IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Jose Angel Islas Anguiano>
<Feb 13 2023>

Outline

- Executive Summary
- Table of Contents
- Introduction
- Methodology
- Results
- Discussion
- Conclusion
- Appendix

Introduction

There are 4 different companies offering space travel services, Space X is the most successful and affordable because it can reuse the first stage of the flight. In this project we examine the data to find the features that impact the most a successful landing to determine the price of each launch of Falcon 9 and the probability of reuse.

Section 1

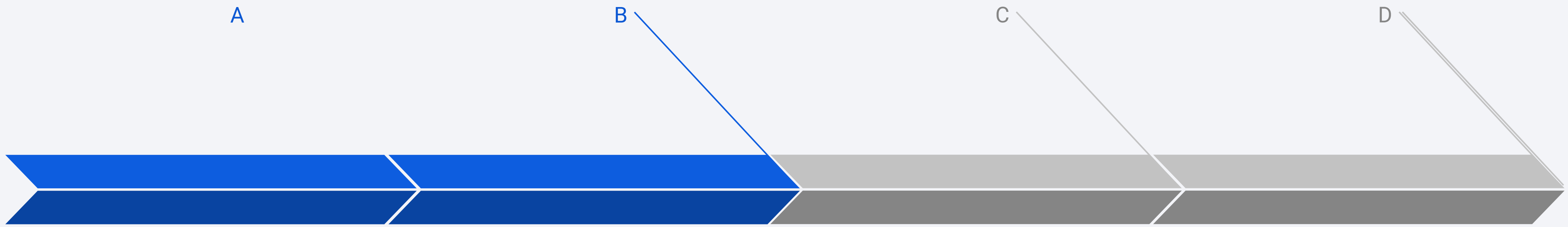
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was requested from SpaceX API and web scraping from Wikipedia
- Performed data wrangling
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models

Data Collection



Request launch data from SpaceX API

Convert Json file to a dataframe. Observe the massive and some unorganized data in the form of IDs.

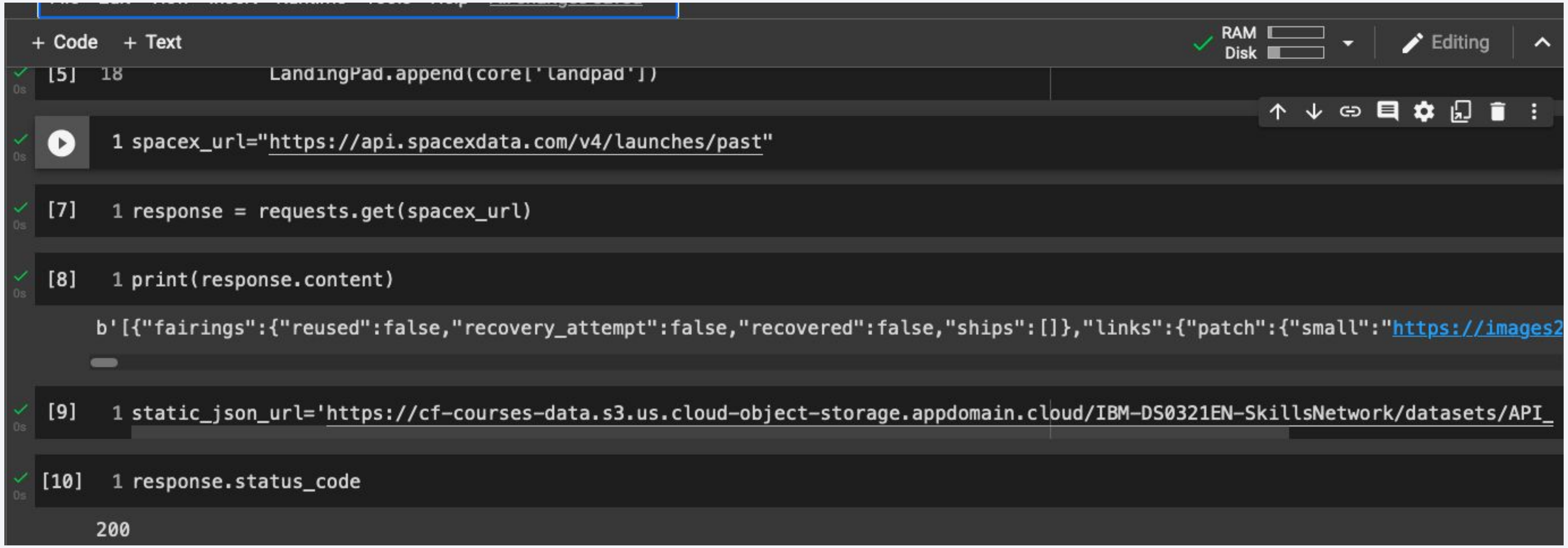
Use IDs to request from the AP data from the columns rocket, payloads, launchpad, and cores.

Create a subset of the data using the features rocket, payloads, launchpad, cores, flight number and date utc. Remove irrelevant rows.

The data was requested from the SpaceX API and converted to a dataframe for exploration. More requests were necessary and some cleaning and wrangling was performed.

Data Collection – SpaceX API

[Github link Data Collection]



Data Collection - SpaceX API

Use customized functions getBoosterVersion(), getLaunchSite(), getPayloadData(), getCoreData() in C to get specific features.

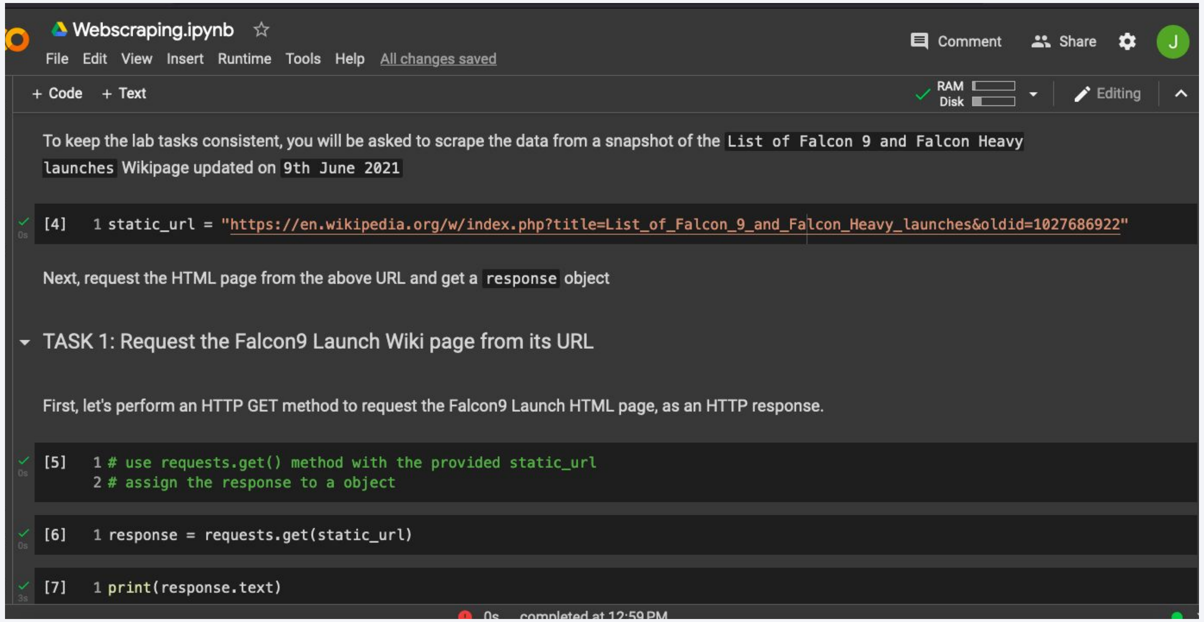
Finally merge the columns of C, D and F the dataframe to work with.

Features: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSire, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

After the API request the dataframe contains the columns: 'FlightNumber', 'Date', 'BoosterVersion', 'PayloadMass', 'Orbit', 'LaunchSite', 'Outcome', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial', 'Longitude', 'Latitude'.

Data Collection - Scraping

- Performed web scraping to obtain Falcon 9 launch records with BeautifulSoup
- Parsed the table and converted it into a pandas dataframe.



[Github link for web scraping]

Data Collection Wrangling

Performed basic wrangling and the most important was to create an Outcome column which assigns 0 if bad landing and 1 if good landing were recorded.

The number of launches at each site, and the number and occurrence of each orbits were calculated as well.

[Github link Data Wrangling]

In [ ]:
```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
bad_outcomes
```

Out[ ]: {'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'}

TASK 4: Create a landing outcome label from Outcome column

Using the `Outcome`, create a list where the element is zero if the corresponding row in `Outcome` is in the set `bad_outcome`; otherwise, it's one. Then assign it to the variable `landing_class`:

In [ ]:
```
def comp(list1, list2):
    for val in list1:
        if val in list2:
            return True
    return False
```

In [ ]:
```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise

landing_class = [1] * 90
i=-1

for val in df["Outcome"]:
    i=i+1
    if val in bad_outcomes:
      landing_class[i]=0
```

This variable will represent the classification variable that represents the outcome of each launch. If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully

In [ ]:
```
df['Class']=landing_class
df[['Class']].head(8)
```

Out[ ]:

	Class
0	0

EDA with Data Visualization

- Explored the data by visualizing the relationship between flight number and launch site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend. These pairs have been select to identify patterns which may affect the outcome of a launch.
- EDA link

EDA with SQL

Performed SQL queries, these are specified in the following section 2.

Github link file

Build an Interactive Map with Folium

Assigned the feature launch outcomes (failure or success) to class 0 or 1: 0 for failure, and 1 for success.

Color-labeled marker clusters identify which launch sites have relatively high success rate or low success rate..

Calculated the distances from a launch site to landmarks railways, coastlines, cities.

[Github link Map with Folium]()

Build a Dashboard with Plotly Dash

Created an interactive dashboard with Plotly dash, please go to section 4.

[SpaceX Dashboard link]

Predictive Analysis (Classification)

Data loaded and normalized using numpy and pandas, then it was split into training and testing data sets.

The machine learning algorithms: Logistic regression, decision tree, SVM,and KNN were used and compared for prediction.

- Github link for classification

Results

Exploratory data analysis

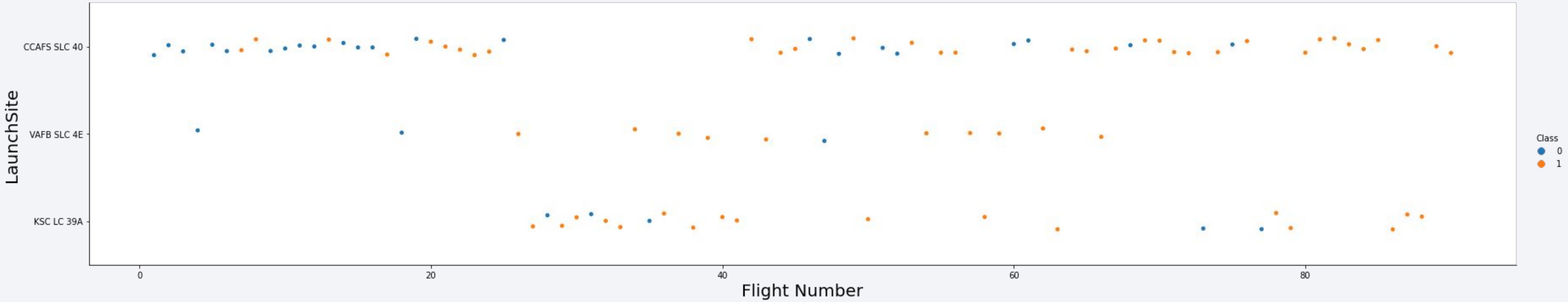
Interactive analytics in screenshots

Predictive analysis (Classification)

Section 2

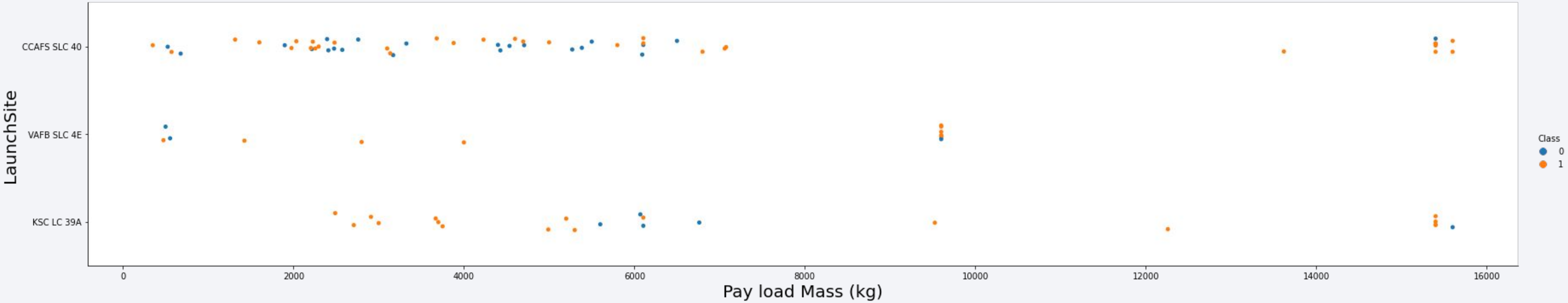
Insights drawn from EDA

Flight Number vs. Launch Site



As the flight number increases, the success is more likely at a specific launch site.

Payload vs. Launch Site

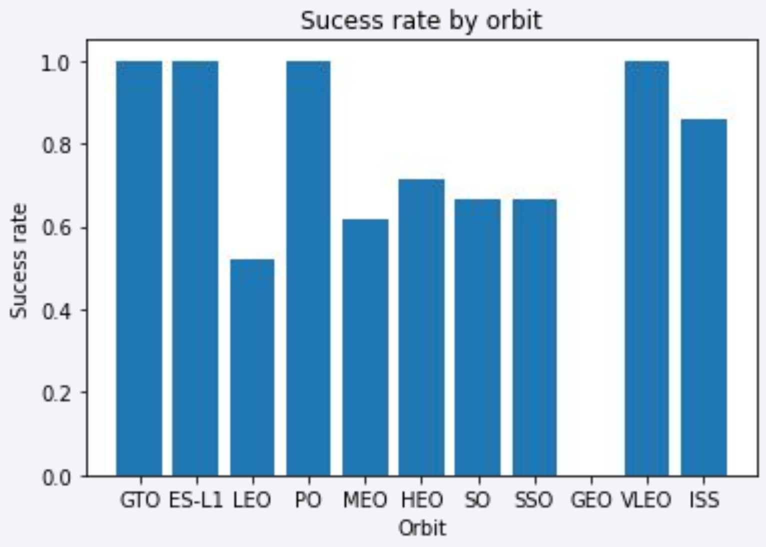


KSC LC 39A has similar more successes for light and heavy Pay loads.

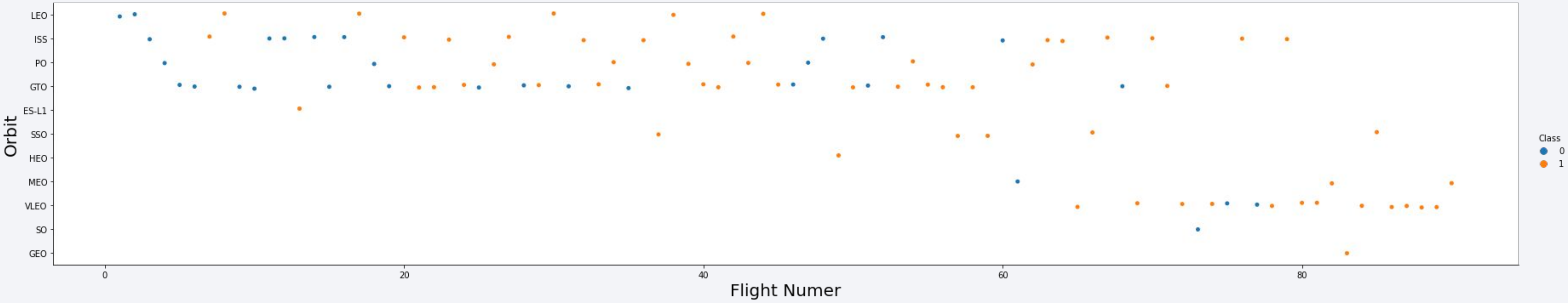
CCAFS SLC 40 clearly performs better for heavy Pay loads.

Success Rate vs. Orbit Type

The orbits GTO, ES-L1, PO, and VLEO have the most success rate.



Flight Number vs. Orbit Type



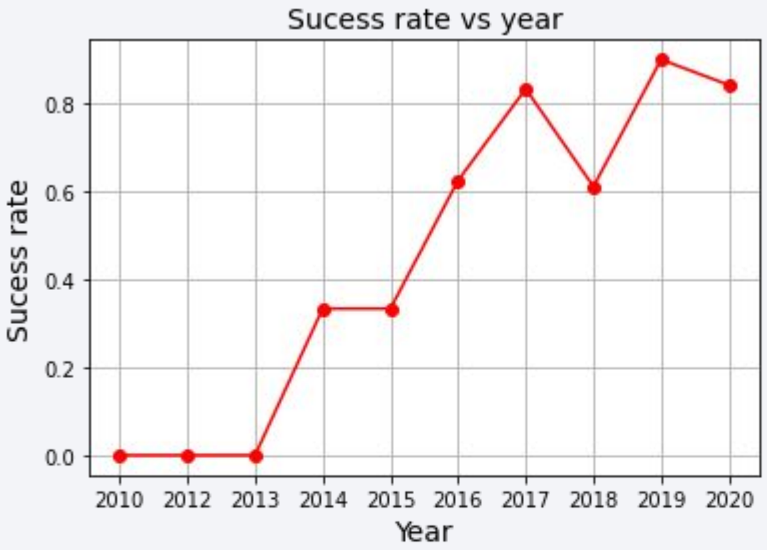
Most of the first flights were failures while most of the last flights were successes. The orbits LEO, ISS, PO, and GTO improved after flight number 20.

Payload vs. Orbit Type

The orbits ISS and PO perform very good after 8000KG while VLEO does not.

Launch Success Yearly Trend

The success rate increases from 2013 until 2020.



All Launch Site Names

DISTINCT is used to show the names of all the unique launch sites

```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total payload carried by boosters from NASA

```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

Date of the first successful landing outcome on ground pad

```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 and less than 6000

```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes

```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

```
The total number of successful mission outcome is:
```

	successoutcome
0	100

```
The total number of failed mission outcome is:
```

	failureoutcome
0	1

Boosters Carried Maximum Payload

List of the names of the booster which have carried the maximum payload mass

```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

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

```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Rank of 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

```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

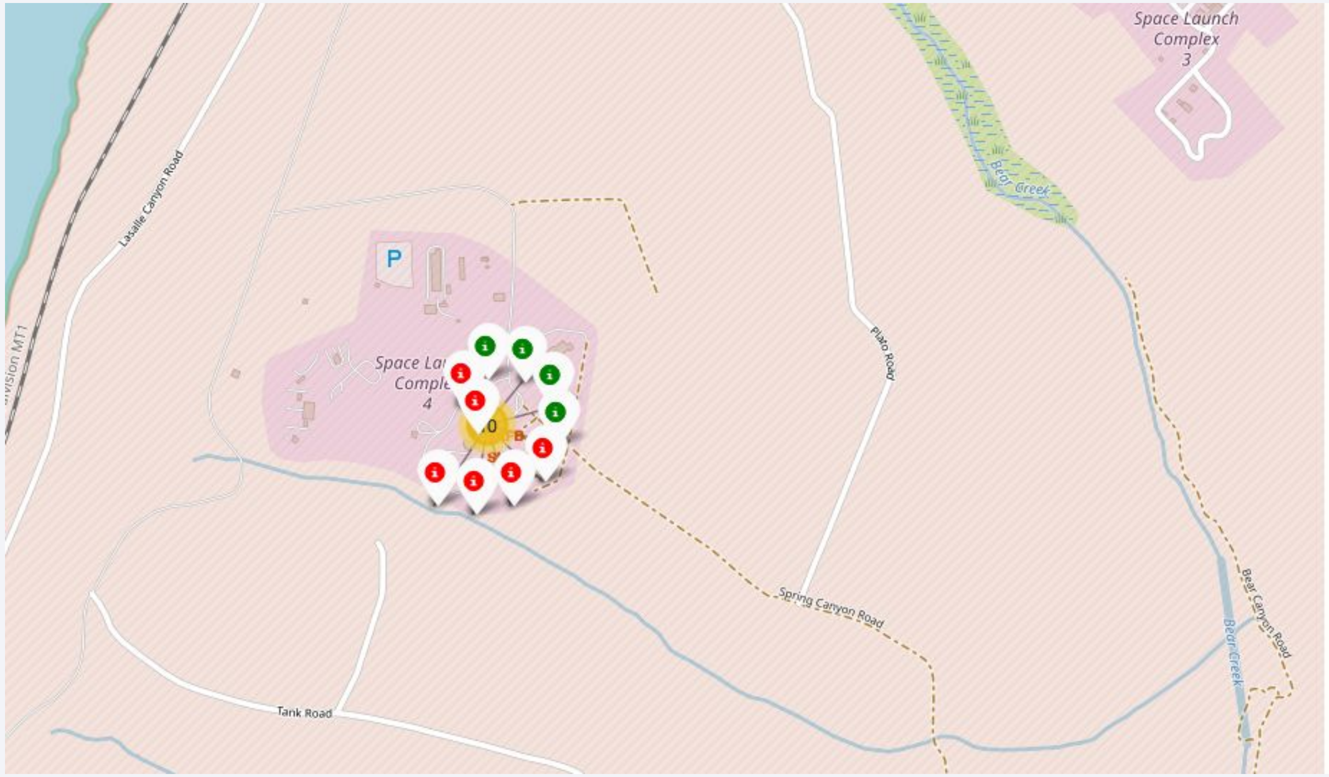
Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites

California and Florida launch sites.

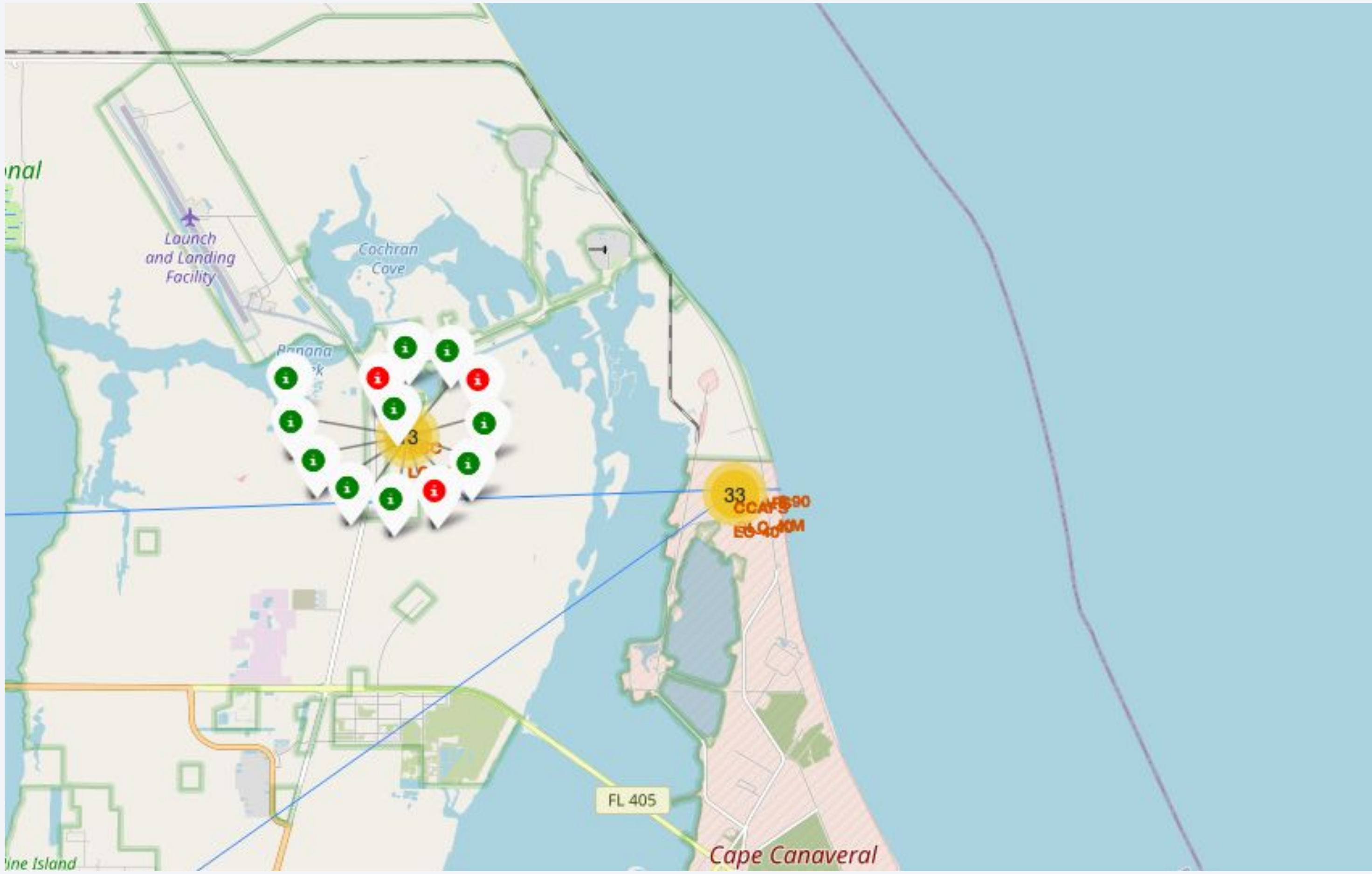
Launch sites by outcome

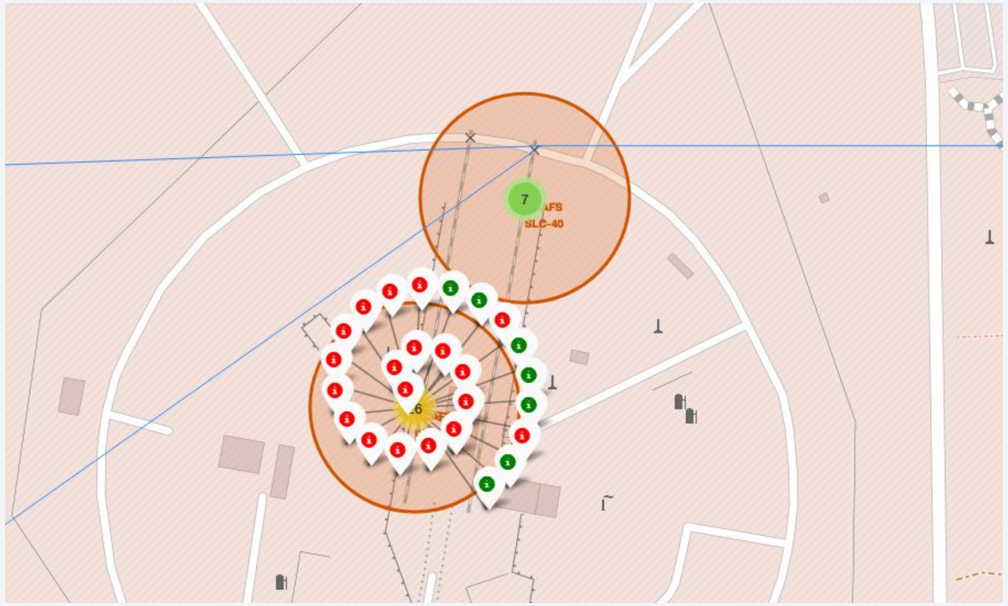


California fails and successes

Launch sites by outcome

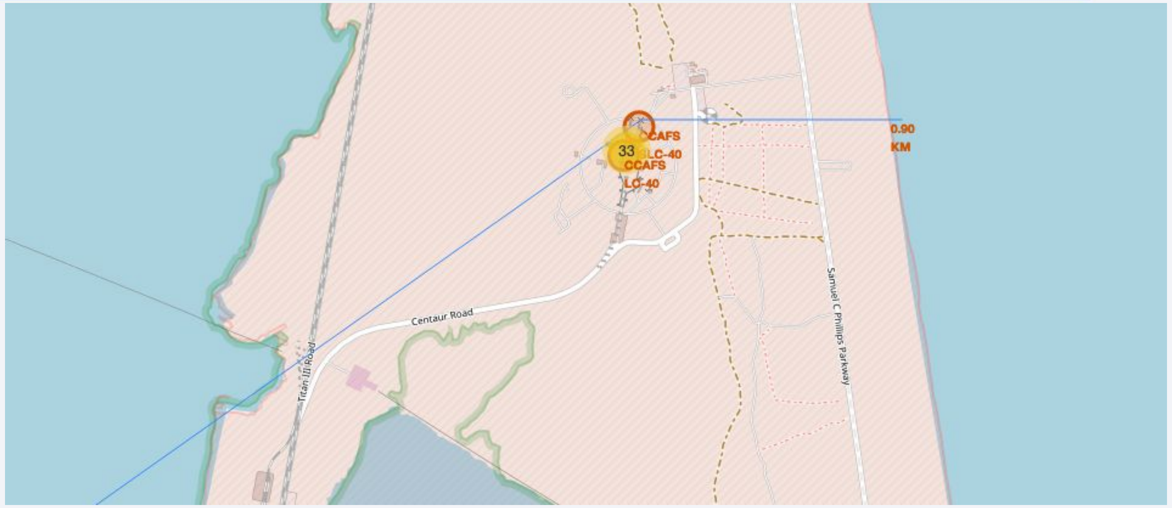
Florida fails and successes launches

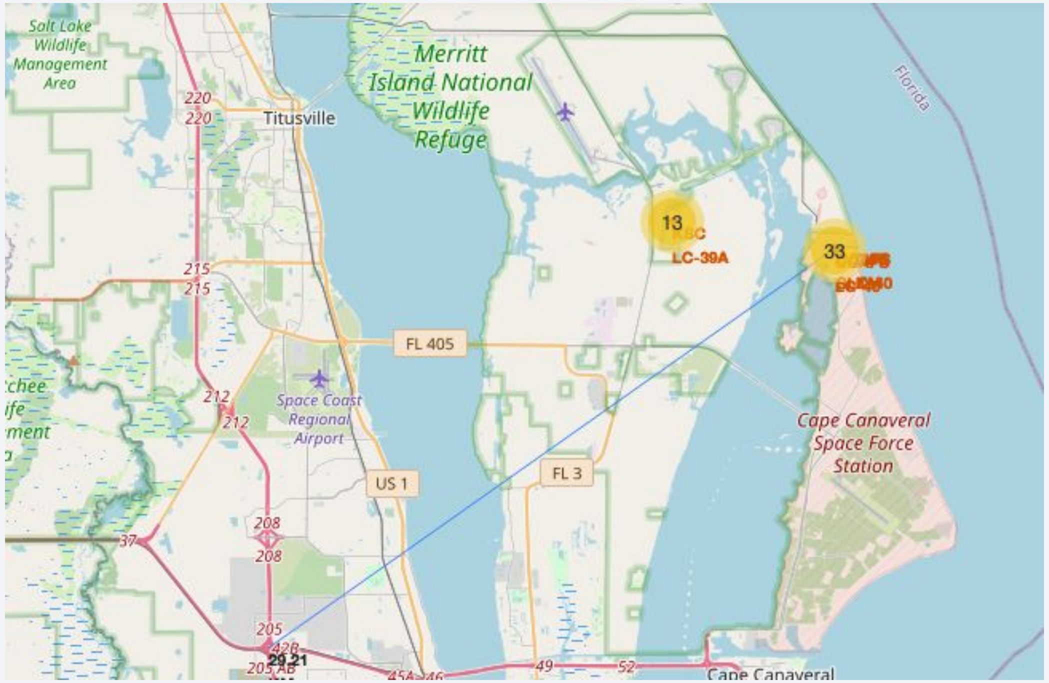




Launch Site distance to landmarks

Distance to sea



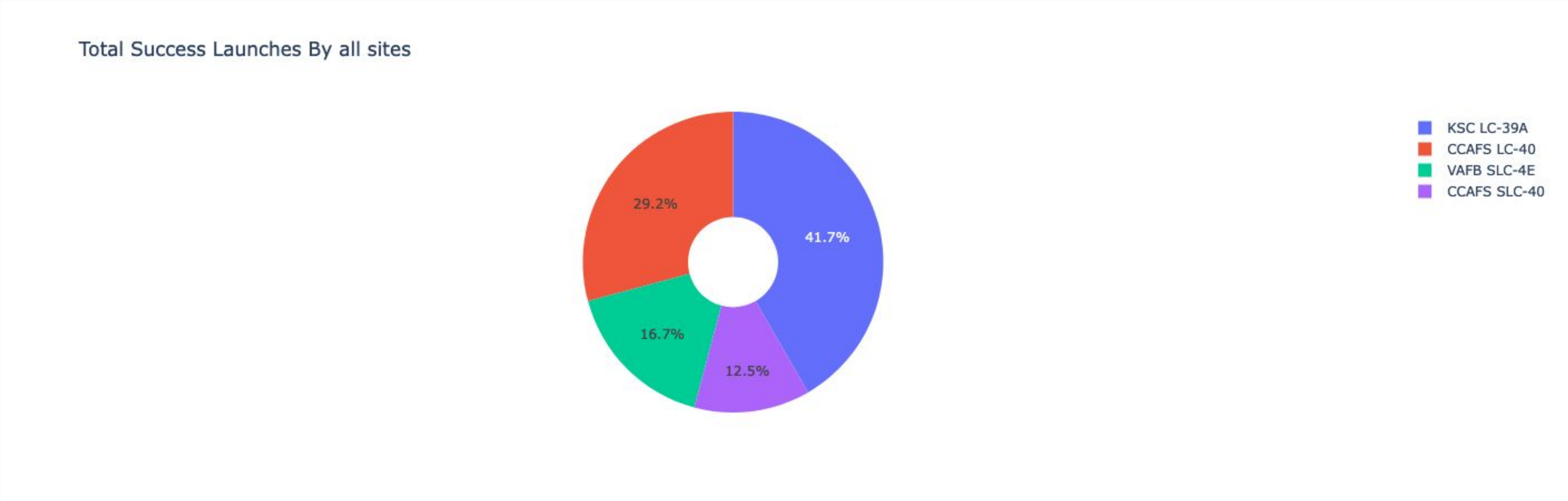


Distance to city

Section 4

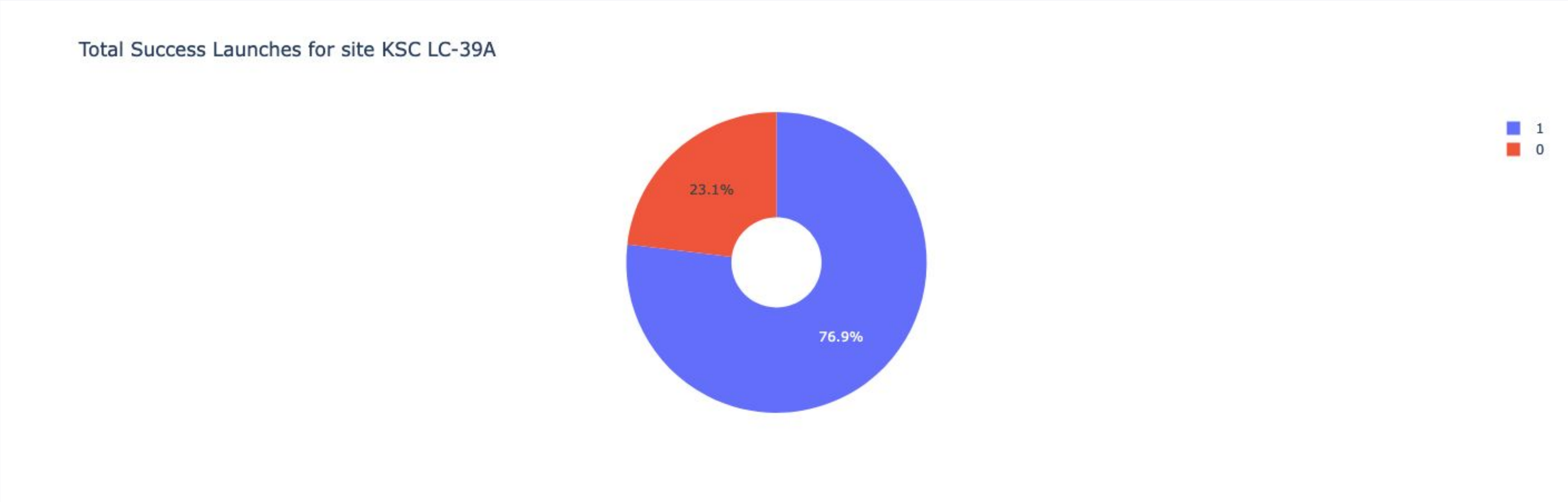
Build a Dashboard with Plotly Dash

Success launches by site



KSC LC-39A has the most successful launches and CCAFS SLC-40 the lowest.

Best success ratio

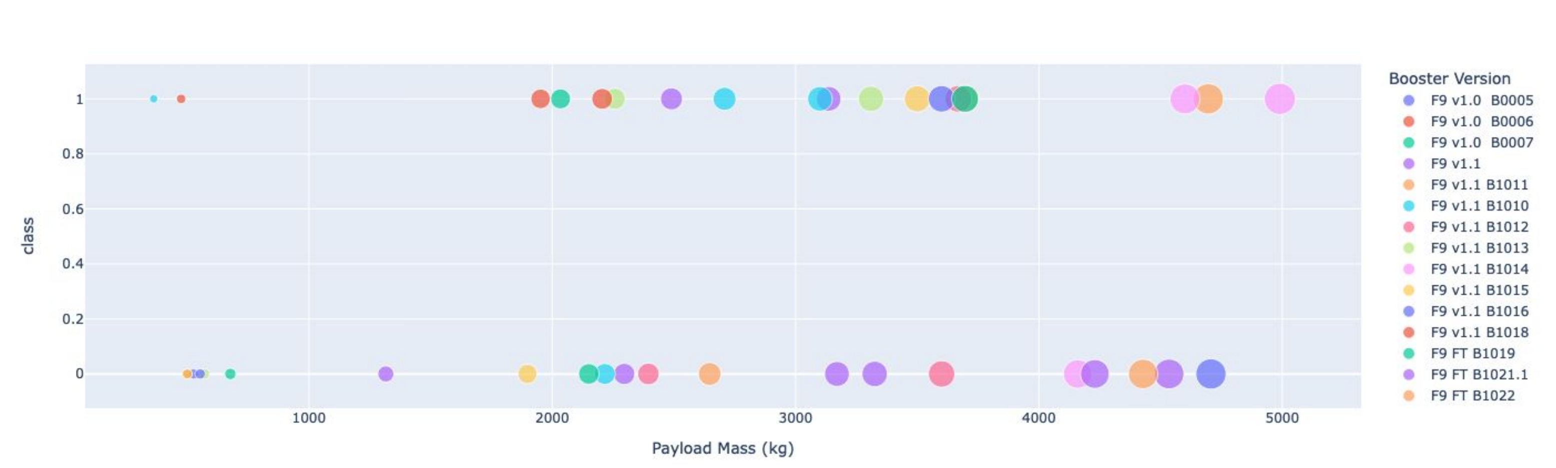


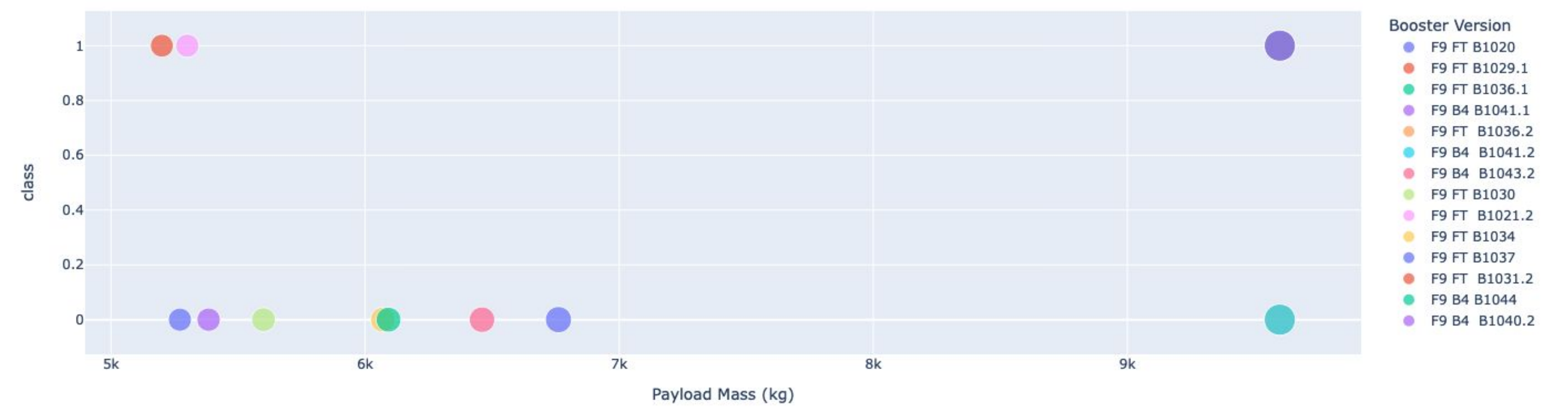
KSC LC-39A has the best success ratio

Payload vs Launch Outcome for all sites

Successes by Payload Mass from 365 Kg to 4990 Kg

Successes by Payload Mass from 5200 Kg to 9600 Kg. Failures dominates the outcome.



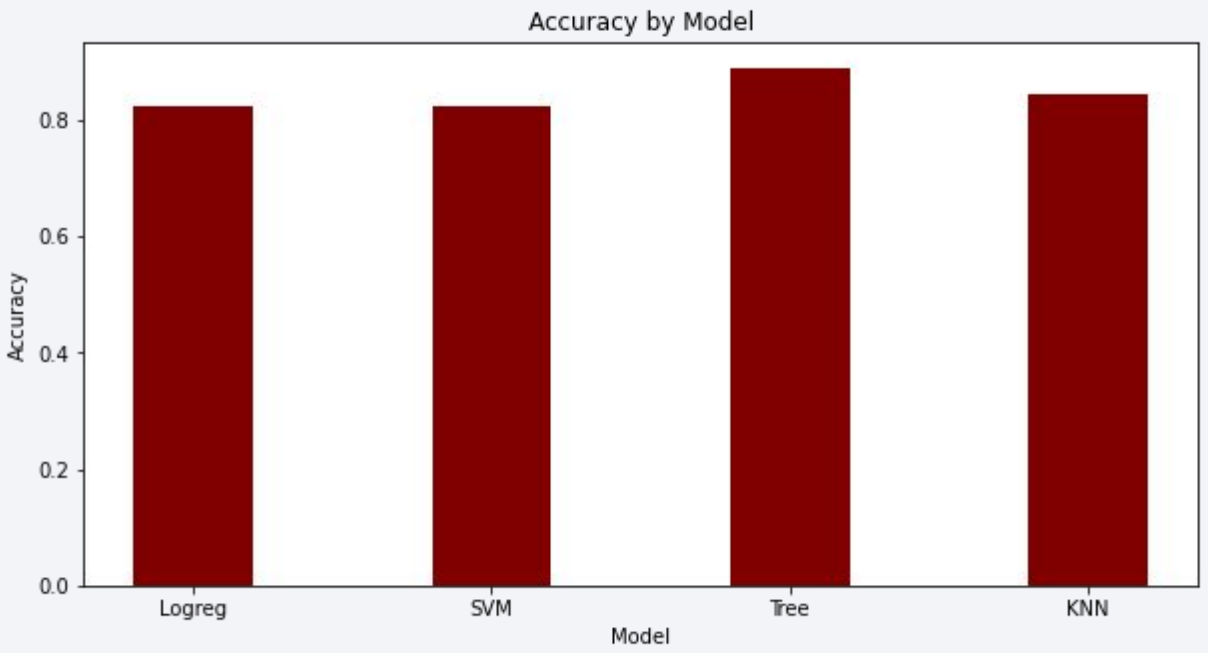


A heavier payload tends to fail more than lighter.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



The best classification model is Decision Tree method with 0.88 score of accuracy.

Confusion Matrix

- The matrix shows that the tree classifier can distinguish between the different classes. The major problem is the false positives, this is an unsuccessful landing marked as successful by the classifier.

Conclusions

- A heavier payload increases the chances of failure.
- Launch success rate started to increase steadily in 2013.
- KSC LC-39A has the most successful launch ratio among all the launch sites.
- Decision tree classifier is the best predictive machine learning algorithm for this project.

Thank you!