IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dipti Kale
05-May-2023

Outline

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

Executive Summary

Summary of methodologies

- The SpaceX REST API was used to collect launch data, including rocket, payload, launch, and landing information, to predict if a rocket will land successfully.
- The data wrangling was done, to clean and transform data in order to use for the exploratory data analysis (EDA).
- EDA was performed using visualization techniques and SQL to understand the SpaceX dataset and prepare for feature engineering.
- Interactive visual analytics using Folium were used to mark launch sites and outcomes on a map, while Plotly Dash visual dashboard helped explore correlations between variables affecting launch costs.
- Predictive analysis was performed using classification models .

Summary of all results

- Overall, these observations suggest that the launch site, orbit type, payload mass can all affect the success rate of rocket launches, thus can be utilized as parameters to predict the launch outcomes.

Introduction

Project background and context

- The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets.
- In the decades since the first rockets flew, the only launch vehicle that was capable of any kind of reuse was the Space Shuttle. And while the orbiter and the solid rocket boosters were recovered after every flight, in-depth inspection and refurbishment was required after each flight. For these reasons, the Space Shuttle cost far more to fly than an equivalent non-reusable rocket. With a total program cost of $196 billion and 135 launches, the Shuttle cost almost $1.5 billion per launch. (It was initially expected to cost about $54 million per launch, adjusted for the USD in 2011.)

Problems we want to find answers

- Determine the price of each launch
- Determine if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API was used to collect launch data, including rocket, payload, launch, and landing information, to predict if a rocket will land successfully.
  - The endpoint we will use is api.spacexdata.com/v4/launches/past.
  - Web scraping technique was used to collect data from Wikipedia page.
- Perform data wrangling
  - Data wrangling was done, specifically the "Outcome" column, to convert landing outcomes to classes (0 for bad outcome, 1 for good outcome) and represent them with the variable "Y".
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Objective of exploratory data analysis using visualization and SQL was to understand SpaceX dataset and preparing data feature engineering.

Methodology - Continued

Executive Summary - Continued

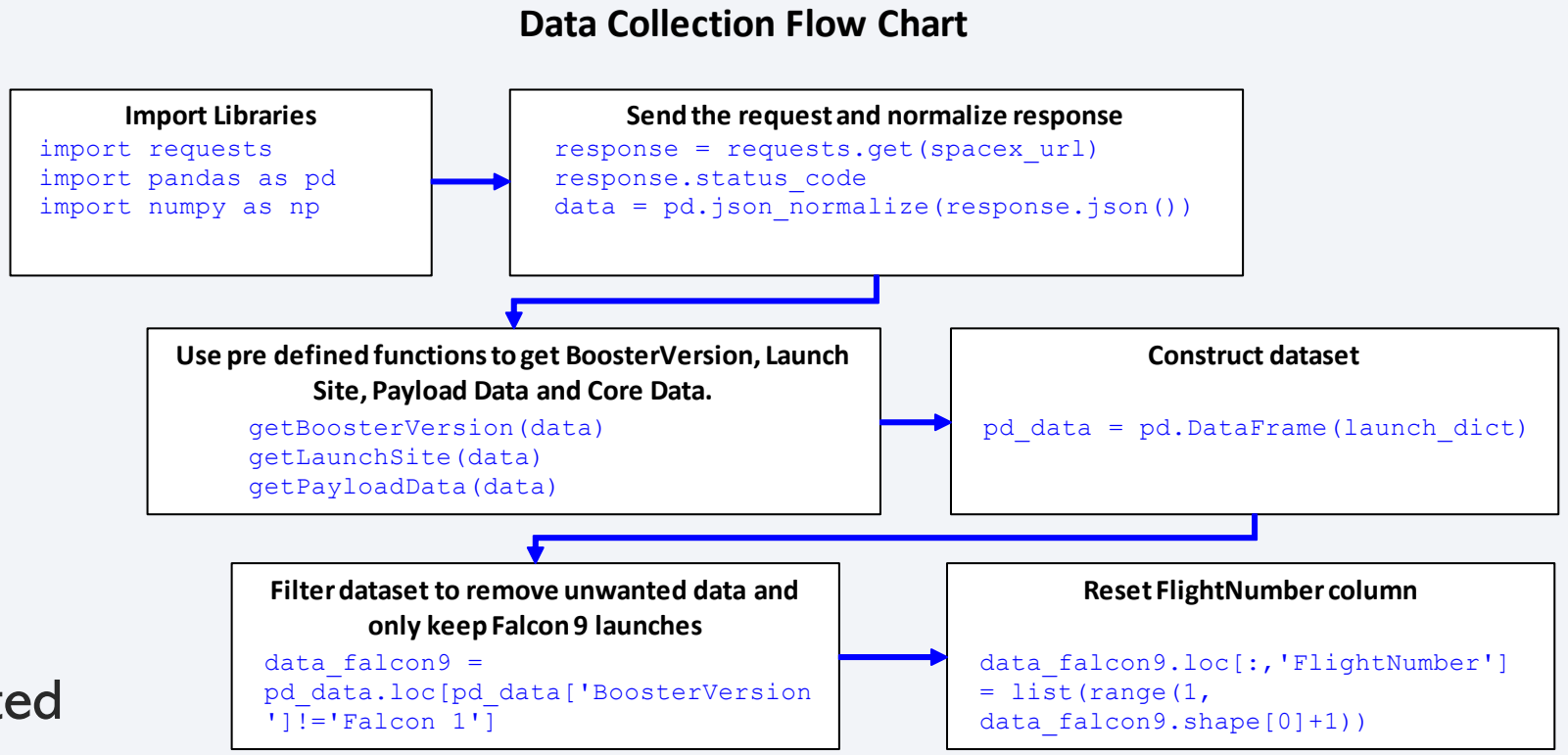
- Perform interactive visual analytics using Folium and Plotly Dash
  - With interactive visual analytics using Folium, objective was to mark all launch sites along with launch outcomes of each launch site on a map. This helped in identifying the optimum launch site.
  - Plotly Dash visual dashboard provides a way to interactively explore and find out correlation between different variables which further impacts the cost of each launch.
- Perform predictive analysis using classification models
  - By splitting the feature engineered data set in to training and test data, various methods where used to predict the launch outcomes and further evaluated by generating confusion matrix to check the accuracy rate of each model.

Data Collection

- SpaceX Falcon launch data is available through API. Also, the website Wikipedia provides some information about the launch testing done over the years since it's first launch.
- The data used for exploratory data analysis was collected using REST API end point
- Web scrapping method was used to collect data from WIKIPEDIA website.

Data Collection – SpaceX API

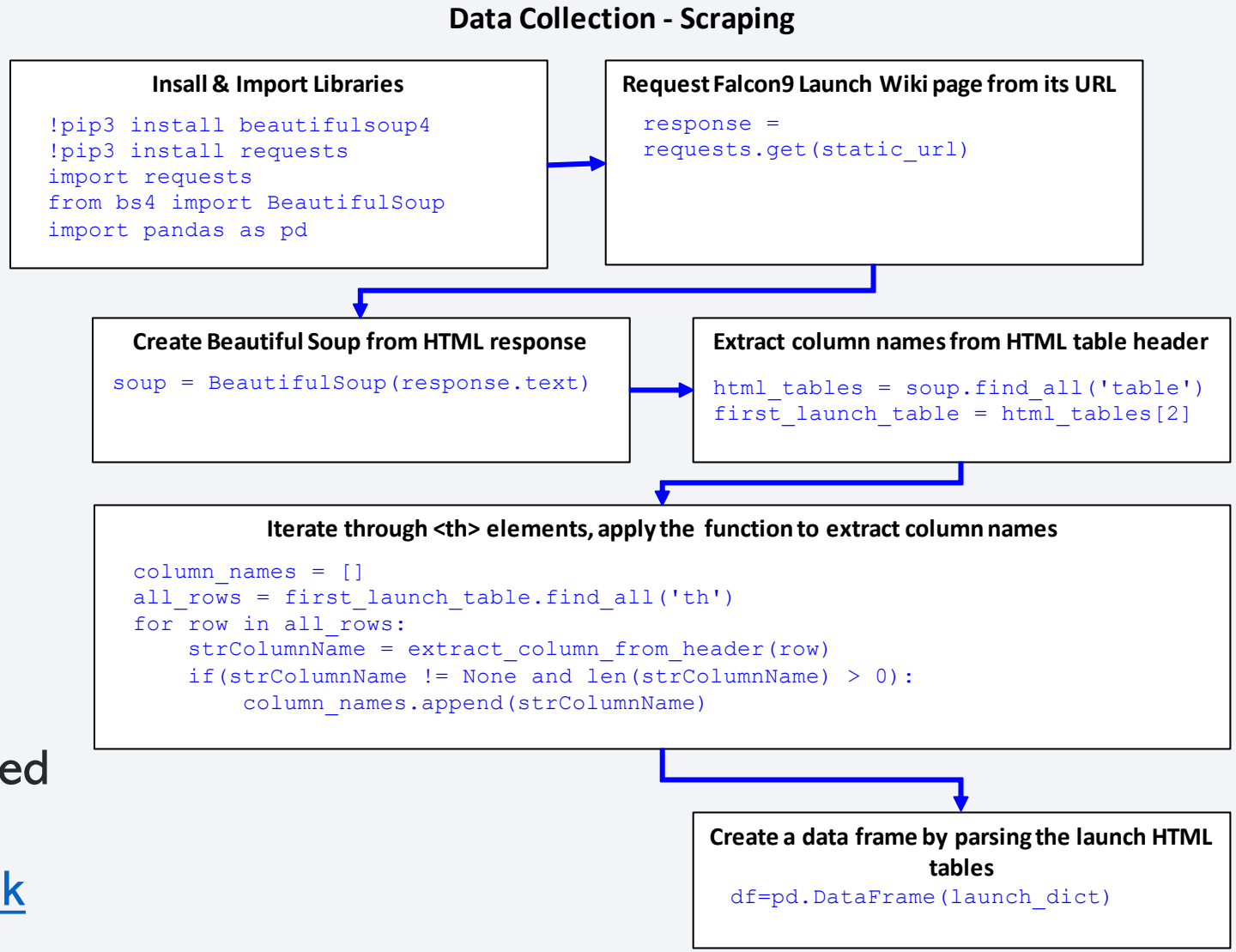
Data collection with SpaceX REST calls



GitHub URL of the completed SpaceX API calls – [01_APIDataCollection_Notebook]

Data Collection - Scraping

Web scraping process flowchart ->



GitHub URL of the completed web scraping notebook - [02_WebScraping_Notebook]

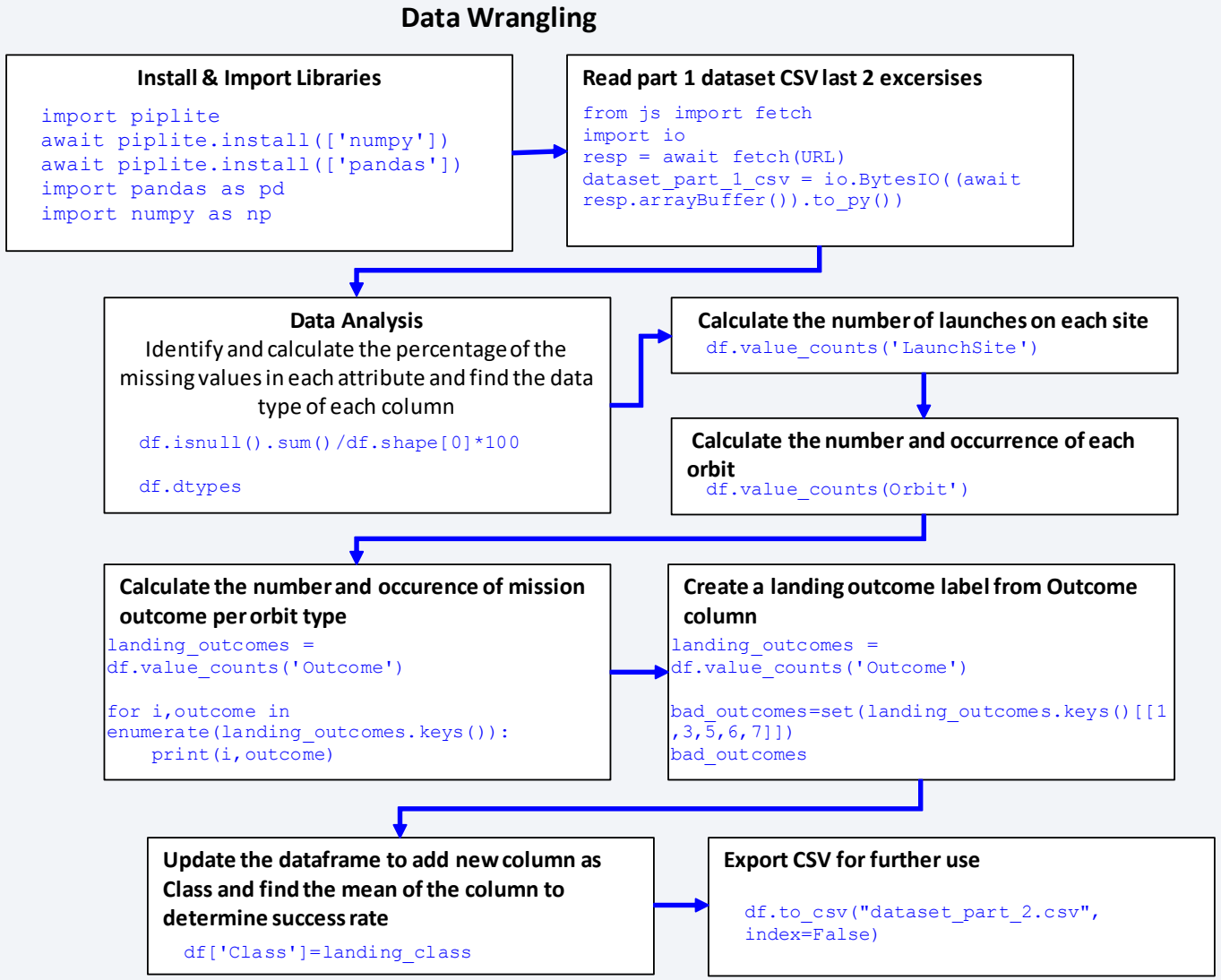
Data Wrangling

During the data wrangling lab, exploratory data analysis was performed to find patterns in the data.

Also, training labels were created and based on the outcome of each launch as 0 or 1 in “Class” column

This class column is utilized in next exercises.

GitHub URL of your completed data wrangling notebook - 03_DataWrangling_EDA



EDA with Data Visualization

Summary of charts that were plotted

1. Scatter plots were plotted using the Seaborn library to visualize the correlation between following variables:
   - Payload mass and Flight Number
   - Launch Site and Flight Number
   - Launch site and Payload mass
   - Orbit and Flight Number
   - Orbit and Payload Mass
2. Bar chart was plotted to visualize the relationship between success rate and each orbit type.
3. Line chart was used to visualize the trend in success rate since year 2010 when the first attempted lunch.

GitHub URL of data visualization notebook – 05_DataVisualization_EDA_Notebook.ipynb

EDA with SQL

Summary of the SQL queries performed:

1. Display the names of the unique launch sites in the space mission.
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS).
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes

Continued on next slide…

EDA with SQL - Continued

Summary - continued

8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
9. 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.
10. Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub URL EDA with SQL notebook - 04_SQL_EDA_Notebook.ipynb

Build an Interactive Map with Folium

Summary of map objects added to a folium map and purpose of using each one of them:

1. Circle and Marker – are used to mark all launch sites on a map.
2. MarkerCluster – are used to mark the success or failed launches for each site on the map.
3. MousePosition – is used to locate the longitude and latitude of the location using the mouse position on the map.
4. PolyLine – is ploted to calculate the distances between a launch site to its proximities.

GitHub URL of completed Folium map notebook -
06_DV_LaunchSiteLocations_Notebook.ipynb

Build a Dashboard with Plotly Dash

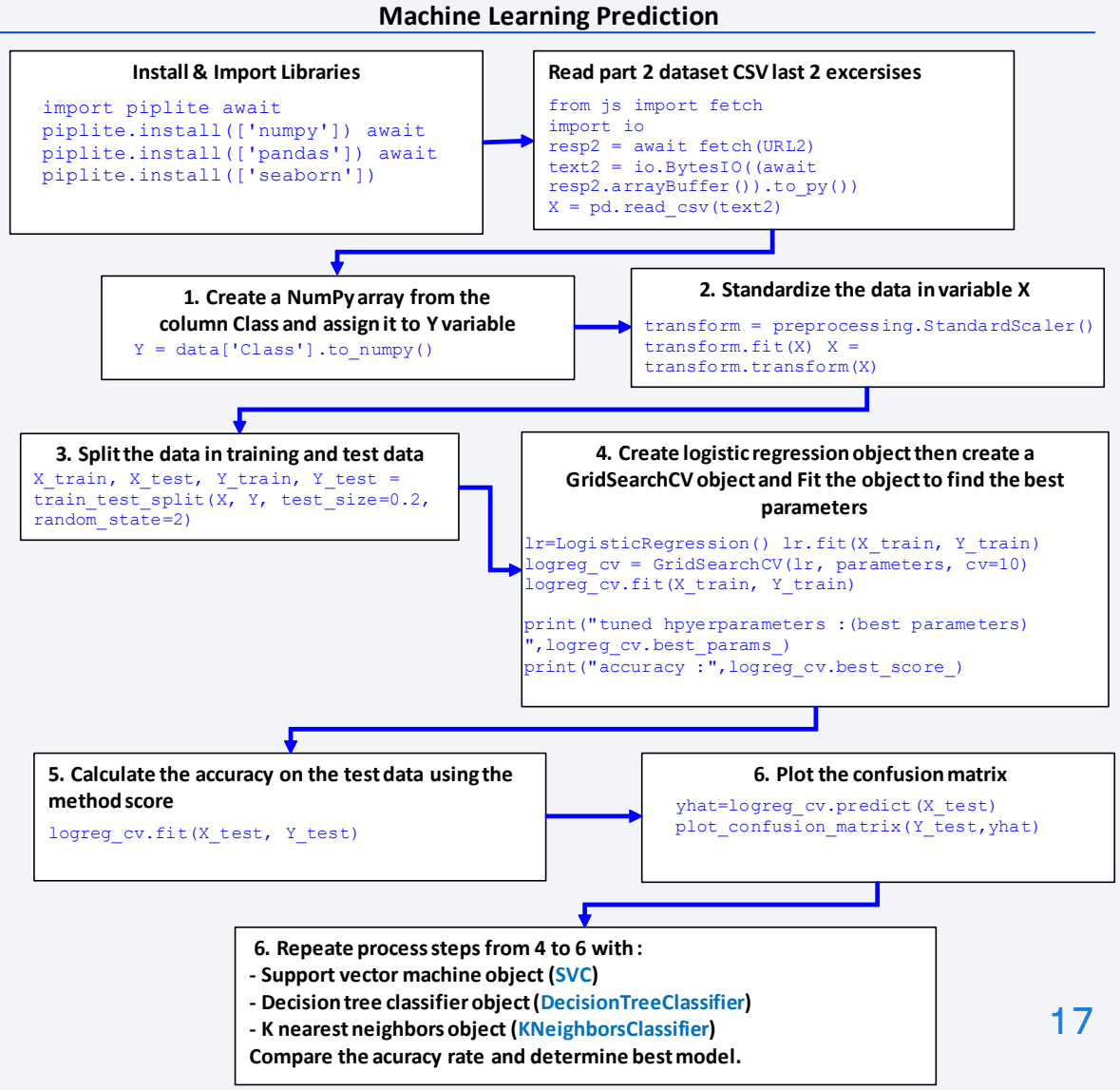
Summary of plots/graphs and interactions on the dashboard

1. Dropdown – a box with options to choose a launch site with option to select “All Sites”
2. Pie chart – launch site and success rate – the pie chart shows the launch site wise successful launches. When a particular launch site is selected from dropdown, the pie chart shows successful and failed launches for the particular site.
3. Slide Bar – the slide bar shows the range of the payload mass for the selected site.
4. Scatter plot – scatter plot shows the correlation between Payload and success for selected launch site or all the launch sites based on the selection in the dropdown box at the top.

- GitHub URL of completed Plotly Dash lab - 07_spacex_dash_app.py

Predictive Analysis (Classification)

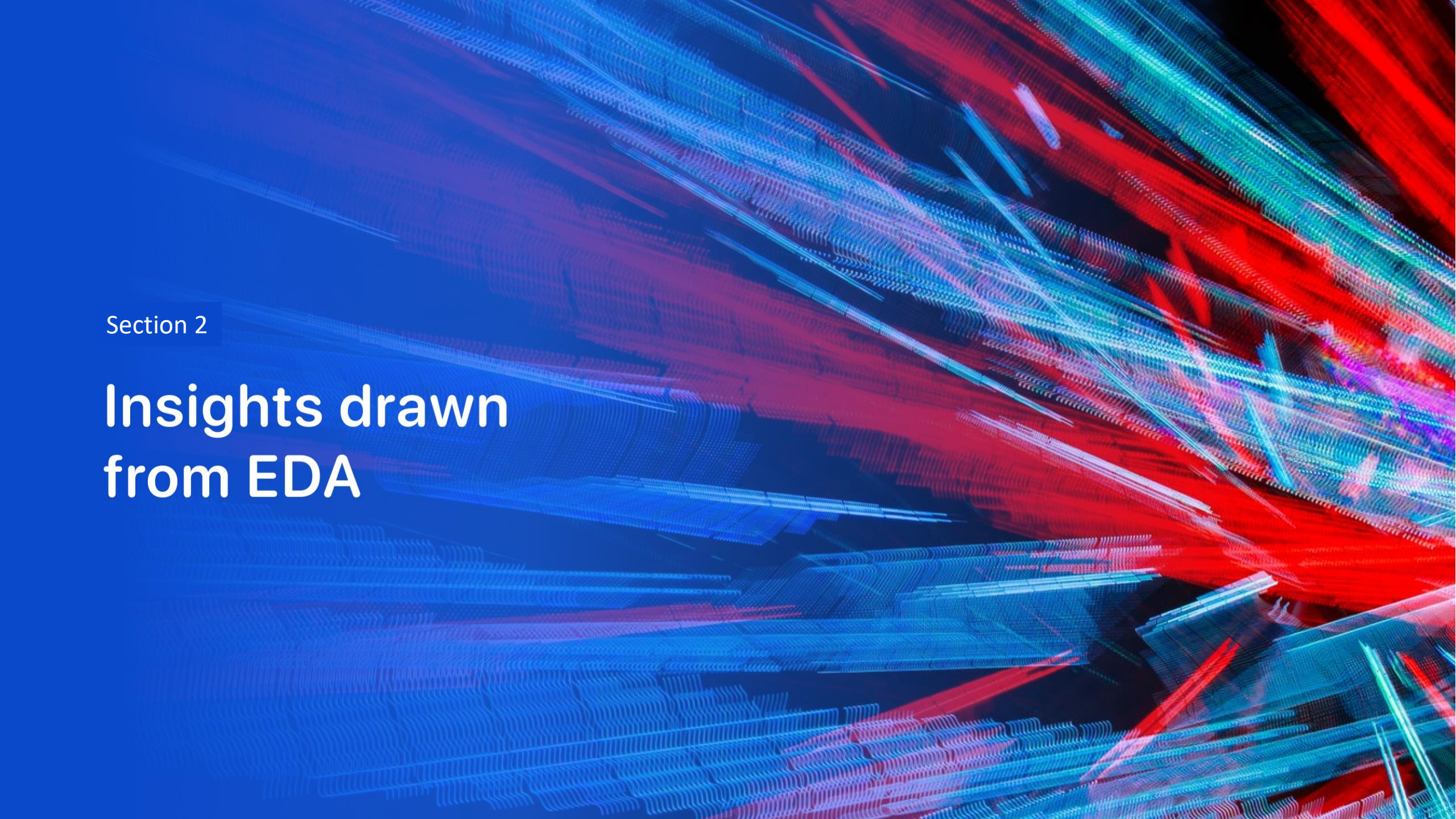
Summary flow chart of classification model building, evaluating of classification model →



GitHub URL of completed predictive analysis lab - [08_ML_Prediction_Part_5.ipynb]

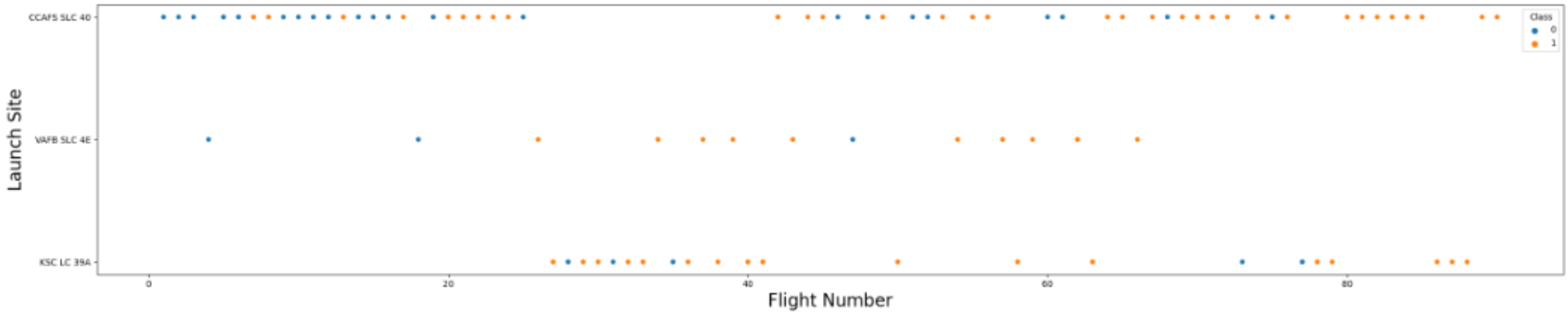
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Scatter plot of Flight Number vs. Launch Site

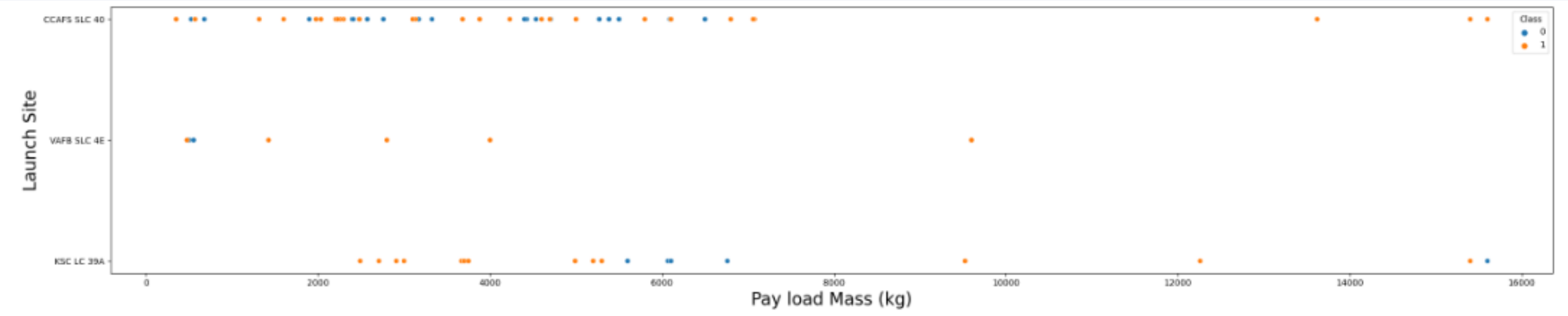


Explanation summary of the scatter plot:

1. The orange dots show successful outcome and blue dot represents failure outcome.
2. We can see that CCAFS LC-40 has the maximum number of flight launches.

Payload vs. Launch Site

Scatter plot of Payload vs. Launch Site



Explanation summary of the scatter plot:

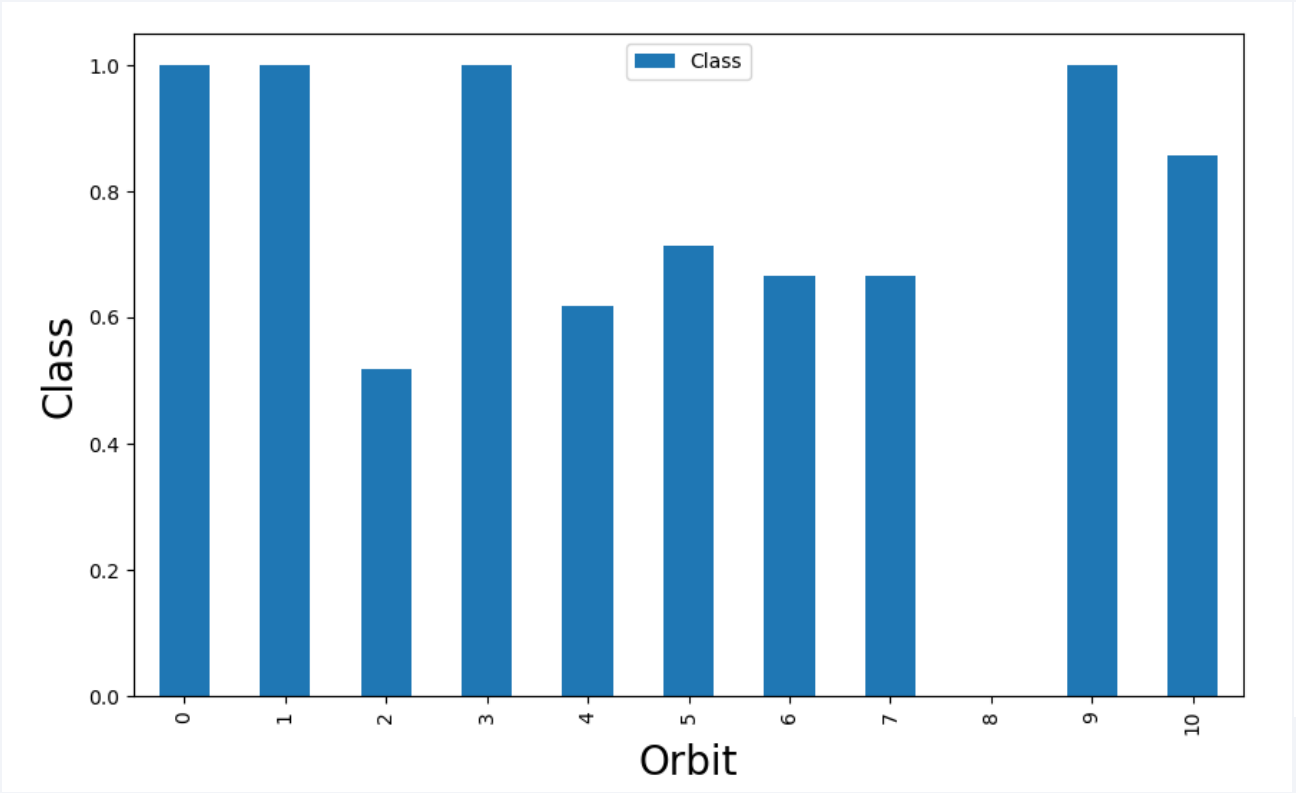
- In this plot also, orange dot represents successful outcome and blue dot represents failure outcome.
- VAFB-SLC launch site has no rockets launched for heavy payload mass (greater than 10000).
- In general, there seems to be weak to no correlation between Launch Site, Payload mass and success rate of the launches.

Success Rate vs. Orbit Type

Bar chart for the success rate of each orbit type

Explanation summary of the Bar chart:

- We can observe that ES-L1, GEO, HEO and SSO orbit types has higher success rates.



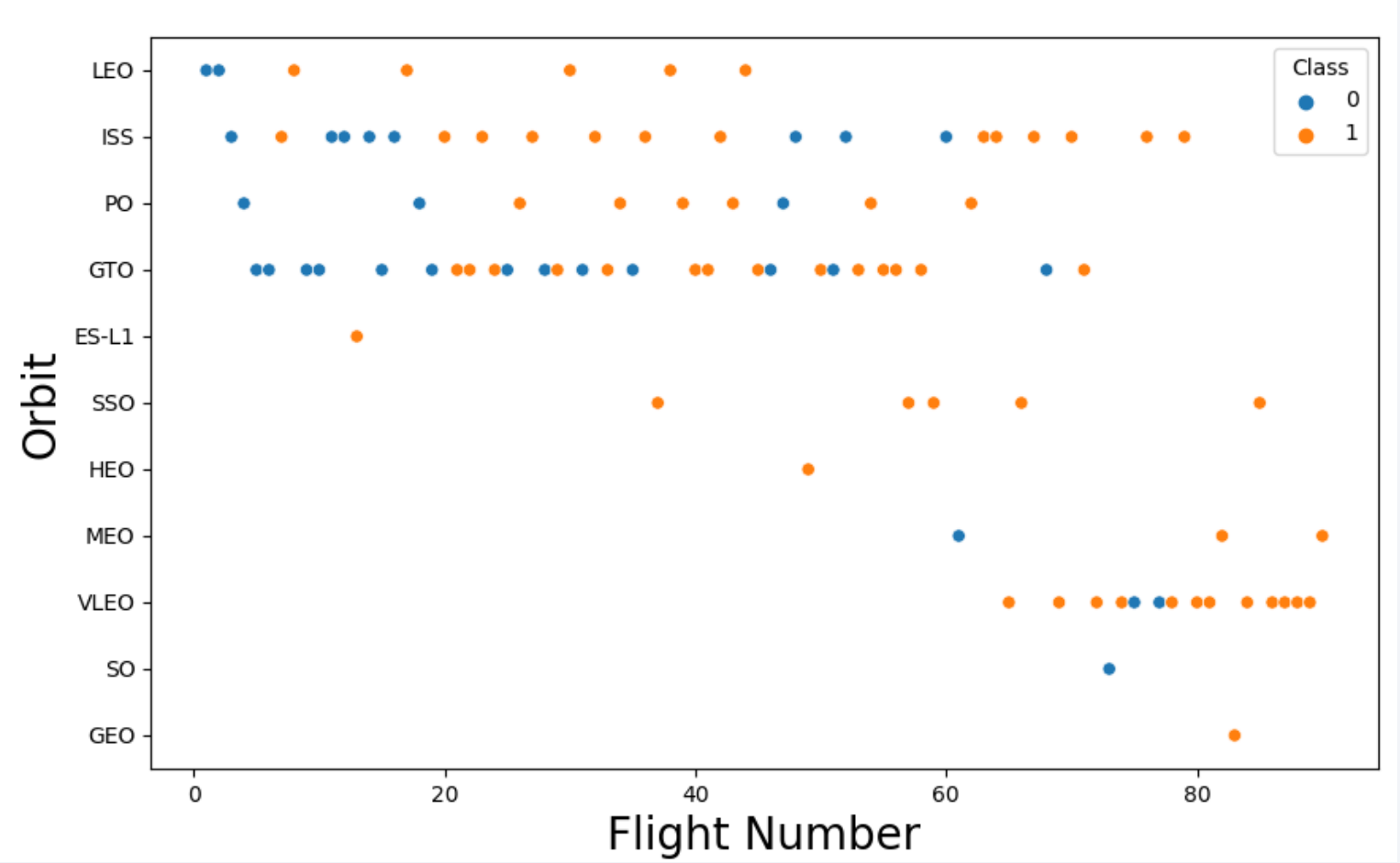
	Orbit	Class
0	ES-L1	1.000000
1	GEO	1.000000
2	GTO	0.518519
3	HEO	1.000000
4	ISS	0.619048
5	LEO	0.714286
6	MEO	0.666667
7	PO	0.666667
8	SO	0.000000
9	SSO	1.000000
10	VLEO	0.857143

Flight Number vs. Orbit Type

Scatter point plot of Flight number vs. Orbit type

Explanation summary of the scatter plot:

- For the LEO orbit the success seems to relate to the number of flights; where as, there seems to be no relationship between flight number when in GTO orbit.

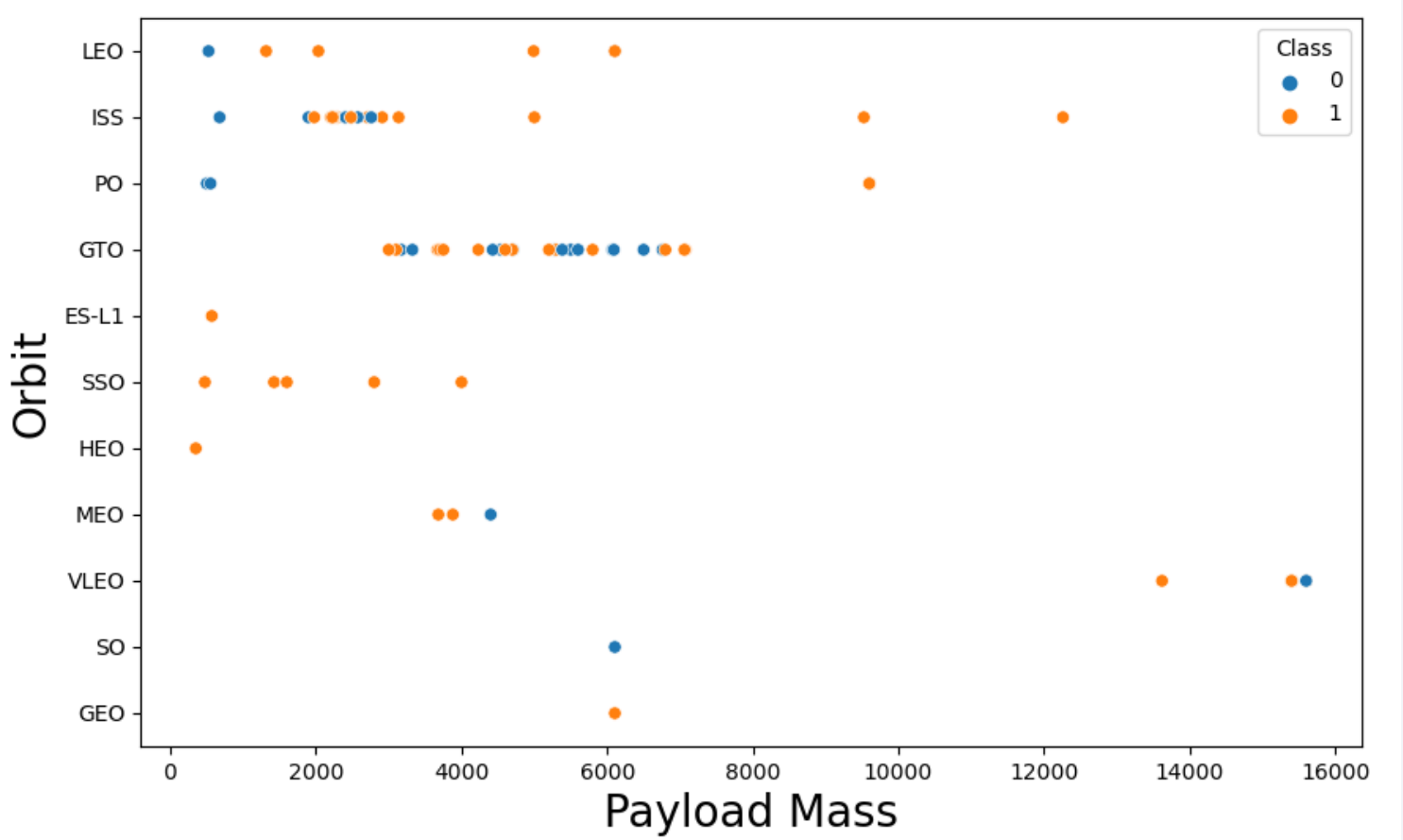


Payload vs. Orbit Type

Scatter point plot of Payload vs. Orbit type

Explanation summary of the scatter plot:

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

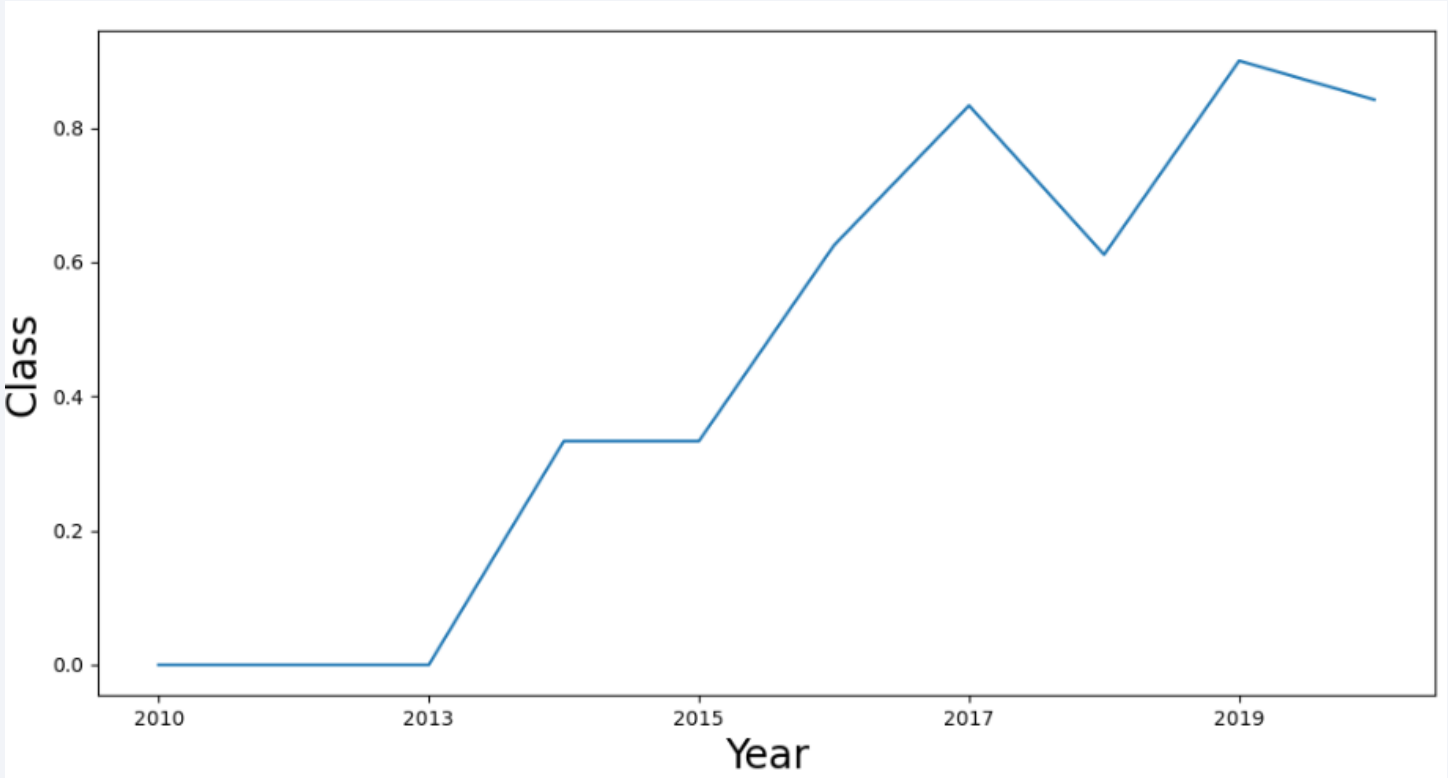


Launch Success Yearly Trend

Line chart of yearly average success rate

Explanation summary of the line chart:

- Since year 2013, average success rate shows significant upward trend
- There is also a noticeable dip in year 2018 and 2020.



All Launch Site Names

Names of the unique launch sites

- The sql query to select Distinct values from “Launch_Site” column returns the unique names of the Launch sites in the entire data table. With this we are able to identify all the site names

```
#%sql select * from SPACEXTBL limit 5
%sql select DISTINCT("Launch_Site") from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

- Using SQL query to match the Launch_Site name which starts with “CCA”, we can use Like phrase with limit option to get only 5 records matching the criteria.

```
%sql select * from SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qual
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight CubeSats, barrel of Brouere
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo f
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX

Total Payload Mass

Calculate the total payload carried by boosters from NASA

- Sum of the Payload_Mass__KG_ column is calculated using SQL query for Customer “NASA” as 45,596 KG

```
%sql select sum("PAYLOAD_MASS__KG_") as PayloadMassInKG from SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

PayloadMassInKG
45596

Average Payload Mass by F9 v1.1

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

- Average payload mass carried by booster version is calculated by using AVG function on Payload_Mass__KG_ column with criteria of booster_version as ‘F9 v1.1’
- The average is noted as 2928.4 kg

```
%sql select AVG("PAYLOAD_MASS__KG_") as AvgPayloadMassInKG from SPACEXTBL WHERE Booster_Version = 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

AvgPayloadMassInKG
2928.4

First Successful Ground Landing Date

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

- MIN function is used on Date column to get the smallest date along with the criteria of Mission outcome as ‘Success’ and Landing outcome like ‘Success (ground pad)’
- The first ground landing date is noted as 01-05-2017

```
%sql select MIN(DATE), * from SPACEXTBL WHERE "Mission_Outcome" = 'Success' AND "Landing _Outcome" LIKE 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

MIN(DATE)	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
01-05-2017	01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)

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

- There are four booster versions which have successfully landed on drone ship with payload mass between 4000 and 6000 KG.

```
%sql select * from SPACEXTBL WHERE "Mission_Outcome" = 'Success' AND "Landing _Outcome" LIKE 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" between 40
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
06-05-2016	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
14-08-2016	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
11-10-2017	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

- Count function is used on Mission outcome column to calculate the number of successful mission outcomes.

```
%sql Select COUNT(*) AS NoOfSuccessMissions from SPACEXTBL WHERE "Mission_Outcome" LIKE '%Success%'
```

```
 * sqlite:///my_data1.db
Done.
```

NoOfSuccessMissions
100

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

Boosters Carried Maximum Payload

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

- Sub query is used to find maximum payload and select the booster versions which carried the max payload

```
%sql Select "Booster_Version", "PAYLOAD_MASS__KG_" from SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

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

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

- There were two failed landing outcomes in drone ship in year 2015

```
%%sql
select substr(Date, 4, 2) AS MonthNames, "Date", "Booster_Version","Launch_Site", "Landing _Outcome"
    from SPACEXTBL WHERE "Landing _Outcome" LIKE '%Failure (drone ship)%' AND substr(Date,7,4)='2015'
```

```
 * sqlite:///my_data1.db
Done.
```

MonthNames	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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

- Ranking of the success outcome count is done by using ‘Order By’ and ‘Desc’ phrases.

```
%%sql
select COUNT("Landing _Outcome") AS nCount, "Landing _Outcome", "Date" from SPACEXTBL
    WHERE "Landing _Outcome" LIKE '%Success%' AND "Date" between '04-06-2010' AND '20-03-2017'
    GROUP By "Landing _Outcome" Order By nCount DESC
```

```
 * sqlite:///my_data1.db
Done.
```

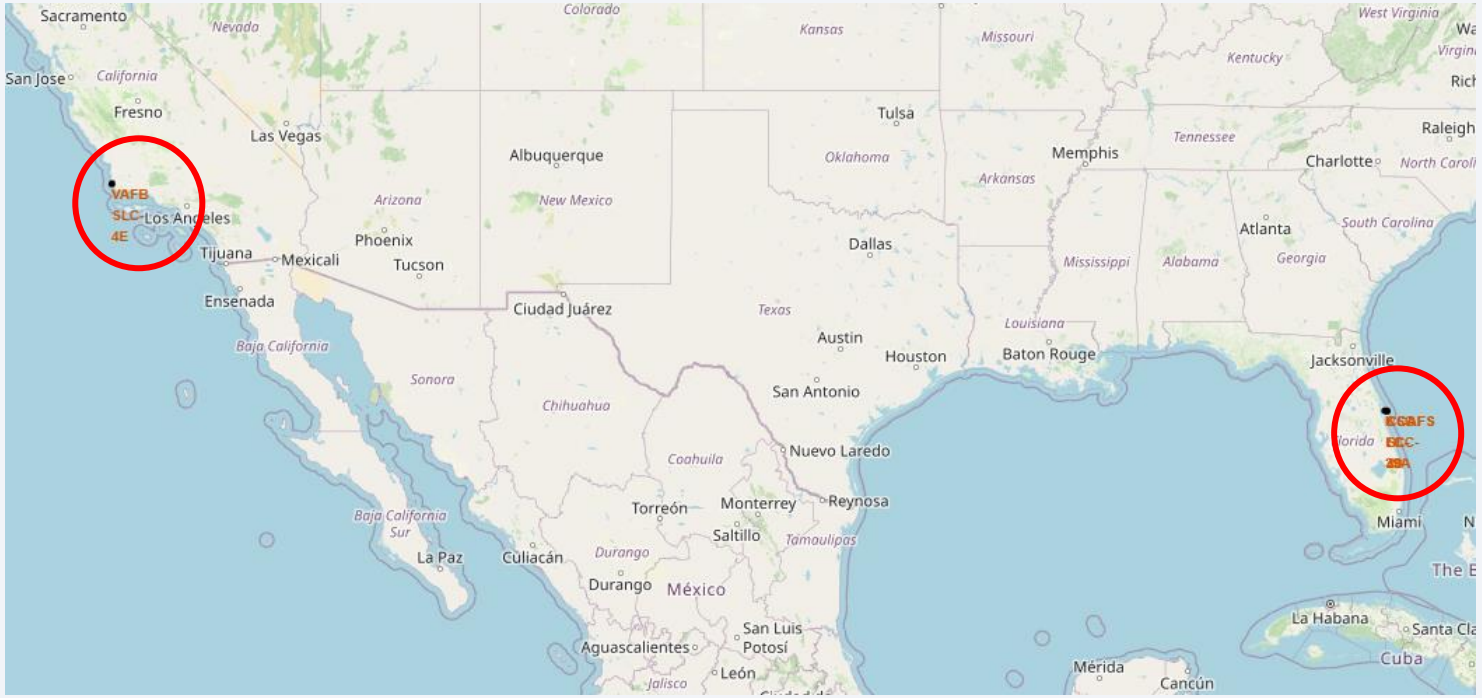
nCount	Landing _Outcome	Date
20	Success	07-08-2018
8	Success (drone ship)	08-04-2016
6	Success (ground pad)	18-07-2016

Section 3

Launch Sites Proximities Analysis

All Launch Site Locations

Screenshot of all launch sites location markers on a global map



Points observed and noted:

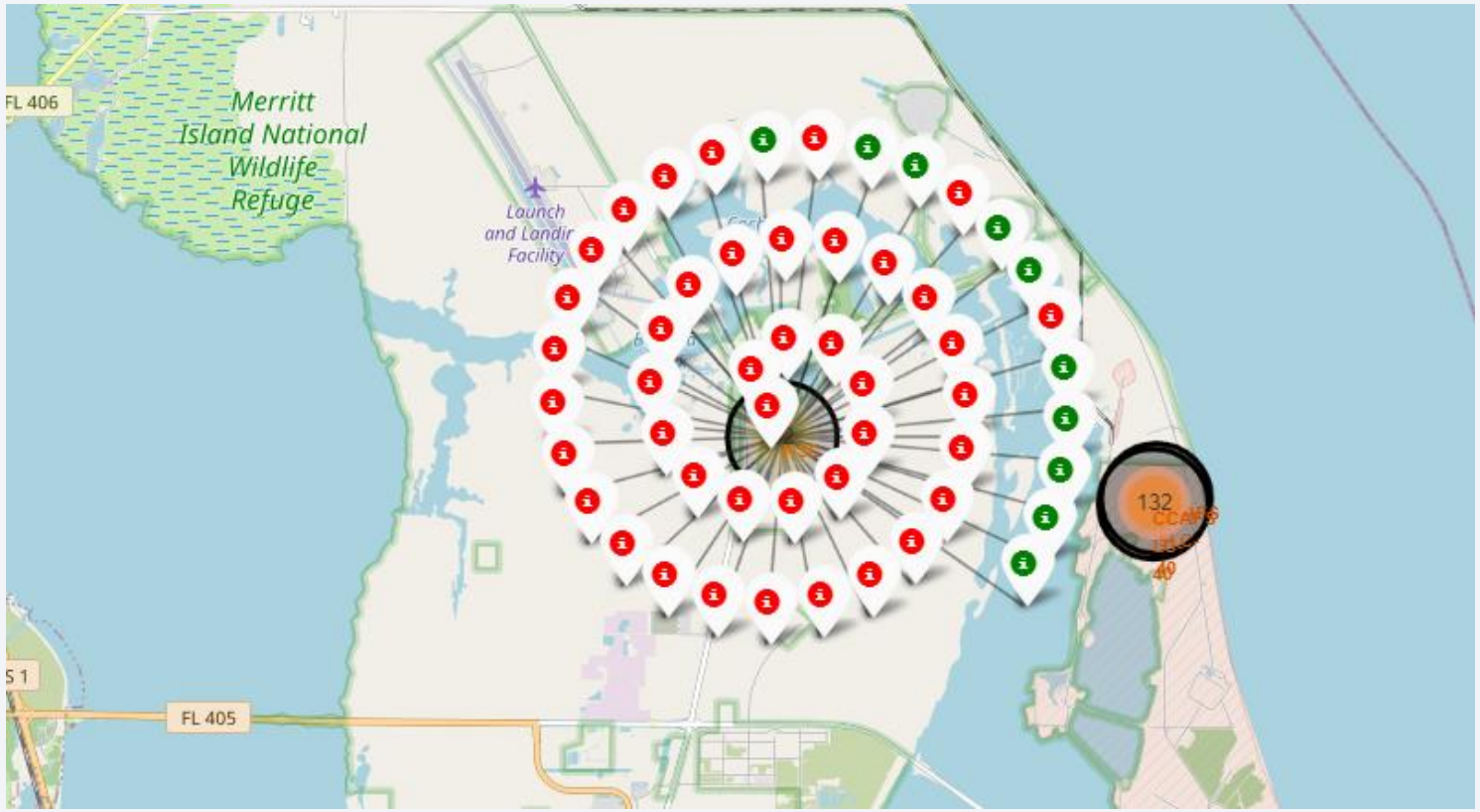
- Launch sites are located near the equator line enabling spacecraft achieve fuel and speed efficiency due to greater rotation speed of the earth near it's equator line. Which means, the launches can be cost effective.
- We can also note that all launch sites are close to the coastal areas in order to mitigate the risk of failed launches making damage on ground surface.

Launch Site KSC LC-39A

Screenshot of KSC LC-39A launch site with color coded launch outcomes

Points observed and noted:

- Cluster of marker for the launch site named KSC LC-39A are showing the red colored markers for failed outcomes and greens ones show the successful outcomes.

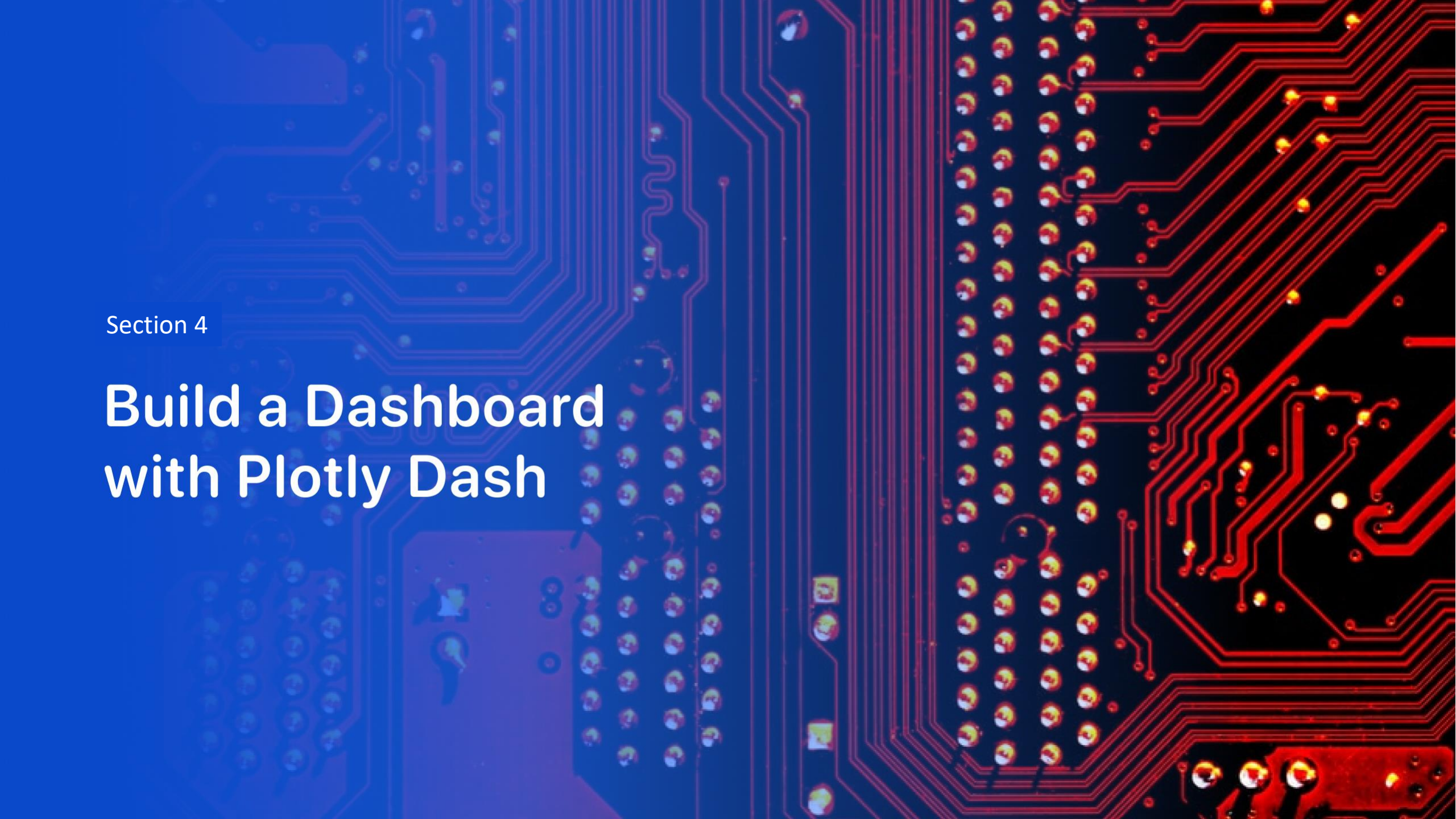


Launch Sites and Distance to Proximities

Screenshot of KSC LC-39A launch site and it's distance to coastline

Points observed and noted:

- The launch sites are away from the towns and heavy population areas and public air ports or public commute facilities like railway stations etc. for the obvious reasons of safety

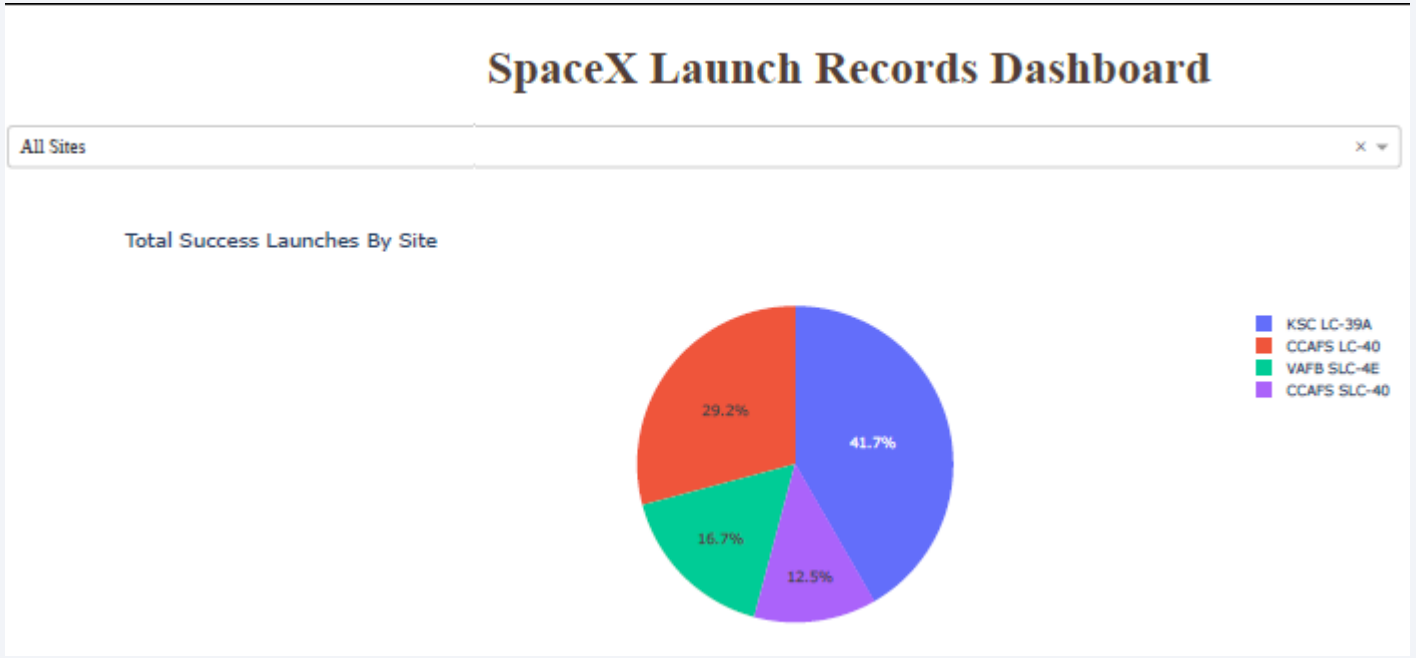

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Sites

Screenshot of pie chart from dashboard showing all sites success launches

Points observed and noted:

- KSC LC-39A site has 41.7% successful launches followed by CCAFS LC-40 site with 29.2%, VAFB SLC-4E with 16.7%, CCAFS SLC-40 site with 12.5% successful launches.

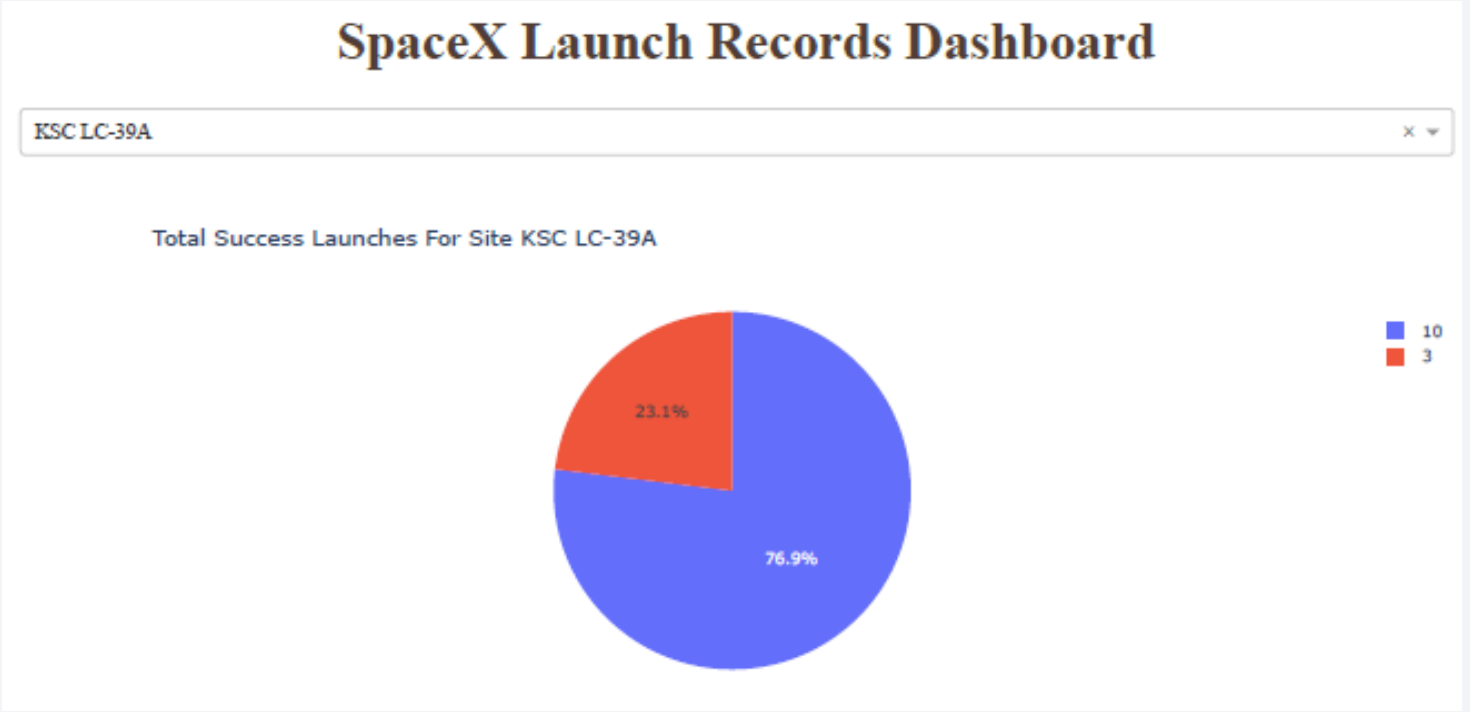


Pie chart for KSC LC-39A

Screenshot of pie chart from dashboard for KSC LC-39A launch site

Points observed and noted:

- KSC LC-39A site shows highest success launch ratio KSC LC-39A site has 76.9% positive outcome

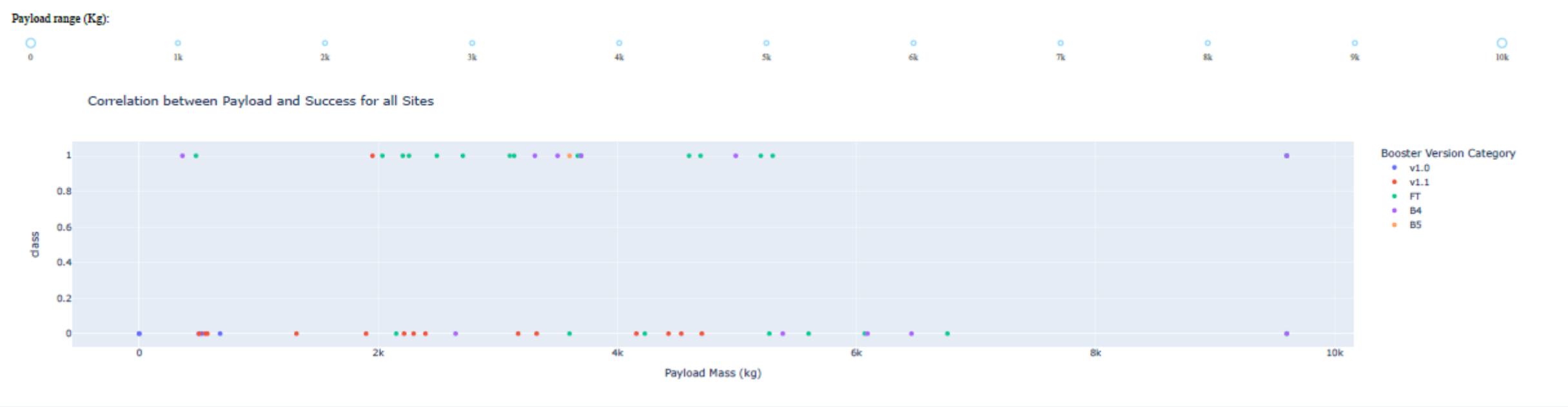


Payload vs. Launch Outcomes

Screenshot of scatter plot from dashboard for all launch sites with range slider

Points observed and noted:

- Payload ranges between 2000 KG and 4000 KG has the highest launch success rate.
- Payload ranges between 0 KG and 2000 KG has the lowest launch success rate.
- Booster version FT has the highest launch success rate.



Section 5

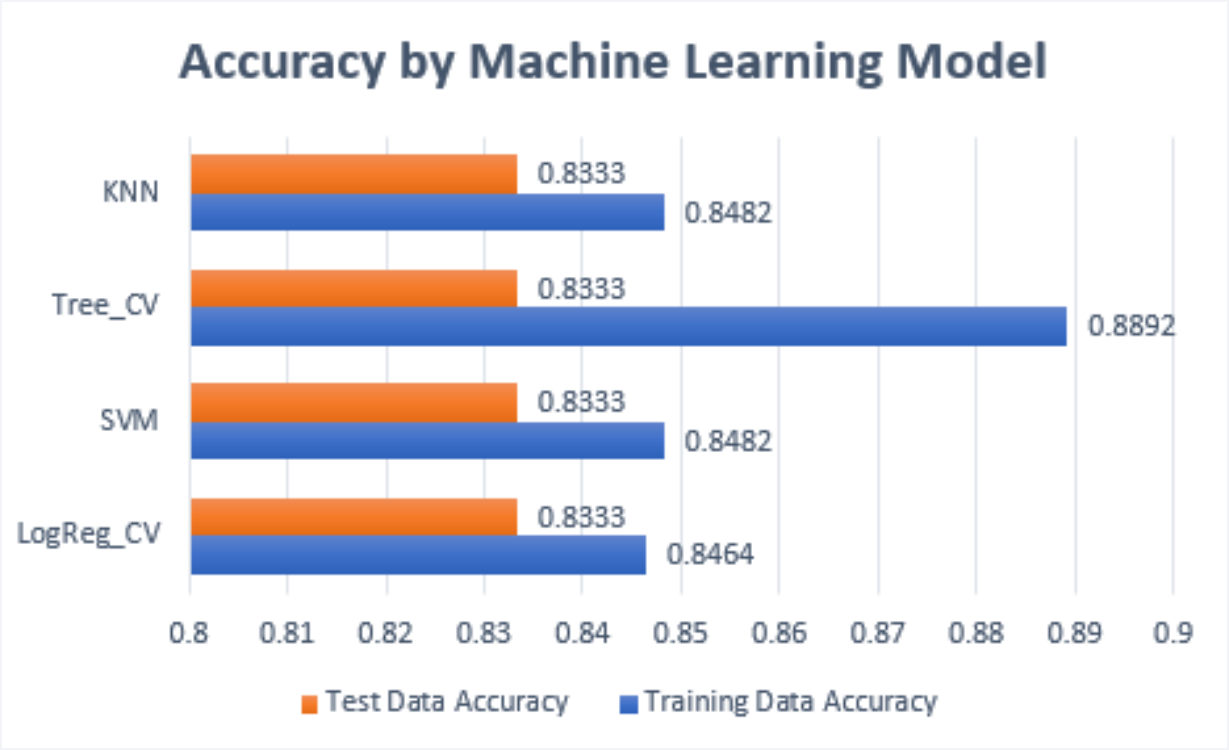
Predictive Analysis (Classification)

Classification Accuracy

Bar chart for accuracy for all built classification models

- All the methods perform equally on the test data. They all have the same accuracy of 0.8333 on the test Data

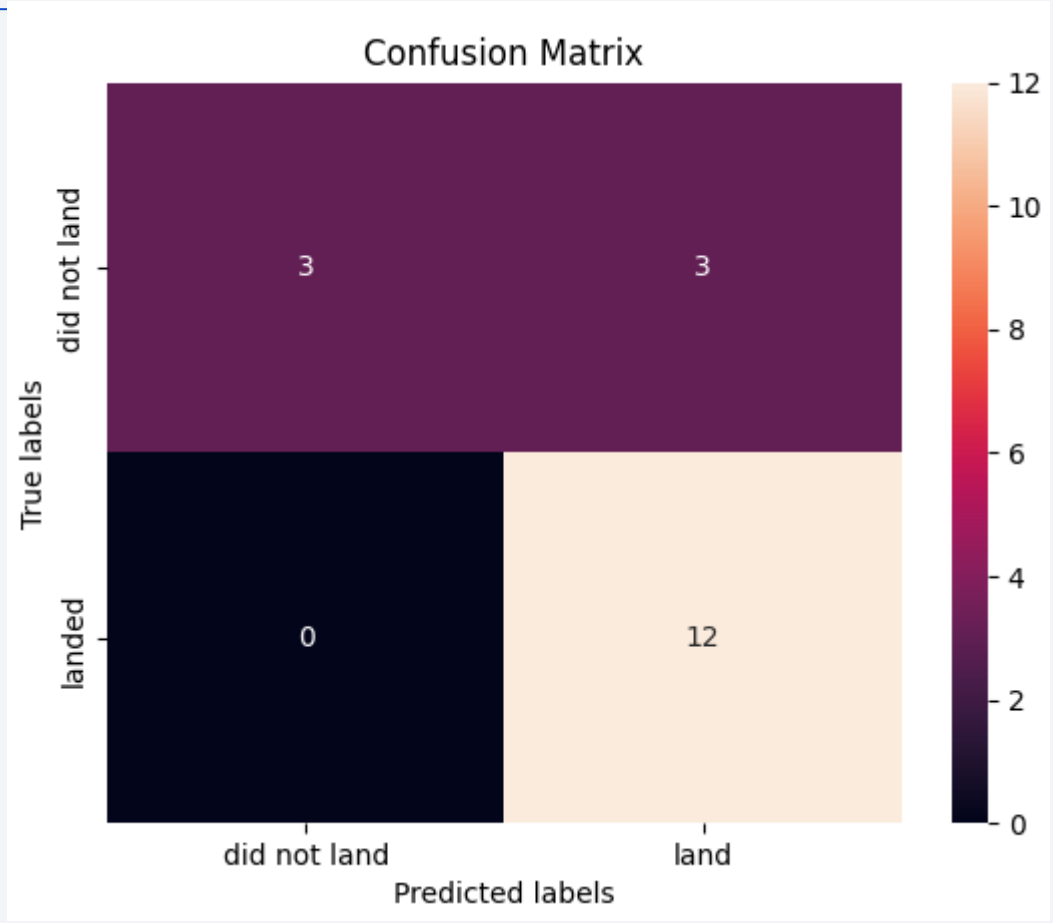
Model Name	Model Short Name	Training Data Accuracy	Test Data Accuracy
Logistic Regression Model	LogReg_CV	0.8464	0.8333
Support Vector Machine	SVM	0.8482	0.8333
Decision Tree Classifier	Tree_CV	0.8892	0.8333
KNeighborsClassifier	KNN	0.8482	0.8333



Confusion Matrix

Confusion matrix for the models

- All models showed same accuracy level on test data.
- The models predicted 12 successful landings when True Label was Successful (True Positive) and 3 unsuccessful landings when True label was Failure (True Negative)
- The model also predicted 3 successful landings when True label was Failure (False Positive)
- The models generally predicted successful outcome.



Conclusions

- CCAFS LC-40 has the most flight launches, while VAFB-SLC has none for heavy payloads.
- ES-L1, GEO, HEO, and SSO orbits have higher success rates, and successful landing rates are higher for heavy payloads in Polar, LEO, and ISS orbits.
- Success rates have been increasing since 2013, with dips in 2018 and 2020.
- Launch sites are located near the equator provide fuel and speed efficiency, making launches cost-effective.
- KSC LC-39A has the highest successful launch rate, followed by CCAFS LC-40, VAFB SLC-4E, and CCAFS SLC-40.

Overall, these observations suggest that the launch site, orbit type, payload mass can all affect the success rate of rocket launches, thus can be utilized as parameters to predict the launch outcomes.

Appendix

- GitHub repository link - https://github.com/diptipkale/ibmcapstone

Thank you!