Winning Space Race with Data Science

Udit Deshpande
2022-09-09

Outline

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

Executive Summary

The objective of this data science engagement was to predict the success/failure of a landing of the Falcon 9 rocket based on the information available. The information required to make these predictions were obtained from SpaceX APIs and some publicly available data points published in Wikipedia.

Based on the data available and features observed by generating various plots with respect to the outcome, it was observed that payload mass, orbit, launch site, location and reused count played a significant part in the determination of the success or failure of the outcome.

After retrieving the data, treating the null values, and appropriately transforming the data, a ‘best’ split ‘entropy’ criterion decision-based tree model was built using grid search to predict the landing success or failure with an accuracy of 94.5%.

Introduction

A company Space Y wanted to analyze the data of the SpaceX Falcon 9 and determine if they could calculate the cost of deploying a satellite in the orbit. A sizable contribution to this cost is based on the outcome of the SpaceX Falcon 9 landing. If the landing is successful, the first phase of the rocket could be reused which could save the costs by approximately $62M. However, a landing failure could increase the cost of the launch.

The main objective of the project was to retrieve the publicly available data related to Falcon 9 launches by SpaceX and determine whether the outcome of the launch was a success or a failure.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the SpaceX APIs for Falcon launches.
  - Web scraping using the Wikipedia pages for Falcon launches and other relevant information.
- Perform data wrangling
  - Data was cleaned for any special characters, formatting errors, etc.
  - Null values were addressed appropriately by assigning mean values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology (Contd.)

Executive Summary

- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

Data Collection

1. Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and
2. Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

Data Collection – SpaceX API

- SpaceX REST API Calls
- Click HERE for the GITHUB permalink

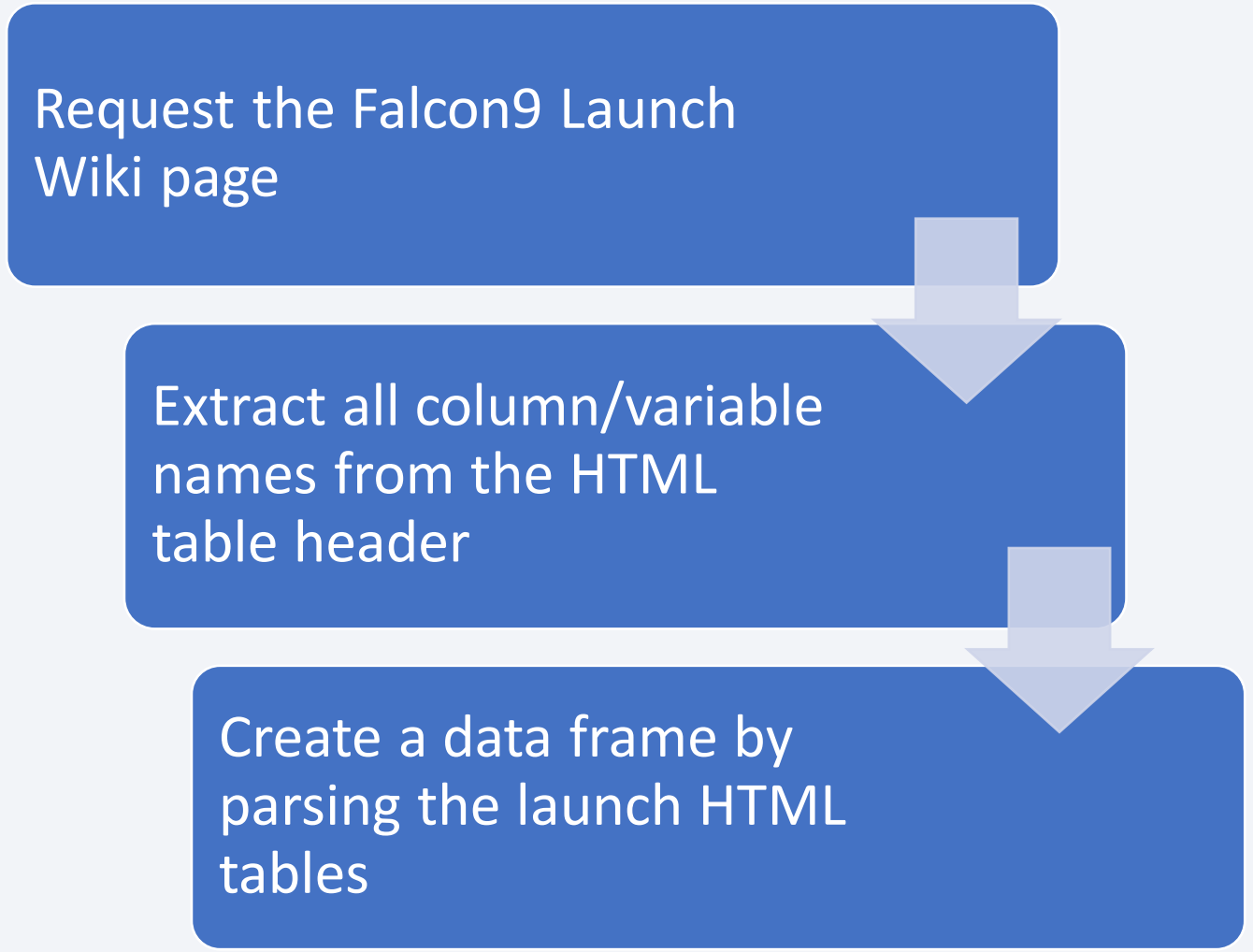
Request API and parse the SpaceX launch data

Filter data to only include relevant data only

Deal with Missing Values with mean fill methodology

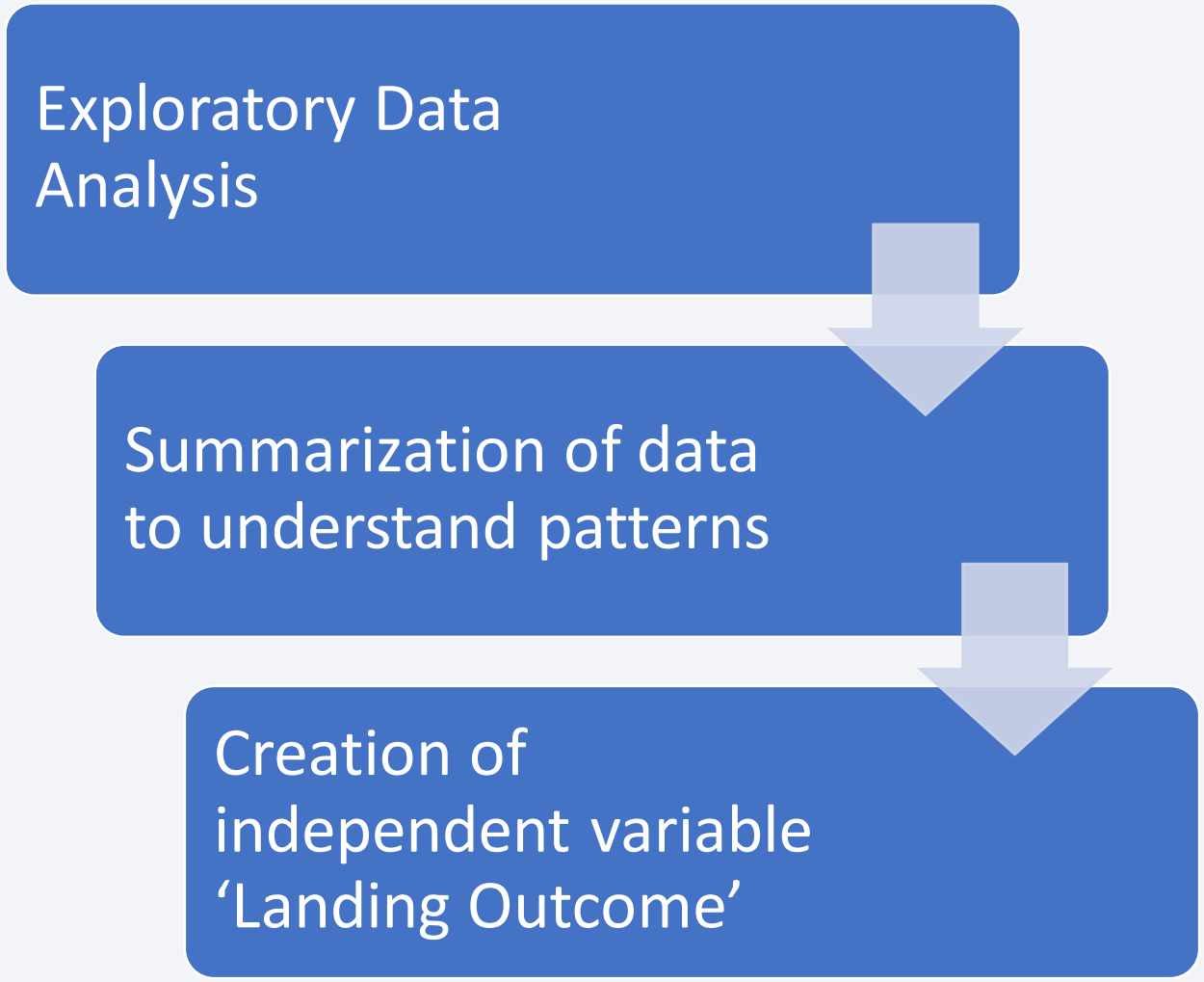
Data Collection - Scraping

- Data from SpaceX launches was obtained from Wikipedia
- Data was downloaded from Wikipedia according to the flowchart and then persisted.
- Click HERE for the GITHUB permalink



Data Wrangling

- Initially Exploratory Data Analysis (EDA) was performed on the dataset to understand the shortcomings in data.
- Summaries of different fields such as launch site level analysis, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated to understand the correlation between variables and the mission outcome.
- Finally, the landing outcome label was created from Outcome column.
- Click HERE for the GITHUB permalink



EDA with Data Visualization

Correlation with various variables present in the dataset were identified using data visualization:

- Payload Mass vs. Flight Number,
- Launch Site vs. Flight Number,
- Launch Site vs. Payload Mass,
- Orbit vs. Flight Number,
- Payload vs. Orbit
- And all the variables mentioned above vs. Success Rate / mission outcome.
- Click HERE for the GITHUB permalink

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA';
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
- Click HERE for the GITHUB permalink

Build an Interactive Map with Folium

- Folium was used to create interactive maps to mark the following:
  - Markers indicate points like launch sites.
  - Circles indicate highlighted areas around specific coordinates and launch sites.
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site.
  - Lines are used to indicate distances between two coordinates.

- Click HERE for the GITHUB permalink

Build a Dashboard with Plotly Dash

- Interactive dashboards were built to understand the correlation between various variables
  - Percentage of launches by site
  - Payload range
- Click HERE for the GITHUB permalink
- Screenshots of the output are attached below:
  - spacex_interactive_01.png
  - spacex_interactive_02.png
  - spacex_interactive_03.png
  - spacex_interactive_04.png

Predictive Analysis (Classification)

- Four classification models used to predict mission outcome:
  - logistic regression,
  - support vector machine,
  - decision tree
  - K nearest neighbors.
- Out of the 4 methods it was observed that the decision tree classifier had a higher success in predicting the mission outcome correctly, thus proving a superior model compared to the other three models.
- Click HERE for the GITHUB permalink

Results

- Space X has been using 4 different launch sites for the Falcon 9 projects.
- The average payload of F9 v1.1 booster version was 2,928 kg.
- The first success landing outcome happened in 2015 five years after the first launch.
- Successful outcomes have been observed while landing in drone ships having payload above than average.
- Almost 98.01% of mission outcomes were successful.
- Only two booster versions failed while landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015.
- Success ratio of the landing outcomes has bettered with the passage of time, especially after 2013.

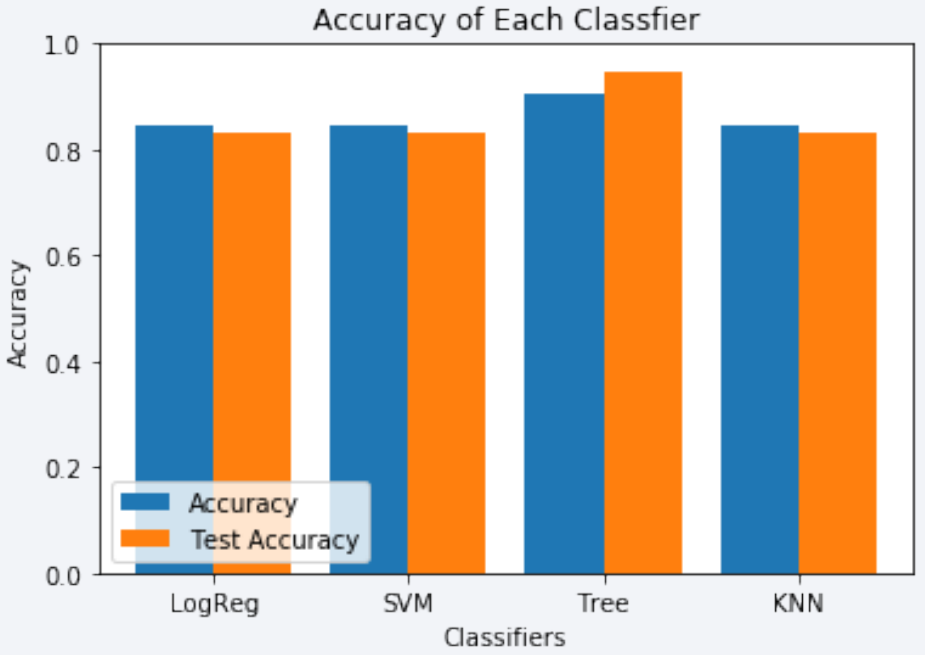
Results

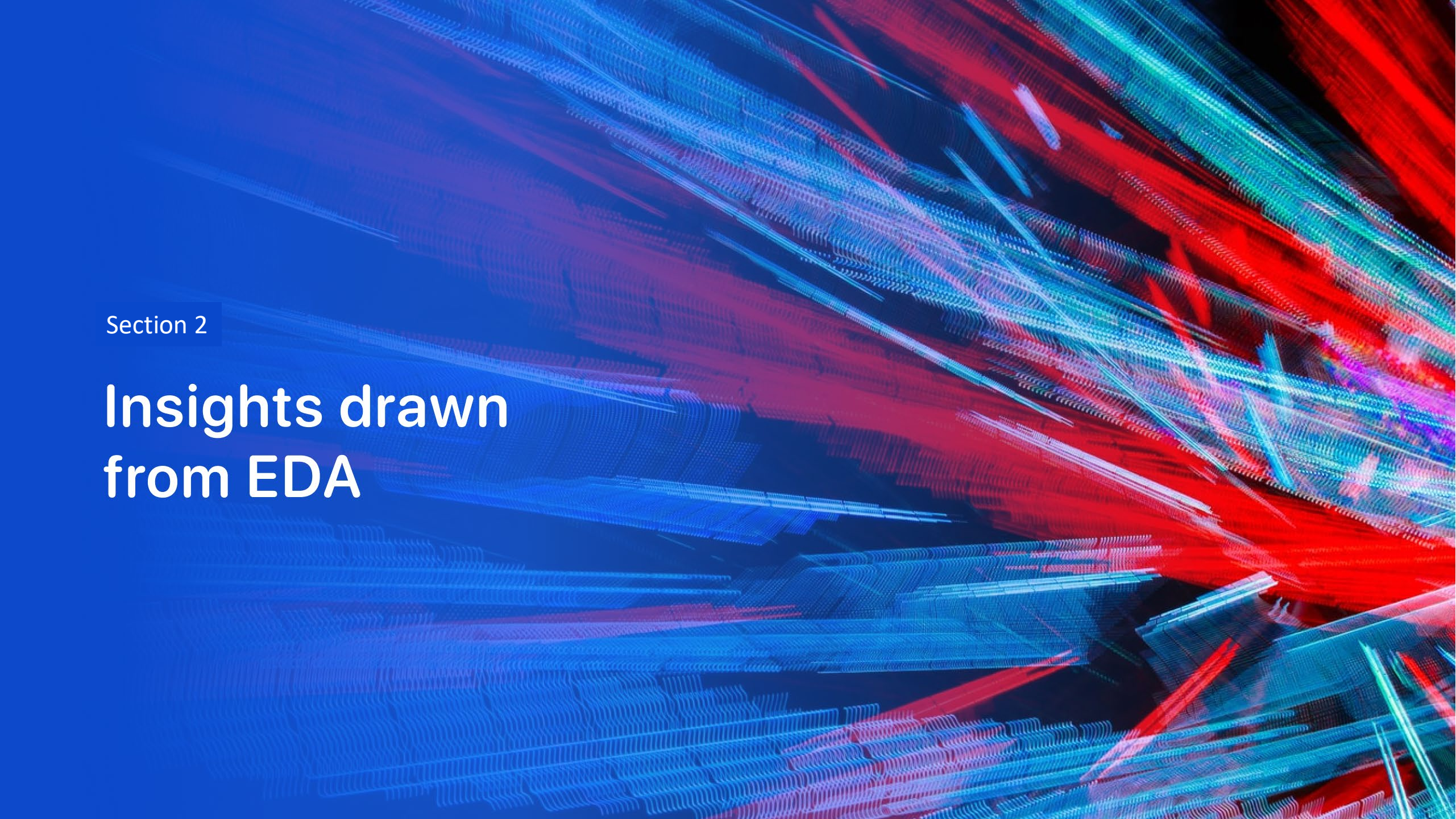
- Geospatial analytics helped understand that launch sites with safety places, near sea, and better logistic infrastructure around have had more launches and a better success ratio.
- Most launches happens at east cost launch sites at the 3 sites (46) than at the west coast (10).

Results

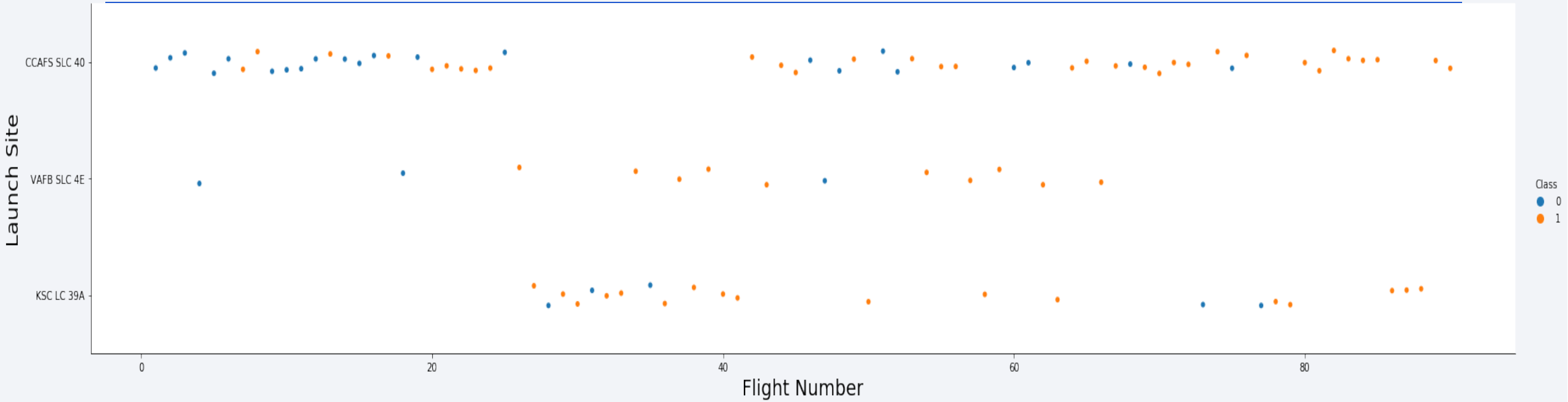
- Decision Tree Classifier was the best model to predict successful landings, having accuracy over 87% for the training data and 94% for the actual test data.

- Other models and accuracies have been displayed in the bar chart.



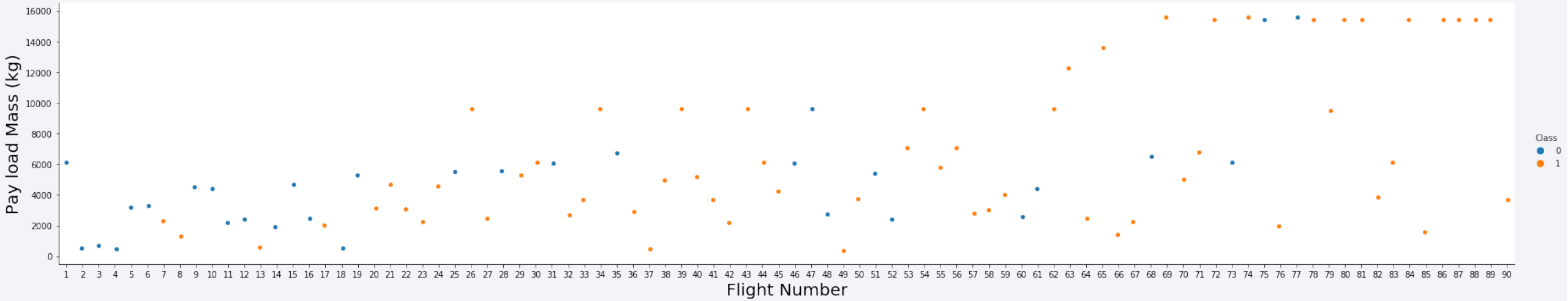

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Observations:
  - Recent flights launched from CCAFS SLC 40 have had more consistent successful outcomes than earlier ones.
  - Site VAFB SLC 4E has had the highest success in between the 3 launch sites.
  - CCAFS SLC 40 has had the highest number of launches, successes and failures.

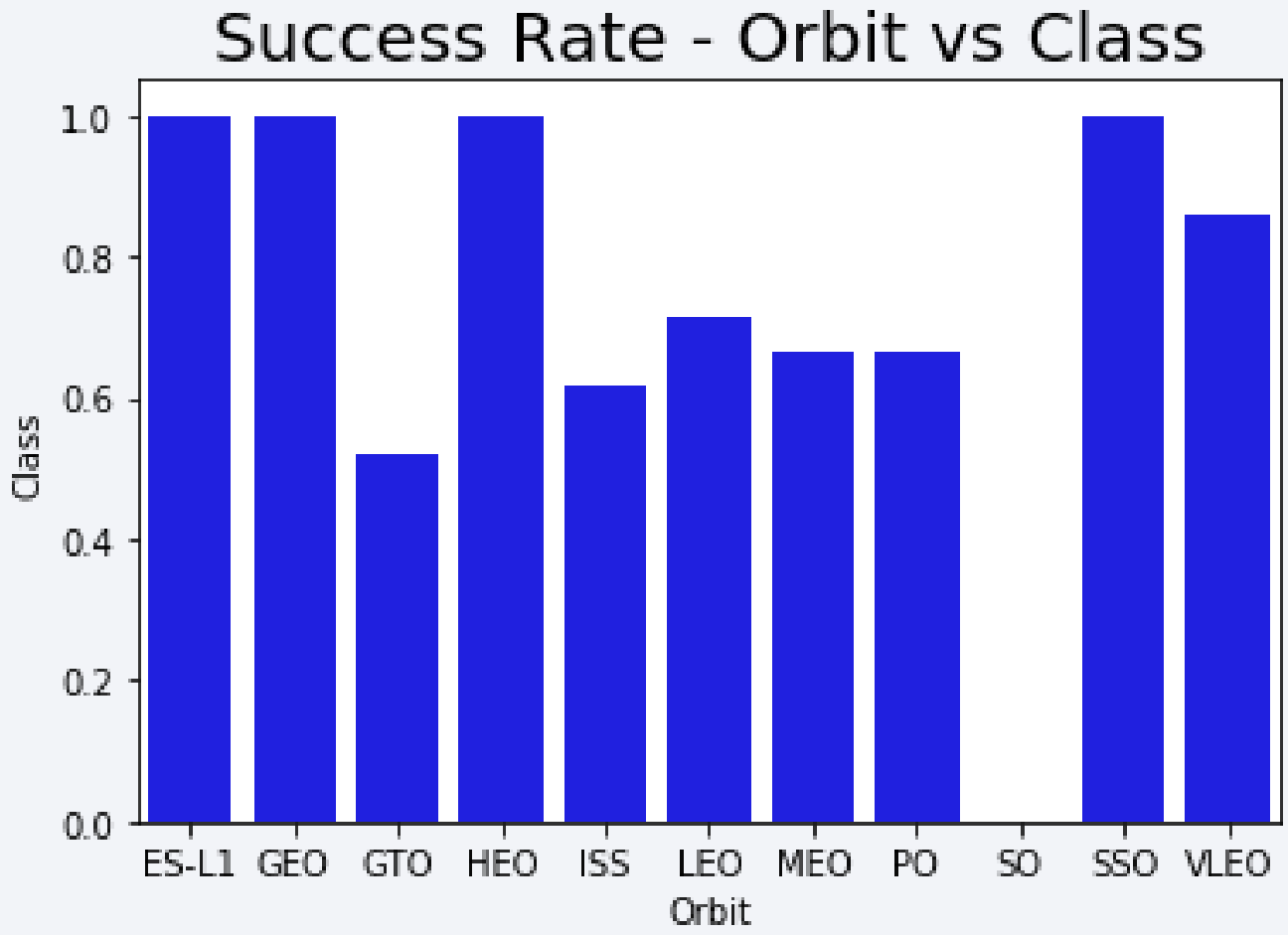
Payload vs. Launch Site



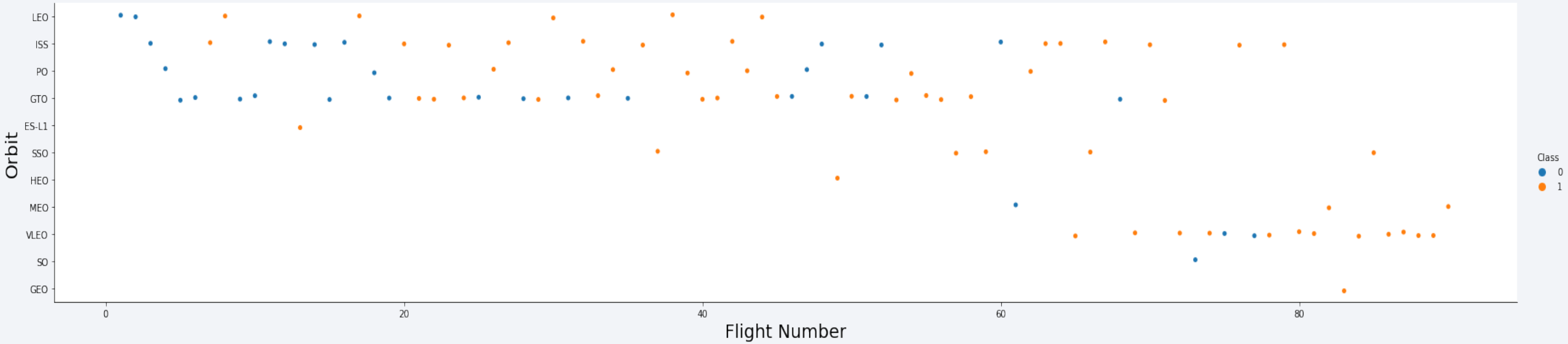
- Observations:
  - Lower payload mass launches seem to have a higher rate of failure than higher payload mass launches.
  - The recent (with higher payload mass) launches have had more successful outcomes than the earlier launches (with lower payload mass).

Success Rate vs. Orbit Type

- This bar chart shows 100% success in the orbits ES-L1, GEO, HEO, SSO, while 0% success with the SO orbit.
- However, the number of flights to the Orbit LEO and GTO has been the highest and that rest of the success rates are based on fewer observations.

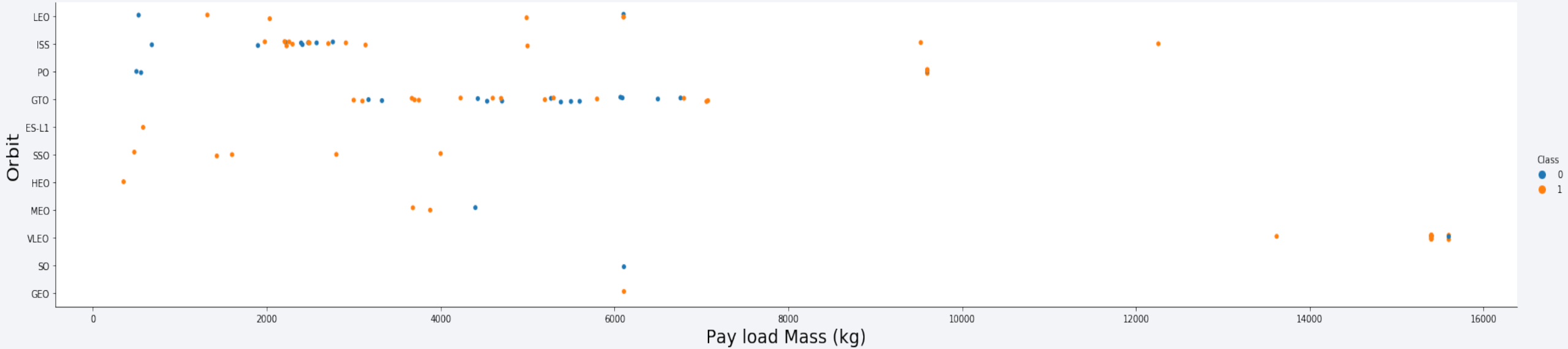


Flight Number vs. Orbit Type



- The highest number of observations were available for the GTO and ISS.
- For all other orbits, SSO, HEO, GEO, and ES-L1 had a 100% success rate, however the number of observations to use orbit as a successful determinant were less.

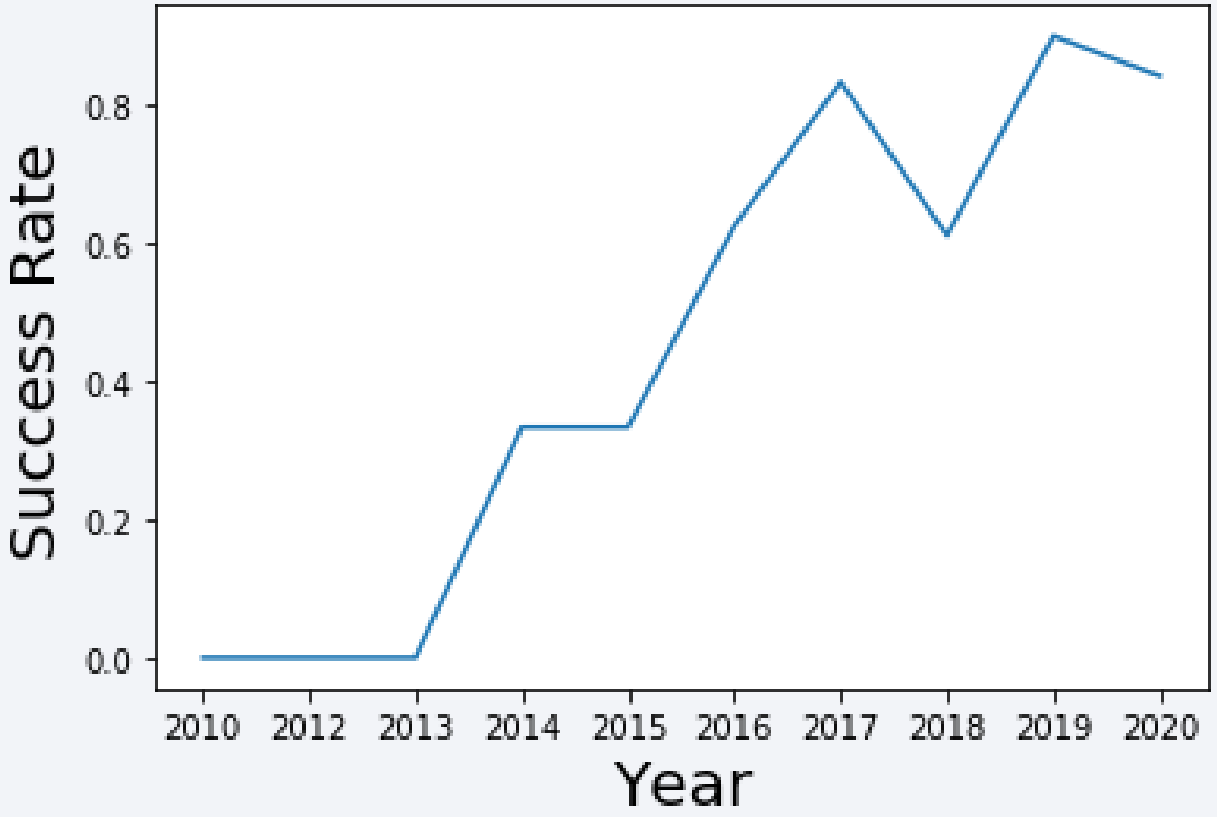
Payload vs. Orbit Type



- It is observed from this scatterplot that a successful outcome is highly evident in mission with a higher payload mass (almost greater than 8000 kgs).
- Also, the concentration of the points in this scatterplot indicate that payload mass and orbits are highly correlated. This means that the payload mass for missions can be approximately determined based on the orbit the mission is planned for.

Launch Success Yearly Trend

- The success rate has been increasing steadily with the passage of time with the highest success in the year 2019.



All Launch Site Names

- There are 4 launch sites present in the data provided and are present in the screenshot.

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

Launch Site Names Begin with 'CCA'

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

```
 * ibm_db_sa://vzl97188:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- All the 5 mission outcomes from the launch site CCFAS LC-40 were successful. However the landing outcomes were a 100% failure (2 out of 2 attempts). Also all the mission were targeted to the LEO or LEO (ISS) orbits.

Total Payload Mass

```
%sql select customer,sum(payload_mass__kg_) as payload_mass from spacextbl group by customer having customer ='NASA (CRS)'
```

```
 * ibm_db_sa://vzl97188:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

customer	payload_mass
NASA (CRS)	45596

- The total payload mass carried by SpaceX for NASA (CRS) is 45,596 kgs.

Average Payload Mass by F9 v1.1

```
%sql select booster_version, avg(payload_mass__kg_) as total_payload_mass from spacextbl group by booster_version having booster
```

```
 * ibm_db_sa://vzl97188:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

booster_version	total_payload_mass
F9 v1.1	2928

- The average payload mass carried by F9 v1.1 was 2928 kgs.

* The screenshot contains a typo. The column header should have been avg_payload_mass and not total_payload_mass.

First Successful Ground Landing Date

```
%sql select min(DATE) as first_successful_landing from spacextbl where mission_outcome ='Success'
```

```
 * ibm_db_sa://vzl97188:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

first_successful_landing
2010-06-04

- The first successful ground landing took place on 2010-06-04.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select distinct booster_version from spacextbl \
where payload_mass__kg_>4000 and payload_mass__kg_<6000 \
and mission_outcome='Success';
```

```
 * ibm_db_sa://vzl97188:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

- Based on the successful mission outcome and payload between 4,000 and 6,000 kgs, it was observed that F9 B5 and FT booster versions were frequently contributing to success which helps us understand that mission outcomes are correlated with the booster versions.

booster_version
F9 B4 B1040.2
F9 B4 B1040.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5 B1058.2
F9 B5B1054
F9 B5B1060.1
F9 B5B1062.1
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1032.2
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1032.1
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome,count(*) as count from spacextbl group by mission_outcome
```

```
 * ibm_db_sa://vzl97188:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

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

- Based on the SpaceX data available, the success rate of the mission outcome was 98.01% (99 out of 101 observations).

Boosters Carried Maximum Payload

```
%sql select BOOSTER_VERSION from SPACEXTBL where payload_mass__kg_=(select max(payload_mass__kg_) from spacextbl)
```

```
 * ibm_db_sa://vzl97188:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

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

- When the data was filtered to identify the booster versions which carried the maximum payload, it was observed that only one booster version F9 B5 was used to carry maximum payload.

2015 Launch Records

```
%sql SELECT date,MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where EXTRACT(YEAR FROM DATE)='2015'
```

```
 * ibm_db_sa://vzl97188:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

DATE	mission_outcome	booster_version	launch_site
2015-01-10	Success	F9 v1.1 B1012	CCAFS LC-40
2015-02-11	Success	F9 v1.1 B1013	CCAFS LC-40
2015-03-02	Success	F9 v1.1 B1014	CCAFS LC-40
2015-04-14	Success	F9 v1.1 B1015	CCAFS LC-40
2015-04-27	Success	F9 v1.1 B1016	CCAFS LC-40
2015-06-28	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
2015-12-22	Success	F9 FT B1019	CCAFS LC-40

- Based on the observations recorded in the year 2015, the mission outcome was 85.17% (6 out 7 attempts).
- One common observation is that all the missions took place at the CCAFS LC-40 (mostly using the F9 v1.1 booster version), which helps us understand that the launch site is a key variable which can help determine the landing outcome.

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

```
%sql select landing__outcome,count(*) \
from (select * from spacextbl \
where  DATE BETWEEN '2010-06-04' AND '2017-03-20') \
group by landing__outcome
ORDER BY 2 DESC;
```

```
 * ibm_db_sa://vzl97188:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

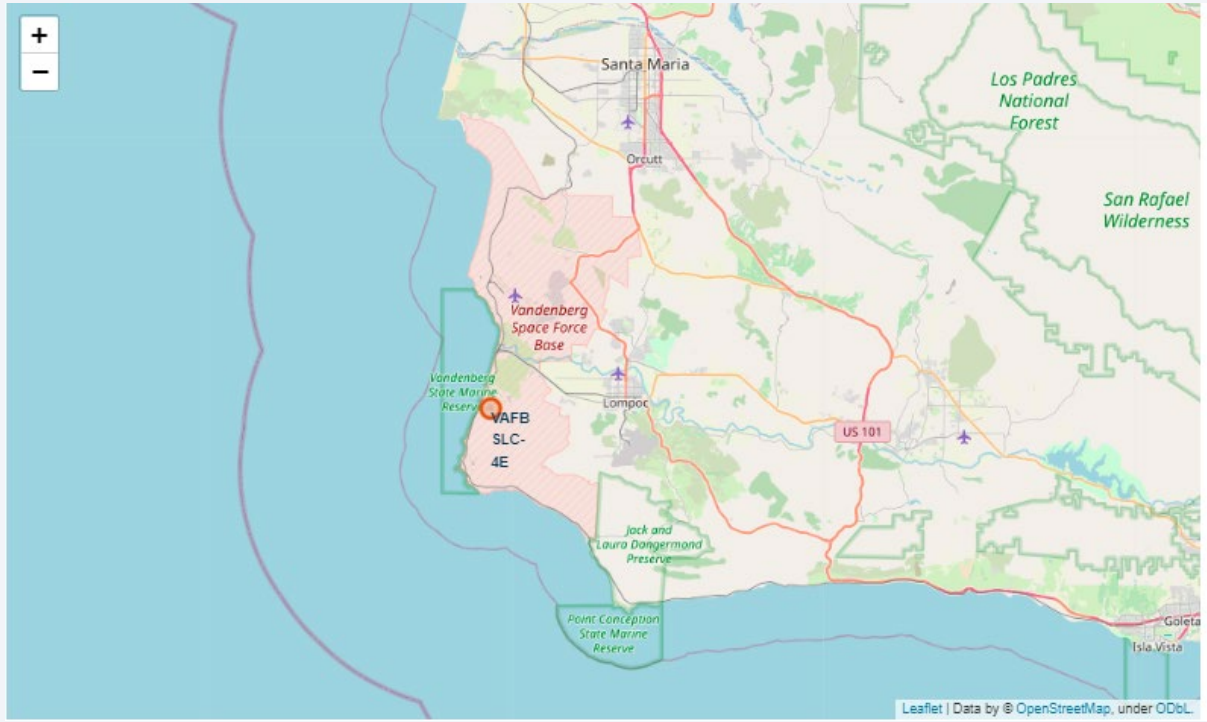
landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

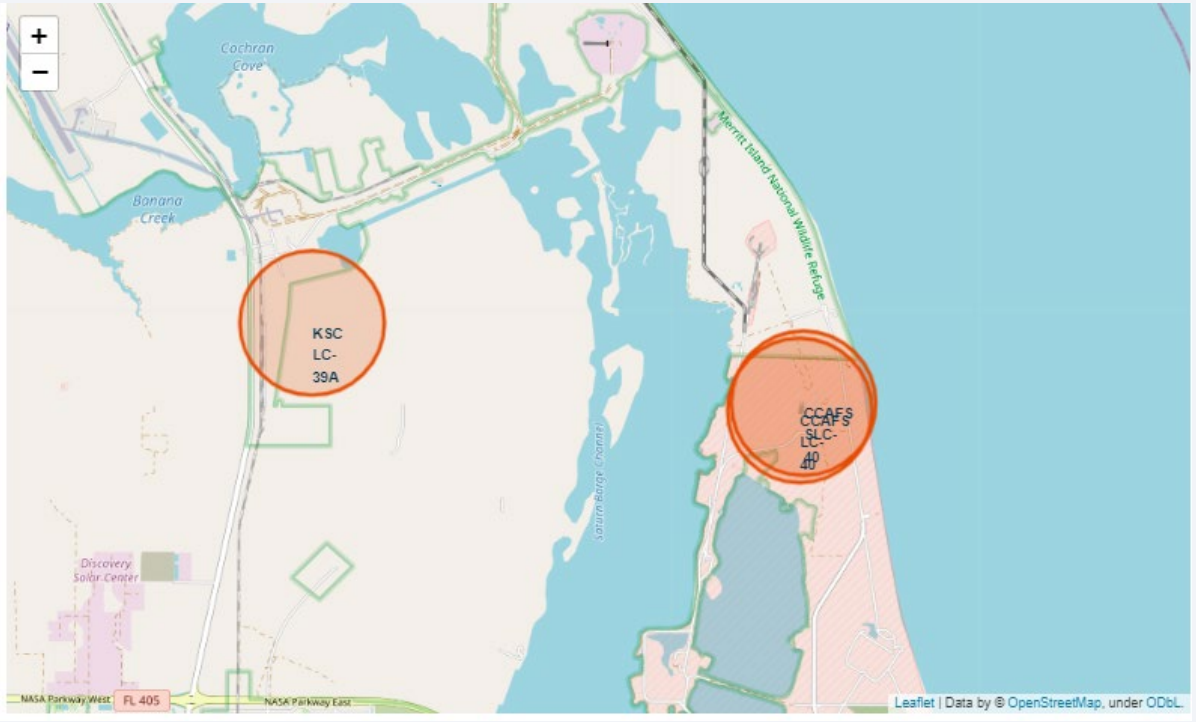
- Based on the query, the most frequent observation for the landing outcome was no attempt. However, the success rate for a drone ship was 50% (5 out of 10 attempts)

Section 3

Launch Sites Proximities Analysis

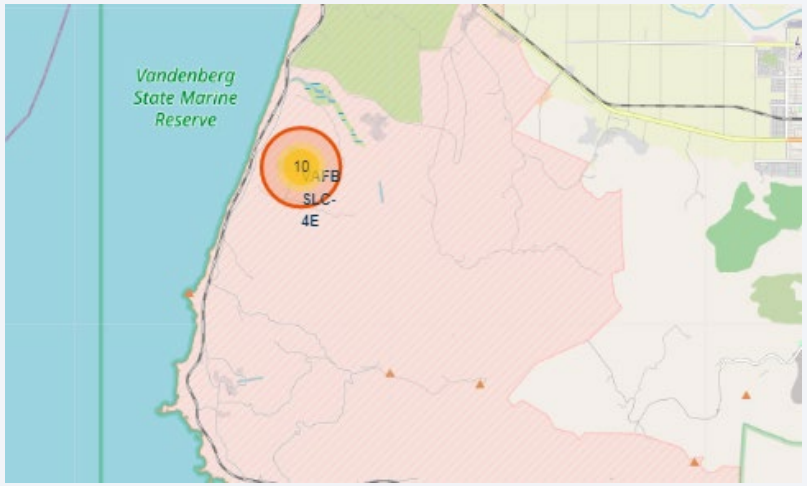
4 Launch Sites on the map of USA



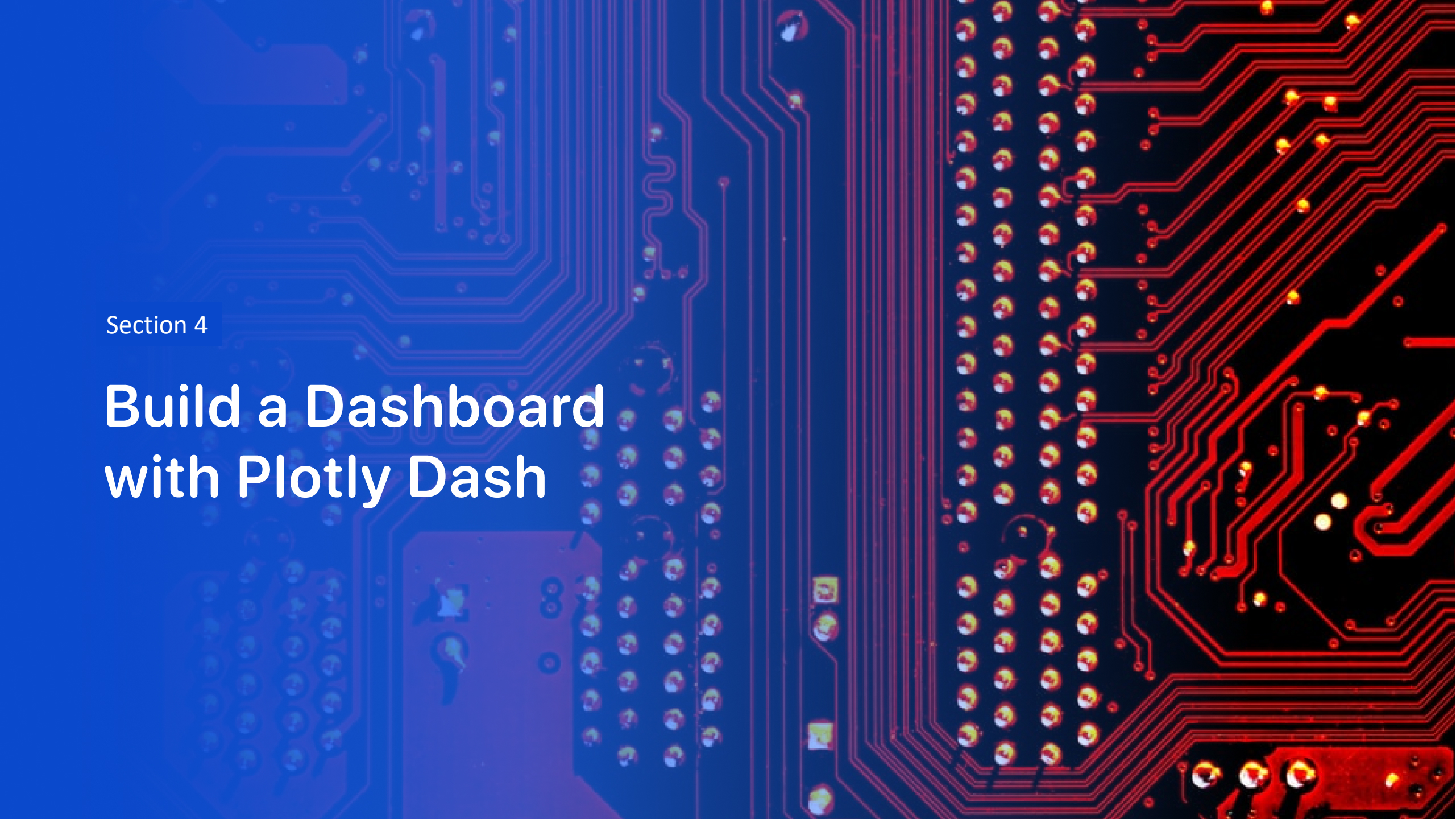


- This map shows the location of the 4 launch sites on the map of USA.
  - CCAFS LC-40
  - CCFAS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

Successful Outcomes at Launch Sites



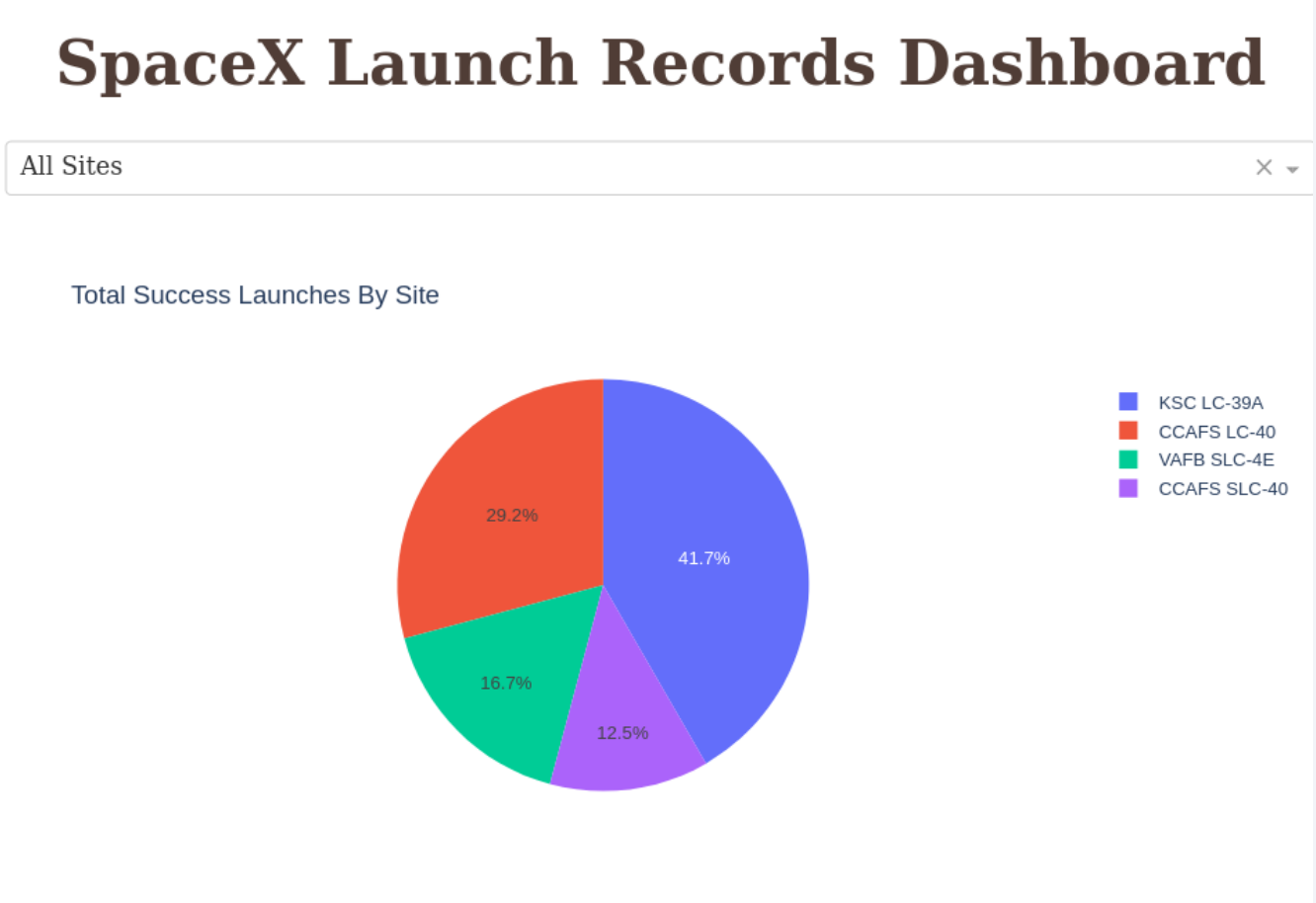
- The launch outcomes are shown in yellow on the map for each launch sites.


Section 4
Build a Dashboard with Plotly Dash

Successful Launches from all Launch Sites

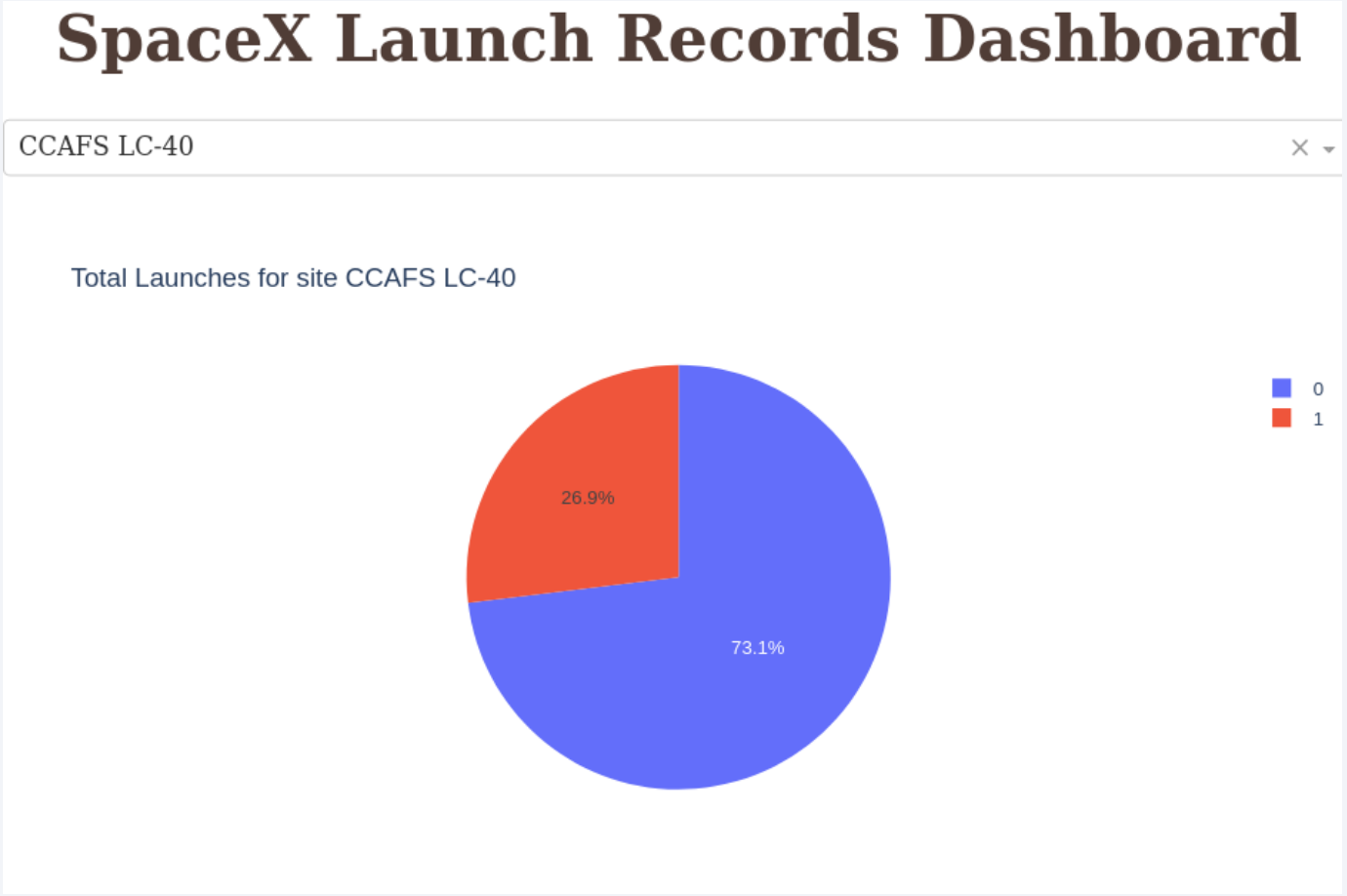
Launch sites seems to be an important contributor in the prediction of a successful outcome.

For example, more successful missions have been observed when launched from KSC LC-39A than CCAFS SLC-40.



Launch Site with the Highest Success Ratio

- The success ratio at CCAFS LC-40 has been the highest among the other launch sites.
- Thus, success of a mission is more likely when launched from CCAFS LC-40 than any other launch site.



Payload vs. Launch Outcome

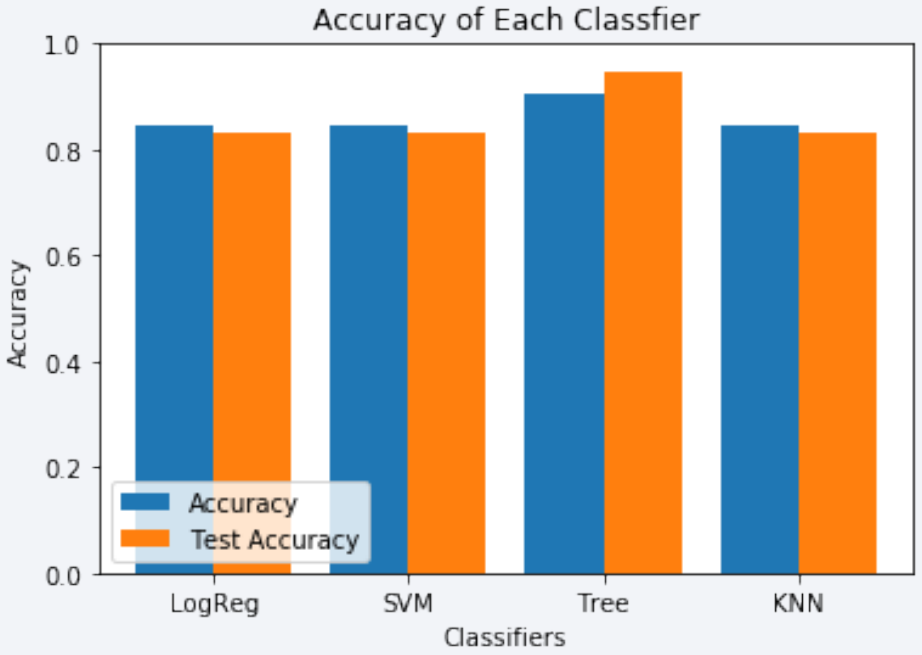
- Booster versions v1.0 and FT share the highest success when compared to any other boosters for a medium payload.
- Also, extremely high and low payloads have had lower chances of success.

Section 5

Predictive Analysis (Classification)

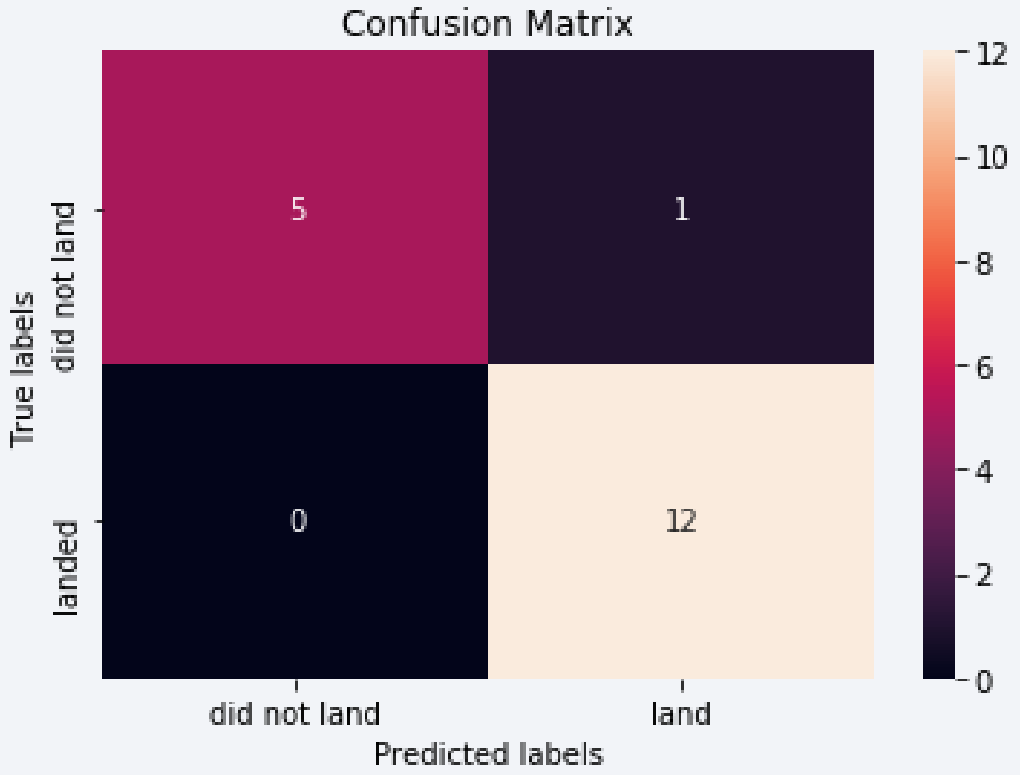
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside.
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracy over 87%.



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves that only 1 value was misclassified, rest all being predicted accurately which proves that decision tree classifier is the best model for this project.



Conclusions

- Variables such as launch site, payload mass, orbit, booster version were highly useful in determining the success of the mission.
- Launches performed from the KSC LC-39A had a higher success ratio than any other site.
- Rockets have been more successful post 2013 and has had a higher success rate compared to the earlier years.
- Launch sites on the east coast have had more launches and hence more success than on the west coast.
- Decision Tree Classifier was the best model with the highest test accuracy.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!