IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix
- GitHub repo here: https://github.com/ThuyThiBird/Applied-Data-Science-Capstone.git

Executive Summary

- The target of this project is to investigate the factors that affected the success of SpaceX missions, such as the payload, launch site, etc.
- SpaceX data was obtained from SpaceX public API and from SpaceX launch data in Wikipedia. The desired data was extracted and transformed for Exploratory Analysis and visualization.
- Various machine learning models were trained and tested (including logistics regression, support vector machines, decision trees, k-nearest-neighbors) in order to find the best model and the most accurate predictors of launch success
- The decision tree model was the most accurate model among all
- GitHub repo: https://github.com/ThuyThiBird/Applied-Data-Science-Capstone.git

Introduction

- In this capstone, we will predict if the Falcon 9 first stage would land successfully. SpaceX adversities F9 rocket launches on its websites with the cost of around 62 M$ while other providers cost at least 165 M$ each, which is much of the savings as SpaceX can reuse the first stage.
- To find the answers, we need to predict as close as possible if the first stage will be successful. From there we can determine the cost of a launch. This information can be used if an alternative company wants to compete with SpaceX for a rocket launch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Public API and webscraping Wikipedia
- Perform data wrangling
  - Handled Missing values as well as converted categorical variables into numeric features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize features, Split train test datasets, Use GridsearchCV to variable models to find the best models.

Data Collection – SpaceX API

Extract requested data
Requests.get(‘spacex_url’)

Decode response into a Pandas dataframe
pd.json_normalize(response.json())

Get the booster versions
getBoosterVersion(data)

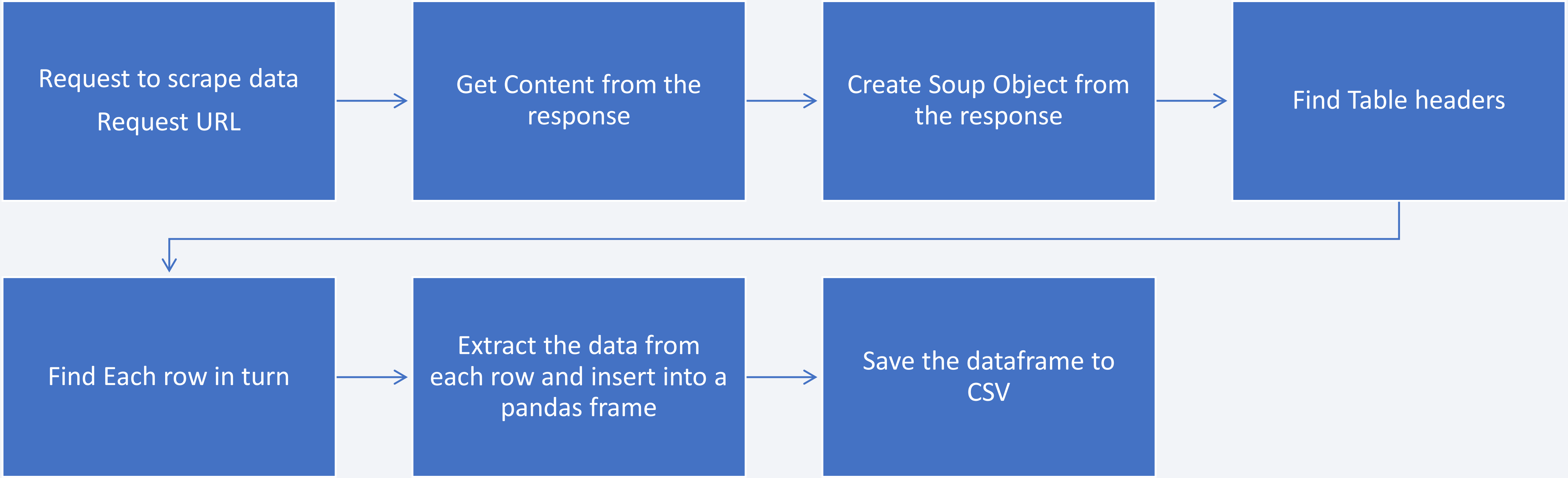
Get Launch Site data
getLaunchSite(data)

Get Payload data
getPayLoadData(data)

Get Core Data
getCoreData(data)

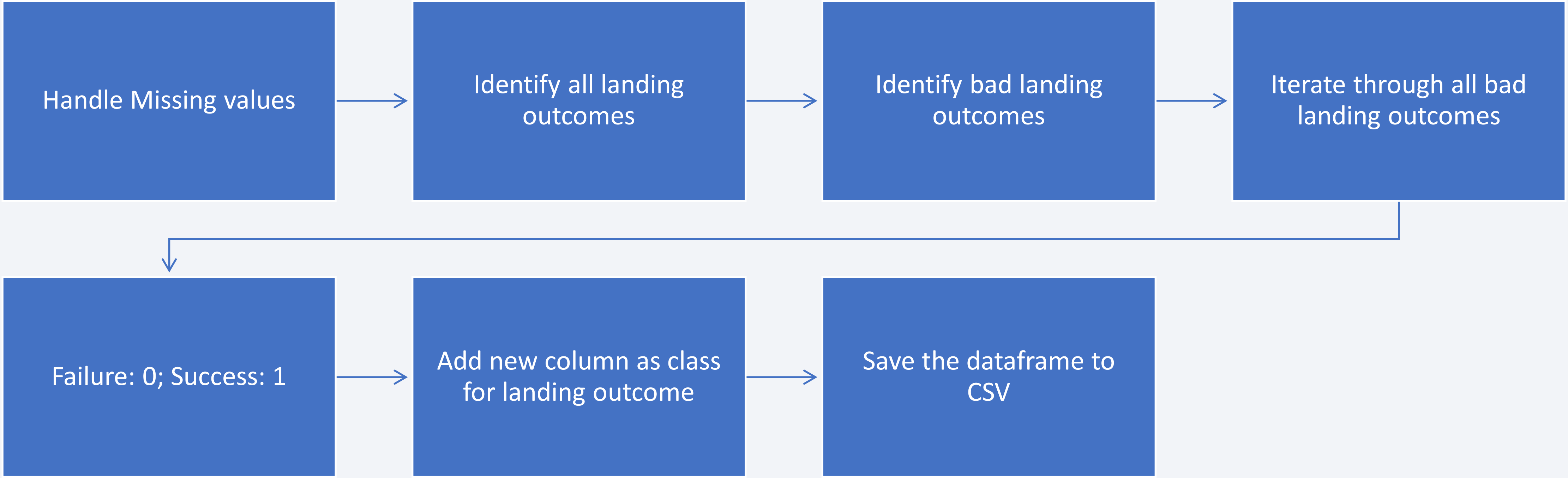
https://github.com/ThuyThiBird/Applied-Data-Science-Capstone/blob/main/1.jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



[Applied-Data-Science-Capstone/2.jupyter-labs-webscraping.ipynb at main · ThuyThiBird/Applied-Data-Science-Capstone (github.com)]

Data Wrangling



[Applied-Data-Science-Capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · ThuyThiBird/Applied-Data-Science-Capstone (github.com)]

EDA with Data Visualization

- We perform EDA using Matplotlib Visualization and Feature Engineering using Pandas
- To be specific, we investigated the relationships between Flight number, Launch Site, Payload Mass & Orbit type using variable graphs styles, including scatter plots, bar chart, line charts.
- The features were encoded using get_dummies() method, which converted categorical variables into binary features.
- GitHub link: Applied-Data-Science-Capstone/4.jupyter-labs-eda-dataviz.ipynb at main · ThuyThiBird/Applied-Data-Science-Capstone (github.com)

EDA with SQL

- Display unique values of launch sites in the space mission
- Display 5 records where launch sites begin with 'CCA'
- Display the total payload mass carried by boosters launched by NASA
- Display average payload mass carried by booster F9 v1.1
- List the date where the first successful launching outcome in ground pad was achieved
- List the name of boosters which have success in drone ship and have payload between 4000 and 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the max payload
- List the records which display the month names, failure landing outcomes in droneship, booster versions, launchsite for the months in 2015
- Rank the count of successful landing outcomes between 4/6/2010 and 20/3/2017
- GitHub link: [Applied-Data-Science-Capstone/3.jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · ThuyThiBird/Applied-Data-Science-Capstone (github.com)]

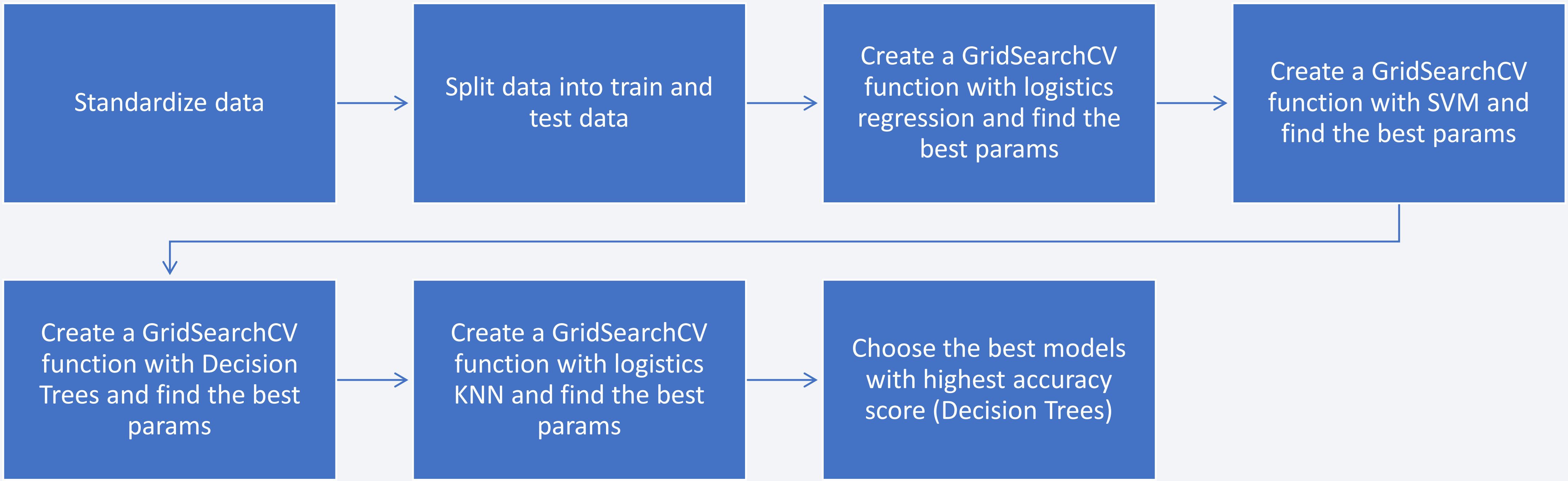
Build an Interactive Map with Folium

- Circles and markers were added to an interative Folium map to indicate launch locations
- A markerCluster was added to visualize the launch outcome at each site, with color codes
- A polyline was drawn between a launch site and another location (coast line, city, high way, metroline) to show the proximity of the two
- GitHub link: [Applied-Data-Science-Capstone/5.lab_jupyter_launch_site_location.ipynb at main · ThuyThiBird/Applied-Data-Science-Capstone (github.com)]

Build a Dashboard with Plotly Dash

- An interactive dashboard which allows users to investigate the effects of launch site, payload mass on the outcome (success or failure)
- Launch site was made selectable from a dropdown menu and a range of payload is selectable using a slider control
- Successful outcomes by launch sites are shown in a pie chart and How launch outcome varied by selected site and payload is shown is a scatter chart
- GitHub link: [Applied-Data-Science-Capstone/6.Spacex_dash_app.py at main · ThuyThiBird/Applied-Data-Science-Capstone (github.com)]()

Predictive Analysis (Classification)



[Applied-Data-Science-Capstone/7.SpaceX_Machine Learning Prediction_Part_5.ipynb at main · ThuyThiBird/Applied-Data-Science-Capstone (github.com)]

Results

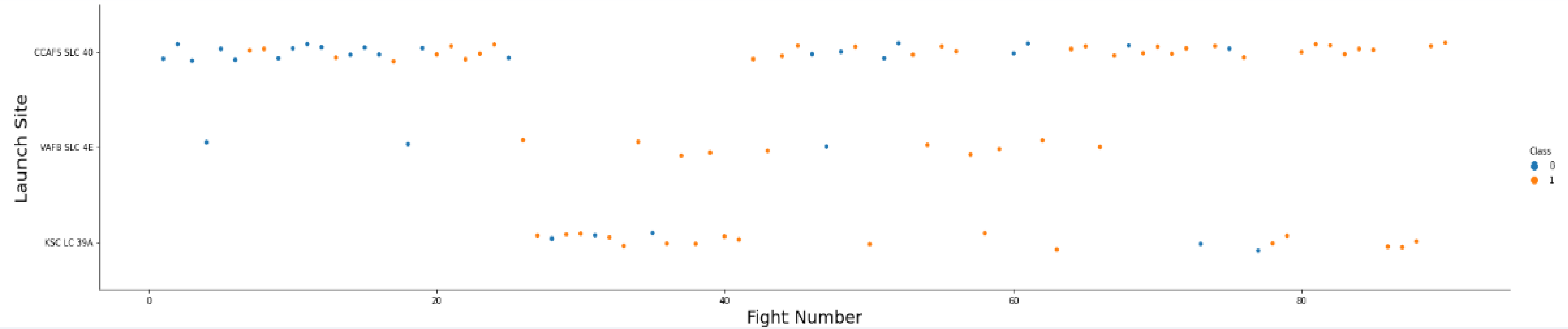
	Logistics Regression	Support vector machines	Decision Tree	KNN
Training score (best params)	0.845	0.848	0.888	0.848
Testing score (best params)	0.833	0.833	0.889	0.833

Section 2

Insights drawn from EDA

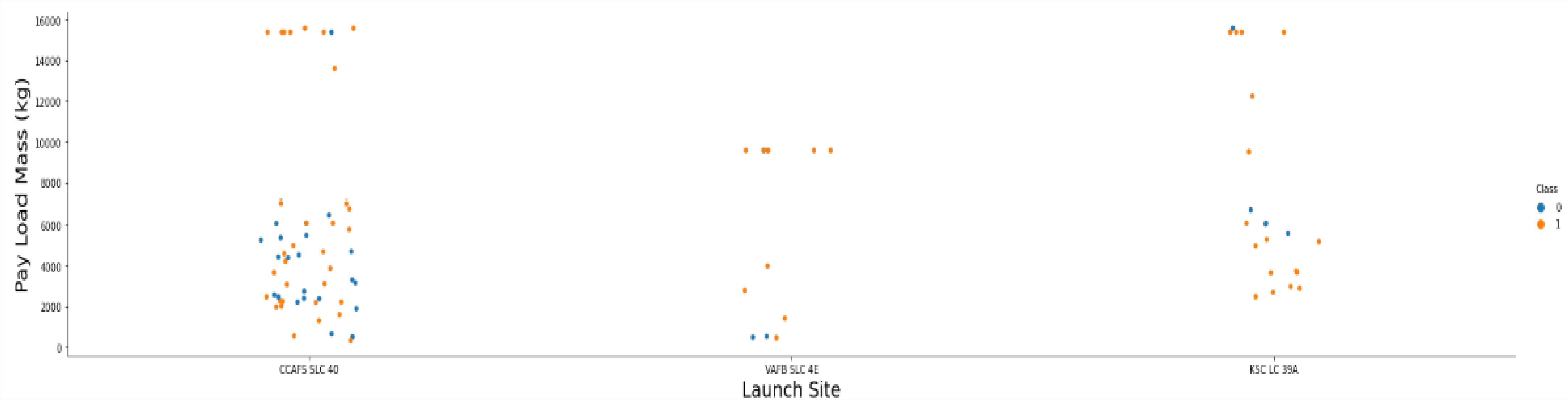
Flight Number vs. Launch Site

- CCAFS SLC 40 had more tests, however, success rate is lower vs the other 2



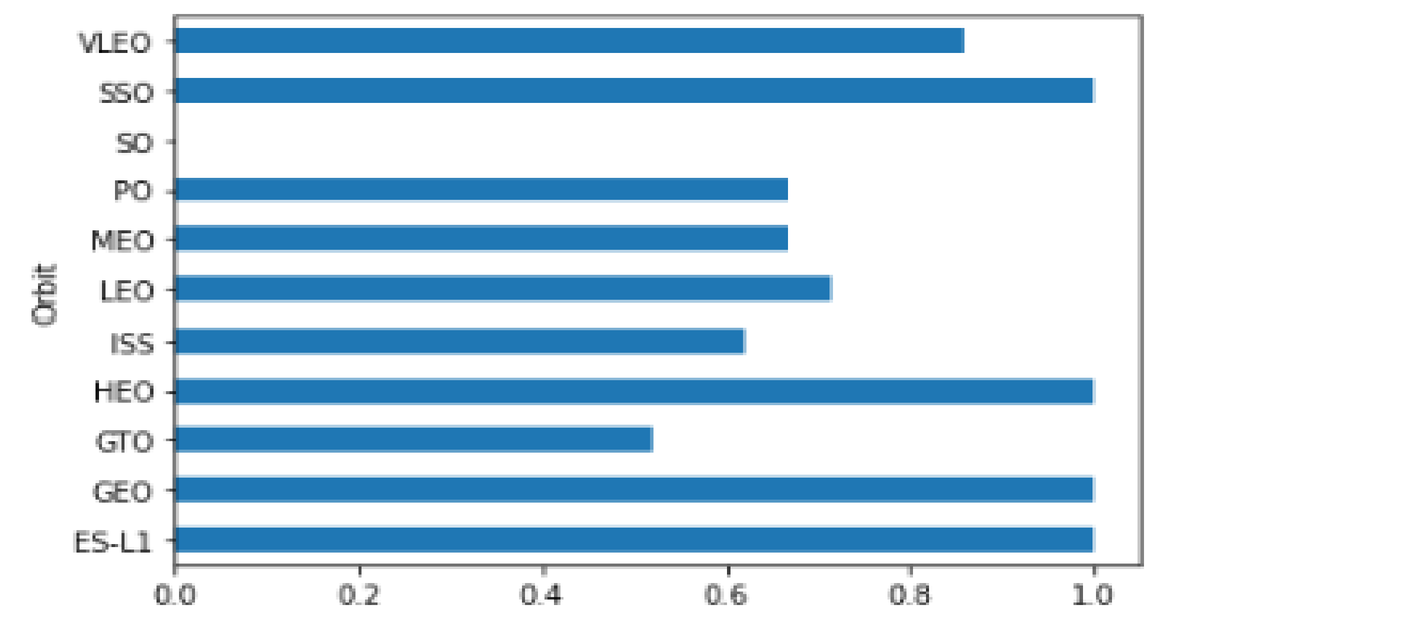
Payload vs. Launch Site

- Launches with heavier payload were more successful



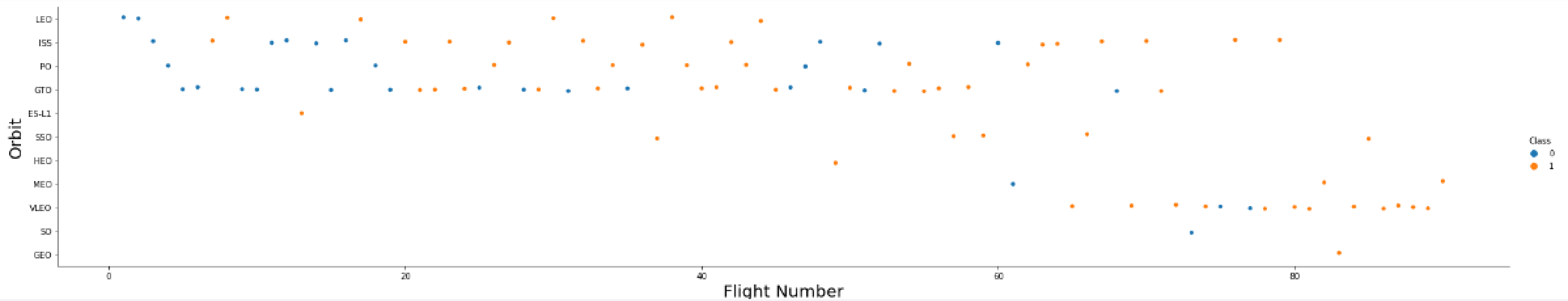
Success Rate vs. Orbit Type

- Orbit type SO has 0% success rate
- Orbit type SSO, HEO, GEO, ES-L1 are 100% successful



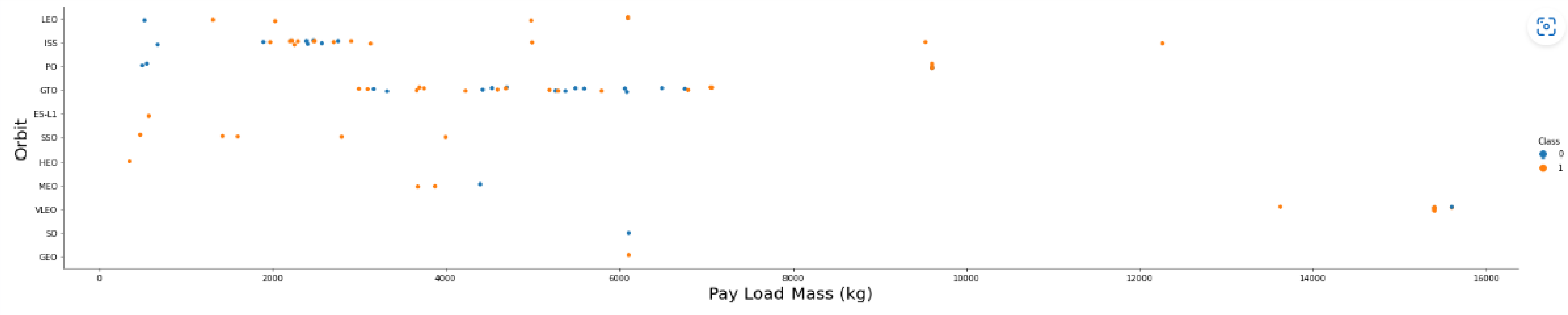
Flight Number vs. Orbit Type

- In all Orbit types, success rate increases when the number of flights increases



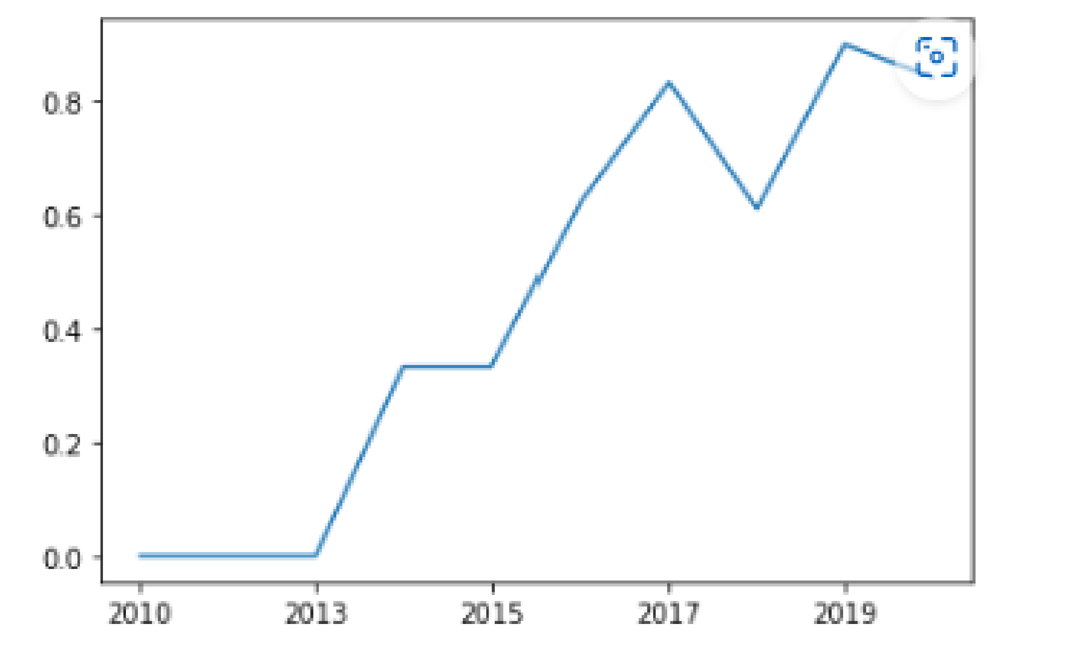
Payload vs. Orbit Type

- No clear patterns



Launch Success Yearly Trend

- Success rate increases by the time



All Launch Site Names

- Using Select distinct

```
%%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'

- Using like ‘%A%’ and limit 5

```
%%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

- Using sum function with filter using like

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

```
%%sql
select sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTBL
where Customer like 'NASA%'
```

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

total_payload_mass
99980

Average Payload Mass by F9 v1.1

- Using average function

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

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

average_payload_mass
2928.4

First Successful Ground Landing Date

- Using min function and filter with like

```
%%sql
select min(Date) from SPACEXTBL where "Landing _Outcome" like "Success%"
```

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

min(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- Using filter with between

```
%%sql
select 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

- Using count with case when to filter

```
%%sql
select count(case when Mission_Outcome like 'Success%' then Mission_Outcome else null end) as Success_mission,
count(case when Mission_Outcome like 'Failure%' then Mission_Outcome else null end ) as Failure_mission
from SPACEXTBL
```

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

Success_mission	Failure_mission
100	1

Boosters Carried Maximum Payload

- Using rank() over in sub query

```
%%sql
select Booster_version, PAYLOAD_MASS__KG_ from
(select *,
 RANK() OVER (ORDER BY PAYLOAD_MASS__KG_ DESC) as Rank from SPACEXTBL order by PAYLOAD_MASS__KG_)
where Rank = 1
```

```
 * 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

- Using substr() to get month and group by

```
%%sql
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)'
group by substr(Date,4,2), "Landing _Outcome", Booster_version, Launch_Site
```

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

month	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

- Using order by to get the outcomes by descending order

```
%%sql
select "Landing _Outcome", count("Landing _Outcome") as count_of_successful_launding
from SPACEXTBL
where Date between '04-06-2010' and '20-03-2017' and "Landing _Outcome" like 'Success%'
group by "Landing _Outcome"
order by count(case when "Landing _Outcome" like 'Success%' then "Landing _Outcome" else NULL end) desc
```

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

Landing _Outcome	count_of_successful_launding
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

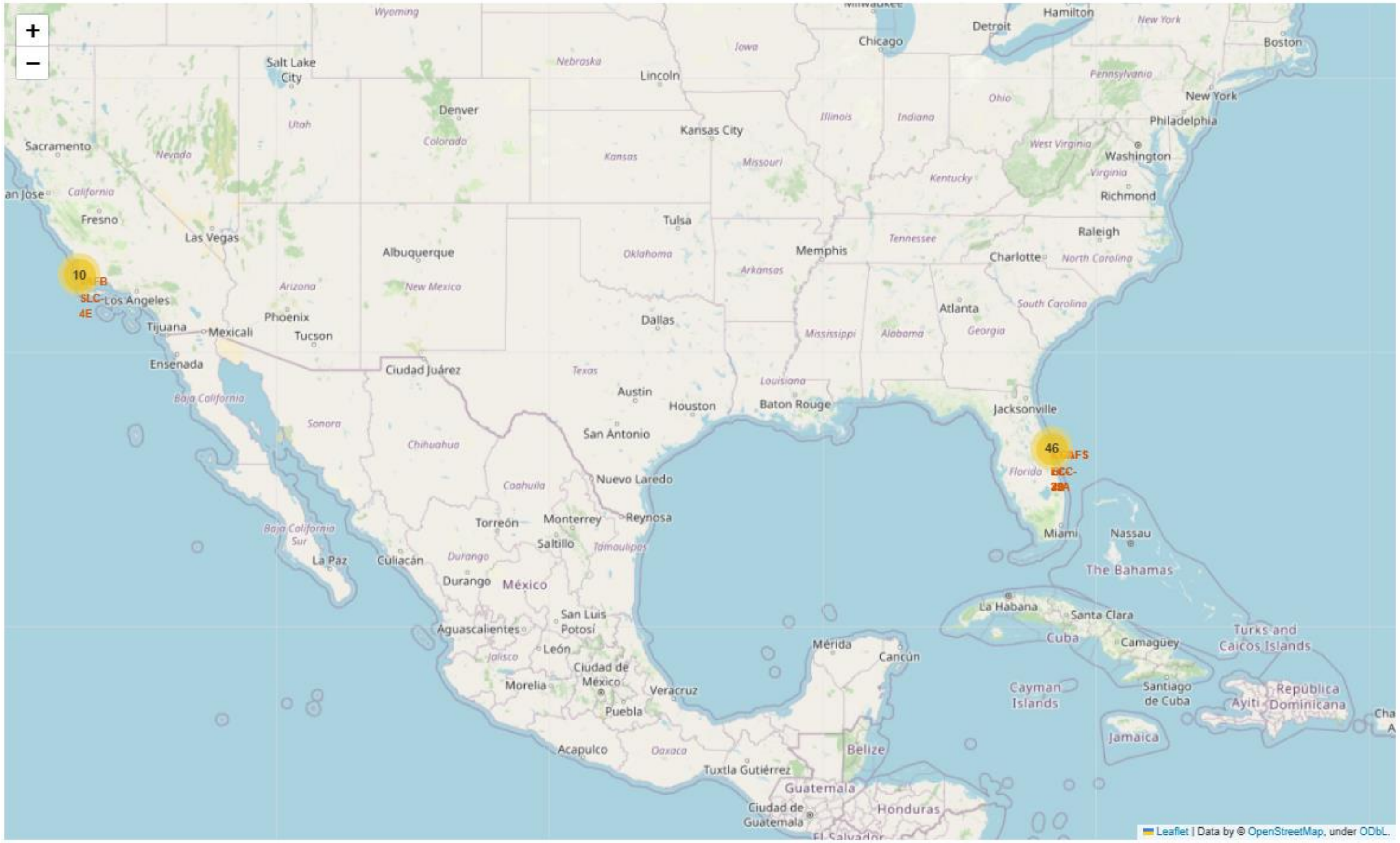
Launch Sites Proximities Analysis

Mark the launching sites

- Mark and Circle 4 launching sites

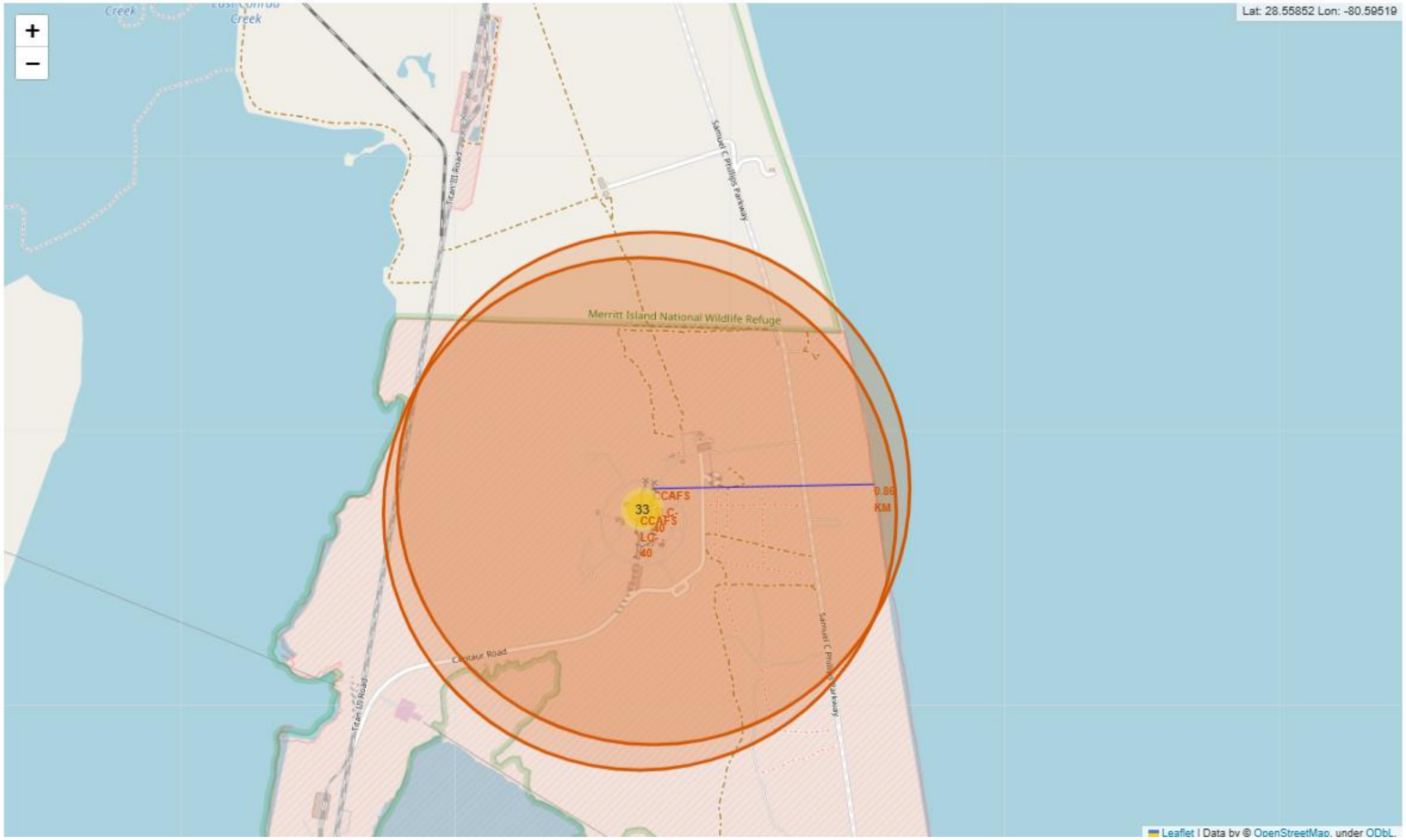
Mark success/failed launches

- Mark success and failed launches in each launching site with color coding of success (green) and failure (red)



Distance between a launch site to its proximities

- Pop up the distance in KM and draw a line between the launch site to the measured location

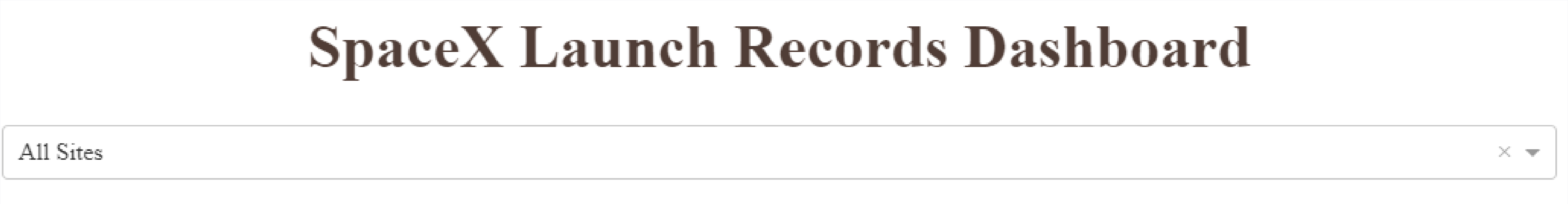


Section 4

Build a Dashboard with Plotly Dash

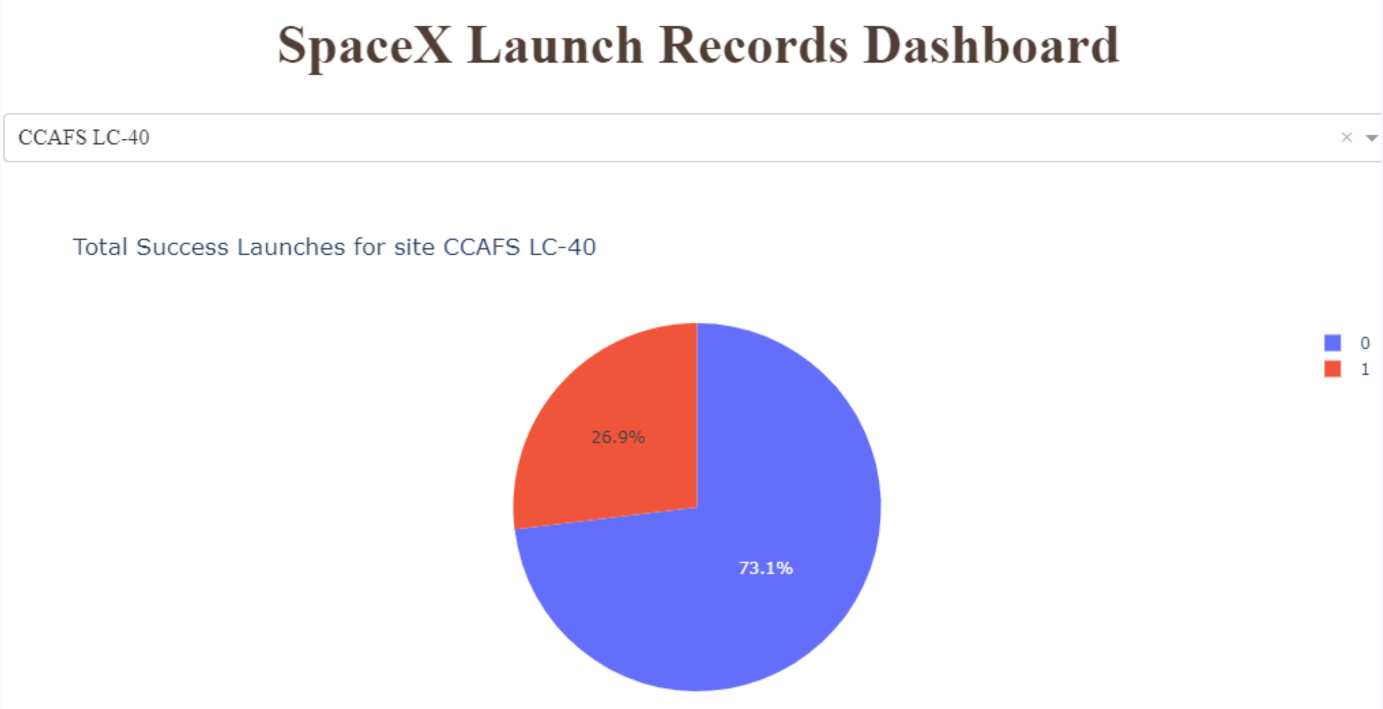
Dropdown menu to choose a launch site

- This allows users to choose a launch site or all launch sites to get interactive results



Pie chart of Success launches

- Total success launch for 1 site



Scatter plot of Success launches with respect to site and payload

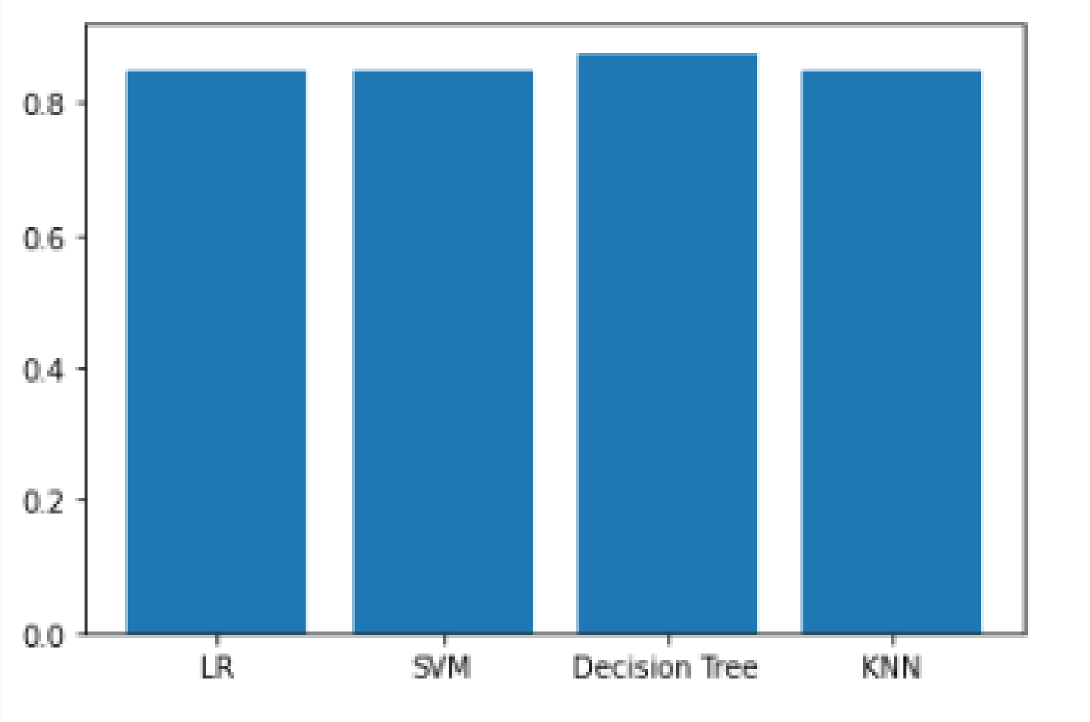
- Success launch by site and payload

Section 5

Predictive Analysis (Classification)

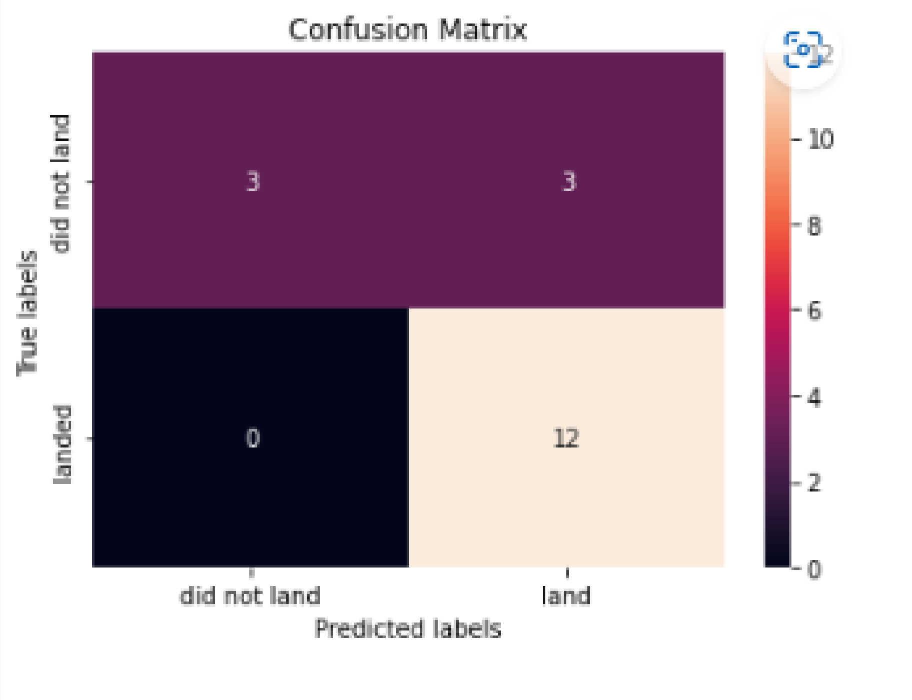
Classification Accuracy

- Decision Tree has the highest accuracy



Confusion Matrix

- 3 Launches did not land and were predicted to have launched
- 12 launches which landed where predicted correctly to land
- 15/18 launches were predicted correctly to launch or not launch



Conclusions

- Decision Tree give the best accurate forecast

Appendix

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 △ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 △ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule;[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulator in place of its engine.								
80	29 January 2020, 14:07[501]	F9 B5 △ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third operational and fourth large batch of Starlink satellites, deployed in a circular 290 km (180 mi) orbit. One of the fairing halves was caught, while the other was fished out of the ocean.[502]								
81	17 February 2020, 15:05[503]	F9 B5 △ B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fourth operational and fifth large batch of Starlink satellites. Used a new flight profile which deployed into a 212 km × 386 km (132 mi × 240 mi) elliptical orbit instead of launching into a circular orbit and firing the second stage engine twice. The first stage booster failed to land on the drone ship[504] due to incorrect wind data.[505] This was the first time a flight proven booster failed to land.								
82	7 March 2020, 04:50[506]	F9 B5 △ B1059.2	CCAFS, SLC-40	SpaceX CRS-20 (Dragon C112.3 △)	1,977 kg (4,359 lb)[507]	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
	Last launch of phase 1 of the CRS contract. Carries *Bartolomeo*, an ESA platform for hosting external payloads onto ISS.[508] Originally scheduled to launch on 2 March 2020, the launch date was pushed back due to a second stage engine failure. SpaceX decided to swap out the second stage instead of replacing the faulty part.[509] It was SpaceX's 50th successful landing of a first stage booster, the third flight of the Dragon C112 and the last launch of the cargo Dragon spacecraft.								
83	18 March 2020, 12:16[510]	F9 B5 △ B1048.5	KSC, LC-39A	Starlink 5 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fifth operational launch of Starlink satellites. It was the first time a first stage booster flew for a fifth time and the second time the fairings were reused (Starlink flight in May 2019).[511] Towards the end of the first stage burn, the booster suffered premature shut down of an engine, the first of a Merlin 1D variant and first since the CRS-1 mission in October 2012. However, the payload still reached the targeted orbit.[512] This was the second Starlink launch booster landing failure in a row, later revealed to be caused by residual cleaning fluid trapped inside a sensor.[513]								
84	22 April 2020, 19:30[514]	F9 B5 △ B1051.4	KSC, LC-39A	Starlink 6 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)

Thank you!