Winning Space Race with Data Science

SAFDER SHAKIL
29TH OF OCTOBER,2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- The following methodologies were used to analyze the data.
- Data Collection using web scraping and SpaceX API;
- Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
- Machine Learning Prediction.

- Summary of all results
- It was possible to collected valuable data from public sources;
- EDA allowed to identify which features are the best to predict success of launchings;
- Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

- The objective is to evaluate the viability of the new company Space Y to compete with Space X.
- Desirable answers:
  - The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;
  - Where is the best place to make launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data from Space X was obtained from 2 sources:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
    - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

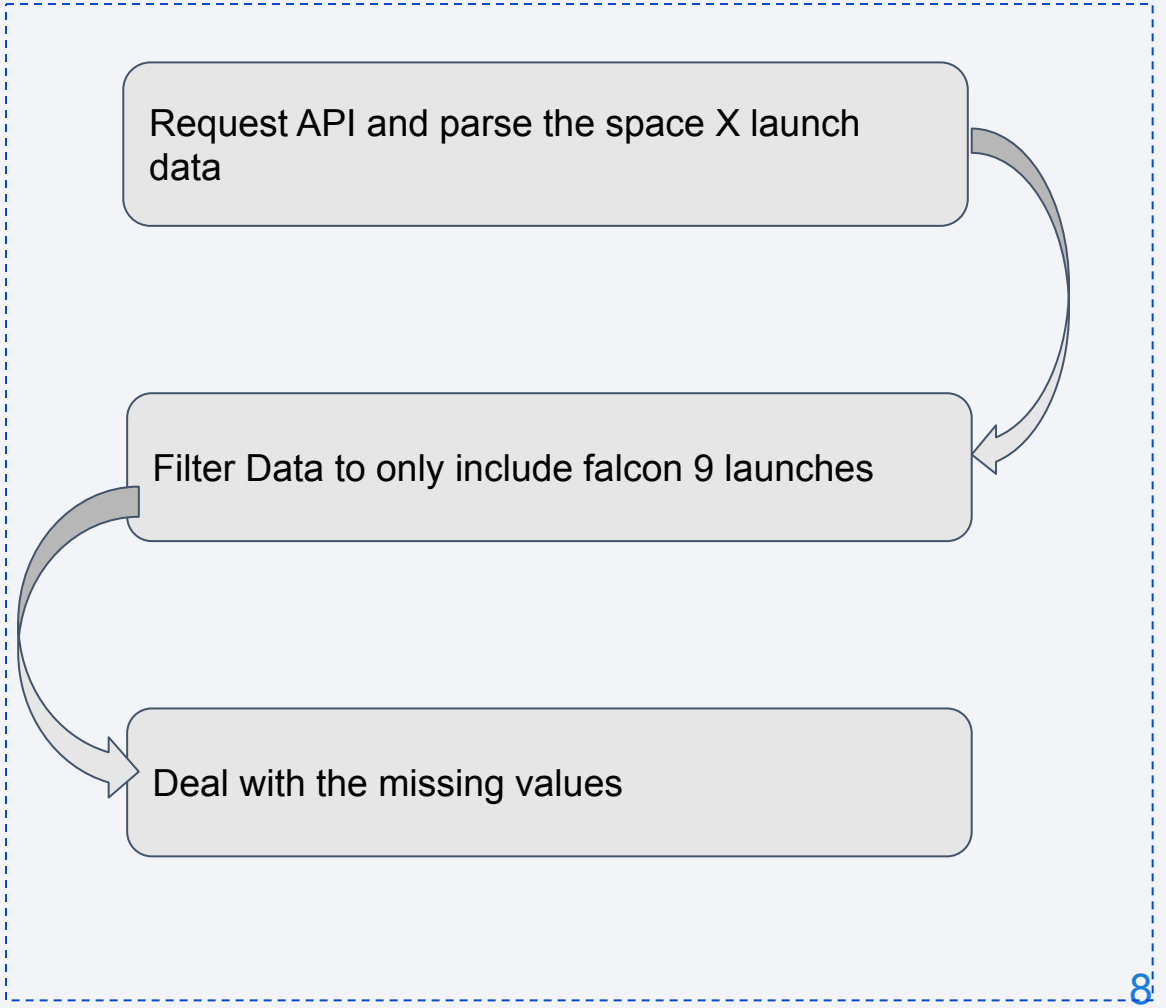
Data Collection

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches), using web scraping technics.

Data Collection – SpaceX API

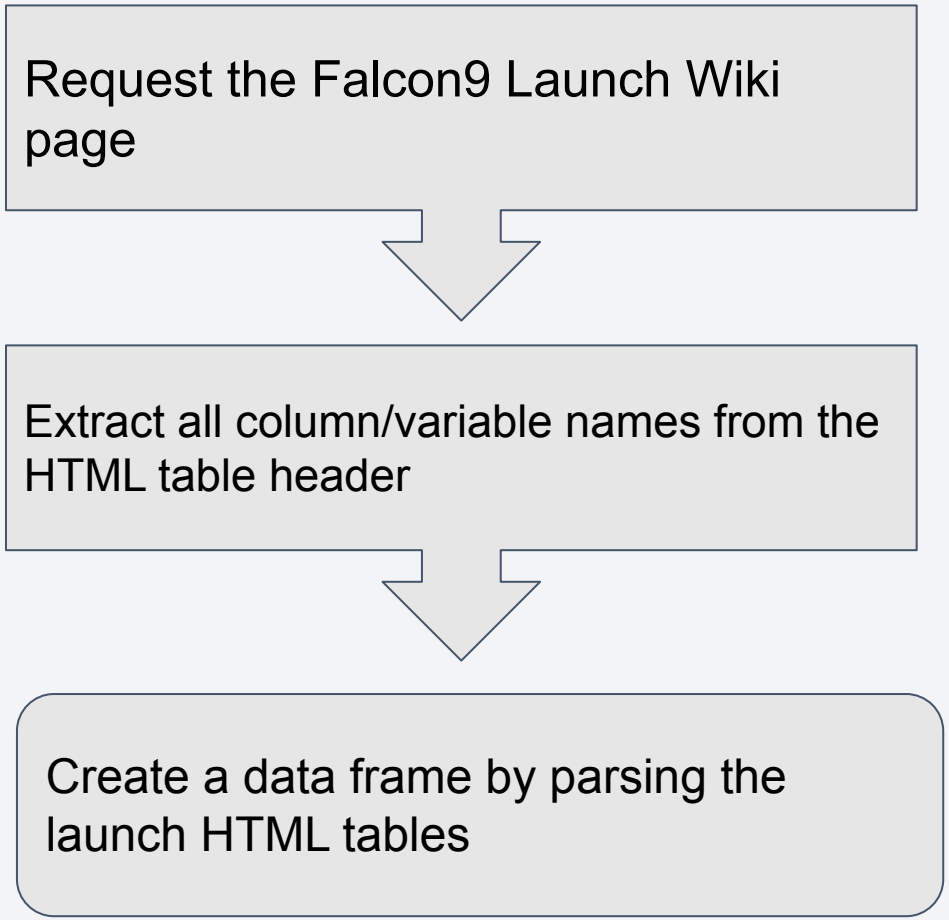
- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.

Gitgub-https://github.com/safder777/Capstone-IBM-Data-Science/blob/main/DATA%20COLLECION%20API%20LAB.ipynb



Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- • Data are downloaded from Wikipedia according to the flowchart and then persisted.
- github link-https://github.com/safder777/Capstone-IBM-Data-Science/blob/main/Data%20Collection%20with%20web%20scraping.ipynb

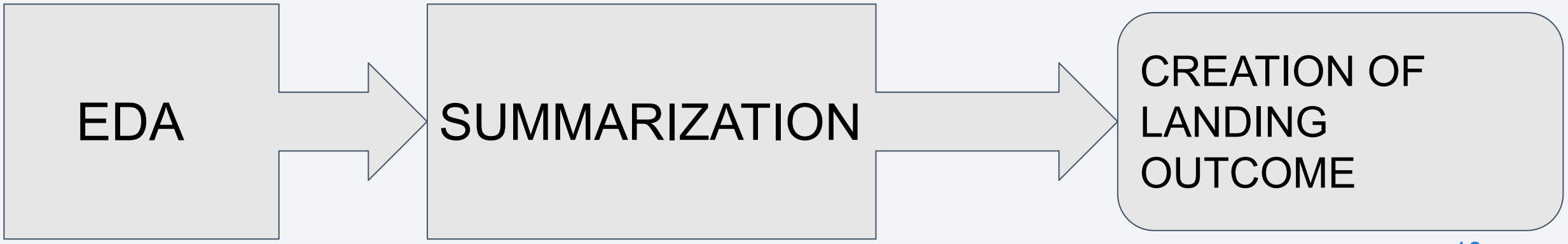


Data Wrangling

Initially some Exploratory Data Analysis (EDA) was performed on the dataset.

- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column

github- https://github.com/safder777/Capstone-IBM-Data-Science/blob/main/Data%20Wrangling.ipynb

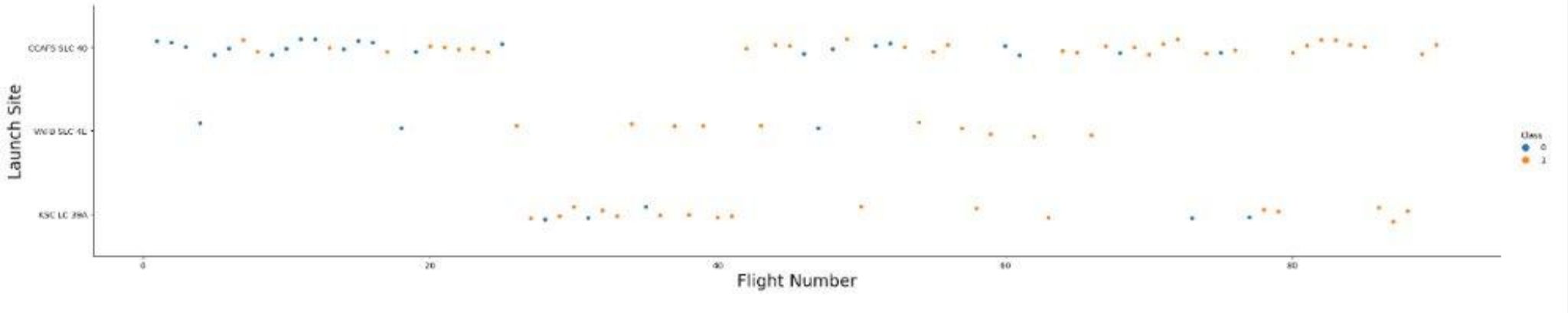


EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:

• Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

GITHUB - https://github.com/safder777/Capstone-IBM-Data-Science/blob/main/jupyter-labs-eda-visualization.ipynb



EDA with SQL

• The following SQL queries were performed:

• Names of the unique launch sites in the space mission;

• Top 5 launch sites whose name begin with the string 'CCA';

• Total payload mass carried by boosters launched by NASA (CRS);

• Average payload mass carried by booster version F9 v1.1;

• Date when the first successful landing outcome in ground pad was achieved;

• Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;

• Total number of successful and failure mission outcomes;

• Names of the booster versions which have carried the maximum payload mass;

• Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and

• 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.

GITHUB- https://github.com/safder777/Capstone-IBM-Data-Science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
- Markers indicate points like launch sites;
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
- Lines are used to indicate distances between two coordinates.

GITHUB-https://github.com/safder777/Capstone-IBM-Data-Science/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

• The following graphs and plots were used to visualize data

• Percentage of launches by site

• Payload range

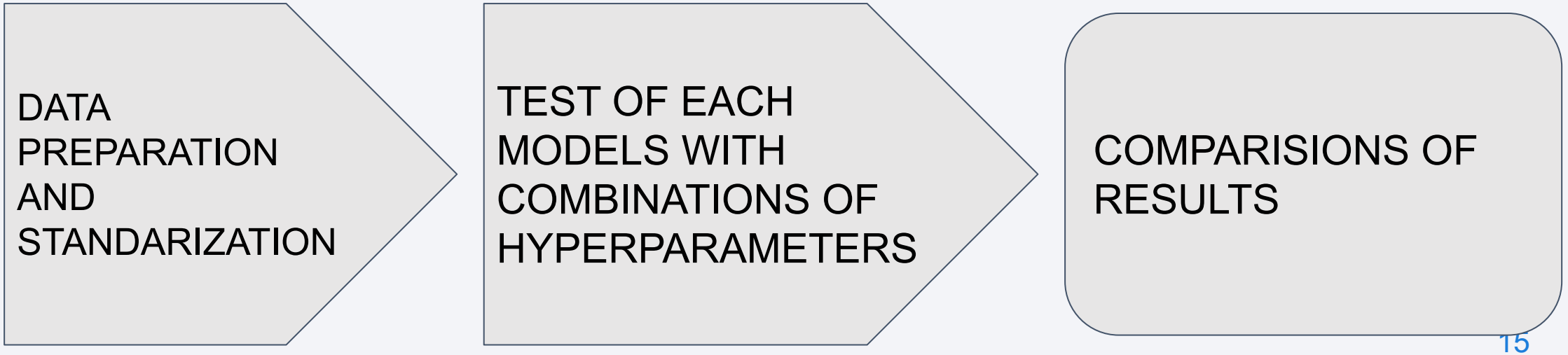
• This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

GITHUB- https://github.com/safder777/Capstone-IBM-Data-Science/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.

GITHUB-
https://github.com/safder777/Capstone-IBM-Data-Science/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



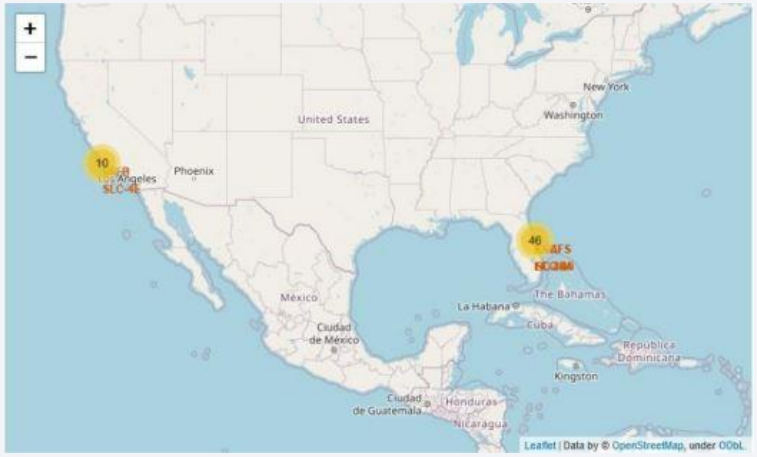
Results

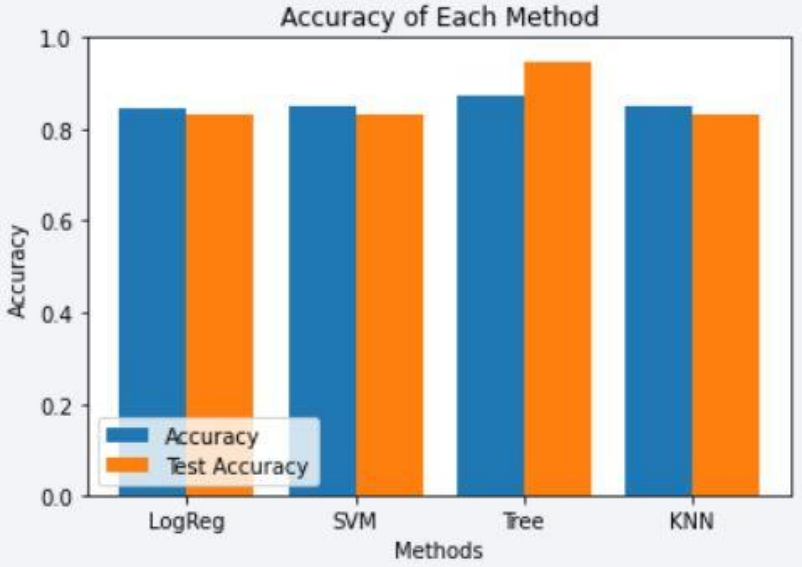
- **Exploratory data analysis results:**
- Space X uses 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

Results

Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.

- Most launches happens at east cost launch sites.

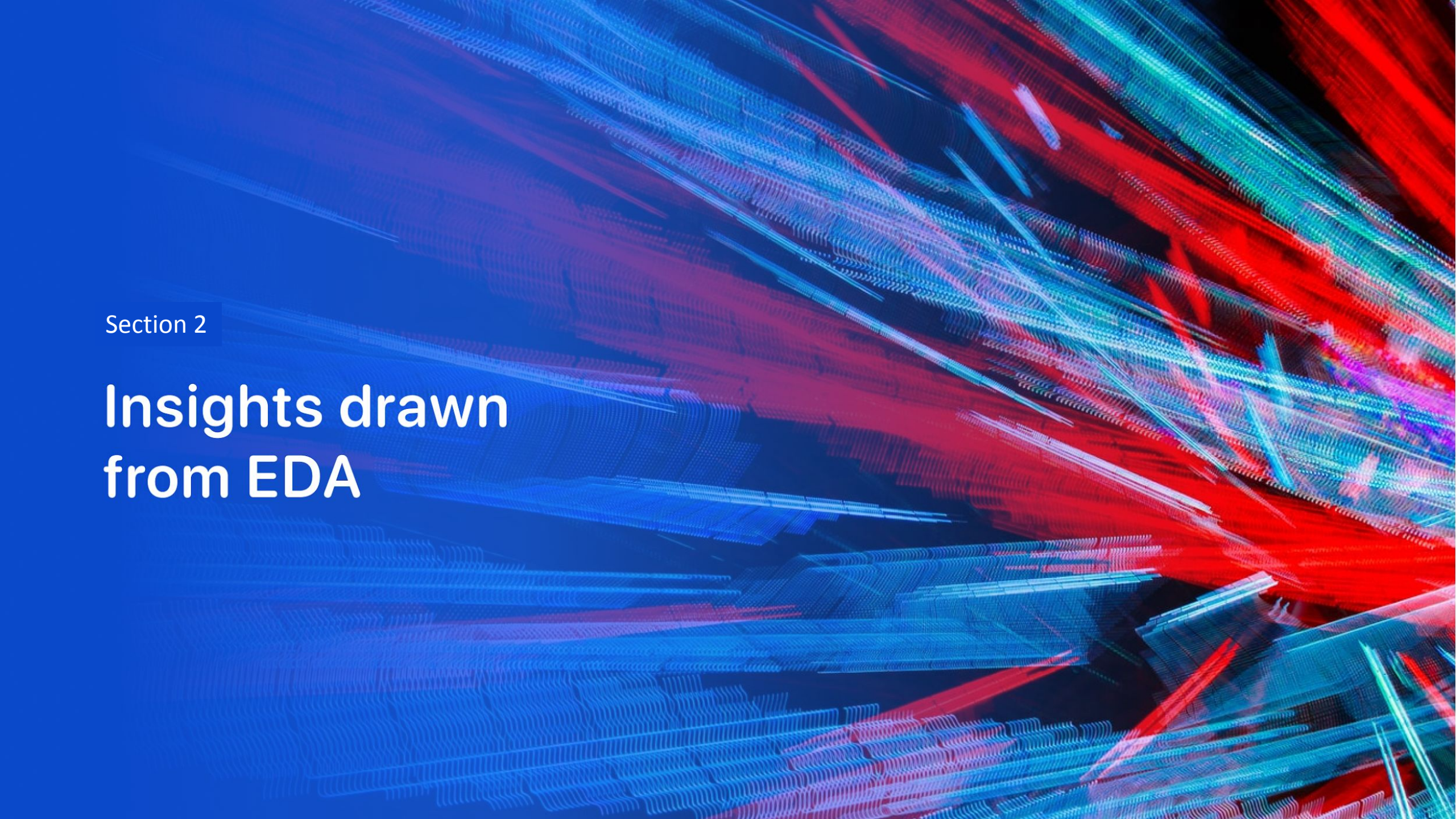




Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.

Section 2

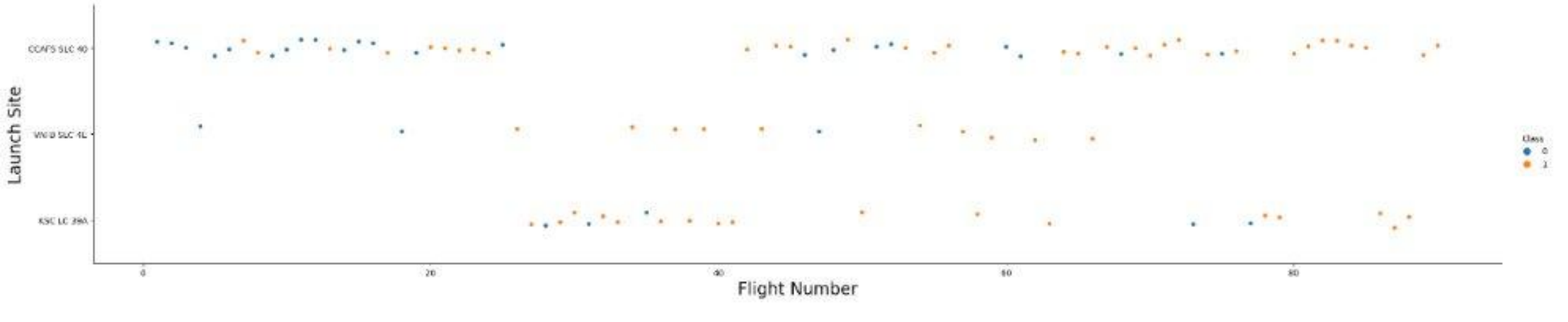
Insights drawn from EDA



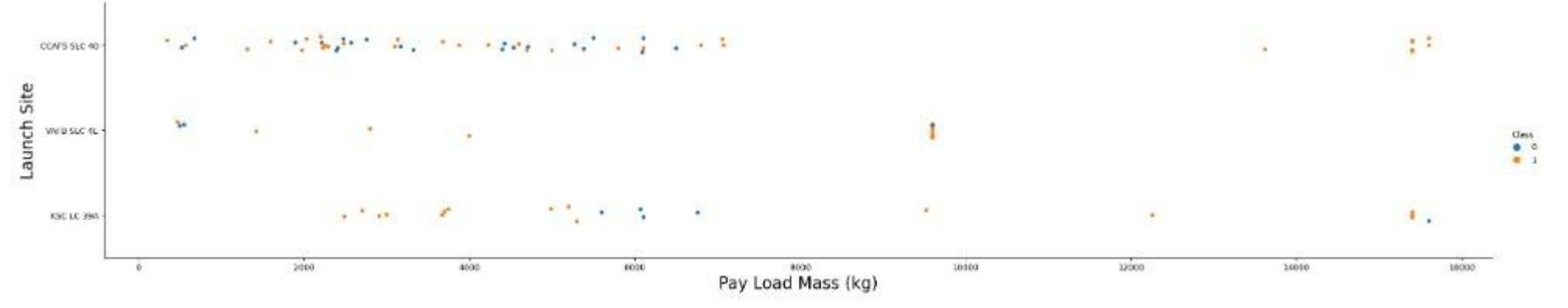
Flight Number vs. Launch Site

According to the plot below, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;

- In second place VAFB SLC 4E and third place KSC LC 39A;
- It's also possible to see that the general success rate improved over time



Payload vs. Launch Site



- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

• The biggest success rates happens to orbits:

• ES-L1;

• GEO;

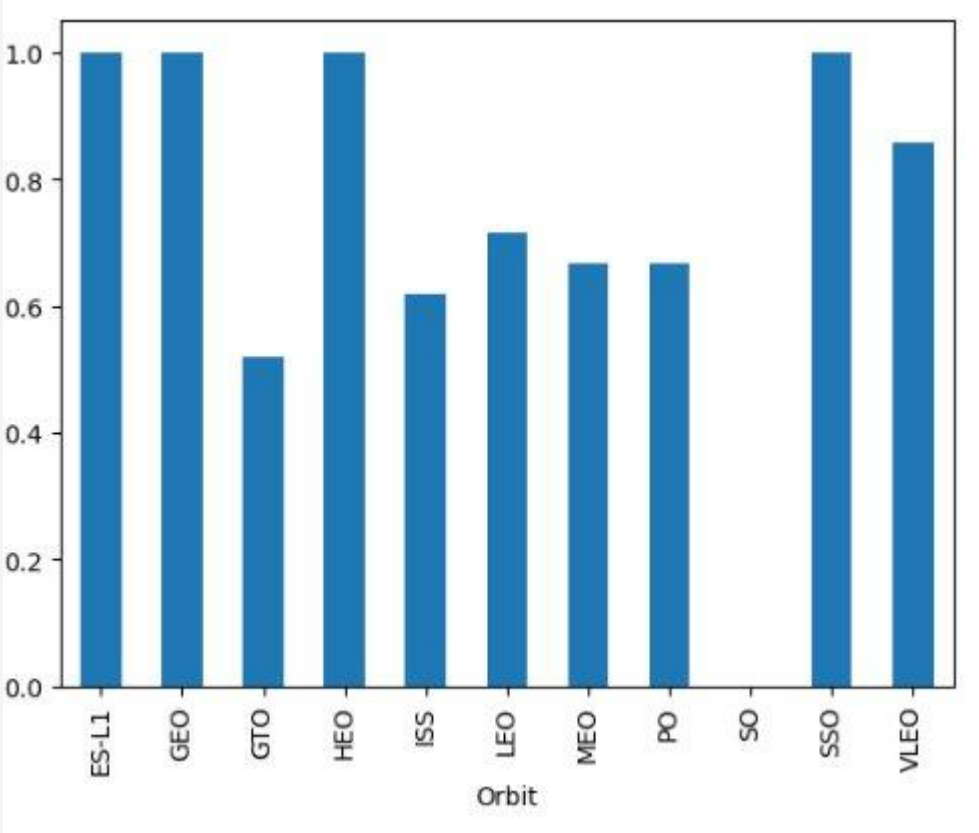
• HEO; and

• SSO.

• Followed by:

• VLEO (above 80%); and

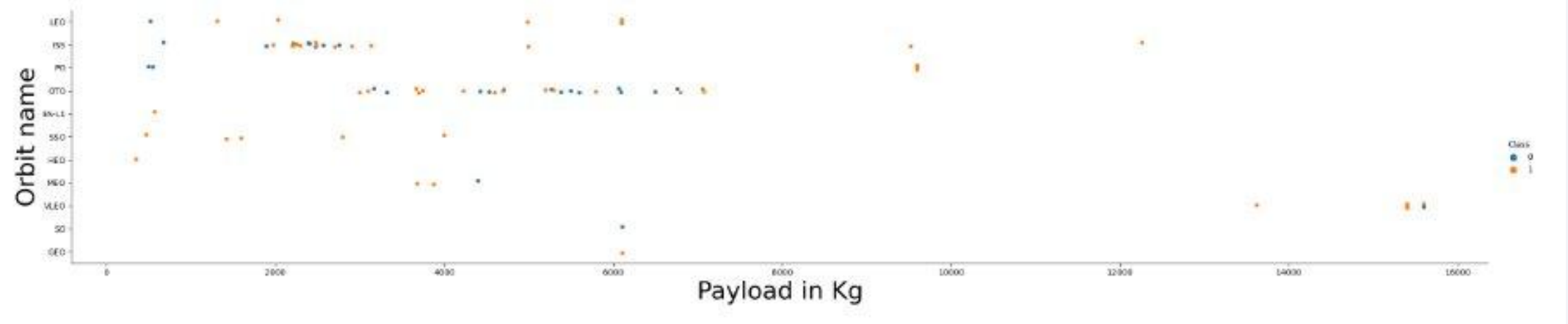
• LFO (above 70%).



Flight Number vs. Orbit Type

- Apparently, success rate improved over time to all orbits;
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency

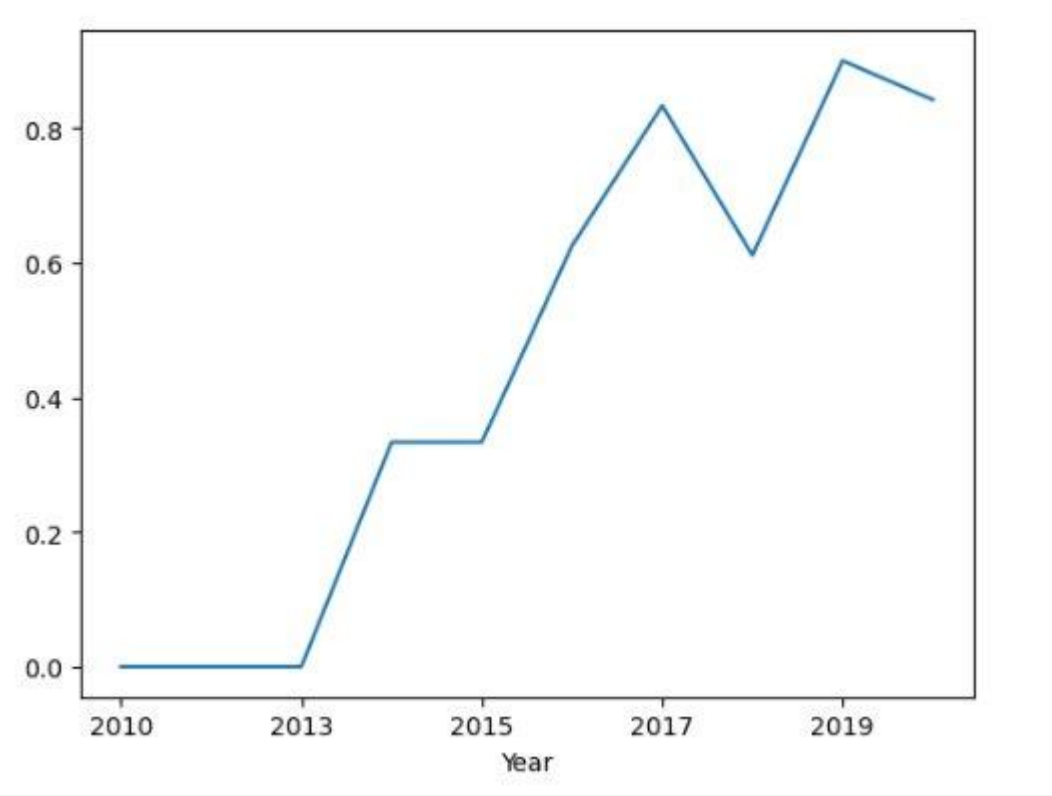
Payload vs. Orbit Type



- Apparently, there is no relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend

The first three years were stagnant, after that the success rate has increased.



All Launch Site Names

- four launch sites are there, they are elected by unique occurrence of launch-sites from data sets.

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

Launch Site Names Begin with 'CCA'

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

- We have the five Cape Carnival launches.

Total Payload Mass

```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%'
```

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

TOTAL_PAYLOAD_MASS
111268

This is the total payload mass carried by the rockets.

Average Payload Mass by F9 v1.1

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

```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS FROM SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1';
```

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

AVG_PAYLOAD_MASS
2928.4

The average payload is calculated by dividing the total mass by the number of launching attempts.

First Successful Ground Landing Date

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

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

min(Date)
01-05-2017

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql

select Booster_Version from SPACEXTBL
where "Landing _Outcome" = "Success (drone ship)"
    and PAYLOAD_MASS__KG_>4000
    and PAYLOAD_MASS__KG_ < 6000
```

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

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

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

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

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

```
[14] %%sql

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

```
 * 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

Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

```
%%sql

select "Landing _Outcome",
    count("Landing _Outcome") as landings
from SPACEXTBL
where Date >= "04-06-2010" and Date <= "20-03-2017"
group by "Landing _Outcome"
order by landings desc
```

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

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017- 03-20


Section 3
Launch Sites
Proximities Analysis

Locations of launch sites on map

- Three in the east
- One in the west
- All in the sout

Display launch outcome by colour.

From the color labels, we can easily see

KSC LC-39A has a rather higher success rate

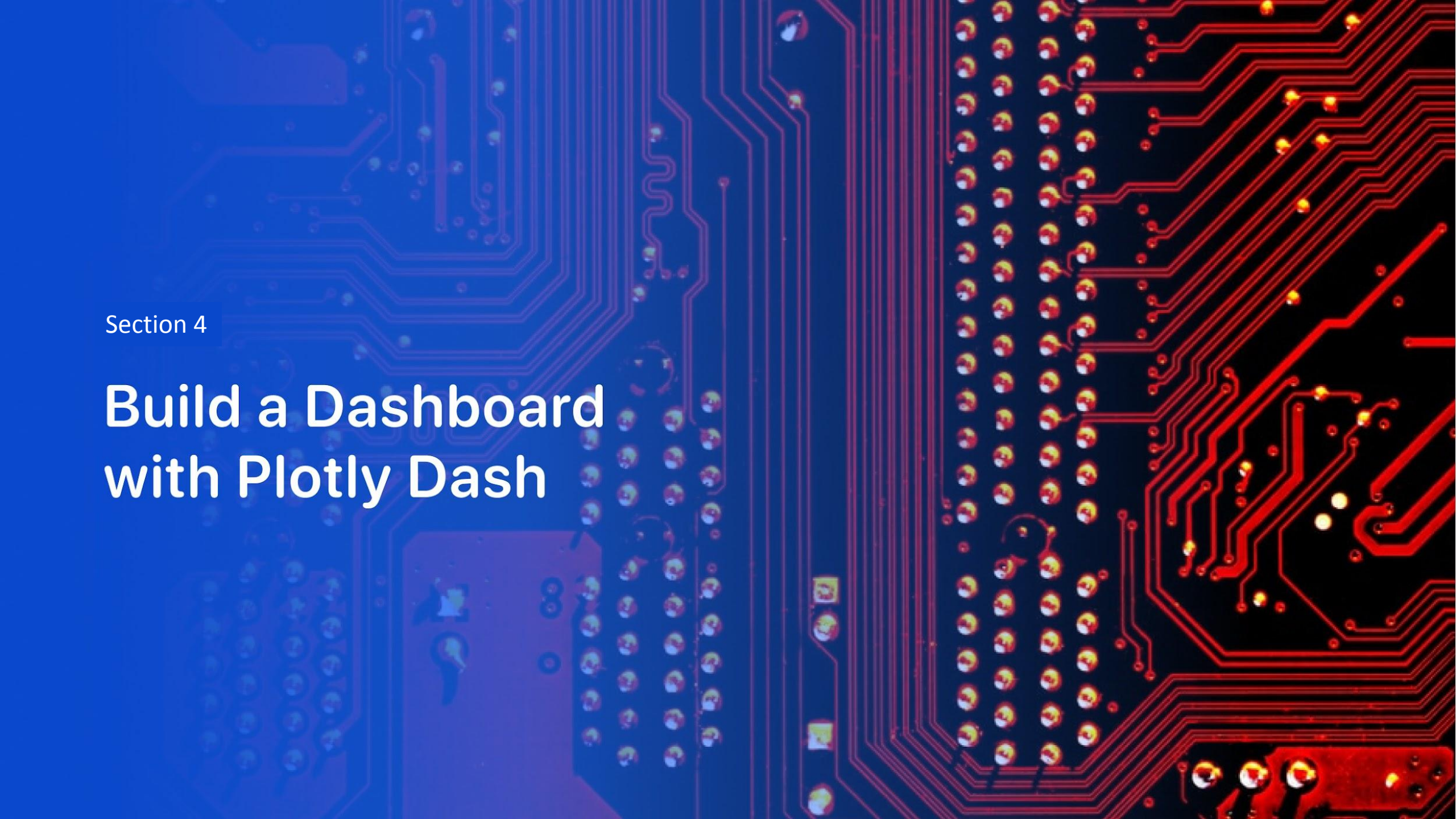
Whereas CCAFS LC-40 and CCAFS SLC-40 have much lower rate

Show distance to proximities.

The distance from KSC LC-39A to the nearest shuttle landing facility is about 4.16 km

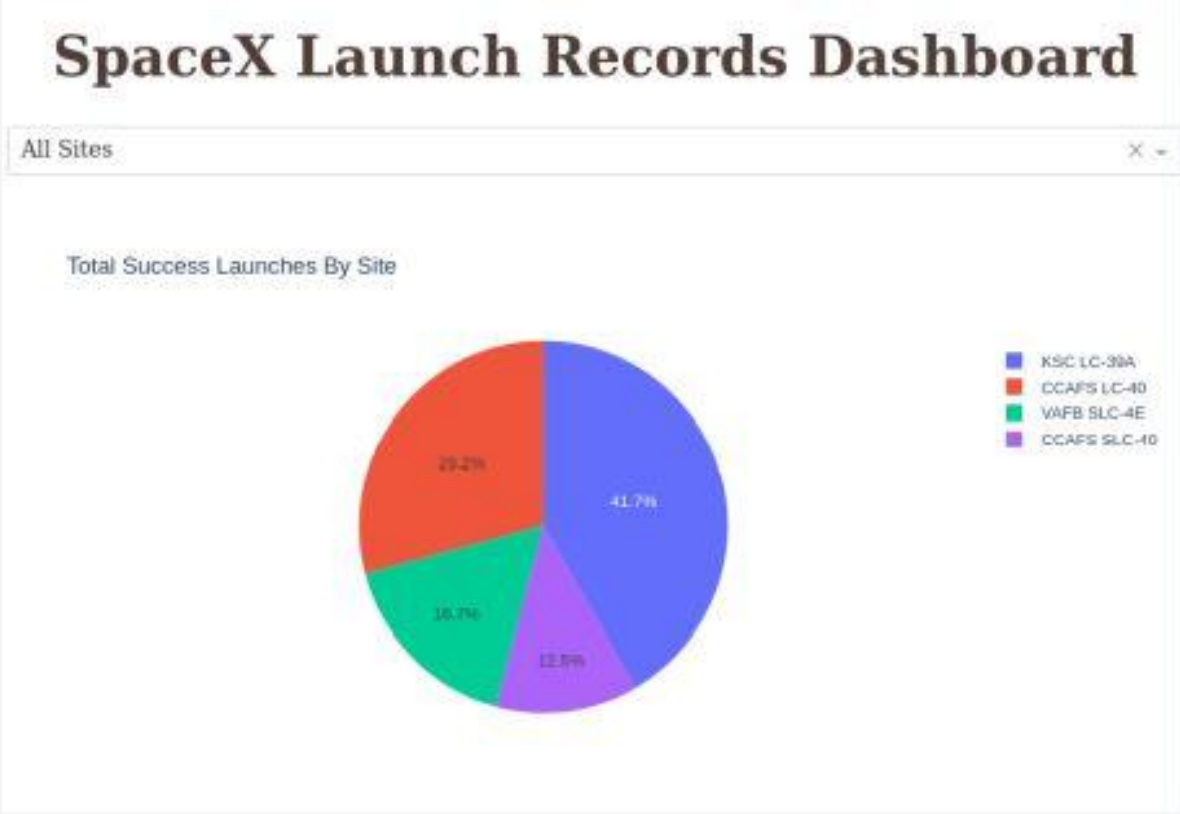
The distance from KSC LC-39A to the nearest highway is less than 1 km.

The distance from KSC LC-39A to the coastline is around 6.5 km


Section 4
Build a Dashboard with Plotly Dash

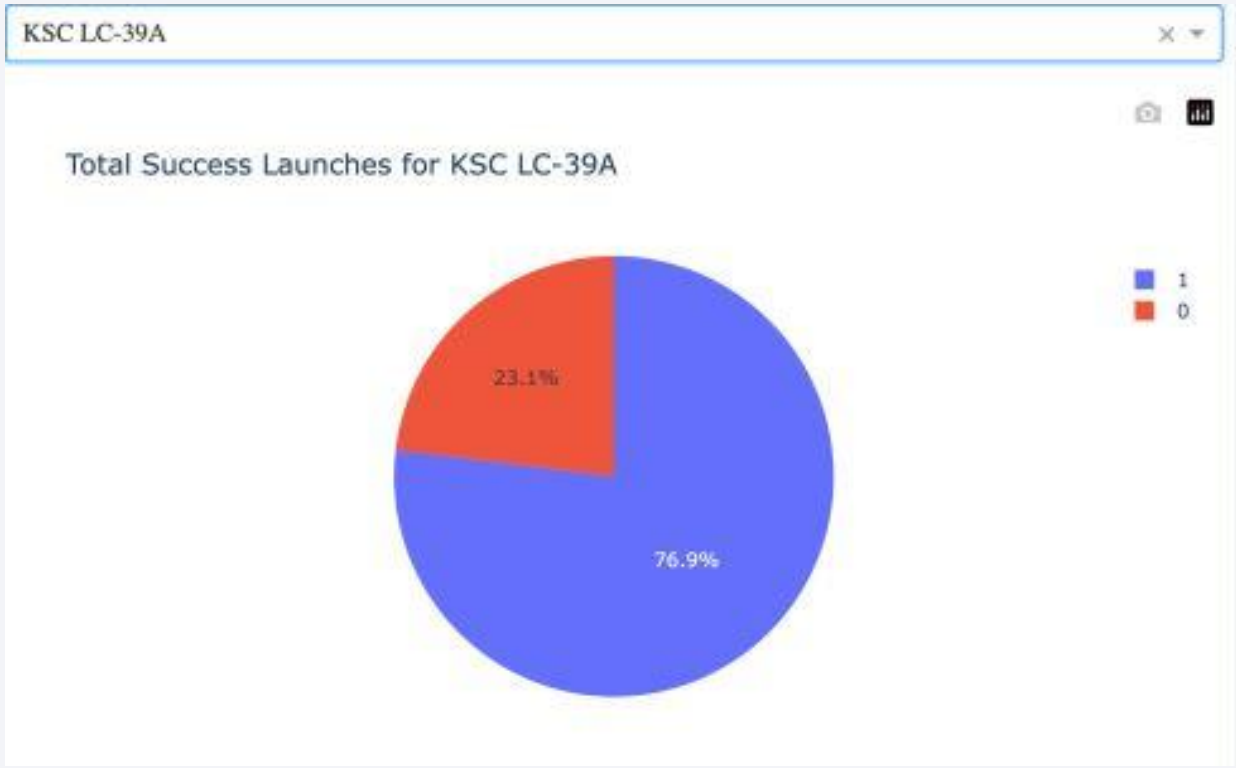
Successful Launches by sites

The place from where launches are done seems to be a very important factor of success of missions

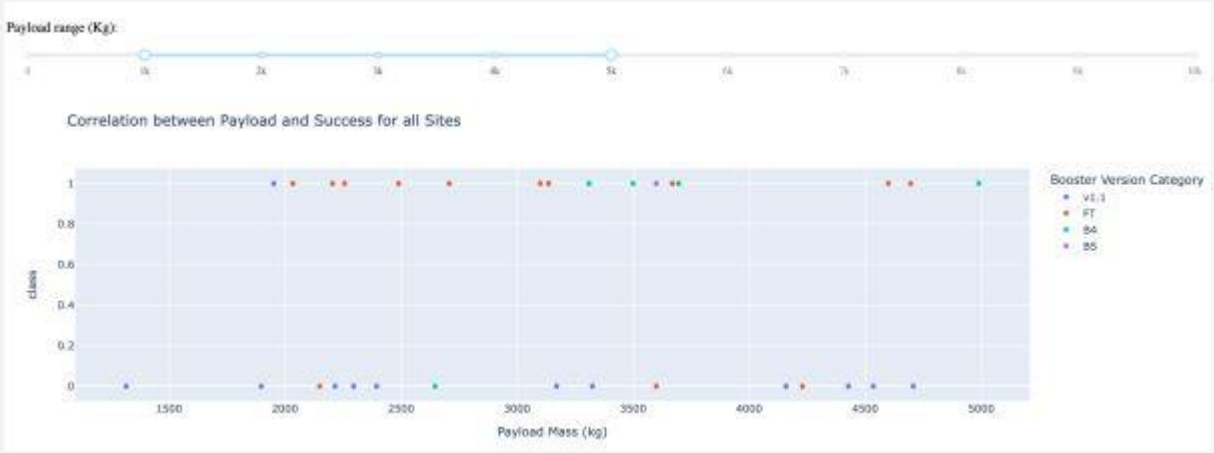


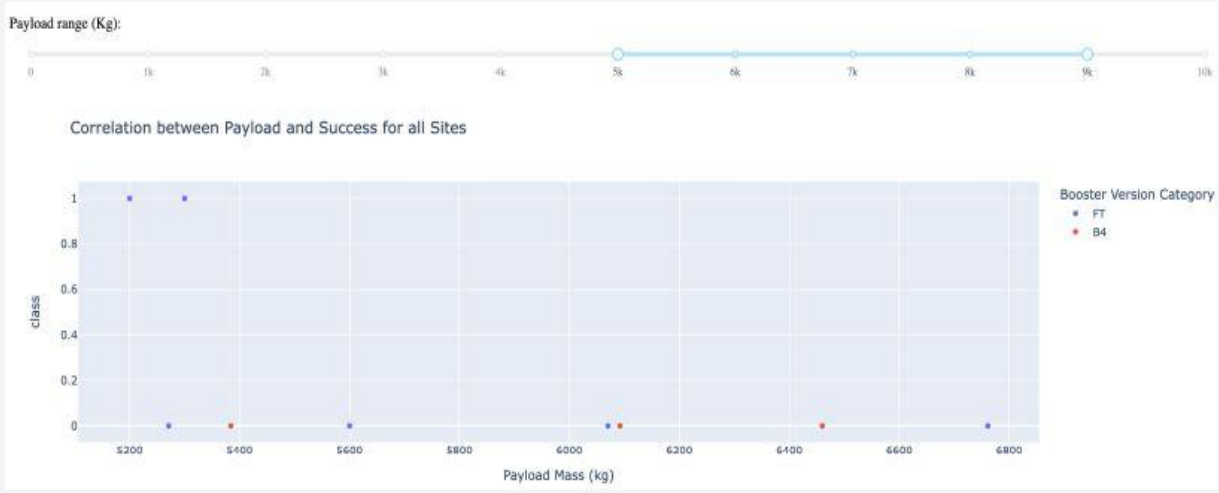
Launch Success Ratio for KSC LC-39A

- 76.9% of launches are successful in this site



Correlation Between Payload and Success





- Payload range in [3000, 4000] has the largest success rate.
- Booster version of FT has the largest success rate

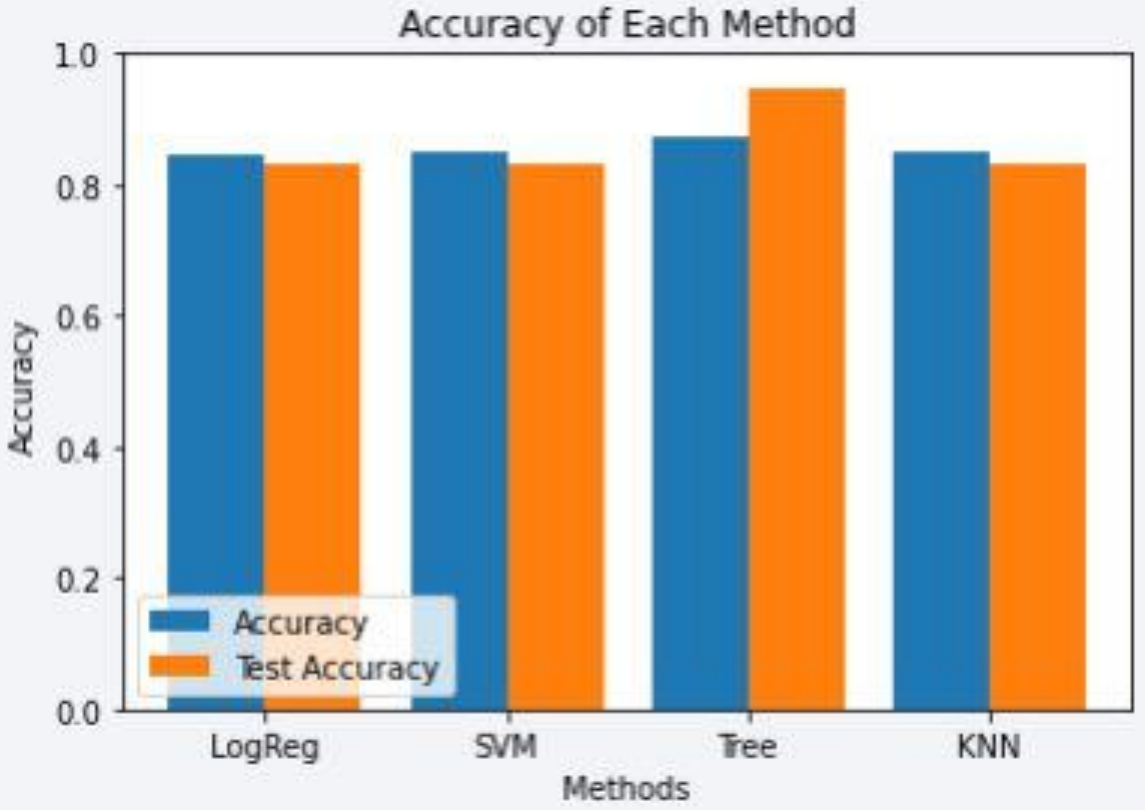
Section 5

Predictive Analysis (Classification)

Classification Accuracy

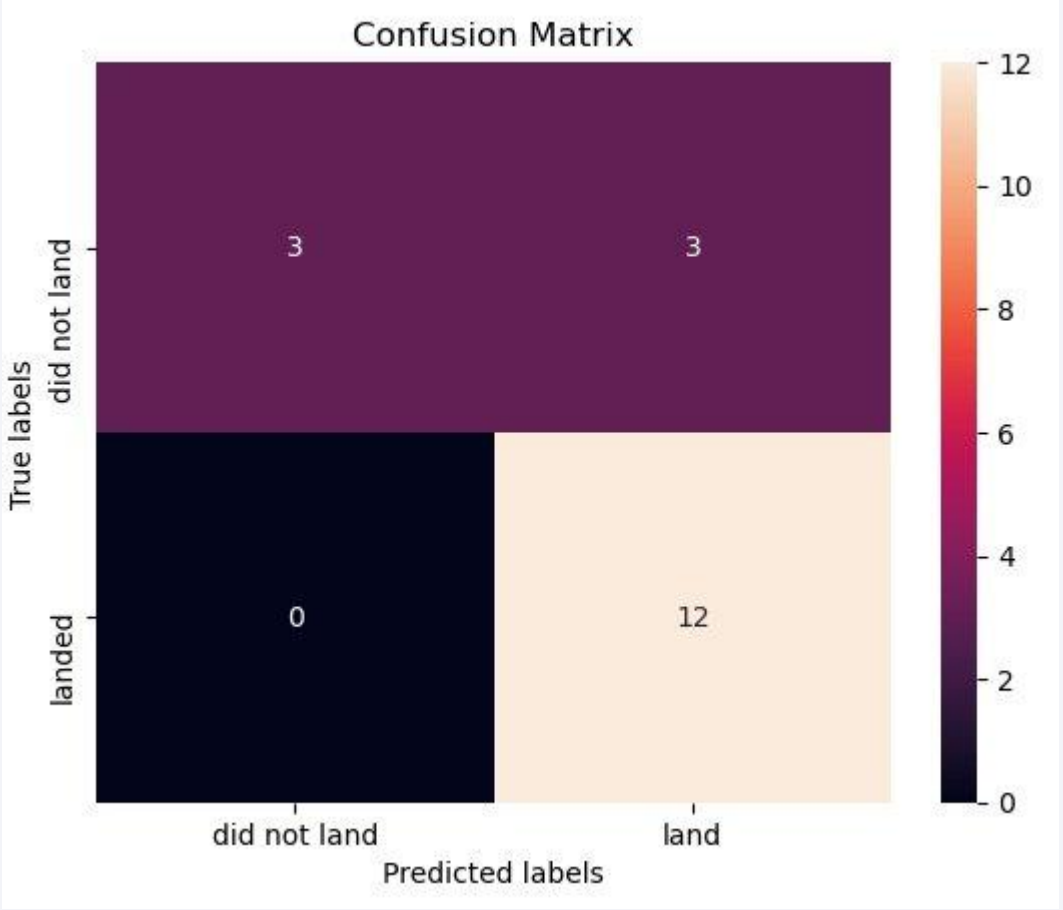
- Four classification models were tested, and their accuracies are plotted beside;

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%



Confusion Matrix

- It show large number of true positives and true negatives rather than false one.



Conclusions

- Different data sources were analyzed, refining conclusions along the process;
- The best launch site is KSC LC-39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- The github link of the entire project is given below:

https://github.com/safder777/Capstone-IBM-Data-Science

Thank you!