IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Konrad Płoszaj

24.07.2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
    - Insights drawn from EDA
    - Launch Sites Proximities Analysis
    - Build Dashboard with Plotly Dash
    - Predictive Analysis (Classification)
- Conclusion
- Appendix

Executive Summary

- During the project data has been collected via SpaceX API, using Web Scrapping from Wikipedia (using Beautiful soup library).
- Missing data has been replace and collected data have been analyzed to find patterns. Additional column class has been created to assess if rocket first stage successfully landed.
- To understand received dataset EDA (Exploratory Data Analysis) has been performed using SQL (SQL lite – built in Jupyter Lab).
- Identified patterns between landing outcomes or success landing rates and independent features have been visualized on charts (matplotlib and seaborn libraries), map (folium library), and interactive dashboard (Dash library).
- Machine Learning Predictions have been created. Data has been previously standardized and splitted into train set and test set. Best parameters has been found by using GridSearchCV method, and train models accuracy have been tested. Used Machine learning models:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K Nearest Neighbors (KNN)

Introduction

- Falcon 9 is a reusable, two-stage rocket designed and manufactured by SpaceX for the reliable and safe transport of people and payloads into Earth orbit and beyond. Falcon 9 is the world's first orbital class reusable rocket. Reusability allows SpaceX to refly the most expensive parts of the rocket, which in turn drives down the cost of space access. Falcon 9 is a reusable, two-stage rocket designed and manufactured by SpaceX for the reliable and safe transport of people and payloads into Earth orbit and beyond. Falcon 9 is the world's first orbital class reusable rocket. Reusability allows SpaceX to refly the most expensive parts of the rocket, which in turn drives down the cost of space access. ~ *https://www.spacex.com/vehicles/falcon-9/*
- Not every launch of Falcon 9 was success -the purpose of the project is learn how to predict if the Falcon 9 first stage will land successfully. To achive the goal multiple independent features have to be analysed to find the best pattern.

Section 1

Methodology

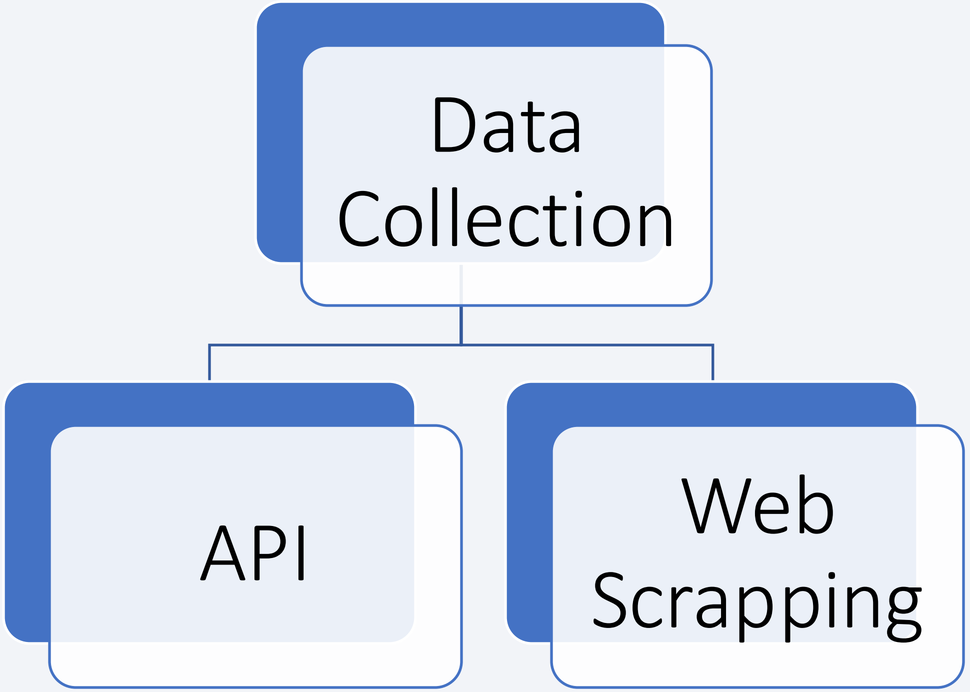
Methodology

Executive Summary

- Data collection methodology:
    - Data has been collected via SpaceX API, using Web Scrapping from Wikipedia (using Beautiful soup library)
- Perform data wrangling
    - Null values have been replaced. Launching sites, dedicated orbites, landing outcomes has been identified. New additional column class has been created to assess if rocket first stage successfully landed.
- 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

- During the project data has been collected via SpaceX API, using Web Scrapping from Wikipedia (using Beautiful soup library).



Data Collection

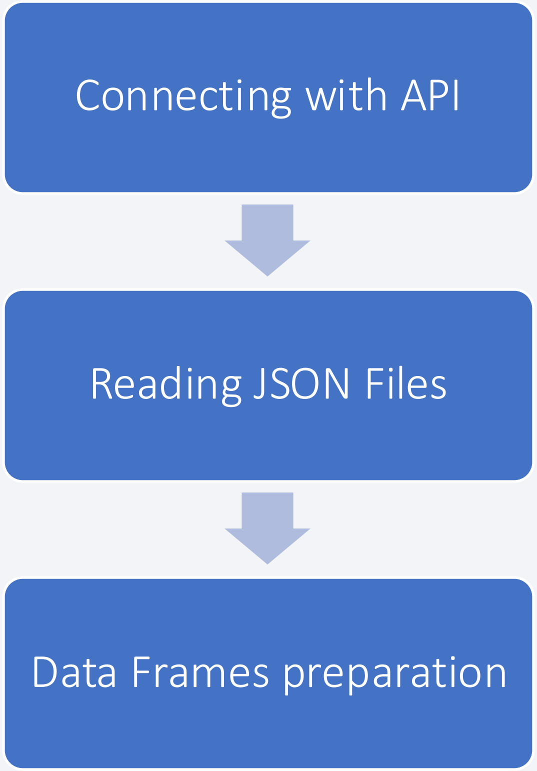
API

Web Scrapping

Data Collection – SpaceX API

- JSON files for rocket launch history, cores, loads, launchpads and rockets have been downloaded from API: https://api.spacexdata.com/v4/ by using get request. After static_json_url has been created from course website.
- Dataframe for collected features has been created, and filtered to only include Falcon 9 Launches
- Collected data contains data up to 13th of November 2020

- jupyter-labs-spacex-data-collection-api.ipynb

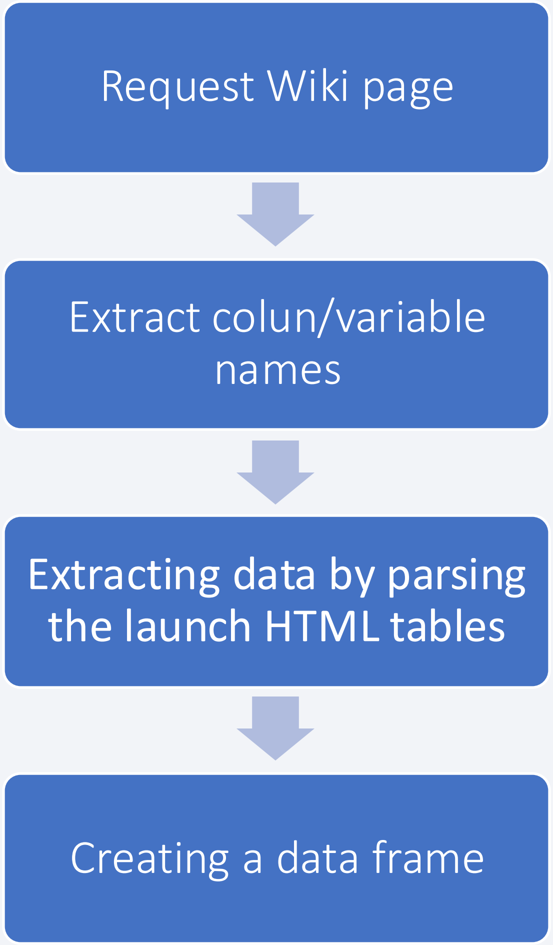

Connecting with API

Reading JSON Files

Data Frames preparation

Data Collection - Scraping

- Requestinng Falcon 9 Wikipedia page from its URL, by get function and using Beautifull soup library
- Extracting all column/variable names from the HTML table header, by using find_all function.
- Extracting data by parsing the launch HTML tables and pasting them(using for loop) to created lists for every column.
- Merge all columns into Data Frame
- Collected data contains data up to 9th of June 2021
- jupyter-labs-webscraping.ipynb


Request Wiki page
Extract colun/variable names
Extracting data by parsing the launch HTML tables
Creating a data frame

Data Wrangling

- Null values have been replaced. Launching sites, dedicated orbites, landing outcomes has been identified. New additional column class has been created to assess if rocket first stage successfully landed.
- Payload mass null values has been replaced by mean Payload mass, and null values for Landing Pad was not used in furthere process.
- Launching sites, dedicated orbites, landing outcomes has been identified.
- New additional column class has been created to assess if rocket first stage successfully landed.
- labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

The purpuse of EDA was to find corelation between particular features, following charts have been used:

1. Scatter plot to visualize the relationship between Flight Number, Payload Mass
2. Scatter plot to visualize the relationship between Flight Number, Launch Site and the outcome
3. Scatter plot to visualize the relationship between Payload Mass, Launch Site and outcome
4. Bar plot to visualize the relationship between success rate of each orbit type
5. Scatter plot to visualize the relationship between FlightNumber and Orbit type and outcome
6. Scatter plot to visualize the relationship between Payload Mass and Orbit type and outcome
7. Line plot to visualize the launch success yearly trend

8.edadataviz.ipynb

EDA with SQL

SQL queries performed:

- Display the names of the unique launch sites in the space mission
- Display 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 succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship 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. Use a subquery
- 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 landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

- jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map shows all launching sites used by SpaceX. Clusters shows how many launches was from every site and additonal markers, which appear after clicking on the site shows succesfull (green) and failed (red) launches. Map shows also distances from the nearest:
    - City
    - Coast
    - Railway
    - Highway
- The objects enable to identify the pattern of locations from SpaceX perform the launches, their succes rate and identify how far from key objects they are located

- lab_jupyter_launch_site_location.ipynb

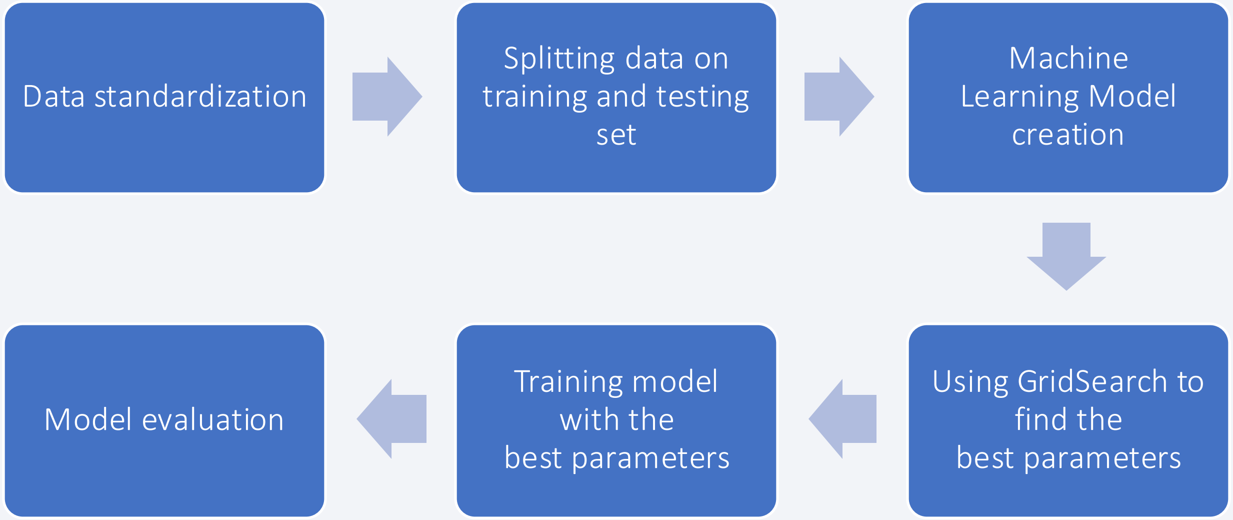
Build a Dashboard with Plotly Dash

- Dashboard shows following charts:
  - Total Success Launches by site (for all sites selected)
  - Success/Failed Launches in selected launch site
  - Correlation between Payload and Success for all sites or selected site
- Selected plots and interactions (possibility to select site and Payload interval) enable to identify launches sites with the largest number of succesfull launches and success rate, and corelation between payload, launch site and outcome.

- dash_interactivity.py

Predictive Analysis (Classification)

- Machine Learning Predictions models have been designed. Data has been previously standardized and splitted into train set and test set. Best parameters has been found by using GridSearchCV method, and train models accuracy have been tested. Used Machine learning models:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K Nearest Neighbors (KNN)



Data standardization

Splitting data on training and testing set

Machine Learning Model creation

Using GridSearch to find the best parameters

Training model with the best parameters

Model evaluation

- Prediction_Part_5.ipynb

Results

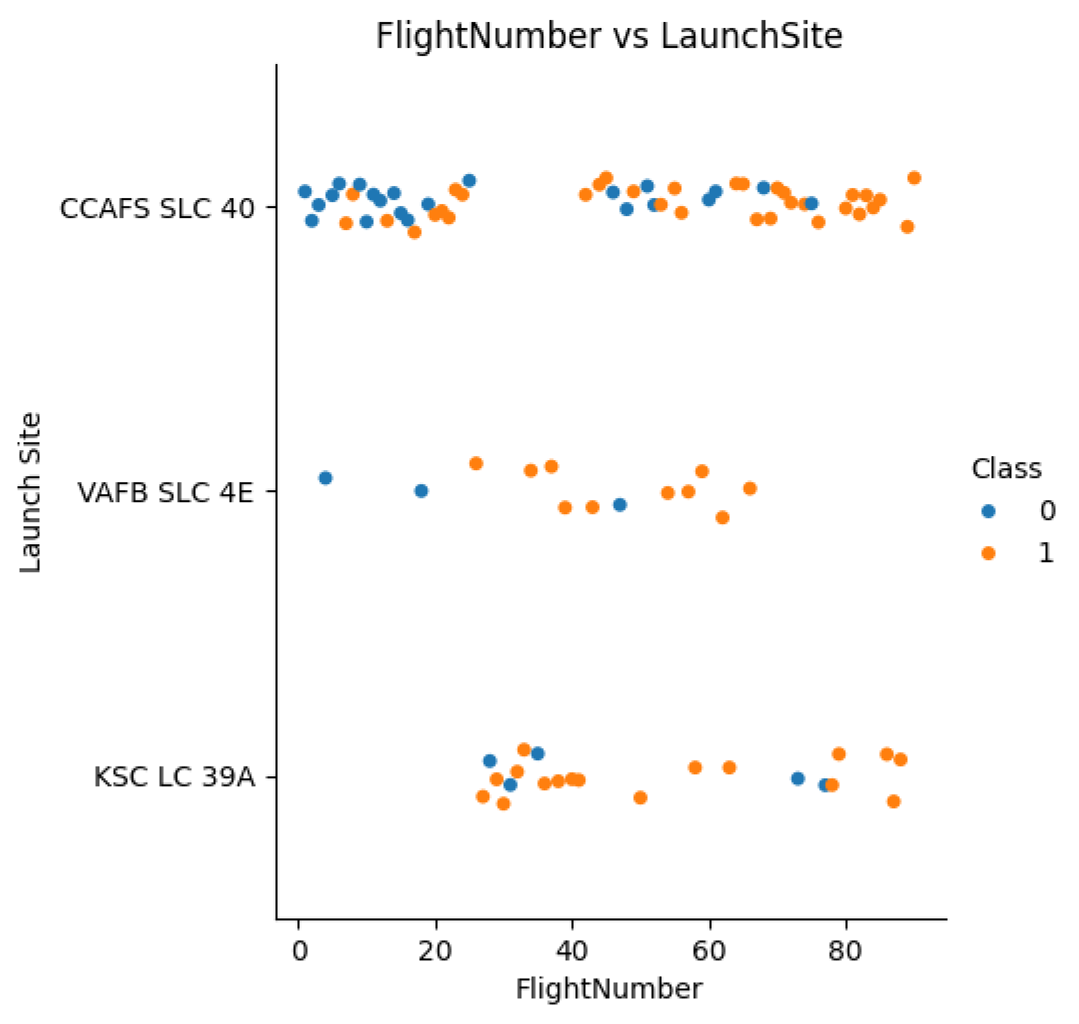
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

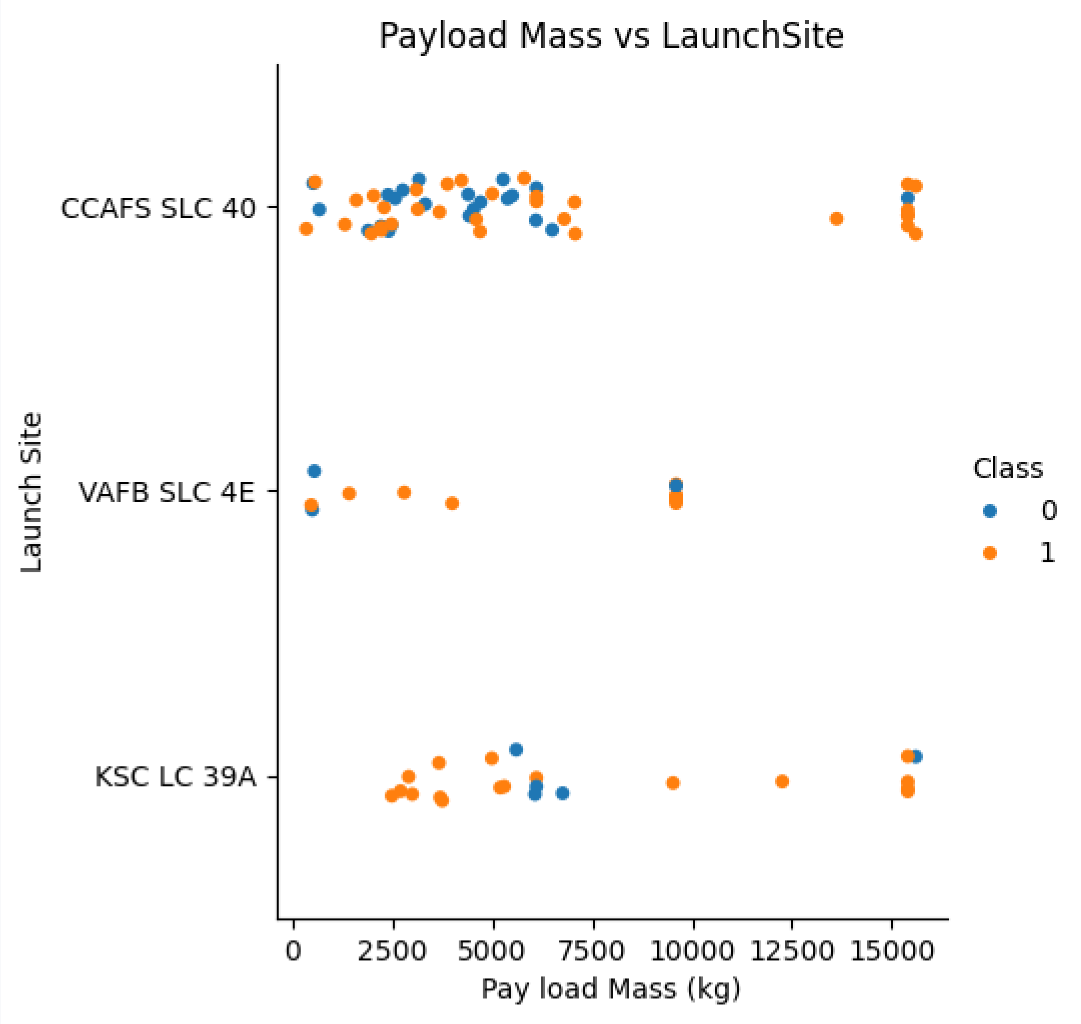
Flight Number vs. Launch Site

- The smallest number of Flights was from VAFB SLC 4E site, and it is no longer used by SpaceX
- The highest number of launches have been performed from CCAFS SLC 40. Approximately between 25th and 40th launch this site was not in use. Instead using of KSC LC 39A has been started.
- Succes probability is correlated with Flight Number- As higher number as higher probability of successfull landing



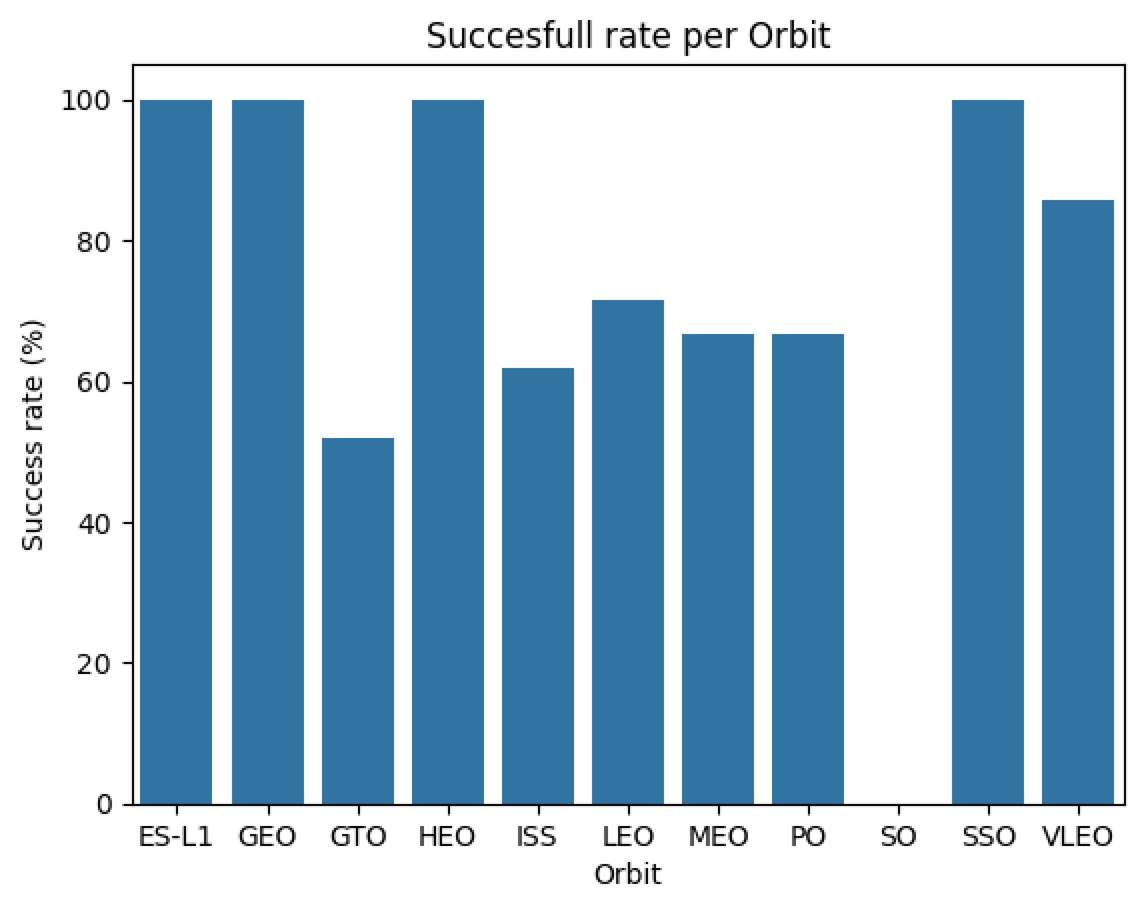
Payload vs. Launch Site

- VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- CCAFS SLC 40 is the most frequent used site to launch haviest payloads
- Limit of the payload is around 15-16 tones



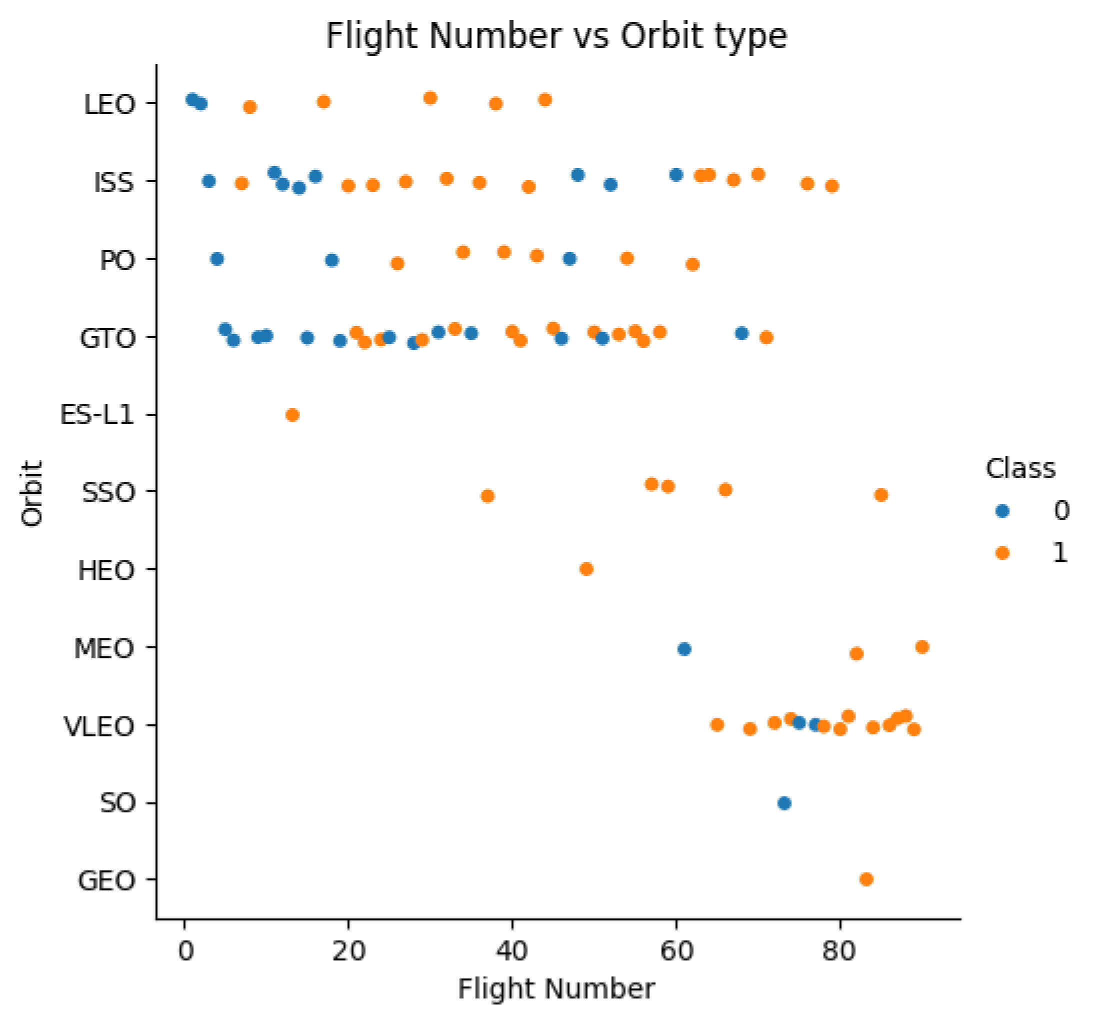
Success Rate vs. Orbit Type

- 100% of success rate have orbites ES-L1, GEO, HEO and SSO. All these orbites had small numer of launches
- The worst rate has SO orbit- 0% success rate - 1 attempt 1 fail



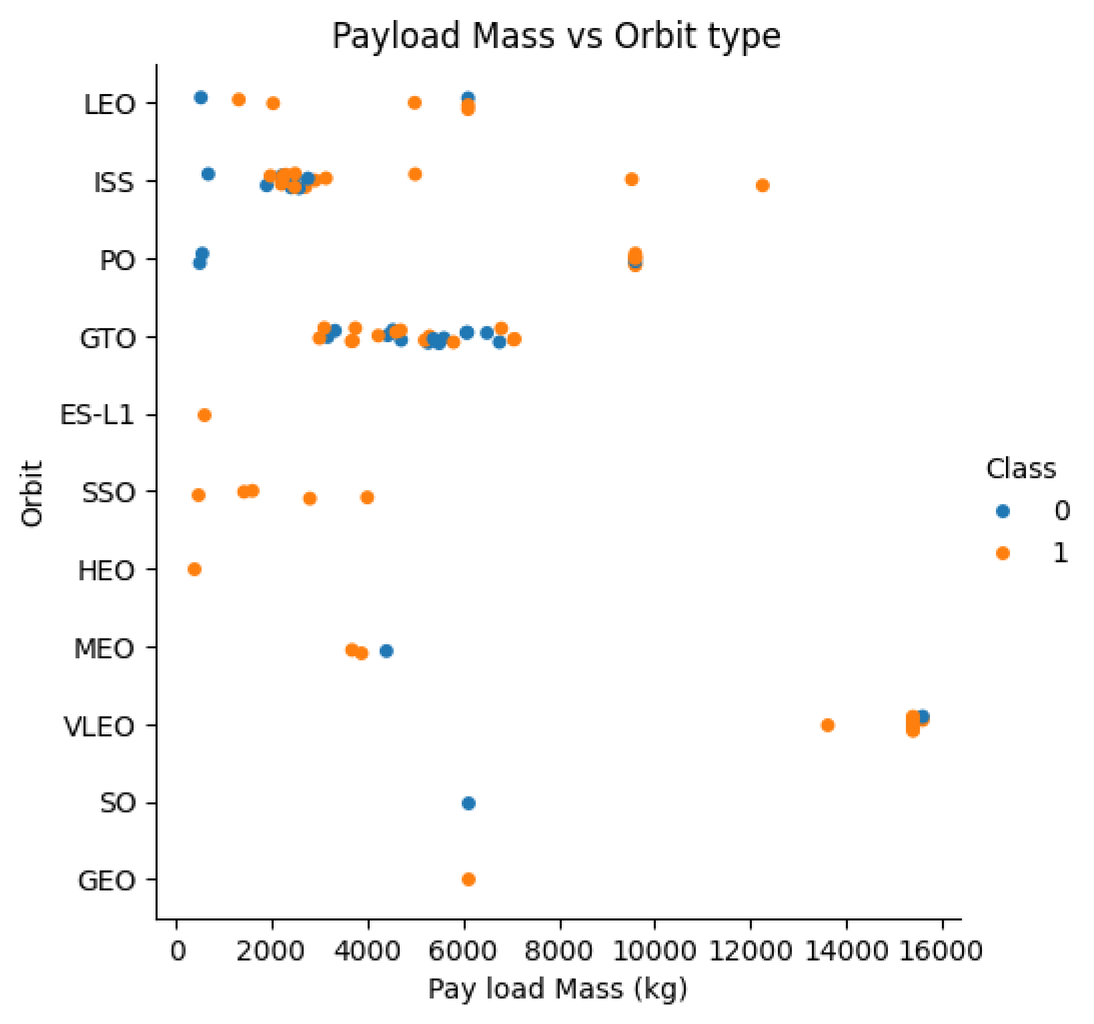
Flight Number vs. Orbit Type

- LEO orbit, success seems to be related to the number of flights. Conversely, in the
- GTO orbit, there appears to be no relationship between flight number and success.
- VLEO orbit has high success rate, but launches on this orbit started after 60th launch



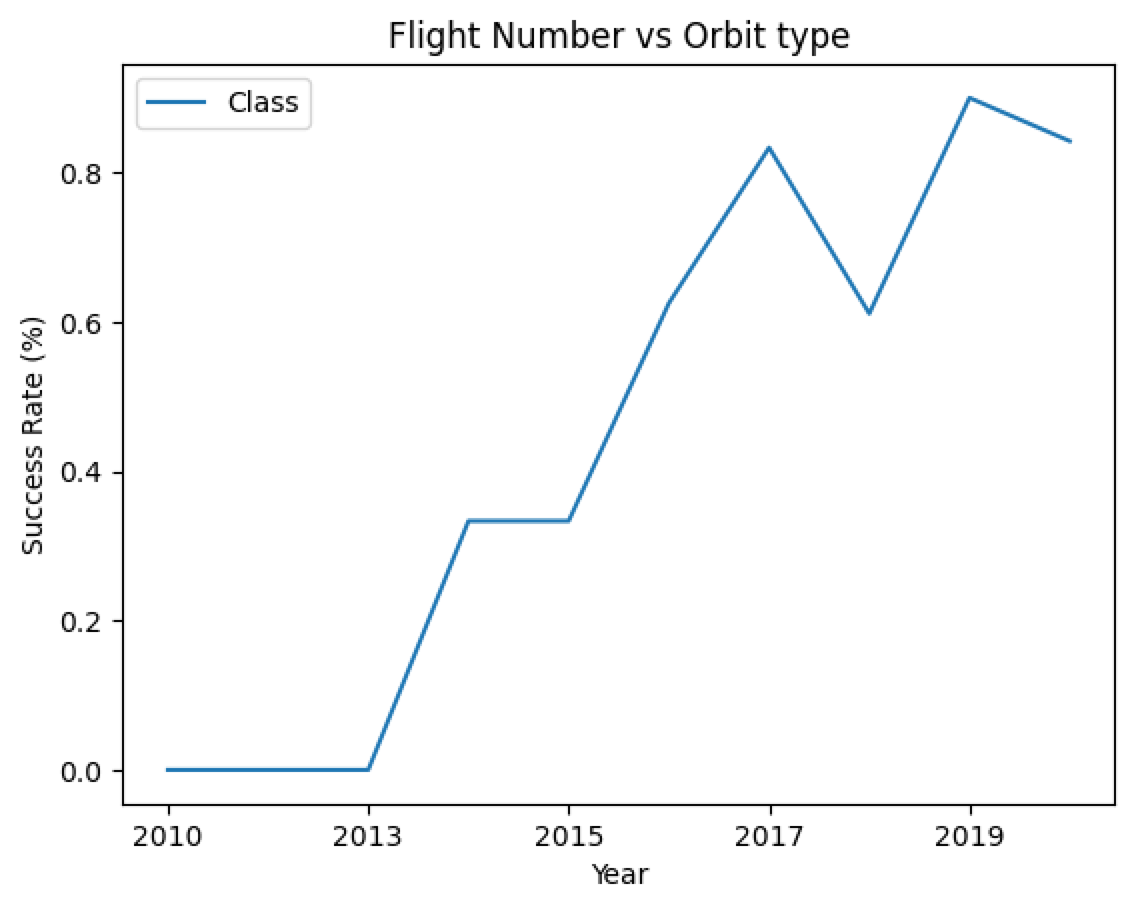
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.
- VLEO orbit has the heaviest payloads



Launch Success Yearly Trend

- sucess rate since 2013 kept increasing till 2020, with small correction in 2018
- In years 2010-2013 success rate was equal 0%



All Launch Site Names

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

- Query:

```
SELECT DISTINCT Launch_site FROM SPACEXTABLE
```

Display Unique Launching sites from data base

- SpeceX perform launches from 4 launch sites.
- CCAFS LC-40 & CCAFS SLC-40 are close to each other

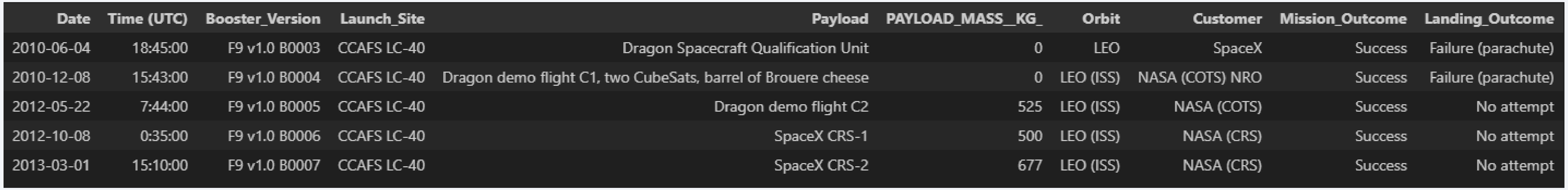
Launch Site Names Begin with 'CCA'

5 first records where launch sites begin with `CCA`

Query:

```
SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

Display all items from data basefor which launch sites starts from 'CCA'. Display only 5



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

- All have been launeched form CCAFS LC-40 on LEO orbit.
- Mission outcome was success, but no one first stage rocket landed successfully

Total Payload Mass

- Total payload carried by boosters from NASA:

  45 596 kg

- Query:

  SELECT SUM(PAYLOAD_MASS__KG_)

  Calculate sum of payload mass

  FROM SPACEXTABLE

  from database

  WHERE Customer == "NASA (CRS)"

  for customer "NASA (CRS)"

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

  2 928.4 kg

- Query:

  SELECT AVG (PAYLOAD_MASS__KG_)

  Calculate averageof payload mass

  FROM SPACEXTABLE

  from database

  WHERE Booster_Version =='F9 v1.1'

  for booster version 'F9 v1.1'

First Successful Ground Landing Date

- The first successful landing outcome on ground pad: ***2015-12-22***
- Query:

```
SELECT MIN (Date)

 Display minimum value for column 'Date'

    FROM SPACEXTABLE

    from database

    WHERE Landing_Outcome == 'Success (ground pad)'

     for successful ground pad landing
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2
- Query:

SELECT Booster_Version *Display list of booste versions*

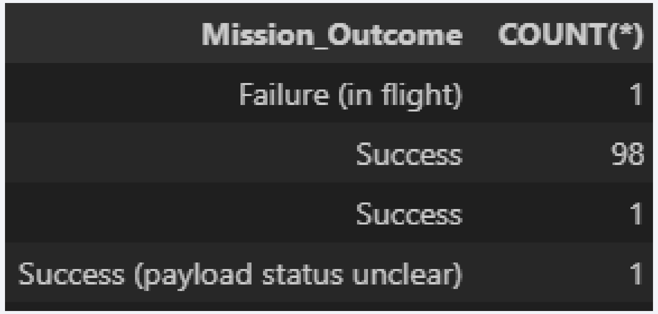
FROM SPACEXTABLE *from database*

WHERE Landing_Outcome == 'Success (drone ship)' *which landed succesfully on drone ship*

AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 *and payload mass was between 4 and 6 tones*

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes



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

- Query:

SELECT Mission_Outcome, COUNT(*) Display mission outcome and their total number

FROM SPACEXTABLE *from database*

GROUP BY Mission_Outcome *grouped by mission outcome*

Boosters Carried Maximum Payload

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

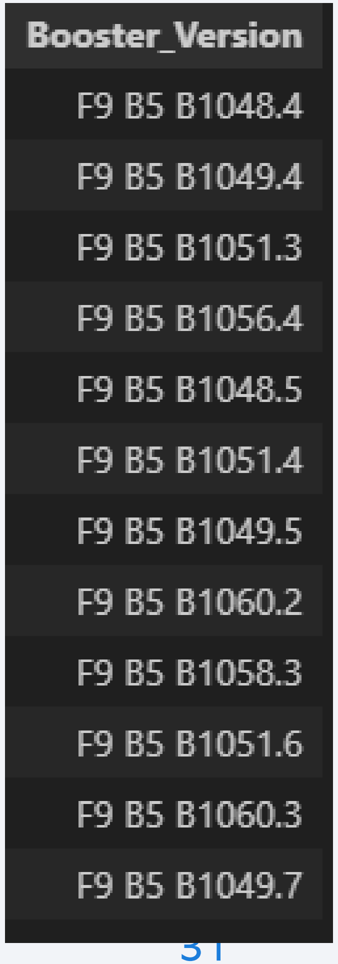
SELECT Booster_Version *Display booster version*

FROM SPACEXTABLE *from database*

WHERE PAYLOAD_MASS__KG_ == *with payload mass equal*

(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

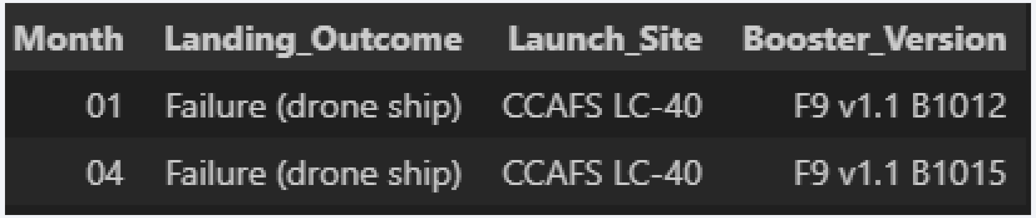
Maximum value of payload mass from the database



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

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:



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

- Query:

SELECT substr(Date,6,2) AS Month,Landing_Outcome, Launch_Site, Booster_Version

Display month (as 2 characters counted from 6th character), launch site and booster version

FROM SPACEXTABLE *from database*

WHERE substr(Date,0,5)=='2015' AND Landing_Outcome == 'Failure (drone ship)'

For year 2015 (selected as first 4 characters of column 'Date') and for landing outcome equal 'Failure (drone ship)'

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

- Query:

SELECT DISTINCT Landing_Outcome,COUNT(*) AS Number_of_Outcomes

Display landing outcome and number of outcomes

FROM SPACEXTABLE *from database*

WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'

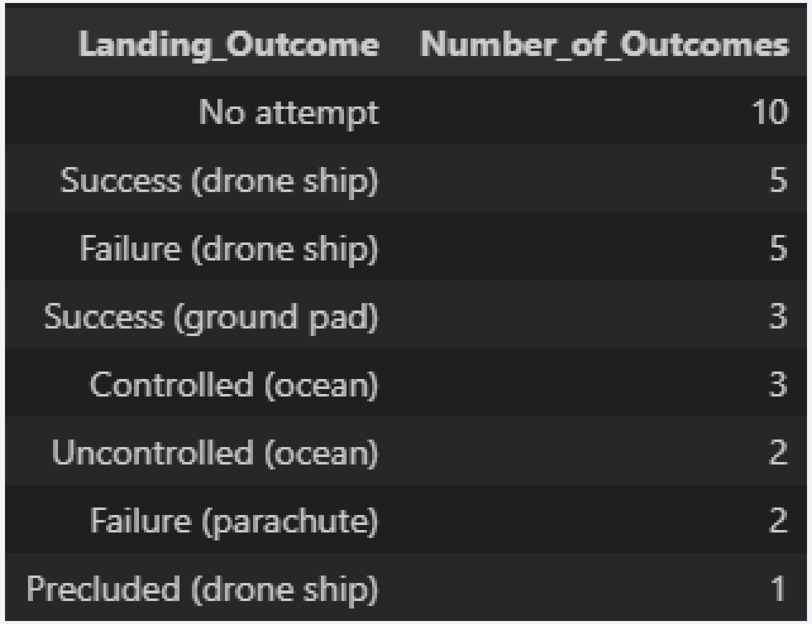
For dates between 4th of June 2010 and 20th of March 2017

GROUP BY Landing_Outcome

Grouped by landing outcome

ORDER BY Number_of_Outcomes DESC

displayed in descending order



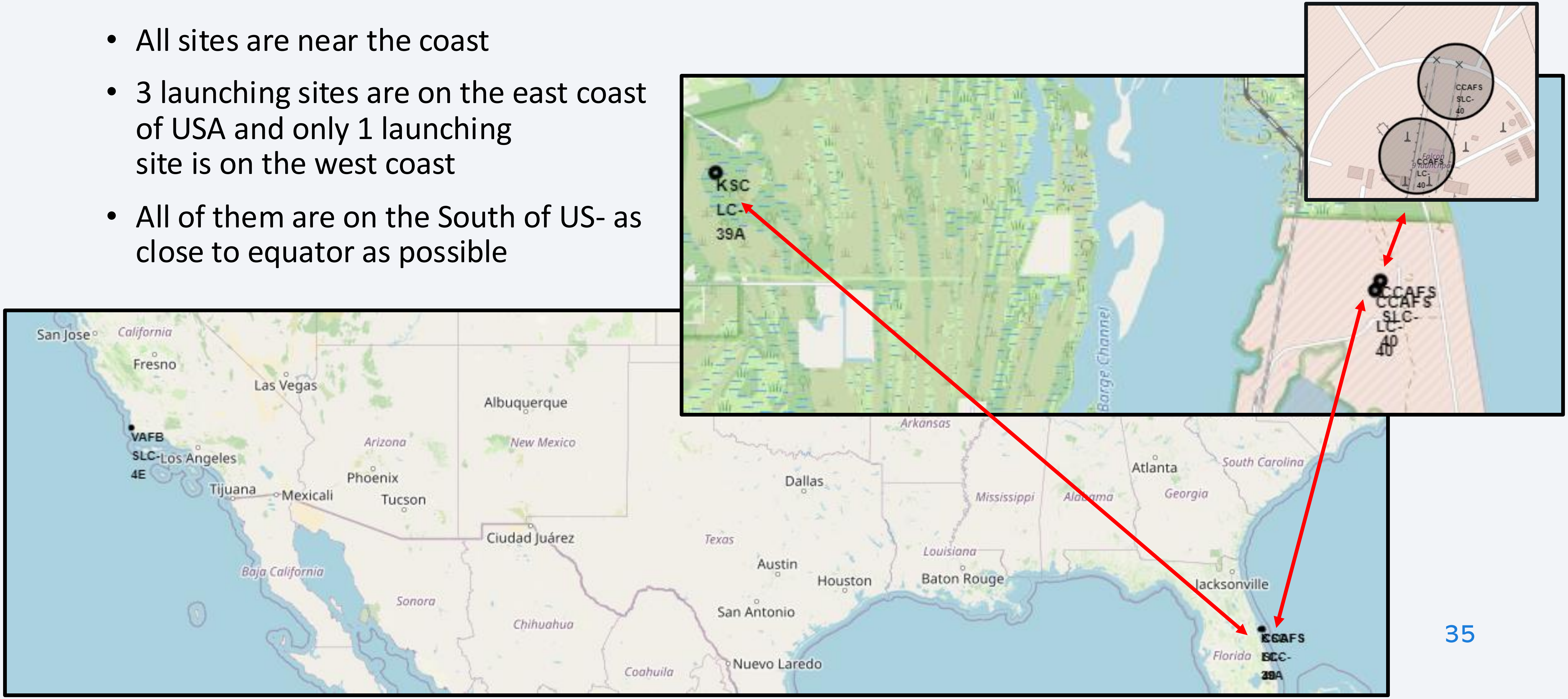
Landing_Outcome	Number_of_Outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

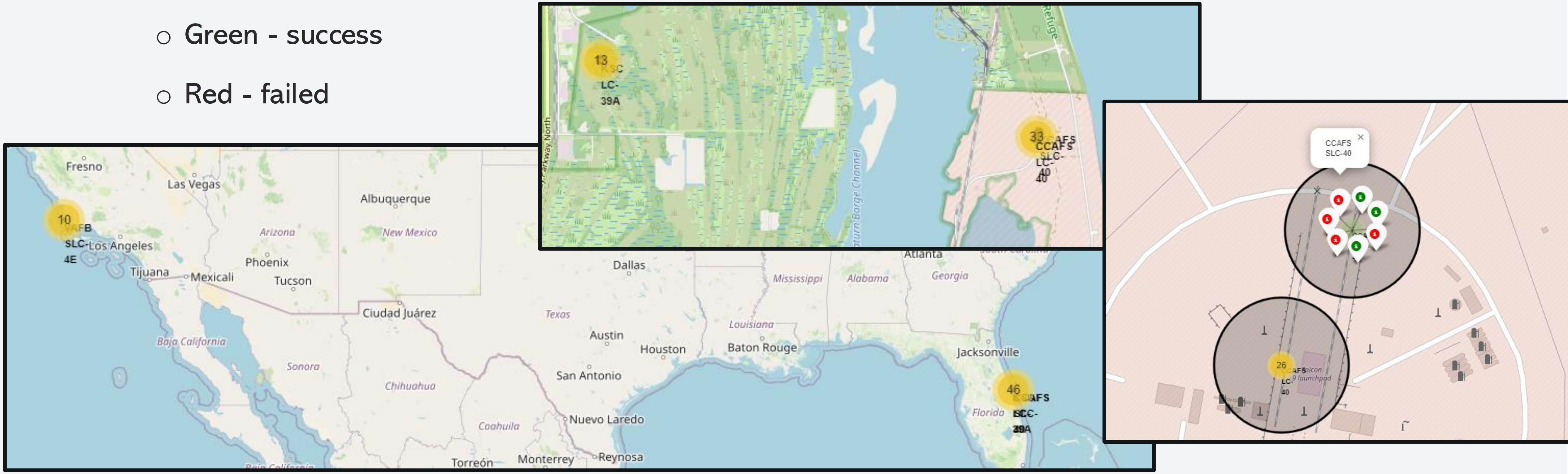
Launching sites used by SpaceX

- All sites are near the coast
- 3 launching sites are on the east coast of USA and only 1 launching site is on the west coast
- All of them are on the South of US- as close to equator as possible



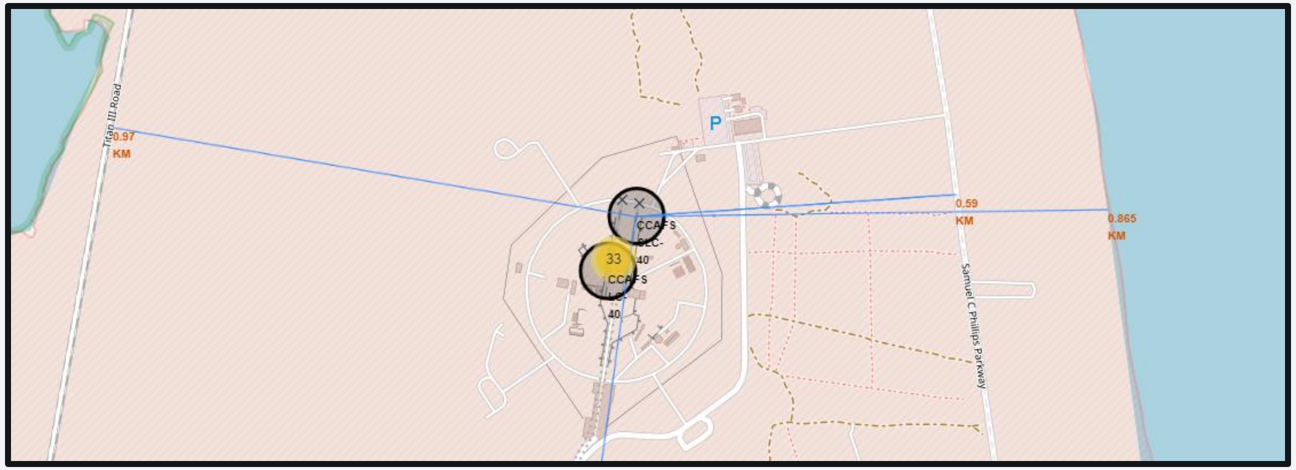
Number of Flights and successful/failed landings

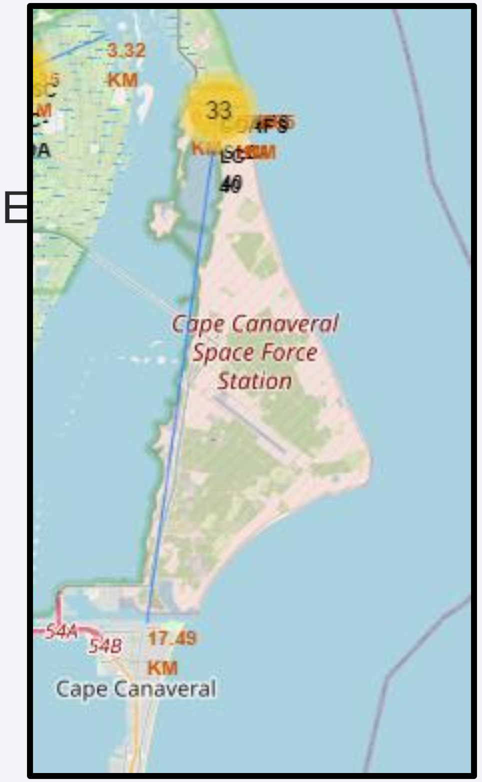
- Map shows clusters with total number of Launches from every launching site
- Additional markers, after clicking on the launching site informa about landing outcome:
  - Green - success
  - Red - failed

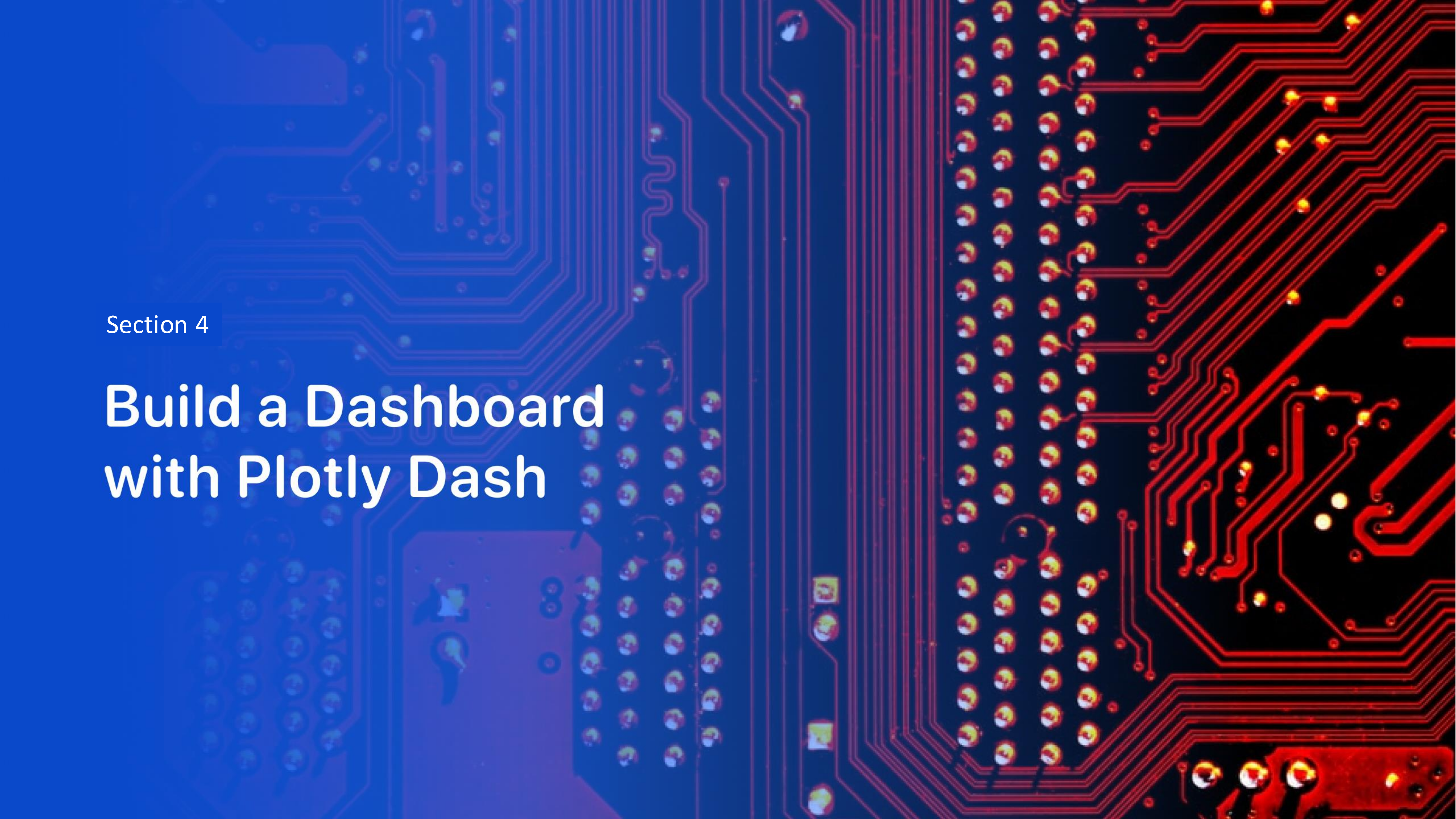


Key objects near the launching sites

- Generated Folium map shows also distances from the nearest key elements:
  - City (They are not far away from cities- around 12-15 km)
  - Coast (All of them are less then 4 km from the coast)
  - Railway (distance from the railway is no more then1 km)
  - Highway (All East coast sites are close to highways – less then 1 km), the exception in West coast launching site -VAFB SLC-4E closest highway is 15 km from the site.







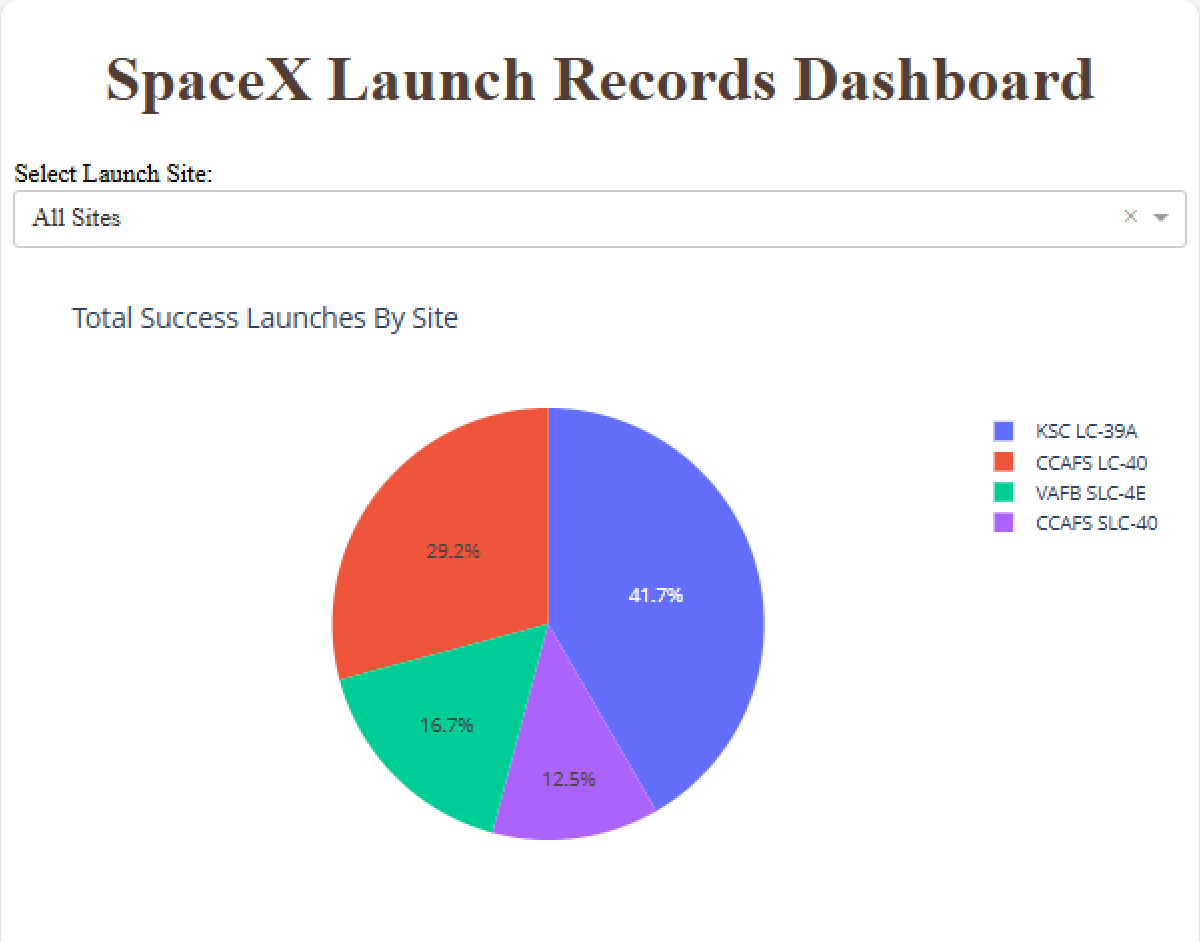
Section 4

Build a Dashboard with Plotly Dash

Total success launches by site

- The highest number of successful launches were performed from site KSC LC-39A – 10 times
- The smallest number successful launches were performed from CCASFS SLC-40 - 3 times
- Launch Sites CCASFS SLC-40 & CCASFS LC-40 (which has near to each other has the same number of succesfull launches as KSC LC-39A

**successful launch is succesful landing of rocket's first stage*



SpaceX Launch Records Dashboard

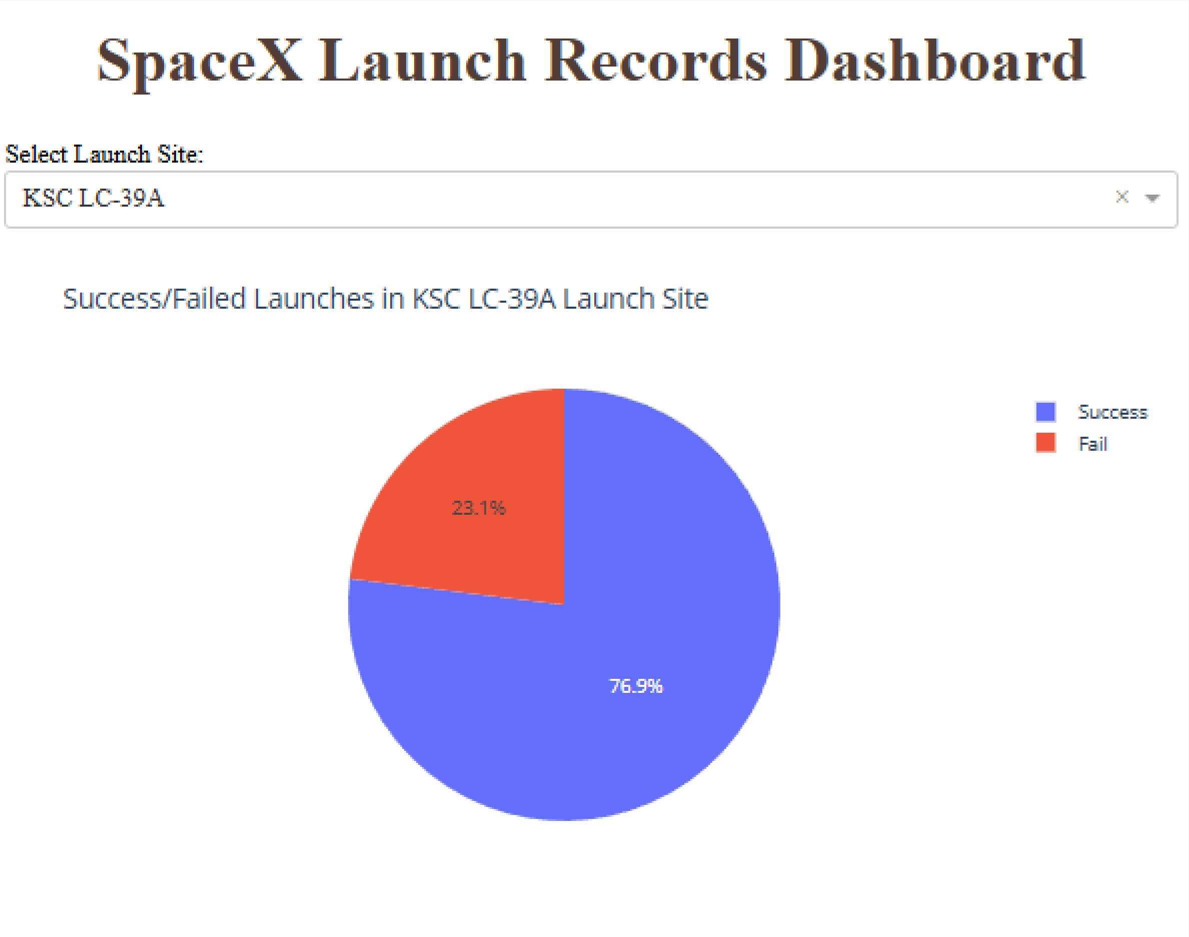
Select Launch Site:

All Sites

Highest launchess success ratio - KSC LC-39A site

- Successful launches are around 77% - 10 times
- Failed launches are around 23% - 3 times

**successful launch is succesful landing of rocket's first stage*

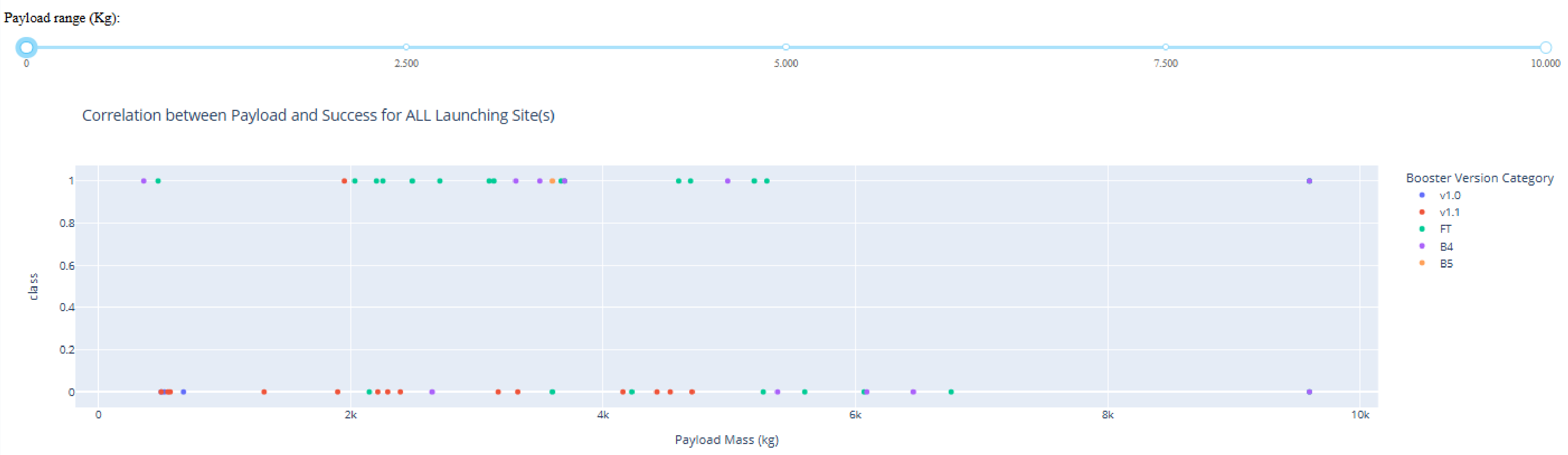


SpaceX Launch Records Dashboard

Select Launch Site: KSC LC-39A

Correlation between Payload Mass and Success

- The highest success rate is for payloads between 2 tones and 4 tones.
- The lowest success rate is for payloads higher then 6 tones - only two launches was successful - 9.600 kg with bossters B4 and FT
- The highest success ratio has booster B5- 100%, the 2ndhas FT, and 3rd has B4
- The lowest success ratio has booster v1.0 - 0%

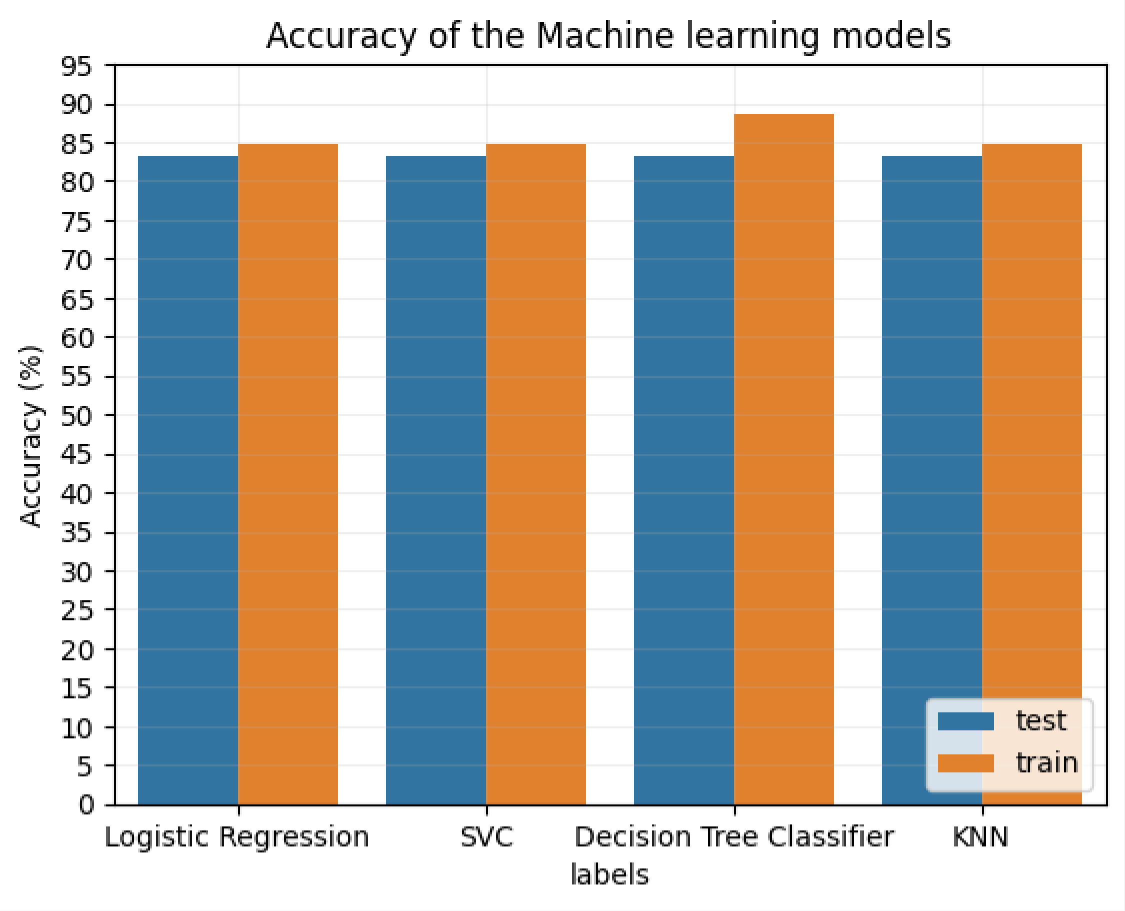


Section 5

Predictive Analysis (Classification)

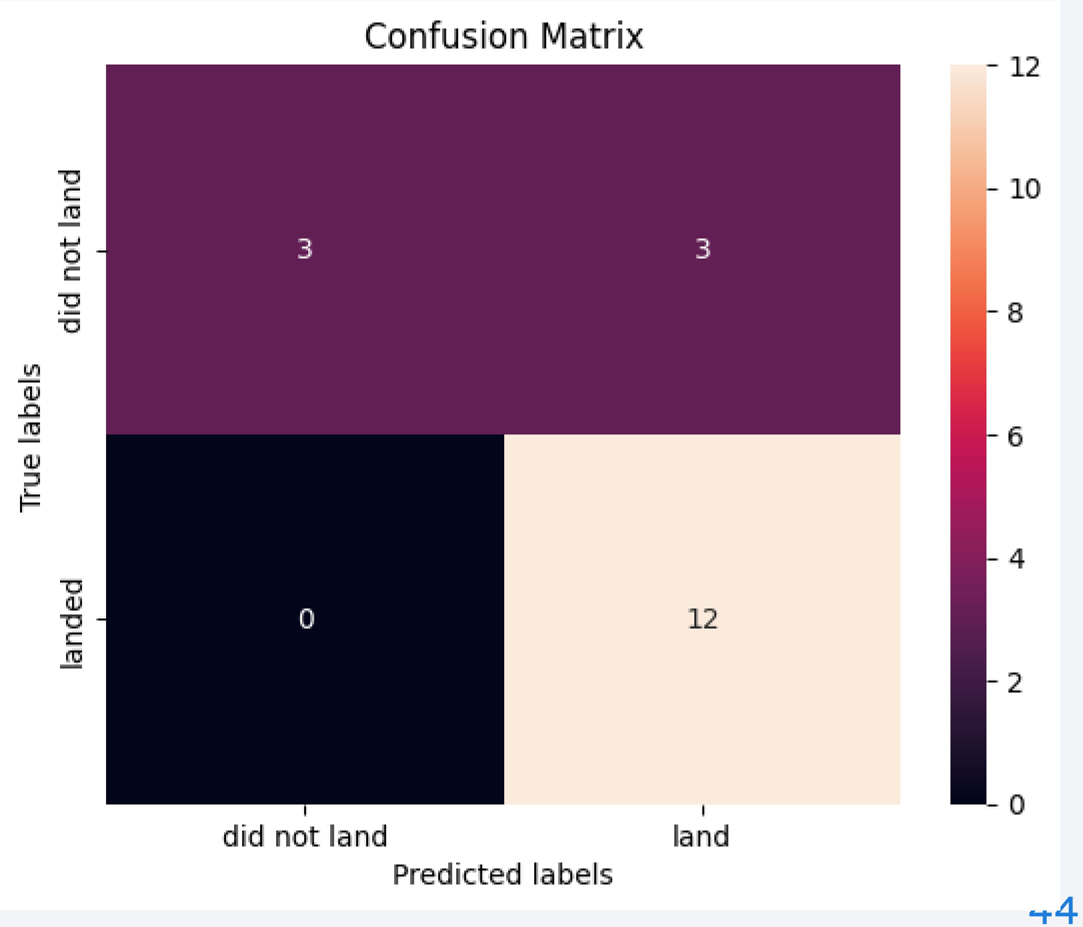
Classification Accuracy

- All models presents the same accuracy on testing data set
- The highest accuracy on traing data set data has Decision Tree Classifier
- Possible reason of the same accuracy on testng data set is too small data set designated for this purpose.



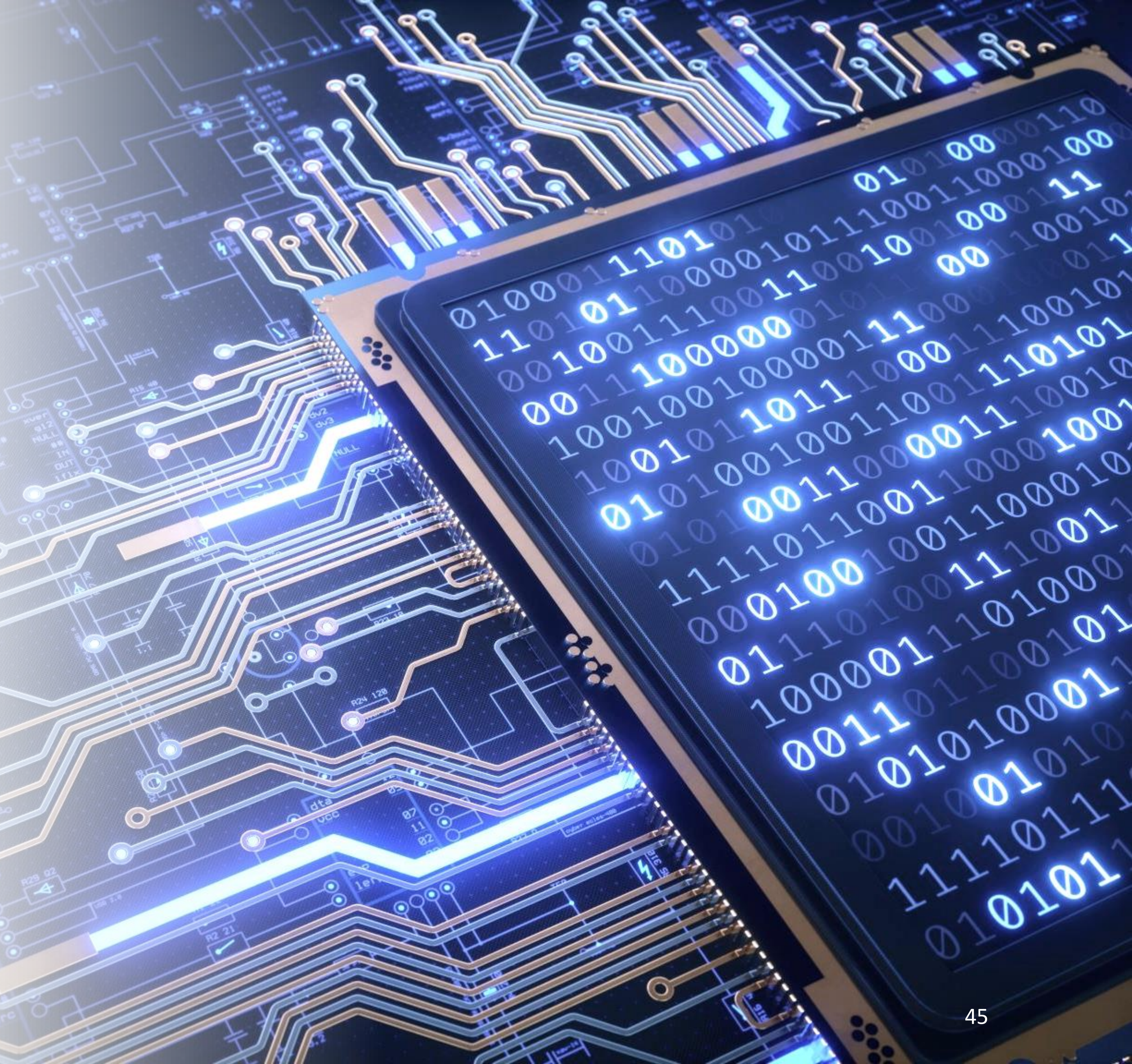
Confusion Matrix

- Confusion matrix is the same for every model
- The biggest problem are false possitive landings (3) - model predicted that they should land, but they did not.

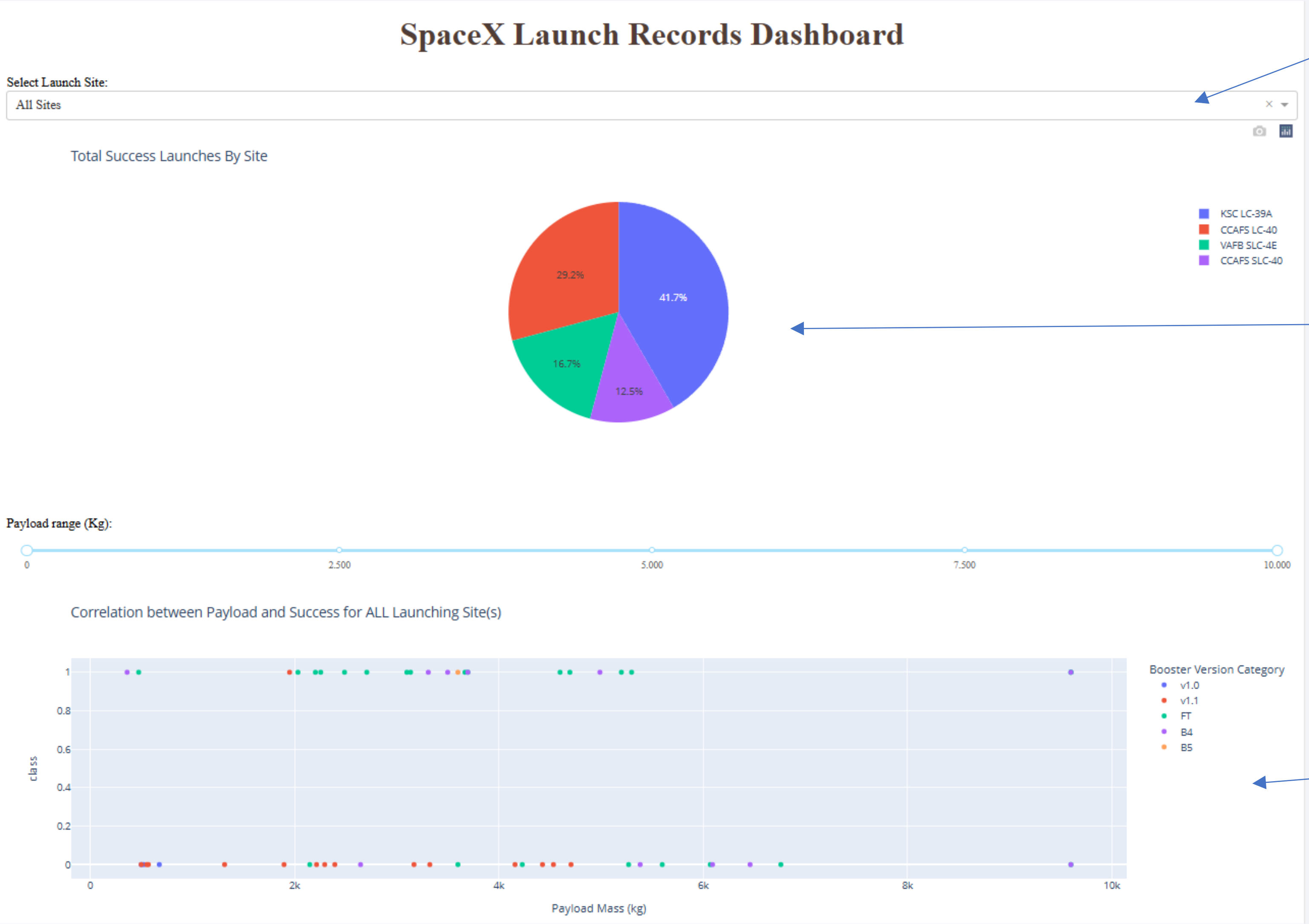


Conclusions

- All methods gave the same result on testing data set.
- The best accuracy on training data setis on Decision Tree Classifier model
- The biggest problem of the model are false positives predictions
- cross validation could be used to decide which model is the best
- Models still should be improved
- Additional data should be collected – another features and data about more launches



Appendix 1: Instruction how to use Dashboard



SpaceX Launch Records Dashboard

Select Launch Site:

All Sites

Total Success Launches By Site

Payload range (Kg):

Correlation between Payload and Success for ALL Launching Site(s)

Possibility to select Launching Site or All launching sites

Pie chart:

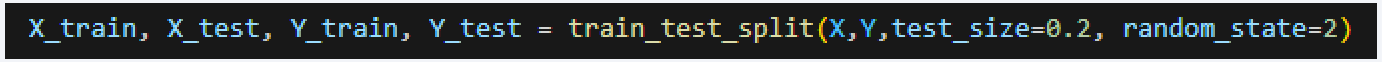
- *Shows successful launches divided per launhcing site - if "All Sites option have been selected"*
- *For selected site shows division on successed / failed launches*

Possibility to selectPayload mass interval, for which user would like to display data

Scatter plot which dhows correlation between payload mass & success for selected site or all sites and for selected payload mass on the slider.

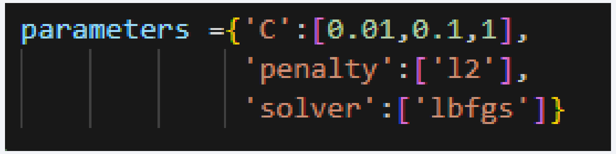
Appendix 2: Machine Learning Parameters

- Train test split:



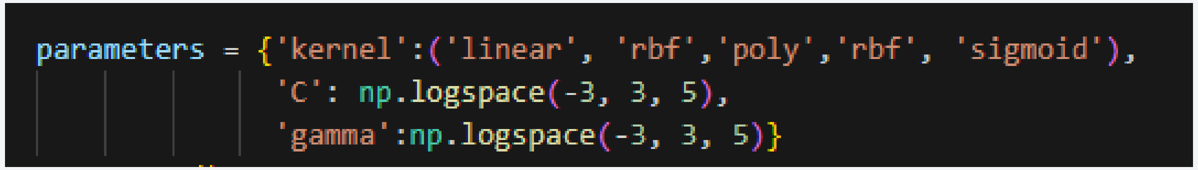
```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.2, random_state=2)
```

- Considered parameters for Logistic regression



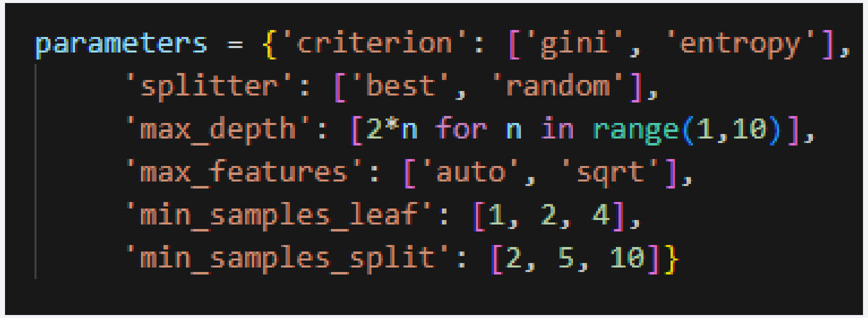
```
parameters ={'C':[0.01,0.1,1],
             'penalty':['l2'],
             'solver':['lbfgs']}
```

- Considered parameters for SVC



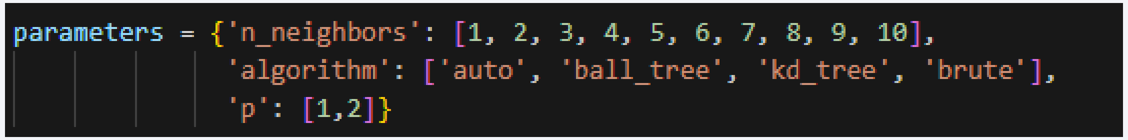
```
parameters = {'kernel':('linear', 'rbf','poly','rbf', 'sigmoid'),
              'C': np.logspace(-3, 3, 5),
              'gamma':np.logspace(-3, 3, 5)}
```

- Considered parameters for Decision Tree



```
parameters = {'criterion': ['gini', 'entropy'],
     'splitter': ['best', 'random'],
     'max_depth': [2*n for n in range(1,10)],
     'max_features': ['auto', 'sqrt'],
     'min_samples_leaf': [1, 2, 4],
     'min_samples_split': [2, 5, 10]}
```

- Considered parameters for KNN



```
parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
              'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
              'p': [1,2]}
```

Thank you!