IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Emmanuel Effanga
9 January, 2024

Outline

Executive Summary

Summary of methodologies

- The IBM Data Science Capstone project aimed to predict if the Falcon 9 first stage will land successfully using advanced data science methodologies. The project followed a comprehensive and structured approach, encompassing the following key methodologies:
    - Data Collection via API
    - Web Scraping
    - Exploratory Data Analysis (EDA) with Data Visualization
    - EDA with SQL
    - Interactive Map with Folium
    - Dashboards with Plotly Dash
    - Predictive Analysis

Summary of all results

- Exploratory Data Analysis results: Through in-depth analysis, significant insights and patterns within the data were uncovered
- Interactive maps and dashboard: Uncovered profound insights within the data, visually represented through interactive maps and dashboards for enhanced comprehension
- Predictive results: The selected model(s) demonstrated high predictive accuracy, effectively addressing the project's primary objectives.

Introduction

Project background and context

- In the realm of space exploration, the cost-effectiveness of rocket launches is a critical factor that can revolutionize the aerospace industry. At the forefront of this endeavor is SpaceX, renowned for its innovative approach in reusing the Falcon 9 rocket's first stage, thereby significantly reducing launch costs. The ability to predict the successful landing of the Falcon 9's first stage emerges as a pivotal element in determining the overall cost of a launch.
- SpaceX, on its website, advertises Falcon 9 rocket launches at a competitive cost of 62 million dollars, a stark contrast to other providers whose offerings soar upwards of 165 million dollars each. This substantial cost disparity primarily stems from SpaceX's groundbreaking achievement of reusing the first stage of the Falcon 9. The potential savings are not merely financial; they represent a paradigm shift in the economics of space travel.

Problems you want to find answers

- Determine if the Falcon 9 first stage will land successfully
- Determine the cost of each launch by gathering information about Space X

Section 1

Methodology

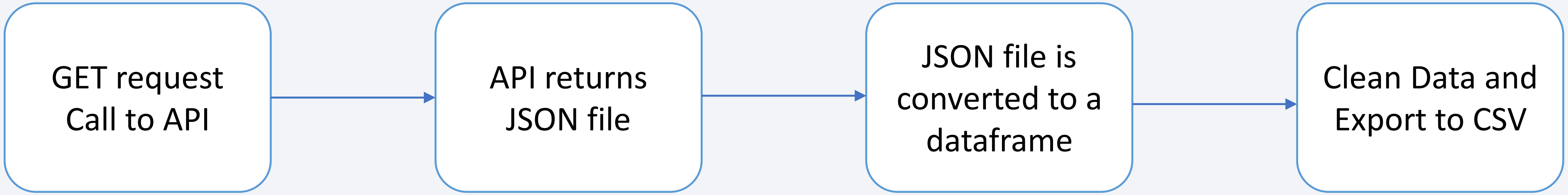
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by sending a GET request to SpaceX REST API
  - Webscraping data from Wikipedia
- Perform data wrangling
  - Transformed the JSON file from the API to a pandas dataframe
  - Removed unnecessary columns and replaced null values with its mean
- 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

- Datasets were collected through the SpaceX REST API available on https://api.spacexdata.com/v4. The data collected from the API were information about past launches, rockets, launchpads, cores and payload data.



- Data was also collected by web scraping the SpaceX Wikipedia sites on https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches using BeautifulSoup python library

Data Collection – SpaceX API

REST API Call

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

Decode response as JSON and convert to a dataframe

```
data = pd.json_normalize(response.json())
```

Filter the dataframe to only include Falcon9 information

```
data_falcon9 = df[(df['BoosterVersion']!='Falcon 1')]
```

Replace missing values with mean

```
mean = data_falcon9["PayloadMass"].mean()
data_falcon9["PayloadMass"].replace(np.nan,mean)
```

Export to csv

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

[Link to Github]()

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- https://github.com/embee430/Data-Projects/blob/d863caee553f532f0143b4389a772002a80e3833/jupyter_labs_webscraping.ipynb

[Link to Github](https://github.com/embee430/Data-Projects/blob/d863caee553f532f0143b4389a772002a80e3833/jupyter_labs_webscraping.ipynb)

Request the Falcon 9 webpage
req = requests.get(static_url)

↓

Create a beautifulsoup object from the html response
soup = BeautifulSoup(req.text, "html.parser")

↓

Extract all table from the html response and assign the third table to a variable

↓

Extract all columns header (th) and data (td) into an empty dictionary and using for loops, extracts data into a dataframe

↓

Export to csv

Data Wrangling

- In the data set, there are several different cases where the booster landed successfully or did not land successfully.
  - True Ocean, True RTLS and True ASDS means the mission outcome was successfully landed to a specific region
  - False Ocean, False RTLS and False ASDS means the mission outcome was unsuccessfully landed to a specific region
- In the data wrangling stage, we converted the outcome into categorical variable 1 and 0 for our model.

[Link to Github]

Calculated number of launches on each launch sites

↓

Calculates the number of occurrence of each orbits

↓

Calculate the number and occurences of mission outcome per orbit type

Create a landing outcome label from Outcome column and append 1 for successful and 0 for unsuccessful

↓

Export the wrangled data to csv

EDA with Data Visualization

The following graphs were used for exploring the SpaceX data in visual form using matplotlib and seaborn libraries.

Scatter Plots

Scatter plots show relationships between variables as follows:

- ❑ Flight Number and Payload Mass
- ❑ Flight Number and Launch Site
- ❑ Payload and Launch Site
- ❑ Flight Number and Orbit type
- ❑ Payload and Orbit type

Bar Graph

Bar graphs show the relationship between numeric and categoric variables

- ❑ Success rate of each orbit type

Line Graph

Line graphs show data variables and their trends

- ❑ Launch success yearly trend

[Link to Github]

EDA with SQL

We performed SQL queries to explore and understand the data based on the questions:

- Displaying the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin with the string 'CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List the names of the booster_versions which have carried the maximum payload mass.
- List the records which will display the month names, faiilure landing_ouutcomes in drone ship, booster versions, launch_site for the months in year 2015.
- Rank the count of successful landiing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

[Link to Github]

Build an Interactive Map with Folium

Folium map object is a map centered on NASA Johnson Space Center at Houson, Texas

- Red circle at NASA Johnson Space Center's coordinate with label showing its name (folium.Circle, folium.map.Marker).
- Red circles at each launch site coordinates with label showing launch site name (folium.Circle, folium.map.Marker,
- folium.features.DivIcon).
- The grouping of points in a cluster to display multiple and different information for the same coordinates
- (folium.plugins.MarkerCluster).
- Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing.
- (folium.map.Marker, folium.Icon).
- Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them.
- (folium.map.Marker, folium.PolyLine, folium.features.DivIcon)

[Link to Github]()

Build a Dashboard with Plotly Dash

A dashboard was created with these components, a dropdown, pie chart, range slider and scatter plot components.

- Dropdown allows a user to choose the launch site or all launch sites (dash_core_components.Dropdown).
- Pie chart shows the total success and the total failure for the launch site chosen with the dropdown component (plotly.express.pie).
- Rangeslider allows a user to select a payload mass in a fixed range (dash_core_components.RangeSlider).
- Scatter chart shows the relationship between two variables, in particular Success vs Payload Mass (plotly.express.scatter)Explain why you added those plots and interactions

[Link to Github]()

Predictive Analysis (Classification)

Data Preparation
Loaded datasets and normalized it
Splits the dataset into test and training set

Model Development
Developed models using different classification models
Trained the model with the training datasets
Used GridSearchCV on all models to find hyperparameters that allows the models to perform best

Model Evaluation
Tested each model with the test data and computed accuracy score with a confusion matrix

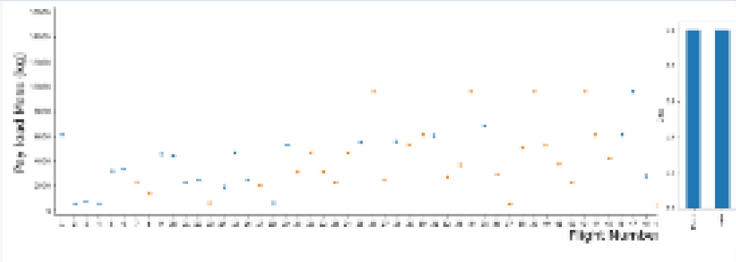
Model Comparison
Compared and chose the model with the best accuracy score

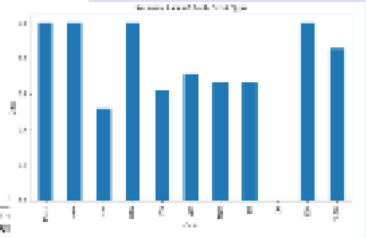
Link to github

Results

Exploratory data analysis results

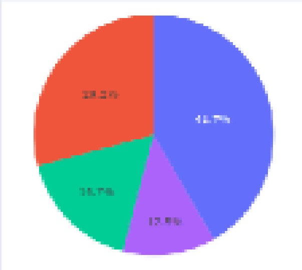
- Samples:

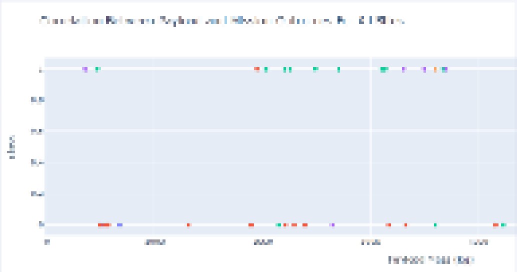




Interactive analytics demo in screenshots

- Samples





Predictive analysis results

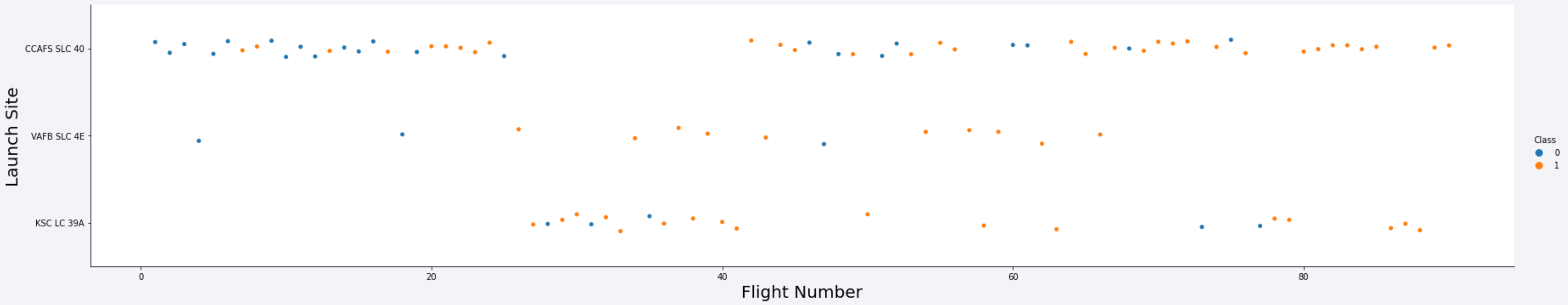
- Samples

	Algo Type	Accuracy Score
2	Decision Tree	0.903571
3	KNN	0.848214
1	SVM	0.848214
0	Logistic Regression	0.846429

Section 2

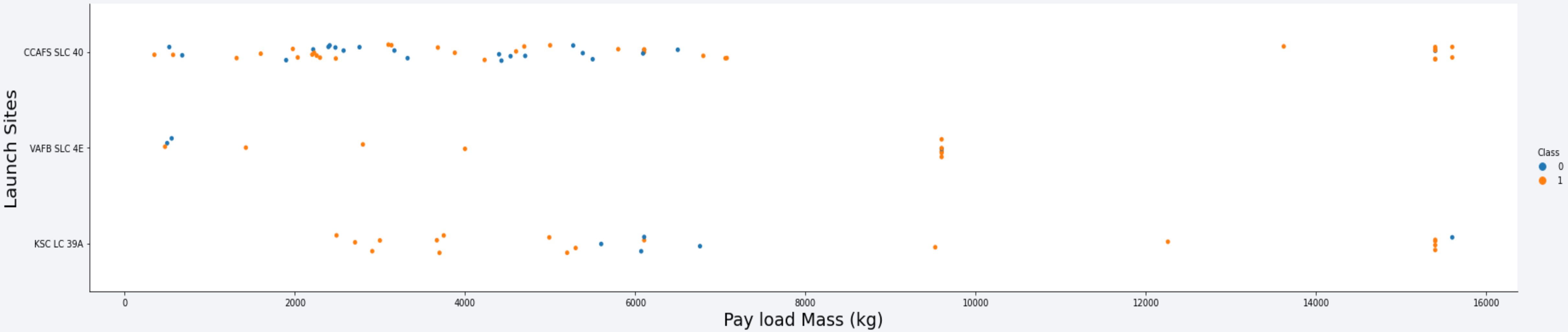
Insights drawn from EDA

Flight Number vs. Launch Site



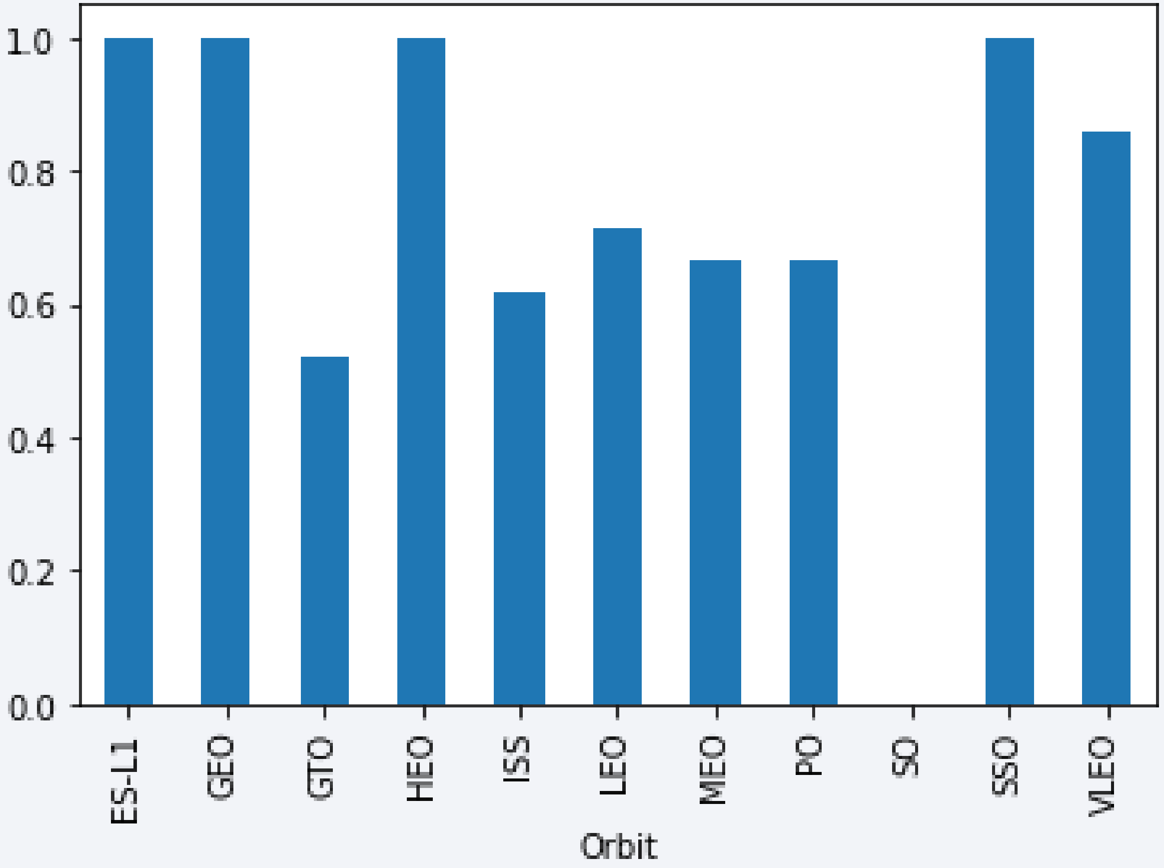
We observed that the launch site CCAFS SLC 40 recorded less success rate compared to other launch sites and also had more launches than other sites.

Payload vs. Launch Site



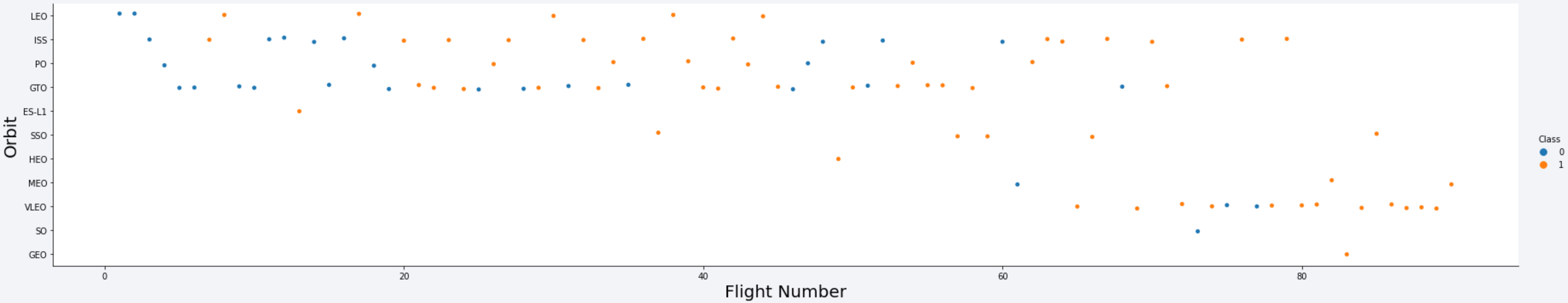
We observed that heavier payload of the rockets were successful from all launch sites

Success Rate vs. Orbit Type



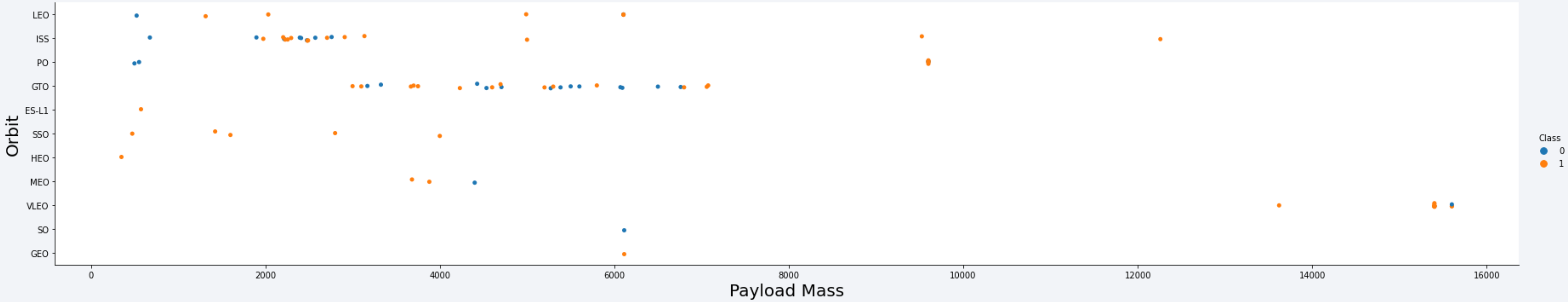
From the chart, we noted launches to the orbit ES-L1, GEO, HEO, SSO had the best success rate.

Flight Number vs. Orbit Type



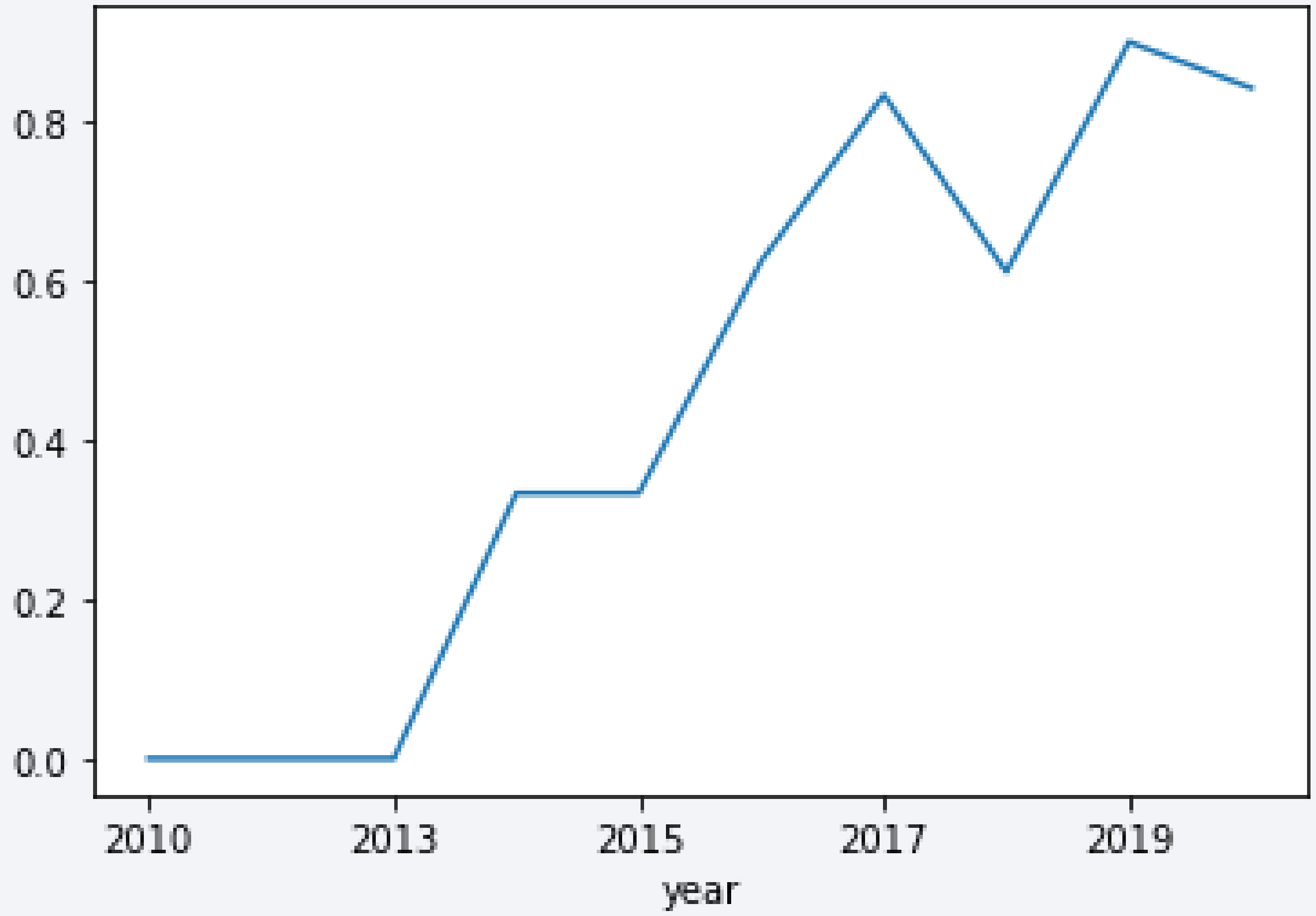
From the chart, we noted that higher launches were more concentrated on the Very Low Earth Orbit (VLEO) and Middle Earth orbit

Payload vs. Orbit Type



We observed that launches of heavier payloads were on the Very Low Earth Orbit (VLEO)

Launch Success Yearly Trend



From the line chart, we can see that the success rates dramatically increased since 2013

All Launch Site Names

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

We selected the distinct values from the launch sites which return 4 distinct launch sites.

```
%sql select distinct launch_site from SPACEX
```

```
 * ibm_db_sa://tgb63017:***@824dfd4d-99de-440d-9
Done.
```

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

Launch Site Names Begin with 'CCA'

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

```
 * ibm_db_sa://tgb63017:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
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

The query returned 5 recorded launches with launch sites names that starts with the string 'CCA'.

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as Total_payload from spacex where customer='NASA (CRS)'
```

```
 * ibm_db_sa://tgb63017:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
Done.
```

total_payload
45596

The total payload mass where the customer is NASA(CRS) is 45,596kg. The query summed all payloads by this customer.

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) from spacex where booster_version='F9 v1.1'
```

```
 * ibm_db_sa://tgb63017:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/BLUDB
Done.
```

1
2928

The query returned 2,928kg as the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

- The first successful landing on ground pad was recorded on the 22^{nd} of December, 2015.

- The date was retrieved using the query

select min(date) from SPACEX where landing__outcome = 'Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are as follows:
- F9 FT B1022
- F9 FT B1026
- F9 FT B1021.2
- F9 FT B1031.2

- The query **select booster_version from spacex where landing__outcome='Success (drone ship)' and payload_mass__kg_ between 4000 and 6000** was used

Total Number of Successful and Failure Mission Outcomes

Failure (in flight):	1
Success:	99
Success (payload status unclear):	1

- The query below grouped the total mission outcomes by the outcome itself.

select mission_outcome, count(mission_outcome) from spacex group by mission_outcome

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

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

- The query below was mixed with a subquery to select the booster version that carried the max payload mass.

select booster_version from spacex where payload_mass__kg_ = (select max(payload_mass__kg_)from spacex)

2015 Launch Records

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

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- The query selected the date, booster version, launch site of failed landing outcome on drone ship

SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM spacex WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(date) = 2015

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

landing__outcome	count_launches
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

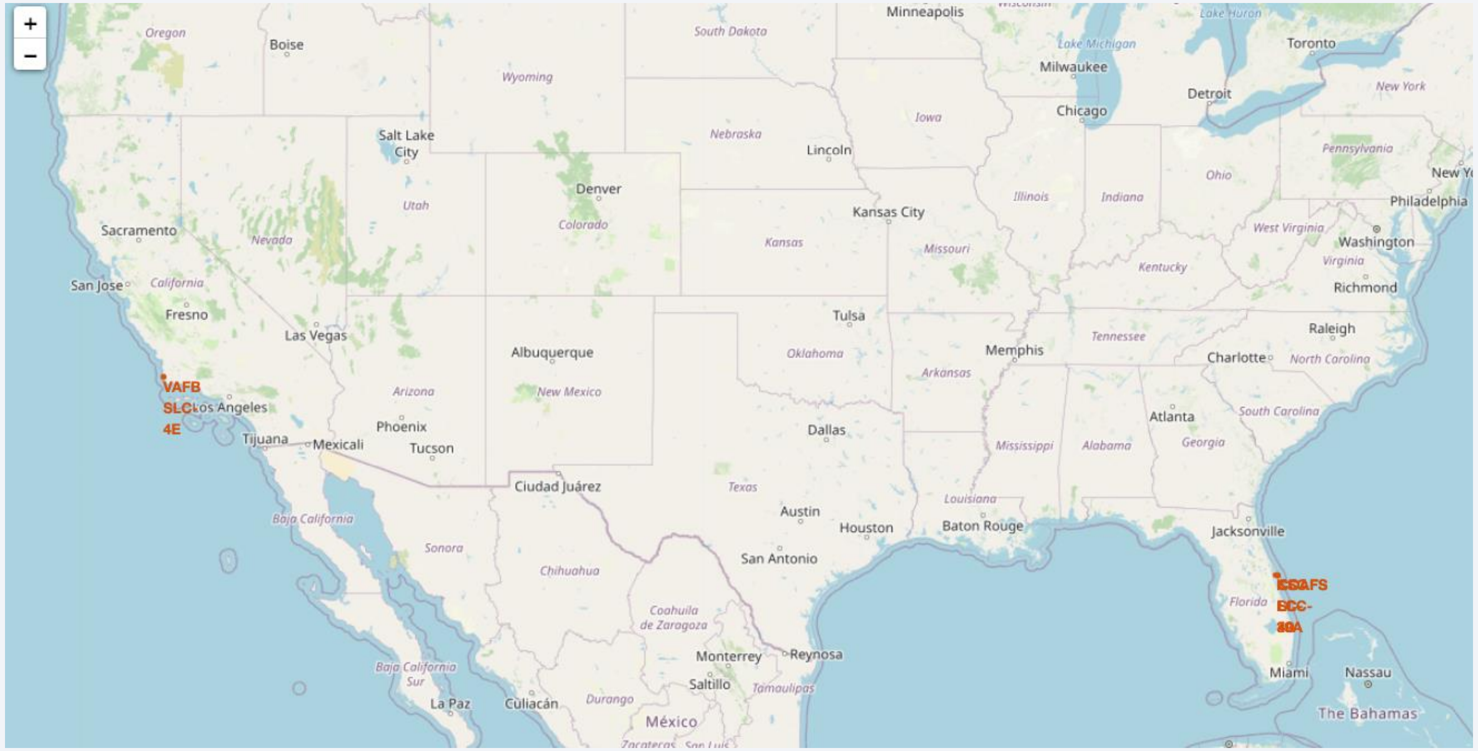
- The query grouped landing outcomes counts in descending order

SELECT LANDING__OUTCOME, COUNT(*) AS COUNT_LAUNCHES FROM spacex WHERE

Section 3

Launch Sites Proximities Analysis

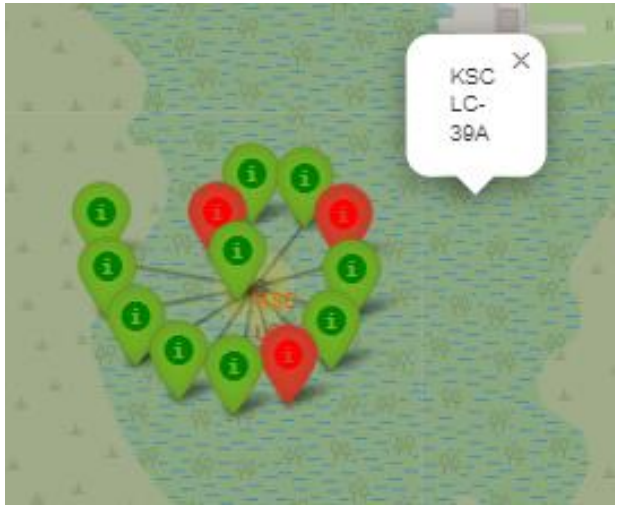
Launch Sites Locations



- 4 distinct launch sites locations were used for the launches. One (1) in Vandenberg Space Force Base LA, 2 in Cape Canaveral Space Force Station and 1 in a creek all in close proximity to the coastline for safety reasons.

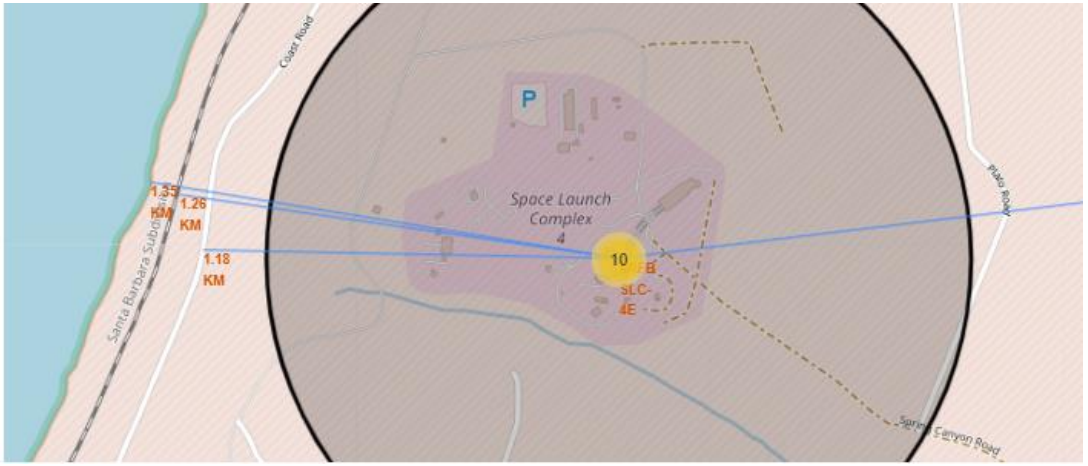
Successful/Failed launches across all launch sites

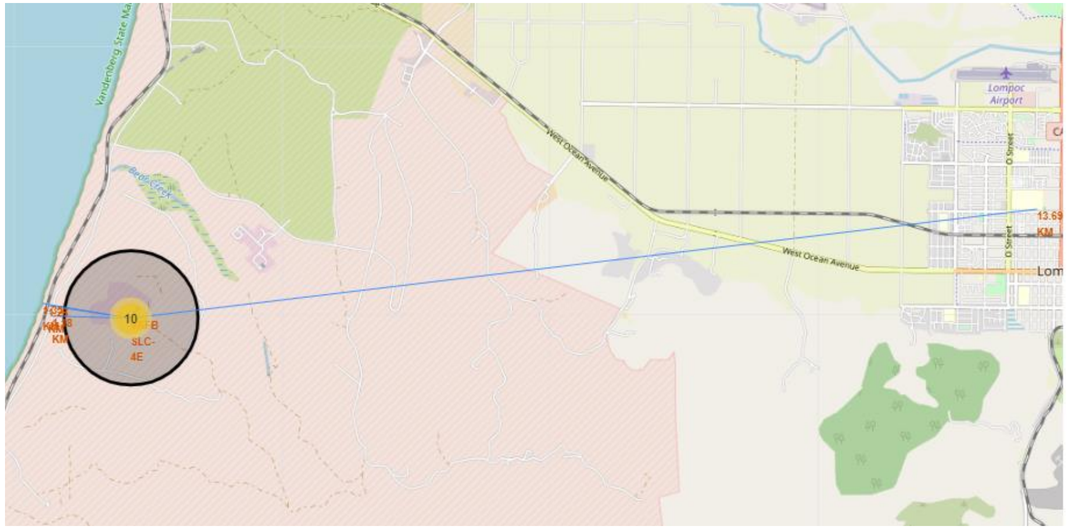
- Launch site KSC LC-39A had the most success rate of launches.



Launch Site Distance to Railways, Highways, Coastline and Closest City

- The launch site is 1.26km to the railways, 1.18km to the highways, 1.35km to the coastline and 13.69km to the closest city, Lompoc.

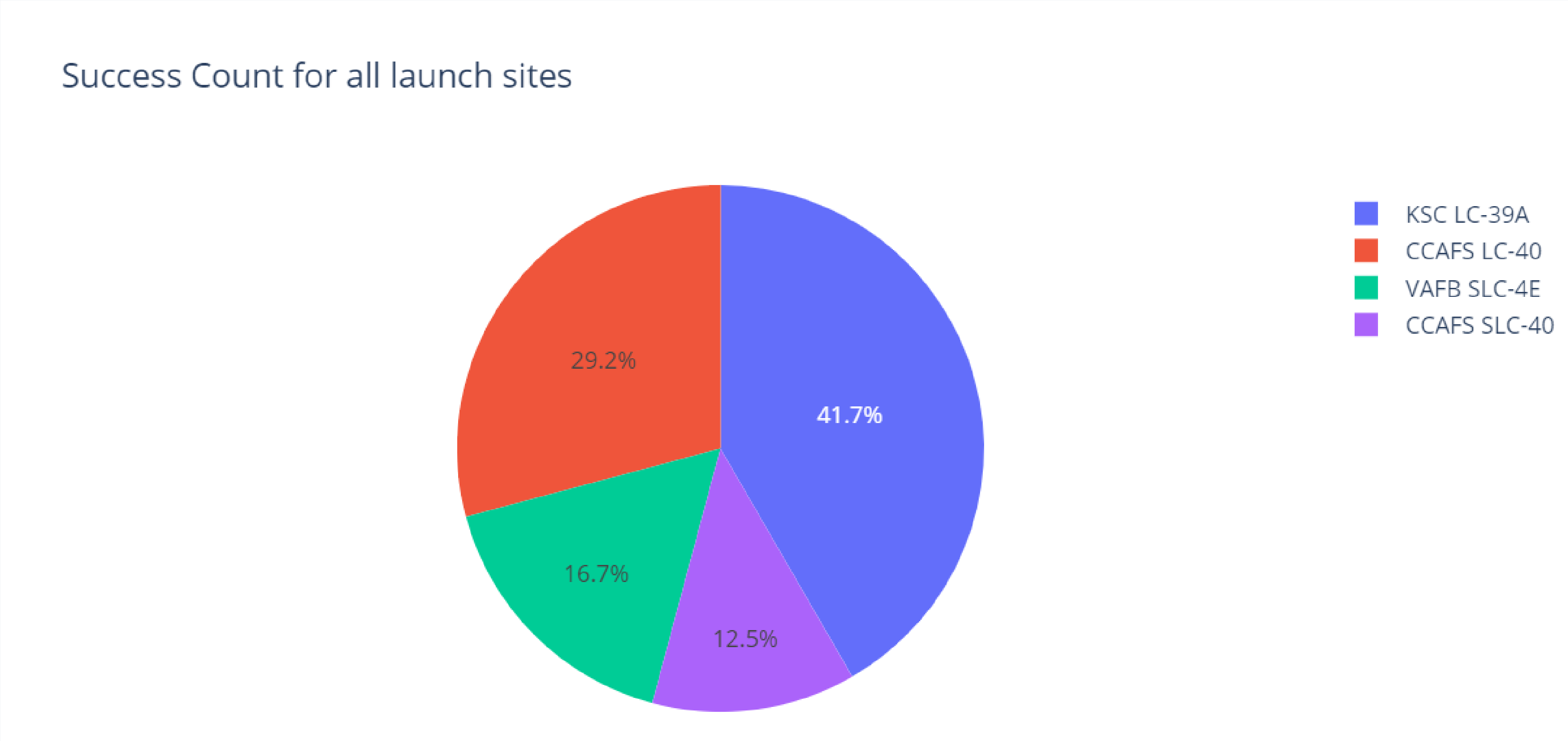




Section 4

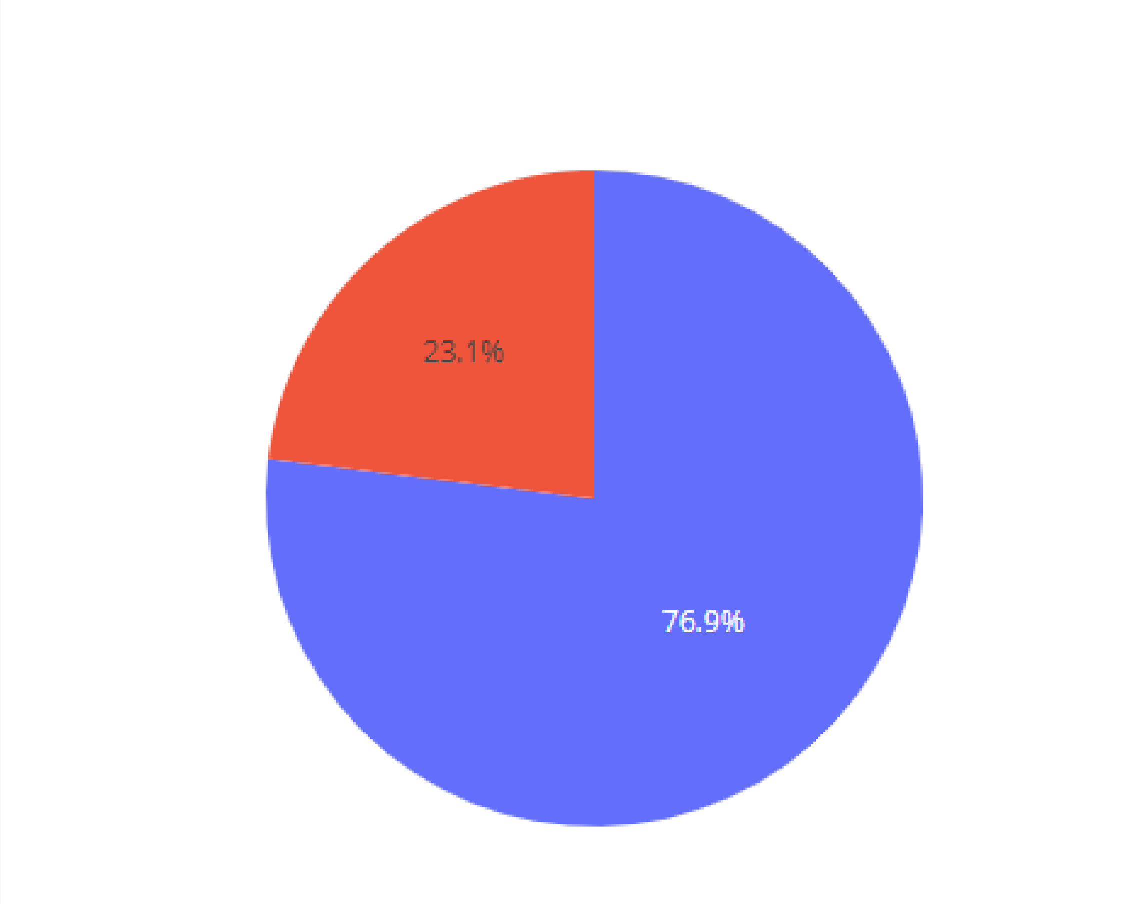
Build a Dashboard with Plotly Dash

Success Count for all Launch Sites



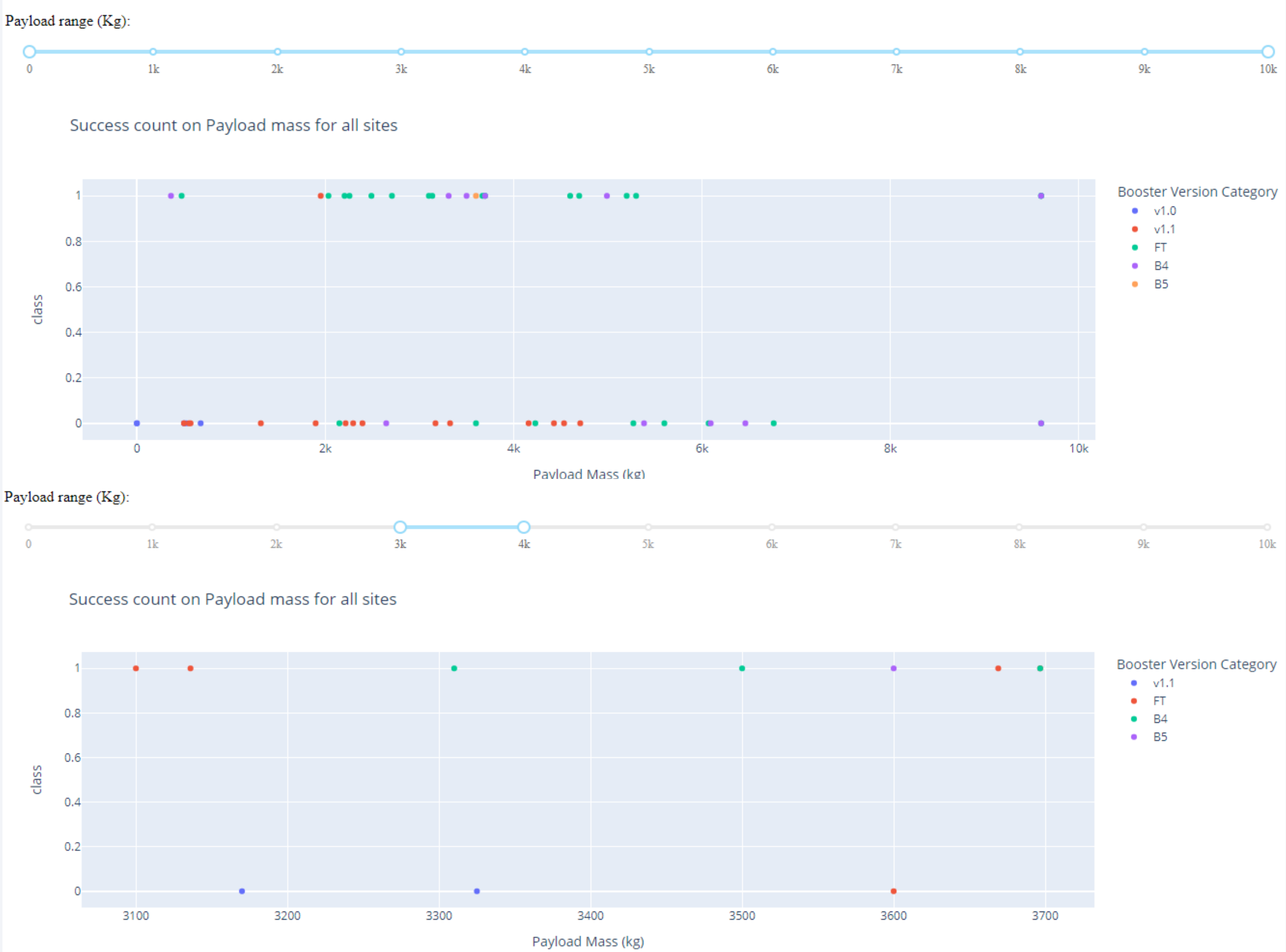
Launch site KSC LC-39A recorded the most successful launch followed by launch site CCAFS LC-40

Launch Site with highest success rate



- Launch site KSC LC-39A had the highest launch success rate with 76.9% success rate and 23.1% unsuccessful launches.

Payload vs Launch Outcome



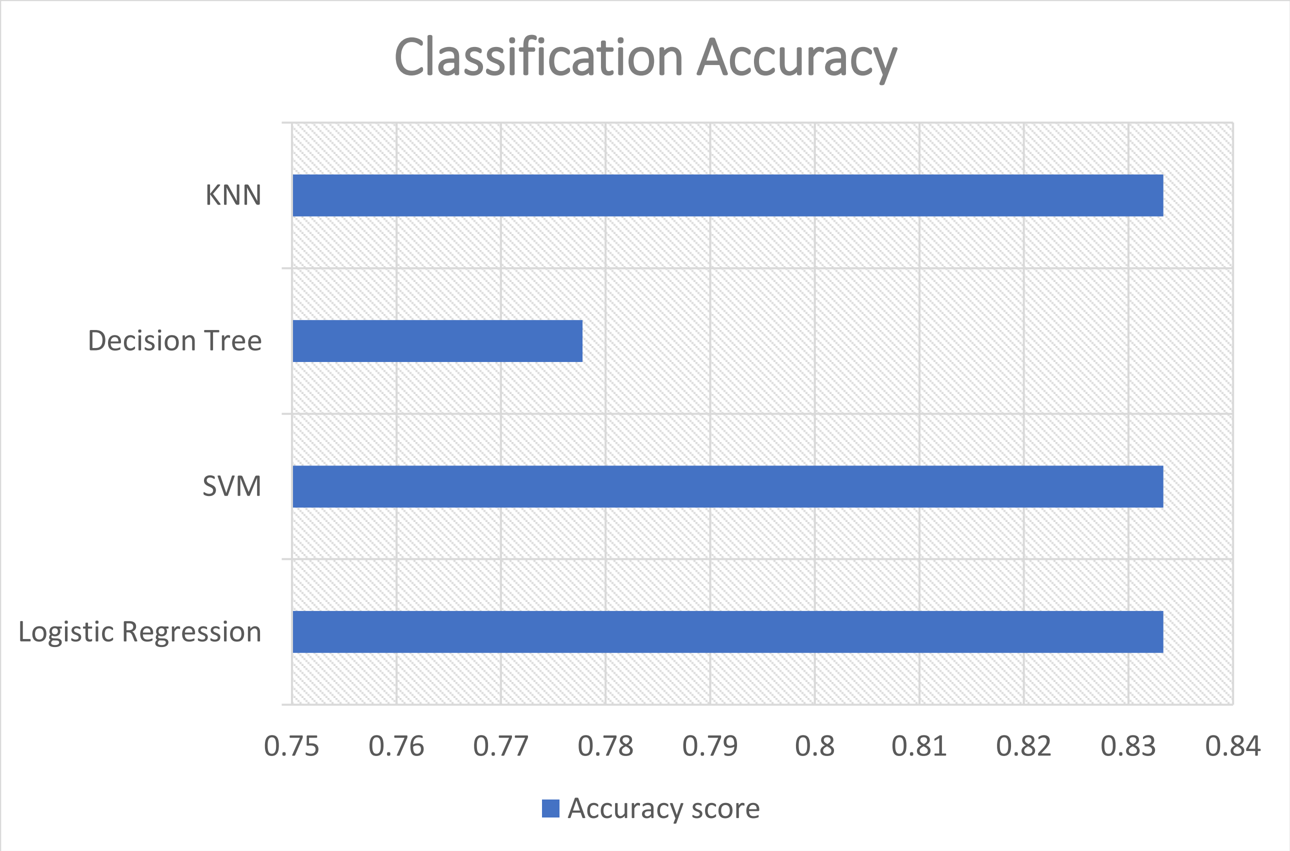
- Payload range of 3000kg – 4000kg had the highest launch successes.
- Lower payload range recorded the most unsuccessful launches.
- F9 Booster version 'FT' had the highest launch success rate

Section 5

Predictive Analysis (Classification)

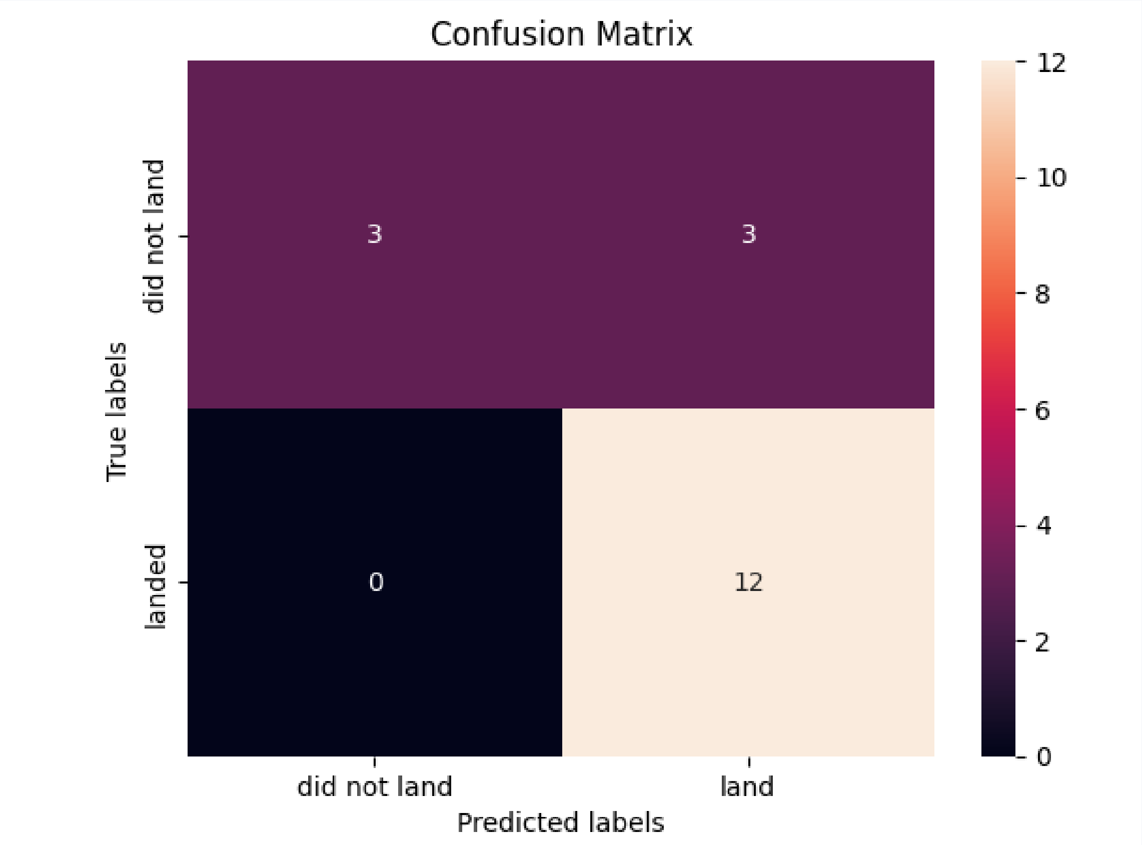
Classification Accuracy

- All the models have the same accuracy score of 83.33% except for the Decision Tree model



Confusion Matrix

- All the models performed the same in predicting if Falcon9 will land or not with accuracy score of 83%



Conclusions

- Successful launches can be achieved by using payload mass between 3000kg to 4000kg
- The chances of the model accurately predicting if the booster with land or not is 83.33%, meaning there is a slight chance of the model to fail in its prediction.
- Successful launches can be achieved by launching the rockets in launch site KSC LC-39A

Appendix

Sample rainfall data in Australia

	Date	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	...	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
0	2/1/2008	19.5	22.4	15.6	6.2	0.0	W	41	S	SSW	...	92	84	1017.6	1017.4	8	8	20.7	20.9	Yes	Yes
1	2/2/2008	19.5	25.6	6.0	3.4	2.7	W	41	W	E	...	83	73	1017.9	1016.4	7	7	22.4	24.8	Yes	Yes
2	2/3/2008	21.6	24.5	6.6	2.4	0.1	W	41	ESE	ESE	...	88	86	1016.7	1015.6	7	8	23.5	23.0	Yes	Yes
3	2/4/2008	20.2	22.8	18.8	2.2	0.0	W	41	NNE	E	...	83	90	1014.2	1011.8	8	8	21.4	20.9	Yes	Yes
4	2/5/2008	19.7	25.7	77.4	4.8	0.0	W	41	NNE	W	...	88	74	1008.3	1004.8	8	8	22.5	25.5	Yes	Yes

Sample launch data

	Flight Number	Date	Time (UTC)	Booster Version	Launch Site	Payload	Payload Mass (kg)	Orbit	Customer	Landing Outcome	class	Lat	Long
0	1	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Failure (parachute)	0	28.562302	-80.577356
1	2	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel o...	0.0	LEO (ISS)	NASA (COTS) NRO	Failure (parachute)	0	28.562302	-80.577356
2	3	2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2+	525.0	LEO (ISS)	NASA (COTS)	No attempt	0	28.562302	-80.577356
3	4	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	No attempt	0	28.562302	-80.577356
4	5	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	No attempt	0	28.562302	-80.577356

Appendix

Sample code snippets

```
#Enter Your Code, Execute and take the Screenshot
x_train, x_test, y_train, y_test = train_test_split( features, Y, test_size=0.2, random_state=10)
```

```
KNN_Accuracy_Score = accuracy_score(y_test, predictions)
KNN_JaccardIndex = jaccard_score(y_test, predictions, pos_label=0)
KNN_F1_Score = f1_score(y_test, predictions, average='weighted')
print('KNN Accuracy Score: ', KNN_Accuracy_Score)
print('KNN Jaccard Index: ', KNN_JaccardIndex)
print('KNN F1 Score: ', KNN_F1_Score)
```

```
KNN Accuracy Score:  0.8183206106870229
KNN Jaccard Index:  0.7901234567901234
KNN F1 Score:  0.802374933635524
```

```
# Add marker_cluster to current site_map
site_map.add_child(marker_cluster)

# for each row in spacex_df data frame
# create a Marker object with its coordinate
# and customize the Marker's icon property to indicate if this launch was successed or failed,
# e.g., icon=folium.Icon(color='white', icon_color=row['marker_color']
for index, record in spacex_df.iterrows():
    # TODO: Create and add a Marker cluster to the site map
    coordinate = [row['Lat'], row['Long']]
    # marker = folium.Marker(...)
    marker = folium.Marker(location=[record['Lat'], record['Long']],icon=folium.Icon(color=record['marker_color'],
        popup=record['class'])
    ).add_to(marker_cluster)
    marker_cluster.add_child(marker)
site_map
```

Thank you!