IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Cao Minh Hieu>
<03/09/2022>

Outline

- Introduction
- Executive Summary
- Methodology
- Results
- Conclusion

Introduction

- 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
- This study will answer the following question:
  - Which launch sites do their rely on? What is their launch frequency?
  - Is there an observable learning curve? Which parameters can we play on to make the learning curve steeper?
- The data collected for this study cover a period ranging from June 2010 to May 2020 and 89 launches.

Executive Summary

The model including, on top of the payload mass and orbit, features of the rocket including its number of flights/re-use count, block version, landing pad coordinates, presence of gridfins and legs (all absent in the earliest version of Falcon 9), help yeld good predictions of the chance of success of a landing, with an accuracy of 83% .

Feature assessment methods (Unvariate Selection, Feature Importance) however show that rocket features (presence of gridfins and legs) and number of flights help most improve success rates, with other variables such as orbit and launch site being neglectable.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- First use SpaceX API
- Secondly use web scraping (Wikipedia)
- Summarize sources in dataframes

Data Collection – SpaceX API

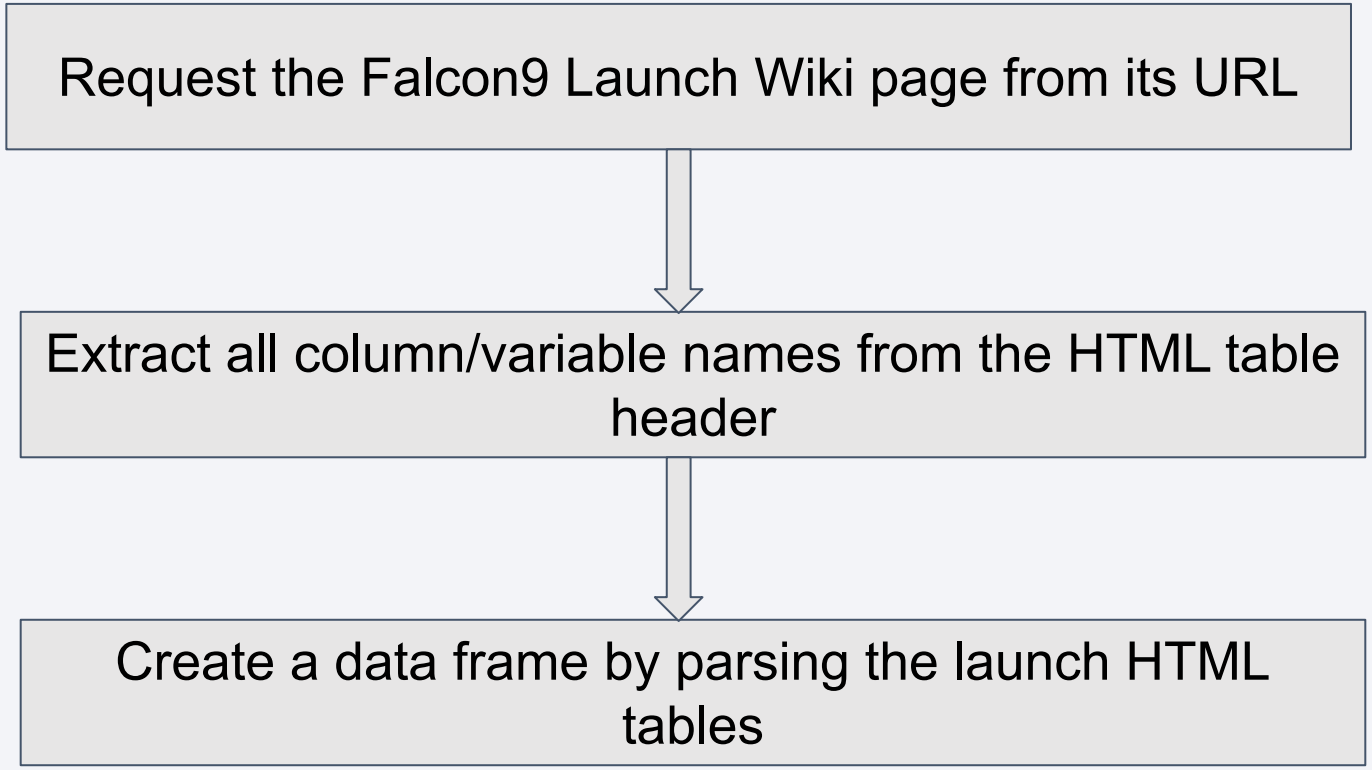
This notebook will show more information to collect data with SpaceX API

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block
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

90 rows × 17 columns

Data Collection - Scraping

This notebook will show more information to collect data with Web Scraping



Data Wrangling

- Three steps for processed data:
  - Filtering on Falcon 9
  - Dealing with missing values
  - Create landing outcome column with numerical values

- This notebook will show more information about data wrangling

EDA with Data Visualization

- All charts in this lab present the correlation between features of data
- This notebook will show more information about data visualization

EDA with SQL

- The SQL queries will show more insights of the data
- This notebook will show more information about that

Build an Interactive Map with Folium

- SpaceX launches rockets from 4 sites East coast : CAFS SLC-40, CCAFS SLC-40, KSC LC-39A West coast : VAFB SLC-4E
- All 4 sites share the same features : Proximity to coastline - Distance from cities, highways and railways
- This notebook will show more information about that

Build a Dashboard with Plotly Dash

Use dashboard for visualize the insights of the data

This notebook will show more information about that Dashboard with Plotly Dash

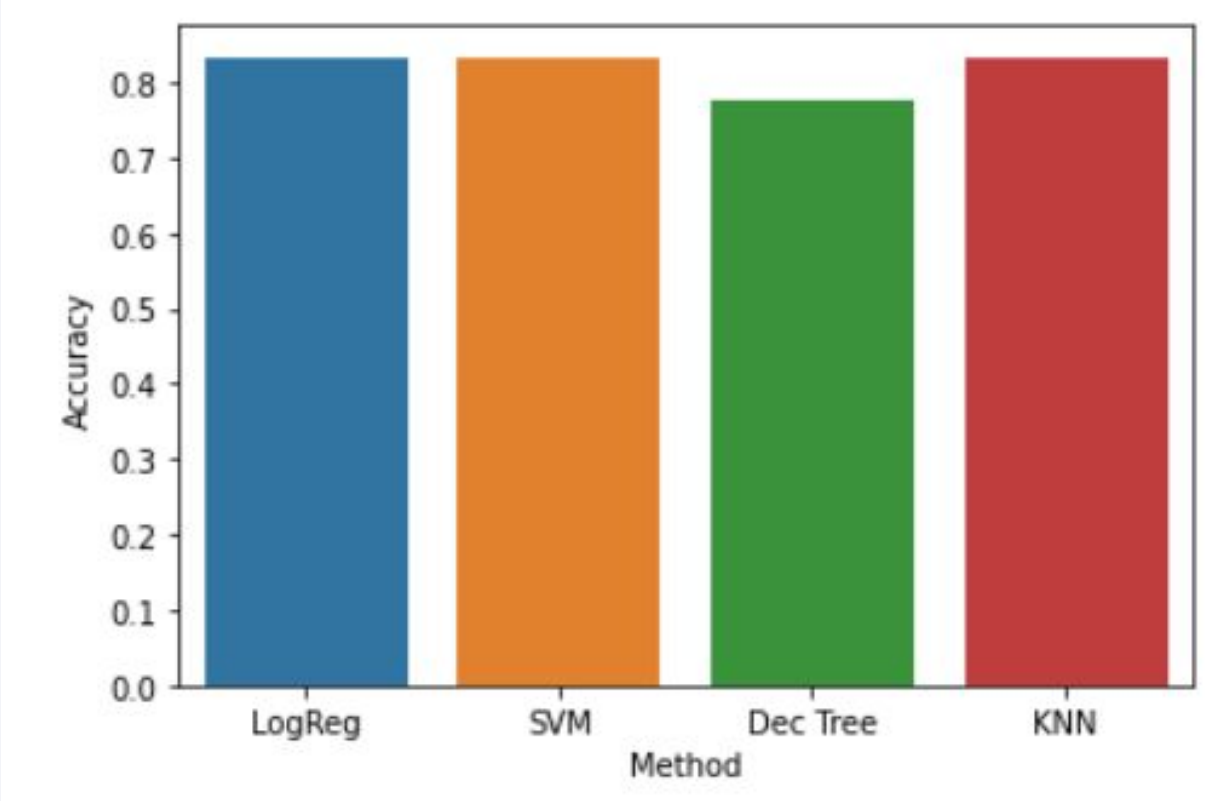
Predictive Analysis (Classification)

- Following for predictive analysis:
  - Create the numpy matrix of features values X, Y (Success/Failure)
  - Standardization of the data
  - Split data for train-test set
  - Use 4 models: Logistic regression, SVM, decision tree, KNN
  - Using Gridsearch to find the best parameters of each models
  - Evaluation results with test set

- This notebook will show more information about that Dashboard with Plotly Dash

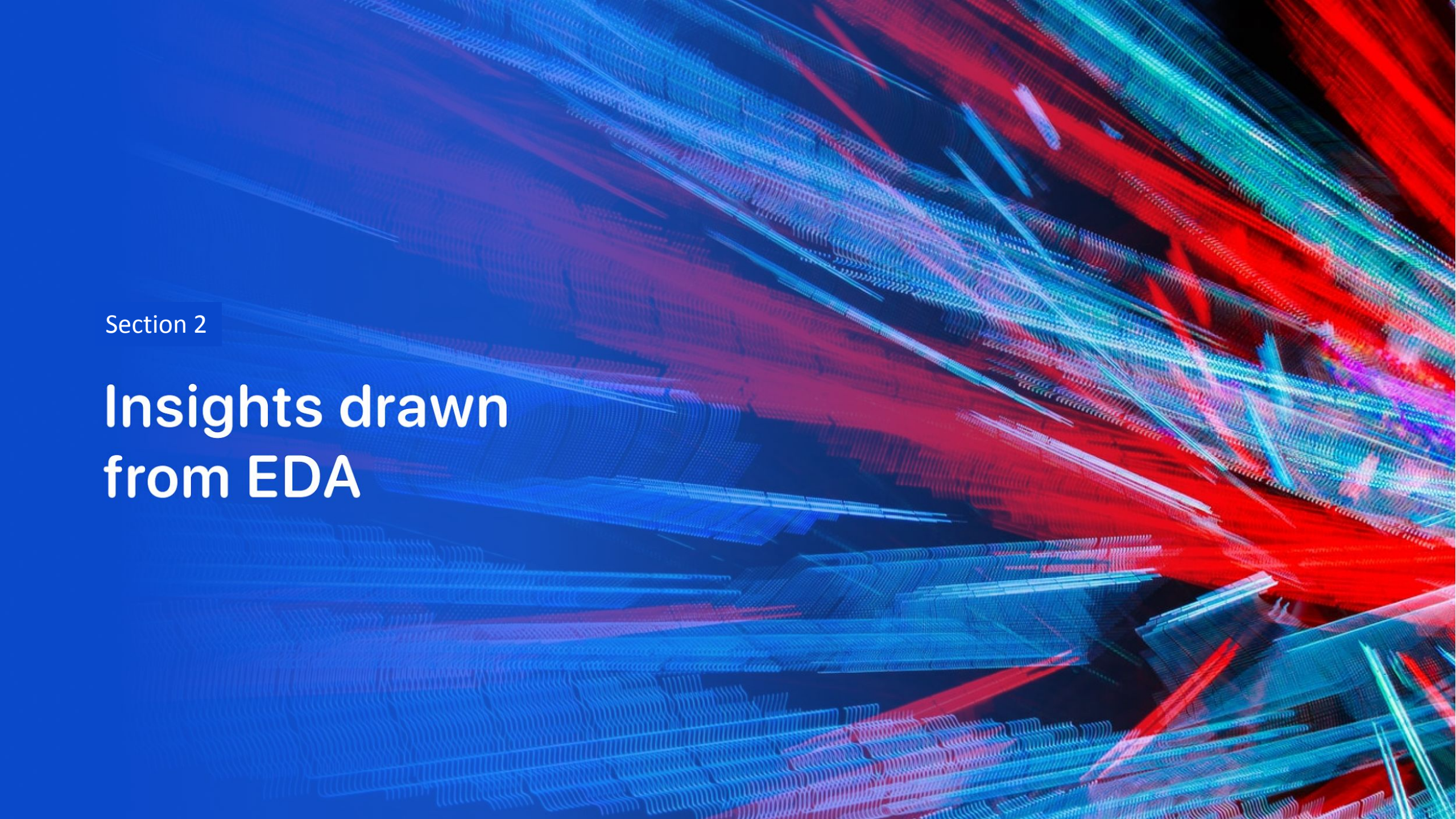
Results

All the methods have the same result, sometimes decision tree have the lower score, but the highest score is 83%



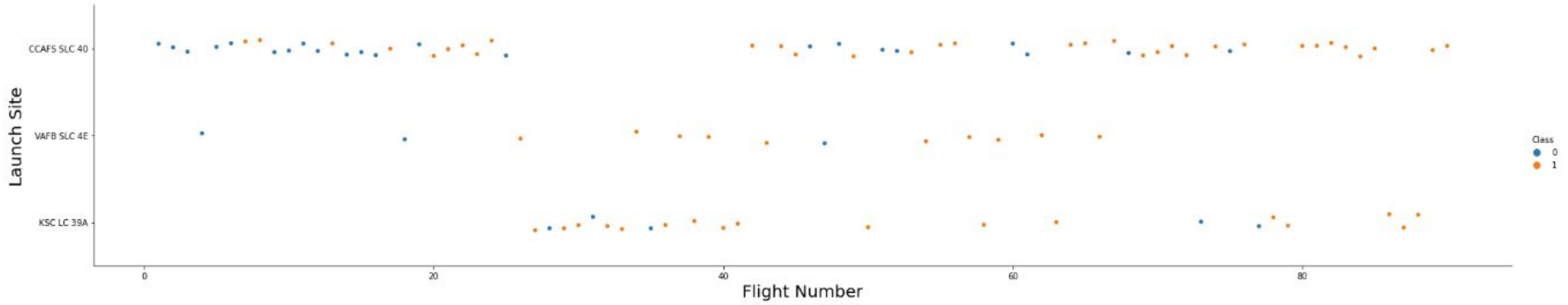
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

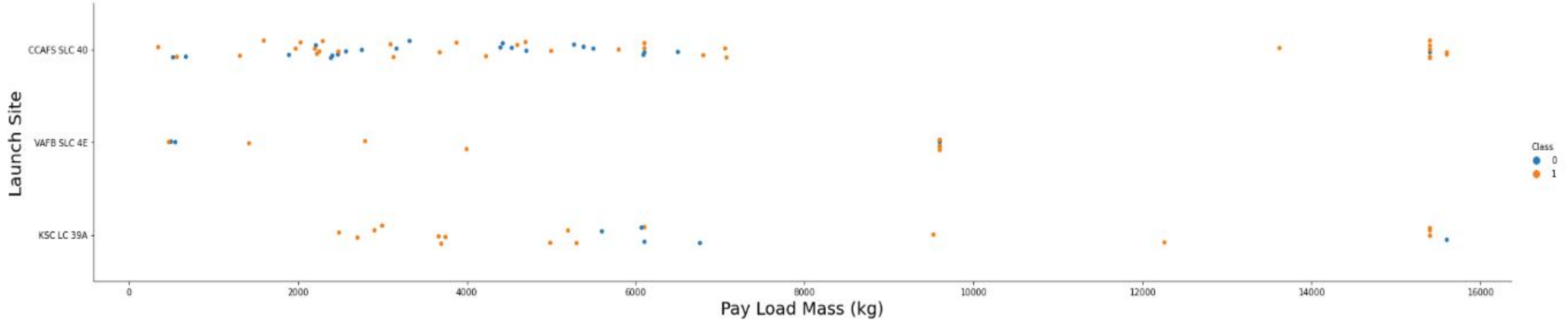


Scatter plot of Flight Number vs. Launch Site

Most of flight launch in CCAFS SLC 40

Payload vs. Launch Site

```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay Load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

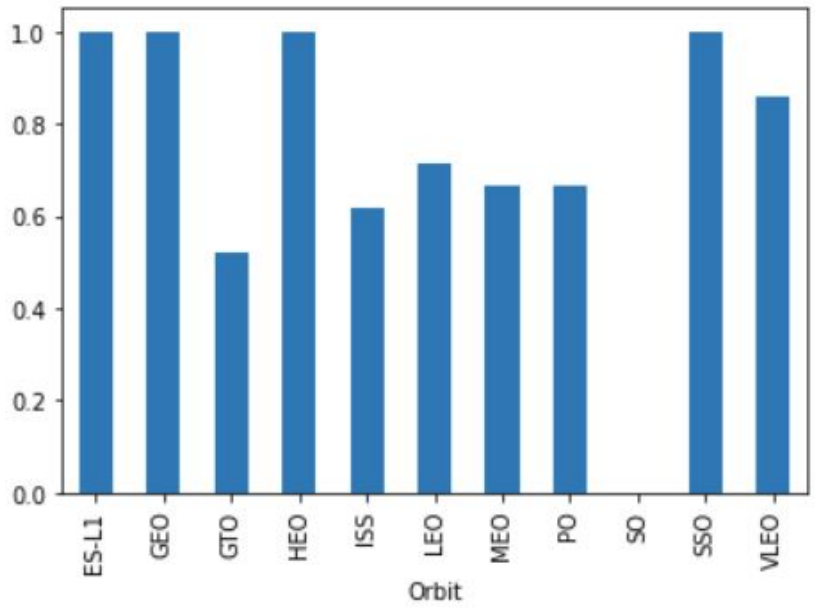


The VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000)

Success Rate vs. Orbit Type

```
# HINT use groupby method on Orbit column and get the mean of Class column
df.groupby(['Orbit']).mean()['Class'].plot(kind='bar')
```

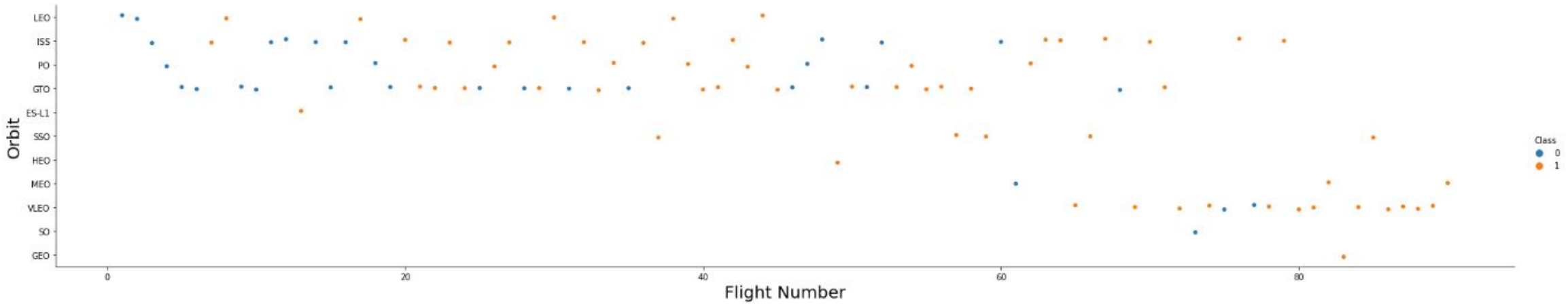
```
<AxesSubplot:xlabel='Orbit'>
```



Orbit Type ES-L1, GEO, HEO, SSO have highest success rate

Flight Number vs. Orbit Type

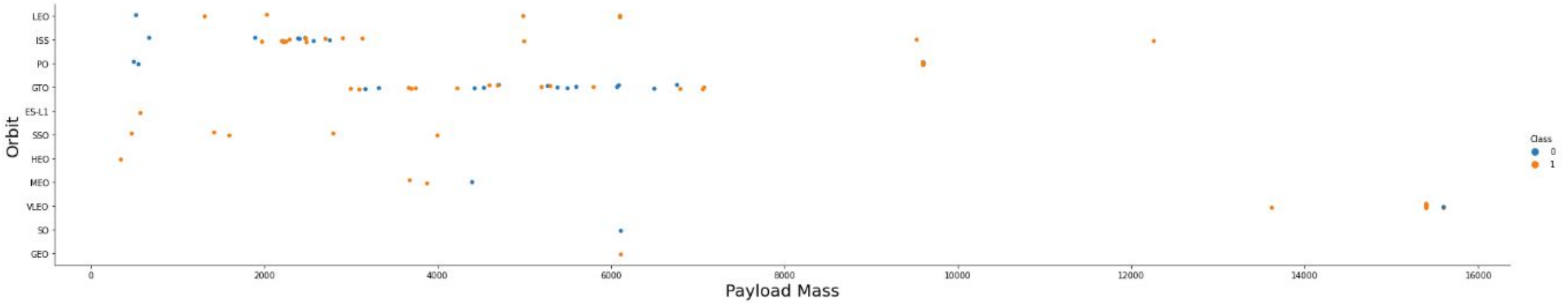
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



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

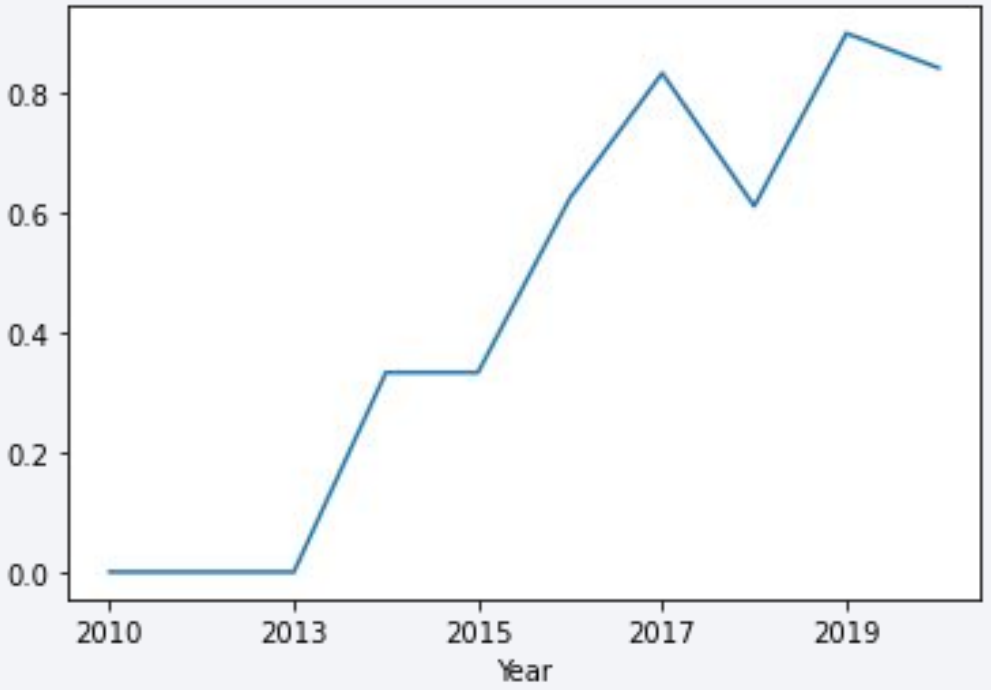
```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



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



The success rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql select distinct(Launch_Site) from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

The names of the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

5 records where launch sites begin with the string 'CCA'

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer='NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
45596

The total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
```

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

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

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad

```
%sql SELECT date from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)' limit 1
```

```
 * sqlite:///my_data1.db
Done.
```

Date
22-12-2015

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select Booster_Version from SPACEXTBL where "Landing _Outcome" = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MAS
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(Mission_Outcome) as "Total outcome" from SPACEXTBL group by Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
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

2015 Launch Records

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, "Landing _Outcome" FROM SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' AND s
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Booster_Version	Launch_Site	Landing _Outcome
10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

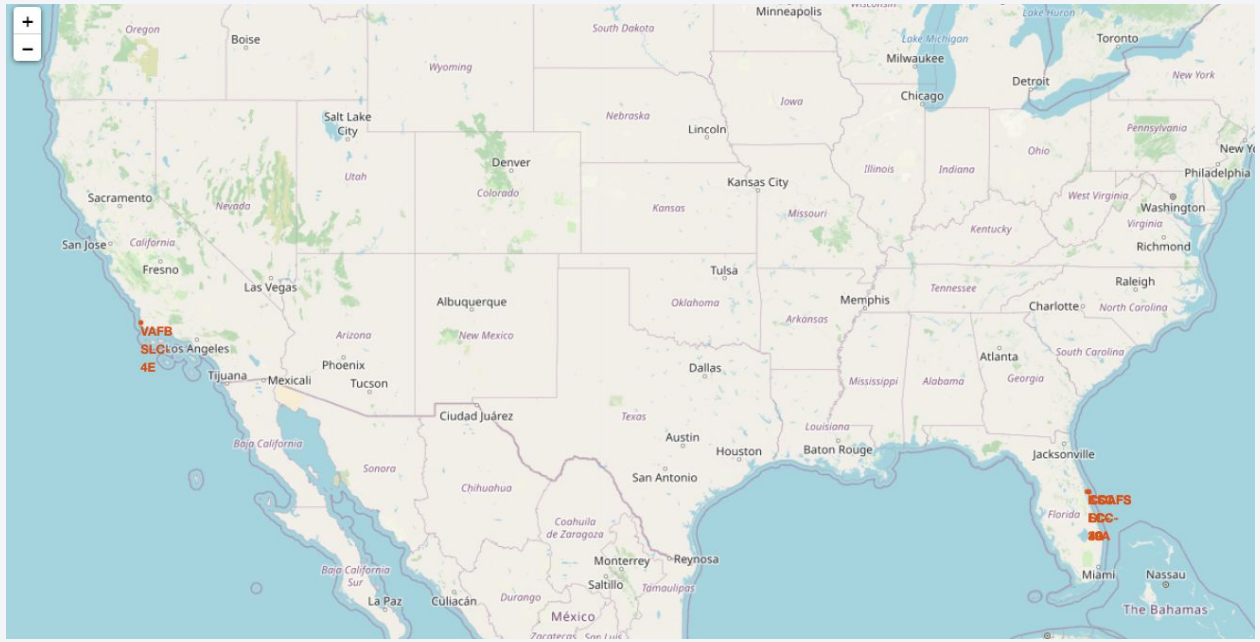
```
%sql select "Landing _Outcome", count("Landing _Outcome") as "count" from SPACEXTBL where date between '04-06-2010' and '20-03-2017' grou
```

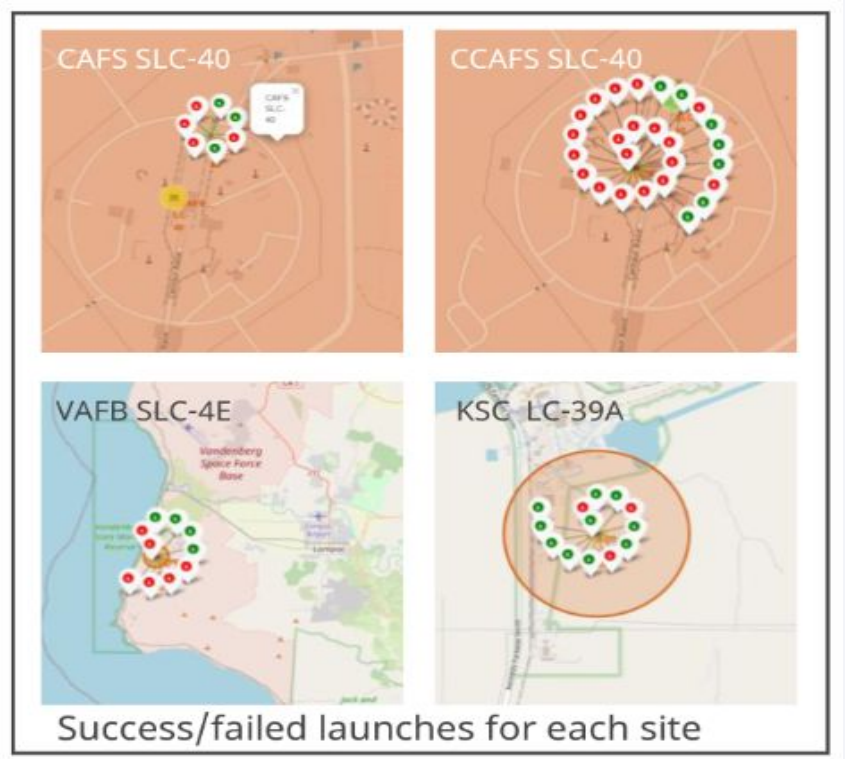
```
 * sqlite:///my_data1.db
Done.
```

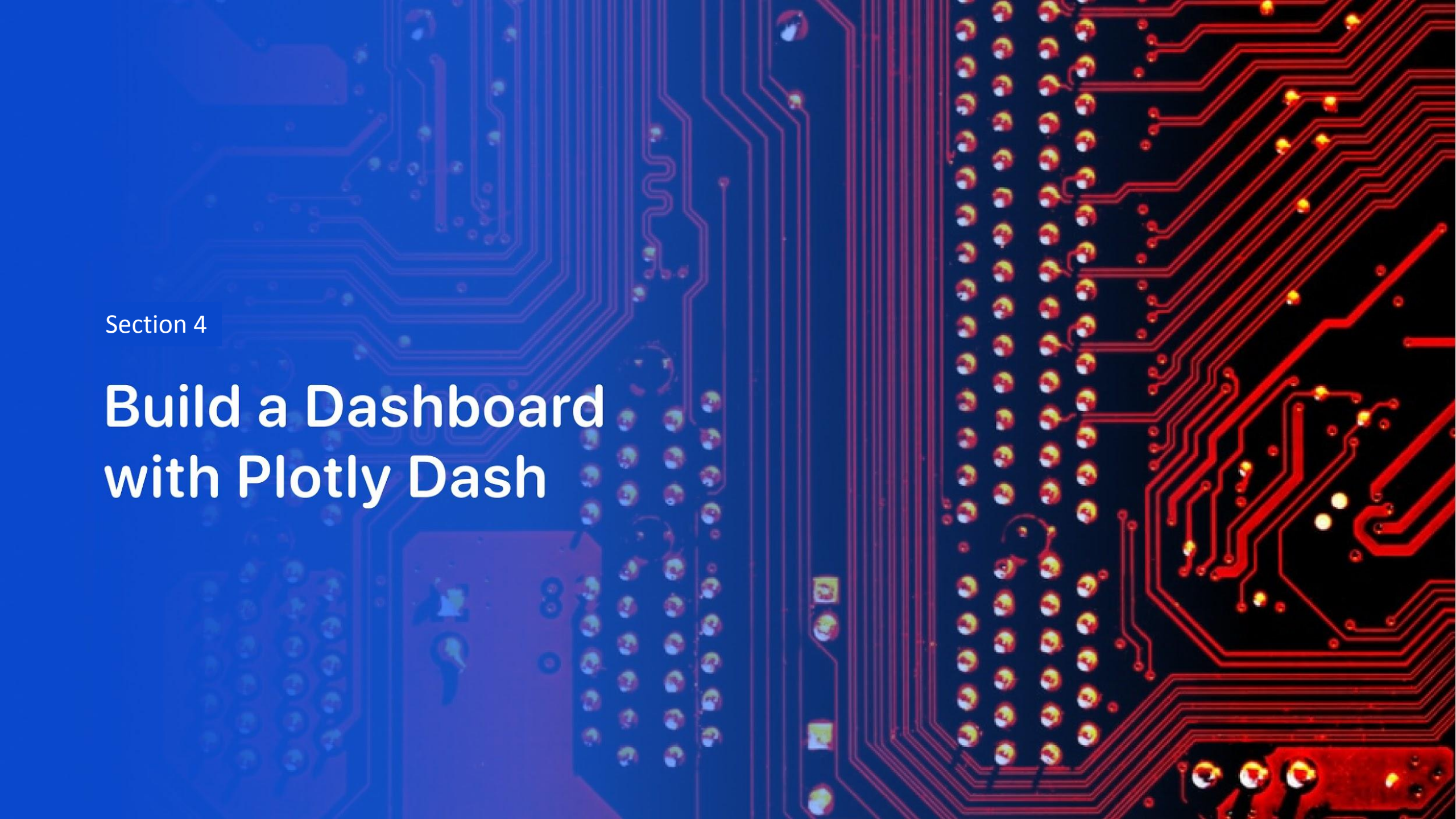
Landing _Outcome	count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1


Section 3
Launch Sites
Proximities Analysis

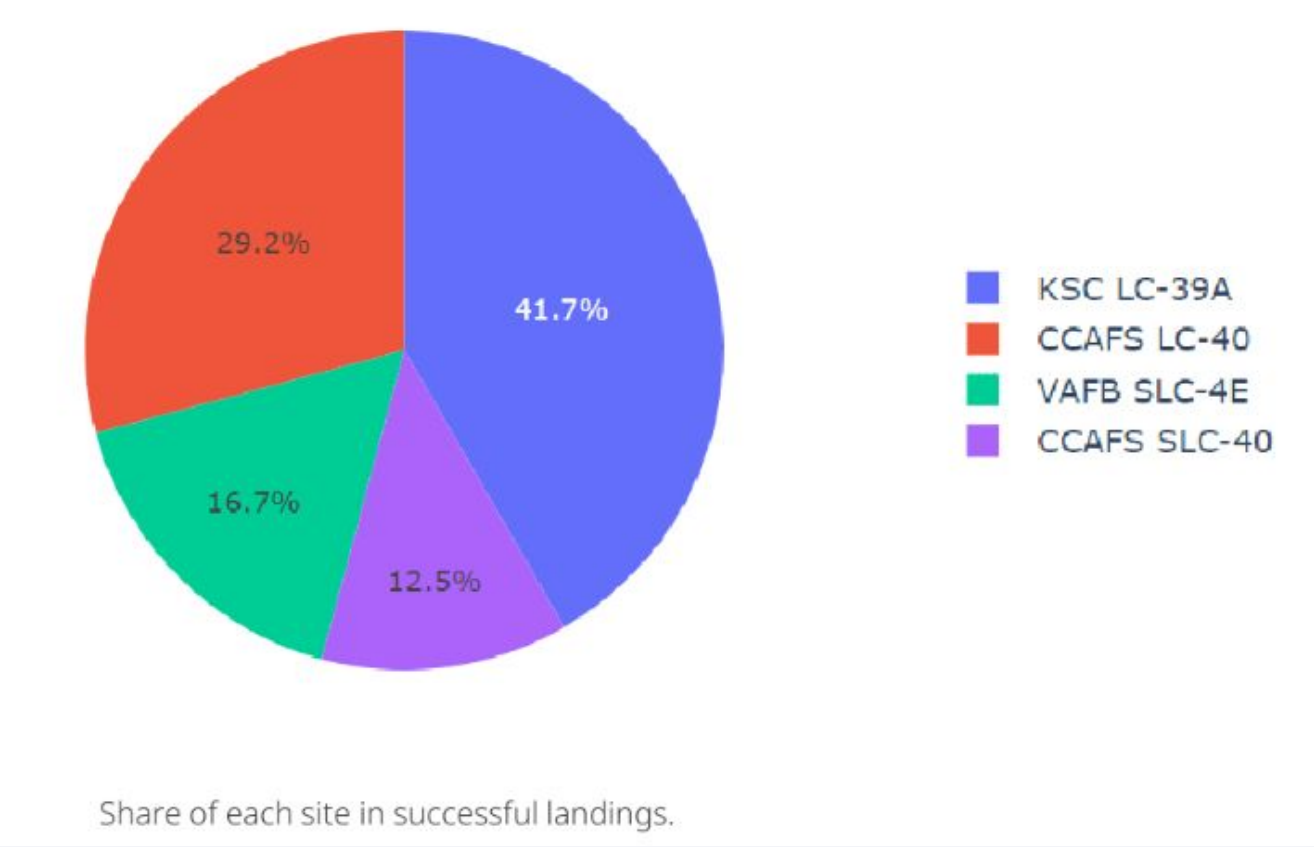
Folium Map All Results



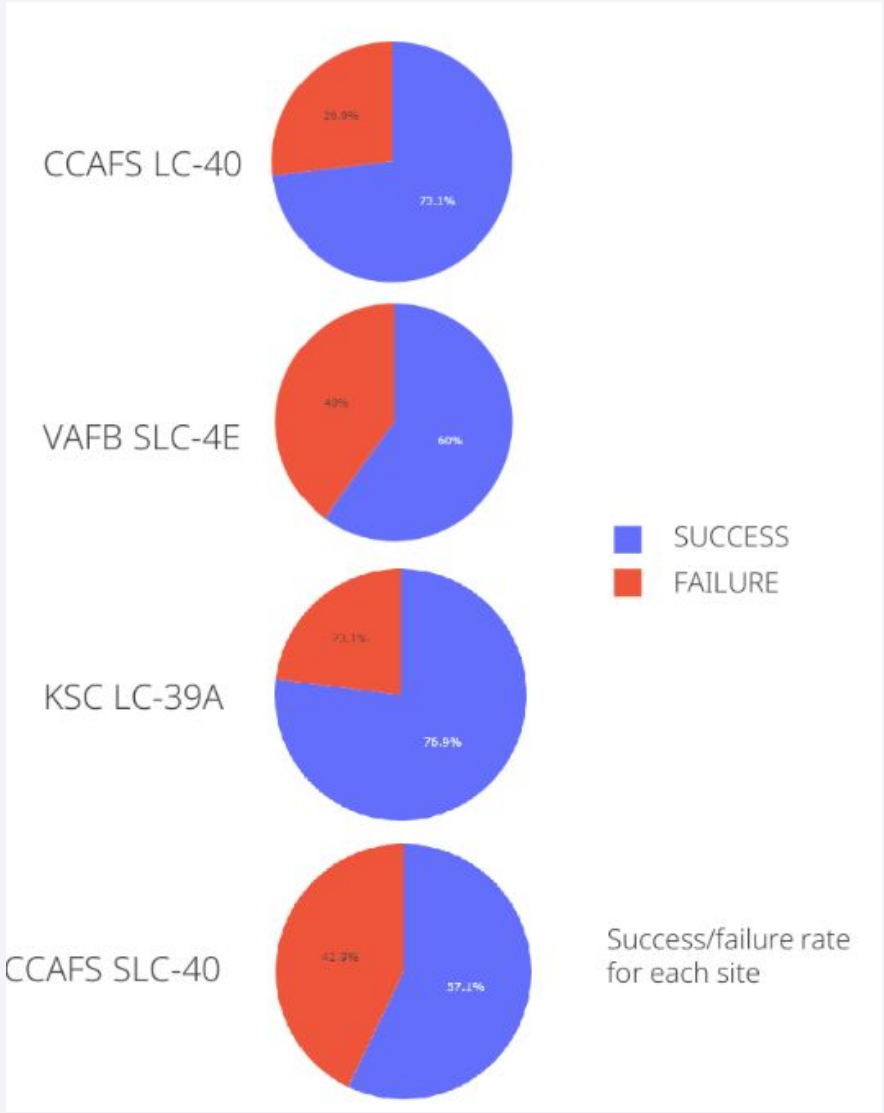



Section 4
Build a Dashboard with Plotly Dash

Dashboard All Results



Share of each site in successful landings.

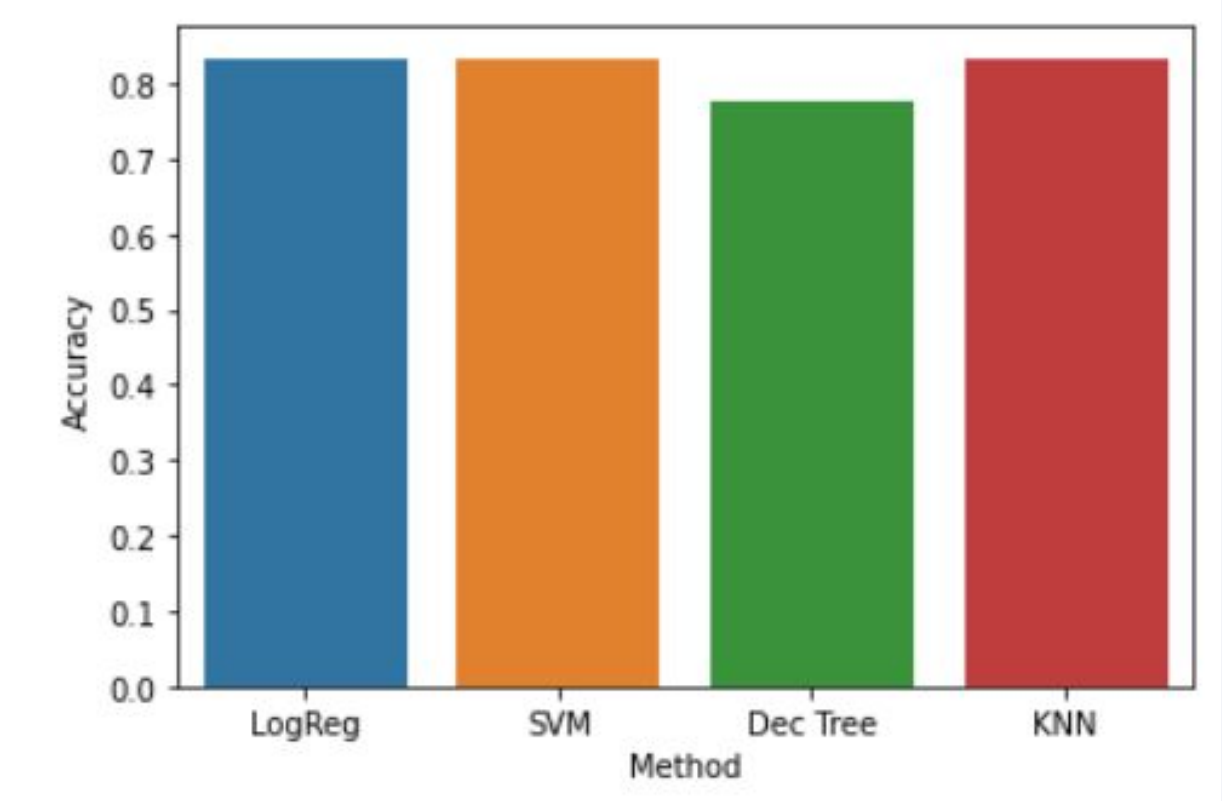


Success/failure rate for each site

Section 5

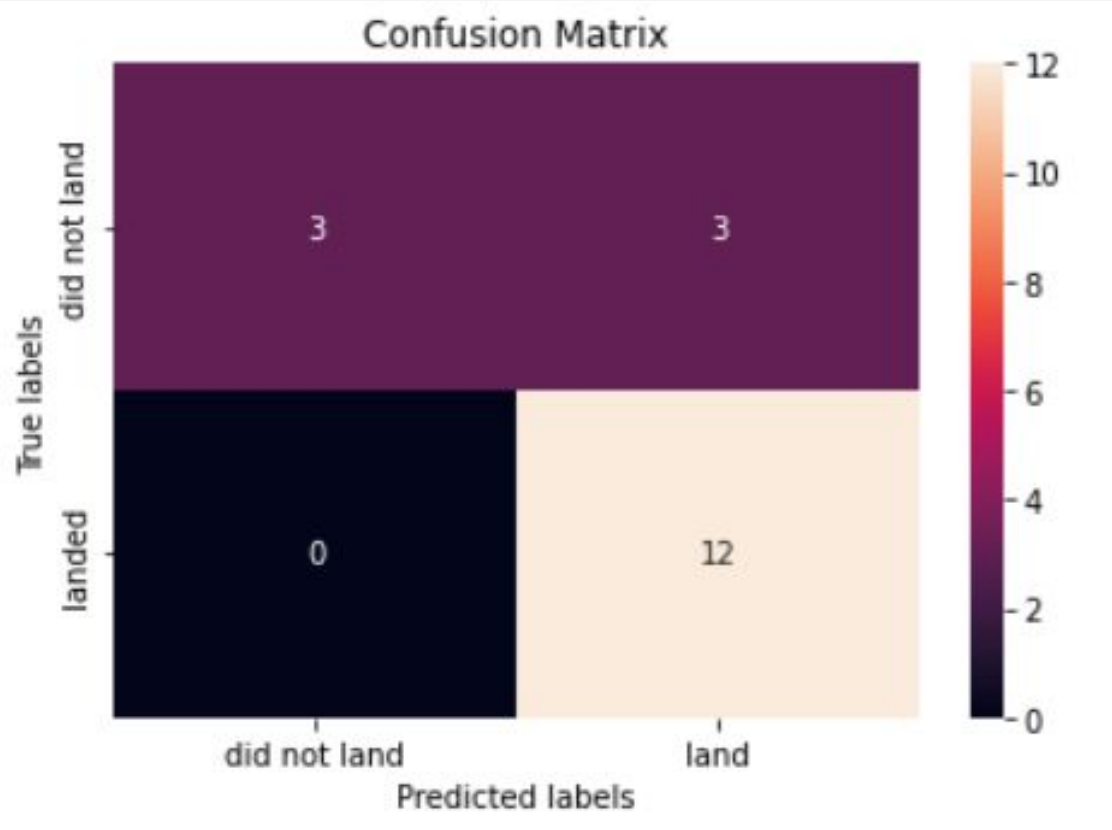
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

- All best model of each method have the same confusion matrix



Conclusions

- The success of SpaceX relies simply on the development of their rockets, with addition of legs and gridfins. The initial versions of the Falcon 9 were devoid of these features and as a result failed to land properly.

- Across the time, SpaceX have dramatically improved their ability to recover rockets. At the same time, they were able to increase the payload mass that could be loaded on top of their rockets.

- This was confirmed by a Feature Importance analysis which confirmed rocket features, number of launches and rocket versions to be the main determinants of a rocket to successfully land.

- The prediction models built on linear regression, SVM, decision tree and KNN all 0.833 accuracy.

- However, launch site and target orbit do not seem to be key attributes for our analysis, and could be considered to be removed to refine the prediction models in the future.

Thank you!