IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Joon Hee Jang
12.17.2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

Project Aim: Predict successful land of SpaceX Falcon 9 first stage

SpaceX advertises Falcon 9 rocket launches on its website with fraction of the cost compared to other providers, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine the outcome of the first stage landing, we can determine the cost of a launch. In this project, data of previous Falcon 9 launch and its outcome is analyzed and use the data to train a machine learning model to predict if the Falcon 9 first stage will land successfully.

There are four different model used to predict the outcome and compared accuracy from cross validation and from test dataset. Decision tree model shows the highest accuracy in cross validation while all model exhibits similar test accuracy score of 0.833.

Introduction

- Project background and context

  SpaceX is an American spacecraft manufacturer, launcher, and a satellite communications corporation. It was founded in 2002 by Elon Musk with the stated goal of reducing space transportation costs. The major cost saving is come from enabling reuse of the first stage rocket. 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.

- Problems you want to find answers

  As the rocket is complex set of engineered machinery, there are always risk of unsuccessful landing outcome. 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. Data of previous Falcon 9 launch and its outcome is analyzed and use the data to train a machine learning model to predict if the Falcon 9 first stage will land successfully.

Successful landing

Unsuccessful landing

Section 1

Methodology

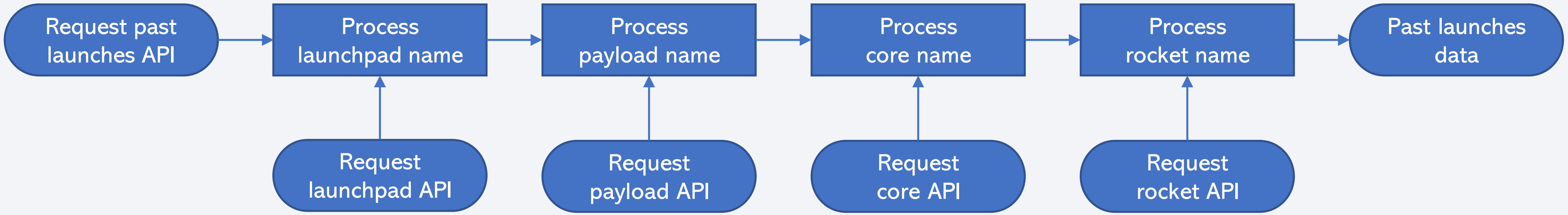
Methodology

- Data collection methodology:
  - Historical data of SpaceX rocket launch and its outcome are collected from SpaceX public database using API.
  - Separately, falcon 9 and falcon heavy launches records are collected from Wikipedia web scraping.
- Perform data wrangling
  - The missing values from collected data has been replaced with mean values.
  - Unnecessary annotation from the data in the table had been removed and formatted correctly.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create dependent and independent variables with train and test data sets.
  - Build machine learning pipeline to predict if the first stage will land from given the data.
  - Find best hyperparameters for Logistic Regression, SVM, Classification Tree, and KNN using GridSearchCV.

Data Collection

- Historical data of SpaceX rocket launch and its outcome from SpaceX API

  API URL Path: https://api.spacexdata.com/v4/



- Falcon 9 and falcon heavy launches records are collected from Wikipedia

  API URL Path: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

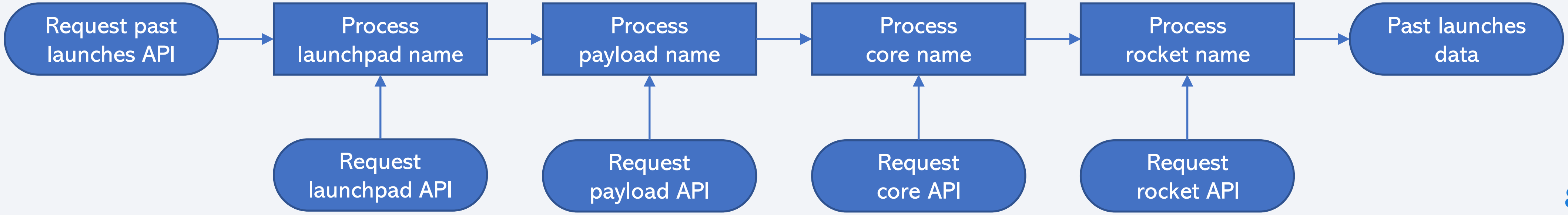
Data Collection – SpaceX API

- Data collection with SpaceX REST

```
spacex_url=https://api.spacexdata.com/v4/launches/past
response = requests.get(spacex_url)
data=pd.json_normalize(response.json())
```

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs
1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False
2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False
3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False
4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False
5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False
...	...	...	...	...	...	...	...	...	...	...
86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True
87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True

- GitHub URL of the completed SpaceX API calls notebook: https://github.com/jhjang101/testrepo/blob/eacd53d1d4e8e00ea55bcf862dd41480496cdb8e/Data%20Collection%20API.ipynb



Data Collection - Scraping

- Web scraping process

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
soup = BeautifulSoup(r, 'html.parser')
```

Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing
1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure
2	CCAFS	Dragon	0	LEO	NASA (COTS)\nNRO	Success	F9 v1.0B0004.1	Failure
3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)	Success	F9 v1.0B0005.1	No attempt\n
4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0006.1	No attempt
5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0007.1	No attempt\n

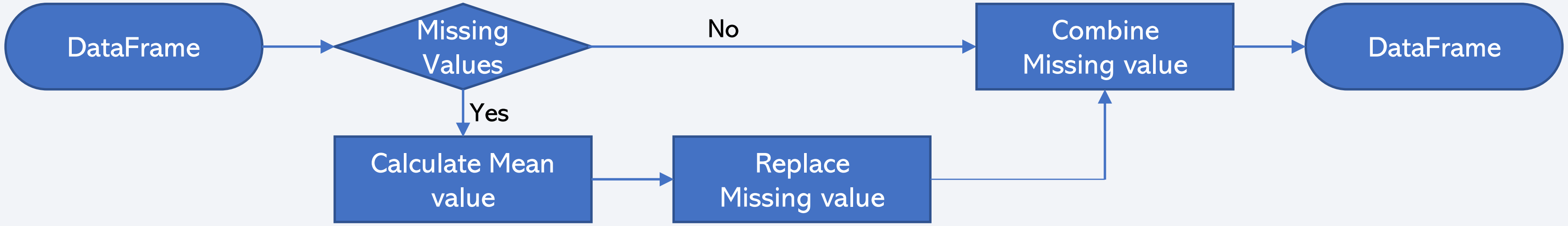
- GitHub URL of the completed web scraping notebook: https://github.com/jhjang101/testrepo/blob/eacd53d1d4e8e00ea55bcf862dd41480496cdb8e/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- The missing values from collected data has been replaced with mean values.

```python
# Calculate the mean value of PayloadMass column
mean_PayloadMass = data_falcon9['PayloadMass'].mean()

# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan, mean_PayloadMass, inplace=True)
```



- Unnecessary annotation from the data in the table had been removed and formatted correctly.

```python
#if it is number save cells in a dictonary
if flag:
    extracted_row += 1
    # Flight Number value
    # TODO: Append the flight_number into launch_dict with key `Flight No.`
    launch_dict['Flight No.'].append(flight_number)
    #print(flight_number)
    datatimelist=date_time(row[0])
```

EDA with Data Visualization

- Flight number vs. Payload - It seems the more massive the payload, the less likely the first stage will return.
- Flight number vs. Launch Site – Higher flight number mostly use CCAFS SLC 40.
- Payload vs. Launch Site – VAFB SLC 4E did not launch high payload (>10,000 kg).
- Orbit vs. Success rate – ES-L1, SSO, HEO, and GEO has 100% success rate.
- Flight number vs. Orbit – MEO, VLEO, SO, and GEO has high flight numbers.
- Payload vs. Orbit – Each orbits has different payload distribution.
- Launch success yearly trend – Sucess rate since 2013 kept increasing till 2020.
- GitHub URL: https://github.com/jhjang101/testrepo/blob/eacd53d1d4e8e00ea55bcf862dd41480496cdb8e/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- SQL queries
  - SELECT DISTINCT(launch_site) FROM SpaceX
  - SELECT * FROM SpaceX WHERE launch_site LIKE 'CCA%' LIMIT 5
  - SELECT SUM(payload_mass__kg_) FROM SpaceX WHERE customer = 'NASA (CRS)'
  - SELECT AVG(payload_mass__kg_) FROM SpaceX WHERE booster_version = 'F9 v1.1'
  - SELECT MIN(DATE) FROM SpaceX WHERE landing__outcome = 'Success (ground pad)'
  - SELECT DISTINCT(booster_version) FROM SpaceX WHERE mission_outcome = 'Success' AND payload_mass__kg_ BETWEEN 4000 AND 6000
  - SELECT COUNT(*) FROM SpaceX WHERE mission_outcome LIKE 'Success%'
  - SELECT DISTINCT(booster_version) FROM SpaceX WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SpaceX)
  - SELECT * FROM SpaceX Limit 50
  - SELECT COUNT(*), landing__outcome FROM SpaceX WHERE (landing__outcome = 'Failure (drone ship)' OR landing__outcome = 'Success (ground pad)') AND
  - BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing__outcome ORDER BY COUNT(landing__outcome) DESC

- GitHub URL
  - https://github.com/jhjang101/testrepo/blob/eacd53d1d4e8e00ea55bcf862dd41480496cdb8e/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Map objects
  - All launch sites – Circle and Markers to show physical locations of launch sites.
  - Success/failed launches for each site – Markers using marker_cluster to show Success/failed launches depending on launch sites and check correlations.
  - Distances between a launch site to its proximities – Markers and Polylines to show Distances between a launch site to railways, highways, and cities and check correlations.

- GitHub URL
  - https://github.com/jhjang101/testrepo/blob/eacd53d1d4e8e00ea55bcf862dd41480496cdb8e/lab_jupyter_launch_site_location.ipynb – Folium not displaying correctly.
  - Alterative PDF access: https://github.com/jhjang101/testrepo/blob/56f56ac41b0c95fd2915b9e207d6a4d7a387b1c3/lab_jupyter_launch_site_location.pdf

Build a Dashboard with Plotly Dash

- Plots
  - Success Count for all launch sites – To show success counts depending on launch sites and visualize correlations.
  - Success Count for Payload Mass (kg) with Booster Version – To show success counts depending on payload and booster version.

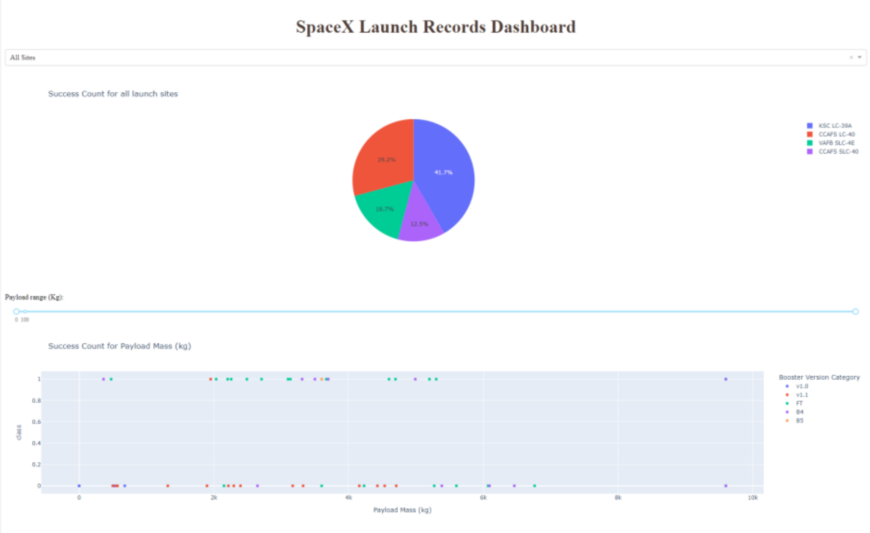
- GitHub URL
  - https://github.com/jhjang101/testrepo/blob/56f56ac41b0c95fd2915b9e207d6a4d7a387b1c3/spacex_dash_app.py
  - Dash PDF: https://github.com/jhjang101/testrepo/blob/f6b7548cd69adee212591bd4bb55a6f4e2915ed2/spacex_dash.pdf

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Results

- Exploratory data analysis shows there are some degrees of correlation between features such as payloads, launch sites, and booster with success rate.
- There are four different model used to predict the outcome and compared accuracy from cross validation and from test dataset. Decision tree model shows the highest accuracy in cross validation while all model exhibits similar test accuracy score of 0.833.



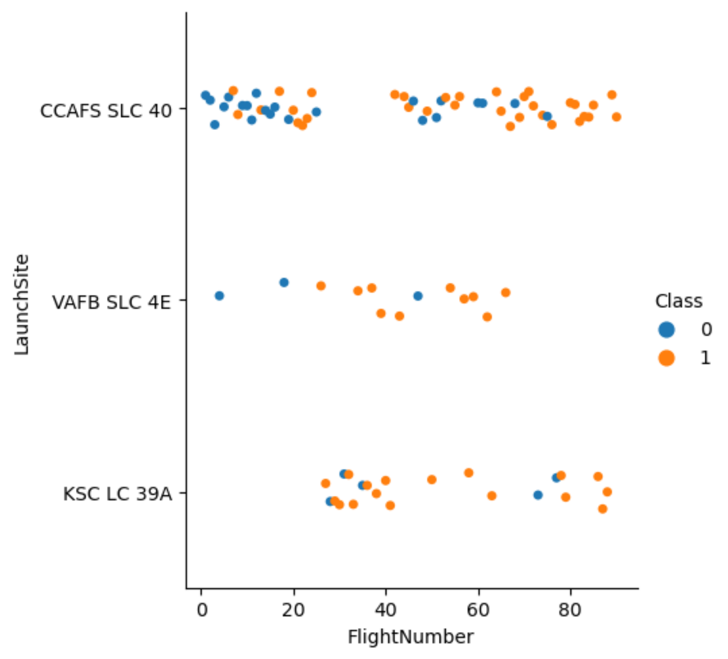
Model	CV Accuracy	Test Accuracy
LogisticRegression	0.846429	0.833333
SVM	0.848214	0.833333
Decision Trees	0.889286	0.833333
KNN	0.848214	0.833333

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

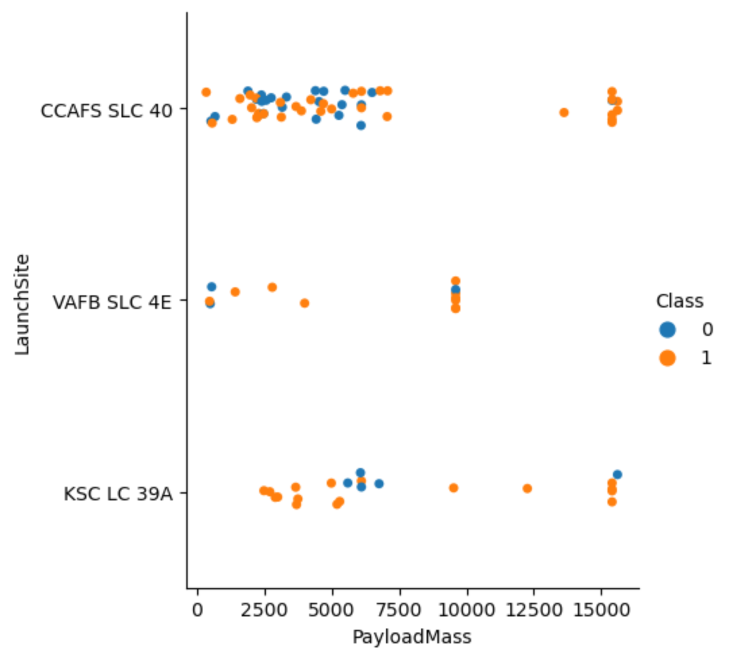
Flight Number vs. Launch Site



- CCAFS SLC 40 launch site has the most number of launches.
- As the flight number increases, success rate become higher.
- VAFB SLC 4E launch site has lowest number of launches.
- VAFB SLC 4E has not been used since flight number 65.
- In low flight numbers, KSC LC 39A launch site was not used.

Payload vs. Launch Site

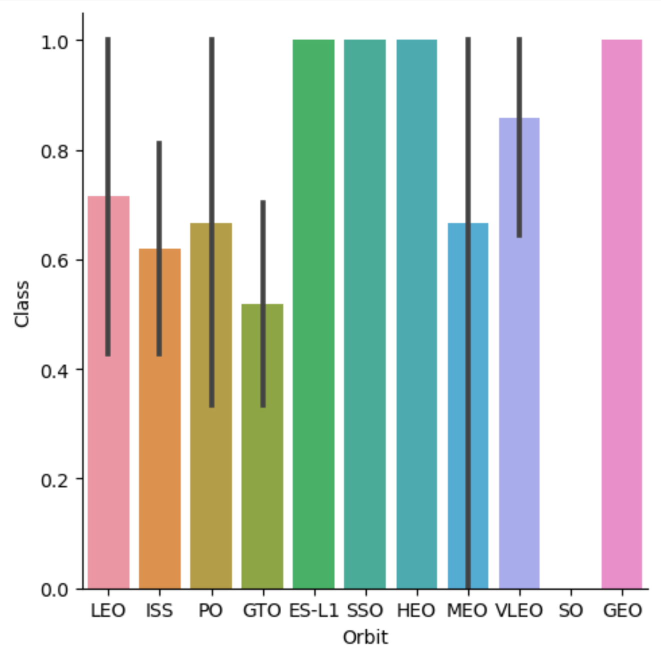
Payload vs. Launch Site



- CCAFS SLC 40 launch site mostly used for low payload or very high payload.
- CCAFS SLC 40 launch site has good success rate for very high payload.
- VAFB SLC 4E launch site was not used for payload more than 10,000.
- KSC LC 39A launch site was not used for payload less than 2,500.
- Higher payload tend to have better success rate but it may be due to CCAFS SLC 40 has relatively low success rate in low payload and CCAFS SLC 40 was mostly used in the low flight number.

Success Rate vs. Orbit Type

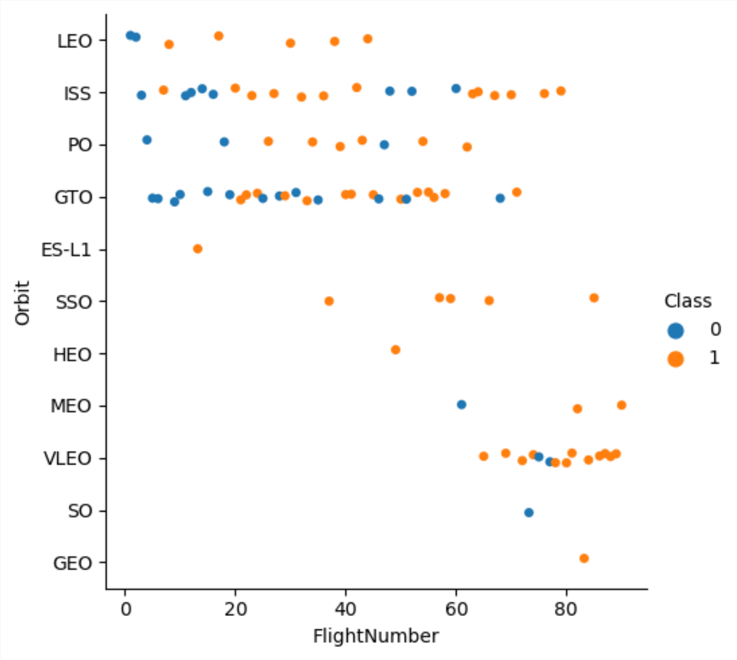
Success rate of each orbit type



- ES-L1, SSO, HEO, and GEO has 100% success rate.
- SO has zero success rate.

Flight Number vs. Orbit Type

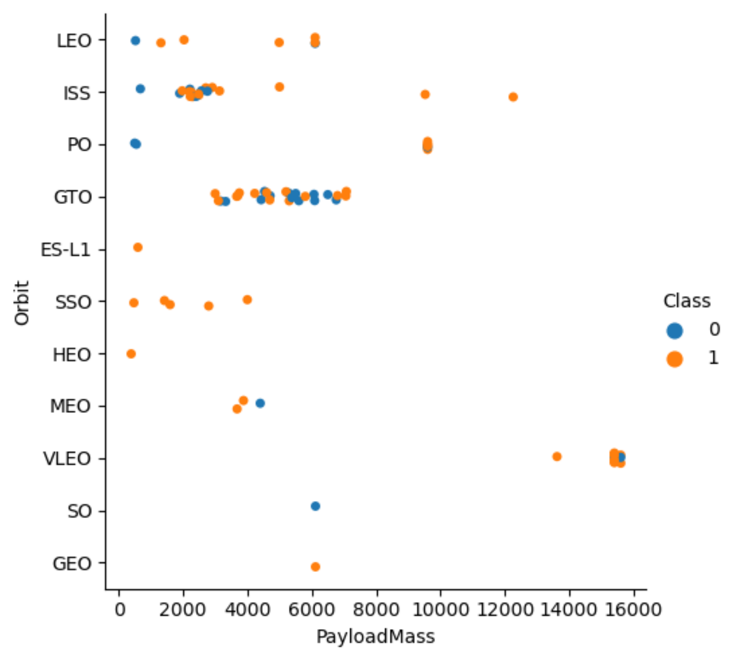
Flight number vs. Orbit type



- MEO, VLEO, SO, and GEO has high flight numbers.
- Low flight number did not lauch to orbit HEO, MEO, VLEO, SO, and GEO.

Payload vs. Orbit Type

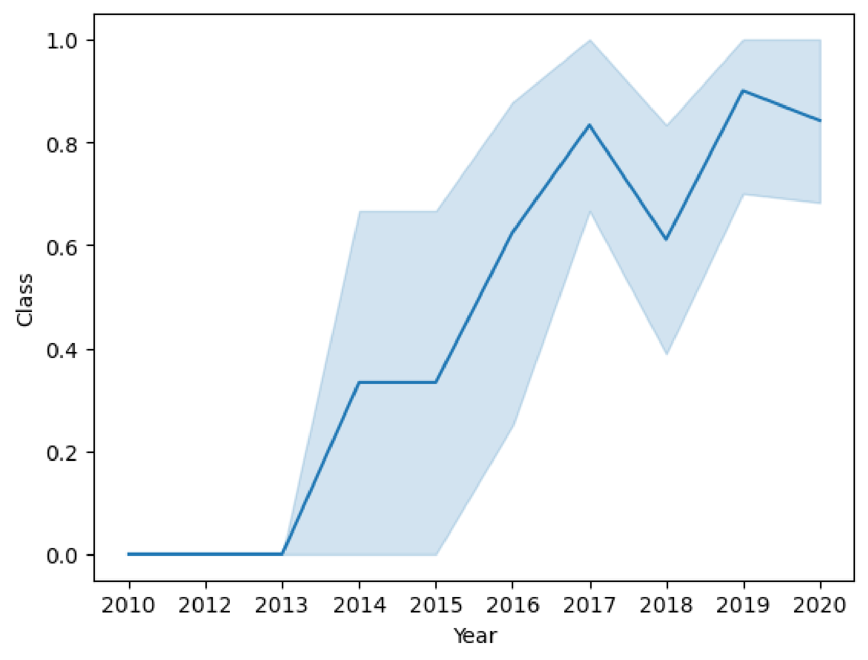
Payload vs. Orbit type



- GTO has payload between 2,500 and 8,000
- VLEO has very high payloads

Launch Success Yearly Trend

Yearly average success rate



- Success rate since 2013 kept increasing till 2020.
- There are no success before 2013.
- Success rate reached plateau at around 0.8 since 2017.

All Launch Site Names

Names of the unique launch sites

```
%%sql
SELECT DISTINCT(launch_site) FROM SpaceX
```

```
 * ibm_db_sa://xqk12606:***@54a2f15b-5c0f-46df-8954-7
Done.
```

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

- There are four unique launch sites

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

Task 2

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

```
%%sql
SELECT * FROM SpaceX
WHERE launch_site LIKE 'CCA%'
LIMIT 5
```

```
 * ibm_db_sa://xqk12606:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/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	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

- No success landing outcome

Total Payload Mass

Total payload carried by boosters from NASA

Task 3

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

```
%%sql
SELECT SUM(payload_mass__kg_) FROM SpaceX
WHERE customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://xqk12606:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj
Done.
```

1
45596

- Total payload from NASA : 45,596 kg

Average Payload Mass by F9 v1.1

average payload mass carried by booster version F9 v1.1

Task 4

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

```
%%sql
SELECT AVG(payload_mass__kg_) FROM SpaceX
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://xqk12606:***@54a2f15b-5c0f-46df-8954-7e38e612c2
Done.
```

1
2928

- average payload mass carried by booster version F9 v1.1: 2,928 kg

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad

Task 5

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

Hint:Use min function

```
%%sql
SELECT MIN(DATE) FROM SpaceX
WHERE landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://xqk12606:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0l
Done.
```

1
2015-12-22

- The dates of the first successful landing outcome on ground pad: 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

Task 6

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

```
%%sql
SELECT DISTINCT(booster_version) FROM SpaceX
WHERE mission_outcome = 'Success' AND payload_mass__kg_ BETWEEN 4000 AND 6000
```

booster_version
F9 B4 B1040.2
F9 B4 B1040.1
F9 B5 B1046.2
F9 B5 B1046.3
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5 B1058.2
F9 B5B1054
F9 B5B1060.1
F9 B5B1062.1
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1032.2
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1032.1
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016

- names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%%sql
SELECT COUNT(*) FROM SpaceX
WHERE mission_outcome LIKE 'Success%'
```

```
 * ibm_db_sa://xqk12606:***@54a2f15b-5c0f-46df-8954-7e38e612
Done.
```

1
100

```
%%sql
SELECT COUNT(*) FROM SpaceX
WHERE mission_outcome LIKE 'Failure%'
```

```
 * ibm_db_sa://xqk12606:***@54a2f15b-5c0f-46df-8954-7e38e612
Done.
```

1
1

- Successful mission outcomes: 100
- Failure mission outcomes: 1

Boosters Carried Maximum Payload

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

Task 8

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

```
%%sql
SELECT DISTINCT(booster_version) FROM SpaceX
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SpaceX)
```

```
 * ibm_db_sa://xqk12606:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databas
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

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

2015 Launch Records

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

```
%%sql
SELECT * FROM SpaceX
WHERE DATE LIKE '2015%'
AND landing__outcome = 'Failure (drone ship)'
```

```
 * ibm_db_sa://xqk12606:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2015-01-10	09:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

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

Task 10

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 desce

```
%%sql
SELECT COUNT(*), landing__outcome  FROM SpaceX
WHERE (landing__outcome = 'Failure (drone ship)' OR landing__outcome = 'Success (ground pad)')
AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY COUNT(landing__outcome) DESC
```

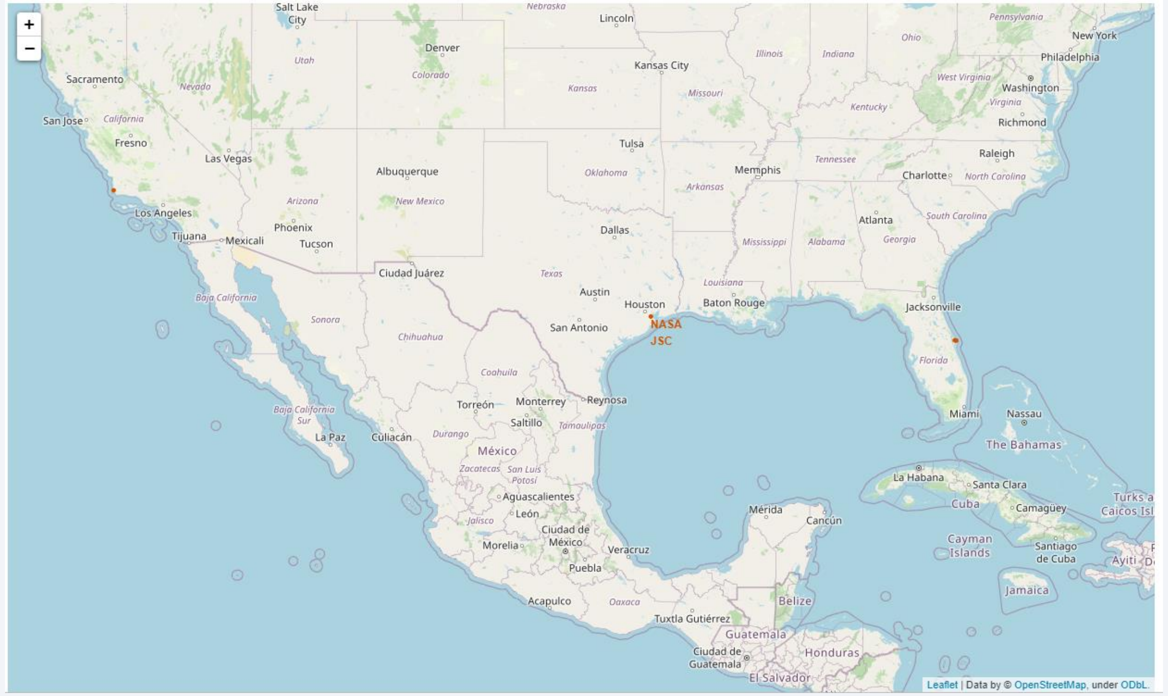
```
 * ibm_db_sa://xqk12606:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

1	landing__outcome
5	Failure (drone ship)
3	Success (ground pad)

Section 3

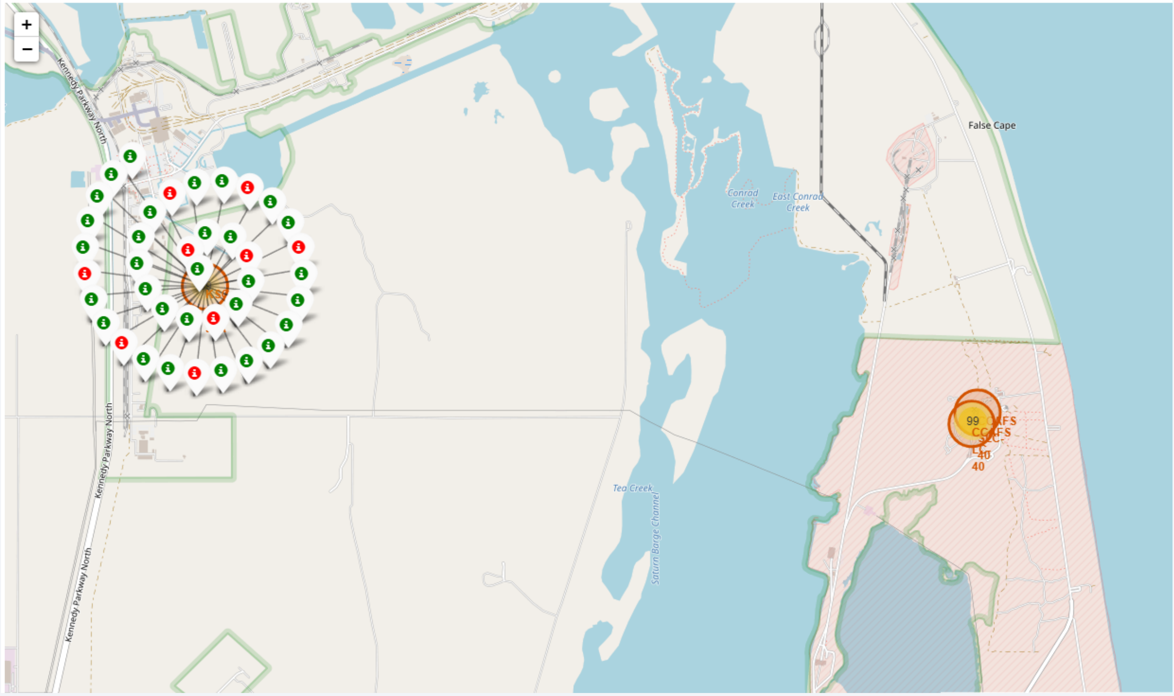
Launch Sites Proximities Analysis

Location of launch site and NASA Johnson Space Center at Houston, Texas



Launch sites are located at California and Florida.

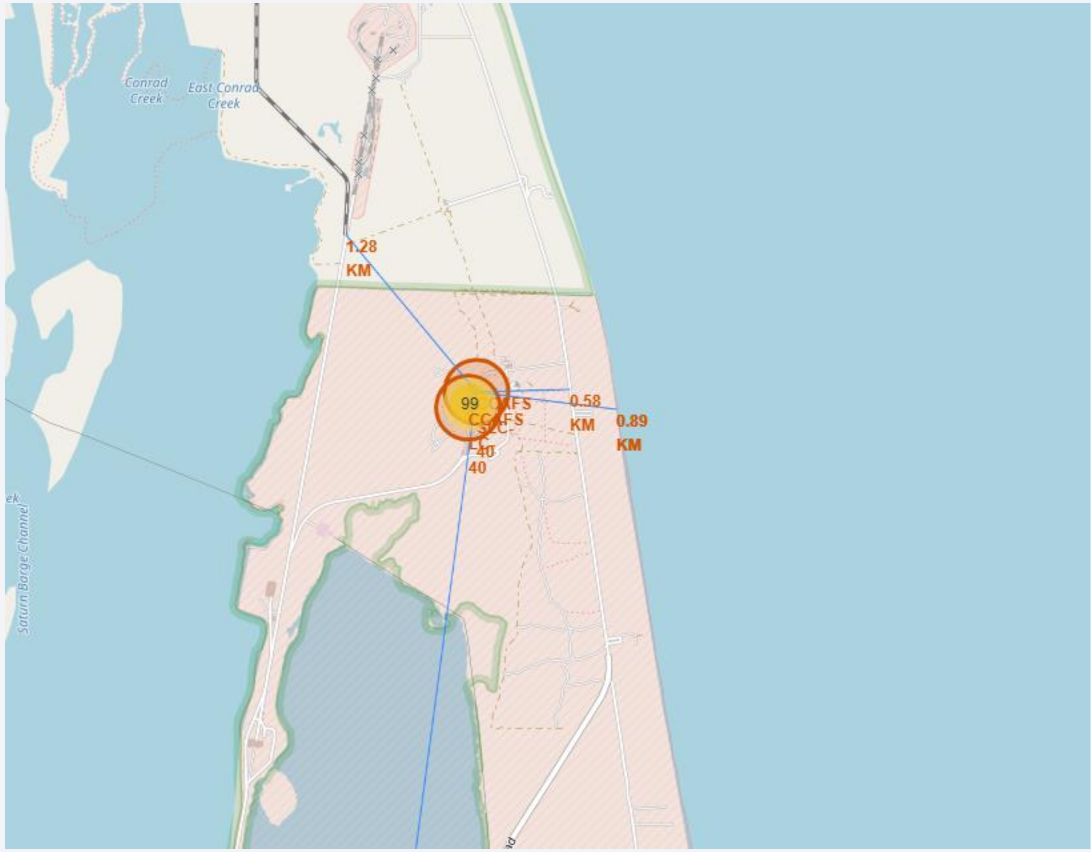
Success/failure for each launch site



Green are success and Red are failure.

Example screen shot shows success/failure at KSC LC 39A

Distances between a launch site to its proximities

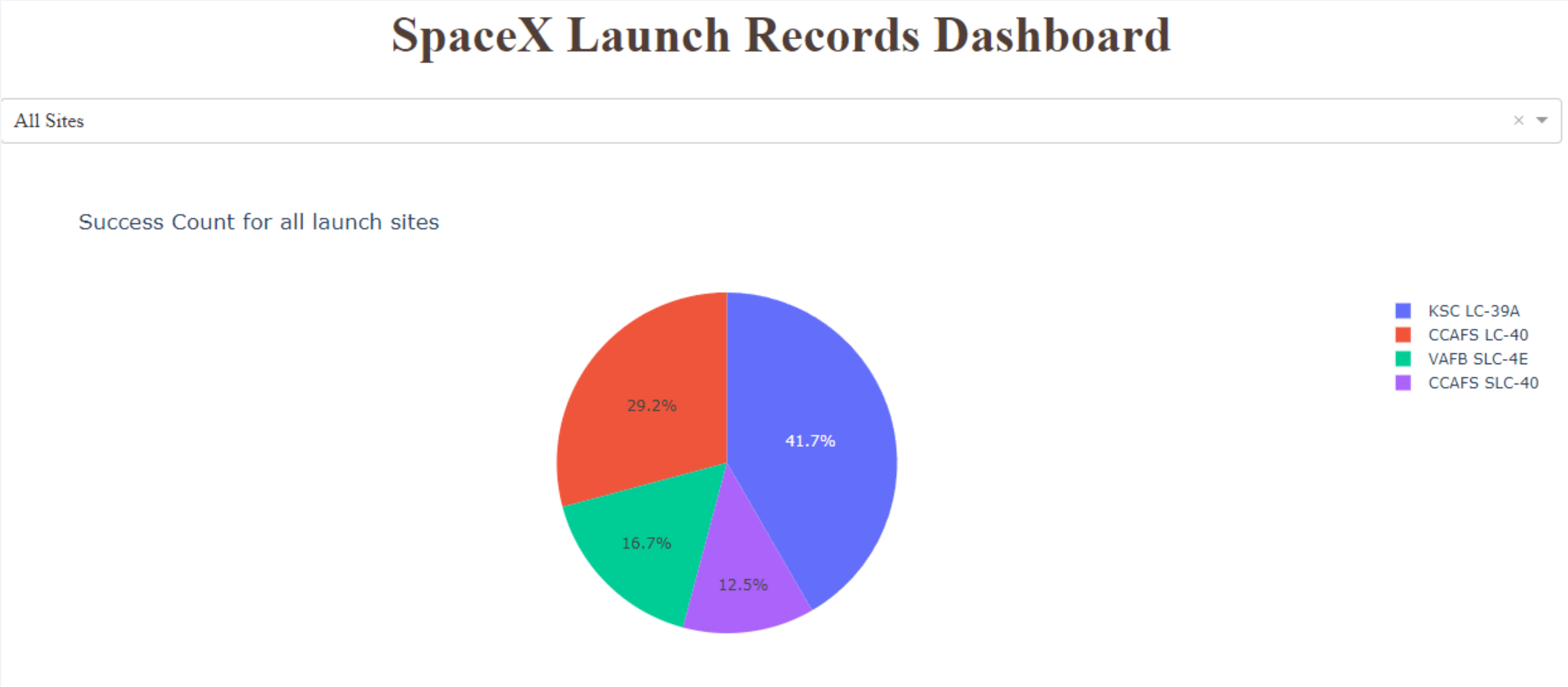


- Markers and Polylines to show Distances between a launch site to railways, highways, and cities and check correlations.

Section 4

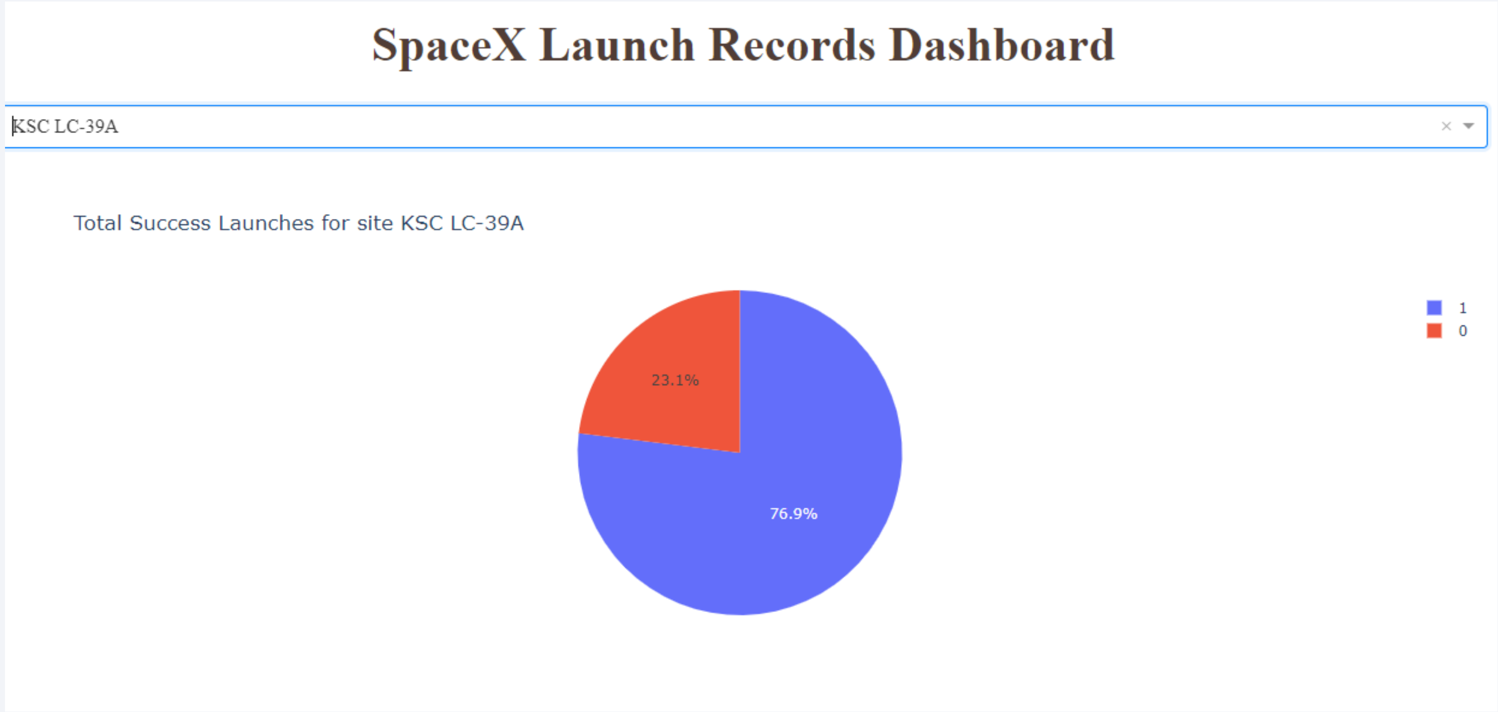
Build a Dashboard with Plotly Dash

launch success count for all sites



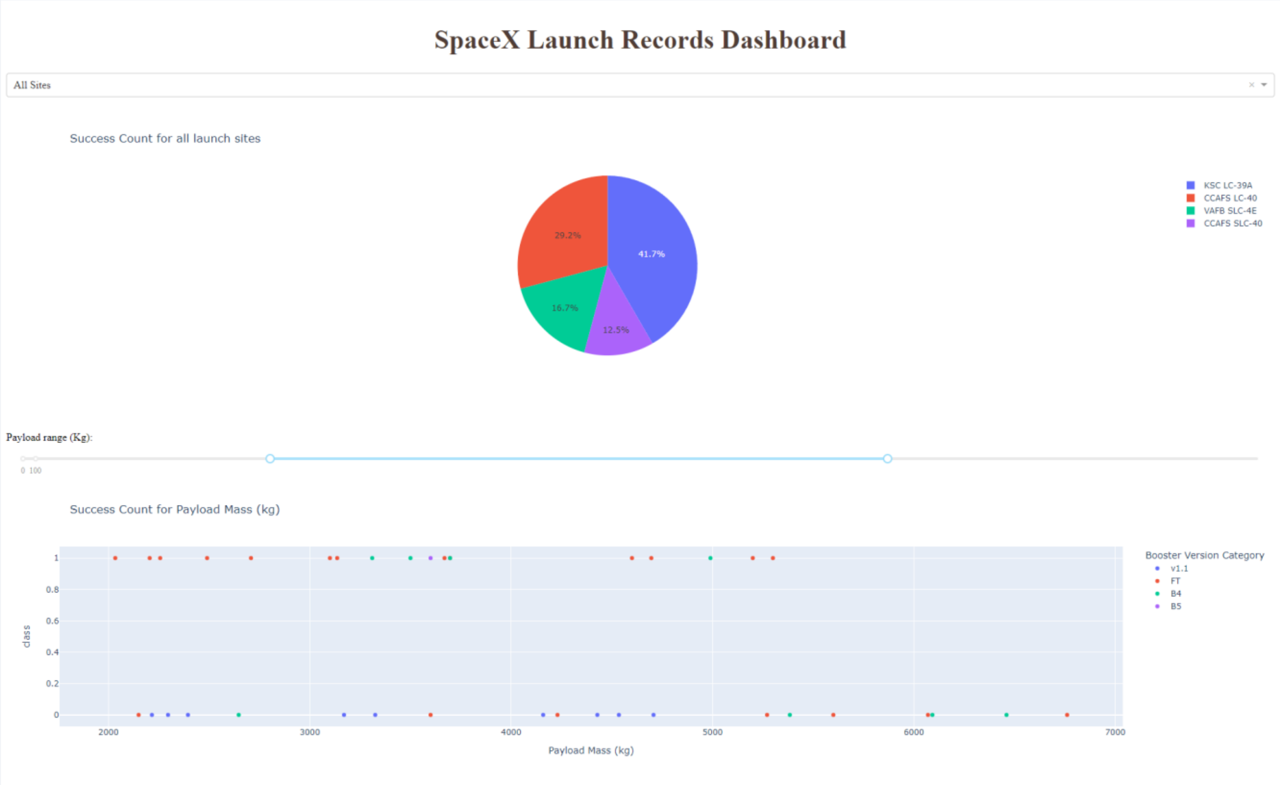
- KSC LC 39A has the most success count

the launch site with highest launch success ratio



- KSC LC 39A has the highest success rate

Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

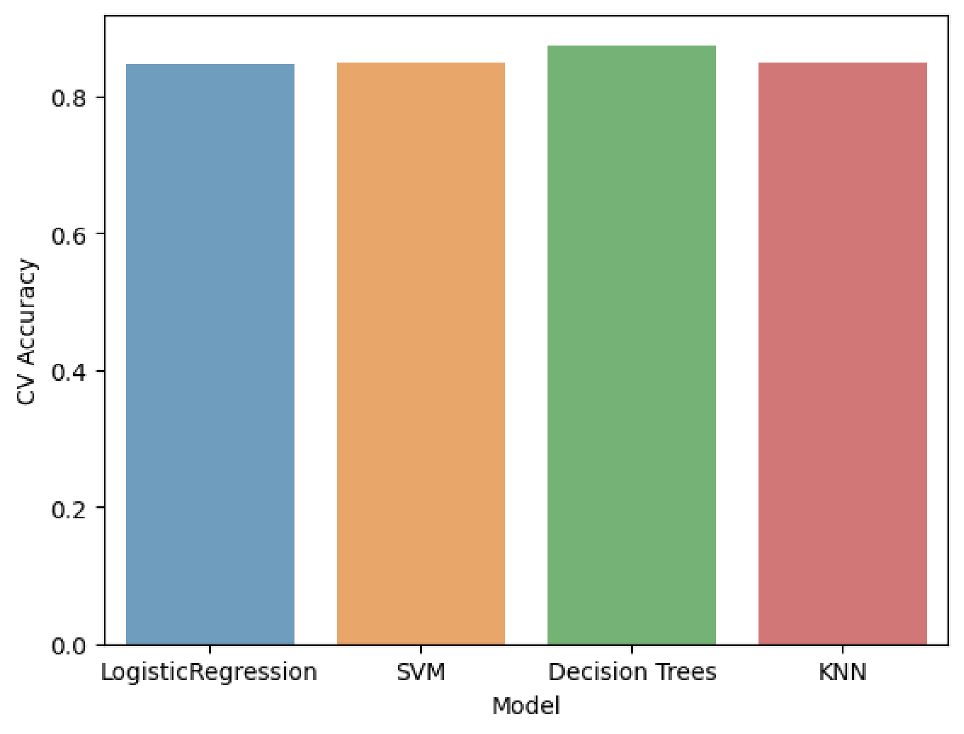


Section 5

Predictive Analysis (Classification)

Classification Accuracy

Model accuracy for all built classification models, in a bar chart

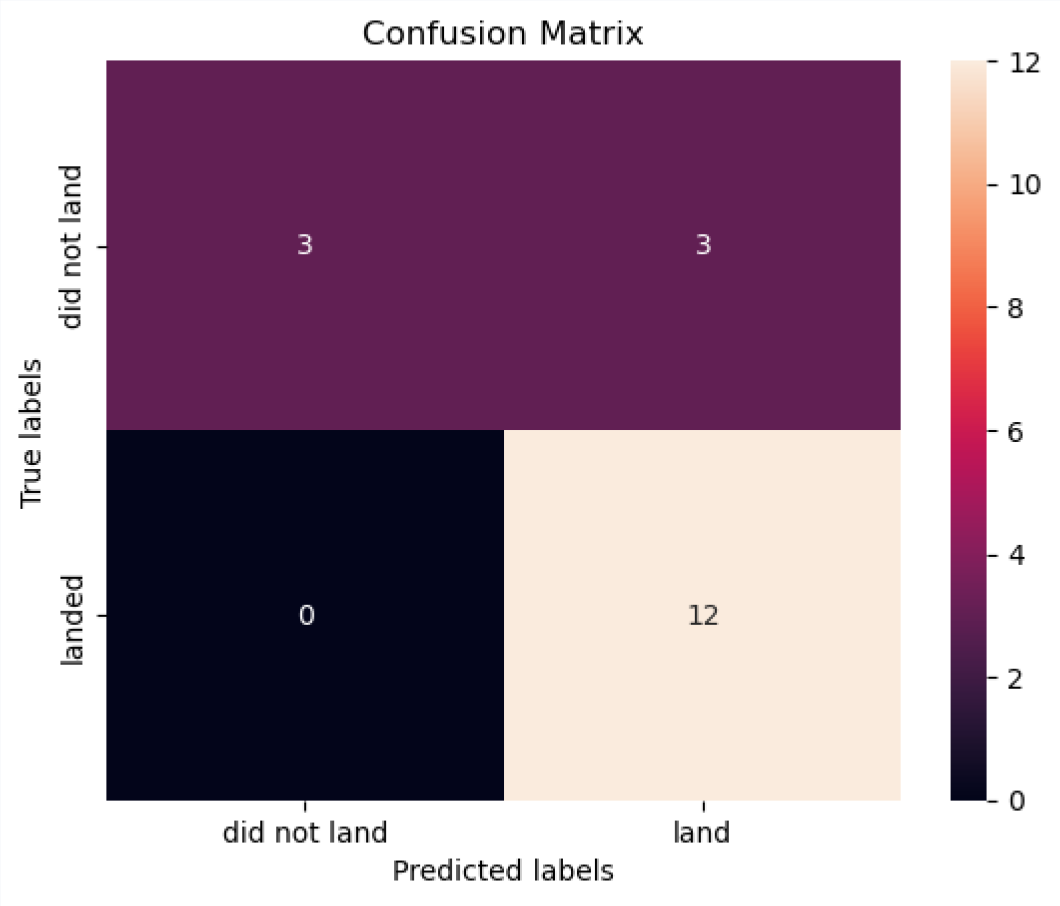


Model	CV Accuracy	Test Accuracy
LogisticRegression	0.846429	0.833333
SVM	0.848214	0.833333
Decision Trees	0.889286	0.833333
KNN	0.848214	0.833333

- Decision Trees has the highest Accuracy.

Confusion Matrix

Confusion matrix of the Decision Trees model



- Among 6 failure test data, 3 of them are correctly predicted as failure and 3 of them are not correctly predicted.
- Among 12 success test data, all of them are correctly predicted.

Conclusions

- Success rate since 2013 kept increasing till 2020 and reached plateau at around 80% success rate since 2017.
- Decision tree model shows the highest accuracy in cross validation while all model exhibits similar test accuracy score of 0.833.
- Decision tree model tend to successfully predict success outcome but not failure outcome.

Appendix

Data Collection API: https://github.com/jhjang101/testrepo/blob/f6b7548cd69adee212591bd4bb55a6f4e2915ed2/Data%20Collection%20API.ipynb

Data Collection with Web Scraping:
https://github.com/jhjang101/testrepo/blob/f6b7548cd69adee212591bd4bb55a6f4e2915ed2/Data%20Collection%20with%20Web%20Scraping.ipynb

EDA with Data Visualization:
https://github.com/jhjang101/testrepo/blob/f6b7548cd69adee212591bd4bb55a6f4e2915ed2/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL: https://github.com/jhjang101/testrepo/blob/f6b7548cd69adee212591bd4bb55a6f4e2915ed2/EDA%20with%20SQL.ipynb

EDA: https://github.com/jhjang101/testrepo/blob/f6b7548cd69adee212591bd4bb55a6f4e2915ed2/EDA.ipynb

Launch_site_location:
https://github.com/jhjang101/testrepo/blob/f6b7548cd69adee212591bd4bb55a6f4e2915ed2/lab_jupyter_launch_site_location.ipynb

Launch_site_location PDF:
https://github.com/jhjang101/testrepo/blob/f6b7548cd69adee212591bd4bb55a6f4e2915ed2/lab_jupyter_launch_site_location.pdf

spacex_dash_app: https://github.com/jhjang101/testrepo/blob/f6b7548cd69adee212591bd4bb55a6f4e2915ed2/spacex_dash_app.py

spacex_dash pdf: https://github.com/jhjang101/testrepo/blob/f6b7548cd69adee212591bd4bb55a6f4e2915ed2/spacex_dash.pdf

SpaceX_Machine Learning Prediction:
https://github.com/jhjang101/testrepo/blob/f6b7548cd69adee212591bd4bb55a6f4e2915ed2/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Thank you!