IBM Developer
SKILLS NETWORK

WINING SPACE RACE WITH DATA SCIENCE

BY TRANG NGUYEN

May 2023

This is the capstone project of the final course in a program called IBM Data Science Professional provided by IBM and Coursera

Outline

Wining Space Race
with Data Science

Executive summary

Summary of all Methodologies and Results

METHODOLOGY

- **Data Collection** using API
- **Data Wrangling** using *pandas and numpy*
- **EDA** and **Feature Engineering** using *SQL*, *pandas,* and *Matplotlib*
- **Interactive Visual Analytics** using map built via *folium* and dashboard built via *plotly dash*
- **Predictive analysis** (*classification*)

RESULTS

- Exploratory Data Analysis result (SQL results, chart and in-array perform result)
- Interactive analytics in screenshots (showing geometrical characteristics and insights via interactive dashboard)
- Predictive Analytics result (the model with best performance and evaluation results)

INTRODUCTION

Project background & Context

Because of SpaceX's first-stage reusability, Falcon 9 rocket launches are advertised on their website for $62 million, which is significantly less than other providers' costs - up to $165 million per launch. By determining whether the first stage successfully landed, launch costs can be determined. This information would be insightful for SpaceX’s competitors who want to compete against SpaceX for bidding.

Questions to be answered

- Which factors have impacts on a successful landing of a rocket? How how they affect the success of the first stage landing?
- What is the best algroithm that can be used to predict succesful landing in this case?

SECTION 1

Methodology

Executive Summary

- **Data collection methodology:**
  - Collected using SpaceX API and web scraping from Wikipedia.
- **Data wrangling**
  - Dealing with missing data (imputation)
  - Encoding categorical features(One Hot Encoding)
- **Exploratory data analysis (EDA)**
  - using visualization
  - using SQL
- **Interactive visual analytics**
  - using maps built by Folium
  - using dashboards built by Plotly Dash
- **Predictive analysis** using *classification models*
  - Building, tuning, and evaluating classification models to achieve optimal outcomes

METHODOLOGY

DATA COLLECTION

Data is collected from 2 sources which are:

- API requests from SpaceX's API
- Web scrapping data from a table in SpaceX's Wikipedia site

Columns obtained from API

'FlightNumber', 'Date', 'BoosterVersion', 'PayloadMass', 'Orbit', 'LaunchSite', 'Outcome', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial', 'Longitude', 'Latitude'

Columns obtained from Wikipedia

'Flight No.', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome', 'Version Booster', 'Booster landing', 'Date', 'Time'

DATA COLLECTION FLOW

API (GitHub link)

Request rocket launch data from SpaceX API

Decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()

Request needed data via API and apply predefined functions to get comprehensible information

Construct obtained data into a pandas dataframe

Collect all needed information via API

DATA COLLECTION FLOW

Web Scraping (GitHub link)

Request the Falcon9 Launch Wiki page from its URL

Extract all column/variable names from the HTML table header with BeautifulSoup

Create an empty dictionary called launch_dict with keys from the extracted column names

Create parsing processes looping through all table ('tr') and table cell ('td') from HTML tables to fill launch_dict

Convert the launch_dict into a Pandas dataframe

Collect all needed information from Wikipedia

DATA WRANGLING

Github

NULL DECTECTION

Identify and calculate the percentage of the missing values in each attribute using isnull()

UNIQUE VALUES AND FREQUENCIES

Calculate unique values and its frequency using value_counts()

TRAINING LABEL

Convert Outcome column into Training Labels with "1" mean the booster successfully landed and "0" mean unsuccessful.

EDA

EDA WITH VISUALIZATION

Multiple charts of different kinds (i.e: Bar chart, Scatter plot, Line chart) were generated to explore the relationship between features, show trends in data overtime for time series data, and compare some specific categories by measured values. Plotted charts show the relationships of: 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, Yearly Trend, etc.

GitHub link

EDA WITH SQL

Performed SQL queries:

- Display the names of the unique launch site 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 boosters which have success in drone ship and hay 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 ailed landing outcomes in drone ship, their booster versions and launch site names 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 017-03-20

GitHub link

INTERACTIVE VISUAL ANALYTICS

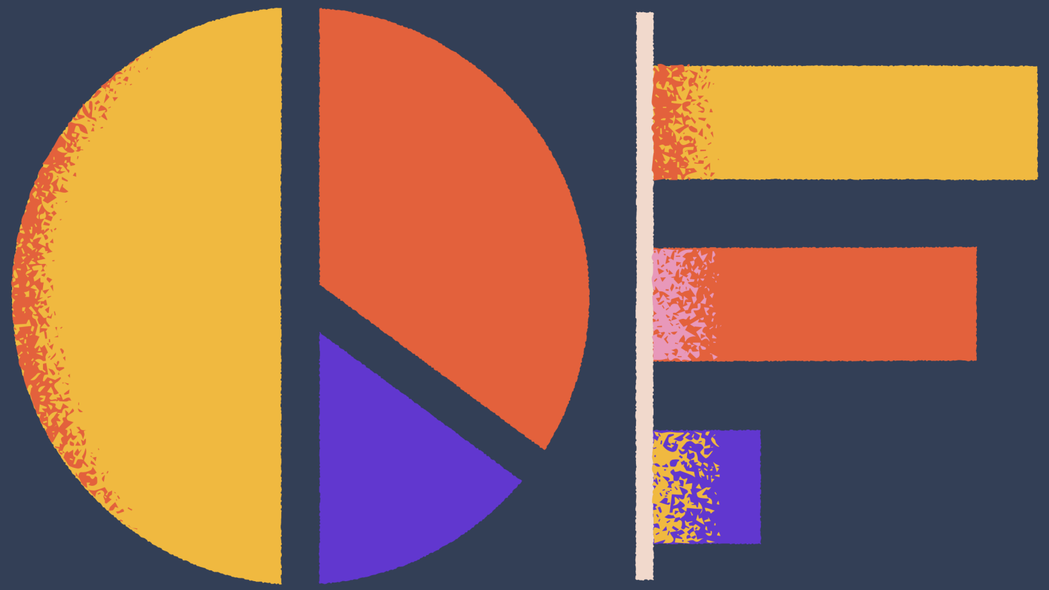
INTERACTIVE FOLIUM

Mark all launch sites, and added map objects such as markers, circles, pop up label to mark the success or failure of launches for each site on the folium map using their latitude and longtitude coordinates => show their geographical locations.

Use the color-labeled marker clusters to identify which launch sites have relatively high success rate.

Identify distances between a sample launch site and its proximities like railway, coastline, highway, and city.

GitHub link



INTERACTIVE VISUAL ANALYTICS

INTERACTIVE DASHBOARD

Pie charts showing the total successful launches for all sites, or the success vs failed counts for certain site selected from dropdown list.

Scatter chart show he correlation between payload mass and launch success, the the range of payload mass selected by payload mass range slider

Launch site selection button and range slider of payload mass range act as data filters to filtered data under certain conditions and plot it via charts.

GitHub link

PREDEICTIVE ANALYSIS FLOW

(GitHub link)

Define Label (Class) and Features (feature set after applied One Hot Encoder)

Starndardize Features usering StandardScaler

Split features and label into train set and test set using train_test_split

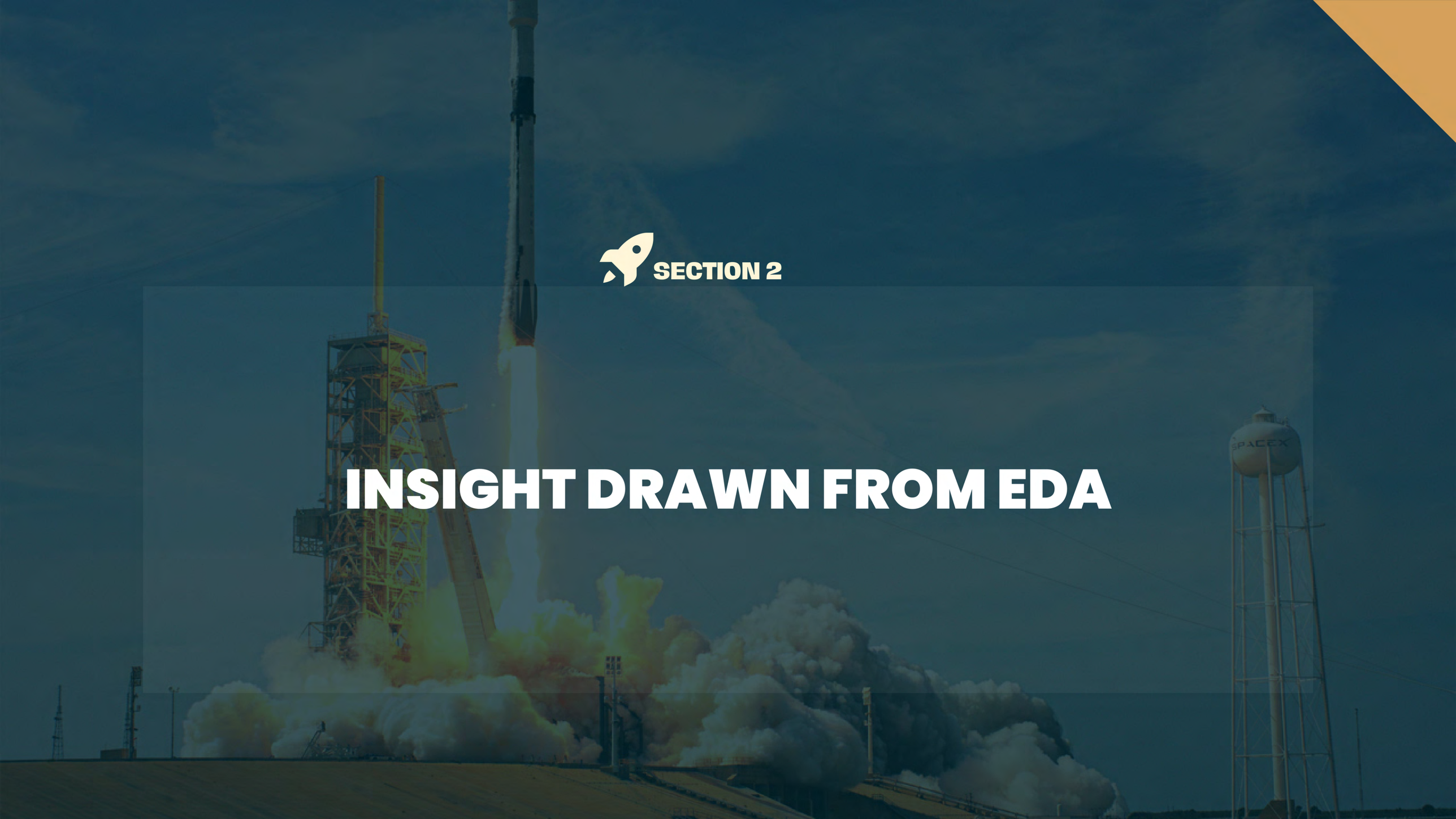
Apply GridSearchCV on Logistic Regression, SVM, Decision Tree and KNN models to find the best parameter set

Evaluate the performace of each model using method .score() and confusion matrix

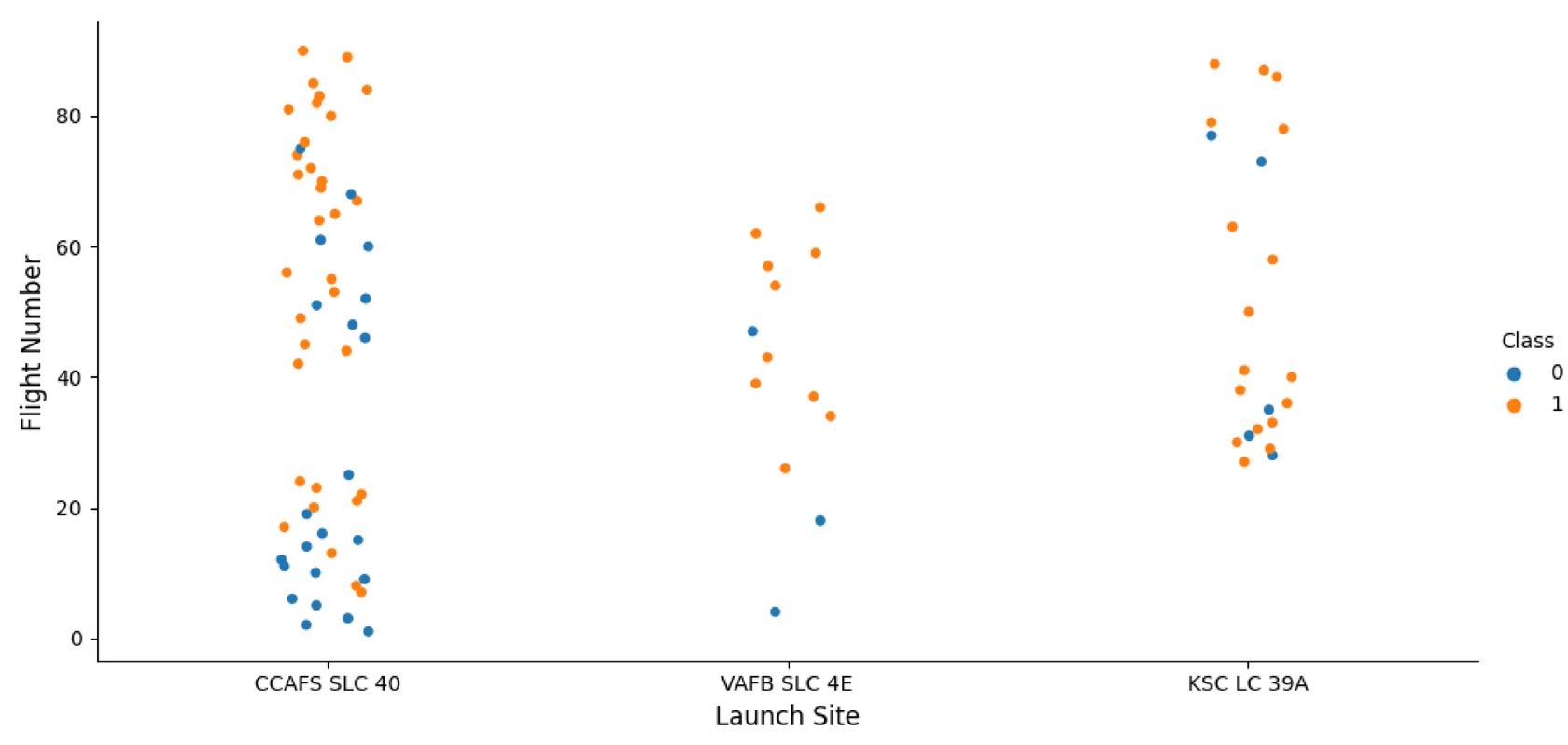
Indicate the model with the best outcome

RESULTS

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


SECTION 2
INSIGHT DRAWN FROM EDA

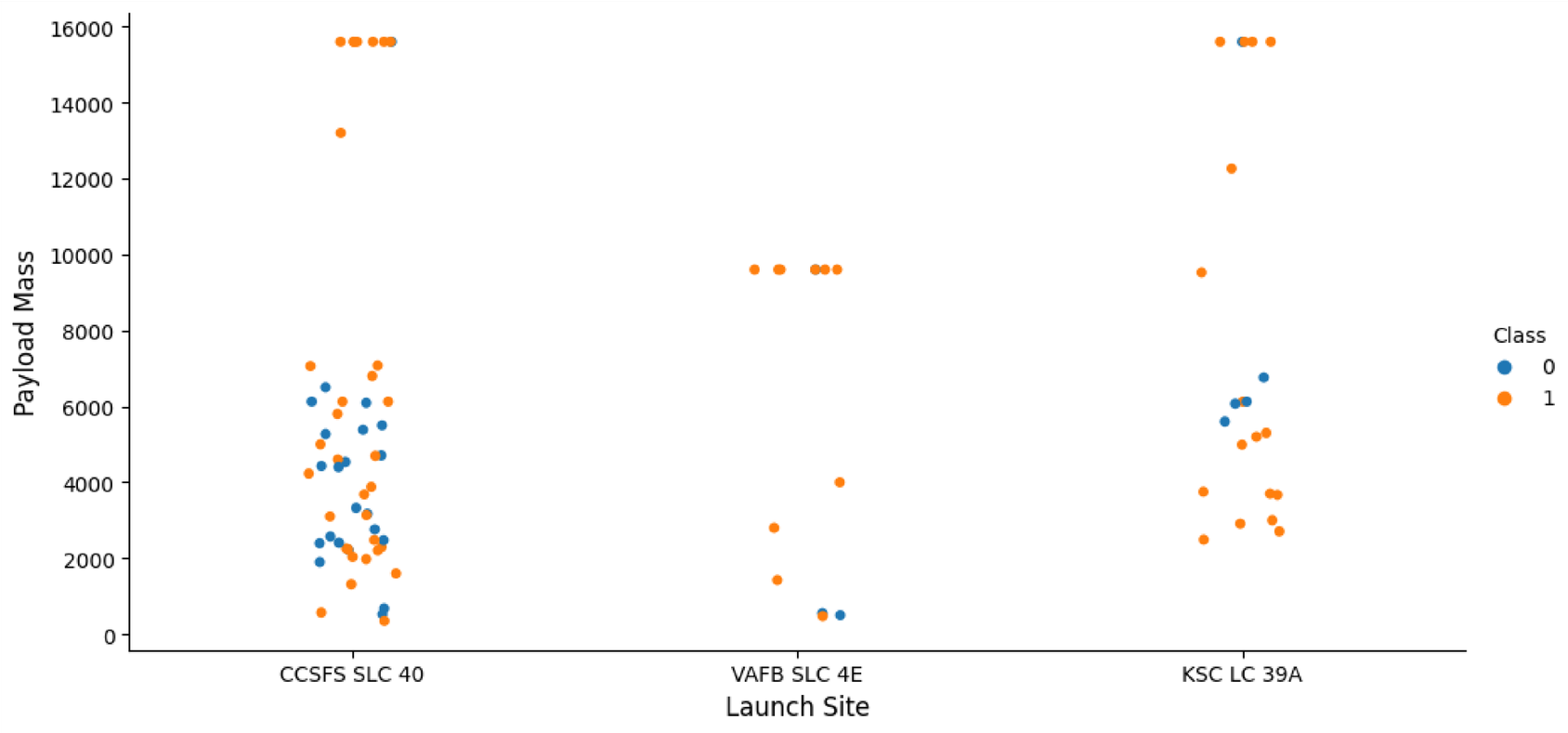
FLIGHT NUMBER VS. LAUNCH SITE



It could be seen from the chart that:

- About half of the launches were implemented at **CCAFS SLC 40** launch site.
- Earlier flights tend to fail at landing while later flights were more likely to succeed.
- The success rate of each landing seems raised with new launches.

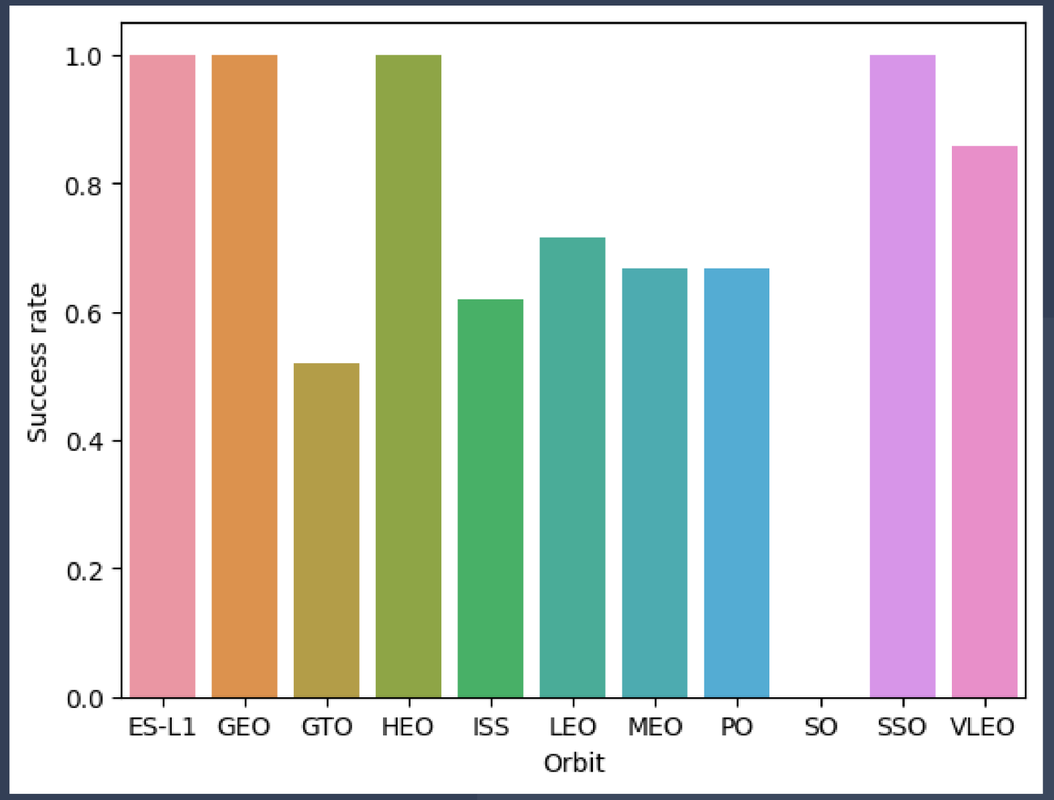
PAYLOAD VS. LAUNCH SITE



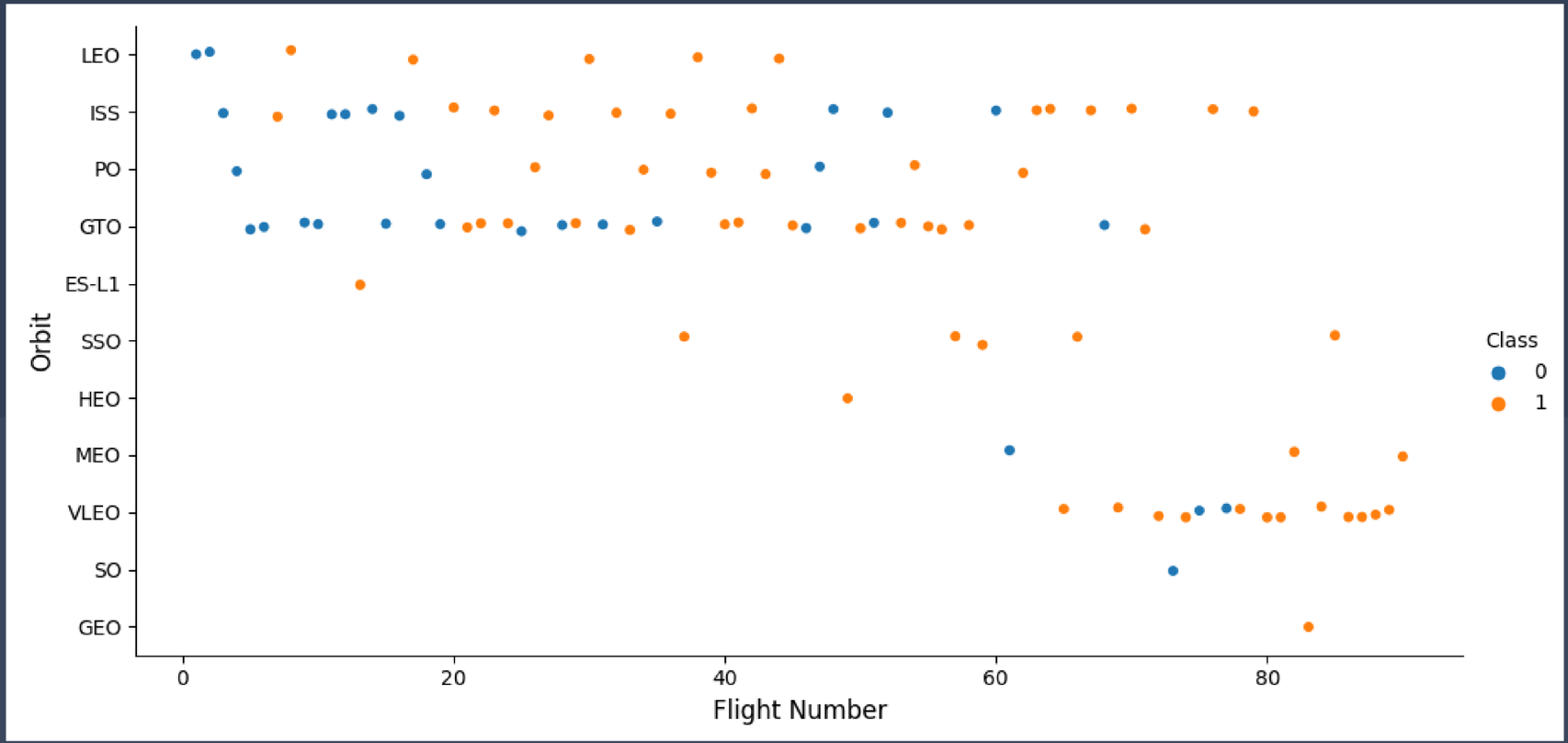
- Payload Mass of launches implemented at **VAFB SLC 4E** is relatively lower then other 2 sites.
- Flights with maximum payload mass are more likely to succeed at landing.
- Most of flights with payload mass over 7000kg were succeed at landing.
- All flights at KSC LC 39A with payload mass under 5500kg were succeed at landing.

SUCCESS RATE VS. ORBIT

- ES-L1, GEO, HEO, and SSO orbit types have the landing success rate of 100%
- In contrast, SO orbit types have the success rate of 0%
- All the remaining orbit types have success rate between 50% and 85%

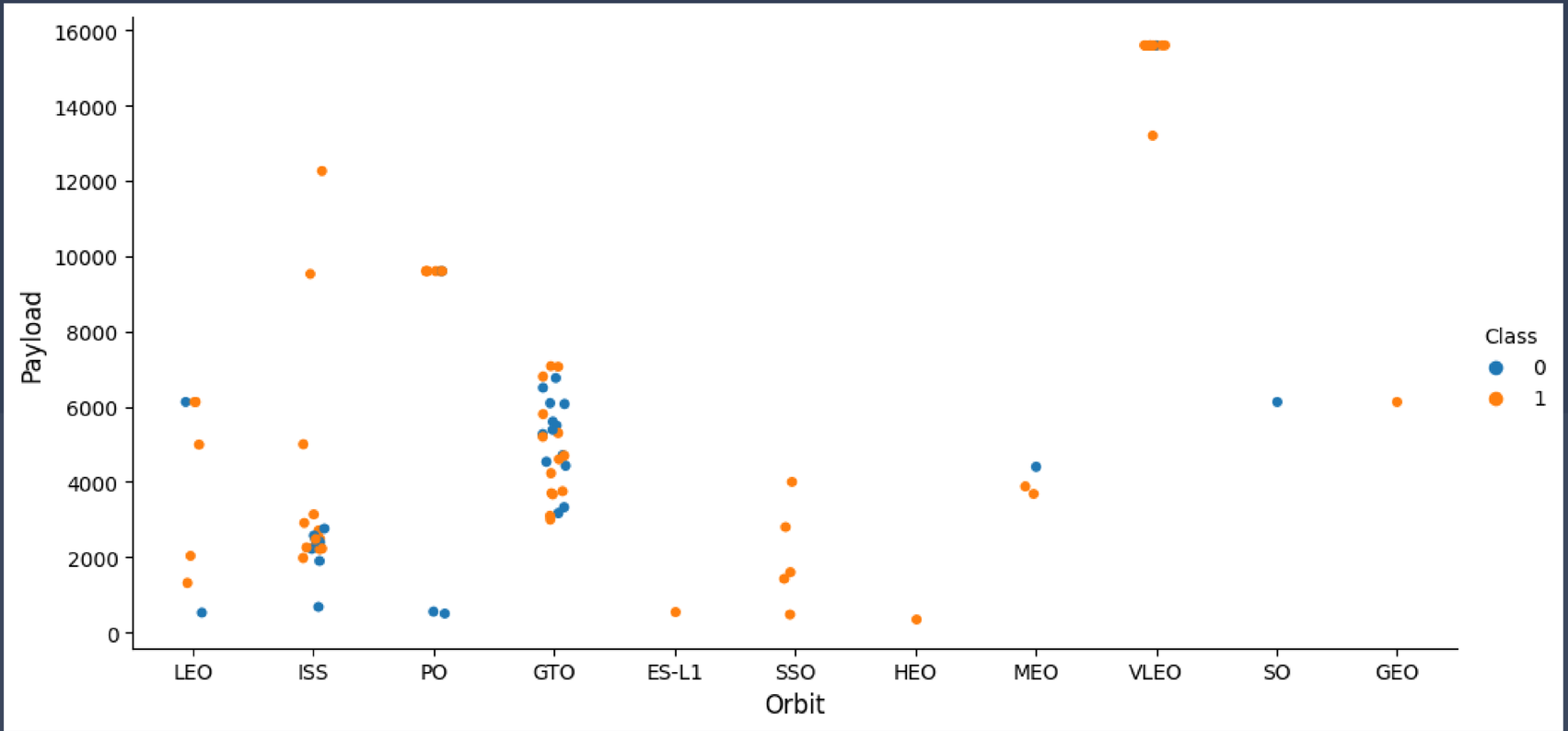


FLIGHT NUMBER VS. ORBIT TYPE



- Earlier launches normally apply LEO, ISS, PO and GTO orbit type while only later launches apply SSO, HEO, MEO, VLEO, SO, and GEO type.
- After first 2 failed LEO landing outcome appears correlated to the flight number.
- Meanwhile no correlation between landing outcome and flight number could be seen at GTO type.

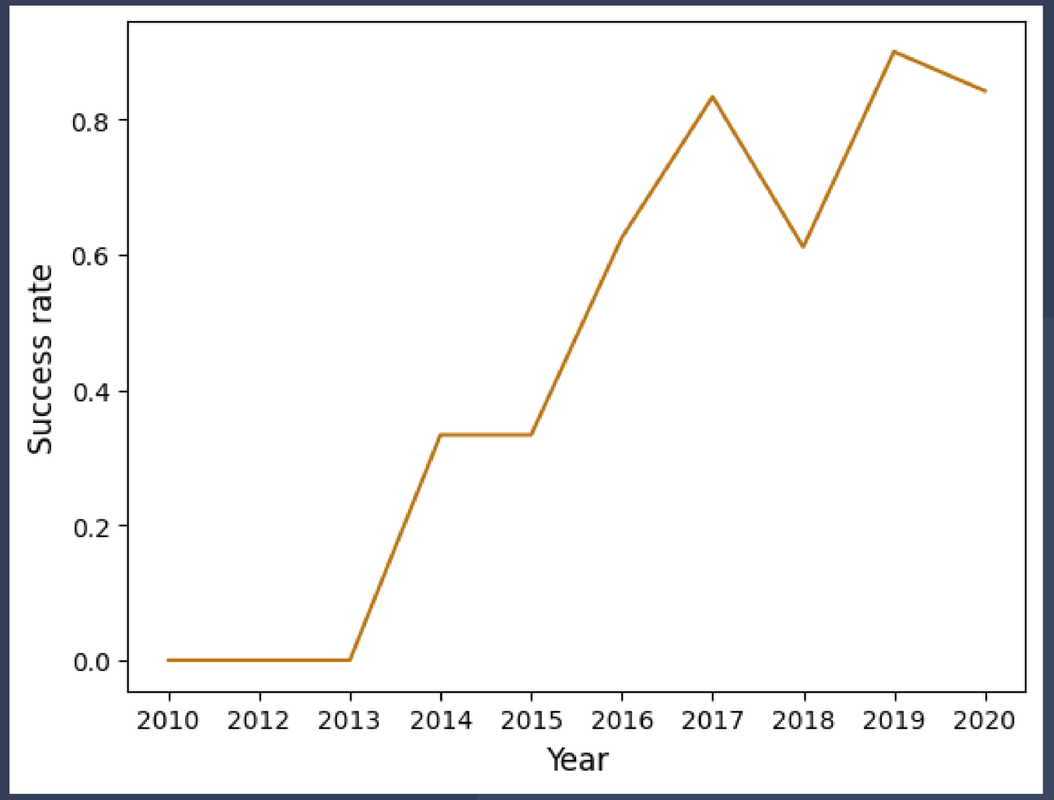
PAYLOAD VS. ORBIT TYPE



- VLEO appears to has the largest payload mass among all orbit types while HEO has the smallest.
- GTO shows no correlation between payload mass and orbit type
- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.

LAUNCH SUCCESS YEARLY TREND

- The sucess rate since 2013 kept increasing till 2020 except a slight decrease in 2018.
- The highest success rate observed in 2019 with more than 90%



```
[13]
1  %%sql
2  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

All launch site names

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

LAUNCH SITE NAMES BEGIN WITH CCA

```
%%sql
select *
from SPACEXTBL
where "Launch_Site" like "CCA%"
limit 5
```

Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Display 5 record where launch sites begin with the string 'CCA' using syntax with "%" in WHERE condition

```
1  %%sql
2  select sum(PAYLOAD_MASS__KG_)
3  from SPACEXTBL
4  where Customer='NASA (CRS)'
5
```

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

sum(PAYLOAD_MASS__KG_)
45596

Total payload mass

Display the total payload mass carried by booosters launched by NASA (CRS) using sum()

```
1  %%sql
2  select avg(PAYLOAD_MASS__KG_)
3  from SPACEXTBL
4  where Booster_Version like "F9 v1.1%"
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

Average Payload Mass by F9 v1.1

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

```
1  %%sql
2  select min("Date")
3  from SPACEXTBL
4  where "Landing _Outcome"="Success (ground pad)"
```

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

min("Date")
01-05-2017

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was achieved

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

```
1  %%sql
2  select "Booster_Version"
3  from SPACEXTBL
4  where "Landing _Outcome"="Success (drone ship)" and "PAYLOAD_MASS__KG_"  between 4000 and 6000
```

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

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

```
%%sql
select "Success" as state, count(*)
from SPACEXTBL
where "Mission_Outcome" like "%Success%"
union

select "Fail" as state, count(*)
from SPACEXTBL
where "Mission_Outcome" like "%Fail%"
```

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

state	count(*)
Fail	1
Success	100

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%%sql
select distinct Booster_Version
from SPACEXTBL
where  PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

```

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

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

Boosters Carried Maximum Payload

List the name of the booster versions which have carried the maximum payload mass

2015 LAUNCH RECORDS

```
%%sql
select substr(Date,4,2), Booster_Version, launch_site
from SPACEXTBL
where "Landing _Outcome"="Failure (drone ship)" and substr(Date,7,4)='2015'
```

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

substr(Date,4,2)	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

```
%%sql
select "Landing _Outcome", count("Landing _Outcome") as count
from SPACEXTBL
where Date between "04-06-2010" and "20-03-2017"
group by "Landing _Outcome"
order by count desc
```

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

Landing _Outcome	count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

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

List

RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

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

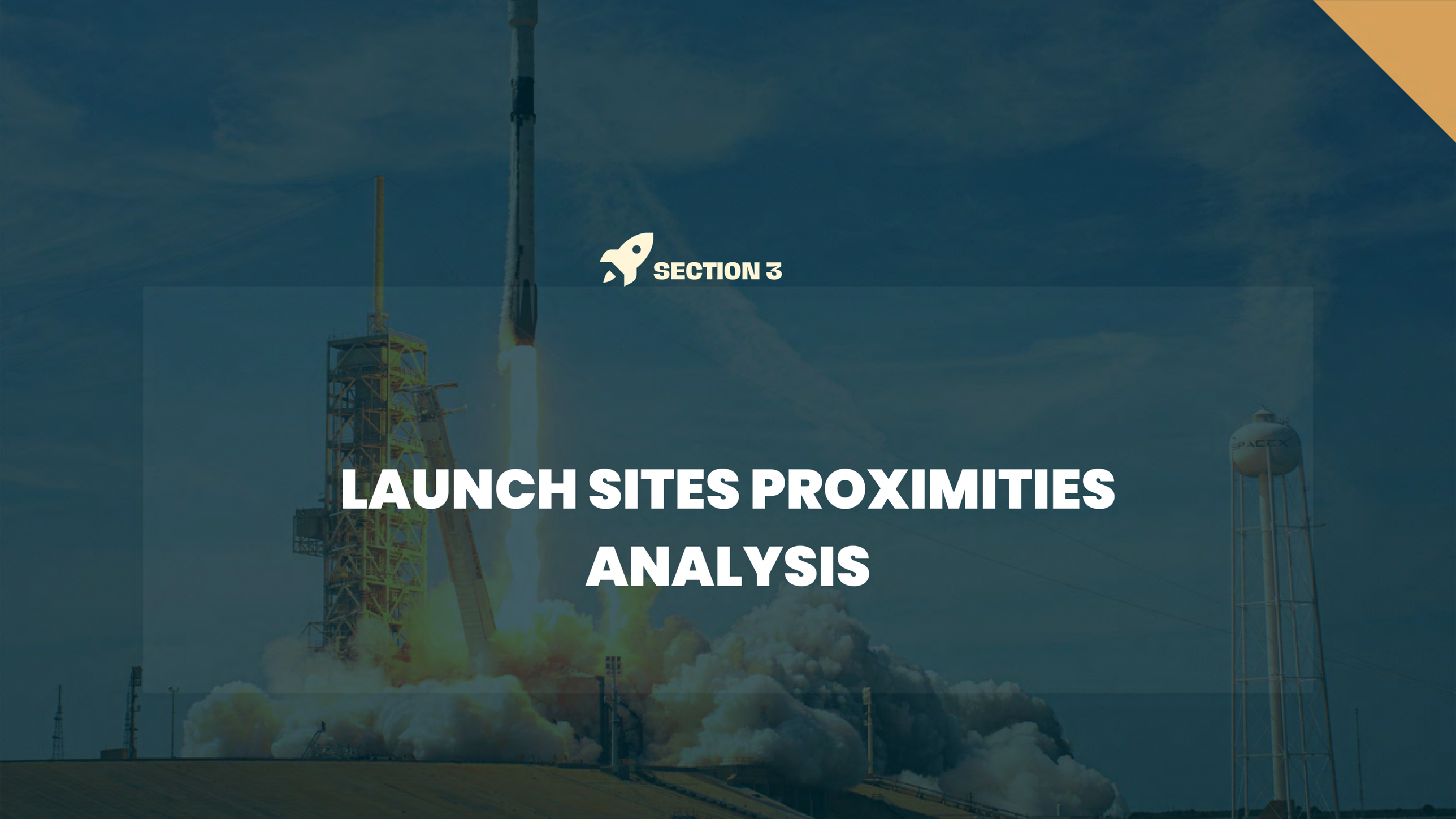
```
%%sql
select "Landing _Outcome", count("Landing _Outcome") as count
from SPACEXTBL
where Date between "04-06-2010" and "20-03-2017" and "Landing _Outcome" like "%Success%"
group by "Landing _Outcome"
order by count desc
```

Python

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

Landing _Outcome	count
Success	20
Success (drone ship)	8
Success (ground pad)	6

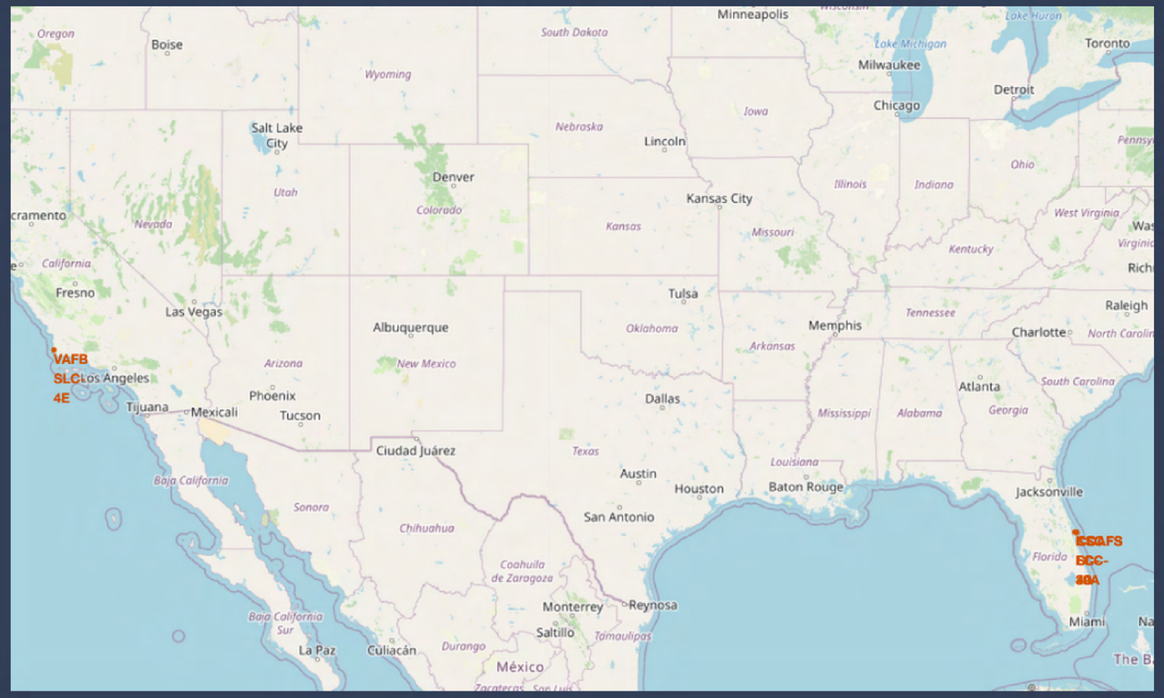
Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.



LAUNCH SITES PROXIMITIES ANALYSIS

MAP

ALL LAUNCH SITES ON GLOBAL MAP



- Most of launch sites are in proximity to the Equator line.
- All launch sites are in very close proximity to the coast

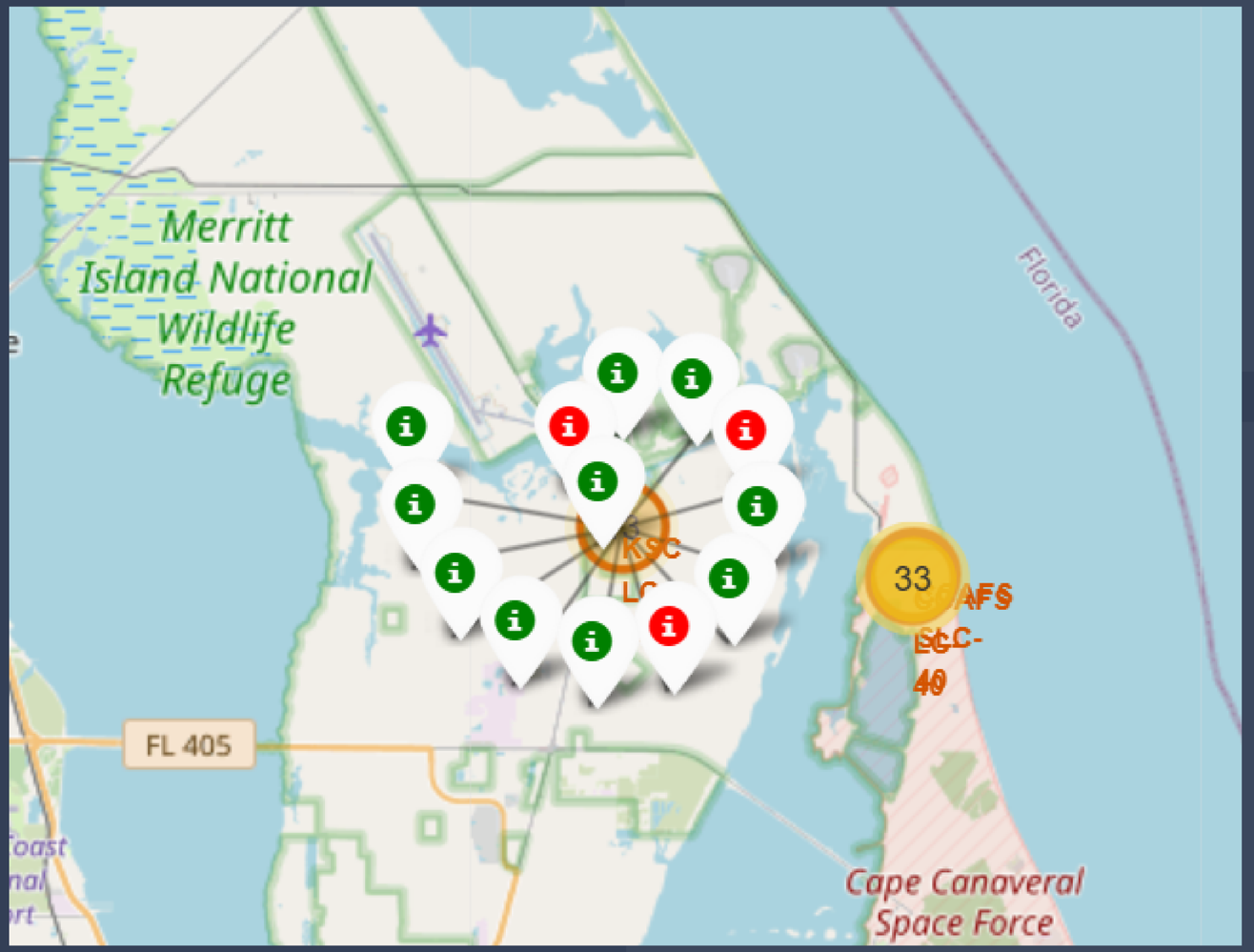
Explanation:

- Launching rockets towards the ocean will minimize the risk of having any debris dropping or exploding near residential areas.
- The speed around the equator will reinforce the rocket's flight speed

MAP

COLOR-LABELED LAUNCH RECORDS

- Green: Successful outcome
- Red: Unccessful outcome
- It could be seen from the map that Launch site KSC LC-39A has a high succress rate

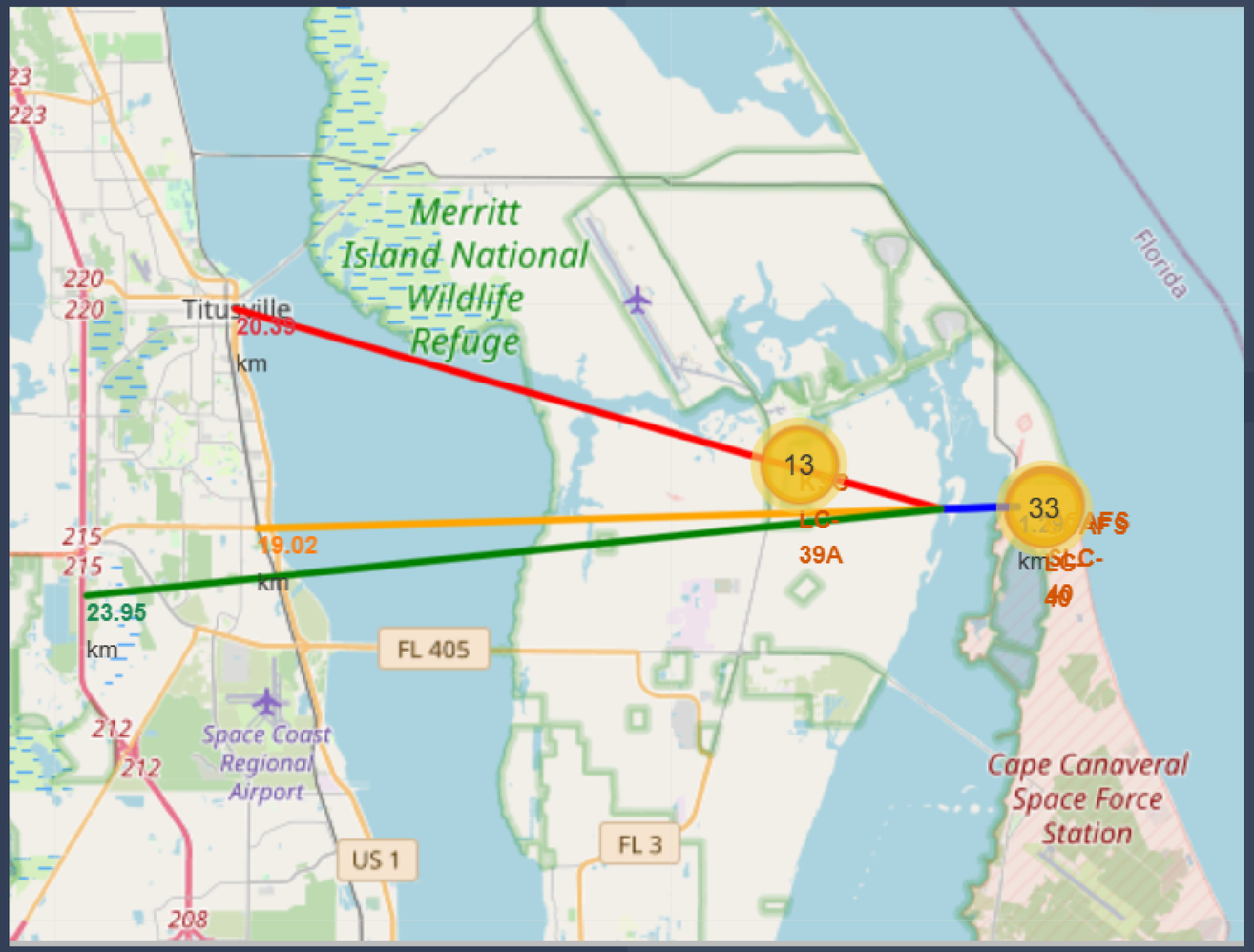


MAP

DISTANCE BETWEEN LAUNCH SITE AND ITS PROXIMITIES

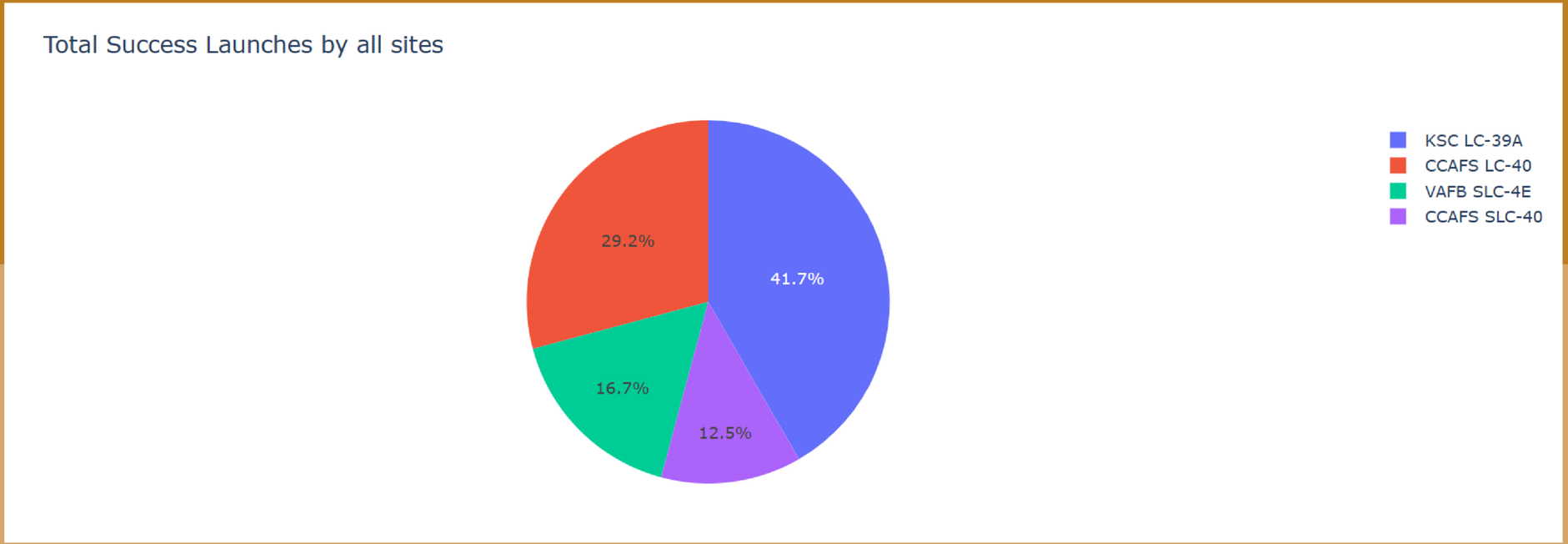
Take launch site CCAFS SLC-40 as an examplt, tt could be seen from the map that:

- Launch site is close to coastline (~1.29km)
- Launch site is relativvely far from railway (19.02km), highway (23.95km) or its closest city Titusville (20.35km)



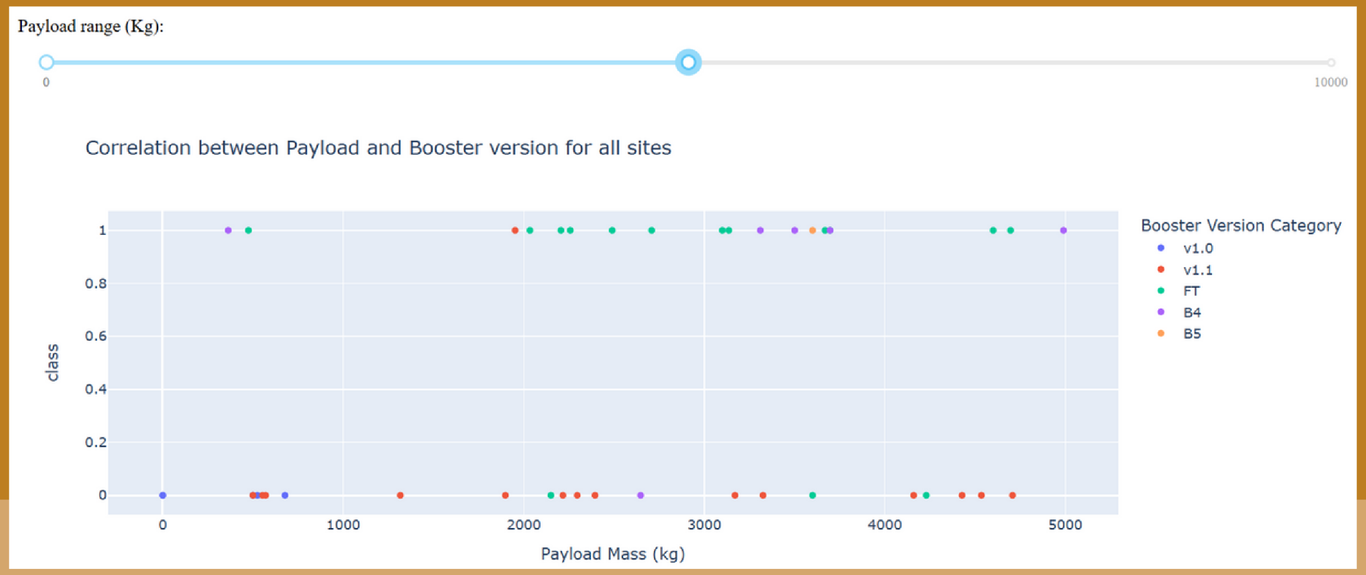

SECTION 4
BUILD A DASHBOARD WITH PLOTLY

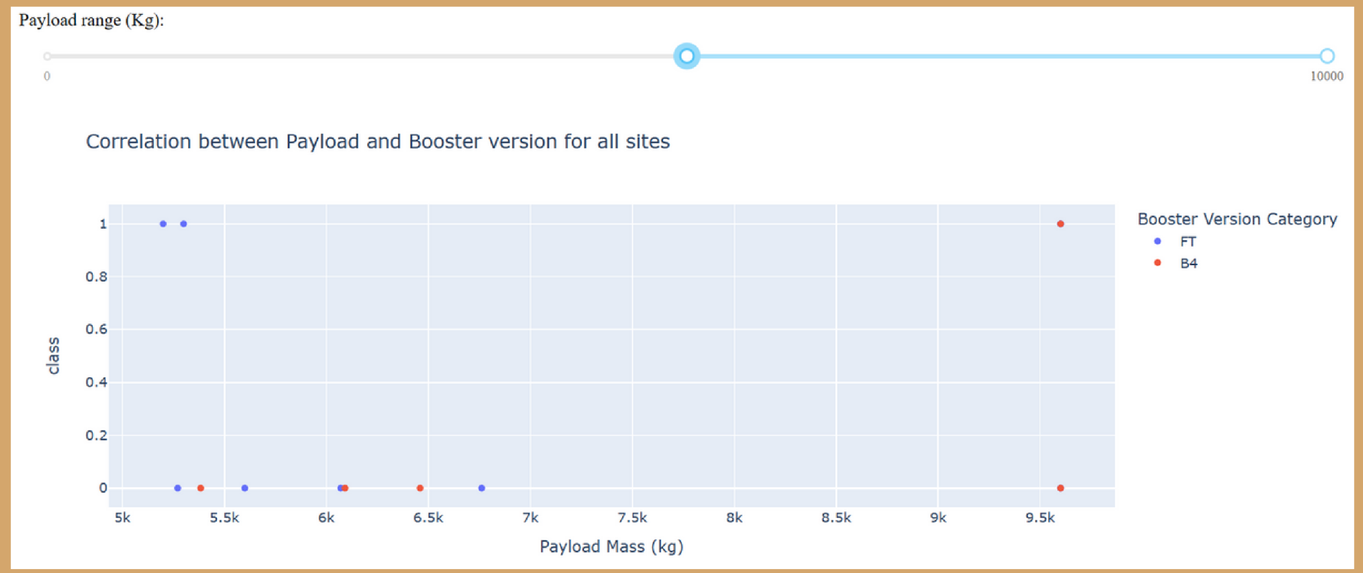
SUCCESS LAUNCHES BY ALL SITES



KSC LC -39A has the most successful launches among all sites while CCAFS LC-40 has the last position

SUCCESS LAUNCHES BY ALL SITES

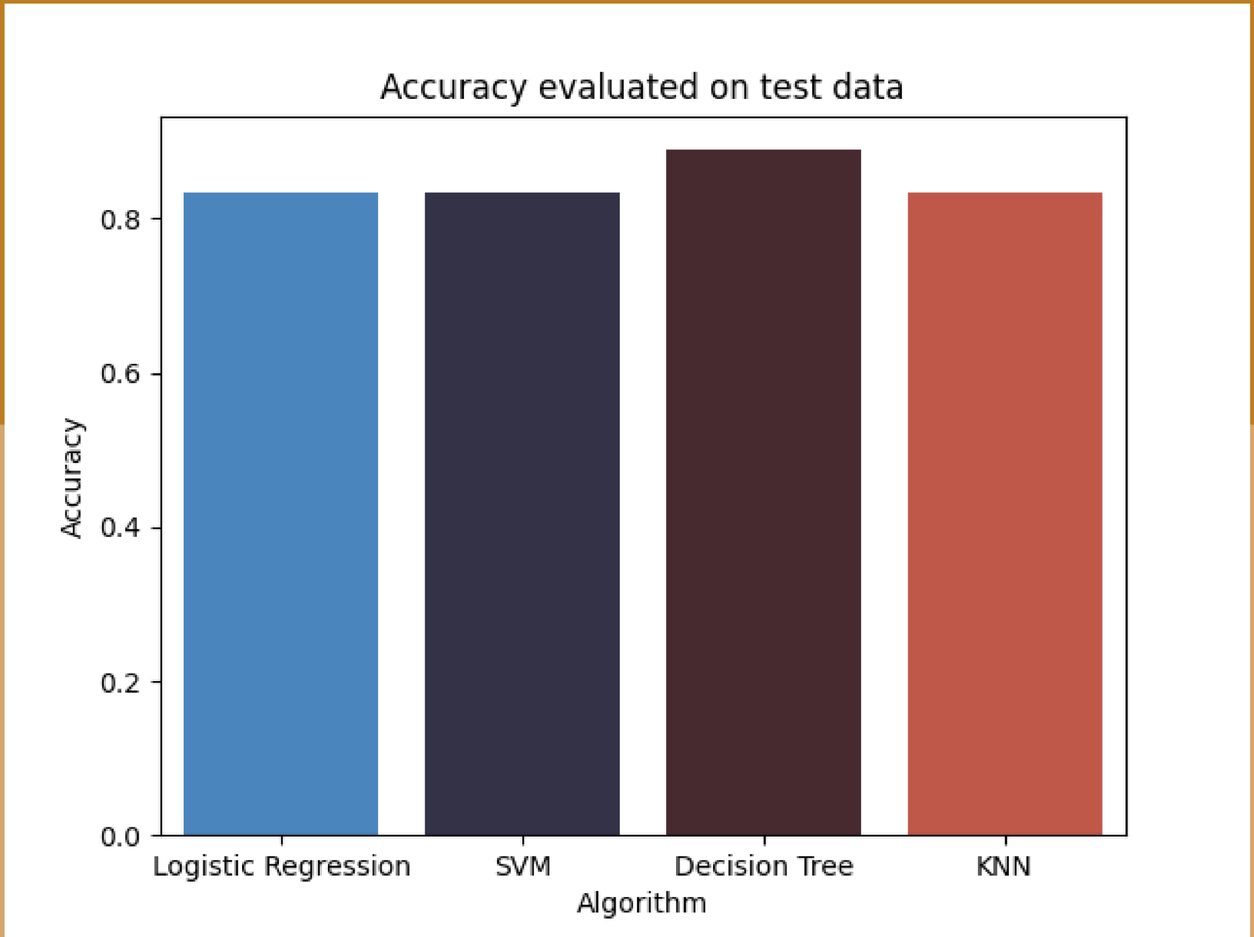




- For relatively heavy payload (>5000kg), only FT and B4 booster version are used while there are many choice for lighter payload (<5000kg)
- Light payload appears to have a higher success rate


SECTION 5
PREDICTIVE ANALYSIS
CLASSIFICATION

SUCCESS LAUNCHES BY ALL SITES



	Acc Train	Acc Test
Logistic Regression	0.846429	0.833333
SVM	0.848214	0.833333
Decision Tree	0.900000	0.888889
KNN	0.848214	0.833333

Based on the bar plot of accuracy evaluated on test set and the dataframe below. It could be easily seen that model built with Decision Tree has the best perform among all used classfication algorithms.

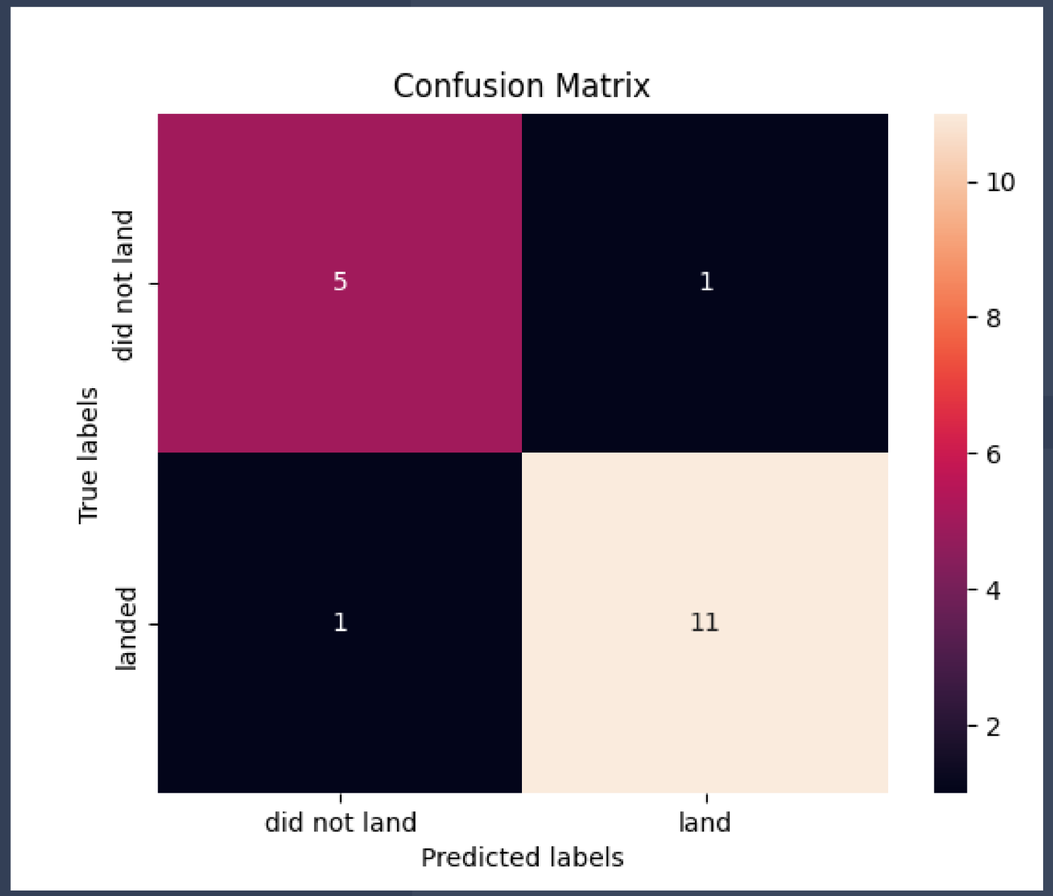
CONFUSION MATRIX

Confusion matrix of Decision Tree classfication model shows that:

- True Negative = 5
- True Positive = 11

While False Negative and False Positive is relatively low as 1.

=> Decision Tree performs well in solving this project problem.



CONCLUSION

01

In general, launches are getting more and more stable (higher success rate over time and flight number)
CCAFS SLC 40

02

Light payload launches appears to have better success probability

03

Launch sites near the Equator (i.e KSC LC 39A) are more frequently used and seems to achived higher success rate in comparison to further site (i.e VAFB SLC 4E)

04

All the launch sites are close to the coastline and far from residental areas

05

ES-L1, GEO, HEO, and SSO orbit types showing the success rate of 100%

06

Decision Tree should be chosen to solve classification state in this dataset as both the accuracy score and confusion matrix showing good evaluation

Appendix

For any questions or concerns

PROJECT GITHUB LINK

https://github.com/IrisGun/IBM-Applied-Data-Science-Capstone

Trang Nguyen

THE END

THANK YOU

Trang Nguyen