IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fiona Tan Yoke Shuan
11 August 2022

Outline

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

Executive Summary

- This study is to understand what contributes to the success of the landing of the first stage of Falcon 9 rockets.
- Data was collected from SpaceX API and webscraping. Data was processed to remove null values and unnecessary features. Exploratory data analysis was performed to observe pattern in data using visualization, SQL queries, interactive dashboard and folium map. Finally, predictive classification models was built, and evaluation performed to select the best model.
- The success rate of landing increases over the year as technology improves. Lighter payload mass generally has higher success rate.
- The best predictive classification model for this study is the decision tree model.

Introduction

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. The goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully

- Problem: What factors determine the success landing of the first stage?

Section 1

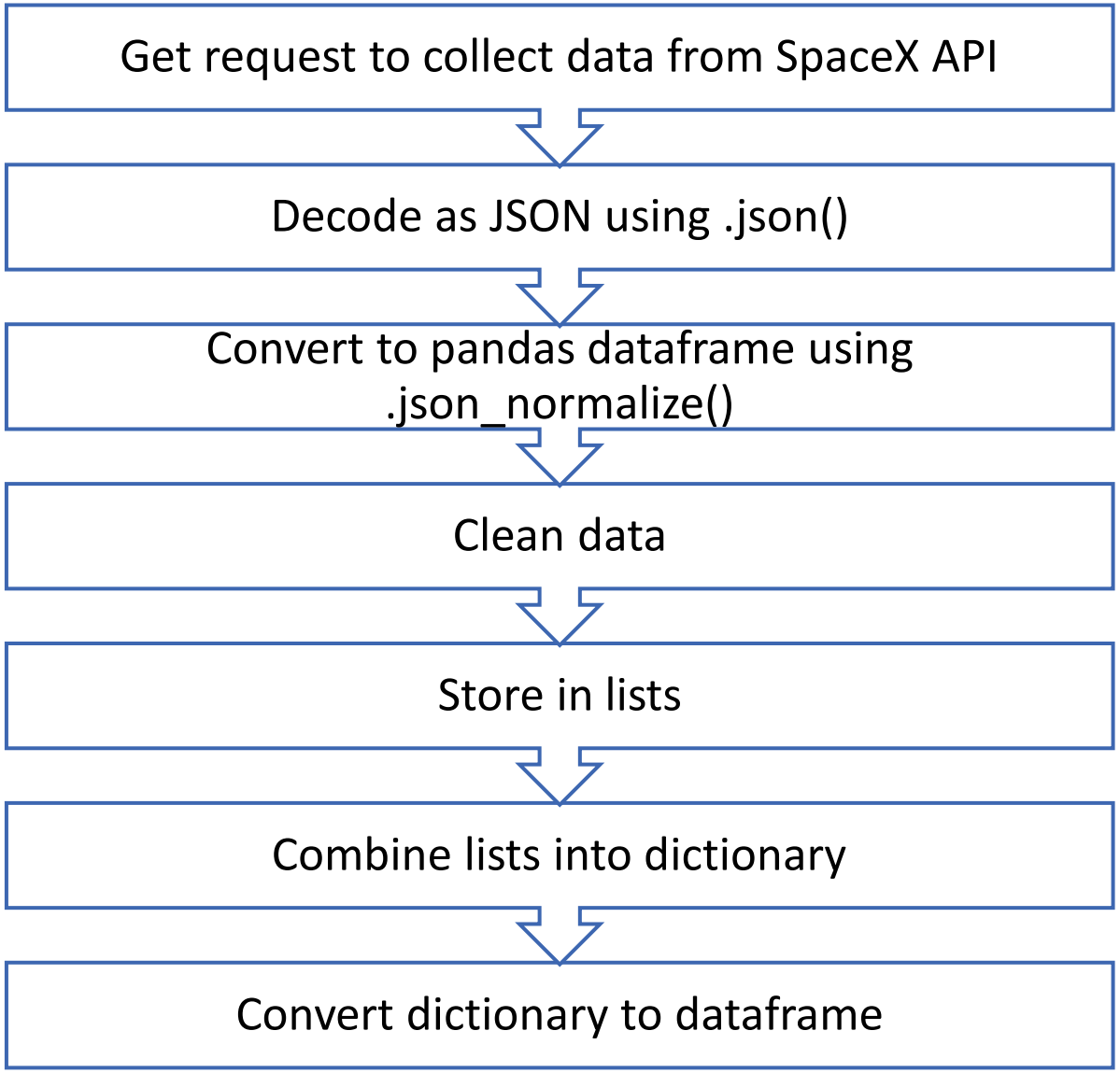
Methodology

Methodology

Executive Summary

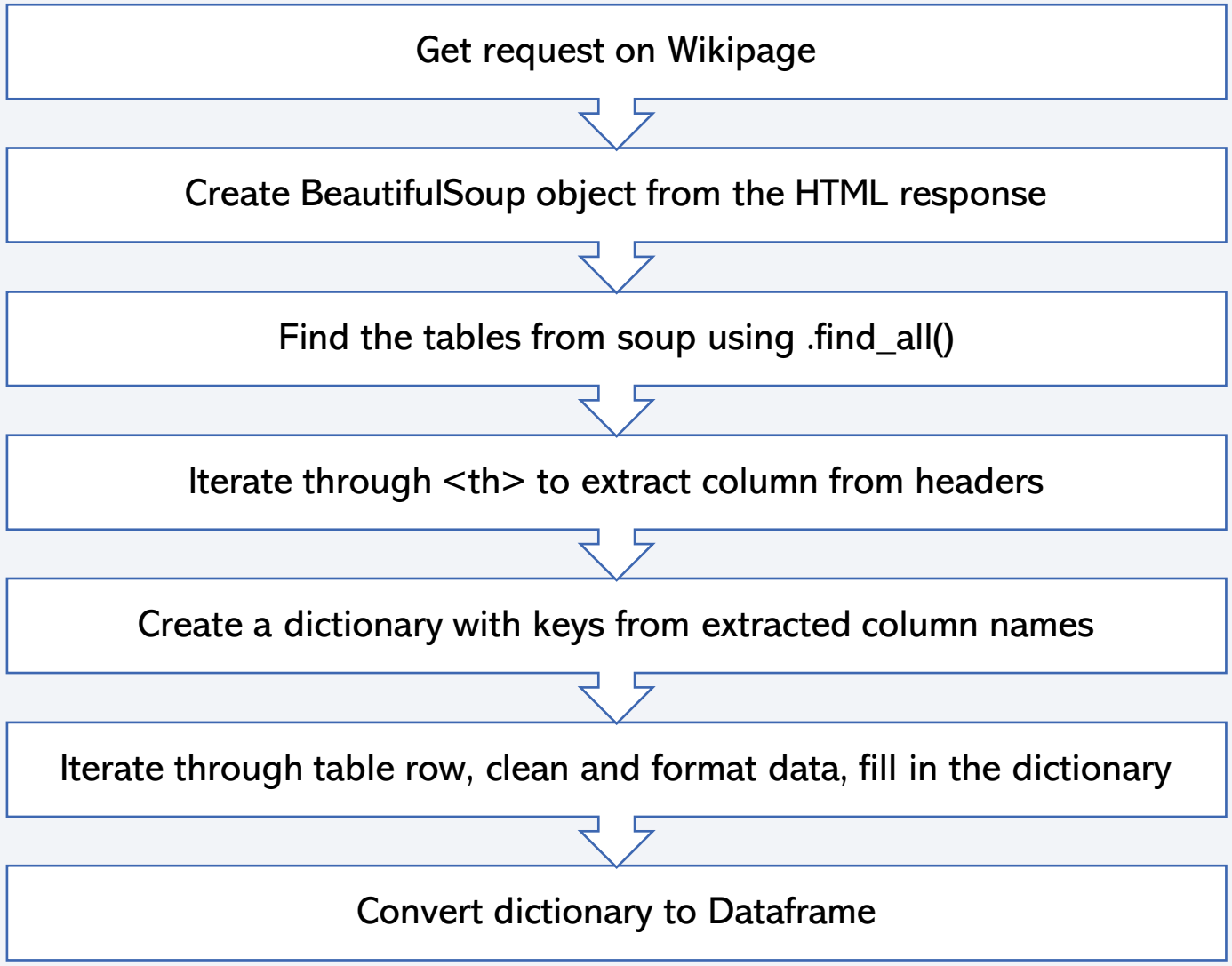
- Data collection methodology:
  - Data was collected from SpaceX API and webscraping from Wikipedia page
- Perform data wrangling
  - Clean null values and remove unnecessary columns
  - Classify landing outcomes into useful binary labels (successful / fail)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Load data, transform and standardize, split into training/testing sets, train models, evaluate model, tune with different hyperparameters, repeat with different models (KNN, Logistic Regression, Decision Tree, Support Vector Machine.

Data Collection – SpaceX API



https://github.com/fionatanys/ibm_capstone/blob/ece3b0075abcbd9cc7a8d3bd04441ad34e547593/Data_collection_API.ipynb

Data Collection - Web Scraping



https://github.com/fionatanys/ibm_capstone/blob/5e722654f8673ceb5eb8b1d4a3af20d94e3bc182/Data_collection_with_webscraping.ipynb

Data Wrangling

Check missing values using .isnull()

Identify types of columns (numerical / categorical) using .dtypes()

Determine number of launches on each site using .value_counts()

Calculate number and occurrence of each orbit using .value_counts()

Calculate the number and occurrence of mission outcome.

Iterate and label bad landing outcomes (second stage did not land successfully)

Assign to new variable landing_class

https://github.com/fionatanys/ibm_capstone/blob/5e722654f8673ceb5eb8b1d4a3af20d94e3bc182/Data_wrangling.ipynb

EDA with Data Visualization

- The following charts were plot to understand how the success rate is influenced by flight number, launch site, payload mass and orbit
  - Flight Number vs Payload
  - Flight Number vs Launch Site
  - Payload vs Launch Site
  - Success Rate vs Orbit Type
  - Flight Number vs Orbit Type
  - Payload vs Orbit Type
  - Launch Success Yearly Trend

https://github.com/fionatanys/ibm_capstone/blob/5e722654f8673ceb5eb8b1d4a3af20d94e3bc182/EDA_with_visualisation.ipynb

EDA with SQL

- The following SQL were performed:
  - Names of unique launch sites
  - Launch site names begin with '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 greater than 4000 but less than 6000
  - 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
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

https://github.com/fionatanys/ibm_capstone/blob/8eb5ecabc721d8747530d6e8bb6ab1b70df11665/EDA_with_SQL.ipynb

Build an Interactive Map with Folium

- Circle and marker are added to mark all launch sites to the map
- Green markers are added for success launches and red markers added for failed launches for each site.
- All fail / success launches are added for each site
- Distances from city, coastline, railway and highway are added to evaluate if all location of launch sites are related to these infrastructure.

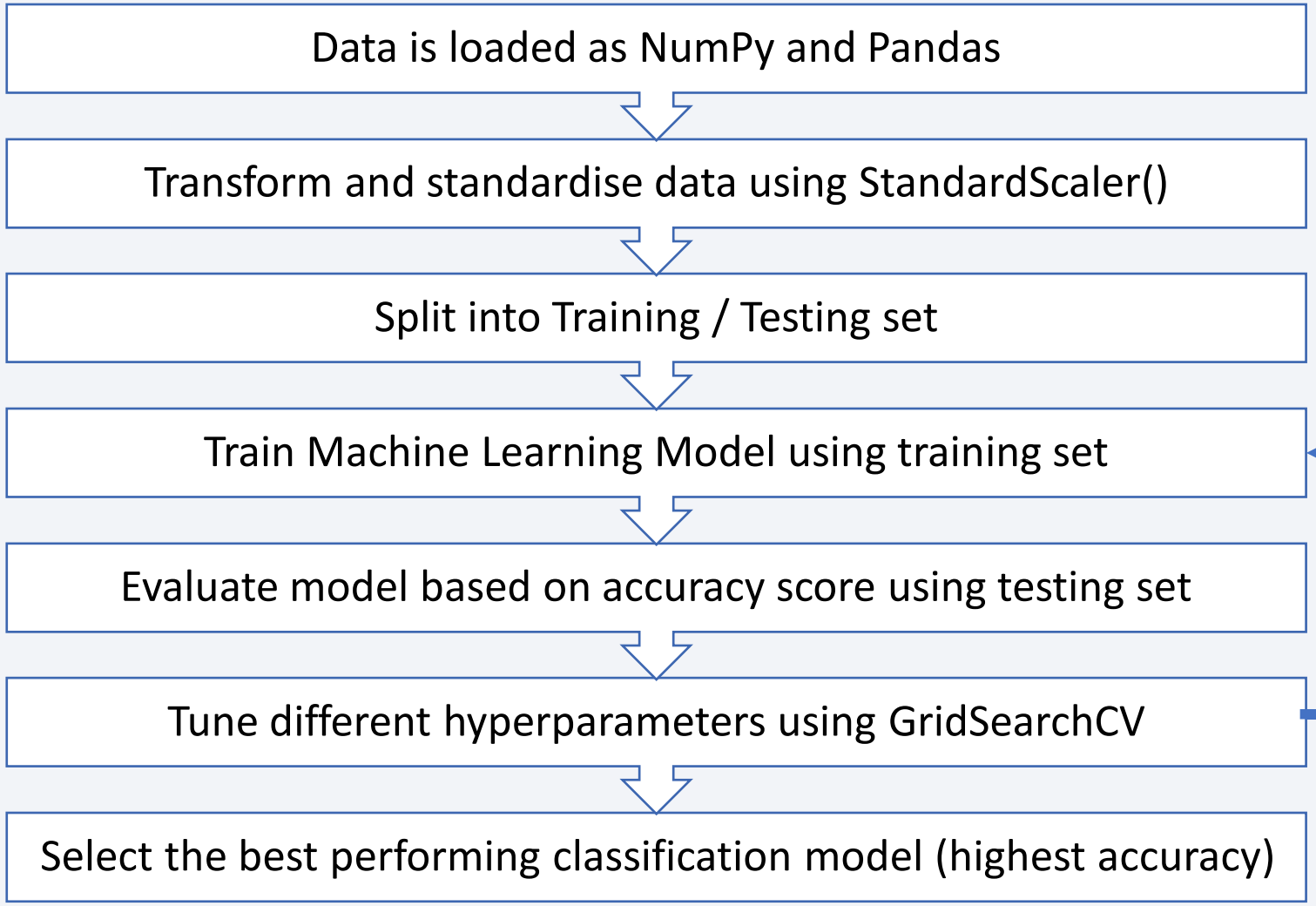
https://github.com/fionatanys/ibm_capstone/blob/5e722654f8673ceb5eb8b1d4a3af20d94e3bc182/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie chart was added to show total success launches for all sites – to show which launch sites has the most success launches
- A dropdown value can be selected to plot success rate for each launch site
- A scatter plot with payload mass range slider, with categories of different booster – to explore influence of payload mass and booster version on the success rate of launches.

https://github.com/fionatanys/ibm_capstone/blob/5e722654f8673ceb5eb8b1d4a3af20d94e3bc182/spacex_dash_app.py

Predictive Analysis (Classification)

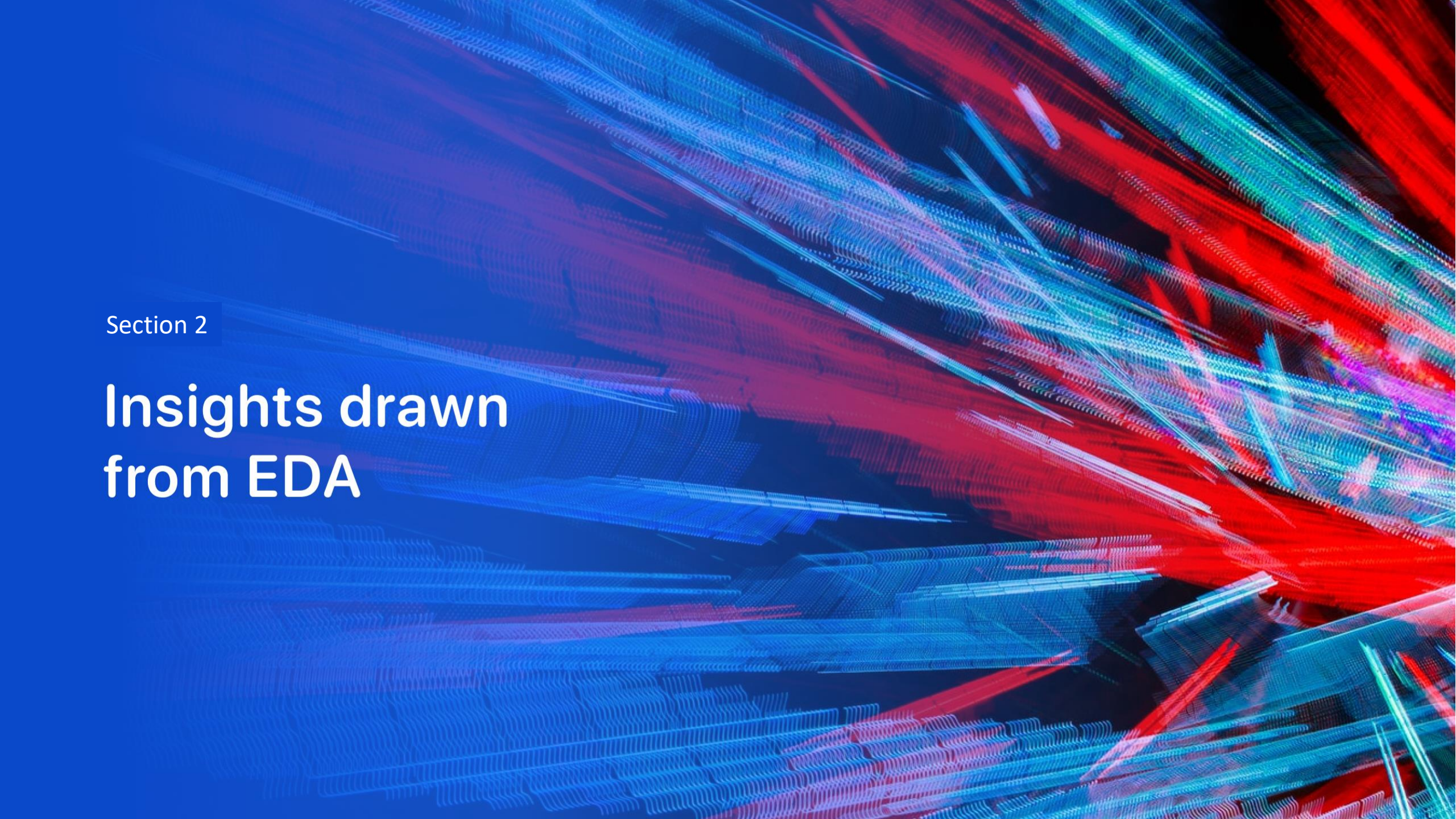


https://github.com/fionatanys/ibm_capstone/blob/ece3b0075abcbd9cc7a8d3bd04441ad34e547593/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Repeat for different ML algorithm:
KNN
Logistic Regression
Decision Tree

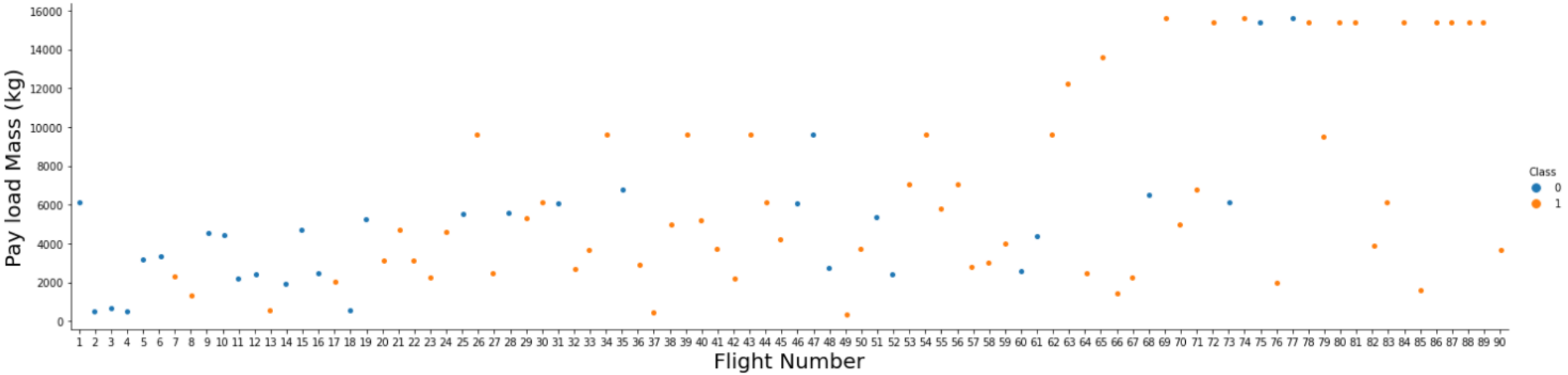
Results

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


Section 2
Insights drawn from EDA

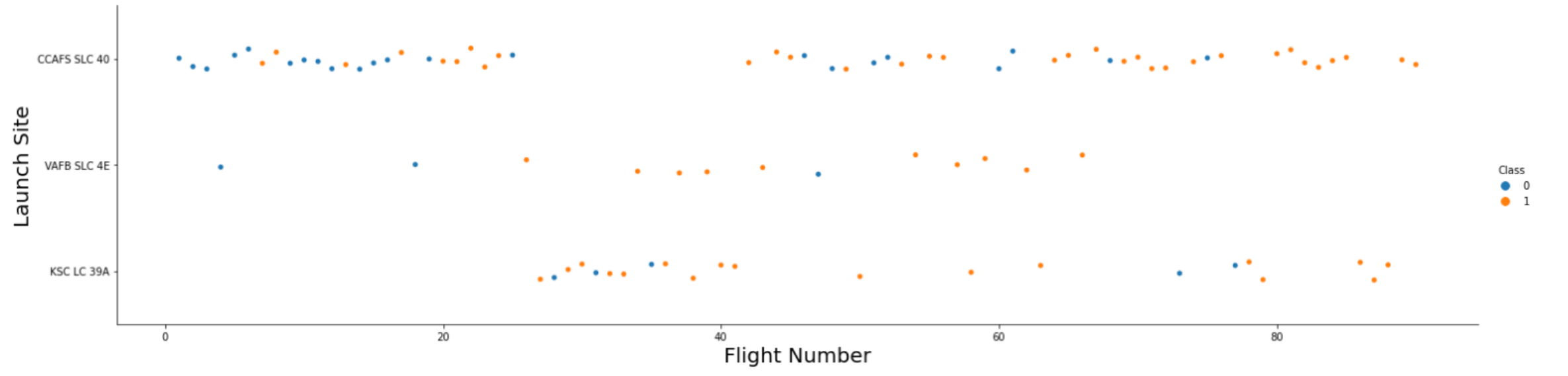
Flight Number vs. Launch Site

- We see that as the flight number increases, the first stage is more likely to land successfully.
- The heavier the payload, the less likely the first stage will return.



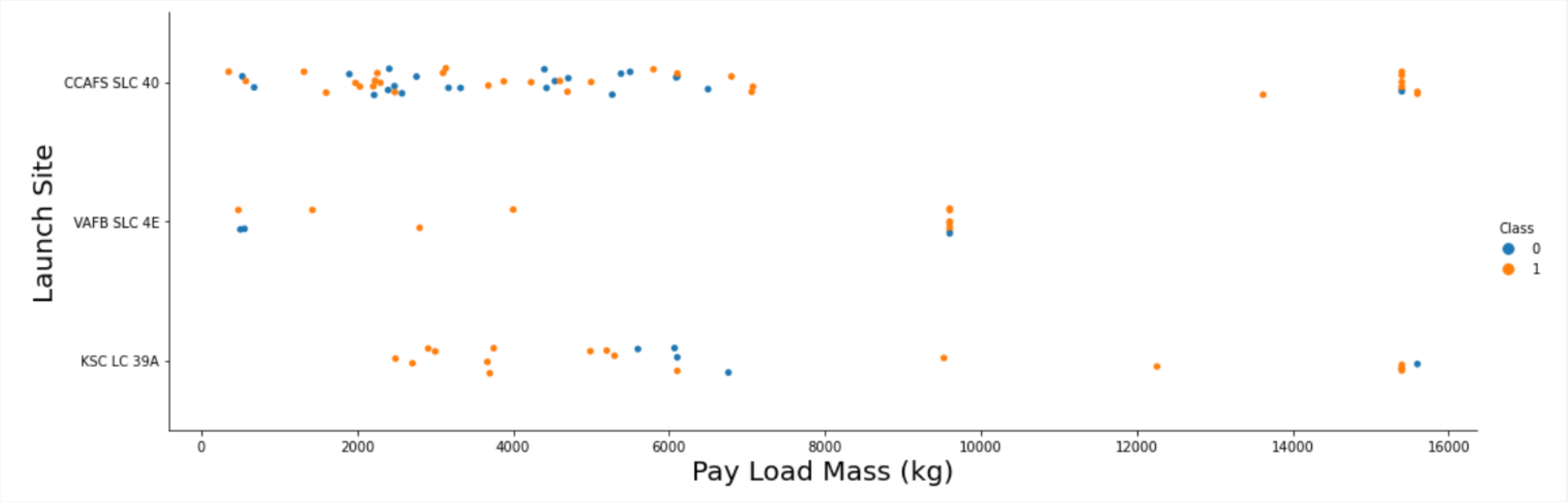
Flight Number vs. Launch Site

- Greater success rate with increase flight number.
- CCAFS SLC 40 has more number of launches rate and high success rate



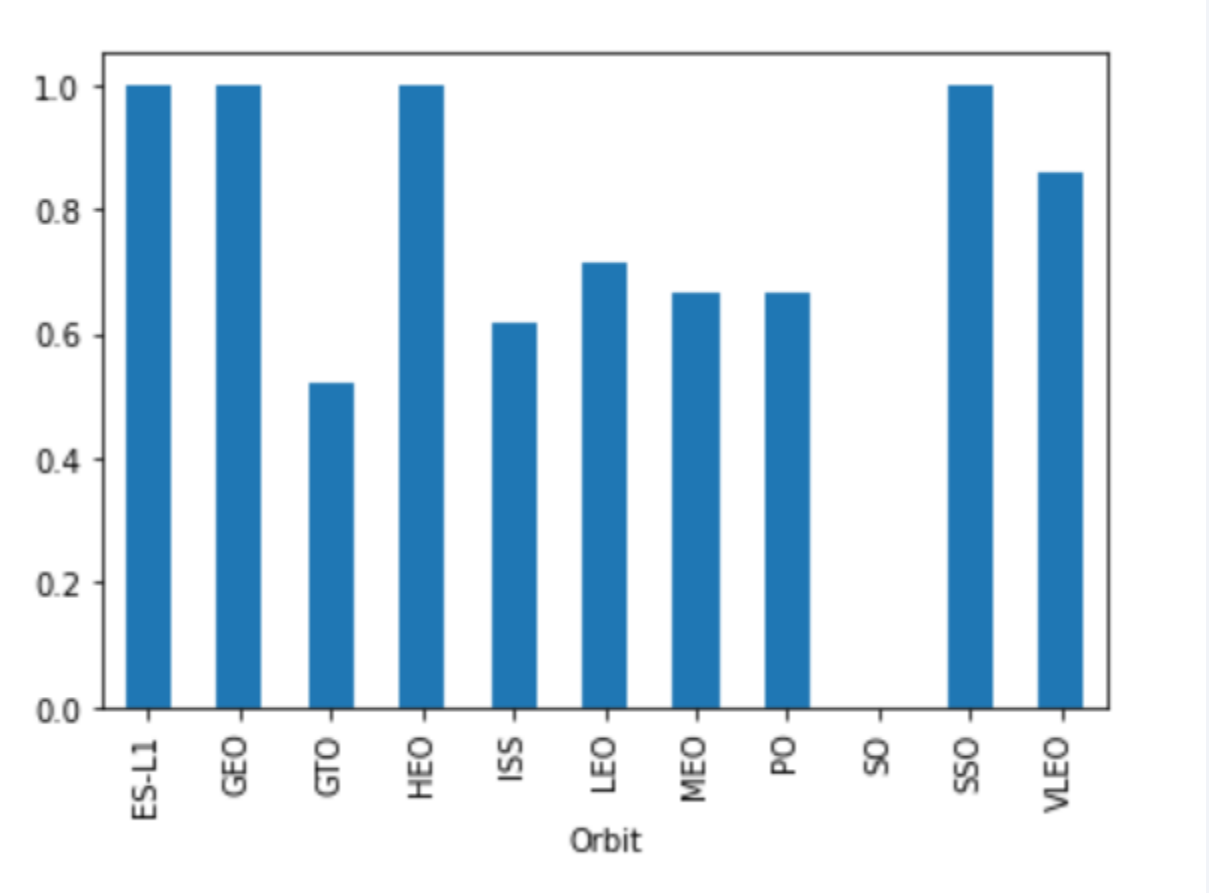
Payload vs. Launch Site

- CCAFS SLC 40 has very high success rate for launches with heavy payload mass (>15000kg)
- VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).
- KSC LC 39A has 100% success rate for launches with light payload mass (<5000kg)



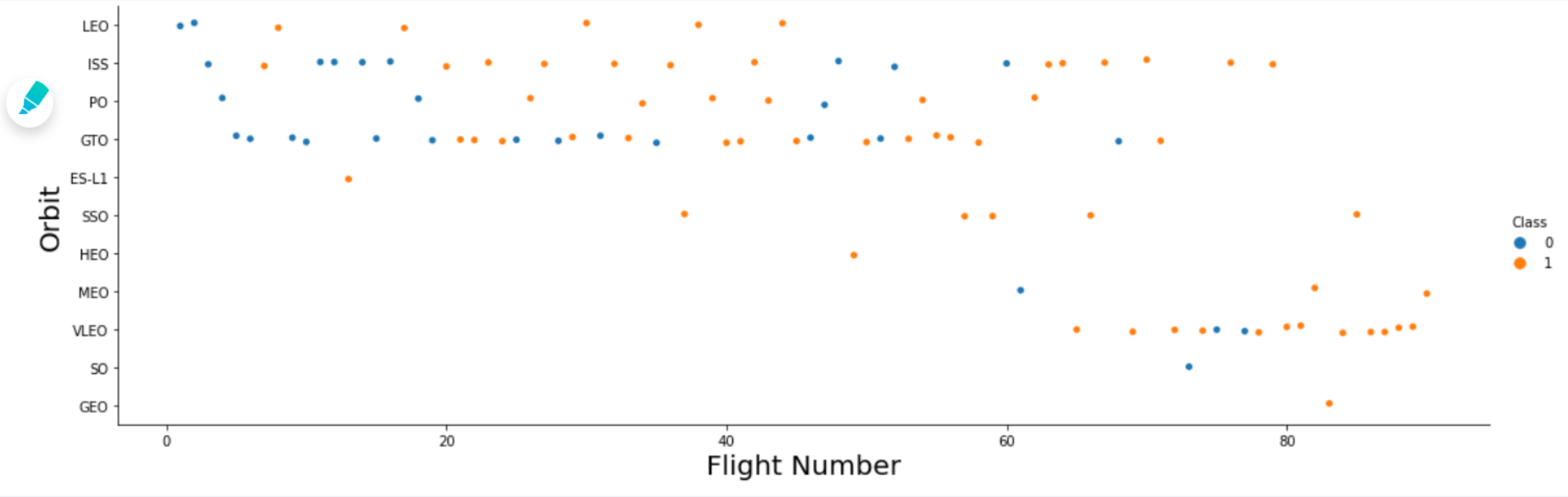
Success Rate vs. Orbit Type

- SO orbit has zero success rate.
- ES-L1, GEO, HEO, SSO Orbit has the highest success rate



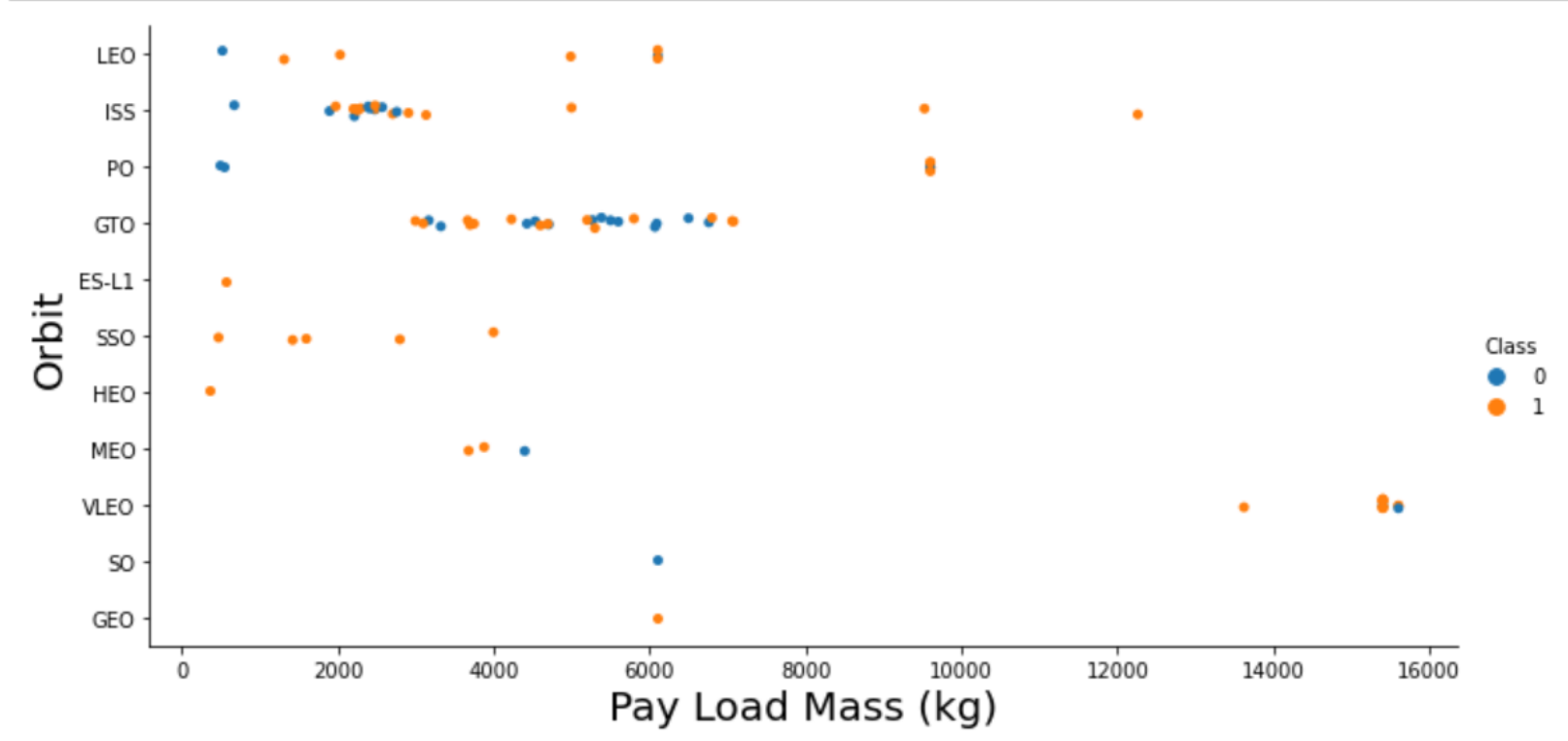
Flight Number vs. Orbit Type

- More VLEO launches in recent years.
- In LEO orbit, success rate increases with flight number.



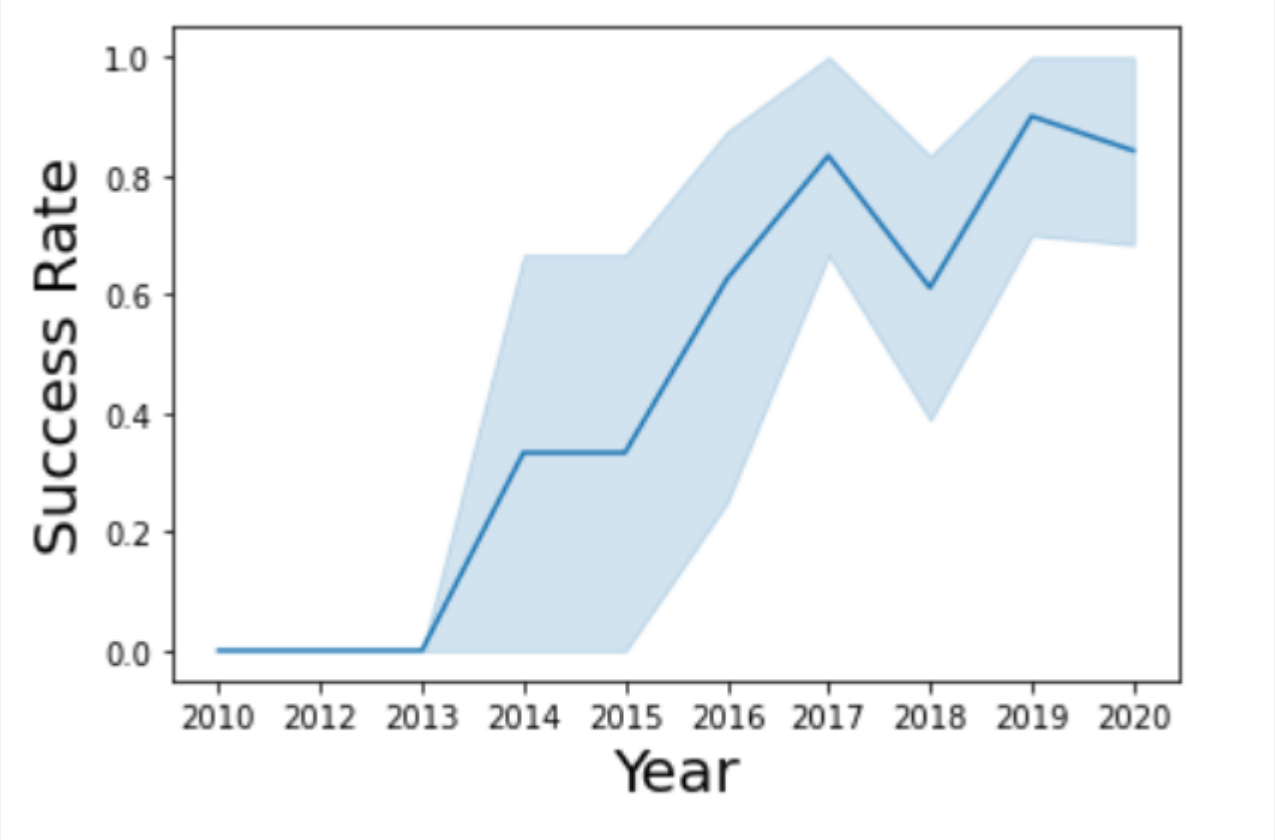
Payload vs. Orbit Type

- VLEO orbit only has launches with heavy payload (>13000kg)
- GTO orbit only has payload between 3000-8000kg
- ISS mostly has payload between 2000-4000kg
- SSO orbit only has payload <4000kg
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- Success rate has increased significantly between 2013 to 2017 and has stabilised between 2017-2020.



All Launch Site Names

- Uses DISTINCT to show unique values of launch sites

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

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

Launch Site Names Begin with 'CCA'

- Uses Where to set conditions; uses LIKE to suggest names start with ‘CCA%’; Uses Limit 5 to limit number of records

```
%%sql
SELECT *
FROM SPACEXDATASET
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	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

Total Payload Mass

- Uses SUM() to calculate total payload mass; uses WHERE to set condition i.e. only boosters from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

1
45596

Average Payload Mass by F9 v1.1

- Use WHERE to filter record with booster version = F9 v1.1,
- Use AVG() to calculate average payload mass

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXDATASET
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

1
2928

First Successful Ground Landing Date

- Use WHERE to select successful landing outcome in ground pad
- Use MIN(DATE) to select the earliest record

```
: %%sql
SELECT MIN(DATE)
FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Use WHERE to select record with success in drone ship
- Use BETWEEN to select range of payload mass

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

Total Number of Successful and Failure Mission Outcomes

- Use GROUP BY to group records according to MISSION_OUTCOME
- Use Count(*) to count number of records

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*)
FROM SPACEXDATASET
GROUP BY MISSION_OUTCOME;
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Use a subquery to find out maximum payload mass
- Use WHERE to condition payload mass = max payload mass from subquery

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

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

- Use WHERE to filter records with failure in drone ship and launch year in 2015
- Use YEAR() to select the year from date.

```
%%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Failure (drone ship)'
    AND YEAR(DATE) = 2015;
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

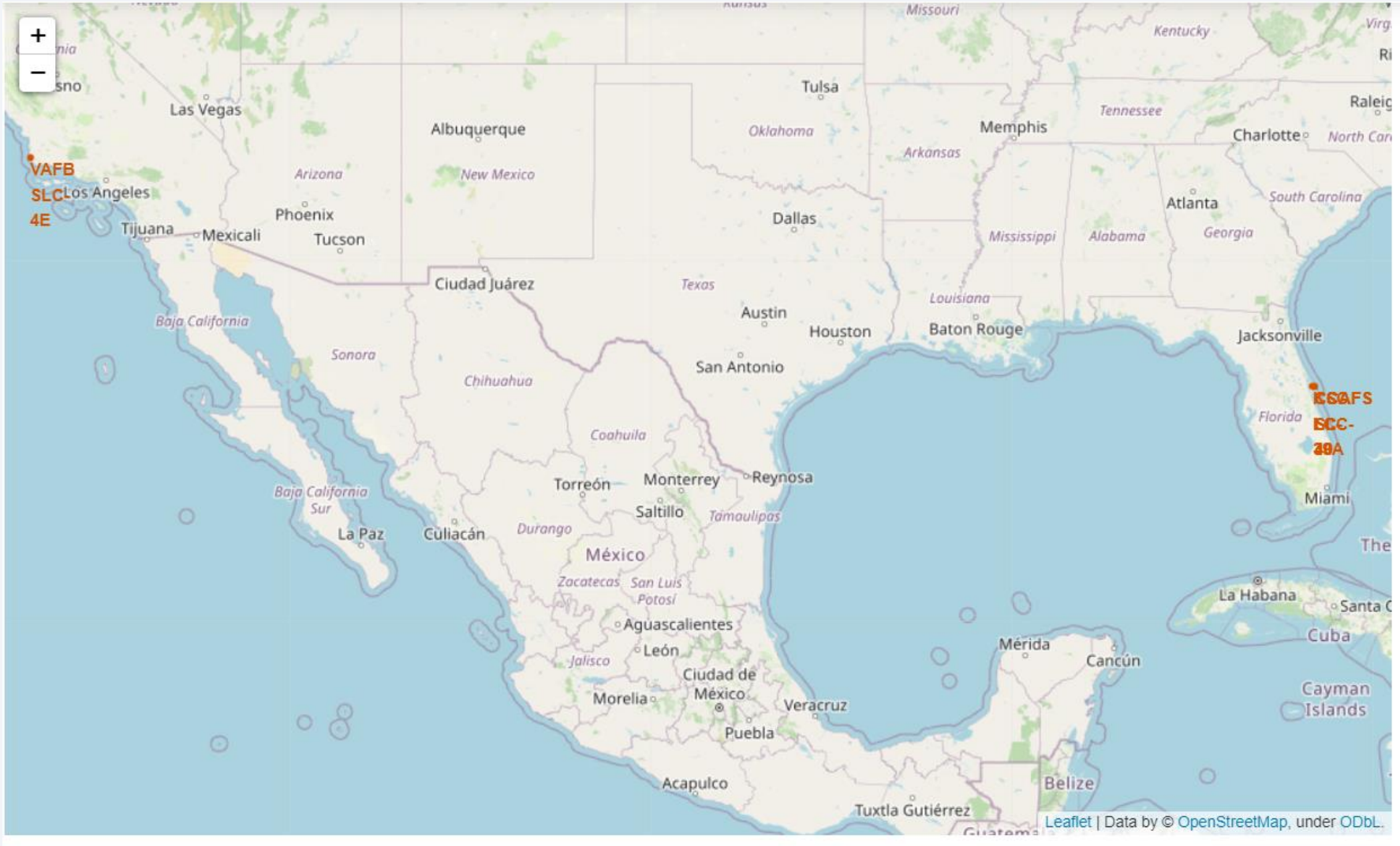
- Use WHERE and BETWEEN to select records between 2010-06-04 and 2017-03-20.
- Use GROUP BY to sort records by landing outcomes
- Use COUNT(*) to count number of records
- Use ORDER BY and DESC to rank the records

```
%%sql
SELECT LANDING__OUTCOME, count(*)
FROM SPACEXDATASET
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY count DESC;
```

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

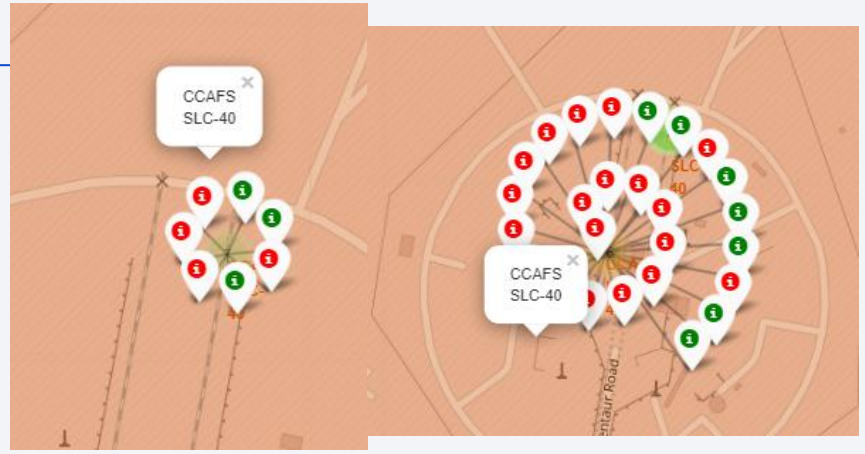

Section 3
Launch Sites
Proximities Analysis

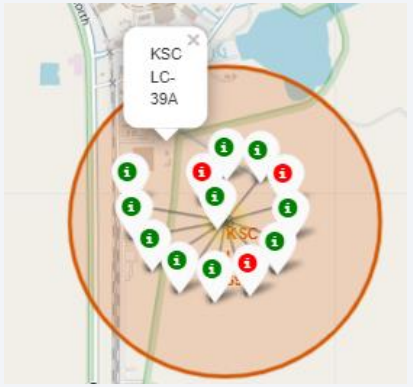
All launch sites marked on map

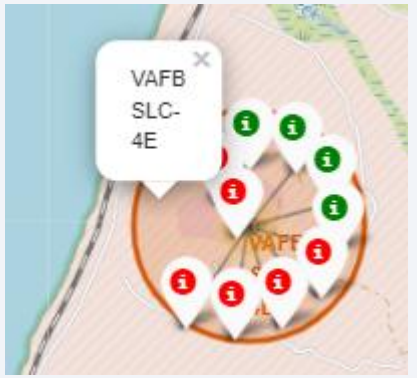


- SpaceX launch sites are near the coastline of the United States of America i.e near Florida and California

Colour-labeled launch record

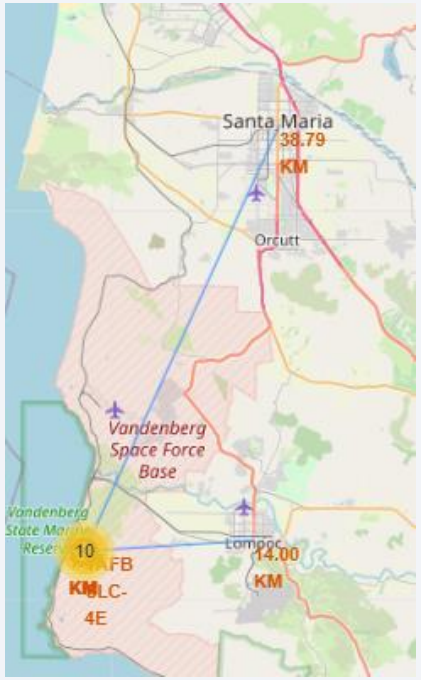


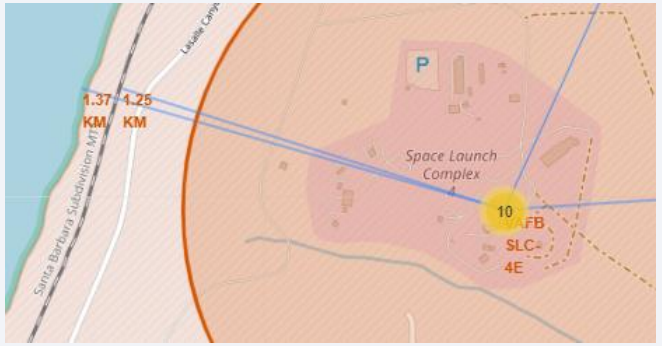


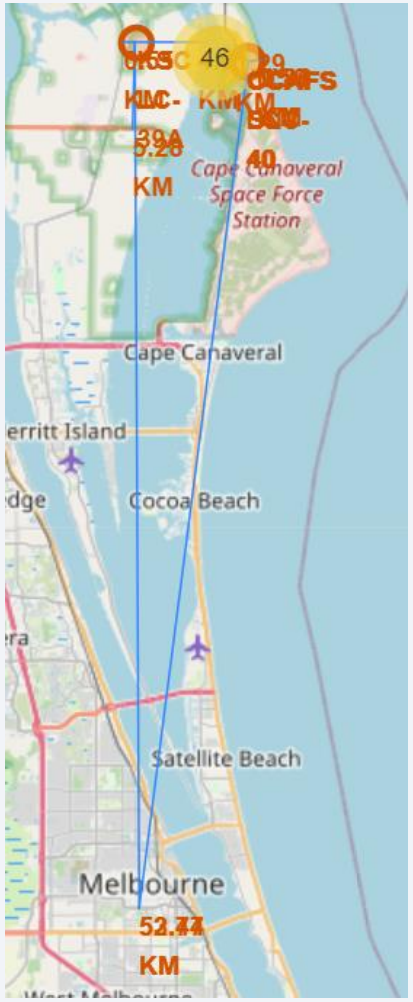


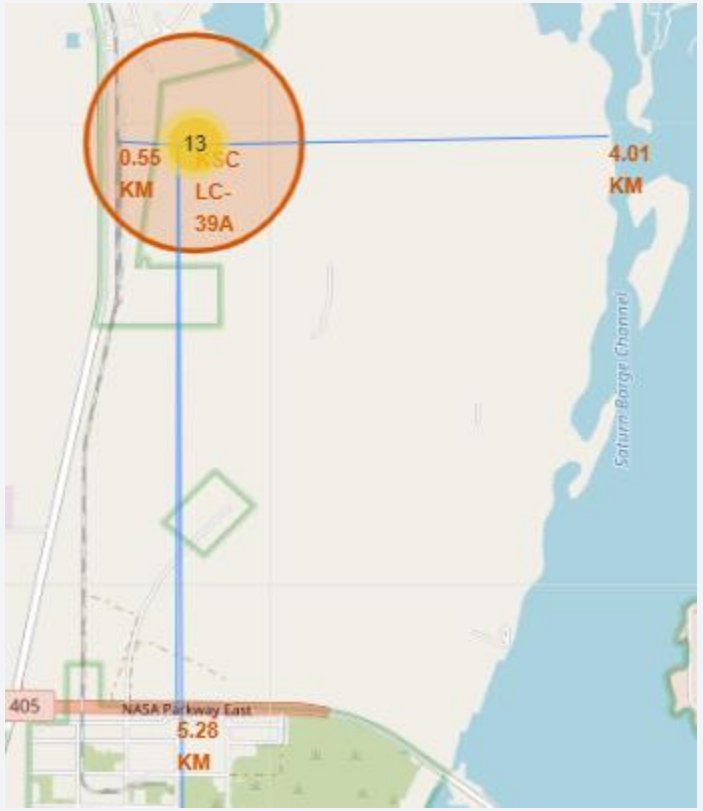
- Red labels indicate failed launches whereas green labels indicate successful launches. It can be observed that KSC LC-39A has highest probability of successful launches even though CCAFS SLC-40 has the most number of launches

Distance from coastline, highway, railway and city





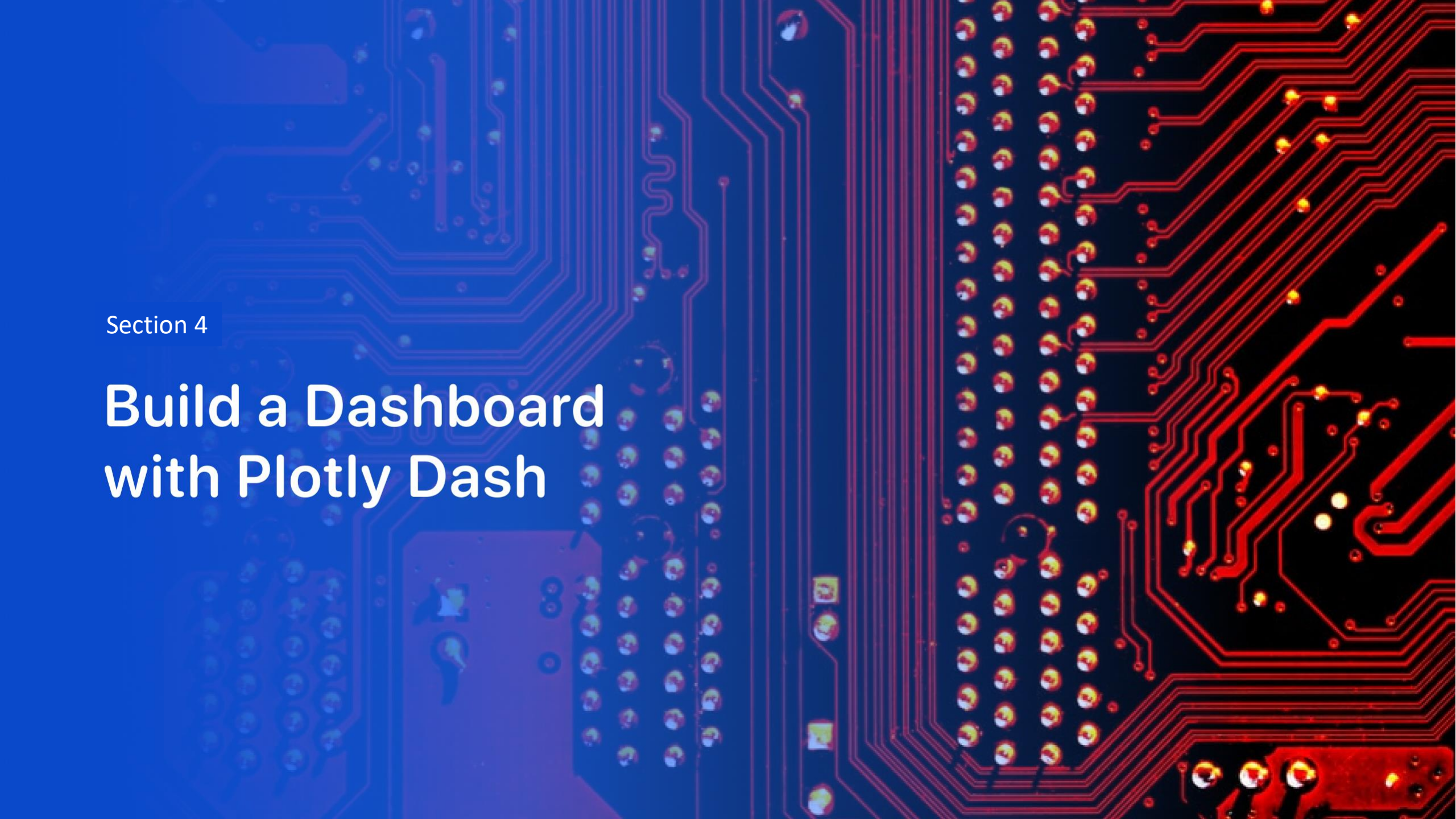




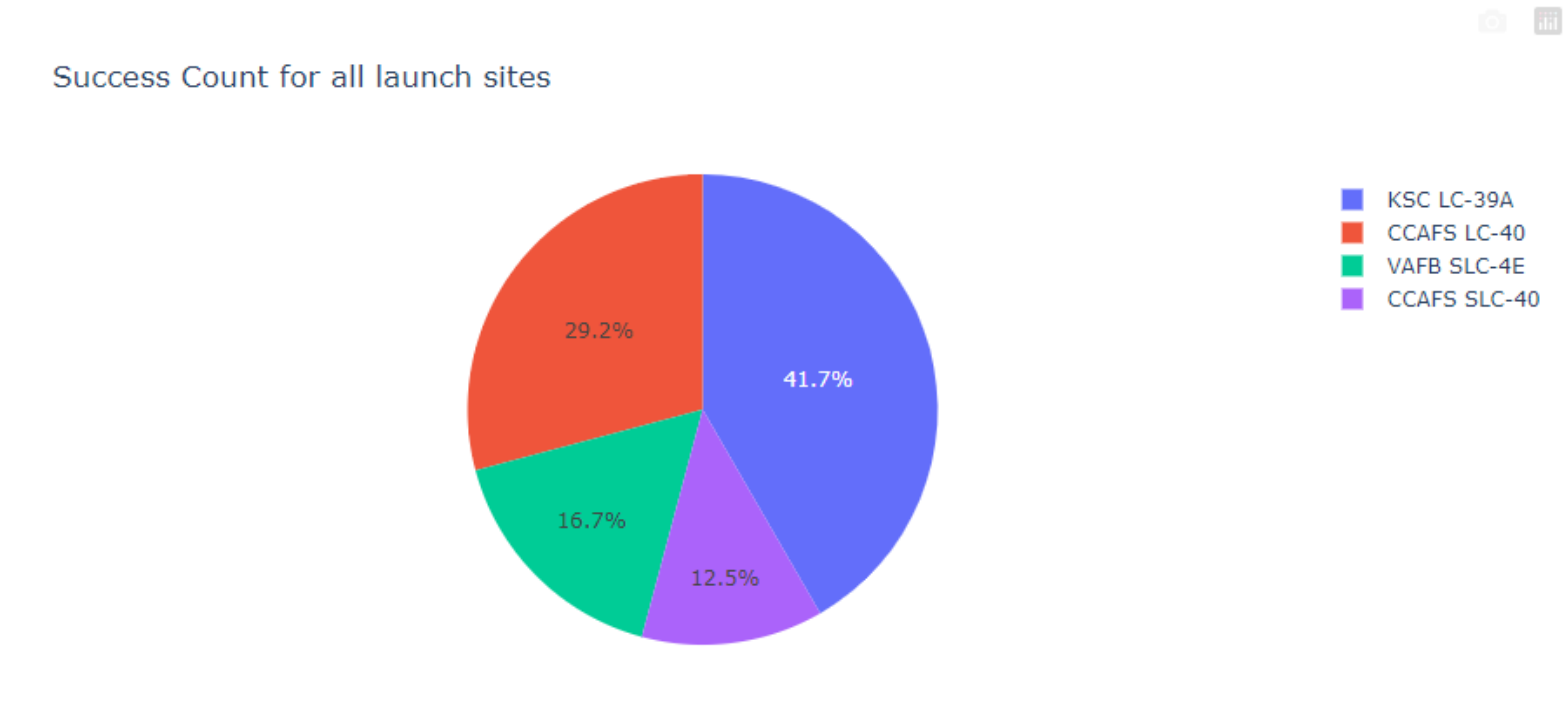
Distance from coastline, highway, railway and city

Launch Site	Distance to Railway	Distance to Highway	Distance to Coastline	Distance to City
CCAFS SLC-40	1.29 km	0.59 km	0.86 km	52.77 km
KSC LC-39A	0.55 km	5.28 km	4.01 km	53.44 km
VAFB SLC-4E	1.25 km	14.00 km	1.37 km	38.79 km

- All launch sites are in very close proximity to railway <1.5km.
- All launch sites are in relatively close proximity to highway <15km.
- All launch sites are in close proximity to coastlines <5km.
- All launch sites are far from cities >35km.

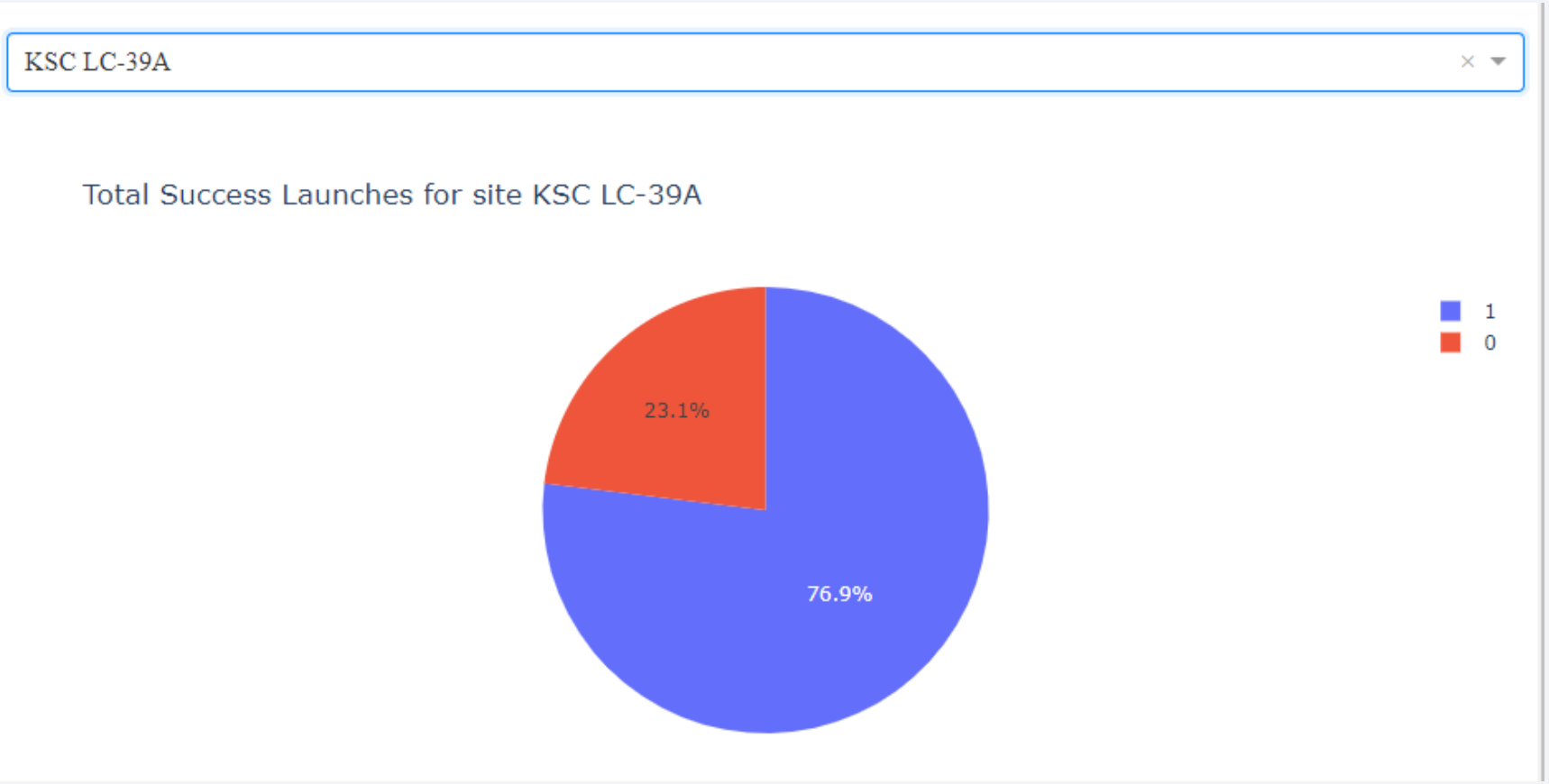

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches for All Launch Sites



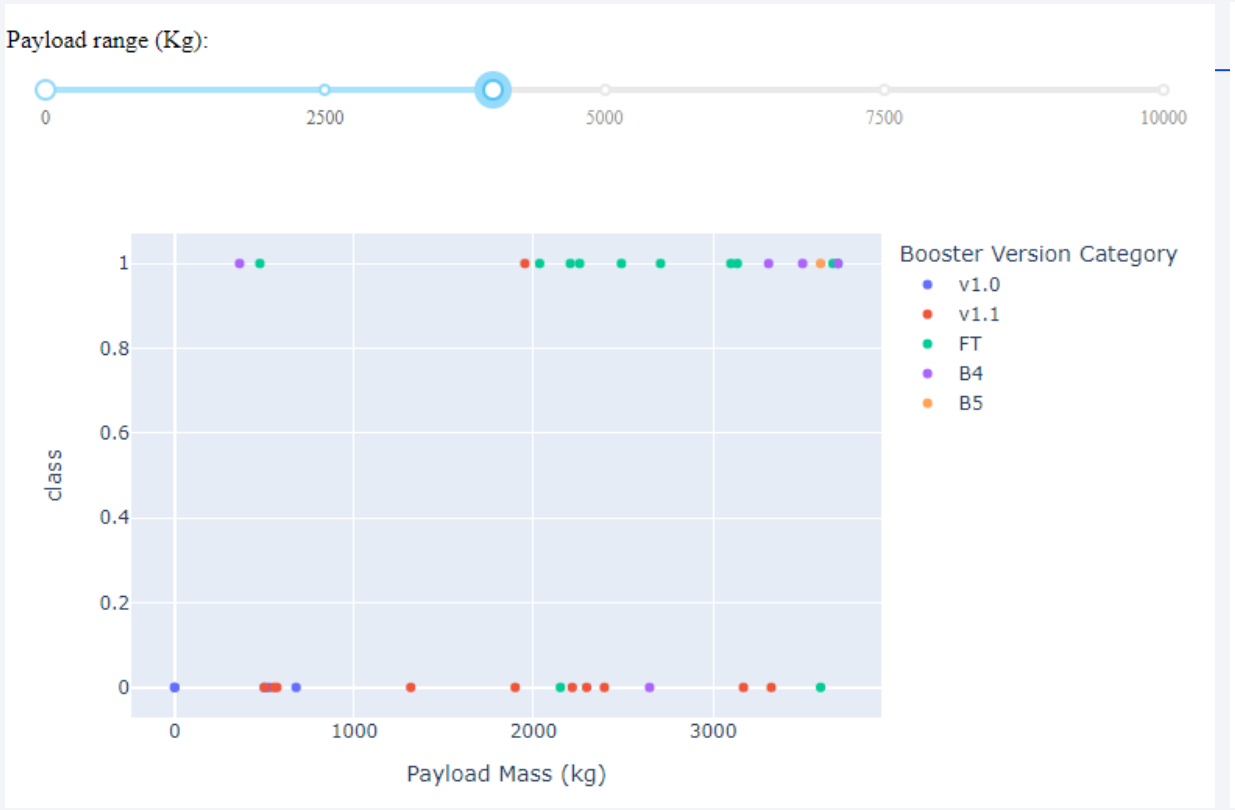
- KSC LC-39A has the most successful launches

Success rate at site KSC LC-39A

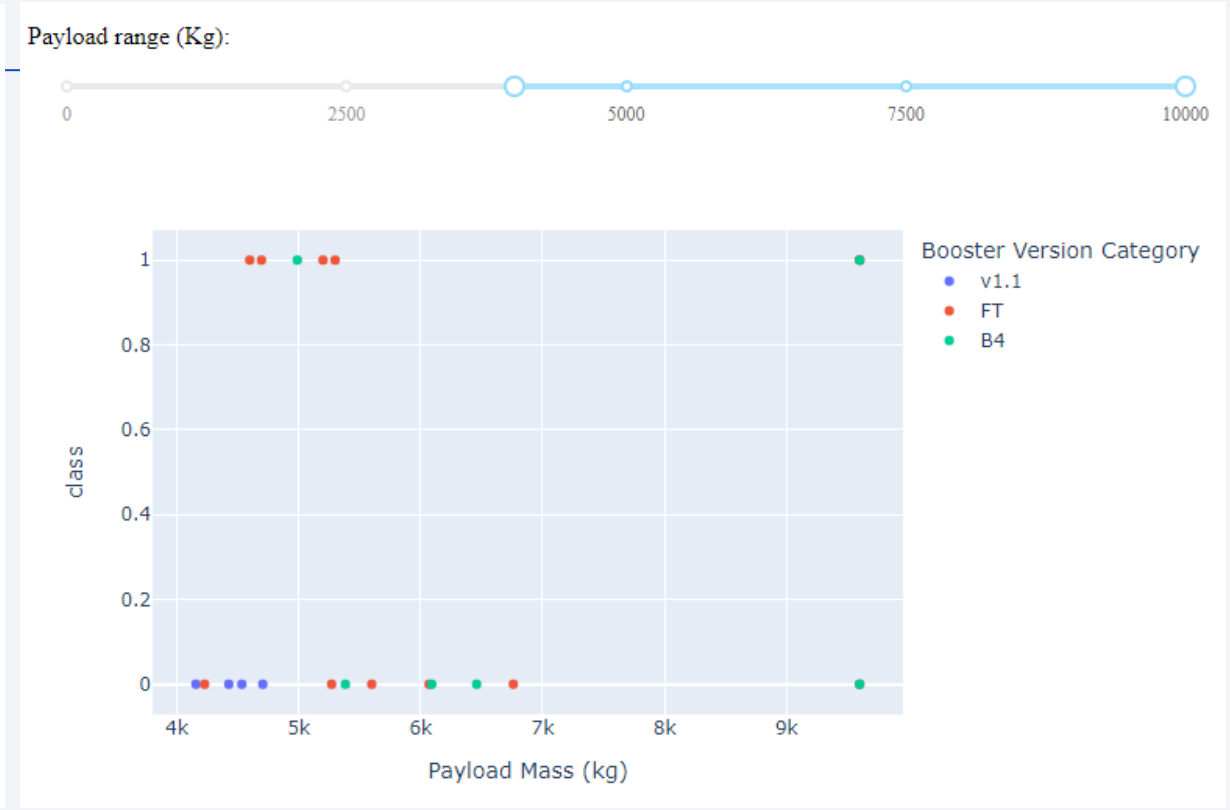


- KSC LC-39 is the launch site with most success, and has 76.9% success rate.

Comparing different payload range



Payload range between 0-4000kg



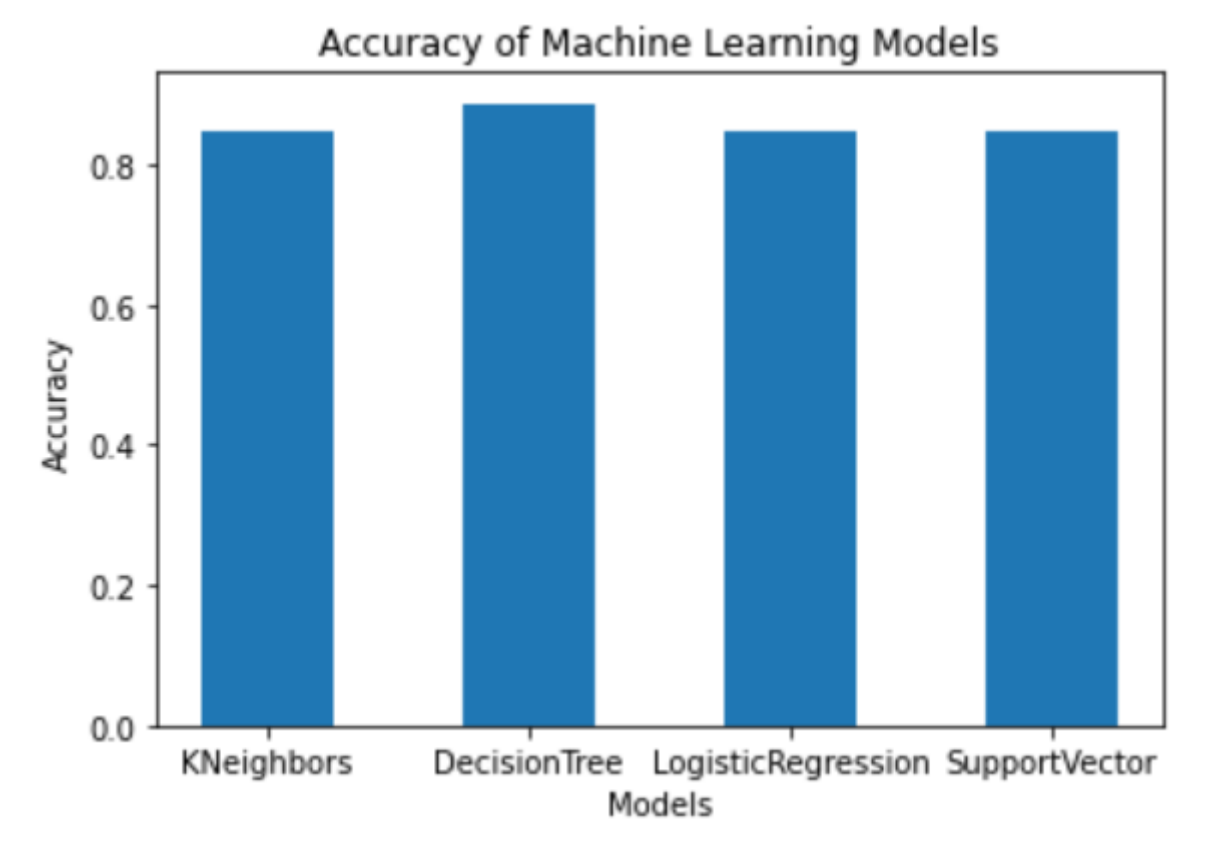
Payload range between 4000-10000kg

- There is higher success rate for lower payload range (between 0-4000kg)
- Some booster version do not have payload greater than 4000kg i.e. Booster v1.0 and B5

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree model has the highest accuracy score of 0.887
- The best parameters is :

{'criterion': 'entropy',

'max_depth': 4,

'max_features': 'sqrt',

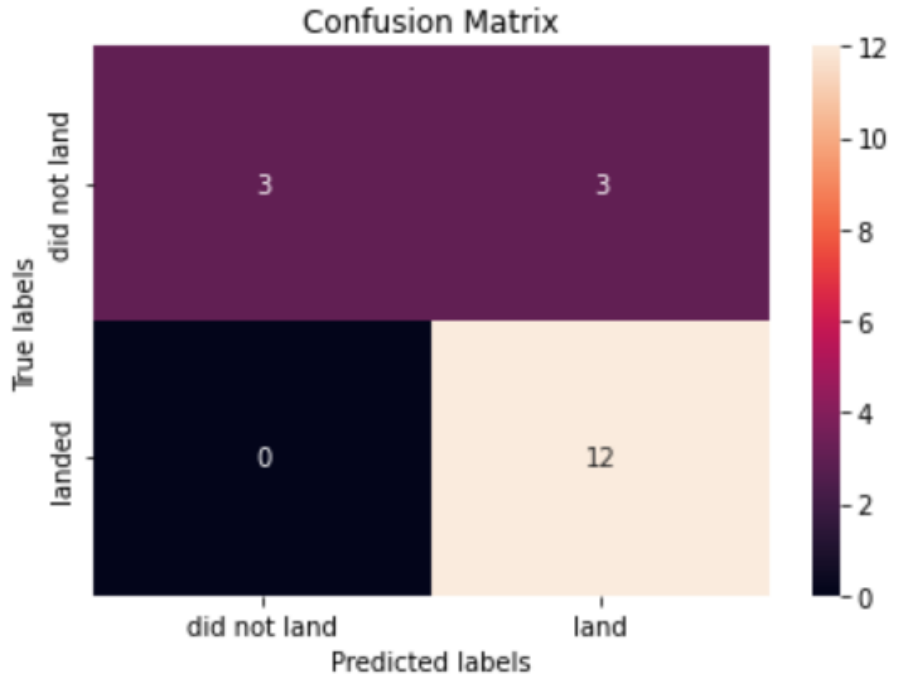
'min_samples_leaf': 4,

'min_samples_split': 10,

'splitter': 'best'}

Confusion Matrix

```
In [36]: yhat = tree_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```



- The confusion matrix shows that the decision tree model predicts well for successful landing. True positive rate = 12/12 =1
- However, the model is bad at predicting failed landing. It also has high false positive rate i.e 3/6 = 0.5 and low true negative rate 3/6 = 0.5.

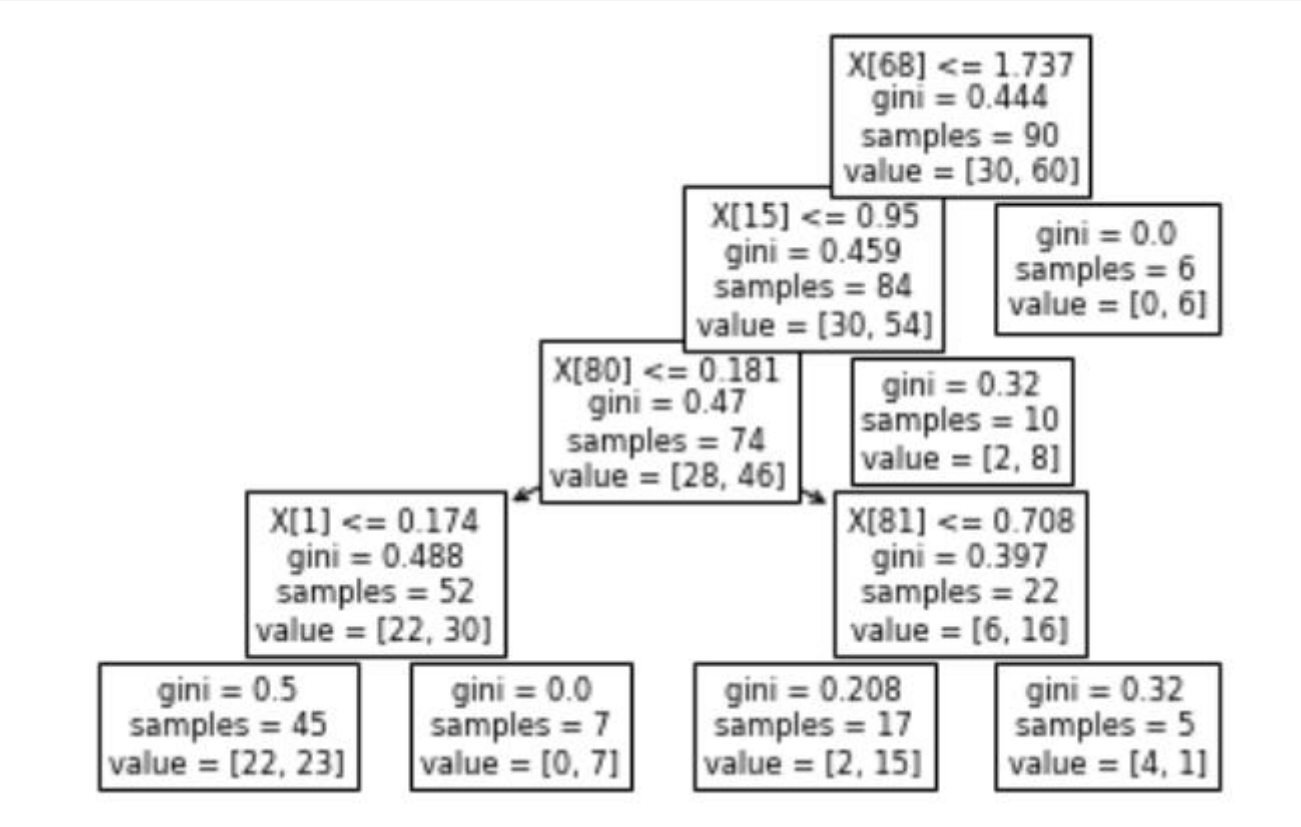
Conclusions

- Lower weighted payloads has higher landing success rate than heavier payloads.
- Launching success rate increase over the years – as more rockets are launch, technology improves and the success rate increases
- Launch sites are generally close to coastline, railway and highway and far from cities.
- KSC LC-39A is the launch site with the most successful launches, with success rate of 76.9%.
- Decision tree is the best performing model compared to KNN, Logistic Regression and SVM, with accuracy of 88.7%.

Appendix – Best Parameters for different models

Machine **Learning Model**	**Accuracy**	**Best parameters**
Logistic Regression	0.8464	{'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
Support Vector Machine	0.8482	{'C': 1.0, 'gamma': 0.031622776601683793, 'kernel': 'sigmoid'}
KNN	0.8482	{'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
Decision Tree	0.8875	{'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'best'}

Appendix – Decision Tree model



Appendix - Dataset

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	Reus
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	

Thank you!