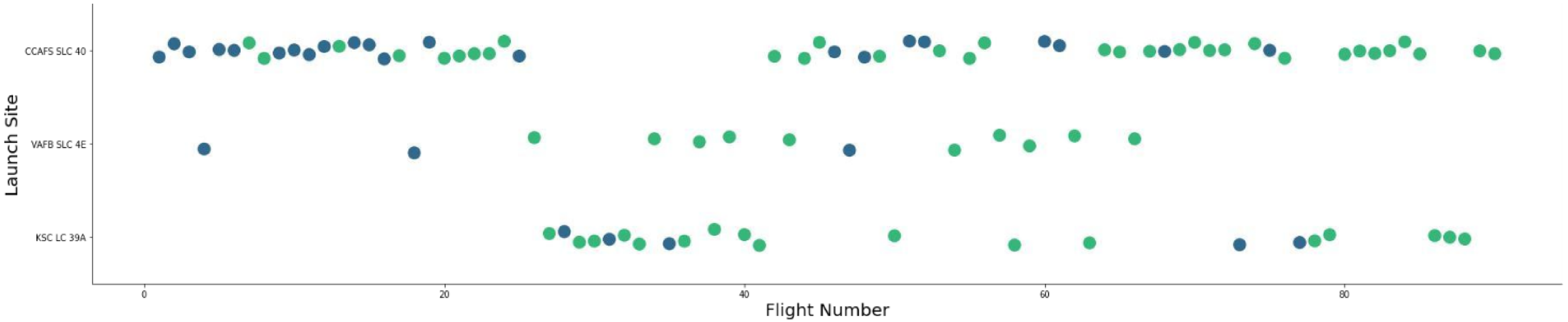
PREDICTIVE ANALYSIS (CLASSIFICATION)



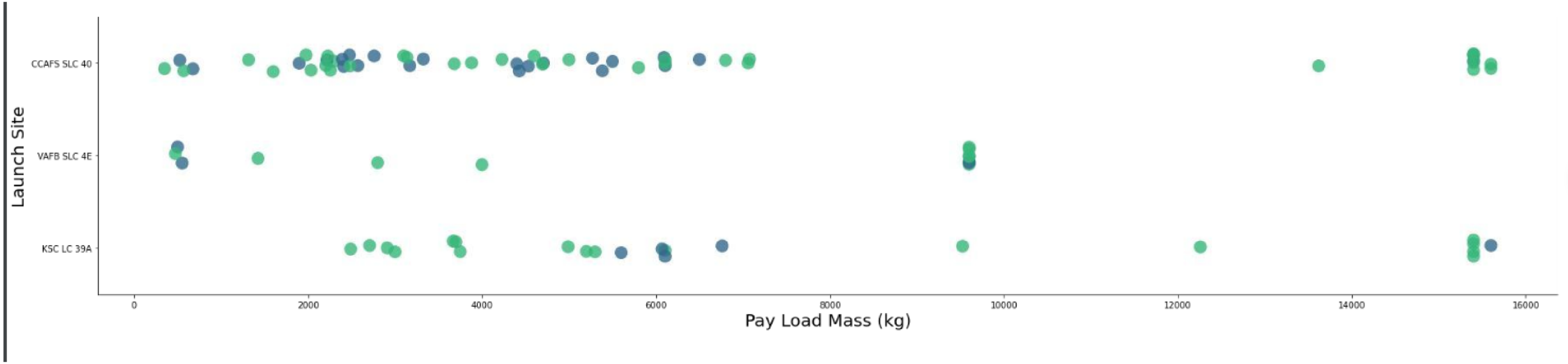
RESULTS


SpaceX Launch Records Dashboard
All Sites
Total Success Launches by Site
41.7%
29.2%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
2500 (Kg)
5000 (Kg)
7500 (Kg)
class
Booster Version Category
v1.0
v1.1
FT
B4
B5

FLIGHT NUMBER VS. LAUNCH SITE



PAYLOAD VS. LAUNCH SITE



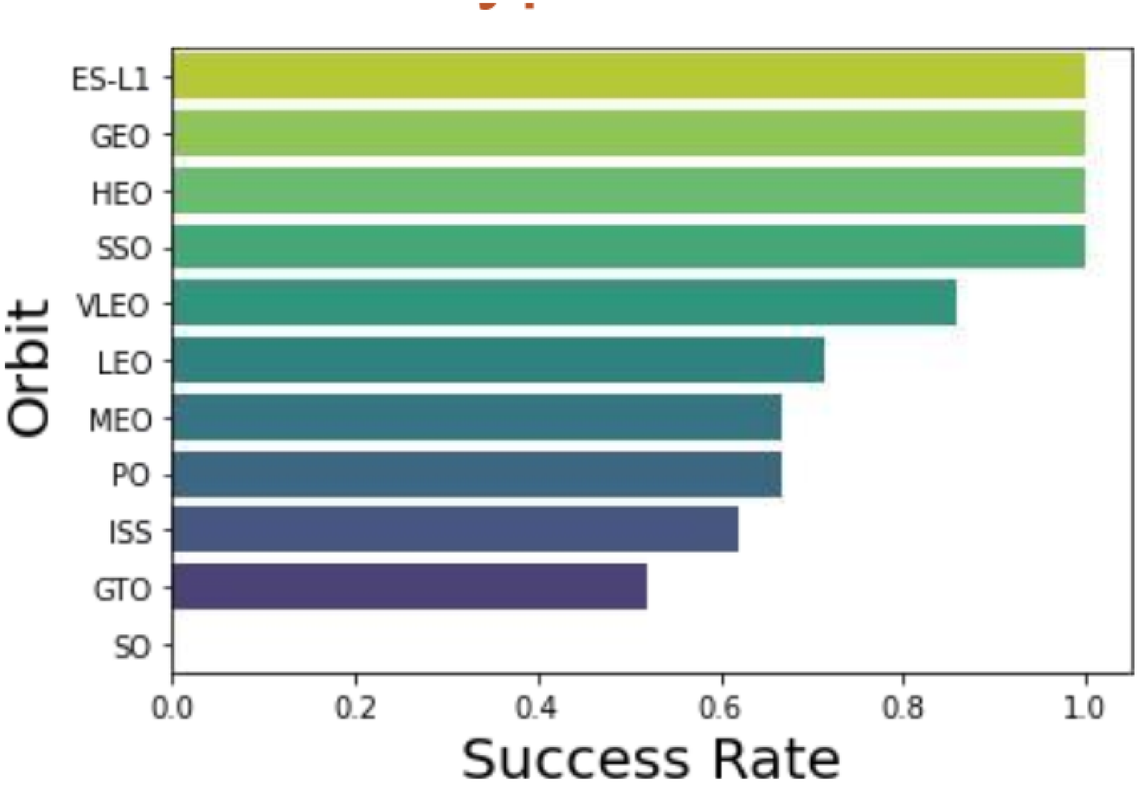
Explanation:

•Foreverylaunchsitethe higher the payloadmass,the higher thesuccessrate.

•Mostofthelaunchеswith payload mass over 7000kg were successful.

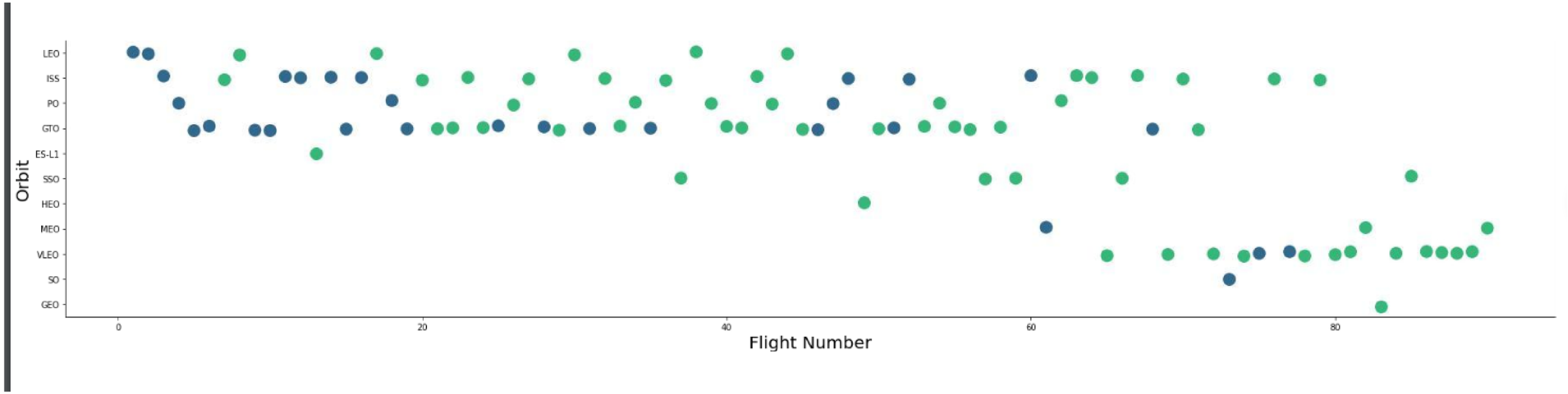
•VAFBSLC 4Ehasthehighest success rate amongthe launch sites.

SUCCESS RATE VS. ORBIT TYPE

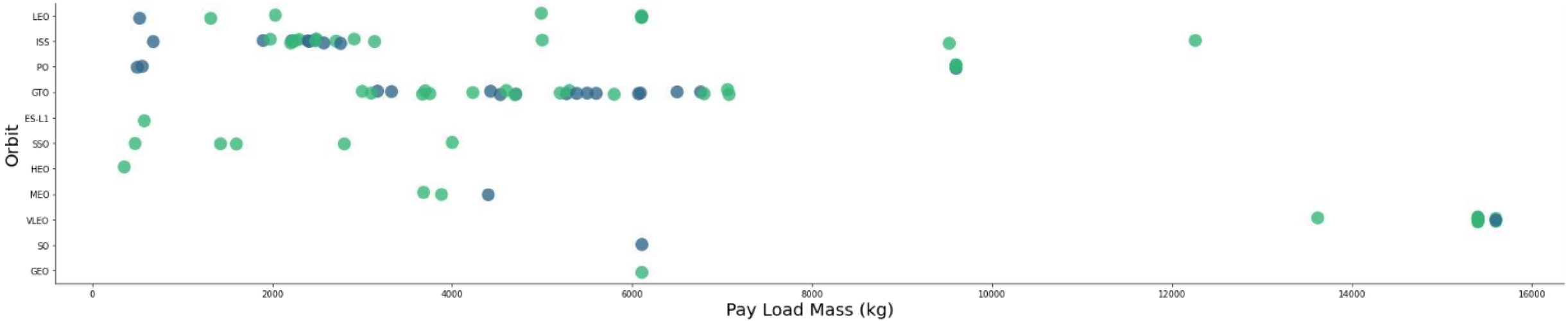


Success Rate Scale with as 0% 0.6 as 60% 1

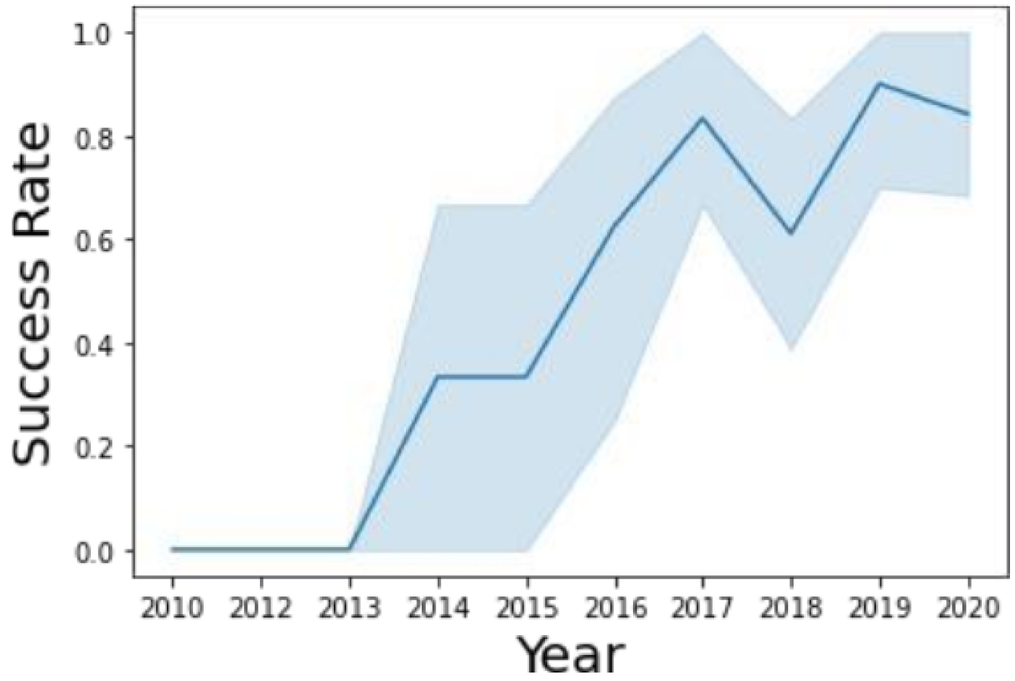
FLIGHT NUMBER VS. ORBIT TYPE



PAYLOAD VS. ORBIT TYPE



LAUNCH SUCCESS YEARLY TREND



95% confidence interval (light blue shading)

AL LAUNCH SITE NAMES

In [14]:
```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

Out[14]:

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

LAUNCH SITE NAMES BEGINNING WITH`CCA

In [9]:
```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' Limit 5
```

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

Out[9]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

LAUNCH SITE NAMES BEGINNING WITH`CCA

In [9]:
```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' Limit 5
```

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

Out[9]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

TOTAL PAY LOAD MASS FROM NASA

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

0]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

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

0]:

TOTAL_PAYLOAD_MASS
45596

AVERAGE PAYLOAD MASS BY F9 V1.1

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

In [11]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

Out[11]:

AVERAGE_PAYLOAD_MASS
2928.4

FIRST SUCCESS FULGROUND PAD LANDING DATE

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

Hint:Use min function

[12]:
```
%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTBL \
    WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) no such column: LANDING__OUTCOME
[SQL: SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTBL      WHERE LANDING__OUTCOME = 'Success (ground pa
d)';]
(Background on this error at: http://sqlalche.me/e/e3q8)
```

SUCCESS FULDRONE SHIPLANDING WITH PAYLOAD BETWEEN 4000 AND 6000

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

In [13]:

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

```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) no such column: LANDING__OUTCOME
[SQL: SELECT BOOSTER_VERSION FROM SPACEXTBL      WHERE (LANDING__OUTCOME = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BET
WEEN 4000 AND 6000);]
(Background on this error at: http://sqlalche.me/e/e3q8)
```

TOTAL NUMBER OF EACH MISSION OUTCOME

In [14]:
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

Out[14]:

Mission_Outcome	TOTAL_NUMBER
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

BOOSTERS THAT CARRIED MAXIMUM PAYLOAD

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

In [15]:
```
%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXTBL \
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[15]:

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 FAILED DRONESHIP LANDING RECORDS

months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [16]:

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

```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) near "FROM": syntax error
[SQL: SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL      WHERE (LANDING__OUTCOME = 'Failure (drone ship)') AND (EXTRACT
(YEAR FROM DATE) = '2015');]
(Background on this error at: http://sqlalche.me/e/e3q8)
```

RANKING COUNTS OF SUCCESS FULLANDINGS 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.

In [17]:

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

* sqlite:///my_data1.db

RANKING COUNTS OF SUCCESS FULLANDINGS 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.

In [17]:

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

* sqlite:///my_data1.db

THANK YOU

19 Jun 2024