Winning Space Race with Data Science

Juan Jose Rivera Lodnoño
2-08-2013

Outline

Executive Summary

SUMMARY OF METHODOLOGIES

- Data collection
- Data wrangling
- EDA with Visualizations and SQL
- Spatial analysis and Dashboarding
- Predictive analysis

SUMMARY OF RESULTS

- High correlation between Launch site, Payload mass and Orbit type and landing success rate.
- Launch sites are very close to coastal zones and railway tracks.
- Best Classification algorithm is a decision tree.
- We can predict with high accuracy the outcome of a booster landing!

Introduction

SUMMARY OF RESULTS.

01

SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars, much of this savings is due to the reuse of the first stage (booster)

02

We want to understand visually and statistically what are the variables that impact on a greater way the successful landing of boosters.

03

We want to predict if a future falcon 9 launch first stage will land successfully on earth or not and potentially determine the cost of a launch

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Request falcon 9 launch records to the SpaceX API and Webscraping of Wikipedia html tables

Perform data wrangling

- Filter for only falcon 9 boosters, replacing of missing values, casting to appropriate data type.

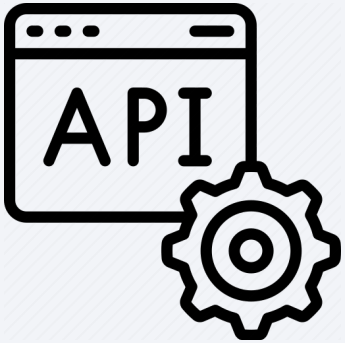
Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Build logistic regression, SVM, Classification trees and KNN, with hyper-parameter tunning and testing with R2 score and confusion matrix

Data Collection – SpaceX API



- Usage of the SpaceX REST API to collect features about flights and landing of mission from SpaceX

API call and conversion of json to pd.DataFrame → Processing of fetched data structure → More API calls for variables like Payload and rockets → Filtering for only Falcon 9 → Replacing of nan values

Link to Data Collection API notebook

Data Collection - Scraping

Usage of BeautifulSoup library to Webscrape Wikipedia html tables with information about features of the mission and the **outcome** of the launch and **landing**

Webscraping wikipage using BeautifulSoap → Usage of regular expressions to extract information → Create a pd.DataFrame with extracted information

Link to web Scraping notebook

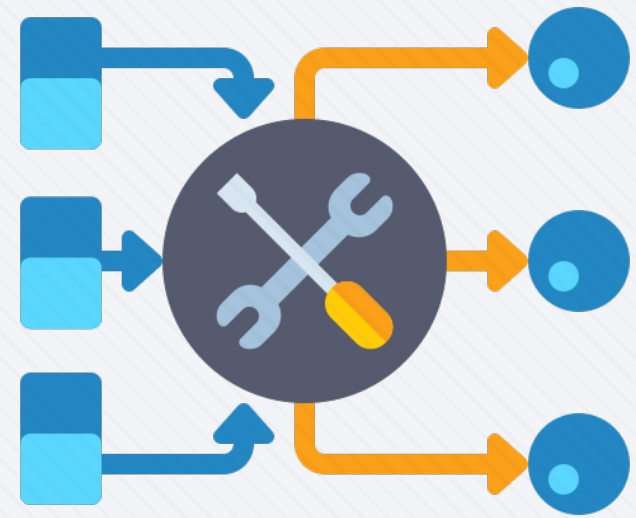
Data Wrangling

Calculate number and percentage of missing values for column LandingPad

Calculate number of missions for each LaunchSite

Calculate number of missions for each Orbit

Convert Landing outcome and type to numeric indicator variable



True ASDS	1
True RTLS	1
True Ocean	1
None None	0
False ASDS	0
False Ocean	0
None ASDS	0
False RTLS	0

Link to data wrangling notebook

EDA with Data Visualization

Scatter plot

Visualize the relationship between variables like payload mass and booter version

Bar charts

Visualize numerical values that change across different categories (Success rate for each Orbit type)

Line charts

Detect trends that ay be present in the data, yearly change of the success rate for booster landing

Link to EDA with data visualization notebook

EDA with SQL

Filter records

Filter records based on success or failure of landing also filtering based on substrings appearing on certain columns

Total results

Total payload mass for certain customers (NASA)

Average results

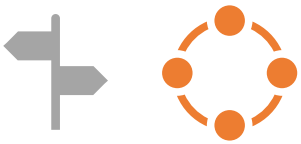
Average payload mass for certain boosters' versions (F9 v1.1)

Date range filtering

Retrieval and count of records happening on a specific date interval.

[Link to EDA with SQL notebook]

Build an Interactive Map with Folium



Circles and Markers

Used to localize Launch sites and give proper naming in the map

Clusters and Icons

Clusters were used to bundle together Launches performed in each launch site and icons were used to categorize successful and failure outcomes

lines

Lines together with markers were used to signal in the map important locations that were close to the launch sites

Link to map notebook

Build a Dashboard with Plotly Dash

Dropdown

Added to filter the data based on the launch site

Bar chart

Used to visualize the success rate for each launch site based on the dropdown filter

Mass range

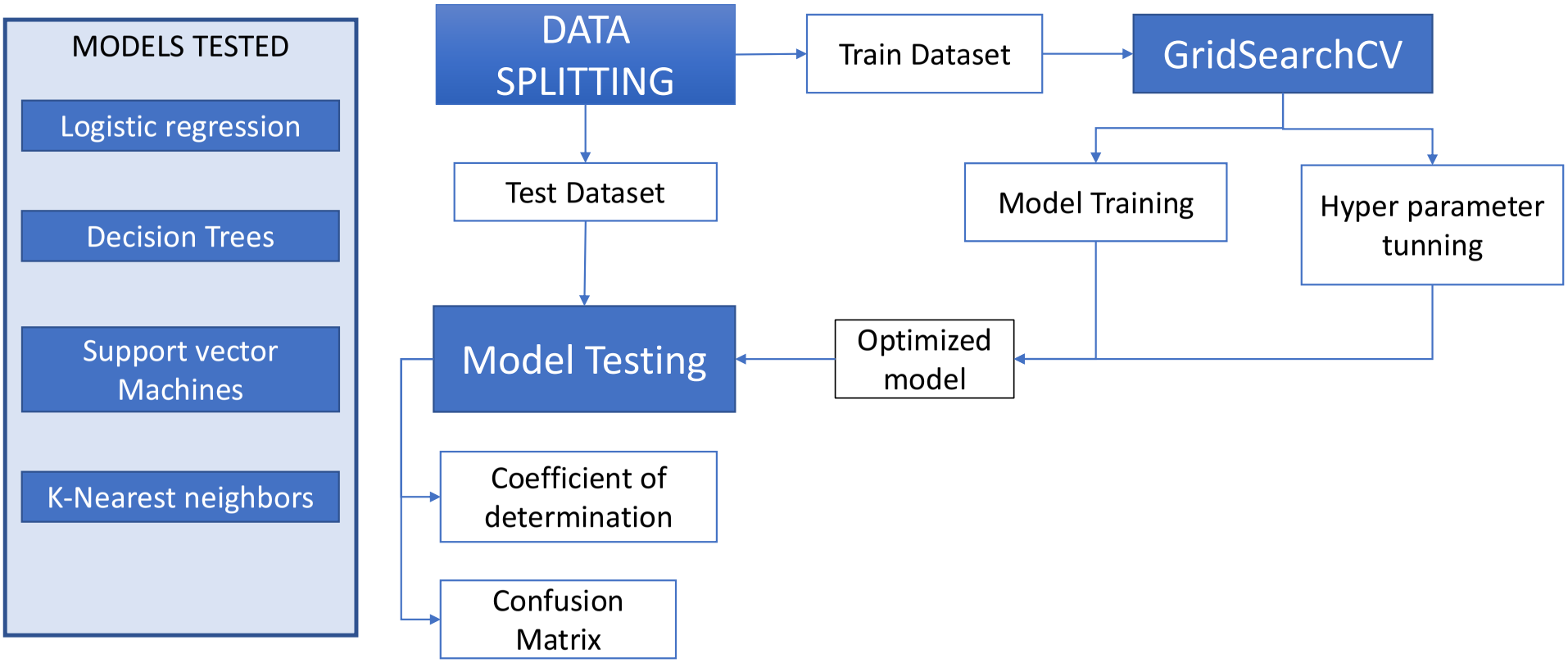
Used to filter the dataset further on a specific payload mass range

Scatter plot

Added to visualize the number of missions that succeeded or failed in landing the booster, filtered by booster version and mass range

Link to Dashboard code

Predictive Analysis (Classification)



Link to Predictive analysis notebook

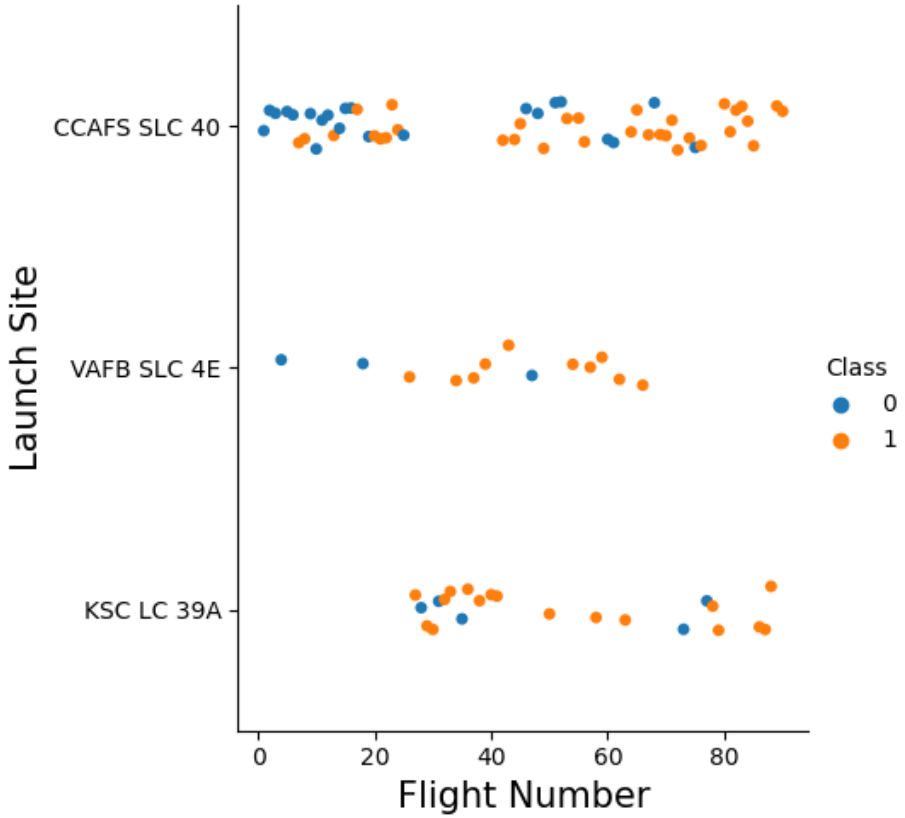
RESULTS

Section 2

Insights drawn from EDA

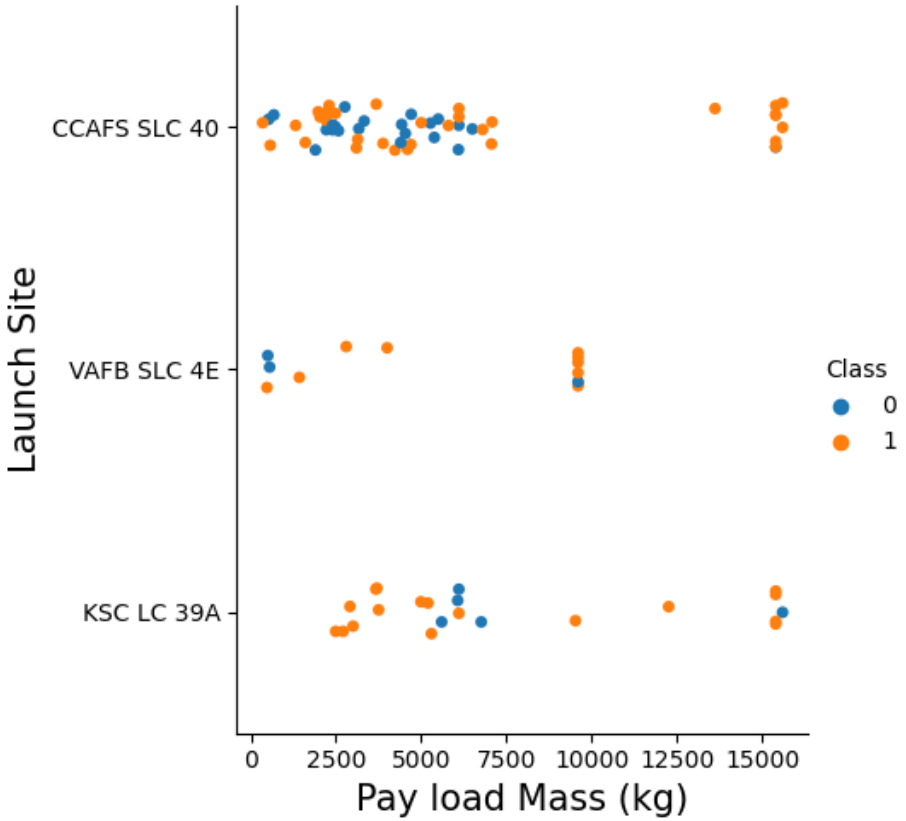
Flight Number vs. Launch Site

- In the x axis we have the Flight number (progresses over time) and on the y axis we have the Launch site, class 0 is failure (blue), class 1 is success (orange)
- As Flight number increases, we see an increase in the success of landing outcomes for every launch site
- The site with the greatest number of launches was CCAFS-SLC 40



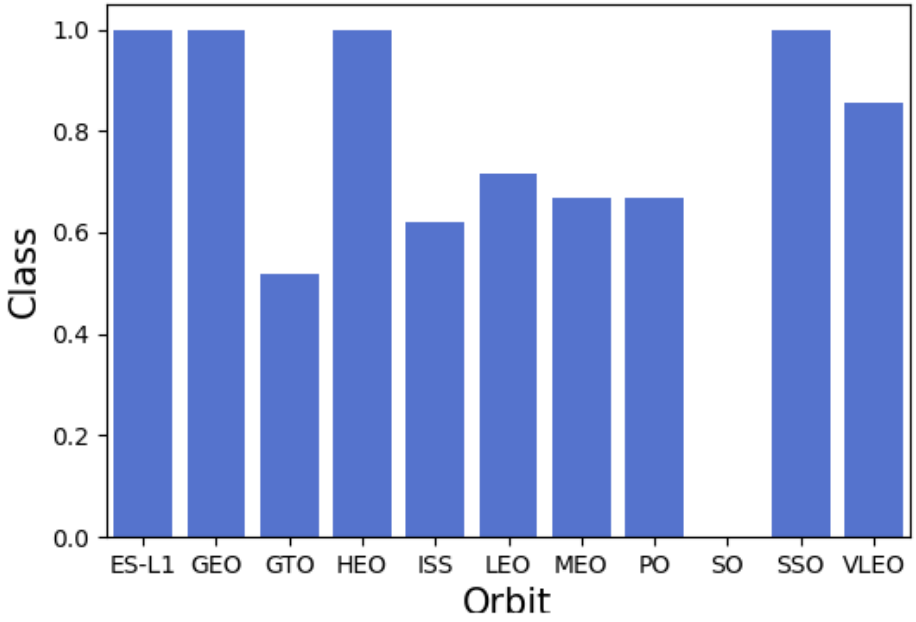
Payload vs. Launch Site

- In the x axis we have payload mass in kilograms, in the y axis we have Launch site type, class 0 is failure (blue), class 1 is success (orange)
- The landing success rate for payloads greater than or equal to 10000 kg greater than the success rate for payload lower than 10000.
- Launch site VAFB SLC 4E was used to launch lighter payloads (<10000)
- For each site, the data seems to cluster around different payload mass intervals.



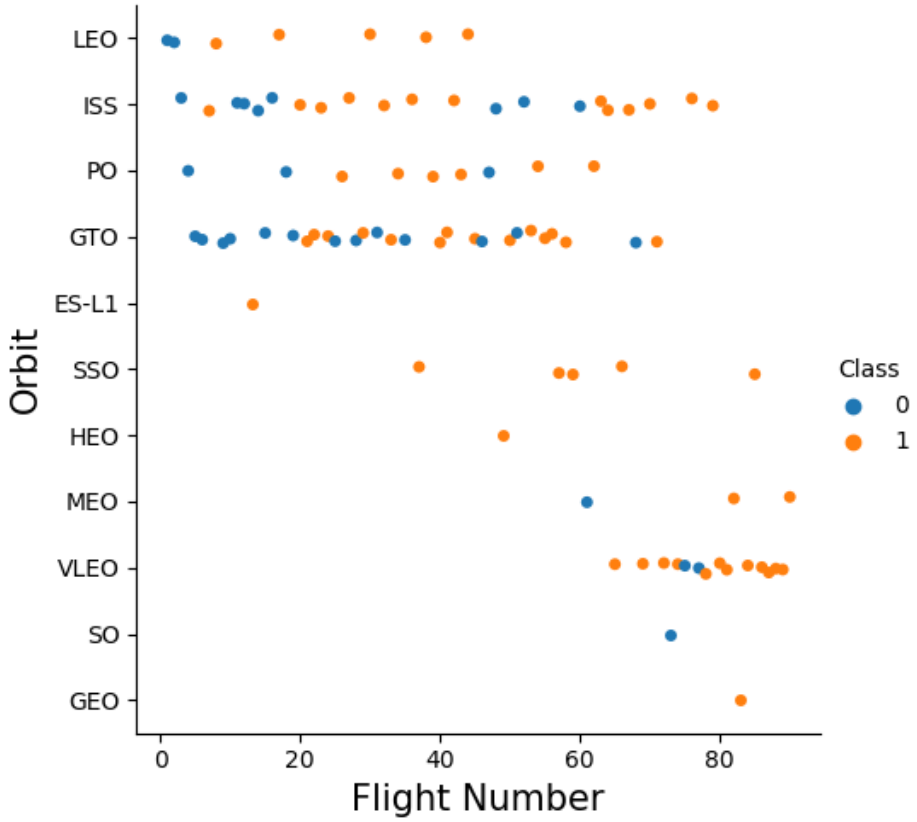
Success Rate vs. Orbit Type

- In the x axis we have the mission orbit type and on the y axis we have the success rate for the missions
- The highest success rates are for orbits ES-L1, GEO and SSO
- The lowest success rates are for orbits GTO, ISS and SO
- 9 out of 11 orbit types have a success rate greater than 50%



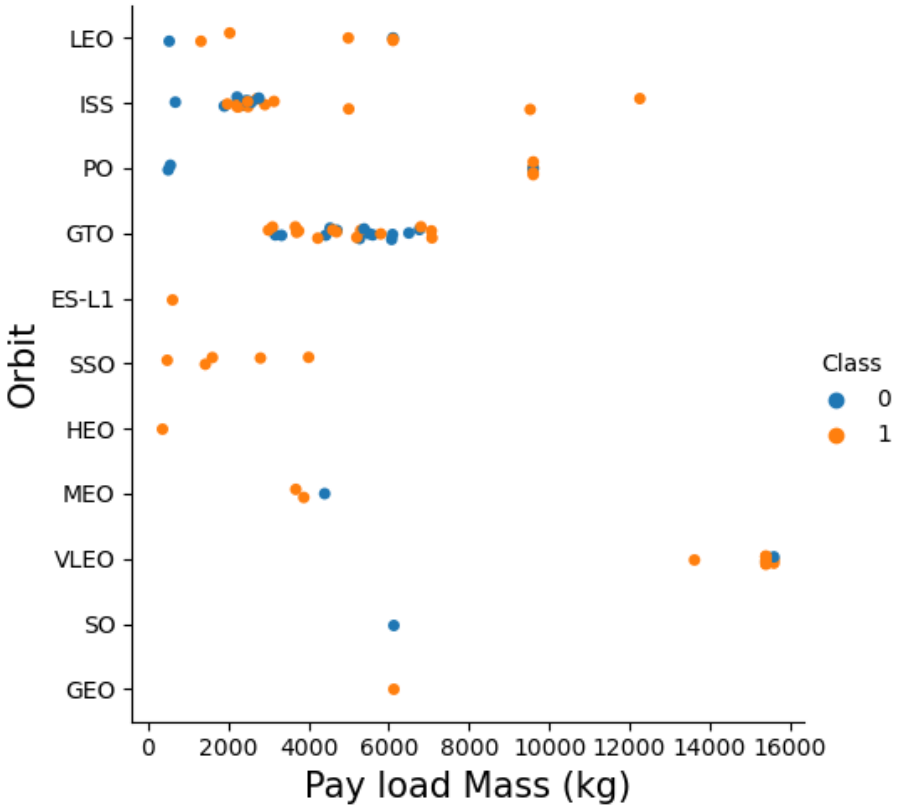
Flight Number vs. Orbit Type

- Show the screenshot of the scatter plot with explanations
- In LEO and PO orbit the Success appears related to the Flight number
- There seems to be no relationship between flight number when in GTO orbit.
- Orbits like ES-L1, HEO, SO and GEO have only been used once in the history of Space X launches.
- More data needs to be collected for orbits with few entries in order to withdraw valid conclusions on the data distribution.



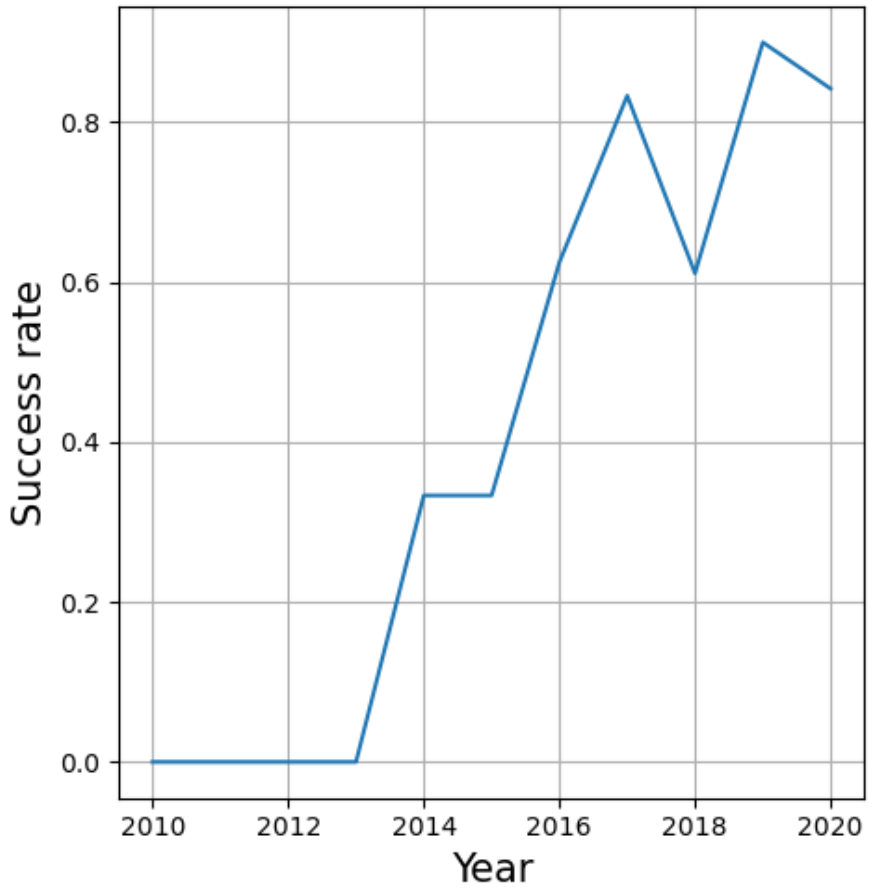
Payload vs. Orbit Type

- With heavy payloads the successful landing rate in the greatest for PO, LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing and negative landing are homogeneous across different payload mass.
- SSO have only positive classes
- Orbit type VLEO is used for very heavy payloads



Launch Success Yearly Trend

- Time Series of success rate, in x axis we have the year and in the y axis we have the success rate
- Since 2013 the success rate have increased until year 2017
- In the period 2017-2018 the success rate showed a downfall
- From 2018 to 2020 the success rate increased again and reached levels achieved previously in 2017.



All Launch Site Names

- We present the names of unique launch sites using the SQL language to create queries to a database

QUERY

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

OUTCOME

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

Launch Site Names Begin with 'CCA'

- We present 5 records where launch site begins with CAA

QUERY

```
%%sql
SELECT "Date", "Time (UTC)", "Booster_Version", "Launch_Site", "PAYLOAD_MASS__KG_ ",
"Orbit", "Customer" , "Mission_Outcome", "Landing _Outcome" FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

OUTCOME

Date	Time (UTC)	Booster_Version	Launch_Site	"PAYLOAD_MASS__KG_ "	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	PAYLOAD_MASS__KG_	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	PAYLOAD_MASS__KG_	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	PAYLOAD_MASS__KG_	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	PAYLOAD_MASS__KG_	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	PAYLOAD_MASS__KG_	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- SQL query use to calculate the total payload carried by boosters from NASA

QUERY

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

OUTCOME

TOTAL PAYLOAD MASS
45,596 Kg

Average Payload Mass by F9 v1.1

- SQL query use to calculate the average payload carried by booster version F9 v1.1 from NASA

QUERY

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE Booster_Version='F9 v1.1';
```

OUTCOME

AVERAGE PAYLOAD MASS FOR F9v 1.1
29,28.4 Kg

First Successful Ground Landing Date

- SQL query used to find the date of the first successful landing outcome on ground pad

QUERY

```
%%sql
SELECT "Landing _Outcome", substr("Date",7,4) ||'-'|| substr("Date",4,2) ||'-'|| substr("Date",1,2) as "Date"
FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%(ground pad)' LIMIT 1;
```

OUTCOME

FIRST SUCCESSFUL LANDING DATE
2015/12/22

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL query used to List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

QUERY

```
%%sql
SELECT "Date", "Time (UTC)", "Booster_Version", "Launch_Site", "PAYLOAD_MASS__KG_ ",
"Orbit", "Customer" , "Mission_Outcome", "Landing _Outcome" FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

OUTCOME

Date	Time (UTC)	Booster_Version	Launch_Site	"PAYLOAD_MASS__KG_ "	Orbit	Customer	Mission_Outcome	Landing _Outcome
06-05-2016	05:21:00	F9 FT B1022	CCAFS LC-40	PAYLOAD_MASS__KG_	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
14-08-2016	05:26:00	F9 FT B1026	CCAFS LC-40	PAYLOAD_MASS__KG_	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	PAYLOAD_MASS__KG_	GTO	SES	Success	Success (drone ship)
11-10-2017	22:53:00	F9 FT B1031.2	KSC LC-39A	PAYLOAD_MASS__KG_	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- SQL query used to calculate the total number of successful and failure mission outcomes

QUERY

```
%%sql
SELECT "Mission_Outcome", COUNT(*) from SPACEXTBL GROUP BY "Mission_Outcome";
```

OUTCOME

Mission Outcome	COUNT
Failure (in flight)	1
Success (Payload status unclear)	99
Success	1

Boosters Carried Maximum Pay load

QUERY

```
%%sql
SELECT DISTINCT("Booster_Version") FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

OUTCOME

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

SQL query used to find the booster versions that carried de maximum payload mass

2015 Launch Records

- 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

QUERY

```
%%sql
SELECT substr("Date", 4, 2) AS MONTH, "Landing _Outcome",
"Booster_Version", "Launch_Site" FROM SPACEXTBL
WHERE substr("Date",7,4) = '2015'
AND "Landing _Outcome" = "Failure (drone ship)";
```

OUTCOME

MONTH	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

QUERY

```
%%sql
SELECT "Landing _Outcome", COUNT(*) as COUNT FROM
(SELECT * FROM SPACEXTBL WHERE substr("Date",7,4) ||'-'|| substr("Date",4,2)
 ||'-'|| substr("Date",1,2) BETWEEN "2010-06-04" AND "2017-03-20"
AND "Landing _Outcome" LIKE 'Success%') GROUP BY
"Landing _Outcome" ORDER BY "COUNT(*)" DESC;
```

OUTCOME

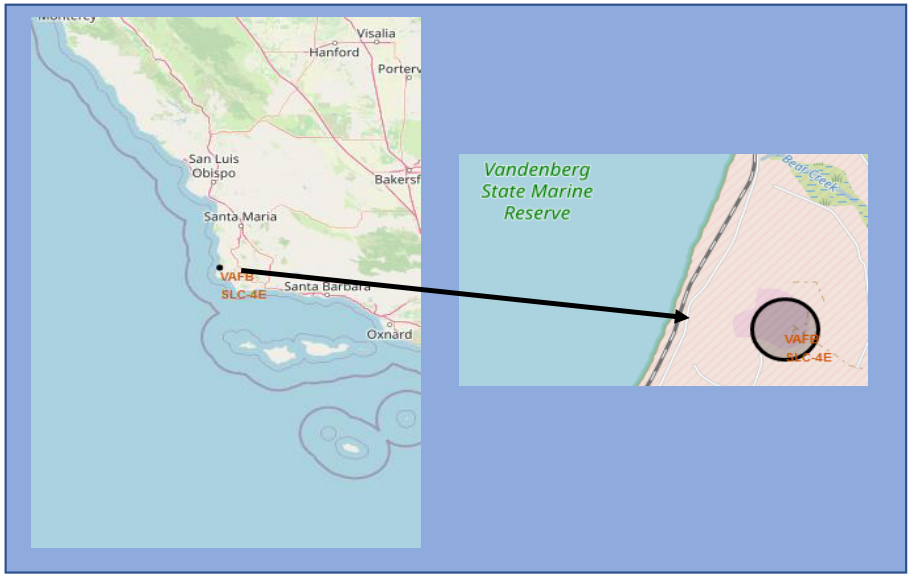
- Rank of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

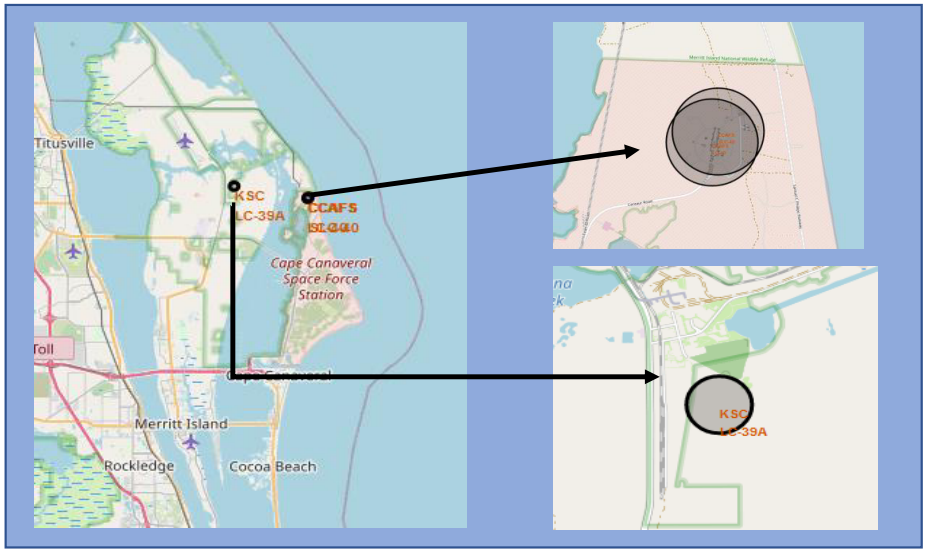
Landing Outcome	COUNT
Success (Ground pad)	3
Success (Drone Ship)	5

Section 3

Launch Sites Proximities Analysis

Locations of launch sites





- The top image shows VAFB SLC-4E Launch site, localized in the coast of California inside the Vanderberg space force base and near the Lompoc town.
- The bottom image shows KSC LC-39A, CCAFS SLC-40 and CCAFS LC-40 launch sites, Localized near the coast of Florida state inside the Cape Canaveral space force station and Merritt island.

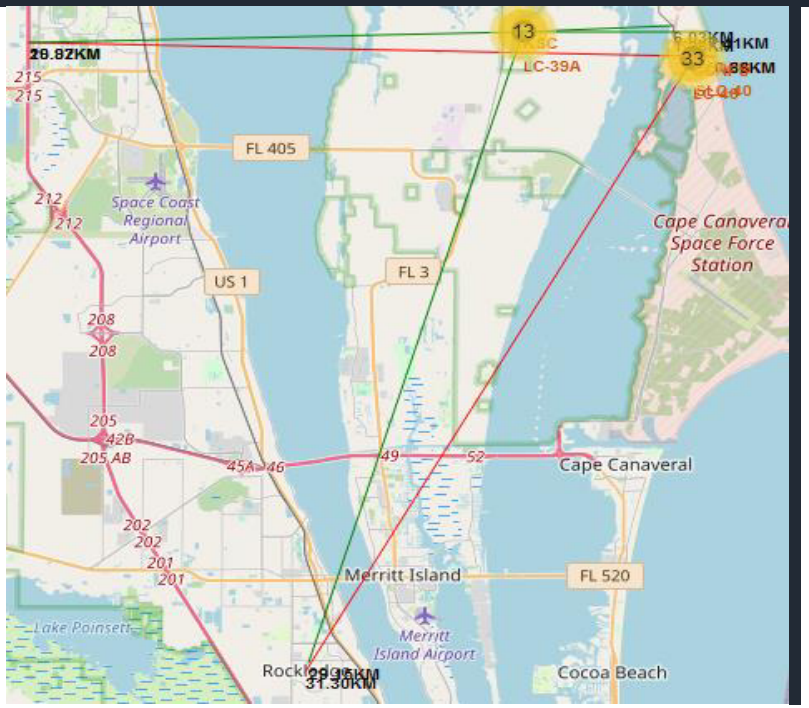
Success and failure markers for launch records

Launch site VAFB SLC-4E has 10 launches in total with 4 (green) being successful landings and 6 being failures (red).

Launch site VKSC LC-39A has 13 launches in total with 10 (green) being successful landings and 3 being failures (red).

CCAFS LC-40(left) with 26 launches, 7 successful and 19 failures and CCAFS SLC-40 (Right) with 7 launches, 3 successful and 4 failures.

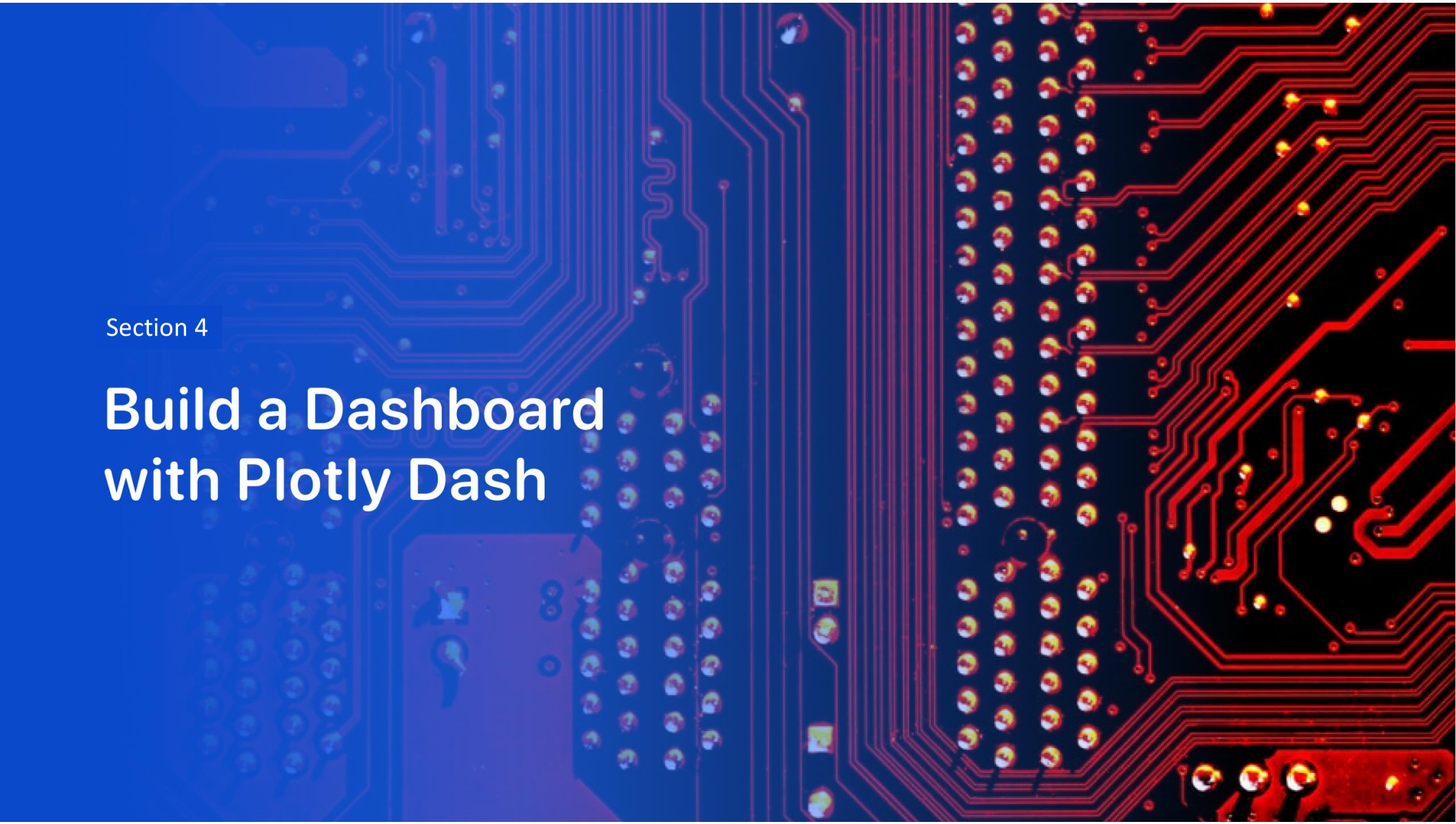
Distance between launch sites and relevant points of interest



Folium map showing launch sites and distance to its proximities such as railway, highway, coastline.

Site	City [Km]	Highway [Km]	Railway [Km]	Coastline [Km]
VAFB SLC -4E	14.01	14.93	1.3	1.44
KSC LC-39A	29.15	28.82	6.03	7.41
CCAFS SLC-40	31.30	28.82	1.28	0.86

- For all launch sites the pattern is similar, they are very close to railways and the coastline and relatively far from Cities and highways.



Section 4

Build a Dashboard with Plotly Dash

All Sites

Total Success Launches by Site

Success rate for all launch sites.

- 47 percent of all the successful landings are from KSC LC-39A launch site.
- CCAFS LC-40 and CCAFS SLC-40 are the launch sites in which most of the missions have taken place.
- Lowest success rate is for CCAFS SLC-40 launch site across all the missions.

SpaceX Launch Records Dashboard

CCAFS SLC-40

Total Succes Launches for Site CCAFS SLC-40

Highest success rate for a specific site.

- Individually CCAFS SLC-40 is the site with the most successful landing in relation with its failed landings

Payload vs. Launch Outcome scatter plot for all sites

- In the payload range of 1000 to 6000 kilograms the booster type FT has the highest success rate across al the rest of booster versions for this payload mass range.

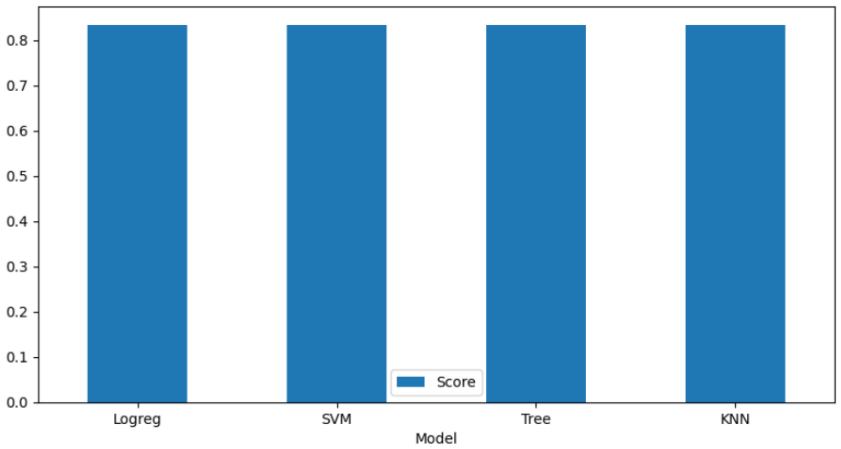
Section 5

Predictive Analysis (Classification)

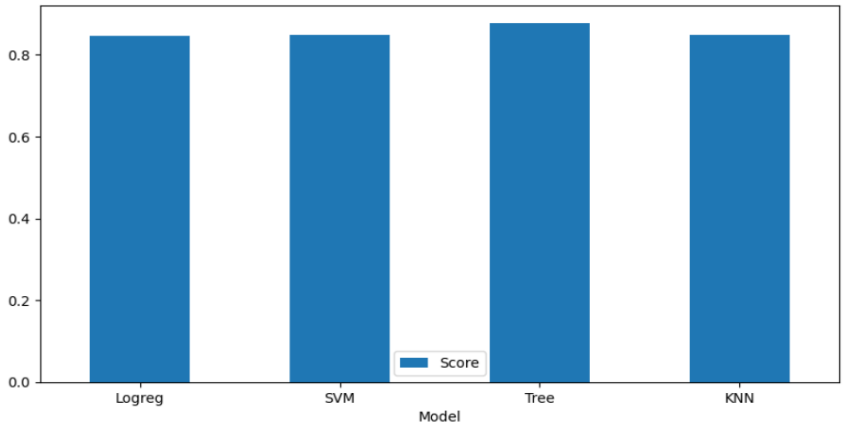
Classification Accuracy

We can see the accuracy for the different models for the test data and for the train data, on the train data all the models performed the same with an accuracy of 0.833, we look at the training data the score is higher which is expected, because of the ambiguity resulting from the same score on the test dataset, I have chosen the best model based on the score obtained in the trins dataset, the highest score is from the **Classification Tree** with 0.87 points

TEST DATA MODEL SCORE

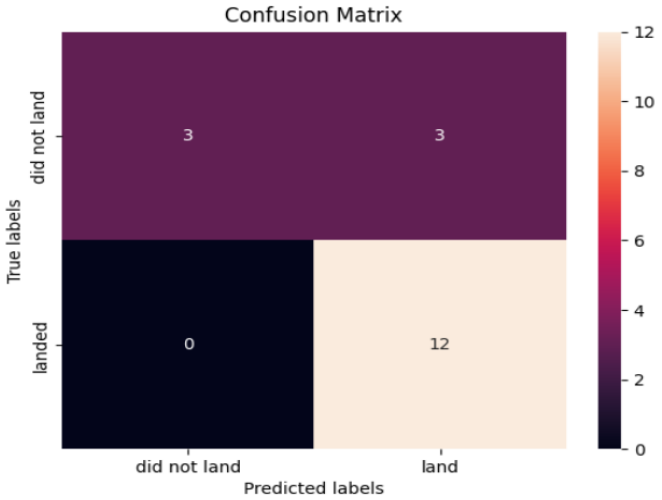


TRAIN DATA MODEL SCORE



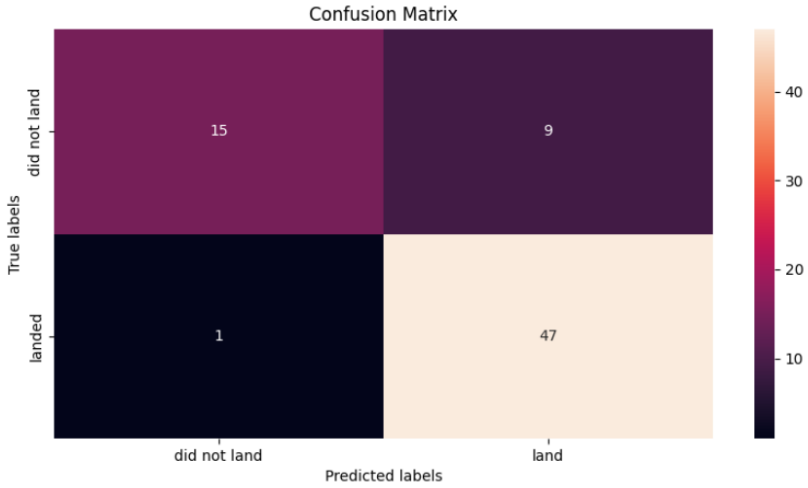
Confusion Matrix

TEST DATA MODEL MATRIX



In this case good predictions, in the case in which the rocket lands or model always predicted that it would land, we have no cases of false negatives and only 3 false positive cases.

TRAIN DATA MODEL MATRIX



In the case in which the rocket lands our model only fails 1/48 to predict this result, on the other hand when the rocket did not land it fails 9/24 times to predict this result

Conclusions

More data should be obtained in posterior studies to make the model and analysis more robust.

The model has better recall tan precision (if the observations are positive then the model will predict positive).

Overall, all the models perform equally on the test data.

The precision issue could mean that we will underestimate the cost of future SpaceX rocket missions.

Thank you!