Winning Space Race with Data Science

Enrique Kishimoto
October 19, 2021

OUTLINE

EXECUTIVE SUMMARY

Summary of methodologies

- Data Collection
- Data Wrangling
- EDA with Visualization
- EDA with SQL
- Building an Interactive maps with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis

Summary of all results

- Preliminary analysis with based on EDA
- Interactive maps and dashboards
- Predictive results

INTRODUCTION

Project background and context

We want to predict if the Falcon 9 first stage will land successfully. 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 are the conditions for a successful landing?
- What are the outcome dependent on different variables with success rate?
- What conditions does SpaceX have to achieve to get the best rocket success landing rate?

METHODOLOGY

METHODOLOGY

Data collection methodology

- Using SpaceX Rest API
- Web Scrapping from Wikipedia

Perform data wrangling

- Data was cleaned from irrelevant columns and transformed using one hot encoding for Machine Learning

Perform exploratory data analysis (EDA) using visualization and SQL

- Plotting: Scatter, Bar and Line graphs to show patterns of data

Perform interactive visual analytics using Folium and Plotly Dash

- Using Folium and Plotly Dash Visualization to build interactive maps and dashboards

Perform predictive analysis using classification models

- How to build, tune, evaluate classification models

DATA COLLECTION

Data was gathered using the SpaceX Rest API.

This API deliver data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

The goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.

The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.

Another popular data source for obtaining Falcon 9 Launch data is web scraping related Wiki pages.

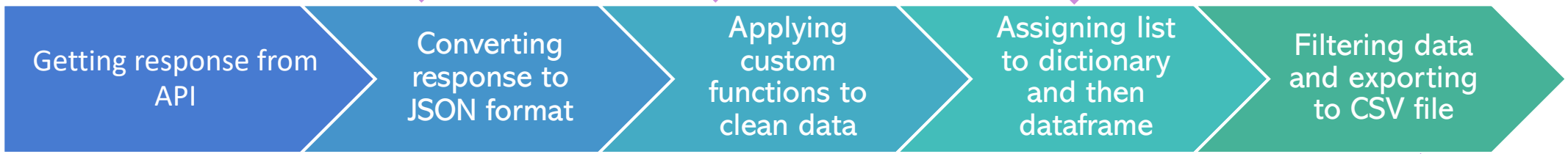
DATA COLLECTION – SPACEX API

GitHub Link

```
response = requests.get(static_json_url)
data = pd.json_normalize(response.json())
```

```
getBoosterVersion(data)
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

```
launch_dict = {'FlightNumber': list(data['flight_number']),
  'Date': list(data['date']),
  'BoosterVersion':BoosterVersion,
  'PayloadMass':PayloadMass,
  'Orbit':Orbit,
  ....
}
data = pd.DataFrame(launch_dict)
```




```
data_falcon9 = data[data.BoosterVersion == 'Falcon 9']
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
response = requests.get(static_json_url)
```

DATA COLLECTION - SCRAPING

GitHub Link


```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
response = requests.get(static_url).text
```

```
soup = BeautifulSoup(response, 'html.parser')
```

```
html_tables = soup.find_all('table')
```

```
column_names = []
elements = soup.find_all('th')
for i in range(len(elements)):
    try:
        name = extract_column_from_header(elements[i])
        if name is not None and len(name) > 0:
            column_names.append(name)
    except:
        pass
```

```
launch_dict= dict.fromkeys(column_names)
```

```
launch_dict['Flight No.'].append(flight_number)
launch_dict['Payload mass'].append(payload_mass)
launch_dict['Launch outcome'].append(launch_outcome)
```

```
df = pd.DataFrame(launch_dict)
```

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

DATA COLLECTION – SPACEX API

GitHub Link

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident.

Ocean means the mission outcome landed to a specific region of the ocean
RTLS means the mission outcome landed to a ground pad
ASDS means the mission outcome landed on a drone ship
True means the mission outcome was successfully
False means the mission outcome was unsuccessfully

We mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

Calculate the # of launches per site

```
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
```

Calculate the # and occurrence of each orbit

```
df["Orbit"].value_counts()
```

```
GTO     27
ISS     21
VLEO    14
PO       9
```

Calculate the # and occurrence of outcome per orbit type

```
landing_outcomes = df["Outcome"].value_counts()
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
```

Create a landing outcome label from Outcome column

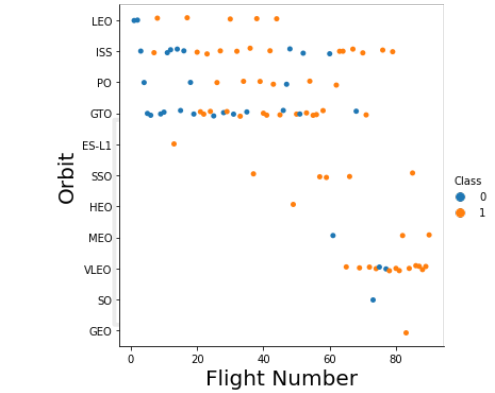
```
landing_class = []
for key,value in df["Outcome"].items():
    if value in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)

df['Class']=landing_class
```

Export data to CSV file

```
df.to_csv("dataset_part_2.csv", index=False)
```

EDA WITH DATA VISUALIZATION

GitHub Link



Scatter Plots to show correlation between these features:

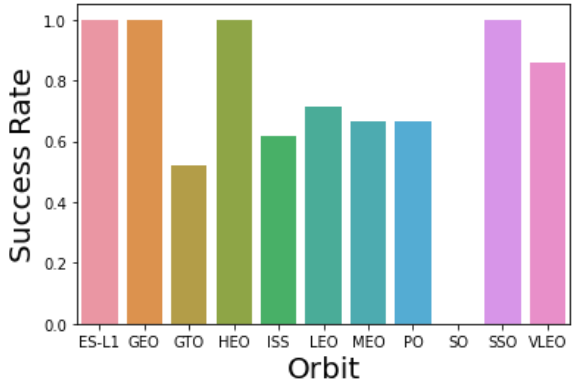
Flight Number vs Payload Mass

Flight Number vs Launch Site

Payload Mass vs Launch Site

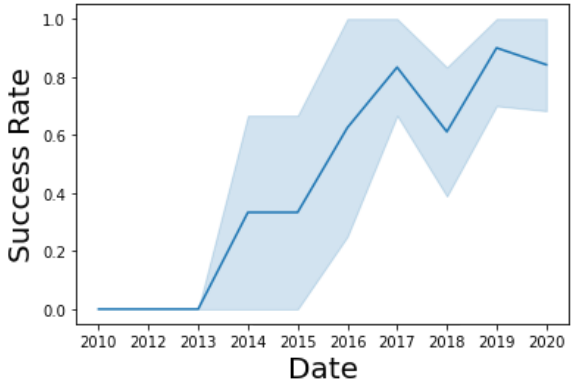
Flight Number vs Orbit Type

Payload Mass vs Orbit Type



Bar Graph to compare sets of data between different groups. In this case, we can determine which orbits have the highest success rate.

Success Rate vs Orbit Type



Line Graph to show trends clearly and we can make predictions.

Success Rate Yearly Trend

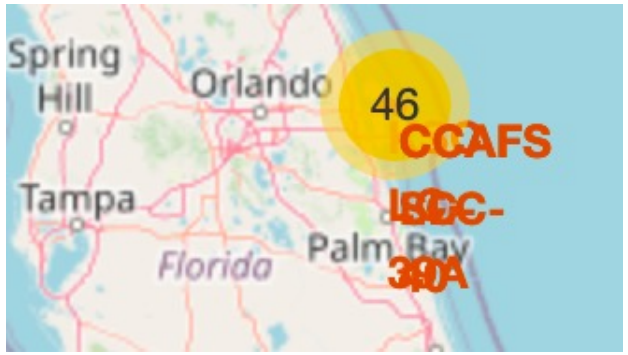
EDA WITH SQL

GitHub Link

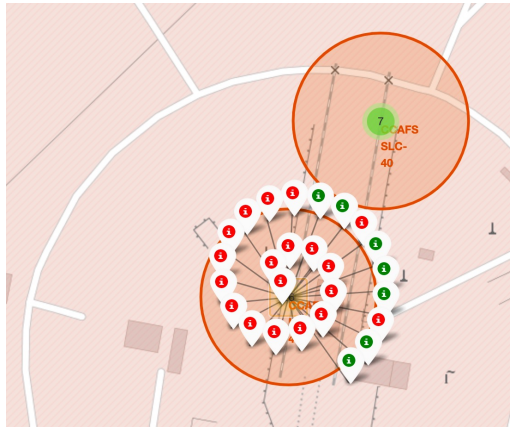
We performed SQL queries to gather information from give dataset

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Ranking 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

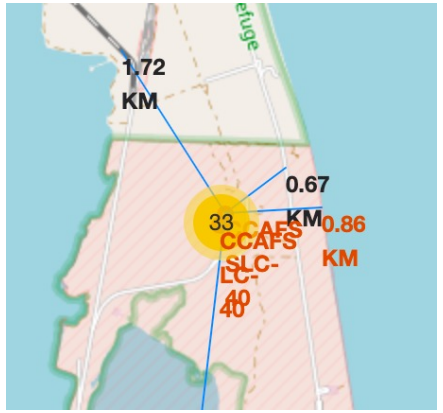
BUILD AN INTERACTIVE MAP WITH FOLIUM



We used latitude and longitude coordinates for each launch site and added Circle Marker around each launch site with a label of the name of the launch site



We used a green marker to determine if a launch was successful and a red marker when a launch was failed.



We used lines to indicate the distances between launch site to its proximities, such as Coastline point, Closest City, Railway and Highway.

BUILD A DASHBOARD WITH PLOTLY DASH

GitHub Link

Pie Chart

- It shows the success rate of all launch sites.
- It can display the proportion between success and fails of given launch site.

Scatter Plot

- It shows the correlation between Mission Outcome and Payload Mass (Kg) for different Booster Versions for all sites or a given launch site.
- The Payload Mass can be filtered by a weight range using the slider

PREDICTIVE ANALYSIS (CLASSIFICATION)

GitHub Link

Building Model

- Preprocessing and standardizing data
- Splitting data into training and test sets
- Optimizing parameters for each model using Grid Search
- Training model and perform Grid Search

Evaluating Model

- Checking accuracy for each model
- Getting best hyperparameters
- Plotting confusion matrix

Finding the best performing Classification Model

- Choosing the model with the best accuracy score

RESULTS

Exploratory data analysis results

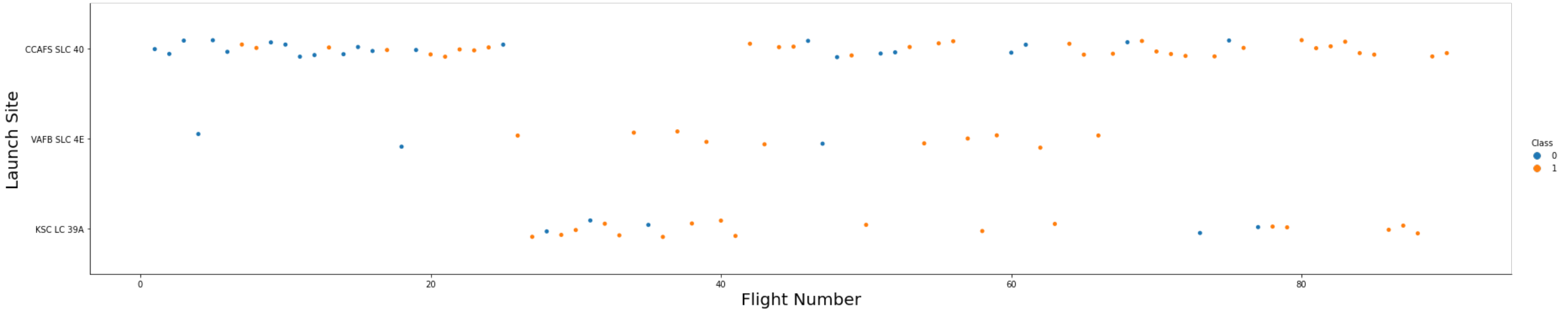
Interactive analytics demo in screenshots

Predictive analysis results

INSIGHTS DRAWN FROM EDA

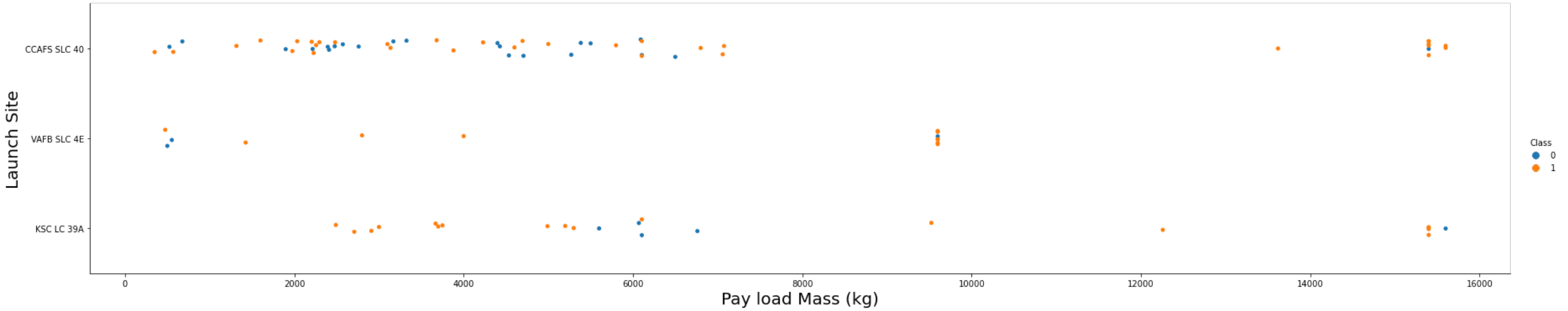
EDA WITH VISUALIZATION

FLIGHT NUMBER VS. LAUNCH SITE



- More recent mission have higher success rate at all launch sites
- Most of the first flights from Cape Canaveral SLC 40 were unsuccessful

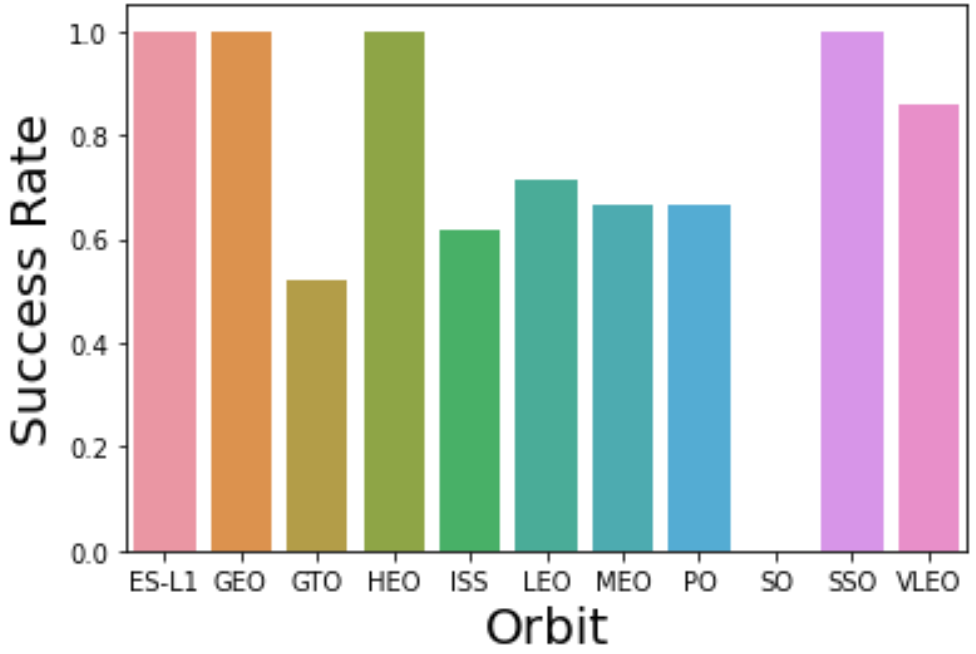
PAYLOAD VS. LAUNCH SITE



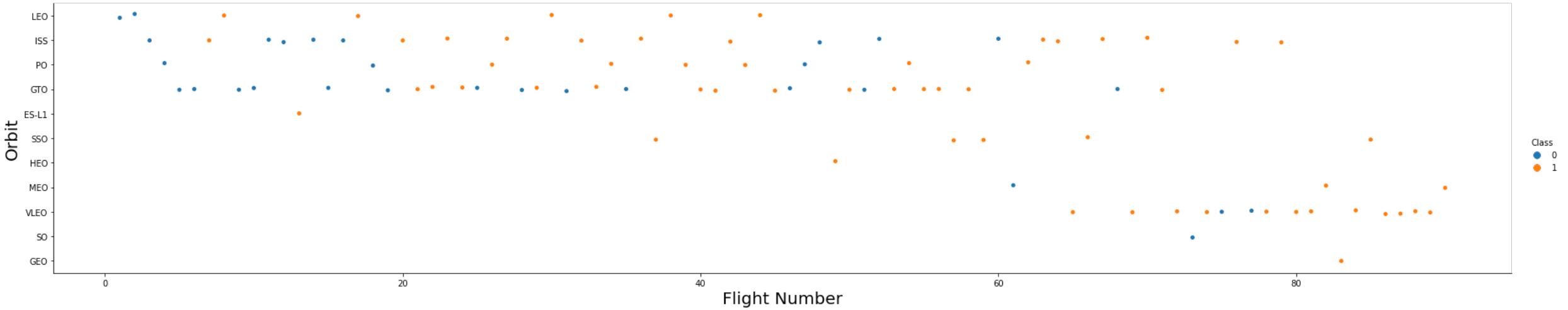
- It seems that as the payload mass increases, the success rate is much higher.
- There is no clear pattern to determine if the launch site is dependent on payload mass for successful launches.

SUCCESS RATE VS. ORBIT TYPE

- All first stages successfully landed from ES-L1, GEO, HEO and SSO orbits.
- It has never been successfully landed from the SO orbit.
- All other orbits had a landing success rate between 50% and 90%.

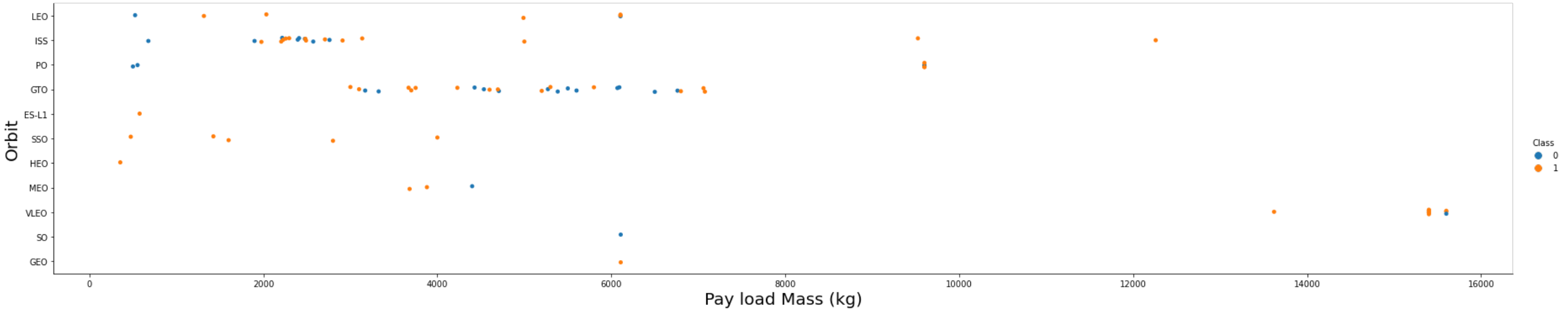


FLIGHT NUMBER VS. ORBIT TYPE



- It seems that GTO orbit has no correlation with flight number
- High flight numbers mainly go to GEO, SO, VLEO, MEO, HEO, SOO orbits
- It seems that as flight number increase on LEO orbit, the success rate is much higher.
- VLEO orbit has a high success rate

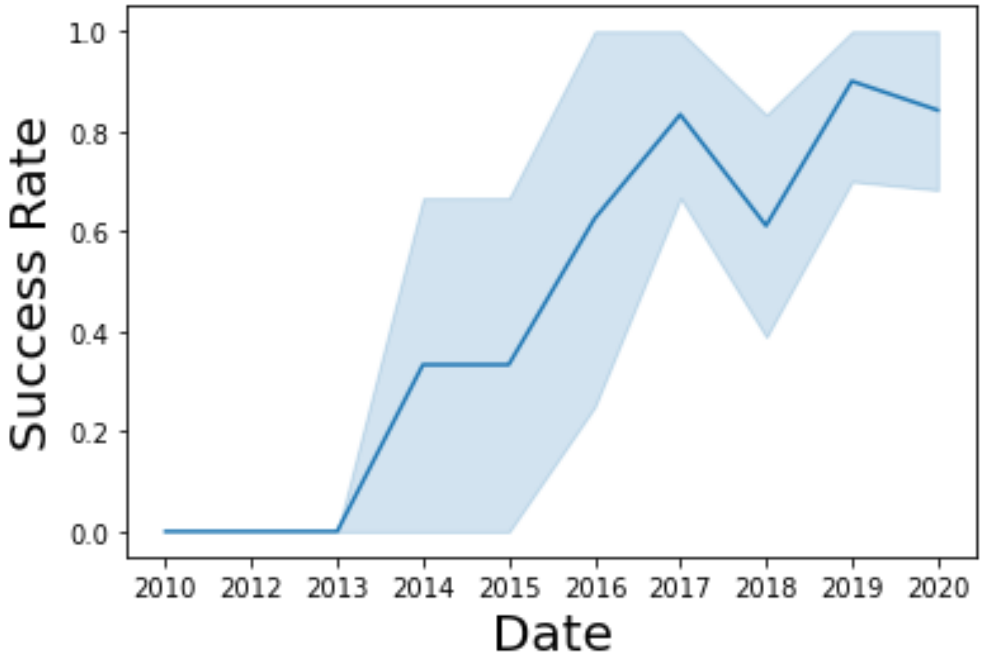
PAYLOAD VS. ORBIT TYPE



- ISS and LEO orbits have higher success rate when payload mass increase.
- It seems that GTO orbit has no correlation with payload mass
- SSO orbit has higher success rate when payload mass is lower

LAUNCH SUCCESS YEARLY TREND

- There are no successful landings before 2014
- Success rate continue to increase from 2014 to 2020



EDA WITH SQL

ALL LAUNCH SITE NAMES

SQL Query

```
select distinct launch_site from spacextbl
```

Query Result

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

Explanation

- We are looking for the names of the unique launch site in the space mission.
- Using `select distinct` statement to retrieve the unique launch site names.

LAUNCH SITE NAMES BEGIN WITH 'CCA'

SQL Query

```
select * from spacextbl
where launch_site like 'CCA%'
limit 5
```

Explanation

- We are looking for 5 records where launch sites begin with CCA.
- Using `like` operator to search a specific pattern in launch site column.
- Using `limit` clause to retrieve 5 records only

Query Result

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

SQL Query

```
select sum(payload_mass__kg_) total_payload_mass
from spacextbl where customer = 'NASA (CRS)'
```

Query Result

total_payload_mass
45596

Explanation

- We are looking for the total payload mass carried by boosters launched by NASA (CRS)
- Using `sum()` function to return the total payload mass
- Using the condition `customer = 'NASA (CRS)'` to filter the records by customer

AVERAGE PAYLOAD MASS BY F9 V1.1

SQL Query

```
select avg(payload_mass__kg_) avg_payload_mass
from spacextbl where booster_version = 'F9 v1.1'
```

Query Result

avg_payload_mass
2928

Explanation

- We are looking for the average payload mass carried by booster version F9 v1.1
- Using `avg()` function to return the average payload mass
- Using the condition `booster_version = 'F9 v1.1'` to filter the records by booster version

FIRST SUCCESSFUL GROUND LANDING DATE

SQL Query

```
select min(date) min_date
from spacextbl
where landing__outcome = 'Success (ground pad)'
```

Query Result

min_date
2015-12-22

Explanation

- We are looking for the date when the first successful landing outcome in ground pad was achieved.
- Using `min()` function to return the smallest value of date.
- Using the condition `landing__outcome = 'Success (ground pad)'` to filter the records by landing outcome.

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

SQL Query

```
select booster_version from spacextbl
where landing__outcome = 'Success (drone ship)'
and payload_mass__kg_ > 4000
and payload_mass__kg_ < 6000
```

Query Result

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

Explanation

- We are looking for the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Using the condition `landing__outcome = 'Success (drone ship)'` to filter the records by landing outcome.
- Using the condition `payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000` to filter the records by payload mass.

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

SQL Query

```
select
(case when mission_outcome like '%Success%' then
'Success' else 'Failure' end) mission_outcomes,
count(*) qty
from spacextbl
group by (case when mission_outcome like '%Success%' then
'Success' else 'Failure' end)
```

Query Result

mission_outcomes	qty
Failure	1
Success	100

Explanation

- We are looking for the total number of successful and failure mission outcomes.
- Using `count(*)` statement to retrieve the number of records.
- Using `group by` statement to group records by mission outcomes.

BOOSTERS CARRIED MAXIMUM PAYLOAD

SQL Query

```
select booster_version from spacextbl
where payload_mass__kg_ = (select
max(payload_mass__kg_) from spacextbl)
```

Explanation

- We are looking for the names of booster version which have carried the maximum payload mass.
- Using the subquery to get maximum payload mass.
- Using `max()` function to return the largest value of payload mass.

Query Result

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

SQL Query

```
select booster_version,launch_site from
spacextbl
where landing__outcome = 'Failure (drone ship)'
and year(date) = 2015
```

Query Result

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Explanation

- We are looking for the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- Using `year()` function to return the year part of a date.

2015 LAUNCH RECORDS

SQL Query

```
select landing__outcome,count(landing__outcome) qty
from spacextbl
where (date between '2010-06-04' and '2017-03-20')
group by landing__outcome order by 2 desc
```

Query Result

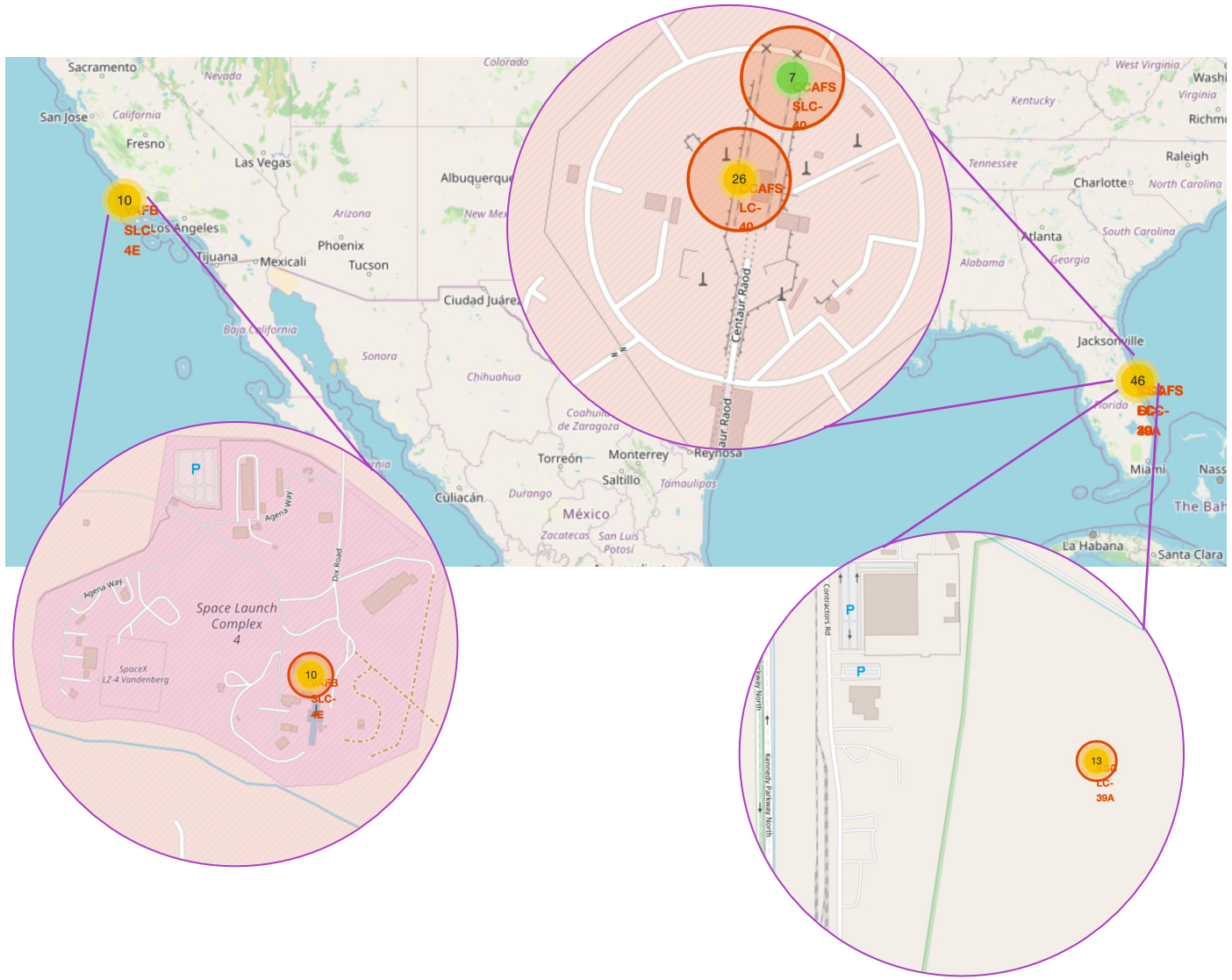
landing__outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Explanation

- We are ranking 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.
- Using **`group by`** statement to group records by landing outcomes.
- Using **`order by 2 desc`** keyword to sort by 2nd column in descending order

LAUNCH SITES PROXIMITIES ANALYSIS

SPACEX LAUNCH SITES



- **Most of launch sites are in proximity to the equator,**

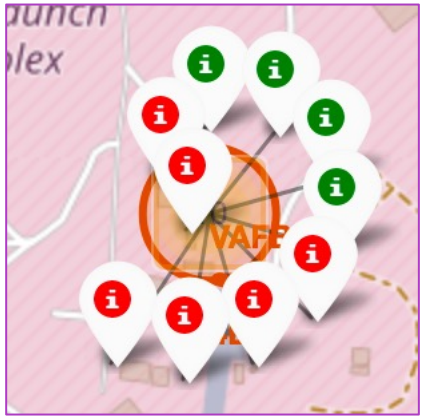
  because anything on the equator is moving at 1670 km/h, reducing the fuel consumption during launches.

- **All launch sites are in very close proximity to the coast,**

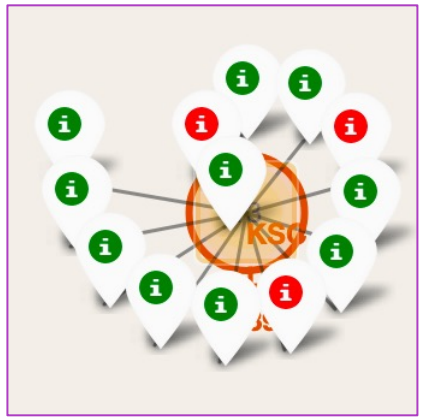
  in order to minimize the risks in the event of an incident.

SPACEX LAUNCH SITE SUCCESS RATE

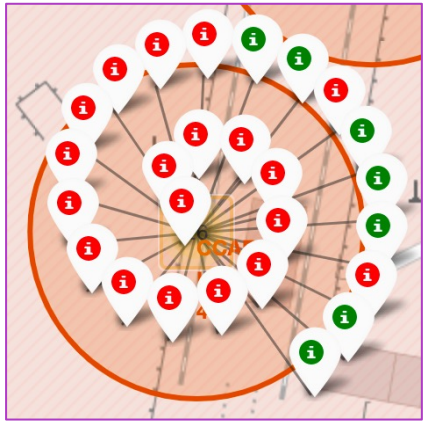
VAFB SLC-4E



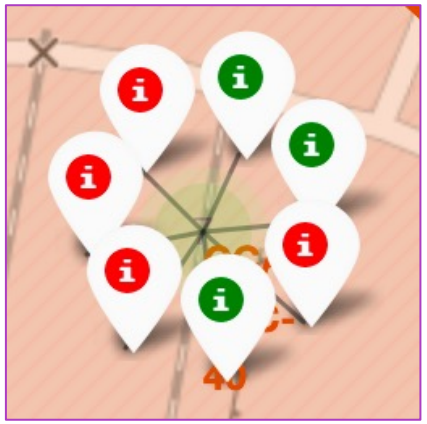
KSC LC-39A



CCAFS LC-40

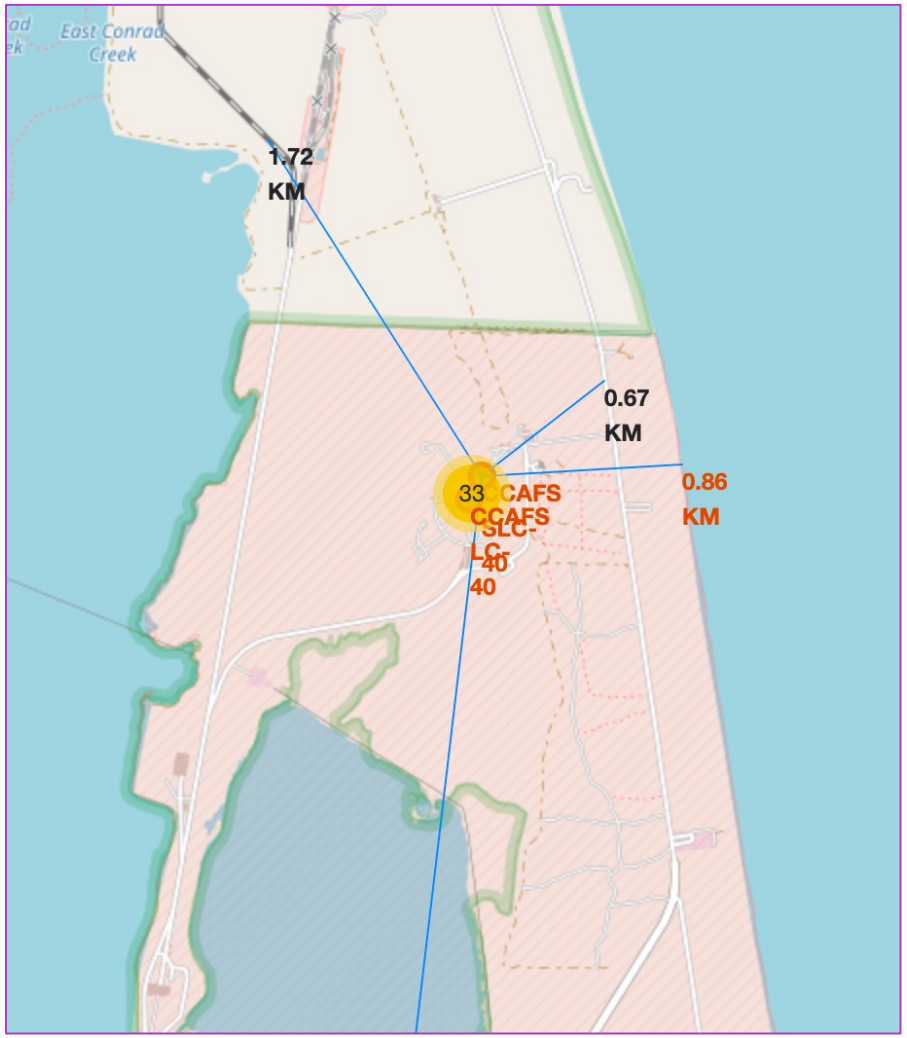


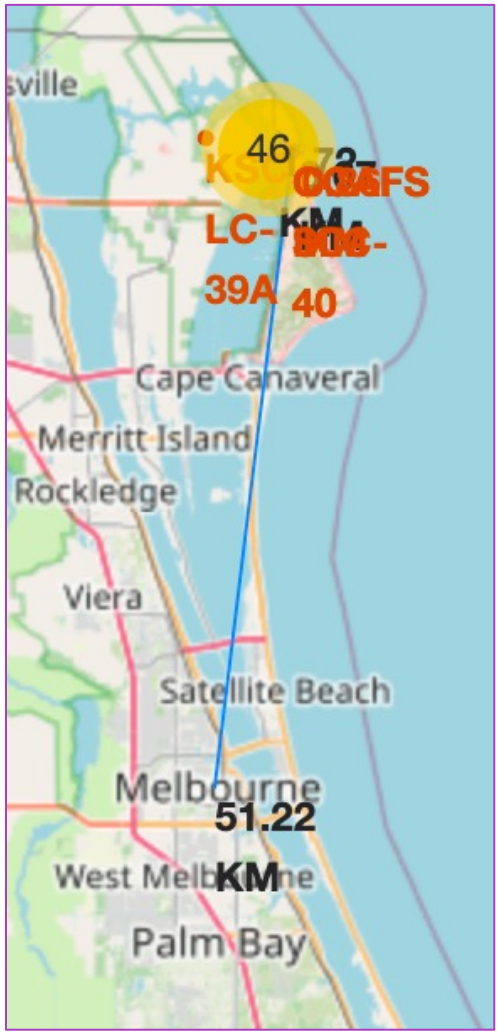
CCAFS SLC-40



- We added a marker for each landing outcomes by launch site.
- Green marker: successful landings
- Red marker: failed landings
- As screenshots shown, KSC LC-39A has the highest success rate

SPACEX LAUNCH SITE SUCCESS RATE

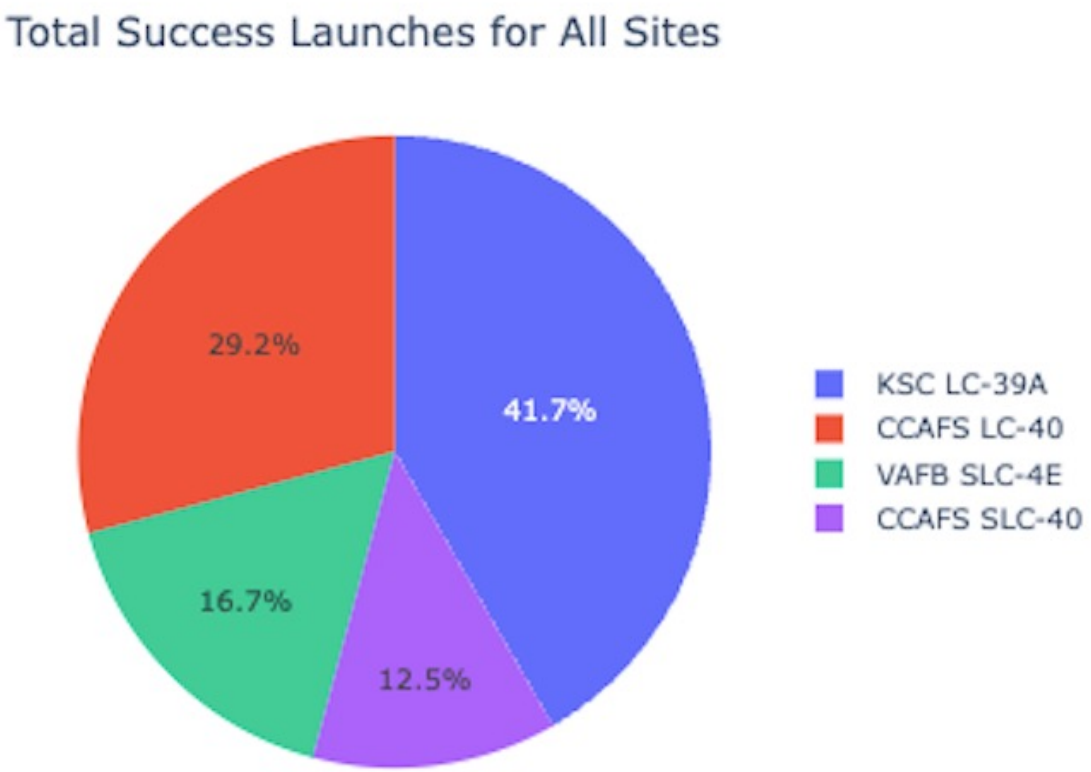




- CCAFS SLC-40 launch site is located away from populated areas and very close to the coast or the ocean, in order to minimize the risks in the event of an incident.
- The CCAFS SLC-40 launch site is located close to highway and railway, to provide access routes, either for logistical purposes or as an emergency evacuation route.

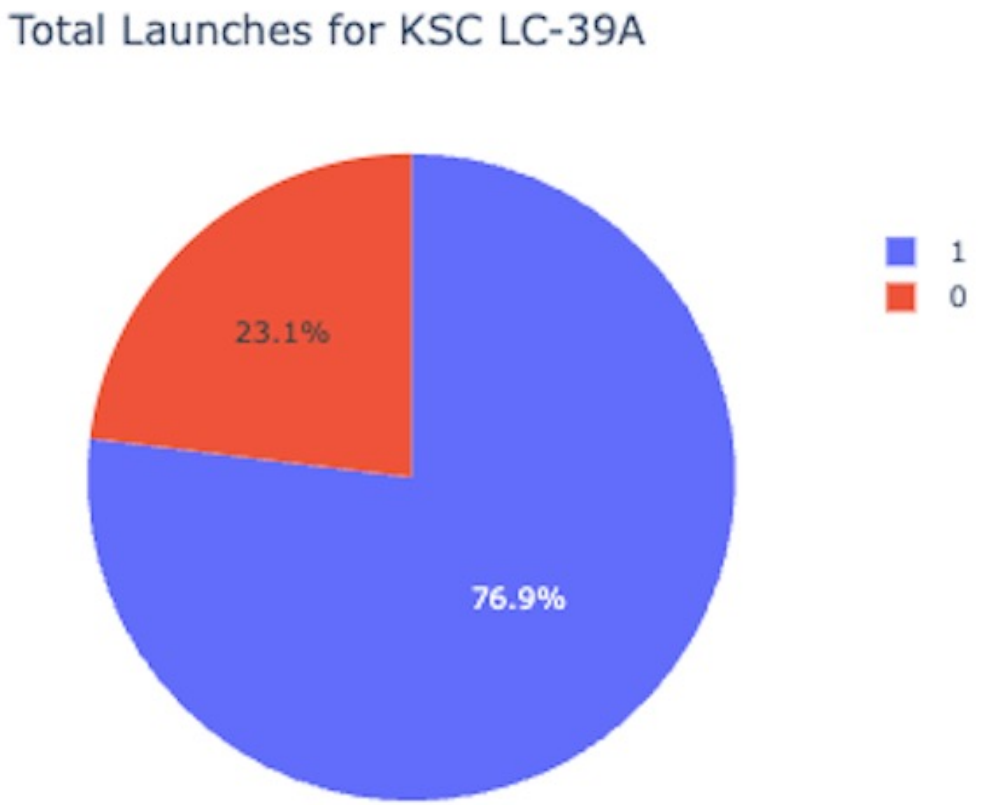

BUILD A DASHBOARD
WITH PLOTLY DASH

DASHBOARD
LAUNCH SUCCESS COUNT FOR ALL SITES



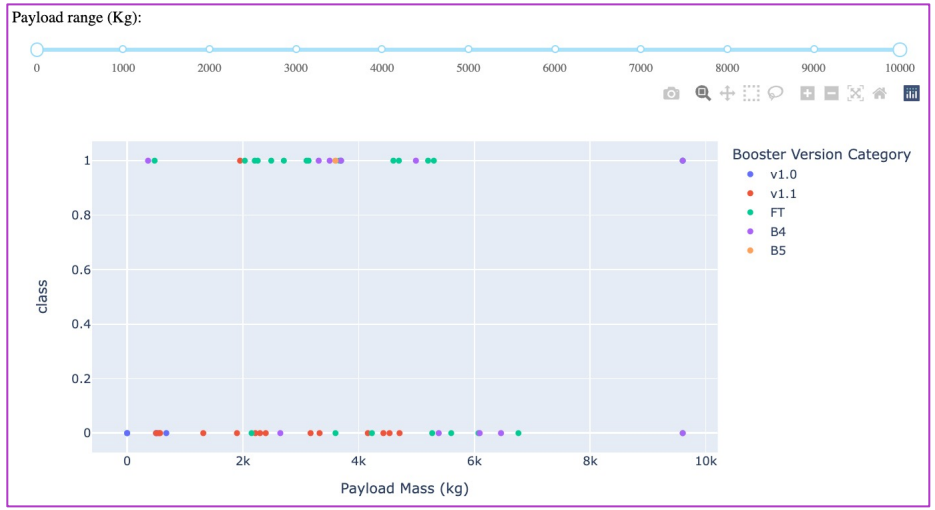
- The pie chart shows the total success launches for all sites.
- KSC LC-39A has the most successful launches

DASHBOARD
LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO

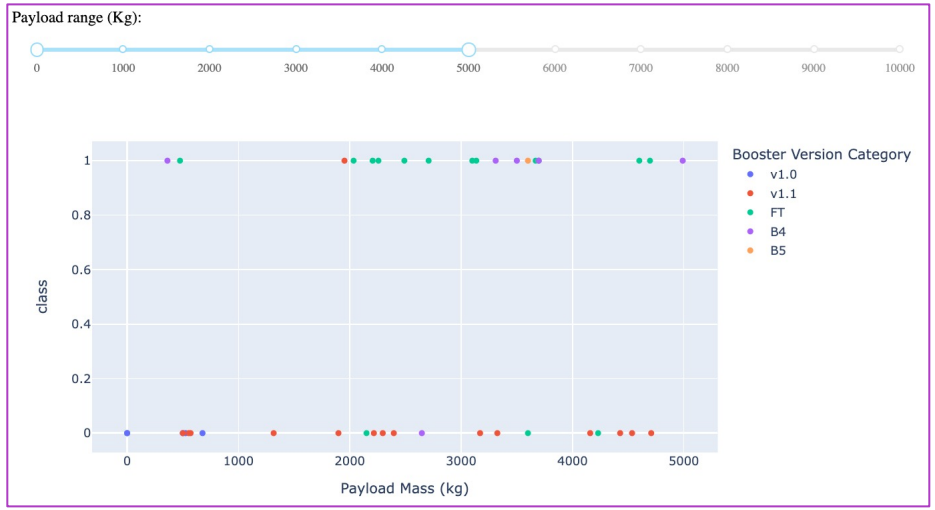


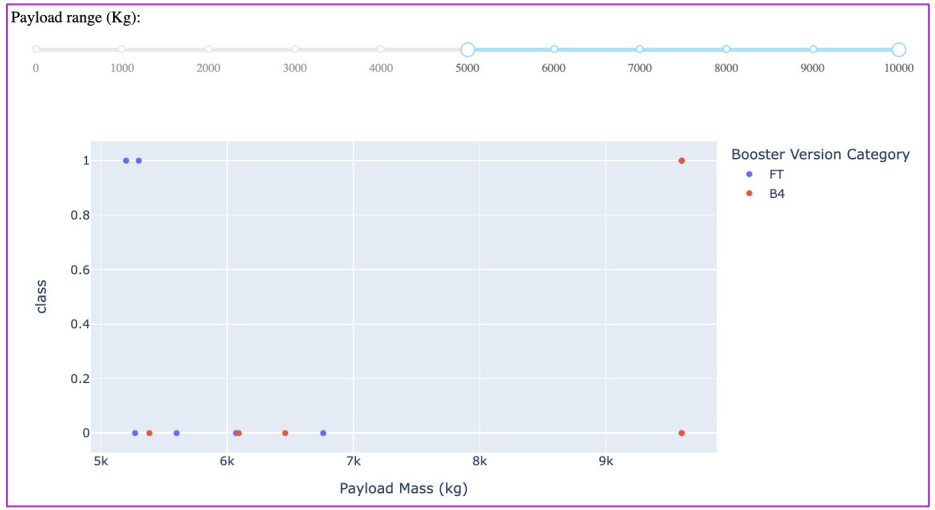
- The pie chart shows the ratio of successful and failed launches by launch site.
- KSC LC-39A launch site has the highest success rate at 76.9%.

DASHBOARD
PAYLOAD VS LAUNCH OUTCOME FOR ALL SITES



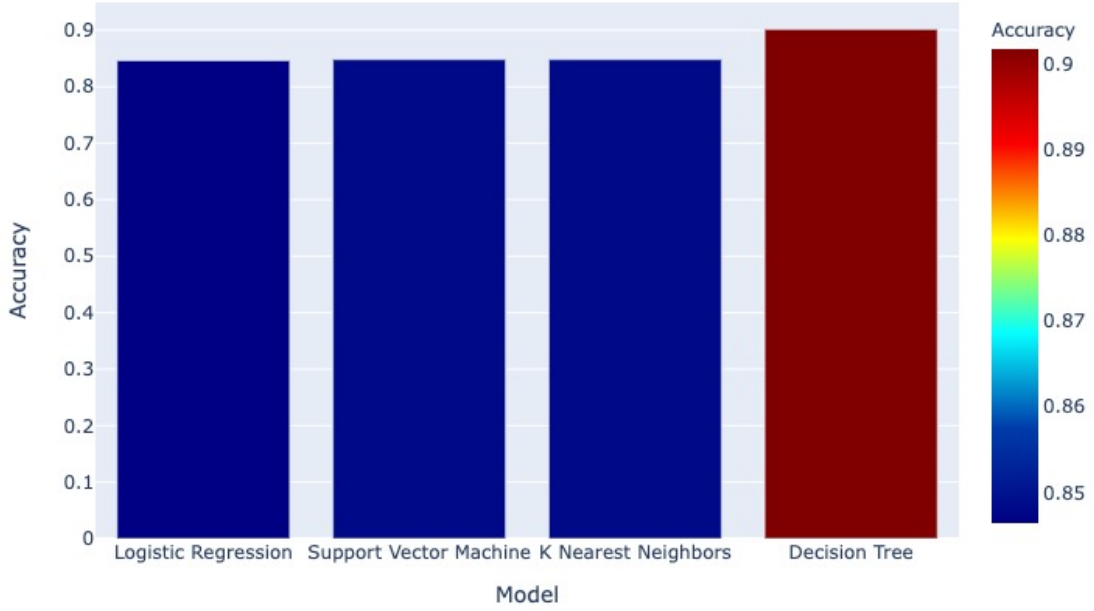
- The scatter plot shows the correlation between launch outcomes and payload mass by booster version.
- Payload mass up to 5000kg has a higher success rate than payload mass over 5000kg
- FT booster version has the highest launch success rate





PREDICTIVE ANALYSIS (CLASSIFICATION)

CLASSIFICATION ACCURACY

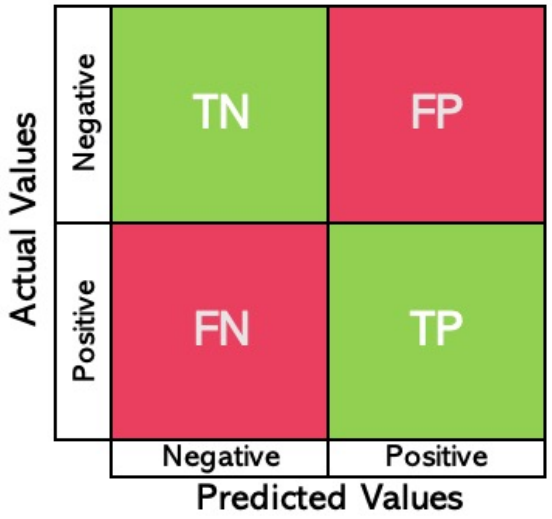


- All models have similar accuracy. However, the **Decision Tree** model performs the best with a score of 0.901786.
- After training four models, each achieved an 83.34% accuracy rate on the test data.

Model	Accuracy	Accuracy on Test Data	Tuned Hyperparameters (best parameters)
Logistic Regression	0.846429	0.833334	{'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
Support Vector Machine	0.848214	0.833334	{'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
K Nearest Neighbors	0.848214	0.833334	{'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
Decision Tree	**0.901786**	0.833334	{'criterion': 'entropy', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'random'}

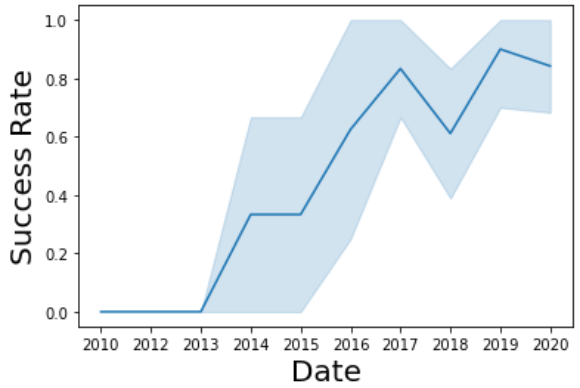
CONFUSION MATRIX

The Decision Tree model has the best balance between matches and false negatives / false positives.

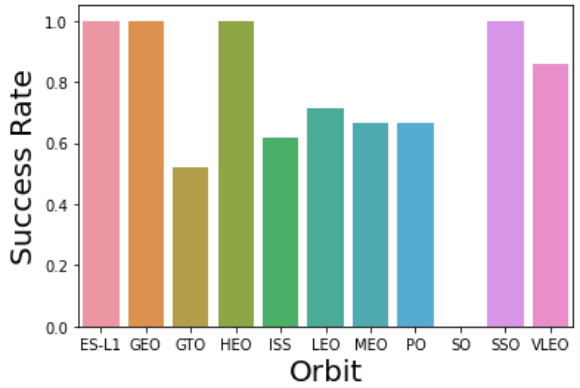


CONCLUSIONS

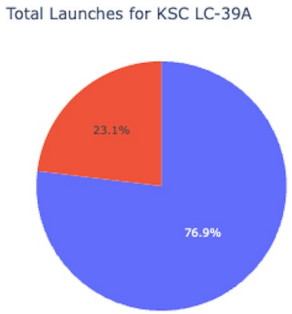
1. SpaceX has improved its launch success over time

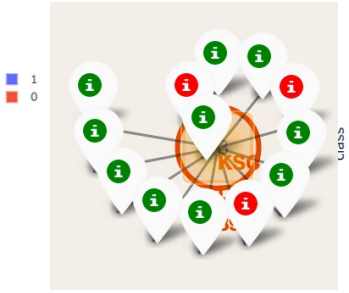


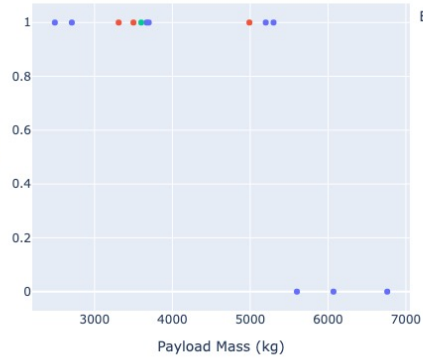
2. All first stages successfully landed from ES-L1, GEO, HEO and SSO orbits



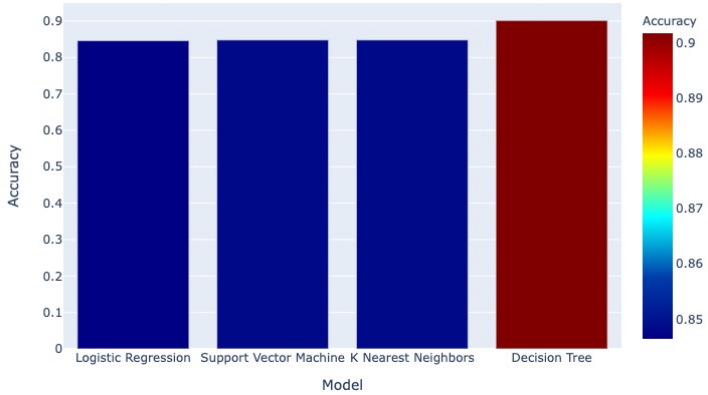
3. KSC LC-39A launch site has the highest success rate, but increasing payload mass seems to have negative impact on success







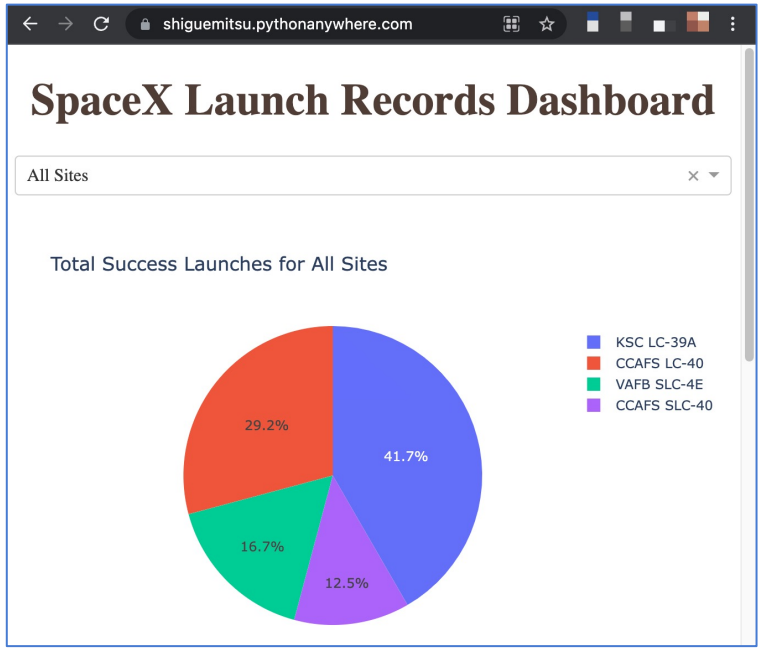
4. Decision Tree model performs the best with a score of 0.901786



Model	Accuracy
Logistic Regression	0.846429
Support Vector Machine	0.848214
K Nearest Neighbors	0.848214
Decision Tree	0.901786

APPENDIX

- I have posted the interactive dashboard via PythonAnywhere.com
- The live website URL is: https://shiguemitsu.pythonanywhere.com/



Thank you!