IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Craig Kent
09/29/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection Using SpaceX API
  - Data Collection through Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis using SQL
  - EDA DataViz using Python Pandas and Matplotlib
  - Launch Sites Analysis with Folium Interactive Visual Analytics and Ploty Dash
  - Machine Learning Landing Prediction
- Summary of all results
  - EDA
  - Interactive Visual Analytics and Dashboards
  - Predictive Analysis

Introduction

- Project background and context

  SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems you want to find answers

  In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

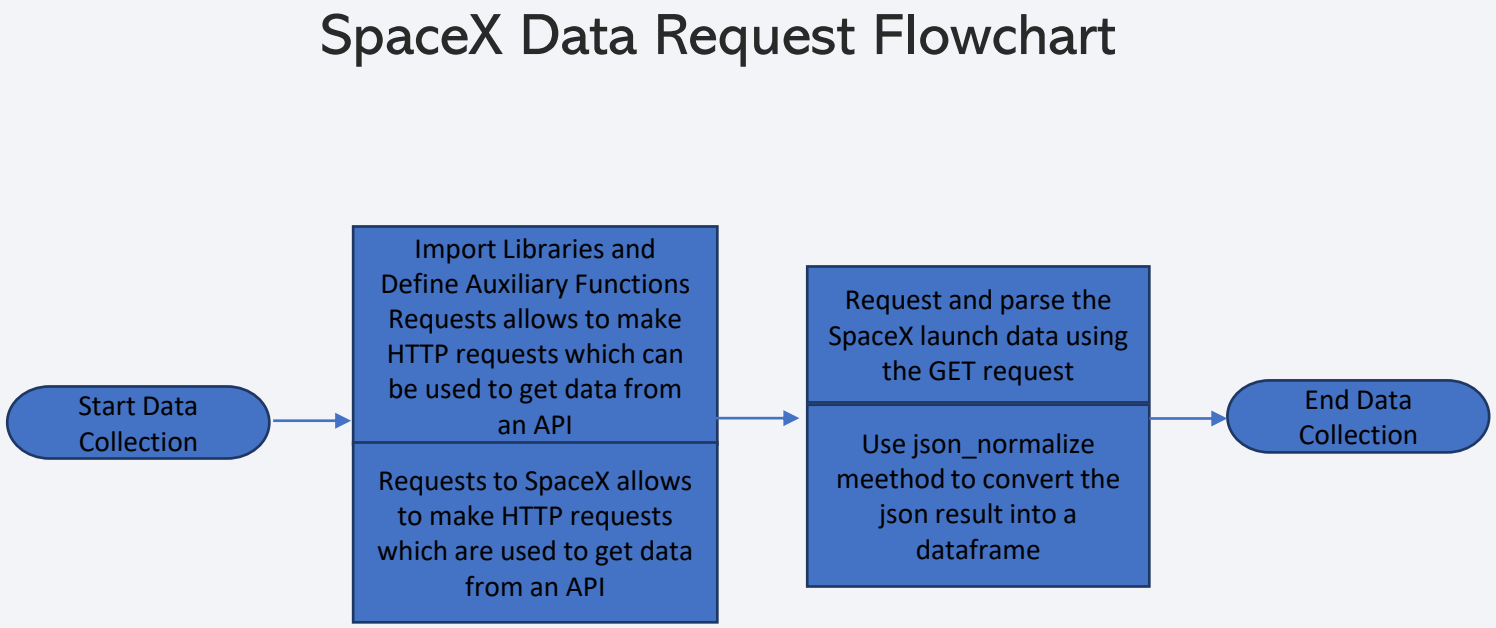
Methodology

Methodology

Executive Summary

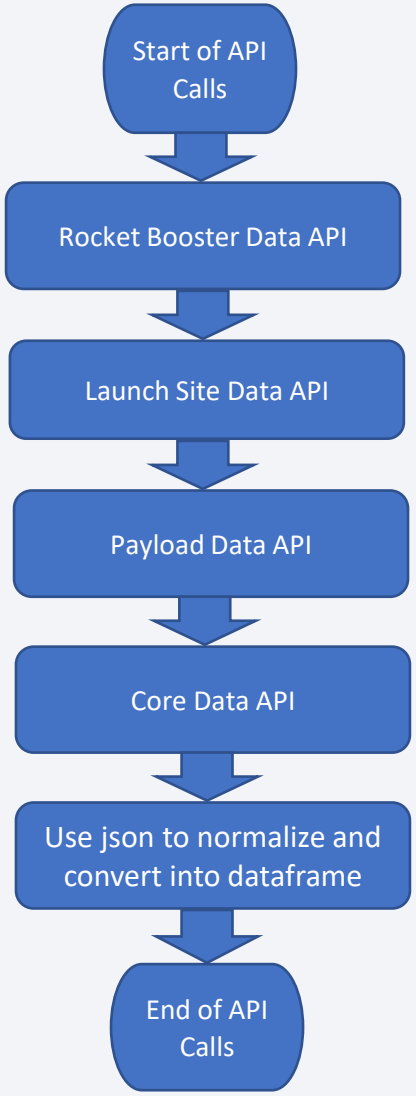
- Data collection methodology:
  - Describes how data sets were collected
- Perform data wrangling
  - Describes how data were processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection



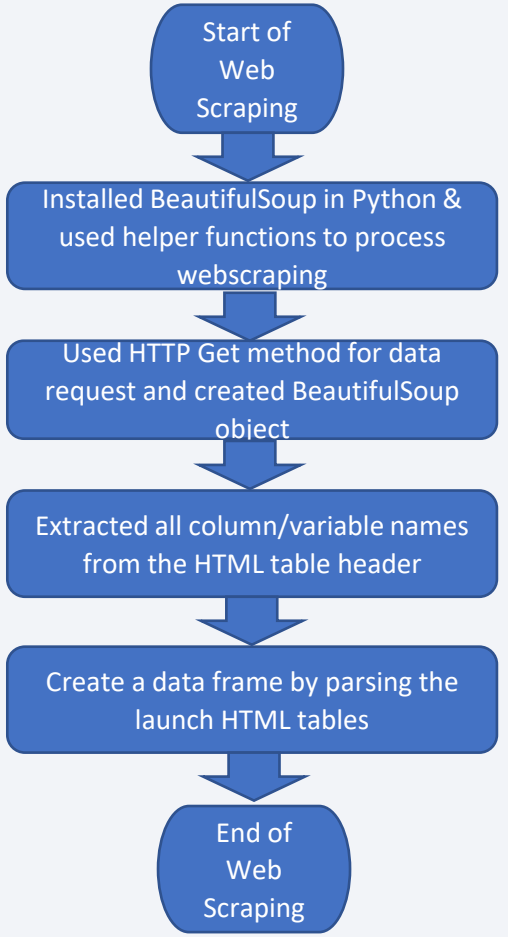
Data Collection – SpaceX API

- The flowchart to the right shows using SpaceX's RESTful API and a get request to parse the data. The from a Json dataset converting the data into a Pandas data frame.
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/ckent123/Data-Science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



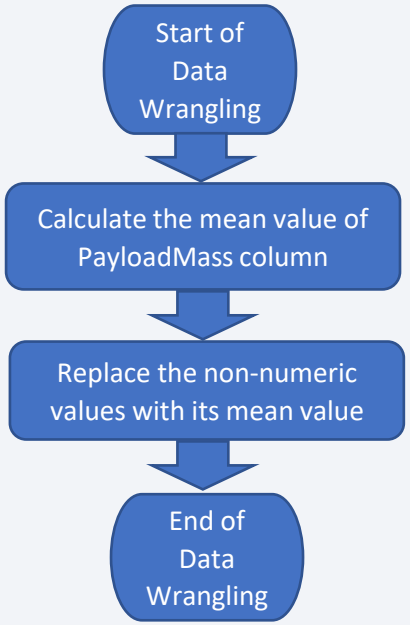
Data Collection - Scraping

- To the right is a flowchart showing Web Scraping to get data on launch records from Wikipedia
- GitHub URL of the completed web scraping notebook: https://github.com/ckent123/Data-Science/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Dealt with missing values in the LandingPad column by having none values represent when landing pads were not used. Calculated the mean for the PayloadMass using the mean() function. Used the mean and the replace() function to replace non-numeric values in the data with the mean that was calculated.
- GitHub URL of completed data wrangling related notebooks: https://github.com/ckent123/Data-Science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



EDA with Data Visualization

- Charts Used:
  - FlightNumber vs. PayloadMass – To show if payload affected the success of the launches as more launches occurred over time.
  - FlightNumber vs LaunchSite – To show if location was a factor in the success of each launch.
  - Payload Vs. Launch Site – To show if the combination of location and payload had an affect on the success of the launch.
  - Payload vs. Orbit – To show if achieving a certain orbit was affected by payload.
- GitHub URL of completed EDA with data visualization notebook: https://github.com/ckent123/Data-Science/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- SQL queries performed:
  - Displayed the names of the unique launch sites in the space mission using Distinct statement.
  - Displayed 5 records where launch sites begin with the string 'CCA' using limit clause.
  - Displayed the total payload mass carried by boosters launched by NASA (CRS) using sum function and group by clause.
  - Displayed average payload mass carried by booster version F9 v1.1 average function.
  - Listed the date when the first successful landing outcome in ground pad was achieved by using min function.
  - Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listed the total number of successful and failure mission outcomes by using trim and count functions and group by clause.
  - Listed the names of the booster_versions which have carried the maximum payload mass using a sub query.
  - Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.
- GitHub URL of completed EDA with SQL notebook: https://github.com/ckent123/Data-Science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium Map Objects:
  - Circle – to highlight specific coordinates
  - Marker – to show successful(green) and failed(red) launches by color
  - Mouse Position – to explore the maps for coordinate information
  - PolyLine – to show distances between points
- GitHub URL of completed interactive map with Folium map: https://github.com/ckent123/Data-Science/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

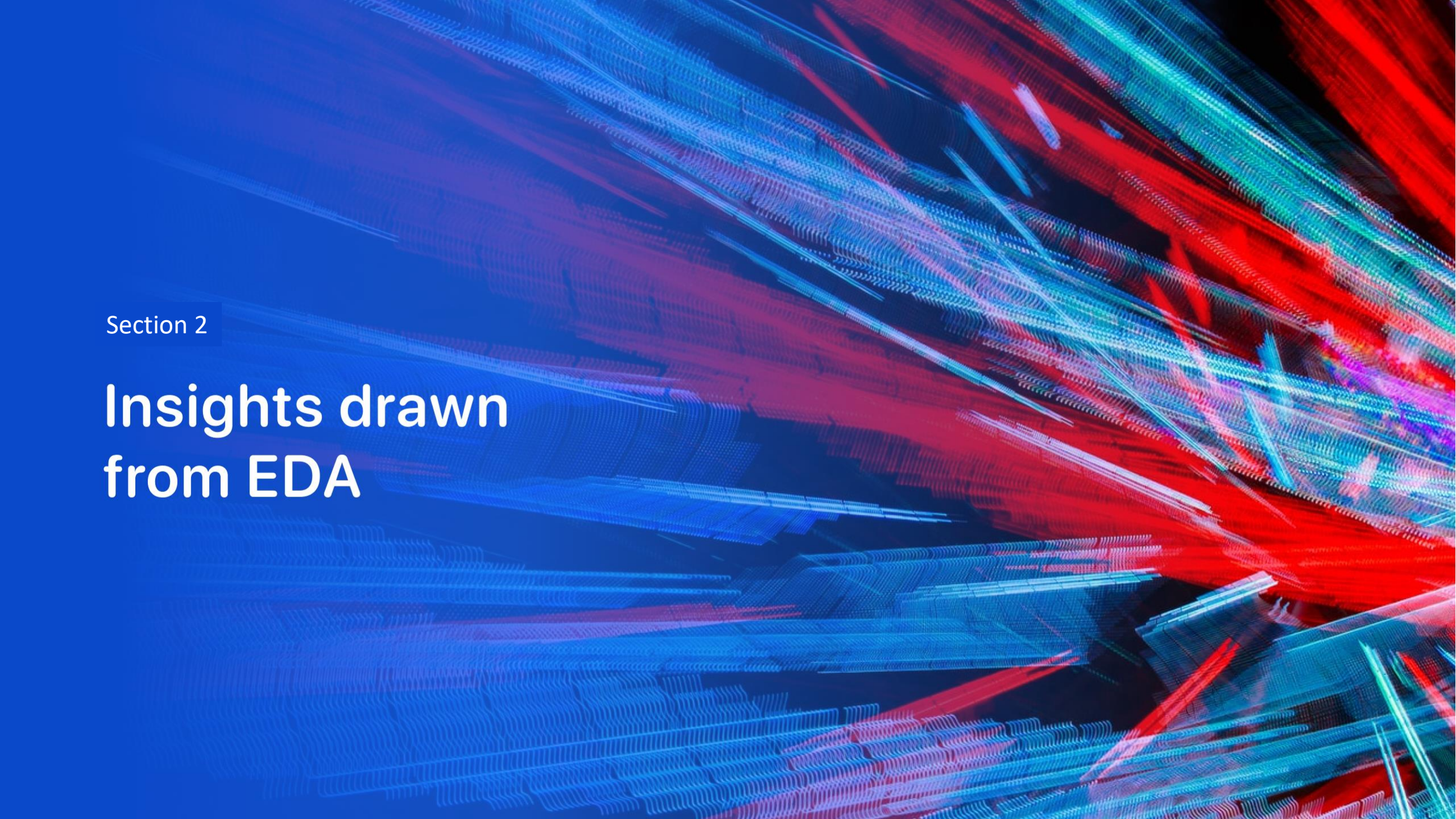
- There is a pie chart and a scatter chart:
  - The pie chart shows successful launches by each site. And has a drop down to select individual sites.
  - The scatter chart shows the payload mass vs. the success(class) of each launch. If further color codes the outcomes according to the booster version. And also lets the user choose a range of payload mass to give in site on outcomes.
- The user interactions were added to give more control and greater depth to the analytics of the chart.
- GitHub URL of completed Plotly Dash lab: https://github.com/ckent123/Data-Science/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Created column for Class (successes)
- Standardized the data
- Found best parameters for SVM, DT, KNN, and LRs.
- Used confusion matrix and accuracy scores to evaluate data.
- GitHub URL of completed predictive analysis lab: https://github.com/ckent123/Data-Science/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

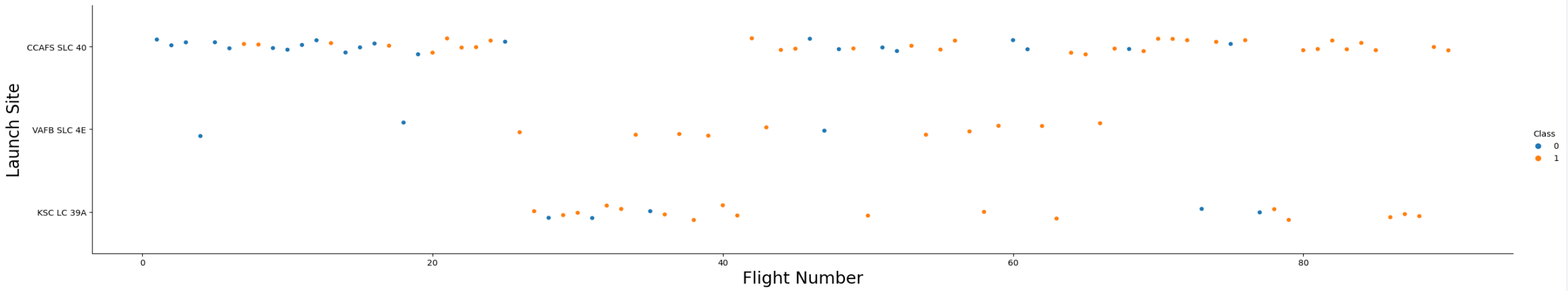
Results

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


Section 2
Insights drawn from EDA

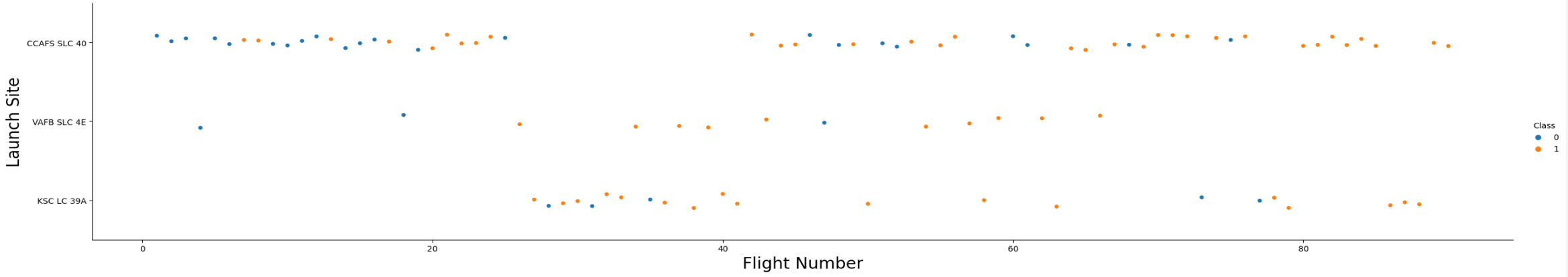
Flight Number vs. Launch Site

- Number vs. Launch Site



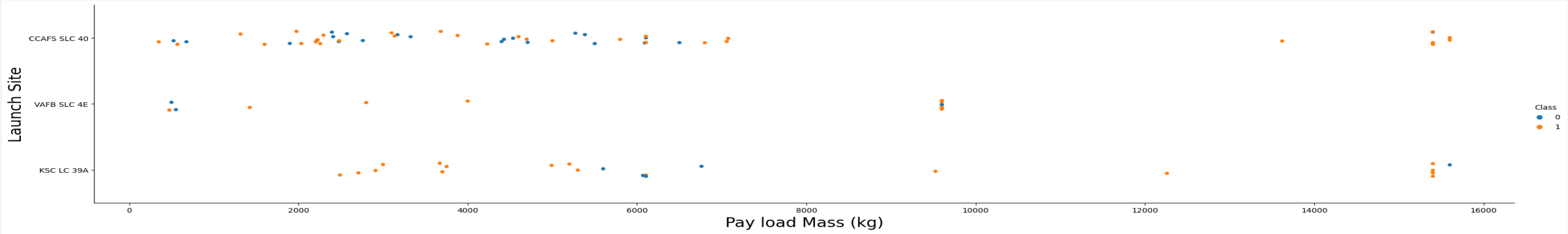
- Scatter plot with explanations

With each flight that is completed no matter what site had the launch, the success of each flight goes up.



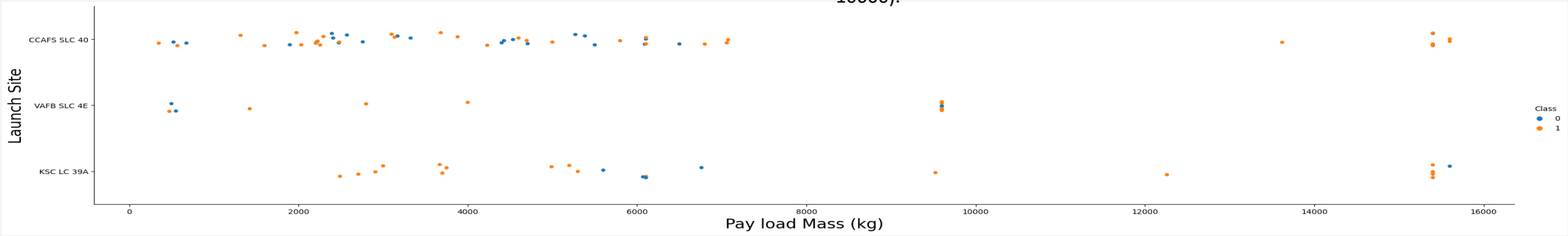
Payload vs. Launch Site

- Payload vs. Launch Site



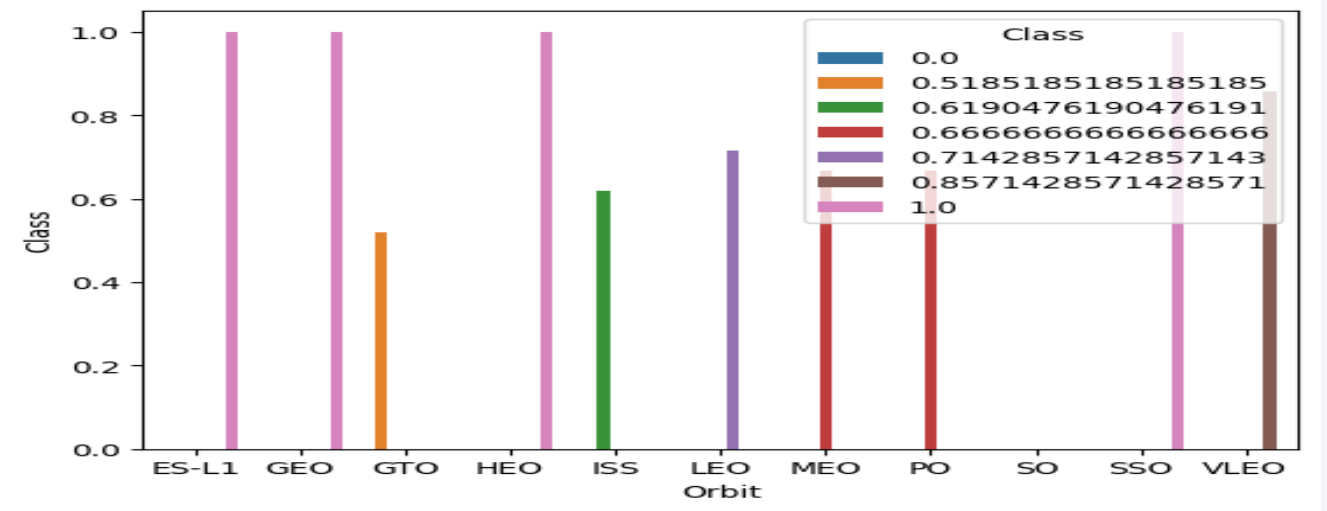
- Scatter plot with explanations

If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).



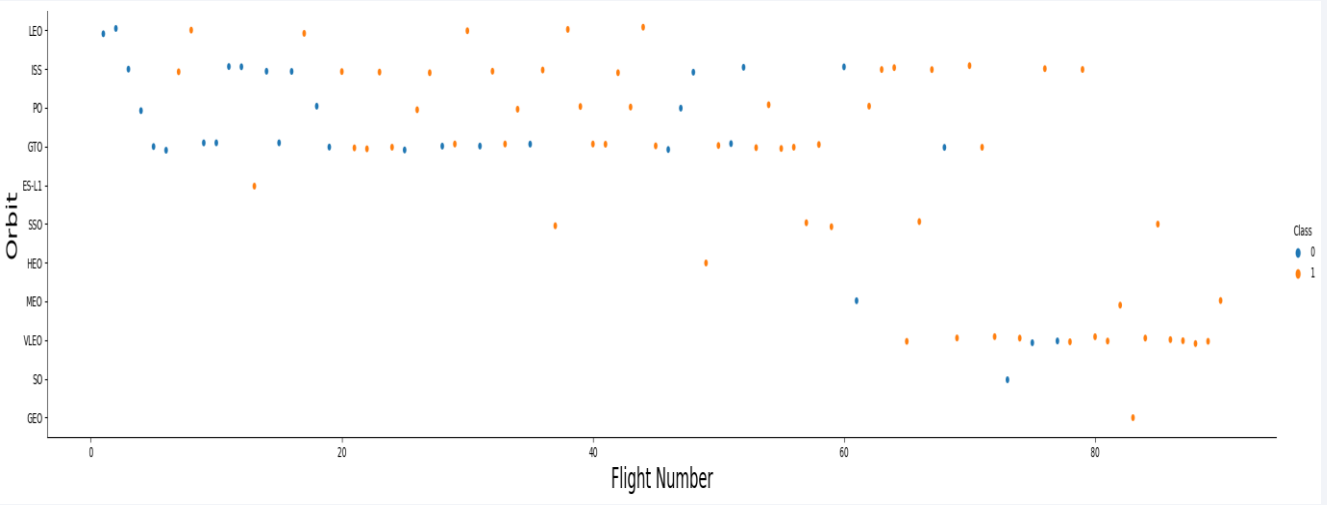
Success Rate vs. Orbit Type

- Success rate of each orbit type



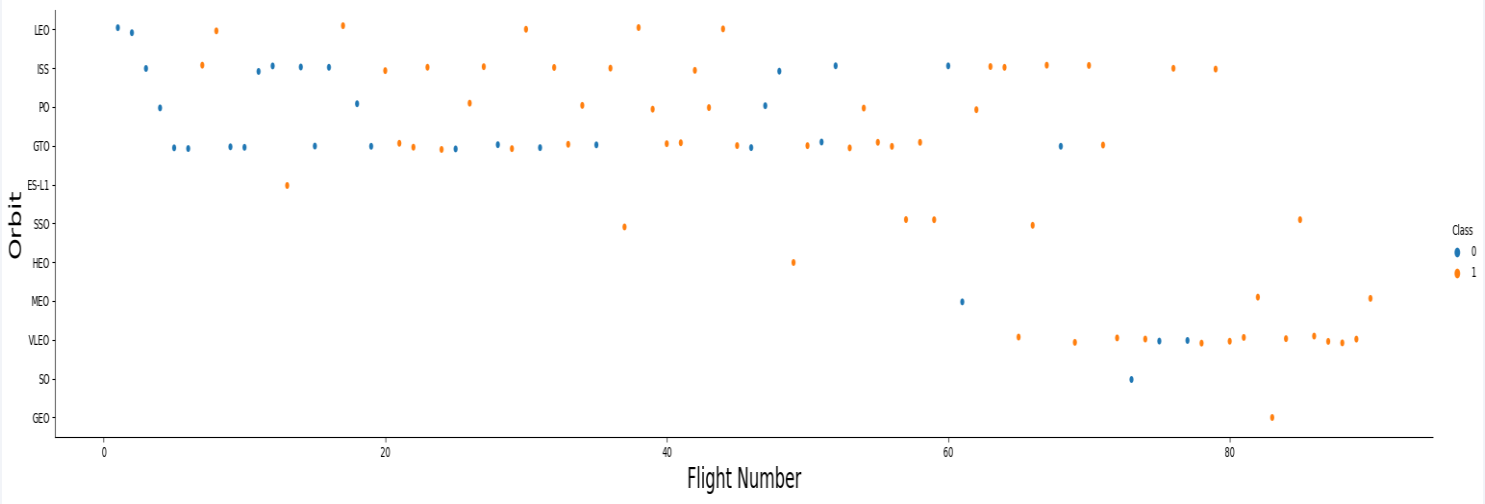
- Scatter plot with explanations

You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



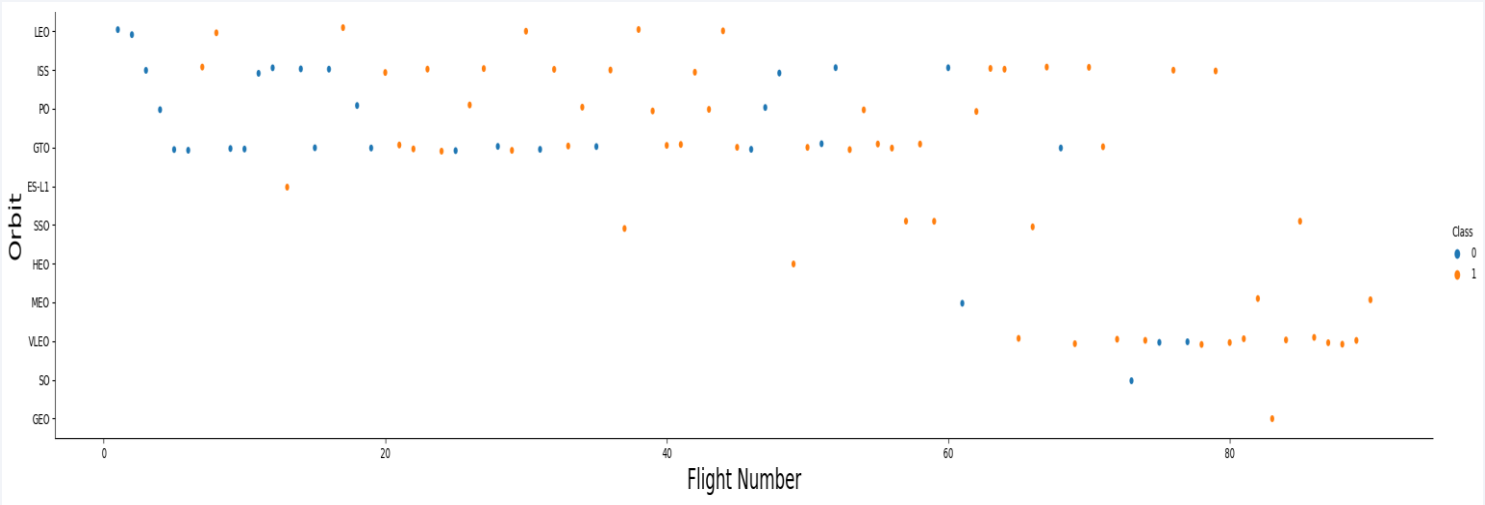
Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type



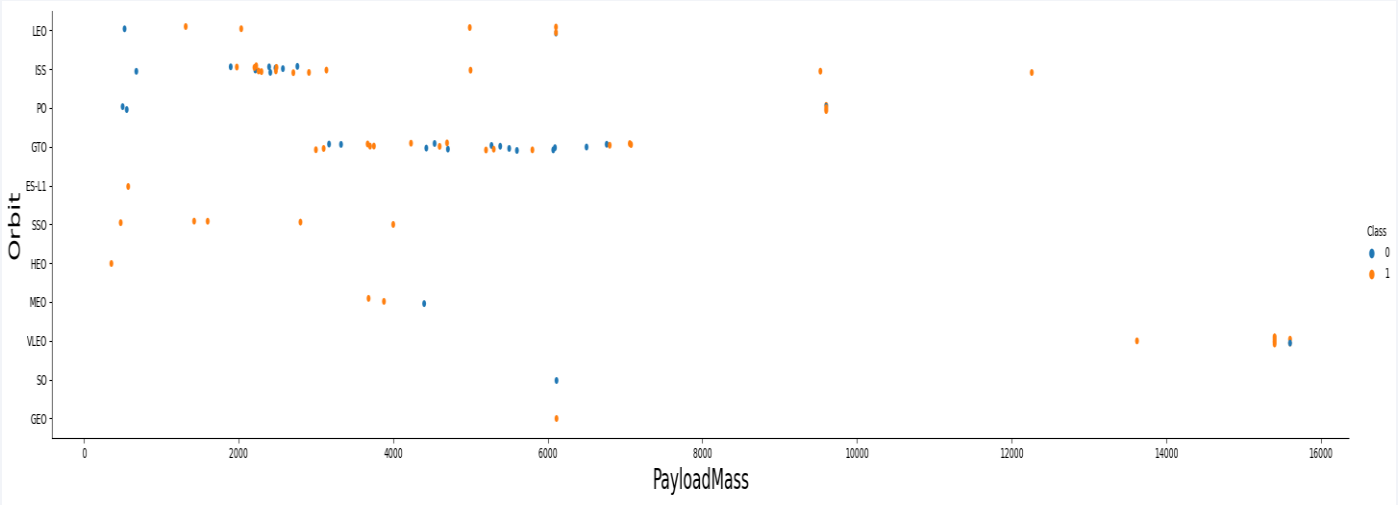
- Scatter plot with explanations

You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



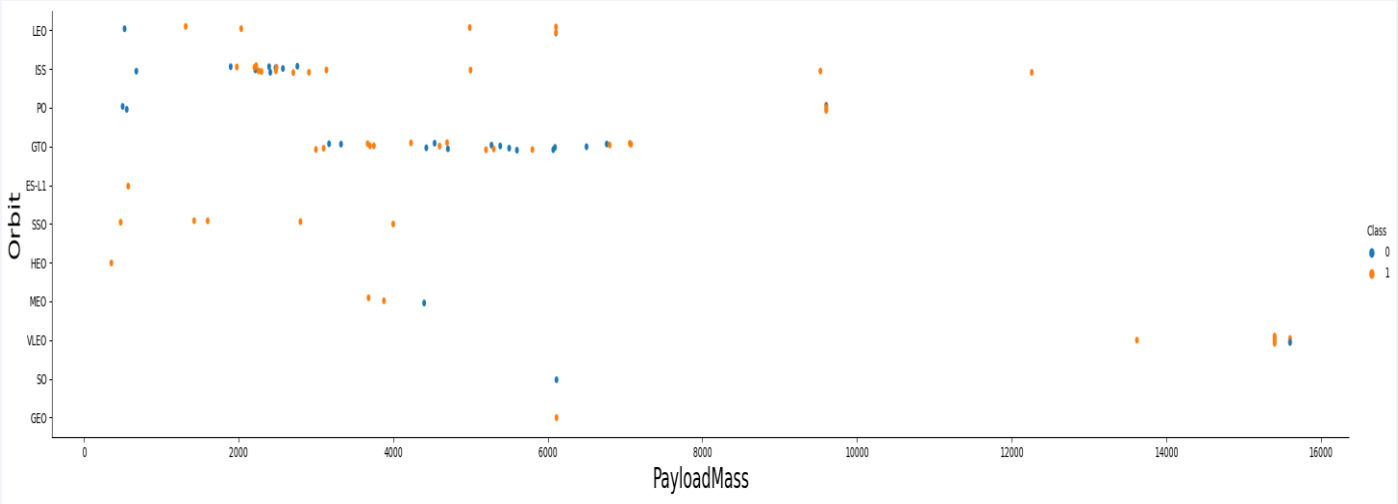
Payload vs. Orbit Type

- Scatter point of payload vs. orbit type



- Scatter plot with explanations

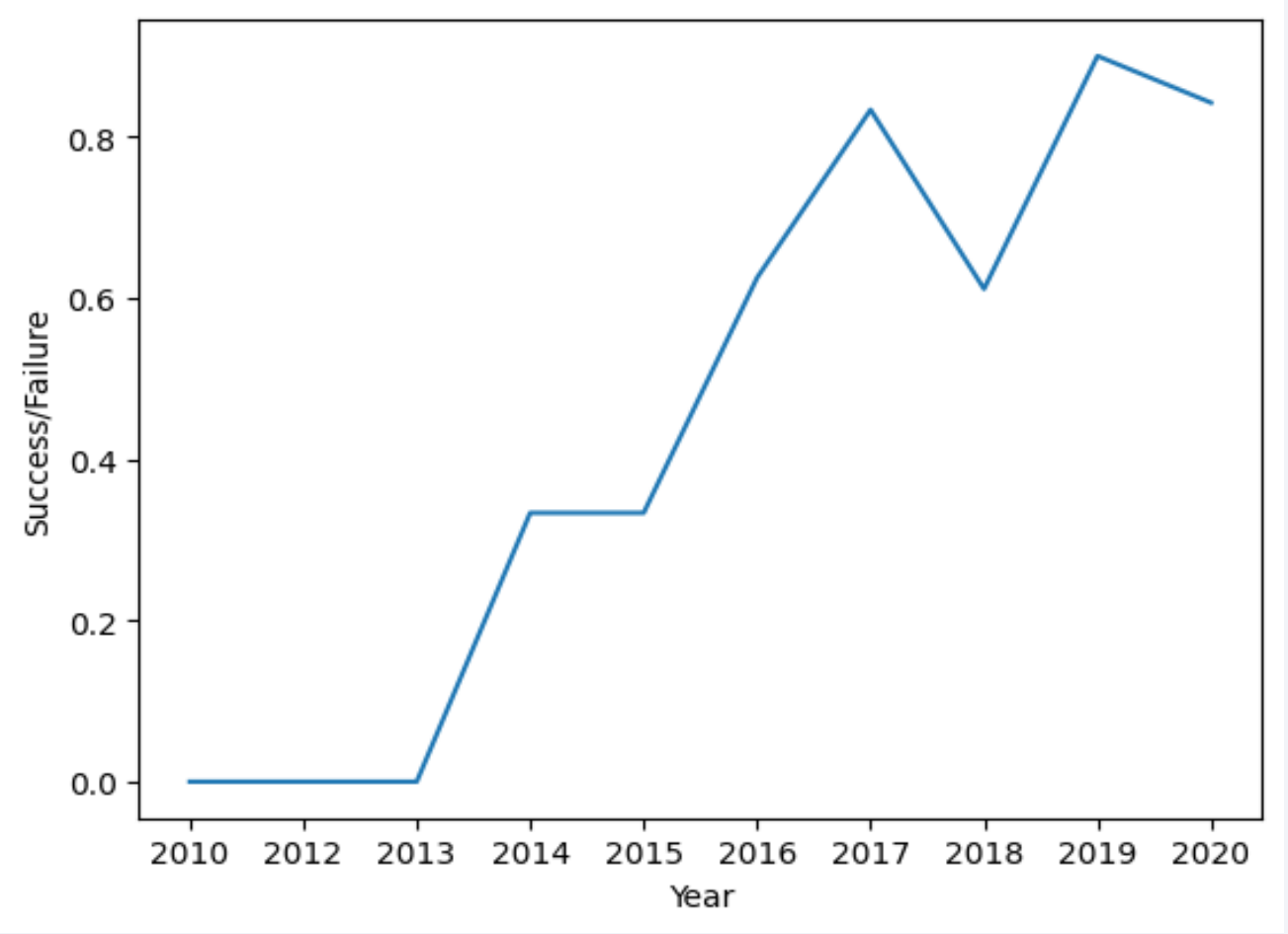
The successful landings are showing in the orbits of ISS, LEO and Polar. And the data shows there isn't a statistical difference with the GTO orbit. There isn't a lot of data on the other orbits.



Launch Success Yearly Trend

- Line chart of yearly average success rate

The line chart shows an increasing success rate starting in 2013 through 2020.



All Launch Site Names

- Find the names of the unique launch sites

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

- The query uses a distinct statement on the Launch_site field to show only the unique data for each site.

%sql select distinct Launch_Site from spacextable

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Used select * to show all fields in record. And used the “like” wildcard operator to show only records that begin with “CCA”.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

%sql select * from spacextable where Launch_Site like 'CCA%' limit 5

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql select customer, sum(PAYLOAD_MASS__KG_) as total_payload from spacextable where customer = 'NASA (CRS)' group by 1
```

- Present your query result with a short explanation here

Customer	total_payload
NASA (CRS)	45596

Filtered for only NASA boosters then used the sum function on the “Payload_Mass__KG_” field to get total payload. It wasn’t needed but there is a group by on the “customer” field. This could help for totaling more than one customer.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%sql select Booster_Version, avg(PAYLOAD_MASS__KG_) as average_payload from spacextable where Booster_Version = 'F9 v1.1' group by 1
```

- Present your query result with a short explanation here

Booster_Version	average_payload
F9 v1.1	2928.4

Filtered for only 'F9 v1.1' boosters then used the “avg” function on the “Payload_Mass__KG_” field to get total payload. It wasn’t needed but there is a group by on the “Booster_Version” field. This could help for totaling more than one Booster.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

date
2015-12-22

- Present your query result with a short explanation here

```
%sql select min(date) as date from spacextable where landing_outcome = 'Success (ground pad)'
```

Filtered for only 'Success (ground pad)' from landing_outcome then used the "min" function on the "date" field to get the first date.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the 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

- Present your query result with a short explanation here

%sql select booster_version from spacextable where landing_outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000

Filtered for 'Success (drone ship)' from landing_outcome and used the between clause for “PAYLOAD_MASS__KG_ ”.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql select trim(mission_outcome) as mission_outcome, count(mission_outcome) as outcome_count from spacextable group by 1
```

- Present your query result with a short explanation here

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

Used the trim function to help with grouping. Aliased “mission_outcome” field for presentation purpose.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

Used a sub query to capture maximum payload. Then joined that sub query to the "spacxtable" table to use as criteria for max payload for each booster version

- Present your query result with a short explanation here

```
%sql select a.booster_version\
from   spacextable a,\
      (select max(PAYLOAD_MASS__KG_) as max_payload\
       from spacextable) b\
where a.PAYLOAD_MASS__KG_ = b.max_payload\
order by 1
```

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Present your query result with a short explanation here

Used the “substr” function to get the month number. Used the between operator for the date range. Used the “like” wildcard for landing outcomes

```
%sql select substr(date,6,2) as month,\
Landing_Outcome,\
Booster_Version,\
Launch_Site\
from spacextable\
where date between '2015-01-01' and '2015-12-31'\
and Landing_Outcome like '%drone ship%'\
and Landing_Outcome like '%Failure%'
```

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- Used trim function to help with grouping. Used the count statement for landing outcome. Used the between function for date range. Used the order clause for descending order.

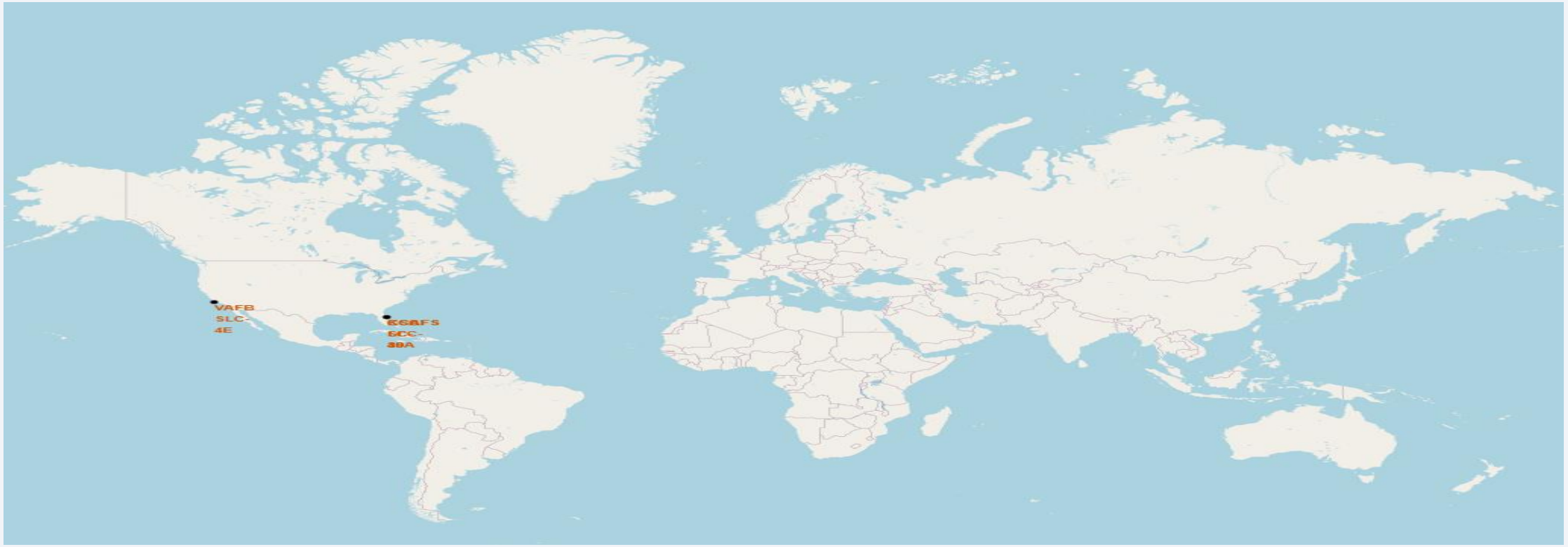
landing_outcome	outcome_counts
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

```
%sql select trim(a.landing_outcome) as landing_outcome,\
    count(a.landing_outcome) outcome_counts\
from spacextable a\
where date between '2010-06-04' and '2017-03-20'\
group by 1\
order by 2 desc
```

Section 3

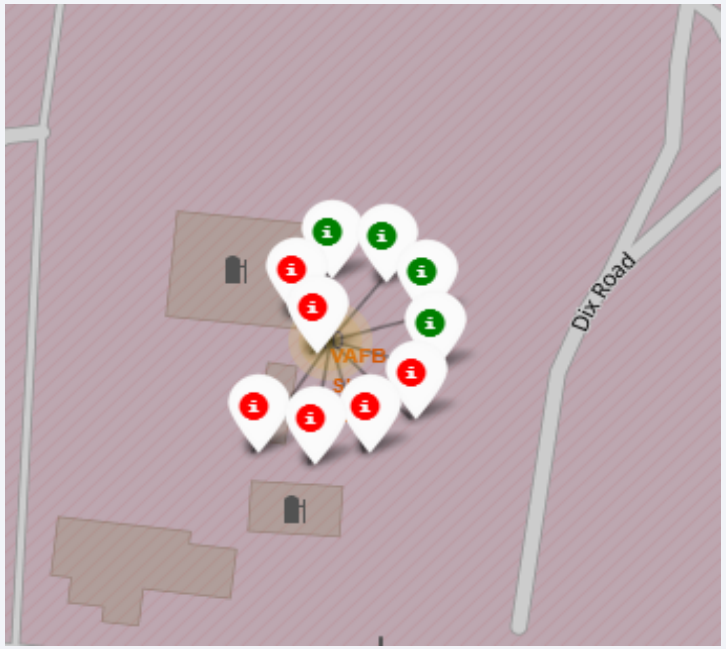
Launch Sites Proximities Analysis

SpaceX Launch Sites

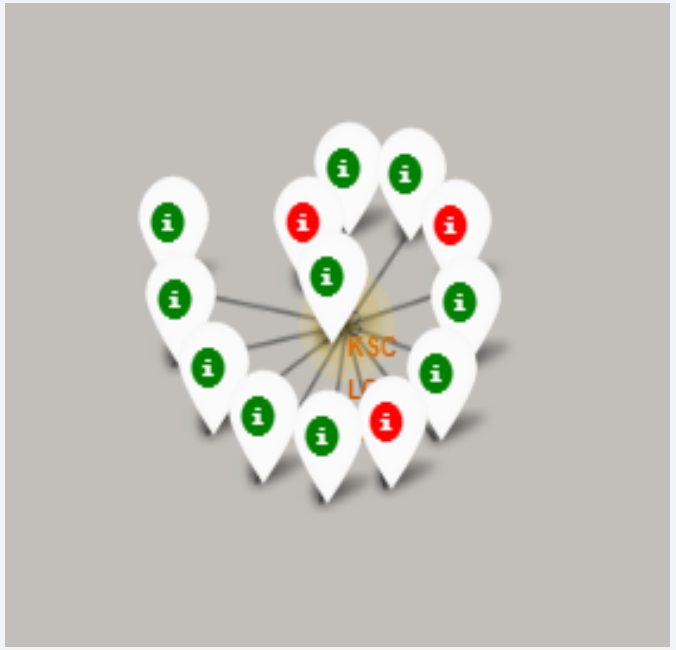


Launch Sites are all in the US. There are three, two in Florida and one in California. All are on the coast.

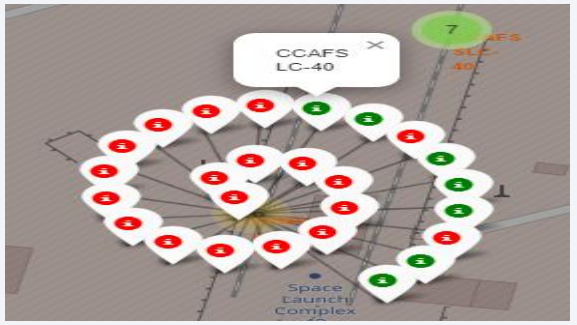
Maps of Launch Sites Showing Results



VAFB
California



KSC
Florida



CCAFS LC & SLC
Florida

Green = Success

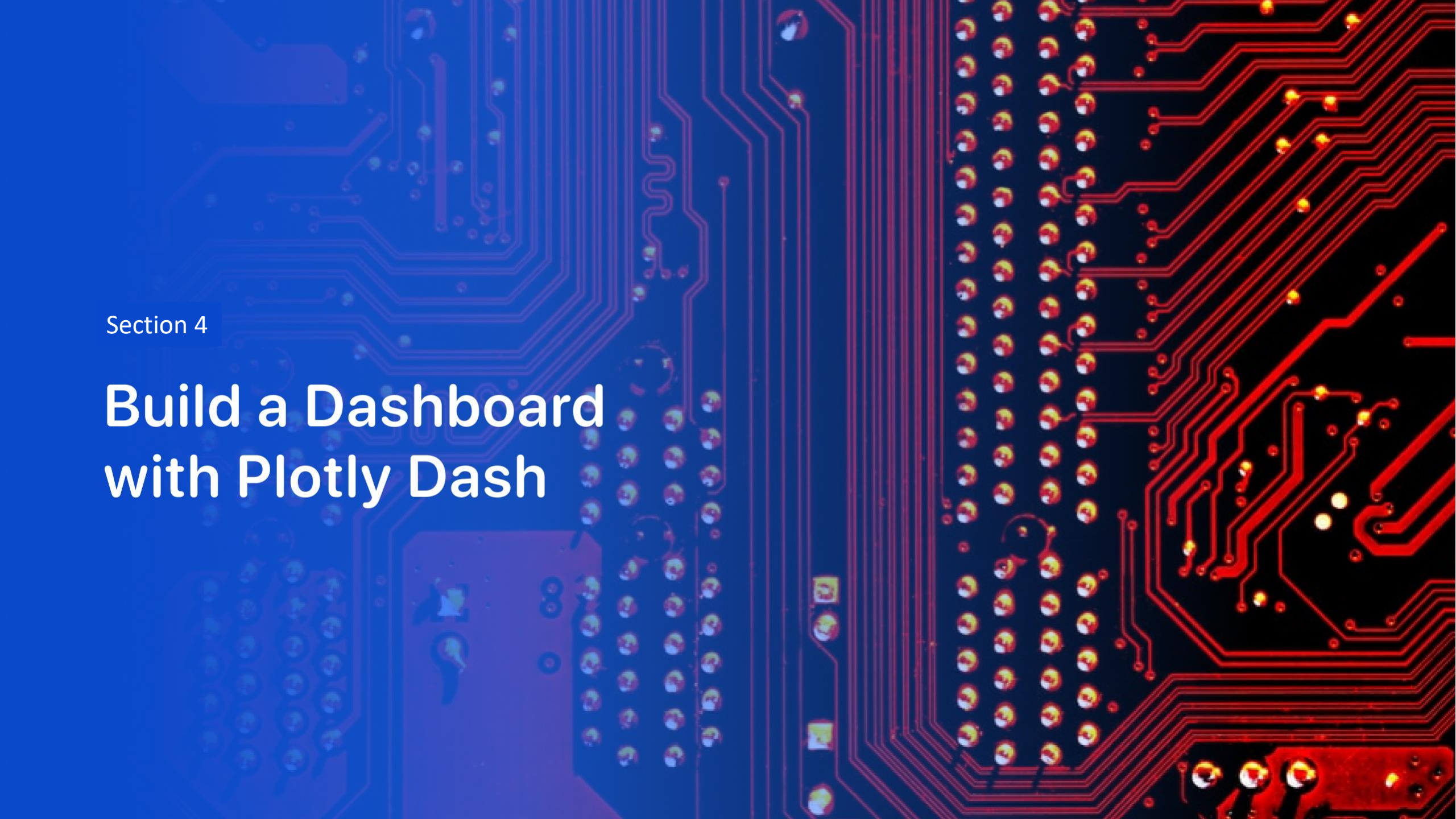
Red = Failed

- KSC site shows the most successes. But the flight numbers should also be considered.

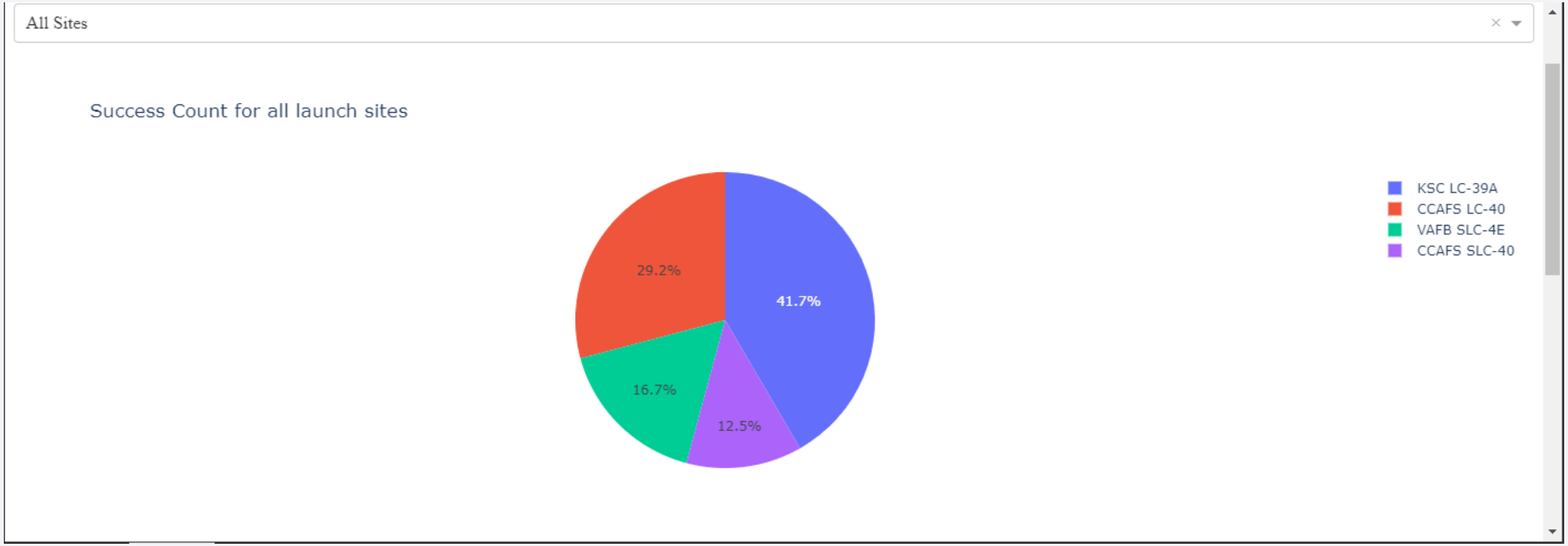
Most Successful Launch Site: KSC LC-39-A

- KSC LC-39-A is approximately 6.63km from the coast. And approximately 0.69km from the train tracks.

- It seems that this launch site may be easily accessible, but yet relatively close to the coastline.

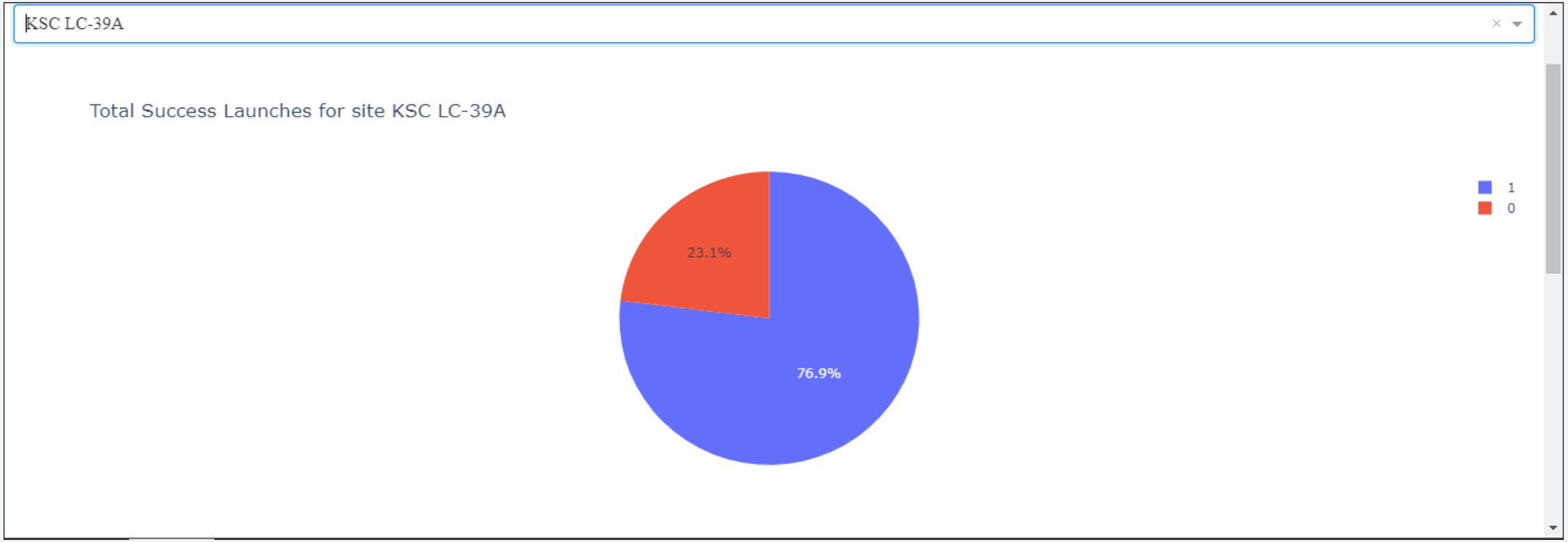

Section 4
Build a Dashboard with Plotly Dash

Launch Site Successes



- The KSC LC-39A has the most success out of all the lunch sites as shown in the pie chart above. Although the flight number should be considered as well.

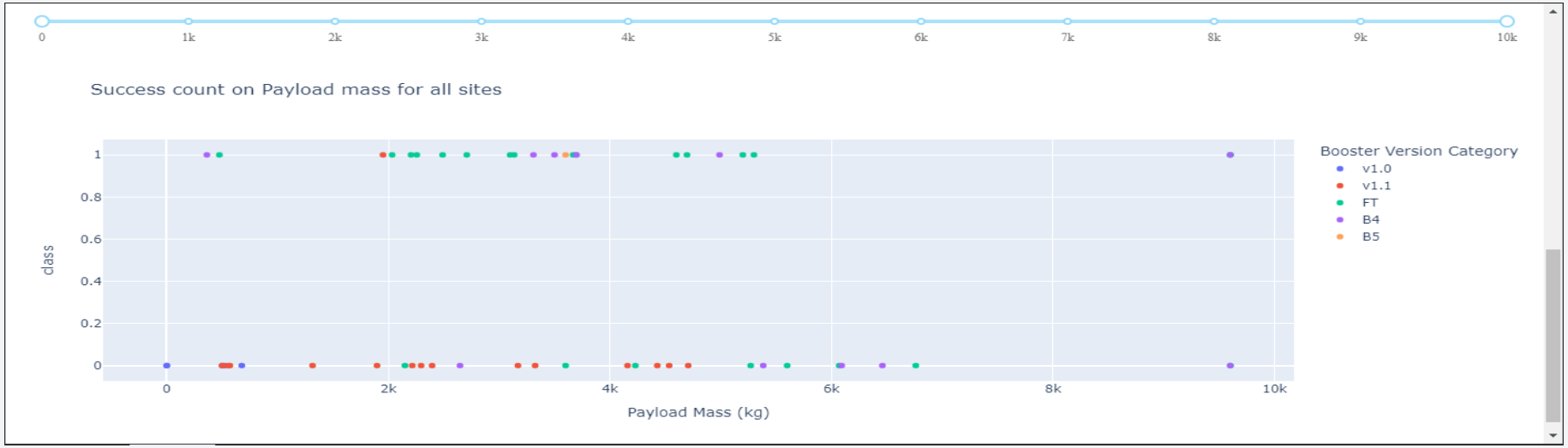
Most Successful Launch Site Chart

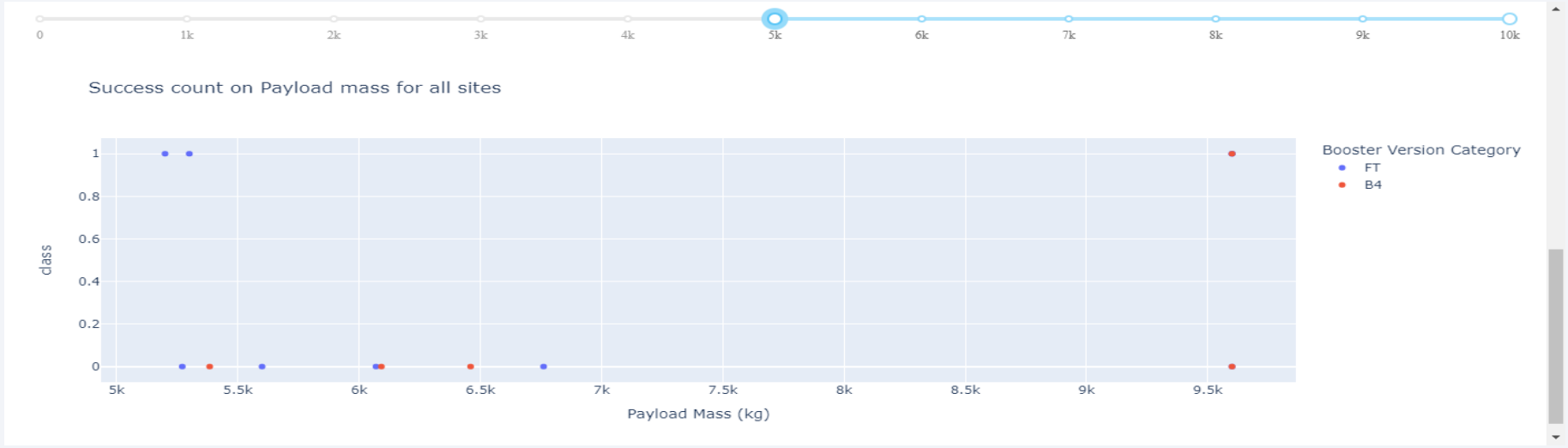


- It shows that just over ¾ of the time lunches are successful at this site.

Payload vs. Launch Outcome

- The first plot graph shows that most of the successful lunches were with the FT booster. The second graph shows that when the payload is over 5000Kg that the FT and B4 boosters where the only ones that could lift the payload and that the success ratio dropped.





Section 5

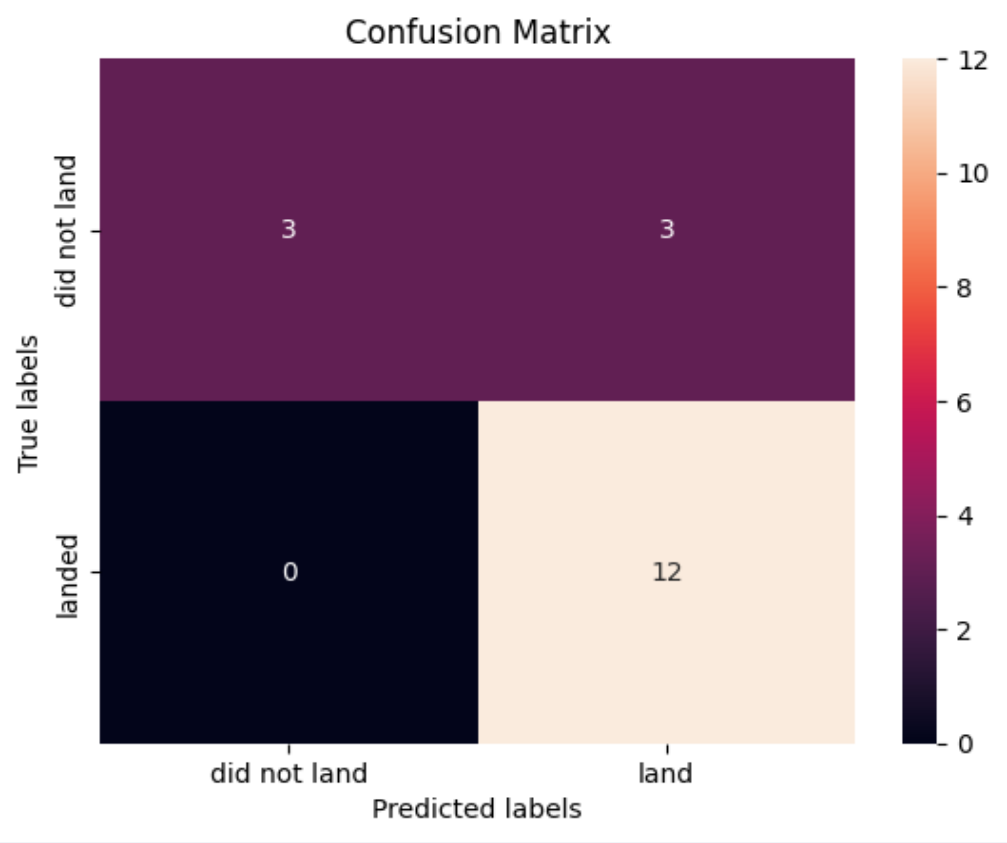
Predictive Analysis (Classification)

Classification Accuracy

	0
Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

As shown in the table above, all methods / Classification performed equally.

Confusion Matrix



- There was false positives but the confusion matrix is the same for all four classifications.

Conclusions

- The most overall conclusion that can be made is the most obvious one. The success rate of the launches and landings have increased from 2013 through 2020.
- The lighter the payload the more successful the launches.
- The launches that strived for certain orbits were more successful. (SSO, ES-L1, GEO & HEO)
- All the predictive models seemed to have the same degree of accuracy.

Appendix

- GitHub link: https://github.com/ckent123/Data-Science/blob/main/ds_capstone_coursera.pdf

Thank you!