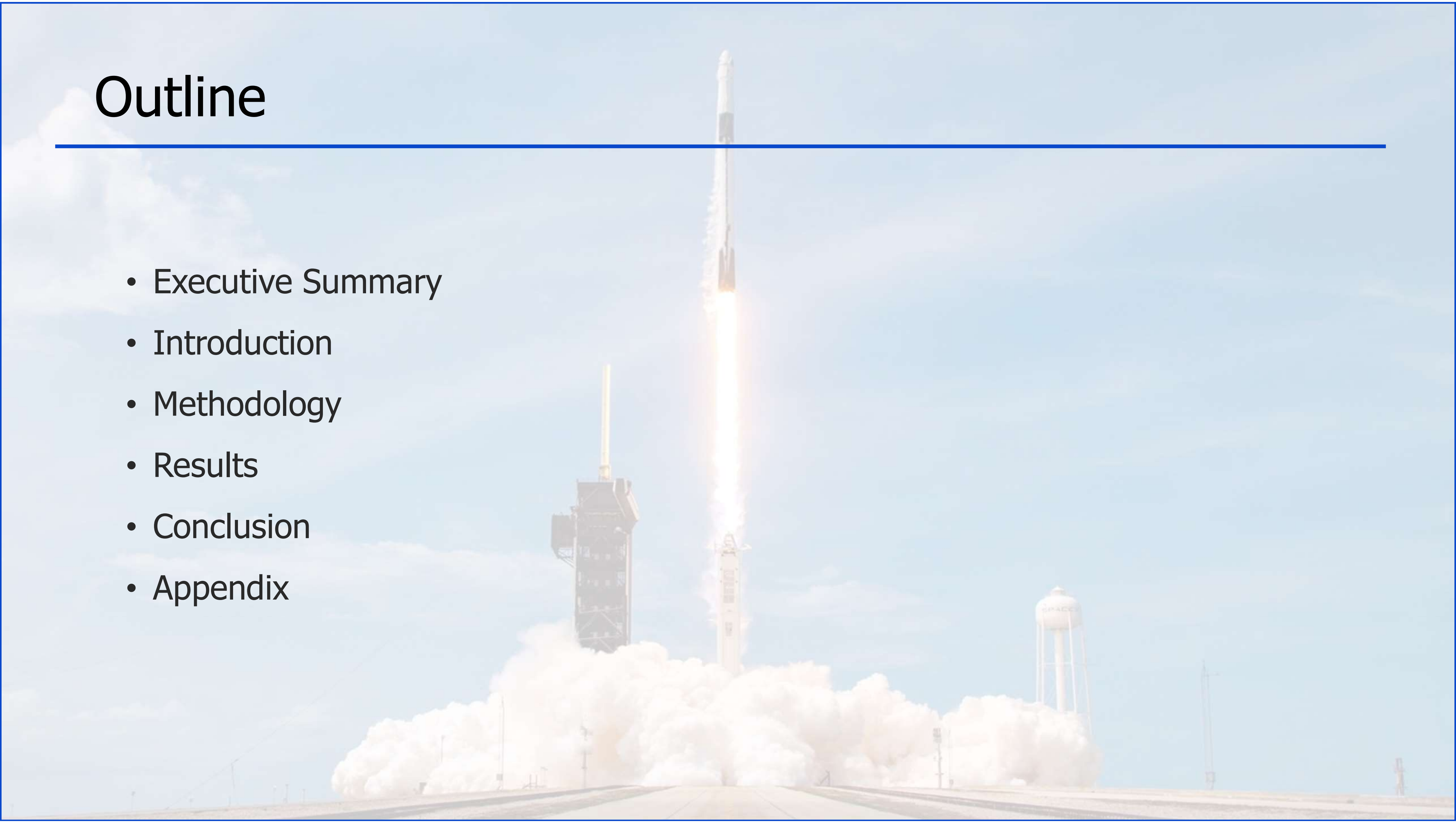
IBM Developer
SKILLS NETWORK

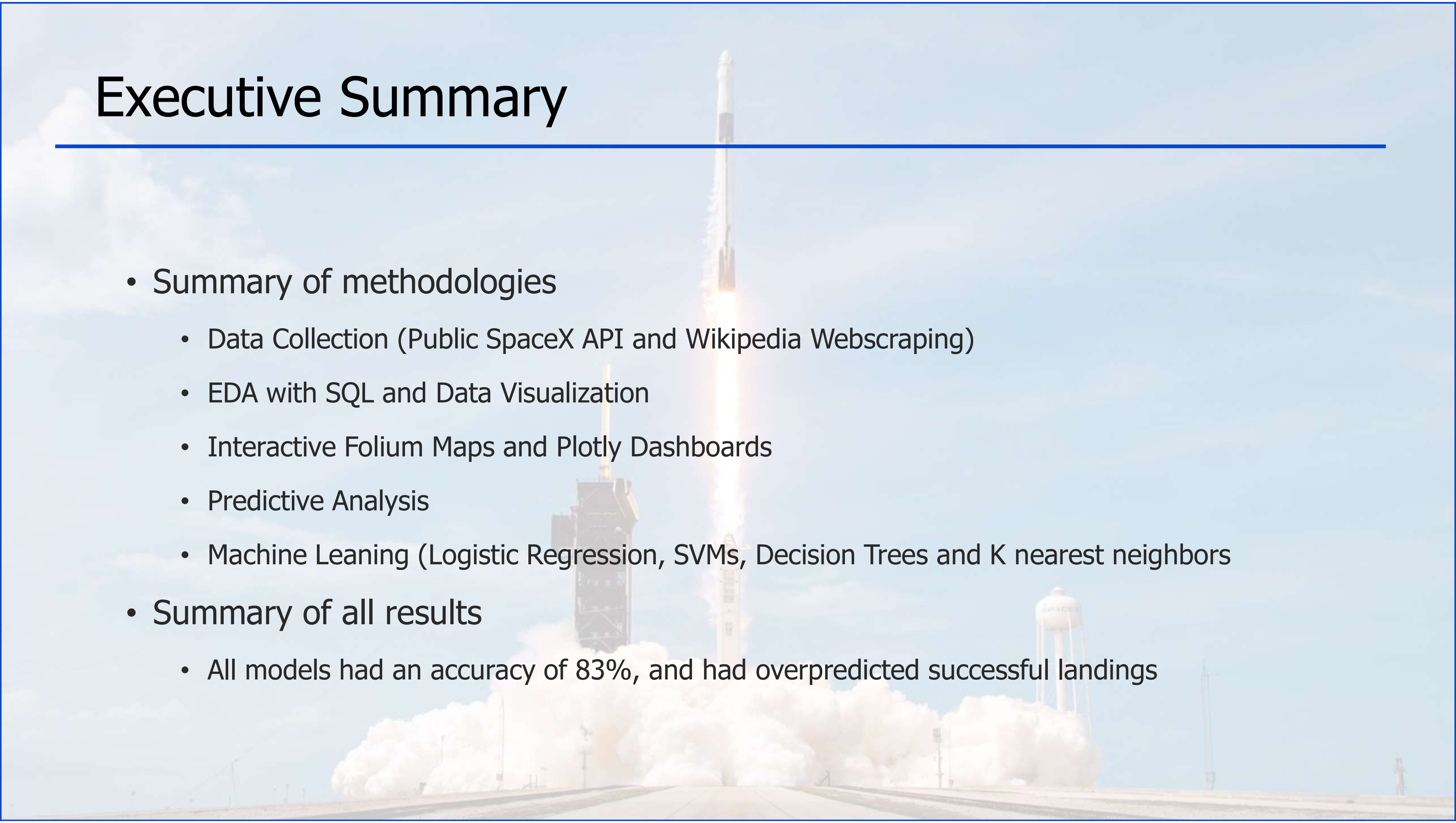
Winning Space Race with Data Science

Muhammad Ibrahim Beg
June 2023



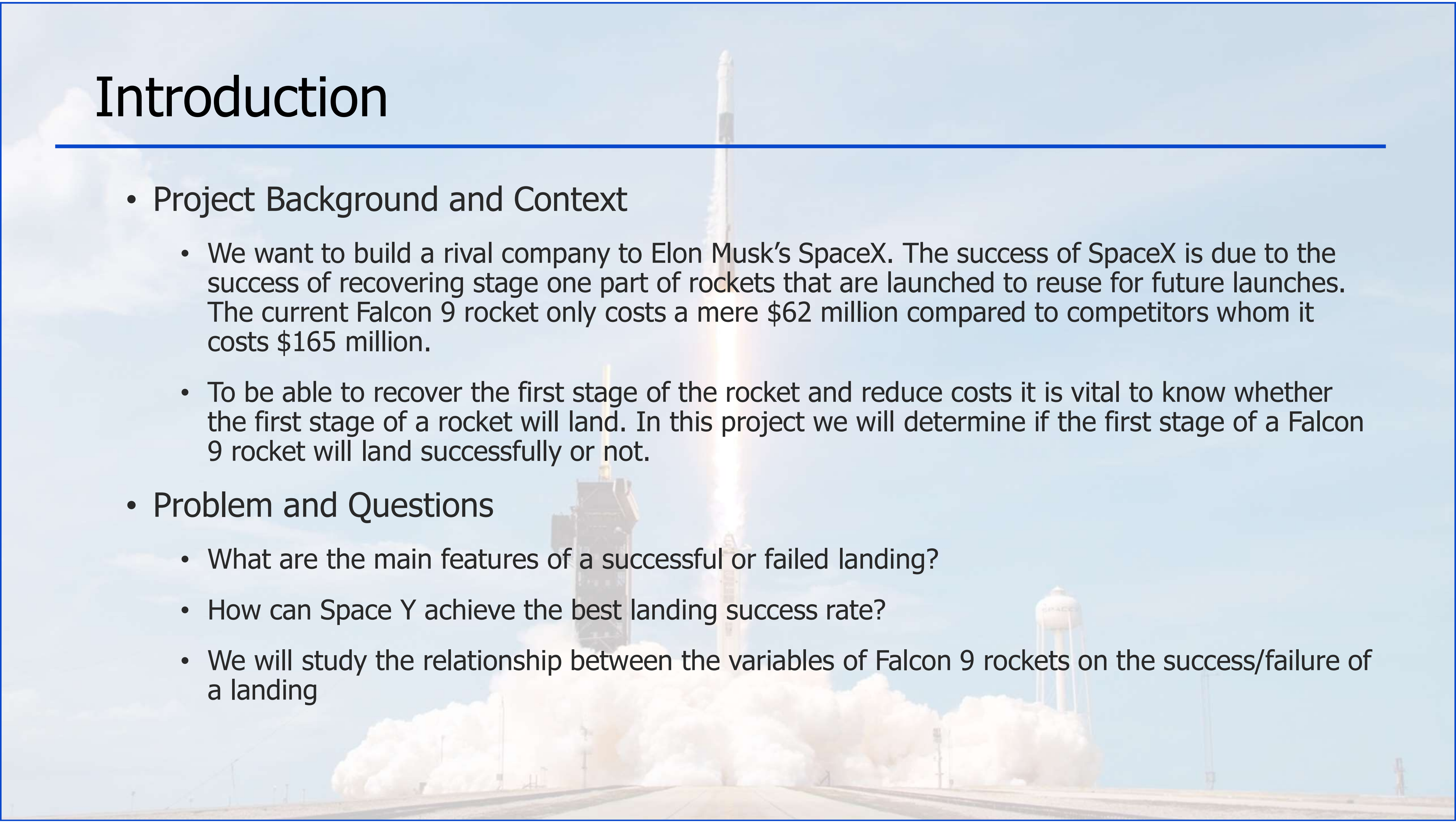
Outline

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



Executive Summary

- Summary of methodologies
  - Data Collection (Public SpaceX API and Wikipedia Webscraping)
  - EDA with SQL and Data Visualization
  - Interactive Folium Maps and Plotly Dashboards
  - Predictive Analysis
  - Machine Leaning (Logistic Regression, SVMs, Decision Trees and K nearest neighbors
- Summary of all results
  - All models had an accuracy of 83%, and had overpredicted successful landings



Introduction

- Project Background and Context
  - We want to build a rival company to Elon Musk’s SpaceX. The success of SpaceX is due to the success of recovering stage one part of rockets that are launched to reuse for future launches. The current Falcon 9 rocket only costs a mere $62 million compared to competitors whom it costs $165 million.
  - To be able to recover the first stage of the rocket and reduce costs it is vital to know whether the first stage of a rocket will land. In this project we will determine if the first stage of a Falcon 9 rocket will land successfully or not.
- Problem and Questions
  - What are the main features of a successful or failed landing?
  - How can Space Y achieve the best landing success rate?
  - We will study the relationship between the variables of Falcon 9 rockets on the success/failure of a landing

Section 1

Methodology

Methodology

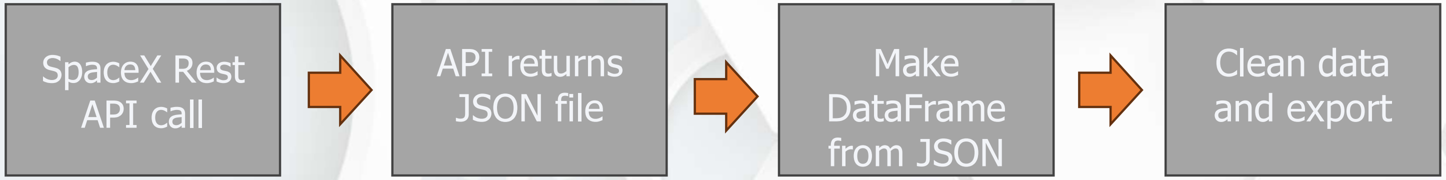
Executive Summary

- Data collection methodology:
  - Public API of SpaceX and Web Scrapping SpaceX Wikipedia
- Perform data wrangling
  - Dropping unnecessary features
  - One Hot Encoding for classification models
- 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 using GridSearchCV

Data Collection

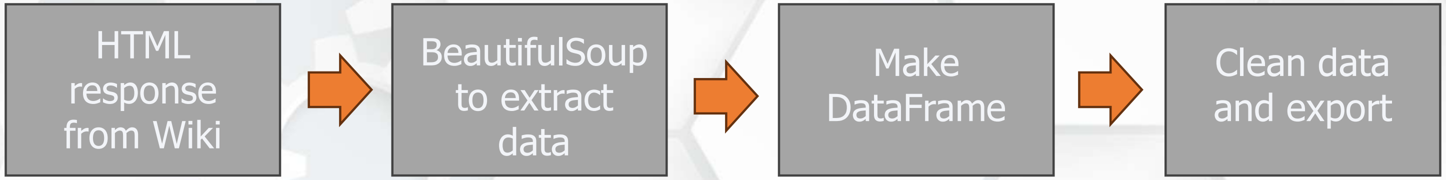
- Data sets were collected from two sources: The Space X Rest API and Webscrapping table data from Wikipedia page on Space X

- The information obtained by the API (api.spacexdata.com/v4/) are rockets, launches, payload information.



SpaceX Rest API call → API returns JSON file → Make DataFrame from JSON → Clean data and export

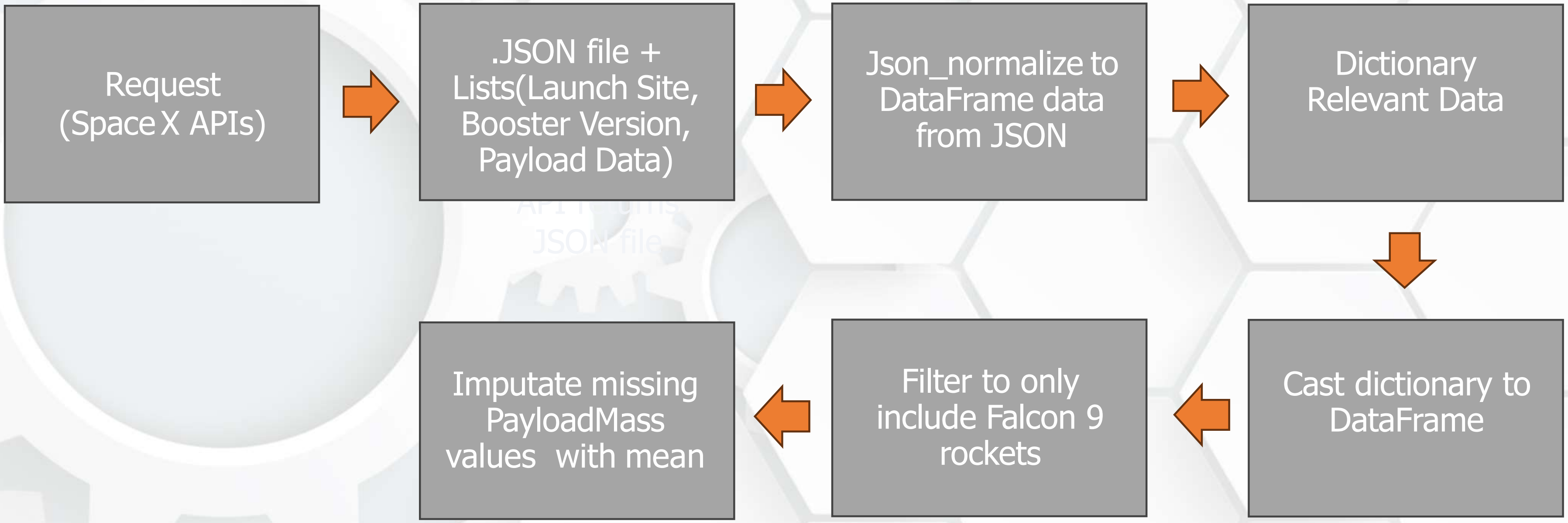
- The information obtained by web scrapping include payload information, launch outcomes, booster landing. (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)



HTML response from Wiki → BeautifulSoup to extract data → Make DataFrame → Clean data and export

Data Collection – SpaceX API

- GitHub URL: https://github.com/ibrahim-beg/Applied-Data-Science-Capstone/blob/671e14447da416241bf24a21d4a6a43bf6c22047/Data%20Collection%20API.ipynb



Request (Space X APIs)

.JSON file + Lists(Launch Site, Booster Version, Payload Data)

Json_normalize to DataFrame data from JSON

Dictionary Relevant Data

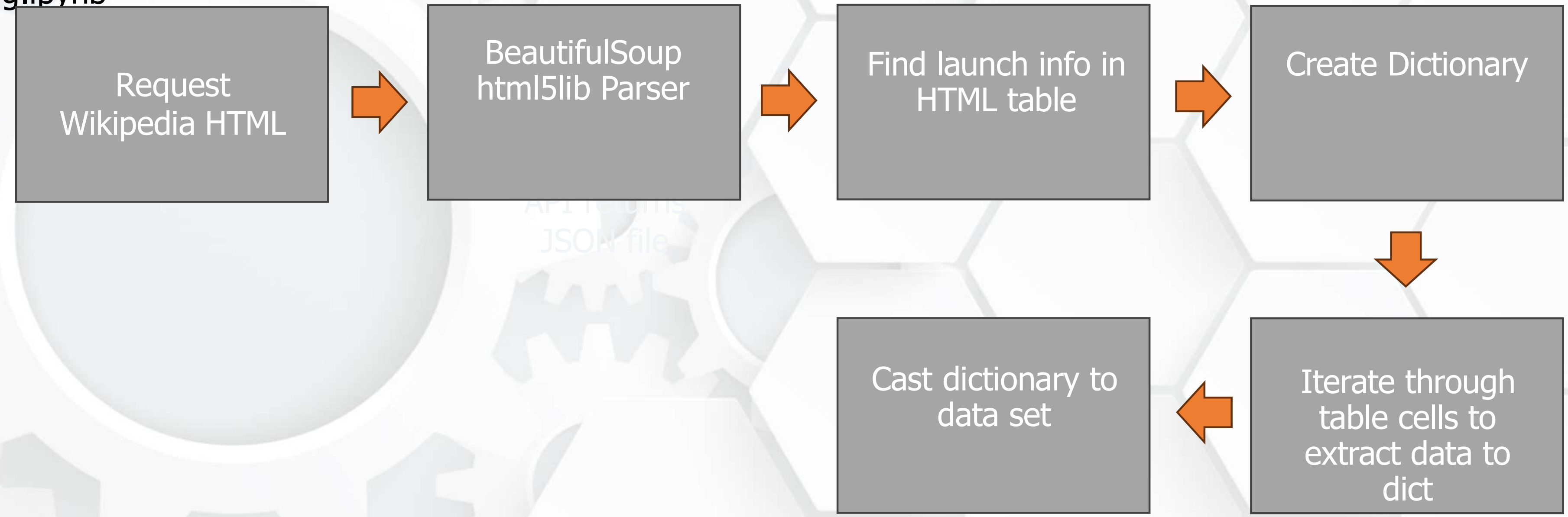
Cast dictionary to DataFrame

Filter to only include Falcon 9 rockets

Imputate missing PayloadMass values with mean

Data Collection – Scrapping

- GitHub URL: https://github.com/ibrahim-beg/Applied-Data-Science-Capstone/blob/671e14447da416241bf24a21d4a6a43bf6c22047/Data%20Collection%20with%20Web%20Scraping.ipynb



Request Wikipedia HTML

BeautifulSoup html5lib Parser

Find launch info in HTML table

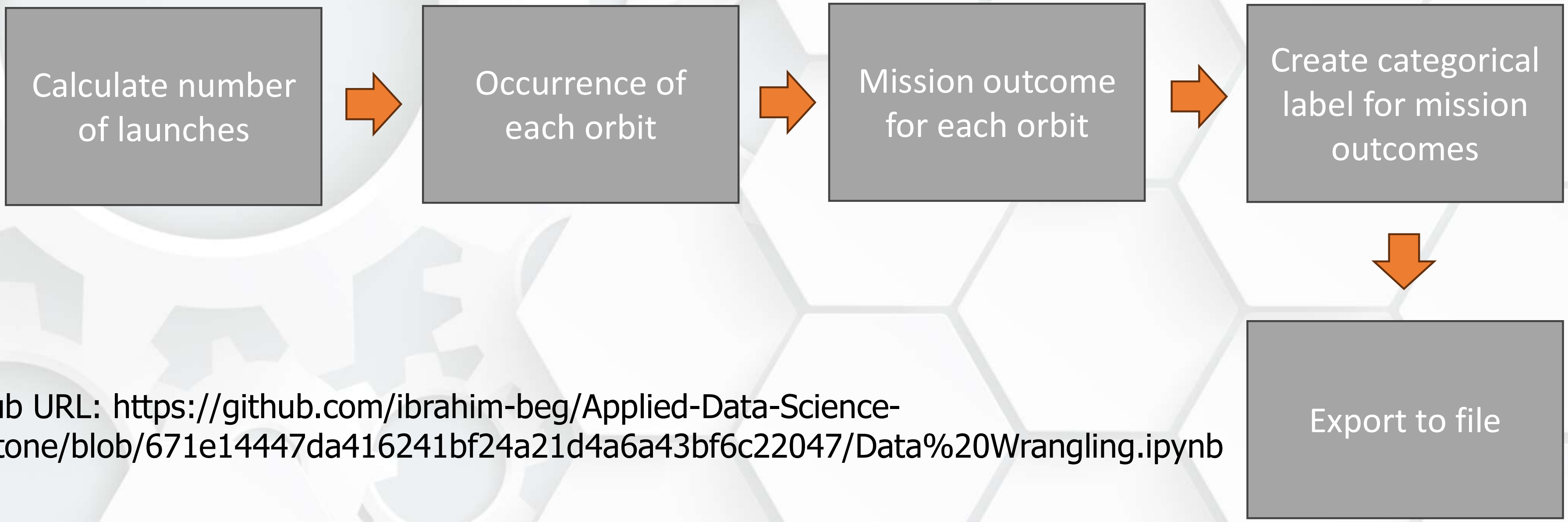
Create Dictionary

Iterate through table cells to extract data to dict

Cast dictionary to data set

Data Wrangling

- There are cases in data where booster did not successfully land.
  - Failed Landings: False Ocean, False RTLS, False ASDS
  - Successful Landings: True ASDS, True RTLS, True Ocean
- We transform these strings into categorical values, 1 for success, 0 for failure.



Calculate number of launches

Occurrence of each orbit

Mission outcome for each orbit

Create categorical label for mission outcomes

Export to file

- GitHub URL: https://github.com/ibrahim-beg/Applied-Data-Science-Capstone/blob/671e14447da416241bf24a21d4a6a43bf6c22047/Data%20Wrangling.ipynb

EDA with Data Visualization

- Various types of charts plotted for different reasons.
- Scatter graphs used to find correlation between different features. Correlation does not necessarily mean causation, but can indicate if two variables have a relationship.
- Scatter graphs:
  - Flight number vs. Launch Site
  - Flight number vs. Payload Mass
  - Payload Mass vs. Launch Site
  - Payload Mass vs. Orbit Type
  - Orbit Type vs. Flight Number
  - Orbit Type vs. Payload Mass
- Bar chart of Success Rate vs. Orbit to compare orbit types to see which had the largest success rate.
- Line graph of Success Rate vs. Year, to see if the company is having more successful landings with each passing year, if the success had been decreasing we could observe which features where changed during this period to assess how it affected success.

- GitHub URL: https://github.com/ibrahim-beg/Applied-Data-Science-Capstone/blob/671e14447da416241bf24a21d4a6a43bf6c22047/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

Summary of SQL queries carried out:

- Unique names of all Launch Sites
- Launch sites beginning with ‘CCA’
- Total Payload mass carried by booster launched by NASA and average payload mass carried by F9 booster
- Dare of first successful landing outcome on ground pad
- Boosters who have payload mass between 4000kg and 6000kg
- Total number of successful and unsuccessful missions
- Which boosters have carried the max payload mass
- GitHub URL: https://github.com/ibrahim-beg/Applied-Data-Science-Capstone/blob/671e14447da416241bf24a21d4a6a43bf6c22047/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Folium map object is a map centered on NASA Johnson Space Center at Houson, Texas, a red circle is used to denote the center of our map
- Red circles to locate the launch site coordinates
- Green markers to show where there were successful landings
- Red markers for locations of failed landings
- Group markings for clusters to show different information for same coordinates
- Markers to show distance between launch site and key locations
- Line plotted between launch site and key locations

- These objects to show geographically represent the data, so we can visualize locations and areas of successful/failed landings.

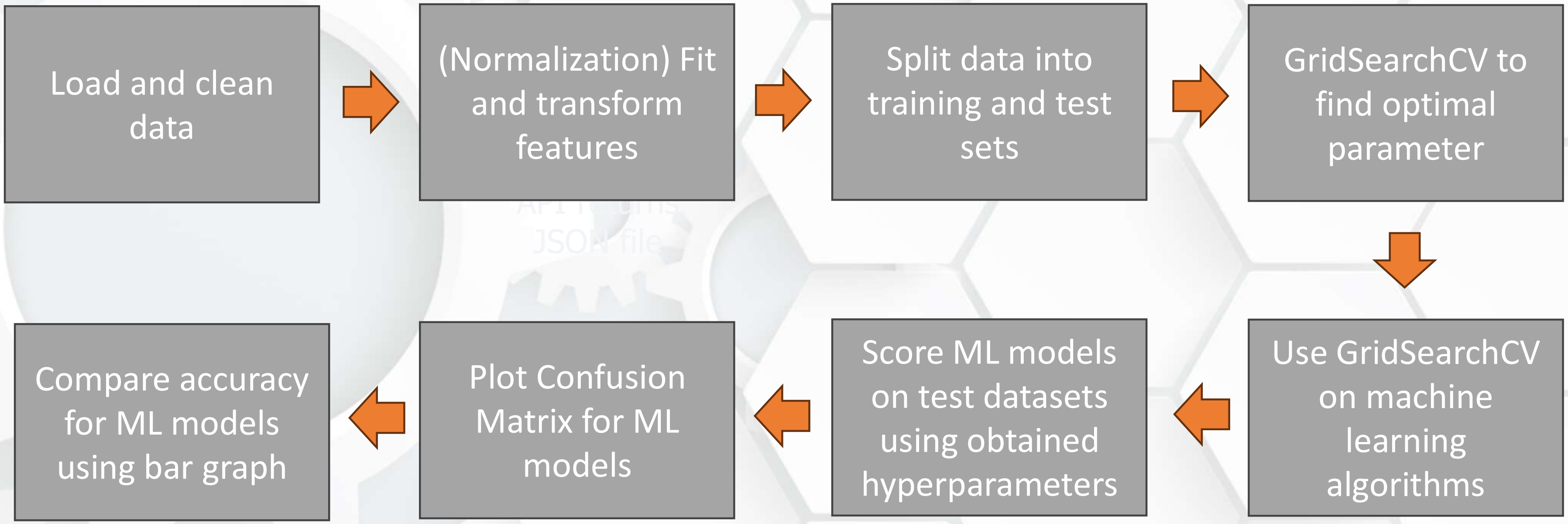
- GitHub URL: https://github.com/ibrahim-beg/Applied-Data-Science-Capstone/blob/671e14447da416241bf24a21d4a6a43bf6c22047/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Dashboard has dropdown, pie chart, range slider and scatter plot components
- Dropdown allows a user to choose the launch site or all launch sites (`dash_core_components.Dropdown`).
- Pie chart shows the total success and the total failure for the launch site chosen with the dropdown component (`plotly.express.pie`).
- Rangeslider allows a user to select a payload mass in a fixed range (`dash_core_components.RangeSlider`).
- Scatter chart shows the relationship between two variables, in particular Success vs Payload Mass (`plotly.express.scatter`). We can see that as payload mass increased our success rate decreased…

- GitHub URL: https://github.com/ibrahim-beg/Applied-Data-Science-Capstone/blob/d7a34f852e9042cc739b3292295cf335810af22b/spacex_dash_app.py

Predictive Analysis (Classification)



Load and clean data

(Normalization) Fit and transform features

Split data into training and test sets

GridSearchCV to find optimal parameter

Use GridSearchCV on machine learning algorithms

Score ML models on test datasets using obtained hyperparameters

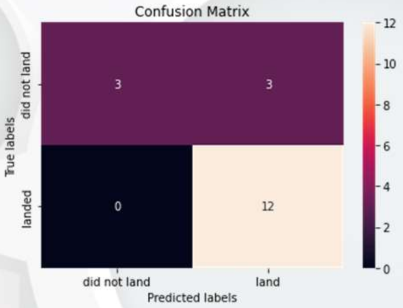
Plot Confusion Matrix for ML models

Compare accuracy for ML models using bar graph

- GitHub URL: https://github.com/ibrahim-beg/Applied-Data-Science-Capstone/blob/d7a34f852e9042cc739b3292295cf335810af22b/Machine%20Learning%20Prediction.ipynb

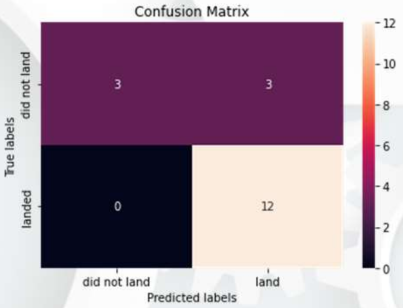
Predictive Analysis Results

- All four machine learning models had the same level of accuracy, so we need further analysis to determine which is the best model.



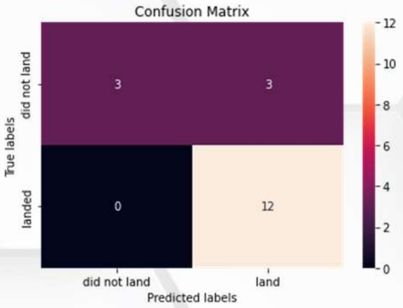
LogReg

Accuracy: 83.33%



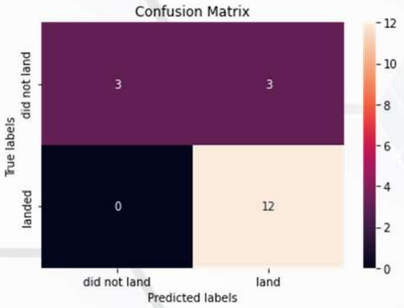
SVM

Accuracy: 83.33%



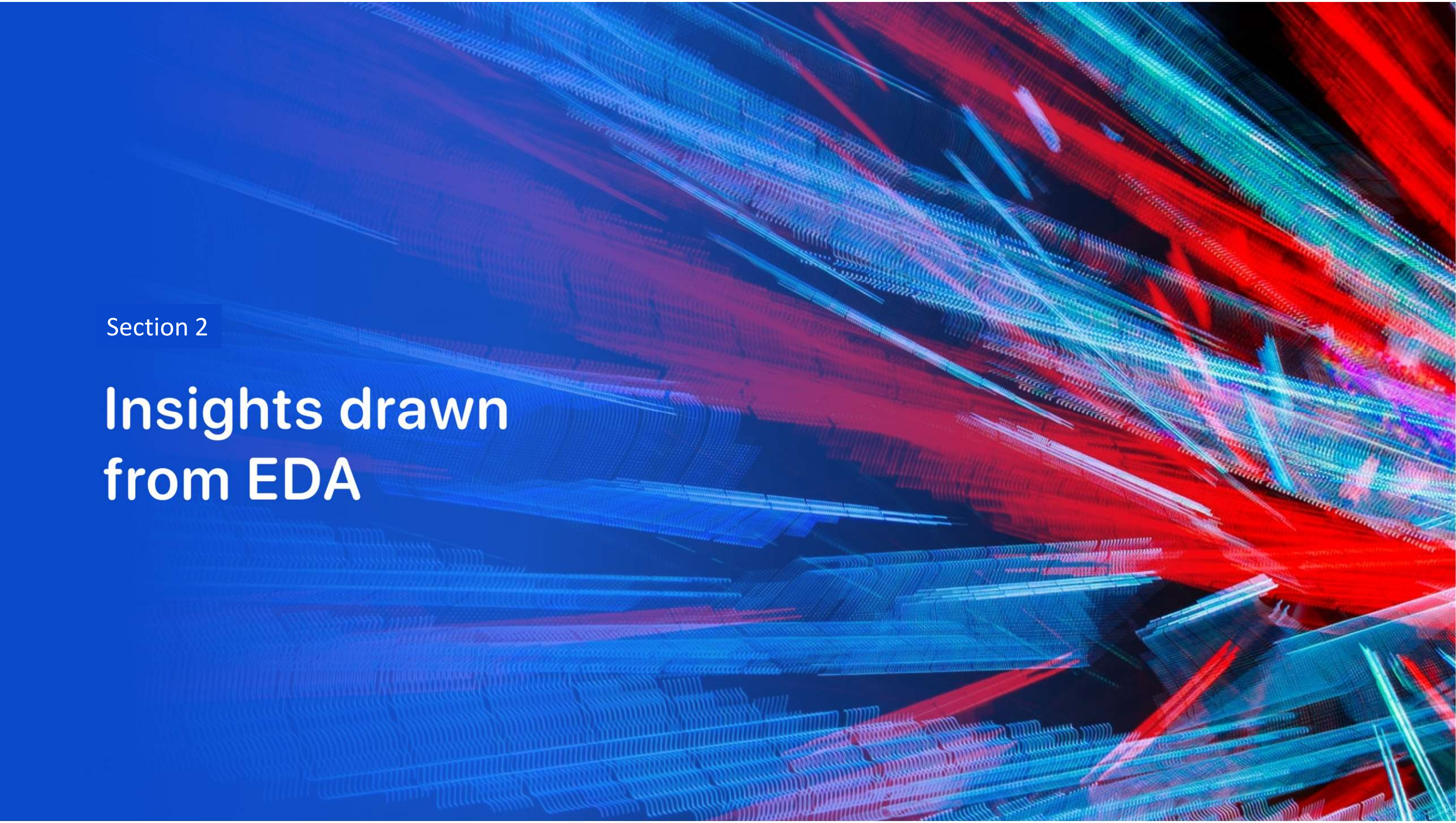
Decision Tree

Accuracy: 83.33%



KNN

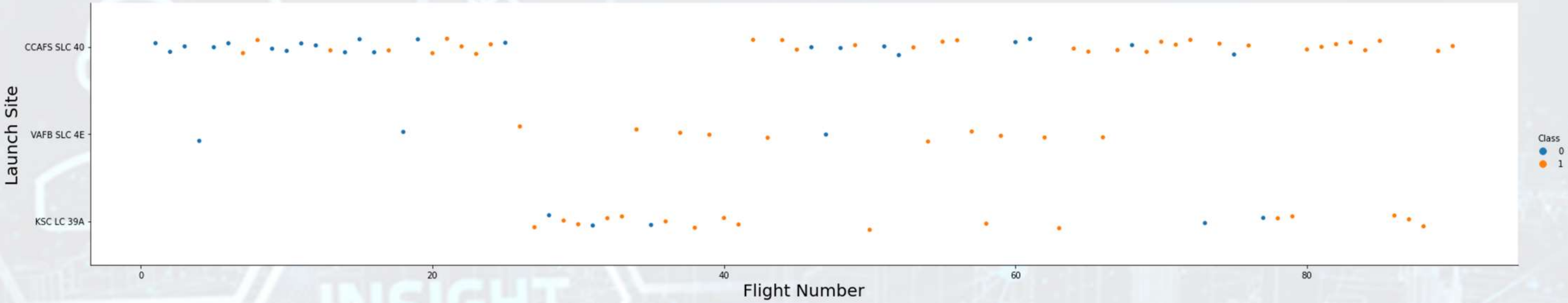
Accuracy: 83.33%



Section 2

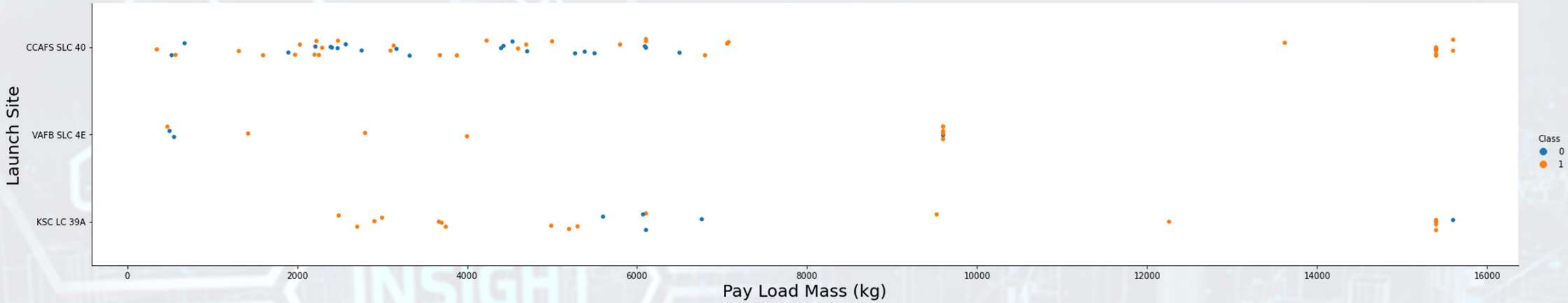
Insights drawn from EDA

Flight Number vs. Launch Site



- Successful launches (blue) and Failed launches (orange)
- Insights:
  - Main Launch Site is CCAFS as most launches take place here
  - As more flights are launched, increase in rate success rate for every site

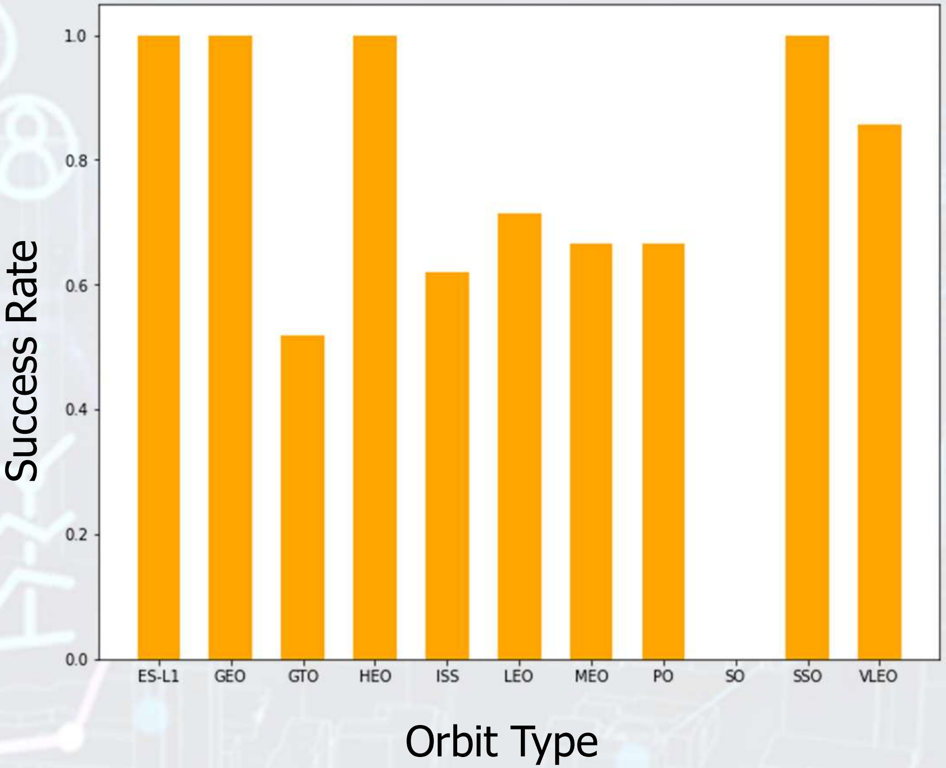
Payload vs. Launch Site



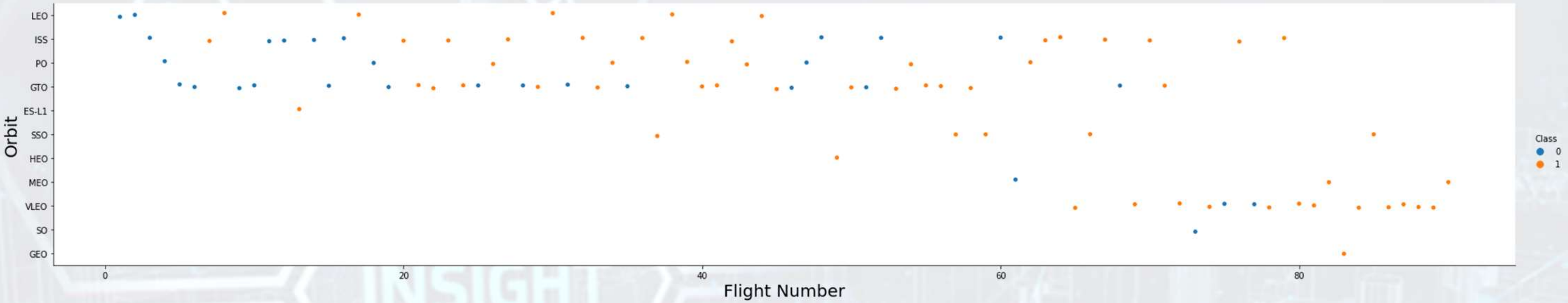
- Insights:
  - Majority of payload mass lie between 0 and 6000kg
  - Majority of payload mass above 6000kg have unsuccessful outcomes

Success Rate vs. Orbit Type

- GTO has largest number of launches but around 50% success rate
- VLEO has above 80% success with 14 launches
- SSO has 100% success rate with 5 launches
- The following sites have only had 1 success launch: ES-L1, GEO and HEO
- SO has a 0% success but has only had 1 launch

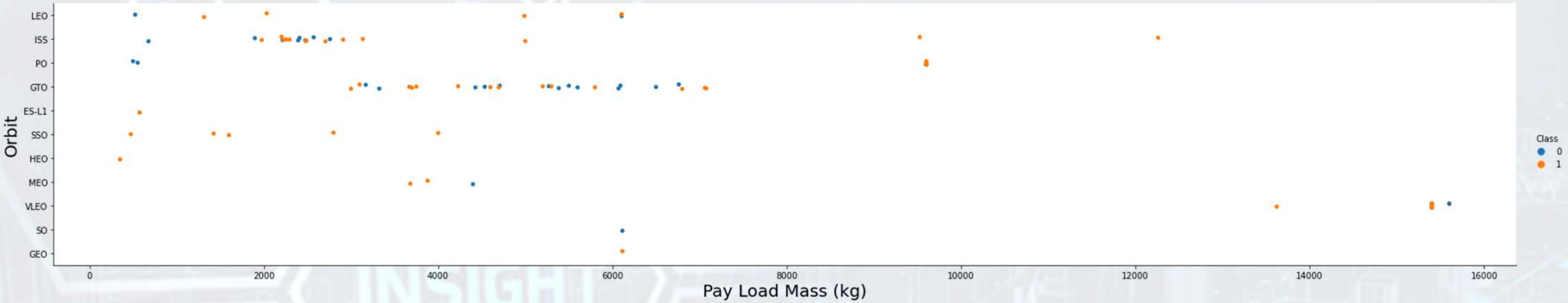


Flight Number vs. Orbit Type



- Insights:
  - The Orbit type preference for launches has shifted over time, this shift has resulted in more successful landings
  - The orbit which were sun-synchronous or lower-orbit seems to have the most success for SpaceX

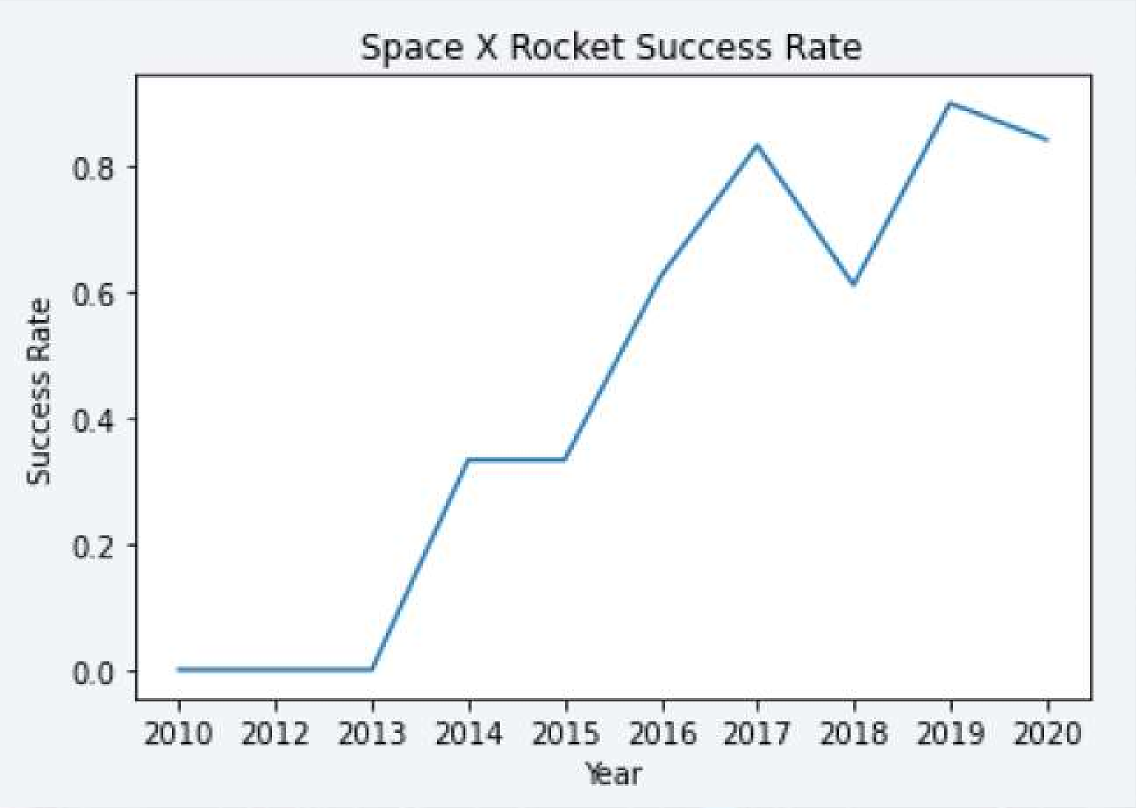
Payload vs. Orbit Type



- Insights:
  - Payload mass correlates with orbit types, for example certain orbit types are not able to handle payload mass over a certain limit (LEO and SSO)
  - VLEO demonstrates that it has some success with large payload masses

Launch Success Yearly Trend

- Since 2013 we see a positive trend, as the success rate has increased except for a small dip in 2018

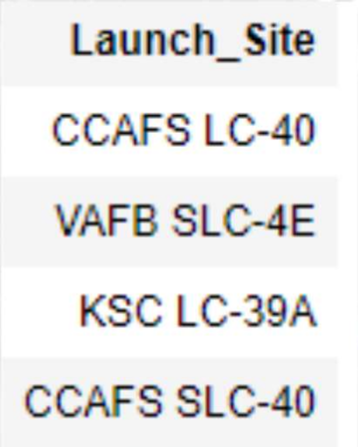


All Launch Site Names

SQL Query

```
%sql SELECT DISTINCT LAUNCH_SITE FROM
SPACEXDATASET;
```

Results



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

Explanation

- The use of DISTINCT in the query allows to remove duplicate LAUNCH_SITE.

Site Names Begin with 'CCA'

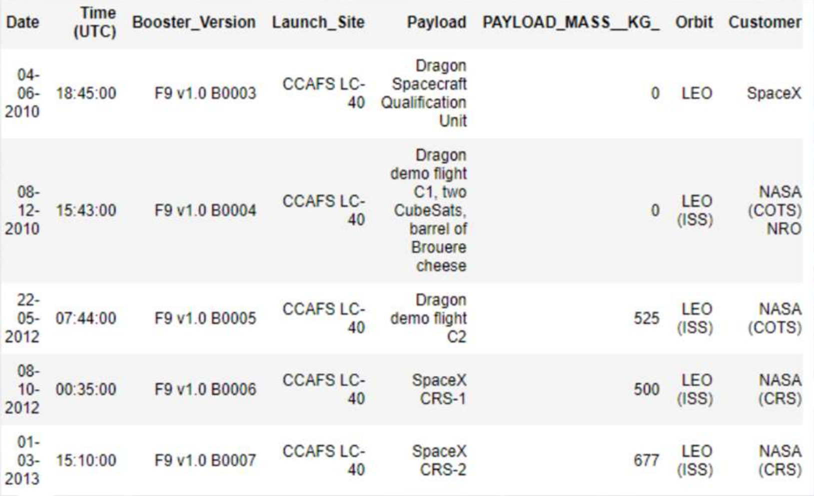
SQL Query

```
%sql SELECT * FROM SPACEXDATASET WHERE
LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Explanation

- The WHERE clause followed by LIKE clause filters launch sitesthat contain the substring CCA.
- LIMIT 5 shows 5 records from filtering.

Results



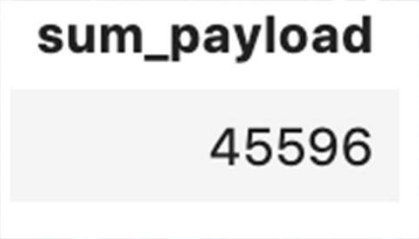
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Total Payload Mass

SQL Query

```
%sql SELECT SUM (payload_mass__kg_) AS
SUM_PAYLOAD FROM SPACEXDATASET WHERE
customer = 'NASA (CRS)';
```

Results



sum_payload
45596

Explanation

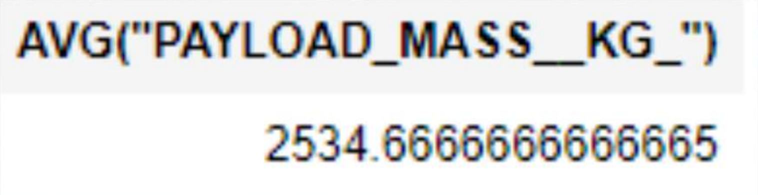
- This query returns the sum of all payload masses where the customer is NASA (CRS).

Average Payload Mass by F9 v1.1

SQL Query

```
%sql SELECT AVG (payload_mass__kg_) FROM
SPACEXDATASET WHERE ‘BOOSTER_VERSION’ LIKE
‘%F9 v1.1%';
```

Results



AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

Explanation

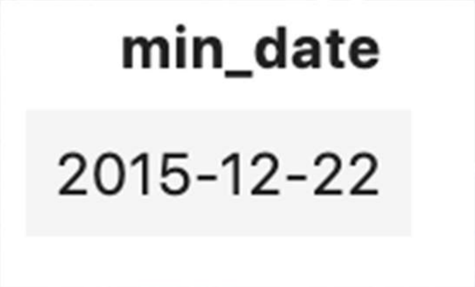
- This query returns the average of all payload masses where the booster version contains the substring F9 v1.1.

First Successful Ground Landing Date

SQL Query

```
%sql SELECT MIN(DATE) AS MIN_DATE FROM
SPACEXDATASET WHERE landing__outcome = 'Success
(ground pad)';
```

Results



```
min_date
2015-12-22
```

Explanation

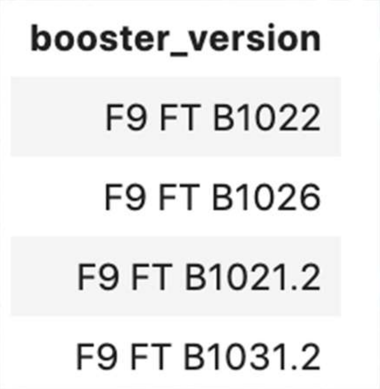
- With this query, we select the oldest successful landing.
- The WHERE clause filters dataset in order to keep only records where landing was successful.
- With the MIN function, we select the record with the oldest date.

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query

```
%sql SELECT booster_version FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)’
AND payload_mass__kg_ > '4000’ AND
payload_mass__kg_ < '6000’;
```

Results



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

Explanation

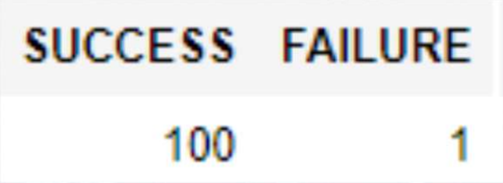
- This query returns the booster version where landing was successful and payload mass is between 4000 and 6000 kg. The WHERE and AND clauses filter the dataset.

Total Number of Successful and Failure Mission Outcomes

SQL Query

```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME")
FROM SPACEXDATASET WHERE "MISSION_OUTCOME"
LIKE '%Success%') AS SUCCESS, \ (SELECT
COUNT("MISSION_OUTCOME") FROM SPACEXDATASET
WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS
FAILURE;
```

Results



SUCCESS	FAILURE
100	1

Explanation

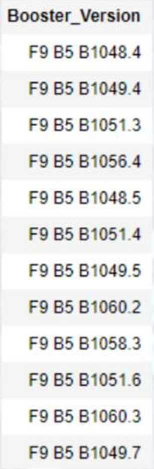
- With the first SELECT, we show the subqueries that return results. The first subquery counts the successful mission. The second subquery counts the unsuccessful mission.The WHERE clause followed by LIKE clause filters mission outcome. The COUNT function counts records filtered.

Boosters Carried Maximum Payload

SQL Query

```
%sql SELECT DISTINCT “BOOSTER_VERSION” FROM
SPACEXDATASET \ WHERE “PAYLOAD_MASS__KG_” =
(SELECT MAX(“PAYLOAD_MASS__KG_”) FROM
SPACEXDATASET ;
```

Results



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

Explanation

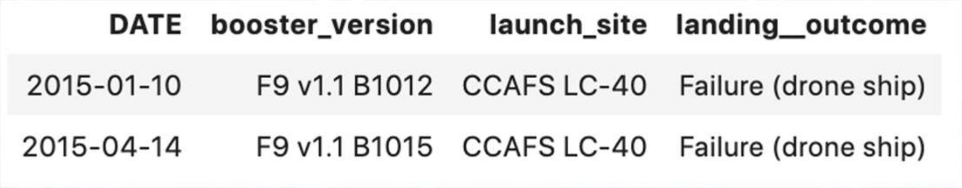
- We used a subquery to filter data by returning only the heaviest payload mass with MAX function. The main query uses subquery results and returns unique booster version (SELECT DISTINCT) with the heaviest payload mass.

2015 Launch Records

SQL Query

```
%sql SELECT Date, booster_version,
launch_site,landing__outcome FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)’ AND
YEAR(Date) = 2015;
```

Results



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

Explanation

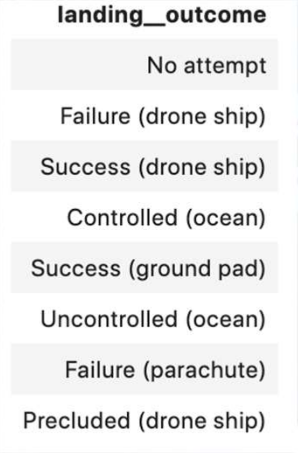
- This query returns month, booster version, launch site where landing was unsuccessful and landing date took place in2015. Substr function process date in order to take month or year. Substr(DATE, 4, 2) shows month. Substr(DATE,7, 4) shows year.

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

SQL Query

```
%sql SELECT landing__outcome FROM SPACEXDATASET
WHERE Date > '2010-06-04' AND Date < '2017-03-20'
GROUP BY landing__outcome ORDER BY
COUNT(landing__outcome) DESC;
```

Results



landing__outcome
No attempt
Failure (drone ship)
Success (drone ship)
Controlled (ocean)
Success (ground pad)
Uncontrolled (ocean)
Failure (parachute)
Precluded (drone ship)

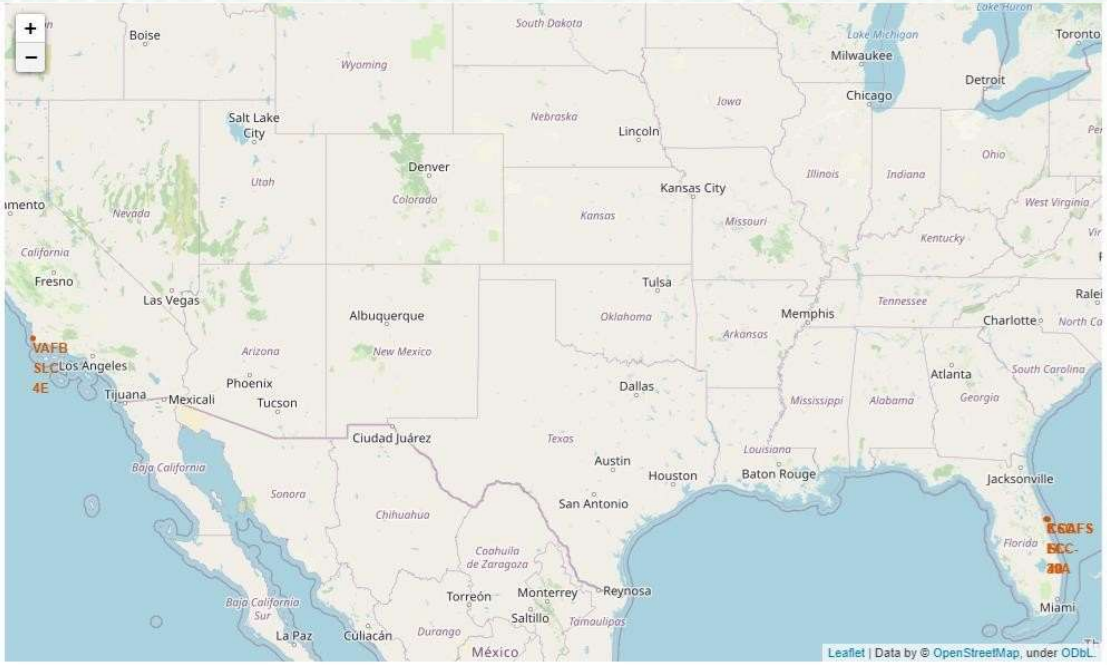
Explanation

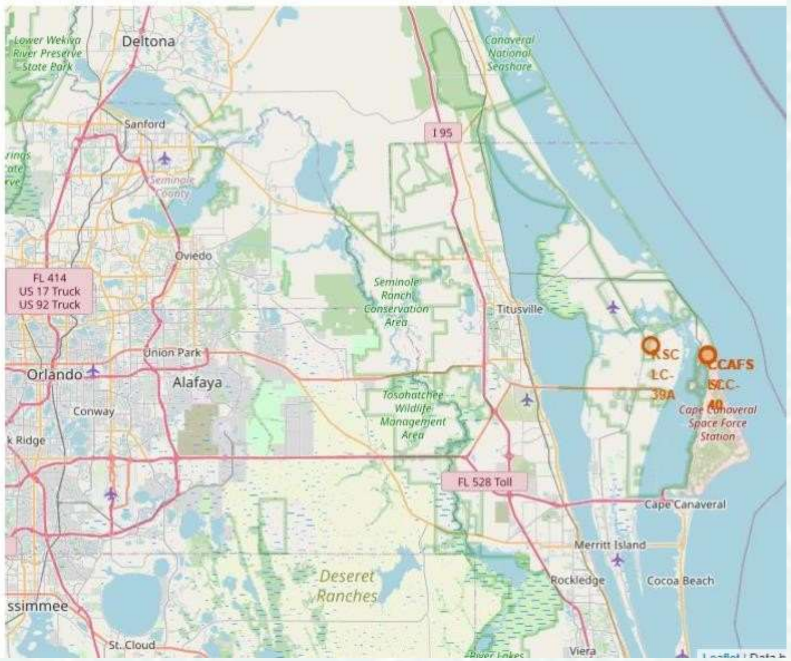
- The GROUP BY clause groups results by landing outcome and ORDER BY COUNTDESC shows results in decreasing order.

Section 3

Launch Sites Proximities Analysis

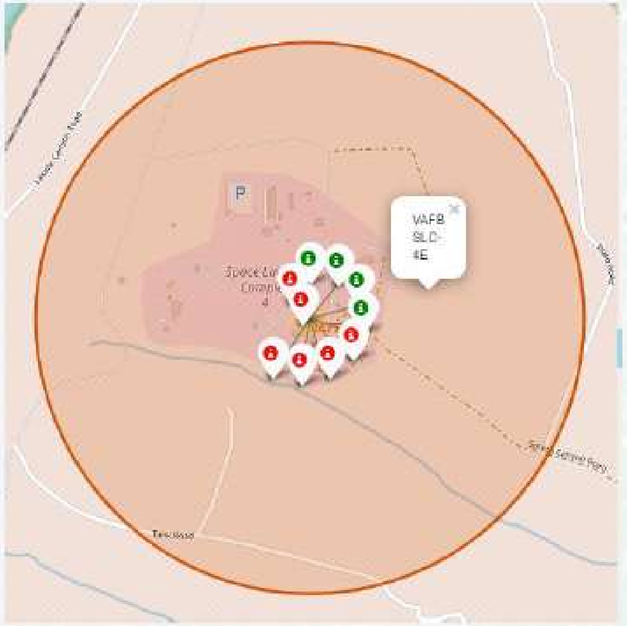
Folium Map – Launch Sites





- The map on the left shows every launch site together on one map
- The map on the right only shows the Florida launch sites as they are relatively close
- All launch sites are stationed near the ocean

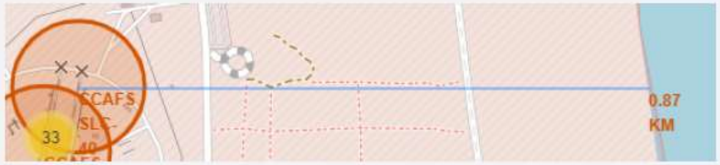
Folium Map – Colour Coded Launch Markers



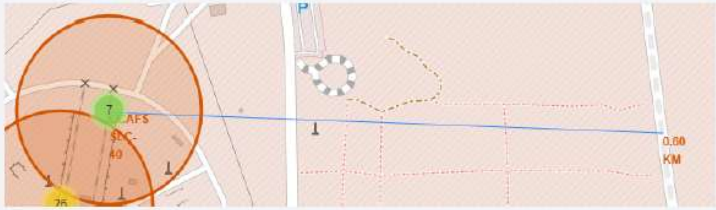
- The cluster of launches from VAFB-SLC-4E, the successful launches have green markers and the unsuccessful launches have red markers.

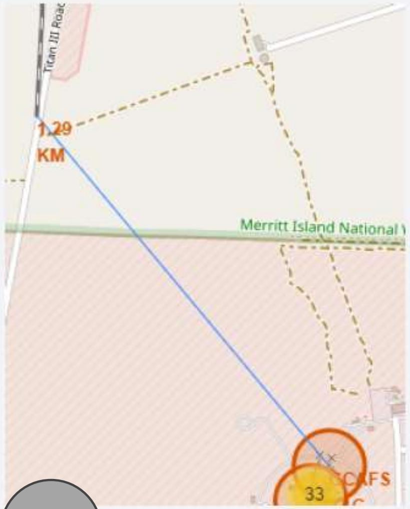
Folium Map – Key Location Proximities

1

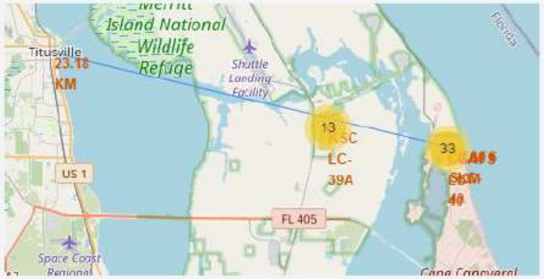


2



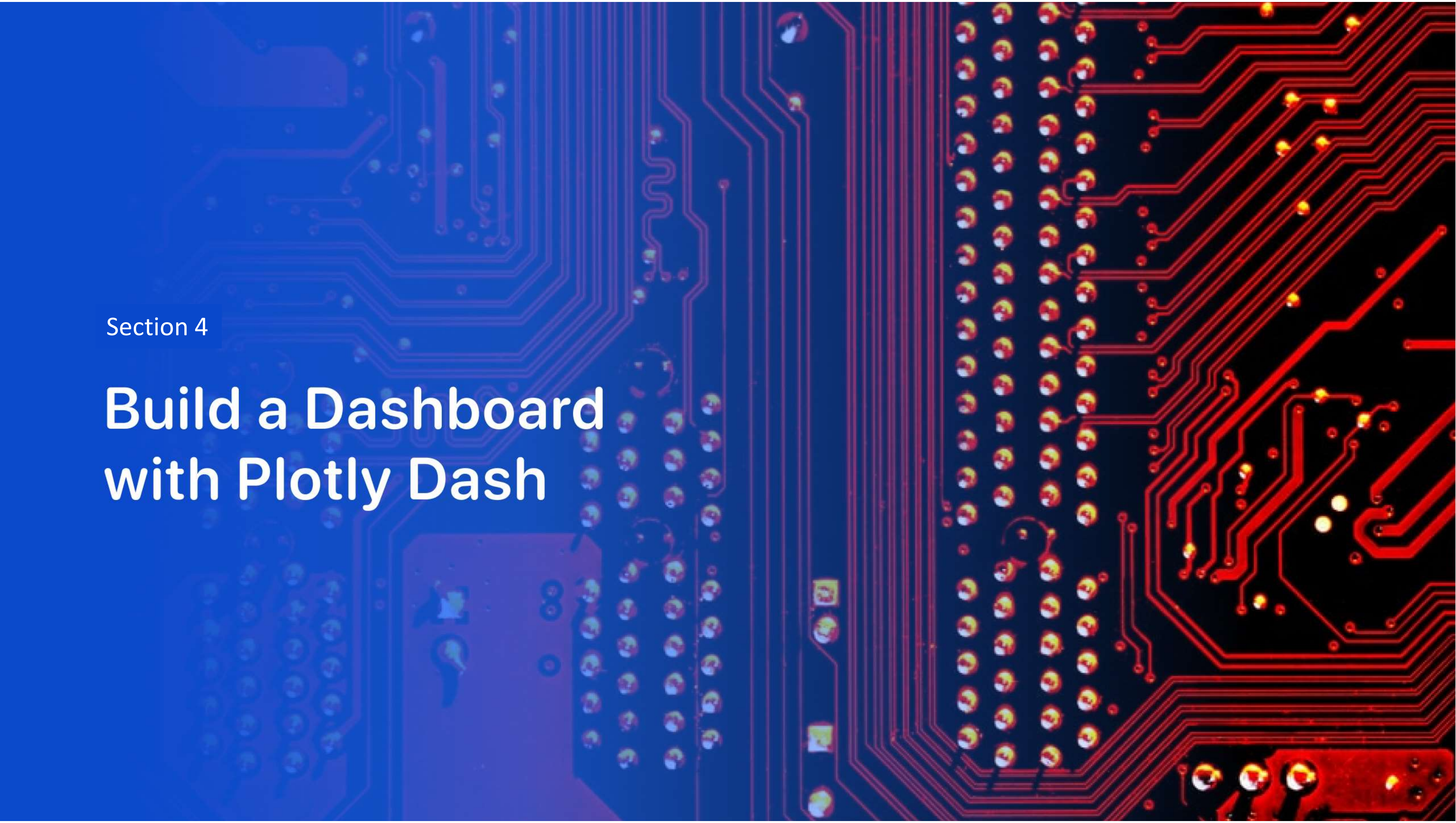


3



4

- Observe the launch site CCFAS SLC-40. In picture 1, we can see that it is close to the coastline and in picture 2 we can see that it close to a highway.
- The launch site CCFAS-SLC-40 is also close to a railway as can be seen in picture 3, however it far away from cities (over 20km to nearest city see picture 4)

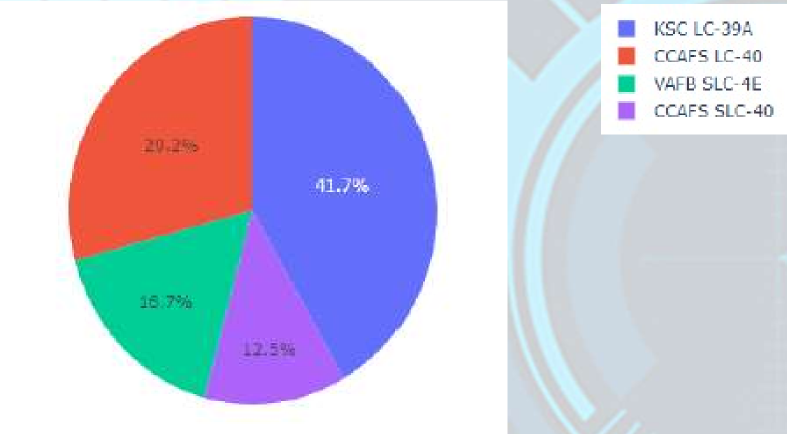


Section 4

Build a Dashboard with Plotly Dash

Launch Success Count

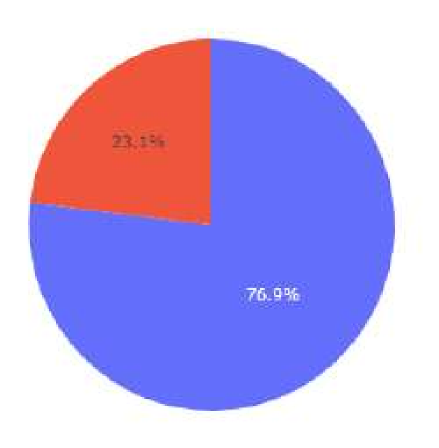
Total Successful Launches by Site



- KSC LC-39A has the greatest number of successful missions, and CCAFS SLC-40 has the fewest. It is important to note the total number of launches per site otherwise this information could be deceiving, KSC LC-39A could have the greatest number of failed missions aswell.

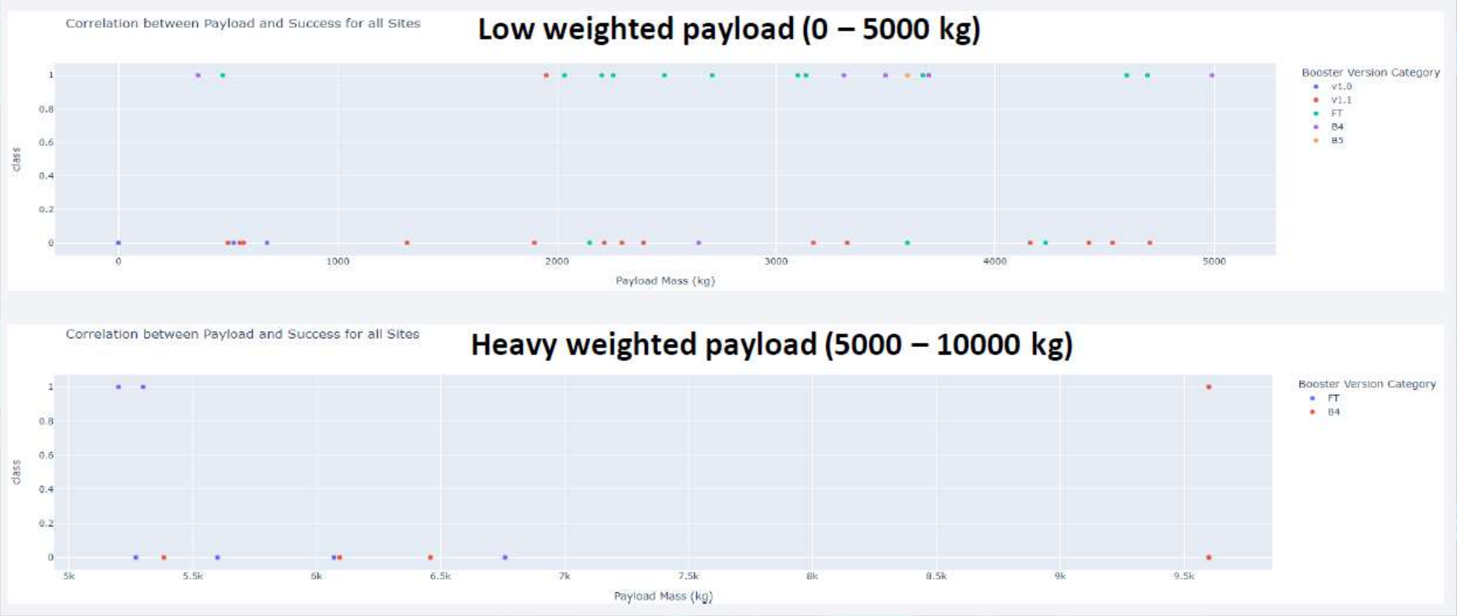
Total Successful launches for Site KSC LC-39A

Total Successful Launches by Site



- The blue is successful launches and the red is failure. We can observe this site has great success rate considering the total number of launches at this site.

Payload Mass vs. Outcomes for all site



Low weighted payload (0 – 5000 kg)

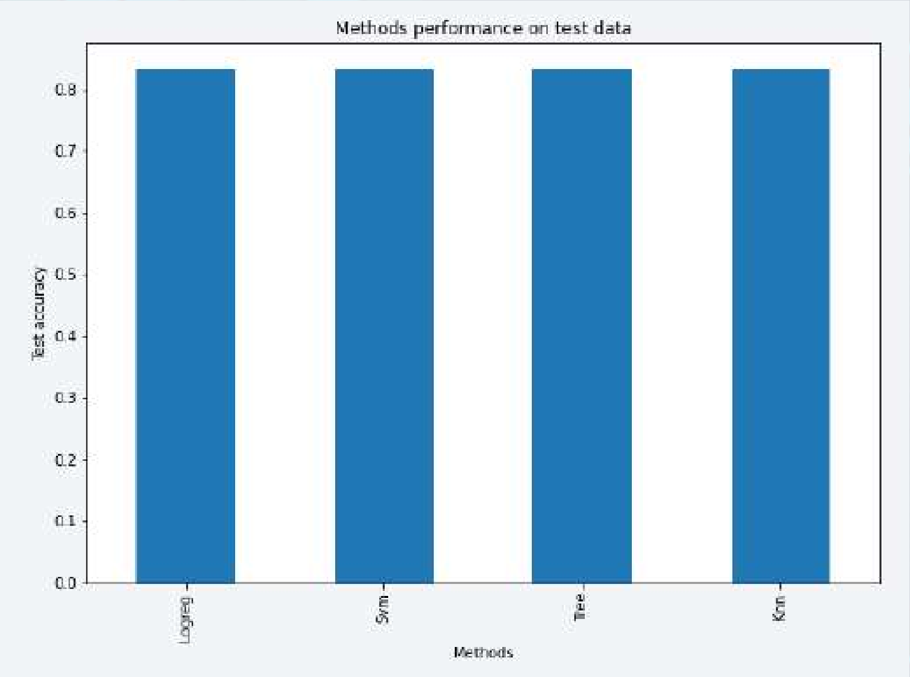
Heavy weighted payload (5000 – 10000 kg)

- A hardly difficult observation, which we could have guessed, is that low weighted payload have a better success rate than heavy weighted payloads. For us to compete with SpaceX, SpaceY needs to become successful at missions with heavier payload masses.

Section 5

Predictive Analysis (Classification)

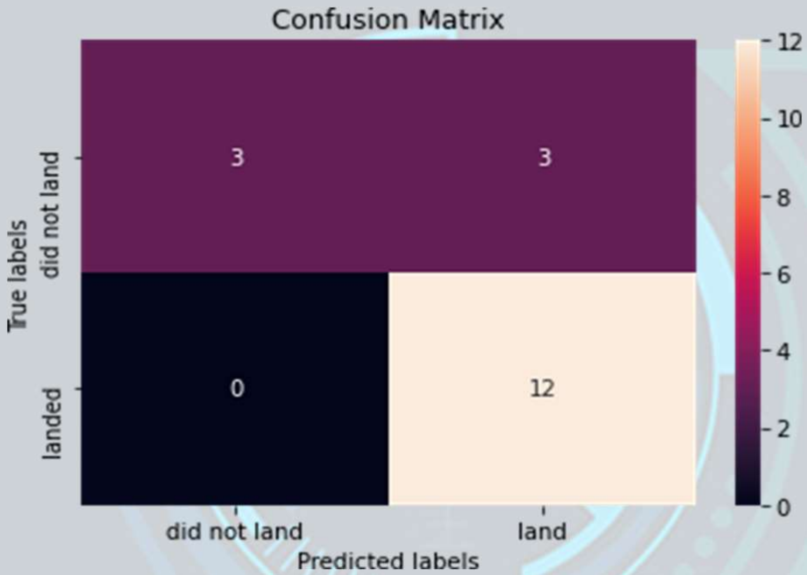
Classification Accuracy



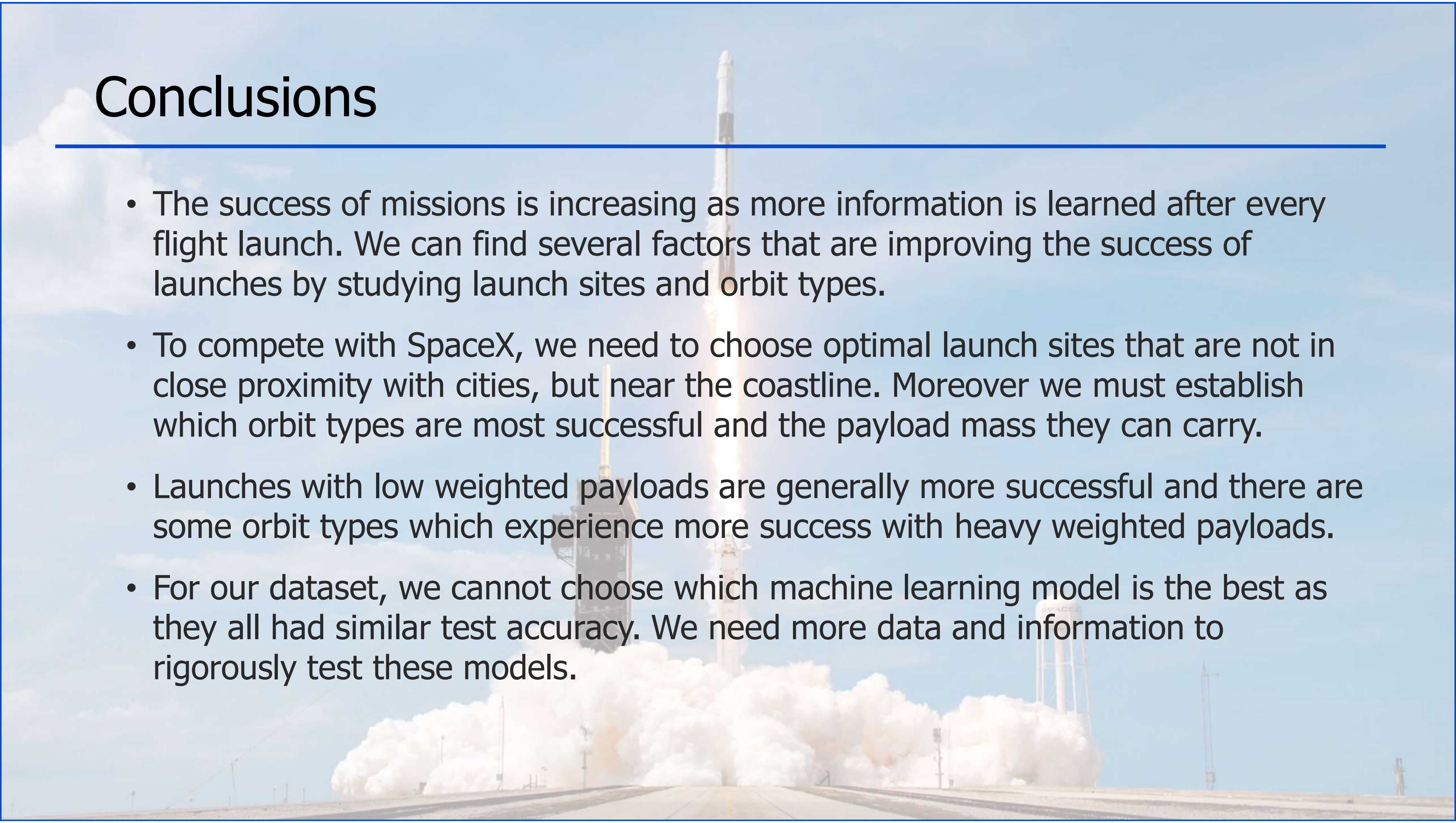
Model	Training data accuracy	Test dataset accuracy
Logistic Regression	0.846429	0.833333
Support Vector Machine	0.848214	0.833333
Decision Tree	0.876786	0.833333
K-nearest neighbours	0.848214	0.833333

- For the test dataset, all models had the same accuracy. We need more data to determine which is the best model.
- However on the training dataset the Decision Tree Classifier had the highest accuracy.
- But we must note that the sample size for our models is very small, so we need to conduct further research to decide the best model to implement.

Confusion Matrix



- The confusion matrix for all four models are identical because the test accuracy was the same for all of them.
- All the models has zero false negatives, however an issue they all had was 3 false positives. We will need to fine tune the models to obtain the best results.



Conclusions

- The success of missions is increasing as more information is learned after every flight launch. We can find several factors that are improving the success of launches by studying launch sites and orbit types.
- To compete with SpaceX, we need to choose optimal launch sites that are not in close proximity with cities, but near the coastline. Moreover we must establish which orbit types are most successful and the payload mass they can carry.
- Launches with low weighted payloads are generally more successful and there are some orbit types which experience more success with heavy weighted payloads.
- For our dataset, we cannot choose which machine learning model is the best as they all had similar test accuracy. We need more data and information to rigorously test these models.

Thank you!