Winning Space Race with Data Science

Liam Stuart
30/09/2024

Outline

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

Executive Summary

- The methodologies used to analyze the data included:
  - Data Collection using the SpaceX REST API and Web Scraping;
  - Exploratory Data Analysis (EDA) using SQL and Data Visualisation;
  - Visual Analytics using Folium and Dash;
  - Machine Learning Prediction.
- Summary of results:
  - EDA Results.
  - Data Visualisation and Interactive Analytics Results.
  - Predictive Analysis Using Machine Learning.

Introduction

- We are assuming the role of a data scientist working for SpaceY, a competitor against SpaceX, and are being asked to predict the price for the launch of a rocket. SpaceX are able to keep costs lower than competitors by reusing the first stage of their launch if the first stage lands.

- The main goal for the project is to predict whether or not the first stage of a Falcon 9 SpaceX rocket will land successfully given prior information about it, such as its Booster Version or Launch Site.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data was collected via the following means:
        - The SpaceX REST API (https://api.spacexdata.com/v4/rockets/)
        - Web Scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
    - The data was then preprocessed to prepare it for the Machine Learning phase. This included cleaning null values, removing irrelevant columns, and creating a new outcome label based on outcome data.

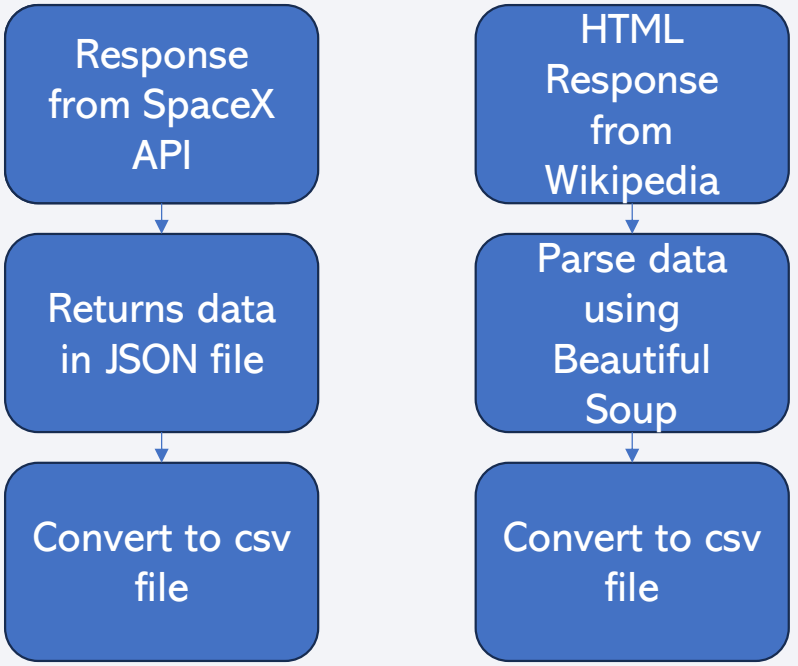
Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models:
  - The data was normalised and then split into training and test sets. Logistic Regression, SVC, Decision Tree, and KNN models were all trained across various parameters and their accuracies evaluated on the test set.

Data Collection

- The two main data collection methods utilized the SpaceX REST API and Web Scraping from Wikipedia.
- The SpaceX API returns the data as a JSON file, which can then be converted into a csv file for further processing.
- The BeautifulSoup package was used for web scraping, where the html data was extracted and then converted to a csv file.



Data Collection – SpaceX API

- Data collection using the SpaceX REST API calls.
- Code: (https://github.com/liam-stuart/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb)

1. Response from API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"

response = requests.get(spacex_url)
```

2. Convert to JSON File, then Pandas DataFrame

```
response=requests.get(static_json_url).json()
data=pd.json_normalize(response)
```

3. Apply functions to extract information

```
getBoosterVersion(data)
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

4. Create new Pandas DataFrame

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

```
data2=pd.DataFrame(launch_dict)
```

5. Filter to Falcon 9 launches then convert to a csv file

```
data_falcon9=data2[data2['BoosterVersion']=='Falcon 9']

data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

Data Collection - Scraping

- Data collection using web scraping from Wikipedia.
- Code: (https://github.com/liam-stuart/Applied-Data-Science-Capstone/blob/main/Web%20Scraping.ipynb)

1. Response from HTML, Creation of BeautifulSoup Object

```
response=requests.get(static_url)
soup=BeautifulSoup(response.content,parser='html.parser')
```

2. Finding tables and extracting column names

```
html_tables=soup.find_all('tr')
column_names = []
cols=first_launch_table.find_all('th')
for c in cols:
    column=extract_column_from_header(c)
    column_names.append(column)
```

3. Convert to Pandas DataFrame, append data (see code for further details)

```
launch_dict= dict.fromkeys(column_names)
del launch_dict['Date and time ( )']
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

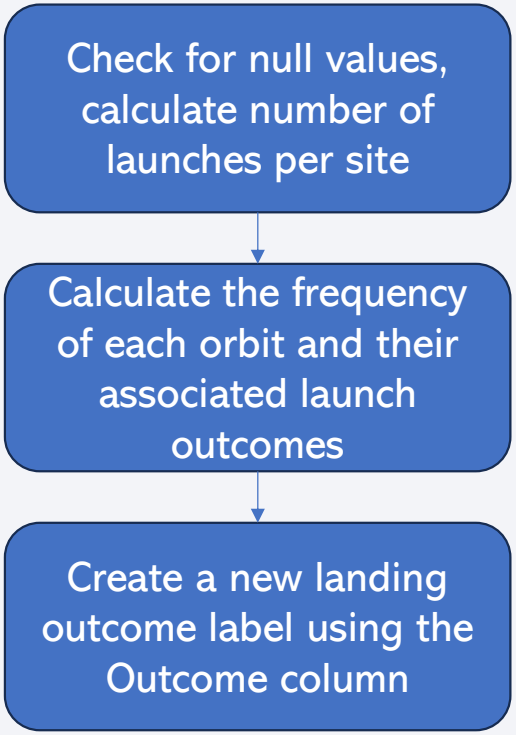
```
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

4. Convert to csv file

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

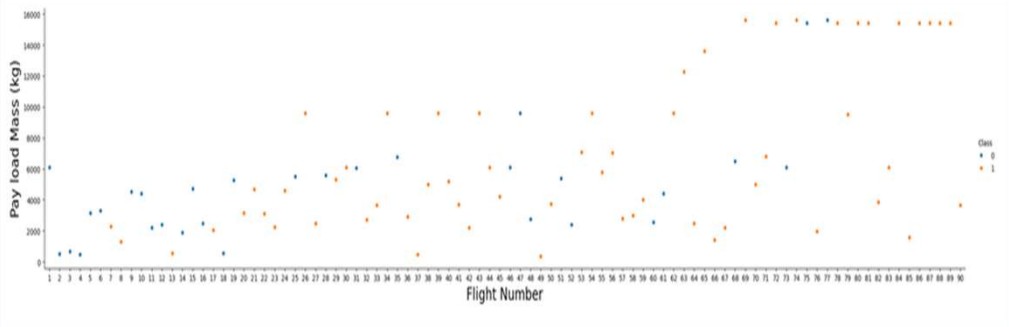
Data Wrangling

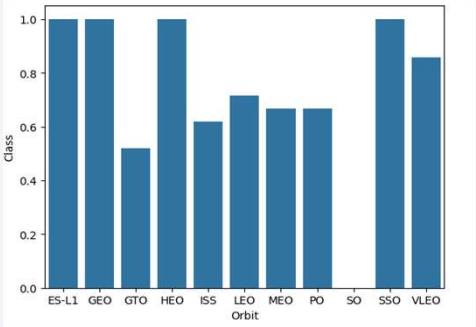
- Flowchart of the data processing stage, including some initial EDA
- Code: (https://github.com/liam-stuart/Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb)

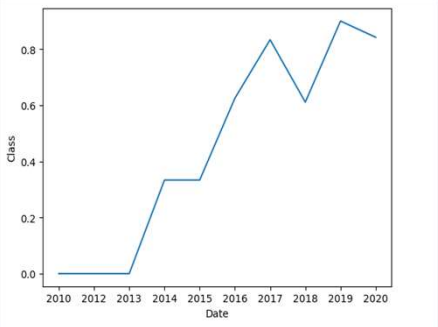


EDA with Data Visualization

- To get a better grasp on the relationships between certain variables, charts were generated to visualize the data. Some are displayed below:







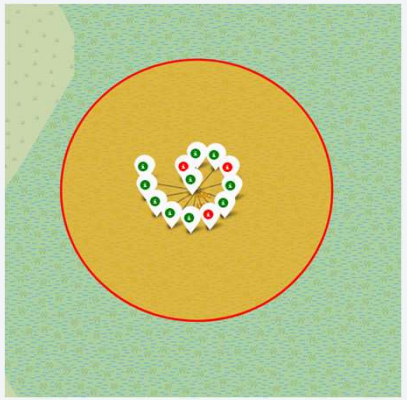
- One immediate conclusion is that over time, the overall success rate of launches showed an increasing trend.
- Code: (https://github.com/liam-stuart/Applied-Data-Science-Capstone/blob/main/Data%20Visualisation.ipynb)

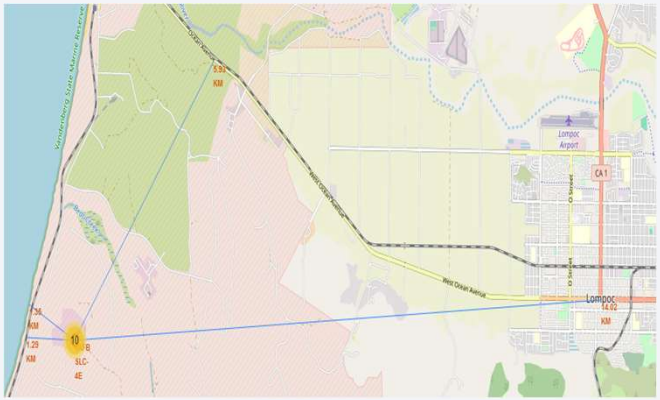
EDA with SQL

- The SQL queries performed displayed the following:
  - Unique launch site names;
  - 5 records where launch sites began with 'CCA';
  - The total payload mass carried by NASA (CRS) boosters;
  - The average payload mass carried by booster version F9 v1.1;
  - The date when the first successful landing outcome in ground pad was achieved;
  - The names of the boosters which have success in drone ship and payload mass between 4000kg and 6000kg;
  - The total number of successful and failure mission outcomes;
  - The names of booster versions which have carried the maximum payload mass;
  - The records which will display the month names, failure landing outcomes in drone ship, booster versions, and launch site for the months in the year 2015;
  - Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.
- Code: (https://github.com/liam-stuart/Applied-Data-Science-Capstone/blob/main/SQL.ipynb)

Build an Interactive Map with Folium

- Map features have been created to identify the optimal launch site location, such as proximity to certain features and success rate for each site.

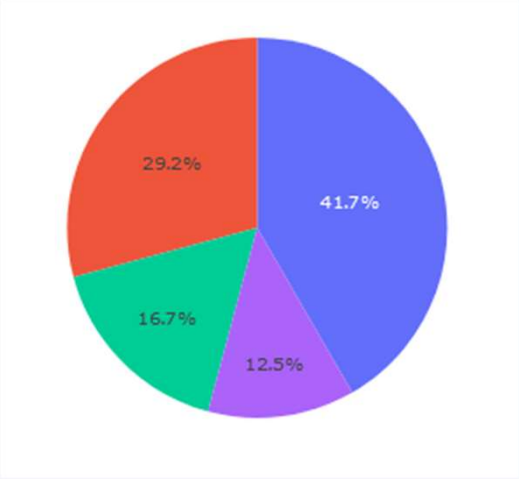


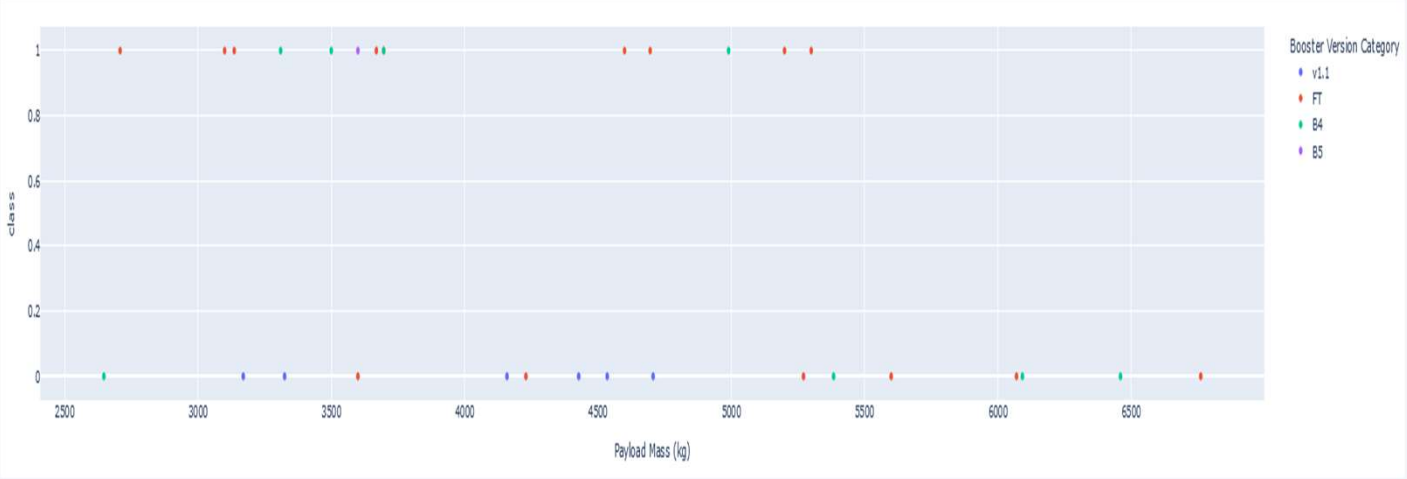


- Code: (https://nbviewer.org/github/liam-stuart/Applied-Data-Science-Capstone/blob/main/Launch%20Site%20Maps.ipynb)

Build a Dashboard with Plotly Dash

- An interactive dashboard has been created using Plotly Dash to easily compare Launch Site success rates, and how Payload Mass affects this.

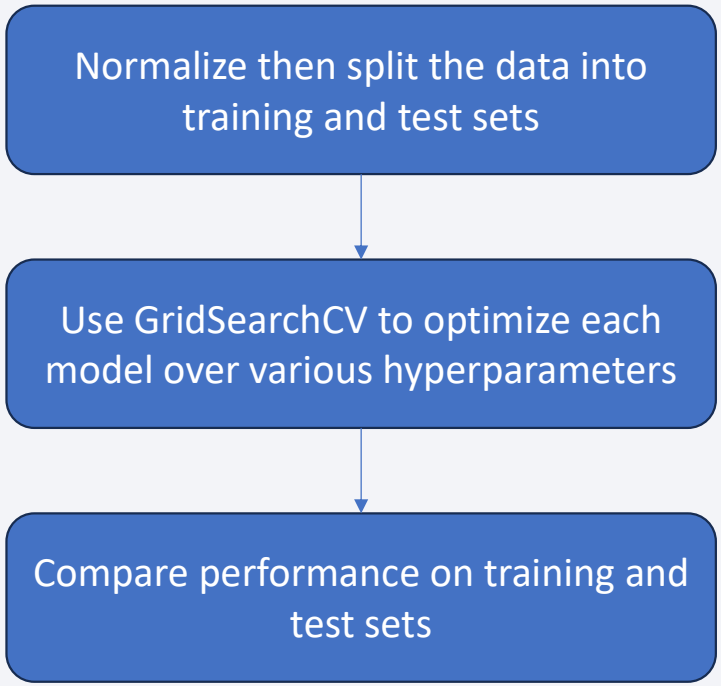




- Code: (https://github.com/liam-stuart/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py)

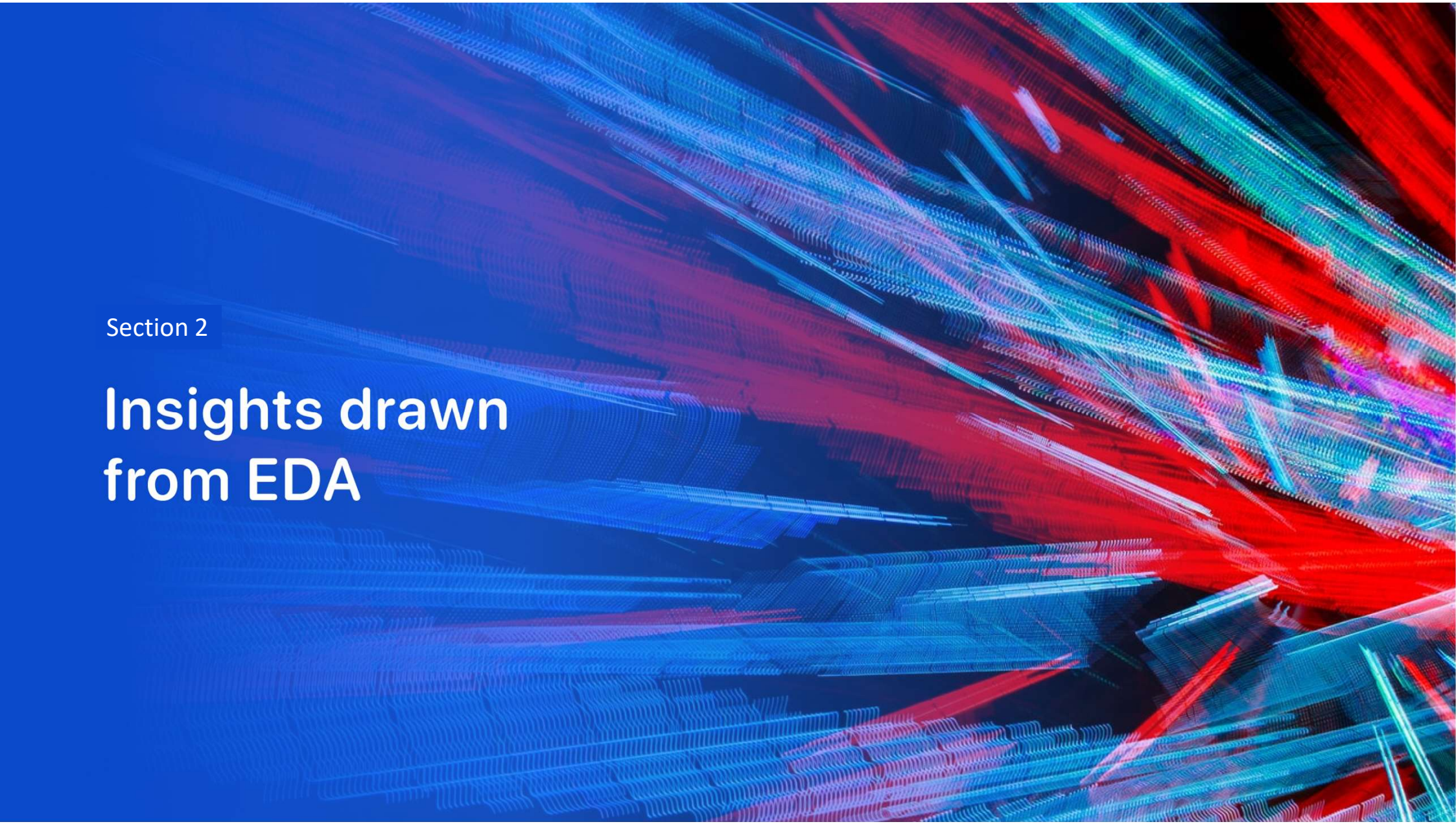
Predictive Analysis (Classification)

- After the data was analyzed, machine learning was then utilized to train models on the data.
- Code: (https://github.com/liam-stuart/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb)

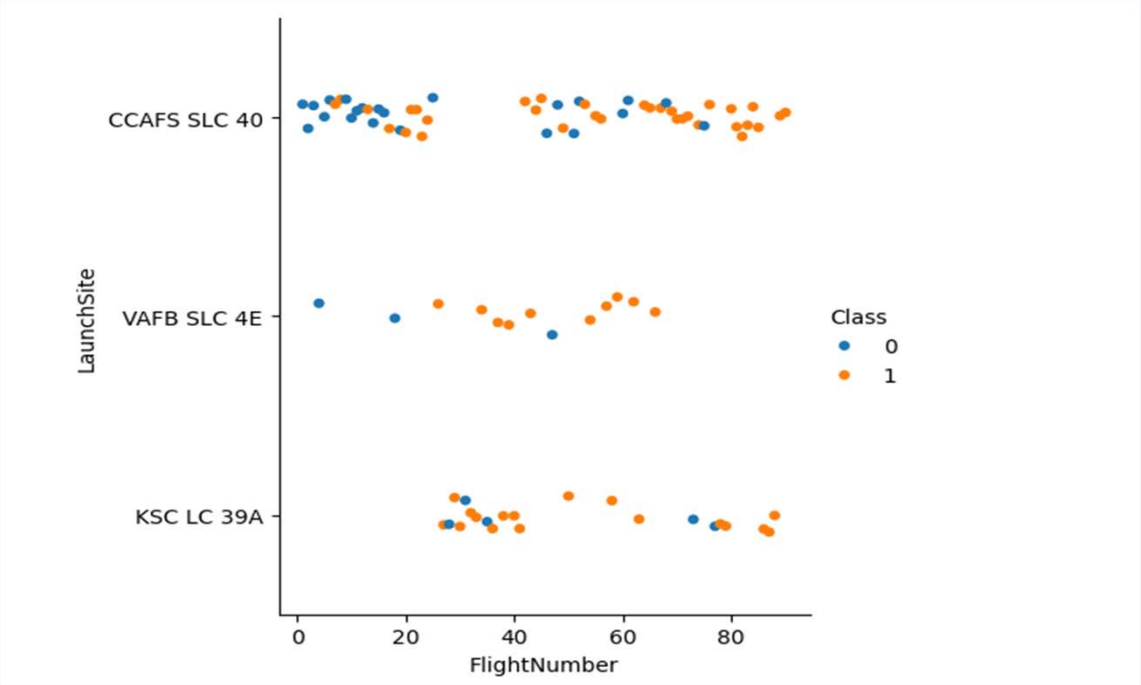


Results

- The Decision Tree model ended up performing best on both the training and test sets.
- Success rates for SpaceX launches increased over time, likely due to more experience garnered from each launch and learning from mistakes.
- Good launch site locations were coastal and near railways for easy access to rocket components, and far away from civilian infrastructure like cities and highways to minimize risk from failed launches.
- Overall, KSC LC-39A had the highest success rate for launches.

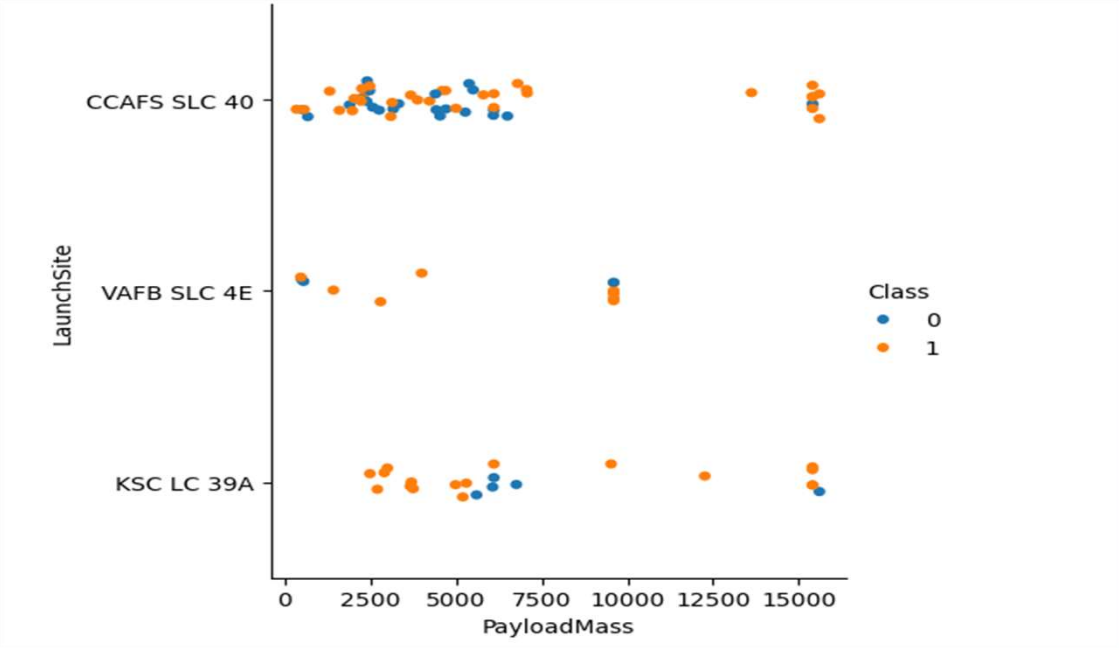

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



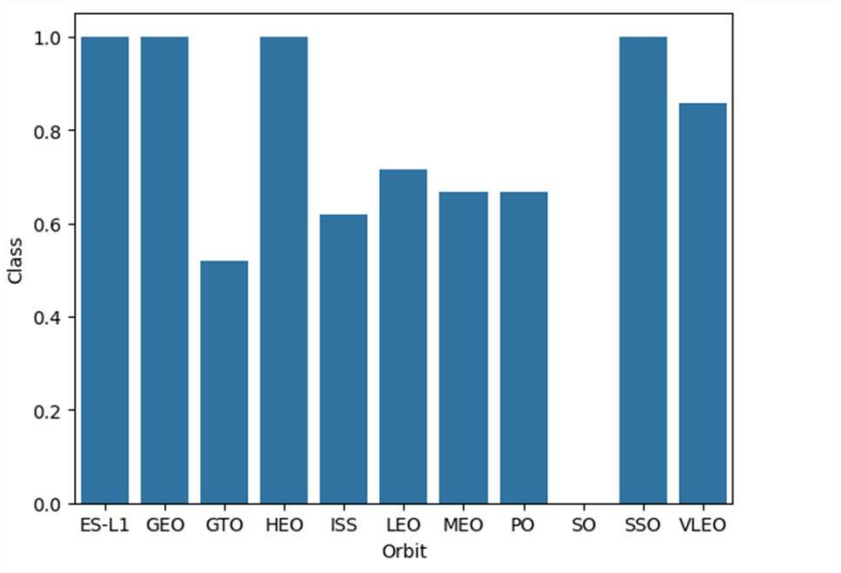
- We can see that as flight number increases, the frequency of successful launches also increases. This trend holds for all sites.

Payload vs. Launch Site



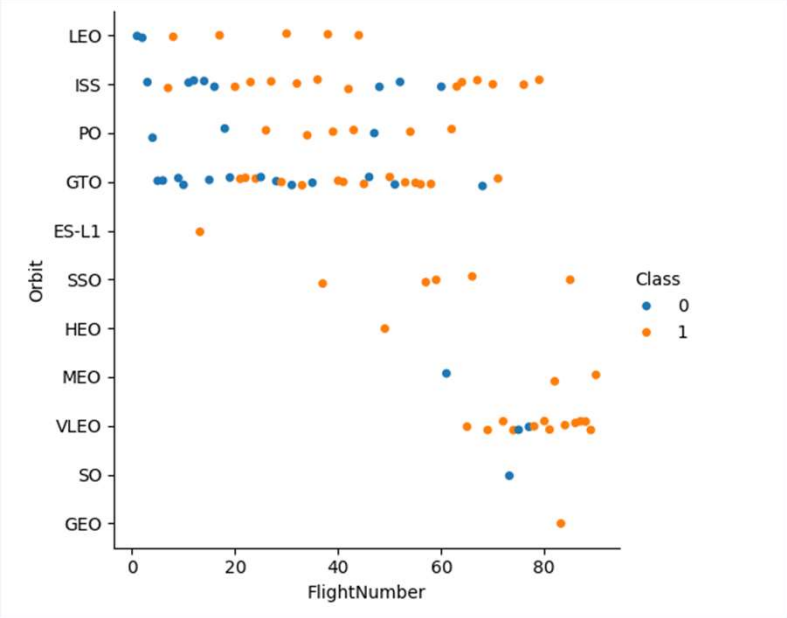
- There does not appear to be a clear relationship between success rate and payload mass. Most of the failures on low masses are localised to CCAFS SLC 40.

Success Rate vs. Orbit Type



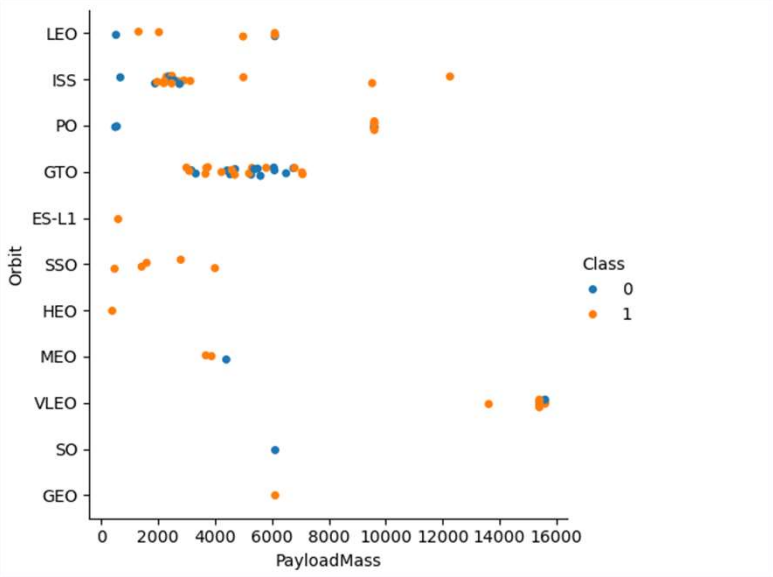
- ES-L1, GEO, HEO and SSO appear to have the highest success rates in terms of orbit, but it should be noted that some of these orbits only appear once or twice in the dataset, skewing their percentages.

Flight Number vs. Orbit Type



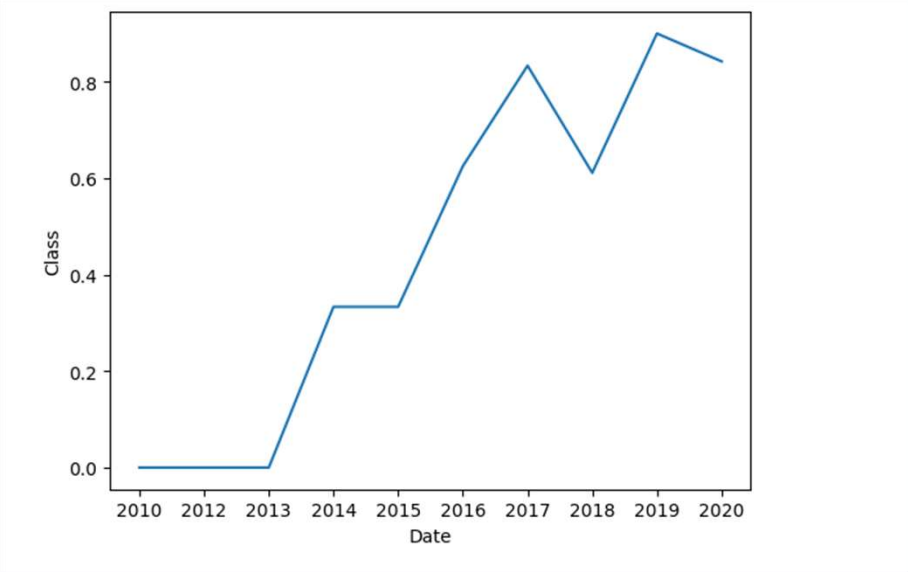
- The graph shows a clear shift towards higher orbit launches as the flight number increases.

Payload vs. Orbit Type



- The graph shows that ISS orbits in the 2000-4000kg range and GTO orbits in the 2000-8000kg range are by far the most common types of launches.

Launch Success Yearly Trend



- The overall success rate shows a clear upward trend over the years since 2013.

All Launch Site Names

- %sql SELECT "Launch_Site" FROM SPACEXTABLE GROUP BY "Launch_Site"

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

A list of all of the unique launch site names.

Launch Site Names Begin with 'CCA'

- %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- A list of 5 different records where the launch site starts with ‘CCA’.

Total Payload Mass

- %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Customer"=="NASA (CRS)“

SUM(PAYLOAD_MASS__KG_)
45596

- The total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

- %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Booster_Version"=="F9 v1.1"Present your query result with a short explanation here

AVG(PAYLOAD_MASS__KG_)
2928.4

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

First Successful Ground Landing Date

- %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE "Landing_Outcome"=="Success (ground pad)“

MIN(Date)
2015-12-22

- The date of the first successful landing outcome in ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT Booster_Version FROM SPACEXTABLE WHERE ("Landing_Outcome"=="Success (drone ship)" AND 4000<"PAYLOAD_MASS__KG_" AND "PAYLOAD_MASS__KG_"<6000)

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

- The names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

- %sql SELECT Mission_Outcome,COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome

Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- The total number of successful and failed mission outcomes.

Boosters Carried Maximum Payload

- %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

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

- List of booster versions which have carried the maximum payload.

2015 Launch Records

- %sql SELECT substr(Date,6,2) as Month,Landing_Outcome,Booster_Version,Launch_Site FROM SPACEXTABLE WHERE (substr(Date,0,5)=='2015' and Landing_Outcome=='Failure (drone ship)')

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

- List of records in year 2015 with landing outcome being a failure (drone ship).

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

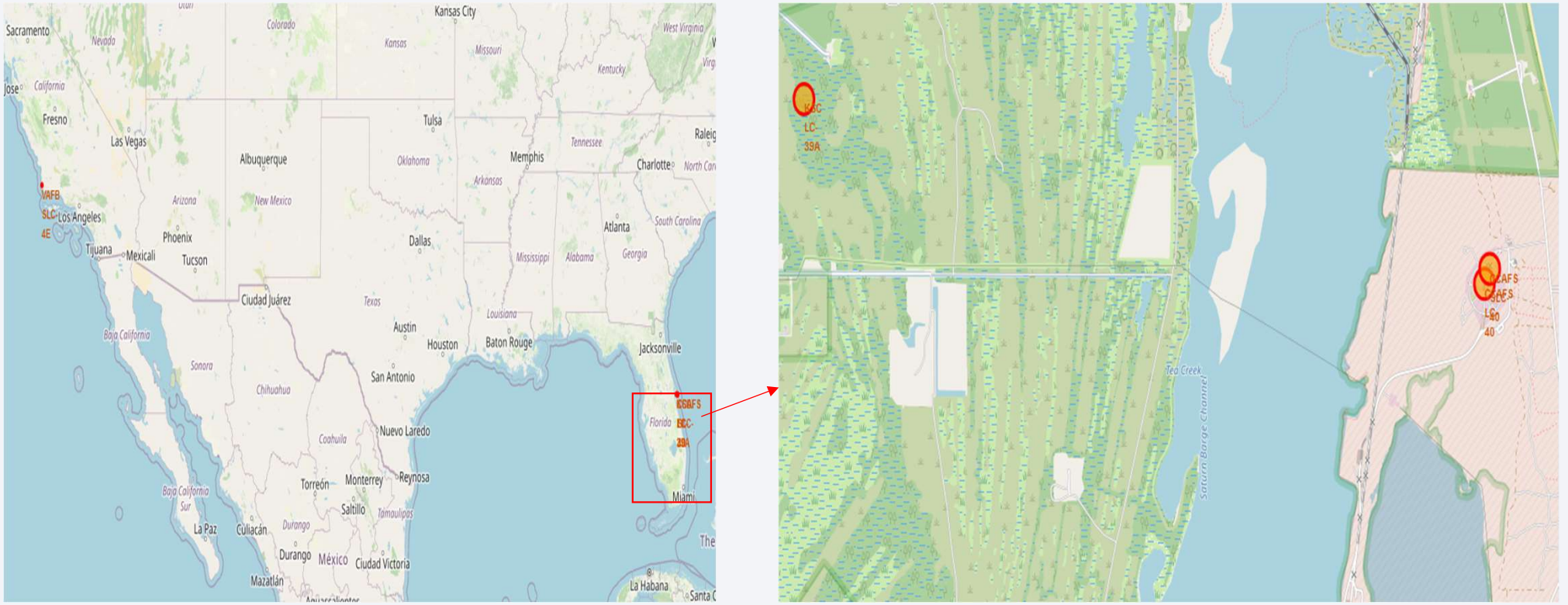
- %sql SELECT Landing_Outcome,COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE (Date>'2010-06-04' AND Date<'2017-03-20') GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC

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

- Count of all landing outcomes in descending order.


Section 3
Launch Sites
Proximities Analysis

Location Markers for Each Launch Site

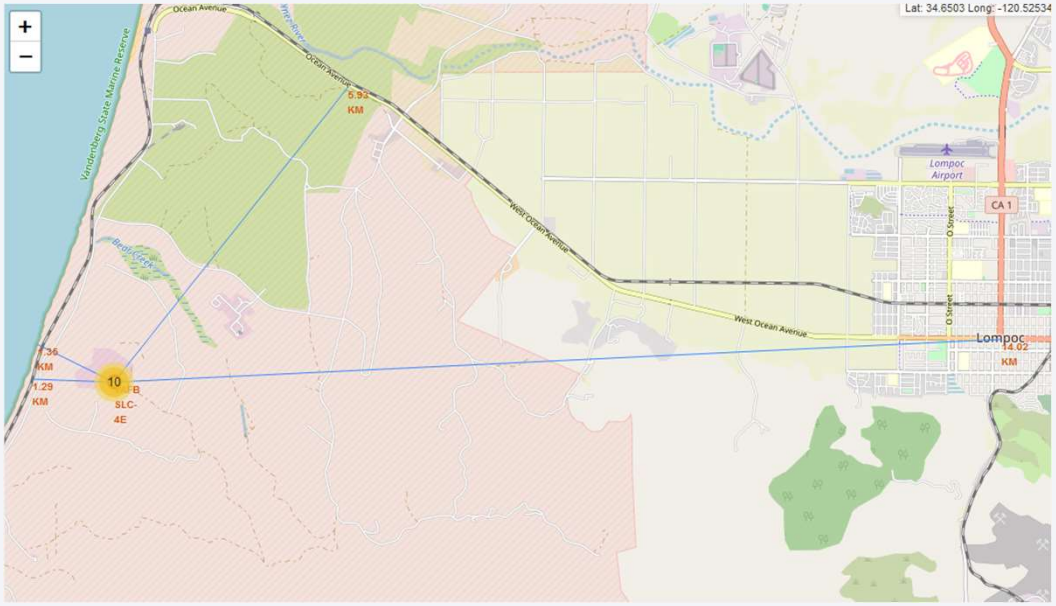


- A map showing all location markers. Due to the close proximity of three of them, a zoom in of that region has also been provided. Clear trend towards coastal locations.

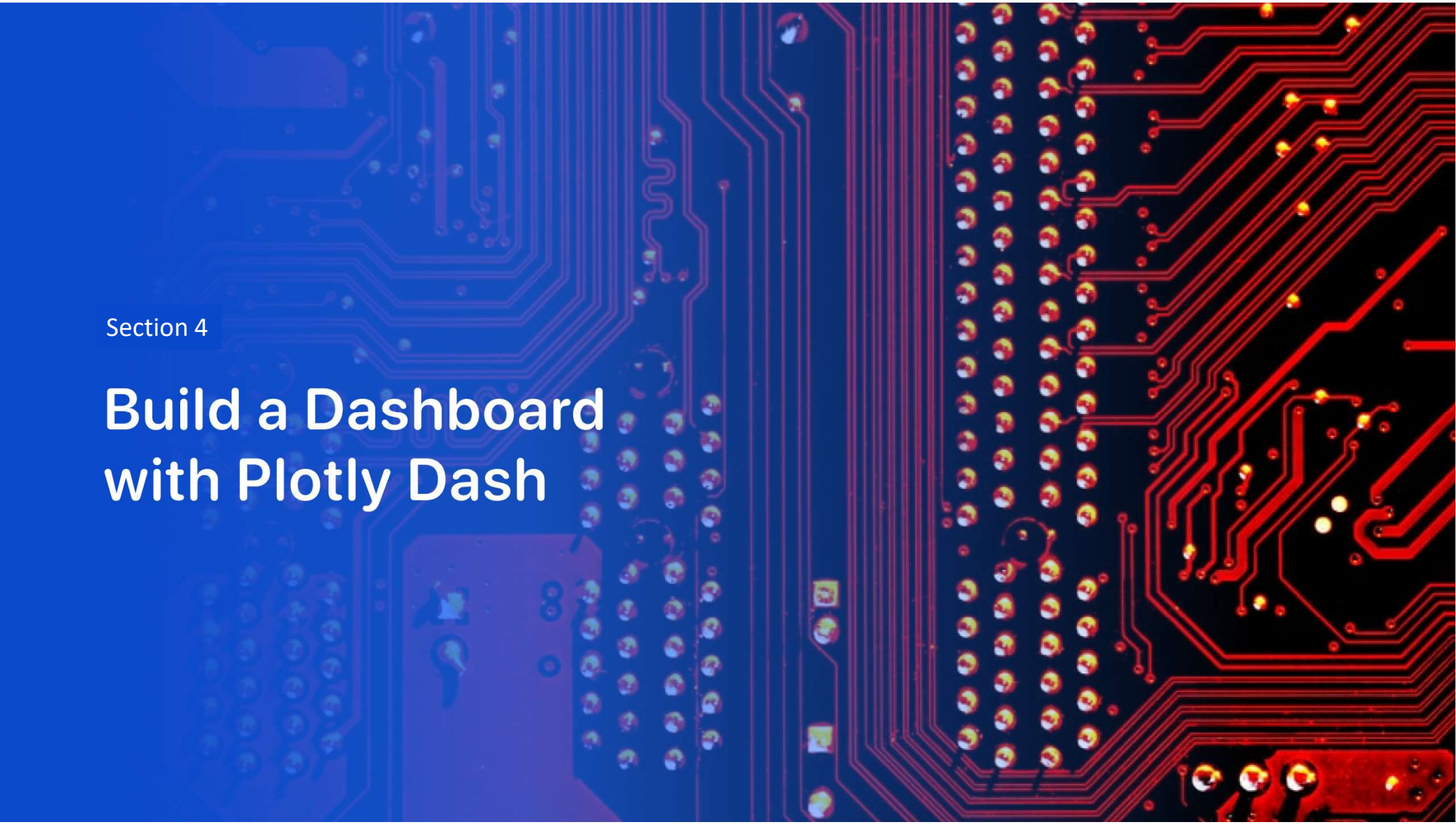
Successful and failed launches per site

- A map showing the successful and failed launches at CCAFS LC-40. Green markers indicate success, red markers indicate failure.

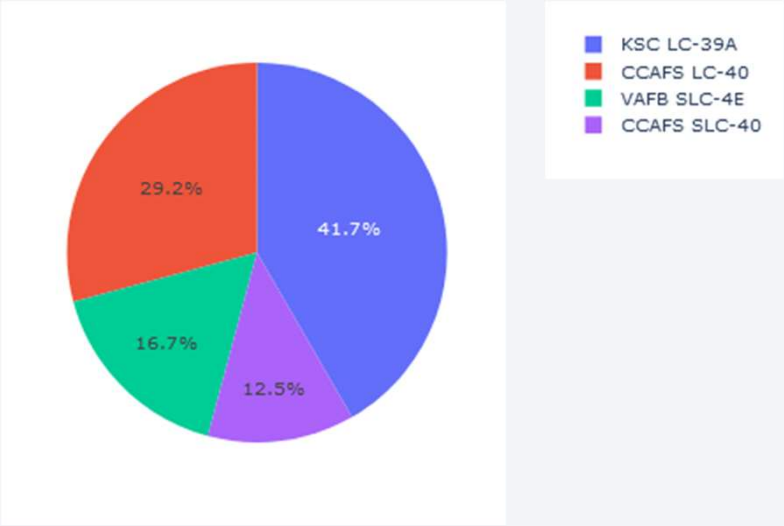
Proximity of launch site to features



- Proximity of VAFB SLC-4E to its nearest railway, coastline, highway and city, with distances marked. Note the proximity to the coast/railway, and further distance from the highway and city.

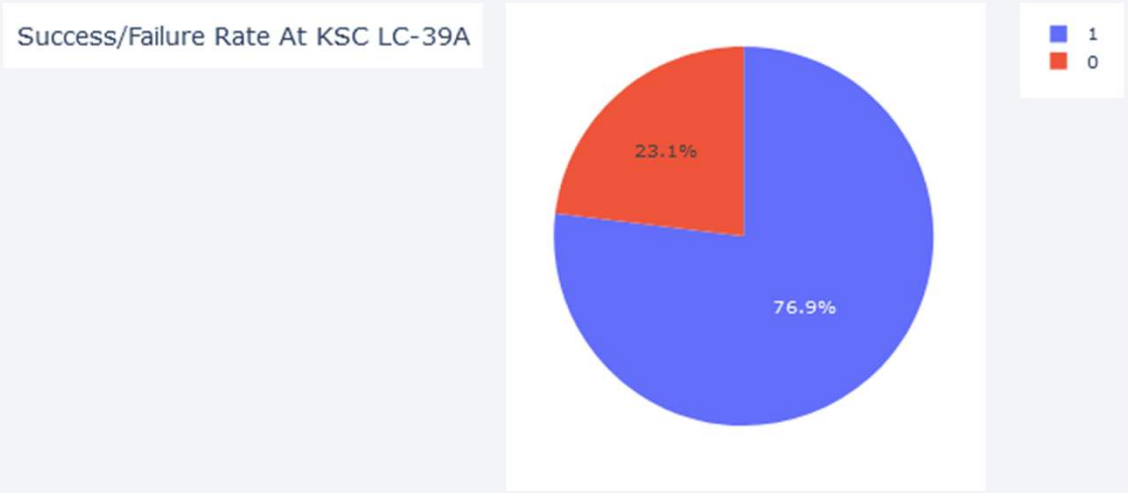

Section 4
Build a Dashboard with Plotly Dash

Launch Site Success Rate Per Site



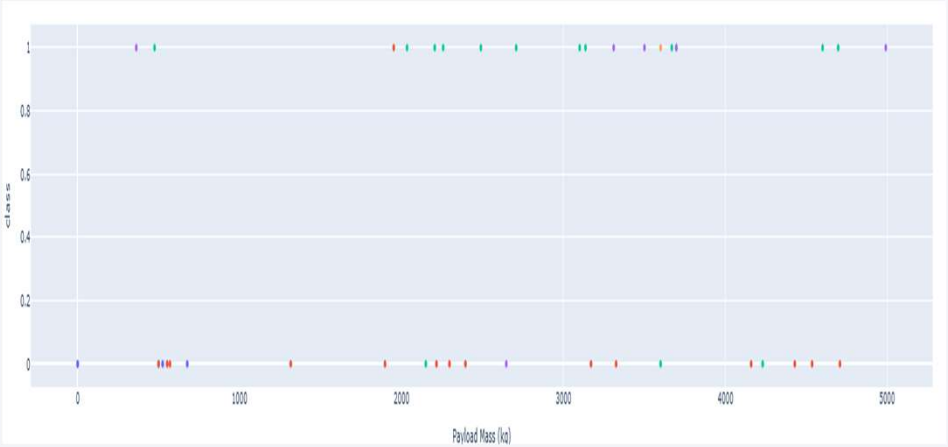
- As we can see, KSC LC-39A has the highest overall number of successful launches, and VAFB SLC-4E has the smallest.

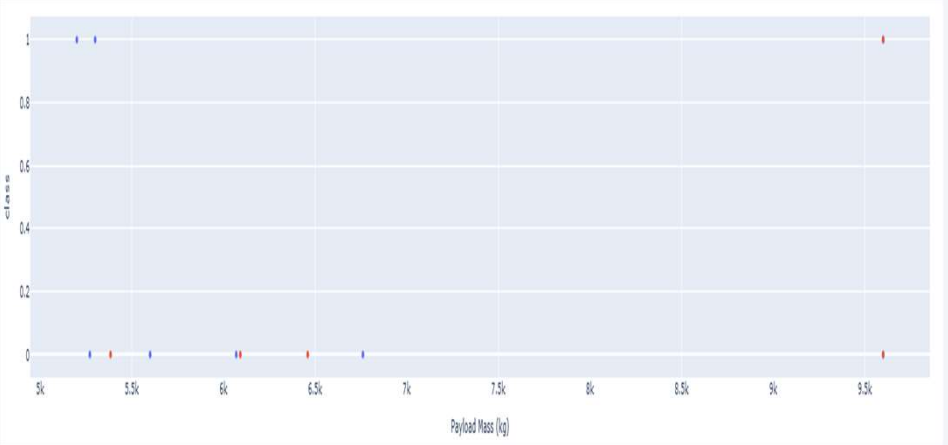
Highest Launch Site Success Ratio



- We can see that KSC LC-39A also has the overall highest success rate, which invites further analysis on what makes this site the most successful in both metrics.

Comparison of Payload Ranges vs. Launch Outcomes



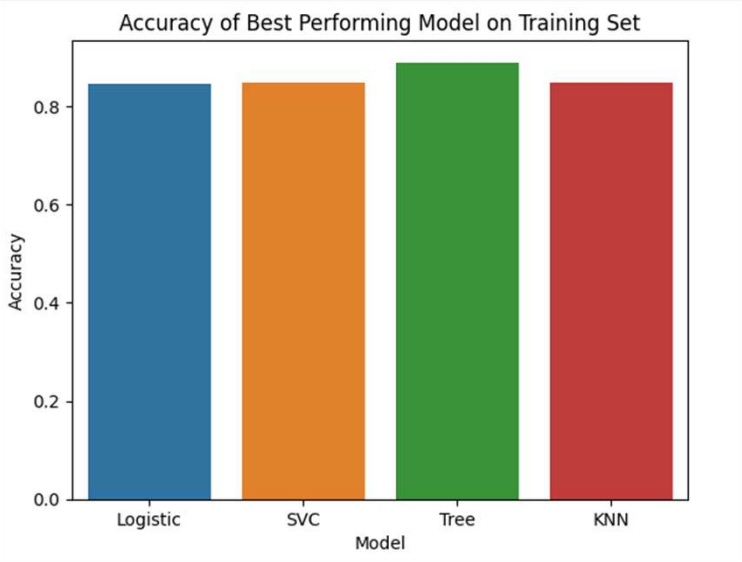


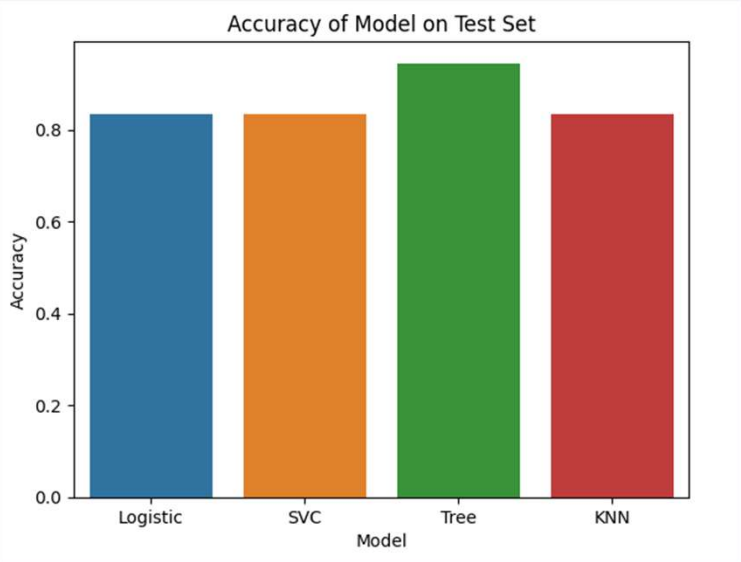
- A comparison of payload ranges between 0-5000kg and 5000-10000kg across all launch sites. Larger payloads appear to have a higher rate of failure.

Section 5

Predictive Analysis (Classification)

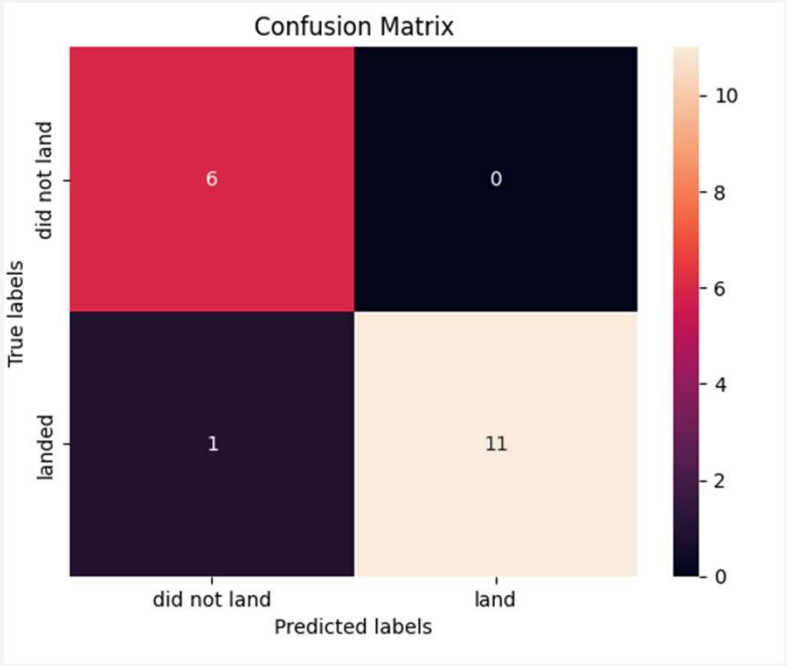
Classification Accuracy





- The logistic, SVC and KNN models all had comparable accuracies. Decision Tree classifier performed the best, with 88.9% accuracy on the training set and 94.4% accuracy on the test set.

Confusion Matrix



- The confusion matrix of the decision tree classifier on the test set. All failure outcomes were correctly classified, and only one false negative was assigned.

Conclusions

- KSC LC-39A has been shown to be the best launch site in terms of landing outcomes, as it has both the highest number of successful launches as well as the highest ratio of success to failure.
- Lower mass payload launches appear to incur less risk.
- The overall success rate of launches has a clear increase over time, likely due to technological improvements and learning from past errors.
- Decision Tree Classification provided the best accuracy on both the training and test sets, making it a good candidate for predictive analysis of future launches.

Appendix

- All code and generated csv files can be located at https://github.com/liam-stuart/Applied-Data-Science-Capstone/tree/main
- The Folium maps will not work there, hence the alternative link provided in this PowerPoint.

Thank you!