IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pavan R
16th Dec 2021

Outline

- Executive Summary Page 3
- Introduction Page 4
- Methodology Page 5
- Results Page 16
- Conclusion Page 45
- Appendix Page 46

Executive Summary

- **Methodology**:
  - Data Source is Wikipedia and https://api.spacexdata.com
  - Data wrangling was done to classify success/failure landing into numerical value
  - Exploratory data analysis (EDA) using visualization and SQL
  - Interactive visual analytics using Folium and Plotly Dash
  - predictive analysis using classification models like Logistic Regression, SVM, Decision Trees and KNN
- **Results**:
  - Success depends on payload mass, orbit.
  - We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%
  - Payload mass and Launch pad success rates are interdependent. Payload mass between 4000 to 10000 has highest success rates.
  - Decision tree has the highest accuracy. Success rate of successful landing of Falcon 9 is 88.93%

Introduction

- 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.
- Determine if the first stage will land successfully so that we can determine the cost of a launch. This helps us to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from Wikipedia about Falcon 9. Web scraping using Python BeautifulSoup package is used to collect and transform data into csv file.
  - Data collection from SpaceX website. https://api.spacexdata.com
- Perform data wrangling
  - Landing outcome could be success or failure. Landing region can be Ocean, ground or a drone ship. We can transformed landing outcome to ‘Class’ as ‘1’ if the booster successfully landed and ‘0’ if it was unsuccessful.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - EDA visualization we performed using seaborn and matplotlib pyplot libraries.
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

Data Collection

- Data was collected from Wikipedia about Falcon 9. Web scraping using Python BeautifulSoup package is used to collect and transform data into csv file.
- Extract data from table titled '**List of Falcon 9 first-stage boosters**'

Data Collection – SpaceX API

- Data collection using Python requests and Pandas packages from SpaceX "https://api.spacexdata.com/v4/launches/past"

- GitHub URL SpaceX API calls notebook

[CS_capstone/DS_SpaceYEval.ipynb at master · PavanKumarR/CS_capstone (github.com)]

response = requests.get(spacex_url)

Use json_normalize meethod to convert the json result into a dataframe

Filter data for 'Falcon 9'

Data Collection – Scraping

- Data collection using Python requests and BeautifulSoup packages

- GitHub URL of the completed SpaceX API calls notebook

[CS_capstone/DS_SpaceY_WebScrap.ipynb at master · PavanKumarR/CS_capstone (github.com)]

```
response = requests.get(static_url)

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

html_tables = soup.find_all('table')

first_launch_table = html_tables[2]
```

Data Wrangling

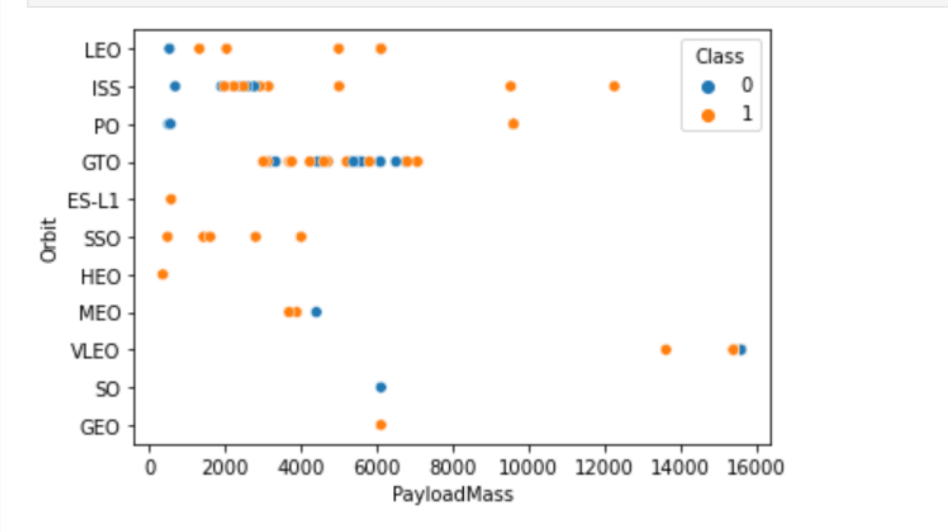
- Dataset processed from data collection using pandas are used
- Creating ‘Class’ variable to define outcome of successful and unsuccessful landing.
- CS_capstone/DS_SpaceY_Data_Wrangler.ipynb at master · PavanKumarR/CS_capstone (github.com)

EDA with Data Visualization

- Exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib

- Successful / Failure landing based

  on PayloadMass vs Orbit



- [CS_capstone/DS_SpaceY_EDA_Visual.ipynb at master · PavanKumarR/CS_capstone (github.com)]

EDA with SQL

- Total payload mass carried by boosters launched by NASA (CRS).
  - select sum(payload_mass__kg_) from CQD18693.SPACEXTBL where customer like 'NASA%'
- Names of the booster_versions which have carried the maximum payload mass.
  - %sql select booster_version from CQD18693.SPACEXTBL \
  - where payload_mass__kg_ = (select max(payload_mass__kg_) from CQD18693.SPACEXTBL)
- CS_capstone/DS_SpaceY_SQL_EDA.ipynb at master · PavanKumarR/CS_capstone (github.com)

Build an Interactive Map with Folium

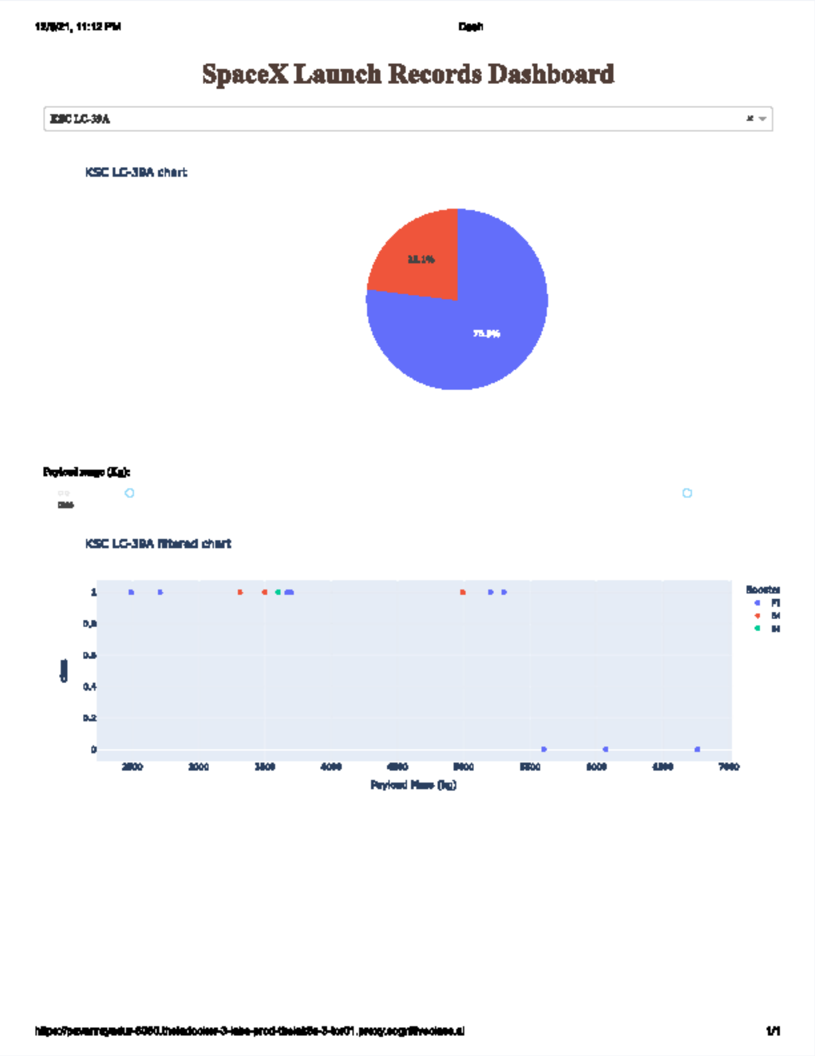
- Mark all Launchsites sites based on their coordinates
- Mark successful launches for each launchpads in ‘green’ and failure launches in ‘red’. This will give us visualization about the failures/success of each launches in geographic view.
- CS_capstone/DS_SpaceY_InteractiveVisual.ipynb at master · PavanKumarR/CS_capstone (github.com)

Build a Dashboard with Plotly Dash

- Pie chart to visualize launch site to success/failure

- Interactive scatter plot to see success/failure vs payload mass

- [CS_capstone/spacex_dash_app.py at master · PavanKumarR/CS_capstone (github.com)]



Predictive Analysis (Classification)

- We split the data into training and testing data using the function train_test_split. The training data is divided into validation data, a second set used for training data; then the models are trained and hyperparameters are selected using the function GridSearchCV
- Tuned for hyperparmeters using
  - Logistic Regression
  - Support Vector Machine
  - Decision tree classifier
  - K nearest neighbors classifier
- CS_capstone/DS_ML_Prediction.ipynb at master · PavanKumarR/CS_capstone (github.com)

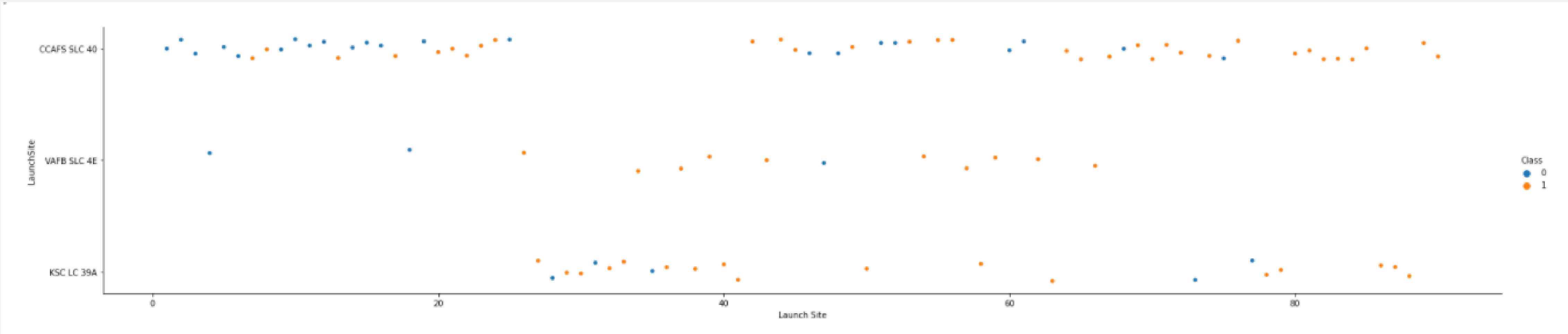
Results

- Success depends on payload mass, orbit.
- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%
- Payload mass and Launch pad success rates are interdependent. Payload mass between 4000 to 10000 has highest success rates.
- Decision tree has the highest accuracy. Success rate of successful landing of Falcon 9 is 88.93%

Section 2

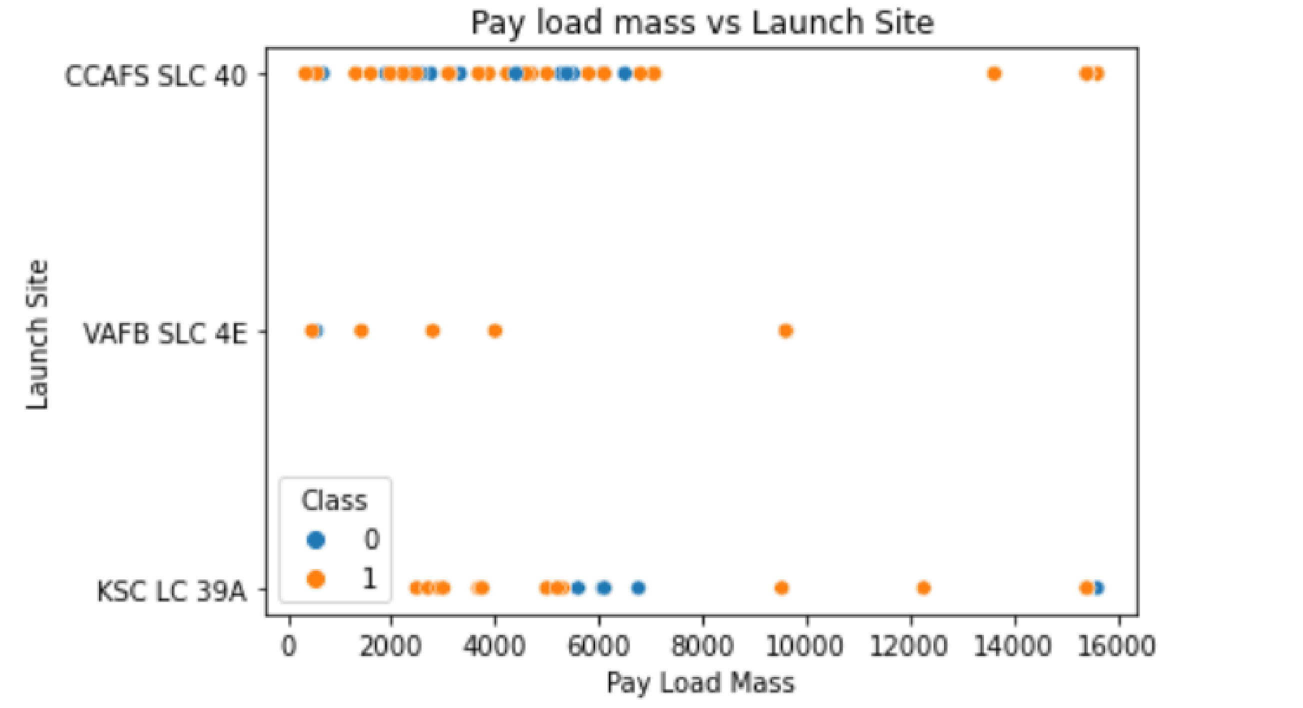
Insights drawn from EDA

Flight Number vs. Launch Site



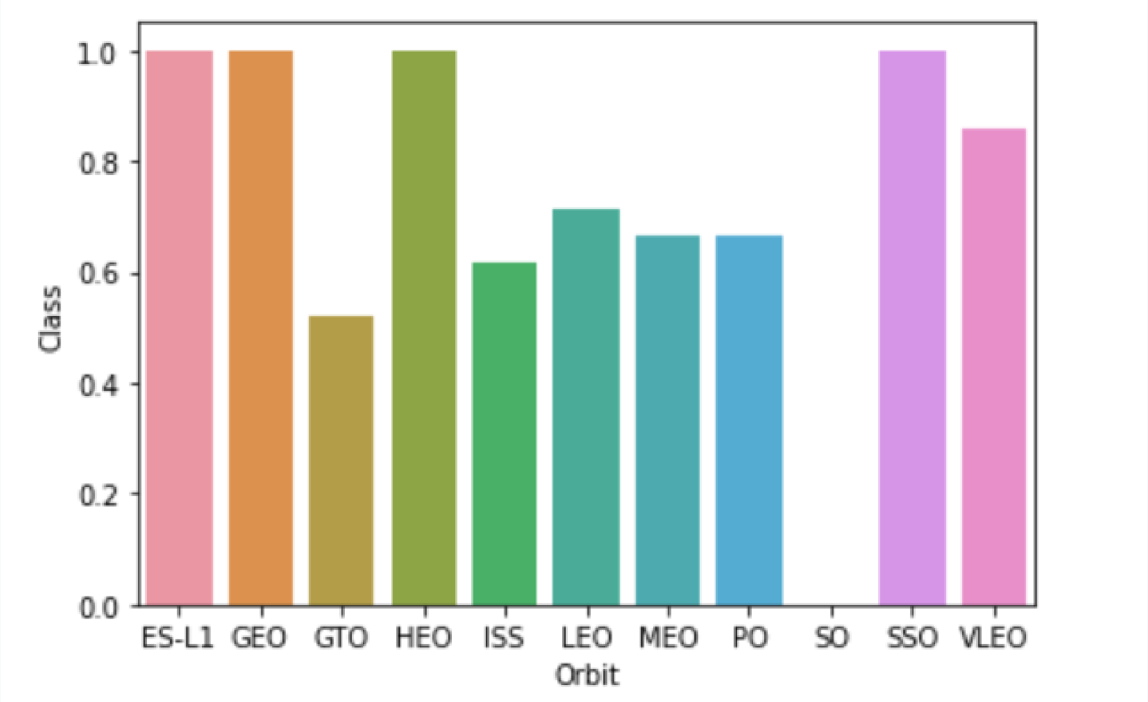
- All flight numbers range are launched from ‘CCAFS LC-40’
- Flight numbers do not have direct correlation with launchpad.
- Flight numbers do not decide success/failure

Payload vs. Launch Site



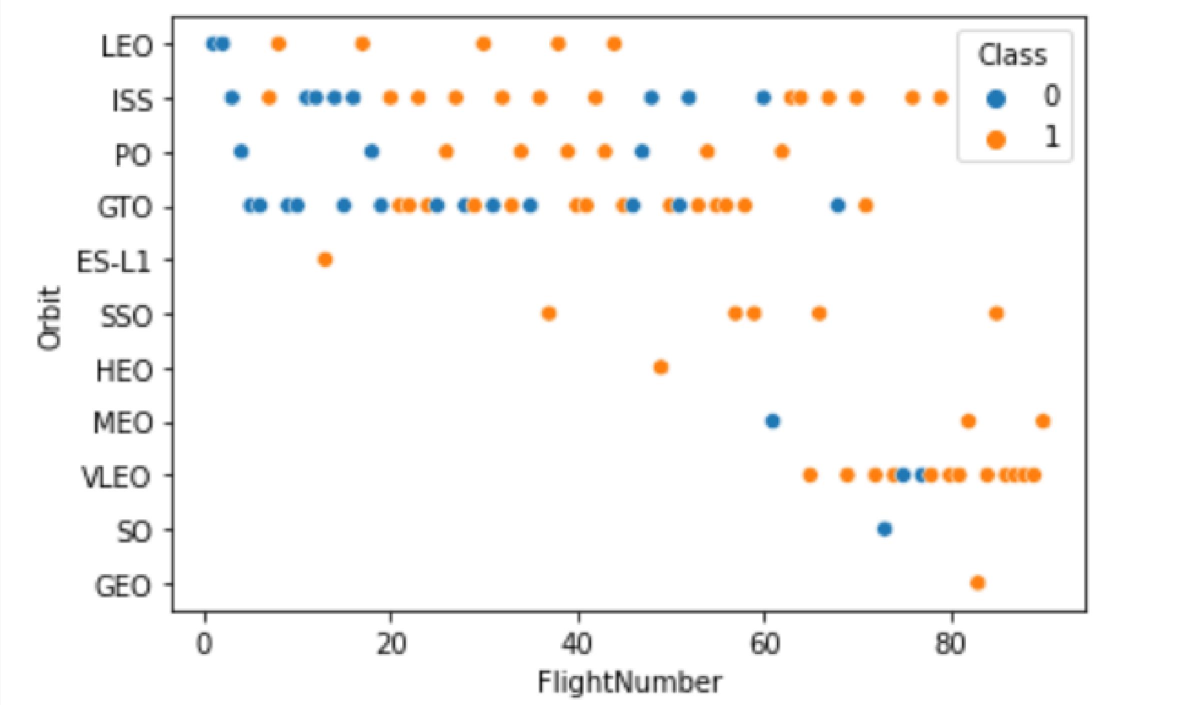
- High payload mass more than 7000 has more success rate

Success Rate vs. Orbit Type



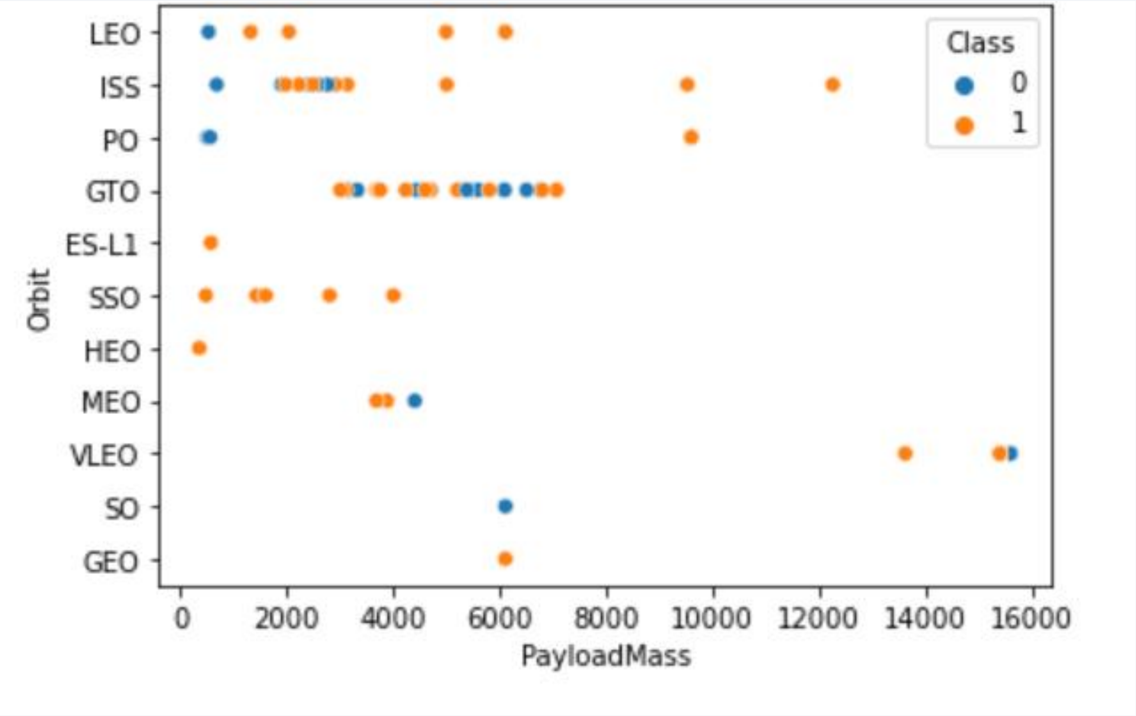
- ES-L1, GEO, HEO and SSO has 100% success rate
- GTO has lowest success rate of 50%
- SO orbit has 100% failure rate

Flight Number vs. Orbit Type



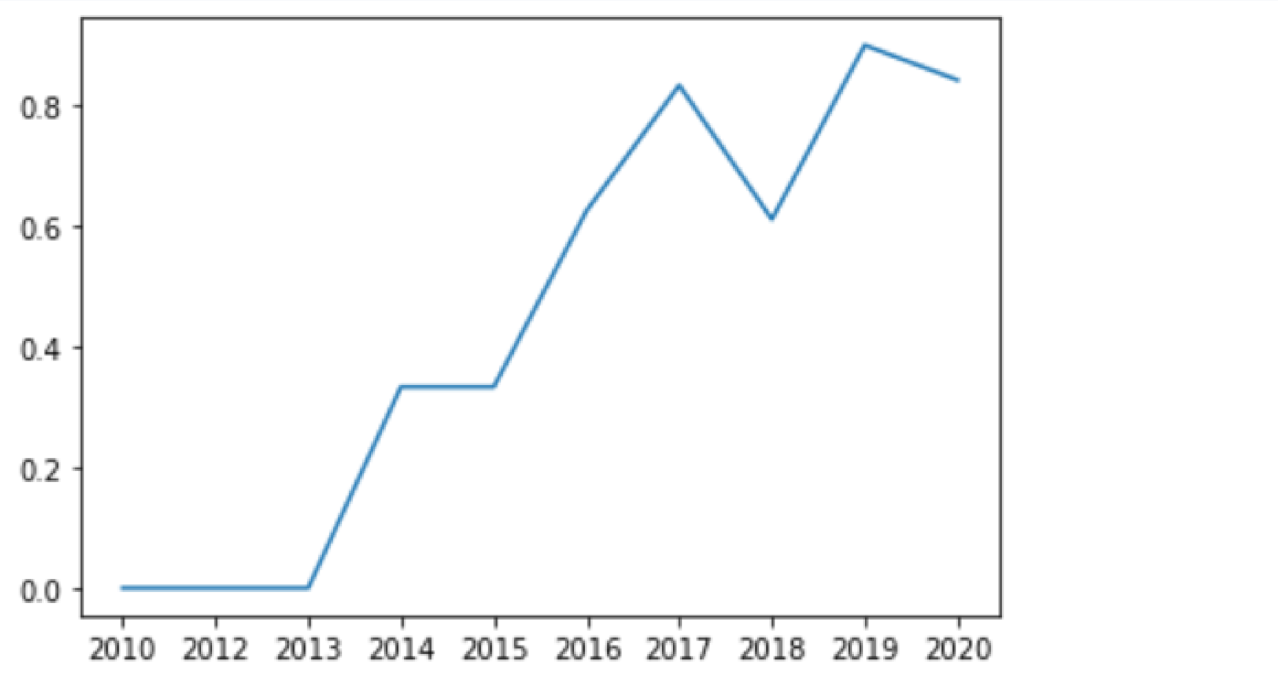
- Most flight number range are used for ISS orbit
- ES-L1, HEO, SO and GEO has used 1 particular flight
- SO orbit with particular flight number resulted in failure

Payload vs. Orbit Type



- Payload between 7000 and 15000 has resulted in success with all orbits
- SSO has 100% success rate orbit with payload used between 750 to 5000

Launch Success Yearly Trend



- Sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Display the names of the unique launch sites in the space mission

In [7]:

```
%sql select distinct launch_site from CQD18693.SPACEXTBL
```

```
 * ibm_db_sa://cqd18693:***@0c77d6f2-5da9-48a9-81f8-86b520b875
b
Done.
```

Out[7]:

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

- Distinct function on a launch_side column will list all unique launch site values

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

In [15]:
```
%sql select * from CQD18693.SPACEXTBL where launch_site like 'CCA%' fetch first 5 rows only
```

```
 * ibm_db_sa://cqd18693:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/blud
b
Done.
```

Out[15]:

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

- Like operator on a column will match substring with '%' to match all characters

Total Payload Mass

Task 3

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

In [16]:
```
%sql select sum(payload_mass__kg_) from CQD18693.SPACEXTBL where customer like 'NASA%'
```

```
 * ibm_db_sa://cqd18693:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.dat
b
Done.
```

Out[16]:

1
99980

- Sum function calculates the total value on payload mass column for the filtered set of rows with customer matching ‘NASA%’

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [18]:
```
%sql select avg(payload_mass__kg_) from CQD18693.SPACEXTBL where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://cqd18693:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
b
Done.
```

Out[18]:

1
2928

- Avg function provides average (i.e. sum of all values / number of values) on payload mass value for booster version matching F9 v1.1

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [21]:

```
%sql select min(date) from CQD18693.SPACEXTBL where landing__outcome like 'Success%'
```

```
 * ibm_db_sa://cqd18693:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.da
b
Done.
```

Out[21]:

1
2015-12-22

- Min function on date or numerical value will provide the least value from the rows with 'Success' landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [23]:
```
%sql select distinct booster_version from CQD18693.SPACEXTBL where payload_mass__kg_ > 4000 and landing__outcome = 'Succe
```

```
 * ibm_db_sa://cqd18693:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/blud
b
Done.
```

Out[23]:

booster_version
F9 B4 B1041.1
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026
F9 FT B1029.1
F9 FT B1036.1

- Unique names of booster version based on ‘Success’ landing

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [25]:

```
%sql select mission_outcome, count(*) from CQD18693.SPACEXTBL group by mission_outcome
```

```
 * ibm_db_sa://cqd18693:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.data
b
Done.
```

Out[25]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [29]:

```
%sql select booster_version from CQD18693.SPACEXTBL \
where payload_mass__kg_ = (select max(payload_mass__kg_) from CQD18693.SPACEXTBL)
```

```
 * ibm_db_sa://cqd18693:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databas
b
Done.
```

Out[29]:

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

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

In [33]:
```
%sql select landing__outcome, booster_version, launch_site from CQD18693.SPACEXTBL \
where year(date) = 2015 and lower(landing__outcome) like 'failure (drone ship)%'
```

```
 * ibm_db_sa://cqd18693:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.a
b
Done.
```

Out[33]:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Filtering data based on multiple columns year and 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

In [39]:

```
%sql select landing__outcome, count(*) from CQD18693.SPACEXTBL \
where date >= '2010-06-04' and date <= '2017-03-20' \
group by landing__outcome \
order by count(*) desc
```

```
 * ibm_db_sa://cqd18693:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bl
b
Done.
```

Out[39]:

landing__outcome	2
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

- Group the filtered data and order by that column will provide the list

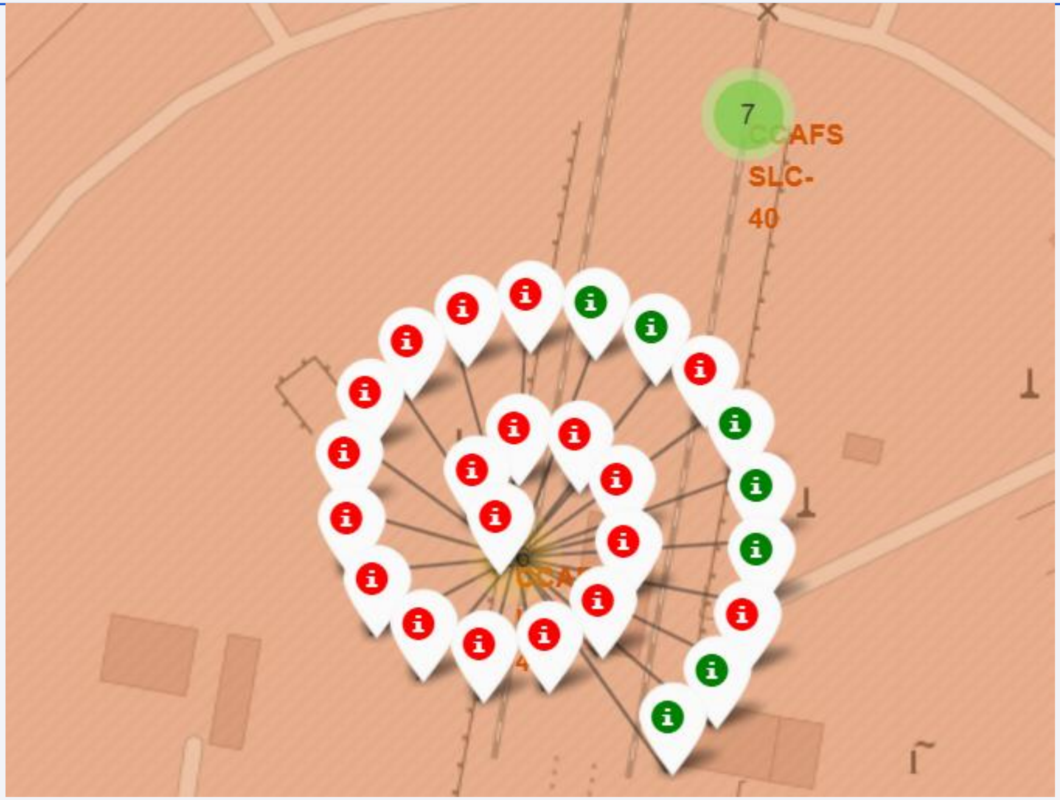
Section 4

Launch Sites Proximities Analysis

All launch sites

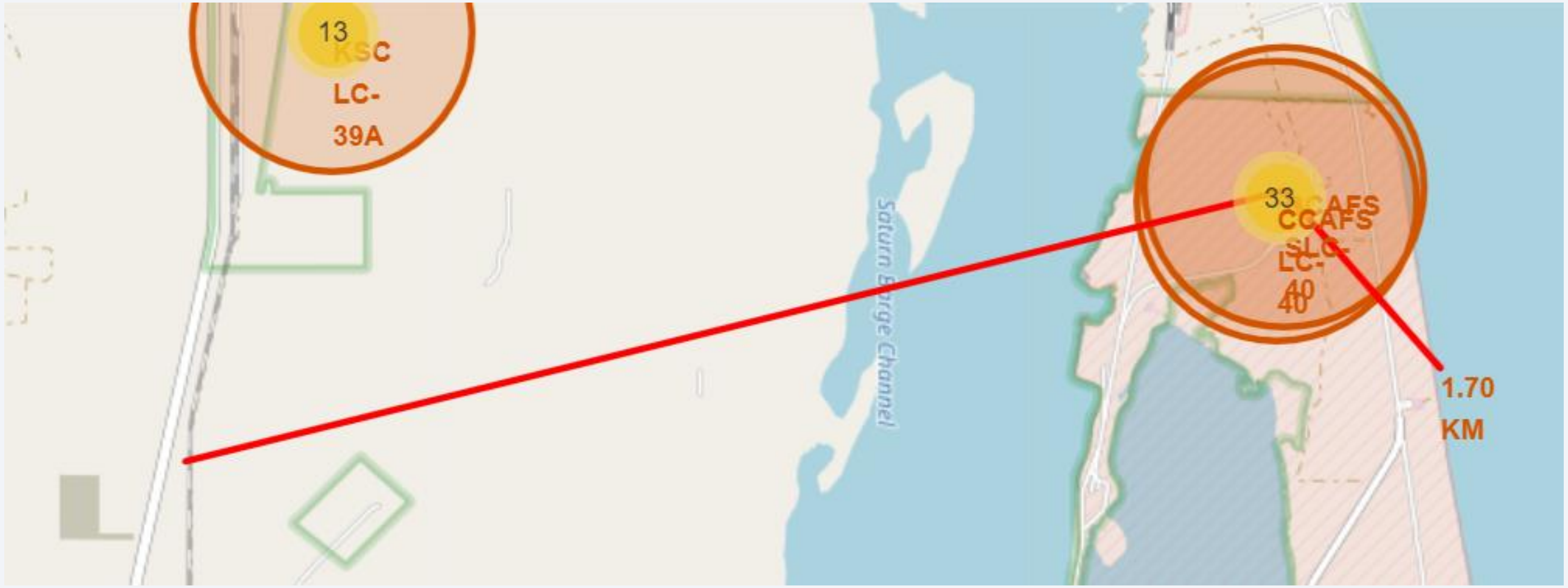
- Launch sites are usually to the coastline

Success/failed launches for each site



- Since launch pad location are shared, they have overlapping markers. Green represents success and Red markers represents failure

Distance to the Coastal line and Railway line

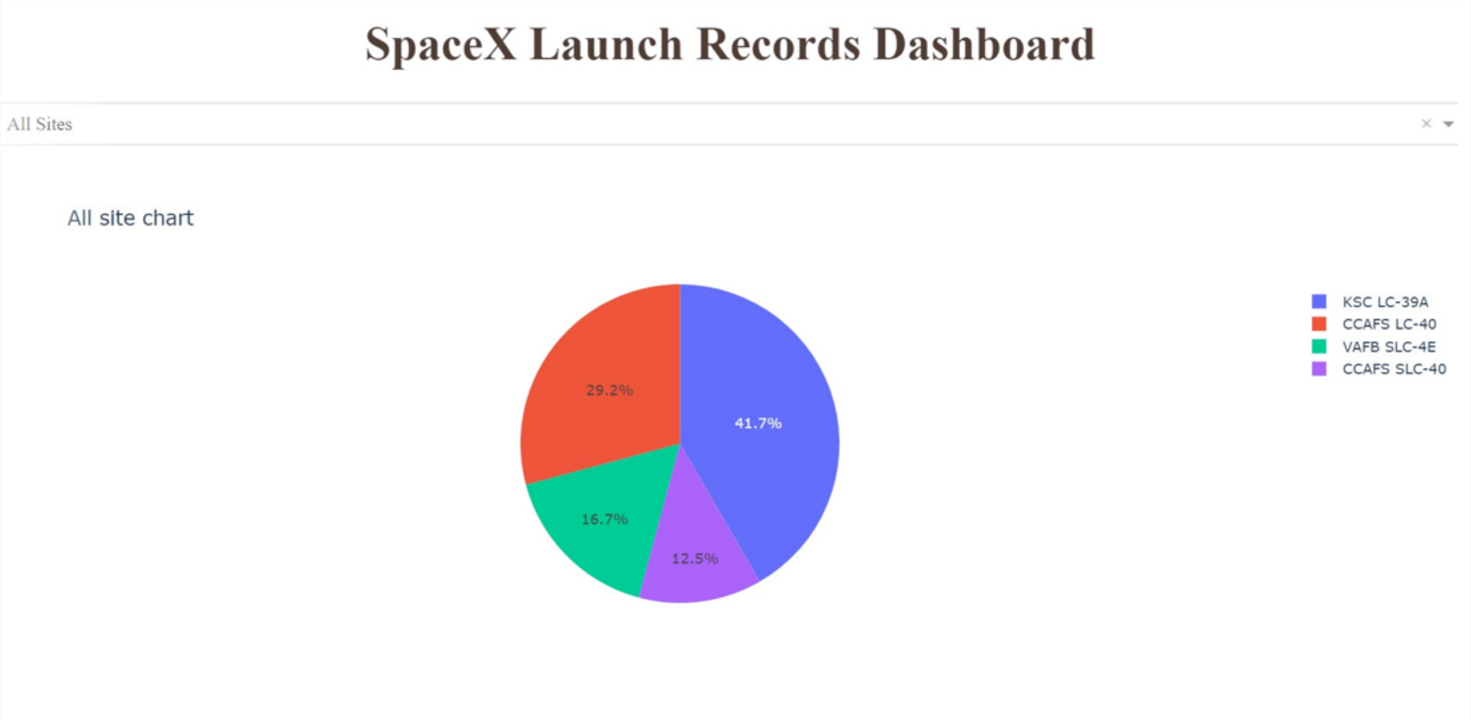


- To differentiate better, coastal point is taken a little farther.

Section 5

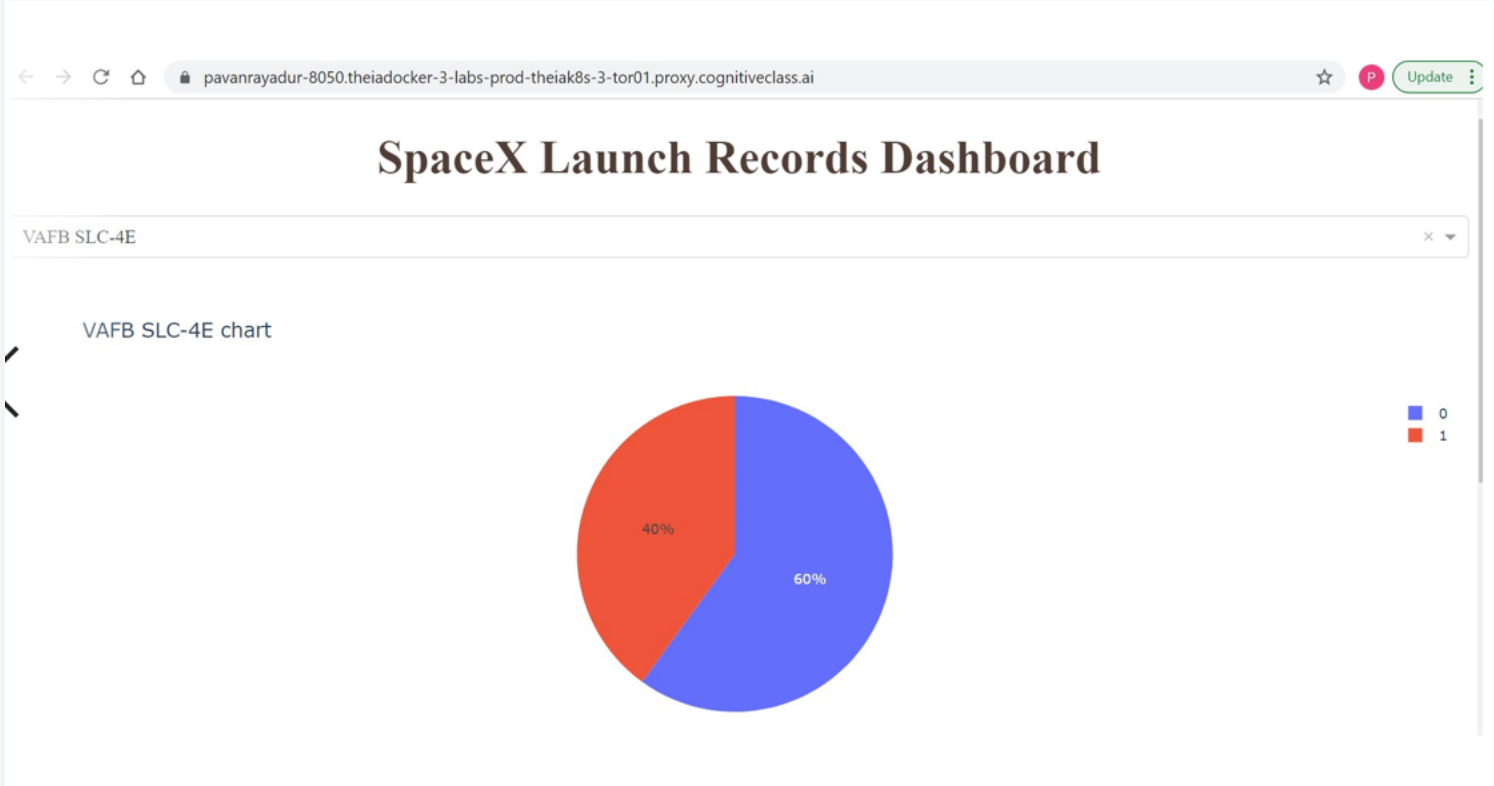
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard for all sites



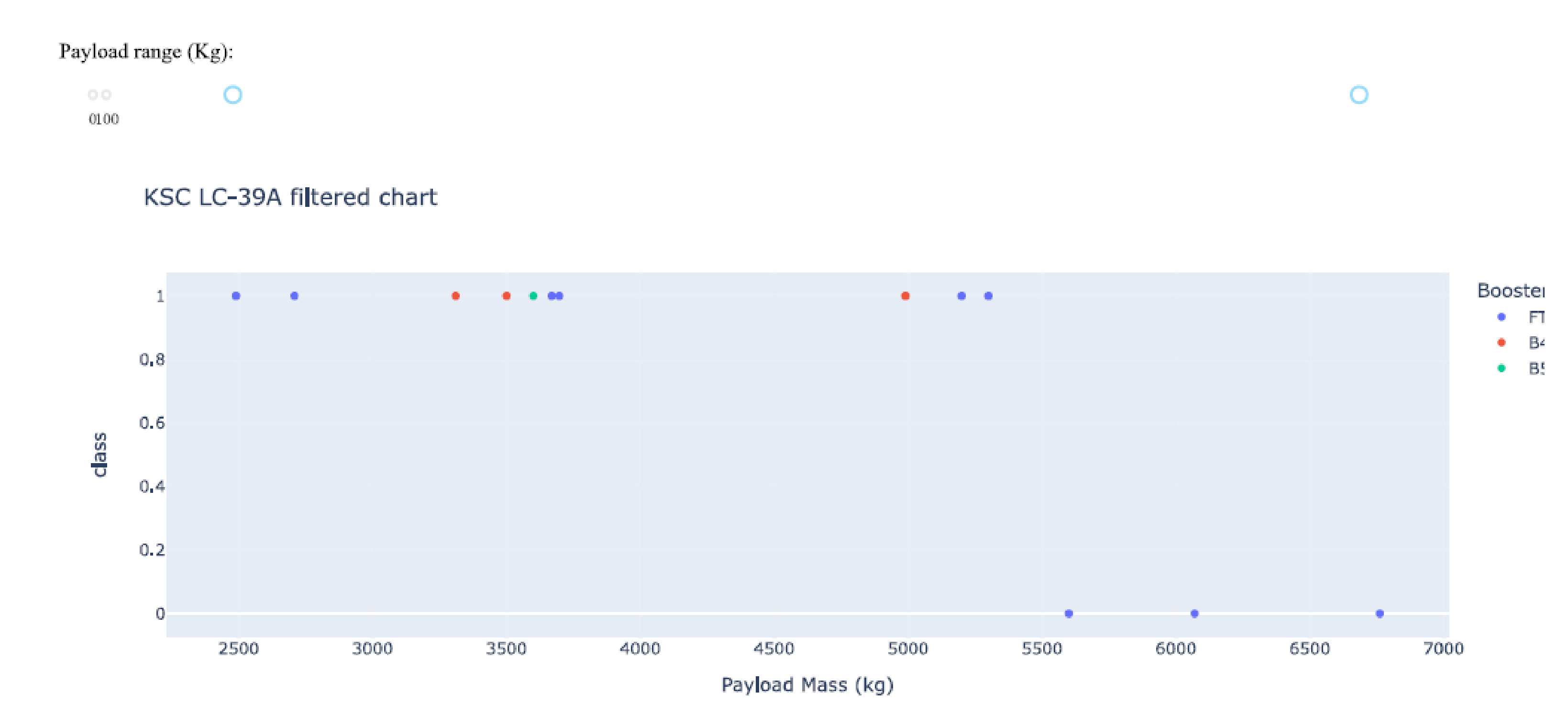
- Chart denotes the launch sites used from overall available data

SpaceX Launch Records Dashboard – VAFB SLC-4E site



- VAFB SLC-4E site has 60% success rate

KSC LC-39A site Payload Success index



- Payload maxx between 5000 and 7000 has failures

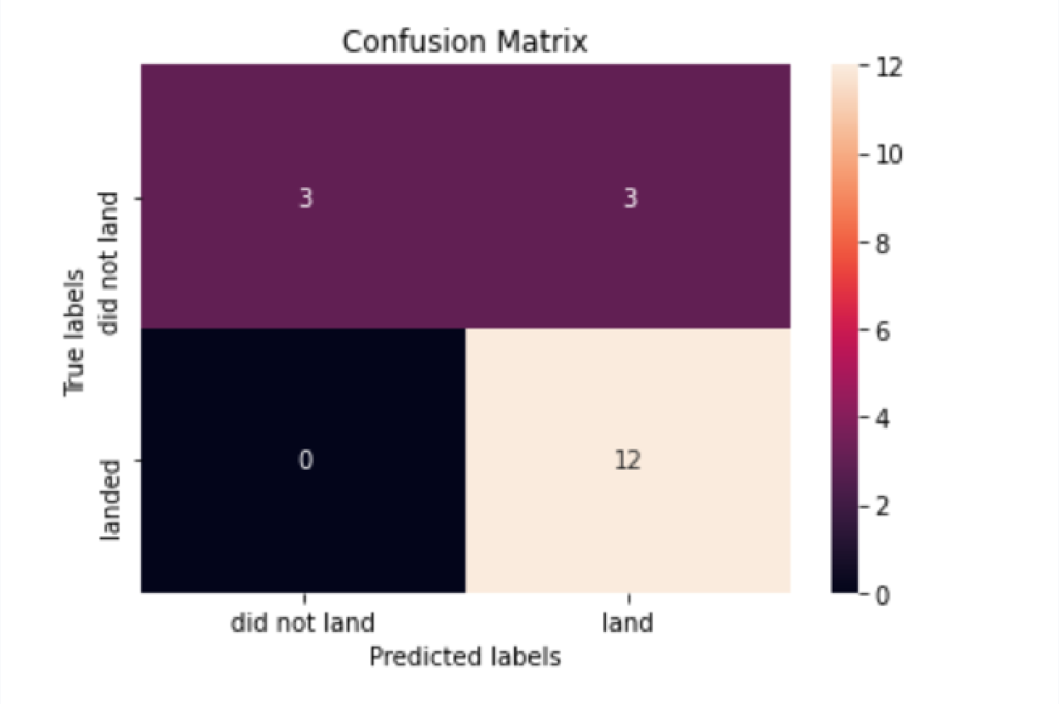
Section 6

Predictive Analysis (Classification)

Classification Accuracy

- All models have an accuracy of 85% and above.
- Decision Trees have the highest accuracy with 88.93% success rates

Confusion Matrix of logistics regression



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

Conclusions

- Success depends on payload mass, orbit.
- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%
- Payload mass and Launch pad success rates are interdependent. Payload mass between 4000 to 10000 has highest success rates.
- Decision tree has the highest accuracy. Success rate of successful landing of Falcon 9 is 88.93%

Appendix

- None

Thank you!