IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jaime Cañellas Oliver
07/12/2021

Outline

Executive Summary

Summary of METHODOLOGIES

1. Data Collection
    1. SpaceX API
    2. Web Scraping
2. Data Wrangling
3. EDA with Data Visualization
4. EDA with SQL
5. Interactive Map with Folium
6. Plotly Dash Board
7. Machine Learning models (LogReg, SVM, KNN, Tree)

Summary of all RESULTS

1. Complete DataFrame of Falcon 9
2. Success Rate
    1. vs. Launch Site
    2. vs. Flight Number
    3. Vs. Orbit
3. Launch Site's Location and Surroundings
4. Charts and Plots
5. Accuracies of the models

Introduction

Manufacturing Cost

SpaceX	Other providers
$62 MILLIONS	Up to $165 MILLIONS

¿HOW IS IT POSSIBLE?

REUSING Stage 1

¿What's next?

In this PowerPoint, we will find out the KEY of a success landing

Section 1

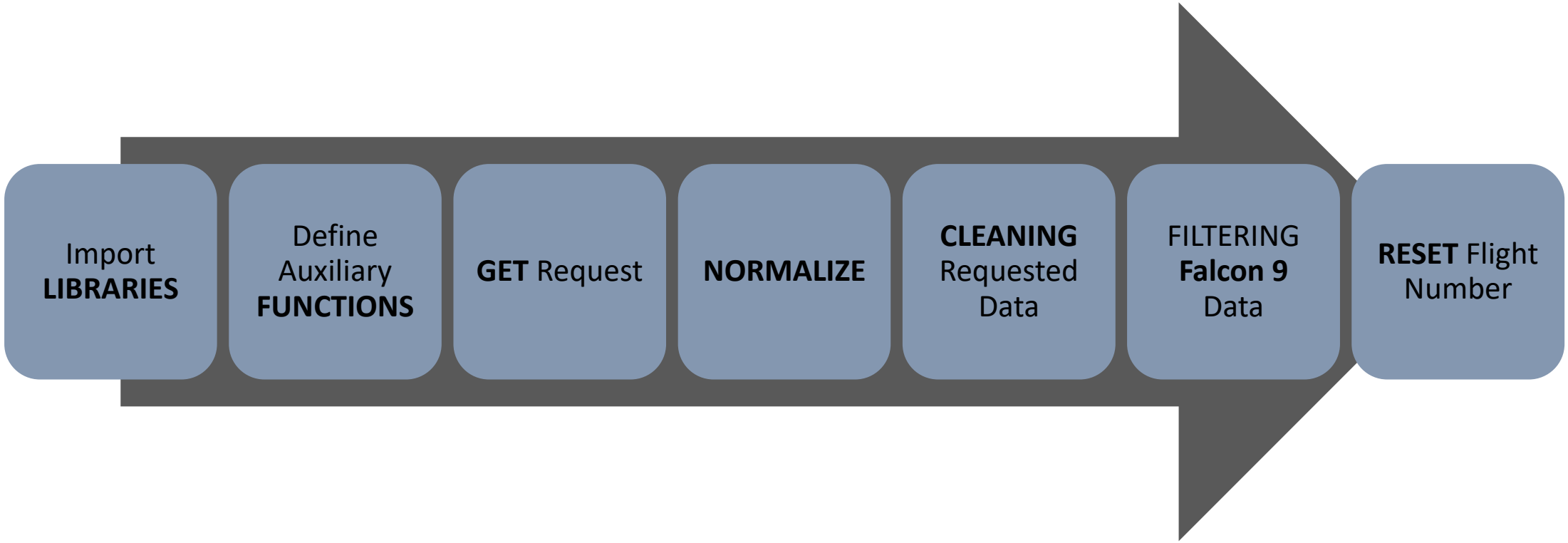
Methodology

Methodology

Executive Summary

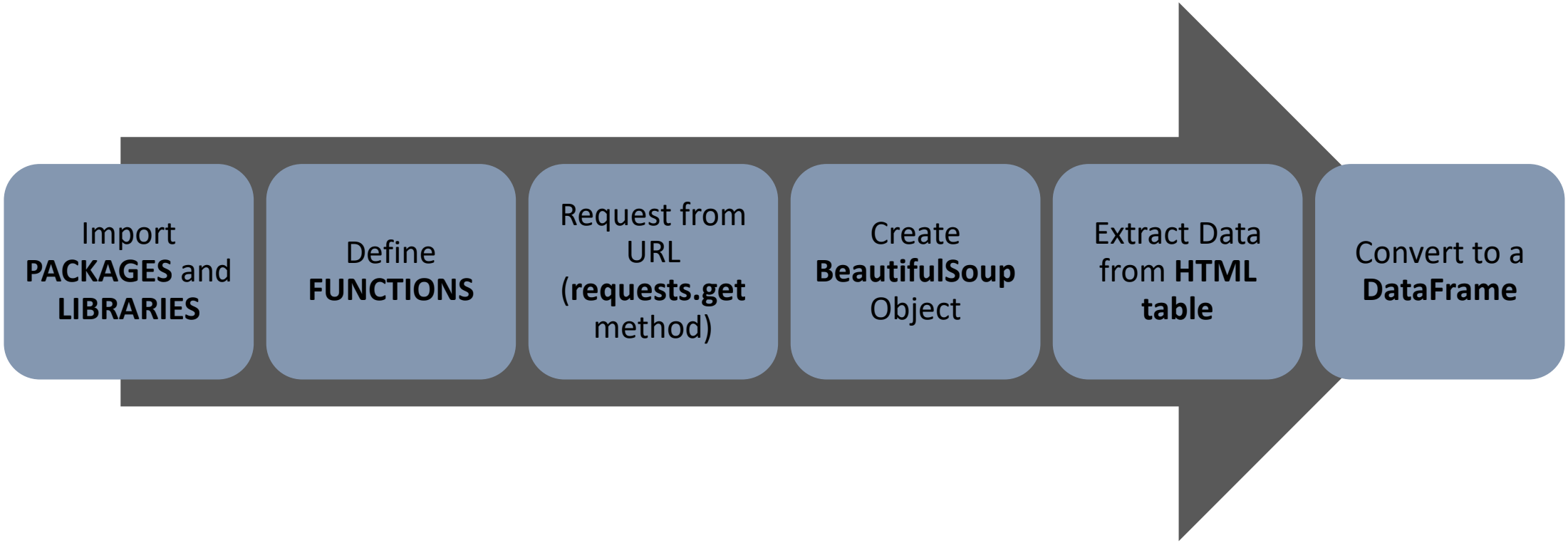
- Data collection methodology (Slides 7-8)
  - SpaceX API
  - Web Scraping
- Perform data wrangling (Slide 9)
  - Filtering, Transforming, Changing and Adding rows/columns ⟶ df['Class'] ⟶ Success Rate
- Perform exploratory data analysis (EDA) using visualization and SQL (Slides 10-12)
- Perform interactive visual analytics using Folium and Plotly Dash (Slides 13 - 14)
- Perform predictive analysis using classification models (Slide 15)
  - Split, Fit, Train, Test, Best parameters, Best score, Score method.

Data Collection- SpaceX API



https://github.com/JCO2502/capsproj/blob/8e812247b929ff40a85085fd4b44f0cc3a19065c/1_C10_CollectDataAPI.ipynb

Data Collection- Web Scraping



https://github.com/JCO2502/capsproj/blob/8e812247b929ff40a85085fd4b44f0cc3a19065c/2_C10_CollectDataWebScrap.ipynb

Data Wrangling

IMPORTANT STEPS

Missing values

Column TYPE

No. LUNCH per SITE

No. LUNCH per ORBT

Landing OUTCOMES (0 True ASDS/1 None None/2 True RTLS/3 False ASDS/4 True Ocean/5 None ASDS/6 False Ocean/7 False RTLS)

AS

TRUE =1 FALSE=0

Add 'Class' Column

SUCCES RATE

https://github.com/JCO2502/capsproj/blob/6cf62df3ea921c588febdaeb216c8c0ac047e61b/3_C10_Data_Wrangling.ipynb

EDA with Data Visualization I

Flight Number vs. Payload Mass	⬆ Flight Number = + Payload Mass & +Success Rate
Flight Number vs. Launch Site	**TOP** = CCAFS SLC 40
Payload vs Launch Site	• VAFB SLC 4E up to **10,000kg** • Most HEAVY launches in **CCAFS SLC 40** (FAIL in lower PayloadMass)
Success Rate vs Orbit	✚ **HIGHER SUCCESS:** ES-L1, GEO, HEO & SSO **(100%)** ━ **LOWER SUCCESS:** GTO & ISS

https://github.com/JCO2502/capsproj/blob/dd6c3763933fab1f7a4b311529f04db3d92be849/5_C10_EDA_DataVis.ipynb

EDA with Data Visualization II

Flight Number vs. Orbit	⬆ Flight Number = ⬆ VLEO Orbit
Payload Mass vs. Orbit	**HIGHEST** Payload Mass = VLEO Orbit
Success Rate vs Year	⬆ 2013-2017 ⬇ 2017-2018 ⬆ 2018-2019 ⬇ 2019-2020

https://github.com/JCO2502/capsproj/blob/dd6c3763933fab1f7a4b311529f04db3d92be849/5_C10_EDA_DataVis.ipynb

EDA with SQL

SQL Queries

- DISTINCT Statement
- WHERE + LIKE '%NAME%' – LIMIT
- SUM() + WHERE
- AVG() + WHERE
- MIN() + WHERE
- COUNT() + GROUPBY
- DISTINCT + WHERE [SUBQUERY MAX()]
- substr(Date, 4, 2) as month + substr(Date,7,4)='2015'
- WHERE + BETWEEN AND + LIKE '%NAME%' – ORDER BY Desc

https://github.com/JCO2502/capsproj/blob/dd6c3763933fab1f7a4b311529f04db3d92be849/4_C10_EDA_SQL_2.ipynb

Build an Interactive Map with Folium

CIRCLE

MARK Launch Site

CLUSTER

Showing in a simple manner the SUCCESS/FAILURE of each launch in each Launch Site

LINES

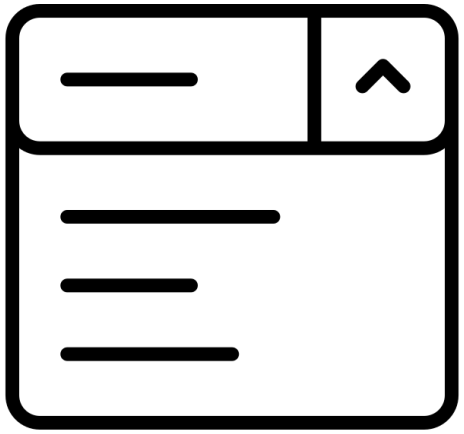
Illustrating the distances among key places surrounding the Launch Site

DISTANCE KM Expressing a numerical fact

https://github.com/JCO2502/capsproj/blob/dd6c3763933fab1f7a4b311529f04db3d92be849/6_C10_InteractiveVisualAnalyticsFolium.ipynb

Build a Dashboard with Plotly Dash

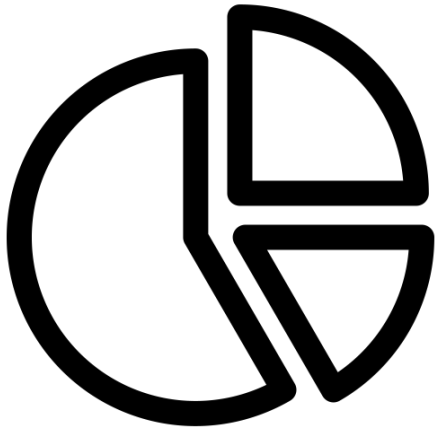
DROPDOWN



Filtering Launch Site

[All sites, CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E]

PIE CHART



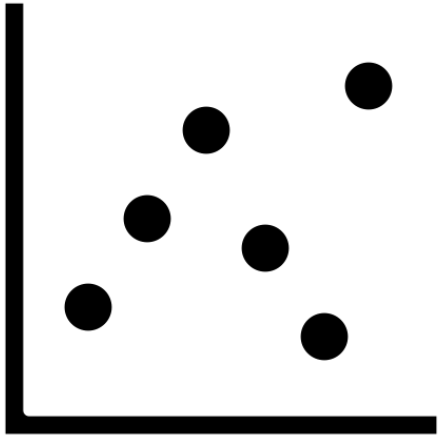
Success Rate vs Launch Site

SLIDER

Filtering Payload Mass

[0-10,000kg]

SCATTER PLOT



Payload Mass vs Succes Rate (ALL sites)

https://github.com/JCO2502/capsproj/blob/dd6c3763933fab1f7a4b311529f04db3d92be849/spacex_dash_app.py

Predictive Analysis (Classification)

Define Y → Normalize and Define X → Split the Data: TRAIN vs TEST set → Run GridSearch CV objects with different classification techniques (Logarithmic Regression, SVM, KNN, Tree) → Calculate their accurancies → Show each Confusion Matrix → Compare Accurancies

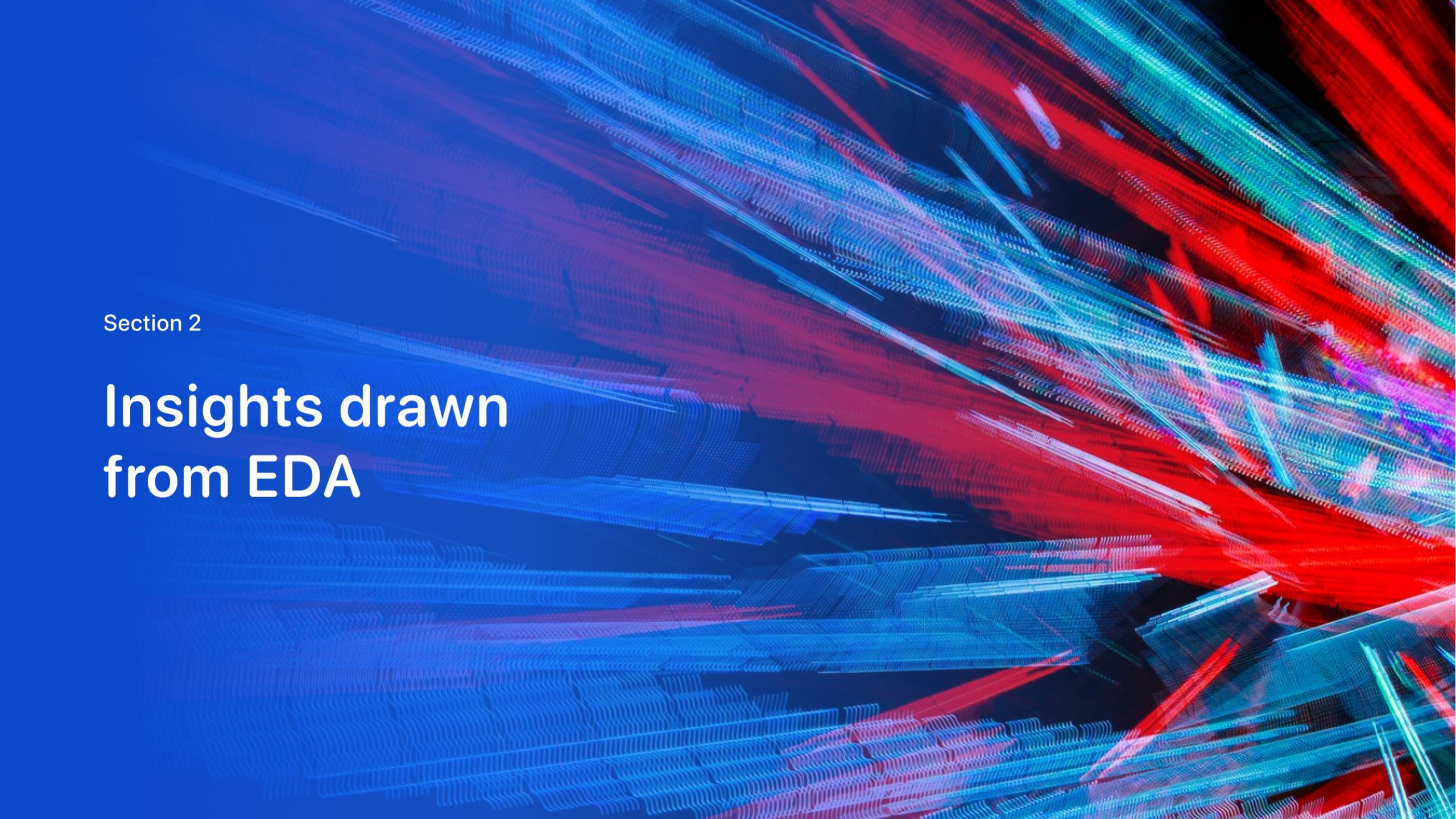
https://github.com/JCO2502/capsproj/blob/dd6c3763933fab1f7a4b311529f04db3d92be849/8_C10_SpaceX_Machine_Learning_Prediction_Part_5.ipynb

Results

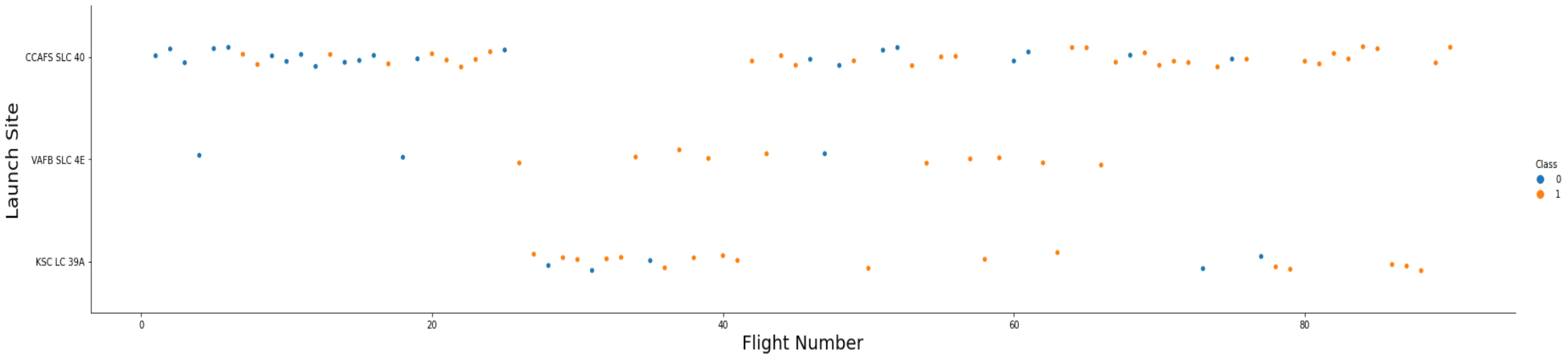
Exploratory data analysis results	Interactive analytics	Predictive analysis results
• Different Launch Sites: ✓ CCAFS LC-40 ✓ KSC LC-39A ✓ VAFB SLC 4E ✓ CCAFS SLC-40 • Most No. Launches: ✓ CCAFS SLC-40 • 100% Success Orbits: ✓ ES-L1 ✓ GEO ✓ HEO ✓ SSO • Success Rate: 0,6667 • Total Payload Mass: 45,596KG • First Successful Landing: 01-05-17	• Launch Site Latitude and Longitude *(Appendix 1)* • Launch Site Plotting • Marker color ⟶ Cluster • Closest Distances: ✓ Coast: 0,87KM ✓ Highway: 0,58KM ✓ City: 23,19KM ✓ Railway: 1,13KM	• 4 models: ✓ Logarithmic Regression ✓ SVM ✓ KNN Neighbors ✓ Tree • Best parameters *(Appendix 2)* • Accurancy *(Appendix 2)* • Score *(Appendix 2)*

Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

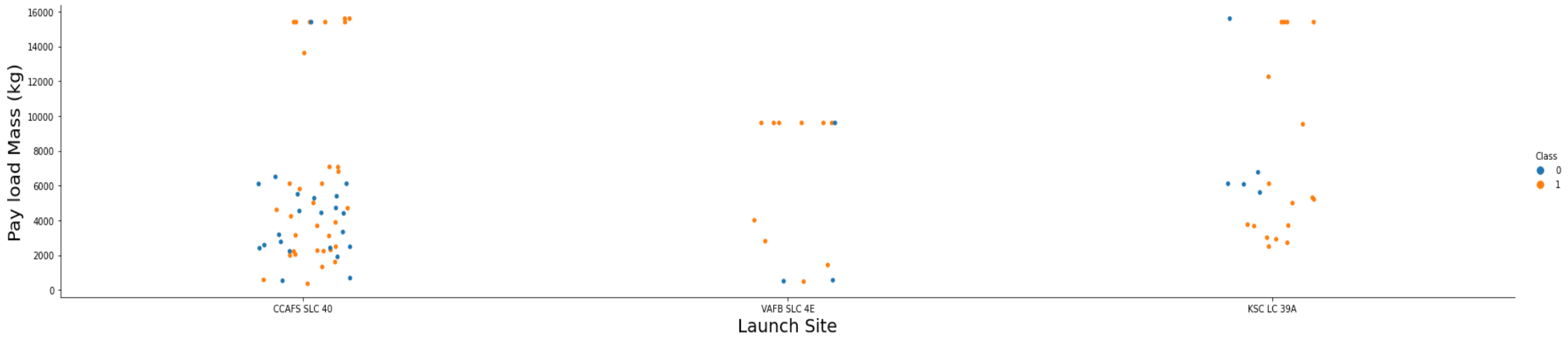


Most launches at CCAFS SLC 40 (the more n° launches, the more successful)

At VAFB SLC 4E, few n° of launches. No one for some launches.

Late starting at KSC LC 39A, however, good successful rate and still working there.

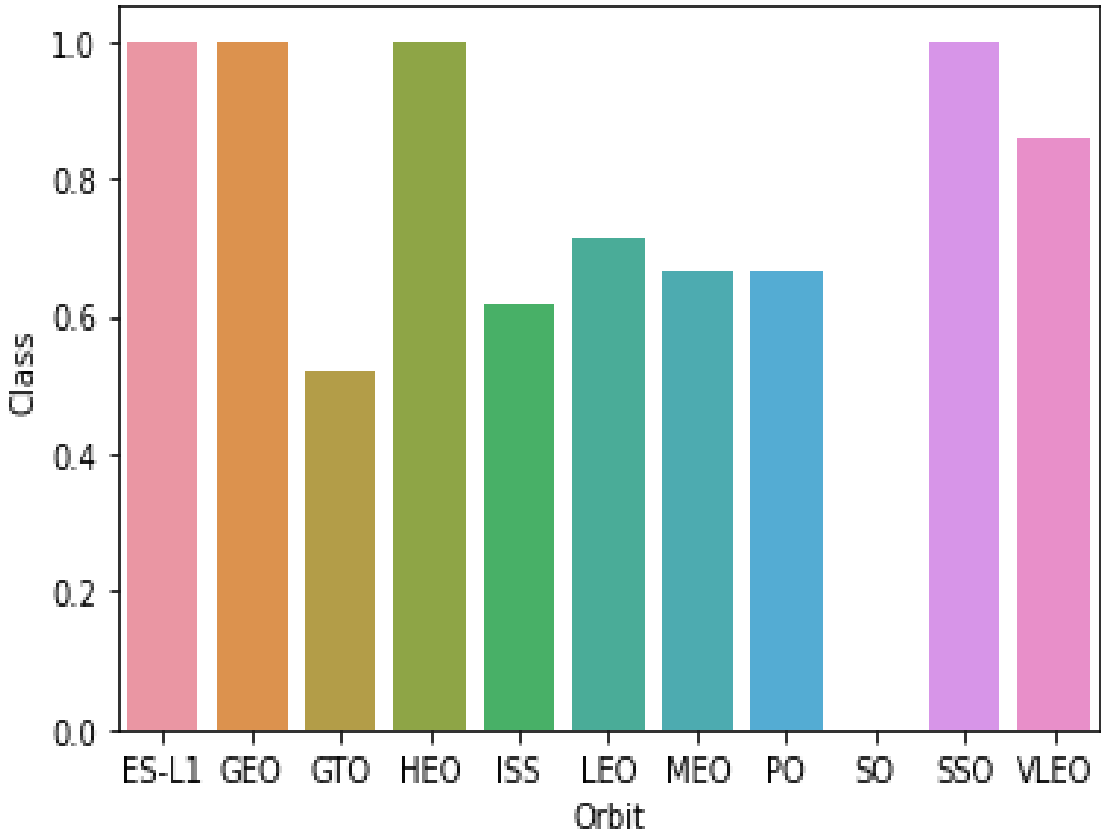
Payload Mass vs. Launch Site



VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

CCAFS SLC 40 more launches than KSC LC 39A although more failure. Both of them reaching heavies payload mass.

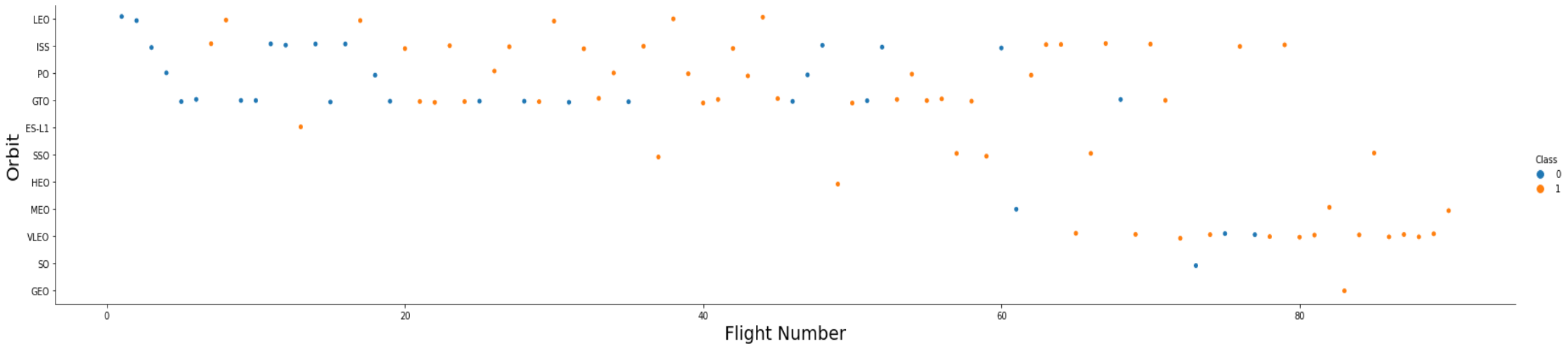
Success Rate vs. Orbit Type



100% Success: ES-L1 / GEO / HEO / SSO

Most % of failure: SO / GTO / ISS

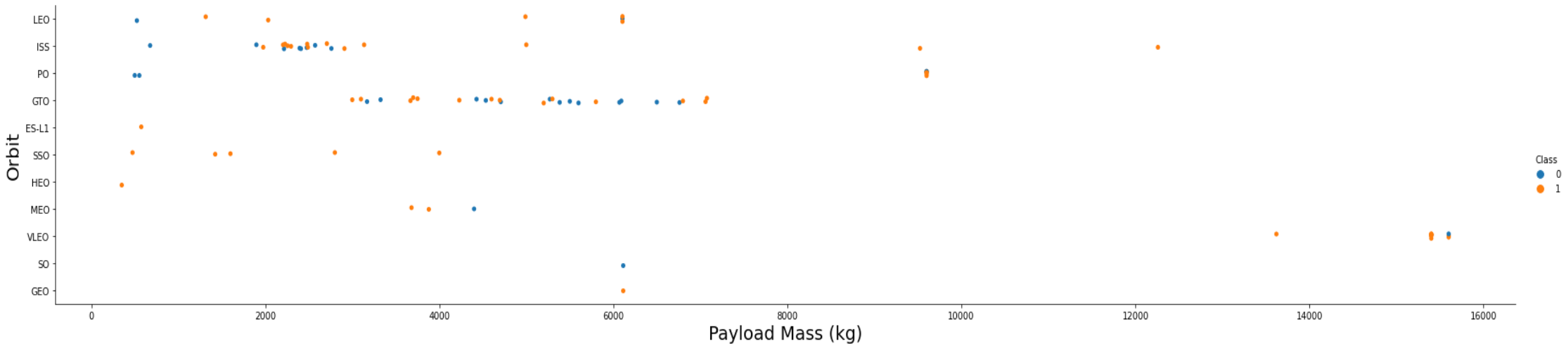
Flight Number vs. Orbit Type



First launches, extremely high rate of failure

Latest launches, most of them at VLEO with a great success rate (13 out of 15)

Payload Mass vs. Orbit Type

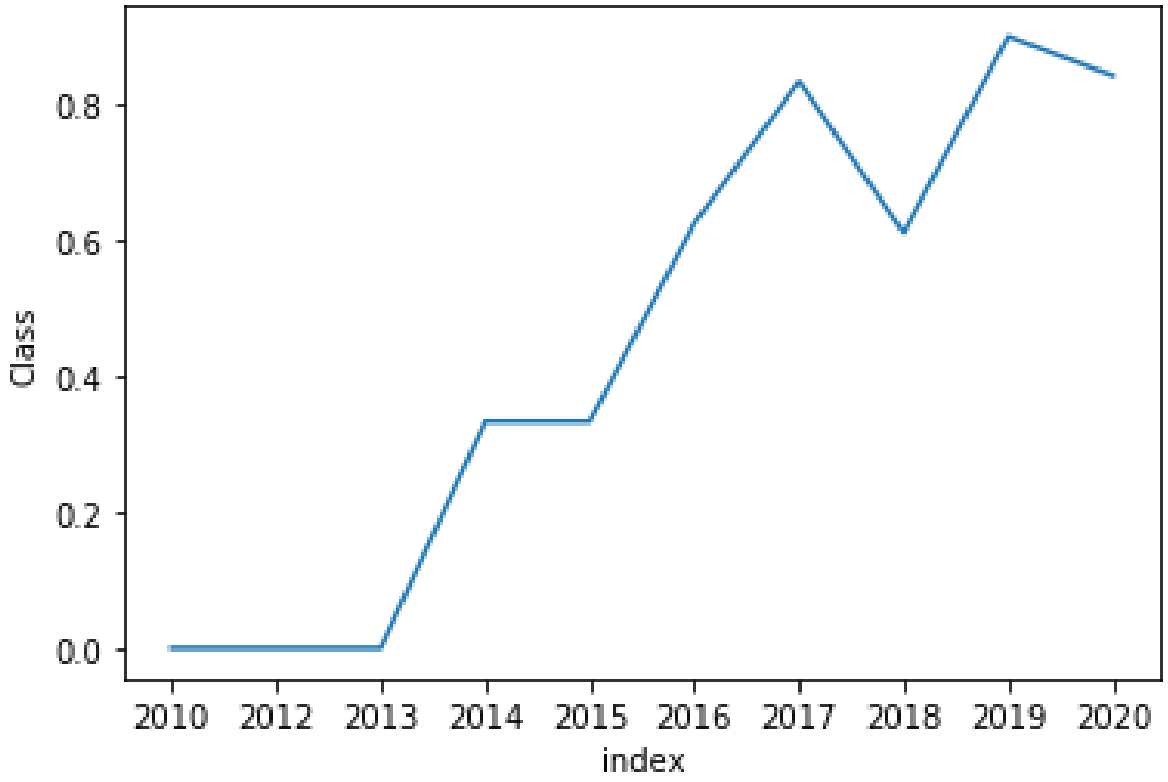


GTO: low successful rate, low payload mass

VLEO: highest payload mass

Payload mass is lower than 8,000kg in most of the orbits

Launch Success Yearly Trend



- First successful launches since 2013
- Highest successful rate in 2019
- Deep decreasing in 2018
- Stagnated growth during 2014

All Launch Site Names

DISTINCT METHOD

%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;

CCAFS LC-40

KSC LC-39A

VAFB SLC-4E

CCAFS SLC-40

Launch Site Names Begin with 'CCA'

LIKE '%%' METHOD + LIMIT

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

CCAFS LC-40 CCAFS LC-40

CCAFS LC-40 CCAFS LC-40

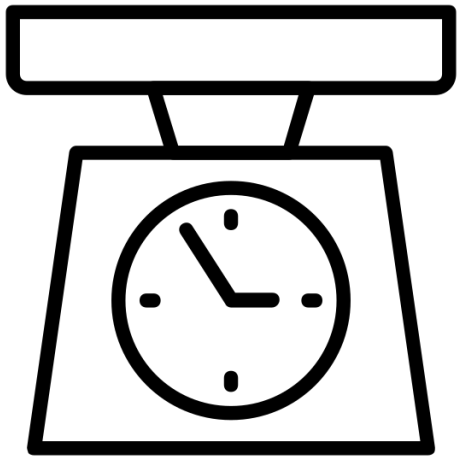
CCAFS LC-40

Total Payload Mass

SUM METHOD + WHERE CLAUSE

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

45,596 KG

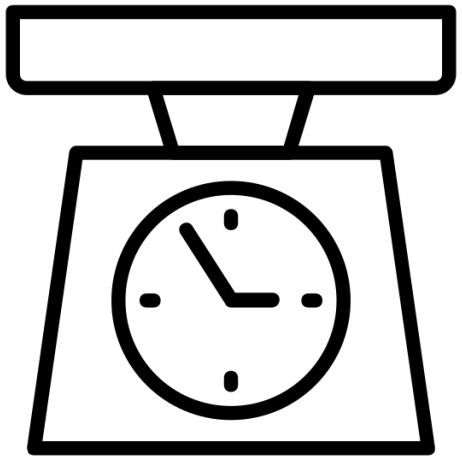


Average Payload Mass by F9 v1.1

AVG METHOD + WHERE CLAUSE

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

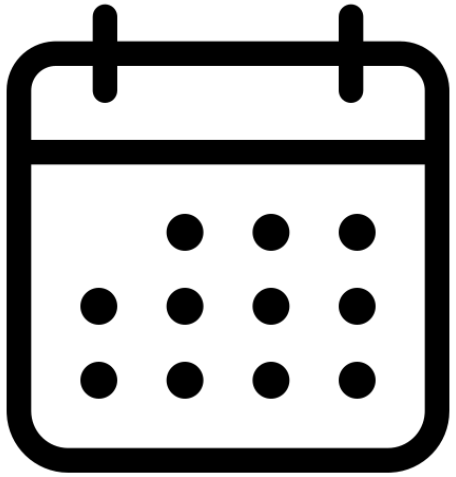
2928.4 KG



First Successful Ground Landing Date

MIN METHOD + WHERE CLAUSE

%sql SELECT MIN(Date) FROM SPACEXTBL WHERE [Landing _Outcome] ='Success (ground pad)'



01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

DISTINCT METHOD + WHERE CLAUSE with AND & BETWEEN-AND

%sql SELECT DISTINCT Booster_Version, PAYLOAD_MASS__KG_, [Landing _Outcome] FROM SPACEXTBL WHERE [Landing _Outcome] ='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

Booster Version	Payload Mass	Landing Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

COUNT METHOD + GROUPBY

%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTBL GROUP BY Mission_Outcome

Mission Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

DISTINCT METHOD + WHERE SUBCLAUSE

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE
PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
LIMIT 5;
```

F9 B5 B1048.4

F9 B5 B1056.4

F9 B5 B1049.4

F9 B5 B1048.5

F9 B5 B1048.4

2015 Launch Records

substr(Date, 4, 2) AS MONTH + SUBSTR(Date,7,4)='2015'

%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)'

MONTH	Landing Outcome	Booster Version	LaunchSite
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

WHERE CLAUSE + BETWEEN-AND +LIKE + ORDER BY Desc

%sql SELECT * FROM SPACEXTBL WHERE Date BETWEEN '04-06-2010' AND '20-03-2017' AND [Landing _Outcome] LIKE '%Success%' ORDER BY Date Desc;

Date	Time (UTC)	Booster Version	Launch Site	Payload	Payload Mass (KG)	Orbit	Customer	Mission Outcome	Landing Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
18-10-2020	12:25:57	F9 B5 B1051.6	KSC LC-39A	Starlink 13 v1.0, Starlink 14 v1.0	15600	LEO	SpaceX	Success	Success
18-08-2020	14:31:00	F9 B5 B1049.6	CCAFS SLC-40	Starlink 10 v1.0, SkySat-19, -20, -21, SAOCOM 1B	15440	LEO	SpaceX, Planet Labs, PlanetIQ	Success	Success
18-07-2016	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
18-04-2018	22:51:00	F9 B4 B1045.1	CCAFS SLC-40	Transiting Exoplanet Survey Satellite (TESS)	362	HEO	NASA (LSP)	Success	Success (drone ship)

Section 4

Launch Sites Proximities Analysis

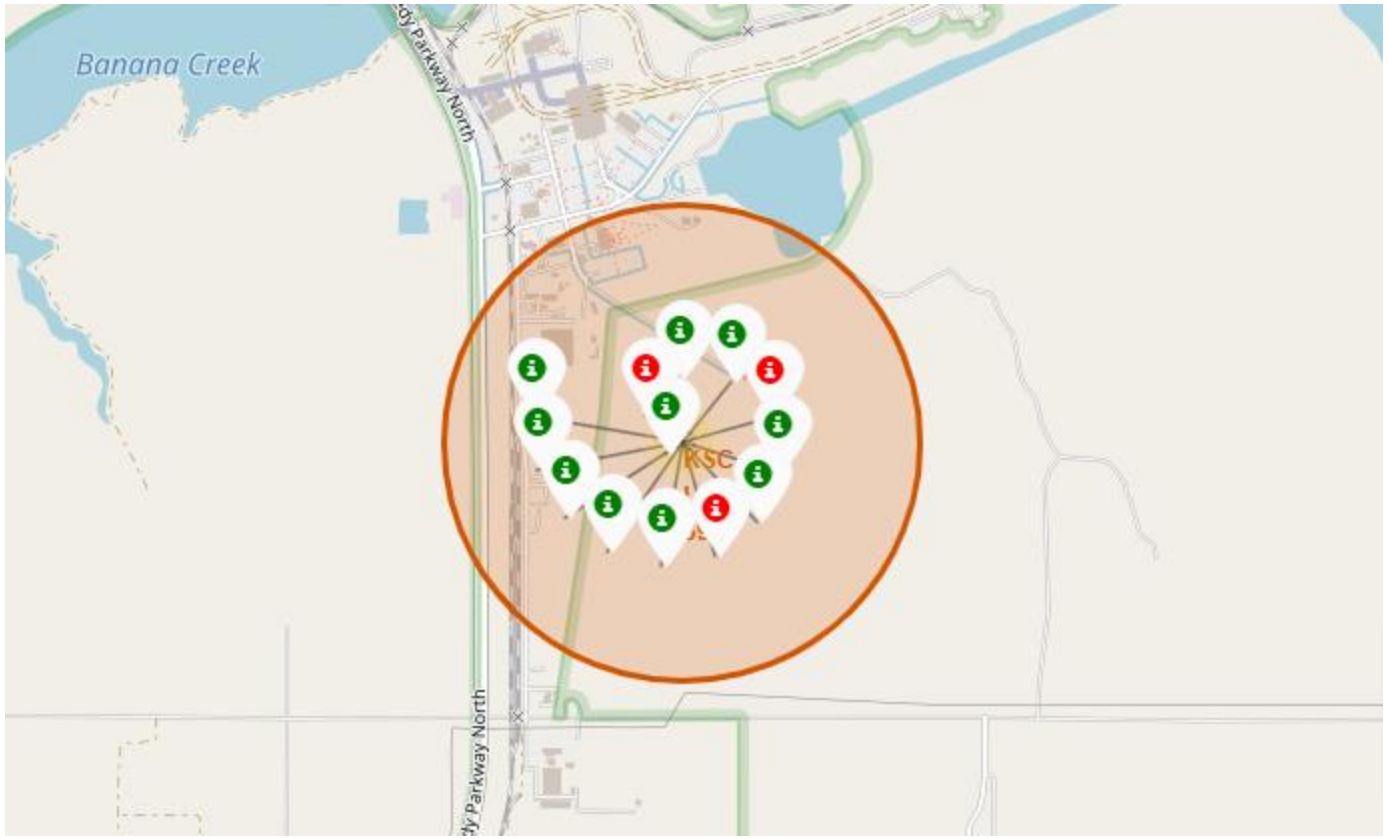
Launch Site Locations

All of them on the coast

3 out of 4 in the east one

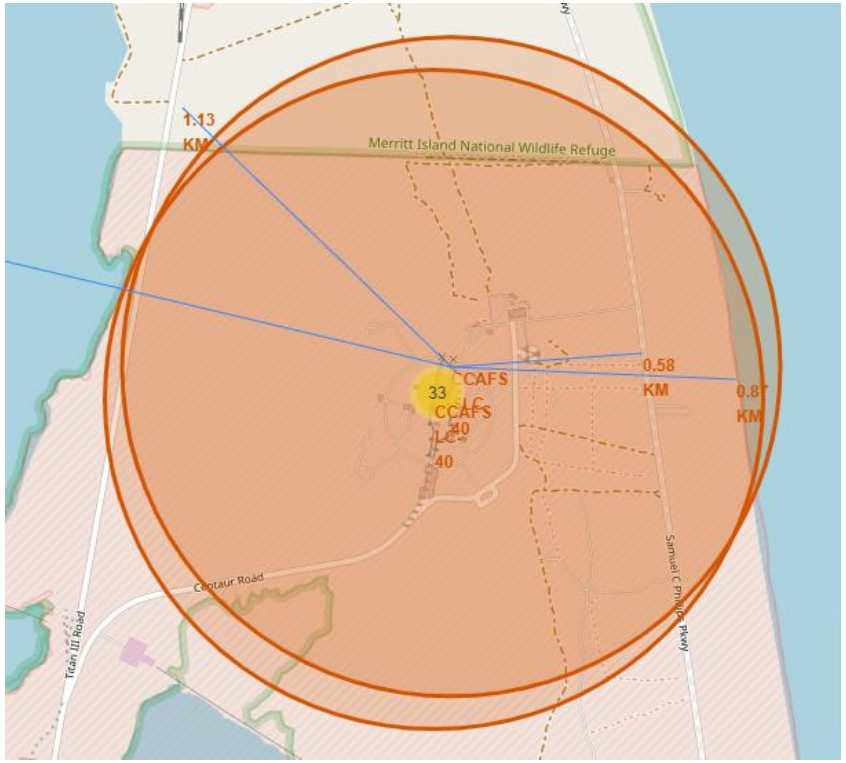
Those on the east cost are pretty close

Launch Site Cluster



Easy and accessible way to group and show the data (each launch site with its success/failure launch)

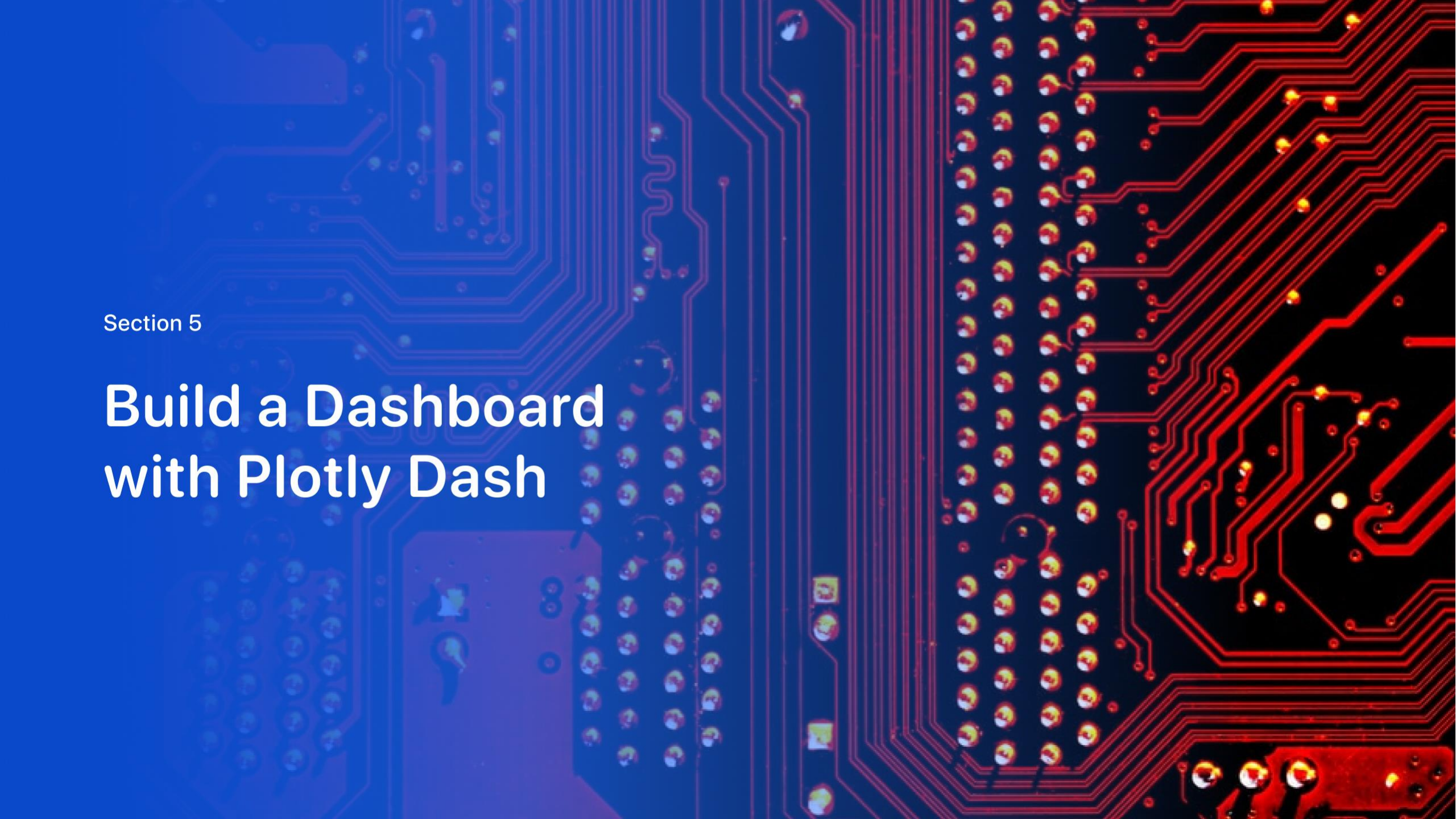
Exploring the surroundings of a Launch Site



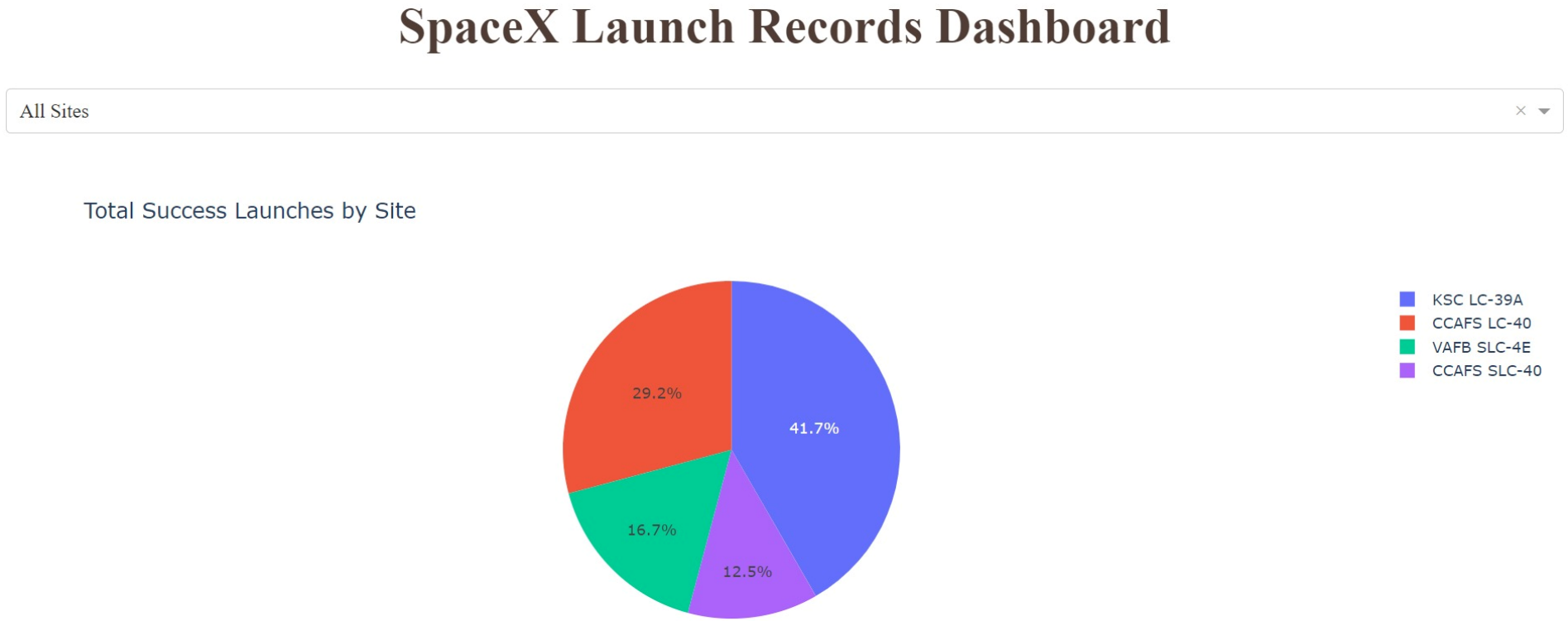
Adding **LINES** and **DISTANCE** to the closest railway, highway, coast and city

Section 5

Build a Dashboard with Plotly Dash

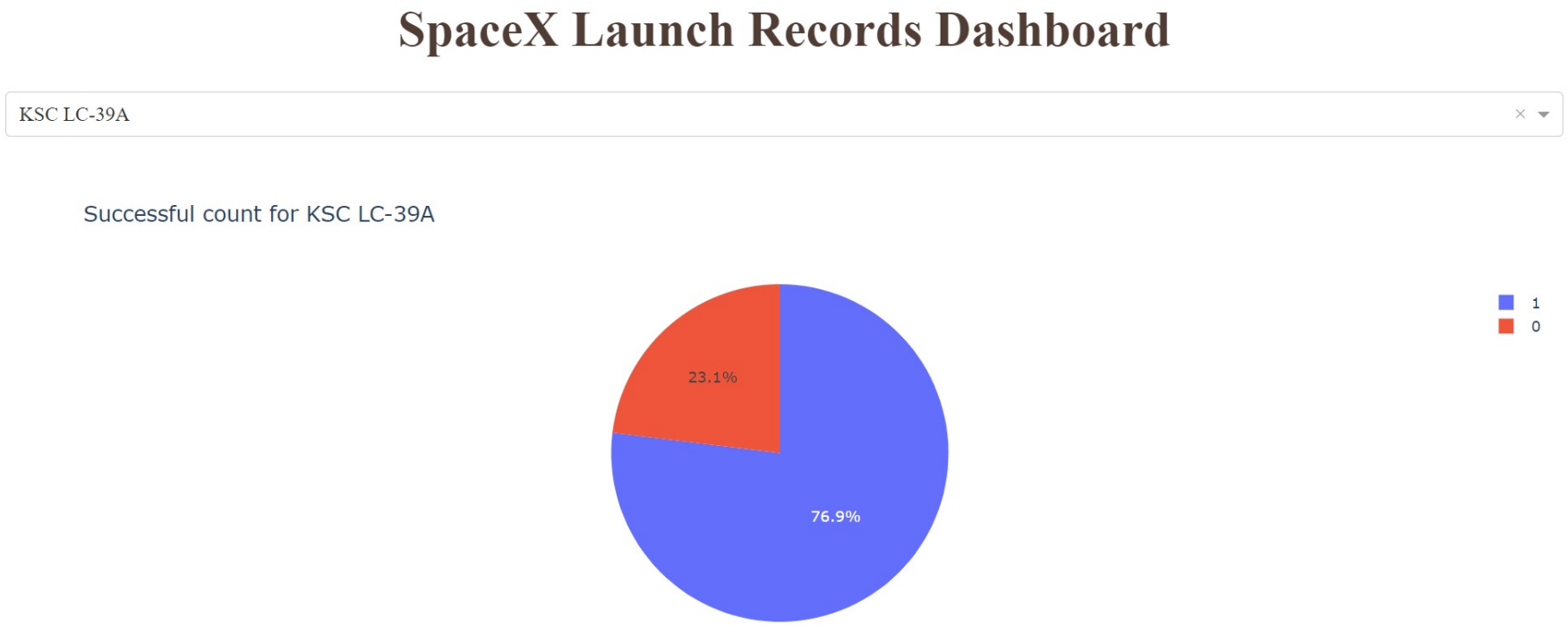


Success Rate vs. Launch Site



- KSC LC-39A: The most successful Launch Site
- CCAFS LC-40's Success Rate = VAFB SLC-4E + CCAFS SLC-40

KSC LC-39A: The way to success



- KSC LC-39A: Although being the most successful one, nearly 25% of the launches fail!
- All of them reach a success rate higher than 55%

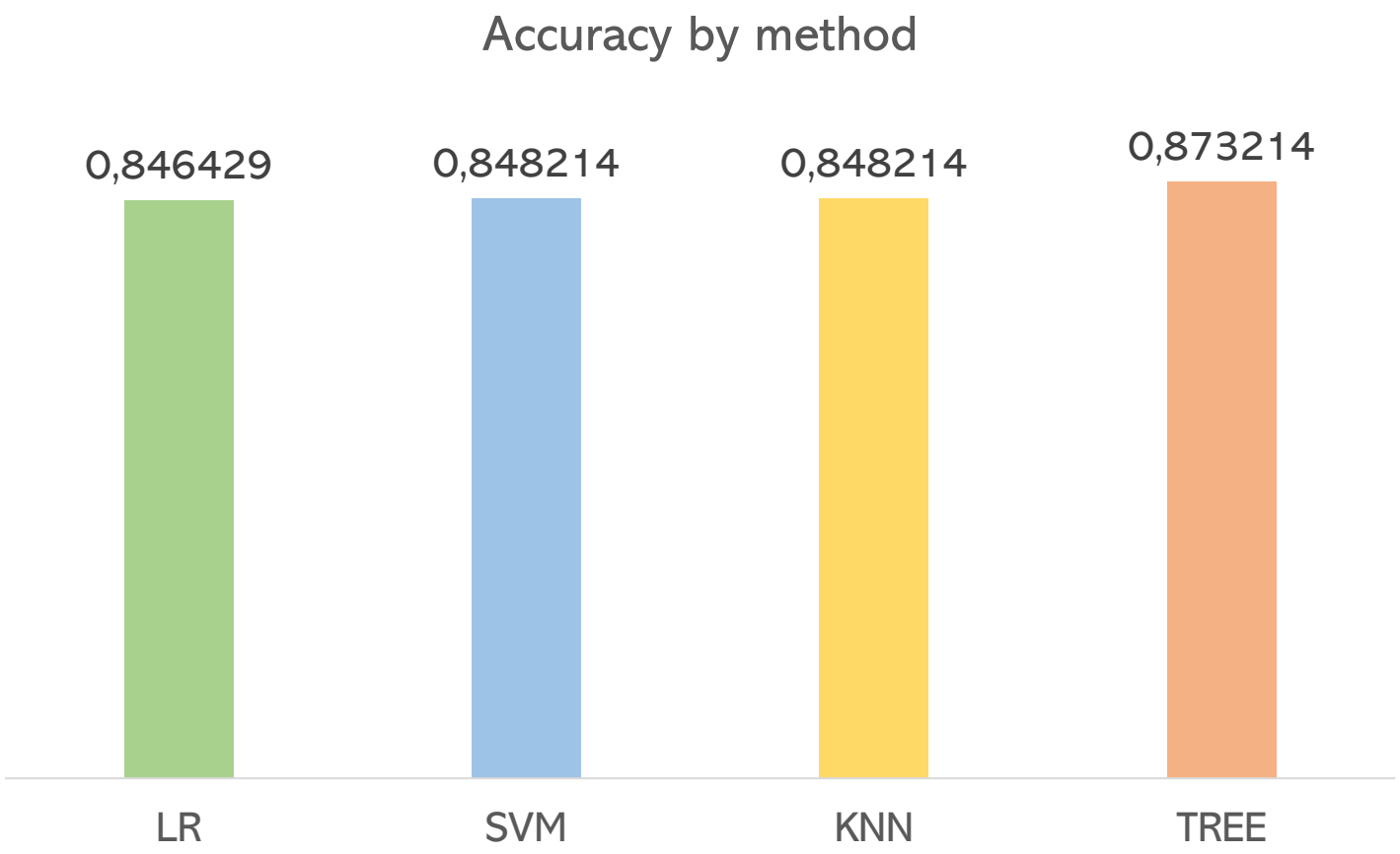
Success Rate per Payload Mass for each Booster Version (2,500-7,500 kg)

- FT: More launches but (3/7 failed)
- B4 & B5: 100% Success Rate in this range (2,500-7,500kg)

Section 6

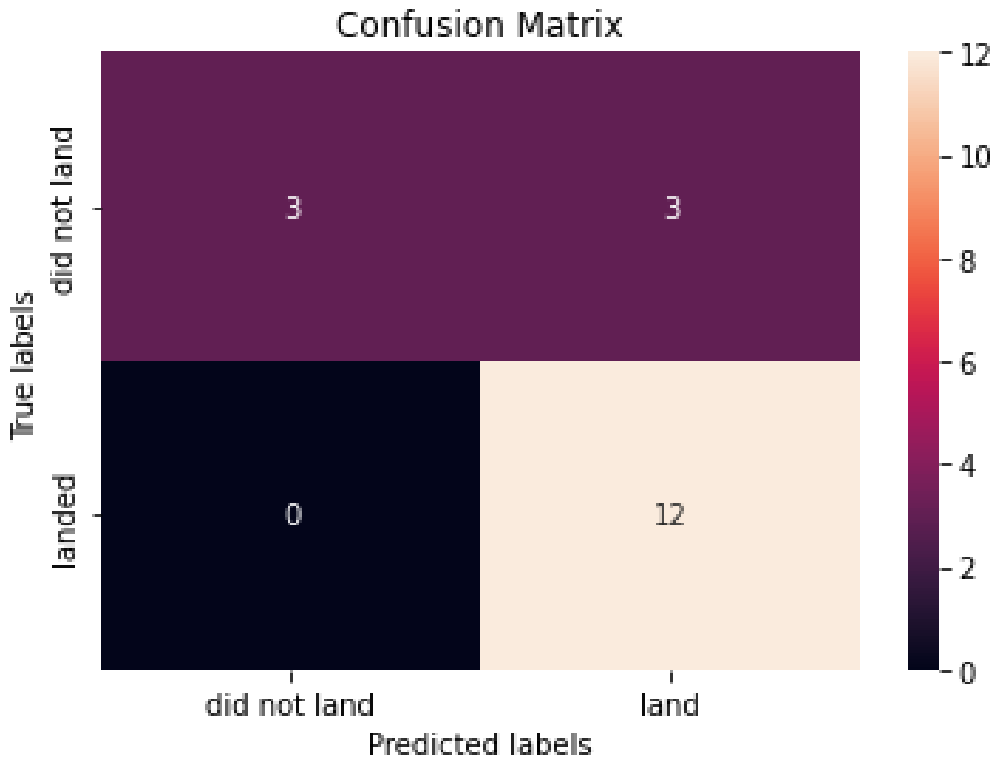
Predictive Analysis (Classification)

Classification Accuracy



Method	Accuracy
LR	0.846429
SVM	0.848214
KNN	0.848214
TREE	0.873214

Confusion Matrix – Tree MODEL



The model predicted:

3 launches as not landed and 15 as landed.

Whereas true labels shows:

6 of them unsuccessfully landed and 12 did it with success.

PRECISION: 0,5

RECALL: 1

ACURRANCY: 0,8334

Conclusions

- **Payload Mass**, **Orbit** and **Launch Site** affect directly to the probability of success
- **KSC LC-39A** is the most successful Launch Site (79,6%)
- All of Launch Sites reach a SUCCESS RATE **higher than 55%**
- **F9 B5 Booster Version** is the one which endure a higher Payload Mass
- **Decision Tree Model** is the most accurate

Keep working hard because as it is shown in the **SUCCESS RATE each year** means a step further in discovering the outer world

Appendix 1

LAUNCH SITE LATITUDE & LONGITUDE

Launch Site	Latitude	Longitude
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610746

Appendix 2

Best Parameters, Score & Accuracy

MODEL	Best Parameters	Score	Accuracy
LOG REG	{'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}	0.83334	0.84642
SVM	{'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}	0.83334	0.84821
KNN	{'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'random'}	0.83334	0.84821
TREE	{'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}	0.83334	0.87321

Thank you!