IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eni Hal

10/21/21

Outline

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

Executive Summary

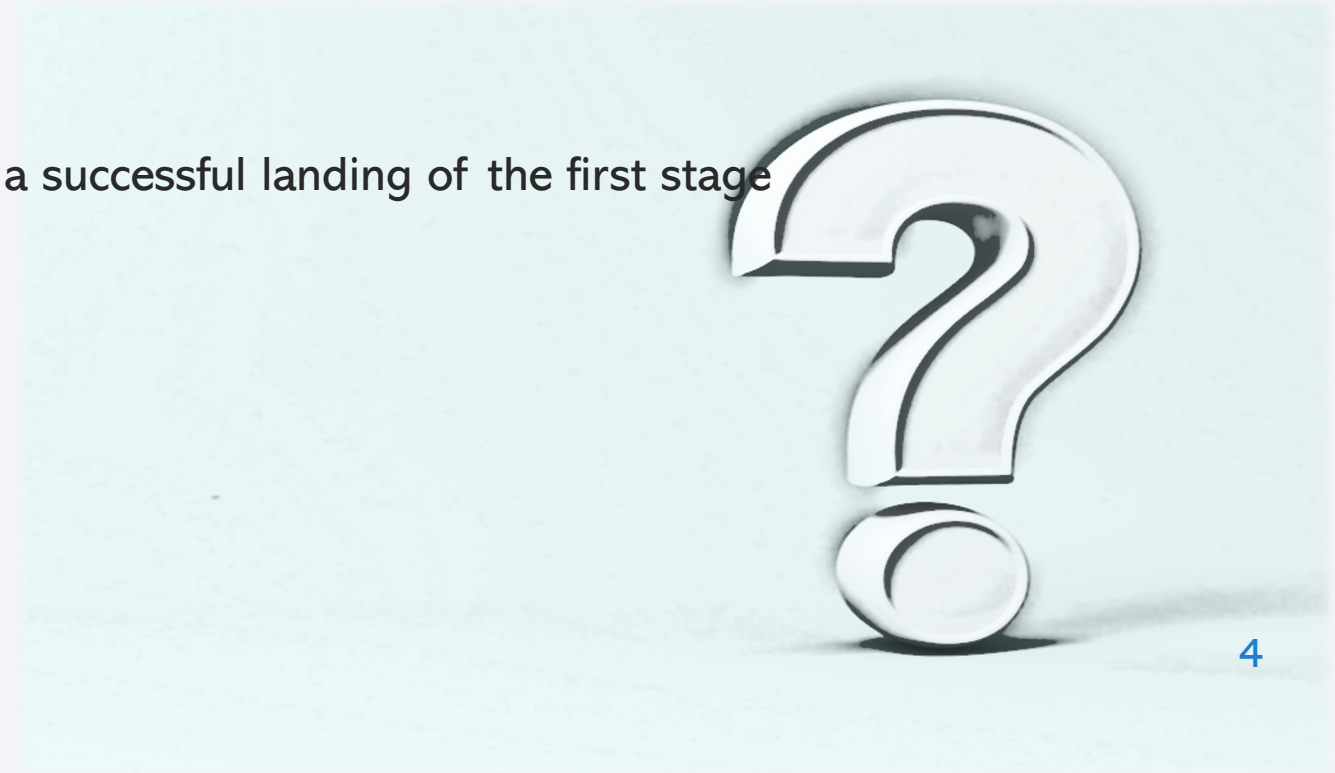
- Summary of methodologies
  - Data collection – utilizing APIs and web scrapping
  - Data wrangling – feature engineering and data cleanup
  - EDA with visualization and SQL
  - Interactive maps of launch sites
  - Dashboard reports with interactive graphics
  - Classification algorithms and prediction
- Summary of all results
  - Conducted extensive EDA with visualizations
  - Decision tree model ended up providing the best accuracy results, during our predictive modeling phase.

Introduction

- Project background and context

  SpaceX has the ability to reuse the first stage part of the rocket, which dramatically reduces the cost of space launches. Competitors' space launches cost around $165M, while SpaceX launches are roughly around $62M. Successful return and landing of the first stage part of the rocket is crucial in cost reduction and we want to evaluate what variables are significant in predicting a successful return landing of the first stage.

- Investigation topics
  - Which variables are significant in predicting a successful landing of the first stage
  - What role does weight, or location play?
  - Which model is the best predictor?
  - What is the accuracy of our chosen model?



Section 1

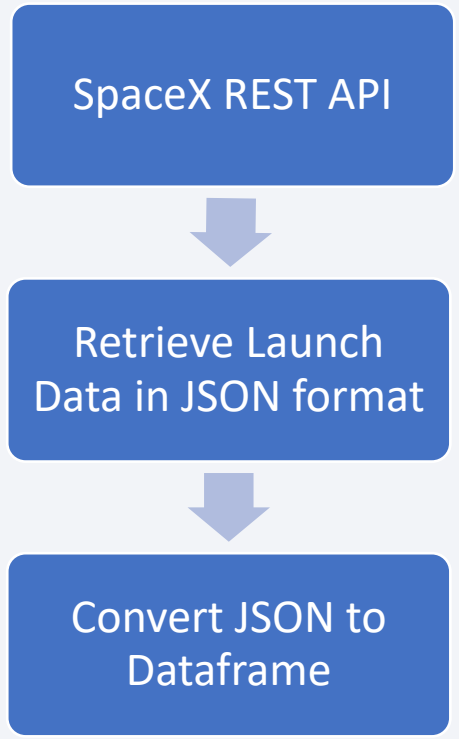
Methodology

Methodology – Executive Summary

- Data collection methodology:
  - Data was obtained via API, specifically SpaceX REST API.
  - Web scrapping
- Perform data wrangling
  - Unnecessary columns were dropped
  - Null values for Payload field were replaced with mean values
  - One-hot encoding of categorical columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, and performance evaluation

Data Collection

- The SpaceX data was obtained via an API, specifically the SpaceX REST API.
- This API provided data regarding launches, the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Data was obtained in JSON format and then converted to a dataframe in Python
- We also had the ability to obtain data via web scrapping, from Wikipedia



Data Collection – SpaceX API

1)

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

```
response = requests.get(spacex_url)
```

2)

```
# Use json_normalize meethod to convert the json result into a dataframe
response.json()
data = pd.json_normalize(response.json())
```

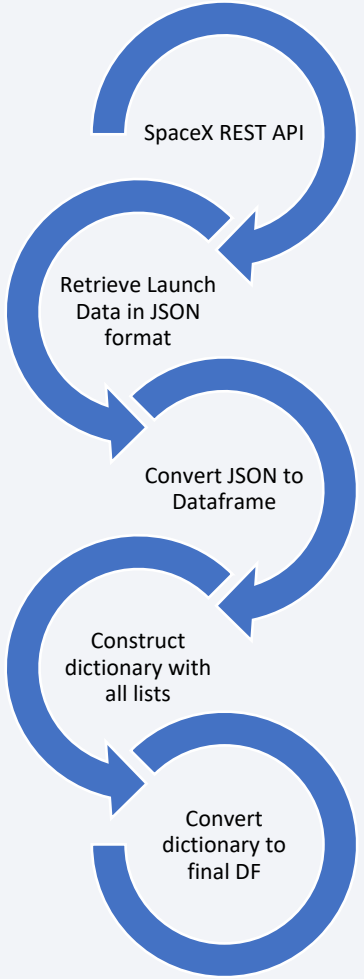
3)

Finally lets construct our dataset using the data we have obtained. We we combine the columns into a dictionary.

```
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}
```

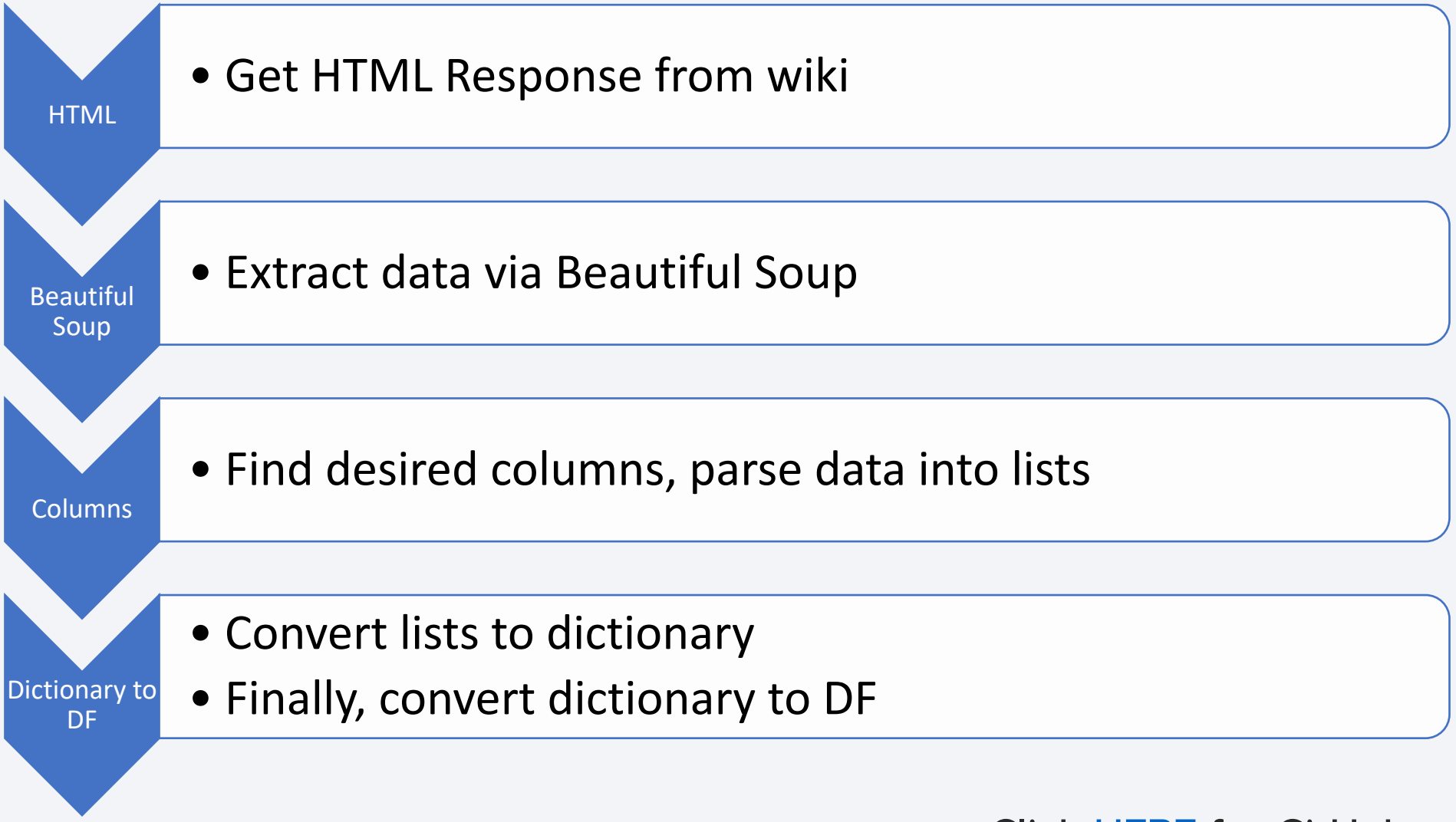
4)

```
# Create a data from launch_dict
df = pd.DataFrame.from_dict(launch_dict)
```



Click HERE for GitHub notebook

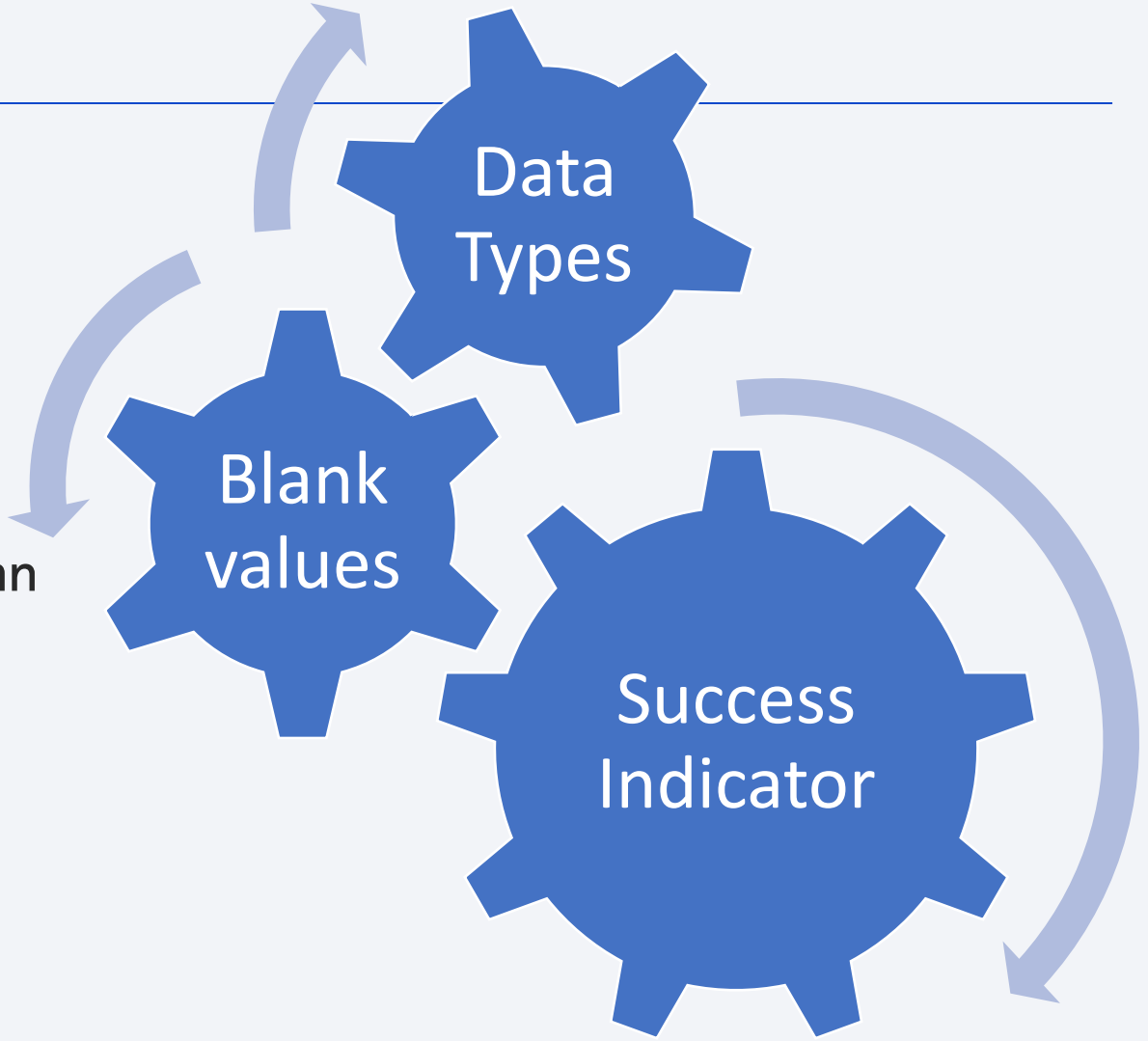
Data Collection – Scraping Process Overview



Click HERE for GitHub notebook

Data Wrangling

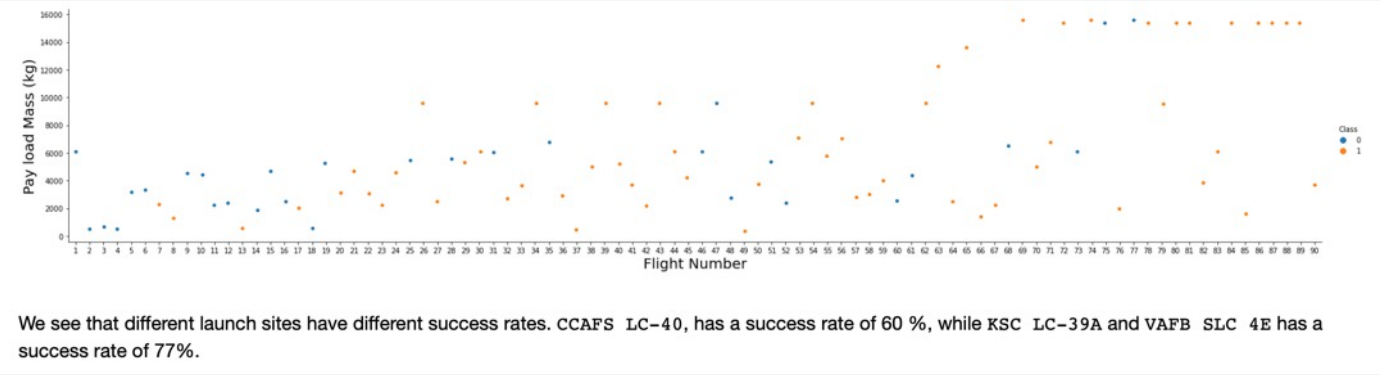
- Evaluated data types
- Evaluated blank values to ensure no issues
- Analyzed launch outcomes by orbit
- Created “Landing_class” indicator column to label successful outcome as 1 and unsuccessful outcomes as 0



Click HERE for GitHub notebook

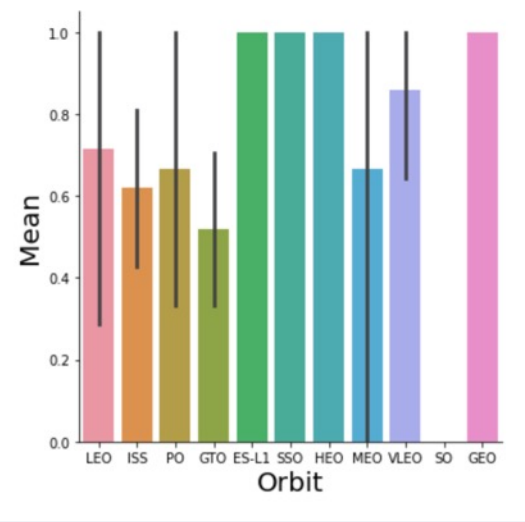
EDA with Data Visualization

- In this EDA we're looking at a number of variables, to determine their impact on a successful return of the first stage.
- Pay-load Mass weight



We see that different launch sites have different success rates. `CCAFS LC-40`, has a success rate of 60 %, while `KSC LC-39A` and `VAFB SLC 4E` has a success rate of 77%.

- Mean success rate by Orbit
  - A number of orbits have very high success rates, while others don't – so we can observe some significant differences



Click HERE for GitHub notebook

EDA with SQL

Below are some of the SQL queries that were performed to analyze the data:

- List of distinct launch sites
- Launch sites that begin with “KSC”
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- List the date where the successful landing outcome in drone ship was achieved
- List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- Please see notebook for more info

Click HERE for GitHub notebook

Build an Interactive Map with Folium

- Following map objects were created using Folium Map package:
  - Circles
  - Lines
  - Lables
- These objects were added to show locations of launch sites, and their distance from certain landmarks – such as coastline
- Objects were also added to mark successes and failures of each launch site
- Please see notebook for further details

Click HERE for GitHub notebook

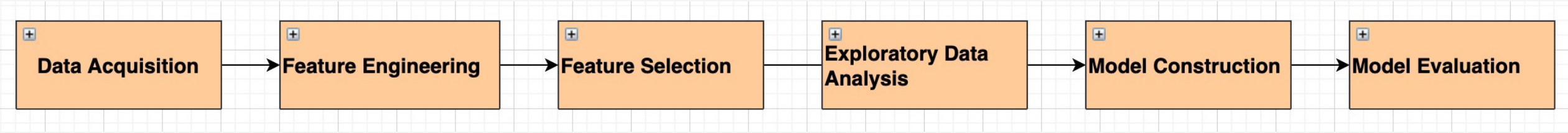
Build a Dashboard with Plotly Dash

Following was added to the dashboard:

- Drop-down list of launch sites
  - This drop down will serve as an input into the pie chart, so we display information for the desired selection
- Pie-chart of success rate of selected launch sites
  - This shows the success rates for one ore all launch sites, and it's interactive
- Scroll bar
  - Allows for scrolling and better viewing
- Scatterplot
  - To see the non-linear pattern of the pay load mass and the various booster versions
  - Allows us to easily see min/max ranges of the data
  - Frequencies of launces by booster version and their success rates

Click HERE for GitHub notebook

Predictive Analysis (Classification)

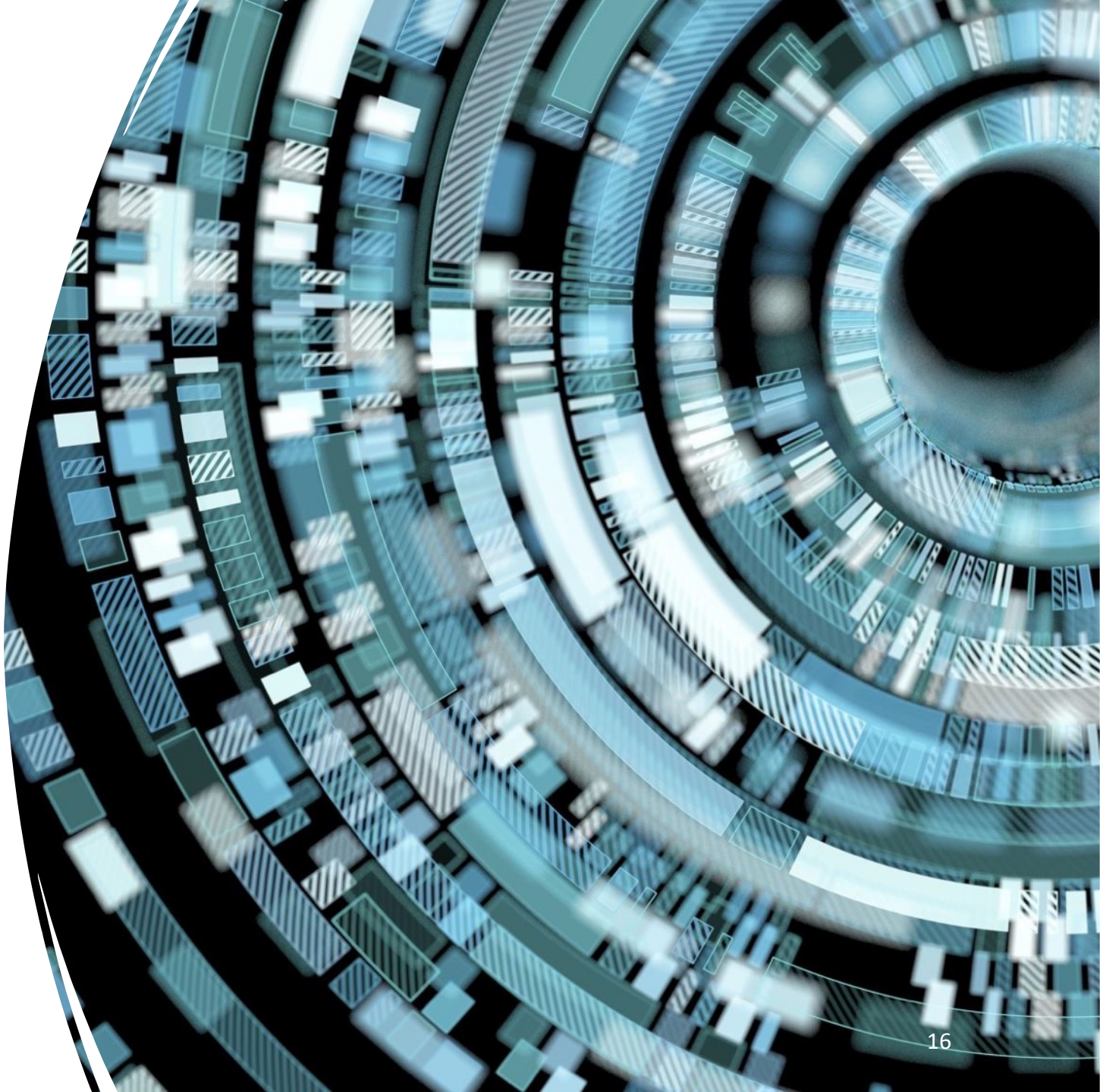


- Laid groundwork with data acquisition, feature engineering, EDA
- Built and evaluated following categorization models:
  - Logistic regression
  - KNN
  - Decision tree
- Modeling Approach and Highlights:
  - Split data into Train/Test
  - Evaluated using accuracy measure and by looking at confusion matrix
  - Optimized accuracy by optimizing parameters using grid search approach

Click HERE for GitHub notebook

Results

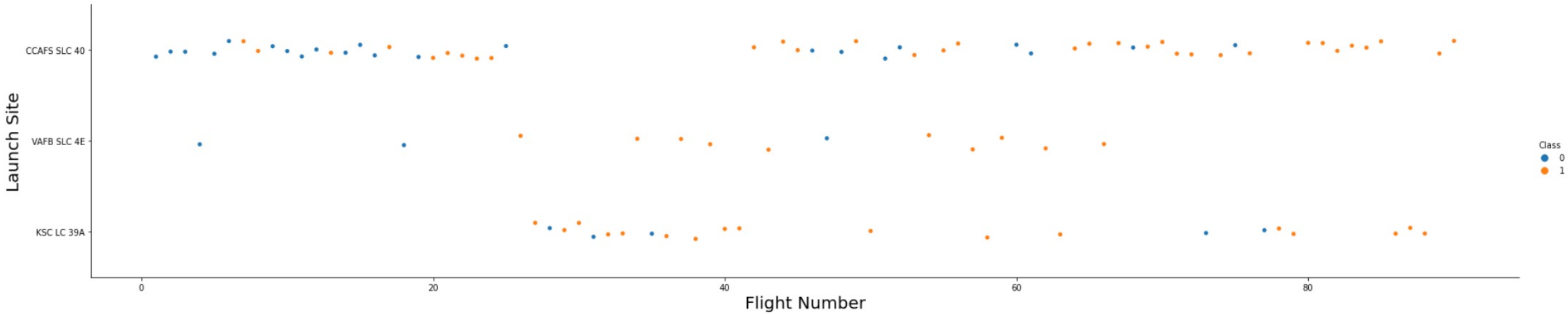
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

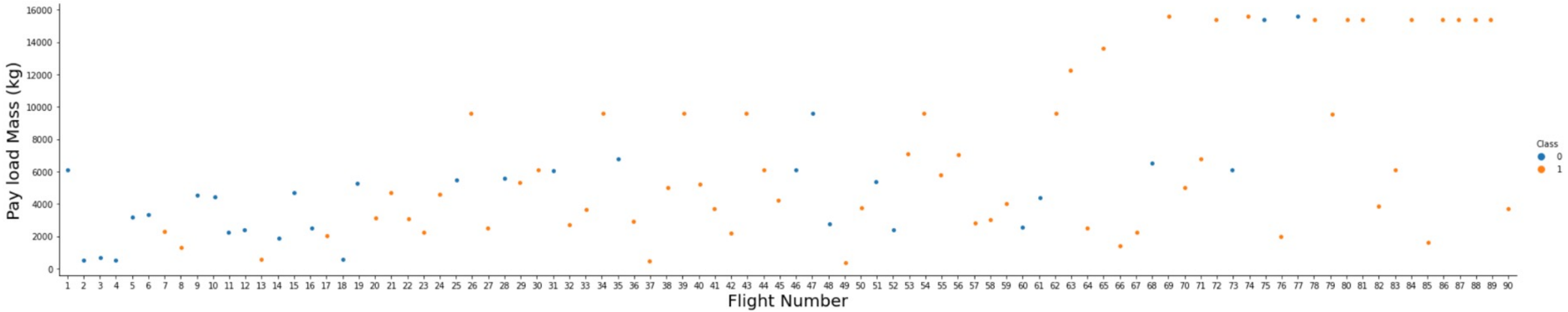
Insights drawn from EDA

Flight Number vs. Launch Site



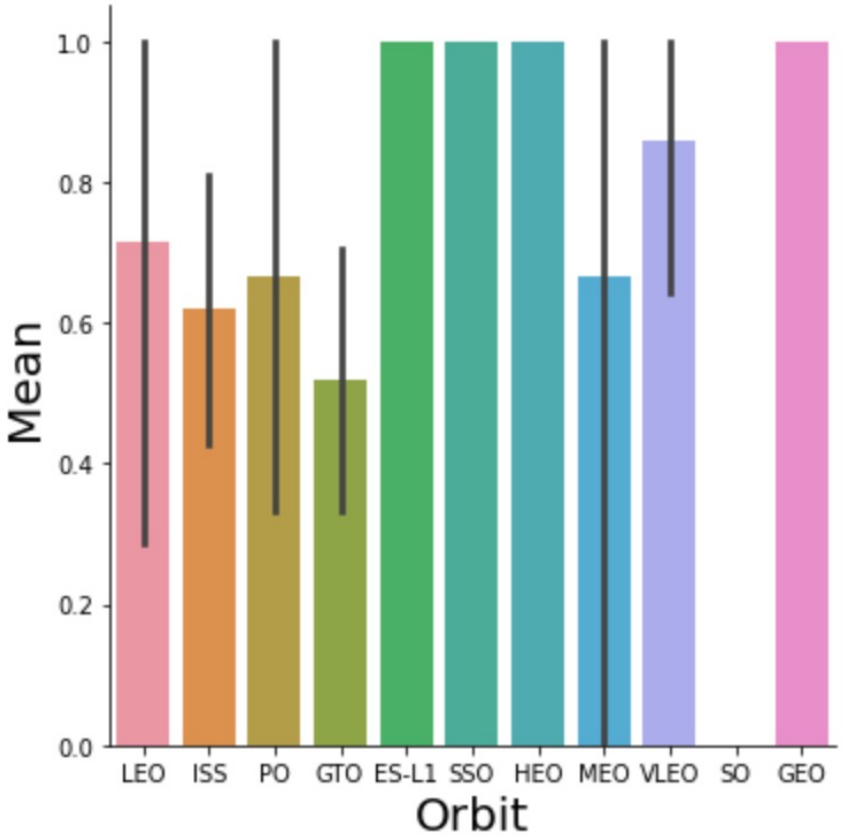
Launch Site "CCAFS LC-40" has the vast majority for flights, and it appears that successes are increasing as flight numbers increase.

Payload vs. Launch Site



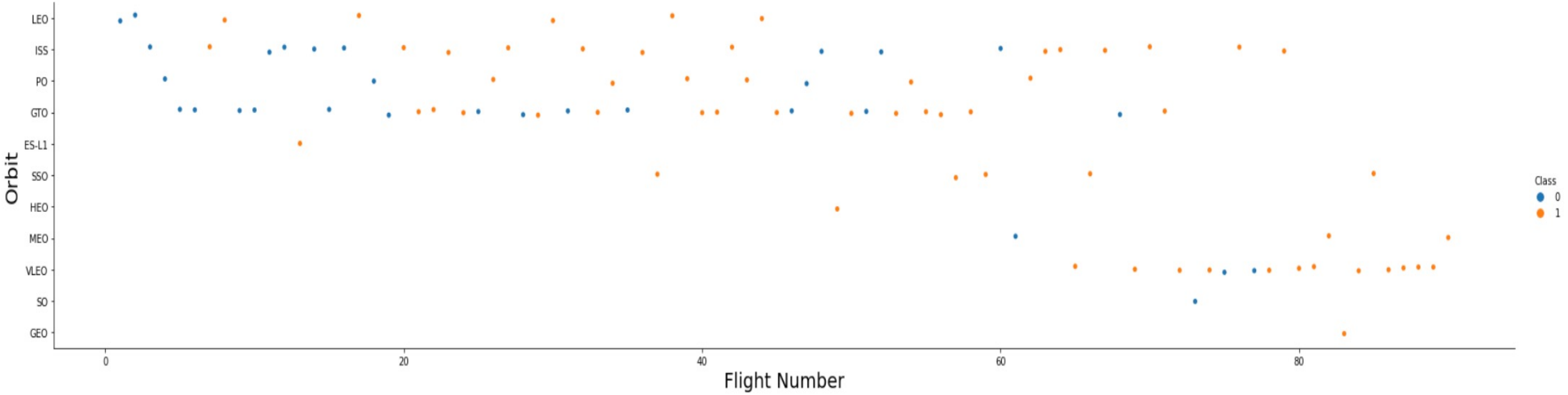
We see that different launch sites have different success rates. `CCAFS LC-40`, has a success rate of 60 %, while `KSC LC-39A` and `VAFB SLC 4E` has a success rate of 77%.

Success Rate vs. Orbit Type



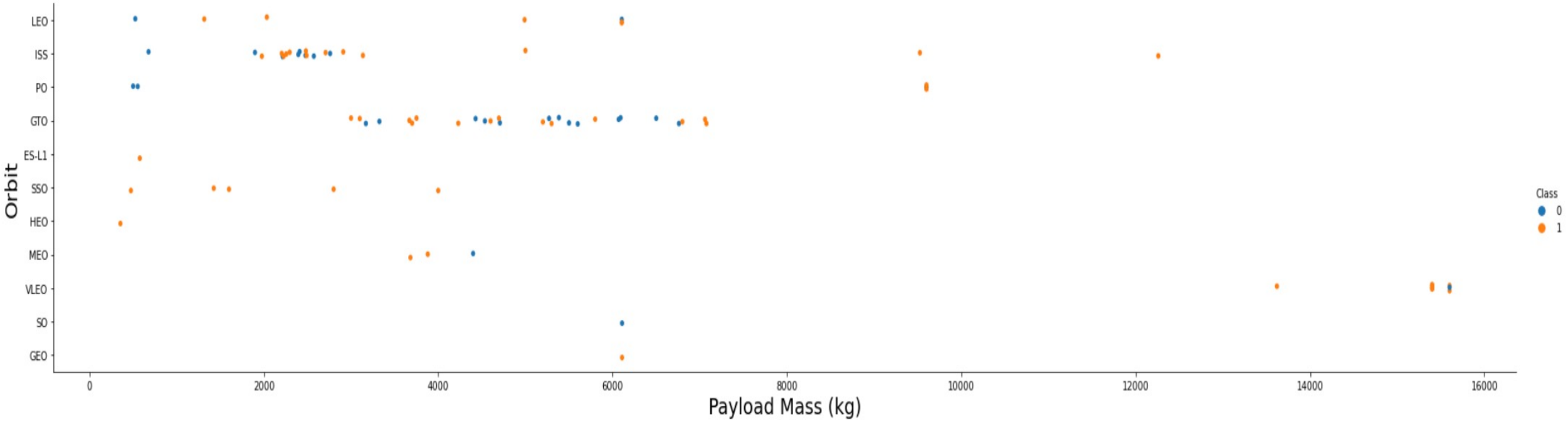
4 Orbits have a 100% success rate (ES-L1, SSO, HEO, and GEO). Another Orbit with high success rate is VLEO, at around 85%. Orbit SO has 0% success rate.

Flight Number vs. Orbit Type



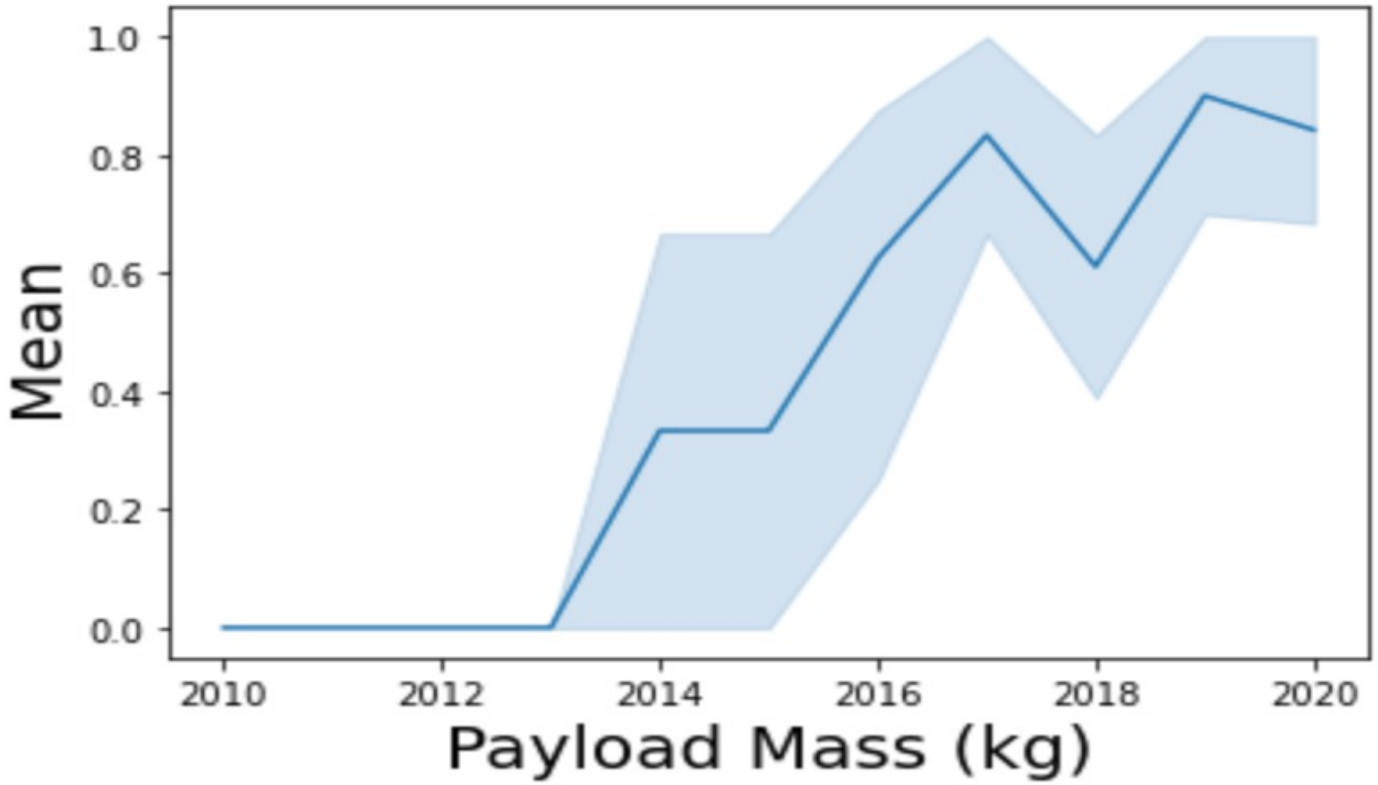
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



We observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%%sql
select distinct launch_site from SPACEXDATA
```

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

- We used SQL distinct clause to obtain a list of all distinct launch sites.

Launch Site Names Begin with ‘KSC'

Display 5 records where launch sites begin with the string 'KSC'

```
%%sql
select *
from SPACEXDATA
where launch_site like 'KSC%'
limit 5
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2/19/17	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
3/16/17	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
3/30/17	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
1/5/17	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
5/15/17	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

- We used the WHERE clause combined with the LIKE clause to filter launch sites beginning with “KSC”.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
select customer, sum(PAYLOAD_MASS__KG_) as Total_Payload
from SPACEXDATA
where customer like 'NASA (CRS)'
group by customer
```

customer	total_payload
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%%sql
select booster_version, avg(PAYLOAD_MASS__KG_) as Avg_Payload
from SPACEXDATA
where booster_version like 'F9 v1.1'
group by booster_version
```

booster_version	avg_payload
F9 v1.1	2928

- We filtered the data by using the WHERE and LIKE clauses. Grouping the data helped aggregate the average payload mass.

First Successful Ground Landing Date

List the date where the succesful landing outcome in drone ship was acheived.

Hint:Use min function

```
%%sql
SELECT min(DATE)
FROM SPACEXDATA
WHERE mission_outcome LIKE 'Success' and landing__outcome like '%(drone ship)'
```

1
1/14/17

- We used min function to get the earliest date, along with filtering criteria.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATA
WHERE mission_outcome LIKE 'Success' and landing__outcome like '%(ground pad)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

booster_version
F9 FT B1032.1
F9 B4 B1040.1

Total Number of Successful and Failure Mission Outcomes

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

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

- We used count(*) to get the count, along with groupby to get the information.

Boosters Carried Maximum Payload

```
%%sql
select booster_version, max(payload_mass__kg_) as Max_Load
from SPACEXDATA
group by booster_version
order by Max_Load desc
```

booster_version	max_load
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2017 Launch Records

```
%%sql
SELECT FORMAT(DATE, 'MMMM') AS Month, mission_outcome, booster_version, launch_site, DATE
FROM SPACEXDATA
WHERE mission_outcome LIKE '%Success%' AND Date like '%/17%'
```

mission_outcome	booster_version	launch_site	DATE
Success	F9 FT B1029.1	VAFB SLC-4E	1/14/17
Success	F9 FT B1031.1	KSC LC-39A	2/19/17
Success	F9 FT B1030	KSC LC-39A	3/16/17
Success	F9 FT B1021.2	KSC LC-39A	3/30/17
Success	F9 FT B1032.1	KSC LC-39A	1/5/17
Success	F9 FT B1034	KSC LC-39A	5/15/17
Success	F9 FT B1035.1	KSC LC-39A	3/6/17
Success	F9 FT B1029.2	KSC LC-39A	6/23/17
Success	F9 FT B1036.1	VAFB SLC-4E	6/25/17
Success	F9 FT B1037	KSC LC-39A	5/7/17
Success	F9 B4 B1039.1	KSC LC-39A	8/14/17
Success	F9 FT B1038.1	VAFB SLC-4E	8/24/17
Success	F9 B4 B1040.1	KSC LC-39A	7/9/17
Success	F9 B4 B1041.1	VAFB SLC-4E	9/10/17
Success	F9 FT B1031.2	KSC LC-39A	11/10/17
Success	F9 B4 B1042.1	KSC LC-39A	10/30/17
Success	F9 FT B1035.2	CCAFS SLC-40	12/15/17
Success	F9 FT B1036.2	VAFB SLC-4E	12/23/17

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

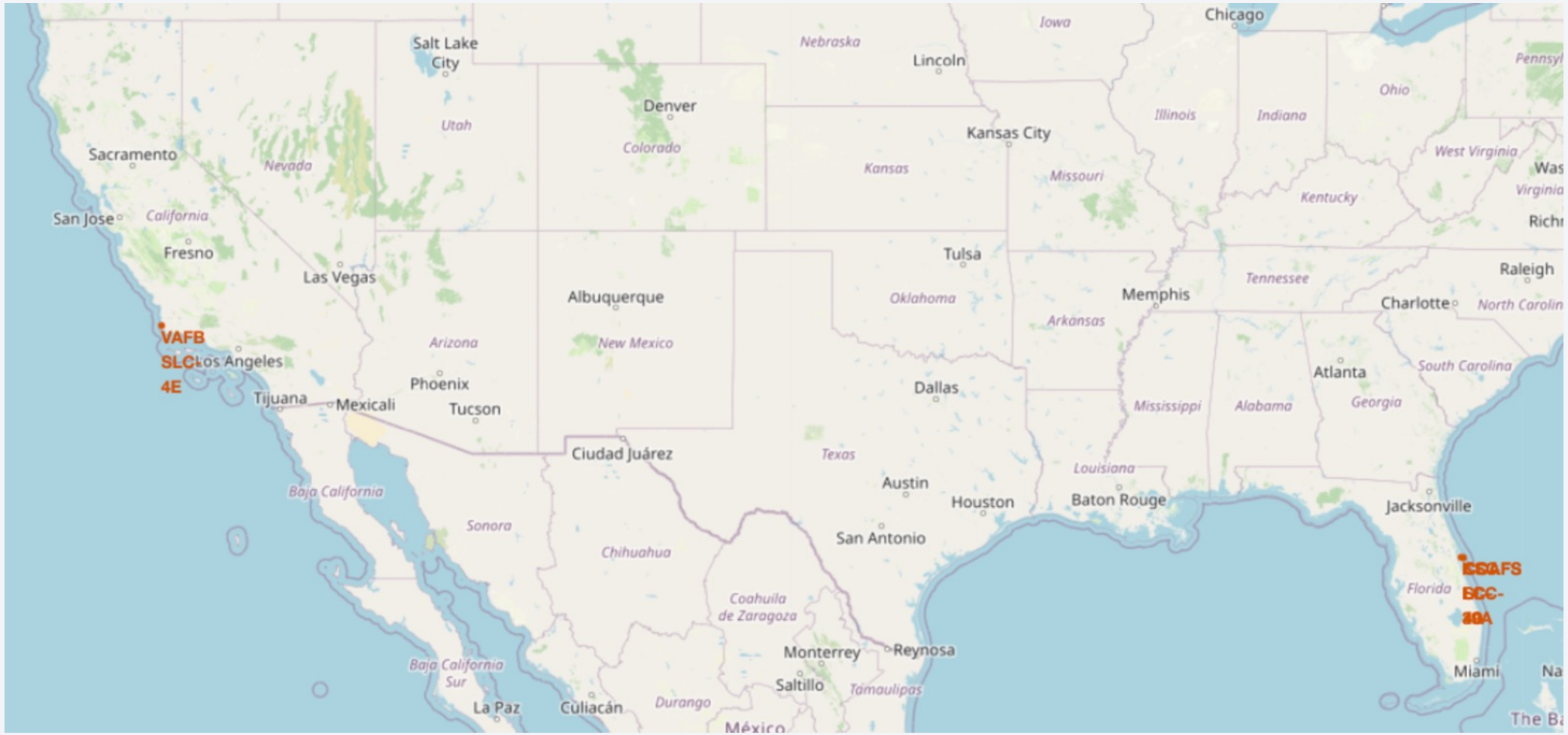
```
%%sql
SELECT landing__outcome, count(*) as count
FROM SPACEXDATA
WHERE landing__outcome LIKE '%Success%' AND (Date between '2010/06/04' and '2017/03/20')
group by landing__outcome
order by count desc
```

- We used count(*) to get the count, along with groupby to get the information by landing_outcome. Order by was utilized to order the results by descending order.

Section 4

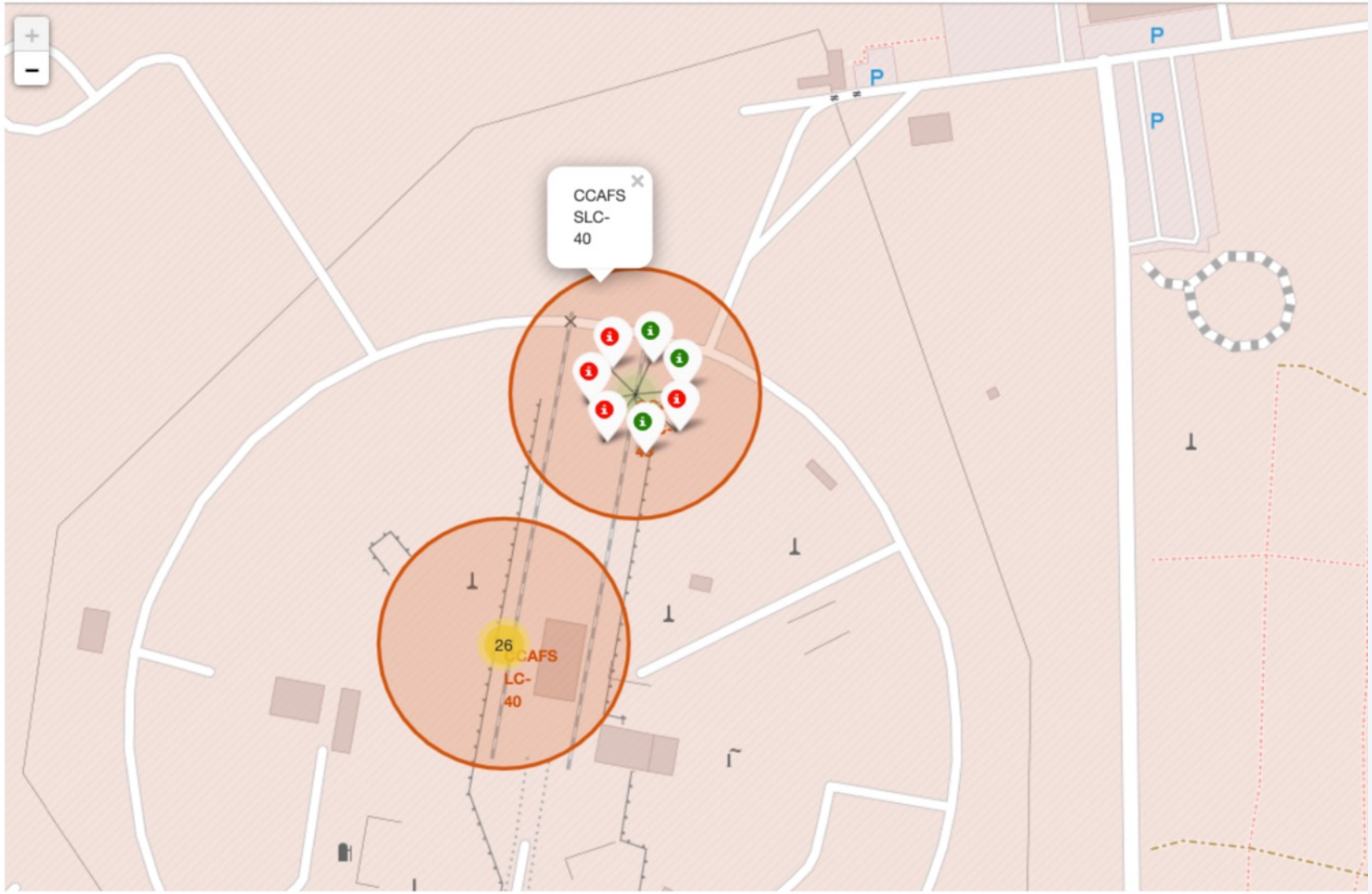
Launch Sites Proximities Analysis

SPACEX Launch Sites in US



- In this map we can see markers for launch sites in Florida and Southern California.

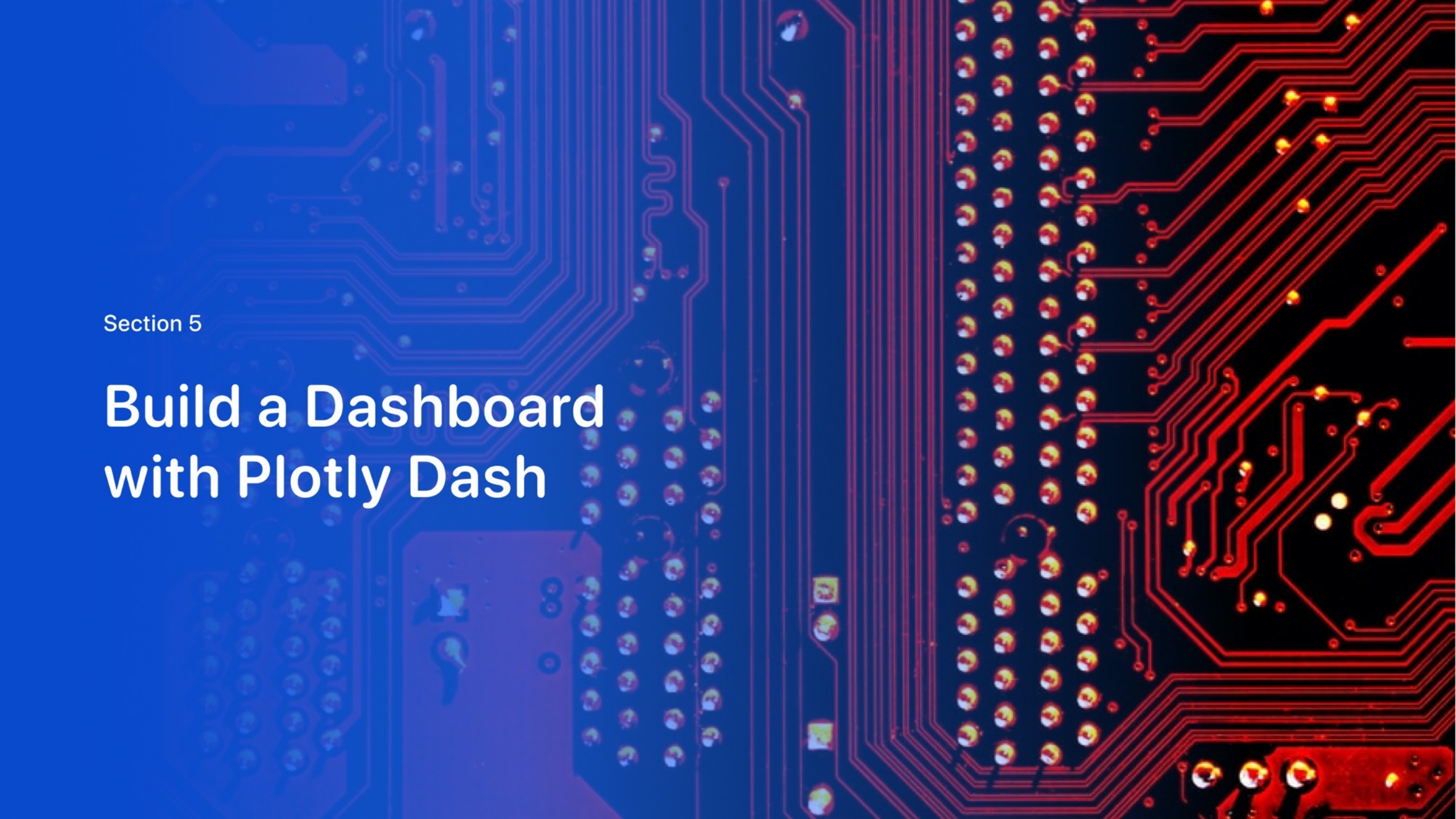
High-Success Rate Launch Site



From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

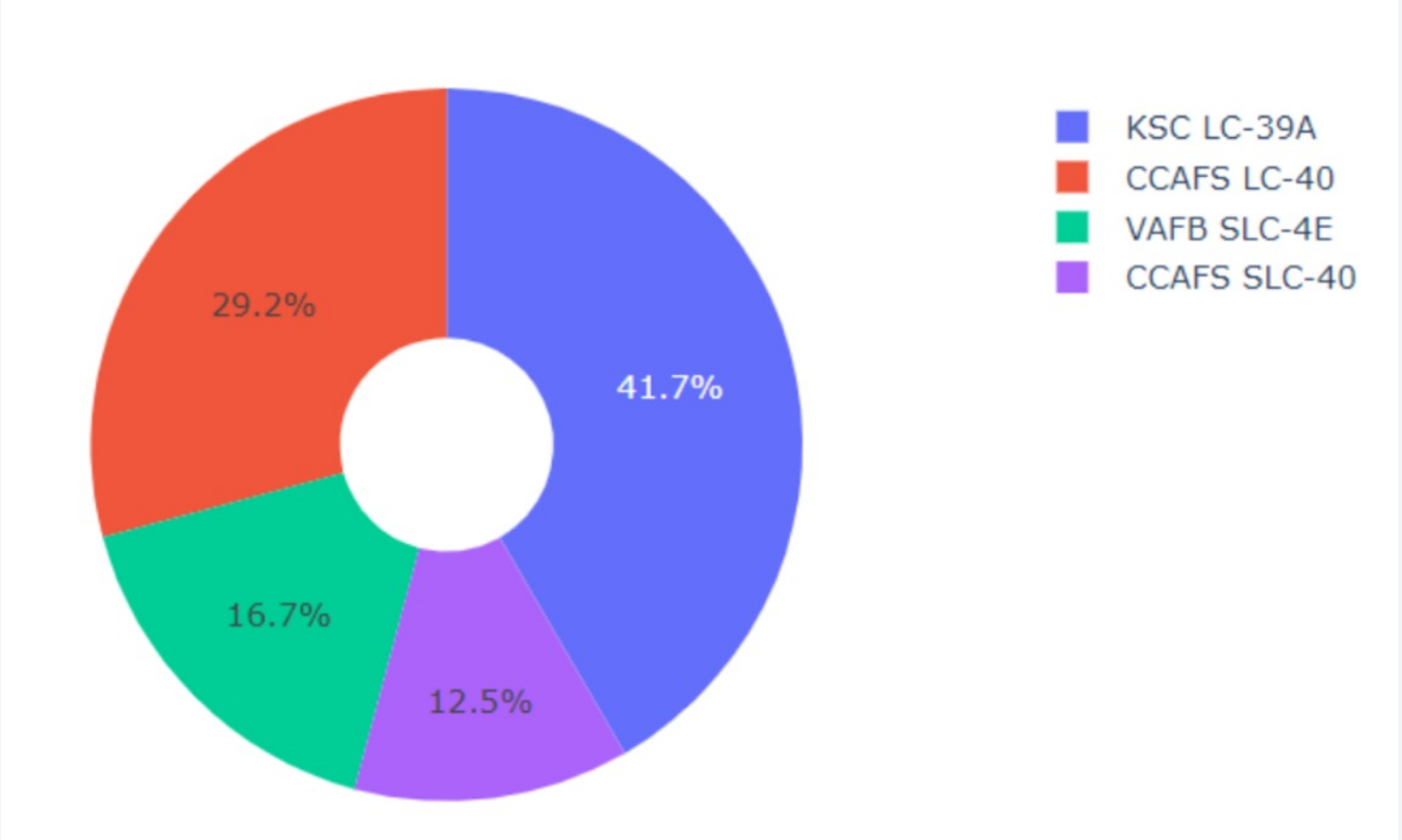
Section 5

Build a Dashboard with Plotly Dash

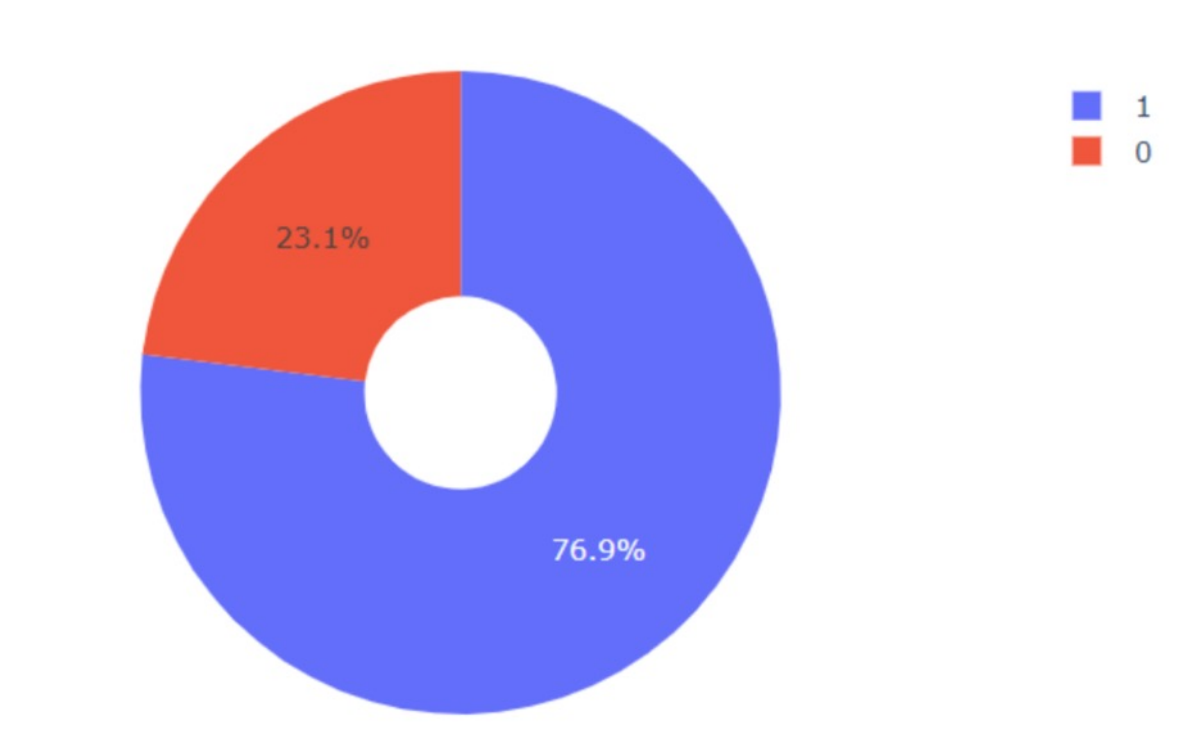


Success Rate by Site

- We can observe that launch site KSC LC-39A had the highest portion of sucesses.



<Dashboard Screenshot 2>



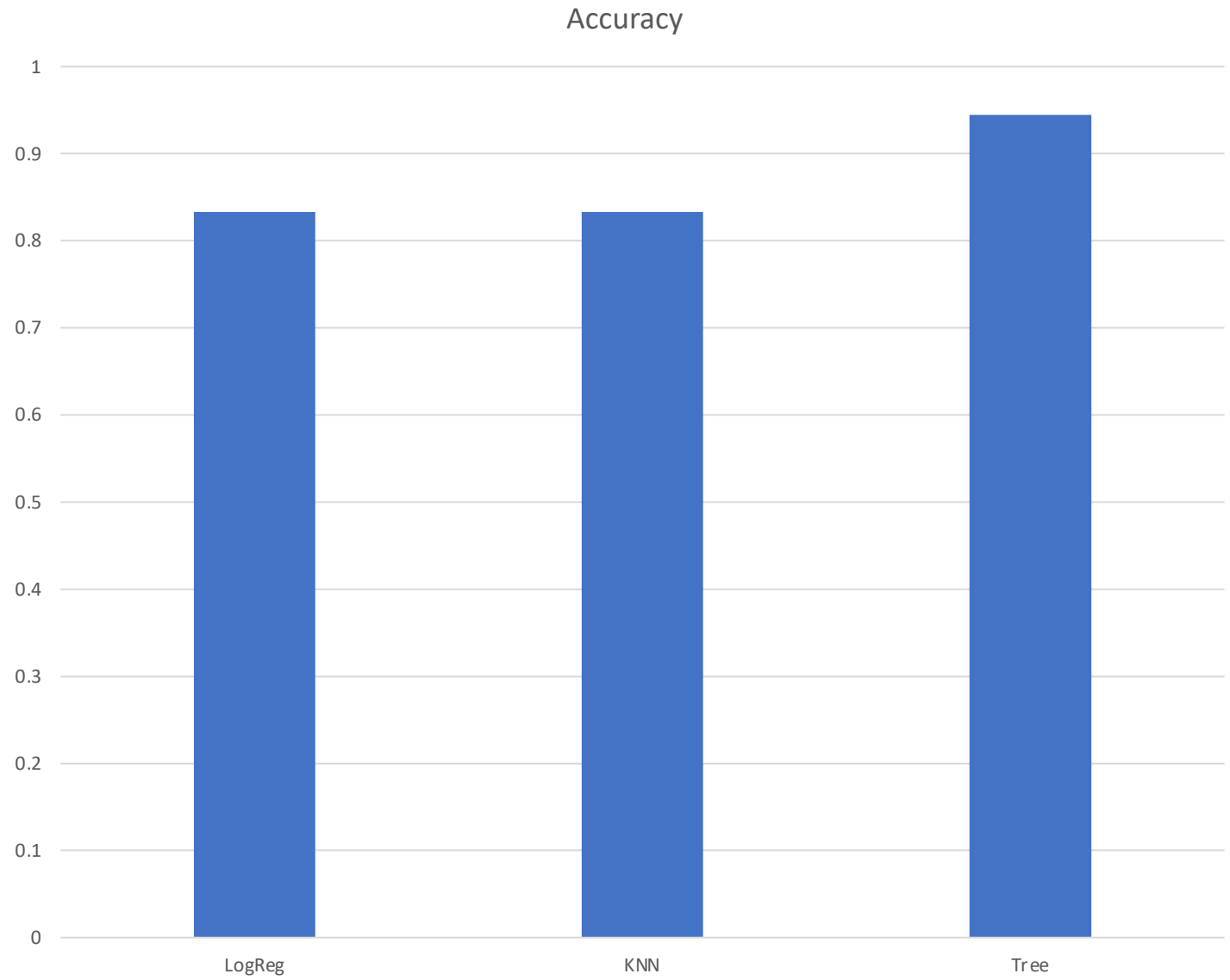
Taking a closer look at site KSC LC-39A, we can see that the overall success rate sits at 76.9% and failure at 23.1%.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

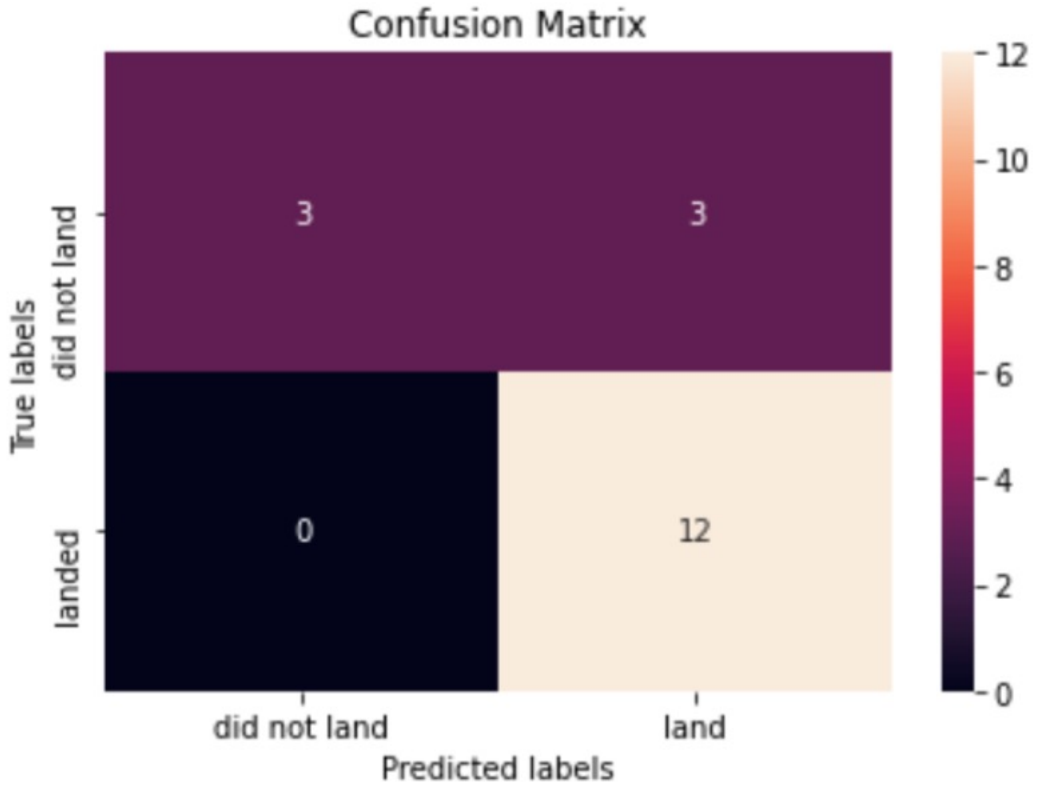
- All three models performed quite well, however the Decision Tree model did have the best accuracy with over 90% accuracy!



Confusion Matrix

- From the confusion matrix we can see that the model performs quite well in predicting true positives. It also predicts 3 true negatives, but struggles a bit with predicting 3 false positives.

```
yhat = svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- We concluded that the decision tree algorithm performed the best, with incredible 93% accuracy. However, we also observed some lack in precision, due to false positives.
- We’ve observed a significant increasing trend in successful launches starting from 2013
- Launch site KSC LC-39A has the highest success rate, with around 73% successful launches
- Payload mass turned out to be also an indicator, and heavier payload mass was associated with more successful launches.

Thank you!