IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Apoorva N
03-04-2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data wrangling
  - EDA with SQL and Visualization
  - Marking launchpads on interactive maps
  - Building interactive dashboards
  - Predictive analysis
- Summary of all results
  - Descriptive EDA results through individual graphs and dashboards
  - Classification results and performance across models

Introduction

- Project background and context
  - 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
- Problems you want to find answers
  - What is the relationship between landing outcome and different variables?
  - Are there any hidden patterns or correlations that can help in better prediction of landing?
  - Ultimately, which model to use to find best prediction for the landing outcome that determines the cost?

Section 1

Methodology

Methodology

Executive Summary

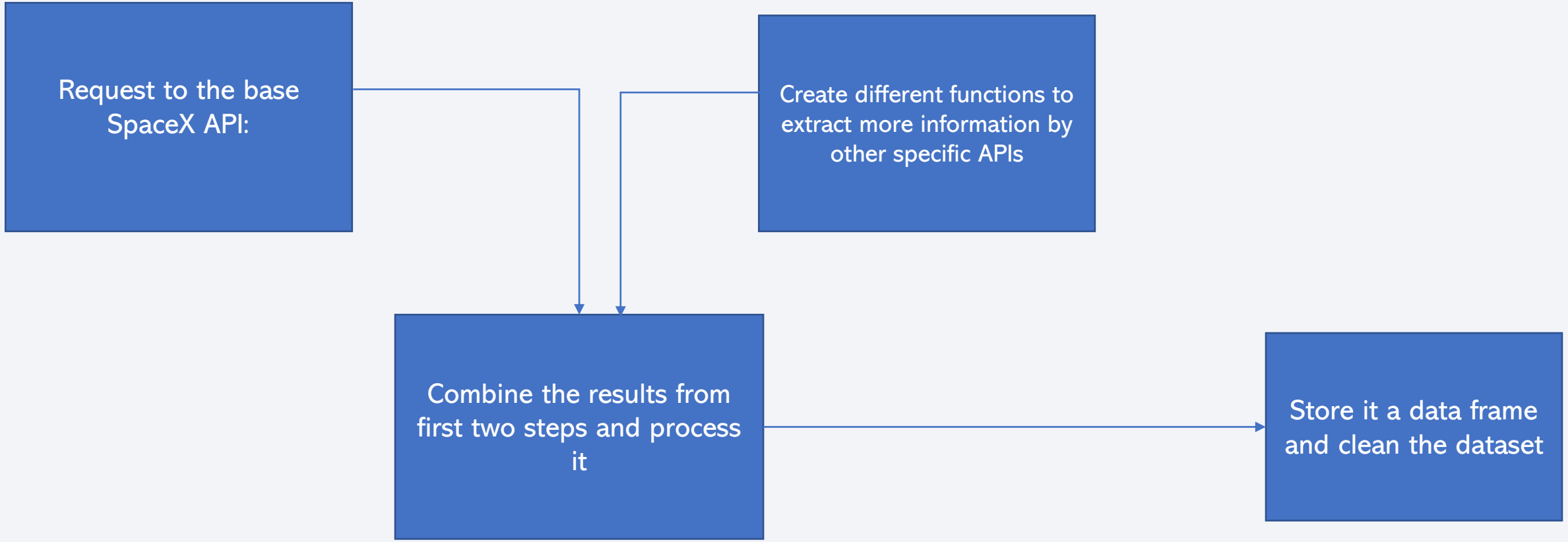
- Data collection methodology:
  - Through web scraping and APIs then loading them into python as data frames for manipulation
- Perform data wrangling
  - Various data cleaning and refining methods used on the data frames, ex- null removal, one hot encoding
- Perform EDA with SQL and visualize through graphs and dashboards
  - Used SQL and visualization tools like interactive dashboards to uncover patterns between different variables impacting landing and finally selecting a list of features for modelling
- Perform predictive analysis using classification models
  - Different models used for predicting the landing which were then tested to choose the best performing one through evaluation scores

Data Collection

- Web Scraping
  - Falcon 9 Launch data was obtained by web scraping related Wiki pages
  - Source: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Was done using BeautifulSoup and response libraries in Python
- APIs
  - Specifically SpaceX RestAPIs were also used to target another endpoint to gather specific information
  - Source/endpoint URLs: api.spacexdata.com/v4/
  - More specifically RestAPIs and python's json() and json_normalize() methods were used

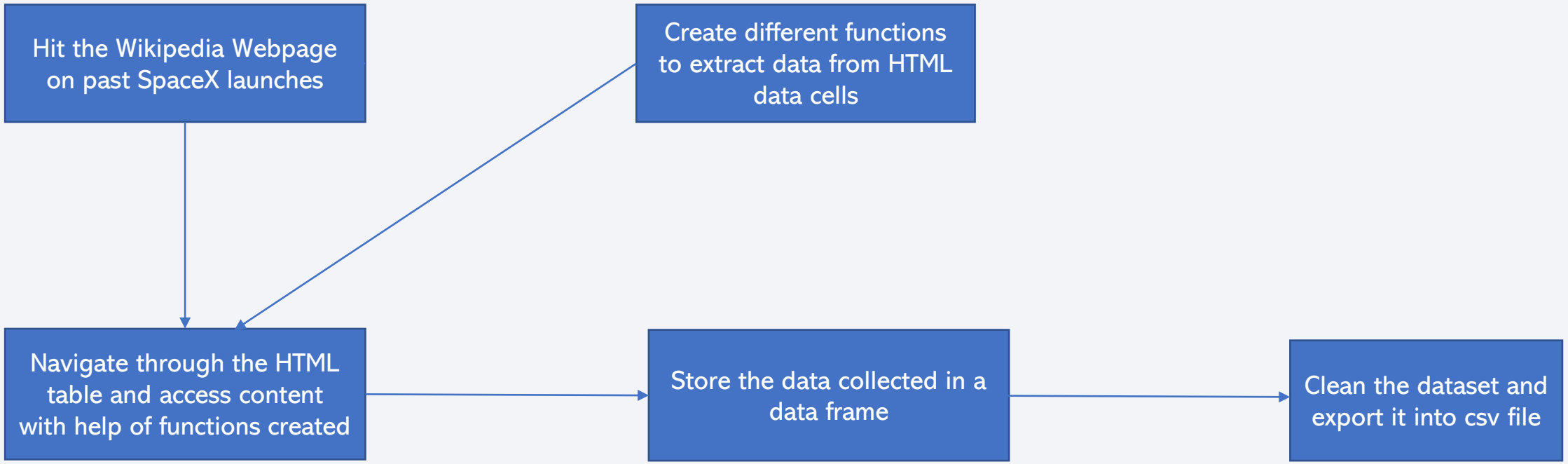
"

Data Collection – SpaceX API



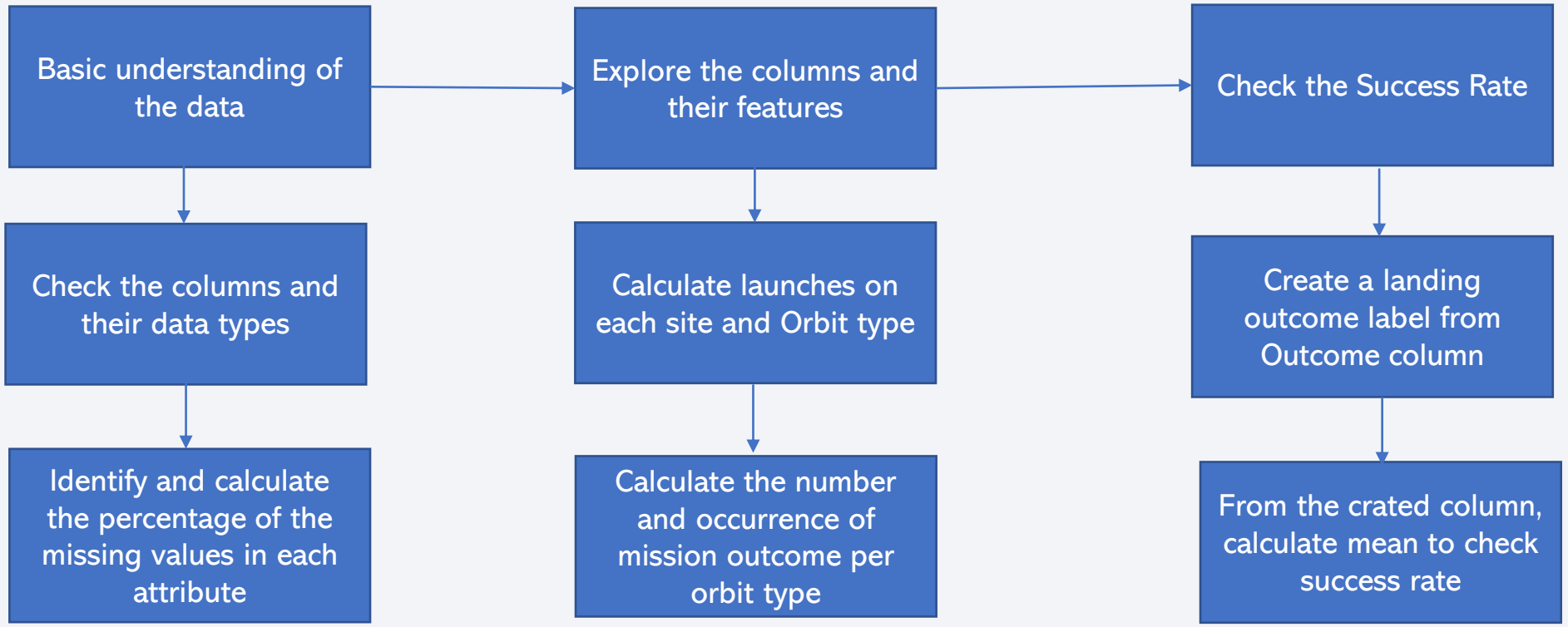
Complete notebook link here

Data Collection - Scraping



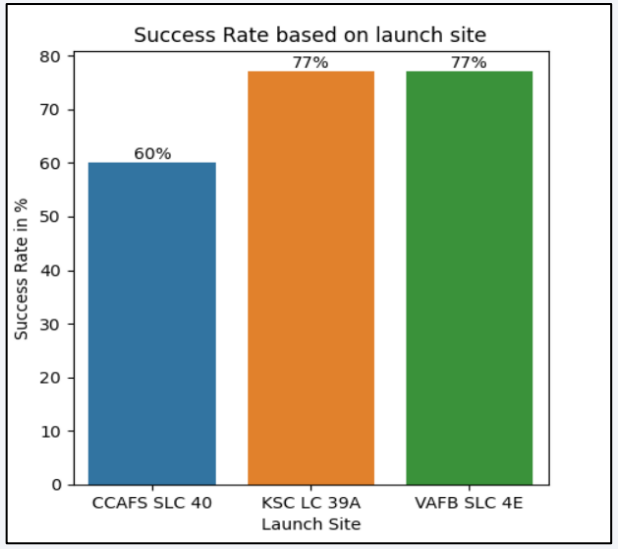
- Complete notebook link here

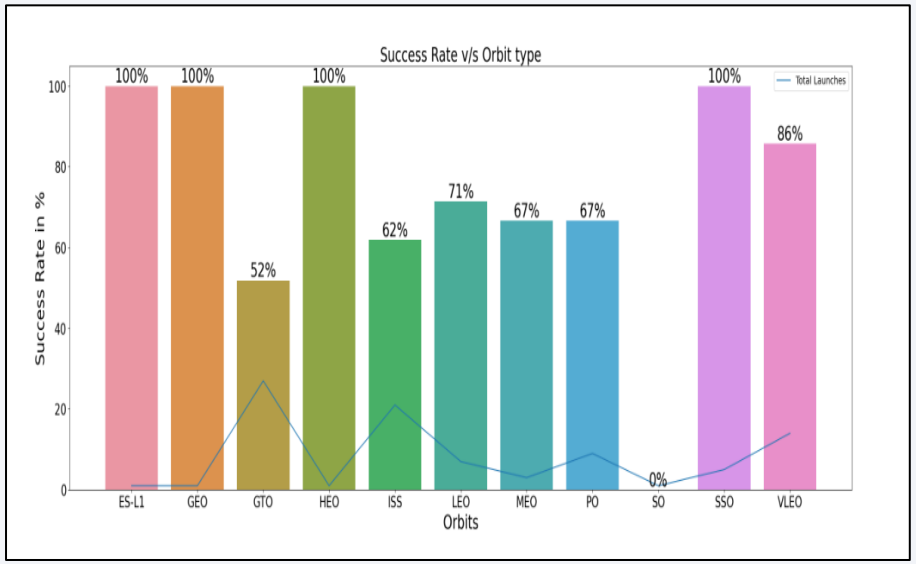
Data Wrangling

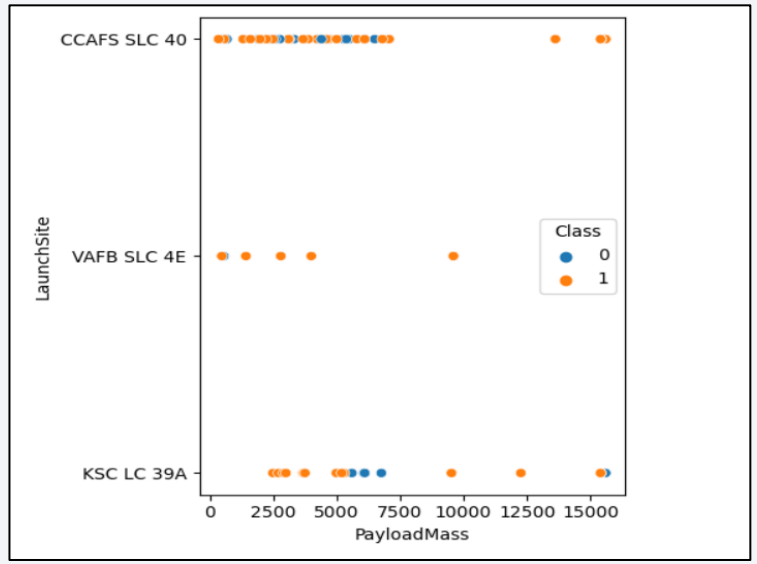


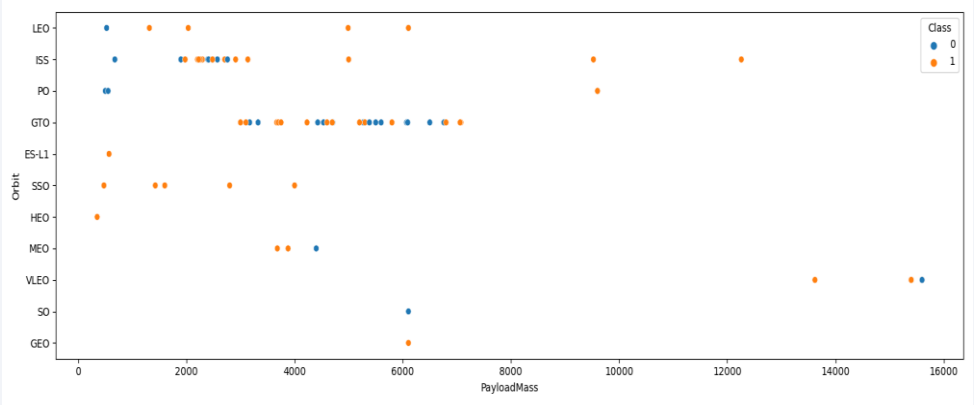
- Complete notebook link here

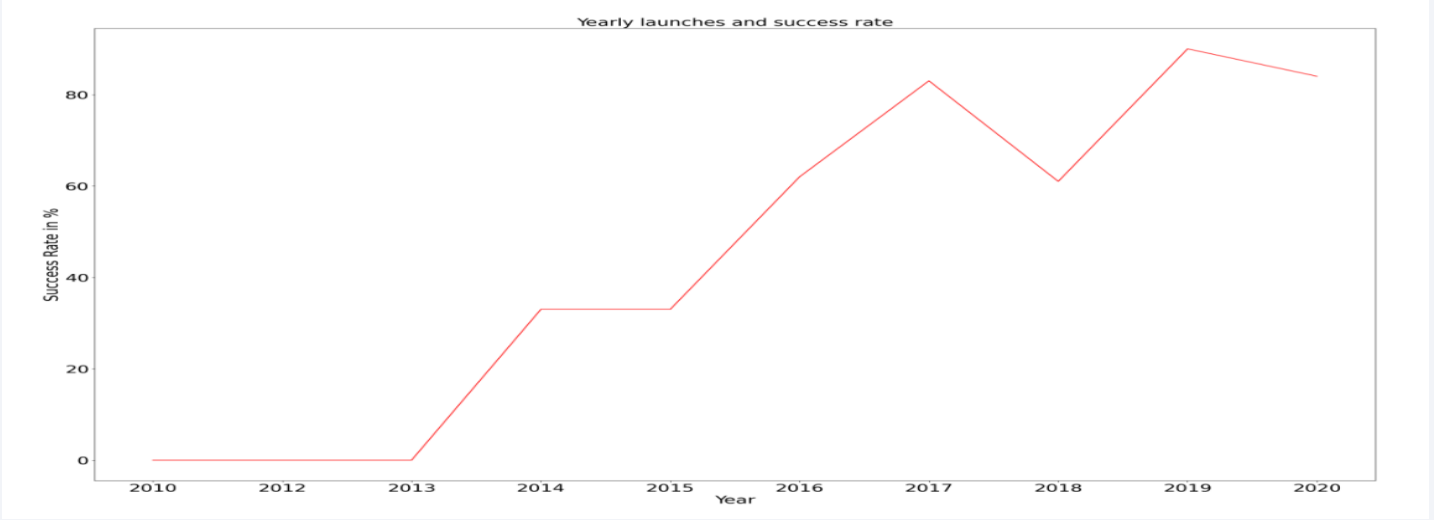
EDA with Data Visualization











- Complete notebook link here

EDA with SQL

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

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

```
%%sql select min(Date_new) from
(
    select *, substr(Date, 7) || '-' || substr(Date, 4, 2) || '-' || substr(Date, 1, 2)  as Date_new
    from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)'
)
```

```
%sql select distinct Booster_Version from SPACEXTBL where  "Landing _Outcome" = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
%%sql select case when Mission_Outcome like 'Success%' then 'Success' when Mission_Outcome like 'Fail%' then 'Failure'
  end as Mission_Outcomes ,count(*) as "count"
from SPACEXTBL group by Mission_Outcomes
```

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

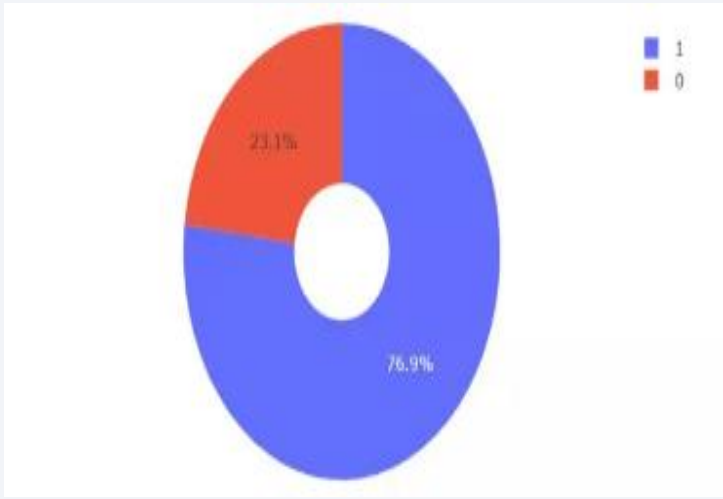
```
%%sql select substr(Date,4,2) as Months, "Landing _Outcome", Booster_Version, Launch_Site
from SPACEXTBL where substr(Date,7,4) = '2015' and "Landing _Outcome" = "Failure (drone ship)"
```

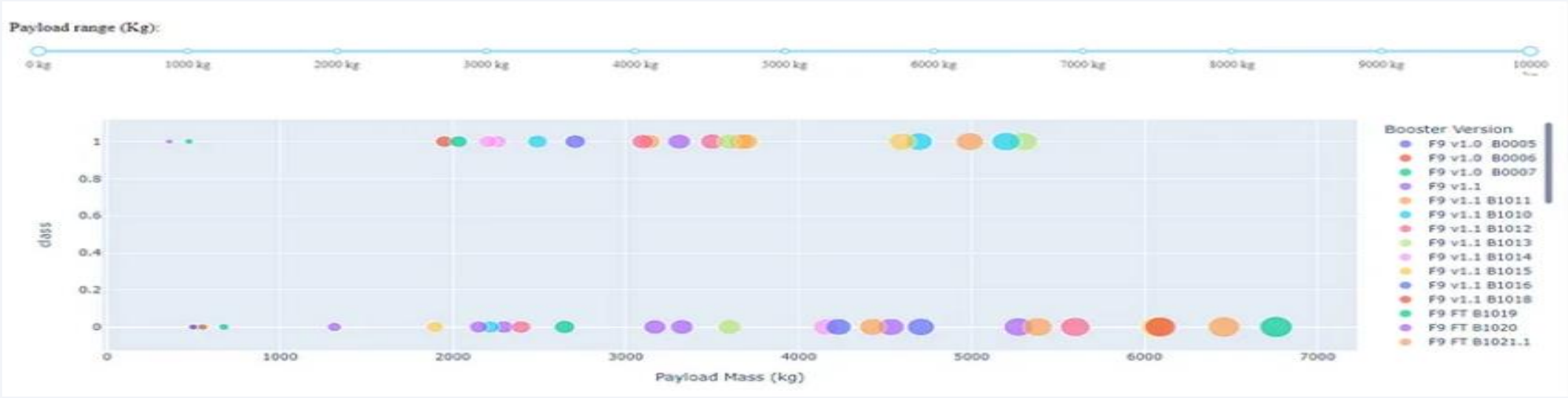
```
%%sql select Date_new, count(*) as successful_landing_outcomes
from (
      select *, substr(Date, 7) || '-' || substr(Date, 4, 2) || '-' || substr(Date, 1, 2)  as Date_new from SPACEXTBL
      where "Landing _Outcome" like 'Success%' and Date_new between '2010-06-04' and '2017-03-20'
      )
group by 1
order by Date_new desc
```


- Complete notebook link here

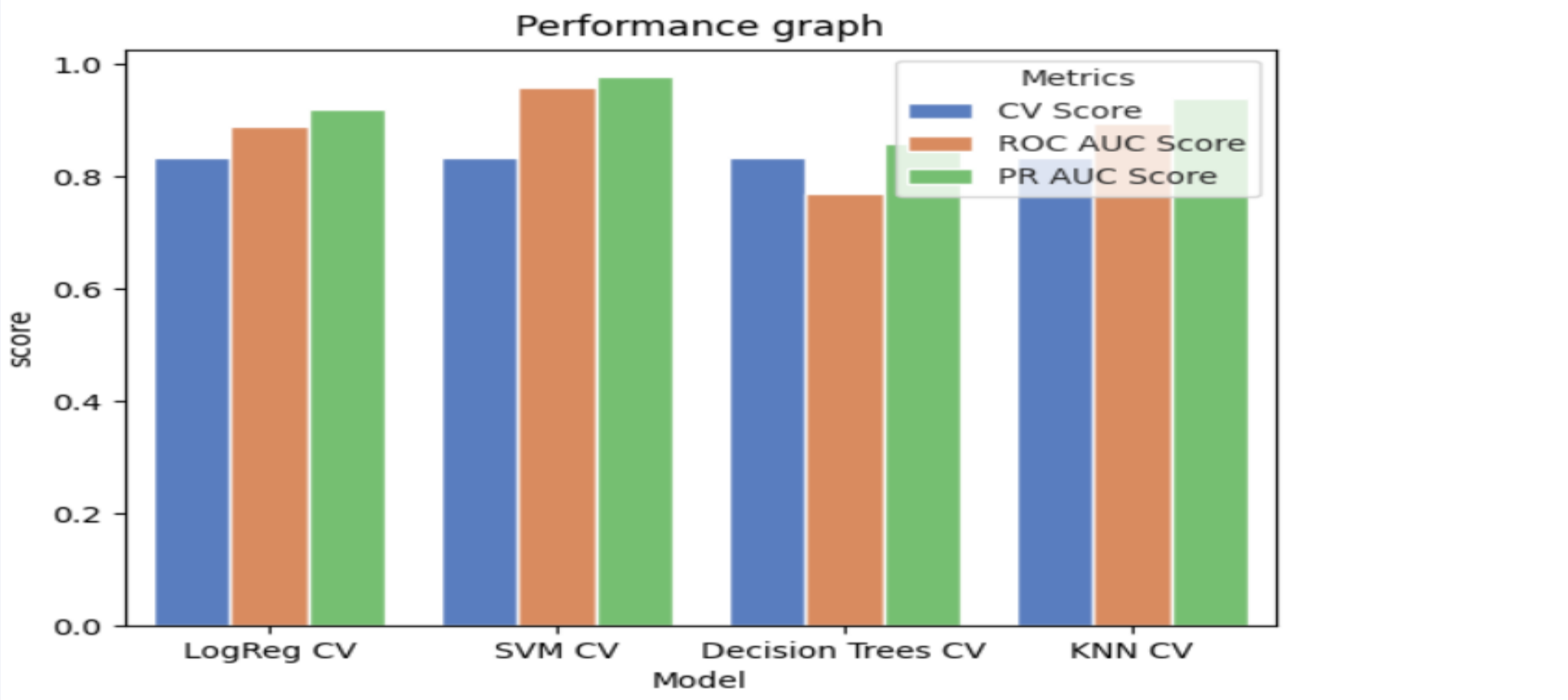
Build an Interactive Map with Folium

Build a Dashboard with Plotly Dash





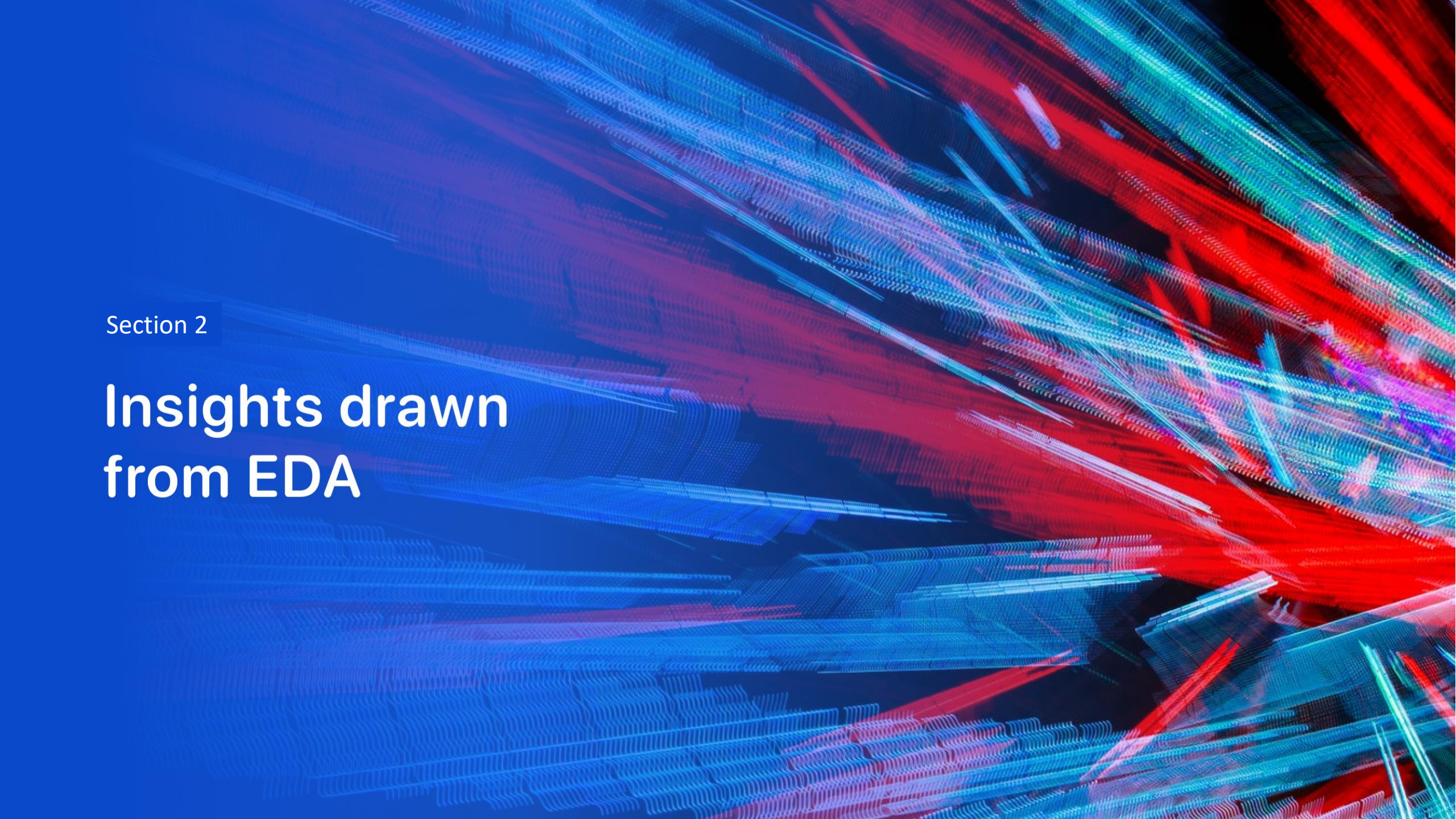
Predictive Analysis (Classification)



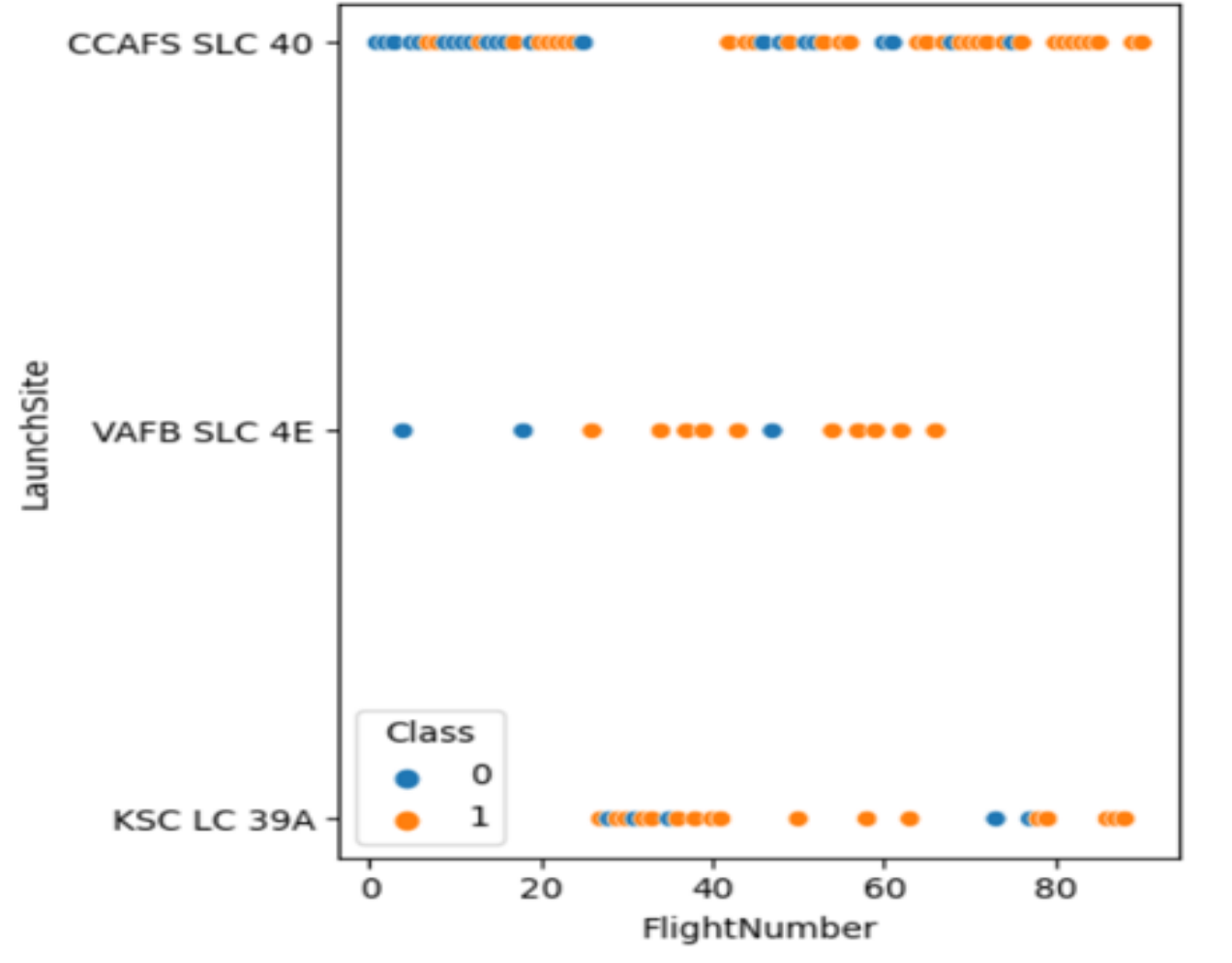
- Complete notebook link [here]

Results

- Volume of Launches from Launch Sites, in descending order: CCAFS SLC 40, KSC LC 39A VAFB SLC 4E
- Lower payload masses seem to have better success rate comparatively but it's not entirely coherent as the number of launches with higher payload masses is less
- Even though by graphs, we can see that ES-L1,GEO,HEO have 100% success rates but the number of launches in those orbits are only a few so best performing orbit is VLEO
- Successful landings are inversely proportional to launch years
- Out of various model used and checked for prediction, SVM CV performs best considering the ROC AUC and PR AUC score

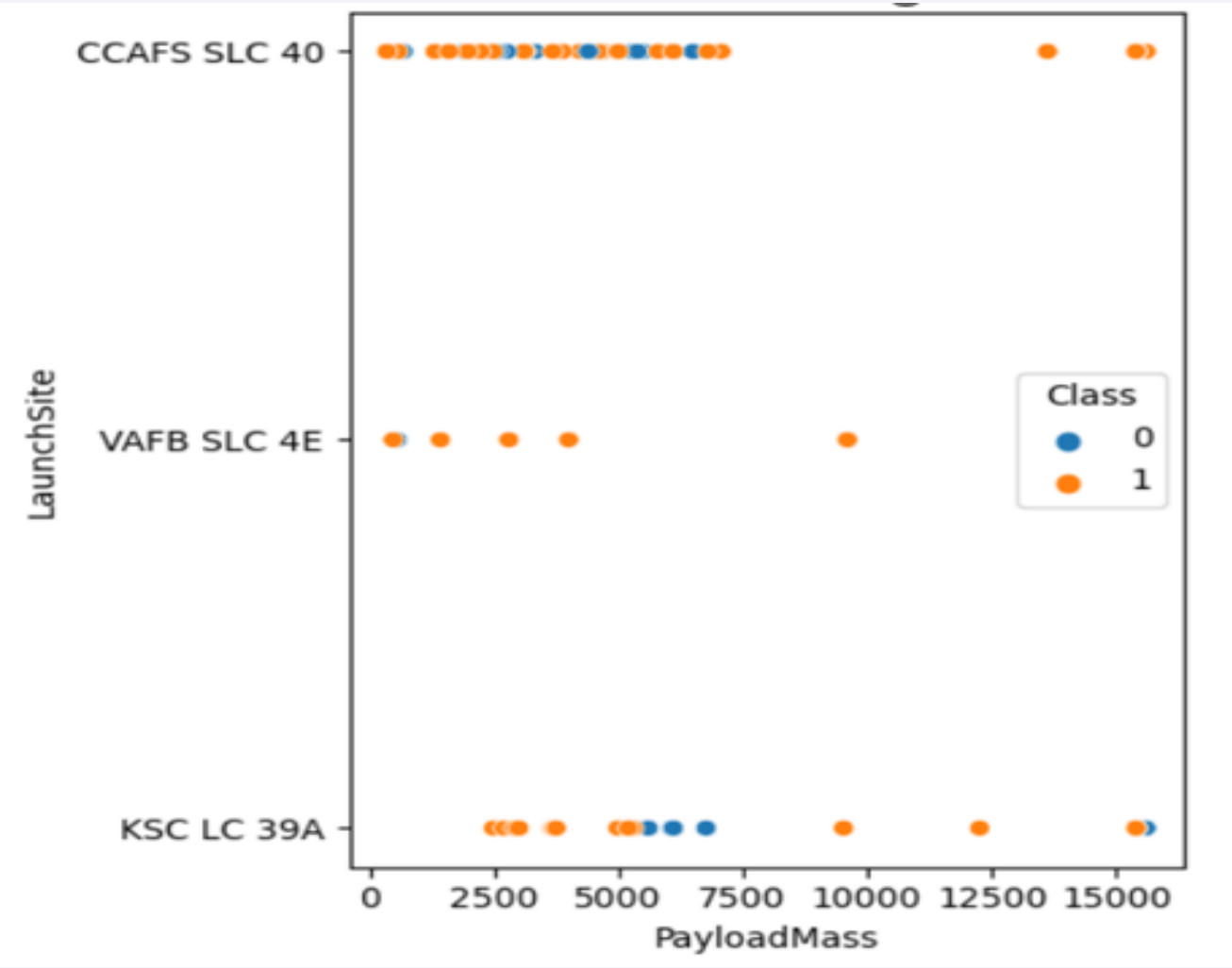

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAFS SLC 40 – has maximum number of launches
- KSC LC 39A has the least failure rate compared to all other launch sites

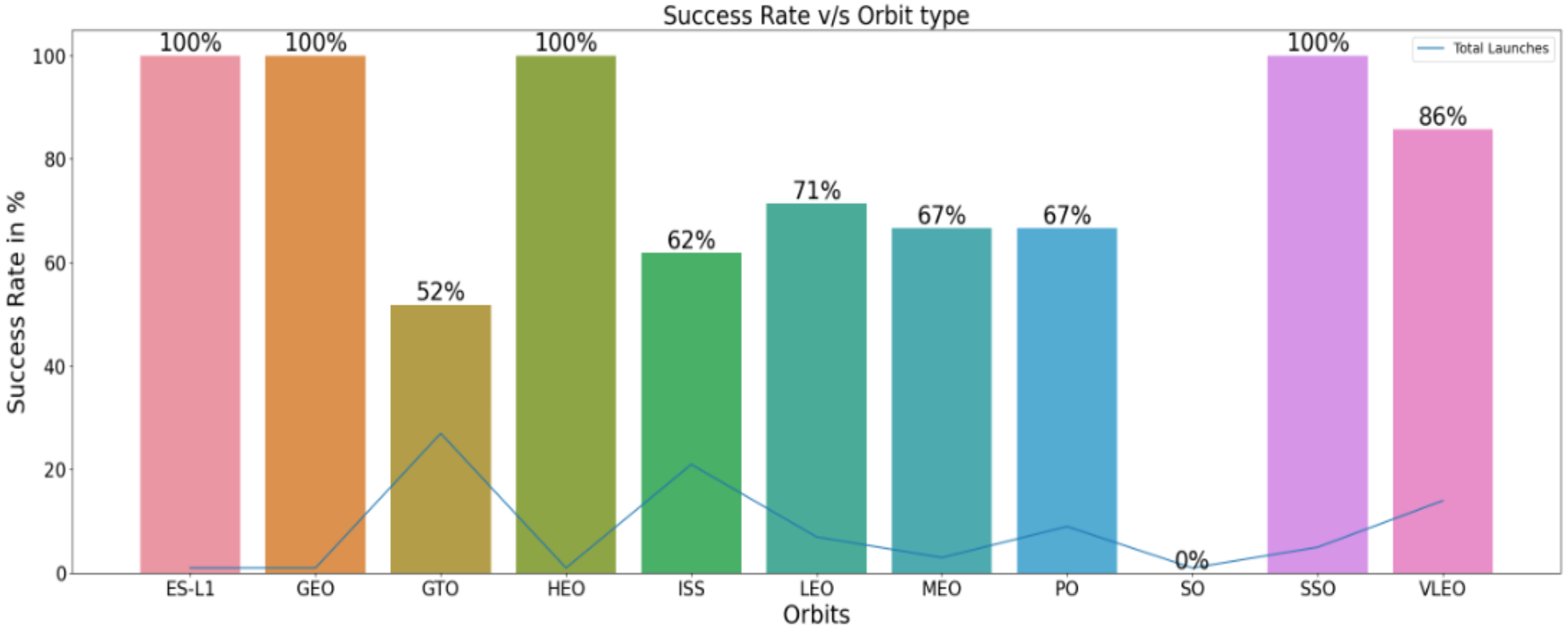
Payload vs. Launch Site



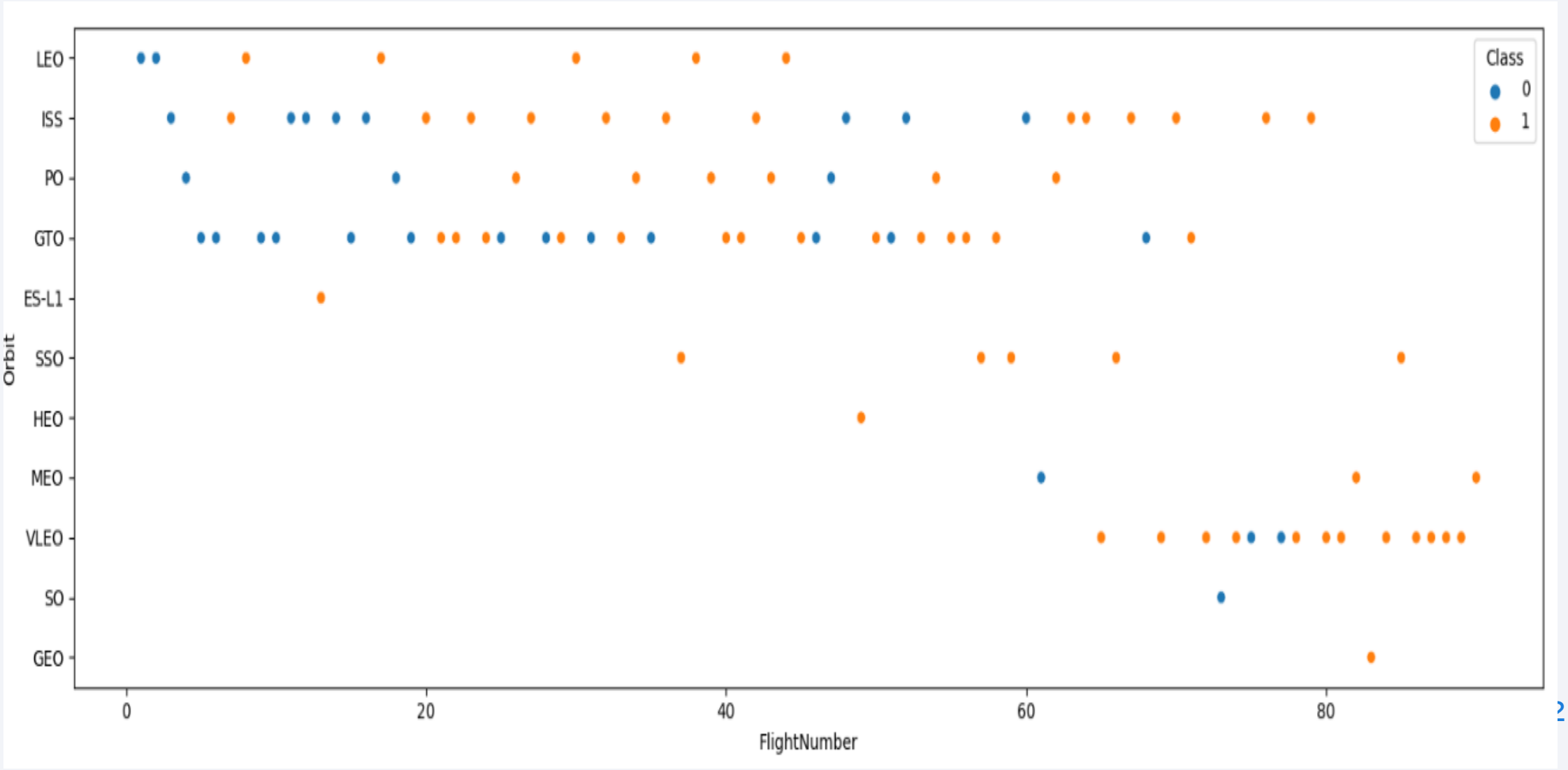
- Lower payload mass seem to have better success rate comparatively
- However, it is not entirely coherent as the number of launches with higher payload mass is less

Success Rate vs. Orbit Type

Even though by graphs, we can see that ES-L1,GEO,HEO have 100% success rates but the number of launches in those orbits are only a few so best performing orbit is VLEO

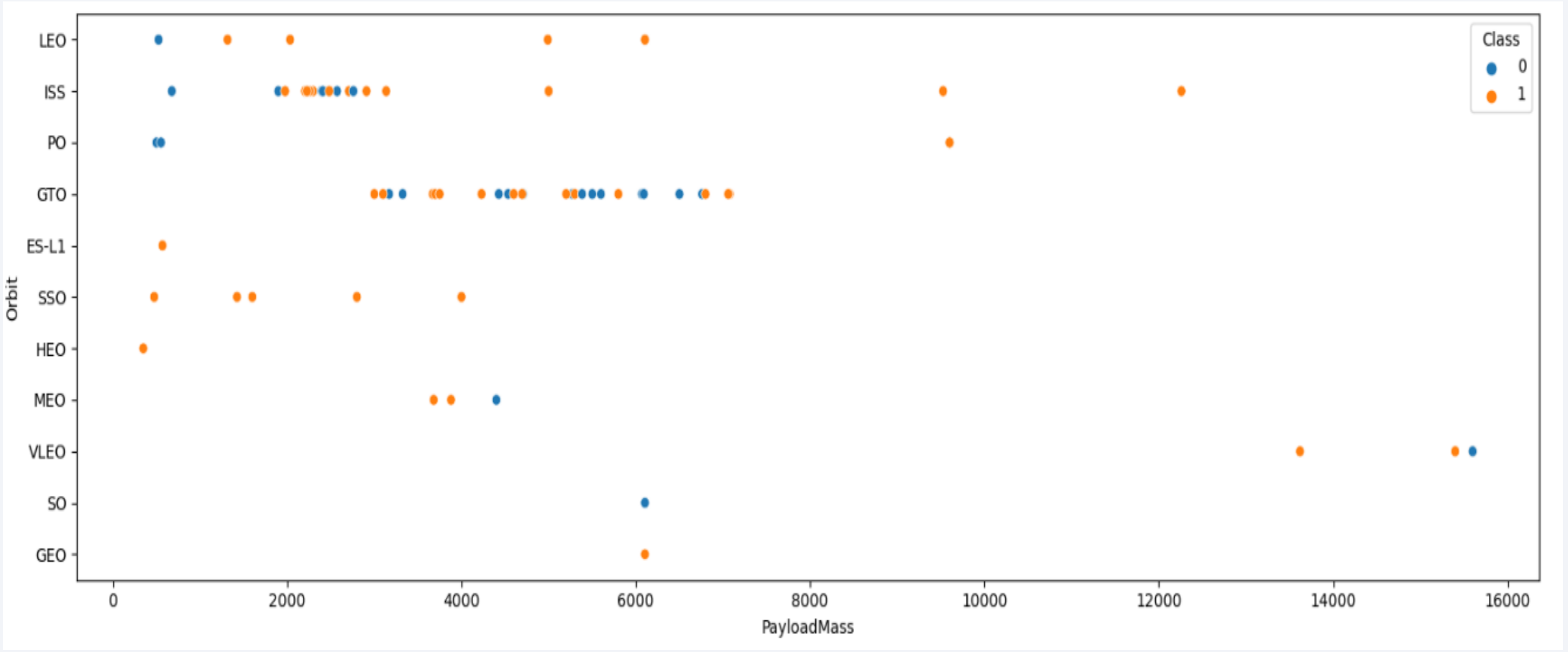


Flight Number vs. Orbit Type



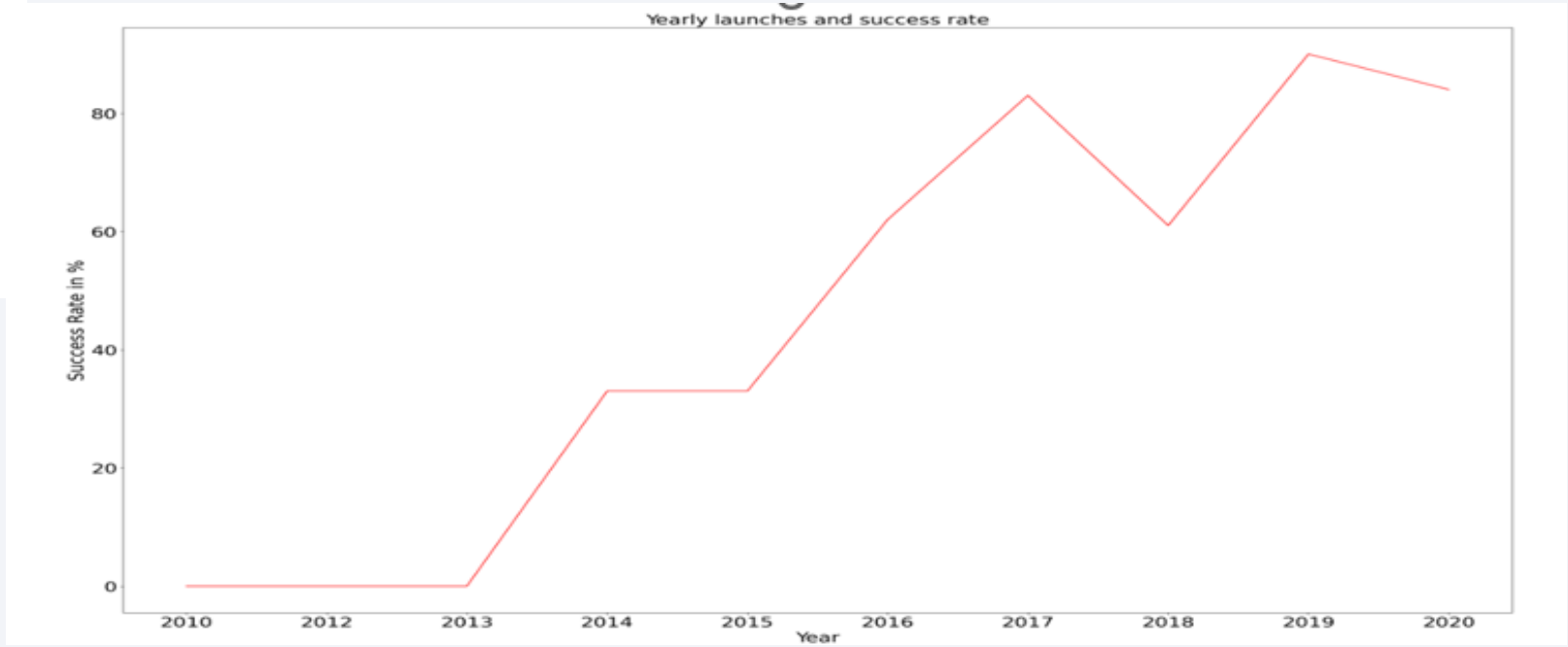
Payload vs. Orbit Type

- They are only a handful of orbit types were high payload mass have been launched



Launch Success Yearly Trend

	Date	Successful_Launches	Total_Launches	Success_Rate
8	2019	9	10	90.0
9	2020	16	19	84.0
6	2017	15	18	83.0
5	2016	5	8	62.0
7	2018	11	18	61.0
3	2014	2	6	33.0
4	2015	2	6	33.0
0	2010	0	1	0.0
1	2012	0	1	0.0
2	2013	0	3	0.0



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

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

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

Average Payload Mass by F9 v1.1

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

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

```
%%sql select min(Date_new) from
(
    select *, substr(Date, 7) || '-' || substr(Date, 4, 2) || '-' || substr(Date, 1, 2)  as Date_new
    from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)'
)
```

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

min(Date_new)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select distinct Booster_Version from SPACEXTBL where  "Landing _Outcome" = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * 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

```
%%sql select case when Mission_Outcome like 'Success%' then 'Success' when Mission_Outcome like 'Fail%' then 'Failure'
  end as Mission_Outcomes ,count(*) as "count"
from SPACEXTBL group by Mission_Outcomes
```

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

Mission_Outcomes	count
Failure	1
Success	100

Boosters Carried Maximum Payload

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

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

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

2015 Launch Records

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%%sql select substr(Date,4,2) as Months, "Landing _Outcome", Booster_Version, Launch_Site
from SPACEXTBL where substr(Date,7,4) = '2015' and "Landing _Outcome" = "Failure (drone ship)"
```

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

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

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

```
%%sql select Date_new, count(*) as successful_landing_outcomes
from (
      select *, substr(Date, 7) || '-' || substr(Date, 4, 2) || '-' || substr(Date, 1, 2)  as Date_new from SPACEXTBL
      where "Landing _Outcome" like 'Success%' and Date_new between '2010-06-04' and '2017-03-20'
      )
group by 1
order by Date_new desc
```

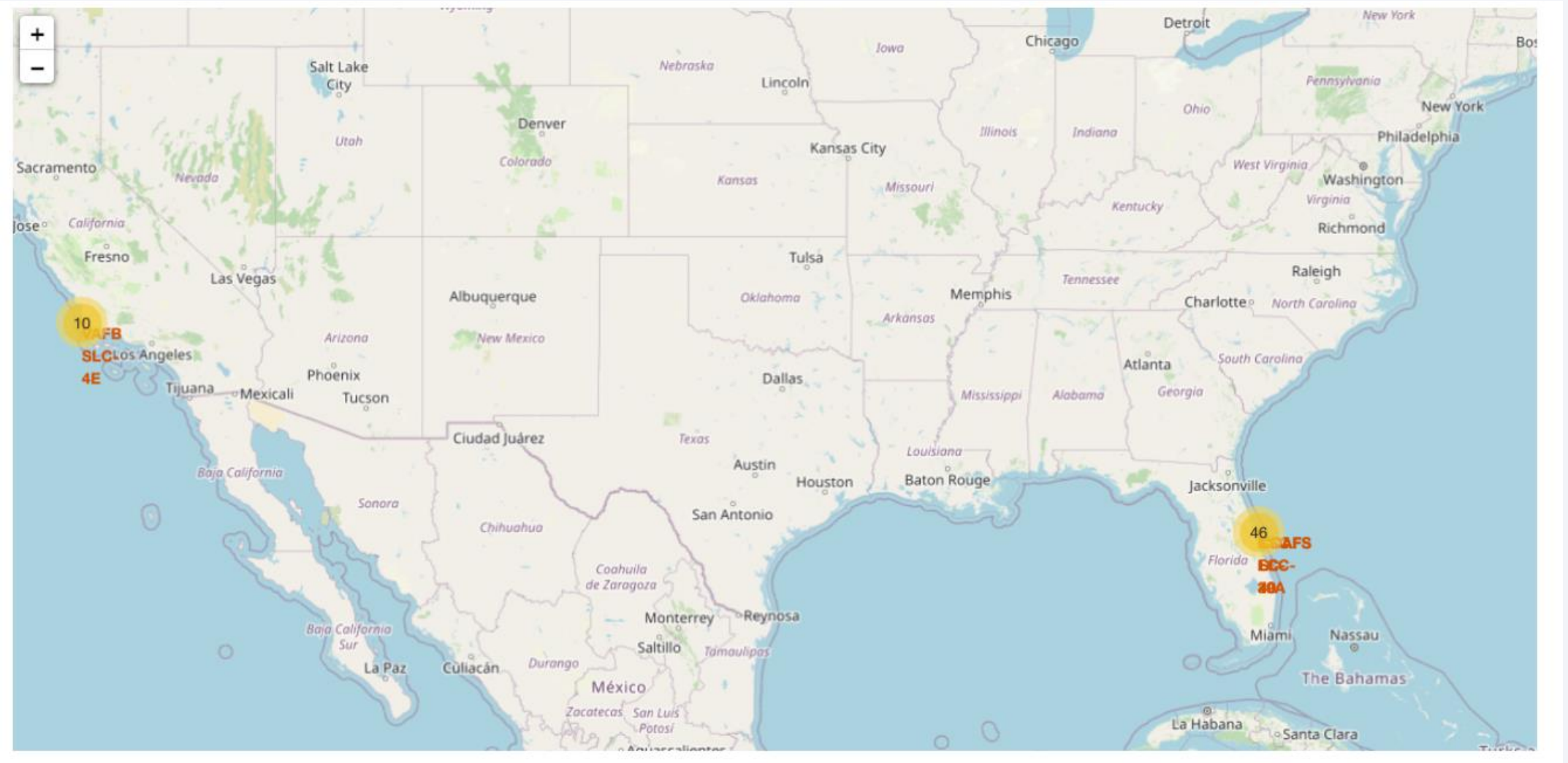
```
 * sqlite:///my_data1.db
Done.
```

Date_new	successful_landing_outcomes
2017-02-19	1
2017-01-14	1
2016-08-14	1
2016-07-18	1
2016-05-27	1
2016-05-06	1
2016-04-08	1
2015-12-22	1

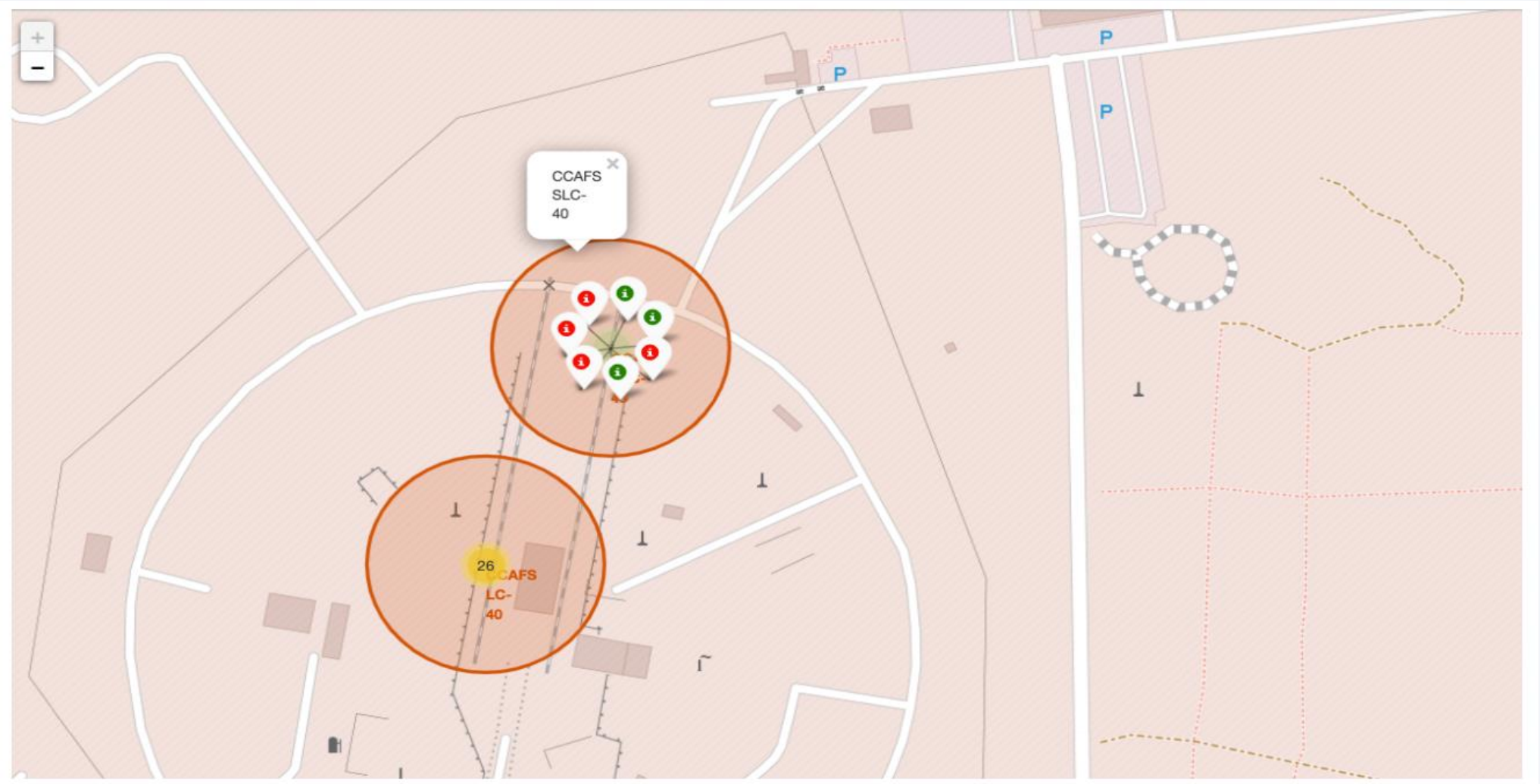
Section 3

Launch Sites Proximities Analysis

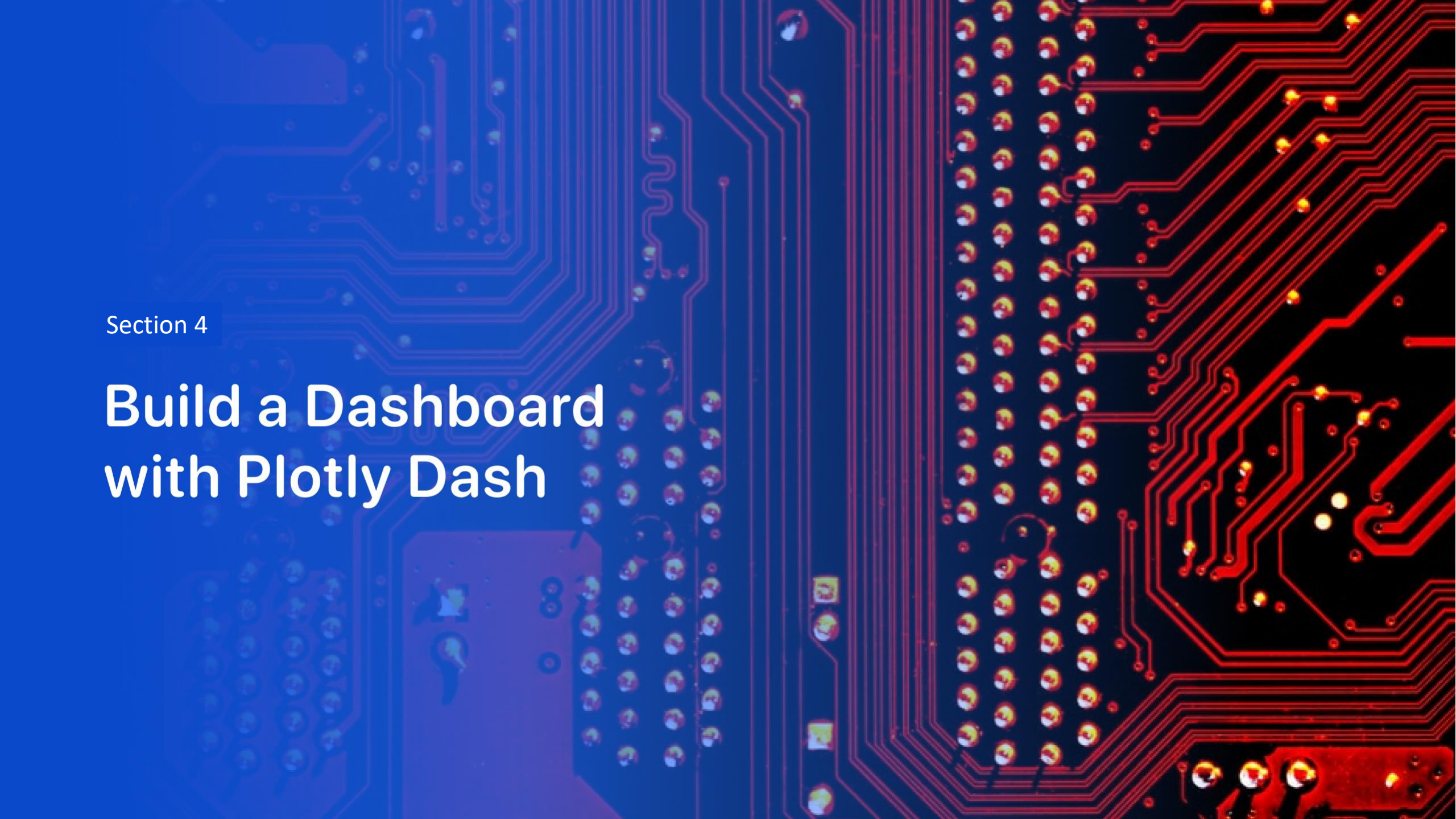
Interactive Maps



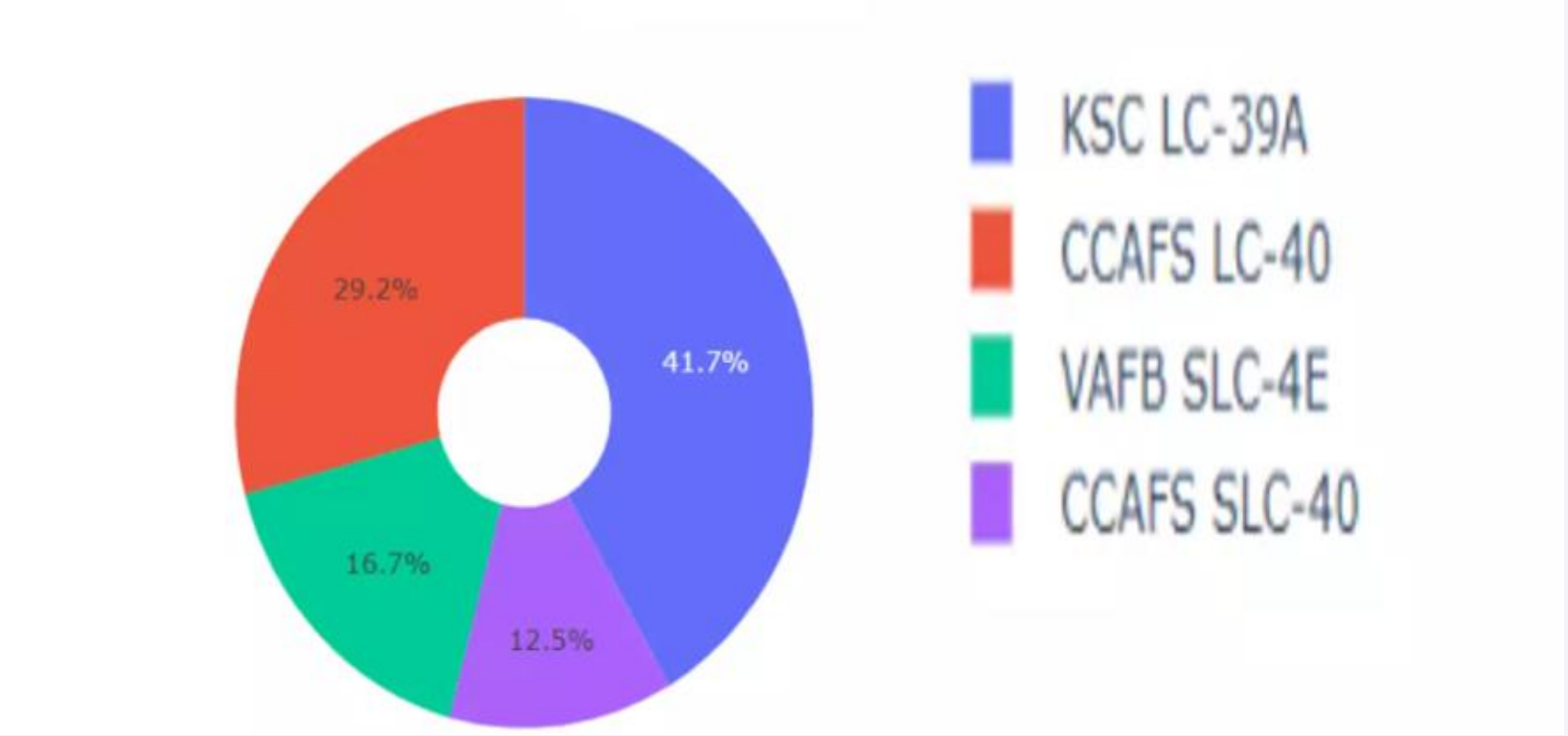
Labeled launch outcomes on the map



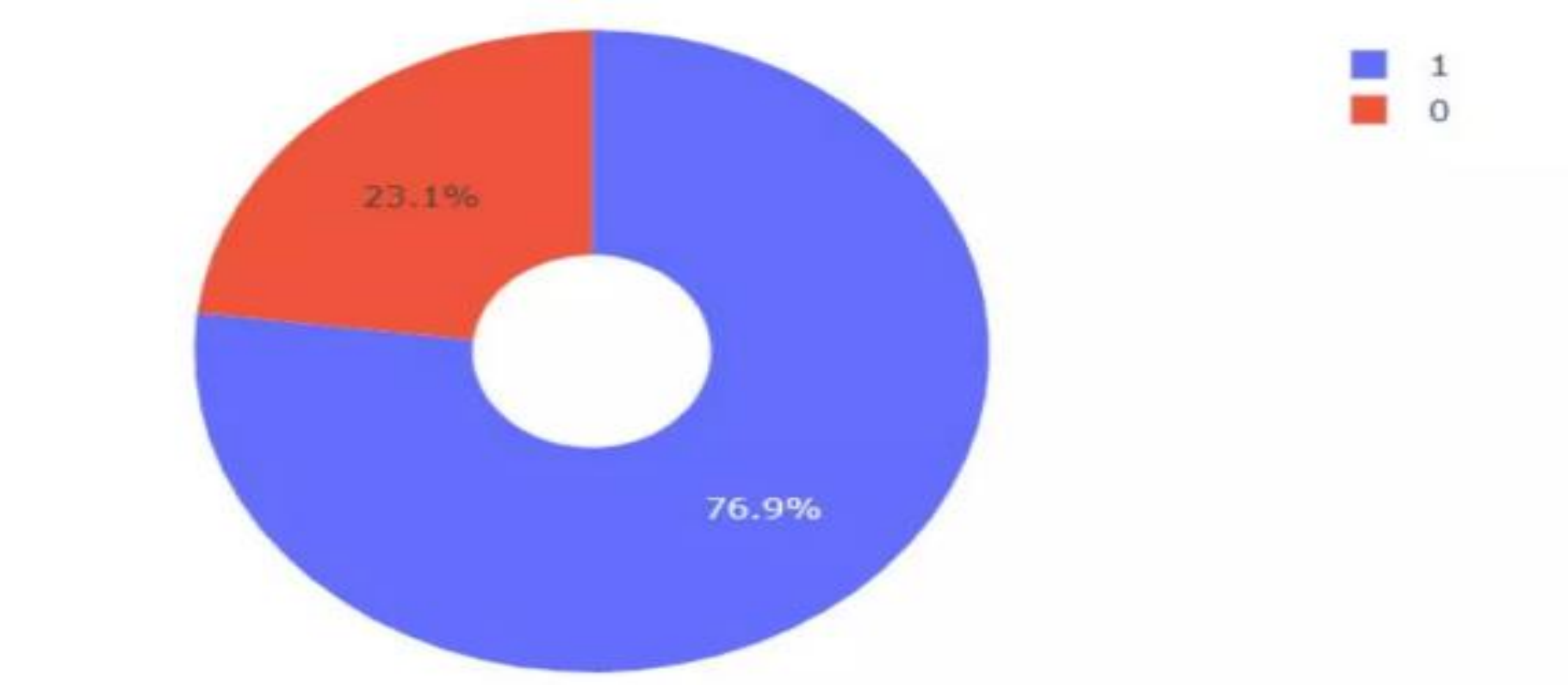
Launch site proximities


Section 4
Build a Dashboard with Plotly Dash

Success rate across launch sites



Launch site with highest success rate



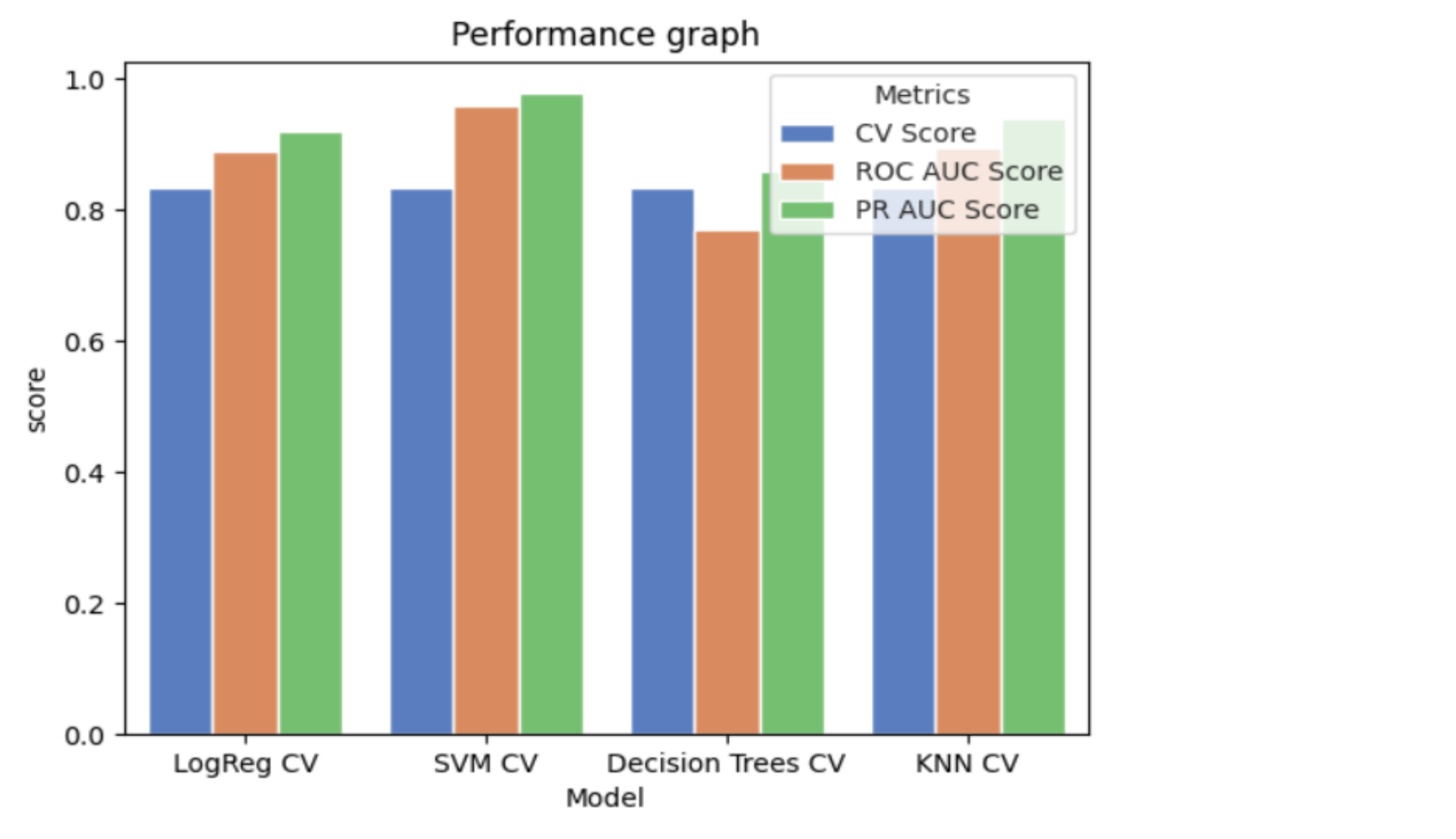
KSC LC 39A has the highest success rate of 76.9%

Payload vs Launch Outcome

Section 5

Predictive Analysis (Classification)

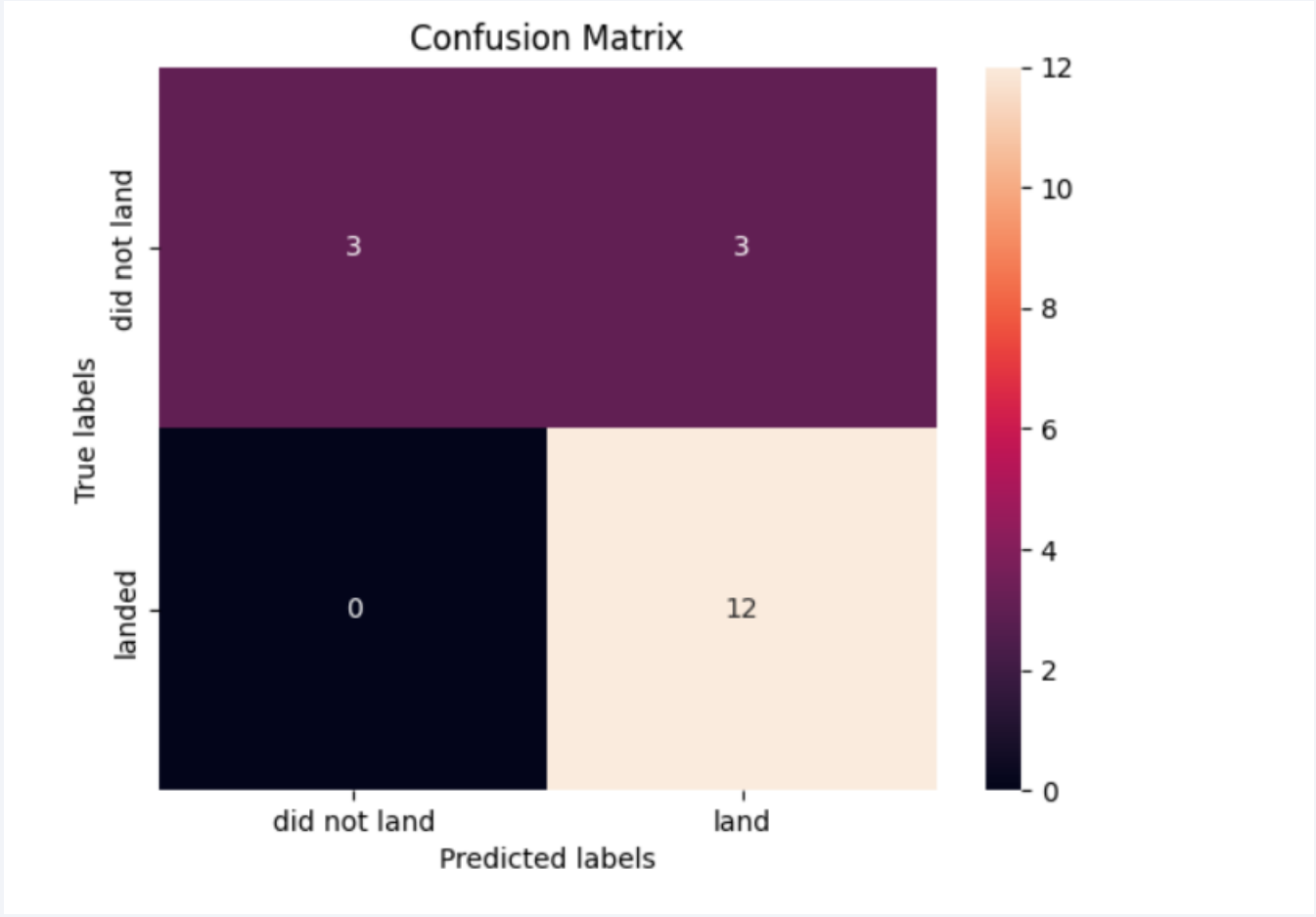
Classification Accuracy



Confusion Matrix

```
print("Accuracy score of svm cv:",svm_accuracy_score )
print("ROC AUC score for lgr cv:",svm_roc_auc)
print("PR AUC:",svm_pr_auc)
```

```
Accuracy score of svm cv: 0.8333333333333334
ROC AUC score for lgr cv: 0.9583333333333334
PR AUC: 0.9781080031080033
```



Conclusions

Model usage:

- Best model to predict is SVC, has ROC AUC and PR AUC score of ~96% and 98~ respectively. Optimal parameters to be used with it are as follows:
    - 'C' 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'

Recommendations:

- The current data is imbalanced as there are only few orbit types/launch sites with high payload launches. More experiments can help predict which orbit types/launch sites are best for what kgs of payload mass launches

Future Scope

- The data we obtained and trained the model on is quite low, better data collection can give a clearer picture of success outcome

Thank you!