IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

Using available SpaceX launch data, exploratory data analysis of the data shows that the three following criteria combined determine mission success where success is both deployment of the payload and retrieval of the booster for reuse.

1. Launch Site KSC LC39-A
2. Payload Mass < 5500 kg
3. Booster type FT

Introduction

SpaceX costs and competitive advantage are determined both by deployment of the payload and successful retrieval of the booster for reuse.

Mission failure arises from:

- Failed launch'
- Failure to deploy payload before reentry
- Loss of the booster into the sea
- Destruction of the booster by a failed landing

Available data includes booster design, launch location and payload but neglects any other variables such as weather, date, etc.

We seek to determine which parameter most strongly affects success, but will not explore *why*.

Section 1

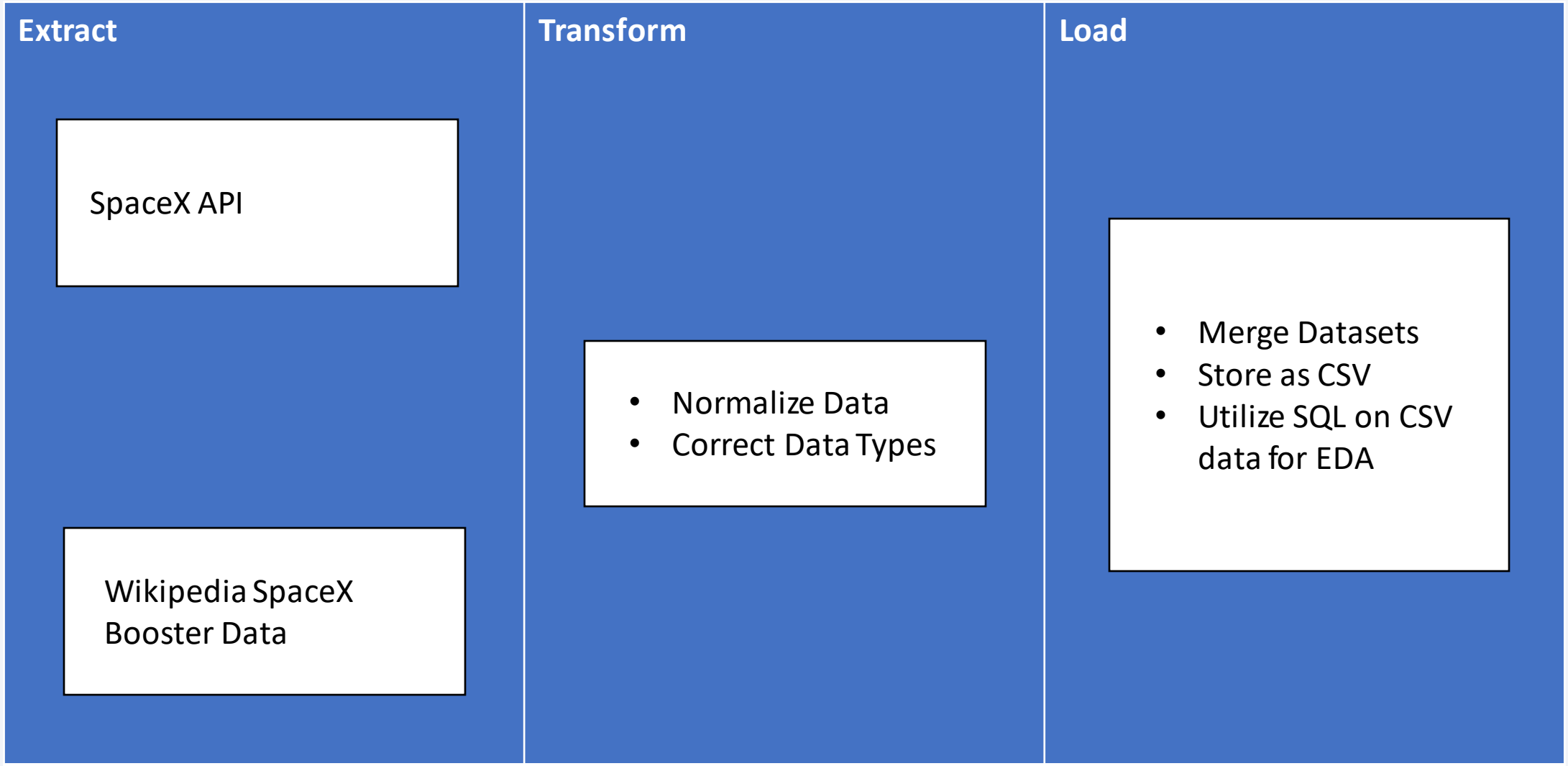
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by dowloading launch data from SpaceX by API
  - API data was combined with web-scraped data from wikipedia data on the launches
- Perform data wrangling
  - Wikipedia and API data were merged to form the complete dataset based upon launch ID.
  - Missing data was dropped
  - Some value types were converted.
- Exploratory data analysis (EDA) was used to identify the most likely parameters
- Perform predictive analysis was used to identify which models are mostly likely to predict the success of future launches

Data Collection



Data Collection – SpaceX API

- Jupyter Notebook in GitHub

```
https://api.spacexdata.com/v4/
getBoosterVersion()
GetLaunchSite()
GetPayloadData()
GetCoreData()
```

```
#Global variables
BoosterVersion = []
PayloadMass = []
Orbit = []
LaunchSite = []
Outcome = []
Flights = []
GridFins = []
Reused = []
Legs = []
LandingPad = []
Block = []
ReusedCount = []
Serial = []
Longitude = []
```

```
data[['rocket', 'payloads', 'launchpad', 'cores',
'flight_number', 'date_utc']]
```

Data Collection - Scraping

- https://github.com/cubeconvict/data-science/blob/main/IBM_capstone/jupyter-labs-webscraping.ipynb

`https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches`

Extract

```
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'] = [] #
Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

Transform

```
date_time()
booster_version()
landing_status()
get_mass()
booster_version()
```

Transform

`spacex_web_scraped.csv`

Data Wrangling

https://github.com/cubeconvict/data-science/blob/main/IBM_capstone/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

https://github.com/cubeconvict/data-science/blob/main/IBM_capstone/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

- Flight number vs Payload (scatter plot)
- Flight Number vs Launch Site (scatter plot)
- Payload vs Launch Site (scatter plot)
- Success rate vs Orbit Type (bar plot)
- Flight Number vs Orbit Type (scatter plot)
- Payload vs Orbit (scatter plot)
- Launch success over time

EDA with SQL

https://github.com/cubeconvict/data-science/blob/main/IBM_capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- Determined distinct Launch Sites
- Filtered Booster Versions to Falcon 9
- Determined date ranges
- Explored types and numbers of Mission Successes
- Explored Landing Outcome types

Build an Interactive Map with Folium

https://github.com/cubeconvict/data-science/blob/main/IBM_capstone/lab_jupyter_launch_site_location.jupyterlite.ipynb

Interactive map created to show:

- Launch locations
- Launch pads
- Launch Success/Failure
- Launch Data
- Launch proximity to local utilities and features

Build a Dashboard with Plotly Dash

https://github.com/cubeconvict/data-science/blob/main/IBM_capstone/dash_app.ipynb

EDA determined that launch location, booster type and payload were features that best determined mission success.

An interactive application was created to allow:

- Selection of launch site
- Selection of Payload Range

Context Sensitive Graphs Include:

- Scatter plot of payload vs success rate
- Success vs Failure based upon site selection

Predictive Analysis (Classification)

https://github.com/cubeconvict/data-science/blob/main/IBM_capstone/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

- Static dataframe was used from class files (dataset_part_3.csv)
- Data Standardized and split into test and train sets
- GridSearchCV used with Logistic Regression
- GridSearchCV used with Support Vector Classifier
- GridSearchCV used with Decision Tree Classifier
- GridSearchCV used with K Nearest Neighbors
- The accuracy of each method was confirmed using confusion matrices and accuracy calculations
- The methods were compared

Logistic Regression accuracy : 0.8357142857142857
Tree accuracy : 0.9321428571428573
KNN accuracy : 0.8767857142857143
SVM accuracy : 0.8625

It was determined that a decision tree was the best machine learning algorithm to predict Launch Success

Results

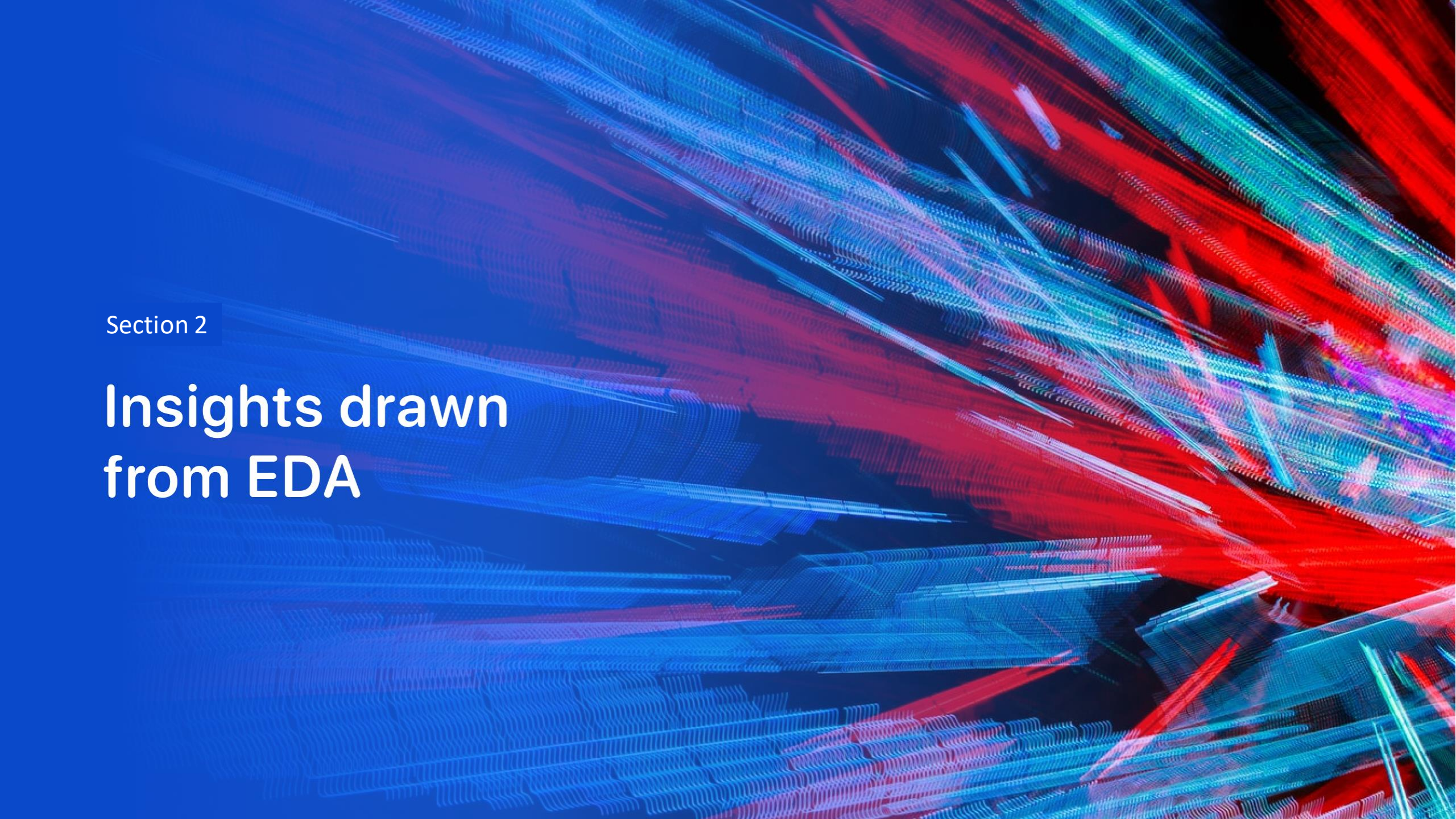
Data shows that the following parameters are the best available predictors of launch success from the available data:

- Launch Site
- Payload Mass
- Booster Type

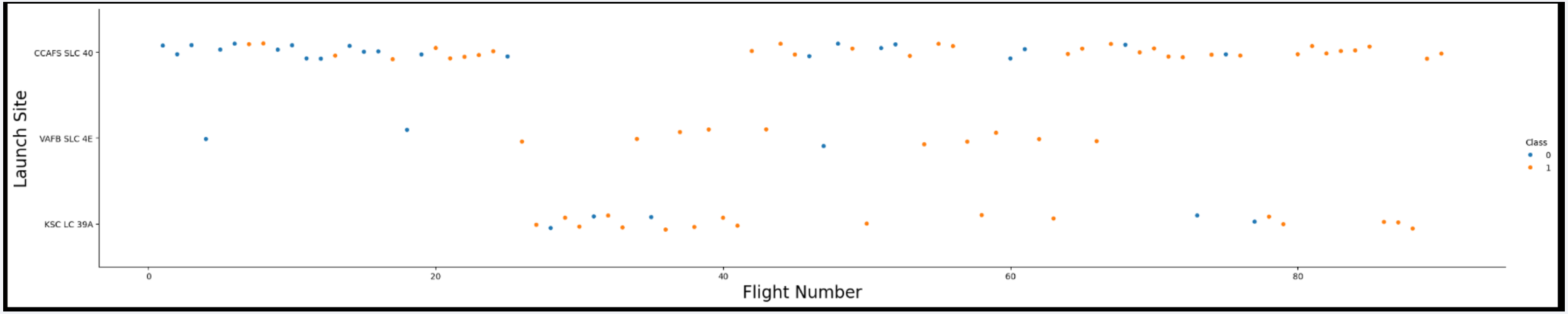
Machine learning shows that a decision tree using these parameters is the best method to predict success of future launches

Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

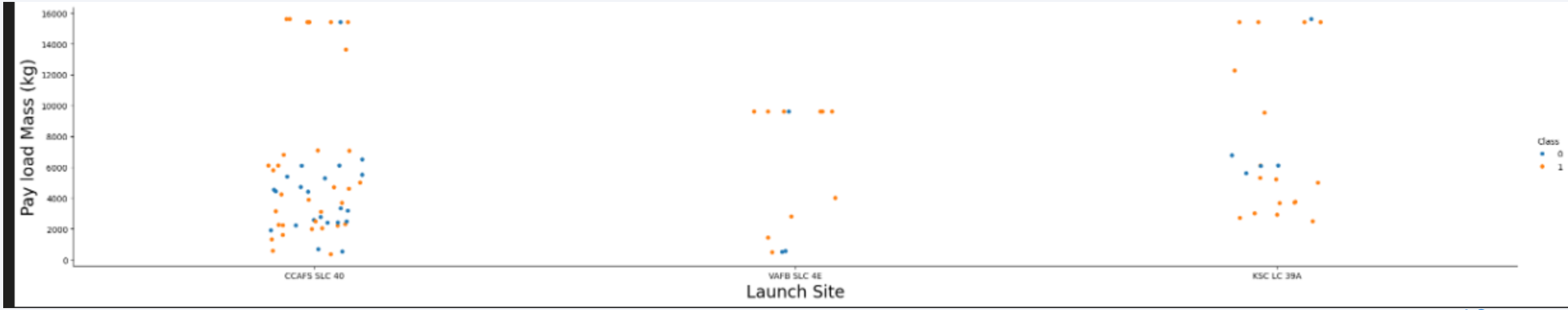


Data shows:

- In most cases a succession of launches were done from each site before changing
- Visual analysis shows that a greater number of launches at the CCAFS SLC 40 site

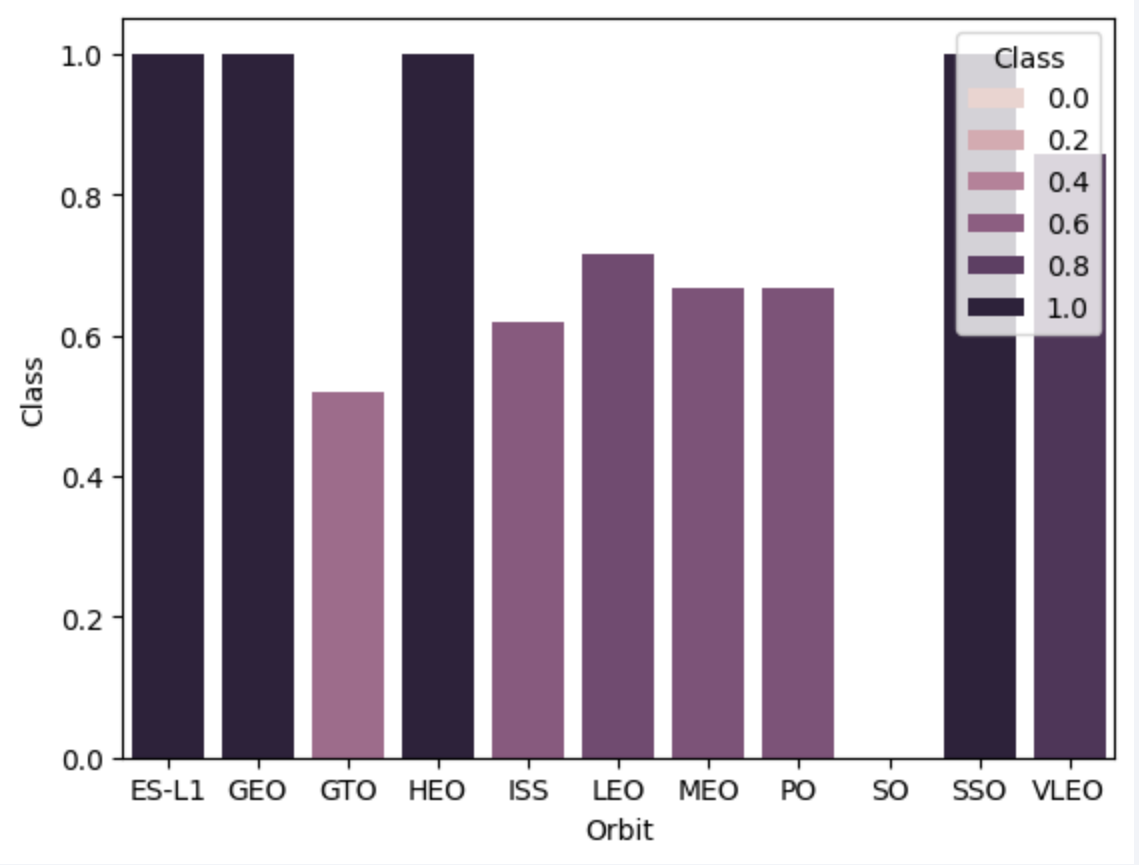
Payload vs. Launch Site

- VAFB SLC 4E did not launch larger payloads but launched the largest number of
- The other two sites launched roughly the same number of larger payloads ~10,000 kg payloads
- CCAFS SLC 40 launched the largest number of payloads



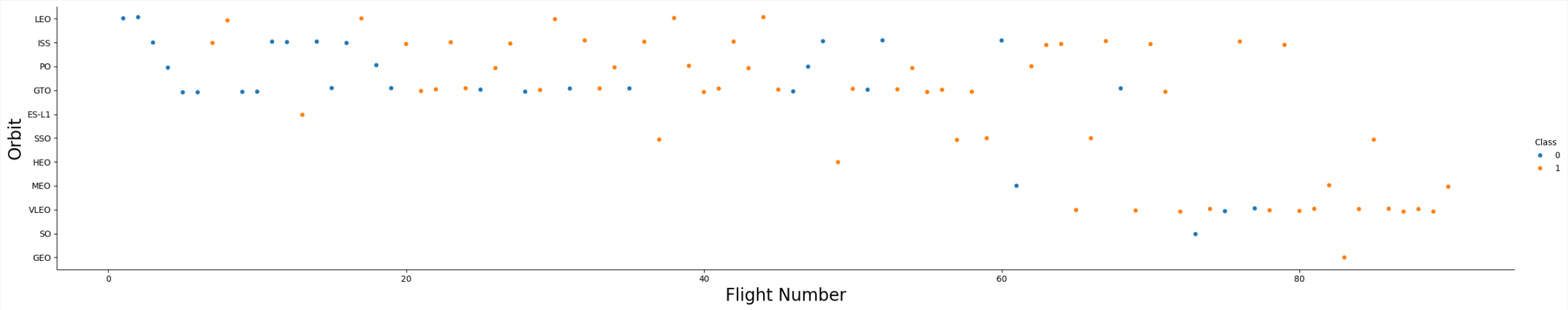
Success Rate vs. Orbit Type

- This examination showed that orbit type did have an effect on mission success
- Four orbit types had almost complete success
- The remaining orbit types were close to 50/50 and so those orbit types were not good predictors of sucess



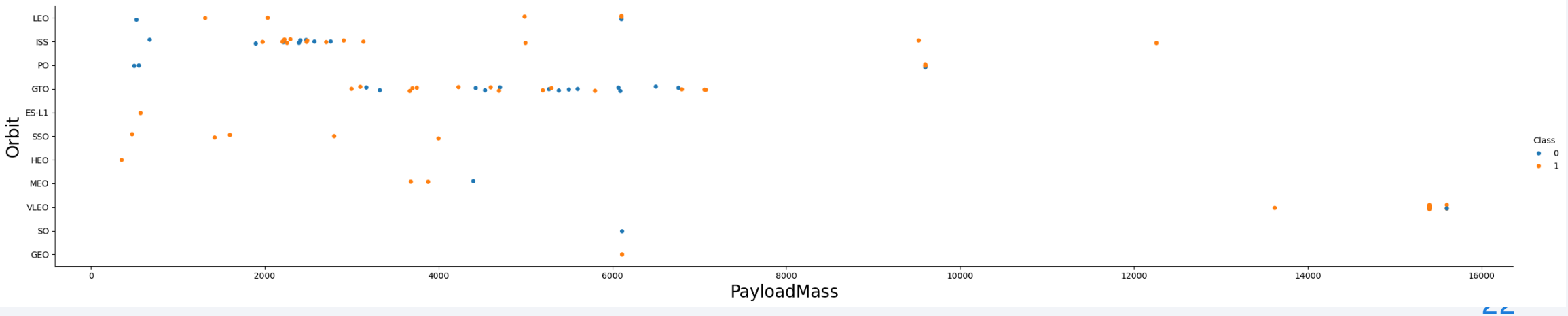
Flight Number vs. Orbit Type

- Since flight number is sequential, we can see that overall orbit type changed over time with visual analysis seeming to indicate that mission success decreased.
- Other analyses did not show that mission success decreased and so this requires further investigation to understand



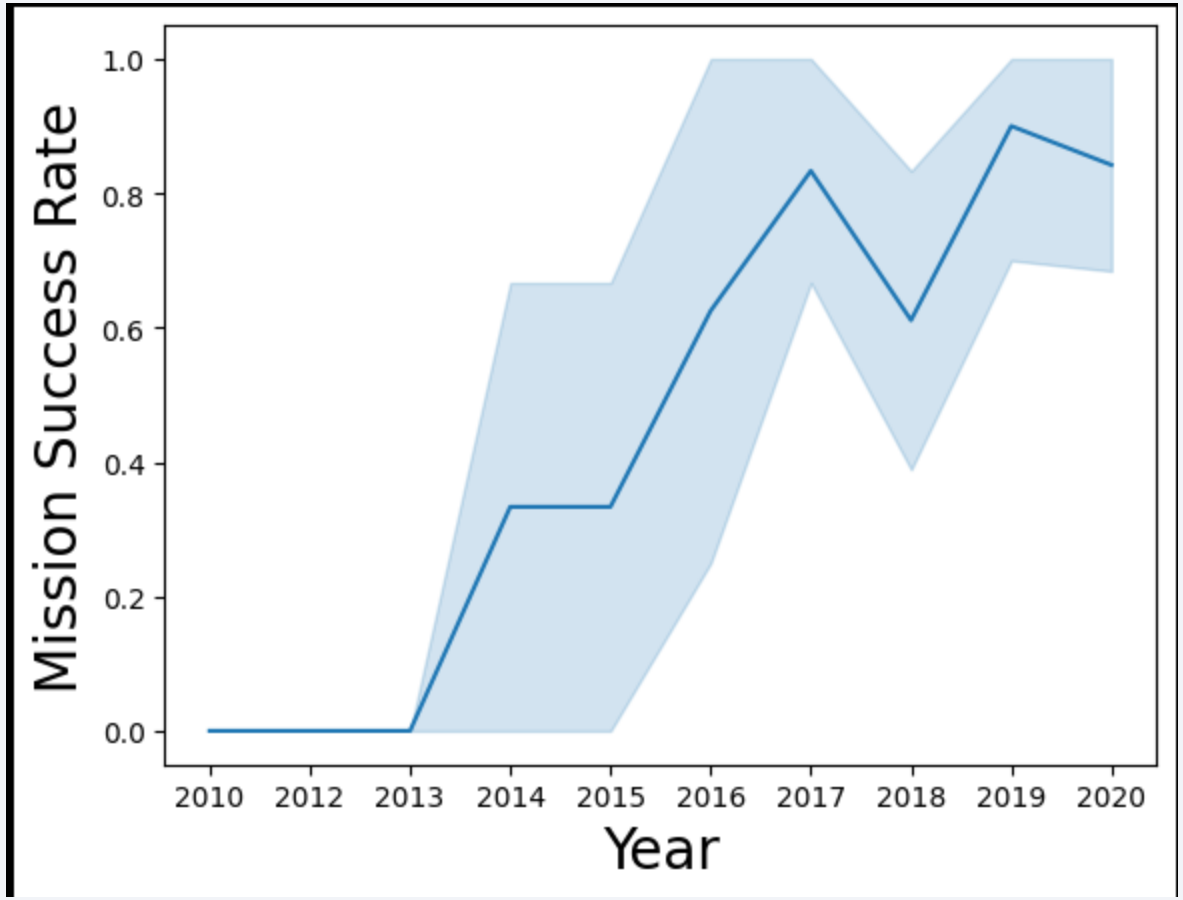
Payload vs. Orbit Type

- Most payloads were below 8000 kg
- Rough visual analysis shows that mission success is greater with lower payload mass



Launch Success Yearly Trend

- Contrary to earlier visual analysis, we can see that mission success rate definitely increase over time
- The greatest increase was between 2015 & 2017
- Improvement levels off after 2017



All Launch Site Names

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

```
select distinct Launch_Site from SPACEXTABLE;
```

The SpaceX Table is a list of all launches and the query selects only the unique launch site names

Launch Site Names Begin with 'CCA'

```
select * from SPACEXTABLE where Launch_Site like "%CCA%" LIMIT 5;
```

The query extracts all launch sites with CCA in the name and returns 5 results

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

Total Payload Mass

```
select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer
= "NASA (CRS)";
```

The query sums payload mass from all launches for NASA and returns 45596 kg

Average Payload Mass by F9 v1.1

```
select AVG(PAYLOAD_MASS__KG_) from
SPACEXTABLE                    where Booster_Version like "F9
v1.1%";
```

The query determines the average payload of all F9 v1.1 boosters to be 2534.7 kg

First Successful Ground Landing Date

```
select min(Date) from SPACEXTABLE where Landing_Outcome like
"%SUCCESS%";
```

The query extracts all launches with "Success" in the outcome text and returns the earliest suceess to have been on 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

- `select distinct Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)';`
- This query did not take into account the mass restrictions, but returns all successful drone landings
- I was unable to reboot the SQL environment after I discovered this error

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

```
select count(Mission_Outcome) as Success  from SPACEXTABLE
where Mission_Outcome like '%Success%';

select count(Mission_Outcome) from SPACEXTABLE where
Mission_Outcome like '%Failure%';
```

- A better method would have been to query the total number of outcomes using a GROUPBY statement
- Success = 100
- Failure = 1
- (Clearly something wrong here)

Boosters Carried Maximum Payload

```
select distinct(Booster_Version) from SPACEXTABLE where
PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from
SPACEXTABLE);
```

Query results seem to indicate the the F9 B5 booster type is designed for larger payloads with the remainder of the name indicating a possible serial number and reuse number

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

2015 Launch Records

- select substr(Date, 6,2) as month, Booster_Version, Landing_Outcome, Launch_Site from SPACEXTABLE where substr(Date,0,5)='2015';

month	Booster_Version	Landing_Outcome	Launch_Site
01	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
02	F9 v1.1 B1013	Controlled (ocean)	CCAFS LC-40
03	F9 v1.1 B1014	No attempt	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1016	No attempt	CCAFS LC-40
06	F9 v1.1 B1018	Precluded (drone ship)	CCAFS LC-40
12	F9 FT B1019	Success (ground pad)	CCAFS LC-40

It looks like when I created this query I neglected to add in the WHERE clause I needed in order to exclude the successful ground pad landing.

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

```
select Landing_Outcome, count(Landing_Outcome) as count from
SPACEXTABLE where date between 2010-06-04 and 2017-03-20;
```

The query automatically returns the result in a ranked form, but a better method would have been to use an ORDER BY clause


Section 3
Launch Sites
Proximities Analysis

SpaceX Launch Sites

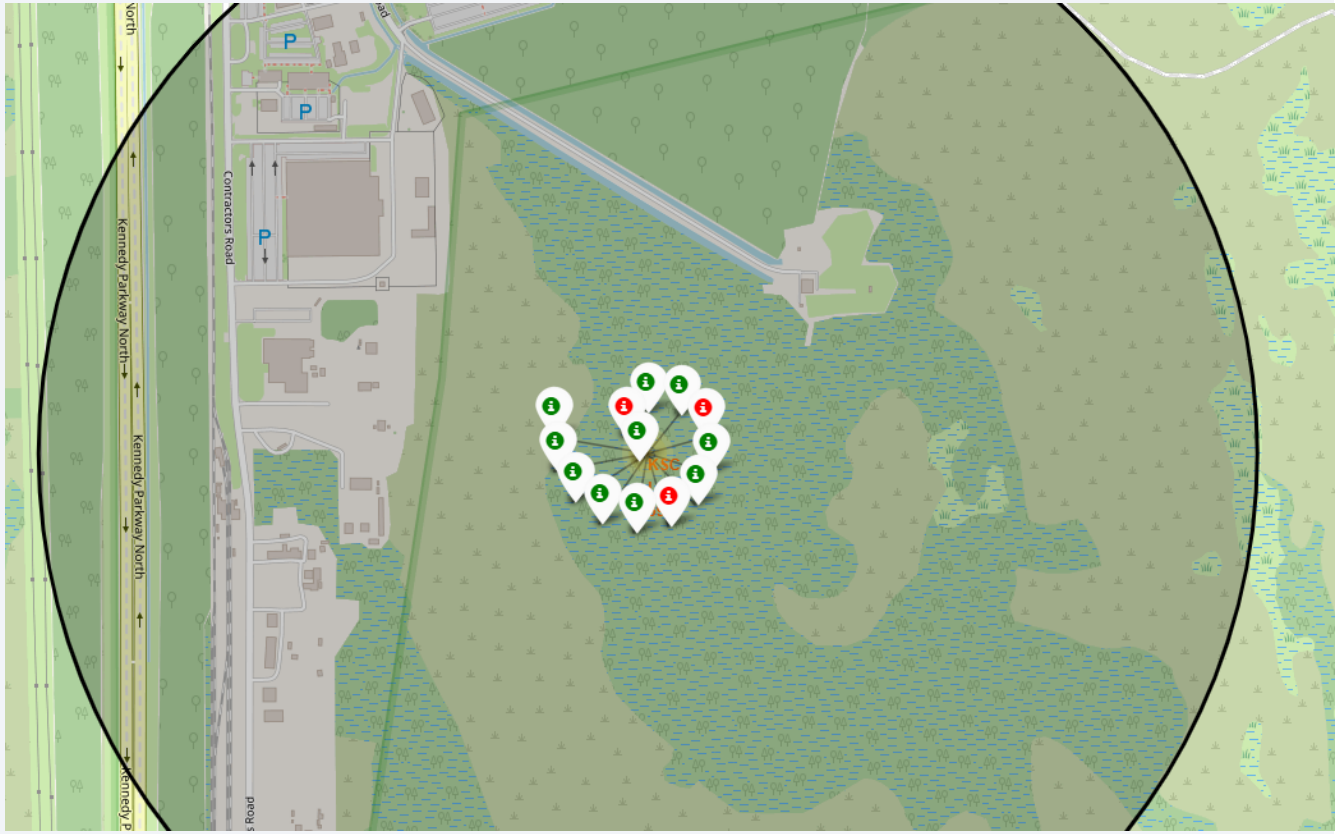
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 1>

- Replace <Folium map screenshot 1> title with an appropriate title
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot

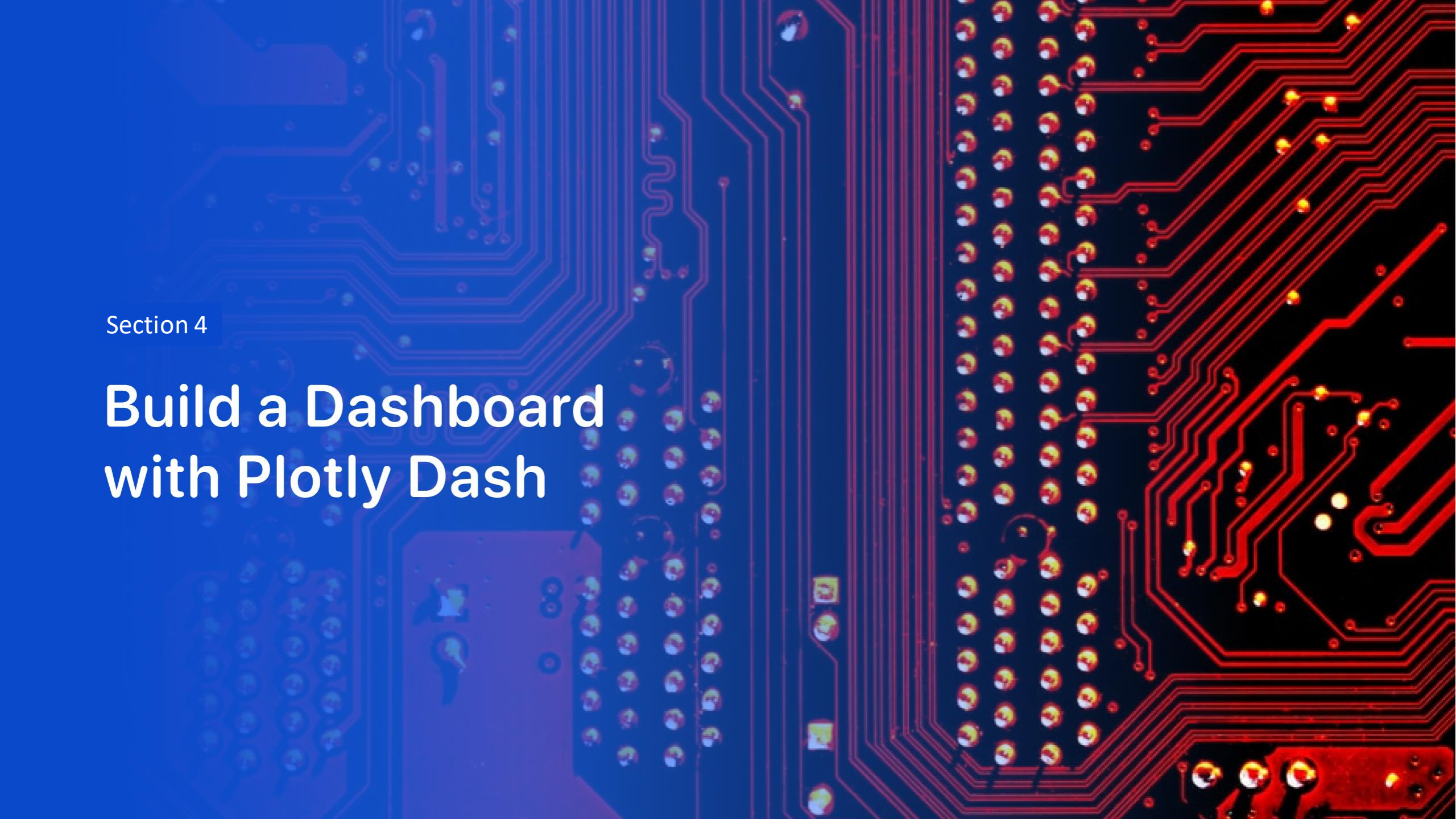
Illustrating launch success within sites

- Green markers indicate a launch success while red indicate mission failures.
- Clicking on markers provides more information

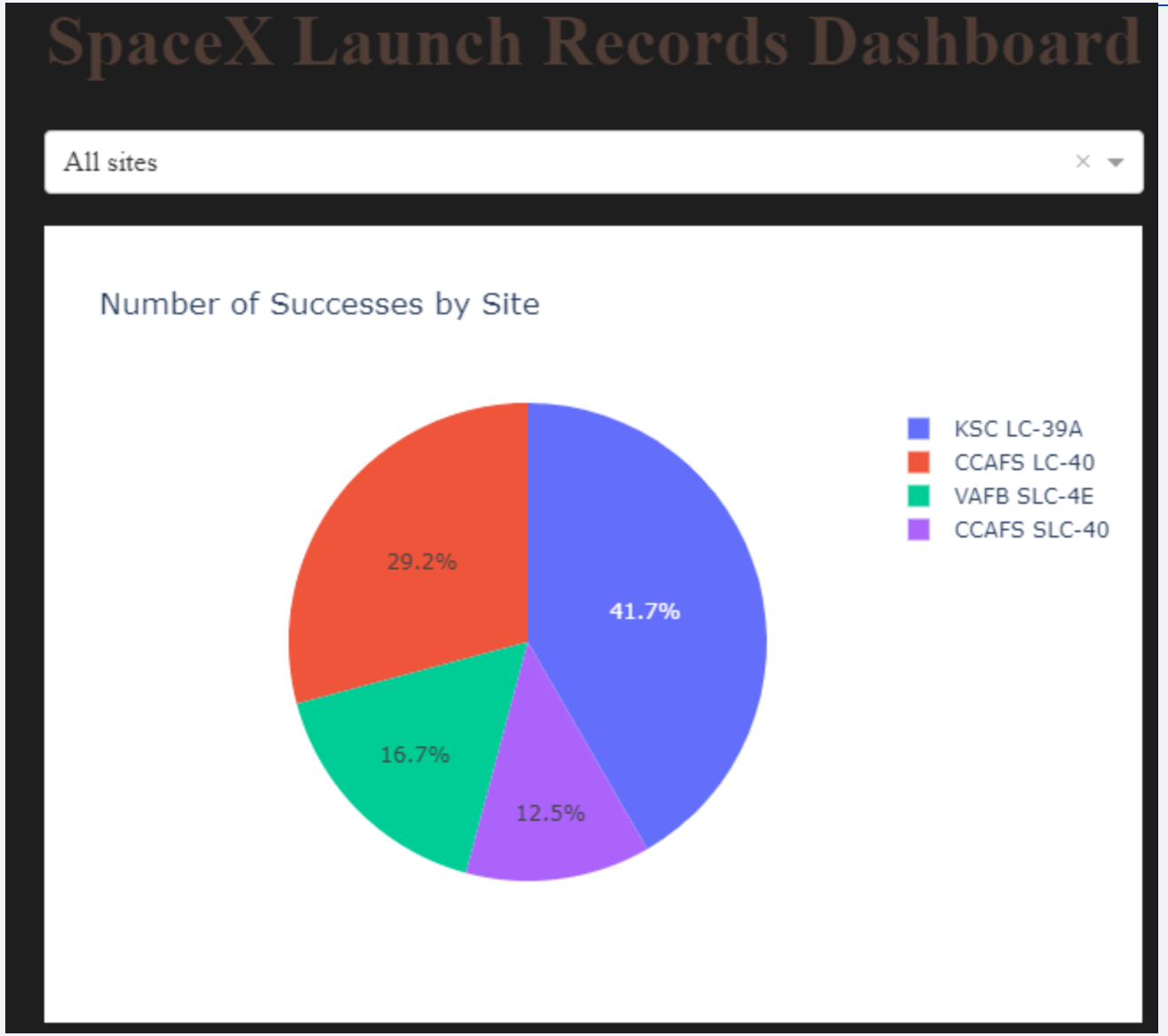


Allowing Adjacency Measurements

- Using the coordinates of a point of interest creates a line between the selected launch site and the point of interest
- Here, you can see that the launch site is 0.70 km from the nearest railway


Section 4
Build a Dashboard with Plotly Dash

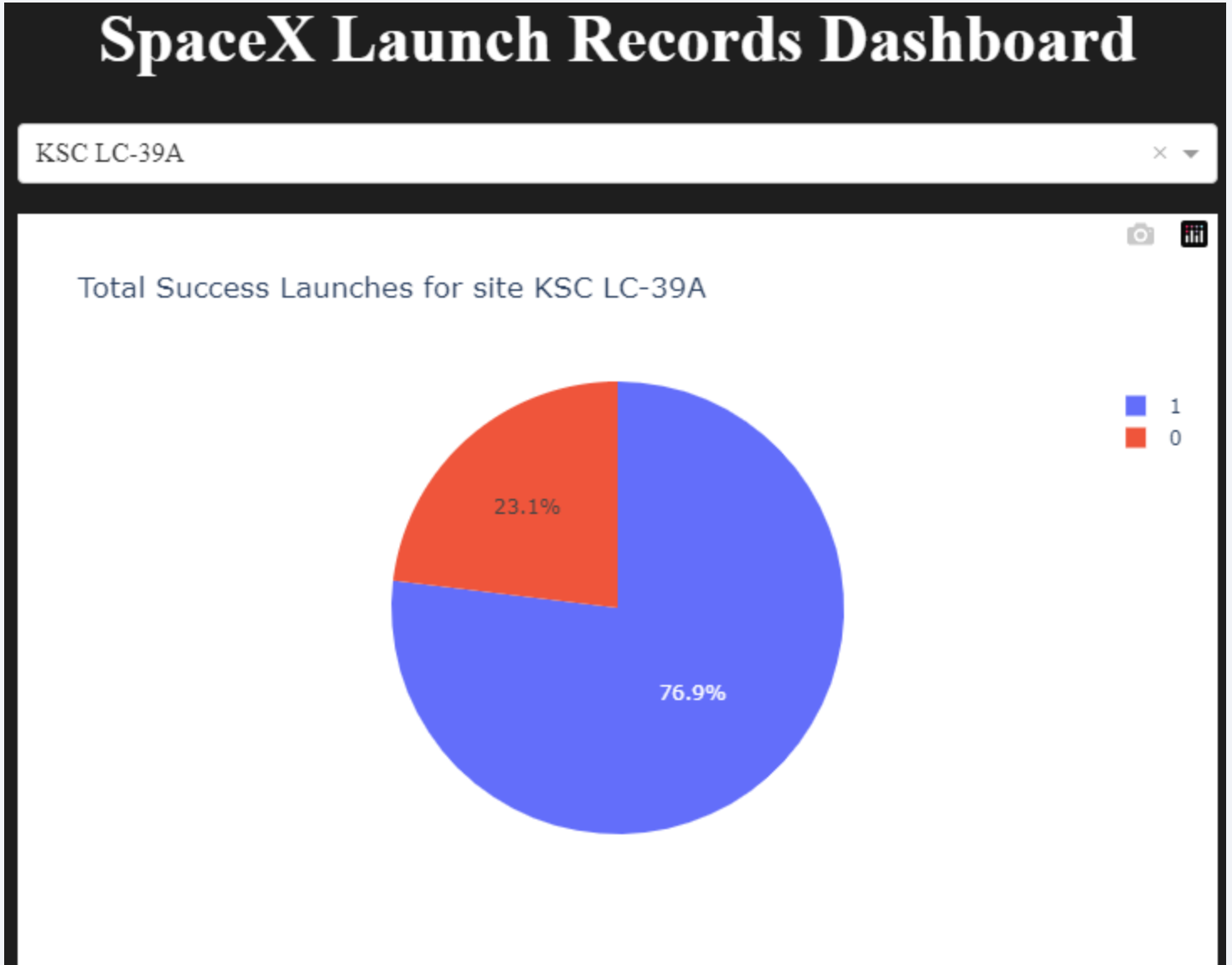
Launch Success by Site



Of the four SpaceX launch sites in this study, the most successful launches were from Kennedy Space Center Launch Complex 39A

Select Site Success

- Selecting a site from the dropdown changes the pie chart to show the success to failure proportions for the selected site.



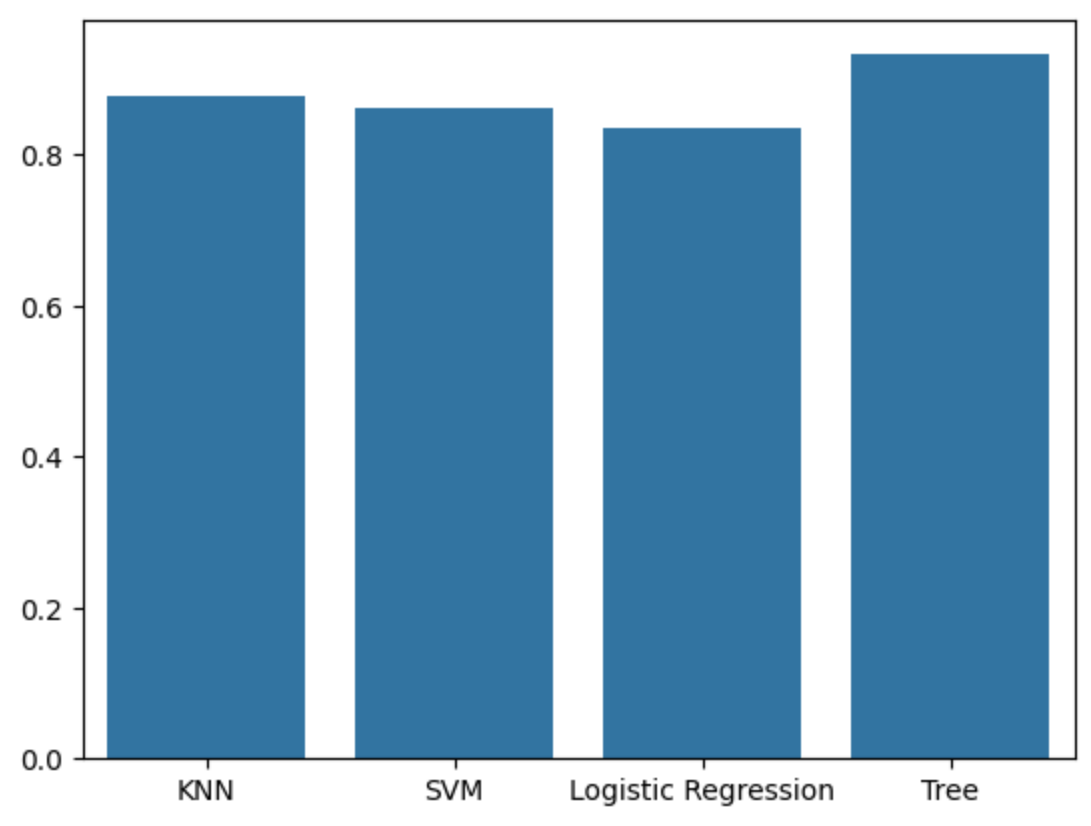
FT Booster Version Capacity

Note that the FT version booster has a severely reduced success rate for payloads over approximately 5500 kg

Section 5

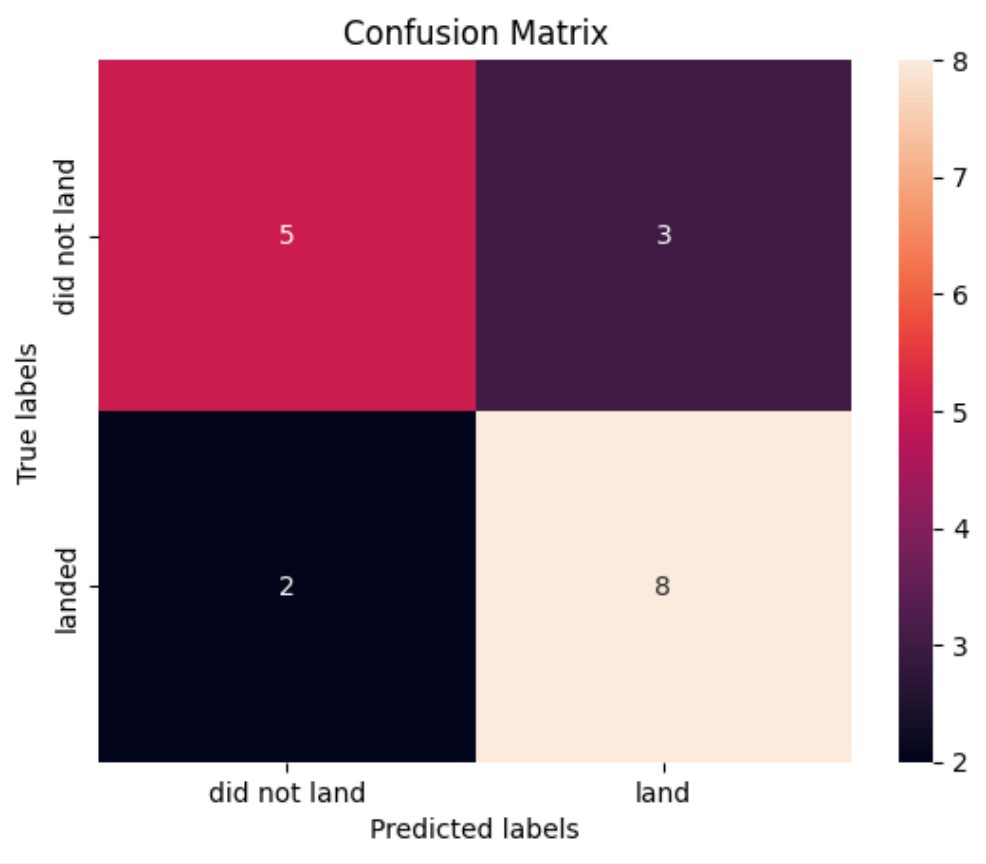
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree classification provides a better prediction of launch success than the other options.

Confusion Matrix



The confusion matrix shows the types of conclusions for each prediction

- True Positive: 8
- True Negative: 5
- False Positive: 3
- False Negative: 2

Conclusions

- A decision tree classifier should be used to determine future launch success
- Investigation should be made as to why launches are significantly better at KSC LC-39A
- Care should be taken to minimize payload mass when possible to maximize booster reuse
- Booster design has significantly improved launch success and root cause should be determined to drive further improvement

Appendix

All relevant code for this analysis can be found in the repository at:
https://github.com/cubeconvict/data-science/tree/main/IBM_capstone

Public data sources used:

- Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- SpaceX data API: https://api.spacexdata.com/v4/rockets/

Thank you!