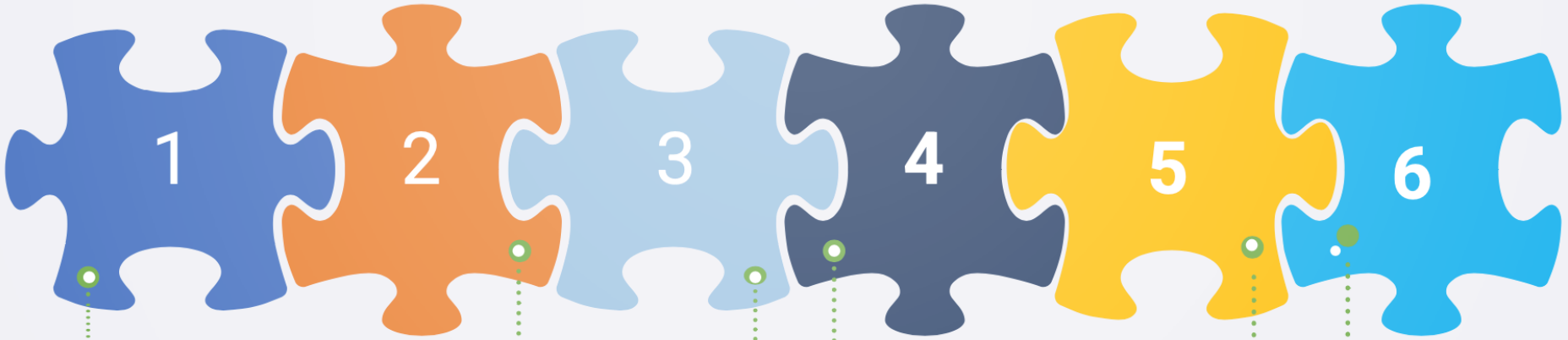
WINNING SPACE RACE WITH DATA SCIENCE

IULIIA SHIPALOVA

04.06.2024

Outline

Executive Summary



Collection data from API and web page

Data Wrangling: transforming and mapping data

Exploratory data analysis (EDA) by using SQL

EDA by visualization

Interactive visual analytics and dashboards

Predictive analysis by machine learning

Summary: Select the best model for future prediction.

Introduction

SpaceX's goal:

- Sending spacecraft to the International Space Station.
- Starlink, a satellite internet constellation providing satellite Internet access.
- Sending manned missions to Space.

Object	Spaces X's Falcon 9 launch
Goal	To determine the price of each launch.
Reason	If we can determine if the first stage will land, we can determine the cost of a launch
Method	By gathering information about Space X and creating dashboards: o To determine if SpaceX will reuse the first stage o Train a machine learning model and use public information to predict if SpaceX will reuse the first stage

Methodology

- Data collection methodology:
  - Require the data from SpaceX API
  - Collect data from a Wikipedia page
- Perform data wrangling
  - Perform EDA to find some patterns
  - Determine what would be the label for training supervised model

Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Compare logistic regression model, support vector machine tree decision classifier, KNN by using GridSearchCV to select the best fit model

METHODOLOGY

SECTION 1

Methodology

- Data collection methodology:
  - Require the data from SpaceX API
  - Collect data from a Wikipedia page
- Perform data wrangling
  - Perform EDA to find some patterns
  - Determine what would be the label for training supervised model

Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Compare logistic regression model, support vector machine tree decision classifier, KNN by using GridSearchCV to select the best fit model

Data Collection

1 API

spacex_url="https://api.spacexdata.com/v4/launches/past"

- Required the data from Space API
- Clean the data

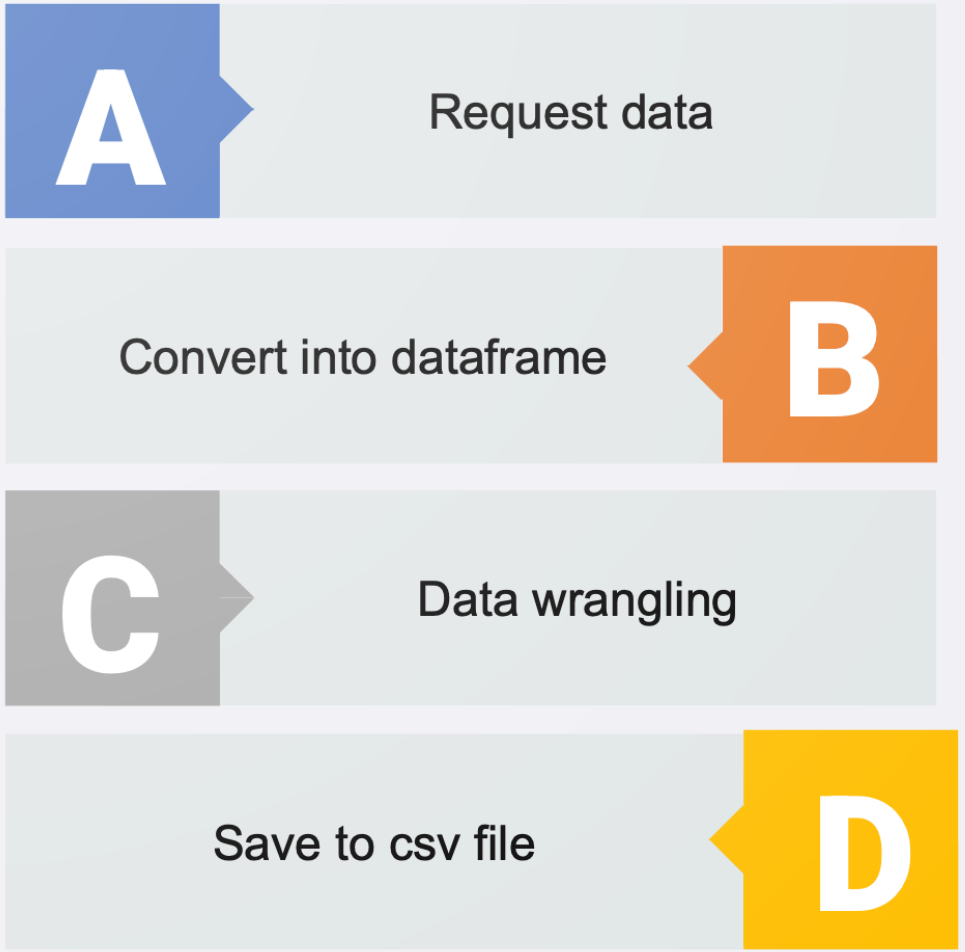
2 Web page

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

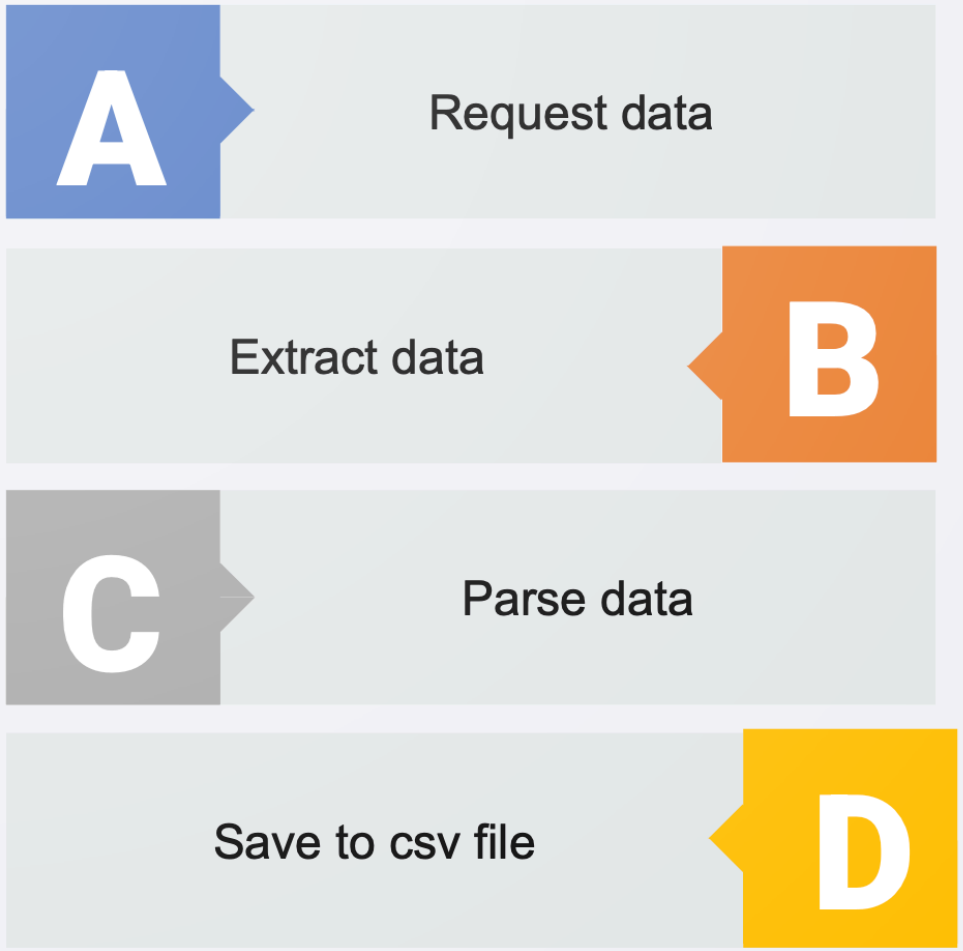
Data Collection – SpaceX API

- Request data from SpaceX API
- Convert the json result into a dataframe
- Filter dataframe to only Falcon 9` launches and data wrangling
- Export to csv



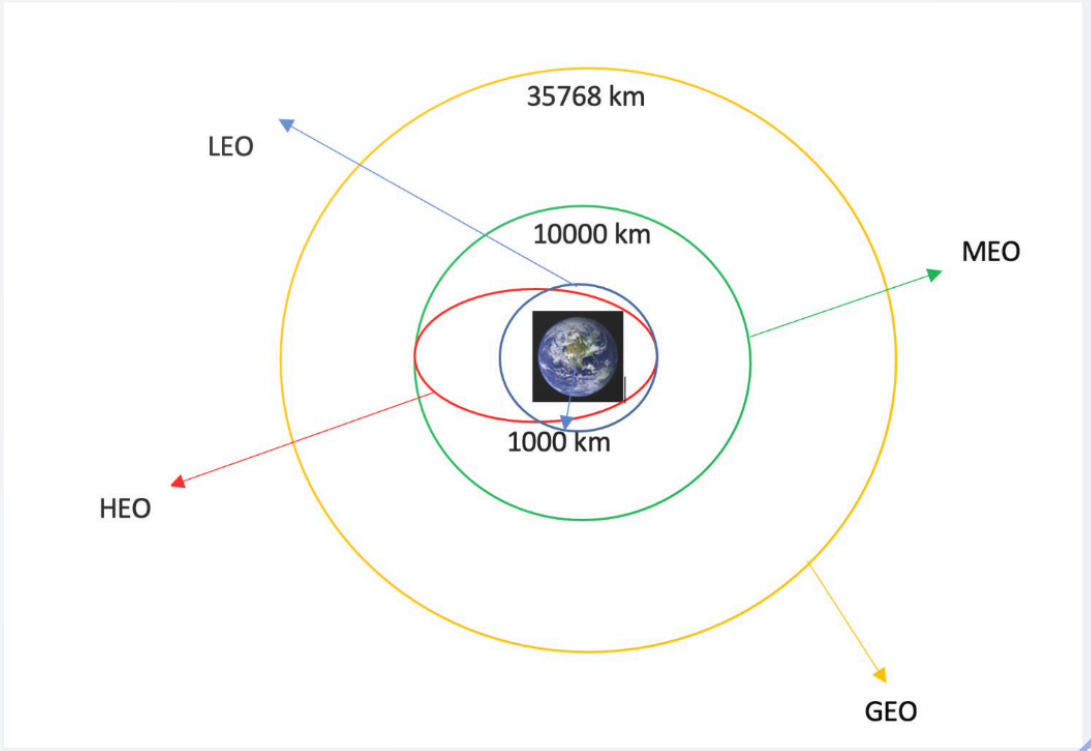
Data Collection – Scraping

- Request the Falcon9 Launch Wiki page from its URL
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables
- Export to csv



Data Wrangling

- Specify the missing value
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of each orbit
- Create a landing outcome label from Outcome column



EDA with Data Visualization

- The relationship between Flight Number and Launch Site -> scatter plot
- The relationship between Payload and Launch Site -> scatter plot
- The relationship between success rate of each orbit type -> bar plot
- The relationship between Flight Number and Orbit type -> scatter plot
- The relationship between Payload and Orbit type -> scatter plot
- The launch success yearly trend -> line chart

- The scatter plot is the best to describle the relation between two categorical data
- The bar plot is the best to compare several categorical data
- The line plot is the best to show the time series data

EDA with SQL

- 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 successful landing outcome in ground pad was achieved.
- 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.
- 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 successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Build an Interactive Map with Folium

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities
  - Whether it is close to the coast
  - Whether it is close to the railway
  - Whether it is close to the highway
  - Whether it is close to the city

- **To find some geographical patterns about launch sites**

Build a Dashboard with Plotly Dash

- A launch site drop-down input component
- A success-pie-chart based on the selected site dropdown
- A range slicer to select payload
- A success-payload-scatter-chart scatter plot based on the selected site dropdown

- **To inspect the relationship of success rate between lauch site and payload**

Predictive Analysis (Classification)

Data wrangling

Data standarization

Split into traning and test datasets

Predictive model evalutation

Predictive model selection

https://github.com/YuliyaShe/project_learning/blob/main/Space_X.ipynb

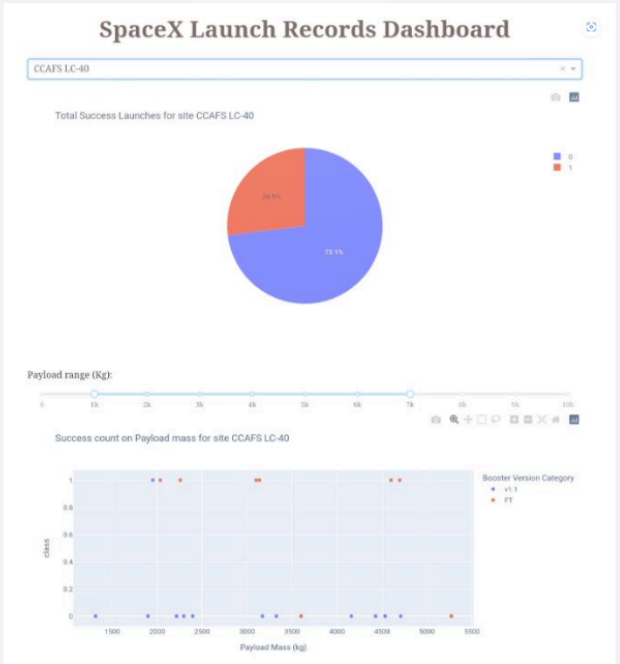
- Logistic regression
- Support vector machine
- Decision tree classifer
- K-nearest neighnors
- K-nearest neighnors

Results

EDA

- KSC LC-39A and VAFB SLC 4E has a success rate of 77%
- VAFB SLC 4E has no payload above 10000 kg
- In the LEO orbit the Success appears related to the number of flights
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS
- The sucess rate since 2013 kept increasing till 2020

Interactive analytics



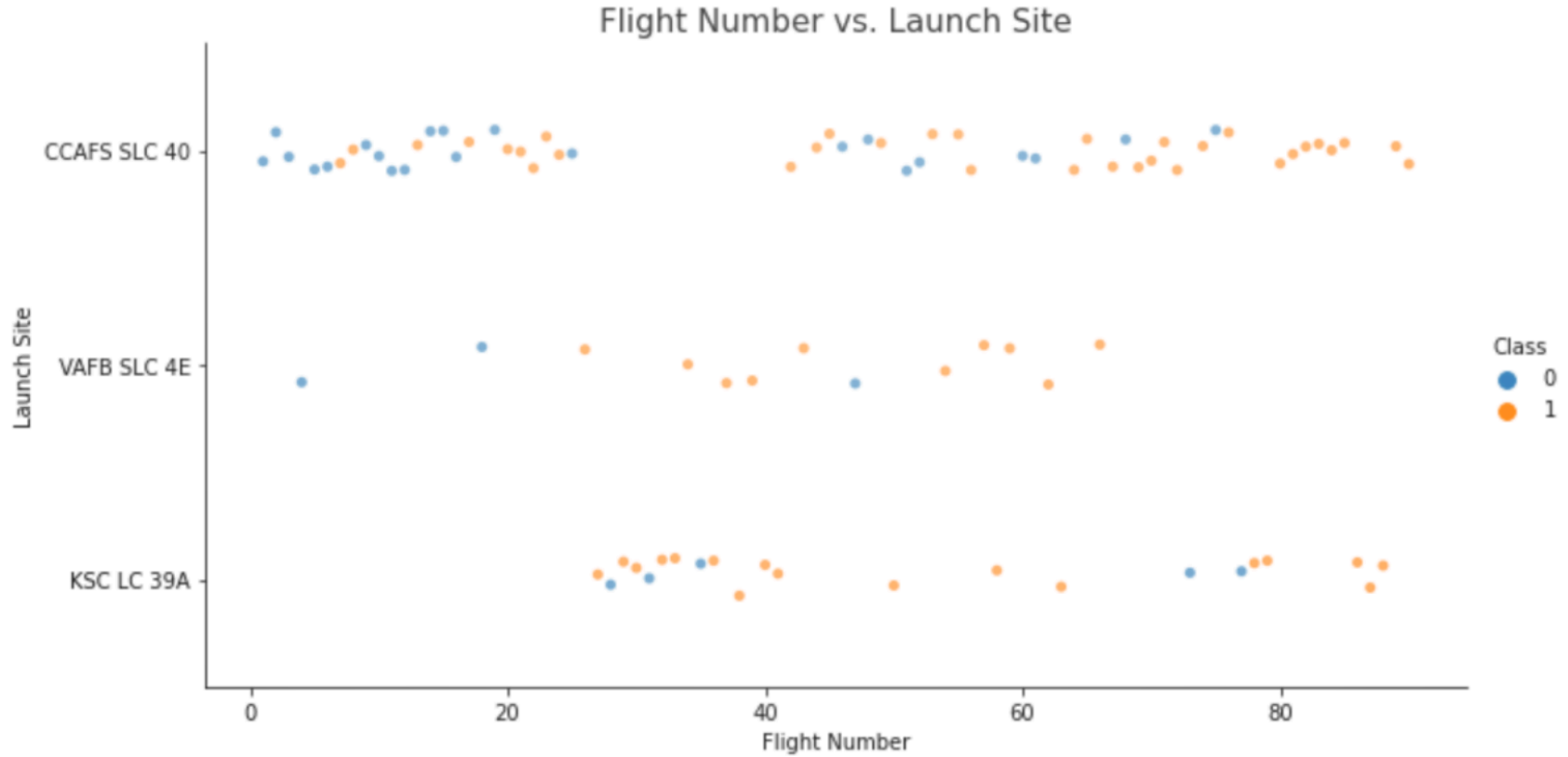
Predictive analysis

According to the decision tree classifier model, the predictive model tells us that there will be 4 ture postive, 7 true negative, 5 false postive and 2 false negative . The accuracy of the mode is around 89% with the best parameters.

INSIGHTS DRAWN FROM EDA

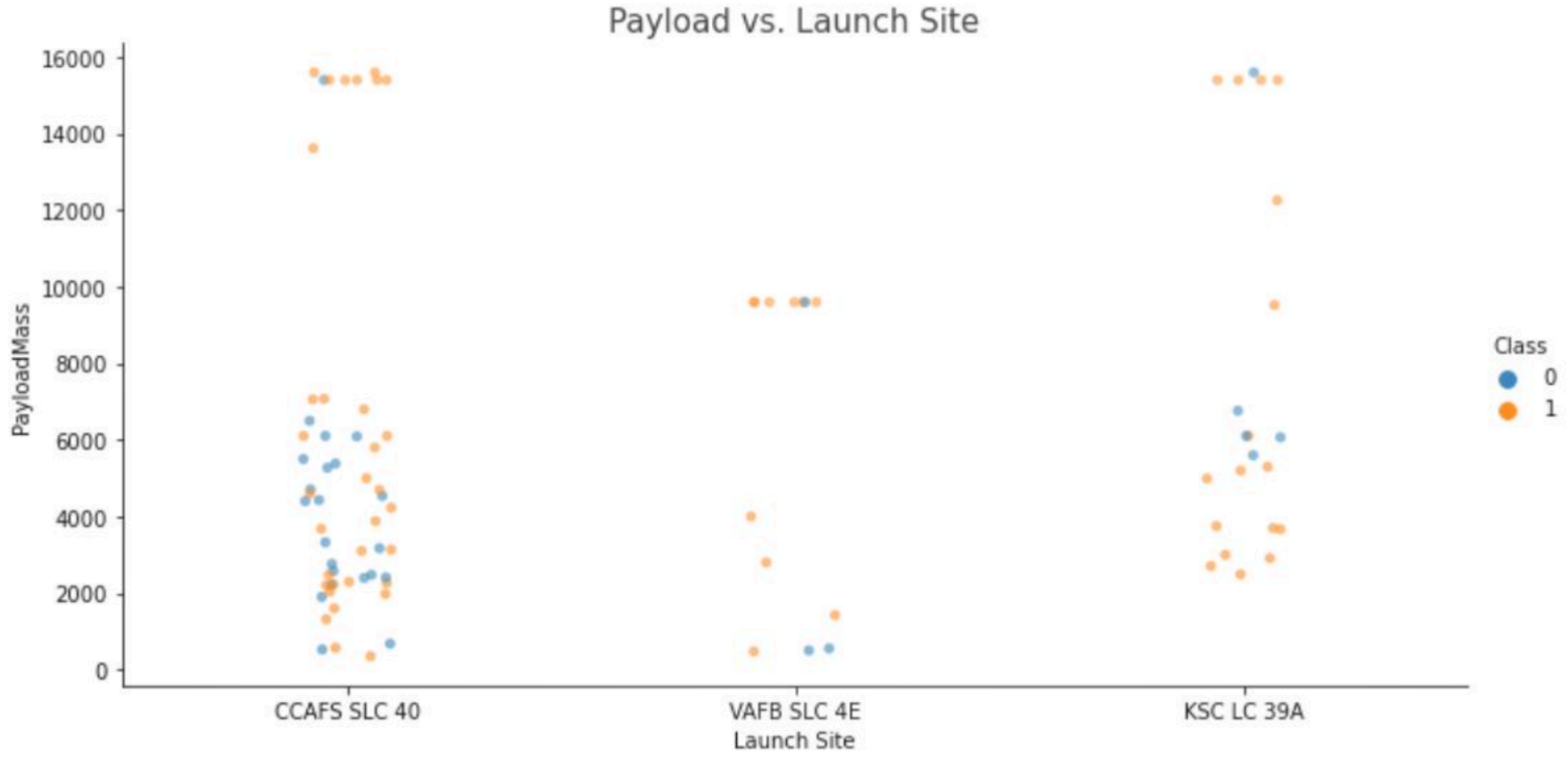
SECTION 2

Flight Number vs. Launch Site



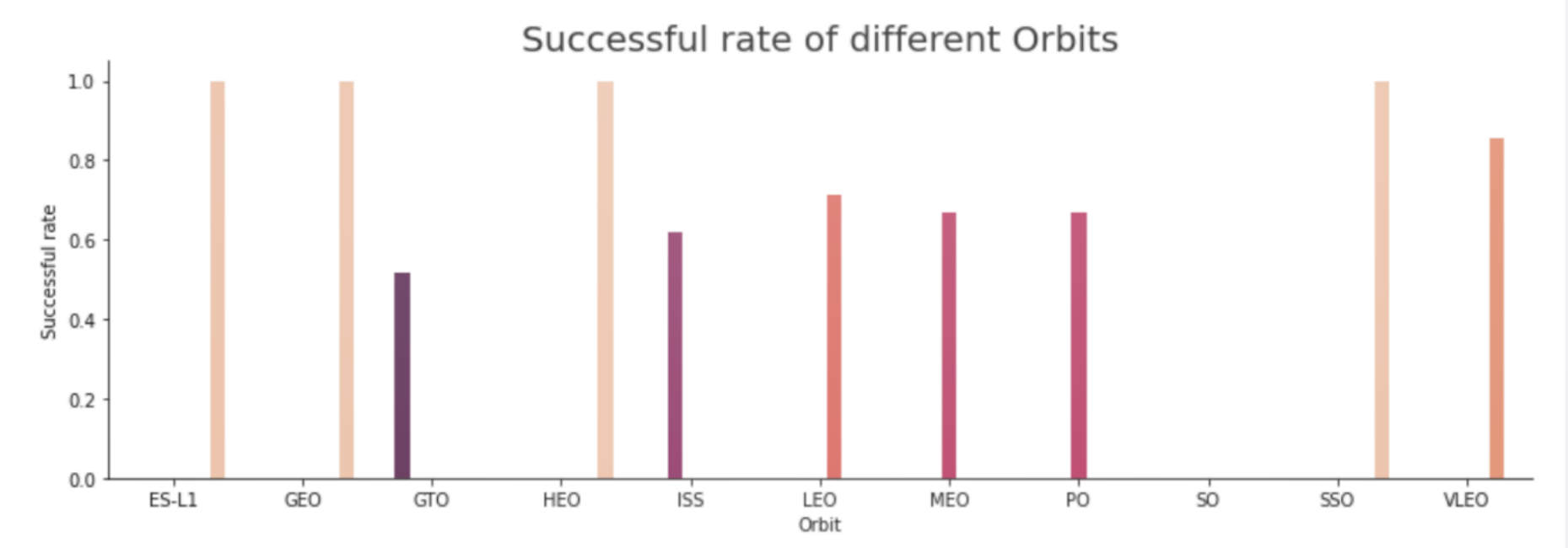
-> *KSC LC 39A has the highest sucessful rate, vice verce CCAFS SLC 40 has the lowest*

Payload vs. Launch Site



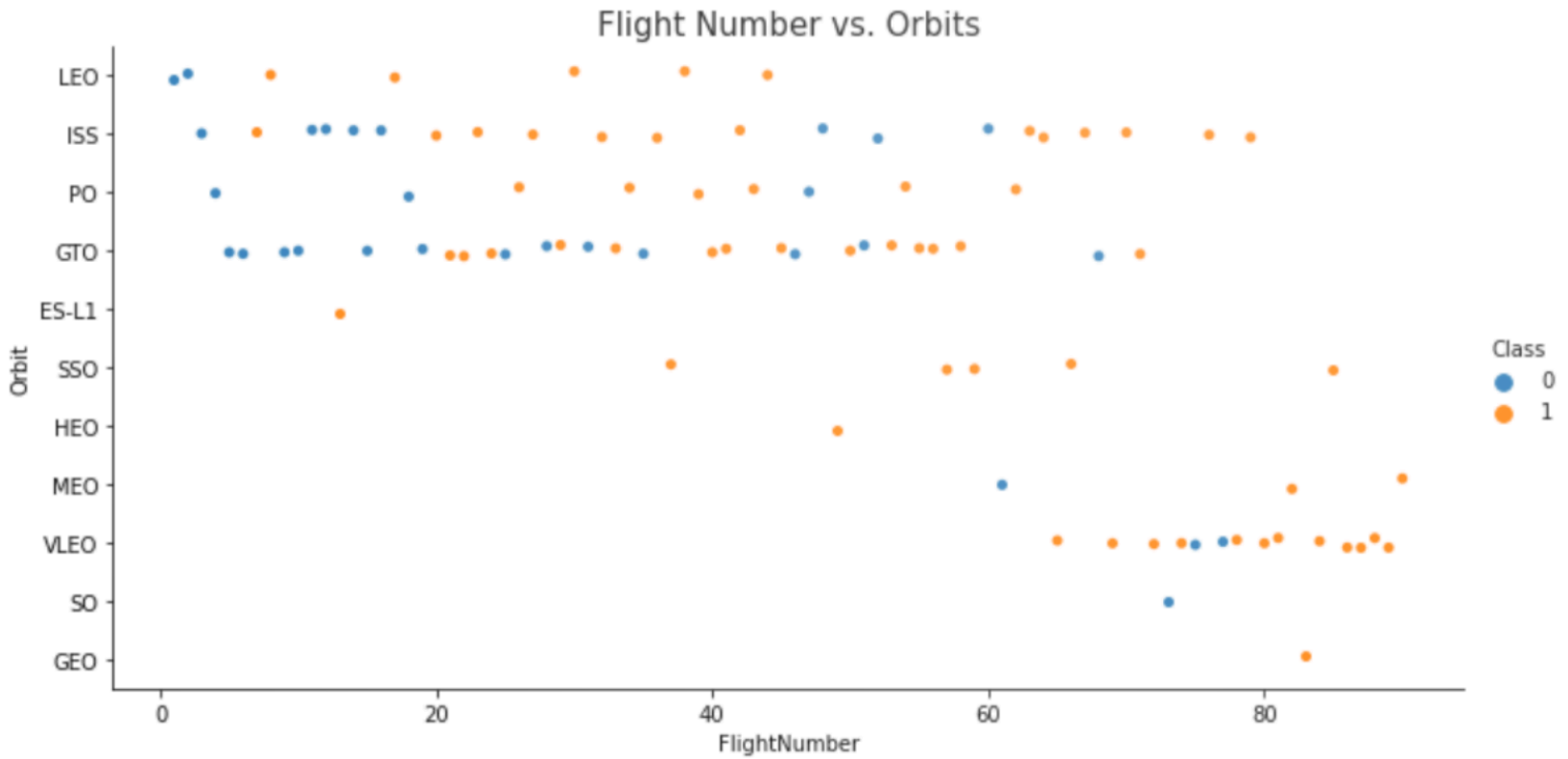
-> *VAFB SLC 4E has no payload above 10000 kg*

Success Rate vs. Orbit Type



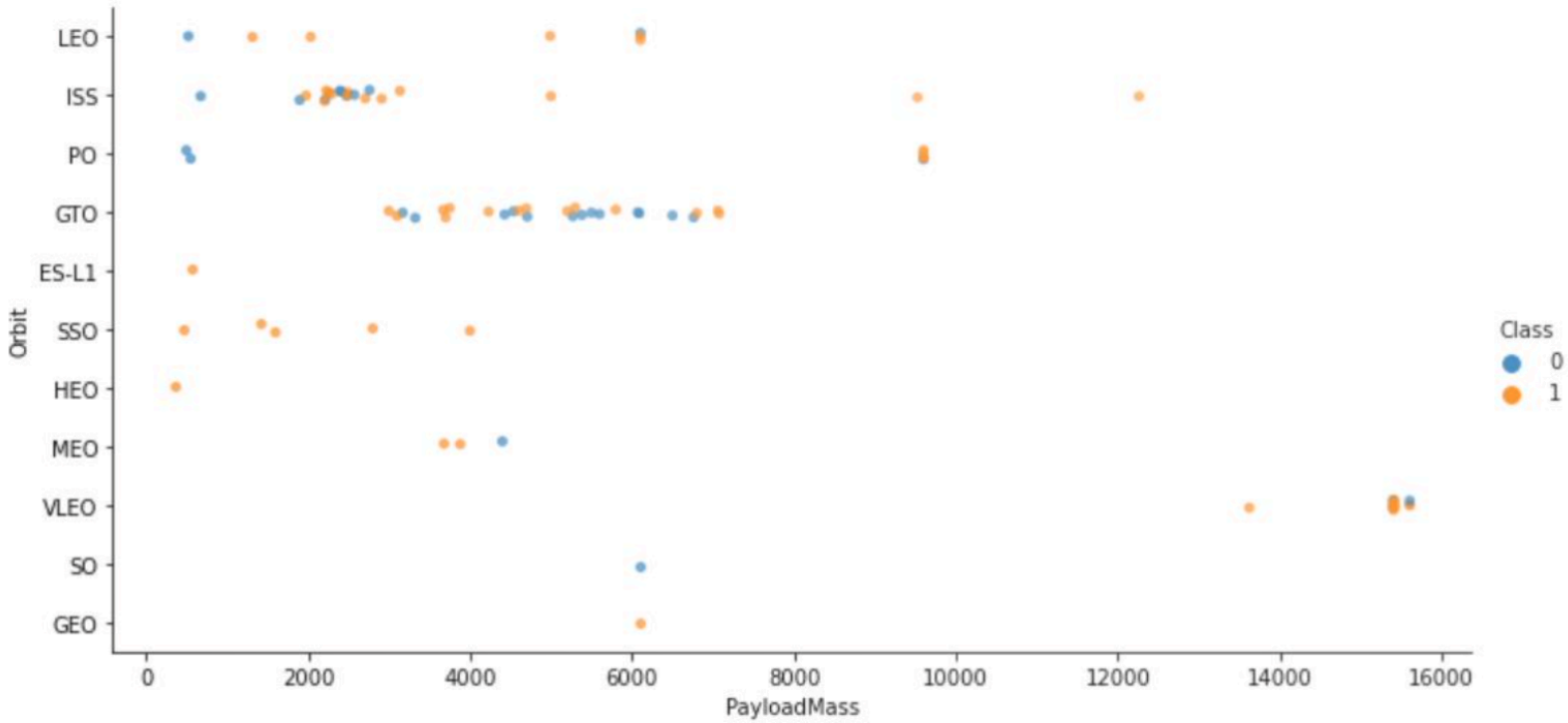
-> ES – L1, GEO, HEO, SSO have the highest success rate

Flight Number vs. Orbit Type



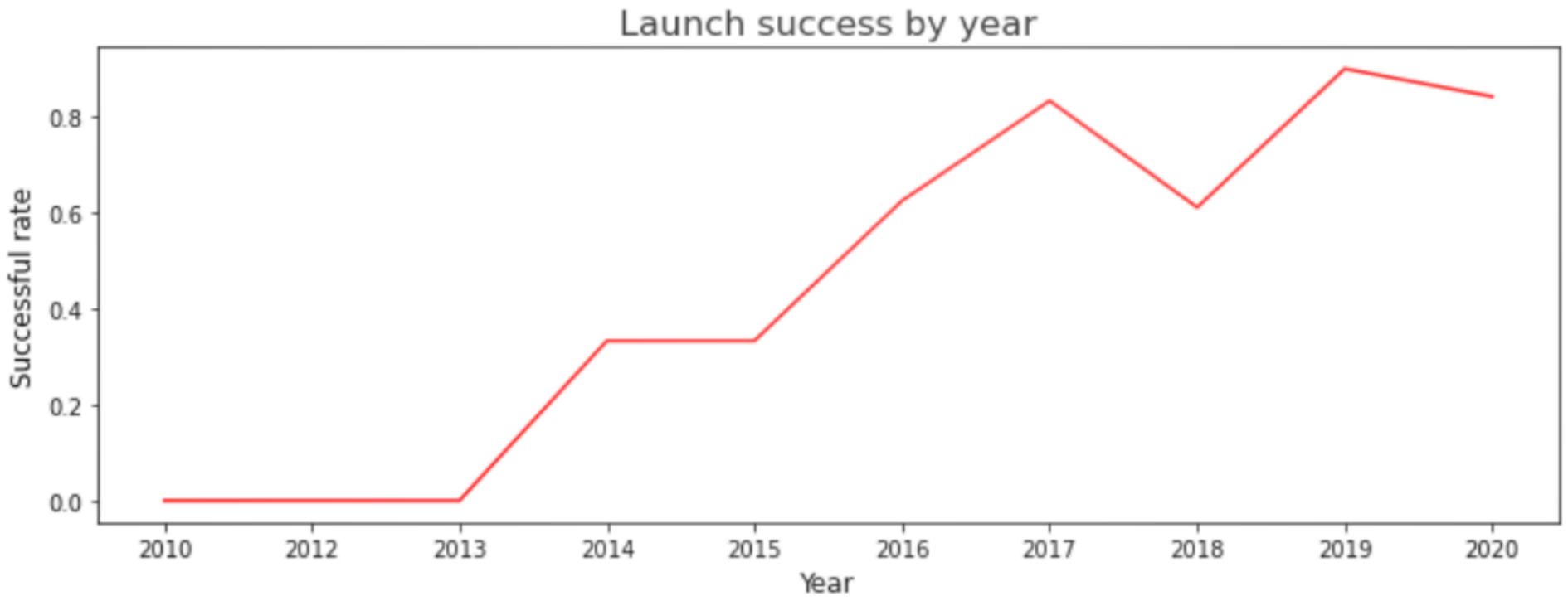
->In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



->With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



->The sucess rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

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

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

-> Total 4 different launch sites

Launch Site Names Begin with 'CCA'

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

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

Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
5:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
3:00	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)
4:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
5:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
0:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

-> The landing outcome are all failure

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload FROM SPACEXTBL WHERE Customer LIKE 'NASA (CRS)';
```

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

total_payload
45596

-> The total payload mass for NASA is 45,596 kg

Average Payload Mass by F9 v1.1

```
%sql SELECT avg(PAYLOAD_MASS__KG_) AS Avg_Payload FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1';
```

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

Avg_Payload
2928.4

-> The average payload mass carried by booster version F9 v1.1 is 2,928.40 kg

First Successful Ground Landing Date

```
%sql SELECT min(date) AS Early_Date from SPACEXTBL where Landing_Outcome LIKE 'Success (ground pad)'
```

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

Early_Date
01-05-2017

-> The first ground landing successful is on 01.05.2017

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT Customer, Landing_Outcome,PAYLOAD_MASS__KG_ FROM SPACEXTBL
WHERE Landing_Outcome ='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

Customer	Landing_Outcome	PAYLOAD_MASS__KG_
SKY Perfect JSAT Group	Success (drone ship)	4696
SKY Perfect JSAT Group	Success (drone ship)	4600
SES	Success (drone ship)	5300
SES EchoStar	Success (drone ship)	5200

-> The most successful landing is by drone ship.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome, Count(*) AS Numbers FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

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

-> There are 1 failure in flight, 99 successes and 1 success wiht unclear payload status.

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version, Max_Payload FROM (SELECT Booster_Version, MAX(PAYLOAD_MASS__KG_)
                                               AS Max_Payload FROM SPACEXTBL GROUP BY Booster_Version)
```

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

Booster_Version	Max_Payload
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092

-> Different booster version has different max payload mass.

2015 Launch Records

```
%sql SELECT SUBSTR(Date,4,2) AS Month, Booster_Version, Launch_site FROM SPACEXTBL
WHERE Landing_Outcome LIKE 'Failure%drone%' AND SUBSTR(Date,7,4) = '2015'
```

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

-> In January and April, 2015 there are launch failure by booster B1012 and B1015.

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

```
%sql SELECT Landing_Outcome, COUNT(*) AS Numbers FROM SPACEXTBL
WHERE Landing_Outcome LIKE 'Success%' AND Date BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY Landing_Outcome ORDER BY Numbers DESC;
```

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

Landing_Outcome	Numbers
Success	20
Success (drone ship)	8
Success (ground pad)	6

-> Between 04-06-2010 and 20-03-2017, there are toally 20 successful landing, 8 successful drone ship landing and 6 successful ground pad landing

LAUNCH SITES PROXIMITIES ANALYSIS

SECTION 3

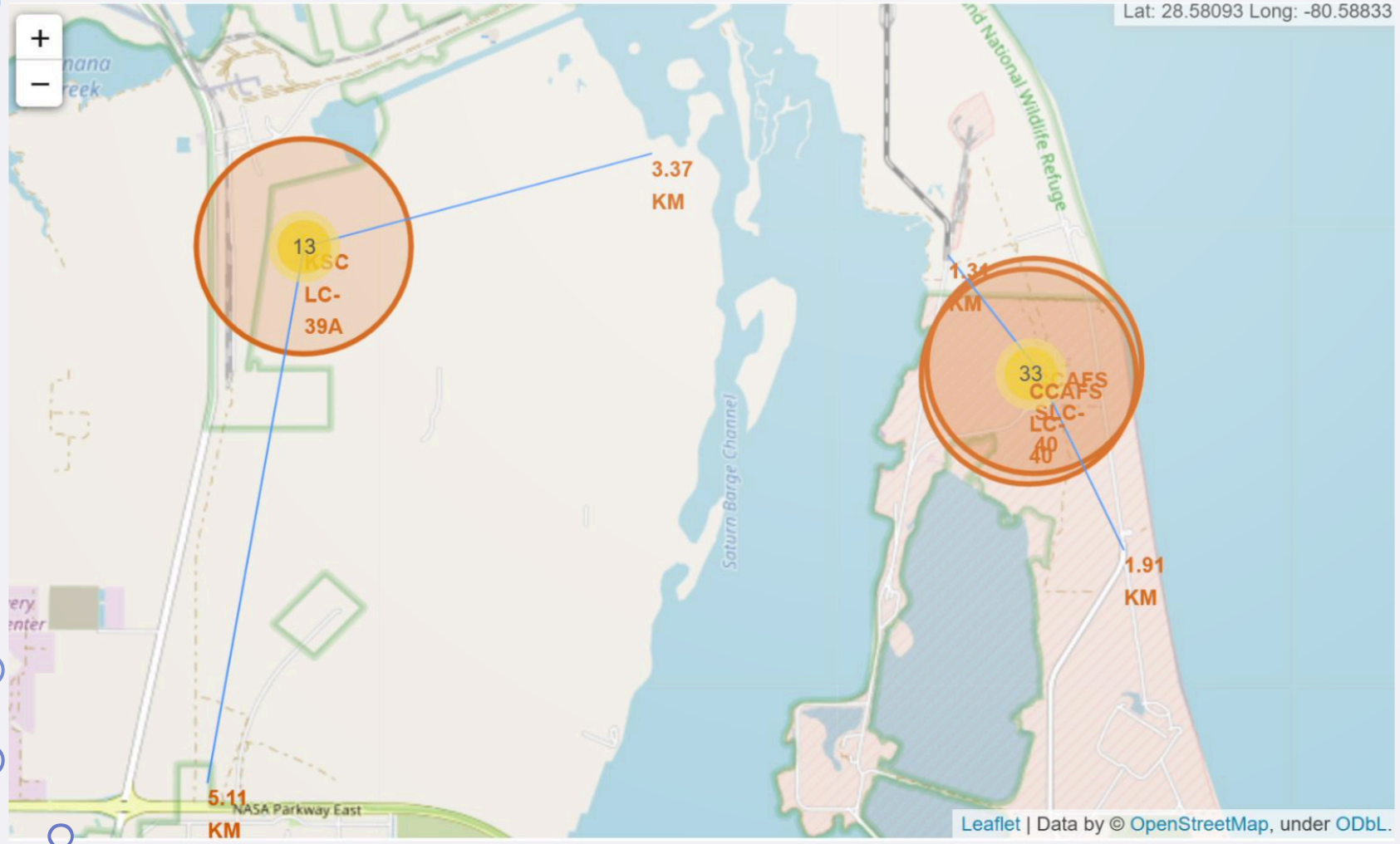
<All launch sites>

- All launch sites in proximity to the Equator line
- All launch sites in very close proximity to the coast

Launch outcome of different site

- Left coast site has 10 trails and right coast site has 46 trails

The proximity of the launch sites

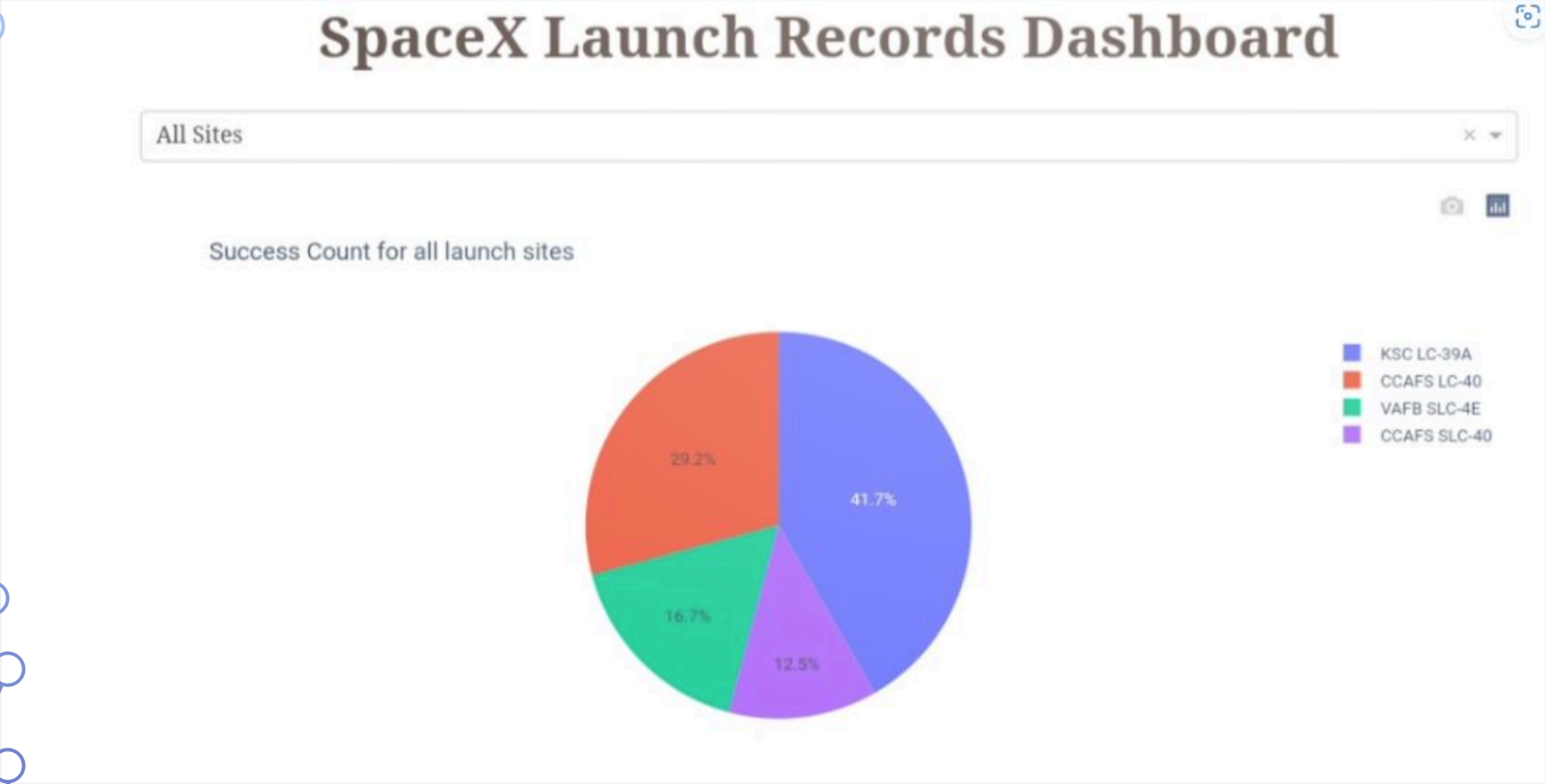


- KSC LC-39 A is 3.37 km far from the coast, and 5.11 km from the city
- CCAFS LC-40 is 1.91 km from the highway and 1.34km from the railway

BUILD A DASHBOARD

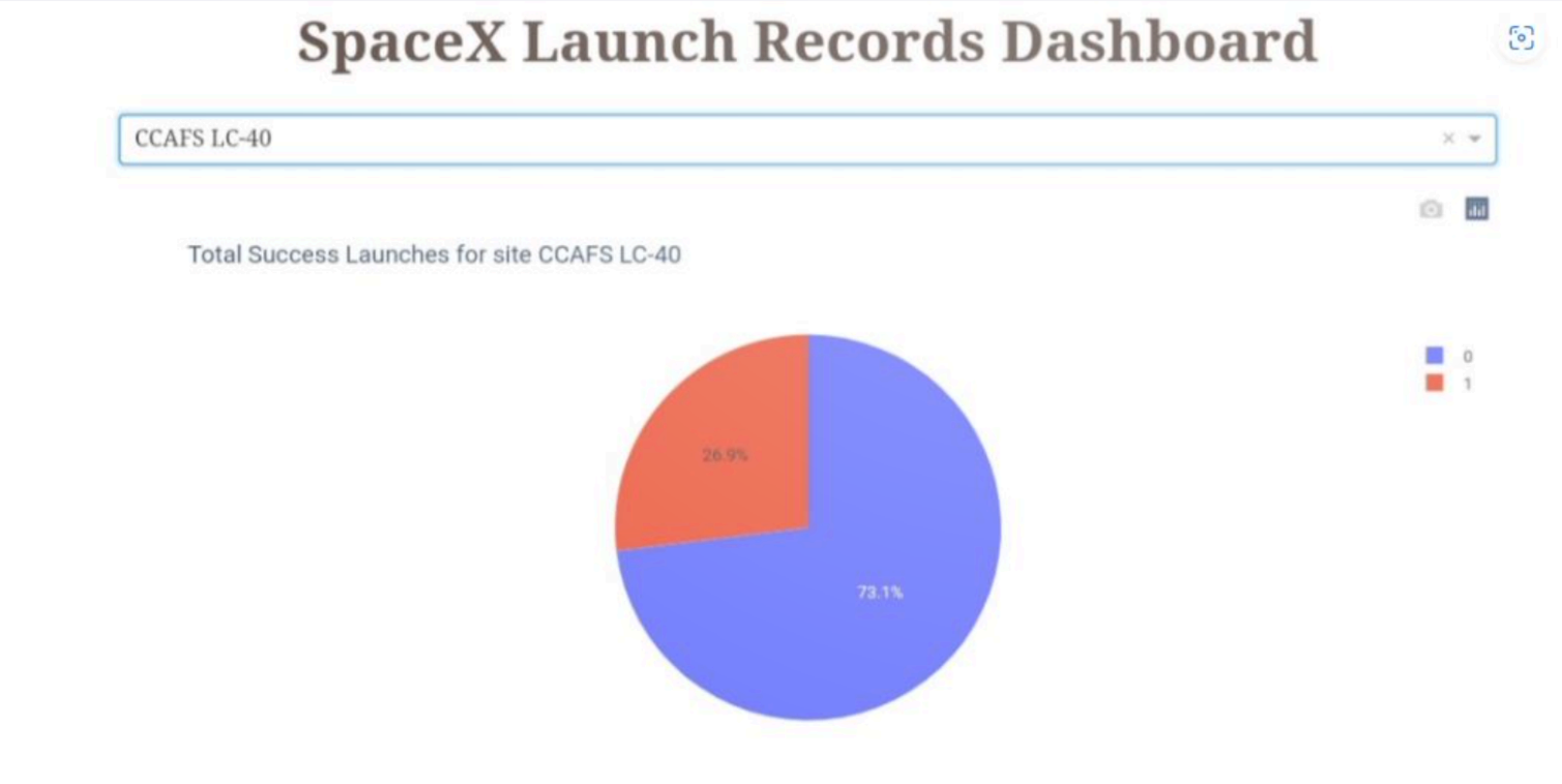
SECTION 4

All site launch



- KSC LC 39A: 41.7%
- CCAFS LC-40: 29.2%
- VAFB SLC -4E: 16.7%
- CCAFS SLC-40: 12.5%

Highest success launch ratio



- Success lauch ratio: 73.1%

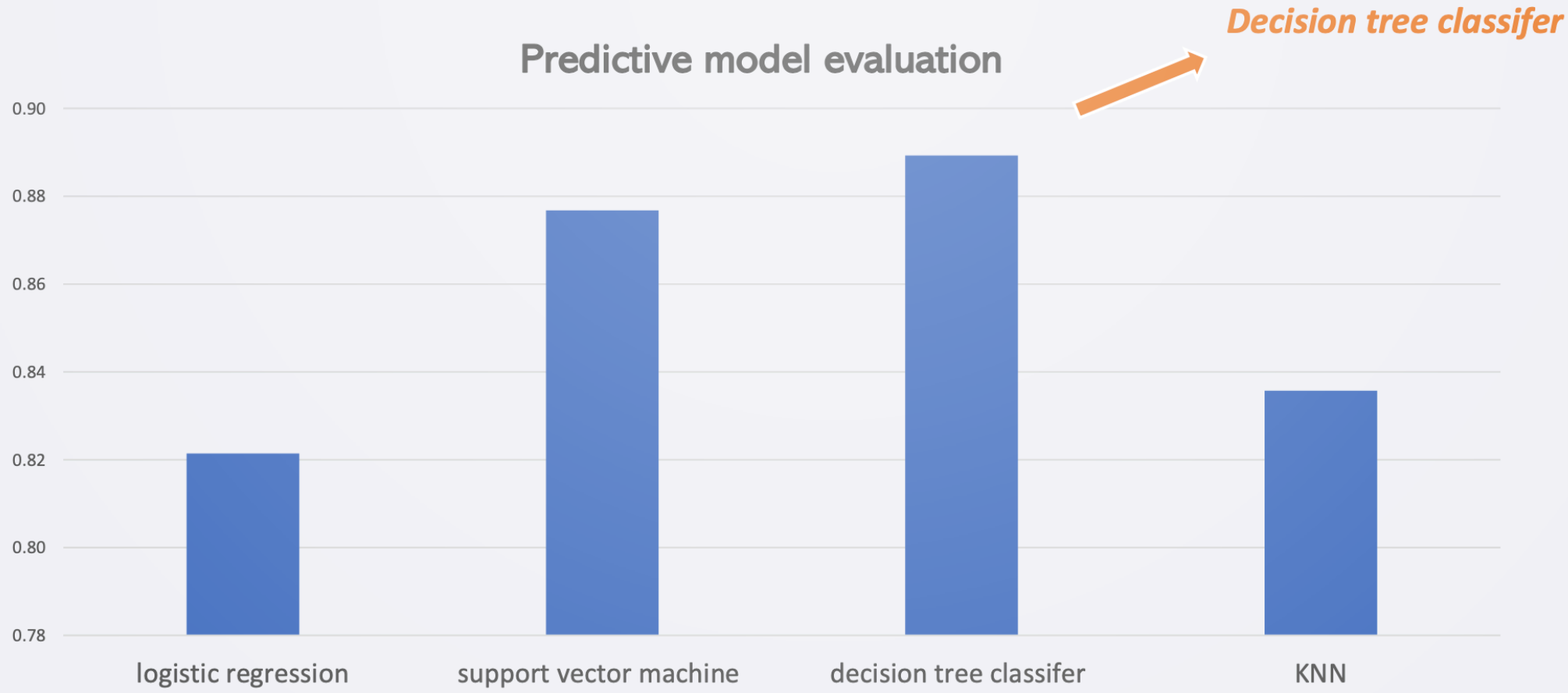
Payload vs. Launch Outcome

- V1.0 can take heaviest payload
- The success land happens between payload from 2k to 5k
- FT has the highest success rate

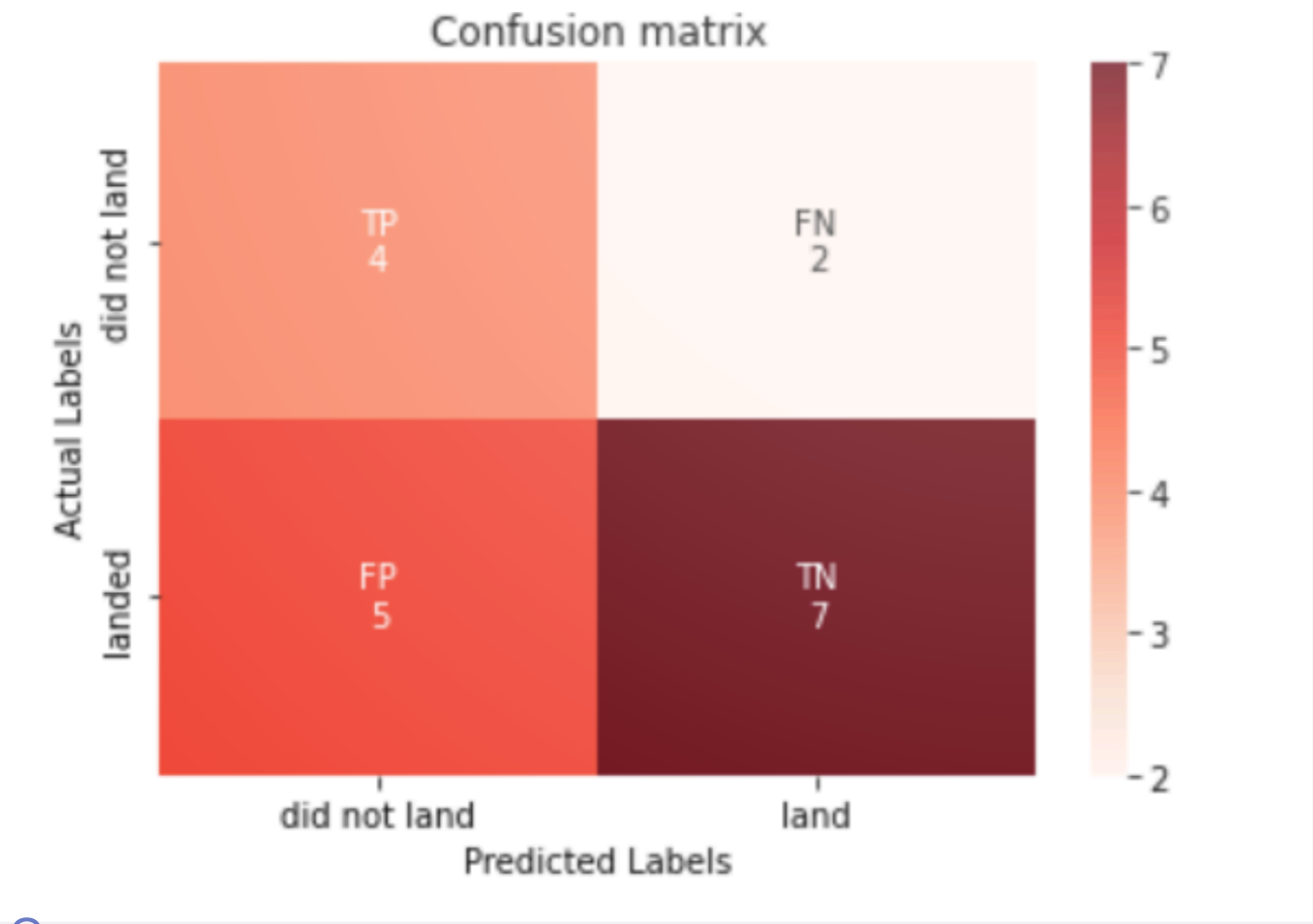
PREDICTIVE ANALYSIS

SECTION 5

Classification Accuracy



Confusion Matrix



- According to the decision tree classifier model, the predictive model tells us that there will be 4 ture postive, 7 true negative, 5 false postive and 2 false negative .

Conclusions

- There is a correlation between launch site and success rate Payload mass is also associated with the success rate.: the more massive the payload, the less likely the first stage will return
- For orbit type, SO has the least success rate while ES-L1, GEO, HEO and SSO have the highest success rate According to the yearly trend
- There has been an increase in the success rate since 2013 kept increasing till 2020
- With best parameter provided, decision tree classifier used in prediction yielded the highest accuracy of 89%. .

Appendix

- https://www.coursera.org/learn/applied-data-science-capstone/home/welcome
- https://github.com/YuliyaShe/project_learning/blob/main/Space_X.ipynb