IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Soumya Paul
02/07/2022

Outline

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

Executive Summary

In this project will apply our data science skills for a private space launch company to determine if it can reuse the first stage of its launch.

- Summary of methodologies
  - Launch Data Collection via API & Web Scrapping
  - Data wrangling / Cleaning
  - Exploratory Data Analysis & Data Visualization
  - Predictive modeling
- Summary of all results
  - Exploratory Data Analysis Results
  - Finding from Interactive Dashboard
  - Predictive Modeling

Introduction

- Project background and context

  SpaceX is holding a remarkable position in the commercial space industry as a result of the low cost of launching a rocket. According to the company, the standard payment plan for a Falcon 9 launch is $62 million; while the cost of other providers exceeds $165 million. Much of the savings is because SpaceX can reuse the rocket's first stage.

- Problems you want to find answers

  ✓ What determines whether a rocket will land successfully?

  ✓ Which variables have the greatest influence on the landing success rate?

  ✓ Can we estimate the cost of a launch by predicting whether the first stage will land?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

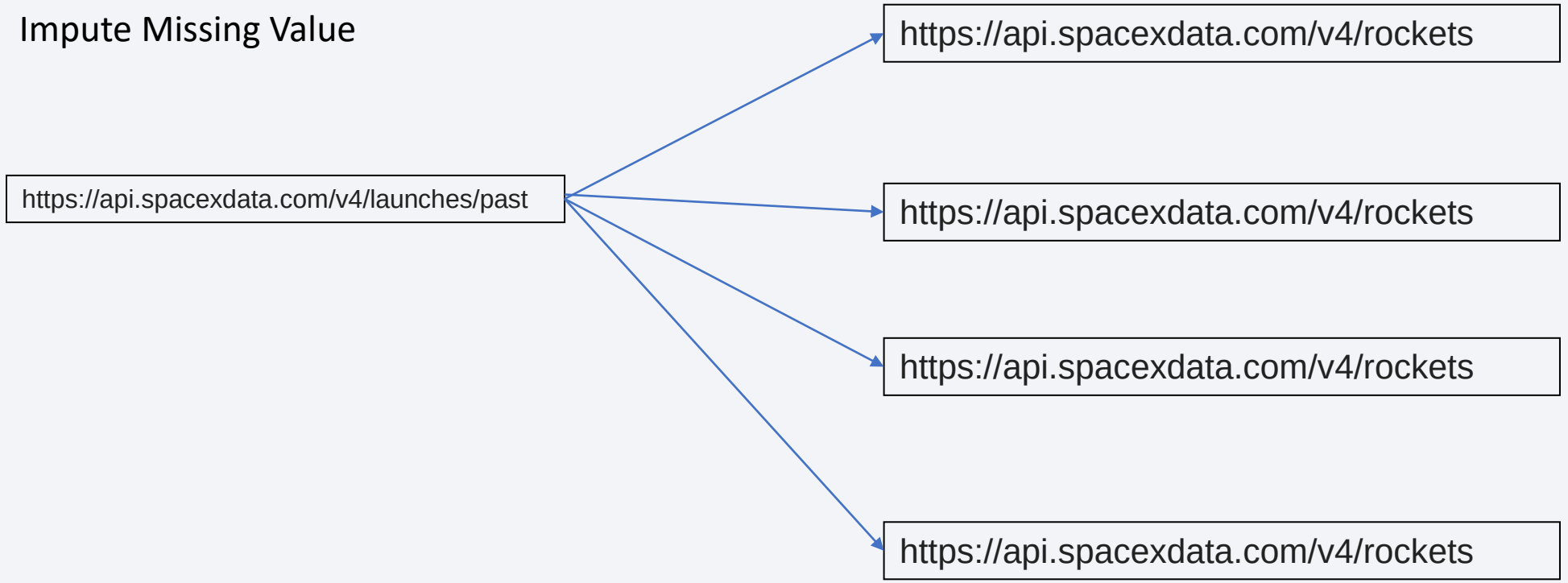
  Data source: SpaceX REST API – HTTP Requests using Python Requests Library.

  Data source: Wikipedia Article – Web scraping with Beautiful Soup Library.

- Perform data wrangling
  - Dealing with missing values, one hot encoding, type cast, transform and cleaning 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 standardization and data set splitting. Hyperparameter tuning using Scikit-Learn's GridSearchCV. Evaluation of Logistic Regression, Decision Tree, KNN and SVM classifiers algorithms.

Data Collection using API

Collect Launch data from APIs using GET Request
Filter the data
Impute Missing Value



Notebook

Data Collection – SpaceX API

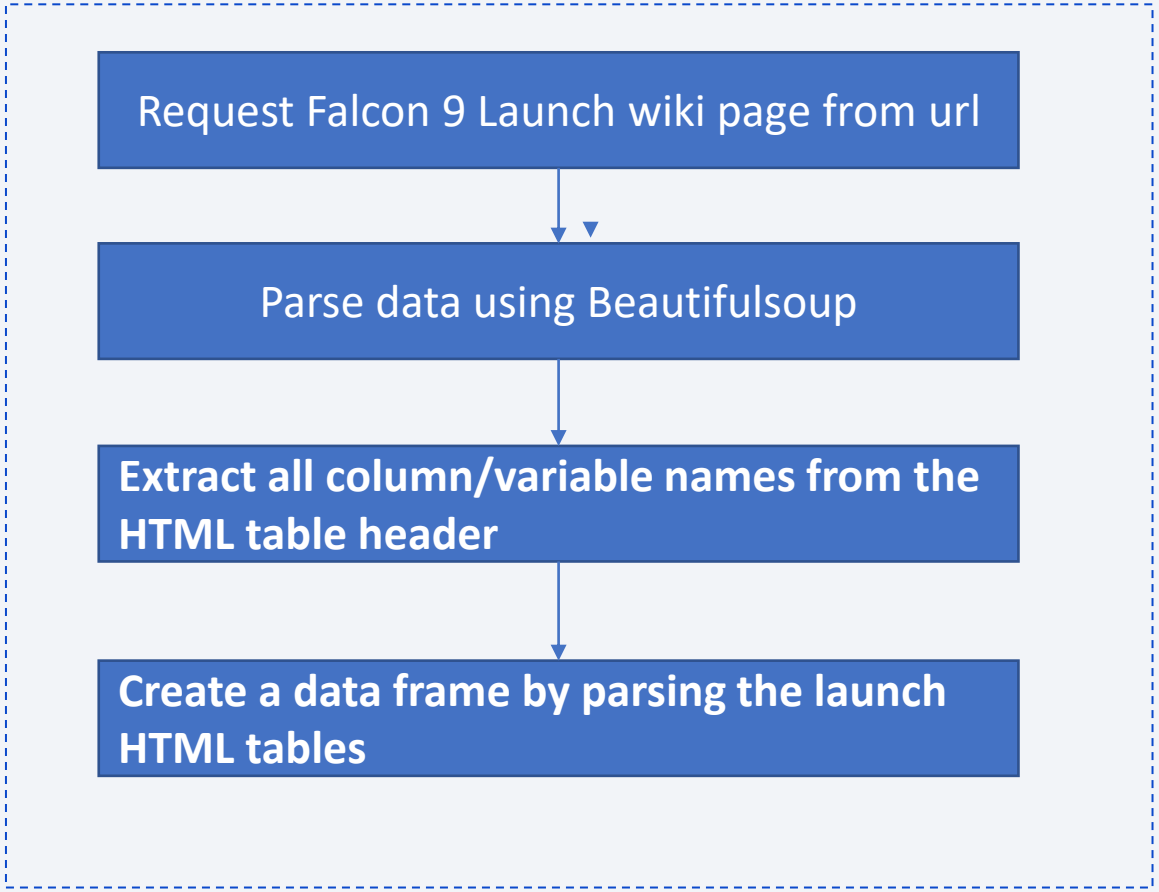
- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

Place your flowchart of SpaceX API calls here

Data Collection - Scraping

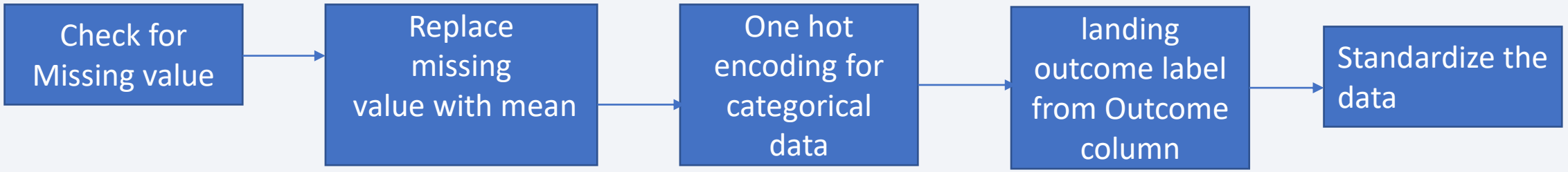
- Notebook



Data Wrangling

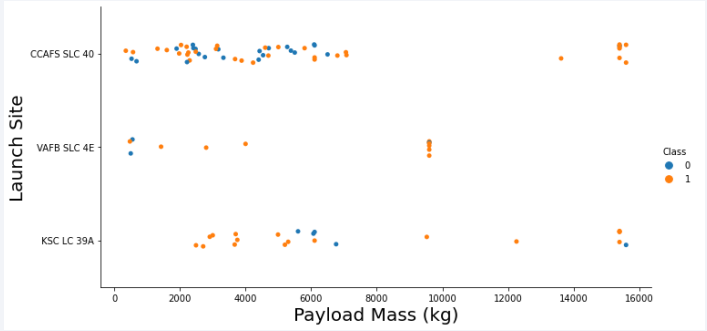
- Missing values replaced with mean
- One hot encoding
- Create a landing outcome label from Outcome column
- Standardize the data
-

Notebook

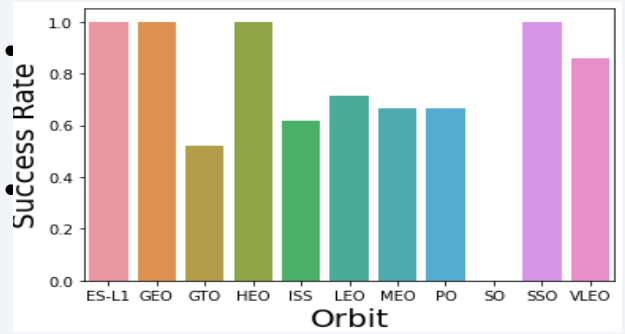


EDA with Data Visualization

- Flight Number and Launch Site
- Flight Number vs Payload Mass
- Payload Mass vs Launch Site
- Orbit vs Flight Number
- Payload Mass vs Orbit Type

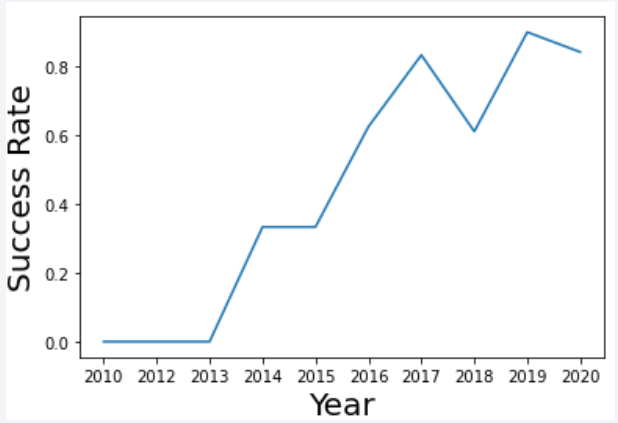


A scatter plot uses dots to represent values for two different numeric variables. Scatter plots are used to observe relationships between variables.



A bar chart presents categorical data with rectangular bars with heights or lengths proportional to the values

that they represent



- A line chart is a visual comparison of how two variables— shown on the

x- and y- axes —are related or vary with each other. It shows related information

by drawing a continuous line between all the points on a grid.

Notebook

EDA with SQL

- **Using SQL queries to answer questions about the data set**

✓ 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

- Notebook

Build an Interactive Map with Folium

- Markers + Circles with Popup that identify launch site by name ○ Shows where launch sites are located
- MarkerCluster of Markers color coordinated by launch status (success/failure) ○ Shows where launches are concentrated
- MousePosition object used to get the coordinates of several points of interests
- PolyLine between launch site to selected coastline point ○ Shows distance between launch site and nearest coastline

- Notebook

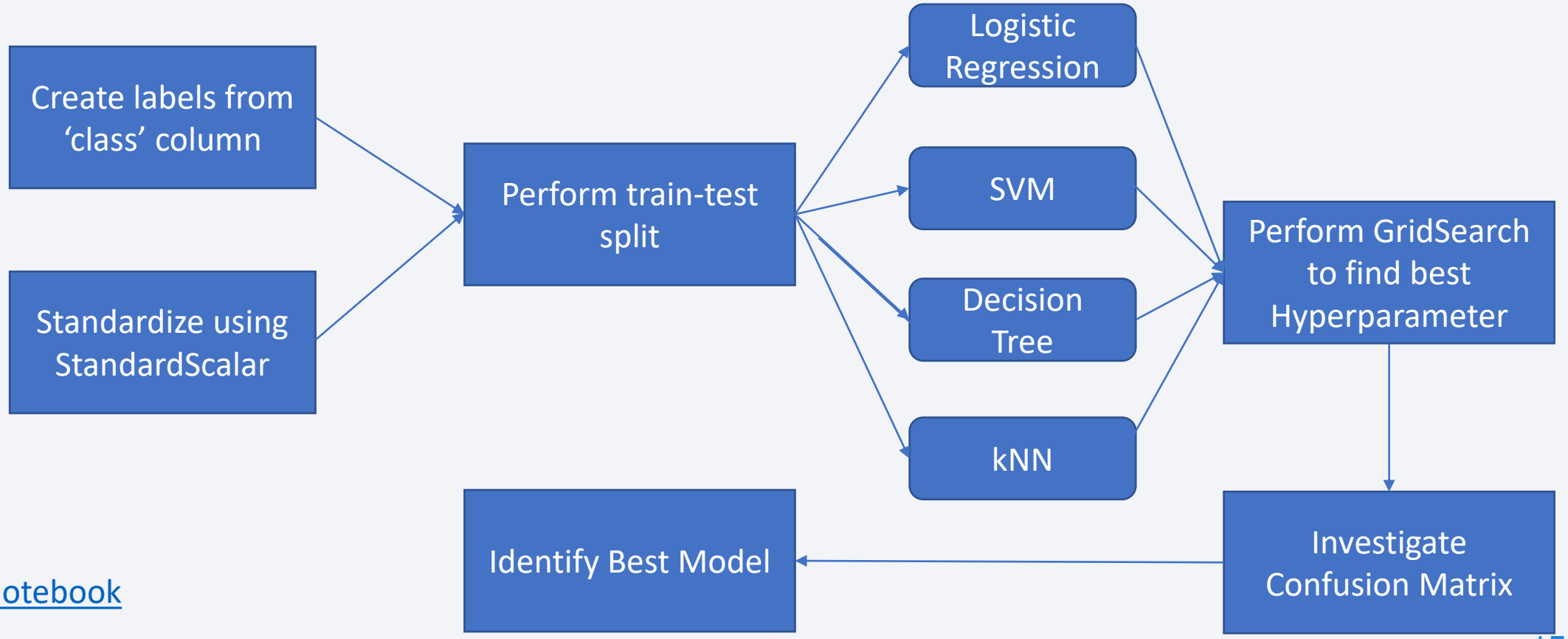
Build a Dashboard with Plotly Dash

The dashboard was build using the Dash framework

- Launch Site Dropdown input component
  - Allows filtering landings by launch site
- Payload RangeSlider input component
  - Allows filtering landings by payload mass
- Landing success rate pie chart
  - Shows the percentage of successful landings by launch site
- Landing outcome vs payload mass scatter plot
  - Shows relationship between payload mass and landing outcome

Notebook

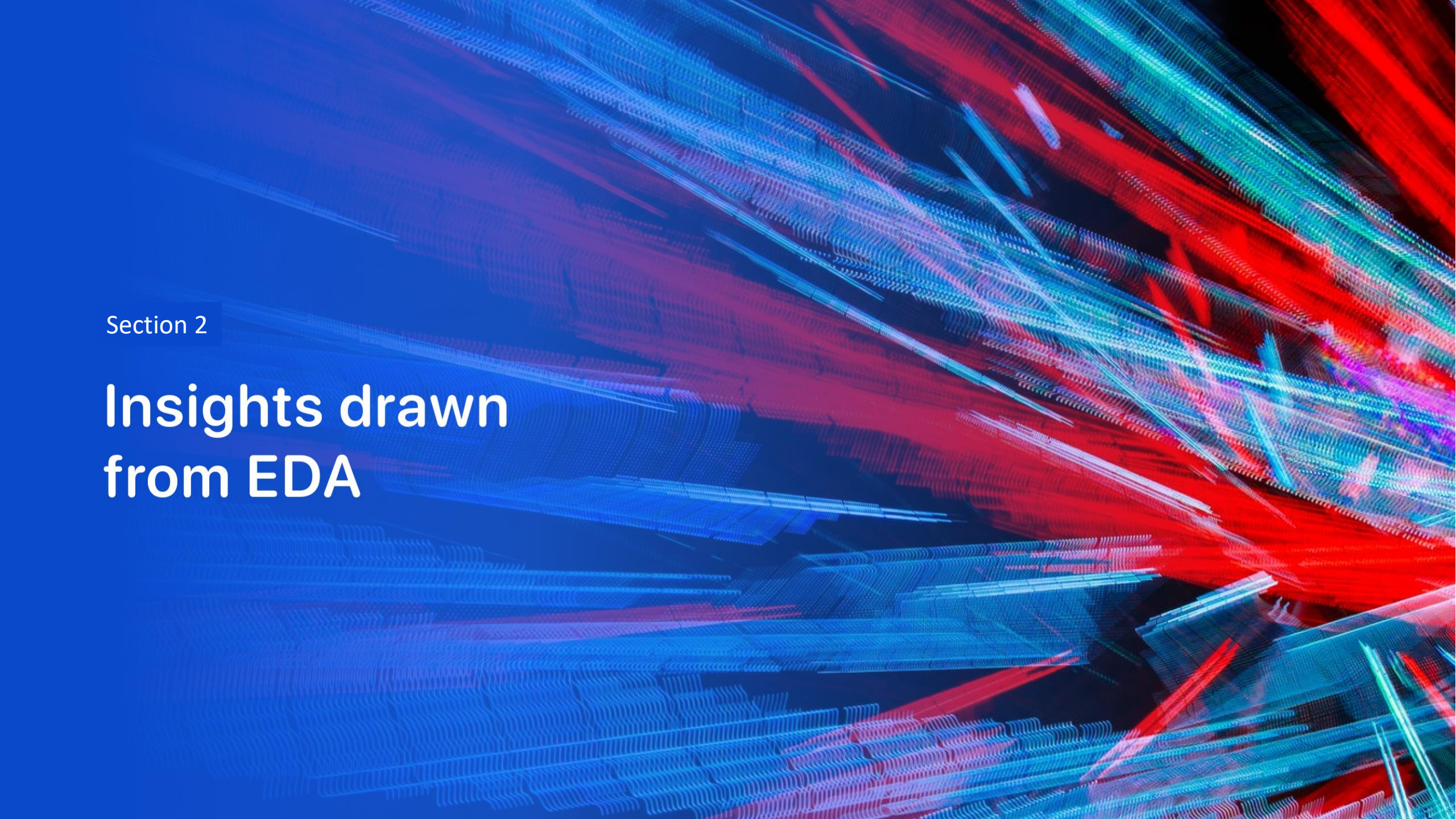
Predictive Analysis (Classification)



Notebook

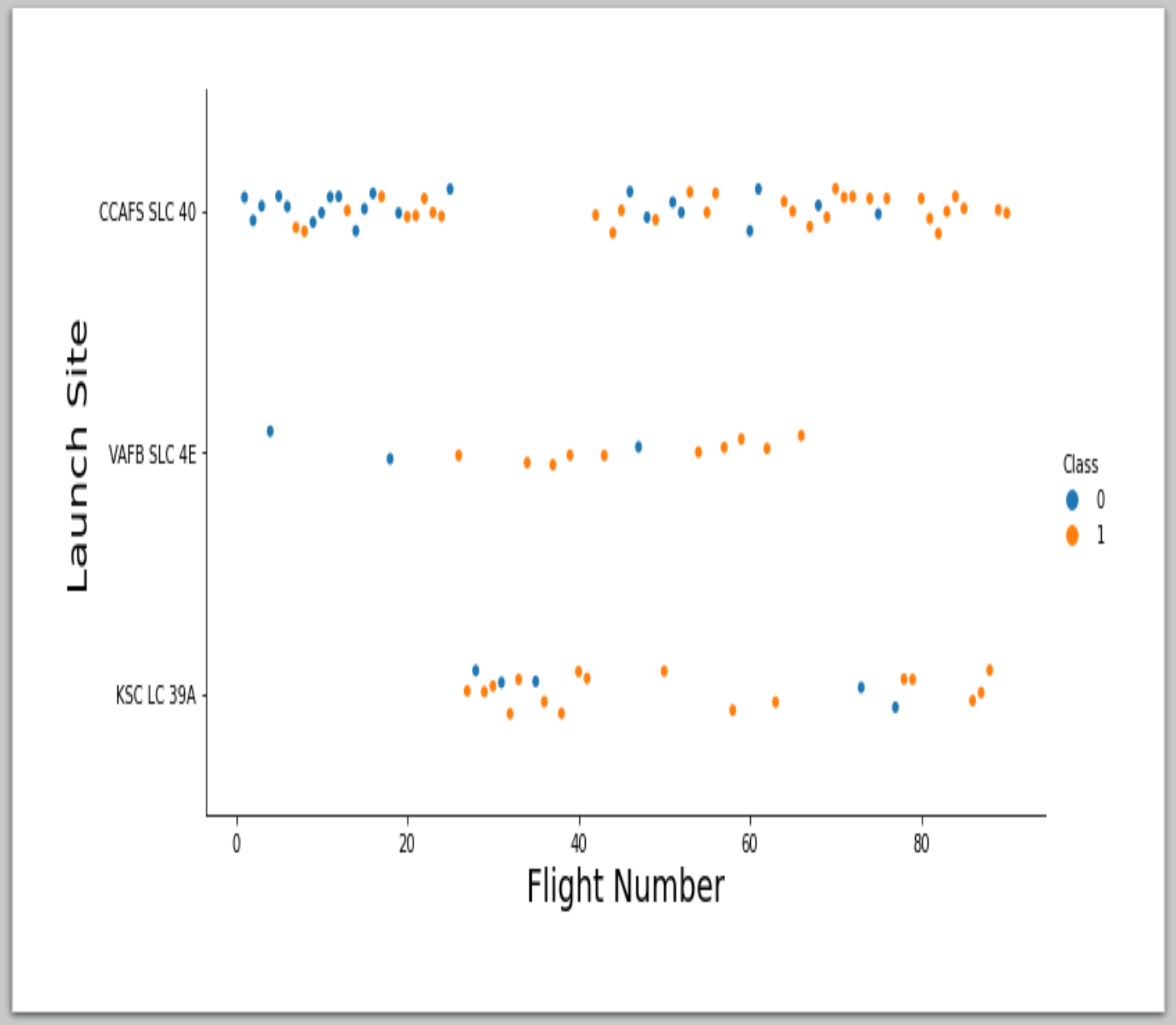
Results

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


Section 2
Insights drawn from EDA

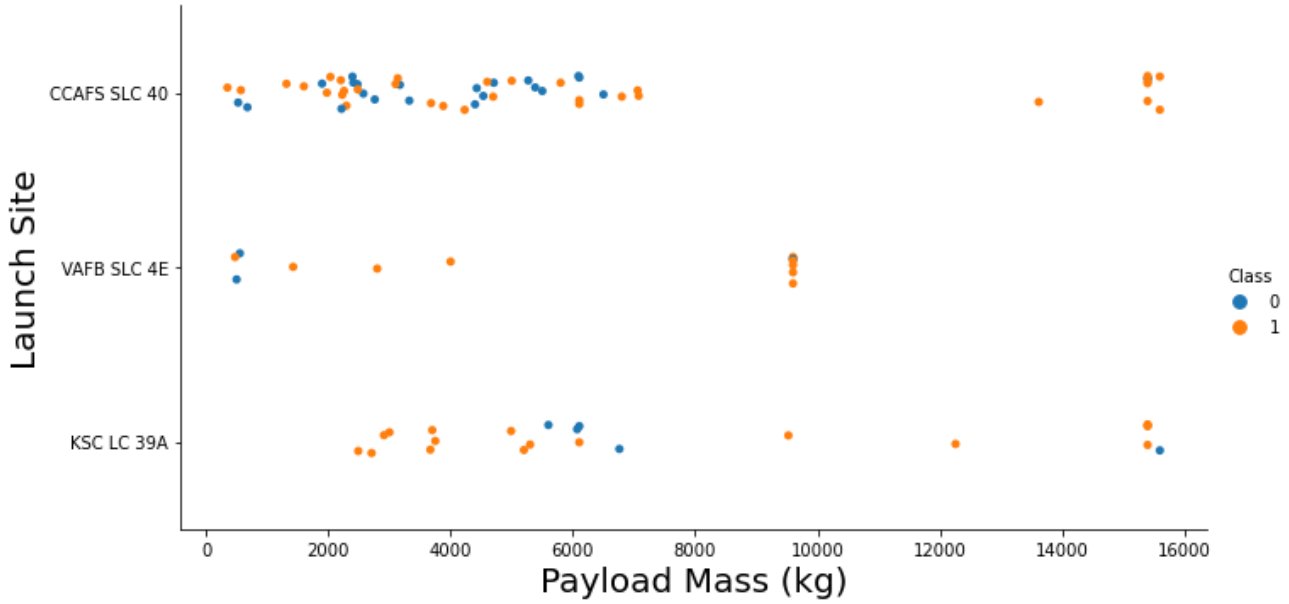
Flight Number vs. Launch Site

- According to the Scatter Plot below, it is possible to notice that the success rate of a launch increased with time.
- Furthermore, it is evident that both CCAFS SLC 40 and KSC LC 39A accumulate more launches and have similar success rates.

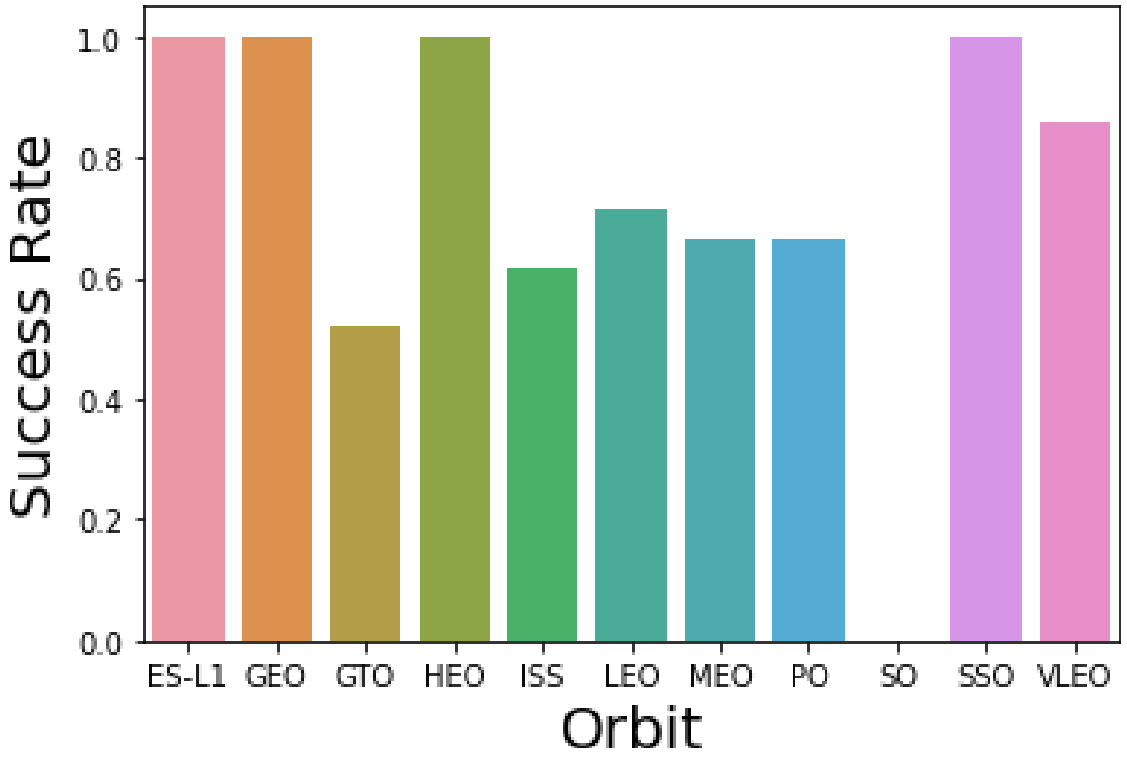


Payload vs. Launch Site

- The graph shows that the success rate for all launch sites are higher when the payload mass is greater than 8000 kg. However, this does not mean that mass is the only variable that increases success rate.
- The majority of successful launches were carrying a payload mass between 2000 kg and 6000 kg.



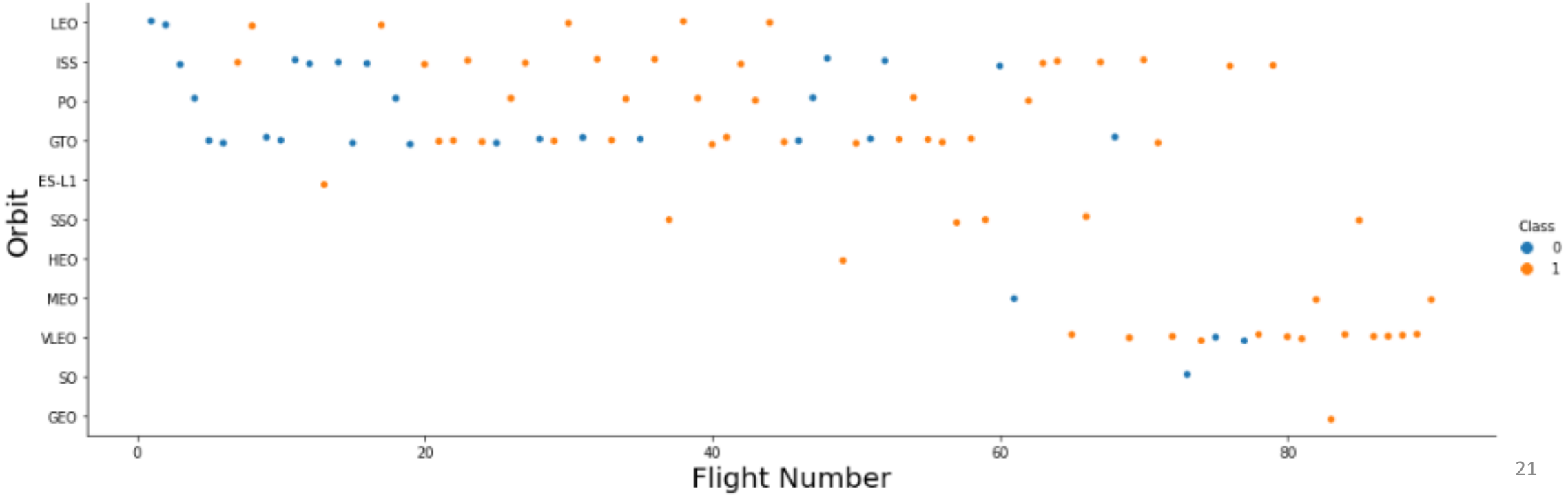
Success Rate vs. Orbit Type



- The Bar Chart shows that the maximum success rate was settled by the ES-L1, GEO, HEO and SS0 orbits.
- The GTO orbit hit the lower success mark, while the others score around 70% in average.

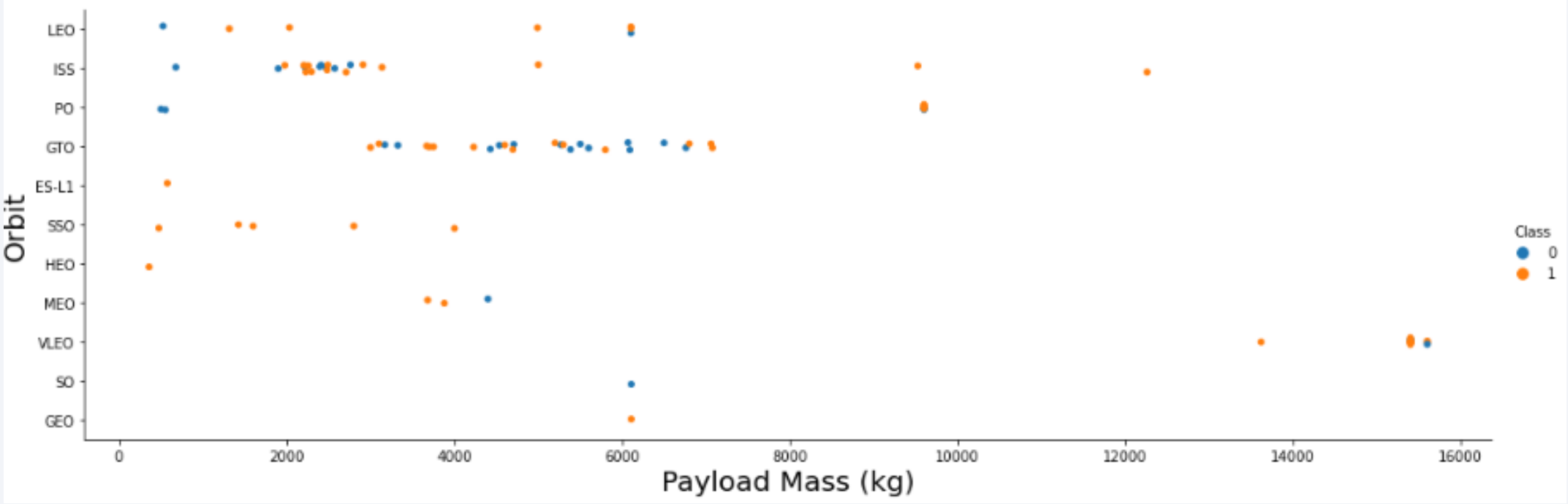
Flight Number vs. Orbit Type

- SSO orbit hit the 100% success rate with five launches
- most of the launches were destined to the ISS, PO, GTO and VLEO orbits. These four orbits together have approximately 65% average success rate.
- More recent launches are concentrated in VLEO and ISS. LEO launches seem to have been discontinued. GTO launches seem to have the highest number of failures. Only one each of SO, GEO, HEO, ES-L1 launch were ever attempted.



Payload vs. Orbit Type

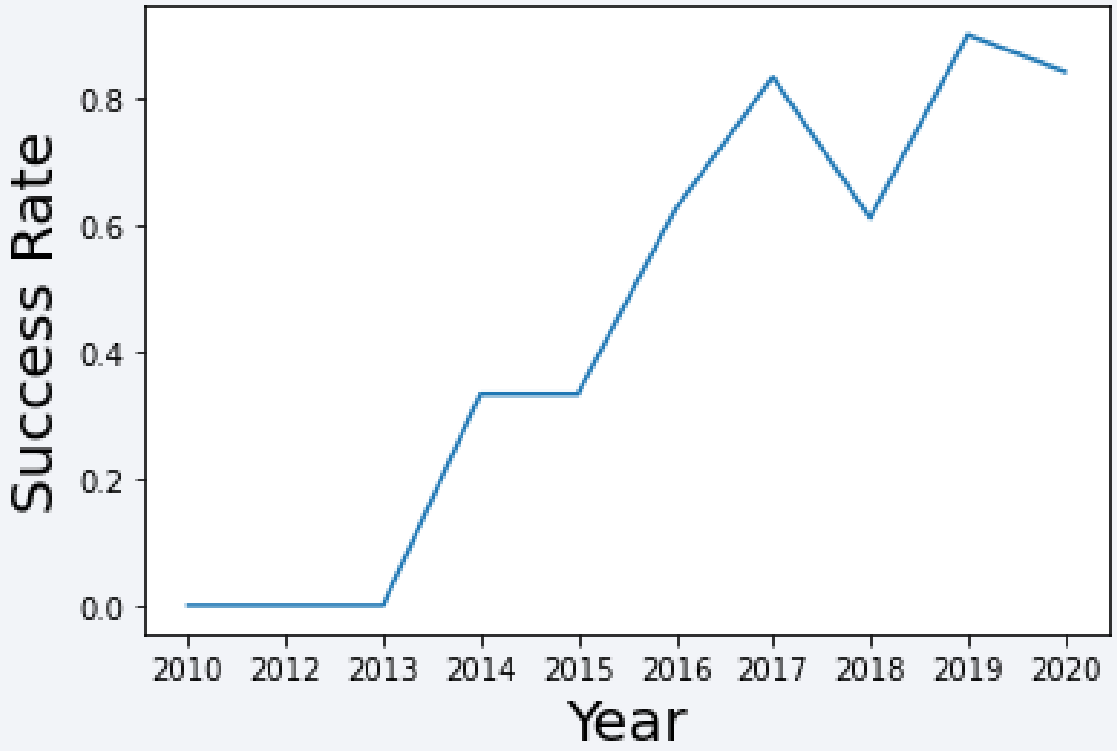
- We observe that the success rate is near 100% for ISS, PO and VLEO orbits when the payload mass exceeds 8000 kg.



- Most successful launches were carrying a payload between 2000 kg and 6000 kg. In the graph we notice a greater concentration of launches in the ISS and GTO orbits.

Launch Success Yearly Trend

- We observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- The DISTINCT statement was used to return only unique values from the launch_site column
- **select distinct LAUNCH_SITE from SPACEXTBL;**

  distinct is SQL function that selects unique values in provided column

  launch_site is the column of interest

  spacextbl is the table of interest

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

Launch Site Names Begin with 'CCA'

- select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' LIMIT 5;
  - The query statement uses the wildcard % after CCA meaning that only the values beginning with CCA will match the condition of the WHERE clause. The number of rows returned will be limited by 5

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

Total Payload Mass

- select sum(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD_MASS from SPACEXTBL where CUSTOMER = 'NASA (CRS)';

sum function sums values of column & '**as**' renames output column to 'TOTAL_PAYLOAD_MASS ;

where clause filters 'NASA (CRS)' from client

TOTAL_PAYLOAD_MASS
45596

Average Payload Mass by F9 v1.1

select avg(PAYLOAD_MASS__KG_) as AVERAGE_PAYLOAD_MASS from SPACEXTBL

The query statement uses the built-in function AVG to calculate the average payload mass for flights with version F9 v1.1 of the booster, filtered by the WHERE clause.

AVERAGE_PAYLOAD_MASS
2928.4

First Successful Ground Landing Date

Query >>> select min(DATE) as FIRST_SUCCESSFUL_LANDING from SPACEXTBL

where LANDING__OUTCOME like 'Success%';

- The built-in function MIN is used to return only the lowest date value
- wildcard % after Success, meaning that only the values beginning with Success will match the condition of the WHERE clause

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SQL >>> select distinct BOOSTER_VERSION from SPACEXTBL
where LANDING__OUTCOME = 'Success (drone ship)' and
PAYLOAD_MASS__KG_ < 6000 and PAYLOAD_MASS__KG_ > 4000;
```

DISTINCT statement was used to return only unique values from the booster_version column

WHERE clause was used to filter records that fulfill the following conditions:

boosters which have success in drone ship

payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

SQL >>> select MISSION_OUTCOME, count(MISSION_OUTCOME) as MISSION_COUNT from SPACEXTBL GROUP BY MISSION_OUTCOME

GROUP BY statement groups rows that have the same values for the mission outcome.

COUNT is used to return the number of rows for each category (or group) of mission outcome.

mission_outcome	mission_count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

SQL >>> select BOOSTER_VERSION, PAYLOAD_MASS__KG_

from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);

Sub-query is used to find maximum Payload mass. This output

is used in where clause to find which Booster_Version carries

this.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

SQL >>> select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE, DATE from SPACEXTBL where LANDING__OUTCOME = 'Failure (drone ship)' and year(DATE) = 2015;

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

- WHERE clause in this statement is used to filter by two conditions.

  First - failed landing outcome in drone ship.

  Second - YEAR to match only the dates where the year is 2015.

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

```
SQL >>> select LANDING__OUTCOME,
count(LANDING__OUTCOME) as LANDING_COUNT
from SPACEXTBL
where DATE between '2010-06-04' and '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY LANDING_COUNT DESC;
```

landing__outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

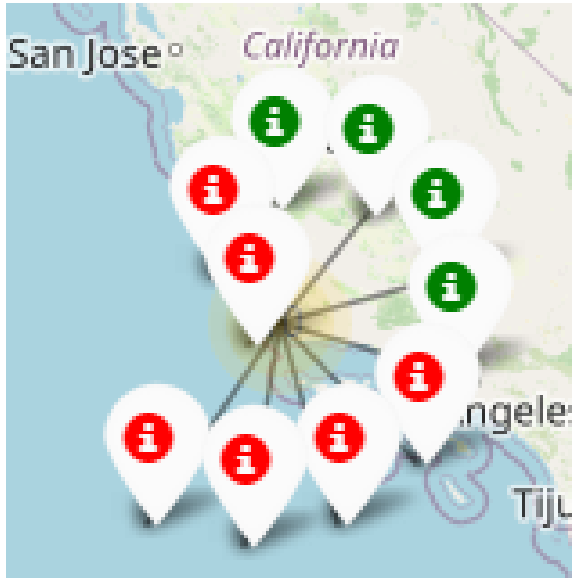
This query uses the group by statement and the built-in function COUNT to calculate the number of landing outcomes. The WHERE clause is used to filter by date and the ORDER BY keyword sorts the result set in descending order.

Section 3

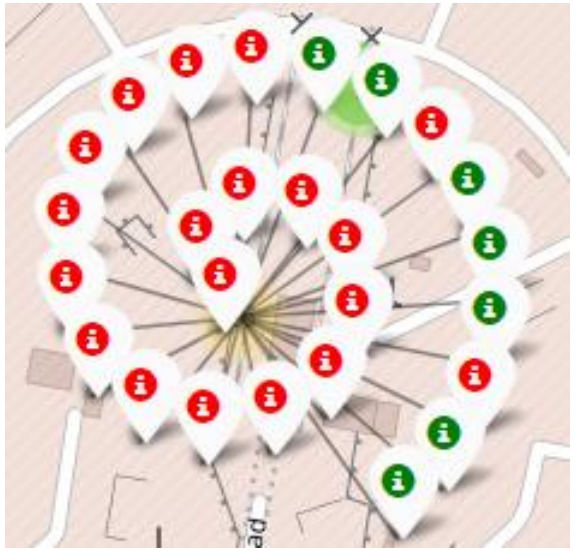
Launch Sites Proximities Analysis

SpaceX Launch Sites Locations

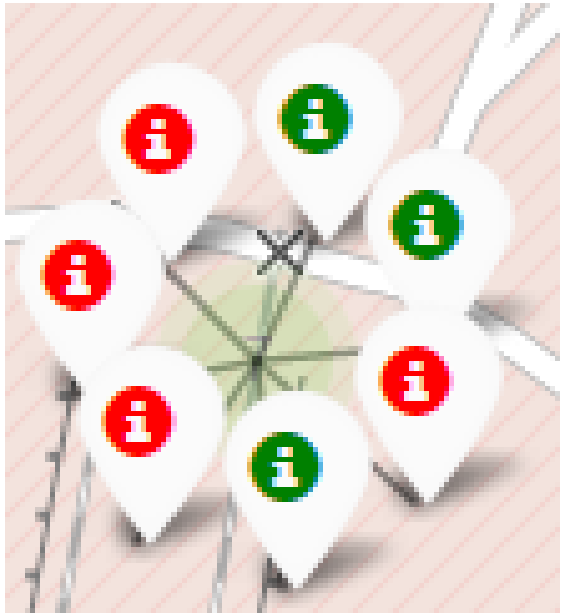
- Launch sites are on either the West or East coast of United States & their latitude is just above the Tropic of Cancer



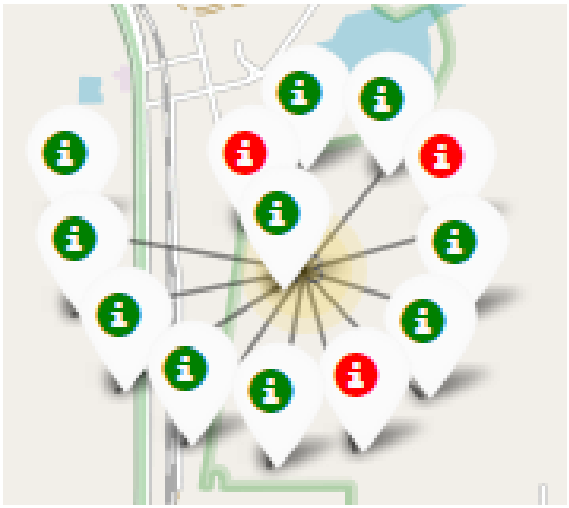
VAFB SLC-4



CCAFS LC-40

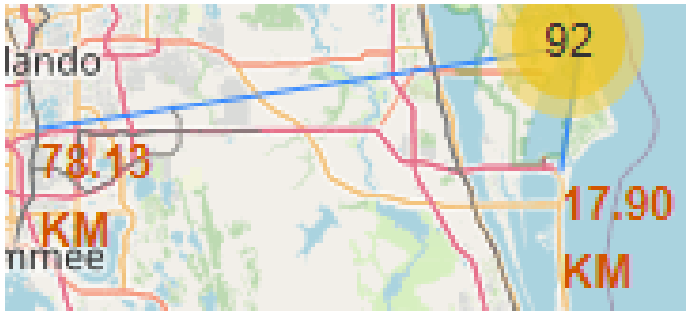


CCAFS SLC-40



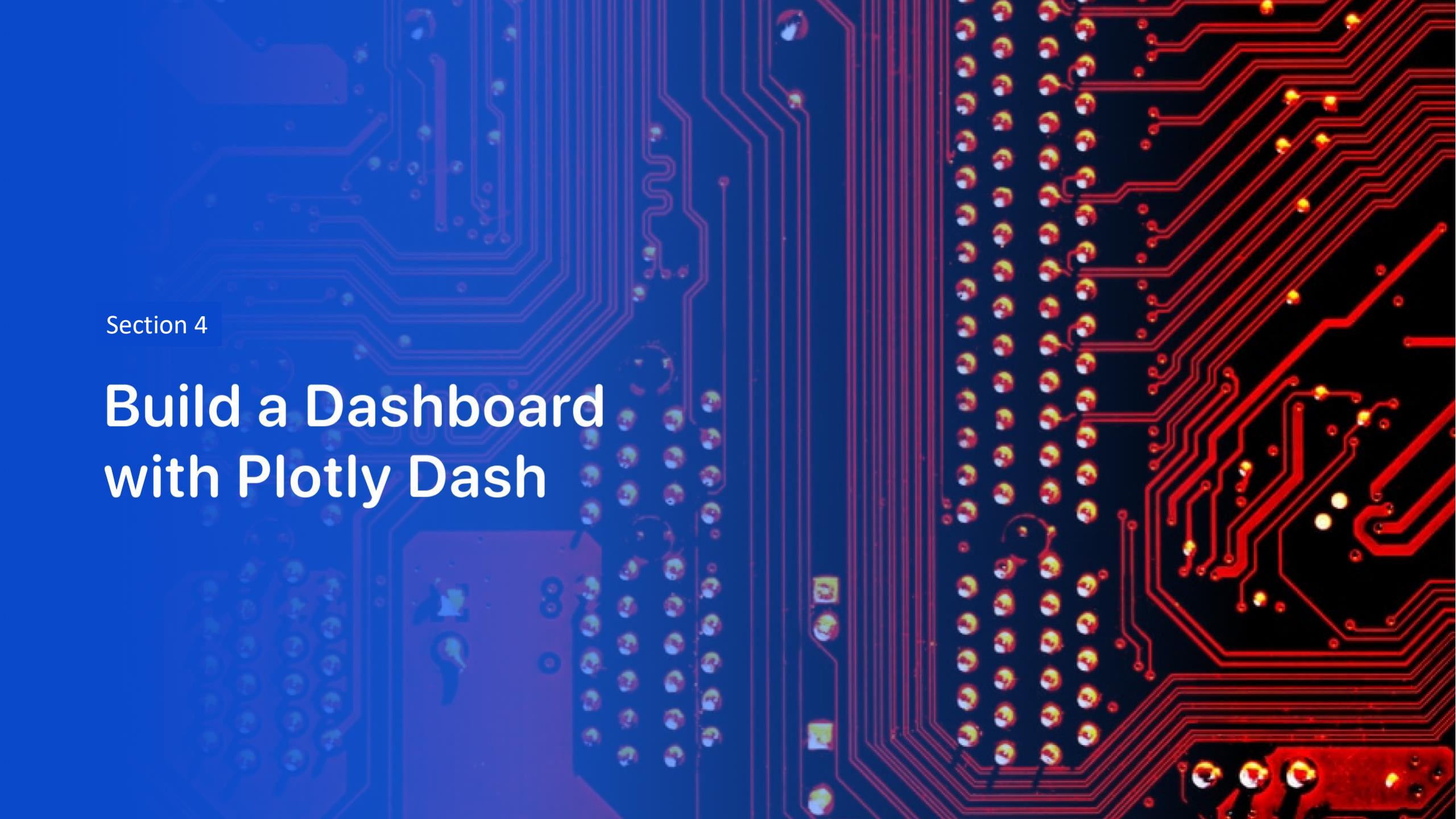
KSC LC-39A

SpaceX Launch Outcomes

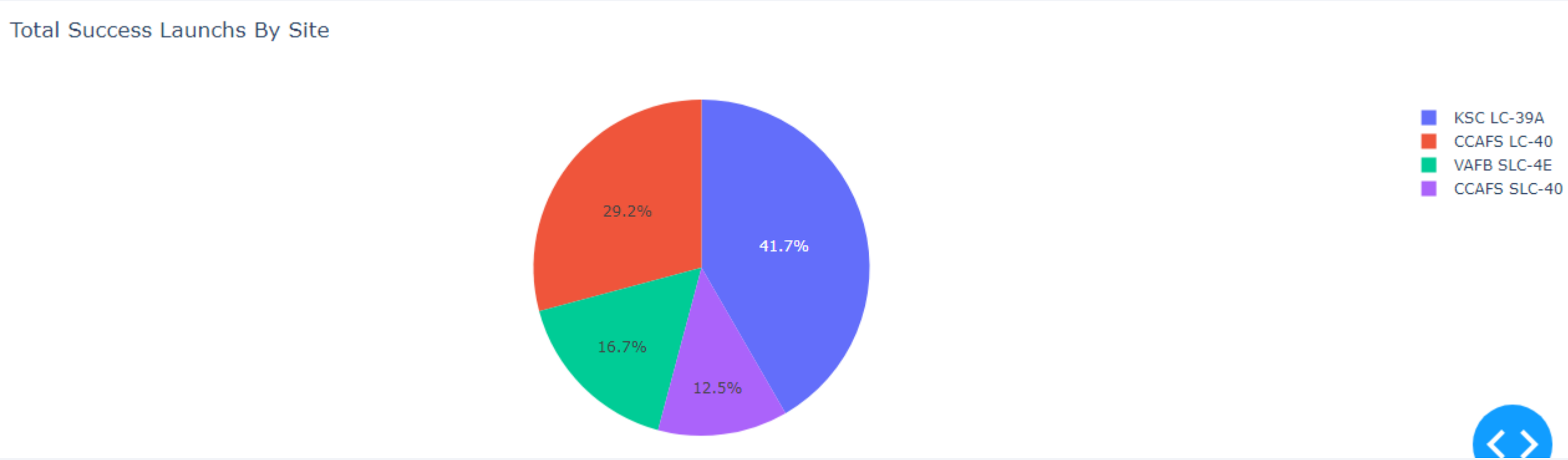


Distances between a launch site to its proximities

- Launch sites are close to coastline but located at a reasonable distance from nearest city, railway & highway, coastline.

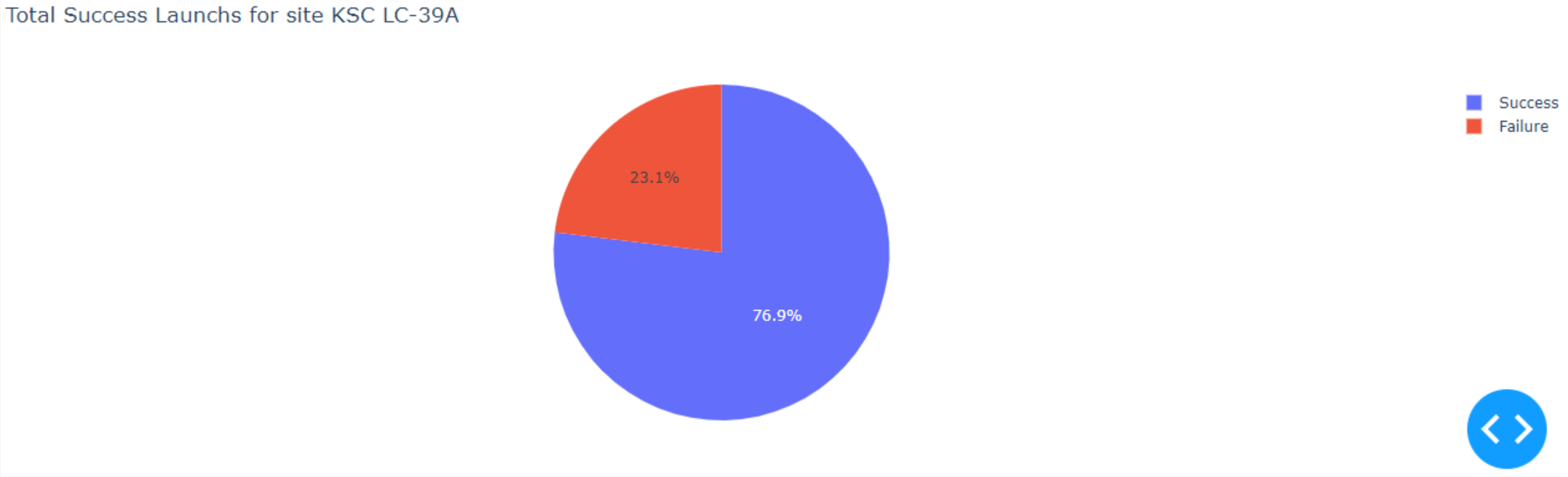

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site



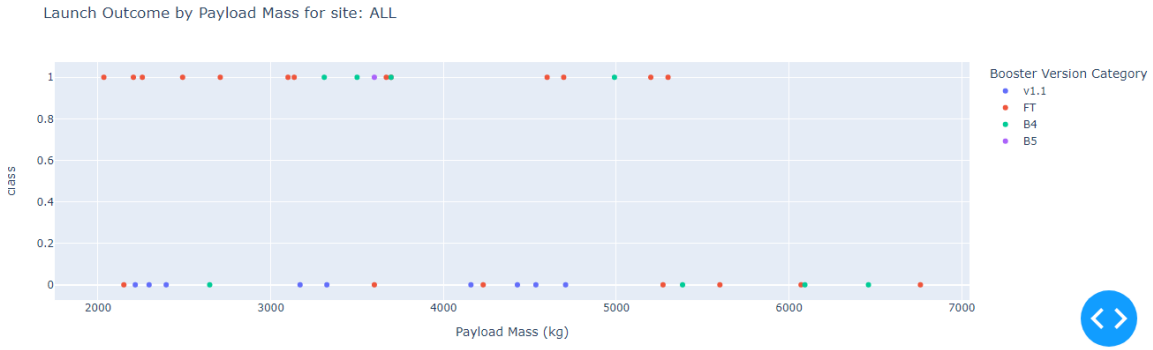
The pie chart confirms previous information, gained from the map markers, that the KSC launch complex has the highest success rate.

Total Success Launches for Kennedy Space Center

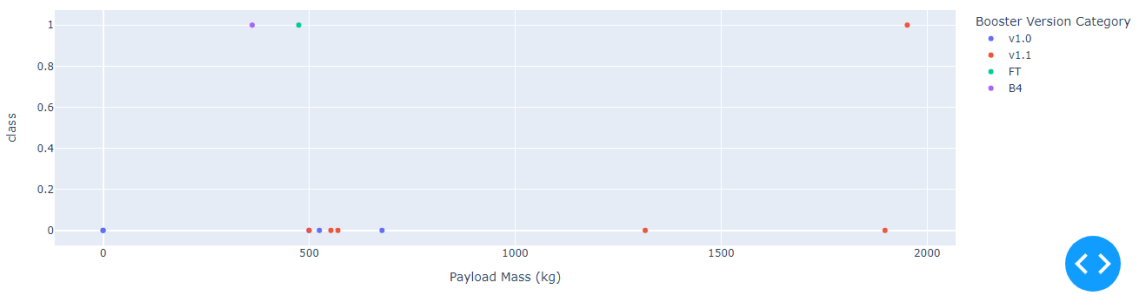


Kennedy Space Center has maximum successful launch rate. Noe : Bur there is no clear evidence that the location is the only reason for successful outcome.

Launch Outcome by Payload Mass



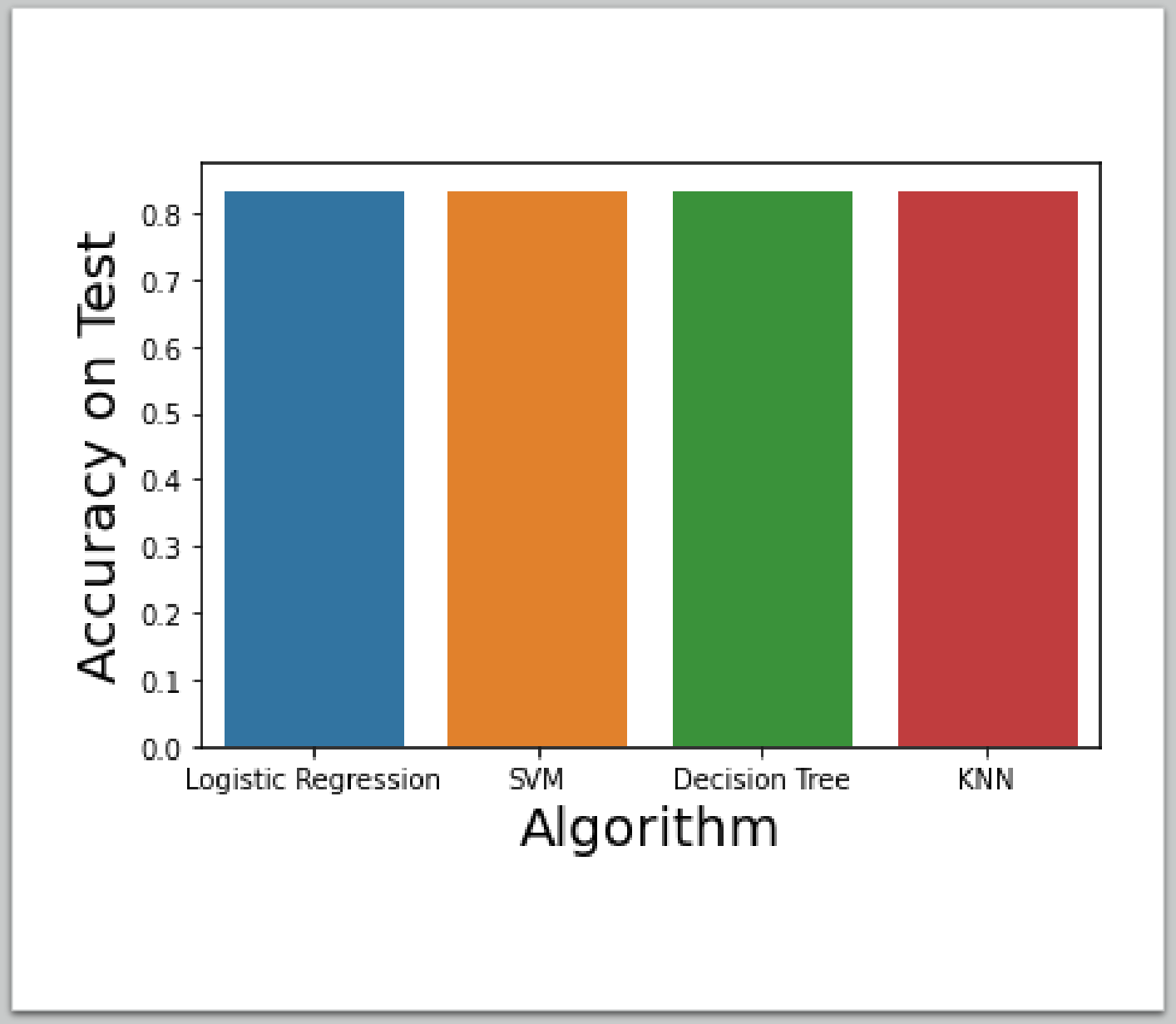
Payload mass, between 2000-6000 kg, the success rate is approximately 52% and the booster version FT is the most used and successful.



Payload mass, lower than 2000 kg, there is a higher failure rate and the booster version v1.1 is the most used.

Section 5

Predictive Analysis (Classification)

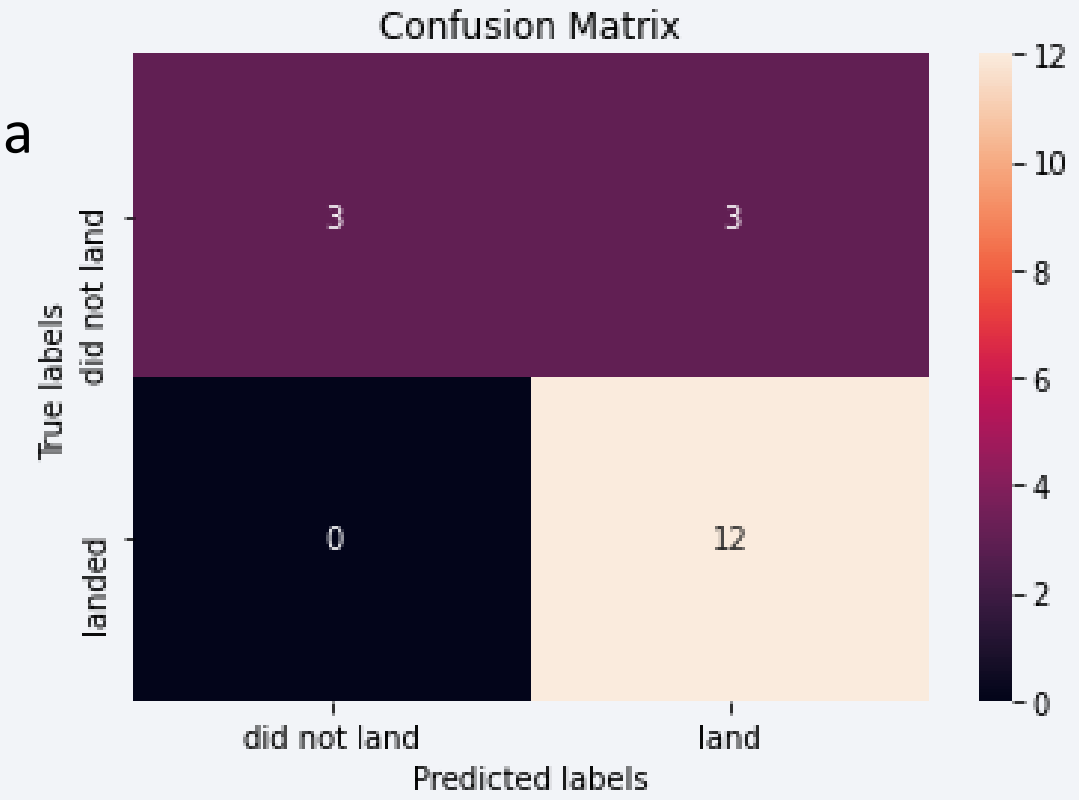


Classification Accuracy

- Based on accuracy performance of all models is same.

Confusion Matrix

- The confusion matrix shows that, given a test set of 18 records, we got 15 correct predictions and 3 incorrect.
- The incorrect prediction is a false positive, also known as a type I error or false alarm.



Conclusions

- **What determines whether a rocket will land successfully?**

  The most important factor is time. The success rate of Falcon 9 launches has grown continuously over the period 2013-2020. Newer SpaceX boosters (e.g. FT, B4, B5) greatly improved launch success rate & can carry more payload.

- **Which variables have the greatest influence on the landing success rate?**

  The Kennedy Space Center (KSC LC-39A) has the highest success rate of 76.9%.

  A payload mass between 2000 kg and 6000 kg is most likely to give a positive outcome.

  Most of the launches were destined to the ISS, PO, GTO and VLEO. These four orbits together have approximately 65% average success rate.

  The new booster versions contributed to a significant improvement in the success rate.

- **Can we estimate the cost of a launch by predicting whether the first stage will land?**

  According to the predictive analysis results it is possible, with a satisfactory confidence margin, to estimate costs based on launch data.

Appendix – Resources, Tools and Service

❖ Python Libraries

Matplotlib

Seaborn

Pandas

NumPy

Plotly Dash

Scikit-learn

❖ Tools

Visual Studio Code

MS PowerPoint

Adobe Photoshop

❖ Services

IBM Cloud

GitHub

Google App Engin

Notebook Link

Thank you!