IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ulrike Eilhauer
11. July 2024

OUTLINE

Executive Summary

Introduction

Methodology

Results

Conclusion

Executive Summary

Summary of **methodologies**

- **Collection** of data using SpaceX API and web scraping on Wikipedia page
- **Wrangle** data to create a success/fail outcome variable
- **Explore and analyze** the data with SQL and visualization techniques
- **Visualize the location** of the launch sites
- **Predict launch success** with four different models: logistic regression, SVM, decision tree and KNN

Summary of all **results**

Exploratory data analysis results

- Launch success **increases over time**
- Site with highest success rate: KSC LC-39A
- Payload with higher success rate < 5000kg
- Booster Version with higher success rate: FT
- Orbit: ES-L1, GEO; HEO and SSO have a 100% success rate.

Interactive analytics demo in screenshots

- Most launch sies are near to the coast and **far enough away from damaging** proximities

Predictive analysis results

- **Decision tree model** is the best predictive model for this case

Introduction

In this capstone project, our goal is to predict the successful landing of the Falcon 9 first stage.

SpaceX promotes Falcon 9 rocket launches on its website, pricing them at 62 million dollars, whereas other providers charge upwards of 165 million dollars each. A significant part of these savings comes from SpaceX's ability to reuse the first stage.

By predicting whether the first stage will land, we can estimate the launch costs. This insight can be valuable for companies looking to compete with SpaceX for rocket launch bids.

Focus on

- Influence of payload mass, launch site, number of flights (over time) and orbits on success of the first stage landing
- Creation of a predicative model for successful landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web Scraping
- Perform data wrangling
  - Finding format with EDA, binary landing outcome label is created
- 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 – SpaceX API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
```

```
# We also want to convert the date_utc to a datetime datatype and then extracting the date leaving the time
data['date'] = pd.to_datetime(data['date_utc']).dt.date
```

Request SpaceX API (json) → Normalize SpaceX API (json) → Create data frame → Requesting more information from API (keeping relevant data) → Create new data frame from altered data → Create dictionary → Filter data frame (Falcon 9 launches) → Replace missing values with the respective mean → Export data to csv

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

https://github.com/UlrikeEE89/spacex_analysis/blob/main/01_jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Web Scraping

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(page, "html.parser")
```

```
#Extract each table
for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders collapsible")):
    # get table row
    for rows in table.find_all("tr"):
        #check to see if first table heading is as number corresponding to launch a number
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.string.strip()
                flag=flight_number.isdigit()
```

Request Falcon 9 launch data from Wikipedia → Create BeautifulSoup object (HTML response) → Extract column names → Collect data (parsing HTML tables)

Collect data (parsing HTML tables) → Create dictionary → Create data frame → Export data to csv

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

https://github.com/UlrikeEE89/spacex_analysis/blob/main/02_jupyter-labs-webscraping.ipynb

Data Wrangling

- EDA (Exploratory Data Analysis) is done to find patterns in the data
  - Number of launches on each site
  - Number and occurrence of each orbit
  - Number and occurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome (0 if bad, 1 if good)
- Data can now be processed with a binary format so that is more accessible for prediction models

https://github.com/UlrikeEE89/spacex_analysis/blob/main/03_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Charts

- Flight Number vs. Payload (scatter plot)
- Flight Number vs. Launch Site (scatter plot)
- Payload Mass (kg) vs. Launch Site (scatter plot)
- Success Rate of each orbit type (bar chart)
- Flight Number vs. Orbit type (scatter plot)
- Payload Mass (kg) vs Orbit Type (scatter plot)
- Success Rate over the years (line plot)

https://github.com/UlrikeEE89/spacex_analysis/blob/main/05_edadataviz.ipynb

EDA with SQL

Queries

- Unique launch sites
- 5 Launch sites that start with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster_versions which have carried the maximum payload mass
- Failed landing outcomes on drone ship, their booster version and launch site from 2015
- Ranking landing outcomes from 2010-06-04 and 2017-03-20, in descending order.

https://github.com/UlrikeEE89/spacex_analysis/blob/main/04_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers for launch sites (incl. popup label)
  - Blue circles at NASA Johnson Space Center’s coordinate
  - Red circles at all launch sites
- Markers for launch outcomes
  - Unsuccessful launches colored in red
  - Successful launches colored in green
- Distances between launch site and proximities
  - Measuring distance from launch site CCFAS SLC-40 and nearest coastline, railway, highway, city.

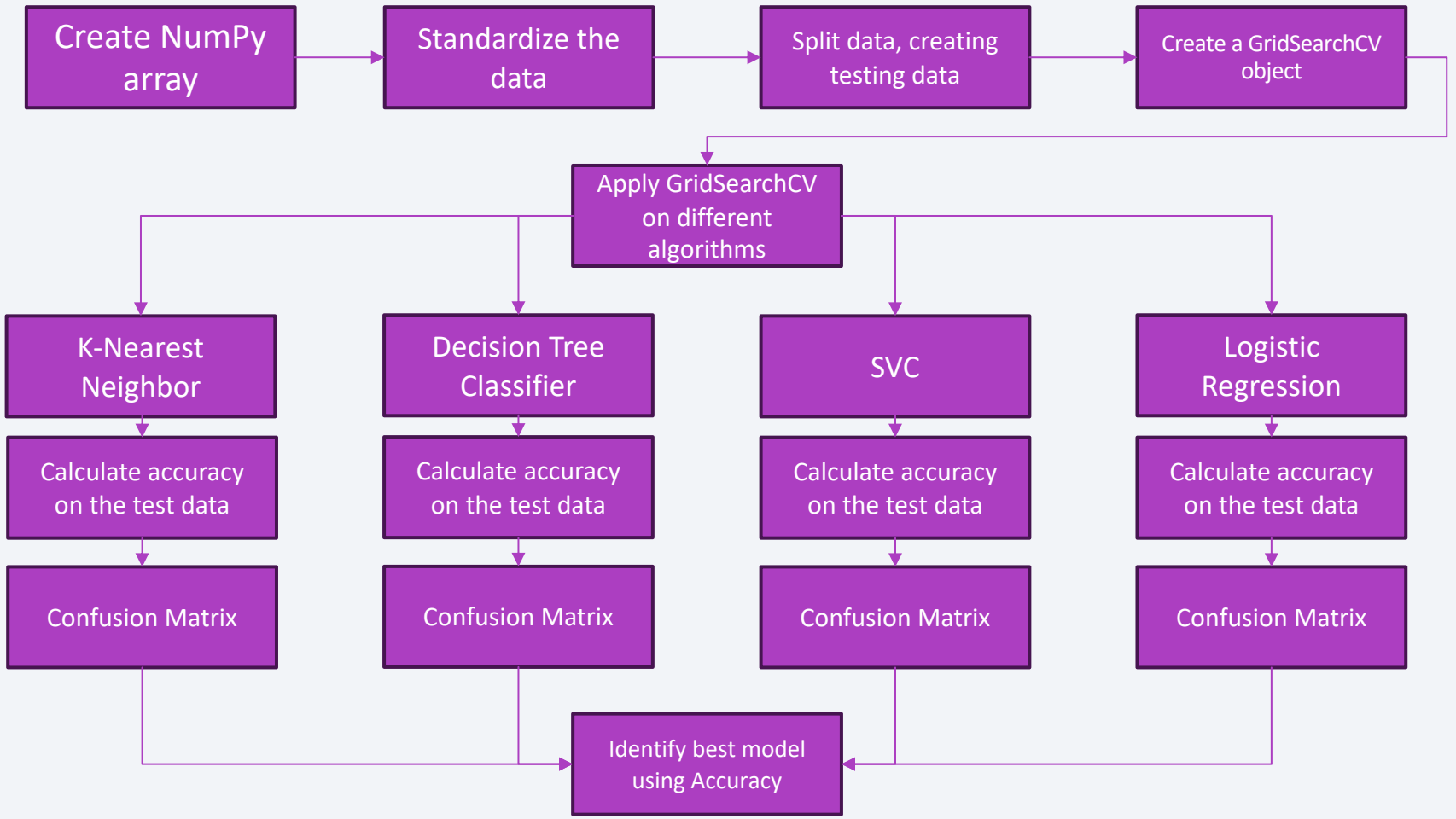
https://github.com/UlrikeEE89/spacex_analysis/blob/main/06_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- **Dropdown** list with lunch sites to be selected by users. All graphs shown are influenced by the dropdown.
- **Pie Chart** that shows the success launches by the selected sites from the dropdown. Possible to see which site had most success launches.
- **Scatter Chart** shows the correlation between payload mass and success rate by booster version. The payload mass can be scaled with a **slider** above the chart.

https://github.com/UlrikeEE89/spacex_analysis/blob/main/07_spacex_dash_app.py

Predictive Analysis (Classification)



https://github.com/UlrikeEE89/spacex_analysis/blob/main/08_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

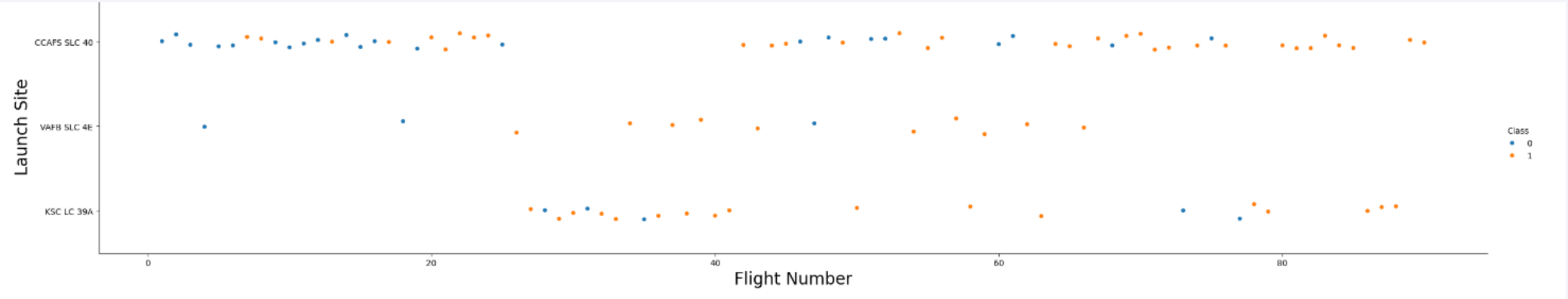
- Exploratory data analysis results
  - Launch success increases over time
  - Site with highest success rate: KSC LC-39A
  - Payload with higher success rate < 5000kg
  - Booster Version with higher success rate: FT
  - Orbit: ES-L1, GEO; HEO and SSO have a 100% success rate.
- Interactive analytics demo in screenshots
  - Most launch sies are near to the coast and far enough away from damaging proximities
- Predictive analysis results
  - Decision tree model is the best predictive model for this case

Section 2

Insights drawn from EDA

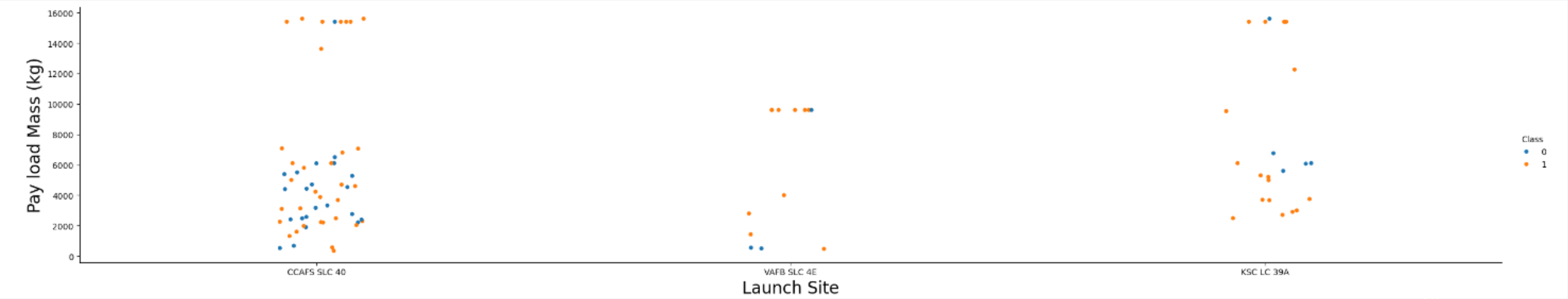
Flight Number vs. Launch Site

- Recent flights have a higher success rate (orange = success, blue = failure)
- Most flights (around 50%) were done from CCFAS SLC 40 launch site
- KSC LC-39A has the highest success rate



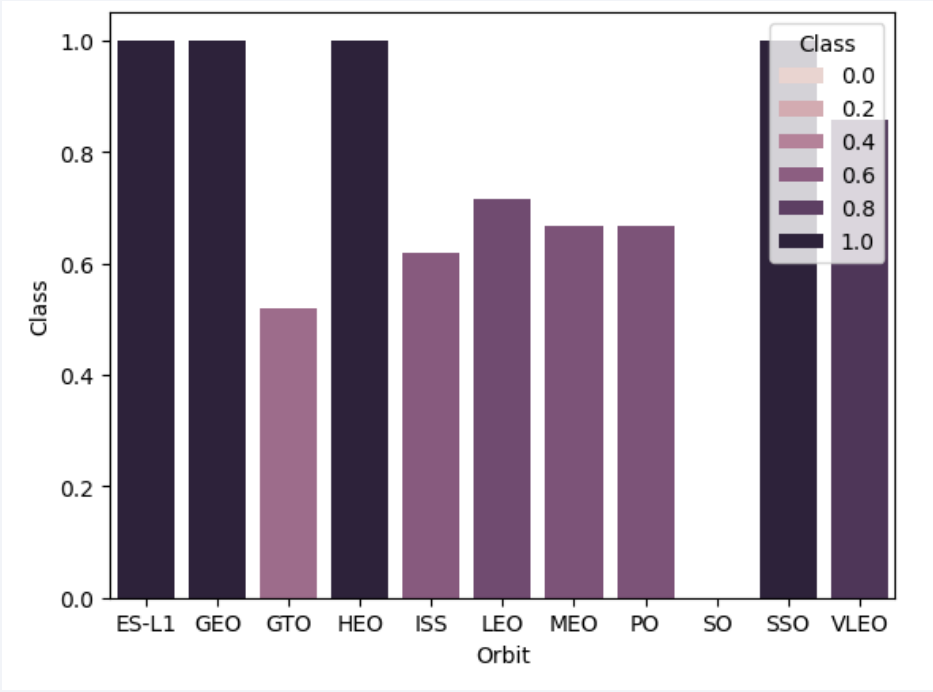
Payload vs. Launch Site

- Payload mass with more than 10000 kg were not launched by VAFB SKC 4E site
- The success rate is higher for higher payload mass



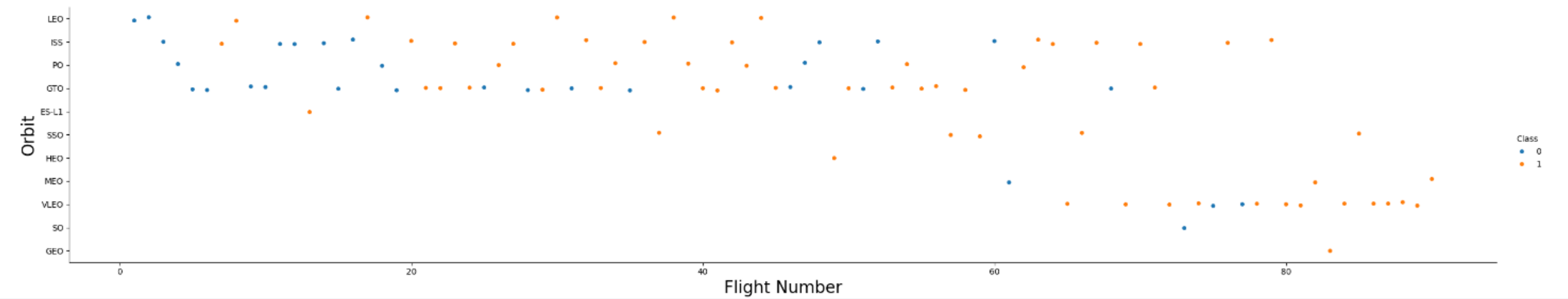
Success Rate vs. Orbit Type

- Orbit: ES-L1, GEO; HEO and SSO have a 100% success rate
- Orbit: GTO, ISS, LEO, MEO and PO have a success rate between 50-70%
- SO has 0% success rate



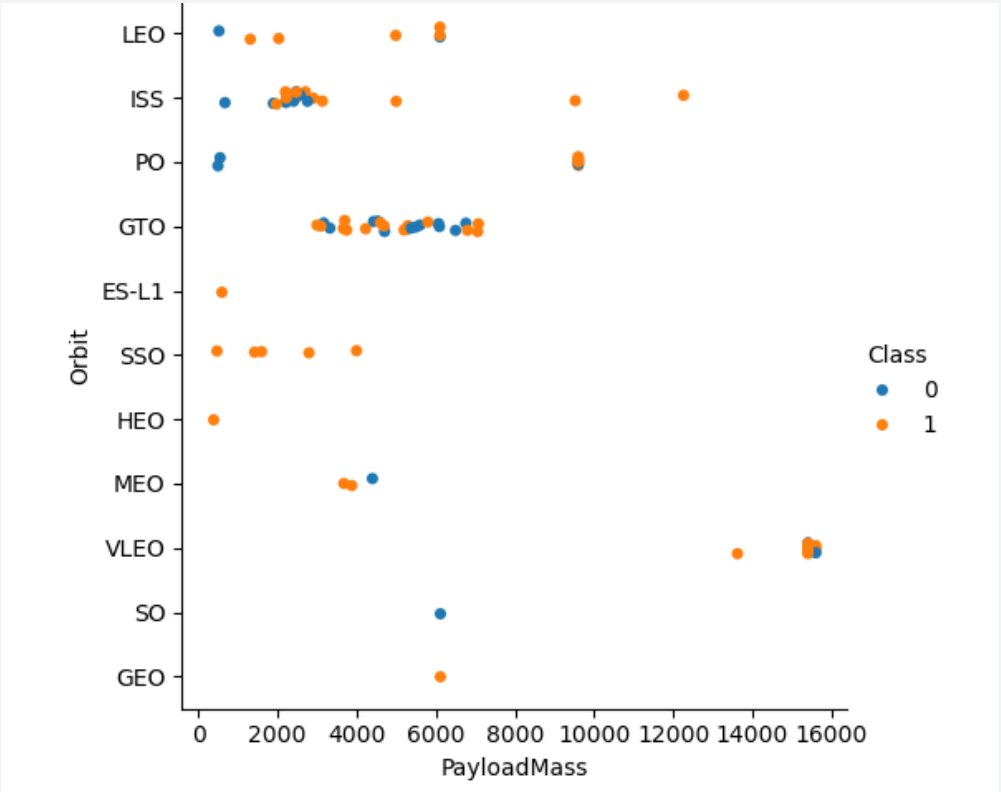
Flight Number vs. Orbit Type

- Success rate increases with number of flights on each orbit



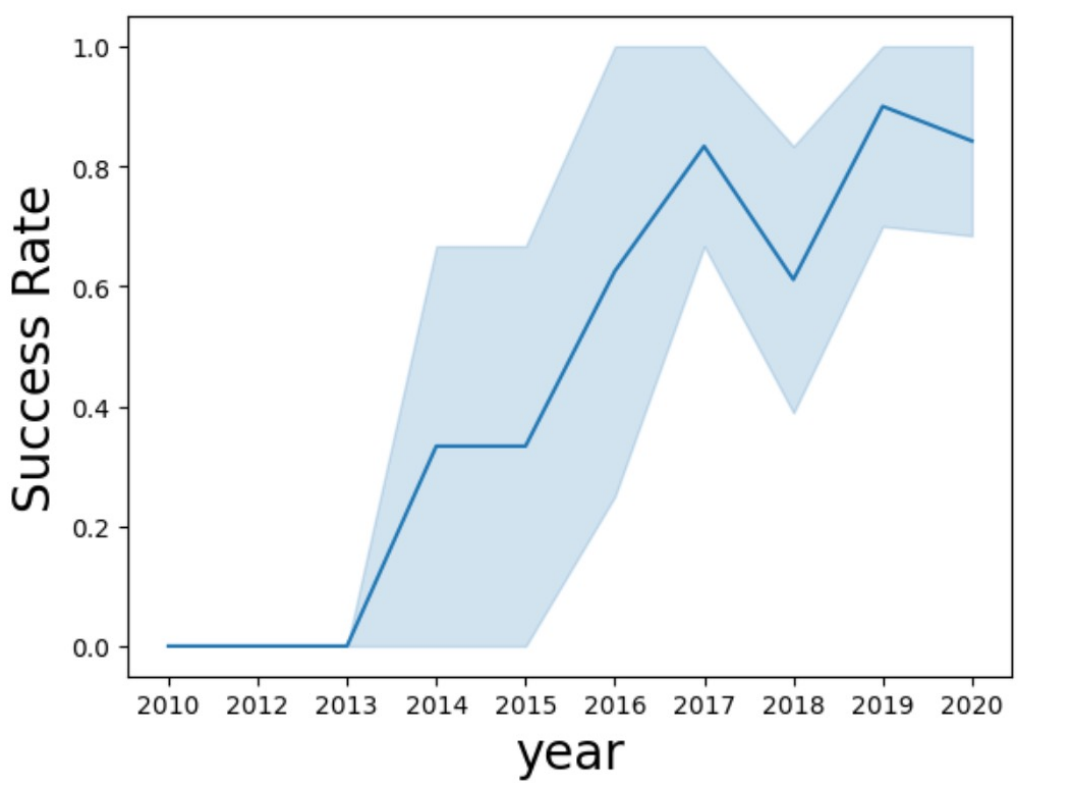
Payload vs. Orbit Type

- Higher payload mass is better for LEO, ISS and PO orbit
- GTO orbit has mixed success, so does not depend on payload mass



Launch Success Yearly Trend

- Overall, a clear increase from 2010 to 2020 in the success rate is visible
- Success rate improved strongly between 2013 and 2017
- Between 2017 and 2018 was a 20% decrease in the success rate



All Launch Site Names

- Unique launch sites:
  - CCFAS LC-40
  - CCFAS SLC-4E
  - KSC LC-39A
  - VAFB SLC-4E

```
%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'

- Find 5 records where launch sites 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
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- 45596kg is the sum of all payload masses from NASA (CRS)

```
%sql SELECT sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL WHERE CUSTOMER= 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
```

payloadmass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- 2928 kg in average carried by booster version F9 v1.1

```
%sql SELECT avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

payloadmass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The first successful landing outcome was 2015-12-22

```
%sql SELECT min(Date) FROM SPACEXTBL WHERE Landing_Outcome LIKE "%ground%"
```

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

min(Date)
2015-12-22

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
- JSCAT-14, JSCAT-16, SES-10, SES-11/ EchoStar 105

```
%sql select * from SPACEXTBL where (LANDING_OUTCOME LIKE "%success (drone%" and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000)
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcome
- 100 Successful missions

```
%sql select MISSION_OUTCOME, count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

boosterversion
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

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

- Unfortunately did not work as expected. No error shown.

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME FROM SPACEXTBL where LANDING_OUTCOME = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome

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

- Unfortunately did not work as expected. No error shown.

```
%sql SELECT LANDING_OUTCOME, count(*) as count_outcomes  FROM SPACEXTBL WHERE DATE between '04-06-2010' and '20-03-2017' group by LANDING_OUTCOME order by count_outcomes DESC;

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

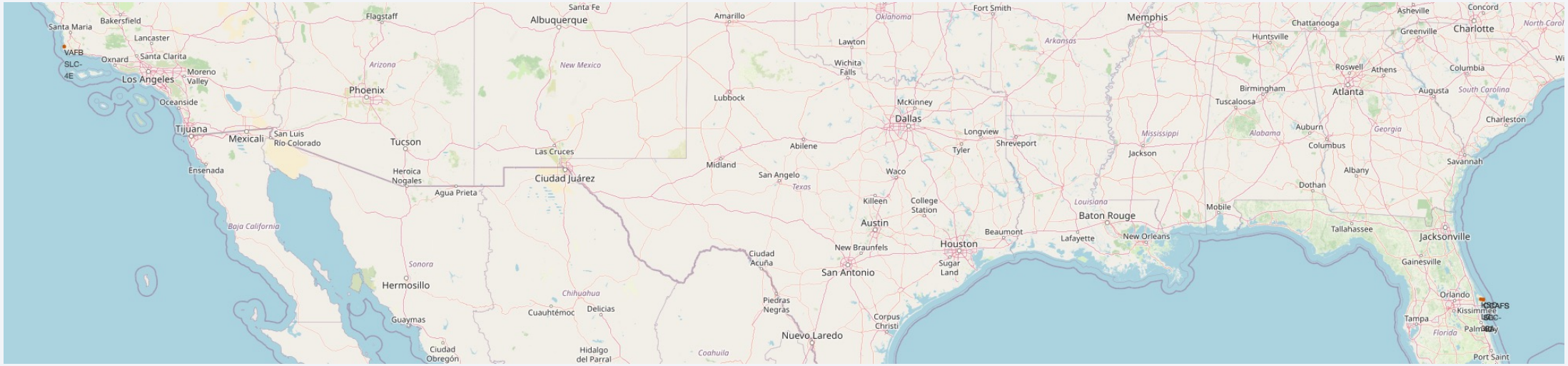
Landing_Outcome	count_outcomes

Section 3

Launch Sites Proximities Analysis

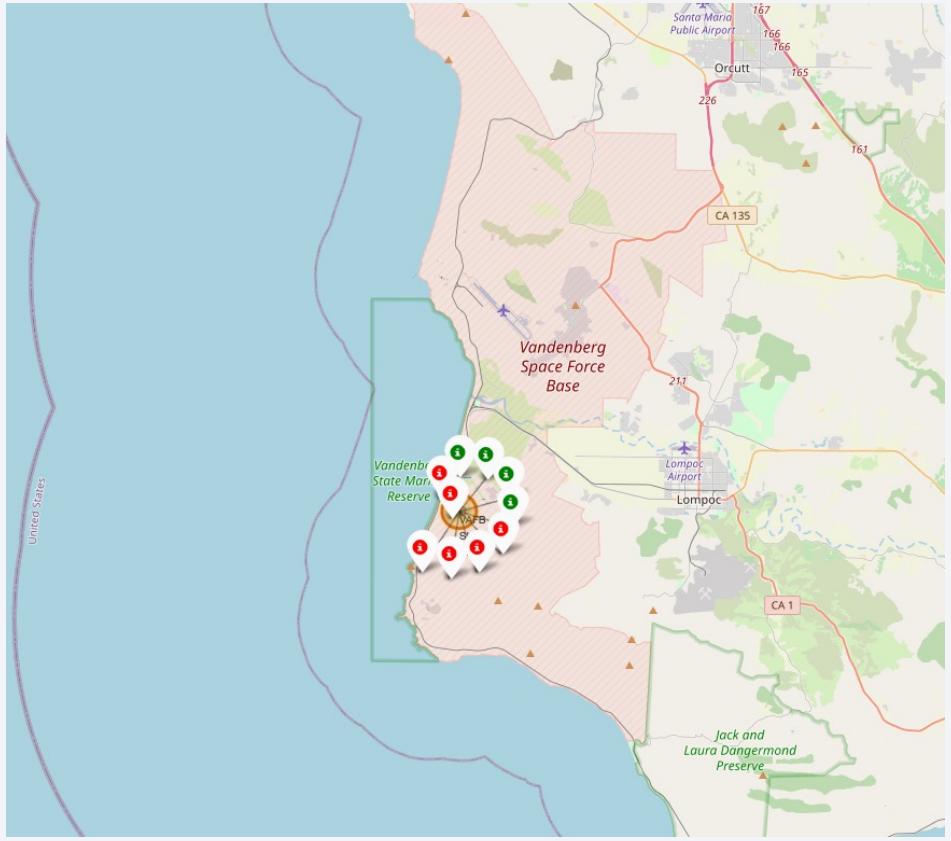
All launch sites' location

- Launch location on the coats of the USA. They are situated near equator.
- 1 location on east coast, 3 locations on west coast.



Launch Outcomes

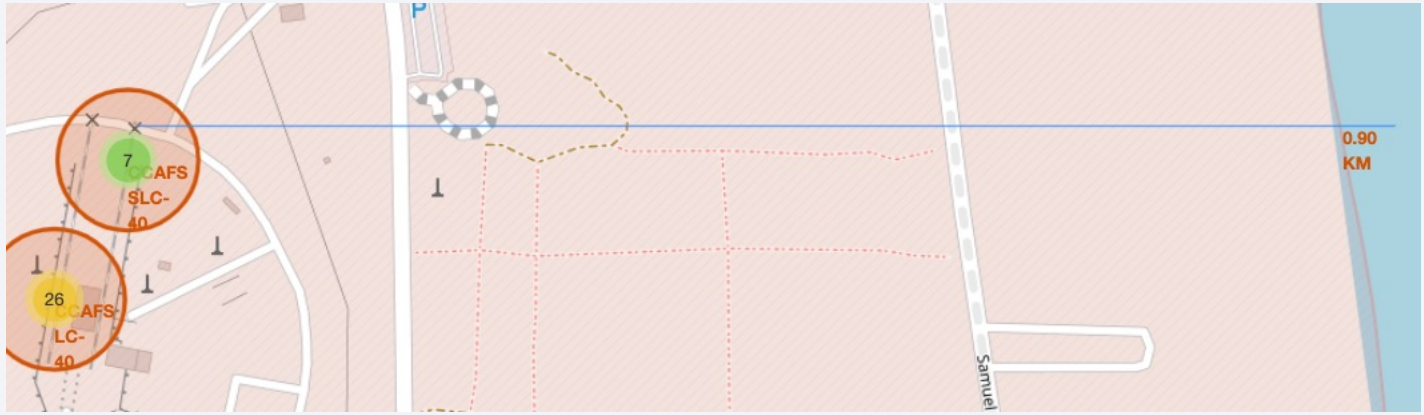
- Green markers for successful launch
- Red markers for unsuccessful launch
- Example in the graph: VAFB SLC-4E with a success rate of 4/10 = 40%



Site distance to proximities

CCAFS SLC-40

- 0.9km from nearest coastline
- 21.9km from nearest railway
- 23.2km from nearest city
- 26.9km from nearest highway

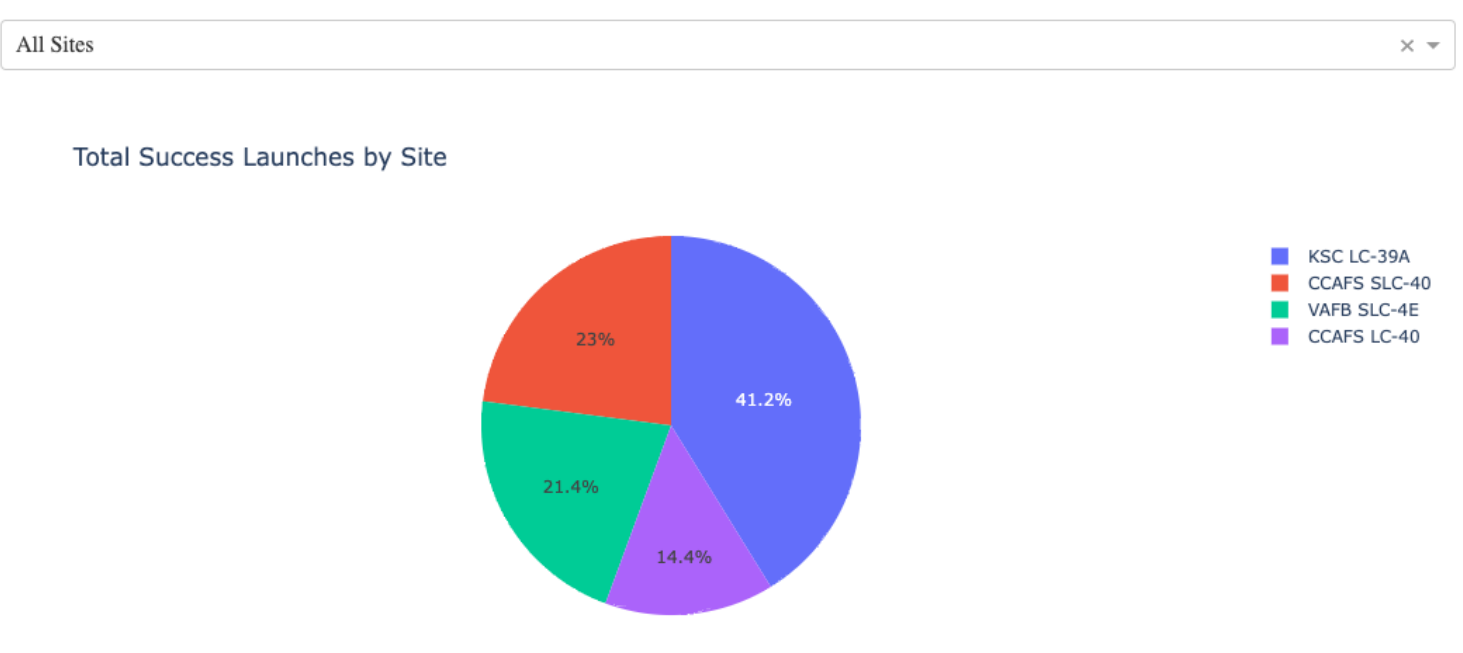


Section 4

Build a Dashboard with Plotly Dash

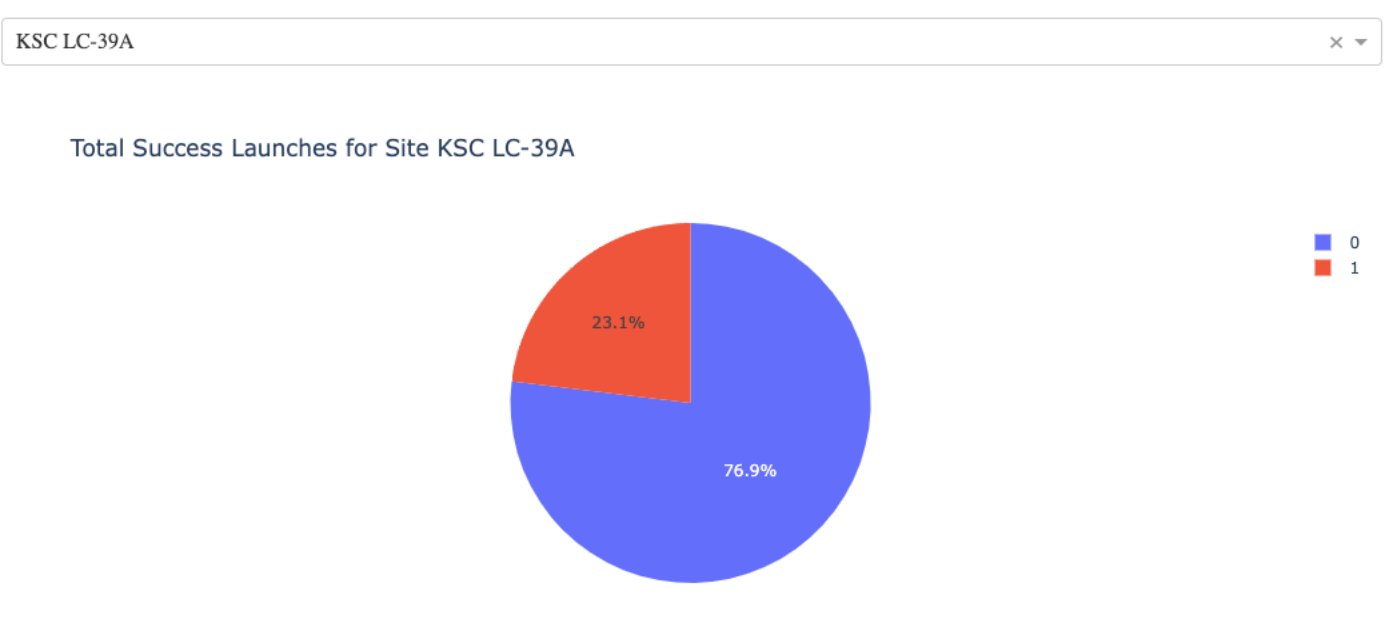
Total Success Launches by Site

- The pie chart represents all success launches by each site as percent of total.
- The site KSC LC-39A has the most successful launches with 41.2%.
- The site CCAFS LC-40 has the least successful launches with 14.4%.



Launches at KSC LC-39A

- KSC LC-39A has among all the space x sites the most successful launches.
- 76.9% of the launches at KSC LC-39A were successful. 23.1% failed.



Payload vs. Launch Outcome scatter plot for all sites

- Payloads lower than 5000 kg have a higher success rate.
- Booster Version Category FT has the highest success rate among booster versions.

Payload range between 3000-5000kg

Payload range between 5000-10000kg

Section 5

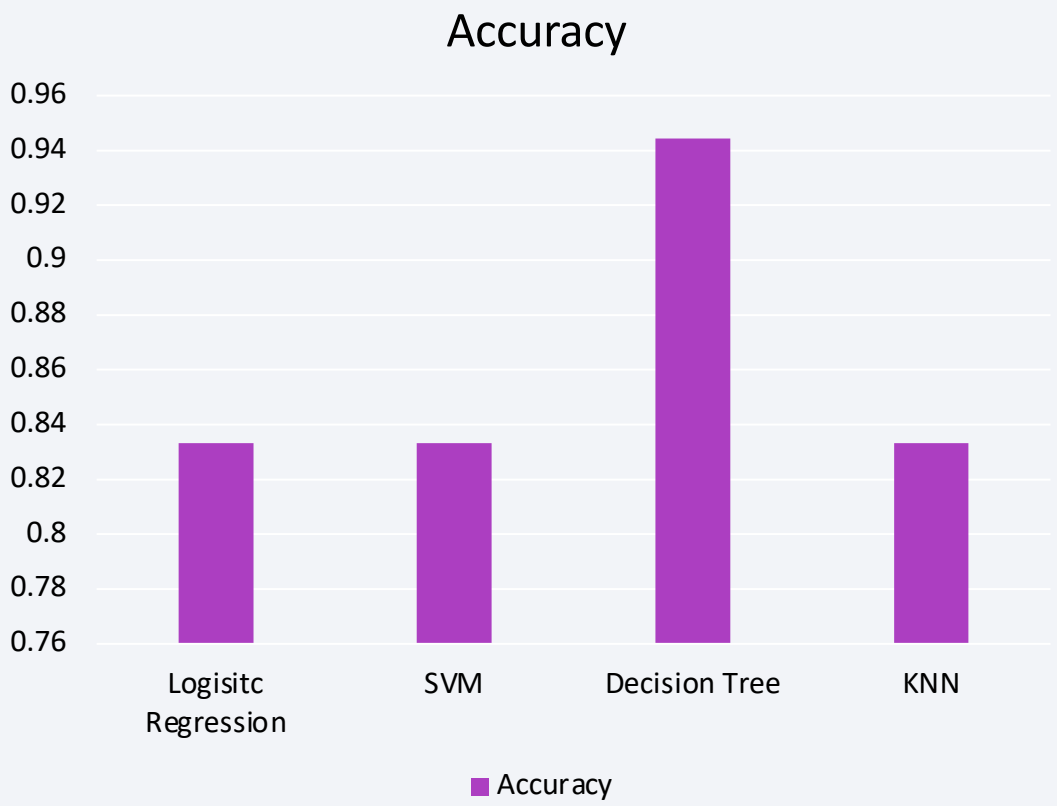
Predictive Analysis (Classification)

Classification Accuracy

- The **decision tree** has the highest accuracy

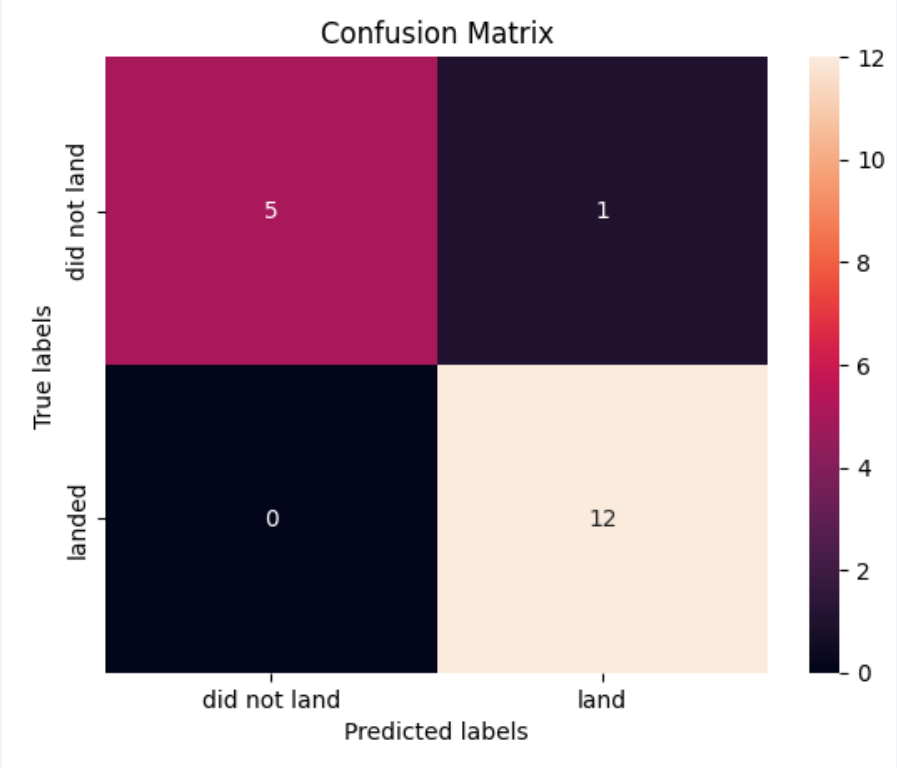
```
print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.9444444444444444
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```



Confusion Matrix

- Only the decision tree confusion matrix is not identical like the others
- The other matrixes have 3 false positives (we predict that launch lands but does not) while the decision tree model has only 1 false positive.



Conclusions

- The chosen prediction models perform similarly on the test set. Only the decision tree model could slightly outperform the others in accuracy.
- Launch success rate:
  - Over time the success rate increases
  - Site with highest success rate: KSC LC-39A
  - Payload with higher success rate < 5000kg
  - Booster Version with higher success rate: FT
  - Orbit: ES-L1, GEO; HEO and SSO have a 100% success rate.
- Launch Sites are close to the coast and enough away from proximities.

Thank you!