



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bonaventure Ouedraogo
October 31, 2022



Outline

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

Executive Summary

- To determine If the first stage will land successfully, we collected data, performed data wrangling, exploratory data analysis (EDA), interactive visual analytics, and predictive analysis.
- We found that different launch sites have different success rates, and the first stage is more likely to land successfully as the flight number increases. More than 4/5 of the launches made from the East coast and closer to the Equator. Finally, 12/18 first stage successful landings are predicted.

Introduction

Space X is competitive in its industry because it is can *lower costs* by reusing the first stage of its rockets. We are tasked with the [Space X Falcon 9 First Stage Landing Prediction](#) to allow Space Y to make [informed decisions](#) about their project to *compete against SpaceX*. To achieve this goal, we will:

- Determine If the [first stage](#) will land successfully, and if Space X will [reuse](#) it by:
 - using Public Information.
 - training a Machine Learning model
- Support our marketing team in determining the [price of each flight](#) by:
 - gathering information about space X
 - and creating dashboards for our teams.

Section 1

Methodology

Methodology

Executive Summary

- Data **collection** methodology: from Space X REST API, and web scraping to collect Falcon 9 historical launch records.
- Perform data **wrangling**: Collected data was sampled and cleaned by dealing with nulls and then combined in a data set.
- Perform **exploratory data analysis** (EDA) using visualization and SQL
- Perform **interactive visual analytics** using Folium and Plotly Dash
- Perform **predictive analysis** using classification models: Models are trained by building a Machine Learning pipeline and evaluated through the confusion matrix using tuned hyperparameters.

Data Collection

Data sets **collection** method:

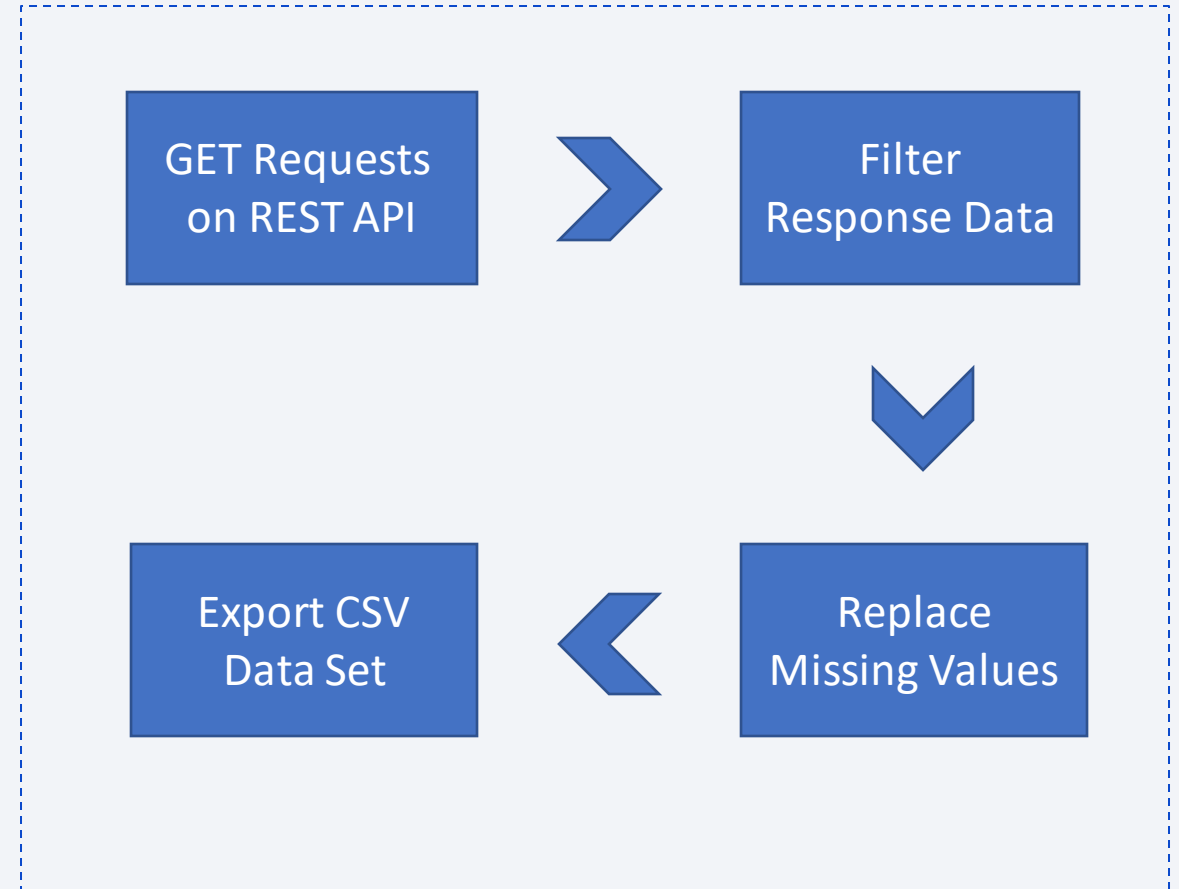
- Data collected using Space X REST API.
- Endpoint URLs were used to retrieve the data of the launches.
- The JSON documents obtained are normalized and converted into Pandas DataFrames
- Web scraping of Wikipedia Falcon 9 records using BeautifulSoup
- Then, parsing and conversion in DataFrames

Data **wrangling** method using an API:

- Targeting different endpoints with the API to gather missing data from prior documents.
- The New data set is filtered and sampled to keep only Falcon 9 launches.
- Replacing some Null values with means and dummy variables using one hot encoding

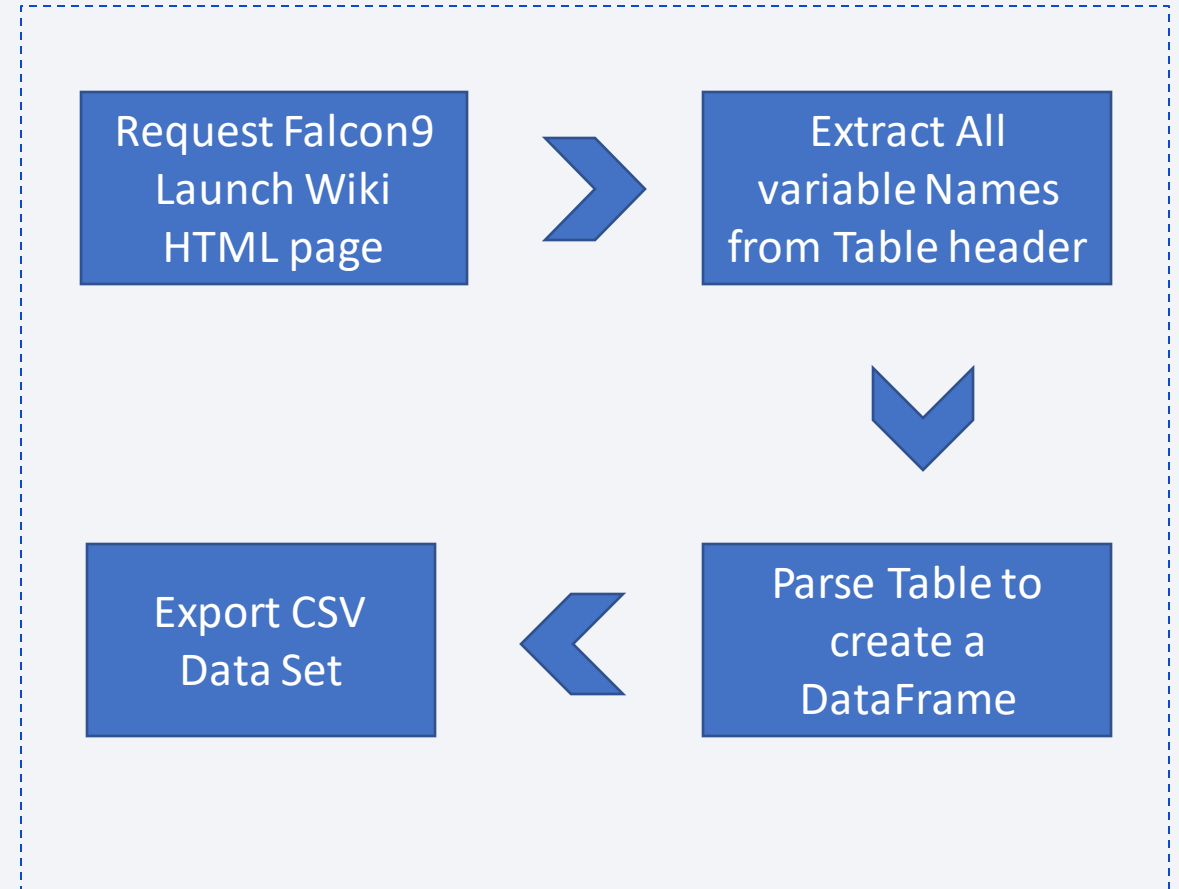
Data Collection – SpaceX API

- To collect Falcon 9 launches data using the Space X REST API, we performed **GET requests**, parsed and filtered the JSON documents from the response, and dealt with the missing values while cleaning. After performing the **basic data wrangling**, we exported the Data Frame as a CSV file.
- GitHub URL of the completed SpaceX API calls notebook: <http://bit.ly/3Um1eW4>



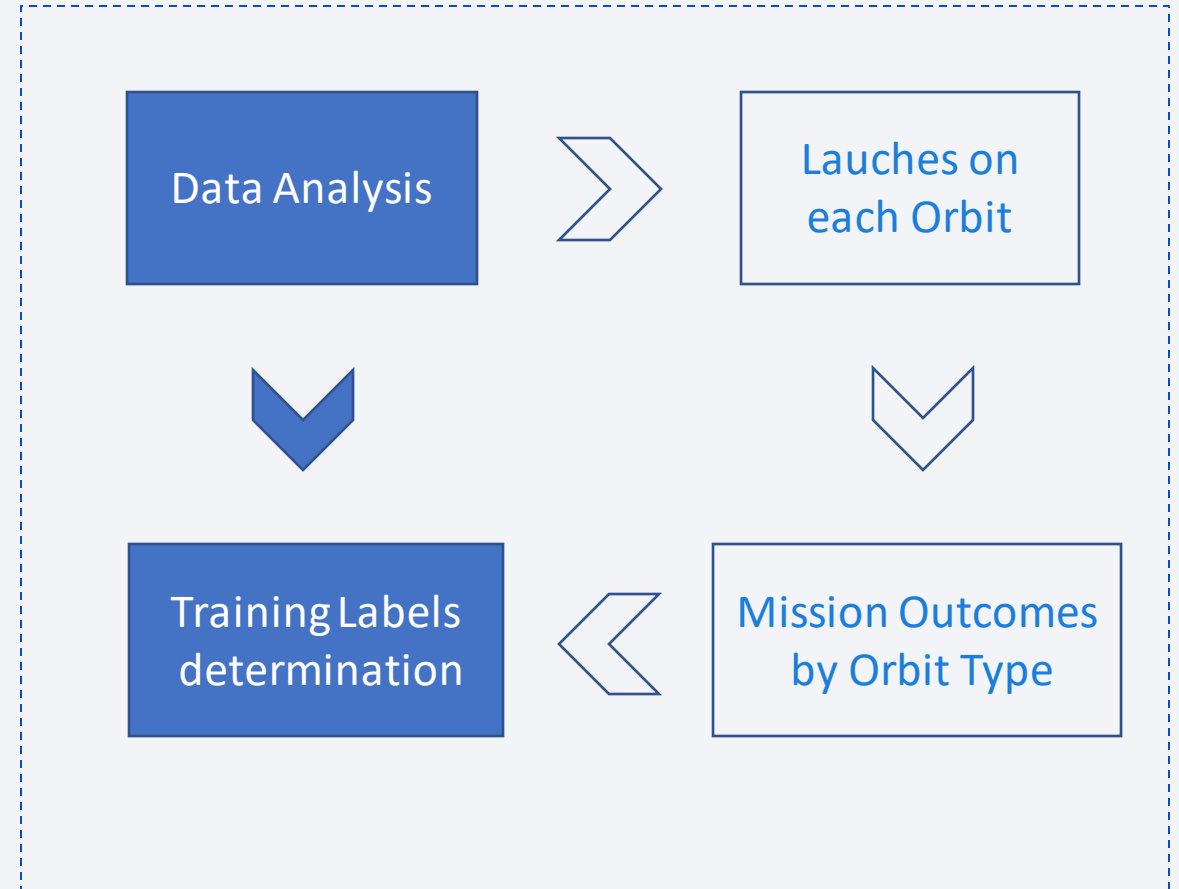
Data Collection - Scraping

- To scrap Falcon 9 launch records with BeautifulSoup, we performed an HTTP **GET request** on the wiki HTML page, extracted all column names from the HTML table header and **parsed** our tables to create a Data Frame. Then exported it in CSV format.
- GitHub URL of the web scraping notebook: <https://bit.ly/3DNqM8O>



Data Wrangling

- To perform the EDA and determine the labels, we calculated the number of launches on each site, their occurrences on each orbit and the occurrences of mission outcomes by orbit type. Then, we created a landing outcome label from Outcome column using one hot encoding.
- GitHub URL of completed data wrangling related notebooks: <http://bit.ly/3T0LWog>



EDA with Data Visualization

- Scatter plots to visualize correlations between the following variables that may affect the launch outcome:

FlightNumber/launch attempts and Payload Mass

Flight Number and Launch Site

Payload and Orbit type

- Bar chart to visualize the relationship between success rate of each orbit type. They are appropriate for comparing different (few) categorical values like the Orbit types.
- Line chart to get the average launch success trend. This chart is used to show change in information over time or/and comparisons between variables.

GitHub URL of your completed EDA with data visualization notebook: <http://bit.ly/3T10G6X>

EDA with SQL

- List the names of the **unique (DISTINCT) launch sites** in the space mission
- List 5 records where launch sites **begin with (LIKE)** the string 'CCA'
- Get the **total (SUM)** payload mass carried by boosters launched by NASA (CRS)
- Get **average (AVG)** payload mass carried by booster version F9 v1.1
- List the date when **the first (MIN)** successful landing outcome in ground pad was achieved.
- List the names of boosters with success in drone ship and have payload mass in **range (BETWEEN)** 4000 – 6000.
- **List the total (GROUP BY)** number of successful and failure mission outcomes.
- List the names of the booster versions which have carried the **maximum (MAX)** payload mass
- List the failed landing outcomes in drone ship, their booster versions, and launch site names in year 2015.
- Rank (ORDER BY) the **count** of landing outcomes between the date 2010-06-04 and 2017-03-20.

GitHub URL of your completed EDA with SQL notebook: <http://bit.ly/3WDsy47>

Build an Interactive Map with Folium

Map objects used:

- **Geographical base map**: to provide context for the data Space X launch data
- **Choropleth maps**: to represent spatial variations of a quantity
- **Markers**: colored to distinguish the launch outcome (class) and clustered to group the information spatially and logically.
- **Circles**: to represent the zone of the launch sites
- **Lines**: to represent the distance between two elements

GitHub URL of your completed interactive map with Folium map: <http://bit.ly/3TfC62l>

Build a Dashboard with Plotly Dash

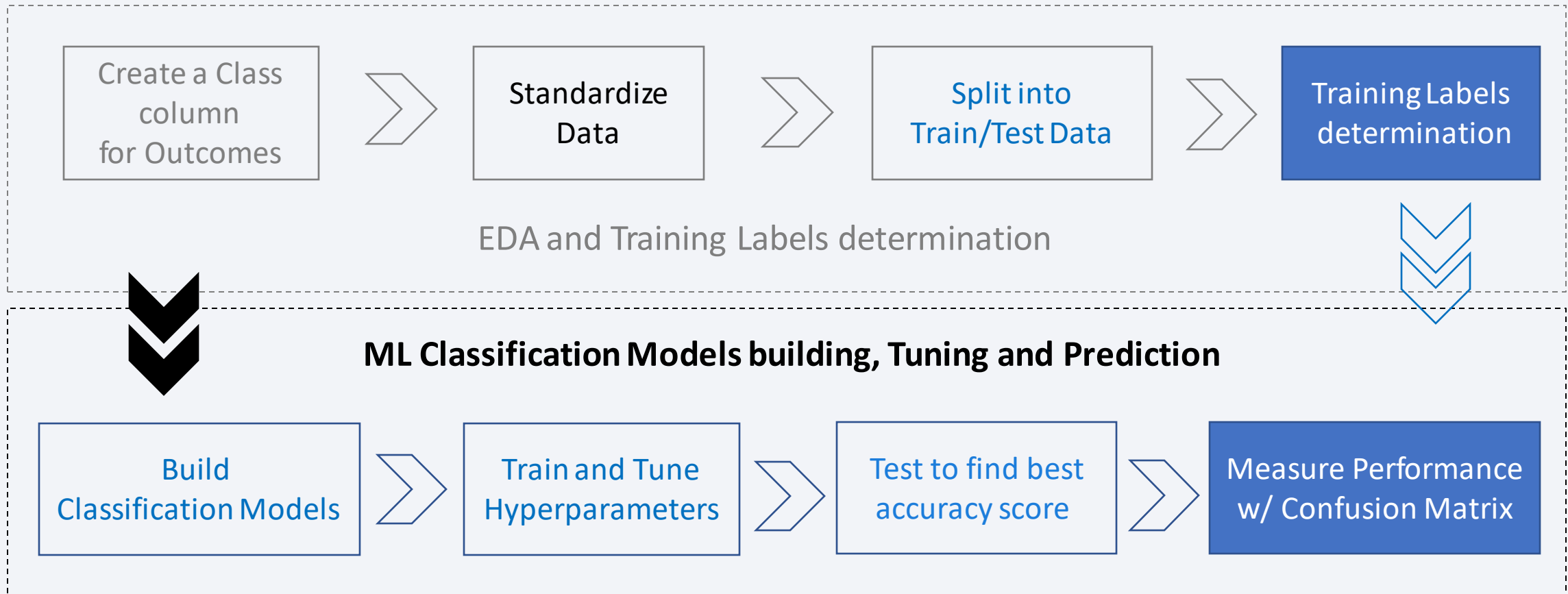
Plots/charts and Interactions used in the Dashboard:

- **Pie charts:** to provide context for the data Space X launch data
- **Scatter plots:** to represent spatial variations of a quantity
- **Drop-down menu:** to select individual launch site views/display data for All sites.
- **Range Slider:** to filter the desired range for Payload Mass (kg) values to display.

Purpose: **ease and speed-up exploration** with interactivity.

GitHub URL of your completed interactive map with Folium map: <http://bit.ly/3TfC62l>

Predictive Analysis (Classification)



GitHub URL of completed predictive analysis lab: <http://bit.ly/3h72cqy>

Results

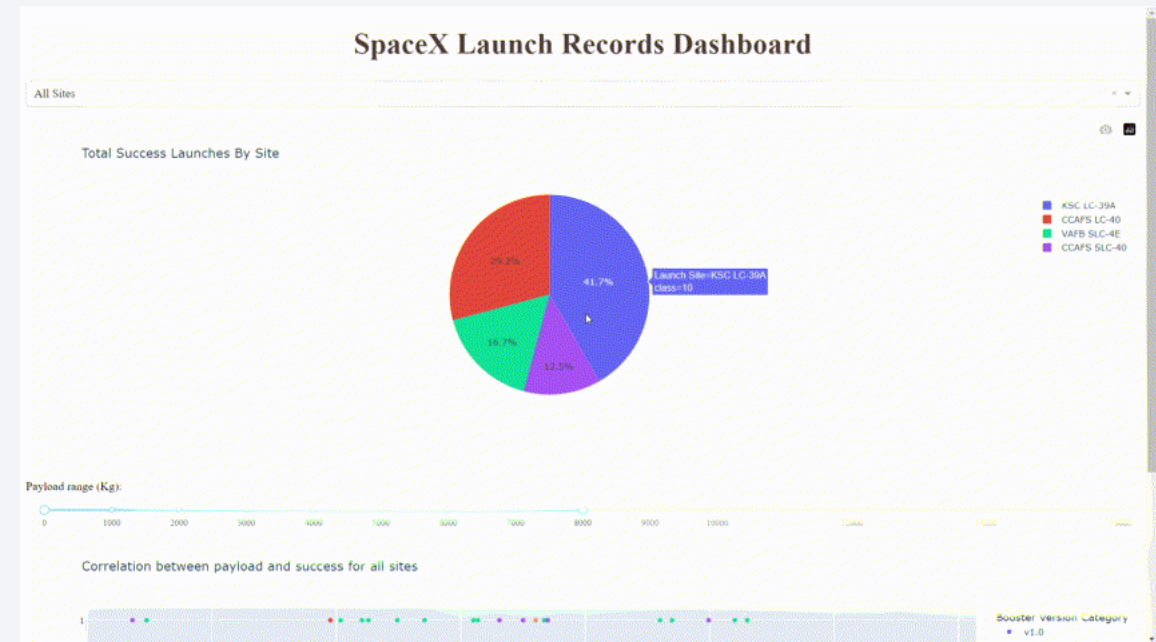
Exploratory Data Analysis results

- Different launch sites have different success rates
- The VAFB-SLC launch site has no rocket launch for heavy payload mass (greater than 10000)
- Launches to ES-L1, SSO, HEO and GEO Orbits have high success rates
- As the flight number increases, the first stage is more likely to land successfully.
- All launch sites are in very close proximity to the coasts

Predictive analysis results

- The tree classifier yields the best results and **12/18 first stage successful landings** are predicted.

Interactive analytics demo in screenshots



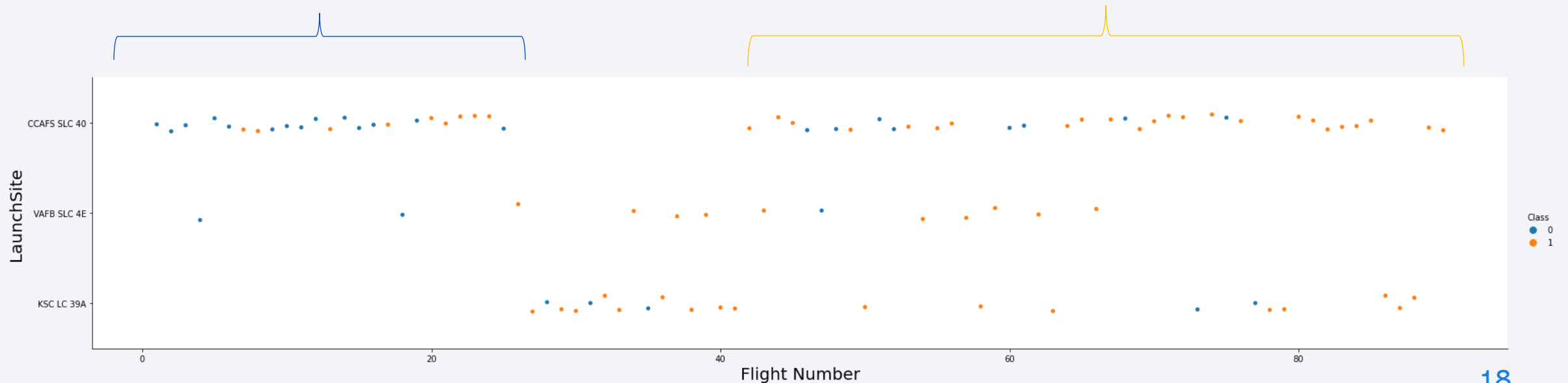
The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of blue and red, creating a sense of motion or data flow. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is high-tech and digital.

Section 2

Insights drawn from EDA

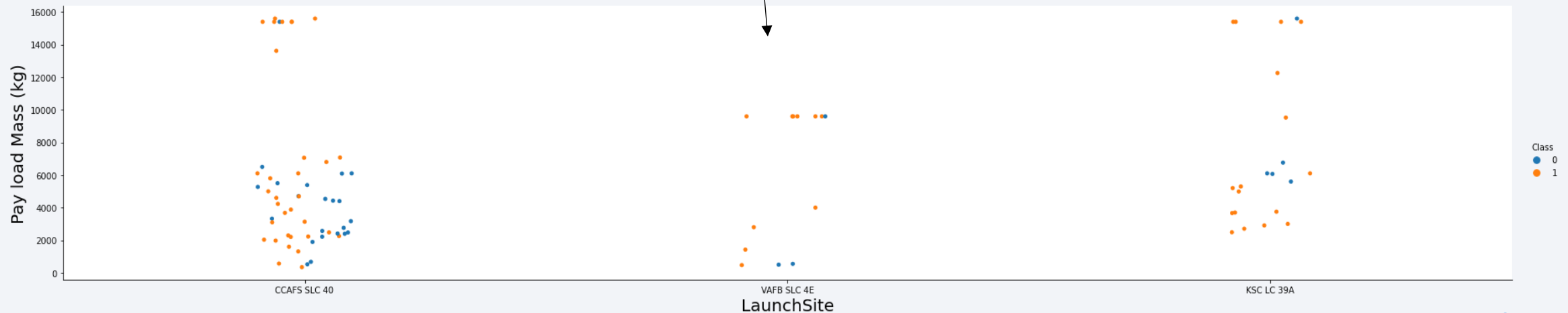
Flight Number vs. Launch Site

- As the flight number increases, the first stage is more likely to land successfully.



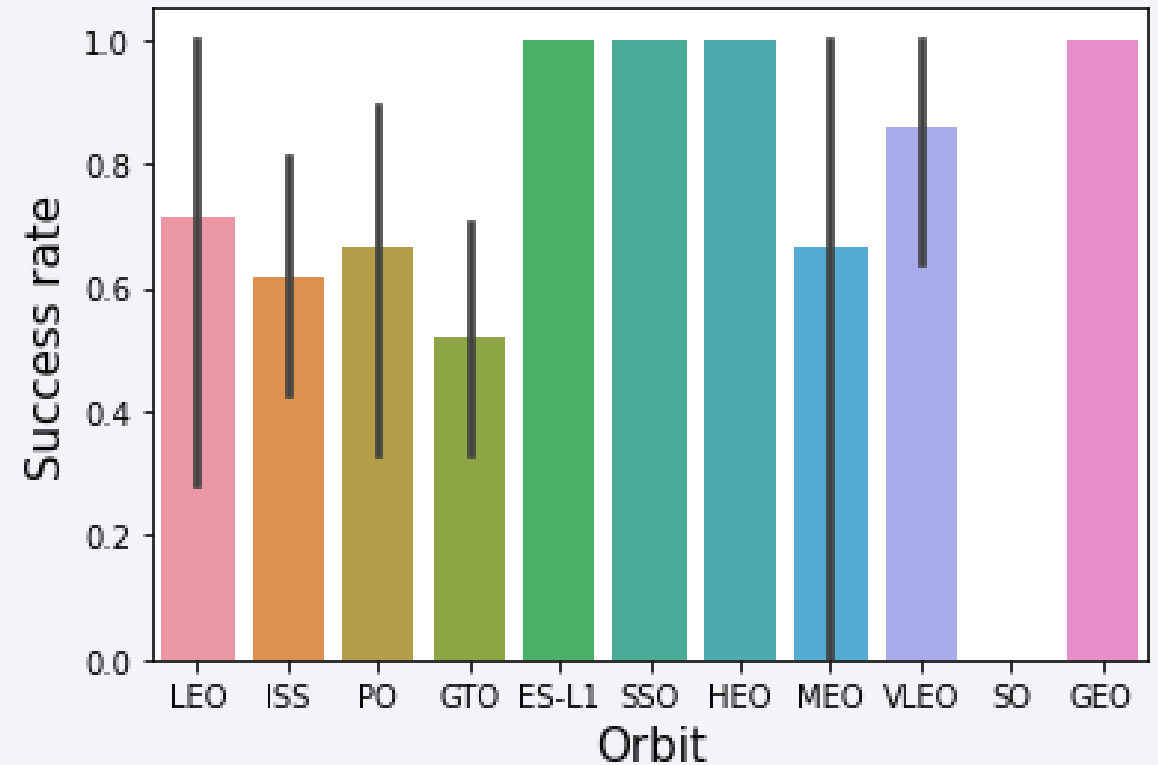
Payload vs. Launch Site

- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).



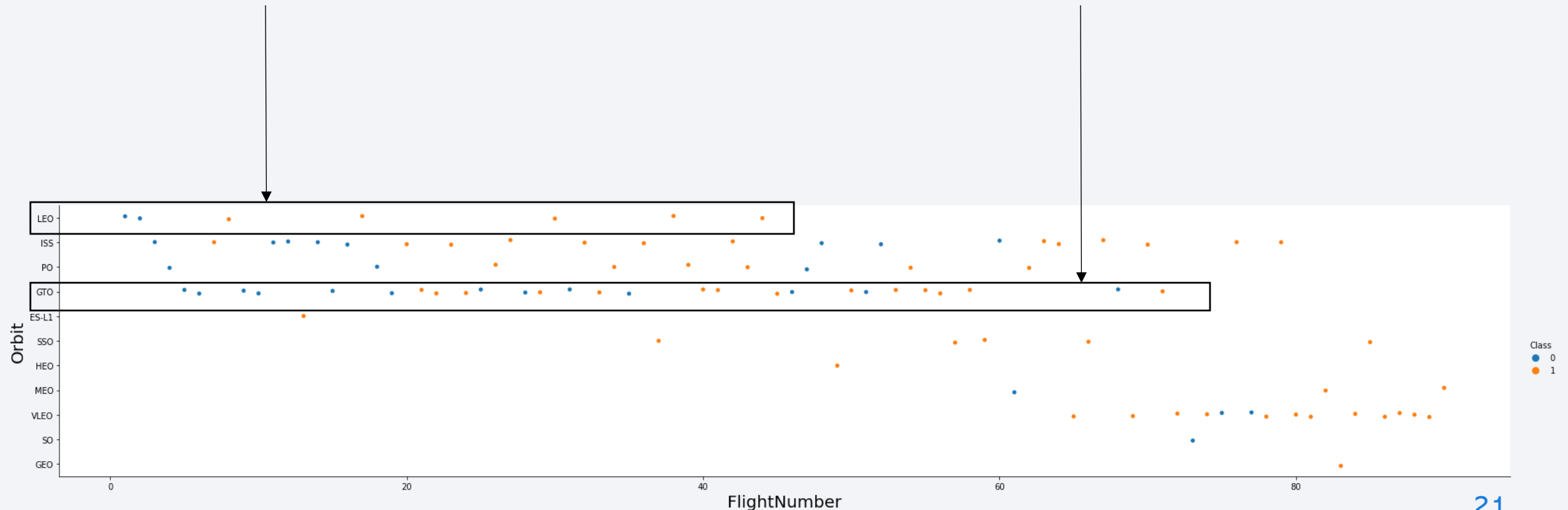
Success Rate vs. Orbit Type

- Launches to ES-L1, SSO, HEO and GEO Orbits have high success rates.
- The unique SO Orbit launch failed.



Flight Number vs. Orbit Type

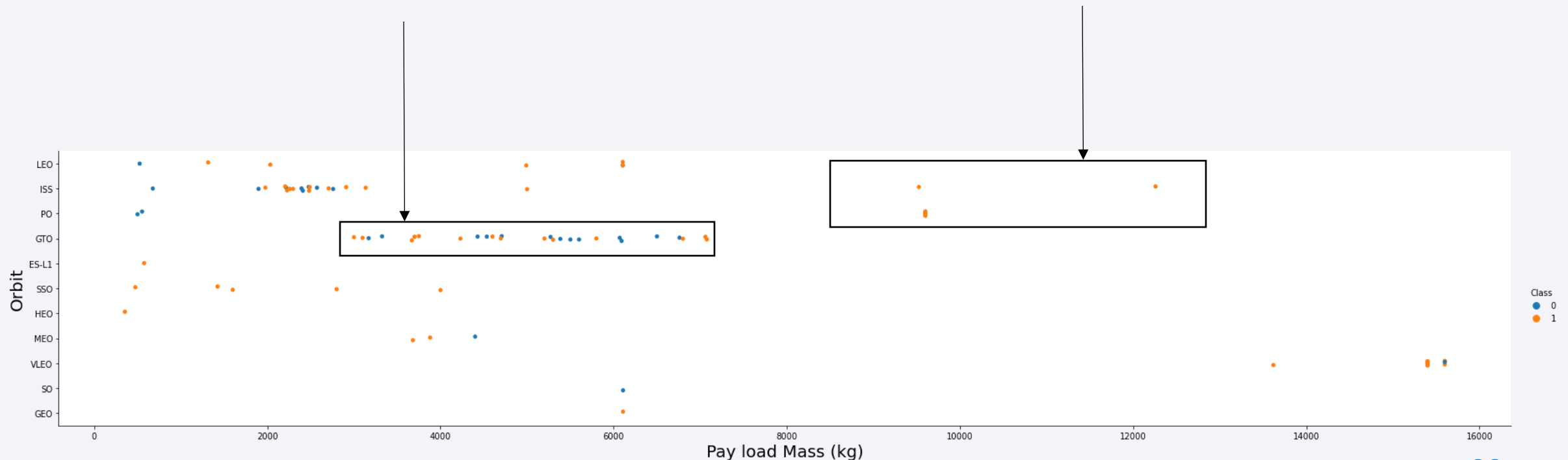
- in the LEO orbit the Success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

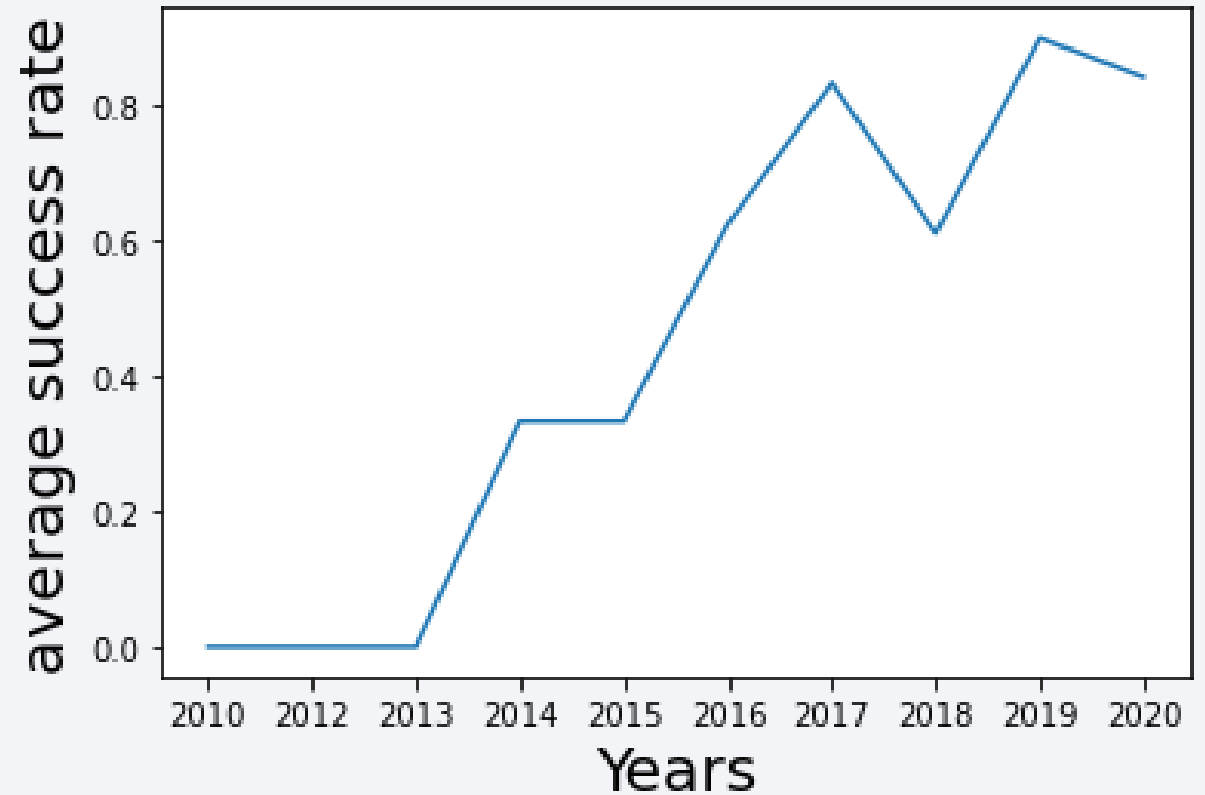
- However, for GTO no insightful conclusion can be drawn because both landing outcomes seems to be scattered randomly.

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.



Launch Success Yearly Trend

- Overall, the success rate since 2013 kept increasing till 2020;
- Despite some setbacks in 2018.



All Launch Site Names

- Names of the unique launch sites

launch_site

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

- SQL query from the DB2 Database PZC04681

%%sql

```
SELECT DISTINCT LAUNCH_SITE  
FROM PZC04681.SPACEXTBL;
```

This query commands to select all distinct values in the LAUNCH_SITE column from the table PZC04681.SPACEXTBL

Launch Site Names Begin with 'CCA'

%%sql

```
SELECT *  
FROM PZC04681.SPACEXTBL  
WHERE LAUNCH_SITE LIKE 'CCA%'  
FETCH FIRST 5 ROWS ONLY;
```

- SQL query from the DB2 Database PZC04681

This query commands to select all (*) values in the LAUNCH_SITE column from the table PZC04681.SPACEXTBL; Where the values start (LIKE "...%") with the letters "CCA".

DATE	time_utc	booster_version	launch_site	payload	payload_mass_kg	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Query results



Total Payload Mass

- Total payload carried by boosters from NASA



- SQL query from the DB2 Database PZC04681

```
%%sql
```

```
SELECT  
    SUM(payload_mass__kg_) AS Total_Payload_Mass_NASACRS  
FROM  
    PZC04681.SPACEXTBL  
WHERE CUSTOMER='NASA (CRS)';
```

This query commands to select, from the table PZC04681.SPACEXTBL, the SUM of payload_mass_kg_ column values and name it (AS) Total_Payload_Mass_NASACRS; But filter (WHERE) to only keep the records that values in the CUSTOMER column = "NASA (CRS)"

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1



- SQL query from the DB2 Database PZC04681

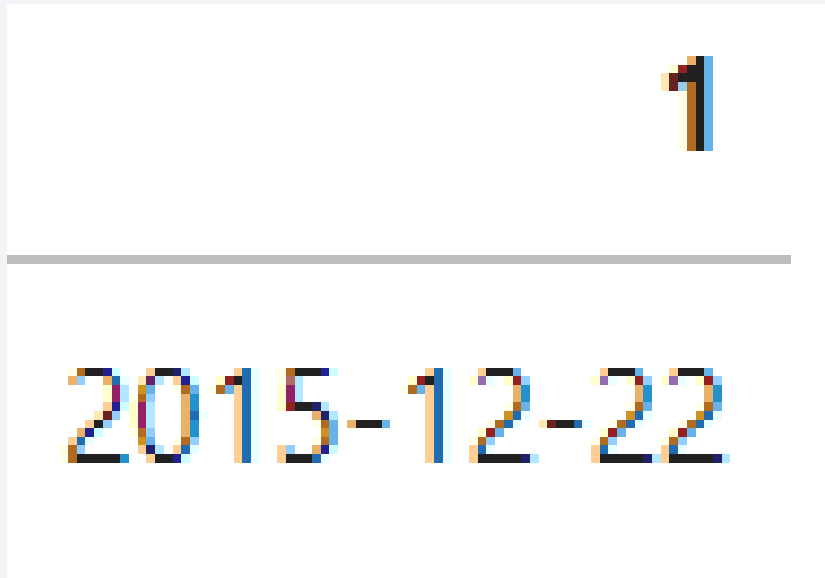
%%sql

```
SELECT
    AVG(payload_mass__kg_) AS Average_Payload_Mass_F9v11
FROM
    PZC04681.SPACEXTBL
WHERE BOOSTER_VERSION='F9 v1.1';
```

This query commands to select, from the table PZC04681.SPACEXTBL, the average (AVG) of payload_mass_kg_ column values and name it (AS) Average_Payload_Mass_F9 v1.1; But filter (WHERE) to only keep the records that values in the BOOSTER_VERSION column = "F9 v1.1".

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad



1

2015-12-22

- SQL query from the DB2 Database PZC04681

```
%%sql
SELECT
    MIN(DATE) AS first_successful_landing_outcome_in_ground_pad
FROM
    PZC04681.SPACEXTBL
WHERE
    LANDING__OUTCOME = 'Success (ground pad)';
```

This query commands to select, from the table PZC04681.SPACEXTBL, the first element (MIN) in the DATE column values and name it (AS) first_successful_landing_outcome_in_ground_pad; But filter (WHERE) to only keep the records that values in the LANDING_OUTCOME column = "Success (ground pad)".

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

booster_version

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

- SQL query from the DB2 Database PZC04681

```
%%sql
SELECT
  BOOSTER_VERSION AS Names_of_Boosters_droneship_gt_ground_4000_lt_6000
FROM
  PZC04681.SPACEXTBL
WHERE
  LANDING__OUTCOME = 'Success (drone ship)' AND (payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000);
```

This query commands to select, from the table PZC04681.SPACEXTBL, booster versions in the BOOSTER_VERSION column values and name it (AS) Names_of_Boosters_droneship_gt_ground_4000_lt_6000; But filter (WHERE) to only keep the records that values in the LANDING__OUTCOME column = "Success (drone ship)" AND payload_mass__kg_ is between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

- Total Number of Successful and Failure Mission Outcomes

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- SQL query from the DB2 Database PZC04681

```
%%sql
SELECT
    COUNT(MISSION_OUTCOME) AS total_number_of_successful_and_failure_mission_outcomes
FROM
    PZC04681.SPACEXTBL
WHERE
    LANDING__OUTCOME LIKE 'Success%'
UNION
SELECT
    COUNT(MISSION_OUTCOME) AS total_number_of_successful_and_failure_mission_outcomes
FROM
    PZC04681.SPACEXTBL
WHERE
    LANDING__OUTCOME LIKE 'Failure%'
```

This query commands to select, from the table PZC04681.SPACEXTBL, the List of numbers (COUNT) of values in the MISSION_OUTCOME column (...); But filter (WHERE) to only keep the records that values in the LANDING__OUTCOME column start with "Success" or (UNION) "Failure".

Boosters Carried Maximum Payload

- Boosters Carried Maximum Payload

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

- SQL query from the DB2 Database PZC04681

```
%%sql
SELECT
    booster_version AS booster_version_maxpayload
FROM
    PZC04681.SPACEXTBL
WHERE
    payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM PZC04681.SPACEXTBL);
```

This query commands to select, from the table PZC04681.SPACEXTBL, the booster versions from the values in the BOOSTER_VERSION column and name it (AS) booster_version_maxpayload; But filter (WHERE) to only keep the records that values in the payload_mass_kg_ column are the maximum.

2015 Launch Records

- 2015 Launch Records

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- SQL query from the DB2 Database PZC04681

```
%%sql
SELECT
    landing_outcome, booster_version, launch_site
FROM
    PZC04681.SPACEXTBL AS SX
WHERE
    LANDING__OUTCOME = 'Failure (drone ship)' AND DATE = (SELECT DATE FROM SX where LIKE "2015%") ;
```

This query commands to select, FROM the table PZC04681.SPACEXTBL, the Landing Outcomes, booster versions, launch sites from their column of the same names; But filter (WHERE) to only keep the records that values in the LANDING__OUTCOME column = "Failure (drone ship) AND the DATE starts with (LIKE ...%)" 2015.

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

- Landing Outcomes Between 2010-06-04 and 2017-03-20
- SQL query from the DB2 Database PZC04681

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

```
%%sql
SELECT
    COUNT(landing__outcome), landing__outcome from SPACEXTBL
FROM
    PZC04681.SPACEXTBL AS SX
WHERE
    DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing__outcome
ORDER BY count(landing__outcome) DESC;
```

This query commands to select, from the table PZC04681.SPACEXTBL, the COUNT of landing outcomes and the values from the so-called column; But filter (WHERE) to only keep the records that values are BETWEEN 4000 and 6000, GROUP BY Landing outcomes types in a descending order (ORDER BY) of the numbers of landing outcomes.

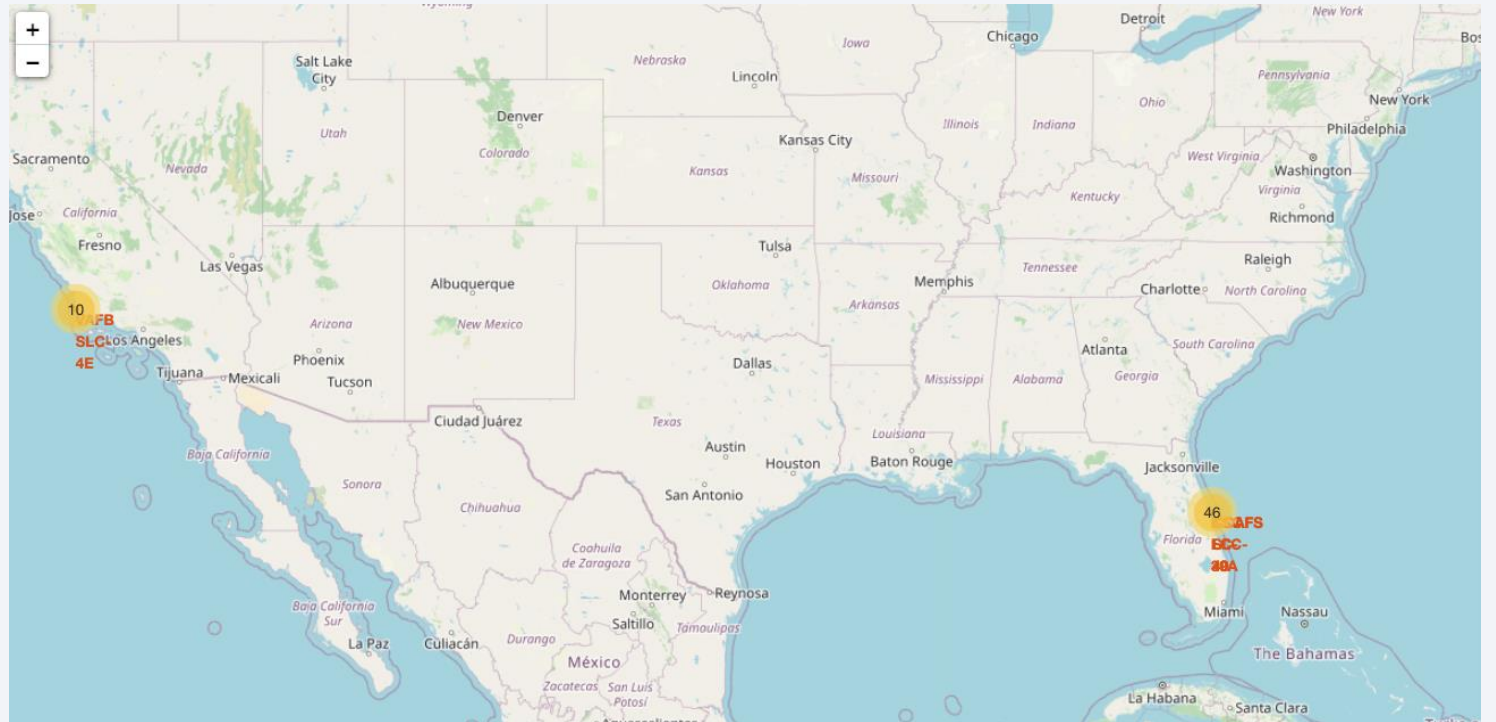
A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a solid blue background on the left and a satellite photograph of Earth on the right. The Earth's surface is dark, with numerous bright yellow and orange lights representing cities and urban areas. The horizon of the Earth is visible as a thin, curved line separating the dark surface from the deep blue of space.

Section 3

Launch Sites Proximities Analysis

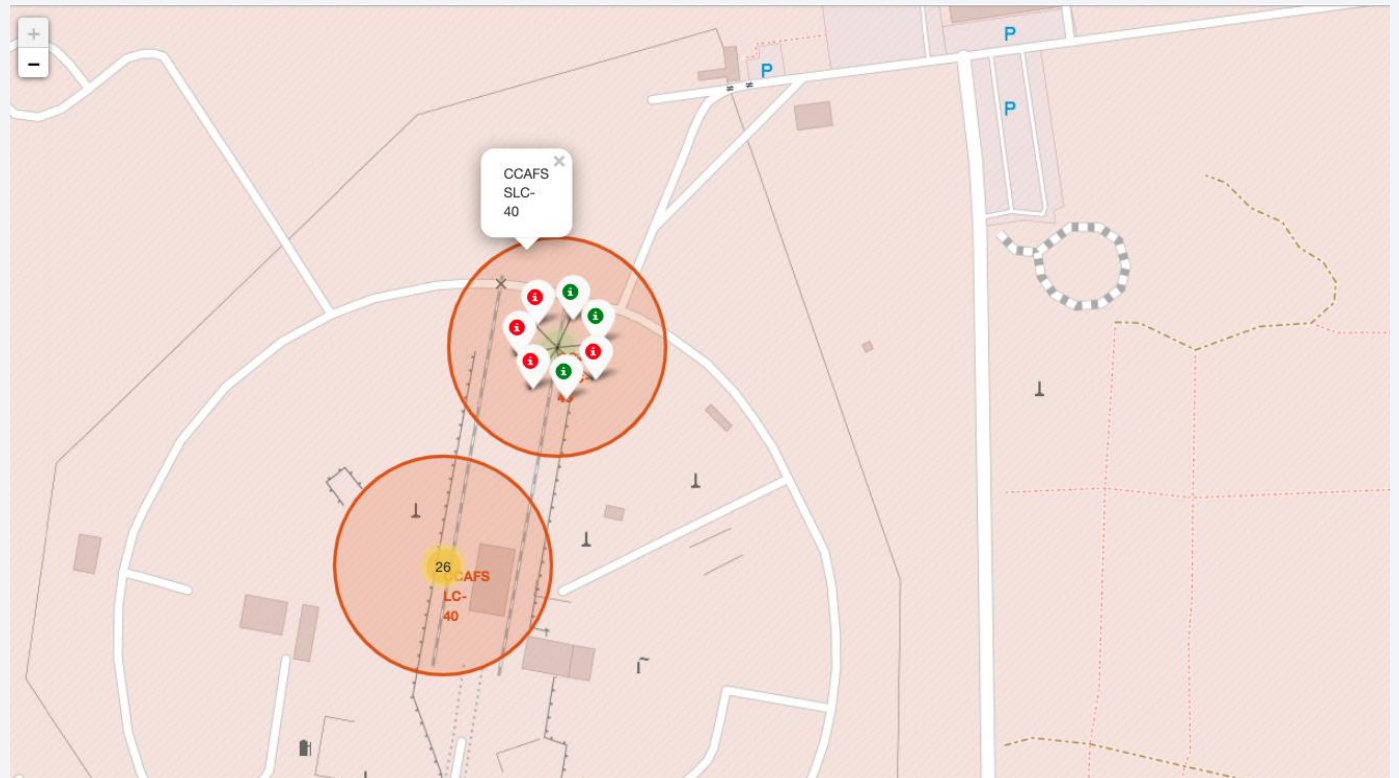
All Space X launch sites locations

- All launch sites are in very close proximity to the coasts. It lowers the risk of damage due to crashes on land.
- There are more than 4/5 of the launches made from the **East coast and closer to the Equator**. Because rockets get an additional boost from the rotational speed of the Earth.



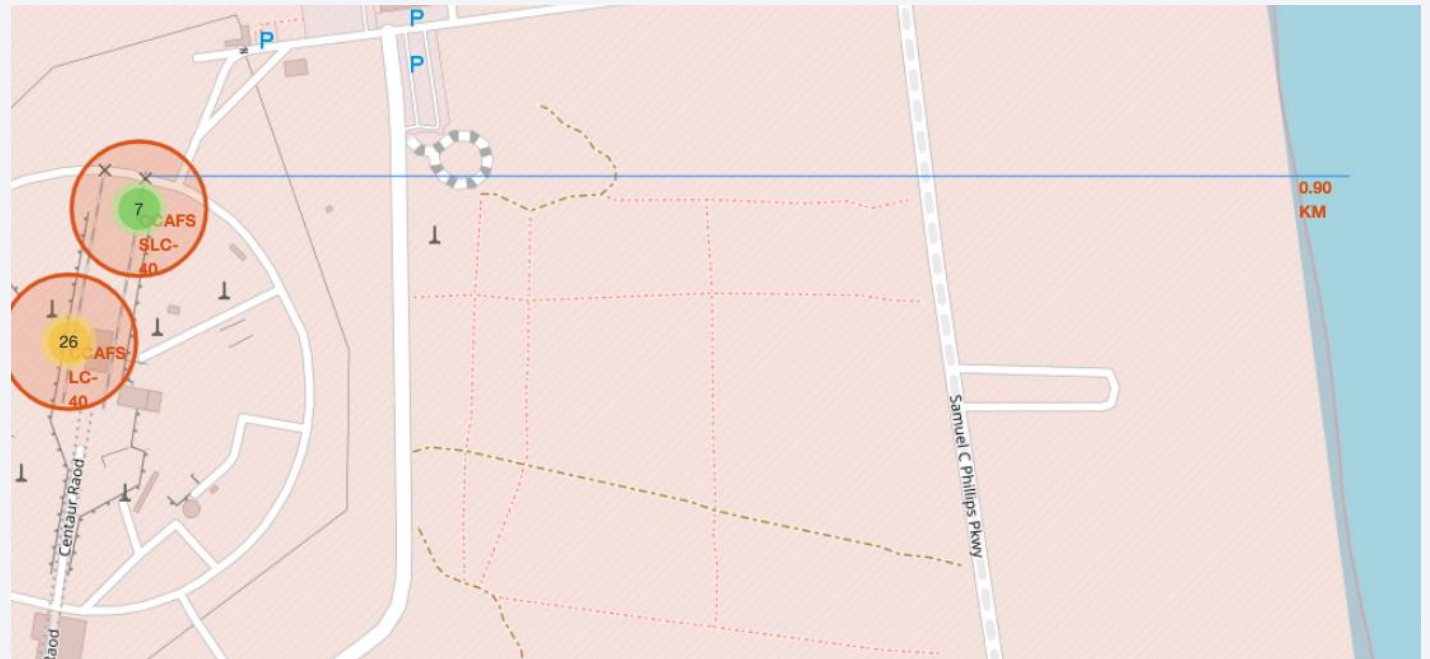
Launch Outcomes representation

- From the color-labeled markers in marker clusters, it is possible to easily identify which launch sites have relatively high/**low success rates**.
- The current site CCAFS SLC-40 has 4/7 failures as per the image.



Proximity of Launch Site

- Exploration of the **proximity** of the launch sites to landmarks like railway, highway, coastline, etc.
- The current site CCAFS SLC-40 is **just 0.90 KM away** from the coastline as per the image.
- It may give operations and logistics advantage.





Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site

Total Success Launches By Site



- This pie chart summarizes the launch success **proportions** for all sites.
- The KSC LC-39A is by far the **site with the most successful launches**.

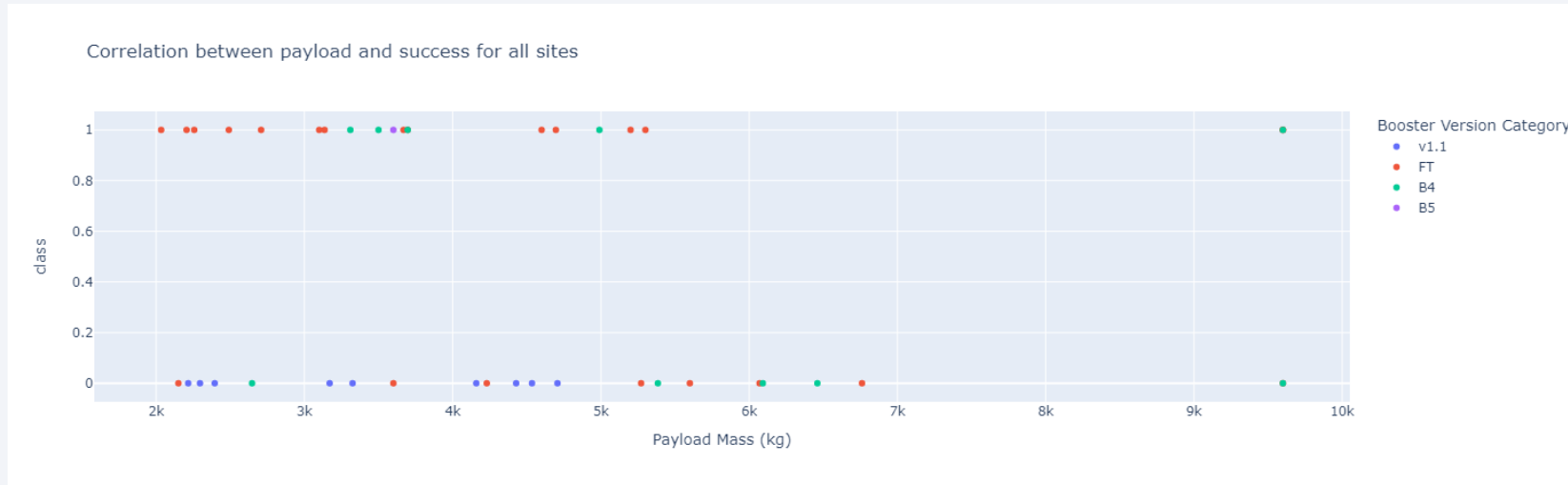
Launch site with highest launch success ratio

Successes vs. failures pie-chart for KSC LC-39A

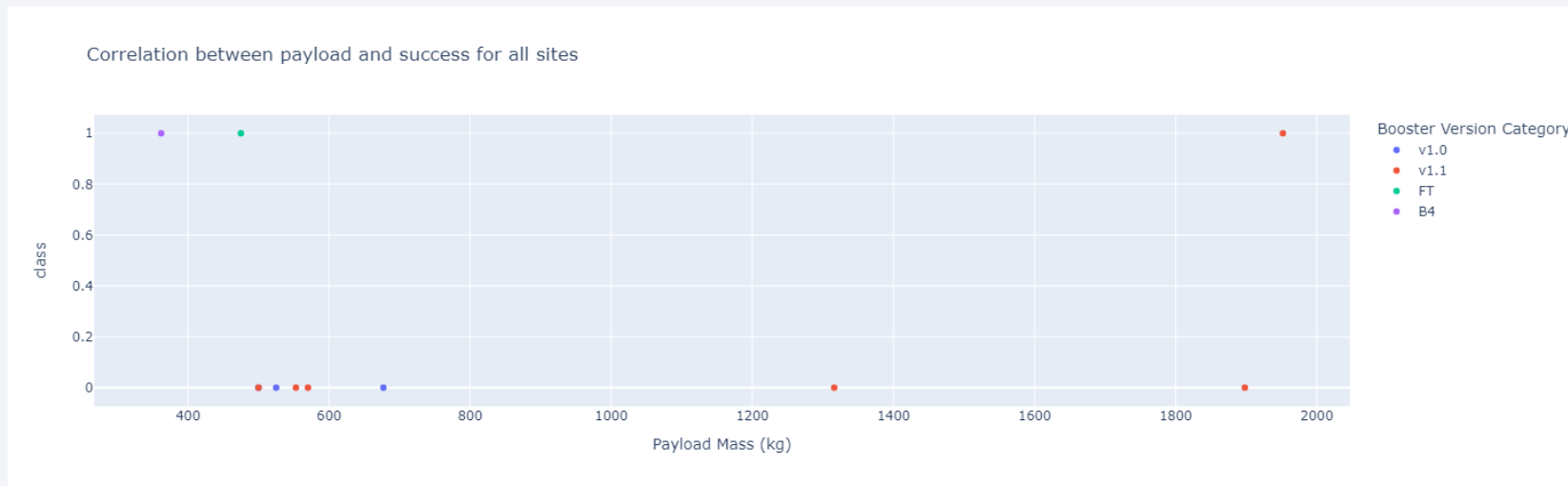


- This pie chart summarizes the **proportions** of successful vs failed launches from KSC LC-39A
- The KSC LC-39A has more than **$\frac{3}{4}$ of successful** launches.

Payload vs. Launch Outcome analysis for all sites



- In the Payload Mass (kg) range 2k – 10k, Booster Version category FT is the most successful, V.1 is unsuccessful and V1.0 is discarded.

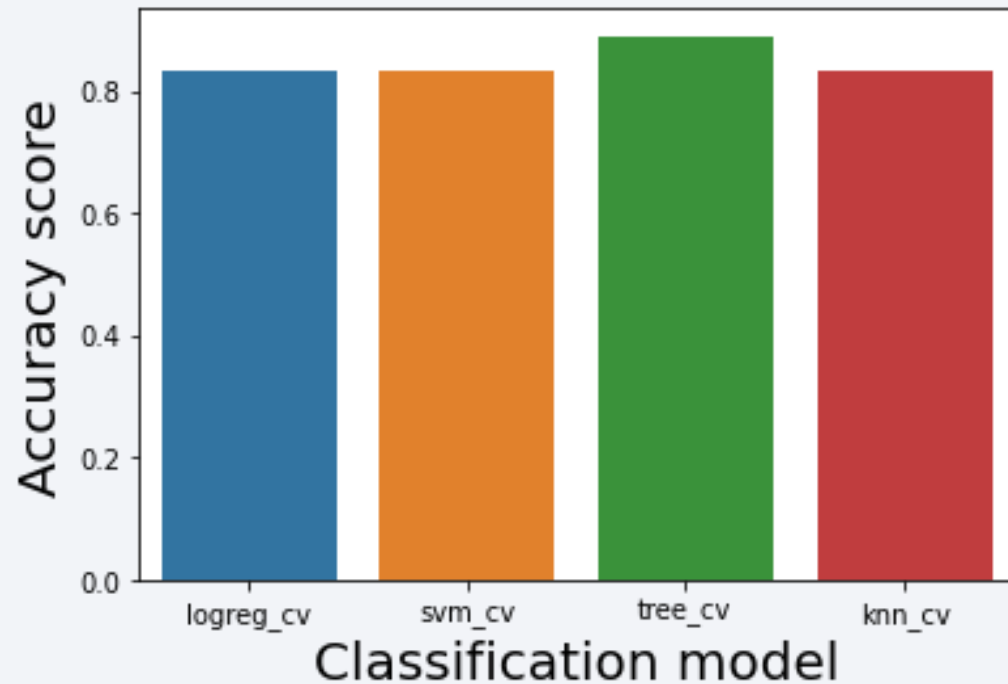


- In the Payload Mass (kg) range 0 – 2k, Booster Version category V1.x seems to yield mitigated success.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- The `tree_cv` model has the highest classification accuracy with a score of 0.88
- This indicates that the `decision tree classifier` gives the best results for our sample data.

Confusion Matrix

- Confusion matrix of the best performing model: **decision tree classifier**



- The confusion matrix measures the **performance** of the ML classification model - decision tree classifier here.

True Positive (land–landed): **12** accurately predicted successful outcomes

False Positive (land–did not land): **1** Error type 1

False Negative (didn't land–landed): **0** Error type 2

True Negative (did not land–did not land): **5** accurately predicted failure outcomes.

Only 1 Error for 17 accurate predictions

Conclusions

- Different launch sites have different success rates
- Mostly, the first stage is more likely to land successfully as the flight number increases
- More than 4/5 of the launches made from the East coast and closer to the
- 12/18 first stage successful landings are predicted.
- Our teams will be able to determine the price of each flight based on this knowledge.

Appendix

- Data set - Panda DataFrame - obtained after performing wrangling

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	
4	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...
89	86	2020-09-03	Falcon 9	15600.000000	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060	-80.603956	28.608058
90	87	2020-10-06	Falcon 9	15600.000000	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058	-80.603956	28.608058
91	88	2020-10-18	Falcon 9	15600.000000	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1051	-80.603956	28.608058
92	89	2020-10-24	Falcon 9	15600.000000	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	B1060	-80.577366	28.561857
93	90	2020-11-05	Falcon 9	3681.000000	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1062	-80.577366	28.561857

90 rows × 17 columns



Thank you!

