IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Serra Işık
16/07/23

Outline

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

Executive Summary

1. **Data Collection and Processing**: The first step would be to gather data about SpaceX's launches. Information such as mission type, payload weight, orbit type, customer, launch outcomes, and whether the first stage was reused would be critical. This data can be obtained from various public sources like SpaceX's website, press releases, mission reports, and third-party databases. After obtaining the data, cleaning and preprocessing will be conducted to ensure it's ready for analysis and modeling.
2. **Exploratory Data Analysis (EDA)**: This step will involve understanding the distributions, relationships, and patterns in the data. This will be achieved using statistical analysis and data visualization techniques.
3. **Cost Analysis**: Utilizing the gathered data and insights from EDA, you can calculate the approximate cost of each launch, considering whether the first stage was successfully reused.
4. **Predictive Modeling**: You will employ machine learning algorithms (for example, decision trees, random forest, logistic regression, or neural networks) to predict whether SpaceX will attempt to reuse the first stage for a given launch. The model will be trained using a portion of the collected data, then validated and tested using other portions. Features for this model can include mission type, payload weight, orbit type, and customer.
5. **Evaluation and Improvement**: The predictive model's performance will be evaluated using appropriate metrics (like accuracy, precision, recall, F1 score). If performance is lacking, you may need to revisit the model and make improvements, which could include feature engineering, hyperparameter tuning, or trying a different algorithm.

Introduction

- The context for this project is the emerging commercial space industry, in which companies like SpaceX are driving down costs and increasing access to space through the use of innovations like reusable rocket stages. By better understanding these cost structures and predicting SpaceX's behavior, your fictional company, Space Y, aims to better position itself within this industry and compete more effectively against established players.

1. **Cost Analysis**: Establish a detailed cost profile for each SpaceX launch. This will be done by considering the various factors that could influence the cost, such as the type of mission, the payload, the target orbit, the customer, and notably, whether or not the first stage is successfully recovered and reused.
2. **Reuse Prediction**: Develop a predictive model that can forecast whether or not SpaceX will attempt to reuse the first stage for a given launch. This prediction could be based on various factors, such as the specifics of the mission and payload, as well as past data on SpaceX's reuse efforts.

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

Data collected from SpaceX API, and cleanedto only contain Falcon 9 flights.

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

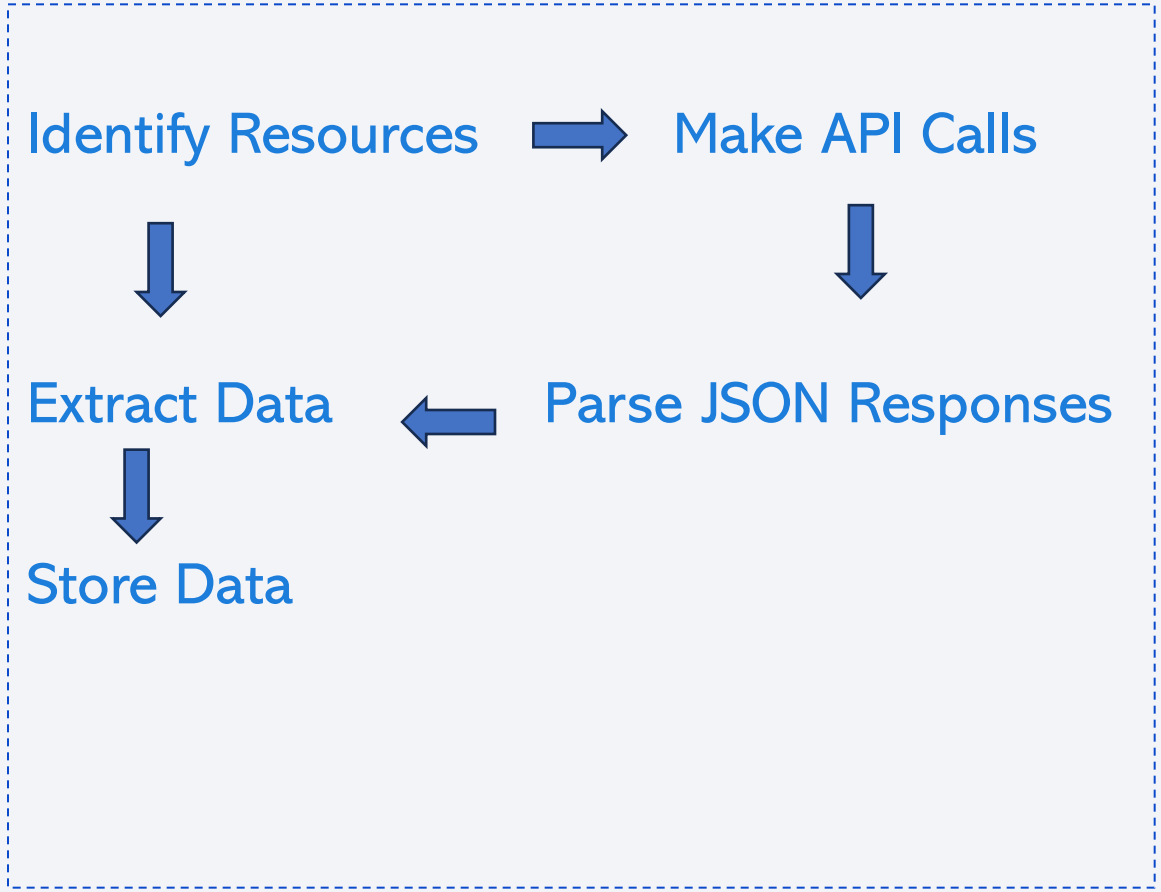
```
response = requests.get(spacex_url)
```

```
data.head()
```

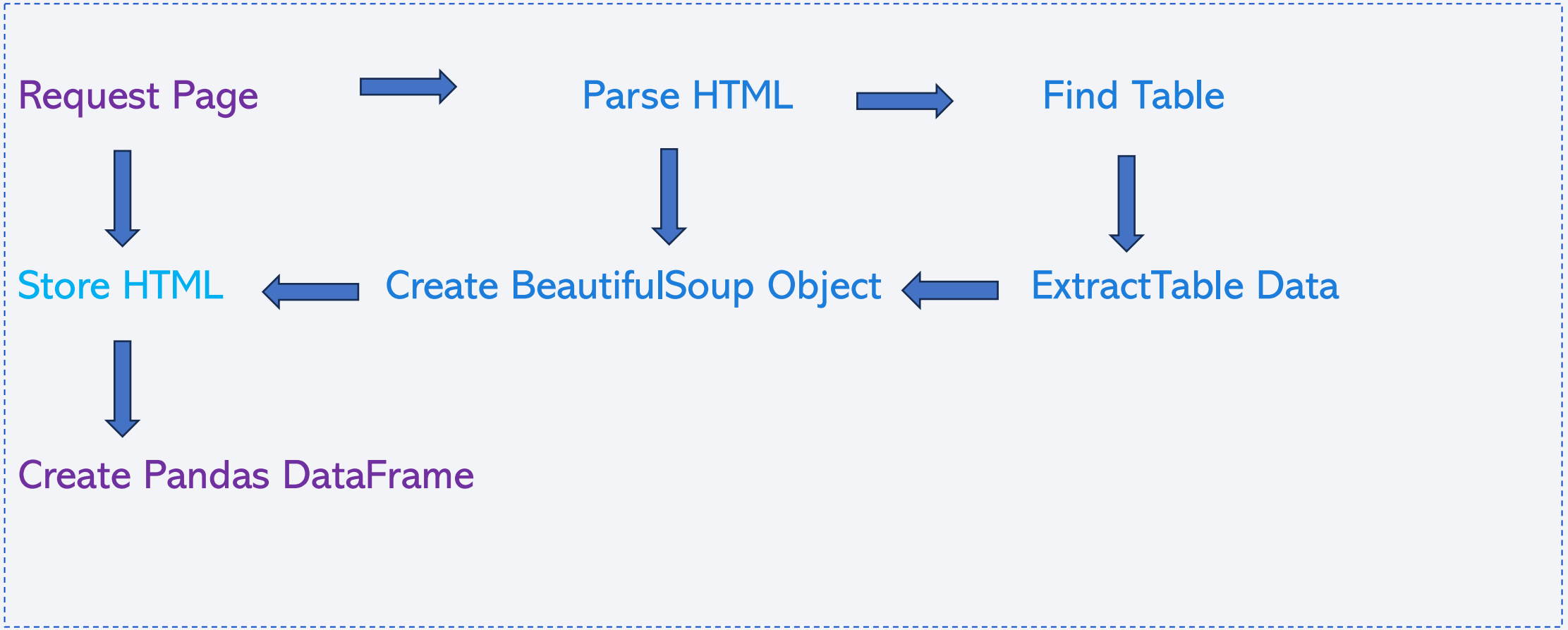
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPa
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	Nor
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	Nor
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	Nor
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	Nor
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	Nor

Data Collection – SpaceX API

- https://github.com/serraisik5/IBM_datascience/blob/0f599253317dcf392d14ecfef5096b91e883c3e8/AppliedDataScience/spaceX_API.ipynb



Data Collection - Scraping



https://github.com/serraisik5/IBM_datascience/blob/0f599253317dcf392d14ecfef5096b91e883c3e8/AppliedDataScience/SpaceX_WebScrapping.ipynb

Data Wrangling

- I wanted to create a binary classification model that predicts whether a booster will land successfully but I need to convert outcomes into training labels:

1. **Define Success Criteria**: landing can occur in three ways: ocean landing (True Ocean), ground pad landing (True RTLS), or drone ship landing (True ASDS).
2. **Create Labels**: create a new variable in your data that is 1 if the landing was successful (in any of the three ways), and 0 otherwise.

- Using the Outcome, create a list where the element is zero if the corresponding row in Outcome is in the set bad_outcome; otherwise, it's one. Then assign it to the variable landing_class.

```
df.head(5)
```

LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0

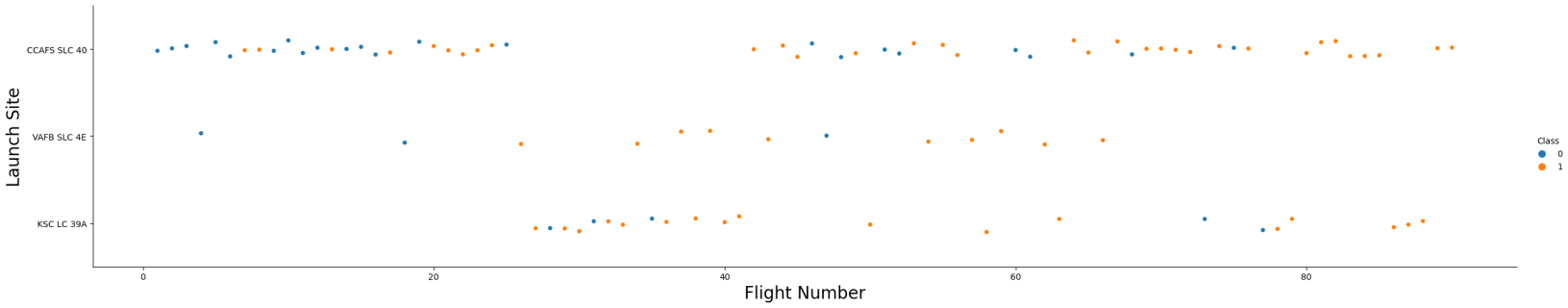
Perform EDA

Visualize Data

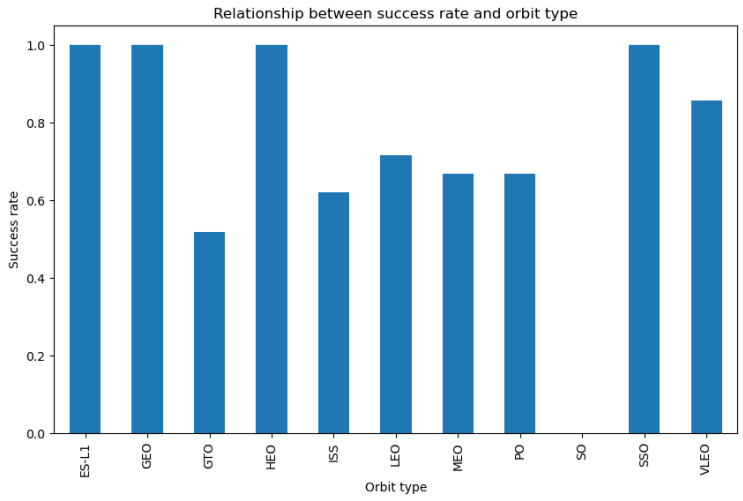
Determine Relationships

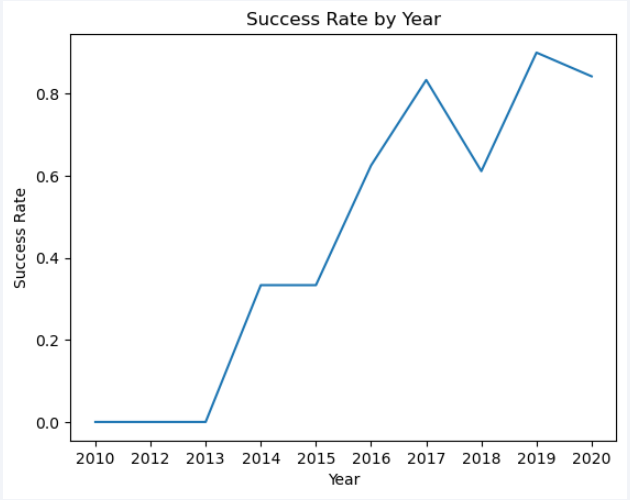
EDA with Data Visualization

- To see how the FlightNumber (indicating the continuous launch attempts.) and LauncSite variables would affect the launch outcome.



- To check if there are any relationship between success rate and orbit type.





https://github.com/serraisik5/IBM_datascience/blob/0f599253317dcf392d14ecfef5096b91e883c3e8/AppliedDataScience/SpaceX_EDAwithVisualization.ipynb

EDA with SQL

https://github.com/serraisik5/IBM_datascience/blob/0f599253317dcf392d14ecfef5096b91e883c3e8/AppliedDataScience/SpaceX_SQL.ipynb

- select DISTINCT "Launch_Site" from SPACEXTBL
- select * from SPACEXTBL where "Launch_Site" like "CCA%" limit 5
- select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where "Customer" like "NASA (CRS)"
- select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where "Booster_Version" like "F9 v1.1%"
- select min("Date"), "Landing_Outcome" from SPACEXTBL where "Landing_Outcome" like "Success%"
- select "Booster_Version", "Landing_Outcome", PAYLOAD_MASS__KG_ from SPACEXTBL where "Landing_Outcome" like "Success (drone ship)%"

and PAYLOAD_MASS__KG_< 6000 and PAYLOAD_MASS__KG_ >4000

- SELECT

  SUM(CASE WHEN "Mission_Outcome" LIKE 'success%' THEN 1 ELSE 0 END) AS Successful_Missions,

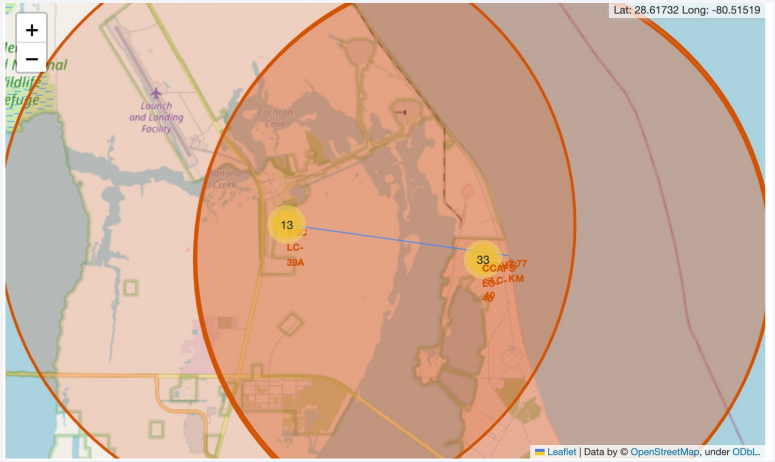
  SUM(CASE WHEN "Mission_Outcome" LIKE 'failure%' THEN 1 ELSE 0 END) AS Failed_Missions FROM SPACEXTBL;

- select "Booster_Version", PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
- SELECT substr(Date, 4, 2) as Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTBL WHERE substr(Date, 7, 4) = '2015' AND "Landing_Outcome" = 'Failure (drone ship)'
- SELECT "Landing_Outcome", COUNT(*) as Count FROM SPACEXTBL where (SUBSTR(Date, 7, 4) || SUBSTR(Date, 1, 2) || SUBSTR(Date, 4, 2)) BETWEEN '20100604' AND '20170320' GROUP BY "Landing_Outcome" ORDER BY Count DESC;

Build an Interactive Map with Folium

- I used folium.Circle to add a highlighted circle area with a text label on a specific coordinate.
- I used marker to mark the location and give information about it.
- I used Marker clusters to simplify the map containing many markers having the same coordinate.
- I used mousePosition to get the coordinates from map.
- I used PolyLine to draw line between coordinates

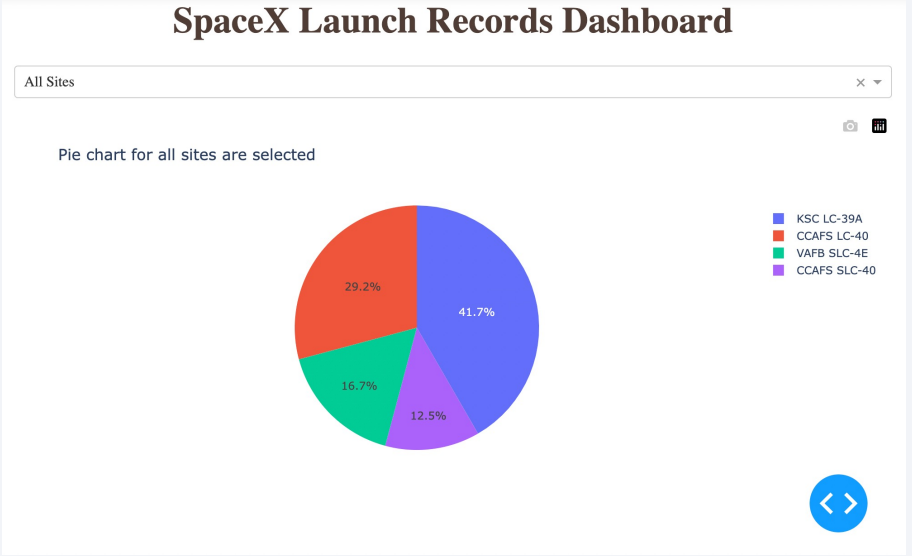
https://github.com/serraisik5/IBM_datascience/blob/0f599253317dcf392d14ecfef5096b91e883c3e8/AppliedDataScience/SpaceX_Dashboard_Map.ipynb



Build a Dashboard with Plotly Dash

- A Launch Site Drop-down Input Component to be able to select site
- A callback function to render success-pie-chart based on selected site dropdown
- A Range Slider to Select Payload
- A callback function to render the success-payload-scatter-chart scatter plot

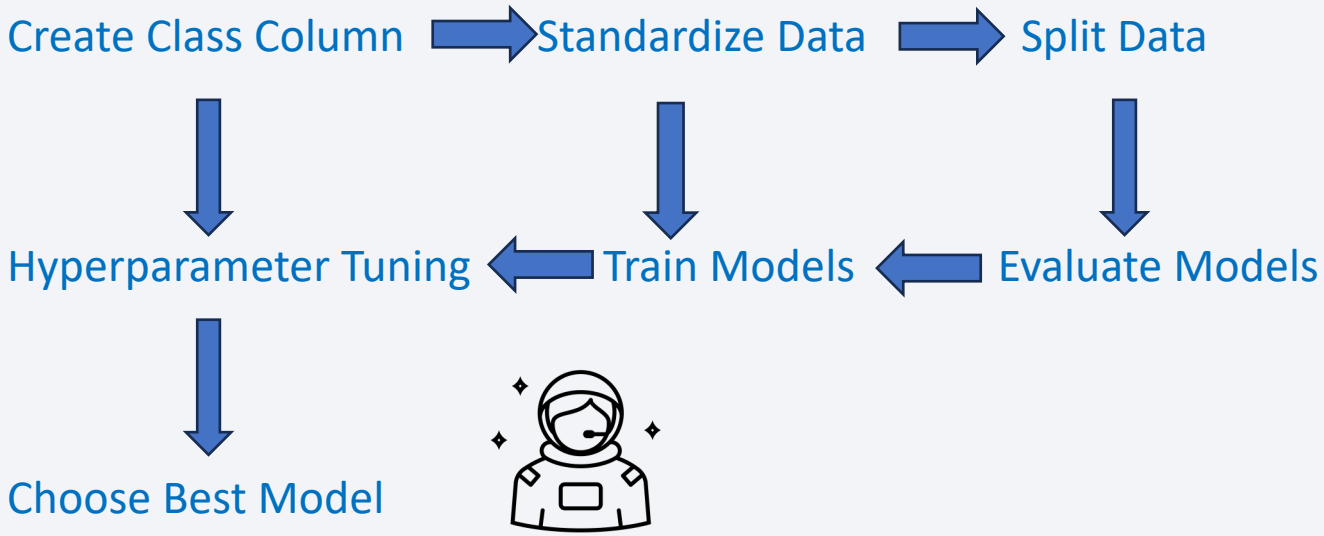
https://github.com/serraisik5/IBM_datascience/blob/0f599253317dcf392d14ecfef5096b91e883c3e8/AppliedDataScience/Spacex_dash.ipynb



Predictive Analysis (Classification)

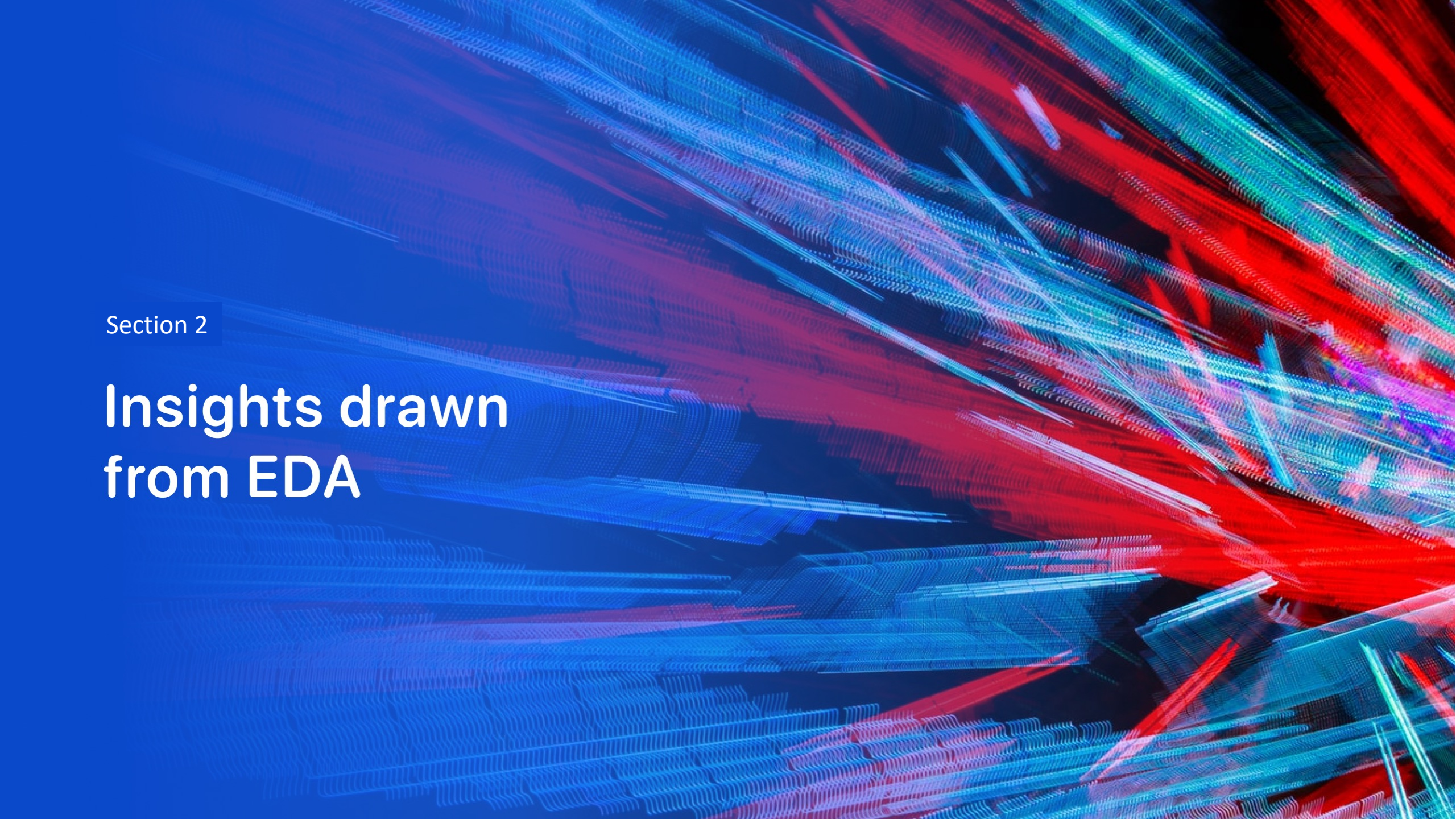
- Create a column for the class
- Standardize the data
- Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Find the method performs best using test data

https://github.com/serraisik5/IBM_datascience/blob/0f599253317dcf392d14ecfef5096b91e883c3e8/AppliedDataScience/SpaceX_ML.ipynb

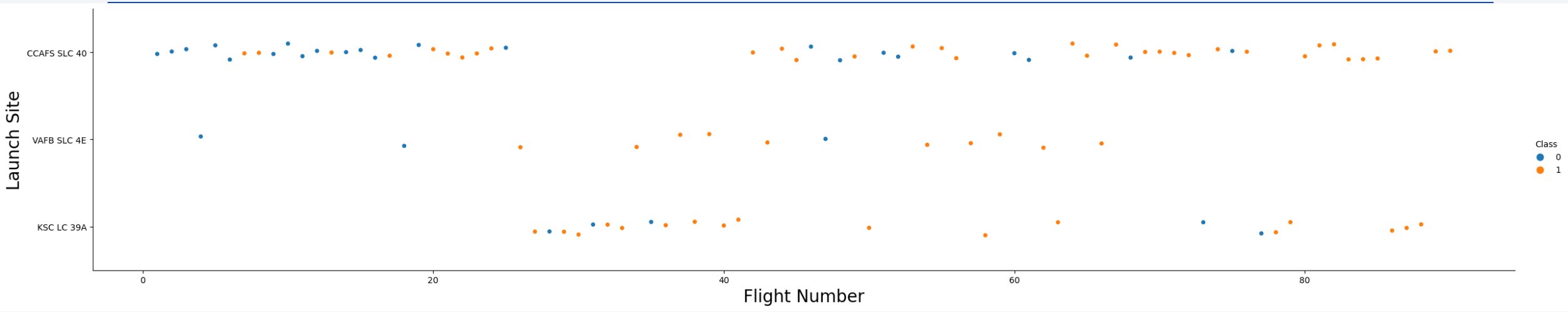


Results

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

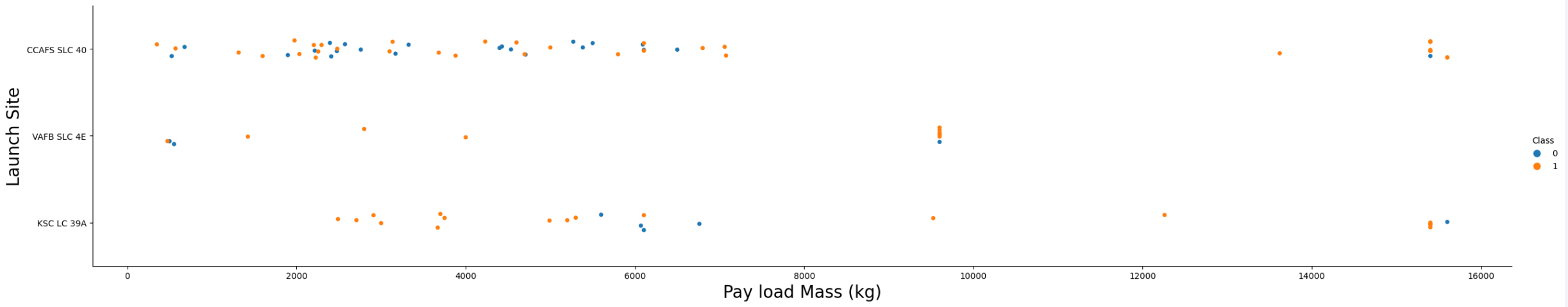

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- We observe same launch site usage in consecutive flights.
- We see that different launch sites have different success rates.

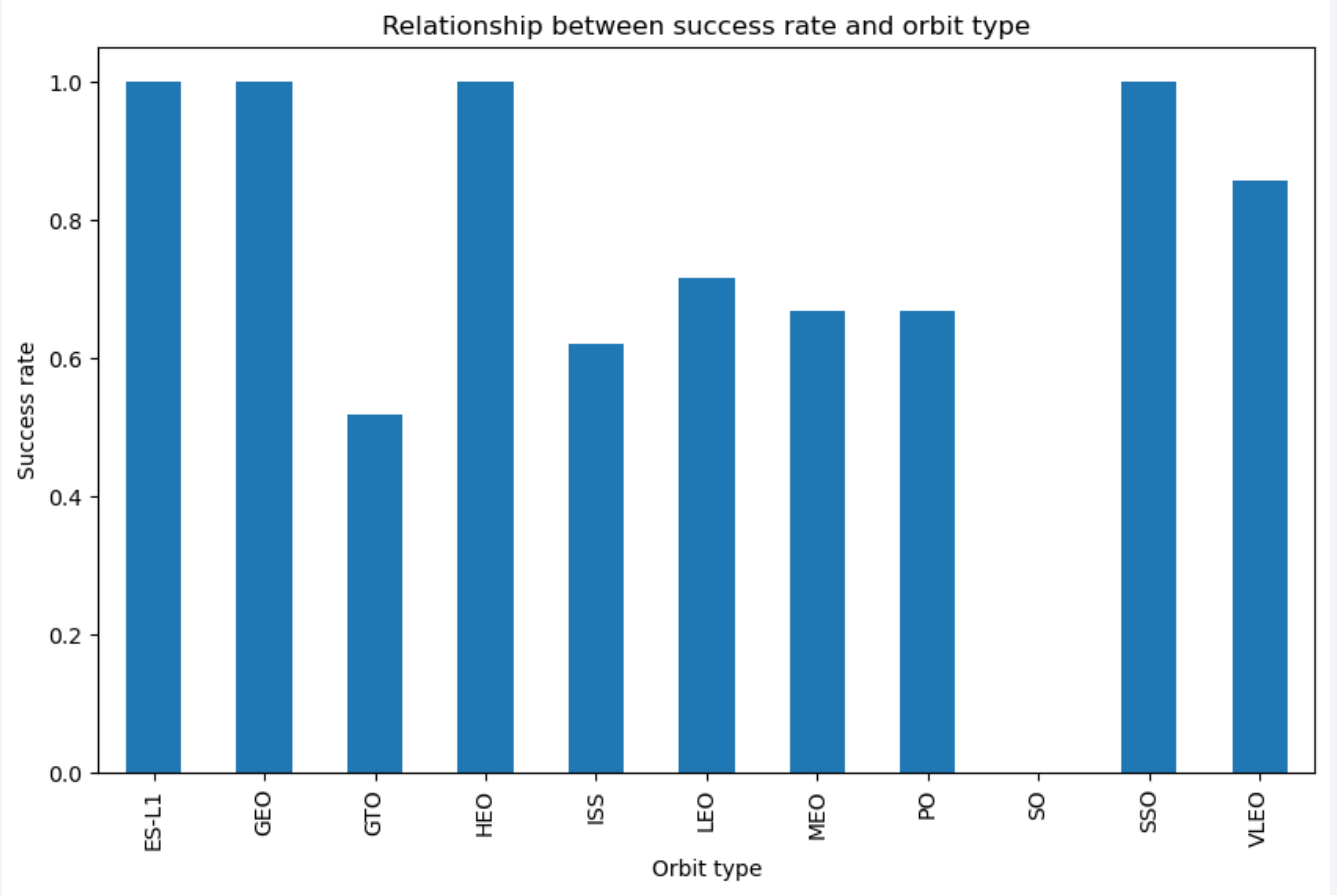
Payload vs. Launch Site



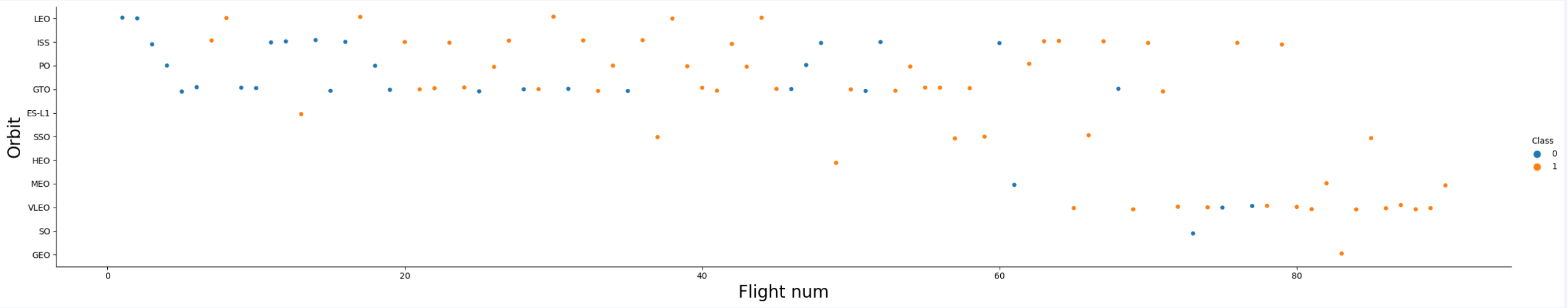
- CCAFS LC-40 has more flights than other launch sites.
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- We see a jump in mass after 8000 kg

Success Rate vs. Orbit Type

- SO has 0 success rate.
- ES-L1 GEO HEO SSO has very high success rate.

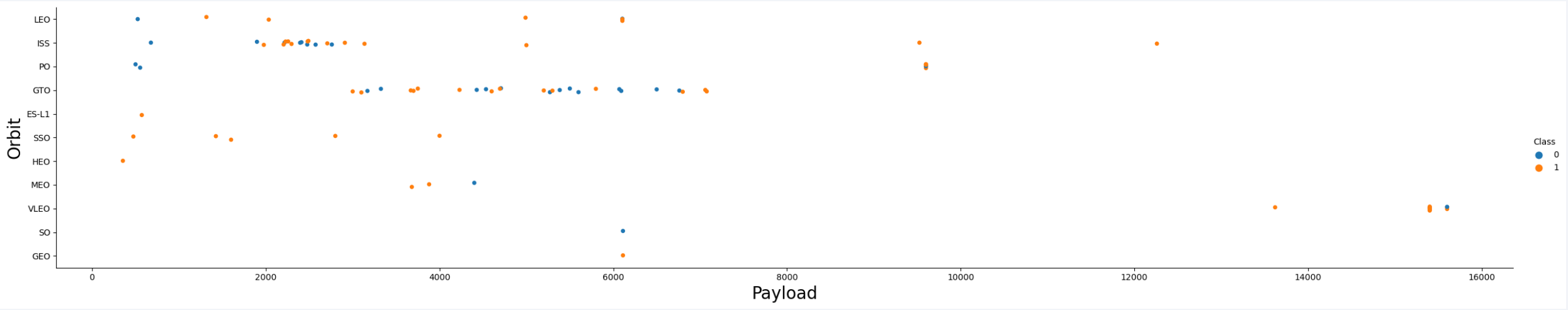


Flight Number vs. Orbit Type



- It 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.
- Each orbit has different number of flights

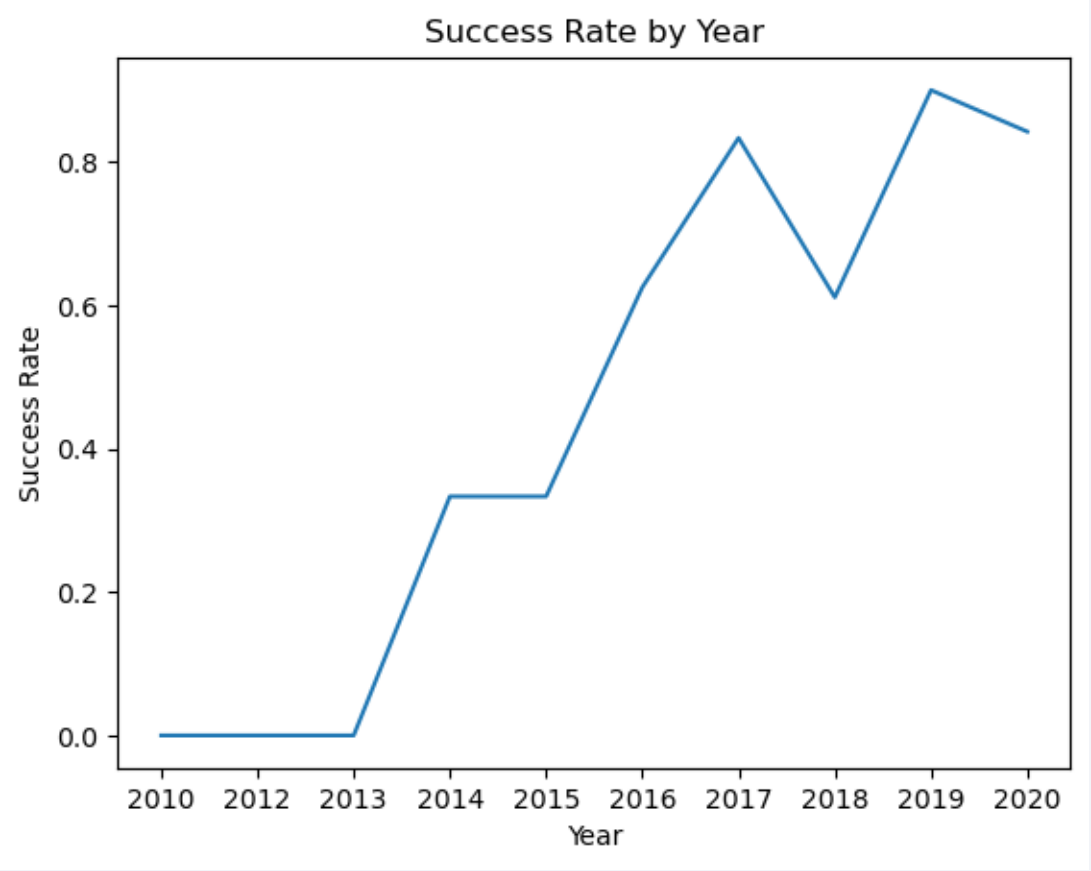
Payload vs. Orbit Type



- 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

- We observe that the sucess rate since 2013 kept increasing till 2020.
- We observe little decrease in 2017.



All Launch Site Names

- There area 4 different launch sites

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

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string '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	Custom
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	Space
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NAS (COT NR
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NAS (COT
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NAS (CR
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NAS (CR

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where "Customer" like "NASA (CRS)"
```

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

sum(PAYLOAD_MASS__KG_)
45596.0

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

First Successful Ground Landing Date

```
%sql select min("Date"), "Landing_Outcome" from SPACEXTBL where "Landing_Outcome" like "S
```

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

min("Date")	Landing_Outcome
01/07/2020	Success

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql select "Booster_Version", "Landing_Outcome", PAYLOAD_MASS__KG_ from SPACEXTBL where "
and PAYLOAD_MASS__KG_< 6000 and PAYLOAD_MASS__KG_ >4000
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696.0
F9 FT B1026	Success (drone ship)	4600.0
F9 FT B1021.2	Success (drone ship)	5300.0
F9 FT B1031.2	Success (drone ship)	5200.0

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT
  CASE
    WHEN "Mission_Outcome" LIKE '%Success%' THEN 'Success'
    WHEN "Mission_Outcome" LIKE '%Failure%' THEN 'Failure'
    ELSE 'Other'
  END AS Outcome,
  COUNT(*) AS Count
FROM SPACEXTBL
WHERE "Mission_Outcome" LIKE '%Success%' OR "Mission_Outcome" LIKE '%Failure%'
GROUP BY Outcome;
```

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

Outcome	Count
Failure	1
Success	100

Boosters Carried Maximum Payload

```
%sql select "Booster_Version", PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ =
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

```
sql SELECT substr(Date, 4, 2) as Month, "Landing_Outcome", "Booster_Version", "Launch_Site"
OM SPACEXTBL
ERE substr(Date, 7, 4) = '2015' AND "Landing_Outcome" = 'Failure (drone ship)'
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
10	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

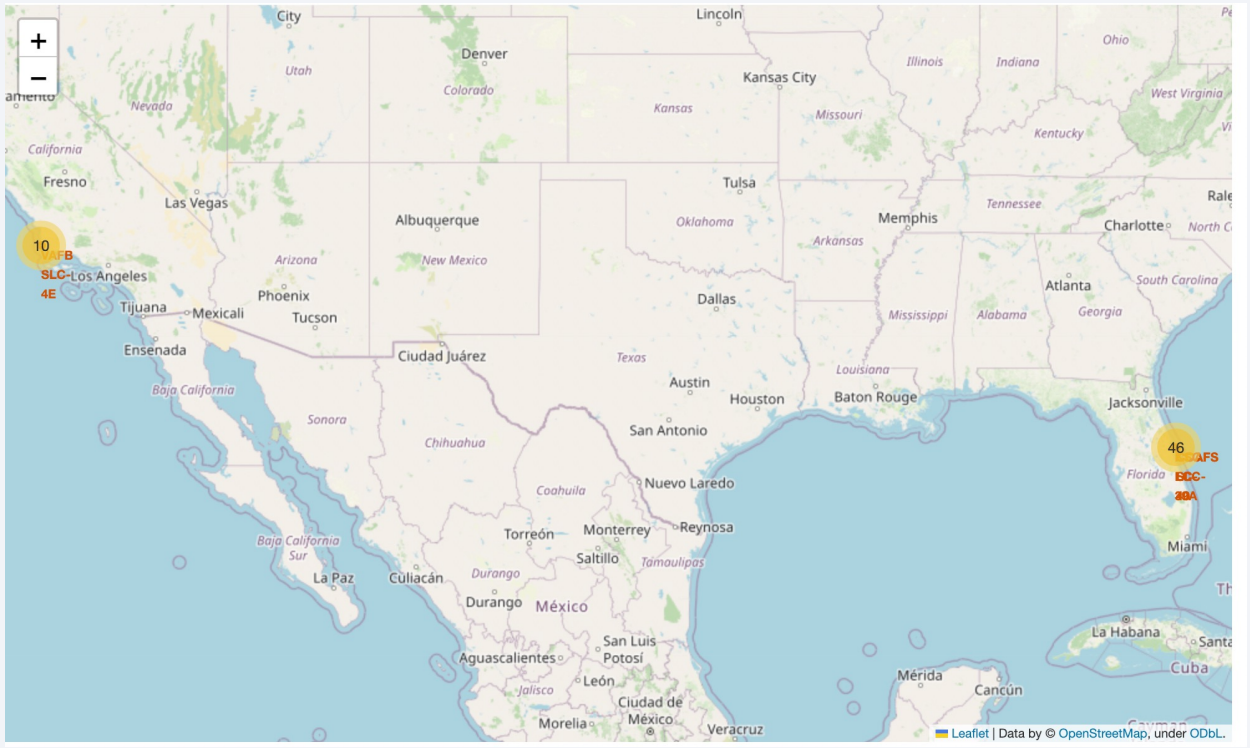
```
%%sql SELECT "Landing_Outcome", COUNT(*) as Count
FROM SPACEXTBL
where (SUBSTR(Date, 7, 4) || SUBSTR(Date, 1, 2) || SUBSTR(Date, 4, 2))
BETWEEN '20100604' AND '20170320'
GROUP BY "Landing_Outcome"
ORDER BY Count DESC;
```

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

Landing_Outcome	Count
No attempt	9
Failure (drone ship)	5
Success (drone ship)	4
Controlled (ocean)	3
Uncontrolled (ocean)	2
Success (ground pad)	2
Failure (parachute)	2
Precluded (drone ship)	1

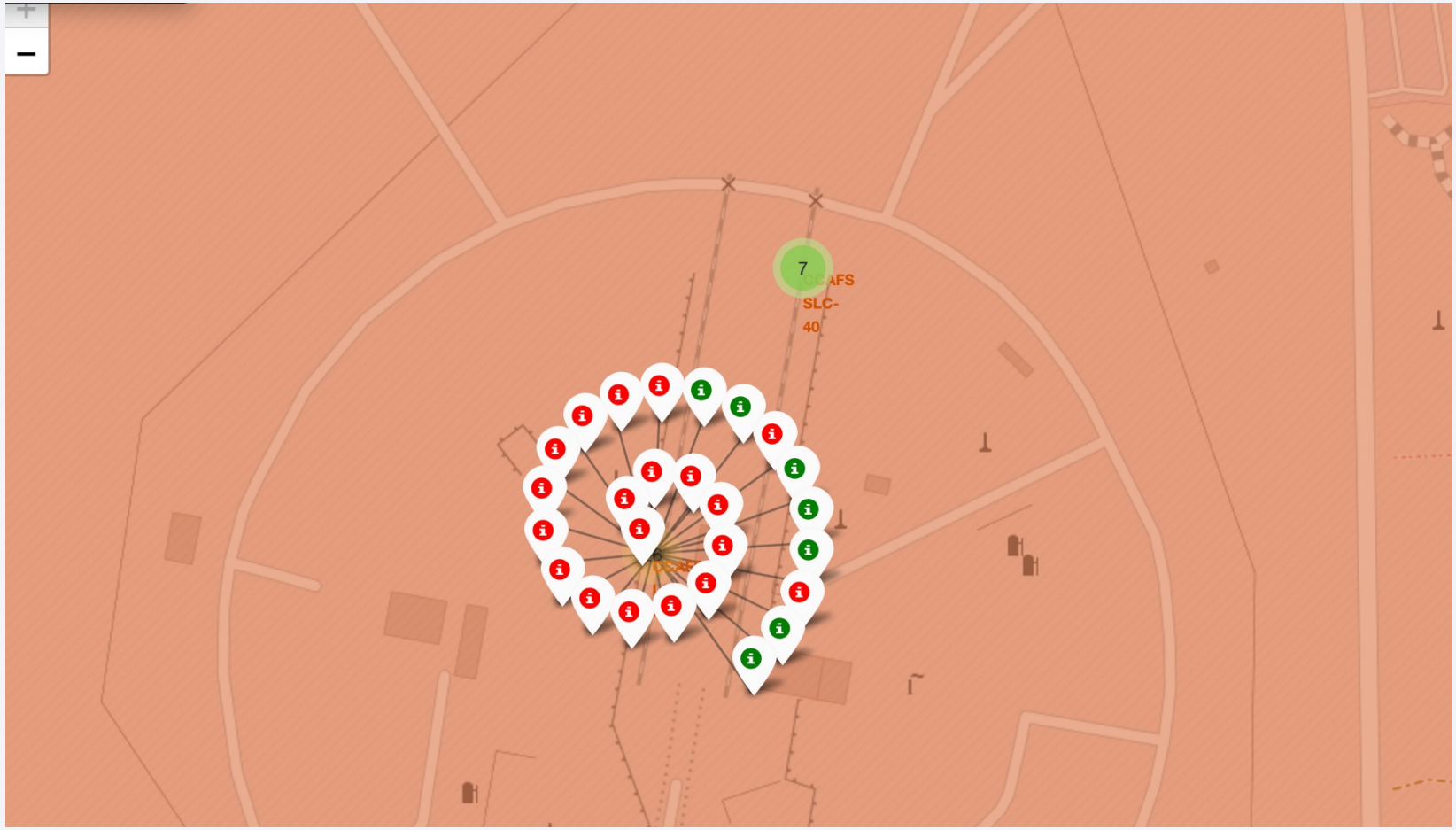

Section 3
Launch Sites
Proximities Analysis

Success/failed launches for each site on the map



Success/failed launches

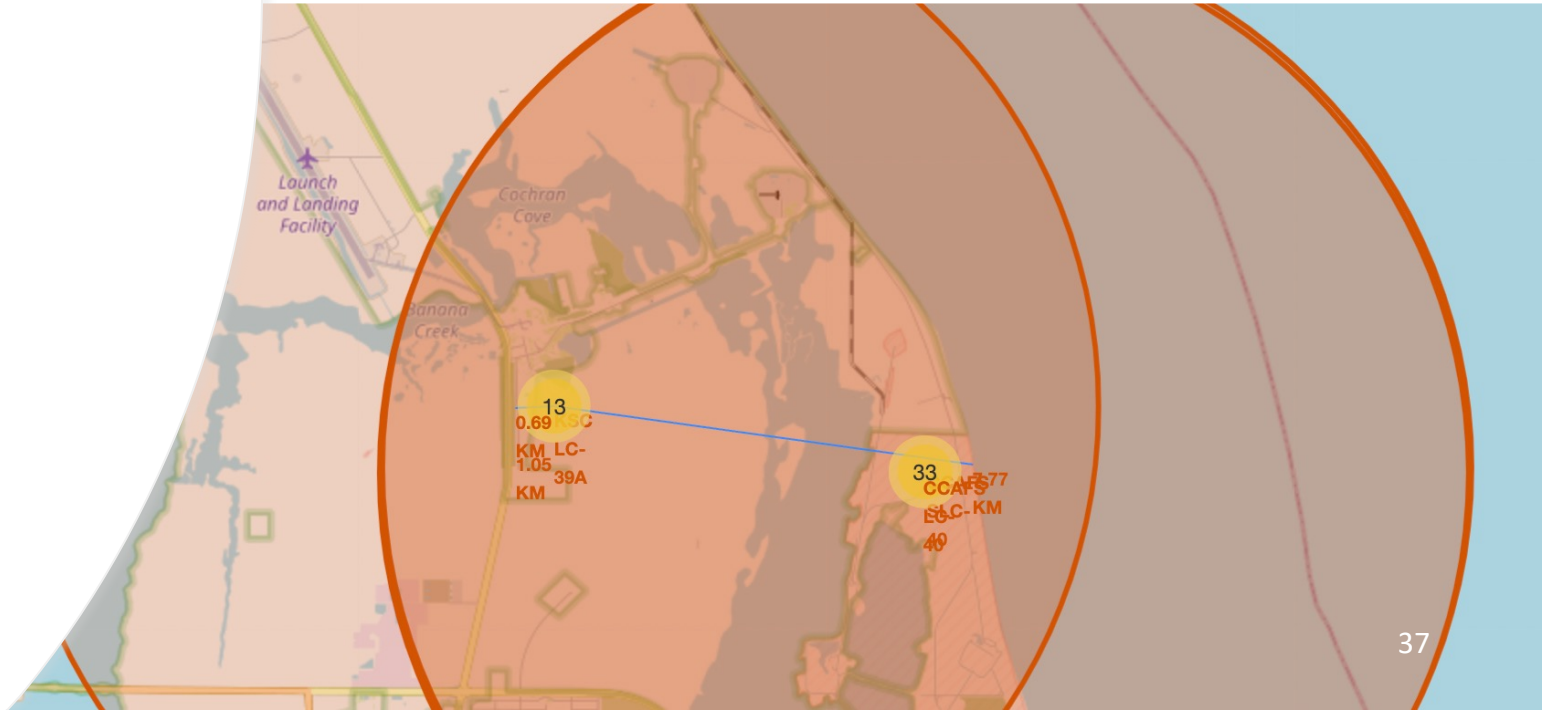
- Red locations are failures, greens are success.
- You can observe, clustered 7 launches in above point.

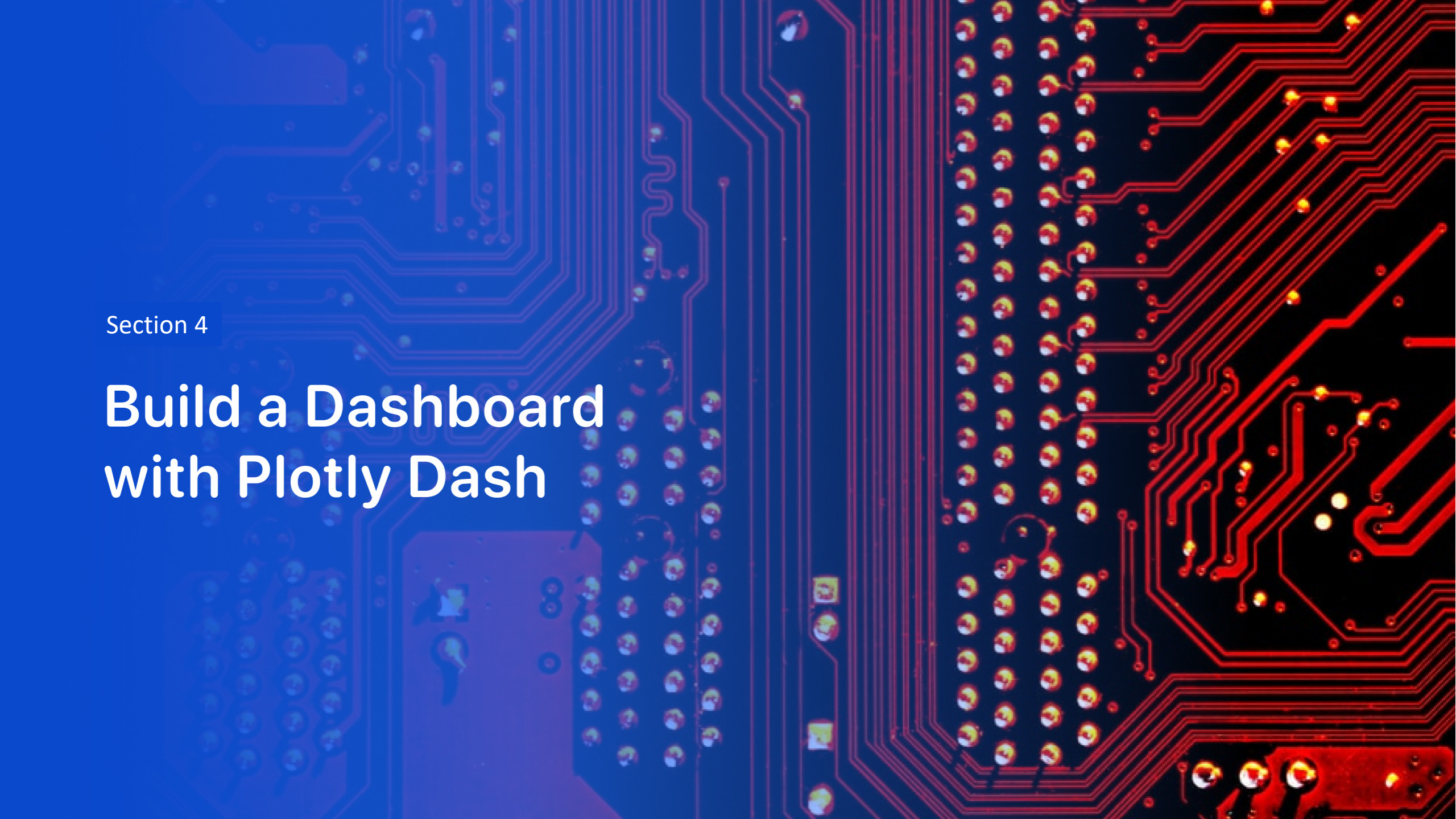


Distances between a launch sites

After you plot distance lines to the proximities, you can answer the following questions easily:

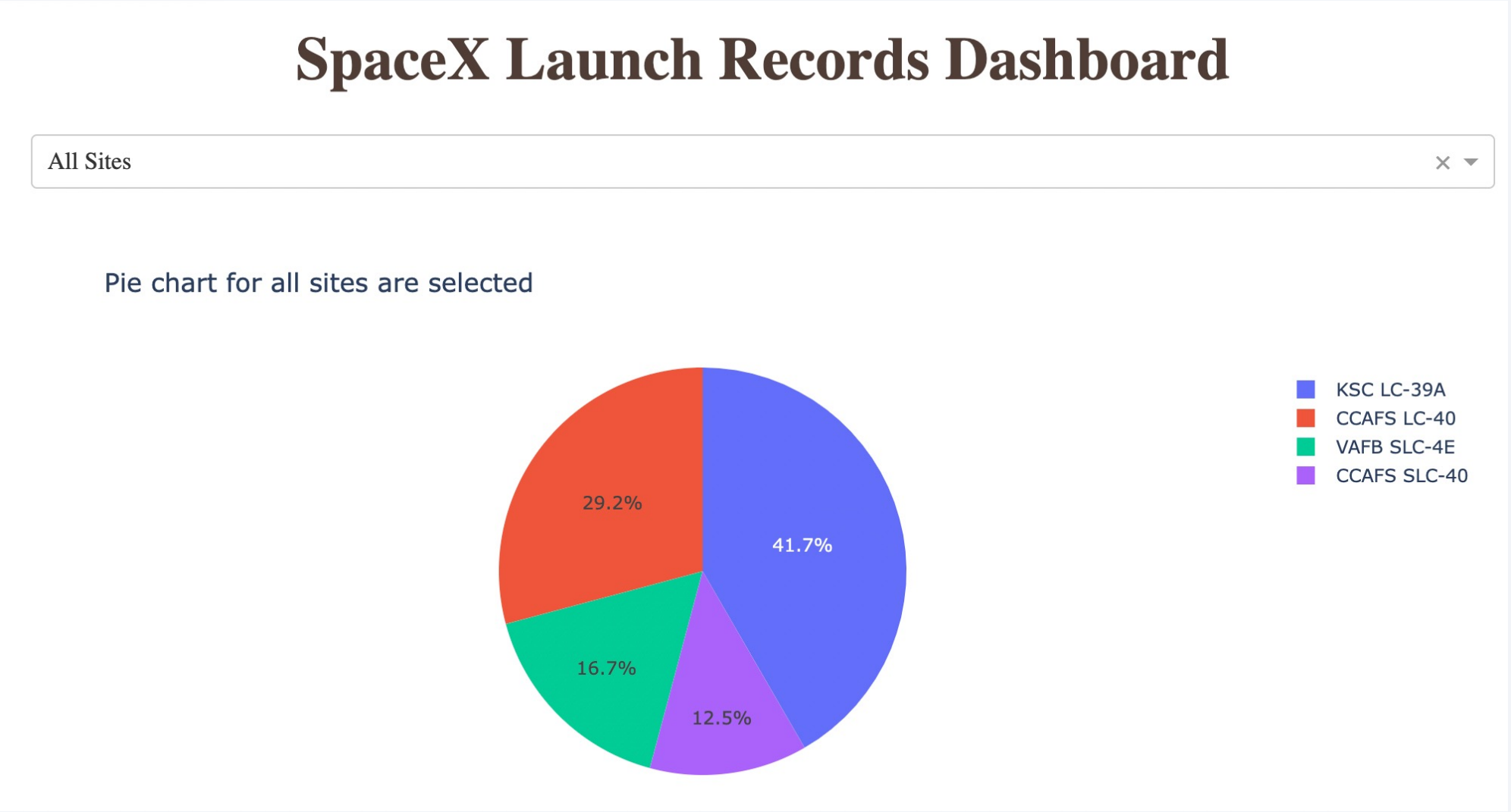
- Are launch sites in close proximity to railways?
- Are launch sites in close proximity to highways?
- Are launch sites in close proximity to coastline?
- Do launch sites keep certain distance away from cities?



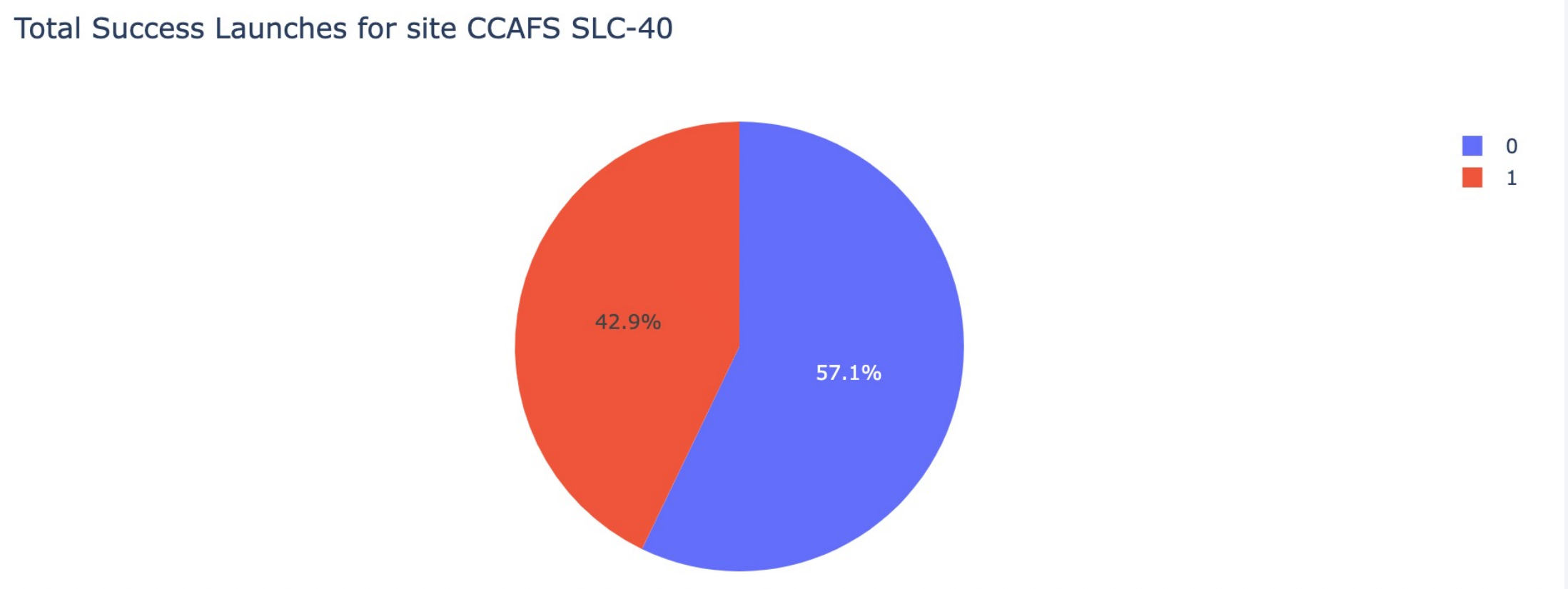

Section 4
Build a Dashboard with Plotly Dash

All Sites

- Highest number of launches from KSC with 41.7%



Highest launch success



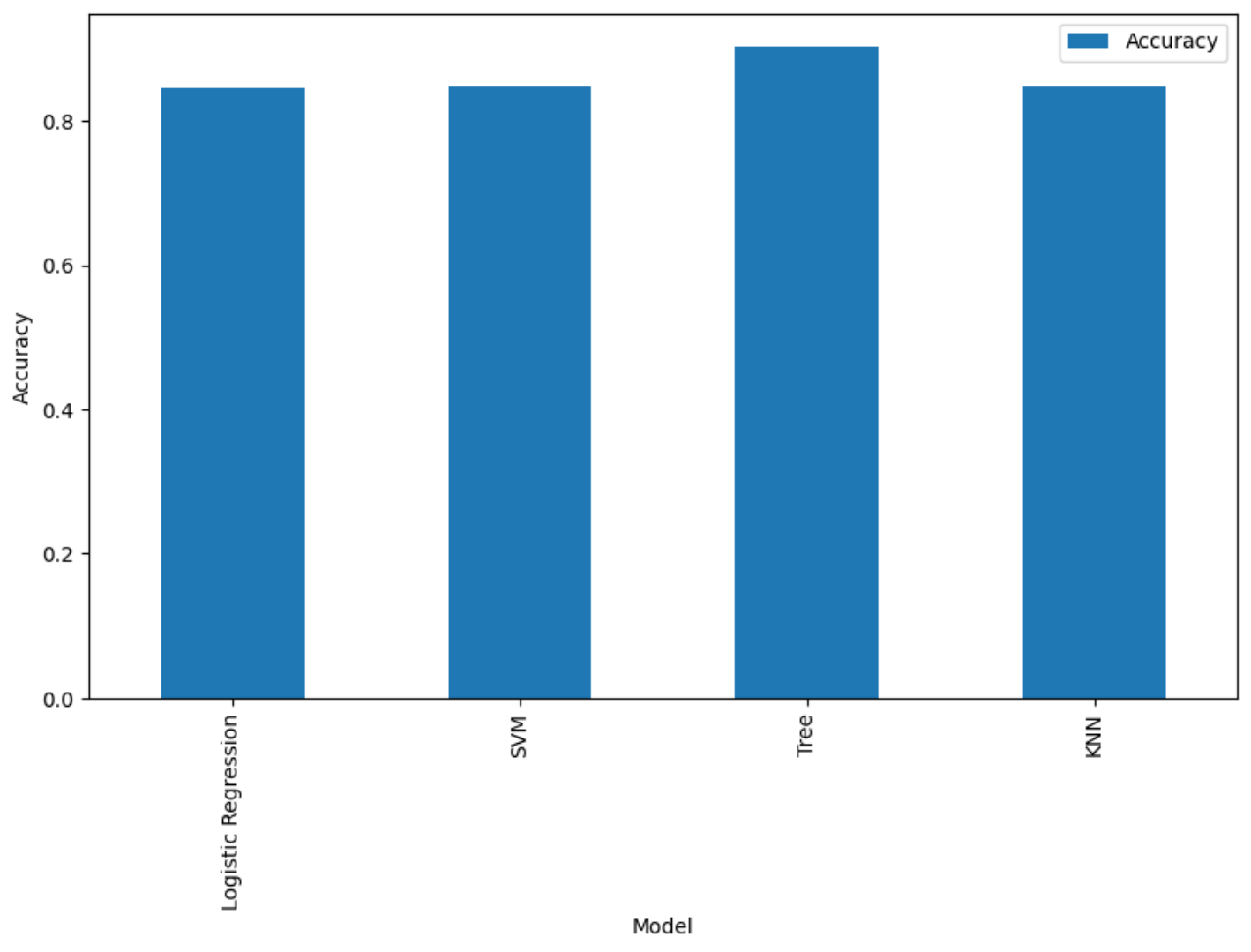
Payload vs. Launch Outcome

Section 5

Predictive Analysis (Classification)

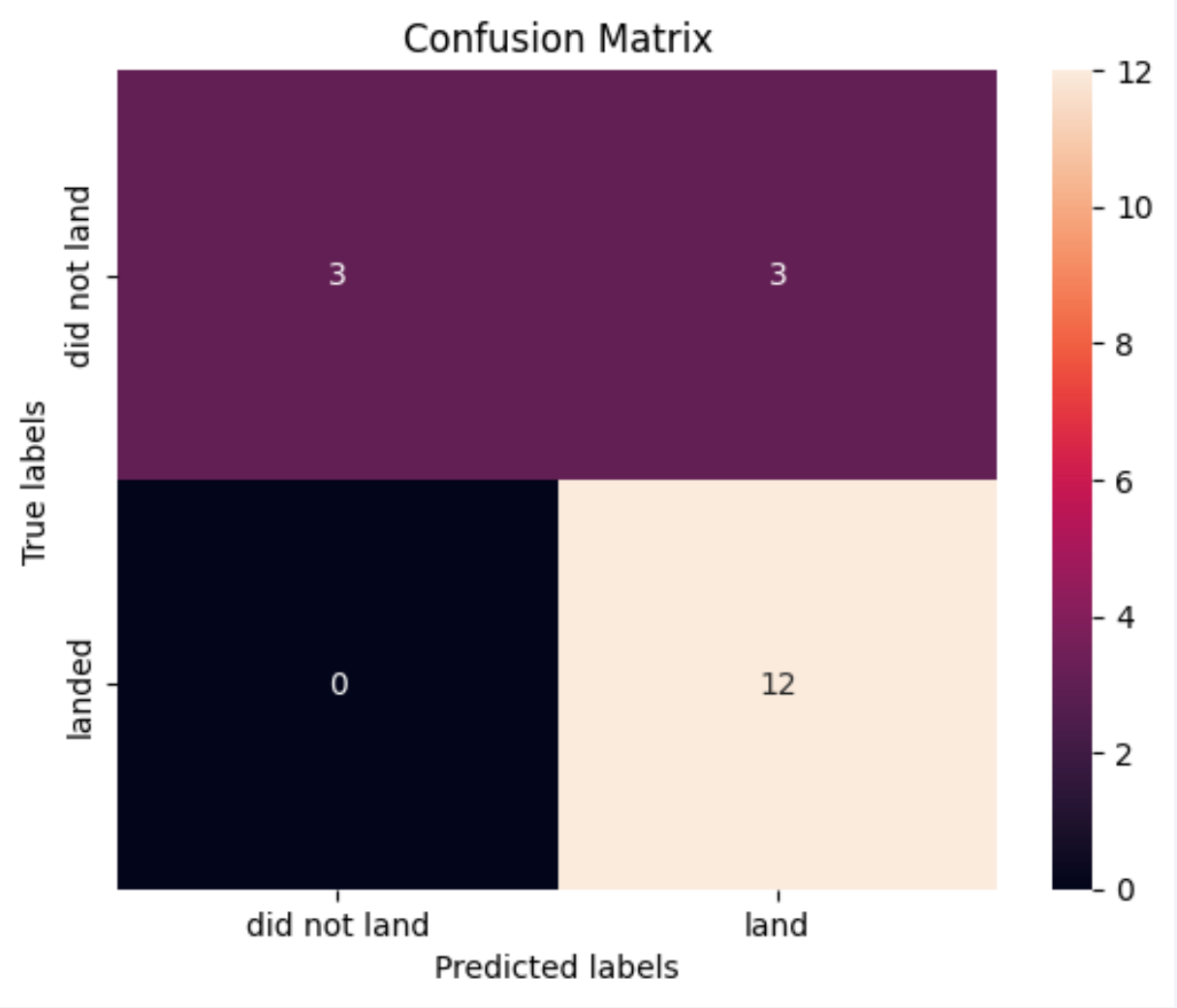
Classification Accuracy

- Decision Tree has highest model accuracy.



Confusion Matrix

- Examining the confusion matrix, we see that Tree can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- All models performed similarly on the test data, but the Decision Tree model showed a higher cross-validation score, it indicates that the Decision Tree model is more stable and generalized better to unseen data.

Appendix

- https://github.com/serraisik5/IBM_datascience/tree/0f599253317dcf392d14ecfef5096b91e883c3e8/AppliedDataScience

Thank you!